Velepodatki – 5 V-jev v praktičnih primerih

Jure Jeraj[1], Urška Nered[1], Stevanče Nikoloski[1,2]
[1] Result, d. o. o., Celovška 182, Ljubljana
[2] Fakulteta za ekonomijo in informatiko, Univerza v Novem mestu, Na Loko 2, 8000, Novo mesto
jure.jeraj@result.si, urska.nered@result.si, stevance.nikoloski@result.si

Izvleček

Vsak posameznik in podjetje dimenzijo velepodatkov dojema malce drugače. A velepodatki se ne merijo kot 5 diskov podatkov, niti kot 5 sodov podatkov. Tudi pretočnih podatkov ne merimo z litri. Velepodatki so miselnost, ki predstavlja temelje za podatkovno gnano poslovanje. Da bi velepodatke resnično uvedli in izkoristili v poslovanju, jih moramo najprej razumeti. Le tako si lahko od njih obetamo sobivanje posla in tehnologije ter ustvarjanje dodane vrednosti.
V tem članku predstavljamo vse dimenzije velepodatkov (t. i. 5 V) in njihovo rabo osvetljujemo na praktičnih primerih. Predstavljamo širši ekosistem velepodatkov ter temelje za gradnjo okolja za velepodatke.

Ključne besede: velepodatki, obdelava v realnem času, podatkovne platforme, pretočni podatki, računalništvo v oblaku

Big Data – 5 V's in practical cases

Abstract

Every individual and company perceives the dimensions of big data a bit differently. Big data is not measured as 5 hard drives nor as 5 data barrels, neither is flow of data measured in litres. Big data is a mindset that forms the foundation for data-driven business. For mass data to be truly introduced to and used in business, one must first understand it. Only then can we expect business and technology to co-exist and deliver added value. In this article, I present all dimensions of big data (i.e., 5 V's) and shed light on its use in practical examples. I also present the broader ecosystem of big data and the foundations for building an environment for big data.

Keywords: Big data, real-time computing, data platforms, data streaming, cloud computing

1 UVOD

Izraz velepodatki (ang. Big Data) je v današnjem času zelo zanimiv in privlačen, a hkrati izredno zavajajoč. Nakazuje namreč na nekaj modernega (»Big« in še »Data«), hkrati pa namiguje, da imamo opravka zgolj z izrazito veliko količino podatkov. Vendar se velepodatki ne merijo zgolj s predponami tera-, peta-, eksa- in podobnimi, velika količina podatkov je le ena izmed njihovih lastnosti. So namreč še mnogo več.

V obstoječi literaturi lahko najdemo veliko različnih definicij koncepta »velepodatki«, ki jih raziskovalci in praktiki (inženirji, razvijalci) uporabljajo v svojih razlagah. Na primer, Cox & Ellsworth (1997) sta velepodatke omenila kot obseg oziroma količino znanstvenih podatkov za vizualizacije. Manyika in drugi (2011) definirajo velepodatke kot »količino podatkov, ki presega zmogljivost tehnologije za učinkovito hrambo, vodenje in procesiranje«. Hashem in drugi (2015) pišejo, da so velepodatki nabor tehnik in tehnologij, ki zahtevajo nove oblike integracije za odkrivanje velikih skritih vrednosti iz raznolikih in kompleksnih obsežnih podatkov.

Največkrat izpostavljena osnovna definicija velepodatkov pravi, da gre za kompleksne podatkovne strukture velikega obsega, ki jih ni moč obdelati na klasičen in tradicionalen način (Awanish, 2021). Po-

manjkljivost te definicije je že v načinu primerjave. Kaj posameznik razume kot klasičen in tradicionalen način? Je to nekaj, kar je bilo vrhunec tehnologije pred 10 leti? Pred 5 leti? Včeraj? Tehnologija se namreč izredno hitro razvija, zato preveč posplošena razlaga ni dobrodošla.

Napačno ali pomanjkljivo razumevanje velepodatkov ne spravlja v zadrego le strokovnjakov ob tehničnih razpravah. Precej večja težava je, kadar zaradi napačnega razumevanja podjetje prehitro zavrne implementacijo orodji in tehnik za izrabo velepodatkov v svoj poslovni model ali pa se prehitro zadovolji z zgolj vpeljavo podatkovne analitike na veliki količini podatkov, npr. z obdelavo 100 milijonov zapisov v podatkovni bazi. Kaj pa dejansko delamo z velepodatki? Že od nekdaj podatke obdelujemo in velepodatki v tem pogledu niso izjema. Razlika je v razvoju sodobnih, ekstremnih tehnik, tehnologij, orodij in konceptov za delo z njimi.

Eden izmed glavnih pogojev, da podjetje postane podatkovno gnano podjetje, je, da je uporaba in izraba velepodatkov poslovna odločitev, za katero trdno stoji miselnost organizacije. Brez podatkovno razvojne miselnosti in dobre podatkovne pismenosti bo imelo podjetje oziroma organizacija ali skupnost velike težave pri uspešni izrabi velepodatkov. Odločitev o uvedbi tehnologij za obdelavo velepodatkov ni tehnična, temveč predvsem poslovna odločitev.

Poučen primer take miselnosti je projekt LoginEko fundacije Login (Login5 Foundation, 2022). Glavni cilj fundacije je »oblikovati prihodnost kmetijstva nove generacije«. Poslovni model predvideva »vzpostavitev novega modela trajnostnega eko kmetijstva velikih razsežnosti«. Temelji na konceptu podatkovno gnanega poslovanja. Snovalci pojasnjujejo, da »podatkovno gnano kmetijstvo« zanje pomeni:

- zajem podatkov iz vseh možnih senzorjev,
- aktualne posnetke vseh polj s pomočjo brezpilotnih letalnikov (dronov),
- zbiranje podatkov mehanizacije v realnem času,
- napredne vremenske postaje na vseh mikrolokacijah,
- sledenje mikroflore tal,
- sledenje vsem kmetijskim dejavnostim,
- centralni informacijski sistem,
- odločanje in učenje na podlagi zbranih podatkov.

To je primer pristopa, ki v celoti podpira uvajanje tehnologij za obdelavo velepodatkov. Pri tem se tehnologije ne gleda zgolj kot strošek, temveč predvsem kot rešitev, ki ustvarja dodano vrednost.

2 VELEPODATKI, OPREDELJENI KOT SKUPEK 5 V-JEV

Podjetja z visoko podatkovno zrelostjo in razvito podatkovno kulturo so se ob prelomu stoletja začela aktivno ukvarjati s snovanjem poenostavljenega razumevanja koncepta »velepodatki«, ki bi ga lahko vpeljali v korporativno podatkovno upravljanje. Eden prvih, Doug Laney iz Meta Group, ki je del Gartnerjeve korporacije, je leta 2001 predstavil koncept velepodatkov kot arhitekturo 3 V-jev (Laney, 2001), in sicer:

- količina oziroma obseg podatkov (ang. volume),
- hitrost prevzemanja, obdelave in posredovanja podatkov (ang. velocity),
- raznolikost v smislu različne strukturiranosti, frekvence in formata podatkov (ang. variety).

A razvoj je pokazal, da ti trije V-ji niso dovolj, saj se je začel pojavljati dvom v zanesljivost zajetih podatkov. V podjetju IBM so naredili raziskavo več različnih podatkovno gnanih podjetjih. Raziskava je pokazala, da 27 % anketirancev vlaga 3,1 trilijon ameriških dolarjev na leto v zagotavljanje ustrezne kakovosti zbranih podatkov ter zanesljivost podatkovnih analiz (IBM, 2014). Glavna ugotovitev je torej bila, da je treba definicijo 3 V-jev nadgraditi z novim, četrtim V-jem (Van Rijmenam, 2014; Allen, 2016):

- verodostojnost podatkov (ang. veracity)
  Z izboljšano podatkovno infrastrukturo ter večjo kakovostjo podatkov in podatkovnih cevovodov so v zadnjem desetletju podatkovno gnana podjetja veliko pridobila z uvedbo koncepta velepodatkov. Tako paradigma velepodatkov pridobiva novo dimenzijo. Sestavni del koncepta zgoraj opisanih 4 V-jev je tako postala tudi:
- vrednost (ang. value).
  V strokovni literaturi se poleg vrednosti pojavljajo še dodatne dimenzije, kot sta na primer variabilnost (ang. variability) in vizualizacija (ang. visualization) (Van Rijmenam, 2013). Ne glede na to, koncept 5 V-jev trenutno predstavlja pot do razumevanja maksimalne in pravilne izrabe velepodatkov (Alabi, 2020).

2.1 Ekosistem velepodatkov

Vsak V zase je preprosto razumeti in o njem lahko razpredamo že z osnovnimi izkušnjami iz sveta in-

formacijske tehnologije. Marsikateri klasični informacijski sistem je možno prikazati kot popoln sistem velepodatkov. A bistvo velepodatkov se skriva v poskusih implementacije s sodobnimi informacijskimi sistemi (sistem v več oblačnih okoljih, dogodkovno orientirane arhitekture …) in z njimi povezanim doseganjem tehnoloških in konkurenčnih prednosti.

Velepodatki gredo z roko v roki z dvema sodobnima tehnikama:

- internetom stvari (ang. IoT – Internet of Things) in
- umetno inteligenco/strojnim učenjem (ang. AI – Artificial Intelligence / ML – Machine Learning).

Oboje pa obkroža še ena ključna dimenzija:

- realni čas (ang. Real Time).

Za lažje razumevanje velepodatkov in njihovo umestitev v sistem si poglejmo tri praktične primere iz realnega življenja:

1. Sprejemanje odločitev ob koncu prvega polčasa nogometne tekme

V skladu s sporazumom, sklenjenim z UEFO za obdobje 2021–2024 glede Lige prvakov ter Evropskim nogometnim pokalom v ženski kategoriji (Carp, 2022), se je morala Mednarodna korporacija za dostavo hrane Just Eat (oziroma hčerinsko podjetje La Nevera Roja) odločiti o spletni oglasni kampanji ob nastopu polčasa nogometne tekme Lige prvakov glede na poročila o prodaji med prvim polčasom tekme v omenjenem tekmovanju.

Za odločevalce je v danem trenutku pomembna predvsem hitrost pridobivanja, shranjevanja in obdelave podatkov. Direktor prodaje ni potreboval le podatkov o prodaji, temveč je za hitro reagiranje moral imeti tudi predlog optimalnega oglaševanja s ključnimi besedami preko sistema Google AdWords. Poleg tega je moral obdelanim podatkom tudi zaupati (BBVA Communications, 2021). Direktorju podatki naslednji dan ne bi pomagali; takrat bi lahko le ugotovil, kaj bi lahko storil (bolje).

2. Celovit sistem pametnih semaforjev

Pametne semaforje poznamo že dolgo, novo dimenzijo pa prinaša sodelovanje celotne infrastrukture za upravljanje semaforjev. V prometu namreč ni vsako (semaforizirano) križišče ločen otoček, temveč tvorijo del širše prometne infrastrukture. Poleg tega tudi vplivajo en na drugega. (Al Nuaimi in drugi, 2015). Za optimalno delovanje mora sistem pridobiti čimbolj celovito in objektivno sliko prometnega stanja na širšem področju (npr. na področju mesta Ljubljana).

Možni viri podatkov so senzorji na semaforjih ter že vgrajena senzorika v cestni infrastrukturi. Možen vir podatkov so lahko tudi vozila javnega prometa, kjer sta natančno znana njihov položaj in hitrost premikanja. Dodaten in pomožen vir podatkov lahko predstavljajo tudi podatki iz storitve Google Zemljevidi. Nenazadnje pa so potencialni vir podatkov kar vsa vozila v prometu oziroma vsa vozila, ki so skladna z novim standardom V2I (ang. Vehicle To Infrastructure).

Na podlagi celotne slike mora sistem pravilno upravljati semaforje, in sicer tako, da ti v najboljši meri optimalno sodelujejo skupaj in dosežejo zadani cilj – kar največjo in hkrati uravnoteženo pretočnost prometa. Dejansko to pomeni izvajanje ukrepov, kot so:

- preprečitev blokiranja križišča,
- čim hitreje sprostiti promet v naslednjem križišču ob zaznavi blokade predhodnega križišča,
- preventivna zaustavitev prometa na začetku vpadnic s ciljem preprečevanja večje zgostitve prometa na koncu vpadnice,
- podobni ukrepi.

Predvsem mora celovit sistem pametnih semaforjev posnemati delo policistov na vsakem križišču – policistov, ki so med seboj povezani (s komunikacijsko povezavo) in imajo dovolj podatkov za odločanje. Policist namreč s svojim vidnim zaznavanjem in odločanjem preprečuje primere zgoraj opisanih neželenih prometnih zamaškov.

3. Odprti podatki skupnosti

Skupnost ima s svojo javno infrastrukturo in javnimi službami (glej tudi prejšnji primer) veliko podatkov. Te podatke lahko – ob zagotovitvi vseh varnostnih standardov – preda v uporabo širši javnosti, ki lahko nato uporabi v namene raziskav in razvoja. (Al Nuaimi in drugi, 2015). Tudi v tem primeru lahko prepoznamo večino V-jev. Dodana vrednost takega pristopa se skriva v povečani intelektualni moči uporabe podatkov, skrajša pa se tudi reakcijski čas ob kriznih situacijah, ker so podatki in infrastruktura deležnikom že na voljo.

2.2 5 V-jev v praksi

Za še boljše razumevanje bomo 5 V-jev pogledali z vidika ideje pametnega semaforja.

1. Volumen

Podatke pridobivamo iz namenskih senzorjev in klasične IoT infrastrukture. Namenske senzorje bi namestili v vse semaforje; za zanesljivost delovanja morajo biti semaforji samozadostni in morajo smiselno delovati tudi, če se iz kateregakoli razloga prekine dotok informacij iz vseh posrednih virov. Že v tem koraku si lahko predstavljamo število križišč, število semaforjev v križiščih, število senzorjev v vsakem semaforju, s katerimi se pokrijejo vse točke križišča, ter predvsem praktično zvezno spremljanje stanja. V takem okolju ni dovolj zajem enega podatka na sekundo; prej govorimo o 100 zajemih na sekundo, kar je v rangu zajema žive slike.

Še bolj pa k volumnu prispevajo vse že obstoječe naprave, ki proizvajajo in imajo možnost deljenja podatkov (kar bi lahko poimenovali celotna posredna IoT infrastruktura). Razlika od prejšnjega dela je, da je ta infrastruktura postavljena iz drugih razlogov ali namenov, vendar se jo vseeno lahko uporabi kot uporabljivi vir za maksimalno in optimalno rešitev.

Po podatkih iz leta 2013 je v Ljubljani delovalo 266 semaforjev (Pandur, 2013). Če predpostavimo uporabo treh senzorjev na posamezen semafor ter 100 zajemov na sekundo, pridelamo v enem dnevu skoraj 7 milijard senzorskih podatkov.

Internet stvari dejansko prinaša veliko razliko, saj si ljudje težko predstavljamo 7 milijard transakcij dnevno. Še posebej, če gre za npr. trgovsko podjetje.

2. Hitrost

Tu dilem ni − semafor mora reagirati hipno. Bistveno je, da so vsi podatki na voljo takoj oziroma v realnem času, ki se meri v milisekundah. To je možno doseči na dogodkovno vodeni arhitekturi, kjer sprožitev neke akcije simultano omogoči pošiljanje podatka v centralni sistem.

3. Raznolikost

Deloma je raznolikost že opisana v sklopu volumna. Bi pa poudarili dve podvrsti raznolikosti. Prva in osnovna je, da lahko podatek o stanju križišča pridobimo iz različnih virov. Pri tem imamo lahko naslednje tipe podatkov:

- strukturirane podatke (npr. namenski senzorji bodo pošiljali zelo strukturirane podatke, saj so izdelani izključno za ta namen),
- polstrukturirane podatke (podatki storitve Google Zemljevidi so lahko deloma strukturirani, saj nimajo natančno identificiranega križišča, temveč se do teh podatkov pride posredno – preko geolokacije),
- nestrukturirane podatke (zajem iz videokamer že postavljenega cestnega nadzora ali v skrajnem primeru iz deljenja slike kamer v vozilih).

Taka opredelitev je daleč najbolj pogosta. Pri tej pa sicer vidim eno pomanjkljivost: lahko nas namreč kaj prehitro zadovolji, če imamo različno strukturirane podatke. Zato podajamo še drugačen pogled – lahko bi rekli kar podzvrst.

Raznolikost lahko namreč razumemo tudi z vidika dogodka, ki ga preučujemo. O njem želimo zbrati čim več raznolikih podatkov iz različnih virov. V našem primeru je preučevani dogodek lahko vozilo v križišču. Ta podatek lahko dobimo iz različnih virov, idealno medsebojno čim bolj neodvisnih.

4. Verodostojnost

Prejšnji trije elementi so glavni argument, zakaj je naslednji V prav verodostojnost. Zaradi količine, hitrosti zajemanja in raznolikosti podatkov se porodi vprašanje o verodostojnosti oziroma zaupanju v podatke. Sprva si lahko predstavljamo odločevalca pred množico grafov in tabel, ki dvomi v podatke in bo vse še dodatno preveril.

V primeru podatkov o pretočnosti križišč lahko dobimo podatke o hitrosti vozil skozi posamezno križišče. Pri tem lahko dobimo manjkajoče in nasprotujoče si podatke. Na primer, da je povprečna hitrost 20 km/h, največja pa 320 km/h. Storitev Google Zemljevidi z dodatno storitvijo Promet nam lahko prikaže, da je križišče zablokirano brez pretočnosti, senzor pa prešteje 100 vozil v minuti.

S človeškim vidom bi si seveda hitro ustvarili pravo sliko in bi lahko ustrezno prečistili podatke oziroma pravilno ukrepali. A v našem primeru to ni možno zaradi količine križišč, njihove medsebojne povezanosti ter reakcijskega časa. V našem ekstremnem primeru nimamo niti realnih možnosti napisati klasičnih algoritmov na način »če to, potem ono«, ker je kombinacij in možnosti preprosto preveč.

Edini pravi odgovor lahko ponudijo storitve umetne inteligence oziroma inteligentne uporabe podatkov. Gre za izredno kompleksne rešitve. Posledično je velikokrat ravno ta točka ključna, zakaj podjetje še vedno nima uvedene rešitve, kot npr. naš primer, pa čeprav je izredno preprosto razložljiv.

5. Vrednost

Vrednost je pogosto področje, ki se ga ob uvajanju rešitev nad velepodatki spregleda. Prejšnje štiri točke so tehnično izredno napredne, posledično pa tudi izredno zanimive za tehnični kader, kar lahko razvijalce potegne v razvoj pretiranih rešitev. Hkrati pa velja tudi obratno – ker ne znamo pravilno izračunati vrednosti, ki nam jo lahko neka rešitev prinese, razumemo vse prejšnje V-je kot strošek namesto kot naložbo.

Pri celovitih pametnih semaforjih vrednost ni zgolj odprava nepotrebnega časa v čakanju v zastojih, temveč je tu še ogljični odtis vozil, ki čakajo v križiščih. Če se tega ne da izračunati, je večja težava v prepoznavi vrednosti v odpravi slabe volje in posledično vseh negativnih posledic vseh v prometu čakajočih udeležencev.

2.3 Meje uporabe velepodatkov

Čeprav so velepodatki izjemno orodje, ki lahko pomaga pri poslovnih odločitvah podatkovno gnanih podjetij, imajo kljub temu svoje meje (Croft, 2014). Preko definicije 5-V in opredelitve ekosistema najlažje ugotovimo, kje so te meje:

- **Dajanje prednosti korelacijam**. Analitiki podatkov uporabljajo velepodatke za ugotavljanje korelacije (ko je ena spremenljivka povezana z drugo). Vendar pa vse te korelacije niso bistvene ali smiselne. Natančneje, samo zato, ker sta dve spremenljivki korelirani oziroma povezani, še ne pomeni, da med njima obstaja vzročna zveza. Na primer, med letoma 2000 in 2009 sta se podobno zmanjšala število ločitev v ameriški zvezni državi Maine in poraba margarine na prebivalca (National Center for Health Statistics, 2002). Vendar margarina in ločitev nimata veliko skupnega.
- **Napačna vprašanja**. Velike podatke je mogoče uporabiti za ugotavljanje korelacije in vpogledov z neskončno paleto vprašanj. Vendar pa je odvisno od uporabnika, da ugotovi, katera vprašanja so smiselna. Če na koncu dobite pravi odgovor na napačno vprašanje, naredite sebi, svojim strankam in svojemu podjetju zelo drago uslugo.
- **Varnost**. Kot pri mnogih tehnoloških prizadevanjih je tudi analitika velepodatkov nagnjena k zlorabam. Podatki, ki jih posredujete tretji osebi, bi lahko prišli do strank ali konkurentov.
- **Prenosljivost**. Ker se večina podatkov, ki jih potrebujete za analizo, skriva za požarnim zidom ali v zasebnem oblaku, je za učinkovito posredovanje teh podatkov skupini za analitiko potrebno tehnično znanje. Poleg tega je morda težko dosledno prenašati podatke strokovnjakom za ponovno analizo.
- **Nedoslednost pri zbiranju podatkov**. Včasih so orodja, ki jih uporabljamo za zbiranje velepodatkovnih množic, nenatančna. Na primer, Google je znan po svojih popravkih in posodobitvah, ki spreminjajo izkušnjo iskanja na različne načine; današnji rezultati iskanja bodo verjetno drugačni od jutrišnjih.

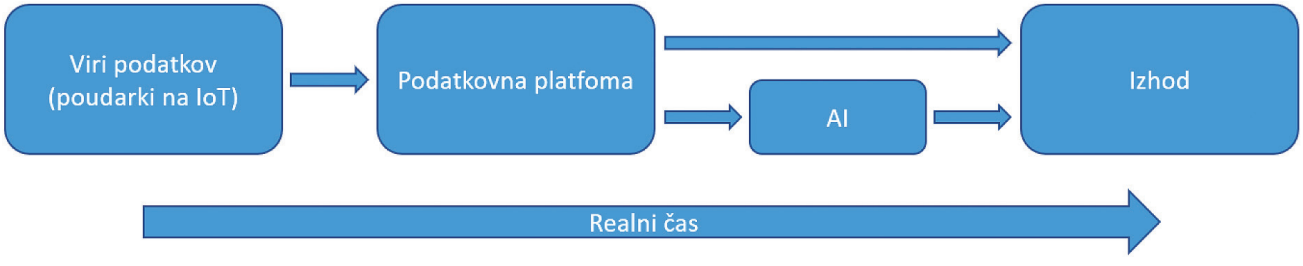
3 PODATKOVNE PLATFORME

Omenili smo že ekosistem velepodatkov: internet stvari ter umetna inteligenca (IoT, AI) v realnem času. Vendar s tem še vedno ne odgovorimo na vprašanje, kaj dejansko implementiramo in kaj je tisto, kar omogoča celotne iniciative velepodatkov. Odgovor so podatkovne platforme.

3.1 Umestitev podatkovnih platform

V spodnji shemi so podatkovne platforme umeščene na najvišjem nivoju.

Zavedati se je treba, da ta shema ni nova, saj obstaja še iz časov pred digitalizacijo poslovanja podjetij. Lahko si predstavljamo, da je podatkovna platforma ustreznica univerzitetne knjižnice, ki hrani ogromno (knjižnega) gradiva, ni pa knjižnica avtor tega gradiva



Slika 1: **Umestitev podatkovnih platform na najvišjem nivoju**.

(ni vir). Knjižnica je hkrati okolje, v katerem se lahko nekaj dela z gradivom (pregleda, izposodi, kopira, obdela …) in na podlagi tega lahko nastane neka (nova) uporabna vsebina ali rezultat (vsebinski izhod).

Podatkovna platforma je torej skupek tehnologij in orodij za zbiranje, shranjevanje in obdelavo podatkov, ki omogoča njihovo uporabo ostalim uporabnikom in orodjem. Ta opredelitev velja za vsa okolja, tudi za okolje velepodatkov. Razlika je le v sestavnih delih (gradnikih) podatkovnih platform.

3.2 Visokonivojska arhitektura podatkovnih platform za velepodatke

Arhitekture so izjemno kompleksne, saj ne obstaja ena arhitektura, ki bi ustrezala vsem potrebam. Podobno kot pri hišah je lahko neka hiša za eno družino idealna, za drugo pa povsem nefunkcionalna. Za dobro arhitekturo je treba imeti celovito znanje o vseh gradnikih podatkovnih struktur, praktične izkušnje na čim bolj reprezentativnih primerih ter sposobnost branja in razumevanja dejanske potrebe konkretnega primera.

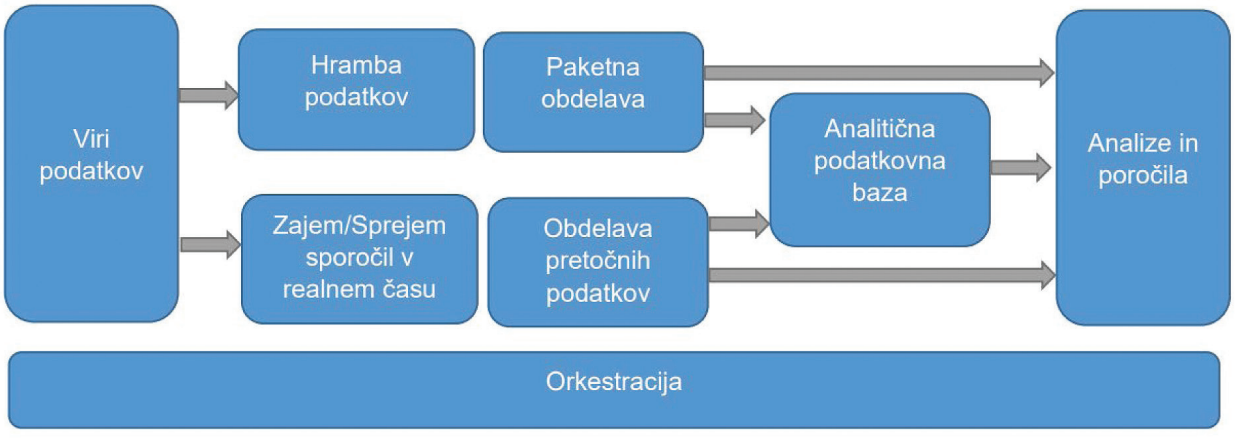
Klasična in še vedno zelo razširjena arhitektura je osredotočena okrog podatkovnega skladišča (ang. Data Warehouse) (Inmon, 2005; Kimball & Ross, 2013). Ko je postalo pomembno, da paketnemu procesiranju dodamo tudi obdelavo podatkov v realnem času, sta se razvili t. i. *Kappa* in *Lambda* arhitekturi (Marz & Warren, 2015; Kreps, 2014). Z nenehnim tehnološkim napredkom in novimi zahtevami se seveda pojavljajo vedno novi arhitekturni vzorci, v katere se v tem članku ne bomo poglabljali, pač pa bomo poskusili nekoliko posplošeno predstaviti različne elemente, ki jih podatkovna platforma lahko ima.

Najbolje je začeti pri osnovah, predvsem pa je vedno priporočljivo iti iz visokonivojske arhitekture v vedno bolj podrobne. Če se v prehodu pripetita nerazumevanje in zmedenost, se je treba vrniti le en korak nazaj in se še nekoliko seznaniti s koncepti tega nivoja arhitekture.

V tej nameri podajamo v spodnji sliki primer visokonivojske arhitekture podatkovnih platform za velepodatke. Predstavljena ideja in diagram sta Microsoftova – sicer pa so visokonivojski koncepti zelo podobni vsepovsod.

Shema prikazuje primer osnovnih visokonivojskih elementov podatkovnih platform.

- **Viri podatkov** (ang. Data Source)
  Brez virov podatkov podatkovne platforme ne morejo obstajati (kot knjižnice ne bi obstajale, če ne bi nihče ustvarjal knjižnega gradiva). Med vire podatkov sodi tudi IoT.
- **Hramba podatkov** (ang. Data Storage)
  V preteklosti so se podatki shranjevali v relacijskih podatkovnih bazah. V platformah za velepodatke ta element nadomešča osnovni datotečni sistem, saj je shranjevanje datotek tako mnogo hitrejše kot zapisovanje v relacijsko bazo. Obstajajo napredni datotečni formati, ki so prilagojeni za velepodatke (npr. Parquet, Orc ipd.). Tak datotečni sitem je primeren tudi za velike binarne datoteke (BLOB formate), kot so slike ali videoposnetki. Tako organizirana shramba podatkov se nato navzven predstavlja kot podatkovno jezero (ang. Data Lake).
- **Paketna obdelava** (ang. Batch Processing)
  Paketno obdelavo si najlažje predstavljamo kot dobro poznani proces ETL. Paketna obdelava po-



Slika 2: **Visokonivojska arhitektura podatkovne platforme. (Microsoft documentation, 2022)**

meni, da v določenih časovnih intervalih (npr. enkrat na dan) obdelamo novo spremenjene zapise ali pa celoten nabor zapisov in jih preoblikujemo v zahtevano končno obliko. Te procese načeloma izvajamo s programskimi okolji za obdelavo velikih količin podatkov (npr. Spark, Databricks ipd.)

- **Zajem/Sprejem sporočil v realnem času** (ang. Real-Time message ingestion)
  Kot smo že omenili, je pri velepodatkih izredno pomemben V tudi hitrost. Največkrat to pomeni, da internet stvari ali dogodkovno orientirana arhitektura (ang. event-driven architecture) omogoča ustvarjanje podatkov v realnem času. V tem primeru mora imeti podatkovna platforma tudi možnost zanesljivega zajema oziroma sprejema teh sporočil/podatkov v realnem času. Najbolj tipičen predstavnik tega elementa je rešitev Apache Kafka.
- **Obdelava pretočnih podatkov** (ang. Stream Processing)
  Ta element je logična posledica prejšnjega in vitalen element za vsak sistem, ki deluje v realnem času. Po zajemu podatkov v realnem času je namreč te treba tudi obdelati in (pred)pripraviti v realnem času. To je povsem nova komponenta, saj nič od obstoječih sistemov ni narejeno na ta način oziroma za ta element. V praksi se največkrat uporablja sisteme Apache Flink ali Apache Spark oziroma njegove nadgradnje/nove generacije Databricks.
- **Analitična podatkovna baza** (ang. Analytical Data Store)
  Gre za bržkone še najbolj tradicionalen element. Tu je mišljeno najbolj osnovno podatkovno skladišče oziroma OLAP-nivo za potrebe klasične podatkovne analitike. Ne glede na ves blišč velepodatkov, klasična podatkovna analitika ne izginja. Še vedno velik del praktične rabe predstavljata klasično analiziranje in predstavitev (vizualizacija) podatkov.
- **Analize in poročila** (ang. Analysis and Reporting)
  Tudi ta element je zelo klasičen. Najlažje ga razumemo kot podatkovno analitiko oziroma obveščanje (ang. Business Intelligence). Zaradi razvoja sodobnih orodij za samopostrežno analitiko pa ta element vendarle ni tako tradicionalen kot prejšnji.
- **Orkestracija** (ang. Orchestration)
  Z dodajanjem novih elementov že v visokonivojsko arhitekturo (sploh pa s številnimi storitvami v dejanski izvedbi) se poveča tudi zahteva po orkestraciji oziroma učinkovitem sodelovanju vseh storitev (ang. Services) med seboj. Zato je ta element izpostavljen in postavljen že kot samostojen v visokonivojski arhitekturi.

V ta del lahko uvrstimo tudi element upravljanja in ravnanja s podatki (ang. Governance), saj se podatki pretakajo čez mnogo storitev, nivojev ter sistemov. S tem se potenčno poveča možnost nastanka šumov v podatkih. V tem primeru pride do izraza dober sistem upravljanja in ravnanja s podatki.

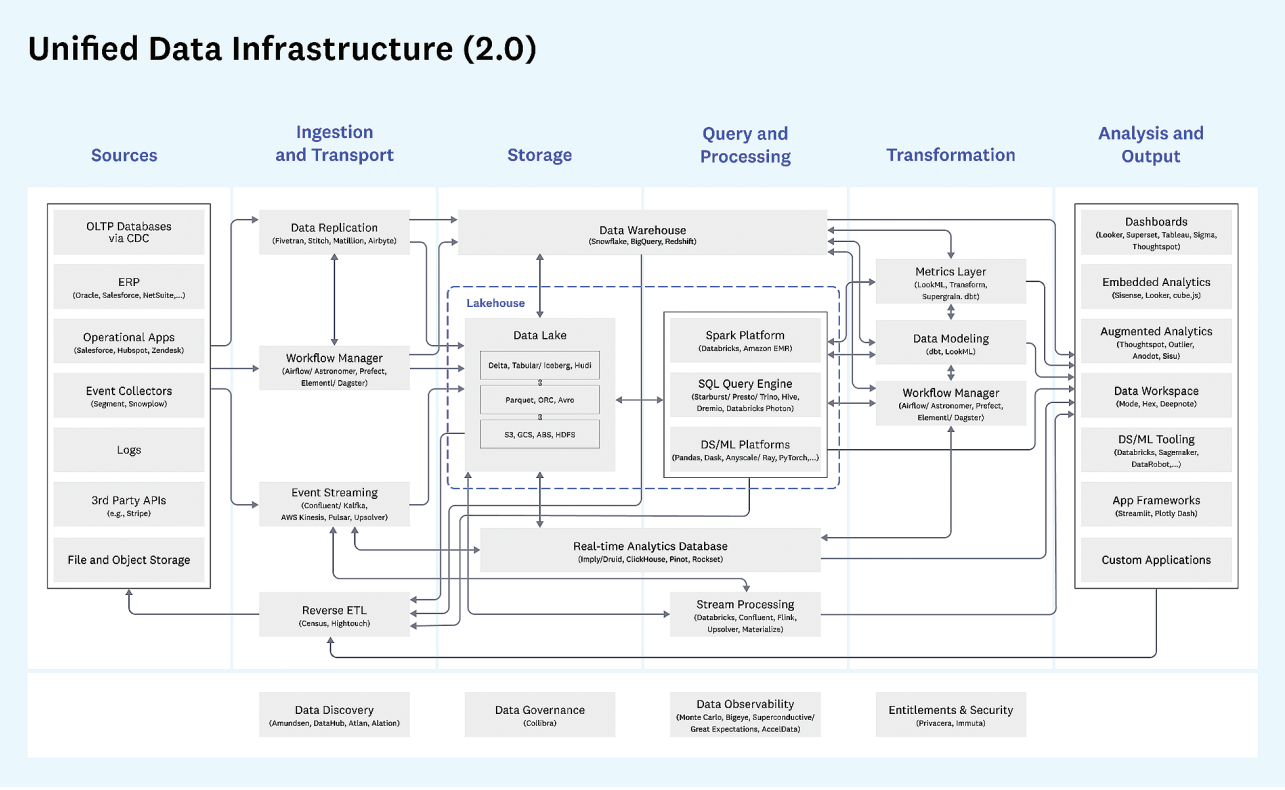
Izkušnje iz klasičnih in tradicionalnih podatkovnih struktur kažejo, da se v bistvu dodajata le dva nova elementa: zajem/sprejem sporočil v realnem času in obdelava pretočnih sporočil. To je seveda povsem razumljivo, saj smo v podpoglavju ekosistema posebej poudarili dimenzijo realnega časa.

3.3 Podrobna arhitektura podatkovnih platform za velepodatke

V visokonivojski arhitekturi pravzaprav nismo naredili velikega preskoka, kar je seveda prav, saj smo izpostavili, da je treba korakati postopoma. Tako stališče pa ne velja za podrobno arhitekturo. Tu se ne dodajo zgolj posamezne storitve, temveč se zelo spremenijo tudi storitve za izvedbo tradicionalnih elementov; podobno kot so se ob pojavu avtomobilov na prelomu 20. stoletja spremenila celotna infrastruktura in pravila. Avtomobili se niso le dodali kočijam, ampak so se morale tudi kočije prilagoditi novim časom.

Pri podrobnih arhitekturah velja enako kot pri visokonivojskih – ni le ene različice resnice. Vendarle pa večina snovalcev stremi k uveljavljenim konceptom in terminologiji, ki pa se na podrobnem nivoju drastično spreminja glede na tradicionalne sisteme.

Za osnovo razlage koncepta sodobnih podatkovnih platform tokrat izhajamo iz ideje, ki jo je podprl članek na portalu Andreesen Horowitz in je prikazana na spodnji sliki:



Slika 3: **Primer podrobne arhitekture podatkovne platforma za velepodatke. (Bornstein, Li, & Casado, 2020)**

Predstavljena arhitektura je neodvisna od velikih proizvajalcev rešitev in ponudnikov storitev (AWS, MS Azure, Google, Oracle …), pri njeni zasnovi pa so sodelovali številni inženirji posameznih predstavnikov rešitev.

Pri tej shemi je zelo ilustrativen način prikaza arhitekture, razdeljene na sklope (stolpce), ki jih lahko uskladimo s prej predstavljenimi visokonivojskimi elementi. V sklopih so našteti ključni elementi ter vodilna orodja oziroma ogrodja, s katerimi je možno izvesti te elemente. Za podroben opis vseh bi potrebovali nov (ali daljši) prispevek, vseeno pa je smiselno izpostaviti tiste, ki prinašajo drastične spremembe:

- **Zajem sprememb v relacijskih bazah** (ang. OLTP Databases via CDC)
  Ta blok pretvarja klasične relacijske podatkovne baze v dogodkovno orientirano arhitekturo. Izraz CDC namreč pomeni zajem sprememb podatkov (ang. *Change Data Capture*), kar pomeni, da lahko vsaka sprememba v relacijski bazi sproži dogodek, ki se ga posreduje na sprejem sporočil v realnem času. S tem blokom lahko vsako obstoječo tradicionalno programsko rešitev precej enostavno pretvorimo v sodobno dogodkovno usmerjeno rešitev. Dobro izpostavljeno orodje za to ponuja Oracle z rešitvijo Oracle Golden Gate, priljubljena odprtokodna alternativa pa je Debezium.
- **Upravljalec podatkovnih operacij** (ang. Workflow Manager)
  Ta blok je neposredna rešitev orkestracije iz visokonivojske arhitekture. Izpostavili smo že, da je orkestracija pomemben element. Obstajajo vgrajena ali neodvisna orodja, katerih glavna naloga je orkestracija vsega dogajanja. Zelo prodorno orodje v tem sklopu je Apache Airflow.
- **Modeliranje podatkov in upravljanje z meta podatki** (ang. Data Discovery, Data Governance)
  Ob razumevanju petih V-jev vemo, da imamo veliko podatkov, ki so hkrati tudi zelo raznoliki. Razumevanje podatkov in upravljanje z njimi je za velike sisteme precejšnje breme, zato sta tu enakovredno poudarjena bloka podatkovnega modeliranja in upravljanja z meta podatki.
- **Podatkovno jezero** (ang. Data Lake)
  Podatkovno jezero je verjetno najtežje razumljiv element te podatkovne strukture. Težava je v idej-

ni opredelitvi, ki se posplošeno nanaša le na kopijo surovih podatkov iz operativnih sistemov. Zato je velikokrat poenostavljeno enačeno kot področje priprave podatkov (ang. staging), kar pravzaprav tudi je (lahko). Vendarle pa je na nivoju podatkovnih platform za velepodatke ta nivo vseeno mišljen za učinkovito shranjevanje vseh podatkov, tudi pol- in nestrukturiranih (spomnimo se na element raznolikosti pri opredelitvi 5-V). Zavedati se moramo, da pri velepodatkih shranjujemo tudi slike, videoposnetke, pomanjkljive podatke, podatke s spreminjajočo se strukturo ali povsem nepoznane podatke in strukture. V tem kontekstu pa se namen podatkovnega jezera loči od področja priprave podatkov (saj se velepodatki ločijo od običajnih podatkov).

- **Podatkovno skladišče** (ang. Data Warehouse)
  Podatkovno skladišče je verjetno daleč najbolj prepoznaven izraz v podani arhitekturi. Kar nas lahko preseneti, je premik od tradicionalnih konceptov do platform velepodatkov. V tradicionalnih sistemih je bila podatkovna platforma enaka podatkovnemu skladišču, sedaj pa je podatkovno skladišče le eden izmed elementov podatkovne strukture.
  To je pa pravzaprav logično, saj že iz visokonivojske arhitekture sledi, da se ne moremo izogniti analitičnim podatkovnih bazam, a te niso več glavni igralec zaradi dimenzije obdelave v realnem času.
- **Kolišče** (ang. Lakehouse)
  Arhitektura kolišča postaja vse bolj prepoznavna, zlasti ko ga je začel podpirati nabor ponudnikov (vključno z AWS, Databricks, Google Cloud, Starbrust in Dremio) ter pionirji podatkovnih skladišč. Temeljna vrednosti kolišča je združiti prednosti podatkovnega jezera in podatkovnega skladišča. Tako dobimo cenovno dostopno shranjevanje podatkov v odprtih formatih, hkrati pa imamo še vedno na voljo napredne funkcionalnosti, npr. ACID transakcije, indeksiranje, verzioniranje podatkov itd. (Armbrust, 2021).
- **Spontane poizvedbe** (ang. Ad Hoc Query Engine)
  Ta element je zanimiv, saj povezuje tradicionalni pristop z ekstremi velepodatkov. Če smo v tradicionalnih sistemih imeli relacijsko bazo, ki je bila na nek način črna škatla, kjer nikjer nismo imeli pravega vpliva na to, kako se bodo podatki shranjevali in kako bomo zares dostopali do njih, imamo zdaj ločene bloke za isto uporabniško izkušnjo. Uporabniki namreč še vedno želijo brskati po podatkih s programskim jezikom SQL. V sodobnih sistemih imamo blok podatkovnega jezera, ki je pravzaprav datotečni sistem s shranjenimi datotekami. Ta blok nam omogoča, da lahko do vsebin teh datotek dostopamo preko jezika SQL. Tako navzven ne spreminjamo uporabniške izkušnje, navznoter pa imamo možnost popolne prilagodljivosti za uporabo velepodatkov.

Katere od vseh elementov, ki jih imamo na izbiro, v arhitekturi svoje podatkovne platforme dejansko uporabimo, je odvisno od naših potreb, pričakovanj in načrtov. Vsekakor lahko začnemo le s klasičnimi elementi, ki jih morda že imamo, nato pa jih ob pravem trenutku nadgradimo z dodatnimi, naprednejšimi.

4 ZAKLJUČEK

Videli smo, da velepodatke težko natančno definiramo, lahko pa jih opredelimo s 5 V-ji (obseg, hitrost, raznolikost, verodostojnost in vrednost). Z razumevanjem vseh teh dimenzij se nam odpre pravi pogled na priložnosti velepodatkov, ki jih mora podatkovno gnano podjetje razumeti, da bi lahko v pravem trenutku še dodatno dvignilo vrednost svojih podatkov in tako povečalo svojo podatkovno zrelost.

Ker je treba tudi tehnično podpreti vse operacije, ki morajo obdelovati vseh 5 V-jev, se v ta namen pojavlja gradnik podatkovne platforme za velepodatke. Ta je skupek tehnologij in orodij za omogočanje klasičnih ter tradicionalnih procesiranj podatkov, mora pa se tudi uspešno spopadati z najbolj ekstremnimi izzivi velepodatkov. Te platforme pa le niso tako zapleten, kot je morda videti na prvi pogled, saj so le naravna evolucija dosedanje poti. Le pogledati jih moramo postopoma iz grobe visokonivojske arhitekture do podrobne logične.

Dejstvo je, da so velepodatki realnost. Predvsem zaradi interneta stvari je podatkov več kot kadarkoli, podatki so mnogo bolj raznoliki, vse več pa je tudi nestrukturiranih podatkov. Ti se ustvarjajo hitreje kot kadarkoli. Pri tem se je treba zavedati, da zgolj proizvedeni in shranjeni podatki predstavljajo predvsem strošek, zato je nujno, da podatke uporabimo tako, da nam prinesejo čim večjo vrednost, saj se le tako smotrno sklene celotna veriga življenjske dobe podatka.

Velepodatki prinašajo popolnoma nove ekstremne dimenzije možnosti in priložnosti obdelave, a

vendar je z dobrim razumevanjem preskok lahko narediti tudi postopoma in z majhnimi koraki. Iz lastnih izkušenj podamo nekaj smernic:

izkoristimo priložnost razumeti ozadja ter jih smiselno vpeljati v našo trenutno realnost in pričakovan prihodnji razvoj, saj bomo tako še uspešnejše transformirali naše poslovanje v podatkovno gnano organizacijo in družbo;

povečajmo zavedanje, da sta podatkovna kultura in zrelost ena izmed načinov, da se razume namen ter korist, ki ga imamo z uvedbo moderne »big data« platforme;

načrtujmo zaposlitev ustreznih kadrov, ki bodo zmožni zgraditi platforme za ravnanje z velepodatki, kot so podatkovni inženirji, znanstveniki in analitiki.

Vsekakor ima uvedba velepodatkov v podjetje pozitivne učinke − če se seveda procesa transformacije lotimo pravilno in po korakih, prilagojenih trenutnemu stanju v podjetju, zmožnostim in potrebam.

VIRI IN LITERATURA

[1] Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. Journal of Internet Services and Applications, 6, 25. doi:10.1186/s13174-015-0041-520

[2] Alabi, M. O. (2020). Big Data, 3D Printing Technology, and Industry of the Future. (I. R. Association, Ured.) Additive Manufacturing: Breakthroughs in Research and Practice, 503-520. doi:10.4018/978-1-5225-9624-0.ch021

[3] Allen, M. (2016). The Challenge of Big Data–It's more than just big files. Pridobljeno 30. oktober 2017 iz Pro2Col: https://pro2col.com/challenge-big-data-more-than-just-big-files

[4] Armbrust, M. et al (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. CIDR '21, Jan. 2021.

[5] Awanish. (19. september 2021). Big Data Tutorial: All You Need To Know About Big Data! Pridobljeno iz Big Data Tutorial: All You Need To Know About Big Data!: https://www.edureka.co/blog/big-data-tutorial

[6] BBVA Communications. (19. september 2021). The five V's of big data. Pridobljeno iz https://www.bbva.com/en/five-vs-big-data/

[7] Bornstein, M., Li, J., & Casado, M. (2020). Emerging Architectures for Modern Data Infrastructure. Pridobljeno iz Future: https://future.com/emerging-architectures-modern-data-infrastructure/

[8] Carp, S. (2022). Uefa's Just Eat sponsorship covers Champions League and Women's Euro. Pridobljeno 9. julij 2022 iz SportsPro SmartSeries: https://smartseries.sportspromedia.com/news/uefa-just-eat-sponsorship-champions-league-womens-euro

[9] Croft, C. (2014). The Limits of Big Data. SAIS Review of International Affairs, 34, 117–120. doi:10.1353/sais.2014.0005

[10] Dehghani, Z. (2019). How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh. Pridobljeno 3. avgust 2022 iz https://martinfowler.com/articles/data-monolith-to-mesh.html

[11] Dice. (2020). Dice Tech Job Report: The Fastest Growing Hubs, Roles and Skills. Pridobljeno iz ISSUE #1: Q1 2020: https://techhub.dice.com/Dice-2020-Tech-Job-Report.html

[12] Edjlali, R., & Beyer, M. (2012). Understanding the Logical Data Warehouse: The Emerging Practice. Gartner Tech. Rep. G00234996.

[13] Hashem, I. A., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of »big data« on cloud computing: Review and open research issues. Information Systems, 47, 98–115. doi:10.1016/j.is.2014.07.006

[14] IBM. (2014). IBM Big Data & Analytics Hub: The Four V's of Big Data. Pridobljeno 26. oktober 2017 iz http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[15] Inmon, W. H. (2005). Building the Data Warehouse 4th Edition. Wiley.

[16] Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: : The Definitive Guide to Dimensional Modeling 3rd Edition. Wiley.

[17] Kreps, J. (2014). I Heart Logs: Event Data, Stream Processing and Data Integration. O'Reilly Media.

[18] Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Delta. Application Delivery Strategies, 1–4.

[19] Linstedt, D., & Olschimke, M. (2015). Building a Scalable Data Warehouse with Data Vault 2.0. Morgan Kaufmann.

[20] Login5 Foundation. (22. June 2022). LoginEKO. Pridobljeno iz https://www.logineko.com/

[21] M. Cox, & D. Ellsworth. (1997). Managing Big Data for Scientific Visualization.

[22] Mansoor, I. (30. June 2022). Netflix Revenue and Usage Statistics (2022). Pridobljeno 8. julij 2022 iz Business of Apps: https://www.businessofapps.com/data/netflix-statistics/

[23] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[24] Marz, N., & Warren, J. (2015). Big Data: Principles and Best Practices of Scalable Realtime Data Systems. Manning.

[25] Microsoft documentation. (2022). Big data architecture style. Pridobljeno 9. julij 2022 iz docs.microsoft.com: https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data

[26] National Center for Health Statistics. (7. februar 2002). National Vital Statistics System. Pridobljeno iz cdc.gov: https://www.cdc.gov/nchs/nvss/marriage-divorce.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fnchs%2Fmardiv.htm

[27] Oracle. (2022). Oracle Cloud Infrastructure (OCI) GoldenGate. Pridobljeno 9. julij 2022 iz https://www.oracle.com/integration/goldengate/

[28] Pandur, S. (2013). Kjer imajo nadzor nad vsemi rdečimi lučmi. Delo.

[29] Priebe, T., Neumaier, S., & Markus, S. (2021). Finding Your Way Through the Jungle of Big Data Architectures. 2021 IEEE International Conference on Big Data (Big Data) (str. 5994–5996). Orlando, FL, USA: IEEE. doi:10.1109/BigData52589.2021.9671862

[30] Van Rijmenam, M. (2013). Why The 3V's Are Not Sufficient To Describe Big Data. Pridobljeno 8. July 2022 iz Datafloq: https://datafloq.com/read/3vs-sufficient-describe-big-data/

[31] Van Rijmenam, M. (2014). Think Bigger: Developing a Successful Big Data Strategy for Your Business. American Management Association.

[32] Wikipedia. (19. september 2021). Pridobljeno iz Big Data: https://en.wikipedia.org/wiki/Big_data
[33] Zaidi, E., Thoo, E., De Simoni, G., & Beyer, M. (2019). »Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration. Gartner, Tech. Rep. G00450706

▪

Jure Jeraj je diplomiral na Fakulteti za računalništvo in informatiko ter opravil specializacijo na Fakulteti za organizacijske vede s področja Organizacije in managementa informacijskih sistemov. Trenutno je zaposlena pri podjetju Result, d.o.o. kot vodja oddelka za podatke in umetno inteligenco. Predvsem v zadnjih 10 letih se osredotoča na informacijske rešitve v povezavi s podatki. Opravljal je naloge arhitekta podatkovnih skladišč in sistemov za poslovno analitiko, naknadno se je specializiral za arhitekturo modernih podatkovnih platform in uporabe velepodatkov. Pridobljeno znanje deli na raznih strokovnih konferencah in ostalih dogodkih. Je tudi programski direktor Foruma podatkovne analitike.

▪

Urška Nered je diplomirala iz fizike na Fakulteti za matematiko in fiziko Univerze v Ljubljani. Med študijem je pridobila izkušnje kot podatkovni analitik pri podjetju Result, d.o.o, kjer se je kasneje tudi zaposlila kot podatkovni inženir. Urška se v svojem delu osredotoča na podatkovne arhitekture ter sodeluje pri številnih domačih in tujih projektih, kjer je prispevala k uspehu in kvaliteti izvedbe.

▪

Stevanče Nikoloski je doktor znanosti na področju informacijske in komunikacijske tehnologije. Pet let je delal na Inštitutu Jožef Stefan v Ljubljani, kot doktorski študent in raziskovalec na oddelku Tehnologije znanja, in sicer na področju umetne inteligence. Svoje raziskovalno delo je objavil v zelo prestižnih revijah in ga predstavil na konferencah. Trenutno je zaposlen kot vodilni podatkovni znanstvenik v podjetju Result, d.o.o. v Ljubljani in je odgovoren za grajenje Data Science kompetenc. Poleg tega uči kot visokošolski učitelj predmeta »Programski inženiring« in »Digitalizacija poslovnih procesov« na Fakulteti za ekonomijo in informatiko na Univerzi v Novem mestu.