Primjena data science koncepta i vizualizacija poslovnih podataka

Pejić, Ružica

Sveučilište Josipa Jurja Strossmayera u Osijeku

Ekonomski fakultet u Osijeku

Preddiplomski studij Poslovne informatike

Ružica Pejić

Primjena data science koncepta i vizualizacija poslovnih podataka

Završni rad

Osijek, 2019.

Sveučilište Josipa Jurja Strossmayera u Osijeku

Ekonomski fakultet u Osijeku

Preddiplomski studij Poslovne informatike

Ružica Pejić

Primjena data science koncepta i vizualizacija poslovnih podataka

Završni rad

Kolegij: Upravljanje informacijskim resursima

JMBAG: 0010217506

e-mail: rpejic@efos.hr

Mentor: prof. dr. sc. Josip Mesarić

Osijek, 2019.

Josip Juraj Strossmayer University of Osijek
Faculty of Economics in Osijek
Undergraduate Study of Business informatics

Ružica Pejić

Data science concept and visualization of business data

Final paper

Osijek, 2019.

IZJAVA
O AKADEMSKOJ ČESTITOSTI,

PRAVU PRIJENOSA INTELEKTUALNOG VLASNIŠTVA, SUGLASNOSTI ZA OBJAVU U INSTITUCIJSKIM REPOZITORIJIMA I ISTOVJETNOSTI DIGITALNE I TISKANE VERZIJE RADA

1. Kojom izjavljujem i svojim potpisom potvrđujem da je završni rad isključivo rezultat osobnoga rada koji se temelji na mojim istraživanjima i oslanja se na objavljenu literaturu. Potvrđujem poštivanje nepovredivosti autorstva te točno citiranje radova drugih autora i referiranje na njih.
2. Kojom izjavljujem da je Ekonomski fakultet u Osijeku, bez naknade u vremenski i teritorijalno neograničenom opsegu, nositelj svih prava intelektualnoga vlasništva u odnosu na navedeni rad pod licencom *Creative Commons Imenovanje – Nekomercijalno – Dijeli pod istim uvjetima 3.0 Hrvatska.*
3. Kojom izjavljujem da sam suglasan/suglasna da se trajno pohrani i objavi moj rad u institucijskom digitalnom repozitoriju Ekonomskoga fakulteta u Osijeku, repozitoriju Sveučilišta Josipa Jurja Strossmayera u Osijeku te javno dostupnom repozitoriju Nacionalne i sveučilišne knjižnice u Zagrebu (u skladu s odredbama Zakona o znanstvenoj djelatnosti i visokom obrazovanju, NN br. 123/03, 198/03, 105/04, 174/04, 02/07, 46/07, 45/09, 63/11, 94/13, 139/13, 101/14, 60/15).
4. Izjavljujem da sam autor/autorica predanog rada i da je sadržaj predane elektroničke datoteke u potpunosti istovjetan sa dovršenom tiskanom verzijom rada predanom u svrhu obrane istog.

Ime i prezime studenta/studentice: Ružica Pejić
JMBAG: 0010217506
OIB: 09478234335

e-mail za kontakt: rrpejic.97@gmail.com

Naziv studija: Preddiplomski sveučilišni studij poslovne informatike

Naslov rada: Primjena *data science* koncepta i vizualizacija poslovnih podataka

Mentor/mentorica rada: prof. dr. sc. Josip Mesarić

U Osijeku, 05. 07. 2019. godine

Potpis Ružica Pejić

Primjena *data science* koncepta i vizualizacija poslovnih podataka

SAŽETAK

Količina podataka koja se svakodnevno stvara u svijetu je svakim danom sve veća, a rast joj je ubrzaniji. Samim time stvaraju se enormne količine podataka koje je potrebno analizirati te iz njih stvoriti korisne informacije koje se kasnije koriste za unapređenje poslovanja.

Data science pruža mogućnosti brzih analiza pomoću softverskih rješenja za vizualizaciju podataka time upotpunjuje pogled na podatke dajući im novu vrijednost. U radu će biti dana teorijska podloga i razlozi korištenja *data scienca*, koncepta kao unapređenja obrade i interpretacije velikih količina, strukturiranih i nestrukturiranih, podataka

U radu će se na izabranom skupu podataka prikazati funkcionalnosti jednog od suvremenih alata za vizualizaciju, te obrazložiti postavljena hipoteza o dodanoj vrijednosti, lakšem razumijevanju i interpretaciji podataka.

Ključne riječi: *big data*, *data science*, vizualizacije, analitika

Data science concept and visualization of business data

ABSTRACT

The amount of data that is generated daily in the world is growing every day, and its growth is accelerating. This creates enormous amounts of data that needs to be analyzed and generate useful information that is later used to improve business.

Data science provides quick analysis capabilities using software data visualization solutions, thus completing the view of the data by giving them a new value. The paper will give the theoretical basis and the reasons for using data sciences, the concept of improving processing and interpretation of large quantities, structured and unstructured, data.

This paper presents the functionality of one of the modern visualization tools in the selected data set and explains the hypothesis about added value, easier understanding and interpretation of the data.

***Key words**: data science, big data, visualization, analytics*

Sadržaj

1. Uvod

Dvadeseto stoljeće je obilježeno značajnim tehnološkim napretkom i novim znanstvenim otkrićima. Početak razvoja informacijsko komunikacijskih tehnologija kakve su danas poznate svijetu bilježi se početkom dvadesetog stoljeća, a razvoj digitalnih tehnologija seže u osamdesete godine dvadesetoga stoljeća. Riječ je o vrlo turbulentnom vremenu kada je informacijsko komunikacijska tehnologija u pitanju. Osim novosti na hardverskom polju (prva osobna računala, početci danas dominantnih tvrtki u svijetu računalnih tehnologija), informacijsko komunikacijske tehnologije su napredovale i u softverskom smislu. Osobito je bitna pojava World Wide Weba kao sredstva umrežavanja računala te njegov razvoj kada je u pitanju količina i vrsta dostupnih informacija. Bitno je napomenuti kako su same tehnologije utjecale na razvoj društva i gospodarstva na globalnoj razini.

Od tada rapidan razvoj informacijsko komunikacijskih tehnologija rezultira stvaranjem enormnih količina podataka, globalnog fenomena koji se naziva *big data*. Riječ je dakle, ne samo o velikoj količini već navedenih podataka, nego i o raznolikosti formata i bilježenju stalnih promjena u vremenskom smislu to jest ažuriranju baza i odnosa među samim podatcima.

Može se zaključiti kako se kao posljedica globalizacije (povezanosti svijeta) javlja velika količina informacija koje je potrebno obraditi na odgovarajući način. Odgovor na tu problematiku daju informacijske znanosti točnije *data science* koncept koji će biti objašnjen u nastavku rada.

1.1 Identifikacija problema i istraživačka pitanja

Kako je već spomenuto, ekonomskim razvojem, globalizacijom i općenito povezivanjem poslovnih subjekata (međusobno, prema klijentima ili državi), stvara se opsežna dokumentacija, bilo u fizičkom ili danas sve više digitalnom obliku. Rješenje velikog broja fizičke dokumentacije je bilo u njihovoj digitalizaciji, što zbog sigurnosti, a što zbog lakše pretrage i analize samih podataka. S obzirom na sve brži razvoj tehnologija, količina podataka, na svjetskoj razini se udvostručuje svake tri godine bez vremenskog i prostornog ograničenja, pa tako ulaskom u informatičku eru količina digitalnih zapisa, dokumenata, podataka iz baza i repozitorija postaje nepregledna te njihova analitika biva sve teža, u krajnjem slučaju nemoguća bez naprednih alata.

Obrada i razumijevanje velikih količina digitalnih podataka je glavna problematika ovoga rada. Kako je poslovna okolina, ali i svijet, sve ubrzaniji, odluke koje menadžeri donose

moraju biti brze i pravilne, te su s toga razvijeni brojni alati za potporu odlučivanju, ali i drugi alati koji su od velikog značaja u poslovnom svijetu odnosno odlučivanju temeljenom na podatcima.

Već spomenuti „veliki podatci" postaju nesagledivi, zbog čega je klasična analitika nad takvim podatcima nemoguća, oduzima puno dragocjenog vremena i ima veliku vjerojatnost pogrešaka koje u poslovnom svijetu mogu biti donijeti velike gubitke, osobito ako je riječ o strateškim odlukama. Zato se danas statistički proračuni i analitika izvode pomoću specijaliziranih modela i njima prilagođenih programskih rješenja, počevši od alata otvorenog koda, do komercijalnih, specijaliziranih softvera koji osim algoritama za statističke proračune imaju mogućnosti vizualnog prikazivanja informacija. Alati obuhvaćaju i različite tipove vizualnih prikaza, vrlo pregledne grafikone i različite karte koje u trenutku postaju interpretabilne i temelj donošenja poslovnih odluka. Nadalje, zbog raznolikosti podatkovnih formata i struktura, razvijeni su softveri koji omogućuju analizu podataka neovisno tome je li riječ o strukturiranim ili nestrukturiranim podatcima, tekstualnim, tabličnim ili multimedijalnim datotekama.

Također je bitno napomenuti kako se razlikuje sama analitika ali i programska rješenja namijenjena podatcima sa pravilnom strukturom, od onih koji su namijenjeni podatcima nepravilne strukture i različitih formata. Tek u slučaju analize velikih nestrukturiranih podataka, njihova vizualizacija postaje ključni element razumijevanja istih, budući da je takva analitika dugotrajnija i kompleksnija od analize podataka s pravilnom strukturom. Davanjem nove vrijednosti podatcima u vidu vizualnog prikaza, efikasno i efektivno se postiže stvaranje kvalitetne informacije koje se mogu upotrijebiti za unapređivanje poslovanja.

Bitno je napomenuti i samu prirodu čovjeka u ovome slučaju. Od osjetila, čovjek se najviše oslanja na vid. Stoga se kaže da su ljudi vizualna bića, te da najlakše pamtimo slike, boje, oblike. Shodno tome, naša vizualna percepcija je otvorila vrata tehnici prikazivanja podataka, bilo da je riječ o starim ručnim metodama, ili digitalnom prikazu podataka. Vizualizacija velikih količina podataka podrazumijeva grafičko prikazivanje statističkih proračuna, trendova i odstupanja putem programskih rješenja, te su u svakom slučaju jednostavnije za interpretaciju i oku ugodnije od tabličnih ili brojčanih zapisa.

Najveći problem kod skladištenja te interpretacije velikih količina podataka su svakodnevna ažuriranja to jest. stvaranje novih setova podataka koji (ne)odgovaraju strukturi već poznate

baze. Brojni su pokušaji da se vizualizacija kompleksnih podatkovnih struktura izvede na primjeren način (oblik grafičke reprezentacije, brzina izvođenja i dinamika promjena).

Polazeći od *data science* kao znanosti o podatcima te podataka sa svojim kvalitativnim i kvantitativnim vrijednostima, u radu će se nastojati doći do odgovora na slijedeća istraživačka pitanja:

„U kojoj mjeri vizualizacija podataka doprinosi brzini donošenja poslovnih odluka?"

„Koliko i na koji način su vizualizirani podatci laki za interpretaciju i razumijevanje njihova međusobnog odnosa?"

„Koji alati su najpogodniji za vizualizaciju podataka?"

1.2 Cilj rada

Cilj rada je je doći do što cjelovitijih odgovora na postavljena istraživačka pitanja. U tu svrhu će se na kompleksnom skupu podataka pokazati funkcionalnost alata za vizualizaciju u smislu bržeg donošenja odluka i interpretabilnosti postignutih izračuna. U matematičko statističkom smislu, alati služe za dobivanje to jest izračun korisnih pokazatelja o poslovanju poduzeća, počevši od najjednostavnijih kao što je deskriptivna statistika do složenih metoda i alata koji obuhvaćaju inferencijalnu statitstiku i druge metode.

Uz izračun navedenih ali i drugih pokazatelja, alati daju i oku ugodne vizualne prikaze podataka u vidu grafikona. Autori programskih rješenja su se pobrinuli da ona budu brza, točna i jednostavna za korištenje te da vizualni prikazi budu laki za interpretaciju.

Na kraju će se dati pregled raspoloživih podataka o alatima za vizualizaciju u pogledu njihovih ključnih karakteristika i prihvaćenosti od strane korisnika.

2. Teorijska podloga i prethodna istraživanja

S obzirom na navedeni rast količine informacija u svijetu, u znanstvenim krugovima se koncem dvadesetog stoljeća počelo govoriti i podatkovnoj znanosti, kao potrebi i rješenju problema nastalih gomilanjem podataka koje je potrebno analizirati. Uzevši u obzir rasprostranjenost podataka i njihovu važnost, isti postaju predmetom istraživanja mnogih znanstvenika. Neovisno o znanstvenoj grani kojoj pripadaju i stavovima prema određenim problemima, stajališta prema podatcima su uglavnom ista to jest, naglašava se njihova važnosti i iskoristivost.

„Sa sigurnošću možemo reći da smo u tijeku jedne nove velike revolucije koja ima i svoje prigodno ime Big Data – Veliki podatci. Iako su termin osmislili znanstvenici iz područja poput astronomije i genomike, Veliki podatci su posvuda. Oni su istovremeno i resurs i alat čiji je glavni zadatak informiranje" (Kocijan 2014:1). Prema tome, vidljivo je kako je pojam velikih podataka prisutan u svim znanstvenim granama počevši od humanističkih znanosti preko tehničkih koje su temelj podatkovne znanosti pa do prirodnih znanosti gdje se u prvom redu govori o medicini kao znanstvenoj grani i djelatnosti čiji su podatci od velikog značaja i gdje su greške u analizi nedopustive.

Isto tako bilježi se značajan „porast broja znanstvenih istraživanja, povećanja broja znanstvenika i istraživača i pojave novih znanstvenih disciplina koji je doveo do eksponencijalnog povećanja broja časopisa, knjiga, zbornika radova s kongresa, disertacija, patenata, tehničkih izvještaja i drugih publikacija u kojima se objavljuju rezultati znanstvenoga rada." (Poropat, Marušić, Štimac 2017:455)

Također, bilježe se svakodnevni podatci o društvu i okolini te se oni koriste kako bi se povećala kvaliteta života. „S povećanjem pozornosti posvećene velikim podacima postupno postaje jasno da veliki geoprostorni podaci imaju važnu ulogu u povećanju sposobnosti čovjeka da prati i razumije društvo i prirodu te da reagira na probleme okoliša s prostornim i vremenskim dimenzijama. Od 17 ciljeva za održivi razvoj do 2030., što ih je UN naveo 2015., najmanje osam ih može na različite načine imati koristi od velikih podataka o Zemlji (Big Earth Data)." (Frančula 2017:97)

Mnogi znanstvenici napominju kako velike količine podataka nisu vezane samo za Internet, te ističu njihovu važnost u financijama, medicini, bioinformatici, obrazovanju, socijalnoj skrbi i ostalim djelatnostima. (Schutt, O'Neil 2014.)

S ciljem unapređivanja poslovanja, korištenje velikih količina podataka u poslovnim subjektima je od velikog značaja u smislu konkurentske prednosti. Analitika velikih količina podataka pomoću suvremenih alata daje daleko korisnije informacije i bolje rezultate po samu organizaciju od tradicionalne statističke analitike. (McAfee, Brynjolfsson 2012:4) Sastavni dio naprednih alata za podatkovnu analitiku su alati za vizualizaciju podataka koji su od velike pomoći prilikom razumijevanja dobivenih rezultata. Dobivene vizualizacije koriste perceptivne sposobnosti ljudi, omogućuju uvid u podatke i olakšavaju donošenje poslovnih odluka. Isto tako, suvremene grafičke tehnike pružaju mogućnost prikazivanja i analize podataka i njihovih svojstava s kompleksnim relacijama. (Fayyad, Grinstein, Wierse 2002:4)

Iz navedenih razmišljanja hrvatskih, ali i svjetskih znanstvenika, vidljivo je kako se podatcima i njihovu značenju predaje velika važnosti ne samo u teoriji nego i u praksi. Isto tako, naglašava se važnost vizualiziranja podataka kao bitnog elementa podatkovne analitike.

3. Metodologija rada

3.1 Klasični alati za vizualizaciju kao potprogrami tabličnih kalkulatora i baza podataka

Da bi se podatci mogli vizualno prikazati, potrebno je imati pravilnu strukturu te najčešće tablični oblik istih. Programska rješenja za stvaranje takvih struktura su tablični kalkulatori i baze podataka.

Od početaka razvoja znanosti o podatcima, alati za grafičko, vizualno prikazivanje su dostupni kao dio programskih rješenja za skladištenje podataka, ali i za njihovu manipulaciju.

Jedni od prvih programa te vrste su Visicalc te Lotus 123 nakon kojih se pojavio *Excel*, program za proračunske tablice koji je razvijen od strane *Microsofta* koji zajedno s drugim alatima čini sastavni dio programskog paketa *Microsoft Office*.

Osim brojnih funkcionalnosti, u *Excel* je moguće spremiti i veliki broj podataka. Naime, tablična struktura Excela, u sastavu *Microsoft Office 365*, omogućava spremanje 1 048 576 redaka, koji su obilježeni brojem te 16 384 stupaca obilježenih slovima, u jednom radnom listu. Ako govorimo o brojnosti radnih listova, broj je ograničen dostupnom memorijom i resursima sustava.

Kreirati jednostavan vizualni prikaz nad jednostavnim skupom podataka je vrlo lako i vremenski nije zahtjevno te je na službenim *Microsoftovim* stranicama za podršku korisnicima moguće pronaći i detaljne upute za izradu grafikona, koje se sastoje od svega pet koraka. Osim toga, *Microsoft Excel* omogućava brojne operacije kosu su potrebne kako bi korisnik dobio prikaz željenih podataka na odgovarajući način. Također, brojne su mogućnosti glede dizajna samih grafičkih prikaza, i što je najvažnije, novije verzije programa nude vrlo jednostavno korisničko sučelje pa je snalaženje u samom programu olakšano.

S obzirom na raširenost upotrebe Excela, kao glavnog tabličnog kalkulatora i najčešće jedinog alata za manipulaciju podatcima i njihovo prikazivanje, te na trend svakodnevnog rasta baza i brojnosti podataka, dolazi do otežane analitike podataka. Ako se uzme u obzir kompleksnost tržišta i izazovi koji se postavljaju pred poslovne subjekte, postaje vrlo teško izvući korisnu informaciju iz nepreglednih tabličnih struktura. Dakako, barijere je moguće premostiti, no za vizualni prikaz velike količine podataka te za filtriranje, posebno obilježavanje i slične dorade, utroši se dosta vremena koje je ekonomski gledano jedan od najvažnijih resursa u svim sferama života, pa tako i u poslovnoj. Zbog navedenih razloga, za prikaze pojedinih

podataka, sve se više pribjegava novijim programskim rješenjima specijaliziranim samo za vizualni prikaz podataka.

3.2 Suvremeni alati za vizualizaciju – novi pristup vizualizaciji

Današnji alati za vizualizaciju podataka, nadmašuju standardne grafičke prikaze koji se koriste u *Microsoft Excel* radnim listovima, prikazujući podatke na sofisticiraniji način kao što su infografike, toplinske i zemljopisne karte. Prikazi mogu biti interaktivni dopuštajući korisnicima da manipuliraju njima te omogućavaju složenije upite i analizu podataka.

Razvojem informacijskih i računalnih tehnologija, alati za vizualizaciju, postali su neizostavan dio većeg sustava, tzv. Poslovne inteligencije (BI – *Business Intelligence*). Uspjeh dva najveća lidera na polju programskih rješenja za vizualizaciju podataka, potisnuo je sa tržišta i iz upotrebe druge alate i tehnologije. Riječ je o alatima *Tableau* i *Qlik*. Prednost ovih programskih rješenja leži u lakšem rukovanju u usporedbi sa tradicionalnim programima za statističku analizu, kod ranijih verzija poslovne inteligencije.

Također, programi za analizu podataka igraju važnu ulogu u naprednoj analitici kroz prediktivnu analitiku i strojno učenje zbog olakšane interpretacije samih modela. Naime, pisanje programskog koda u svrhu predviđanja budućih vrijednosti iz postojećih (prediktivna analitika) zahtjeva vizualizaciju dobivenih rezultata, što osigurava ispravan rad modela.

Brojne su prednosti i funkcije modernih programskih rješenja za vizualizaciju, osobito kada je u pitanju velika količina podataka. Stoga će u radu biti prikazan praktični empirijski primjer, vizualizacije poslovnih podataka, pomoću alata *Tableau*, kako na brz i jednostavan način napraviti oku ugodnu vizualizaciju. (Rouse 2017.)

4. *Data science* – problematika

4.1 Objašnjenje pojmova i nastanak fraze „*data science*“

„*Data Science* (podatkovna znanost) se može definirati kao interdisciplinarno područje čiji je glavni zadatak ekstrakcija informacija i znanja iz podataka bilo da oni dolaze u strukturiranom ili možda nestrukturiranom obliku.“ (*Data Science Croatia* 11.11.2015.)

Podatkovnu znanost karakteriziraju tri pojma; velika količina podataka, strojno učenje i vizualizacije. Također, znanstvene grane koje obuhvaća su statistika, matematika, napredna analitika. Na neki način se može reći kako je podatkovna znanost u današnjem obliku vrlo mlada s obzirom da se njen razvoj temelji na razvoju tehnologije.

Može se reći kako su korijeni podatkovne znanosti duboko povezani i isprepleteni sa statistikom i statističkim modelima, te koristeći napredak u tehnologiji i druge već navedene koncepte, izrasta u samostalnu znanost koja sve više dobiva na značaju. Veliku ulogu u razvoju ove znanosti ima razvoj svijeta i društva u cjelini, povezanost svijeta, globalizacija i Internet, ne samo kao masovni medij nego nužnost čovjeka 21. stoljeća. Sve su te pojave rezultirale stvaranjem globalnog fenomena – *Big Data*.

Nastanak podatkovne znanosti rezultirao je pojavom podatkovnih stručnjaka, tzv. *Data scientista*. Prema tome, postoje i različite podjele te klasifikacije podatkovnih stručnjaka, primjerice podjela na profesije zadužene za izgradnju resursa te profesije koje koriste iste te resurse. Kada je riječ o podatkovnoj znanosti, zanimanja koja ona obuhvaća su brojna.

Menadžer velikih podataka posrednik je između svih članova tima koji radi na Velikim podatcima ali i između tima i ostalih službi u organizaciji. Njegov je zadatak kreirati protokole kojima se opisuju procesi kojima se omogućuje ponovna identifikacija neidentificiranih objekata, ali i nadgledanje tih procesa.

Nadalje, može se govoriti o podatkovnim znanstvenicima koji imaju dobre statističke, matematičke, programerske, vizualizacijske i pripovjedačke sposobnosti; analitičarima koji prikupljaju i organiziraju podatke; vizualizatorima; arhitektima; inženjerima, agentima podatkovnih promjena te konzervatorima izvora velikih podataka. (Kocijan 2014:20-22)

4.1.1 Vremenska linija razvoja podatkovne znanosti (data science)

Godine **1962**, John Tukey pisao je o promjenama u svijetu statistike, govoreći „...proučavajući matematičke i statističke sadržaje, imao sam razloga za pitanja i sumnju... Počeo sam osjećati kako je moj glavni interes analitika podataka“ Autor se ovom mišlju

odnosio na spajanje statističkih i računalnih znanosti, shodno vremenu, kada je količina podataka postala tolika da su bili potrebni sati i dani računanja, ne bili se dobio željeni rezultat" (Foote 14.12.2016.).

Peter Naur je, **1974**. godine, prvi upotrijebio izraz *data science*, dajući vlastitu definiciju nove znanstvene discipline: „Znanost koja se bavi podatcima, kada se jednom uspostavi, dok se odnos prema onome što podatci predstavljaju delegira na druga područja i znanosti." (Press 28.05.2013.). Kao posljedica novih inicijativa o povezanosti podataka, metoda i modela za njihovu analizu i računalnih tehnologija osnovano je **1977**. IASC - *International Association for Statistical Computing* (hrv. Međunarodno udruženje za statističko računalstvo) u čijem statutu se ističe: "Naša misija je povezati tradicionalnu statističku metodologiju, modernu računalnu tehnologiju, ekspertno znanje, u svrhu prevođenja podataka u informacije u znanje" (Lauro 1996:191)

Kasnije godine, **1980-e i 1990-e**, protekle su u znanstvenom definiranju znanosti o podatcima te njezine važnosti i potrebe, te povezivanja tradicionalnih i novih načina obrade i analize podataka.

Izraelski znanstvenik, sa područja menadžmenta, Jacob Zahavi je **krajem 20. stoljeća** ukazivao na problematiku vezanu za tadašnju količinu podataka: „Konvencionalne statističke metode rade dobro sa manjim skupovima podataka. Međutim, današnje baze podataka, mogu uključivati milijune redaka i stupaca. ... Novi tehnološki izazovi leže u razvoju modela koji mogu uraditi bolji posao, analizirati podatke, detektirati nelinearne veze i interakcije među elementima." (Volugaris 2014:21)

Prema nekim izvorima, prvu definiciju Velikih podataka kao fenomena daje Diebold (**2000.**) koji navodi: „U zadnje je vrijeme dosta dobre znanosti, bez obzira je li u pitanju fizika, biologija ili sociologija, bilo prisiljeno suočiti se – od čega je često i profitirala – s fenomenom Velikih podataka. Veliki podatci odnose se na eksploziju u količini (a katkad i kvaliteti) dostupnih i potencijalno relevantnih podataka, uglavnom kao posljedica skorih i besprimjerenih napredaka u tehnologiji zapisivanja i pohranjivanja podataka."(Kocijan, 2014:2)

Software-as-a-Service (SaaS) je uspostavljen **2001**. godine, što je omogućilo distribuciju aplikacija putem interneta. Iste godine William S. Cleveland, izražava potrebu za podatkovnim „znanstvenicima" (*data scientist*), glede upoznavanja potreba budućnosti.

Godinu dana kasnije, objavljen je prvi *data science* časopis, publikacija fokusirana na područja kao što su aplikacije, Internet, *data systems*. (Foote 2016.)

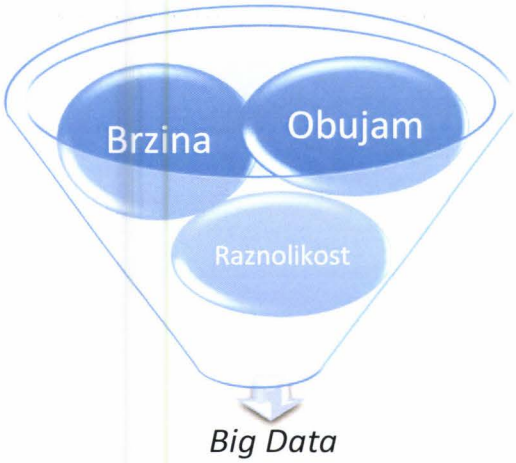
Spomenuta „titula" *data scientist*, **2008**. godine postaje viralna i sveprisutna riječ govorenog jezika.

Oglasi za zapošljavanje *data scinetist*a, se **2011**. Godine povećavaju za 15,00%. Također se bilježi znatan porast broja konferencija i seminara posvećenih podatkovnoj znanosti i velikim količinama podataka. Iste godine James Dixon, promovira koncept *Data Lakes*, umjesto *Data Warehouses*, govoreći o prvom terminu kao mjestu spremanja sirovih neobrađenih i nestrukturiranih podataka, dok isti ti podatci pravilnim rasporedom postaju dio „skladišta podataka".(Woods 26.01.2015.)

Do **2013**. Godine, prema IBM-u je količina svjetskih podataka porasla za 90% (2011.-2013.). (Jacobson 24.04.2013.) Idućih godina, količina podataka na svjetskoj razini je dosegla enormnu količinu, a tehnologija sve više teži razvoju umjetne inteligencije, strojnog učenja i „pametnih" uređaja. S obzirom na svoju raširenost, ali i raširenost podataka, *data science* je postala neizostavan dio poslovnih i akademskih istraživanja. Osim toga, ova znanost se primjenjuje i u drugim znanstvenim granama te gospodarskim sektorima, počevši od biologije i genetike, preko inženjerstva i astronomije pa do ekonomije (ponajprije digitalne ekonomije) i vlada. (Foote 2016.)

4.1.2 Definicija pojma „*big data*" kroz 3V

Većina definicija velikih količina podataka fokusira se samo na njihov obujam, koji iako je jedna od glavnih karakteristika ovoga fenomena, nije jedini koji ga opisuje u potpunosti. Brzina i raznolikost, zajedno sa obujmom podataka čine tri glavne karakteristike *big data*, to jest 3V(eng. *volume, variety, velocity*), kao što je prikazano na slici 1 (veličina ili obujam, brzina i raznolikost).



Slika 1 Grafički prikaz 3V

S mišlju o velikoj količini podataka, uvriježeno je mišljenje kako, kada je riječ o memoriji, podatci zauzimaju terabajte a ponekad i petabajte prostora. Količina podataka može biti određena brojanjem transakcija, datoteka ili tablica. Isto tako, neke organizacije, korisnijim smatraju vremensku dimenziju kretanja varijabli. Djelokrug velike količine podataka, reflektira se također na njezinu određenost u smislu razlikovanja podataka namijenjenih analizi od klasičnih skladišta podataka.

Ono što velike podatke uistinu čini velikima je to što dolaze iz puno različitih izvora, čiji se broj neprestano povećava. Mnogi od njih su zapravo web izvori, uključujući zapise o prijavama u različite sustave, posjete web stranicama i društvene mreže kao nepresušan izvor kako informacija tako i samih podataka. Dakako, veliki broj organizacija sakuplja podatke iz okoline, no za većinu njih je to postalo opterećenje, te pribjegavaju korištenju sofisticiranih alata za njihovu analizu. U tom pravcu razmišljanja, mogu se izdvojiti strukturirani i nestrukturirani podatci. No, osim toga, razvojem *web*a, podatci dolaze u raznim drugim oblicima kao što su slikovni i zvučni zapisi to jest multimedijalne datoteke.

Veliki podatci, također mogu biti opisani kroz brzinu kojom nastaju odnosno kojom se umnožavaju. Riječ je o brzini „dostave" podataka s mnogih uređaja ili senzora, kao što su: robotski uređaji u proizvodnji, toplinski senzori, mikrofoni i video kamere. Sakupljanje velike količine podataka u stvarnom vremenu, nije novost. Naime, mnoge organizacije i poslovni subjekti pribjegavaju tome u cilju najboljeg mogućeg iskorištavanja *web*a za unaprjeđenje svog poslovanja. Pomoću mnogih senzora različitih vrsta, već ionako velika količina podataka postaje sve veća te se u konačnici može reći kako se brzina kao karakteristika velikih količina podataka ogleda u stalnom ažuriranju baza i skladišta podataka. (*Russom* 2011:6-8)

Kada je riječ o raznolikosti podataka, njihovoj veličini i brzini, može se reći kako svaka na svoj način i u jednakoj mjeri čini velike podatke onime što oni jesu.

4.1.3 Razlika između velikih podataka i standardnih podataka

Bitno je napomenuti kako velike podatke ne odlikuje samo veliki broj zapisa u bazama podataka, te shodno tome podatci se mogu podijeliti na velike podatke(VP) i standardne podatke.(SP) „Pojam 'puno podataka' odnosi se na velike zbirke zapisa jednostavnoga formata" koje je nadalje moguće analizirati, ali i izmjenjivati u smislu brisanja, nadopunjavanja ili promjena same strukture baze. Prema tome, razlike se mogu razvrstati prema određenim kriterijima:

1. ***ciljevi***: SP daje odgovore na specifična pitanja s jedinstvenim ciljem dok VP daju čitav spektar odgovora s promjenjivim ciljem.

2. ***lokacija***: SP se uglavnom nalaze unutar određene organizacije; VP mogu biti raspoređeni na različite lokacije.

3. ***struktura i sadržaj***: SP su najčešće strukturirani podatci; VP su uglavnom nestrukturirani te sadrže podatke različitih formata u prvom redu misleći na multimedijalne podatke.

4. ***priprema:*** SP uglavnom priprema korisnik tih podataka; VP uglavnom pripremaju timovi stručnjaka koji u najčešćem broju slučajeva nisu korisnici istih.

5. ***životni vijek***: SP imaju kratak vijek postojanja (oko 7 godina); VP ne zastarijevaju

6. ***mjerenja***: SP se mjeri pomoću jednog protokola; VP se mjeri pomoću različitih protokola misleći pri tome na različite vrste analiza, tehnika i alata te u konačnici njihove ishode. (Berman 2013:89)

7. ***reproduciranje***: projekti koji koriste SP mogu se lako reproducirati; projekti koji koriste VP rijetko kada se mogu reproducirati.

8. ***financijsko ulaganje***: financije uložene u projekt koji se bazira na SP su uglavnom male za razliku od financijskog ulaganja u projekte bazirane na VP.

9. ***introspekcija***: Pojedinačni SP mogu se lako locirati u bazi pomoću redaka i stupaca; Kada je riječ o velikim podatcima, rabi se tehnika introspekcije koja se zasniva na objektno orijentiranom programiranju to jest samoprepoznavanju objekata, njihovih svojstava i vrijednosti.(Berman 2013:60)

10. ***analiza***: Kod SP analiza se može vršiti nad svim podatcima istovremeno: VP zahtijeva složeniji postupak analize (izvlačenje, pregledavanje, smanjivanje, ‚normalizacija, transformacija, interpretacija te ponovna analiza). (Kocijan 2014:4)

4.2 Znanje u podatcima; Data- Information-Knowledge-Whisdom koncept

Kako bi se bolje objasnila znanost o podatcima, potrebno je razlikovati podatke od informacija, znanja i mudrosti (DIKW koncept). „Koncept DIKW-hijerarhije (engl. *Data-Information-Knowledge-Wisdom hierarchy*) osmišljen je kako bi oslikao odnos između danas nezaobilaznih pojmova u znanstvenom i praksiološkom diskursu – podatka, informacije, znanja i mudrosti."(Bosančić 2017:2)

U području upravljanja znanjem ovaj pojam često se naziva hijerarhijom mudrosti, piramidom mudrosti, te hijerarhijom ili pak piramidom znanja. znanja „Piramida (ili trokut) najčešći je grafički oblik koji u literaturi reprezentira DIKW-hijerarhiju (slika 2).

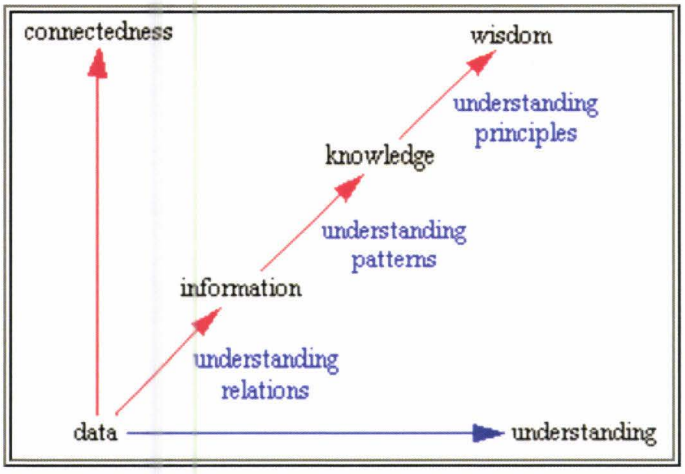
Iznad „sloja podataka" nalazi se „sloj informacija" značajno manjeg volumena (ili površine); iznad sloja informacija nalazi se „sloj znanja" još manjeg volumena (odnosno površine), a iznad „sloja znanja" najmanji je od sva četiri grafički prikazana koncepta, „sloj mudrosti"; no upravo tom sloju pripada vrh piramide. „ (Bosančić 2017:3)

Slika 2 Hijerarhija DIKW koncepta

Počevši od vrha, hijerarhije, termini se, prema većini znanstvenika, definiraju na sljedeći način. Mudrost je sposobnost stvaranja novih znanja iz postojećih znanja, dok se inteligencija ogleda u postizanju efikasnosti, to jest usavršavanja postupka i sredstava u cilju ostvarenja zamisli. Nadalje, znanje podrazumijeva *know-how* i to je ono što omogućava pretvaranje

informacija u instrukcije za rad. Informacije odgovaraju na pitanja „što, tko, kada i gdje“ te daju smisao podatcima koji se nalaze na dnu piramide. Oni se mogu definirati kao simboli koji prezentiraju obilježja pojedinih objekata, događaja i njihovog okruženja. Podatci su u svojoj biti, zapravo rezultati istraživanja.

Za sirove podatke je karakteristično da generiraju najmanju razinu razumijevanja, te međusobno nisu povezani dok, mudrost sa svim pripadajućim svojstvima odražava najveću razinu povezanosti i razumijevanja. Prema tome, može se zaključiti kako kretanjem od dna piramide prema vrhu, razina razumijevanja kako i međusobne povezanosti podataka je sve veća, što je grafički prikazano na slici 3. (Bellinger, Castro, Mills 2004.)



Slika 3 DIKW - povezanost i razumijevanje,
Izvor: http://www.systems-thinking.org/dikw/dikw.htm

Uzimajući u obzir sva svojstva DIKW koncepta, mnogi autori progovaraju o manjkavostima istoga. Riječ je o kritikama DIKW koncepta: (Bosančić 2017:16)

- kritika logičkih pretpostavki DIKW-hijerarhije i modela koja u obzir uzima „odnos koji je uspostavljen između ključnih pojmova u konceptu i međusobnu logičku nekonzistentnost i nesvodivost.“(Bosančić 2017:16)
- kritika epistemoloških pretpostavki DIKW-hijerarhije i modela – nejasnoće vezane uz primjenu DIKW hijerarhije na ljudsko/računalno stjecanje znanja

- Kritika simboličnog prikaza i metaforične interpretacije daje DIKW konceptu uz simbolično i metaforično značenje, razlikujući pri tome pojam simbola od pojma metafore.

4.3 Suvremene podatkovne strukture poslovnih podataka: baze i skladišta podataka

Kroz prethodni tekst je navedeno kako organizacije ponekad prave razliku između skladišta i baza podataka u svrhu naprednije analize te donošenja poslovnih odluka (sustavi za potporu odlučivanju). Može se reći kako ideja skladištenja podataka leži u izdvajanju podataka iz operativne baze u posebne baze namijenjene kompleksnijim analizama. Prema tome, skladište se može definirati kao „subjektno orijentiran, integrirani, nepromjenjiv, a vremenski dinamičan skup podataka za potporu odlučivanja ."(Ćurko 2001:843) Tako uređeni podatci omogućavaju dobivanje potrebnih i nužnih informacija kroz analitiku i upite. „Najčešći izvor podataka za sustav skladišta podataka jesu transakcijske odnosno relacijske baze podataka." (Mekterović, Brkić 2017:7) „U transakcijskim bazama podataka i bazama klijenata nalaze se velike količine podataka. Međutim, transakcijske baze podataka su goleme, pa se za potrebe analize podataka izabire uzorak na kojem će se one vršiti." (Panian 2007:159)

Shodno tome, možemo razlikovati transakcijske sustave od samih skladišta podataka u više dimenzija ili karakteristika koje su prikazane na tablici ispod.

Transakcijski sustav	Skladište podataka
sadrži trenutne podatke	sadrži povijesne podatke
pohranjuje detaljne podatke	pohranjuje detaljne i sumarne podatke
podaci su promjenjivi	podaci su postojani
velika učestalost transakcija	srednja ili mala učestalost transakcija
predvidljivi načini korištenja (ponavljaju se)	nepredvidljivi načini korištenja
orijentiran ka dnevnim operacijama i vođenju poslovnog sustava	orijentiran ka analizi podataka
potpora dnevnim, operativnim odlukama	potpora strateškim odlukama
poslužuje velik broj operativnih korisnika	poslužuje manji broj korisnika obično pozicioniranih u upravljačkim strukturama poduzeća (mada postoji trend sve veće dostupnosti skladišta podataka svim članovima poduzeća kao potpora svih vrsta odluka)
izuzetno važna raspoloživost	manje važna raspoloživost
težište na pohranjivanju podatka	težište na dobavljanju informacija

Slika 4 Razlike transakcijskog sustava i skladišta podataka, Izvor: https://www.fer.unizg.hr/_download/repository/SKRIPTA_-_Skladista_podataka_i_poslovna_inteligencija.pdf

Sustavi za potporu odlučivanju, dio su poslovne inteligencije te, ukoliko su zasnovani na konceptu skladišta podataka, omogućuju pravovremeno dobivanje kvalitetne informacije i donošenje ispravnih odluka analitikom podataka.

„Polazeći od činjenice da su poslovni podaci koje organizacije rutinski prikupljaju obavljajući svoje poslovne aktivnosti heterogeni, može se uočiti kako postoje dvije velike skupine izvora podataka, i to:

- Vanjski izvori podataka: podatci pristižu iz okruženja organizacije, odnosno s tržišta na kojima te organizacije djeluju.
- Unutarnji izvori podataka: podaci nastaju realizacijom poslovnih procesa unutar same organizacije." (Panian 2007:1)

Definicija same baze podataka u računalskom smislu govori da je „baza podataka skup međusobno povezanih podataka, pohranjenih u vanjskoj memoriji računala." (Manger 2003:3)

„Shema baze podataka opisuje predmete koji su prikazani u bazi podataka i odnose među njima, dok je model baze podataka skup pravila koji određuje kako može izgledati logička struktura baze i čini osnovu za koncipiranje, projektiranje i implementiranje baze." (Poropat, Marušić, Štimac 2017:455)

Računalni programi koji se koriste za upravljanje bazama podataka, ispitivanje, i njihovo pretraživanje, oblikovanje te rad s podatcima, jednim imenom se nazivaju Sustavima za upravljanje bazama podataka (DBMS – engl. *Database Management System*).

Može se reći kako su podatci u bazama podataka organizirani u skladu s modelom baze podataka. Danas najrašireniji model je relacijski model baza podataka koji se zasniva na pojmu relacije to jest veze među podatcima koji se prikazuju u pravokutnim tablicama.

Jedna od glavnih karakteristika baza podataka je ta da se podatcima može pristupiti isključivo putem posebnih programskih jezika (SQL - *Structured Query Language*), upitnih jezika za rad s bazama podataka. „Jezik je originalno razvijen u IBM-u 70-ih godina prošlog stoljeća, te je postao glavni jezik za rad s bazama podataka." (Mujadžević 2016:3)

Jezici kojima korisnici komuniciraju sa sustavima za upravljanje baza podataka moguće je podijeliti u tri skupine.

- Jezik za opis podataka (*Data Description Language* – DDL) , koji služi za definiranje podataka i veze među podatcima, na logičkoj razini.
- Jezik za manipuliranje podatcima (*Data Manipulation Language* – DML) omogućava "manevriranje" po bazi, te jednostavne operacije kao što su upis, promjena, brisanje ili čitanje.

- Jezik za postavljanje upita (*Query Language* - QL). Služi neposrednom korisniku za interaktivno pretraživanje baze te po svom obliku podsjeća na govorni engleski jezik.

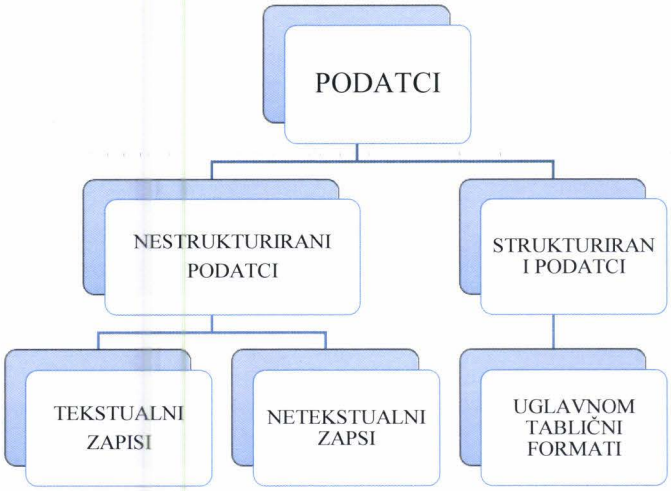
Objekt iz stvarnog svijeta koji se želi predstaviti u bazi podataka naziva se entitetom. Entitet može biti stvaran objekt ili osoba (primjerice student) ili pak apstraktni objekt ili događaj (pohađanje nastave). Tip entiteta određen je njegovim atributima.

Ako se u obzir uzme kompleksnost nestrukturiranih podataka iz transakcijskih baza, te raznolikost podatkovnih formata, organizacije se nalaze pred izazovom kada je njihova analitika u pitanju. Kako bi se dobili željeni formati podataka, pogodni za analitiku, organizacije uglavnom koriste ETL tehniku koja se ogleda u ekstrakciji podataka iz transakcijskih baza, njihovoj transformaciji i prebacivanju u skladišta podataka. (Twain 2019.)

Nadalje, su zbog kompleksnosti SQL upita i same sintakse upitnog jezika, razvijena OLAP (*online analytical processing*) programska rješenja koja krajnjem korisniku omogućuju lakši pristup podatcima iz baze u svrhu različitih analiza. Zbog pojednostavljenja čitavog procesa analize, podatci iz skladišta podataka se organiziraju takozvane OLAP kocke koje sadrže podatke kategorizirane prema dimenzijama (klijenti, vremenski period i drugo)i podatcima iz dimenzija, što uvelike olakšava samu analizu podataka te reducira potrebno vrijeme za istu. (Rouse 2018.)

4.4 Problem s nestrukturiranim podatcima

Prema strukturi, podatci mogu biti strukturirani, nestrukturirani i polustrukturirani. Kada je riječ o analitici, najviše posla i vremena iziskuje analitika i obrada nestrukturiranih podataka. Razlog tomu je raznolikost formata podataka iz baze, počevši od tabličnih struktura, preko teksta, raznih izvještaja i weba, pa do multimedijalnih datoteka kao što su slike, zvuk ili videozapisi.



Slika 5 Podjela podataka prema strukturi

Da bi se takvi podatci analizirali, potrebno je znatno više energije, zbog kompleksnih postupaka raščlambe podatkovnih jedinica, te specijalizirani alati, namijenjeni isključivo za analitiku nestrukturiranih podataka. Bitno je napomenuti, kako je jedna od glavnih aktivnosti prilikom analize nestrukturiranih podataka, određivanje njihove buduće strukture to jest. njihova konverzija kako bi se olakšala daljnja uporaba istih te stvaranje kvalitetne informacije u svrhu donošenja odluka. (Taylor 28.03.2018.)

U svrhu napredne analitike velikih količina podataka, vodeće tvrtke na ovome području razvijaju različite procedure, softvere i standarde namijenjene analitici nestrukturiranih podataka. IBM-ov UIMA standard je istaknut kao vodeći kada je u pitanju ovakva vrsta analitike.

UIMA (engl. *Unstructured Information Management Architecture*) je ujedno OASIS[1] standard namijenjen sadržajnoj analitici, razvijen od strane IBM-a. UIM aplikacije su sustavi koji analiziraju nestrukturirane podatke, kako bi otkrili, organizirali i dostavili relevantne podatke krajnjim korisnicima.

U analitici nestrukturiranih podataka, aplikacije vrše različite vrste analiza, koristeći različite tehnike, uključujući statističke obrade i obrate podataka temeljene na prirodnom govorenom jeziku, strojno učenje i ontologiju. UIMA radni okvir podržava kreiranje, otkrivanje, kompoziciju i uvođenje širokog spektra mogućnosti analize i njihovo povezivanje sa strukturiranim informacijskim servisima i uslugama, kao što su baze podataka ili tražilice.

Veliko, ali ne i jedino područje za tekstualnu analizu je poboljšanje pretraživanja teksta. Otkrivanjem važnih pojmova i tema unutar dokumenata, semantičke tražilice pružaju

[1] ***Organization for the Advancement of Structured Information Standards*** je globalni neprofitni konzorcij koji radi na razvoju, usklađivanju i prilagodbi otvorenih standarda sigurnosti, tehnologije, interneta.

mogućnost pretraživanja pojmova i odnosa umjesto ključnih riječi, kao što je slučaj s klasičnim tražilicama gdje se ubraja i Google. Mnoge ovakve aplikacije su u mogućnosti analizirati i čitave kolekcije dokumenata s velikom preciznošću, uočavajući i bilježeći podatkovne obrasce.

Za vršenje ovakvih analiza i detekcija sadržaja odgovorna je dio sveukupne arhitekture sustava tzv. Arhitektura obrade zbirki koja dopušta praćenje tijeka podataka, od izvora do odredišta, dakle, od čitača zbirki, kroz set algoritama za analizu, sve do krajnjeg korisnika. (Bank, Schierle 2012:3482-3484)

4.5 Analitika velikih količina podataka

Analitika koja se bavi velikim količinama podataka ispituje ih kako bi se dobili skriveni obrasci, korelacije i druge informacije vezane za setove podataka. S današnjom tehnologijom, moguće je analizirati enormne količine podataka kojima raspolažu različite organizacije i odgovor dobiti gotovo trenutačno u odnosu na napor koji je potrebno uložiti kada su tradicionalne metode analitike u pitanju i stariji sustavi poslovne inteligencije. (Russom 2011:5)

Aplikacije specijalizirane za analitiku velikih količina podataka omogućuju analitičarima, podatkovnim znanstvenicima, statističarima, programerima prediktivnih modela i drugim analitičkom profesionalcima, temeljitu analizu rastućih transakcijskih baza koje uključuju različite podatkovne formate (koje tradicionalni sustavi poslovne inteligencije nisu u mogućnosti prepoznati i analizirati). Spomenuti formati uključuju mješavinu polusturkturiranih i nestrukturiranih baza, primjerice sadržaj društvenih mreža, e-mailove kupaca, strojne podatke detektirane pomoću raznih senzora te *Internet of things*[2]. (Rouse 2018.)

Nadalje, navedene tehnike i tehnologije pomažu organizacijama pri donošenju zaključaka, a samim time i poslovnih odluka. Bitno je napomenuti kako ovakva detaljna analitika pomaže organizacijama u iskorištavanju svih dostupnih podataka (gledajući na podatke kao na resurs) i prepoznavanju novih mogućnosti. (Davenport, Dyché 2013:20)

Vođeni specijaliziranim analitičkim sustavima i softverom, kao i snažnim računalnim sustavima, analitika velikih količina podataka nudi razne poslovne prednosti,uključujući nove

[2] **Internet stvari** (engl. *Internet of things*) označava povezivanje uređaja putem interneta. Uglavnom su to kućanski uređaji kao što su hladnjaci, perilice i sl.

mogućnosti zarade, efektivniji marketing, bolju uslugu klijentima, poboljšanu operativnu učinkovitost i konkurentsku prednost. (Rouse 2018.)

Važnost analitike velikih podataka može se sažeti u tri temeljne prednosti koje ovakva analitika pruža.

1. **Smanjenje troškova** – Tehnologije velikih količina podataka, kao što su Hadoop[3] ili analitika u oblaku donose značajne prednosti u vidu troškova,ali i utrošenog vremena.

2. **Brže i bolje donošenje odluka** – Koristeći analitiku u memoriji[4], u kombinaciji s analizom novih izvora podataka, organizacije mogu u vrlo kratkom vremenu analizirati podatke i donositi odluke.

3. **Novi proizvodi i usluge** - S mogućnošću mjerenja potreba kupaca i zadovoljstva, kroz analitiku dolazi moć zadovoljavanje potreba klijenata. Bitno je napomenuti da, s velikom analizom podataka, sve više tvrtki (organizacija) stvara nove proizvode koji u većoj mjeri zadovoljavaju potrebe kupaca, kroz personalizaciju proizvoda i usluga. (Davenport, Dyché 2013.)

S obzirom na prednosti koje donosi analitika velikih podataka, sve je veći broj tvrtki koje pribjegavaju iskorištavanju podataka kao resursa. Također, mnoge tvrtke specijalizirane za razvoj alata za analitiku velikih podataka, nude i specijalizirana programska rješenja ovisno o djelatnosti svojih klijenata. Prema tome, različite vrste organizacija vrše različite analize, različitim postupcima i metodama.

Primjerice, programska rješenja namijenjena analitici medicinskih podataka, koristeći umjetnu inteligenciju i *Internet of Medical Things*[5] (IoMT), otkriva se potencijal za poboljšanje brzine i učinkovitosti u svakoj fazi kliničkog istraživanja isporukom inteligentnijih, automatiziranih rješenja. Također, velike podatke iskorištavaju sustavi zdravstvene zaštite, na način da se zapisi o pacijentima, zdravstvenim planovima, zapisi o osiguranju i mnogi drugi podatci analiziraju velikom brzinom te pružatelji zdravstvenih usluga mogu pružiti dijagnoze ili mogućnosti liječenja s malim vremenskim odmakom.

---

[3] **Hadoop** je radni okvir koji omogućuje distribuiranu obradu velikih skupova podataka preko klastera računala pomoću jednostavnih modela za programiranje.

[4] **Analitika u memoriji** ili *In-memory analytics* ili je metodologija poslovne inteligencije koja se koristi u rješavanju kompleksnih i vremenski osjetljivih poslovnih scenarija.

[5] **Internet medicinskih stvari** (engl. Internet of Medical Things) je skup medicinskih uređaja i aplikacija povezanih s zdravstvenim IT sustavom kroz računalne mreže.

Kada je riječ o bankarstvu, financijske institucije, prikupljaju i analitičkim pristupom dobivaju uvid u velike skupove uglavnom nestrukturiranih podataka (e-mailovi, pozivi, podatci s društvenih mreža, podatci s provedenih anketa i razni drugi formati i izvori) kako bi donijele ispravne financijske odluke. Potrebna je velika preciznost ako su kvantitativni podatci u pitanju kako ne bi došlo do pogrešnog tumačenja istih te loših posljedica po samu organizaciju. Ovakva, napredna, analitika omogućuje takav pristup te eliminaciju preklapajućih podataka i redundantnih podataka i sustava. (Ewen 04.03.2019.)

U industriji, to jest proizvodnji dobara i usluga, borba s velikom količinom podataka je dio svakodnevice, počevši od složenih opskrbnih lanaca preko interneta stvari, do ograničenja rada i kvarova na opremi. Zato je analitika velikih podataka ključna u prerađivačkoj industriji, jer je konkurentnim organizacijama omogućila da otkriju nove mogućnosti uštede troškova i stvaranja većih prihoda.

Trgovačke organizacije analize velikih podataka koriste za poboljšanje usluga, ponude proizvoda kao i njihovu različitost. Može se reći kako je korištenjem tehnologija za analizu velikih podataka, služba za korisnike uvelike napredovala te je u mogućnosti ispuniti očekivanja kupaca. Naoružani beskrajnim količinama podataka iz programa lojalnosti kupaca, kupovnih navika i drugih izvora, trgovci ne samo da imaju dubinsko razumijevanje svojih kupaca, već mogu predvidjeti trendove, preporučiti nove proizvode i povećati profitabilnost.

Određene vladine agencije suočavaju se s velikim izazovom koji se ogleda u pooštravanju proračuna bez ugrožavanja kvalitete ili produktivnosti. Takva aktivnost osobito je zabrinjavajuća kada su u pitanju agencije i organizacije za provedbu zakona koje su u stalnoj borbi za smanjenje stope kriminala s relativno oskudnim resursima. Zbog toga mnoge agencije koriste analitiku velikih podataka pri čemu tehnologija pojednostavljuje poslovanje, a agenciji daje cjelovitiji pogled na kriminalne aktivnosti. (de Fremery 28.02.2018)

Iz prethodno opisanih slučajeva i djelatnosti u kojima se koristi analitika velikih količina podataka, može se zaključiti kako sve više organizacija pribjegava iskorištavanju podataka kao resursa te kako je u vremenu globalizacije teško ostati neprimijećen i nezabilježen u nekoj od baza, uzevši u obzir okolinu i trendove. Bitno je napomenuti kako su ljudi okruženi gomilom javno dostupnih podataka koja tek čeka da bude iskorištena na pravi način.

5. Vizualizacija – ključ razumijevanja velikih podataka

Osjetilo koje je najbrže i zauzima najveći dio čovjekove percepcije je oko, a bitno je napomenuti kako je čovjek svjestan svega 0,7% onoga što zapravo percipira u okolini. (Rosling 2006.) Oko je posebno osjetljivo na uzorke u varijaciji boje, oblika i uzorka. Kombinirajući vizualne prikaze sa kontekstom i podatcima, dobiva se nova perspektiva i mijenjaju se gledišta i percepcija okruženja pojedinca.

Primjer tome je dao Hans Rosling u svom izlaganju (Rosling 2006.) o usporedbi nataliteta, veličine obitelji i životnog vijeka čovjeka prema zemljama u kojima žive. Interaktivnom vizualizacijom, Rosling pokazuje kako su zemlje u razvoju tzv. zemlje trećeg svijeta, od 1960-ih pokazale znatan pomak kada je riječ o ovim pokazateljima.

Ono što je izlagač htio reći je da ako se ne promatra cjelovita slika to jest podatci, ljudi ne mogu primijetiti promjene u okolini, kako one ekonomske prirode, tako i one društvene. Primjerice, vizualnim prikazom korelacije među parametrima zdravstva, općenito kvalitete zdravstvenog sustava, i bruto domaćeg proizvoda po glavi stanovnika, uvodeći dimenziju vremena, može se promatrati rast i razvoj država svijeta te ih svrstati u određene grupe. No, ono što im je svima zajedničko je da se gotovo sve države kroz vrijeme kreću prema gornjem desnom kutu koordinatnog sustava. (Rosling 2006.) Na slici 6 je prikazan interaktivni grafikon, zaustavljen na 1985. godini za koju se smatra da je godina velikog gospodarskog rasta na svjetskoj razini, ali i povećanog nataliteta.(Strarr 05.05.2015.)

Slika 6 Rasti i razvoj država svijeta tijekom vremena, Izvor: https://www.gapminder.org/tools/#$stateStime$value=1985;&marker$select@$country=ind&trailStartTime=1985;&$country=rus&trailStartTime=1985;&$country=chn&trailStartTime=1985;&$country=egy&trailStartTime=1985;&$country=usa&trailStartTime=1985;;;;&chart-type=bubbles

„Fraza koja kruži u svijetu podataka, „podatci su nova nafta", generira važnost podataka u današnjem svijetu. Govori o tome kako su podatci vrsta sveprisutnog resursa koji se može iskorištavati za dobivanje inovacija i novih uvida".(McCandless 2010.)

Mccandless govori o promjeni fraze, dajući podatcima sasvim novu dimenziju, nazivajući ih novim tlom (McCandless 2010.). Također, govori kako su podatci poput novog, kreativnog medija, da naizgled svi podatci izgledaju kao hrpa beznačajnih brojeva i nepovezanih činjenica te da se pravilnom obradom podataka, mogu se pojaviti zanimljive stvari i otkriti različiti uzorci koji vrlo često nisu vidljivi kada su klasične metode analize u pitanju.

Podatke je potrebno vizualizirati i dizajnirati na način da bitni uzorci i poveznice postanu jasne te da promatraču omoguće usredotočenje na bitne informacije. Vizualiziranje podataka može dati jako brzo rješenje svakodnevnih društvenih problema te se često može dobiti jasnoću ili odgovor na jednostavno pitanje veoma brzo.

David McCandless se poziva na riječi svoga mentora "Dopusti da ti sklop podataka promijeni misaoni sklop", što svjedoči o važnosti pravilne interpretacije podataka. Bitno je napomenuti kako je za pravilnu interpretaciju podataka, za istinitu informaciju, potreban pravilan koncept obrade podataka te odgovarajući način prikaza istih. (McCandless 2010.)

Primjerice, apsolutne brojke u svijetu kakvog danas poznajemo ne daju potpunu sliku i nisu istinite koliko bi mogle biti. U tom slučaju su potrebne relativne brojke koje su povezane s ostalim podatcima da bi mogli vidjeti potpuniju sliku, što u konačnici dovodi do toga da pojedinac ispravno interpretira podatke pretvorene u informaciju i u krajnjem slučaju promijeni gledište. Također, za istraživanja i analize se najčešće koriste podatci izraženi u prosječnim vrijednostima, no prema mnogim autorima, to nije preporučljivo jer u jednom uzorku mogu postojati velike razlike u vrijednostima podataka, to jest ekstremno niske i visoke vrijednosti, što u konačnici dovodi do krive interpretacije. Prema tome, može se zaključiti kako je jedini način za razumijevanje kompleksnih podataka vizualno prikazivanje istih i korištenje relativnih vrijednosti.

Vizualizacija informacija je određena vrsta sažimanja znanja. To je način da se ogromna količina informacija i razumijevanja stisne u mali prostor te se neki uzorci u podatcima mogu vidjeti samo uz pomoć vizualizacija. Može se reći kako se vizualiziranjem podataka oni pretvaraju u krajolik koji pojedinac može istraživati svojim očima te se dobiva svojevrsna informacijska karta koja služi kao orijentir u nepreglednim bazama, ali i postaje ključni element interpretacije istih. (McCandles 2010.)

5.1 Vrste i pravila vizualiziranja

U vizualizaciji podataka velikih razmjera, mnogi istraživači koriste ekstrakciju značajki i geometrijsko modeliranje što uvelike smanjuje veličinu podataka prije stvarnog prikazivanja podataka. Odabir pravilnog prikazivanja podataka također je vrlo važan kada su u pitanju veliki podatci zbog pravilne interpretacije.

Kada je u pitanju način vizualiziranja, najčešće se koriste osnovne metode vizualnog prikaza skupova podataka, kao što su: tablice s isticanjima, histogrami, dijagram rasipanja, linijski, stupčasti, tortni grafikon, dijagram površine, toka, mjehurićasti dijagram, višestruki niz podataka, vremenski slijed, Vennov dijagram, dijagram toka podataka, odnos entiteta itd. Unatoč širokoj lepezi vrsta grafičkih prikaza, klasični prikazi ponekad nisu dovoljni za efikasnu analizu, te su podložni pogrešnoj interpretaciji.

Za naprednije analitike i što bolje iskorištavanje podataka kao resursa često se koriste manje poznati načini vizualiziranja kao što su:

- metoda paralelnih koordinata koja prikazuje multidimenzionalnost podataka, to jest koristi se za isticanje podataka u mnogim dimenzijama,
- graf površine ili karte – efektivna metoda za označavanje hijerarhije gdje površina jednog elementa označava jednu mjeru dok primjerice boja označava drugu mjeru promatranog podatka,
- stablo stošca -metoda koja prikazuje hijerarhijske podatke organizirane u tri dimenzije,
- semantičke mreže – grafička prezentacija logičkih veza između različitih koncepta.

Bitno je napomenuti kako vizualizacije mogu biti i interaktivne, što uključuje promjenu veličine, filtriranje, fokus na različite detalje itd. Za pravilnu izradu interaktivne vizualizacije potrebno je pratiti četiri osnovna koraka:

1. selekcija – označavanje potrebnih podataka,
2. povezivanje – potrebno je i korisno povezati podatke kroz različite poglede,
3. filtriranje – fokus na odabrane parametre i promatranje samo ti podataka,
4. preuređivanje i preslikavanje - preraspodjela prostornog rasporeda informacija je vrlo učinkovit u stvaranju različitih uvida. (Wang, Wang, Alexander 2015:34;35)

Da bi vizualizacija bila uspješna, to jest da bi prenijela poruku te rezultirala ispravnom interpretacijom, potrebno je pratiti određena pravila. Prvenstveno je potrebno da sam autor

zna što želi prikazati i koju informaciju treba dobiti. Isto tako, bitno je da publika kojoj je namijenjena vizualizacija bude upoznata s tematikom i prepozna vrijednost plasiranih informacija. Nadalje, potrebno je da vizualizacija bude jasna, s jasnim fontom i isticanjem te ispravnim načinom prikaza.

5.2 Alati za vizualizaciju

Većina alata za vizualizaciju velikih podataka koristi Hadoop platformu koja je učinkovita za kvalitetnu analizu, no nedostatak je manjak vizualiziranja. Noviji alati namijenjeni su u jednakoj mjeri i analitici i vizualizaciji, te su u daleko većoj prednosti u odnosu na starije inačice:

- ***Pentaho*** je dio poslovne inteligencije i podržava kreiranje dashboarda, i analitiku.
- ***Flare*** je alat za vizualizacije koji funkcionira u *Adobe Flash Playaru.*
- ***Platfora*** konvertira sirove nestrukturirane podatke u strukturirane, pogodne vizualizaciji.
- ***Many Eyes***, online je alat razvijen od strane IBM-a za kreiranje interaktivnih vizualizacija. (Wang, Wang, Alexander 2015:36).

Jedan o najistaknutijih programskih rješenja za vizualizaciju podataka, je ***Tableau***, alat namijenjen analitici velikih skupova podataka koji uključuje različite vrste vizualizacija te podržava različite formate datoteka.

Alat je prepoznat 2018. kao jedna od lidera u Gartnerovom magičnom kvadrantu koji je prikazan na slici 7. (Ajenstat 27.02.2018)

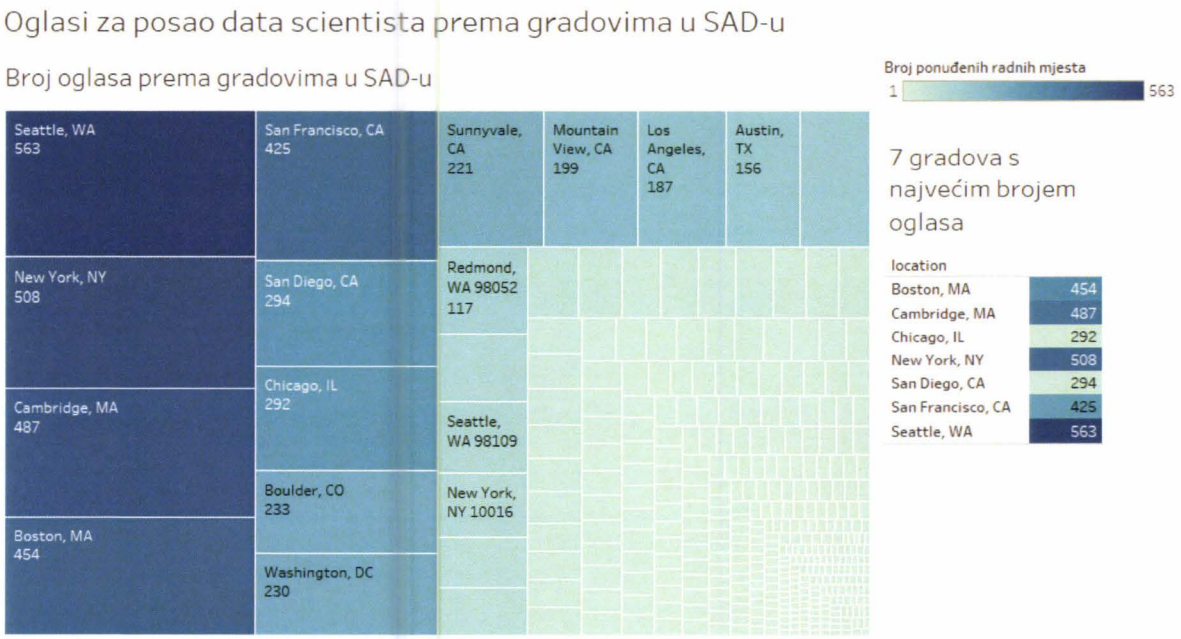
Slika 7 Gartnerov kvadrant 2018.,
Izvor: https://cdnl.tblsft.com/sites/default/files/blog/mq_2018-500_0.png

6. Praktični primjer vizualizacije velike količine podataka uz upotrebu alata *Tableau*

U ovome dijelu rada bit će prikazane vizualizacije izrađene u programu Tableau, koristeći velike strukturirane skupove poslovnih podataka xlxs formata. Podatke slične analiziranima koriste brojne organizacije za unapređenje poslovanja.

6.1 Analiza potražnje za podatkovnim znanstvenicima u kolovozu 2018. godine

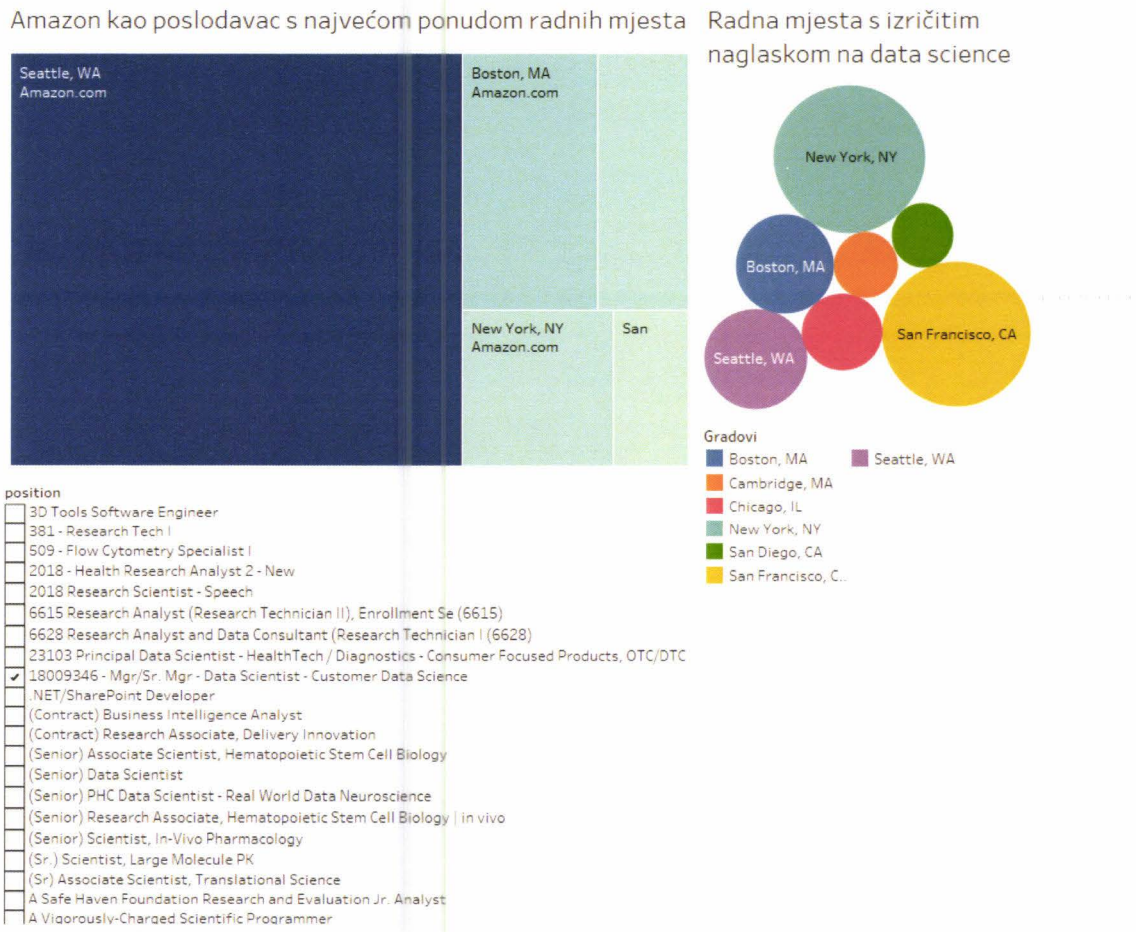
Već je spomenuto kako su za analizu velikih količina podataka potrebni podatkovni stručnjaci koji će rješavati probleme vezane za raznolikost i veličinu podataka. Iz vizualizacije je lako zaključiti i protumačiti kolika je potražnja i gdje je najveća potražnja za *data scientist*ima. Slijedi prikaz vizualizacije (slika 8).



Slika 8 Vizualizacija potražnje za data scientistima u SAD-u, Izvor: samostalni rad

Slika 8 prikazuje takozvani *dashboard* to jest kontrolnu ploču s dvije vizualizacije u jednom slikovnim prikazu koja je kreirana kako bi podatci bili što razumljiviji. S velikom lakoćom se mogu interpretirati podatci prikazani površinskim grafikonom koji je istovremeno obilježen i bojama koje obilježavaju stupnjevitu kvantitativnu promjenu. Nadalje, isti ti podatci su prikazani u tabličnom formatu te su također gradijentno osjenčani i filtrirani samo gradovi s najvećom potražnjom poslova. Ovaj filter bit će korišten i u daljnjoj analizi zbog praktičnosti i preglednosti vizualizacija.

Svaki oglas za posao ima i kategoriju opisa posla gdje točno naglašava kakve sposobnosti eventualni zaposlenik treba posjedovati. Vizualizacija koja prikazuje potražnju za poslovima koji u svom opisu sadrže izričiti naglasak na *data science* je prikazana slikom 9, je te tom *dashboardu* dodana i vizualizacija najvećeg poslodavca i lokacije kojoj je eventualni radnik potreban.

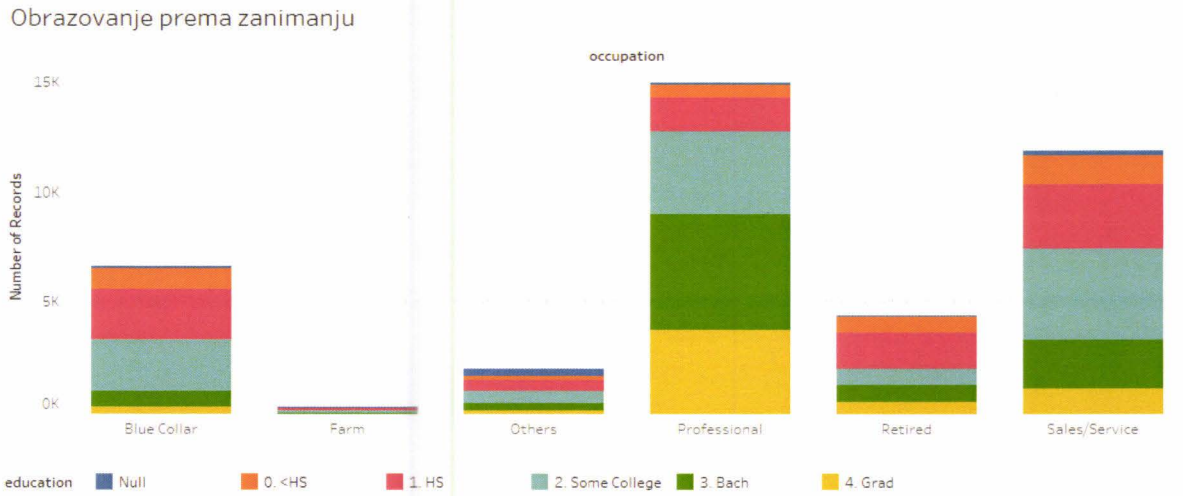


Slika 9 Detaljnija vizualizacija potražnje za data scientistima u SAD-u, Izvor: samostalni rad

Iz vizualnog prikaza je jasno vidljivo u kojim gradovima se izričito traže *data scientisti* te kako je Amazon najveći poslodavac s lokacijom u Seattleu. Program Tableau omogućava korisniku laku manipulaciju nad vizualizacijom u vidu promjene filtera i dobivanja željenih podataka.

6.2 Analiza informacija o kupcima

Kako bi unaprijedile svoje poslovanje, mnoge tvrtke prikupljaju podatke o svojim kupcima i klijentima. U drugom primjeru analize skupa podataka prikazani su upravo podatci klijenata jedne trgovine.



Obrazovanje prema zanimanju

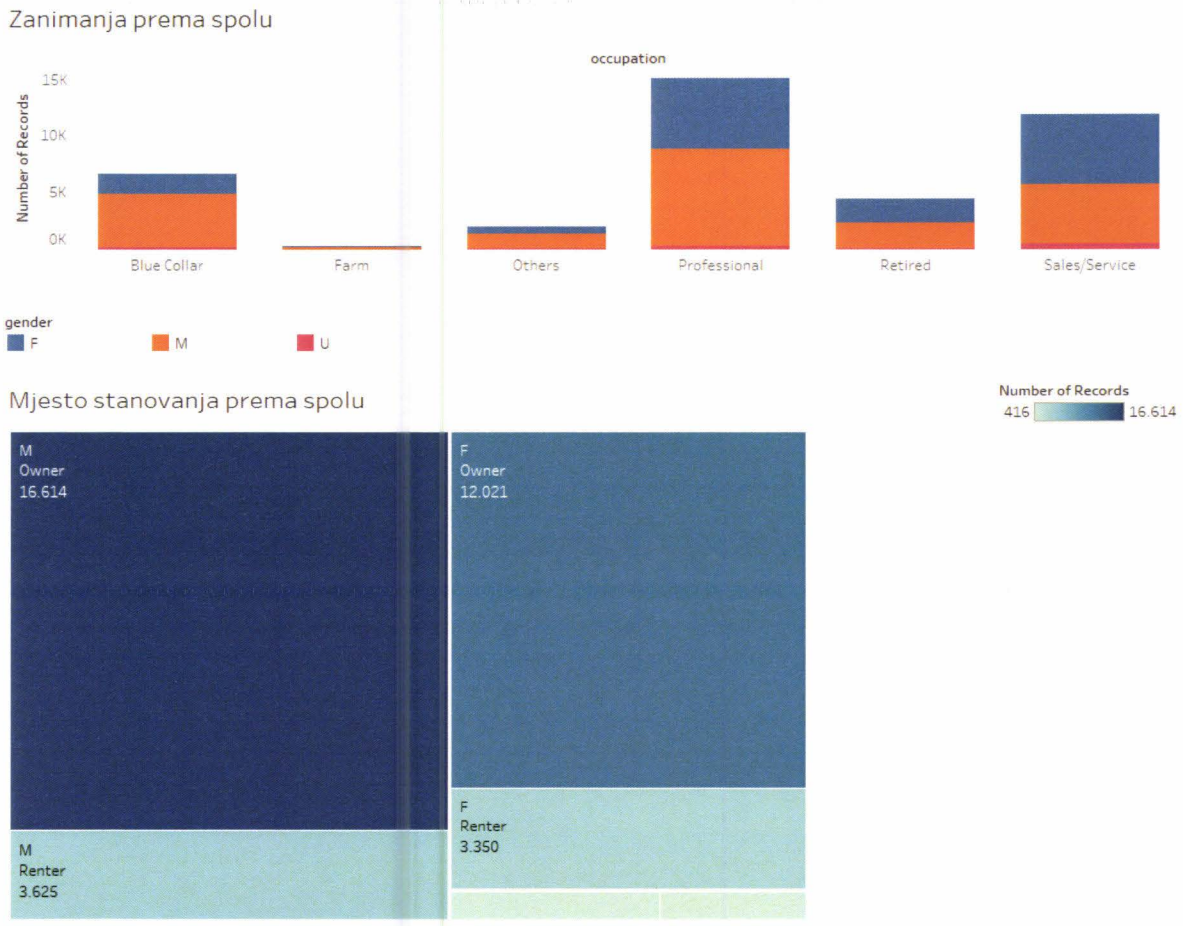
Broj zapisa
5 5.342

education	occupation: Blue Collar	Farm	Others	Professi..	Retired	Sales/Se..
Null	93	5	300	116	33	194
0. <HS	984	77	209	584	746	1.248
1. HS	2.262	75	441	1.580	1.572	2.898
2. Some College	2.231	109	568	3.626	682	4.184
3. Bach	719	48	305	5.342	755	2.098
4. Grad	332	15	183	3.688	553	1.145

Slika 10 Vizualizacija podataka o kupcima,
Izvor: samostalni rad

Na ovom primjeru je vidljivo kako zanimanje kupaca ove trgovine uglavnom odgovora njihovu obrazovanju te kako najveći broj kupaca ima visoko obrazovanje te se bavi profesionalnim djelatnostima.

Nadalje, podatci koji su vrlo zanimljivi su oni koji se tiču podjele prema spolovima, počevši od obrazovanja, do posjedovanja nekretnine, što je prikazano slikom 11 na kojoj je vidljivo kako je općenito veći broj muškaraca u svim poljima obrazovanja i te kako isti ti muškarci posjeduju neku nekretninu, a približno jednak broj i muškaraca i žena živi u iznajmljenom stanu ili kući.



Slika 11 Vizualizacija podataka o kupcima prema spolu,
Izvor: samostalni rad

7. Rasprava

U radu je pobliže objašnjen problem velikih količina podataka i poteškoće na koje nailaze menadžeri kada je njihova analiza u pitanju. Podatkovna znanost daje odgovore na mnoge izazove koji stoje pred analitičarima i znanstvenicima.

Problem kod kojeg se nailazi kada su podatci u pitanju je u prvom redu kognitivne prirode, budući da je prije kreiranja same vizualizacije, ona apstraktna te korisnik prvenstveno mora znati što želi dobiti kako bi vizualizacija bila uspješna.

Alati za rad s velikim količinama podataka su napredni i imaju brojne mogućnosti kada su vizualizacije u pitanju. Konkretno, Tableau ima ponuđen širok spektar raznih vizualizacija i pogodnosti za korisnika kada je analiza u pitanju. Također, vrlo je jednostavan za korištenje i rezultati koje daje su u vrlo kratkom roku interpretabilni, može se reći i trenutno. Program također sadrži mogućnost kreiranja prezentacija od već kreiranih vizualnih prikaza podataka te je kao takav u velikoj prednosti u odnosu na druge klasične programe za vizualizaciju i analizu kao što je *Microsoft Excel*.

Isto tako, Tableau prepoznaje različite podatkovne formate te je u mogućnosti prikazati puno bolji i širi spektar podataka. Osim toga, Tableau ima ugrađen algoritam za prepoznavanje geografskih širina i dužina i u mogućnosti je prikazati podatke putem geografske karte ako je to potrebno. Iako Tableau posjeduje takav algoritam, nije uvijek u mogućnosti prikazati podatke na taj način. Tako da, na tom polju ima mjesta za poboljšanje alata. Isto tako, program ponekad nije u mogućnosti prepoznati nazive stupaca u bazi te ih šifrira. Podatci tada ne mogu biti interpretirani i analizirani budući da im nije dodijeljena dimenzija preko koje će biti razlikovani.

8. Zaključak

Danas, u vremenu sve većeg i bržeg razvoja informacijsko komunikacijskih tehnologija, stvaraju se velike količine podataka čija analiza predstavlja nužnost kako bi podatci bili analizirani i upotrijebljeni na pravi način. Vizualizacija istih tih podataka postala je sastavni dio analitike. Razlog tomu je nepreglednost čistih brojki te njihova otežana interpretacija koja oduzima vrijeme i novac. U poslovnom svijetu analize se vrše gotovo na dnevnoj razini, a ako se uzme u obzir sama količina i kompleksnost podataka, vrijeme je uistinu bitan čimbenik u analizi.

Bitno je napomenuti kako veliki podatci nisu kompleksni samo zbog svoje veličine to jest obujma te se osvrnuti na 3V (*volume, variety, velocity*), dakle obujam, raznolikost i brzinu kojom se podatci i baze mijenjaju. Također kompleksnosti analiza doprinosi i struktura skupova podataka i pitanje kako na najbrži mogući način analizirati nestrukturirane podatke bez gubitka njihove kvalitete. Upravo zbog kompleksnosti samih podataka i njihove nesagledivosti, bitno je primijeniti tehniku vizualiziranja kako bi se podatci tranformirali u informacije.

Podatkovna znanost je dala programska rješenja za analitiku različitih podatkovnih skupova i formata. Među suvremene alate za vizualizaciju podataka ubraja se Tableau koji je prepoznat kao jedan od lidera prema Garterovom magičnom kvadrantu. Tableau nudi brojne mogućnosti korisniku, počevši od prepoznavanja različitih podatkovnih formata, preko široke palete mogućih grafičkih prikaza pa do manipulacija nad kreiranim vizualizacijama. Može se reći kako je Tableau vrlo jednostavan alat i kao takav je poželjan dio sustava za potporu odlučivanju, a samim time i poslovne inteligencije bez koje je moderno poslovanje nezamislivo.

LITERATURA

1. Ajenstat, Francois. Tableau named a leader in the Gartner Magic Quadrant for six years in a row. 27.02.2018. dostupno na: https://www.tableau.com/about/blog/2018/2/tableau-named-leader-gartner-magic-quadrant-six-years-row-82534 [pristupljeno 11.06.2019.]
2. Bank, Mathias. Schierle, Martin. A Survey of Text Mining Architectures and the UIMA Standard. 2012. dostupno na: http://www.lrec-conf.org/proceedings/lrec2012/pdf/183_Paper.pdf [pristupljeno 03.07.2019.]
3. Bellinger, Gene. Castro, Durval_. Mills, Anthony. Data, Information, Knowledge, and Wisdom. 2004. dostupno na: http://www.systems-thinking.org/dikw/dikw.htm [pristupljeno 05.20.2019.]
4. Berman, Jules J. Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information. 2013. dostupno na: https://books.google.hr/books?id=gEho0DI8a2kC&pg=PA52&lpg=PA52&dq=introspection+in+big+data&source=bl&ots=OaUyBRNn2_&sig=ACfU3U0HnDN77wOVEf-GklHsOxPVvgr2lQ&hl=hr&sa=X&ved=2ahUKEwjL_oqs1v_iAhUGpIsKHeTNChMQ6AEwEXoECAkQAQ#v=onepage&q&f=false [pristupljeno 05.06.2019.]
5. Bosančić, Boris. DIKW hijerarhija: za i protiv. 2017. dostupno na:https://hrcak.srce.hr/195861[pristupljeno 02.06.2019.]
6. Ćurko, Katarina. Skladište podataka – sustav za potporu odlučivanju. 2001. dostupno na: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=45126 [pristupljeno 05.05.2019.]
7. Dan, Woods. James Dixon Imagines A Data Lake That Matters. 26.01.2015. dostupno na: https://www.forbes.com/sites/danwoods/2015/01/26/james-dixon-imagines-a-data-lake-that-matters/#31c409f84fdb [pristupljeno 08.05.2019.]
8. Data science Croatia. 11.11.2015. dostupno na: http://datascience.com.hr/2015/11/11/organiziramo-prvi-meetup-na-temu-znanosti-o-podacima/ [pristupljeno 05.05.2019.]
9. Davenport, Thomas H., Dyche, Jill. Big data in big companies. 2013. dostupno na: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/bigdata-bigcompanies-106461.pdf [pristupljeno 10.06.2019.]
10. David McCandless, The beauty of dana visualization. 2010. dostupno na: https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization#t-40488 [pristupljeno 08.05.2019.]
11. de Fremery, Rose. Big Data and Government: How the Public Sector Leverages Data Insights. 28.02.2018. dostupno na: https://hortonworks.com/article/big-data-and-government-how-the-public-sector-leverages-data-insights/ [pristupljeno 03.07.2019.]
12. Ewen, James. How Big Data Is Changing The Finance Industry. 04.03.2019. dostupno na: https://www.tamoco.com/blog/big-data-finance-industry-analytics/ [pristupljeno 03.07.2019.]
13. Fayyad, Usama. Grinstein, Georges G. Wierse, Andreas. Information visualization in data mining and knowledge discovery. 2002. dostupno na: https://books.google.hr/books?hl=hr&lr=&id=rYFvnyPRwkgC&oi=fnd&pg=PR5&dq=data+visualization&ots=6eBEkpOwI3&sig=wi-BTbi4iFKvC8crHNJtp66TUDI&redir_esc=y#v=onepage&q=data%20visualization&f=false [pristupljeno 03.07.2019.]
14. Foote, D. Keith. A brief history of data science. 14.12.2016. dostupno na: https://www.dataversity.net/brief-history-data-science/# [pristupljeno 07.05.2019.]

15. Frančula, Nedjeljko. Digitalna Zemlja i veliki podatci. 2017. dostupno na: https://bib.irb.hr/datoteka/877613.Digitalna_Zemlja_i_veliki_podaci.pdf [pristupljeno 15.06.2019.]
16. Hans Rosling, The best stats you've ever seen, 2006. dostupno na: https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen [pristupljeno 08.05.2019.]
17. Jacobson, Ralph. 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?. 24.04.2013. dostupno na: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/ [pristupljeno 08.05.2019.]
18. Kocijan, Kristina. "Big Data: kako smo došli do Velikih podataka i kamo nas oni vode." (2014). dostupno na: http://darhiv.ffzg.unizg.hr/id/eprint/5064 [pristupljeno 15.05.2019.]
19. Lauro, Carlo. Computational statistics or statistical computing, is that the question?. 1996. dostupno na: http://www.mat.ufrgs.br/~viali/estatistica/mat2274/material/textos/1-s2.0-0167947396889201-main.pdf [pristupljeno 05.05.2019.]
20. Manger, Robert. Baze podataka. 2003. dostupno na: http://jadran.izor.hr/~dadic/EKO/baze-podataka.pdf[pristupljeno 11.06.2019.]
21. McAfee, Andrew. Brynjolfsson, Erik. Big data: The management revolution. 2012. dostupno na: http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf [pristupljeno 03.07.2019.]
22. Mekterović, Igor. Brkić, Ljiljana. Skladišta podataka i poslovna inteligencija. 2017. dostupno na: https://www.fer.unizg.hr/_download/repository/SKRIPTA_-_Skladista_podataka_i_poslovna_inteligencija.pdf [pristupljeno 23.05.2019.]
23. Mujadžević, Edin. Uvod u SQL. 2016. dostupno na: https://www.srce.unizg.hr/files/srce/docs/edu/osnovni-tecajevi/d301_polaznik.pdf [pristupljeno 01.06.2019.]
24. Panian, Željko. Poslovna inteligencija Studije slučajeva iz hrvatske prakse. 2007. dostupno na: https://bib.irb.hr/datoteka/481181.PISSHP_-_Glavnina_teksta.pdf [pristupljeno 30.05.2019.]
25. Poropat, Goran. Marušić, Martina. Štimac, Davor. Sustavno pretraživanje baza podataka. 2017. dostupno na: https://urn.nsk.hr/urn:nbn:hr:184:180292 [pristupljeno 10.06.2019.]
26. Press, Gil. A very short history of data science. 28.05.2013. dostupno na: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#7131adc555cf[pristupljeno 05.05.2019.]
27. Rouse, Margaret. Big data analytics. 2018. dostupno na: https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics [pristupljeno 15.05.2019.]
28. Rouse, Margaret. OLAP (online analytical processing).2018. dostupno na: https://searchdatamanagement.techtarget.com/definition/OLAP [pristupljeno 15.06.2019.]
29. Rouse, Margaret. What is data visualization. 2017. dostupno na: https://searchbusinessanalytics.techtarget.com/definition/data-visualization [pristupljeno 03.06.2019.]
30. Russom, Philip. Big data analytics. 2011. dostupno na: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf [pristupljeno 15.05.2019.]
31. Schutt, Rachel. O'Neil, Cathy. Doing data science. 2014. dostupno na: https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html[pristupljeno 03.07.2019.]
32. Taylor, Christine. Structured vs. unstructured data. 28.03.2018. dostupno na: https://www.datamation.com/big-data/structured-vs-unstructured-data.html [pristupljeno 12.05.2019.]

33. Taylor, Twain. 5 modern ETL tools for microservices data integration.2019. dostupno na: https://searchmicroservices.techtarget.com/feature/5-modern-ETL-tools-for-microservices-data-integration [pristupljeno 15.06.2019.]
34. Voulgaris, Zacharias. Data scientist : the definitive guide to becoming a data scientist. 2014. dostupno na: https://www.worldcat.org/title/data-scientist-the-definitive-guide-to-becoming-a-data-scientist/oclc/881183082/viewport [pristupljeno 08.05.2019.]
35. Wang, Lidong, Guanghui Wang, and Cheryl Ann Alexander.;Big data and visualization: methods, challenges and technology progress. 2015. dostupno na: https://pdfs.semanticscholar.org/2975/4e4295a9ce4d51937c0712d6482634474628.pdf [pristupljeno 07.05.2019.]
36. Tableau. Dostupno na: https://www.tableau.com/ [pristupljeno 05.05.2019.]

POPIS SLIKA