DISTRIBUIRANO PROCESIRANJE VELIKIH PODATAKA U POSLOVNIM PROCESIMA

Krizmanić, Gordana

Master's thesis / Diplomski rad

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Faculty of Economics in Osijek / Sveučilište Josipa Jurja Strossmayera u Osijeku, Ekonomski fakultet u Osijeku**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:145:311460

Sveučilište Josipa Jurja Strossmayera u Osijeku
Ekonomski fakultet u Osijeku
Diplomski studij *Poslovna Informatika*

Gordana Krizmanić

DISTRIBUIRANO PROCESIRANJE VELIKIH PODATAKA U POSLOVNIM PROCESIMA

Diplomski rad

Osijek, 2021

Sveučilište Josipa Jurja Strossmayera u Osijeku
Ekonomski fakultet u Osijeku
Diplomski studij *Poslovna Informatika*

Gordana Krizmanić

DISTRIBUIRANO PROCESIRANJE VELIKIH PODATAKA U POSLOVNIM PROCESIMA

Diplomski rad

Kolegij: Sustavi poslovne inteligencije

JMBAG: 0010135166
e-mail: gkrizmanic@efos.hr

Mentor: doc.dr.sc. Slobodan Jelić

Osijek, 2021

Josip Juraj Strossmayer University of Osijek
Faculty of Economics in Osijek
Graduate Study Business Informatics

Gordana Krizmanić

DISTRIBUTED BIG DATA PROCESSING IN BUSINESS PROCESSES

Graduate paper

Osijek, 2021

IZJAVA

O AKADEMSKOJ ČESTITOSTI, PRAVU PRIJENOSA INTELEKTUALNOG VLASNIŠTVA, SUGLASNOSTI ZA OBJAVU U INSTITUCIJSKIM REPOZITORIJIMA I ISTOVJETNOSTI DIGITALNE I TISKANE VERZIJE RADA

1. Kojom izjavljujem i svojim potpisom potvrđujem da je diplomski (navesti vrstu rada: završni / diplomski / specijalistički / doktorski) rad isključivo rezultat osobnoga rada koji se temelji na mojim istraživanjima i oslanja se na objavljenu literaturu. Potvrđujem poštivanje nepovredivosti autorstva te točno citiranje radova drugih autora i referiranje na njih.
2. Kojom izjavljujem da je Ekonomski fakultet u Osijeku, bez naknade u vremenski i teritorijalno neograničenom opsegu, nositelj svih prava intelektualnoga vlasništva u odnosu na navedeni rad pod licencom *Creative Commons Imenovanje – Nekomercijalno – Dijeli pod istim uvjetima 3.0 Hrvatska.*
3. Kojom izjavljujem da sam suglasan/suglasna da se trajno pohrani i objavi moj rad u institucijskom digitalnom repozitoriju Ekonomskoga fakulteta u Osijeku, repozitoriju Sveučilišta Josipa Jurja Strossmayera u Osijeku te javno dostupnom repozitoriju Nacionalne i sveučilišne knjižnice u Zagrebu (u skladu s odredbama Zakona o znanstvenoj djelatnosti i visokom obrazovanju, NN br. 123/03, 198/03, 105/04, 174/04, 02/07, 46/07, 45/09, 63/11, 94/13, 139/13, 101/14, 60/15).
4. izjavljujem da sam autor/autorica predanog rada i da je sadržaj predane elektroničke datoteke u potpunosti istovjetan sa dovršenom tiskanom verzijom rada predanom u svrhu obrane istog.

Ime i prezime studenta/studentice: Gordana Krizmanić

JMBAG: 0010135166

OIB: 65348616342

e-mail za kontakt: goga.krizmanic@gmail.com

Naziv studija: Diplomski sveučilišni

Naslov rada: Distribuirano procesiranje velikih podataka u poslovnim procesima

Mentor/mentorica diplomskog rada: doc.dr.sc. Slobodan Jelić

U Osijeku, 2021 godine

Potpis Gordana Krizmanić

Distribuirano procesiranje velikih podataka u poslovnim procesima

SAŽETAK

U ovom diplomskom radu analizirat će se veliki podaci, te distribuirano procesiranje velikih podataka, a fokus će se staviti na utjecaj velikih podataka u poslovnim procesima. Ljudi svakodnevno generiraju ogromnu količinu raznorodnih podataka na temelju kojih se može doći do skrivenih informacija. Te skrivene informacije predstavljaju znanje na temelju kojih kompanije mogu donositi bolje poslovne odluke i unaprijediti svoje poslovanje, države mogu spriječiti različite vanjske ili unutarnje napade, dok u medicini mogu spasiti ljudske živote. Nakon što se podaci prikupe potrebno ih je prebaciti u upotrebljiv oblik, odnosno potrebno ih je procesirati kako bi se nad njima mogle vršiti različite analize. Apache Hadoop je alat otvorenog koda koji se koristi za spremanje, procesiranje i analizu velikih podataka. Stoga, u prvom dijelu ovog diplomskog rada fokus je stavljen na definiranje znanosti o podacima, velikim podacima, te alata Apache Hadoop i MapReduce tehnike procesiranja i njihovih mogućnosti. U drugom dijelu rada će se na praktičnom primjeru pokazati kako se veliki podaci procesiraju pomoću alata Apache Hadoop i MapReduce tehnike procesiranja. Zaključak donosi osvrt na procesiranje velikih podataka, kako se veliki podaci mogu pretvoriti u znanje pomoću alata za distribuirano procesiranje i omogućiti kompanijama da poboljšaju svoje poslovne procese, te što se očekuje u budućnosti kada su u pitanju veliki podaci i poslovni procesi.

Ključne riječi: Znanost o podacima, Veliki podaci, Apache Hadoop, MapReduce, Poslovni procesi

Distributed big data processing in business processes

ABSTRACT

In this graduate paper, big data and distributed big data processing will be analyzed, and the focus will be on the impact of big data in business processes. People generate a huge amount of diverse data daily based on which hidden information can be obtained. This hidden information represents the knowledge based on which companies can make better business decisions and improve their business, states can prevent various external or internal attacks, while in medicine they can save lives. After the data is collected, it is necessary to transfer it to

a usable form, ie it is necessary to process it so that various analyzes can be performed on it. Apache Hadoop is an open-source tool used to store, process, and analyze big data. Therefore, in the first part of this paper, the focus is on defining the data science, big data, and the Apache Hadoop tool and MapReduce processing techniques and their capabilities. The second part of the paper will show a practical example of how large data is processed using the Apache Hadoop tool and MapReduce processing technique. The conclusion looks at big data processing, how big data can be turned into knowledge using distributed processing tools and enable companies to improve their business processes, and what is expected in the future when it comes to big data and business processes.

Keywords: Data Science, Big data, Apache Hadoop, MapReduce, Business processes

SADRŽAJ

1 Uvod

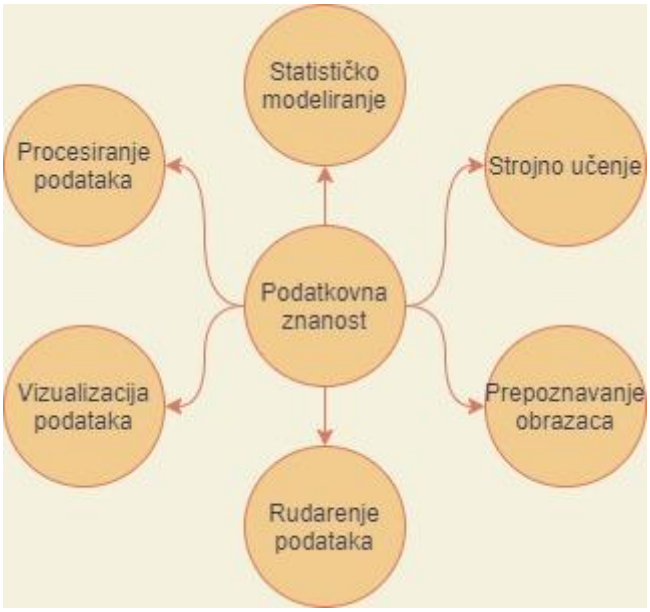
Razvojem informacijske tehnologije količina generiranih podataka eksponencijalno raste. Ti podaci sadrže skrivena znanja koja treba otkriti i upotrijebiti. Razvoj podatkovne znanosti donosi metode i tehnologije kojima se iz velike količine podataka može izvući znanje. Jedan od glavnih ciljeva podatkovne znanosti je pronaći skriveno znanje koje se krije u podacima, a na temelju kojih će poslovni subjekti donositi odluke, poboljšati svoje poslovanje i na kraju ostvariti konkurentsku prednost. Promjene u okolini se događaju brzo i poslovni subjekti moraju biti sposobni brzo se prilagoditi tim promjenama. Veliki podaci kriju te informacije, međutim volumen podataka koji se prikuplja je ogroman, te svakodnevno eksponencijalno raste, a tradicionalni alati nemaju sposobnost spremanja i procesiranja tako velike količine podataka. Jedan od načina kako doći do znanja iz velikih podataka je distribuirano procesiranje. Hadoop, čiji je razvoj počeo još 2003. godine, je alat otvorenog koda koji se koristi za distribuiranu pohranu i procesiranje velikih podataka, a njegove dvije glavne komponente su HDFS i MapReduce. HDFS je distribuirani datotečni sustav koji omogućava spremanje velike količine podataka na klasteru računala, a radi na master-slave principu. MapReduce je model programiranja za paralelno obrađivanje velikih podataka, a temelji se na generiranju izlaza na osnovi ulaznih podataka i sastoji se od dva dijela, odnosno zadatka, *Map* i *Reduce*.

Poslovni procesi u organizacijama nisi uvijek standardizirani, neki poslovni procesi su dinamični i zahtijevaju da se brzo prilagođavaju promjenama u okruženju organizacije. Zahvaljujući velikim podacima organizacije mogu otkriti promjene koje će aktivirati i promjene u njihovim poslovnim procesima. Veliki utjecaj na poslovne subjekte i njihove poslovne procese imala je pojava COVID-19 virusa koja je paralizirala cijeli svijet. Poslovni subjekti koji su svoje poslovne procese uspjeli brzo prilagoditi novim promjenama su nastavili poslovati, dok ih je dosta moralo zatvoriti svoja poduzeća, jer se nisu uspjeli dovoljno brzo prilagoditi promjenama u okolini. U ovom radu pokazat će se kako se koriste Hadoop i distribuirano procesiranja velikih podataka koristeći podatke o COVID-19 pandemiji, odnosno podacima o mjerama koje su vlade zemalja poduzele kako bi se suzbila pandemija, te što su tvrtke morale poduzeti kako bi se prilagodile novonastaloj situaciji. Podaci su prikupljeni u periodu između 01.01.2020. godine do 06.08.2021. godine i nalaze se u csv formatu.

2 Teorijska podloga i prethodna istraživanja

2.1 Podatkovna znanost

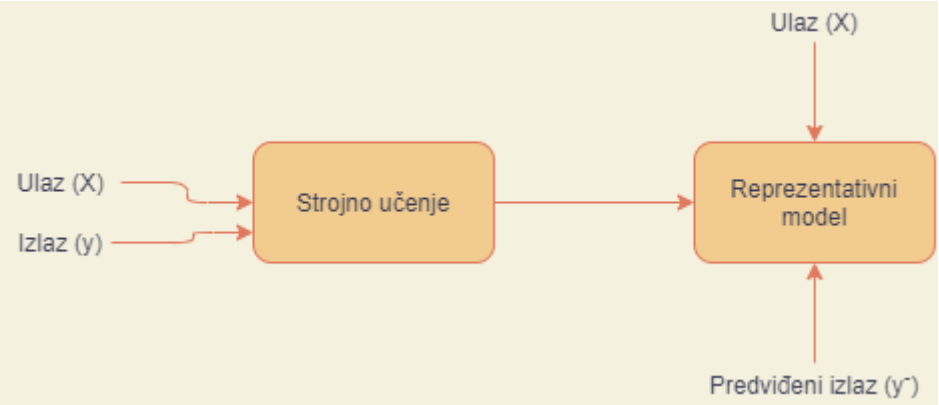
Svakodnevno se generira velika količina podataka na temelju kojih se može doći do skrivenih informacija. Razvoj tehnologije omogućio je lakše i brže pronalaženje skrivenih znanja iz velike količine generiranih podataka. Podatkovna znanost sadrži različite tehnike pomoću kojih podatkovnih znanstvenici pronalaze različite obrasce i veze među podacima. Kotu i Deshpande (2018) smatraju da je podatkovna znanost interdisciplinarno područje koje izvlači vrijednost iz podataka.



Slika 1 Podatkovna znanost (Madbouly, Al-falluji, 2019)

Jedan od glavnih ciljeva podatkovne znanosti je pronaći skriveno znanje koje se krije u podacima, a na temelju kojih će se donositi odluke. Kako bi se to postiglo potrebno je pomoću različitih obrazaca pronaći veze među podacima. Pronalaženje znanja provodi se u nekoliko koraka (Cooper, 2018):

1. Definiranje problema
2. Prikupljanje podataka potrebnih za rješavanje problema
3. Procesiranje podataka
4. Istraživanje podataka
5. Analiziranje podataka
6. Komuniciranje rezultata



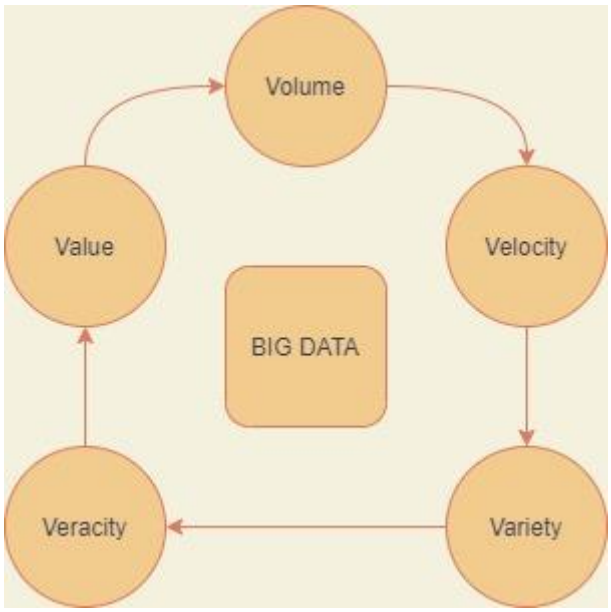
Slika 2 Modeli u podatkovnoj znanosti (Cooper, 2018)

Analiza podataka uključuje izgradnju modela koji će predvidjeti izlaz na temelju ulaznih varijabli, odnosno dati rješenje za postavljeni problem, ali i pomoći da se utvrde veze između ulaznih varijabli i izlaza.

2.2 Veliki podaci

Podaci se prikupljaju sa različitih sustava koji se koriste, spremaju u bazu podataka, šalju na analizu i transformiraju u čitljiv oblik. Volumen podataka koji se prikuplja je ogroman, te svakodnevno eksponencijalno raste, a tradicionalni alati nemaju sposobnost spremanja i procesiranja tako velike količine podataka. Podaci koji se prikupljaju mogu biti (Mayer-Schönberger, Cukier, 2013):

- Strukturirani – format strukturiranih podataka je unaprijed definiran. Primjer strukturiranih podataka je tablica u relacijskoj bazi podataka.
- Polustrukturirani – struktura i tip podataka nije unaprijed definiran, odnosno nemaju unaprijed definiranu shemu. Primjer polustrukturiranih podataka su podaci zapisani u XML dokumentu, JSON dokumentu.
- Nestrukturirani – format nestrukturiranih podataka je nepoznat. Nestrukturirani podaci mogu uključivati tekstualne podatke, video i audio podatke

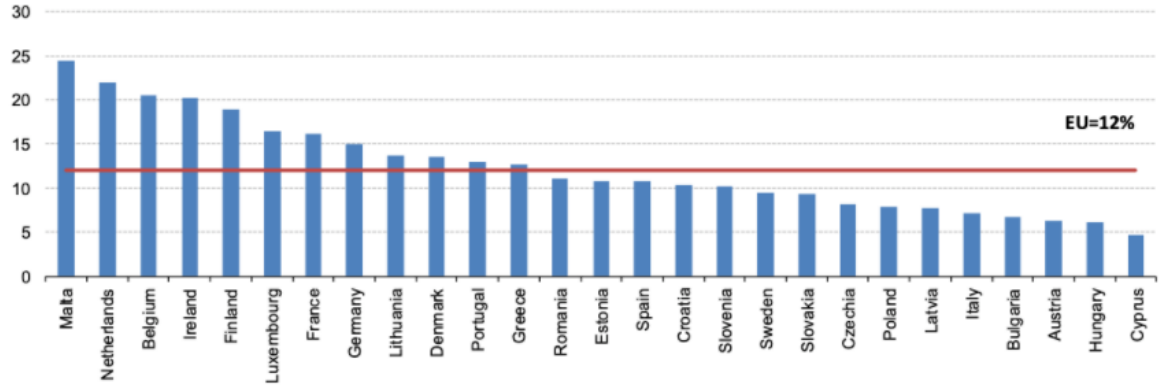


Slika 3 Veliki podaci - 5V (IIBA, 2021)

Veliki podaci definirani su pojmom 5V, a koji se odnosi na (IIBA, 2021):

1. *Volume* (Volumen) – odnosi se na količinu generiranih i procesiranih podataka
2. *Velocity* (Brzina) – odnosi se na brzinu kojom se generiraju novi podaci
3. *Variety* (Raznolikost) – odnosi se na raznolikost izvora podataka, formata podataka, tipova podataka
4. *Veracity* (Vjerodostojnost) – odnosi se na pouzdanost, autentičnost, dostupnost podataka
5. *Value* (Vrijednost) – odnosi se na vrijednost koja će se dobiti iz podataka

Prema podacima Europske komisije (2021) tvrtke u Europi se sve više bave analizom velikih podataka, unutar tvrtke (8%) ili angažiranjem vanjskih suradnika (5%). U Hrvatskoj je analizu velikih podataka koristilo oko 10% tvrtki.



Slika 4 Korištene analize velikih podataka u EU (Europska komisija, 2021)

Tvrtke u EU koje analiziraju velike podatke najčešće koriste podatke prikupljene preko (Europska komisija, 2021):

- mobilnih uređaja – 49% tvrtki
- društvenih mreža – 45% tvrtki
- vlastitih pametnih uređaja ili senzora – 29% tvrtki
- drugi izvori – 26% tvrtki

2.3 NoSQL

NoSQL baze podataka su nestrukturirane baze podataka, a razvile su se iz potrebe za spremanjem, te dostupnošću velike količine podataka. Njihova glavna prednost u odnosu na relacijske baze podataka je brzina, agilnost, a njihova distribuirana arhitektura otvara mogućnost vodoravnog skaliranja, odnosno dodavanja novih poslužitelja.

Slika 5 Modeli podataka prema kompleksnosti i povezanosti (The Enlightened DBA, 2014)

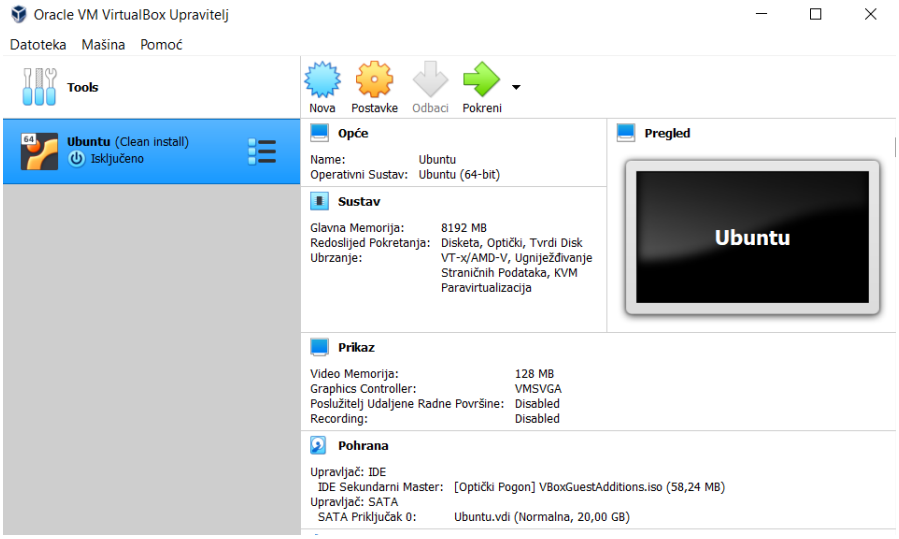
NoSQL baze razlikuju se prema modelu i metodama na temelju kojih spremaju podatke. Obzirom na to, noSQL baze dijele se na (The Enlightened DBA, 2014):

- Tablične (*Tabular*) baze podataka – podaci se spremaju u tablice, redove i dinamičke kolone. Redovi ne moraju imati iste kolone. Primjer tabličnih baza podataka su Cassandra i HBase.
- Ključ/Vrijednost (*Key/Value*) baze podataka – temelje se na principu ključ i vrijednost, gdje se do vrijednosti dolazi preko ključa. Najčešće se koriste za spremanje velike količine podataka kojima se pristupa preko jednostavnih upita. Popularne baze podataka su Redis i DynamoDB.
- Dokument (*Document*) baze podataka – podaci se spremaju u dokumente u obliku JSON ili XML formata. Vrijednosti podataka mogu biti tekstualni, numerički, polja ili objekti. Omogućavaju vertikalno skaliranje kako bi podržale velike količine podataka. Jedna od popularnih dokument baza podataka je MongoDB.
- Graf (*Graph*) baze podataka – baze podataka koje se temelje na teoriji grafova, a koriste se za spremanje kompleksnih i povezanih podataka. Temelje se na principu čvorova i

veza gdje čvorovi predstavljaju entitete, a veze informacije o odnosima između čvorova. Primjer graf baza podataka su Neo4j i JanusGraph.

2.4 VirtualBox

VirtualBox je alat otvorenog koda koji se koristi za virtualizaciju i može se instalirati na bilo koji operativni sustav, Windows, Linux, Mac OS. VirtualBox može istovremeno pokretati više virtualnih mašina, a na svakoj može biti instaliran drugi operativni sustav. Na slici 6 prikazano je sučelje VirtualBox alata koji ima jednu virtualnu mašinu na kojoj je instaliran Ubuntu.



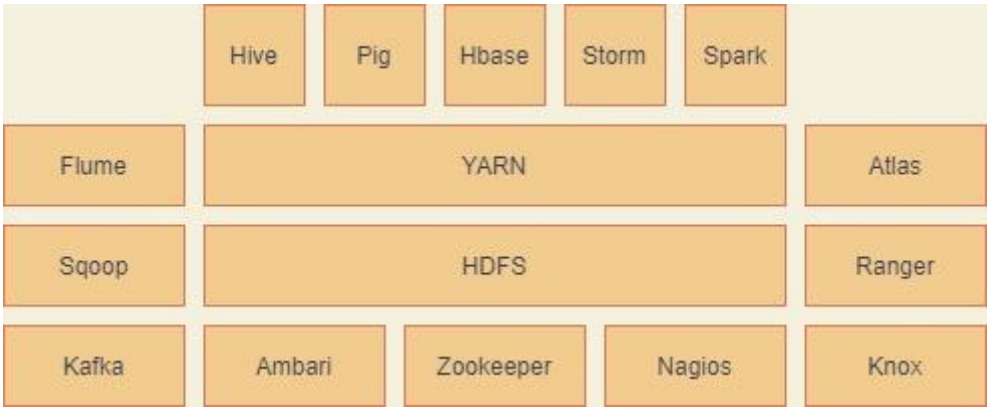
Slika 6 VirtualBox upravitelj (vlastita izrada)

Glavne značajke alata su (Oracle Corporation, n.d.):

- Prenosivost – ista virtualna mašina se može pokrenuti na Windows operativnom sustavu, ali i na Linux operativnom sustavu
- Dodaci za goste (*guest additions*) – paket koji omogućava dodatne integracije i komunikaciju sa sustavom domaćina
- Hardverska podrška
- Mogućnost kreiranja snimaka svake virtualne mašine
- Grupiranje virtualnih mašina
- Udaljeni pristup na bilo koju virtualnu mašinu koja je trenutno pokrenuta
- Čista arhitektura i modularnost

2.5 Apache Hadoop

Hadoop je alat otvorenog koda koji se koristi za distribuiranu pohranu i procesiranje velikih podataka, a njegove dvije glavne komponente su HDFS i MapReduce. Njegov razvoj započeo je 2003. godine u sklopu projekta Apache Nutch, a 2006. godine Doug Cutting ga izdvaja kao zasebni projekt kada počinje njegov intenzivniji razvoj. Već 2011. godine Yahoo koristi Hadoop na preko 40000 računala.



Slika 7 Hadoop alati (Shrivastava, Tanmay, 2016)

Hadoop platforma sastoji se od alata otvorenog koda, te komercijalnih alata koji se koriste za rad s velikim podacima, a možemo ih podijeliti u slijedeće skupine što je prikazano na slici 7 (Shrivastava, Tanmay, 2016):

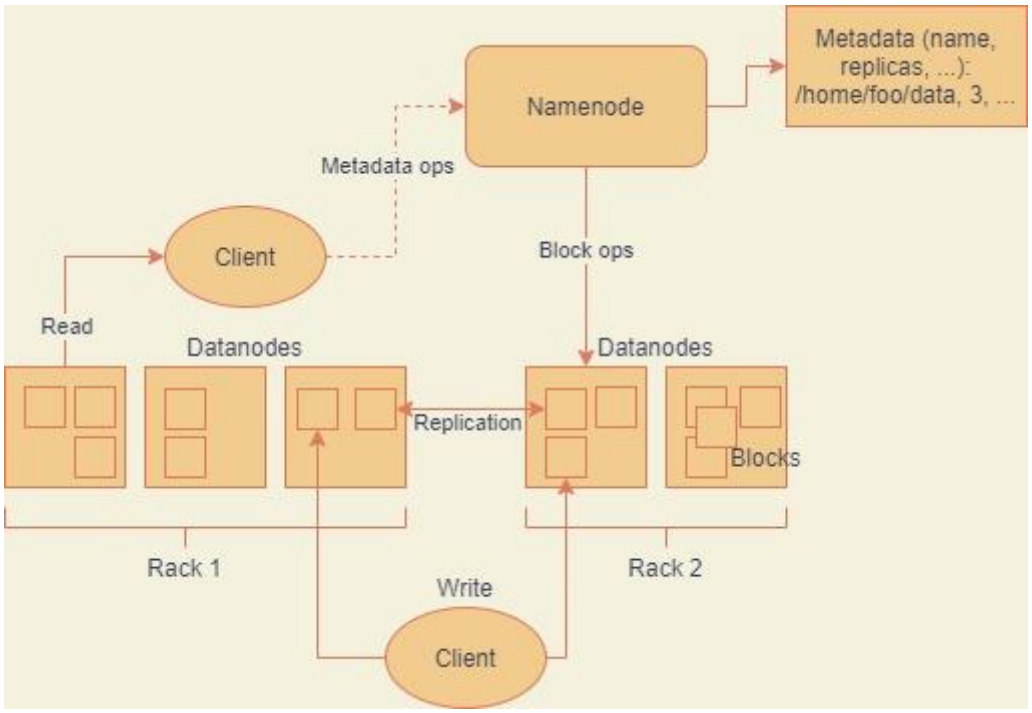
- Alati za unos podataka – alati koji se koriste za čitanje podataka sa izvora i njihovo učitavanje na HDFS. Ovoj skupini pripadaju:
  - Flume
  - Sqoop
  - Kafka
- Alati za pristup podacima – alati koji se koriste za pristup podacima na Hadoop klasterima radi analize podataka. Ovoj skupini pripadaju:
  - Hive
  - Pig
  - Hbase
  - Storm
  - Spark
- Alati za nadzor infrastrukture i resursa – alati koji se koriste za nadzor i osiguranje optimalnog rada Hadoop klastera. Ovoj skupini pripadaju:
  - Ambari
  - Zookeeper

- Nagios

- Alati za upravljanje podacima – alati koji osiguravaju integritet, sigurnost, te upotrebljivost svih podataka. Ovoj skupini pripadaju:
  - Atlas
  - Ranger
  - Knox

2.5.1 HDFS

HDFS je distribuirani datotečni sustav koji omogućava spremanje velike količine podataka na klasteru računala, odnosno sustavu umreženih računala preko brze lokalne mreže pomoću koje računala međusobno komuniciraju. Arhitektura HDFS sustava koja je prikazana na slici 8 dizajnirana je na *master-slave* principu.



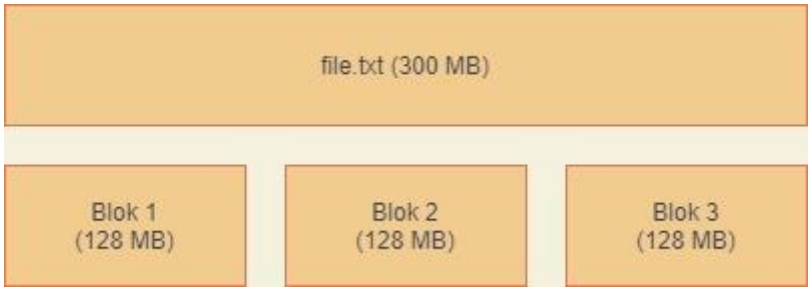
Slika 8 Arhitektura HDFS sustava (Borthakur, 2021)

Master-slave arhitektura HDFS klastera sastoji se od dva čvora, glavni čvor i podatkovni čvor gdje je glavni čvor master, a podatkovni čvor slave. U HDFS klasteru može postojati samo jedan *master*, odnosno glavni čvor, dok podatkovnih čvorova može biti više, obično jedan po čvoru u klasteru. Glavni čvor (*NameNode*) zadužen je za (Borthakur, 2021):

- Operacije otvaranja, zatvaranja i preimenovanja datoteka i direktorija
- Mapiranje blokova u podatkovnim čvorovima

Podatkovni čvor (*DataNode*) zadužen je za (Borthakur, 2021):

- Čitanje i pisanje datotečnih sustava koje su dobili od klijenata
- Kreiranje, brisanje i replikaciju blokova



Slika 9 HDFS blokovi (Shrivastava, Tanmay, 2016)

Podaci se u HDFS sustav spremaju u blokove u podatkovnim čvorovima. Zadana vrijednost svakog bloka je 128 MB, a blok se uvijek sprema na jednom računalu, odnosno sustav neće nikada podijeliti blok i spremiti ga na više računala. Slika 9 prikazuje kako će HDFS sustav podijeliti datoteku od 300 MB na blokove. HDFS sustav sprema kopije blokova na različitim podatkovnim čvorovima kako bi u slučaju kvara diska ili kvara na mreži moglo pristupiti datotekama.

2.5.2 MapReduce

MapReduce je model programiranja za paralelno obrađivanje velikih podataka, a temelji se na generiranju izlaza na osnovi ulaznih podataka. MapReduce se sastoji od dva zadatka, *Map* i *Reduce*. Ta dva zadatka pokreće posao (*job*) koji se dijeli na *JobTracker* i *TaskTracker* gdje *TaskTracker-a* može biti više.

JobTracker je odgovoran za (Borthakur, 2021):

- Zakazuje izvršavanje poslove na odabranom *TaskTracker-u*
- Vrši nadzor zadataka
- Ponovno pokreće zadatak u slučaju da se zadatak ne izvrši uspješno

TaskTracker je odgovoran za (Borthakur, 2021):

- Izvršava zadatke koje je dobio od *JobTracker-a*
- Javlja *JobTracker-u* u slučaju da se zadatak nije uspješno izvršio

Map i Reduce zadaci se izvode slijedno, a njihovi ulazi i izlazi su u obliku ključ-vrijednost (*key-value pairs*). Obrada podataka u MapReduce modelu se izvršava paralelno preko više računalnih čvorova. Izvršavanje zadatka u MapReduce modelu (Borthakur, 2021):

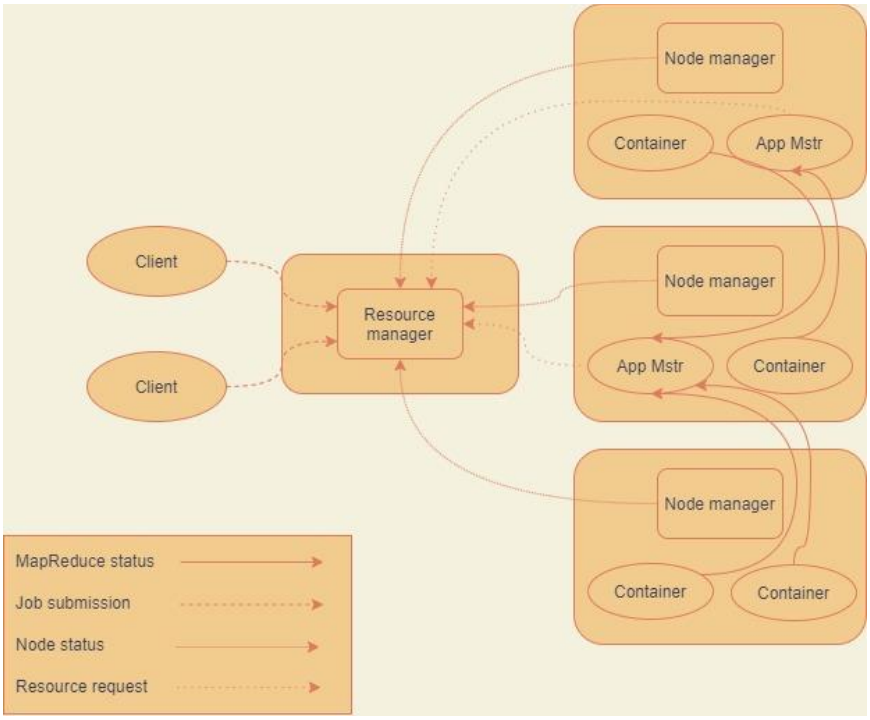
1. Map zadatak – ulaz je skup podataka u obliku ključ-vrijednost i rastavlja ih u drugi skup podataka koji su isto oblika ključ-vrijednost.

2. Reduce zadatak – ulaz su podaci iz Map zadatka oblika ključ-vrijednost. Reduce zadatak kombinira ulazne podatke u manji set podataka oblika ključ-vrijednost.

(ulaz) <k1, v1> -> **map** -> <k2, v2> -> **combine** -> <k2, v2> -> **reduce** -> <k3, v3> (izlaz)

2.5.3 YARN

Od 2012. godine YARN se pridružuje Apache Hadoop sustavu, a smjestio se između HDFS-a i sustava za procesiranje koji se koriste za pokretanje aplikacija. Prema Apache Hadoop (2021) njegova temeljna ideja je razdvojiti funkcionalnosti upravljanja resursima i praćenja poslova u zasebne pozadinske procese (*deamons*).



Slika 10 YARN sustav (Apache Hadoop, 2021)

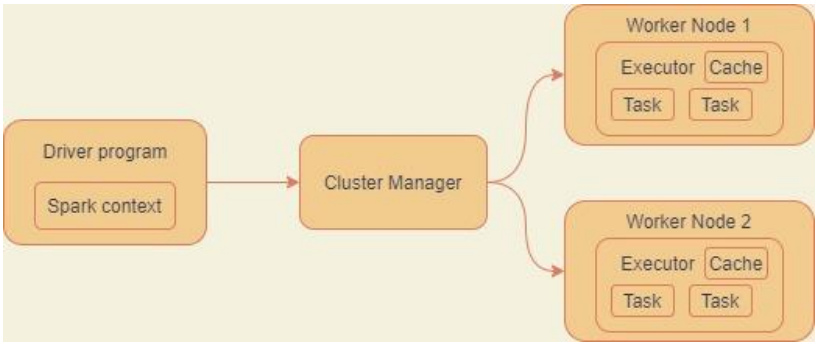
Na slici 10 prikazana je arhitektura YARN sustava koja se sastoji od sljedećih komponenti:

1. *ResourceManager* – jedan globalni koji zaprima i raspoređuje zadatke i dodjeljuje im resurse
2. *NodeManager* – jedan na svakom čvoru koji nadgleda izvršavanje zadataka i javlja njihov status *ResourceManager-u*
3. *ApplicationMaster* – kreira se jedan za svaku aplikaciju koji traži da mu se dodijele resursi i u suradnji sa *NodeManager-om* radi na izvršavanju i nadgledanju zadataka
4. *Resource container* – dodjeljuje resurse aplikacijama, te je pod kontrolom *NodeManager-a*

2.6 Apache Spark

Apache Spark je alat otvorenog koda koji se koristi za procesiranje velikih podataka. Procesiranje velikih podataka odvija se u memoriji, odnosno podaci se zapisuju na disk tek nakon što je procesiranje podataka završeno. Apache Spark sastoji se od slijedećih komponenti (Shrivastava, Tanmay, 2016):

- Machine Learning (MLib)
- Graphs (GraphX)
- Spark Streaming
- SparkSQL



Slika 11 Komunikacija između komponenti (Shrivastava, Tanmay, 2016)

Na slici 11 prikazana je komunikacija između komponenti kada se aplikacija izvršava na Spark klasteru, a prema Shrivastava, Tanmay (2016) u tom procesu sudjeluju tri komponente:

1. Upravitelj (*Driver program*) – pokreće glavnu funkciju aplikacije i stvara *SparkContext* koji koordinira obradu podataka
2. Izvršitelj – izvršava zadatke i rezultate šalje upravitelju
3. Upravitelj grozdom (*Cluster Manager*) – dodjeljuje resurse izvršitelju

Spark daje puno bolje performanse kada računalo ima dovoljno memorije da u nju stanu svi podaci koje treba obraditi u odnosu na MapReduce, ali u slučaju kada računalo nema dovoljno memorije da spremi sve podatke, performanse Sparka opadaju i bolje je koristiti MapReduce.

2.7 Poslovni procesi

Svaki posao koji se obavlja u nekoj organizaciji može se smatrati procesom, zapošljavanja djelatnika, naručivanje robe, otvaranje projekta, narudžbe kupca. Prema von Rosing i dr. (2014), poslovni proces je skup zadataka i aktivnosti koji se sastoji od zaposlenika, materijala, strojeva, sustava i metoda koji su strukturirani na takav način kako bi dizajnirali, napravili i isporučili proizvode ili usluge potrošačima. Organizacije definiraju poslovne procese kako bi opisali kako se nešto u organizaciji radi, te kako bi osigurala kontinuirano poboljšanje,

zadovoljstvo djelatnika, ali i zaštitila klijente. Poslovni proces sastoji se od niza koraka, odnosno aktivnosti koje treba odraditi kako bi se proces završio. Aktivnosti u poslovnom procesu ne izvršavaju se proizvoljno, već u nekom određenom slijedu. Svaki poslovni proces ima svoj početak i kraj. Poslovni proces aktiviran je nekim događajem, odnosno odlukom, uvjetom ili vremenom koji na kraju procesa rezultiraju proizvodom, uslugom ili informacijom. Karakteristike poslovnih procesa (Mesarić, Šebalj, 2019):

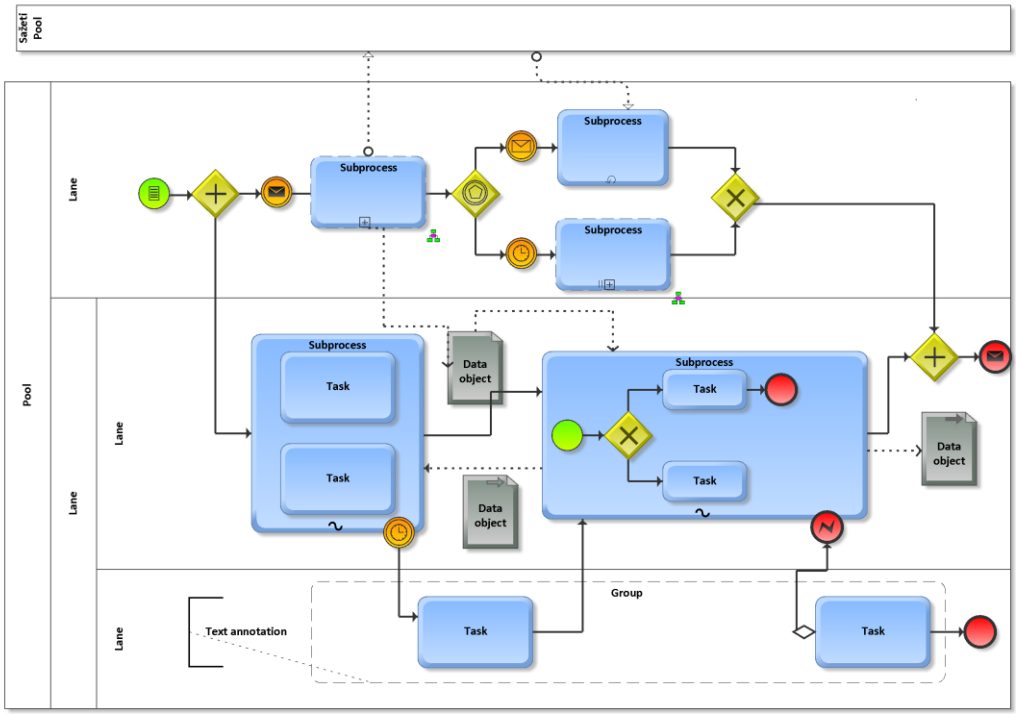
- Hijerarhija
- Mjerljivost
- Vlasnici procesa i klijenti
- Vidljivost
- Procese treba promatrati kao cjelinu
- Ponovljivost
- Prilagodljivost
- Mogućnosti za automatizaciju

Poslovne procese u poduzeću potrebno je analizirati i dokumentirati, te opisati sve njegove značajke kako bi se izbjeglo pogrešno interpretiranje i nepoštivanje ustanovljenih poslovnih procesa u poduzeću, te razumjelo kako posao teče kroz proces. Nakon što su postojeći procesi analizirani i dokumentirani, moguće je utvrditi da li su optimalno postavljeni i da li postoje područja koja se mogu poboljšati. Poslovni procesi u poduzeću ne bi trebali biti nepromjenjivi, nego bi se trebali mijenjati i prilagođavati ovisno o promjenama unutar, ali i izvan poduzeća. Kada su procesi u poduzeću jasno definirani može se i kvalitetno upravljati poduzećem.

2.7.1 Modeliranje poslovnih procesa

Poslovni proces se obično prikazuju pomoću različitih dijagrama poslovnih procesa. Jedna od opće prihvaćenih metoda dokumentiranja poslovnih procesa je pomoću BPMN norme. Modeliranje poslovnih procesa odnosi se na izradu dijagrama na kojem se prikazuju aktivnosti i slijed njihova događanja u poslovnom procesu. Na tržištu postoje različiti softverski alati za grafičko modeliranje poslovnih procesa koji se oslanjaju na BPMN normu. BPMN norma je grafički standard za modeliranje poslovnih procesa koja omogućuje definiranje poslovnih procesa od početka do kraja. BPMN norma je napravljena na način da bude razumljiva svima, od projektnih menadžera koji vode projekte, poslovnih analitičara, te razvojnih inženjera koji trebaju razviti rješenje. Razvoj dijagrama koristeći BPMN normu omogućit će svima

sudionicima razumijevanje procesa, te će smanjiti mogućnost nesporazuma između osoba uključenih u proces.



Slika 12 BPMN dijagram izrađen u ARIS express-u (Mesarić, Šebalj, 2019)

Na slici 12 prikazan je BPMN dijagram izrađen u ARIS express-u koji sadrži elemente koji se mogu koristiti za prikaz poslovnih procesa u poduzeću. BPMN norma sastoji od sljedećih elemenata (Mesarić, Šebalj, 2019):

- Plivaće staze
  - Bazeni – grafički kontejner koji grupira skup aktivnosti za jednog sudionika
  - Staze – particija unutar bazena, a koristi se za grupiranje aktivnosti sudionika
- Objekti tokova
  - Događaji – predstavljeni kružićem, a odnosi se na neki događaj u procesu
  - Aktivnosti – predstavljeni zaobljenim pravokutnikom, a odnosi se na zadatak i pod proces
  - Prolazi – predstavljeni rombom, a predstavljaju točku odluke
- Povezujući objekti
  - Sekvencijski tok – predstavljen punom strelicom, a koristi se za prikaz slijeda aktivnosti u procesu
  - Tok poruka – predstavljen crtkanom strelicom sa kružićem na početku i strelicom na kraju, a koristi se za prikaz toka poruka između sudionika u procesu

    - Asocijacija – predstavljena točkastom linijom s otvorenom strelicom, a koristi se za pridruživanje artefakata objektima tokova
    - Poruka – koristi se za određivanje komunikacije među sudionicima u procesu
- Artefakti
    - Grupa – koristi se za grupiranje aktivnosti
    - Zabilješka – koristi se za opis dijela procesa
- Podaci
    - Podatkovni objekt – prikazuje dokumente koji se isporučuju iz procesa ili aktivnosti
    - Podatkovni ulazi – prikazuje ulazne podatke u proces
    - Podatkovni izlazi – prikazuje izlazne podatke iz procesa
    - Podatkovna pohrana – odnosi se na baze podataka u koje se spremaju podaci u procesu

2.7.2 Veliki podaci u poslovnim procesima

Poslovni procesi u organizacijama nisi uvijek standardizirani, neki poslovni procesi su dinamični i zahtijevaju da se brzo prilagođavaju promjenama u okruženju organizacije. Zahvaljujući velikim podacima organizacije mogu otkriti promjene koje će aktivirati i promjene u njihovim poslovnim procesima. Korištenje velikih podataka omogućava organizacijama kontinuirano usavršavanje i prilagođavanje poslovnih procesa koji će im omogućiti konkurentsku prednost i povećanje profita. Organizacije koriste velike podatke kako bi:

- Bolje upravljale otpadom,
- Bolje upravljale kupcima i razumjele njihove potrebe
- Bolje upravljale talentima, te poboljšale procedure prilikom zapošljavanja
- Razvile proizvode koji su potrebni kupcima
- Poboljšale procese u proizvodnji

2.7.3 Kako veliki podaci mogu poboljšati poslovne procese

Analiza podataka odnosi se na pregled i utvrđivanje povezanosti, trendova i obrazaca u velikim skupovima podataka, a u kombinaciji sa poslovnim procesima omogućava poboljšanja unutar organizacije i ostvarivanje konkurentske prednosti, te u konačnici profit. Analizom velikih podataka unutar poslovnih procesa mogu se otkriti mjesta u procesima koja nisu prilagođena

promjenama u okolini. Veliki podaci omogućuju poduzeću da razumije potrošače, povećaju svoju efikasnost, smanje troškove, donose bolje poslovne odluke.

3 Metodologija rada

3.1 Predmet istraživanja

Predmet ovog diplomskog rada je prikazati kako se poslovni procesi mogu poboljšati korištenjem velikih podataka. Fokus se stavlja na alat Hadoop koji se koristi za distribuirano procesiranje velikih podataka i njegove mogućnosti. Pomoću alata Apache Hadoop napravljena je analiza podataka o mjerama koje su uvedene prilikom pojave COVID-19 virusa, te je analiziran njihov utjecaj na poslovne procese u poduzeću.

3.2 Postavljanje i konfiguriranje Apache Hadoop alata

Hadoop alat instaliran je na virtualnoj mašini na kojoj je instaliran operativni sustavu Linux, verzija sustava Ubuntu (64-bit). Za virtualizaciju je korišten alat VirtualBox. U tablici 1 prikazani su resursi koji su dostupni Ubuntu virtualnoj mašini na koju je instaliran Hadoop alat, Radna memorija i prostor na disku se po potrebi mogu povećati.

Tablica 1 Ubuntu virtualna mašina

Operativni sustav	Ubuntu 20.04.2.0 LTS
Prostor na disku	20GB
CPU	Intel(R) Core(TM) i7-8550 CPU @1.80GHz
Radna memorija	8192MB

Prije postavljanja Hadoop alata potrebno je postaviti Javu. Postavljena je verzija Jave *Java SE Development Kit 8* (jdk1.8.0_301.). Prije nego što se postavi Hadoop, potrebno je konfigurirati SSH, kako bi Hadoop mogao upravljati čvorovima na klasteru. Hadoop alat preuzet je sa stranice The Apache Software Foundation, a preuzeta je zadnja stabilna verzija *hadoop-3.3.1-src.tar.gz* koja je raspakirana u datoteku na disku. Konfiguriranje Hadoop-a zahtjeva dodavanje konfiguracije u .bashrc datoteku koja se nalazi u Home direktoriju:

- Postaviti lokaciju Hadoop direktorija
  #Set HADOOP_HOME
  export HADOOP_HOME=/home/goga/Downloads/hadoop
- Postaviti lokaciju Java direktorija

#Set JAVA_HOME

export JAVA_HOME=/home/goga/Downloads/jdk1.8.0_301

Konfiguriranje HDFS-a zahtjeva dodavanje konfiguracije:

- Java direktorij u hadoop-env.sh datoteku koja se nalazi u Hadoop direktoriju
  export JAVA_HOME=/home/goga/Downloads/jdk1.8.0_301
- Dodavanje svojstava u core-site.xml datoteku
  - hadoop.tmp.dir – roditeljski direktorij za ostale privremene direktorije
  - fs.defaultFS – ime pretpostavljenog datotečnog sustava

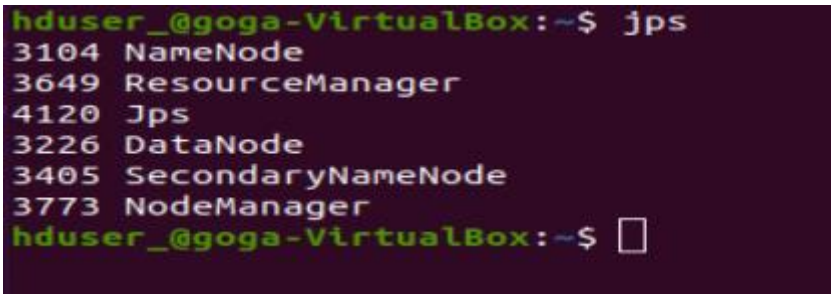
Konfiguriranje MapReduce-a zahtjeva konfiguriranje:

- Dodavanje svojstva u mapred-site.xml datoteku u Hadoop direktoriju
  - mapreduce.jobtracker.address – postavlja se vrijednost porta i poslužitelja
- Dodavanje svojstava u hdfs-site.xml datoteku u Hadoop direktoriju
  - dfs.replication
  - dfs.datanode.data.dir

Nakon što su dodane sve konfiguracije, a prije prvog pokretanja Hadoop-a potrebno je formatirati HDFS. Formatiranje HDFS-a pokreće se korištenjem komande *$HADOOP_HOME/bin/hdfs/ namenode -format*. Za pokretanje Hadoop-a koriste se sljedeće komande:

- *$HADOOP_HOME/sbin/start-dfs.sh*
- *$HADOOP_HOME/sbin/start-yarn.sh*

Pokretanjem naredbe *jsp* vidjet ćemo koji su sve servisi pokrenuti. Na slici 13 prikazani su pokrenuti servisi u Hadoop okolini.



Slika 13 Hadoop servisi (vlastita izrada)

Za zaustavljanje Hadoop-a koriste se sljedeće komande:

- *$HADOOP_HOME/sbin/stop-dfs.sh*
- *$HADOOP_HOME/sbin/stop-yarn.sh*

3.3 Analiza ulaznih podataka

Ulazni podaci koji će se koristiti u ovom diplomskim radu javno su dostupni na stranici Microsoft-a, a preuzeti skup podataka *Oxford COVID-19 Government Response Tracker*, sadrži podatke o COVID-19 pandemiji, odnosno informacije o mjerama koje su vlade zemalja poduzele kako bi se suzbila pandemija (Hale i dr., 2020). Set podataka sadrži sljedeće varijable koje će se analizirati koristeći Hadoop:

- Zatvaranje radnih mjesta
    - 0 – nema mjera
    - 1 – preporučuje se zatvaranje ili rad od kuće
    - 2 – zahtjeva zatvaranje ili rad od kuće za neke djelatnosti
    - 3 – zahtjeva zatvaranje ili rad od kuće za sve osim trgovina, zdravstvenih ustanova
- Otkazivanje javnih događaja
    - 0 – nema mjera
    - 1 – preporučuje se otkazivanje
    - 2 – zahtjeva se otkazivanje
- Ograničavanje putovanja unutar zemlje
    - 0 – nema mjera
    - 1 – preporučuje se ne putovati
    - 2 – zabrana putovanja
- Ograničavanje ulaska u zemlju
    - 0 – nema ograničenja
    - 1 – pregledi na granici
    - 2 – karantena za neke zemlje
    - 3 – zabrana za neke zemlje
    - 4 – zabrana za sve zemlje ili zatvaranje granice

Podaci su prikupljeni u periodu između 01.01.2020. godine do 06.08.2021. Podaci se nalaze u polustrukturiranom obliku, odnosno CSV obliku. Na slici 14 prikazan je primjer ulaznih podataka, prvih 5 redova, te 13 kolona.

	countryname	countrycode	date	c1_school_closing	c1_flag	c2_workplace_closing	c2_flag	c3_cancel_public_events	c3_flag	c4_restrictions_on_gatherings	...	h5_investment_in_vaccines	m1_wildcard	confirmedcases
0	Aruba	ABW	2020-01-01	0.0	False	0.0	False	0.0	False	0.0	...	0.0	NaN	NaN
1	Aruba	ABW	2020-01-02	0.0	False	0.0	False	0.0	False	0.0	...	0.0	NaN	NaN
2	Aruba	ABW	2020-01-03	0.0	False	0.0	False	0.0	False	0.0	...	0.0	NaN	NaN
3	Aruba	ABW	2020-01-04	0.0	False	0.0	False	0.0	False	0.0	...	0.0	NaN	NaN
4	Aruba	ABW	2020-01-05	0.0	False	0.0	False	0.0	False	0.0	...	0.0	NaN	NaN

Slika 14 Primjer ulaznih podataka (vlastita izrada)

Na slici 15 prikazana je deskriptivna analiza ulaznih podataka. Deskriptivna analiza ulaznih podataka obuhvaća:

- Ukupni broj podataka
- Aritmetičku sredinu koja opisuje centar distribucije podataka, odnosno koliko u prosjeku iznosi promatrana varijabla
- Standardnu devijaciju koja daje prosječno kvadratno odstupanje od prosjeka
- Minimalnu vrijednost varijabli ulaznih podataka
- Kvartile
    - donji kvartil – 25% podataka ima manju ili istu vrijednost
    - gornji kvartil – 75% podataka ima veću ili istu vrijednost
- Medijan koji dijeli niz podataka na dva jednaka dijela, gdje 50% podataka ima vrijednost medijana i manje od te vrijednosti, a drugih 50% podataka ima vrijednost medijana i više od te vrijednosti
- Maksimalnu vrijednost varijabli ulaznih podataka

	c1_school_closing	c2_workplace_closing	c3_cancel_public_events	c4_restrictions_on_gatherings	c5_close_public_transport	c6_stay_at_home_requirements	c7_restrictions_on_internal_movement
count	176397.000000	176322.000000	176409.000000	176359.000000	176395.000000	176251.000000	176283.000000
mean	1.703181	1.417492	1.353877	2.489734	0.575895	0.982996	0.949388
std	1.106997	0.961944	0.755813	1.591861	0.665080	0.862949	0.828706
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000
50%	2.000000	2.000000	2.000000	3.000000	0.000000	1.000000	1.000000
75%	3.000000	2.000000	2.000000	4.000000	1.000000	2.000000	2.000000
max	3.000000	3.000000	2.000000	4.000000	2.000000	3.000000	2.000000

Slika 15 Deskriptivna analiza ulaznih podataka (vlastita izrada)

4 Opis istraživanja i rezultati istraživanja

4.1 Obrada podataka pomoću MapReduce tehnike procesiranja

Ulazni podaci koji se obrađuju u sklopu ovog istraživanja nalaze se u covid_policy_tracker.csv dokumentu, a koji sadrži podatke o poduzetim mjerama kako bi se spriječilo širenje virusa u svijetu. Varijable koji će se analizirati u ovom istraživanju su:

- c2_workplace_closing
- c3_cancel_public_events
- c7_restrictions_on_internal_movement
- c8_international_travel_controls

Cilj ove analize je pomoću MapReduce tehniku procesiranja vidjeti koje države provode mjere za sprječavanje širenja virusa, te koliko su te mjere restriktivne. Za procesiranje podataka potrebno je napisati sljedeće java datoteke:

- CovidCountryDriver.java
- CovidMapper.java
- CovidCountryReducer.java

Na slici 16 prikazana je datoteka CovidCountryDriver.java koja je odgovorna za pokretanje MapReduce zadatka u Hadoop sustavu. U klasi se definira naziv zadatka, ulazni i izlazni tipovi podataka, te naziv klasa za *mapper* i *reducer*:

- Naziv zadatka: CovidPerCountry
- Ulazni tip podataka: Text
- Izlazni tip podataka: Text
- Naziv mapper-a: CovidCountry.CovidMapper
- Naziv reducer-a: CovidCountry.CovidCountryReducer

Unutar CovidCountryDriver klase slijedeća linija koda pokreće MapReduce zadatak:

JobClient.runJob(job_conf)

```
package CovidCountry;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;

public class CovidCountryDriver {
    public static void main(String[] args) {
        JobClient my_client = new JobClient();

        JobConf job_conf = new JobConf(CovidCountryDriver.class);
        job_conf.setJobName("CovidPerCountry");
        job_conf.setOutputKeyClass(Text.class);
        job_conf.setOutputValueClass(Text.class);
        job_conf.setMapperClass(CovidCountry.CovidMapper.class);
        job_conf.setReducerClass(CovidCountry.CovidCountryReducer.class);
        job_conf.setInputFormat(TextInputFormat.class);
        job_conf.setOutputFormat(TextOutputFormat.class);

        FileInputFormat.setInputPaths(job_conf, new Path(args[0]));
        FileOutputFormat.setOutputPath(job_conf, new Path(args[1]));

        my_client.setConf(job_conf);
        try {
            JobClient.runJob(job_conf);
        } catch (Exception e) {
            e.printStackTrace();
        }
```

Slika 16 CovidCountryDriver.java (vlastita izrada)

Na slici 17 prikazana je datoteka CovidMapper.java koja se sastoji od CovidMapper klase koja nasljeđuje MapReduceBase klasu i implementira Mapper sučelje. Unutar klase nalazi se map metoda koja zaobilazi (*override-a*) map metodu iz bazne klase. Map metoda prihvaća četiri parametra i svaki put kada se poziva map metoda prosljeđuju joj se vrijednosti u obliku ključ-vrijednost (*key-value*). Vrijednost primljena u *value* argumentu se razdvaja po delimiteru ',' u polje stringova. Izlazna vrijednost metode je ključ koji sadrži podatke o nazivu zemlje i vrijednost koja sadrži podatke o vrijednostima koji će se analizirati:

output.collect(new Text(covidData[0]), new Text(covidData[5]));

```
package CovidCountry;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class CovidMapper extends MapReduceBase implements Mapper
<LongWritable, Text, Text, Text> {

        @Override
        public void map(LongWritable key, Text value, OutputCollector
<Text, Text> output, Reporter reporter) throws IOException {

                String valueString = value.toString();
                String[] covidData = valueString.split(",");

                output.collect(new Text(covidData[0]),new
Text(covidData[5]));
        }
}
```

Slika 17 CovidMapper.java (vlastita izrada)

Na slici 18 prikazana je datoteka CovidCountryReducer.java koja se sastoji od CovidCountryReducer klase koja nasljeđuje MapReduceBase klasu i implementira Reduce sučelje. Unutar klase nalazi se reduce metoda koja zaobilazi (*override-a*) reduce metodu iz bazne klase. Metoda reduce ima dva ulazna parametra tipa *Text* i u metodu se prosljeđuju u obliku ključ-vrijednost (key-value), a prosljeđuju se iz klase CovidMapper.class kao <key, value> koji su Text tip podatka. Zadnja dva parametra metode reduce su izlazne vrijednosti u obliku ključ-vrijednost (*key-value*) koje generira metoda.

```
package CovidCountry;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;


public class CovidCountryReducer extends MapReduceBase implements
Reducer<Text, Text, Text, Double> {

        @Override
        public void reduce(Text t_key, Iterator<Text> values,
OutputCollector<Text, Double> output, Reporter reporter) throws IOException
{
                Text key = t_key;

                int noMeasures = 0;//0.0
                int recomendClosing = 0;//1.0
                int requireClosing = 0;//2.0
                int requireClosingAll = 0;//3.0
                int noValue = 0;

                LinkedList<String> valuesList = new LinkedList<String>();
```

Slika 18 CovidCountryReducer.java (vlastita izrada)

Na slici 19 prikazan je izlaz koji generira metoda reduce i koji je oblika ključ-vrijednost (*key-value*).

```
                double sum = noMeasures + recomendClosing + requireClosing
+ requireClosingAll;
                double percentageNoMeasures = noMeasures / sum * 100;
                double percentageRecomendClosing = recomendClosing / sum *
100;
                double percentageRequireClosing = requireClosing / sum *
100;
                double percentageRequireClosingAll = requireClosingAll /
sum * 100;
                double percentageNoValue = noValue / sum * 100;


                output.collect(key, percentageNoMeasures);
        }
}
```

Slika 19 CovidCountryReducer output (vlastita izrada)

Java datoteke potom se kompajliraju kako bi se mogle koristiti tijekom obrade. Kompajliranje java datoteka izvršava se pokretanjem naredbe u terminalu:

javac -d . CovidMapper.java CovidCountryReducer.java CovidCountryDriver.java

Kompajliranje java datoteka kreirat će tri nove datoteke u novom direktoriju s ekstenzijom class:

- CovidCountryDriver.class
- CovidMapper.class
- CovidCountryReducer.class

Nakon kompajliranja kreirana je Manifest.txt datoteka koja sadrži naziv glavne klase, CovidCountry.CovidCountryDriver od kojeg polazi procesiranje i koji sadrži podatke o zadatku koji treba izvršiti. Nakon toga kreirana je izvršna *jar* datoteka pomoću naredbe:

jar cfm CovidPerCountry.jar Manifest.txt CovidCountry/.class*

Kako bi se podaci mogli procesirati na Hadoop sustavu potrebno je csv datoteku kopirati u HDFS, distribuirani datotečni sustav:

$HADOOP_HOME/bin/hdfs dfs -copyFromLocal ~/inputMapReduce /

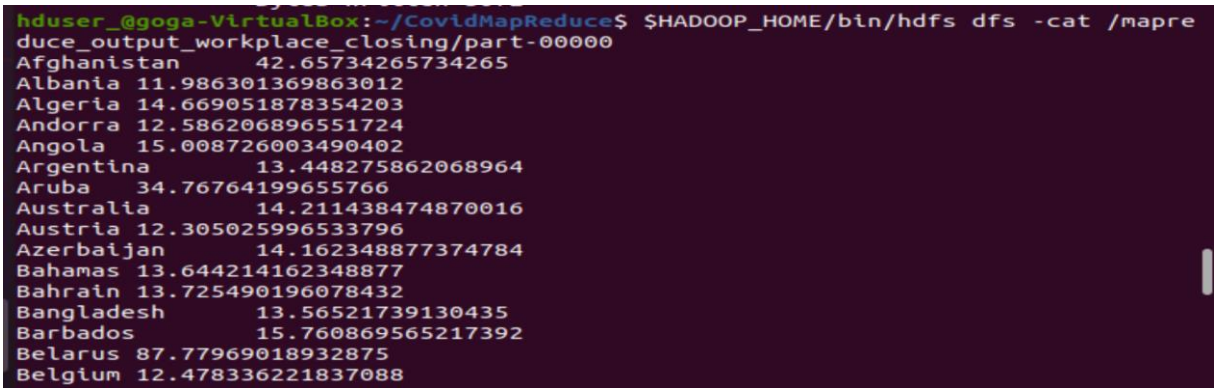
MapReduce posao pokreče se pomoću naredbe kojoj se prosljeđuje izvršna jar datoteka i lokacija csv datoteke u HDFS-u.

$HADOOP_HOME/bin/hadoop jar CovidPerCountry.jar /inputMapReduce /mapreduce_output_workplace_closing

Kada je procesiranje gotovo kreiran je direktorij naziva mapreduce_output_1 sa dokumentima koji sadrže analizirane podatke. Podacima se može pristupiti spajanjem na http://localhost:9870 iz web preglednika ili preko naredbe:

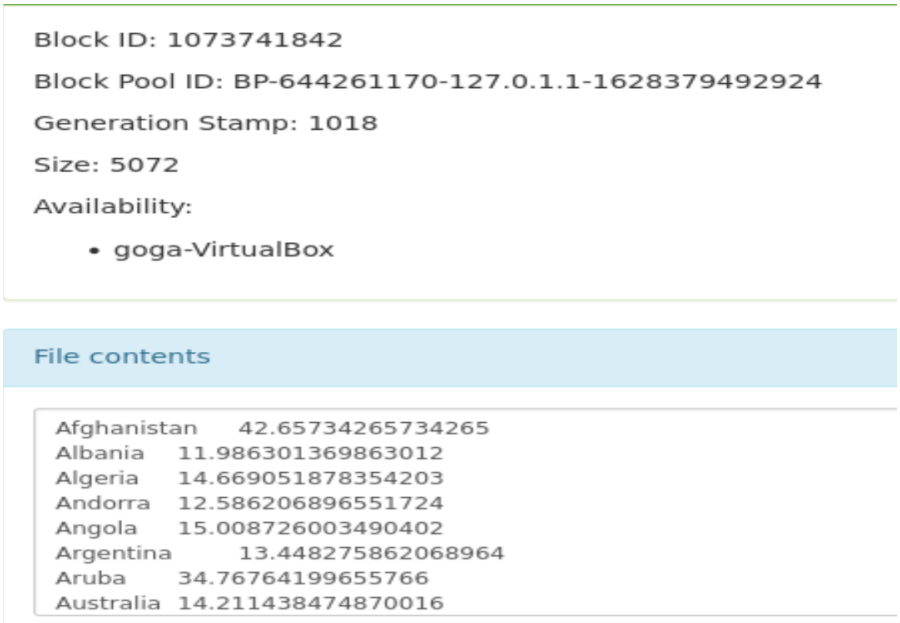
$HADOOP_HOME/bin/hdfs dfs -cat /mapreduce_output_workplace_closing/part-00000

Output MapReduce procesiranja koji je dobiven izvršavanjem naredbe prikazan je na slici 20.



Slika 20 MapReduce output preko naredbe (vlastita izrada)

Na slici 21 prikazan je output MapReduce procesiranja preko web preglednika.

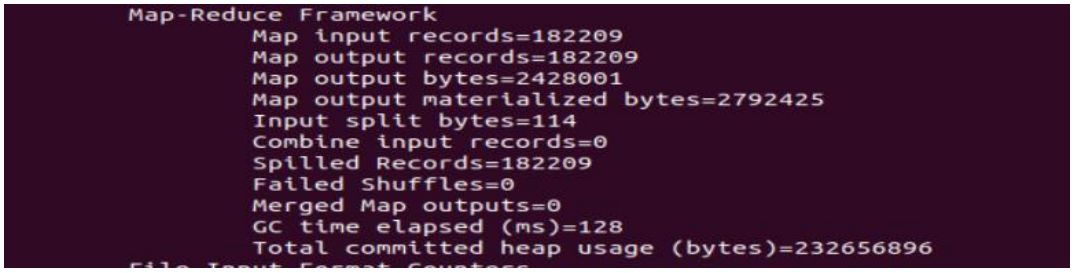


Slika 21 MapReduce output iz web preglednika (vlastita izrada)

4.2 Analiza rezultata

Datoteka, covid_policy_tracker.csv, koja je prebačena na HDFS datotečni sustav smještena je na jednom datotečnom bloku od 128 MB obzirom da je njena 33MB. Nakon što je pokrenut zadatak za procesiranje datoteka je podijeljena u logičke dijelove, a za svaki logički dio dodjeljuje se jedan *mapper*, odnosno broj poslova koji će obrađivati map metodu. U slučaju koji se analizira je jedan logički dio kojem je dodijeljen jedan *mapper*, za učitanu količinu podataka dovoljan je samo jedan posao koji je obradio map dio procesiranja. Na slici 22 prikazana je statistika nakon map procesiranja, gdje je:

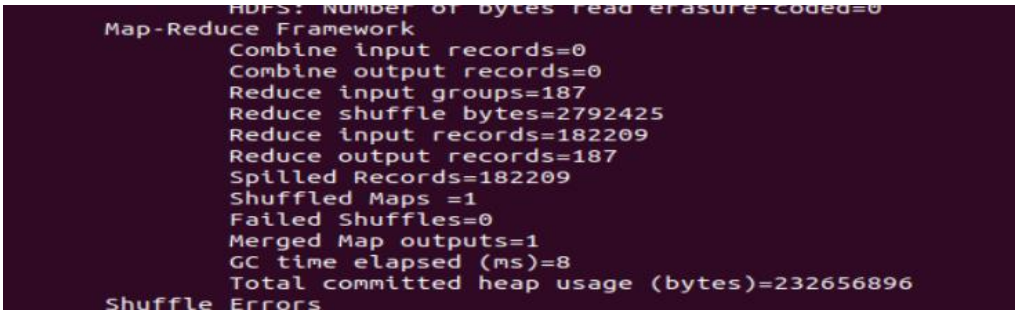
- Map input records – odnosi se na broj vrijednosti procesiranih u map dijelu, a kojih je 182209.
- Map output records – odnosi se na broj izlaznih vrijednosti iz map procesiranja
- Map output materialized bytes – odnosi se na veličinu izlaznih podataka u bajtima zapisanih na disku, 2792425 bajta (2.6631 MB).



Slika 22 Map statistika (vlastita izrada)

Nakon map dijela slijedi reduce dio, koji prima podatke u obliku ključ-vrijednost od *mapper-a*. Na slici 23 prikazana je statistika reduce procesiranja gdje je:

- Reduce input groups – odnosi se na broj ključeva (*key*), odnosno broj jedinstvenih vrijednosti proslijeđenih iz map dijela procesiranja kojih je 187. Za svaki ključ pokrenut je jedan reduce.
- Reduce input records – odnosi se na broj vrijednosti (*value*) proslijeđenih iz map dijela procesiranja kojih je 182209
- Reduce output records – odnosi se na broj izlaznih vrijednosti



Slika 23 Reduce statistika (vlastita izrada)

Koristeći distribuirano procesiranje napravljena je analiza mjera koje su zemlje provodile tijekom pandemije. Mjere koje su se provodile imale su veliki utjecaj na poslovne subjekte i njihove poslovne procese, te su se brzo morali prilagoditi novonastaloj situaciji kako bi mogli uspješno poslovati. U tablici 2 prikazani su rezultati analize za mjeru c2_workplace_closing. Prema rezultatima analize za zemlje sa područja bivše Jugoslavije najviše se je tražilo da se zatvore radna mjesta za neke sektore ili kategorije radnika ili da se uvede rad od kuće. Nakon toga slijedi preporuka za zatvaranje radnih mjesta ili uvođenje rada od kuće. Pretpostavka je da su se mjere mijenjale kako se je mijenjao i broj zaraženih u zemljama pa prema tome u periodu kada je broj zaraženih pao došlo je i do promjene mjera koje su zemlje provodile, a u nekom periodu čak i ukinule.

Tablica 2 c2_workplace_closing

Zemlja	**0.0 nema mjera**	**1.0 preporučuje se zatvaranje ili rad od kuće**	**2.0 zatvaranje ili rad od kuće (neke djelatnosti)**	**3.0 zatvaranje ili rad od kuće (osim nužnih djelatnosti)**	**Nema podataka**
Albanija	11.97%	38.70%	49.32%	0%	0%
Bosna i Hercegovina	13.26%	34.73%	42.41%	9.60%	1.92%
Hrvatska	25.47%	24.44%	43.55%	6.54%	0.52%
Kosovo	12.69%	18.70%	67.58%	1.03%	0.17%
Srbija	18.37%	10.23%	66.03%	5.37%	1.21%
Slovenija	29.50%	7.37%	56.78%	6.35%	0.17%

U tablici 3 prikazani su rezultati analize za mjeru c3_cancel_public_events. Prema rezultatima analize u zemljama sa područja bivše Jugoslavije najviše se ukidaju javna događanja, osim u slučaju Hrvatske gdje je to samo preporuka i provodila se je 65% vremena. Mjere se mijenjaju ovisno o promjeni broja zaraženih, ali i prilagodbi organizatora javnih događanja. Javni događaji su preko 70% vremena bili otkazani u Albaniji, Srbiji i Sloveniji.

Tablica 3 c3_cancel_public_events

Zemlja	**0.0 nema mjera**	**1.0 preporučuje se otkazivanje**	**2.0 otkazivanje**	**Nema podataka**
Albanija	11.64%	10.45%	77.91%	0%
Bosna i Hercegovina	12.04%	39.79%	48.17%	1.92%
Hrvatska	22.20%	64.54%	13.25%	0.52%
Kosovo	12.01%	39.79%	48.20%	0.17%
Srbija	12.82%	16.12%	71.06%	1.21%
Slovenija	13.38%	13.89%	72.73%	0.17%

U tablici 4 prikazani su rezultati analize za mjeru c7_restrictions_on_internal_movement. Prema rezultatima analize u zemljama sa područja bivše Jugoslavije mjera se je najmanje provodila, odnosno najveći dio vremena je bilo dozvoljeno kretati se unutar zemlje. Prema napravljenoj analizi, postojao je period kada je putovanje unutar zemlje bilo zabranjeno, u Kosovu čak 47% vremena, u Sloveniji 35% vremena, dok je u drugim zemljama taj postotak oko 20% i manji. U Srbiji je u 42% vremena preporuka bila da se ne putuje unutar zemlje.

Tablica 4 c7_restrictions_on_internal_movement

Zemlja	**0.0 nema mjera**	**1.0 ne preporučuje se putovati**	**2.0 zabrana putovanja**	**Nema podataka**
Albanija	76.71%	0%	23.29%	0%
Bosna i Hercegovina	72.43%	15.36%	12.22%	1.92%
Hrvatska	86.23%	0%	13.77%	0.52%

Kosovo	53.17%	0%	46.83%	0.17%
Srbija	47.83%	41.77%	10.40%	1.21%
Slovenija	59.35%	5.49%	35.16%	0.17%

U tablici 5 prikazani su rezultati analize za mjeru c8_international_travel_controls. Prema rezultatima analize u zemljama sa područja bivše Jugoslavije najviše su se uvele kontrole na granici jer je za prelazak granice bio potreban test na COVID-19, potvrda o cijepljenju ili preboljenju COVID-19 virusa. Hrvatska je najveći dio vremena imala zabranu ulaska u zemlju za neke zemlje. Mjere su se mijenjale ovisno o broju zaraženih pa je i bilo perioda kada su zemlje potpuno zatvorile svoje granice.

Tablica 5 c8_international_travel_controls

Zemlja	**0.0 nema ograničenja**	**1.0 kontrole na granici**	**2.0 karantena za neke zemlje**	**3.0 Zabrana za neke zemlje**	**4.0 Zabrana za sve zemlje (zatvaranje granice)**
Albanija	25.88%	32.88%	13.81%	27.43%	13.62%
Bosna i Hercegovina	18.66%	58.51%	0%	22.83%	3.80%
Hrvatska	6.65%	35.62%	11.55%	46.18%	13.70%
Kosovo	14.12%	44.14%	27.24%	14.51%	15.90%
Srbija	20.58%	76.70%	1.75%	0.97%	12.04%
Slovenija	11.84%	12.52%	40.31%	35.33%	0%

5 Rasprava

Poslovni procesi se odnose na niz logički povezanih zadataka i aktivnosti koji se sastoji od zaposlenika, materijala, strojeva, sustava i metoda, a čija je svrha zadovoljenje potreba klijenata za robom ili uslugom. Poduzeća definiraju poslovne procese kako bi opisali kako se nešto radi u poduzeću, te koje je korake potrebno odraditi kako bi se proces završio. Poslovni procesi u poduzeću trebali bi biti fleksibilni i reagirati na promjene unutar i izvan poduzeća. Poslovne procese je potrebno konstantno pratiti i mjeriti kako bi se utvrdila njihova učinkovitost. Različiti vanjski čimbenici mogu utjecati na postavljene poslovne procese u poduzeću i smanjiti njihovu

učinkovitost. Jedan od takvih čimbenika bila je pojava COVID-19 virusa koja je imala utjecaj na poslovanje poduzeća i na njihove poslovne procese koji su se trebali prilagoditi novonastaloj situaciji, odnosno novoj normalnoj situaciji i zaštiti kupce uz istovremeno zadržavanje visoke razine kvalitete. Osim zaštite kupaca, tvrtke su trebale pružiti zaštitu i svojim zaposlenicima. Obzirom na mjere koje su se uvodile u državama, tvrtke su morale identificirati koje poslovne procese trebaju prilagoditi kako bi kupcima i zaposlenicima pružile potrebnu zaštitu.

Koristeći distribuirano procesiranje napravljena je analiza mjera koje su zemlje provodile tijekom pandemije. Mjere koje su se provodile imale su veliki utjecaj na poslovne subjekte i njihove poslovne procese, te su se brzo morali prilagoditi novonastaloj situaciji kako bi mogli uspješno poslovati. Svaka od tih mjera utjecala je na poslovanje poduzeća koja su bila primorana prilagoditi svoje poslovne procese ovisno o mjerama koje su se uvodile, te restrikcijama po svakoj mjeri. Mjere koje su se analizirale u sklopu ovog rada, a odnosile su se na zemlje s područja bivše Jugoslavije su:

- Mjere koje su se provodile za dolazak djelatnika na radna mjesta (c2_workplace_closing)
- Mjere koje su se provodile za javna okupljanja (c3_cancel_public_events)
- Mjere koje su se provodile za kretanja unutar države (c7_restrictions_on_internal_movement)
- Mjere koje su se provodile za putovanja izvan države (c8_international_travel_controls)

Za svaku od navedenih mjera napravljena je analiza kako su se mjere provodile u državama bivše Jugoslavije u periodu između 01.01.2020. godine do 06.08.2021.

Za mjeru c2_workplace_closing, koja se odnosi na postupanje poduzeća za dolazak radnika na radna mjesta, od 42% do 68% vremena djelatnici poduzeća nisu dolazili na radna mjesta, nego im je omogućen rad od kuće ili je poduzeće bilo zatvoreno. Manje od 10% vremena sva poduzeća, osim nužnih djelatnosti, su morala poslati svoje djelatnike da rade od kuće ili zatvoriti poduzeće.

Za mjeru c3_cancel_public_events, koja se odnosi na otkazivanje javnih događanja, preko 70% vremena svi javni događaji su bili otkazani u Albaniji, Srbiji i Sloveniji. U Hrvatskoj su javni događaju bili otkazani 13% vremena, dok je preporuka da se javni događaji otkažu vrijedila 65% vremena.

Za mjeru c7_restrictions_on_internal_movement, koja se odnosila na kretanje unutar zemlje, u 48% do 86% vremena se nije provodila u svim zemljama bivše Jugoslavije. Prema napravljenoj analizi, postojao je period kada je putovanje unutar zemlje bilo zabranjeno, u Kosovu čak 47%

vremena, u Sloveniji 35% vremena, dok je u drugim zemljama taj postotak oko 20% i manji. U Srbiji je u 42% vremena preporuka bila da se ne putuje unutar zemlje.

Za mjeru c8_international_travel_controls, koja se je odnosila kontrolu putovanja izvan zemlje, uvele su se kontrole na granici, jer je za prelazak granice bio potreban test na COVID-19, potvrda o cijepljenju ili preboljenju COVID-19 virusa, od 13% u Sloveniji do 77% u Srbiji. Hrvatska je 47% vremena imala zabranu ulaska u zemlju za neke zemlje. Slovenija je 40% vremena uvela obaveznu karantenu za neke zemlje. Mjere su se mijenjale ovisno o broju zaraženih pa su u jednom periodu zemlje potpuno zatvorile svoje granice.

Mjere koje su se provodile u analiziranom periodu značajno su utjecale na poslovanje poduzeća i njihove poslovne procese. U ovom periodu poduzeća su morala prilagoditi poslovne procese kako bi opstala na tržištu. Prema rezultatima analize, mjere koje su se provodile utjecale su na smanjenje interakcije među ljudima što je rezultiralo time da su se tvrtke morale prilagoditi i pronaći nove načine kako ponovno ostvariti komunikaciju sa potrošačima, a da ih pri tome ne ugroze. Tvrtke su unutar svojih poslovnih procesa trebale identificirati područja koja zahtijevaju fizičku interakciju sa kupcem i zamijeniti sa beskontaktnom. Poslovni procesi koji su se trebali mijenjati bili su vezani za komunikaciju, plaćanje, trebali su se mijenjati u komunikaciji, plaćanju, transportu, te raznim uslugama za kupce. Obzirom da je interakcija među ljudima bila ograničena, komunikacija je postala beskontaktna preko aplikacija za slanje direktnih poruka, društvenih mreža. Promjene su se dogodile i prilikom plaćanja, gdje su se kupci ohrabrivali da ne koriste gotovinu, nego kartice. Restorani omogućuju naručivanje i plaćanje proizvoda putem interneta i mobilnih aplikacija sa dostavom na kućna vrata.

Novo normalna situacija koja je pogodila cijeli svijet, nagnala je poduzeća da prilagode dosadašnje načine poslovanja i pruže svojim korisnicima nove vrijednosti. Poduzeća trebaju pratiti aktivnosti u svojim poslovnim procesima, analizirati ih i kontinuirano tražiti prostore za poboljšanja.

5.1 Prednosti i nedostatci distribuiranog procesiranja velikih podataka

Volumen podataka koji se prikuplja je ogroman i svakodnevno eksponencijalno raste. Ti podaci mogu se pretvoriti u znanje i stvaranje konkurentske prednosti. Analizom velikih podataka dobivaju se uvidi koji mogu pomoći:

- menadžerima da donesu bolje poslovne odluke i time povećaju profit poduzeća
- da se poboljšaju i optimiziraju poslovni procesi u poduzeću
- da se poveća efikasnost, te smanje troškovi

- razumjeti ponašanje kupaca

Obzirom na volumen velikih podataka koji se koriste, distribuirano procesiranje velikih podataka ima i svoje nedostatke:

- tvrtke trebaju osigurati dovoljno hardverskih resursa što utječe na povećanje troškova Jedan od problema je i
- potrebno je obratiti pažnju na kvalitetu podataka koji mogu biti netočni, te dovesti do krivi zaključaka i loših poslovnih odluka
- nedostatak stručnih ljudi na tržištu rada

5.2 Mogućnost implementacije novih funkcionalnosti

U ovom radu napravljena je analiza utjecaja vanjskim čimbenika koji utječu na poslovne procese koristeći distribuirano procesiranje velikih podataka pomoću alata Hadoop. Daljnje mogućnost za implementaciju novih funkcionalnosti mogle bi ići u smjeru analize podataka iz samih poslovnih procesa i otkrivanja točaka u poslovnim procesima koji se mogu optimizirati. Obzirom na podatke koji se svakodnevno generiraju i mijenjaju moglo bi se analizirati da li postoje mjesta u poslovnim procesima zahtijevaju daljnje prilagodbe kako bi poslovni procesi bili što otporniji na utjecaje iz okoline.

5.3 Mogućnosti drugih alata za distribuirano procesiranje velikih podataka

Apache Spark je još jedan alat otvorenog koda koji se koristi za distribuirano procesiranje velikih podataka. Za razliku od Apache Hadoop-a koji podatke za obradu drži, dohvaća i obrađuje koristeći stalnu memoriju, odnosno tvrdio disk, Apache Spark koristi radnu memoriju za procesiranje velikih podataka, a podaci se zapisuju na disk tek nakon što je procesiranje podataka završeno. Apache Spark daje puno bolje performanse, što se tiče dohvaćanja i obrade podataka, u slučajevima kada računalo ima dovoljno memorije da u se u njega učitaju podaci. Osim što omogućuje veliku brzinu obrade i dohvaćanja podataka, Apache Spark je prilično lak za korištenje, te proširiv. Obzirom da je Apache Spark alat koji se koristi za obradu podataka, on koristi distribuirane sustave kao što su HDFS i Cassandra za pohranu podataka.

Apache Storm je još jedan alat otvorenog koda koji se koristi za procesiranje toka podataka u stvarnom vremenu. Procesiranje toka podataka odnosi se na konstantno dodavanje novih podataka u procesiranje kako bi se nad njima vršile analize i dobivali novi rezultati. Apache Storm je otporan na kvarove pa u slučaju da se dogodi kvar na jednom čvoru, zadaci se automatski prenose na drugi čvor.

5.4 Budućnost velikih podataka u poslovnim procesima

Analiza velikih podataka pomaže poslovnim subjektima da uoče mjesta u poslovnim procesima koja se nisu prilagođena promjenama u okolini. Poslovni procesi koji se mogu brzo prilagoditi promjenama u okolini omogućit će stvaranje dodatne vrijednosti i pomoći poslovnim subjektima da ostvare konkurentsku prednost i generiranje većih profita. Prema istraživanjima sve veći broj poslovnih subjekata u Europi se bavi analizom velikih podataka, unutar tvrtke ili angažiranjem vanjskih suradnika. Korištenje velikih podataka smatra se ključem uspjeha, a oni koji prvi otkriju znanja koja skrivaju veliki podaci ostvarit će i veće profite. Konstantan tehnološki napredak utječe na poslovanje poduzeća i motivira ih da se konstantno prilagođavaju. Nove tehnologije omogućavaju proizvodima komunikaciju s drugim uređajima, prikupljanje i slanje podataka, što dodatno povećava količinu novih podataka koji se mogu analizirati. Prema tome, tvrtke bi u budućnosti mogle sve više koristiti velike podatke kako bi poboljšale svoje poslovne procese.

6 Zaključak

Razvojem informacijske tehnologije količina generiranih podataka eksponencijalno raste. Ti podaci sadrže skrivena znanja koja se koriste kako bi se razumjela ponašanja potrošača, donijele bolje poslovne odluke, optimizirali poslovni procesi, povećala produktivnost, smanjili troškovi, te ostvarila konkurentska prednost i veći profiti.

Jedan od načina kako otkriti znanje iz velikih podataka je distribuirano procesiranje pomoću alata Hadoop i njegove dvije glavne komponente su HDFS i MapReduce. Apache Hadoop je alat otvorenog koda koji omogućava paralelno procesiranje skupa podataka, a osim Apache Hadoop-a na tržištu su dostupni i drugi alati koji se mogu koristiti za distribuirano procesiranje velikih podataka. Tako napravljena analiza podataka može se primijeniti za optimizaciju poslovnih procesa koji su dinamični i zahtijevaju da se brzo prilagođavaju promjenama u okruženju organizacije. Koristeći analize iz velikih podataka, organizacije mogu otkriti promjene koje će aktivirati i promjene u njihovim poslovnim procesima.

Veliki utjecaj na poslovne subjekte i njihove poslovne procese imala je pojava COVID-19 virusa koja je paralizirala cijeli svijet. Poslovni subjekti koji su svoje poslovne procese uspjeli brzo prilagoditi novim promjenama su nastavili poslovati, dok ih je dosta propalo jer se nisu uspjeli dovoljno brzo prilagoditi. Stoga je vrlo važno pratiti promjene i imati mogućnost brzo prilagoditi svoje poslovne procese. Iako je COVID-19 virus tema koja je stara malo više od

godinu dana može se naći priličan broj istraživanja koja obrađuju njegov utjecaj na poslovanja poduzeća i kako su se ona prilagodila novonastaloj situaciji, te zaštitile svoje kupce i zaposlenike. Poslovni subjekti koji su se brzo reagirali i prilagođavali se obzirom na promjene koje su se događale i koje se još uvijek događaju nastavile su s radom i ostvarile konkurentsku prednost. Zaključak koji se može izvući iz novonastale situacije je da poslovni procesi u poduzeću trebaju imati mogućnost brzog prilagođavanja na promjene u okruženju.

Literatura

Apache Hadoop, 2021. *Apache Hadoop YARN*. [Online]. Available at: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html. [pristupljeno: 27.06.2021.].

Borthakur, D., 2021. *HDFS Architecture Guide*. [Online]. Available at: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. [pristupljeno: 27.06.2021.].

Cooper, S., 2018. *Data Science from Scratch.* [Online]. Available at: https://www.scribd.com/read/385898411/Data-Science-from-Scratch-The-1-Data-Science-Guide-for-Everything-A-Data-Scientist-Needs-to-Know-Python-Linear-Algebra-Statistics-Coding-Applic. [pristupljeno: 07.06.2021.].

Europska komisija, 2021. *Europska komisija.* [Online]. Available at: https://ec.europa.eu/croatia/basic/everything_you_need_to_know_about_big_data_technology_hr. [pristupljeno: 21.06.2021.].

Hale, T., Webster, S., Petherick, A., Phillips, T., Kira, B., 2020. *Oxford COVID-19 Government Response Tracker*. [Online]. Available at: https://docs.microsoft.com/en-us/azure/open-datasets/dataset-oxford-covid-government-response-tracker?tabs=azure-storage. [pristupljeno: 21.06.2021.].

IIBA, 2021. *Guide to Business Data Analytics.* [Online]. Available at: https://www.scribd.com/read/507022693/Guide-to-Business-Data-Analytics. [pristupljeno: 21.06.2021.].

Kotu, V., Deshpande, B., 2018. *Data Science: Concepts and Practice.* [Online]. Available at: https://www.scribd.com/read/394443557/Data-Science-Concepts-and-Practice. [pristupljeno: 07.06.2021.].

Madbouly, M., Al-falluji, R., 2019. *Researchgate.* [Online]. Available at: https://www.researchgate.net/publication/332120693_Chapter_1_Data_Mining_A_First_View. [pristupljeno: 18.06.2021.].

Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. [Online]. Available at: https://www.scribd.com/book/249309568/Big-Data-A-Revolution-That-Will-Transform-How-We-Live-Work-and-Think. [pristupljeno: 21.06.2021.].

Mesarić, J., Šebalj, D., 2019. *Merlin.* [Online]. Available at: https://moodle.srce.hr/2020-2021/mod/resource/view.php?id=1726867. [pristupljeno: 29.07.2021.].

Oracle Corporation, n.d. *Oracle VM VirtualBox.* [Online]. Available at: https://www.virtualbox.org/manual/UserManual.html. [pristupljeno: 24.08.2021.].

Shrivastava, A., Tanmay, D., 2016. *Hadoop Blueprints.* [Online]. Available at: https://www.scribd.com/read/365184972/Hadoop-Blueprints. [pristupljeno: 22.06.2021.].

The Enlightened DBA, 2014. *DBA's Guide to NoSQL*. [Online]. Available at: https://www.scribd.com/book/241642040/DBA-s-Guide-to-NoSQL. [pristupljeno: 21.06.2021.].

von Rosing, M., von Scheel, H., Scheer, A.-W., 2014. *The Complete Business Process Handbook: Body of Knowledge from Process Modeling to BPM, Volume 1*. [Online]. Available at: https://www.scribd.com/book/282650848/The-Complete-Business-Process-Handbook-Body-of-Knowledge-from-Process-Modeling-to-BPM-Volume-1.
[pristupljeno: 26.07.2021.].

Popis slika

Popis tablica

Prilozi

Naziv dokumenta	Dokument
Specifikacija podataka	 Oxford COVID-19 Government Respons
Ulazni podaci	 covid_policy_tracker.c sv