BIG DATA

Almedina Hatarić, MA; e-mail: almedina.hataric@iu-travnik.com
Nehad Gaši, BA; e-mail: nehad.gasi@iu-travnik.com
Internacionalni univerzitet Travnik, Travnik, Bosna i Hercegovina

Sažetak: "Big Data" je popularni termin koji se koristi da bi se opisao eksponencijalni rast i dostupnost struktuiranih i nestruktuiranih podataka. S obzirom da veći broj podataka dovodi do tačnijih analiza, "Big Data" ima veliki značaj i za poslovanje i za društvo, kao i sam internet. "Big Data" ne predstavlja jedinstvenu tehnologiju, već komunikaciju novih i starih tehnologija koje pomažu kompanijama da steknu djelotvoran uvid u određene podatke. Koncept "Big Data" sadrži skup povezanih komponenti koje omogućavaju organizacijama da koriste podatke za praktične potrebe i rešavaju niz poslovnih problema. Ovo uključuje IT infrastrukturu potrebnu za podršku "Big Data"; analitiku koja se primenjuje na podatke; tehnologiju potrebnu za projekte "Big Data"; povezane skupove vještina; i stvarne slučajeve za koje je potrebno da se koristi "Big Data".
Ključne riječi: podaci, informacione tehnologije, poslovanje, Internet, organizacija

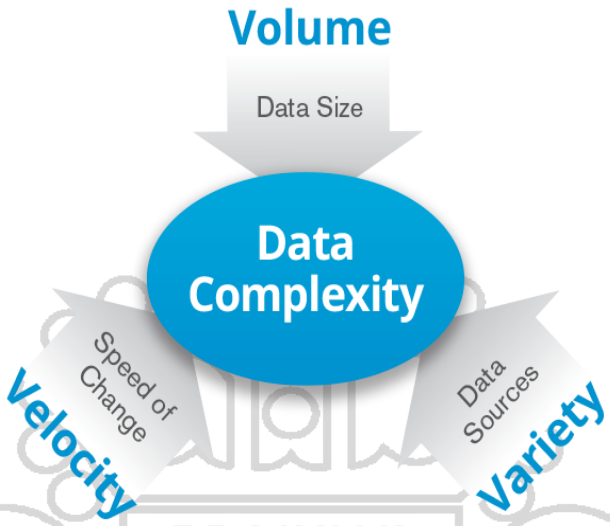
BIG DATA

Abstract: Big Data is a popular term used to describe the exponential growth and availability of structured and unstructured data. As more data leads to more accurate analyzes, Big Data is of great importance both to business and society, as well as to the Internet itself. Big Data is not a unique technology, but a communication of new and old technologies that help companies gain effective insight into certain data. The Big Data concept includes a set of related components that enable organizations to use data for practical needs and solve a variety of business problems. This includes the IT infrastructure needed to support Big Data; data analytics; technology required for Big Data projects; related skill sets; and the actual cases that need to use Big Data.
Keywords: Big Data, IT, business, Internet, organization

1. UVOD

U današnje vrijeme, broj korisnika Interneta kontinuirano se povećava te je život bez njega postao nezamisliv. Internet je, između ostaloga, omogućio brži i efikasniji prijenos podataka, povezivanje ljudi te stvaranje novih poslovnih modela. Internet je također, zajedno sa razvojem tehnologije poluvodiča i komunikacijske tehnologija, omogućio povezivanje fizičkih objekata u koje su ugrađeni senzori, programi i mogućnost povezivanja, što se naziva Internet objekata. Big Data postali su vrlo popularna tema u području brojnih znanosti. Razvojem tehnologije stvorili su se uvjeti u kojima se granica memorijskih ograničenja sistema polako gubi, vremenski intervali između generiranja podataka mogu biti gotovo pa proizvoljno mali, a podaci ne moraju postovati neku strogu strukturu. Sistemi za rad s velikim podacima mogu pohranjivati podatke koji pristižu u masivnoj količini, veoma brzo i u različitim oblicima. Ništa od navedenog više ne predstavlja problem. Ako postoji potreba za pohranom vee količine podataka, system se može lagano proširiti. To svojstvo naziva se skalabilnost i zajedničko je svim sistemima za rad s velikim podacima. S obzirom da veći broj podataka dovodi do tačnijih analiza, „big data" ima veliki značaj i za poslovanje i za društvo, kao i sam internet. Preciznije,

odnosno tačnije analize dovode do pouzdanijih odluka što može značiti veću operativnu efikasnost, smanjenje troškova i smanjenje rizika. Sada već učestala definicija spominje 3V dimenzije „Big data“:



Ilustracija 1: 3V Big Data

1. **Volume** (količina)

Mnogi faktori doprinose povećavanju obima podataka. Podaci bazirani na transakciji pohranjivani tokom godina. Nestruktuirani podaci proizišli iz društvenih mreža. Povećanje količine senzora i „machine-to-machine“ podataka koji se prikupljaju. U prošlosti je postojao problem skladištenja prekomjernog obima podataka. Ali sa smanjenjem troškova skladištenja pojavljuju se druga pitanja, npr. kako odrediti relevantne u okviru velikog broja podataka i kako koristiti analitiku za stvaranje vrijednosti od relevantnih podataka.

2. **Velocity** (brzina)

Podaci protiču neviđenom brzinom i moraju se blagovremeno obraditi. RFID tagovi, senzori i pametno mjerenje (smart metering) dovode do potrebe da se bave disperzijom podataka u skoro realnom vremenu. Za mnoge organizacije je izazov da reaguju dovoljno brzo na brzinu podataka.

3. **Variety** (raznovrsnost)

U današnje vrijeme podaci dolaze u različitim formatima. Struktuirani, brojčani podaci u tradicionalnim bazama podataka. Informacije kreirane od „line-of“ poslovnih aplikacija. Nestruktuirani tekstualni dokumenti, e-mail, video, audio, podaci o dionicama i finansijske transakcije. Upravljanje, spajanje i uređivanje različitih vrsta podataka je nešto sa čime se mnoge organizacije još uvijek bore.

Međutim, neki uzimaju u obzir još dvije vrste dimenzija:

4. **Variability** (promjenljivost)

S obzirom na povećanje brzine i raznovrsnosti podataka, njihovi tokovi mogu biti u nedoslijednosti sa periodičnim rastom . Dnevni, sezonski i event-triggered nagli porast protoka podataka mož biti izazovni za upravljanje. Čak i više ukoliko su nestrukutirani podaci uključeni.

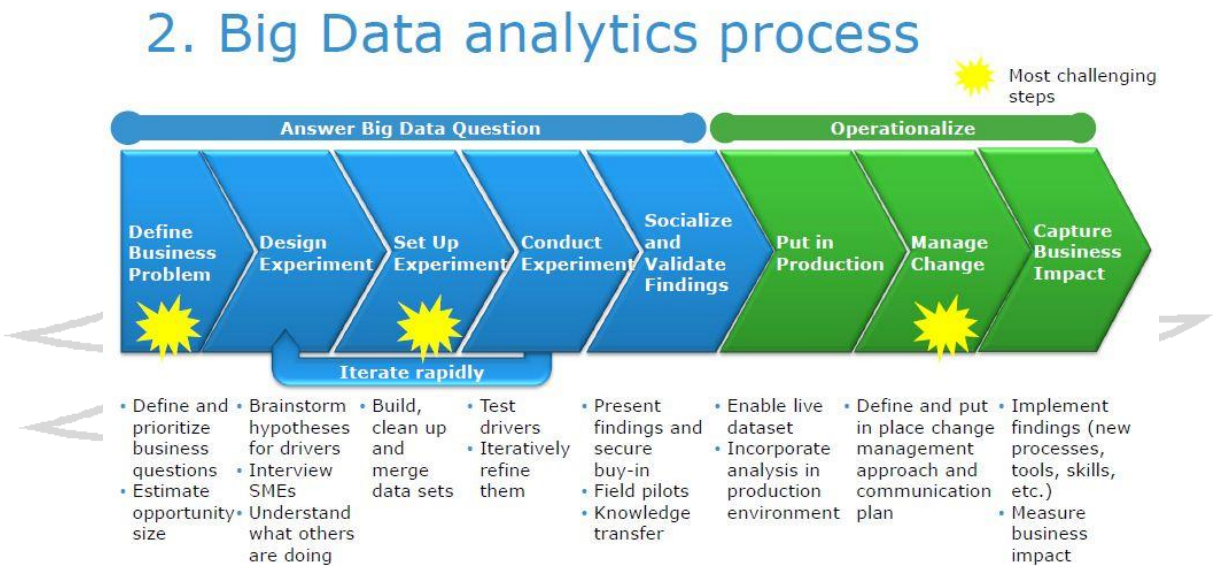
5. **Complexity** (složenost)

U današnje vrijeme podaci dolaze iz raznovrsnih izvora. Još uvijek je veliki poduhvat povezati, očistiti i transformisati podatke kroz sistem. Međutim, neophodno je da se podaci povežu, da se hijerarhijski poredaju ili se mogu naći van kontrole.

2. UPOTREBA BIG DATA

Osnovno pitanje nije dobijanje velike količine podataka, već upotreba podataka koji su uračunati. Obećavajuća vizija je da će organizacije moći uzimati podatke iz bilo kojeg izvora, prikupiti relevantne podatke i analizirati ih da bi se pronašli odgovori koji omogućavaju smanjenje troškova, smanjivanje vremena, razvoj novih proizvoda i optimiziranje ponuda, kao i pametnija odlučivanja u toku poslovanja. Na primjer, kombinovanjem „Big data“ i „high-powered“ analitika, moguće je:

- Utvrditi uzroke neuspjeha, probleme i nedostatke u najbržem mogućem periodu, te tako potencijalno sačuvati milione godišnje;
- Optimizovati rute za hiljade dostavljačkih vozila dok su još na putu;
- Proizvesti maloprodajne kupone na mjestu prodaje na osnovu prošlih i sadašnjih kupovina kupca;
- Poslati prilagođene preporuke na mobilne uređaje dok su kupci na pravom mjestu da iskoriste prednosti te ponude;
- Preračunati kompletan rizik portofolija u nekoliko minuta;
- Brzo identifikovati najvažnije kupce;
- Koristiti „clickstream“ analizu i „data mining“ za otkrivanje prevara.



3. IZVORI BIG DATA

Izvori „Big data“ se svake godine povećavaju, ali uglavnom spadaju u jednu od tri grupe:

1. **Streaming data,** što uključuje podatke koji dolaze do IT sistema sa mreže ili povezanih uređaja. Organizacija može da analizira ove podatke čim stignu i može donositi odluke o tome koje podatke da zadrži, koje ne, te šta zahtijeva dalje analize.
2. **Podaci sa društvenih mreža,** koji predstavljaju sve atraktivniji izvor informacija, naročito za marketing, prodaju i funkcije podrške. Ovi podaci su često u nestruktuiranim ili polustruktuiranim oblicima, tako da i pored tolikog obima podataka analiza i korištenje informacija predstavlja jedinstven izazov.
3. **Javno dostupni izvori,** kao što su CIA World Factbook, ili Portal otvorenih podataka Evropske Unije (European Union Open Data Portal) predstavljaju izvor ogromne količine podataka.

4. BIG DATA TEHNOLOGIJE

Organizacijama je omogućeno da koriste veliku količinu podataka zahvaljujući velikom broju tehnoloških dostignuća. „Big data“ tehnologije, pored toga što prikupljaju veliku količinu podataka, omogućavaju izvlačenje vrijednosti tih podataka kao i njihovo bolje razumijevanje. Bilo je neophodno pronaći tehnologije koje će imati sposobnost da efiksno obrade velike količine podataka, a da to ne zahtijeva velike troškove. Prvi koji su došli do rješenja i napravili proboj u tehnologiji „Big data“ su Yahoo!, Google i Facebook i donijeli su promjene na tržištu upravljanja podacima. Pojavila se nova generacija u upravljanju podacima kojoj su doprinijele MapReduce, Hadoop, Big Table i Apache Spark tehnologije.

4.1. MapReduce

MapReduce predstavlja efikasno rješenje za veliku količinu podataka pomoću distribuiranih, paralelnih algoritama u klasteru. „Map“ rukovodi programerskim zadacima tako što ih balansirano raspoređuje i oporavlja od eventualnih grešaka, dok „reduce“ predstavlja funkciju koja spaja sve elemente nazad zajedno. Ove dvije funkcije se često koriste u funkcionalnom programiranju, međutim njihova uloga u MapReduce sistemu nije ista kao inače. „Map“ vrši sortiranje i filtriranje podataka, a „reduce“ vrši agregaciju. Ime „MapReduce“ je prvobitno bilo u vlasništvu Google tehnologije, ali se od tada generalizovalo.

4.2. Hadoop

Tehnologija koja se najčešće vezuje za Big data jeste Hadoop. Nastala je 2005. Godine i dizajnirana je tako da radi na jeftinijim hardverskim resursima, kao što je „commodity hardver“. Služi za skladištenje i procesiranje velike količine podataka i sastoji se iz četiri dijela:

-Hadoop common-niz biblioteka i konfiguracionih fajlova,

-HDFS-fajl sistem koji je zadužen za skladištenje podataka u klasteru,

-MapReduce-model za procesiranje podataka

-Yarn-zadužen za raspodjelu resursa i upravljanje poslovima

Osim ove četiri komponente Hadoop se oslanja na specijalizovane alate za prikupljanje podataka (Flume, Kafka, Sqoop), procesiranje podataka (Pig, Hive, Storm…), upravljanje (Ambari, Falcon…).

4.3. Big table

Big table je rješenje koje je predviđeno da upravlja skalabilnim struktuiranim podacima koji su organizovani u tabele. Predstavlja višedimenzionalnu mapu koja služi za mapiranje dva

proizvoljna stringa i vremenski trenutak u vezani niz bitova. Namjenjen je za čuvanje velike količine podataka na običnim serverima. Big Table je predviđen za rad na preko stotinu hiljada mašina. 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. (ovu rečenicu sam kopirala, dakle nisam je mjenjala, pa ti vidi da nekako izmjeniš)

4.4. Apache Spark

Apache Spark je platforma za obradu podataka, sa dodatnim modulima za mašinsko učenje, streaming i grafičku obradu. Obradu vrši u radnoj memoriji, što znači da je jako brz. Ukoliko podatci ne mogu da stanu u memoriju, Apache Spark ih premješta na disk, što dovodi do brže obrade nego da je samo na disku.

Koncept Apache Spark-a je RDD (Resilient Distributed Datasets) - kolekcija objekata rasprostranjenih kroz klaster RAM-u ili na disku, za koje je karakterističan paralelizam i automatski oporavak. Python, Scala, Java i od skoro R su jezici u kojima mogu da se pišu Spark programi.

5. ZAKLJUČAK

Razvojem globalne internetske mreže sve je lakše dijeliti i sakupljati ogromne količine podataka. Sa sve moćnijim i bržim alatima analitičari pokušavaju da idu u korak sa svakodnevnim povećanjem količine sirovih podataka i da ih efikasno obrade i primjene. Big data dobija svoj puni poitencijal tek kada se pravilno obradi i iskoristi. Podaci koji čine spremišta Big Data mogu da potiču iz izvora koji uključuju veb lokacije, društvene medije, stone i mobilne aplikacije, naučne eksperimente i - sve češće - senzore i druge uređaje na internetu stvari (IoT). Koncept Big Data sadrži skup povezanih komponenti koje omogućavaju organizacijama da koriste podatke za praktične potrebe i rešavaju niz poslovnih problema. Ovo uključuje IT infrastrukturu potrebnu za podršku Big Data; analitiku koja se primenjuje na podatke; tehnologiju potrebnu za projekte Big Data; povezane skupove veština; i stvarne slučajeve za koje je potrebno da se koristi Big Data.