Big Data

Bertoša, Gordana

Master's thesis / Diplomski rad

2015

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka, Faculty of Humanities and Social Sciences / Sveučilište u Rijeci, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:186:207052

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-11-26**

FILOZOFSKI FAKULTET U RIJECI

Sveučilište u Rijeci
University of Rijeka

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Humanities and Social Sciences - FHSSRI Repository

zir.nsk.hr

UNIRI DIGITALNA KNJIŽNICA

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U RIJECI
ODJEL ZA INFORMATIKU

Gordana Bertoša

BIG DATA

DIPLOMSKI RAD

Rijeka, srpanj 2015.

SVEUČILIŠTE U RIJECI

ODJEL ZA INFORMATIKU

Gordana Bertoša

Studij: Njemački jezik i književnosti
i informatika

Diplomski studij

BIG DATA

DIPLOMSKI RAD

Mentor:
doc. dr.sc. Marina Ivašić-Kos

Rijeka, srpanj 2015.

ODJEL ZA INFORMATIKU
Povjerenstvo za diplomske ispite
Rijeka, 9.2.2015.

Z A D A T A K
za diplomski rad

Pristupnik: **Gordana Bertoša** Matični broj:

Naziv zadatka: *Big Data*

Sadržaj zadatka: Objasniti što je *Big Data*. Opisati glavne karakteristike *Big Data* i probleme koji se javljaju kod implementacije aplikacija za analizu i rad s podacima koji se smatraju *Big Data*. Dati tipične primjere sustava koji koriste *Big Data*.

Zadano: 9.2.2015

Mentor:

doc. dr. sc. Marina Ivašić Kos

Predsjednik Povjerenstva:

Zadatak preuzeo: 9.3.2015.

(potpis pristupnika)

Dostaviti:
- Predsjednik Povjerenstva
- Mentor
- Djelovođa povjerenstva
- Evidencija studija
- Pristupnik

IZJAVA

Izjavljujem da sam diplomski rad s naslovom *Big Data* izradila samostalno uz stručno vodstvo mentora doc. dr. sc. Marine Ivašić Kos. U radu sam koristila literaturu koja je navedena na kraju diplomskog rada. Rad je pisan u duhu hrvatskog jezika.

U Rijeci, srpanj 2015. Potpis

_____

Sadržaj

1. Uvod

Big Data je fenomen koji se zadnjih godina proširio i ima utjecaja u svim područjima. Taj fenomen uvidjele su mnoge tvrtke, te ga počele istraživati i primjenjivati u svom poslovanju. Ogromna količina podataka koja svakim danom raste treba biti istražena kako bi se ti podaci iskoristili u najbolju moguću svrhu.

Središnji cilj ovog diplomskog rada je teoretski prikaz i opis tog fenomena, te opis problema prilikom implementacije kao i utjecaj *Big Data* u područjima poslovanja.

Sam diplomski rad sastoji se od 12 glavnih dijelova.

Na početku samog rada opisan je pojam *Big Data*, te su opisane dimenzije koje ga sačinjavaju. Na samim primjerima je prikazano koliko daleko seže ta velika količina podataka. Nadalje je u radu prikazan utjecaj poslovne inteligencije na veliku količinu podataka. Velika količina podataka može se pojavljivati u različitim oblicima, te je predstavljeno koji se problemi prilikom toga javljaju. Zatim je prikazano kako prikupljanje, organiziranje i analiza velike količine podataka ima utjecaj na poslovanje. Prikazani su i pozitivni primjeri primjene *Big Data* analitike. Kako se značaj velike količine podataka počeo otkrivati zadnjih nekoliko godina, tvrtke nisu pripremljene na to. Iz tog razloga opisano je kako da se tvrtke pripreme na tu količinu podataka, kako da pripreme svoje sustave, arhitekturu i proračun. Zatim je prikazano kako se implementacijom i primjenom *Big Data* može postići veća konkurentnost na tržištu. Opisani su primjeri koji pokazuju kako je primjena Big Data aplikacija poboljšala poslovanje unutar i izvan tvrtke, povećala prihode, te smanjila troškove.

Jedna od najvećih društvenih mreža Facebook isto je uvidjela prednosti primjene *Big Data*. Dan je uvid na koji način Facebook iskorištava ovaj fenomen u svoju korist.

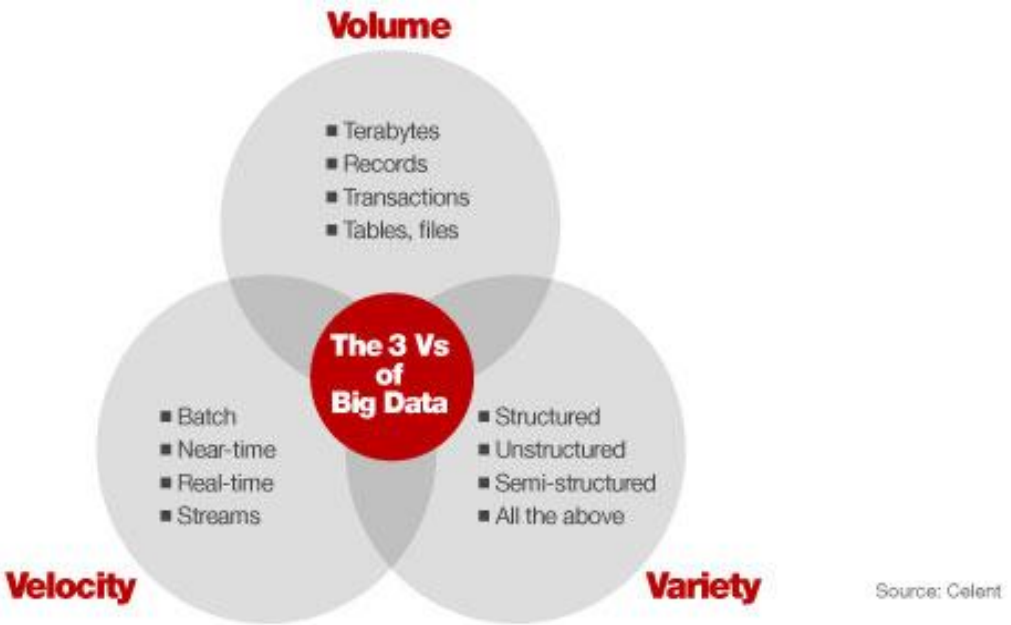
Pred kraj dan je uvid u sigurnost *Big Data*, pod time se smatra sigurnost podataka i korisnika tih podataka, te su prikazani i rizici koji su s njima povezani.

Za kraj je prikazana povezanost *Big Data* i umjetne inteligencije, gdje je opisan poznati sustav umjetne inteligencije Watson koji koristi veliku količinu podataka u kratkom vremenu.

2. Definicija pojma „Big Data“

Sam pojam „*Big Data*“ odnosi se na veliku količinu podataka. To su strukturirani i nestrukturirani podaci skupljeni tijekom vremena koje je teško obraditi pomoću tradicionalnih alata za obradu podataka i standardnog statističkog softvera. Količine tih podataka su velike, brzo rastu ili se ne uklapaju u strukturu baze podataka i nije ih moguće obraditi standardnim načinima obrade podataka. [Dumbill, 2012., str.3][1]

Taj problem javlja se u današnjim tvrtkama koje svakodnevno stvaraju sve veći broj podataka koji su u svojoj najosnovnijoj formi, polustrukturirani ili nestrukturirani. Pretpostavlja se da pohranjeni podaci sadrže informacije koje bi mogle biti korisne za poslovanje tvrtke, pa one pohranjuju sve veći broj podataka. Međutim, javlja se novi problem, s povećanjem količine pohranjenih podataka smanjuje se postotak podataka koje tvrtke mogu obraditi. [Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.3][2] Za bolje razumijevanje pojma *Big Data* potrebno je opisati njegove karakteristike, odnosno 3 dimenzije i to volumen, brzina i raznovrsnost podataka (Slika 1 i Slika 2).



Slika 1. 3 V model

[1] Dumbill, 2012., str.3

[2] Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.3

Slika 2. Prikaz 3V modela

Kako je već rečeno, tvrtke pohranjuju sve više podataka koje je teško obraditi u prihvatljivom i zahtijevanom vremenskom roku i zato je potrebno prilagoditi ili razviti nove tehnologije.[3]

Prema istraživanju iz 2001. godine od ***META Group*** (sada ***Gartner***), analitičar Doug Laney definirao je veliku količinu podataka kao trodimenzionalnu, odnosno opisao njihov volumen (količine podataka), brzinu (brzine podataka unutar i izvan) i raznovrsnost (raspon tipova podataka i izvora). Gartner, a i veliki dio industrije, i dalje koristiti taj model "3Vs" za opisivanje *Big data*. U 2012. godini, Gartner je ažurirao svoju definiciju, tako da sada glasi: "*Big Data* je veliki volumen, velika brzina, i/ili velika raznovrsnost podataka, koja zahtijeva nove oblike obrade da bi se omogućilo bolje donošenje odluka, uvid u otkrića i optimizaciju procesa“. Osim toga, pojedine organizacije su dodale novu dimenziju u opisu tih podataka. Dodale su dimenziju vjerodostojnost tih novonastalih podataka, kako bi bolje opisale značenje velike količine podataka.

[3] Big Dana i nova tehnoologija URL:http://en.wikipedia.org/wiki/Big_data

2.1. Karakteristike

Big Data mogu se opisati prema sljedećim karakteristikama[4]:

- Volume - Volumen
- Velocity - Brzina
- Variety - Različitost
- Variability - Varijabilnost
- Veracity - Vjerodostojnost

2.2. Volumen

Količina podataka koja se svakodnevno pojavljuje neprestano raste. Nekada se ta količina mjerila u megabytima, dok danas to seže i do petabyta, a očekuje se da će do 2020. godine ta količina porasti i do mjere zetabyt i to čak do 35 zetabyta. [Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.5][5] Najviše je količina podataka porasla u razdoblju od 2010. do 2012. godine i to čak za 90%.

Tekstualnih podatak ima nekoliko kilobyta, pjesma ima nekoliko megabyta, cijeli fim ima nekoliko gigabyta. [Soubra, 2012.][6] Primjeri koji dokazuju kako volumen podataka raste su Twitter koji svakodnevno stvara više od 7 terabyta podataka, ili elektronička pošta koje se po minuti pošalje 200 miliona. [Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.5][7] Facebook koji ima preko milijardu korisnika koji po minuti stvaraju 650.000 različitih sadržaja ili podijele oko 35 000 Likeova, uz to preko 600 miliona korisnika pristupa toj društvenoj mreži preko mobilnog uređaja, te se time dnevno stvara preko 10 terabyta podataka. [Klein, Tran – Gia, Hartmann, 2013.][8] Na YouTube se svake minute postavlja 72 sata video sadržaja. [Soubra, 2012.][9] Avioni godišnje proizvode oko 2,5 biliona terabyta podataka pomoću senzora koji su instalirani u njihove motore. [van Rijmenam, 2013.][10]

[4] Podijela karakteristika BD. URL: http://en.wikipedia.org/wiki/Big_data
[5] Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.5
[6] Soubra, 2012.
[7] Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.5
[8] Klein, Tran – Gia, Hartmann, 2013.
[9] Soubra, 2012.
[10] van Rijmenam, 2013.

Svakog dana, svakog sata se količina podataka povećava i time njihov volumen. Spremaju se i skladište različite vrste podataka. Neki od tih su financijski podaci, medicinski, pravni, bankarski, privatni i individualni i drugi. Jedan od izvora produkcije podataka je i osobno računalo kojeg skoro svako kućanstvo ima i na kojem se pohranjuje velika količina privatnih podataka. [Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.6][11]

Zbog znatnog povećanja volumena podataka, tvrtke imaju problema kako sa procesuiranjem, analizom, obradom i razumijevanjem tih podataka. U prošlosti su podaci u tvrtkama bili generirani i analizirani od strane zaposlenika te tvrtke, dok u sadašnje vrijeme to rade i zaposlenici, partneri i klijenti tvrtke. Za velike tvrtke i kompanije ili grupe kompanija podaci se analiziraju metodama dubinske analize podataka. [Soubra, 2012.][12] Dostupnost podataka raste, a mogućnost procesuirana opada, te zbog toga velika količina podataka ostane neistražena. Tvrtke ne znaju koji se podaci nalaze u tom neistraženom dijelu, a podaci su im bitni kako bi mogli bolje razumjeti svoje poslovanje, klijente i donositi odluke koje su bitne za tvrtku. Baze podataka su se povećale do razine petabyta. Tvrtke moraju izvagati koju vrijednost podaci imaju i da li su ti podaci doista vrijedni troškova i resursa koji su potrebni za njihovu obradu. [Klein, Tran – Gia, Hartmann, 2013.][13]

[11] Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.6
[12] Soubra, 2012.
[13] Klein, Tran – Gia, Hartmann, 2013.

2.3. Brzina

Brzina kao jedna od karakteristika koja obilježavaju *Big Data*, može biti promatrana s dva aspekta. S jedne strane brzina se odnosi na veliku brzinu najstajanja podataka u kratkom roku, a s druge strane brzina se odnosi i na samu brzinu obrade sakupljenih podataka. [Klein, Tran – Gia, Hartmann, 2013.][14] Prije nastanka velike količine podataka bilo je uobičajeno da se podaci obrađuju u jednom koraku, te da se ti ažurirani podaci dobiju svaku večer ili čak svaki tjedan. Prije se uzimao i samo jedan dio dolaznih podatka, koji je bio poslan serveru na obradu i čekao se rezultat obrađivanja. Serverima i računalima trebala je znatna količina vremena za obrađivanje podataka i za ažuriranje baze podataka. [van Rijmenam, 2013.][15] Takav način rada bio je nekad moguć jer je brzina dolaznih podataka bila sporija od vremena potrebnog za skupnu obradu podataka i zbog toga što su rezultati koji su nastali bili korisni bez obzira ako je njihova isporuka kasnila. [Soubra, 2012.][16]

U današnje vrijeme kada je sveprisutna velika količina podataka i brzina dolaznih podataka, brzina njihove obrade je jako bitna. Uz to podaci danas imaju jako kratak rok trajanja, stoga bi poduzeća trebale dolazne podatke obrađivati u skoro realnom vremenu. Brzom obradom podataka tvrtke bi mogle dobiti dodatne informacije koje su im potrebne za donošenje odluka. Važnost leži u povratnoj informaciji, te se razmatra i obrađivanje podataka i tijekom njihova prikupljanja. Dva su glavna razloga zašto se razmatra da se podaci obrađuju neposretno prilikom stvaranja i prikupljanja. Jedan je taj da baza podataka mora imati prostor za pohranu ostalih podataka. Kod velikog i brzog dolaska podataka nije moguće sve podatke pohraniti i kako bi se ostavilo prostora u skladištu za pohranu, bilo bi potrebno odmah obrađivati podatke kako oni dolaze. Drugi razlog je zbog ukazane potrebe za trenutnom povratnom informacijom, što je uobičajenija pojava kod raznih mobilnih aplikacija i online igrica. [Dumbill, siječanj 2012.][17]

Za efektivno i učinkovito korištenje velike količine podataka potrebno je obrađivati podatke prilikom stvaranja, a ne tek onda kada se oni pohrane i kada su u mirovanju. [Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.9][18]

[14] Klein, Tran – Gia, Hartmann, 2013.
[15] van Rijmenam, 2013.
[16] Soubra, 2012.
[17] Dumbill, siječanj 2012.
[18] Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.9

Za brzu obradu tih podataka tvrtke trebaju novu tehnologiju, kako bi dobili što kvalitetnije i najbolje informacije koje su korisne za samu tvrtku. [Wang, John, 2014., str.316][19]

[19] Wang, John, 2014., str.316

2.4. Različitost

Različitost predstavlja raznovrsnost svih dolaznih podataka, koji mogu biti tekstualni podaci, video podaci, slike, multimedijski podaci, podaci sa raznih internet portala, MP3 i slično. Svi ti podaci mogu biti strukturirani, polustrukturirani ili čak nestrukturirani i zbog te raznovrsnosti podataka i različite strukturiranosti podataka, tradicionalni sistemi baza podataka imaju poteškoća u obradi samih podataka.

Kod relacijskih baza podataka podaci se spremaju pomoću relacija u tako zvane tablice gdje svaki dio tablice odgovara određenom skupu podataka. Kako bi takvo spremanje i obrađivanje podataka bilo moguće, podaci trebaju biti strukturirani. Jedan od primjera strukturiranih podataka su podaci o kupcima ili proizvođačima jer kod tih podataka znamo kojem tipu podataka oni pripadaju. U te podatke spadaju ime, prezime, datum rođenja, te znamo da ime i prezime spadaju u tekstualne podatke, datum spada u brojčane podatke, te su time ti podaci strukturirani i lakše ih je obrađivati i dobiti potrebne informacije iz njih.

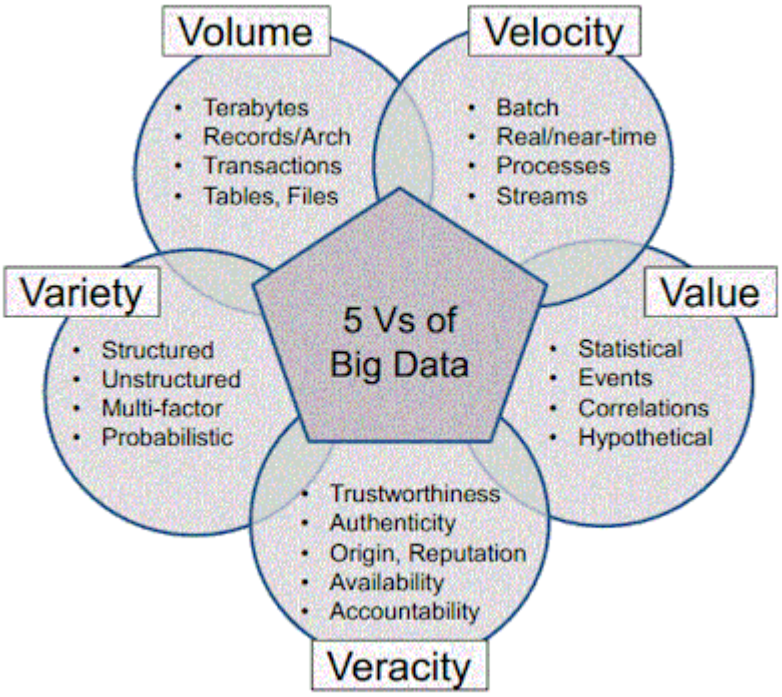
U polustrukturirane podatke spadaju oni podaci kod kojih je samo jedan dio podatka strukturiran, dok je drugi nestrukturiran. Primjer polustrukturiranih podataka je elektronička pošta (eng. *E-mail*). Strukturirani dio kod elektroničke pošte je glava poruke jer pošiljatelj poruke, primatelj poruke i predmet poruke imaju određenu strukturu. Samo tijelo elektroničke pošte je nestrukturirano jer sadržaj koji on može sadržavati može biti različite strukture, kao i prilozi koji se mogu slati pomoću elektroničke pošte.

Nestrukturirani podaci mogu se podijeliti u tri kategorije i to po tome između koga se odvija komunikacija. Tako imamo nestrukturirane podatke koji nastaju iz komunikacije koja se odvija između osoba, to može biti putem socijalnih mreža, foruma, blogova. Zatim imamo podatke koji se razvijaju iz komunikacije između ljudi i usluga odnosno ljudi i strojeva, tu spada online trgovina, razni automati kao što su bankomati, automati za plaćanje parkinga, korištenje mobilnih uređaja. U treću kategoriju spadaju podaci koji nastaju iz komunikacije između samih strojeva i primjeri za to su GPS uređaji, uređaji za snimanje slike ili zvuka, senzorni uređaji, nadzorne kamere.[Klein, Tran – Gia, Hartmann, 2013.][20]

Kako bi tvrtke i poduzeća dobili potrebne informacije koje utječu na donošenje odluka i veću dobit, potrebno je da svi tipovi podataka budu analizirani i obrađeni. [Eathon, deRoos, Deutsch, Lapis i Zikopoulos, 2012., str.8]

[20] Klein, Tran – Gia, Hartmann, 2013.

Kako se sam *Big Data* razvija i proširuje na različite domene primjene, tako se definicija ovog pojma proširuje. Osnovne karakteristike ovog problema koje se mogu opisati sa karakteristikama kao što su veliki volumen, velika brzina i velika različitost podataka, proširene su sa novim karakteristikama[21]: vjerodostojnost i varijabilnost. (Slika 3)



Slika 3. Prošireni V model

Varijabilnost (eng. Variability) se odnosi na promjenjivost podataka u relativno kratkom vremenu čime se ometa učinkovitost procesa obrade i upravljanja podacima.
Vjerodostojnost (eng. Veracity) znači da kvaliteta podataka može biti raznolika. Točnost analize ovisi o vjerodostojnosti izvora podataka.

[21] Karakteristike BD. URL: http://en.wikipedia.org/wiki/Big_data

3. Poslovna inteligencija[22]

Informacijska tehnologija sve brže napreduje i kako bi postigla konkurentnost bitno je brzo donošenje odluka, a u tome joj pomaže poslovna inteligencija.

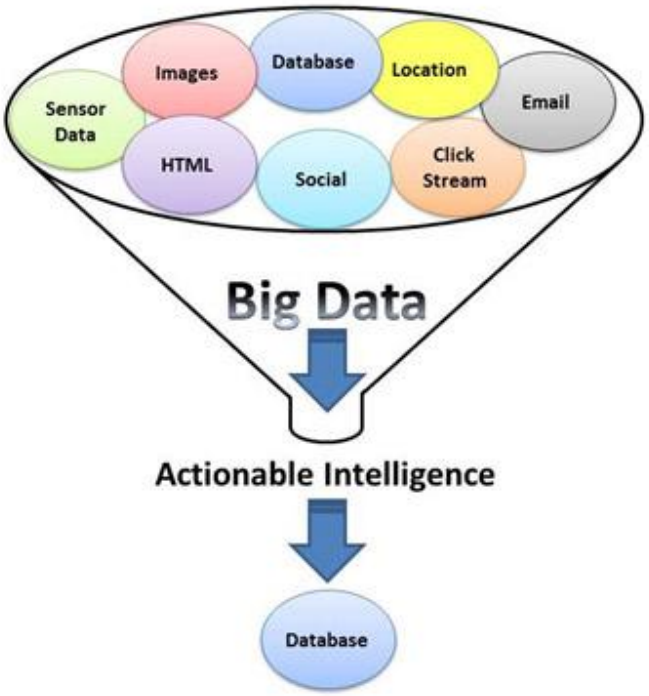
Poslovna inteligencija (eng. BI) je skup tehnika i alata za transformaciju sirovih podataka u smislene i korisne informacije koje se mogu dalje analizirati. Metode poslovne inteligencije rukuju velikom količinom nestrukturiranih podataka s ciljem da se omogući njihovo tumačenje i donošenje zaključaka koji se mogu koristiti u poslovnim procesima. Prepoznavanje novih mogućnosti i provedba učinkovite strategije može pružiti tvrtkama prednost na konkurentnim tržištima i dugoročnu stabilnost. Metode poslovne inteligencije daju uvid u prijašnje, trenutne i buduće stanje poslovanja.

Funkcije poslovne inteligencije su: izvješćivanje, on-line analitička obrada, podataka, istraživanje podataka, složena obrada događaja, upravljanje poslovnim performansama, benchmarking, analitika predviđanja i normativne analitike.

Poslovna inteligencija se može koristiti i za podršku širokog spektra poslovnih odluka u rasponu od operativnih do strateških. Osnovne poslovne odluke odnose se na pozicioniranje proizvoda na polici ili određivanje cijena. Strateške poslovne odluke uključuju prioritete, ciljeve i smjerove na najširoj razini. U svim slučajevima, poslovna inteligencija je najučinkovitija kada kombinira podatke dobivene direktno sa tržišta u kojem posluje (vanjski podaci) s onima unutar poduzeća, koji se koriste u poslovanju, kao što su financijski i operativni podaci (interni podaci). Kombinacija vanjskih i internih podataka pruža potpuniju sliku koja, stvara "inteligenciju" koja se ne može stvoriti od pojedinačnog skupa podataka.

Naglim napretkom u razvoju mobilnih uređaja, neizbježna je primjena metoda poslovne inteligencije na tabletima i pametnim telefonima (eng. smartphone). Razvijen je novi trend u poslovanju pod nazivom mobilna poslovna inteligencija (eng. MBI-Mobile Business Intelligence) što je zapravo primjena poslovne inteligencije na mobilnim uređajima. Jedan od primjera je softverski alata za mobilnu poslovnu inteligenciju tvrtke MicroStrategy u svrhu kreiranja kontrolnih ploča (eng. Dashboard). Isto tako paralelno sa fenomenom *Big Data*, razvijaju se i novi trendovi u poslovanju koji se koriste u poslovnoj inteligenciji kao što je računarstvo u „oblaku“ i razvoj socijalnih mreža. (Slika 4)

[22] Poslovna inteligencija. URL: http://en.wikipedia.org/wiki/Business_intelligence



Slika 4.Poslovna inteligencija

4. Nestrukturirani podaci[23]

Nestrukturirani podaci su podaci koji nemaju određenu strukturu, te ih nije moguće pohraniti u tradicionalnom obliku. Dok su strukturirani podaci pohranjeni u bazi podataka po tipu podatka.

Nestrukturirani podaci često uključuju tekstualne i multimedijske sadržaje. Primjeri tih podataka su poruke e-pošte, video, fotografije, audio datoteke, prezentacije, web stranice i mnoge druge vrste poslovnih dokumenata. Ove vrste datoteka mogu imati unutarnju strukturu, ali oni se i dalje smatraju nestrukturiranima jer podaci koje sadrže ne mogu se uvrstiti u bazu podataka.

Stručnjaci procjenjuju, da je 80 do 90% podataka, bilo koje organizacije nestrukturirano. Ta količina nestrukturiranih podataka u organizacijama značajno raste i to često mnogo puta brže u odnosu na strukturirane podatke.

4.1. Istraživanje nestrukturiranih podataka

Mnoge organizacije vjeruju da njihovi nestrukturirani podaci sadrže informacije koje bi im mogle pomoći u donošenju boljih poslovnih odluka. Nestrukturirani podaci teško se analiziraju. Da bi se taj problem riješio, organizacije su počele koristiti različita softverska rješenja koja su namijenjena za pretraživanje nestrukturiranih podataka kako bi izdvojili korisne informacije iz tih podataka. Glavna prednost tih alata je sposobnost u prikupljanju bitnih informacije koje mogu pomoći u poboljšanju poslovnog uspjeha.

Kako volumen nestrukturiranih podataka vrlo brzo raste, mnoge organizacije počele su primjenjivati hardverska ili softverska rješenja kako bi što kvalitetnije upravljale i pohranile nestrukturirane podatke.

[23] Nestrukturirani podaci. URL: http://www.webopedia.com/TERM/U/unstructured_data.html

4.2. Nestrukturirani podaci i Big Data

Strukturirani podaci pohranjeni su u bazi podataka po tipu podatka koji sadrže, te se te baze nazivaju relacijske baze podataka i ti podaci su relacijski podaci. Tako su primjerice podaci o telefonskim brojevima klijenata, poštanski brojevi mjesta i brojevi kreditnih kartica brojčani podaci koji imaju istu formu i mogu se upisati u relacijsku tablicu. Nasuprot njima, između nestrukturiranih podataka ne mogu se jednostavno definirati relacijski odnosi i ne mogu se pohraniti u naprijed definirane modele podataka.

Osim strukturiranih i nestrukturiranih podataka, tu je i treća kategorija i to polustrukturirani podaci. Polustrukturirani podaci su podaci koje se ne nalaze u relacijskoj bazi podataka, ali imaju neke elemente strukturiranih podataka koji olakšavaju njihovu analizu. Kao primjeri polustrukturiranih podataka mogu se smatrati XML dokumenti i NoSQL baze podataka.

Pojam *Big Data* usko je povezan s nestrukturiranim podacima. Mnogi od analitičkih alata za analizu *Big Data* mogu obrađivati nestrukturirane podatke.

4.3. Upravljanja nestrukturiranim podacima

Organizacije koriste različite softverske alate za organiziranje i upravljanje nestrukturiranim podatcima. Neki od alata su:

- ***Big Data alati***: softver kao što je ***Hadoop*** koji može obraditi nestrukturirane i strukturirane podatke, koji su izuzetno veliki, vrlo složeni i brzo se mijenjaju.
- ***Softver za poslovu inteligenciju*** *(eng. Business intelligence software)*: poznat kao BI. Obuhvaća alate za istraživanje podataka, kontrolu i izvještavanje, a sve u svrhu obrade strukturiranih i nestrukturiranih podataka, za donošenjem što kvalitetnijih poslovnih odluka
- ***Alati za integraciju podataka:*** ovi alati kombiniraju podatke iz različitih izvora, tako da oni mogu biti pregledani ili analizirani samo od jedne aplikacije. Ponekad posjeduju sposobnost da povežu strukturirane i nestrukturirane podatke
- ***Sustavi za upravljanje dokumentima:*** također se nazivaju i "*poslovni sustavi za upravljanje sadržajem*". Sustavi za upravljanje dokumentima mogu pratiti, pohranjivati i dijeliti nestrukturirane podatke koji su spremljeni u datoteke u obliku dokumenta

- ***Rješenja za upravljanje informacijama***: ovaj tip softvera prati strukturirane i nestrukturirane poslovne podatke tijekom njihovog vijeka trajanja
- ***Alati za traženje i indeksiranje:*** ovi alati izdvajaju određene podatke iz nestrukturiranih datoteka podataka kao što su dokumenti, web stranice i fotografije

4.4. Tehnologija nestrukturiranih podataka

Grupa pod nazivom Organizacija za promicanje standarda strukturiranih informacija (Oasis) (eng. *Organization for the Advancement of Structured Information Standards)*, objavila je publikaciju *Unstructured Information Management Architecture (UIMA) standard*. UIMA definira neovisnu platformu podatkovne reprezentacije i sučelja za softverske komponente ili usluge pod nazivom *Analytics,* koja analizira nestrukturirane podatke i dodjeljuje značenje nestrukturiranim podacima.

Mnoge organizacije složile su se da je **Hadoop** postao standard za upravljanje sa velikom količinom podataka. Hadoop je open source projekt kojim upravlja Apache Software Foundation[24].

[24] Open source URL: http://www.apache.org/

5. Big Data Analitika[25]

Big Data analitika se odnosi na proces prikupljanja, organiziranja i analizu velikih količina podataka kako bi se definirali obrasci i pronašle korisne informacije. *Big Data* analitika pomože organizacijama da bolje razumiju informacije sadržane u podacima, te time utvrde koji su podaci najvažniji za poslovanje i buduće poslovne odluke. *Big Data* analitičari uglavnom traže informacije koje dobivaju iz analize podataka.

5.1. Izazovi Big Data analitike

Analiza velike količine podataka organizacijama predstavlja izazov. Veliki volumen i heterogenost podataka koji se prikupe u organizaciji mogu se kombinirati, uspoređivati i analizirati kako bi se iz njih izvukle korisne informacije.

Prvi izazov je pristup podacima koje neka organizacija pohranjuje i čuva na različitim mjestima, ali često i u različitim sustavima. Drugi *Big Data* izazov je u stvaranju platforme na koju će se moći pohranjivati nestrukturirani podaci, na isti način na koji se pohranjuju strukturirani podaci.

5.2. Pregled analitike Big Data[26]

Nitko nije siguran s kojom količinom podataka raspolaže, ali bivši predsjednik Google CEO Eric Schmidt tvrdio je da stvaramo podatke veličine čitave ljudske povijesti svaka dva dana. *"Bilo je 5 exabyta podataka stvorenih od početka civilizacije do 2003. godine"*, rekao je Schmidt prije par godina, *"ali da se tolika količina podataka danas stvara svaka dva dana je zastrašujuća, a tempo porasta stvaranja podataka je u porastu."*

RJMetrics predsjednik Robert J. Moore izjavio je u TEDx da je *"23 exabyta podataka zabilježeno u 2002. Mi sada snimamo i prenosimo toliko informacija svakih sedam dana."*

Gartner smatra da će veličine podataka rasti 65% u sljedećih pet godina, dok IDC tvrdi da će se svjetski podaci duplicirati svakih godinu i pol. IDC navodi da smo u 2011. godini izradili 1,8 zettabyta (ili 1.8 triliona GBs) podataka, što je dovoljno podataka da se popune 57,5 biliona - 32GB Apple iPada.

[25] Analitika BD. URL: http://www.webopedia.com/TERM/B/big_data_analytics.html
[26] Pregled analitike BD. URL: http://www.datamation.com/applications/big-data-analytics-overview.html

Tempo stvaranja podataka sigurno će se povećavati, pogotovo zato što je komunikacija među uređajima postala jeftinija i sve učestalija

EMC Corporation (NYSE: EMC) objavila je 2011. godine rezultate istraživanja koje je sponzorirao *IDC Digital Universe studij*, *"Vađenje vrijednost iz kaosa***"**. Tim je istraživanjem utvrđeno da se svjetska količina podataka udvostručila i da se to događa svake dvije godine. Ova studija je mjerenjem i predviđanjem količine stvorenih i kopiranih podataka na godišnjoj razini, analizirala posljedice za pojedince, poduzeća i IT profesionalce.

Neke od činjenica utvrđenih u studiji:

Volumen podataka od 1,8 zettabyta može se usporediti činjenicama:

- svaka osoba u Sjedinjenim Američkim Državama tweet-a - 3 tweeta u minuti za 26.976 godina će to biti non stop,
- svaka osoba na svijetu ima više od 215 milijuna visoke razlučivosti MRI skeniranja po danu,
- više od 200 milijardi HD filmova (svaki u trajanju od 2 sata) – što znači da jednoj osobi treba 47 milijuna godina da pogleda sve filmove i to pod uvjetom da gleda filmove 24/7,
- količina podataka potrebna za popuniti 57,5 - 32GB Apple iPada. Sa tim brojem iPada mogli bismo:
    - napraviti **iPad zid**, dužine 6.000 kilometara, 20 metara visok i koji bi se protezao od Aljaske-Anchoragea do Floride-Miami.
    - izgraditi **Veliki iPad kineski zid** – koji bi bio dvostruko viši o izvornog,
    - izgraditi 10 metara visok zid oko Južne Amerike,
    - pokriti 86% od Mexico City-a,
    - izgraditi planinu 25 puta veću od planine Fuji u Japanu.

Rezultati studije pokazali su i:

- vještine, iskustvo i sredstva za upravljanje ogromnom količinom podataka i resursa su u raskoraku sa svim područjima rasta. Tijekom sljedećeg desetljeća (do 2020. godine), IT odjeli u svijetu će doživjeti:
    - 10 puta veći broj poslužitelja (virtualnih i fizičkih),
    - 50 puta veću količina podataka kojom se upravlja,

    - 75 puta veći broj datoteka koji šalju podatke u digitalni svemir, a koje rastu čak i brže od samog podatka u sve više i više ugrađenih sustava, kao što su senzori u odjeći, u mostovima ili medicinski uređaji,
    - 1,5 puta veći broj IT stručnjaka koji će trebati sa svim tim podacima upravljati.
- Veću količinu Cloud (oblak) računanja troškova i operativne učinkovitost. Dok cloud (oblak) računanje danas čini manje od 2% IT potrošnje, IDC procjenjuje da će do 2015. godine gotovo 20% podataka biti u doticaju s cloud (oblak) računanjem usluga, te da će biti pohranjeni ili obrađeni u cloudu (oblaku). Možda će se čak 10% zadržavati u cloudu (oblaku).
- da *Digitalni Shadow* ima vlastiti um, što znači da količina podataka koje pojedinac stvara kao što su osobni dokumenti, fotografije, skidanje glazbe i slično je daleko manje od količine podataka koje se stvaraju o njima u digitalnom svemiru.

Zaključak je da *"kaotični volumen podataka i dalje raste i donosi beskrajnu količinu prilika za upravljanjem transformacijskim promjenama, bilo da su one društvene, tehnološke, znanstvene ili gospodarske,"* kazao je Jeremy Burton, voditelj marketinga, EMC Corporation. Zaključak Burtona je da *"Big Data prisiljava promjene u načinu poslovanja, promjene u upravljanju i promjene u pronalaženju najbitnije stavke koji nosi, a to je* ***informacija****."*

Ostali ključni nalazi studije:

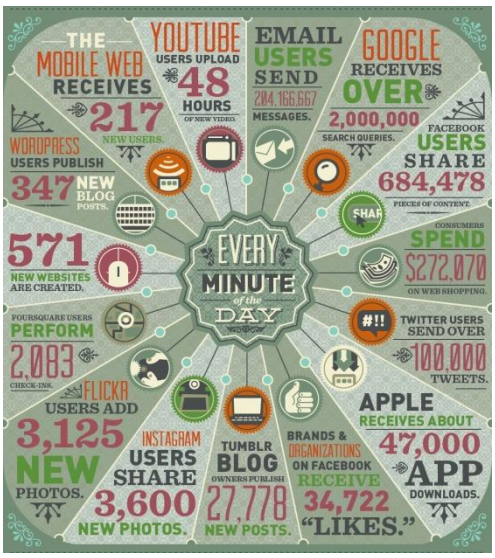
- Novo snimanje, pretraživanje, otkrivanje i alati za analizu mogu pomoći organizacijama da steknu uvide u svoje nestrukturirane podatke, što čini više od 90% digitalnog svemira. Ovi alati mogu stvoriti podatke o podacima automatski, baš kao što je i rutinsko prepoznavanje lica, koje pomaže označavanju Facebook fotografija. Podaci o podacima ili metapodacima, rastu dvostruko brže od digitalnog svemira kao cjeline.
- Alati za poslovnu inteligenciju rade u realnom vremenu.
- Novi alati za upravljanje pohranom podataka su sposobni smanjiti troškove dijela gdje spremamo podatke, kao što su de-dupliciranje, auto-rangiranje i virtualizaciju, a time nam pomažu pri odluci što se pohranjuje, a i daje nam rješenja za upravljanje sadržajem.

- Novi sigurnosni postupci i alati mogu pomoći tvrtkama identificirati podatke koje treba osigurati i na kojoj razini sigurnosti, a onda odrediti određene načine zaštite od prijetnji softverskim sustavima upravljanja, a sve u cilju sprječavanja prijevara i u cilju zaštite ugleda.
- Cloud računanje rješenja može biti i javno i privatno. Oni pružaju tvrtkama agilnost i fleksibilnost, u usporedbi s tradicionalnim IT okruženjima. Dugoročno gledano, to će biti ključni alat za rješavanje složenosti digitalnog svemira.
- Cloud računanje je omogućavanje potrošnje IT kao Servisa. A on će sa fenomenom *Big Data* motivirati organizacije da IT konzumiraju kao vanjski servis nasuprot unutarnjem ulaganju u infrastrukturu.
- Gigabajt pohranjenih podataka može generirati petabajte ili više prolaznih podataka koje obično ne pohranjujemo (npr. digitalne TV signale gledamo, ali ne snimamo, glasovne digitalne pozive koji su digitalno zapisani u jezgrama mreže za vrijeme trajanja poziva).
- Za manje od trećine podataka može se reći da ima barem minimalnu sigurnost ili zaštitu, samo oko pola podataka koje trebamo zaštititi su zaštićeni.

Prema IBM, svaki dan stvoramo 2,5 quintillion bajtova podataka. IBM tvrdi da eksponencijalni rast podataka znači da je 90% od podataka koji danas postoje u svijetu se stvorio u posljednje dvije godine.

Svake minute svakoga dana stvaramo (Slika 5):

- više od 204 milijuna e-mail poruka,
- više od 2 milijuna Google upita za pretraživanje,
- 48 sata novih YouTube videa,
- 684.000 bitova sadržaja dijelimo na Facebooku,
- više od 100.000 tweets,
- na e-trgovinu se troši $ 272.000,
- na Instagramu se dijeli 3.600 slika,
- gotovo 350 novih članova na WordPress blogu.



Slika 5. Prikaz stvaranja količine podataka

5.3. Big Data zahtijeva visoke analitičke performanse[27]

Analiza velikih količina podataka obično se izvodi pomoću specijaliziranih programskih alata i aplikacija. U aplikacije spadaju aplikacije za analitiku, pronalaženje podataka, dubinske analize teksta, predviđanja i optimizaciju podataka. Koristeći ove alate i aplikacije omogućava se obrada iznimno velike količine podataka, te se time olakšava i raspodjela podataka na relevantne i korisne, te se njihovom analizom dalje poboljšava donošenje kvalitetnijih odluka u poslovanju.

5.4. Primjeri korištenja Big Data analitike danas

Kako se tehnologija za pristup podacima, ali i njihova analiza poboljšava, tako se korištenja analitike u poslovanju, za samo poslovanje može transformirati na sve moguće načine. Prema *Datamation*, današnji napredak u analiziranju velike količine podataka omogućuje istraživačima da se dekodiranje ljudskog DNK obavi u samo nekoliko minuta. Pomoću analitike podataka moguće je predvidjeti gdje teroristi planiraju napasti, moguće je za određene bolesti sa velikom vjerojatnoćom odrediti koji gen je odgovoran za to i moguće je odrediti koji oglasi vam se najviše sviđaju na Facebooku.

[27] Analitika BD. URL: http://www.webopedia.com/TERM/B/big_data_analytics.html

Drugi primjer dolazi iz jedne od najvećih mobilnih usluga u svijetu. Francuska tvrtka Orange objavila je svoje podatke za projekt razvoja zdravlja i sigurnosti i to podatke svojih pretplatnika u Obali Bjelokosti. Podaci su korišteni od 2,5 milijarde evidencija pretplatnika, koji uključuju podatke o pozivima i SMS porukama razmijenjenih između 5 milijuna korisnika i ti podaci su anonimni. Nakon što su istraživači obradili podatke slali su ih u Orange kako bi ti podaci poslužili kao temelj za razvoj projekata za poboljšanje javnog zdravlja i sigurnosti. Na temelju tog projekta su predloženi i drugi projekti. Jedan od uključenih projekata je i onaj koji je pokazao kako poboljšati javnu i osobnu sigurnost za praćenje podataka s mobitela na karti, prateći gdje su ljudi otišli tj. gdje se nalaze nakon nekog hitnog slučaja ili događaja, drugi projekti su pokazali kako koristiti mobilne podatke za praćenje širenja bolesti.

Također mnoge tvrtke koriste *Big Data* analitike kako bi olakšali donošenje poslovnih odluka koje bi utjecale na povećanje prodaje, povećate učinkovitosti i poboljšate poslovanja, usluge kupcima i upravljanje rizicima.

Na primjer tvrtka Webopedia, ispitala je oko 540 poslovnih managera koji su trebali iskoristiti *Big Data* analitiku za poboljšanje poslovanja. Oko polovice svih ispitanika izjavilo je da su primjenom *Big Data* analitike poboljšali odnos sa klijentima i da im je to pomoglo u razvoju proizvoda i stjecanju određene konkurentske prednosti.

Dodatno 62% ispitanika izjavilo je da su koristili *Big Data* analitiku za povećanje brzine i za smanjivanje složenosti poslovanja.

6. Uvriježene predodžbe o Big Data[28]

Mnoge tvrtke su već provele implementaciju *Big Data* aplikacije. Te aplikacije koriste velike količine podataka, pa se sastoje od hibridnog hardvera i softvera za pohranu i pristup podacima i od sofisticiranog softverskog sučelja koje prihvaća upite poslovnih analitičara, pristup pohrani podataka, te daje odgovore koji se koriste za razumijevanje potrebe kupaca, pojednostavljenje poslovanja i povećanje profitabilnosti.

Kao što su se priče o uspjehu i neuspjehu pojavile u vijestima i tehničkim publikacijama, pojavilo se i nekoliko uvriježenih predodžbi vezanih za *Big Data koje nisu sasvim utemeljene.* U ovom poglavlju prikazane su neke od tih predodžbi povezanih s implementacijom *Big Data* u poslovanju.

6.1. Big Data aplikacije ne mogu raditi samostalno

Pretpostavka da *Big Dana* aplikacije ne mogu raditi samostalno je pogrešna. Analiziranje poslovnih podataka je uobičajeno, osobito u tvrtkama koje već imaju skladišta podataka. Pohrana podataka sadrži vremenski ovisne snimke operativnih podataka, a trenutni statusi podataka i analitičkih izvješća ovise o dimenzijama u skladištu.

Na primjer preferencije kupaca su entiteti prema kojima analitičari određuju i kategoriziraju podatke. U prodaji oni uključuju vrijeme, položaj, tip kupaca, trgovine, odjele. Upit kupaca koji se odnosi na kupnju elektroničkih predmeta u maloprodaji u nekoliko država tijekom božićnih blagdana uključuje dimenziju vrste proizvoda (elektronički uređaji), trgovine, zemljopis (država) i vrijeme (božićnih blagdana). Svaka dimenzija daje drugačiji način prikazivanja podataka, a može dati tragove koji se odnose na želju kupaca, dostupnost proizvoda u trgovinama ili profitabilnosti.

Big Data aplikacije zahtijevaju takve dimenzije. Kada su ti podaci pohranjeni i održavaju se u skladištu podataka, prirodno je da se integriraju u modele podatke dotičnog skladišta i *Big Data* aplikacije.

Prirodni ishod takvih integracija je da će se vršiti nadogradnja skladišta podataka, tako da se analitičkim upitima može obuhvatiti podatke u skladištima. Dobri modeli podataka poduzeća i sveobuhvatni rječnici podataka su vrlo važni i nužni.

[28] Mitovi o BD. URL: http://www.databasejournal.com/features/db2/exploding-the-myths-of-big-data.html

Nadogradnja skladišta uključuje dodavanje novih dimenzije, prikupljanje podataka iz novih operativnih sustava i skladištenje velikih objekata, kao što su skenirane slike i XML. Veliki, složeni predmeti ne smiju biti izravno analizirani od programskog paketa poslovne inteligencije, ali osnovni podaci mogu biti pohranjeni u skladištu podataka. Na primjer, XML dokumenti mogu se dekodirati nekim sustavima za upravljanje bazama podataka i pohranjivati se u bazi podataka kao tablica. Ova tablica podataka može se zatim analizirati pomoću BI softvera.

6.2. Osim Big Data hardvera i softvera nove metode analize nisu potrebne

Pretpostavka da za veliku količinu podataka nove metode analize nisu potrebne je pogrešna. Svaka implementacija *Big Data* od bilo kojeg IT poduzeća će izazvati značajne troškove pored ulaganja u *Big Data* hardver i softvera.

Big Data podacima se mora moći odrediti veličina. To se odnosi na sposobnost sustava da reagira na veće količine podataka, veće brzine prijenosa podataka i sve veći broj korisnika podataka. Početni simptomi ovog problema će usporiti vrijeme odziva. Dugi rad zahtijeva vrijeme i produžuje vrijeme transakcije.

Za mnoge aplikacije ta pitanja će se percipirati kao kapacitet povezivanja i odgovor će biti dodavanje više procesora, više memorije, a i diskova većeg kapaciteta pohrane. Međutim, u *Big Data* okruženju takav pristup nije dovoljan. Većina hibridnih hardvera i softvera za *Big Data* koje dobavljači osiguravaju, ovisi o vlasničkim metodama za pohranu podataka, uključujući i kompresije podataka, masivne paralelne obrade i koordinaciju sa sustavom za upravljanje bazom podataka (DBMS). Skaliranje u ovom okruženju zahtijeva da se ponovno razmišlja na koji način su podaci projektirani i pohranjeni, uključujući i moguću de-normalizaciju podataka, logičko particioniranje, inteligentnije upite za ponovno zapisivanje i dodatno analiziranje SQL upita.

Neki *Big Data* sustavi sadrže složene tipove podataka kao što su podaci u Extensible Markup Language (XML), audio i video podaci, skenirane slike. Aplikacije *Big Data* trebale bi analizirati ove vrste podataka, radi sakupljanja i drugih operacija.

Za provedbu toga moraju biti uključene odgovarajuće metode i alati.

Drugi problem korištenih metoda je što se u razvoju aplikacija i testiranju najčešće ne koriste testovi na *Big Data* podacima.

Vrlo je bitno da se sustav redovito održava dodavanjem novih izvora podataka i pohranjivanja, čišćenjem i arhiviranjem nevažnih podataka, praćenjem performansi i planiranjem kapaciteta.

6.3. Big Data aplikacije ne zahtijevaju podešavanje performansi

Pretpostavka da velika količina podataka nema potrebu za podešavanjem performansi je pogrešna. Očekuje se da *Big Data* aplikacije trebaju kratko vrijeme za pristup podacima i da imaju sposobnost da brzo i jednostavno izvrše analizu velike količine podataka.

Ključ za performanse *Big Data* ovisi o samim podacima. IT sustavi i dalje moraju pribavljati podatke iz operativnih sustava, transformirati ih, i učitati ih u svoje *Big Date* aplikacije. Što se više podataka zahtijeva, to treba puno više rada na sustavima potpore da bi se osigurali ažurirani podaci.

Prikupljanje i ažuriranje podataka iz operativnih sustava kao što je kopiranje datoteka i ekstrahiranje baza podataka, brisanje nepotrebnih podataka i slično, zahtjeva vrijeme, pa zbog takvih procesa trebaju se ugoditi performanse *Big Data* aplikacije.

Također i kako se količina ulaznih podataka koji se dnevno učitavaju stalno povećava, povećava se i vrijeme učitavanja, tako da treba predvidjeti vrijeme i za to.

Važna karika u određivanju performansi *Big Data* aplikacije su broj upita za pristup *Big Data* podacima i skladištu podataka, odnosno broj korisnika.

Većina skladišta podataka koristi DBMS-ov alat za optimizaciju koji mjeri trošak pristupa podacima pa se takav pristup koristi i u *Big Data* aplikacijama.

7. Priprema poduzeće za Big Data[29]

Primjena *Biga Data* je sada prilično uobičajena u velikim organizacijama. Primjena počinje kao dio projekta informatizacije poduzeća koji će izdvojiti, pohraniti i analizirati velike količine postojećih podataka kako bi se smanjili troškovi, predviđanja kupaca i kako bi se ubrzalo vrijeme postavljanja proizvoda na tržište i kako bi se predvidjeli zahtjevi za određenim proizvodima i kapacitetima proizvodnje.

Teško je *Big Data* aplikacije uključiti u postojeću IT infrastrukturu. Osim zahtjeva za energijom i za hlađenjem novih hardvera, koji bi podržali novu primjenu *Big Data*, potrebno je i druga IT područja pripremiti. Potreba za pohranom podataka, veći kapacitet prijenosa podataka i zahtjevi koji će se postaviti na postojeći hardver i softver glavni su čimbenici koji određuju koja dodatna oprema će biti potrebna za postojeće aplikacije.

7.1. Primjer sustava koji podržava Big Data

Jedno od programskih rješenja koje omogućuje korištenje *Big Data* je IBM DB2 Analytics Accelerator (IDAA), hibrid hardvera i softvera iz IBM-a. Taj hardver uključuje višestruka terabajt diskovna polja za pohranu podataka, kao i high-speed mrežu kabela za prijenos podataka. Nakon što su podaci pohranjeni može im se pristupiti kao da je baza podataka.

Sustav za upravljanje bazom podataka (DB2 u ovom slučaju) će upravljati upitima za podacima. Kada se pristupa podacima koji se pohranjuju lokalno, upiti će se izvoditi vrlo brzo pa je to prednost koja navodi određenu organizaciju na odluku o kupnji dodatnog diskovnog prostora. Analiza velikih količina podataka zahtijeva puno vremena. Analitički upiti u standardnim bazama podataka mogu trajati satima, a isti ti upiti mogu se izvesti u samo nekoliko minuta ili sekundi ako se pristupa podacima koji se nalaze na memorijskim diskovima.

Ova kompleksna i hardverska rješenja zahtijevaju mnogo električne energije. Potrebno je preispitati može li postojeća električna mreža podnijeti taj dodatni teret s obzirom na snagu električne energije.

[29] Poduzeća i BD. URL: http://www.databasejournal.com/features/db2/preparing-your-enterprise-for-big-data.html

7.2. Zahtjevi arhitekture

Dodatno treba preispitati i slijedeće:

- Hoće hardver pohraniti samo proizvedene podatke? Ako je tako, znači li to da se analitički upiti moraju testirati na proizvedenim podacima? Ako ne, kako će takvi upiti biti ispitani i koji podaci?
- Ako *Big Data* rješenja jesu (ili će postati) kritična za svoju organizaciju, da li će se morati uzeti u obzir da se instalira hardver u Disaster Recovery (oporavku od katastrofe) okruženju? Ako je tako, kako će zadržati veliku količinu podataka na Disaster Recoveryu i sinhronizirati ih sa trenutnim podatacima?
- Ako je odluka da se može pohraniti samo jedan dio trenutno proizvedenih podataka, koji kriteriji će se koristiti za učitavanje podataka u hardver? Drugim riječima, ako podaci u hardveru omogućuju velike brzine upita, koji podaci će biti tamo pohranjeni?

Ostale arhitekture uvjetuju potrebu za obradom i prijenosom velikih količina podataka. Podaci moraju biti izvađeni iz izvornih sustava, provjereni, transformirani, a potom učitani u baze podataka i na memorijske diskove. Veće količine protoka podataka će dovesti do sljedećeg:

- Ekspanzije mreže, uključujući i moguće dodatne paralelne kanale podataka za prijenos podataka;
- Novih i većih medija za pohranu, najčešće u obliku niza diskova, kako za spremanje osnovnih podataka, taka i za stvaranje sigurnosnih kopija (backups);
- Nadogradnje za administraciju baza podataka i automatizirane procese, kao što su sigurnosne kopije baze podataka, povratak, preustrojavanje, održavanje indeksa i slično;
- Potrebe za dodatnim osobljem za upravljanje i nadzor.

7.3. Nadogradnja arhitekture skladišta podataka

Na početku analize velike količine podataka potrebno je pregledati i nadograditi trenutno okruženje skladišta podataka. Poslovni analitički upiti koriste se za analizu velike količine podataka, koji obično zahtijevaju raščlambu podataka po kategoriji ili dimenziji.

Svako analitičko rješenje velike količine podataka zahtijeva integraciju sa skladištem podataka, a time i integracije po sljedećim pitanjima:

- Podaci. Koliko vremena će trebati za učitavanje novih podataka? Hoće li uzastopna ažuriranja podataka biti sinhronizirana ili će biti potrebno učitavanje tijekom noćnih ciklusa?
- Stara arhiva podataka. Koliko često se stari ili neiskorišteni podaci brišu ili arhiviraju? Kako će to utjecati na podatke na diskovima? Na koji način će se izvršiti brisanje velike količine podataka diskova?
- Oporavak od katastrofe. Kako će se izvršiti sigurnosna kopija podataka na diskovima? Da li ima dovoljno prostora na drugim medijima za pohranu kako bi se zadržali svi podaci? U slučaju katastrofe, koliko dugo će trajati oporavak svih podataka?
- Performanse i rast. Ako sve veći broj korisnika postavlja sve više zahtjeva i upita, hoće li hardver to uspješno odrađivati? Kako će se pratiti takvi zahtjevi i da li postoje mogućnosti za ugađanjem performansi hardvera?

7.4. Plan za integraciju

Priprema tvrtke za veliku količinu podataka je zahtijevan plan. Jedna od najčešćih metoda je da se započne s najvećim pitanjem integracije i to skladištem podataka. Treba razmisliti o tome kako podaci prolaze kroz skladište podataka. Jedna od uobičajenih načina preslikavanja tih protoka je ekstrahiranje, transformiranje i opterećenje. Ekstrahiranje se odnosi na dotok podataka iz izvornih sustava, transformiranje uključuje sve izmjene, ispravke i sažetak podataka, a opterećenje uključuje učitavanje podataka u skladište. Treba razmisliti o tome kako će rješenja za veliku količinu podataka utjecati na svaku točku.

7.5. Izvorni sustavi

U izvorne sustave spadaju temeljni operativni sustavi. Ti sustavi procesuiraju narudžbe, računovodstvo, izvršavanje transakcija s kupcima, naplatu i slično. Ti sustavi produciraju transakcijske podatke, od kojih su neki izvađeni i poslani u skladište podataka. *Big Data* rješenja mogu uključivati dodatne informacije na ovim izvadcima ili mogu uključivati potpuno nove izvore podataka.

IT organizacije mogu odlučiti da se implementacija *Big Data* vrši na vađenju

podataka izravno iz produkcijskih sustava za neposrednu pohranu i analizu. Međutim, postoji nekoliko nedostataka za takvu ideju, tu su podaci koji još uvijek nisu pretvoreni, jako puno elemenata podataka mogu biti nepotpuni ili čak mogu i nedostajati. Osim toga, podaci još nisu pohranjeni u skladište podataka, tako da može biti otežano usklađivanje *Big Data* i pohrane podatake. I kao posljednje, osoblje koje podržava proizvodne programe ne mora biti stručno za pomaganje u analitici.

Ako se treba provesti implementacija *Big Data* na ovom mjestu, potrebno je uključiti osoblje za potporu produkcijskih sustava od samog početka. Njih treba uključiti u diskusijama i dizajnerskim odlukama, uključujući i to kako će se podaci uspoređivati i spajati i unutar trenutnih podataka. Treba inzistirati na dokumentiranju svih polja i na promjenama u rječniku podataka.

7.6. Kretanje podataka i transformacija

Neki elementi internih podataka mogu imati pogrešne ili nevažeće stavke. Na primjer podaci koji sadrže sve nule u datumskom polju i koji se neće moći obuhvatiti u analitičkim upitima koji zahtjevaju pretraživanje po mjesecu ili godini. Ostali problemi uključuju izgubljene podatke ili podatke koji zahtijevaju verifikaciju prema drugom sustavu.

Kod vanjskih podataka koji su sakupljeni od dobavljača, sa web stranica i slično česti problem je nepotpunost podataka, nebrojčani znakovi u numeričkim poljima, te slobodna tekstualna polja koja se moraju raščlaniti da bi se od njih dobili podaci (kao što adresna polja sadrže adresu, grad, državu i poštanski broj). Ova polja s podacima zahtijevaju transformacijsku logiku za rješavanjem problema podataka i dodjeljivanje zadanih vrijednosti.

Mnoge transformacije su standardizirane kao na primjer u slučaju ispravanog datuma, polje se može postaviti na vrijednost 2099/12/31.

7.7. Učitavanje podataka u skladište

Učitavanje podataka u skladište podataka je mjesto na kojem se većina *Big Data* implementira. IT stručnjaci definiraju hardver kojim se mogu izvršiti i skladišni procesi tako da se osim učitavanja podataka sa diskova čitaju i podaci iz skladišta baze podataka.

Središnja točka integracije rješenja *Big Data* u postojeći IT prostor je koordinacija

između *Big Data* i skladišta kojom se treba definirati hoće li i u kojem slučaju podaci u skladištu baze podataka biti dostupni.

Big Data rješenja uvijek zahtijevaju usporedbu s veličinom prikupljenih podataka i postojećim dimenzijama skladišta podataka, da li u zemljopisnom području ili vremenskom periodu tako da je bliska koordinacija između analitičara skladišta podataka i *Biga Data* presudan faktor za uspjeh.

7.8. Planiranje proračuna za Big Data[30]

Big Data softver, hardver i analiza poslovnih rješenja su sastavni dio implementiranja *Big Data* sustava. Poduzeća su preplavljena sa ponudama trgovaca za rješavanjem *Big Data* problema za koja nisu niti znali da postoji.

Konkurenti skupljaju podatke da bi definirali potrebe kupaca, da bi odredili nove kategorije proizvoda i da bi povećali profit. Uobičajeni programi uključuju skladištenje i analizu prodanih podataka kupcima, web interakcijama, očitavanja strojnih senzora, i još mnogo toga.

7.8.1 Scaling up - povećanje opsega

Scaling up – povećanje opsega, je izraz koji se koristi za opisivanje sposobnosti sustava da reagira na veće količine podataka, brži prijenos podataka i sve veći broj upita korisnika. *Skaling up* problemi obično se manifestiraju kao produženo vrijeme transakcije. Pristupanje podacima zahtijeva duže vrijeme i korisnici percipiraju da je reagiranje na upit loše. Najčešći odgovor korisnika na to je da je to pitanje resursa i kapaciteta. IT timovi moraju dodavati više hardverske memorije, više bržih procesora, veće i brže diskove i kalibrirati high-speed mreže.

U *Big Data* okruženju IT tehničke ekipe su osigurale velike kapacitete, high-speed pohranu podataka i to obično u obliku hibridnog hardver/softver uređaja kojeg se može iznajmiti. Zašto bi se u tim okolnostima netko brinuo o scaling up-u?

[30] Planiranje proračuna. URL: http://www.databasejournal.com/features/db2/it-budget-planning-for-big-data.html

Razlog tome je način na koji *Big Data* koristi scaling up tijekom vremena. Najčešće se koriste za pohranu velike količine vremenski ovisnih podataka, kao što su višemjesečne transakcije klijenata. Ti podaci se zatim analiziraju pomoću sofisticiranih softvera za poslovne inteligencijske analize (eng.*business intelligence analytics*), koje uključuju kombinaciju naprednih analitičkih softvera i high-speed rezultata obrade u upitima.

Kako podaci i analitika postaju sve vrjedniji, više korisnika postavljaju veće i složenije upite. Brzo vrijeme odaziva dovodi do više upita nego što je uobičajeno. Ad-hoc upiti postaju redovita mjesečna izvješća, potom tjedna, a onda svakodnevna. Količina rada povećava se eksponencijalno. Odjednom, se dolazi do granice *Big Data* aplikacije.

Zbog toga IT mora imati proračun za eventualni *scaling up*. Uspješan projekt će, dovesti do zahtjeva za većim kapacitetom i resursima, tako da treba predvidjeti kako brzo će se to dogoditi i te procjene za proširenjem treba uključujući u proračun.

7.8.2. Scaling Out

Proračun za *scaling out* se smatra proračunom za osoblje. Potrebno je razumjeti podatke kroz razne arhitekture, a to podrazumijeva standardiziranu dokumentaciju, razvoj najbolje prakse i potencijalnu integraciju podataka.

Big Data nisu samo velika, komprimirana inačica jednostavnih poslovnih podataka. Većina operativnih sustava današnjice sadrže nove i složene vrste podataka, kao što su veliki objekti (Large objects - LOBs) koji sadrže audio zapise, slike, videa i skenirane dokumente. Mnogi sustavi pohrane podataka sadrže vlastite opise na XML jeziku (Extensible Markup Language). *Big Data* znači ponovno korištenje trenutnih modela podataka i integraciju podataka s različitim vrstama.

7.8.4. Obučavanje osoblja

Big Data aplikacije dolaze s novim tehnologijama, a time i uvjetima za nova područja stručnosti osoblja. Neki od tih uvjeta su:

- Podrška za instalaciju, konfiguraciju, tuning i nadogradnju bilo koje posebne namjene hardvera ili softvera, kao što su *Big Data* primjene.

- Podrška za poslovnu inteligencijsku analitiku, uključujući i upite za ugađanjem performansi, ispravljanje pogrešaka, a možda i iskustvo u novom analitičkom softveru.
- Poznavanje nekoliko srodnih operativnih sustava, a posebno onih iz kojih će se izvor podataka povući u *Big Data* aplikaciju.

Proračun za osoblje ne uključuje samo trening za trenutno osoblje, nego možda i stjecanje novih zaposlenika ili savjetodavnih usluga. Rani zahtjevi uključuju specifičnosti implementacije *Big Data* hardvera, softvera i aplikacije. Ubrzo nakon toga, trebat će osoblje za organizaciju podrške za potporu analitike korisnika, kao i osoblje za obavljanje nadzora performansi, tuninga i planiranje kapaciteta.

7.8.5. Proračun za operativne sustave

Treba voditi računa i o aktualnim operativnim aplikacijama jer su ti sustavi primarni izvor naših *Big Data*. Podaci izvađeni iz operativnih sustava mogu biti nepotpuni i netočni. Arhitekti baza podataka znaju da su oni već pripremljeni i učitani procesima za preuzimanje operativnih podataka i da im je uklonjena ili "očišćena" vrijednost (eng.value).

Isti ili slični postupci moraju se implementirati za popularizaciju *Big Data,* da bi se to ostvarilo, moramo razumjeti naše podatke. Većina osoblja operativnih sustava će biti upoznata s *Big Data* ili analitikom istih, dok eksperti koji koriste *Big Data* neće znati postojeće sustave.

7.8.6. Čišćenje/za arhivu

U jednom trenutku će podaci u *Big Data* postati stari ili neupotrebljivi zbog svoje starosti. Čak i osobna high-speed pohrana podatke može biti preopterećena sa previše podataka. Usklađenost ili regulatorni zahtjevi za sigurnost podataka primjenjuju se na operativne sustave i eventualno skladište podataka, ali ne i za *Big Data*.

U proračun treba uzeti u obzir i čišćenje ili procese za arhiviranje. To će uzeti vremena za razvoj i vremena za izvršenje čišćenja koje može biti značajno jer mnogi dobavljači nude rješenja za brzo učitavanje podataka i brze upite, ali ne i za brzo uklanjanje podataka.

7.8.7. Oporavak od katastrofe

Okruženje za oporavak od katastrofe omogućuje pokretanje operativnih sustava, kada jedna ili više primarnih stranica nisu dostupne. Mnoge IT aplikacije ne uključuju *Big Data* kod planiranja oporavka od katastrofe jer se prioritet daje operativnim sustavima, a pokrenuti *Big Data* upiti se ne smatraju kritičnima.

Kada IT implementira *Big Data*, postoji početno razdoblje treninga i niskog korištenja. U nekom trenutku, dovoljan broj internih korisnika je toliko ovisan o njihovoj analitici da dobiju etiketu- kritičan. Odjednom, IT menadžment mora osigurati rješenje za oporavak od katastrofe.

Proračun je varljiv proces, balansira trenutno i dostupno osoblje, novac, te povrat investicije. Za jedinstven proračun, potrebno je za *Big Data* implementaciju, zajednički obrazac: novi procesi, nove pohrane podataka, nove analitike. Uvođenje novina znači promjenu, ali isto tako dio IT osoblja može biti otporno na promjene.

8. Zašto su Big Data nova konkurentska prednost?[31]

Mnogi vjeruju da su *Big Data* novost kojom će neke tvrtke pokušati iskoristiti druge tvrtke da postanu najbolje u svojoj branši. Naravno postoje i skeptici, ali mnogi koji vjeruju u mogućnosti i iskoristivost *Big Data*, će iskoristiti alate za što kvalitetniju uporabu *Big data* za poboljšanje poslovnih odluka, a time i da budu ispred konkurencije.

Podaci su sada utkani u svaki sektor i funkciju globalne ekonomije, naravno i kod drugih bitnih faktora proizvodnje, kao što su imovine i ljudski kapital. Mnoge moderne ekonomske aktivnosti jednostavno se ne mogu održati bez njih. Korištenje *Big Data* postat će temelj konkurencije i rasta za pojedine tvrtke, povećavajući produktivnost i stvarajući značajnu vrijednost za svjetsko gospodarstvo smanjivanjem otpada i povećanjem kvalitete proizvoda i usluga.

Do sada velika količina podataka preplavljuje naš svijet. Istraživanje iz McKinsey Global Institute (MGI) i McKinsey & Company Business Technology ureda, prikazuju da do sada ukupna količina generiranih, pohranjenih i istraživanih podataka za nalaženje iskoristivih informacija, će znakovito djelovati na ekonomske odnose između poslovnih subjekata, vlade i potrošača.

Povijest prijašnjih trendova u IT investicijama i inovacijama, te njihov utjecaj na konkurentnost i produktivnost snažno ukazuju na to da *Big Data* može imati sličnu moć na tržištu. Pretpostavke koje su dopustile prethodne valove IT, omogućuju inovacije za povećanje produktivnosti, tj, tehnološke inovacije praćene usvajanjem komplementarnih inovacija upravljanja na mjestu za *Big Data*.

Sve tvrtke moraju uzeti vrlo ozbiljno *Big Data* i njihov potencijal za stvaranje vrijednosti, ako se žele i dalje natjecati o ostati konkurentni. Neki trgovci obuhvaćajući *Big Data* podatke, vide potencijal za povećanjem operativne marže od 60%.

8.1. Big Data: Nova konkurentska prednost

Korištenje *Big Data* postaje ključan način u vodećim tvrtkama da nadjačaju svoje kolege u drugim tvrtkama. U većini industrija, osnovana je konkurencija i novi će sudionici podjednako utjecati na podatke, vođeni strategijama za inovacijama, natjecanjem, ali i stvaranjem vrijednosti.

[31] BD i konkurentnost. URL: http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/

U zdravstvu, data pioniri, su analizirali zdravstvene ishode lijekova, koji su bili široko propisivani i nastojali su otkriti prednosti i rizike koji nisu bili vidljivi tijekom ograničenih kliničkih ispitivanja. Ostali prvi korisnici *Big Data*, su na temelju podataka iz senzora ugrađenih u proizvode od dječjih igračaka, pokušali utvrditi koliko se ti proizvodi zapravo koriste u stvarnom svijetu. Takva saznanja pomažu u stvaranju novih ponuda usluga i dizajna budućih proizvoda.

Big Data pomaže u stvaranju novih prilika za rast i posve nove kategorije tvrtki, poput onih koje skupljaju i analizirati podatke industrije. Mnoge od njih biti će tvrtke koje se nalaze u sredini velikih informacijskih tokova, u kojima podaci o proizvodima i uslugama, kupcima i dobavljačima, sklonostima potrošača i namjerama, se mogu dohvatiti i potom analizirati. Lideri sa naprednim idejama, u pojedinim sektorima će početi agresivno graditi svoje vlastite organizacije *Big Data* podataka.

U realnom vremenu i sa visokom frekvencijom, priroda podataka je vrlo važna. Na primjer, sposobnost procijene podataka kao što je povjerenje potrošača, i to odmah, nešto što se prije moglo učiniti uglavnom naknadno, sada postaje mogućnost da se intenzivno koristi, dodajući znatnu moć predviđanja.

8.2. Pet načina iskoristivosti Big Data

1) *Big Data* mogu stvoriti transparentnu vrijednost informacija. Još uvijek postoji vrlo značajna količina informacija koja još uvijek nije pohranjena u digitalnom obliku, npr, podaci koji su na papiru, ili podaci koji su vrlo nepristupačni za pretraživanje putem mreže. Došlo se do saznanja da se 25% truda, u samom znanju članova radne skupine, sastoji u traženju podataka, a potom u prebacivanju na drugu lokaciju (ponekad virtualnu). Ovaj trud predstavlja značajan izvor neučinkovitosti.
2) Organizacije mogu stvoriti i pohraniti više transakcijskih podataka u digitalnom obliku, mogu prikupiti više točnih i detaljnih informacija o učinku na sve, od zaliha proizvoda do bolovanja i time stvarati varijabilnosti i povećati učinkovitost. Određene vodeće tvrtke koriste svoje sposobnosti za prikupljanje i analizu *Big Data*, da bi potom provodile kontrolirane eksperimente u svrhu poboljšanja u donošenju kvalitetnijih i boljih odluka u upravljanju.
3) *Big Data* omogućuju sve užu segmentaciju kupaca, a time i puno kvalitetnije i prilagođene proizvode ili usluge.

4) Sofisticirana analitika može znatno poboljšati donošenje odluka, smanjiti rizike i ukazati na vrijednosti koje bi inače ostale skrivene.
5) *Big Data* se mogu iskoristiti za razvoj nove generacije proizvoda i usluga. Na primjer, proizvođači na temelju podataka dobivenih od senzora ugrađenih u proizvode, stvaraju inovativne usluge nakon prodaje, a time pridonose proaktivnom održavanju, kako bi izbjegli kvarove na novim proizvodima.

8.3. Stvorene vrijednosti upotrebom Big Data

Ako bi, američki zdravstveni sustav, koristio *Big Data* kreativno i učinkovito, za poboljšanje učinkovitosti i kvalitete mogao bi stvoriti vrijednost višu od 300 milijardi $ svake godine. U razvijenim gospodarstvima Europe, vladini dužnosnici mogu stvoriti više od € 100 milijardi (123bn $) u operativnim poboljšanjima učinkovitosti samo pomoću *Big Data*, a to ne uključuje korištenje naprednih analitičkih alata za smanjenje prijevara i pogrešaka.

Nisu samo tvrtke i organizacije te koje bi mogle profitirati od onoga što *Big Data* može stvoriti. Potrošači također mogu imati vrlo značajne koristi. Na primjer, korisnici usluga koje omogućuju određivanje lokacije, mogu ostvariti 600 milijardi $ potrošačkog suficita.

Korištenje pametnog usmjeravanja (smart routing) je jedna od najviše korištenih aplikacija pomoću informacija u realnom vremenu. Kako penetracija pametnih telefona raste, a time i aplikacije za navigaciju (free navigation) koje su uključene u istima, očekuje se da će rasti i korištenje aplikacija za navigaciju. Do 2020. godine, očekuje se da će više od 70% mobitela imati GPS mogućnosti, u usporedbi sa 20 posto u 2010. godini. Procijenjeno je da će do 2020. godine potencijal globalne vrijednosti pametnog usmjeravanja (smart routing) u obliku vremena i uštede goriva biti oko 500 milijardi $. To je ekvivalent za uštedu vozačima od 20 miijardi sati na cesti, ili 10 do 15 sati svake godine za svakog putnika, te oko 150 milijardi $ na potrošnju goriva.

Neki od najznačajnijih potencijala za generiranje vrijednosti iz *Big Data* nalazi se u kombiniranju odvojenih baza podataka. Američki zdravstveni sektor, na primjer, ima četiri glavne baze podataka – *klinička*, *aktivnost (potraživanja) i troškovi*, *farmaceutskih i medicinskih proizvoda* i *podaci o ponašanju i raspoloženju pacijenta* od kojih je svaka pohranjena i s kojom se upravlja u različitim institucijama.

U medicinskom sektoru *Big Data* tehnike omogućuju analiziranje podataka o stvarnim vremenima medicinskih tretmana, njihovim troškovima i zdravstvenim ishodima, te je time moguće usmjeravati liječnike na one tretmane koji daju najbolje rezultate i koji su najisplativiji. *MGI* procjenjuje da će se ako američki zdravstveni sektor bude u potpunosti koristio te raspoložive tehnike koje *Big Data* nudi godišnja produktivnost povećati za dodatnih 0,7%. No, za ostvarivanje tog poticaja produktivnosti zahtijevati će se kombinacija podataka iz različitih izvora i to često od organizacija koje nemaju povijest dijeljenja podataka. Skup podataka, kao što su zapisi o pacijentima i klinički zahtjevi morali bi biti integrirani.

Na taj način bi se ostvarila korist ne samo za razne industrije koje prate zdravstveni sektor, već i za pacijente, koji će imati širi i jasniji pristup većim različitim zdravstvenim podacima, što ih činiti više informiranima. Pacijenti bi mogli usporediti ne samo cijene lijekova, liječenja i liječnika, nego i njihovu relativnu učinkovitost, te tako im omogućiti da izaberu najučinkovitije i ciljane lijekove, potencijalno čak i prilagođene njihovom osobnom genetskom i molekularnom make-up-u. Za dobivanje tih širokih prednosti, korisnici zdravstvenih usluga morali bi prihvatiti nešto drugačiji kompromis između njihove privatnosti i prednosti koje će donijeti veće uvezivanje podataka.

Osjetljivost oko privatnosti i sigurnost podataka su samo jedna od prepreka koje tvrtke i vlade moraju prevladati, ako žele imati gospodarske koristi od iskoristivosti *Big Data*. Jedan od najvažnijih izazova je značajan nedostatak ljudi s vještinama za analizu *Big Data*. Do 2018. godine, Sjedinjene Američke Države mogla bi se suočiti s nedostatkom od 140 do 190 tisuća ljudi s velikim analitičkim treninzima (u statistici ili strojnog učenja) i još oko 1,5 milijuna ljudi s menadžerskim i kvantitativnim sposobnostima da bi mogli učinkovito uokviriti i interpretirati analize podataka.

Tu su i mnogi tehnološki problemi koje treba riješiti za stvaranje *Big Data*. Postojeći sustavi i nespojivi standardi i formati, vrlo često spriječavaju integraciju podataka i primjenu više sofisticirane analitike koje stvaraju vrijednost. U konačnici, korištenje velikih digitalnih skupova podataka zahtijeva skup tehnologija preuzimanja podataka iz pohrane i računanja kroz analitike i vizualizacije softverskih aplikacija.

Prije svega, dostupnost podacima treba proširiti. Tvrtke će morati pristupati podacima trećih osoba, primjerice, poslovnih partnera ili klijenata, i integrirati ih u svoje. Vrlo važna kompetencija, za data-vođene organizacije, u budućnosti će biti mogućnost stvaranja snažnih vrijednosti propozicija za druge, uključujući i potrošače, dobavljače i potencijalno čak i konkurente, da dijele podatke.

Sve dok tvrtke i vlade razumiju moć i snagu *Big Data*, za povećanje produktivnosti, bolje vrijednosti za potrošače i kao val rasta u globalnom gospodarstvu, trebalo bi to biti dovoljno jak poticaj za njih da djeluju snažno da prevladaju prepreke za njegovo korištenje. Na taj način će se osloboditi putevi za nove konkurentnosti među tvrtkama, veću učinkovitost u javnom sektoru, koja će omogućiti bolje usluge, čak i u suzdržanim fiskalnim vremenima, i omogućiti tvrtkama, pa čak i cijelim ekonomijama da budu više produktivniji.

8.4. Big Data su velika stvar

Doba *Big Data* bi mogla donijeti nova načela upravljanja. U ranim danima profesionalizacije korporativnog upravljanja, čelnici su otkrili da minimalna učinkovita ljestvica je ključ određivanja natjecateljskog uspjeha. Isto tako, buduće konkurentske prednosti, su uvjeriti tvrtke da ne pohrane što više i što kvalitetnije podatke, već i da ih kvalitetno upotrijebe. Da bi se to stvarno i reflektiralo na određene interese, te razmišljajući o pet pitanja koja slijede, biti će lakše prepoznati kako pravilno korištenje velike količine podataka ima stažan utjecaj.

8.4.1. Što se događa u svijetu radikalne transparentnosti, s vrlo dostupnim podacima?

Kako podaci postanu lako dostupni svim sektorima, to može ugroziti tvrtke koje su se oslonile na vlasničke podatke kao natjecateljsku imovinu. Primjer toga može biti industrijska nekretnina i trgovina informacijskih asimetrija, kao što su povlašteni pristup transakcijama i čvrstom poznavanju ponude i ispitivanju ponašanja kupaca. U posljednjih nekoliko godina on-line stručnjaci, vlasnici podataka i analitike počeli su zaobilaziti agente, dozvoljavajući kupcima i prodavačima razmjenu stajališta o vrijednosti nekretnina i stvaranje paralelnih izvora podataka.

Jedan od velikih izazova je i činjenica, da se ogromne količine podataka mnogih tvrtki često gomilaju i skrivaju u odjelima, koji se bavi istima, kao što su istraživanje i razvoj, inženjering, proizvodnja, usluga ili servis, i da ne dopuštaju njihovu trenutnu eksploataciju. Razmjena informacija unutar poslovnih jedinica također mogu biti problem. Mnoge financijske institucije pate od vlastitog neuspjeha za razmjenom podataka između različitih područja poslovanja, kao što su financijska tržišta, upravljanje novcem i kreditiranjima. Često to sprečava te tvrtke za stvaranjem koherentnih pogleda pojedinačnih kupaca ili razumijevanje veza između financijskih tržišta.

Neki proizvođači pokušavaju natjerati otvaranje tih odjela. U naprednim proizvodnim sektorima, kao što je primjerice automobilska industrija, dobavljači iz cijelog svijeta stvaraju tisuće komponenti. Više integriranih platformi podataka sada omogućuje tvrtkama i njihovim partnerima u lancu nabave da surađuju tijekom faze dizajna, te je to ključna odrednica konačnih troškova proizvodnje.

8.4.2. Da ste mogli provjeriti sve svoje odluke, kako bi to promijenilo način na koji se natječete na tržištu?

Big Data donosi mogućnosti korištenja bitno različitih vrsta odluka. Koristeći kontrolirane eksperimente, tvrtke mogu testirati hipoteze i analizirati rezultate za usmjeravanje investicijskih odluka i poslovnih promjena. Eksperimentiranje može pomoći menadžerima da razlikuju uzročnost od obične povezanosti, čime se smanjuje varijabilnost rezultata, a poboljšavaju se financijske vrijednosti i učinkovitost proizvoda.

Opsežna eksperimentiranja mogu uzeti mnoge oblike. Vodeće online tvrtke, na primjer, su kontinuirano testeri. U nekim slučajevima, oni objavljuju dio svojih stavova na svojim web stranicama za provođenje istraživanja, te tako potiču korisnike i posjetitelje tih stranicama da se uključe u istraživanje na način da daju svoja mišljena, preporuke pa i kritike. Tako dobiveni stavovi od posjetitelja web stranica u sebi sadrže čimbenike koji se dalje koriste u poboljšanju i privlačenju više korisnika ili povećavanju prodaje. Tvrtke koje prodaju fizičku robu, koriste istraživanje odluka, ali *Big Data* može pogurati ovaj pristup na novu razinu. McDonald's, na primjer, je opremio neke svoje restorane sa uređajima koji skupljaju operativne podatke kao što su praćenje interakcije s klijentima, promet u restoranima i uzorak naručivanja.

Tamo gdje takvi kontrolirani eksperimenti nisu izvedivi, tvrtke mogu koristiti "prirodne" eksperimente za identificiranje izvora varijabilnosti u izvedbi. Jedna Vladina organizacija, na primjer, prikuplja podatke od više skupina zaposlenih koji rade slične poslove na različitim mjestima. Jednostavno stavljanje na raspolaganje tih podataka potaknulo bi poboljšanje performansi radnika koji su zaostajali u svom radu.

8.4.3. Kako bi se Vaše poslovanje promijenilo, ako bi koristili Big Data za prilagodbu u realnom vremenu?

Tvrtke okrenute kupcima odavno koriste podatke za dio kupaca i za ciljane kupce. *Big Data* dopušta velik korak izvan onoga što se donedavno smatralo da je najsuvremeniji način, stvarajući što veću moguću personalizaciju u realnom vremenu. Nova generacija trgovaca će biti u mogućnosti pratiti ponašanje pojedinih kupaca kroz strujanje Internet klikova. Time će moći ažurirati njihove želje i zahtjeve i modelirati njihovo vjerojatno ponašanje u stvarnom vremenu. Oni će tada biti u stanju prepoznati kada su kupci blizu odluke o kupnji i pogurati transakciju za željenim proizvodom do završetka, ujedno nudeći određenu nagradu kao korist kupnje. Ovo ciljanje u realnom vremenu povećat će kupnju marginalnih proizvoda od svojih najvrjednijih kupaca.

Maloprodaja je očito industrija za prilagođavanje vodećih podataka jer količina i kvaliteta raspoloživih podataka dobivenih tijekom kupovina putem Interneta, društveno-mrežnih razgovora i u novije vrijeme interakcije pomoću pametnih telefona su se povećale. No i ostali sektori, također, mogu imati koristi od novih aplikacija podataka, zajedno s rastućom sofisticiranosti analitičkih alata za podjelu kupaca u sve više otkrivenih mikrosegmenata.

8.4.4. Kako Big Data može povećati ili čak zamijeniti upravljanje?

Big Data proširuje moguće domene aplikacijskih algoritama i strojno-posredovane analize. Pojedinim proizvođačima algoritmi primjerice analiziraju senzorske podatke s proizvodnih linija, stvarajući samoregulirajuće procese za smanjivanje otpada, izbjegavajući skupe (a ponekad i opasne) ljudske intervencije. U naprednim, "digitalnim" naftnim poljima, instrumenti stalno očitavaju podatke o uvjetima na glavi bušotine, cjevovodima i mehaničkim sustavima. Te podatke analizira skup računala, koji procjenjuju sve rezultate u realnom vremenu operacijskog centra zbog prilagodbe protoka ulja, a sve u cilju optimizacije proizvodnje i smanjivanja zastoja. Jedna od vodećih naftnih tvrtki je na taj način smanjila operativne i kadrovske troškove za 10 do 25 %, dok je povećala proizvodnju za 5%

Proizvodi od fotokopirnih uređaja do mlaznih motora sada mogu generirati tok podataka kako bi se pratila njihova upotreba. Proizvođači mogu analizirati dolazne podatke i, u nekim slučajevima, automatski izvršiti popravak softverskih propusta ili poslati predstavnika servisa da izvrši popravak. Neki proizvođači računalnog hardvera sakupljaju i analiziraju skupljene podatke, da bi izvršili preventivne popravke u proizvodnji, te da bi time spriječili nezadovoljstvo i neuspjehe koji bi im mogli poremetiti poslovanje sa klijentima. Podaci se također mogu koristiti za promjenu proizvoda i time sprječavati buduće probleme ili je moguće da kupci dobivaju informacije o novim generacijama proizvoda.

Poanta je poboljšati performanse, upravljanje rizicima, te sposobnosti da se ono što bi inače ostalo skriveno. Kako cijena senzora, komunikacijskih uređaja i analitičkih softvera i dalje pada, sve više i više tvrtki će se pridruživati ovoj menadžerskoj revoluciji.

8.4.5. Može li se stvoriti novi poslovni model koji se temelji na podacima?

Big Data je i miješanje novih kategorija tvrtki koje pristupaju informacijama upravljanjem poslovnim modelima. Mnoge od tih tvrtki igraju posredničke uloge u vrijednosnim lancima gdje se pronalaze i stvaraju vrijedni "ispušni podaci" proizvedeni u poslovnim transakcijama. Jedna prijevoznička tvrtka, primjerice, prepoznala je da u tijeku svog poslovanja, vrši prikupljanje ogromnih količina podataka o globalnim isporukama pošiljaka. Osjetivši priliku, ona je stvorila jedinicu koja prodaje podatke za dopunu poslovnih i gospodarskih prognoza.

Druga globalna tvrtka naučila je jako puno kroz analitiku svojih podatke kao dio proizvodnog obrta koji je odlučio stvoriti poslovanje kako bi napraviti slično poslovanje za druge tvrtke. Sada tvrtka prikuplja od trgovina i opskrbnih centara podatke za nekoliko proizvodnih kupaca i prodaje softverske alate za poboljšanje performansi. Ovo uslužno poslovanje sada nadmašuje tvrtke za proizvodnju poslovanja.

9. Zašto su Big Data – Big deal (veliki stvar)[32]

Podaci sada kolaju u svakodnevnom životu od mobitela, kreditnih kartica, televizora i računala, od infrastrukture gradova, od senzora kojim su zgrade opremljene, vlakova, autobusa, zrakoplova, mostova i tvornica. Podaci teku tako brzo da je ukupna akumulacija u protekle dvije godine toliko velika da, podaci koji sadrže cjelokupnu evidenciju ljudske civilizacije, izgledaju iznimno maleno. "Ovdje je *Big Data* revolucija", kaže Weatherhead-ov sveučilišni profesor Gary King, ali nije količina podataka revolucionarna, "Big data revolucija je da sada možemo nešto korisno učiniti s podacima."

Revolucija je u poboljšanim statističkim i numeričkim metodama, a ne u eksponencijalnom rastu pohranjivanja podataka ili čak u računalnom kapacitetu, objašnjava King-a. Udvostručenje računalne snage svakih 18 mjeseci nije ništa u usporedbi s velikim algoritmom tj. setom pravila koja se mogu koristiti za rješavanje problema tisuću puta brže nego pomoću mogućih konvencionalnih računalnih metoda. Kolega od profesora King, suočen s tolikom količinom podataka, shvatio je da će trebati 2 milijuna $ računala da bi ih analizirao. Umjesto toga, King i njegovi diplomirani studenti su došli sa algoritmom u roku od dva sata kako bi učinili istu stvar za 20 minuta na laptopu, jednostavan primjer, ali ilustrativan.

Novi načini povezivanja skupova podataka su odigrali veliku ulogu u stvaranju novih spoznaja. Kreativan pristupu vizualizaciji podataka često dokazuje da su sastavni dio procesa stvaranja znanja jer su ljudi daleko bolji od računala u određivanju obrazaca. Mnogi od alata koji su sada razvijeni se mogu koristiti širom disciplina koje naoko djeluju posve različito kao što je astronomija i medicina. Među studentima, postoji ogroman apetit za nova područja. Harvardski tečajevi u znanosti podataka prošle jeseni privukli su 400 studenata, od prava, gospodarstva, vlade, dizajna i medicine, tako i fakulteta inženjerstva i primijenjene znanosti (SEAS), pa čak i MIT. Nastavnici su objavili da će Harvard School of Public Health (HSPH) uvesti novi magisterij iz računalne biologije i kvantitativne genetike iduće godine, vjerojatno preteča Ph.D. programa. Na fakultetu inženjerstva i primijenjene znanosti govorili su o organiziranju magisterija u znanosti podataka.

[32] BD kao velik posao. URL: http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

Između fakultetskih kolega, King izvješćuje da "pola članova odjela vlade rade neku vrstu analize podataka, usporedno sa Odsjekom za sociologiju i dobrim dijelom ekonomije, više od polovice Škola javnog zdravlja i veliki dio na Medicinskom fakultetu. "Čak i zakon je uzet od strane pokreta za empirijsko istraživanje koji je društvena znanost ", kaže on, te navodi da je teško naći prostor na koji nije bilo utjecaja.

King opisuje kako je Kevin Quinn, bivši profesor i član uprave na Harvardu, objavio natječaj uspoređujući njegov statistički model i kvalitativne procjene 87 profesora prava, da bi vidio kako može najbolje predvidjeti ishod svih predmeta Vrhovnog suda u godini. "Profesori prava znaju sudsku praksu i što je svaki od sudaca odlučio u prethodnim slučajevima, tako da su oni znali sudsku praksu i sve argumente," King podsjeća. "Quinn i njegov suradnik, Andrew Martin [izvanredni profesor političkih znanosti na Sveučilištu Washington], prikupili su šest sirovih varijabli na cijeli niz prethodnih slučajeva i napravili su analizu." Navodi da to nije natjecanje, nego kad god postoji dovoljno podataka koje se može kvantificirati, modernim statističkim metodama će se nadmašiti pojedinac ili mala grupa ljudi svaki put.

U marketingu, poznati načini korištenja *Big Data* uključuju "*preporuke*" poput onih koje se koriste u tvrtkama kao što su Netflix i Amazon. Te tvrtke na temelju prethodnih interesa i kupnji kupaca koriste preporuke za kupnju. Prema ciljevima i odabiru proizvoda tijekom kupnje, koristili su se algoritmi koji su otkrili kada su žene trudne i to praćenjem kupnje predmeta kao što bezmirisni losioni, da bi se potom ponudili posebni popusti i kuponi od tih vrijednih pokrovitelja. Tvrtke koje izdaju kreditne kartica došle su do neobičnih zaključaka tijekom istraživanja podataka za procjenu rizika neplaćanja i to da ljudi koji kupuju jastučići protiv grebanja, za njihov namještaj su vrlo odlučni za njihovo plaćanje.

U javnoj sferi, postoje sve vrste aplikacija pomoću kojih se dolazi do korisnih podataka, tako tu spadaju i raspodjela policijskih resursa za predviđanje gdje i kada će se najvjerojatnije dogoditi zločini. Zatim pronalazak povezanosti između kvalitete zraka i zdravlja ili korištenje genomske analize kako bi se ubrzao uzgoj usjeva poput riže koja je otporna na sušu. U više specijaliziranim istraživanjima, primjer toga je stvaranje alata za analizu ogromnih skupova podataka u biološkim znanostima. U takvo jedno istraživanje uključen je i izvanredni profesor organske i evolucijske biologije Pardis Sabeti koji je proučavanjem ljudskog genoma iz milijarde baznih parova, identificirao gen koji je postao vrlo prominentan tijekom ljudske evolucije za utvrđuje svojstva, kao što su sposobnost da se probavi kravlje mlijeko ili otpornost na bolesti poput malarije.

King je sam nedavno razvio alat za analizu tekstova na društvenim medijima. "Sada nastaju milijarde postova socijalnih medija svaka dva dana, što predstavlja najveći porast, u svojstvu ljudske rase, da se izrazi u bilo kojem trenutku u dosadašnjoj povijesti svijeta," kaže on. Niti jedna osoba ne zna i ne razumije sve jezike. No, statističke metode razvijene od strane Kinga i njegovih učenika, koji su testirali svoje funkcije na postovima napisanih kineskim jezikom, to bi sada omogućavala.

King je također osmislio i proveo "ono što je nazvao najveći pojedinačni eksperimentalni dizajn za procjenu socijalnog programa u svijetu, dosada (*Seguro Popular*.)", rekao je Julio Frenk, dekan HSPH. "Cijela moja karijera je vođena na temeljnim uvjerenjem da su znanstveno izvedeni dokazi najmoćniji instrument, mi moramo osmisliti prosvijetljenu politiku i proizvesti pozitivne društvene promjene," kaže Frenk, koji je bio ministar zdravstva u Meksiku. Kada je preuzeo dužnost 2000. godine, više od polovice zdravstvenih izdataka, bili su plaćeni koje je narod platio iz svog džepa svake godine, četiri milijuna obitelji su bile upropaštene glede katastrofalnih zdravstvenih troškova. Frenk je provodio zdravstvene reforme koje su implementirane i pozitivno ocijenjene, te je stvorio nove javne sheme zdravstvenog osiguranja, *Seguro Popular*. Zahtjev za procjenu programa (za koji on kaže da je projektiran da košta 1% BDP-a od dvanaestog po veličini gospodarstva u svijetu) sagrađen je na pravima. Dakle, Frenk je angažirao najadekvatniju osobu da provede procjenu i to Gary Kinga.

S obzirom na komplikacije prilikom provođenja eksperimenta *Seguro Popular*, za vrijeme trajanja programa, King je morao pronaći nove analitičke metode. Frenk ga je nazvao *"veliki akademski rad*“. *Seguro Popular* se proučavao i simulirao u desetcima zemalja diljem svijeta, zahvaljujući u velikoj mjeri na činjenici da su tome pridonijela vrlo rigorozna istraživanja *Big Data*. "King je izgradio nevjerojatno originalan dizajn", objašnjava Frenk. Zato što je King uspoređivao zajednice koje su dobivale zdravstveno osiguranje u prvoj fazi (implementacija je trajala sedam godina) sa demografski sličnim zajednicama koje nisu i rezultati su bili "vrlo jaki", kaže Frenk. Nakon samo 10 mjeseci Kingova studija pokazala je da je Seguro Popular uspješno zaštitio obitelj od katastrofalnih rashoda zbog teške bolesti, a njegov rad daje smjernice za potrebna poboljšanja, kao što je javna promidžba za uporabom preventivne skrbi.

Osobno King kaže da se mogućnosti *Big Data* mogu iskoristi u korist društva da bi napredovalo puno više od onoga što je do sada. Google je analizirao pojmove u klasterima pretraživanja po regijama u Sjedinjenim Državama za predviđanje izbijanja gripe brže nego što je bilo moguće pomoću bolničkih prijemnih zapisa.

King navodi da je to bio pilot projekt, ali ipak da je to tek mali dio od onoga što se može učiniti kada bi bilo omogućeno da akademski istraživači pristupe informacijama u posjedu tvrtke. Tvrtke sada imaju više socijalno-znanstvenih podataka od akademika. Napravljen je pomak od nedavne prošlosti, kada je to bilo obrnuto. Kada bi društveni znanstvenici mogli koristiti taj materijal, kaže on, moglo bi se riješiti sve vrste problema. King izvješćuje da se čak i u akademskim krugovima podaci ne dijele na mnogim poljima. "Postoje čak i studiji na ovom sveučilištu u kojima ne možete analizirati podatke, osim ako ne napravite originalne kolektore od podataka koautora.", kaže King.

Potencijal za učiniti korisno i dobro, je možda nigdje više, nego u javnom zdravstvu i medicini, kaže King , te navodi da *"Ljudi su doslovno umirali svaki dan"*, jednostavno zato što se podaci ne dijele.

10. Kako Facebook upravlja sa Big Data[33]

Facebook ima vrlo jasan pristup, a to je uradi sam, dizajnira vlastite poslužitelje i umrežavanje, te dizajnira i gradi svoje podatkovne centre. Njegovo osoblje piše većinu svojih aplikacija i stvara gotovo sve vlastite middleware (softver koji djeluje kao most između operacijskog sustava ili baze podataka i aplikacija, osobito na mreži). Sve svojim operativnim radom ujedinjuje u jedan iznimno velik sustav koji se koristi od strane internih i eksternih korisnika podjednako.

To je vjerojatno zato što je malo IT tvrtki moralo posluživati više od 950 milijuna registriranih korisnika. Od toga vrlo visoki postotak u realnom vremenu, svakodnevno. Mali broj njih moraju prodavati oglašavanje za oko 1 milijun klijenata ili imaju desetke novih proizvoda u svom poslu, a sve to u preciznom vremenu.

Na primjer, Facebookov-a skupina ljudskih resursa, računovodstveni ured, Mark Zuckerberg na e-mail, pa čak i mi sami preko laptopa provjeravamo svoj status. Svi se koristimo, istim divovski podatkovnim centralnim sustavom koji okružuje Zemlju u njezinoj moći i opsegu.

10.1. Sve što Facebook radi, uključuje i Big Data

"Dakle, sve što mi radimo ispada da je u vezi s *Big Data*", rekao je Jay Parikh, potpredsjednik infrastrukture inženjerstva u Facebooku. "To utječe na svaki sloj našeg pohranjivanja, pričali smo o poslužiteljima, pohrani podataka, umrežavanjima i podatkovnom centru, kao i ukupnom softveru, operacijama, vidljivosti, alatima - sve to dolazi zajedno u ovoj jednoj aplikaciji koju moramo pružiti svim našim korisnicima. ", navodi J. Parikh.

Treba shvatiti koliko *Big Data* ima utjecaja na njihovo poslovanje, navodi Parikh.

Prikupljeni podaci se trebaju iskorištavati, ako ne oni se samo gomilaju, a Facebook želi iskoristiti sve podatke koje prikupi.

Facebook ne zna uvijek što želi učiniti s popisom korisnika, Web statistikama, geografskim informacije, fotografijama, pričama, porukama, linkovima, video zapisima i svim ostalim što tvrtka prikuplja, ali želi sve prikupiti.

"Želimo znati tko posjećuje stranice, kojim aktivnostima se bave na njima, što rade na stranici.", rekao je Parikh.

[33] Facebook i BD: URL: http://www.eweek.com/c/a/Data-Storage/How-Facebook-Is-Handling-All-That-Really-Big-Data-423736

10.2. Facebook radi u *Storage-Buying Mode*

Svoj prvi, u potpunom vlasništvu, podatkovni centar, Facebook je otvorio u proljeće 2011. godine u Prineville-u, Oregon, i to nakon dvije i pol godine izgradnje. Izgrađen je za potrebe Facebooka i koristi arhitekturu tvrtke *Open Compute Project*. Ona ima dvije velike zgrade, 330.000 četvornih metara, jedna se koristi za svakodnevno poslovanje, a druga kao hladnjača.

Ako zapitate nekoga na Facebooku koliko skladištenja tvrtka radi u bilo kojem trenutku, nikada nećete dobiti izravan odgovor, jer iskreno to se nikada ne zna.

Recimo samo da Facebook nikada ne napušta tj. da se stalno nalazi u Storage-Buying Mode.

Facebook je pokrenuo OCP 7. travnja 2011. To je bez presedana pokušaj open-source specifikacije koja zapošljava, za svoje hardvere i podatkovne centre, učinkovitu snagu u socijalnoj mreži koja obuhvaća više od 950 milijuna ljudi.

U sklopu projekta, Facebook je objavio specifikacije i mehanički dizajn, koji se koristio za izgradnju matičnih ploča, napajanja, poslužiteljskih šasija, poslužitelja i baterijskih ormara, za svoj podatkovni centra. Taj presedan je dovoljan, za tvrtku Facebook, da raste na ljestvici, ali društvena mreža je također otvoren izvor za podatkovne centre električnih i mehaničkih konstrukcija.

10.3. Facebook ne radi particije podataka (Ključno pravilo pohrane)

Iznad i izvan svega, sa dobro dokumentiranom sigurnosti Facebook nastavlja bitku sa rukovanjem ogromnom količinom podataka koji dolaze u Prineville i podacima koje se rentaju od drugih podatkovnih centara.

Facebook je dosta rano uspostavio da se infrastruktura njihovih podataka dijeli u cijeloj tvrtki. Tu infrastrukturu nije lako razdijeliti. Oni su uspjeh ostvarili skaliranjem ograničenja tog sustava, zbog vlastitog rasta, te iz tog razloga žele i dalje držati sve zajedno.

Mnoge tvrtke koriste jednostavniji način razdijele i to u timove jer zajedno ne mogu to napraviti. Tako sustav dijele na sve manje dijelove i unutar samih timova na još manje dijelove. Takvo razdjeljivanje centraliziranih informacijskih sustava na manje dijelove, jednostavno dodaje veću kompleksnost, troškove i vrijeme osoblja.

Facebook ne radi na takav način, Parikh navodi da je njegov tim uvijek u potrazi za načinima kako ubrzati analizu opterećenja *Big Data*.

Velike količine financijskih sustava, sada dobivaju odgovore u vremenu od mikro ili nano-sekundi. To je vrsta konkurentske prednosti koje ograđeni fondovi dobivaju da budu u mogućnosti obrađivati velike količine podataka u realnom vremenu, rekao je Parikh.

To se odnosi i na pristup udruženih sustava. Parikhovo mišljenje je da ne treba imati nikakav otpor da netko pristupa podacima druge organizacije, ako će pomoći da većoj prodaji ili boljoj učinkovitosti.

11. Sigurnost Big Data[34]

U eksploataciji *Big Data* vrlo je bitan detalj njihova sigurnost kao i sigurnost terminalnih senzora koji služe za njihovo prikupljanje.

11.1. Big Data sa sigurnosne točke gledišta

Big Data je iznimno popularna tema o kojoj se govori, ali što je ono o čemu se stvarno raspravlja? Iz perspektive sigurnosti, tu su dva različita pitanja:

- zaštita tvrtke i njenih korisnika/kupaca u kontekstu *Big Data,*
- pomoću tehnike *Big Data* izvršiti analizu, pa čak i predvidjeti, sigurnosne incidente.

11.1.1. Osiguranje/zaštita Big Data

Mnoge tvrtke već koriste *Big Data* za marketing i istraživanje, ali ne mogu imati temeljna prava, pogotovo iz perspektive sigurnosti. Kao i sa svim novim tehnologijama, sigurnost se čini da je u zaostatku.

Big Data štete će biti vrlo velike, s potencijalom za još ozbiljniju reputacijsku štetu i na pravne posljedice.

Većina organizacija već se suočila s implementacijom ovih koncepta. Moramo identificirati vlasnika za izlazne procese *Big Data*, ali i za one neobrađene podatke.

Vrlo mali broj organizacija će vjerojatno izgraditi okruženje *Big Data* u vlastitom okruženju, tako da će Cloud (oblak) i *Big Data* biti neraskidivo povezani. Tako su i mnoge tvrtke svjesne, da pohranjivanje podataka u Cloudu (oblaku) ne uklanja njihovu odgovornost za njihovom zaštitom od regulatorne i komercijalne perspektive.

Tehnike kao što su attribute based encryption mogu biti potrebne kako bi se zaštitili osjetljivi podaci i primijenile na kontrole pristupa (kao atributi od samih podataka, a ne iz sredine u kojoj su pohranjeni). Danas su mnogi od tih koncepata nepoznati u poduzećima.

[34] Sigurnost BD. URL: http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view

11.1.2. Upotreba Big Data u svrhu sigurnosti/zaštite

Upotreba *Big Data* za otkrivanje prijevara je vrlo atraktivna za mnoge organizacije. Postoji komercijalna zamjena na raspolaganju za postojeće sustave upravljanja.

Uzimajući ideje za budućnost, izazov otkrivanja i sprječavanja trajne prijetnje, prije nego se one dogode, je moguće uz određeni način analize *Big Data*. Ove tehnike mogu imati ključnu ulogu u pomaganju otkrivanja prijetnje u ranoj fazi, pomoću više sofisticiranih analiza uzorka, a kombinirajući i analizu od više izvora podataka.

Današnje logiranje je često zanemareno, osim ako dođe do kakve greške. *Big Data* pruža priliku za konsolidaciju i analizu logiranja automatski iz više izvora, a ne samo izolirano, tako da bi se sustavi mogli poboljšavati kroz stalne prilagodbe i kroz učinkovito učenje "dobrog" i "lošeg" ponašanja.

Integracija podataka od fizičkih sigurnosnih sustava, kao što su izgradnja kontrole pristupa, pa čak i video nadzor, mogao bi značajno poboljšati zaštitu ali i socijalni inženjering se treba uzeti u obzir u procesu sigurnosti. To predstavlja mogućnost znatno naprednijeg otkrivanja prijevare i kriminalnih aktivnosti.

11.1.3. Rizici povezani s Big Data tehnologijama

Rizici koji se pojavljuju su:

- To je nova tehnologija za većinu organizacija, a bilo koja tehnologija koja se dobro ne razumije može dovesti novih problema
- Implementacija *Big Data* sadrži otvoreni izvorni kod, s potencijalom za neprepoznatim *back doors* (stražnjim vratima) i zadanim vjerodajnicama
- Napadi na čvorove u klasteru, ne moraju biti pregledani i od poslužitelja adekvatno kontrolirani i spriječeni
- Korisničko ovjeravanje autentičnosti i pristup podacima s više mjesta ne mogu biti dovoljno kontrolirani

Postoji značajna prilika za zlonamjernim unosom podataka i neodgovarajućim vrednovanjem podataka

11.2. Hakerska revolucija

Vrijeme usamljenih hakera koji su u mraku svoje sobe pokušavali kompromitirati neki računalni sustav odavno je prošlo. Takvom se aktivnošću sada bave dobro organizirane skupine nerijetko sponzorirane od nekih zemlja. Mobilne platforme su područje koje još nije osvojeno i kompromitirano. Tu još nije zabilježen veći upliv malicioznog koda, ali je s obzirom na njihovu rasprostranjenost, otvorenost koda pojedinih platformi i neupućenost krajnjih korisnika samo pitanje vremena kad će antivirusni softver na pametnom telefonu biti obvezna stavka kao što je danas za klasičnu desktop platformu. Što naravno može ugroziti *Big Data* podatke i same korisnike.

12. Umjetna inteligencija i Big Data[35]

Fanovima znanstveno-fantastične serije Zvjezdane staze sigurno je poznato brodsko računalo koje na pitanja postavljena normalnim ljudskim govorom odgovara na engleskom jeziku kao da je riječ o živoj osobi. Radnja te serije zbiva se u 24. stoljeću, a malo tko zna da već danas postoji računalo koje može na govorom postavljene upite odgovoriti na isti način.

Riječ je o IBM-ovom sustavu umjetne inteligencije pod nazivom Watson (nazvanom po IBM-ovom prvom izvršnom direktoru Thomasu J. Watsonu). Watson je razvijen pomoću različitih programskih jezika, a pogon ima na SUSE Linux Enterprise serveru uz pomoć Hadoop platforme koja mu omogućuje distribuirano upravljanje računalnim resursima. Taj sustav koristi 2.880 POWER7 procesorskih jezgri koje rade na taktu od 3,5 GHz te 16 TB RAM-a.

Zahvaljujući Hadoop platformi i iznimno snažnoj hardverskoj platformi, Watson u jednoj sekundi može obraditi količinu podataka ekvivalentnu milijunu knjiga, a vrijednost mu je procijenjena na 3 milijuna dolara. Watsonovu bazu znanja izgradili su IBM-ovi inženjeri, a sastoji se od goleme količine dokumenata, rječnika, enciklopedija (uključujući i kompletni sadržaj Wikipedije) i sličnih materijala koji sačinjavaju više od 200 milijuna stranica strukturiranog i nestrukturiranog sadržaja koji zauzima više od 10 TB diskovnog prostora.

Jasno je da dizajniranje jednog takvog sustava za običan kviz baš i nema smisla tako da je Watson zaposlen i na mnogo ozbiljnijim zadaćama. Prošle su godine različite tvrtke dizajnirale posebne aplikacije koje su mogle koristiti Watsonovu umjetnu inteligenciju (kroz Watson API koji im je IBM stavio na raspolaganje) i ponuditi odgovore na određena pitanja iz točno određenih područja poput medicine ili prava. Takve vanjske tvrtke zajedno s IBM-om sačinjavaju takozvani Watsonov ekosustav gdje IBM vanjskim suradnicima stavlja na raspolaganje Watson platformu, dok oni pronalaze potencijalnu praktičnu primjenu.

IBM je samo ove godine uložio više od milijardu dolara u daljnji razvoj te platforme koja će se razvijati u smjeru različitih cloud-based usluga. S brzinom obrade od 80 TeraFLOP-a, Watson nije ni među prvih petsto svjetskih superračunala, no ono što ga izdvaja od ostalih je algoritam to jest umjetna inteligencija uz pomoć koje na računalni jezik prevodi, analizira i zatim nudi odgovor na postavljeno pitanje.

[35] Umjetna inteligencija i BD. URL: www.infotrend.hr/clanak/2015/1/suvremeni-proroci,82,1125.html

Ljudi umjetnu inteligenciju često pogrešno poistovjećuju s nekakvim automatiziranim sustavima. Sam pojam inteligencije uključuje mogućnost donošenja vlastitih odluka koje ponekad i neće biti u skladu s onime što bismo željeli od nekog sustava. Još je 1943. godine pisac znanstvene fantastike Isaac Asimov naveo tri zakona robotike koji bi se mogli primijeniti i na umjetnu inteligenciju, a prvi glasi da robot ili umjetna inteligencija ne smiju ni na koji način svojom akcijom dovesti u opasnost ljudsko biće.

Koliko god se umjetna inteligencija činila naprednom, nikada ne smijemo smetnuti s uma da ona ipak nema osnovne ljudske osjećaje poput empatije i osjećaja za dobro i zlo tako da bismo uvijek trebali biti iznimno oprezni pri određivanju koji ćemo stupanj samostalnosti dati takvim platformama i koji i koliki pristup podacima.

13. Zaključak

Sa početkom analitike podataka tvrtke i organizacije su uočile i zaključile da se svakodnevno povećavaju podaci te da to povećanje raste sve više, tako da se podaci multipliciraju. To utječe na povećavanje volumena podataka, na brzinu nastajanja i na različitost podataka. Pojedine tvrtke i organizacije su uvidjele njihov značaj i počele ih iskorištavati u svoju korist. Postepeno to postaje novi standard. Trenutačni od većih problema predstavlja velika heterogenost podataka, njihova veličina koja zahtijeva vrijeme za obradu podataka koje bi trebalo biti puno brže. Ujedno je potrebna i bolja i brža analitika za raščlanjivanje potrebnih i nepotrebnih podataka. S obzirom na to sve više je potrebno razvijanje sustava i opreme za pohranu i obradu novonastalih podataka.

Big Data uvjetuje napredak tehnologije ili nadogradnju postojeće tehnologije kako bi se novonastala velika količina podatka mogla pohraniti, obraditi i analizirati. Pravilno i adekvatno korištenje novonastalih tehnologija uvjetovat će potrebu za dodatnom edukacijom postojećeg osoblja, ali i potrebom za novim stručnim osobljem.

Zaključak je da će se trenutni problemi koje tvrtke imaju prilikom implementacije *Big Data* aplikacija s vremenom biti riješeni i da će se Big Data pokazati korisnim za samu tvrtku, njezine proizvode i poslovanje.

14. Literatura

1. Dumbill Edd. *Planning for Big Data* 2012., str.3.

2. Eathon Chris; deRoos Dirk; Deutsch Thomas; Lapis George; Zikopoulos Paul. *Understending Big Data* 2012., (str. 3 – 10)

3. Big Dana i nova tehnoologija URL:http://en.wikipedia.org/wiki/Big_data

4. Podijela karakteristika BD. URL: http://en.wikipedia.org/wiki/Big_data

5. Soubra, 2012. URL: http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

6. Klein,Tran-Gia,Hartmann, 2013. URL: http://www.gi.de/service/informatiklexikon/detailansicht/article/big-data.html

7. van Rijmenam, 2013. URL: https://datafloq.com/read/3vs-sufficient-describe-big-data/166

8. Dumbill, siječanj 2012. URL: https://beta.oreilly.com/ideas/what-is-big-data

9. Wang, John. *Encyclopedia of Business Analytics and Optimization.* 2014., str.316

10. Karakteristike BD. URL: http://en.wikipedia.org/wiki/Big_data

11. Poslovna inteligencija. URL: http://en.wikipedia.org/wiki/Business_intelligence

12. Nestrukturirani podaci. URL: http://www.webopedia.com/TERM/U/unstructured_data.html

13. Open source URL: http://www.apache.org/

14. Analitika BD. URL: http://www.webopedia.com/TERM/B/big_data_analytics.html

15. Pregled analitike BD. URL: http://www.datamation.com/applications/big-data-analytics-overview.html

16. Analitika BD. URL: http://www.webopedia.com/TERM/B/big_data_analytics.html

17. Mitovi o BD. URL: http://www.databasejournal.com/features/db2/exploding-the-myths-of-big-data.html

18. Poduzeća i BD. URL: http://www.databasejournal.com/features/db2/preparing-your-enterprise-for-big-data.html

19. Planiranje proračuna. URL: http://www.databasejournal.com/features/db2/it-budget-planning-for-big-data.html

20. BD i konkurentnost. URL: http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/

21. BD kao velik posao. URL: http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

22. Facebook i BD: URL: http://www.eweek.com/c/a/Data-Storage/How-Facebook-Is-Handling-All-That-Really-Big-Data-423736

23. Sigurnost BD. URL: http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view

24. Umjetna inteligencijaURL: www.infotrend.hr/clanak/2015/1/suvremeni-proroci,82,1125.html