Факултет инжењерских наука

Универзитета у Крагујевцу

Назив студијског програма: Машинско инжењерство

Ниво студија: Мастер академске студије

Модул: Информатика у инжењерству

Предмет: Пројектовање информационих система и база података

Број индекса: 392/2013

Александра Д. Митровић

Big Data – Карактеристике и могућности примене

Мастер рад

Комисија за преглед и одбрану:

1. **Проф. Др Милан Ерић - ментор**
2. _____
3. _____

Датум одбране:_____

Оцена:_____

У оквиру мастер рада са темом "*BigData – Карактеристике и могућности примене***" кандидат треба да:**

Кроз уводна теоријска разматрања везана за објашњење појма BigData, архитектуре BigData, технологије BigData, карактеристике и могућности BigData, прикаже BigData као полазиште за управљање и анализу података различитих типова и облика. Рад треба да обухвати и конкретне примере премене BigData, односно студије случаја. На крају дати закључна разматрања и списак коришћене литературе.

1. Rainer K., Turban E.: **Introduction to Information Systems**, John Wiley & Sons, Inc., 2009
2. Лазаревић Б., ……: **Базе података**, ФОН Београд, Београд 2003.
3. Soumendra Mohanty, …: **Big Data Imperatives**, Springer Science+Business Media, New York, 2013
4. Peter Zadrozny, Raghu Kodali: **Big Data Analytics**, Springer Science+Business Media, New York, 2013

Ментор:

Prof. Dr Milan D. Erić

Крагујевац, Септембар 2015

Изјављујем да сам овај мастер рад урадила самостално, служећи се стеченим знањем и искуством као и наведеном литературом.

Александра Митровић 392/2013

Садржај

Резиме

У овом раду је објашњен данас веома актуелан појам, али не и довољно обрађен на нашем подручију, који се зове Big Data. Кроз овај рад приказане су главне карактеристике Big Data, модели података који се односе на велику количину података, технологије које су способне да обраде тако велику количину података, а што је најважније дата је и анализа ових података. Оно што је суштина рада је да објасни и прикаже које су користи од Big Data, како је на прави начин искористити и применити на пословање, што се може видети кроз дате примере.

Кључне речи: Big Data, Hadoop, Big Data анализа, MapReduce, HDFS, Twitter, Сензор подаци.

Abstract

In this paper explains today very actual term, but not enough processed on our territory, which is called Big Data. Through this work are explained the main characteristics of Big Data, data models that apply to large data, technologies that are capable of processing such a large amount of data, which is the most important data and analysis of these data. What is the essence of this paper is to explain and present the benefits of Big Data, how to properly use and apply to the operations, which can be seen through the examples.

Keywords: Big Data, Hadoop, Big Data Analysis, MapReduce, HDFS, Twitter, Sensor Data.

1.Увод

Количина података свакодневно расте веома великом брзином и то из више разлога. Најпре, расте број извора података јер све више корисника генерише и оставља своје податке, како на послу, тако и на друштвеним мрежама. Није занемарљива ни нова количина података која се добија аутоматским очитавањем нових сензора, система за надзор и регулације и сличних система. Поврх свега, количина података за сличне информације расте из године у годину јер расту резолуције фотографија и видео материјала који су све чешћи носиоци информација.

Прикупљање, класификација, складиштење и обрада података огроман су изазов за компаније које желе на прави начин да искористе све ресурсе које имају. Иако је у основи то технички проблем, он има снажне пословне консеквенце – ако организација не анализира довољно брзо податке које прикупља, може закаснити с доношењем важних одлука. Последице кашњења обично значе лошије пословне резултате. Додатну тежину овом проблему даје чињеница да је већина података неструктурирана, односно не прикупља се у базе података, већ настаје помало стихијски, а чине га различити типови докумената, фотографије, видео и други фајлови.

Када има толико информација у толико различитих облика, немогуће је размишљати о управљању подацима на традиционалан начин. Иако је одувек постојало много података, разлика је у томе што данас већина тога постоји, а варира само у врсти и начину обраде. Организације, више него икада раније, проналазе начин да искористе ове информације. Дакле, о управљању подацима мора се мислити другачије и то је изазов, а уједно и шанса, за Big Data.

У овом раду је представљен појам Big Data, архитектура система која служи за обраду, складиштење и анализу података, модели података, а такође су дати и примери примене Big Data, као и анализе које помажу да се побољша свака врста пословања.

1.1. Историја - развој база података

Настанак база података се везује за Herman Holerith-a. Он је 1884. год пријавио патент систем за аутоматску обраду података (АОП), који је приказан на Слици 1. АОП је служио за попис становништва у САД. Подаци су били на бушеним картицама који су се ручно убацивали у уређај за очитавање, а обрада података вршила се пребројавањем. Дотадашња обрада података пописа трајала је 10-ак година, а са овим изумом време обраде смањило се на шест недеља. Идеја АОП-а је била да се сваки становник САД представља са низом од 80 карактера - име, годиште итд. Од Holerith - ове компаније настао је данашњи IBM [9].

Слика 1. Машина Herman Holerith-a за обраду податка 1884.године [21]

Након Другог светског рата, у компанијама и владиним институцијама почели су се појављивати први електронски рачунари. Они су се често користили управо за једноставне линеарне базе података, најчешће за рачуноводство. Ипак, врло брзо, богати купци су почели да захтевају више од њихових екстремно скупих машина. Све је то водило до раних база података. Занимљиво, ове ране апликације су наставиле да користе Hollerith - ове бушене картице, незнатно модификоване у односу на оригинални дизајн. Нефлексибилност поља исте дужине, базе података покретане 80 колонским бушеним картицама, учиниле су ране рачунаре метом напада и шала и потпуном мистеријом за обичног човека [10].

Већина првобитних база података се односила на специфичне програме написане за специфичне базе података. За разлику од модерних система који могу бити примењени на потпуно различите базе података, ови системи су били уско повезани за базу података да би осигурали брзину на уштрб флексибилности. Системи управљања базама података су се први пут појавили током 1960-тих година и наставили су да се развијају током следећих деценија. У већини случајева, период увођења је дуго трајао, скоро деценију пре наведене године почетка употребе. На пример, релациони модел је први пут дефинисан од стране E.F.Codd у тексту објављеном 1970 године. Међутим, релациони модел није био широко прихваћен све до 1980-тих година [10].

1960'те

Системи засновани на датотекама су имали доминантну улогу, али и поред тога први системи за управљањем базом података су уведени у овој деценији. Најпре су се користили само код великих пројеката, као што је био пројекат слетања Apollo-а на Месец. Такође први кораци ка стандардизацији предузети су у овој деценији, формирањем DTB Grupa (Data Base Task Group) [9].

1970'те

Систем за управљањем базом података је постала комерцијална ствар, великим делом због потребе за системом који ће моћи да управља сложеним структурама података, као сто су рачуни фабрика при набавци сировина. Ови модели се сматрају првом генерацијом система за управљањем базом података. Многи од тих система су у употреби и дан данас, али имају неколико великих недостатака [9]:

- Тежак приступ подацима. За приступ и најједноставнијим подацима били су потребни изузетно сложени програми.

- Веома ограничена независност података, тако да се програми нису могли изоловати од промена у формату података.

- Да би се превазишла ова ограничења Edgar.F.Codd је развио релациони модел. Овај модел одваја логички модел од физичког начина смештања података.

1980’те

Релациони модел је доживео широку комерцијалну употребу у пословном свету. Са релационим моделом сви подаци су представљени у форми табеле. Релативно једноставан програмски језик четврте генерације, назван SQL, коришћен је за добијање информација. Овим моделом је обезбеђен једноставан приступ подацима и оним људима који нису били програмери. Овај модел је такође имао погодност за клијент/сервер обраду, паралелни пренос података и употребу графичког корисничког интерфејса (GUI) [9].

1990’те

Развој рачунарских мрежа и клијент/сервер обрада као и појава мултимедијалних података (графика, звук, слика и видео запис) постали су уобичајна ствар деведесетих. Самим тим што су подаци постали све сложенији морало је да се пронађе ново решење за системе база података. Тако су настали системи који су окренути ка објекту, и они се сматрају трећом генерацијом [9].

2000'те

У овој деценији се јављају нови правци за развој система за базе података, као што су [9]:

- Могућност управљања све сложенијим типовима података. Ови типови укључују и мултидимензионалне податке, који су већ добили на важности у апликацијама складиштења података,
- Децентрализоване базе података,
- Примена вештачке интелигенције ће олакшати приступ подацима и необученим корисницима,
- Развој нових техника и алгоритама за анализу података – анализа складишта података,
- Заштита података.

На основу претходне хронологије може се уочити развој база података и напредак у овој области кроз деценије. Ми смо сведоци у овој текућој деценији невероватном развоју података, а заправо то доводи и до развоја Big Data, појма који је управо описан у овом раду и који је тренутно актуелан и релативно нов, како у свету тако и код нас, а који не представља ништа друго од енормне количине података. Питање које следи је шта урадити са том количином, како обрадити и искористи те податке (о чему ће бити речи у наредним поглављима).

Слика 2. Савремени кластери [21]

На Слици 2 је приказана једна врста кластера. Кластер рачунари се састоје од скупа слабо или чврсто повезаних рачунара који раде заједно, тако да се у многим аспектима они могу посматрати као један систем. У овом одељку неће бити речи о кластерима и супер компијутерима, циљ ове слике је да покаже како се технологија, односно системи за скалдиштење и обраду података развијали од настанка прве машине за обраду података до данас.

1.2. Шта је Big Data?

Big Data представља податаке које су оне количине, која превазилази могућности уобичајено коришћеног софтвера за складиштење, обраду и управљање подацима [4].

Ако се традиционална база података посматра као колекција онда се може рећи да је Big Data колекција колекција у потпуно различитим форматима, облицима и технологијама. Исто тако, на први поглед, није једноставно разумети како их међусобно смислено повезати да могу вратити смислену информацију. Колекције могу бити толико сложене да не морају бити организоване у колоне и редове већ могу бити представљене низом неструктурираних података као што је низ информација које се размењују путем Facebook или Twitter порука, или неких других друштвених мрежа. На Слици 3 су приказани различити типови извора података, који генеришу огромну количину и различите врсте података [3].

Слика 3. Различити типови извора података [23]

Када се сумирају постојеће дефениције Big Data, може се рећи да је Big Data појам који се односи на велику количину података која се генерише коришћењем производа, услуга и апликација. Велике количине података се генеришу несвесно од стране корисника. Како би се употребили ти подаци, потребно их је сачувати и омогућити анализу у неком реалном времену.

Ипак, споменуте дефиниције не могу у потпуности објаснити шта је Big Data. Зато појам Big Data се најједноставније може објаснити уз помоћ карактеристика, приказаних на Слици 4, а које се у литератури називају „3V“: Volume (количина), Velocity (брзина) и Variety (разноврност).

Слика 4. Карактеристике Big Data [3]

У зависности од литературе, односно аутора, могу се пронаћи различити „V“- ови, међутим споменуте карактеристике (Volume, Variety и Velocity) су основне и спомињу се свуда, па се углавном уз њих додају и остале карактеристике.

У оквиру следећег поглавља свака карактеристика биће детаљно објашњена.

1.3. Карактеристике Big Data

Обим онога што се данас посматра као Big Data је широк, а дефиниције су често нејасне, понекад контрадикторне. Најшире прихваћена дефиниција термина Big Data проистекла је из анализе Мета Групе (садашњи Гартнер) која је рађена 2001 године и по тој дефиницији под појмом Big Data се подразумева информациони ресурс велике количине, велике брзине и велике разноврсности података који захтева нове и иновативне методе обраде и оптимизације информација, побољшање увида у садржај података и доношења одлука [3].

С обзиром на то да се спомињу информациони ресурси велике количине, велике брзине и велике разноврсности података, Big Data технологије се не могу разматрати без осврта на споменуте „V“- ове. Поред основних карактеристика Volume (количине), Velocity (брзина) и Variety (разноврности) у оквиру овог поглавља споменуте су још и Veracity (истинитост) и Value (вредност).

1.3.1. Volume (количина)

Веома је битно питање колико је заправо то података када се говори о великој количини података. Много фактора доприноси увећању обима података (трансакциони подаци складиштени годинама, текстуални подаци који константно надолазе са друштвених мрежа, итд.). У прошлости је прекомерна количина података стварала проблеме око складиштења, али са данашњим ценама меморијских уређаја то више не представља проблем. Ипак, други проблеми се јављају, укључујући одређивање важности одређених података у великој гомили [3].

Volume је најочигледнији појам који описује Big Data. Сам појам величине се мења из године у годину, па се тако прешло на речи ZB (zettabajti). На Слици 5 приказане су јединице које описују величину великих података. Осим раста у величини, долази и до пада цене саме меморије тј. дискова. Према истраживањима 2009. године било је 0.8 ZB података у свету, 2010. године бројка је прешла 1 ZB, а у 2011. години говорило се о 1.8 ZB. Такође, очекује се раст ових бројева и до 35 ZB до 2017. године. Овим подацима стиче се утисак да се ради о заиста огромним количинама података, које треба анализирати. Све ово не значи да се научници или аналитичари баве целокупном количином података, већ указује на могући проблем приликом анализе тих података, где управо технологија Big Data може помоћи [3].

Naziv	Ozn	Binarni prikaz	Decimalni prikaz
byte	B	2^0=1 byte	10^0=1
kilobyte	KB	2^{10}=1.024 byte (B)	10^3=1.000
megabyte	MB	2^{20}=1.048.576 B	10^6=1.000.000
gigabyte	GB	2^{30}=1.073.741.824 B	10^9=1.000.000.000
terabyte	TB	2^{40}=1.099.511.627.776 B	10^{12}=1.000.000.000.000
petabyte	PB	2^{50}=1.125.899.906.842.624 B	10^{15}=1.000.000.000.000.000
exabyte	EB	2^{60}=1.152.921.504.606.846.976 B	10^{18}=1.000.000.000.000.000.000
zettabyte	ZB	2^{70}=1.180.591.620.717.411.303.424 B	10^{21}=1.000.000.000.000.000.000.000
yottabyte	YB	2^{80}=1.208.925.819.614.629.174.706.176 B	10^{24}=1.000.000.000.000.000.000.000.000
brontobyte	BB	2^{90}=1.237.940.039.285.380.274.899.124.224 B	10^{27}=1.000.000.000.000.000.000.000.000.000
geopbyte	GeB	2^{100}=1.267.650.600.228.229.401.496.703.205.376 B	10^{30}=1.000.000.000.000.000.000.000.000.000.000

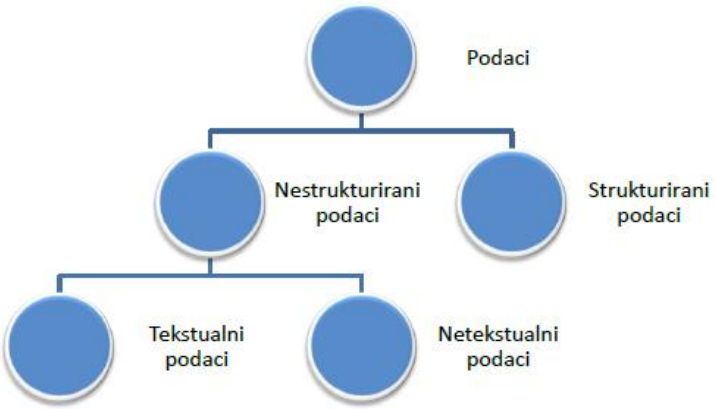
Слика 5. Јединице које описују огромну количину података [3]

Конкретно на неким друштвеним мрежама као што је на пример Twitter, према подацима из 2014.године, имао је око 645 милиона корисника, око 58 милиона твитова дневно у просеку, а у секунди се генерише 9100 твитова. Према истом истраживању, Facebook је имао око 1,3 милијарде корисника који су активни на месечној бази те је имао пораст од 22% што се тиче броја корисника гледајући период из 2012.године и 2013.године. Facebook, такође има око 54 милиона страница (Facebook Pages), док се на сваких 20 минута пошаље 3 милиона порука [3].

Све наведено јасно доказује зашто је битно систематски и правовремено размишљати о управљању тако великим количинама података. Као решење је изабрана Big Data технологија од којих ће неке бити објашњене у оквиру овог рада.

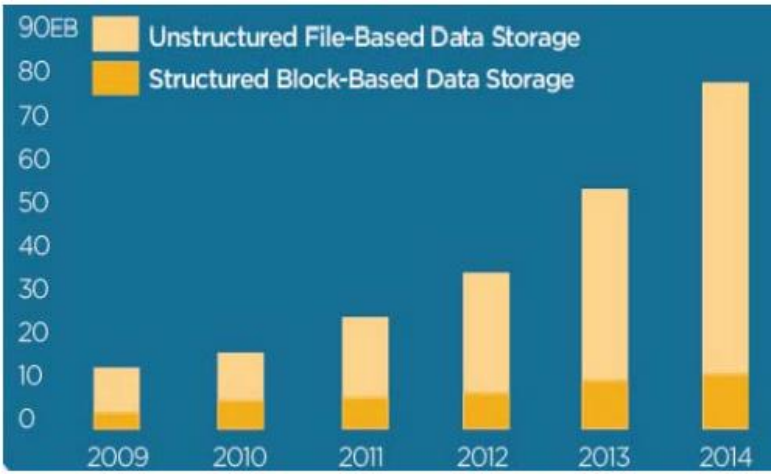
1.3.2. Variety (разноврност)

У овом делу се говори о разноликости пристиглих података. Док су у традиционалним базама података подаци обично добро структурирани, код Big Data то не мора бити случај. Big Data обухвата и неке поступке обраде и анализе текстуалних података који нису структурирани. Око 50 процената података у Big Data системима је неструктурирано. То конкретније речено значи да уместо уобичајених унапред дефинисаних података, попут трансакцијских података о продаји или о купцима, овде у систем улазе и различити текстуални подаци, слике, видео записи, подаци са друштевених мрежа, логови, сензорски податци и слично. Подаци који се користе могу бити интерни или из екстерних извора, прикупљени ван организације. Према истраживању IBM-а о коришћењу Big Data технологије објављеном 2013. године, између 30 и 40 процената података које организације користе потичу из спољних извора [3].



Слика 6. Врсте података у у Big Data системима [3]

На Слици 6 се може видети концептуална подела података с тим да је у овом случају, за потребе овог рада, занимљива грана неструктурираних података. Осим текстуалних може се видети да постоје нетекстуални неструктурирани подаци, а они представљају графике и слике (фотографије, илустрације, X-zrake, MRI...). Ово не упућује на то да неструктурирани подаци немају структуру већ да њихове субкомпоненте немају структуру (коментари, слике ...). Битно је усредсредити се на све податке које треба комбиновати како би се повећала њихова вредност. Пример би биле телекомуникације компаније које у својим позивним центрима свакодневно примају позиве који представљају неструктуриране податке који се могу комбиновати са структурираним подацима (историју трансакција, ...) и тако се добија веома персонализовани модел купца помоћу комбинације структурираних и неструктурираних података.

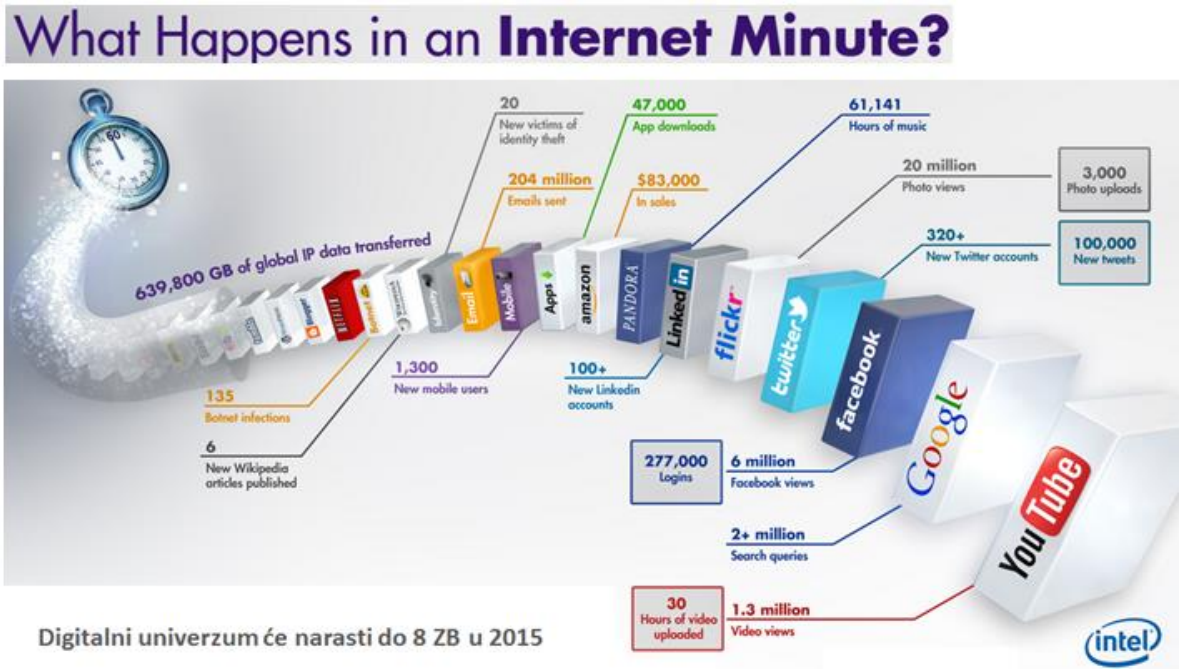


Слика 7. Однос сачуваних структурираних и неструктурираних података [5]

На Слици 7 приказано је како су се са годинама повећавали неструктуирани подаци, што је и логично јер се технологија развијала, повећавао се број друштвених мрежа, подаци очитани са сензора, размена порука преко напредних технологија, и тако се увећавао број различтих типова података.

1.3.3. Velocity (брзина)

Брзина обраде података представља брзину производње и генерисања података и брзину којом подаци морају бити обрађени да би задовољили одређене критеријуме. Правовремено реаговање и брза обрада података представљају велики изазов и за највеће компаније на свету [3].



Слика 8. Количина података која настане на Интернету за 1 минут [23]

Што се брзине тиче, већ је наведена кратка статистика у делу Volume која описује генерисање података на друштвеним мрежама на дневном и годишњем нивоу. Тако на пример Twitter има 9100 твитова у секунди. Ради се заиста о великом броју твитова које треба анализирати. Наравно, кориснике неће занимати свих 9100 твитова из те секунде, већ доста мањи број, али треба приликом израде апликације то имати на уму. Што значи да технологије које се баве великом количином подака морају да поседују способност веома брзе обраде и анализе података.

1.3.4 . Veracity (истинитост)

Према Зикопоулосу (аналитичар података у IBM-у) трећина руководећих људи који у пословању доносе одлуке не верује својим информацијама. Ова чињеница говори да би требало да се обрати пажња на квалитет података и информација који се добијају. Пре свега мисли се на несигурне податке који пристижу, податке који су недоследни, нејасни. Потребно је истражити колико су истинити подаци које добијамо. Такође, оно што је веома битно је чињеница колико се може веровати семантичким моделима да добро раде свој део задатка. Због тога је потребно константно пратити како модел, тако и цео систем, функционише и стално бити у потрази за нелогичностима које би систем могао генерисати. Што се семантичког модела тиче, било да се одабере метод речником или машинским учењем, оптимизација је нешто чему треба тежити. Поготово кад се говори о машинском учењу. Најмања промена у моделу може довести до пада или раста тачности модела, потребно је утврдити како се научени модел понаша са новим подацима, како различити алгоритми раде, који је најбољи алгоритам за дате податке [4].

1.3.5. Value (вредност)

Већина аутора спомене наведена четири „V“- а, али забораве најважније „V“, а то је Value. Value (вредност) је веома важна јер Big Data се не може посматрати само као огромна количина података која настаје великом брзином за веома кратко време, Big Data-у треба посматрати као ресурс, који треба да буде искористив.

Вредност Big Data се огледа у томе што анализом оваквих података пословне организације могу да побољшају своје посовање, јер уз помоћ нових технологија фирмама су веома приступачна мишљења (подаци) корисника, која су веома битна за сваку организацију, тако да анализом тих података организцаије могу да унапреде на пример своје производе или услуге.

Конкретније која је вредност Big Data, приказана је кроз примере у оквиру овог рада.

2. Менаџмент информационих технологија

Успешно функционисање и управљање организацијом не може се остварити без одговарајућих података, информација и знања. Стога се подаци, информације и знања могу схватити као својеврсни ресурс организације попут ресурса радне снаге, материјала, енергије, финансија и других [6].

2.1. Основе менаџмента ИТ-а

Информационе технологије (ИТ) су закорачиле у све видове друштва и живота, тако да је данас тешко наћи било какав пословни процес у коме бар неки његов део није подржан информационим техонологијама. Нажалост, ретке су организације које су успеле да добију пуне ефекте од примене информационих технологија. Углавном се примена информационих технологија своди на подршку традиционалном начину рада. Мали број организација је успешно извршио реинжењеринг пословних процеса са циљем аутоматизације својих пословних процеса применом савремених достигнућа ИТ [6].

Да би се искористиле све предности ИТ мора се извршити интеграција знања о томе како функционише цела организација са знањима из одговарајућих информационих технологија[6].

Веома је битно доћи до сазнања како функционише организација, без ограничења која могу бити проузрокована постојећим функцијама. То захтева промену начина посматрања организације, усмеравање пажње са појединих функција и активности на процесе који покривају комплетно пословање [6].

Процесни приступ омогућава ИТ специјалистима да лакше схвате дешавања у пословном процесу, као и који су циљеви и резултати појединих процеса. Менаџмент мора поседовати основна знања из ИТ како би могао схаватити решења ИТ специјалиста, а такође и иницирати одређене идеје. Заједнички рад на BPR (Business Process Reengineering) захтева од менаџмента велику промену у схватањима како треба управљати пословним процесом [6].

Примена ИТ проузрокује појаву нових правила, која су понекад у потпуности у супротности са до тада важећим правилима [6]:

- повећава се доступност и брзина протока докумената и информација,
- подижу се и уједначавају способности и квалитет рада запослених,
- одлучивање улази у радне обавезе сваког запосленог,
- истовремено су доступне предности централизације и децентрализације система,
- географска разуђеност не захтева ангажовање додатнх ресурса,
- потрошчи могу одређене услуге конзумирати самостално помоћу Интернет комуникације
- преиспитивање планова у сваком тренутку омогућено је тренутним сумирањем информација.

Менаџменту предузећа је потребна одговарајућа информациона подршка приликом избора, развоја, реализације и контроле реализације конкуретних пословних стратегија, што би требалло да доведе до повећања ефикасноти у пословању предузећа и што би резултирало остварењу екстра профита [6].

Информатизација организације се може дефинисати као примена рачунара и других информатичких и комуникационих средстава и уређаја у аутоматизацији информационих процеса и токова и аутоматизацију информационе комуникације организације са њеним релевантним окружењем. Циљ информатизације организације је аутоматизација основних процеса. У то спада постизање унутрашње информационе интегрисаности организације и њене спољашње информационе интегрисаности са релевантним окружењем [6].

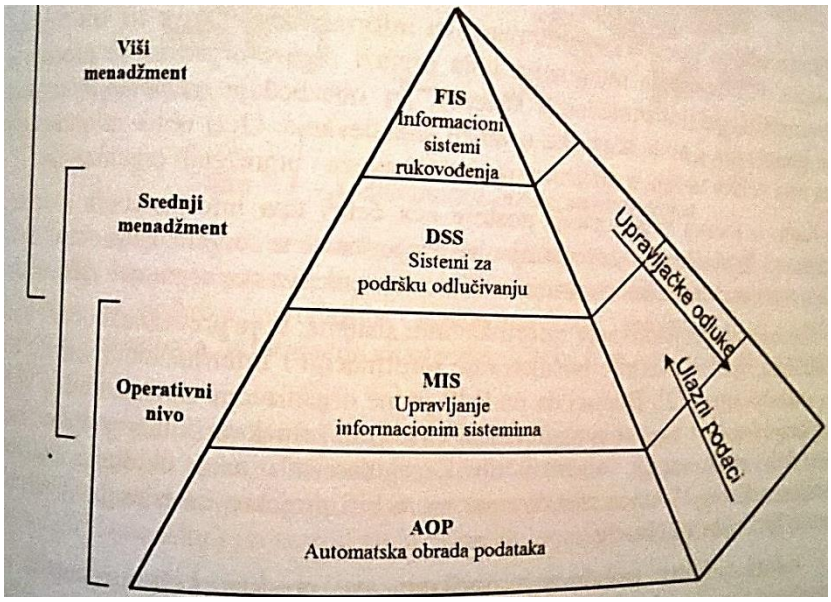
За изграђивање компјутеризоване организације битно је успоставити везу планирања база података и информационих система одозго на доле са локализваним обликовањем система у разним функционалним областима организације. Неопходно је обезбедити да се то локализовано обликобање развија у складу са информационом архитектуром која описује базе података и информационим системом потребним за успешно спровођење пословних планова организације. Да би се у информационом стратешком планирању могло одредити које су базе података и информациони ресурсипотребни, неопходно је сагледати перспективу организације као целине, односно стећи стратешку визију организације [6].

Одлучивање се у једној организацији може поделити на три нивоа [6]:

- оперативни,
- средњи и
- виши ниво.

У зависности од нивоа одлучивања постоје различити информациони системи (ИС) су [6]:

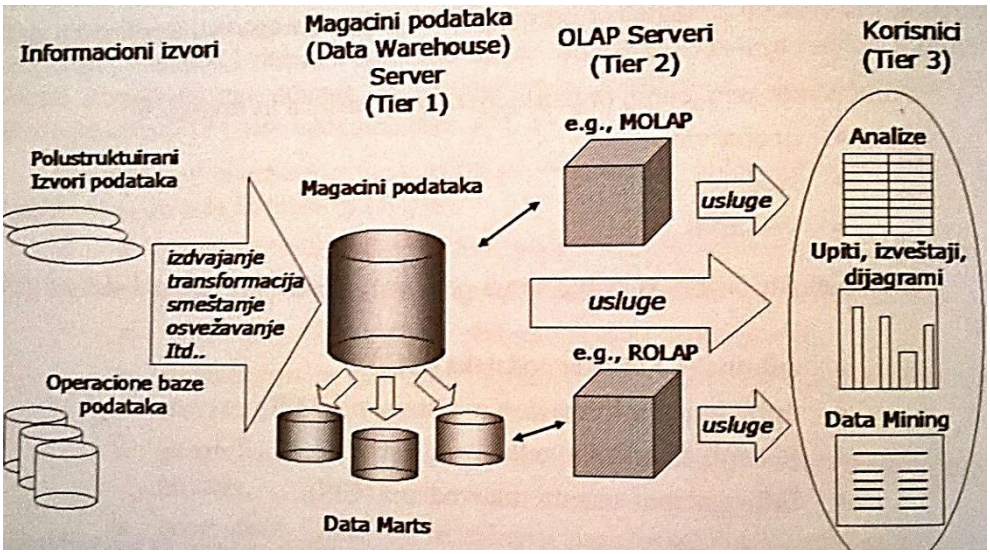
- аутоматска обрада података (AOP - Automatic Data Processing),
- управљање информационим системима (MIS - Management Information System),
- систем за подршку одлучивању (DSS - Decision Support System),
- информациони систем руковођења (EIS - Executive Information System).



Слика 9. Информациони систем са хијерахијским нивоима у предузећу [6]

На Слици 9 је приказано повезивање информационих система са хијерахијским нивоима у предузећу, на којима се они користе [6].

Између нивоа одлучивања постоји вертикални проток информација. Информације које иду са нижег нивоа представљају улазне податке за информациони систем на вишем нивоу. Информације које иду са вишег нивоа представљају управљачке одлуке за нижи ниво. Вертикални токови информација између управљачких нивоа називају се вертикална интеграција информационих система. На Слици 10 приказана је структура система за подршку при одлучивању [6].



Слика 10. Структура система за подршку при одлучивању [6]

Хоризонтална интеграција подразумева информационе токове на истом нивоу одлучивања. Хоризонтална интеграција која прелази оквире организације представља још савременији вид информационих система, јер обезбеђује размену информација између компанија и свих учесника у ланцу снабдевања. Овај облик информационих система има велику улогу [6].

Када у једној организацији постаје сва четири типа информационих система и вертикална и хоризонтална интеграција система, тада та организација има добро пројектован и савремен информациони систем, који покрива све сегменте рада [6].

Оптимално пројектовање информационих система, које превазилази област једне организације, захтева тачније посматрање информација и информационих система дуж ланца снабдевања. Будући да на ИМС једне организације значајан утицај имаји информације и послови који су у надлежности екстерних субејката (купаца, испоручилаца, партнерских организација, нормативних и регулаторних тела, околине, друштвене заједнице, тржишта), ИС ланца снабдевања мора бити пројектован тако да укључи ове информације и њихову обраду [6].

Основне карактеристике информационих система у ланцу снабдевања су [6]:

- Примена информација у свим карикама ланца снабдевања је веома интезивна,
- Информациона повезаност свих процеса у једној организацији,
- Информациона повезаност организације и окружења,
- Постоји велики број повратних информационих токова,
- Квалитет робе дирктно утиче на пласман робе на тржишту.

Информације у ланцу снабдевања су крајње интезивне, потичу из разних процеса и разних организација што намеће високе захтеве за електорнском комуникацијом у информационим системима ланца снабдевања. Осим тога планирање и развој појединих задатака су временски уско међусобно испреплетани, тако да је перменетни ток информација у информационим системима ланца снабдевања меродаван за њихову фукнционалну способност – капацитет. Архитектура информационих система у ланцу снабдевања мора бити флексибилна [6].

2.2. Big Data – информациони менаџмент

Big Data је нови појам у менаџменту који се односи на велику количину података које карактерише велика разноликост и велики проток (фреквенција). Ти и такви подаци захтевају иновативне форме процесирања и обраде ради унапређења процеса доношења одлука и оптимизације пословних процеса у организацији. Big Data могу бити веома корисни за предиктивну анализу чиме би имали веома велики утицај на раст и развој пословања. Технологије су присутне и расположиве су за све, али само мали број компанија је способан да добро уклопи предности технологија са сопственим корпоративним снагама и могућностима. Питање је зашто је веома важно управљање мноштвом података, њихова доступност, анализа и примена. Суштина је како искористити мноштво података који се свакодневно продукују у компанији кроз све сегменте њеног пословања и деловања [12].

Раст количине података у пословању компанија последица је пада цена хард дискова, повећања њиховог капацитета, повећања степена дигитализације и информатизације пословних процеса и раста количине „великих" фајлова за складиштење, као што су видео фајлови и системски лог фајлови. Big Data се односи на расположиве податке о понашању корисника, о тржишту, о продаји, итд. Big Data су истовремено и проблем и шанса, прилика и могућност да компанија постане иновативнија, конкурентнија и продуктивнија. Big Data, као способност да се сачувају (ускладиште) и анализирају огромне количине података, има велике последице на ИТ сектор у компанији и оне који управљају базама података. Шта ће се урадити са тим подацима, шта ће ти подаци рећи о корисницима, пословним процесима, о перформансама компаније? У наредних неколико година биће потпуно видљива и јасна разлика у погледу конкурентности између компанија које разумеју и користе Big Data и компанија које нису свесне значаја Big Data и не знају шта би урадиле са њим(а). Иако свест о важности Big Data расте, данас је мали број компанија у свету способно да Big Data искористе на прави начин, једне од њих су Google и Facebook. Огромне могућности за коришћење Big Data имају и војно-безбедоносно структуре, телекомуникационе компаније, малопродајни и велепродајни ланци, финансијске, здравствене и производне организације [12].

Подаци се могу класификовати по количини (обиму), по разноликости (врсти) и по брзини којом се „производе". Без анализе, подаци нису употребљиви као вид новог знања тј. не представљају знање. Анализом података одређује се њихова права вредност, која може бити пресудна у диференцирању компаније у односу на конкуренцију. Информације могу постати вредне уколико су транспарентне и употребљиве. На основу њих се доносе боље одлуке, унапређују перформансе компаније и њених производа и/или услуга, идентификују циљне групе и области у којима треба деловати [13].

Big Data може помоћи да се прецизније сегментирају корисници, креирају бољи пакети услуга и направе успешније промоције. Најважнији проблем је како обрадити велику количину података и инфомација на начин да омогући унапређење пословања и искуства купца, уз задржавање безбедности података и приватности корисника. На који начин треба дизајнирати и трансформисати пословне процесе – нпр. продају? Како нпр. податке о понашању корисника на Интернету и њиховим трансакцијама ускладити са новим производима и сервисима. Или, како искористити податке о недостатку сервиса или производа, као и времену чекања корисника на производ или услугу, да би се у компанији оптимизовало коришћење ресурса и корисници добили услугу што је пре могуће. Циљ је да подаци којима се располаже доведу до правих одлука (у право време) и најбољих планова [13].

Да би се одржала конкурентност треба повезати пословне процесе и информације. Процес није ништа друго до организовање људи, ресурса и активности да би се дошло до жељеног циља и резултата. Добар процес треба да буде надоградив, прилагодљив и отпоран да може да обради повећану количину информација. Успешан процес је заснован на квалитативном, а не квантитативном побољшању у одлучивању. Често се могу видети да фантастичне апликације, системи и технологије не доносе и не обезбеђују очекивани бољитак и напредак. Основни разлог тога је што компаније нису редизајнирале своје процесе и обучиле своје запослене [13].

У Big Data свету, креирање процеса треба да дефинише да ли ће се само пливати у мору информација или ће се ти подаци искористити да организације буду успешније. Програмске алате који помажу и омогућавају да се огромна количина података учини смисленом не користе програмери или дизајнери база података, већ бизнисмени који знају шта представљају ти подаци и чему могу да служе. Алати за Big Data морају да укажу на нове захтеве пословања и правац у коме треба деловати у процесу управљања компанијом [13].

2.3. Могућности Big Data

Данашње технологије не само да подржавају складиштење ових података већ дају могућност да се добијени подаци разумеју и да се искористи њихова вредност. Ово помаже организацијама да поправе своје пословање и профит. На пример, уз помоћ ових технологија могуће је [13]:

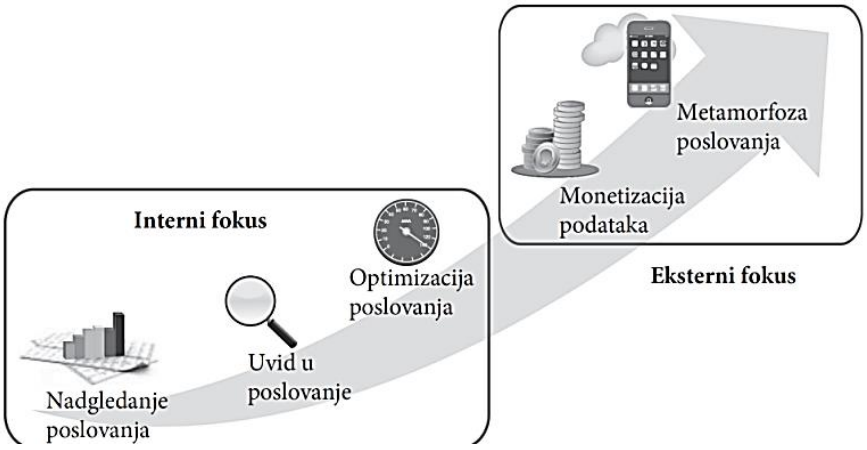
- Анализирати милионе тржишних производа да би се одредила оптимална цена, увећао профит или ослободило складиште.
- Прерачунавати ризике у минути и на тај начин се прилагођавати променама.
- Истраживање података везаних за потрошачке навике и потребе и на тај начин повећавање профита, подршке у изборним кампањама итд.
- Идентификовање најозбиљнијх купаца.
- Генерисање малопродајних купона за потрошаче, базираних на претходним куповинама. Ово осигурава већи откуп робе.
- Слати адекватне понуде мобилних провајдера на мобилне телефоне у правом тренутку, када ће корисник моћи да их искористи на најбољи начин.
- Анализирање података са средстава јавног информисања због сагледавања трендова.
- Одређивање главних проблема у функционисању мрежа и машинских сензора.

2.3.1. Big Data Business Model Maturity Index

Да би организације на прави начин искористиле све могућности које Big Data пружа, оне морају да сагледају фазу коју су достигле применом Big Data технологија, техника и алата и уоче које још могућности стоје пред њима. За ту сврху од помоћи је Big Data Business Model Maturity Index који је представио Bill Schmarzo, експерт са вишедеценијским искуством у области база података, аналитике и рада са великим подацима [11].

Фокус индекса је на свим оним пословним процесима до чије трансформације долази применом Big Data технологија, техника и алата. Значај индекса се огледа у томе што на основу њега организације могу стећи увид у којој се фази тренутно налазе и шта је то што би још могле да остваре са применом Big Data како би побољшале процес креирања вредности од прикупљених података. Big Data Business Model Maturity Index има пет фаза при чему су прве три фазе интерно оријентисане и односе се на оптимизовање интерних пословних процеса, док су последње две фазе екстерно оријентисане на купце, производе, тржиште [11].

На Слици 11 су приказане фазе Big Data Business Model Maturity Index-a, које ће појединачно бити објашњене у даљем тексту.



Слика 11: Фазе Big Data Business Model Maturity Index-a [11]

Фаза 1: Надгледање пословања

У фази "Надгледања пословања" организације имају традиционалне системе пословне интелигенције и складишта података помоћу којих надгледају своје пословање кроз различите извештаје, прате кључне индикаторе перформанси и користе могућности аутоматизованих упозорења на незадовољавајуће или лоше перформансе. У оквиру ове фазе постојећи информациони системи и технологије примењују се за [11]:

- Праћење трендова.
- Компарацију са претходним периодима
- Мерила резултата као што су развој бренда, перформансе производа, финансијски резултати, задовољство купаца.
- Процентуално учешће/ процентуални удео: удео на тржишту, учешће одређеног производа у укупној продаји и сл.

Фаза 2: Увид у пословање

Увид у пословање представља фазу у којој компаније почињу да прикупљају, обрађују и анализирају нове неструктуиране изворе података применом сложенијих статистичких метода како би се идентификовале пословне активности/процеси којима треба посветити више пажње и које треба третирати као кључне пословне процесе [11].

Организације које се налазе у овој фази имају бројне користи. На пример, маркетинг сектор може користити информациони систем да идентификује зашто нека маркетиншка кампања има већи ефекат од друге и усмери будуће инвестиције ка ефективнијим кампањама. Такође, праћењем активности купаца/клијената који су укључени у програме лојалности могу се идентификовати они који имају низак ниво активности (куповине) и они се могу стимулисати за већи ниво активности слањем e-mail-a са купонима на основу којих добијају одређене попусте [11].

Сам процес транзиције организације из фазе 1 у фазу 2 није једноставан и подразумева одређено време и напор. Пре свега, неопходно је разумети како корисници користе постојеће извештаје да идентификују пословне проблеме и могућности и како на основу њих доносе пословне одлуке. Након тога је потребно сагледати који су то спољни, нови извори података који су њима неопходни и како се они могу прикупити и обрађивати [11].

Фаза 3: Оптимизација пословања

Оптимизација пословања представља фазу у којој организација већ почиње да примењује комплексније аналитичке моделе којима аутоматски оптимизује своје пословне операције и процесе. Пример је усмеравање маркетиншког буџета на кампање које су се показале као најефективније [11].

Да би организација прешла из фазе 2 у фазу 3 потребно је да се сагледају све оне области које су дефинисане као кључне области оптимизације у фази "Увида у пословање" и да се идентификује листа процеса и активности које су кандидати за оптимизацију у зависности од њиховог пословног и финансијског утицаја на организацију [11].

Фаза 4: Монетизација података

Монетизација података представља фазу у којој организације на основу расположивих података траже могућности за развој новог пословног модела и креирање вредности. У овој фази расте значај свих оних података који потичу из различитих извора како би се даље креирали високо интелигентни производи као што су аутомобили који помоћу сензора запажају и уче понашање возача и аутоматски подешавају седиште, ретровизоре, контролну таблу итд [11].

Фаза монетизације података је карактеристична по томе што је најпре потребно идентификовати жеље и потребе циљних купаца детаљном и свеобухватном анализом њиховог понашања праћењем структуираних али и свих оних неструктуираних података из различитих извора. Прикупљене податке организације морају структуирати и трансформисати како би над њима могле да спроводе разне аналитичке операције [11].

Фаза 5: Метаморфоза пословања

Фаза метаморфозе представља последњу фазу коју организације могу достићи применом Big Data. У овој фази долази до промене фокуса са производа на креирање платформе/екосистема над којом ће се развијати или примењивати различити производи. За долазак у ову фазу потребно је уложити пуно времена у проучавање потреба и жеља купаца и клијената, разумети и сагледати постојећи екосистем и како се уклопити у њега, развити платформу на којој се могу развијати многобројни производи и услуге [11].

У погледу фаза које се могу достићи применом Big Data битно је истаћи да се неке организације крећу веома опрезно и успорено јер немају јасну визију на који начин желе да примене Big Data, одакле треба да почну, коју технологију треба да имплементирају и на који начин. Са друге стране, бројне организације се упуштају у брзо прихватање Big Data и његову интеграцију у постојеће пословне процесе желећи да прате савремене технолошке трендове, немајући јасну визију и стратегију. Оно што је битно знати је да не постоји јединствени оквир, рецепт, правило, процедура којих би организације требало да се придржавају приликом примене Big Data концепта. Свака организација је специфична и јединствена и самим тим различитим брзинама прихвата и примењује Big Data за креирање конкурентске предности [11].

2.4. Утицај Big Data на пословање и како се може искористити

Пре свега, треба нешто рећи о стратегији. Најједноставније би се могло рећи да је стратегија жељена али реална будућност компаније. Жељена јер одражава циљеве које менаџмент жели да оствари стратегијом, а реална јер менаџмент мора да сагледа шта се дешава у окружењу и да ли компанија има довољно потенцијала да оствари своје жеље.

У компанији постоје три нивоа на којима се доносе стратегије [19]:

- Стратегија за ниво предузећа (corporate-level strategy) којом се одређује будући развој компаније: компанија се одлучује за пословно подручје у којем ће обављати своје пословање.

- Стратегија за ниво пословних јединица (business-level strategy) које представљају организациони подсистеми који имају своје окружење и конкуренте са којима се суочавају. Свака пословна јединица усваја сопствену стратегију која мора бити у складу са стратегијом на нивоу компаније.

- Стратегија за ниво функционалних јединица (functional level strategy) где менаџери функционалних јединица доносе стратегије како би реализовали циљеве који су им задати.

Big Data има утицај на формулисање све три врсте стратегија обзиром да оне подразумевају анализу окружења, идентификовање могућности, вредновање и избор најбоље стратегијске алтернативе [19].

Једно је сигурно, а то је да компаније које уводе Big Data технологије формулишу стратегије којима настоје да изграде кључне способности за бржи и квалитетнији процес одлучивања како би успеле да креирају вредност на основу поплаве података са којима су суочене. Прикупљени подаци провучени кроз различите софтвере и приказани у различитим бојама на различите начине у суштини не значе ништа уколико се на основу њих не донесу пословне одлуке [19].

Без обзира на то које циљеве себи постави, компанија која имплементира Big Data -у мора имати план на основу ког ће компанија прикупљати жељене податке, обрађивати их и помоћу њих доносити одлуке. Експлозија дигиталних података ставља пред менаџере велике изазове од којих је један на врху по тежини изазов које податке треба прикупљати.

Што се тиче нивоа пословних функција: набавка, производња, маркетинг, продаја, финансије, информационе технологије, људски ресурси, истраживање и развој. Свака од тих функција има своје стратегије којима реализује циљеве. И свака од тих функција може да има велике користи од Big Data-е.

Једна од првих функција у компанији која се мења због Big Data-е јесте функција информационих технологија. Нова стратегија треба да припреми план шта је то што се мора урадити да би се имплементацијом Big Data технологија остварили циљеви компаније. Ту углавном спадају одлуке о неопходној инфраструктури – хардверу и софтверу, времену потребном за импелементацију и трошковима. Менаџерима стоји на располагању широка понуда алата различитих карактеристика и произвођача која се свакодневно увећава. Такође, сама функција информационих технологија формулише своју стратегију како да се на основу имплементиране технологије прикупљају, обрађују, приказују и чувају подаци [12].

Прва функција која уочава потребу да примењује Big Data -у јесте маркетинг функција. Маркетинг подразумева 4 корака: анализу потенцијалних купаца, привлачење њихове пажње, постизање заинтересованости купаца и прихватање постојеће понуде. Сва 4 корака зависе од маркетиншких активности организације. Big Data нуди велике потенцијале због обухвата свих оних података са друштвених мрежа, блогова, умрежених уређаја који одражавају потребе и жеље потрошача, њихове коментаре, евентуалне примедбе и слично. Клијенти бројних туристичких агенција у свету се већ могу похвалити да добијају све боље сугестије за рекреацију, хотеле, разне туристичке аранжмане, наравно, иза завесе је Big Data која процењује њихове потребе, навике, жеље, интересовања [12].

Big Data је корисна за све индустрије, а у малопродаји је посебно значајна за предвиђање навика купаца, оптимизацију ланца снабдевања, до праћења резултата промоције производа. Мање компаније, такође, могу да користе ове анализе и стекну предност над конкуренцијом [12].

У системима који функционишу у реалном времену, често је неопходно да се за веома кратко време анализирају подаци који се брзо мењају. Један од таквих примера је заштита од превара у банкарству или коцкању. Да би ухватили корак и успешно обавили тај задатак, стручњацима у ИТ одељењима је потребан радикално нови приступ. Такође, поставља се питање шта је од те огромне количине података релевантно за компанију? Зато се у подједнако важне циљеве убрајају селектовање и чување само оних података који заиста имају вредност за предузеће [12].

Где се Big Data користи? Кратак одговор је – свуда. Не постоји индустрија или одељење у компанији којем велика база података није од користи, од пољопривреде, преко производње, до малопродаје. У студији коју је IDC (International Data Corporation) спровео са Аустријским технолошким институтом (AIT) за потребе Министарства за иновације у Аустрији, идентифковано је 18 привредних грана у којима коришћење великих база података може бити корисно. Али кључно питање није да ли би нека компанија или њена подружница могле да имају користи од велике базе података, већ – како те податке употребити? Јер, њихово коришћење захтева пуно знања о индустрији у којој компанија послује, као и знања како добро управљати подацима и конкретним технологијама као што су Hadoop (софтверске библиотеке које омогућавају коришћење података на великој мрежи компјутера уз коришћење једноставних техника). Да би се велике базе података максимално користиле, потребно је анализирати рад компаније из различитих углова [12].

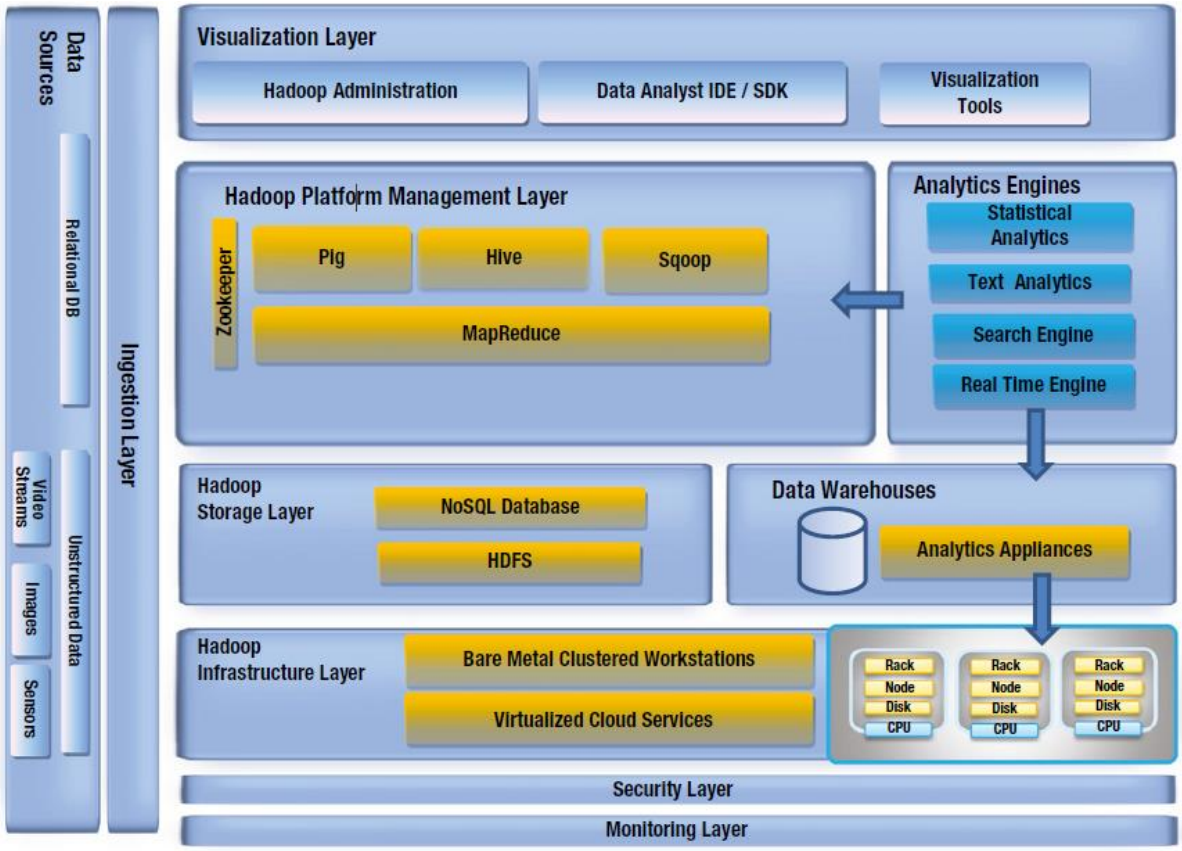
Различита одељења унутар компаније могу, такође, имати користи од великих база података. Тако ће запослени у одељењу маркетинга боље разумети своје купце ако буду анализирали њихове потребе и начин на који реагују, на пример, преко друштвених мрежа. Одељења људских ресурса могу да стекну драгоцене увиде у погледу запошљавања нових људи, тако што ће анализирати друштвене мреже или било које друге расположиве податке. Јер, послодавцима је данас важно да пронађу кадрове који, поред стручних квалификација, деле вредности предузећа и могу добро да се уклопе у тимски рад. Ако је реч о важности велике базе података за одељење финансија, финансијски директор компаније не само да жели да упореди пословање предузећа у односу на исти период претходне године, већ му је битно да пројектује и какви ће резултати бити на крају године. Данас постоји више информатичких решења која омогућавају да се подаци једноставно "изваде" из базе ради упоређивања резултата пословања са истим периодом у претходној години, и комбинују са уговореним наруџбинама/ продајом, у систему управљања продајом. Овакве анализе омогућавају боље разумевање и предвиђање развоја посла и значајно доприносе томе да се елиминише било каква случајност [12].

3. Архитектура Big Data апликације

До сада је више пута споменуто да постоји велики број извора података, да подаци настају великом брзином, да се ради о великој количини података раличитог типа, зато је потребно да организације имају системе тако развијане да би се добили корисни закључци од тих података.

Управљање подацима, складиштење и методе анализе морају бити другачије од досадашњег традиционалног управљања, како би донело вредност организацији, јер велике податке треба посматрати као ресурс који се може искоритити, али потребно је знати како.

Пре него што се крене у анализу великих података, битно је да ли се у оквиру архитектуре налазе сви потребни елементи, јер без правилног подешавања тих елемената тешко да се може доћи до правих вредности. Архитектура система за управљање великом количином података треба да на брз и јефтин начин обради тако велику количину података.



Слика 12. Пример архитектуре за Big Data [2]

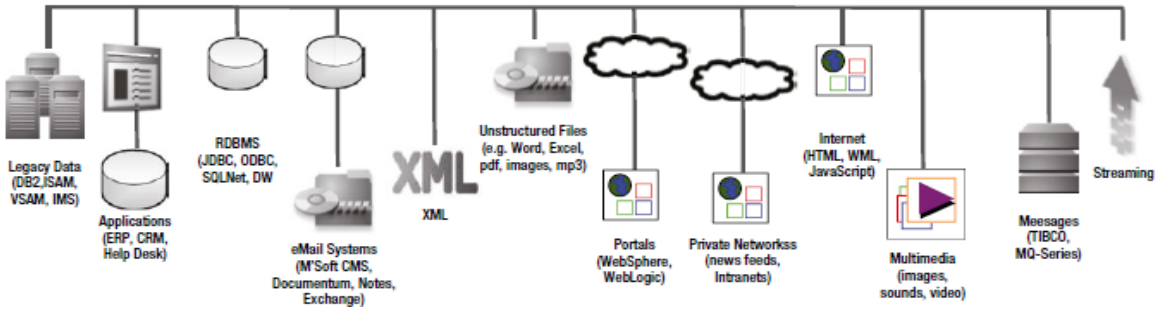
На Слици 12 приказане су компонете које би требало да садржи сваки систем који обрађује велику количину података. Наравно, у зависности од сврхе коју систем обрађује, системи се разликују, али ово је општи преглед. Могу се изабрати open source frameworks или packaged licensed products како би се у потпуности искористила функционалност различитих компоненти. Као што се може видети на слици, овај тип архитектуре садржи неколико „слојева“ (layer-а), посматрано, с лева на десно на Слици 12 то су [2] :

- Извори података *(Data Sources)*
- Слој који се односи на унос података у кластер односно у HDFS *(Ingestion layer)*
- Слој Hadoop складиштења података *(Hadoop Storage Layer)*
- Слој система за управљање заснован на Hadoop платформи *(Hadoop Platform Management Layer)*
- Слој Hadoop инфраструктуре *(Hadoop Infrastructure Layer)*
- Слој визуализације *(Visualization Layer)*
- Заштитни слој *(Security Layer)*
- Слој за праћење *(Monitornig Layer)*
-

3.1. Извори података *(Data Sources)*

Прави проблеми тумечења великих података почињу још у у изворима самих података, јер су извори различити, а самим тим различите су количине података које пристижу различитим брзинама. Све „V“-ове споменуте у уводном делу, треба укључити у један скуп, и анализирати. Ове велике скупове података називају се још и језара података, које треба истражити и од којих треба направити базене, односно направити образац по коме ће се ти исфилтрирани подаци слати даље у Hadoop Framework. Ово заправо предаставља задатак алата, који су задужени за слој уноса података у фајл (ingestion layer) [2] .

На Слици 13 су приказане различите врсте извора података.



Слика 13.Извори података [2]

3.1.1. Извори структуираних података

Иако се чини да су структуирани подаци добро познати, заправо у свету Big Data добијају нову улогу. Развој технологије омогућава појаву нових извора структуираних података - често у реалном времену и у великим количинама. Извори података се деле у две категорије [4]:

- Рачунарски или машински генерисани - појам се обично односи на податке које производи машина без људског утицаја.

- Људски генерисани - ово су подаци које обезбеђују људи у интеракцији са рачунарима.

Неки стручњаци тврде да постоји и трећа категорија која представља хибрид између две наведене категорије.

Машински генерисани структуирани подаци укључују:

- Сензорске податке:

Примери за ову врсту података су: радио фреквенцијске таласи (RFID), паметни мерачи (нпр. електронска бројила за мерење потрошње електричне енергије), подаци са медицинских уређаја, GPS podaci. На пример, RFID убрзано постаје популарна технологија. Користе се минијатурни рачунарски чипови да би се уређаји пратили са удаљености. Пример овога је праћење контејнера са производима од једне до друге локације. Када пријемник добије информације оне могу бити прослеђене серверу где ће бити анализиране. Компаније су заинтересоване за ову технлогију због управљања транспортом робе и контролу инвентара. Још један пример извора сензорних података су паметни телефони који имају сензоре као што је GPS који могу бити коришћени за разумевање понашања потрошача на нови начин [4].

- Web лог податке:

Када сервери, апликације, мреже и слично раде они бележе различите податке о својој активности. Количина ових података може постати огромна, а ови подаци могу бити искоришћени за, на пример, предвиђање нарушавања безбедности [4].

- Податке у тренутку продаје:

Када радник на каси очита бар код било ког производа који купујете, генеришу се сви подаци везани за производ. Ако се размисли колико људи свакодневно купује различите производе може се схватити колико је количина ових података огромна [4].

- Финансијске податке:

Доста финансијских система су данас програмирани, њихов рад се заснива на предефинисаном скупу правила што аутоматизује процес. Подаци о трговању на берзи су добар пример овога. Садрже структуиране податке као сто су ознака компаније и вредност у доларима. Неки од ових података су машински генерисани а неки људски генерисани [4].

Примери људски генерисаних структуираних података укључују [4]:

- Улазне податке:

Ово је било који тип података који човек може унети у рачунар, као што је име, презиме, године старости, приход, одговори на анкете и слично. Ови подаци могу бити коришћени за разумевање основног понашања потрошача.

- Клик податке:

Сваки пут када се кликне на линк на сајту подаци се генеришу. Ови подаци могу бити анализирани да би се одредило понашање потрошача и обрасци куповине.

- Податке везане за игре:

Сваки потез који се направи у игри може бити забележен. Ови подаци могу бити корисни за разумевање како крајњи корисници играју игру.

Неки од ових података не морају бити велики сами по себи, као што су профилни подаци. Међутим, када се обједине подаци милиона корисника који шаљу информације, количина података постаје огромна. Додатно, много ових података је везано за време у ком се генеришу што може бити корисно за разумевање образаца који имају потенцијал за предвиђање исхода. Поента је да ове информације могу бити од вредности и могу бити коришћене у различите сврхе [4].

3.1.2. Извори неструктуираних података

Неструктуирани подаци су подаци који не прате неки дефинисани формат. Ако је 20% података који су доступни предузећима структуирано, преосталих 80% је неструктуирано. Неструктуирани подаци су заправо подаци који се најчешће срећу. До скоро, међутим, технологија није подржавала друге начине рада са овим подацима осим складиштења и ручне обраде. Неструктуирани подаци се могу наћи свуда. Заправо, већина људи и организација функционише на основу неструктуираних података. Као и у случају структуираних података и неструктурирани подаци могу бити машински или људски генерисани [4].

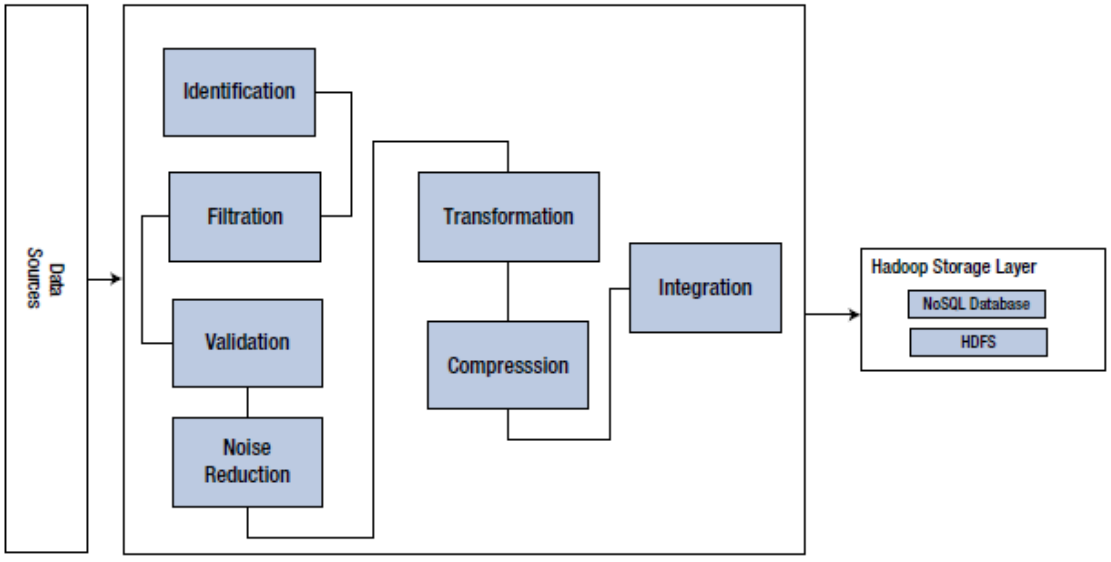
Неки примери машински генерисаних неструктуираних података су [4]:

- Сателитске слике: Ово укључује податке о временским приликама или податке које владе прикупљају приликом сателитског надгледања. На пример, ГооглеЕартх поседује огромну количину сателитских снимака које обрађује и спаја на одговарајуди начин.

- Научни подаци: ово укључује сеизмичке слике, атмосферске податке, физику високих енергија, итд.

3.2. Унос података са извора података у HDFS *(Ingestion Layer)*

Ingestion Layer је део архитектуре који има задатак да изабере валидне податке из гомиле података. Што значи да треба да буде у стању да обради огроман обим података веома брзо. Требало би да има способност да потврди, очисти, трансформише, смањи, и интегрише податаке и пошаље на даљу обраду. Уколико овај део система није адекватно адаптиран цео систем ће бити нестабилан, јер алати који се налазе у овом слоју дужни су да елеминишу непотребне податке [2].



Слика 14. Особине слоја за унос података у Hadoop Storage Layer [2]

Слој уноса података учитава коначне релеванте податке на дистрибуирани Hadoop слој за складиштење, односно уноси податке HDFS. Особине које треба да садржи овај слој су следећи, а приказани су на Слици 14 [2]:

- Идентификација *(Identification)* - Идентификовање различитих формата података или пренос подразумеваног формата у неструктуриране податке.
- Филтрирање *(Filtration)* – Филтрирање долазних информација које су релавантне за предузеће
- Провера и анализа *(Validation)*- Провера података, да ли су подаци валидни за даљу обраду.
- "Смањење буке*" (Noise Reduction)* - Обухвата чишћење података и минимизирање поремећаја.
- Трансформација *(Transformation)* - Може укључивати дељење, денормализаију или сумирање податка.
- Компресија *(Compression)* подразумева смањење величине података, при чему се не губи релевантност података у процесу. То не би требало да утиче на резултате анализе након компресије.
- Интеграција *(Integration)* - Представља интегрисање коначних података у следећи слој а то је Hadoop storage layer (HDFS и NoSQL базе података).

Дакле, приликом уноса податка најпре треба обратити пажњу на брзину података, односно треба утврдити којом брзином пристижу подаци у систем. Генерална подела података у том смислу би била на статичне и нестатичне податке [2].

Статични подаци су они који се не повећавају након уноса. Прави пример за овакве податке су генетски подаци. Могу се наравно додавати нови подаци у систем, али када се унесу подаци, анализе се врше само над њима. На пример добије се 1TB генетских података и све анализе врше се над њима. За статичне податке је лакше одабрати алат за трансфер у Hadoop кластер из простог разлога што трансфер није чест [2].

Насупрот статичним, нестатични подаци се стално увећавају. Добар пример су социјалне мреже или log подаци – то значи да ако је постојало 2TB log података, за кратко време настаће још толико. У оваквим ситуацијама мора се пажљиво одабрати алат како би се процес аутоматизовао [4].

Веома је важно да се осим иницијалног уноса података у HDFS предвиди колики ће прилив бити убудуће и на које време. То је битно кад се на пример подаци унесу у HDFS данас, па тек за месец дана, и такав процес је паметније аутоматизовати него трошити време на писање сложених скрипти. Ако са друге стране подаци стижу свакога дана, свакако је паметније да је систем аутоматизован. Количина је битна из разлога планирања канала преноса. Наиме, неки алати дају могућност одабира пропусног опсега. Ако пристижу мале количине података, онда треба дефинистати параметре према томе како се не би оптеретио систем .

Постоји неколико начина за пренос података у Hadoop кластер, односно у HDFS . Неки од алата на тржишту који олакшавају унос података су: *Apache Sqoop, Flume, Storm, Korišćenje HDFS komandi.*

3.2.1. Алати за пренос података у HDFS

Sqoop је алат који служи за пренос података из и у Hadoop кластер. Његова посебна намена је да ради са релационим базама података. На Слици 15 приказан је лого Sqoop-а, а који заправо представља и његову улогу, а то је пренос података из релационих база података у Hadoop фајл систем. Осим трансфера података у HDFS, Sqoop може директно да ради са Hive-ом или HBase-ом. Предност Sqoop-а је што брзо преноси податке унутар HDFS-а [19].

Слика 15. Лого алата Sqoop [19]

Осим што је брз, Sqoop је јако флексибилан, па уколико се не треба да се уносе цела база, може се изабрати табела која је потребна, при чему није потребно знати њен назив, јер могу да се излистају табеле. Sqoop се користи у ситуацијама када се користи алат за визуелизацију а при томе нема конектор са Hadoop-ом, а подаци морају да буду у некој бази. Подаци који се извозе из Hadoop-а помоћу Sqoop-а су структуирани. Исто тако, често се Hadoop кластер повезује са ERP (Enterprise Resource Planning) системима који користе неку релациону базу и опет се примењује Sqoop. Споменути примери односе се на трансфер података ван Hadoop-а. Међутим, Sqoop се може користити за трансфер података унутар Hadoop-а. Ако је база такве величине да традиционални алати не могу да изађу на крај са њом, онда се користи Sqoop [19].

Интересантан пример за коришћење Sqoop-а је ако се сакупљају подаци који су сензорки и у log-у постоје кординате. Приликом анализе се ти бројеви могу претворити у име локације, у случају да постоји табела која описује координате и име места. Једноставно се уносе табела у Sqoop, укрсте се подаци и са лакоћом се добију читљиви резулатати. Ово је добар пример да се покаже како користити релационе базе са Hadoop-ом, када имамо падатке који се стално понављају. Можда је боље рећи да се овакве табеле могу користити као шифараници, а при томе не заузимају пуно места, лако се уносе у Hadoop и лако се могу искористити са неким неструктуираним подацима [19].

Apache Flume је дистрибуиран и изузетно поуздан сервис за сакупљање, агрегацију и транспорт великих количина података. Првенствено је осмишљен за пренос података, али са еволуцијом Hadoop екосистема постаје одличан сервис за тзв. streaming податке. Има једноставну архитектуру, јако је отпоран на губитке података и има могућност брзог опоравка. Може се рећи да су три најбитнија дела Flume-a source (извор), channel (канал) i sink (одредиште) [19].

Најлакше је описати овај алат примером из реалног света. Уколико имамо огромну шуму на врху планине и треба да се посече и транспортују стабла до града у подножију, питање је како то урадити. Одговор је уз помоћ природе, односно уз помоћ брзе планинске реке која се спушта до града. Дакле, потребно је ставити стабла у воду и вода ће без муке пренети стабла у град. Аналагно томе, у Big Data свету постоје подаци, односно стабла, шума је извор стабала, а извор података може бити нека друштвена мрежа, милион сензора или лог фајл неког јако посећеног сајта. Наравно, град где се обрађују стабла, у овом дигиталном свету, је HDFS. Отуда и следећа Слика 16 која представља лого овог алата.

Слика 16. Лого алата Flume [19]

Flume се не користи за пренос релационе базе података у HDFS, а уколико су у питању streaming подаци, Flume је одличан избор. Подаци који се могу сакупљати овим алатом, а да при томе буду корисни при анализи и доношењу одлука су подаци са Twitter streaming друштвене мреже, а такође се може користити и за велику количину log података [19].

3.2. Складиштење података и велики подаци- *Distributed (Hadoop) Storage Layer*

Традиционална складишта узимају доста времена и доста труда који се улаже у чишћење података, обогаћивање, метаподатке, master data management итд. То аутоматски подразумева и велики квалитет тих података. Ради се о скупом процесу чији је исход висока вредност и широка примена. С друге стране наши Big data системи ретко пролазе тако стриктне фазе предпроцесирања, јер су скупи те се рад на овим системима углавном своди на истраживања и откривање него на вредност података. Ово је кључна разлика између складишта података (подизање квалитета података за израду квалитетних извештаја) и Big data (проналазак занимљивих података релативно јефтиним процесом, које даље можемо убацити у скупа складишта података). Дакле, технологија Big data "копа" кроз огромне количине прљавих података у потрази за златом и кад га пронађе оно се прочисти, структурира и напуни у складиште података чиме се искоришћава његова пуна вредност [2].

Што се тиче софтвера за складиштење и обраду великих података тренутно је најактуелнији Hadoop. Неке фирме попут IBM-а и HP-а су узеле делове Hadoopa и створиле своје платформе, али изучавајући ову област, Hadoop се провлачи свуда. На основу до сада одржаних конференција везаних за Hadoop, може се рећи да је он постао стандардни софтвер за Big Data. Наравно, ту је Google од кога је све кренуло, али њихов софтвер није за употребу од стране других компанија [2].

Укратко речено Hadoop је open-source software framework Apache фондације. Служи за складиштење и процесирање великих количина података. Hadoop је настао 2005. године од стране Doug Cutting и Mike Cafarella. Име је добио по слону, играчки Cutting-овог сина. Написан је у Java програмском језику. Када се каже велика количина, мисли се на више TB (terabytes), а можда је боље рећи и PB (petabytes). У данашње време све је више података и кључна ствар је да се из података извуку информације, да се употребе и да се стекне профит. Друштвене мреже, као што су Facebook и Twitter, морају обрадити податке тако комплексно, а притом то треба урадити што брже да би победили конкуренцију [2].

Најбитнији делови Hadoop-а су:

- HDFS (Hadoop Distributed File System) – фајл систем
- MapReduce модел

3.2.1. HDFS- Hadoop дистрибуирани фајл

HDFS је веома битна компонента јер је управо то фајл који је способан да поднесе огромну количину података. То је практично фајл систем који користи Hadoop и овде је дефинисано како се подаци складиште, копирају и читају. HDFS је заслужан за могућност лаког складиштења огромне количине података [19].

Слика 17. Лого Hadoop дистрибуираног фајл система [19]

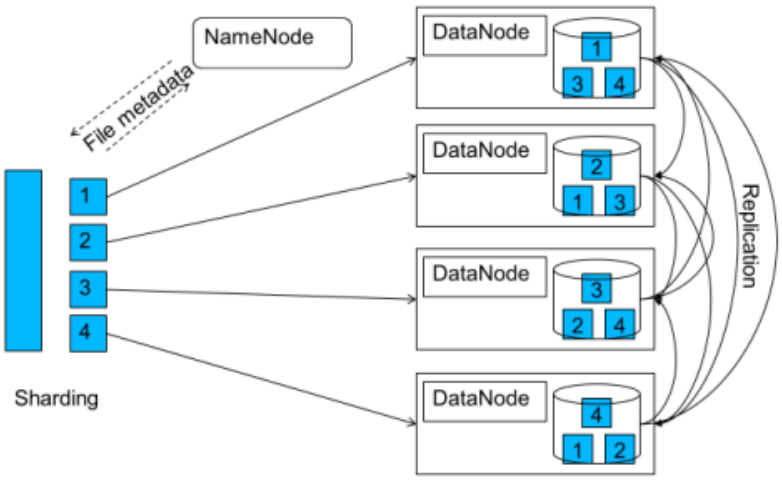
HDFS је Hadoop дистрибуирани фајл систем и саставни је део Hadoop-а. Дизајниран је тако да може да ради на било којој хардверској структури. Овај фајл систем је јако отпоран на грешке и није хардверски захтеван. Погодан је за складиштење велике количине података [19].

Што се тиче одлика самог HDFS-а, намењен је за стотине, односно хиљаде сервера који имају различите компоненте за које постоји вероватноћа отказивања, што одавде следи да је главни циљ HDFS архитектуре да брзо открије грешке и да их аутоматски уклони.

HDFS је намењен за баш велике количине података, типично један фајл који се увезе у њега је реда GB (gigabytes) и више, што значи да је брзина уписа података веома битна. Има једноставан модел за приступ фајловима – упиши једном, читај више пута. HDFS није намењен за интеракцију са корисником, него је више окренут несметаној обради података од стране других алата. Ако се директно приступи машини на коју је инсталиран, тешко или готово никако се могу наћи фајлови који су унети, а да се прочитају у људима видљивом формату [19].

NameNode i DataNode

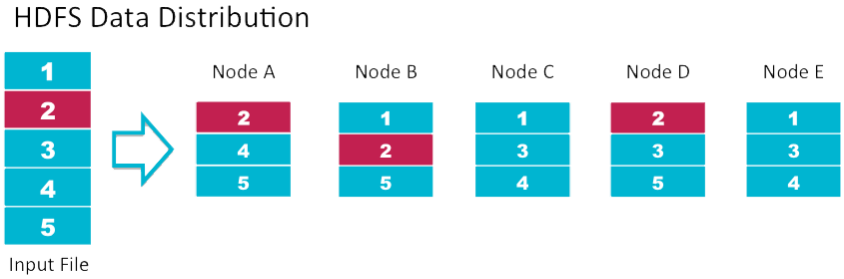
Након неких основних информација о HDFS, питање је како заправо све то ради. HDFS је дизајниран тако да има master/slave архитектуру, што је приказано на Слици 18. То би значило да један Hadoop кластер има један и само један NameNode и више DataNode-ова. Обично је један сервер мастер и на њему се инсталира NameNode, а на осталима DataNode-ови. Мастер сервер на коме је инсталиран NameNode контролише приступ фајловима и управља именским простором (The File System Namespace), који подржава традиционалну хијерархију. Корисник може да креира директоријум и у њему складишти фајл, може да мења име директоријума, да га брише и премешта. Исто важи и за фајл. Корисник управља фајловима на мастер серверу, види директоријуме и фајлове. Даље се фајлови деле у блокове и складиште на остале сервере на којима је инсталиран DataNode. DataNode служи за складиштење блокова, дозвољава креирање блокова, брисање и понављање (веома битна могућност) [19].



Слика 18. Master/slave архитектура [19]

Уколико се претходно објашњење сумира, то би изгледало овако: NameNode чува мета податке блокова и ту је да помогне да крајњи корисници виде фајл, а не блокове који нису читљиви, док DataNode чува податке. Ово би се могло посматрати као да се у NameNode-у чувају адресе блокова, име и број копија [19].

Понављање блокова је веома битно. HDFS има ту могућност да се блокови уписују више пута, што је приказано на Слици 19. Ово смањује ризик од губљења података. Подразумеван број копија је три, што значи да ће се сваки блок копирати три пута. То не мора да буде на једном серверу, већ на свим серверима у кластеру на којима је инсталиран DataNode. Тако се смањује ризик губљења података. За 1TB података ће бити потребно 3TB простора, али добијате сигурност, што је посебно важно ако се ради о битним фајловима. Могуће је подесити да број копирања буде и већи, па да се тако додатно осигурају подаци. Одавде се извлачи и закључак да је непотребно инсталирати Hadoop на једном серверу [19].



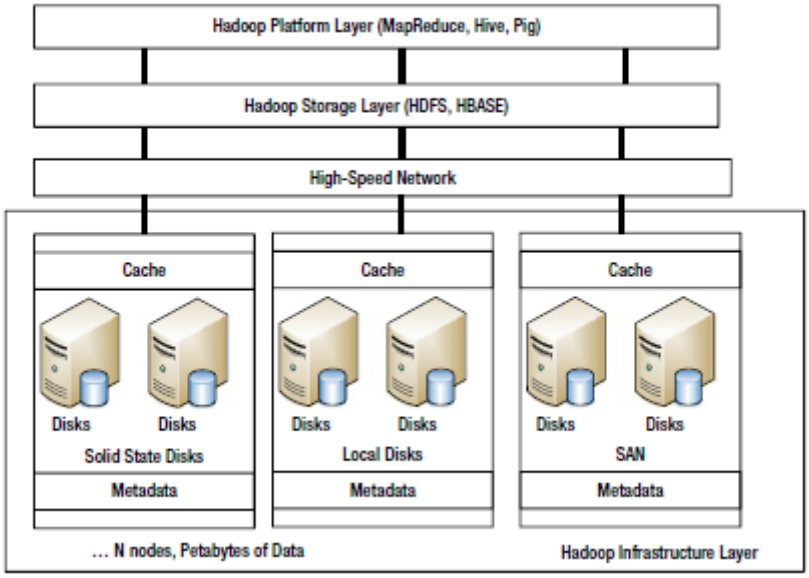
Слика 19. Понављање блокова на серверу [19]

Комуникација између сервера у кластеру се одвија помоћу TCP/IP протокола, па самим тим и NameNode комуницира са DataNode-овима. Поред тога DataNode-ови периодично шаљу такозвани Heartbeat NameNode-у како би све функционисало у најбољем реду [19].

3.3. Hadoop платформа систем за управљање *(Hadoop Platform Management Layer)*

Ово је слој који обезбеђује алате и упитне језике за приступ NOSql базама података користећи HDFS складиштени фајл систем, који седи на врху слоја физичке Hadoop инфраструктуре [2].

Са развојем компјутерске технологије, сада је могуће управљати огромним количинама података којима су претходно могли руковати само суперкомпјутери, што је представљало велики трошк. Пошто су цене система (CPU, RAM и DISK) опале, као резултат тога, нове технике за дистрибуирање су постале основа. На Слици 20 је приказана једна како изгледа једна од платформи [2].



Слика 20. Платформа Big Data архитектуре [2]

Hadoop и Map Reduce су нове технологије које омогућавају предузећу складиштење, приступ и анализу огроманих количина података у скоро реалном времену, тако да оне могу да уновче предности поседовања велике количине података. То значи да се ове технологије хватају у коштац са једним од најосновнијих проблема - способност да се обради огромна количина података ефикасно, економично и благовремено [2].

3.3.1. Map Reduce

MapReduce је програмски модел за процесирање великих количина података. Битно је напоменути да се његов алгоритам извршава паралелно и дистрибуирано. Ближе говорећи, ако се посматра кластер од три сервера, алгоритам ће се извршавати паралелно, односно у исто време на сва три сервера [19].

MapReduce је систем базиран на YARN-у и представља модел паралелног процесирања велике количине података. Ова компонента је посебно зсначајна Hadoop програмерима јер све што се програмира за Hadoop, било да се ради у Javi, Pythonu, Pigu или било ком програмском језику, мора да поштује овај модел [19].

Када би се систем састојао само од Hadoop-а са HDFS-ом и Common пакета, то би била само платформа за складиштење велике количине података. Поставља се питање зашто само складиштити податке, то је глупо, а и скупо. Зато Hadoop има посебну компоненту а то је MapReduce [19].

Овај модел се може посматрати и као две одвојене целине, део Map и део Reduce.

- *Map* – Maper служи да сортира и филтрира податке. Ово може да буде и неко једноставно сортирање, све зависи шта је потребно. То може бити сортирање студената по презимену, на пример.

- *Reduce* – Reducer сумира податке које је обрадио Maper, односно на датом примеру сада лакше можемо избројати студенте по неком критеријуму.

Maper или Reducer се могу срести и у функционалном програмирању. На пример, обичан Bubble sort је Maper који сортира податке. Међутим, вредност овог модела је управо оно што заједно чине Mapere и Reducere [19].

Пре него што је почео да се користи у Hadoop-у, MapReduce модел се користио у Google-у, али није био толико искоришћен. Права моћ овог модела се види када се комбинује са HDFS-ом [19].

У Hadoop верзији 1 MapReduce је служио за управљање ресурсима и за процесирање података, док је сада, у верзији 2, управљање ресурсима преузела компонента YARN [19].

Требало би напоменути логику овог модела, односно структуру података коју поштује MapReduce. Map функција и Reduce функција поштују структуру пар (кључ, вредност), о којој ће бити речи у поглављу 4. Map функција на почетку прима један пар (кључ, вредност) у неком домену (података) и враћа листу парова у другачијем домену [19].

Map (key_1, value_1) —> list (key_2, value_2)

Ово се извршава паралелно. Када се заврши тај део, MapReduce сакупља све исте кључеве из листе и групише их. Након груписања наступа Reduce функција која се примењује паралелно на сваку групу [19].

Reduce (key_2, list (value_2)) —> list (value_3)

Разлика у односу на функционално програмирање је то што Reduce функција враћа лист (кључ, вредност) као листу свих вредности, док код функционалног програмирања добијамо само једну вредност [19].

Оно што представља праву вредност овог модела је његова комбинација са HDFS. Већ је писано о HDFS-у. Дакле, сви подаци у HDFS-у се складиште на DataNode-у као блокови и на NameNode- у се чувају мета подаци о тим блоковима. Логички се намеће закључак да када су подаци подељени у мањим блоковима, лакше их је обрадити. Са оваквим складиштењем података лакше је Map функцији да их групише. Битно је да MapReduce модел ради са подацима, односно блоковима, и то никако не значи да се за пар (кључ, вредност) узима било шта из NameNoda. Пар (кључ, вредност) значи "групиши ми сва имена студената по години студија из свих блокова (без копираних)", а онда Reduce све то спакује и добију се резултати са свих блокова у било ком серверу изабраног кластера [19].

MapReduce модел има практичну примену у Hadoop-у, односно може се рећи да програмер има додира са овом компонентом. Као што је напоменуто, Hadoop само са HDFS-ом је BigData складиште, што значи да ту постоје само подаци и ништа више Данас није битно имати новац да би имали моћ, потребно је имати информацију коа се може ваљано искористити. MapReduce је битан део претварања података у информације. Данас велике фирме попут Facebook-a, Twitter-a и осталих користе управо Hadoop како би сазнале навике и жеље својих корисника и то искористили да побољшају услугу [19].

У Hadoop кластеру MapReduce се односи на job (posao) koji se дели на мање делове или задатке (tasks). Апликација прихвата посао те га додељује одређеном Hadoop кластеру који покреће JobTracker. Он комуницира са NameNode како би сазнао где се све подаци који су потребни налазе унутар кластера те се посао (job) дели на мање делове тј. taskove или како је раније речено map task и reduce task за сваки node. Такође, битан је и TaskTracker којем је посао да прати статус сваког task-a или задатка. Ако задатак не успе његов статус се шаље JobTrackeru, који ће исти тај задатак поновно доделити новом node-у унутар кластера [19].

3.3.2. Програмски језици– пројекти повезани са Hadoop-ом

Pig

Име је добио по животињи која је позната по томе што може појести готово све (свиња) те се учинило згодним тако назвати овај програмски језик, јер може управљати с било којом врстом података. Pig се састоји од два појма: PigLatin – језик и другог дела који пружа окружење првоме да се изврши. Циљ Piga је поједноставнити MapReduce програме. Кораци како му то полази за руком су LOAD, TRANSFORM, DUMP i STORE. Како би Pig могао радити програм му мора рећи које податке да користи, а то се ради преко наредбе LOAD `data_file`, затим креће манипулација помоћу TRANSFORM, где је могуће филтрирати, груписати, спајати податке итд., затим на крају долазе DUMP и STORE чија употреба зависи да ли желимо приказати резултате на екрану (DUMP) или похранити податке за даљу анализу (STORE). Након овога потребно је покренути Pig унутар Hadoop окружења и то помоћу три начина : уграђивањем у скрипту, уграђивањем у Java програм или преко Pig command line званог Grunt. Који год одабрали, на крају долази до извршавања мап и редуце задатака чиме се испуњава циљ Piga, а то је поједностављивање целог процеса [19].

Hive

Pig увелико помаже да се поједностави цели процес, као што се може видети у прретходном тексту, али и даље је нешто што се мора научити и савладати. Како би олакшали цели процес још више, створен је Hive. Он помаже људима који су до сада радили на SQL-у, да уз слично окружење боље искористе Hadoop окружење. SQL девелопери при раду с њим користе тзв. HQL – Hive Query language који има одређена ограничења, али је и даље веома користан. Његове наредбе се деле на MapReduce послове и извршавају се кроз Hadoop кластер. Hive се базира на Hadoop и MapReduce операцијама, али постоје неке разлике. Због тога што је Hadoop направљен за секвенцијално скенирање, очекују се упити којима треба дуго да се изврше. Уколико је потребан брз одговор онда ово представља проблем. Друго, Hive је read-based што обично укључује велики део писаних операција [19].

ZooKeeper

Ради се о Open Source Apache пројекту који осигурава централизирану инфраструктуру и услугу која осигурава синкронизацију кроз кластер. Уколико имамо и мању количину сервера нужна је централизација кад се говори о управљању, а поготово кад се ради о великом броју сервера. ZooKeeper сервер чува копије стања целог система и сваки клијент комуницира преко једног ZooKeeper сервера (може их бити више), како би вратио или надоградио информацију о синхронизацији [19].

Hbase

Hbase је систем за управљање базама података које су оријентисане ка колонама оријентисане и покреће се над HDFS-ом. Веома је користан за рашчлањене податке, који су веома чести у Big Data случајевима. Битно је нагласити да Hbase не подржава SQL и није релацијска база података. Његове апликације су написане у Javi. Сам систем је веома сличан традиционалним базама података, а главна разлика су column famlies који омогућава похрањивање елемената column фамилија заједно. Код традиционалних система се колоне одређеног реда складиште заједно. Сама шема Hbase је веома флексибилна и веома је лако променити column families. Hbase слично HDFS-овом NameNode i MapReduce-ovom JobTrackeru и TaskTrackeru, има master node (управља кластером) и region server (складишти делове таблица и врши операције над подацима) [19].

HCatalog

Apache HCatalog је алат који омогућава лакше управљање складиштењем података и табелама на Hadoop-у, што пружа могућност корисницима разних алата за анализу података да лакше читају и пишу податке. Практично је HCatalog слој на Hadoop-у који омугућава презентовање података са HDFS-а у виду табела и ослобађа кориснике брига о томе где и у ком формату су сачувани подаци. HCatalog може да прикаже податке који су у RCFile формату, текстуалне фајлове, секвентне фајлове и све их приказује у табеларном облику. Још једна корисна ствар која долази уз HCatalog је и REST API који омогућава приступ табелама и спољним корисницима [19].

HCatalog подржава читање и писање фајлова у форматима за које је могуће написати Hive SerDe (Serializer-Deserializer), али подразумевани формати су већ споменути. Такође, JSON су подржани као подразумевани. HCatalog је направљен да ради са Hive metastore сервисом и укључује Hive DDL компоненте. Такође HCatalog пружа и интерфејс за Pig и MapReduce. HCatlog заједно са Hive-ом веома личи на познати SQL, релационе базе података, разлика је у нијансама [19].

3.5. Big Data инфраструктура (*Hadoop Infrastructure Layer)*

Слој који подржава слој за складиштење података (Storage Layer) је физичка инфраструктура, која представља основу за рад целог система великих података. За разлику од инфраструктуре традиционалних података, ова инфраструктура мора бити другачија, како би подржала све „V“ - ове (количину, брзину и разноликост) Big Data [2].

The Hadoop physical infrastructure layer (HPIL) заснован је на дистрибуираном рачунарском моделу. То значи да подаци могу да буду физички сачувани на различитим локацијама и повезани су кроз мреже и дистрибуиране фајл системе. Традициналне апликације предузећа су изграђене на основу вертикалног скалирања хардвера и софтвера. Оваква врста је скупа за велике податке [2].

HADOOP и HDFS могу управљати овим слојем у виртуелном облак окружењу или преко дистрибуиране решетке робних сервера преко брзог gigabit network. Организације данас могу да бирају између две врсте облака. Традиционални облака нуди виртуелне машине (VMS) који су изузетно једноставане за коришћење, али са апстрактним диском, меморијом и процесором, док Bare Metal облаци су у суштини физички сервери [2].

3.6. Заштитни слој *(Security Layer)*

Big Data пројекти су веома подложни безбедносним питањима због дистрибуиране архитектуре, због коришћења једноставаног модела за програмирање и open source алата. Међутим, безбедност мора да се спроводе на начин који не штети перформансе, скалабилност или функционалност, а требало би да буде релативно једноставнан за управљање и одржавање. Да би се спровела безбедност, потребно је осмислити такав Big Data систем, који ће минимум да [2]:

- Садрже аутентичне чворове користећи протоколе као што је Kerberos
- Омогућава file-layer шифровање
- Користи алатке као што су Chef or Puppet за валидацију приликом распоређивања сетова података
- Обезбеђује да сва комуникација између чворова буде безбедна - на пример користећи Secure Sockets Layer (SSL), TLS....

3.7. Систем за праћење *(Monitornig Layer)*

Системи за праћење морају да обезбеде контролу, односно праћење великих дистрибуираних кластера који су распоређени у обједињеном режиму. Машине морају да комуницирају са алатом за праћење преко високих нивоа протокола, као што су XML [2].

3.8. Визуелизација резултата *(Visualization Layer)*

Постоје два пута која воде до ваљане визуализације података који су претходно обрађени Hadoop-ом, односно неким алатом из Hadoop екосистема. Први је неко бесплатно решење, а други, наравно, оно које се плаћа. Постоје разлике између ова два решења, и једно и друго имају мане и предности. Сигурно бесплатно решење звучи примамљиво јер се не мора издвојити новац за њега, али друга страна медаље је да је обично теже одржавати и користити такав алат, као што је на пример D3.js. Предност је што је потпуно бесплатна и нуди готово све што је потребно за визуализацију. Други избор је онај који се плаћа. Алати који се истичу су Tableau, наравно Microsoft алати за визуализацију и BI. Мада, углавном зависи од система до система који ће се алат користити [2].

4. Модели података

Због различитих извора, а одатле и због различитих типова података, није могуће користити моделе база података, као што су на пример традиционални релациони модели. Неки подаци су структуирани и сачувани у релационим базама података док су неки други подаци, укључујуди документа, слике и видео записе, неструктуирани. Компаније такође имају задатак да размотре нове изворе података које генеришу машине као што су сензори. Други извори информација су они који генеришу људи као што су подаци из друштвених медија и click-stream подаци добијени са разних сајтова. Поред тога, доступност и прихватање нових, мобилних уређаја, уз сталан приступ глобалној мрежи доводе до нових извора података [9].

Big Data технологије окрећу се NoSQL базама података, које наравно не искључују традиционалне базе, већ их на неки начин надограђују због различитих, нових типова података. У савременом развоју софтвера термин нерелационе базе података се односи на системе за управљање колекцијама података које [9]:

- немају строгу статичку структуру података,
- немају исцрпну проверу услова интегритета,
- не користе упитни језик SQL.

Назив NoSQL настао је 1998. године управо зато што је SQL симбол релационих база података. Било је покушаја да се зову и NonRel базе, али је NoSQL преовладао. Данас се тумачи као “Not only Sql”. Без обзира на буквално значење, NoSQL се користи данас као општи термин за све базе података и складишта података које не прате популарне и добро утврђене принципе релационих база података и често се односе на велике скупове података којима је потребно брзо и ефикасно приступити и мењати их на Интернету. То значи да NoSQL није јединствени производ, чак ни јединствена технологија [9].

4.1. NoSql базе података

Недостатак релационих база података представља релативно висока цена читања због нередундантности и строге структуре података. Отежано је дистрибуирање пре свега због конзистентности података. Скупа је промена структуре због повезаности структуре са употребом и оптимизацијама [9].

Данашње Web апликације сусрећу се са многим проблемима. Конкурентни корисници у огромном броју који се тешко одређује, дневно производе податке реда терабајта и петабајта. Обрада података је отежана, јер корисници у сваком тренутку врше измену тих података. Оптерећење је неравномерно са непредвидивим порастом, неравномерно велика динамичност, непрекидно додавање нових могућности, промена постојећих компоненти [9].

Нема јединствене дефиниције, али се може рећи да је нерелациона база података је база података која не почива на релационом моделу података или га се бар не држи чврсто. Лако се дистрибуира, хоризонтално је скалабилна тј. лако подноси значајне промене шеме [9].

Пошто је споменута скалабилност, потребно је објаснити шта значи овај појам. Скалабилност је могућност апликације да поднесе повећање захтева и броја корисника, а да сама апликација не мора да се мења. Што је апликација скалабилнија она ће лакше поднети повећан проток података. Циљ којим теже сви пројектанти система јесте да се постигне линеарност у брзини одговора на захтев и количине података са којима се манипулише [9].

Када се говори о самом хардверу постоји хоризонтална скалабилност и вертикална скалабилност [9].

- Вертикална скалабилност је када је апликација смештена на једном серверу, а на повећан проток се реагује тако што се серверу додаје меморија, јачи процесор, нова језгра или додатни хард диск.

- Хоризонтална скалабилност је идеалније решење, посебно за велике системе. Додавањем нових чворова систем наставља да ради као до сада само са новим играчем (чвором) у тиму. Чвор представља један сервер.

Иако концепти NoSql база података могу да подсећају на релационе базе података (каталог-табела, вредност-ред,...) заправо постоје значајне разлике. Структура вредности обично није строго предефинисана, не инсистира се на нередундантности, практично ретко има природних кључева, не постоји референцијални интегритет.

Друге честе карактеристике су да нема статичку шему, лако се реплицира, поседује једноставан API (Application Programming Interface), омогућава огромне количине података, почива на скупу принципа BASE (Basically Available, Soft state, Eventual consistency), а не ACID (Atomicity, Consistency, Isolation, Durability) [16].

Овде је битно навести и CAP (Consistency, Availability, Partition tolerance) теорему која каже да је немогуће у дистрибуираним рачунарским системима да се истовремено омогуће све три гаранције [9]:

1) Конзистентност - да сви чворови виде исте податке у исто време

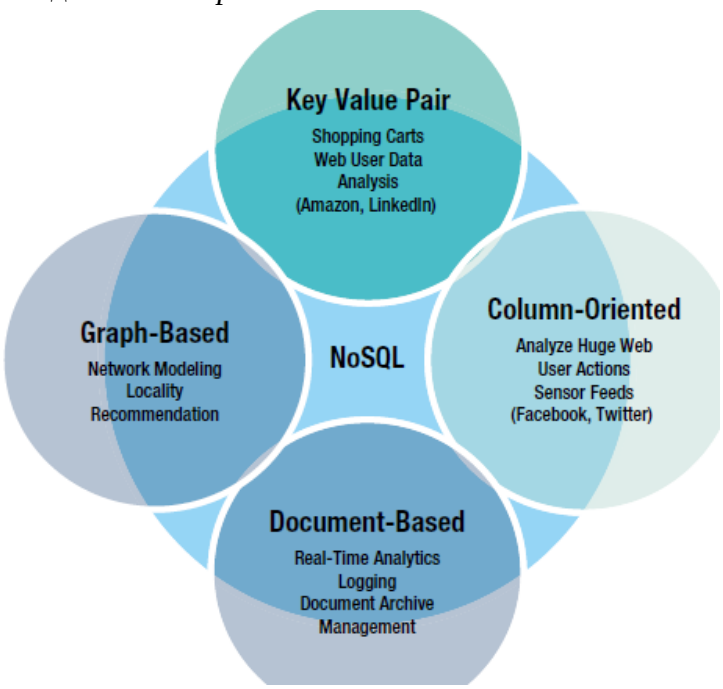
2) Расположивост - сваки захтев прима одговор да ли је успешно извршен или не

3) Прихватање раздвојености - систем наставља рад без обзира на поруку о неуспеху на неком делу система

4.2. Врсте модела података за Big data

Постоје 4 типа база података у односу на типове података, која ће појединачно бити објашњена, а то су [9]:

1. Кључ/вредност базе података – *Key/Value DS*
2. Уређено сортирана складишта оријентисана ка колонама - *Column-oriented DS*
3. Базе података оријентисане ка документима – *Document DS*
4. Графовске базе података - *Graph DS*



Слика 21. Типови NoSql података [1]

4.2.1. Уређено сортирана складишта оријентисана ка колонама

Google-ов Bigtable представља модел у коме се подаци чувају уређени према колонама, што представља контраст у односу на РСУБП (Системи за управљање релационим базама података), где се подаци чувају уређени у редовима. Овако се подаци смештају много ефикасније зато што ако нема података за дату колону они просто неће бити уписани, док код РСУБП се уписују NULL вредности [9].

Свака јединица може да се посматра као скуп парова кључ/вредност, где се свака јединица идентификује са примарним идентификатором, који се често назива редни кључ. Јединице података су распоређене у уређено-сортираном поретку према редном кључу [16].

Пример:

Имамо просту табелу која садржи поља: ime, prezime, zanimanje, post_kod, pol, и податке:

ime: Jovan
prezime: Jovanović
post_kod: 11000
pol: muški

ime: Petar
post_kod: 18000

Редни кључ првог податка је 1, а другог 2. Подаци се чувају сортирано по колонама, тако да ће податак са редним кључем 1 бити уписан пре податка са редним кључем 2.

Подаци у Bigtable и његовим клоновима су смештени у секвенцијалном редоследу. Када подаци напуне један чвор, они се деле у више чворова. Подаци су сортирани не само на једном чвору већ и на свим чворовима који чине један велики секвенцијални скуп. Подаци и даље постоје у редоследу толерантном на грешке где се одржавају 3 копије истих података. Већина клонова одржава дистрибуирани систем датотека за чување података на диску. Дистрибуирани систем датотека омогућава да се подаци чувају на кластерима [9].

Hbase је популарно, open-source, уређено према колонама, складиште података које је моделовано на основу идеја Google-овог Bigtable-а. Овај алат објашњен је у Глави 3 овог рада.

4.2.2. Кључ/вредност базе података

Хеш мапа или асоцијативни низ је најједноставнија структура података која може чувати скуп кључ/вредност података. Такве структуре података су популарне зато што је потребно O(1) времена за приступ подацима. Кључ из пара је јединствена вредност у скупу помоћу које се брзо приступа подацима [9].

Кључ/вредност парови су различитих типова: неки чувају податке у меморији, а неки пружају могућност перзистенције података на диску. Кључ/вредност парови се могу дистрибуирати у кластеру чворова [9].

Једна од познатијих база овог типа је Oracle-ова BarkeleyDB где су и кључ и вредност низови бајтова. Engine базе података не придаје значење кључевима или вредностима, већ узима парове из низова бајтова и враћа исте позиву. BarkeleyDB допушта да се подаци кеширају у меморији и пребаце на диск када нарасту. Такође, постоји и индексирање кључева за бржи преглед и приступ [9].

Други тип кључ/вредност парова који је у општој употреби је кеш. Кеш пружа пресек стања у меморији најкоришћенијих података у апликацији. Циљ кеша је да редукује улазно-излазне операције диска. Кеш системи могу бити рудиментарне структуре мапа или робустни системи са политикама истека података из кеша [9].

Ови типови кључ/вредност парова пружају строгу конзистенцију података. Међутим постоје и оне које истичу доступност испред конзистенције у дистрибуираном распоређивању. Најпознатији представник те класе је Amazonov Dynamo који обећава изузетну доступност и скалабилност и представља окосницу Amazonovog система за складиштење података. Најважнији концепт овог система је евентуална конзистентност. Евентуална конзистентност подразумева да могу постојати мали интервали неконзистентности између реплицираних чворова када се подаци ажурирају између peer-to-peer чворова [9].

Најпознатија open-source имплементација овог система је Appache Cassandra развијена од стране Facebook-a, а касније донирана Appache фондацији. Имплементирана је у Javi. Издата је под Apache License version 2, а од познатих компанија је користе Facebook, Digg, Reddit, Twitter [9].

4.2.3. Базе података оријентисане ка документима

Складишта оријентисана ка документима нису системи за управљање документима. Реч документ сугерише лабаво структуиране скупове кључ/вредност парова у документима, обично JSON (JavaScript Object Notation). JSON јесте текстуални стандард намењен размени података у форми читљивој и разумљивој и за човека и за рачунар. JSON формат је изведен из JavaScript језика, али сам формат није зависан од било ког конкретног програмског језика. MongoDB документе складишти у формату који се назива BSON, и користи га за размену података са апликацијама на мрежи. BSON формат се заснива на JSON-у и настао је од скраћенице "Binary JSON". База третира документ као целину и избегава поделу документа на саставне кључ/вредност парове. На нивоу колекције, ово омогућава складиштење различитих скупова докумената заједно у једну колекцију [9].

Најпознатија open-source имплементација овог типа је MongoDB10. Креирана је од стране 10Gen компаније, имплементирана у C++. Упитни језик је направљен у JavaScript-у, али подсећа помало на Sql. Издата је под GNU Affero GPL лиценцом, а од познатих компанија је користе FourSquare, SAP, MTV, SourceForge, Github [16].

Функционалности MongoDB:

Индексирање - индекси обезбеђују високе перформансе извршења операција код најчешће коришћених упита. MongoDB индекс представља структуру података која омогућава брзо лоцирање докумената на основу вредности одређених поља у документима [9].

Агрегација података - MongoDB пружа функцију agregate() која се користи за агрегацију података и која се позива над одређеном колекцијом. Користе се и следећи оператори: sort (сортирање докумената), skip (игнорише одређени број докумената на почетку колекције), match (филтрирање докумената на основу вредности поља), group (груписање докумената), limit (дефинише број докумената који пролазе кроз филтер), project (одређује поља из документа која ће бити укључена у резултат) и unwind (трансформише низ) [9].

Механизам репликације - овај механизам омогућује синхронизацију података између већег броја MongoDB инсатнци. Постоји само једна примарна инстанца, а све остале су секундарне. Репликација има задатак да обезбеди редундантност, повећа доступност и олакша неке административне задатке попут прављења резервних копија. MongoDB обезбеђује и механизам ако откаже примарна инстанца, механизмом гласања бира се нова примарна инстанца [9].

Механизам дељења - овај механизам омогућује партиционисање колекције докумената и њихову дистрибуцију на већи број инстанци. Идеја која стоји иза механизма дељења је да се повећа капацитет система [9].

Map/Reduce - омогућује паралелну обраду већих количина података [9].

Ad-hoc упити - Динамички (ad hoc) упити омогућују рад са MongoDB- ом на сличан начин као са РДБМС-ом. Самим тим, ова карактеристика је погодна при преласку са неког РДБМС система на MongoDB [9].

GridFS механизам - овај механизам омогућује памћење великих докумената. Конкретно, MongoDB може да ради са документом величине 16MB. Због овог ограничења користи се GridFS механизам који документ дели у низ комада. За чување великих докумената користе се две колекције: колекција „files" која садржи метаподатке о документу и колекција „chunks" која садржи делове документа. Величина делова на које се документ дели обично је 256KB [9].

4.2.4. Графовске базе података

Графовске базе података користе графовске структуре са чворовима (који представљају ентитете), гранама (које повезују чворове) и својствима (представљају атрибуте) за репрезентацију и чување података. По дефиницији графовска база је сваки систем код којег сваки елеменат садржи директни показивач на суседни елемент и није потребан никакав преглед индекса. Ове базе су моћан алат за упите попут на пример израчунавање најкраћег пута између два чвора у графу [9].

Једна од познатијих база овог типа је FlockDB11. Креирана је од стране Twiter-a и користи се за складиштење листе суседних пратилаца на Twiter-у. Имплементирана је у Scala програмском језику и издата под Apache License version 2 [9].

NoSql не представља решење за све проблеме и има својих недостатака. Суштина је да се NoSql решења понашају добро приликом скалирања, када подаци расту у великим количинама и потребно је да буду дистрибуирани на више чворова у кластеру. Процесирање великих количина података представља једнако велики изазов и захтева нове методе. Нерелационе базе података неће у блиској будућности избацити релационе. То уопште није и њихов циљ. Оне допуњују РСУБП и у зависности од потреба ће се одлучити који системи ће се користити, док је у великим компанијама, попут Facebook-a, пракса да се користе и једне и друге.

5. Анализа и методологија велике количине података

Услед интензивног развоја информатичке инфраструктуре скоро све фирме, а посебно оне веће, чувају велике количине података о пословању, сваком клијенту и кретањима у окружењу. Дневни унос информација које велике фирме похрањују у своје базе података, мери се терабајтима. У један терабајт стане довољно текста за око два милиона књига. Извори тих информација су различити (интерни, екстерни, аналитички), информације могу бити атрибутивне или нумеричке, могу се односити на факторе које утичу на пословање фирме, интерне процедуре, на кориснике услуга предузећа (потрошаче), пословање конкуренције, пословну околину [8].

Као што је до сада више пута споменуто, такви сирови подаци, неадекватно структурирани, различитих формата, немају претерано велику употребну вредност. Неопходно их је припремити, анализирати и на основу тога доћи до информација (знања) која могу предузећу обезбедити остварење пословног успеха.

Поред споменуте архитектуре и њених елемената у претходном одељку, веома битну улогу има анализа података, као и методе које се примењују. Обзиром на чињеницу да се ради о великим количинама података, просто је немогуће да човек сам врши анализе. Анализе се препуштају за то посебно развијеним програмима (од којих су неки споменути у одељку 4 овог рада у оквиру Hadoop архитектуре). Нова врста технологије чији циљ је управо решавање проблема са којим су се фирме суочиле јесте Business Intelligence. Business Inteligence (BI) обухвата широки скуп апликација и технологија за прикупљање података, лак приступ подацима и експертску анализу података, а у циљу обезбеђивања адекватне подршке процесу одлучивања. BI представља фамилију производа у коју спадају : OLAP (Online Analytical Processing) производи, Data Mining производи и производи за креирање извештаја [8].

Data mining је најважнији производ из фамилије Business Intelligence производа , чија је сврха проналажење скривених образаца у подацима, повећавање њихове употребљивости и трансформација тих података у корисно знање [8].

Постоји неколико дефиниција Data Mining-a. Data Mining (DM) се може дефинисати као процес проналажења скривених законитости и веза међу подацима. То је техника претраживања података у циљу идентификације тражених узорака и њихових међусобних релација. Једноставно речено, DM је поступак издвајања интересантних, нових и потенцијално корисних информација или узорака, садржаних у великим базама података, а све у циљу доношења исправних пословних одлука [8].

DM је мултидисиплинарно подручје које обухвата: базе података, експертне системе, теорију информација, статистику, математику, логику и читав низ других подручја [8].

Са развојем рачунарске технологије, предузећа су могла чувати све веће количине података у својим базама, тако да је омогућена комерцијална употреба великог броја DM техника у сврхе пословног одлучивања [8].

RAZDOBLJE	EVOLUCIJSKI KORACI	POSLOVNI UPITI	TEHNOLOGIJA	KARAKTERISTIKE
1960-te	Prikupljanje podataka	Koliki je ukupni prihod kompanije u posljednjih pet godina?	Kompjutori, trake, diskovi	Statična isporuka povijesnih podataka
1980-te	Pristup podacima	Kolika je bila prodaja po određenim prodajnim jedinicama na području Dalmacije u proteklom mjesecu?	Relacijske baze podataka, SQL, ODBC	Dinamična isporuka povijesnih podataka jedne razine
1990-te	Skladištenje podataka i sustavi za potporu odlučivanju	Kolika je bila prodaja u određenim prodajnim jedinicama na području Dalmacije u protekom mjesecu. Istraži (drill down) lokalitet Split	OLAP, multidimenzijske baze podataka, skladište podataka	Dinamična isporuka povijesnih podataka s više razina
Danas	Rudarenje podataka	Što se može dogoditi s prodajom na lokalitetu Split u sljedećem mjesecu? Zašto?	Napredni algoritmi, multiprocesorski kompjutori, masivne baze podataka	Predvidiva i proaktivna isporuka informacija

Слика 22. Приказ четири револуционарна корака која су пружила могућност брзих и прецизних одговора какве данас захтева савремено пословање [8]

DM подразумева коришћење софистицираних алата за анализу, а они могу укључивати статистичке моделе, математичке алгоритме, методу машинског учења, базе података и сл. Процес еволуције од података до корисних информација и нових сазнања ишао је корак по корак, што је приказано на Слици 23 [8].

5.1. Поређење Data Mining-а са традиционалним статистичким моделом

Као софистицирани систем за подршку одлучивању, DM користи најсавременије статистичке и математичке моделе за анализу података о пословању предузећа и његовим потрошачима, како би се открили потенцијални проблеми и шансе. Помоћу DM менаџери долазе до корисних информација и знања, неопходних за ефикасно управљање. Data mining се разликује од класичних статистичких метода по томе што се не одвија по унапред утврђеним правилима, већ показује креативност у анализи података и на тај начин може да открије нова, неочекивана правила. Data Mining је процес откривања нових знања и информација из прикупљених података. Може се помислити да је иста ситуација и са употребом класичних статистичких метода за анализирање података. Међутим, на тај начин се стварају одређене претпоставке, тврдње (хипотезе), и траже се подаци који ће то потврдити или оспорити. Са друге стране, употреба Data Mining-а подразумева анализу података, а да пре те анализе истраживач није дефинисао одређене тврдње или претпоставке везано за појаву која се анализира. Једноставно се поставља питање на које се тражи одговор, а препушта се Data Mining алгоритмима да дефинишу обрасце, и пруже одговоре [8].

DM при томе подразумева анализу података које потичу из различитих извора, из различитих организационих јединица предузећа (продаје, производње, финансија) и различитих информационих система (платформи). DM не само да омогућава извлачење, трансформацију и учитавање разлицитих информација у јединствену базу података, већ омогућава предузећима и менаџерима да анализирају нпр. понашање потрошача на бази 100 и више обележја, док традиционални статистички модели омогућавају истовремено посматрање 3 или 4 оваква обележја [8].

Док традиционалне статистичке анализе почивају на тестирању хипотеза, DM се ослања на софтверско моделирање, којим се утврђују везе и међузависности великог броја појава, и обезбеђује знање за решавање проблема, унапређивање пословања и предвиђање [8].

Неке технике и модели DM су уско повезани са онима у статистици, као нпр. модели линеарне регресије и временских серија, али се углавном употребљавају доста сложенији и флексибилнији програми и на њима засновани модели. Две технике DM, неуронске мреже и стабло одлучивања, могу анализирати истовремено и до 200 независних промењивих. Са друге стране, ово није могуће са моделима нпр. вишеструке регресије [8].

Користећи традиционалне статистичке методе, аналитичар себи може да постави питање : „Јесу ли потрошачи са већим приходима лојалнији неком супермаркету од оних који имају ниже приходе ? “ - нулта хипотеза ће бити одбачена или неће. DM са друге стране, може омогућити знатно више података и бољи увид у факторе који утичу на лојалност, од оних које сазнајемо тестирањем хипотеза. Анализом података путем

DM, могу се добити груписани подаци о потрошачима према томе: да ли имају Клубску кредитну картицу, живе даље од 10 миља од маркета, имају 2 аутомобила... и њиховој лојалности према групама [8].

Сви DM модели се углавном састоје из Независних променљивих (predictors) и Зависних променљивих (responses). Тако нпр, компаније за осигурање аутомобила могу скупљати податке о потрошачима, о величини њихове породице, кредитном рејтингу. Ове информације (независне променљиве) могу се употребити да се предвиде губици по појединим групама потрошача, или да се одреди који потрошачи ће највероватније купити нови прозивод фирме (зависне променљиве) [8].

Два основна услова за избор софтвера су величина базе података, и комплексност питања на које се тражи одговор. Јасно је да за веће количине података које се анализирају и сложенија питања за које се траже одговори, морају се користити и моћнији програми [8].

DM се може применити у свим оним областима где се располаже великом количином података чијом анализом се желе открити одређена правила, законитости и везе. Стога треба поменути и концепт Data Werhousing-a, који користе све велике светске компаније у циљу интеграције података у једну базу, на основу које крајњи корисници могу спроводити ad-hock анализе, правити извештаје, предвиђати и доносити одлуке. Концепт Data Werhousing-a (Складиштења података) има за циљ прикупљање и дистрибуцију информација кроз предузеће, уз омогућавање мултидимензионалног приступа подацима какав је данас неопходан за доношење пословних одлука [8].

5.2. Фазе у процесу Data Mining-а

Животни циклус једног Data Mining пројекта се састоји из следећих осам корака [8] :

1. **Сакупљање података** – је обично први корак у Data Mining пројекту. Пословни подаци су ускладиштени у бројним системима, интернету, базама података компанија, и први корак обично представља пренос релевантних података у базу података где се подаци анализирају. Понекад постоји и складиште података што олакшава даљи рад али у великом броју случајева подаци који су сакупљени могу бити недовољно корисни за анализу те се због тога неопходни подаци морају сакупити из других извора. Након што се сакупе, подаци се могу семпловати да би се смањила величина тренинг скупа података. У многим случајевима, обрасци који су пронађени на скупу од 50 000 купаца су исти као и они пронађени на тренинг скупу од 1 000 000 купаца.

2. **Филтрирање података и трансформација** – је најинтензивнији корак у Data Mining пројекту кад су ресурси у питању. Циљ филтрирања података је одстрањивање ирелевантних и сувишних информација из скупа података. То подразумева уклањање дуплих и непотпуних података, њихову транформацију и јединствен систем података, изабирање подгрупа података, одређивање броја промењивих са којима је могуће радити. Циљ трансформације података је промена изворног податка у другачији формат типа података. Постоје различити технике које се могу применити за корак филтрирања и трансформацију података, а најчешће коришћене су: трансформација типова података, непрекидна трансформација колона, груписање, рад са вредношћу која недостаје, брисање абнормалних случајева итд.

3. **Креирање и избор модела** – је трећи корак који се примењује након филтрирања и трансформације података. Тек када се подаци филтрирају и када се променљиве трансформишу у погодне типове података, може се започети са креирањем модела. Пре креирања модела треба да разумети циљ Data Mining пројекта и врсту Data Mining задатка који ће се користити. За сваки Data Mining проблем постоји неколико одговарајућих алгоритама. Прецизност алгоритма зависи од природе података као што су: број стања атрибута који се користе за предвиђање, пренос вредности сваког атрибута, веза између атрибута итд. У овом почетном делу пројекта потребно је саставити тим пословних аналитичара који су експерти у одређеној области.

4. **Процена квалитета модела** – У делу креирања модела креира се скуп модела користећи алгоритме и технике DM-а , али након креирања мора се извршити и евалуацију тог модела. Постоји неколико популарних алата за евалуацију квалитета модела. Најпознатији је лифт дијаграм. Он користи већ истрениран модел како би предвидео вредности које ће се добити из скупа података који се тестира. На основу вредности које се добију и вероватноће он графички приказује модел на дијаграму.

5. **Креирање извештаја** – Након креирања модела и евалуације квалитета тог модела врши се креирање извештаја који се достављају менаџерима на увид. Већина Data Mining алата има особину креирања извештаја који омогућује корисницима да генеришу претходно дефинисан извештај са текстуалним и графичким детаљима Data Mining модела. Постоје два основна типа извештаја: извештаји о пронађеним обрасцима и извештаји о предвиђеним вредностима модела.

6. **Оцењивање модела** – У многим Data Mining пројектима, проналажење образаца и модела је само пола посла; коначни циљ је употреба тог модела за предвиђање. Предвиђање се још назива и scoring у Data Mining терминологији. Да би се добиле предвиђене вредности мора постојати већ истренирани модел и скуп нових података.

7. **Интеграција Data Mining модела у апликацију** – Интегрисање Data Mining модела у пословне апликације представља поновну примену пословне интелигенције на пословни систем тј. затварање петље за анализу. Све више пословних апликација укључује и Data Mining компоненту а предности Data Mining-а су велике. На пример CRM (Customer Relationship Management) апликације могу имати Data Mining особине које групише купце у сегменте, ERP (Enterprise Resource Planning) апликације могу имати Data Mining особине које им користе да предвиде обим производње. Online књижара може дати потенцијалним купцима препоруке књига. Интегрисање Data Mining особина, поготово компоненте за предвиђање у апликације један је од битнијих корака Data Mining пројекта. Ово је кључни корак за увођење дата мининг-а у масовну употребу.

8. **Управљање моделом** – Одржавање статуса Data Mining модела представља прави изазов. Сваки Data Mining модел има свој животни циклус. У неким областима примене обрасци су релативно стабилни и модели не захтевају учестало поновно тренирање модела. Али у многим областима обрасци се мењају често. Трајање једног Data Mining модела је ограничено. Нова верзија модела се мора правити често. Одређивање прецизности модела и креирање нових верзија овог модела би требало бити постигнуто коришћењем аутоматизованих процеса.

5.3. Методе Data mining-a

Аналитичке технике које се користе у DM, у великом броју случајева су одавно познате математичке технике и алгоритми које су коришћене годинама пре тога. Технике које се најчешће примењују углавном су изведене из три главне области: статистике, машинског учења и база података [8].

Одређени алгоритми попут регресије и стабла одлучивања преузети су из статистике. С обзиром да се Data Mining базира на откривању образаца понашања из анализираних података неки алгоритми су преузети из области машинског учења попут неуронских мрежа које се изузетно успешно примењују код класификације и регресије и онда када су везе међу атрибутима нелинеарног типа. Генетски алгоритми престављају још једну технику која се користи за класификацију и кластеровање. Али је развијено и много нових алгоритама, метода и софтвера. Такође постоји и неколико скалабилних верзија алгоритама класификације и кластеровања који користе технике база података, укључујући и Microsoft-ов алгоритам кластеровања[8].

Уопштено говорећи, све Data Mining технике се могу поделити у две групе[8].:

1) Discovery data mining – технике за откривање нових знања (информација)
2) Predictive data mining – технике за предвиђања

Како би се проблеми решавали што брже и тачније, развијен је велики број техника, алгоритама и метода DM-a, у неколико последњих година. Све су оне сврстане под истим називом – Data Mining технике [8].

Неке од техника DM су [8]. :

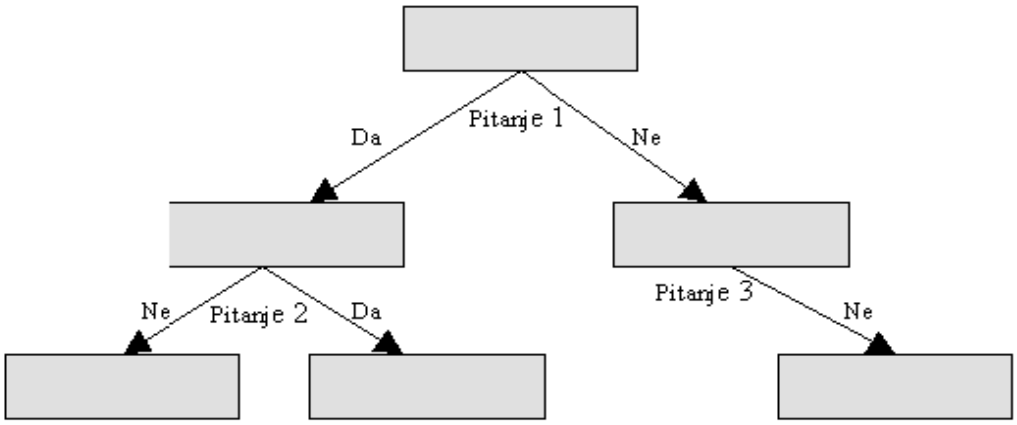
1. Стабло одлучивања (Decision Tree)

Decision Tree је веома популаран метод за класификацију и одлучивање. То је техника одлучивања која се темељи на односима између стратегија и стања, а користи се за решавање проблема у финансијама, банкарству, маркетингу, осигурању. Коришћењем серије питања и правила за категоризацију података, предвиђају се исходи.

Стабло одлучивања настаје гранањем као последица испуњења услова класификацијских питања. Свако питање ће поделити податке у подскупове који су хомогенији него виши скуп. Ако питање има два одговора, тада ће као одговор на питање настати два подскупа (бинарно стабло). Колико питање има одговора толико ће подскупова настати. Самим тим врши се класификација појединих података. Предвиђање понашања појединог клијента може се извести на темељу његовог припадања поједином скупу (у који је сврстан на основу низа питања и услова), за који се зна како ће се понашати.

Приликом изградње стабла одлучивања важно је знати поставити право питање. Питање је утолико боље, уколико ће се њиме боље организовати подаци, односно уколико ће се након тога створити подскупови који су хомогенији. Модели који се базирају на стаблу одлучивања разликују се по алгоритмима који захтевају обележја појединих података на бази којих се креирају питања. Стабла одлучивања се веома примењују на релацијским базама података (нпр. SQL).

Пример стварања стабла одлучивања приказан је на Слици 23 .



Слика 23. Пример изградње бинарног стабла одлука [8]

2. Метода најближег суседа (Nearest neighbor classification)

Nearest neighbour classification једна од најстаријих техника која се примењује у Data Mining-у за класификацију података. Због свог начина рада, који је сличан људском начину размишљања, ова метода је једна од најједноставнијих. Темељи се на тражењу података који имају најсличинија својства и познато понашање. Податак који има најсличнија својства је најближи сусед, па се претпоставља да ће се слично и понашати. Питање алгоритма је како одредити ко је најближи сусед. Један од најједноставнијх начина је употреба еуклидске геометрије у н-димензионалном простору.

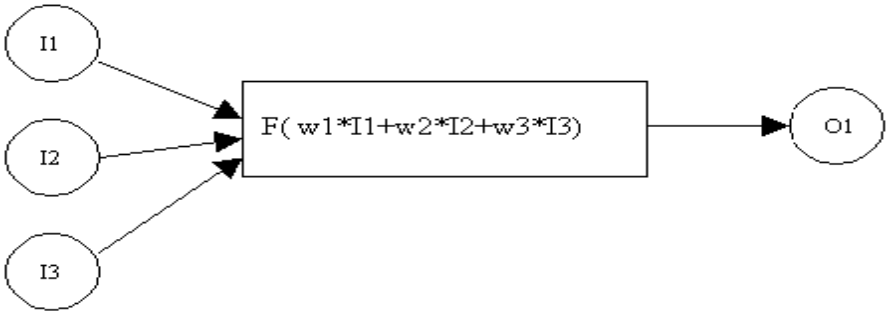
Како би метода била што тачнија, потребно је у бази података наћи што сличнији податак (за који је потребно што тачније познавати понашање), што захтева велике количине података. За разлику од осталих техника, овде не постоји процес учења како би се креирао модел. Подаци који се користе за учење су у ствари модел. Када се појави нови податак, алгоритам анализира све податке у бази, како би нашао подгрупу случајева која најбоље одговара том случају, и на основу тога врши предвиђање.

3. Неуронске мреже (Neural networks)

То је техника DM замишљена да делује слично људском мозгу. Као што људски мозак након процеса учења извлачи одређене претпоставке на основу ранијих запажања, тако и ове мреже предвиђају промене и дешавања у систему након процеса учења. DM на основу ове технике почиње „учењем“ мреже помоћу података који су већ познати, а који се односе на вредност коју желимо прогнозирати. Након тога знање се проверава, све дотле док резултати провере не буду задовољавајући. Цео процес се у основи своди на следеће : Прво се неуронској мрежи дају одређени подаци за које се већ знају излазне вредности. На основу ових података неуронске мреже препознају обрасце и правила. Затим се на ознову ових образаца и функција истражују гомиле података које предузећа имају у својим базама.

Пример: Компаније које се баве издавањем платних картица располажу огромним подацима о својим корисницима, процесу одобравања и трансакцијама. DM омогућава утврђивање веза и правила међу подацима. Ако компанија нпр. зна да од 3000 захтева за картице постоји 100 покушаја преваре, коришћењем неуронских мрежа, утврђују се обрасци на основу којих се препознају ови покушаји. Ови обрасци се након тога користе за испитивање свих база података компаније, утврђивање и препознавање превара. Такође, проверавају се и саме трансакције при плаћањима. На основу утврђених шема понашања потрошача (шта купује, где купује, колико троши), систем одређује вероватноћу сваке трансакције и шаље контролорима поруку у колико треба неку од трансакција проверити.

Неуронске мреже су најкомпликованија метода (како за употребу, тако и за примену), али дају најтачније моделе. Нуронске мреже настале су проучавањем и покушајима имитирања рада мозга и нервног система човека (и других животиња). Основна ћелија неуронских мрежа (неурон) приказана је на Слици 24.



Слика 24. Блок шема неурона [8]

Неурон свој излаз темељи на комбинацији низа улаза помножених с одговарајућим тежинама. Неуронска мрежа састоји се од низа неурона који су међусобно повезани. Приликом пројектовања неуронске мреже потребно је одредити структуру (број неурона и њихове међусобне везе). Да би створили модел предвиђања употребом неуронских мрежа потребно је дефинисати тежине појединих веза. То се постиже тренингом неуронске мреже. Дају јој се тестни подаци и затим се коригује одговор који даје, ако је нетачан. Неуронска мрежа ће тада кориговати тежине појединих веза између неурона. Ако је претходни неурон дао тачан одговор вези према њему, тежина ће се повећати, док ће се у супротном смањити. С временом неуронска мрежа учи, па са повећањем броја тренинга даје све тачније резултате.

4. Fuzzy логика (Fuzzy logic)

Fuzzy логика је веома сложен појам, односно веома тешко је одредити тачну дефиницију. Да би се најлакше објаснио овај појам, упоредићу га са конвенцијалном логиком. Основа класичне логике, коју је дефинисао Аристотел, заснива се на јасним и прецизно утврђеним правилима, а почива на теорији скупова. Скупови имају јасно дефинисане границе. Неки елемент може да припада неком скупу или да не припада. И овакви скупови се дефинису као Crisp тј .јасни, бистри.

Код Fuzzy логике, није јасно дефинисана припадност елемента неком скупу, већ се мери у процентима. Скалирани, ови проценти узимају вредност од 0 до 1. Као пример могу се узети дани у недељи, и од тога направити два подскупа – радне дане и викенд.

По Crisp логици – понедељак, уторак, среда, четвртак и петак – припадају радним данима, и њих би обележавали бројем 1. У викенд дане спадају субота и недеља, и обележавамо их са 0. По Fuzzy логици, ситуација би била другација. Петак, је једним делом радни дан, а другим делом почетак викенда, тако да он припада једним делом (нпр 0,75) радном дану, а другим делом (0,25) викенд данима. Слична је ситуација и за недељу, јер се недељом увече људи припремају за радну недељу. Тако да се истинитост сваке тврдње у Fuzzy логици мери у процентима.

Ова логика је јако блиска људској перцепцији о многим стварима у животу. Сама техника се спроводи као симулација људског резоновања и размишљања, при чему се дозвољава рачунару да се понаша мање прецизно.

5. Меморисјки засновано расуђивање (Memory based reasoning)

Меморијски засновано расуђивање је техника DM која се користи за предвиђање и класификацију. Слична је техници неуронских мрежа, са разликом што MBR тражи сличне податке, али при том не утврђује обрасце и правилности у подацима. Пример употребе – Уколико доктор има пацијента са сличним симптомима болести као и код ранијих пацијената, он ће на основу искуства дати дијагнозу.

6. Кластеринг (Clustering)

Техника груписања која омогућава груписања података који су слични. Груписања су у ствари разврставања елемената у скупове, у којима се постиже највећа сличност података (сегментација купаца – по старости, занимању, дохотку, потрошњи).

При подели морају бити задовољена 2 критеријума :

1) свака група представља хомоген скуп – слични подаци,

2) сваки скуп се мора разиковати од осталих скупова – значајне разлике у подацима.

7. Анализа потрошачке корпе (Market Basket Analysis)

Market Basket Analysis често се назива и груписање по сличности. Користи се за проналажење група артикала који се најчешће заједно купују у једној трансакцији. Анализом потрошачке корпе, утврђује се вероватноћа да ће потрошач купити производ Б, уколико је при једној куповини већ купио производ А. Модел се широко употребљава у тржишним центрима и супермаркетима.

Као пример наводи се DM које је спровео Wall Mart. Анализирање продаје између 17-19 h поподне, утврђено је да су два прозивода која су најчешће заједно куповано Пиво и Пелене. На бази овог податка, менаџери си захтевали измештање витрине са Пивом ближе полицама са Пеленама. Као резултат продаја је повећана за 15%.

Ова метода се такође користи за анализе продаје у маркетима на различитим локацијама, продајама по различитим данима, годишњим добима. а све у циљу прилагођавања арсортимана и услуга како би се увећала продаја.

8. Rule indication

Употреба ове методе заснива се на проласку кроз базу података употребљавајући логичке функције на варијаблама, и рачунајући вероватноћу појаве таквог догађаја, појединих записа, како би се дошло до скривених информација. Како би се могло доћи до скривених информација, потребно је проћи кроз што више могућих међусобних комбинација варијабли (све комбинације), што драстично успорава и поскупљује ову методу. Ако се одбацују поједине варијабле као неважне, тада постоји могућност да не уочи веза између појединих података и модел учинити мање тачним. Осим с техничке стране, претраживање сличности појединих података по свим варијаблама често даје огроман број повезаности између појединих података, па је понекад потребан још један пролаз кроз добијени резултат како би се изоловали они закључци који су најинтересантнији.

Модели који се базирају на употреби rule inidication показали су се међу тачнијима (тачније дају неуронске мреже), али су за разлику од неуронских мрежа једноставнији за коришћење.

9. Метода К Најближег Суседа (K Nearest neighbors)

Побољшање у односу на методу најближег суседа је у томе што се посматра понашање неколико сличних података, а не само један. Самим тим (статистички) може се тачније предвидети понашање и својства појединог податка. Овакав алгоритам је врло лако имплементирати.

10. Остали алгоритми

Постоји низ других алгоритама на којима се темеље модели за Data Mining, али они се мање користе од горе наведених. Неки од њих су:

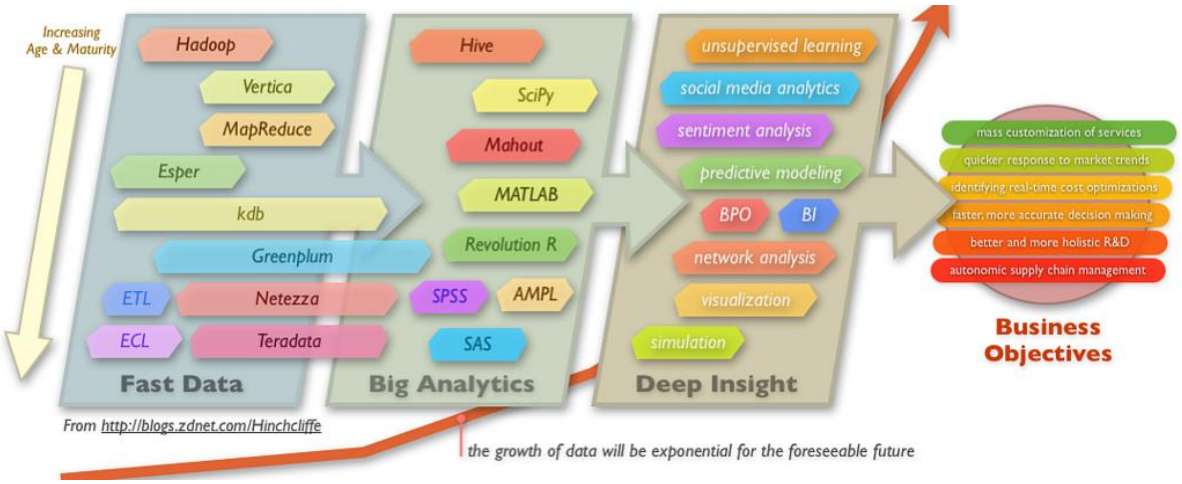
- K-means clustering
- Генетски алгоритми
- Самоорганизујуће мапе (енгл. Self organized maps)

5.4. Big Data анализа

Пре свега треба поменути да су предузећима данас доступни јефтинији DM програми и софтвери. Неки од најпознатијих су IBM-ов Intellgent Miner, Oracle-ов Darwin, SAS-ов Institute Enterprise Miner i SPSS-ов Clementine. Ова нова генерација DM програма не захтева ангажовање експерата, и не захтева од менаџера детаљно познавање статистике. Њихово коришћење и примена је данас доста поједностављено [8].

Драстично су смањени и трошкови припреме и трансформације података. Како свака база података има своје формате записа, а поједина решења алгоритама која се користе за DM користе своје формате записа, често је преношење података из базе података у алгоритам за Data Mining процедура која је раније одузимала доста времена. Током 90-тих година, чак 80% труда односило се на припрему података [8].

Иако аналитика података као егзактна наука често укључује и осећај то јест искуство аналитичара у одређеним условима може се говорити и о уметности анализе података. Зато је сасвим извесно да ће се убудуће јављати све напредније аналитичке методе које ће све прецизније упућивати на пословне смерове којима би се требало кретати. Програми и софтвери који данас постоје за анализу података су описана у глави 4, а заједно чине целокупан систем за прикупљање, обраду и анализу података, тако да са таквим информацијама могу унапредити пословни процеси, и утицати на побољшање у пословању. У наредној глави, биће објашњени примери како све то практично функционише.



Слика 25. Технике за анализу Big Data [17]

На Слици 25 се могу видети технике, односно алати актуелни за обраду и анализу великих података, као што су MATLAB, SAS и R, а неки од релативно нових техника карактеростичних за Big Data су Apache Hive и Mahout [17], мада је о њима већ било речи у глави 4 овог рада.

6. Big Data кроз анализу конкретних примера

Како би се у потпуности разумео појам Big Data, као и његова сврха, потребно је приказати неки реалан пример из праксе. Анализа наредних примера приказује која је корисна страна Big Data, и како је треба на прави начин искористити на пољу пословања.

Пример 1 говори о томе како мишљење корисника друштвених медија (Twitter-а) може утицати на маркетинг, у овом случају филмске продукције, док Пример 2 представља анализу података учитаних са сензора, у овом случају сензора која мере температуру са система за климатизацију и вентилацију, и како анализе таквих података могу утицати на побољшање температура у стамбеним зградама, као и на маркетинг фирми које производе системе за климатизацију и вентилацију.

Треба напоменути да су примери преузети са Hortonworks-овог сајта, тако да се систем базира на њиховој платформи. Тако да је потребно инсталирати и покренути Hortonworks Sandbox као и Hortonworks ODBC driver.

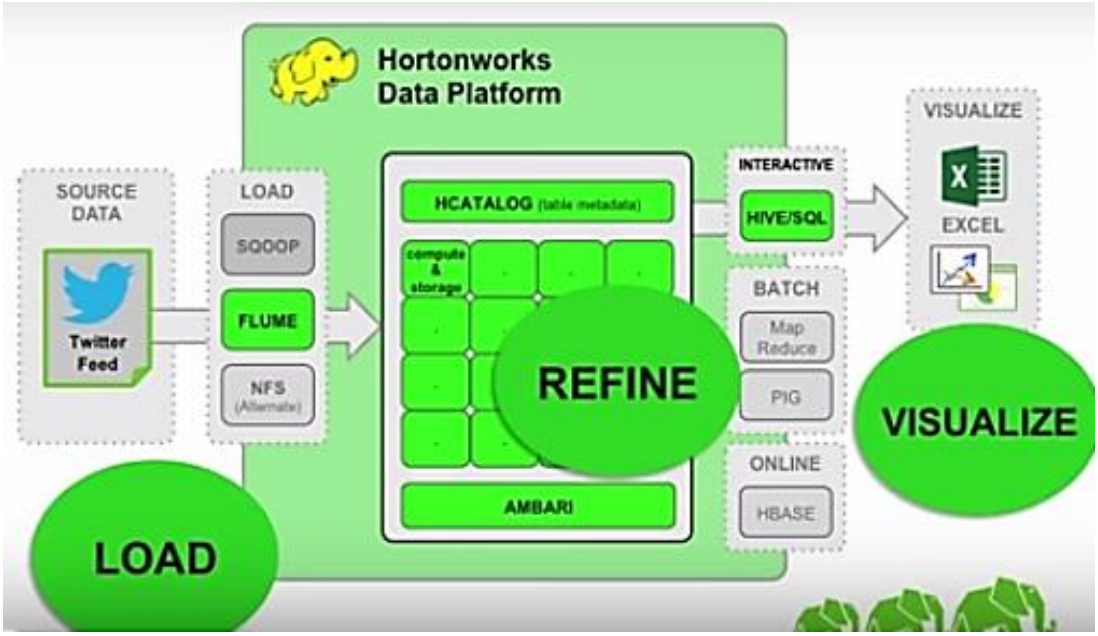
Пример 1: Анализа података прикупљених са друштвене мреже Twitter

У Примеру 1 је приказана анализа података са друштвене мреже Twitter. Анализирани су подаци корисника који су дали своје мишљење о новом филму Iron Man 3. Уз помоћ до сада приказаних Hadoop алата у овом раду, биће објашњено како су убачени подаци у HDFS, како су анализирани и визуелно представљени, а најбитније чему ти резултати служе.

Суштина овог примера је да се прикаже до сад све теоријски објашњено у овом раду, како се анализом велике количине података, у овом случају података са друштвених мрежа, може унапредити пословање, односно како се може искористити анализа мишљења корисника ради бољег пословања.

На Слици 26 је приказано из чега се састоји систем који служи за прикупљање, обраду и визуелизацију података, чији је извор друштвена мрежа Twitter. Као што се може видети систем се састоји од три дела а то су LOAD (учитавање података), REFINE (обрада података), VISUALIZE (визуелизација резултата).

У оквиру дела који служи за учитавање података приказан је алат Flume, уз помоћ којег су подаци учитани у HDFS. HCatalog и Hive се налазе у делу обраде и постављања упита, чија је то и улога, док се за визуелни приказ резултата користио Microsoft Excel.



Слика 26.Систем за обраду Twitter података

Може се рећи да су поступци по којима је обрађен овај пример следећи:

1. Убациавње података са Twitter-а у HDFS помоћу Flume
2. Коришћење Hcatalog како би се добио релациони приказ података
3. Коришћење Hive за упите и за прераду података
4. Увоз података у Microsoft Excel помоћу ODBC конектора
5. Визуелизација резултата помоћу Powerview

Sentiment Data

Подаци који су добијени са извора као што је Twitter називају се Sentiment Data. Sentiment Data су неструктурирани подаци који представљају мишљење, емоције и ставове који су садржани у изворима као што су поруке на друштвеним медијама, блоговима, online продавницама, поруке настале као резултат интеракције купаца. Организације користе анализу Sentiment података како би разумели како се људи осећају у одређеном моменту, такође ови подаци служе како би се пратило како се та мишљења мењају током времена.

Предузећа анализирају Sentiment Data како би побољшали:

- Производ – На пример анализирају које производе купци купују често или на пример предвидети који ће производ купити у будућности.

- Услугу – На пример хотели или ресторани могу увидети да ли је њихова услуга добра или лоша у односу на мишљење корисника.

- Конкуренцију – Анализом мишљена могу се уочити у којој области људи виде компанију као бољу (или лошију) од одређене конкуренције компаније.

- Репутацију - Шта јавност заиста мисли о предузећу. Да ли је репутација предузећа позитивна или негативна.

У овом примеру фокусираћу се на лансирање производа на тржиште, односно како је јавност одреаговала на ланисрање производа. У овом случају производ је филм (Iron Man 3). Како се публика (јавност) осећа и како се том анализом може побољшати промовисање филма биће речи у тексту који следи.

Кораци по којима је реализован овај пример су следећи:

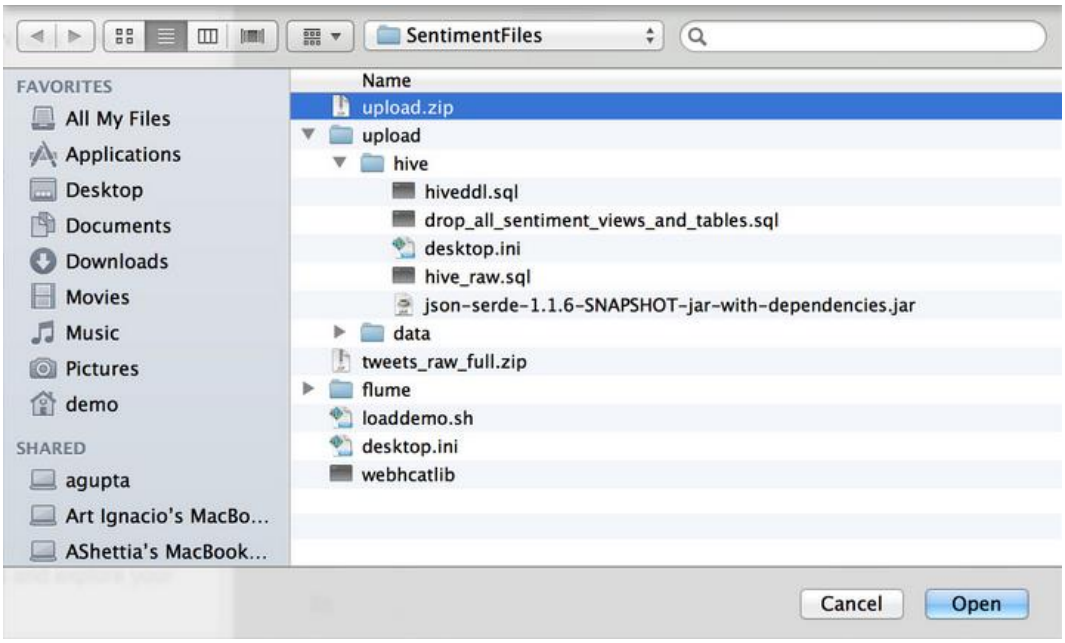
1. Преузеимање фајлова који садрже Sentiment Data
2. Убацивање Twitter података у Hortonworks Sandbox
3. Копирање Hive скрипте на Sandbox
4. Покретање Hive скрипте ради регулисања необрађених података
5. Убацивање обрађених Sentiment података у Excel
6. Визуелно представљање резултат уз помоћ Excel Power View.

Корак 1.

Twitter подаци су добијени коришћењем Hortonworks-овог Flume. Flume се може користити као лог агрегатор, који прикупља податаке из многих различитих извора и премешта у централизовано складиште података. У овом случају, Flume је коришћен за снимање stream података са Twitter-а, који се сада може учитати у Hadoop дистрибуирани систем фајл (HFDS).

Корак 2.

Након Корака 1 следи учитавање Twitter података у Sandbox. На слици 27 се може видети како се учитавају Sentiment подаци преко File Browser. Датотека са подацима је постављена у Sandbox виртуелну машину и аутоматски распакована у директоријум.



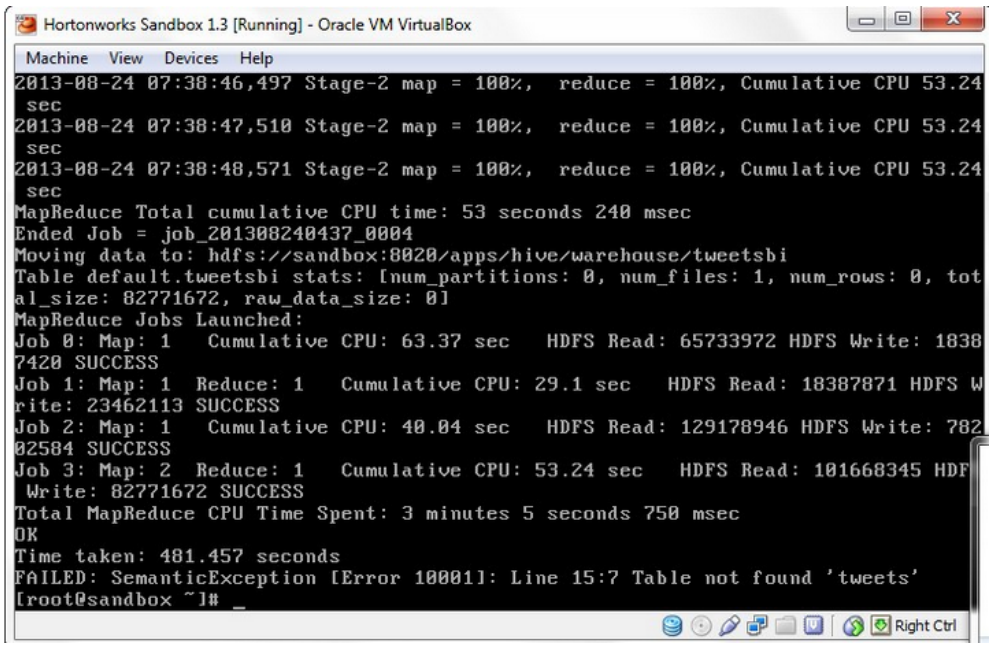
Слика 27. Учитавање Twitter података у Sandbox

Корак 3.

У овом кораку се користи SCP (Secure Copy Protocol) за копирање hiveddl.sql фајла у Sandbox. Процедура која је описана у овом раду односи се на систем Windows 7. На Windows-у је потребно инсталирати WinSCP апликацију. Након инсталације потребно је отворити апликацију и унети следеће податке: Host name: 127.0.0.1, Port: 2222, User name: root. Након конфигурације треба прекопирати Sentiment Data фајл у root фолдер на Sandbox.

Корак 4

Линије текста које се појављују након притиска на `hive -f hiveddl.sql` представља скрипту покренутих MapReduce послова, што је приказано на Слици 28.

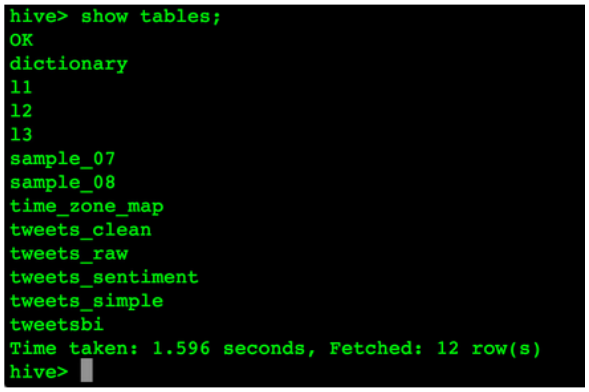


Слика 28. Скрипта која обавља MapReduce послове

Hiveddl.sql скрипта обавља следеће кораке како би се прецизирали подаци:

- Претвара сирове (необрађене) Twitter податке у табеларну форму.
- Пореди негативне и позитивне коментаре за сваког корисника Twitter-а, а затим добија позитивне, негативне или неутралне Sentiment вредности за сваки „Tweet“.
- Креира нову табелу за сваки „Tweet“ са Sentiment вредностима.

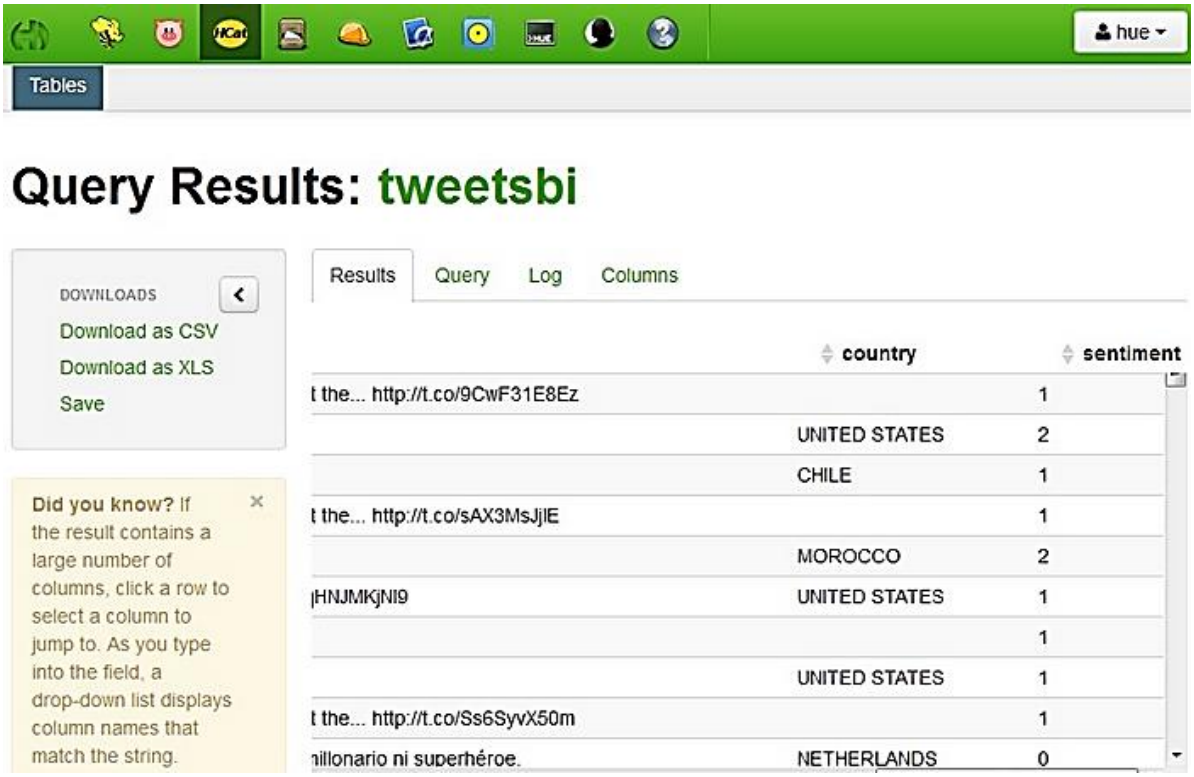
Након тога користи се Hive командна линија како би се видели тражени подаци.



Слика 29. Команде писане у Hive-у

На Слици 29 се може видети команда „show tables” која приказује табелу са такозваним твитовима. Подаци се могу прегледати и користећи команду „select * from limit 10;” што значи да граница 10 даје 10 евиденција уместо целе табеле.

Затим се користи HCatalog како би се брзо прегледали подаци. Када се отвори Sandbox HUE кориснички интерфејс кликне се на Hcatalog, након чега се селектују „tweetsbi” табеле, а затим се кликне на Browse Data. „Tweetsbi” табеле су табеле које су креиране од стране Hive скрипта која садржи колону са сваком Sentiment вредношћу, што је приказано на Слици 30.

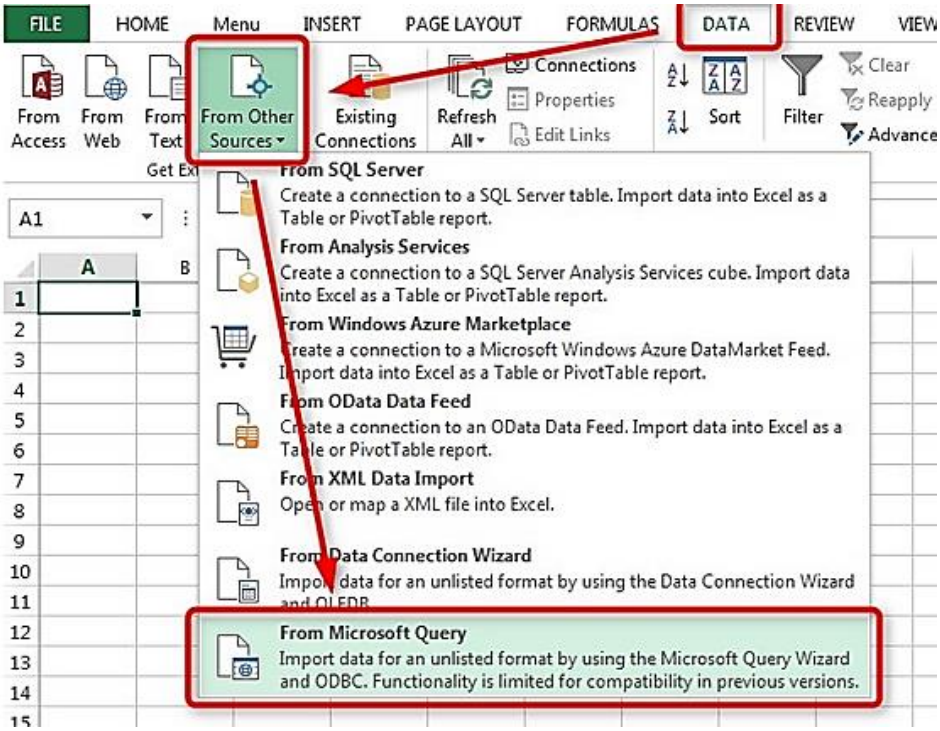


Слика 30. Приказ резултата након упита

На Слици 30 се може видети да табела садржи колоне које се односе на земљу у којој је твит („Tweet“) настао, као и да ли су утисци позитивни, негативни или неутрални, а то је означено са 1, 2 и 0.

Корак 5.

Након што су подаци „пречишћени“ потребно их је пребацити у Excel. А то се може урадити на начин који је приказан на Слици 31.



Слика 31. Експортовање резултата у Excel

Импортовани подаци изгледају као на Слици 32, где се може видети да табела садржи назив државе и вредност твита.

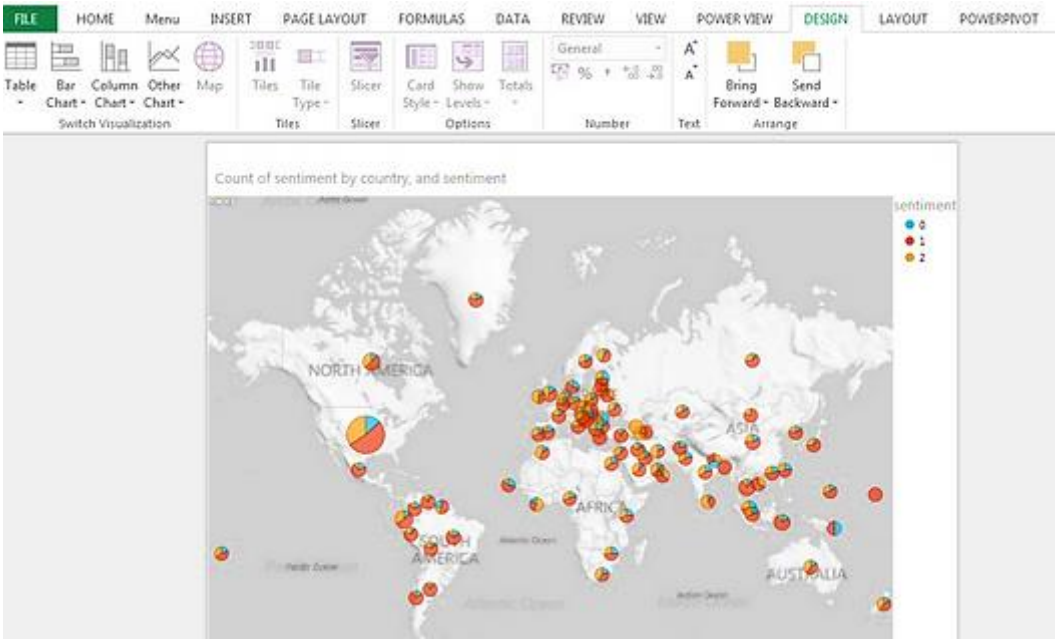
	id	ts	country	sentiment
2	3.30044E+17	5/2/2013 19:39		1
3	3.30044E+17	5/2/2013 19:39	UNITED STATES	2
4	3.30044E+17	5/2/2013 19:39	CHILE	1
5	3.30044E+17	5/2/2013 19:39		1
6	3.30044E+17	5/2/2013 19:39	MOROCCO	2
7	3.30044E+17	5/2/2013 19:39	UNITED STATES	1
8	3.30044E+17	5/2/2013 19:39		1
9	3.30044E+17	5/2/2013 19:39	UNITED STATES	1
10	3.30044E+17	5/2/2013 19:39		1
11	3.30044E+17	5/2/2013 19:39	NETHERLANDS	0
12	3.30044E+17	5/2/2013 19:39	THAILAND	1
13	3.30044E+17	5/2/2013 19:39	UNITED STATES	2
14	3.30044E+17	5/2/2013 19:39		1
15	3.30044E+17	5/2/2013 19:39	UNITED STATES	2
16	3.30044E+17	5/2/2013 19:39	INDONESIA	1
17	3.30044E+17	5/2/2013 19:39	ARGENTINA	1
18	3.30044E+17	5/2/2013 19:39	UNITED STATES	2
19	3.30044E+17	5/2/2013 19:39	THAILAND	1
20	3.30044E+17	5/2/2013 19:39		1
21	3.30044E+17	5/2/2013 19:39	PAKISTAN	2
22	3.30044E+17	5/2/2013 19:39	UNITED STATES	1
23	3.30044E+17	5/2/2013 19:39	IRAQ	1

Слика 32. Импортовани подаци

Након што су Twitter sentiment подаци импортовани у Excel потребно је још визуелно приказати резултате и анализирати, што представља следећи и последњи корак у овој обради примера.

Корак 6.

Визуелизација података као што је већ поменуто олакшава преглед и анализу добијених резултата. У наредном делу биће приказано какви су утисци, односно sentiment data варирали од земље до земље, након емитовања филма. Уз помоћ Excel Power View резултати се могу приказати као на Слици 33.



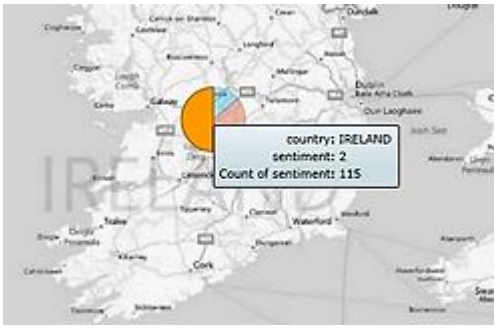
Слика 33. Визуелизација резултата

На Слици 33, односно на графикону, плавом бојом су приказани негативни, наранџастом позитивни, и жутом неутрални sentiment подаци.

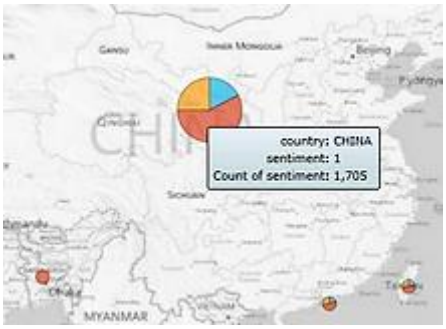
Сада се може урадити детаљнија анализа на појединачним и карактеристичним земљама. Тако на пример уколико се увећају подаци, односно графикон који се односи на податке из Ирске може се уочити да половина прикупљних података даје позизивне утиске о филму што се може видети на Слици 34. б).

Следећа је анализа података на територији Мексика. У Мексику, око једне петине твитова изразио негативно осећање (приказано у плавој боји), а само мали део твеетова су били позитивни. Већина твитова из Мексика су били неутрални, као што је приказано у црвеној боји на Слици 34. а)

а)



б)



в)

г)

Слика 34. Појединачни приказ резултата за:

а) Мексико, б) Ирску, в) Кину и г) САД

Кинески студио DMG.co је финансирао снимање филма Iron Man 3 и глумци су познати глумци из Кине, тако да уколико се анализира дијаграм са подацима из Кине (Слика 36. в)), може се уочити да је највећи део неутралан, и доста више је позитивних од негативних утисака, што је и логично.

САД је највеће тржиште, па уколико се погледају подаци тамо може се видети погледамо сентимента податке тамо да релативно велики број укупних твеетова долазе из US. Половина твитова у САД показује неутралне утиске, са релативно малом количином негативних, што је приказано на Слици 34. г).

Анализом ових података може се у будуће побољшати маркетинг од неких на пример наредних издања Iron Man-а. Овај пример је имао циљ да покаже колико је важно направити такав систем који ће обрадити велике податке, а што је најважније приказати и анализарати све те податке, и увидети како је производ утицао на кориснике, у овом случају како је публика реаговала на филм, што не значи да се овакав принцип не може применити на било које тржиште, односно производ. Коришћењем овакве анализе компаније би постале успешније.

Пример 2: Анализа података очитаних са сензора

У Примеру 2 је приказана анализа података чији су извори сензори машина, конкретно сензора који мере температуру. Уз помоћ до сада приказаних Hadoop алата у овом раду, биће објашњено како су убачени подаци у HDFS, како су анализирани и визуелно представљени, а најбитније чему ти резултати служе, као и у претходном примеру, при чему се у овом случају разликују извори података и понеки алат који је коришћен за обраду.

Овај пример ће приказати како анализом података учитаних са сензора који се налазе у клима уређајима и мере температуру, могу утицати на то како да се обезбеди отимална температура у стамбеним зградама, и како на прави начин одабрати прави модел клима уређаја.

Sensor Data

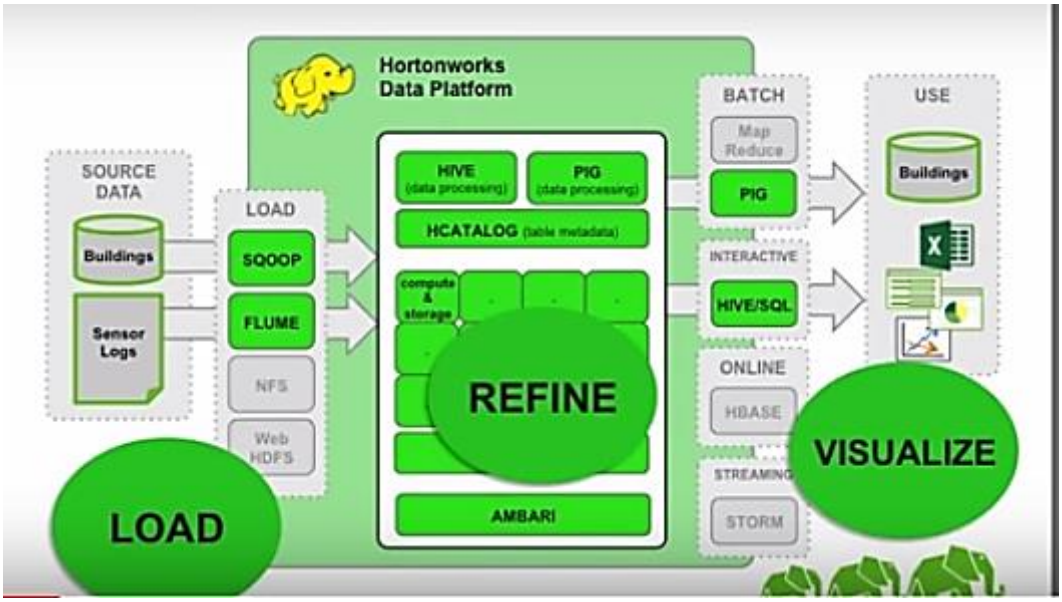
Сензор је уређај који мери физичке величине и претавра их у дигитални сигнал. Податке које очитава сензор називају се Sensor Data.

Сензори прикупљају података са многих извора, а могу се користити за:

- Праћење машине или инфраструктуре, као што су вентилациони системи, мостови, бројила или авионски мотори. Подаци очитани са оваквих система се могу користити за предиктивну аналитику, поправку или замену пре него што дође до евентуалног квара.

- Праћење природних појава као што су метеоролошке појаве, подземни притисци током експлоатације нафте, или праћење виталне статистике пацијената током опоравка након неког медицинског поступка.

Конкретно, у овом примеру обрађени су и анализирани су подаци очитани са сензора који су уграђени у системе који служе за грејање, вентилацију, тачније речено климатизацију (HVAC) у 20 великих објеката широм света. Као и у Примеру 1 ради се о Hortonworks-овој платформи и алатима већ поменутим у овом раду. На Слици 35 се могу видети алати који су потребни за реализацију овог примера.



Слика 35. Систем за обраду сензор података

Да се ни би понављали технике који су већ објашњени у Примеру 1 кроз кораке описане у наредном делу биће објашњено оно што се у систему разликује од претходног примера.

У кратким цртама оно што је урађено током обраде ових сензорских података је:

- Убациавње података са сензора у HDFS помоћу Flume
- Убацивање структуираних података у HDFS помоћу Sqoop
- Коришћење Hcatalog како би се добио релациони приказ података
- Коришћење Hive и Pig за прераду података
- Увоз података у Microsoft Excel помоћу ODBC конектора
- Визуелизација резултата помоћу Powerview

Слично као и у претходном примеру, кораци којима се потребни за реализацију овог примера су следећи:

1. Преузимање фајлова са сензор подацима.

2. Убацивање сензор података у Hortonworks Sandbox.

3. Покретање Hive скрипте за обраду сензор података

4. Убацивање обрађених података у Microsoft Excel.

5. Визуелно представљање податка уз помоћ Excel Power View.

Корак 1.

Потребни фајлови са сензор подацима преузети су директно са Hortonworks-овог сајта. У оквиру тог фолдера налезе се две врсте фајлова, а то су:

- ***HVAC.csv*** – садржи жељену температуру заједно са измереном, односно стварном температуром. Темеператрни подаци су добијени уз помоћ Flume. Flume се може користити као лог агрегатор, за прикупљање података из многих различитих извора и за премештање у централизовано складиште података. У овом случају, Flume је коришћен за снимање података лог сензора, након чега се подаци могу учитати у Хадооп дистрибуирани фајл систем (HDFS).

- ***building.csv*** – садржи „објекате“ табела базе података. Apache Sqoop се користи за пренос ове врсте података из структуриране базе у HDFS.

Корак 2.

У кораку 2 се користи Sandbox Hue и уз помоћ Hcatalog креирају се табеле са подацима из поменутих фајлова. За сваки фајл креира се засебна табела, односно добијају се две табеле приказане на Слици 36 и Слици 37.

Уколико се погледају ове две табеле може се видети да прва садржи колоне као што су време, место, жељена температура, стварна температура а друга друга табела садржи моделе клима уређаја, назив земље у којој се налази зграда, старост зграде.

Query Results: hvac

DOWNLOADS

Download as CSV

Download as XLS

Save

Did you know? If the result contains a large number of columns, click a row to select a column to jump to. As you type into the field, a drop-down list displays column names that match the string.

Results | Query | Log | Columns

	date	time	targettemp	actualtemp	system	systemage	buildingid
0	6/1/13	0:00:01	66	58	13	20	4
1	6/2/13	1:00:01	69	68	3	20	17
2	6/3/13	2:00:01	70	73	17	20	18
3	6/4/13	3:00:01	67	63	2	23	15
4	6/5/13	4:00:01	68	74	16	9	3
5	6/6/13	5:00:01	67	56	13	28	4
6	6/7/13	6:00:01	70	58	12	24	2
7	6/8/13	7:00:01	70	73	20	26	16
8	6/9/13	8:00:01	66	69	16	9	9
9	6/10/13	9:00:01	65	57	6	5	12

Слика 36. Приказ табеле „hvac“

Query Results: building

DOWNLOADS

Download as CSV

Download as XLS

Save

Did you know? If the result contains a large number of columns, click a row to select a column to jump to. As you type into the field, a drop-down list displays column names that match the string.

Results | Query | Log | Columns

	buildingid	buildingmgr	buildingage	hvacproduct	country
0	1	M1	25	AC1000	USA
1	2	M2	27	FN39TG	France
2	3	M3	28	JDNS77	Brazil
3	4	M4	17	GG1919	Finland
4	5	M5	3	ACMAX22	Hong Kong
5	6	M6	9	AC1000	Singapore
6	7	M7	13	FN39TG	South Africa
7	8	M8	25	JDNS77	Australia
8	9	M9	11	GG1919	Mexico
9	10	M10	23	ACMAX22	China

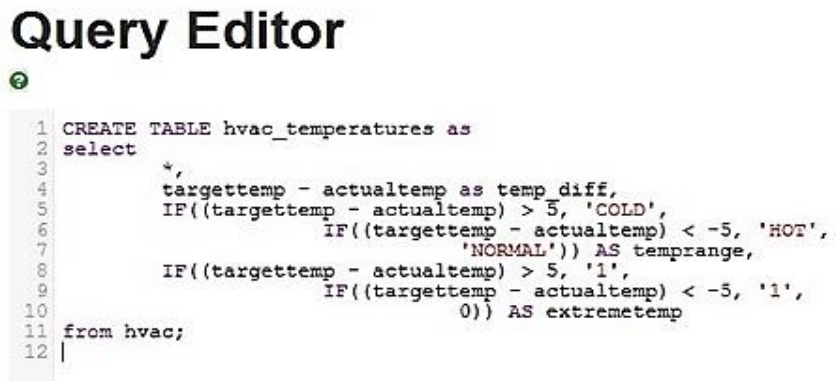
Слика 37. Приказ табеле „building“

Корак 3.

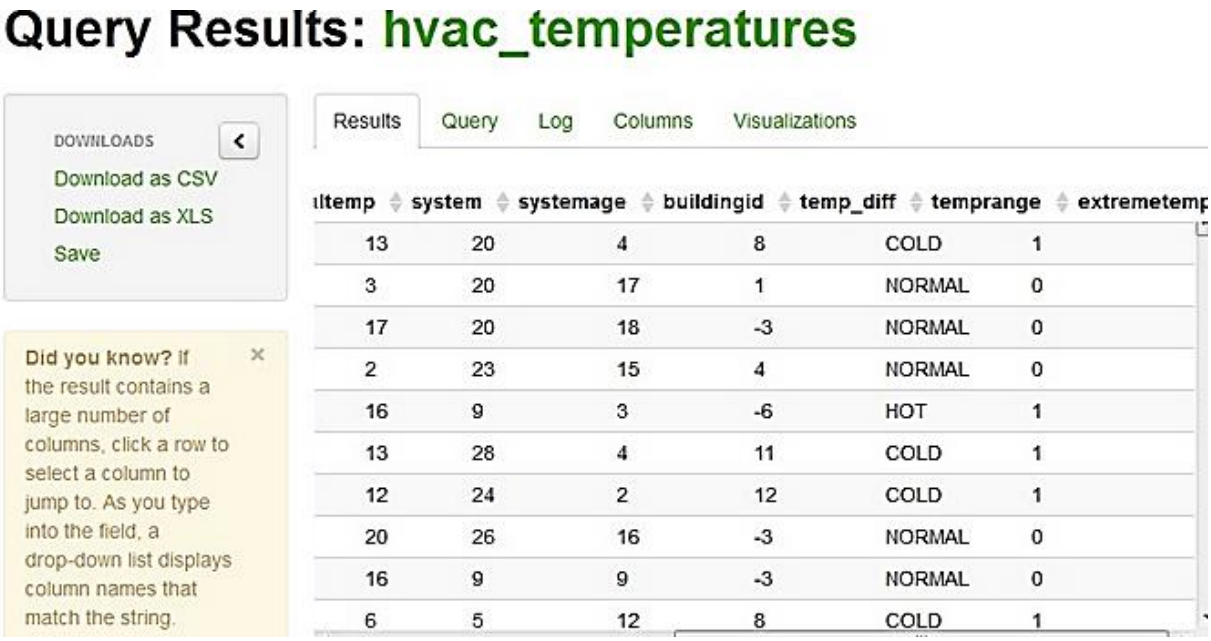
У Кораку 3 користе се две Hive скрипте како би се побољшали резултати, које имају задатак да остваре три циља а то су:

- Смањивање трошкова грејања и хлађења
- Одржавање оптималне температуре у згради
- Утврђивање модела клима уређајам односно који су модели поуздани, и одредити оне који нису и заменити их одговарајућим.

Најпре, прво се одређује да ли је стварна температура за 5 степени виша или нижа у односу на жељену температуру. Упит који се поставља је приказан на Слици 38 и односи се на табелу hvac. Резултати овог упита приказани су на Слици 39.



Слика 38. Упит за креирање табеле 1

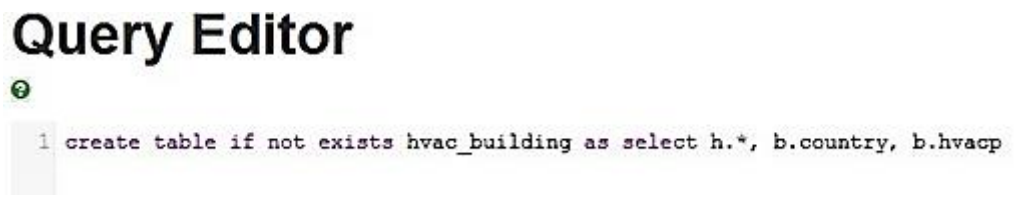


ıltemp	system	systemage	buildingid	temp_diff	temprange	extremetemp
13	20	4	8	COLD	1	
3	20	17	1	NORMAL	0	
17	20	18	-3	NORMAL	0	
2	23	15	4	NORMAL	0	
16	9	3	-6	HOT	1	
13	28	4	11	COLD	1	
12	24	2	12	COLD	1	
20	26	16	-3	NORMAL	0	
16	9	9	-3	NORMAL	0	
6	5	12	8	COLD	1	

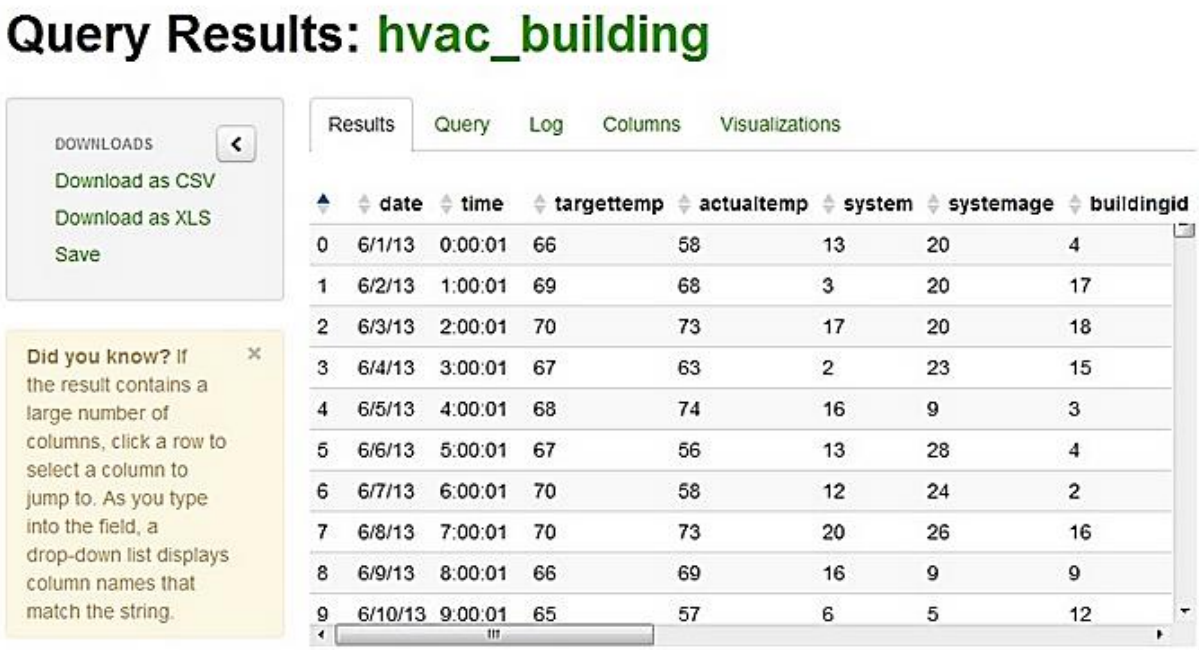
Слика 39. Резултат упита 1

Уколико је темепратура нижа за 5 степени од жељене онда је температура рангирана са COLD (hladno), уколико је виша од 5 степени онда се у поље „temprange" уписује HOT (топло) ,а уколико се поклопају стварна и жељена температура онда се уноси вредност NORMAL. Ако је температура ван нормалног опсега, у поље "extremetemp" додељује се вредност 1, у супротном њена вредност је 0.

Сада се поставља упит за другу врсту података, који је приказан на Слици 40, а резултат овог упита приказан је на Слици 41.



Слика 40. Упит за креирање табеле 2



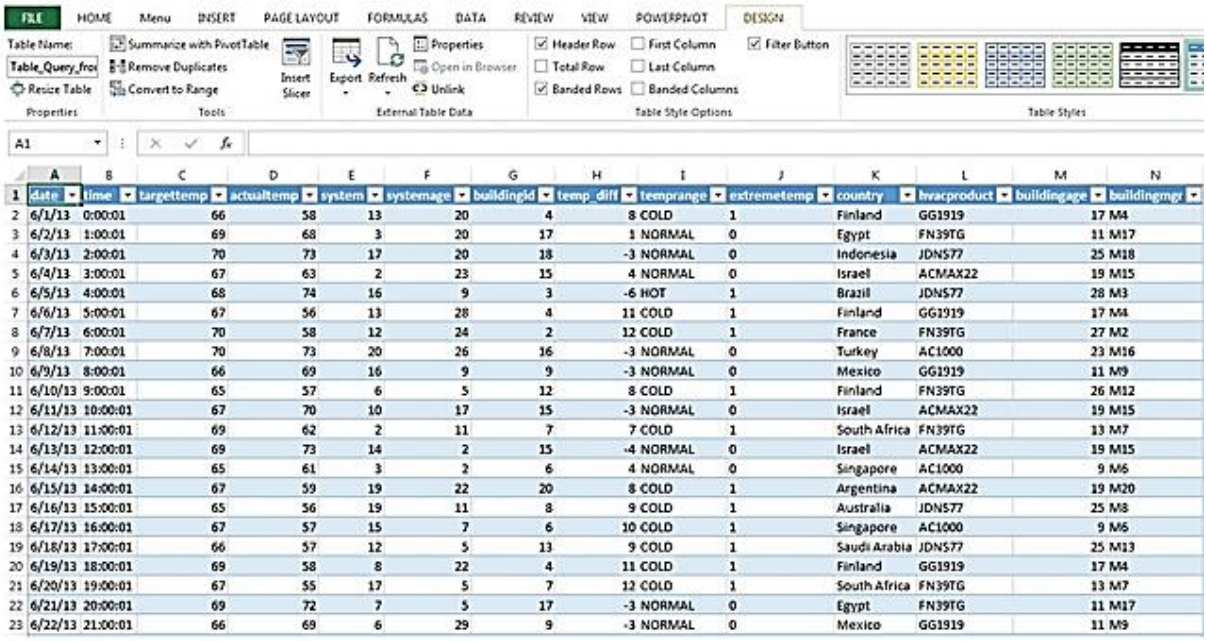
	date	time	targettemp	actualtemp	system	systemage	buildingid
0	6/1/13	0:00:01	66	58	13	20	4
1	6/2/13	1:00:01	69	68	3	20	17
2	6/3/13	2:00:01	70	73	17	20	18
3	6/4/13	3:00:01	67	63	2	23	15
4	6/5/13	4:00:01	68	74	16	9	3
5	6/6/13	5:00:01	67	56	13	28	4
6	6/7/13	6:00:01	70	58	12	24	2
7	6/8/13	7:00:01	70	73	20	26	16
8	6/9/13	8:00:01	66	69	16	9	9
9	6/10/13	9:00:01	65	57	6	5	12

Слика 41. Резултат упита 1

Након овог дела подаци се могу убацити у Excel и анализирати.

Корак 4.

На исти начин као у претходном примеру се увозе подаци у Excel и након тог дела добија се табела приказана као на слици, коју чине претходне две табеле спојено:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	date	time	targettemp	actualtemp	system	systemage	buildingid	temp_diff	temprange	extremetemp	country	hvacproduct	buildingage	buildingmgr
2	6/1/13	0:00:01	66	58	13	20	4	8	COLD	1	Finland	GG1919	17	M4
3	6/2/13	1:00:01	69	68	3	20	17	1	NORMAL	0	Egypt	FN39TG	11	M17
4	6/3/13	2:00:01	70	73	17	20	18	-3	NORMAL	0	Indonesia	JDNS77	25	M18
5	6/4/13	3:00:01	67	63	2	23	15	4	NORMAL	0	Israel	ACMAX22	19	M15
6	6/5/13	4:00:01	68	74	16	9	3	-6	HOT	1	Brazil	JDNS77	28	M3
7	6/6/13	5:00:01	67	56	13	28	4	11	COLD	1	Finland	GG1919	17	M4
8	6/7/13	6:00:01	70	58	12	24	2	12	COLD	1	France	FN39TG	27	M2
9	6/8/13	7:00:01	70	73	20	26	16	-3	NORMAL	0	Turkey	AC1000	23	M16
10	6/9/13	8:00:01	66	69	16	9	9	-3	NORMAL	0	Mexico	GG1919	11	M9
11	6/10/13	9:00:01	65	57	6	5	12	8	COLD	1	Finland	FN39TG	26	M12
12	6/11/13	10:00:01	67	70	10	17	15	-3	NORMAL	0	Israel	ACMAX22	19	M15
13	6/12/13	11:00:01	69	62	2	11	7	7	COLD	1	South Africa	FN39TG	13	M7
14	6/13/13	12:00:01	69	73	14	2	15	-4	NORMAL	0	Israel	ACMAX22	19	M15
15	6/14/13	13:00:01	65	61	3	2	6	4	NORMAL	0	Singapore	AC1000	9	M6
16	6/15/13	14:00:01	67	59	19	22	20	8	COLD	1	Argentina	ACMAX22	19	M20
17	6/16/13	15:00:01	65	56	19	11	8	9	COLD	1	Australia	JDNS77	25	M8
18	6/17/13	16:00:01	67	57	15	7	6	10	COLD	1	Singapore	AC1000	9	M6
19	6/18/13	17:00:01	66	57	12	5	13	9	COLD	1	Saudi Arabia	JDNS77	25	M13
20	6/19/13	18:00:01	69	58	8	22	4	11	COLD	1	Finland	GG1919	17	M4
21	6/20/13	19:00:01	67	55	17	5	7	12	COLD	1	South Africa	FN39TG	13	M7
22	6/21/13	20:00:01	69	72	7	5	17	-3	NORMAL	0	Egypt	FN39TG	11	M17
23	6/22/13	21:00:01	66	69	6	29	9	-3	NORMAL	0	Mexico	GG1919	11	M9

Слика 42. Приказ података који су увезени у Excel

Корак 5.

У овом кораку се врши визуелизација података као и њихова анализа. Прво је урађена визуелизација за податке оних зграда које су најчешће ван оптималног опсега температуре.

На Слици 43 је приказано како су зграде у Финској са укупно 814 сензора измериле температуру која је била виша или нижа од жељене температуре. Насупрот томе, Немачки уред на пример много боље обавља посао одржавања идеалне температуре у зградама, код којих је са само 363 сензора очитана температура ван идеалног опсега.

Слика 43. Приказ резултата за Финску

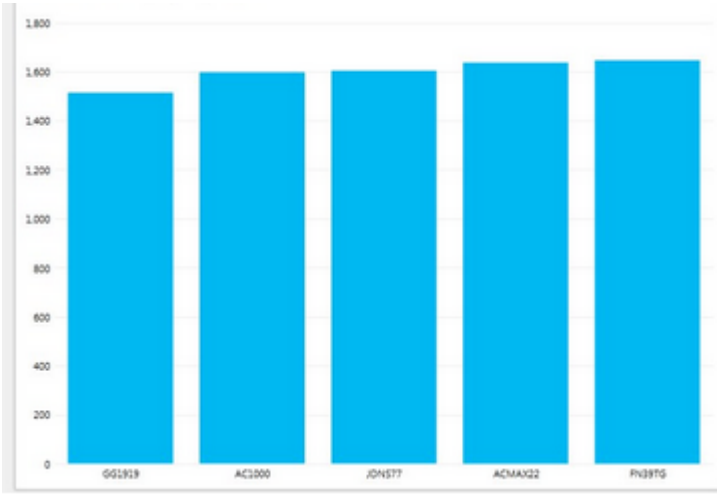
Слика 44. Приказ резултата 1

Слика 45. Приказ резултата 2

На Слици 44 се види да зграде у Финској и Француској поседују температуру која је изнад жељене. Док на Слици 45 се може видети да је ситуација у Финској и Индонезији испод жељене температуре.

Анализа ових података може помоћи и у тумачењу модела система за вентилацију и климатизацију. У табелама се могу уочити неколико модела ових система, и може се видети који модел је релативно поуздан да одржава оптималну температуру.

	A	B
1	extremetemp	hvacproduct
2	1	GG1919
3	0	FN39TG
4	0	JDN577
5	0	ACMAX22
6	1	JDN577
7	1	GG1919
8	1	FN39TG
9	0	AC1000
10	0	GG1919
11	1	FN39TG
12	0	ACMAX22
13	1	FN39TG
14	0	ACMAX22
15	0	AC1000
16	1	ACMAX22
17	1	JDN577
18	1	AC1000
19	1	JDN577
20	1	GG1919
21	1	FN39TG
22	0	FN39TG
23	0	GG1919



Слика 46. Модели система и резултати истраживања

На Слици 46. се може видети да је GG1919 модел клима уређаја регулише температуру најпоузданије, док је FN39TG није успео да одржи одговарајући температурни опсег 9% чешће него GG1919. Дакле, може се видети који модел заправо најпоузаданији а који је најмање поуздан, па на основу ових закључака могу се побољшати системи за вентилацију и климатизацију у различитим зградама, односно објектима.

7. Закључак

Big Data се примењује све више у свим областима пословања и као таква пружа конкурентску предност. Игнорисање Big Data ће ставити предузеће у позицију ризика и могућности заостајања за конкуренцијом. Како би остала кокурентна, предузећа ће у свом пословању морати да прикупљају све више података из нових извора како би добили што бољи увид у пословање. Кроз рад је више пута споменуто, а посебно у поглављу 2, како Big Data утиче на пословање и како је искористити.

Big Data као појам је још увек неистражен и биће потребно уложити још више напора у његово истраживање. Заправо може се рећи да Big Data није појам који је нов у подручју информационе технологије, него логичан наставак развоја технологије. Такође треба бити опрезнан при коришћењу Big Data као технологије која ће решити све пословне проблеме те се мора сматрати као додатак постојећој ИТ инфраструктру и пословању.

Поред технологије која се развила и која ће се развијати а која подржава обраду велике количине података, веома је значајна и сама анализа. Она зависи од аналитичара више него од саме технологије. Као што напоменух технологија се развила, и још више ће се развијати али битно је знати искоритити све то што „техника“ обради.

Big Data како са собом носи многе предности, тако има и веома велике мане. С обзиром да су све више и више доступни разни подаци о корисницама, доводи се у питање и етика, као и заштита можда неких приватних података самих корисника. Јер сваки човек представља извор података, а да тога није ни свестан.

Са стране компанија и фирми појам Big Data треба искористи на што бољи могући начин како би фирме напредовале и усавршиле пословање, као што је приказано на примерима у овом раду, док са стране корисника треба имати на уму да на пример сваки „клик“ на Интернету ствара податак који може бити искоришћен.

Сматрам да треба створити код људи свест о томе шта је Big Data, колико брзо напредује, како функционише и како је на прави начин треба искористити, које су њене добре а које лоше стране. Уколико се створи свест о Big Data, она ће представљати прави ресурс који се може искористити и уновчити.

Референце

[1] Soumendra Mohanty, Madhu Jagadeesh, Harsha Srivatsa: Big Data Imperatives, 2011.

[2] Nitin Sawant , Himanshu Shah: Big Data Application Architecture Q&A, 2011

[3] Живко Кртић: Дипломски рад „Data i semantička analiza: Iskorištavanje vrijednosti nestrukturiranih podataka u poslovanju“, Сплит 2014.

[4] Никола Станковић, Драган Ђурђевић, Марко Макарић, Срђан Терзић: Семинарски рад „Big Data“, Универзитет у Београду, 2012.

[5] Растко Павловић, Ратко Дејановић: Big Data и пословна интелигенција, Бања Лука, 2014.

[6] Проф.др Славко Арсовски, Интегрисани системи менаџмента,Факултет инжењерских наука, Крагујевац, 2013.

[7] Хрвоје Габелица, Живко Крстић: Истраживачки рад „Primjena Big Data podataka i rudarenja teksta u suvremenom poslovanju“, Економски факултет, Сплит, 2013.

[8] Велида Кијевчанин, Шабан Грачанин: Самостални истраживачки рад „DATA MINING“, Универзитет у Крагујвцу, 2009.

[9] Немања Вељковић: Нерелационе базе података NoSql, Универзитет у Београду Математички факултет, 2013

[10] http://www.raf.edu.rs/citaliste/istorija/3624-razvoj-baza-podataka (приступљено: 23.08.2015.)

[11] http://istorija-razvoja-baze-podataka.blogspot.com/search?updated-min=2013-01-01T00:00:00%2B01:00&updated-max=2014-01-01T00:00:00%2B01:00&max-results=1 (приступљено: 23.08.2015.)

[12] http://www.hadoop-srbija.com/big-data-business-model-maturity-index/ (приступљено: 15.08.2015.)

[13] http://bif.rs/2014/06/potencijal-velikih-baza-podataka-kako-predvideti-ponasanje-kupca (приступљено: 27.08.2015.)

[14] http://www.infotrend.hr/clanak/2013/11/big-data---izazovi-implementacije,78,1029.html (приступљено: 17.08.2015.)

[15] http://hortonworks.com/solutions/ (приступљено: 16.09.2015.)

[16] http://www.computermuseum.li/Testpage/Hollerith-Tabulator-1890.htm (приступљено: 23.08.2015.)

[17] https://www.flickr.com/photos/dionh/7550578346/(приступљено: 03.08.2015.)

[18] http://hortonworks.com/solutions/ (приступљено: 15.09.2015.)

[19] http://www.hadoop-srbija.com/ (приступљено: 03.09.2015.)

[20] http://www.computermuseum.li/Testpage/Hollerith-Tabulator-1890.htm (приступљено: 23.08.2015.)

[21] https://pt.wikipedia.org/wiki/Supercomputador (приступљено: 23.08.2015.)

[22] http://thesocialmediamonthly.com/empowered-data-inspired-words-successful-use-social-media/ (приступљено: 21.09.2015.)

[23] http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html (приступљено: 21.09.2015.)