O podatcima velikog obujma

Brkljača, Katarina

Sveučilište u Rijeci, Odjel za informatiku

Preddiplomski jednopredmetni studij informatike

Katarina Brkljača

O podatcima velikog obujma
(eng. About Big Data)

Završni rad

Mentor: prof. dr. sc. Patrizia Poščić

Rijeka, rujan 2020.

Sadržaj

Sažetak

U ovom završnom radu objasnit će se osnovni pojmovi, struktura i karakteristika Big Data danas. Nadalje, bit će opisan povijesni nastanak i dan pregled Big Data. Kroz primjenu Big Data u praksi navest će se nekoliko najuspješnijih tvrtki koje koriste Big Data analizu (Facebook, Microsoft, Google), a zatim će se dati usporedba prednosti i nedostataka korištenja Big Data. Posljednji osvrt bit će na primjerima alata za analizu podataka velikog obujma (R, SQL, Python,).

Ključne riječi

Podatak, strukturirani podatak, nestrukturirani podatak, polu-strukturirani podatak, Big Data, alati, analiza

1.Uvod

Čovječanstvo konačno uči iz informacija koje može prikupiti kao dio našeg izvanvremenskog zadatka da shvatimo svijet i naše mjesto u njemu. Zato veliki podatci jesu velika stvar. Podatci se nalaze svugdje: u knjigama, na Internetu, društvenim medijima, zdravstvenim industrijama, čak i u ljudima. *Podatak* je postao ključno sredstvo društvenog i gospodarskog razvoja. Mogućnost prikupljanja i analiziranja podataka danas, uvelike se razlikuje od onoga do prije nekoliko godina. Svaki novi podatak omogućuje nešto novo, bolje i drugačije. Jedan od najbrže rastućih izvora velikih podataka za analizu jesu društveni mediji. Važnu ulogu imaju i podatci u zdravstvenim industrijama, automobilskim industrijama, marketingu, poljoprivredi i mnogim drugim [2].

Društvo nije toliko napredovalo. I dalje se podatci čuvaju na diskovima, USB-ovima, memorijskim karticama, ali danas se može sačuvati puno više podataka nego ikad prije. Što se tiče povijesnog napretka, u posljednjih nekoliko godina ljudsko znanje je razvilo više podataka nego u cijeloj postojanosti čovječanstva. Podatci su prešli iz nečega što je statično u nešto fluidno i dinamično. *Analiza* podataka, posebno analiza podataka velikog obujma koja uključuje rad s ogromnim, stalno promjenjivim i vrlo složenim skupovima podataka, znatno je teža negoli prije nekoliko desetljeća [1]. Koncept velikih podataka dobio je zamah početkom 2000-ih godina kada je analitičar Doug Laney artikulirao glavnu definiciju velikih podataka kao 3V.

Big Data tehnologija enormno raste upravo zbog brzine stvaranja novih podataka i potrebom za pohranu i obradu istih. Zašto danas tražilice izbacuju upravo informacije koje nas zanimaju? Ili se na web stranicama pojavljuju razni oglasi i reklame o kojima se istražuje, razgovara i/ili čak razmišlja? Odgovor će uvijek biti isti – zbog velike količine podataka. U digitalnom svijetu, ti podatci su pohranjeni samo jednim klikom. Podatci su postali veliki kapital. Ako pogledamo neke od najvećih svjetskih tehnoloških tvrtki, veliki dio vrijednosti koje nude potiče iz analize vlastitih podataka u svrhu razvitka i učinkovitosti novih proizvoda [1].

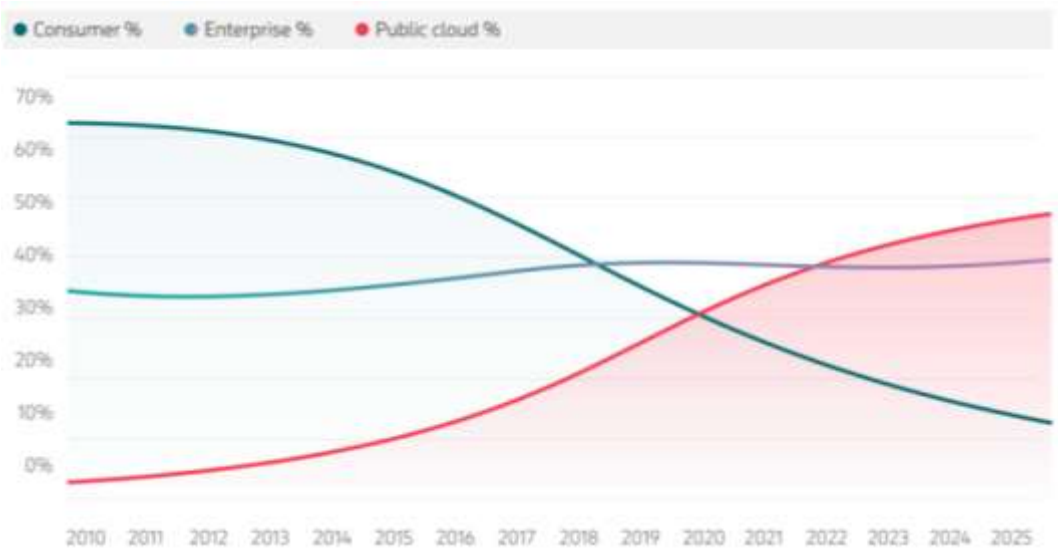
Veliki problemi s podatcima zahtijevaju nove tehnologije i alate za pohranu, analizu, upravljanje i ostvarivanje poslovne dobiti. Neograničena količina podataka znatno utječe na novu nadolazeću, nepoznatu i bližu budućnost koja bi ubrzo mogla zamijeniti ljudsku snagu i znanje s nevidljivom tehnologijom sastavljenom od milijardu podataka. Potrebno je znati iskoristiti sve prednosti novoga svijeta, ali isto tako pokušati popraviti ključne nedostatke kako bi u svemu bili puno uspješniji.

Tema ovog završnog rada je upoznati Big Data kroz njegove karakteristike, strukturu, prednosti i nedostatke, primjenu u praksi te alate za analizu velikih podataka.

2.Općenito o podatcima velikog obujma

Podatak je bilo koja činjenica ili poruka koja može prenijeti informaciju, ali se ne mora iskoristiti, dok je informacija protumačeni podatak. Može biti u obliku znakovnog, brojevnog ili grafičkog prikaza. U dosadašnjim istraživanjima naglasak je bio više na nešto vidljivo oku, na hardverski dio, dok zapravo mnogo bitniju ulogu ima ono manje uočljivo, a to su podatci. Što je zapravo Big Data? *Big Data* je rezultat korištenja dvaju različitih područja, a to su strojno učenje i računarstvo u oblaku. Temelji se na 3 glavna atributa podataka, volumen, brzina i raznolikost. Koristi se za opisivanje ogromne količine strukturiranih i nestrukturiranih podataka, zapravo Big Data je beskonačna zbirka podataka velikog obujma koji eksponencijalno raste. 90% svjetskih digitalnih podataka je stvoreno u posljednje 2 godine [10].

Veliki podatci se mogu prikupiti iz svih javno dijeljenih komentara na društvenim mrežama i web stranicama, putem upitnika i anketa, kupovine proizvoda i elektroničke prijave, itd. Najčešće se pohranjuju u računalnim bazama podataka, a analiziraju se pomoću posebno dizajniranog softvera za obradu velikih složenih skupova podataka [8]. Na slici 1 prikazano je gdje se sve pohranjuje Big Data i u kojim količinama. Vidljivo je da kod potrošača postotak opada kako vrijeme odmiče, a udio velikih podataka poduzeća ima blagu krivulju koja ide ka porastu. Suprotno potrošačima, pohrana u oblaku implicitno raste od 2015. godine, dok se u 2020. godini nalazi na jednakoj količini pohrane podataka kao i kod potrošača. Više podataka se pohranjuje u oblaku i u poduzetničkim bazama podataka, nego u potrošačkim.



Slika 1 - Gdje se sve pohranjuje Big Data

Veliki podatci pružaju nove uvide koji otvaraju vrata ka drugim mogućnostima i poslovnim modelima, a dijeli se u 3 koraka [10]:

1. Integriranje – podatci koji se prikupljaju dolaze iz različitih izvora i aplikacija. Za vrijeme integracije potrebno je unijeti podatke, izvršiti njihovu obradu i provjeriti jesu li formatirani i dostupni u obliku prilagođenim za poslovne analitičare.

2. Upravljanje – svaki prikupljeni podatak treba se i pohraniti. To može biti u oblaku, prostoriji, magnetskoj pohrani ili na bilo koji mogući način. Najvažnije je da svaki pohranjeni podatak bude dostupan na temelju traženog zahtjeva od strane korisnika.

3. Analiziranje – vrši se kod pristupa podatcima. Postupkom analize može se unaprijediti sustav ili izgraditi modeli podataka pomoću umjetne inteligencije i strojnog učenja.

2.1.Vrste podataka

Big Data se mogu dobiti u više oblika, uključujući strukturirane i nestrukturirane podatke kao što su financijski podatci, tekstualne i multimedijske datoteke i genetička preslikavanja. Većina Big Data je nestrukturirana ili polu-strukturirana što zahtijeva različite tehnike i alate za obradu i analizu o kojima nešto više ima u odlomku *Alati* za analizu podataka velikog obujma.

Strukturirani podatci se sastoje od informacija kojima već upravlja organizacija u svojim bazama podataka i proračunskim tablicama – prikazano na slici 2 u stupcu 'Employee'. To je tradicionalni podatak koji se organizira i u skladu je s formalnom strukturom podataka. Takva vrsta podataka se može pohraniti u relacijsku bazu podataka, npr. izvod banke koji sadrži datum, vrijeme, iznos, itd. Strukturirani podatci stoje iza većine trenutnih poslovnih aplikacija i skladišta podataka. Podatci u relacijskim bazama podataka i proračunskim tablicama su osnovni primjer strukturiranih podataka.

Employee_ID	Employee_Name	Gender	Department
2365	Rajesh Kulkarni	Male	Finance
3398	Pratibha Joshi	Female	Admin
7465	Shushil Roy	Male	Admin
7500	Shubhojit Das	Male	Finance
7699	Priya Sane	Female	Finance

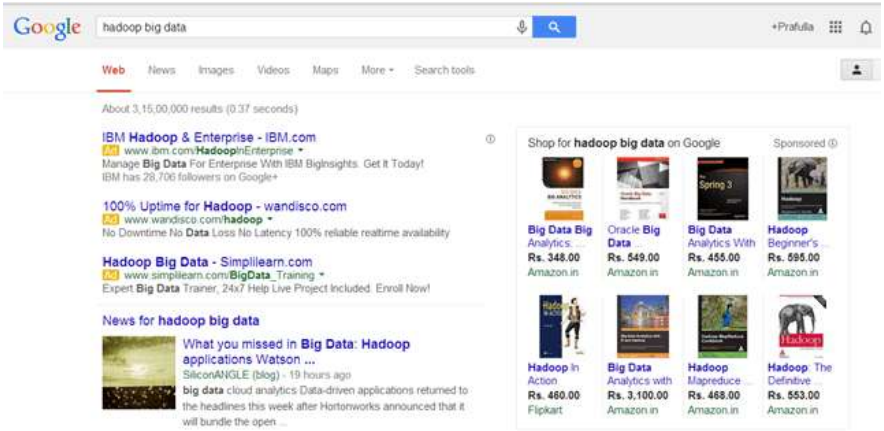
Slika 2 - Primjer strukturiranih podataka

Polu-strukturirani podatci jesu polu-organizirani podatci. Ne podudaraju se s formalnom strukturom podataka. Primjer polu-strukturirane baze podataka implementirane bez tablica su tzv. Dokument baze podataka. Nemaju shemu, te ih se definira kao samo opisujuće što znači da nije potreban opis strukture niti definiranje podataka. Npr. podatci senzora, JSON datoteke, XML datoteke – prikazano na slici 3.

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35
</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</
age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29
</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26
</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35
</age></rec>
```

Slika 3 - Primjer polu-strukturiranih podataka u XML-u

Nestrukturirani podatci su neorganizirani podatci i ne spadaju u određeni model ili format. Ne uklapaju se u strukturu redova i stupaca relacijske baze podataka. Uključuju podatke prikupljene iz raznih izvora društvenih medija koji pomažu institucijama u prikupljanju podataka o potrebama korisnika. Primjer nestrukturiranih podataka su. slike, videozapisi, tekstualne datoteke, e-poruke, itd [11] kao što je prikazano na slici 4. Specifično za nestrukturirane podatke je da se moraju najprije čistiti kako bi se mogli koristiti u bilo koju svrhu za analitiku.



Slika 4 - Primjer nestrukturiranih podataka

2.2.Karakteristike podataka

Glavni razlog prikupljanja podataka nije u tome što postoji dovoljno kapaciteta za pohranjivanje podataka velikog obujma, brzine i raznolikosti, nego je cilj pronaći optimalno rješenje za neki istraživački ili poslovni problem, a to je traženje djelotvorne inteligencije. Svaka velika stvar s kojom se upravlja, treba biti okarakterizirana na način da razumije svoje organiziranje. Od 1997. godine dodani su atributi, 3V ili Gartnerova interpretacija, koji čine Big Data potpunom, to su volumen (eng. volume), brzina (eng. velocity) i raznolikost (eng. variety) [4].

Volumen (količina) podataka je vrlo bitna. Veliki podatci traže i veći udio obrade nestrukturiranih podataka niske gustoće. Što je više podataka prikupljeno, to je znatno lakša ponuda i potražnja sličnijih proizvoda ili usluga.

Brzina prikupljanja podataka je ključan faktor kod algoritama za upravljanje podacima „u hodu“. Uz količinu podataka, vrlo je važna brzina generiranja i prikupljanja podataka u što manje vremena, odnosno što većom brzinom.

Raznolikost informacija odnosi se na dostupne podatke. Porastom broja podataka, novi podatci dolaze u nestrukturiranom obliku koji zahtijevaju dodatnu obradu. Označava raznolikost nekompatibilnih i nedosljednih formata i struktura podataka.

Iako količina velikih podataka privuče najviše pažnje, općenito brzina i raznolikost podataka daju precizniju definiciju Big Data. Moguće je da veličina bude relativno mala, a opet previše raznolika, ili može biti jednostavna, a imati ogroman broj podataka. Tijekom posljednjih nekoliko godina, IBM-ovi znanstvenici raščlanjuju podatke u 4 dimenzije, odnosno na vrh Douglas Laney-evih 3V, dodaju još jednu, istinitost (eng. veracity).

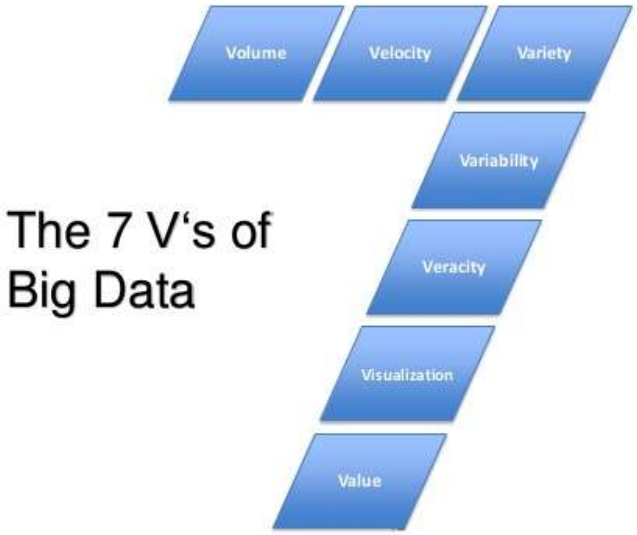
Istinitost je karakteristika koja za dostupne podatke, osiguravajući relevantnost, visoku kvalitetu, integritet i točnost, razdvaja one koji su bitni od onih koji su nebitni. Nastala je kao odgovor na pitanje kvalitete i izvora s kojima su se klijenti suočavali koristeći Big Data inicijativu.

Nedugo nakon IBM-a, Microsoft proširuje broj karakteristika na 6V dodajući još vrijednost (eng. value) i vidljivost (eng. visibility) podataka [4]. Uz 6V postoji još jedan pod nazivom varijabilnost (eng. variability) podataka – slika 5.

Vrijednost se odnosi na to koliko su korisni podatci u odlučivanju. Naime, ukoliko imamo puno podataka, nema nikakve koristi ako iz njih ne izvučemo neke uvide. Potrebno je izdvojiti vrijednost velikih podataka koristeći odgovarajuću analitiku.

Vidljivost ili vizualizacija podataka pruža mogućnost analize procesa u cijelosti od početnog kontakta do njegova ispunjenja. Omogućuje trgovcima da brzo istaknu proizvode i usluge, štedeći na vremenu i financijama.

Varijabilnost je stupanj u kojem se podatkovne točke u statističkoj distribuciji ili skupu podataka razlikuju, tj. variraju, od prosječne vrijednosti u mjeri u kojoj se razlikuju jedna od druge. Te mjere se pronalaze metodama otkrivanja anomalija i vanjskih oblika kako bi se došlo do smislene tematike. Do varijabilnosti dolazi zbog prikupljanja podataka iz različitih izvora.



Slika 5 - 7 karakteristika Big Data

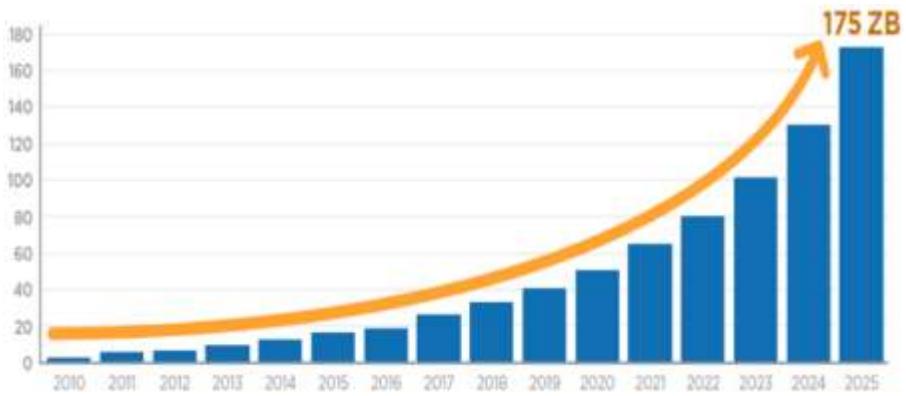
Primarni razlog zašto se Big Data naglo razvijao posljednjih godina je taj što povećava dugoročnu vrijednost poduzeća. Vrijednost se bilježi kako u neposrednoj socijalnoj ili novčanoj dobiti, tako i u obliku strateške konkurentske prednosti. Ako je dozvoljen kapacitet tehnologije, ne postoje granice ili ograničenja za podatke. Svaka od gore navedenih definicija namjerava opisati određeni problem samo s jednog aspekta Big Data i vrlo je restriktivan.

3.Povijesni razvoj Big Data

Big Data nije nešto sasvim novo za današnje društvo. Stoljećima su ljudi pokušavali analizirati podatke kako bi mogli donositi važne odluke. Egipćani su svoje podatke sačuvali u Aleksandrijskoj knjižnici, dok su stari Rimljani vodili evidenciju o optimalnoj raspodjeli vojske. Iz evolucijske perspektive, veličina Big Data se uvijek razvijala.

Tijekom 90-ih obujam informacija se često mjerio u terabajtima. Većina organizacija je analizirala strukturirane podatke u redovima i stupcima i koristila relacijske baze podataka i skladišta podataka za upravljanje velikim trgovinama podataka o poduzeću. U sljedećem desetljeću je došlo do širenja različitih vrsta izvora podataka pa su se podatci počeli mjeriti u petabyte-ima [1].

Prvi podatkovni centar je izgrađen u SAD-u 1965. godine radi pohranjivanja milijuna poreznih prijava i otisaka prstiju. Termin *Big Data* počeo se koristiti u ranim 1990.-ima bez imena, kada je prva kompanija, *Teradata*, analizirala i pohranila 1 terabajt podataka. Ista ta kompanija je 2007. godine instalirala prvi RDBMS[1] (eng. Relational Database Management System) temeljen na petabajtima. Svijetu je prvi put izraz Big Data predstavio Roger Mougalas 2005. godine [3]. Na početku 2006. godine razvijen je Hadoop, okvir otvorenog koda posebno stvoren za pohranu i analizu velikih skupova podataka. Razvoj otvorenog koda je bio vrlo ključan za rast velikih podataka jer omogućuje lakšu uporabu i jeftiniju pohranu ukupne količine podataka. Do 2008. godine su se uglavnom koristili strukturirani tipovi podataka. U posljednja 2 desetljeća brzina i volumen, kojima se generiraju podatci, promijenili su se izvan ljudskog očekivanja. 2013. godine ukupna količina podataka u svijetu je iznosila 4,4 zettabajta, što se u 2020. povećalo na 44 zettabajta, odnosno 10 puta više(44 zettabajta ≈ 44 bilijuna GB) [4]. Također, 2020. godine se stvara oko 1,7 megabajta novih podataka svake sekunde [2]. Slika 6 prikazuje predviđanje rasta podataka u zettabyte-ima prema IDC-u. Ako bismo mogli preuzeti 175 zettabyte-a na današnjem najvećem tvrdom disku, trebalo bi 12,5 milijardi pogona [11]. Sve više dominira cloud pohrana.



Slika 6 - Predviđeni rast podataka do 2025. godine

[1] RDBMS je softverski sustav koji se koristi za održavanje relacijskih baza podataka.

4. Primjena Big Data u praksi

U ovom poglavlju objasnit će se primjena Big Data u praksi, a naglasak će biti na najuspješnijim tvrtkama i društvenim mrežama za analizu podataka velikog obujma.

Svako poduzeće mora prodati proizvode i usluge da bi preživjelo. Kako bi to postiglo, mora pronaći kupce kojima će prodati. Tradicionalno, to je učinjeno oglašavanjem na "emitirani" način: novine, radio, televizija i prikaz oglašavanja rade na principu da ako se oglas postavi na najistaknutije mjesto, veliki broj ljudi će ga vidjeti i neki će od njih vjerojatno biti zainteresirani za ono što se nudi. Za veliku multinacionalnu tvrtku jasno je da će TV spot za vrijeme Super Bowla povećati njezinu izloženost i staviti svoj brend pred potencijalne kupce. Dok mala tvrtka mora mnogo detaljnije razmisliti o najučinkovitijem načinu trošenja svog ograničenog marketinškog proračuna. Takve tvrtke sebi ne mogu priuštiti pokriće svih osnova, tako da alati koji im mogu pomoći da utvrde tko su njihovi kupci i gdje ih pronaći mogu biti od velike koristi.

Slika 7 - Big Data u praksi

Slika 7 grafički prikazuje gdje se sve nalazi Big Data. Stvari poput društvenih medija, web knjiga, glazbe, videa, senzora su povećali količinu podataka u realnom svijetu. Ti podatci postali dostupni za analizu. Ono što se dijeli, čita, prenosi, šalje, gleda putem pametnih telefona, kompjutera, laptopa, iPada, sve se to analizira i negdje pohranjuje kako bi nam ponovna dostupnost tih podataka bila olakšana.
Širenje internetskog svijeta u posljednja 2 desetljeća omogućila je poduzetnicima i oglašivačima da unaprijede eksponiranje vlastitih proizvoda i usluga. Velikim tvrtkama poput Facebook-a, Google-a i Microsoft-a podatci su primarni proizvod. Ovisno o količini prikupljenih podataka, stopa usluge se povećava. Što je više podataka, bolja je kvaliteta, produktivnija je usluga i veća je zarada, a samim time i koriste se bolji alati za analizu još veće količine podataka. Društveni mediji ubrzavaju inovacije, ostvaruju uštedu troškova i jačaju brendove kroz masovnu suradnju. Facebook, kao najveća društvena mreža s 1,5 milijardi aktivnih mjesečnih korisnika, ima pristup daleko većem broju korisničkih podataka

nego bilo tko drugi. Google usluge mogu pratiti posjete svakog pojedinog korisnika na svim web stranicama. Microsoft ima snažne rezultate za pravilno predviđanje glavnih trendova u računarstvu.

Slika 8 - Grafički prikaz primjene Big Data

Slika 8 prikazuje u kojim se sve granama danas primjenjuje Big Data tehnologija. Veliki podatci se koriste u zdravstvu za mapiranje izbijanja bolesti i testiranje alternativnih tretmana. NASA koristi Big Data za istraživanje svemira. Glazbena industrija zamjenjuje intuiciju sa Big Data studijima. Veliki podatci se koriste i u suzbijanju cyber kriminala. Automobilska industrija se toliko razvila da za upravljanje volanom više nije potreban čovjek. Sve više filmova koristi CGI, računalno generiranje slike, za stvaranje bolje vizualnosti, efekata i postizanje željenog rezultata ubacivanjem holograma ili simulatora. Upotrebom Big Data tehnologije, poljoprivrednici također mogu povećati produktivnost i prihode vlastitog uzgoja. Jedan od takvih primjera su inteligentne sonde koje prate razinu vode u zemlji.
Povrh svega, imamo podatke sa svih senzora kojima smo okruženi. Pametni telefoni imaju senzore koji govore gdje smo (GPS), kojom se brzinom krećemo (akcelerometar), kakvo je vrijeme oko nas (barometar), kojom silom pritišćemo touch screen (senzor osjetljiv na dodir) i još puno toga. U svijetu, do 2020. godine postoji preko 60 milijardi pametnih telefona, što znači da su svi puni senzora koji prikupljaju podatke. Uz mobitele, postoje i pametne televizije, satovi, brojila, frižideri, automobili, žarulje, tenisice. Sve to ukazuje da će količina i raznolikost podataka u svijetu porasti do nezamislivih razina [1].

Zbog velikog broja primjene, Big Data obuhvaća sve vrste industrije, od zdravstvene zaštite, financija i osiguranja, do akademskog neprofitnog sektora. Postoje različiti načini pomoću kojih se vrijednost organizacije može ocijeniti na temelju prikupljenih podataka i kako poduzeća mogu utjecati na postizanje lakšeg rasta ili učinkovitosti. Poduzeća mogu zabilježiti razvoj vlastite vrijednost iz sakupljenih podataka na jedan od sljedećih načina [9]:

1) Stvaranje transparentnosti
   Korištenje podataka organizacije za određivanje budućih odluka čini organizaciju sve transparentnijom i probija jaz između različitih odjela. Podatci velikog obujma se analiziraju kroz različite granice koje mogu utvrditi različite neučinkovitosti. Primjerice, u proizvodnji Big data može pomoći identificirati mogućnost poboljšanja u odjelima za istraživanje i razvoj kako bi se novi proizvodi brže plasirali na tržište.

2) Otkrivanje podataka
   Poduzeća sve više stvaraju i pohranjuju podatke u digitalnom obliku. Na taj način postaje sve više dostupnih podataka o uspješnosti. U industriji osiguranja Big Data može pomoći odrediti profitabilne proizvode i pružiti poboljšane načine izračuna premije osiguranja.

3) Segmentacija i prilagođavanje
   Analiza velikih podataka pruža poboljšanu priliku za prilagodbu ponude na tržištu određenim segmentima kupaca u cilju povećanja prihoda. Podatci o korisničkom ponašanju omogućuju izgradnju različitih korisničkih profila koji se u skladu s tim mogu ciljano ponašati.

4) Moć automatizacije
   Algoritmi koji analiziraju skupove velikih podataka mogu se upotrijebiti za zamjenu ručnih odluka i izračunavanja radnog intenziteta automatiziranim odlukama. Automatizacija može optimizirati poslovne procese i poboljšati točnost ili vrijeme odziva.

5) Inovacije i novi proizvodi
   Podatci velikog obujma mogu otkriti obrasce koji identificiraju potrebu za novim proizvodima ili povećati dizajn trenutnih proizvoda ili usluga. Analizirajući podatke prema kupnji ili količini pretraživanja, organizacije mogu prepoznati potražnju proizvoda za koje organizacija možda nije svjesna.

4.1. FACEBOOK

Danas najpopularnija svjetska mreža društvenih medija nastala je 2005. godine. Uz Google, Facebook je vjerojatno jedina tvrtka koja posjeduje visoku razinu detaljnih podataka o korisnicima. Najposjećenija je web stranica na svijetu s Google-ove tražilice, a najčešća stvar koju Google koristi za pretraživanje je Facebook. Slika 9 pokazuje povezanost Facebook korisnika diljem svijeta i način na koji se industrijske tvrtke koriste platformama za društvene medije kako bi pratile o svojim uslugama i proizvodima nadzirući što korisnici pretražuju, komentiraju, lajkaju i objavljuju na Facebook-u i bilo kojoj web stranici.

Slika 9 - Facebook analiza podataka

2010. godine Facebook počinje koristiti HBase kao svoju korisničku infrastrukturu za razmjenu poruka u koju je smješteno 350 milijuna korisnika koji su mjesečno slali 15 milijardi poruka. HBase je baza podataka otvorenog koda, Slika 10, koja se temelji na Hadoop-u, distribuiranom sustavu otvorenog koda – Slika 11, a smještena je u strukturiranom sloju za pohranu. Tablica 1 predstavlja detaljnije karakteristike HBase i Hadoop-a. Posjeduje karakteristike visokih performansi, pohrane u stupcima, skalabilnosti i čitanja i pisanja u stvarnom vremenu [4]. Oslanja se na tehnologiju otvorenog koda za softver napisan u PHP-u i pokreće MySQL baze podataka. Njegovi programeri su stvorili HipHop[2] za MySQL kompajler, što prevodi PHP kod u C++ tijekom izvođenja, pritom omogućavajući brže izvođenje koda i redukciju opterećenja CPU-a [2].

Svakih 60 sekundi prenese se 136 tisuća fotografija, objavi 510 tisuća komentara i 293 tisuće statusa [12]. Iako nam isprva ti podatci možda ne znače mnogo, ali zato Facebook zna tko su naši prijatelji, kako izgledamo, čime se bavimo, što volimo, gdje se nalazimo i još mnogo toga. Neki istraživači kažu da Facebook ima dovoljno podataka da nas poznaje bolje od terapeuta.

[2] HipHop (HPHPc) je prekinuti PHP transpiler kreiran od strane Facebook-a. Prevodi PHP kod u C++, kompajlira se u binarni zapis i izvodi kao izvršna datoteka.

Slika 11 - Logo HBase

Slika 10 - Logo Hadoop

Tablica 1 – Karakteristike HBase i Hadoop

HBASE	HADOOP
Baza podataka otvorenog koda orijentirana na stupce i retke. Dizajnirana za pohranu ogromne količine podataka. Pruža lakši pristup podatcima kojima se može manipulirati koristeći MapReduce infrastrukturu. Izgradnja jednostavnih upita vrši se pomoću sučelja. Koristi HDFS sustav za pohranu – ima sposobnost pohrane velike količine podataka kroz distribuirane čvorove otporne na greške [3].	Najistaknutiji i najkorišteniji alat u industriji velikih podataka. Distribuirani sustav otvorenog koda za pohranu i upravljanje strukturiranih i nestrukturiranih podataka. Upravlja pohranom i analizom velikih podataka putem povezanih baza podataka i poslužitelja [2]. Sastoji se od 4 dijela: HDFS (eng. Hadoop Distributed File System), MapReduce, YARN, Libraries

Neki od načina utvrđivanja ponašanja korisnika [12]:

1. Praćenje kolačića – uz pomoć kolačića za praćenje Facebook „nadzire“ sve svoje korisnike. Ukoliko je korisnik prijavljen putem svog korisničkog računa i istovremeno pregledava druge web stranice, Facebook ga može pratiti.

2. Facijalno prepoznavanje ili prepoznavanje lica – jedno je od najnovijih Facebook-ovih ulaganja.

3. Označavanje „tagiranje“ prijedloga - kada podijelimo neku fotografiju, nudi se mogućnost dodavanja oznake pojedine osobe koja se na slici nalazi. Odnosno, kada se prepozna lice osobe, automatski nam dolazi pitanje „Who is this?“.

4. Analiza lajkova – kod ove vrste utvrđivanja ponašanja zanimljivo je da na temelju onoga što nam se sviđa, i tomu dodijelimo prikladnu oznaku – lajk, Facebook može precizno predvidjeti našu inteligenciju, zadovoljstvo životom, emocionalnu stabilnost, religiju, status odnosa, dob, političke stavove itd.

Nekoliko primjera kako Facebook koristi Big Data:

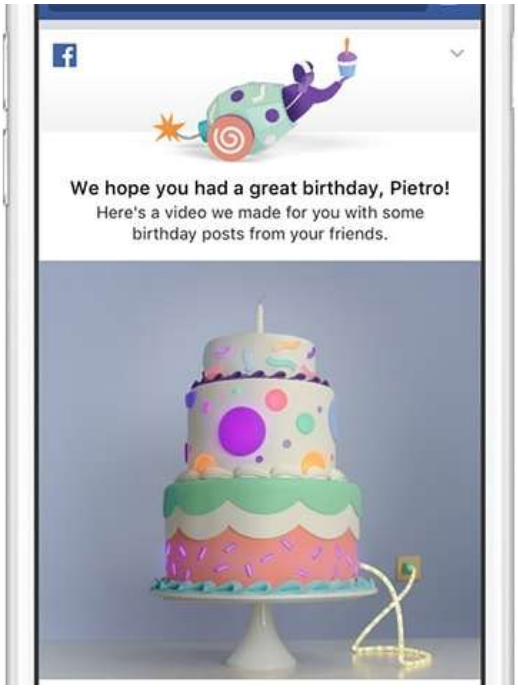
- Jedan od svakidašnjih i najčešćih primjera korištenja Big Data tehnologije su zasigurno rođendanske čestitke i drugi događaji. Ukoliko *scrollate* po Facebook-u i naletite na događaj poput koncerata, proslave ili bilo kakvih okupljanja i reagirate potvrdnim odgovorom, Facebook će taj dan poslati obavijest kao podsjetnik – slika 12.

Slika 12 - Primjer rođendanskog podsjetnik

- Uspješno je povezao političku aktivnost s angažmanom korisnika kada su izašli sa društvenim eksperimentom stvorivši naljepnicu koja omogućuje korisnicima da na svojim profilima proglase da „su glasali“ – slika 13.

Slika 13 - Primjer oznake za glasanje

- Povodom svoje 10. obljetnice, svojim korisnicima je ponudio mogućnost gledanja i dijeljenja videa koji prati tijek aktivnosti na njihovim društvenim mrežama od datuma registracije do danas. Naziva se „Flashback“, a sastoji se od zbirki fotografija i objava koji su dobili najviše lajkova i komentara. Prikazano na slici 14.



Slika 14 - Primjer Flashback-a

2012. godina je bila vrlo profitabilna za Facebook. Dogodile su se mnoge znatne promjene u količini podataka, kao i u pohrani tih istih.

- Kompanija je proizvela preko 200 terabajta podataka u sat vremena što je povećalo generiranje podataka u posljednje 2 godine za 90% [4]
- U 1 sekundi širom svijeta objavljeno je preko 700 ažuriranja statusa od strane korisnika

U ožujku 2013. Facebook je objavio novu značajku pod nazivom "Pretraživanje grafikona" koja omogućuje korisnicima i programerima da pretražuju grafičke prikaze na društvenim mrežama za ljude sa sličnim interesima, hobijima i zajedničkim lokacijama [1].
U tipičnoj implementaciji RDBMS-a, novi stupci zahtijevaju uključivanje DBA (eng. Database Analytics) kako bi se promijenila struktura tablice. Izvješće o marketingu društvenih medija za 2019. godinu navodi da je Facebook prva društvena platforma za marketing.

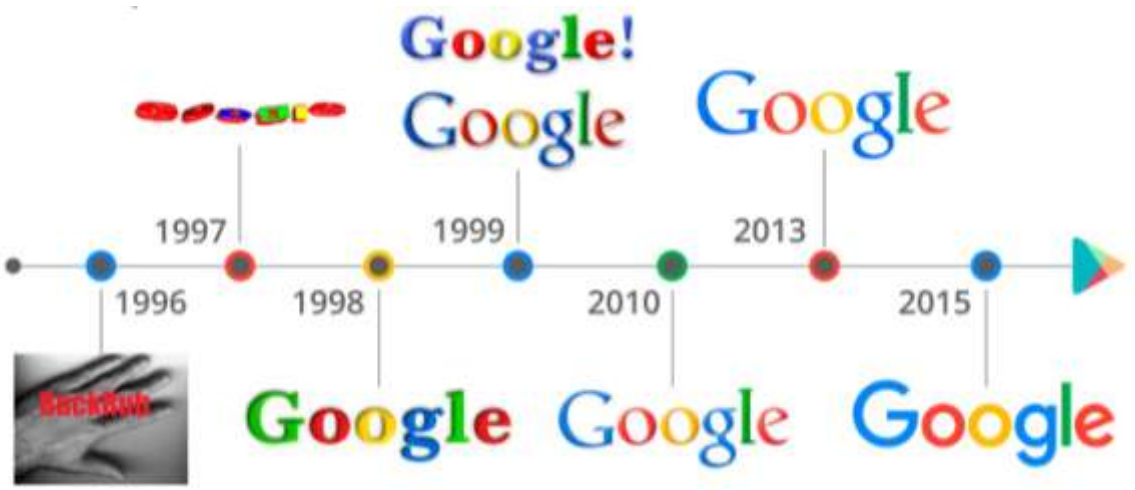
LOŠE STRANE FACEBOOK-A

Facebook je uvijek svoje korisnike uvjeravao da se informacije dijele samo uz njihovo dopuštenje i anonimnost kada se prodaju trgovcima. Međutim, čini se da problemi i dalje postoje. Od nastanka Facebook-a postojao je visoki stupanj zabrinutosti o privatnosti – slika 15. Mnogi korisnici se žale da postavke privatnosti nisu jasno objašnjene ili su prekomplicirane jer korisnici mogu jednostavno nenamjerno dijeliti svoje podatke. Da bi se osigurala privatnost, svi podatci se anonimno dodaju u sustave što znači da je korisničko ime uklonjeno i zamijenjeno jedinstvenim identifikacijskim kodom koji se ne može pridružiti. Nešto više o dobrim i lošim stranama Big Data tehnologije u *4.1. Prednosti Big Data* i *4.2. Nedostaci Big Data*.

Slika 15 - Privatnost na Facebook-u

4.2. GOOGLE

Najveća multinacionalna korporacija koju su osnovali Larry Page i Sergey Brin 1998. godine. Kao studenti, izradili su tražilicu 1996. godine pod nazivom „*BackRub*“ čiji je cilj bio razvoj tehnologija koje bi mogle stvoriti univerzalnu digitalnu biblioteku. Evolucija Google-ovog loga traje od njegova nastanka pa sve do danas. Slika 16 prikazuje lentu vremena 8 različitih promjena loga. Google je dobio naziv po broju 10^{100} koji ukazuje na beskonačnu količinu informacija koje postoje. Nagli razvoj zahvaljujući brojnim uslugama koje nudi svojim korisnicima doživljava 2000. godine. Tijekom godina, lansirao je niz proizvoda uključujući karte (eng. Google Maps[3]), gmail, internet preglednik Chrome i operacijski sustav za Android mobilne uređaje koji je besplatan za proizvođače pametnih telefona. Na početku je pretraživanje bilo moguće isključivo na engleskom jeziku, dok se danas broj povećao na 149 različitih jezika.

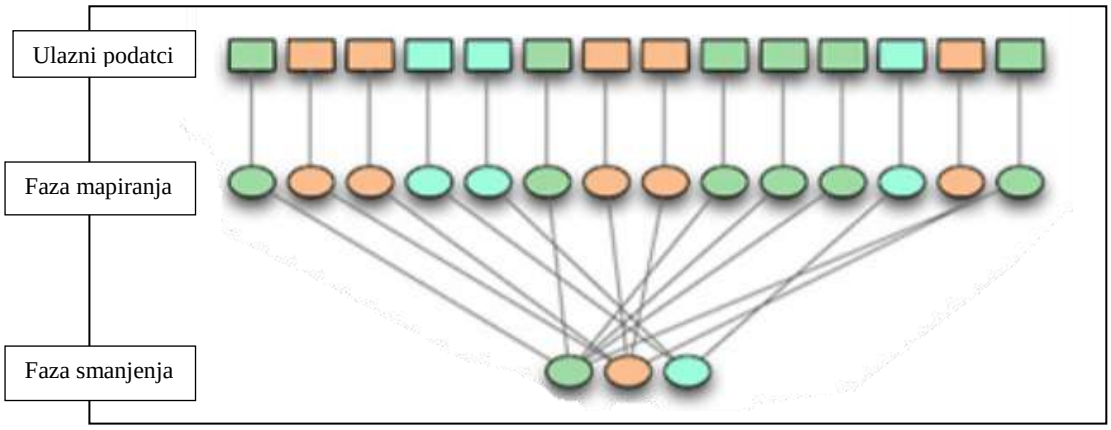


Slika 16 - Logo evolucija Google-a

Znanost podataka nije samo postojanje podataka ili nagađanje o tome što bi podatci mogli značiti; radi se o testiranju hipoteza i provjeri jesu li zaključci koji se izvlače iz podataka valjani. Prepoznavanje govora je oduvijek bio nerješiv problem. Međutim, Google je uspio integrirati glasovno pretraživanje u svoju tražilicu koristeći prikupljene glasovne podatke. Uz to, izrađen je i rječnik uobičajenih pravopisnih grešaka kod pretraživanja te njihovi ispravci i konteksti u kojima se pojavljuju. *PageRank*[4] algoritam je među prvima koristio podatke izvan same stranice; posebno broj veza koji upućuju na stranicu. Praćenjem tih veza, učinio je Google pretraživanja mnogo korisnijima. Za vrijeme svinjske gripe 2009. godine, Google je mogao pratiti napredak epidemije prateći pretraživanja na temu povezanih s gripom [3].

[3] Google karte dostupne su na stranici https://www.google.com/maps. To je usluga koja nudi satelitski prikaz gradova, ulica, naselja, cesta, putokaza za automobile, bicikliste i pješake, itd.
[4] PageRank (PR) je algoritam koji Google koristi za rangiranje web stranica u rezultatima tražilice. Ime je dobio po osnivaču Google-a, Larry-u Page. Način mjerenja važnosti web stranica.

Da bi ostvario izazov obrade podataka velikog obujma, Google je kreirao paralelni programski model, *MapReduce* kao pokretača većeg dijela današnje velike obrade podataka. Ima mogućnost da preuzme upit preko skupa podataka, podijeli ga i pokrene paralelno izvođenje na mnogo čvorova. Računanje izvodi kao evaluaciju matematičkih funkcija.



Slika 17 - Način rada MapReduce-a

Rad MapReduce-a sastoji se od ulaznih podataka, faze mapiranja, obrade, faze smanjenja i izlaznih podataka - prikazano na Slici 17. Zadatak je da prebroji jedinstvene riječi u dokumentu, što znači da se u fazi preslikavanja ili mapiranja svaka riječ identificira i dodjeljuje joj se broj 1. U fazi smanjenja se broji ukupan zbroj svake riječi. Da bi se zadatak izvršio, faza preslikavanja i faza smanjenja moraju se pridržavati ograničenja koja omogućuju paralelno izvođenje. Prevođenje upita u jedan ili više MapReduce koraka nije intuitivan proces.

Kod rješavanja problema pomoću MapReduce-a imamo [3]:

- Učitavanje podataka – podatci moraju biti izvađeni iz izvora, strukturirani tako da budu spremni za obradu i učitani u sloj za spremanje programa MapReduce
- MapReduce – dohvaća podatke iz pohrane, obrađuje ih i vraća rezultate u pohranu
- Izvođenje rezultata – preuzimanje podataka iz pohrane i predstavljanje ljudima

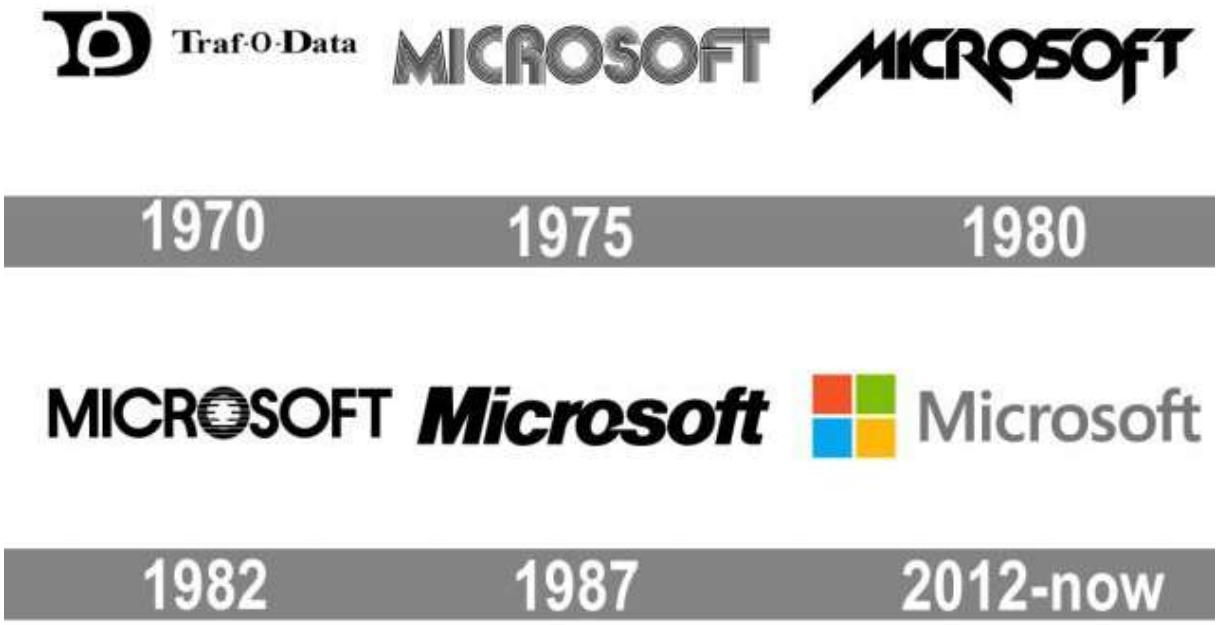
BigTable

Dizajn HBase je zasnovan prema Google-ovom dokumentu BigTable iz 2006. godine. BigTable je distribuirani sustav za pohranu i upravljanje strukturiranih podataka čiji je razvoj počeo 2004 godine. Sustavi poput tražilice, karata i YouTube-a koji ostavljaju velike količine podataka na dohvat ruke, temelje se na vlastitom okviru baze podataka i analiza koji se nazivaju BigTable i BigQuery [2]. To je skalabilna baza podataka visokih performansi izgrađena na Google-ovom datotečnom sustavu i koristi se u mnogim uslugama visoke skalabilnosti. Tek 2015. godine javna verzija je postala dostupna kao usluga – prikaz loga BigTable-a na Slici 18. Koriste ga razni Googleovi proizvodi: Google Analytics, Google Earth, Orkut,…

Slika 18 - Logo BigTablea

4.3.MICROSOFT

Softverska tvrtka koju su osnovali 1975. godine Bill Gates i Paul Allen. Glavna ideja je bila razvoj računalnog softvera poput operacijskih sustava, razvojnih alata, uredskih programa, baza podataka, itd [6]. Na tržištu operacijskih sustava za osobna računala ima daleko vodeću ulogu. Velika razlika između onoga što su ljudi sposobni zamisliti i onoga što su sposobni izgraditi dovela je do pojave mnogih poduzeća koja nude DaaS[5] (eng. Data as a service) ili SaaS[6] (eng. Software as a service) rješenja. Tu se Microsoft nalazi na 1. mjestu, baš kao što su nudili operativne sustave poput MS-DOS[7]-a, Windows[8]-a, komercijalne softvere za produktivnost poput Office[9]-a i web preglednika [2].



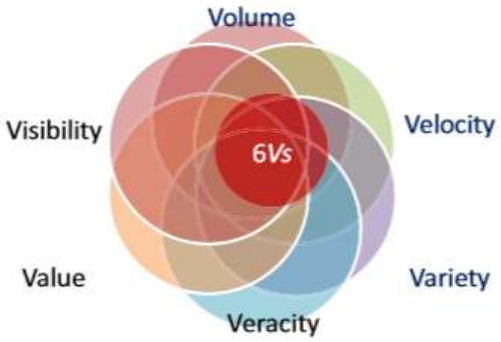
Slika 19 - Logo Microsoft-a

Kao i Google, Microsoft je također mijenjao svoj logo s godinama – slika 19. Najraniji se pojavio 1975. godine. 1970. godine, *Traf-O-Data*, je bilo poslovno partnerstvo Bill Gatesa, Paul Allena i Pol Gilberta. Cilj je bio da se pročitaju podatci s brojača prometnica i stvore izvještaji za inženjere prometa što je bilo ključno za stvaranje Microsoft korporacije nekoliko godina kasnije. Posljednja verzija loga se pojavila 2012. godine. Označava prelazak s klasičnog sučelja na moderne pločice od kojih svaka predstavlja jedan od glavnih proizvoda tvrtke – Windows, Xbox, Bing i Office.

[5] Daas – strategija upravljanja podatcima koja koristi oblak za isporuku podataka za pohranu, obradu i analitiku putem Interneta.
[6] Saas - infrastruktura oblaka i aplikacije smještene u njoj su dostupne korisniku.
[7] MS-DOS(MicroSoft Disk Operating System) je proizvod tvrtke Microsoft. Operacijski sustav koji se isključivo bavio diskovima ili tekstualnim sučeljima.
[8] Microsoft Windows je naziv za grupu operacijskih sustava tvrtke Microsoft. Postoji 27 klijentskih verzija Windowsa, a najnovija je Windows 10.
[9] Microsoft Office je programski paket namijenjen uredskoj obradi podataka. Najnovija verzija je iz 2019.

Kako bi povećao poslovnu vrijednost, Microsoft je produžio 3V atribute na 6V atribute dodavši varijabilnost, istinitost i vidljivost [4]. Prikazano na slici 20.

1. Količina podataka predstavlja opseg podataka
2. Brzina podataka označava analizu strujnih podataka
3. Raznolikost označava različite oblike podataka
4. Istinitost se fokusira na pouzdanosti izvora podataka
5. Varijabilnost se bazire na složenosti podataka
6. Vidljivost naglašava cjelovitost podataka kako bi se mogla donijeti informativna odluka



Slika 20 - Microsoft-ovih 6 atributa

Microsoft, kao Facebook i Google također prikuplja podatke svojih korisnika. To čini na način na koji koristimo računalo, učestalost pada softvera, ažuriranje Windowsa, Office-a ili nekog drugog programa, izvršavanje raznih zadataka na računalu. Manjak vlastite platforme za pametne uređaje doveo je Microsoft u nepovoljan položaj pokraj Google-a i Apple-a.

5. Prednosti i nedostaci Big Data

Ogromne količine podataka nose svoje pozitivne i negativne strane. Općenito, kada postoji više podataka o nekom proizvodu, osobi, predmetu, kompaniji, to ljudski um čini informiranijim i bogatijim. Već se otprilike zna što će se kupiti, s kim se družiti, gdje se zaposliti, na koji način nešto funkcionira i kako se koristi. Veća količina podataka pomaže u podjeli između onoga što je dobro i onoga što je loše. Mogućnost prikupljanja informacija ne dolazi sama od sebe. Potrebno je uložiti određeni napor, vrijeme i novac za ostvarenje dobiti. Isto to vrijedi i u marketingu. Imati više podataka o potencijalnim kupcima, omogućuje tvrtkama da prilagode svoje proizvode i napore kako bi stvorili najvišu razinu zadovoljstva i ponovili poslovanje. Onim tvrtkama koje imaju mogućnost prikupljanja veće količine podataka, otvorena su vrata za dublju i bogatiju analizu.

5.1. Prednosti Big Data

Većina tvrtki smatra da su prednosti velikih podataka jako velike, te su one sumirane u Tablici 2.

Tablica 2 - Prednosti Big Data

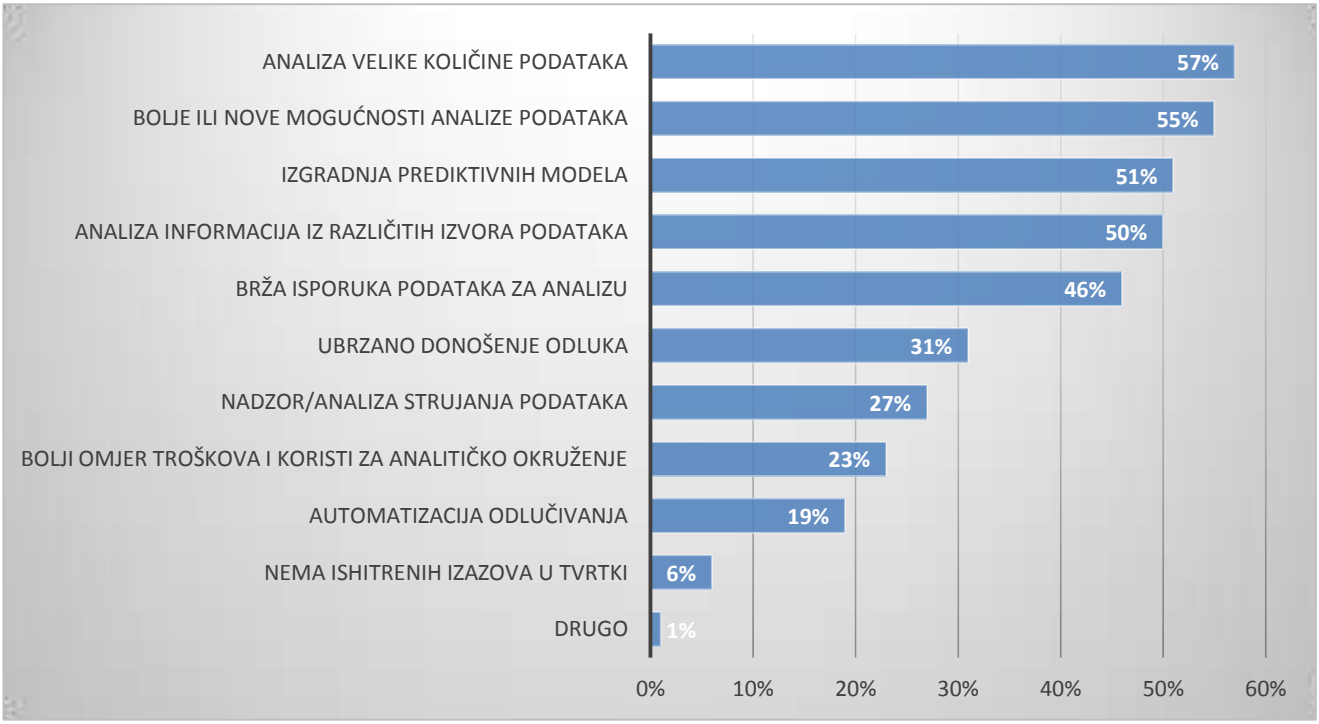
Donošenje boljih odluka
Povećanje produktivnosti
Smanjenje troškova
Poboljšanje korisničke usluge
Otkrivanje prijevara
Povećanje prihoda
Povećanje agilnosti
Veće inovacije
Veća brzina tržišta

Svako dobro odlučivanje je bitno i korisno. Donošenje boljih odluka dovodi bližem ostvarenju cilja. Analiza, donositeljima poslovnih odluka, pruža uvide temeljene na potrebnim podatcima kako bi im pomogle za rast i razvoj tvrtke. Istraživanjem dobavljača Syncsort[10] je utvrđeno da 59,9 % ispitanika od velikih alata koriste Hadoop i Spark za povećanje produktivnosti poslovnih korisnika [7]. Moderni alati omogućuju analitičarima bržu analizu više podataka, što povećava osobnu produktivnost. Velika analiza podataka pomaže tvrtkama da smanje svoje troškove. Pojedine organizacije odabiru smanjenje troškova kao svoj glavni cilj analize velikih podataka kako bi postigle što bolji uspjeh u što manjem financijskom režimu. Socijalni mediji, sustavi upravljanja odnosa s kupcima pružaju poduzećima informacije o svojim kupcima kako bi ih iskoristili za bolje opsluživanje tih kupaca. Jedna od glavnih prednosti sustava za veliku analizu podataka je što su izvrsni u otkrivanju obrazaca i anomalija. Sposobnosti otkrivanja prijevara mogu bankama i tvrtkama s kreditnim karticama

[10] Syncsort je globalna softverska tvrtka specijalizirana za velike podatke, proizvode za brzo sortiranje podataka, kvalitetu podataka i ponudu podataka o lokaciji za IBM, Hadoop, Microsoft, UNIX, Linux i mainframe sisteme.

dati mogućnost da uoče ukradene kreditne kartice ili lažne kupnje čak i prije nego vlasnik kartice sazna da nešto nije uredu. Kada kompanije koriste velike podatke za donošenje boljih odluka i pružanje poboljšane usluge kupcima, povećani prihod je često prirodni rezultat.

Organizacije koriste analitiku velikih podataka iz različitih razloga. Slika 21 prikazuje koji su najutjecajniji razlozi [14]:



Slika 21 - Zašto tvrtke koriste Big Data analizu

Analiza velike količine podataka se sastoji od volumena, raznolikosti i brzine. Te tri ključne karakteristike čine najveći udio postotka zbog kojih razne tvrtke koriste analitiku podataka velikog obujma. Sljedeća dva razloga, bolje ili nove mogućnosti analize podataka i izgradnja budućih modela omogućavaju smanjenu neizvjesnost i povećanu predvidljivost nadolazećih proizvoda. Te dvije statistike se vode kao važne komponente u provođenju inicijative za velike podatke i izvlačenju većih vrijednosti iz podataka.

5.2. Nedostaci Big Data

Tvrtke moraju biti u mogućnosti obraditi veće količine podataka, za vrijeme dok određuju podatke koji predstavljaju pozitivne signale u usporedbi sa onima koje zahtijevaju troškove. S druge strane, iako je bolja analiza velike količine podataka pozitivnija, postoje i mnogi drugi razlozi koji dovode do preopterećenosti i velikih troškova, odnosno gubitaka tvrtke. U Tablici 3 prikazani su najveći nedostaci za analizu velikih podataka [7]:

Tablica 3 – Nedostaci Big Data

Potreba za talentom
Kvaliteta podataka
Potreba za kulturnim promjenama
Usklađenost
Cybersecurity rizici
Nagle promjene
Potreba za hardverom
Trošak
Poteškoće s integriranjem naslijeđenih sustava

Nedostatak velikog broja vještina za analizu i obradu podataka je najveći izazov za velike podatke. Zapošljavanje ili osposobljavanje osoblja može značajno povećati troškove, a postupak stjecanja vještina za analizu i obradu velikih podataka može trajati dosta dugo vremena. Zbog toga su svugdje manje-više traženi „gotovi proizvodi“. Kako bi se ljudima pružilo što pouzdanije i originalnije, ali isto tako točnije informacije, javlja se potreba za kvalitetom podataka. Prije same uporabe prikupljenih podataka za analizu, mora se osigurati točnost, relevantnost i pravilni oblik informacija. Spremanje velikih podataka, osobito osjetljivih podataka, može učiniti tvrtke privlačnijom metom za cyber napad, što se smatra najvećom prijetnjom podataka. Potreban prostor za pohranu podataka, prijenos u analitičke sustave i iz njih, te izračunavanje resursa za obavljanje analitike je skupo za održavanje. Mnogi današnji alati za obradu podataka velikog obujma se oslanjaju na tehnologiju otvorenog koda koja drastično smanjuje troškove softvera, ali se zato povećavaju troškovi osoblja, hardvera, održavanja i usluga.

6.Alati za analizu podataka velikog obujma

Novi alati i tehnologije omogućuju stvaranje i upravljanje velikim skupovima podataka i okruženjem za pohranu u kojem se nalaze. Često su ti alati ograničeni za analizu u memoriji na radnim površinama koja analizira uzorke podataka, a ne čitavu populaciju skupa podataka [1]. Struktura podataka je jedan od faktora koji diktira alate i analitičke tehnike. Npr. ovisno o tome planira li tim analizirati tekstualne ili transakcijske podatke, potrebni su različiti pristupi i alati. Loša strana alata je da mogu imati ograničenja kod primjene modela za vrlo velike skupove podataka, što je uobičajeno za Big Data. Alati za pomoć su prvenstveno fokusirani na statističku analizu ili softver za vađenje podataka.
Primjeri besplatnih komercijalnih alata za istraživanje, modeliranje i prezentaciju podataka su SAS Entreprise Miner, SPSS Modeler, Matlab, Alpine Miner.

Primjeri besplatnih alata ili alata otvorenog koda za fazu planiranja modela jesu:

6.1. R alat

Popularan programski jezik za istraživanje, analitiku i vizualizaciju podataka. Projekt je osmišljen 1992. godine, a prvu verziju implementirali su Ross Ihaka i Robert Gentleman 1995. godine. Ima mogućnost povezivanja sa bazama podataka i izvršavanja statističkih ispitivanja i analiza za Big Data putem otvorenog koda. Sadrži gotovo 5000 paketa za analizu podataka i grafičku reprezentaciju. Izvodi se na Linux i Windows poslužitelju te na SQL serveru. Može se lako implementirati na drugim poslužiteljima. Logo alata prikazan je na slici 22.

Slika 22 - Logo R alata

Načini na koje je R[11] smislio rješenja za podatke velikog obujma [13]:

a) Eksplicitni paralelizam – istodobno pokretanje nekoliko računanja dijeljenjem podataka na više jezgara. Eksplicitno znači naglasiti sustavu svaki korak koji treba slijediti. Neki od primjera su: Rmpi (eng. Message Passing Interface), Snow, SnowFall, sfCluster.

b) Implicitni paralelizam – za razliku od eksplicitnog paralelizma, gdje korisnik kontrolira većinu postavki klastera, s ovom vrstom paralelizma je izbjegnut najveći dio neredovitog rada u postavljanju sustava i distribuciji podataka. Najpoznatiji primjeri paketa u R koji koriste implicitni paralelizam i mogu se koristiti za rukovanje s velikim podatcima su MapReduce i Hadoop.

[11] Preuzimanje alata R moguće je na stranici: https://cloud.r-project.org/

6.2.SQL (Structured Query Language) alat

Strukturirani upitni jezik koji se koristi za interakciju s relacijskim bazama podataka. Nastao je 1974. godine. Od sredine 1970-ih kada su relacijske baze podataka dominirale svijetom, SQL je bio sveprisutni način pristupa i manipuliranja podatcima. Više nije alat namijenjen programerima, administratorima baza podataka i analitičarima, nego je integriran u sve aplikacije koje koriste baze podataka. Jezik upita je jezik koji korisnicima omogućuje pristup i obradu baza podataka. Ima izravan odnos prema relacijskoj algebri – kombinacija odabira, projekcije, kartezijanski proizvodi i više operacija koje se mogu izvoditi na tablicama. Karakteristike SQL-a su jednostavnost korištenja, neproceduralnost i mogućnost interaktivnog i klasičnog programiranja. Slika 23 prikazuje logo SQL alata za analizu podataka velikog obujma.

Slika 23 - Logo SQL alata

Umjesto pitanja „kako?“, relacijski model pomoću SQL jezika naglašava „Što?“ treba dohvatiti. Tablica se kreira jednom naredbom i odmah nakon kreiranja dostupna je za korištenje. Kao jezik, SQL objedinjuje jezik za rukovanje podatcima (eng. Data Manipulation Language), jezik za opis podataka (eng. Data Definition Language) i jezik za upravljanje podatcima (eng. Data Control Language). Omogućava stvaranje i mijenjanje strukture baze podataka, dodavanja prava korisniku za pristup bazama podataka ili tablicama, traženje informacija i mijenjanje sadržaja baze podataka.

Za rješavanje različitih upita potrebno je poznavati SQL jezik, odnosno razne tipove podataka, naredbe, operatore, znakove i funkcije. Kod kreiranja tablica određuje se naziv stupaca te tip podatka koji će biti spremljen. Za izradu tablica koriste se naredbe CREATE TABLE, ALTER TABLE, DROP TABLE,… Neki od tipova podataka su CHAR, VARCHAR2, LONG, DATE, TIMESTAMP, NUMBER,… prikazano na slici ispod.

```
mysql> create table saveMovie (
    -> id int not null auto_increment,
    -> Name varchar(30) not null unique,
    -> Object1 varchar(60),
    -> Object2 varchar(60),
    -> Object3 varchar(60),
    -> Comment text,
    -> primary key (id));
Query OK, 0 rows affected (0.00 sec)

mysql> describe saveMovie;
+---------+-------------+------+-----+---------+----------------+
| Field   | Type        | Null | Key | Default | Extra          |
+---------+-------------+------+-----+---------+----------------+
| id      | int(11)     |      | PRI | NULL    | auto_increment |
| Name    | varchar(30) |      | UNI |         |                |
| Object1 | varchar(60) | YES  |     | NULL    |                |
| Object2 | varchar(60) | YES  |     | NULL    |                |
| Object3 | varchar(60) | YES  |     | NULL    |                |
| Comment | text        | YES  |     | NULL    |                |
+---------+-------------+------+-----+---------+----------------+
6 rows in set (0.00 sec)

mysql>
```

Slika 24 - Primjer izrade tablice u SQL alatu

Ukoliko je potrebno analizirati ili koristiti već postojeće podatke iz tablice, koristi se naredba SELECT. Za još opsežnije uvjete pretraživanja koriste se logički operatori AND, OR i NOT ili metoda LIKE, IS NULL, ORDER BY (ASC ili DESC), GROUP BY,… prikazano na slici ispod.

```
MySQL 5.5 Command Line Client
+----------+-----------------------+
| emp_name | Minimum working hour  |
+----------+-----------------------+
| Ajeet    |                    12 |
| Ayan     |                    10 |
| Milan    |                     9 |
| Ruchi    |                     6 |
+----------+-----------------------+
4 rows in set (0.00 sec)

mysql>
mysql> SELECT emp_name, AVG(working_hours) AS "Average working hour"
    -> FROM employees
    -> GROUP BY emp_name;
+----------+-----------------------+
| emp_name | Average working hour  |
+----------+-----------------------+
| Ajeet    |               12.0000 |
| Ayan     |               10.0000 |
| Milan    |                9.0000 |
| Ruchi    |                6.0000 |
+----------+-----------------------+
4 rows in set (0.00 sec)

mysql>
```

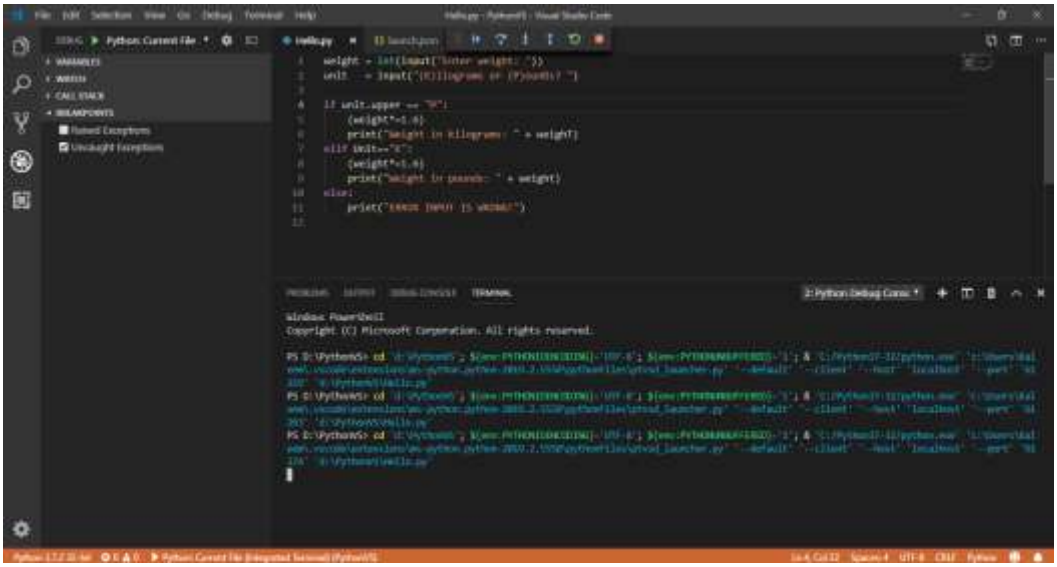
Slika 25 - Primjer grupiranja podataka

6.3.PYTHON alat

Programski jezik otvorenog koda 80-ih godina koji nudi alate za strojno učenje i analize, kao što su scikit-learning, numpy, scipy, pande i vizualizacija podataka pomoću matplotlib-a. Izumio ga je Guido van Rossum. Jezični konstruktori i objektno orijentirani pristup nastoje olakšati programeri u pisanju koda za velike i male projekte. Programiranje u Python-u zahtijeva manje redaka koda u odnosu na druge jezike dostupne za programiranje. Ima strukturu gniježđenja koja se temelji na uvlačenju redaka. Može obraditi dugotrajne zadatke u kratkom vremenom intervalu. Kako nema ograničenja za obradu podataka, može izračunati podatke u prijenosnom računalu, oblaku i radnoj površini. Primjer Python[12] loga prikazan je na slici 26.

Slika 26 - Logo Python alata

Pojavom platforme Anaconda[13], promijenila se brzina izvođenja programa u Python-u zbog čega je postao jedna od najpopularnijih opcija u industriji. Podržava više platformi te se može izvoditi na Windowsu i Linuxu. Python nudi upotrebu više biblioteka što ga čini poznatim jezikom u području znanstvenog računarstva, a takva karakteristika je potrebna za Big Data analizu. [15] Primjer Python sučelja prikazano je na slici ispod.



Slika 27 - Python sučelje

[12] Za detaljniju dokumentaciju alata Python otvoriti stranicu https://www.python.org/downloads/

[13] Anaconda je besplatna open-source distribucija programskog jezika Python i R za znanstveno računarstvo. Cilj joj je pojednostaviti upravljanje i implementaciju paketa.

7.Zaključak

Tijekom proteklog desetljeća veliki podatci su evoluirali i uvelike olakšali rješavanje problema u poslovanju i društvu. Dostupnost Big Data i novog softvera za upravljanje informacijama i analitičkim podatcima proizveli su jedinstven trenutak u povijesti analize podataka. Važnost velikih podataka se ne odnosi na količinu podataka već što se s njima može napraviti. Prednosti velikih podataka nose mnogo veću težinu, a njegove primjene u poslovanju, zdravstvu, prometu, upravljanju i šire trebaju se poticati. Zajedno s velikim podatcima dolazi i potencijal za otvaranje velikih uvida za svaku industriju, od velike do male. Većina podataka je danas nestrukturirana ili polu-strukturirana što povećava upotrebu raznih alata za obradu i analizu. Programska paradigma Hadoop i MapReduce imaju značajnu bazu na području analize, a takva se upotreba sve više povećava. Hadoop se brzo razvio u jedno od glavnih pokretača istraživanja i ekonomije. Tvrtke poput Facebooka, Googlea i Microsofta dnevno proizvode velike količine podataka, a za njihovu analizu i pohranu koriste alate poput R alata, Pythona ili SQL-a.

Iako postoji mnogo pozitivnih strana korištenja Big Data tehnologije, također postoje i one loše. Bitno je da velike podatke prilagodimo našim, ljudskim potrebama. Kako je veličina podataka porasla do nevjerojatnih razmjera, ljudska sposobnost donošenja potpuno intuitivnih odluka je smanjena. Kao rezultat toga, donošenje odluka utemeljenih na podatcima postalo je sve učestalije kako bi se osigurao razuman put ka uspjehu. Organizacije se „utapaju“ u podatcima, a trebale bi se prebaciti na potpuna algoritamska predviđanja usmjerena prema Big Data kako bi preživjele visoki konkurentni svijet. To stvara izazov iskorištavanja ogromne količine novih izvora podataka. Fokus se promijenio od pokušaja da se shvati pojam ovog fenomena do pronalaska opipljive vrijednosti u njegovoj primjeni. Ono će promijeniti način života, rada i razmišljanja, a takav sustav je potrebno prilagoditi vlastitim mogućnostima.

Na kraju, kada se sumiraju sve prednosti i nedostaci Big Data, većina organizacija smatra da prednosti nadmašuju nedostatke. Međutim, relativne nedostatke i prednosti podataka velikog obujma uvijek je potrebno pažljivo razmotriti prije pokretanja nekog projekta velikih podataka. Korištenjem različitih alata za analizu podataka velikog obujma, podatci velikog obujma sada imaju sredstva i održiv kontekst koji mogu koristiti u različite svrhe u poslovnom procesu organizacije.

Literatura

1. Wiley & Sons, J. (2015.). *Data Science & Data Analytics.* Indianapolis: Wiley.
2. Marr, B. (2016.). *Big Data in practice.* United Kingdom: Wiley.
3. Radar, O. (2011.). *Big Data Now.* Sebastopol: O'Reilly.
4. Buyya, R., Calheiros, R. N., & Dastjerdi, A. V. (2016.). *Big Data Prinicples and Paradigms.* Cambridge: Elsevier.
5. Oracle. *What is Big Data?* Dohvaćeno 3.8.2020. iz What is Big Data?: https://www.oracle.com/big-data/what-is-big-data.html
6. Wikipedia. *Microsoft*. Dohvaćeno 4.8.2020. iz: https://hr.wikipedia.org/wiki/Microsoft
7. Datamation. *Big Data pros and cons*. Dohvaćeno 3.8.2020. iz: https://www.datamation.com/big-data/big-data-pros-and-cons.html
8. Investopedia. *Definicija Big Data*. Dohvaćeno 8.8.2020. iz: https://www.investopedia.com/terms/b/big-data.asp
9. Framework. *Value of Big Data*. Dohvaćeno 8.8.2020. iz: https://www.bigdataframework.org/value-of-big-data/
10. Oracle. *What is Big Data*. Dohvaćeno 8.8.2020. iz: https://towardsdatascience.com/what-is-big-data-lets-answer-this-question-933b94709caf
11. Datanami. *Global datasphere*. Dohvaćeno 9.8.2020. iz: https://www.datanami.com/2018/11/27/global-datasphere-to-hit-175-zettabytes-by-2025-idc-says/
12. Simplilearn. *How Facebook is Using Big Data*. Dohvaćeno 9.8.2020. iz: https://www.simplilearn.com/how-facebook-is-using-big-data-article
13. Jigsawacademy. *Handling Big Data using R*. Dohvaćeno 10.8.2020. iz: https://www.jigsawacademy.com/handling-big-data-using-r/
14. BI-SURVEY. *Comapnies Use Big Data Analytics*. Dohvaćeno 9.8.2020. iz: https://bi-survey.com/companies-use-big-data-analytics
15. Science, T. d. *Why is Python programming a perfect fit for Big Data*. Dohvaćeno 24.8.2020. iz: https://towardsdatascience.com/why-is-python-programming-a-perfect-fit-for-big-data-5ac54ee8f95e

Popis slika

Popis tablica

ODJEL ZA INFORMATIKU
SVEUČILIŠTE U RIJECI

Rijeka, 18. lipnja 2020.

Zadatak za završni rad

Pristupnik: Katarina Brkljača

Naziv završnog rada: O podatcima velikog obujma Naziv

završnog rada na eng. jeziku: About Big Data

Sadržaj zadatka:

Podatci su postali ključno sredstvo za gospodarstvo i naše društvo. Big Data tehnologija, odnosno tehnologija velikih podataka, predstavlja velike mogućnosti jer pomaže u razvoju novih kreativnih proizvoda i usluga. Cilj rada je objasniti osnovne pojmove, povijesni razvoj, strukturu i karakteristike Big Data. Također, potrebno je navesti područja primjene te navesti prednosti i nedostatke korištenja ove tehnologije. Na kraju treba se osvrnuti i na alate za rad s podacima velikog obujma.

Mentor

Prof. dr.sc. Patrizia Poščić

Voditelj za završne radove

Dr. sc. Miran Pobar

Zadatak preuzet: 19. lipnja 2020.

Katarina Brkljača

(potpis pristupnika)