Podatkovni inženiring – vzpon, razvoj in prihodnost

Jure Jeraj, Anžej Curk

Result d.o.o., Ljubljana, Slovenija
jure.jeraj@result.si, anzej.curk@result.si

Podatkovni inženiring je doživel neverjeten vzpon in razvoj v zadnjih letih, postajajoč ključen gradnik sodobnega poslovanja in tehnološkega napredka. Ta kompetenca se osredotoča na oblikovanje, razvoj in vzdrževanje robustnih podatkovnih arhitektur, ki omogočajo učinkovito obdelavo, shranjevanje in analizo velikih količin podatkov. Je odgovor na eksponentno rast količine podatkov, ki jih proizvajajo organizacije in uporabniki po vsem svetu. Je privedel do nastanka številnih tehnologij in orodij, ki omogočajo boljše obvladovanje podatkov. To vključuje napredne podatkovne baze, porazdeljene sisteme za obdelavo podatkov, orodja za integracijo in transformacijo podatkov ter platforme za orkestracijo in upravljanje podatkovnih tokov. Prav tako so se razvile metodologije in prakse, ki omogočajo avtomatizacijo in standardizacijo procesov podatkovnega inženiringa ter zagotavljajo doslednost in zanesljivost. Podatkovni inženiring postaja pomembne člen vsake moderne organizacije ali družbe, ki so prepoznale možnost inovativnih rešitev na podlagi moči podatkov.

Ključne besede:

podatkovni inženiring

velepodatki

upravljanje s podatki

procesiranje v realnem času

nOps

Univerzitetna založba Univerze v Mariboru

ISBN 978-961-286-772-0
DOI https://doi.org/10.18690/um.feri.8.2023.26

1 Uvod

Danes se podatkovni inženiring jemlje kot kritična komponenta modernih podatkovnih struktur. Te so posledica ekstremne rasti [1] količine podatkov v zadnjih letih, predvsem kot posledica vseh mobilnih naprav, interneta stvari, socialnih omrežij, oblačnega procesiranja ter tudi zaradi naglega razvoja poslovne inteligence v preteklem desetletju. Podatkov pa ni samo več, prihajajo hitreje in se pojavljajo v manj strukturiranih oblikah. Klasične podatkovne baze in klasični ETL procesi tega enostavno ne zmorejo več, zato je bilo potrebno poseči k zahtevnejšim rešitvam, te pa so za seboj potegnile tudi povsem novo kompetenco, in sicer podatkovni inženiring.

Podatkovni inženiring se tako povezuje s pojavitvijo velepodatkov. Velepodatki so na preprost način opredeljeni kot tiste zbirke podatkov, ki jih ni mogoče obdelovati s klasičnimi in tradicionalnimi metodami [2]. In če ni možno delati tradicionalno, so potrebne nove veščine. In tako se pojavi podatkovni inženiring kot ga opredeljujemo danes.

Sicer pa ne smemo spregledati, da se podatkovni inženiring sramežljivo pojavlja že od devetdesetih let dvajsetega stoletja z uveljavljanje podatkovnih zbirk in njihove analize. A vseeno pravi razmah se je zgodil s pojavitvijo velikih tehnoloških korporacij kot so Facebook, AirBnb, Amazon, Apple, Netflix… Njihov poslovni model v osnovi temelji na produktivizaciji in monetizaciji podatkov. In kjer so podatki osnovno sredstvo obstoja, je tudi inženirska skrb za podatke na najvišji prioriteti vodstva.

V povezavi z zgoraj omenjenimi korporacijami je objavljen zanimiv blog članek [3] Maxime Beauchemina, ki pravi, da se je leta 2011 pridružil Facebooku kot inženir poslovne analitike. Ko ga je leta 2013 zapustil je bil podatkovni inženir, pri tem pa ni niti napredoval niti ni zamenjal delovnega mesta. Enostavno je Facebook v tem času ugotovil, da je delo, ki ga je Maxime opravljal, presega običajno delo inženirja poslovne analitike. Zato so začeli razvijati nove veščine in tako ustvarili novo kompetenco.

Za zanimivost, podobno pot sva opravila tudi avtorja tega članka. Sredi prejšnjega desetletja sva se pridružila ekipi mednarodnega podjetja kot ETL razvijalca, nekaj let kasneje pa ugotavljava, da se lahko opiševa kot podatkovna inženirja. A postala sva pozorna še na en element – če sva pred leti delala skoraj enake zadeve, zdaj opravljava različne zadeve in se niti ne moreva več primerjati med seboj. To nakazuje, da je podatkovni inženiring postala zelo široka vloga in se lahko tudi znotraj nje pojavljajo določene specializacije. To širino bova poskusila na kratko predstaviti v tem prispevku, v smeri povzetka, kako se je ta kompetenca razvijala v preteklih letih.

Osredotočila se bova predvsem na elemente s katerimi sva imela izkušnje in jih dobro poznava: to pa povezala s ključnimi prelomnicami in mejniki, ki jih je zelo dobro povzel že pred nama tudi Tobias Macey v blog prispevku o vzponu podatkovnega inženiringa [4].

2 Apache License 2.0

Ob pogledu nazaj na vse ključne komponente, ki so pripeljali do današnjega stanja, lahko ugotovimo, da se je veliko inicialnega razvoja zgodilo v omenjenih velikih korporacijah. Pri tem pa je ključno, da so te korporacije zelo hitro svoje rešitve objavile kot odprtokodne, skoraj vedno pod licenco Apache License 2.0.

Apache License 2.0 je ena najbolj razširjenih odprtih licenc za programske rešitve. Prvič je bila objavljena s strani Apache Software Foundation, organizacije, ki spodbuja razvoj odprte programske opreme. Pred ustanovitvijo Apache Software Foundation leta 1999 je bila skupina razvijalcev, znana kot Apache Group, ključna pri ustvarjanju projekta Apache HTTP, ki je postavil temelje za številne druge odprtokodne projekte. Danes ta licenca omogoča svobodno uporabo, distribucijo in spreminjanje odprtokodne programske opreme, kar je prispevalo k razcvetu odprtih tehnologij po vsem svetu.

Te rešitve so tako močne in razširjene, da veliki ponudniki oblačnih rešitev, pa tudi komercialnih komponent uporabljajo kot osnovo za svoje storitve. S to licenco imajo namreč prost dostop do kakovostnih in preverjenih odprtih tehnologij, ki so že bile razvite in preizkušene v skupnosti. Nato nadgradijo te osnovne odprtokodne rešitve

z lastnimi funkcijami, prilagoditvami in izboljšavami, ki ustrezajo njihovim specifičnim potrebam in zahtevam trga. S tem pristopom lahko ponudniki oblačnih storitev hitro razvijajo in ponujajo inovativne in konkurenčne rešitve, ki so hkrati stabilne in zanesljive za njihove stranke.

Zaradi tega imajo podatkovni inženirji še nekaj manj opaznih prednosti. Skupnost za podporo posameznih komponent je mnogo večja, saj je v veliko osnovnih primerih že dovolj podpora za odprtokodno rešitev in nato lahko rešuješ tudi primere na oblačni storitvi. Konkretni primer je lahko AWS Athena (storitev za ad hoc SQL poizvedbe nad podatkovnim jezerom), ki izhaja iz odprtokodne rešitve PrestoDB (tudi objavljen pod Apache License 2.0). Za samo sintakso in specifike AWS Athene je tako povsem dovolj, da uporabljamo dokumentacijo PrestoDB-ja.

Prednost pa je tudi v tem, da so visokonivojske arhitekture med seboj primerljive, predvsem na vseh večjih ponudnikih oblačnih storitve, kar pomeni, da se lahko inženirji hitreje prilagodijo na drugega ponudnika. Možna pa je tudi kombinacija originalnih odprtokodnih rešitev in lastnih komercialnih storitev teh ponudnikov, kar močno poveča fleksibilnost razvoja.

3 Apache Hadoop – velike količine podatkov

Apache Hadoop je prvo uveljavljeno ogrodje oz. zbirka orodij za obdelavo velikih količin podatkov. Razvijati sta ga začela Doug Cutting in Mike Caferalla leta 2004 v času porasta spletnih iskalnikov; glavni namen je bil narediti možnost porazdeljene obdelave podatkov. Doug Cutting se je leta 2006 pridružil Apache Software Foundation, s tem pa se je Hadoop prenesel v skupnost odprtokodnih projektov, posledično je doživel nagel razvoj na področju obdelave velike količine podatkov.

Temeljna prednost Apache Hadoopa je njegova sposobnost obvladovanja velikih količin podatkov na porazdeljen način preko ključne paradigme, imenovane MapReduce. Ta metoda razčleni velike naloge na manjše, ki jih nato porazdeli na več vozlišč v grozdu. Vsako vozlišče neodvisno izvaja naloge, nato pa rezultate združi v končni izhod. To omogoča izjemno paralelno obdelavo, ki poveča učinkovitost in zmogljivost sistema.
Druga ključna komponenta Apache Hadoopa je distribuiran sistem za shranjevanje podatkov Hadoop Distributed File System (HDFS). HDFS omogoča shranjevanje podatkov na več vozliščih v grozdu, kar zagotavlja vzdržljivost podatkov in visoko razpoložljivost. Zaradi te zasnove je Hadoop odporen na izpad posameznih vozlišč in omogoča nemoteno delovanje tudi pri okvarah strojne opreme.
Apache Hadoop je tako postal (in pravzaprav še vedno ostaja) osrednji del ekosistema za obdelavo podatkov v velikih količinah. Je odprtokoden in ima še vedno zelo veliko skupnost, ki ga uspešno razvija. Mogoče ni več prva izbira za končne uporabnike, je pa svojima konceptoma MapReduce in HDFS uporaben v drugih modernejših ogrodjih.

4 Apache Kafka – podatki v realnem času

Naslednji izziv pri velepodatkih je bila hitrost generiranje podatkov in potreba po upravljanju z njimi praktično takoj. Rešitev se je ponudila leta 2011, ko je Linkedin svojo interno rešitev naredil odprtokodno in jo predal Apache Software Foundationu.

Apache Kafka je danes praktično nepogrešljiva v ekosistemu obdelave podatkov v realnem času. Ena ključna prednost Apache Kafka je njegova sposobnost obvladovanja ogromnih tokov podatkov na visoki ravni zmogljivosti in nizki latentnosti. Deluje na principu porazdeljenega sistema, kjer se podatki shranjujejo v podatkovnih vrečah, imenovanih teme (ang. topics), in so nato poslani na različne porabnike (ang. consumers), ki jih obdelujejo skladno s svojimi potrebami. Ta arhitektura omogoča večkratno branje podatkov in preprosto horizontalno razširjanje sistema, kar zagotavlja visoko razpoložljivost in vzdržljivost.

Apache Kafka se pogosto uporablja kot osrednji gradnik v arhitekturi sistema za obdelavo podatkov v realnem času, kjer se podatki neprekinjeno zajemajo iz različnih virov, kot so senzorji, strežniki, aplikacije in druge naprave. Ti podatki se nato prenašajo v realnem času prek Kafka tokov, kar omogoča hitro obdelavo, analizo in posredovanje na druge ciljne sisteme ali aplikacije.
Poleg tega se Kafka odlično integrira z drugimi odprtokodnimi orodji za obdelavo podatkov, kot sta Apache Spark in Apache Hadoop (in vsemi komercialnimi izpeljankaj kot so npr. Cluodera, Databricks ipd.).

5 Podatkovna jezera in oblačne storitve

Prejšnja dva elementa omogočata hitrejšo obdelavo ogromne količine strukturiranih in tudi nestrukturiranih podatkov. Vpeljuje programerske tehnike v obdelavo podatkov, ki pa odstopajo od metod klasičnih relacijskih baz kot so npr. Oracle, IBM DB2 ali MS SQL Server. Enostavno je osnovna arhitekturna zasnova relacijskih baz namenjena za zagotavljanje (poenostavljeno) transakcijske konsistentnosti podatkov pri kompleksnih aplikativnih sistemih, ki temeljijo na digitalizaciji in avtomatizaciji procesov. Medtem ko zgoraj omenjeni koncepti so osredotočeni na upravljanje in obdelavo podatkov samih.

Če se spomnimo, ena izmed ključnih komponent Hadoopa je HDFS, ki je v bistvu prilagojen datotečni sistem. To pa pravzaprav pripelje do tega, da se pri obdelavi velepodatkov začenjamo ukvarjati tudi s samimi koncepti in sistemi za shranjevanje (s čimer se pri relacijskih bazah pravzaprav ne ukvarja niti bazni administrator). Zato je bilo to nekaj novega in to je tudi osnova za nastanek izraza podatkovno jezero. Ta se je začel uveljavljati šele po letu 2011, torej kar 5 let kasneje od prve verzija Hadoopa in HDFS-ja.

Podatkovna jezera so tako predvsem shrambe v katero lahko zelo hitro zapisujemo strukturirane, polstrukturirane in nestrukturirane podatke. Pri podatkovnih jezerih lahko podatke zapisujemo izredno hitro, ker se ne preverja konsistenca zapisov. Tak sistem omogoča boljšo skalabilnost, hkrati je tudi cenejši, ker ima manj kompleksnejših elementov. Te namreč nadomeščamo z ogrodji in rešitvami kot sta Hadoop, Kafka ….

Z že do zdaj povedanim lahko razumemo, da je z vsemi temi komponentami prišlo do situacije, ko je mnogo več elementov odvisnih od same arhitekturnih rešitev. S tem pa se je odprla novo področje, in sicer upravljanje skalabilnosti. Z možnostjo nadzora posameznih komponent se lahko bolje izrablja procesorska, strežniška in infrastrukturna moč. Potrebe po tem pa se lahko močno spreminjajo čez dan, čez teden, lahko tudi čez leto.

Ker so organizacije vedno bolj odvisne od tega procesiranja, so morale zagotavljati računalniško in mrežno infrastrukturo za pokrivanje največjih špic delovanja. Te pa se pojavljajo le občasno, kar pripelje do situacije, da je bila večino časa pripravljena infrastruktura neizkoriščena.

To pa je en od razlogov, da so se zelo razširile oblačne storitve, specifično tudi na področju podatkovnega inženiringa. Možnost skalabilnosti na manjši časovni enoti (lahko tudi na nivoju posamezne ure ali dneva) je izredno pomembna, hkrati pa tudi lastnost, ki jo ponudniki ponujajo na način »plačaj kolikor uporabljaš«.

Pri tem pa se je treba zavedati, da za vsako področje (npr. sistemska administracija, bazna administracija, varnost…) dobite storitev najvišje kvalitete. Glavni razlog, da je vse to možno, je lastnost, da je so mejni primeri oz. špice v zahtevah redke in se lahko lažje porazdelijo med tisoče različnih organizacij.

Smatramo, da je vzpon modernih podatkovnih struktur in oblačnih storitev povezan in da bi verjetno težko uspela en brez drugega. Podatkovna jezera in oblačne storitve so tako danes enako ključni gradnik v sodobni obdelavi podatkov kot predhodno omenjene komponente.

6 Orkestracija sistemov in podatkovni katalog

Verjetno imamo v vsakem sistemu skrite junake. Avtorja smatrava, da so to v našem primeru sistemi za orkestracijo opravil in sistemi za nadzor na izvajanje teh opravil ter dokumentiranje vseh meta podatkov na kateremkoli nivoju naše arhitekture.

Sodobne arhitekture na področju upravljanja s podatki so sestavljene iz različnih komponent. Tudi na oblačnih storitvah imamo veliko število storitev (npr. AWS ima skupno več kot 300 različnih storitev), s tem, da imamo vedno možnost uporabljati tudi odprtokodne rešitve. Ni pa niti nenavadno, da se uporabljajo storitve različnih ponudnikov oblačnih storitev.

Število različnih storitev in heterogenost povečujeta možnost, da se pri procesiranju podatkov kje zalomi. Izraz, ki se je uveljavil za nadzor, pregled in urejanje tega področja je orkestracija opravil. Sam izraz nakazuje na usklajeno delovanje orkestra. Vsak glasbenik orkestra mora točno vedeti kaj igra v katerem delu nastopa je, kaj se je in kaj se bo zgodilo. Večji kot je orkester, zahtevnejše je. Tako se povečuje vloga dirigenta. Povsem enako je v modernih podatkovnih rešitvah, saj imamo veliko komponent in (mikro) rešitev, ki jih moramo pravilno povezati v celoto. Vlogo dirigenta prevzemajo namenska orodja in rešitve za orkestracijo.

Trenutno zelo popularno orodje je Apache Airflow. Sledi vsem trendom iz področja, predvsem pa je izredno povezljiv s preostalimi rešitvami fundacije Apache. Povezljivost je pa ključna za dobro orkestracijo. V povezavi s spodobnim uporabniškim vmesnikom nam daje to možnost, da imamo lahko z enim (neodvisnim) orodjem nadzor nad vsemi akcijami in opravili pri našem upravljanju s podatki. Na ta način lahko vidimo zelo zemljevid naših opravil, poročila o izvajanju, ozka grla ali pa identificirati težave še predno bi jih opazili končni uporabniki.

Orkestracija je podatkovnim inženirjem še blizu; pravzaprav je samo močna nadgradnja nekdanjih »cron« opravil ali podobnih rešitev. Na drugi strani pa se zdi, da je podatkovni katalog nekaj, kar je vedno prvi kandidat, da se ga preskoči.

V razdrobljenem in heterogenem sistemu je pomembno, da imamo centraliziran sistem za preglednost in usklajenost podatkov oz. meta podatkov. V večjih organizacijah so procesi ločeni med posameznimi oddelki ali enotami, istočasno pa so oddelki, ki sprejemajo odločitve na podlagi podatkov, ki niso pod njihovim lastništvom. To so navadno večje organizacije, kjer je zaradi velikosti tudi komunikacija med enotami otežena zaradi široke organizacijske strukture.

V teh primerih je podatkovni katalog edina stična točka, da se lahko posamezne enote sporazumevajo med seboj kaj točno pomeni posamezen podatek. Ali pa kakšno je poslovno pravilo za posamezen kazalnik. Velja pripomniti, da je zelo dobra lastnost podatkovnih analitikov in znanstvenikov, da se vedno znajdejo in najdejo podatke, za svoje delo. A včasih samo iz podatkov samih ni možno razbrati, kaj predstavljajo. Posledično pomanjkljivo ali ceno napačno razumevanje lahko pripelje tudi do netočnih, zavajajočih ali celo kritično napačnih analiz, poročil ali ukrepov. Najbolj klasični primeri – za vsakega, ki je delal pri mednarodnih korporacijah – so časovni pasovi, valute in merske enote. Namreč hitrost 40 ni enaka v Nemčiji ali v Veliki Britaniji. A brez dodatnih informacij ni mogoče predvideti, ali so podatki v sistemu že pretvorjeni v enotno mersko enoto ali ne.

Namen podatkovnega kataloga je, da poveže sorodne podatke. In tudi da pove lastnosti posameznega podatka, predvsem če je časovno, geografsko ali kako drugače povezan.

Najina izkušnja je, da ravno za področje podatkovnega kataloge je še zelo malo res učinkovitih orodij, predvsem odprtokodnih. Trenutno je verjetno najbolj napredna odprtokodna rešitev DataHub, ki je tudi objavljen pod licenco Apache License 2.0.

7 Evolucija poslovne analitike v podatkovno analitiko

Uvodoma smo podali primer, kako se je podatkovni inženiring razvil iz poslovne analitike. Zanimivo je pogledati kakšen je po vseh teh spremembah današnji položaj poslovne analitike. Predvsem bi radi poudarili, da imamo tu malo smole s slovenskimi prevodi. Pri nas se je namreč – mogoče celo preveč – uveljavil izraz poslovna analitika; dobesedni prevod bi namreč bil poslovna inteligenca (ang. Business intelligence). To omenjamo, ker se danes mednarodno več uporablja podatkovni analitik (ang. Data Analytic). Ni pa samo preimenovana vloga, tudi vrline so se bolj specializirale na zmožnost iskanja, razumevanje, priprave in obdelave podatkov.

Praktično to pomeni nadaljnjo delitev med tehničnimi in poslovnimi vlogami. Izzivi poslovnih priložnosti se ves čas premikajo k skrajnostim; če je bilo pred 10 leti še izziv, kako pregledati vse lastne podatke, je danes izziv kako vpeljati inteligentne metode ne le nad lastnimi podatki, temveč v kombinaciji s komercialnimi, odprtimi in pravzaprav vsemi možnimi podatki.

V današnjem svetu se pa odpira dodatna vrlina interpretacije podatkov. Če smo včasih opredelili strokovnjaka za poslovno inteligenco kot most med tehničnim in poslovnim svetom, lahko danes ugotovimo, da je ta most vrlina posameznika, da razume podatke in jih zna smiselno predstaviti poslovnim uporabnikom in odločevalcem. To vlogo danes lahko opravljajo tako napredni poslovni analitiki kot razumevajoči podatkovni analitiki. Ta vez je vedno obstajala in bo tudi v prihodnje obstajala; spremembe so predvsem del zahtevanih specializacij kot odgovor razvoja, ki smo do zdaj že prikazali.

8 nOps

Ne moremo se izogniti vzporednicam med inženirjem za razvoj programske opreme in podatkovnim inženirjem. Pravzaprav so vzporednice zelo močne; podatkovni inženiring je nekje na stopnji kot je bil razvoj programske opreme npr. pred 15, 20 leti. Zato ne čudi, da se nekatere aktivnosti razvijajo po zelo podobnih poteh. Eno izmed področij je nOps, ki je nadpomenka raznim ostalim Ops sistemom, kot je npr. verjetno najbolj znan DevOps. Posledično se na področju podatkov poskuša vzpostaviti izraz DataOps.

Sami smo mnenja, da se ta izraz ne bo uveljavil, predvsem iz nekaj razlogov.

Prvi je, da tehnično gledano ni tako zelo drugačen od DevOps. Podatkovni inženiring je namreč le specializirana veja razvoja programske opreme, ki je specializirana za področje podatkov. Zato so tudi operacije za razvoj teh rešitev podobne operacijam za razvoj programske opreme. Iz tega vidika je DataOps pravzaprav zelo soroden kot DevOps.

Drugi razlog pa je, da so se same operacije nad kvaliteto podatkov, orkestracijo in lastnostmi že razvile v okviru lastnosti celovitega upravljanja s podatki (ang. Data Governance) in je velika možnost, da se to področje ne bo razumelo kot del DataOps. S tem pa potem

Obstaja pa še tretji razlog, kjer se pa težje razume vpliv na potencialni razvoj izraza DataOps. Prej in bolj se je namreč že uveljavilo področje MLOps, ki pokriva operacije nad strojnim učenje. To pa obsega tudi delo s podatki, saj brez urejenih in smiselnih podatkov ni metod strojnega učenja. Avtorja tega članka, kljub temu, da se identificirava kot podatkovna inženirja, bova težko sprejela izraz DataOps, ker smatrava, da se ga uveljavlja malo na silo, po vzoru ostalih uveljavljenih Ops zadev.

A ne glede na najino mnenje, aktivna operativna aktivnost na področju podatkov – kvaliteta, orkestracija in dokumentacija – je, je bila in bo ena izmed ključnih elementov upravljanja podatkov.

9 Prihodnost (podatkovnega) inženirstva

Kljub temu, da je napovedovanje prihodnjih trendov nehvaležno, pa vseeno lahko podava nekaj najinih razmišljanj, o prihodnosti podatkovnega inženirstva.

Iz prispevka sledi, da so se večje spremembe začele nekje od leta 2005. V naslednjih 10 letih se je izvedlo veliko nadgradenj in rešitev, ki so zavzele prazen prostor na področju upravljanja z velepodatki. Tega so omogočile evolucije programske opreme in infrastrukturne zmožnosti, predvsem oblačne storitve. V tem obdobju so se pripravili odgovori na enormen količine podatkov, procesiranje v realnem času in upravljanje tudi z nestrukturirani podatki. S tem smo zadostili potrebam velepodatkov, ki so osnovno opredeljeni s 3V definicijo (volumen [ang. volume], hitrost [ang. velocity] in raznolikost [ang. variaty]).

V prihodnosti ne pričakujemo takih strateških sprememb, kot so se zgodile v zadnjih 15 letih. Se bo pa še vedno razvijala in sledila trendom predvsem iz vidika združevanja še več podatkov ter poudarka na varnosti in sistemom za upravljanje s podatki. Verjetno bo veliko vzporednic s trendi razvoja programske opreme. Ogrodja, rešitve in knjižnice se bodo ves čas posodabljale, prilagajale na nove okoliščine in sledila ostalim potrebam po podatkih. Razvoj zagotovo ne bo izostal, a zdi se, da je trenutno predvsem čas za ustrezne za maksimizacijo izrabe tehnologije na strani poslovnega sveta.

Se pa nakazuje zanimiv trend, ki se dogaja v zadnjih letih - razvoj vloge inženirja strojnega učenja (ang. ML engineer) oz. na splošno vzpon inženirstva. Trenutno je namreč val navdušenja nad umetno inteligenco iz vseh področij. Lahko je to strojno učenje, lahko je to generativna umetna inteligenca, lahko kaj drugega.

Vse to navdušenje se bo preneslo do porasta večje količine modelov za reševanje najrazličnejših izzivov. S tem pa se bo pojavila težava vpeljati te rešitve v vsakdanjo, produkcijsko uporabo, najverjetneje na podlagi velepodatkov in obdelavo v realnem času.

Ocenjujemo, da se bo v bližnji prihodnost razvilo veliko različnih inženirskih smeri, a v nekem trenutku se bo ta raznolikost začela združevati in upamo, da se bo usmerilo v pot, da bomo razumeli, da imamo le ene inženirstvo z različnimi specializacijami. Ker konec koncev, veliko je skupnih elementov med inženirjev programske opreme, podatkovnim inženirjem ali pa inženirjem strojnega učenja.

10 Sklepna misel

V uvodnih mislih tega članka sva avtorja podala najino izkušnjo prehoda iz ETL razvijalca v podatkovna inženirja ter ugotovitve o raznolikosti te vrline. V prispevku sva želela prikazati obširen razvoj področja na katerem delujeva. Desetletje ali dve nazaj je bil IT sektor osredotočen na aplikacije – digitalno podporo procesom in njihovo avtomatizacijo. Ekosistem se je delil na transakcijsko (OLTP) in analitično (OLAP) področje, kjer pa je bil analitičen svet podrejen transakcijskem.

Nato pa so se začele vzpenjati korporacije, ki so se zavedale moči podatkov in so svoje poslovne modele ustvarile na podlagi inovativne izrabe podatkov. Kar naenkrat se je zgodilo, da organizacija ne sloni na zgolj na transakcijskih sistemih, temveč da so tudi podatki pomemben del generiranja prihodkov.

Zato so te korporacije začele usmerjati razvoj rešitev v smeri izrabe vseh podatkov – tudi nestrukturiranih - v najkrajšem možnem času. S tem so kreirale ekosistem velepodatkov, hkrati pa ponudile rešitve za njihovo procesiranje. Ena izmed pozitivnih posledic pa je tudi vpeljava novih vrlin, ki so združila kot vloga podatkovnega inženirstva.

Podatkovni inženiring je zdaj že uveljavljena in prepoznana vloga. Namenjena je organizacijam, ki se zavedajo, da morajo za svojo odpornost uporabljati več podatkov, tudi odprte in komercialne. In da morajo podatke procesirati hitreje od konkurence.

Za zaključek pa moramo biti vseeno tudi realni – podatkovni inženirji ne bodo nikoli zvezde organizacij verjetno niti IT ekip. Bodo pa ključni člen za uspešno produkcijsko vpeljavo metod umetne inteligence v organizacije in skupnosti.

Literatura

[1] https://www.gartner.com/en/documents/3898487, Gartner Magic Quadrant for Data Management Solutions for Analytics, obiskano 25. 7. 2023

[2] JERAJ Jure, NERED Urška, NIKOLOSKI Stevanče "Velepodatki – 5V-jev v praktičnih primerih", Uporabna informatika. Letnik 31, Številka 1 (maj 2023), str. 4 – 14.

[3] https://medium.com/free-code-camp/the-rise-of-the-data-engineer-91be18f1e603, The Rise of the Data Engineer, obiskano 25. 7. 2023

[4] https://www.dataengineeringpodcast.com/six-year-retrospective-episode-361, Reflecting On The Past 6 Years Of Data Engineering, obiskano 25. 7. 2023