UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Nina Cvetko

Pristopi k analizi nestrukturiranih velikih podatkov (Big Data)

Diplomsko delo

Ljubljana, 2017

UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Nina Cvetko

Mentor: doc. dr. Luka Kronegger

Pristopi k analizi nestrukturiranih velikih podatkov (Big Data)

Diplomsko delo

Ljubljana, 2017

ZAHVALA

Diplomsko delo je posvečeno moji hčerki Evelini. Hvala ti za inspiracijo in za vse prespane noči.

Iskrene zahvale gredo tudi moji družini, partnerju in prijateljem za motivacijo in podporo med študijem.

Za vso pomoč in nasvete se zahvaljujem tudi mentorju doc. dr. Luki Kroneggerju in intervjuvancem, ki so s svojim prispevkom dodali vrednost diplomskemu delu.

Pristopi k analizi nestrukturiranih velikih podatkov

Izraz »Big Data« v originalu pomeni veliko količino podatkov, ki je ni možno učinkovito procesirati z običajnimi metodami in orodji. Namen diplomskega dela je spoznati bolj specifičen del velikih podatkov – nestrukturirane podatke in spoznati možne pristope k uporabi in analizi tovrstnih podatkov. Na primeru Instituta Jožefa Stefana je prikazano, kako lahko organizacija pristopa k analizi nestrukturiranih podatkov, katere faze v procesu se jim zdijo najbolj zahtevne in kateri so največji izzivi, s katerimi se pri delu z nestrukturiranimi velikimi podatki srečujejo. Ker so poleg uporabe nestrukturiranih velikih podatkov zanimivi tudi vzroki za njihovo redko uporabo, sta v primerjavi s pristopom Instituta Jožefa Stefana predstavljena tudi alternativna raziskovalna pristopa v dveh večjih mednarodnih podjetjih. Izsledki kažejo, da so glavni razlogi za redko uporabo nestrukturiranih velikih podatkov predvsem v tem, da potencialni uporabniki (še) ne vidijo potrebe po opravljanju tega tipa analiz oziroma se jim zdi razmerje med investicijo in vrednostjo nesprejemljivo. S poenostavljanjem pristopov in programskih orodij v prihodnosti bo tovrstna analiza lahko postala bolj prijazna za širši krog potencialnih uporabnikov.

Ključne besede: veliki podatki, nestrukturirani podatki, analiza podatkov, proces analize.

Approaches to Unstructured Big Data Analysis

»Big Data« means great volumes of data, which can not be processed by regular methods and tools. The aim of the thesis is to present a specific part of Big Data – unstructured data and to recognise and determine possible usability and analysis approaches. In the case of Institut Jožefa Stefana we can see how this organisation approaches unstructured data analysis, which phase in the procedure is more difficult for them and what are the greatest challenges they have to face while analysing unstructured data. Because I was also interested in reasons for rare usage of unstructured data I included a comparison of approaches of Institut Jožefa Stefana and alternative approaches of two multinational companies. Findings suggest that the reasons for rare usage are mostly in potential users not recognising any need for implementing this type of analysis. In their opinion the ratio between investment and the value is unacceptable. By simplifying approaches and programmer tools in the future the analysis can become much more usable for a wider range of users.

Key words: Big Data, unstructured data, data analysis, analysis process.

KAZALO VSEBINE

KAZALO SLIK

1 UVOD

Izraz Veliki podatki (Big Data) je v zadnjih letih postal zelo priljubljen, vendar zanj še nista vzpostavljena enotna definicija ali koncept. Po navadi je kot ena izmed lastnosti velikih podatkov navedena 'raznolikost', ki velike podatke deli na strukturirane in nestrukturirane (v nekaterih virih tudi polstrukturirane) podatke. Večinski delež velikih podatkov predstavljajo nestrukturirani podatki, ki so kljub težavnosti za zbiranje in analizo lahko pravi rudnik informacij za tiste, ki imajo znanje in finančne ter tehnične zmožnosti za izkoristek njihovega potenciala.

Namen diplomske naloge je spoznati področje velikih nestrukturiranih podatkov in spoznati možne pristope k uporabi in analizi tovrstnih podatkov. Za predstavitev teoretičnega ozadja sem proučila različne pisne vire, s pomočjo katerih sem oblikovala združen koncept nestrukturiranih velikih podatkov in jih umestila v širšo sliko velikih podatkov. Povzela sem prednosti in izzive pri uporabi nestrukturiranih podatkov, v nadaljevanju pa sem se osredotočila na področje analize nestrukturiranih podatkov in podrobneje predstavila arhitekturo analize, analize različnih spletnih virov, proces analize nestrukturiranih podatkov ter izzive, ki se lahko ob analizi pojavljajo.

Pristop k analizi nestrukturiranih velikih podatkov sem želela raziskati tudi s pomočjo lastne raziskave, zato sem opravila intervju z zaposlenim na Institutu Jožefa Stefana, ki tovrstne podatke analizira. Z raziskavo sem želela ugotoviti, na kakšen način organizacija pristopa k analizi nestrukturiranih podatkov. Podrobneje me je zanimalo, kako pri njih poteka postopek analize podatkov, katere faze v procesu so najbolj zahtevne in zamudne ter kateri so največji izzivi, s katerimi se srečujejo pri delu z nestrukturiranimi velikimi podatki. Zanimalo me je tudi njihovo strokovno mnenje o vzrokih za redkost uporabe teh podatkov in kakšne so njihove napovedi za prihodnost na področju analize podatkov.

Kot rečeno je analiza nestrukturiranih velikih podatkov na slovenskem trgu za zdaj še precej redka, zaradi česar se pojavi naslednje raziskovalno vprašanje, in sicer »Zakaj je kljub poslovnim in raziskovalnim koristim, ki jih nestrukturirani veliki podatki lahko prinesejo, zanimanje za ta tip raziskav na slovenskem trgu tako majhno?«.

Predpostavljam, da so glavni vzroki v specifičnosti znanj, zaradi česar je za analizo nujno potreben primerno izobražen kader, ki za osnovne analize klasičnih baz podatkov ni nujen pogoj. Zelo verjetno ima tako vpliv tudi težavnost analize in zbiranja tovrstnih podatkov.

Ker so me zanimali tudi vzroki za neuporabo nestrukturiranih velikih podatkov, sem intervjuvala tudi zaposlena v večjih mednarodnih podjetjih, ki bi glede na naravo svojega dela lahko imela interes za delo s takšnimi podatki, vendar jih ne uporabljajo. Na podlagi njunih odgovorov sem predstavila še alternativna pristopa – analizo običajnih nestrukturiranih podatkov in analizo bolj strukturiranih velikih podatkov, ki sta tudi pri nas v praksi pogostejši in bolj vpeljani, ter njune značilnosti primerjala s pristopom k analizi nestrukturiranih velikih podatkov.

2 VELIKI PODATKI (BIG DATA)

»Veliki podatki« (»Big Data«) so relativno nov koncept, za katerega kljub široki uporabi termina ne obstaja enotna, splošno sprejeta definicija. Lahko se nanaša na družbeni fenomen, baze podatkov, tehnike analiziranja, tehnologije shranjevanja podatkov, procese, infrastrukture in etične dileme (De Mauro 2014; Japec in drugi 2015).

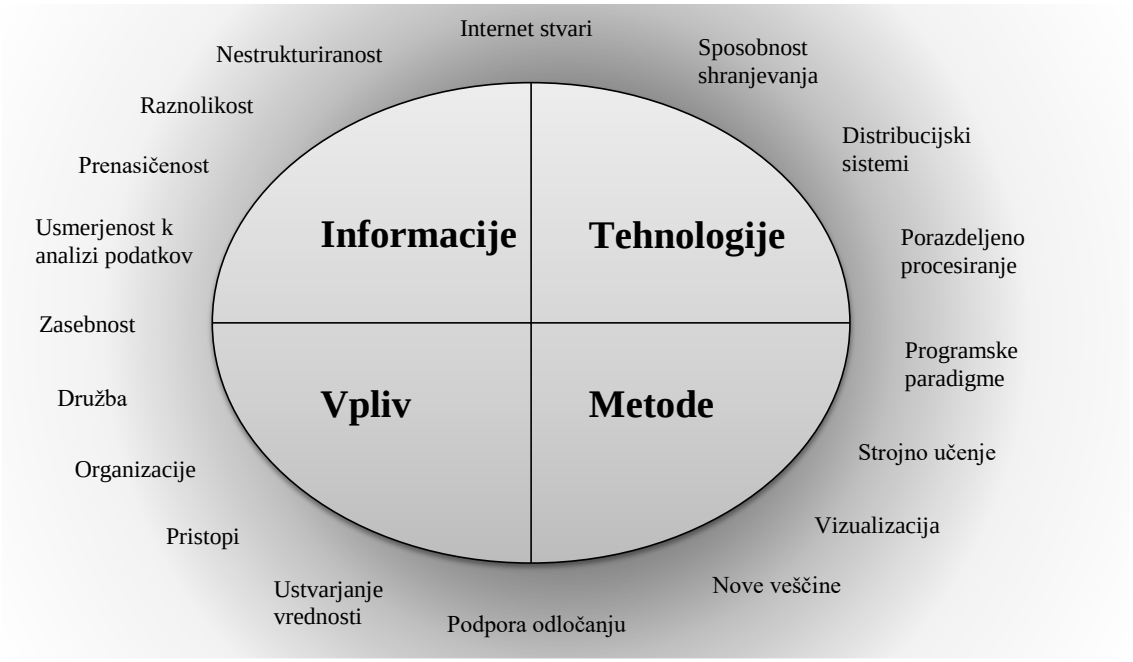
Najbolj priznana in široko uporabljena je opredelitev velikih podatkov, ki jo je leta 2001 podal Laney (2001). V njej sicer ne omenja dobesedno velikih podatkov, ampak govori o tridimenzionalni rasti podatkov glede na velikost, raznolikost in hitrost, kar so pozneje poimenovali »3 V modek«.

»V originalu je izraz »Big Data« pomenil veliko količino podatkov, ki je ni možno učinkovito procesirati z običajnimi metodami in orodji« (Kaisler in drugi 2013, 995). Veliki podatki izvirajo s področja fizike in astronomije, kjer so zaradi značilnosti podatkov začeli prvi uporabljati tehnike današnjega koncepta Big Data. V zadnjem času se te tehnike uporabljajo tudi za analize ekonomskih in socialnih sistemov, kjer so bile prej v ospredju anketne raziskave, eksperimenti in etnografije (Japec in drugi 2015).

De Mauro in drugi (2014) so z analiziranjem pojavljanja izraza »Big Data« identificirali štiri ključne teme, na katere se veliki podatki lahko nanašajo:

- Informacije: v kolikšni meri so lahko informacije ustvarjene in dostopne.
- Tehnologije: veliki podatki se v literaturi pogosto povezujejo s tehnologijo, ki glede na velike količine podatkov in kompleksnost operacij omogoča njihovo uporabo.
- Metode: za analizo nestrukturiranih velikih podatkov niso več dovolj tradicionalne statistične tehnike in so potrebne zahtevnejše metode procesiranja.
- Vpliv: predstavlja vpliv velikih podatkov na družbo, organizacije in podjetja.

Slika 2.1: Štiri ključne teme velikih podatkov



Vir: povzeto po De Mauro in drugi (2014).

Večina definicij velikih podatkov se sicer nanaša na tehnološke pristope in tehnološke zahteve. Za procesiranje velikih podatkov je namreč potrebna velika računalniška moč in arhitektura za shranjevanje, manipulacijo in analizo podatkov, s katero potem lahko izvlečemo vrednost iz širokega spektra različnih podatkov (Gantz in Reinsel 2011; De Mauro in drugi 2014).

2.1 VIRI VELIKIH PODATKOV

Velike podatke lahko glede na izvor ločimo na notranje (tiste, ki nastajajo znotraj organizacije) in zunanje (tiste, ki jih organizacija lahko pridobi za analizo) (Baesens in drugi 2016). Sprva so včasih namenjeni za drugo uporabo (kot se pozneje uporabijo) ali nastajajo celo kot stranski produkt nastajanja drugih podatkov. Takšni podatki so sekundarni, nereaktivni in nastajajo nenačrtovano, kot taki pa so še posebej uporabni npr. za merjenje mnenj in obnašanja na družbenih omrežjih (Japec in drugi 2015).

Obstaja več vrst virov velikih podatkov (Aggarwal 2013; Kaisler in drugi 2013; Katal in drugi 2013; Japec in drugi 2015):

- Družbena omrežja: z analizo mnenj posameznikov na družbenih omrežjih podjetja pridobijo mnenja (potencialnih) strank, kar lahko uporabijo za načrtovanje strategij in večanje dobička.
- Osebni podatki: primer so podatki sledilnih naprav.
- Senzorični podatki: nastajajo pri zaznavanju lokacije preko mobilnega telefona in GPS podatkov, pri okoljskih pristopih za zaznavanje vremena in spremljanje stopnje onesnaženosti okolja in različnih vojaških pristopih za zaznavanje nenavadnih aktivnosti.
- Transakcijski podatki: sem spadajo podatki o finančnih transakcijah – plačilih, nakazilih in dvigih. Ti tipi podatkov so predvsem uporabni za banke, večje trgovine in spletna družbena omrežja (kjer lahko uporabniki plačujejo za oglaševanje, igranje določenih igric ipd.).

  S transakcijskimi podatki lahko ciljamo tudi na podatke, ki so vključeni v procese transakcije na spletu, kot je število klikov na povezave ali potek prenašanja informacij (npr. deljenje povezave) med uporabniki spleta.
- Administrativni podatki: so bolj strukturirani in definirani kot podatki iz drugih virov.

2.2 LASTNOSTI VELIKIH PODATKOV

V literaturi (Beal 2017; Japec in drugi 2015; Kaisler in drugi 2013; Russom 2011) lahko zasledimo naslednje lastnosti velikih podatkov (3V model):

- Velika količina podatkov (Volume): nanaša se na količino podatkov, razpoložljivih za analizo, ki jo lahko merimo v:
  - bajtih (po navadi terabajtih, petabajtih),
  - času (npr. nekatera podjetja shranjujejo za toliko let podatkov, kolikor je določeno z zakonom),
  - s številom datotek/tabel,
  - številom transakcij.

  Količina zelo hitro narašča zaradi več orodij, kjer nastajajo podatki (mobilne aplikacije, senzorji, družbena omrežja), in zaradi boljših možnosti shranjevanja in

prenašanja teh podatkov. Z naraščanjem količin podatkov upada njihova vrednost, skladno z njihovo starostjo, tipom, kvantiteto ipd.

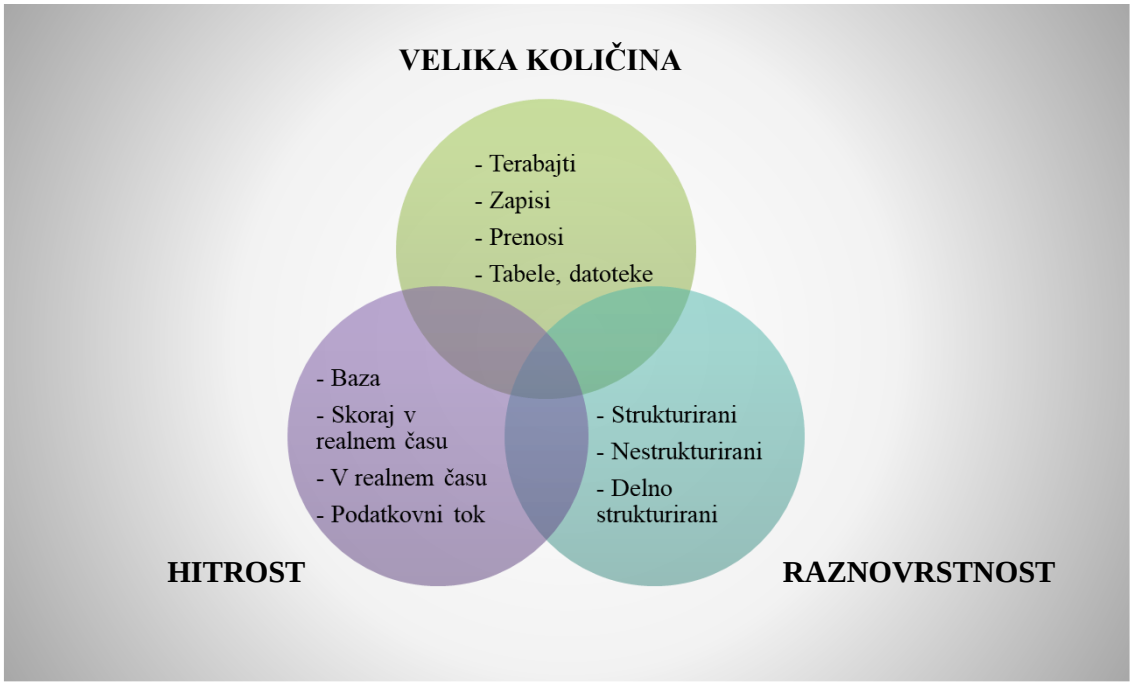
- Hitrost (Velocity): gre za merjenje hitrosti ustvarjanja, prenašanja in zbiranja podatkov. Podatke lahko glede na hitrost ločimo na:
    - bazo,
    - skoraj sočasne (near time),
    - sočasne (real time),
    - tok (streams).

  Zaradi hitrega naraščanja količin podatkov je potreben izjemno hiter dostop, podatki se namreč lahko glede na potrebe dodajajo v baze v različnih časovnih intervalih (lahko recimo na vsako uro ali tudi do nekaj tisoč na sekundo). Pogosto je pri obdelavi velikih podatkov zahtevana takojšnja razpoložljivost podatkov (»real time«) ali tok (stream) za nadaljnjo analizo, ki mora biti pogosto prav tako izvršena v najkrajšem možnem času za omogočanje hitre podpore odločitvam.

- Raznovrstnost (Variety): »se nanaša na kompleksnost formatov, v kakršnih obstajajo veliki podatki« (Japec in drugi 2015, 8). Ločimo strukturirane, nestrukturirane in delno strukturirane velike podatke. Strukturirani podatki zavzemajo fiksno mesto znotraj datoteke ali zapisa. Ker jih lahko razvrstimo v relacijske baze, omogočajo hitrejšo in lažjo analizo kot druge oblike velikih podatkov (Taylor 2017).

  Nestrukturirani podatki prihajajo iz različnih virov v različnih formatih (besedilo, zvok, slika, videoposnetki ipd.), zaradi česar je strukturiranje in povezovanje teh podatkov precej zahtevno in predstavlja enega izmed največjih izzivov pri analizi velikih podatkov (Taylor 2017).

Slika 2.2: Model 3 V



Vir: povzeto po Russom (2011).

Nekateri avtorji (De Mauro in drugi 2014; Japec in drugi 2015; Kaisler in drugi 2013; Katal in drugi 2013) so model 3V kasneje dopolnili še z drugimi karakteristikami, kot so:

- nihanje (Variability): nekonsistentnost toka podatkov skozi čas;
- vrednost (Value): meri uporabnost podatkov pri (poslovnih) odločitvah;
- resničnost (Veracity): zaupanje v točnost podatkov;
- kompleksnost (Complexity): povezovanje različnih virov, čiščenje in pretvarjanje podatkov in
- nestrukturiranost (Unstructuredness) (Intel in Suthaharan v De Mauro in drugi 2014).

3 NESTRUKTURIRANI VELIKI PODATKI

»Nestrukturiranost podatka se nanaša na informacijo, ki nima vnaprej določenega podatkovnega modela oziroma ne ustreza relacijskim tabelam« (Bakshi 2012, 1). Nestrukturirani podatki (Bakshi 2012; Japec in drugi 2015; Taylor 2017) so lahko proizvedeni s strani človeka:

- besedilne datoteke,
- videoposnetki,
- fotografije,
- zvočne datoteke,
- e-pošta,
- objave na družbenih omrežjih,
- sporočila na mobilnih telefonih.

Drugi nestrukturirani podatki pa nastanejo s pomočjo naprav:

- senzorski podatki,
- znanstveni podatki,
- digitalni nadzor,
- telemetrija.

Nestrukturirani podatki so heterogeni in spremenljivi, njihova količina pa se povečuje hitreje kot pri strukturiranih velikih podatkih. Njihova prednost se kaže v cenejšem (elektronskem) zbiranju podatkov, saj so izvedbe tradicionalnih tipov raziskav lahko precej drage in zamudne. Nasprotno pa se stroški precej povišajo pri fazi čiščenja in procesiranja nestrukturiranih podatkov, prav tako se v tej fazi pojavi potreba po spremembi in povečanju kadra ter s tem povečanje stroškov (Das in Kumar 2013; Japec in drugi 2015). Čeprav so nestrukturirani veliki podatki zaradi svojih značilnosti zahtevni za analizo, pa ponujajo priložnost za raziskovanja in opazovanja, ki jih drugače ne bi mogli izvesti.

4 UPORABA NESTRUKTURIRANIH VELIKIH PODATKOV

4.1 PREDNOSTI PRI DELU Z NESTRUKTURIRANIMI VELIKIMI PODATKI

Veliki podatki se med drugim lahko uporabljajo za namene pisanja zakonov in posledično zmanjšanje kriminala, policija pa lahko z uporabo lastne baze podatkov in javnih baz bolj učinkovito posreduje, kjer je to potrebno. Uporabni so tudi za izboljševanje zdravstvenih pripomočkov in analizo medicinskih zapisov in slik za zaznavanje vzorcev, s katerimi lahko bolezen zaznajo prej in učinkoviteje razvijajo zdravila in načine zdravljenja (Marr 2017).

Analiza senzorskih velikih podatkov omogoča predvidevanje in hitrejše reagiranje na naravne katastrofe. Primer je predvidevanje potresov, sledenje premikov večjih neviht ali tornadov. NASA velike podatke uporablja tudi za analizo podatkov o vesolju in odkrivanje novih planetov (Marr 2017).

Nestrukturirani veliki podatki nastajajo tudi tam, kjer si posamezniki olajšajo vsakodnevno življenje z omogočanjem spletnega nakupovanja ali lažjega načrtovanja dopustov (Marr 2017). Nastajajo tudi pri upravljanju spletnih mest, med drugim pa omogočajo tudi ciljanje (targetiranje) uporabnikov, zaradi česar se lahko posamezniko m ob uporabi spleta prikazujejo prilagojeni oglasi, recimo glede na njihovo zgodovino iskanja na spletu ali zadnje nakupe. Ti podatki imajo izjemen potencial tudi na področju javnomnenjskih raziskav (Japec in drugi 2015).

Največja prednost velikih podatkov, ki jo Japec in drugi (2015) omenjajo v svojem članku, je, da povečini ti podatki že obstajajo v neki obliki in zato ni potrebno zbiranje podatkov, ki je lahko zelo drago in prepočasno, kadar želimo sprejeti hitre odločitve. Drugo prednost predstavljajo veliki podatki, ki so posledica zapisovanja (merjenja) dogodkov v realnem času in so neposredno takrat tudi na voljo, kar prav tako predstavlja veliko prednost pri sprejemanju hitrih odločitev v poslovnem procesu.

4.2 IZZIVI PRI DELU Z NESTRUKTURIRANIMI VELIKIMI PODATKI

4.2.1 ZASEBNOST IN VARNOST

Osebne informacije posameznika lahko v kombinaciji z notranjimi velikimi bazami podatkov osvetlijo nove informacije o posamezniku, ki so zasebne in posameznik morda ne želi, da bi kdo vedel zanje. Prav tako lahko informacije, za katere se posamezniki ne zavedajo, da so lahko komu dostopne, dajejo vpogled v njihovo življenje. Podjetja in organizacije jih nato lahko zbirajo in uporabijo za različne namene (npr. oglaševanje). Sporna je tudi uporaba velikih podatkov v pravne namene, saj se s tem poveča možnost diskriminacije brez vednosti posameznika, ki zato tudi nima možnosti zagovora oziroma obrambe. Pojavi se lahko problem družbene stratifikacije med posamezniki, ki znajo velike podatke obrniti v svojo korist, in tistimi, ki te možnosti nimajo (Katal in drugi 2013).

Čeprav so v bazo, ki jo za namene analize nestrukturiranih podatkov uporabljajo, večinoma zajeti podatki o dejanjih mnogih posameznikov, je kljub temu možen vpliv informacije na posameznika. Tej nevarnosti se lahko delno izognemo z anonimizacijo postopka, vendar obstaja možnost tudi obratnega postopka, s katerim bi te informacije lahko pridobili. Drug etični problem je pridobivanje informacij o posamezniku, na podlagi katerih lahko predvidevamo prihodnje obnašanje in dejanja, kar bi lahko vplivalo na posameznikovo svobodno voljo (Katal in drugi 2013).

4.2.2 DOSTOP DO PODATKOV IN DELJENJE INFORMACIJ

Če so podatki uporabljeni za odločanje, je pomembno, da so dostopni pravočasno, celostno in točno. Do velikih baz nestrukturiranih velikih podatkov imajo povečini dostop le večja podjetja in ta imajo tudi popoln nadzor nad tem, kdo lahko dostopa do katerih informacij. Od podjetij sicer težko pričakujemo, da bodo z drugimi podjetji delila svoje podatke, saj s tem ogrožajo zasebnost svojih strank in svojo prednost na trgu. Vendar pa to ustvarja vedno večji digitalen razcep med podjetji, sprememba podatkov v tekmovalno

prednost je namreč tista, zaradi katere so veliki podatki nova revolucija v poslovnem svetu (Katal in drugi 2013).

4.2.3 SHRANJEVANJE IN PROCESIRANJE PODATKOV

Družbena omrežja in senzorske naprave proizvajajo ogromne količine podatkov, ki jih ne moremo shraniti na običajen razpoložljiv prostor. En način reševanja tega problema je shranjevanje v »oblaku« (Cloud), pri čemer gre za shranjevanje podatkov na server zunaj organizacije, ki ga ta najame oziroma zakupi. Problem pri uporabi oblaka predstavlja dolgotrajno nalaganje podatkov in posodabljanje, zato ta rešitev ni vedno najbolj optimalna. Prenosu podatkov v postopku od shranjevanja do procesiranja se lahko izognemo tako, da procesiramo na mestu, kjer shranjujemo, prenesemo pa samo rezultate. Druga rešitev pa je prenos samo tistih podatkov, ki so za nas pomembni (Katal in drugi 2013).

Čeprav je prostor za shranjevanje podatkov relativno poceni, so stroški ustvarjanja in vzdrževanja sistemov za analizo velikih podatkov dragi. Stroški za analizo velikih podatkov vsebujejo vsaj medij (disk) za shranjevanje podatkov, aktivne računalniške komponente (procesor in delovni pomnilnik) in infrastrukturo (server, hlajenje, elektrika, dostop do interneta in varnostni sistemi). Stroški lahko nanesejo tudi več sto tisoč evrov, zaradi česar se včasih bolj splača uporaba zunanjih računalniških sestavov. Tehnologija na področju velikih podatkov se namreč zelo hitro razvija, oprema je zato lahko zelo hitro zastarela (Japec in drugi 2015).

Za shranjevanje velikih podatkov se lahko uporabljajo tradicionalne relacijske baze (RDBMS) ali NoSQL baze, ki so še posebej uporabne za nestrukturirane velike podatke, ki niso tabelarični. Analiza velikih podatkov omogoča fleksibilnost, kar pomeni omogočanje uporabe različnih analitičnih scenarijev na enakem delu velikih podatkov (Cuzzocrea in drugi 2011). Za ta namen je pomembno kombiniranje prednosti tradicionalnih relacijskih baz podatkov in novejših NoSQL baz, ki omogočajo predstavljanje in upravljanje horizontalnih delov podatkov (Cattell v Cuzzocrea in drugi 2011, 102).

4.2.4 *ANALIZA PODATKOV*

Analiza nestrukturiranih podatkov zahteva veliko sposobnosti in znanja. Način analize je odvisen od tega, kakšne rezultate želimo oziroma kakšne odločitve želimo sprejeti. To lahko storimo na dva načina: lahko vključimo velike količine podatkov v analizo ali se vnaprej odločimo, kateri podatki so za nas pomembni (Katal in drugi 2013).

Oblikovanje sistemov za obdelavo podatkov in analize podatkov na način, ki bi nam omogočal izvleči relevantne informacije za osnovanje poslovnih odločitev, je zaradi pospešene rasti količin velikih podatkov oteženo. Zato se je pomembno izogniti nagnjenosti ljudi k prepoznavanju vzorcev tam, kjer jih pravzaprav ni, ampak jih vidimo samo zaradi velike količine podatkov, kjer je veliko povezav (Kaisler in drugi 2013).

4.2.5 *ZAHTEVANE SPRETNOSTI*

Ker gre za novo in razvijajočo se tehnologijo, so potrebne raznolike sposobnosti s tehničnega, analitičnega, interpretativnega in kreativnega področja. Za razvoj znanja so potrebni izobraževalni programi znotraj organizacij in usmeritev smeri študija na področje velikih podatkov na univerzah (Katal in drugi 2013).

Delo z velikimi podatki zahteva vsaj štiri profile zaposlenih (Japec in drugi 2015):

- Strokovnjak za domene: uporabnik, analitik ali vodja s poglobljenim znanjem o podatkih, njihovi pravilni uporabi in omejitvah. Ta profil je še posebej pomemben pri uporabi nestrukturiranih podatkov.
- Raziskovalec: član ekipe z izkušnjami s področja tradicionalnih metod raziskovanja in statistike. Pripomore k primerni integraciji velikih podatkov v javnomnenjske raziskave.
- Računalniški znanstvenik: tehnično izobražen član ekipe z znanjem programiranja in tehnologij za procesiranje podatkov. Pomembne so predvsem kompetence v okolju, ki uporablja ukazne vrstice, znanje programskih jezikov, delo z bazami podatkov in naprednimi analitičnimi orodji.

- Sistemski administrator: član ekipe, ki je odgovoren za definiranje, ustvarjanje in vzdrževanje infrastrukture za shranjevanje in analizo velikih podatkov.

Kljub temu podjetja pogosto zaposlujejo posameznike, ki opravljajo vsa našteta dela.

4.2.6 TEHNIČNI IZZIVI

Razvijanje popolnih naprav ali programske opreme ni možno, zato je treba napake omejiti, kolikor je mogoče. To je zelo zahtevna naloga, ki vključuje tudi precej visoke stroške (Katal in drugi 2013).

V zadnjih letih so se spremenile tehnologije, ki se lahko uporabljajo za shranjevanje podatkov. Trde diske (HDD) so zamenjali pogoni SSD (Solid State Drives), ki nimajo enakega delovanja pri zaporednem in naključnem prenosu podatkov. Kot že omenjeno, lahko rešitev ponuja tudi shranjevanje v oblaku, ki pa s seboj prinaša izzive pri analizi podatkov (Katal in drugi 2013).

Pomembno je zagotoviti shranjevanje kvalitetnih in uporabnih podatkov za boljše rezultate in lažje sprejemanje poslovnih odločitev. Problem se pojavi pri ugotavljanju kvalitete podatkov, ugotavljanju potrebne količine podatkov za sprejemanje odločitev in ugotavljanju točnosti podatkov (Katal in drugi 2013). Med podatki, ki jih želimo analizirati, se namreč pogosto pojavijo tudi velike količine nepovezanih podatkov. Filtriranje, čiščenje in izločanje nepovezanih podatkov je pri analizi podatkov izjemno pomembno, saj ima velik vpliv na kvaliteto končne analize (Cuzzocrea in drugi 2011).

Nestrukturiranost velikih podatkov ne vpliva le na tipične probleme z integracijo, temveč tudi na razvoj in oblikovanje analitičnih orodij. Za namene pomenske analize je namreč pomembno, da nestrukturirane podatke pretvorimo v primeren, strukturiran format. Kljub temu so skladišča velikih nestrukturiranih podatkov pogosto precej nestrukturirana, še posebej, ko gre za podatke s spletnih družbenih omrežij, biološke podatke itd. v primerjavi z nestrukturiranimi podatki, ki so popularni v tradicionalnih BI orodjih (XML, RDF) (Cuzzocrea in drugi 2011).

5 ANALIZA NESTRUKTURIRANIH VELIKIH PODATKOV

Analiza velikih podatkov združuje uporabo velikih količin podrobnih informacij in napredne analitike – nabora različnih orodij, vključno s tistimi, ki so osnovani na temeljih napovedne analitike, statistike, umetne inteligence, procesiranja naravnega jezika in podatkovnega rudarjenja. Napredna analiza nam pomaga ugotoviti, kaj se je spremenilo in kako naj na spremembo reagiramo, hkrati pa je najboljši način za odkrivanje novih segmentov strank, razumevanje sezonske prodaje itd. (Russom 2011).

Tehnike za raziskovanje nestrukturiranih velikih podatkov niso nove, pojavile so se že v 90. letih. Spremenila se je le potreba po analizi podatkov, vedno več organizacij namreč priznava prednost, ki jim jo lahko prinesejo pridobljene informacije (Russom 2011).

Analiza nestrukturiranih podatkov s pomočjo novih tehnologij postaja vedno bolj dostopna, tako cenovno kot tudi praktično. Novi pristopi z uporabo moči paralelnega procesiranja spreminjajo način upravljanja in analiziranja podatkov. Uporabljajo razširljivo ne-relacijsko arhitekturo, razdeljene delovne okvire za procesiranje in ne-relacijske in paralelne relacijske baze (Das in Kumar 2013).

5.1 ANALIZA RAZLIČNIH VRST NESTRUKTURIRANIH PODATKOV

5.1.1 ANALIZA BESEDIL

Najbolj pogost format shranjevanja je besedilo (e-pošta, poslovni dokumenti, spletne strani in družbena omrežja), zaradi česar ima v poslovnem svetu večjo vrednost kot strukturirani podatki. Na splošno se analiza besedil nanaša na interdisciplinaren postopek, s katerim izločimo uporabne informacije in znanje iz nestrukturiranih besedil s pomočjo strojnega učenja, statistike, računalniškega jezika in podatkovnega rudarjenja. Večina sistemov rudarjenja besedil je osnovana na izrazih v besedilih in procesiranju naravnega jezika (NLP), ki omogoča računalnikom analizo, interpretacijo in celo generiranje besedil (Chen in drugi 2014).

5.1.2 *ANALIZA SPLETNIH PODATKOV*

Cilj analize spletnih podatkov je pridobivanje, izločanje in ocenjevanje informacij iz spletnih dokumentov in strani za pridobivanje uporabnih informacij. Analiza spletnih vsebin skozi baze, pridobivanje informacij, procesiranje naravnega jezika in rudarjenja besedil vključuje analizo različnih tipov podatkov, kot so besedilo, slike, zvok, videoposnetki, kode, metapodatki in hiperpovezave (Chen in drugi 2014).

5.1.3 *ANALIZA VEČPREDSTAVNOSTNIH VSEBIN*

Količine večpredstavnostnih vsebin, kot so slike, zvok in videoposnetki, naraščajo z veliko hitrostjo. Zaradi njihove heterogene narave in posledično velike količine razpoložljivih informacij pa je izločanje uporabnih informacij velik izziv (Chen in drugi 2014).

Pristopi k raziskovanju tovrstnih vsebin vključujejo (Chen in drugi 2014):

- Povzemanje večpredstavnostnih vsebin:
    - povzemanje zvoka: z izločanjem pomembnih besed ali fraz iz metapodatkov;
    - povzemanje videoposnetkov: lahko je statično ali dinamično. Pri statičnem uporabimo ključno zaporedje slik ali na kontekst občutljive pomembne ključne slike. Pri dinamičnem pa uporabimo serijo slik, pri katerih uporabimo različne metode za naraven gladek potek povzetka.
- Večpredstavnostno označevanje: gre za vstavljanje oznak, s katerimi lahko opišemo slike ali videoposnetke na osnovnem in pomenskem nivoju ter s pomočjo katerih lahko upravljamo, povzemamo in pridobivamo večpredstavnostne podatke. Ker ročno označevanje zahteva veliko časa, je bolj uporabno avtomatsko označevanje, pri čemer pa se lahko pojavlja problem pomenskih razlik.
- Popisovanje večpredstavnostnih vsebin: vključuje opisovanje, shranjevanje in organiziranje večpredstavnostnih informacij in pomoč uporabnikom, ki lahko priročno in hitro pregledajo večpredstavnostne vire. Poteka skozi pet postopkov – strukturno analizo (razdelitev videa v več pomenskih strukturnih elementov),

izločanje značilnosti (nadaljnje rudarjenje po značilnostih ključnih slik, objektov in gibanja), podatkovno rudarjenje, razvrščanje/označevanje vsebin ter poizvedba/iskanje.

- Predlogi v večpredstavnostnih vsebinah: gre za pristop priporočanja specifičnih večpredstavnostnih vsebin uporabnikom glede na njihove preference, s katerim lahko učinkovito zagotovimo personalizirane storitve. Sisteme za priporočanje lahko razdelimo na tiste s poudarkom na vsebini (s to metodo identificiramo glavne značilnosti uporabnikov in njihovih zanimanj ter jim priporočamo vsebine s podobnimi značilnostmi) in tiste s poudarkom na skupinskem filtriranju (identificiranje skupin s podobnimi interesi in priporočamo vsebine članom skupine glede na njihovo vedenje).
- Zaznavanje dogodkov v večpredstavnostnih vsebinah: z uporabo tehnologij lahko zasledimo pojavljanje dogodka v videoposnetkih s pomočjo orodij, ki vsebujejo besedilne opise, povezane s koncepti in primeri videoposnetkov. Ta pristop je še v razvoju, v glavnem pa se uporablja na področju športnih dogodkov in novic.

5.1.4 ANALIZA SPLETNIH OMREŽIJ

Analiza družbenih spletnih omrežij se je razvila s procesi kvantitativne analize in sociološke analize omrežij v razvijajočo se spletno analizo družbenih omrežij (Hirsch in Watts v Chen in drugi 2014, 196). Priljubljena spletna družbena omrežja (npr. Facebook, Twitter, LinkedIn) vsebujejo izjemno velike količine povezanih podatkov in vsebinskih podatkov (Chen in drugi 2014).

- Strukturna analiza povezav med podatki: podatke lahko predstavimo z grafi, na katerih so točke (uporabniki) in povezave, ki jih povezujejo. Z metodami, kot je razvrščanje na temelju značilnosti, verjetnostnih metodah in linearnih enačbah, lahko tudi vnaprej predvidevamo verjetne povezave med dvema točkama (Scellato in drugi v Chen in drugi 2014, 197).
- Vsebinska analiza: znana tudi kot analiza družbenih omrežij. Vključuje analizo besedil, večpredstavnostnih vsebin, pozicioniranja in komentarjev na spletnih družbenih omrežjih (Chen in drugi 2014).

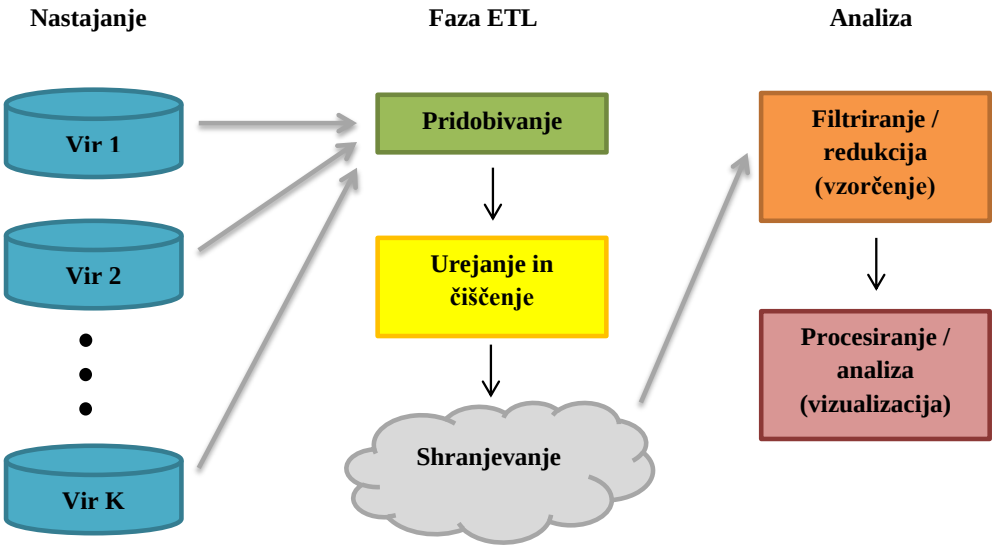
5.1.5 ANALIZA MOBILNIH (SENZORSKIH) PODATKOV

Mobilni podatki imajo edinstvene značilnosti, kot sta zaznavanje premikanja in zmožnost spreminjanja lokacije. Mobilni telefoni omogočajo možnost uporabe mobilnih socialnih skupnosti kadarkoli in kjerkoli. Mobilne socialne skupnosti označujejo skupino posameznikov z enakimi interesi (zdravje, zabava), ki se zbere na družbenem omrežju, člani pa sledijo nekemu skupnemu cilju in se posvetujejo med seboj. Tehnološki napredek na področju brezžičnih senzorjev, mobilne komunikacijske tehnologije in procesiranja podatkov v toku omogočajo celo nadziranje posameznikovega zdravstvenega stanja v realnem času (Chen in drugi 2014).

5.2 PROCES ANALIZE NESTRUKTURIRANIH VELIKIH PODATKOV

Obstaja več pristopov, s katerimi lahko pristopimo k procesiranju analize nestrukturiranih podatkov. Osnoven model za proces analize podatkov je naslednji (Japec in drugi 2015):

Slika 5.1: Model procesa analize podatkov



Vir: povzeto po Japec in drugi (2015).

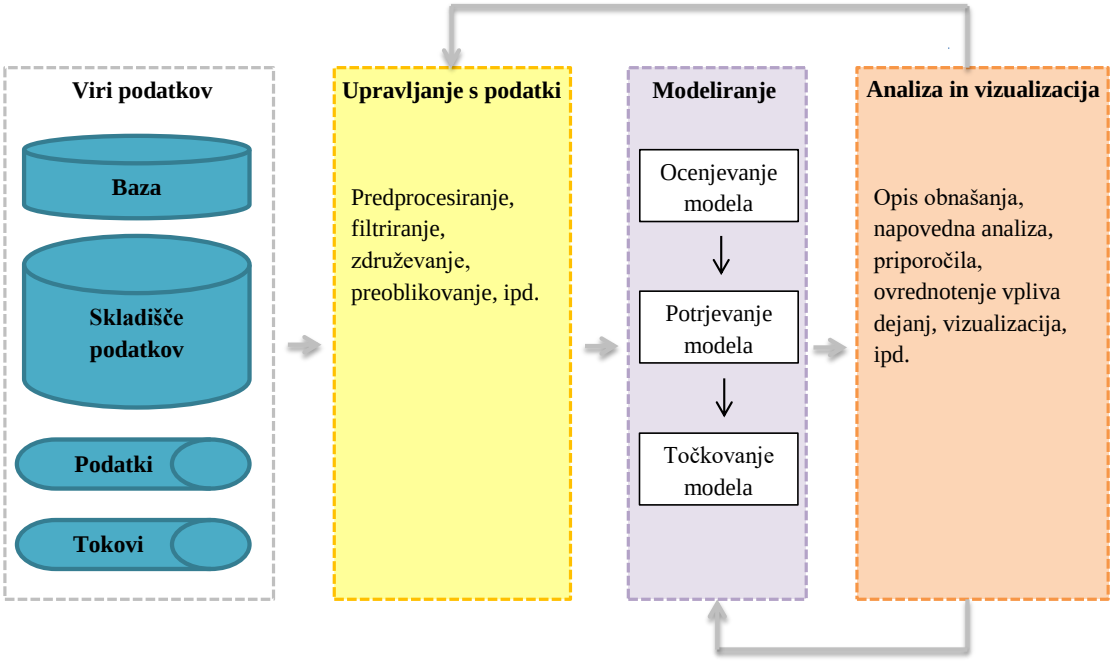
V fazi nastajanja podatkov so podatki ustvarjeni v nekem viru, namerno ali nenamerno (Japec in drugi 2015).

V drugi fazi (ETL) podatke najprej prenesemo iz virov, jih potrdimo in shranimo. Potem te podatke prevedemo, kodiramo, združimo/razdružimo in uredimo. V nadaljevanju podatke združimo in shranimo v podatkovno skladišče (Japec in drugi 2015).

V fazi analize so podatki spremenjeni v informacije s pomočjo filtriranja/redukcije, ki vključuje izbris nezaželene vsebine, združevanje značilnosti za ustvarjanje novih, vzorčenje. V fazi procesiranja, analize in vizualizacije pa so podatki analizirani in pripravljeni za interpretacijo ter izločanje uporabnih informacij (Japec in drugi 2015).

Assuncao in drugi (2014) so v svojem delu prikazali tradicionalne faze analitičnega delovnega načrta za delo z velikimi podatki. Faze pri obeh načrtih so si načeloma precej podobne:

Slika 5.2: Faze delovnega načrta za delo z velikimi podatki



Vir: povzeto po Assuncao in drugi (2014).

Najprej potrebujemo dostop do podatkov, v nadaljevanju jih filtriramo, združimo, transformiramo in izvedemo vse druge potrebne operacije za pripravo podatkov, na koncu pa jih analiziramo s pomočjo različnih pristopov in prikažemo s pomočjo vizualizacije. Dodana pa je faza modeliranja, torej preverjanje nastavljenega modela za delo z velikimi podatki.

V nadaljevanju so podrobneje predstavljene posamezne faze v procesu analize podatkov.

5.2.1 *DOLOČANJE CILJEV RAZISKAVE*

Določanje ciljev zbiranja podatkov je prvi korak procesa. Podatke, ki jih želimo uporabiti, je treba izbrati glede na strategijo oziroma potrebe poslovanja, kot so operacije, odločanje in načrtovanje. Vnaprej je treba določiti vire, tip, količino in kvaliteto podatkov, kriterije ocenjevanja in specifikacije, pa tudi pričakovane cilje (Cai in Zu 2015).

5.2.2 *Generiranje in pridobivanje podatkov*

Po določanju ciljev raziskave sledi pridobivanje podatkov. Obstaja več načinov pridobivanja podatkov, kot so integracija podatkov, iskanje po spletu, spletni pajki, metode posrednikov in podobno (Cai in Zu 2015). Ko zberemo surove podatke, jih skozi učinkovite mehanizme prenesemo v primeren sistem za shranjevanje podatkov, ki podpira različne analitične pristope (Chen in drugi 2014).

Pridobivanje podatkov je relativno nezahtevno, kljub temu pa je treba v mislih imeti tudi kvaliteto podatkov in ne samo količino (Cai in Zu 2015). V nasprotnem primeru lahko pride do simptoma prevelike samozavesti pri uporabi Big Data (Big Data hubris), kjer raziskovalec verjame, da količina velikih podatkov izniči njene druge pomanjkljivosti (Japec in drugi 2015).

5.2.3 *Ocenjevanje kvalitete podatkov*

Ker uporabljeni nestrukturirani veliki podatki niso vedno ustvarjeni s strani uporabnikov, je pomembno posebno pozornost nameniti kvaliteti podatkov. Cai in Zu (2015) sta predlagala dvonivojski koncept za merjenje standarda kvalitete podatkov, pri katerem so glavni kriteriji:

- razpoložljivost (dostopnost, časovnost, avtorizacija);
- uporabnost (definicija/dokumentacija, verodostojnost, metapodatki);
- zanesljivost (natančnost, integriteta, konsistenca, popolnost, revizija);
- pomembnost (ustreznost);
- predstavnost (struktura, primernost za branje).

Preverjanje kvalitete uporabljenih podatkov je zelo pomemben korak, saj na kvalitetnih podatkih temelji celotna raziskava in veljavnost končnih rezultatov. Kljub temu se lahko izkaže za zelo zahtevnega predvsem pri uporabi nestrukturiranih podatkov, bodisi pridobljenih s spletnih omrežij, kjer je pogosto vprašljiva verodostojnost, bodisi pri podatkih s področja biologije, kjer so problem lahko visoka nestrukturiranost in različni formati (Cai in Zu 2015).

5.2.4 *ČIŠČENJE PODATKOV*

Namen čiščenja podatkov je izboljševanje kakovosti skozi zaznavanje in odstranjevanje napak, kot so napake v podatkih, manjkajoče informacije, nekonsistence in nasičenost (Cai in Zu 2015). Zbrani nestrukturirani podatki lahko včasih vključujejo nepomembne oziroma neuporabne podatke, ki brez potrebe povečujejo količino podatkov in potreben prostor za shranjevanje ter vplivajo na kasnejšo analizo (Chen in drugi 2014). Čiščenje podatkov lahko glede na metode vpeljevanja razdelimo na ročno vpeljevanje, pisanje posebnih programov, čiščenje podatkov nepovezano s specifičnim programom in reševanje problema tipa specifične domene (Cai in Zu 2015).

5.2.5 *PREVERJANJE MODELA*

Na podlagi razpoložljivih podatkov si pripravimo model (načrt) procesa, po katerem nameravamo obdelovati, analizirati in interpretirati podatke. V tej fazi poteka ocenjevanje, pri katerem že pripravljene podatke uporabimo za testiranje parametrov. Ko je model ocenjen, mora biti še pred dokončno uporabo potrjen z uporabo (najpogosteje originalnih) podatkov in specifičnih metod. V zadnji točki model privzamemo in uporabimo na podatkih za ustvarjanje opisov, napovedi (poslovnih) in priporočil (Assuncao in drugi 2014).

5.2.6 *ANALIZA IN ANALITIČNE METODE*

Namen analize je izločanje uporabnih vrednosti, predlogov za poslovno odločanje in vpeljevanje poslovnih odločitev. Je zadnja in najbolj pomembna faza pri pridobivanju vrednosti iz velikih podatkov. Analiza velikih podatkov vključuje analitične metode za tradicionalne in velike podatke, analitično arhitekturo velikih podatkov in programsko opremo za rudarjenje in analiziranje velikih podatkov (Chen in drugi 2014).

Obstaja več analitičnih metod. Za nestrukturirane podatke je pomembno indeksiranje, ki je učinkovita metoda za krčenje stroškov shranjevanja in branja diskov. Indeksiranje izboljšuje vstavljanje, brisanje, spreminjanje podatkov, tako v tradicionalnih relacijskih bazah kot tudi na nestrukturiranih podatkih. Slabost indeksiranja je to, da prinaša dodatne stroške za shranjevanje indeks datotek, ki jih je treba vzdrževati med nalaganjem podatkov (Chen in drugi 2014).

Za analizo nestrukturiranih velikih podatkov je pomembno tudi paralelno procesiranje, ki je v primerjavi s tradicionalnim procesiranjem hitrejše, saj se naloga porazdeli na več ločenih procesov, ki so neodvisno in istočasno izvedeni (Chen in drugi 2014).

5.3 IZZIVI PRI PROCESU ANALIZE NESTRUKTURIRANIH VELIKIH PODATKOV

Velike količine večdimenzionalnih in nestrukturiranih podatkov prinašajo nove izzive za analitike.

1. Eden izmed glavnih izzivov pri nestrukturiranih podatkih se pojavi že pri generiranju podatkov. Ti podatki namreč niso ustvarjeni z instrumenti in metodami, ki bi bile prirejene za ustvarjanje veljavnih in zanesljivih podatkov namenjenih znanstvenim analizam, ampak so pogosto stranski proizvod procesov, katerih namen ni vedno v skladu s tistim, ki ga imajo analitiki (Japec in drugi 2015).
2. Avtomatski sistemi za generiranje vsebin na spletu predstavljajo drug izziv. Posledično je kvaliteta podatkov lahko slabša, analitiki pa morajo biti pozorni na omejitve pri analizi in skušati minimizirati učinek napak na njihove rezultate (Japec in drugi 2015).
3. Naslednji izziv pri procesu analize nestrukturiranih podatkov predstavlja dinamičnost algoritmov, ki jih je treba nenehno razvijati in izboljševati za zagotavljanje boljše uporabniške izkušnje. Kritičen problem predstavlja točka (»koleno«), kjer delovanje algoritma neha naraščati linearno z viri procesiranja ali začne celo upadati. Reševanje tega problema pogosto zahteva nov algoritem ali izboljšavo originalnega, tako da se »koleno« premakne naprej po lestvici (Kaisler in drugi 2013). Algoritmi so pogosto lastniški in so zato redko javno merjeni za

natančnost, zaradi česar po navadi njihova neuspešnost ni znana javnosti (Japec in drugi 2015).

4. Procesiranje nestrukturiranih podatkov v realnem času predstavlja velik izziv. Velikost in kompleksnost teh podatkov namreč presega običajne tehnične sposobnosti zajemanja, upravljanja in procesiranja teh podatkov v doglednem času in stroškovnem okviru (Wu in drugi 2014).
5. Naslednji izziv se nanaša na iskanje ključnih podatkov, ki omogočajo iskanje rešitev v okviru prostora problema. Napovedna analiza, s katero želimo najti pravi odgovor, je zelo zapletena (Kaisler in drugi 2013). Vendar pa je povsem enostavno potrditi, da gre res za pravi odgovor, ko enkrat vemo, kje se nahaja »igla v senu« (Felten v Kaisler in drugi 2013, 1002). Problem lahko omilimo z zmanjšanjem »sena« (naše množice podatkov) ali izboljšamo naše iskanje, analizo in postopke sprejemanja odločitev (Kaisler in drugi 2013).
6. Problem predstavlja tudi procesiranje velikih količin podatkov v kvalitetne podatke. S tem problemom se soočamo predvsem z različnimi algoritmi s poudarkom na razumevanju (strojnem učenju). Z njimi lažje vidimo celo sliko, interpretiramo in prepoznamo vzorce obnašanja (Kaisler in drugi 2013).
7. Zaradi zelo velikih količin heterogenih podatkov predstavlja velik izziv ugotavljanje, katere podatke pravzaprav imamo in kako jih analizirati (Kaisler in drugi 2013). V primeru napovedne analize problem predstavljajo tudi nepričakovani dogodki, ki jih vnaprej ne moremo predvidevati in imajo velik vpliv na obnašanje populacije, ta sprememba pa lahko potem močno vpliva na novonastale podatke (Taleb v Kaisler in drugi 2013, 1002).

To so le nekateri izmed najpogostejših izzivov, ki se v praksi pojavljajo pri procesu analize nestrukturiranih velikih podatkov.

6 PRISTOP K ANALIZI NESTRUKTURIRANIH PODATKOV V PRAKSI

Ker sem želela pridobiti informacije o tem, kako delo z nestrukturiranimi podatki poteka v praksi, sem izvedla kvalitativno raziskavo. Prvotna ideja za raziskavo je bila sicer primerjava več različnih pristopov k analizi nestrukturiranih podatkov v praksi in izzivov, ki se pri tem pojavljajo, a so se pri tem pojavile ovire.

Kontaktirala sem več slovenskih podjetij, ki se ukvarjajo z javnomnenjskimi raziskavami, telekomunikacijami, razvojem spletnih aplikacij in celo Statistični urad RS. Za te tipe podjetij oziroma organizacij sem domnevala, da bi glede na naravo dela lahko imela interes za delo z nestrukturiranimi velikimi podatki. Ker se je izkazalo, da je na slovenskem trgu analiza velikih nestrukturiranih podatkov zelo redka, žal ni bilo možno pridobiti dovolj intervjuvancev za medsebojno primerjavo. Kljub temu mi je uspelo pridobiti informacije o delu z nestrukturiranimi podatki s pomočjo Instituta Jožef Stefan, kjer se s tovrstnimi raziskavami ukvarjajo v raziskovalne namene.

»Institut Jožef Stefan izvaja vrhunske raziskave in razvoj tehnologij, kot so nanotehnologije, novi materiali, biotehnologije, tehnologije vodenja in proizvodnje, komunikacijske tehnologije, računalniške tehnologije in tehnologije znanja, okoljske tehnologije in reaktorske tehnologije« (Institut Jožef Stefan 2017).

Ker je dejstvo, da gre za tako redko prakso, lahko zanimivo predvsem z vidika *»kaj so tiste ovire, da se podjetja, ki bi lahko imela interes na tem področju, ne spustijo v tovrstne raziskave«,* sta bila izvedena še dva intervjuja z večjima mednarodnima podjetjema. Na primeru teh dveh podjetij bosta v primerjavi z analizo nestrukturiranih velikih podatkov predstavljena tudi alternativna pristopa analize podatkov, ki se v praksi na našem trgu pogosteje pojavljata, kot sta analiza nekoliko bolj strukturiranih velikih podatkov (Ekipa2, hčerinska družba Outfit7) in analiza manjših količin nestrukturiranih podatkov (Ipsos).

Outfit7 je eno izmed najhitreje rastočih mednarodnih podjetij na svetu, ki se ukvarja z razvojem zabavnih digitalnih vsebin. Najbolj znano je po svojem karakterju Talking Tom (govoreči Tom). Njihove aplikacije so prejele več nagrad, štejejo pa že več kot 6,5

milijarde prenosov aplikacij in 348 milijonov mesečno aktivnih uporabnikov (Outfit7 2017).

Mednarodno podjetje Ipsos je tretja največja raziskovalna agencija na svetu, ki se ukvarja z raziskavami trga in svetovanjem. Podjetje deluje v 88 državah sveta in ima več kot 5000 strank (Ipsos 2017).

6.1 RAZISKOVALNA METODA

Za metodo raziskovanja je bila uporabljena kvalitativna raziskava. Kvalitativna raziskava je osnovana na določenih predpostavkah oziroma tezah, ki jih oblikujemo na podlagi zbiranja gradiva. V nasprotju s kvantitativno raziskavo nas pri kvalitativni raziskavi ne zanimajo frekvence določenega pojava, ampak izkušnje, torej kako posamezniki določene stvari razumejo, si razlagajo in nanje reagirajo (Mesec 1998).

Med najpogostejše kvalitativne pristope spada tudi metoda poglobljenega intervjuja, pri kateri poteka pogovor o vnaprej določeni temi. Še posebej uporabna je pri pridobivanju mnenja strokovnjaka glede določene teme oziroma trendov na njegovem strokovnem področju. Pri tem so vprašanja dovolj odprta, da omogočajo izražanje mnenj in pogledov, z njihovo pomočjo pa lahko ugotovimo vzroke za določen pojav ali dejanja (Mediana 2009).

Izvedla sem delno strukturirane poglobljene intervjuje z vnaprej določenimi glavnimi vprašanji, ki sem jim glede na situacijo in odgovore intervjuvancev dodala še kakšno dodatno vprašanje. Delno strukturiran vprašalnik so sestavljala naslednja vprašanja, ki so bila zastavljena vsem intervjuvancem:

1. Koliko zaposlenih v vašem podjetju sodeluje pri analizi podatkov? Kateri oddelki?
2. Katere programe uporabljate pri delu s podatki?
3. Katere vrste podatkov analizirate?
4. Kako v vašem podjetju poteka postopek analize podatkov, od samega pridobivanja podatkov do vizualizacije/zaključnega poročila?

5. Katera faza v postopku vam vzame največ časa? Katera faza je najbolj zahtevna?
6. Kateri so po vašem mnenju največji izzivi, s katerimi se srečujete pri analizi podatkov? Kako jih rešujete?
7. Ali podatke uporabljate tudi za napovedovanje trendov (predictive analysis)? Če da, ste morda že imeli primer, kjer je prišlo do napačnega napovedovanja na podlagi preteklih podatkov? Kaj je bil vzrok in kako ste temu prilagodili vaš pristop?
8. Kako se bo po vašem mnenju v prihodnosti razvijalo analiziranje podatkov?

6.2 POTEK IZVEDBE INTERVJUJEV

Intervjuji so bili izvedeni v juliju in avgustu 2017. Vsi intervjuvanci so pokazali visoko pripravljenost na sodelovanje in profesionalen odnos, zaradi česar nisem imela težav z izvedbo intervjuja. Intervjuja z Mojco Klenovšek Podobnik (podjetje Ipsos) in Martinom Žnidaršičem (Institut Jožef Stefan) sta bila izvedena osebno v prostorih podjetja/organizacije, intervju z Antejem Odićem (Ekipa2, hčerinska družba Outfit7) pa preko e-pošte. Ker ni šlo za občutljivo temo (kot so merjenja stališč in podobno) in ker so bila bistvena vprašanja določena vnaprej, se mi izvedba preko spleta ni zdela sporna. Edina opazna razlika je dolžina odgovorov in s tem posledično manjša količina pridobljenih informacij kot pri osebnih intervjujih, kar pa je pri spletnem intervjuju pričakovano.

6.3 ANALIZA INTERVJUJEV

6.3.1 INSTITUT JOŽEF STEFAN

Na Institutu Jožef Stefan se na Odseku za tehnologije znanja z analizo nestrukturiranih podatkov ukvarja kakšnih šest oseb od štiridesetih na odseku. Analizirajo predvsem besedila, majhen del pa se jih ukvarja tudi z analizo rentgenskih slik. Z analizo

nestrukturiranih podatkov se sicer ukvarjajo občasno, odvisno od projektov. Z njo se ukvarjajo tudi nekateri drugi odseki, na celotnem Institutu Jožef Stefan (IJS) se tako z nestrukturiranimi velikimi podatki ukvarja približno 20–30 zaposlenih (Žnidaršič 2017).

V literaturi se poudarjajo predvsem uporabnikom bolj prijazni in splošno znani programski vmesniki (npr. Hadoop in MapReduce), na IJS pa poleg teh **za namene analize besedilnih nestrukturiranih podatkov uporabljajo predvsem svojo programsko opremo in namenske knjižnice v Pythonu (od katerih sicer nekatere tudi sledijo principom MapReduce)**. Uporabljajo NLTK, knjižnico za obdelavo naravnega jezika in knjižnice, ki se uporabljajo za bolj specifične stvari, kot je detekcija sentimenta, knjižnice za oblikoslovno označevanje besedil (part – of – speech (POS) tagging), knjižnica scikit-learn za strojno učenje in aktivno strojno učenje ipd. V uporabi imajo tudi kakšne svoje programe, kot je Latino, ki ga je v C# razvil njihov sodelavec Miha Grčar. Imajo tudi svoja orodja, ki so zaradi prijaznih grafičnih vmesnikov namenjena tudi začetniškim uporabnikom, kot sta TextFlows (za obdelavo besedil) in ClowdFlows (za raznovrstne podatke, primeren tudi za delo na velikih podatkih zaradi možnosti paralelnega procesiranja) (Žnidaršič 2017).

Analizirane **vrste nestrukturiranih velikih podatkov** so večinoma besedila, v manjši meri pa tudi (predvsem medicinske) slike in EKG signali. Vendar gre pri slednjih dveh samo za nestrukturiran tip podatka, ne pa tudi velike količine (Žnidaršič 2017).

Ko sem Žnidaršiča (2017) vprašala, **kako v njihovem podjetju poteka postopek analize nestrukturiranih velikih podatkov, od samega pridobivanja podatkov do vizualizacije/zaključnega poročila,** je poudaril, da je še pred pridobivanjem podatkov nekaj korakov, pri čemer je prvo razumevanje problema, ki je zelo pomemben korak. Potreben je razmislek, ali lahko z analizo nekih velikih podatkov res dosežemo želeni namen. Z analizo problema lahko torej ugotovimo, ali je analiziranje tako velikih količin podatkov sploh potrebno.

Naslednji korak je preverjanje, ali te podatke imamo oziroma ali imamo na voljo prave podatke. Pogosto se namreč lahko zgodi, da ima nekdo kvalitetne in dobre podatke, ki pa niso ustrezni za želeno raziskavo. Tudi v primeru, ko imamo razpoložljive približno prave podatke, se lahko zgodi, da niso dovolj natančni, recimo v primeru, ko so agregirani (Žnidaršič 2017).

V samem začetku je tako treba ugotoviti, ali rešujemo pravi problem, imamo prave podatke, nato pa se podatke lahko začne zbirati. Pri tem sta spet zelo pomembna razmislek in načrtovanje vnaprej, še posebej pri delu z velikimi podatki. Najverjetneje bo namreč potrebno shraniti zelo velike količine podatkov, kar je treba že vnaprej oceniti in pripraviti ustrezno računalniško opremo (Žnidaršič 2017).

Pomemben je tudi format zapisa, če bi se namreč odločili za gol podatek, potrebujemo manj prostora, kot če bi zraven zapisovali še metapodatke, ki lahko potreben prostor povečajo za nekajkrat. Format je pogosto stvar dogovora s partnerji, ker pa se večkrat podatki shranjujejo za uporabo različnih uporabnikov, je potreben razmislek o tem, kaj vse bomo zapisovali, da ne bomo zavzeli preveč pomnilnika ali imeli premalo podatkov (Žnidaršič 2017).

Žnidaršič (2017) pove, da za analizo, ki je naslednji korak v procesu, uporabljajo raznorazne algoritme na različne načine. Zelo radi uporabljajo tudi strojno učenje, kjer se želijo naučiti nekih tipičnih modelov in narediti klasifikatorje. V primeru analize besedil je nekaj tisoč besedil potrebno ročno označiti, ali gre za pozitivnega ali negativnega, neprimeren govor, diplomatski govor ipd., kar je dolgotrajen postopek. Ko se ta postopek zaključi, se lahko včasih zgodi, da ob analizi prvih rezultatov ugotoviš, da bi potreboval še kakšne dodatne informacije. Pride do povratne zanke v procesu, to namreč popraviš in se vrneš nazaj na zbiranje. To velja tudi za druge korake, iz vsakega greš z določeno verjetnostjo še malo nazaj za prilagajanje in izboljševanje.

Katera faza je najbolj zahtevna, je zelo odvisno od problema, včasih je to lahko ugotavljanje, kaj sploh je raziskovalni problem, naslednjič sestavljanje posebnega algoritma. **Največ časa** jim vzame faza čiščenja podatkov, ki je zelo pomembna za nadaljnjo analizo. Čiščenje podatkov je včasih lahko še posebej zahtevno pri obdelavi besedil, ker so lahko znotraj besedila nepravilne besede, znaki, izpeljanke, skrajšane ali združene besede. Problem lahko nastane tudi, ko gre za besedila, označena z besedo, ki ima lahko več pomenov ali ko so vsebovani kakšni drugi nepričakovani podatki. Čiščenje podatkov je pomembno tudi v primeru, ko je zajem podatkov luknjast (npr. pri senzorskih podatkih) (Žnidaršič 2017).

Včasih se izkaže za zahtevno tudi povezovanje raziskovanja s stroko. V primeru raziskovanja specifične teme, recimo s področja poslovanja, je namreč potrebno tudi

poznavanje konteksta in terminologije, kar rešujejo s sodelovanjem z eksperti s teh področij (Žnidaršič 2017).

Med **izzivi pri analizi nestrukturiranih velikih podatkov** je tako bilo omenjeno shranjevanje tako velikih količin podatkov, pri čemer je še posebej velik izziv zadosti hitro shranjevanje in hiter priklic ter odločanje o shranjevanju metapodatkov. Dodaten izziv predstavlja vzdrževanje, torej, da podatkovne toke pobirajo brez prekinitev oziroma kar se da neprekinjeno, torej tudi med vsemi potrebnimi posodobitvami, pa izpadi elektrike in podobnimi nevšečnostmi (Žnidaršič 2017).

Za napovedovanje trendov (predictive analysis) je znano, da v določenih primerih ni možno napovedati z veliko natančnostjo. Problem nastane takrat, kadar so napovedi tako napačne, da so neuporabne. Žnidaršič (2017) pravi, da se je v tem primeru najbolje vrniti na prejšnje korake, skušati izboljšati algoritem, spremeniti čiščenje podatkov, morda uporabiti še kakšne dodatne podatke in s tem ugotoviti, ali je točnost boljša in ocena napake nižja. V najslabšem primeru se lahko kdaj opazi napaka že na samem začetku, recimo uporaba neustreznih podatkov.

Včasih na problem lahko vplivajo tudi drugi dejavniki, ki se jih ne da izmeriti ali ne vemo, da sploh obstajajo. Takrat je potrebno oceniti vpliv dejavnika, če se izkaže za ključno stvar, je ta podatek nujno potreben, v nasprotnem primeru bo raziskava nerelevantna (Žnidaršič 2017).

Največjo vrednost določenih nestrukturiranih podatkov, kot so viri besedil na spletu, Žnidaršič (2017) vidi v tem, da v realnem času lahko pridobimo informacijo o tem, kaj ljudje mislijo. Pri tem tipu podatkov je prednost tudi to, da izvemo osebna mnenja ljudi.

Kljub prednostim analize nestrukturiranih velikih podatkov je njihova analiza na slovenskem trgu zelo redka. Razlog vidi v tem, da analiza tovrstnih podatkov za zdaj še ni ključnega pomena in je lahko uporabljena kot neka dodatna aktivnost. Lažje in verjetno bolj učinkovito je namreč tudi analiziranje marsikaterih strukturiranih podatkov, ki jih imajo podjetja na razpolago. Izjema so posebni primeri, kjer so nestrukturirani podatki edini, ki jih zanimajo. Za te primere tudi na slovenskem trgu počasi prihaja do vedno večjega zanimanja in delovanja v tej smeri (Žnidaršič 2017).

V prihodnosti bo analiziranje nestrukturiranih velikih podatkov po mnenju Žnidaršiča (2017) vedno bolj prikladno. Z razvojem orodij, sistemov in pristopov bo

imelo vedno več podjetij in posameznikov dostop do tega, da bodo lahko analizirali ta tip podatkov. Razvilo se bo tudi analiziranje in zaznavanje bolj specifičnih pojavov, recimo bolj podrobnih sentimentov, kar je za določen tip raziskav zelo pomembno.

S tem, ko pristopi postajajo bolj zreli za analizo, postajajo vedno bolj zreli tudi za generiranje. Avtomatsko generiranje nestrukturiranih podatkov obstaja že zdaj in bo tudi v prihodnosti vedno pogostejše. Primer takšnih objav so reklamna obvestila, spam e-maili ipd. Problem pri analizi se pojavi zato, ker v določenih primerih ni jasno vidno, ali je neko besedilo ustvaril človek ali avtomatski sistem (Žnidaršič 2017).

6.3.2 *Primerjava s pristopi podjetij Ipsos in Ekipa2 (hčerinska družba Outfit7)*

V podjetju Ipsos se z analizo nestrukturiranih podatkov ukvarjata dve zaposleni na oddelku za kvalitativne raziskave. V nasprotju s pristopom k analizi velikih nestrukturiranih podatkov za to delo ne uporabljajo nobenih programov, ker se jim zdijo časovno neučinkoviti. Za analizo različnih spletnih vsebin (slike, videoposnetki, besedila) so v nekaterih primerih uporabljali le Excel zaradi priročnih vmesnikov, ki te podatke pretvorijo v prepis (Klenovšek Podobnik 2017). V podjetju Ekipa2 (hčerinska družba Outfit7), kjer s strukturiranimi velikimi podatki dela osem zaposlenih na oddelku za analitiko, pa podobno kot na IJS uporabljajo predvsem lastno infrastrukturo in programe, ki jih razvijajo s pomočjo programskih orodij Python, R, Google Big Query in Jupyter (Odić 2017; Žnidaršič 2017).

Pridobivanje nestrukturiranih podatkov v podjetju Ipsos je določeno z metodo, upoštevajo torej zakonitosti za zbiranje kvalitativnih podatkov (Klenovšek Podobnik 2017). Pri podjetju Ekipa2 (hčerinska družba Outfit7) je za strukturirane velike podatke ta proces določen z vzpostavljeno infrastrukturo, podatki se skozi avtomatiziran proces ves čas zbirajo in shranjujejo v podatkovnih bazah. Ker uporabljajo le določen tip podatkov, problem načrtovanja ni tako zahteven kot pri nestrukturiranih velikih podatkih, kjer se mora, odvisno od tipa podatkov in problema, vsakič znova prilagajati načrtovanje raziskave. So pa opazne podobnosti pri načinu analize, oboji namreč uporabljajo različne algoritme strojnega učenja (Odić 2017; Žnidaršič 2017).

Medtem ko je pri nestrukturiranih velikih podatkih velik poudarek na načrtovanju in zbiranju podatkov, je pri analizi velikih strukturiranih podatkov in nestrukturiranih podatkov večji poudarek na interpretaciji rezultatov (Klenovšek Podobnik 2017; Odić 2017). Interpretacija je namreč stvar konteksta, ki je ni mogoče učinkovito reševati samo s pomočjo metode in tehnologije, ampak je stvar osebnega vlaganja in kreativnosti. Ta del je pri opisanih alternativnih pristopih tudi časovno najbolj zamuden in zahteven (Klenovšek Podobnik 2017), medtem ko je pri nestrukturiranih velikih podatkih najbolj zamudno čiščenje podatkov in načrtovanje raziskave (Odić 2017). Odić (2017) pravi, da skušajo v podjetju Ekipa2 (hčerinska družba Outfit7) ta problem v čim večji meri reševati s pomočjo lastne in obstoječih infrastruktur, namenjenih za analizo podatkov.

Kar je skupno vsem pristopom, je povezovanje raziskovanja s kontekstom raziskave, potrebno je namreč dobro poznavanje terminologije in nekoliko podrobneje spoznavanje področja, ki ga raziskujemo, da lahko raziskavo ustrezno načrtujemo in rezultate logično interpretiramo (Klenovšek Podobnik 2017; Odić 2017; Žnidaršič 2017).

Pri napovedni analizi sta si pristopa podjetja Ekipa2 (hčerinska družba Outfit7) in IJS zelo podobna, ker oba uporabljata velike podatke, pridobljene v realnem času. V obeh primerih poudarjajo učinkovito načrtovanje v fazi razvoja algoritmov, v primeru napačnega napovedovanja pa vračanje na prejšnje korake in ponovno prilagajanje pristopa. Kljub temu so pri napovedni analizi napačno napovedovanje in napake sestavni del napovedovanja, ki se jim v nekaterih primerih ne da popolnoma izogniti, v drugih primerih pa se jim da izogniti s pravilno evalvacijo in testiranjem (Odić 2017; Žnidaršič 2017).

Razloge za redko uporabo nestrukturiranih podatkov vsi intervjuvanci vidijo predvsem v tem, da dejansko ni takšne potrebe po analizi tega tipa podatkov (Klenovšek Podobnik 2017; Odić 2017; Žnidaršič 2017). Podjetju Ekipa2 (hčerinska družba Outfit7) že strukturirani podatki omogočajo neomejeno število analiz, optimizacij in eksperimentov za podporo odločanja in izboljševanja aplikacij, zaradi česar ne vidijo potrebe po uporabi nestrukturiranih podatkov (Odić 2017). Podjetje Ipsos pa trenutno po tem tipu raziskave na slovenskem trgu nima nekega povpraševanja s strani klientov, saj je zanje to prevelika investicija za premajhno uporabno vrednost (Klenovšek Podobnik 2017). Analiza nestrukturiranih podatkov je tako v tem trenutku rezervirana za tiste, ki za določeno raziskavo potrebujejo prav ta tip podatkov.

Njihove napovedi za prihodnost analize podatkov predvidevajo hitrejši razvoj orodij, metod in algoritmov (Klenovšek Podobnik 2017; Odić 2017; Žnidaršič 2017). To bo po mnenju IJS uporabo približalo več podjetjem in posameznikom, podjetje Ipsos pa izpostavlja tudi skrajšan čas analize nestrukturiranih podatkov zaradi preusmerjanja enostavnih nalog v avtomatizirane sisteme (Klenovšek Podobnik 2017). V podjetju Ekipa2 (hčerinska družba Outfit7) izpostavljajo tudi razvoj stroke in večje priznavanje pomembnosti analitike in uporabe podatkov za podporo odločitvam (Odić 2017).

7 SKLEP

Ko govorimo o velikih podatkih, moramo biti pozorni na različne uporabe termina. Na splošno je za opisovanje lastnosti velikih podatkov v znanstvenih krogih uveljavljen Laneyev (2001) 3 V model, ki izpostavlja velikost količin, hitrost in raznovrstnost podatkov, temu pa so nato drugi avtorji dodali še različne lastnosti glede na svoje lastno področje raziskav.

Nestrukturirani veliki podatki so zaradi svoje heterogenosti in spremenljivosti zahtevni za analizo. Kljub temu so lahko zelo uporabni za različne namene, kot je pisanje zakonov in posledično zmanjšanje kriminala, izboljševanje zdravstvenih pripomočkov in analizo medicinskih zapisov in slik, predvidevanje in hitrejše reagiranje na naravne katastrofe (Marr 2017). Olajšajo tudi vsakodnevno življenje posameznikov, izjemen potencial pa imajo tudi na področju javnomnenjskih raziskav (Japec in drugi 2015).

Veliki podatki omogočajo hitro sprejemanje (poslovnih) odločitev, ker povečini že obstajajo v neki obliki in zato ni potrebno zbiranje podatkov, pogosto pa so podatki tudi posledica zapisovanja (merjenja) dogodkov v realnem času in so neposredno takrat tudi na voljo (Japec in drugi 2015). Na IJS se jim to pridobivanje podatkov v realnem času zdi največja vrednost velikih nestrukturiranih podatkov (Žnidaršič 2017).

Analiza nestrukturiranih velikih podatkov združuje uporabo velikih količin nestrukturiranih podatkov in napredne analitike (Russom 2011). V praksi se na področju nestrukturiranih velikih podatkov na IJS najpogosteje uporablja analiza besedil. Podatki se lahko analizirajo z različnimi programskimi orodji, med katerimi sta najbolj široko uporabljena Hadoop in MapReduce, na IJS pa v te namene uporabljajo predvsem svojo programsko opremo in knjižnice v Pythonu (od katerih nekatere sicer tudi sledijo principom MapRedce) (Žnidaršič 2017).

Na IJS poudarjajo, da je še pred pridobivanjem podatkov bistveno razumevanje problema in razmislek, ali lahko sploh z določenim tipom podatkov uspemo raziskati naš problem. Tudi pri zbiranju je pomembno imeti v mislih računalniško opremo in drugo infrastrukturo, ki jo bomo potrebovali za shranjevanje in procesiranje podatkov, kar je redko poudarjeno v literaturi (Žnidaršič 2017). Postopku analize nestrukturiranih velikih podatkov se tudi na splošno v večini pregledanih virov posvečajo šele od faze generiranja

oziroma pridobivanja podatkov naprej, zelo malo besed pa je namenjenih fazam, ki potekajo na začetku načrtovanja projekta in so prav tako (če ne še bolj) pomembne za končne rezultate analize in aplikacijo na raziskovalni problem.

Izzivi se pri procesu analize nestrukturiranih podatkov pojavljajo že v fazi generiranja podatkov, saj gre za podatke, ki niso bili ustvarjeni z instrumenti in metodami, posledično pa sta njihova kvaliteta in tudi avtentičnost vprašljivi. Obstajajo namreč že tako dobri avtomatski sistemi generiranja besedil in drugih tipov reklam, da je včasih težko ločiti, ali je neko sporočilo avtomatsko generirano ali ne. Ta problem se bo po mnenju IJS v prihodnosti lahko še povečal, saj bodo na voljo vedno boljša orodja, sistemi in pristopi za delo z nestrukturiranimi podatki (Žnidaršič 2017).

Čeprav je velik del procesa analize podatkov avtomatiziran, je treba nekatere dele opraviti po dolgi poti in ročno. Eden izmed takih korakov je klasificiranje podatkov, pri katerem je treba ročno označiti nekaj tisoč besedil, če gre za pozitivno/negativno ali druge lastnosti. Najbolj zahtevna faza pri procesiranju podatkov je odvisna od problema, največ časa pa vzame faza čiščenja podatkov. Zahtevno je lahko tudi raziskovanje specifičnega področja, ki nas zanima, zato je potrebno v fazi načrtovanja sodelovati s strokovnjaki, ki se na specifike spoznajo (Žnidaršič 2017).

V praksi se izzivi pojavljajo na področju hitrega shranjevanja velikih količin podatkov in hitrega priklica podatkov. Odločanje o tem, katere metapodatke želimo dodatno shranjevati poleg osnovnega podatka, da pridobimo vse potrebne informacije, predstavlja izziv, ki je redko omenjen, je pa pomemben za načrtovanje potrebne programske opreme in pomnilnika. Drug primer iz prakse, ki po navadi ni posebej izpostavljen, pa predstavlja vzdrževanje neprekinjenega podatkovnega toka, ki je zelo pomembno za kvaliteto končnih zbranih podatkov kot celote (Žnidaršič 2017).

Predpostavljala sem, da sta za analizo nestrukturiranih velikih podatkov potrebna specifično znanje in kader, kar bi lahko bil eden izmed glavnih razlogov za redko uporabo nestrukturiranih velikih podatkov na slovenskem trgu. Izkazalo se je, da so glavni razlogi za redko uporabo nestrukturiranih podatkov predvsem v tem, da potencialni uporabniki (še) ne vidijo potrebe po opravljanju tega tipa analiz oziroma se jim zdi razmerje med investicijo in vrednostjo nesprejemljivo (Klenovšek Podobnik 2017; Odić 2017; Žnidaršič 2017). Na slovenskem trgu je tako analiza nestrukturiranih velikih podatkov za zdaj rezervirana le za tiste, ki za določeno raziskavo potrebujejo prav ta tip podatkov. S

poenostavljanjem pristopov in programskih orodij v prihodnosti bo tovrstna analiza lahko postala bolj prijazna za širši krog potencialnih uporabnikov.

8 LITERATURA

1. Aggarwal, Charu C. 2013. *Managing and Mining Sensor Data*. Dordrecht: Kluwer Academic Publishers.
2. Assuncao, Marcos D., Rodrigo N. Calheiros, Silvia Bianchi, Marco A. S. Netto in Rajkumar Buyya. 2014. Big Data Computing and Clouds: Trends and Future Directions. *Journal of Parallel and Distributed Computing* (79–80): 3–15.
3. Baesens, Bart, Ravi Bapna, James R. Marsden, Jan Vanthienen in J. Leon Zhao. 2016. Transformational issues of big data and analytics in networked business. *MIS Quarterly* 40 (4): 807–818.
4. Bakshi, Kapil. 2012. *Considerations for Big Data: Architecture and Approach*. Dostopno prek: http://ieeexplore.ieee.org.nukweb.nuk.uni-lj.si/xpls/icp.jsp?arnumber=6187357 (30. avgust 2017).
5. Cai, Li in Yangyong Zu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal* 14 (2): 1–10.
6. Chen, Min, Shiwen Mao in Yunhao Liu. 2014. Big Data: A Survey. *Mobile Networks and Applications* 19 (2):171–209.
7. Cuzzocrea, Alfredo, Il-Yeol Song in Karen C. Davis. 2011. *Analytics over Large-Scale Multidimensional Data:The Big Data Revolution!* Dostopno prek: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.706.5675&rep=rep1&type=pdf (30. avgust 2017).
8. Das, T. K. in P. Mohan Kumar. 2013. BIG Data Analytics: A Framework for Unstructured Data Analysis. *International Journal of Engineering and Technology* 5 (1): 153–156.
9. De Mauro, Andrea, Marco Greco in Michele Grimaldi. 2014. *What is Big Data? A Consensual Definition and a Review of Key Research Topics.* Dostopno prek: https://www.researchgate.net/publication/265775800_What_is_Big_Data_A_Consensual_Definition_and_a_Review_of_Key_Research_Topics?channel=doi&linkId=54e61d170cf277664ff2f0b4&showFulltext=true (30. avgust 2017).
10. Gantz, J. in E. Reinsel. 2011. *Extracting Value from Chaos. IDC's Digital Universe Study*. Dostopno prek: https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (30. avgust 2017).

11. Institut Jožef Stefan. 2017. *O institutu*. Dostopno prek: https://www.ijs.si/ijsw/V000/IJS (30. avgust 2017).
12. Ipsos. 2017. *About us*. Dostopno prek: https://www.ipsos.com/en/about-us (30. avgust 2017).
13. Japec, Lili, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil in Abe Usher. 2015. *AAPOR Report on Big Data*. Dostopno prek: https://www.aapor.org/getattachment/Education-Resources/Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.aspx (30. avgust 2017).
14. Kaisler, Stephen, Frank Armour, J. Alberto Espinosa in William Money. 2013. *Big Data: Issues and Challenges Moving Forward*. Dostopno prek: https://www.computer.org/csdl/proceedings/hicss/2013/4892/00/4892a995.pdf (30. avgust 2017).
15. Katal, Avita, Mohammad Wazid in R. H. Goudar. 2013. Big Data: Issues, Challenges, Tools and Good Practices. Dostopno prek: http://www.stat.purdue.edu/~doerge/BIOINFORM.D/SPRING16/KatalWazidGoudar_2013.pdf (30. avgust 2017).
16. Klenovšek Podobnik, Mojca. 2017. Intervju z avtorico. Ljubljana, 27. julij.
17. Laney, Doug. 2001. *Application Delivery Strategies*. META Group. Dostopno prek: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (30. avgust 2017).
18. Marr, Bernard. 2017. The Complete Beginner's Guide To Big Data In 2017. *Forbes*, 14. marec. Dostopno prek: https://www.forbes.com/sites/bernardmarr/2017/03/14/the-complete-beginners-guide-to-big-data-in-2017/#58fafe087365 (30. avgust 2017).
19. Mediana. 2009. *Metode kvalitativnega raziskovanja*. Dostopno prek: http://www.mediana.si/raziskovalne-metode/metode-kvalitativnega-raziskovanja/ (30. avgust 2017).
20. Mesec, Blaž. 1998. *Uvod v kvalitativno raziskovanje v socialnem delu*. Ljubljana: Visoka šola za socialno delo.
21. Odić, Ante. 2017. Intervju z avtorico. Ljubljana, 11. avgust.
22. Outfit7. 2017. *We take fun seriously*. Dostopno prek: https://outfit7.com/about-us/ (30. avgust 2017).

23. Russom, Phillip. 2011. *Big data analytics*. TDWI Best Practices Report, četrto četrtletje 2011. Dostopno prek: ftp://ftp.software.ibm.com/software/tw/Defining_Big_Data_through_3V_v.pdf (30. avgust 2017).
24. Taylor, Christine. 2017. Structured vs. Unstructured Data. *Datamation*, 3. avgust. Dostopno prek: http://www.datamation.com/big-data/structured-vs-unstructured-data.html (30. avgust 2017).
25. Wu, Xindong, Xingquan Zhu, Gong-Qing Wu in Wei Ding. 2014. *Data Mining with Big Data*. Dostopno prek: http://lansainformatics.com/wp-content/plugins/project-mgt/file/upload/pdf/2440Data-mining-with-big-data-pdf.pdf (30. avgust 2017).
26. Žnidaršič, Martin. 2017. Intervju z avtorico. Ljubljana, 10. avgust.