Primena Hadoop okvira u analizi velikih skupova podataka

Aleksandar Simović, Zoran Ćirović
Visoka škola elektrotehnike i računarstva strukovnih studija
Beograd, Srbija
asimovich@gmail.com, zoran.cirovic@gmail.com

Sadržaj— **U radu su opisani problemi analize, skladišta i obrade velikih skupova podataka. Prikazana je primena Hadoop okvira u Big Data okruženju. Analiziran je način rada HDFS i MapReduce komponenti kao i neizbežnost za njihovim korišćenjem.**

Ključne reči – e-poslovanje; Big Data; Hadoop okvir

I. UVOD

Sa eksponencijalnim rastom količine dostupnih informacija na vebu, pitanje skladištenja, odziva, analize i pružanja korisnih, relevantnih izlaznih podataka, sve više se postavlja [1]. IDC (engl. International Data Corporation) je procenila veličinu ukupnog digitalnog univerzuma na 0,18ZB za 2006. godinu sa rastom za 2011. godinu na 1,8ZB. To je približno reda veličine jednog terabajt diska za svaku osobu na svetu [2]. Ovi podaci dolaze iz mnogih izvora:

- *Njujorška berza generiše oko jednog terabajta novih trgovinskih podataka na dan*
- *Facebook obradi oko 10 milijardi fotografija na dan, koristeći 1PB skladišta*
- *Ancestri.com [3] skladišti oko 2,5PB podataka*
- *Internet arhiva [4] skladišti oko 2PB podataka na dan i raste po stopi od 20TB mesečno*

Pitanje skladištenja podataka digitalnog univerzuma danas nije kako ih samo čuvati, već kako ih ekonomično, sigurnosno i inteligentno obraditi i analizirati. Analiza podataka omogućava kompanijama stvaranje novih vrednosti i smernica budućih poslovnih procesa. U 2012. godini, 28% ukupno uskladištenih i analiziranih podataka je moglo stvoriti novu vrednost kompanijama. Procene su da je od ukupnog broja, tek pola obrađeno i analizirano. Do 2020. godine, procenat dostupnih, obradivih podataka digitalnog univerzuma bi mogao da iznosi 45%. Ta količina dostupnih i korisnih informacija treba biti inspiracija programerima, menadžerima i analitičarima za usvajanje i primenu Big Data tehnologije u praksi.

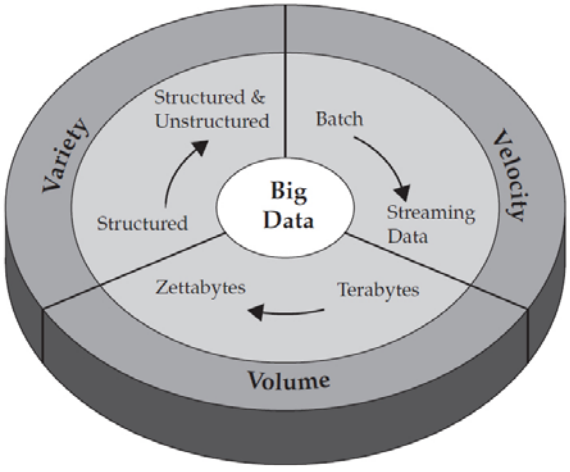
Kompanije danas prikupljaju podatke bez jasne vizije i plana njihovog iskorišćenja i upotrebljivosti [5]. Kako bi podatke pretvorili u korisne informacije i prema njima definisali buduće strategije, potrebna su im nova znanja i nove veštine što rezultuje profitabilnije poslovanje kompanije. Iskoristljivost velikih tokova informacija može radikalno poboljšati performanse organizacionih poslovnih procesa za šta je potrebna promena pristupa donošenju odluka.

Odluke vođene podacima su bolje odluke. Korišćenje Big Data koncepta omogućava rukovodiocima odlučivanje utemeljeno na dokazima a ne na osnovu intuicije. Digitalne kompanije kao sto su Google i Amazon vladaju novonastalim konceptom, ali potencijal da pridobiju konkurentsku dobit od njih može biti čak i veći za manje kompanije.

Upravljački izazovi su realni. Donosioci odluka treba da prihvate donošenje odluka koje su zasnovane na dokazima. Njihove kompanije moraju da zaposle naučnike koji mogu da nađu strukturalni obrazac podataka i upotrebe ga kao informaciju od koristi. Danas bi sve organizacije trebalo da redefinišu svoje poimanje procena, mišljenja i rasuđivanja [6].

II. BIG DATA KONCEPT

Termin Big Data je u istraživačkom izveštaju ispred tadašnje Meta grupe 2001. godine, upotrebio analitičar Douglas Laney [7], i definisao izazove koje donose velike količine dostupnih podataka kroz tri dimenzije – povećanje obima podataka, brzine obrade podataka i raznovrsnost tipova podataka i njenih izvora (engl. 3Vs – Volume, Velocity, Variety) koje zahtevaju nove načine procesiranja i obrade kako bi se omogućilo bolje donošenje poslovnih odluka, otkrića značajnih informacija i optimizaciju poslovnih procesa.



Slika 1. 3Vs – Volume, Velocity, Variety

A. *Volume*

Generisanje velike količine podataka – 90% svih podataka, nastali su u protekle 2 godine. Od danas, količina podataka na globalnom nivou biće udvostručena svake dve godine. Internet stvari (engl. IoT – Internet of Things) sa senzorima stvaraju podatke svake sekunde širom planete i imaju veliki doprinos konstantnom i eksponencijalnom širenju digitalnog univerzuma. Era trilion senzora je tek pred nama.

Od 2012. godine, oko 2.5EB se stvara svakog dana i taj broj se udvostručava svakih 40 meseci. Više podataka prođe putem Interneta svake sekunde nego što je količina globalnog digitalnog univerzuma bila pre 20 godina. To daje mogućnost kompanijama da rade sa više PB podataka iz samo jednog seta podataka. Procenjeno je da Walmart sakupi više od 2.5PB podataka svakog dana od transakcija sa njihovim kupcima; dok Facebook ima najveći klaster koji skladišti preko 100PB informacija.

B. *Velocity*

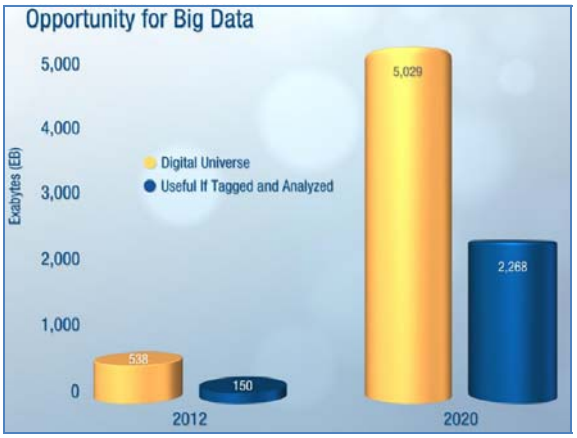
Brzina pristizanja novih podataka je značajnija od velike količine podataka jer informacije u realnom vremenu omogućavaju kompanijama da budu produktivnije od konkurenata. Brz uvid u podatke obezbeđuje konkurentsku prednost. Termin brzina se u kontekstu odnosi na brzinu generisanja podataka, odnosno koliko se brzo podaci stvaraju i procesiraju da zadovolje zahteve i izazove.

Primera radi, Google obradi preko 60 milijardi veb strana u danu a na osnovu upita ključne reči ili fraze generiše stranicu sa rezultatima pretrage (engl. SERP – Search Engine Results Page) za ≈0,2 sekunde.

C. *Variety*

Raznovrsnost podataka Big Data koncepta kroz formu poruka, ažuriranja profila i slika istaknutih na društvenim mrežama, čitanja senzora, GPS signala sa mobilnim telefonima i sl. Mnogi od većine važnih Big Data izvora su relativno novi. Ogromna količina informacija sa društvenih mreža su, primera radi, stare koliko i same mreže; kao što je Facebook ustanovljen 2004 a Twitter 2006. Isto važi i za smart telefone i druge mobilne uređaje koji sada obezbeđuju izuzetno veliki protok podataka. Razvoj je toliko brz i ovi uređaji sveprisutni, da treba pomenuti da je iPhone otkriven pre samo 7 godina (Model: A1203 – 29. Jun 2007. godine). Prema tome, strukturirane baze podataka koje su do skoro skladištile veliku većinu korporativnih informacija danas su nepodesne za Big Data skladišta i analitičku obradu.

U isto vreme, stalno opadanje troškova svih elemenata računarstva: skladišta, memorije, obrade, propusnog opsega mreže i sl, znači da će intenzivan skup raznovrsnosti podataka brzo postati ekonomski isplativ. Sve više poslovnih aktivnosti su digitalizovane. Novi izvori informacija i pristupačna oprema, kombinuju se i otvaraju novu eru elektronskog poslovanja u kojoj postoji velika količina digitalnih informacija koje se tiču bilo koje teme od poslovnog interesa. Elektronske prodavnice na Internetu, društvene mreže, mobilni telefoni, GPS i ostale elektronske komunikacije, proizvode velike količine podataka svojim dnevnim, uobičajenim aktivnostima.



Slika 2. Mogućnosti Big Data koncepta

Masivni skupovi raznovrsnih podataka složenih struktura predstavljaju teškoće za skladištenje, analizu i vizualizaciju rezultata za dalje procese rada [8]. Dostupni podaci su često nestrukturirani podaci, neorganizovani u bazama i veliki, noseći signale i informacije od značaja koji čekaju da budu pročitani i iskorišćeni za poslovne interese. Analitika je donela rigorozne uslove i tehnike za procese donošenja odluka, po automatizmu stvarajući Big Data koncept nezaobilaznim delom poslovnih procesa kompanija. Kao što je Google direktor istraživanja Peter Norving naznačio: Mi nemamo bolje algoritme. Mi samo imamo više podataka.

Termin Big Data se ne odnosi samo na količinu podataka izraženu u velikim jedinicama: petabajtima ili eksabajtima, već na bilo koju količinu podataka koja prevazilazi mogućnost postojećeg sistema da ih obradi i analizira.

III. NASTANAK HADOOP SISTEMA

Big Data sistem zasnovan na Hadoop okviru ima mogućnost obrade nestrukturiranih, polustrukturiranih i strukturiranih podataka. Velike količine podataka, izražene i u PB zastupaju termin Big Data. Tako velike količine podataka moguće je obraditi, skladištiti i njima upravljati koristeći specijalizovane programe dizajnirane za tu namenu.

Primera radi, danas Amazon koji koristi sistem preporuke putem kolaborativnog filtriranja radi pronalaženja sličnosti za generisanje liste, obrađuje 541TB podataka na preko 5 milijardi (10^9) veb strana [9]; dok se Google na početku novog milenijuma suočio sa novim izazovom i novom misijom: Organizovati informacije globalno i kontinuirano na Internetu bilo da su one vezane za pretraživanje podataka ili indeksiranje veb strana.

Kako je sa velikim i brzim rastom dostupnih veb sajtova Google servis postajao sve popularniji, bio je primoran da obradi sve veću količinu brzo pristižućih podataka. Tada nijedan dostupan komercijalni softver nije mogao da upravlja generisanom količinom podataka spremnih za obradu, tako da se Google susreo sa činjenicom da njihova kastomizovana infrastruktura dostiže granice svoje skalabilnosti [10].

Kako bi rešili novonastali problem, inženjeri Google-a dizajniraju i prave novu infrastrukturu za procesiranje podataka sa dva ključna servisa: (1) Google fajl sistem (engl. GFS – Google File System) – pouzdano i skalabilno skladište

brzo dolazećih podataka sa obezbeđenjem tolerantnosti na greške (engl. Fault-Tolerant); i (2) MapReduce – sistem za procesiranje podataka koji omogućava deljenje i paralelno obavljanje posla na velikom broju servera.

2004. godine Google objavljuje naučno-istraživački članak [11] opisujući svoj rad. Ubrzo, Doug Cutting, dobro poznati programer softvera otvorenog kôda odlučuje da isproba opisano rešenje. U to vreme se Cutting sa veb pretraživačem Nutch, na projektu na kome je tada radio susreo sa istim problemima velikih količina generisanih podataka i brzine njihovog indeksiranja, koje je zapravo i navelo Google da razvije MapReduce. Cutting zamenjuje infrastrukturu skladištenja i obrade sa novom implementacijom baziranom na MapReduce-u nazivajući novi softver Hadoop.

Rukovodstvo Yahoo-a, tada direktnog konkurenta Google-a, zainteresovani za Cutting-ov rad, angažuju ga i investiraju u razvoj Hadoop-a sa strateškim ciljem: Hadoop će biti softver otvorenog kôda, slobodnog za preuzimanje, ažuriranje i nadogradnju od strane velike zajednice programera i IT konstruktora koji će raditi na njegovom daljem usavršavanju.

Hadoop je Open Source projekat i posluje pod pokroviteljstvom Apache Software Fondacije danas. Kompanije koje razvijaju proizvode, derivate Hadoop-a i značajni njegovi distributeri na tržištu koji nude i usluge tehničke podrške kroz nadogradnju Hadoop modula su: Cloudera, Hortonworks, IBM, Pentaho, Fico, Jaspersoft, Apache Bigtop, Cascading, Amazon Elastic MapReduce, Azure HDInsight i drugi. Sa realnošću potreba industrije da skladišti, obrađuje i analizira podatke, razvoj Apache Hadoop-a su pratila tri ključna trenda.

Prvo, veliki data centri su drugačije izgrađeni danas u odnosu na onih od pre jedne decenije. Umesto velikih, centralizovanih servera, organizacije kupuju i pokreću kolekcije potrebnih, ugradnih rekova na serverima, od kojih svaki ima više procesora i diskove kapaciteta izraženih u TB. Starije centralizovane i deljene aplikacije, sporo i slabo rade u ovakvim uslovima. Hadoop je napravljen to da iskoristi.

Drugo, raznovrsnost i kompleksnost raspoloživih podataka je veoma brzo porasla. Menadžment relacionih baza podataka (RDBMS) i struktuirani data menadžment sistemi koji su i dalje veoma rasprostranjeni gube mogućnost obrade velikih logova, tekstova, slika, audio i video zapisa, senzorskih zapisa i drugih kompleksnih tipova podataka koji velikom brzinom pristižu na servere i data centre danas. Ova raznolikost je i kvalitativne i kvantitativne prirode. Kompleksni podaci ne mogu biti lako iskorišćeni u bazama struktuiranih redova i kolona. Informacije ove vrste su velike, a brzina kojom se generišu raste. Inteligentne aplikacije i novi softverski sistemi poput Hadoop-a pomažu u praćenju rasta i analizi digitalnog univerzuma.

Konačno, nove tehnike za obradu podataka moraju biti sposobne za izdvajanje vrednosti i uvid u informacije sakrivene u tim podacima. Poslovanje veb kompanija se ranije baziralo na kombinaciji tipova podataka koji se beleže, (npr. postavljanje veb logova koji beleže aktivnosti korisnika ili ažuriranje naloga registrovanog korisnika) i omogućavanje uvida u informacije iz različitih izvora. Platforma kao što je Hadoop, koja omogućava analitiku kombinovanjem strukturiranh i nestrukturiranih, kompleksnih podataka; da obradi i analizira izvore i tipove podataka, veliki je iskorak i napredak informacionih tehnologija i upravljanja informacijama.

Ova tri trenda – (i) prelazak na skalabilnu, elastičnu infrastrukturu; (ii) kompleksnost i raznolikost dostupnih podataka; (iii) snaga i vrednost koje dolaze iz kombinovanja različitih podataka za obavljanje sveobuhvatne analize, čini Hadoop kritičnom novom platformom za informacione kompanije koje manipulišu velikim količinama podataka. Primera radi, banke i osiguravajuća društva koriste algoritme praćenja ponašanja komitenata na osnovu istorije transakcija radi otkrivanja sumnjivih aktivnosti. Kompanije okrenute elektronskoj trgovini imaju mogućnost da poboljšaju tačnost svojih sistema za preporučivanje proizvoda, kombinujući transakcione podatke sa logovima aktivnosti korisnika generisanim na veb serveru ili drugim mrežnim uređajima. U energetskom sektoru, kompanija Tennessee Valley Authority - TVA je izgradila sistem prikupljanja i analize podataka pod nazivom Open Phasor Data Collector – OpenPDC na Hadoop platformi [12].

OpenPDC prati količinu dolazećih podataka od senzora vezanih za sistem proizvodnje električne energije. Ovi senzori, raspoređeni oko električne mreže, prate i izveštavaju funkcionalnost i status svakog generatora. Pažljivo praćenje i brz odgovor na moguće promene na mreži, omogućava Hadoop ekosistemu da smanji ili spreči kvarove, kao i da bolje upravlja raspoloživim kapacitetima. Ovo je smanjilo operativne troškove kompanije i obezbedilo društveno odgovorno poslovanje sa ciljem bolje kontrole staklene bašte, emisionih gasova i drugih ekoloških rizika.

Apache Hadoop je okvir koji omogućava distribuiranu obradu velikih skupova podataka širom klastera računara koristeći jednostavan model programiranja. Skalabilan je i dizajniran da se proširi sa pojedinačnog servera na hiljade mašina i da pritom svaka omogućava izračunavanja i skladištenje informacija. Umesto da se oslanja na hardver da bi isporučio visoku dostupnost, sistem je dizajniran da detektuje i upravlja kvarovima na aplikacionom nivou, čime usluga postaje dostupna i raspoloživa na vrhu klastera kompjutera, od kojih svaki od njih može biti sklon kvaru.

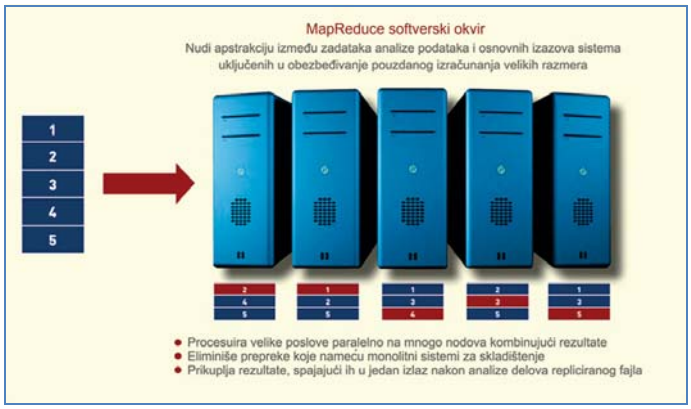
IV. KOMPONENTE HADOOP TEHNOLOŠKOG OKVIRA

Kao Google MapReduce sistem na kojem je bio zasnovan, Hadoop se sastoji od dve glavne (engl. Core) komponente: (1) File Store – HDFS, i (2) Distribuirani sistem za obradu – Distributed Processing System – MapReduce.

Podaci se čuvaju na HDFS-u koji obezbeđuje skalabilno, prilagodljivo, otporno na greške skladište. HDFS detektuje i kompenzuje greške na serveru ili probleme na disku. Skladišti fajlove na više servera u klasteru. Fajlovi se razgrađuju po blokovima i svaki blok je kopiran na više od jednog servera (obično 3).

Ovaj replikacija omogućava: (a) otpornost na greške – gubitak jednog diska ili servera neće uništiti datoteku, i (b) performanse – bilo koji blok može da se pročita sa jednog ili

Slika 3. Distribuirano skladište – File Store – HDFS



Slika 4. MapReduce softverski okvir

više servera pritom poboljšavajući protok. HDFS stvara više replika svakog bloka podataka i distribuira ih na računare širom klastera, omogućavajući pouzdan i brz pristup.

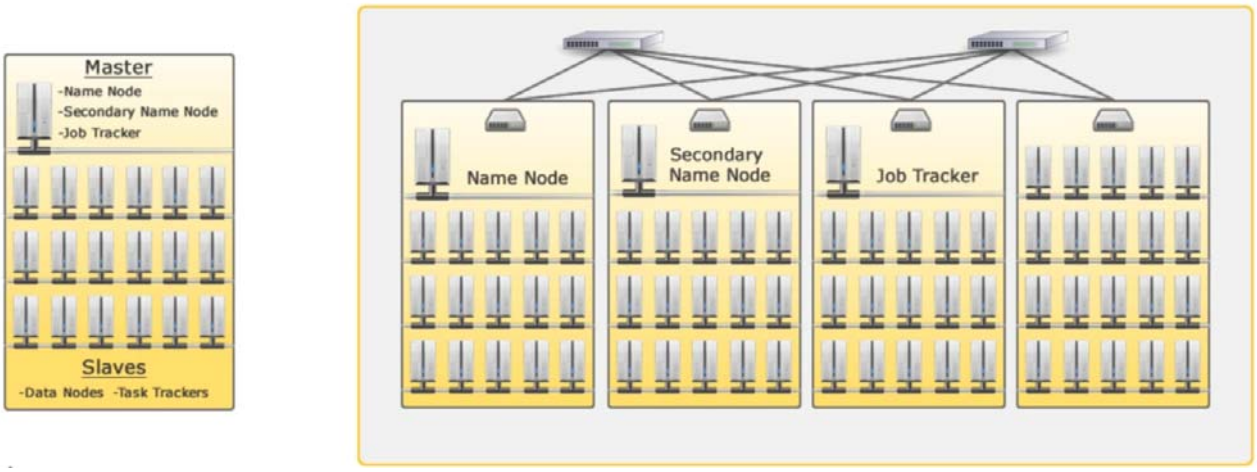
HDFS obezbeđuje dostupnost podataka tako što kontinuirano prati rad servera u klasteru, i blokova kojima oni upravljaju. Individualni blokovi podležu proveri i kontroli rada. Kada se blok pročita, utvrđuje se ispravnost (da li je vrednost koja je zabeležena, ispravna). Ukoliko je blok oštećen, biva obnovljen sa jednim od njegovih replika a ukoliko server ili disk otkaže, svi podaci koje skladišti repliciraju se na drugi nod ili druge nodove u klasteru iz čitave kolekcije replika. Kao rezultat, HDFS toleriše i nadoknađuje moguće greške u klasteru. Međutim, kako klaster postaje veći, postoji i veća šansa da neki od servera otkaže. U slučaju neuspeha HDFS-a, organizacije mogu manje ulagati u servere, puštajući softver da kompenzuje hardverske probleme.

HDFS omogućava povoljno, dostupno i pouzdano skladištenje podataka, međutim, sam nije dovoljan da bi se stvorio odgovarajući nivo adaptacije Hadoop ekosistema koji ga karakteriše poslednjih nekoliko godina. Eksponencijalni rast ukupne količine dostupnih podataka je doveo do mnogih vitalnih izazova u modernom elektronskom poslovanju. Postojeći sistemi su postali neadekvatni za obradu tako velikih skupova podataka.

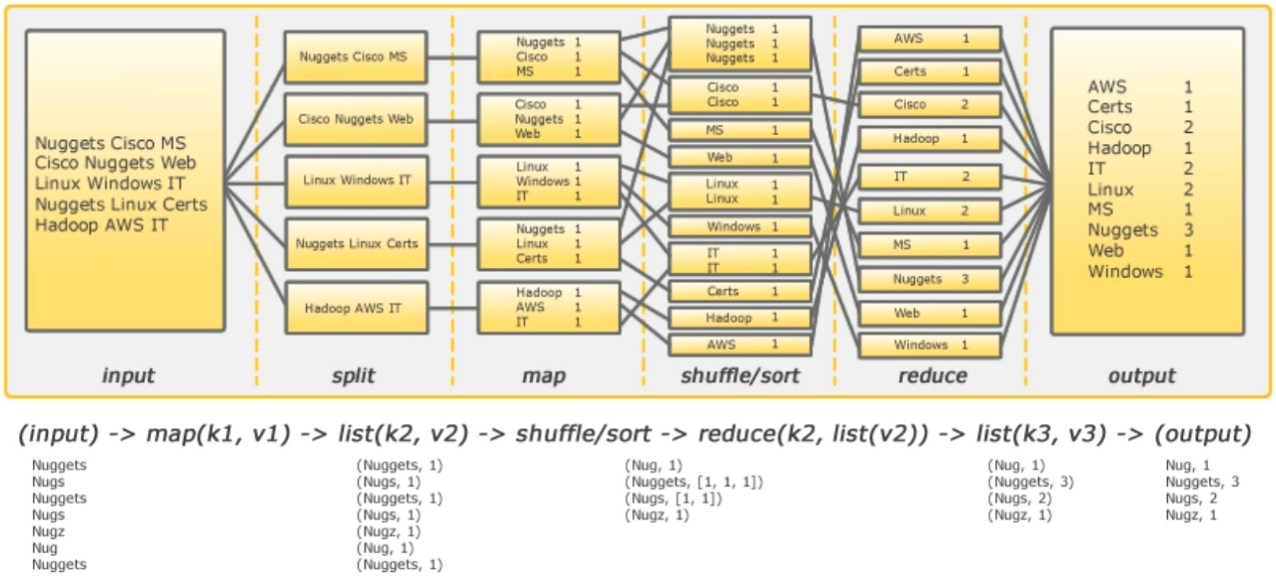
Da bi to prevazišao, Google je kao pionir predstavio, tada novi model programiranja – MapReduce, drugu glavnu Hadoop komponentu za paralelnu obradu podataka. MapReduce uključuje važnu softversku komponentu planiranja posla (engl. Job Scheduler). Planer je odgovoran za izbor servera koji pokreće poslove i za izvršavanje planiranja više poslova na deljenom klasteru.

Planer konsultuje Name Node za lokaciju svih blokova koji čine fajl ili fajlove potrebne za izvršenje posla. Svaki server se upućuje da radi analizu kôda lokalnog bloka ili blokova. Infrastruktura MapReduce procesiranja obuhvata apstrakciju koja se zove ulazna podela (engl. Input Split); rascep koji omogućava da svaki blok bude podeljen na individualne, pojedinačne zapise.

Kôd koji definiše posao mapera, praktično omogućava programerima da pišu i pokreću program koji radi direktno na svakom DataNode serveru u klastreru. Kôd spoznaje format podataka koji su skladišteni u svim blokovima i može sprovesti algoritme za obradu; npr. broj pojavljivanja jedne reči u datom tekstu; detekciju obrazaca, paterna; mašinsko učenje; prepoznavanje lica i dr.



Slika 5. HDFS arhitektura



Slika 6. MapReduce – Shuffle and Sort

Na kraju faze posla mapera (Map) rezultati se sakupljaju i filtriraju redukcijom (Reduce). MapReduce garantuje da će podaci biti sortirani i dostavljeni reduktoru, tako da izlazi svih mapera se prikupljaju i prosleđuju procesu sortiranja (engl. Shuffle and Sort). Sortiran izlaz se zatim prenosi na deo redukcije za obradu. Rezultati su upisuju nazad u HDFS.

Zbog replika ugrađenih u HDFS, MapReduce je u stanju da pruži i druge korisne funkcije. Na primer, ako jedan od servera koji je uključen u MapReduce posao radi sporo u odnosu na druge koji su posao već završili, planer može pokrenuti drugu instancu tog istog zadatka na drugom serveru u klasteru. Dakle, preopterećeni nodovi u klasteru neće zaustaviti ni usporiti MapReduce posao.

Jedna od važnih karakteristika MapReduce poslova i ključna prednost je mogućnost obrade PB podataka u datom roku i dobijanja traženog odgovora. Korisnici mogu čekati na rezultat minut ili sat, ali mogu postavljati upite na koje je jednostavno nemoguće bilo odgovoriti pre nego što je MapReduce postao dostupan.

Osobine Big Data koncepta su uticale na stvaranje Hadoop sistema sa komponentama HDFS i MapReduce sa pratećim razvojnim modulima. Do tada, problem velikih setova podataka nije mogao da reši nijedan komercijalni ni istraživački sistem. Platforma se sada koristi da podrži veliku raznovrsnost aplikacija. Ove aplikacije nužno ne karakterišu ogromni skupovi podataka. Umesto toga, one trebaju ključne osobine koje Hadoop nudi.

Hadoop je konsolidovana platforma za skladištenje svih vrsta podataka. Naravno, postoje nezavisne relacione baze koje postoje na tržištu i danas, i ti sistemi će ostati u upotrebi u godinama koje dolaze sa zadatkom da tačno rešavaju probleme za koje su i namenjeni. Hadoop ih dopunjuje, komplementaran je sa postojećim sistemima i obezbeđuje novo skladište gde se strukturirani podaci sa složenim, nestrukturiranim mogu lako kombinovati.

Drugo, Hadoop obezbeđuje znatno veće skladište po mnogo nižoj ceni nego postojeći sistemi utemeljenog softvera i hardvera koje je teško zameniti zbog njihove široke upotrebe. Omogućava pouzdano skladište sa neizbežnom silaznom krivom cene koštanja kako distributeri poboljšavaju performanse sistema. Konačno, MapReduce eksploatiše distribuiranu arhitekturu HDFS skladišta radi prilagodljivog, skalabilnog i pouzdanog, paralelno-orjentisanog algoritma za procesiranje datih zadataka. Korisnici nisu ograničeni na mali skup algoritama omogućenih od strane tradicionalnih RDBM sistema. Hadoop sistem se za skladištenje može programirati. Analitičari mogu obrađivati i analizirati podatke koristeći procesore povezane sa diskovima na kojima se ti podaci i nalaze. Ova kombinacija je nova – konsolidacija svih tipova podataka; niža cena usluge i pouzdana platforma za čuvanje koja isporučuje brzo i paralelno, distribuirano izvršavanje. Hadoop nudi mogućnost eksploatacije podataka i izvlačenja korisnih informacija na način koji ranije nije postojao.

V. ZAKLJUČAK

MapReduce i Hadoop su nastali kada i Google i Yahoo, preuzimajući rešavanje inženjerskih i analitičkih problema – jezgra njihovog poslovanja: indeksiranja više od milijarde veb stranica. Tehnologija ima neprocenjive vrednosti. Dokazana je vrednost i u drugim, neočekivanim oblastima, kao što je rendering mapa, provera pravopisa, poboljšanje izgleda strane, izbor reklama i tako dalje. Alat opšte namene koji omogućava analizu svih podataka koji se ranije nisu mogli čak ni zamisliti. Pruža programerima lak pristup svim podacima i stvara iznenađujuće veliki broj poslovnih poboljšanja.

Hadoop softver omogućava distribuiranu obradu i procesiranje velikih zapremina podataka obrađujući ih preko klaster nodova koristeći module koje je potrebno implementirati u Hadoop ekosistem. Sa Hadoop infrastrukturom, obezbediće se upravljanje i analiziranje strukturiranih, polustrukturiranih i nestrukturiranih podataka iz različitih izvora.

LITERATURA

[1] A. Simović, „Sistemi preporuke u e-trgovini“, Impact of the Internet on Business Activities in Serbia and Worldwide, Singidunum University International Scientific Conference Sinteza, Beograd, 2014.

[2] J. Gantz, D. Reinsel, „The digital universe in 2020: Big data“, IDC Analyze the Future, 2012.

[3] www.ancestry.com – datum pristupa: 5. januar 2016.

[4] www.archive.org – datum pristupa: 5. januar 2016.

[5] P. Staletić, A. Simović i M. Lutovac, “Elektronska prodavnica korišćenjem open-source softvera,” TELFOR, Beograd, 2010.

[6] A. McAfee, E. Brynjolfsson, „Big data: the management revolution“, Harvard business review, 2012.

[7] L. Douglas, „3D data management: Controlling data volume, velocity and variety“, META Group Research Note 6, 2001.

[8] S. Sagiroglu, D. Sinanc, „Big data: A review", Collaboration Technologies and Systems, International Conference IEEE, 2013.

[9] www.aws.amazon.com/datasets/41740 – datum pristupa: 6. februar 2016.

[10] M. Olson, „Hadoop: Scalable, flexible data storage and analysis“, IQT Quart, 2010.

[11] D. Jeffrey, S. Ghemawat, „MapReduce: simplified data processing on large clusters“, Communications of the ACM , 2008.

[12] J. Patterson, „The Smart Grid: Hadoop at the Tennessee Valley Authority“, Cloudera Blor, 2009.

ABSTRACT

The paper describes the problems of analysis, storage and processing of large data sets today. It shows the application of Hadoop framework in Big Data environments. We analyzed HDFS and MapReduce components as well as the inevitability of their use.

BIG DATA ANALYSIS USING HADOOP FRAMEWORK

Aleksandar Simović, Zoran Ćirović