УНИВЕРЗИТЕТ У БЕОГРАДУ
ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Милан Н. Симаковић

СИСТЕМ ЗА НАДГЛЕДАЊЕ ПЕРФОРМАНСИ МРЕЖЕ КАБЛОВСКОГ ОПЕРАТОРА ЗАСНОВАН НА ТЕХНОЛОГИЈИ ВЕЛИКИХ ПОДАТАКА

докторска дисертација

Београд, 2022.

UNIVERSITY OF BELGRADE
SCHOOL OF ELECTRICAL ENGINEERING

Milan N. Simaković

BIG DATA SYSTEM FOR CABLE OPERATOR NETWORK PERFORMANCE MONITORING

Doctoral Dissertation

Belgrade, 2022.

ПОДАЦИ О МЕНТОРУ И ЧЛАНОВИМА КОМИСИЈЕ:

МЕНТОР:

др Зоран Чича, ванредни професор,
Универзитет у Београду, Електротехнички факултет

др Дејан Драјић, ванредни професор,
Универзитет у Београду, Електротехнички факултет

ЧЛАНОВИ КОМИСИЈЕ:

др Зоран Чича, ванредни професор,
Универзитет у Београду, Електротехнички факултет

др Дејан Драјић, ванредни професор,
Универзитет у Београду, Електротехнички факултет

др Предраг Иваниш, редовни професор,
Универзитет у Београду, Електротехнички факултет

др Ненад Јевтић, ванредни професор,
Универзитет у Београду, Саобраћајни факултет

др Милош Цветановић, ванредни професор,
Универзитет у Београду, Електротехнички факултет

Датум усмене одбране:_____

ЗАХВАЛНИЦА

Овом приликом изражавам искрену захвалност ментору, професору др Зорану Чичи, на великој и несебичној помоћи, подршци и инспирацији током читавих докторских студија. Такође, желим да се захвалим и ментору, професору др Дејану Драјићу, на великој помоћи током израде тезе и веома корисним саветима. Такође, захваљујем се и члановима комисије који су својим сугестијама допринели у коначној изведби ове дисертације.

Велику захвалност изражавам пријатељима на константној подршци и охрабривању током студија.

Коначно, неизмерну захвалност за безусловну подршку, посебно у најкритичнијим моментима школовања, дугујем својој породици, супрузи Ини, сестри Кристини, мајци Армине и оцу Новици без којих не би ни било ове дисертације.

СИСТЕМ ЗА НАДГЛЕДАЊЕ ПЕРФОРМАНСИ МРЕЖЕ КАБЛОВСКОГ ОПЕРАТОРА ЗАСНОВАН НА ТЕХНОЛОГИЈИ ВЕЛИКИХ ПОДАТАКА

САЖЕТАК:

Велики телекомуникациони оператори пружају сервис милионима корисника користећи комплексну и хетерогену телекомуникациону инфраструктуру која садржи огроман број различитих мрежних уређаја. Надгледање перформанси таквих мрежа је с једне стране веома битан ради постизања високог квалитета рада мреже и великог степена задовољства корисника, а с друге стране је веома изазован задатак који није лако испунити. Надгледање перформанси мреже подразумева периодично прикупљање података са сваког мрежног уређаја, њихово складиштење и даљу обраду. Коришћењем прикупљених података, мрежни оператори су у стању да ефикасно детектују, локализују и решавају проблеме у мрежи како би унапредили квалитет и перформансе. Пошто се ради о огромним количинама података потребно је користити технологију великих података. У овој дисертацији представљен је скалабилан и флексибилан систем за надгледање перформанси мреже кабловског оператора заснован на технологији великих података.

У првом делу дисертације је дат преглед алата отвореног кода који се користе у решењима заснованим на технологији великих података. Такође, представљени су и изазови надгледања перформанси мрежа са становишта великих података са посебним нагласком на кабловске мреже. У наставку дисертације је детаљно описан систем за надгледање перформанси мреже кабловског оператора, при чему су детаљно описани сви делови система, као и ток података кроз сам систем. Предложени систем је у стању да прикупи, складишти и обрађује податке са огромног броја уређаја. Скалабилност система пружа могућност једноставног проширења капацитета и процесорске снаге на сваком слоју архитектуре система у циљу подршке, како и самог проширења постојеће мреже, тако и даљих интеграција са другим доменима и мрежама. Како у мрежи постоје и уређаји са којих се не могу прикупљати подаци, представљен је алгоритам за процену стања таквих уређаја на основу података прикупљених из мреже. Додатно, предложен је механизам за детекцију и локализацију отказа у мрежи кабловског оператора. Предложени систем је успешно имплементиран у реалној кабловској мрежи, при чему су у дисертацији наведени и проблеми који су се појавили приликом имплементације. Систем се може применити и у другим системима са великим бројем уређаја (на пример, интернет ствари) где је неопходно прикупљати временске серије што је описано у другом делу дисертације. Изложени су и потенцијали за даљи развој предложеног система којима се може проширити функционалност система и изван основне функције надгледања перформанси мреже.

КЉУЧНЕ РЕЧИ: кабловска мрежа, телекомуникациони оператори, технологија великих података, надгледање перформанси, детекција кварова, локализација кварова, интернет ствари

НАУЧНА ОБЛАСТ: Техничке науке – Електротехника и рачунарство

УЖА НАУЧНА ОБЛАСТ: Телекомуникације – телекомуникационе мреже

УДК БРОЈ: 621.3

BIG DATA SYSTEM FOR CABLE OPERATOR NETWORK PERFORMANCE MONITORING

ABSTRACT:

Large telecom operators provide service to millions of users using a complex and heterogeneous telecommunication infrastructure comprising huge number of various network devices. Performance monitoring of such networks is very important for achieving high levels of performance and user satisfaction, but on the other hand it represents a very challenging task. Performance management requires a periodic data collection from all network devices, as well as collected data storing and processing. Using the collected data, telecom operators can efficiently detect, localize, and solve network problems to further improve network's quality and performance. Big data technology needs to be used given the huge amount of collected data. This thesis proposes scalable and flexible big data based system for performance monitoring of cable operator networks.

The first part of the thesis provides open-source big data tools overview. Also, big data challenges in network performance monitoring are given and discussed with special attention to cable networks. Next in thesis, the proposed big data system for cable network performance monitoring is explained in detail. All parts and components of the system are explained as well as the data flow through the overall system architecture. The proposed system is capable of collecting, storing and processing data from large number of devices. System scalability offers simple capacity and processing power expansion on every layer of system architecture to support future network expansions, but also to support future system integration with other domains and networks. Since there are network devices that are not capable of generating data for collection, an algorithm for performance estimation of such network elements based on collected data from other network devices is proposed. Moreover, a mechanism for failure detection and localization is proposed for cable networks. The proposed system is succesfully deployed in real cable network. Discussion regarding the problems that emerged during the system deployment is given in thesis. The proposed system can be used in other networks with a huge number of devices where time-series collection is required (e.g., Internet of Things) which is dicussed in the later part of thesis. Also, potentials for future expansions of supported system features is discussed providing use cases beyond performance monitoring.

KEYWORDS: cable network, telecom operators, big data technologies, performance monitoring, failure detection, failure localization, Internet of Things

SCIENTIFIC FIELD: Technical sciences – Electrical and computer engineering

SCIENTIFIC SUBFIELD: Telecommunications – communication networks

UDK NUMBER: 621.3

САДРЖАЈ

1. Увод

Експоненцијални развој комуникационих и интернет технологија отварају ново поглавље у развоју човечанства. Интернет технологије омогућавају брзу и једноставну размену информација, а ниска цена хардвера обезбеђује приступачност овог сервиса сваком човеку на планети. Према [1], преко 89% светске популације користило је услуге мобилне телефоније крајем 2021. године, а више од 66% користило је интернет [2]. Пандемија изазвана COVID-19 вирусом додатно је убрзала прихватање и коришћење дигиталних технологија [3,4] што је највише видљиво кроз одржавање школске наставе и рада на даљину. Услед значаја интернет технологија и велике конкуренције на тржишту, пружаоци телекомуникационих услуга су приморани да константно унапређују квалитет мреже како би обезбедили сервис високе доступности, високог протока и малог кашњења.

Приступ интернету крајњим корисницима је омогућен коришћењем различитих мрежа и медијума за пренос. Најпопуларније технологије за пренос података у данашње време користе мобилне, DSL (*Digital Subscriber Line*), HFC (*Hybrid Fiber-Coaxial*) и оптичке мреже. У оквиру ове тезе ће посебно бити разматране HFC мреже јер је управо за њих развијен систем за надгледање перформанси заснован на технологији великих података. HFC мреже су еволуирале од традиционалних мрежа кабловске телевизије како би биле у стању да понуде претплатницима шири спектар сервиса, то јест тзв. *Triple-play* сервисе (сервиси телефоније, телевизије и интернета). HFC мреже користе DOCSIS (*Data over cable service interface specification*) стандард. DOCSIS 3.1 подржава брзине до 10 Gb/s у *downstream* и до 1 Gb/s у *upstream* смеру [5]. DOCSIS 4.0 ће подржати чак и веће протоке и комуникацију у пуном дуплексу [6]. С тим у вези, HFC мреже, поред тога што већ пружају својим претплатницима услуге високог квалитета, очекује перспективна будућност обзиром на то да се активно ради на њиховом даљем унапређењу.

Велики телекомуникациони оператори пружају сервис милионима корисника. Телекомуникационе мреже таквих оператора имају велику и комплексну инфраструктуру, која се састоји од бројних мрежних елемената и линкова, како би се пружио сервис великом броју корисника. Увођење нових технологија и нових сервиса, додатно компликује постојећу инфраструктуру. Надгледање и одржавање таквих хетерогених, комплексних и великих телекомуникационих мрежа представља изузетно изазован посао. Мрежни елементи и корисничка опрема, CPE (*Customer-Premises Equipment*), могу пружити много корисних информација о свом стању, али и о стању мреже уколико би се прикупљали подаци са истих. Како би се омогућили сервиси високог квалитета корисницима, неопходно је обезбедити квалитетан рад мреже, а у ту сврху је неопходно поставити платформу за надгледање перформанси мреже. Слично као и сама телекомуникациона мрежа, неопходно је да ова платформа буде високо доступна. Платформа за надгледање перформанси мреже доноси између осталог и бржу детекцију кварова и решавање таквих проблема што свакако доприноси квалитетном раду мреже. Међутим, надгледање перформанси тако великих и комплексних мрежа, узевши притом разноврсност мрежних елемената, представља изузетно изазован посао. Платформа за надгледање перформанси мреже треба да прикупља различите типове података са уређаја који се налазе у телекомуникационој мрежи. Прикупљени подаци

се додатно обрађују како би се дошло до корисних информација. Добијени резултати (на пример, детекција загушења линка, детекција отказа у мрежи…) се користе за одржавање рада мреже на жељеном нивоу квалитета. Додатно, прикупљени подаци се могу обрадити са циљем даљег унапређења мреже у будућности. На овај начин, телекомуникациони оператори могу ефикасно планирати будућа проширења мреже као што су, на пример, инсталација линкова вишег капацитета у појединим деловима мреже у циљу спречавања предвиђених загушења, понуде нових сервиса клијентима, замена опреме итд.

Треба имати у виду да је количина прикупљених података огромна јер платформа за надгледање перформанси мора непрестано да надгледа мрежу и прикупља податке с великог броја уређаја. Традиционалне методе за складиштење и обраду огромне количине података су практично бескорисне услед ограничености ресурса и недостатка скалабилности [7]. Због немогућности складиштења огромне количине података и осталих хардверских ограничења, већина генерисаних података бива одбачена и анализа се врши само над подскупом података. Уколико би постојала могућност за обрадом читавог скупа података, добијени резултати би били знатно прецизнији. Како би решиле проблеме са складиштењем и обрадом огромне количине података задржавајући притом високе перформансе, представљене су технологије великих података [7,8]. У данашње време, многи телекомуникациони оператори имплементирају платформе засноване на технологији великих података како би истражили и максимално искористили прикупљене податке из њихових мрежа [9-12].

У овој дисертацији биће представљено иновативно системско решење за надгледање перформанси телекомуникационих мрежа уз конкретну примену у HFC мрежама. Важно је нагласити да предложено решење успешно оперише у једној реалној HFC мрежи што додатно потврђује резултате ове дисертације. Предложено решење покрива комплетан ток података, од прикупљања, преко складиштења и обраде, до конкретне примене. Како би се подржале велике количине података и обезбедио систем високе доступности и скалабилности, предложено решење је засновано на технологији великих података. Све компоненте предложене у овом решењу су отвореног кода, па самим тим није потребно додатно лиценцирање чиме се постиже велика уштеда приликом имплементације и одржавања. Поред платформе, у дисертацији ће бити представљен алгоритам за детекцију и локализацију кварова неинтелигентних уређаја (са којих се не могу директно прикупљати подаци) у HFC мрежама заснован на технологији великих података и подацима прикупљеним са интелигентних уређаја. Предложено решење се може користити не само у другим телекомуникационим мрежама, већ и у другим системима где се прикупљају и посматрају временске серије (на пример, IoT (*Internet of Things*) мрежама). Коначно, у дисертацији су представљене и додатне могућности платформе као и будући правци истраживања.

Дисертација је организована на следећи начин. У другом поглављу представљена је технологија великих података. Главне карактеристике ове технологије и изазови данашње обраде података су представљени кроз популарни ”5V” концепт. Поред тога, представљена је архитектура дистрибуираних система и дат преглед најбитнијих алата заснованих на овом принципу. Додатно, дискутована је могућност имплементације система заснованих на технологији великих података у различитим окружењима, од стандардног серверског, преко хибридног до решења заснованог потпуно на *cloud* технологијама. Коначно, дискутовани су тренутни изазови са којима се технологија великих података тренутно сусреће.

Преглед телекомуникационих мрежа дат је у трећем поглављу. Преглед обухвата анализу и дефиницију најпознатијих мрежа као и тумачење изазова приликом надгледања

њихових перформанси. Поред тога, представљен је преглед литературе у смислу тренутно постојећих решења надгледања перформанси телекомуникационих мрежа као и употреба технологије великих података у овим мрежама. У четвртом поглављу представљене су HFC мреже. Поред прегледа ових мрежа и њихових карактеристика, представљени су сви мрежни елементи и појмови релевантни за ово истраживање. Поред тога, дат је преглед изазова за надгледање перформанси ових мрежа представљен кроз ”5V” концепт великих података.

Пето поглавље представља главни допринос дисертације. Наиме, у овом поглављу представљена је архитектура система за надгледање перформанси HFC мрежа заснована на технологији великих података. Најпре су дефинисани циљеви платформе на основу претходно анализираних изазова. Након тога, представљена је слојевита архитектура платформе уз опис сваке од компоненти као и сваког слоја понаособ. Поред тога, дат је преглед тока података кроз систем, од прикупљања, преко складиштења и обраде, до употребе. Додатно, представљене су све релевантне метрике у HFC мрежи и дат је предлог за њихово прикупљање. За ефикасан упис и брзо читање, представљена је специјализована шема података. Посебна пажња посвећена је колектору података обзиром на то да исти успоставља комуникацију са великим бројем уређаја, а самим тим представља потенцијално уско грло. У циљу употребе прикупљених података дат је преглед свих врста агрегација које предложена платформа подржава. Поред тога, дат је предлог агрегације података за процену стања неинтелигентних уређаја у мрежи. Додатно, представљени су изазови који су се појавили током имплементације система у једној реалној HFC мрежи уз предлог конкретних решења. Коначно, посебна пажња посвећена је безбедности података и приватности корисника.

У шестом поглављу представљен је механизам за детекцију и локализацију отказа неинтелигентних елемената у HFC мрежама. Најпре је представљен опис проблема са којим се сређу кабловски оператори. Додатно, дат је преглед тренутних решења у литератури на ову тему. Након тога, представљен је развијени алгоритам за детекцију и локализацију отказа неинтелигентних елемената. Алгоритам се састоји из два дела, детекција и локализација проблема. Детекција проблема се фокусира на идентификацију тренутка дешавања отказа. Обзиром на то да број активних корисника мреже константно флуктуира, овај изазов није тривијалан. Локализација проблема фокусира се на детекцију уређаја који је узрочник отказа мреже или њеног дела. Предложени алгоритам упоређен је са тренутно доступним решењима у литератури.

Седмо поглавље се бави додатним могућностима предложене платформе. Једна од могућности платформе представља додатну анализу прикупљених података у циљу откривања нових међузависности података, а самим тим и примена истих за решавање нових типова проблема. Поред тога, дискутовано је о интеграцији нових домена. Домени интеграције представљају прикупљање података с других типова уређаја или прикупљање другог типа података са постојећих уређаја (на пример, WiFi). Посебна пажња дата је примени машинског учења за аутоматску детекцију деградације перформанси сигнала. Обзиром на то да платформа обезбеђује централизовано место за складиштење велике количине података, исти могу послужити за тренирање и евалуацију модела машинског учења. Коришћење ових модела може помоћи у проактивном сервисирању мреже, тј. спречавању отказа.

Осмо поглавље представља будуће правце развоја. Обзиром на то да је платформа првенствено развијена за ефикасно складиштење временских серија, један од праваца ће бити примена постојеће платформе за прикупљање метрика из других мрежа, конкретно

мобилних, DSL и IoT. Посебна пажња ће бити дата интеграцији IoT и HFC мрежа у циљу развоја специјализованог решења паметних кућа и паметних градова. Последње поглавље сумира резултате дисертације и њене доприносе.

2. ТЕХНОЛОГИЈА ВЕЛИКИХ ПОДАТАКА

Количина генерисаних података се константно повећава услед експанзије интернета и континуалног развоја рачунарских технологија. Велика количина података као и њихова брзина генерисања доводи до потешкоћа у обради истих. Традиционалне методе складиштења и обраде података постале су практично бескорисне због недостатка ресурса и скалабилности и нису у стању да испрате експоненцијални тренд раста података. Због поменутих ограничења, већина генерисаних података бива одбачена што доводи до тога да се анализа података изводи само на малом узорку података. Анализом целокупног скупа података могуће је остварити бољи увид у податке, а самим тим генерисати боље и квалитетније закључке. У циљу омогућавања складиштења и обраде велике количине података, IT (*Information Technology*) свет улази у нову еру технологије великих података (енгл. *big data*). У наставку дисертације ће се за технологију великих података користити термин "биг дата" обзиром да се исти, због велике популарности, одомаћио у српском језику.

Биг дата технологија је претходних година била веома актуелна и атрактивна тема у свету и у складу с тим се развијала (и развија) изузетно брзо. Практични проблеми који се јављају у пракси и њихова разноврсност управљају динамиком и развојем различитих техничких решења у овој области. Термин биг дата се први пут помиње још 1990-их у тренутку када се препознаје велики тренд раста података као предстојећи изазов [13]. Овај термин је сам по себи превише апстрактан. Обзиром на то да је ова грана информационих технологија релативно нова, не постоји јединствена дефиниција која би једнозначно дала објашњење шта биг дата представља. Према Гартнеру, биг дата представљају информациона средства са особинама волумена, брзине и/или разноврсности која захтевају иновативне начине за обраду информација за унапређено откривање вредности са циљем бољег доношења одлука и аутоматизације процеса [14]. IBM (*International Business Machines*) сматра да је биг дата јединствено дефинисана кроз "5V" парадигму – ***Volume***, ***Variety***, ***Velocity***, ***Veracity*** и ***Value***. Неки везују биг дата за конкретне вредности као што су терабајти и зетабајти. Термин биг дата треба везивати за ону количину података којом у датом тренутку није могуће управљати коришћењем традиционалних метода и алата услед неких од поменутих "5V" изазова. Количина података је сама по себи релативна категорија и не може се везати за конкретне нумеричке вредности. Оно што се сматра биг дата конкретно зависи од компаније и њене могућности за управљање подацима. За неке организације суочавање са гигабајтима података по први пут може бити знак за разматрање биг дата технологија, док за друге десетине, па чак и стотине терабајта се сматрају за значајну количину података [15]. Такође, оно што се у данашње време сматра биг дата, већ сутра може бити релативно мала количина података.

2.1. Карактеристике биг дата

Главни изазови данашње обраде података су представљени кроз "5V" концепт, што уједно представља и њене карактеристике [16]. "5V" концепт се односи на термине *Volume* (волумен), *Variety* (разноврсност), *Velocity* (брзину), *Veracity* (поверење) и *Value* (вредност) о чему ће бити више речи у наставку.

Константан тренд раста броја уређаја и њихово усавршавање доводи до константног повећања количине података како на локалном, тако и на глобалном нивоу (*Volume* – волумен). Уређаји као што су мобилни телефони, камере, аутомобили, телевизори, машине у индустрији и здравству, доприносе расту података које се може запазити у данашње време [17]. На једном упрошћеном примеру из авио индустрије се може видети проблем велике количине података. Авио компаније прикупљају податке са авиона како би проучавали његово понашање у току рада, вршили даља унапређења, детектовали проблеме у раду његових компоненти и на тај начин спречили катастрофалне догађаје. У једној летилици се у просеку налази око 6000 сензора. Према наводима Stephen Brobst-a, техничког директора Teradata, један боинг авион генерише 10 TB података по сваком млазном мотору за 30 минута рада [18]. С тим у вези, за једночасовни лет, летилица Боинг 737 са два млазна мотора, генерише око 40 TB података. Другим речима, само један лет трајања једног сата захтева обраду 40 TB података. Према [19], у свету се на дневном нивоу оствари око 100000 комерцијалних летова па је очигледно да се ради о огромној количини података чак и ако се посматра само на дневном нивоу. На основу овог једноставног примера се може видети битност *Volume* изазова у биг дата дефиницији.

Velocity се односи на брзину којом се подаци генеришу. Генерисани подаци, како малих тако и великих брзина, морају бити обрађени у очекиваним временским оквирима. Ова димензија отвара нову грану у биг дата технологијама која се бави обрадом тзв. токова података (енгл. *data streaming*) у реалном времену. У оквиру ове димензије, могу се идентификовати два изазова који се устаљено срећу у пракси.

У многим случајевима, нема довољно времена за складиштење података и њихову накнадну обраду. Уместо тога, анализу и обраду података је неопходно обавити у реалном времену, тј. податке је неопходно обрадити одмах након што су исти генерисани. Циљ је смањити што је могуће више време од тренутка када је податак генерисан, до тренутка када је исти обрађен. У већ поменутом примеру о прикупљању података са летилице, лако је закључити да се обрада података мора обављати у току лета како би се правовремено идентификовали потенцијални проблеми и спречила катастрофа. Уколико би се подаци у оваквом примеру обрађивали накнадно, генерисани резултати у сценарију катастрофалног догађаја не би имали вредност јер се несрећа већ догодила. У оваквом примеру, узевши у обзир количину података коју је потребно обрадити у унапред дефинисаном временском прозору, процесорска моћ представља лимитирајући фактор.

Још један изазов који спада под *Velocity* димензију тиче се експлозивних налета података (енгл. *data bursts*). Експлозивни налети података не представљају само изазов приликом складиштења, већ и приликом обраде података. На пример, једна базна станица се налази у близини градске арене. У нормалним околностима, ова базна станица покрива подручје око арене уобичајеним просечним капацитетом. У складу с тим, ова базна станица генерише просечну, и унапред очекивану, количину података. С друге стране, у току манифестација које се одржавају у градској арени (као што су, на пример, концерти, изложбе, спортска дешавања), подручје поменуте базне станице посећује знатно већи број корисника. Последично, количина података генерисаних у зони базне станице се експлозивно повећава стотинама, па чак и хиљадама пута за време трајања манифестације. Након завршетка манифестације, саобраћај се враћа у своје просечно стање. Традиционални системи нису у стању да изађу у сусрет оваквим проблемима што је сигнал да је потребно потражити решење у биг дата технологији.

У данашње време, постоји велики број различитих извора података. Сви подаци се могу класификовати у три категорије: структурирани, полуструктурирани и неструктурирани. Структурирани подаци су они код којих су формат и структура унапред познати (на пример, базе података или CSV (*Comma-Separated Values*) фајлови). Полуструктурирани подаци су они код којих постоји неко правило у формату и начину слагања информација, али због превелике флексибилности их није могуће сврстати у структуриране изворе. Неки од полуструктурираних типова података су лог фајлови, XML (*Extensible Markup Language*) и JSON (*JavaScript Object Notation*) фајлови. Неструктурирани подаци су они који не поседују никакав унапред познат формат. Таквих фајлова је највише, најразноврснији су и представљају највећи изазов у обради. Неки од њих су документи, имејлови, веб странице, слике, видео садржај и многи други. Поред тога, узевши у обзир велики број различитих формата фајлова који данас постоји, обрада података се даље усложњава. У најчешћем броју случајева податке није могуће користити директно, већ је неопходно претходно извршити обраде. Додатно, корелација података добијених из различитих извора је у данашње време незаобилазна. Велика разноврсност и могућност складиштења, обраде и комбиновања различитих типова и структура података додатно повећава комплексност система што представља *Variety* изазов биг дата технологија.

Veracity представља фактор поверења (енгл. *confidence factor*) генерисаних резултата. Ова карактеристика се односи на несигурност података, неповерење у извор података и неповерење у процес који генерише резултате. Несигурност и неповерење у извор података се може десити услед различитих фактора и генерално је нешто што се у пракси свакодневно среће. Грешке у куцању текста, недостатак података због проблема у комуникацији, кашњење података су само неки од мноштва разлога због којих може доћи до проблема у подацима. У складу са тим, *Veracity* изазов се јавља када се није могуће потпуно ослонити на генерисане резултате. Обично, због недостатка ресурса, анализа се врши на случајном узорку података што доводи до несигурности и недовољно прецизних резултата. Због поменутих разлога, постоји потреба за посматрањем целокупног скупа података уместо узорка. С друге стране, целокупан скуп доноси биг дата проблеме по питању складиштења и обраде података.

Value (вредност) изазов се односи на способност претварања података у конкретну употребну вредност. Лако је упасти у замку и користити биг дата технологије за складиштење података без већег разумевања шта са њима радити и на који начин их искористити у циљу унапређења пословања. Уколико није могуће издвојити вредност из података, поставља се питање чему уопште и њихово иницијално прикупљање. Из велике количине разноврсних података неопходно је извући корисне информације. Ово представља посао дата аналитичара и често може бити веома изазован.

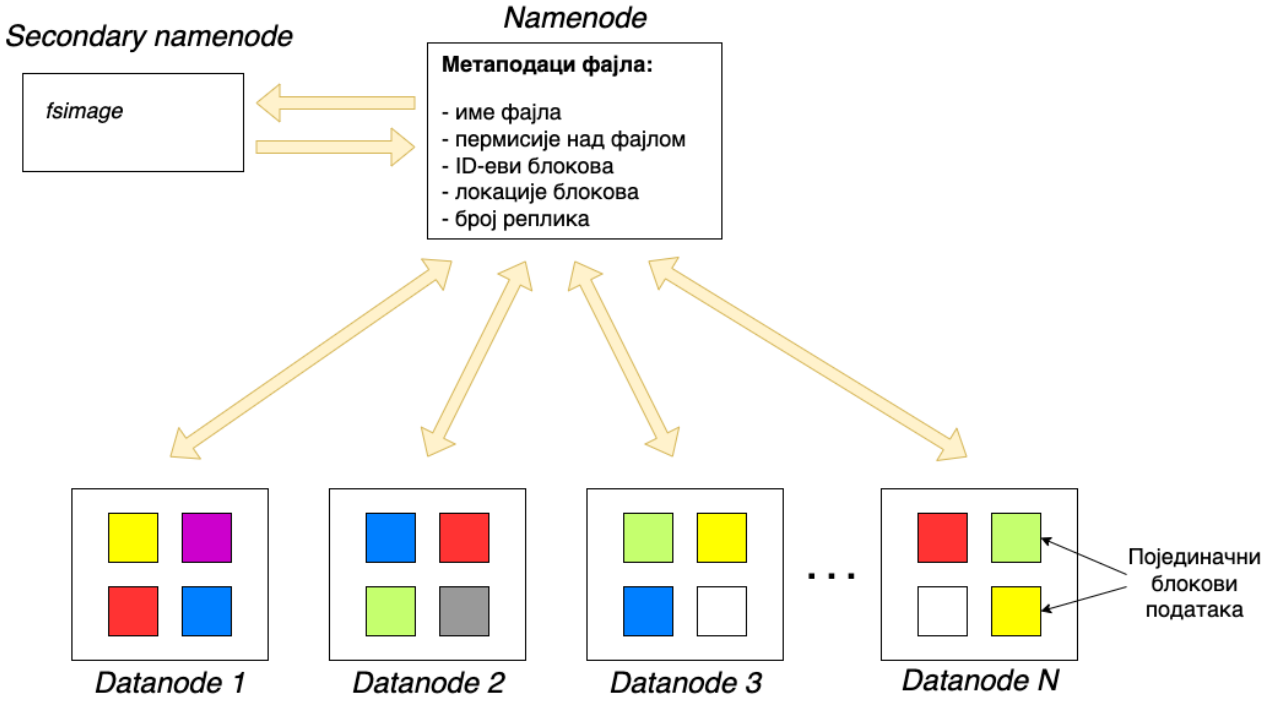
2.2. Архитектура и алати

Прве техничке реализације биг дата платформе настају након што Google објављује научни рад познат као ”MapReduce” 2004. године [20]. MapReduce концепт пружа модел за паралелно програмирање који је предвиђен за обраду огромних количина података. У овом раду представљена је парадигма обраде података у два корака, *map* и *reduce*. Објављени рад стиче велику популарност и бива прихваћен од стране Apache Hadoop фондације у Hadoop пројекат. Чињеница да су биг дата технологије у успону инспирисала је и друге компаније да развију и понуде своја биг дата решења. Сва та решења у ствари представљају модификацију оригиналне Apache Hadoop платформе. Неки од најпознатијих произвођача су Microsoft [21], Amazon [22], Google [23], IBM [24], Oracle [25], Cloudera [26] и многи други.

Hadoop је пројекат отвореног кода (енгл. *open-source project*) основан од стране Apache Software Foundation. Он се састоји од мноштва малих потпројеката који припадају категорији инфраструктуре за дистрибуирано програмирање [27]. Биг дата је дистрибуирана, скалабилна платформа за складиштење и обраду огромних количина података генерисаних великим брзинама које није могуће обрадити на традиционалан начин. То је комбинација различитих технологија и алата који међусобно сарађују. Заједно, ти алати сачињавају Hadoop екосистем. Узевши у обзир да је платформа у константном развоју, број ових алата се непрестано повећава. Сва друга комерцијална решења која се у овом тренутку могу наћи у понуди су проистекла од иницијалног Apache Hadoop-а. Преглед биг дата технологија и алата дат је у [28]. У овом раду представљена је Apache фондација и дат преглед најважнијих алата из Hadoop екосистема. Поред тога, представљене су и различите дистрибуције биг дата система заједно са серверским и *cloud* имплементацијама. С тим у вези, у наставку ће бити представљене основне, уједно и најважније биг дата компоненте Hadoop екосистема. Поред тога, биће дат опис компоненти која ће бити коришћене за предложени систем за надгледање перформанси мреже кабловског оператора.

Један од две најважније компоненте Hadoop екосистема је *Hadoop Distributed File System* (HDFS). HDFS је радни оквир (engl. *framework*) који обезбеђује јединствени дистрибуирани фајл систем преко више различитих сервера који заједно формирају кластер за чување података [29]. На слици 2.2.1 приказана је архитектура HDFS система. HDFS је заснован на мастер/слејв (енгл. *master/slave*) архитектури где *Namenode* представља мастера, а *Datanode*-ови представљају слејвове (којих има *N* у систему). *Namenode* дели податке, које је потребно уписати на HDFS, на блокове и управља њиховим складиштењем. Секундарни *Namenode* складишти информације о трансакцијама унутар фајл система, а које се користе за опоравак фајл система у случају метаподатака који садрже физичке грешке (*corrupted metadata*). На тај начин, *Namenode,* заједно са секундарним *Namenode*-ом, решава проблем јединствене тачке отказа (енгл. *Single Point of Failure* (SPOF)). *Datanode*-ови се користе за физичко складиштење блокова података. Главна предност овог фајл система је, у поређењу с другим, могућност складиштења великих фајлова који не могу бити сачувани ни на једном другом медијуму услед хардверских ограничења. Поред тога, овај фајл систем обезбеђује високу доступност и скалабилност. На овај начин, проширење капацитета је могуће додавањем нових *Datanode*-ова у кластер чиме се може, теоретски, постићи неограничен капацитет. Висока доступност, отпорност на грешке и отказ хардвера омогућена је репликацијом података. Сваки блок података на HDFS-у је сачуван у неколико копија (конфигурациони параметар фактор репликације је најчешће n=3) на различитим *Datanode*-овима. То значи да у сваком тренутку истовремено може отказати n-1 *Datanode*-ова, а да подаци и даље буду доступни. Поред тога, HDFS имплементира контролну суму за све сачуване податке како би обезбедио интегритет истих. Сви наведени механизми заједно пружају високу доступност и на хардверском и на софтверском нивоу. Укупан капацитет пропорционалан је агрегираном капацитету дефинисаном на *Datanode*-овима, а обрнуто пропорционалан фактору репликације у складу са једначином 2.2.1. Уколико је потребно, нови *Datanode*-ови се могу додати у кластер без гашења читавог система чиме се постиже већи капацитет и процесорска моћ.

$$HDFS_{capacity} = \frac{\sum_{i=1}^{N} Datanode_{capacity\,i}}{n} \tag{2.2.1}$$

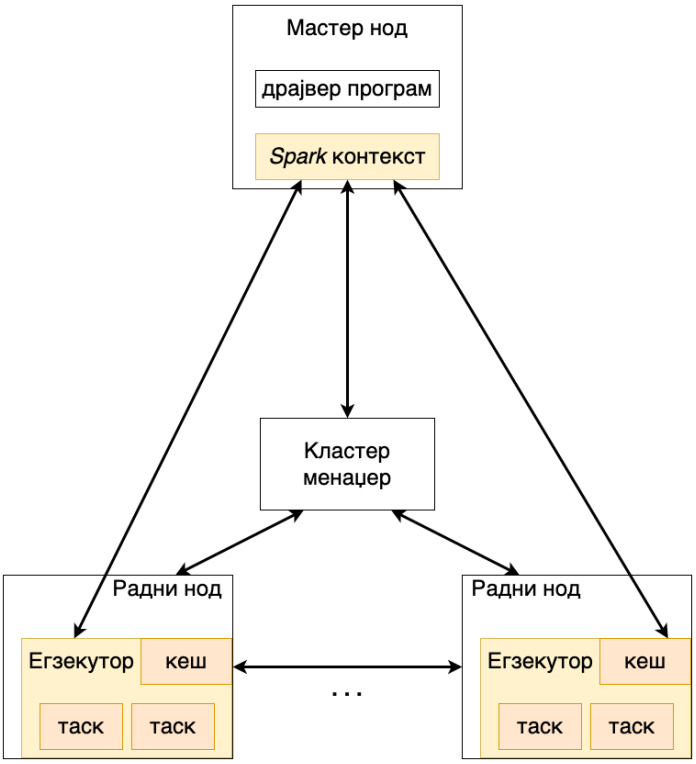


Слика 2.2.1. Архитектура HDFS система.

Следећа главна компонента Hadoop биг дата кластера је MapReduce. То је програмски модел за обраду великих скупова података који користи паралелни, дистрибуирани алгоритам на кластеру [30]. У данашње време, MapReduce је замењен YARN-ом (*Yet Another Resource Negotiator*, MRv2) који представља ажурирану верзију MapReduce-а (MRv1). Apache YARN представља менаџера ресурса и пратиоца послова на кластеру. YARN раздваја слој за управљањем ресурса од слоја за обраду података. Ова компонента омогућава различитим софтверским алатима и различитим програмима да деле заједничке, дељене, ресурсе. YARN динамички заузима ресурсе и организује редослед извршавања послова [31]. Биг дата обрада података реализована је уз помоћ *map/reduce* парадигме. Током *map* процеса, сваки *Datanode* изводи калкулације над блоковима података од интереса. На овај начин се постиже дистрибуирано програмирање и теоретски неограничена процесорска моћ. Резултати *map* дела се прикупљају на *Namenode*-у где се врши *reduce* део програма, то јест врши се агрегација *map* резултата и калкулише се коначан резултат. MapReduce је компонента базирана на Java програмском језику. Међутим, како би се обезбедила флексибилност, креирани су интерфејси који омогућавају коришћење других програмских језика за писање MapReduce алгоритама. Ови интерфејси се ослањају на основну MapReduce компоненту. Неки од најпознатијих алата су Hive (сличан SQL-у (*Structured Query Language*)), Pig (апстрактно програмирање на високом нивоу), Pydoop (python), JAQL (*Java Query Language*).

Spark је још један алат за обраду података. У односу на класичан MapReduce, Spark представља алтернативу за обраду података. Због своје природе, Spark је једноставнији за коришћење и постиже боље перформансе од 10 до 100 пута у односу на претходне генерације (у односу на одређене MapReduce апликације) [30]. С тим у вези, овај радни оквир је, у данашње време, један од најпопуларнијих за дистрибуирану обраду података. За разлику од MapReduce-а, Spark је концентрисан за обраду података у меморији чиме се постижу поменуте супериорне перформансе. На овај начин се добија уштеда у времену јер се избегава записивање међурезултата на физички медијум. Spark може да ради на различитим окружењима као што су Hadoop, Apache Mesos, Kubernetes или као независан кластер [32].

Архитектура Spark-а, приказана на слици 2.2.2., слична је архитектури HDFS-а. Мастер нод садржи драјвер програм који управља апликацијом креирањем Spark контекст објекта. Spark контекст објекат сарађује са кластер менаџером (YARN у предложеном решењу) како би добио ресурсе за обављање задатака. Радни (енгл. *worker*) нодови извршавају задатке (енгл. *tasks*) и враћају резултате мастер ноду. Обрада података у меморији омогућава Spark-у да се, поред класичне обраде, бави обрадом података у реалном времену (стримовима података). Поред тога, због своје велике популарности и једноставности коришћења, овај алат се користи и за обраду структурираних података (Spark SQL) као и за машинско учење (Spark Mlib) [32]. Како би писање програма било једноставније, Spark обезбеђује једноставне API-је (*Application Programming Interface*) у Scala, Java, Python, R и SQL програмским језицима [32]. Додатно, постоји могућност коришћења Spark-а путем интерактивних *shell*-ова у Scala и Python-у чиме се постиже још један ниво једноставности приликом обраде великих скупова података [30].



Слика 2.2.2. Архитектура Spark-а.

Обзиром на то да је Hadoop иницијално развијен као систем за складиштење података и њихову пост-обраду, он није био у стању да одговори на изазове по питању обраде и складиштења података у реалном времену. Такође, постоји велики број структурираних извора података који се због своје величине не могу чувати у традиционалним базама података. Чак и да је могуће сачувати све податке у релационој бази, величине табела би биле огромне, а перформансе читања би биле лоше. Постоје поједина професионална решења која би могла да парирају овом захтеву (на пример, IBM Netezza [33], Teradata [34], Vertica [35]), али је цена лиценци ограничавајући фактор [36]. Због тога, Hadoop породица је обогаћена системима за складиштење структурираних података. Један од таквих система је Apache HBase. HBase је не-релациона (NoSQL) дистрибуирана база података. Ова база података користи HDFS као фајл систем за складиштење табела. Он представља биг дата компоненту која служи за складиштење временски осетљивих (енгл. *time-sensitive*)

структурираних података. Према [37], HBase је дизајнирана на начин да оперише великим табелама, са милионима колона и милијардама редова. Додатно, HBase је оптимизована за насумичан приступ приликом уписа и ишчитавања података. Очигледно, HBase је погодна база за процесуирање података у реалном времену као што су временске серије. Додатно, високоперформантни случајни приступ подацима обезбеђује читање података са минималним кашњењем што је у овој дисертацији битно да се задовоље захтеви предложеног система по питању перформанси читања.

OpenTSDB (TSDB - *Time Series Database*) је дистрибуирана база података специјализована за временске серије. OpenTSDB поједностављује процес складиштења и анализирања велике количине временских серија генерисаних од стране уређаја (на пример, сензора, сервера, апликација). OpenTSDB користи HBase сервисе за упис и читање података [38]. HBase користи високоперформантну шему табела која је оптимизована за брзе агрегације података сличних временских серија и оптимизована је по питању заузећа простора. Коришћењем приступних тачака OpenTSDB-а (HTTP (*Hypertext Transfer Protocol*), API, TELNET (*Teletype Network*) или уграђени софтверски веб интерфејс), подацима се може приступати директно, без потребе за приступ HBase-u. Више различитих инстанци OpenTSDB-а може бити инсталирано на више различитих сервера како би се постигла висока доступност. OpenTSDB инстанце користе заједничке HBase табеле и раде независно једна од друге. Додатно, овакав приступ обезбеђује дистрибуцију оптерећења читања и писања података. Према [38], један податак се састоји од имена метрике, времена у UNIX (*Uniplexed Information and Computing System*) формату, вредности метрике и скупа тагова који су представљени у виду *key/value* парова који детаљније описују очитану метрику.

Биг дата екосистем је растао током времена и постало је компликовано надгледати рад свих његових компоненти. Због тога, развијен је Zookeeper. Apache Zookeeper је систем који се користи за координацију и управљање кластера и његових компоненти обезбеђујући на тај начин високу доступност сервиса. То је централизован сервис за надгледање конфигурација, именовање и обезбеђивање синхронизације у дистрибуираним сервисима [39]. Zookeeper поједностављује процес развоја, чини га агилним и на тај начин потпомаже развоју робусних имплементација кластера и система [30].

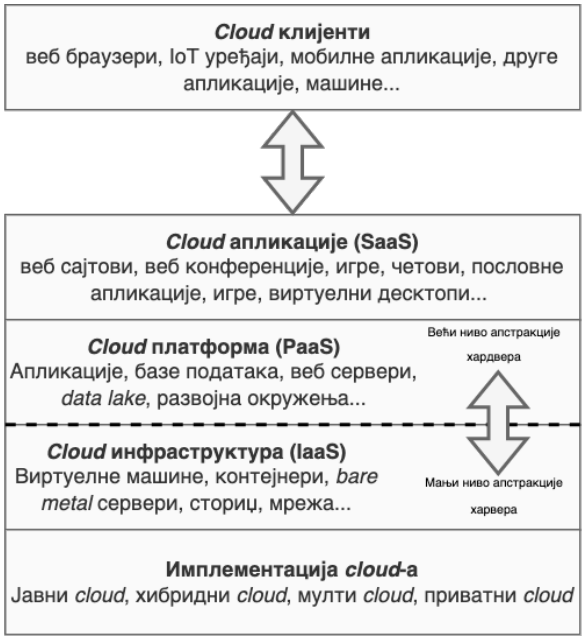
2.3. *Cloud* технологије

Cloud computing представља могућност коришћења рачунарских ресурса, посебно за складиштење и обраду података, у виду сервиса. *Cloud* представља систем за пружање поменутих услуга најчешће дистрибуиран на више локација (дата центара). Ови системи не представљају ништа друго него сервис развијен на биг дата технологијама који се корисницима нуди као инфраструктура за развој сопствених решења. Он елиминише потребу за надгледањем хардвера, додељеног простора и софтвера. Коришћење биг дата технологија на свом персоналном хардверу и њихово одржавање може бити веома захтевно [40]. Данашњи велики *cloud* провајдери (Microsoft Azure [21], AWS (*Amazon Web Services*) [22], Google Cloud [23], IBM [24]) нуде готова биг дата решења на *cloud*-у.

Постоји доста предности у оваквом приступу у односу на класична серверска решења. Једна од најважнијих предности је агилност и флексибилност по питању ширења инфраструктуре. У сваком тренутку се може користити минимално потребна количина ресурса, једноставно додавати нови, уклањати постојећи и на тај начин вршити оптимизацију трошкова. Већи број сервера дистрибуираних на различите локације обезбеђују високу поузданост, доступност и брз опоравак од катастрофалних догађаја (енгл. *disaster recovery*).

Коришћењем *cloud* сервиса није потребно поседовање и одржавање хардвера због чега компанија може да усмери све своје ресурсе на развијање нових сервиса чиме се постиже већа продуктивност.

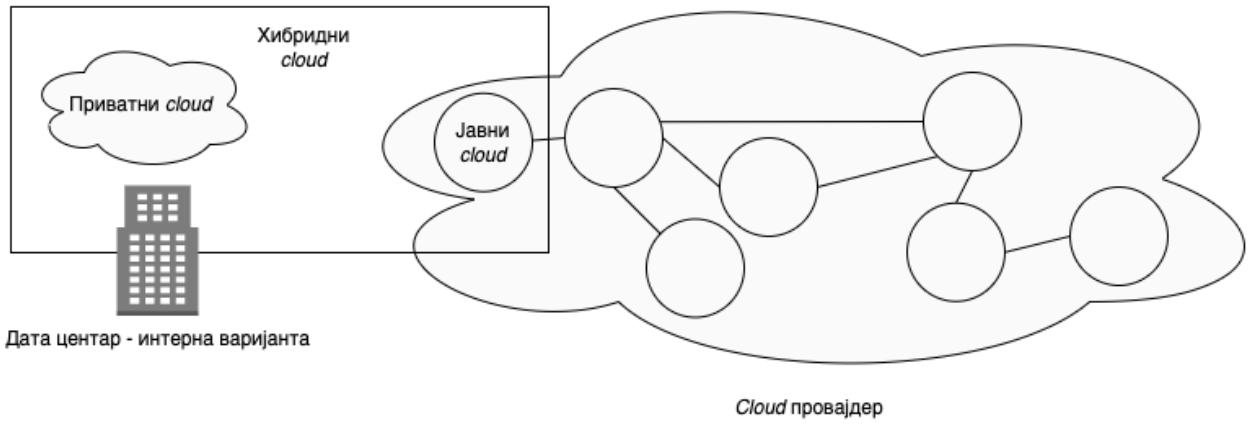
Постоје три стандардна сервисна модела за коришћење *cloud* технологија: *Infrastructure as a Service* (IaaS), *Platform as a Service* (PaaS) и *Software as a Service* (SaaS) [41]. Ови модели нуде побољшану апстракцију за коришћење *cloud* сервиса чиме се постиже још један ниво флексибилности. Сервисни модели приказани су у виду слојевите структуре на слици 2.3.1. Иако су ови слојеви на слици представљени у виду стека, њихово појединачно коришћење је потпуно независно. На пример, клијент може користити IaaS сервисе имплементиране на хардверу без да претходно треба сам да подиже инфраструктуру (тај део ће бити урађен од стране *cloud*-а аутоматски). IaaS се односи на физичку инфраструктуру која се може користити као сервис. На пример, на захтев клијента *cloud* хипервизор креира виртуелну машину која му се додељује на коришћење. PaaS представља могућност коришћења платформе као сервис. Корисник PaaS-а не управља хардвером и оперативним системом на ком се налази платформа [41]. SaaS користи софтвер као сервис. На овај начин, клијенту је загарантован високо доступан и перформантан софтвер без потребе за управљањем хардвером, оперативним системом и самим софтвером. У последње време све више постаје популарна парадигма *Function as a Service* (FaaS) или обрада ”без сервера” (енгл. *serverless computing*). Идеја ове парадигме се састоји у извршавању кода као сервис без икаквог надгледања хардвера. Приликом креирања захтева за извршавање, *cloud* аутоматски креира ресурсе, покреће код, скалира ресурсе (уколико је неопходно услед рачунарске комплексности кода који се покреће) и враћа резултате.



Слика 2.3.1. *Cloud* сервисни модели.

Постоји неколико различитих варијанти имплементације *cloud*-а. Приватни *cloud* представља коришћење платформе која се потпуно налази на страни корисника (компаније). Комплетна имплементација овог *cloud*-а се врши на серверима компаније и нема никакве везе са јавним *cloud*-ом. Овај приступ даје компанији потпуно управљање хардвером, али и смањену флексибилност. Јавни *cloud* представља потпуно коришћење горе поменутих *cloud* сервиса. Хибридни приступ представља комбинацију приватног и јавног *cloud*-а. Ово је

веома заступљен вид употребе јер даје компромис између контролисања сопственог хардвера, чувања поверљивих података у својој приватној мрежи (често је то веома битан параметар због законских регулатива) али и предности флексибилности коју пружају јавне *cloud* платформе. Поред тога, овај приступ је заступљен у ситуацијама када се врши миграција постојећег система на *cloud* и обратно. На слици 2.3.2 је приказана имплементација све три варијанте.



Слика 2.3.2. Варијанте *cloud* имплементације.

Cloud технологије су се, у односу на IT технологије, релативно скоро појавиле. У овом тренутку, још увек постоје изазови са којима се ова врста технологије сусреће. Један од највећих недостатака је безбедност. Одговорност за недозвољен приступ је одговорност пружаоца услуга па не постоји потпуна контрола над овим сегментом од стране клијената. Поред тога, постоји недостатак транспарентности по питању локације и начина складиштења осетљивих података. Поменута флексибилност се одражава и на све остале аспекте *cloud* сервиса. То је највећи компромис са којим компаније морају да се сложе приликом преласка на *cloud*. Аутоматско скалирање хардвера може довести до непредвидивости по питању трошкова. Уколико се не користи на оптималан начин, *cloud* технологије могу бити скупље од традиционалне *in-house* имплементације. Поред тога, још увек не постоји обезбеђена гарантована брзина преноса и кашњење током преноса. Оно што је чињеница јесте да све више компанија прелази на *cloud* решења упркос недостацима и такав тренд ће се наставити и у наредним годинама.

2.4. Примена биг дата технологије

Развој економије и индустрије повлачи са собом генерисање огромне количине података. Смањење цена уређаја и описмењавање становништва доводи до повећане употребе модерних технологија. Према [1], у 2021. години број мобилних претплатника у свету премашује 8.6 милијарди. Масовна употреба сензора и развој технологије значајно утичу на светски раст података. Интуитивно, велика количина података са собом доноси и велику количину информација, а самим тим и велики потенцијал за убрзани развој области у којој се примењује. Могућност обраде великих количина података различитих типова отвара нова поглавља и игра важну улогу у откривању нових информација и сазнања [42]. Биг дата технологија је веома важна у многим сферама, од компанија, преко фабрика, до здравства. Коришћењем ове технологије, могуће је унапредити пословање једне компаније, а самим тим

директно утицати на већи профит. Велики број компанија је започео истраживања у овој области, чиме биг дата додатно добија на важности.

На основу великог броја различитих алата доступних у биг дата портфолију, биг дата технологија се може користити за решавање различитих врста проблема. Најједноставнија употреба биг дата јесте коришћење исте као централизовано решење за складиштење података. Уз помоћ Spark-a, прикупљени подаци се могу комбиновати и агрегирати како би се извукле скривене информације и генерисао бољи увид у податке. Додатно, коришћењем Spark-a могуће је вршити обраду стримова података у реалном времену, као и могућност дизајнирања модела машинског учења и предиктивне анализе над огромним количинама података.

Табела 2.4.1. Примена биг дата технологија груписана по индустријама.

Индустрија	Биг дата примена
ИТ	Анализа лог фајлова Оптимизација мреже Предикција отказа
Телекомуникације	Надгледање перформанси мреже Наплата и оптимизација тарифа Предикција прекида уговора 360 степени преглед претплатника
Банкарски сектор и осигурање	Детекција преваре Анализа операција Анализа клијената Предикција прекида уговора 360 степени преглед претплатника
Држава	Национална безбедност Предикција и спречавање прекршаја Сајбер безбедност Научна истраживања Контрола пореза
Продаја (традиционална и интернет)	Наплата Разумевање клијента Нагледање историје производа Предикција у смислу препоруке артикла клијенту 360 степени преглед клијента Класификација клијената
Интернет ствари (IoT)	Паметне куће Паметни градови (интелигентно осветљење, детекција саобраћајне гужве, контрола семафора, контрола квалитета ваздуха…)
Анализа социјалних медија	Анализа осећања клијената Анализа јавног мњења Анализа задовољства клијената
Индустрија	Детекција неочекиваних догађаја у производном ланцу Контрола квалитета Предикција отказа машина
Здравство	Анализа здравствених података Израда клиничке слике Предикција срчаног удара Детекција болести

Биг дата представља актуелну тему у многим гранама индустрије као што су, на пример, мобилне мреже [9-11][43-46], софтверски дефинисане мобилне мреже [47], енергетика [48,49], здравство [50], IoT [51-54], паметни градови [55-57], паметне куће [58,59], банкарство [60-62], малопродаја и е-трговина [63-65], индустрија [66], авио индустрија [67,68], државне институције [69] и још многим другим. Различити сценарији употребе се суочавају са различитим биг дата проблемима. На пример, неки сценарији се фокусирају само на складиштење података и агрегацију [70,71], неки се баве обрадом стримова података у реалном времену [72,73], неки се фокусирају на ефикасну претрагу [74] и тако даље. Детаљан преглед примене биг дата технологија у ”сајбер-физичким” системима дат је у [75] при чему су разматрани многи аспекти попут разноврсности извора података и колектора, кеширања, безбедности, али и еколошких проблема. Под ”сајбер-физичким” системима подразумевају се сви паметни рачунарски системи који су контролисани и надгледани алгоритмима што у данашње време представља већину модерних система. Разматрање примене биг дата технологије у различитим индустријама представљено је у [28]. Табела 2.4.1 заснована је управо на разматрањима у [28] и илуструје могућности примене биг дата технологије груписано по различитим индустријским гранама. Иако је приказан само мали делић могућности примена, посматрањем табеле 2.4.1 се може закључити да је потенцијал биг дата технологије огроман.

2.5. Актуелни изазови у биг дата

Узевши у обзир да је биг дата технологија релативно млада, као и чињеницу оперисања над великим количинама података са различитих извора, постоји још доста изазова на које треба одговорити. Неки од тих изазова су стриктно техничке природе, док су други више организационе. Неки од најважнијих изазова са којима се тренутно среће биг дата технологија су анализирани у [28]. У наставку ће бити дат пресек неких од најважнијих и најпознатијих изазова из ове области.

Обзиром на то да се подаци могу комбиновати из различитих извора, њиховом комбинацијом, могуће је открити информације о клијенту које директно задиру у његову приватност. Иако то иницијално није био план, већ је циљ био користити податке за унапређење пословања, неопрезно коришћење података може довести до оваквог феномена [27]. Праћење кретања, социјалне активности, приватне поруке, су само неке од великог броја примера у којима приватност може бити угрожена неопрезним коришћењем података. Уколико се генерисана информација користи за угрожавање особе, ово директно води до нарушавања безбедности лица.

Претпоставимо да је једна компанија усвојила биг дата технологију и прикупила велику количину података различитих типова са различитих извора. Једном кад су подаци прикупљени, поставља се питање шта са подацима радити, које информације из њих извући и на који начин. Аналитички процес који се бави анализом података у циљу екстракције корисних информација се назива дата мајнинг (енгл. *data mining*). Поред тога, потребно је повезати прикупљене информације са пословним процесима и схватити какве је акције потребно извршити на основу нових сазнања у циљу побољшања пословања. Ово често није тривијалан посао и захтева добро доменско знање. За дата мајнинг, није довољно само програмерско знање, већ је потребно и доменско. Врло често, инжењери који познају и једно и друго су веома ретки што може успорити време до реализације новог пословног процеса. Овај изазов се може решити формирањем тима људи са техничким и доменским знањем. Свака компанија поседује свој специфичан скуп података и своје интерне процесе. Због тога,

развијени модели за једну компанију често нису директно применљиви за другу што ограничава њихову даљу примену и захтева прављење решења ”по мери”.

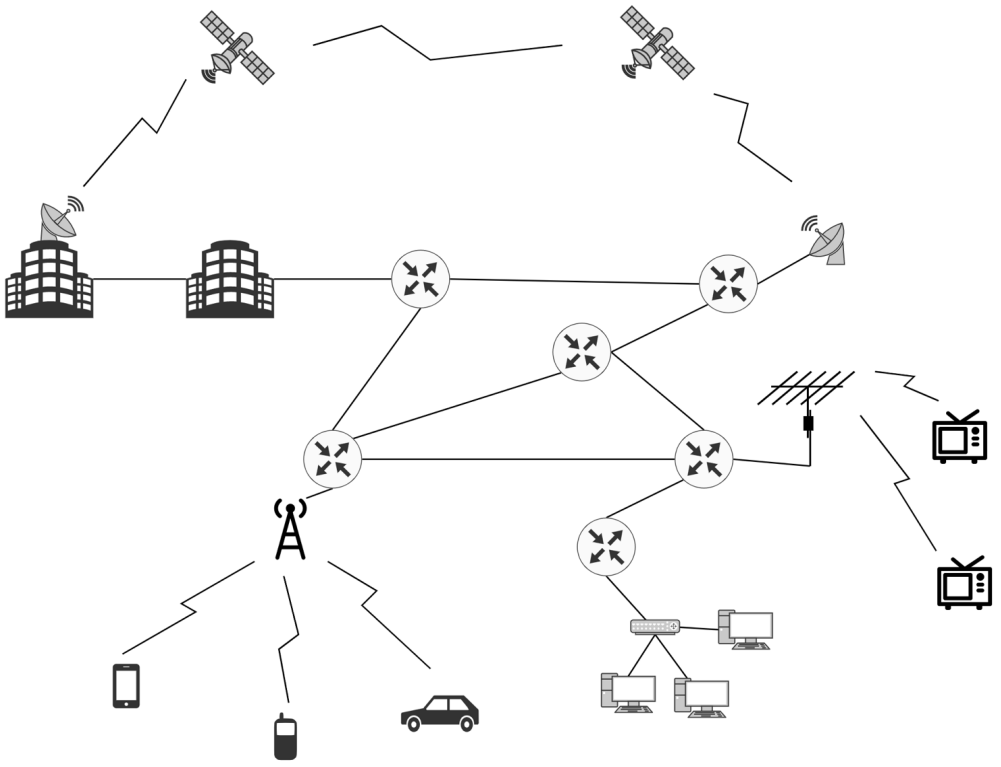
Велика количина података поред корисних, по теорији великих бројева, доноси и велику количину невалидних информација. Квалитет података је још једна ствар на коју треба обратити посебну пажњу. Узроци за проблеме у квалитету података могу бити разни. Неки од примера су промена модела на извору, која није испраћена на слоју података, непрепознавање проблема, конфликт у подацима, неконзистентност. Уколико се не води рачуна и ради са подацима лошег квалитета, може доћи до контрапродуктивности и, уместо унапредити, уназадити пословање.

Интензиван развој биг дата технологија води до отварања нових радних позиција. Велики захтев тржишта и чињеница да је ова технологија доста нова доводи до мањка стручног кадра. Неопходно потребно знање једног софтвер инжењера се шири са чисто техничког на истраживачко и аналитичко [27]. Како би надоместили овај недостатак и омогућили инжењерима стицање нових знања, велике компаније и *open-source* организације креирају велики број онлајн курсева и радионица. Додатно, универзитети полако уводе биг дата у предавања како би студентима представили ову парадигму у циљу боље припреме за рад на овом пољу.

Обзиром на то да су тренутно све сфере индустрије, од науке, здравства, па до интернет технологија у наглом развоју, све чешће се поставља питање узрочно последичне везе истих са утицајем на животну средину. Велика количина података подразумева већи број крајњих уређаја који генеришу податке, а самим тим и више система који исте обрађују и складиште. Ово безусловно доводи до повећања потрошње енергије, од производње компоненти, до обраде података што доприноси глобалном загревању. Анализа утицаја биг дата технологија на екологију је приказана у [76]. У [76] се разматра да ли биг дата технологија заиста има утицај на екологију и на који начин. Поред детаљног прегледа литературе, дат је предлог разматрања утицаја кроз дефинисање метрика енергетске ефикасности и ефективне енергетске ефикасности. Утицај је анализиран кроз целокупан циклус података, од генерисања, преко складиштења и обраде.

3. НАДГЛЕДАЊЕ ПЕРФОРМАНСИ У МРЕЖАМА

Телекомуникације представљају пренос информација различитим технологијама и медијумима за пренос. Данашње савремене телекомуникације еволуирале су у комплексне електричне, електромагнетске, оптичке и бежичне системе који имају за циљ решавање изазова који се тичу преноса великих количина информација на велике удаљености без штетних губитака изазваних различитим факторима попут шумова, различитих врста интерференција и сметњи. Телекомуникације су у данашње време постале стандард без ког је даљи развој човечанства незамислив. Због огромног протока информација, великог броја сервиса и корисника који се ослањају на телекомуникације, од есенцијалне важности је обезбедити мрежу високог капацитета и високе доступности. Узевши у обзир хетерогеност и комплексност данашњих мрежа, ово може бити веома изазован посао. У овом поглављу је дат кратак преглед телекомуникационих мрежа са фокусом на мреже које се користе за пружање сервиса интернета крајњим корисницима. Обзиром на распрострањеност и велики утицај ових мрежа, посебна пажња је посвећена представљању значаја надгледања њихових перформанси. Популарност и значај ове теме се може видети кроз велики број радова из области надгледања перформанси телекомуникационих мрежа. Стога ово поглавље даје преглед релевантне литературе из области надгледања перформанси мрежа како би се добио увид у тренутно стање у свету у овој области. На крају овог поглавља дата је кратка дискусија која представља доприносе дисертације у односу на постојеће радове из ове области.



Слика 3.1. Пример мешовите телекомуникационе мреже.

Телекомуникационе мреже представљају групу уређаја и чворова, међусобно повезаних телекомуникационим линковима који се користе за пренос информација. Телекомуникациони линкови могу користити различите медијуме за пренос попут ваздуха, бакарних или оптичких каблова. Линкови могу имплементирати различите технике за комуникацију и усмеравање података. Комуникација између уређаја се може обавити директно или преко посредника у више скокова. Како би рутирање у мрежи било могуће, сваком чвору се додељује мрежна адреса у циљу идентификације. Неке од најпознатијих телекомуникационих мрежа су интернет, компјутерске мреже, телефонске мреже (PSTN (*Public Switched Telephone Network*)), бежичне радио мреже, јавне мобилне мреже и кабловске мреже. Поред тога, у зависности од примене и медијума за пренос, у данашње време постоји велики број различитих типова мрежа. На пример, сателитске мреже се користе за двосмерну комуникацију и уз помоћ радио линкова у стању су да повежу чворове удаљене по неколико хиљада километара. Радио и ТВ (*Телевизија*) мреже служе за бежично дистрибутивно емитовање сигнала и представља једносмерну комуникацију са једним предајником и великим бројем пријемника. SON (*Self-Organizing Network*) представљају аутоматске самоорганизујуће мреже. Ова парадигма обезбеђује самостално планирање, конфигурацију и управљање мреже на основу алгоритама и свог окружења. Различити типови мрежа могу бити повезани. На слици 3.1 дат је илустративан пример мешовите телекомуникационе мреже.

Брз развој технологија и интернета форсирају динамичан развој телекомуникација и гурају их изван граница за које су иницијално предвиђене. На пример, кабловске мреже су иницијално предвиђене за пренос ТВ сигнала док су се исте у данашње време развиле до нивоа пружања сервиса телевизије, интернета и телефоније. У првој генерацији мобилне телефоније, главни фокус је био на обезбеђивању аналогног канала комуникације између два претплатника. Даљим развојем, мобилне мреже мењају свој фокус са комуникације на пакетски пренос а самим тим и повећање капацитета линка као једног од главних параметара успеха. У четвртој, LTE (*Long Term Evolution*), генерацији мобилних мрежа, капацитет линка је повећан до 100 Mb/s уз значајно смањење кашњења пакета. Пета генерација мобилних мрежа даље повећава капацитет, смањује кашњење и обезбеђује подршку IoT уређајима.

Због високе конкуренције на тржишту телекомуникација, захтевних сервиса и клијената, неопходно је обезбедити мрежу високих капацитета и високе доступности. Неопходно је вршити константан надзор мреже како би се клијентима обезбедио високо доступан сервис и како би се правовремено детектовали проблеми и исти уклонили што је пре могуће. Надгледање перформанси мреже доноси комплетан увид у стање читавог система, од извора података до крајњег клијента. Константно праћење параметара мреже и константно унапређење у дугорочном смислу побољшава квалитет мреже. Континуалним прикупљањем метрика могуће је имати увид у историјско стање мреже и пратити трендове свих параметара. Уз помоћ прикупљених података могуће је приоритизовати радове на мрежи у складу са реалним стањем и потребама. Посматрањем истих могуће је правовремено детектовати и решити проблеме уског грла у мрежи. С тим у вези, могуће је вршити правовремена проширења избегавајући загушења линкова. Праћење мреже у реалном времену и увид у податке може помоћи приликом детекције и локализације појединих кварова. У ситуацијама групног отказа, могуће је обавестити клијенте о квару пре него што и исти пријаве квар и на тај начин поправити корисничко искуство, па чак и кад мрежа не ради. Поред тога, у ситуацијама жалби клијената, поред тренутног увида у рад његове CPE опреме, могуће је имати и историјски увид који помаже у проналажењу узрока проблема.

Комплексност и хетерогеност данашњих мрежа као и велики број клијената отежавају процес надгледања мреже. Велика количина генерисаних података отежава процес складиштења и брзог приступа. Појавом биг дата технологије, отварају се нове могућности по питању праћења перформанси мрежа, складиштења и обраде података. Постоји доста научних радова где се, применом биг дата технологија, предлажу решења за надгледање перформанси мрежа и решавање конкретних проблема применом прикупљених података. Поред тога, примена техника машинског учења у овој области је такође стекла изузетну популарност.

Примена биг дата аналитике у мобилним мрежама представљена је у [9]. Најпре је математички дефинисан дата модел коришћењем теорије насумичних матрица које су касније повезане са подацима из мобилне мреже. Такође, дат је предлог биг дата радног оквира за складиштење и обраду података. Поред тога, објашњена је присутност биг дата технологије у мобилним подацима и дат предлог примене саобраћајних, локацијских и других података. Решење за надгледање мрежа и анализирање система заснованом на Hadoop технологији предложено је у [43]. Аутори такође истичу проблем скалабилности постојећих готових решења што је и била главна мотивација рада. Систем је дизајниран у вишеслојној архитектури. Предложени систем је имплементиран у језгру мреже (енгл. *core network*) у 2G и 3G мобилним мрежама. Решење за надгледање перформанси мобилних мрежа засновано на биг дата технологијама предложено је у [44]. Предложено решење користи HDFS фајл систем и фокусира се на пост-обраду података. Прикупљени подаци се прикупљају у XML формату коришћењем Apache Flume-a. Предложено решење је тестирано на тестним сетовима из мобилне 5G мреже. Како би се симулирао реалан систем, вршена је репликација постојећег дата сета. Платформа за прикупљање података мобилног интернета заснована на биг дата технологијама представљена је у [77]. Предложено решење користи HDFS и HBase за складиштење података. Предложено решење имплементирано је на кластеру ког чини 188 сервера и које се користи у продукционом окружењу. У поређењу са Oracle базом података, решење [77] даје боље резултате приликом генерисања паралелних упита. Решење предложено у [78] фокусира се на анализи CDR-ова (*Call Detail Record*) и конкретној примени у циљу откривања понашања мобилне мреже. Поред тога, у циљу обраде скупа података заснован на CDR-овима предложено је биг дата решење засновано на HDFS, MapReduce, HBase и Hive. Софтверски дефинисана мобилна мрежа која подржава анализу и ефикасно складиштење података предложена је у [47]. Ово решење се пореди са традиционалним мобилним мрежама и предлаже циклус управљања подацима у циљу ефикасног складиштења и анализе. Детаљна анализа и класификација различитих извора података у мобилним мрежама представљена је у [46]. Извори података се посматрају из перспективе ”5V” биг дата изазова. Поред тога, подаци су такође анализирани из угла животног циклуса, тј. генерисање, пренос, агрегација и конкретна примена. Еволуција мобилних мрежа и биг дата технологија дискутована је у [79]. Аутор наглашава битност истраживања и анализе података уз одговарајуће предности и недостатке. Предности се односе на све вредности које анализа података може донети док се недостаци тичу угрожавања приватности.

Примена биг дата технологија у енергетици и паметним мрежама предложена је у [49]. Конкретна мрежа са које се прикупљају подаци садржи велики број сензора и паметних бројила за даљинско очитавање потрошње електричне енергије. Обзиром на то да се подаци константно прикупљају из мреже, разматра се обрада података у реалном времену. Надгледање квалитета сигнала у HFC мрежама анализирано је у [80] где је предложен специфичан алгоритам за анализу перформанси. Наиме, алгоритам је заснован на

тродимензионој анализи података прикупљених са CMTS-а (*Cable Modem Termination System*) и кабловских модема. Прикупљени параметри се приказују у 3D простору на основу чега се може закључити узрок проблема у мрежи.

У случају мрежа које садрже огроман број уређаја, биг дата је неизбежна за надгледање перформанси. Међутим, прикупљени подаци се могу користити и за друге сврхе. Детаљан преглед за примену биг дата технологија у жичним и бежичним мрежама представљен је у [81]. Неки од дискутованих аспеката примене обухватају предикцију саобраћаја, унапређење QoS-а (*Quality of Service*), сајбер безбедност, оптимизацију перформанси [81]. За детекцију отказа у мобилним мрежама коришћење XDR-ова (*Extended Detection and Response*) предложено је у [82], док је анализа трендова протока како би се предвидело загушење линка предложено у [83].

Оператори мобилних мрежа суочени су са огромним растом података у својим мрежама. Због комплексности мрежа, захтева за сервисом високих перформанси и високе конкуренције, мобилни оператори су приморани да константно усавршавају своје мреже. Поред тога, увођење 5G мрежа, IoT уређаја и M2M (*Machine to Machine*) сервиса додатно повећава количину прикупљених података. Решење за прикупљање и обраду података за мобилне операторе засновано на биг дата технологијама предложено је у [45]. Решење је засновано на "ламбда" [84] архитектури које истовремено омогућава прикупљање и обраду података и у реалном времену, и у пост-обради. Поред тога, предложен је начин обраде података у циљу анонимизације, побољшане безбедности контроле приступа.

Постоји велики број научних радова на тему обраде података прикупљених из мреже телекомуникационих оператора. Позитиван тренд раста броја радова на ову тему додатно потврђује популарност области, али још увек постоји простор за даља унапређења и доприносе. Највећи број постојећих радова је фокусиран на прикупљање и обраду података из мобилних мрежа због њихове популарности и броја активних корисника. Надгледање перформанси мрежа, попут HFC и DSL мрежа, је неправедно запостављено иако и оне пружају сервисе великом броју корисника. Систем за надгледање перформанси мрежа предложен у овој дисертацији је намењен и тестиран у HFC мрежама. Али, предложени систем је креиран флексибилно тако да може да подржи и друге типове мрежа, попут мобилне, DLS, IoT и многих других где је фокус на прикупљању временских серија. Конкретна примена у HFC мрежама, као и флексибилност предложеног система, која је је дискутована у другом делу тезе, представљају доприносе имајући у виду постојеће радове из ове области. Такође, радови из области надгледања перформанси телекомуникационих мрежа, али и примене биг дата технологија у телекомуникационим мрежама су типично фокусирани на поједине проблеме и изазове, као што се може приметити из прегледа литературе датом у овом поглављу. Мали број радова се бави решењима у виду свеобухватне биг дата архитектуре, а чак и у тим радовима тестирања су вршена у лабораторијским окружењима, док систем предложен у овој дисертацији је тестиран у реалном окружењу које увек доноси и непредвиђене ситуације и сценарије. Ово такође представља додатан допринос дисертације постојећој литератури ове области.

Иако је биг дата препозната као технологија са много потенцијала у телекомуникационим системима, у литератури се иста тек понегде примењује и то за решавање одређеног проблема. Наиме, не постоји јединствено решење довољно скалабилно и робусно које би могло да служи као централизовано место где би се имплементирала решења за поједине проблеме заснована на подацима из мреже. С друге стране, систем предложен у овој дисертацији представља целовито решење за складиштење, обраду и брз

приступ подацима. Исти се може применити за имплементацију било каквог алгоритма који је заснован на подацима. Поред система за надгледање перформанси мреже, у дисертацији је предложен и механизам за детекцију и локализацију кварова у мрежи заснован на реалним подацима у мрежи, и који је верификован у пракси. Овај механизам представља допринос у области детекције и локализације кварова, пошто се већина радова у овој области бави или детекцијом или локализацијом отказа што је дискутовано у шестом поглављу. Додатно, теме попут процена стања уређаја са којих се не могу прикупљати подаци није адекватно обрађена у литератури. У оквиру дисертације, предложен је алгоритам за оцењивање таквих уређаја што такође представља новину и један од доприноса ове дисертације.

4. HFC МРЕЖЕ

HFC представља једну од најисплативијих мрежних технологија за пружање сервиса крајњим корисницима која је добијена комбиновањем најбољих карактеристика оптичке и коаксијалне мреже. Ова технологија је резултат еволуције кабловске телевизије (CATV (*Cable Television*)). Традиционалне CATV мреже су се користиле за емитовање ТВ сигнала. Како би се повећао капацитет ових мрежа, поузданост, квалитет слике, отпорност на шум, али и смањио трошак одржавања, CATV оператори замењују коаксијалне каблове оптичким влакнима у великим деловима своје мреже. Наиме, оптички каблови се полажу ближе оператору, док у деловима ближим кориснику остаје коаксијална мрежа. CATV мреже модификоване на описан начин се називају HFC мрежама [85]. Развојем интернет технологија, појавила се потреба за двосмерном комуникацијом преко CATV инфраструктуре како би се корисницима, поред телевизије, пружио и интернет сервис. Узевши у обзир чињеницу да су делови мреже замењени оптиком, HFC мрежа је била у стању да одговори на овакав изазов. Данашње HFC мреже пружају екстремно високе протоке корисницима и проширују свој портфолио сервиса у такозвани ”*triple play service*” који обухвата емитовање ТВ сигнала, широкопојасни приступ интернету и IP (*Internet Protocol*) телефонију [86].

HFC мрежу сачињавају различити типови мрежних елемената и уређаја повезаних оптичким и коаксијалним кабловима на одређени начин. На слици 4.1 приказан је пример физичке архитектуре HFC мреже. У наставку ће бити упрошћено описана HFC мрежа заједно са одговарајућим уређајима који су релевантни за ово истраживање. У складу с тим, биће уопштено описан ток података кроз HFC мрежу у оба смера.

CMTS је уређај који се користи за пружање сервиса високог протока корисницима, као што су телевизија, кабловски интернет и VoIP (*Voice over IP*). Овај уређај се типично налази у централама кабловског оператора. Један CMTS, зависно од своје конфигурације, може да опслужује хиљаде корисника. Он је задужен за додељивање динамичких IP адреса корисничким уређајима, као и за прослеђивање саобраћаја у оба смера. Другим речима, CMTS прикупља долазни мултиплексирани саобраћај од корисника и рутира га даље ка интернету и обратно. У зависности од потреба, овај уређај се може понашати и као свич и као рутер [87]. CPE уређаји повезани на један (исти) CMTS не могу комуницирати директно између себе. Постоје интегрисани и модуларни типови CMTS-а, где модуларни дају могућност додатног скалирања капацитета у складу са потребама. Постоји велики број различитих произвођача CMTS-ова од којих су најпознатији Cisco, Casa, Arris, Teleste и Huawei.

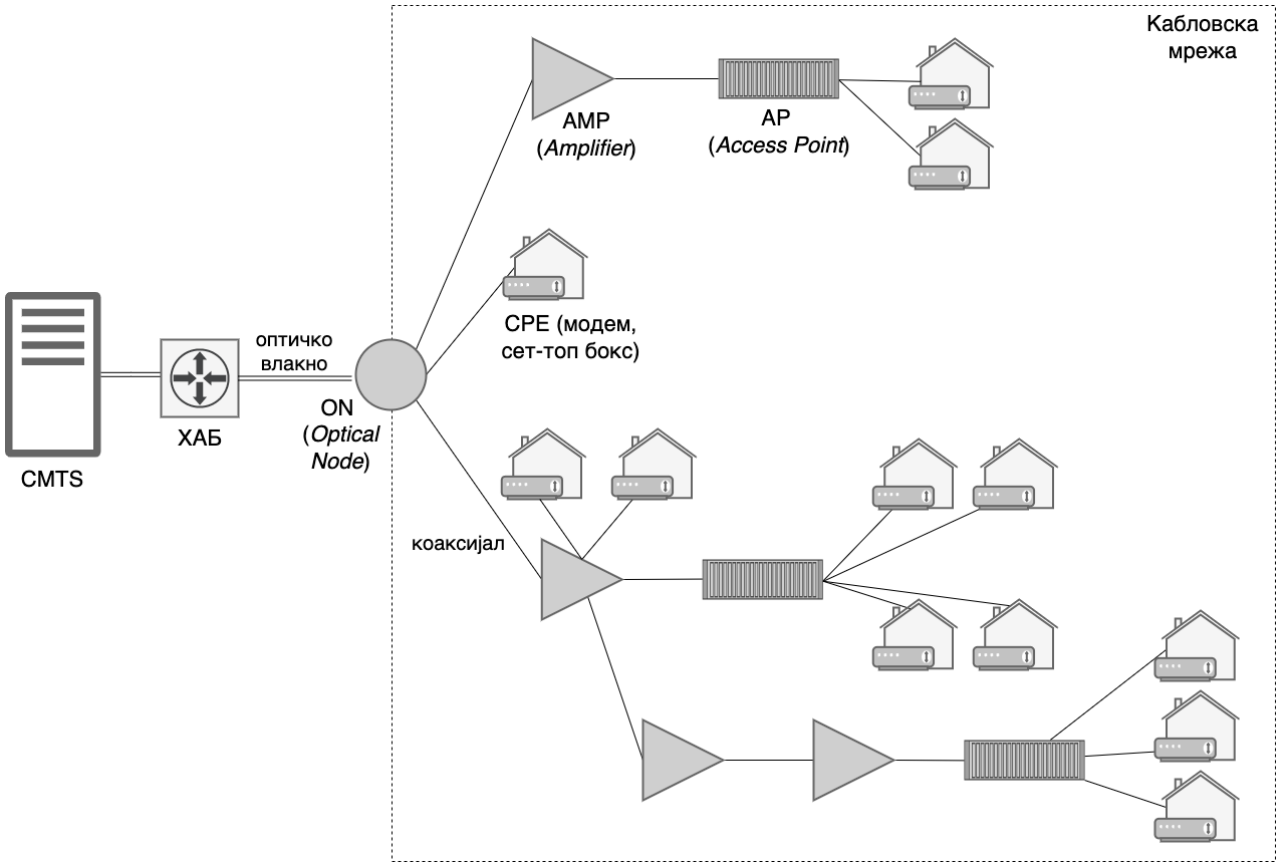
ON (*Optical Node*) представља уређај који конвертује оптички сигнал у електрични и обратно у *downstream*, тј. *upstream* смеру, респективно. Могућност конвертовања сигнала обезбеђује приближавање оптичког влакна ближе корисницима и на тај начин више протоке и мање кашњење. ON је у стању да комбинује и мултиплексира сигнале који долазе са више извора и на тај начин додатно оптимизује проток.

Како би се превазишло слабљење коаксијалног кабла и пасивни губици, неопходно је вршити периодично појачавање сигнала како би исти могао да стигне до крајњих корисника. У ту сврху се користе AMP-ови (*Amplifiers*) тј. појачавачи. У зависности од типа мреже и дужине каблова, у мрежи се може налазити више појачавача.

Када сигнал дође до приступне тачке, на пример, улаза у зграду (или другу мању логичку јединицу која спаја мањи број корисника), користи се AP (*Access Point*) који даље раздваја сигнал према крајњим корисницима. Зависно од конвенције, понегде се ови уређаји називају и "тапови" (енгл. *taps*) [88].

Како би крајњи корисници били у могућности да користе дигитално модулисани сигнал, на њиховој локацији се инсталира CPE опрема. Ова опрема конвертује сигнал у формат компатибилан пријемницима (на пример, ТВ сигнал). У зависности од типа сервиса, то може бити CM (*Cable Modem*), STB (*Set-Top Box*) и VoIP телефон. CM обезбеђује интернет сервис, STB телевизију, а VoIP уређаји телефонију путем интернета. Због једноставности, у наставку овог рада ће се под CPE подразумевати и CM и STB.

На основу тога да ли се са уређаја могу прикупљати директно подаци као и могућност директне комуникације, у мрежи је извршена подела на интелигентне и неинтелигентне уређаје. Уређаји који могу да врше мерења (генеришу метрике) и комуницирају са централном платформом у циљу прикупљања података (метрика) спадају у интелигентне уређаје, што су у случају HFC мрежа CMTS и CPE опрема. У неинтелигентне уређаје спадају ON, AMP и AP уређаји. Са ових уређаја није могуће прикупљање података, нити они имају способност мерења (генерисања метрика). Више речи о овој подели ће бити у наставку.



Слика 4.1. Пример физичке архитектуре HFC мреже.

Као што је напоменуто, комуникација у HFC мрежи се одвија у оба смера. У *downstream* смеру, регионални CMTS прима сигнал (ТВ сигнал, телефонија и интернет), модулише га и комбинује у јединствени сигнал који се дистрибуира путем оптичког влакна [36]. Оптички сигнал пристиже на ON преко хаба. ON конвертује оптички сигнал у електрични и дистрибуира га путем кабловске мреже. Кабловска мрежа представља тополошку структуру у виду стабла сачињена од појачавача AMP и сплитера повезаних коаксијалним каблом. Појачавачи се користе како би се превазишло слабљење електричних сигнала у коаксијалном каблу. Коначно, преко приступне тачке, сигнал преко AP долази до крајњег корисника, тј. до корисничке мрежне опреме (CPE). У супротном, *upstream*, смеру, сигнал се фреквенцијски мултиплексира, шаље преко коаксијалне мреже до ON, конвертује у оптички сигнал и шаље назад до CMTS-а [36]. Са слике 4.1 се такође може видети да се CPE опрема може повезивати не само преко AP, већ се може повезивати директно и на AMP и ON. Ово суштински не усложњава даљу логику, али је неопходно напоменути.

Широкопојасни сервиси високих брзина стандардизовани су од стране DOCSIS-а. DOCSIS представља стандард за транспорт сигнала за HFC мрежу уз дефинисање модулација. Иницијална верзија DOCSIS 1.0 стандарда дефинисана је 1998. године од стране ITU (*International Telecommunication Union*) и подржавала је брзине до 10 Mb/s у *upstream*-у и до 40 Mb/s у *downstream*-у. У верзији 1.1 DOCSIS почиње да подржава и VoIP сервисе (VoIP 1.1). У верзији 2.0 долази до повећања капацитета на *upstream*-у проширењем фреквенцијског опсега и подршком нових модулационих шема. Значајна унапређења уводи DOCSIS 3.0 где се *downstream* и *upstream* капацитети повећавају до 1 Gb/s, тј. 200 Mb/s, респективно. Најзаслужнија функционалност за повећање капацитета је опција здруживања канала (енгл. *channel bonding*) уз коју је могуће остварити неколико конекција у паралели, а самим тим и остварити веће протоке. Поред ове функционалности, представљена је и подршка за IPv6. Следећа верзија 3.1 значајно повећава капацитете, док је тренутна верзија, DOCSIS 3.1 *full duplex* (преименована касније у DOCSIS 4.0), прерасла у пун дуплекс комуникациони систем са капацитетом до 10 Gb/s [36][89].

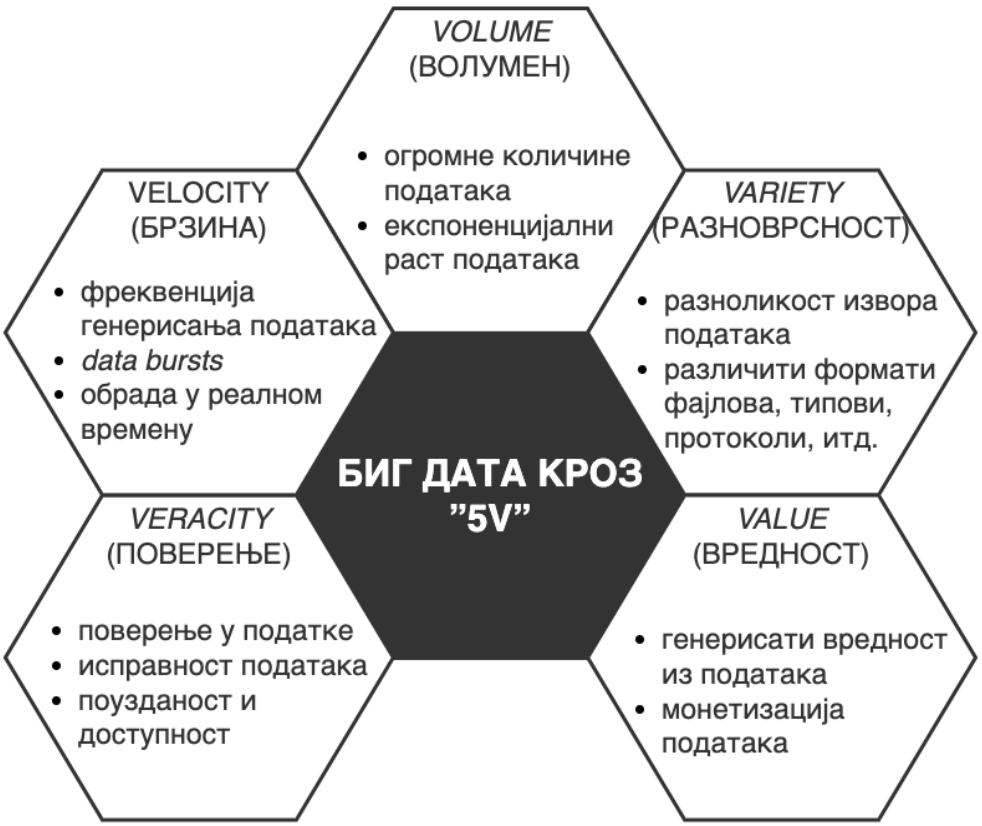
Иако HFC мреже и DOCSIS стандард постоје више од 20 година, због већ постојеће изграђене инфраструктуре и високих протока које пружа, и даље се активно ради на усавршавању и дефинисању нових верзија стандарда. У наставку су представљена нека од тренутно активних истраживања на ову тему. Увођењем DOCSIS 3.1, новог комуникационог стандарда у кабловским мрежама, значајно је повећана брзина преноса сигнала. Ово је омогућено коришћењем OFDM-а (*Orthogonal Frequency-Division Multiplexing*) и модулација вишег реда. OFDM омогућава подешавање носиоца по сваком кабловском модему што обезбеђује велику флексибилност за оптимизацију протока, што није био случај у претходној верзији стандарда (DOCSIS 3.0). Овај потенцијал препознат је и искоришћен у [90,91]. Алгоритам за динамичку конфигурацију носиоца по кабловском модему заснован на машинском учењу предложен је [90]. Овај алгоритам је тестиран на 10 реалних скупова података уз показане боље резултате. Слична анализа урађена је у [91] уз предлагање сопственог алгоритма за управљање профилом корисника. Оба приступа ослањају се на податке прикупљене из мреже. Решавање проблема феномена само-интерференције (уређај сам себи изазива интерференцију) у DOCSIS 3.1 стандарду је дискутовано у [92,93]. Емитовани сигнали у двосмерној комуникацији се међусобно изобличују. Обзиром на то да је количина интерференције и њена природа унапред позната, предложене су технике за самопроцену у циљу минимизације поменутог феномена што је и доказано у лабораторијским условима. Најновији интернет трендови показују повећану потражњу за комуникацију у *upstream* смеру. Док је DOCSIS 3.1 најновији стандард у HFC мрежама, исти

и даље имплементира асиметричне линкове за *upstream* и *downstream*. Постоји неколико научних радова који предлажу проширење постојећег стандарда како би се остварила симетрична комуникација [94,95]. Пун дуплекс екстензија DOCSIS 3.1 стандарда и преглед изазова груписаних по физичком, MAC (*Media Access Control*) домену и системском нивоу дата је у [94], док је пун дуплекс стандард који покушава да реши све изазове, симулиран у лабораторијским условима, предложен у [95].

4.1. Изазови у надгледању перформанси HFC мрежа

Обзиром на величину и комплексност генерално свих типова телекомуникационих мрежа, њихово надгледање може бити веома изазован посао. Међутим, систем за надгледање перформанси предложен у овој дисертацији је превасходно прилагођен HFC мрежама, а такође управо у једној великој HFC мрежи се предложени систем и успешно користи са веома добрим резултатима. Отуда се ово потпоглавље бави прегледом изазова са становишта надгледања перфороманси HFC мреже са великим бројем корисника. Управо имајући у виду изазове описане у овом потпоглављу је систем за надгледање перформанси и дизајниран, а који ће бити описан детаљно у следећем поглављу.

Изазови приликом надгледања перформанси HFC мрежа груписани су у [36] што је и један од резултата рада на овој дисертацији. У наставку ће бити детаљно представљени изазови и задаци са којима се суочава предложени систем (платформа) који треба да подржи прикупљање, смештање и обраду метрика са интелигентних уређаја у мрежи. Ови изазови ће бити представљени кроз призму ”5V” концепта који, као што је већ поменуто, обухвата *Volume*, *Variety*, *Velocity*, *Veracity* и *Value* изазове. На слици 4.1.1 представљени су наведени изазови у поменутом ”5V ” концепту.



Слика 4.1.1. Изазови у надгледању кабловских мрежа кроз биг дата ”5V” концепт.

Volume (Волумен) изазов се односи на енормну количину прикупљених података. Узевши у обзир број уређаја у HFC мрежи, овај изазов је очигледан. На пример, под претпоставком да у HFC мрежи постоји један милион CPE уређаја и да се све метрике прикупљају једном у сат времена и уколико је у просеку сваки уређај повезан на два *upstream*-а, онда се два милиона одбирака прикупи у једној итерацији за једну метрику. Уколико треба складиштити прикупљене податке за последњих шест месеци, онда око 8.73 милијарди одбирака ће бити сачувано само за једну метрику. Овај једноставан пример илуструје *Volume* изазов приликом надгледања перформанси HFC мреже [36].

Очигледно, складиштење прикупљених података представља *Volume* изазов у овом случају. Узевши у обзир количину података коју треба ускладиштити и природу временских серија, традиционалне релационе базе нису у стању да одговоре на захтеве по питању капацитета и скалабилности. Због тога, као једно од потенцијалних решења намеће се коришћење дистрибуиране не-релационе базе података. Поред тога, перформансе приликом читања података представљају значајан фактор приликом одабира одговарајућег решења за складиштење. Подсетимо се да је пожељно да одабрана база буде заснована на технологији отвореног кода из разлога економичности таквог приступа. Чак и уколико се дође до решења на све поменуте изазове и одабере се оптимална база, поставља се питање на који начин је потребно уписати податке у циљу постизања максималних перформанси читања на нивоу посматраног уређаја.

Variety (разноврсност) изазов се односи на структуриране и неструктуриране податке генерисане од стране различитих извора и њихово прикупљање из HFC мреже. Ову мрежу сачињавају различити типови уређаја што се могло уочити и на слици 4.1. Поред тога, у мрежи се могу наћи уређаји различитих произвођача (на пример, неки од CMTS произвођача су Cisco, CASA, ARRIS, Motorola, Huawei). Додатно, чак и уређаји истог произвођача се могу међусобно разликовати. На пример, исти уређај се може разликовати по инсталираној верзији софтвера или се могу разликовати по скупу могућности и операција које обављају (на пример, Cisco uBR, cBR, Remote PHY). Типови података се такође могу разликовати у зависности од улоге уређаја у мрежи и његових способности. На пример, модеми са уграђеним WiFi модулом, у поређењу са оним без овог модула, садрже и метрике везане за WiFi функционалности. Додатно, надгледани уређаји могу комуницирати коришћењем различитих протокола. SNMP (*Simple Network Management Protocol*), IPDR (*IP Detail Record*) и FTP (*File Transfer Protocol*) протоколи су најчешће коришћени за ову сврху. Међутим, који ће протокол бити коришћен зависи од конфигурације и могућности надгледаног уређаја. Поред тога, често је потребно прихватити податке од стране екстерних извора (подаци из других система, друге апликације…) што додатно повећава разноврсност прикупљених података. Горе поменута разноврсност извора података представља *Variety* изазов који представља веома битан аспект који се мора узети у обзир приликом дизајна система за надгледање перформанси HFC мреже.

Velocity (брзина) изазов се тиче фреквенције прикупљања података. Већа фреквенција прикупљања података пружа бољу грануларност и више детаља о стању HFC мреже, али захтева више ресурса за складиштење и обраду података. Увек мора да постоји компромис између фреквенције прикупљања података и физичких ресурса. Приликом прикупљања података, треба развити механизам који је у стању да дозволи привремено прикупљање података са вишом фреквенцијом у циљу детектовања специфичних проблема који се повремено јављају. Описани изазов захтева пажљиво бирање фреквенције прикупљања података приликом надгледања перформанси како би се постигао оптималан баланс између физичких ресурса и квалитета информација добијених прикупљеним подацима. Додатно,

платформа треба да буде у стању да подржи сваки вид промене фреквенције и повећања хардвера не нарушавајући постојеће прикупљене податке.

Процес прикупљања података представља веома битан изазов предложеног система зато што највише утиче на његове перформансе. У ствари, изазов прикупљања метрика представља комбинацију *Volume*, *Variety* и *Velocity* изазова. Колекциони слој система захтева значајну количину ресурса узевши у обзир количину и разноврсност надгледаних уређаја у HFC мрежи (*Volume* и *Variety* изазови). Што је већа фреквенција прикупљања метрика, то је више ресурса потребно (*Velocity* изазов). Чињеница да CPE опрема претплатника није високо доступна и могу постојати проблеми у конекцији из различитих разлога (на пример, претплатник је искључио CPE уређај, нестанак струје, проблеми у мрежној конекцији) који додатно повећавају време прикупљања метрика. Додатно, CPE опрема може имати додељену IP адресу из динамичког опсега због чега је потребно развити механизам за надгледање и ажурирање мапирања додељених IP адреса.

Veracity (поверење) изазов се тиче фактора поверења, то јест да ли се прикупљеним подацима може веровати. Уколико подаци недостају, постоји нејасноћа у вези узрока који је то изазвао. На пример, узрок може бити покварени уређај, недоступност уређаја изазвана нестанком струје или нестанком интернета, недоступност колектора који врши прикупљање. Овакве ситуације нису само теоретске, већ се у пракси свакодневно дешавају и могу се очекивати. На пример, уколико кабловски оператор реконфигурише уређај без да је претходно обавестио платформу за надгледање перформанси, могуће је да ће се прикупити нетачни подаци (на пример, промена офсета на уређају чија се вредност коригује у колектору).

Value (вредност) изазов се односи на могућност претварања огромне количине прикупљених података у корисну информацију. Једном кад су подаци прикупљени, врши се њихова даља обрада у циљу екстракције информација. Прикупљени подаци могу садржати информације које на први поглед нису видљиве. Због тога, инжењер који пише програм за обраду података треба да добро познаје исте како би могао да извуче корисну информацију и то на оптималан начин.

HFC мрежа се састоји од великог броја различитих уређаја, као што је дискутовано у *Variety* делу. Сви ти уређаји се могу груписати у интелигентне (они који могу да мере перформансе статистике и шаљу их даље) и неинтелигентне (они који нису у стању да обаве интелигентне радње). Уређаји као што су ON, AMP и AP спадају у групу неинтелигентних уређаја. Међутим, уколико је позната топологија HFC мреже, прикупљени подаци са CPE опреме се могу искористити како би се извршила оцена стања неинтелигентних уређаја. Како би се овај податак добио, неопходно је корелисати податке о мрежној топологији са подацима прикупљеним са CPE. *Value* изазов представља знање потребно за креирање поменутог алгоритма у циљу добијања најбоље могуће процене. Пример процене ће бити представљен у поглављу 5.2.4.1.

Уколико дође до проблема у мрежи, HFC оператори, на основу пријављених кварова од стране претплатника и свог искуства, проналазе квар. Овакав приступ често захтева превише времена и директно утиче на доступност и поузданост сервиса. Још један од задатака платформе јесте аутоматизација овог процеса. *Value* изазов у овом случају је креирање механизма који ће бити у стању да, на основу прикупљених података, аутоматски, у реалном времену детектује отказ мреже и у корелацији са топологијом, детектује место узрока проблема. На овај начин, време за решавање проблема је минимизовано. Овај конкретан пример ће бити детаљно објашњен у шестом поглављу.

Преглед изазова по “5V” концепту представља добру полазну тачку при дизајнирању система. Изазови детектовани и дефинисани на овај начин представљају техничке и нетехничке захтеве система. Задовољити сваки од ових захтева, а притом узети у обзир и аспекте флексибилности и скалабилности је веома изазовно јер су, поред тежине захтева система, често и сами захтеви међусобно контрадикторни. Додатно, пре самог дизајна система, неопходно је упознати се и са свим постојећим биг дата алатима и њиховим могућностима, па применити само оне који оптимално могу да испуне сва дефинисана очекивања како би се правилно креирала архитектура система. Чак и када се све узме у обзир и дизајнира систем, приликом пуштања у рад и имплементације у реалној мрежи, може доћи до непредвиђених проблема што је такође један од аспеката који је разматран у оквиру дисертације.

5. АРХИТЕКТУРА СИСТЕМА ЗА НАДГЛЕДАЊЕ ПЕРФОРМАНСИ МРЕЖЕ

У овом поглављу представљен је предложени систем за надгледање перформанси мреже заснован на биг дата технологији са применом у кабловским мрежама. Систем је реализован у виду биг дата платформе, па ће се у наставку дисертације користити скраћеница ПНПМБД (*Платформа за Надгледање Перформанси Мреже заснована на Биг Дата технологији*) за означавање предложеног система, односно платформе. Смернице за развој ПНПМБД су били ”5V” изазови са којима се срећу HFC оператори, а који су представљени и објашњени у претходном поглављу.

Архитектура ПНПМБД представља главни допринос ове дисертације и, самим тим, ово поглавље представља централни део дисертације. Резултати рада на ПНПМБД су публиковани у [96]. Поред архитектуре решења, у оквиру овог поглавља садржани су и додатни доприноси дисертације као што су:

- дефинисана основна листа метрика за прикупљање из HFC мреже,
- дата колектор за HFC мрежу који користи SNMP за повезивање на уређаје у мрежи,
- дефинисање високоперформантне шеме за брзо читање података,
- процена стања неинтелигентних уређаја,
- искуство током имплементације у реалној мрежи.

Ово поглавље је организовано на следећи начин. Најпре су, на основу ”5V” изазова, дефинисани циљеви платформе. Након тога, представљена је слојевита архитектура ПНПМБД, где је детаљно описан сваки слој понаособ. Поред саме архитектуре, приказан је и објашњен ток података, од извора, преко платформе до крајњег корисника. У циљу надгледања перформанси HFC мреже, дата је препорука основних метрика за прикупљање са CMTS и CPE уређаја. Потом, представљена је шема података која решава циљеве по питању брзине читања података из система. Обзиром да се ради о прикупљању података са великог броја уређаја, посебна пажња посвећена је дизајну дата колектора. Прикупљене податке, поред сировог облика, могуће је и агрегирати како би се добио увид у стање мреже или дела мреже из другог угла, па је стога један део овог поглавља посвећен образложењу на који начин ПНПМБД подржава ову функционалност. Имплементација решења у пракси се може доста разликовати од теоријске поставке. Један део овог поглавља је посвећен свим изазовима који су се појавили током имплементације решења у продукцији као и начинима на који су исти били адекватно превазиђени. Коначно, урађена је и анализа предложеног ПНПМБД по питању безбедности и приватности података.

5.1. Циљеви ПНПМБД

Како би било могуће развити ефикасну биг дата платформу за конкретне примене, неопходно је дефинисати листу циљева коју иста мора да испуњава током експлоатације у реалној мрежи. Анализа циљева је веома битна јер се на тај начин адресирају главни изазови

пре самог развоја платформе. Због тога, у наставку су представљени циљеви које ПНПМБД мора да испуни.

Један од главних циљева ПНПМБД је успостављање централизоване платформе за надгледање перформанси. ПНПМБД треба периодично да прикупља податке са надгледаних уређаја (извори података). Одбирци прикупљани у периодичним временским интервалима се другачије називају временске серије. ПНПМБД треба да буде у стању да прикупља податке са теоретски бесконачног броја извора података (у пракси стотина милиона уређаја). Време приступа подацима за поједини уређај треба да буде реда величине секунде. Поред тога, резолуција (периода) прикупљања података треба такође да буде реда величине секунде. Периода прикупљања података треба да буде конфигурабилан параметар како би било могуће прикупљати различит број тачака за различите метрике у зависности од њихове променљивости и важности. На пример, периода прикупљања метрике снаге сигнала треба да буду реда величине минута, док се споропроменљиве метрике попут регистрованог броја корисника могу прикупљати на 30 минута. Време складиштења прикупљених података се такође може разликовати у односу на њихову важност. Архитектура платформе мора бити дефинисана на начин да буде скалабилна и лако проширива у циљу подршке будућих проширења, интеграције са новим доменима, додавања нових уређаја. Поред тога, платформа треба да буде заснована на бесплатним *open-source* компонентама како би се минимизовао трошак одржавања. Коначно, ПНПМБД треба да буде високо доступна, то јест треба да буде омогућен приступ, а уједно и задржана веродостојност података у случају отказа било ког дела система.

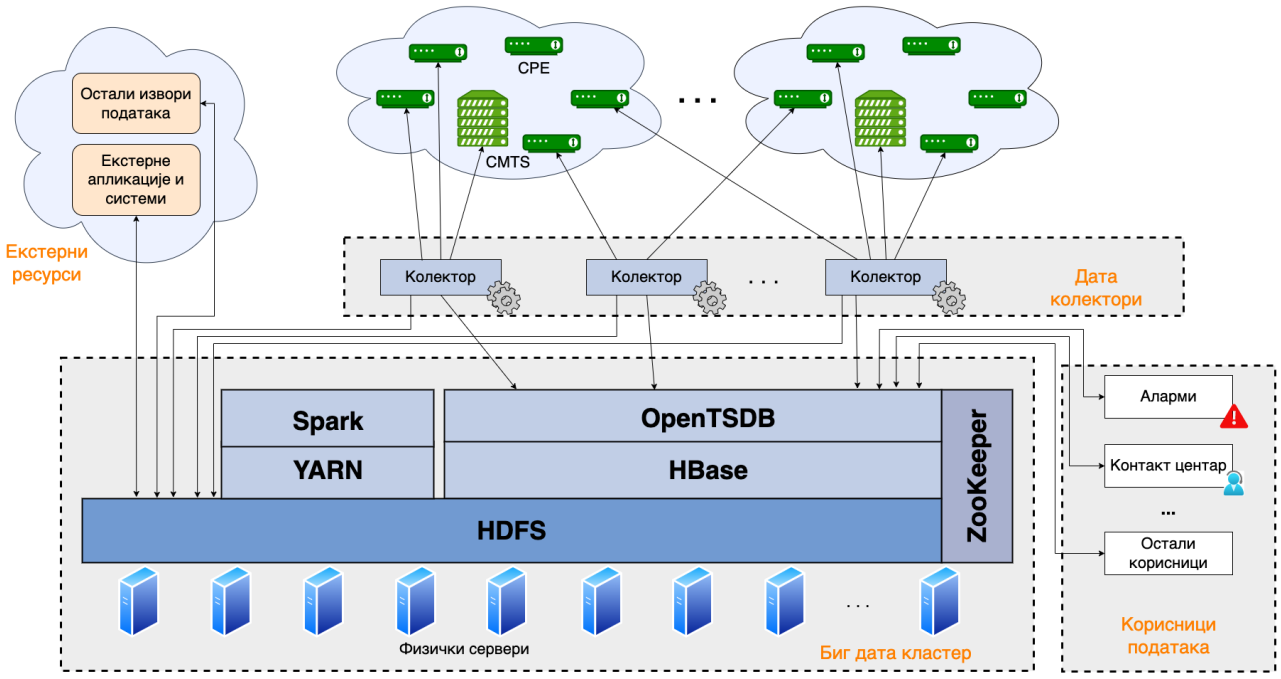
Прикупљени подаци ће се примарно користити за надгледање перформанси уређаја. То подразумева могућност посматрања временских серија уређаја у циљу увида у тренутно стање одређеног дела мреже или уређаја, али и његове историје. Још један од важних захтева који треба бити подржан јесте могућност промене резолуције посматрања дела мреже, од посматрања појединачног елемента, преко логичке групе елемената, до јединственог прегледа на нивоу целе компаније. Коришћењем ове функционалности, мрежни оператор ће бити у стању да лако детектује проблематичне делове мреже и у складу с тим правовремено планира њено одржавање и проширење. Један од важних циљева ПНПМБД је ефикасна детекција отказа и локализација проблема. Због тога, логика за детекцију отказа треба бити развијена над прикупљеним подацима. На пример, ова логика треба да детектује нагле падове у броју активних корисника и укаже на одређени део мреже где се потенцијално налази узрок пада који ће даље бити анализиран како би се локализовао проблем.

ПНПМБД је прилагођена HFC мрежи. Због тога, перформансне метрике ће се прикупљати са CMTS и CPE уређаја. Један CMTS може опслуживати хиљаде претплатника у зависности од његовог типа и конфигурације. На пример, CMTS Cisco uBR10000 серије у стању је да подржи до 64000 претплатника [97]. Обично, један оператор кабловске мреже има десетине, па чак и стотине CMTS-ова зависно од броја претплатника, дефинисаног протока и конфигурације. С друге стране, број претплатника је много већи од броја CMTS-ова. На пример, према [98], Unitymedia је на крају 2018. године имала 6283000 видео и 3615500 интернет претплатника. Због чињенице да и CM и STB представљају корисничку опрему и да у просеку сваки претплатник има оба сервиса, број уређаја за надгледање перформанси је два пута већи од броја претплатника. Листа најважнијих метрика за надгледање перформанси предложена је у [36] што је један од резултата рада на дисертацији. Листа садржи метрике и за CMTS и за CPE уређаје. ПНПМБД треба да буде у стању да обради све метрике од тако великог броја уређаја. У наставку поглавља ће бити дата проширена листа неопходна за надгледање комплетне мреже.

5.2. Биг дата платформа за надгледање перформанси HFC мрежа

У овом потпоглављу ће бити представљена ПНПМБД архитектура и образложена улога сваке од компоненти. Додатно, биће описан ток прикупљања података кроз архитектуру и начин на који ПНПМБД постиже надгледање перформанси HFC мреже.

Логичка архитектура ПНПМБД-а је приказана на слици 5.2.1. Архитектура се састоји из следећих слојева: надгледана опрема, дата колектори (тј. слој за прикупљање података), биг дата кластер (тј. слој за складиштење и обраду података) и корисници података.



Слика 5.2.1. Логичка архитектура ПНПМБД.

Надгледана опрема представља све уређаје у HFC мрежи као што су CMTS-ови, кабловски модеми, сет-топ боксеви, који су у стању да комуницирају са колектором као и да врше мерења над самим собом. Кабловски модеми и сет-топ боксеви представљају корисничку (CPE) опрему инсталирану на локацији претплатника. Све перформансне метрике са свих мрежних уређаја (CMTS и CPE) се прикупљају путем SNMP протокола. SNMP протокол је изабран због једноставности зато што су сви надгледани уређаји у конкретној мрежи подржавали овај протокол. Табела 5.2.1 приказује листу најважнијих метрика које се могу прикупљати са CMTS и CPE уређаја у циљу надгледања перформанси HFC мреже. Приказане метрике у табели су дате за Cisco CMTS уређај, док су CPE метрике приказане за Cisco кабловски модем. У табели нису наведене метрике за друге произвођаче јер би у супротном табела била превелика (*Variety* проблем). Најчешће је ова листа подржана од стране већине произвођача и може се евентуално разликовати за пар метрика и у вредности OID-а (*Object Identifier*). Додатни детаљи (као опис OID-а и конкретне очекиване вредности) се могу наћи у [99,100]. Додатно, у табели 5.2.1 је дат предлог периоде прикупљања за сваку метрику. Препорука је дата на основу искуства током имплементације, динамике променљивости, и реалних потреба оператора са циљем оптимизације простора. Напомена је да се ове периоде могу произвољно мењати конфигурацијом дата колектора.

Табела 5.2.1. Перформансне метрике.

Извор података	Домен	Перформансна метрика	Периода прикупљања [min]
CMTS	окружење	*cpmCPUTotal5sec* (свеукупно искоришћење процесора у последњих 5 секунди)	5
CMTS	окружење	*ciscoMemoryPoolFree* (слободна RAM меморија)	5
CMTS	окружење	*ciscoEnvMonTemperatureStatusValue* (тренутно мерење температуре)	5
CMTS	CPE групно	*cdxCmtsCmTotal* (укупан број CM-ова)	1
CMTS	CPE групно	*cdxCmtsCmActive* (укупан број активних CM-ова)	1
CMTS	CPE групно	*cdxIfUpChannelCmRegistered* (укупан број регистрованих и активних CM-ова на *upstream*-у)	1
CMTS	CPE групно	*cdxCmtsCmRegistered* (укупан број регистрованих и активних CM-ова на MAC домену)	15
CMTS	интерфејси	*docsIfSigQSignalNoise* (однос сигнал/шум за одређени канал)	5
CMTS	интерфејси	*ccsUpSpecMgmtCNR* (однос носилац/шум за одређени канал)	5
CMTS	интерфејси	*ifInErrors* (укупан број пакета са грешком)	5
CMTS	интерфејси	*ifHCInOctets* (укупан број *upstream* октета примљених на интерфејсу)	5
CMTS	интерфејси	*ifHCOutOctets* (укупан број *downstream* октета послатих на интерфејсу)	5
CMTS	интерфејси	*docsIfUpChannelFrequency* (централна фреквенција канала асоцираног одговарајућем *upstream* интерфејсу)	30
CMTS	*cpe_cmts*	*docsIfCmtsCmStatusValue* (CM вредност статуса)	60
CMTS	*cpe_cmts*	*ccsFlapTotal* (тотални број прескока)	60
CMTS	*cpe_cmts*	*docsIf3CmtsCmUsStatusSignalNoise* (однос сигнал/шум са CM-а на овом *upstream* каналу)	60
CMTS	*cpe_cmts*	*docsIf3CmtsCmUsStatusRxPower* (примљена снага на овом *upstream* каналу)	60
CMTS	*cpe_cmts*	*docsIfDownChannelPower* (операциона предајна снага)	60
CMTS	*cpe_cmts*	*docsIfCmStatusTxPower* (операциона предајна снага за дефинисани *upstream* канал)	60
CMTS	*cpe_cmts*	*docsIfSigQSignalNoise* (просечан однос сигнал/шум на *upstream* нивоу)	60
CMTS	*cpe_cmts*	*ifInOctets* (укупан број примљених октета од CM)	60
CMTS	*cpe_cmts*	*ifOutOctets* (укупан број послатих октета ка CM)	60
CPE	*cpe*	*docsIfDownChannelPower* (пријемна снага)	60
CPE	*cpe*	*docsIfCmStatusTxPower* (предајна снага ка повезаном *upstream* каналу)	60
CPE	*cpe*	*docsIfSigQSignalNoise* (однос сигнал/шум за *downstream* канал)	60
CPE	*cpe*	*ifInOctets* (укупан број примљених октета)	60
CPE	*cpe*	*ifOutOctets* (укупан број послатих октета)	60

Дата колектори су одговорни за прикупљање података са надгледане опреме, форматирање порука и слање истих на биг дата кластер. Међутим, битност података прикупљених са CMTS-а и CPE-а нису исте. Додатно, одзив и доступност CMTS-а је значајно већа у поређењу са CPE-овима. Због тога, регуларна периода прикупљања података за ова два се разликује што се примењује и за перформансне метрике (видљиво у табели 5.2.1 ако се погледају вредности периоде прикупљања). У ПНПМБД, периода прикупљања за CPE уређаје, T_{CPE}, је подешена на један сат, док периода прикупљања за CMTS, T_{CMTS}, варира у опсегу од 1 до 60 минута зависно од метрике која се прикупља. Напомена је да се ови параметри могу произвољно мењати у складу са реалним захтевима мрежног оператора.

Додатно, периода прикупљања током решавања конкретних проблема се по потреби може спустити на ниво секунде.

Слој за прикупљање података користи менаџера који управља листом надгледаних уређаја и додељује колектор уређају. Алгоритам за доделу је једноставан. Менаџер користи конфигурацију колектора како би приступио листи CMTS-ова и CPE-ова повезаних на њих. Конфигурација дата колектора представља резултат *mac-ip-mapping* механизма који ће бити описан касније у секцији 5.2.5. Менаџер прослеђује листу CMTS-ова на *round-robin* начин. Када се са сваког CMTS-а прикупи листа CPE уређаја који су повезани на њега, менаџер шаље податке о CPE на неки од слободних колектора предвиђених за прикупљање података са модема и сет-топ боксева. Колектор се бира на случајан начин како би се извршила правилна дистрибуција. Прослеђена листа, поред MAC адреса, садржи и IP адресе CPE опреме. IP адресе су неопходне за комуникацију са уређајима у случају коришћења SNMP протокола. Један колектор (са 4 процесорска језгра и. 8 GB RAM меморије) у стању је да подржи прикупљање података са до 25000 уређаја за периоду прикупљања T_{CPE} од једног сата. Прикупљени подаци се обогаћују одговарајућим таговима што је објашњено у секцији 5.2.3. Након што се подаци обогате, исти се шаљу на OpenTSDB коришћењем веб сокета. Додатно, копија прикупљених података се у паралели чува и на HDFS-у како би се омогућила пост-агрегација података.

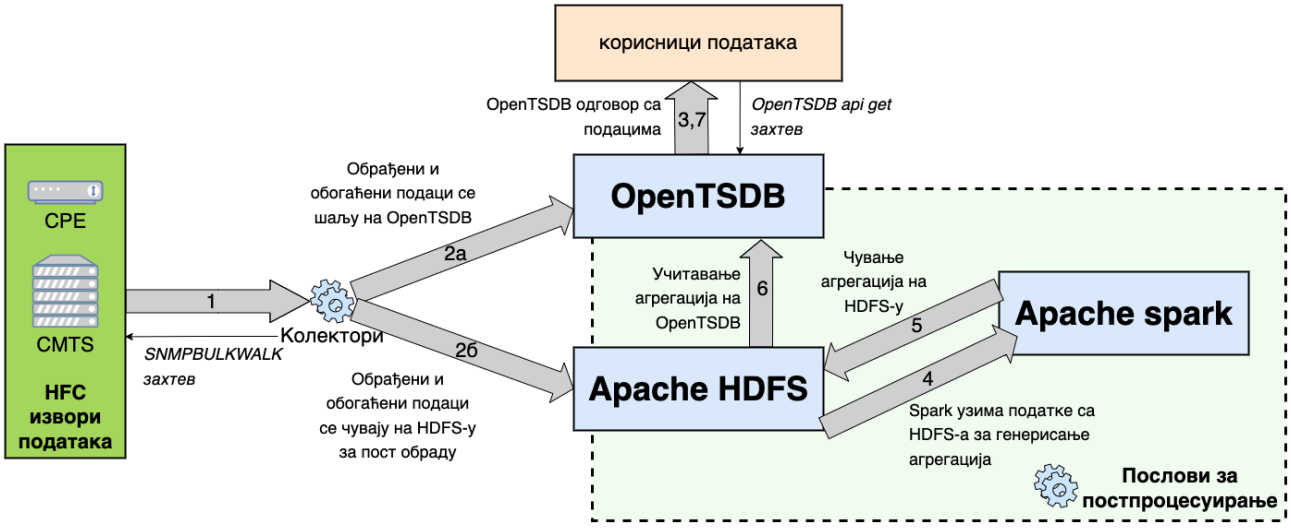
Биг дата кластер извршава операције складиштења и обраде података. Поред података прикупљених са уређаја, овај кластер може додатно да подржи и податке из екстерних апликација. Информације добијене од обраде прикупљених података се користе од стране корисника података. Корисници података могу бити различити извештаји, алармни системи, NOC (*Network Operations Center*) дешборди, call центар извештаји, екстерни системи…

На слици 5.2.1 се види да се биг дата кластер састоји од неколико биг дата алата: OpenTSDB, Apache HBase, HDFS, Apache Spark, Hadoop YARN и Zookeeper. Колектори шаљу прикупљене податке OpenTSDB-у. OpenTSDB прихвата приспеле податке, чита, проверава исправност формата порука и складишти их у одговарајуће HBase табеле. HBase складишти фајлове табела на HDFS, а HDFS складишти фајлове на физичке хард дискове. Spark изводи агрегације над подацима. YARN алоцира ресурсе за послове процесуирања и омогућава различитим радним оквирима да користе заједнички хардвер кластера. YARN се у случају ПНПМБД-а користи као ресурс менаџер послова пост агрегација. Zookeeper обезбеђује синхронизацију између дистрибуираних сервиса. Предложени биг дата кластер задовољава све иницијално постављене захтеве по питању робусности платформе, скалабилности, високе доступности, отпорности на грешке и перформанси. Биг дата кластер представљен на слици 5.2.1 представља централни део ПНПМБД архитектуре.

Дијаграм тока података приказан на слици 5.2.2 сумира претходно описан процес прикупљања и обраде података од извора до крајњих потрошача (корисника података). Колектор шаље захтеве надгледаним уређајима и прихвата податке од извора података (1). Обрађени и обогаћени подаци шаљу се OpenTSDB-у коришћењем веб сокета (2а). У паралели, колектор уписује копију података у фајл и отпрема га на HDFS (2б). Подаци у OpenTSDB-у су спремни за конзумацију одмах након пристизања (3). Подаци сачувани на HDFS се даље обрађују у пост-обради коришћењем Apache Spark радног оквира (4). Агрегације се снимају назад на HDFS (5) и шаљу на OpenTSDB (6) за даљу конзумацију (7).

Прикупљени подаци се користе за надгледање перформанси HFC мреже. Као што је дискутовано у ранијим поглављима постоје бројне предности надгледања перформанси мреже и могућности употребе прикупљених података за извлачење вредних информација.

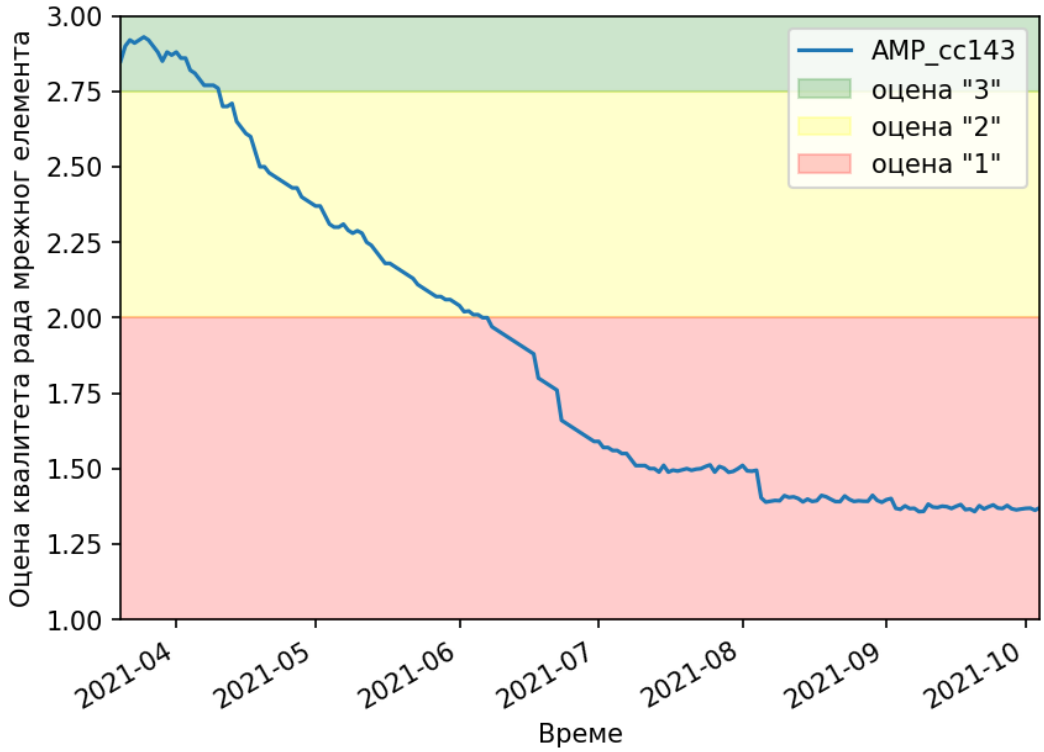
Једна од очигледних предности је ефикасна детекција и локализација отказа у мрежи. Друга предност, веома битна за операторе, представља детекцију слабих тачака у мрежи у циљу превенције и свеукупног побољшања перформанси HFC мреже, а самим тим и побољшања корисничког QoE (*Quality of Experience*).



Слика 5.2.2. Дијаграм тока података у ПНПМБД.

Подаци који се тичу броја онлајн модема на CMTS-овима се константно проверава. Уколико се детектује нагли пад у броју онлајн модема за неки одређени CMTS, истог тренутка се обавештава систем за алармирање. Након тога се врши активација механизма за локализацију проблема. Више речи о овом механизму биће у шестом поглављу, које је и посвећено детекцији и локализацији отказа у HFC мрежи.

Подаци прикупљени са надгледаних уређаја се могу користити за дефинисање квалитета рада уређаја, QoO (*Quality of Operation*). Међутим, у HFC мрежи постоје и неинтелигентни мрежни елементи (ON, AMP, AP) који директно утичу на квалитет мреже, а чији се параметри о квалитету не могу прикупити. Због тога, у циљу процене QoO таквих уређаја, врши се пост-процесуирање података прикупљених са CPE комбинованих са мрежном топологијом. Резултати процене се уписују у OpenTSDB и доступни су за конзумацију као и све друге метрике прикупљене са уређаја. Коришћењем посебно развијене веб апликације, мрежни оператори су у могућности да изврше разне врсте специфичних агрегација над подацима како би проверили стање HFC мреже. Један пример, коришћен у ПНПМБД, је QoO процена надгледаних уређаја. QoO уређаја се може проценити за један период прикупљања и након тога агрегирати на различитим временским интервалима (дневно, недељно, месечно) у циљу свеукупне провере перформанси. Оваква информација се може користити за предиктивно надгледање мреже. У случају кад поједини мрежни елемент има лошу оцену дужи временски период (али и даље ради), такав уређај представља слабу тачку мреже коју је потребно на време заменити или поправити. На слици 5.2.3 је приказан пример квалитета рада појачавача AMP_cc143. Квалитет рада овог уређаја добијен је агрегацијом метрика прикупљених са CPE уређаја који се налазе хијерархијски испод њега. Са слике се може видети како квалитет овог уређаја континуално опада. Коришћење ПНПМБД-а у реалним мрежама је показало да континуално превентивно надгледање перформанси мреже повећава укупан квалитет мреже у дугорочном смислу.



Слика 5.2.3. Оцена квалитета рада мрежног елемента AMP_cc143.

Као што је дискутовано у 2.3, постоји неколико начина за имплементацију биг дата платформе. У конкретном случају, ПНПМБД користи локални хардвер и Cloudera 5.14 дистрибуцију. Имплементација је урађена на 2 namenode/6 datanode кластеру где сваки сервер користи 32 језгра Intel Xeon CPUE5-2630 @ 2.4GHz и 64GB RAM. Током тестирања коришћен је локални хардвер, а не *cloud* технологије због постојања истог и уштеде средстава. Cloudera дистрибуција обезбеђује једноставну инсталацију жељених биг дата компоненти, а самим тим и штеди време потребно за основну поставку система што је главни разлог због ког је иста одабрана приликом развоја.

5.2.1. Метрике

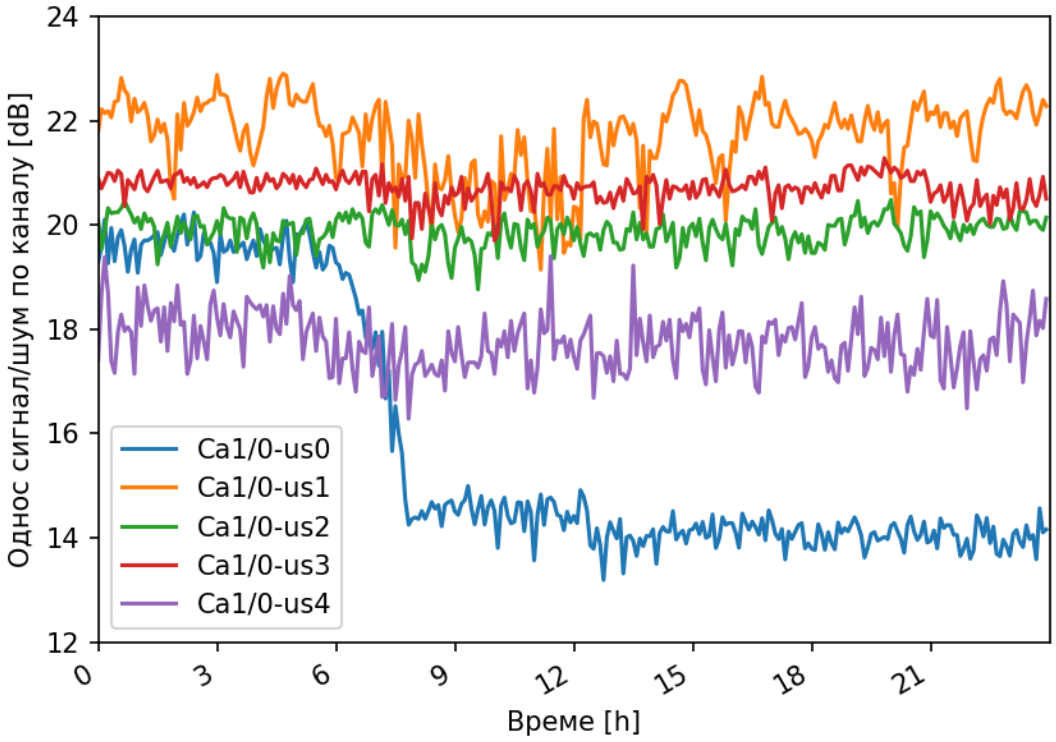
Приликом надгледања перформанси мрежа, неопходно је дефинисати групу метрика за надгледање које ће заједно дати комплетан увид у стање мреже из свих углова. Које ће се тачно метрике прикупљати и са којих уређаја, представља резултат иницијалног истраживања, али и доменског познавања мрежа, у овом конкретном случају HFC мрежа, али и могућности уређаја да врше мерење и слање перформанси, односно метрика. Најбољи могући сценарио би био када би сваки мрежни елемент могао да измери и пошаље своје стање и перформансне статистике. На жалост, постоје елементи у мрежи као што су ON, AMP, AP, који немају такве функционалности. Једини уређаји у HFC мрежи који су у стању да пруже податке за надгледање су CMTS и CPE (CM и STB) уређаји [36].

CMTS је уређај који се користи за пружање сервиса корисницима. Обзиром на то да је овај уређај централни за HFC мрежу или део HFC мреже, и да опслужује велики број корисника, од велике је важности да овај уређај ради квалитетно и без проблема. Због своје комплексне улоге у мрежи, овај уређај има могућност мерења својих перформанси у

различитим доменима. Метрике које се прикупљају са CMTS-ова и које су релевантне за ово истраживање спадају у следеће домене:

- Окружење,
- Групне метрике о корисницима (CPE групно),
- Интерфејси (*upstream, downstream,* MAC домен),
- CPE метрике (*cpe_cmts*).

Статистике о окружењу дају информацију о генералном стању CMTS-а као што су, на пример, оптерећење процесора, заузеће меморије и температура. Групне статистике о корисницима дају свеукупне информације о повезаним уређајима. Неке од метрика које спадају у ову групу су број регистрованих, активних, онлајн CPE уређаја. Информације о *upstream*-овима, *downstream*-овима и MAC доменима спадају под метрике о интерфејсима. Неке од метрика из ове групе су SNR (*Signal to Noise Ratio*) и CNR (*Carrier to Noise Ratio*). Коначно, на CMTS-у се могу прикупљати и метрике о самим корисницима. Ове метрике идентичне су метрикама прикупљаним са CPE опреме, али на пријему CMTS-а. Неке од метрика из ове групе су SNR и снага сигнала на пријему на CMTS страни [36]. На слици 5.2.1.1, приказан је пример једне метрике која се прикупља за CMTS-а. Наиме, овде се може видети однос SNR за један CMTS за његове *upstream*-ове. У конкретном примеру се може видети значајан пад односа сигнал/шум за *upstream Ca1/0-us0* почев од 6 часова.



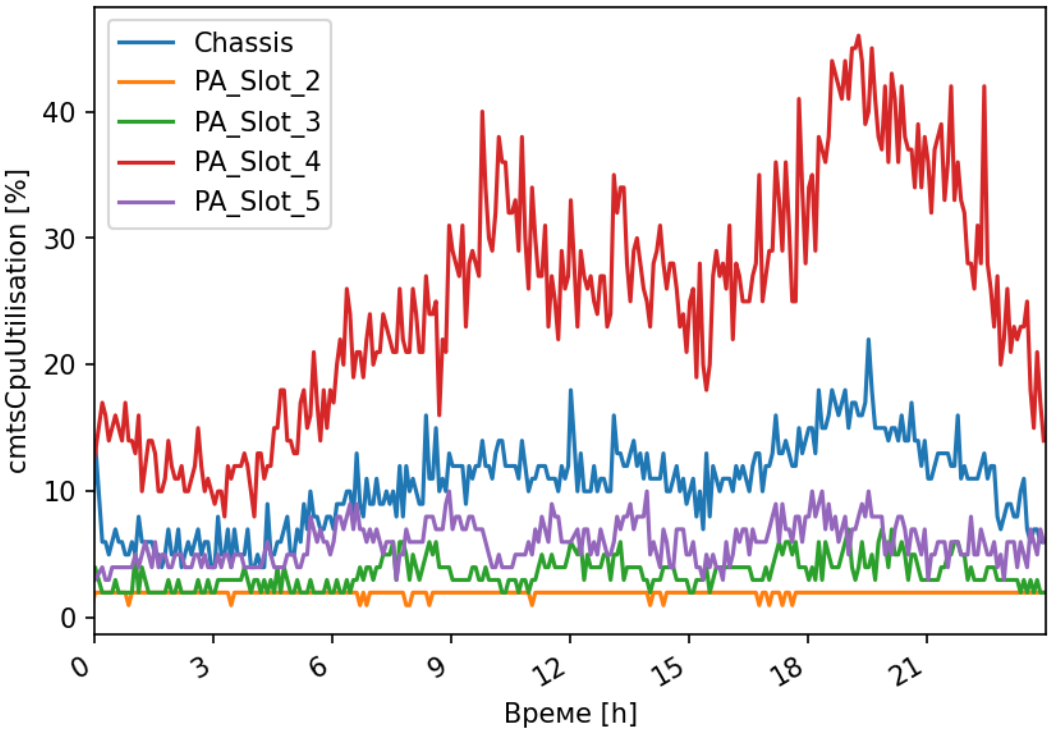
Слика 5.2.1.1. Однос сигнал/шум по каналу.

CPE статистике дају информације о сваком повезаном кабловском модему и сет-топ боксу. Ове статистике су нарочито важне јер дају увид у стање мреже из корисничког угла. Поред тога, ове метрике се могу користити приликом решавања проблема за појединачни уређај, тј. за појединачног корисника. Неке од метрика које се могу прикупљати са ових уређаја су пријемна и предајна снага сигнала на CPE, број примљених и послатих октета…

У табели 5.2.1 у потпоглављу 5.2 дат је преглед листе најважнијих метрика које се могу прикупљати са CMTS и CPE уређаја у циљу надгледања перформанси HFC мреже. Ова листа представља препоруку и, у складу са реалним потребама, може се проширити другим метрикама [99,100]. Приликом додавања нове метрике треба обратити пажњу којој групи иста припада и у складу с тим доделити јој одговарајуће тагове што ће бити детаљно описано у секцији 5.2.3.

5.2.2. Дата шема

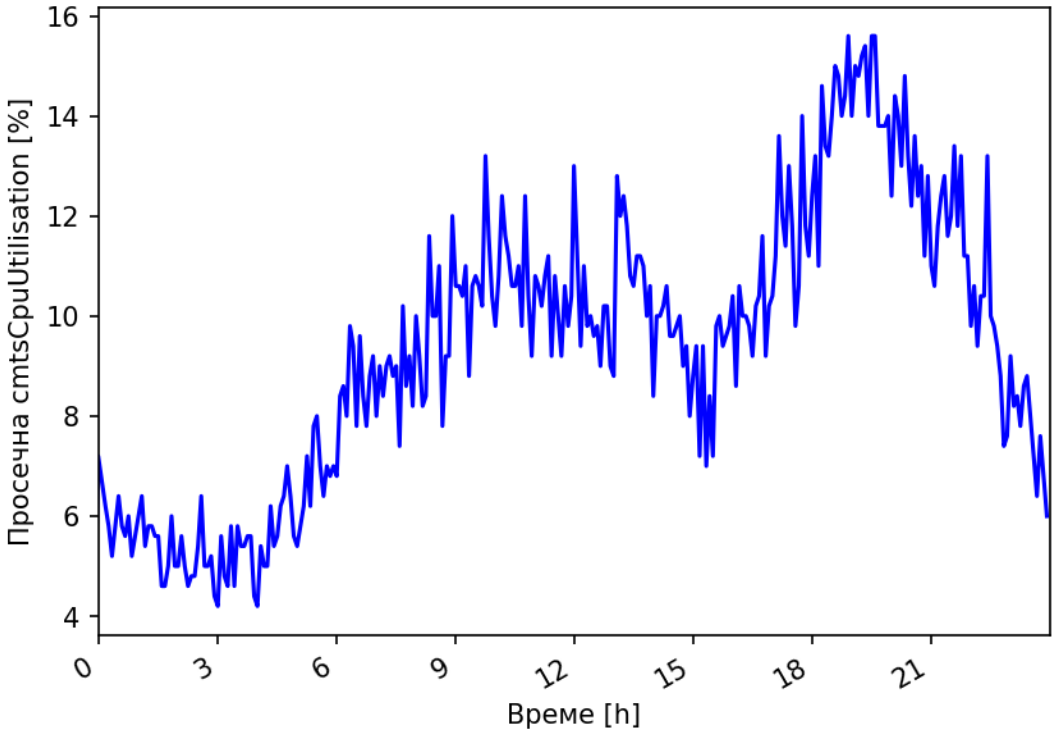
Као што је објашњено у потпоглављу 5.2, ПНПМБД користи OpenTSDB да складишти перформансне метрике прикупљене из HFC мреже. Према [38], OpenTSDB пружа неколико начина за анализу и манипулацију над прикупљеним подацима. У овој секцији су дискутоване могућности за имплементацију шеме података, а дат је и предлог шеме података која побољшава перформансе упита што представља и један од доприноса ове дисертације.



Слика 5.2.2.1. Оптерећење процесора CMTS-а по језгру.

Коришћењем тагова, могуће је раздвојити временске серије са различитих извора. На овај начин, прикупљени подаци за једну одређену метрику и различит скуп тагова се могу лако посматрати, што појединачно што групно [38], коришћењем опција за филтрирање и групацију. На пример, на слици 5.2.2.1 је приказано CMTS CPU (*Central Processing Unit*) искоришћење по сваком језгру (таг *entity_name*) за један дан. Приказани график на овој слици генерисан је коришћењем сирових података прикупљених са TEST-CMTS CMTS-а који има пет процесорских језгара (таг ”*entity_name*” приказан у легенди на слици 5.2.2.1). Пример сирових података се може видети на слици 5.2.2.3 где се уједно може видети и таг ”*entity_name*”. Обзиром да приликом генерисања упита није вршена филтрација, све временске серије постоје на графику. Коришћењем филтрације по тагу, могуће је посматрати само појединачну серију уместо комплетног скупа. На пример, уколико би филтар био подешен на *entity_name=Chassis*, график на слици 5.2.2.1 би приказао само одговарајућу

временску серију. Поред филтрације, друга основна функционалност OpenTSDB-а је груписање. Груписање агрегира више временских серија у једну по дефинисаном типу здруживања (просечна вредност, максимум, минимум…) [38]. Пример груписања је приказан на слици 5.2.2.2. На слици 5.2.2.2 је могуће видети просечно искоришћење процесора за исти скуп података који је приказан на слици 5.2.2.1. Просечно CPU искоришћење се добија аутоматски од стране OpenTSDB-а изостављањем тага ”*entity_name*” из упита и подешавање ”*avg*” типа агрегације. Поред стандардних агрегационих типова, OpenTSDB подржава даунсемплинг (енгл. *downsampling*) као и неке друге типове агрегације о којима се може више прочитати у документацији производа.



Слика 5.2.2.2. Просечна вредност оптерећења процесора CMTS-а.

```
cmtsCpuUtilisation 1584658859 6 entity_name=Chassis device_name=TEST-CMTS device_type=CMTS company=comp_A
cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation 1584658859 4 entity_name=PA_Slot_3 device_name=TEST-CMTS device_type=CMTS company=comp_A
cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation 1584658859 2 entity_name=PA_Slot_2 device_name=TEST-CMTS device_type=CMTS company=comp_A
cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation 1584658859 14 entity_name=PA_Slot_4 device_name=TEST-CMTS device_type=CMTS company=comp_A
cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation 1584658859 3 entity_name=PA_Slot_5 device_name=TEST-CMTS device_type=CMTS company=comp_A
cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation 1584658859 68 entity_name=PA_Slot_1 device_name=TEST-CMTS2 device_type=CMTS company=comp_A
cmts_name=TEST-CMTS2 entity_type=CPU manufacturer=Huawei device_ip=192.168.212.99
cmtsCpuUtilisation 1584658859 42 entity_name=PA_Slot_2 device_name=TEST-CMTS2 device_type=CMTS company=comp_A
cmts_name=TEST-CMTS2 entity_type=CPU manufacturer=Huawei device_ip=192.168.212.99
```

Слика 5.2.2.3. Пример прикупљених података форматираних за OpenTSDB.

Слика 5.2.2.3 садржи временске одбирке за једну метрику (*cmtsCpuUtilisation*) са два различита уређаја (TEST-CMTS и TEST-CMTS2) прикупљене у једном временском интервалу (1584658859 у *epoch* формату). Када корисник креира упит за прикупљање података за одређену метрику, OpenTSDB извршава филтрацију података на основу прослеђеног скупа тагова и дефинисаног временског оквира. Време извршавања упита је директно пропорционално броју одбирака прикупљених за метрику од интереса у једној итерацији. У HFC мрежама, постоје милиони CPE уређаја од којих сваки има више интерфејса са којих се прикупљају подаци. Због тога, извлачење података за један конкретан уређај и његове ентитете (интерфејсе) може бити екстремно споро и самим тим не може да задовољи иницијалне захтеве по питању перформанси читања података.

Како би се превазишао овај проблем, развијена је нова шема података. У новој шеми се јединствени идентификатор уређаја (то може бити име уређаја, MAC адреса, назив хоста (*hostname*) за CMTS) припаја имену метрике. У примеру са слике 5.2.2.3, метрика *cmtsCpuUtilisation* прикупљена за уређаје TEST-CMTS и TEST-CMTS2 ће бити сачувана кроз нове метрике под именом *cmtsCpuUtilisation_TEST-CMTS* и *cmtsCpuUtilisation_TEST-CMTS2*, као што је приказано на слици 5.2.2.4. На овај начин, нова метрика ће бити креирана за сваки уређај. Ово је разумна одлука јер се упити у највећем броју случајева креирају за одређени уређај, а не за групу уређаја (на пример, неки од извештаја су кол центар извештаји или извештаји за решавање проблема за одређени CPE). Предложена шема побољшава перформансе упита у просеку 1300 пута за један конкретан уређај што је добијено на основу изведених експеримената. За експерименте је коришћено милион уређаја, сваки са четири интерфејса. Иницијално време упита за један уређај је било 307.041 секунда што је предложеном шемом података смањено на 0.234 секунде. Тестирање је обављено на кластеру чије су спецификације дате на крају описа архитектуре ПНПМБД, тј. непосредно испред секције 5.2.1.

```
cmtsCpuUtilisation_TEST-CMTS 1584658859 6 entity_name=Chassis device_name=TEST-CMTS device_type=CMTS
company=comp_A cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation_TEST-CMTS 1584658859 4 entity_name=PA_Slot_3 device_name=TEST-CMTS device_type=CMTS
company=comp_A cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation_TEST-CMTS 1584658859 2 entity_name=PA_Slot_2 device_name=TEST-CMTS device_type=CMTS
company=comp_A cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation_TEST-CMTS 1584658859 14 entity_name=PA_Slot_4 device_name=TEST-CMTS device_type=CMTS
company=comp_A cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation_TEST-CMTS 1584658859 3 entity_name=PA_Slot_5 device_name=TEST-CMTS device_type=CMTS
company=comp_A cmts_name=TEST-CMTS entity_type=CPU manufacturer=Cisco device_ip=192.168.88.123
cmtsCpuUtilisation_TEST-CMTS2 1584658859 68 entity_name=PA_Slot_1 device_name=TEST-CMTS2 device_type=CMTS
company=comp_A cmts_name=TEST-CMTS2 entity_type=CPU manufacturer=Huawei device_ip=192.168.212.99
cmtsCpuUtilisation_TEST-CMTS2 1584658859 42 entity_name=PA_Slot_2 device_name=TEST-CMTS2 device_type=CMTS
company=comp_A cmts_name=TEST-CMTS2 entity_type=CPU manufacturer=Huawei device_ip=192.168.212.99
```

Слика 5.2.2.4. Пример података форматираних за OpenTSDB са модификованом шемом.

Недостатак предложене модификације шеме је немогућност креирања групних упита над уређајима за одређени таг. Један пример таквог упита је извлачење свих CPE уређаја одређеног типа. Међутим, упити овог типа нису временски критични у оперативном раду, односно није неопходно генерисати резултате истог тренутка. Због тога, овакве агрегације се могу вршити коришћењем Spark дневних агрегација.

Уколико се у решењу користи предложена шема, треба бити опрезан са бројем креираних метрика. Према [38], метрике су у HBase-у кодиране коришћењем 3 бајта, дајући

2^{3*8}-1=16777215 јединствених метрика. Међутим, овај параметар је конфигурабилан од OpenTSDB верзије 2.2, па се једноставно може повећати. Предлог за решење овог изазова дат је у секцији 5.2.5.

5.2.3. Дата колектори

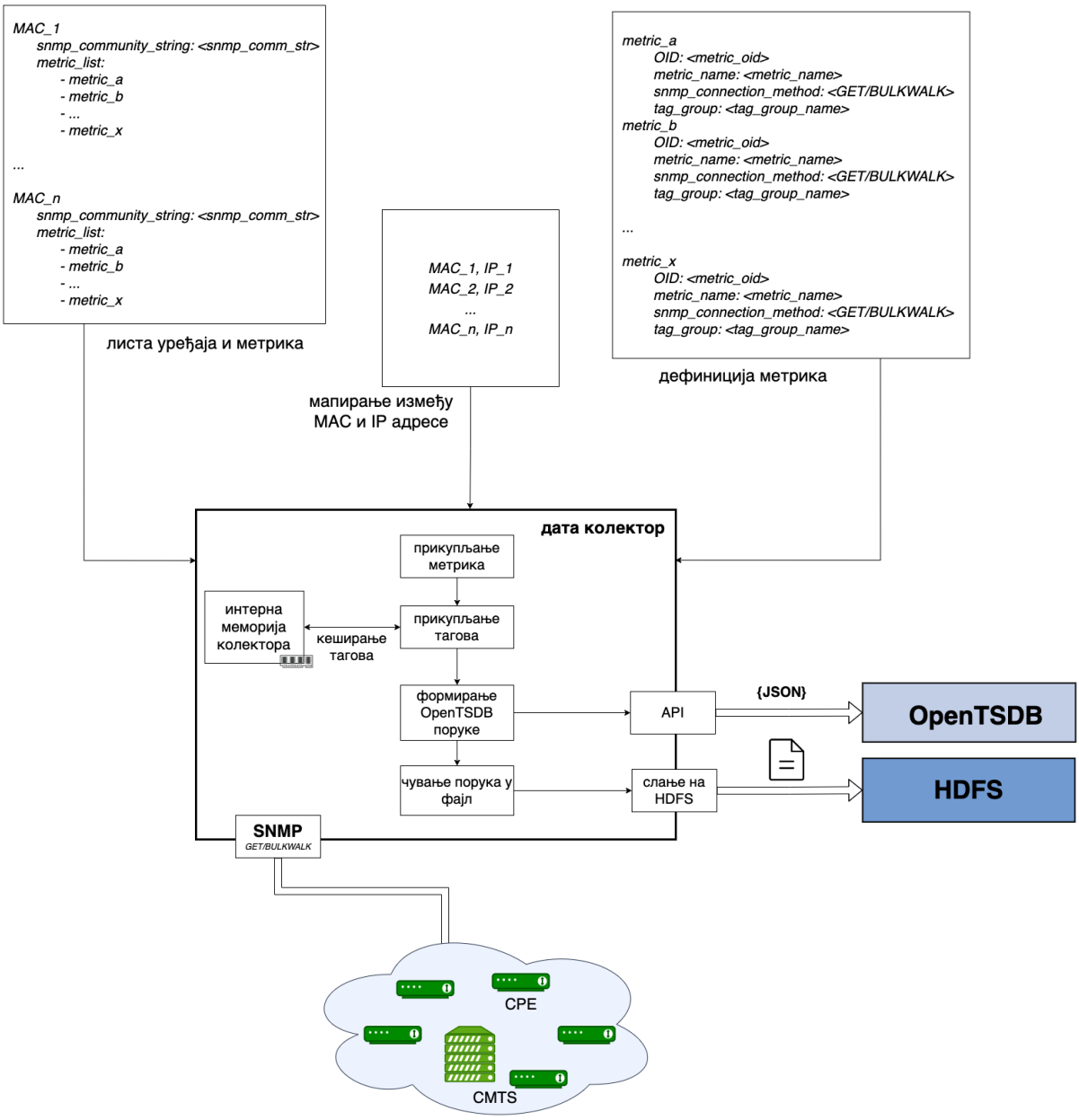
Након што је иницијално постављена архитектура за складиштење, потребно је дефинисати слој за прикупљање података, односно дата колекторе. Дата колектори су компоненте које прикупљају податке са уређаја, обогаћују их, формирају поруке и шаљу на слој за складиштење података, што су OpenTSDB и HDFS у случају ПНПМБД-а. Постоје јавно доступни колектори који се могу користити [38]. Међутим, колектори овог типа нису у стању да покрију специфичности које су неопходне у ПНПМБД-у. Због тога, неопходно је било развити нов дата колектор што представља један од доприноса ове тезе.

Полазне претпоставке за развој колектора проистичу из ”5V” изазова. Наиме, приликом развоја колектора неопходно је обратити пажњу на неколико ствари. Колектор треба да буде у стању да ефикасно прикупља податке са великог броја уређаја. Током развоја колектора потребно је узети у обзир недоступност уређаја што је валидан случај јер укључивање и искључивање CPE опреме зависи од клијента и самим тим ти уређаји се не могу сматрати високо доступним. Колектор треба да буде ефикасан и да обезбеди могућност прикупљања на различитим периодама, од нивоа секунда, до нивоа сата. У зависности од конфигурације HFC мреже, CPE опрема може добити динамичку IP адресу за комуникацију. С тим у вези, неопходно је константно пратити мапирање између уређаја и његове адресе.

Поред метрика, дата колектор је задужен и за прикупљање одређеног скупа тагова. Тагови су информација у виду *key/value* парова које једнозначно дефинишу извор мерења [38]. Измерене метрике немају много вредности без тагова. Постоје два типа тагова, обавезни и опциони. Током развоја дата колектора, неопходно је истражити све тагове који описују једно мерење, идентификовати обавезне и филтрирати само оне од интереса. Обавезни тагови су неопходни како би се једнозначно дефинисао извор података. У примеру приказаном на слици 5.2.2.3 обавезни тагови су *device_name* и *entity_name* јер раздвајају метрику искоришћења процесора и по уређају и по језгру. Уколико су неки од обавезних тагова изостављени, прикупљена мерења ће имати исти скуп тагова што ће изазвати преклапање у подацима, па самим тим до нетачне и неконзистентне информације. Опциони тагови се користе да додатно обогате мерења. Што више тагова постоји, то се више различитих агрегација може извести. На пример, произвођач и модел CPE уређаја нису обавезни у току прикупљања података. Међутим, ови тагови се могу касније користити у агрегацијама и анализама које би дале увид у перформансе уређаја сваког произвођача и модела што је значајна информација приликом набавке нове опреме. Не треба претеривати са опционим таговима, већ треба искористити оне који заиста имају смисла у том тренутку. Уколико се испостави да је неки таг важан, а није додат приликом иницијалног развоја дата колектора, исти се може накнадно додати без угрожавања претходно прикупљених података. Треба имати у виду да ће новопридодати таг бити доступан од тренутка додавања, али не и за претходну историју.

Размена информација између дата колектора и извора података је заснована на клијент-сервер комуникацији где су дата колектори клијенти, а извори података сервери. Обзиром на то да је у питању развој дата колектора за специфичну намену, интеграција са изворима података се може обавити на разне начине, зависно од могућности извора. REST API (*Representational State Transfer Application Programming Interface*), JDBC (*Java Database Connectivity*), SNMP, HTTP, *subscriber-broker* комуникација су само неки од могућих начина

комуникације. Обзиром да и CMTS и CPE опрема подржавају SNMP протокол за комуникацију, у наставку је предложена архитектура дата колектора заснована на овом протоколу.



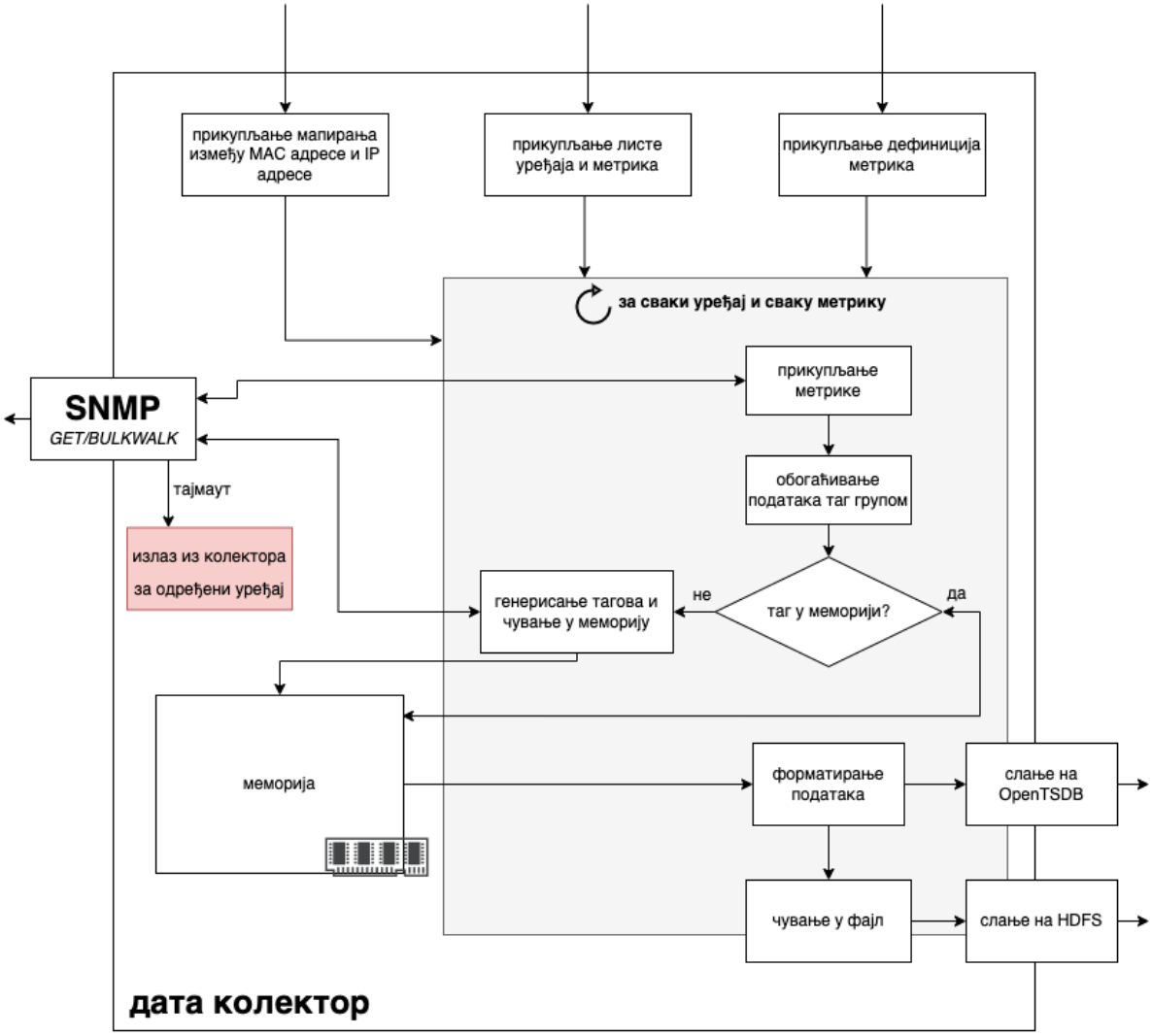
Слика 5.2.3.1. Архитектура дата колектора.

Архитектура колектора приказана је на слици 5.2.3.1. За конфигурацију колектора, потребна су три улазна фајла. Фајл са дефиницијом метрика садржи све неопходне информације за прикупљање једне метрике. За једну метрику неопходно је познавати који ће се OID користити, који метод прикупљања (SNMPGET или SNMPBULKWALK), које је име метрике и таг група која ову метрику дефинише. Таг група представља груписање свих тагова, и неопходних и опционих, који ће се придружити прикупљеној метрици. Мапирање тагова се дефинише у дата колектору и специфично је за сваки тип уређаја, верзију софтвера и произвођача. Пример дефиниције једне метрике дат је у табели 5.2.3.1. Фајл са дефиницијом уређаја садржи списак уређаја са којих ће се прикупљати подаци на одређеном

дата колектору и које метрике се прикупљају за који уређај. Имена метрика морају одговарати именима дефинисаним у фајлу за прикупљање. Трећи неопходни фајл је неопходан само за CPE опрему и представља мапирање између MAC адресе и IP адресе. Овај податак се може лако добити пропитивањем CMTS-а по одговарајућим OID-има. Док је фајл са дефиницијом метрика заједнички за све колекторе, остала два специфична су за сваку инстанцу дата колектора јер исти дефинишу листу уређаја са којих ће дотична инстанца прикупљати податке.

Табела 5.2.3.1. Пример дефиниције метрике.

CPU_Cisco
OID: 1.3.6.1.4.1.9.9.109.1.1.1.1.3 ***metric_name:*** cpmCPUTotal5sec ***snmp_connection_method:*** SNMPBULKWALK ***tag_group:*** environment_cisco

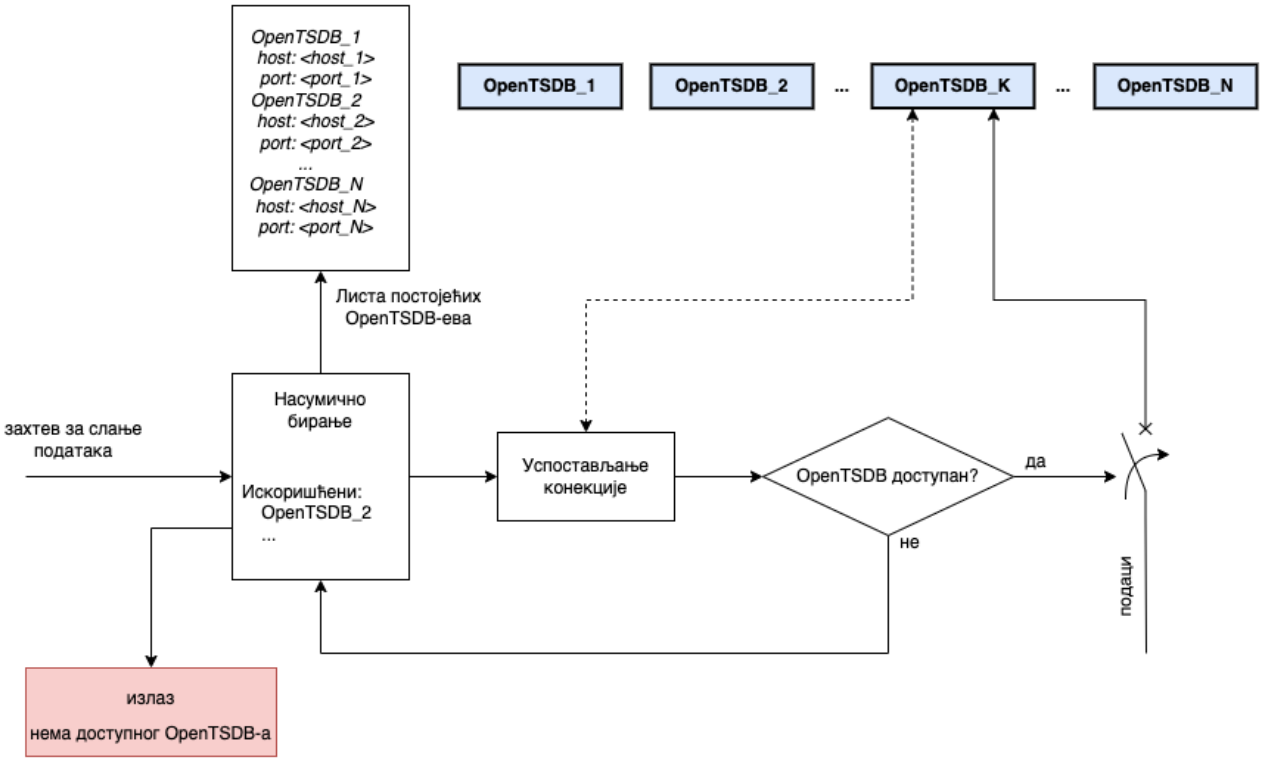


Слика 5.2.3.2. Детаљна шема дата колектора.

Дата колектор очитава листу за прикупљање и секвенцијално покушава да прикупи дефинисане метрике за сваки уређај. Најпре се прикупља метрика. Уколико уређај одговори

дата колектору, поред очитане метрике, то значи да је уређај доступан и иде се даље у наставак прикупљања. Након тога се врши очитавање тагова. Прикупљени тагови се смештају у кеш меморију како би исти могли бити искоришћени и за друге метрике. На овај начин се добија на перформансама дата колектора јер се штеди време на непотребној поновној комуникацији. Кад су сви тагови прикупљени, врши се формирање OpenTSDB порука и њихово слање на дефинисану базу. У паралели, подаци се такође смештају у фајл и њихова копија се чува на HDFS-у.

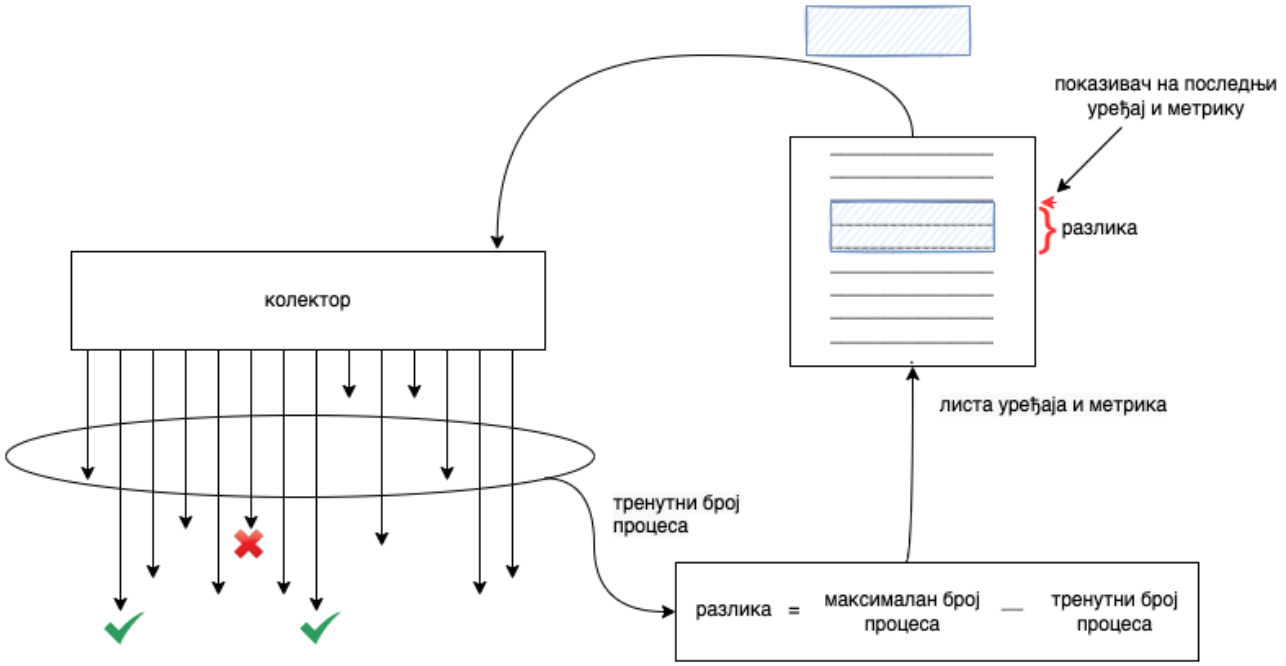
На слици 5.2.3.2 је приказана детаљна шема дата колектора. За сваки уређај и сваку метрику колектор покушава најпре да успостави контакт слањем SNMP упита за метрику. Уколико надгледани уређај не одговори за дефинисано време и број поновних покушаја, сматра се да је уређај недоступан и прескаче се прикупљање података за исти. Након што је метрика прикупљена, за дефинисану групу тагова, врши се прикупљање једног по једног у циљу формирања поруке. Пре него што се изврши упит, проверава се да ли потребан таг већ постоји у меморији. Уколико постоји, таг се додељује метрици, а у супротном се врши упит ка уређају и резултати смештају у меморију. Обзиром на то да се једним упитом може добити више тагова одједном и да метрике имају велики број заједничких тагова, оваква поставка значајно штеди време приликом прикупљања, а и растерећује надгледани уређај од непотребних (редудантних) упита. Након тога се врши форматирање поруке у складу са OpenTSDB форматом и врши се слање и чување у фајл након чега и слање на HDFS. Време очитавања поруке се бележи тик пре него што се пошаље упит за метрику ка уређају. Веома је важно да време буде што ближе времену очитавања метрике. Метрике попут бројача (на пример, улазних и излазних октета) су посебно осетљиве уколико време није добро усклађено са временом прикупљања и може довести до пикова и аномалија приликом рачунања протока.



Слика 5.2.3.3. Слање података на OpenTSDB.

Пре него што се подаци пошаљу преко веб сокета, дата колектор проверава да ли је OpenTSDB доступан како би могао да успостави конекцију. Наиме, како би се обезбедила висока доступност OpenTSDB-а, у предложеној архитектури инсталирано је више инстанци које деле заједничке HBase табеле. Провера OpenTSDB-а се врши због могућности загушења услед велике количине долазног саобраћаја. Приликом слања, дата колектор приступа листи OpenTSDB-ева и насумично бира један, проверава његову доступност и шаље податке. Уколико одабрани OpenTSDB тренутно није доступан, бира се следећи из листе на идентичан начин. Поступак се понавља док се не пронађе доступна инстанца. Насумично бирање OpenTSDB-а обезбеђује униформну дистрибуцију долазног саобраћаја. Уколико се деси да нема доступног OpenTSDB-а, дата колектор обуставља слање података. Наведени принцип слања података на OpenTSDB је илустрован на слици 5.2.3.3.

Због специфичне дата шеме која се користи (метрике су даље подељене у *metric_device-name* као што је описано у секцији 5.2.2.), подаци сачувани у OpenTSDB нису погодни за пост агрегације. Како би се превазишао овај изазов, дата колектор, поред података послатих на OpenTSDB, чува копију података у текстуалном фајлу која се шаље директно на HDFS. Колико ће се дуго подаци чувати на HDFS-у зависи од потреба агрегације кабловског оператора. Најчешће се ти подаци чувају до 10 дана. Треба имати у виду да се период чувања сирових података може повећати, али да то директно утиче на хардверске захтеве по питању меморије за складиштење.



Слика 5.2.3.4. Паралелно прикупљање података.

Приликом прикупљања података, значајно време се потроши на чекање уређаја на одговор. Секвенцијално прикупљање података (један по један уређај) може довести до неефикасног искоришћења хардвера дата колектора. Како би се максимизовало искоришћење хардвера, дата колектор је имплементиран да паралелно прикупља податке. На слици 5.2.3.4 је приказан начин функционисања овог механизма. Дата колектор у паралели прикупља податке са више уређаја. Број уређаја дефинисан је максималним бројем процеса. Овај број може да варира тј. зависи од хардвера (броја процесорских језгара, меморије) на ком је дата колектор инсталиран и потребно га је дефинисати емпиријски за развијени дата

колектор. На сваких пар секунди врши се пребројавање тренутног броја процеса и рачуна се разлика, тј. број нових уређаја са којих се могу прикупљати подаци. Приступа се листи уређаја и метрика и узима се нови скуп уређаја за прикупљање. Како би се пратио пролазак кроз листу уређаја користи се показивач који указује на последњи уређај и метрику. Механизам се понавља док се не покрене прикупљање за све уређаје.

Дата колектор користи одговарајући *community* стринг и IP адресу за комуникацију са уређајима. Комуникација се обавља путем SNMPGET и SNMPBULKWALK зависно од тога да ли се у одговору очекује једна или више порука. Због својих супериорних перформанси за комуникацију се користи SNMPBULKWALK уместо класичног SNMPWALK-а. У циљу ограниченог приступа, препорука је конфигурисати SNMP протокол надгледане опреме само са дозволама читања података. Поред читања, SNMP протокол има могућност писања, тј. постављања вредности коришћењем SNMPSET. Овај параметар се може користити приликом задавања наредби уређају, на пример, за извршавања спектралне анализе. У првом кораку се пошаље SNMPSET наредба, уређај изврши мерење, а у другом кораку, са задршком од пар секунди, се очитају спектралне вредности. Овакав начин комуникације није коришћен у основном дата колектору и представља специфичан случај употребе. Овим се само наглашава додатна могућност коришћења што може представљати полазну тачку за даља проширења функционалности колектора.

SNMP конфигурација по питању тајмаута и поновних покушаја може значајно да утиче на свеукупне перформансе дата колектора. У ситуацијама када је надгледани уређај високо доступан, као што је CMTS у случају HFC мреже. Поновни покушај и тајмаут се за овакве уређаје могу поставити на више вредности (на пример, тајмаут 2 секунде, 3 поновна покушаја). Међутим, приликом надгледања CPE опреме, не постоји гаранција да је опрема заиста и доступна. Поред тога, ред величине CPE опреме у HFC мрежи може бити у стотинама хиљада, па чак и милионима. То значи да се значајан део времена проведе у чекању недоступних уређаја, а самим тим врши се резервација и трошење рачунарских ресурса без потребе. Тајмаут и број поновних покушаја треба одабрати са великом пажњом како би се минимизовао овај феномен. Тестирање у реалним условима је показало да тајмаут од једне секунде и само један покушај успоставе комуникације даје одличан резултат за пропитивање CPE опреме. Уколико се уређај не одазове након једне секунде, велика је вероватноћа да је уређај недоступан. У оваквом приступу постоје и уређаји који су недоступни или због загушења у мрежи или тренутног преоптерећења, међутим, показало се у пракси да је број таквих уређаја занемарљиво мали. Предлог иницијалне поставке параметара за комуникацију путем SNMP протокола по типу уређаја за постизање најбољих могућих перформанси по питању одзива уређаја и лажно недоступних уређаја представљен је у табели 5.2.3.2. Предложени параметри добијени су тестирањем над реалном мрежом.

Табела 5.2.3.2. Предлог поставке SNMP параметара у зависности од типа уређаја.

Извор података	Тајмаут [секунда]	Број поновних покушаја
CMTS	2	3
CPE	1	1

Дата колектори се креирају по домену интеграције. Домен интеграције представља логичку јединицу за коју се врши надгледање. На пример, интеграционе јединице могу бити DOCSIS, MPLS (*Multiprotocol Label Switching*), UPS (*Uninterruptible Power Supply*), QoS... Интеграциони домени се у појединим случајевима раздвајају даље у произвођаче, моделе

уређаја, па чак и верзије инсталираног софтвера на уређајима. На пример, Cisco CMTS-ови имају различите OID-е и другачију MIB (*Management Information Base*) структуру стабла него CASA или Motorola CMTS-ови. Додатно, Cisco Remote PHY CMTS платформа се разликује од стандардне Cisco HFC имплементације [101].

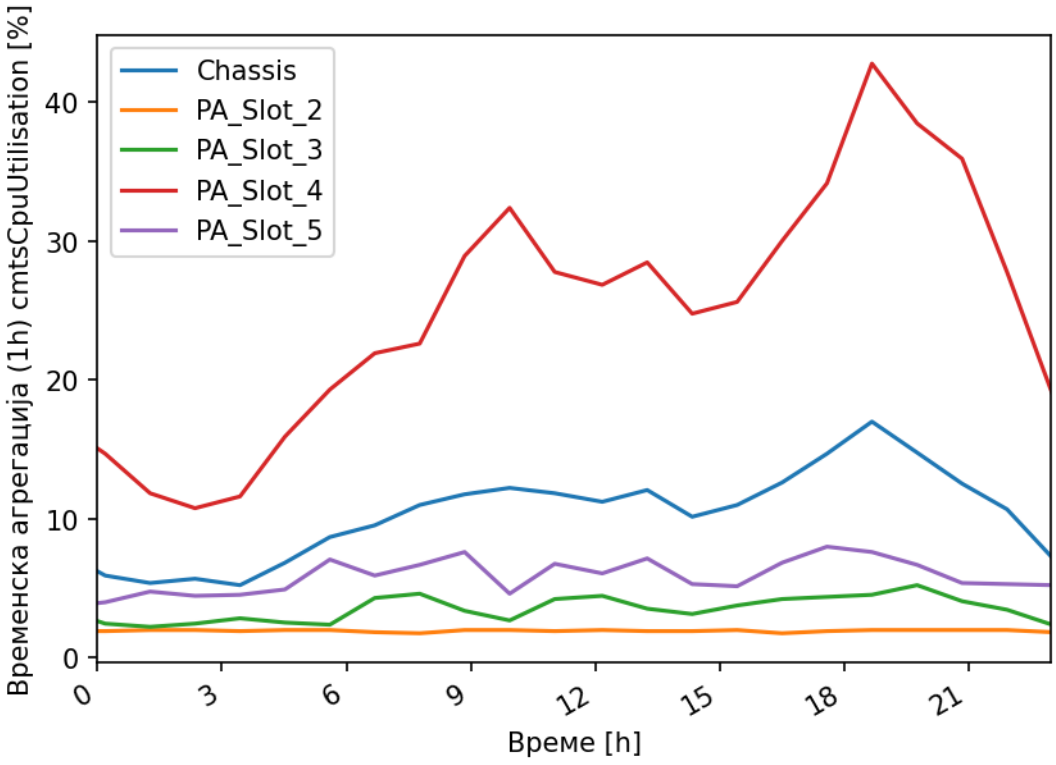
Интеграциони домени од интереса за ПНПМБД су DOCSIS домен за CMTS и CPE опрему. DOCSIS домен покрива перформансне метрике дефинисане у табели 5.2.1, као што су:

- корисничке, метрике окружења и статистике интерфејса за CMTS,
- CPE статистике прикупљене са CMTS-а (метрике прикупљене на CMTS-у, а тичу се CPE),
- CPE статистике прикупљене директно са CPE опреме.

Домени као што су MPLS, WiFi, IoT и други представљају део будућих истраживања. Приликом развоја дата колектора, посебна пажња је дата перформансама колектора због великог броја надгледаних уређаја и оптимизацији хардвера потребног за прикупљање података.

5.2.4. Агрегације података

Сирови подаци прикупљени директно из мреже су корисни за надгледање перформанси, решавање проблема и осталих свакодневних операција. Прикупљени подаци се могу агрегирати како би пружили бољи увид у стање мреже. Коришћењем одговарајућих агрегација, телекомуникациони оператори су у стању да добију информације које могу бити значајне за доношење даљих одлука и планирања. У оквиру ове секције су приказане све могућности ПНПМБД-а када је реч о агрегацијама.



Слика 5.2.4.1. Временски агрегирана просечна вредност оптерећења процесора CMTS-а.

У зависности од циља који се жели постићи, постоји неколико различитих типова агрегације. Агрегације се могу поделити у временске и просторне. Временске агрегације представљају агрегације података прикупљених током одређеног временског периода (на пример, часовно, дневно, недељно, месечно) у једну тачку. Просторна агрегација представља агрегацију различитих извора података у једном тренутку.

На слици 5.2.4.1 приказан је пример временске агрегације. Агрегација је извршена над подацима оптерећења процесора приказаном оригинално на слици 5.2.2.1. За сваки од процесора израчунато је просечно оптерећење на нивоу једног сата. Са слике се може закључити како временска агрегација може помоћи у ситуацијама сигнала са великом варијансом тј. великим варијацијама.

На слици 5.2.2.2 представљен је пример просторне агрегације. Процесорско оптерећење сваког језгра TEST-CMTS-а агрегирано је у јединствену просечну вредност за сваку итерацију прикупљања. На овај начин, генерисана је информација о свеукупном оптерећењу CMTS-а. Посматрање оптерећења процесора је много једноставније посматрањем ове агрегације уместо појединачног посматрања јер се, на овај начин, ублажавају привремени пикови оптерећења појединих процесорских језгара. Зависно од времена извршавања агрегација, постоје пост агрегације и стрим агрегације (агрегације у реалном времену). Пост агрегације се спроводе након што су подаци прикупљени. С друге стране, стрим агрегације користе тек генерисане податке и врше агрегацију у истом тренутку. У наставку ће бити представљене само пост агрегације обзиром на то да се исте тренутно користе у ПНПМБД. Агрегације стримова представљају део будућег истраживања.

ПНПМБД користи биг дата алате за агрегацију како би изашао на крај с великом количином података. У поређењу са традиционалним методама за обраду података, биг дата врши дистрибуирано програмирање на *master/slave* начин. Овај приступ је скалабилан јер користи дистрибуцију оптерећења на више сервера. У оквиру ПНПМБД решења користи се Apache Spark за агрегацију података.

У ПНПМБД-у постоје два главна типа агрегације:

- OpenTSDB агрегације
- пост агрегације

OpenTSDB је у стању да обави и временске и просторне агрегације коришћењем уграђених функционалности. Просторне агрегације се регулишу груписањем тагова приликом креирања упита. Временске агрегације се извршавају коришћењем уграђене ”*downsampling*” функционалности. Коришћењем ове две функционалности, OpenTSDB обезбеђује једноставан интерфејс за извршење комплексних и ефикасних агрегација над сировим подацима. Хиљаде тачака се агрегирају на нивоу секунде. OpenTSDB је посебно згодан у ситуацијама када је потребна агрегација за један уређај (CMTS или CPE у HFC мрежи). Постоји доста начина на које је OpenTSDB у стању да агрегира временске серије. Детаљи о самим функционалностима овог алата се могу пронаћи у [38].

Пост агрегације се користе за пружање дубљег увида у велике скупове података. Пост агрегације су по својој природи споре због огромне количине података која се обрађује, па самим тим и нису временски осетљиве. Apache Spark се користи за пост агрегације у ПНПМБД. У пракси, постоји велики број разних примера где се ове агрегације могу користити. У HFC мрежама, ове агрегације се користе за пружање бољег увида у стање мреже из неколико различитих нивоа. Нивои посматрања зависе од конкретних примера. Један пример нивоа хијерархије за посматрање агрегираних података је: *upstream*, MAC

домен, CMTS, град, област, држава и компанија, од најниже агрегације ка вишој, респективно. Нивои до CMTS-а се користе за посматрање статистика одређеног уређаја, док се нивои изнад користе за анализе пословања компаније. Највише пост агрегација се извршава над подацима прикупљених са CPE јер ти подаци дају увид у стање мреже из угла корисника тј. претплатника.

Један пример агрегације је доступност мреже на основу CPE података. Оваква агрегација даје мрежним операторима увид у стање мреже на различитим нивоима из претплатничког угла, а самим тим пружа увид у корисничко искуство, тј. QoE. Ова метрика се креира на основу времена доступности CPE опреме где се најпре за сваки CPE уређај рачуна доступност на дневном нивоу. Још један пример агрегације је надгледање перформанси уређаја који нису у стању да пруже статистике о свом статусу (неинтелигентни уређаји). Прикупљени подаци са CPE се могу комбиновати са мрежном топологијом чија агрегација може дати увид у стање таквих уређаја (ON, AMP, AP). Овај пример ће бити детаљно обрађен у наставку.

Агрегације података се у ПНПМБД могу користити не само за надгледање перформанси мрежа, већ и за многе друге ствари. На пример, сирови подаци са CPE се могу комбиновати са подацима о уговору претплатника. Први потенцијални пример употребе је детектовање клијената са премијум налогом и лошим квалитетом сигнала што би побољшало корисничко искуство, а самим тим и превентивно спречило најважније кориснике да напусте оператора. Други пример употребе је идентификација великих потрошача интернета са малим претплатничким пакетима у циљу понуде већих пакета. Ових пар једноставних примера показују обећавајући потенцијал ПНПМБД решења за потребе унапређења пословања које не спадају директно у надгледање перформанси мреже.

Уколико су резултати агрегације форматирани за OpenTSDB, исти се могу слати директно коришћењем импорта из фајла. Овај метод се ретко користи у дата колекторима обзиром да се хардвер између њих готово никада не дели. С друге стране, импорт фајлова је изузетно користан када је потребно унети резултате Spark агрегација. Ово је могуће јер су у предложеној архитектури ПНПМБД-а и Spark и OpenTSDB инсталирани на истим физичким серверима. Подаци се уписују директно што додатно растерећује веб сокете OpenTSDB-а, па самим тим додатно се повећава доступност за дата колекторе.

i) Процена стања неинтелигентних мрежних елемената

Посебан скуп изазова приликом имплементације ПНПМБД-а тиче се комбинације прикупљених података са информацијом о мрежној топологији. Надгледање перформанси мреже се може обавити над уређајима који су способни да изврше мерење и пошаљу резултате дата колекторима, Такви уређаји у HFC мрежама су CMTS и CPE опрема. Међутим, постоји и друга група (неинтелигентних уређаја) као што су ON, AMP, AP који су значајни за надгледање перформанси, али који нису у стању да изврше претходно поменуте акције интелигентних уређаја. Ови елементи су подједнако битни као и интелигентни уређаји и такође утичу на квалитет сигнала. У наставку је описан механизам за процену стања неинтелигентних мрежних елемената што представља један од важнијих доприноса ове дисертације.

Предуслов за предложени механизам је постојање информације о топологији мреже, тј. мапирање између мрежних елемената и корисничких уређаја. Оцењивање стања неинтелигентних мрежних елемената се врши на основу већ прикупљених података са CPE опреме која се у топологији налази испод посматраног уређаја. Ти подаци се обрађују и повезују са топологијом мреже. Најпре се на основу метрика дефинише јединствена оцена за

један CPE уређај, а након тога се добијене оцене комбинују с топологијом и процењује се стање неинтелигентног уређаја. Како би се добили усредњени резултати, али и приказао комплекснији метод за агрегацију, предложено оцењивање ће бити приказано над подацима прикупљеним у последња 24 сата. У складу са потребама, предложени механизам се може прилагодити за различите периоде, од нивоа једног очитавања до вишедневних агрегација. Обзиром да се ради о великој количини података, агрегација података се изводи коришћењем Apache Spark радног оквира. Због своје комплексности, поступак агрегације је подељен у неколико корака.

Пре него што се крене у сам поступак агрегације, неопходно је дефинисати које метрике ће бити коришћене за дефинисање оцене стања једног CPE уређаја. У конкретном примеру агрегације користиће се метрике које дефинишу снагу сигнала на пријему и предаји, тј.

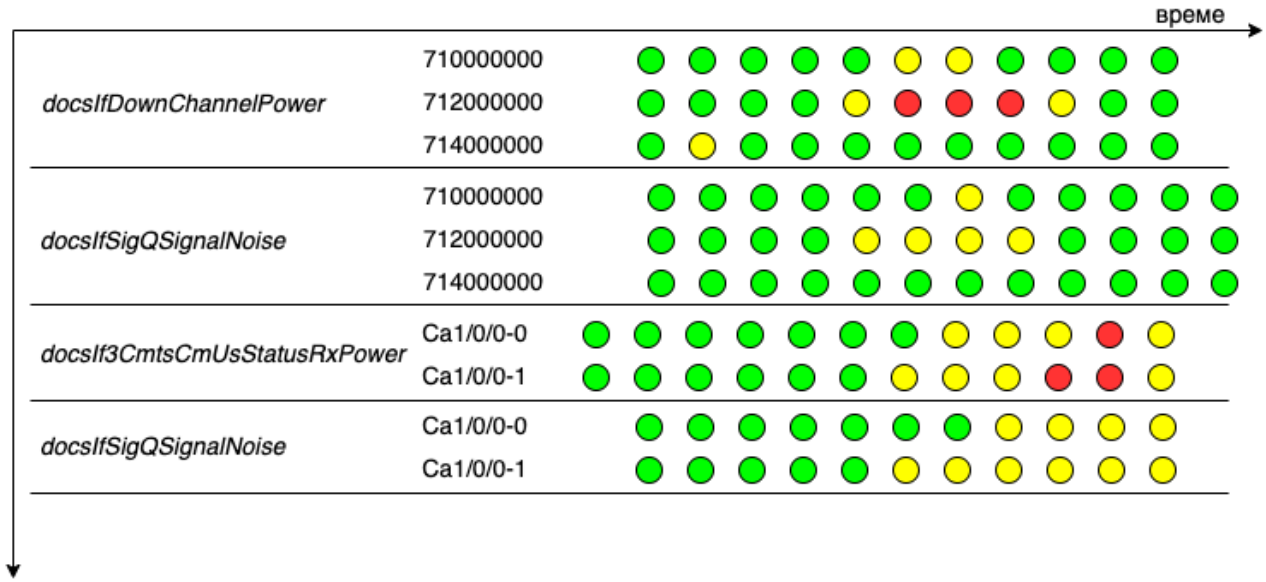
- *docsIfDownChannelPower* - пријемна снага на CPE опреми,
- *docsIfSigQSignalNoise* - однос сигнал/шум за *downstream* канал на CPE опреми,
- *docsIf3CmtsCmUsStatusRxPower* - примљена снага на *upstream* каналу на CMTS-у послата од CPE опреме (тренутно подржано само за DOCSIS3 модеме),
- *docsIfSigQSignalNoise* - просечан однос сигнал/шум на *upstream* каналу на CMTS-у послата од CPE опреме.

Напомена: пошто друга и четврта метрика имају исти назив, а прикупљају се са различитих уређаја (друга са CPE уређаја, а четврта са CMTS уређаја), у наставку текста, испред метрика наведених у листи ће бити стављени префикси *cpe* и *cpe_cmts* да се означи тиме метрика прикупљена са CPE и CMTS уређаја, респективно. Користиће се *cpe_cmts*, а не само *cmts* као префикс да би се нагласило да је у питању метрика прикупљена са CMTS уређаја али која се односи на одређени CPE уређај.

У првом кораку потребно је извршити појединачно оцењивање свих одбирака одабраних метрика. Оцењивање метрика је неопходно како би се све метрике довеле на јединствену скалу квалитета и на тај начин поједноставио процес њиховог даљег комбиновања. Сваком одбирку потребно је доделити оцену 1, 2 или 3 у зависности од квалитета перформансне метрике. Табела 5.2.4.1.1 приказује предлог прагова за оцењивање метрика коришћених у овом алгоритму. Прагови се могу даље мењати у зависности са конфигурацијом одређеног мрежног оператора као и коришћеног стандарда. Слика 5.2.4.1.1 приказује пример првог корака оцењивања метрика у складу са табелом 5.2.4.1.1. Додељене оцене 1, 2, 3 су приказане црвеном, жутом и зеленом бојом, респективно.

Табела 5.2.4.1.1. Оцењивање појединачних метрика.

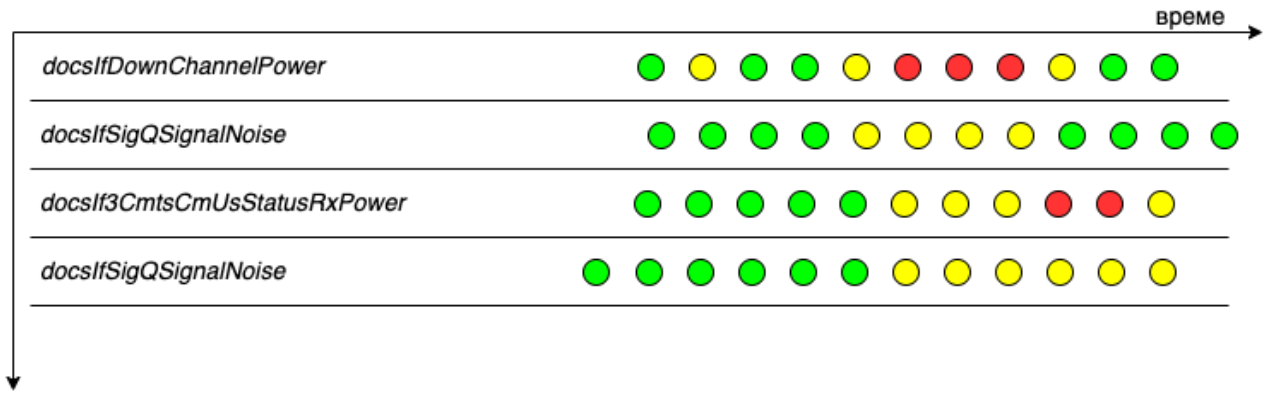
Метрика	Праг за оцену 1	Праг за оцену 2	Праг за оцену 3
cpe.docsIfDownChannelPower	$X \leq -10$ или $X \geq 17$	остало	$-6 \leq X \leq 10$
cpe.docsIfSigQSignalNoise	$X \leq 28$	остало	$X \geq 33$
cpe_cmts.docsIf3CmtsCmUsStatusRxPower	$X \leq -21$ или $X \geq 21$	остало	$-11 \leq X \leq 11$
cpe_cmts.docsIfSigQSignalNoise	$X \leq 21$	остало	$X \geq 28$



Слика 5.2.4.1.1. Први корак оцењивања метрика.

Обзиром на то да се један CPE може повезати на више *upstream*, тј. *downstream*, канала, неопходно је за сваку метрику одредити јединствену оцену у датом тренутку. На овај начин се омогућава комбиновање метрика са различитим бројем интерфејса. На представљеном примеру, посматрани CPE уређај је повезан на 3 *downstream* и 2 *upstream* канала. Бирање јединствене оцене у датом тренутку се врши методом одабира најмање оцене (5.2.4.1.1). На слици 5.2.4.1.2 су представљене метрике након селекције јединствене метрике.

$$KPI\_Mark_i = \min\left(KPI\_interface_i\right) \tag{5.2.4.1.1}$$



Слика 5.2.4.1.2. Метрике након селекције јединственог одбирка.

Након што су метрике оцењене, потребно је исте међусобно комбиновати како би се добила јединствена оцена. Комбинација метрика се најпре врши у једном временском тренутку, а након тога се врши комбинација добијених међурезултата. Пре него што се крене у комбинацију метрика, неопходно је извршити њихово поравнање по времену. Метрике јесу прикупљане у истим периодима времена, али се њихово време прикупљања може

разликовати на нивоу секунда или минута. Ова неједнакост времена се може видети и на сликама 5.2.4.1.1 и 5.2.4.1.2. Поред тога, неопходно је одбацити одбирке за које не постоје све метрике, што ће у датом примеру бити крајњи одбирци *cpe.docsIfSigQSignalNoise* и почетни одбирци *cpe_cmts.docsIfSigQSignalNoise*. Поступак поравнања времена се врши на нивоу периоде прикупљања података. У наставку је дат пример кода (писан у python 3.8) који врши заокруживање времена на основу периоде.

```
import datetime

def round_time_minutes(epoch_time: int, round_minute: int) -> int:
    if round_minute > 60:
        raise ValueError('Round minute cannot be greater than 60')
    round_minute = round_minute if round_minute else 1
    t = datetime.datetime.utcfromtimestamp(epoch_time) \
        .replace(tzinfo=datetime.timezone.utc)
    rounded_time = t.replace(second=0, microsecond=0, minute=0) \
        + datetime.timedelta(minutes=round_minute*(t.minute//round_minute))
    return int(rounded_time.timestamp())
```

Заокруживање се врши на почетни тренутак периоде. За конкретан пример 1650437638 (2022-04-20 6:53:58AM) и за период од 60 минута (што је предложено у табели 5.2.1), заокружено време ће бити 1650434400 (2022-04-20 6:00:00AM). За исто то време, али за период заокруживања од 15 минута, заокружено време ће бити 1650437100 (2022-04-20 6:45:00AM). На слици 5.2.4.1.3 се могу видети одбирци након временског поравнања.

Израчунавање оцене једног CPE уређаја у једном тренутку i врши се комбиновањем метрика по формули (5.2.4.1.2). Свакој метрици j се придружује одговарајући тежински фактор W_j. Коришћењем тежинског фактора могуће је контролисати утицај једне метрике на коначну процену стања уређаја. На слици 5.2.4.1.3 представљене су нумеричке вредности оцена за дати пример узевши вредност тежинског фактора за сваку метрику W_j=1.

$$Mark_i = \frac{\sum_j W_j \cdot metr_j}{\sum_j W_j} \qquad (5.2.4.1.2)$$

Здружене оцене у једном тренутку се поново процењују као добре, средње и лоше, тј. заокружују на оцене 3, 2, 1, респективно. Табела 5.2.4.1.2 приказује коришћене прагове приликом заокруживања.

Табела 5.2.4.1.2. Процена стања на основу здружене оцене на нивоу једног дана.

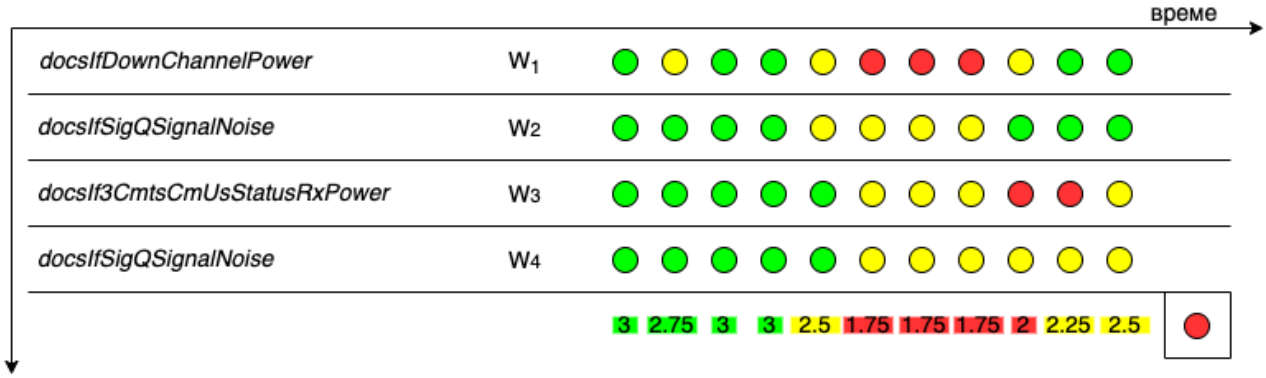
Стање	Праг за оцењивање
1	$X<2$
2	остало
3	$X\geq 2.75$

Последњи корак у одређивању стања једног CPE уређаја представља комбинацију претходно заокружених здружених оцена. Коначно стање се одређује према табели 5.2.4.1.3.

Табела 5.2.4.1.3. Одређивање коначног стања CPE уређаја.

Стање	Праг за оцењивање
1	Више од 20% одбирака има оцену 1.
2	Остало
3	Више од 85% одбирака има оцену 3.

На слици 5.2.4.1.3 приказани су поступци добијања здружене оцене у једном тренутку, заокруживање и одређивање коначног стања једног CPE уређаја. Приликом прорачуна коришћени су једнаки тежински фактори за сваку метрику W_j=1. Уређај из датог примера добија коначну оцену 1 јер је више од 30% времена био ”црвен”.



Слика 5.2.4.1.3. Коначно оцењивање CPE уређаја.

Након што су сви CPE уређаји оцењени, врши се комбинација њихових оцена како би се проценило стање неинтелигентних уређаја. Предуслов за то је постојање тополошких информација које би једнозначно дефинисале који CPE уређаји се налазе хијерархијски испод ког неинтелигентног елемента. Када је све испуњено, за сваки мрежни елемент се врши пребројавање оцена CPE уређаја и додељивање коначне оцене. Коначна оцена за један неинтелигентни мрежни елемент се додељује према правилима дефинисаним у табели 5.2.4.1.4.

Табела 5.2.4.1.4. Одређивање стања неинтелигентних мрежних елемената.

Стање	Праг за оцењивање
1	Више од 50% CPE уређаја има оцену 1.
2	Остало
3	Више од 70% CPE уређаја има оцену 3.

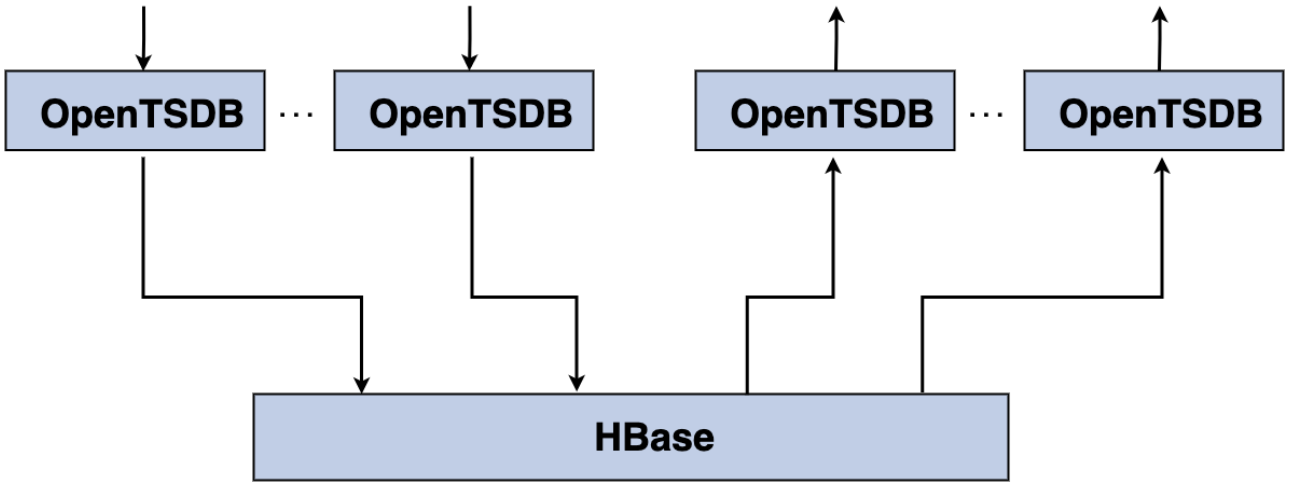
Представљени механизам креиран је на основу познавања понашања неинтелигентних мрежних елемената. Предложени прагови у овој дисертацији се могу додатно фино подешавати у складу са реалним потребама оператора и понашања његове мреже. Напомена је да су прагови предложени у овој дисертацији добијени емпиријски на основу тестова вршених у једној реалној HFC мрежи.

5.2.5. Искуство приликом имплементације у реалној мрежи

Током развоја ПНПМБД-а, главни циљ је био обезбедити високо доступну, скалабилну и ефикасну платформу која је у стању да складишти огромне количине података, задовољи захтеве по питању перформанси уписа и читања података, али и по питању могућности агрегације података. ПНПМБД архитектура предложена у овом поглављу задовољава све поменуте захтеве. Међутим, током имплементације, односно, инсталације ПНПМБД у реалној HFC мрежи појавило се неколико непланираних и непредвиђених изазова. У наставку ће бити представљени најважнији проблеми који су се јавили, као и конкретни начини на који су исти превазиђени.

Један скуп изазова током реализације ПНПМБД платформе тицао се конфигурације и имплементације OpenTSDB-а. Први проблем који се појавио је било загушење OpenTSDB-а услед огромне количине долазног саобраћаја који је требао бити смештен у HBase. Као додатна последица овог проблема, упити за читање података су били заглављени у реду за извршење. Први приступ у решавању овог проблема био је повећавање радне меморије OpenTSDB-у подешавањем одговарајућих *Xms* и *Xmx* Java параметара. Овај приступ је поправио перформансе, али није сасвим отклонио проблем.

Други приступ, приказан на слици 5.2.5.1, представља додавање више OpenTSDB инстанци. Овај приступ је комплетно решио проблем са загушењем. У овом приступу, све инстанце OpenTSDB-а деле заједничке HBase табеле. HBase се брине о приступу и самим тим одржава конзистентност података и табела. Сваки дата колектор шаље податке на насумично одабрану OpenTSDB инстанцу. На овај начин, извршена је једнака дистрибуција саобраћаја на све OpenTSDB инстанце за упис. Поред тога, како би се комплетно раздвојило читање и упис података, додат је посебан скуп OpenTSDB инстанци за читање чиме се постигло значајно унапређење у перформансама читања података. Како би се минимизовало кашњење у преносу, OpenTSDB-еви су физички инсталирани на HBase серверима, тј. на *namenode*-овима и *datanode*-овима. Овај приступ је потпуно решио проблем са загушењем OpenTSDB-а услед превелике количине саобраћаја.



Слика 5.2.5.1. Вишеструки OpenTSDB слој за упис и читање података.

Други OpenTSDB проблем тицао се UID-а (*Unique Identifier*) метрика описаном у секцији 5.2.2. OpenTSDB креира UID како би сачувао нову метрику. Број бајта који се користи за ову сврху дефинисан је параметром ”*UID size parameter*”. Уобичајена вредност овог параметра је 3 бајта. Коришћењем шеме података предложене у 5.2.2, скуп могућих

параметара се може брзо потрошити. Један приступ за решавање овог проблема је повећање параметра *UID size parameter* на 4 бајта. На тај начин, могући број комбинација за генерисање UID-а је 2^{32}-1= 4294967295. Стога, како би се проблем подржаног броја метрика превентивно спречио, неопходно је проценити број потребних метрика пре уписа података и подесити поменути параметар у складу са тим. Међутим, чак и са овим предложеним решењем (повећање вредности параметра *UID size parameter*), постоји вероватноћа да се сви UID бројеви потроше због очекиваних проширења мреже и додавања нових домена интеграције. Ово директно утиче на параметар скалабилности ПНПМБД-а и могућност интеграције са новим доменима. У једном тренутку кад се подеси поменути параметар, исти се више не може мењати. Дакле, уколико се у будућности појави потреба за повећањем овог параметра, неопходно ће бити рекреирати табеле са новом вредности UID величине што ће довести до губитка података. Због тога, предложен је следећи приступ који решава горе поменути проблем.

У другом приступу, који представља један од доприноса ове дисертације, предлаже се коришћење посебних HBase табела по интеграционом домену са одговарајућим скупом OpenTSDB инстанци уместо централизованог коришћења једног скупа HBase табела. Овај приступ би требало применити не само приликом интеграције нових домена, већ и у случају интеграције нових компанија и у било ком другом случају где нема потребе да се подаци чувају заједно. Предложено решење доноси неколико бенефита:

1) генерално, главна табела која садржи податке ће бити додатно подељена што додатно повећава перформансе уписа и читања.
2) за сваки скуп HBase табела се може дефинисати засебна величина UID-а у циљу боље процене долазног саобраћаја.
3) бољи "*multi-tenancy*" приступ. Свака држава и свака подкомпанија која користи платформу ће имати сопствени скуп табела. Уколико је потребно обрисати податке неке земље/компаније из кластера, то се може урадити једноставним брисањем одговарајућег скупа табела. У супротном, поступак брисања би био веома компликован због тога што је једна метрика представљена комбинацијом имена метрике и уређаја.
4) Флексибилност времена чувања података. Време чувања података се врши на нивоу HBase табела подешавањем одговарајућег параметра ("*time to live*" параметар). Коришћењем предложеног приступа, сваки скуп табела може имати засебно дефинисано време за чување.

Следећи изазов током ПНПМБД имплементације се тицао коришћења IP адресе за идентификацију CPE уређаја у HFC мрежи. CPE уређај у HFC мрежи, у зависности од конфигурације HFC мреже, може имати динамичку IP адресу. Приликом повезивања CPE на мрежу, исти добија локалну IP адресу из предефинисаног скупа. Овај приступ није обавезан, али је најчешће коришћен у кабловским мрежама. Први изазов је само надгледање CPE опреме јер иста може временом мењати своју IP адресу. Додатно, као последица, IP адреса не може бити јединствени идентификатор уређаја у овом случају. Уместо тога, MAC адреса се користи као јединствени идентификатор уређаја. Поред тога, IP адреса је неопходна како би се вршила комуникација SNMP протоколом са CPE уређајем приликом прикупљања података. Како би се решио поменути изазов, предложен је *mac-ip-mapping* механизам описан у наставку, а који такође представља један од доприноса ове тезе.

Информација о мапирању између MAC и IP адресе се периодично прикупља са одговарајућег CMTS-а. За прикупљање MAC адресе и IP адресе модема користе се CMTS

OID-и *docsIfCmtsCmStatusMacAddress* и *docsIfCmtsCmStatusIpAddress* [99,100], респективно. Прикупљена информација се користи за креирање конфигурације дата колектора. Ова конфигурација се користи како би се успоставила веза између надгледаних CPE приликом прикупљања података. Механизам *mac-ip-mapping* се покреће неколико пута на дан, најчешће шест, како би се освежило мапирање. Овај приступ има један недостатак. CPE уређај који у међувремену промени IP адресу ће бити недоступан до следећег покретања механизма за освежавање. Срећом, у пракси се показало да се адреса једног CPE веома ретко мења. Оваква промена се може десити приликом поновног покретања уређаја. Чак и тада, у већини случајева, CPE уређај ће добити исту IP адресу. Како би се потпуно елиминисао овај феномен, потребно је ускладити фреквенцију извршавања *mac-ip-mapping* са фреквенцијом прикупљања података са CPE опреме. Ово је оправдано јер број упита који *mac-ip-mapping* механизам генерише је занемарљиво мали у поређењу са укупним бројем упита које изврши дата колектор у току дана. Због тога, може се сматрати да *mac-ip-mapping* механизам не утиче на перформансе CMTS-а.

Следећи скуп изазова током имплементације ПНПМБД се тиче процеса прикупљања података. Због великог броја надгледаних уређаја, постоји више дата колекторских инстанци. Сваки дата колектор прикупља податке са толико уређаја колико може поднети за дефинисану периоду прикупљања (уз остављени мали резервни капацитет у случају неких непредвиђених ситуација). Приликом прикупљања података, дата колектор обрађује редом, један по један уређај док не обради све њему додељене уређаје. У случају када надгледани уређај није доступан или је мрежа успорена, време за прикупљање података је дуже него уобичајено. У ситуацијама групног отказа (на пример, нестанак струје или губитак конекције), читава група уређаја постаје недоступна за прикупљање података. Ово значајно повећава време потребно за пролазак кроз листу уређаја приликом прикупљања и самим тим се може десити да дата колектор не буде у стању да заврши прозивање свих уређаја у једном циклусу. Иако се претходни циклус није завршио, биће покренуто прикупљање у следећем циклусу што може изазвати загушење читавог сервера.

Како би овај изазов био превазиђен, уместо иницијалног серијског, предложено је прикупљање података са више уређаја у паралели. Сваки уређај се процесуира у посебној нити (енгл. *thread*). Уколико је појединии уређај недоступан, то ће утицати само на њему додељену нит. Поред тога, овај приступ вишеструко убрзава процес прикупљања података. Овај приступ захтева сервере са више физичких ресурса, али мањи број дата колектора. Поред решавања проблема недоступних уређаја, овај приступ у глобалу смањује број процесора и RAM меморије. Међутим, треба бити пажљив са дефинисањем броја паралелних процеса. Пракса је показала да треба подесити 30-50 уређаја по једном процесорском језгру за постизање оптималних перформанси. Уколико се одабере превелики број паралелних процеса, може доћи до загушења дата колектора. Треба имати у виду да број паралелних процеса директно зависи од архитектуре и сложености дата колектора. Због овог разлога, како би се одредио оптималан број паралелних процеса, препорука је извршити тестове за конкретно развијени дата колектор.

Подаци прикупљених са CPE су есенцијални приликом решавања проблема у мрежи и надгледања перформанси. Сваки CPE је повезан на CMTS и одређени скуп *upstream*-ова. На жалост, надгледани CPE не зна на који је CMTS и *upstream* повезан. Због тога, било би значајно обогатити податке прикупљене са CPE овом информацијом. Како би се подаци са CPE опреме обогатили подацима са CMTS-а, предложен је *add-upstream-info* механизам који такође представља један од доприноса ове дисертације. Овај механизам врши колекцију података са CMTS-а и прикупља све неопходне податке којим је потребно обогатити податке

са CPE опреме. Приликом прикупљања креира се примарни кључ од комбинације CMTS-а, *upstream* фреквенције и MAC адресе CPE уређаја. Овај примарни кључ се касније користи од стране дата колектора који прикупља податке са CPE како би се приступило подацима за обогаћивање. Резултати *add-upstream-info* механизма, то јест подаци за обогаћивање, се чувају у виду фајла и шаљу на колекторски сервер. Колектор приликом прикупљања података очитава *upstream* фреквенције, комбинује их са MAC адресом како би пронашао податке за обогаћивање. Кад је *add-upstream-info* механизам развијен и имплементиран, појавио се проблем са спорим очитавањем генерисаног фајла за претрагу. У просеку, један дата колектор (4 језгра, 8 GB RAM) прикупља податке са 25000 уређаја. Ово значи да ће генерисани фајл имати до 100000 линија. Интуитивно, овај фајл је велик и спор за читање. Како би се превазишао овај проблем, предложен је следећи метод. Уместо креирања јединственог великог фајла за један дата колектор, креира се више мањих фајлова и то на начин да један фајл садржи информације о једном CPE уређају. Име фајла би била MAC адреса и самим тим би се омогућио једноставан приступ одређеном фајлу за један уређај. Овај приступ је комплетно анулирао проблем великог фајла.

5.3. Безбедност и приватност података

Један од неопходних захтева ПНПМБД платформе је подршка безбедности и приватности података. Иако се ова дисертација не фокусира на аспекте безбедности и приватности, у овом потпоглављу су описана и ова два аспекта да би се истакао њихов значај, али и стекао увид у тренутно стање у свету.

Безбедност података представља заштиту података без обзира на тип садржаја информације. Безбедност је заснована на поверљивости, интегритету и доступности података. Поверљивост података представља заштиту података од недозвољеног приступа, њихово обелодањивање или крађу. Постоји неколико начина на који се подаци у ПНПМБД могу енкриптовати. Први и најнижи начин енкрипције представља енкрипцију диска. Ова енкрипција штити од физичке крађе или губљења диска. Други начин енкрипције представља заштиту на нивоу апликације. Овај заштитни слој спречава насилан приступ, али додаје додатан ниво комплексности апликацији. Последњи начин енкрипције представља коришћење сервиса за енкрипцију података (енгл. *data-at-rest encryption*). Овај сервис омогућава заштиту одабраних фајлова и директоријума и креирање такозваних ”зона енкрипције” [29]. Поред поменутих начина енкрипције, ПНПМБД омогућава дефинисање полиса за приступ корисника. На овај начин се дефинишу дозволе за приступ за сваког корисника или групу корисника и постиже се још један слој безбедности на већ енкриптован HDFS слој. Интегритет података представља прецизност, тачност и веродостојност сачуваних података. За сврху интегритета, ПНПМБД користи уграђене HDFS функционалности за управљање блоковима података уз фактор репликације подешен на вредност три. Висока доступност података обезбеђена је подешавањем високе доступности на свакој од компоненти ПНПМБД-а.

5.3.1. Приватност података

Приватност у технологији података, посебно у биг дата, све више добија на значају последњих неколико година. У наставку ове секције је дат кратак преглед изазова по питању приватности података и предложених решења у литератури, а потом ће у следећој секцији бити представљен један пример значаја аспекта приватности у ПНПМБД.

Преглед изазова по питању очувања приватности у биг дата дат је у [102]. Одавде се може видети тренд у расту интересовања о заштити приватности, посебно у областима као што су анализа социјалних мрежа, система, дата мајнинг и многим другим. Изазови по питању приватности локације за мобилне апликације су дискутовани у [103]. Проблем је представљен не само за провајдере апликација, већ и за екстерне кориснике података који могу индиректно да израчунају локацију. Додатно је представљена забринутост о сервисима који продају податке о локацији екстерним компанијама. Праћење локације људи и трајекторије кретања препознат је као озбиљан проблем у приватности података. Како би се заштитили подаци који садрже GPS (*Global Positioning System*) координате, алгоритам за маскирање података који додаје шум у оригиналну локацију предложен је у [104]. Услед брзог развоја технологије, IoT мреже постају све више популарне, а самим тим све више података је генерисано из ових мрежа. Генерисани подаци могу садржати осетљиве информације по питању приватности. Механизам за заштиту приватности података у индустријским IoT системима заснован на моделу стабла предложен је у [105]. Велики потенцијал у подацима о локацији корисника је препознат у маркетиншкој индустрији. Због своје популарности, и ова грана индустрије се често ослања на биг дата технологије. Метод предложен у [106] предлаже биг дата анализу података о локацији корисника која се истовремено брине и о њиховој приватности. Овај метод користи алгоритам за кластеризацију и локацијску ентропију како би се детектовале најактивније локације.

Компаније препознају приватност података као озбиљан проблем и користе различите методе како би исти превазишли. Модели за заштиту приватности података су предложени како за традиционалне [107], тако и за биг дата системе [108]. Модели за биг дата технологије подржавају све слојеве података, тј. колекцију, складиштење и употребу података. Шеме модела података оптимизованих по питању приватности података представљене су у [108]. У динамичним окружењима, подаци се деле између различитих тимова, како интерних унутар једне компаније, тако и екстерних подизвођача. У таквим окружењима где је велика флуктуација људи и података могуће је доћи до нарушавања приватности података. Узевши у обзир да је у већини земаља приватност података дефинисана и заштићена законом, компаније често одлуче да одступе од коришћења података што за последицу има значајан утицај на њихово пословање и ефикасност. Модел заснован на биг дата технологијама који омогућава дељење и истраживање података задржавајући истовремено приватност предложен је у [109]. Механизми за заштиту приватности података са становишта *k-anonymity*, *l-diversity* и *t-closeness* дискутовани су у [110]. За сваки од механизама дискутоване су предности и недостаци. Поред тога, предложен је механизам за заштиту приватности заснован на комбинацији сва три модела. Коришћење "*differential privacy*" приступа у вишеслојној биг дата архитектури за очување приватности података дискутовано је у [111].

Приватност података, тј. приватност информација, представља могућност контроле крајњег корисника по питању који његови подаци ће се скупљати и на који начин ће се ти подаци користити. Приватни подаци, у зависности од типа апликације, могу бити имејл, локација, историја онлајн претраге, подешавања итд. Узевши у обзир комплексност модерних апликација и система, прикупљање информација (на пример, локације) је неопходно како би се кориснику обезбедило најбоље могуће искуство током коришћења апликације. Међутим, апликације често прикупљају више информација него што им је заиста потребно што доводи до могућности нарушавања приватности корисника. Прикупљени кориснички подаци се могу користити или унутар компаније како би се унапредили интерни процеси и сервиси или продати другим компанијама у виду скупова података. Због

недостатка безбедности унутар компаније могуће су провале и крађа корисничких података [112]. Приватност података се често меша са термином безбедности података. Безбедност података представља заштиту од приступа података недозвољеним лицима, док се приватност података тиче техника прикупљања и регулација које обезбеђују заштиту корисничких информација.

Како би се контрола над подацима приближила крајњим корисницима и истовремено компаније ограничиле по питању прикупљања и употребе података, многе владе широм света усвојиле су законе који регулишу како се подаци могу користити, чувати и штитити. Неке од најважнијих регулација су GDPR (*General Data Protection Regulation*) [113], CCPA (*California Consumer Privacy Act*) [114] и PIPL (*Personal Information Protection Law*) [115]. GDPR је закон за регулацију приватности података у Европској унији. GDPR даје јасне инструкције како подаци треба да се прикупљају, преносе, складиште и штите. Поред тога, овај закон даје корисницима контролу да управљају својим подацима, тј. "право да буду заборављени" (енгл. "*right to be forgotten*"). Ова законска регулатива приморава све компаније које прикупљају корисничке податке да имплементирају механизме за једноставно брисање свих корисничких података без додатног кашњења [113]. CCPA је закон за регулацију приватности података на територији Сједињених Америчких Држава. Овај закон говори о томе да корисник треба да буде свестан који лични подаци се прикупљају као и давање дозвола компанији за продају његових приватних података. Додатно, CCPA пружа компанијама смернице како се овај закон може имплементирати [114]. PIPL је закон за заштиту приватности података у Народној Републици Кини. Поред регулација које преписују GDPR и CCPA, PIPL даје посебан акценат на локализацију података. Наиме, овај закон прописује да се одређени приватни подаци морају физички налазити у Народној Републици Кини [115]. Поред поменутих регулатива, постоје многи други закони који су фокусирани на одређене индустрије. На пример, HIPAA (*Health Insurance Portability and Accountability Act*) је закон у Сједињеним Америчким Државама који дефинише како је потребно управљати здравственим подацима пацијената [116].

Како би обезбедиле приватност података, компаније користе различите технологије за трансформацију истих. Анонимизација и псеудоанонимизација података представљају једне од најпопуларнијих техника трансформације података. Ове две технике представљају трансформацију корисничког UID-а у скуповима података. Анонимизација представља енкрипцију у једном смеру. Уколико се неки податак маскира анонимизацијом, не постоји начин за његову декрипцију тј. враћање оригиналног садржаја. С друге стране, псеудоанонимизација представља процес маскирања података где се исти може декриптовати. По погледу безбедности, псеудоанонимизација је мање сигурна од анонимизације и треба је користити опрезно. Ова техника се обично користи за маскирање података који нису UID, већ поједини атрибути. Неке од најпопуларнијих техника за псеудоанонимизацију су [117]:

- енкрипција – маскирање података шифровањем,
- мешање (енгл. *shuffling*) – мешање података у оквиру једне колоне како би се раздвојио корисник од својих атрибута,
- потискивање (енгл. *suppression*) – уклањање осетљивих колона из скупа података,
- редукција (енгл. *redaction*) – комплетно уклањање делова из скупа података са осетљивим подацима.

Поред поменутих техника постоје и многе друге за заштиту података. На пример, генерализација података представља промену вредности у једној колони одговарајућим

опсегом вредности. На пример, вредност поља година једног корисника, 34, ће бити модификована у опсег 30-40. Синтетизација података представља начин за генерисање потпуно новог скупа података на основу оригиналног скупа коришћењем техника машинског учења. Новокреирани скуп симулира понашање оригиналног скупа. Иако су поменуте технике корисне са становишта приватности података, оне нарушавају информацију и смањују ентропију. Ово је посебно видљиво приликом агрегације података.

5.3.2. Приватност података у ПНПМБД

Прикупљени подаци се углавном користе за надгледање HFC мреже. Међутим, прикупљени подаци могу садржати информације које се могу користити и за друге сврхе. На пример, на основу оствареног саобраћаја, детекције вршних вредности саобраћаја у одређеном временском периоду, могуће је проценити тип локације CPE опреме (на пример, радни простор, кафе или стан). Мрежни оператори могу да прикупљају податке о WiFi мерењима са модема за сличне анализе. Када се прикупљају WiFi подаци, могуће је прикупити податке о клијентима повезаним на WiFi (њихову MAC адресу). Анализа оваквих података може дати увид у кретање појединих корисника. На пример, један мобилни телефон је био повезан на један WiFi уређај, након тога на други и тако даље. Иако подаци нису прикупљани у циљу праћења корисника, ове информације неизоставно постоје унутар прикупљених података. У наставку ће бити описан пример нарушавања приватности корисника кроз прикупљене WiFi податке. Овакав, али и многи други примери несавесне употребе података могу угрозити приватност корисника.

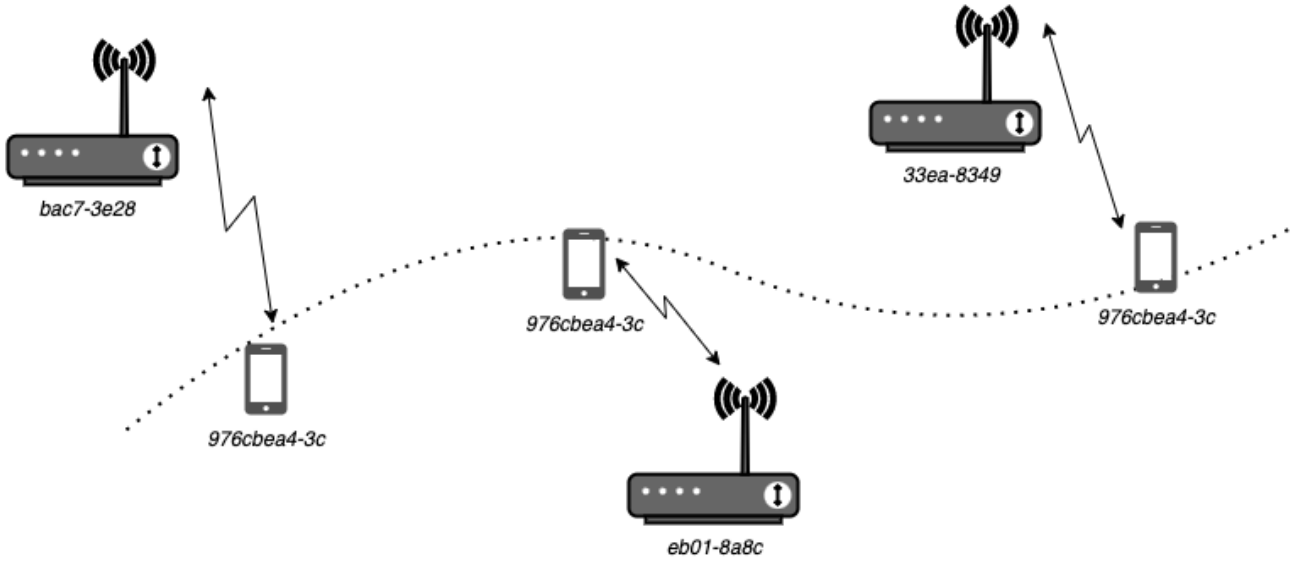
Услед развоја технологија и своје флексибилности, веома велики проценат клијената користи услуге интернет сервиса путем WiFi конекције (мобилни телефони, таблети, лаптоп рачунари). Иако оператори допремају сигнал гарантованог квалитета до приступне тачке, тј. CPE уређаја, може се десити да крајњи корисник има слаб проток. Постоји доста фактора који могу довести до овог феномена од којих су неки дебљина зидова, величина простора, интерференција. Ова појава доводи до незадовољства корисника и повећаног броја жалби. Како би се побољшао квалитет WiFi сигнала, поједини оператори, иако уговором нису обавезни, врше корекције које би требале да поправе WiFi сигнал. Неке од корекција су премештање модема на другу локацију, додавање екстендера сигнала у циљу бољег покривања простора, реконфигурација модема (промена радног канала и његове ширине) како би се избегла интерференција са суседним уређајима итд.

Како би проактивно детектовали локације са лошим квалитетом WiFi сигнала, а самим тим и потврдили успех након акције, оператори прикупљају податке о квалитету WiFi сигнала. Модеми који у себи садрже WiFi рутер у стању су да генеришу такву врсту информације. На пример, SA-RG-MIB садржи све информације везано за податке који се могу прикупљати са Cisco модема [99,100]. Поред основних информација, овде се налазе и MAC адресе повезаних уређаја. Пример једног поједностављеног WiFi скупа података дат је у табели 5.3.2.1.

Из примера се може видети да су колоне ”модем” и ”клијент MAC” шифроване како би се задовољиле регулативе за заштиту приватности. Обзиром да се приликом шифровања користила иста информација (MAC адреса клијента) на сваком модему креиран је исти шифровани стринг. Уколико се исти уређај (нпр. мобилни телефон) повеже на више WiFi мрежа које су у власништву истог оператора, могуће је извршити праћење корисника као што је представљено у табели 5.3.2.1 и на слици 5.3.2.1. Овде се може видети како се уређај ”*976cbea4-3c*” лоцира на различитим локацијама у току дана. Иако на први поглед праћење корисника није била намера оператора, ова информација се налази у подацима.

Табела 5.3.2.1. Пример поједностављеног WiFi скупа података.

модем	Ауто канал	канал [#]	ширина канала [MHz]	клијент MAC	RSSI [dBm]	*upstream* проток [Kb/s]	*downstream* проток [Kb/s]	Време
bac7-3e28	no	10	20	*976cbea4-3c*	-45	1000	2000	2022/05/02 06:12:02
bac7-3e28	no	10	20	*7df94a87-50*	-50	1000	3000	2022/05/02 06:12:02
bac7-3e28	no	10	20	*438893fc-0d*	-38	2000	2000	2022/05/02 06:12:03
...								
eb01-8a8c	yes	2	40	*83bf8e62-64*	-32	65000	130000	2022/05/02 12:45:17
eb01-8a8c	yes	2	40	*976cbea4-3c*	-58	65000	130000	2022/05/02 12:45:18
eb01-8a8c	yes	2	40	*a9eaaf27-e6*	-29	20000	100000	2022/05/02 12:45:18
...								
33ea-8349	yes	5	20	*976cbea4-3c*	-48	1000	24000	2022/05/02 20:13:55
33ea-8349	yes	5	20	*370f5d2d-d9*	-51	1000	24000	2022/05/02 20:13:55



Слика 5.3.2.1. Праћење корисника коришћењем WiFi података са модема.

Уколико се примени исти алгоритам за шифровање података, што је пример у овом случају, за сваки повезани уређај ће се креирати исти шифровани назив. Иако није познато о ком се тачно уређају ради, могуће је праћење његовог кретања. Обзиром на то да је циљ прикупљања података унапређење мреже и корисничког искуства, а не праћење корисника, неопходно је применити другачији механизам за шифровање. Механизам предложен у [118] сугерише да се за шифровање крајњег корисника, уместо његовог јединственог идентификатора (5.3.2.1), користи комбинација његовог идентификатора и модема на који је повезан у датом тренутку (5.3.2.2). На овај начин, шифрована вредност за истог клијента ће бити различита на другим модемима, а самим тим и онемогућено његово праћење задржавајући, притом, све информације неопходне за надгледање квалитета сервиса.

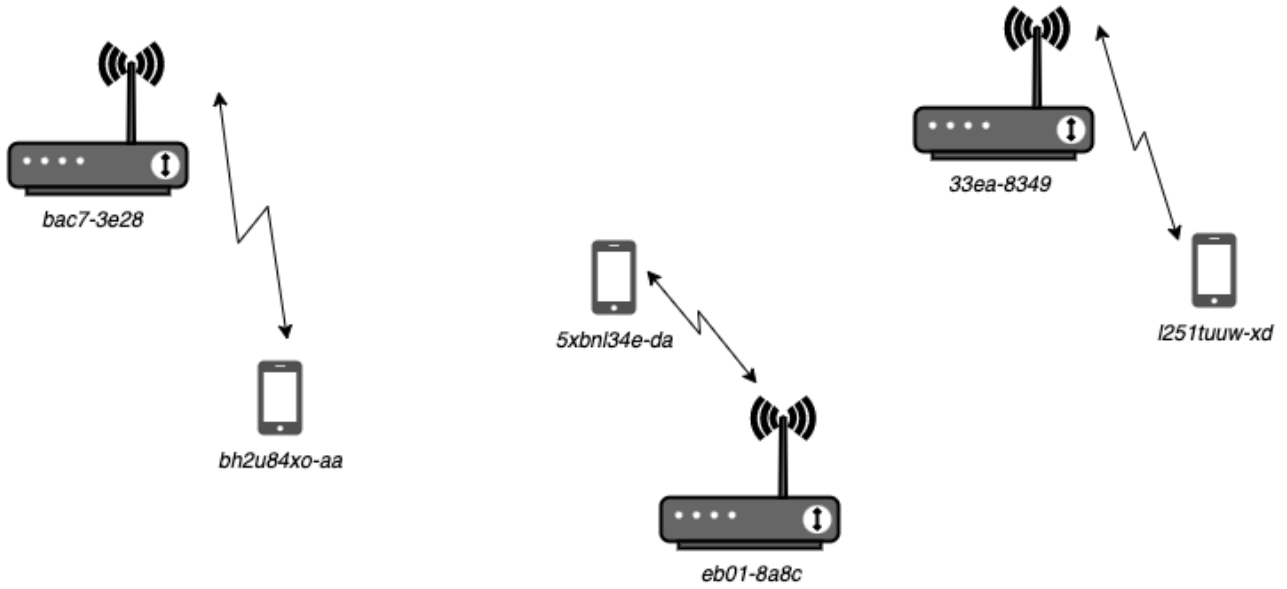
$$шифровани\ MAC = \textbf{\textit{hash\_function}}(klijent\ MAC) \tag{5.3.2.1}$$

$$шифровани\ MAC = \textbf{\textit{hash\_function}}(modem + klijent\ MAC) \tag{5.3.2.2}$$

Пример скупа података и одговарајућа мапа кретања након шифровања механизмом (5.3.2.2) приказани су у табели 5.3.2.2 и на слици 5.3.2.2, респективно. Уређај који је одговарао идентификатору ”*976cbea4-3c*” у табели 5.3.2.1, сада има другачију вредност на сваком модему. Уколико се исти корисник поново повеже на исти модем, биће генерисан исти идентификатор, а самим тим и задржана информација о квалитету за поједини уређај. Ово има предности у ситуацијама када је узрочник лошег квалитета сигнала кориснички уређај (нпр. мобилни телефон, лаптоп...), а не модем. У случају да је потребно елиминисати и овај податак, у формулу за шифровање 5.3.2.2 могуће је додати идентификатор сесије или време очитавања податка.

Табела 5.3.2.2. Пример симплификованог скупа података са унапређеним механизмом шифровања.

модем	Ауто канал	канал [#]	ширина канала [MHz]	клијент MAC	RSSI [dBm]	*upstream* проток [Kb/s]	*downstream* проток [Kb/s]	Време
bac7-3e28	no	10	20	*bh2u84xo-aa*	-45	1000	2000	2022/05/02 06:12:02
bac7-3e28	no	10	20	*j832z60s-k7*	-50	1000	3000	2022/05/02 06:12:02
bac7-3e28	no	10	20	*u1ahnq51-xi*	-38	2000	2000	2022/05/02 06:12:03
...								
eb01-8a8c	yes	2	40	*lsqlxgdq-my*	-32	65000	130000	2022/05/02 12:45:17
eb01-8a8c	yes	2	40	*5xbnl34e-da*	-58	65000	130000	2022/05/02 12:45:18
eb01-8a8c	yes	2	40	*uwd1bs2i-hf*	-29	20000	100000	2022/05/02 12:45:18
...								
33ea-8349	yes	5	20	*l251tuuw-xd*	-48	1000	24000	2022/05/02 20:13:55
33ea-8349	yes	5	20	*8cbdagls-vt*	-51	1000	24000	2022/05/02 20:13:55



Слика 5.3.2.2. Онемогућено праћење корисника из WiFi података са модема.

Приликом развоја решења за складиштење перформансних метрика и креирање одговарајуће шеме података, треба обратити пажњу на приватност корисника. Приватност корисника је у многим земљама дефинисана законским регулативама. Приликом имплементације, поред законских регулатива, неопходно је применити и доменско знање како би се потпуно заштитила приватност корисника. Претходни пример приказује како се

приватност може нарушити иако је испоштована законска регулатива и како се иста, применом доменског знања, може поправити задржавајући исту ентропију података.

6. ДЕТЕКЦИЈА И ЛОКАЛИЗАЦИЈА ОТКАЗА У HFC МРЕЖАМА

Један заједнички проблем свим пружаоцима телекомуникационих услуга су мрежни откази. Мрежни откази су чести, а разлози за то могу бити разноврсни. Нестанак струје, оптички кабл пресечен услед радова, отказ мрежне опреме су само неки од проблема који се свакодневно дешавају. С тим у вези, мрежни оператори морају да превазиђу два главна изазова: детекцију отказа и локализацију квара. Како би се одговорило на ове изазове, у овој дисертацији је предложен механизам за детекцију и локализацију отказа у кабловским тј. HFC мрежама, ДЛОКМ (*Детекција и Локализација Отказа у Кабловским Мрежама*). Овај механизам је развијен над ПНПМБД и представља конкретну примену исте. ДЛОКМ представља један од главних доприноса ове дисертације, при чему су резултати рада на ДЛОКМ публиковани у [119]. У овом поглављу је описан начин функционисања ДЛОКМ уз детаљно објашњење модула за детекцију и за локализацију квара.

Већина радова из области HFC мрежа је фокусирана на унапређење перформанси или кроз детекцију или кроз корекцију проблема одређених делова, тј. појединих елемената, мреже. Циљ предложеног ДЛОКМ је детекција и локализација проблема, без обзира на тип уређаја. Проблем детекције и локализације проблема у мрежама представља важан аспект било које комуникационе мрежне технологије. Због тога, у наставку је дат преглед релевантних радова за детекцију и локализацију кварова.

Детаљан преглед аспеката надгледања мрежа дат је у [120]. Веома битан део било ког система за надгледање мрежа је ефикасно управљање проблемима које укључује и детекцију и локализацију [120]. Важност познавања мрежне топологије за ефикасну детекцију и локализацију је наглашена у [120]. Детекција и локализација отказа у оптичким мрежама је популарна тема због тога што су обично ове мреже имплементиране у језгру мреже (енгл. *core network*) сервис провајдера телекомуникационих услуга [121-126]. Постоје две врсте отказа који се могу детектовати, "меки" и "тврди". Тврди откази представљају комплетан отказ уређаја или линка (на пример, пресечено оптичко влакно), док меки откази представљају деградацију перформанси (на пример, услед старења уређаја) [121]. Тврди откази утичу на перформансе истог тренутка и лакше их је детектовати. Међутим, меки откази су такође важни јер утичу на свеукупне перформансе једне мреже. У овом конкретном примеру, фокус ДЛОКМ-а је на детекцији и локализацији тврдих отказа. Међутим, треба имати у виду да се применом предложене ПНПМБД платформе, прикупљених података и биг дата технологија, може вршити и детекција меких отказа, као што је објашњено на примеру у одељку *i* секције 5.2.4 где је објашњена процена стања неинтелигентних уређаја у HFC мрежи (праћењем оцене рада уређаја кроз време се може уочити евентуална деградација перформанси). Најчешће се за детекцију меких отказа у оптичким мрежама користе механизми машинског учења и неуралних мрежа [121-123]. Слично, машинско учење се такође може користити и за локализацију отказа [124-126]. Предложена решења заснована на машинском учењу захтевају одређене податке (као густина спектралне снаге [122], узорак BER-а (*Bit Error Rate*) [123], усмерене оптичке путање [124],

средње време између отказа [125] итд.) из оптичке мреже како би могли да врше детекцију и локализацију проблема.

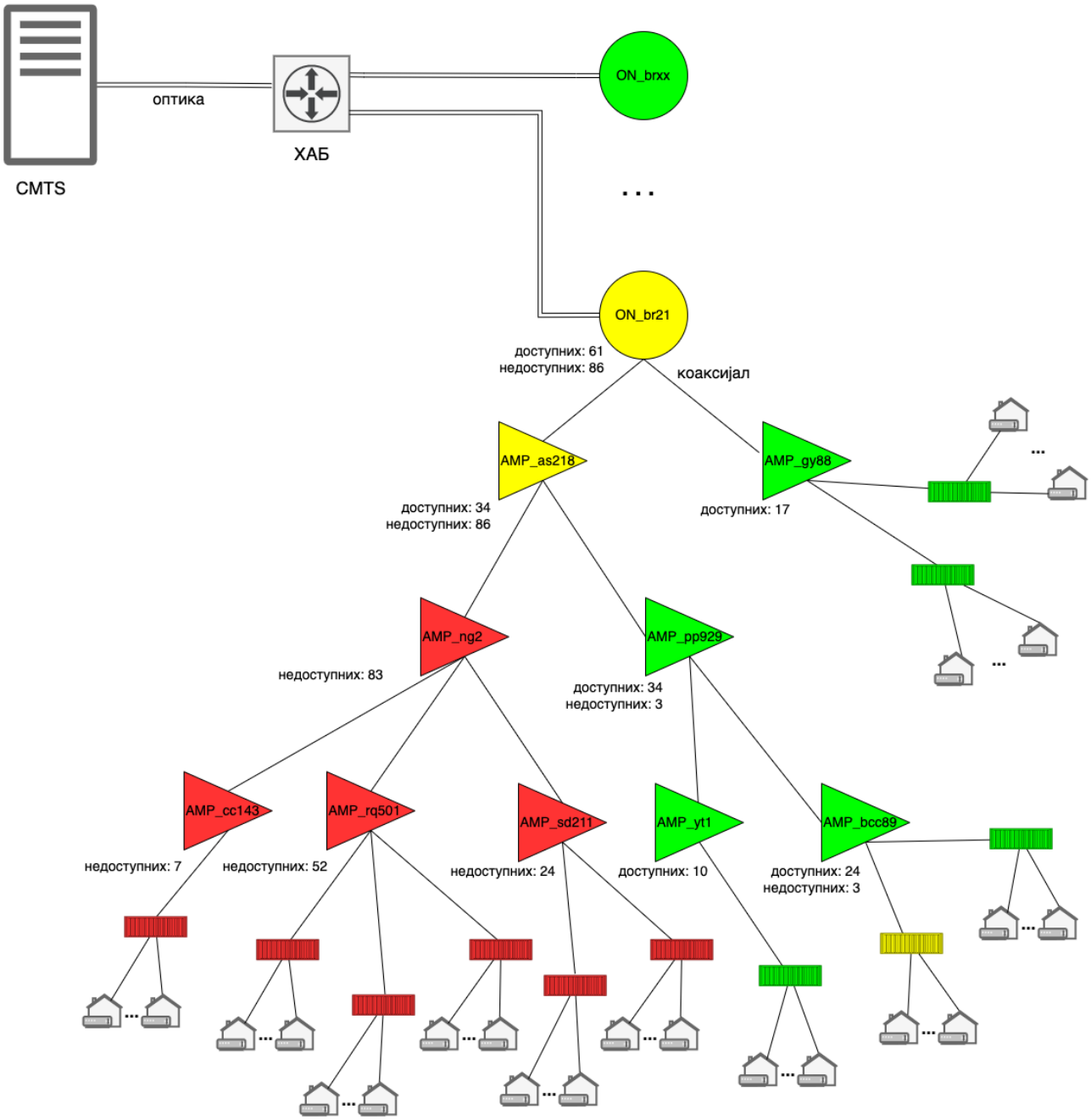
Детекција и локализација отказа су важне за рачунарске мреже као и за дата центре. Развијено је много протокола и механизама за брзу детекцију и промену руте, посебно у транспортним (*core*) деловима мреже као што су MPLS механизми за промену руте [127]. Преглед техника за дијагностиковање отказа мреже представљен је у [128]. Дијагностиковање отказа може бити активно и пасивно. Пасивне технике се ослањају на агенте за детекцију инсталираним на мрежној опреми где исти могу сигнализирати различите типове проблема. Активне технике се ослањају на слање тест (*probe*) сигнала у мрежу како би се детектовао и локализовао проблем. Оптимална селекција тест сигнала је главни аспект ових техника [129]. Поред аларма генерисаних од стране мрежних уређаја, могу се користити и лог фајлови за детекцију и локализацију проблема [120][130]. Лог фајлови се могу користити не само за детекцију отказа, већ и за њихову предикцију. Главни недостатак овог приступа је у томе што су логови неструктурирани подаци и што формат логова може зависити од произвођача, али и од верзије инсталираног софтвера на мрежном уређају [130]. Ово повећава комплексност обраде логова и захтева ажурирање механизма за обраду сваки пут када се појави нови тип уређаја у мрежи или када се ажурира верзија система. Дата центри представљају рачунарске мреже са огромним бројем уређаја. За постизање високе доступности битно је брзо детектовати и локализовати отказ. Сада је детекција још тежа јер између уређаја постоје редундантни линкови и рутирање по вишеструким путањама [131]. Коришћење активних тест сигнала за брзу детекцију и локализацију проблема предложено је у [131]. Постоји неколико радова који се баве детекцијом и локализацијом проблема у приступној мрежи. Решење за детекцију отказа засновано на RADIUS (*Remote Authentication Dial-In User Service*) протоколу предложено је за xDSL (*X Digital Subscriber Line*) приступну мрежу [132]. Решење за локализацију за FTTH (*Fiber to the Home*) мреже предложено је у [133].

Не постоји много радова на тему употребе биг дата технологија за детекцију и локализацију отказа. Углавном, ови радови покривају мобилну мрежу [82,83]. HFC мреже су у литератури углавном адресиране са становишта техничког аспекта. Фокус радова је на унапређењу и оптимизацији уређаја, протока итд. Међутим, нема много радова који се баве проблемом детекције и локализације отказа на нивоу комплетне HFC мреже. Додатно, употреба биг дата технологија у HFC мрежама такође није адекватно покривена у литератури. Постоје радови који покривају употребу биг дата технологија у другим комуникационим мрежама, али HFC мреже имају специфичности које их разликују од класичних мрежних технологија. Додатно, не постоје радови који дискутују примену биг дата технологије у другим комуникационим мрежама који истовремено покривају и детекцију и локализацију отказа.

6.1. Детекција и локализација отказа заснована на биг дата технологији

Један од најважнијих захтева у случају HFC јесте детекција отказа мреже и то у што је могуће краћем временском периоду. Откази се надгледају на нивоу CMTS-а. CMTS је централизовани уређај за један одређен део мреже, високо доступан и са брзим одзивом за пропитивање и, због тога, најприкладнији за детекцију отказа. За ову сврху ДЛОКМ користи *cdxCmtsCmRegistered* метрику [36]. Метрика *cdxCmtsCmRegistered* приказује број онлајн и активних CPE уређаја по MAC домену. Обзиром на то да се број CPE уређаја константно мења, потребно је дефинисати динамичку вредност прага детекције отказа како би се избегле лажне детекције отказа. Због тога је дефинисано следеће правило: отказ је детектован када је

број активних CPE уређаја мањи за 15% од просечне вредности броја активних CPE уређаја генерисане на основу двадесет претходних мерења за одређени CMTS и MAC домен. Ово правило је креирано на основу тестирања у реалним условима. У пракси се показало да овај приступ даје одличне резултате за детекцију отказа док уједно и минимизује број лажних детекција.



Слика 6.1.1. Пример топологије са детекцијом и локализацијом проблема.

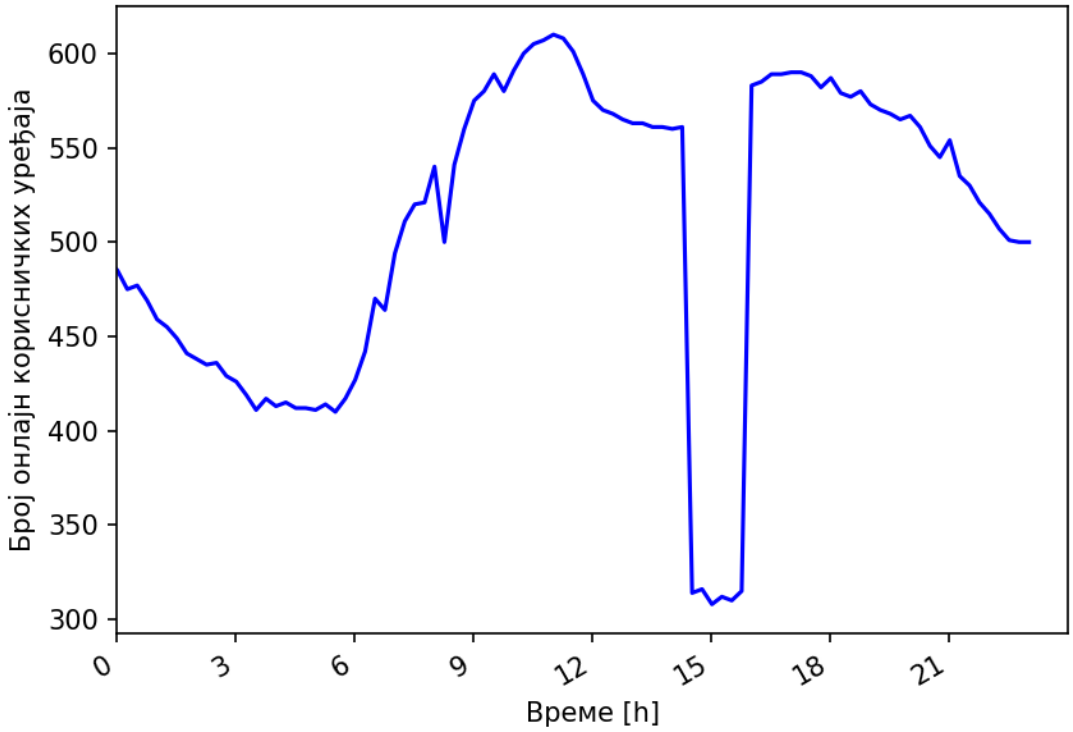
Након што је отказ детектован, покреће се ДЛОКМ за локализацију отказа. Локализација отказа помаже HFC мрежним операторима да ефикасно пронађу и уклоне проблеме у мрежи. Најпре, ДЛОКМ покушава да оствари конекцију са сваким CPE уређајем који припада проблематичном CMTS-у и MAC домену. Информација о томе да ли је веза успостављена или није се корелише са мрежном топологијом. ДЛОКМ механизам се креће

од дна ка врху по мрежној хијерархији, то јест од CPE уређаја ка ON-у и покушава да пронађе главни уређај који је узрочник отказа. Мрежни уређај на највишем хијерархијском нивоу, а испод ког није доступан ниједан CPE уређај је узрочник проблема.

На слици 6.1.1 је приказан пример локализације проблема. Црвеном бојом су приказани мрежни елементи испод којих су сви CPE уређаји недоступни. Зеленом бојом су приказани уређаји где је већина CPE уређаја доступно, а жута где је више од пола уређаја недоступно. Топологија у виду стабла олакшава локализацију узрочника проблема. Узрочник проблема представља уређај који је највиши у хијерархији испод ког су сви уређаји недоступни. У приказаном примеру, појачавач *AMP_ng2* је детектовани узрочник проблема. *AMP_as218* није узрочник проблема јер се испод њега налазе и доступни CPE уређаји. Са становишта проблема, не мора да значи да је конкретно детектовани уређај у квару, већ да се квар налази у његовој околини. Ово усмерава теренске инжењере на тачну област где се проблем налази.

6.1.1. Детекција отказа

У делу за детекцију отказа, ДЛОКМ константно ослушкује број активних CPE уређаја и, у случају наглог пада, покреће део за локализацију квара. Број активних уређаја није сталан и флуктуира кроз време. На слици 6.1.1.1, дат је пример броја активних CPE уређаја на једном CMTS-у и његовом MAC домену за један дан. У конкретном примеру се може видети пад активних CPE уређаја, између 15 и 17 часова, изазван отказом у мрежи. У овој ситуацији механизам треба да детектује тренутак кад је настао проблем, као и тренутак кад је проблем разрешен.



Слика 6.1.1.1. Број онлајн корисничких уређаја.

Очигледно, број активних CPE уређаја расте и пада у зависности од доба дана. Иако је на први поглед интуитивно са слике 6.1.1.1 кад се десио отказ, неопходно је дефинисати

математички модел који би могао да опише овакву ситуацију. Како би се детектовао отказ, предложен је део за детекцију проблема који функционише на следећи начин. Најпре, израчуна се број активних корисника N_{onl_avg} у последњих K итерација очитавања сигнала:

$$N_{onl\_avg} = \frac{1}{K}\sum_{i=1}^{K} N_{onl}[T_{curr} - i] \tag{6.1.1.1}$$

где је T_{curr} тренутни период очитавања, а $N_{onl}[j]$ је број активних корисника у интервалу j. K се може подесити да буде произвољно, у зависности од понашања одређене мреже. За реалну HFC мрежу где се подаци прикупљају на један минут, одређено је да се за K=20 добијају најпрецизнији резултати коначног модела за детекцију. Обзиром на то да се свака мрежа разликује, као и понашање корисника у различитим земљама, препорука је тестирати ову вредност приликом имплементације ДЛОКМ како би се одредио оптималан број за мрежу у којој се овај механизам имплементира.

Отказ у мрежи дефинисан је испуњењем следећег услова:

$$Failure = \frac{N_{onl}[T_{curr}]}{N_{onl\_avg}} \le 1 - Threshold\_d,\ \ 0 < Threshold\_d < 1 \tag{6.1.1.2}$$

где је *Threshold_d* као што име сугерише, праг за детекцију отказа. Тестирањем на реалној HFC мрежи показало се да *Threshold_d*=0.15 даје оптималне резултате. Треба имати у виду да су формуле (6.1.1.1) и (6.1.1.2), као и предложене вредности параметара K и *Threshold_d* резултати спровођења експеримената над реалним подацима HFC мреже и са реалним отказима. За тестирану мрежу, предложени алгоритам детектује преко 95% мрежних отказа. Овај приступ се такође може користити и у другим HFC мрежама. Међутим, зависно од многих параметара као што су, на пример, држава, социјално окружење где је HFC мрежа имплементирана, понашање клијената (култура нације, навике, радно време итд.) може бити другачије. Због тога, очекивано понашање и облик криве временских серија метрике *cdxCmtsCmRegistered* на дневном нивоу се може разликовати. Због тога, предлог је приликом имплементације урадити додатно фино подешавање параметара како би се постигли најбољи могући резултати. У циљу постизања најбољих могућих резултата алгоритма, неопходно је одредити оптимални *Threshold_d* пре саме имплементације ДЛОКМ за HFC мрежу. Предложени параметри су подешени за једног кабловског оператора у Европи.

Сличан метод прагова се може искористити за детекцију ситуација када је проблем решен. Ово је посебно корисна функционалност у ситуацијама када се детектовани проблем реши сам од себе. Нестанак струје на неколико минута и поновни долазак је само један од мноштва примера када се овакве ситуације могу догодити. У случају интеграције ДЛОКМ-а са постојећим системом за управљање алармима (енгл. *fault management system*), два догађаја као што су проблем и разрешење проблема се могу комбиновати и претходно креирани проблем се може аутоматски уклонити. Услов за детекцију резолуције проблема је следећи:

$$Resolution = \frac{N_{onl}[T_{curr}]}{N_{onl\_avg}} \ge 1 + Threshold\_r,\ \ 0 < Threshold\_r < 1 \tag{6.1.1.3}$$

где је *Threshold_r* праг за детекцију резолуције проблема. Једино ограничење овог алгоритма јесте да се проблем мора десити пре него што се примени детекција резолуције (6.1.1.3). Другим речима, проблем најпре мора бити детектован, а тек након тога се може активирати механизам за детекцију разрешења. Треба имати у виду да имплементирани ДЛОКМ користи и део за детекцију резолуције проблема.

У табели 6.1.1.1 приказан је пример детекције и резолуције једног проблема који је трајао неколико минута. У табели су приказана само релевантна мерења $N_{onl}[T_{curr}]$ и N_{onl_avg} због поверљивости скупа података. Свакако, пример у табели је довољан да илуструје предложени алгоритам. Црвени ред показује тренутак када је нарушен *Threshold_d* зато што је у овом тренутку вредност *Failure* параметра пала испод 0.85. Уколико би отказ трајао више од *K* итерација, онда би, према (6.1.1.2) у неком тренутку *Threshold_d* престао да пријављује детекцију проблема зато што би се вредност просечног броја CPE, N_{onl_avg} , смањила према смањеном броју CPE. Међутим, овакво понашање није проблем зато што детекција проблема не мора нужно да буде детектована све време током трајања проблема, већ је довољно детектовати први тренутак пада уређаја. У примеру у табели 6.1.1.1, такође је приказан пример детекције разрешења проблема. Зелени ред приказује тренутак када се детектовала резолуција проблема (јер је вредност *Resolution* параметра већа од 1.15). Овај пример показује да се разрешење проблема може детектовати чак и када проблем траје неколико итерација.

Табела 6.1.1.1. Пример детекције и резолуције проблема.

T_{curr}	$N_{onl}[T_{curr}]$	N_{onl_avg}	$N_{onl}[T_{curr}]/N_{onl\_avg}$
1584659988	417	455.55	0.915377017
1584660048	413	451.15	0.915438324
1584660108	415	446.9	0.928619378
1584660168	412	442.6	0.930863082
1584660228	120	424.05	0.282985497
1584660288	121	405.85	0.298139707
1584660348	121	388.15	0.311735154
1584660408	118	370.2	0.318746623
1584660468	429	368.2	1.165127648
1584660528	427	366.6	1.164757229
1584660588	442	365.95	1.207815275

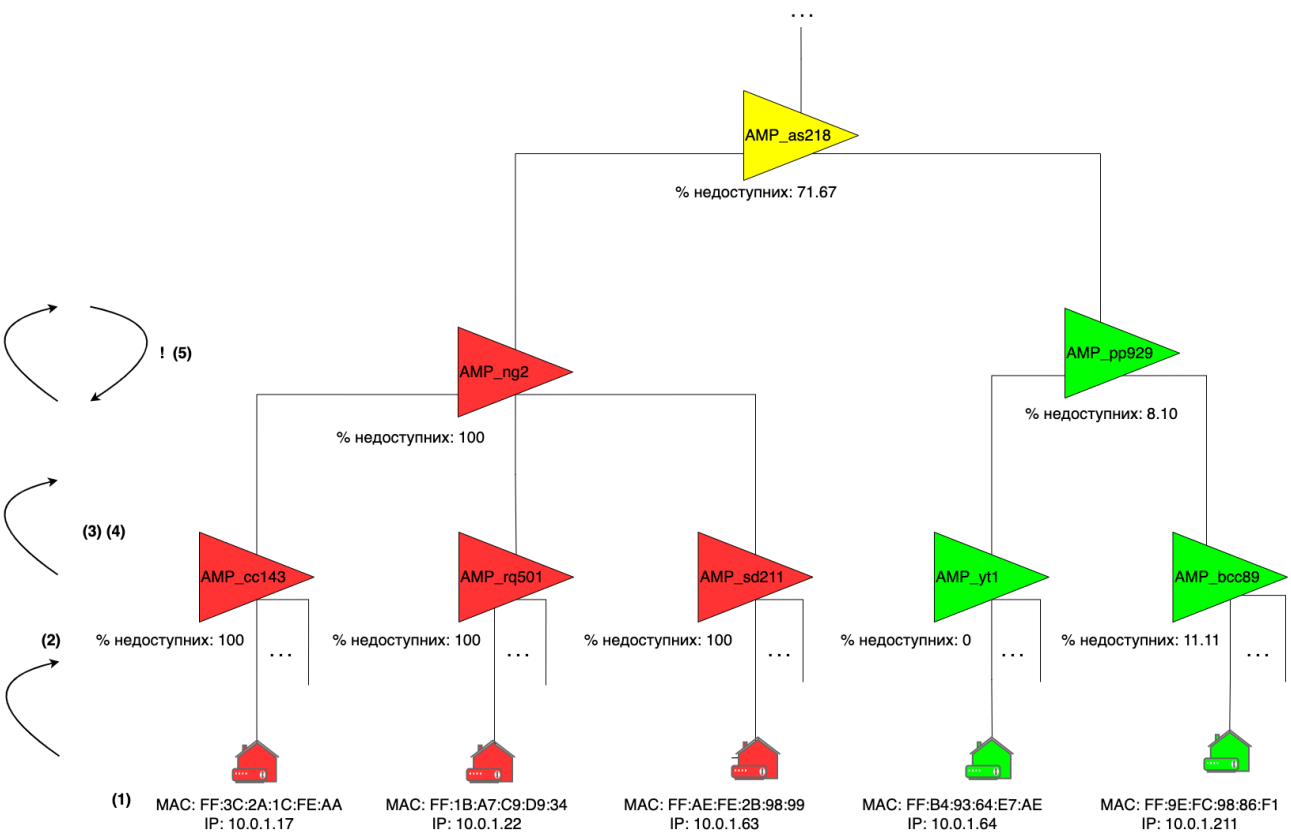
6.1.2. Локализација отказа

Након што се детектује отказ у мрежи за један CMTS и MAC домен, покреће се алгоритам за локализацију отказа. Главни циљ овог дела је детекција мрежног елемента (AP, AMP или ON) који представља узрок проблема. Предуслов за реализацију овог алгоритма је информација о топологији мреже као и информација о томе на који мрежни елемент је повезан сваки CPE уређај (на пример, модем FF:3B:2A:8C:AE:DC повезан је на *AP_cc43*). Информација може бити у било ком облику (на пример, табела у бази или CSV фајл са везом уређаја у релацији родитељ-дете). Поред тога, неопходно је постојање везе MAC домена и CPE уређаја за сваки CMTS као и IP адресе за сваки CPE уређај. Овај део није проблематичан јер се може лако добити пропитивањем CMTS-а за метрике *fordocsIf3CmtsCmRegStatusMdIfIndex*, *docsIfCmtsCmStatusIpAddress* и *docsIfCmtsCmStatusMacAddress* [99,100] и њиховим међусобним комбиновањем.

Алгоритам за локализацију састоји се из следећих корака:

1) Проверити доступност свих CPE уређаја који су повезани на проблематичан MAC домен. Пропитивање доступности се врши *ping* алатом.
2) Здружити резултате из 1. корака са топологијом и израчунати проценат недоступних CPE уређаја за сваки мрежни елемент.
3) Кренути од дна топологије и проверавати вредност израчунатог процента доступности. Уколико је за одређени мрежни елемент 100% CPE уређаја недоступно, отићи хијерархијски један ниво изнад и поновити поступак.
4) Понављати 3. корак док проценат недоступних CPE уређаја не опадне. То значи да је алгоритам дошао до узрочника проблема.
5) Вратити се један корак назад и проверити ко од деце детектованог мрежног елемента има недоступност од 100% и вратити резултат. Изаћи из алгоритма.
6) Уколико се дође до врха топологије, а вредност недоступности CPE уређаја је и даље 100%, значи да је узрочни мрежни елемент оптички чвор. Вратити оптички чвор и изаћи из алгоритма.

Због чињенице да је већина CPE уређаја константно укључена, овај алгоритам готово да не даје лажно-позитивне резултате и прецизно детектује мрежни елемент који је узрочник проблема. Треба имати у виду да не мора нужно значити да је мрежни елемент проблем, већ се проблем може десити и у његовој околини (на пример, исечен кабл или слично). Детекција мрежног уређаја у огромној мери поједностављује решавање проблема слањем једне екипе на терен на тачно одређену локацију уместо слања више екипа и ручне претраге проблема.



Слика 6.1.2.1. Пример локализације проблема.

Овај алгоритам креиран је анализом података које је могуће добити са CMTS-а и анализом података о инвентару мрежног оператора. Предложени алгоритам је тестиран и имплементиран у реалној HFC мрежи. У наставку ће бити представљен пример егзекуције алгоритма и локализације проблема над топологијом представљеној на слици 6.1.1. На слици 6.1.2.1 визуелно су приказани кораци алгоритма. Представљени пример је поједностављен и вештачки генерисан, али осликава реално понашање мреже. Најпре, врши се мапирање између MAC домена и CPE уређаја (MAC адресе и IP адресе) који су прикупљени са одређеног CMTS-а. Пример мапирања:

MAC_domain1, 10.0.1.17, FF:3C:2A:1C:FE:AA

MAC_domain1, 10.0.1.22, FF:1B:A7:C9:D9:34

MAC_domain1, 10.0.1.63, FF:AE:FE:2B:98:99

MAC_domain1, 10.0.1.64, FF:B4:93:64:E7:AE

...

MAC_domain1, 10.0.1.211, FF:9E:FC:98:86:F1

MAC_domain2, 10.0.1.214, FF:DA:E3:5A:ED:79

Узевши у обзир да се ова информација споро мења у времену, може се прикупљати периодично, не у тренутку локализације, и на тај начин додатно смањити време извршења алгоритма. Тополошко мапирање приказује релацију између CPE MAC адресе и мрежног елемента на који је CPE повезан у топологији:

FF:3C:2A:1C:FE:AA, AMP_cc143

FF:1B:A7:C9:D9:34, AMP_rq501

FF:AE:FE:2B:98:99, AMP_sd211

FF:B4:93:64:E7:AE, AMP_yt1

...

FF:9E:FC:98:86:F1, AMP_bcc89

У првом кораку алгоритма, врши се провера доступности свих CPE уређаја повезаних на MAC домен. За проверу доступности користи се мрежни алат “ping”. CPE уређаји који одговоре на упит (редови испод са ”*ttl*” и ”*time*”) су онлајн, док се остали сматрају недоступнима. Пример *ping* резултата:

ping 10.0.1.17, Request timeout for icmp_seq 0

ping 10.0.1.22, Request timeout for icmp_seq 0

...

ping 10.0.1.63, Request timeout for icmp_seq 0

...

ping 10.0.1.64, 64 bytes from 10.0.1.64: icmp_seq=0 ttl=64 time=0.053 ms

ping 10.0.1.211, 64 bytes from 10.0.1.214: icmp_seq=0 ttl=252 time=22.557 ms

У другом кораку, алгоритам комбинује резултате из првог корака са мрежном топологијом и рачуна проценат недоступних CPE уређаја. Пример резултата другог корака:

AMP_cc143, 100

AMP_rq501, 100

AMP_sd211, 100

AMP_yt1, 0

AMP_bcc89, 11.11

AMP_ng2, 100

AMP_pp929, 8.10

AMP_as218, 71.67

AMP_gy88, 0

ON_br21, 62.7

У трећем и четвртом кораку, алгоритам се креће по топологији одоздо на горе и покушава да пронађе елемент испод ког су сви CPE уређаји недоступни. У овом примеру, заснованом на слици 6.1.2.1, *AMP_ng2* је узрочник проблема.

6.1.3. Поређење

У овој секцији ће бити извршено поређење предложеног ДЛОКМ са постојећим решењима из ове области. Узевши у обзир да су постојећа решења у тренутној литератури дизајнирана за друге мрежне технологије, а не HFC, директна нумеричка компарација није потпуно могућа. Свака мрежна архитектура, па чак и имплементација истих типова мреже, се могу доста разликовати што генерално отежава директно поређење различитих решења, не само у овом случају. Због тога, одабрана су решења која углавном циљају приступну мрежу [132], [133] као и решење које користи лог фајлове за детекцију отказа [130] обзиром на то да се овај приступ може користити у већини мрежних технологија. Поред тога, ДЛОКМ ће такође бити поређен са PNM (*Proactive Network Maintenance*) представљеним у [134] јер се овај алгоритам такође бави HFC мрежама и детекцијом и локализацијом отказа у њима. Коначно, потребно је нагласити да ДЛОКМ посматра отказе из угла крајњег корисника што је очигледно из објашњења алгоритма представљеног раније у овом поглављу. Ова чињеница је важна јер је директно повезана са корисничком QoE.

Решење за детекцију отказа за xDSL приступно-агрегациону мрежу представљено је у [132]. xDSL приступне мреже су веома сличне HFC мрежама у смислу да обе користе тополошке мреже у виду стабла. Корисничка опрема (CPE) типично остварује PPP (*Point-to-Point Protocol*) сесије са сервером за широкопојасни даљински приступ. За прикупљање информација о корисничким сесијама (на пример, почетак и крај сесије) користи се RADIUS. Логови са RADIUS сервера се онда користе за детекцију отказа. Овај приступ је сличан предложеном ДЛОКМ решењу. Отказ ће бити детектован уколико проценат PPP раскинутих сесија у одређеном временском интервалу пређе унапред дефинисани праг. Праг је дефинисан као проценат активних корисника на DSLAM-у (*Digital-Subscriber-Line-Access-Multiplexer*). Предложена вредност прага је нешто испод 100%. Узевши у обзир чињеницу да је DSLAM последњи корак према корисничкој опреми у приступној мрежи у *downstream* смеру, ова вредност прага је веома висока. Овако висок праг захтева исцрпну претрагу RADIUS логова што представља значајан проблем са обрадом података. Због овог разлога, користи се додатни праг. Овај праг је подешен на 30-40% од најмањег капацитета картице како би се надгледала комплетна мрежа. Чим се пређе овај праг, иде се на проверу

иницијалног прага како би се избегло непотребна обрада и оптерећење сервера. Овај приступ се може користити за локализацију отказа (на пример, за детекцију отказа неког свича), али ово није дискутовано у [132]. У наставку ће решење предложено у [132] бити референцирано као DMPF (*Detection of Mass PPP Failures*) зато што је ово име коришћено од стране аутора рада. Када се пореди са ДЛОКМ, главни недостатак DMPF-а је чињеница да се у предложеном решењу не користе биг дата технологије. Аутор потврђује да је количина RADIUS логова огромна због великог броја остварених сесија у мрежи, посебно уколико се у обзир узме чињеница да се RADIUS логови морају чувати историјски. Велику количину података није лако обрадити, што је и главни недостатак DMPF-а који би могао профитирати од биг дата технологија. Додатно, детаљи везани за алгоритам локализације нису приказани у [132]. Коначно, још један потенцијални недостатак DMPF-а је чињеница да мора да се користи PPP у приступним мрежама. Неки xDSL уређаји користе DHCP (*Dynamic Host Configuration Protocol*), па самим тим у овим ситуацијама DMPF се не може користити.

Лог фајлови се такође користе за детекције отказа у [130]. У овом решењу посматрају се IP мреже, али је решење генерализовано и на друге мреже. Логови се прикупљају са мрежних уређаја уз помоћ NMS-а (*Network Management System*). Обрадом логова, могуће је детектовати, па чак и предвидети, отказе. Главни изазов у овом приступу је структура логова. Лог фајлови могу имати другачије структуре у зависности од типа уређаја, произвођача, па чак и исти уређаји се могу разликовати у верзијама фирмвера која је на њима инсталирана итд. Такође, логови спадају у неструктуриране податке. Све ово води до потешкоћа у обради података као и потребе за константним ажурирањем алгоритма кад год се појави нови тип лог фајла (на пример, додавање уређаја новог произвођача). Такође, у случајевима кад у мрежи постоје уређаји који не прикупљају податке о себи (као што су на пример, AMP-ови у HFC мрежама), не постоји решење како реконструисати стање ових уређаја, чак је упитно и да ли је то могуће. Механизми који користе перформансне метрике прикупљене са уређаја би били боље решење обзиром на то да се ту подаци структурирају, а и многи мрежни елементи подржавају ову функционалност. Коришћење структурираних података би смањило захтеве по питању обраде података. Ово је важна предност ДЛОКМ-а у односу на све приступе засноване на лог фајловима.

Детекција и локализација отказа у FTTH мрежама дискутована је у [133]. У овом раду је разматрана TDM-PON (*Time Division Multiplexing Passive Optical Network*) мрежа. Фокус у [133] је локализација проблема у оптичким мрежама и ONU (*Optical Network Unit*) уређајима. У пракси, техничари користе OTDR (*Optical Time-Domain Reflectometer*) уређаје на местима ONU уређаја како би идентификовали растојање до исеченог оптичког влакна (локализација проблема). Међутим, овај процес који изискује излазак техничара на терен је веома временски захтеван и потражује много ресурса. Ови разлози су охрабрили ауторе за аутоматизацију овог процеса додавањем SRE (*Switchable Reflective Element*) уређаја. SRE уређаји се даљински управљају из контролног центра и могу се конфигурисати у моду за рефлектовање сигнала. На овај начин, провера оптичких влакана може бити аутоматизована и проблем детектован. На крају, представљено је централизовано решење за детекцију отказа, али нема много детаља о томе како предложени механизам заиста врши детекцију отказа. Очигледно, решење у [133] је стриктно фокусирано на FTTH и не може се генерализовати на друге мреже. Поред тога, предложено решење детектује и локализује кварове само на PON (*Passive Optical Network*) делу мреже, а не и на агрегационом делу као што то раде ДЛОКМ и DMPF.

PNM за HFC мреже предложен је у [134]. Идеја PNM механизма је базирана на идентификацији слабих делова мреже који би отказали у скоријој будућности како би исти

могли бити проактивно замењени. PNM механизам користи спектралне податке прикупљене са *Full-Band Capture* алата, детектује потенцијалне проблеме и групише уређаје који су погођени истим проблемом. Поред тога, PNM механизам прави разлику између ситуација када је детектовани проблем на улазу или излазу, тј. у стамбеној јединици или ван ње. PNM користи машинско учење без надзора користећи *k-means* алгоритам. С друге стране, ДЛОКМ детектује и локализује проблеме у тренутку њиховог појављивања. Локализација проблема се изводи директним пропитивањем CPE уређаја и, користећи корелацију уређаја са тополошким информацијама, омогућава ДЛОКМ-у да прецизно израчуна позицију проблема. Модели машинског учења увек имају простора за погрешне предикције што се може сматрати као недостатак PNM механизма. Генерални закључак је да су PNM и ДЛОКМ комплементарни механизми који се могу користити заједно у мрежи – ДЛОКМ за детекцију тврдих отказа у мрежи, а PNM за детекцију меких отказа. У случајевима тврдих отказа, ДЛОКМ би детектовао отказ и покренуо механизам за његову локализацију, док би се PNM користио за даље унапређење већ исправне мреже у циљу њеног даљег усавршавања и превентивног решавања (”меких”) отказа.

Резиме поређења ДЛОКМ са другим решењима приказан је у табели 6.1.3.1. Имати у виду да колона ”адаптивност” подразумева могућност коришћења решења у другим мрежним технологијама. Под колоном ”покривеност” се подразумева покривеност мреже по питању детекција/локализација проблема.

Табела 6.1.3.1. Резиме поређења решења.

	Мрежна топологија	**Типови отказа**	**детекција/ локализација**	**Адаптивност**	**Покривеност**
ДЛОКМ	HFC	тврди	оба	средња/висока	висока
DMPF [132]	xDSL	тврди	оба	средња	висока
Лог базирани [130]	IP	тврди	оба	висока	средња
SRE базирани [133]	TDM-PON	тврди	оба	ниска	средња
PNM [134]	HFC	меки	детекција	ниска	висока

Због тога што ДЛОКМ покрива детекцију и локализацију тврдих отказа, већина решења разматрана у овој секцији су такође фокусирана на тврде отказе и њихову детекцију и локализацију. Међутим, локализација DMPF-а није објашњена у [132]. Додатно, у решењу базираном на SRE [101] није дато много детаља за детекцију отказа. PNM [134] је једино решење намењено меким отказима, али са друге стране PNM је једино решење, поред ДЛОКМ, које је намењено HFC мрежама. ДЛОКМ и решење засновано на логовима су најадаптивнији зато што се они ослањају на SNMP протокол и лог фајлове, што се често користи у мрежним технологијама и њиховим уређајима. Наравно, оба решења би захтевала одређене модификације приликом имплементације, обзиром на то да се метрике и структура лог фајлова разликују између мрежних технологија. У случајевима мрежа које садрже петље, неопходно је урадити додатне адаптације над техником локализације у ДЛОКМ-у. Због овог разлога је у табели 6.1.3.1 стављена оцена средње/високе адаптивности за ДЛОКМ. DMPF је такође адаптиван зато што се ослања на RADIUS. С друге стране, решење базирано на SRE [133] и PNM [134] се више фокусирају на веома одређене мрежне технологије што их чини неадаптивним за друге. ДЛОКМ, DMPF и PNM су у стању да покрију све елементе у једној мрежи, због чега је њихова покривеност оцењена као висока. Решење засновано на SRE има добру покривеност на PON делу мреже, али не разматра агрегациони део мреже. Покривање уређаја који нису у стању да генеришу логове није дискутовано у [130]. Зависно од информације прикупљене из логова других уређаја, могуће би било реконструисати

информацију о стању уређаја који не генеришу логове, али ово у великој мери зависи од информације садржане у самим логовима.

7. ДОДАТНЕ МОГУЋНОСТИ ПНПМБД

ПНПМБД решење је развијено са намером да обезбеди високо доступну, скалабилну платформу за складиштење велике количине података са фокусом на податке временских серија и са конкретном применом прикупљања и обраде телекомуникационих перформансних метрика. Предложени дата колектори и шема података развијени су и оптимизовани за перформансне метрике из HFC мреже, конкретно са CMTS и CPE опреме. Прикупљени подаци се користе у два основна облика. Први облик представља посматрање сирових временских серија у циљу детаљног увида у сваки уређај и део мреже. Други облик се односи на посматрање стања мреже или њеног већег дела на основу прикупљених података. Посматрање ове врсте се добија агрегацијом сирових података по различитим нивоима. Агрегације могу дати ширу слику о квалитету пруженог сервиса. Поред тога, уз помоћ прикупљених података могуће је добити другачију слику о мрежи, као што је, на пример, детекција и локализација отказа у мрежама, описана у претходном поглављу. Поред поменутих примена, ПНПМБД се, заједно са прикупљеним подацима, може искористити и за развој других функционалности у циљу решавања различитих изазова са којима се оператори сређу. У наставку овог поглавља је представљено пар примера додатних могућности ПНПМБД.

7.1. Анализа постојећих података

Услед велике количине података као и информација које се крију унутар истих, често није очигледно видети разне међузависности између података. Скривене информације често могу бити врло корисне и донети додатну вредност компанији. Додатна вредност се може огледати у унапређењу постојеће мрежне инфраструктуре или директној монетизацији кроз пружање нових сервиса клијентима. Анализом података се баве дата аналитичари који имају добро доменско знање, у овом случају познавање кабловских мрежа, као и способност да самостално претражују и комбинују податке. Када се у маси података пронађе корисна информација и докаже њена вредност, иста се уврштава у платформу креирањем скупа агрегација и, уколико је потребно, извештаја. У наставку ће бити представљена два примера која демонстрирају наведено, а која врше анализу прикупљених података из HFC мреже дефинисаним у поглављу 5 у табели 5.2.1.

Један од података који се прикупља са CPE уређаја јесте укупан број примљених и послатих октета (*ifInOctets* и *ifOutOctets*). На основу агрегације ова два податка, може се одредити укупан послати и примљени саобраћај за сваког корисника на месечном нивоу. Добијени резултати се могу повезати са подацима о претплатничким уговорима. Претплатнички уговори, поред основних информација, садрже и податак о максималном дефинисаном протоку у оба смера. На основу ова два података могу се детектовати најзахтевнији претплатници са малим уговореним максималним протоцима. Резултат ове агрегације се може користити у маркетиншке сврхе у циљу понуде пакета већих протока најзахтевнијим корисницима и на тај начин повећати профит.

Сваки CPE уређај поседује MAC адресу. MAC адреса представља јединствени идентификатор уређаја. Приликом регистрације CPE уређаја на мрежу, CMTS проверава да

ли се CPE уређај (његова MAC адреса) налази у листи претплатника. Уколико је CPE уређај регистрован, он бива пријављен и додељује му се IP адреса. У супротном, CMTS одбија захтев за успостављање везе. Иако је MAC адреса јединствени идентификатор, иста се може променити у конфигурацији уређаја. Ово може довести до злоупотребе саобраћаја тако што се недозвољеном CPE уређају промени MAC адреса у адресу неког од претплатника и претплатнички CPE уређај и недозвољени CPE уређај се налазе на различитим CMTS-овима. Обзиром на до да CMTS-ови немају могућност међусобне комуникације, оба ће приликом пријаве регистровати да се ради о претплатничком уређају и дозволити саобраћај. Овакав сценарио је теоретски могућ уколико није имплементирана додатна заштита од стране оператора. На основу агрегације постојећих података и одговарајућих тагова може се креирати механизам заштите за описану ситуацију. Детектовани CPE уређаји који злоупотребљавају сервис су они који су у једном тренутку доступни на више од једног CMTS-а. Овај податак се може даље обогатити информацијом о претплатнику како би се идентификовало који корисник је прави претплатник, а који је клон. CMTS-у на ко м се налази клонирани CPE уређај се може послати сигнал за прекид везе са идентификованим уређајем. Цео процес се може аутоматизовати и на тај начин спречити злоупотреба.

7.2. Интеграција нових домена

ПНПМБД је оптимизована за складиштење временских серија. Развијени дата колектори подржавају прикупљање перформансних метрика са уређаја из HFC мреже с опцијом агрегације података. Међутим, могућности предложене платформе су далеко веће од тога.

Поред HFC домена интеграције, могуће је извршити интеграцију и са другим доменима у оквиру мреже оператора. Интеграција домена подразумева развој дата колектора који ће бити у стању да комуницира и прикупља податке са нових типова уређаја. Један пример интеграције је додавање нових типова метрика са већ интегрисаних уређаја. Пример интеграције овог типа јесте прикупљање WiFi података са кабловских модема. Уколико оператор поседује део мреже који користи DSL, иста се такође може додати у постојећу платформу. Уколико оператор жели да прошири портфолио и понуди корисницима IoT уређаје као део система паметних кућа, иста платформа се може искористити и за складиштење и обраду ових података. Конкретан пример примене платформе у IoT описан је у следећем поглављу.

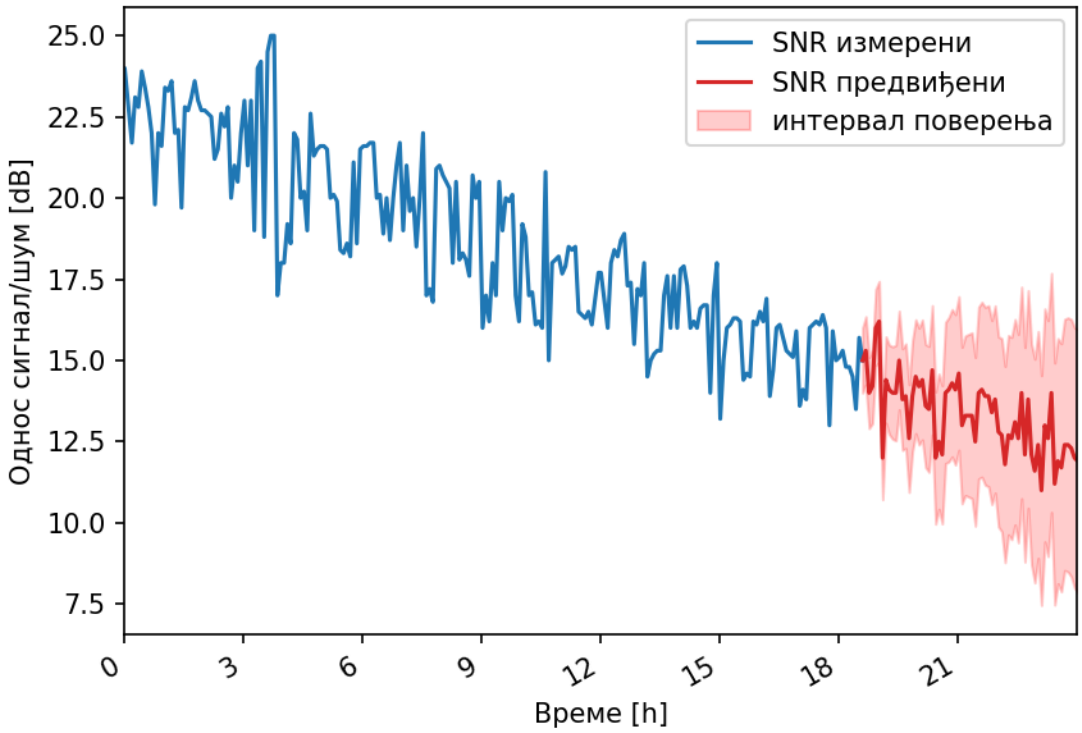
Обзиром на то да је један од слојева ПНПМБД архитектуре HDFS, она се може користити за складиштење свих врста података и фајлова. Могућност интеграције омогућава комуникацију са другим системима. С тим у вези, ПНПМБД платформа се може користити као место за историјско одлагање података и као *data lake*. *Data lake* концепт подразумева централизовано складиштење података са свих система на једно место. Дата аналитичари истражују податке складиштене у *data lake*-у и покушавају да пронађу међузависности и корисне информације које могу унапредити пословање компаније. Скалабилност HDFS-а обезбеђује ширење капацитета платформе додавањем нових дискова и сервера па је самим тим обезбеђен теоретски неограничен капацитет.

7.3. Примена алата машинског учења

Прикупљени подаци служе за надгледање перформанси мреже. Подаци се користе како би се стекао увид у стање мреже као и ефикасно решили детектовани проблеми и откази. Поред тога, прикупљени подаци се могу користити за предиктивно надгледање

мреже коришћењем техника машинског учења. Коришћење машинског учења може предвидети деградацију сигнала пре него што до ње дође. Проактивним корекцијама на мрежи се могу спречити ситуације отказа и лошег квалитета сигнала.

Примена машинског учења се врши у два корака. У првом кораку се одвија тренирање модела, а у другом примена истренираног модела над тренутним подацима. Тренирање модела машинског учења врши се на основу историјских података прикупљених метрика. Приликом тренирања, модел учи понашање временских серија. Да би модел био што боље истриниран, пожељно је користити што више историјских података. Обзиром на то да ПНПМБД поседује велику количину историјских података, овај захтев је аутоматски испуњен. Истренирани модел у стању је да предвиди вредност сигнала са одређеним интервалом поверења. На слици 7.3.1, дат је пример предикције односа сигнал/шум. Плава линија представља измерену вредност, док црвена приказује предвиђену вредност метрике. Поред тога, на слици 7.3.1 је приказан интервал поверења предикције. Са слике се може видети како интервал поверења опада с временом. То значи да ће предикције са највећом вероватноћом бити тачне за времена близу очитавања, а мање тачне како се предвиђа даље кроз време. Током примене алгоритма, тренирани модел упоређује предвиђену вредност са стварно очитаном вредношћу и коригује се у складу са тим. На тај начин осигурано је да је модел истриниран најсвежијим подацима што максимизира његову релевантност.



Слика 7.3.1. Предикција односа сигнал/шум.

Са техничке стране гледишта, ПНПМБД-у је неопходно додати библиотеке машинског учења како би ова функционалност била могућа. У данашње време постоји велики број библиотека машинског учења које се могу применити у ову сврху. Неке од најпознатијих су Spark MLlib [135], PyTorch [136] и TensorFlow [137]. Поменуте библиотеке поседују конекторе ка Apache Spark, па се за тренирање модела може једноставно користити процесорска моћ кластера.

8. Будући правци развоја

Паметне куће и паметни градови представљају IoT концепте који су веома популарни у последње време [55][58]. Узевши у обзир број становника и чињеницу да сваки од њих може имати неколико IoT уређаја и сензора у својим домовима, очигледно је да је генерисана количина података огромна. Због тога, биг дата технологија представља једно од најподеснијих решења за складиштење и обраду овакве количине података [53]. Већина резиденцијалних корисника покривени су неким телекомуникационим оператором (HFC, PON, ADSL (*Asymmetric Digital Subscriber Line*) итд.). Већина телекомуникационих оператора већ користе биг дата платформе како би прикупљали и обрађивали податке из своје мреже. Те постојеће платформе се могу искористити и за прикупљање и складиштење података са IoT уређаја. На овај начин се може искористити постојећа инфраструктура и за IoT сврхе. Интеграцијом постојећих биг дата решења телекомуникационих мрежа, могуће је развити високо економично и ефикасно решење за паметне градове.

У претходним поглављима представљено је решење засновано на биг дата технологији за прикупљање перформансних метрика из мреже кабловских оператора са подршком детекције и локализације отлаза у мрежи. У овом поглављу ће бити представљен начин за проширење ПНПМБД за подршку IoT уређаја на локацији корисника. Проширење не захтева архитектурне промене постојеће платформе, већ захтева развој дата колектора који ће подржавати нову групу уређаја. Прикупљени подаци се могу користити за различите сврхе као, на пример, нотификација у случају прекорачених прагова (детекција дима и пожара), извештавање о сензору или логичкој групи сензора, аналитика о квалитету ваздуха.

Постоји велики број истраживања на тему IoT мрежа и њихова интеграција са биг дата технологијом. Решења паметних кућа заснована на IoT технологији представљају једну од најважнијих и најатрактивнијих области за развој обзиром да се њиховом применом може побољшати квалитет живота становника. Решења паметних кућа обухватају различите типове сензора и уређаја зависно од потреба корисника, параметара стамбеног објекта као и географске локације где се исти налази. У решењима за негу старијих лица, користе се различити здравствени сензори у виду наруквица и других уређаја који се носе обезбеђујући на тај начин праћење здравственог стања и кретања пацијента [138]. Како би се надгледало окружење, могуће је инсталирати различите сензоре као, на пример, сензоре за температуру, влажност ваздуха, осветљености, детекторе дима/пожара итд. На основу мерења могуће је контролисати температуру и осветљеност унутар стамбеног објекта као и вршити одговарајућа алармирања. Поред тога, коришћењем одговарајућих сензора, могуће је вршити надгледање потрошње електричне енергије и воде као и детекцију цурења, тј. плављења. За повећање сигурности дома, могуће је инсталирати аудио и видео сензоре као и детекторе покрета [59].

Пројекти паметних градова настоје да реше многе урбане проблеме као што су загађење ваздуха, надгледање градске буке, гужве у саобраћају, системе за детекцију паркинг места, потрошњу енергије, управљање смећем, градско осветљење, помоћ старијим грађанима итд. [56]. Ово подразумева коришћење различитих технологија (биг дата, IoT, WSN (*Wireless Sensor Networks*)) и *cloud* решења. Један од начина комуникације великог

броја сензора у градским областима је WiFi. Неколико различитих решења за планирање градских WiFi мрежа је представљено у [139]. WiFi као решење за пренос података је избор многих паметних градова због тога што је овај протокол постао стандард за комуникацију многих паметних уређаја као што су рачунари, паметни телефони, сатови итд. Многи изазови приликом развоја система паметних градова су описани у [55] где су као најважнији препознати безбедност и приватност, мрежа паметних сензора и анализа биг дата података. Анализа и будући правци у развоју биг дата технологија за подршку паметних градова је дата у [140].

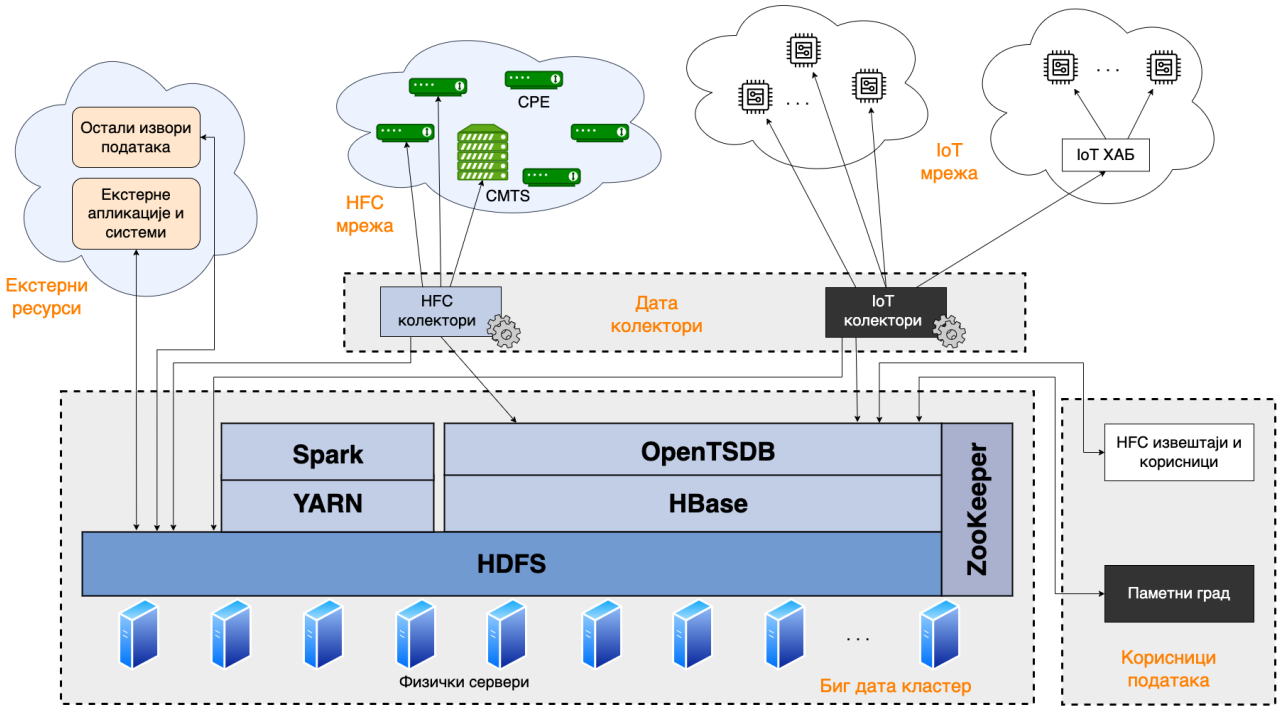
IoT концепт је један од најважнијих трендова у модерним телекомуникацијама. IoT је кључан део многих паметних решења као што су паметни градови, паметне зграде, паметна агрикултура, саобраћај, здравство итд. [54]. Узевши у обзир огромну количину података које овакви системи генеришу, биг дата је препозната као кључна технологија за могућност постојања IoT [54][141]. Због овога, постоји велики број научних радова (на пример, [54][142]) који се баве применом биг дата технологије у IoT. Примена машинског учења и комплексне обраде стримова података у сензорским (IoT) мрежама дискутована је у [143]. Коришћењем ових технологија, предложено је решење за предикцију загушења линка. Предложено решење комбинује неколико различитих извора података и дефинише правила одлучивања. Коначно, предикција загушења линка се врши на основу тренираног модела. Решење за складиштење временских серија података прикупљених са сензора дато је у [52]. Предложени систем представља скалабилну, модуларну платформу за складиштење сензорских очитавања. За комуникацију са сензорима користи се MQTT (*Message Queuing Telemetry Transport*), REST API протокол и веб сокет док се за складиштење података користи InfluxDB. Примена машинског учења над временским серијама предложена је у [144]. Аутори предлажу неколико различитих алгоритама машинског учења над временским серијама у циљу предикције цене акција на берзи.

Велики број научних радова на тему паметних градова показује да је ова тема најпривлачнија у имплементацији IoT система. Преглед биг дата решења у паметним градовима дат је у [57]. Индустријска производња такође представља једну од области за дигитализацију и примену IoT система у циљу побољшане ефикасности. Примена биг дата и IoT технологија у индустрији је дискутована у [145], у фармама и агрикултури у [146], а у здравству у [147]. Иако многа IoT решења раде у синергији са биг дата системима, постоји велики број изазова које треба решити [148].

8.1. Проширена ПНПМБД архитектура за подршку IoT мрежа

Предложена ПНПМБД састоји се из слојевите архитектуре и користи се за прикупљање перформансних метрика из телекомуникационих мрежа. Читава архитектура је развијена за ефикасно складиштење, приступ и обраду података временских серија са конкретном применом у телекомуникационим кабловским мрежама. Слој за прикупљање података (дата колектори) садржи информацију специфичну за домен интеграције, прикупља податке из мреже, форматира их на дефинисани начин и прослеђује делу за складиштење. У случају додавања новог типа уређаја, модела или софтверске верзије, неопходно је извршити адаптацију дата колектора. IoT мреже су веома сличне кабловским мрежама. Уређаји у мрежи су у стању да прикупљају и прослеђују перформансне метрике. Поред тога, ових уређаја у једној мрежи, као и CPE опреме, има веома много што представља биг дата изазов. У наставку је дат опис на који начин је могуће проширити постојећу ПНПМБД за прикупљање података из IoT мреже.

На слици 8.1.1, приказана је архитектура ПНПМБД с подршком за IoT мреже. Додатни модули обележени су црном бојом. Предложена архитектура представља проширену верзију архитектуре предложене у [149] која представља резултат рада на дисертацији. У [149] IoT уређаји су везани за локацију претплатника, док слика 8.1.1 приказује генерализовану варијанту проширења ПНПМБД подршком за IoT. Са слике 8.1.1 се види да је главна модификација коју је потребно урадити развој новог дата колектора за прикупљање података из IoT мреже. Наиме, сваки тип уређаја има своје специфичности приликом комуникације, прикупљања података и форматирања истих у складу са дефинисаним форматом. Предност ПНПМБД је у њеној флексибилности. Услед великог броја различитих типова уређаја у телекомуникационим мрежама (различити типови опреме, произвођачи, модели итд.), ова платформа је дизајнирана на начин да ефикасно реши биг дата *Variety* изазов тј. изазов ”разноврсности”. Обзиром на то да се и у IoT мрежама, као и у HFC, прикупљају подаци у виду временских серија, IoT уређаји се могу посматрати као нови тип с колекторске тачке гледишта. Због тога, развој новог дата колектора представља минорни посао у односу на развој комплетне платформе. Што се тиче повећања простора и процесорске моћи биг дата кластера, неопходно је само извршити скалирање хардвера додавањем нових сервера у складу са очекиваним саобраћајем који ће бити додат.



Слика 8.1.1. Архитектура ПНПМБД са подршком за IoT мреже.

Након што су подаци прикупљени, потребно је развити одговарајуће агрегације и извештаје неопходне за кориснике, на пример, окружење паметних градова као што је приказано на слици 8.1.1. IoT корисници могу бити интерни и екстерни у зависности од типа интеграције са биг дата платформом. На пример, биг дата платформа се може користити, поред класичног система за надгледање перформанси, као и IoT сервисни хаб управљан од стране мрежног оператора (интерни корисник података). Поред тога, биг дата платформа може обезбедити приступ екстерним корисницима података као што су, на пример, државна платформа паметног града или екстерне компаније као дистрибутери електричне енергије,

воде и гаса. Ове агрегације се могу креирати на постојећој архитектури без физичких модификација. Приликом развоја агрегација и извештаја, неопходно је консултовати се са корисницима података како би се обезбедиле агрегације у складу са њиховим потребама.

У зависности од типа IoT мреже, постоје два могућа приступа за прикупљање података. Оба приступа захтевају развој специфичног дата колектора. Један приступ је прикупљање података директно са IoT уређаја. Други приступ представља IoT мреже где су IoT уређаји на једној локацији повезани преко хаба (на пример, један хаб у стану и сви IoT уређаји су повезани преко њега). У другом приступу, дата колектор се повезује само са хабом што смањује број конекција потребних за прикупљање података са сензора. Додатна предност коришћења приступа са хабом јесте једноставнија структура дата колектора. Један хаб прикупља податке са више IoT уређаја, па се самим тим једним позивом могу прибавити подаци са свих њих. Поред тога, хаб се напаја сталним извором електричне енергије што не мора бити случај са IoT уређајима (могу користити батерије, соларни или неки други извори енергије), што додатно поједностављује дата колектор јер се исти не мора бринути о ситуацији када је неки сензор недоступан. IoT мреже које користе хаб дају доста флексибилности. Уређаји се могу повезати на хаб коришћењем различитих протокола (на пример, ZigBee или Bluetooth). Поред тога, уколико IoT уређаји користе WiFi за комуникацију са централним хабом, хаб се може интегрисати директно у кабловски модем обзиром на то да данашњи модели модема имају уграђен WiFi модул у случају кад је IoT мрежа на локацији претплатника HFC мреже. Оба приступа су илустровани на слици 8.1.1.

Проширење постојеће ПНПМБД за подршку IoT уређаја отвара могућности примене паметног надгледања. IoT и сензорске мреже представљају изворе података временских серија који се могу прикупљати, складиштити и процесуирати у ПНПМБД. Као што је дискутовано, постоје две могућности за комуникацију са IoT уређајима, директно и преко IoT хаба. У пракси постоје и архитектуре где се ова два приступа могу комбиновати у хибридни модел.

Једна од најважнијих и најзаступљенијих паметних мерења базираних на IoT у стамбеним јединицама су надгледање квалитета ваздуха у простору, потрошње електричне енергије и потрошње воде. Уређаји за надгледања ових метрика генеришу периодични и мали пакетски саобраћај. Најзаступљеније метрике приликом надгледања квалитета ваздуха у простору су CO, CO_2, PM (*Particle Matters*) честице, VoC (*Volatile Organic Compounds*) материје, температура и релативна влажност. Поред основних, неки уређаји подржавају и надгледање O_2, CH_4, H_2S, NH_3. На тржишту већ постоји велики број јефтиних уређаја који подржавају ова мерења [150]. Ови уређаји се могу користити за надгледање квалитета ваздуха и детекцију пожара. У нормалним условима, прикупљање података је подешено на 5 минута, док је за период алармирања могуће је подесити 1 минут. Очекивана количина саобраћаја је 0.5 KB по једном мерењу. У случају надгледања на 1 минут, очекивана количина саобраћаја је 720 KB на дневном нивоу, тј. 21.6 MB на месечном. Уколико се оригинални податак обогати додатним информацијама (на пример, име корисника, модел, тип итд.) количина података ће бити мало већа. Неопходан простор за складиштење података за један месец ће бити око 5.7 TB под следећим претпоставкама: свака стамбена јединица у просеку има 2 IoT уређаја, прикупљена количина података са једног уређаја је просечно 30 MB на нивоу једног месеца, 100000 стамбених јединица је покривено HFC мрежом.

Телекомуникациони оператори могу обезбедити приступ сировим подацима коришћењем API-ја. Ова информација може бити превише детаљна и збуњујућа за корисника који нема потребно знање за тумачење истих. Уместо тога, подаци се могу агрегирати и

визуелно представити у извештајима који би корисницима били доступни путем веб портала. Извештаји би били развијени на начин да крајњем кориснику буду интуитивно јасне вредности метрика. Поред тога, коришћењем веб портала, корисник може итеративно подесити жељене нотификације, рачунати индекс квалитета ваздуха итд. У случају различитих модела и верзија IoT уређаја, оператори могу креирати агрегације података по произвођачу и моделу како би имали увид у перформансе различитих типова уређаја па самим и понудили клијентима боље уређаје. У случају уређаја за мерење квалитета ваздуха, у стамбеним јединицама могуће је инсталирати и сензоре за надгледање спољашњег квалитета, на пример, на терасама или прозорима. Подаци прикупљени са ових уређаја се могу обрадити и агрегирати како би се добио увид у квалитет ваздуха у различитим деловима града, квалитет ваздуха у зависности од висине у случају вишеспратних зграда. Ове информације могу бити значајне локалним заједницама у циљу прављења мапа загађења, детекције критичних зона и укупног побољшања квалитета ваздуха у градовима.

Док надгледање квалитета ваздуха захтева фреквентно надгледање, потрошња електричне енергије [151] и воде [152] се може надгледати једном дневно (око 0.4 KB података по једном мерењу у оба случаја, што значи око 12 KB на месечном нивоу). Мерења потрошње електричне енергије и воде се могу складиштити и обрађивати у ПНПМБД. Прикупљена мерења се могу представити у виду графикона у извештајима. Поред тога, крајњи корисник се може обавестити у случајевима прекорачења дефинисаног прага потрошње на месечном нивоу. Концепт паметних градова подразумева међусобно повезивање грађана, локалне власти и компанија које пружају услуге грађанима. Ове компаније треба да врше очитавања потрошње на месечном нивоу како би могли исту да наплате грађанима. Уколико се врше поменута мерења у складиште у ПНПМБД, онда би се ови подаци могли агрегирати и слати директно компанијама. На овај начин се добија аутоматизовано и даљинско очитавање потрошње, а самим тим и уштеда средстава, као и детекција аномалија која могу алармирати истовремено и компаније и кориснике. У случају вишеспратних зграда, мерна бројила обично нису у стамбеним јединицама. У таквим случајевима, уколико је зграда покривена HFC мрежом, додатни кабловски модем се може инсталирати који би покривао само мерна бројила.

Као што је већ објашњено, предложени примери употребе IoT уређаја нису превише захтевни по питању складиштења података па би ПНПМБД могла једноставно да подржи велики број корисника. Увид у прикупљена мерења обезбеђује корисницима корисне информације о квалитету ваздуха као и детаљан увид у потрошњу енергије и воде, а самим тим простор за оптимизацију трошкова. Предложени концепт олакшава интеграцију паметних домова и може се једноставно проширити на интеграцију паметних градова са циљем прављења мапа загађења и буке (уређаји за квалитет ваздуха се могу монтирати на спољашњим деловима стамбених јединица), даљинско очитавање потрошње електричне енергије и воде, а самим тим и оптимизација трошкова компанија. С друге стране, предложена IoT интеграција пружа прилику телекомуникационим операторима проширење сервисног портфолија и понуду нових паметних сервиса корисницима што доводи до повећања профита и задовољства клијената.

Потребни IoT уређаји су јефтини и једноставни за инсталацију и управљање. Услед пораста загађености ваздуха у густо насељеним областима, потреба за поузданим и приступачним системима за праћење квалитета ваздуха је у порасту. Јефтини *low-cost* комерцијални сензори представљају добру почетну тачку за праћење квалитета ваздуха с обзиром на њихову доступност на тржишту и ниску цену, али заостају у смислу тачности и поузданости измерених података у односу на јавне мерне станице. С друге стране, јавне

мерне станице су скупе, лоциране на фиксним местима и захтевају значајну количину редовног одржавања. На основу доступних информација од продаваца уређаја са овим сензорима, просечан однос цена је између 1:20 и 1:25, односно цена једне јавне станице за праћење квалитета ваздуха је упоредива са ценом 20-25 јефтиних уређаја за исти скуп посматраног загађивача. Јавна референтна станица додатно захтева знатно већи ниво одржавања. Пример једног могућег јефтиног уређаја за праћење квалитета ваздуха може се наћи на [153]. У посматраном случају димензије уређаја су 180×180×265 mm^3 (за ралику од јавних мерних станица које су смештене у великим контејнерима [154]) тежина 1,5 кг и потрошња енергије 2,5 W. Напајање се може реализовати преко адаптера за 220 W, пуњиве батерије или преко соларног панела. Подржане су различите технологије преноса података: јавне мобилне мреже (2G, 3G, 4G), WiFi, а у зависности од коришћених сензора величина пакета са подацима је у распону 0.1 - 0.5 KB.

Перформансе *low-cost* сензора (тачност мерења) веома су осетљиве на радне услове околине, тј. релативну влажност и температуру ваздуха, због процеса детекције гаса (коришћени сензори су електрохемијски), који укључује прилично сложене реакције у зависности од услова околине, а одговарајуће хемијске реакције такође варирају од доба дана/ноћи, што додатно умањује перформансе сензора. Произвођачи сензора, генерално обезбеђују корекционе факторе за температуру и релативну влажност, међутим, за спољашње услове у којима се релативна влажност и температура могу значајно променити на дневној и сезонској основи, потребно је имплементирати софистицираније корекције. Сензор за надгледање суспендованих честица (оптички сензор), с друге стране, на већим вредностима релативне влажности ваздуха има мању тачност мерења с обзиром да тада долази до повећања дијаметра суспендованих честица услед „лепљења" честица влаге на суспендоване честице. Такође, сваки сензор има своју сопствену осетљивост на температуру и релативну влажност ваздуха који утичу на тачност мерења.

Поменута осетљивост сензора на температуру и релативну влажност ваздуха се тешко може моделовати једноставном функцијом, и када се ради о моделовању различитих зависности неопходно је применити комплексније алате тј. машинско учење (ML (*Machine Learning*)). Помоћу моћних алата ML-а могуће је прецизније моделовати зависности сензора од температуре и релативне влажности ваздуха и на тај начин омогућити прецизније и поузданије резултате мерења. Параметри околине, као што су релативна влажност и температура ваздуха (уколико су доступни, од користи би били и подаци о брзини и правцу ветра на мерном месту), неопходни су као улази за алгоритам за одговарајућу калибрацију. У раду [155] тестирано је више различитих алгоритама машинског учења: Линеарна Регресија, *Support vector machine* Регресија, *Ada Boost* Регресија, *Random Forest* Регресија, Регресија помоћу неуралних мрежа, и показано је да примена ML може значајно повећати тачност мерења *low-cost* сензора. У раду [156] аутори су користећи *low-cost* сензоре, испитивали утицај закључавања у Београду за време пандемије COVID 19 током 2020. године и евалуирали примену неуралних мрежа за корекцију измерених вредности сензора. Поменуте сензоре је могуће користити у затвореним просторијама, чиме се може пратити квалитет ваздуха у становима, школама, канцеларијама, фабричким халама итд. На основу свега наведеног, може се закључити да је могуће користити јефтиније уређаје за праћење квалитета ваздуха јер је могуће унапредити њихову прецизност одговарајућим алгоритмима. Отуда, оператори могу понудити својим корисницима овакве економичније уређаје у оквиру својих пакета што даје веће изгледе да се усвоје IoT решења у домовима корисника.

Поред поменутих типова сензора, постоје и многи други који се могу инсталирати за надгледање различитих параметара. Кроз пример сензора за квалитет ваздуха и потрошњу

електричне енергије и воде представљен је концепт примене IoT мреже у паметним домовима и паметним градовима уз конкретне примере примене.

8.2. Биг дата архитектура за подршку мобилних мрежа

Мобилни оператори, као и кабловски, су због своје велике популарности такође суочени са проблемом огромне количине података коју треба сачувати и обрадити. Мобилне мреже представљају комплексне слојевите системе који захтевају високу доступност и квалитет рада. Висока конкуренција на тржишту, висока очекивања крајњих корисника по питању поузданости сервиса и интернет протока су само неки од покретача за константно унапређење мреже. Поред тога, увођењем 5G мобилних мрежа, мобилни оператори повећавају спектар сервиса, посебно у IoT сфери, што даље повећава количину прикупљених података и доноси нове изазове.

Због огромне количине података која се генерише и складишти, биг дата технологије представљају логичан избор приликом развоја система за прикупљање и обраду података. Иако већ постоји велики број радова на тему примене биг дата технологија у мобилним мрежама [9], [43-47], [77,78], ова тема је и даље актуелна како због константног развоја мобилних мрежа, тако и због развоја биг дата технологија. Тренд развоја ових система креће се у правцу (поред стандардног складиштења и обраде података) минимизације времена између генерисања податка и генерисања повратне информације на основу истог. Због тога, технологије обраде стримова података у реалном времену све више добијају на значају.

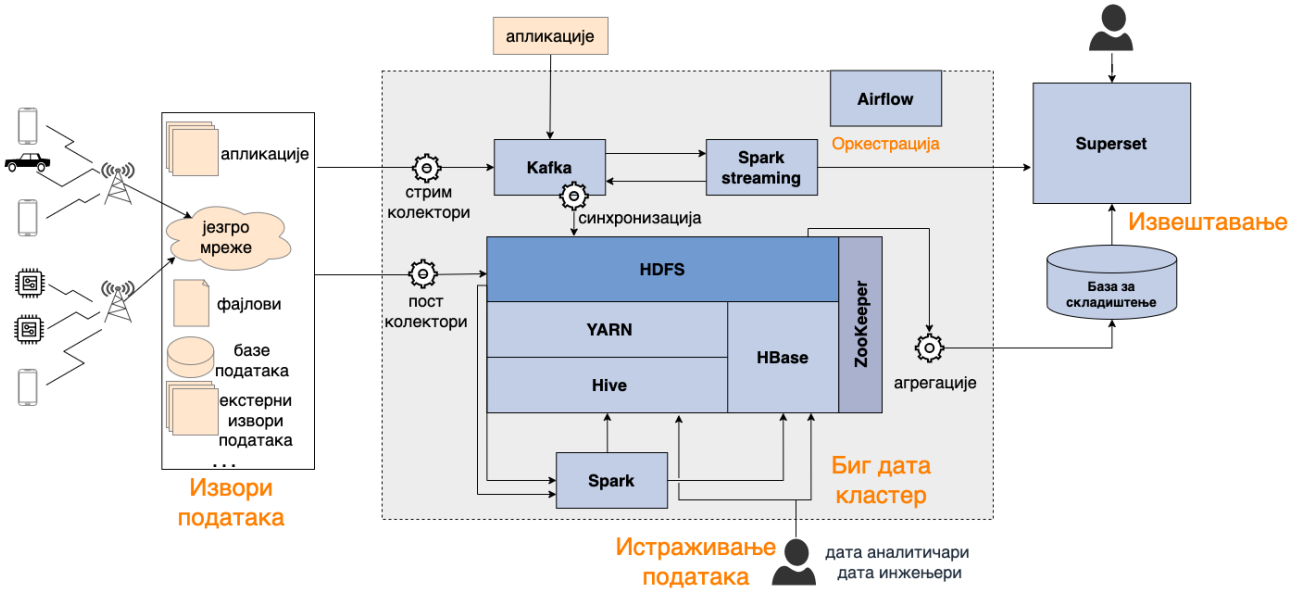
У наставку ће бити представљен нацрт архитектуре за обраду података прикупљених из мобилних мрежа као један од праваца даљег развоја. Иницијална архитектура предложена је у [82]. У наставку истраживања фокус ће бити даља оптимизација платформе предложене у [82], интеграција са новим системима и имплементација нових шема и агрегација неопходних за решавање специфичних потреба оператора.

Табела 8.2.1. Извори података у мобилним мрежама.

Група	Извори података
Језгро мреже	Подаци са мрежних компоненти (рутери, свичеви, *firewall*), мрежна опрема (ћелије, SGSN-ови (*Serving GPRS* (*General Packet Radio Service*) *Support Node*), GGSN-ови (*Gateway GPRS Support Node*)), CDR, XDR
Апликације	Наплата сервиса, мобилне апликације, промоције, веб сајт, системи за бригу о корисницима, управљање уређајима, управљање наруџбинама, IoT сервиси
Екстерни извори података	*Google analytics*, рекламни сајтови

Узевши у обзир комплексност мобилних мрежа, постоје многи извори података који се могу сматрати као добри кандидати за прикупљање и обраду. Неки од најзначајних извора података приказани су у табели 8.2.1. Листа предложених система за прикупљање зависи од инфраструктуре провајдера и варира у складу са тим. На основу количине информација коју прикупљени подаци носе као и брзину генерисања, извори података се могу поделити у пост и стрим. Пост подаци се прикупљају у предефинисаним периодама, на пример, једном дневно. С друге стране, стрим извори константно генеришу податке, те их је неопходно прикупити и обрадити истог тренутка кад су и генерисани, или што је могуће пре. Неки од најважнијих кандидата за стриминг прикупљање у мобилним мрежама су CDR-ови, XDR-

ови, лог фајлови, рекламни сајтови, IoT итд. Добри кандидати за пост обраду су споропроменљиви извори података као, на пример, системи за бригу о корисницима, управљање уређајима, управљање наруџбинама итд. Треба имати у виду да се сваки извор података може третирати као стрим и обрнуто у зависности од реалних потреба прикупљања.



Слика 8.2.1. Биг дата архитектура за подршку мобилних мрежа.

Нацрт система за прикупљање података у мобилним мрежама заснован на биг дата технологијама представљен је на слици 8.2.1. Предложена архитектура прати ламбда (λ) приступ. Наиме, архитектура се састоји из две гране, стрим и периодичне. На основу типа извора података, сваки бива обрађен једном од ове две гране. На стрим грани, подаци се прослеђују на Kafka-у, систем за управљањем редовима података, након чега бивају обрађени у реалном времену. За обраду података у реалном времену користи се Spark streaming. Резултати агрегација у реалном времену су или директно доступни некој апликацији или се уписују у Kafka-у у нови ред за чекање одакле се могу конзумирати од стране других система. Поред тога, основни стрим извори података се периодично уписују из Kafka-е на HDFS за периодичну обраду и историјско складиштење. У грани пост обраде, подаци се прикупљају и обрађују периодично у зависности од осетљивости података (нпр. сваког сата, дневно, недељно итд.). Уколико су сачувани подаци на HDFS-у структурирани, истим се може приступити користећи SQL и Hive. Полуструктурирани подаци којима је неопходан брз приступ и упис (као што су временске серије перформантних метрика у случају кабловских мрежа) ће бити уписани у HBase. Након што су подаци агрегирани, резултати агрегације ће бити уписани у базу за складиштење (енгл. *data warehouse*). Резултати сачувани овде ће бити доступни слоју за извештавање, као што је Superset у предложеној архитектури. За пост агрегацију предложен је Spark. За дистрибуирано оркестрацију свих послова складиштења и обраде података предложен је Airflow.

9.ЗАКЉУЧАК

У данашње време, модерне телекомуникационе мреже омогућавају међусобну комуникацију и дељење информација сваког појединца на планети, а самим тим директно утичу на убрзани развој човечанства. Висока конкуренција на тржишту телекомуникација и све већа очекивања од стране корисника су кључни мотиви за обезбеђивањем сервиса високог квалитета и високе доступности. Како би одговорили на ове изазове, телекомуникациони оператори надгледају перформансе својих мрежа што представља један од главних параметара за њихово даље унапређење. Чак и у ситуацијама привременог отказа мреже, пожељно је решити исте у што краћен временском року. У случају мреже лошег квалитета и поузданости, оператори ризикују губитак корисника, па чак и сношење финансијских последица у случају правних корисника.

Обзиром на велике количине података и на комплексност мрежа, прикупљање и складиштење података је веома изазован посао. С тим у вези, биг дата технологија се намеће као логично решење за развој система за надгледање перформанси. Биг дата технологија представља релативно нов концепт који се доказао у многим областима где постоји један или више ”5V” биг дата изазова као што су, на пример, електродистрибуција, IoT, трговина, производна индустрија итд. Услед велике примене и брзог развоја, биг дата технологија постаје довољно зрела да постане кандидат за продукциона окружења, а самим тим и за надгледање перформанси мрежа. Постоје комерцијална решења на тржишту, као, на пример, SolarWinds NPM (*Network Performance Monitor*) [157] и ManageEngine OpManager [158], која делимично могу решити претходно поменуте проблеме, али је цена таквих решења често ограничавајући фактор за многе телекомуникационе операторе, посебно у мање развијеним земљама. Поред тога, проблем готових комерцијалних решења је у њиховој затворености, прилагодљивости и флексибилности, па се самим тим намеће развој система заснован на технологији отвореног кода и његовој оптимизацији за конкретну намену.

У овој дисертацији је представљено ПНПМБД решење, које је засновано на биг дата технологији, намењено надгледању перформанси HFC мрежа. Први допринос у овој дисертацији је анализа постојећих решења у литератури на тему прикупљања перформансних метрика у телекомуникационим мрежама, као и анализа постојећих изазова. Анализа изазова прикупљања метрика уопштено у телекомуникационим мрежама, а и специфично у HFC мрежама, представљена је кроз биг дата ”5V” модел изазова. Други, а уједно и највећи, допринос представља ПНПМБД који се у пракси користи за надгледање перформанси у HFC мрежама. При томе, ПНПМБД архитектура је заснована на биг дата алатима отвореног кода па је самим тим економски исплатив и не захтева лиценцирање. Предложено решење је скалабилно и одговара на ”5V” биг дата изазове у надгледању перформанси телекомуникационих мрежа. Капацитет и процесорска моћ платформе се могу једноставно скалирати додавањем одговарајућег хардвера. Предложено решење се може користити и у другим телекомуникационим мрежама, поред HFC мрежа, где постоји проблем великих података и где се врши прикупљање временских серија као, на пример, мобилне, DSL и IoT мреже. У зависности од потреба, предложено решење се може имплементирати на физичким серверима, хибридној или *cloud* инфраструктури. Предложено решење може

комуницирати са другим системима, па се самим тим може једноставно интегрисати у постојећи екосистем телекомуникационих оператора. Поред прикупљања и обраде перформансних метрика из HFC мреже, овај систем се може користити као централизована платформа за историјско складиштење и обраду, агрегацију података из било којих других извора, што представља трећи допринос.

ПНПМБД је имплементирана за надгледање перформанси у HFC мрежама. Током имплементације решене су специфичности везане за HFC мреже што уједно представља четврти допринос ове дисертације. Ово подразумева предлог метрика за прикупљање, развој шеме података високих перформанси за упис и читање и решавање изазова доделе динамичке адресе CPE уређајима. Поред перформанси уређаја са којих су подаци прикупљени, предложен је алгоритам за процену стања уређаја са којих није могуће прикупљање података (на пример, AP, AMP и ON) што представља пети допринос дисертације. Алгоритам користи податке прикупљене са CPE уређаја и, на основу топологије, повезује исте са неинтелигентним уређајима и врши процену стања. Шести допринос ове дисертације представља ДЛОКМ, иновативни приступ за брзу детекцију и локализацију отказа у мрежама коришћењем биг дата платформе. Неинтелигентни мрежни елементи по пут ON, AMP и AP нису у стању да прикупљају метрике и поделе исте са платформом. Коришћењем информације о топологији мреже, ДЛОКМ може детектовати и локализовати проблематичне мрежне уређаје. На овај начин је обезбеђено комплетно аутоматизовано надгледање ове врсте уређаја. Додатно, трошкови надгледања мреже су смањени јер предложени механизам користи постојећу биг дата платформу која прикупља и обрађује податке па самим тим нема потребе за додатним хардвером. Главни бенефити овог механизма су ефикасна аутоматизација детекције отказа у HFC мрежама и локализација проблема над мрежним елемената који се не могу надгледати директним прикупљањем података.

Будућа истраживања се могу груписати у два правца. Први правац, дискутован у седмом поглављу, се бави додатним могућностима развијене ПНПМБД. Под тим се подразумева анализа прикупљених података у циљу детекције скривених информација које могу додатно да унапреде пословање оператора HFC мрежа, повећају задовољство претплатника и др. Поред тога, биће размотрена интеграција нових домена у развијену платформу. Интеграција подразумева прикупљање нових типова података релевантних за HFC операторе, као, на пример, подаци о квалитету WiFi мрежа на локацији корисника. Такође, биће размотрена и употреба алгоритама машинског учења за потребе предикције различитих ситуација и трендова у HFC мрежи. На пример, примена алгоритама машинског учења у циљу аутоматске детекције деградације квалитета сигнала. Технике машинског учења се могу користити за превентивно спречавање перформансних проблема, као што су, на пример, детекција деградације SNR-а или загушење капацитета линка. Други правац развоја, дискутован у осмом поглављу, се тиче примене постојеће платформе за интеграцију у друге домене, конкретно IoT и мобилне мреже. Обзиром на то да сваки корисник на локацији има све више уређаја који су у стању да комуницирају (сензори температуре, влаге, квалитета ваздуха, детектори присуства, мерачи протока воде, потрошње електричне енергије, паметни прекидачи итд.), постојећа платформа се може проширити за пружање IoT сервиса паметних кућа и паметних градова претплатницима HFC мреже оператора. Поред тога, могућа је и генерализација IoT подршке, уз неопходне адаптације, тако да постојећа платформа може да се примењује у било којим другим IoT мрежама, независно од телекомуникационог оператора. Поред HFC и IoT, постојећа архитектура се може проширити за подршку мобилним мрежама. Обзиром на то да мобилне мреже имају више

различитих типова података, неопходно је проширити постојећи систем додатним компонентама. На пример, у циљу подршке обраде података неопходно је додати Kafka и Spark streaming. Поред тога, за складиштење агрегација, могуће је додати централизовану базу података која ће се користити и за извештавање. Као и код имплементације HFC, доста изазова се очекује током самог развоја подршке за мобилне мреже што ће такође бити покривено даљим истраживањем.

ЛИТЕРАТУРА

[1] S. O'Dea, „Forecast number of mobile users worldwide from 2020 to 2025," July 2021. Available: https://www.statista.com/statistics/218984/number-of-global-mobile-users-since-2010/ [Accessed: June 29, 2022.]

[2] „Internet World Stats". Available: https://www.internetworldstats.com/stats.htm [Accessed: June 29, 2022.]

[3] A. Russell, E. Frachtenberg, „Worlds Apart: Technology, Remote Work, and Equity," *Computer*, vol. 54, no. 7, pp. 46-56, July 2021, doi: 10.1109/MC.2021.3074121.

[4] V. Das Swain, K. Saha, G. Abowd, M. De Choudhury, „Social Media and Ubiquitous Technologies for Remote Worker Wellbeing and Productivity in a Post-Pandemic World," in proc. of *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, Atlanta, GA, USA, Oct. 2020, pp. 121-130, doi: 10.1109/CogMI50398.2020.00025.

[5] T. Nguyen, H. Nguyen, J. Eric Salt, B. Berscheid, „Zero-CP OFDM for DOCSIS-Based CATV Networks," *IEEE Transactions on Broadcasting*, vol. 65, no. 4, pp. 727-741, Dec. 2019, doi: 10.1109/TBC.2019.2904853.

[6] J. Schnitzer et al., „Toward Programmable DOCSIS 4.0 Networks: Adaptive Modulation in OFDM Channels," *IEEE Transactions on Network and Service Management*, vol.18, no. 1, pp. 441-455, Mar. 2021, doi: 10.1109/TNSM.2020.3044850.

[7] M. Chen, S. Mao, Y. Liu, „Big data: A Survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171-209, Apr. 2014, doi: 10.1007/s11036-013-0489-0.

[8] C. Emani, N. Cullot, C. Nicolle, „Understandable Big data: A Survey," *Computer Science Review*, vol. 17, pp. 70-81, Aug. 2015, doi: 10.1016/j.cosrev.2015.05.002.

[9] Y. He, F. Richard Yu, N. Zhao, H. Yin, H. Yao, R. Qiu, „Big Data Analytics in Mobile Cellular Networks," *IEEE Access*, vol. 4, pp. 1985-1996, Mar. 2016, doi: 10.1109/ACCESS.2016.2540520.

[10] A. Garcia, M. Toril, P. Oliver, S. Luna-Ramirez, R. Garcia, „Big Data Analytics for Automated QoE Management in Mobile Networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 91-97, Aug. 2019, doi: 10.1109/MCOM.2019.1800374.

[11] Y. Zhang et al., „A Novel Big Data Assisted Analysis Architecture for Telecom Operator," in proc. of *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS)*, Shenyang, China, Oct. 2019, pp. 611-615, doi: 10.1109/IUCC/DSCI/SmartCNS.2019.00128.

[12] Y. Jia, K. Chao, X. Cheng, L. Xu, X. Zhao, L. Yao, „Telecom Big Data Based Precise User Classification Scheme," in proc. of *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*

(SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, UK, Aug. 2019, pp. 1517-1520, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00273.

[13] J. Mashey, „Big Data and the Next Wave of InfraStress," Usenix 1998.

[14] D. Laney, „3D Data Management: Controlling Data Volume, Velocity and Variety," META Group, Feb. 2001.

[15] J. Guterman, „Big Data, Release 2.0: Issue 11," Radar, O'Reilly, Jun. 2009.

[16] A. Jain, „The 5 V's of big data – IBM Watson Health Perspectives," Sept. 2016. Available: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/ [Accessed: June 29, 2022.]

[17] D. deRoss, P. C. Zikopoulos, B. Brown, R. Coss, R. Melnyk, „Hadoop For Dummies," John Wiley & Sons, Inc., 2014.

[18] Businessweek, „Sensor Networks Top Social Networks for Big Data," Sept. 2010. Available: https://cacm.acm.org/news/99109-sensor-networks-top-social-networks-for-big-data/fulltext [Accessed: June 29, 2022.]

[19] „Number of commercial flights tracked by Flightradar24, per day (UTC time)," May 2022. Available: https://www.flightradar24.com/data/statistics [Accessed: June 29, 2022.]

[20] J. Dean, S. Ghemawat, „MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107-113, Jan. 2008, doi: 10.1145/1327452.1327492.

[21] Microsoft, „Microsoft Azure". Available: https://azure.microsoft.com/en-us/ [Accessed: June 29, 2022.]

[22] Amazon Web Services, „Amazon Web Services". Available: https://aws.amazon.com/ [Accessed: June 29, 2022.]

[23] Google, „Google Cloud Platform". Available: https://cloud.google.com/ [Accessed: June 29, 2022.]

[24] IBM Corp., „IBM Cloud". Available: https://cloud.ibm.com/ [Accessed: June 29, 2022.]

[25] Oracle, „Oracle Cloud Infrastructure (OCI)". Available: https://www.oracle.com/cloud/ [Accessed: June 29, 2022.]

[26] Cloudera, Inc., „Cloudera". Available: https://www.cloudera.com/ [Accessed: June 29, 2022.]

[27] A. Katal, M. Wazid, R. Goudar, „Big Data: Issues, Challenges, Tools and Good Practices," in proc. of *2013 Sixth International Conference on Contemporary Computing (IC3)*, Noida, India, Aug. 2013, pp. 404-409, doi: 10.1109/IC3.2013.6612229.

[28] M. Simakovic, Z. Cica, „Big Data Applications and Challenges," in proc. of Infoteh 2016, Jahorina, BiH, Mar. 2016, pp. 675-678.

[29] The Apache Software Foundation, „HDFS Architecture," May 2022. Available: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html [Accessed: June 29, 2022.]

[30] „The Hadoop Ecosystem Table - GitHub". Available: https://hadoopecosystemtable.github.io/ [Accessed: June 29, 2022.]

[31] The Apache Software Foundation, „Apache Hadoop YARN," May 2022. Available: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html [Accessed: June 29, 2022.]

[32] The Apache Software Foundation, „Apache Spark". Available: https://spark.apache.org/ [Accessed: June 29, 2022.]

[33] IBM Corp., „Netezza Performance Server". Available: https://www.ibm.com/products/netezza [Accessed: June 29, 2022.]

[34] Teradata, „Teradata". Available: https://www.teradata.com/ [Accessed: June 29, 2022.]

[35] „Vertica". Available: https://www.vertica.com/ [Accessed: June 29, 2022.]

[36] M. Simakovic, I. Masnikosa, Z. Cica, „Performance monitoring challenges in HFC networks," in proc. of *2017 13th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, Nis, Serbia, Oct. 2017, pp. 385-388, doi: 10.1109/TELSKS.2017.8246305.

[37] The Apache Software Foundation, „Apache HBase". Available: https://hbase.apache.org/ [Accessed: June 29, 2022.]

[38] The OpenTSDB Authors, „OpenTSDB," Sept. 2021. Available: http://opentsdb.net/ [Accessed: June 29, 2022.]

[39] The Apache Software Foundation, „Apache Zookeeper". Available: https://zookeeper.apache.org/ [Accessed: June 29, 2022.]

[40] I. Hashem, I. Yaqoob, N. Anuar, S. Mokhtar, A. Gani, S. Khan, „The rise of "Big Data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, no. 1, pp. 98-115, Jan. 2015, doi: 10.1016/j.is.2014.07.006.

[41] P. Mell, T. Grance, „The NIST Definition of Cloud Computing (Technical report)," National Institute of Standards and Technology: U.S. Department of Commerce, doi: 10.6028/NIST.SP.800-145.

[42] J. Shaw, „Why "Big Data" Is a Big Deal," Harvard Magazine, Mar. 2014.

[43] J. Liu, F. Liu, N. Ansari, „Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop," *IEEE Network*, vol. 28, no. 4, pp. 32-39, Jul. 2014, doi: 10.1109/MNET.2014.6863129.

[44] D. Martinez-Mosquera, R. Navarrete, S. Lujan-Mora, „Development and Evaluation of a Big Data Framework for Performance Management in Mobile Networks," *IEEE Access*, vol. 8, pp. 226380-226396, Dec. 2020, doi: 10.1109/ACCESS.2020.3045175.

[45] M. Simakovic, Z. Cica, I. Masnikosa, „Big Data Architecture for Mobile Network Operators," in proc. of *2021 15th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, Nis, Serbia, Oct. 2021, pp. 283-286, doi: 10.1109/TELSIKS52058.2021.9606290.

[46] X. Cheng, L. Fang, L. Yang, and S. Cui, „Mobile Big Data: The Fuel for Data-Driven Wireless," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1489-1516, Oct. 2017, doi: 10.1109/JIOT.2017.2714189.

[47] J. Wen, V. Li, „Big-Data-Enabled Software-Defined Cellular Network Management," in proc. of *2016 International Conference on Software Networking (ICSN)*, Jeju, South Korea, May 2016, pp. 1-5, doi: 10.1109/ICSN.2016.7501923.

[48] M. Ghorbanian, S. Dolatabadi, P. Siano, „Big Data Issues in Smart Grids: A Survey," *IEEE Systems Journal*, vol. 13, no. 4, pp. 4158-4168, Dec. 2019, doi: 10.1109/JSYST.2019.2931879.

[49] H. Jiang, K. Wang, Y. Wang, M. Gao, Y. Zhang, „Energy big data: A survey," *IEEE Access*, vol. 4, pp. 3844-3861, 2016, doi: 10.1109/ACCESS.2016.2580581.

[50] M. Chen, J. Yang, L. Hu, M. Hossain, G. Muhammad, „Urban Healthcare Big Data System Based on Crowdsourced and Cloud-Based Air Quality Indicators," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 14-20, Nov. 2018, doi: 10.1109/MCOM.2018.1700571.

[51] T. Wang et al., „An Intelligent Dynamic Offloading from Cloud to Edge for Smart IoT Systems with Big Data," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2598-2607, Apr. 2020, doi: 10.1109/TNSE.2020.2988052.

[52] J. Rafferty et al., „A Scalable, Research Oriented, Generic, Sensor Data Platform," *IEEE Access*, vol. 6, pp. 45473-45484, July 2018, doi: 10.1109/ACCESS.2018.2852656.

[53] S. Khare, M. Totaro, „Big Data in IoT," in proc. of *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, July 2019, pp. 1-7, doi: 10.1109/ICCCNT45670.2019.8944495.

[54] M. Bansal, I. Chana, S. Clarke, „A Survey on IoT Big Data: Current Status, 13 V's Challenges, and Future Directions," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1-59, Nov. 2021, doi: 10.1145/3419634.

[55] A. Syed, D. Sierra-Sosa, A. Kumar, A. Elmaghraby, „IoT in Smart Cities: A Survey of Technologies, Practices and Challenges," *Smart Cities*, vol. 4, no. 2, pp. 429-475, Mar. 2021, doi: 10.3390/smartcities4020024.

[56] J. Temperton, „Bristol is making a Smart City for Actual Humans," Mar. 2015. Available: http://www.wired.co.uk/news/archive/2015-03/17/bristol-smart-city [Accessed: June 29, 2022.]

[57] I. Hashem et al., „The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748-758, Oct. 2016, doi: 10.1016/j.ijinfomgt.2016.05.002.

[58] A. Nag, M. Alahi, N. Afsarimanesh, S. Prabhu, S. Mukhopadhyay, „IoT for smart homes," in book Sensors in the Age of the Internet of Things: Technologies and applications, pp. 171-199, 2019.

[59] M. Alam, M. Reaz, M. Ali, „A Review of Smart Homes - Past, Present, and Future," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1190-1203, Nov. 2012, doi: 10.1109/TSMCC.2012.2189204.

[60] F. Yao, Y. Wang, „Financial Innovation System of Commercial Banks Based on Big Data Technology," in proc. of *2021 6th International Conference on Smart Grid and Electrical Automation (ICSGEA)*, Kunming, China, May 2021, pp. 303-306, doi: 10.1109/ICSGEA53208.2021.00074.

[61] S. Sun, D. Hu, Z. Zhou, X. Hu, Q. Shao, „Application and research of Big data analysis in commercial Banks," in proc. of *2020 International Conference on Big Data and Social*

Sciences (ICBDSS), Xi'an, China, Aug. 2020, pp. 76-79, doi: 10.1109/ICBDSS51270.2020.00025.

[62] K. Wong, R. Wong, „Big Data Quality Prediction on Banking Applications: Extended Abstract," in proc. of *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, Oct. 2020, pp. 791-792, doi: 10.1109/DSAA49011.2020.00119.

[63] S. Alrumiah, M. Hadwan, „Implementing Big Data Analytics in E-Commerce: Vendor and Customer View," *IEEE Access*, vol. 9, pp. 37281-37286, Mar. 2021, doi: 10.1109/ACCESS.2021.3063615.

[64] H. Gao, „Research on the Development of Retail E-commerce in China from the Perspective of Big Data," in proc. of *2020 International Conference on Big Data Economy and Information Management (BDEIM)*, Zhengzhou, China, Dec. 2020, pp. 87-90, doi: 10.1109/BDEIM52318.2020.00029.

[65] A. Sato, R. Huang, „From Data to Knowledge: A Cognitive Approach to Retail Business Intelligence," in proc. of *2015 IEEE International Conference on Data Science and Data Intensive Systems*, Sydney, NSW, Australia, Dec. 2015, pp. 210-217, doi: 10.1109/DSDIS.2015.106.

[66] R. Alguliyev, R. Aliguliyev. M. Hajirahimova, „Big data integration architectural concepts for oil and gas industry," in proc. of *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, Baku, Azerbaijan, Oct. 2016, pp. 1-5, doi: 10.1109/ICAICT.2016.7991832.

[67] Y. Luo, H. Zhao, B. Xiong, „Research on Air Conditioning Performance Monitoring and Trend Prediction of A320 Aircraft Based on Big Data Analysis," in proc. of *2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, Changsha, China, Oct. 2021, pp. 375-379, doi: 10.1109/ICCASIT53235.2021.9633479.

[68] A. Crespino, A. Corallo, M. Lazoi, D. Barbagallo, A. Appice, D. Malerba, „Anomaly detection in aerospace product manufacturing: Initial remarks," in proc. of *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, Bologna, Italy, pp. 1-4, Sep. 2016, doi: 10.1109/RTSI.2016.7740644.

[69] L. Wang, „Research on Tax Collection and Administration Based on Big Data Analysis," in proc. of *2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Vientiane, Laos, Jan. 2020, pp. 679-682, doi: 10.1109/ICITBS49701.2020.00149.

[70] A. Siddiqa, A. Karim, A. Gani, „Big data storage technologies: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 8, pp. 1040-1070, Aug. 2017, doi: 10.1631/FITEE.1500441.

[71] S. Boubiche, D. Boubiche, A. Bilami, H. Toral-Cruz, „Big Data Challenges and Data Aggregation Strategies in Wireless Sensor Networks," *IEEE Access*, vol. 6, pp. 20558-20571, May 2018, doi: 10.1109/ACCESS.2018.2821445.

[72] I. Flouris, N. Giatrakos, A. Deligiannakis, M. Garofalakis, M. Kamp, M. Mock, „Issues in Complex Event Processing: Status and Prospects in the Big Data Era," *Journal of Systems and Software*, vol. 127, pp. 217-236, May 2017, doi: 10.1016/j.jss.2016.06.011.

[73] F. Gurcan, M. Berigel, „Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges," in proc. of *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, Oct. 2018, pp. 1-6, doi: 10.1109/ISMSIT.2018.8567061.

[74] P. Srinavasa Rao, M. Krishna Prasad, K. Thammi Reddy, „An Efficient Keyword Based Search of Big Data Using Map Reduce," *Journal of Advances in Information Technology*, vol. 8, no. 3, pp. 159-164, Aug.2017, doi: 10.12720/jait.8.3.159-164.

[75] R. Atat, L. Liu, J. Wu, G. Li, C. Ye, Y. Yang, „Big Data Meet Cyber-Physical Systems: A Panoramic Survey," *IEEE Access*, vol. 6, pp. 73603-73636, Nov. 2018, doi: 10.1109/ACCESS.2018.2878681.

[76] J. Wu, S. Guo, J. Li, D. Zeng, „Big Data Meet Green Challenges: Greening Big Data," *IEEE Systems Journal*, vol. 10, no. 3, pp. 873-887, Sep. 2016, doi: 10.1109/JSYST.2016.2550538.

[77] W. Huang, et al., „Mobile Internet Big Data Platform in China Unicom," *Tsinghua Science and Technology*, vol. 19, no. 1, pp. 95-101, Feb. 2014, doi: 10.1109/TST.2014.6733212.

[78] D. Jiang, Y. Wang, Z. Lv, S. Qi, S. Singh, „Big Data Analysis-based Network Behavior Insight of Cellular Networks for Industry 4.0 Applications," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1310-1320, Feb. 2020, doi: 10.1109/TII.2019.2930226.

[79] M. Musolesi, „Big Mobile Data Mining: Good or Evil?," *IEEE Internet Computing*, vol. 18, no. 1, pp. 78-81, Jan.-Feb. 2014, doi: 10.1109/MIC.2014.2.

[80] T. Benhavan, K. Songwatana, „HFC network performance monitoring system using DOCSIS cable modem operation data in a 3 dimensional analysis," in proc. of *The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering (JICTEE)*, Chiang Rai, Thailand, Mar. 2014, pp. 1-5, doi: 10.1109/JICTEE.2014.6804074.

[81] M. Hadi, A. Lawey, T. El-Gorashi, J. Elmirghani, „Big data analytics for wireless and wired network design: A survey," *Computer Networks*, vol. 132, pp. 180-199, Feb. 2018, doi: 10.1016/j.comnet.2018.01.016.

[82] C. I, Y. Liu, S. Han, S. Wang, G. Liu, „On Big Data Analytics for Greener and Softer RAN," *IEEE Access*, vol. 3, pp. 3068-3075, Aug. 2015, doi: 10.1109/ACCESS.2015.2469737.

[83] A. Sahni, D. Marwah, R. Chadha, „Real time monitoring and analysis of available bandwidth in cellular network-using big data analytics," in proc. of *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, Mar. 2015, pp. 1743-1747.

[84] J. Lin, „The Lambda and the Kappa," *IEEE Internet Computing*, vol. 21, no. 5, pp. 60-66, Sep.-Oct. 2017, doi: 10.1109/MIC.2017.3481351.

[85] C. Bisdikian, K. Maruyama, D. Seidman, D. Serpanos, „Cable Access Beyond the Hype: On Residential Broadband Data Services over HFC Networks," *IEEE Communications Magazine*, vol. 34, no. 11, pp. 128-135, Nov. 1996, doi: 10.1109/35.544203.

[86] J. Park, M. Lee, „QoS Provisioning Method for Downstream VoIP Service Flows in HFC Networks," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 448-453, May 2007, doi: 10.1109/TCE.2007.381714.

[87] Cisco, „Cable Modem Termination Systems (CMTS)". Available: https://www.cisco.com/c/en/us/tech/broadband-cable/cable-modem-termination-systems-cmts/index.html [Accessed: June 29, 2022.]

[88] Cisco, „Cisco 1.25 GHz Surge-Gap Flexible Solutions Taps Data Sheet," Mar. 2017. Available: https://www.cisco.com/c/en/us/products/collateral/video/traditional-size-taps/datasheet-c78-733863.html [Accessed: June 29, 2022.]

[89] Cable Television Laboratories, Inc., „Data-Over-Cable Service Interface Specifications (DOCSIS)," Ver. CM-SP-eDOCSIS-I28- 150305.

[90] A. Gaydashenko, S. Ramakrishnan, „A Machine Learning approach to maximizing Broadband Capacity via Dynamic DOCSIS 3.1 Profile Management," in proc. of *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA, Dec. 2019, pp. 341-345, doi: 10.1109/ICMLA.2019.00064.

[91] S. Abedin et al., „A Novel Approach for Profile Optimization in DOCSIS 3.1 Networks Exploiting Traffic Information," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 578-590, doi: 10.1109/TNSM.2019.2901879.

[92] M. Baek, J. Song, J. Jung, „Design and Performance Verification of Time-Domain Self-Interference Estimation Technique for DOCSIS 3.1 System with Full Duplex," in proc. of *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Valencia, Spain, June 2018, pp. 1-4, doi: 10.1109/BMSB.2018.8436658.

[93] M. Baek, J. Song, O. Kwon, J. Jung, „Self-Interference Cancellation in Time-Domain for DOCSIS 3.1 Uplink System With Full Duplex," *IEEE Transactions on Broadcasting*, vol. 65, no. 4, pp. 695-701, Dec. 2019, doi: 10.1109/TBC.2019.2897738.

[94] B. Berscheid, C. Howlett, „Full Duplex DOCSIS: Opportunities and Challenges," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 28-33, Aug. 2019, doi: 10.1109/MCOM.2019.1800851.

[95] W. Coomans, H. Chow, J. Maes, „Introducing Full Duplex in Hybrid Fiber Coaxial Networks," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 74-79, Mar. 2018, doi: 10.1109/MCOMSTD.2018.1700011.

[96] M. Simakovic, Z. Cica, D. Drajic, „Big-Data Platform for Performance Monitoring of Telecom-Service-Provider Networks," *Electronics*, vol. 11, no. 14, an. 2224, July 2022, doi: 10.3390/electronics11142224.

[97] Cisco, „Cisco uBR10012 Universal Broadband Router Hardware Installation Guide," June 2017. Available: https://www.cisco.com/c/en/us/td/docs/cable/cmts/ubr10012/installation/guide/hig.html [Accessed: June 29, 2022.]

[98] Unitymedia GMBH, „Unitymedia Q4 2018 Report," Dec. 2018. Available: https://www.libertyglobal.com/wp-content/uploads/2019/03/Unitymedia-Q4-2018-Report.pdf [Accessed: June 29, 2022.]

[99] „Global OID reference database". Available: http://oidref.com/ [Accessed: June 29, 2022.]

[100] Circitor, „MIB files repository,“ June 2022. Available: https://www.circitor.fr/Mibs/Mibs.php [Accessed: June 29, 2022.]

[101] Cisco, „Advantage Remote PHY“. Available: https://www.cisco.com/c/en/us/solutions/service-provider/industry/cable/advantage-remote-phy.html [Accessed: June 29, 2022.]

[102] A. Cuzzocrea, „Privacy-Preserving Big Data Stream Mining: Opportunities, Challenges, Directions,“ in proc. of *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, USA, Nov. 2017, pp. 992-994, doi: 10.1109/ICDMW.2017.140.

[103] M. Damiani, C. Cuijpers, „Privacy Challenges in Third-Party Location Services,“ in proc. of *2013 IEEE 14th International Conference on Mobile Data Management*, Milan, Italy, June 2013, pp. 63-66, doi: 10.1109/MDM.2013.67.

[104] H. Liu, W. Di, „Application of Differential Privacy in Location Trajectory Big Data,“ in proc. of *2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Vientiane, Laos, Jan. 2020, pp. 569-573, doi: 10.1109/ICITBS49701.2020.00125.

[105] C. Yin, J. Xi, R. Sun, and J. Wang, „Location Privacy Protection Based on Differential Privacy Strategy for Big Data in Industrial Internet of Things,“ *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3628-3636, Aug. 2018, doi: 10.1109/TII.2017.2773646.

[106] S. Wang, R. Sinnott, S. Nepal, „Privacy-protected place of activity mining on big location data,“ in proc. of *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, USA, Dec. 2017, pp. 1101-1108, doi: 10.1109/BigData.2017.8258035.

[107] C. Wu, Y. Guo, „Enhanced user data privacy with pay-by-data model,“ in proc. of *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, Oct. 2013, pp. 53-57, doi: 10.1109/BigData.2013.6691688.

[108] X. Feng, „The Optimization of Privacy Data Management Model In Big Data Era,“ in proc. of *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China, Mar. 2021, pp. 2656-2660, doi: 10.1109/IAEAC50856.2021.9390675.

[109] Y. Canbay, Y. Vural, S. Sagiroglu, „Privacy Preserving Big Data Publishing,“ in proc. of *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, Ankara, Turkey, Dec. 2018, pp. 24-29, doi: 10.1109/IBIGDELFT.2018.8625358.

[110] R. Mahesh, T. Meyyappan, „Anonymization technique through record elimination to preserve privacy of published data,“ in proc. of *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, Salem, India, Feb. 2013, pp. 328-332, doi: 10.1109/ICPRIME.2013.6496495.

[111] K. M. Shrivastva, M. A. Rizvi, S. Singh, „Big Data Privacy Based on Differential Privacy a Hope for Big Data,“ in proc. of *2014 International Conference on Computational Intelligence and Communication Networks*, Bhopal, India, Nov. 2014, pp. 776-781, doi: 10.1109/CICN.2014.167.

[112] Cloudflare, „What is data privacy?“. Available: https://www.cloudflare.com/learning/privacy/what-is-data-privacy/ [Accessed: June 29, 2022.]

[113] „Complete Guide to GDPR Compliance“. Available: https://gdpr.eu/ [Accessed: June 29, 2022.]

[114] „California Consumer Privacy Act“. Available: https://oag.ca.gov/privacy/ccpa [Accessed: June 29, 2022.]

[115] China Briefing Team, „The PRC Personal Information Protection Law,“ Aug. 2021. Available: https://www.china-briefing.com/news/the-prc-personal-information-protection-law-final-a-full-translation/ [Accessed: June 29, 2022.]

[116] „Health Insurance Portability and Accountability Act of 1996“. Available: https://www.cdc.gov/phlp/publications/topic/hipaa.html [Accessed: June 29, 2022.]

[117] E. Devaux, „Which data protection methods do you need to guarantee privacy?,“ Sep. 2020. Available: https://www.statice.ai/post/data-protection-techniques-need-to-guarantee-privacy [Accessed: June 29, 2022.]

[118] M. Simakovic, Z. Cica, D. Drajic, „Location Privacy Improvements in Telecommunication Data Management Systems,“ in proc. of *IcETRAN 2022*, Novi Pazar, Serbia, June 2022, pp. 385-388.

[119] M. Simakovic, Z. Cica, „Detection and Localization of Failures in Hybrid Fiber–Coaxial Network Using Big Data Platform,“ *Electronics*, vol. 10*,* no. 23, an. 2906, Nov. 2021, doi: 10.3390/electronics10232906.

[120] S. Lee, K. Levanti, H. Kim, „Network Monitoring: Present and Future,“ *Computer Networks*, vol. 65, no. 2, pp. 84-98, June 2014, doi: 10.1016/j.comnet.2014.03.007.

[121] L. Shu, Z. Yu, Z. Wan, J. Zhang, S. Hu, K. Xu, „Dual-Stage Soft Failure Detection and Identification for Low-Margin Elastic Optical Network by Exploiting Digital Spectrum Information,“ *Journal of Lightwave Technology*, vol. 38, no. 9, pp. 2669-2679, May 2020, doi: 10.1109/JLT.2019.2947562.

[122] H. Lun et al., „Soft failure identification in optical networks based on convolutional neural network,“ in proc. of *45th European Conference on Optical Communication (ECOC 2019)*, Dublin, Ireland, Sep. 2019, pp. 1-3, doi: 10.1049/cp.2019.1138.

[123] S. Shahkarami, F. Musumeci, F. Cugini, and M. Tornatore, „Machine-Learning-Based Soft-Failure Detection and Identification in Optical Networks,“ in proc. of *2018 Optical Fiber Communications Conference and Exposition (OFC)*, Angers, France, Mar. 2018, pp. 1-3, doi: 10.1364/OFC.2018.M3A.5.

[124] K. Mayer et al., „Soft Failure Localization Using Machine Learning with SDN-based Network-wide Telemetry,“ in proc. of *2020 European Conference on Optical Communications (ECOC)*, Brussels, Belgium, Dec. 2020, pp. 1-4, doi: 10.1109/ECOC48923.2020.9333313.

[125] T. Panayiotou, S. Chatzis, G. Ellinas, „Leveraging statistical machine learning to address failure localization in optical networks,“ *Journal of Optical Communications and Networking*, vol. 10, no. 3, pp. 162-173, Mar. 2018, doi: 10.1364/JOCN.10.000162.

[126] Z. Li, et al., „Demonstration of Fault Localization in Optical Networks Based on Knowledge Graph and Graph Neural Network," in proc. of *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, San Diego, CA, USA, Mar. 2020, pp. 1-3, doi: 10.1364/OFC.2020.Th1F.5.

[127] W. Gray, A. Tsokanos, R. Kirner, „Multi-Link Failure Effects on MPLS Resilient Fast-Reroute Network Architectures," in proc. of *2021 IEEE 24th International Symposium on Real-Time Distributed Computing (ISORC)*, Daegu, South Korea, June 2021, pp. 29-33, doi: 10.1109/ISORC52013.2021.00015.

[128] A. Dusia, A. Sethi, „Recent Advances in Fault Localization in Computer Networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 3030-3051, doi: 10.1109/COMST.2016.2570599.

[129] Y. Qiao, X. Qiu, L. Cheng, L. Meng, „A Methodology Used to Optimize Probe Selection for Fault Localization," in proc. of *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, Miami, FL, USA, Oct. 2010, pp. 1-5, doi: 10.1109/GLOCOM.2010.5684146.

[130] T. Kimura, A. Watanabe, T. Toyono, K. Ishibashi, „Proactive failure detection learning generation patterns of large-scale network logs," in proc. of *2015 11th International Conference on Network and Service Management (CNSM)*, Barcelona, Spain, Nov. 2015, pp. 8-14, doi: 10.1109/CNSM.2015.7367332.

[131] C. Tan, et al., „Netbouncer: active device and link failure localization in data center networks," in proc. of *USENIX Conference on Networked Systems Design and Implementation*, Boston, MA, USA, Feb. 2019, pp. 599-613.

[132] P. Zych, „Network failure detection based on correlation data analysis," *AEU - International Journal of Electronics and Communications*, vol. 77, pp. 27-35, July 2017, doi: 10.1016/j.aeue.2017.04.014.

[133] M. Ab-Rahman, N. Boon Chuan, M. Safnal, K. Jumari, „The overview of fiber fault localization technology in TDM-PON network," in proc. of *2008 International Conference on Electronic Design*, Penang, Malaysia, Dec. 2008, pp. 1-8, doi: 10.1109/ICED.2008.4786674.

[134] E. Gibellini, C. Righetti, „Unsupervised Learning for Detection of Leakage from the HFC Network," in proc. of *2018 ITU Kaleidoscope: Machine Learning for a 5G Future (ITU K)*, Santa Fe, Argentina, Nov. 2018, pp. 1-8, doi: 10.23919/ITU-WT.2018.8598128.

[135] Apache Software Foundation, „Spark MLlib," June 2016. Available: https://spark.apache.org/mllib/ [Accessed: June 29, 2022.]

[136] „PyTorch". Available: https://pytorch.org/ [Accessed: June 29, 2022.]

[137] „TensorFlow". Available: https://www.tensorflow.org/ [Accessed: June 29, 2022.]

[138] B.J. Cheng, M. Jamil, R. Ambar, M. Wahab, A. Ma'radzi, „Elderly Care Monitoring System with IoT Application," in book Recent Advances in Intelligent Information Systems and Applied Mathematics, pp. 525-537, 2020.

[139] L. Zhang, L. Zhao, Z. Wang, J. Liu, „WiFi Networks in Metropolises: From Access Point and User Perspectives," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 42-48, May 2017, doi: 10.1109/MCOM.2017.1600262.

[140] S. Atitallah, M. Driss, W. Boulila, H. Ghezala, „Leveraging Deep Learning and IoT big data analytics to support the smart cities development: Review and future directions," *Computer Science Review*, vol. 38, pp. 100303, Nov. 2020, doi: 10.1016/j.cosrev.2020.100303.

[141] J.D. Chimeh, „Compelling Services for 5G Creation," in proc. of *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Seoul, South Korea, Apr. 2020, pp. 1-6, doi: 10.1109/WCNCW48565.2020.9124735.

[142] E. Ahmed et al., „The role of big data analytics in Internet of Things," *Computer Networks*, vol. 129, no. 2, pp. 459-471, Dec. 2017, doi: 10.1016/j.comnet.2017.06.013.

[143] A. Akbar et al., „Real-Time Probabilistic Data Fusion for Large-Scale IoT Applications," *IEEE Access*, vol. 6, pp. 10015-10027, Feb. 2018, doi: 10.1109/ACCESS.2018.2804623.

[144] F. Wang, M. Li, Y. Mei, W. Li, „Time Series Data Mining: A Case Study With Big Data Analytics Approach," *IEEE Access*, vol. 8, pp. 14322-14328, Jan. 2020, doi: 10.1109/ACCESS.2020.2966553.

[145] D. Mourtzis, E. Vlachou, N. Milas, „Industrial Big Data as a Result of IoT Adoption in Manufacturing," *Procedia CIRP*, vol. 55, pp. 290-295, 2016, doi: 10.1016/j.procir.2016.07.038.

[146] N. Misra, Y. Dixit, A. Al-Mallahi, M. Bhullar, R. Upadhyay, A. Martynenko, „IoT, big data and artificial intelligence in agriculture and food industry," *IEEE Internet of Things Journal*, vol. 9, no. 9, pp. 6305-6324, May 2020, doi: 10.1109/JIOT.2020.2998584.

[147] A. Ahad et al., „Technologies Trend towards 5G Network for Smart Health-Care Using IoT: A Review," *Sensors*, vol. 20, no. 14, pp. 4047, July 2020, doi: 10.3390/s20144047.

[148] S. Kumar, P. Tiwari, M. Zymbler, „Internet of Things is a revolutionary approach for future technology enhancement: a review," *Journal of Big Data*, vol. 6, no. 1, pp. 1-21, Dec. 2019, doi: 10.1186/s40537-019-0268-2.

[149] M. Simakovic, Z. Cica, D. Drajic, „Introducing IoT to Big Data Platform for Network Performance Monitoring," in proc. of *IcETRAN 2022*, Novi Pazar, Serbia, June 2022, pp. 385-388.

[150] Shandong Renke Control Technology Co.,Ltd., „Professional AQI Sensor for Indoor and Outdoor". Available: https://www.renkeer.com/product/aqi-sensor/ [Accessed: June 29, 2022.]

[151] Acrel, „Wireless Power Meter, ADW300". Available: https://acrel-power.com/3-iot-wireless-power-meter.html [Accessed: June 29, 2022.]

[152] Z. Che Soh, M. Shafie, M. Shafie, S. Sulaiman, M. Ibrahim, S. Abdullah, „IoT Water Consumption Monitoring & Alert System," in proc. of *2018 International Conference on Electrical Engineering and Informatics (ICELTICs)*, Banda Aceh, Indonesia, Sep. 2018, pp. 168-172, doi: 10.1109/ICELTICS.2018.8548930.

[153] ekoNET air quality device. Available online: https://ekonet.solutions/air-monitoring/ [Accessed June 29, 2022.]

[154] HORIBA, „Ambient air quality monitoring station". Available online: https://www.horiba.com/at/process-environmental/products/system-engineering/air-quality-monitoring-system/ [Accessed June 29, 2022.]

[155] I.Vajs, D. Drajic, N. Gligoric, I. Radovanovic, I. Popovic, „Developing Relative Humidity and Temperature Corrections for Low-Cost Sensors Using Machine Learning,“ *Sensors*, vol. 21, no. 10, an. 3338, May 2021, doi: 10.3390/s21103338.

[156] I. Vajs, D. Drajic, Z. Cica, „COVID-19 Lockdown in Belgrade: Air Pollution Impact and Evaluation of the Neural Network Model for the Correction of Low-Cost Sensors' Measurements,“ *Applied Sciences*, vol. 11, no. 22, an. 10563, Nov. 2021, doi: 10.3390/app112210563.

[157] SolarWinds Worldwide, LLC „SolarWinds Network Performance Monitor“. Available: https://www.solarwinds.com/network-performance-monitor [Accessed: June 29, 2022.]

[158] „ManageEngine OpManager“. Available: https://www.manageengine.com/network-monitoring/ [Accessed: June 29, 2022.]

СПИСАК СКРАЋЕНИЦА

ADSL	Asymmetric Digital Subscriber Line
AMP	Amplifier
AP	Access Point
API	Application Programming Interface
AWS	Amazon Web Services
BER	Bit Error Rate
CATV	Cable Television
CCPA	California Consumer Privacy Act
CDR	Call Detail Record
CM	Cable Modem
CMTS	Cable Modem Termination System
CNR	Carrier-to-Noise Ratio
CPE	Customer-Premises Equipment
CPU	Central Processing Unit
CSV	Comma-Separated Values
DHCP	Dynamic Host Configuration Protocol
DMPF	Detection of Mass PPP Failures
DOCSIS	Data Over Cable Service Interface Specifications
DSL	Digital Subscriber Line
DSLAM	Digital-Subscriber-Line-Access-Multiplexer
FTP	File Transfer Protocol
FTTH	Fiber to the Home
FaaS	Function as a Service
GDPR	General Data Protection Regulation
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
HFC	Hybrid Fiber-Coaxial
HIPAA	Health Insurance Portability and Accountability Act
HTTP	Hypertext Transfer Protocol
IBM	International Business Machines
IP	Internet Protocol
IPDR	Internet Protocol Detail Record
IT	Information technology
ITU	International Telecommunication Union
IaaS	Infrastructure as a Service
IoT	Internet of Things
JAQL	Java Query Language
JDBC	Java Database Connectivity

JSON	JavaScript Object Notation
LTE	Long Term Evolution
M2M	Machine to Machine
MAC	Media Access Control
MIB	Management Information Base
ML	Machine Learning
MPLS	Multiprotocol Label Switching
MQTT	Message Queuing Telemetry Transport
MRv1	MapReduce version 1
MRv2	MapReduce version 2
NOC	Network Operations Center
NMS	Network Management System
NPM	Network Performance Monitor
NoSQL	Not only SQL
OFDM	Orthogonal Frequency-Division Multiplexing
OID	Object Identifier
ON	Optical Node
ONU	Optical Network Unit
OpenTSDB	Open Time Series Database
OTDR	Optical Time-Domain Reflectometer
PIPL	Personal Information Protection Law
PM	Particle Matters
PNM	Proactive Network Maintenance
PON	Passive Optical Network
PPP	Point-to-Point Protocol
PSTN	Public Switched Telephone Network
PaaS	Platform as a Service
QoE	Quality of Experience
QoO	Quality of Operation
QoS	Quality of Service
RAM	Random Access Memory
REST API	Representational State Transfer Application Programming Interface
RSSI	Received Signal Strength Indication
SGSN	Serving GPRS Support Node
SNMP	Simple Network Management Protocol
SNR	Signal to Noise ratio
SON	Self-Organizing Networks
SPOF	Single Point of Failure
SQL	Structured Query Language
SRE	Switchable Reflective Element
STB	Set-Top Box
SaaS	Software as a Service
TDM-PON	Time Division Multiplexing Passive Optical Network
TELNET	Teletype Network Protocol
UID	Unique Identifier
UNIX	Uniplexed Information Computing System

UPS	Uninterruptible Power Supply
VoC	Volatile Organic Compounds
VoIP	Voice over Internet Protocol
WiFi	Wireless Fidelity
WSN	Wireless Sensor Networks
XDR	Extended Detection and Response
XML	Extensible Markup Language
YARN	Yet Another Resource Negotiator
xDSL	X Digital Subscriber Line
ДЛОКМ	Детекција и Локализација Отказа у Кабловским Мрежама
ПНПМБД	Платформа за Надгледање Перформанси Мреже заснована на Биг Дата технологији
ТВ	Телевизија

СПИСАК СЛИКА

СПИСАК ТАБЕЛА

Биографија аутора

Милан Симаковић је рођен 3.11.1990. у Јеревану, Република Јерменија. Основну школу „Ђорђе Јовановић“ завршио је у Селевцу, а средњу машинско-електротехничку школу „Гоша“ завршио је у Смедеревској Паланци, обе са одличним успехом. Електротехнички факултет у Београду је уписао 2009. године на којем је и дипломирао 2013. године на модулу за Телекомуникације. Мастер академске студије је уписао 2013. године на Електротехничком факултету у Београду, модул системско инжењерство и радио комуникације. Диплому мастер инжењера електротехнике и рачунарства је стекао 2014. године. Докторске академске студије је уписао на Електротехничком факултету у Београду, 2015. године - модул Телекомуникације. Област истраживања током докторских студија и рада на дисертацији су биле биг дата технологије и примена истих у телекомуникацијама, односно у телекомуникационим мрежама. Као резултат рада на овим истраживањима, објавио је два рада у међународним часописима са SCI листе, четири рада на међународним конференцијама и један рад на домаћој конференцији. Милан Симаковић је тренутно запослен у компанији „Grid Dynamics“, на позицији биг дата инжењера.

образац изјаве о ауторству

Изјава о ауторству

Име и презиме аутора Милан Симаковић

Број индекса 2015/5011

Изјављујем

да је докторска дисертација под насловом

Систем за надгледање перформанси мреже кабловског оператора заснован на технологији великих података

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 11.07.2022.

образац изјаве о истоветности штампане и електронске верзије докторског рада

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Милан Симаковић

Број индекса 2015/5011

Студијски програм Електротехника и рачунарство

Наслов рада Систем за надгледање перформанси мреже кабловског оператора заснован на технологији великих података

Ментор проф. др Зоран Чича, проф. др Дејан Драјић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 11.07.2022.

образац изјаве о коришћењу

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Систем за надгледање перформанси мреже кабловског оператора заснован на технологији великих података

која је моје ауторско дело

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 11.07.2022.

1. **Ауторство.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.