MASOVNI PODATKI – 5 V-jev v PRAKTIČNIH PRIMERIH

Jure Jeraj
Result, d.o.o., Celovška 182, Ljubljana
jure.jeraj@result.si

Povzetek

Vsak posameznik in podjetje dimenzijo masovnih podatkov dojema malce drugače. A masovni podatki se ne merijo kot 5 diskov podatkov, niti kot 5 sodov podatkov, niti pretočnih podatkov ne merimo z litri. Masovni podatki so miselnost, ki predstavlja temelje za podatkovno gnano poslovanje. Da bi masovne podatke resnično uvedli in izkoristili v poslovanju, jih moramo najprej razumeti. Šele takrat si lahko od njih obetamo sobivanje posla in tehnologije ter ustvarjanje dodane vrednosti.

V tem članku predstavljam vse dimenzije masovnih podatkov (t. i. 5 V) in njihovo rabo osvetljujem na praktičnih primerih. Predstavljam širši ekosistem masovnih podatkov ter temelje za gradnjo okolja za masovne podatke.

A ni vse le v tehnologiji – podjetja potrebujejo tudi strokovnjake za podatke. To so predvsem podatkovni znanstveniki, podatkovni inženirji in podatkovni analitiki, ki bodo podatke zbrali, obdelali in osmislili ter dali na razpolago poslovnim uporabnikom. Šele takrat se bodo podjetja v praksi šla podatkovno gnano poslovanje.

Abstract

BIG DATA AND DATA-DRIVEN BUSINESS

Every individual and company perceives the dimension of big data a little differently. Big data is not measured as 5 hard drives, nor as 5 data barrels, nor is flow data measured in liters. 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 the business and technology to coexist and deliver added value. In this article, I present all dimensions of big data (i.e. 5 V) and shed light on its use on practical examples. I also present the broader ecosystem of big data and the foundations for building an environment for big data. But it's not all about technology itself, companies also need data experts. Data scientists, data engineers and data analysts are the ones who will collect, process and make sense of the data and make it available to business users. Only then will companies be able to put data-driven business in practice.

Ključne besede

masovni podatki, podatkovne platforme, računalništvo v oblaku, podatkovno gnano poslovanje, internet stvari, pretočni podatki, obdelava v realnem času

Key words

Big Data, data platforms, cloud computing, data-driven business, IoT – internet of things, data streaming, real-time computing

UMESTITEV MASOVNIH PODATKOV

Izraz masovni podatki (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 le z izrazito veliko količino podatkov. Vendar tega ne moremo enačiti, podobno kot masa ni enako teža. Masovni podatki se ne merijo zgolj s predponami tera-, peta-, eksa- in podobnimi. So namreč mnogo več, velika količina podatkov je le ena izmed lastnosti masovnih podatkov.

Napačno ali pomanjkljivo razumevanje masovnih podatkov ne vodi v zadrego le ob tehničnih razpravah. Bistvo večja težava je, kadar zaradi napačnega razumevanja podjetje prehitro zavrne implementacijo masovnih podatkov v svoj poslovni model, ali pa se prehitro zadovolji zgolj z vpeljavo podatkovne analitike na veliki količini podatkov, npr. obdelavo 100 milijonov zapisov v podatkovni bazi.

Sledi vprašanje, kaj sploh je implementacija masovnih podatkov?

Masovnih podatkov namreč ne implementiramo, prav tako jih ne moremo označiti kot projekt masovnih podatkov. Največkrat izpostavljena osnovna definicija masovnih podatkov pravi, da gre za kompleksne podatkovne strukture velikega obsega, ki jih ni moč obdelati na klasičen in tradicionalen način. (Wikipedia, 2021) (Awanish, 2021)

Pomanjkljivost te definicije je že v samem 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, saj lahko pripelje do izjav, kakršna je uvedba masovnih podatkov.

Kaj dejansko delamo z masovni podatki? Že od nekdaj jih obdelujemo. Tudi sedaj, saj podatkov nismo nikoli implementirali, izvajali smo le aktivnosti s podatki. Masovni podatki v tem pogledu niso izjema. Razlika je v razvoju sodobnih, ekstremnih tehnik, tehnologij, orodij in konceptov za delo z masovnimi podatki. Te predstavljam v nadaljevanju.

MASOVNI PODATKI, OPREDELJENI KOT SKUPEK 5 V-JEV

Pot v razumevanje maksimalne in pravilne izrabe masovnih podatkov se začne s pravilno opredelitvijo. V strokovni literaturi so masovni podatki opredeljeni kot skupek 5 V-jev, in sicer:

- *Volume* (volumen, obseg podatkov),
- *Velocity* (hitrost prevzemanja, obdelave in posredovanja podatkov),
- *Variety* (raznolikost podatkov; različni podatki, različne strukturiranosti o istem objektu),
- *Veracity* (verodostojnost podatkov; kako močno lahko zaupamo podatkom),
- *Value* (vrednost podatkov; obdelani podatki nam morajo (do)dati poslovno vrednost).

Čar masovnih podatkov je preplet vseh V-jev. Pri tem niti ni nujno, da je vsak posamezen V ekstremen, vendar prav zmnožek teh »ekstremnosti« predstavlja mejo med podatki in masovnimi podatki.

Ekosistem masovnih podatkov

Vsak V zase je preprosto razumeti in o njem lahko razpredamo z že osnovnimi izkušnjami iz sveta informacijske tehnologije. Dejansko je možno marsikateri klasičen informacijski sistem prikazati kot popoln sistem masovnih podatkov. A čar pravega pomena masovnih podatkov 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.

Masovni podatki 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*).

Preko definicije 5-V in opredelitve ekosistema najlažje ugotovimo, kje so (ekstremne) meje masovnih podatkov.

Trije primeri masovnih podatkov v realnem življenju

1. Sprejemanje odločitev ob koncu prvega polčasa nogometne tekme

   Mednarodna korporacija za dostavo hrane Just Eat (oz. hčerinsko podjetje La Nevera Roja) se je moralo odločiti o spletni oglasni kampanji ob nastopu polčasa nogometne tekme Lige prvakov glede na poročila o prodaji med prvim polčasom tekme v omenjenem tekmovanju.[1] 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. (The five V's of big data, 2021)

   Direktorju podatki ne bi pomagali naslednji dan; 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 eno na drugega. (Al Nuaimi, Al Neyadi, Mohamed, & Al-Jaroodi, 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 sami senzorji na semaforjih ter že vgrajena senzorika v cestni infrastrukturi. Možen vir podatkov so lahko tudi vozila javnega prometa, kjer sta natančno znana njihovo 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 celovite 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.
   - In podobni ukrepi.

   Pod črto mora celovit sistem pametnih semaforjev posnemati delo policistov na vsakem križišču – policistov, ki so med seboj povezani (s komunikacijsko povezavo) in imajo dovolj za odločanje. Policist namreč s svojim vidnim zaznavanjem in odločanjem preprečuje ravno primere zgoraj opisanih neželenih prometnih zamaškov.[1]

3. Odprti podatki skupnosti

   Skupnost s svojo javno infrastrukturo in javnimi službami (glej tudi prejšnji primer) ima veliko podatkov. Te podatke lahko – ob zagotovitvi vseh varnostnih standardov – preda v uporabo širši javnosti, ki lahko nato te podatke uporabi v namene raziskav in razvoja. (Al Nuaimi, Al Neyadi, Mohamed, & Al-Jaroodi, 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.

[1] Kot zanimivost lahko navedem, da se z zelo sorodnim področjem ukvarja slovenski projekt iPOT (https://ipot.si), kjer si lahko z zgornjim primerom lažje predstavljate dejavnost, ideje in cilje tega projekta.

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 vsi že obstoječe naprave, ki proizvajajo in imajo možnost deljenja podatkov (kar bi lahko poimenovali celotna posredna IoT infrastruktura). Razlika od prejšnjega dela je v tem, da je ta infrastruktura postavljena iz drugih razlogov ali namenov, vendar se jo vseeno lahko uporabi kot uporabljivi viri 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 – ljudje si težko predstavljamo 7 milijard transakcij dnevno – npr. v trgovskem podjetju.

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 poudaril dve podvrsti raznolikosti.

   Prva in osnovna je, da lahko podatek o stanju križišča pridobimo iz različnih virov. Pri tem imamo lahko:

   - 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 video kamer že postavljenega cestnega nadzora ali v skrajnem primeru iz deljenja slike kamer v vozilih).

Taka opredelitev je daleč najbolj pogosta. Pri tej pa vidim sicer eno pomanjkljivost; lahko nas namreč kaj prehitro zadovolji in smo prehitro zadovoljni, če imamo različno strukturirane podatke.

Zato podajam še drugačen pogled, lahko bi rekli kar podzvrst. Raznolikost lahko 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 čim manj medsebojno povezanih.

Zakaj je tak pogled dvoplastni podatek pomemben? Ker menim, da je bolj pomembno dobiti npr. podatke iz petih strukturiranih in med seboj povsem nepovezanih sistemov kot pa petih različno strukturiranih podatkov iz istega vira.

4. Verodostojnost

Prejšnji trije elementi so glavni argument, zakaj je naslednji V prav verodostojnost. Zaradi količine, hitrosti zajemanja in raznolikosti podatkov se porodi vprašanje verodostojnosti oziroma zaupanja 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. Npr. 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, 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 niti nimamo 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 strojnega učenja. 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 velikokrat področje, ki se ga ob uvajanju rešitev nad masovnimi podatki 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 še da izračunati, je večja težava v prepoznavi vrednosti v odpravi slabe volje in posledično vseh negativnih posledic vseh v prometu čakajočih udeležencev.

   In ravno ta element lepo opiše, da so masovni podatki pravzaprav miselnost podjetja, organizacije, skupnosti … Miselnost, ki jo lahko opredelimo kot podatkovno razvojno miselnost.

MASOVNI PODATKI KOT PODATKOVNO RAZVOJNA MISELNOST

Brez podatkovno razvojne miselnosti bo imelo podjetje oziroma organizacija ali skupnost velike težave pri uspešni izrabi masovnih podatkov. Odločitev o uvedbi tehnologij za obdelavo masovnih podatkov ni tehnična odločitev, temveč je predvsem poslovna odločitev, za katero trdno stoji miselnost organizacije. Če te ni, bo investicija v tehniko zgolj strošek, v katerega se bo vedno dvomilo, to pa bo (dodatno) zaviralo možnost uvida dodane vrednosti.

Uvedbo tehnologij in predvsem rešitev za obdelavo masovnih podatkov je nemogoče natančno oceniti, saj ne gre za enkratni projekt, ki bi ga lahko natančno in vnaprej opredelili. Gre za miselnost in vizijo.

Poučen primer take miselnosti je projekt LoginEko fundacije Login.[2] 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 (t. i. dronov).
- Zbiranje podatkov mehanizacije v realnem času.
- Napredne vremenske postaje na vseh mikrolokacijah.

[2] Več o fundaciji lahko preberete na http://www.login5.org ter o projektu na: http://www.logineko.com. Velja pa na tem mestu omeniti, da sta fundacija in posledično projekt ustanovljena in vodena v Sloveniji.

- Sledenje mikroflore tal.
- Sledenje vsem kmetijskim dejavnostim.
- Centralni informacijski sistem.
- Odločanje in učenje na podlagi zbranih podatkov.

Omenjena fundacija ima vse našteto javno objavljeno na spletni strani, kar še dodatno potrjuje njeno miselnost.

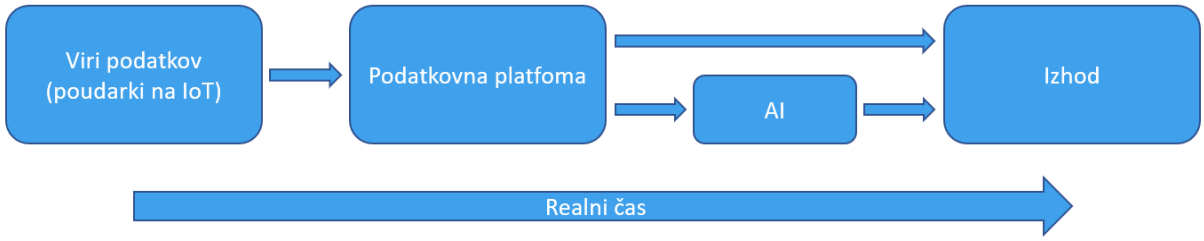
To je eden (in mogoče celo edini) pristop, ki zares podpira uvajanje tehnologij za obdelavo masovnih podatkov. Pri tem se tehnologije ne bo gledalo kot zgolj strošek, temveč predvsem kot rešitev, ki ustvarja dodano vrednost.

PODATKOVNE PLATFORME

Omenili smo že ekosistem masovnih podatkov: 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 masovnih podatkov.

Umestitev podatkovnih platform

Odgovor so podatkovne platforme. Na najvišjem nivoju jih lahko umestimo z naslednjo shemo:



Slika 1: Umestitev podatkovnih platform na najvišjem nivoju.

Zavedati se moramo, da ta shema ni nova, saj obstaja še iz časov pred digitalizacijo poslovanja podjetij. Če znamo res abstraktno razmišljati, je podatkovna platforma ustreznica univerzitetne knjižnice, ki hrani ogromno (knjižnega) gradiva, ni pa knjižnica avtor tega gradiva (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 hramba podatkov ter skupek tehnologij in orodij, ki omogočajo zbiranje, shranjevanje in obdelavo podatkov ter omogočajo njihovo uporabo ostalim

uporabnikom in orodjem. Ta opredelitev velja za vsa okolja, tudi za okolje masovnih podatkov. Razlika je le v sestavnih delih (gradnikih) podatkovnih platform.

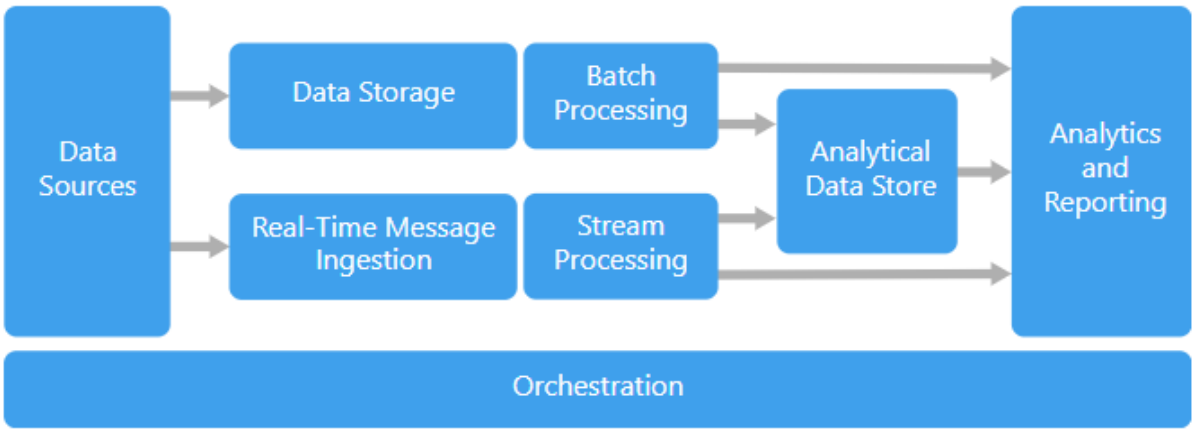
Visokonivojska arhitektura podatkovnih platform za masovne podatke

Arhitekture so izredno nehvaležna stvar, saj zanje ne velja empirična znanost, kjer imamo ali prav ali narobe. Podobno kot pri hišah, ena hiša je lahko za eno družino idealna, taista hiša pa bo za drugo družino povsem nefunkcionalna. Za dobro arhitekturo je treba imeti izredno celovito znanje o vseh gradnikih podatkovnih struktur, dejanske izkušnje na čim bolj reprezentativnih primerih ter sposobnost branja in razumevanja dejanske potrebe konkretnega primera.

Začeti velja 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 podajam primer visokonivojske arhitekture podatkovnih platform za masovne podatke. Predstavljena ideja in diagram sta Microsoftova – sicer pa so visokonivojski koncepti zelo podobni vsepovsod. Več na strani: *https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data*

Izjemoma sem pri arhitekturah pustil tudi izvirne (angleške) nazive. Predvsem zato, ker jih je težko natančno prevesti, in zato, ker se s temi platformami uvajajo novi izrazi in je prav, da jih tudi usvojimo.



Slika 2: Visokonivojska arhitektura podatkovne platforme.

Shema prikazuje primer osnovnih visokonivojskih elementov podatkovnih platform.

- Viri podatkov (ang. *Data Source*)

Brez virov podatkov podatkovne platforme ne morejo obstajati (kot niti knjižnice ne bi obstajale, če ne bi niče ustvarjal knjižnih gradiv). Med vir podatkov sodi tudi IoT.

- Hramba podatkov (ang. *Data Storage*)

  V preteklosti so se podatki shranjevali v relacijskih podatkovnih bazah. V platformah za masovne podatke 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 masovne podatke (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 pomeni, 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 jeziki ali programskimi okolji za obdelavo velikih količin podatkov (npr. Java, Scala, Python; Spark, Databricks ipd.)

- Zajem/Sprejem sporočil v realnem času (ang. *Real-Time message ingestion*)

  Kot smo se že seznanili, je pri masovnih podatkih izredno pomembna dimenzija 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 podatkovna platforma imeti 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 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 oz. OLAP-nivo za potrebe klasične podatkovne analitike. Ne glede na ves blišč masovnih podatkov klasična podatkovna analitike 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 oz. 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 oz. učinkovitem sodelovanju vseh storitev (ang. *Services*) med seboj. Zato je ta element izpostavljen in postavljen že kot samostojen element v visokonivojski arhitekturi.

  V ta del lahko mirno 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.

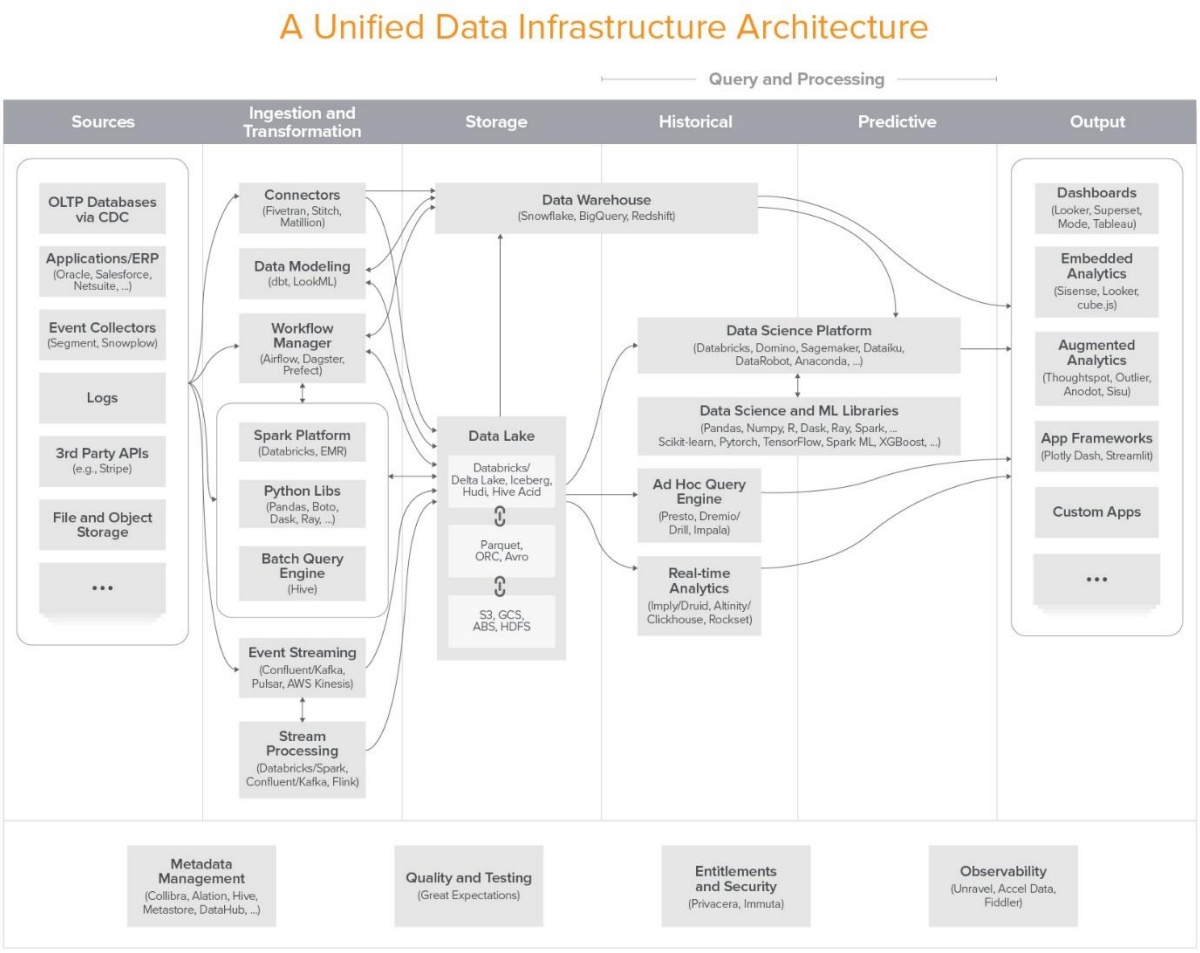
Izkušnje iz klasičnih in tradicionalnih podatkovnih struktur kažejo, da se v bistvu dodajata le dva nova elementa: zajem/sprejem sporočil v realnem času in obdelava pretočnih sporočil. Kar je seveda povsem razumljivo, saj smo v podpoglavju ekosistema posebej poudarili dimenzijo realnega časa.

Podrobna arhitektura podatkovnih platform za masovne podatke

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 se je 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. Ta pa se na podrobnem nivoju drastično spreminja glede na tradicionalne sisteme.

Za osnovo razlage koncepta sodobnih podatkovnih platform tokrat izhajam iz ideje, ki jo je podprl članek na portalu Andreesen Horowitz. Več o tem na strani: *https://a16z.com/2020/10/15/the-emerging-architectures-for-modern-data-infrastructure/*



Slika 3: Primer podrobne arhitekture podatkovne platforma za masovne podatke.

Pri tej shemi je zelo ilustrativen način prikaza arhitekture, ki je razdeljen na nivoje (stolpce), na posamezne elemente v teh nivojih ter na vodilna orodja oz. ogrodja, s katerimi je možno izvesti te elemente.

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.

Za vmesni korak k razumevanju arhitekture se poglobimo na nivoje, ki jih ta arhitektura prinaša:

- Viri (ang. *Sources*)

  Enak nivo kot pri visokonivojski arhitekturi.

- Zajem in transformacija (ang. *Ingestion and Transformation*)

  Osnovni namen je pridobitev podatkov iz operativnih (prometnih) sistemov, shramba v podatkovni platformi ter pretvorba v vsebinsko obliko, primerno za analiziranje in poročanje.

V osnovi je ta nivo primerljiv z ELT principi. A pozor, tega se ne sme zamešati z ETL. Tu se podatki po pridobitvi najprej shranijo ter šele nato pripravijo v obliko, ustrezno za naslednje korake obdelave.

- Hramba (ang. *Storage*)

  Na tem nivoju se ukvarjamo z dejanskim shranjevanjem podatkov. Poudarek je na ravnovesju med čim manjšimi stroški, skalabilnostjo in zmožnostjo analitike nad veliko količino podatkov.

- Poizvedovanje in napovedovanje (ang. *Query and Processing*)

  Na tem nivoju omogočimo infrastrukturo, pogoje in okolje za podatkovne analitike in znanstvenike, da lahko učinkovito opravljajo svoje delo. Omogočiti moramo okolje, da lahko z najnovejšimi pristopi, programskimi jeziki ter koncepti pridobijo uporabne uvide ter analize, ki organizaciji ustvarjajo dodano vrednost.

- Izhod (ang. *Output*)

  Nivo, ki omogoča vse rezultate, vso vsebino, vso vrednost posredovati in omogočiti odjemalcem. Na tem nivoju omogočimo prikaz, vključevanje ali enostaven dostop do rezultatov podatkovne platforme.

Naslednji nivoji so že zelo konkretni in za podroben opis ter lastnosti bi verjetno potrebovali nov (ali daljši) prispevek. Vseeno pa je smiselno izpostaviti tiste elemente, ki prinašajo drastične spremembe.

- 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 relativno enostavno pretvorimo v sodobno dogodkovno usmerjeno rešitev. Dobro izpostavljeno orodje za to ponuja Oracle z rešitvijo Oracle Golden Gate (https://www.oracle.com/integration/goldengate/).

- Workflow Manager

  Ta blok je neposredna rešitev orkestracije iz visokonivojske arhitekture. Izpostavili smo, da je orkestracija pomemben element. Obstajajo vgrajena ali neodvisna orodja, katerih glavna naloga je orkestracija vsega dogajanja. Izredno prodorno orodje v tem sklopu je Apache Airflow.

- Data Modelling in Metadata Management

  Ob razumevanju petih V-jev nam je znano, da imamo veliko podatkov, ki so hkrati tudi zelo raznoliki. Razumevanje nad podatki in upravljanje z njimi je za velike sisteme izredno breme. Zato ne čudi, da sta tu enakovredno poudarjena bloka podatkovnega modeliranja in upravljanja z meta podatki.

Sta pa dva bloka v nečem drugačna. Iz lastnih izkušenj lahko potrdim, da so na tem področju najbolj učinkovite lastne aplikacije, saj je v praksi pogosto prisotnih veliko specifik in prilagajanje namenskih storitev in orodij navadno vzame mnogo več časa kot razvoj nečesa lastnega.

- Data Lake

  Podatkovno jezero je verjetno najtežje razumljiv element te podatkovne strukture. Težava je v idejni 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 masovne podatke 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 masovnih podatkih 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 izredno loči od področja priprave podatkov (saj se masovni podatki ločijo od običajnih podatkov).

- Data Warehouse

  Podatkovno skladišče je verjetno daleč najbolje prepoznaven izraz v podani arhitekturi. Kar nas lahko preseneti, je premik od tradicionalnih konceptov do platform masovnih podatkov. V tradicionalnih sistemih je bila podatkovna platforma enaka podatkovnemu skladišču. Sedaj pa je podatkovno skladišče le eden izmed elementov podatkovne strukture.

  Kar 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.

- Ad Hoc Query Engine

  Ta element je izredno zanimiv, saj povezuje tradicionalni pristop z ekstremi masovnih podatkov. Če smo v tradicionalnih sistemih imeli relacijsko bazo, ki je bila na nek način črna škatla – 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 v bistvu 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 masovnih podatkov.

  Na tem mestu dodajam lastno izkušnjo, da lahko tak pristop pripelje tudi do zavajajoče izkušnje, ko se podatkovno jezero izenači s podatkovnim skladiščem. V tem primeru pride nehote do »zlorabe« namembnosti, kar pa dolgoročno lahko pripelje do izjemnih težav in stroškov. Podatkovna jezera so namreč stroškovno optimizirana za shranjevanje podatkov, niso primerna za dimenzioniranje in indeksiranje podatkov. Zato lahko nastanejo visoki stroški, če želimo indeksiran dostop do podatkov, ki so optimizirani zgolj za učinkovito shranjevanje.

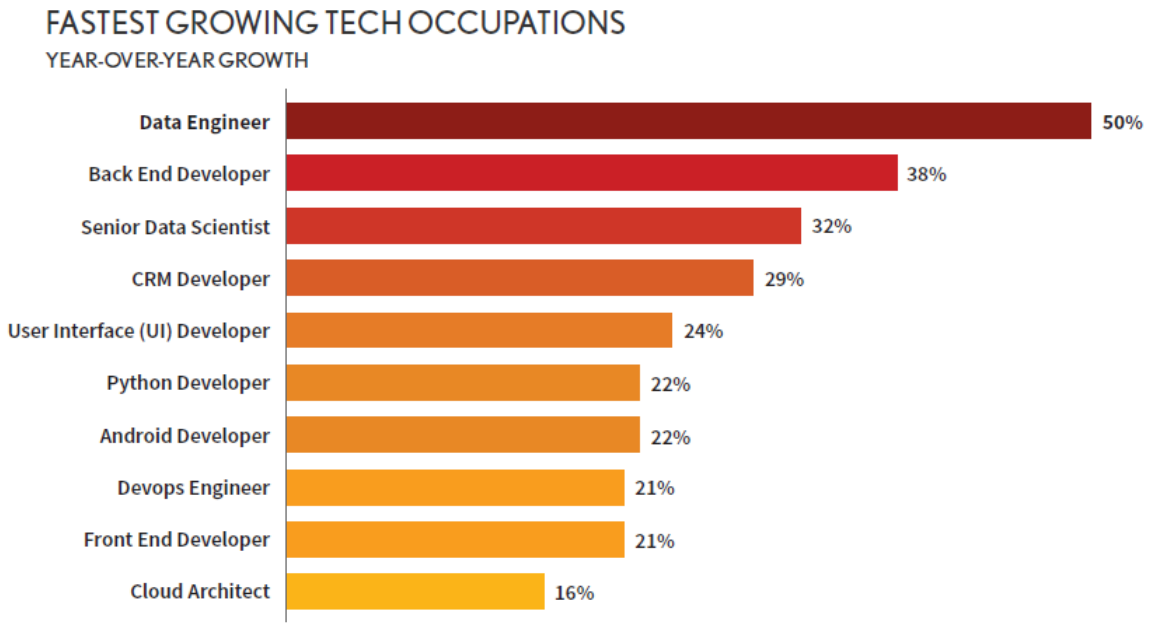
S temi bloki bi tudi zaključili pregled tehničnih elementov.

Sklepam, da se je ob prebranem marsikomu postavilo še eno dodatno vprašanje: kdo je sploh sposoben omenjene zadeve implementirati in jih vpeljati v poslovanje?

Podjetja torej potrebujejo tudi strokovnjake za podatke. To so predvsem podatkovni znanstveniki, podatkovni inženirji in podatkovni analitiki.

PODATKOVNI INŽENIRJI, ANALITIKI IN ZNANSTVENIKI

Po podatkih poročila The Dice 2020 Tech Job Report[3] je bilo v letu 2020 v Združenih državah Amerike najhitreje rastoče delovno mesto Podatkovni inženir (ang. *Data Engineer*).



Slika 4: Najhitreje rastoča tehnološka delovna mesta v ZDA v 2020.

Podatkovni inženir je različica programskega inženirja (ang. *software engineer*) oz. nam bližje programerja. Podatkovni inženir je dokaj nova opredelitev, zelo pogosto se pojavlja v navezi z masovnimi podatki. Razumljivo – glede na vse posebnosti masovnih podatkov nam je verjetno jasno, da potrebujemo posebne kompetence in lastnosti, da se vse predstavljeno vpelje v prakso.

Kljub temu da podatkovni inženir v nazivu vsebuje besedo podatek, pa se dejansko ne ukvarja s podatki, temveč je njegova glavna naloga vzpostaviti podatkovne platforme za masovne podatke. Glede na podrobno arhitekturo lahko hitro ugotovimo, da pravzaprav pokriva izredno

[3] Več o tem lahko preberete na: https://techhub.dice.com/Dice-2020-Tech-Job-Report.html

široko področje in mora razumeti veliko različnih nivojev in elementov. S tem pa se tudi pojasni korelacija med rastjo masovnih podatkov in rastjo potreb podjetij po podatkovnih inženirjih.

Šele ko ima podjetje postavljeno ustrezno podatkovno okolje, pridejo na vrsto podatkovni znanstveniki in podatkovni analitiki. To pa so tudi trije ključni poklici za celovito in uspešno delo na podatkovnih platformah.

ZAKLJUČEK

Masovni podatki so že globoka realnost. Predvsem zaradi interneta stvari, je podatkov več kot kadarkoli, podatki so mnogo bolj raznoliki, vse več je nestrukturiranih podatkov. Ti se kreirajo 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 življenje dobe podatka.

Pri tem pa ne smemo prezreti, da v današnjem času se še dodatno dviga vrednost podatkom njihova obdelava in kreiranje vrednost v realnem času.

Ravno zaradi tega so masovni podatki opredeljenimi s 5 V-ji (volumen, hitrost, raznolikost, verodostojnost in vrednost). Z razumevanjem vseh teh dimenzij se nam odpre pravi pogled na priložnosti masovnih podatkov.

S tem pa se pojavi izziv, kako tehnično podpreti vse operacije, ki morajo obdelovati vseh 5 V-jev. V ta namen se pojavlja gradnik podatkovne platforme za masovne podatke. Ta je skupek elementov in servisov za omogočanje tako klasičnih ter tradicionalnih procesiranj podatkov kot se tudi uspešno spopada z najbolj ekstremnimi izzivi masovnih podatkov.

Te platforme pa le niso tako zapletene kot verjetno zgledajo na prvi pogled, saj so le naravna evolucija dosedanje poti, le pogledati jih moramo postopoma iz grobe visokonivojske arhitekture do podrobne logične.

Masovni podatki prinašajo popolnoma nove ekstremne dimenzije možnosti in priložnosti obdelave podatkov, a vendar je z dobrim razumevanjem preskok lahko narediti tudi postopoma in z majhnimi koraki.

Iz lastnih izkušenj svetujem, da se pri pojavi masovnih podatkov ne sme le zamahniti z roko in reči, da to ni za nas, temveč si dajmo priložnost razumeti ozadja ter jih smiselno vpeljati v našo trenutno realnost in tudi pričakovan prihodnji razvoj. S tem bomo še uspešnejše transformiralo naše poslovanje v podatkovno gnano organizacijo in družbo.

VIRI IN LITERATURA

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*.

Awanish. (2021, September 19). *edureka!* Retrieved from Big Data Tutorial: All You Need To Know About Big Data!: https://www.edureka.co/blog/big-data-tutorial

Pandur, S. (2013). Kjer imajo nadzor nad vsemi rdečimi lučmi. *Delo*.

The five V's of big data. (2021, September 19). Retrieved from https://www.bbva.com/en/five-vs-big-data/

Wikipedia. (2021, September 19). Retrieved from Big Data: https://en.wikipedia.org/wiki/Big_data