УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ“ – ШТИП

ФАКУЛТЕТ ЗА ИНФОРМАТИКА

ШТИП

Дијана Лапевска

АНАЛИЗА НА АКТИВНОСТИТЕ НА MOODLE БАЗА НА ПОДАТОЦИ, ПРЕД И ПОСЛЕ ПОЈАВАТА НА ПАНДЕМИЈАТА COVID-19

-МАГИСТЕРСКИ ТРУД-

Штип, ноември 2021

Претседател:	**Проф. д-р Татјана Атанасова Пачемска** **Факултет за информатика** **Универзитет „Гоце Делчев“ Штип**
Ментор:	**Проф. д-р Зоран Здравев** **Факултет за информатика** **Универзитет „Гоце Делчев“ Штип**
Член:	**Доц. д-р Александра Стојанова** **Факултет за информатика** **Универзитет „Гоце Делчев“ Штип**

Датум на одбрана: _____

Датум на промоција:_____

Рецензирани и објавени трудови:

1. Dijana Lapevska, Aleksandar Velinov, Zoran Zdravev (2021).
   ANALYSIS OF MOODLE ACTIVITES BEFORE AND AFTER THE COVID-19 PANDEMIC – CASE STUDY AT GOCE DELCHEV UNIVERSITY. *BJAMI, year2021, volume IV, number 1.*

Краток извадок

Образовните системи ширум светот се соочија со предизвик без преседан, при што беше неопходно да се обезбеди образование од далечина преку мешавина на технологии, со цел да се обезбеди континуитет на студирање и учење базирано на наставна програма за сите. Тоа предизвика миграција на досегашниот начин на едукација, односно наставата со физичко присуство веќе беше заменета со учење на далечина преку интернет. Затворањето на училиштата беше наложено како дел од препораките за јавно здравје за да се спречи ширењето на Covid-19 од февруари 2020 година во повеќето земји.

Учењето на далечина, вклучително и настава и учење преку интернет, се изучува и применува со децении, но со појавата на пандемијата тоа стана единствен начин за продолжување на образовниот процес. Бројни истражувачки студии, теории, модели, стандарди и критериуми за евалуација се фокусираат на квалитетно учење преку интернет, настава преку интернет и дизајн на курсеви преку интернет.

Во овој контекст, во 2020 година, студискиот процес на Универзитетот „Гоце Делчев“ кој се одвиваше со физичко присуство беше променет со учење од далечина поради новосоздадената ситуација предизвикана од пандемијата Covid-19. Примарната цел на ова истражување е да се анализира бројот на активности на корисниците на Moodle платформата пред и по пандемијата. Системот е-учење Moodle се користи скоро 10 години. За таа цел е извршена анализа на податоците од базата на податоци на Moodle користејќи алатки за големи податоци. Според добиените резултати, вкупниот број активности во 2020 година е зголемен за три пати во споредба со истиот период во 2019 година. Од истражувањето исто така беа добиени резултати за поединечните активности на наставничкиот кадар и студентите, односно резултати за одредени модули кои беа анализирани. Истите тие резултати покажуваат дека има разлика во бројот на активностите на корисниците на Moodle платформата.

Клучни зборови: *големи податоци, Moodle, систем за електронско учење, Covid-19*

Abstract

The closure of the schools was ordered as part of public health efforts to prevent the spread of Covid-19 from February to May 2020 in most countries. Education systems around the world have faced an unprecedented challenge, with the need to provide distance education through a mix of technologies in order to ensure continuity of study and curriculum-based learning for all. This caused a migration to the current way of education, ie physical presence teaching has already been replaced by distance learning through the Internet.

Online education, including online teaching and learning, has been studied and practiced for decades, but with the onset of the pandemic it has become the only way for the educational process to unfold. Numerous research studies, theories, models, standards, and evaluation criteria focus on quality online learning, online teaching, and online course design.

In this context, in 2020, the study process at the University "Goce Delchev" in Stip was moved to distance learning due to the newly created situation caused by the pandemic Covid-19. The primary purpose of this study is to analyze the number of activities of users of the Moodle platform before and after the pandemic. The Moodle learning system has been in use for almost 10 years. For this purpose, the data from the Moodle database was analyzed using big data tools. According to the obtained results, the total number of activities in 2020 has increased by 3 times compared to the same period in 2019. The research also obtained results for the individual activities of the teaching staff and students, ie results for certain modules that were analyzed. The same results show whether there is a difference in the number of activities of users of the Moodle platform.

Keywords: *Big data, Moodle, e-learning system, Covid-19*

Содржина

1. Вовед 10
2. Цел на истражувањето 12
3. Големи податоци 14
3.1. Карактеристики на големи податоци 14
3.2. Примена на големите податоци и нивна обработка и анализа 16
4. Анализа на големи податоци со алгоритми за машинско учење 17
4.1. Машинско учење 17
4.1.1. Учење во длабочина 18
4.1.2. Надгледувано учење 19
4.1.2.1. Регресија 19
4.1.3. Учење без надзор 23
5. Алатки за анализа на големи податоци 28
5.1. Hadoop 28
5.1.1. Архитектура на Hadoop 29
5.2. Ниво на обработка на податоци 34
5.2.1. MapReduce 34
5.2.2. Hadoop YARN 35
5.3. Дистрибуции на Hadoop 36
5.3.1. Cloudera 36
5.3.2. Hortonworks 37
5.4. Ниво за управување со податоци 38
5.4.1. Apache Ambari 38
5.4.2. Hue 39
5.4.3. ZooKeeper 40
5.4.4. Avro 40
5.4.5. Oozie 41
5.5. Ниво на пристап до податоци 41
5.5.1. Hive 41
5.5.2. Hcatalog 42
5.5.3. Apache Pig 43
5.5.4. Sqoop 43

5.5.5. JAQL 45
5.5.6. Flume 46
5.5.7. Chukwa 46
6. Едукативно податочно рударење 47
7. Истражување 49
7.1. Систем за управување со учење (LMS) 49
7.2. Moodle 49
7.2.1. Moodle за време на Covid-19 50
7.2.2. Кориснички улоги на Moodle 52
7.2.3. Активности на Moodle 53
7.4. Фази на истражувачката работа 56
7.4.1. Собирање на податоци 57
7.4.2. Препроцесирање и трансформација на собраните податоци и креирање на податочен сет погоден за понатамошна обработка 57
7.4.3. Обработка на добиените податоци 58
7.4.4. Евалуација и анализа на добиените резултати 61
8. Заклучок 79
9. Користена литература 82

Слики

Слика 1. Водопаден модел за текот на истражувачката работа 13
Слика 2. Видови на големи податоци 15
Слика 3. Процес на машинско учење 17
Слика 4. Пример за линеарна регресија 20
Слика 5. Дрво на одлучување 22
Слика 6. Различни групи на исти податоци 25
Слика 7. Компоненти на Hadoop 30
Слика 8. Компоненти на HDFS 32
Слика 9. Архитектура на Hadoop имплементирана од Hortonworks 37
Слика 10. Архитектура на Apache Ambari 39
Слика 11. Кориснички интерфејс на Hue 40
Слика 12. Use Case дијаграм за Table Storage layer и SQL пребарувач 42
Слика 13. Use Case дијаграм за користење на Sqoop за преместување податоци 45
Слика 14. Работна околина на Ambari 59
Слика 15. Графикон на тек на сите активности за анализа на податоците 60
Слика 16. Разлика помеѓу активностите на Moodle во периодот пред пандемијата и за време на пандемијата 62
Слика 17. Разлика во однос на вкупната годишна активност на корисниците на Moodle 63
Слика 18. HiveQL упит за добивање на резултати на поединечни модули 64
Слика 19. Hive делот од Hadoop со интерпретирани кодови за селекција на корисниците 67
Слика 20. Блок дијаграм за добивање на сите активности на корисниците 67
Слика 21. Корисници на Moodle 68
Слика 22. Вкупни активности на наставнички кадар и студенти 69
Слика 23. Бројот на активности на корисниците во модулот форум 74
Слика 24. Бројот на активности на корисниците во модулот квиз 74
Слика 25. Бројот на активности на корисниците во модулот задачи 75
Слика 26. Бројот на активности на корисниците во модулот избор 75
Слика 27. Бројот на активности на корисниците во модулот книга 76
Слика 28. Бројот на активности на корисниците во модулот разговор 76
Слика 29. Бројот на активности на корисниците во модулот речник 77

Табели

Табела 1. Избрани табели за обработка и подготовка на податочниот сет ... 58
Табела 2. Разлика во активностите на Moodle за различни компоненти на годишно ниво ... 65
Табела 3. Вкупните активности на наставнички кадар и студенти ... 68
Табела 4. Вкупни активности на студентите ... 72
Табела 5. Вкупни активности на професорите ... 72
Табела 6. Резултати од активностите на корисниците за избраните модули 73

1. Вовед

Денешното општество е сè повеќе дигитализирано и меѓусебно поврзано. Дигиталната технологија го менува секој аспект од нашите животи. Тоа носи големи трансформации во општеството кои вклучуваат промени во начинот на пристап до услугите, интеракција со други луѓе, добивање и споделување информации и правење промени во природата и организацијата на работата. Дигиталниот свет постепено навлезе во областа на образованието, а технологијата сè повеќе се користи во високото образование за стекнување на знаења и вештини преку нови иновативни методи. Новонастаната ситуација која започна да владее насекаде низ светот поради пандемијата предизвикана од Covid-19, предизвика актуализирање на дигитализацијата. Образовниот систем драстично се промени, се спроведе посебен начин на е-учење, односно учење на далечина и дигитални платформи. Брзиот раст и популарноста на интернетот, едукацијата од далечина и пандемската криза направија и онлајн образованието исто така да расте брзо. Со порастот на дигитализацијата се бележи и значителен пораст на податоците кои ги содржи истата, што всушност ја претставува и клучната точка која е цел на ова истражување.

Предмет на анализа во овој магистерски труд се податоците добиени од Moodle базата на податоци на Универзитетот „Гоце Делчев“ – Штип. Бидејќи текот на наставата во 2020 година се изведуваше преку учење на далечина преку интернет, тоа придонесе за зголемување на електронските активности на професорите и студентите. Во тој контекст беа поставени целите на ова истражување, а тоа се:

- Дали има зголемување на активностите во Moodle базата на податоци во 2020 година?
- Доколку постои зголемување на активностите, колкава е разликата помеѓу 2019 и 2020 година?
- Колку изнесува разликата на месечно и годишно ниво на активностите?

- Дали постои зголемување на активностите поединечно кај професорите и студентите?
- Кај кои модули се забележува зголемување на активностите?

Кога станува збор за анализа на големи податоци, неопходно е да се споменат техниките и алатките за машинско учење кои претставуваат главна компонента во процесот на обработка на големи податоци. Во *Поглавје 4* се претставени видовите на машинско учење и дел од алгоритмите кои се применуваат при машинско учење.

Поради големината и комплексноста на големите податоци, анализата на собраните податоци не може да се прави со традиционалните алатки и технологии, па поради тоа дошло до појава на нови решенија и креирање на дополнителни алатки кои овозможуваат процесирање на ваквите големи податоци. Дел од алатките и нивниот начин на работа се презентирани во *Поглавје 5.*

Во *Поглавје 6* е објаснето кои техники се применуваат за извлекување на суштински информации од огромните волуменски податоци во образовните системи.

Во *Поглавје 7* се презентирани фазите од истражувачката работа, користени алатки, дел од имплементираните програмски кодови потребни за анализа на големите податоци и добиените резултати. Чекорите во фазите од истражувањето се претставени преку графикони и табели. Исто така, претставен е и текот на анализата на добиените податоци, како и исходот од секој од реализираните чекори.

Во *Поглавје 8* се наведени финалните сознанија и се претставени добиените резултати. Направена е паралела помеѓу добиените резултати, во однос на бројот на активности на Moodle системот за е-учење пред и за време на пандемијата. Во завршниот дел од магистерската теза е дадено размислување врз основа на добиените резултати, за тоа како истите влијаат во образовниот систем и дали постои разлика во текот на едукативниот процес помеѓу наведениот период.

2. Цели на истражувањето и тек на истражувачката работа

Основна цел на оваа истражувачка работа е анализа на големи податоци добиени од системот за електронско учење – Moodle на Универзитетот „Гоце Делчев“ – Штип, односно обработка на податоците во период од 2019 година, пред почетокот на пандемијата и 2020 година, со појавата на пандемијата. Податоците се добиени од MySQL базата за податоци. Бидејќи станува збор за период во кој образовниот процес изврши миграција и се одвиваше само електронски, бројот на активностите на корисниците беше зголемен, па така станува збор за огромна количина на податоци. Поради тоа за анализа на податоците беа потребни посебни алатки за обработка на големи податоци.

Примарната идеја на истражувачката работа се состои во добивање на конкретни резултати, кои ќе ја прикажат генералната слика за употребата на системот за електронско учење во наведениот период, односно добивање на статистички податоци за бројот на активности на системот за е-учење во периодот на пандемија и пред него. По утврдување на добиените резултати, целта е да се добијат конкретни информации за бројот на реализираните активности на одредени модули на Moodle системот за електронско учење.

Примарната цел на истражувањето е:

- Селекција на потребните податоци од базата на податоци, со цел создавање на податочен сет за понатамошната анализа;
- Обработка на добиените податоци и нивно претставување во формат погоден за понатамошна анализа;
- Евалуација и анализа на добиените резултати;
- Донесување на заклучоци врз основа на добиените резултати.

Слика 1. Водопаден модел за текот на истражувачката работа
Figure 1. Waterfall model for the course of research work

На *слика 1* е прикажан водопаден модел за чекорите кои се преземени за реализација на истражувачката работа. Како што може да се забележи, најпрво се врши собирање на податоците од базата на податоци на Moodle, а потоа од истите тие податоци се креира податочен сет којшто е потребен за обработка за да се добијат потребните резултати. По добивањето на резултатите од обработката на податочниот сет се врши евалуација и нивно интерпретирање преку графикони и дијаграми, за на крај да се дојде до потребниот заклучок, односно што е и примарната цел на оваа истражувачка работа.

3. Големи податоци

Бидејќи информатичката технологија во контекст на иновациите се развива под влијание на современиот начин на живот, собирањето дигитални податоци расте експоненцијално. Денес постои огромна количина на податоци што се генерира секојдневно во производството, бизнисот, науката и нашиот личен живот. Правилната обработка на податоци може да открие ново знаење за пазарот, општеството и околината и да овозможи навремено да се реагира на новите можности и промени. Растот на обемот на податоци во дигиталниот свет бара подобра компјутерска инфраструктура за нивна обработка. Конвенционалните технологии за обработка на податоци, како што се бази на податоци и складишта на податоци, стануваат несоодветни за количината на податоци со која се среќаваат секојдневно.

Овој нов предизвик предизвика создавање на специфични технологии за обработка на големи податоци [1]. Големите податоци се податоци што го надминуваат капацитетот за обработка на конвенционалните системи за бази на податоци. Количината на податоци е премногу голема, се движат премногу брзо или не се вклопуваат во структурата на архитектурата на базата на податоци. За да можат таквите податоци да станат корисни, потребна е модерна технологија за обработка. Големите податоци, поради нивната сложеност, бараат нови можности, алатки и модели за управување со информации и нивни надворешни и внатрешни текови. Трансформацијата на голема количина на податоци во стратешки ресурси е предуслов за задоволување на потребите на идните корисници. Од друга страна, предизвиците поставени од големата количина на податоци бараат промена во бизнис моделите и човечките ресурси [2].

3.1. Карактеристики на големи податоци

Заеднички елементи [3] на големите податоци се:

- Волумен - што карактеризира голема количина на податоци, кои се собрани, обработени за да ги направат податоците погодни за анализа;

• Брзина - што го карактеризира континуираното собирање на големи количини на податоци, во реално време;

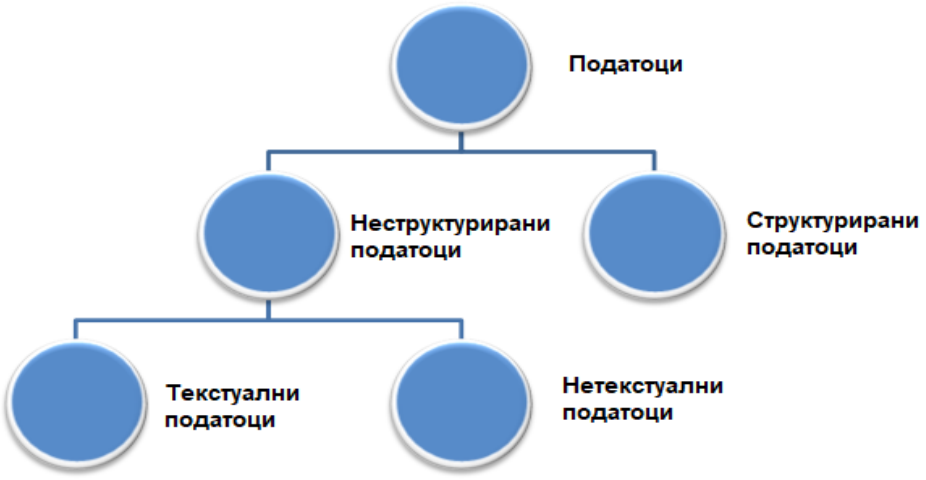
• Разновидност - која ја карактеризира разновидноста на формите и изворите на податоци.

Постојат уште неколку варијации на оваа поделба на карактеристики на големи податоци, но најраспространети и најпознати, се оние кои се дадени погоре. Корените на големите податоци беа поставени пред петнаесет години и постојат два вида податоци:

• структурирани податоци и

• неструктурирани податоци.

Структурираните податоци најчесто се собираат во контролирани случаи, како на пример во компании и организации. Неструктурираните податоци доаѓаат од неорганизирани извори. Исто така, постојат податоци што се полуструктурирани. Полуструктурирани податоци се податоците што не се во согласност со податочниот модел, но имаат одредена структура. Тоа се податоците што не живеат во рационална база на податоци, но имаат некои организациски својства што го олеснуваат анализирањето.



Слика 2. Видови на големи податоци
Figure 2. Types of big data

Сликата погоре ја покажува концептуалната поделба на податоците, но во овој случај за целите на оваа магистерска теза, користени се структурирани податоци. Покрај структурираните податоци може да се види дека има и неструктурирани податоци и тие претставуваат графикони или слики (фотографии, илустрации, рендгенски зраци итн.)

3.2. Примена на големите податоци и нивна обработка и анализа

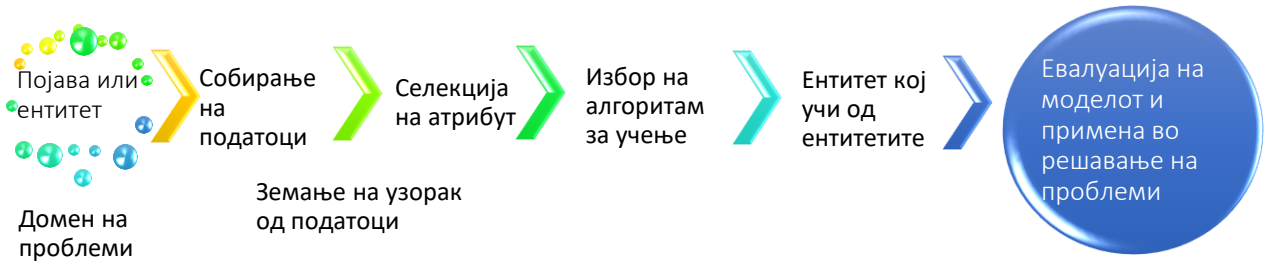
Технологијата за големи податоци помага во ситуации кога традиционалните методи не се во можност да го сторат тоа. Ако количината на неструктурирани податоци што треба да се анализираат расте и ако анализата треба да се направи во реално време, сигурно постои област каде што решение се технологиите за обработка на големи податоци. Постојат различни алатки за обработка на податоци, но деловните проблеми не се доволно подготвени за да ги користат, па затоа е неопходно деловниот проблем да се подели на помали делови. Информатичката технологија може да се користи за пронаоѓање на информативни податоци во рамките на голема количина на податоци. Секоја година расте бројот на начини да се пресметаат сличностите преку компјутер. Долгорочната анализа на сетот на податоци ќе даде резултати, но резултатите не гарантираат генерализација надвор од набљудуваниот сет. Неопходно е да се најде начин да се избегне фокусирање во рамките на само еден сет, за кој има процеси, алгоритми и методи за процена. Дополнително, пред да се донесат брзи заклучоци, неопходно е да се идентификуваат можните скриени фактори.

Со оглед на својата големина, големите податоци зафаќаат голем простор, но со добро ракување со нив, односно со анализа и обработка може да дојдеме до голем број резултати и одговори, кои се потребни. Анализата на податоци се користи во различни академски дисциплини како што се економија, статистика, психологија, социологија и природни и општествени науки [4]. Постојат различни модели со кои може да се анализираат податоците, а за секој сет на податоци за која може да се врши анализа може да се примени различен модел.

4. Анализа на големи податоци со алгоритми за машинско учење

4.1. Машинско учење

Машинско учење [5] претставува гранка на вештачката интелигенција која се занимава со техники и методи кои овозможуваат компјутерите и другите машини да развиваат интелигентни апликации кои се способни да учат, без притоа да бидат експлицитно програмирани за таа намена. Како област на вештачката интелигенција машинското учење е збир на парадигми, алгоритми, теоретски резултати и апликации од различни области на вештачката интелигенција и еволутивни модели, но и други области од математиката.



Слика 3. Процес на машинско учење
Figure 3. The process of machine learning

Машинското учење се заснова на принципот на користење на генерички алгоритми на големи сетови на податоци, што резултира со предвидување одредена вредност или создавање логика за корелација помеѓу дадените податоци. Односно, одреден сет на податоци се вметнува во генерички алгоритам, кој потоа, без да напише специфичен код за даден проблем, ја гледа логиката меѓу внесените информации и може да се примени на неколку различни множества податоци. Се користи при препознавање на лице, препознавање објекти на слики или видеоклипови (на пр. аномалии на Х-зраци), потоа автономно возење автомобил, игри на табла, компјутерски игри, во квизови, класификација на текст, превод, анализа на социјалните медиуми, препознавање говор итн.

Како и за секој метод, постојат различни начини за обука на алгоритми за машинско учење, секој со свои предности и недостатоци. Во машинското учење

постојат два вида на податоци – означени податоци и неозначени податоци. Означените податоци ги имаат и влезните и излезните параметри во целосно читлива шема за машина, но бараат многу повеќе труд за да ги означат податоците, за почеток. Податоците без ознаки имаат само еден или ниеден параметар во машински читлива форма.

Машинското учење се базира на два основни типа на генерички алгоритми – надгледувано учење и учење без надзор.

4.1.1. Учење во длабочина

Длабокото учење е функција на вештачка интелигенција која го имитира работењето на човечкиот мозок при обработка на податоци и создавање обрасци за употреба при донесување одлуки. Длабокото учење е подмножество на машинското учење во вештачката интелигенција кое има мрежи способни да учат без надзор од податоците што се неструктурирани или без ознаки [6]. Исто така, тоа е познато како длабоко невронско учење или длабока невронска мрежа. Алгоритмите за машинско учење користат структурирани, односно означени податоци за да направат предвидувања, што значи дека специфични карактеристики се дефинирани од влезните податоци за моделот и се организирани во табели. Ова не мора да значи дека не користи неструктурирани податоци, туку тоа само значи дека доколку се случи генерално поминува низ одредена претходна обработка за да се организира во структуриран формат. Длабокото учење се заснова на хиерархиски принцип, па поради тоа при негова примена во анализата на податоци, соодветно е истата количина на податоци да се поедностави, со помош на семантичко индексирање, означување на податоци, пронаоѓање на корисни информации и задачи.

Длабокото учење елиминира некои од претходните обработки на податоци што обично се вклучени во машинското учење. Овие алгоритми можат да внесат и обработуваат неструктурирани податоци, како текст и слики, при што се автоматизира екстракција на карактеристики, отстранувајќи дел од потребата од човечки експерти. На пример, доколку е потребно да се анализираат сет фотографии од различни миленичиња истите тие да се категоризираат како

„мачка“, „куче“, „хрчак“ и слично, алгоритмите за длабоко учење можат да одредат кои карактеристики (на пример, ушите) се најважни за да се разликуваат наведените животни, едно од друго. Во машинското учење, оваа хиерархија на карактеристики е воспоставена рачно од човечки експерт.

4.1.2. Надгледувано учење

Надгледуваното учење е тип на машинско учење во кое машините се обучуваат користејќи добро „означени“ податоци за обука и врз основа на тие податоци, машините го предвидуваат излезот. Етикетираните податоци значат дека некои влезни податоци се веќе означени со точниот излез. Во надгледуваното учење, податоците за обука што им се даваат на машините работат како супервизор кој ги учи машините правилно да го предвидат излезот. Го применува истиот концепт како што учи ученикот под надзор на наставникот. Тоа подразбира формирање сет на податоци и нивна обука со користење на посебен софтвер. Софтверот ги применува набљудуваните врски помеѓу податоците во комплетот за обука на нов сет на податоци, што не е виден порано. Проблемите што се решаваат со методот на надгледувано учење се поделени на регресивни и класификациски. Во проблеми со регресија, целта е да се поврзат внесените променливи и да се прикаже резултатот со континуирана функција. Пример за ова претставува предвидувањето на цената на недвижностите врз основа на претходното искуство во следењето на трансакциите со недвижен имот [7]. Во проблемите со класификацијата, целта е да се прикаже резултатот во форма на дискретна излезна вредност.

4.1.2.1. Регресија

Постојат два вида на регресија:

- Линеарна регресија
- Логистичка регресија.

4.1.2.1.1. Линеарна регресија

Линеарната регресија [8] е еден од наједноставните и најчесто користени модели на машинско учење. Претставува алгоритам за машинско учење базиран

на надгледувано учење. Врши регресивна задача. Регресијата моделира целна вредност на предвидување врз основа на независни променливи. Најчесто се користи за откривање на врската помеѓу променливите и предвидувањето. Различни модели на регресија се разликуваат врз основа на видот на односот помеѓу зависните и независните променливи, и бројот на независни променливи што се користат. Постојат различни начини за воведување. Едната е веројатноста и открива повеќе за овој метод отколку другите. Од аспект на веројатност, основната претпоставка е претпоставка за нормална распределба на целната променлива *y*, со оглед на вредностите на атрибутот *x*. Ова е претставено со следната формула:

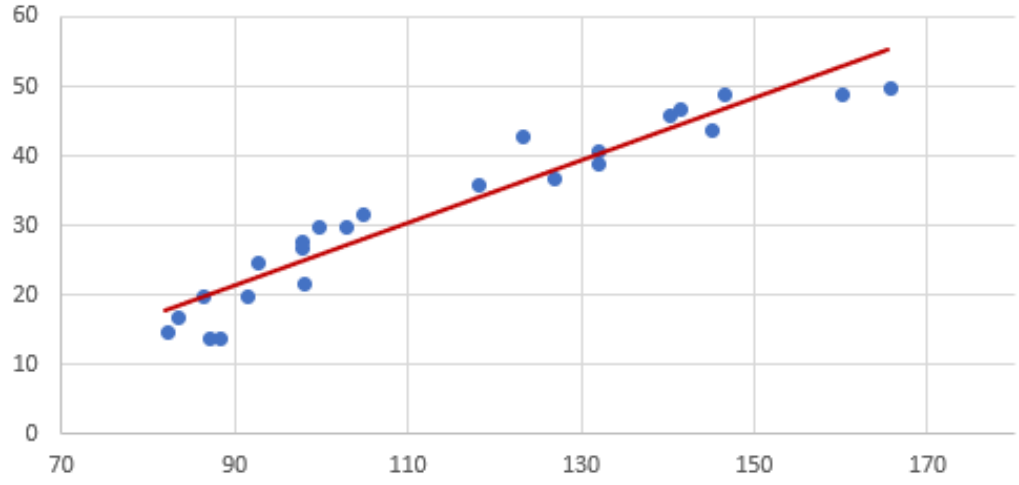
$$p(y|x) = N(f(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - f(x))^2}{2\sigma^2}\right)$$

Во зависност од формата на оваа функција, можно е да се добијат многу различни модели. Технички, наједноставната форма на моделот е линеарна:

$$f_\omega(x) = \omega \cdot x$$

Математички, линеарната регресија е дефинирана со оваа формула:

$$y = bx + a + \varepsilon$$



Слика 4. Пример за линеарна регресија
Figure 4. Example of Linear Regression

4.1.2.1.2. Логистичка регресија

Логистичката регресија [9] е надгледуван алгоритам за класификација на учење кој се користи за да се предвиди веројатноста за целна променлива. Природата на целната или зависната променлива е дихотомна, што значи дека би имало само две можни класи. Со едноставни зборови, зависната променлива е бинарна по природа и има податоци кодирани како 1 (означува да) или 0 (означува не).

Математички, логистички регресивен модел предвидува *P (Y = 1)* како функција на *X*. Логистичката регресија користи формула, многу слична на линеарната регресија. Влезните вредности *(x)* се комбинираат линеарно користејќи тежини или вредности на коефициентот за да се предвиди излезна вредност *(y).* Клучна разлика од линеарната регресија е тоа што излезната вредност што се моделира е бинарна вредност (0 или 1), а не нумеричка вредност. Подолу е даден пример за логистичка регресија:

$$y = \beta 0 + \beta 1 \times x1 + \beta 2 \times x2 + \cdots \beta_n x_n$$

Пример формула за веројатност:

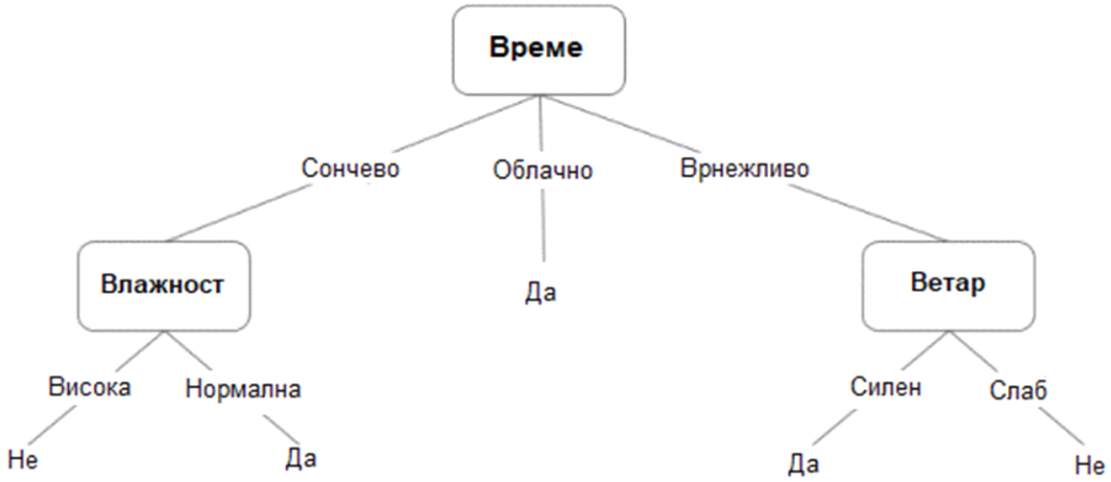
$$p = \frac{e^{logit(p)}}{1} + e^{logit(p)}$$

4.1.2.2. Дрво за одлучување

Алгоритам на дрвото за одлучување припаѓа на семејството на надгледувани алгоритми за машинско учење. Може да се користи и за проблем со класификација, како и за проблем со регресија. Целта на овој алгоритам е да се создаде модел што ја предвидува вредноста на целната променлива, за што дрвото на одлуки ја користи претставата за да го реши проблемот во кој листот јазол одговара на ознака на класа и атрибутите се претставени на внатрешниот јазол на дрвото [10]. Претпоставки што се прават при употреба на дрвото за одлучување.

- Во почетокот, целата колекција на податоци се смета како основна;
- Се претпочитаат карактеристичните вредности да бидат категорични;

- Се користи статистички метод за подредување на атрибути како надворешен јазол или внатрешен јазол.



Слика 5. Дрво за одлучување
Figure 5. Example of Decision Tree

Слика 5 илустрира пример за дрво за одлучување. Можеме да видиме дека секој јазол претставува атрибут или карактеристика, а гранката од секој јазол го претставува исходот на тој јазол. Крајно се претставени лисјата на дрвото каде што се донесува конечната одлука. Ако карактеристиките се континуирани, внатрешните јазли можат да ја тестираат вредноста на карактеристиката според прагот. Дури и кога некои примероци може да имаат непознати вредности, може да се користи методот дрво за одлучување.

Ентропијата контролира како дрвото за одлучување донесува одлука да ги подели податоците. Таа влијае на тоа како дрвото за одлучување ги поставува своите вредности. Вредностите на ентропијата се движат од 0 до 1. Помала вредност на ентропија не се пресметува. Формулата за ентропија е:

$$H(s) = -probability\ of\ log_2(p+) - -probability\ of\ log_2(p-)$$

Каде што:

(p+) - % позитивно

(p-) - % негативно

4.1.2.3. Бајесов алгоритам

Бајесов алгоритам (Naive Bayes) е модел за машинско учење кој се користи за голем обем на податоци. Дава многу добри резултати кога станува збор за задачи за NLP, како што е сентименталната анализа [11]. Тоа е брз и некомплициран алгоритам за класификација. Бајесовиот алгоритам се базира на моделирање на распределбата на целната променлива *y* по дадени вредности на променливата *x,* користејќи ја Бајесовата формула:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Општо земено, целта на Бајесовиот алгоритам е да ја процени задната распределба *(p (θ | x))* со оглед на веројатноста *(p (x | θ))* и претходната распределба, *p (θ).* Веројатноста е нешто што може да се процени од податоците за обука.

4.1.3. Учење без надзор

Учење без надзор [12] е техника на машинско учење во која корисниците не треба да го надгледуваат моделот. Наместо тоа, му овозможува на моделот да работи самостојно за да открие обрасци и информации што претходно не биле откриени. Главно се занимава со податоци без ознаки. Се однесува на заклучување на основни обрасци од необележана база на податоци без никаква референца за означени резултати или предвидувања. При учење без надзор не се користи сет за обука, се анализираат само техники за наоѓање скриено знаење во комплетот. Во случај на учење без надзор може да се формираат структури на податоци, групи. Кластерите се направени врз основа на односот на променливите во податоците. Целта е да се забележат законитости меѓу податоците и нема повратни информации врз основа на предвидените резултати, односно не се знае што е вистина, а што не. Затоа, потенцијалното решение за проблемот не може да се оптимизира со овој метод. Моделите на учење без надзор се користат за три главни

задачи - групирање, поврзување и намалување на димензионалноста. Алгоритмите за учење без надзор им овозможуваат на корисниците да извршуваат посложени задачи за обработка во споредба со надгледуваното учење. Учењето без надзор може да биде понепредвидливо во споредба со другите природни методи на учење.

Постојат неколку методи за учење без надзор, но кластерирањето е една од најчесто користените техники за учење без надзор. Кластерирањето се однесува на процесот на автоматско групирање на податоци со слични карактеристики и нивно доделување на „кластери“.

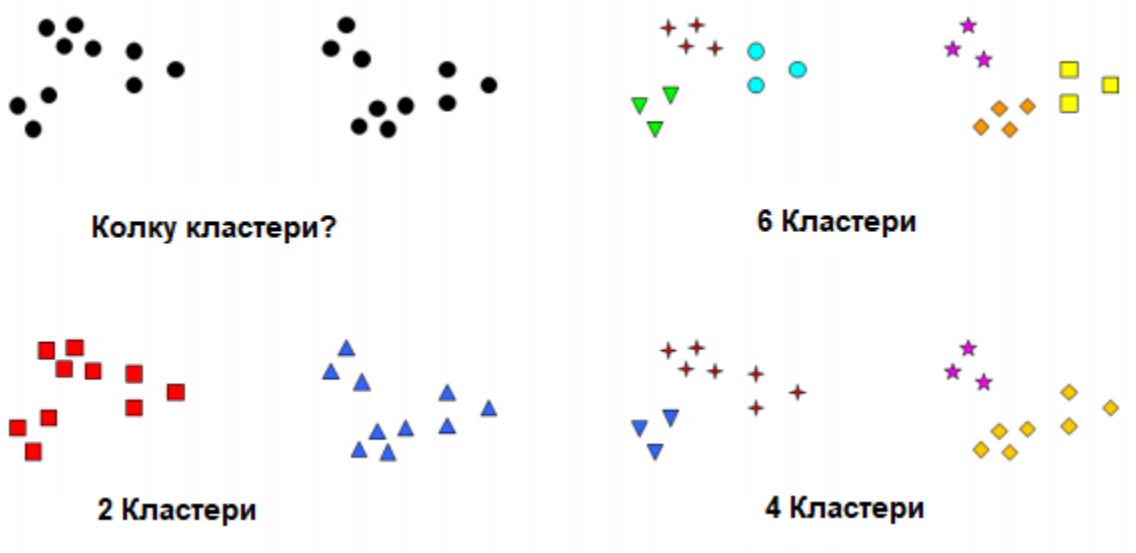
4.1.3.1. Кластерирање

Како што кажува и името, кластерирањето вклучува поделба на податоци во повеќе групи со слични вредности. Со други зборови, целта на групирањето е да се одделат групи со слични карактеристики и да се спојат заедно во различни групи. Идеално, тоа е имплементација на човечката когнитивна способност во машините што им овозможува да препознаат различни предмети и да ги разликуваат меѓу нив врз основа на нивните природни својства. За разлика од луѓето, многу е тешко за машината да идентификува дали станува збор за јаболко или портокал, освен ако не е соодветно обучена на огромна релевантна база на податоци. Оваа обука се постигнува со алгоритми за учење без надзор, особено со кластерирање.

Кластерирањето е важен концепт кога станува збор за учење без надзор [13]. Главно се занимава со наоѓање на структура или модел во збирка некатегоризирани податоци. Алгоритмите за групирање без надзор за учење ги обработуваат податоците и наоѓаат природни групи, доколку постојат во податоците. Исто така, може да се модифицира колку кластери треба да ги идентификуваат алгоритмите. Тоа овозможува приспособување на грануларноста на овие групи. На пример, во случај на обработка на огромен број продажби цели групи на податоци може да се заменат со нивните претставници. Ова не е идеално од гледна точка на квалитетот на добиениот предвидлив модел, но од гледна точка на пресметковната и меморискатa ефикасност може да биде профитабилно. Исто така, за да се осигури дека различни слоеви имаат слично дистрибуирани податоци за време на вкрстена валидација, податоците може прво да се групираат, а потоа

да се формираат n слоеви со делење на секој од кластерите на n еднакви делови кои се класифицирани во различни слоеви. Организирањето на податоците во кластери помага во идентификување на основната структура на податоците и наоѓа примени во индустријата.

Очигледно, кластерирањето може да биде корисно и како техника за решавање проблеми и како техника за претходна обработка. Како што покажува примерот прикажан на слика 6, неколку различни групирања, честопати со различна грануларност, можат да се идентификуваат во еден сет. Во исто време, ваквите случаи не се резултат на недоволно разгледување на дефиницијата за групирање, туку на разновидноста на контекстите во кои може да се изврши групирање и целите што кластерирањето треба да ги постигне.



Слика 6. Различни групи на исти податоци
Figure 6. Different sets of the same data

4.1.3.2. K-means

Алгоритмот за групирање K-means се користи за пронаоѓање групи кои не се експлицитно означени во податоците. Ова може да се искористи за да се потврдат деловните претпоставки за тоа какви типови групи постојат или да се идентификуваат непознатите во комплексни податоци [14].

Алгоритмот k-means значи наоѓање на k кластери во податоците претставени со k центроиди на овие кластери, од кои секој се добива со просек на елементите на даден кластер. Центроидот или геометрискиот центар на рамнинска фигура е аритметичка средна позиција на сите точки на сликата. Оваа претпоставка го прави алгоритмот применлив само за податоци како што се вектори. Под одредени услови постојат генерализации на алгоритмот за други видови податоци, но тие нема да бидат дискутирани. Почетните k центроиди се избираат по случаен избор (иако, ако корисникот знае нешто за структурата на нивните податоци, тие може да се дадат однапред), а потоа се повторуваат следните чекори:

- Распоредување на сите инстанци во нови кластери со спојување на секој пример до најблискиот центроид;

- Пресметување на новите центроиди како просек од случаите поврзани со нив.

Овие чекори се извршуваат сè додека се менуваат центроидите. Кога центроидите се исти во две последователни повторувања, алгоритмот престанува.

4.1.3.3. Мешавина од нормални дистрибуции и EM алгоритам

Мешавината од нормални дистрибуции претставува нешто пософистициран модел на групирање од моделот k-means. Основната претпоставка е дека податоците може да се поделат на голем број релативно компактни групи чија форма може добро да се опише со нормални распределби со различни просеци. Просеците јасно ги дефинираат позициите на кластерите, додека матриците за коваријанса ја опишуваат нивната форма и ориентација во просторот. Областите со еднаква густина во кластер во овој случај се елипсоиди. Во овој модел дистрибуцијата е малку посложена отколку во претходниот, но има природно распаѓање во поедноставни дистрибуции [15].

Прво, кластерот може да се избере по случаен избор, а потоа да се избере точка од тој кластер според нормалната распределба што му одговара. Ако C е бројот на множества на броевите *p1* ,. . . , *pC* важи *pi ≥ 0 u PC i = 1 pi = 1,* тогаш *(p1, ..., pC)* ја

претставува мултиномната распределба низ кластерите. Густината на распределба по примери може да се запише со следната формула:

$$p(x) = \sum_{i=1}^{C} p_i N(x; \mu, \Sigma_i)$$

каде што *N* е густината на нормалната распределба со повеќе променливи претставена со формулата:

$$N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Кога вредностите на *zi* за секој од примерите се познати, параметрите *μk* и *Σk* на нормалната распределба на кластерот *k* се проценуваат со стандардни емпириски проценки, просекот на случаите за кои *zi = k* се применува со матрицата на коваријанса помеѓу сите атрибути пресметани врз основа на овие случаи. Во такви ситуации може да се претпостави дека постојат таканаречени латентни или скриени променливи z кои не се набљудуваат. Кога се дадени емпириски податоци и за xi за z се користи алгоритам за максимизирање на очекувањата или накратко EM алгоритам.

Подолу е претставен EM алгоритамот во општа форма. Следната формула се однесува на алгоритамот за функцијата за сигурност на параметарот:

$$l(\omega) = log\mathrm{L}(\omega) = \sum_{i=1}^{N} logp_{\omega}(x_i) = \sum_{i-1}^{N} log \int_{z} p_{\omega}(x_i, z)dz$$

Овој проблем не е лесно да се реши, бидејќи алгоритамот не може да помине низ збирот. Решението се состои во набљудување на функцијата за веродостојност во однос на заедничката распределба на променливите *x* и *z*. Со оглед на тоа што покрај вредностите на набљудуваните променливи *x* вредностите на латентните променливи *z* соодветствуваат на секоја инстанца, ова е сосема легитимно, но бидејќи вредностите на *z* не се познати, таква функција не може да се пресмета, туку претставува случајна променлива.

5. Алатки за анализа на големи податоци

5.1. Hadoop

Hadoop е проект на Apache, каде што сите компоненти се достапни преку лиценцата за отворен извор на Apache [16]. Hadoop обезбедува дистрибуиран датотечен систем и рамка за анализа и трансформација на многу големи множества податоци со помош на програмскиот модел MapReduce. Важна карактеристика на Hadoop е поделбата на податоците и пресметката на многу (илјадници) домаќини и извршување на апликациски пресметки паралелно близу до нивните податоци. Hadoop Distributed File System (HDFS) е компонента на датотечниот систем на Hadoop. HDFS ги зачувува метаподатоците за датотечниот систем и податоците за апликацијата одделно. Hadoop е дизајниран да се зголемува од единечни сервери до илјадници машини, од кои секоја нуди локално пресметување и складирање. Наместо да се потпира на хардвер за да се испорача голема достапност, библиотеката е дизајнирана да открива и да се справува со дефекти во слојот на апликацијата, така што испорачува високодостапна услуга на врвот на кластер компјутери, од кои секоја може да биде склона кон дефекти. Hadoop во своето јадро содржи две главни компоненети, на коишто се сведува целокупното функционирање и работа на Hadoop, HDFS и MapReduce.

Како и другите дистрибуирани датотечни системи, како PVFS , Luster и GFS, така и HDFS ги зачувува метаподатоците на наменски сервер, наречен NameNode. Податоците за апликацијата се зачувуваат на други сервери наречени DataNodes. Сите сервери се целосно поврзани и комуницираат едни со други користејќи протоколи базирани на TCP. За разлика од Luster и PVFS, DataNodes во HDFS не користат механизми за заштита на податоци како RAID за да ги направат податоците издржливи. Наместо тоа, како GFS, содржината на датотеката се реплицира на повеќе DataNodes за сигурност. Додека се обезбедува издржливост на податоците, оваа стратегија има дополнителна предност што ширината на опсегот на пренос на податоци се множи и има повеќе можности за лоцирање на пресметките во близина на потребните податоци. Неколку дистрибуирани

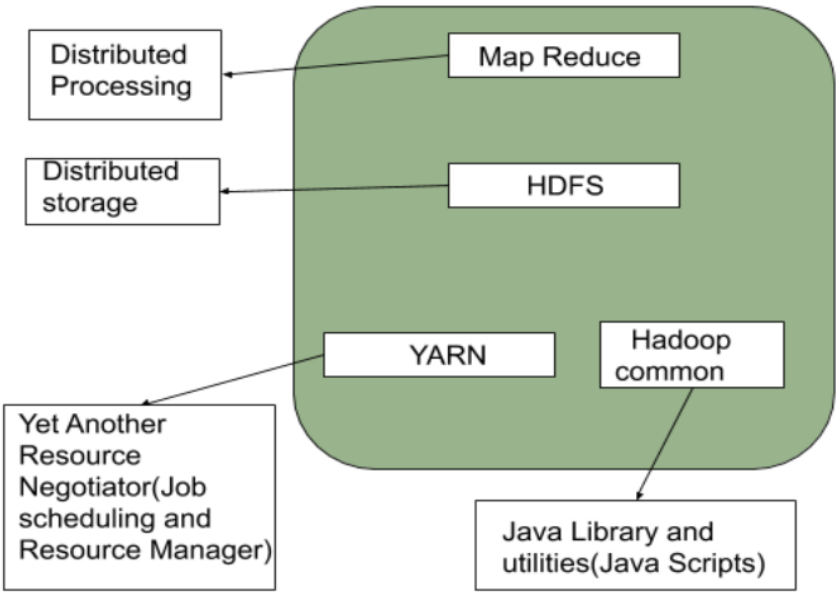
датотечни системи имаат или истражуваат вистински дистрибуирани имплементации на именскиот простор.

Постојат многу алатки на пазарот што се поврзани со Hadoop, а компании како Facebook и Microsoft развиваат свои, што се исто така достапни за инсталација. Денес има компании кои обезбедуваат бесплатни услуги, односно платформи базирани на Hadoop со веќе изграден екосистем, што го олеснува изборот на потребните проекти за работа и процесот на инсталација на персонален компјутер. Меѓу најпознатите се: Cloudera, Apache, MapR Technologies, IBM и други.

5.1.1. Архитектура на Hadoop

Hadoop се користи за складирање и имплементација на големите сетови на податоци. Како што беше споменато, јадрото на Hadoop се состои од дел за складирање – HDFS и дел за обработка - MapReduce. Hadoop ги дели датотеките во големи блокови и ги дистрибуира меѓу јазлите во компјутерскиот кластер. Рамката MapReduce и HDFS обично се на ист сет јазли, што овозможува рамката да распоредува задачи на јазли што содржат податоци. Рамката Map Reduce се состои од еден главен JobTracker и еден slave TaskTracker по јазол.

Слика 7 ги претставува компонентите на Hadoop. JobTracker е од суштинско значење бидејќи ги извршува сите задачи на MapReduce на различни јазли во кластерот, односно во оние јазли што веќе ги содржат податоците или барем се наоѓаат во истата решетка како и јазлите што ги содржат податоците. JobTracker е услуга во рамките на Hadoop која е одговорна за преземање на барањата на клиентите. Ги доделува на TaskTrackers на DataNodes каде што потребните податоци се локално присутни. Ако тоа не е можно, JobTracker се обидува да ги додели задачите на TaskTrackers во истата решетка каде што податоците се локално присутни.



Слика 7. Компоненти на Hadoop
Figure 7. Hadoop components

Таквиот пристап ја фаворизира локалноста на податоците, јазлите манипулираат со податоците што ги поседуваат овозможувајќи побрза обработка на податоците и поголема ефикасност во однос на конвенционалните архитектури на суперкомпјутери кои се потпираат на паралелен датотечен систем каде што податоците и пресметките се поврзани со брза мрежа [17].

• Пакет Hadoop Common - содржи библиотеки и услуги за други модули.

Hadoop Common пакет

Пакетот Hadoop Common ги содржи потребните архиви на Java или JAR-датотеки и скрипти што се користат за извршување на Hadoop операции. Овој пакет содржи изворен код и документација. Исто така, овој пакет ги содржи сите потребни елементи за да може Hadoop да комуницира со други алатки. Структурата на пакетот се менува поради воведувањето на нови верзии.

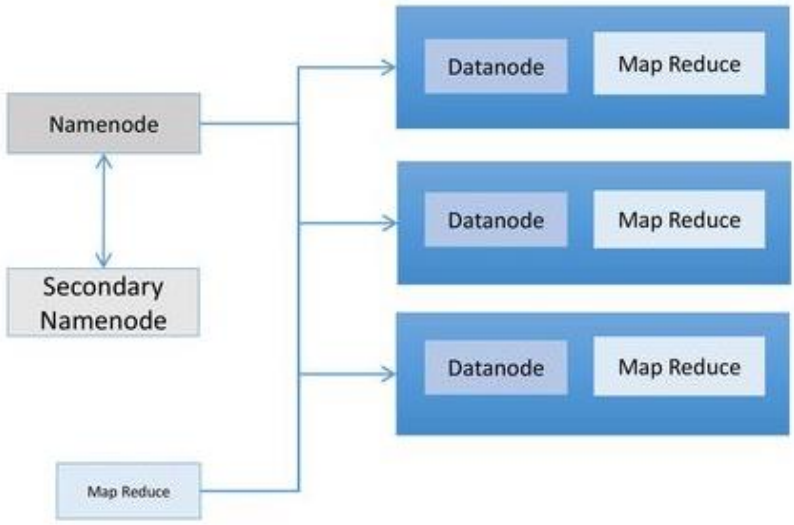
5.1.1.2. Дистрибуиран датотечен систем Hadoop (HDFS)

HDFS е дистрибуиран датотечен систем и компонента на Hadoop [18]. Тоа е всушност систем што дефинира како се чуваат, копираат и читаат податоците. HDFS овозможува складирање на податоци во кластер на дистрибуиран начин. Ги

дели големите датотеки со кои работи во блокови, кои се поставени на предефинирана големина од 128 MB. Секоја машина во кластерот може да пристапи до дел од потенцијално голема датотека и да ја обработи. Покрај тоа, има вградена репликација, не се потребни дополнителни машини, а идејата е дека ако машината престане да работи тогаш обработката не запира, туку продолжува, па затоа постои толеранција на грешки. Односно, копија од еден блок постои на друга машина наречена јазол.

HDFS овозможува лесно складирање на големи количини податоци. Тој е дизајниран да работи на која било хардверска инфраструктура, иако нејзината вистинска моќ се гледа на серверите. Системот е толерантен на грешки и не е хардверски интензивен. Бидејќи е наменет за сервери, поточно за стотици или илјадници сервери, кои имаат различни компоненти и за кои постои можност за дефект, главната цел на архитектурата на HDFS е брзо откривање на грешките и автоматско отстранување на истите. Тој е дизајниран да манипулира со големи количини на податоци (гигабајти-GB) и има едноставен модел за пристап до датотеки - „напиши еднаш - вчитај повеќе пати“. HDFS не е наменет за интеракција на корисникот, туку за непречена обработка на податоците од други апликации. HDFS има архитектура господар/роб (master/slave). Како што беше споменато, компјутерскиот кластер кај Hadoop се состои од единствен NameNode и повеќе DataNodes. Обично еден сервер е главен и NameNode е инсталиран на него, а DataNode на други сервери. Главниот сервер на кој е инсталиран NameNode го контролира пристапот до датотека и управува со именскиот простор на датотечниот систем што ја поддржува традиционалната хиерархија. Корисникот може да создаде директориум и да ја зачува датотеката во него, може да го смени името на директориумот, да го избрише или премести. Истото важи и за датотеките. Корисникот управува и со датотеки на главниот сервер и има пристап до директориуми и датотеки. Понатаму, датотеките се поделени во блокови и се чуваат на други сервери на кои е инсталиран DataNode. DataNode се користи за складирање блокови, дозволува создавање, бришење и репликација на блокови. Едноставно кажано, NameNode ги зачувува метаподатоците за блокирање и им помага на крајните корисници да ја видат датотеката, а не блоковите што не се

читаат од корисникот, додека DataNode складира податоци. Ова исто така може да се толкува како складирање на адреси на блокови, име и број на копии во NameNode. HDFS има можност да пишува блокови повеќе пати, што го намалува ризикот од загуба на податоци. Обично содржи три копии од податоците што значи дека секој блок ќе се копира три пати. Ова не мора да се случи само на еден сервер, туку на сите сервери во кластерот на кој е инсталиран DataNode (на пример, за 1 TB податоци ќе бидат потребни 3 TB простор). Комуникацијата помеѓу серверите во кластерот се одвива со помош на протоколот TCP/IP и затоа NameNode комуницира со DataNodes, така што тие периодично испраќаат импулси т.н. *име на јазли*. Во однос на репликацијата на блокот, корисниците се безбедни во случај на дефект на еден или дури два роб-сервери, бидејќи секогаш ќе останат со друга копија од податоците. Меѓутоа, ако главниот сервер престанал да работи, корисникот нема метаподатоци, во тој случај ова претставува проблем. Иако се бараат решенија и за ова (можност за копирање на главниот сервер - создавање на секундарен NameNode), засега потенцијалната можност за откажување на главниот сервер е единствениот недостаток на HDFS. На слика 8 се прикажани компонентите на HDFS.



Слика 8. Компоненти на HDFS
Figure 8. HDFS components

5.1.1.2.1. Name Node

Јазолот за име (Name Node) е контролорот кој го контролира целиот датотечен систем. Така, секое барање што доаѓа во датотечниот систем, како што е креирање на директориум, креирање на датотека, читање и пишување во датотека, ќе помине низ јазолот за име. Значи, јазолот за име во суштина управува со датотечниот систем. Има n меморија и таа содржи блокови со обвивки. Секогаш кога ќе се стави датотека во HDFS, Name Node ќе ја раздели датотеката, ќе ги отвори блоковите и ќе се шири низ јазлите на податоците [19]. Јазлите за името знаат дека сите блокови се наоѓаат во кластерот.

5.1.1.2.2. Data Node

Јазолот за податоци (Data Node) е работна сила на системот. Тие ја извршуваат целата операција на блоковите. Тие ќе добијат инструкции од јазолот за име каде да ги постават блоковите и како да ги постават блоковите. Доколку е потребно податоците да бидат надвор од кластерот, тогаш се доделува команда до кластерот за податоци. Клиентот всушност комуницира со јазол за име. Се доделува команда на јазолот за име за добивање на блоковите каде што се лоцирани податоците. Јазолот за име ги испраќа информациите до клиентот. Тој клиент директно комуницира со јазолот за име, каде што јазлите за податоци ги опслужуваат тие блокови директно до клиентот. Јазлите за податоци се исто така одговорни за репликација. Додека јазолот за име е контролорот што ги испраќа информациите каде да се реплицира, јазолот за податоци е оној што прави физичка репликација.

5.1.1.2.3. Secondary NameNode

Secondary NameNode е специјално посветен јазол во кластерот кај HDFS чија главна функција е да презема контролни точки на метаподатоците на датотечниот систем присутен кај NameNode. Од неговото име може да се заклучи дека Secondary NameNode е еден вид резервен сервер кој ќе почне да дејствува како NameNode во случај NameNode да пропадне, но тоа не е така. Всушност, Secondary NameNode може да се замисли како асистент на NameNode кој зема дел од

работниот товар на NameNode [20]. Наредниот директориум на Secondary NameNode има ист распоред како и тековниот директориум на главниот NameNode.

5.2. Ниво на обработка на податоци

5.2.1. MapReduce

MapReduce е софтверска рамка за лесно пишување апликации кои паралелно обработуваат огромни количини на податоци (множества на податоци од повеќе терабајти) на големи кластери, илјадници јазли на хардвер којшто е сигурен, толерантен на грешки. MapReduce е софтверски модел за обработка на големи количини на податоци чиј алгоритам се извршува паралелно и се дистрибуира (на пример, во компјутерски кластер од три сервери, алгоритмот ќе се изврши паралелно, односно истовремено на сите три сервери). MapReduce е најкористената компјутерска рамка за анализа на големи податоци. Работата на MapReduce обично ги дели податочните сетови на независни парчиња, кои се обработуваат од задачите зададени на мапата на сосема паралелен начин [21]. Рамката ги подредува излезите на мапите, кои потоа се внесуваат во задачите за намалување. Обично и влезот и излезот на работата се чуваат во датотечниот систем. Рамката се грижи за закажување задачи, нивно следење и повторно извршување на неуспешни задачи.

Работата на MapReduce претежно се одвива во 2 фази:

- Map Phase,
- Reduce Phase.

Функцијата Map обработува блок од базата на податоци како пар (клуч, вредност) и произведува излез на мапа во форма на листа на парови (клуч, вредност). Map се користи за едноставно или сложено сортирање и филтрирање на податоци (на пр. сортирање на ученици по презиме). Средните вредности се групирани заедно врз основа на истиот клуч на пример 2k, а потоа се преминува на функцијата за намалување. Reducer ги комбинира податоците обработени од Map, на пример собира колку пати се повторува еден збор во дадени реченици. Функцијата за намалување го зема средниот клуч 2k заедно со неговите поврзани

вредности и ги обработува за да произведе нова листа на вредности како конечен излез. MapReduce е високоскалабилен компјутерски модел кој може да им овозможи на илјадници евтини компјутерски стоки да се користат како ефективна компјутерска платформа за дистрибуирани и паралелни пресметки.

Библиотеките за MapReduce се напишани на многу програмски јазици со различни разновидни подобрувања. Влезните фрагменти се состојат од парови на клучни вредности. Излезот за мапирање потоа служи како влез за фазата на намалување. Задачата за намалување го комбинира резултатот во одреден излез на парот со клучна вредност и ги запишува податоците во HDFS.

Сите податоци се зачувуваат како блокови на DateNodes, а метаподатоците за тие блокови се зачувуваат на NameNodes [22]. Кога податоците се поделени на блокови полесно може да се процесираат, така што складирањето на податоци во блокови го олеснува групирањето на функцијата Map. Бидејќи HDFS има повеќе податочни јазли (DataNode) на кои податоците се поделени и складирани во блокови, можно е да се користи компјутерската моќ на секој од овие јазли и да се извршуваат задачи на нив. Така, секој јазол може да извршува задачи „Map“ или „Reduce“ и бидејќи секој јазол содржи повеќе податоци, можно е да се очекува извршување на задачите во исто време за различни блокови на податоци.

Користејќи ги ресурсите на повеќе меѓусебно поврзани машини, MapReduce ефикасно се справува со голема количина структурирани и неструктурирани податоци. MapReduce доделува фрагменти од податоци низ јазлите во кластерот на Hadoop. Целта е да се подели базата на податоци на делови и да се користи алгоритам за истовремена обработка на тие делови. Паралелната обработка на повеќе машини значително ја зголемува брзината на ракување со дури и количината од податоци изразена во петабајти.

5.2.2. Hadoop YARN

Во првата генерација на Hadoop компонентата YARN не постоеше, но нејзината работа беше составен дел од MapReduce. YARN беше воведен како нова компонента во втората генерација на Hadoop чија цел беше да се раздели

претходната MapReduce работа на два дела за да се олесни користењето на целата платформа [23]. Се состои од две компоненти: Scheduler и Applications Manager кои заедно го сочинуваат Resource Manager. Со одделување на овој процес во нова компонента, MapReduce се користеше само за обработка на податоци. Друга можност што се појави со втората генерација на Hadoop, односно со YARN, е дека може да се започнат голем број апликации напишани за Hadoop. Ова имаше особено влијание врз деловниот свет, бидејќи со можноста за правење повеќе работи паралелно, се создаваше конкуренција на пазарот.

Scheduler е компонента која се грижи за распределбата на ресурсите на апликациите што работат. Важно е дека се земаат предвид само ресурсите, односно не се грижи за статусот на апликацијата што се извршува, така што работата на апликацијата не се следи. Бидејќи се грижи само за распределбата на ресурсите, не се зема предвид дали настанала грешка или кодот е лош, што значи дека ресурсите ќе бидат распределени на апликацијата сè додека нејзината работа не биде прекината од корисникот или некоја друга компонента.

ApplicationsManager управува со апликациите напишани за Hadoop. Неговата работа е да преговара и да ја прифати задачата. Да се преговара значи да се испитаат ресурсите и да се направат заклучоци за тоа што треба прво да се направи. Тој исто така е одговорен за ресетирање на работата или апликацијата доколку се појави грешка.

5.3. Дистрибуции на Hadoop

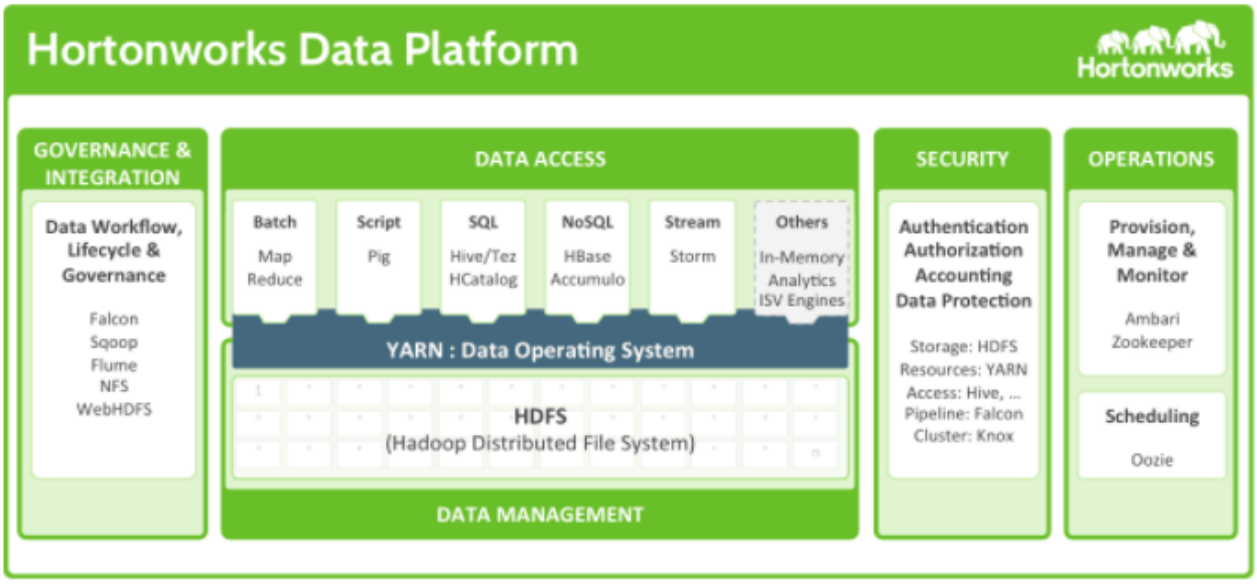
5.3.1. Cloudera

Cloudera е аналитичка платформа составена од интегрирани технологии со отворен извор, кои им помагаат на корисниците да извлекуваат заклучоци од нивните податоци [24]. Формирајќи облак со податоци за организацијата, таа го става управувањето со податоците на дофат на аналитичарите, со приспособливост и еластичност за управување со секој обем на работа. Cloudera исто така им нуди на корисниците транспарентност во целиот животен циклус на

податоци и флексибилност на приспособувањето преку неговата отворена архитектура.

5.3.2. Hortonworks

Платформата за податоци Hortonworks [25] е деловно решение создадено во 2011 година во САД. Hortonworks е ориентирана кон управување со податоци што овозможува централизирана архитектура и извршување на индиректни, интерактивни апликации во реално време паралелно со дистрибуирани групи на податоци. Изграден е врз основа на проектот Apache Hadoop и поддржува сеопфатен сет на алатки кои се однесуваат на основните барања за безбедност, деловно работење и управување со податоци. Претставува бесплатна платформа, применлива за оперативните системи Windows, Mac или Linux и бара минимално знаење за програмирање, што го прави совршена алатка за почетници. За да го олесни учењето и програмирањето со Hadoop, податочната платформа Hortonworks нуди бесплатно преземање и инсталирање на системот Hortonworks Sandbox. Единствениот предуслов е корисникот веќе да има инсталирано управувачи со виртуелни машини како VirtualBox или VMware.



Слика 9. Архитектура на Hadoop имплементирана од Hortonworks
Figure 9. Hadoop Ecosystem Architecture implemented by Hortonworks

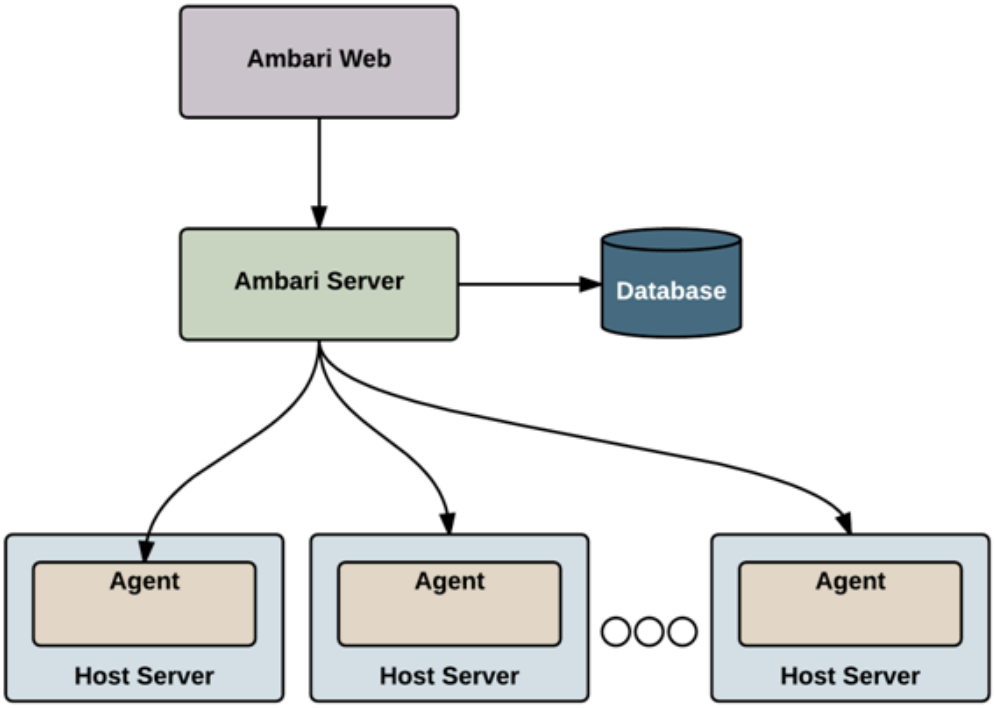
Префрлувајќи се на веб прелистувач корисникот наидува на графички интерфејс кој овозможува лесно пребарување, комуникација и пристап до најновите информации поврзани со Hortonworks Sandbox. Една од важните почетни активности е најава во системот на Ambari.

5.4. Ниво за управување со податоци

5.4.1. Apache Ambari

Следењето и одговорот на проблемите се двете главни активности што клиентот ги очекува од давателот на услуга кој управува со нивото на податоци и платформата. Apache Ambari [26] претставува платформа со отворен код за обезбедување, управување и следење на кластери од Apache Hadoop.

Apache Ambari, како дел од платформата за податоци на Hortonworks, им овозможува на претпријатијата да планираат, инсталираат и безбедно да го конфигурираат HDP што ќе го олесни обезбедувањето на тековно одржување и управување со кластерите, без разлика на нивната големина. Насочен е кон поедноставување на управувањето со Hadoop преку развој на софтвер за обезбедување, управување и следење на кластерите на Apache Hadoop. За да се визуализира напредокот, како и статусот на секоја апликација што работи преку кластерот на Hadoop, Ambari обезбедува интуитивен и лесен графички кориснички интерфејс. Неговиот конзистентен и безбеден интерфејс овозможува да биде прилично ефикасен во оперативната контрола. Серверот Apache Ambari прима комуникации од агентот Ambari. Откако агентот Ambari е доделен на зоната за пристап, тој се регистрира со серверот Ambari. Агентот потоа обезбедува статус за почеток на работа на серверот. За подобро разбирање на тоа како работи Ambari, на слика 10 е прикажана деталната архитектура на Apache Ambari [27]:



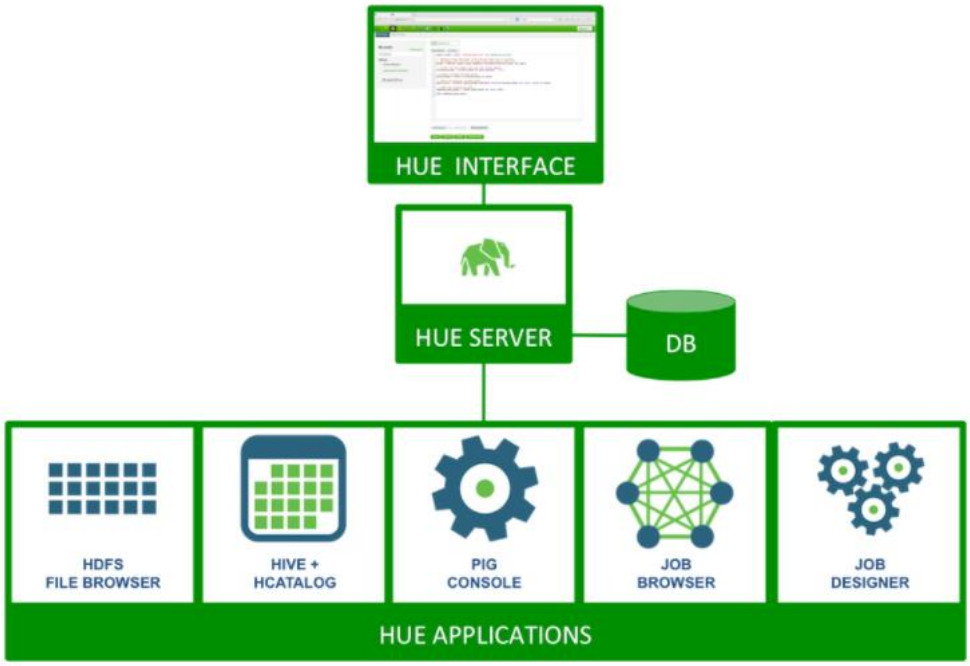
Слика 10. Архитектура на Apache Ambari
Figure 10. Apache Ambari architecture

Apache Ambari следи архитектура на господар/роб каде што главниот јазол им дава инструкции на јазлите на робовите да извршат одредени дејства и да ја пријават состојбата на секое дејство. Главниот јазол е одговорен за следење на состојбата на инфраструктурата. За да го направите ова, главниот јазол користи сервер за бази на податоци, кој може да се конфигурира за време на поставувањето.

5.4.2. Hue

Hue е алатка со отворен извор на интернет за Hadoop и неговиот екосистем е прикажан на слика 11. Напишан е во Python и ги поддржува најчестите алатки на екосистемот на Hadoop [28]. Hue е одличен избор за анализа на податоци, бидејќи нема употреба на терминали и командна линија. Најважните карактеристики на Hue се пребарувачот, Hadoop shell, администраторски права, уредникот Impala, прелистувач на датотеки HDFS, Pig, Hive, веб-интерфејс на Ozzie и пристап до

Hadoop API. Овој распоред на веб-кориснички интерфејс им помага на корисниците да ги прелистуваат датотеките, слично како корисник на Windows што ги лоцира своите датотеки на неговата машина.



Слика 11. Кориснички интерфејс на Hue
Figure 11. Hue user interface

5.4.3. ZooKeeper

ZooKeeper [29] е централизирана дистрибуирана услуга за координација на дистрибуирани апликации. Првично е развиен од Yahoo, а подоцна стана дел од екосистемот Hadoop. Услугите што ги нуди ZooKeeper вклучуваат управување со конфигурација, синхронизација, именување и членство во група. HBase, Flume и HDFS HA (голема достапност) сите зависат од ZooKeeper.

5.4.4. Avro

Apache Avro [30] е рамка за моделирање, сериско уредување и остварување повици од далечина (RPC). Avro дефинира компактен и брз бинарен формат на податоци за поддршка на интензивни апликации за податоци и обезбедува поддршка за овој формат на различни програмски јазици, како што се Java, C, C ++ и Python. Avro обезбедува ефикасна компресија на податоци и складирање на

различни јазли на Apache Hadoop. Во рамките на Hadoop, Avro пренесува податоци од една програма или јазик на друг.

5.4.5. Oozie

Apache Oozie [31] е систем за распоредување на работниот тек дизајниран да работи и управува со работни места во кластери на Hadoop. Тоа е сигурен, еластичен и скалабилен систем за управување кој може да се справи со ефикасно извршување на голем обем на работни текови. Активностите на работниот тек имаат форма на насочени ациклични графови. Oozie може да поддржува разни видови на Hadoop работни алатки, вклучително и MapReduce, Pig, Hive, Sqoop и Distcp.

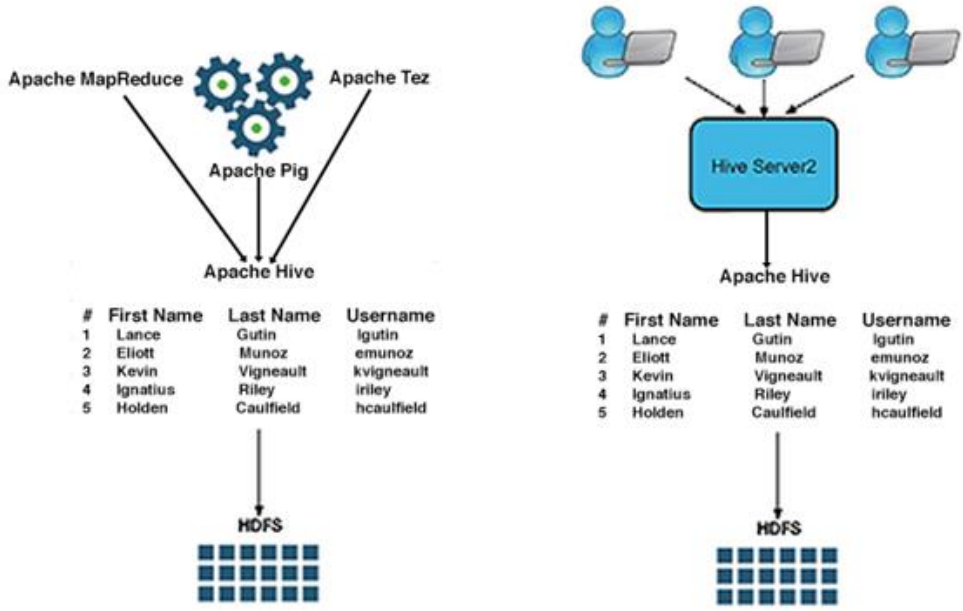
5.5. Ниво за пристап до податоци

5.5.1. Hive

Apache Hive е алатка што се користи за складирање и обработка на големи количини на податоци кај Hadoop [32]. Hive обезбедува нешто што се нарекува HiveQL. The Hive Query Language (HiveQL) е програмски јазик за барања на Hive за обработка и анализа на структурирани податоци од Metastore. Hive лесно може да се интегрира со постојните алатки користејќи го JDBC или ODBC интерфејсот, односно можно е да се поврзе со Microsoft Excel и слично. Се карактеризира со организација и складирање на големи групи податоци од различни извори и обезбедување на можност на корисниците за пребарување, структурирање и анализа на податоци за деловна интелигенција. Начинот на работа на Hive е многу сличен на релацискиот модел, при што табелите се слични на него, а податоците се организирани од поголеми до помали единици. До податоците се пристапува преку пребарувања кои се слични на SQL. За разлика од познатите бази на податоци, Hive не поддржува бришење и ажурирање на податоците. Причината за ова е што податоците во HDFS можат да се внесуваат, но не и да се менуваат. Постојат два случаи за примарна употреба на Hive:

- Table storage layer овозможува многу компоненти на HDP и основните технологии, како што се Apache Hive, Apache HBase, Apache Pig, Apache MapReduce

и Apache Tez да се потпираат на Hive како слој за складирање на податоци. На слика 12 е прикажан Use Case дијаграм за Table storage layer и SQL пребарувач.



Слика 12. Use Case дијаграм за Table Storage layer и SQL пребарувач
Figure 12. Use Case diagram for Table Storage layer and SQL browser

- SQL пребарувач, каде што администраторите на Hadoop, деловните аналитичари и научниците за податоци користат Hive за да извршат SQL пребарувања преку Hive CLI и од далечина преку клиент кој се поврзува со Hive преку HiveServer2.

5.5.2. Hcatalog

Apache HCatalog [33] е алатка што го олеснува управувањето со складирање на податоци и табели на Hadoop и им олеснува на корисниците да пишуваат и читаат податоци. Во практична смисла, HCatalog е слој на Hadoop што овозможува да се прикажуваат HDFS податоци во табеларна форма. HCatalog поддржува читање и пишување во датотеки преку Hive SerDe (Serializer-Deserializer) формати како RCFile, CSV, JSON.

5.5.3. Apache Pig

Apache Pig [34] е платформа што им овозможува на корисниците на Hadoop да напишат сложени трансформации на MapReduce со користење на едноставен скриптирачки јазик. Apache Pig ја преведува Pig Latin скриптата во програма на MapReduce што работи на YARN за да има пристап до податоците кои се складирани во HDFS. Pig е дизајнирана да извлече код на MapReduce напишан во Java програмскиот јазик. За разлика од SQL кој е декларативен програмски јазик, Pig е последователен јазик што значи дека начинот на кој се пишува програмата дефинира како ќе се трансформираат податоците. Скриптите напишани во овој јазик можат да бидат графови, што значи дека е можно да се напишат сложени трансформации со повеќе влезови и излези. Pig може да работи во две состојби, локална и MapReduce. Оваа алатка е наменета првенствено за ETL (Extract-Transform-Load) работа на необработени податоци и итеративна обработка на податоци.

5.5.4. Sqoop

Sqoop [35] е алатка за пренос на податоци помеѓу релациони бази на податоци и Hadoop. Apache Sqoop е дел од екосистемот на Hadoop. Бидејќи многу податоци треба да се пренесат од системи за релациона база на податоци на Hadoop, има потреба од специјална алатка за брзо извршување на оваа задача. Apache Sqoop е алатка која сега е широко користена за пренос на податоци од Relational Database Management System (RDBMS) во екосистемот на Hadoop за обработка со MapReduce. Апликацијата MapReduce не е во можност директно да пристапи до податоците што наоѓаат во надворешни релациони бази на податоци. Користејќи го Sqoop, податоците може да се преместат во HDFS/Hive/Hbase од MySQL/PostgreSQL/Oracle/SQL/Server/DB2 и обратно. При употреба на Sqoop се извршуваат следните постапки:

- Командата Sqoop import увезува табела од RDBMS во HDFS, односно секој запис од табелата RDBMS се смета за посебен запис во HDFS. Записите може да се складираат како текстуални датотеки и истите резултати ќе се добијат од HDFS,

при што се добива исходот во формат RDBMS, а процесот се нарекува извоз на табела;

- Го испраќа барањето до релациона база на податоци да испрати враќање на информациите за метаподатоците за табелата;

- Sqoop создава и зачувува команди за увоз и извоз за да добие подобар исход, обезбедувајќи точни резултати;

- Ги одредува параметрите за идентификување и потсетување на зачуваната работа, што помага да се создадат релевантни резултати од точка до точка;

- Повторното повикување или повторното извршување се користи во дополнителниот увоз, што може да ги увезе ажурираните редови од табелата од RDBMS во HDFS и обратно.

Алатката за увоз увезува индивидуални табели од RDBMS во HDFS. Секој ред во табелата се третира како рекорд во HDFS. Сите записи се чуваат како текстуални податоци во текстуални датотеки или како бинарни податоци во датотеки Avro и Sequence. Алатката за извоз извезува множество датотеки од HDFS назад во RDBMS. Датотеките дадени како влез во Sqoop содржат записи, кои се нарекуваат како редови во табелата. Тие се читаат и анализираат во збир на записи и се разграничуваат со разграничувач одреден од корисникот.

Алатката Sqoop може да помогне во увозот на структурирани податоци од релациони бази на податоци и NoSQL системи. Процесот за користење на Sqoop за преместување податоци е прикажан на слика 13:

Слика 13. Use Case дијаграм за користење на Sqoop за пресметување на податоци
Figure 13. Use Case for using Sqoop to move data into Hive

5.5.5. JAQL

JAQL е декларативен јазик на врвот на Hadoop кој обезбедува јазик за пребарување и поддржува масивна обработка на податоци. Ги претвора прашањата од високо ниво во работни места на MapReduce. Тој е дизајниран да бара полуструктурирани податоци базирани на JSON (Java-Script Object Notation) формат. Значи, JAQL како Pig не бара шема на податоци. JAQL обезбедува неколку вградени функции, основни оператори и I/O адаптери. Ваквите карактеристики обезбедуваат обработка на податоци, складирање, преведување и претворање на податоците во JSON формат.

JAQL е најмногу сличен со јазиците и системите за обработка на податоци, дизајнирани за архитектури, како што се MapReduce. Моделот на податоци на JAQL поддржува делумна шема, која помага во транзиција на скриптите од истражувачки во производствени фази на анализа. Особено, корисниците можат да содржат такви промени во дефинициите на шемата без потреба да ги изменат постојните прашања или да ги реорганизираат податоците.

5.5.6 Flume

Apache Flume е високосигурна и дистрибуирана услуга што може да се користи за автоматско собирање на огромни податоци за пренос од различни извори и пренесување во HDFS. Првично била развиена за да собира податоци за стриминг од веб дневник, но сега може да се користи за собирање сетови на податоци од различни извори и нивен пренос во HDFS. Flume е високосигурна, дистрибуирана и приспособлива алатка. Главно е дизајнирана да копира стриминг податоци (дневник на податоци) од различни веб-сервери на HDFS. Архитектурата на Flume главно вклучува извор, посредник кој ги доставува податоците до HDFS, канал што ги поврзува изворот и посредникот и агент.

Flume е рамка која се користи за преместување на дневникот податоци во HDFS. Општо земено, настаните и податоците за дневникот се генерираат од серверите за евиденција и овие сервери имаат агенти Flume што работат на нив. Овие агенти ги примаат податоците од генерирачките на податоци. Податоците во овие агенси ќе се соберат со среден јазол познат како *колектор*. Исто како агентите, може да има повеќе колектори во Flume.

5.5.7. Chukwa

Chukwa е систем за собирање податоци изграден на врвот на Hadoop. Целта на Chukwa е да следи големи дистрибуирани системи. Таа користи HDFS за да собира податоци од различни даватели на податоци и MapReduce за да ги анализира собраните податоци. Ја наследува приспособливоста и робусноста на Hadoop. Обезбедува интерфејс за прикажување, следење и анализирање на резултатите. Chukwa нуди флексибилна и моќна платформа за големи податоци. Тоа им овозможува на аналитичарите да соберат и анализираат групи на големи податоци, како и да ги следат и прикажуваат резултатите. За да се обезбеди флексибилност, Chukwa е структурирана како процес од собирање, фази на обработка, како и дефинирани интерфејси помеѓу фазите.

6. Едукативно податочно рударење

Главната цел на образовните системи е обезбедување знаење и вештини за студентите да ја создадат својата идна кариера во одреден период. Начинот на кој образовните системи ја реализираат оваа цел е од големо значење за социјалниот и за економскиот напредок. Проценките за тешкотиите и проблемите играат важна улога во широк спектар на образовните системи, вклучително и одредување на редоследот на проблемите презентирани на студентите и толкување на добиените одговори. Точноста на овие метрики е важна, бидејќи тие можат да ја одредат релевантноста на образовното искуство. За образовните системи кои содржат големи количини на сурови податоци, овие набљудувања може да се користат за да се направи проценка на перформансите на учење на студентите, во кој правец и насока тие се одвиваат, да се направи евалуација на материјалите за учење, но и да се детектираат нетипичните однесувања на студентите.

Едукативното податочно рударење [36] е нов тренд во областа на рударството на податоци, кој се фокусира во ископување корисни модели и откривање на корисни знаења од образовните информациски системи, како што се системи за прием, системи за регистрација, управување со курсеви и сите други системи кои се занимаваат со студенти на различни нивоа на образование, од училишта до колеџи и универзитети.

Анализата на податоците и информациите на учениците им овозможува да се класифицираат учениците, да се создадат дрва за одлучување или правила на асоцијација, да се донесат подобри одлуки, да се подобрат перформансите на учениците е интересно поле на истражување. Сето ова главно се фокусира на анализа и разбирање на образовните податоци на учениците што укажуваат на нивните образовни перформанси и генерира специфични правила, класификации и предвидувања за да им помогне на учениците во нивните идни образовни перформанси.

Проектите за едукативно податочно рударење се спроведуваат со цел да се откријат обрасци на релевантни и корисни информации во големи количини на податоци. Ова е направено со развој на четири фази, кои обично се:

- Филтрирање податоци;
- Избор на променливи;
- Извлекување знаење;
- Толкување и евалуација.

Класификацијата е еден од најчесто проучуваните проблеми од истражувачите од областа на едукативното податочно рударење и машинското учење. Се состои во предвидување на вредноста на категоричен атрибут и класата врз основа на вредностите на другите атрибути (атрибутите за предвидување). Постојат различни методи на класификација, како што се:

- Статистичката класификација е процедура во која одделни ставки се сместуваат во групи врз основа на квантитативните информации за карактеристиките својствени за ставките (наведени како променливи, знаци итн.).
- Дрвото за одлучување е збир на услови организирани во хиерархиска структура. Тоа е предвидлив модел во кој инстанцата се класифицира со следење на патот на задоволени услови од коренот на дрвото до достигнување на лист, што ќе одговара на ознаката на класата. Дрвото за одлучување лесно може да се претвори во збир на правила за класификација.
- Индукцијата на нејасни правила применува нејасна логика со цел јазично да се протолкуваат основните податоци. За целосно да се опише нејасен систем треба да се одреди правилна основа (структура) и нејасни партиции (параметри) за сите променливи.
- Невронска мрежа, позната и како паралелно дистрибуирана мрежа за обработка, е компјутерска парадигма која е лабаво моделирана по кортикалните структури во мозокот. Се состои од меѓусебно поврзани елементи за обработка наречени јазли или неврони кои работат заедно за да произведат излезна функција.

7. Истражување

7.1. Систем за управување со учење (LMS)

Систем за управување со учење (Learning Management System - LMS) [37] е софтверска апликација или веб-базирана технологија која се користи за планирање, имплементација и оценување на одреден процес на учење. Се користи за практики за е-учење и во најчеста форма се состои од два елементи: сервер кој ја извршува основната функционалност и кориснички интерфејс што го управуваат инструктори, професори и администратори. Како систем што се состои од низа функционалности, LMS може да се примени во институции како што се училиштата и големите компании. Негова цел е да им обезбеди на класот, институцијата или компанијата централизирана средина за учење базирана на компјутер во најкус можен рок, без оглед на нивната улога во одредена институција, нивното претходно знаење и многу други фактори од кои биле зависни дотогаш.

Обично, системот за управување со учење му обезбедува на инструкторот начин да создаде и испорача содржина, да го следи учеството на студентите и да ги процени перформансите на студентите. Системот за управување со учење, исто така, може да им обезбеди на студентите способност да користат интерактивни карактеристики, како што се групни дискусии, видеоконференции и форуми за дискусија. Врз основа на параметрите потребни за следење на процесот на учење снимен од LMS, можно е во секое време да се следи напредокот на индивидуален вработен, ученик или група и на крајот од образовниот процес сигурно да се измерат и анализираат индивидуалните перформанси.

7.2. Moodle

Moodle [38] претставува онлајн платформа за учење на далечина, дизајнирана за збогатување на искуството и знаењето на студентите, давајќи им пристап до материјали за учење, курсеви, активности, испити, тестови итн., а сето тоа преку интернет. Moodle е платформа дизајнирана да им обезбеди на предавачите, администраторите и учесниците (студентите) безбеден и интегриран систем за создавање персонализирана средина за учење. Наменета за поддршка и

на наставата и на учењето, под раководство на педагогијата за социјална конструкција, во текот на повеќе од 10 години развој, платформата Moodle стана заедничка околина што ги зајакнува наставата и учењето. Предноста на Moodle е тоа што е бесплатен софтвер со отворен извор и секој може да го преземе, прошири или модифицира за комерцијални и некомерцијални проекти без надоместок за лиценцирање. Moodle обезбедува најфлексибилен сет на алатки кои поддржуваат мешано учење и 100% курсеви преку интернет. Посветеноста кон заштитата на податоците и приватноста на корисникот се изразува преку постојано ажурирање и имплементација на компонентите задолжени за заштита од неовластен пристап, загуба на податоци и злоупотреба. Со стандардниот интерфејс, овозможена е компатибилност со мобилни уреди, разни веб-прелистувачи и разни оперативни системи.

7.2.1. Moodle за време на Covid -19

Covid-19 е акроним за коронавирусна болест. Вирусот е откриен кон крајот на 2019 година во кинеската провинција Вухан и за само неколку месеци доведе до прогласување на пандемија од Светската здравствена организација (СЗО). Вирусот Covid-19 се прошири низ целиот свет, зафаќајќи ги скоро сите земји и територии [39]. Првично се сметаше дека вирусот е фатален само за постарата популација, но денес се смета дека е фатален за сите луѓе со ослабен имунолошки систем и придружни болести, без оглед на нивната возраст. Поради тоа една од препораките за заштита од вирусот е социјалната дистанца. Мерките за социјално дистанцирање поради пандемијата Covid-19 доведоа до затворање на училишта, институти за обука и високообразовни објекти во повеќето земји.

Пандемијата на вирусот Covid-19 имаше сериозен ефект врз промените во образовните процеси ширум светот. Пандемијата официјално беше прогласена во март 2020 година, отворајќи глобална криза во сите области, вклучително и во образованието. Глобалната пандемија Covid-19 паралелно со глобалната здравствена криза создаде криза во образованието, при што беше препорачано учење преку интернет и одржување на социјална дистанца. Тоа предизвика промена на наставата, од часови со физичко присуство во онлајн часови, без

доволно време за планирање и подготвување виртуелни образовни програми. Наставниците ги реконструираа своите образовни планови и развиваа вештини за настава во виртуелна средина.

Платформата Moodle го олесни пристапот до содржините за учење, создавање курсеви, форумска комуникација, задачи и многу други активности кои придонесуваат за создавање на потполно нова слика за процесот на учење. Овој систем за управување со материјали за учење стана широко користен во образованието, од основните училишта до универзитетите, со што стана ново секојдневие во образовниот процес со цел да се решат последиците од затворањето на образовните институции. Целта на Moodle може да се сфати како додаток на традиционалното предавање во училница, но исто така и за создавање динамични заедници за учење преку интернет, како и за создавање на голем број курсеви во еден единствен систем.

Бидејќи методот на традиционално учење мигрираше на учење преку интернет, а бесплатните веб-платформи станаа достапни за студентите и професорите, како што се: Microsoft Teams, Zoom, Skype и Google Classroom, од почетокот на новата учебна година, 1 октомври 2020 година, наставата на факултетите и колеџите продолжи да се спроведува само електронски. Ова овозможи студентите и професорите успешно да ја завршат студиската година на поразличен начин од претходните години користејќи ја платформата за електронско учење Moodle.

Иако некои од недостатоците на онлајн образованието вклучуваат сложена употреба на онлајн платформи и софтвер и потреба од технолошко знаење, стекнувањето на нови знаења преку интернет, ова не го смени принципот на образованието. Образовниот процес сепак бараше од студентите да ги извршуваат своите должности, како што се домашни работи, семинарски работи и полагање испити.

Интерактивната врска помеѓу професорите и студентите и сите модули како што се квизови, задачи, анкети и лекции се само дел од активностите што овозможија реализација на образовните процеси. Ова ја зголеми употребата на

платформата Moodle. Таа овозможува пристап до предавања на професорите и нивно следење од страна на студентите. Преку Moodle платформата, стручната литература им беше достапна на студентите и во електронска форма, а професорите континуирано снимаа и изработуваа дополнителни материјали за да ги ажурираат курсевите.

7.2.2. Кориснички улоги на Moodle

Улогите на платформата Moodle ги претставуваат нивоата на уредничките права, односно секоја улога има одредени одобрувања и ограничувања во својата работа (освен улогата на администратор). Основните достапни улоги и нивното објаснување се наведени подолу.

Менаџер - Практично ја претставува административната улога, но дејството е ограничено само на одреден курс во кој е активен, односно во кој е запишано лицето што ја држи сметката со оваа улога. Оневозможен е пристап и уредување на курсеви на лице кое не е запишано со улога „менаџер“.

Наставник - Корисниците со улогата „наставник“ можат да ги уредуваат своите курсеви, да додаваат наставни материјали, да оценуваат студенти, да гледаат статистички податоци за курсеви, да испраќаат е-пошта итн.

Асистент - Корисниците со улога на „асистент“ имаат идентичен авторитет на курсевите и исто така можат да ги уредуваат своите курсеви, да додаваат наставни материјали, да додаваат студенти, да гледаат статистика на курсеви, да испраќаат е-пошта итн.

Студент - Корисниците со улога на „студент“ можат да ги прегледаат курсевите на кои се запишани, да прегледуваат наставни материјали, да решаваат тестови за знаење, да користат алатки за комуникација и соработка, да испраќаат документи (на пр. семинарски работи) итн.

Регистриран гостин – Овие корисници имаат исти овластувања како студентите кога станува збор за преглед на наставните материјали, но тие не можат да прават тестови, да користат алатки за комуникација или да испраќаат какви било документи кои се поврзани со едукативниот процес.

7.2.3. Активности на Moodle

Форум (Forum) - Им овозможува на учесниците да водат асинхрони дискусии, односно да разговараат за различни теми за подолг временски период. Постојат неколку видови на форуми за избор, како што е стандардниот форум каде што секој може да започне нова дискусија во секое време, форум каде што секој студент може да иницира само една дискусија или форум „Прашање и одговор“ каде што студентите прво мора да го објават својот одговор пред да можат да ги видат одговорите на другите студенти. Предавачот може да им овозможи на студентите да достават приложени датотеки со своите пораки. Приложените датотеки се прикажуваат во пораката.

Избор (Choice) - Модулот за активност на избор му овозможува на предавачот да постави прашање и да им понуди на студентите избор помеѓу неколку можни одговори. Резултатите од изборите можат да бидат објавени откако студентите ќе ги дадат своите одговори или по одреден датум и тие можат да останат необјавени. Исто така, резултатите можат да бидат објавени со имиња на студенти или анонимно. Изборот може да се искористи: како брзо истражување за поттикнување размислување за некоја тема, за брзо тестирање на студентите за да видат колку добро разбираат одреден материјал или за да им се олесни на учениците да донесуваат одлуки, на пр. им овозможува на студентите да гласаат за понатамошниот тек на курсот, изборот на теми за семинарска работа итн.

Лекција (Lesson) - Модулот за лекција му овозможува на професорот да ја презентира содржината или практичните активности на интересен и флексибилен начин. Професорот може да ја искористи лекцијата за да создаде линеарна серија од страници со содржина или упатства што нудат различни патеки или опции за ученикот. Во зависност од избраниот одговор, како и од тоа како професорот го организирал часот, студентите можат да напредуваат на следната страница, да бидат вратени на претходната страница или пренасочени на друга страница од лекцијата. Лекциите може да се оценуваат, а оценките се внесуваат во книгата за оценки. Лекцијата може да се искористи: за независно учење на нови содржини, за сценарија или симулации, вежби за донесување одлуки, за корекција на наученият

материјал, со различни групи прашања во зависност од одговорите дадени на почетните прашања.

Соба за разговор (Chat) - Им овозможува на учесниците да водат дискусии за синхрони текстови во реално време. Собата за разговор може да биде еднократна активност или може да се повторува во исто време секој ден или секоја недела. Сесиите за разговор се зачувани и можат да бидат достапни на секого за прегледување или ограничени на оние корисници кои имаат можност да ги видат дневниците на сесиите. Собата за разговор им овозможува на студентите да разговараат со професорите и на тој начин да не заостануваат зад материјалот, за да им помогне на студентите да се подготват за тест каде што професорот или другите студенти ќе можат да поставуваат примери на прашања слични на оние што студентите ќе ги имаат на тестот.

Речник (Glossary) - Им овозможува на учесниците да креираат и одржуваат списоци со поими и нивните дефиниции или да собираат и организираат ресурси и информации. Професорите може да дозволуваат датотеки да се додадат на поимите во речникот како прилог. Приложените датотеки се прикажуваат во рамките на дефиницијата за поимот. Поимите во речникот може да се пребаруваат или прелистуваат по азбучен ред или по категорија, датум или автор. Внесените термини можат автоматски да се одобрат за објавување или професорот мора да ги одобри пред да станат видливи за секого. Ако е овозможен филтерот за автоматско поврзување на речникот, терминте автоматски ќе се поврзат каде и да се појават зборови или фрази од терминот (или клучни зборови што дополнително го дефинираат терминот) во текот на курсот. Професорот може да дозволи коментирање на поимите, како и нивно оценување од страна на професорот или асистентот, како и од страна на студентите (оценување од врсници). Речниците може да се користат на различни начини, како што се: заеднички креирани клучни термини во банка, простор за средби каде што новите студенти додаваат свои имиња и лични информации, „корисни совети“ - каталог на примери за најдобра практика на различни теми, простор во споделување на корисни видеоклипови, слики или аудио датотеки, каталог на факти што треба да ги запомните.

Задачата (Assignment) - Модулот за задачи дава можност професорот да им задава задачи на студентите, да ги собира нивните дела, да ги оценува, како и да им испраќа повратни информации. Кога прегледуваат задачи, професорите можат да остават коментари, повратни информации и да објавуваат датотеки, како што се прегледување и обележување на студентски трудови со коментари, посебни документи со коментари или повратни информации. Задачите се оценуваат на нумеричка или приспособена скала за оценување. Завршните оценки се запишуваат во книгата за оценки. Овој модул може да се искористи така што студентите, како нивна работа, можат да предаваат разни дигитални содржини (датотеки), како што се текстуални документи, табеларни пресметки, слики, презентации или аудио и видеоклипови. Алтернативно, задачата може да бара од студентите да го внесат текстот директно во уредникот за текст. Задачата може да се користи и како потсетник на студентите за „вистинската“ задача што треба да ја извршат офлајн, надвор од страницата, без да бараат од нив да достават каква било дигитална содржина. Студентите можат да го предаваат трудот индивидуално или како членови на група. Модулот за задачи најчесто се користи за предавање и оценување на семинарски работи.

Квиз (Quiz) - Им овозможува на професорите да дизајнираат тестови кои се состојат од различни типови прашања, вклучувајќи повеќе избор, совпаѓање, краток одговор, да-не прашања, нумерички прашања итн. Професорот може да дозволи тестот да се реши повеќе пати, со мешан редослед на прашања или случајно избрани прашања од банката, при секој обид. Исто така е можно да се постави временско ограничување за решавање на тестот. Секој обид се оценува автоматски, со исклучок на есејскиот тип на прашање, а оценката се запишува во книгата за оценки. Професорот може да избере дали и кога на студентите ќе им бидат прикажани совети, повратни информации и точни одговори. Тестовите може да се користат: како завршни испити, како мини-тестови откако студентите ќе имаат задача да прочитаат одреден текст, или на крајот од некоја тема (област), како подготовка за завршниот испит, користејќи прашања од претходните испити, за да обезбедат непосредна повратна информација за достигнувањата, за независна проценка на знаењето итн.

Директориум (Directory) - Модулот папка им овозможува на предавачите да прикажуваат бројни, логички меѓусебно поврзани датотеки во рамките на еден директориум. На овој начин се намалува потребата од движење на насловната страница на курсот или лекцијата, односно да се пополни со бројни датотеки.

Книга (Book) - Му овозможува на професорот да создаде ресурс, односно едукативен материјал со поголем број страници, во формат сличен на книга, со повеќе поглавја. Книгите можат да содржат мултимедијални датотеки, како и текст, и се корисни за прикажување подолги делови на едукативен материјал, кои можат да се поделат на повеќе делови. Книгата може да се користи за: презентација на едукативна содржина за одделни предмети, како прирачник за студенти или како репрезентативно портфолио на студентска работа.

Натписот (Inscription) - Модулот за наслови овозможува вметнување на текст и мултимедијална содржина на насловната страница на курсот или како поднаслов во рамките на лекцијата помеѓу врските кон други активности и ресурси. Написите се многу флексибилни и ако се користат внимателно може да помогнат да се подобри изгледот на насловната страница на курсот и лекциите за да им се олесни навигацијата на студентите. Може да се користат натписи за: да се оддели долг список на активности по наслов или слика, да се прикажат вградени звучни датотеки, видеа и слично.

7.4. Фази на истражувачката работа

Фази на истражувачка работа:

1. Собирање на податоци;
2. Препроцесирање и трансформација на собраните податоци и креирање на податочен сет погоден за понатамошна обработка;
3. Обработка на добиените податоци;
4. Евалуација и анализа на добиените резултати;
5. Донесување заклучоци.

7.4.1. Собирање на податоци

Примарната цел на ова истражување е да се анализира бројот на активности на корисниците на Moodle платформата пред и по пандемијата. За реализација на истражувањето коешто е опфатено во овој магистерски труд се користени податоци од базата на податоци на системот за електронско учење на Универзитет „Гоце Делчев“ - Штип којшто е базиран на Moodle платформата. Системот има MySQL база којашто има повеќе од 300 табели, кои содржат податоци почнувајќи од 2012 година. За ова истражување беа користени податоци од соодветни табели од 2019 и 2020 година. Тоа претставуваат годините кои ги содржат потребните информации за реализација на истражувањето. За реализација на првобитната идеја на ова истражување најпрво беше потребно да се обезбедат податоците кои беа потребни за понатамошна анализа и обработка. Од базата на податоци на Moodle платформата беше потребно да се направи селекција за потребниот временски период. Во текот на собирањето на податоци неопходно беше да се познаваат особеностите на податоците кои се добиваат од едукативните системи, а исто така да се познаваат и информациите кои може да доаѓаат од различни извори на податоци.

7.4.2. Препроцесирање и трансформација на собраните податоци и креирање на податочен сет погоден за понатамошна обработка

Обезбедените податоци во првата фаза, собирање на податоци, вообичаено не се добиваат во форма која е погодна за понатамошна анализа на големите податоци. Податочниот сет добиен по оваа фаза треба да се подготви за понатамошна анализа. Добиените резултати ги претставуваат посебно активностите за 2019 година и посебно за 2020 година, па за таа цел најпрво беше потребно да се издвојат податоците од базата според годините. Најпрво податоците беа издвоени според временскиот период кој беше потребен за анализа, односно според годините 2019 и 2020. Од Moodle базата на податоци која содржи огромен број на табели, за овој магистерски труд беа користени табелите кои се прикажани во табела 1.

Табела 1. Избрани табели за обработка и подготовка на податочниот сет
Table 1. Selected tables for processing and preparation of the data set

mdl_logstore_standard_log	Сите активности на корисниците-логови
mdl_users	Информации за сите корисници на Moodle
mdl_role	Назначување на улогата на професор, асистент, студент
mdl_role_assignment	Назначена улога за сите корисници

Во оваа фаза која ги опфаќа препроцесирањето и трансформацијата на собраните податоци, многу важен дел е правилниот избор на табели. Оваа почетна активност за текот на истражувачката работа и добивањето на потребните резултати може да се каже дека е од огромна важност, бидејќи изборот на погрешни табели може да влијае на добивање на погрешни информации за корисниците, а со тоа да се отежни процесот на истражување.

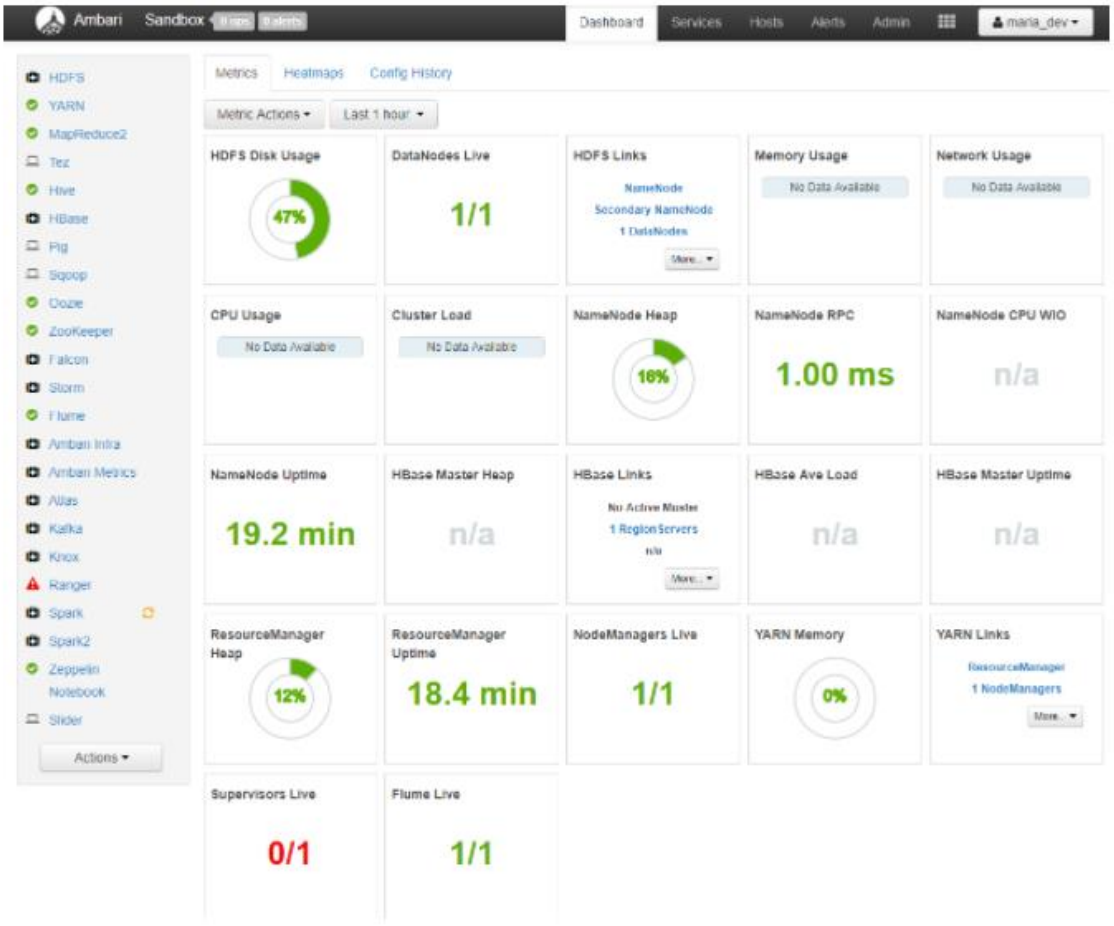
Најголем број на податоци во текот на истражувачката работа беа искористени од табелата *mdl_logstore_standard_log* која содржи записи за сите активности на корисниците. Испитувањето на податоците се изврши со поврзување на потребните табели и задавање на соодветни упити за на крај да се добијат потребните информации. Важен дел од подготовката на податочниот сет за анализа претставува поврзувањето на потребните табели од базата на податоци и извршување на соодветни упити заради извлекување на потребните податоци.

Вчитаните табели се прикажуваат во Hue делот како Hive табели. Подетално објаснување за начинот на пренесување на табелите е објаснет понатаму во текот на магистерскиот труд, каде што е прикажана и обработката на добиените податоци.

7.4.3. Обработка на добиените податоци

За анализа на активностите на корисниците на Moodle беше користена платформата Hortonworks. Како што беше спомнато и претходно, станува збор за

платформа која е наменета за управување со големи податоци што обезбедува централизирана архитектура за водење индиректни, интерактивни апликации во реално време, паралелно со дистрибуирани множества на податоци. Платформата поддржува сеопфатен сет на алатки кои се однесуваат на основните барања за безбедност, деловно работење и управување со податоци, кои придонесуваат за успешна обработка и анализа на податоците. Платформата за податоци Hortonworks Sandbox беше преземена од официјалната веб-страница. Единствениот предуслов беше претходно да имаме инсталирано VirtualBox, VMware или некој друг софтвер за виртуелизација. За потребите на ова истражување беше инсталиран VirtualBox. Најавувањето на Hortonworks се извршува преку веб-интерфејс, каде што беше користен Ambari. На слика 14 во продолжение е прикажана работната околина на Ambari сервисот.

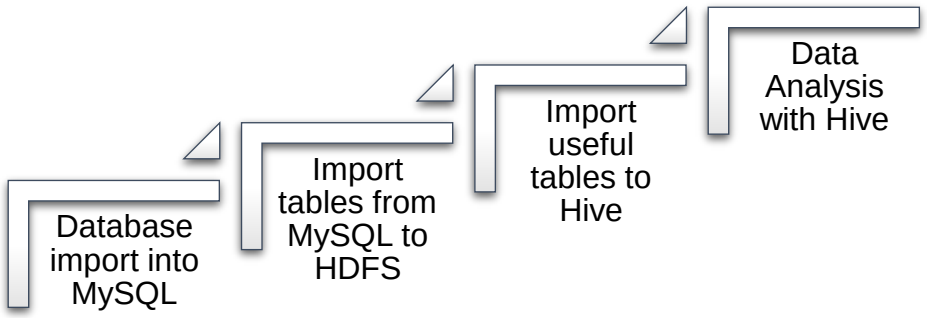


Слика 14. Работна околина на Ambari
Figure 14. Ambari desktop

Од платформата Hortonworks беа користени алатките Sqoop и Hive за да се добијат потребните резултати, а потребните табели беа импортирани во делот на Hive.

Со помош на Sqoop алатката беше извршено увезување на големите податоци од релациониот систем за управување со бази на податоци (RDBMS) во системот за дистрибуирани датотеки на Hadoop (HDFS).

Apache Hive е алатката којашто беше користена за складирање и обработка на големата количина на податоци. За текот на понатамошната анализа беа користени HiveQL упити, јазик за пребарување сличен на SQL за пребарувања на големи податоци.



Слика 15. Графикон на тек на сите активности за анализа на податоците
Figure 15. Graph of all data analysis activities

На слика 15 можеме да го видиме дијаграмот на проток на сите активности кои беа извршени за анализа на податоците од Moodle. Во следниот чекор, користејќи ја алатката Sqoop, се внесени избраните табели со бази на податоци од MySQL во HDFS. Пренесувањето на податоци беше извршено со кодот испишан во командна линија којшто е прикажан во продолжение:

```
sqoop import --connect jdbc:mysql://localhost:3306/moodle --username root -P --table mdl_assign_grades --split-by id -m 1 --hive-import --mysql-delimiters;
```

На тој начин импортираните табели беа складирани во Hive и беа подготвени за понатамошна обработка и анализа. Неминовен дел од подготовката на податочниот сет потребен за анализа на податоците претставува поврзувањето на

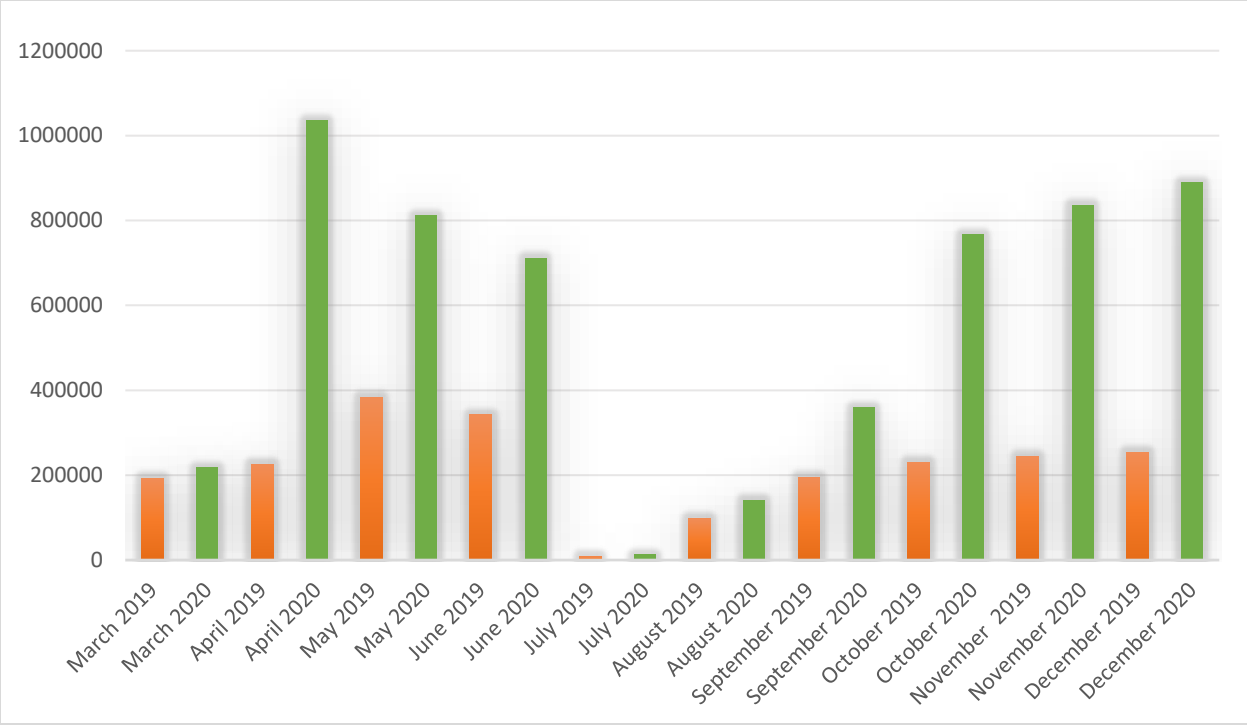
веќе избраните табели од базата на податоци и извршување на соодветни упити врз нив.

Преку задавање на посебни HiveQL упити се изврши обработка на табелите, со што беа добиени потребните резултати, кои ги даваа информациите и податоците потребни за истражувањето. Добиените податоци најпрво беа издвоени според годините, со цел да се добијат првичните резултати. По издвојувањето на податоците според годините беше извршена анализа на месечните активности за да се утврди дали постои разлика во добиените податоци и доколку постои за колку пати е зголемена или намалена истата. Од месечните активности е направена збирна табела, преку која може да се забележат разликите помеѓу активностите на годишно ниво. По извршените годишни и месечни анализи, со посебни HiveQL упити беше направено испитување за вкупните активности на наставничкиот кадар и студентите, како и нивните активности во посебните модули кои се содржат во Moodle базата.

7.4.4. Евалуација и анализа на добиените резултати

Фокусот на истражувањето е ставен на бројот на активностите на корисниците на Moodle платформата. Видот на корисници на Moodle кои се од интерес за истражувањето се наставничкиот кадар (професори и асистенти) и студентите. Графиконот на слика 16 ги претставува првичните резултати кои се добиени од извршената обработка и анализа на податоците кои веќе беа подготвени во претходната фаза. Графиконот прикажува паралелна споредба помеѓу месечните активности на платформата Moodle од март 2020 година со појавата на пандемијата и затворањето на универзитетите и 2019 година, кога образовниот процес се одвиваше со физичко присуство. За секој од месеците посебно се претставени добиените резултати, а со тоа се има увид каде постои најголема разлика помеѓу вкупните активности на месечно ниво.

Мора да се напомене дека овде не е извршена селекција на посебните активности на модулите, односно резултатите ги прикажуваат вкупните активности на професорите и на студентите.

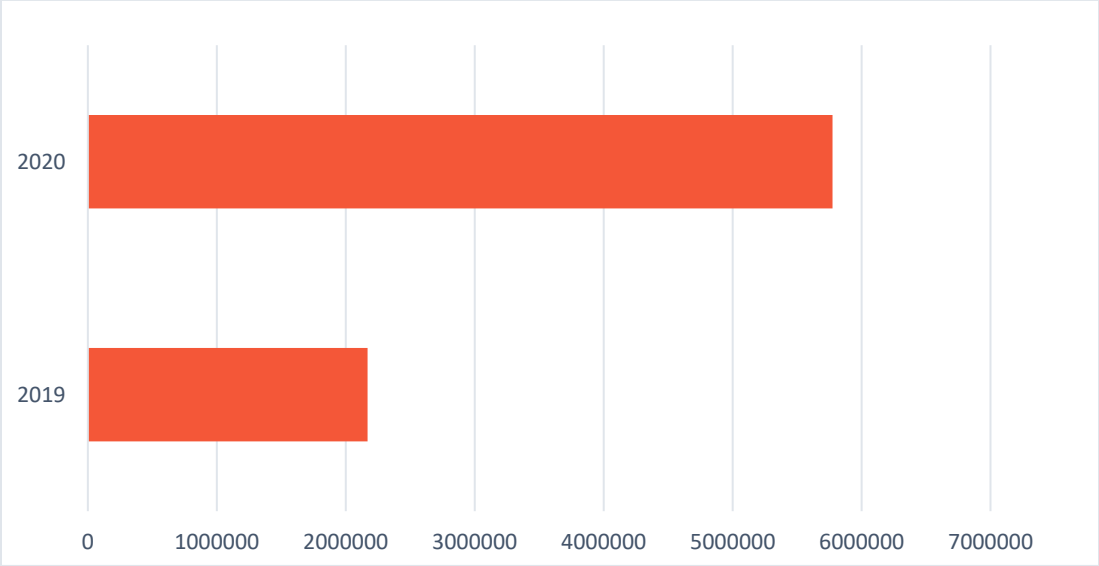


Слика 16. Разлика помеѓу активностите на Moodle во периодот пред пандемијата и за време на пандемијата
Figure 16. Difference between Moodle activities in the pre-pandemic period and during the pandemic

Март 2020 година беше земен како клучен почеток на пандемијата и исто така претставува почеток на нов начин на функционирање на образовниот процес. Споредувајќи ги месечните активности може да се забележи дека најголемата е разликата помеѓу април 2020 година и април 2019 година. Може да се забележи дека активностите се зголемени над 1.000.000 во април 2020 година, во однос на април 2019 година. Април го бележи најголемиот раст на користење и употреба на Moodle платформата, бидејќи воедно го претставува почетокот на електронскиот начин на функционирање на образовниот процес. Корисниците на Moodle платформата поминуваа повеќе време користејќи ги модулите и извршувајќи задачи, бидејќи адаптирањето и започнувањето на нов тек на образовниот процес бараше од нив да креираат и создадат нов начин на употреба на едукативните материјали, а истите потоа да бидат оценети. Периодот помеѓу мај и јуни, исто така,

забележува пораст поради големиот број на активности за текот на академската година, вклучувајќи ги колоквиумите и испитните сесии. Месеците јули, август, септември и октомври не прават голема разлика поради крајот од претходната академска година и почетокот на новата академска година, односно поради годишните одмори. Во ноември и декември бројот на активности е исто така зголемен, бидејќи текот на пандемијата продолжи со ист интензитет како и на самиот почеток, па така академската година продолжи да се одвива по електронски пат.

Слика 17 покажува споредба на вкупните активности на корисниците на Moodle помеѓу 2019 и 2020 година. Овие резултати се добиени на тој начин што ги покажуваат збирните резултати од месечните активности. Што значи дека тие ги прикажуваат вкупните резултати за претходно направената месечна анализа.



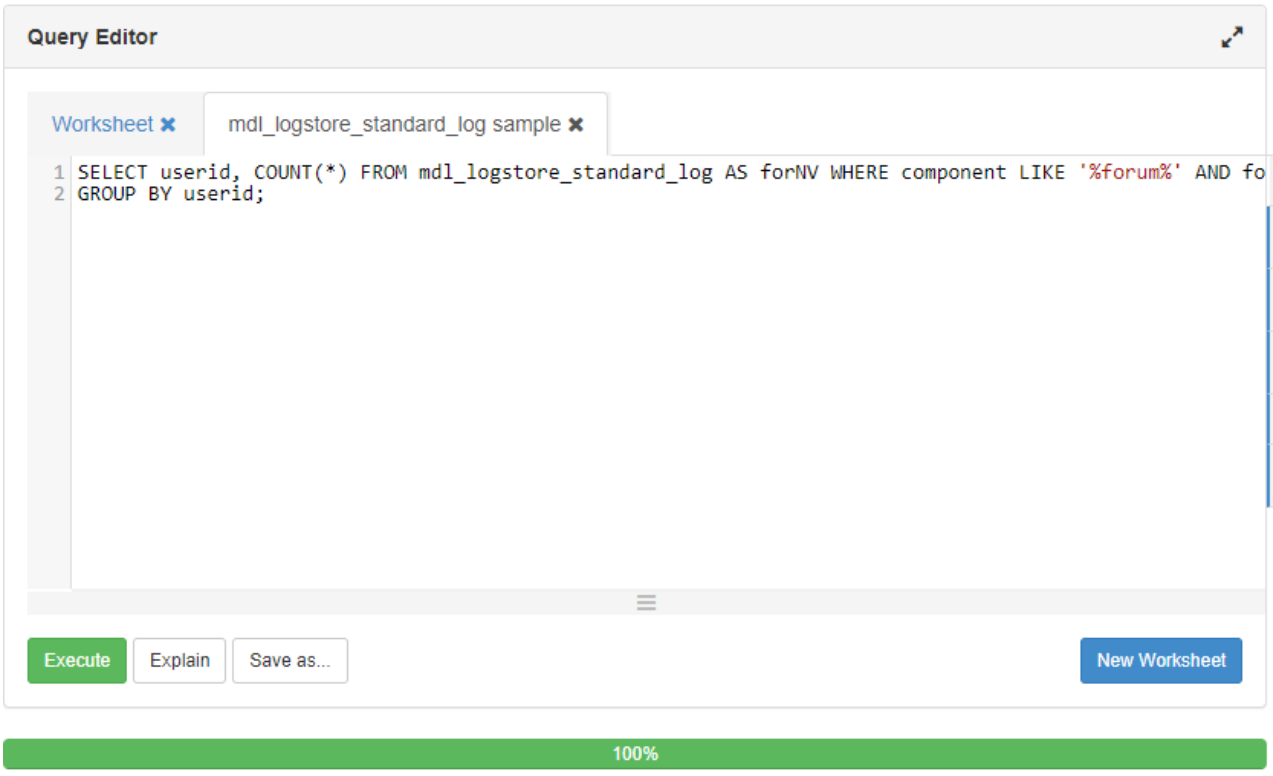
Слика 17. Разлика во однос на вкупната годишна активност на корисниците на Moodle
Figure 17. Difference in relation to the total annual activity of Moodle users

Може да се забележи дека во 2020 година има околу 3 пати зголемување на вкупните активности на корисниците.

Откако беше направена месечна и годишна анализа на вкупните активности, следниот чекор беше анализа на вкупните активности на Moodle за различни компоненти, како што се: *Forum, Quiz, Assignment, Choice, Book, Chat* и *Glossary.* Сите овие активности се содржани во табелата *mdl_logstore_standard_log*. Оваа

табела содржи записи за секоја активност на корисниците на Moodle. Од табелата со вкупните активности на корисниците е извлечен вкупниот број на активности на сите корисници за соодветните модули. Бројот на активности се однесува на корисниците кои имаат улога како наставнички кадар и студенти.

Со задавање на соодветни HiveQL упити за секој од избраните модули беше добиена одредена вредност, која ги претставуваше вкупните активности за истиот тој модул. На слика 18 е прикажана работната околина на Hive, каде што е впишан дел од кодот од извршените упити за добивање на активностите. Истиот принцип за поставување на упити беше поставен за добивање на резултатите за сите модули, со тоа што секој од упитите беше извршен посебно.



Слика 18. HiveQL упит за добивање на резултати за поединечни модули
Figure 18. HiveQL query for getting results on individual modules

Во табела 2 можеме да ги видиме конечните резултати и разликата во годишните активности за различни модули.

Табела 2. Разлика во активностите на Moodle за различни компоненти на годишно ниво

Table 2. Difference in Moodle's activities for different components on an annual basis

Year	*Forum*	*Quiz*	*Assign*	*Choice*	*Book*	*Chat*	*Glossary*
2020	250 380	2 030 404	1 706 870	10 818	9 325	75 876	6 503
2019	34 615	240 932	168 321	6413	7 398	6 128	4 932

Модулот Forum го означува бројот на активности на форумот и бројот на активности што вклучуваат активно учество во комуникацијата. Бидејќи овозможува повеќе начини на комуникација помеѓу корисниците, тој е еден од модулите со најголем раст за време на пандемијата, така што разликата пред пандемијата и со појавата достигнува дури над 200.000 активности.

Модулот Quiz обично содржи прашања и нивни одговори. Повеќето од професорите го зголемија начинот на оценување преку електронски квизови и тестови. Иако до претходната година електронските тестови беа актуелни, во 2020 година тие станаа неопходни за оценување на студентите и успешно завршување на семестрите. Сите овие опции овозможуваат избор на начинот на креирање на испити од страна на професорите, што придонесува за разлика од 1 800 000 активности.

Модулот Assignment содржи активности на професорите кои поставуваат задачи, односно тие поставуваат проблем за кој студентите се задолжени да најдат решение. На тој начин, професорите ја проценуваат активноста на студентите при извршувањето на домашните задачи. Физичкото присуство и домашните задачи сега се заменети со електронска домашна работа од сите видови. Затоа, можеме да забележиме дека бројот на активности во модулот за задачи се зголемил на 1 000 000.

Модул *Choice* - професорот поставува прашање и одредува избор на повеќе одговори. Можеме да забележиме зголемување на активностите затоа што може да биде корисно како брза анкета за стимулирање на размислувања за некоја тема,

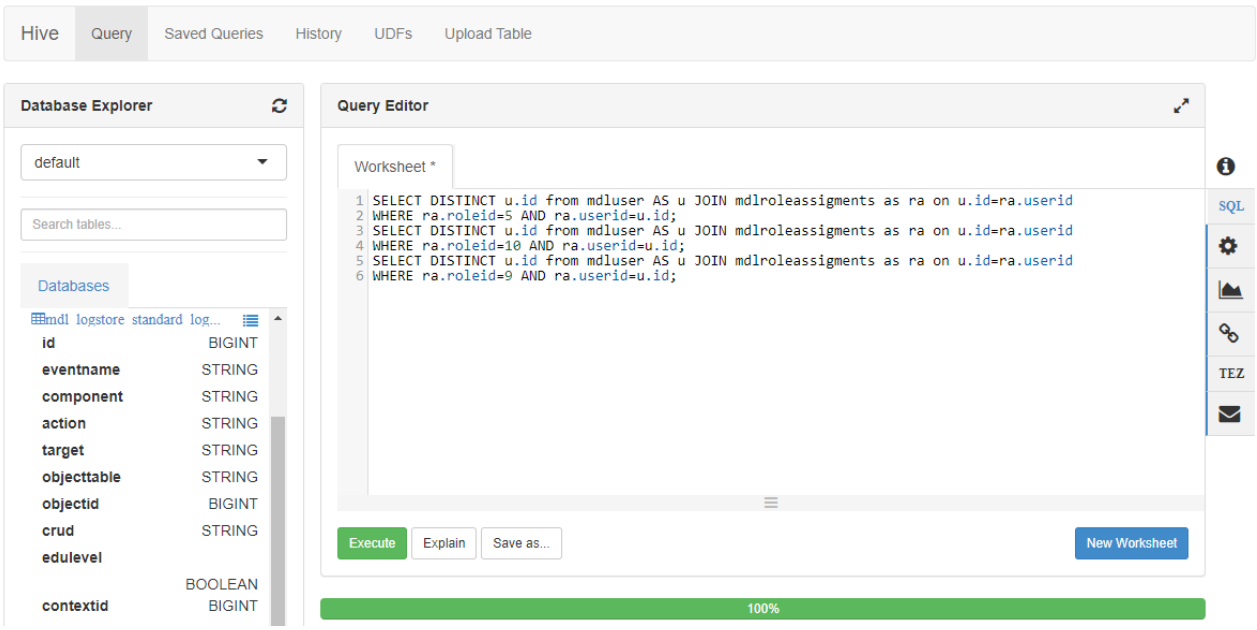
да се овозможи на часот да се гласа за насока за курсот или да се собере согласност за истражување. Може да се забележи дека има зголемување на бројот на активности за 4 000 активности повеќе во 2020 година, во однос на 2019 година.

Модулот *Book* има поголем број активности за 2020 година. Иако до претходната година многу од професорите користеа е-книги и материјали за учење, сепак во 2020 година може да се забележи дека има значително зголемување од 3 000 активности повеќе во 2020 година, за разлика од 2019 година. Поставувањето и користењето содржина за е-учење е неизбежен процес за студирање во време на пандемија.

Модулот *Chat* претставува број на реализирани разговори. Заради почитување на мерките и одржување на социјалната дистанца, меѓусебната комуникација помеѓу професорите и студентите е ограничена. Активностите во овој модул се зголемиле во 2020 година за 70 000 во споредба со 2019 година.

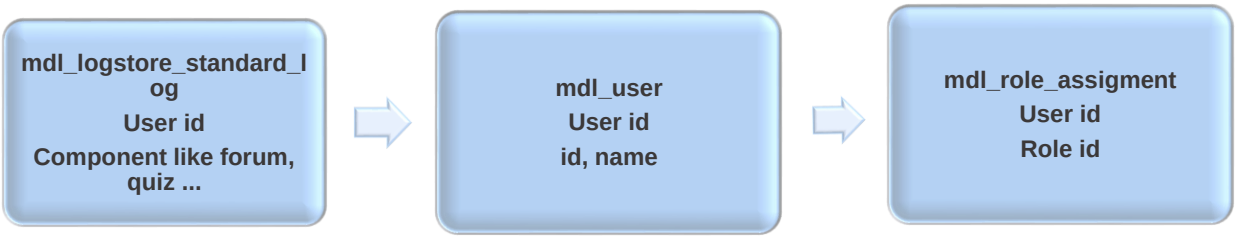
Како и кај претходните модули, зголемувањето на активноста може да се забележи и кај модулот *Glossary*. Овој модул им овозможува на корисниците да креираат и одржуваат список со изјави, како што е речник. Кај овој модул разликата во активностите помеѓу 2019 и 2020 година е околу 2 000.

Покрај вкупните активности на наставничкиот кадар и студентите, се анализирани и активностите посебно на наставничкиот кадар и посебно на студентите. За таа цел беа користени опциите кои ги нуди Hortonworks платформата, односно како што спомнавме претходно, беше користена алатката Sqoop со која истовремено се овозможува и креирање на табела во Hue делот на Hortonworks. Откако табелите беа креирани на нив се извршуваат потребните HiveQL упити со цел добивање на бараните податоци. Најпрво беше потребно да се издвојат оние корисници кои имаат улога на професори, асистенти и студенти. На слика 19 се прикажани кодовите од извршените упити со кои се врши селекција и извлекување на корисниците кои имаат улога на наставнички кадар и студенти.



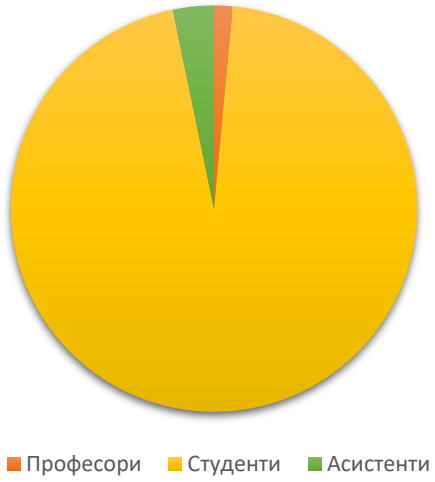
Слика 19. Hive делот од Hadoop со интерпретирани кодови за селекција на корисниците
Figure 19. Hive section of Hadoop with interpreted user selection codes

Најпрво беше потребно да се издвојат од табелата *mdl_user* корисниците кои имаат улога како наставнички кадар и посебно корисниците кои имаат улога на студенти. Збирните резултати за наставничкиот кадар се добиени, така што е извршена посебна селекција за професорите, а посебна за асистентите. Ова е направено со извршување на упит за извлекување на податоци од табелата која содржи информации за корисниците во системот *mdl_user,* како и табелата *mdl_role_assignment* која содржи информации за улогата на секој од корисниците. Овие две табели меѓусебно се поврзани со атрибутот *userid* (слика 20).



Слика 20. Блок дијаграм за добивање на сите активности на корисниците
Figure 20. Block diagram for obtaining all user activities

Вкупниот број на корисници кои се наставнички кадар, добиен од извршените испитувања, изнесува 376, со тоа што претставува вкупен резултат од асистентите и професорите, а вкупниот број на студенти изнесува 7 593. На слика 19 се претставени добиените резултати за корисниците на Moodle.



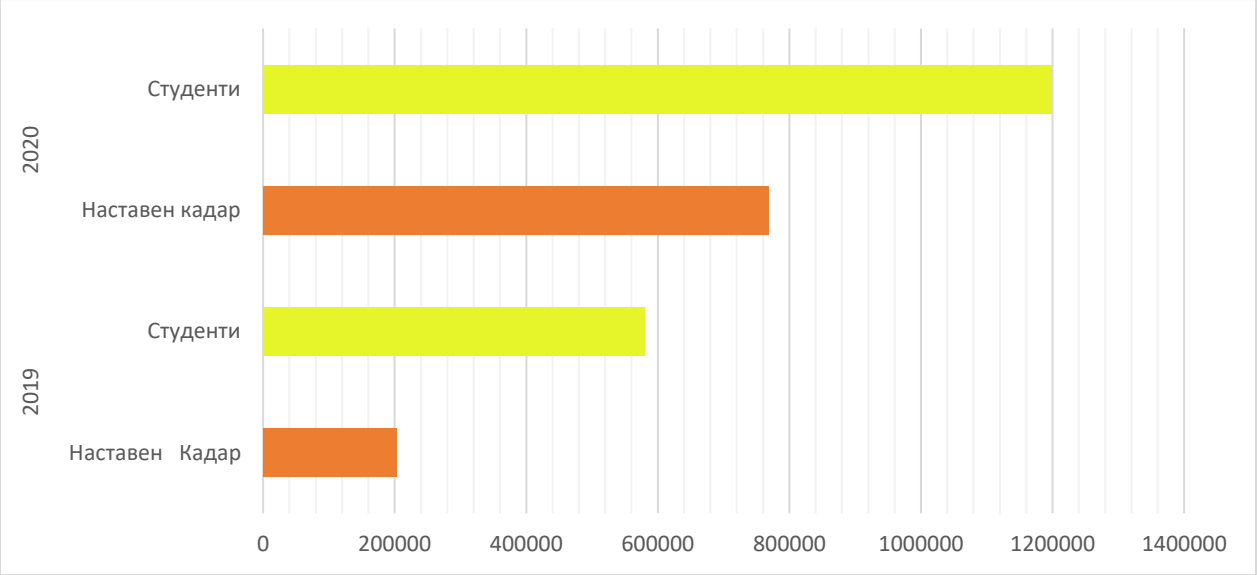
Слика 21. Корисници на Moodle
Figure 21. Moodle users

Откако беше добиен бројот за наставничкиот кадар и бројот за студентите, беше направена анализа на активностите кои се однесуваат посебно на наставничкиот кадар и посебно на студентите. Во *mdl_logstore_standard_log* табелата се наоѓаат записи кои се однесуваат само за професорите и само за студентите. Во добиените резултати беа вклучени дури и оние најавувања кои немаа никаква активност на некој од посебните модули на Moodle платформата. Резултатите, исто така, беа добиени од извршените анализи кои се однесуваат поединечно за 2019 и 2020 година. Добиените резултати се прикажани во табела 3.

Табела 3. Вкупни активности на наставнички кадар и студенти
Table 3. Total activities of teaching staff and students

Година	*2019*	*2020*
Наставнички кадар	203 120	768 020
Студенти	580 420	1 197 904

Од графиконот на следната слика може да се забележи дека во 2020 година вкупните активности на студентите достигнуваат скоро до вредност од 1 200 000, додека пак за наставничкиот кадар достигнуваат вредност од 800 000. За истиот период во 2019 година активностите на студентите достигнуваат вредност до 600 000, а на наставничкиот кадар до 200 000. Исто како и за вкупните активности на годишно ниво, може да се забележи дека и поединечно анализираните резултати на наставничкиот кадар покажуваат дека бројот на активности е зголемен до три пати. Слика 22 го претставува графиконот на вкупните активности на наставничкиот кадар и студентите.



Слика 22. Вкупни активности на наставнички кадар и студенти
Figure 22. Total activities of teaching staff and students

Од вкупните активности на наставничкиот кадар и студентите преку HiveQL упит беа извлечени активностите за посебните модули: *Forum, Choice, Assignment, Quiz, Book, Chat* и *Glossary.* Дополнително, заради зголемување на брзината на процесирање како и редуцирање на комплексноста на упитите, пребарувањата и анализата на табелите беа направени на тој начин што со неколку поедноставни упити користејќи ја табелата *mdl_logstore_standard_log* беа создадени привремени помали помошни табели. Во продолжение е прикажан дел од кодот од упитот со кој се врши спојување на секоја од помошните помали табели во кои се наоѓаат информации од записите за вкупниот број на активностите:

SELECT a.userid, a.firstname, a.faculty, f.for1, forNV, ch1, chNV, ass1, assNV, qui1, quiNV, bo1, boNV, cha1, chaNV, glo1, gloNV

FROM assprof a left JOIN forum for on (a.userid=f.userid) left JOIN forumNV for.NV on (a.userid=forNV.userid)

left JOIN choiche ch on (a.userid=ch.userid)

left JOIN choicheNV chNV on (a.userid=chNV,userid)

left JOIN assign ass on (a.userid=ass.userid)

left JOIN assignV assNV on (a.userid=assNV.userid)

left JOIN quiz qui on (a.userid=qui.userid)

left JOIN quizNV quiNV on (a.userid=quiNV.userid)

left JOIN book ass bo (a.userid=bo.userid)

left JOIN bookNV boNV (a.userid=boNV.userid)

left JOIN chat cha (a.userid=cha.userid)

left JOIN chatNV chaNV (a.userid=chaNV.userid)

gorup by a.userid, a.firstname, a.lastname, a.faculty, f.for1, forNV, ch1, chNV, ass1, assNV, qui1, quiNV, bo1, boNV, cha1, chaNV, glo1, gloNV, order by a.firstname desc;

Добиените табели ги содржат податоците за индивидуалните корисници за одредена активност на некој модул. Сите податоци во табелите се поврзани меѓусебе со атрибутот *userid,* којшто беше искористен за поврзување со новокреираните помошни табели. Добиените резултати од горенаведениот упит се претставени во .csv документ кој во себе содржи голем број на NULL вредности. Сите тие беа соодветно заменети и обработени на начин кој не би влијаел негативно врз крајниот резултат на извршената анализа.

Пребарувањата и анализите на селектираните табели од Moodle базата на податоци беа поделени на неколку поедноставни упити на новогенерираните

помошни табели, кои се значително помали и овозможуваат подобрување на перформансите и брзината на извршување на процесирањата на табелите. Со обработка на овие вредности беа добиени конечните табели, односно поттабели во кои се наоѓаат информации за вкупниот број на активности на корисниците кои се од интерес за истражувањето. Тоа претставува само дел од извршената обработка на податочниот сет за анализа на големите податоци.

Бидејќи станува збор за обработка на податоци за чија анализа е важен временскиот период, истата таа обработка беше извршена посебно за годините во наведениот период, потребен за истражувачката работа. Дополнително, бидејќи од интерес се посебните активности на наставничкиот кадар и студентите, текот на обработка и анализа на податоците се одвиваше во тој правец што податочниот сет беше посебно креиран за наставничкиот кадар, а посебно за студентите. Од извршените упити беа добиени две табели за 2019 и две табели за 2020 година, кои поединечно ги прикажуваат активностите на наставничкиот кадар и студентите. Секоја од табелите ги содржи модулите кои беа избрани и поединечено анализирани. Од нив беа извлечени и поединечно анализирани активностите за секој од модулите.

Поттабелите го претставуваат конечниот податочен сет, кој така подготвен понатаму е користен за прикажување на конечните резултати. Тие беа искористени за прикажување на резултатите за поединечните модули. Ги содржат потребните информации за активностите на наставничкиот кадар и студентите, кои ги претставуваат корисниците кои се предмет на анализата. За секоја од годините и корисниците беше добиена посебна табела. Во тој контекст, подолу во табела 4 и табела 5 се прикажани дел од поттабелите кои се добиени при обработка на крајниот податочен сет. Тие се добиени со анализи кои беа извршено посебно за временскиот период кој беше од интерес на истражувањето.

Табела 4. Вкупни активности на студентите
Table 4. Total student activities

Userid	for1	forNV	ch1	chNV	ass1	assNV	qui1	quiNV	bo1	boNV	cha1	chaNV
10546	24	null	null	null	38	16	null	null	null	null	null	45
10558	6	3	null	null	null	Null	null	null	null	null	null	null
10573	90	9	null	null	22	8	42	12	null	null	null	21
10593	654	null	null	null	62	22	35	8	null	null	null	67
10645	162	null	null	null	Null	Null	null	null	null	null	null	null
10646	47	null	null	null	null	Null	null	null	null	null	null	null
10653	28	null	null	null	95	23	null	null	null	null	8	null
10811	403	2	null	null	56	19	null	null	null	null	22	46
10822	68	34	null	null	10	Null	null	null	null	null	null	70
10855	27	null	null	null	null	Null	null	null	null	null	null	null

Табела 5. Вкупни активности на професорите
Table 5. Total activities of the professors

Userid	for2	for2NV	ch2	ch2NV	ass2	ass2NV	qui2	qui2NV	bo2	bo2NV	cha2	cha2NV
15117	8	null	null	null	74	10	null	null	null	null	4	54
15399	1024	840	null	null	895	223	null	null	null	null	null	null
1554	90	null	null	null	65	13	null	null	null	null	null	21
169	953	437	87	14	null	Null	null	null	null	null	null	12
1998	18	9	null	null	null	null	9	3	null	null	null	null
2008	97	13	74	56	null	null	null	null	null	null	null	null
20098	null	null	null	null	null	null	null	null	null	null	null	25
20880	234	123	null	Null	456	234	null	null	null	null	null	42
21927	22	64	null	null	null	null	null	null	19	13	null	19
22053	145	78	192	19	null	null	null	null	null	null	null	null

Добиените резултати се прикажани посебно за 2019 и 2020 година, со цел подетално да биде прикажана разликата помеѓу активностите во избраните модули.

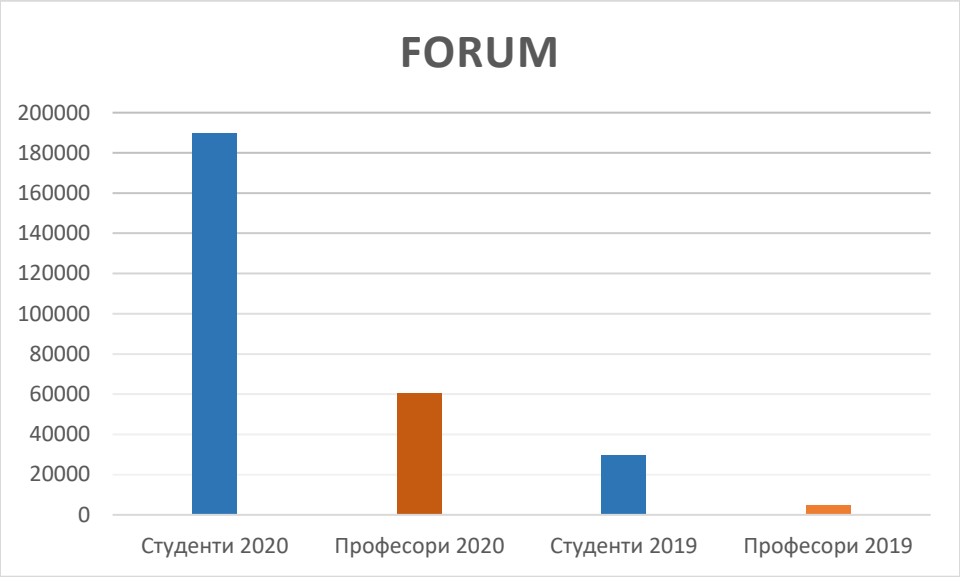
Крајната цел на овој магистерски труд е да се прикажат резултатите за поединечните активности на наставничкиот кадар и студентите и да се заклучи дали и колкава е разликата пред и со појавата на пандемијата. Во табела 6 се прикажани резултатите кои содржат информации за максималната вредност за бројот на активности на секој од модулите. Резултатите, исто така, се прикажани во посебни колони за 2019 и 2020 година и е претставена посебна селекција на корисниците за временскиот период, односно годините. Може да се забележи дека исто како и добиените резултати за вкупните активности на сите корисници и овде во 2020 година има скоро до три пати зголемување на активностите. Најголема разлика може да се забележи кај модулите *Forum, Quiz* и *Assignment*, додека пак помала разлика може да се забележи кај модулите *Book* и *Glossary.*

Табела 6. Резултати од активностите на корисниците за избраните модули
Table 6. Results from the activities of the users for the selected modules

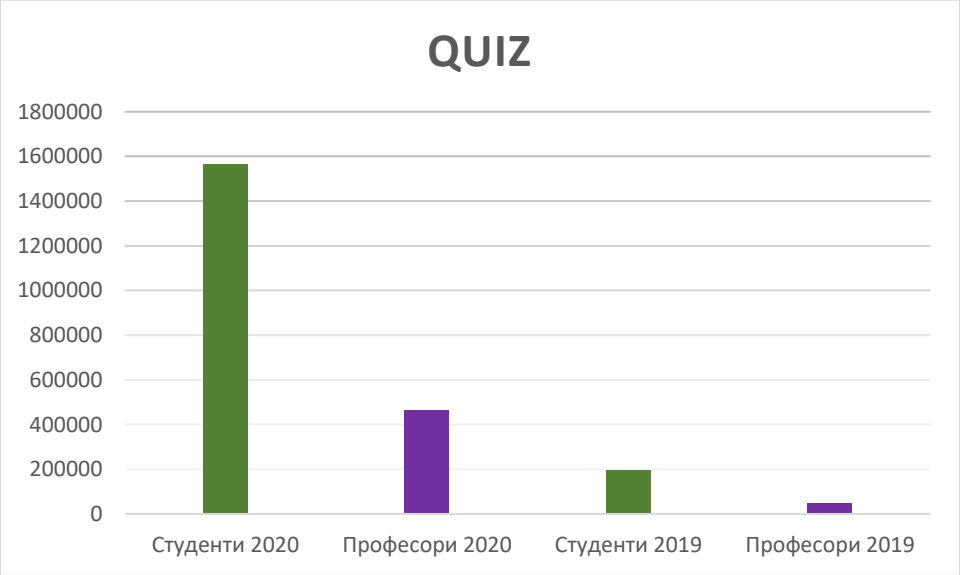
Година	2019		2020	
Корисници	*Наставнички кадар*	*Студенти*	*Наставнички кадар*	*Студенти*
Forum	3 545	29 815	60 346	189 534
Quiz	44 312	195 816	475 203	1 562 199
Assign	43 705	125 613	386 013	1 294 855
Choice	1 124	4 816	3 052	7 346
Book	1 028	6 219	1 214	7 614
Chat	618	5 114	11 564	63 128
Glossary	1 456	3 187	1 379	4 565

Заради подетална анализа и добивање на попрецизна слика за поединечните активности на корисниците е направена поделба, така што секоја од активностите *Forum, Quiz, Assignment, Choice, Book, Chat, Glossary* е претставена поединечно за секој од корисниците. За секој од подолу прикажаните графикони од

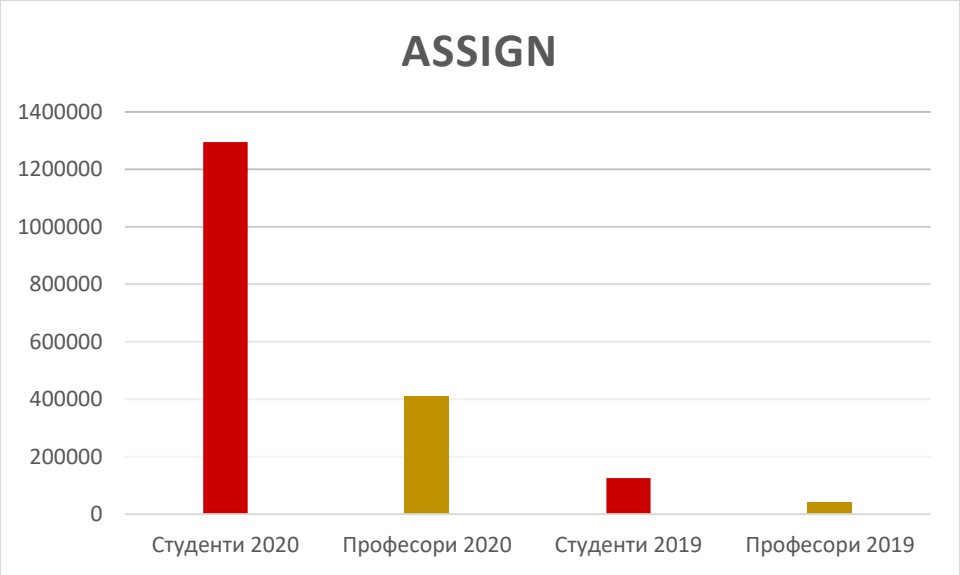
слика 23 до слика 29 на x-оската се претставени корисниците и годината која ги прикажува резултатите, а на y-оската се прикажани вредностите.



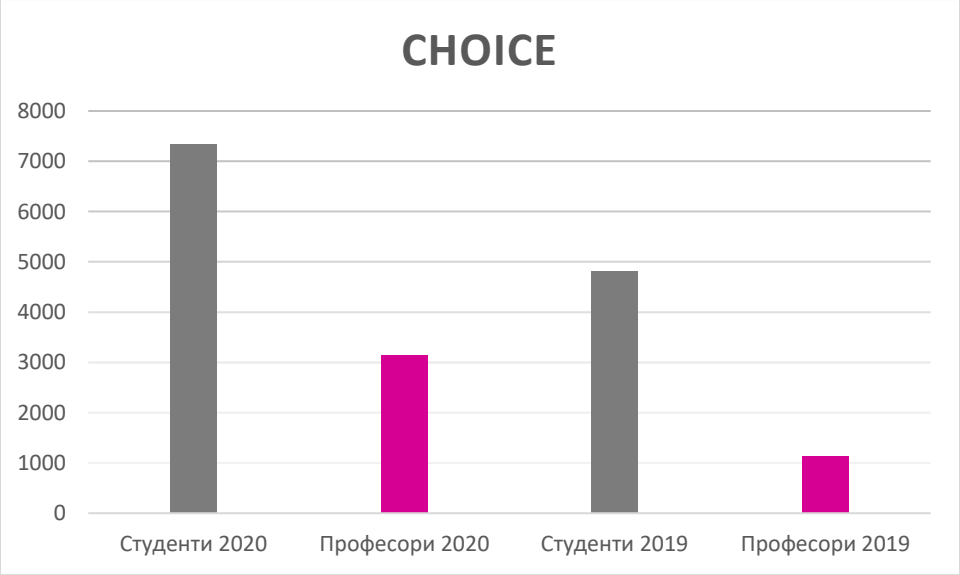
Слика 23. Бројот на активности на корисниците во модулот форум
Figure 23. The number of user activities in the forum module



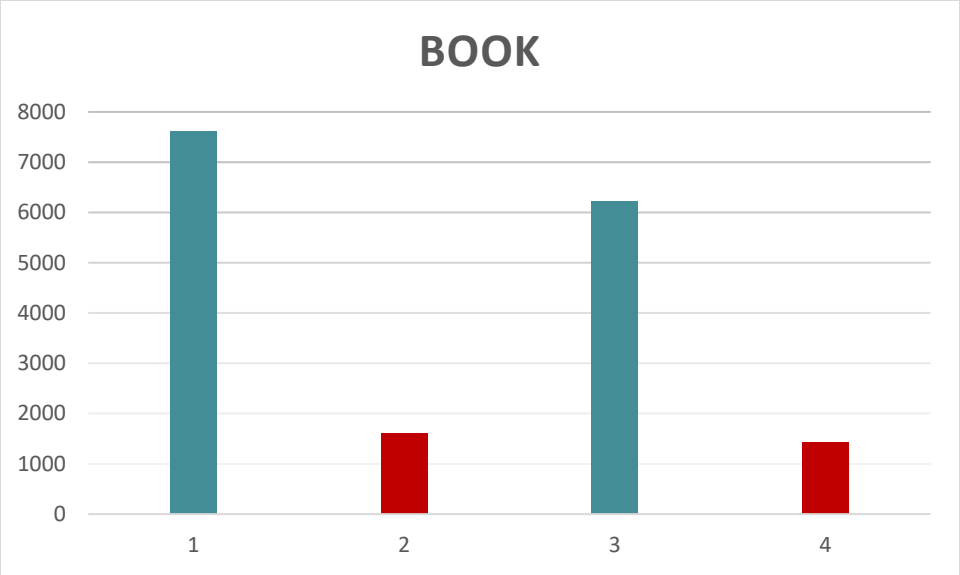
Слика 24. Бројот на активности на корисниците во модулот квиз
Figure 24. The number of user activities in the module quiz



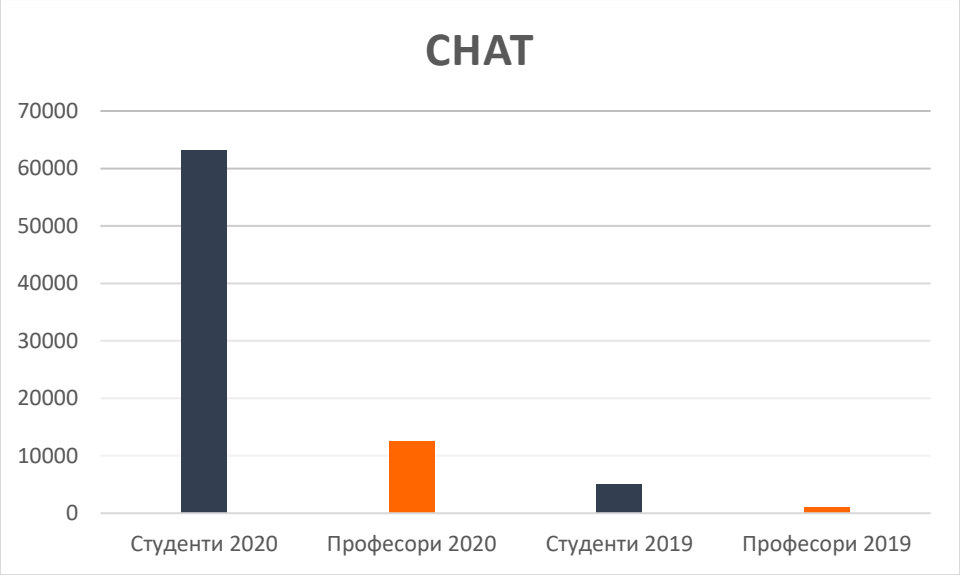
Слика 25. Бројот на активности на корисниците во модулот задачи
Figure 25. The number of user activities in the assignment module



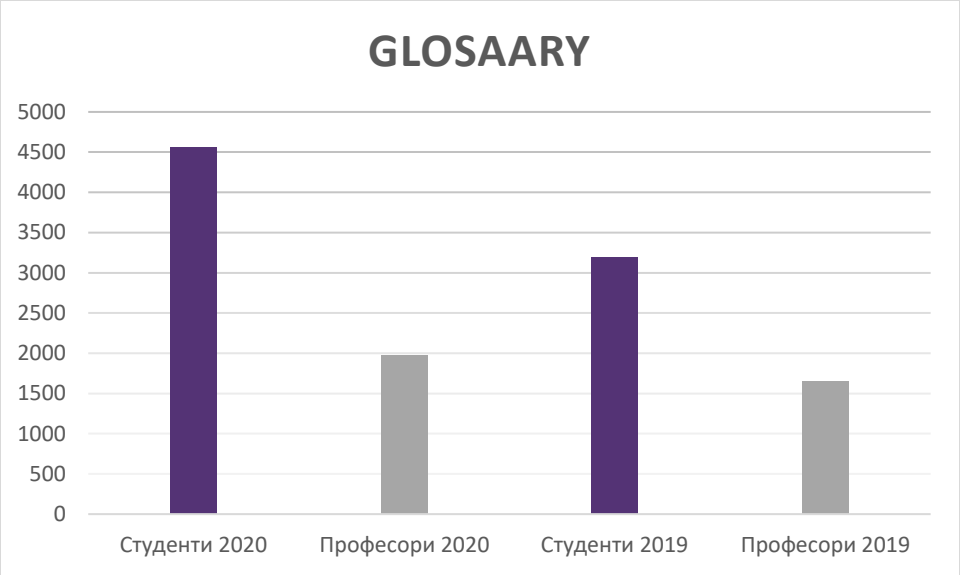
Слика 26. Бројот на активности на корисниците во модулот избор
Figure 26. The number of user activities in the choice module



Слика 27. Бројот на активности на корисниците во модулот книга
Figure. 27 The number of user activities in the book module



Слика 28. Бројот на активности на корисниците во модулот разговор
Figure 28. The number of user activities in the chat module



Слика 29. Бројот на активности на корисниците во модулот речник
Figure 29. The number of user activities in the glossary module

Со овие графички прикази кои се однесуваат на секој од модулите посебно многу лесно може да се воочи колку и каде е настанато зголемување на активностите.

Онлајн учењето обезбеди различни инструкции кои се предводени од предавачите, односно од наставничкиот кадар во образовните институции. Инструкциите можат да бидат синхрони (комуникација каде што учесниците комуницираат во исто време со видеоконференции) или асинхрони (временски разделена комуникација како е-пошта, форма на Google, стриминг видеосодржини, објавување белешки за предавања и платформи за социјални медиуми). Поради тоа, начинот на изведување на наставата беше овозможен со користење на истите тие инструкции. Од резултатите кои се добиени и прикажани може да се забележи дека најголемиот број на зголемување на активностите се кај модулите кои се однесуваат на активности како што се домашни задачи, форуми кои овозможуваат комуникација, тестови, е-пошта и слично. Сите овие активности се неопходни за текот на едукативниот процес, а истите тие бележат зголемување како кај наставничкиот кадар, така и кај студентите. Она што може да се забележи согласно со горенаведените табеларни и графички прикази е дека достигнувањето на

високиот број на активности во 2020 година го прикажува порастот на активности на системот за електронско учење Moodle, поради појавата на пандемијата. Добиените резултати може да се користат и како предмет за анализа за тоа на кој начин пандемијата влијае на начинот на живот. Иако дигитализацијата и учењето на далечина заземаа сè поголем замав во последните неколку години, појавата на пандемијата го поттикнува прашањето дали во иднина ќе стане нов начин на функционирање на едукативниот процес во образовните установи и институции.

8. Заклучок

Образовниот систем никогаш порано не се соочил со ситуација како сега. Кризата која беше предизвикана од ширењето на пандемијата Covid-19 ги истакна потребите и предизвиците во такви услови да се обезбеди право на студирање и да се продолжи со текот на наставата. Оваа криза покрај несигурноста и стравот кој го предизвика, поттикна потрага по нови решенија за организација и имплементација на наставата и учењето. Додека учењето на далечина се сметаше како алтернатива на традиционалното учење, за време на пандемијата тоа стана суштински елемент за одржување на универзитетските активности. Програмите за електронско учење не се нови во процесот на учење, бидејќи постојат и додипломски и постдипломски програми кои се целосно испорачани електронски (на пример, учење на далечина, учење преку интернет или комбинација од двете). Но, со појавата на пандемијата која го зафати целиот свет, овој тип на учење на далечина стана најважна и најмоќна алатка во управувањето со образованието. Со целосна надградба и употреба на сѐ она што го овозможуваат системите за електронско учење, заедно со бројните информатички достигнувања претставуваат причина за овозможување на значајна промена во образовниот систем, како и сите негови компоненти.

Онлајн учењето, далечинското и континуираното образование станаа лек за оваа глобална пандемија без преседан и покрај предизвиците поставени и за професорите и за студентите. Премин од традиционално учење со физичко присуство на онлајн учење може да биде сосема поинакво искуство за учениците и професорите, на кое тие мора да се приспособат со малку или без никакви други алтернативи. Алатките за електронско учење одиграа клучна улога за време на оваа пандемија, помагајќи им на училиштата и универзитетите да го олеснат учењето на студентите за време на затворањето на универзитетите и училиштата.

Искуството на цела година со образование на далечина на прво место го истакна предизвикот за откривање на активностите на студентите и професорите за време на пандемијата и пред почетокот на пандемијата. Менаџирањето и анализата на податоци претставува основа за еден од најголемите бенефити односно за добивање на вистински патоказ за идните чекори во работењето, но и

насоки за подобри резултати. Појдовен момент за решавање на предизвикот поврзан со големите податоци е нивната обработка. Со правилен избор на алатки и нивна употреба за анализа и обработка на големите податоци се добиваат посакуваните резултати, односно се дава приказ за користењето на различните модули за тоа кој од нив наоѓа најголема примена во учењето на далечина за време на пандемијата. Исто така се добиваат информации за можностите кои ги нудат системите за електронско учење. Методите и техниките за обработка на податоци се многу значаен момент, бидејќи од тоа зависи исходот на добиените резултати. Односно, од огромна важност е изборот на техниката за обработка на податоците, кој треба да биде направен врз основа на видот на податоци кои се предмет на обработка. Изборот на техниката за обработка на податоци треба да биде направен врз основа на видот на податоците кои се предмет на обработка.

При обработка на податочниот сет резултатите се добивани и анализирани последователно и поединечно, почнувајќи од вкупните активности на корисниците на Moodle платформата, па завршувајќи сè до поединечните активности на корисниците. Добиените резултати, покрај тоа што ја покажуваат разликата во активностите на Moodle платформата, пред почетокот на пандемијата и со појавата на пандемијата, даваат информации и за тоа како појавата на пандемијата Covid-19 влијаеше на текот на образованието.

Од овој магистерски труд може да се заклучи дека со обработка на податоците од системите за електронско учење се утврдува начинот на функционирање на едукативните процеси, односно може да се добијат информации од поединечните активности на сите корисници на платформата, па сè до информации од типот на разликата помеѓу вкупните активности на годишно ниво на корисниците. Со оглед на тоа што добиените резултати покажуваат дека бројот на активности во 2020 година е зголемен три пати повеќе во однос на 2019 година, се добива одговор на првото поставено прашање кое е од интерес на ова истражување - дали има зголемување на активностите во 2020 година. Исто така, врз основа на понатамошната анализа која се однесува на поединечните активности на наставничкиот кадар и студентите, може да се заклучи дека во 2020

година има зголемување за секој од модулите. Тоа укажува дека и кај наставничкиот кадар и кај студентите има зголемување до три пати повеќе во 2020 година, во однос на 2019 година.

Со реализација на истражувачки и аналитички аткивности од овој тип се добиваат информации кои може да се искористат за оптимизација на комуникацијата и активностите помеѓу професорите и студентите. Преку овие истражувања се добива претстава за користењето на различните модули, а како резултат на тоа се добива јасна слика за текот на наставата која се одвива во текот на едукативниот процес во образованието. Иако ова истражување е направено во текот на пандемијата, може да се заклучи дека дури и со поминување на тековната криза предизвикана од пандемијата долгорочно може да се очекуваат трајни промени во образовниот систем, поттикнати од ова искуство.

9. Користена литература

[1] M. G. M, "Big data: The next big thing in innovation.," *Research-technology management,* pp. 64-67, 2013.

[2] H. P. E., „What is Big Data and why is so important?," *Journal of Educational Technology Systems,* pp. 159-171, 2014.

[3] J. S. S. B. F. a. H. S. L. Johnson, "Big data facilitation, utilization, and monetization: Exploring the 3Vs in a new product development process," *ournal of Product Innovation Management,* pp. 640-658, 2017.

[4] A. C. Alexeev Michael, „The impact of institutional quality on manufacturing sectors: A panel data analysis.," *Economic Systems ,* 2021.

[5] X.-D. Zhang, „Machine learning," *A Matrix Algebra Approach to Artificial Intelligence,* pp. 223-440, 2020.

[6] Y. B. A. C. Goodfellow Ian, „Machine learning basics," *Deep learning,* pp. 98-164, 2016.

[7] D. Fumo, "Towards Data Science," 5 May 2017. [Online]. Available: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861.

[8] S. Y. Kumari Khushbu, „Linear regression analysis study," *Journal of the practice of Cardiovascular Sciences,* pp. 33-34, 2018.

[9] L. Connelly, „Logistic regression," *Medsurg Nursing 29.5,* pp. 353-354, 2020.

[10] A. A. Charbuty Bahzad, "Classification based on decision tree algorithm for machine learning," *ournal of Applied Science and Technology Trends, 2(01),* pp. 20-28, 2021.

[11] S. Prabhakaran, "Machine learning," 4 November 2018. [Online]. Available: https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/.

[12] P. Sayan, „Unsupervised Learning and Other Tools for Data Stream Mining," в *Practical Machine Learning for Streaming Data with Python*, Berkeley, Apress, 2021, pp. 97-113.

[13] V. S. P. W. M. A. Garcia-Dias R, "Clustering analysis," in *Machine learning*, Academic Press, 2020, pp. 227-247.

[14] M.-S. Y. Sinaga Kristina P, „Unsupervised K-means clustering algorithm,“ *EEE Access ,* том 8, pp. 80716-80727, 2020.

[15] B. C. L, „The EM algorithm: theory, applications and related methods,“ *Lecture Notes, University of Massachusetts,* 2017.

[16] H. K. S. R. R. C. Konstantin Shvachko, The Hadoop Distributed File System, Incline Village, NV, USA: IEEEE, 2007.

[17] J. A. Alam Anam, "Hadoop architecture and its issues," *2014 International Conference on Computational Science and Computational Intelligence.,* vol. 2, pp. 288-291, 2014.

[18] S. S. J. W. Mackey Grant, "Improving metadata management for small files in HDFS," *2009 IEEE International Conference on Cluster Computing and Workshops,* pp. 1-4, 2009.

[19] B. Dhruba, "HDFS architecture guide," *Hadoop apache project,* no. 2, pp. 1-13, 2008.

[20] D. G. Ghazi Mohd Rehan, "Hadoop, MapReduce and HDFS: a developers perspective," *Procedia Computer Science 48,* pp. 45-50, 2015.

[21] S. Kyuseok, „MapReduce algorithms for big data analysis,“ *Proceedings of the VLDB Endowment ,* pp. 2016-2017, 2012.

[22] S. Sinha, "Fundamentals of MapReduce with MapReduce Example," Edureka, 15 Novembar 2016. [Online]. Available: https://medium.com/edureka/mapreduce-tutorial-3d9535ddbe7c.

[23] M. R. N. K. Kavyashree T, "Hadoop Yarn Big Data," *Journal of Advanced Database Management & Systems,* vol. 7, no. 3, pp. 18-26, 2021.

[24] N. Vaidya, "Big Data and Hadoop," Edureka, 22 May 2019. [Online]. Available: https://www.edureka.co/blog/cloudera-hadoop-tutorial/.

[25] M. Olagunju, "Application of big data for distribution and consumption of power," *Telkomnika,* vol. 4, no. 19, 2021.

[26] H. A. M. K. Iyengar Samaya Pillai, "Big data analytics in healthcare using spreadsheets," in *Big Data Analytics in Healthcare*, Cham, Spinger, 2020, pp. 155-187.

[27] P. S. V. Naresh Kumar, "Apache Ambari architecture," O'reilly, 2021. [Online]. Available: https://www.oreilly.com/library/view/modern-big-data/9781787122765/f9e02281-de42-437c-afd2-f76de30c58c7.xhtml.

[28] P. G. Ritu Ratra, "Big Data Tools and Techniques: A Roadmap for Predictive Analytics," *IJEAT,* vol. 9, no. 2, p. 2249 – 8958, 2019.

[29] "Hadoop Zookeeper Tutorial," ProjectPro, [Online]. Available: https://www.dezyre.com/hadoop-tutorial/zookeeper-tutorial.

[30] "Apache Avro," Techopedia, [Online]. Available: https://www.techopedia.com/definition/30298/apache-avro.

[31] X. Waibel, "Lesser-Known Tips on Apache Oozie," Towards data science, 24 Novembar 2019. [Online]. Available: https://towardsdatascience.com/lesser-known-tips-on-apache-oozie-1e9bee9169da.

[32] S. Scott, "Hive architecture," in *Practical hive*, Berkeley, Apress, 2016, pp. 37-48.

[33] V. Balaswamy, "HCatalog," in *Beginning Apache Pig*, Berkeley, Apress, 2016, pp. 103-113.

[34] V. Balaswamy, Beginning Apache Pig, Berkeley: Apress, 2016.

[35] V. Deepak, „Using apache sqoop,“ в *Pro Docker*, Berkeley, Apress, 2016, pp. 151-183.

[36] H. A.-S. W. M. F. Aldowah Hanan, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and informatic,* vol. 37, pp. 13-49, 2019.

[37] S. L. W. Watson William, "An argument for clarity: What are learning management systems, what are they not, and what should they become," 2007.

[38] H. W. Rice William, Moodle, Birmingham: Packt publishing, 2006.

[39] A. V. Z. Z. Dijana Lapevska, "Analysis of Moodle activites before and after the Covid-19 Pandemic - Case study at Goce Delchev University," *Bjami,* vol. 4, pp. 51-58, 2021.