HADOOP Ekosistem: Analiza alata i primena platforme u industriji
HADOOP Ecosystem: Components and Applications

Marija Tošić[1], Aleksandra Tašković[1], Valentina Janev[2]
Univerzitet Metropolitan[1]
Institut Mihajlo Pupin, Univerzitet u Beogradu[2]

Sadržaj – *Rad opisuje softversko okruženje za obradu velikih količina podataka Apache Hadoop, njegove komponente koji upotpunjavaju i proširuju mogućnosti Hadoop-a, kao i njegovo korištenje u praksi. Da bismo u potpunosti razumeli prednosti Hadoop-a, potrebno je sagledati najpre razliku između paralelnog i distribuiranog računanja, načine čuvanja (HDFS arhitekturu), upravljanje resursima i procesiranja podataka, princip horizontalne skalabilnosti, itd. Rad dalje analizira prednosti i nedostaci, kao i primenu Hadoop ekosistema u različitim industrijama (finansijski sektor, maloprodajni sektor i zdravstvo).*

Abstract - *This paper describes the Hadoop software environment for processing big data, its components that complement and expand the capabilities of Hadoop, as well as its use in practice. To fully understand the advantages of Hadoop, it is necessary to first consider the difference between parallel and distributed computing, storage methods (HDFS architecture), resource management and data processing, the principle of horizontal scalability, etc. The paper further analyses advantages and disadvantages and utilization of Hadoop ecosystems in different various industries including financial sector, retail sector and healtcare.*

1. UVOD

Upravljanje i analiza podataka oduvek je predstavljala najveći izazov za sve organizacije u svim poljima industrije. Brzi razvoj digitalnih tehnologija, računarstvo u oblaku (eng. *Cloud computing*), tehnologije inteligentnih IoT (srp. Internet stvari, eng. *Internet of Things*) uređaja, društvene mreže, video, audio i geolokacione usluga stvorili su mogućnosti za prikupljanje / akumuliranje velike količine podataka. Dok su se u prošlosti korporacije bavile statičnim, centralno uskladištenim podacima, danas organizacije prikupljaju strukturitane i nestrukturirane podatke iz različitih izvora i organizuju njihovu čuvanje i obradu na distribuirani način. Tako, na primer, zahtevi za razvojem naprednih aplikacija koje odlikuje pouzdanost, distribuiranost i skalabilnost, ne mogu se realizovati primenom tradicionalnih baza podataka. Zato se razvijaju novi pristupi za skladištenje, brzu pretragu i analizu velikih količina (eng. *Big data*) podataka u realnom vremenu, zasnovani na *Big data* tehnologijama [1].

Da bi se omogućilo pouzdano i skalabilno skladištenje velikih količina podataka,neophodno je obezbediti čuvanje i upravljanje fajlovima u distribuiranom okruženju. Za to se koriste distribuirani fajl sistemi koji omogućuju jednostavan pristup fajlovima na različitim lokacijama, replikaciju fajlova između servera i kompresiju podataka optimizovanu za transfer kroz mrežu sa ograničenom propusnom moći. Primeri za implementaciju distribuiranih fajl sistema su: *Google File System (GFS), Hadoop distributed file system (HDFS), GlusterFS.*

Cilj ovog rada je analiza softversko okruženje za obradu velikih količina podataka *Apache Hadoop* [2], njegove komponente koji upotpunjavaju i proširuju mogućnosti *Hadoop*-a, kao i njegovo korištenje u praksi. Da bi se razumela potrebe za alatima ekosistema *Hadoop* u poglavlju 2 najpre definišemo pojmove veliki podaci, paralelna obrada i distribuirana obrada. Dalje, poglavlju 3 predstavlja rezultate analiza komponenti (HDFS, YARN i MapReduce), Poglavlje 4 ukazuje na prednosti i slabosti Hadoop-a, dok poglavlju 5 diskutuje njihovu primenu u praksi.

2. BIG DATA KONCEPT

Veliki podaci predstavljaju veliku količinu podataka velike različitosti koja zahtevaju isplative, inovativne oblike informacija za poboljšani uvid i donošenje odluka. Potreba za primenom *Big data* tehnologija često se objašnjava korišćenjem tri „V“ modela, po kome su glavne karakteristike podataka [3]:

- Obim podataka (eng. *Volume*)
- Raznovrsnost podataka (eng. *Variety*)
- Brzina (eng. *Velocity*).

Danas se podaci nalaze u velikom broju različitih formata, npr. tradicionalne baze podataka, tekstualne fajlove, e-mail, video, audio, podatke o finansijskim transakcijama, itd. Prema nekim procenama oko 80 procenata podataka nije numeričkog tipa, ali oni i dalje moraju biti uključeni u procedure analize i donošenja odluka u vezi sa njima. Mnogo faktora doprinosi uvećanju obima podataka. Brzina kojom se generišu veliki podaci zahteva da se podaci takođe vrlo brzo obrade, a raznolikost velikih podataka znači da sadrže različite vrste podataka, uključujući strukturirane, polustrukturirane i nestrukturirane podatke. Obim, brzina i raznolikost velikih podataka zahtevaju nove, inovativne tehnike i okvire za prikupljanje, čuvanje i obradu podataka, zbog čega je stvoren i *Apache Hadoop*.

2.1 Paralelna naspram distribuiranoj obradi podataka

Razumevanje paralelne obrade i distribuirane obrade pomoći [4] će razumeti kako se *Apache Hadoop* koristi u analitici velikih podataka. Budući da i paralelna obrada i distribuirana obrada uključuju razbijanje računanja na

manje delove, može doći do zabune između njih dvoje. Razlika između paralelnog i distribuiranog računanja je u memorijskoj arhitekturi, gde se kod paralelnog računanje istovremeno koristi više od jednog procesora za rešavanje problema, dok se za distribuirano računanje istovremena koriste više računara za rešavanje problema. Zadaci paralelnog računanja pristupaju istom memorijskom prostoru, dok zadaci distribuiranog računara ne mogu, jer se distribuirano računanje zasniva na disku, umesto na memoriji. Neki zadaci raspodeljenog računara rade na jednom računaru, a neki na više njih.

2.2 Ostale karakteristike velikih podataka

Kada govorimo o karakterisitkama, bitno je napomenuti još dve bitne dimenzije [5]:

- Promenljivost: Koliko su podaci podložni promenama. Kao dodatak velikim količinama i brzinama obrade podataka, tok podataka može postati prilično nepravilan sa vremenom. To se može objasniti nekom popularnom pojavom u sredstvima javnog informisanja, gde se jedan isti podatak ponavlja nebrojeno puta. Ovakvi izuzeci su jako teški za obradu, pogotovu kad se uzme u obzir skorašnji rast popularnosti socijalnih mreža.
- Složenost: Koliko su podaci teški za obradu. Kada se bavimo velikim količinama podataka, oni uobičajeno dolaze iz različitih izvora i nalaze se u različitim oblicima. Sa eksplozijom razvoja senzora, pametnih uređaja i socijalnih mreža podaci su postali složeni prvenstveno zato što sada ne uključuju samo tradicionalne struktuirane podatke, već i nestruktuirane ili polustruktuirane podatke.

Struktuirani podaci opisuju podatke koji su grupisani u relacione sheme(redovi i kolone u okviru standardnih baza podataka). Organizacija ovih podataka daje mogućnost izvršavanja jednostavnih upita koji mogu vratiti korisne informacije za poslovanje. Polustruktuirani podaci predstavljaju podatke za koje se ne može redi da su grupisani u neku fiksiranu shemu. Podaci su često nerazdvojivi i sadrže oznake koje pomažu pri hijerarhijskom organizovanju ovakvih podataka.

3. HADOOP EKOSISTEM

Hadoop je softver otvorenog koda (eng. *Open Source)* pod pokroviteljstvom *Apache Software* Fondacije. Hadoop se može slobodno preuzimati, ažurirati i nadograđivati od strane velike zajednice programera i IT konstruktora koji rade na njegovom daljem usavršavanju. Razvoj Hadoop-a je započet u 2006. godini kao projekat kompanije Yahoo, a kompanije koje razvijaju proizvode, derivate Hadoop-a i značajni njegovi distributeri na tržištu su: Cloudera, Hortonworks, IBM, Pentaho, Fico, Jaspersoft, Apache Bigtop, Cascading, Amazon Elastic MapReduce, Azure HDInsight i drugi. HADOOP ima nekoliko komponenti:

- Hadoop sistem distribuiranih datoteka (HDFS, eng. *Hadoop Distributed File System*), koji datoteke skladišti u izvornom formatu Hadoop i paralelizuje ih kroz klaster;
- YARN (eng. *Yet Another Resource Negotiator*), komponenta koja raspoređuje izvođenje aplikacije; i
- MapReduce, algoritam koji zapravo paralelno obrađuje podatke.

Hadoop je izgrađen u Javi i dostupan mu je putem mnogih programskih jezika za pisanje MapReduce koda, uključujući Python. Pored ovih osnovnih komponenti, Hadoop takođe uključuje:

- Sqoop, koji relacione podatke premešta u HDFS,
- Hive- interfejs nalik SQL-u koji omogućava korisnicima da pokreću upite na HDFS-u; i
- Mahout, za mašinsko učenje.

Pored upotrebe HDFS-a za skladištenje datoteka, Hadoop se sada takođe može konfigurisati da koristi S3 segmente ili Azure blobs kao ulaz. Dostupan je ili putem otvorenog koda kroz distribuciju Apache-a ili putem dobavljača kao što je Cloudera (najveći Hadoop dobavljač po veličini i obimu), MapR ili HortonWorks.

3.1 HDFS-Hadoop Distributed File System

Hadoop Distributed File System (Sistem distribuiranih datoteka) predstavlja distribuirano skladište za Hadoop. HDFS ima topologiju master-slave. Master je vrhunska mašina u kojoj su robovi jeftini računari. Datoteke velikih podataka dele se na broj blokova. Hadoop skladišti ove blokove distribuirano na klasteru podređenih čvorova. Na masteru imamo sačuvane metapodatke. HDFS pokreće dva čvora :

- NameNode (čvor imena):
  - Radi na glavnoj mašini.
  - Odgovoran je za održavanje, nadzor i upravljanje DataNode-om (čvorom podataka).
  - Beleži metapodatke datoteka kao što su lokacija blokova, veličina datoteke, dozvola, hijerarhija itd.
  - Beleži sve promene u metapodacima kao što su brisanje, kreiranje i preimenovanje datoteke u evidencijama uređivanja.
  - Redovno prima otkucaje srca i blokira izveštaje od DataNodes-a.

- DataNode (čvor podataka): -
  - DataNode radi na pomoćnom računaru.
  - U njemu se čuvaju stvarni podaci o poslovanju.
  - Služi zahtevu za čitanje i pisanje od korisnika.
  - DataNode vrši osnovni posao stvaranja, umnožavanja i brisanja blokova na naredbi NameNode.
  - Nakon svake 3 sekunde šalje otkucaje srca NameNodeu izveštavajući o stanju HDFS- a.

3.2 MapReduce

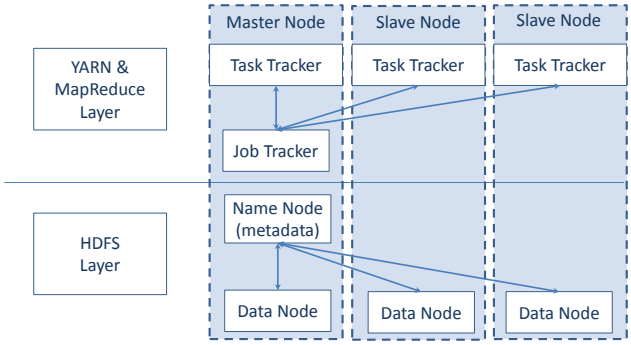
MapReduce je sloj za obradu podataka Hadoop-a. Podatke obrađuje u dve faze:

- Faza mapiranja - Ova faza primenjuje poslovnu logiku na podatke. Ulazni podaci se pretvaraju u parove ključ / vrednost, nešto nalik rečniku podataka.
- Faza primene agregacije (Reduce faza) - uzima kao ulaz izlaz mape faze. Primenjuje agregaciju, odnosno bilo koju računicu, zasnovanu na ključu parova ključ

/ vrednost. Ova primena vraća podatke različite strukture od ulaza.

MapReduce radi na sledeći način:

- Klijent određuje datoteku ili varijablu za unos u funkciju Map.
- Mapa funkcija definiše ključ i vrednost iz ulazne datoteke. Izlaz funkcije mape je ovaj par ključ / vrednost.
- MapReduce framework sortira par ključ / vrednost iz funkcije map.
- Okvir spaja primere koje imaju isti ključ.
- Reduktori dobijaju ove spojene parove ključ / vrednost kao ulaz.
- Reduktor primenjuje agregatne funkcije (sumiranje, oduzimanje, prebrojvanje, množenje...) na par ključ / vrednost.
- Izlaz iz reduktora se zapisuje u HDFS.



Slika 1. Grafički prikaz glavnih komponenti

MapReduce je sloj za obradu podataka Hadoop-a. To je softverski framework koji omogućava pisanje aplikacija za obradu velike količine podataka. MapReduce paralelno pokreće ove aplikacije na klasteru nižih mašina. Posao MapReduce sadrži brojne zadatke mapiranja gde svaki zadatak se izvršava samo na delu podataka. Ovo raspoređuje opterećenje po klasteru. Funkcija zadataka Mape je učitavanje, raščlanjivanje, transformacija i filtriranje podataka. Svaki zadatak smanjenja radi na podskupu rezultata iz zadataka mape. Zadatak ReduceMap primenjuje grupisanje i agregiranje na ove posredne podatke iz zadataka mape, po principu rečnika kao što je ranije objašnjeno. Ulazna datoteka MapReduce procesa postoji na HDFS-u. Ulazni format odlučuje kako podeliti ulaznu datoteku na ulazne podele. Podela unosa nije ništa drugo do bajtno orijentisani prikaz dela ulazne datoteke. Nakon toga se ovi delovi učitavaju od strane procesa mapiranja. Zadatak mapiranja se izvodi na čvoru na kojem su relevantni podaci. Podaci se ne moraju kretati mrežom i obrađivati lokalno.

3.3 YARN

YARN ili *Yet Another Resource Negotiator* je nivo upravljanja resursima i ima sledeće komponente:

- Menadžer Resursa,
- Menadžer čvorova
- Podnosilac posla

Osnovni princip koji stoji iza YARN-a je upravljanja resursima i raspoređivanja, odnosno nadgledanja poslova u zasebnim čvorovima. U YARN postoji jedan globalni ResourceManager i ApplicationMaster po aplikaciji. Unutar YARN okvira imamo dva čvora ResourceManager i NodeManager. ResourceManager arbitrira resurse među svim konkurentskim aplikacijama u sistemu. Posao NodeMangera je nadgledanje upotrebe resursa od strane kontejnera i prijavljivanje istog ResourceMangeru. Resursi su poput CPU-a, memorije, diska, mreže i tako dalje. ApplcationMaster pregovara o resursima sa ResourceManager-om i radi sa NodeManger-om da izvrši i nadgleda posao.

4. PREDNOSTI I SLABOSTI HADOOP-A

4.1 Prednosti Hadoop-a

Hadoop je jednostavan za upotrebu, skalabilan i isplativ. Uz ovo, Hadoop ima mnogo prednosti, kao što su na primer sledeće:

1. Različiti izvori podataka

Hadoop prihvata razne podatke. Podaci mogu poticati iz različitih izvora kao što su razgovori putem e-pošte, društvenih mreža, itd., A mogu biti u strukturiranom ili nestrukturiranom obliku. Hadoop može izvući vrednost iz različitih podataka. Hadoop može da prihvati podatke u tekstualnoj datoteci, XML datoteci, slikama, CSV datotekama itd.

2. Isplativost

Hadoop je ekonomično rešenje jer koristi grupu skromnog hardvera za skladištenje podataka. Ovako skroman hardver je jeftina mašina, pa troškovi dodavanja čvorova u okvir nisu mnogo visoki. U Hadoop-u 3.0 imamo samo 50% dodatnih troškova, za razliku od 200% u Hadoop2.k-u. Ovo zahteva manje mašina za skladištenje podataka jer su se suvišni podaci znatno smanjili.

3. Performanse

Hadoop sa distribuiranom obradom i arhitekturom distribuiranog skladišta obrađuje ogromne količine podataka velikom brzinom. Hadoop je čak i najbržu mašinu pobedio superračunar 2008. godine. Datoteku ulaznih podataka deli na više blokova i podatke u tim blokovima čuva preko nekoliko čvorova. Takođe deli zadatak koji korisnik podnosi na različite pod-zadatke koji dodeljuju ovim radničkim čvorovima koji sadrže potrebne podatke i koji se pod-zadatak paralelno izvode, poboljšavajući tako performanse.

4. Tolerantan na kvarove

U Hadoop-u 3.0 tolerancija grešaka obezbeđuje se brisanjem. Na primer, 6 blokova podataka proizvode 3 bloka pariteta pomoću tehnike kodiranja brisanjem, tako da HDFS skladišti ukupno ovih 9 blokova. U slučaju kvara bilo kojeg čvora, pogođeni blok podataka može se oporaviti korišćenjem ovih delova blokova i preostalih blokova podataka.

5. Visoko dostupno

U Hadoop 2.k, HDFS arhitektura ima jedan aktivni NameNode i jedan pomoćni NameNode, pa ako NameNode padne, onda imamo pomoćni NameNode na koji možemo računati. Ali Hadoop 3.0 podržava višestruki rezervni NameNode, čineći sistem još dostupnijim, jer može nastaviti da funkcioniše u slučaju da dva ili više NameNode padne.

6. Nizak mrežni saobraćaj

U Hadoopu je svaki posao koji je korisnik podneo podeljen na određeni broj nezavisnih podzadataka i ti podzadaci se dodeljuju čvorovima podataka, premeštajući tako malu količinu koda u podatke, umesto da premeštaju ogromne podatke u kod, što dovodi do niski mrežni saobraćaj.

7. Velika propusnost

Propusnost predstavlja količinu posla urađen u jedinici vremena. Hadoop skladišti podatke distribuirano, što omogućava lako korišćenje distribuirane obrade. Dati posao se deli na manje poslove koji paralelno rade na delovima podataka dajući tako veliku propusnost.

8. Open Source

Hadoop je tehnologija otvorenog koda, tj. izvorni kod je dostupan. Izvorni kod možemo modifikovati tako da odgovara određenim zahtevima.

9. Skalabilan

Hadoop radi na principu horizontalne skalabilnosti, odnosno možemo dodati celu mašinu u klaster čvorova i da ne menjamo konfiguraciju mašine poput dodavanja RAM-a, diska i tako dalje, što je poznato kao vertikalna skalabilnost. Čvorovi se mogu dodavati Hadoop klasteru u hodu što ga čini skalabilnim okvirom.

10. Jednostavnost upotrebe

Hadoop framework brine o paralelnoj obradi, programeri MapReduce ne moraju da brinu o postizanju distribuirane obrade, to se automatski vrši na pozadini.

11. Kompatibilnost

Većina nove tehnologije Big Data koja je u nastajanju kompatibilna je sa Hadoop-om poput Spark-a. Oni imaju mehanizme za obradu koji preko Hadoop-a rade kao pozadina, odnosno koristimo ih kao platforme za skladištenje podataka.

12. Podržani više jezika

Programeri mogu da kodiraju koristeći mnoge jezike na Hadoop-u poput C, C ++, Perl, Python, Rubi i Groovy.

4.2 Slabosti Hadoop-a

1. Problem sa malim datotekama

Hadoop je pogodan za mali broj velikih datoteka, ali kada je reč o aplikaciji koja se bavi velikim brojem malih datoteka, Hadoop ovde ne uspeva. Mala datoteka nije ništa drugo do datoteka koja je znatno manja od veličine bloka Hadoop-a koja po podrazumevanoj vrednosti može biti 128 MB ili 256 MB. Ovaj veliki broj malih datoteka preopterećuje NameNode jer čuva prostor imena sistema i otežava funkcionisanje Hadoop-a.

2. Ranjiva po prirodi

Hadoop je napisan u Javi koja je široko korišćeni programski jezik, pa ga zato cyber kriminalci lako koriste, što čini Hadoop ranjivim na bezbednosne provale.

3. Prekovremena obrada podataka

U Hadoop-u se podaci čitaju s diska i zapisuju na disku, što čini radnje čitanja / pisanja veoma skupim kada imamo posla sa tera i petabajtima podataka. Hadoop ne može da vrši proračune u memoriji, zbog čega dolazi prekovremena obrada podataka.

4. Podržava samo serijsku obradu

U osnovi, Hadoop ima mehanizam za serijsku obradu koji nije efikasan u obradi u trenutku. Ne može da proizvede izlaz u realnom vremenu sa malom kašnjenjem. Radi samo na podacima koje prikupimo i sačuvamo u datoteci unapred pre obrade.

5. Iterativna obrada

Hadoop ne može sam da izvrši iterativnu obradu. Mašinsko učenje ili iterativna obrada ima ciklični protok podataka, dok Hadoop ima podatke koji teku u lancu faza, gde izlaz u jednoj fazi postaje ulaz druge faze.

6. Bezbednost

Iz sigurnosnih razloga, Hadoop koristi Kerberos autentifikaciju kojom je teško upravljati. Nedostaje šifrovanje na nivoima skladišta i mreže što je glavna stvar koja zabrinjava.

5. PRIMENA HADOOP PLAFORME U INDUSTRIJI

5.1 Finansijski sektor

Finansijski sektor [6,7] je jedan od glavnih korisnika Hadoop-a. Jedan od primarnih slučajeva upotrebe platforme, bilo je modeliranje rizika, kako bi se rešilo pitanje banki u boljem procenjivanju kupaca i tržišta od starijih sistema.

Hadoop je pomogao finansijskom sektoru da održi bolju evidenciju rizika nakon ekonomskog pada 2008. godine. Pre toga, svaka regionalna filijala banke održavala je zastareli okvir skladišta podataka izolovan od globalnog entiteta. Podaci kao što su provera i čuvanje transakcija, detalji hipoteke na kući, transakcije kreditnim karticama i drugi finansijski detalji svakog klijenta, bili su ograničeni na lokalne sisteme baza podataka, zbog čega banke nisu uspele da oslikaju sveobuhvatan portfolio rizika svojih klijenata.

Nakon ekonomske recesije, većina finansijskih institucija i nacionalnih monetarnih udruženja započele su održavanje jedinstvenog Hadoop klastera, koji sadrži više od petabajta finansijskih podataka prikupljenih iz više sistema preduzeća i stare baze podataka. Zajedno sa objedinjavanjem, banke i finansijske institucije počele su da uvlače i druge izvore podataka - kao što su evidencija poziva klijenata, evidencije ćaskanja i veb stranice, prepiska e-pošte i drugi. Kada se takva neviđena skala podataka analizira uz pomoć Hadoop-a, MapReduce-a i tehnika kao što su analiza sentimenta, obrada teksta, podudaranje obrazaca, kreiranje grafikona; banke su mogle da identifikuju iste klijente iz različitih izvora, zajedno sa tačnom procenom rizika.

Predviđanje ponašanja na tržištu je još jedan klasičan problem u finansijskom sektoru. Različiti tržišni izvori, kao što su berze, banke, odeljenje prihoda i prihoda, tržište hartija od vrednosti - i sami imaju ogroman obim podataka, ali njihov učinak je međuzavisan. Hadoop pruža tehnološku okosnicu stvaranjem platforme na kojoj se podaci iz takvih izvora mogu kompajlirati i obrađivati u realnom vremenu.

5.2 Zdravstvo

Zdravstvena industrija [8] koristi velike podatke za lečenje bolesti, smanjenje medicinskih troškova, predviđanje i upravljanje epidemijama i održavanje kvaliteta ljudskog života, prateći velike zdravstvene indekse i pokazatelje.

Kompleksnost i obim zdravstvenih podataka primarna je pokretačka snaga prelaska sa starih sistema na Hadoop u zdravstvenoj industriji. Korišćenje Hadoop-a na takvoj skali podataka, pomaže u lakom i brzom predstavljanju podataka, dizajniranju baze podataka, kliničkoj analizi odluka, ispitivanju podataka i toleranciji grešaka.

Hadoop kao sistem baza podataka omogućava skladištenje nestrukturiranih zdravstvenih podataka u svom izvornom obliku. Za milijarde medicinskih kartona, Hadoop pruža neograničenu paralelnu obradu podataka, toleranciju grešaka i skladištenje za veliku količinu nestrukturiranih skupova podataka. Budući da HDFS1 i MapReduce imaju mogućnost obrade terabajta podataka, čini Hadoop nezamenljivim za probleme velikih podataka u zdravstvenom sektoru.

Kao studiju slučaja, navešćemo kompaniju, za zdravstvenu informacionu tehnologiju, koja je bila dužna da sačuva istorijske zahteve i doznake u proteklih sedam godina. Korišćenje tradicionalnog sistema baza podataka za čuvanje ovih podataka stvorilo je probleme u zadržavanju podataka tokom obrade miliona zahteva dnevno. Stvorili su rešenje dozvoljavajući sistemu Hadoop da skladišti masovne podatke. Kompleksni procesi normalizacije podataka, evidentiranja terabajta informacija i traženja podataka za analitiku su nakon toga nesmetano obavljeni na Hadoop sistemu koji je razvio Cloudera.

Big Data i Hadoop tehnologije se takođe primenjuju u poslovima zdravstvenog osiguranja. Korišćenje distribuiranog sistema baza podataka u zdravstvenim obaveštajnim aplikacijama, pomaže osiguravajućim društvima, bolnicama i korisnicima da povećaju svoju produktivnost, kreiranjem inovativnih poslovnih rešenja. Na primer, ako kompanija za zdravstveno osiguranje treba da proceni opštu starost stanovništva ispod koje pojedinci nisu podložni određenim bolestima, mogu stvoriti profitabilne politike pogodne za različite delove populacije. Za takvu procenu potrebna je obrada ogromnih količina podataka, uključujući lekove, geografske regije, evidenciju nege pacijenata, bolesti, simptome itd. Hadoop i MapReduce pokazaće se ekonomičnim u obradi tako masovnih nestrukturiranih informacija.

Cilj upotrebe Hadoop-a u zdravstvu je skladištenje i analiza medicinskih podataka koji se mogu iskoristiti za procenu trendova javnog zdravstva sa populacijom od milijarde ljudi, kao i namerno stvaranje mogućnosti lečenja za pojedinačne pacijente prema njihovim potrebama.

5.3 Maloprodajni sektor

Bilo koji trgovac na veliko, koji radi analizu podataka o transakcijama, mora da sastavi ogromne količine podataka o transakcijama na prodajnom mestu, koji dolaze iz različitih izvora podataka, sa ciljem predviđanja potražnje, povećanja dobiti i stvaranja ciljanih marketinških i promotivnih kampanja. Maloprodajna analitika [9] je jedan od glavnih potrošača industrije skladištenja podataka i takođe je odgovorna za njene inovacije i rast. To je zbog mogućnosti prikupljanja i skladištenja daleko više podataka o potrošačima, njihovom ponašanju i potrošnji - kako na mreži tako i u prodavnicama. Istorijski zapisi o prodaji, uz pomoć Hadoop-a i MapReduce-a, koriste se za povećanje marže profita i prodaje.

Novi podaci koje trgovci danas generišu, zahtevaju sofisticiranu obradu poput obrade jezika, prepoznavanja obrazaca, analize raspoloženja itd. Stoga tradicionalni sistemi za upravljanje bazama podataka više nisu isplativa platforma za čuvanje složenih podataka namenjenih za takvu analizu.

Rešenje ovog problema je jednostavno, učitavajući istorijsko transakciono mesto podataka o prodaji u Hadoop klasteru. Na toj osnovi se može napraviti analitička aplikacija koristeći Hive, MapReduce i Apache Spark. Omogućavaju nam sistem otporan na kvarove sa malim kašnjenjem, koji se može koristiti za analizu velike količine podataka po uporedivoj ceni.

Postoji nekoliko primera maloprodajnih kompanija koje koriste velike podatke kroz Hadoop. U ovom odeljku ćemo napomenuti primere Etsy i Sears. Etsy je tržište na mreži, dok Sears ima prodavnice na mreži i u prodavnicama cigle i maltera. Obe ove kompanije trebale su da analiziraju velike količine podataka dnevnika, za marketinške kampanje, upravljanje prodajom, upravljanje zalihama, itd. Usluge Amazon Elastic MapReduce su korišćene za stvaranje Hadoop klastera. Podaci su čuvani i analizirani, kako bi procenili ponašanje potrošača, preporuku za pretragu, plasman proizvoda, upravljanje zalihama, ciljani marketing, promociju proizvoda itd.

Big Data i Hadoop se koriste u maloprodajnoj industriji za sledeće slučajeve korišćenja:

- Maloprodajna analitika za predviđanje zaliha.
- Maloprodajna analitika za dinamičke cene proizvoda.
- Maloprodajna analitika za efikasnost lanca snabdevanja.
- Maloprodajna analitika za ciljanu i prilagođenu promociju i marketing.
- Maloprodajna analitika u otkrivanju i prevenciji prevara.

Uz sve veću interakciju preko društvenih medija i maloprodajnih kanala, kupci upoređuju proizvode, usluge i cene za više prodavaca na mreži i u prodavnicama. Ovakvim ponašanjem potrošači mogu brzo da se prebace sa jednog trgovca na drugog, što je apsolutno neophodno za kompanije u maloprodajnom sektoru da prisluškuju ove informacije i vode evidenciju o curenju u prodajnom sektoru. Potrebno je da maloprodajna kompanija koristi analitiku velikih podataka u maloprodaji, koristeći Hadoop da bi razumela ponašanje kupaca u kupovini.

6. ZAKLJUČAK

Analitika velikih podataka može biti dugotrajna, komplikovana i računski zahtevna, bez odgovarajućih alata, okvira i tehnika. Kada je obim podataka prevelik za obradu i analizu na jednoj mašini, korišćenje softverskih alata može pojednostaviti zadatak paralelnom obradom i distribuiranom obradom. Kao što je opisano u ovom radu Hadoop ima distribuirani sistem datoteka (HDFS), što znači da datoteke podataka mogu da se čuvaju distribuirano, na više mašina. Sistem datoteka je prilagodljiv, s obzirom da se serveri i mašine mogu dodati

za smeštaj sve većeg obima podataka. Hadoop je odličan za serijsku obradu, ali je neefikasan za iterativnu obradu pošto se podaci čuvaju na disku ili u oblaku. Zbog toga se Hadoop kombinuje sa *Apache Spark*-om koji koristi elastične distribuirane skupove podataka.
U ovom radu smo izdvojili tri industrijska domena koji koriste komponente Hadoop ekosistema. Više detalja o korišćenju Hadoopa za obradu velikih podataka može se naći na stranicama kompanija kao što su Facebook, eBay, ORACLE, Yahoo.

ZAHVALNICA

Ova studija je delimično realizovana u okviru EU H2020 projekra LAMBDA (Learning, Applying, Multiplying Big Data Analytics. GA 809965).

LITERATURA

[1] Janev, V. "Ecosystem of Big Data", in V. Janev, D. Graux, H. Jabeen, E. Sallinger (Eds) Knowledge Graphs and Big Data Processing, pp. 3--19, Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-53199-7_1

[2] Apache Hadoop, https://hadoop.apache.org/

[3] Laney, D. "3D data management: controlling data volume, velocity, and variety". Application Delivery Strategies, Meta Group (2001)

[4] Buyya, R., Vecchiola, C., Thamarai, S. "Mastering Cloud Computing: Foundations and Applications Programming", ISBN 978-0-12-411454-8, Morgan Kaufmann, 2013. https://doi.org/10.1016/C2012-0-06719-1.

[5] Lakshen, G., Vraneš, S., Janev, V."Big data and quality: A literature review", Proceedings of the 24th Telecommunications Forum (TELFOR), 2016. https://doi.org/10.1109/telfor.2016.7818902

[6] Grygoriev, N. "Big Data in Financial Services: Trends for 2020", https://devsdata.com/big-data-financial-services/

[7] Sharma, S., Sharma, T., Kotak, B., Hasotkar, A. "Big Data Analysis in Banking Sector", International Journal of New Technology and Research (IJNTR)ISSN:2454-4116, Volume-5, Issue-10, October 2019Pages69-72, https://doi.org/10.31871/IJNTR.5.10.37

[8] Dibya JyotiBora "Big Data Analytics for Intelligent Healthcare Management", Advances in ubiquitous sensing applications for healthcare, Pages 43-57, 2019.

[9] Eric T. Bradlow, Manish Gangwar, Praveen Kopalle, Sudhir Voleti, "The Role of Big Data and Predictive Analytics in Retailing", Journal of Retailing, Volume 93, Issue 1, Pages 79-95, 2017. https://doi.org/10.1016/j.jretai.2016.12.004.