UDC: 004.722 Info M str. 4-10

HADOOP CLOUD TEHNOLOGIJE
HADOOP CLOUD TECHNOLOGIES

Dejan Hadži-Milosavljević, Dušan Starčević

REZIME: Zahtevi za razvojem naprednih aplikacija elektronskog poslovanja, 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 podataka u realnom vremenu, zasnovani na Big data tehnologijama. Ispravno i potpuno objašnjenje šta je Hadoop i kakva je njegova povezanost sa Cloud-om, sledi iz proučavanja upravo Big data tehnologije.Hadoop obezbeđuje pouzdano i skalabilno skladištenje velikih količina različitih podataka, upravljanje fajlovima u distribuiranom okruženju, kao i distribuirano računarstvo od koga se zahteva da omogući Big data (tri „V“ model). Sa velikim rastom količine podataka u poslednje vreme, razmatra se skladište podataka zasnovanog u cloud-u, pri čemu bi se iskoristila elastičnost cloud-a da se napravi sistem koji je dinamički skalabilan. U radu će biti opisana tehnologija Hadoop-a u cloud (Hadoop kao servis) i ukazaće se na prednosti ove tehnologije.

KLJUČNE REČI: Big Data tehnologija, Hadoop, Cloud računarstvo, Hadoop u Cloud-u

ABSTRACT: Requirements for the development of advanced e-business applications, characterized by reliability, distribution and scalability, can not be realized using traditional databases. That’s why new approaches for storage, fast search and analysis of large amounts of data in real time, based on Big data technologies are being developed. The correct and complete explanation of what Hadoop is and what his relationship with Cloud is, follows from the study of Big Data technology. Hadoop provides reliable and scalable storage of large amounts of different data, management of files in a distributed environment, and distributed computing from which it is required to enable Big data (three “V” model). With a large increase in data volume lately, a Hadoop in Cloud is being considered, using the elasticity of the Cloud technology to create a system that is dynamically scalable. This paper deals with the Hadoop in Cloud (Hadoop as a Service) and the benefits of this technology.

KEY WORDS: Big Data Technology, Cloud Computing, Hadoop as a Service

1. UVOD

Razvojem tehnologija mobilnog poslovanja, interneta, inteligentnih uređaja i društvenih medija povećava se količina podataka koja se čuva u informacionim sistemima preduzeća. Zahtevi za razvojem naprednih aplikacija elektronskog poslovanja, 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 podataka u realnom vremenu, zasnovani na Big data tehnologijama. Ispravno i potpuno objašnjenje šta je Hadoop i kakva je njegova povezanost sa Cloud-om, sledi iz proučavanja upravo Big data tehnologije. Razlike između Big Data i onoga što se tradicionalno nazvalo analizom su: ogromna količina podataka do kojih se sada ima pristup, veća brzina prikupljanja podataka i veća raznovrsnost različitih podataka.

Količina dostupnih podataka je neverovatno velika. Prema [1,2], pre četiri godine količina podataka je bila u opsegu od nekoliko eksabajtova (2,5Eb) i udvostručavala se svakih 40 meseci. Dakle, više podataka prolazilo je kroz Internet nego što je bilo uskladišteno na celom Internetu pre dvadesetak godina.

Brzina kojom se kreiraju podaci ponekad je važnija od količine podataka jer ona uslovljava sposobnost da se reaguje na velike količine podataka u realnom vremenu.

Raznolikost podataka se naglo povećala pojavom društvenih medija. Danas postoji mnogo relativno novih izvora podataka. Kada se ranije govorilo o podacima, mislilo se na podatke koji se nalaze u relacijskim bazama podataka, dok danas postoje nove vrste podataka kao što su poruke, tweet-ovi, ažurirani statusi, fotografije na društvenim mrežama, podaci iz senzora, lokacije ili GPS koordinate iz mobilnih telefona itd.

Big data, Hadoop i Cloud tehnologije analizirane su od strane analitičara, stručnjaka za razvoj aplikacija i profesionalnih isporučioca aplikacija. Prema njihovoj proceni, Hadoop naširoko prodire i napreduje u preduzećima. Hadoop Cloud tehnologija tj. Hadoop u Cloud-u, idealano odgovara za obezbeđivanje moćnog Big data za obradu velikih paralelnih skupova podataka. Cloud poseduje sposobnost da obezbedi fleksibilnu i agilnu računarsku platformu potrebnu za velike podatke, kao i mogućnost korišćenja velike računarske snage.

U ovom radu najpre će biti posvećena pažnja Big Data tehnologiji, a zatim projektu Hadoop. Hadoop je postao standard za upravljanje i obradu podacima reda veličine stotine TB (terabajt) i reda velićine PB (petabajt). Zbog velike eksplozije podataka u poslednje vreme, postoje ideje o skladištenju velikih podataka u cloud-u jer se ono razlikuje od tradicionalnog Hadoop skladištenja, a pri tome se koristi elastičnost cloud-a kako bi se napravio sistem koji je dinamički skalabilan. U radu će biti opisano smeštanje Hadoop-a u cloud, i najveća pažnja biće posvećena tom cloud servisu - Hadoop u cloud-t tj. Hadoop kao servis.

2. HADOOP OKVIR ZA BIG DATA

Koncept Big Data predstavlja jedan od najaktuelnijih pojmova u oblasti informacionih tehnologija. Big Data u prevodu znači velika količina podataka, pri čemu se govori o redu veličine terabajtili petabajt. Podaci dolaze iz različitih izvora, kao

što su društvene mreže, slike, senzori, logovi, video zapisi, i mnogi drugi. Pored podataka iz osnovne delatnosti i transakcija organizacije, ovde spadaju i mnogobrojni podaci iz spoljnih izvora, i koji ne moraju biti direktno povezani za procese koji se odvijaju u organizaciji. Podaci mogu biti strukturni i nestrukturni.

Zahvaljujući novim tehnologijama, uređajima i komunikacijama, količina podataka proizvedenih od strane čovečanstva brzo raste svake godine. Količina podataka koji su dobijeni do 2003. godine iznosila je 5 milijardi gigabajta. Ako su skladišta podataka u obliku diskova, oni mogu popuniti čitav fudbalski teren. Isti iznos količine podataka stvoren je svaka dva dana u 2011. godini, a svakih deset minuta 2013. godine. Ova stopa i dalje raste enormno. Iako su sve ove dobijene informacije značajne i mogu biti korisne pri obradi, zanemaruju se. Podaci se proizvode u različitim uređajima i aplikacijama i mogu biti u tri oblika: Strukturni podaci (relacijski podaci), Semi strukturni podaci (XML podaci), Nestrukturni podaci (Word, PDF, Text, Media Logs). Tako veliki podaci uključuju ogromnu količinu, veliku brzinu i različitu raznovrsnost podataka i zato se potreba za primenom Big data tehnologija često objašnjava korišćenjem tri „V“ modela, po kome su glavne karakteristike Big data:

– Obim podataka (Volume)
– Raznovrsnost podataka (Variety)
– Brzina (Velocity).

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) i GlusterFS.

Standardni mehanizmi pretrage ne zadovoljavaju u pogledu brzine obrade podataka kada se primenjuju u Big data okruženju, zbog čega je realizovan jedan od novih pristup pod nazivom MapReduce. Googl je 2010. patentirao ovaj algoritam koji pretražuje podatke uređene po parovima (ključ, podatak). Algoritam se koristi kao osnovni mehanizam za pretraživanje i izveštavanje u većini Big data baza podataka.

Tehnologije prenosa podataka važne su za pružanje preciznije analize, zahvaljujući kojima se može doći do konkretnijeg donošenja odluka a samim tim i do veće operativne efikasnosti, smanjenja troškova i smanjenja rizika za posao. Da bi se iskoristila moć velikih podataka, potrebna je infrastruktura koja može da upravlja i obrađuje ogromne količine strukturnih i nestrukturnih podataka u realnom vremenu i može zaštititi privatnost i sigurnost podataka. Glavni Big data izazovi u vezi sa podacima su sledeći: snimanje podataka, organizovanje podataka, skladištenje, pretraživanje, deljenje, transfer podataka, analiza podataka, i prezentacija podataka. Da bi ispunili ove izazove, organizacije obično uzimaju pomoć od poslovnih servera na sledeće načine.

– Tradicionalni pristup - U ovom pristupu, preduzeće će imati računar za čuvanje i obradu velikih podataka. Ovde će se podaci čuvati u RDBMS-u kao što su Oracle Database, MS SQL Server ili DB2, i sofisticirani softverski programi mogu biti pisani za interakciju sa bazom podataka, procesiranje potrebnih podataka i prezentaciju korisnicima za svrhu analize. Postoji sledeće ograničenje da bi ovaj pristup dobro funkcionisao, a to je da se ima manji obim podataka koji mogu biti smešteni na standardnim serverima baze podataka.
– Google-ovo rešenje - Google je rešio ovaj problem pomoću algoritma koji se zove MapReduce. Ovaj algoritam deli zadatak na male delove i dodeljuje te delove mnogim računarima koji su povezani preko mreže i prikuplja rezultate da bi obrazovali konačni skup podataka rezultata. Dakle, Google MapReduce rešenje podrazumeva da se više različitih poslovnih radnji obavlja na pojedinačnim serverima sa većim kapacitetom, a centralizovan računar prikuplja rezultate njihovih obrada i formira krajnji rezultat.

Doug Cutting, Mike Cafarella i tim preuzeli su rešenje koje je obezbedio Google, i 2005. godine pokrenli su open source projekat pod nazivom Hadoop koji je dao najveće rezultate kao Big Data softver. Ovaj prijekat je sada kao Apache Hadoop registrovani zaštitni znak Apache Software Foundation. Hadoop pokreće aplikacije koristeći algoritam MapReduce gde se podaci obrađuju paralelno na različitim CPU čvorovima. Ukratko, Hadoopov okvir je dovoljno sposoban za razvoj aplikacija koje se mogu pokrenuti na klasterima računara i mogu obavljati potpunu statističku analizu za ogromne količine podataka.

3. HADOOP ARHITEKTURA

Hadoop je open source okvir koji obezbeđuje skladištenje i obradu velikih podataka u distribuiranom okruženju preko klastera računara pomoću jednostavnih programskih modula, i omogućava pretraživanje i analizu velikih količina podataka. Dizajniran je kako za pojedinačne servere, tako i za na hiljade umreženih mašina, od kojih svaka nudi lokalnu računarsku i skladišnu opremu. Hadoop je napisan u programskom jeziku Java i, kao open source softver, dostupan je korisnicima koji sa open source licencom mogu da menjaju, prepravljaju i poboljšavaju sadržaj njegovog izvornog koda.

Srž Hadoop-a čine njegove komponente, koje su moćne svaka za sebe, i zajedno čine jednu savršenu platformu. Hadoop se sastoji od sledećoh komponenti:

(i) Hadoop Distributed File System (HDFS)

HDFS je zasnovan na Google File Sistem-u (GFS) i pruža distribuirani sistem datoteka koji je dizajniran da radi i na velikim klasterima od po hiljadu malih računarskih mašina na pouzdan način i otporan je na greške. HDFS je jako bitna komponenta i on definiše kako se podaci skladište, kopiraju

i čitaju. HDFS je zaslužan za mogućnost lakog skladištenja ogromne količine podataka. Takođe, obezbeđuje i veliki protok podataka, kao i pristup aplikacijskim podacima.

(ii) Hadoop MapReduce

Komponenta Hadoop MapReduce je softverski okvir za lako pisanje aplikacija koje na visokopouzdan način paralelno obrađuju velike količine podataka na klasterskom hardveru primenom MapReduce algoritma. Naziv komponente tj. termin MapReduce odnosi se na sledeća dva različita zadatka koje Hadoop programi izvršavaju:

– Map Task: Ovo je prvi zadatak koji prihvata ulazni podatak i pretvara ga u skup podataka, pri čemu se pojedinačni podaci iz tog skupa razdvajaju i dodeljuje im se uređeni par (ključ, vrednost podataka).
– Reduce Task: Ovaj zadatak kao svoje ulaze ima izlaze iz zadatka MapTask tj.skup uređenih parova (ključ, vrednost podataka). Kombinujući uređene parove smanjuje ulazni skup i kao rezultat ima smanjeni skup uređenih parova (ključ, vrednost podataka). Zadatak Reduce Task uvek se izvršava nakon zadatka Map Task.

Obično su ulazni i izlazni skup uređenih parova čuvaju u fajl sistemu. Okvir MapReduce brine o raspoređivanju zadataka, nadgledanju i ponovnom izvršavanju neuspelih zadataka.

(iii) Hadoop YARN

Komponenta YARN (Yet Another Resource Navigator) je okvir za upravljanje resursima u klasteru i za upravljanje izvršenjem poslova. Kada se radi sa Hadoop-om svaka aktivnost se smatra poslom bez obzira da li je to proces, izvršavanje, pokretanje, ili neko kontrolisanje. YARN služi da upravlja tim poslovima u odnosu na resurse. YARN je uveden sa Hadoop-om 2.0. kao poboljšanje osobina MapReduce-a. U Hadoop-u 1.0 sve što sada obavlja YARN obavljao je MapReduce.

(iv) Hadoop Common

Komponenta Hadoop Common je paket koji sadrži Java biblioteke i uslužne programe koje zahtevaju druge Hadoop komponente. Ove biblioteke obezbeđuju sistem datoteka i operativni sistem na apstraktnom nivou, i sadrže potrebne Java datoteke i skriptove potrebne za pokretanje Hadoop-a. Ovaj paket takođe sadrži izvorni kod i dokumentaciju, kao i sve potrebne elemente za komuniciranje Hadoop-a sa ostalim alatima.

Hadoop platforma ima mogućnost da se, po potrebi, proširi dodatnim alatima koji čine Hadoop ekosistem. U ove alate spadaju i HDFS i MapReduce, koji su i komponente platforme, ali se vode i kao Apache projekti. Na Hadoop platformu moguće je dodati još dosta raznih alata, uglavnom razvijenih od strane Apache fondacije. Ti alati su sposobni da komuniciraju sa Hadoop komponentama i njihov rad može da bude nezavistan, a neki od njih su pravljeni samo za Hadoop. Većina alata se ne koristi samostalno već samo uz Hadoop platformu. Hadoop je dovoljan ako se samo skladištite podaci i da to bude brzo, i možda eventualno da se kasnije analiziraju podaci. Za to je dovoljan samo HDFS, i nema se potreba za dodatnim alatima tj. za ekosistemom. Moguće je instalirati samo HDFS, i tako se štedi na vremenu podešavanja klastera. Naknadno je moguće i instalirati MapReduce i da se dobije samo Hadoop kao celina. Za analizu podataka, bilo da su oni strukturni ili ne, služi MapReduce i moguće je za potrebe analize da se individualno pišu Java programi. Ako se želi da se koristi pun potencijal Hadoop-a i alata koji se dodaju na njega, moguće je napraviti sopstveni Hadoop ekosistem ili koristiti rešenja koja su već napravljena tako da uključuju sve one alate koji se obično koriste u poslovnim analizama.

Na tržištu postoji mnogo alata koji su u relaciji sa Hadoop-om. Većina ih je pod Apache licencom, ali i firme poput Facebook-a i Microsoft-a razvijaju neke svoje alate koji mogu da se instaliraju. Alati koji koji dolaze uz Apache Hadoop su:

– Distribuirani Fajl Sistemi (HDFS)
– Distribuirano programiranje (MapReduce, Apache Pig, Apache Tez)
– NoSQL baze podataka (Apache HBase, Apache Accumulo)
– SQL baze podataka (Apache Hive, Apache HCatalog)
– Unošenje podataka (Apache Flume, Apache Sqoop, Apache Storm)
– Programiranje servisa (Apache Zookeeper)
– Zakazivanje (Apache Oozie, Apache Falcon)
– Mašinsko učenje (Apache Mahout)
– Bezbednost (Apache Knox)
– System Deployment (Apache Ambari, HUE).

Najveće pogodnosti korišćenja Hadoop platforme i njenih dodatanih alata su [3]:

1. Mogućnost skladištenja i brze obrade velikih količina bilo koje vrste podataka - Ovo je ključna stavka s obzirom na to da se stalno povećavaju količina podataka i raznovrsnost podataka, posebno u društvenim mrežama i na internetu.
2. Računarska snaga - Hadoop-ov distribuirani računarski model brzo obrađuje velike podatke, što više računarskih čvorova se koristi, to je veća snaga za procesiranje.
3. Visoka pouzdanost - Obrada podataka i rad aplikacija zaštićeni su od otkaza hardvera. Ako otkaže neki računarski server u mrežnom čvoru, zadaci se automatski preusmeravaju na druge čvorove čime se otklanja ovaj nedostatak u distribuiranom računarskom sistemu i obezbeđuje njegovo kontinualno funkcionisanje. Inače, više kopija svih podataka se automatski čuvaju.
4. Fleksibilnost - Za razliku od tradicionalnih relacionih baza podataka, pre skladištenja podataka, ne treba da se preprocesuiraju podaci.. Može da se skladišti što više podataka, koliko se želi, i da se odluči kako će se kasnije koristiti. To se podrazumeva i za podatke kao što su tekst, slike i video zapisi.
5. Mali troškovi - Open-source radni okvir je besplatan i koristi se hardversko skladište za čuvanje velikih količina podataka.

6. Prilagodljivost - Uz samo malo dodatne administracije, jednostavno se može razviti sosptveni sistem da obrađuje više podataka prostim dodavanjem serverskih čvorova.

Izazovi usled upotrebe Hadoop platforme su [4]:

1. MapReduce programiranje nije podjednako odgovarajuće za sve probleme:

MapReduce programiranje je dobro za jednostavne informacije i probleme koji se mogu rasporediti na nezavisne jedinice, ali nije efikasano za iterativne i interaktivne analitičke zadatke. MapReduce je intenzivan na fajlove. Zbog toga što čvorovi međusobno ne komuniciraju, osim prilikom sortiranja i mešanja, iterativni algoritmi zahtevaju da se kompletiraju faze sa više planskih mešanja i redukovanih sortiranja. Ovo dovodi do kreiranja više fajlova između MapReduce faza i nije efikasno za napredne analitičke računare.

2. Postoji široko prihvaćen probelem nedostatka talenata:

Teško je naći programere koji imaju dovoljno Java sposobnosti, a da istovremeno budu produktivni sa MapReduce-om. To je jedan od razloga zbog čega su se provajderi distribucije trkali da stavljaju relacijsku (SQL) tehnologiju na vrhu Hadoop-a. Mnogo je lakše pronaći programera sa SQL veštinama nego sa MapReduce veštinama.

3. Bezbednost podataka:

Izazov usredsređen na pitanja o bezbednosti podataka, uvek je pristutan usled pojavljivanja novih alata i tehnologija. Protokol autentifikacije Kerberos je odličan izbor za bezbednost Hadoop-ovog okruženja.

4. Potpuno upravljanje podacima:

Hadoop nema alatke za upravljanje podacima, za "čišćenje" podataka, nema metapodatke jednostavne za upotrebu. Posebno nedostaju alati za kvalitet podataka i standardizaciju

4. HADOOP KAO SERVIS U CLOUD-U

Cloud tehnologija zasniva se na tome da svi podaci i računarski resursi - aplikacije, dokumenti, hardver, i dr., dostupni korisniku u svakom trenutku, pod uslovom da je prethodno uspostavljena internet veza. Cloud predstavlja uslugu dostavljanja servisa umesto samog proizvoda. Cloud u svakom trenutku pruža mogućnost pristupa aplikacijama, podacima, servisima za čuvanje podataka i ne traži od korisnika poznavanje fizičke lokacije sistema koji pruža servis. Cloud je već uveliko prisutan u oblicima koji veliki broj korisnika svakodnevno koristi, kao što su npr. brojne društvene mreže, e-mail servisi ili tzv. pametni telefoni itd.

Cloud je usluga koja nudi neograničene količine svih resursa (hard disk, procesor, memorija i dr.) onda kada su korisniku zaista potrebni i u onoj meri koja odgovara korisničkim potrebama, i to tako da sve korisnik može samostalno da kontroliše. S obzirom na to da korisnik nije vlasnik infrastrukture, samim tim nije dužan ni da je održava. Korisnik plaća iznajmljivanje usluga i onih resursa koji su mu u određenom trenutku potrebni. Ukoliko mu je potrebna neka poslovna aplikacija, on je iznajmi i nema potrebe da je kupuje. Kupovinom, nema promene hardvera i svih ostalih aktivnosti koje se tu podrazumevaju – u ovom slučaju potreban je samo pristup cloud servisu. Korisnik plaća samo ono što koristi i onoliko koliko to koristi. Kada mu usluge za koje se opredelio više nisu potrebne, jednostavno prestaje sa njihovim korišćenjem.

Osnovne komponente Cloud tehnologije su:

(i) Softver kao Cloud servis (SaaS)

Softver kao servis podrazumeva da određeni softver i aplikacija koju se koriste nisu instalirani u lokalu već na nekom drugom mestu. Treba napomenuti da SaaS predstavlja opšti naziv za model isporučivanja softvera kao servisa i ne koristi se samo u Cloud tehnologiji već se može sresti i kod drugih tehnologija.

(ii) Infrastruktura kao Cloud servis (IaaS)

Infrastruktura kao servis je virtuelni server uz koji se vežu sve relevantne usluge kao što su procesorska snaga, memorija, prostor na disku i ostalo. To znači da se korisnik brine za sve od operativnog sistema, aplikacije, antivirus softvera pa do samih podataka.

(iii) Platforma kao Cloud servis (PaaS)

Platforma kao servis je usluga koja podrazumeva korišćenje operativnog sistema putem Interneta, bez potrebe za prevlačenjem i instalacijom. Korisnik se oslanja na usluge provajdera u pogledu svega što se tiče platforme, uključujući i eventualna proširenja u kasnijoj fazi, kada mu narastu potrebe. PaaS nudi različite kombinacije usluga u Cloud-u za podršku svim fazama razvojnog ciklusa aplikacije među kojima su integrisano razvojno okruženje, kontrola izvornog koda, kontrola verzija, praćenje izmena koda, interaktivni testovi za više korisnika i ostale. PaaS rešenja su razvojne platforme u kojima su razvojni alati smešteni u Cloud i kojima se pristupa pomoću web pretraživača.

Osnovni modeli Cloud tehnologije su:

(i) Javni Cloud

Cloud provajder omogućava putem interneta pristup resursima kao što su aplikacije, skladišta za podatke i drugi resursi dostupni za javnost, nezavisno da li se radi o pojedincima ili organizacijama. Usluge mogu biti besplatne ili se koristi model plaćanja po korišćenju. Infrastruktura se nalazi u vlasništvu provajdera i nije dostupna za uvid ili kontrolu korisnicima. Infrastruktura javnog cloud-a podrazumeva deljene resurse za korisnike. Najčešće provajderi omogućavaju pristup preko Interneta, direktna komunikacija nije moguća. Delovi javnog cloud-a mogu biti i pod isključivom upotrebom samo jednog korisnika, čineći tako privatni centar podataka.

(ii) Privatni Cloud

Privatni cloud napravljen je isključivo za upotrebu jednog klijenta, rezervisan centar podataka za tog klijenta, koji može biti unutar organizacije ili hostovan od strane claud provaj-

dera. IT službe kompanija ili provajder claud usluga grade privatni cloud i upravljaju njime. Organizacije koje poseduju privatni cloud imaju potpunu kontrolu nad strukturom cloud-a.

(iii) Hibridni Cloud

Strukturu oblaka čine dva ili više različitih oblaka (privatni ili javni) koji ostaju jedinstveni entiteti, ali su međusobno povezani standardizovanim ili prikladnim tehnologijama koje omogućavaju efikasan prenos podataka ili aplikacija. Hibridni oblaci povezuju javne i privatne modele oblaka. Mogućnost proširivanja privatnog oblaka s resursima javnog oblaka može se koristiti za održavanje uslužnih nivoa kako bi se lakše izdržala velika opterećenja. Hibridni oblak se takođe može koristiti za upravljanje planiranim velikim opterećenjima. Hibridni oblaci se susreću sa složenošću određivanja kako raspodeliti aplikacije po javnom i privatnom oblaku

Cloud computing je model koji omogućava sveobuhvati, klasični, meržni pristup većem broju računarskih resursa preko Interneta. Pruža servise kao što su Softver, Platforma i Infrastrukturas. Hadoop je open source projekat koji omogućava distribuiranu obradu velikih skupova podataka u klasterima čvorova koja se zasniva na HDSF i MapReduce konceptima. Dakle, zbog brzine rasta podataka, može se zamisliti Hadoop kao servis koji se pokreće na cloud computing-u kako bi pružila distribuiranu obradu podataka. Na Slici 1 dat je prikaz jednog cloud-a sa servisima od kojih je jedan servis i Hadoop.

Hadoop kao servis (HaaS), poznat i kao Hadoop u cloud-u, predstavlja okvir koji skladišti i analizira podatke u cloud-u koristeći Hadoop. Korisnici ne moraju da ulažu ili instaliraju dodatnu infrastrukturu kada koriste ovu tehnologiju, jer je HaaS obezbeđen i upravljan vendorom odnosno provajderom cloud-a.

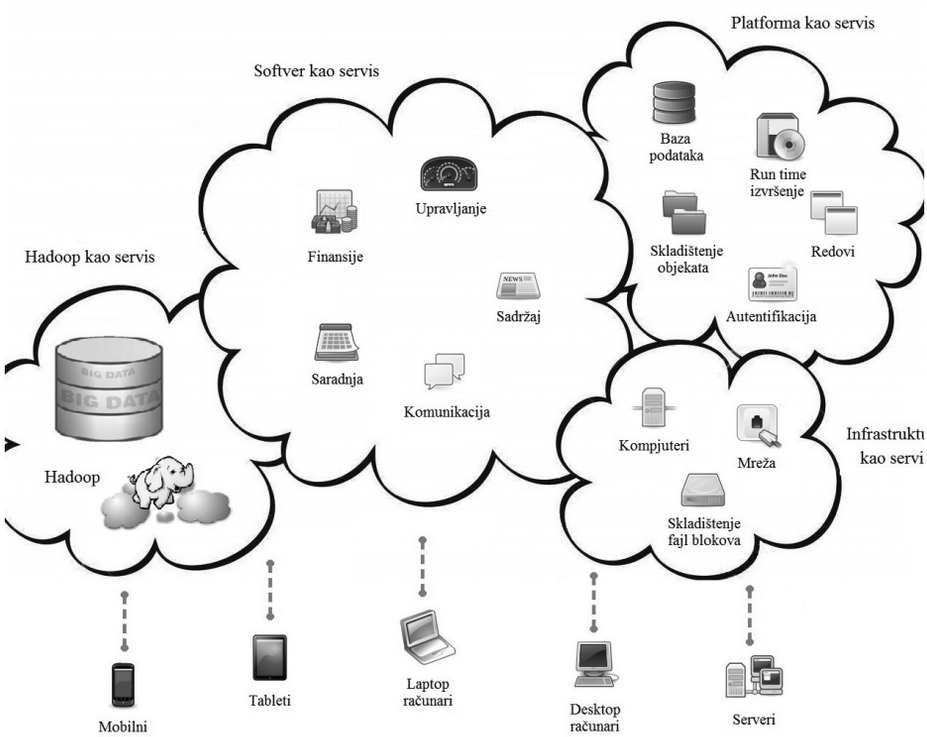
Hadoop okvir za analizu podataka omogućuje analizu velikih, nestrukturiranih skupova podataka. Hadoop-ov mehanizam za skladištenje, Hadoop Distributed File Sistem, distribuira radne zadatke na više čvorova tako da se mogu paralelno obrađivati. Jedan od nedostataka korišćenja Hadoop open source-a je da zahteva poseban skup in-house znanja i veština koje mnoge organizacije nemaju ili ne mogu sebi priuštiti. Hadoop, kao provajder usluga, integriše privatne programe sa Hadoopovim okvirom kako bi organizacijama olakšao korišćenje iobično uključuje mogućnosti upravljanja i podrške. Većina ponuda HaaS-a zasnovane su u cloud-u, a cene se određuju najčešće po klasteru po satu.

Provajderi servisa Hadoop u cloud-u (HaaS) nude različite funkcije i podršku, uključujući sledeće [5]:

– Podrška za Hadoop okruženje
– Hadoop upravljanje klasterom
– Alternativni programski jezici
– Prenos podataka između klastera
– Prilagodljive i korisničke kontrolne table i manipulacije sa podacima
– Bezbednosne funkcije.

Karakteristike servisa HaaS koje treba da obezbedi njegov provajder su sledeće [6]:

– Podaci treba da se čuvaju trajno u HDFS-u
– Elastičnost za prilagođavanje različitim radnim poslovima



Slika 1. *Hadoop u cloud-u*

– Sposobnost oporavka od neuspeha obrade bez ponovnog pokretanja čitavog procesa
– Automatsko konfigurisanje okruženja zasnovanu na poslovima.

Big data analiza i cloud servisi danas su najveći tehnološki trendovi. To su dva od četiri ključna stuba u IDC-ovoj trećoj platformi što je opisano u [7]. Sa rastućom popularnošću javnih cloud-a, kao što su AVS i Azure, preduzeća svih veličina odlučuju se na pokretanje svih ili pojedinih poslovnih zadataka u cloud-u radi postizanja poslovne agilnosti, uštede troškova i bržih inovacija.

Pošto preduzeća počinju da procenjuju analitiku na Hadoop-u, nameće se pitanje da preduzeća mogu da pokrenu Hadoop u cloud-u bez ikakvih negativnih kompromisa. Veruje se da će Hadoop dugoročno živeti u hibridnom cloud-u. U osnovnim razmatranjima o tome kako bi se uspešno primenjivao Hadoop u cloud-u, postoje tri opcije [8]:

– Opcija 1: Hadoop-kao-usluga u javnom cloud-u

Rešenja kao što su Amazon EMR (Elastic MapReduce) i Azure HDInsight obezbeđuju brz i jednostavan način pokretanja MapReduce, bez potrebe ručnog instaliranja Hadoop klastera u cloud-u.

– Opcija 2: Unapred izgrađen Hadoop u javnom cloud-u

Distribucije Hadoop-a, kao što su Cloudera CDH, IBM BigInsights, MapR, i Hortonvorks HDP mogu se postaviti u cloud-e i pokrenuti na javnim cloud-ima kao što su AVS, Rackspace, MS Azure i IBM Softlaier.

– Opcija 3: Izgradnja sopstvenog Hadoop-a u javnom cloud-u

Javni cloud-i nude rešenja Infrastruktura kao servis (IaaS) kao što su AVS i EC2 koji se primenjuju kako bi izgradili i upravljali sopstvenim Hadoop klasterom u oblaku.

Sve tri opcije mogu biti dobre za različite slučajeve korišćenja analitike i snažno dopunjavati lokalnu primenu Hadoop-a u privatnoj cloud infrastrukturi. Na primer, lokalno korišćenje Hadoop-a dobar je izbor kada su izvorni podaci na lokalnom nivou tj. na serverima u spostvenim lokalnim prostorima. Ova opcija obično zahteva ETL (extract, transform, load) iz različitih diskretnih izvora i kreće se od stotina terabajta do nekoliko petabajta u kapacitetu. Dok opcija Hadoop-a u javnom cloud-u je dobra kada se podaci generišu u cloud-u kao npr. analiziranje Tvitter podataka, i dobra je za analitiku na zahtev ako je isplativo, sigurno i lako da se redovno migriraju izvorni podaci iz lokala do cloud-a.

U 2013. godini postalo je očigledno da je glavna tema Hadoop u cloud-u. Bilo je puno novih objavljenih proizvoda i projekata koji se odnose na pokretanje Hadoop-a u cloud okruženjima. Postoje sledeći razlozi zbog čega ima smisla pokretanje ovog modela [8].

1. Smanjenje troškova inovacija
2. Brzo nabavljanje velikih resursa
3. Efikasno upravljanje poslovima
4. Upravljanje različitim potrebama za resursima
5. Izvršavanje u blizini podataka
6. Pojednostavljivanje Hadoop operacija

Veruje se da će Hadoop živeti u hibridnom cloud-u. Kada se razmišlja o Hadoop analitici u cloud-u i odabiru javnog cloud rešenja za Hadoop analitiku, razmatraju se sledeća pitanja [9]:

– Garantovanje konzistentnih performansi
– Visoka pouzdanost slična onoj kod Hadoop-u koji je za lokalnu upotrebu
– Fleksibilno i ekonomično skaliranje računarskih resursa
– Zagarantovana mrežna propusnost potrebna za Hadoop operacije
– Fleksibilno i isplativo skaliranje skladištenja
– Šifriranje podataka - Bezbednost podataka bi bila obezbeđena šifrovanjem podataka i prilikom prenosa podataka u cloud, i obrnuto, kao i prilikom mirovanja podataka, i to na nivou Hadoop klastera i na nivou fajlova tj. nivou Hadoop distibuiranog fajl sistema (HDFS). Pri šifriranju podataka koriste se različiti algoritmi od jednostavnih pa do kriptografskih algoritama kao što su: simetrični algoritmi, ssimetrični algoritmi, kriptografske funkcije za sažimanje - heš funkcije, Digitalni omot, digitalni potpis, digitalni pečat i digitalni sertifikat, kao i drugi načini zaštite [10,11].
– Lako i ekonomično dobijanje podataka iz claud-a o Hadoop analitici
– Lako upravljanje infrastrukturom

5. IZBOR REŠENJA ZA HADOOP U CLOUD-U

Hadoop kao servis u cloud-u olakšava pristup velikim aplikacijama i projektima. Ako kompanija ima puno podataka, onda bi Hadoop trebalo da bude odličan izbor za platformu njenog infomacionog sistema. Hadoop je izabran od internetskih kompanija kao što su Google i Yahoo, a kao najpopularniji i najpoznatiji veliki sistem za upravljanje podacima sada je prisutan i u preduzećima. Postoje dva velika razloga za to: Prvi razlog je što kompanije imaju mnogo više podataka za upravljanje, a Hadoop je odlična platforma, posebno za kombinovanje starih podataka i novih nestrukturiranih podataka. Drugi razlog je što mnoge softverske kuće nude podršku i usluge oko Hadoop-a čineći ga boljim za preduzeća.

Većina firmi procenjuje da od podataka koje imaju analiziraju samo 12% podataka, dok 88% podataka ostavljaju po strani. To znači da firme manji deo svojih podataka često analiziraju, i iz tog razloga firme imaju potrebu za Hadoop-om u cloud-u.

Analitičari Forrester-a, jedne od najuticajnijih istraživačkih i savetodavnih firmi na svetu koja radi sa poslovnim i tehnološkim liderima na razvijanju strategija za kompanije, Mike Gualtieri i Noel Yuhanna napisali su nedavno u Wave Report-u o tržištu Hadoop-a [3]: “Hadoop je nezaustavljiv pošto njegov open-source razvoj napreduje naglo i sve dublje u arhitekture za upravljanje podacima za preduzeća” ... “Forrester veruje

da je Hadoop neophodna platforma za podatke za velika preduzeća, čineći kamen temeljac bilo koje fleksibilne platforme za upravljanje budućim podacima. Ako kompanija ima puno strukturnih, nestrukturnih, i/ili binarnih podataka, Hadoop postaje dobar izbor za kompaniju". Forrester je analizirao i procenio devet prodavaca koji nude Hadoop usluge, i ukazao na prednosti i nedostakte svakog od njih. Takođe je zaključeno da uz ovih devet tehnoloških titana, ne postoji jasni tržišni lider među mladim kompanijama.

Hadoop kao open source projekat može svako slobodno da preuzme. Mnoge kompanije iz IBM-a na Amazon Web Services, Microsoft i Teradata su upakovale Hadoop u distribucije i servise koji su odgovarajući za korisnike. Svaka kompanija uzima malo drugačiju strategiju, ali ključna stvar je da Hadoop ima mogućnost da distribuira rad preko potencijalnih hiljada servera, obezbeđujući da veliki podaci budu podaci kojima se može upravljati. To su sledeće kompanije: Amazon Web Services, Cloudera, Hortonworks, IBM, Intel, MapR Technologies, Microsoft, Pivotal Software i Teradata.

6. ZAKLJUČAK

Hadoop kao open source okvir koji obezbeđuje skladištenje i obradu velikih podataka u distribuiranom okruženju preko klastera računara i koji omogućava pretraživanje i analizu velikih količina podataka ima sledeće prednosti: mogućnost da se bilo koja vrsta velikih podataka može skladištiti i obrađivati, velika računarska snaga, visoka pouzdanost, fleksibilnost, mali troškovi, prilagodljivost. Ali postoje sledeći izazovi usled upotrebe Hadoop platforme: MapReduce programiranje nije podjednako odgovarajuće za sve probleme, postoji široko prihvaćen probelem nedostatka talenata, bezbednost podataka, potpuno upravljanje podacima. Upravo zbog toga teži se smeštanju Hadoop u cloud, a i iz sledećih pogodnosti cloud računarstva: smanjenje troškova inovacija, efikasno upravljanje poslovima, upravljanje različitim potrebama za resursima, izvršavanje u blizini podataka, pojednostavljivanje Hadoop operacija, elastičnost za prilagođavanje različitim radnim poslovima, sposobnost oporavka od neuspeha obrade bez ponovnog pokretanja čitavog procesa, automatsko konfigurisanje okruženja zasnovano na poslovima.

Gledano sa strane korisnika, najbitnije je to što cloud nudi neograničene količine svih resursa onda kada su korisniku zaista potrebni i u onoj meri koja odgovara korisničkim potrebama, i to tako da sve korisnik može samostalno da kontroliše. Korisnik nije vlasnik infrastrukture, samim tim nije dužan ni da je održava. On plaća iznajmljivanje usluga i onih resursa koji su mu u određenom trenutku potrebni. Ukoliko mu je potrebna neka poslovna aplikacija, on je iznajmi i nema potrebe da je kupuje. Korisnik plaća samo ono šta koristi i onoliko koliko to koristi. Kada mu usluge za koje se opredelio više nisu potrebne, jednostavno prestaje sa njihovim korišćenjem.

Korisnicima se nude rešenja Hadoop u cloud-u svetski poznatih kompanija među kojima su Hortonworks, Cloudera i MapR Technologies, jer one kao primarnu delatnost imaju Hadoop, ali i druge kompanije kao što su IBM, Intel, AWS, Microsoft, Teradata koje su poznatije po po nekim drugim tehnologijama.

Tehnologije Hadoop u cloud-u mogu se primeniti u brojnim oblastima IT: u poslovanju kompanija i preduzeća, u medicini, obrazovanju, u transportu i saobraćaju, u meteorologiji, u naučnim istraživanjima, u javnoj upravi, u finansijskim institucijama kao što su banke i osiguravajuće kuće itd. IT sistemi predstavljaju senzorske mreže u kojima se generiše velika količina podataka. Primena Hadoop u cloud-u u IT rešenjima omogućuje pouzdano, distribuirano i skalabilno čuvanje i korišćenje velikih količina podataka, uz očuvanje bezbednosti podataka i privatnosti korisnika.

Veruje se da će dugoročno rešenje biti Hadoop u hibridnom cloud-u koje će za različite slučajeve upotrebe koristiti prednosti i opcije sa lokalnim data centrom i opcije sa raspoređivanjem u cloud-u.

REFERENCE

[1] A. McAfee, E. Brynjolfsson, "Big Data: The Management Revolution", Harvard Business Review, Harvard, October 2012;
[2] G. Menegaz, "What is Hadoop, and how does it relate to cloud?", IBM Cloud computing news, Harvard Business Rewiev, Harvard, May 2014;
[3] B. Butler, "Nine Hadoop companies you should know", Network World, Framingham, March 2014;
[4] N. Carr, "Cloud Computing", Accademic Room, Harvard Innovation Lab, Harvard, 2013;
[5] M. Rouse, "Hadoop as a service (HaaS)", TechTarget - SearchCloudStorage, Atlanta, April 2016;
[6] B. Golden, "As IDC Sees It, Tech's 'Third Platform' Disrupts Everyone", CIO, Framingham, March 2014;
[7] S. Sinha, A. Josh, "8 Questions to Ask about Hadoop in the Cloud", TDWI, Renton, January 2016;
[8] H. Yamijala, "6 Reasons Why Hadoop on the Cloud Makes Sense", Thought Works, Chicago, November 2013;
[9] G. Piatetsky-Shapiro, "Hadoop as a Service: 18 Cloud Options", KDnuggets, GTE Labs, Stamford, April 2015;
[10] Ž. Gavrić, V. Mišković, D. Starčević, "Tehnologije vodenog žiga", Info M, Vol. 60/2016, Fakultet organizacionih nauka, Beograd, 2016;
[11] P. Čisar, "Opšti aspekti kvantne kriptografije", Info M, Vol. 54/2015, Fakultet organizacionih nauka, Beograd, 2015.

Dejan Hadzi-Milosavljević, Raiffeisen banka Srbije
Kontakt: dejan.hadzi-milosavljevic@raiffeisenbank.rs
Oblasti interesovanja: Računarske mreže, Informacione tehnologije, Projektovanje informacionog sistema za bankarsko poslovanje

Dušan Starčević, Fakultet organizacionih nauka Univerziteta u Beogradu
Kontakt: dusan.starcevic@fon.bg.ac.rs
Oblasti interesovanja: Računarske mreže, Multimedijalne komunikacije, Multimedije