UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

ZAKLJUČNA STROKOVNA NALOGA VISOKE POSLOVNE ŠOLE

UVEDBA PODATKOVNEGA SKLADIŠČA V IZBRANEM PODJETJU

Ljubljana, marec 2022 MUHAMED ORAŠČANIN

IZJAVA O AVTORSTVU

Podpisani Muhamed Oraščanin, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Uvedba podatkovnega skladišča v izbranem podjetju, pripravljenega v sodelovanju s svetovalcem doc. dr. Luko Tomatom

I Z J A V L J A M

1. da sem predloženo delo pripravil samostojno;

2. da je tiskana oblika predloženega dela istovetna njegovi elektronski obliki;

3. da je besedilo predloženega dela jezikovno korektno in tehnično pripravljeno v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani, kar pomeni, da sem poskrbel da so dela in mnenja drugih avtorjev oziroma avtoric, ki jih uporabljam oziroma navajam v besedilu, citirana oziroma povzeta v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani;

4. da se zavedam, da je plagiatorstvo – predstavljanje tujih del (v pisni ali grafični obliki) kot mojih lastnih – kaznivo po Kazenskem zakoniku Republike Slovenije;

5. da se zavedam posledic, ki bi jih na osnovi predloženega dela dokazano plagiatorstvo lahko predstavljalo za moj status na Ekonomski fakulteti Univerze v Ljubljani v skladu z relevantnim pravilnikom;

6. da sem pridobil vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označil;

7. da sem pri pripravi predloženega dela ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil soglasje etične komisije;

8. da soglašam, da se elektronska oblika predloženega dela uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;

9. da na Univerzo v Ljubljani neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve predloženega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja predloženega dela na voljo javnosti na svetovnem spletu preko Repozitorija Univerze v Ljubljani;

10. da hkrati z objavo predloženega dela dovoljujem objavo svojih osebnih podatkov, ki so navedeni v njem in v tej izjavi.

V Ljubljani, dne _____ Podpis študenta: _____

KAZALO

KAZALO TABEL

KAZALO SLIK

SEZNAM KRATIC

angl. – angleško
ETL – (angl. Extract, Transform, Load and Loag); Zajemanje, preoblikovanje in nalaganje
SWOT – (angl. Strengths, Weaknesses, Opportunities, Threats); Prednosti, slabosti, priložnosti in nevarnosti
OLAP – (angl. online analytical processing); Sprotna analitična obdelava podatkov
ERP – (angl. Enterprise Resource Planing); Celovite programske rešitve
TDWI – (angl. Transforming Data with Intelligence); Spreminjanje podatkov z inteligenco
SVOT – (angl. Single Version of the Truth); Edinstvena verzija resnice
DSA – (angl. Data Staging Area); Področje za obdelavo podatkov
CSV – (angl. Comma-Separated Values); Z vejico ločene vrednosti
SQL – (angl. Structured Querry Langugage); Strukturiran povpraševalni jezik
ER – (angl. Entity Relationship Diagram); Diagram entitet povezav

UVOD

Podatki so v podjetjih ključnega pomena za analizo poslovanja in odločanja, vendar morajo biti pravilno shranjeni in hitro dosegljivi uporabnikom, saj v nasprotnem ne predstavljajo prednosti, temveč lahko predstavljajo celo slabost. Sama uporabnost in dosegljivost podatkov sta ob večjih količinah slednjih lahko ničelna, zato podjetja, ki poslujejo z večjo količino podatkov in potrebujejo hitre poizvedbe, nemalokrat v svoje poslovanje uvedejo uporabo podatkovnega skladišča (Golfarelli & Rizzi, 2009).

Podatke iz poslovanja podjetja in zunanjih virov tako shranjujemo v sistem, ki mu rečemo podatkovno skladišče. Po definiciji je podatkovno skladišče zbirka entitetno usmerjenih, integriranih, časovno odvisnih in ne spreminjajočih se podatkov, ki so namenjeni podpori odločitvenim procesom (Inmon, 2002).

Vsakodnevne odločitve temeljijo na natančnosti podatkov. Slabi in nepravilni podatki lahko pripeljejo do zavajajočih in napačnih informacij, ki lahko škodujejo podjetju (Vasudev, 2015).

Pri svojem delu sem skušal odgovoriti na raziskovalno vprašanje, ali uvedba podatkovnega skladišča z namenom analiziranja podatkov, podjetju prinaša konkurenčno prednost? Namen zaključne naloge je tako na primeru uvedbe podatkovnega skladišča za namen podatkovne analitike v računovodskem servisu prikazati možne koristi, ki jih podjetjem lahko prinese uporaba podatkovnega skladišča. Pri tem sem zasledoval naslednje cilje:

- opredeliti značilnosti podatkovnih skladišč;
- predstaviti proces pridobivanja, preoblikovanja in nalaganja podatkov (angl. Extract, Transform, Load, ETL);
- predstaviti pomen analitike masovnih podatkov v podatkovnih skladiščih;
- predstaviti potek uvedbe podatkovnega skladišča v izbrani računovodski servis;
- prepoznati koristi uvedbe podatkovnega skladišča v izbrani računovodski servis.

Zaključna naloga je razdeljena na teoretični in empirični del. V teoretičnem delu sem predstavil podatkovna skladišča, njihovo strukturo in arhitekturo, primerjavo pristopov pri gradnji, proces zajemanja, preoblikovanja in nalaganja ter same tehnološke zahteve podatkovnih skladišč. V empiričnem delu sem predstavil podjetje, načrt, zahteve, uvedbo podatkovnega skladišča v računovodski servis ter koristi, ki jih uvedba podatkovnega skladišča prinaša podjetju.

Pri obravnavi raziskovalnih problemov sem uporabil kvalitativno, opisno metodo znanstvenoraziskovalnega dela, s katero sem preučil in analiziral domačo in tujo strokovno in znanstveno literaturo, članke, objave, publikacije in ostale relevantne vire. Koristi uvedbe podatkovnega skladišča sem s pomočjo analize PSPN (angl. Strengths, Weaknesses,

Opportunities, Threats, v nadaljevanju SWOT) prikazal na praktičnem primeru izbranega podjetja.

Zaključno nalogo sem razdelil na 4 poglavja. V prvem poglavju sem predstavil vrste podatkovnih skladišč, njihov namen, podrobno arhitekturo podatkovnih skladišč in predstavil najpogostejše, tako teoretično kot tudi vizualno. V nadaljevanju sem primerjal pristope pri gradnji podatkovnih skladišč. V prvem delu sem predstavil tudi proces ETL ter njegovo vlogo in na koncu prvega dela še tehnološke zahteve aktivnih podatkovnih skladišč. V drugem poglavju sem predstavil vrste in metode analitike masovnih podatkov, analitična orodja, spletno analitično obdelavo (angl. Online Analytical Process, v nadaljevanju OLAP) in podatkovni model ter možnost uporabe analitike masovnih podatkov v podatkovnih skladiščih. V tretjem poglavju zaključne naloge sem predstavil podjetje, kjer bi uvedli podatkovno skladišče, sam načrt uvedbe, značilnosti in tehnološke zahteve načrtovane uvedbe, analizo pripravljenosti za podporo pri upravljanju podatkovnih skladišč in na koncu še PSPN analizo koristi uvedbe podatkovnega skladišča.

1 PODATKOVNA SKLADIŠČA

Potreba po sistemih, ki bodo ponujala pomoč pri odločanju, se je pojavila v začetku 80 let prejšnjega stoletja, ko sta podjetji ACNielsen in IRI predstavili sistem, ki bo podpiral prodajo z informacijskim sistemom. Ob koncu istega desetletja, ko IBM System Journal objavi članek "An architecture for a business and information systems", se prvič pojavi izraz "poslovno podatkovno skladišče" (angl. Business Data Warehouse) in pojasni proces skladiščenja podatkov, ki je kot dobra praksa znan še danes (Koren, 2010).

Rainardi (2008) je podatkovno skladišče opisal kot sistem, ki pridobiva in združuje podatke iz primarnih transakcijskih sistemov v dimenzijsko oziroma normalizirano podatkovno skladišče. Dodal je, da se v podatkovnih skladiščih po navadi najdejo podatki stari več let, ki se s pomočjo sistema uporabljajo za analitične poizvedbe.

Hranjenje vseh ključnih podatkov za poslovno odločanje organizacije je primarni namen podatkovnega skladišča. Celovite programske rešitve (angl. Enterprise Resource Planing, v nadaljevanju ERP), transakcijski sistemi, sistemi za planiranje, aplikacije za podporo poslovanja in ostali operativni programi so le nekateri viri, iz katerih se v podatkovno skladišče zbirajo podatki. ERP predstavlja celovito programsko rešitev, ki podjetjem ponuja enostavno podporo na vseh področjih poslovnih procesov znotraj in zunaj organizacije (planiranje, proizvodnja, prodaja, distribucija, računovodstvo itd ...) z optimalno uporabo virov, zaradi česar ERP sistemi veljajo kot pogost in kakovosten vir podatkov, ki jih skladiščimo v podatkovnih skladiščih (Kovačič, 2004). Podpora poslovanja in vsakodnevne odločitve so z uvedbo podatkovnih skladišč skupaj z orodji za predstavitev podatkov in ad-hoc analizami kakovostnejše. Podatkovno skladišče zagotavlja, da se zaposleni izognejo nepotrebnemu ponavljajočemu se delu pri pripravi podatkov, hkrati pa zmanjša verjetnost

napak pri sami pripravi, modeliranju in na koncu predstavitvi podatkov. Z uvedbo podatkovnega skladišča imajo zaposleni več časa, ki ga lahko namenijo višjim miselnim procesom in uporabi že pripravljenih podatkov za poslovanje (Olenik, 2019).

1.1 Značilnosti podatkovnih skladišč

Podatkovna skladišča so se razvila zaradi potrebe po shranjevanju vedno večjih količin podatkov in potrebe po nadaljnji analizi le teh. Cilj podatkovnih skladišč je pridobljene podatke preoblikovati v poslovni vir znanja, ne glede na način pridobitve teh podatkov. V primeru neusklajenosti podatkov lahko pride do napačnih izvlečkov, ki lahko pripeljejo do napačnih odločitev in neuspeha podjetja, zato je treba skrbeti za usklajenost vseh podatkov po pravilih in lastnostih, ki so že vnaprej določena, saj potem lahko podatke lažje uporabimo za nadaljnje poizvedbe. Te lastnosti najdemo v definiciji podatkovnih skladišč, ki nam jo je postregel Inmon (2005), ki podatkovna skladišča razlaga kot integrirano, entitetno usmerjeno, časovno odvisno in nespremenljivo zbirko podatkov, ki jo lahko uporabimo kot podlago za poslovne odločitve.

Integriranost podatkov pomeni, da je podatkovno skladišče centralizirano, usklajeno, urejeno in da ima uporabnik podatkovnega skladišča konsistenten in enoten pogled na podatke, čeprav so podatki pridobljeni iz med seboj nedoslednih virov (Kimball & Caserta, 2004).

Entitetna usmerjenost, zaradi lažje preglednosti in boljše usklajenosti, organizira in ureja podatkovno skladišče po oddelkih v podjetju, kot so vodstvo, skladiščenje ali računovodstvo. Podatkovno skladišče za vsak oddelek ima določene entitete (npr. cena, stranka, količina, izdelek itd.) (Coronel, Morris & Rob, 2011).

Časovna odvisnost ukazuje, da so podatki v podatkovnih skladiščih v preteklem času. Podatki se v podatkovna skladišča nalagajo periodično (enkrat dnevno, tedensko itd.), poleg tega pa jih lahko prikažemo oziroma predstavimo v različnih časovnih formatih (dan, teden, mesec ,...). Najbolj razširjena praksa je, da se podatki prikazujejo po mesecih, četrtletjih, polletjih in letih (Kimball & Ross, 2002).

Nespremenljivost ali obstojnost pravi, da podatki, zapisani v podatkovnih skladiščih, niso namenjeni spreminjanju in brisanju, zato lahko rečemo, da ti podatki predstavljajo zgodovino podjetja. Povedano drugače, ko se "živi" podatki vpišejo v podatkovno skladišče ob osvežitvi, postanejo zgodovinski podatki in niso več uporabni v operativnih aplikacijah. Sistem omogoča pregledovanje in shranjevanje podatkov, starih tudi več kot 10 let (Coronel, Morris & Rob, 2011).

1.2 Arhitektura podatkovnih skladišč

Ko začnemo govoriti o gradnji podatkovnega skladišča, se moramo med prvimi koraki vprašati po izboru arhitekture za izgradnjo. Najprej moramo določiti, kakšno podatkovno skladišče sploh želimo imeti in šele nato na arhitekturo. Najbolj uporabljeni in priljubljeni pristopi so "od zgoraj navzdol", "od spodaj navzgor" in hibridni pristop. Od primera do primera je odvisno, kateri pristop je bolj učinkovit (Rangarajan, 2016).

Od samega pristopa je v nadaljevanju odvisna arhitektura podatkovnega skladišča. Pristop »od zgoraj navzdol« zagovarja Inmon, nastane pa centralizirana arhitektura. Pristop "od spodaj navzgor" zagovarja Kimball, nastane pa distribuirana arhitektura. Hibridni pristop je mešan pristop prvih dveh in tako dobimo tretjo, federativno arhitekturo (Stoilkovič, 2009).

1.2.1 Centralizirana arhitektura

Pristop "od zgoraj navzdol" gleda na podatkovno skladišče kot ključni dejavnik celotnega analitičnega okrožja. Najprej se iz vira podatkov dobavijo in transformirajo vsi podatki, nato pa se shranijo v skladišče podatkov. Od tod so podatki povzeti, dimenzionirani in distribuirani v en ali več področnih podatkovnih skladišč (angl. Data Marts) za potrebe različnih enot znotraj neke organizacije. V določeno področno skladišče so všteti vsi podatki, ki so relevantni za določeno poslovno področje, korporacijski oddelek ali kategorijo uporabnikov. Področna skladišča se pogosto nazivajo z nesamostojnimi oziroma odvisnimi skladišči, saj vse svoje podatke izvažajo iz centraliziranega skladišča podatkov (Pirc, 2007). Slika 1 prikazuje centralizirano arhitekturo.

Slika 1: Centralizirana arhitektura

Prirejeno po Pirc (2007).

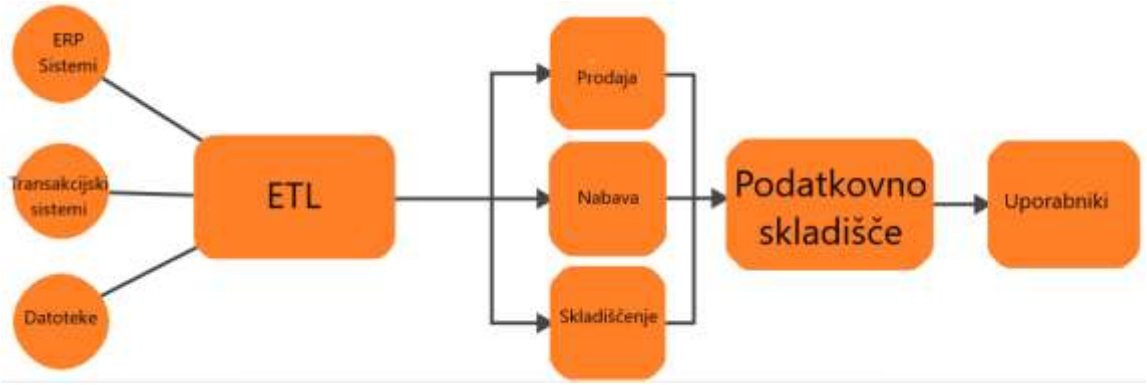
Čeprav področna skladišča v bistvu sploh niso potrebna, so koristna za sisteme skladiščenja podatkov v srednje velikih podjetjih, ker (Seiner, 2007):

- se uporabljajo kot gradniki, medtem ko se podatkovno skladišče postopoma razvija;
- označujejo informacije, ki jih zahteva določena skupina uporabnikov za izdelavo poizvedb;
- so lahko bolj učinkovita zaradi manjše količine podatkov od centraliziranega skladišča.

1.2.2 Distribuirana arhitektura

V pristopu "od spodaj navzgor" je cilj ustvariti podatkovno skladišče, ki mora biti hitro in razumljivo. Ta pristop sprejeme in transformira vse podatke, ampak so ta najprej razvrščena v področno skladišče in šele po tem v osrednje podatkovno skladišče. Za razliko od pristopa "od zgoraj navzdol", področna podatkovna skladišča v tem pristopu vsebujejo atomske in zgodovinske podatke, ki jih uporabniki potrebujejo ali jih bodo potrebovali v prihodnosti. Distribuirana arhitektura zmanjša odvečnost podatkov in olajša razširitev obstoječih dimenzijskih modelov za prilagoditev novih območij. Podatki so bili modelirani z zvezdno shemo za optimizacijo uporabnosti in uspešnosti poizvedb (Kimball, Ross, Thornthwaite, Mundy & Becker, 1998). Slika 2 prikazuje distribuirano arhitekturo.

Slika 2: Distribuirana arhitektura



Prirejeno po Kimball, Ross, Thornthwaite, Mundy & Becker (1998).

1.2.3 Federativna arhitektura

Federativni oziroma hibridni pristop poskuša združiti najboljše specifikacije iz prvih dveh pristopov. Poskuša izkoristiti hitrost in uporabniško orientiranost pristopa od spodaj navzgor, brez da bi škodovali integraciji, ki je značilna za pristop od zgoraj navzdol. Če sta Anmon in Kammball najglasnejša zagovornika prvih dveh pristopov je Pieter Mimno, neodvisni svetovalec, ki predaja o spreminjanju podatkov z inteligenco (angl. Transforming Data with Intelligence, TDWI), najbolj glasen zagovornik federativne arhitekture (Seiner, 2007).

Hibridni pristop veleva, da se prva 2 tedna namenita razvoju poslovnega modela v tretji normalni formi, preden se razvije prvi podatkovni program. Nekaj prvih področnih podatkovnih skladišč je tudi oblikovanih v tretji normalni obliki, ampak se uporabljajo s pomočjo fizičnih modelov zvezdne sheme. Ta dvojni modelni pristop združuje poslovni model brez žrtvovanja uporabnosti in uspešnosti izvedbe poizvedbe zvezdne sheme (Seiner, 2007).

Hibridni pristop temelji na orodju ETL za shranjevanje in upravljanje poslovnih in lokalnih podatkovnih modelov v področnih podatkovnih skladiščih ter sinhronizacijo razlik med njimi. To na primer omogoča lokalnim skupinam, da razvijejo lastne definicije in pravila za podatkovne elemente, ki izhajajo iz poslovnega modela, ne da bi pri tem žrtvovali dolgoročno integracijo. Po implementaciji prvih nekaj področnih podatkovnih skladišč se začne polniti centralno podatkovno skladišče. Atomski podatki se nato prenesejo iz področnih podatkovnih skladišč v osrednje podatkovno skladišče in redundantni viri podatkov se konsolidirajo, kar prihrani čas, denar in vire za obdelavo organizacije. Organizacije običajno zapolnijo podatkovno skladišče, ko uporabniki podjetja zahtevajo, da se atomski podatki prikažejo v več regionalnih podatkovnih skladiščih (Seiner, 2007). Slika 3 prikazuje federativno arhitekturo.

Slika 3: Federativna arhitektura

Prirejeno po Seiner (2007).

1.3 Primerjava pristopov pri gradnji podatkovnega skladišča

Vse tri pristope sem primerjal z vidika prednosti in slabosti. Začnemo pri prednostih prvega pristopa, "od zgoraj navzdol", ki ponuja integrirano, fleksibilno arhitekturo, ki podpira nizvodne strukture analitičnih podatkov. To pomeni, da podatkovno skladišče predstavlja izhodišče za vse podatke, vključujoči doslednost in standardizacijo, da bi organizacije dosegle tako imenovano "edinstveno verzijo resnice" (angl. Single Version of the Truth, v nadaljevanju SVOT). SVOT je tehnična zasnova, ki opisuje idealno skladiščenje podatkov, ki ima eno centralizirano bazo podatkov ali vsaj distribuirano sinhronizirano bazo podatkov, ki shranjuje vse podatke organizacije v dosledni in ne-redundantni obliki. Atomski podatki

v skladišču omogočajo organizacijam, da ponovno uporabljajo te podatke na kakršen koli način, kako bi zadovoljile nove in nepričakovane poslovne potrebe. Na primer, podatkovno skladišče se lahko uporablja kot bogat vir podatkov za namene statistike in raznih poročil. Poleg tega uporabniki lahko pošljejo poizvedbo po več funkcionalnih oziroma poslovnih pogledih na podatke (Seiner, 2007).

Slabost prvega pristopa je, da lahko implementacija podatkovnega skladišča vzame več časa in da stane več, kakor bi stali ostali pristopi, posebej v začetnih korakih. Razlog za to je, ker organizacije morajo, preden sploh začnejo razvijati aplikacije in poročila, najprej ustvariti logičen in podroben model podatkov podjetja, kot tudi fizično infrastrukturo, ki bo zajemala prostor ter mesto podatkovnega skladišča (Naeem, 2020).

Glavna prednost drugega pristopa, "od spodaj navzgor", je ta, da se osredotoča na ustvarjanje enostavnih, fleksibilnih struktur podatkov s pomočjo dimenzijskih, zvezdinih shem. Druga prednost je ta, da podatkovna skladišča shranjujejo zgodovinske in atomske podatke, kar posledično pomeni, da uporabniki z "rudarjenjem" dostopajo do vseh podrobnih ali transakcijskih podatkov že v področnem skladišču podatkov. Eden izmed problemov tega pristopa je, da zahteva od organizacije, da vsilijo uporabo standardnih dimenzij in dejstev za zagotovitev integracije in SVOT. Enkrat, ko so področna podatkovna skladišča logično po razporejena znotraj fizične baze podatkov, je integracijo lahko doseči. Toda v porazdeljeni, decentralizirani organizaciji je morda preveč zahtevati od oddelkov in poslovnih enot, da se držijo in ponovno uporabljajo priporočila in pravila za izračun dejstev. Tukaj lahko organizacija teži k ustvarjanju "neodvisnih" ali ne integriranih podatkovnih skladišč (Seiner, 2007).

Glavna prednost hibridnega pristopa je, da kombinira hitro tehniko razvoja znotraj okvirjev arhitekture podjetja, toda dopolnjevanje podatkovnega skladišča je lahko v tem pristopu zelo destruktiven proces, ki ne prinaša nikakršne očitne vrednosti in zato morda nikoli ne bo mogel biti stroškovno učinkovit. Poleg tega, le nekaj orodij za poizvedbe, lahko dinamično in inteligentno išče atomske podatke v eni bazi podatkov in zbirne podatke v drugi bazi podatkov. Obstaja možnost, da celo uporabniki ne bodo bili prepričani kdaj morajo poslati poizvedbo v katero bazo (Jindal & Acharya, 2019).

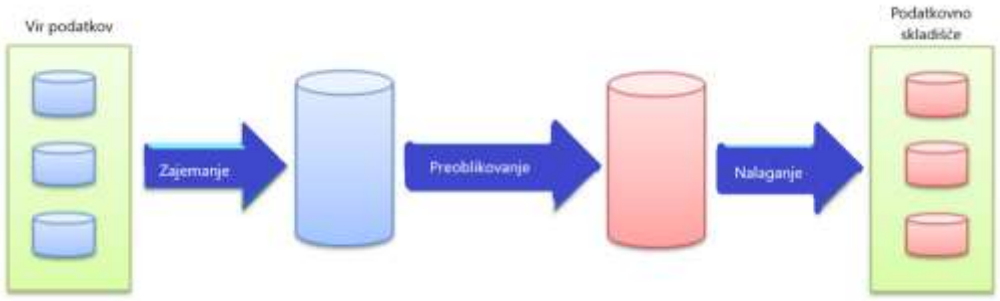
1.4 Proces ETL

Podatki v podatkovnih skladiščih morajo biti usklajeni in morajo ustrezati prej določenim pravilom. Za integracijo podatkov pred vpisom v podatkovno skladišče podatke prečistimo s procesom, ki mu pravimo ETL proces. Proces ETL združuje postopke zajemanja (pridobivanja), preoblikovanja in nalaganja podatkov in se kot, odgovorni za operacije, odvija v ozadju arhitekture podatkovnega skladišča.

Običajno se pridobivajo samo podatki, ki se razlikujejo od predhodno dokončanega procesa ETL, kar pomeni, da se pridobijo le podatki, ki so na novo dodani, posodobljeni ali izbrisani.

Po tej fazi se pridobljeni podatki prenesejo v prostor za posebne namene v podatkovnih skladiščih, ki ga imenujemo področje za obdelavo podatkov (angl. Data Staging Area, DSA), kjer se odvija njihovo preoblikovanje, homogenizacija in samo čiščenje. Posebno pozornost je treba nameniti potrebnemu preoblikovanju, preverjanju in čiščenju podatkov, saj prav to najbolj vpliva na uspešno delovanje podatkovnega skladišča. Najpogosteje uporabljena preoblikovanja vključujejo filtre in preverjanja, ki zagotavljajo, da so podatki, preneseni v podatkovno skladišče, skladni s pravili in omejitvami integritete, kot tudi sheme preoblikovanja, ki zagotavljajo, da se podatki ujemajo s ciljno shemo podatkovnega skladišča. Na samem koncu se podatki shranjujejo v osrednje podatkovno skladišče in področna podatkovna skladišča (Vassiliadis & Simitsis, 2009).

Slika 4: Vizualni prikaz procesa ETL



Vir: Faizaan (2015).

Kot je razvidno na sliki 4 je prvi korak v procesu ETL postopek pridobivanja podatkov. Postopek je lahko precejšnji sistemski zalogaj, saj pogosto pridobivamo podatke, ki imajo različen format, strukturo podatkov in nemalo krat so pridobljeni iz različnih virov. V tem koraku se vsi podatki preverijo in pretvorijo v obliko, ki je predhodno določena. Pogosto se morajo, za naslednjo fazo, podatkovne baze iz izvornih relacijskih in ne-relacijskih pripraviti s težavnostjo. Glede na tehnološko infrastrukturo in naravo izvornega sistema (relacijska baza podatkov, preglednica, spletna stran itd.) ter količino podatkov, ki jih je treba obdelati, se lahko za korak iskanja podatkov sprejmejo različna pravila. Končna točka v koraku pridobivanja podatkov zaradi varnosti in razlogov za implementacijo omrežja vključuje potrebo po šifriranju in stiskanju podatkov prenesenih iz vira v shranjevanje (Kimball & Caserta, 2004).

Drugi korak je preoblikovanje podatkov s pomočjo funkcij na podlagi določenih pravil, kar v nadaljevanju procesa omogoči shranjevanje pridobljenih podatkov. Od posameznega vira podatkov in samega poslovnega modela podjetja je odvisno kako zahteven bo proces preoblikovanja. Najbolj pogosti procesi, ki se odvijajo v tej fazi, so prevajanje in čiščenje podatkov, računanje novih vrednosti, združevanje podatkov iz več različnih virov in nalaganje samo določenih stolpcev podatkov ali razcepljanje le teh v več različnih stolpcev (Kimball & Caserta, 2004).

Zadnji korak je nalaganje pretvorjenih podatkov v ustrezne tabele. Doslednost podatkov je treba vzdrževati, ker se lahko zapisi spremenijo ob nalaganju. Nalaganje se lahko izvede na dva načina. Prvi način je z osvežitvijo, ko se podatki podatkovnega skladišča popolnoma prepišejo, kar pomeni, da se stari podatki zamenjajo z novimi. Drugi način so posodobitve, ko se v podatkovnem skladišču dodajo le tiste spremembe, ki so se zgodile na izvornih podatkih (Kimball & Caserta, 2004).

1.5 Tehnološke zahteve aktivnih podatkovnih skladišč

Za zadovoljivo obdelavo podatkovnega skladišča so potrebne nekatere tehnološke lastnosti. To vključuje robusten jezikovni vmesnik, podporo za kompleksne ključe in podatke spremenljive dolžine, kot tudi možnost upravljanja velikih količin podatkov, upravljanje podatkov na različnih medijih, enostavno indeksiranje in spremljanje podatkov in sodelovanje s široko paleto tehnologij. Poleg tega je nujno vzporedno shranjevanje in dostop do podatkov, nadzor nad metapodatki skladišča, učinkovito nalaganje repozitorija, učinkovita uporaba indeksov, kompaktno shranjevanje podatkov, selektiven izklop upravljalnika zaklepanja, izvedba obdelave samo indeksa in hitro vzpostavitev (vrnitev) podatkov iz pomnilnika. Sodobni sistemi izpolnjujejo vse te zahteve (Inmon, 2005).

2 ANALITIKA MASOVNIH PODATKOV V PODATKOVNIH SKLADIŠČIH

Podatkovna skladišča so v bistvu le zbirke podatkov, ki so namenjene samo za branje, toda s pomočjo poizvedovalnih in analitičnih programov, lahko te podatke uporabimo za različne namene. Ker skladišča običajno vsebujejo veliko količino podatkov, govorimo o masovnih podatkih (angl. Big Data). Proces zbiranja in obdelave tako strukturiranih kot nestrukturiranih podatkov imenujemo tehnologija masovnih podatkov. Ti podatki so lahko zbrani v različnih oblikah, denimo slikah, besedilih, videoposnetkih in drugih (Stepinac, 2014).

Termin, masovni podatki, predstavlja veliko količino podatkov, ki so se lahko zbirali daljše časovno obdobje in so lahko strukturirani ali nestrukturirani. Masovne podatke je zelo težko analizirati v tradicionalnih analitičnih programih in standardnih statističnih sistemih, saj so količine velike, hitrorastoče in se pogosto razlikujejo od obstoječih struktur podatkovnih baz (Dumbill, 2012, str. 3).

Samo zbiranje podatkov je zahvaljujoč se internetu postalo relativno hitro in lahko, saj vsak uporabnik, zavedajoč se ali ne, pušča podatke za seboj. Ti podatki lahko ponudijo odgovor na marsikatero vprašanje, poleg tega pa se lahko uporabniku, na podlagi teh odgovorov, ponudijo razne promocije in ponudbe, ki bi ga lahko zanimale. Podatkov je vse več, posledično tudi potrebe po analizi le teh. Zahvaljujoč se tehnologiji masovnih podatkov, možne so razne analize, simulacije in razvoji. Mnoga podjetja tehnologijo masovnih

podatkov uporabljajo za razvoj uspešnega poslovanja ter sprotnega reševanja poslovnih problemov (Stihović, 2015).

Masovne podatke po navadi prepoznamo po treh karakteristikah, ki jim rečemo »3V«, ki predstavljajo količino podatkov (angl. Volumen), ki se zbirajo in pripravljajo za analitiko, hitrost (angl. Velocity), ki predstavlja kontinuirano zbiranje velikih količin podatkov v realnem času in raznolikost (angl. Variety), ki govori, da so podatki po navadi raznolike oblike ter raznolikega izvora (Stepinac, 2014).

Kasneje se pojavijo še spremenljivost (angl. Variability), ki predstavlja nekonsistentnost podatkov, verodostojnost (angl. Veracity), ki predstavlja točnost in kakovost podatkov ter kompleksnost (angl. Complexity), ki predstavlja povezovanje baz med seboj (Callegaro & Yang, 2017).

Masovni podatki zavzemajo prostor, toda z dobrim upravljanjem in kakovostno analizo slednjih lahko podjetje pride do številnih rezultatov in odgovorov, ki jih išče. Skupaj z računalnikom in orodji, ki so mu na voljo, lahko uporabnik analizira masovne podatke, jih obdeluje in v njih odkriva logične vzorce in pomembna pravila. Celotnemu postopku rečemo analitika masovnih podatkov (angl. Big Data Analytics) (Vujnovac, 2015).

Fawcett in Provost (2013) trdita, da analitika masovnih podatkov, katere namen je izboljšati odločanje, dopolnjuje pristop, ki se zanaša na intuicijo vodilnih oseb v podjetjih. Raziskave kažejo, da produktivnost podjetja sorazmerno narašča s povečanjem izkoriščanja podatkovnih virov. Poleg tega uporaba analitike masovnih podatkov povečuje vse ekonomske kazalnike konkurenčnosti podjetja. V okviru predstavljenega konteksta se ločijo odločitve, za katere so nujna odkritja v zbirkah podatkov in odločitve, ki se ponavljajo, kjer najmanjše povečanje natančnosti ima vpliv pri sprejemu odločitve. Analitika masovnih podatkov vodi do avtomatizacije številnih odločitev že v računalniških sistemih. Finance in telekomunikacije so bile prve, ki so sprejele avtomatizacijo te vrste. Podjetja v številnih panogah uporabljajo podatkovne vire za povečanje konkurenčne prednosti, dobička in zmanjšanja stroškov.

2.1 Vrste in metode analitike masovnih podatkov

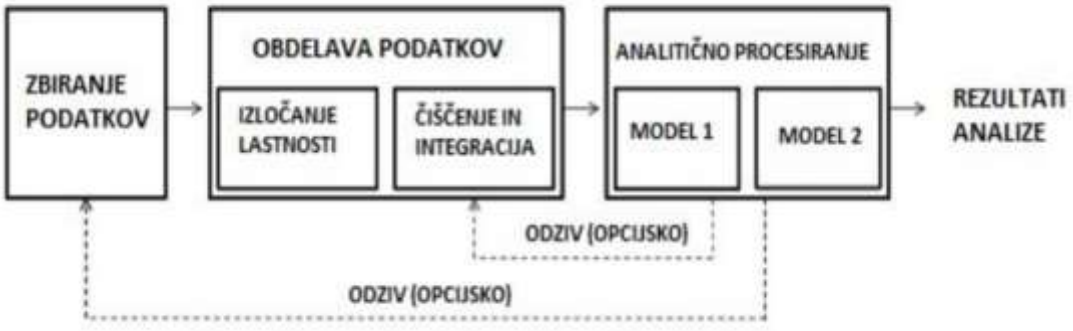
Proces analitike masovnih podatkov se začne z zbiranjem podatkov, nadaljuje z izločanjem lastnosti in čiščenjem slednjih, konča pa se z analitičnim procesiranjem in algoritmi (Aggarwal, 2015). Isti avtor analitiko masovnih podatkov opredeli na naslednji način:

1. prvi korak je zbiranje podatkov, ki je temelj uspešne analize podatkov. Posamezen način zbiranja podatkov lahko v nadaljevanju izključi posamezne metode analitike, zato je pomembno, da se vzpostavi učinkovit načrt zbiranja podatkov;
2. drugi korak je čiščenje podatkov, ki je pomembno zaradi raznolikosti izvornih podatkov. Ključno je, da se podatki pripravijo na način, da bodo ustrezali metodi analize podatkov,

ki se bo uporabila v nadaljevanju procesa, saj je od nje odvisno, katere lastnosti podatkov so ključne za našo analizo. Čiščenje podatkov ter izpostavljanje izbranih lastnosti potekata istočasno, tako da na koncu dobimo urejeno podatkovno zbirko, pripravljeno za samodejno obdelovanje v izbranem programu;

3. tretji korak je izbor analitične metode, na podlagi katere se analizirajo pripravljeni podatki in dobijo odgovori na zastavljena vprašanja.

Slika 5: Analitika masovnih podatkov po korakih



Vir: Aggarwal (2015).

Na sliki 5 je razvidno, da so faze med seboj zelo povezane in da je v vsaki fazi zelo pomembno, kako uspešna je bila predhodna faza. Ključnega pomena je načrtovanje analitike masovnih podatkov, hkrati pa je tudi pomembna prilagoditev podatkov izbrani analitični metodi in samemu znanju, ki ga ima uporabnik (Aggarwal, 2015).

Tehnike analitike masovnih podatkov delimo na usmerjene in neusmerjene. Prve so zasnovane kot »dober« analitični model in imajo jasno zastavljeno ciljno spremenljivko, ki analizo vodi proti točno določenemu odgovoru oziroma rešitvi. Druge se izvajajo na celotni podatkovni zbirki z namenom pridobitve smiselnih in uporabnih znanj, ki bodo prispevala k boljšemu razumevanju povezav v samem poslovanju (Berry & Linoff, 2011).

Poznamo tudi drugo delitev, ki tehnike analitike masovnih podatkov deli na opisne in napovedovalne. Prve kažejo povezave med podatki v podatkovnih skladiščih, kar nam omogoča boljši vpogled in razumevanje podatkov. Druge pa nam omogočajo primerjavo vrednosti trenutnih podatkov in njihovimi rezultati, v kolikor bi dodali nove podatke v sistem ali spremenili stare. Sama tehnika podatkovne analitike se izbere na podlagi tipa, količine in samega problema, ki ga analiziramo oziroma želimo odpraviti (Han, Kamber & Pei, 2012).

Pri tehniki opisovanja se analitiki trudijo najti načine, da opišejo povezave in trende, ki so v podatkih. Analitični modeli morajo biti kar se da transparentni, kar pomeni, da rezultati opisujejo jasne pravilnosti, ki se lahko intuitivno razložijo in interpretirajo. Pomembna značilnost deskriptivne tehnike je tudi združevanje podatkov v skupine, ki temeljijo na

podobnih značilnostih, kar olajša razumevanje podatkov, nadzor in povzemanje. Neke tehnike lahko razlikujemo po tem, koliko je njihova interpretacija transparentna. Recimo tehnika odločitvenih dreves je ljudem zelo razumljiva in pogosto krat zelo intuitivna, medtem ko so nevronske mreže zaradi nelinearnosti in kompleksnosti modelov namenjene izključno področnim strokovnjakom (TechDifferences, brez datuma).

Naslednji tehniki sta tehniki ocenjevanja in napovedovanja, ki na podlagi neodvisne spremenljivke, ki bi lahko imele vpliv na ciljno spremenljivko, ocenjujeta približno vrednost številske ciljne spremenljivke. Enostaven primer so ocene učencev v osnovni in srednji šoli, kjer lahko uporabimo regresijsko analizo, ki bo izračunala oceno, ki jo bo imel dijak v srednji šoli na podlagi ocen, ki jih je dosegel v osnovni šoli (Larose & Larose, 2014).

Tehnika klasifikacije ima cilj natančno predvideti rezultat za vsako spremenljivko v podatkih. Tehnika klasifikacije se začne z naborom podatkov, ki so že razvrščeni v podzvrsti. Na primer, tehnika klasifikacije, ki predvideva kreditno tveganje, bi oceno lahko razvila na podlagi dalj časa opazovanih podatkov poslovnih partnerjev. Poleg zgodovinske bonitetne ocene lahko podatki prikažejo zgodovino zaposlovanja, lastništvo stanovanj ali najemnino, leto zaposlitve ali število in vrsto opravljenih naložb (Witten, Frank, Hall & Pal, 2016).

Tehnika grozdenja ureja podatke po skupnih karakteristikah. Grozdenje je zaradi svoje enostavne uporabnosti in praktičnosti zelo priljubljena tehnika. Tehnika grozdenja se najbolj pogosto uporablja pri segmentaciji trga po kupcih podobnih lastnosti za namen lansiranja proizvoda v prodajo. Rezultat tehnike grozdenja je skupina podatkov, ki vsebujejo podobne vrednosti (Klepac & Mršić, 2006).

Tehnika asociacijskih pravil ima cilj zaznati vse povezave med dvema oziroma večjim številom podatkov. Pogosto je uporabljena v poslovnem svetu, kjer jo imenujejo tudi analiza podobnosti oziroma analiza potrošnikove košarice (Larose & Larose, 2014).

Drevo odločanja je grafični prikaz rešitve, ki temelji na številnih pogojih, ki so podani. Drevo odločanja služi odločanju na podlagi različnih podatkov. Lahko se uporablja za preprosto napoved o kakovosti hrane v določeni restavraciji ali za kompleksnejše analize, kot je vpliv alkohola na mladostno prebivalstvo. Drevesa odločanja ponujajo drugačen nabor orodij za premagovanje ovir in privedejo do konkretnega in smiselnega rezultata (Kozak, 2019).

2.2 Analitična orodja

Orodja za analitiko masovnih podatkov omogočajo reševanje problemov klasifikacije, grozdenja, asociativnih pravil in ostalih tehnik analitike masovnih podatkov. Obstajajo orodja, ki so brezplačna in komercialna.

Prvo orodje, ki smo ga opisali, je orodje Apache Hadoop. Ta je odprtokodni projekt, ki predstavlja skupek orodij za shranjevanje in vzporedno obdelavo masovnih podatkov, shranjenih v računalniških gručah. Gradniki Hadoop Apache so (Ferle, 2013):

- tehnologija shranjevanja podatkov (angl. Hadoop Distributed File System), ki predstavlja implementacijo po principu distribuiranega datotečnega sistema;
- sistem, , ki dodeljuje, nadzira, upravlja računalniške vire;
- uporabniške aplikacije vozlišč v gruči (angl. Yet Another Resource Negotiator);
- MapReduce, ki predstavlja programski model in implementacijo programskega ogrodja za obdelavo podatkov.

MongoDB je sistem, ki predstavlja odprtokodno ne relacijsko podatkovno bazo usmerjeno v dokument. Sistem podatke shranjuje v obliki JSON dokumentov, slednji pa so lahko različnih oblik. Čeprav so podatki zapisani v C, C++ in JavaScriptu lahko MongoDB zaradi platforme neodvisnosti integriramo s številnimi programskimi jeziki, kar je pripeljalo do tega, da je prav za MongoDB razvito kar nekaj orodij, ki funkcionalnost sistema še nadgradi (Taylor, 2022a).

Talend je orodje, ki pod svojim okriljem skriva več drugih orodij, ki omogočajo integracijo masovnih podatkov, njihovo obvladovanje, zagotavljanje kakovosti in podporo sorodnim storitvam. Talend (Talend.com, brez datuma) je odprtokodno orodje, ki uporabnikom daje popolno svobodo z namenom kar se da izboljšane individualne prilagodljivosti. V novejših izdajah programske opreme najdemo vedno več komponent, ki povezujejo in omogočajo uporabo različnih virov in storitev.

Apache Cassandra je odprtokoden, distribuiran in decentraliziran sistem za shranjevanje in upravljanje masovnih podatkov, ki so v danem trenutku lahko shranjeni po celem svetu. Napisan je v programskem jeziku Java in je hibrid podatkovnih shem, ki iz ostalih modelov vzame najboljše lastnosti in jih združuje v super model (Slapnik, 2012).

Apache Stark je orodje, ki je namenjeno distribuiranemu obdelovanju masovnih podatkov. Hitrost, zmožnost, enostavnost uporabe in povezljivost so ključne lastnosti tega orodja. Visoko hitrost programu omogoča shranjevanje med rezultatov v spominu (angl. In-Memory saving), kar je denimo veliko hitreje kot MapReduce, omenjen pri orodju Hadoop, ki zapisuje in bere z diskov. Velik obseg je možen, ker Spark ima 4 shrambe, ki so hkrati funkcionalni deli Sparka. To so Spark Sqlm, ki strukturira podatke, Spark Streaming, ki obdeluje tok podatkov, Spark MLlib, namenjen strojnemu učenju, in GraphX, ki obdeluje podatke, shranjene v obliki grafov. Spark platforma omogoča uporabo jezikov Scala, Java, Python in R. Zaradi same strukture podatkov je platforma zelo enostavna za uporabo. Spark je platforma za obdelavo podatkov in ne njihovo shranjevanje, zato spark, funkcionira v simbiozi z drugimi tehnologijami, kot so Apache Hadoop, Cassandra, MongoDB in drugi (Nevajdić, 2021).

Power BI, je zbirka med seboj povezanih orodij, ki oblikujejo celoten sistem, namenjen analitiki masovnih podatkov. Program se je sprva uporabljal na področju vizualizacije podatkov, nato pa še na področju ETL procesov in interaktivnih funkcij. Program pri obdelovanju različnih tipov podatkov ponuja vizualni vpogled in možnost »rezanja«, hkrati pa samodejno povezuje podatke med seboj (How, 2020).

Orodje Tableau ima nalogo prikazati podatke v obliki, ki bo uporabniku najbolj razumljiva. Orodje je sposobno zelo hitrega samodejnega analiziranja podatkov. Tableau po navadi rezultate prikaže v obliki nadzorne plošče ali delovnih listov. Tableau od uporabnika ne zahteva nikakršnega strokovnega znanja, niti ni uporabno zahteven program, kar se orodju šteje v prid. Pri uporabi orodja je pomembna uporabnikova sposobnost, da podatke in rezultate, ciljnemu avditoriju, predstavi na način, da bodo nudili odgovore na zastavljena vprašanja (Taylor, 2022b).

2.3 OLAP in podatkovni model

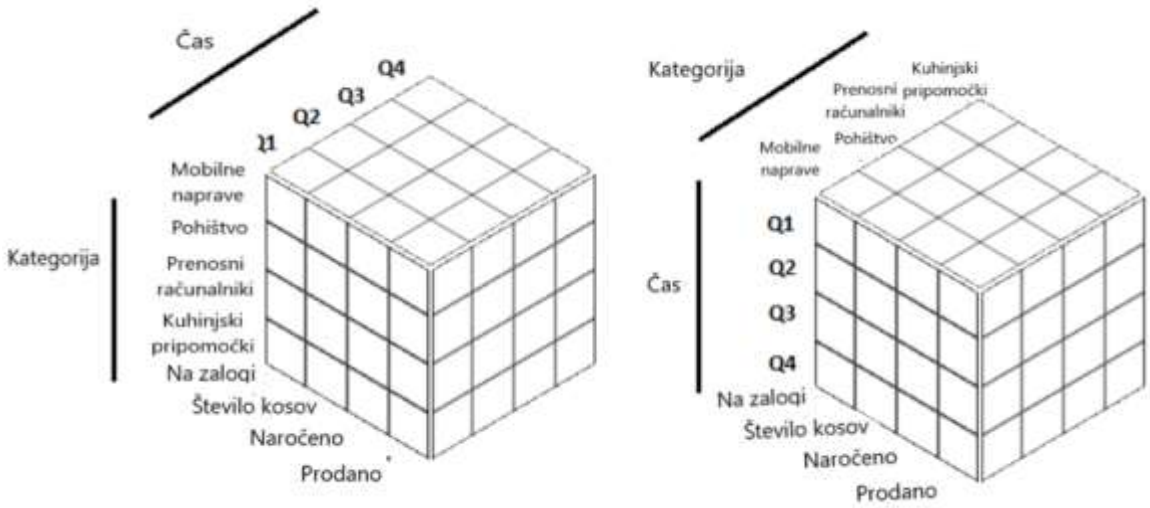
Transakcijske baze podatkov so zaradi relacijskega modela zelo učinkovite pri obdelavi podatkov, toda, zaradi velikega števila relacij, podatkovna skladišča postajajo prezapletena za učinkovito izvajanje analiz in pregledovanja slednjih. To je priznal tudi sam začetnik relacijskih podatkovnih skladišč Edgar F. Codd, kar ga je napeljalo, da prispeva k razvoju boljše tehnološko analitične rešitve. Leta 1993 je tako nastalo delo z imenom »Providing On-Line Analytical Processing to User Analysts«, v katerem se je prvič omenil izraz OLAP. Po definiciji OLAP predstavlja kategorijo programske tehnologije, ki analitikom, menedžerjem in vodstvu organizacij omogoča vpogled v podatke s hitrim, doslednim in interaktivnim pristopom. Informacije, pridobljene iz neobdelanih podatkov, se prikazujejo in odražajo v širokem razponu možnega prikaza na način, ki si ga je uporabnik zamislil (Altaplana, brez datuma).

OLAP uporabnikom ponuja hitro izvedbo poizvedb, česar je zmožen zaradi specializiranega indeksiranja in struktur skladišč. Če podatki niso pridobljeni preko ETL tehnike, se odsvetuje neposreden prenos iz virov podatkov v OLAP dimenzijske kocke, saj jih OLAP zaradi neusklajenosti ne more pravilno obdelati in daje napačne rezultate, kar pripelje do napačnih odločitev podjetja. OLAP sistem ni smiselno uporabljati, če nimamo obsega zgodovinskih podatkov, kajti potem jih ne moremo primerjati z realnim transakcijskimi podatki in posledično ne moremo ugotavljati trendov, niti boljše razumeti poslovanja organizacije. OLAP brez podatkovnih skladišč, kjer so podatki shranjeni, točni, popolni, pravočasni in konsistentni, ne obstaja oziroma nima kakovostnih rezultatov (Kimball & Caserta, 2004).

Večdimenzionalni model OLAP je sistem, ki nam omogoča pogled na podatke s strani, kot si želimo. V osnovi obstajajo tri operacije, ki jih uporabljamo pri uporabi OLAP kocke. Prva funkcija je vrtenja podatkov (angl. Pivoting), prikazana na sliki 6, ki omogoča, da se OLAP

kocka obrača okoli svoje osi in uporabniki vidijo podatke iz različnih perspektiv. Primer je recimo, da so dimenzije, ki so prej bile v vrsticah sedaj prikazane v stolpih ali obratno.

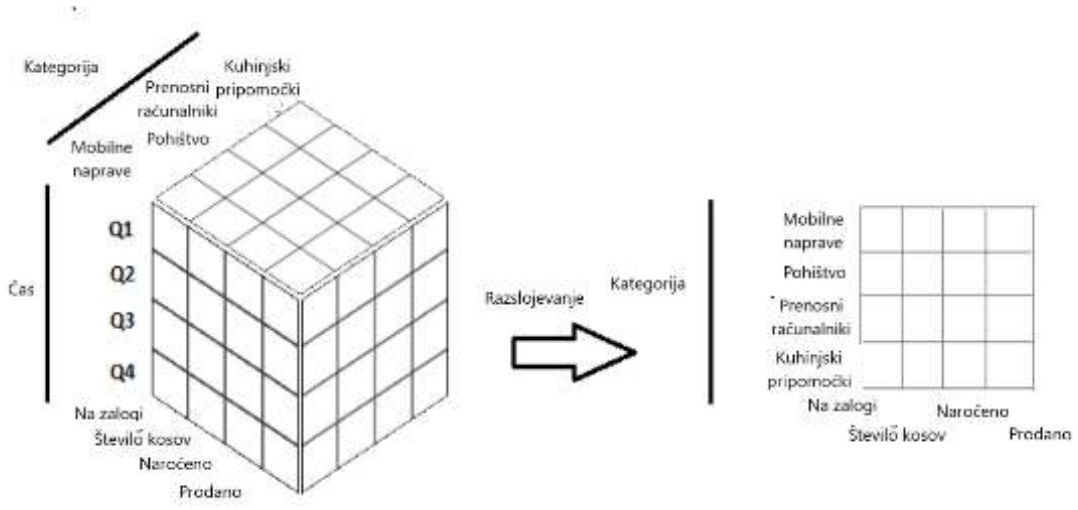
Slika 6: Vrtenje podatkov



Vir: Nitesh (2017).

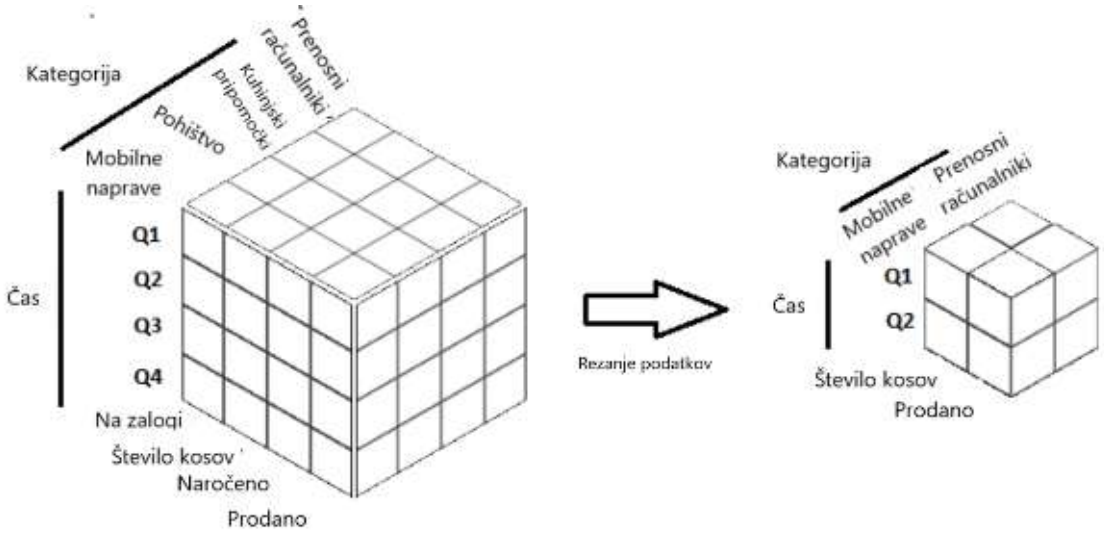
Razslojevanje in rezanje podatkov (angl. Slice and Dice) sta funkciji, prikazani na sliki 7 in sliki 8, ki omogočata, da OLAP kocko razslojujemo vodoravno oziroma navpično ali pa lahko vzamemo del glavne kocke kot pod kocko.

Slika 7: Razslojevanje podatkov



Vir: Nitesh (2017).

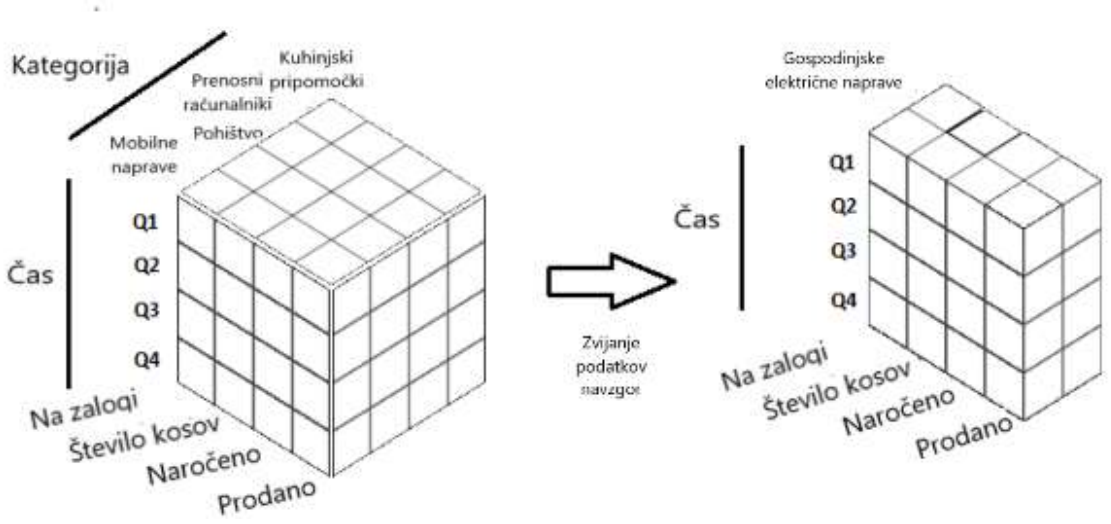
Slika 8: Rezanje podatkov



Vir: Nitesh (2017).

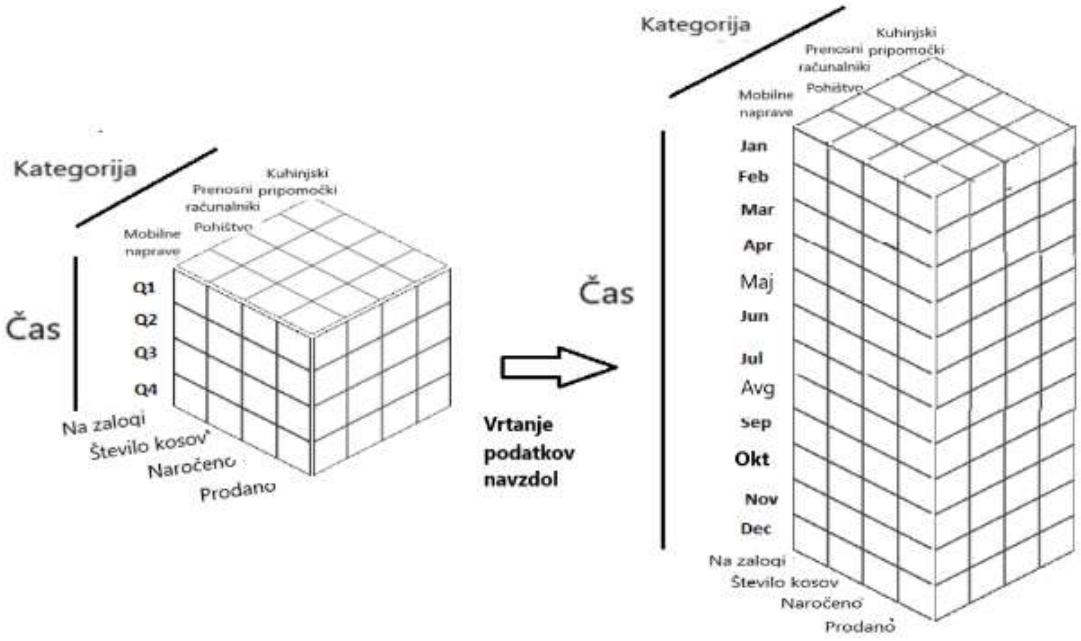
Podroben pogled na podatke oziroma vrtanje in zvijanje (angl. Drill up and Drill down) je funkcija, ki omogoča podrobnejši pogled po hierarhiji navzgor ali po hierarhiji navzdol. Na sliki 9 so mobilni telefoni in prenosniki po hierarhiji navzgor postali elektronske naprave, medtem ko so na sliki 10 po hierarhiji navzdol kvartali postali meseci.

Slika 9: Zvijanje podatkov navzgor



Vir: Nitesh (2017).

Slika 10: Vrtanje podatkov navzdol



Vir: Nitesh (2017).

2.4 Možnosti uporabe analitike masovnih podatkov v podatkovnih skladiščih

Podjetja imajo v podatkovnih skladiščih 90 % zastarelih in samo 10 % operativnih podatkov. Z uvedbo tehnologije analitike masovnih podatkov bi podjetje bilo sposobno uporabiti tudi zastarele podatke. Uvajanju analitike masovnih podatkov v podatkovna skladišča organizacij pogosto rečemo uvajanje poslovne inteligence, ki bo izboljšala poslovanje podjetja, saj bo podpora, ki jo bo prinesla sama uvedba, izvršnim posameznikom veliko pomagala pri odločitvah. V praksi to pomeni, da se bodo vsi podatki, shranjeni v podatkovnih skladiščih, obdelovali v celoti, kar bo rezultiralo s koristnimi informacijami pomembnimi za odločitve in operativne aktivnosti podjetja (Panian & Klepac, 2003).

V podjetjih velja, da je delovna sila, najbolj pomemben gradnik poslovanja. Takoj za delovno silo najdemo gradnik v pogledu informacij, ki pripomorejo k boljšemu poslovanju podjetja. Torej, ko lahko podjetje sprejema odločitve na podlagi pravočasnih in natančnih informacij, lahko podjetje izboljša svojo uspešnost. Analitika masovnih podatkov v lastnih podatkovnih skladiščih pravzaprav pospeši odločanje, hkrati pa izboljša uporabniško izkušnjo z zagotavljanjem pravočasnega in ustreznega odziva na težave s strankami (Ranjan, 2005).

Prva možnost uporabe orodij za analitiko masovnih podatkov v podatkovnih skladiščih je proces OLAP, ki bi se zaradi svoje večdimenzionalnosti in sposobnosti lahko uporabil v prav vseh segmentih poslovanja, najbolj primeren pa bi bil za napovedovanje in anticipacijo procesov, ki bi se utegnili zgoditi v prihodnosti. Na primer OLAP orodje bi bilo zmožno, na

podlagi podatkov v podatkovnih skladiščih, napovedati prodajo po določenih proizvodih ali po geografskih področjih za neko določeno vremensko obdobje.

Druga možnost uporabe analitike masovnih podatkov je rudarjenje podatkov (angl. Data mining), ki je definirano kot zajemanje podatkov iz ogromnih zbirk podatkov. Rudarjenje je postopek črpanja znanja iz podatkov. Ta možnost je namenjena predvsem najbolj naprednim uporabnikom, ki redno izvajajo korelacijo in analizo trendov, hkrati pa izdelujejo specifične projekcije trendov in dogodkov v prihodnosti. Pristop rudarjenja podatkov se lahko uporablja v namen analize trgov, odkrivanje goljufij, zadrževanje uporabnikov, nadzor proizvodnje ali znanstvene raziskave (Panian in drugi, 2007).

Tretja možnost so poslovna poročila (angl. Enterprise reporting), ki so namenjena operativnemu poslovnemu poročanju menedžerjem iz različnih področij podjetja. Poročila so lahko namenjena tudi strankam podjetja. Pri pripravi poslovnih poročil se lahko podatki črpajo neposredno iz podatkovnega skladišča podjetja, kar poslovna poročila klasificira kot najbolj uporabljen pristop analitike masovnih podatkov (Panian in drugi, 2007).

Zadnji pristop, ki smo ga opisali, so Ad Hoc poizvedbe, ki omogočajo analizo poslovnih podatkov do nivoja operativnih transakcijskih podrobnosti. Primarni cilj Ad Hoc pristopa je dati uporabnikom možnost sestave popolnoma novih poročil, brez kakršnihkoli omejitev v poročilu. Ad Hoc poizvedbe uporabnikom omogočajo dostop do katerega koli podatka v podatkovnem skladišču z uporabo OLAP orodja, hkrati se lahko podatki filtrirajo, tako da se dobijo čim bolj točni podatki. Podatkom se lahko tudi filtrirajo lastnosti, brez da bi se v podatkovnem skladišču karkoli porušilo (David, 2021).

3 UVEDBA PODATKOVNEGA SKLADIŠČA V IZBRANEM PODJETJU

Poglavje opisuje uvedbo podatkovnega skladišča v podjetje, ki se primarno ukvarja z računovodstvom, knjigovodstvom, revizijsko dejavnostjo in davčnim svetovanjem. Podjetje je bilo ustanovljeno leta 2004 in od takrat dalje, dejavnost opravlja na sedežu v Ljubljani. Podjetje po velikosti uvrščamo v mikropodjetja. Prednosti, slabosti, priložnosti in nevarnosti uvedbe podatkovne skladišča sem predstavil s PSPN analizo v zadnjem delu tega poglavja.

3.1 Uvedba podatkovnega skladišča

Podjetje za opravljanje računovodskih storitev uporablja programsko rešitev VASCO, ki omogoča izvoz podatkov v formatu besedilne datoteke, ki vsebuje z vejico ločene vrednosti (angl. Comma-Separated Values, v nadaljevanju CSV). CSV je oblika besedilne datoteke, ki je zasnovana za prikaz tabelarnih podatkov. Uvoz podatkov v podatkovna skladišča v CSV obliki omogoča Microsoft SQL Server Express 2019 s pomočjo SQL Server

Management Studio programske rešitve, zato sem v njej tudi pripravil izgradnjo podatkovnega skladišče, za podjetje.

Microsoft SQL Server 2019 Express je zmogljiv, zanesljiv in brezplačen sistem za upravljanje podatkov, ki zagotavlja hitro in zanesljivo shranjevanje podatkov za lahka spletna mesta in namizne aplikacije (Microsoft, brez datuma a).

Po uvozu podatkov v SQL Server Express 2019 v izvorni obliki, se lahko začne postopek njihovega preoblikovanja. V postopku preoblikovanja sem bil pozoren predvsem na napake, ki so nastale ob uvozu. Na podlagi ugotovljenih napak smo v strukturiranem povpraševalnem jeziku za delo s podatkovnimi bazami (angl. Structured Querry Langugage, v nadaljevanju SQL) določili navodila, tako da do istih napak ne bo prišlo pri naslednjem uvažanju. Ko pregledamo, določimo pravila in preoblikujemo podatke, lahko začnemo s postopkom nalaganja, s čimer se zaključi ETL proces.

Po končanem spoznavnem delu z izvornimi podatki je bil izdelan podatkovni model podatkovnega skladišča v zvezdini shemi, z eno tabelo dejstev in nekaj dimenzijskimi tabelami, ki se vežejo na tabelo dejstev. Pred samo izgradnjo podatkovnega skladišča je potrebno izdelati primerno obliko logičnega modela. Za ta namen je bil uporabljen diagram entitet povezav (angl. Entity Relationship Diagram, v nadaljevanju ER). V ER diagramu se oblikujejo medsebojne povezave in atributi modela, ki predstavlja strukturo, na podlagi katere se bo izgradilo podatkovno skladišče.

Vsi postopki se izvajajo jeziku SQL. Ko so bile na podlagi ER diagrama oblikovane dimenzijske tabele in tabela dejstev, se je pričelo polnjenje podatkov v dimenzijske tabele, kar je pogoj, da se lahko pristopi k polnjenju podatkov v tabelo dejstev. V dimenzijske tabele se shranijo edinstveni identifikatorji podatkov, ki pripomorejo k povezovanju tabele dejstev s pravimi podatki. Pri vpisovanju podatkov v dimenzijske tabele je za hierarhijo kategorij treba vsak zapis razdeliti na 3 dele, čemur sledi zapis podatkov v pripadajočo dimenzijsko tabelo.

Za izdelavo poročil in vizualni prikaz podatkov iz uvoženih podatkov v skladišče podatkov je bila uporabljena platforma Microsoft Power Bl. Po zagonu aplikacije je treba izbrati želeni vir podatkov, ki ga želimo analizirati ter pripraviti poročilo. V primeru izbranega podjetja so se podatki uvozili iz sistema SQL Server. Uvoženi podatki predstavljajo osnovo za pripravo poročila in različne vizualizacije,

3.2 Značilnosti in tehnološke zahteve uvedbe podatkovnega skladišča

V tabeli 1 so opisane tehnološke zahteve za namestitev programske opreme, potrebne za uvedbo podatkovnega skladišča

Tabela 1: Tehnološke zahteve za namestitev programske opreme, potrebne za uvedbo podatkovnega skladišča

Programska oprema	Microsoft® SQL Server® 2019 Express	SQL Server Management Studio	Power BI Report Server
Operacijski sistem	Windows 10, Windows Server 2016, Windows Server 2019	Windows Server 2022 (64-bit), Windows 11 (64-bit) Windows 10 (64-bit) Windows 8.1 (64-bit) Windows Server 2019 (64-bit) Windows Server 2016 (64-bit) Windows Server 2012 R2 (64-bit) Windows Server 2012 (64-bit) Windows Server 2008 R2 (64-bit)	Windows Server 2019 Datacenter Windows Server 2019 Standard Windows Server 2016 Datacenter Windows Server 2016 Standard Windows 10 Home Windows 10 Professional Windows 10 Enterprise Windows 11
Procesor (Hitrost)	Vsaj 1 GHz	Vsaj 1.8 GHz	Vsaj 1.4 GHz
RAM	Vsaj 512 MB	Vsaj 2 GB	Vsaj 1 GB
Trdi disk	Vsaj 4.2 GB prostega mesta	Vsaj 2 do 10 GB prostega mesta	Vsaj 1 GB

Vir: Microsoft (brez datuma b, c, d).

Za uvedbo podatkovnega skladišča je pomembno, da so osebe, ki podatkovno skladišče uvajajo, kompetentne. Za upravljanje podatkovnega skladišča je v prihodnosti predvidena pomoč strokovne osebe, saj zaradi potrebe po znanju uporabe SQL jezika, osebe v podjetju niso kompetentne za vzdrževanje podatkovnega skladišča. Poročila in vizalizacije na podlagi že napolnjenega podatkovnega skladišča bodo v podjetju tudi v prihodnje pripravljali s programom Power BI report.

3.3 SWOT analiza koristi uvedbe podatkovnega skladišča v izbrano podjetje

Z analizo operativnih in zgodovinskih podatkov, informacij pomembnih za prihodnost in vseh drugih relevantnih dejavnikov v povezavi s podjetjem, se lahko zelo dobro pripravimo na možno prihodnost oziroma možne izhode ali poslovne izzive. Eden izmed celovitih ocenjevanj podjetij je analiza SWOT (Pučko, 1996).

Prva prednost uvedbe podatkovnega skladišča v računovodsko podjetje je predvsem enostavnejši dostop do bistveno večje količine podatkov, kot je to bila praksa brez

podatkovnega skladišča. Poleg tega so vsi podatki natančni, razen, ko zaradi nenatančnih postavljenih pravil pride do nenatančnosti podatkov. Naslednja prednost je, da lahko uporabniki izdelavo in uporabo poročil prilagajajo svojim potrebam in jih neposredno uporabljajo pri vsakdanjem delu.

Slabost uvedbe podatkovnega skladišča je, da je lahko kontrola natančnosti podatkov zelo dolgotrajen in drag postopek, saj je potrebno veliko analitičnega znanja o poslovnih procesih. Priprava poročil je naslednja slabost, ki je lahko zaradi kadrovske nekompetentnosti usodna, saj lahko terja veliko časa, ki ne bo obrestovan z rezultati in kakovostnimi poročili.

Prvo priložnost uvedbe podatkovnega skladišča vidim v tem, da bi podjetju uspelo na tedenski ravni pripravljati avtomatska poročila, ki bodo tako zaposlenim kot tudi strankam razumljiva in koristna. Naslednjo priložnost vidim v tem, da bi se na podlagi podatkov iz podatkovnih skladišč pripravljala poročila za stranke po njihovih merilih, kar jim sicer vzame veliko časa. Še eno priložnost vidim v hitrosti pridobivanja informacij, saj bi se ta takrat, ko bi izločili vse nepotrebne elemente iz poizvedb, povečala.

Nevarnost uvedbe podatkovnega skladišča je, da bodo poročila zaradi velike količine informacij, nejasna in da bo vzdrževanje takšnega podatkovnega skladišča postalo prezahtevno in predrago. Naslednja nevarnost je, da so lahko poročila, zaradi slabe in hitre nastavitve pravil, neuporabna. Zadnja nevarnost je, da bo večjo količino podatkov nemogoče predstaviti na jedrnat način. V tabeli 2 najdemo povzetek PSPN analize.

Tabela 2: SWOT analiza uvedbe podatkovnega skladišča

Prednosti	Slabosti
• Velika količina podatkov • Natančnost podatkov • Enostavna in razumljiva poročila	• Dolgotrajna kontrola nastavitev • Dolgotrajen postopek priprave poročil • Neuporabna poročila
Priložnosti	**Nevarnosti**
• Avtomatska, razumljiva poročila • Manjši stroški priprave poročil • Hitrejša priprava poročil	• Prezahtevna in predraga priprava poročil • Nerazumljiva poročila

Vir: lastno delo.

SKLEP

Podatkovna skladišča so sedanjost in prihodnost. Analitika masovnih podatkov v podatkovnih skladiščih je v veliko korist podjetjem, ki tehnologijo znajo uporabiti na pravi način in lahko pomaga tako organizacijam kot tudi vsesplošni družbi razumeti dogajanje na podlagi podatkov. Uporaba podatkovnega skladišča lahko veliko pripomore k uspešnosti podjetja, saj se kakovost tako operativnih kot strateških odločitev lahko bistveno poveča.

Uvedba podatkovnega skladišča ni enostavna. Podjetje potrebuje dobrega arhitekta in analitika, ki bo vzpostavil kakovostno in uporabno podatkovno skladišče. Uvedba podatkovnega skladišča v podjetje običajno zahteva visoko začetno investicijo, stroški izdelave in vzdrževanja podatkovnega skladišča pa so sicer sorazmerni z velikostjo skladišča in količino podatkov.

Podjetja morajo pred uvedbo podatkovnega skladišča dobro premisliti o namenu uvedbe, saj je v nasprotnem lahko podatkovno skladišče popolnoma brezpredmetno ali neuporabno, uvedba takšnega podatkovnega skladišča pa ima lahko usodne posledice za poslovanje podjetja.

Pri izdelavi in uvedbi skladišča v podjetje morajo sodelovati strokovnjaki s potrebnimi izkušnjami in znanjem. To je še posebej pomembno v podjetjih, kjer uvedba podatkovnega skladišča predstavlja pomemben mejnik v poslovanju oz. je ključna za nadaljnji razvoj in usmeritev podjetja skladno z njihovo strategijo.

LITERATURA IN VIRI

1. Aggarwal, C. C. (2015). *Data mining: the textbook.* New York, Springer International Publishing. Springer. doi: 10.1007/978-3-319-14142-8.
2. Altaplana. (brez datuma). *Olap and olap server definitions*. Dne 29. januarja 2022 pridobljeno iz http://www.altaplana.com/olap/glossary.html#PAGE%20DISPLAY
3. Berry, M. J. & Linoff, G. S. (2011). *Data mining techniques: for marketing, sales, and customer relationship management.* New York, USA: John Wiley & Sons.
4. Callegaro, Y., Yang, Y. (2017). *The Role of Surveys in the Era of "Big Data". Dne 23.03.2022 pridobljeno iz:* https://link.springer.com/chapter/10.1007/978-3-319-54395-6_23
5. Coronel, C., Morris, S. & Rob, P. (2011). *Database Systems: Design, Implementation and Management* (9. izd.). Boston, USA: Cengage Leaning.
6. David, M. (2021, 25. januar). *What is Ad Hoc Analysis and How Does it Work?*. Dne 29. januarja 2022 pridobljeno iz: https://chartio.com/learn/data-analytics/what-is-ad-hoc-analysis/
7. Dumbill, E. (2012). *Planning for Big Data*. Sebastopol, USA: O'Reilly Media

8. Faizaan, Y. (2015, 3. junij). *ETL (Extract, Transform, and Load) Process & Concept* [objava na blogu]. Dne 13. januarja 2022 pridobljeno iz https://blog.appliedinformaticsinc.com/etl-extract-transform-and-load-process-concept/

9. Fawcett, T., Provost, F. (2013). *Data Science and Its Relationship to Big Data and Data-Driven Decision Making*. Dne 23.03.2022 pridobljeno iz: https://www.researchgate.net/publication/256439081_Data_Science_and_Its_Relationship_to_Big_Data_and_Data-Driven_Decision_Making

10. Ferle, M. (2013). Hadoop in MapReduce. *Monitor Pro*. Dne 27. januarja 2022 pridobljeno iz: http://www.monitorpro.si/156498/praksa/hadoop-in-mapreduce/

11. Golfarelli, M. & Rizzi, S. (2009). *Data warehouse design. Modern principles and Methodologies*. Bologna: Seps.

12. Han, J., Kamber, M. & Pei, J. (2012). *Data Mining: Concepts and Techniques*. San Fransisco: Morgan Kaufmann Publishers.

13. How, M. (2020). *The modern data warehouse in Azure: building with speed and agility on Microsoft's Cloud Platform*. Berkeley, CA, USA: Apress.

14. Inmon, W. H. (2002). *Building the Data Warehouse* (3. izd.). New York: John Willey & Sons, Inc

15. Inmon, W. H. (2005). *Building the data warehouse* (4. izd.). New York: Wiley.

16. Jindal, R. & Acharya, A. (2019, november). *Federated Data Warehouse Architecture*. Dne 8. februarja 2022 pridobljeno iz: https://idoc.pub/documents/federated-data-warehouse-architecture-134w9djjgzn7

17. Kimball, R. & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*. Indianapolis: Wiley Publishing, Inc.

18. Kimball, R. & Ross, M. (2002). *The Data Warehous Toolkit*. Indianapolis: Wiley Publishing, Inc.

19. Kimball, R., Ross, M., Thornthwaite, W., Mundy, J. & Becker, B. (1998). *The Data Warehouse Lifecycke Toolkit*. Indianapolis, USA: John Wiley & Sons, Inc.

20. Klepac, G. & Mršić, L. (2006). *Poslovna inteligencija kroz poslovne slučajeve*. Zagreb: Lider press ; TIM press

21. Koren, R. (2010). *Polnjenje podatkovnih skladišč s pomočjo orodja oracle warehouse builder* (diplomsko delo). Ljubljana: Fakulteta za matematiko in fiziko.

22. Kovačič, A. (2004). *Prenova in informatizacija poslovanja*. Ljubljana: Ekonomska fakulteta.

23. Kozak, J. (2019). *Decision Tree and Ensemble Learning Based on Ant Colony Optimization*. Katowice: Springer International Publishing.

24. Larose, D. T. & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. London: John Wiley & Sons.

25. Microsoft. (brez datuma a). *Microsoft® SQL Server® 2019 Express*. Dne 6. februarja 2022 pridobljeno iz: https://www.microsoft.com/en-us/Download/details.aspx?id=101064

26. Microsoft. (brez datuma b). *tehnološki pogoji za inštalacijo Microsoft® SQL Server® 2019 Express*. Pridobljeno 6. februarja 2022 iz https://www.microsoft.com/en-us/Download/details.aspx?id=101064

27. Microsoft. (brez datuma c). *tehnološki pogoji za inštalacijo Power BI Report Server*. Dne 6. februarja 2022 pridobljeno iz: https://docs.microsoft.com/en-us/power-bi/report-server/system-requirements

28. Microsoft. (brez datuma d). *tehnološki pogoji za inštalacijo SQL Server Management Studio (SSMS)*. Dne 6. februarja 2022 pridobljeno iz: https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15

29. Naeem, T. (2020, 3. februar). *Data Warehouse Concepts: Kimball vs. Inmon Approach* [objava na blogu]. Dne 6. februarja 2022 pridobljeno iz: https://www.astera.com/type/blog/data-warehouse-concepts/

30. Nevajdić, V. (2021). *Analiza velikih količina podataka korištenjem Apache Spark platforme* (magistrsko delo). Zagreb: Prirodoslovno-matematički fakultet.

31. Nitesh. (2017, 26. julij). *OLAP Operations Tutorial*. Dne 29. januarja 2022 pridobljeno iz: https://cracklogic.com/olap-tutorial/#Slice

32. Olenik, B. (2019). *Analiza tehnik poslovne analize za področje poslovne inteligence* (diplomsko delo). Ljubljana: Fakulteta za računalništvo in informatiko.

33. Panian Ž. in drugi (2007). *Poslovna inteligenca: studije slučajeva iz hrvatske prakse*, Narodne novine.

34. Panian, Ž. & Klepac, G. (2003). *Poslovna inteligencija*. Zagreb: Masmedia.

35. Pirc, D. (2007). Podatkovna skladišča v mednarodnem podjetju. Univerza v Ljubljani: Ekonomska fakulteta

36. Pučko, D. (1996). *Strateško upravljanje*. Ljubljana: Ekonomska fakulteta.

37. Rainardi, V. (2008). *Building a Data Warehouse*. New York: Apress.

38. Rangarajan, S. (2016, 1. september). *Data Warehouse Design - Inmon versus Kimball*. Dne 12. januarja 2022 pridobljeno iz: https://tdan.com/data-warehouse-design-inmon-versus-kimball/20300

39. Ranjan, J. (2005). Business intelligence: Concepts, components, techniques and benefits. *Journal of theoretical and applied information technology, 9*(1), 60-70.

40. Seiner R. (2007, 29. maj). Four Ways to Build a Data Warehouse, Dne 12. januarja 2022 pridobljeno iz: http://tdan.com/four-ways-to-build-a-data-warehouse/4770

41. Slapnik, A. (2012). *Uporaba objektno-relacijskega preslikovanja nad Apache Cassandra podatkovno bazo* (diplomsko delo). Ljubljana: Fakulteta za računalništvo in informatiko.

42. Stepinac, L. (2014, 12. maj). *Što je to zapravo Big Data i gdje se primjenjuje.* Dne 23. 01. 2022 pridobljeno iz: https://www.ictbusiness.info/poslovna-rjesenja/sto-je-to-zapravo-bigdata-i-gdje-se-primjenjuje

43. Stihović, V. (2015). *Alati za analitiku velike količine podataka (big data)* (diplomsko delo). Pula: Univerza v Puli, Oddelek za ekonomijo in turizem.

44. Stoilkovič, M. (2009). *Arhitektura informacijskega sistema za upravljanje Solventnosti II v zavarovalnicah* (magistrsko delo). Maribor: Ekonomsko-Poslovna fakulteta.

45. Talend.com. (brez datuma). *Why Talend?*. Dne 29. januarja 2022 pridobljeno iz https://www.talend.com/why-talend/

46. Taylor, D. (2022a, 12. februar). *What is Tableau? Uses of Tableau Software Tool*. Dne 29. januarja 2022 pridobljeno iz: https://www.guru99.com/what-is-tableau.html

47. Taylor, D. (2022b, 5. marec). *What is MongoDB? Introduction, Architecture, Features & Example?*. Dne 5. marca 2022 pridobljeno iz: https://www.guru99.com/what-is-mongodb.html

48. Tech Differences. (brez datuma). *Difference Between Descriptive and Predictive Data Mining*". Dne 27. januarja 2022 pridobljeno iz: https://techdifferences.com/difference-between-descriptive-andpredictive-data-mining.html

49. Vassiliadis, P. & Simitsis, A. (2009, januar). *Extraction, transformation, and loading. Research Gate.* Dne 13. januarja 2022 pridobljeno iz: https://www.researchgate.net/publication/267421465_Extraction_transformation_and_loading

50. Vasudev, M. (2015, 21. februar). *What is Bad Data and its SIde-Effects*. Dne 11. januarja 2022 pridobljeno iz https://www.business2community.com/big-data/bad-data-side-effects-01164045

51. Vujnovac, E. (2015). *Utjecaj velike količine podataka na znanstvenonastavni rad* (diplomsko delo). Univerza Josipa Juraja v Osijeku: Filozofska fakulteta.

52. Witten, I. H., Frank, E., Hall, M. A. & Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4. izd.). San Francisco: Morgan Kaufmann Publishers.