Univerzitet u Beogradu
Matematički fakultet

SEMINARSKI RAD

Metodologija stručnog i naučnog rada

Tema:

Big Data

Studenti:
Nikola Stanković, 1033/2012
Dragan Đurđević, 1059/2012
Marko Makarić, 1090/2012
Srđan Terzić, 1054/2012

Profesor:
dr Vladimir Filipović

SADRŽAJ

1 UVOD

Upravljanje i analiza podataka oduvek je predstavljala najveći izazov za sve organizacije u svim poljima industrije. Preduzeća su se dugo borila da pronađu pragmatičan pristup za sakupljanje informacija o svojim klijentima, proizvodima i uslugama. Kada su kompanije imale samo šačicu kupaca koji su kupovali isti proizvod na isti način stvari su bile prilično jasne i jednostavne. Vremenom, preduzeća i tržišta su porasla pa je stvar postala mnogo komplikovanija. Da bi preživele ili stekle neku prednost nad klijentima u odnosu na konkurenciju, ove kompanije su dodavale nove linije proizvoda i menjale način pružanja usluga.

Problemi oko podataka nisu ograničeni samo na polju preduzeća. Na primer, organizacije koje se bave razvojem i istraživanjem imale su problema da dobiju dovoljno računarske moći da bi pokrenule sofisticirane modele ili obradile slike i druge izvore naučnih podataka. Zaista, suočavamo se sa mnogo problema kada je reč o podacima. Neki podaci su struktuirani i sačuvani u relacionim bazama podataka dok su neki drugi podaci, uključujući dokumenta, slike i video zapise, nestruktuirani. Kompanije takođe moraju da razmotre nove izvore podataka koje generišu mašine kao što su senzori. Drugi izvori informacija su oni koji generišu ljudi kao što su podaci iz društvenih medija i click-stream podaci dobijeni sa raznih sajtova. Pored toga, dostupnost i prihvatanje novih, moćnijih mobilnih uređaja, uz stalan pristup globalnoj mreži dovešće do novih izvora podataka.

Iako se svaki izvor podataka može nezavisno upravljati i pretraživati, trenutno je za kompanije najveći izazov da nađu smislen presek svih tih podataka različitih tipova. Kada imate toliko informacija u toliko različitih oblika, nemoguće je razmišljati o upravljanju podacima na tradicionalan način. Iako smo oduvek imali mnogo podataka, razlika je u tome što danas većina toga postoji, a varira samo u vrsti i načinu obrade. Organizacije, više nego ikada ranije, pronalaze način da iskoriste ove informacije. Dakle, o upravljanju podacima mora se misliti drugačije i to je izazov, a ujedno i šansa, za big data-u.
Big data se može definisati kao bilo koja vrsta izvora podataka koja ima najmanje sledeće tri zajedničke karakteristike:

1. Izuzetno velika količina podataka
2. Izuzetno velika brzina podataka
3. Izuzetno široka raznovrsnost podataka

Big data je važna zato sto omogućava organizacijama da sakuplja, skladišti, upravlja i obrađuje velike količine podataka velikom brzinom. Big data nije samostalna tehnologija, nego je to kombinacija poslednjih 50 godina evolucije tehnologije.

2 ŠTA JE BIG DATA

Big data ne predstavlja jedinstvenu tehnologiju, već kombinaciju novih i starih tehnologija koje pomažu kompanijama da steknu delotvoran uvid u obrađene podatke. U stvari, Big data predstavlja mogućnost upravljanja velikim količinama različitih podataka razumnom brzinom i u odgovarajućem vremenskom okviru da bi se omogućila analiza tih podataka u realnom vremenu. Kao što smo ranije napomenuli, za big data su karakteristične tri stvari:

- Količina: Koliko podataka
  Mnogo faktora doprinosi uvećanju obima podataka (transakcioni podaci skladišteni godinama, tekstualni podaci koji konstantno nadolaze sa društvenih mreža, itd.). U prošlosti je prekomerna količina podataka stvarala probleme oko skladištenja, ali sa današnjim cenama memorijskih uređaja to više ne predstavlja problem. Ipak, drugi problemi se javljaju, uključujući određivanje važnosti određenih podataka u velikoj gomili.

- Brzina: Koliko brzo su podaci obrađeni
  Brzina obrade podataka predstavlja dve stvari. Prva je brzina proizvodnje i generisanja podataka, a druga je brzina kojom podaci moraju biti obrađeni da bi zadovoljili određene kriterijume. Pravovremeno reagovanje i brza obrada podataka predstavljaju veliki izazov i za najveće kompanije na svetu.

- Raznovrsnost: Koliko različitih tipova podataka imamo
  Danas se podaci nalaze u velikom broju različitih formata. Tu imamo 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.

Takođe, kada govorimo o karakterisitkama, bitno je napomenuti još dve bitne dimenzije:

- 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. U velikom broju slučajeva je pogubno uparivati, pročišćavati i transformisati te podatke na bilo koji način. Ipak, neophodno je izvršiti povezivanje odnosa među podacima i hijerarhijama podataka, jer u suprotnom količina podataka može da izmakne kontroli.

Podaci se dobavljaju iz gomile 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. Pod ovim nazivima podrazumevamo sledeće:

- 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 reći da su grupisani u neku fiksiranu shemu. Podaci su često nerazdvojivi i sadrže oznake koje pomažu pri hijerarhijskom organizovanju ovakvih podataka.
- Nestruktuirani podaci su uglavnom podaci koje je teško ubaciti u relacione tabele baza podataka radi analize ili ili izvršavanja upita nad njima. Podaci ovakvog tipa predstavljaju slike, audio i video fajlove.

3 GDE SE KORISTI

Razvoj tehnologija koje se koriste za obradu velikih količina podataka doprineo je razvoju pojedinih oblasti gde se takve analize mogu iskoristiti. Na primer, veliki napredak se vidi u oblasti zdravstva ili saobraćaja. U zdravstvu, može se pratiti broj prevremeno rođene dece i u zavisnosti od dobijenih podataka procenjivati kada je potrebna određena intervencija. Kod saobraćaja, analizom velike količine podataka koje generišu kamere postavljene na autoputevima, moguće je predvideti i regulisati gužve i zakrčenja. Takođe, može se smanjiti broj saobraćajnih nezgoda, štedeti gorivo , pa čak voditi računa i o zagađenju.

Ipak, glavni problem ne predstavlja prikupljanje velikih količina podataka(oni su već oko nas), već izvlačenje korisnih informacija iz tih podataka. Današnje tehnologije ne samo da podržavaju skladištenje ovih podataka već daju mogućnost da se dobijeni podaci razumeju i da se iskoristi njihova vrednost. Ovo pomaže organizacijama da poprave svoje poslovanje i profit. Na primer, uz pomoć ovih tehnologija moguće je:

- Analizirati milione tržišnih proizvoda da bi se odredila optimalna cena, uvećao profit ili oslobodilo skladište.
- Preračunavati rizike u minuti i na taj način se prilagođavati promenama.
- Istraživanje podataka vezanih za potrošačke navike i potrebe i na taj način povećavanje profita, podrške u izbornim kampanjama itd.
- Identifikovanje najozbiljnijh kupaca.
- Generisanje maloprodajnih kupona za potrošače, baziranih na prethodnim kupovinama. Ovo osigurava veći otkup robe.
- Slati adekvatne ponude mobilnih provajdera na mobilne telefone u pravom trenutku, kada će korisnik moći da ih iskoristi na najbolji način.
- Analiziranje podataka sa sredstava javnog informisanja zbog sagledavanja trendova.
- Određivanje glavnih problema u funkcionisanju mreža i mašinskih senzora.

Klasični primeri generisanja velikih količina podataka:

- Sistemi radio frekvencija generišu 1000 puta više podataka od tradicionalnih bar kod sistema.
- 10 000 transakcija plaćanja kreditnom karticom se obavi svake sekunde u svetu.
- Walmart obrađuje više od milion korisničkih transakcija u satu.
- 340 miliona tvitova se pošalje dnevno. To je približno 4 000 tvitova u sekundi.
- Facebook ima više od 901 miliona aktivnih korisnika koji svakodnevno generišu podatke svojom međusobnom interakcijom.
- Više od 5 milijardi ljudi zove, šalje poruke tvituje i surfuje internetom na mobilnim uređajima.

4 IZVORI PODATAKA

4.1 Izvori struktuiranih podataka

Iako se čini da su struktuirani podaci dobro poznati, zapravo, struktuirani podaci u svetu Big data pristupa dobijaju novu ulogu. Razvoj tehnologije omogućava pojavu novih izvora struktuiranih podataka - često u realnom vremenu i u velikim količinama. Izvori podataka se dele u dve kategorije:

- Računarski ili mašinski generisani: pojam mašinski generisanih podataka se obično odnosi na podatke koje proizvodi mašina bez ljudskog uticaja.

- Ljudski generisani: ovo su podaci koje obezbeđuju ljudi u interakciji sa računarima.

Neki stručnjaci tvrde da postoji i treća kategorija koja predstavlja hibrid izmedju dve navedene kategorije. Medjutim, ovde će nas interesovati samo navedene.

Mašinski generisani struktuirani podaci mogu da uključuju:

- Senzorske podatke:
  Primeri uključuju radio frekvencijske ID (RFID) oznake, pametne merače (npr. elektronska brojila za merenje potrošnje električne energije), podatke medicinskih uredjaja, GPS podatke. Na primer, RFID ubrzano postaje popularna tehnologija. Koriste se minijaturni računarski čipovi da bi se uređaji pratili sa udaljenosti. Primer ovoga je praćenje kontejnera sa proizvodima od jedne do druge lokacije. Kada prijemnik dobije informacije one mogu biti prosleđene serveru gde će biti analizirane. Kompanije su zainteresovane za ovu tehnlogiju zbog upravljanja transportom robe i kontrolu inventara. Još jedan primer izvora senzornih podataka su pametni telefoni koji imaju senzore kao sto je GPS koji mogu biti korišćeni za razumevanje ponašanja potrošača na novi način.

- Web log podatke:
  Kada serveri, aplikacije, mreže i slično rade oni beleže različite podatke o svojoj aktivnosti. Količina ovih podataka može postati ogromna, a ovi podaci mogu biti iskorišćeni za, na primer, predviđanje narušavanja bezbednosti.

- Podatke u trenutku prodaje:
  Kada radnik na kasi očita bar kod bilo kog proizvoda koji kupujete, generišu se svi podaci vezani za proizvod. Ako se razmisli koliko ljudi svakodnevno kupuje različite proizvode može se shvatiti koliko je količina ovih podataka ogromna.

- Finansijske podatke:
  Dosta finansijskih sistema su danas programirani, njihov rad se zasniva na predefinisanom skupu pravila što automatizuje proces. Podaci o trgovanju na berzi su dobar primer ovoga. Sadrže struktuirane podatke kao sto su oznaka kompanije i vrednost u dolarima. Neki od ovih podataka su mašinski generisani a neki ljudski generisani.

Primeri ljudski generisanih struktuiranih podataka mogu da uključuju:

- Ulazne podatke: Ovo je bilo koji tip podataka koji čovek može uneti u računar, kao što je ime, prezime, godine starosti, prihod, odgovori na ankete i slično. Ovi podaci mogu biti korišćeni za razumevanje osnovnog ponašanja potrošača.

- Klik podatke: svaki put kada se klikne na link na sajtu podaci se generišu. Ovi podaci mogu biti analizirani da bi se odredilo ponašanje potrošača i obrasci kupovine.

- Podatke vezane za igre: svaki potez koji se napravi u igri može biti zabeležen. Ovi podaci mogu biti korisni za razumevanje kako krajnji korisnici igraju igru.

Neki od ovih podataka ne moraju biti veliki sami po sebi, kao što su profilni podaci. Međutim, kada se objedine podaci miliona korisnika koji šalju informacije, količina podataka postaje ogromna. Dodatno, mnogo ovih podataka je vezano za vreme u kom se generišu što može biti korisno za razumevanje obrazaca koji imaju potencijal za predviđanje ishoda. Poenta je da ove informacije mogu biti moćne i mogu biti korišćene u različite svrhe.

4.2 Izvori nestruktuiranih podataka

Nestruktuirani podaci su podaci koji ne prate neki definisani format. Ako je 20% podataka koji su dostupni preduzećima struktuirano, preostalih 80% je nestruktuirano. Nestruktuirani podaci su zapravo podaci koji se najčešće sreću. Do skoro, međutim, tehnologija nije podržavala druge načine rada sa ovim podacima osim skladištenja i ručne obrade. Nestruktuirani podaci se mogu naći svuda. Zapravno, većina ljudi i organizacija funkcioniše na osnovu nestruktuiranih podataka. Kao i u slučaju struktuiranih podataka i nestrukturirani podaci mogu biti mašinski ili ljudski generisani. Neki primeri mašinski generisanih nestruktuiranih podataka su:

- Satelitske slike: Ovo uključuje podatke o vremenskim prilikama ili podatke koje vlade prikupljaju prilikom satelitskog nadgledanja. Na primer, GoogleEarth poseduje ogromnu količinu satelitskih snimaka koje obrađuje i spaja na odgovarajući način.
- Naučni podaci: ovo uključuje seizmičke slike, atmosferske podatke, fiziku visokih energija, itd.

5 Tehnologija

Veliki broj novih tehnoloških dostignuća omogućava organizacijama da iskoriste veliku količinu podataka kao i da ih efikasno analiziraju. Neke od karakteristika su:

- Jeftino i veliko skladište za podatke, uz mogućnost serverske obrade.
- Brži procesori.
- Dostupne mogućnosti za veliku memoriju, kao što je Hadoop.
- Nove tehnologije vezane za skladištenje i obradu podataka, namenjene baš za velike i obimne podatke, uključujući i nestruktuirane podatke.
- Paralelnu obradu, klasterovanje, MPP, virtualizaciju, velika grid okruženja, visok nivo propusnosti i mogućnosti povezivanja
- Rad u “oblaku” i druga fleksibilna rešenja za rad sa resursima.

Tehnologije koje se svrstavaju pod “Big data” tehnologije ne podržavaju samo mogućnost prikupljanja velike količine podataka, one daju mogućnost za razumevanje tih podataka kao i izvlačenje nekih vrednosti. Glavni cilj svih organizacija koje imaju pristup kolekcijama velikih podataka trebalo bi da bude to da iskoriste većinu relevantnih podataka u svom poslovanju za donošenje raznih poslovnih odluka.

Sa razvojem računarskih tehnologija, danas je moguće upravljati ogromnim količinama podataka, koje su ranije mogle da se obrađuju i koriste jedino uz pomoć superračunara i to uz veliki trošak. Cene sistema su opale i kao rezultat nove tehnike za distribuiranu obradu su trenutno u fokusu upotrebe. Pravi proboj u tehnologiji Big data desio se kada su kompanije kao što su Yahoo!, Google, i Facebook došle do saznanja da mogu da zarade od velikih količina podataka koje su njihovi proizvodi generisali. Ove kompanije su bile pred zadatkom da nađu način u vidu nekih novih tehnologija koje će im omogućiti da čuvaju, pristupaju, obrađuju i analiziraju ogromne količine podataka u realnom vremenu, na takav način da mogu prilično da zarade i na pravi način iskoriste količinu podataka koju poseduju i koji učestvuju u njihovim mrežama. Njihova rešenja koja su nastala su dovela do promena na tržištu upravljanja podacima. Posebno, novine koje su doneli MapReduce, Hadoop i Big Table pokazale su se kao varnice koje su dovele do neke nove generacije upravljanja podacima. Ove tehnologije apostrofiraju jedan od najfundamentalnijih problema, a to je sposobnost obrade velikih količina podataka na efikasan i blagovremen način, na način koji je isplativ i koji ne zahteva velike troškove.

5.1 MapReduce

MapReduce je rešenje koje je predstavio Google kao način efikasnog izvršavanja skupa funkcija nad ogromnim količinama podataka na serijski način. Komponenta “map” raspoređuje programerski problem ili zadatak na veliki broj sistema i rukovodi postavljanju zadataka na način koji podrazumeva balansirano opterećenje i upravlja oporavkom od grešaka. Nakon što završi distribuirana obrada, poziva se druga funkcija nazvana “reduce”, koja spaja sve elemente nazad zajedno, da bi obezbedila rezultat. Jedan primer MapReduce upotrebe mogao bi da bude zadatak da se odredi koliko stranica knjige je napisano na svakom od nekih 50 različitih jezika. MapReduce je programerski model za obradu velikih skupova podataka pomoću paralelnih, distribuiranih algoritama u klasteru. Jedan MapReduce program obuhvata Map() proceduru koja vrši filtriranje i sortiranje (na primer sortiranje studenata po imenu u redove, po jedan red za svako ime) i Reduce() procedura koja vrši operaciju agregacije (na primer broj studenata u svakom redu). MapReduce sistem (može se reći i infrastruktura ili frejmvork) upravlja distribuiranim serverima i uopšteno celim procesom. Sistem izvršava različite zadatke paralelno, upravlja svim komunikacijama kao i prenosu podataka između različitih delova sistema, u isto vreme obezbeđujući sistem od redundantnosti i grešaka. Inspiracija za model je proistekla iz map i reduce funkcija koje se često koriste u funkcionalnom programiranju iako njihova uloga u MapReduce sistemu nije ista kao što je u njihovom originalnom obliku. MapReduce biblioteke se pišu na raznim programskim jezicima. Besplatna implementacija koja je popularna je Hadoop organizacije Apache.

5.2 Big Table

Big Table je rešenje razvijeno od strane kompanije Google, kao distribuirani sistem za skladištenje podataka koji je predviđen da upravlja veoma skalabilnim struktuiranim podacima. Podaci su organizovani u tabele sa redovima i kolonama. Za razliku od tradicionalnog relacionog modela baze podataka, Big Table predstavlja proređenu, distribuiranu i trajnu sortiranu višedimenzionu mapu. Big Table je namenjen za čuvanje velikih količina podataka na običnim serverima. Big Table mapira dva proizvoljna stringa (ključ koji se odnosi na red i ključ koji se odnosi na kolonu) i vremenski trenutak (dakle imamo trodimenzionalno mapiranje) u neki vezani niz bitova. Big Table je predviđen da može da ide do nivoa petabajta, rad na preko stotinu hiljada mašina koji omogućava jednostavno dodavanje novih mašina u sistem i njihovo momentalno uključenje u rad na način koji ne zahteva nikakvo ponovno konfigurisanje ili prekid u radu sistema.

5.3 Hadoop

Inovatori sistema za pretraživanje kao što su Yahoo! i Google su bili pred zadatkom da nađu način kako da izvuku smisao i neku vrednost iz ogromnih količina podataka koje njihovi sistemi prikupljaju. Ove kompanije su bile pred izazovom da u isto vreme razumeju koje informacije prikupljaju, kao i kako da te informacije uklope u svoje poslovanje i poboljšaju svoje poslovanje, a samim tim i prihode. Hadoop dozvoljava kompanijama da na lak način upravljaju velikim količinama podataka. Hadoop omogućava da veliki problemi budu razbijeni na manje tako da analiza može da se izvrši brzo i jeftino. Razbijanjem tih velikih problema na manje delove koje je posle moguće obrađivati paralelno, i po završetku obrade te informacije se prikupljaju i grupišu radi izdavanja krajnjih rezultata. Hadoop je softverski frejmvork izveden iz MapReduce i BigTable sistema. Hadoop dozvoljava aplikacijama baziranim na MapReduce sistemu da se izvršavaju na velikim klasterima običnog hardvera. Hadoop je dizajniran da paralelizuje obradu podataka koristeći čvorove za povećanje brzine izračunavanja i smanjenje odziva. Hadoop se sastoji od dve glavne komponente, visoko skalabilnog distribuiranog fajl sistema koji podržava i količinu podataka koja se meri u petabajtima,dok je druga komponenta MapReduce sistem.

6 ZAKLJUČAK

Godinama su organizacija sakupljale transakciono struktuirane podatke i koristile batch obradu da stave reprezentativne uzorke u tradicionalnu relacionu bazu podataka. Analiza ovakvih podataka je retrospektivna i istraživanje se vrši na skupovima podataka. Poslednjih nekoliko godina, nove tehnologije su omogućile poboljšano sakupljanje, skladištenje i analizu podataka po jeftinijoj ceni. Organizacije sada mogu sakupiti više podataka iz mnogo više izvora (blogovi, audio i video fajlovi). Opcije za optimalno skladištenje i obradu podataka su se drastično proširile i tehnologije, kao što su MapReduce i in-memory computing, obezbeđuju visoko optimizovane mogućnosti za različite poslovne svrhe. Analiza podataka može biti izvršena u realnom vremenu ili veoma blizu realnog vremena obrađujući ceo skup podataka a ne reprezentativne uzorke. Dodatno, broj opcija da se tumače i analiziraju podaci se takođe povećao uz korišćenje različitih tehnologija za vizuelizaciju. Svi ovi izumi predstavljaju kontekst u koji je smešten big data. Big data obično obuhvata skupove podataka veličine daleko veće za obradu od onih sa kojima rade najčešće korišćeni softverski alati čiji je cilj da prikupljaju, upravljaju i procesiraju podatke u određenom periodu vremena koje ima neki prag tolerantnosti. Veličine koje se pominju kada se pomene big data

predstavljaju nešto slično kao pokretna meta, jer samo gledajući 2012. godinu te veličine se kreću od nekoliko desetina terabajta do nekoliko petabajta podataka koji se nalaze u jednom skupu podataka. Ta meta nastavlja da se kreće zahvaljujući konstantnom napretku i u tradicionalnim relacionim sistemima i u novim bazama podataka kao što je NoSQL i njihovim sposobnostima da rukuju sa sve većim količinama podataka.

7 REFERENCE

- Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman - *Big Data for Dummies*, 2013.
- Srinath Perera, Thilina Gunarathne - *Hadoop MapReduce Cookbook*, 2013.
- Wikipedia, Big Data page: http://en.wikipedia.org/wiki/Big_data
- SAS Big Data page: http://www.sas.com/big-data/