PRIRODNO MATEMATIČKI FAKULTET U BANJOJ LUCI
STUDIJSKI PROGRAM MATEMATIKA I INFORMATIKA
SMJER INFORMATIKA

ANALITIKA VELIKIH PODATAKA

Stundent:
Jovana Mika

Mentor:
doc dr. Dragan Matić

Sadržaj

Uvod

Razvoj informacionih tehnologija i interneta unaprijedio je generisanje velike količine podataka i informacije kao što su: sadržaj na društvenim medijima (Facebook, You Tube itd.), državna administracija, poslovne aktivnosti (industrija, bankarstvo..) , namjenska industrija, internet prodaja, mobilni telefoni itd.

Analizom velikih količina podataka (eng. Big Data) omogućava se predviđanje ishoda događaja, procesa, posebno ljudskog ponašanja. Što dovodi do unaprijeđenja procesa donošenja odluka, otkrivaju se prevare i prijetnje, pronalaženja novih izvori prihoda, usavršavaju se procesi i operacije, generišu nove informacije i slično.

Istorija velikih podataka

Koncept velikih podataka postoji već dugi niz godina. Međutim i prije samog uvođenja pojma "veliki podaci", pedesetih godina 20. vijeka, preduzeća su koristila osnovnu analitiku .

Izraz "veliki podaci" koristi se od početka 1990-tih godina. Iako nije tačno utvrđeno ko je prvi uveo ovaj termin, smatra se da je to bio John R. Mashey.
U suštini Big Data nije nešto sasvim novo i nepoznato. I prije su ljudi koristili analize podataka i tehnike analitike kako bi podrzali proces donošenja odluka.
[1]***Primjer:*** Stari Egipćani oko 300. godine prije Hrista su pokušali objediniti sve podatke u Aleksandrijskoj knjižari. Takođe, Rimsko carstvo je analiziralo statistiku svoje vojske kako bi utvrdili optimalnu raspodjelu za svoje vojnike.

Do velikih promjena dolazi u posljednja dva vijeka. Brzina i generisanje podataka su se toliko promjenili da čak izlaze van granica ljudskog razumijevanja.
Ukupna količina podataka u svijetu izmjerena 2013. godine iznosila je 4,4 zettabajta. Do 2020. godine smatra se da je taj broj porastao na 44 zettabajta. Čak i uz najnaprednije tehnologije koje su prisutne u današnje vrijeme, nemoguće je analizirati sve te podatke. To je bio jedan od glavnih razloga pretvaranja tradicionalne analize podataka u velike podatke.

Sam razvoj Big Data-e možemo ilustrovati preko tri glavne faze evolucije podataka. Svaka od te tri faze ima svoje karakteristike i mogućnosti koje ćemo navesti u nastavku.

Prva faza

Analiza podataka, analitika i Big Data potiču iz dugogodišnjeg upravljanja bazama podataka. Prva faza oslanja se u suštini na tehnike pohrane podataka, ekstrakcije i optimizacije. Temeljna komponenta prve faze je upravo to, upravljanje i pohrana podataka. To je temelj moderne analize podataka koji nam je i danas poznat.

Druga faza

Počinje da se razvija početkom 2000 – te godine. Značajna je pojava interneta koji nudi jedinstvene zbirke podataka i mogućnost obrade istih. Javlja se i potreba za organizacijom kako bi se pronašao novi pristup i rješenje za pohranu podataka kako bi analiza podataka bila učinkovitija. Zbog porasta podataka javlja se potreba za novim tehnologijama za obradu podataka.

[1] Primjer preuzet iz literature

Treća faza

Smatra se da obuhvata period razvoja Internet stvari (engl. Internet od things), porast snage računara i razvoj interneta, potrebu za većom pohranom podataka i sl.
2005. godine kreiran je Hadoop koji može da obradi Big Data-u. Temeljio se na na softverskom okviru s otvorenim izvorima i spojenim s Google –ovim MapReduce-eom. To je bila samo jedna od tehnologija za obradu velikih podataka, koja se javlja u trećoj fazi. U nastavku će detaljnije biti objašnjena primjena baza velikih podataka, tehnologije koje se koriste i slično, a vezane su za treću fazu.

BIG DATA PHASE 1	BIG DATA PHASE 2	BIG DATA PHASE 3
Period: 1970-2000	Period: 2000-2010	Period: 2010-present
DBMS-based, structured content: • RDBMS & data warehousing • Extract Transfer Load • Online Analytical Processing • Dashboards & scorecards • Data mining & statistical analysis	Web-based, unstructured content • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis	Mobile and sensor-based content • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization • Human-Computer-Interaction

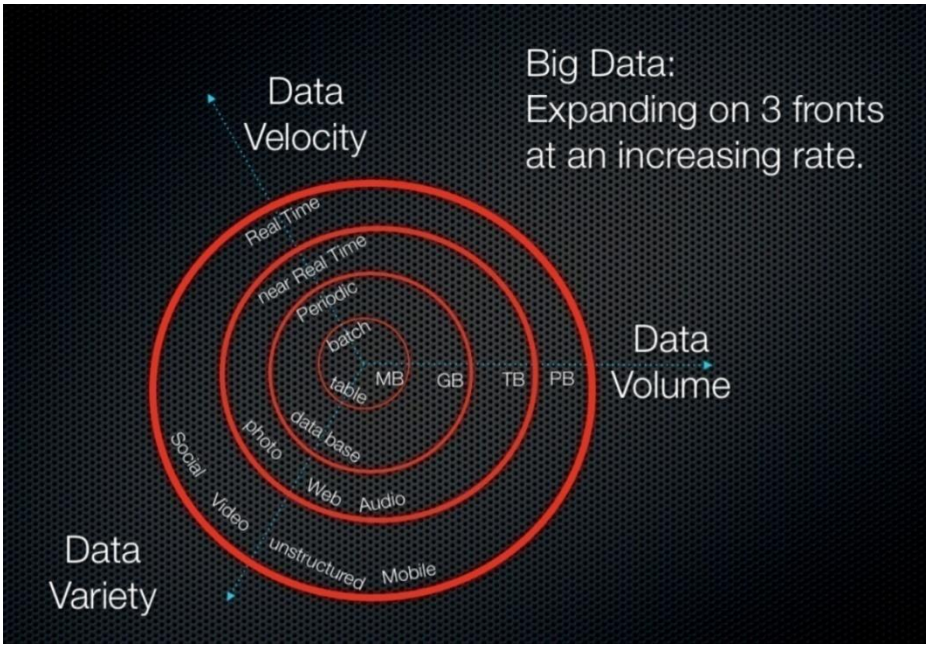
Slika 1.1 Kratki pregled istorije Big Data-e

Definicija i domen Big Data

Big Data (srp. veliki podaci) predstavlja sistem zasnovan na određenoj informacionoj tehnologiji, a odnosi se na velike količine strukturisanih ili nestrukturisanih podataka. Označava skupove velikih podataka. Potreba za obradom većih skupova podataka dovela je do pretvaranja tradicionalne obrade podataka u velike podatke.

Karakteristike velikih podataka:

- Obim (engl. volume)
- Brzina opticaja (engl. velocity)
- Raznovrsnost (engl. variety)
- Složenost (engl. complexity)
- Vjerovatnost (engl. probability)
- Osjetljivost (engl. sensibility)
- Kvalitet (engl. quality)
- Vrijednost (engl. value)



Slika2.1 Grafički prikaz osnovnih dimenzija Big Data

Big Data uvodi mnoge promjene u oblasti biznisa, raznih inustrija i drugih dijelova našeg života. Ne odnosi se samo na količinu velikih podataka, već i na brzinu i način na koji se kreiraju i kori – ste. Objasnimo neke od bitnih karakteristika velikih podataka:

VELIČINA

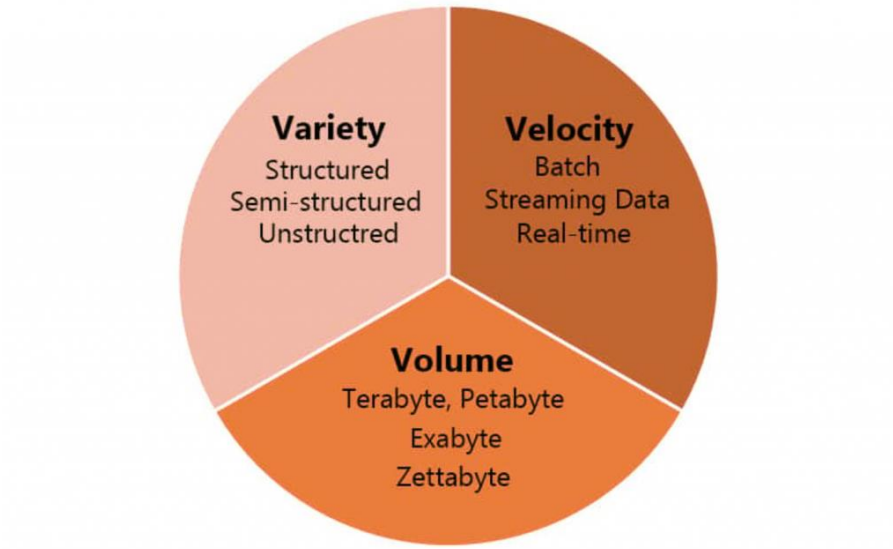
Veličina je jedna od komponenti Big Data. Javlja se potreba za generisanjem sve veće količine podataka. Ako bi se uzeli svi podaci koji su nastali u svijetu od početka mjerenja 2000. godine , možemo zaključiti da se ista količina podataka sada generiše svakih nekoliko minuta. Više od 90% podataka na svijetu je nastalo u posljednjih nekoliko godina.

BRZINA

Brzina je takođe jedna od bitnijih komponenti Big Data. Svakog minuta pošalje se preko 200 miliona e-mailova, klikne se skoro 2 miliona lajkova na Facebooku, pošalje se preko 300 hiljada poruka na nekoj društvenoj mreži i slično.
Takođe samo preko Google-a se pretraži oko 3.5 milijardi pretraga dnevno.

STRUKTURA

Struktura je treća komponenta Big Data. U prošlosti se uglavnom oslanjalo na struktuirane podatke, tipove podataka koje smo mogli da ubacimo u tabele i organizujemo brzo. Manje struktuirani podaci kao što su tekstualni fajlovi, fotografije, video sadrzaji i slično, većinom su ignorisani.
Međutim, danas imamo mogućnost da koristimo i analiziramo i takve podatke, uključujući i pisani tekst, izgovorene riječi, čak i ton u našem glasu, kao i biometrijske podatke i video sadržaj.



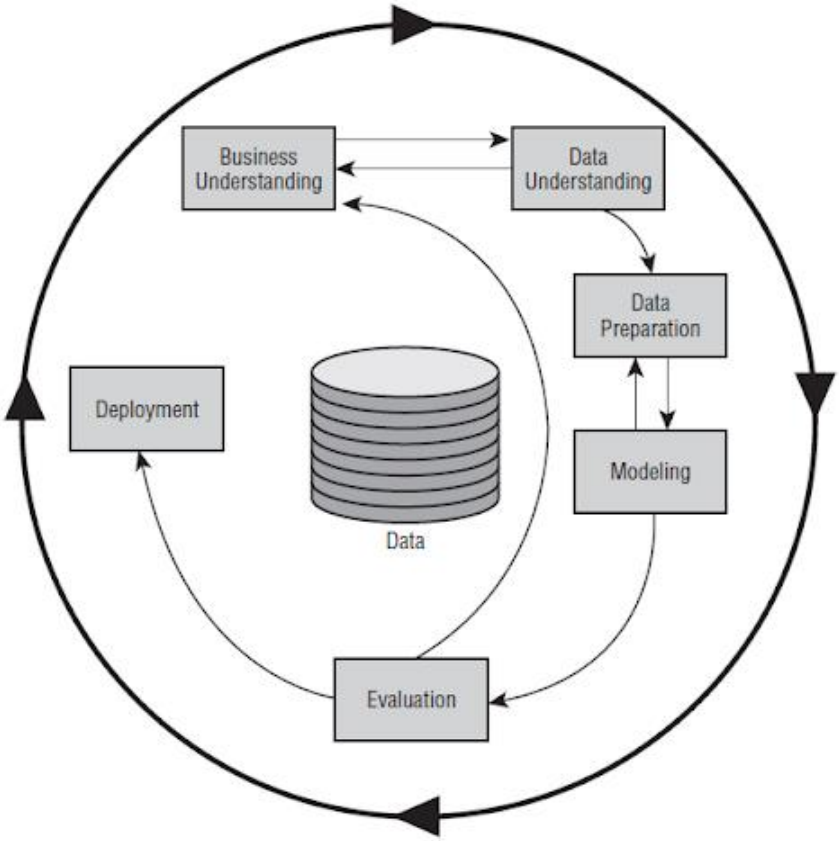
Slika 2.2. 3V model Big Data

Big Data analitika

Analitika velikih baza podataka predstavlja analiziranje i inteligentno donošenje odluka na osnovu prikupljenih podataka. Termini Big Data i Big Data analitika nisu isključivi i formiraju jednu komplementarnu cjelinu.

Analitika velikih podataka jeste postupak predviđanja ponašanja objekata ili uočavanja trendova na osnovu serija prikupljenih struktuiranih i nestruktuiranih podataka. Analitika je jedan od osvnovnih dijelova industrijskih operacija u mnogim privrednim granama, kao na primer: maloprodaja, veleprodaja, zdravstvo, bankarstvo, transport i slično. Omogućava organizacijama da iskoriste svoje podatke na najbolji mogući način i da ih koriste za otkrivanje nekih novih mogućnosti.

Analitika može da se odnosi na osnovne aplikacije poslovne inteligencije ili naprednije analitike kao što su one koje koriste naučne organizacije. Rudarenje podataka je jedno od naprednijih tipova analitike podataka .



Slika 3.1. Rudarenja podataka

Analitika podataka može uključivati istraživačku analizu podataka (ili analizu numeričkih podataka koji imaju kvantifikovane promjenljive koje se mogu statistički uporediti) za razliku od kvalitativne analize podataka(koja se fokusira na nenumeričke podatke kao što su video, slike i tekst) .

Postoje dvije cjeline postupka analiziranja analitike:

1. **oblikovanje/modeliranje uz analizu**
2. **interpretacija**

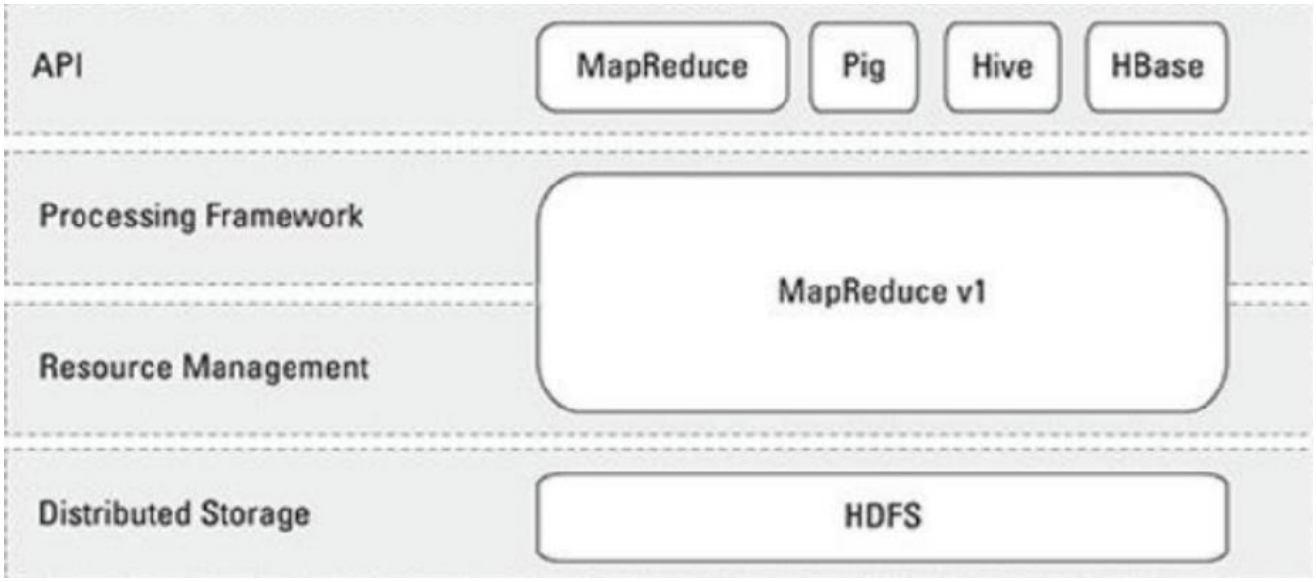
To su tehnike kojima se stiče iskustvo na osnovu velikih baza podataka. Među osnovnim tehnikama Big Data analitike mogu se izdvojiti:

- ***Analitika teksta*** – dobijanje informacija iz natpisa u drustvenim medijima, e-pošti, blogovima, online forumima, poslovnim dokumentima, vijestima i dr. U upotrebi su tehnike poput algoritamske ekstrakcije podataka, sažimanja teksta, servisa za odgovore na pitanje (npr. Siri) i sl.
- ***Audio analitika*** – prikupljanje informacija iz snimljenih razgovora (npr. pozivni centri kompanija, bolničko dijagnostikovanje mentalnih poremećaja ili utvrđivanje stanja novorođenčeta na osnovu boje, visine i jačine glasa) .
- ***Video analitika*** – generisanje informacija na osnovu video sadržaja (npr. siguronosno nadgledanje objekata, prisluškivanje).
- ***Analitika društvenih medija*** – sticanje informacija na osnovu podataka objavljenim na društvenim medijima (npr. identifikovanje zajednica, društveni uticaj natpisa na pojedince).
- ***Predviđanje*** – donošenje odluka na osnovu prikupljenih i obrađenih podataka i iz njih izvedenih informacija upotrebom statističkih tehnika poput kretanja prosječnih vrijednosti (npr. predviđanje sledeće kupovine potrošača, ili predviđanje videa koji bi mogao zanimati datog korisnika).

Tehnologije specifične za Big Data

EKOSISTEM HADOOP

Hadoop je jedna od tehnologija koja je blisko povezana sa Big Data-om. Projekat Apache Hadoop razvija softver otvorenog koda za distribuirano računarstvo. Napisan je u programskom jeziku Java.
Softverska biblioteka Hadoop predstavlja okvir koji omogućava distribuiranu obradu velikih količina podataka preko klastera računara pomoću jednostavnih programskih modela.
Dizajniran je da skalira od jednog servera pa sve do hiljade servera, od kojih svako nudi lokalno računanje i skladištenje.



Slika 4.1. Ključni elementi Hadoop platforme

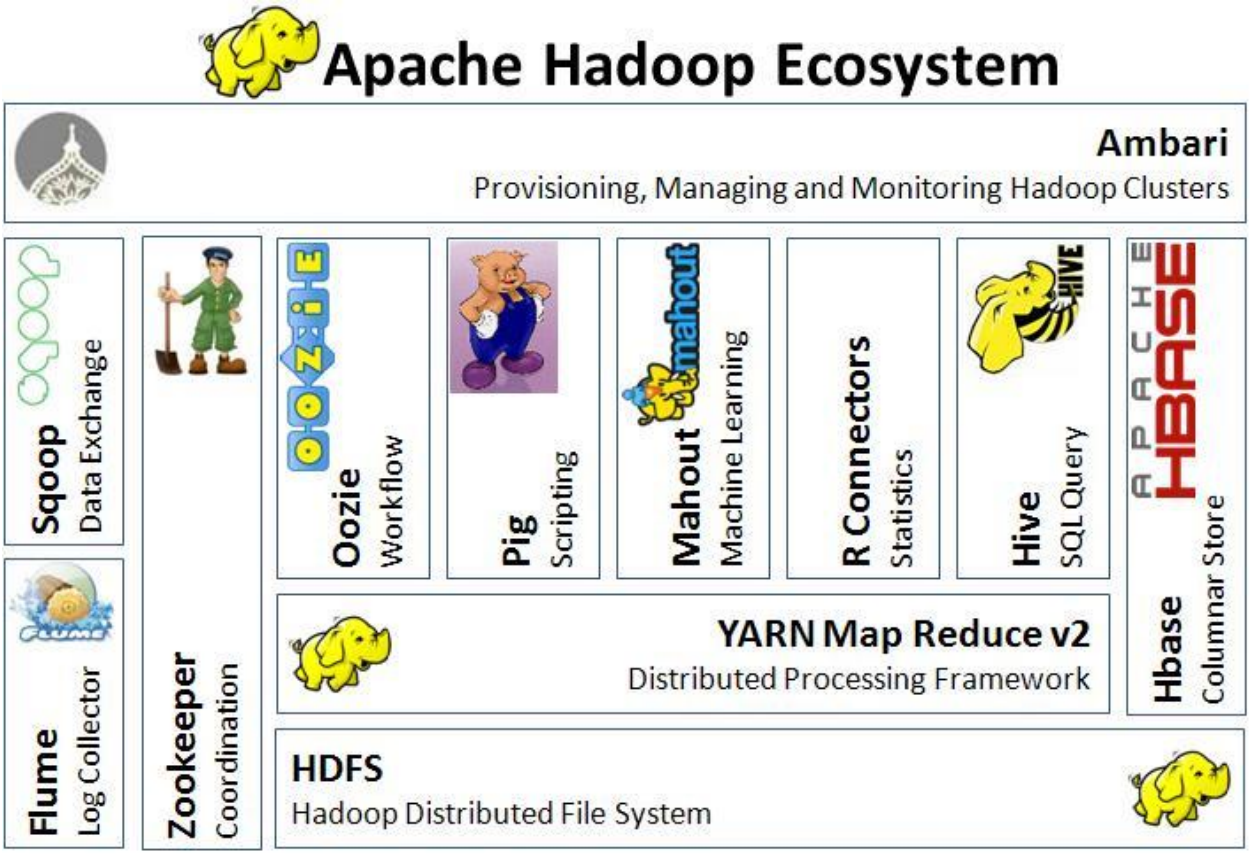
Hadoop najbolje radi sa srednje velikim brojem ekstremno velikih datoteka (većih od 500MB) . Poželjan model upotrebe je "*piši jednom čitaj često*", što znači da se podaci upisuju jednom, a mogu se čitati više puta.
Dozvoljene operacije su kreiranje nove datoteke, dodavanje na kraj datoteke, brisanje, preimenovanje.

Kako Hadoop ekosistem funkcioniše?

Hadoop sistem zapisuje podatke jednom, a zatim ih čita i ponovo koristi više puta.
Hadoop se sastoji iz četiri dijela:

1.Hadoop Common – predstavlja niz biblioteka potrebnih za rad
2. HDFS – distribuirani fajl sistem koji skladisti podatke u klasteru
3.Map Reduce – model za procesiranje podataka
4.Yarn – Hadoop operativni sistem, zadužen za podjelu resursa i upravljanjem poslovima

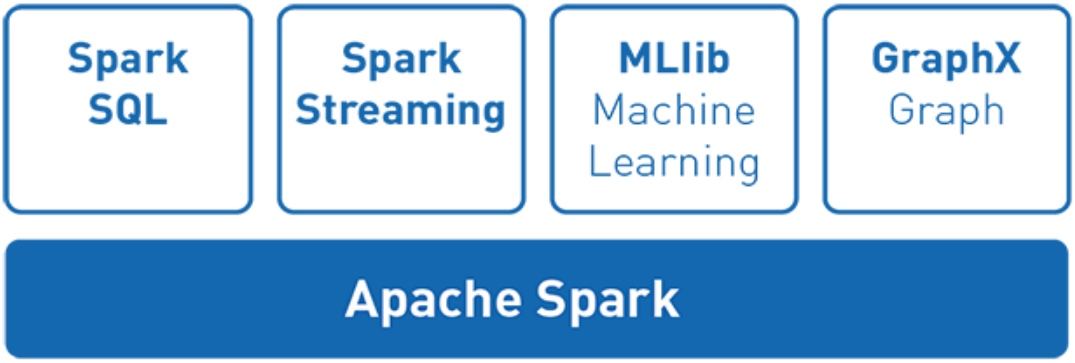
Svi podaci se čuvaju u formi blokova, i postoje 3 kopije kroz klasteru. Što omogućava da ako dođe do greške odnosno otkaza neke mašine, postoje još 2 kopije.
Takođe postoje i neki specijalizovani alati koji se koriste za prikupljanje podataka, upravljanje podacima i slično. Neki od takvih alata su : Kafka, Flume, Falcon.



Slika 4.2. Hadoop ekosistem

APACHE SPARK

Dio ekosistema Hadoop, Apache Spark je radni okvir otvorenog koda za klastersko računarstvo. Koristi se kao motor za obradu Big Data u okviru Hadoop-a. Spark je postao jedno od ključnih okruženja za distribuiranu obradu podataka i može da se koristi na različite načine. Moguće je povezivanje sa Java, Scala, Python i R programskim jezicima, takođe podržava SQL, podatke u strimovanju, mašinsko učenje i obradu grafova.



Slika 4.2. Spark razvojni okvir

Osnovne funkcionalnosti:

- upravljanje zadacima
- upravljanje memorijom
- oporavak od otkaza
- interakcija sa sekundarnom memorijom

Gledano iz perspektive programera Spark program pokreće dva tipa programa:

1. Glavni program (engl. Driver program) koji pokreće različite paralelne operacije nad klasterom i vrši sažimanje parcijalnih računara.
2. Izvršni program (engl. wokers) koji izvršavaju iste operacije nad različitim participama podataka.

JEZERA PODATAKA

Jezera podataka predstavljaju baze podataka u kojima se čuvaju velike količine neobrađenih podataka u izvornom formatu dok podaci ne budu potrebni kod poslovnih korisnika. Često se koriste jezera podataka zasnovana na Hadoop tehnologiji.
Dizajnirana su tako da korisnicima olakšaju pristup velikim količinama podataka kada se javi potreba za tim.

NOSQL BAZE PODATAKA

NoSQL baze podataka napravljene su za posebne modele podataka. NoSQL baze podataka čuvaju podatke i upravljaju njima na način koji omogućava veliku brzinu rada i veliku fleksibilnost. Za razliku od SQL baza podataka , mnoge NoSQL baze podataka mogu da se skaliraju horizontalno preko stotina ili hiljada servera. Između ostalog koriste se za izradu web stranica i modernih mobilnih aplikacija.

BAZE PODATAKA U MEMORIJI

Baza podataka u memoriji (IMDB - in-memory data base) je sistem za upravljanje bazama podataka koji se za skladištenje podataka oslanja na glavnu memoriju, a ne na disk.Jedna od bitnih prednosti IMDB-a je brzina pretrage. Baze podataka u memoriji su brže od baza podataka optimizovanih za diskove, što je važno za analize Big Data i stvaranje skladišta podataka i centara podataka. Tek se očekuje rast globalnog tržišta baza podataka.

VJEŠTINE ZA BIG DATA

Rad sa Big Data-om zahtjeva posebne vještine. Mnoge od vještina podrazumijevaju znanje u radu sa nekom od tehnologija koje smo već pomenuli, a kao što su Hadoop, Apache Spark, NoSQL baze podataka i slično. Takođe, značajno je poznavanje nauka o podacima, analize podataka, rudaranje, vizualizacija podataka, poznavanje opšteg programiranja kao i poznavanje raznih struktura i algoritama.

Još uvijek je prisutan nedostatak stručnih ljudi koji imaju potrebne vještine za rad sa Big Data-om, tako da je jedan od problema velikih kompanija upravo pronalazak stručnog kadra.

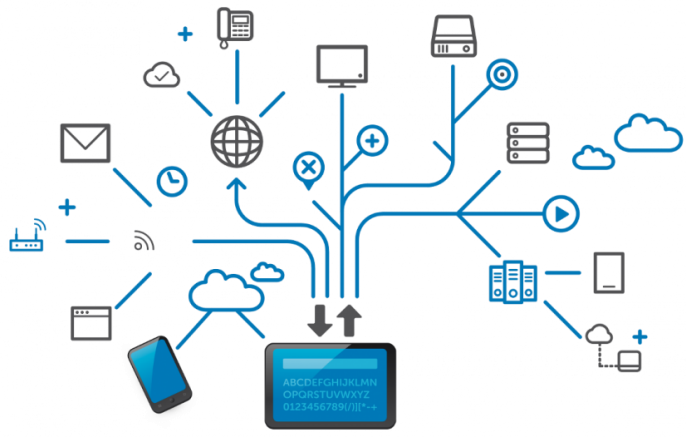
Primjena Big Data

Iako prevladava mišljenje da se Big Data analitika u najvećoj mjeri primjenjuje u oblasti infromacionih tehnologija, činjenica je da se koristi i u drugim sferama poslovanja, ali i u svakodnevnom životu.
Osnovnu primjenu ipak nalazi u oblasti informacionim tehnologijama tj. oblasti obrade digitalnih podataka. Digitalni svijet predstavljda najdinamičniji dio poslovnog svijeta današnji - ce. U digitalnom svijetu se svakog trenutka generišu terabajti podataka, što kroz korištenje društvenih mreza i mreža za dijeljenje sadržaja, što kroz poslovnu razmjenu podataka. Svaka kompanija kao što su npr. apoteke , svakodnevnim korištenjem interneta ostavljaju trag u digitalnom svijetu i generišu velike količine podataka koji adekvatno mogu biti obrađeni jedino upotrebom Big Data analitike.

Primjer: Privatna apoteka može imati naloge na svim društvenim mrežama, može imati i svoj sajt koji ima određenu posjećenost, može na dnevnoj osnovi vršiti na stotine upita na Google pretraživaču i može imati aktivnu E-mail komunikaciju. Sve ove aktivnosti generišu velike serije digitalnih podataka koji ne govore samo o načinu upotrebe informacionih resursa nego I o karakteristikama osoba sa kojima se vrši interakcija i karakteristikama samih zaposlenih.

Upotreba informacionih tehnologija i rad u digitalnom svijetu otvaraju mogućnosti analize podataka u cilju unapređenja poslovanja kompanija i svih učesnika na tržištu.

PRIMJENA U TELEKOMUNIKACIJI



Slika 4.1. Telekomunikacija - umrežavanje

Big Data veliku primjenu ima i u oblasti telekomunikacije. Pored toga što je infromatički uvezana, današnjica je uvezana i preko telekomunikacionih sredstava. Više se ne govori o dva pojma pojedinačno već se uvodi termin "informaciono – telekomunikacione tehnologije ". Najvazniji aspekt telekomunikacije su mobilne komunikacije. Mobilne mreže su u današnjem vremenu posebno rasprostranjene. Operatori mobilnih mreža svakog dana generišu ogromne količine podataka u vidu telefonskih razgovora, SMS i MMS poruka, mobilnog interneta i slično.

Kao i kod informacionih tehnologija i telekomunikacije omogućavaju velikim farmaceutskim kompanijama, da prikupe i analiziraju velike količine podataka koje su vezane za kupce, potrošače, i to samo na osnovu korištenja mobilnih komunikacija. Da bi se svi ti podaci pretvorili u informacije koje možemo koristiti potrebni su veliki kapaciteti Big Data analitike.

PRIMJENA U FINANCIJSKOM SEKTORU

S obzirom da su skoro sve financijske transakcije danas automatizovane, podaci iz financijskog sektora su veoma upotrebljivi u Big Data analitici. Samo u okviru jednog minuta u svijetu se obave preko 3.500 transakcija na berzama i nekoliko miliona bankarskih transakcija. Takođe svaka firma dnevno najčešće generiše na stotine financijskih transakcija prodajom i nabavkom robe i usluga i realizacijom plaćnja za razne namjene. Big Data analitika potpunu obradu podataka, u najkraćem mogućem roku i da na osnovu njih prikazuje dobijene rezultate koji se mogu koristiti na dnevnoj osnovi za unapređivanje poslovanja.

BIG DATA U VOJNOM SEKTORU

Širok spektar upotrebe velikih baza podataka se ogleda u sledećim vojno-ogranizacionim podsistemima:

- Vazduhoplovne jedinice
- Sajber- bezbjednost
- Komandovanje/kontrola
- Podsistem raketnog navođenja
- Pomorski sistem
- Podsistem radara i senzora
- Obuke personala

[2]*Primjer:* Tokom jednog celog dana misije jednostavne složenosti, bespilotna letjelica dostavlja centrali 10 terabajta podataka od čega je samo 5% predmet analize dok se ostatak skladišti. Nepostojanje uslova za analizu preostalog dela od 95% podataka minimizuje ukupan kvalitet donijetih odluka. Bespilotna letelica prikuplja video-audio podatke u toku leta i pri nailasku na neprijateljske objekte ili jedinice preko video snimka dolazi do algoritamskog „isčitavanja“ čime se identifikuje vrsta naoružanja, brojnost i položaj neprijatelja, te prosleđuje predlog rešenja komandnoj jedinici na osnovu analiziranih faktora rizika. Istovremeno, oslanjanjem na Big Data analitička rešenja povećava se sigurnost pilota u toku naleta usled poboljšanog predviđanja rizika u toku samog leta. Potrošnja goriva aviona i letelica je smanjena.

[2] Primjeri preuzeti iz literature

Zaključak

Big Data analitika ima sve veću primjenu u skladištenju i obradi podataka. Ne može da bude zamjena za sisteme poslovne inteligencije i skladište podataka, ali kroz postojeće sisteme donosi neke nove vrijednosti. Big Data se može posmatrati kao novi element u sistemu poslovne inteligencije koji može donijeti novo znanje, a sa njim i novu vrijednost. Vidjeli smo i da je zabilježen porast količine podataka kroz odredjen vremenski period, pa očekujemo da će se taj broj u budućnosti znatno povećati.

Literatura

1. Velike Baze Podataka – Big Data, Primena u Vojno-bezbednosnom sistemu http://www.odbrana.mod.gov.rs/odbrana-stari/vojni_casopisi/arhiva/VD_2018-3/70-2018-3-17-Milojevic.pdf
2. Big Data i poslovna inteligencija - https://infoteh.etf.ues.rs.ba/zbornik/2014/radovi/RSS-3/RSS-3-10.pdf
3. Uvod u analitiku velikih podataka – Aleksandar Kartelj
4. Big Data Analytics - https://www.sas.com/en_us/insights/analytics/big-data-analytics.html