BIG DATA IN VARNOST

Avtor: Tomaž Gambiroža

Visoka šola za poslovne vede, Management in informatika (2. stopnja)

Povzetek

Definicija, ki je nastala leta 2018 pravi, da so veliki podatki tisti, za obdelavo katerih so potrebna vzporedna računalniška orodja (Big data – Wikipedia, b.l.). Big data oziroma veliki podatki bi lahko postali novo gonilo svetovnih, gospodarskih in družbenih sprememb. Z naraščanjem količine zbranih podatkov na svetovnem merilu se približujemo prelomni točki za velike tehnološke spremembe, ki bi lahko prinesle nove načine upravljanja naših financ, izobraževanja, zdravja in ostalih področij. Medtem ko narašča kompleksnost podatkov, ter tudi obseg, hitrosti, verodostojnost in raznolikost, pa bodo spremembe zelo odvisne od naših zmožnosti da pridobimo neko dodano vrednost preko analize velikih podatkov. Big data Analytics tako predstavlja velik izziv pri oblikovanju razširljivih algoritmov, metodologij, aplikacij in sistemov v katere bi se te podatke integriralo in iz njih na učinkovit način pridobivalo skrito oziroma dodano vrednost ter znanje. Uporaba velikih podatkov pa s seboj prinaša tudi nove varnostne izzive. Trije glavni vidiki varnosti velikih podatkov so: varnost podatkov, varnost informacij in spremljanje varnosti. Potrebno je zagotoviti celovitost sistema, kvalitetno upravljanje velikih podatkov ter varnost kibernetskega prostora. Nujno je zagotavljanje spremljanja v realnem času, da bi v najkrajšem možnem času odkrivali varnostne grožnje in neobičajno vedenje, močno zaščito zaupnih informacij ter nadzor dostopa, s katerim preko identifikacije, preverjanja pristnosti in avtorizacije nadziramo kdo ima dostop do katerih podatkov.

Ključne besede: Big data/Veliki podatki, varnostni izzivi, optimizacija informacijske varnosti v organizaciji

Uvod

Big data je izraz, ki opisuje velike količine podatkov, ki jih je težko upravljati (tako strukturiranih kot nestrukturiranih), obenem pa vsakodnevno preplavljajo podjetja. Nanaša se na podatke, ki so tako veliki, hitri in kompleksni, da jih je nemogoče obdelovati s tradicionalnimi metodami.

Izzivi analize velikih podatkov vključujejo zajemanje, shranjevanje, analizo, iskanje, skupno rabo, prenos, vizualizacijo, poizvedovanje, posodabljanje, zasebnost in vir podatkov. Predhodno je omogočala le opazovanje in vzorčenje. Pogosto vključuje podatke z velikostmi, ki presegajo zmogljivosti tradicionalne programske opreme za obdelavo podatkov v sprejemljivem času.

Uporaba izraza BIG DATA se največkrat nanaša na uporabo napovedne analitike, analitike vedenja uporabnikov in drugih naprednih metod analize podatkov, ki omogočajo da se iz velike količine podatkov izvleče neka dodana vrednost. Analiza podatkov lahko opazi poslovne trende, prepreči bolezni, pomaga pri boju proti kriminalu in tako dalje.

Definicija, ki je nastala leta 2018 pravi, da so veliki podatki tisti, za obdelavo katerih so potrebna vzporedna računalniška orodja (Big data – Wikipedia, b. l.).

Kaj pomeni izraz Big data?

Big data so kombinacija nestrukturiranih, polstrukturiranih in strukturiranih podatkov, ki jih zbirajo organizacije. Te podatke nabiramo za pridobivanje vpogledov in uporabo v projektih strojnega učenja, napovednem modeliranju in drugih naprednih analitičnih aplikacijah. Uporabljamo jih za izboljšanje poslovanja, ustvarjanje prilagojenih marketinških kampanj in zagotavljanje boljših storitev za naše stranke. Vse to povečuje našo vrednost. Podjetjem lahko zagotovijo dragocen vpogled v svoje stranke, ki se nato uporabi za izboljšanje trženjskih tehnik za povečanje vključenosti strank (Big Data and analytics - Definitions, value, trends and applications, b.l.).

Uporabljajo se na primer na energetskih in medicinskih področjih. V medicini se lahko veliki podatki uporabljajo za prepoznavanje dejavnikov tveganja za bolezni ali pa jih zdravniki uporabijo za pomoč pri diagnostiki bolezni pri bolnikih (Big data - What it is and why it matters, SAS, b. l.). Energetska industrija uporablja velike podatke za sledenje električnim omrežjem, uvedbo upravljanja tveganj ali real time analizo tržnih podatkov.

Shranjevanje in dostopanje do velike količine informacij oziroma podatkov v analitične namene obstaja že dolgo časa, toda koncept velikih podatkov je pravzaprav pridobil pravi zagon v začetku 2001, ko je industrijski analitik Doug Laney artikuliral (Big Data – BuiltIn, b. l.) zdaj prevladujočo definicijo petih V-jev:

VOLUME - Prostornina

Nanaša se na količino pridobljenih oziroma obstoječih podatkov. Obseg je osnova velikih podatkov, saj je začetna velikost in količina podatkov, ki jih zbiramo. Če je ta količina dovolj velika, lahko podatke uvrstimo v velike podatke. Je relativna vrednost, saj se bo spreminjala glede na razpoložljivo računalniško moč, ki je na trgu.

VELOCITY - Hitrost

Hitrost se nanaša na to, kako hitro se podatki generirajo oziroma zbirajo in kako hitro se ti podatki premikajo. To je zelo pomemben vidik za podjetja, ki potrebujejo hiter pretok podatkov, da bi bili na voljo ob pravem času za sprejemanje najboljših možnih odločitev.

Organizacije, ki uporabljajo velike podatke bodo imele stalen in velik tok podatkov, ki se pošiljajo in ustvarjajo na nek končni cilj. Pridobivamo jih lahko iz različnih virov kot so: stroji, omrežja, pametni telefoni, družbeni mediji, itd. Podatke je potrebno hitro prebaviti in analizirati, včasih tudi v realnem času.

V nekaterih primerih je bolje imeti omejen nabor podatkov ki se zbirajo, kot pa zbrati več podatkov kot jih organizacija zmore v nekem doslednem času analizirati.

VARIETY - Raznolikost

Raznolikost se nanaša na raznolikost tipov podatkov. Organizacija lahko pridobi podatke iz številnih različnih virov podatkov, ki se lahko razlikujejo med seboj po vrednosti. Podatki prihajajo tako iz virov v podjetju kot tudi zunaj njega. Izziv raznolikosti se nanašana na standardizacijo in distribucijo vseh podatkov, ki se zbirajo.

Kot smo omenili že prej, so podatki lahko strukturirani, ne strukturirani ali polstrukturirani. Nestrukturirani podatki se ne prilegajo običajnim podatkovnim modelom. Polstrukturirani podatki so podatki, ki niso organizirani ampak imajo povezane informacije kot so metapodatki, kar olajša obdelavo v primerjavi z nestrukturiranimi podatki. Strukturirani podatki pa so organizirani in se jih lahko vnaša direktno v repozitorij, ki ima točno določen format. Ponavadi so ti podatki najbolj primerni oziroma dostopni za učinkovito obdelavo in analizo podatkov.

VERACITY – Resničnost oz. verodostojnost

Resničnost oziroma verodostojnost se nanaša na kakovost in točnost podatkov. Zbrani podatki imajo lahko manjkajoče vrednosti, lahko so netočni ali pa ne morejo zagotavljati resnično jasnost podatkov. Resničnost oziroma verodostojnost se na splošno nanaša na to, koliko zaupamo v zbrane podatke.

Velika količina podatkov lahko povzroči več zmede kot v jasnosti, če so podatki nepopolni. Kot primer, v medicini je lahko zdravje pacienta ogroženo, če so podatki o tem katera zdravila jemlje pacient nepopolni.

VALUE – Vrednost

Nanaša se na vrednost podatkov, ki jo lahko veliki podatki zagotavljajo in se neposredno nanaša tudi na to, kaj lahko organizacije dejansko storijo z zbranimi podatki.

Sposobnost pridobivanja vrednosti iz velikih podatkov je nujna, saj se vrednost podatkov lahko sorazmerno povečuje z našo sposobnostjo razumevanja teh podatkov in vpoglede, ki jih lahko pridobimo iz njih.

Organizacije lahko za zbiranje in analizo podatkov uporabljajo ista orodja, način kako iz teh podatkov pridobijo vrednost pa naj bi bil za vsako podjetje unikaten oziroma prilagojen za njihove potrebe.

Prednosti uporabe Big data

Podatkovni analitiki uporabljajo različne vrste podatkov za sprejemanje boljših oziroma izboljšanih poslovnih odločitev z razumevanjem nakupovalnih vzorcev in vedenja svojih strank. Podatkovno rudarjenje, napovedna analitika in strojno učenje so le nekatere od na novo razvitih tehnik, ki so v uporabi za doseganje novih vpogledov v neizkoriščene podatke na podlagi katerih se skuša izboljšati poslovne procese v nekem podjetju (Big Data Analytics – IBM, b. l.).

Uporaba velikih podatkov zmanjša stroške podjetja

Podjetjem lahko uporaba velikih podatkov pomaga upravljati svoje zaloge in svojim strankam zagotavljati boljše storitve dostave. Kot primer lahko navedemo Amazon, ki je že integriral tehnike velikih podatkov za optimizacijo dobavne verige. To jim omogoča da svojim strankam nudijo raven storitev brez primere. Zaloga se nabavlja glede na zgodovino nakupov in načrtovanju oziroma predvidevanju materialnih potreb glede na povpraševanje. To zmanjša čas in stroške pošiljanja. Upoštevajo se tudi dejavniki, kot so letni časi, vreme, gospodarske razmere in podobno.

Personalizirana izkušnja

V neki anketi je 87% kupcev dejalo, da so v primeru prilagojene nakupovalne izkušnje pripravljeni kupiti več (Berthiaume, 2019). Strategije personaliziranih izkušenj vključujejo pošiljanje prilagojene e-pošte uporabnikom kjer jim ponujajo posebne popuste in ponudbe, prikazovanje ciljnih oglasov različnim skupinam ljudi, implementacija strategij za večjo ali navzkrižno prodajo posameznikom in tako dalje. Kot primer lahko spet navedemo Amazon, ki je odličen primer uspešne uporabe analize velikih podatkov za ustvarjanje visokih prihodkov.

Ob brskanju po izdelkih po Amazonu se vam prikazujejo seznami priporočil (Customers who viewed this item also viewed; popular products similar to this item, ipd), ki so ustvarjeni na podlagi Amazonovih podatkovnih baz spletnih nakupovalcev. Glede na zgodovino brskanja vsakemu kupcu nudijo prilagojena priporočila. Rezultat je osupljiv, saj je za približno 35% kumulativnih prihodkov podjetja Amazon zaslužen algoritem za priporočilo izdelkov.

Boljša pomoč in podpora strankam

S spremljanjem povprečne odzivne hitrosti lahko osebje za pomoč strankam poveča splošno pravočasnost odziva. S pošiljanjem vprašalnikov in zbiranjem povratnih informacij svojih strank zagotavljajo informacije iz prve roke za izboljšanje kakovosti in zmanjšanje možnosti slabih storitev. S spremljanjem podatkov, kot je čas dostave lahko lastniki spletnih trgovin prepoznajo težave v procesu dostave in se izognejo težavam pri transportu (Oracle, b. l.).

Optimizacija cen

Orodja za analizo podjetjem omogočajo, da si ogledajo in spremljajo cene konkurentov v realnem času. To sicer zahteva velik nabor podatkov z vsemi cenami konkurentov katere je potrebno ažurno posodabljati saj se tržna cena nenehno spreminja. Orodje Octoparse (Octoparse, b. l.) lahko izvleče podrobnosti o izdelkih z spletnih trgovin kot so: Amazon, eBay, BestBuy, Walmart in nato podatke izvozi v formate CSV, JSON, Excel ali jih posreduje preko API-jev. Z Octoparse-ovimi WEB SCRAPING predlogami se pridobivajo bistveni podatki o izdelku (ime, cena, barva, ocena, opis, teža, slike izdelka in še marsikaj drugega.

Varnost Big data

Veliki podatki predstavljajo tako priložnosti kot tudi izzive. Obstajajo določene pomankljivosti, kot so varnostna vprašanja ki bi lahko podjetja pri delu z občutljivimi informacijami spravila v težave. Varnost velikih podatkov je stalna skrb, saj so uvedbe velikih podatkov dragocene tarče morebitnih vsiljivcev. En sam napad izsiljevalske programske opreme lahko povzroči da izgubimo dostop do podatkov oziroma bi morali za ponovno pridobitev dostopa do njih plačati odkupnino. Kar je še huje, lahko nekdo ki nepooblaščeno dostopa baz podatkov iz njih podatke črpa in prodaja naprej. V nadaljevanju bom opisal nekaj izzivov s katerimi se pri varnosti srečujemo.

Shranjevanje podatkov

Podjetja navadno veliko količino podatkov shranjujejo v oblak oziroma Cloud data storage, s katerim poenostavijo premikanje podatkov in pospešijo poslovanja. Vendar pa se tukaj pojavljajo tveganja lahko eksponentna. Najmanjša napaka pri nadzorih dostopa do podatkov lahko omogoča nezaželjene dostope in pridobitev množice občutljivih podatkov. Posledica tveganj je ta, da večja tehnološka podjetja zdaj podatke shranjujejo tako lokalno kot v oblaku da zagotavljajo varnost in prilagodljivost. Kritične informacije oziroma podatke se shranjuje v lokalne baze podatkov, manj občutljivi podatki pa se zaradi lažje uporabe shranjujejo v oblak. Za izvajanje varnostnih politik v lokalnih bazah podatkov podjetja potrebujejo in najemajo strokovnjake za kibernetsko varnost.

Lažni podatki

Ustvarjanje lažnih podatkov je resna grožnja za podjetja ki podatke zbirajo, ker tratijo čas analitikom, ki bi ga drugače lahko ta čas porabili za prepoznavanje ali reševanje drugih težav. Lažni podatki lahko povzročajo zmanjšano proizvodnjo ali druge procese, ki so potrebni za vodenje podjetja. Podjetja so primorana biti kritična do podatkov s katerimi upravljajo in obdelujejo za izboljšanje poslovnih procesov.

Eden od načinov za izogib lažnim podatkom je potrditev virov podatkov z rednimi ocenami in vrednotenje modelov strojnega učenja z različnimi testnimi nabori podatkov, da bi med njimi našli anomalije.

Kaj pa in kako s kreativnostjo? Kdaj se nam porajajo nove ideje in nova razmišljanja? Kdaj se brez strahu postavljamo pred izzive? Pred izzive se postavljamo takrat, ko smo prosti in svobodni v svojih dejanjih in v svojem mišljenju. Takrat smo najbolj kreativni, ne poznamo več samega sebe in se ne čudimo nad našimi dejanji in nad našim obnašanjem. Kot otrok se igramo s čimer koli, kot otrok smo zmožni kuhinjski pokrov spremeniti v vesoljsko ladjo.

Zasebnost podatkov

Eden večjih izzivov v digitalnem svetu je zagotavljanje zasebnosti podatkov. Potrebno je zaščititi osebne in občutljive podatke pred kršitvami, namerno in nenamerno izgubo podatkov ter pred kiberntskimi napadi. Upoštevati je potrebno stroga načela zasebnosti podatkov in s tem okrepiti varstvo podatkov. Splošna pravila so poostren nadzor nad shrambami podatkov, poznavanje svojih podatkov, ažurno varnostno kopiranje podatkov, varovanje omrežja pred nepooblaščenimi dostopi, redno izvajanje ocen tveganja in usposabljanje uporabnikov o varnosti in zasebnosti podatkov.

Upravljanje s podatki

Posledice kršitve varnosti so lahko grozljive za podjetja od ranljivosti kritičnih poslovnih informacij do popolnoma kompromitiranih baz podatkov. Uvajanje čim višje stopnje varnosti za baze podatkov je ključnega pomena. Tisti najboljši sistemi za upravljanje z bazami podatkov so opremljeni z različnimi kontrolami dostopov. Potrebno je upoštevati tako fizično varnostno prakso kot tudi varnostne ukrepe ki temeljijo na opremi za zaščito shranjenih podatkov. Pri tem nam pomaga šifriranje podatkov, segmentiranje in particioniranje podatkov in implementacija zaupanja vrednih strežnikov. Priporočljivo je bazo podatkov opremiti tudi z orodji, ki bazo podatkov spremljajo in nam sporočijo če zaznajo da je baza podatkov kompromitirana.

Nadzor dostopa

Nadzorovanje dostopov, s katerim določamo kateri zaposleni lahko dostopajo ali urejejajo katere podatke nam omogoča ne samo zagotavljanje celovitosti podatkov ampak tudi njihovo zasebnost. Seveda pa upravljanje nadzorov dostopa ni enostavno, še posebaj če govorimo o večjih podjetjih ki imajo več sto zaposlenih. Identity access managment (IAM) opravlja nadzor pretoka podatkov preko identifikacije, preverjanja pristnosti in avtorizacije (Oracle, b. l.). Sledenje ustreznim standardom ISO je dobro izhodišče za zagotovitev da organizacije sledijo najboljšim praksam uporabe IAM.

Zastrupitev podatkov

Obstaja več rešitev za strojno učenje, na primer klepetalni roboti, ki se učijo in nadgrajujejo na ogromni količini podatkov. Prednosti so, da se med uporabo in interakcijo z uporabniki nenehno izboljšujejo. To lahko vodi do zastrupitve podatkov, način kako napadati modelov strojnega učenja. Ponavadi se šteje kot napad na integriteto, saj z napadom spremenimo oziroma vnašamo podatke, ki vplivajo na sposobnost, da strojno učenje zagotovi pravilne napovedi. Rezultati teh napadov so lahko poškodovanje logike modela za strojno učenje, manipulacija podatkov, vbrizgavanje podatkov (injection). Najboljši način za premagovanje teh napadov je zaznavanje izstopajočih podatkov, s čimer se vbrizgani (injected) podatki v takoimenovanem vadbenem bazenu ločijo od obstoječe distribucije podatkov.

Kraja zaposlenih

Napredne podatkovne kulture so zaposlenim omogočile, da imajo dostop do določenih kritičnih poslovnih informacij. Čeprav spodbuja demokratizacijo, je tveganje da bi zaposleni ponesreči ali namerno "leakal" podatke veliko. Kraje podatkov z strani zaposlenih se dogajajo tako v velikih korporacijah kot tudi v startup podjetjih. Da bi se čim bolj efektivno izgonili kraji podatkov, bi morala podjetja izvajati pravne politike skupaj z varovanjem omrežij z VPN-ji. Prav tako je priporočljiva uporaba storitve Desktop as a Service (Daas). DaaS omogoča pregledno in sledljivo upravljanje naprav, kar pomeni da je mogoče prej prepoznati tveganja oziroma namere kraje podatkov in jih omejiti. Prav tako omogoča analitiko zmogljivosti naprav in nam pomaga pri zamenjavi in nadgradnji opreme, ki bi lahko bila nevarna za kibernetsko varnost.

Zaključek

Nove tehnologije imajo skoraj vedno tako dobre kot slabe lastnosti. Prednosti uporabe velikih podatkov so recimo: zmanjšanje stroškov podjetij s pomočjo optimizacije dobavnih verig, zalog, načrtovanju naročil, personalizirana izkušnja za kupce, kjer jim z pomočjo analize velikih podatkov potem pošiljajo prilagojene ponudbe in posebne popuste ter sezname priporočil, ki so ustvarjeni na podlagi podatkovnih baz. Na podlagi podatkov se lahko optimizira podpora in pomoč uporabnikom, saj se z spremljanjem podatkov (na primer čas dostave izdelka) lahko prepoznajo in odpravijo težave v delovnem procesu. Orodja za analizo prav tako lahko spremljajo konkurente in na podlagi podatkov se potem optimizira in ustvarja konkurenčna cena izdelkov ali storitev. Tudi na drugih področjih pa uporaba analize Big data omogoča nešteto različnih prednosti.

Obstajajo pa tudi določene slabe lastnosti predvsem v okviru varnosti občutljivih podatkov, s katerimi upravljajo podjetja. Veliki podatki so dragocene tarče kibernetičnih napadalcev, ki z uporabo izsiljevalske programske opreme lahko povzročijo izgubo dostopa do teh podatkov, ali pa z nepooblaščenim dostopom do podatkovnih baz iz njih črpajo podatke in jih širijo oziroma prodajajo naprej.

Na podlagi navedenih pomislekov, je očitno zakaj podjetja in druge organizacije, ki te tehnologije uporabljajo, vidijo varnost velikih podatkov kot glavno skrb. Dobra novica je ta, da lahko z pravimi viri, informacijami, kvalificiranimi in izobraženimi delavci, dobro strategijo obvladovanja in predanostjo celovitosti in zaščiti zasebnih podatkov zlahka rešimo oziroma se izognemo večini tovrstnih tveganj.

Viri in literatura

Big Data Analytics. IBM. (b.l.). Pridobljeno 25.01.2022 s svetovenega spleta: https://www.ibm.com/analytics/big-data-analytics

Big Data and analytics: Definitions, value, trends and applications. (b.l.). Pridobljeno 23.01.2022 s svetovenega spleta: https://www.i-scoop.eu/big-data-action-value-context/

Big Data. BuiltIn. (b.l.). Pridobljeno 25.01.2022 s svetovenega spleta: https://builtin.com/big-data

Big data. Wikipedia. (b.l.). Pridobljeno 23.01.2022 s svetovenega spleta: https://en.wikipedia.org/wiki/Big_data

Big data: What it is and why it matters. SAS. (b.l.). Pridobljeno 25.01.2022 s svetovenega spleta: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

What is Big Data? Oracle. (b.l.). Pridobljeno 23.01.2022 s svetovenega spleta: https://www.oracle.com/big-data/what-is-big-data/

Berthiaume, D. (2019). Study: The omnichannel features customers most want are. Pridobljeno 23.1.2022 s svetovnega spleta: https://www.chainstoreage.com/technology/study-the-omnichannel-features-customers-most-want-are

Easy web scraping for anyone. Octoparse. (b.l.). Pridobljeno 24.1.2022 s svetovnega spleta: https://www.octoparse.com/