UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

INTEGRACIJA PODATKOVNIH VIROV V POSLOVNOINTELIGENČNO OKOLJE ZA ANALIZO DIGITALNEGA TRŽENJA

Ljubljana, april 2025 UROŠ PREVORČNIK

IZJAVA O AVTORSTVU

Podpisani Uroš Prevorčnik, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Integracija podatkovnih virov v poslovnointeligenčno okolje za analizo digitalnega trženja, pripravljenega v sodelovanju s svetovalcem red. prof. dr. Jurijem Jakličem

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.

11. da sem preveril verodostojnost informacij, ki izhajajo iz zapisov na podlagi uporabe orodij umetne inteligence.

V Ljubljani, dne 6.5.2025

Podpis študenta: Uroš Prevorčnik Digitalno podpisal Uroš Prevorčnik Datum: 2025.05.06 21:21:20 +02'00'

KAZALO

KAZALO TABEL

KAZALO SLIK

KAZALO PRILOG

SEZNAM KRATIC

angl. – angleško
AIMQ – (angl. Methodology for Information Quality Assessment); Metodologija za zagotavljanje kakovosti informacij
API – (angl. Application Programming Interface); Programski vmesnik

BI – (angl. Business Intelligence); Poslovna inteligenca
CRM – (angl. Customer Relationship Management); Management odnosov s strankami
CSV – (angl. Comma-separated Values); Običajni format za besedilno datoteko, ki vsebuje vrednosti ločene z vejico
DDM – (angl. Dependent Data Marts); Odvisna področna podatkovna skladišča
DM – (angl. Data Mart); Področno podatkovno skladišče
DQA – (angl. Data Quality Assessment); Določilo kakovosti podatkov
DQAF – (angl. Data Quality Assessment Framework); Okvir za oceno kakovosti podatkov
DWH – (angl. Data Warehouse); Podatkovno skladišče
DWQ – (angl. Datawarehouse Quality Methodology); Metodologija za zagotavljanje kakovosti podatkovnega skladišča
EDW – (angl. Enterprise Data Warehouse); Organizacijsko podatkovno skladišče
ELT – (angl. Extract - Load - Transform); Pridobivanje-nalaganje-preoblikovanje
ETL – (angl. Extract - Transform - Load); Pridobivanje-preoblikovanje-nalaganje
GA4 – (angl. Google Analytics 4); Platforma Google Analytics 4
GAU – (angl. Google Analytics Universal); Platforma Google Analytics Universal
HDQM – (angl. Heterogenous Data Quality Methodology); Metodologija za zagotavljanje kakovosti heterogenih podatkov
HTML – (angl. Hypertext Markup Language); Označevalni jezik za oblikovanje večpredstavnostnih dokumentov
IDM – (anlg. Independent Data Marts); Neodvisna področna podatkovna skladišča
IP – (angl. Information Production); Proizvodnja informacij
IP – (angl. Iternet Protocol); Internetni protokol
IQ – (angl. Information Quality); Kakovost informacij
JSON – (angl. JavaScript Object Notation); Standardni format za pomnjenje in izmenjavo podatkov v tekstovni obliki
KPI – (angl. Key Performance Indicators); Ključni kazalniki uspeha
OLAP – (angl. Online Analytical Processing); Sprotne analitične obdelave podatkov
PDF – (angl. Portable Document Format); Format datoteke, ki je neodvisen od računalniškega okolja in medoperacijsko prenosljiv
ROI – (angl. Return on Investment); Donosnost naložbe
SGML – (angl. Standardized Generalized Markup Language); Standardizirani splošni označevalni jezik
SSMS – (angl. SQL Server Management Studio); Integrirano okolje za delo s SQL-om
SQL – (angl. Structured Query Language); Strukturiran povpraševalni jezik za delo s podatkovnimi bazami
TDQM – (angl. Total Data Quality Management); Celovito upravljanje kakovosti podatkov
TIQM – (angl. Total Information Quality Management); Celovito upravljanje kakovosti informacij
TQM – (angl. Total Quality Management); Celovito upravljanje kakovosti
UML – (angl. Unified Modeling Language); Enovit modelirni jezik

1 UVOD

Digitalno trženje za zadovoljevanje raznovrstnih potreb strank uporablja različne trženjske kanale, od iskalnikov (angl. search engines), e-pošte, blogov, družbenih omrežij, do spletnih mest izdelkov. Na ta način ustvarja učinkovit oglaševalni in komunikacijski ekosistem. Da se vzpostavi jasna trženjska struktura, je potrebnega veliko generiranja, analiziranja in recikliranja podatkov. Ročna obdelava te množice kompleksnih informacij največkrat ne omogoča njihovega popolnega razumevanja, zato privede do nepravilnih odločitev in posledičnih izgub. Da bi se temu izognili in ohranili urejen delovni tok, se v rutine digitalnega trženja vključijo različna orodja za obdelavo podatkov, bolj znana kot poslovna inteligenca (angl. Business Intelligence, v nadaljevanju BI) (Bhosale in drugi, 2020).

Zadostna količina informacij je postala ključni dejavnik uspeha na vseh področjih človeške dejavnosti. Integrirane informacije, pridobljene iz različnih virov in prilagojene določeni stopnji podrobnosti, so nepogrešljiva podpora v procesu odločanja. Pri tem se lahko količina in struktura virov časovno spreminjata, enako kot stopnja podrobnosti. Pomembno je zagotoviti najkakovostnejše informacije, tj. resnične in celovite, ter jih dostaviti končnemu uporabniku ob pravem času. Za izpolnitev teh zahtev so bile v podjetjih in institucijah uvedene rešitve za poslovno inteligenco (Tvrdíková, 2007).

Dandanes je pomen doseganja in vzdrževanja visokega standarda kakovosti podatkov splošno priznan tako med praktiki kot raziskovalci. Kakovostni podatki se, na račun svojega vpliva na poslovanje, pogosto obravnavajo kot dragoceno sredstvo (Cichy in Stefan, 2019). Kakovost lahko po mnenju Askham in drugih (2013), Batini in drugi (2009) ter Scarisbrick-Hauser in Rouse (2007) ocenimo s pomočjo različnih dimenzij: celovitosti, konsistentnosti, edinstvenosti, veljavnosti, točnosti in pravočasnosti. Pipino in drugi (2002) dodajajo še dostopnost, primerno količino podatkov, verodostojnost in varnost.

Seveda pa uspešno podjetje poleg kakovostnih podatkov potrebuje tudi učinkovit način za njihovo upravljanje in analiziranje. Najpreprostejša rešitev je shraniti podatke na eno mesto, od koder se jih nato uporablja za nadaljnje analize in poročila. Rešitev je poznana pod imenom podatkovno skladišče (angl. Data Warehouse, v nadaljevanju DWH) (Tvrdíková, 2007). Skozi čas postajajo podatki, ki se uporabljajo pri odločanju, vse kompleksnejši, so heterogenih formatov in prihajajo iz razpršenih virov. V okolju z vse številčnejšimi izzivi za organizacije, je ključno imeti podatkovno skladišče, ki združuje heterogene vire (Boufares in Salem, 2012).

Podjetja so prisiljena upravljati z vse večjimi količinami podatkov. Za enostaven dostop se ti podatki iz različnih virov integrirajo z uporabo procesa pridobivanja - preoblikovanja - nalaganja (angl. Extract - Transform - Load, v nadaljevanju ETL) ali procesa pridobivanja - nalaganja - preoblikovanja (angl. Extract - Load - Transform, v nadaljevanju ELT). Po mnenju Sreemanthy in drugi (2020) je proces integracije podatkov najosnovnejši korak.

Integracija podatkov iz več zunanjih virov predstavlja izziv na številnih področjih. Na področju digitalnega trženja je ta problem še posebej izrazit, ker je večina virov zunanjih, kar pomeni, da razvijalec nanje nima vpliva. Podatki iz teh virov prihajajo v različnih oblikah in z različnimi stopnjami kakovosti. Za čim uspešnejšo analizo na področju digitalnega trženja torej potrebujemo proces, ki nam omogoči čim boljšo kakovost podatkov iz več podatkovnih virov pri njihovi integraciji.

Namen magistrskega dela je torej pomagati podjetjem in posameznikom doseči čim boljšo kakovost podatkov, ki se jih za namene analize trženja integrira v BI okolje iz več podatkovnih virov. Priporočila, navedena v delu, so uporabna tudi na drugih področjih, ne le v digitalnem trženju.

Cilji magistrskega dela so:

- pripraviti pregled obstoječe literature na temo poslovne inteligence in BI orodij na področju digitalnega trženja;
- pripraviti pregled načinov integriranja podatkov iz različnih tržnih kanalov in možnosti avtomatizacije procesa ETL;
- pripraviti pregled obstoječe literature o doseganju kakovosti podatkov;
- izvesti analizo procesa integracije podatkov iz več podatkovnih virov v BI okolje na konkretnem primeru;
- pripraviti priporočila za zagotavljanje kakovosti podatkov pri integraciji iz več podatkovnih virov za področje digitalnega trženja v izbranem podjetju, ki pa se bodo lahko uporabljala tudi v podobnih situacijah z več različnimi zunanjimi viri.

Magistrsko delo je razdeljeno na dva dela, teoretični in empirični del. V prvem delu sem analiziral literaturo o poslovni inteligenci na področju digitalnega trženja, o procesu integriranja podatkov iz več virov in procesu ETL, ki omogoča integracijo in zagotavljanje kakovosti podatkov. Kot literaturo sem v teoretičnem delu uporabil knjige, znanstvene članke, publikacije in internetne vire. Z analizo literature bralcu predstavim koncept poslovne inteligence in BI orodij, ter proces ETL in avtomatizacijo procesa, kjer bo poudarek na korakih implementacije (DWH in DWH v oblaku). Predstavim dimenzije za zagotavljanje kakovosti podatkov in okvirje (angl. framework). Cilj prvega dela je bralca seznaniti s samo tematiko in procesom.

V empiričnem delu sem, na podlagi analize primera, analiziral posamezne korake procesa implementacije in podal korake za zagotavljanje kakovosti podatkov pri integraciji iz več podatkovnih virov po teh korakih. Analiziral sem primer analitičnega poročila za spremljanje ključnih kazalnikov uspeha na področju digitalnega trženja iz več podatkovnih virov, z namenom ugotavljanja vpliva ustreznosti pristopa k integraciji zunanjih podatkov za zagotavljanje kakovostnih informacij. Na podlagi sinteze, ugotovitev pregleda literature in analize primera sem podal priporočila za zagotavljanje kakovosti podatkov pri integraciji iz več podatkovnih virov.

2 POSLOVNA INTELIGENCA IN DIGITALNO TRŽENJE

2.1 Poslovna inteligenca

Negash (2004) je v začetnih letih obstoja BI zapisal, da se poslovna inteligenca uporablja za razumevanje razpoložljivih sposobnosti v podjetju. Uporablja se za razumevanje trendov in za odločitve v zvezi z izbiro prihodnje usmeritve na trgih, izbiro tehnologij in regulativnega okolja, v katerem podjetje konkurira. Uporablja se tudi za analizo dejanj konkurentov in posledic teh dejanj.

Čeprav je izraz poslovna inteligenca relativno nov, so računalniški sistemi, namenjeni poslovni inteligenci, v takšni ali drugačni obliki obstajali že pred štiridesetimi leti. BI je zaobjela funkcionalnosti sistemov za podporo odločanju (angl. descision support), izvršilnih informacijskih sistemov (angl. executive information systems) in informacijskih sistemov za upravljanje (angl. management information systems) (Thomsen, 2003).

Bhosale in drugi (2020) opredelijo BI kot tehnologije in prakse, ki pomagajo pri akumulaciji podatkov, njihovi sintezi in njihovi analizi, da bi predstavili koristne poslovne informacije z namenom povečanja dobička in izogibanja izgubam. Osnovni cilj poslovne inteligence je avtomatizacija procesa odločanja in zmanjšanje možnosti napak pri presoji.

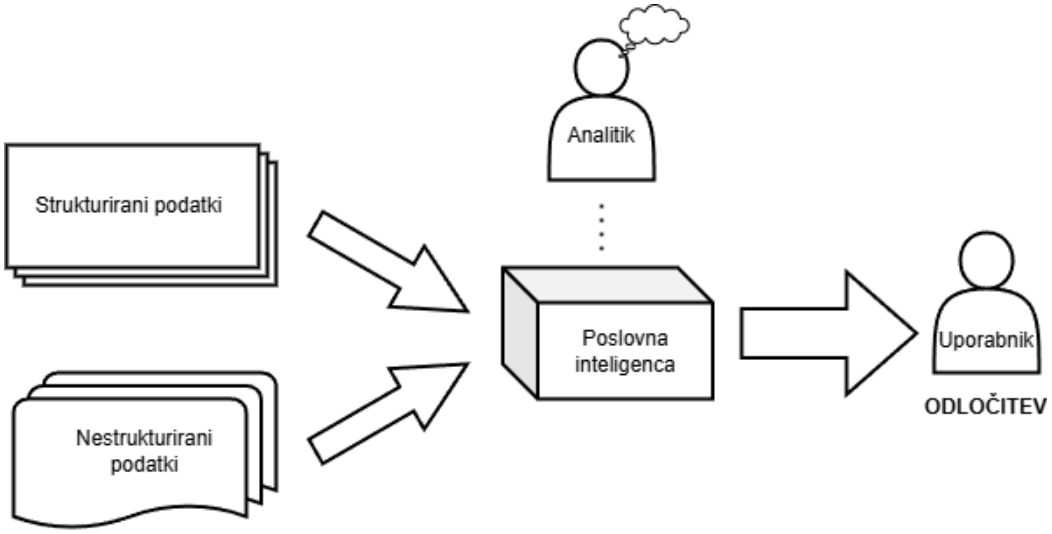
Namen poslovne inteligence je olajšati vodstveno delo z zagotavljanjem pravočasnoih in kakovostnejših vhodnih podatkov v procesu odločanja. Včasih BI pomeni odločanje v realnem času, torej takojšnji odziv. Večinoma pa se nanaša na skrajšanje časovnega okvira odločanja, da je torej inteligenca še vedno uporabna za odločevalca, ko napoči čas za odločitev (Langseth in Nithi, 2003).

Včasih so bile tehnologije BI kompleksne, težko dostopne in namenjene strokovnim analitikom in odločevalcem v podjetju. Z napredkom BI orodij se ta omejitev zmanjšuje, saj orodja postajajo uporabniku prijaznejša in bolj integrirana v vsakdanje poslovne procese. Z razvojem BI se je pojavil tudi pojem »BI za množice«, ki omogoča širšo dostopnost analitičnih orodij za zaposlene na vseh ravneh podjetja, ne le za podatkovne analitike in strokovnjake (Gehra in drugi, 2005).

BI dandanes podpira raznolikost informacijskih vhodov, ki so na voljo pri sprejemanju pravih odločitev. Informacijski vhodi so različni podatkovni viri, ki pa proizvajajo podatke v različnih oblikah. Podatki so ponavadi strukturirani ali nestrukturirani, poznamo pa tudi delno strukturirane podatke. Med strukturirane podatke uvrščamo podatke iz podatkovnih skladišč, podatke, ki so rezultat podatkovnega rudarjenja, celovitih programskih rešitev (angl. Enterprise Resource Planing, v nadaljevanju ERP), sprotne analitične obdelave podatkov (angl. Online Analytical Processing, v nadaljevanju OLAP), iz orodij za upravljanje odnosov s strankami (angl. Customer Relationship Management, v nadaljevanju CRM) in druge. Med nestrukturirane podatke uvrščamo npr. video posnetke, besedila, slike

in grafe. Vloga BI pri sprejemanju odločitev na podlagi raznolikih informacijskih vhodov je razvidna iz slike 1.

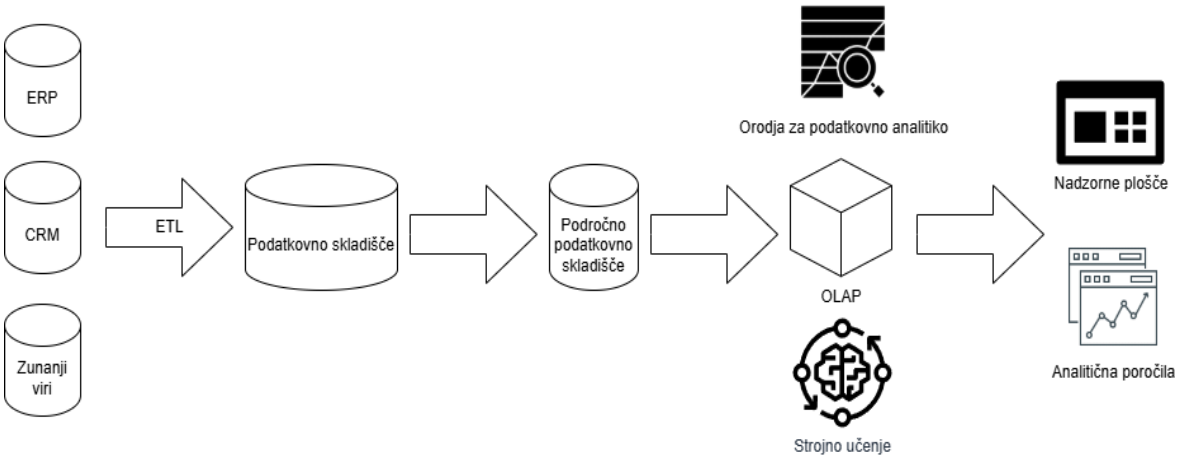
Slika 1: Poslovna inteligenca



Vir: lastno delo na podlagi Negash (2004).

Glede na obliko in število virov podatkov izberemo arhitekturo BI, ki najbolje ustreza pogojem in ciljem. Arhitektura BI je okvir za različne tehnologije, ki jih organizacija uporablja za izvajanje aplikacij BI in analitike. Vključuje informacijsko-tehnološke (IT) sisteme in programska orodja, ki se uporabljajo za zbiranje, integracijo, shranjevanje in analizo podatkov BI, ter predstavitev informacij o poslovnih operacijah in trendih (Yasar, 2023). Primer arhitekture BI s strukturiranimi podatkovnimi viri je viden na sliki 2. Iz slike je razviden tudi proces ETL, ki pomaga zagotavljati kakovost podatkov in bo natančneje opisan v nadaljevanju.

Slika 2: Primer arhitekture poslovne inteligence s strukturiranimi viri podatkov



Vir: lastno delo na podlagi Yasar (2023).

2.2 Digitalno trženje

2.2.1 Digitalni mediji

Digitalne medije je mogoče razumeti tako kot tehnološki pojav, ki omogoča digitalno komunikacijo, kot tudi kot širši kulturni in družbeni fenomen. Po eni strani jih Chaffey in Ellis-Chadwick (2019) opredeljujeta kot medije, v katerih komunikacija poteka prek vsebin in interaktivnih storitev, ki jih omogočajo digitalne platforme, kot so internet, svetovni splet, mobilne naprave in digitalna televizija. Po drugi strani pa Dewdney in Ride (2014) poudarjata, da digitalnih medijev ne gre razumeti le kot skupek tehnoloških orodij. Po njunem mnenju predstavljajo digitalni mediji presečišče tehnoloških, kulturnih in konceptualnih vidikov, kar pomeni, da so neločljivo povezani z načinom, kako jih ljudje uporabljajo, interpretirajo in vključujejo v vsakdanje življenje. Tako digitalni mediji niso le infrastruktura za distribucijo vsebin, temveč tudi orodje kulturnega izražanja, družbenega sodelovanja in medijske transformacije, ki spreminjajo odnose med ustvarjalci, potrošniki in vsebinami samimi.

Dandanes je za razvoj uspešne digitalne strategije, potrebno razumevanje zelo kompleksnega in visoko-konkurenčnega nakupovalnega okolja. Za razvoj primerne strategije, ki bi dosegla in vplivala na ciljno množico kupcev, se tržniki največkrat poslužujejo treh glavnih tipov medijski kanalov (Cahffey in Ellis-Chadwick, 2019):

- Plačani mediji (angl. paid media). So mediji v katere podjetja investirajo z namenom povečanja dosega ali konverzij preko iskanja, prikazovalnih oglasnih omrežij (angl. display ad networks) ali partnerskega trženja (angl. affilliate marketing).
- Lastni mediji (angl. owned media). So mediji v lasti blagovne znamke oziroma podjetja. Na spletu to vključuje spletna mesta podjetij, bloge, e-poštne sezname, mobilne aplikacije ali prisotnost na družbenih omrežjih kot so Facebook, LinkedIn in X.
- Zasluženi/pridobljeni mediji (angl. earned media). Tradicionalno ime za publiciteto, ustvarjeno preko odnosov z javnostmi, s ciljem vplivanja na ključne posameznike za povečanje prepoznavnosti blagovne znamke ali podjetja. Danes mednje uvrščajo tudi ustno izročilo, ki ga je mogoče spodbuditi preko viralnega trženja (angl. viral marketing) na družbenih medijih, ter pogovorov na družbenih omrežjih, blogih in drugih skupnostih. Zaslužene medije lahko razumemo tudi kot deljenje zanimive vsebine, ki jo ustvarijo različni partnerji, kot so založniki, blogerji, vplivneži in zagovorniki strank. Po drugi strani jih lahko obravnavamo kot različne oblike pogovorov med potrošniki in podjetji, ki potekajo tako na spletu kot izven njega.

2.2.2 Digitalno trženje

Digitalno trženje je definirano kot doseganje trženjskih ciljev z uporabo digitalnih medijev, podatkov in tehnologije (Cahffey in Ellis-Chadwick, 2019).V praksi se digitalno trženje

osredotoča na upravljanje različnih oblik spletne prisotnosti podjetja, kot so spletna mesta podjetja, mobilne aplikacije in strani podjetja na družbenih medijih. Te platforme so največkrat integrirane z metodami spletnega komuniciranja, vključno s trženjem s pomočjo spletnih iskalnikov (angl. search engine marketing), trženjem preko družbenih medijev (angl. social media marketing), spletnim oglaševanjem (angl. online advertising), e-poštnim trženjem in partnerstvom z drugimi spletnimi stranmi. Vse omenjene metode se uporabljajo s ciljem pridobivanja novih strank in zagotavljanja storitev obstoječim strankam. S tem pomagajo razvijati odnose s strankami preko upravljanja teh odnosov (več informacij o odnosu in orodjih za upravljanje odnosov s strankami bo predstavljenih v nadaljevanju).

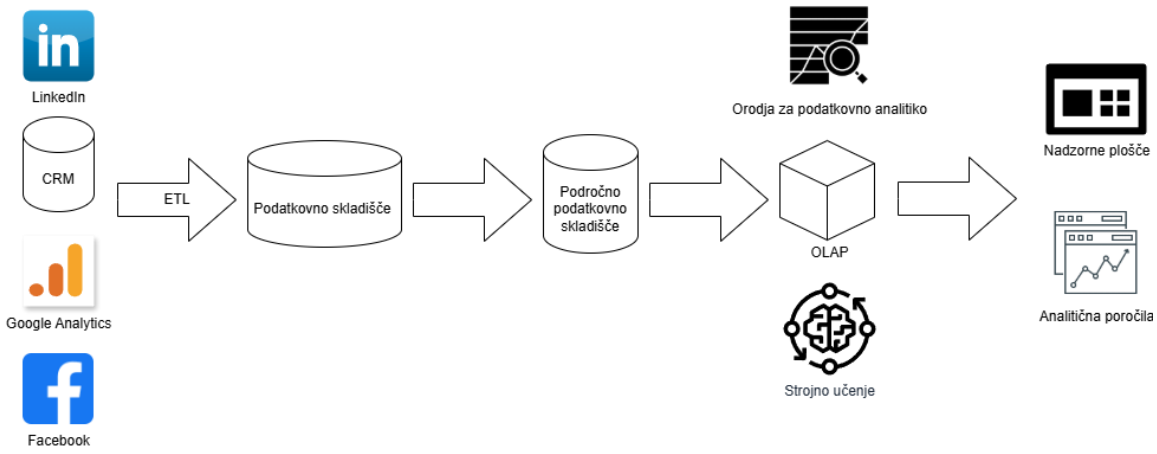
Digitalno trženje se pogosto obravnava kot sinonim za internetno trženje ali e-trženje, kar je napačno. Internet kot medij je le eden izmed številnih načinov doseganja strank (Sawicki, 2016). Za uspeh digitalnega trženja je še vedno potrebna integracija naštetih novodobnih tehnik s tradicionalnimi mediji, kot so tisk, televizija, neposredna pošta ter človeška prodaja in podpora, kot del večkanalnega trženjskega komuniciranja (Cahffey in Ellis-Chadwick, 2019).

2.3 Poslovna inteligenca v digitalnem trženju

Digitalno trženje postaja ključni proizvajalec informacij, ki zagotavlja konkurenčnost podjetja. Za tem stoji BI podjetja, opredeljen kot kombinacija tehnologij, arhitektur, ljudi, procesov in metodologij, ki pretvarjajo surove podatke v uporabne poslovne informacije. V digitalnem trženju se glavne tehnologije poslovne inteligence uporabljajo za poročanje, spletno analitično obdelavo, analitiko (preteklo in napovedno) ter rudarjenje podatkov in besedil. BI običajno zahteva integracijo podatkov iz različnih izvornih sistemov v podatkovno skladišče ter ustvarjanje definicij metapodatkov in podatkovnih modelov za podporo tej integraciji. Uporablja se tudi za analizo ter zagotavljanje podatkov in analiz na mestih, kjer so ti potrebni. Torej prek poročil ali posamičnih prenosov podatkov na kateri koli platformi ali napravi, ki jo uporabniki potrebujejo oziroma uporabljajo (Tvrdíková, 2007).

Izziv, ki se pojavi pri uporabi BI v digitalnem trženju, je integracija podatkov iz več podatkovnih virov, kot je prikazano na sliki 3. Ti viri so pogosto tuji viri (LinkedIn, Google Analytics, Facebook), ki niso v lasti podjetja ali uporabnika. Struktura podatkov se lahko kadarkoli spremeni, kar predstavlja dodaten izziv za integracijo in zagotavljanje kakovosti podatkov. Pri uporabi BI v digitalnem trženju je zato ključen korak proces ETL kot enega od možnih načinov zagotavljanja kakovosti podatkov. Več o tem procesu bo predstavljeno v nadaljevanju.

Slika 3: Primer poslovnointeligenčne arhitekture v digitalnem trženju



Vir: lastno delo na podlagi Yasar (2023).

Uspešnost digitalnega trženja temelji na ustrezno razvitem BI, ki omogoča hitro pridobivanje vpogleda v trenutno stanje digitalnega trženja. Na račun obstoja le »ene resnice«, se z uporabo BI prihrani čas, oblikujejo se boljše strategije ter načrti in sprejemajo boljše odločitve (Tvrdíková, 2007).

Poleg tehnološke podpore, ki jo nudi poslovna inteligenca, je za uspešno izvedbo digitalnega trženja ključno tudi spremljanje učinkovitosti s pomočjo ustrezno izbranih ključnih kazalnikov uspešnosti (angl. key performance indicators, v nadaljevanju KPI). KPI-ji predstavljajo temelj za merjenje uspeha digitalnih kampanj in omogočajo podjetjem vpogled v dejanski učinek trženjskih aktivnosti. Na podlagi kvantitativnih (npr. stopnja konverzije, število obiskov, donosnost naložbe (angl. Return on Investment - ROI)) in kvalitativnih kazalnikov (npr. uporabniška izkušnja, A/B testiranja) lahko podjetja natančneje analizirajo vedenje uporabnikov in optimizirajo svoje strategije (Saura in drugi, 2017).

Z uporabo KPI-jev in podatkovne analitike v okviru BI orodij podjetja pridobijo zanesljive in ažurne podatke, ki prispevajo k boljšemu odločanju ter povečanju konkurenčne prednosti. KPI-ji tako služijo kot vezni člen med tehnično arhitekturo poslovne inteligence in strateškimi cilji digitalnega trženja. Pravilno definirani in spremljani kazalniki omogočajo ne le merjenje uspešnosti, temveč tudi pravočasno prilagajanje trženjskih aktivnosti glede na spreminjajoče se vedenje uporabnikov in tržne pogoje (Saura in drugi, 2017).

2.4 Orodja in tehnologije poslovne inteligence

Včasih so bila orodja poslovne inteligence zapletena in nedostopna, na voljo so bila samo velikim podjetjem in podatkovnim znanstvenikom. Dandanes so dostopna vsakomur, ki želi dostopati do podatkov, jih vizualizirati ali analizirati (Grabbe, 2022). Orodja za poslovno inteligenco, ki jih upravljajo strokovnjaki, lahko v digitalnem trženju omogočijo odlične rešitve (Bhosale in drugi, 2020).

Orodja oziroma aplikacije BI Frankenfield (2024) deli na naslednje kategorije:

- Razpredelnice (angl. spreadsheet). So vrste aplikacij za računanje, analizo, organizacijo in hranjenje podatkov v obliki razpredelnice, npr. Excel, Google Sheets.
- Aplikacije za poročila in poizvedbe. So vrste aplikacij za pridobivanje, urejanje, agregacijo in prikaz podatkov, npr. Tableau, Power BI, SiSense.
- Aplikacije OLAP. So vrste aplikacij, ki omogočajo uporabnikom interaktivno analizo večdimenzionalnih podatkov z različnih zornih kotov. Aplikacije izvajajo tri osnovne analitične operacije: konsolidacijo (angl. roll-up), vrtanje v globino (angl. drill-down) ter rezanje in kockanje (angl. slicing and dicing).
- Konsolidacija je operacija, ki združuje podatke. Običajno to pomeni zbiranje in združevanje podatkov v večje enote ali višjo raven hierarhije. Na primer, prodajni podatki, ki so bili na začetku razdeljeni po dnevih, se lahko združijo v mesečne ali letne skupke.
- Vrtanje v globino je nasprotje konsolidacije. To pomeni premikanje z višje ravni hierarhije podatkov na nižjo, npr. iz letnih podatkov v mesečne ali dnevne, da bi pridobili podrobnejše informacije.
- Rezanje in kockanje omogoča uporabnikom, da pogledajo podatkovno kocko z različnih perspektiv. "Rezanje" se nanaša na izbiro ene določene ravni kocke in prikazovanje podatkov na tej ravni, medtem ko "kockanje" omogoča uporabnikom, da podatke analizirajo s treh različnih dimenzij v istem času, npr. Microsoft Analysis Service, Cognos, Essbase.
- Aplikacije za ustvarjanje nadzornih plošč. So vrste aplikacij, ki omogočajo izdelavo prikazov, ki ponujajo vpogled v KPI-je, pomembne za določen cilj ali poslovni proces, npr. Google Analytics, Tableau, Power BI, Excel.
- Aplikacije za podatkovno rudarjenje (angl. data mining). So vrste aplikacij, ki s pomočjo strojnega učenja, statistike, umetne inteligence odkrivajo in pridobivajo vzorce v naborih podatkov, npr. RapidMiner, Microsoft Analysis Services, R.

Digitalni tržniki se močno zanašajo na podatke za sprejemanje odločitev in oblikovanje kampanj, ki neposredno vplivajo na prihodke podjetja. BI orodja odpravljajo ročno analizo podatkov in potrebo po tehnični podpori, zaradi česar je za tržnike odločanje hitrejše kot kadarkoli. Po mnenju Grabbe (2022) med najbolj priljubljene vrste uporabe orodij sodijo:

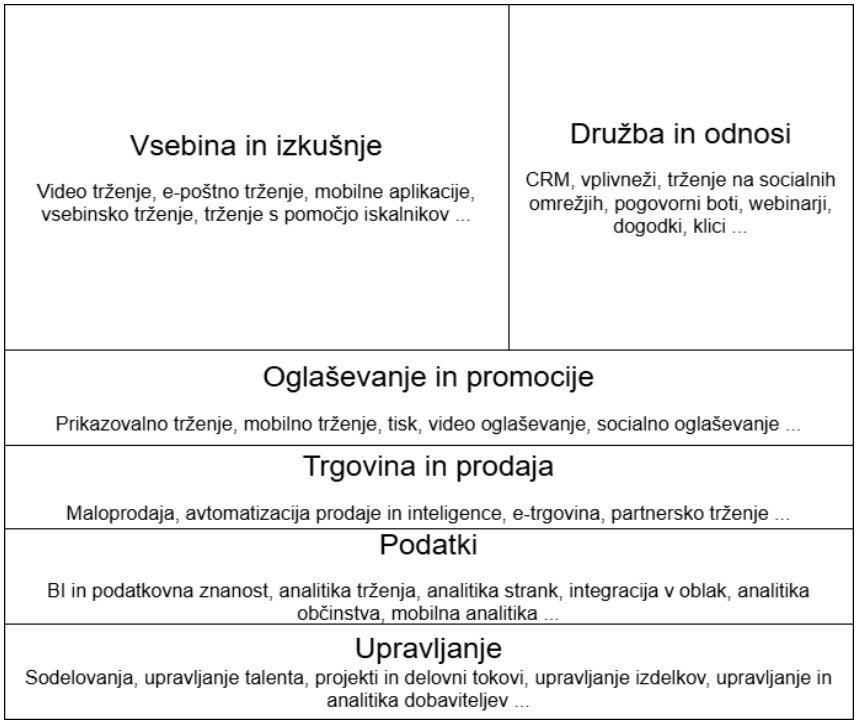
- sledenje in analiza spletnega obnašanja strank;
- zmanjševanje ročne analize podatkov;
- merjenje uspešnosti s pomočjo KPI-jev;
- ocenjevanje donosnosti trženjskih kampanj.

Z uporabo naštetih BI orodij lahko podjetja v BI okolju pridobijo boljši vpogled v vedenje strank, trende in učinkovitost trženjskih aktivnosti. Njihova uporaba v digitalnem trženju presega zgolj analizo podatkov – omogočajo tudi napovedovanje vedenja uporabnikov,

personalizacijo vsebin ter pravočasno prilagajanje strategij glede na spremembe na trgu. Povezava BI orodij z analitiko v realnem času omogoča sprejemanje hitrih, a premišljenih odločitev. Študije potrjujejo, da je prispevek BI orodij k digitalnemu trženju največji v organizacijah, ki svojo strategijo gradijo na globokem razumevanju strank. Uspešna uporaba BI orodij torej ni odvisna le od tehničnih zmogljivosti, temveč tudi od sposobnosti podjetja, da razume potrebe svojih strank in prilagaja svoje trženjske aktivnosti na podlagi teh vpogledov (Salah in Alzghoul, 2024).

Aktualne tehnologije digitalnega trženja je Brinker (2023) prikazal na zemljevidu dostopnih tehnologij. Število prikazanih tehnologij je 11038, kar predstavlja 11% povečanje v primerjavi z enakim obdobjem leta 2022.Zaradi lažjega pregleda in ločitve tehnologij na kategorije in podkategorije, sem Brinkerjevo shemo poenostavil v sliki 4. Vključuje 6 glavnih kategorij in 42 podkategorij, od katerih so naštete le najbolj zastopane.

Slika 4: Poenostavljen zemljevid tehnologij na področju digitalnega trženja



Vir: lastno delo na podlagi Brinker (2023).

3 INTEGRACIJA PODATKOV IZ VEČ VIROV

Kot omenjeno v uvodu, je integracija podatkov iz več virov izziv, s katerim se podjetja srečujejo v različnih panogah, vključno z digitalnim trženjem. Pred predstavitvijo procesa integracije predstavljam osnovne pojme, ki se v tem procesu pojavljajo.

3.1 Podatki in nabori podatkov

3.1.1 Podatki

V najosnovnejši obliki je podatek abstrakcija resničnega subjekta (osebe, predmeta ali dogodka). Izrazi, kot so spremenljivka, značilnost in atribut se pogosto izmenično uporabljajo za označevanje posamezne abstrakcije. Vsak subjekt je običajno opisan z več atributi. Na primer, knjiga ima lahko naslednje atribute: avtorja, naslov, temo, žanr, založnika, ceno, datum objave, število besed, število poglavij, število strani, izdajo in tako naprej (Kellerher in Tierney, 2018). Podatke razvrščamo na različne načine, ločimo jih lahko npr. po tipu in strukturi.

3.1.1.1 Tipi podatkov

Obstaja veliko različnih tipov atributov, za vsak tip pa je primerna točno določena vrsta analize. Razumevanje in prepoznavanje različnih tipov atributov in zanje primernih analiz je temeljna veščina podatkovnega znanstvenika. Standardni tipi podatkov so numerični, nominalni in ordinalni (Kellerher in Tierney, 2018):

- **Numerični atributi** opisujejo merljive količine, ki so predstavljene z uporabo celih ali realnih vrednosti. Lahko jih merimo na intervalni lestvici ali lestvici razmerja. Atribute intervalne lestvice merimo na lestvici s fiksnim, vendar poljubnim intervalom in izhodiščem - npr. meritve datuma in časa. Za atribute intervalne lestvice je primerno uporabljati operacije razvrščanja in odštevanja, druge aritmetične operacije niso primerne. Lestvice razmerja so podobne intervalnim, a s pravim ničelnim izhodiščem. Vrednost nič pomeni, da količine, ki jo merimo, ni. Posledično lahko vsako vrednost na lestvici opisujemo kot večkratnik druge vrednosti.
- **Nominalni atributi** (znani tudi kot kategorični) prevzamejo vrednosti iz končnega nabora. Te vrednosti so imena za kategorije, razrede ali stanja stvari. Primeri nominalnih atributov vključujejo vrste digitalnih medijev (plačani, lastni, pridobljeni). Binarna atribut je poseben primer nominalnega atributa, kjer je nabor možnih vrednosti omejen na le dve vrednosti. Na primer, imeli bi lahko binarni atribut »konverzija«, ki opisuje, ali je bila konverzija narejena (ja (angl. true)) ali ne (ne (angl. false)). Nominalnih atributov ni mogoče razvrščati ali na njih uporabljati aritmetičnih operacij. Nominalni atribut je možno razvrstiti po abecednem redu, vendar je to ločena operacija od navadnega razvrščanja.
- **Ordinalni atributi** so podobni nominalnim, a je možno njihovim kategorijam določiti rang. Na primer, atribut, ki opisuje odgovor na vprašanje ankete, bi lahko prevzel vrednosti: zelo nezadovoljen, nezadovoljen, nevtralen, zadovoljen in zelo zadovoljen. Pri teh vrednostih obstaja naravno urejanje, ki si sledi od »zelo nezadovoljen« do »zelo zadovoljen«. Pomembna značilnost ordinalnih podatkov je, da ni zagotovila o enaki razdalji med temi vrednostmi. Na primer, kognitivna razdalja med »zadovoljen« in »zelo

zadovoljen« ni enaka razdalji med nekaterimi drugimi vrednostmi. Zato na ordinalnih atributih ni primerno uporabljati aritmetičnih operacij.

3.1.1.2 Struktura podatkov

Po strukturi so lahko podatki strukturirani, nestrukturirani (Kellerher in Tierney, 2018) ali delno strukturirani (Batini in drugi, 2009):

- **Strukturirani podatki** so podatki, ki jih je možno shraniti v tabelo, pri čemer ima vsak vnos v tabeli enako strukturo (nabor atributov). Primer so lahko demografski podatki za populacijo, kjer vsaka vrstica v tabeli opisuje eno osebo in sestoji iz istega nabora demografskih atributov (ime, starost, datum rojstva, naslov, spol, izobrazbena raven, zaposlitveni status itd.). Strukturirane podatke je mogoče enostavno shranjevati, organizirati, iskati, prerazporediti in združiti z drugimi strukturiranimi podatki. Na njih je relativno enostavno uporabljati znanost o podatkih (angl. data science) saj so že po definiciji v formatu, ki je primeren za vključitev v analitični zapis.
- **Nestrukturirani podatki** so podatki, kjer ima vsak vnos v naboru podatkov svojo strukturo, ta struktura pa ni nujno enaka pri vsakem vnosu. Na primer, nabor podatkov spletnih strani, kjer ima vsaka spletna stran svojo edinstveno strukturo. Nestrukturirani podatki so pogostejši od strukturiranih. Besedilne zbirke (e-pošta, objave na platformah X, besedilna sporočila, objave, romani itd.) so primer nestrukturiranih podatkov, prav tako pa tudi zbirke zvoka, slik, glasbe, video posnetkov in multimedijskih datotek. Razlika v strukturi med različnimi elementi pomeni, da je nestrukturirane podatke v njihovi surovi obliki težko analizirati. Pogosto lahko iz nestrukturiranih podatkov izluščimo strukturirane podatke z uporabo tehnik umetne inteligence, kot so obdelava naravnega jezika (angl. natural-language processing), strojno učenje (angl. machine learning) ter, digitalna obdelava signalov in računalniški vid (angl. computer vision). Vendar pa sta implementacija in testiranje teh postopkov preoblikovanja podatkov draga in časovno potratna, ter lahko znatno povečata finančne stroške in časovne zamude pri projektih.
- **Delno strukturirani podatki** so podatki, ki imajo nepravilno strukturo. Nekateri objekti morda nimajo vseh atributov, drugi imajo več pojavov istega atributa, isti atribut ima lahko v različnih objektih različne tipe. Semantično povezane informacije so lahko v različnih objektih predstavljene na različne načine. Številne aplikacije shranjuje svoje podatke v nestandardnih formatih podatkov. Tak primer so zastareli sistemi in strukturirani dokumenti, ki vsebujejo označevalni jezik za oblikovanje večpredstavnostnih dokumentov (angl. hypertext markup language - HTML) ali pa standardizirani splošni označevalni jezik (angl. standardized generalized markup language - SGML) itd. Drug primer je integracija heterogenih virov podatkov. Pogosto ti viri pripadajo zunanjim organizacijam ali partnerjem in niso pod nadzorom aplikacije. Njihova struktura je le delno znana in se lahko spremeni brez obvestila (Suciu, 1998).

3.1.2 Metapodatki

Metapodatki so podatki, ki opisujejo ostale podatke (Kellerher in Tierney, 2018). Na primer podatki o posameznem stolpcu. Če ima stolpec naziv »Ime«, potem je metapodatek lahko stolpec »Ime« vsebuje imena delavcev. Metapodatki so ključni za sodobne informacijske sisteme, saj omogočajo boljše iskanje, povezovanje in dolgoročno upravljanje podatkov (Riley, 2017).

3.1.3 Nabori podatkov

Nabor podatkov sestavljajo podatki, ki se nanašajo na zbirko subjektov, pri čemer je vsak subjekt opisan s sklopom atributov. V svoji najosnovnejši obliki je nabor podatkov organiziran v n * m podatkovno matriko, imenovano analitični zapis, kjer je n število subjektov (vrstic) in m število atributov (stolpcev). V podatkovni znanosti se izraza nabor podatkov in analitični zapis pogosto uporabljata izmenično, pri čemer je analitični zapis posebna predstavitev nabora podatkov (Kellerher in Tierney, 2018).

3.2 Podatkovni viri

Podatkovni vir je lokacija, iz katere izvirajo podatki, ki jih uporabljamo. Vir podatkov je lahko začetno mesto, kjer se podatki ustvarijo ali kjer se fizične informacije prvič digitalizirajo. Kot vir služijo tudi najbolj prečiščeni podatki, dokler do njih dostopa drug postopek in jih uporablja (Talend, brez datuma).

Dandanes se organizacije za podporo pri odločanju poslužujejo več vrst podatkovnih virov. Čeprav več vrst virov predstavlja izziv za zagotavljanje kakovosti, ponuja boljše možnosti za analitiko in sprejemanje odločitev.

Vire podatkov lahko razdelimo na več načinov, ena od delitev je na zunanje (angl. external data), notranje (angl. internal data) in odprte vire (angl. open data) ter analitiko tretje osebe (angl. third-party analytics) (FutureLearn, brez datuma):

- **Zunanji viri podatkov** lahko vključujejo skoraj vse, od zgodovinskih in demografskih podatkov, tržnih cen ali vremenskih razmer, do trendov na družbenih omrežjih. Organizacije uporabljajo zunanje podatke za analizo in modeliranje ekonomskih, političnih, socialnih ali okoljskih dejavnikov, ki vplivajo na njihovo poslovanje (FutureLearn, brez datuma). V primeru digitalnega trženja so to najpogosteje podatki iz družbenih omrežij (X, Facebook, LinkedIn).
- **Notranji podatki** so podatki, ki nastanejo v okviru poslovnih procesov v podjetju oziroma jih proizvedejo notranji podatkovni viri. Podjetje ima lahko npr. na voljo strojno generirane podatke iz senzorjev ali naprav, ki sodelujejo pri izdelavi izdelka, ali pa jih zabeleži sam izdelek. V primeru digitalnega trženja naprej pomislimo na podatke iz CRM sistema, v primeru proizvodnega ali prodajnega podjetje pa na ERP sistem.

- Če podjetje ni zmožno samo zajemati podatkov, se lahko poslužuje **analitike tretje osebe**. Storitve spletne analitike tretjih oseb lahko zagotavljajo stroškovno učinkovito zbiranje in analizo. Tipičen primer v digitalnem trženju je orodje Google Analytics, ki bo predstavljeno v nadaljevanju.
- **Odprti viri podatkov** so dostopni vsem in jih je mogoče brezplačno uporabljati. Podatki so dostopni v različnih formatih ali pa že agregirani, kar je lahko prednost ali pa slabost. Odprte vire lahko tako v digitalnem trženju kot v drugih panogah uporabljamo za primerjavo uspešnosti med različnimi podjetji.

3.3 Homogenost in heterogenost

Homogenost in heterogenost podatkov sta v podatkovni znanosti definirani na več ravneh. Ti značilnosti lahko pripišemo podatkom, naborom podatkov, podatkovnim virom in celo podatkovnim bazam ter skladiščem.

3.3.1 Homogenost/heterogenost podatkov in naborov podatkov

Heterogeni podatki so podatki z visoko variabilnostjo tipov in formatov podatkov. Pogosto so lahko dvoumni in slabe kakovosti zaradi manjkajočih vrednosti, velike redundance podatkov in netočnosti (Wang, 2017).

Homogeni podatki so skladni po tipu in formatu. Homogeni nabori podatkov so tisti, ki vsebujejo atribute enakega tipa ali področja. Če imamo tabelo, v kateri so vsi stolpci numeričnega tipa, je ta nabor homogen (Huang, 2020).

3.3.2 Homogenost/heterogenost podatkovnih virov

Heterogenost v kontekstu virov podatkov se nanaša na obstoj različnih in raznolikih virov podatkov. Heterogeni viri podatkov lahko izvirajo iz različnih sistemov, platform ali so različnih formatov (Chanial in drugi, 2018, str. 4). Posledične razlike lahko predstavljajo izziv pri integraciji ali analizi tovrstnih podatkov.

Homogenost v kontekstu virov podatkov pomeni doslednost in enotnost virov podatkov. Homogeni viri podatkov izvirajo iz istih ali podobnih sistemov, platform ali so istih formatov (Huang, 2020). Enotnost poenostavi proces zbiranja, integracije in analize podatkov.

3.4 Podatkovno skladišče

Skladiščenje podatkov ni izdelek ampak postopek za zbiranje in upravljanje podatkov iz različnih virov, z namenom pridobivanja enotnega, podrobnega vpogleda na del ali celotno dejavnost (Gardner, 1998). Podatkovno skladišče je glavno nahajališče zgodovinskih podatkov organizacije in je optimizirano za poročanje in analizo (Rifaie in drugi, 2008).

Podatkovna skladišča so zasnovana za podporo združevanju in integraciji podatkov tako iz različnih notranjih kot tudi zunanjih podatkovnih virov (Strand in Benkt, 2004, str. 272).

Podjetja se za implementacijo podatkovnih skladišč odločajo iz različnih razlogov. Najpogostejši razlogi so: hiter dostop, enostavna uporaba ter učinkovito deljenje in vzdrževanje kakovostnih in pravočasnih poslovnih informacij, ki jih potrebujejo za doseganje uspeha pri svojih poslovnih ciljih (Rifaie in drugi, 2008).

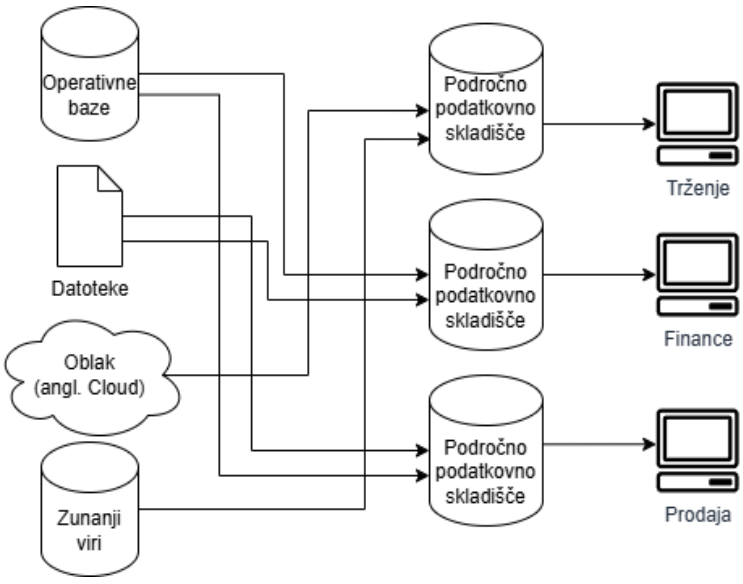
3.4.1 Arhitekture podatkovnega skladiščenja

Kljub pomembnosti in naraščajočim izkušnjam z uporabo skladiščenja podatkov, so mnenja o najprimernejši oz. najuporabnejši arhitekturi podatkovnih skladišč vedno izrazito deljena. V primeru nepravilne odločitve pri izbiri arhitekture lahko nastopijo težave, kot so pomanjkanje skalabilnosti, težave z zmogljivostjo in ustvarjanje več verzij resnice (Ariyachandra in Watson, 2010).

3.4.1.1 Arhitektura neodvisnih področnih podatkovnih skladišč

Arhitektura neodvisnih področnih podatkovnih skladišč (angl. Independent Data Marts - IDM), prikazana na sliki 5, je prilagojena tako, da ima vsak oddelek ali uporabnik na voljo podatke iz posameznega področnega podatkovnega skladišča, ki mu je namenjen. Podatki prihajajo direktno iz virov, transformacije pa se opravijo skladišča. Arhitektura sicer pokriva potrebe posameznih oddelkov, ne zagotavlja pa ene verzije resnice celotne organizacije (Ariyachandra in Watson, 2010).

Slika 5: Arhitektura neodvisnih področnih podatkovnih skladišč

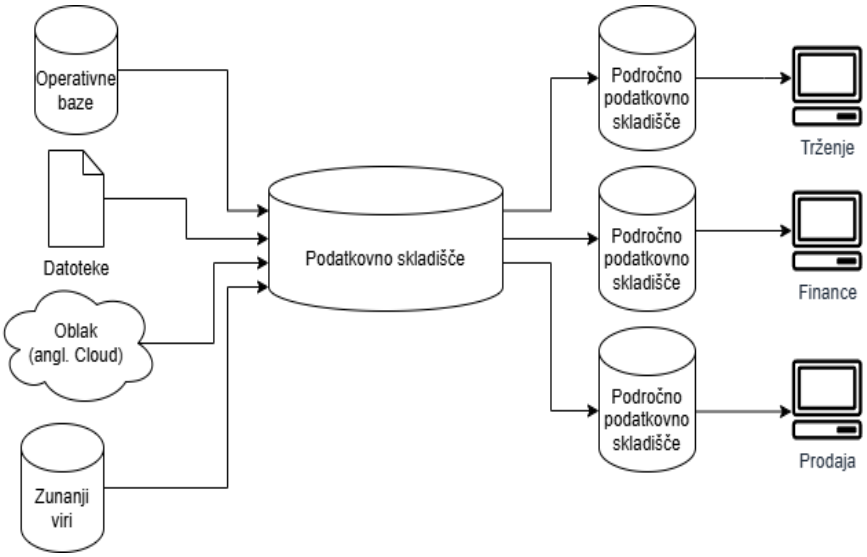


Vir: lastno delo na podlagi Gardner (1998) in Hajmoosaei in drugi (2011).

3.4.1.2 Arhitektura odvisnih področnih podatkovnih skladišč

Arhitektura odvisnih področnih podatkovnih skladišč (angl. Dependent Data Marts - DDM), prikazana na sliki 6, je zgrajena tako, da se podatki najprej naložijo v podatkovno skladišče, od tam pa specifični in relevantni podatki naprej v posamezna področna podatkovna skladišča. Proces nalaganja podatkov je poenostavljen, saj so podatki očiščeni že pri nalaganju v podatkovno skladišče. Tudi v tem primeru so pokrite potrebe posameznikov, ni pa zagotovljena ena verzija resnice celotne organizacije (Ariyachandra in Watson, 2010).

Slika 6: Arhitektura odvisnih področnih podatkovnih skladišč

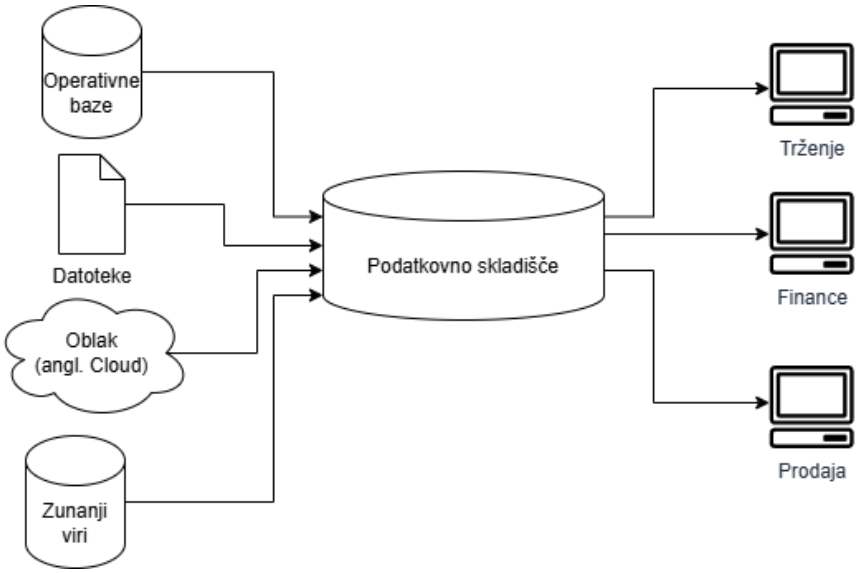


Vir: lastno delo na podlagi Gardner (1998) in Hajmoosaeiin drugi (2011).

3.4.1.3 Arhitektura organizacijskega podatkovnega skladišča

Arhitektura organizacijskega podatkovnega skladišča (angl. Enterprise Data Warehouse, v nadaljevanju EDW), prikazana na Slika 7 7, omogoča dostop do podatkov vsem deležnikom. Zagotavlja tudi eno verzijo resnice celotne organizacije in omogoča analitiko med posameznimi oddelki (Ariyachandra in Watson, 2010). EDW služi kot osrednje skladišče za zgodovinske in trenutne podatke ter je zasnovano tako, da omogoča robustno, prilagodljivo in razširljivo podatkovno infrastrukturo. Ključni podatki so strukturirani na način, ki omogoča učinkovito poročanje in podporo odločanju, z uporabo tehnik, ki se razlikujejo od tistih v operativnih sistemih. Ena od pomembnih prednosti EDW arhitekture je, da razbremeni operativne sisteme potrebe po shranjevanju zgodovinskih podatkov, saj se ti centralizirano hranijo v podatkovnem skladišču. Uspešna implementacija EDW običajno zahteva sodelovanje tehničnega osebja in poslovnih uporabnikov, saj se razvoj izvaja postopoma, z iterativnimi preverjanji in izboljšavami glede na uporabniške potrebe (Rifaie in drugi, 2008).

Slika 7: Organizacijsko podatkovno skladišče



Vir: lastno delo na podlagi Rifaie in drugi (2008).

3.4.2 Podatkovno skladiščenje v oblaku

V zadnjih letih se je pojavilo podatkovno skladišče v oblaku (angl. Cloud Data Warehouse), nov tip podatkovnega skladišča z vidika oblikovanja in implementacije. Model se je razvijal postopoma, kar vpliva na aplikacijske in poslovne domene, razvit pa je bil za nadzor nad podatki velikega obsega. Omogoča povečanje in zmanjšanja kapacitet v kateremkoli trenutku in nima omejitev glede števila uporabnikov. Po mnenju Ur Rehman in drugi (2018) je podatkovno skladiščenje v oblaku prihodnost podatkovnega skladiščenja. Primerjava med tradicionalnim in podatkovnim skladiščem v oblaku je prikazana v tabeli 1.

Tabela 1: Primerjava tradicionalnega in podatkovnega skladišča v oblaku

Lastnosti	**Tradicionalno podatkovno skladišče**	**Podatkovno skladišče v oblaku**
Stroški	Višji začetni stroški za strojno opremo in licenciranje; stalni stroški vzdrževanja in podpore.	Lahko nižji začetni stroški, mesečno ali letno plačilo na podlagi porabe; nižji stroški podpore.
Upravljanje	Popolna kontrola in odgovornost za vzdrževanje in upravljanje.	Deljena odgovornost s ponudnikom; manjša kontrola, manj obveznosti.
Nove funkcionalnosti	Počasnejše uvajanje novih funkcij, možna zamuda pri nadgradnjah, potreba po ročni implementaciji.	Hitrejše uvajanje novih funkcij, redne nadgradnje s strani ponudnika.
Skalabilnost	Težavnejše in dražje nadgradnje; potreba po fizičnih spremembah ali nakupu dodatne opreme.	Hitra in dinamična prilagodljivost brez potrebe po fizičnih spremembah.

se nadaljuje

Tabela 1: Primerjava tradicionalnega in podatkovnega skladišča v oblaku (nad.)

Lastnosti	**Tradicionalno podatkovno skladišče**	**Podatkovno skladišče v oblaku**
Nadzor nad uspešnostjo	Več orodij za nadzor, popolna kontrola nad optimizacijo, potreba po več znanja in orodjih.	Orodja prilagojena oblaku, odvisnost od ponudnika za reševanje večjih težav.
Varnost	Popolna odgovornost za varnost od strojne opreme do aplikacij; potreba po stalnem usposabljanju in investiranju v varnost.	Deljena odgovornost s ponudnikom; korist od naprednih varnostnih orodij ponudnika.
Revizija in regulativa	Ponudnik običajno ponuja potrdila in poročila o skladnosti, deljena odgovornost za pridobitev vseh dokazov o skladnosti s predpisi.	Popolna odgovornost za skladnost, vendar z neposrednim dostopom do vseh potrebnih podatkov in orodij za revizijo.

Vir: lastno delo na podlagi Foot (2024).

Med oblačne rešitve za podatkovno skladiščenje med drugim spadajo Amazon Redshift, Google BigQuery in Snowflake, ki so zelo skalabilne ter ponujajo boljše razmerje med ceno in zmogljivostjo v primerjavi z alternativami. Te platforme so optimizirane za obdelavo velikih količin podatkov (Seenivasan, 2022).

3.5 Proces pridobivanja, preoblikovanja in nalaganja podatkov

Proces pridobivanja, preoblikovanja in nalaganja obstaja že od zgodnjih sedemdesetih let prejšnjega stoletja. Sprva je bil ETL proces ročno kodiran in prožen, kar pomeni, da je bil celoten postopek dolgotrajen in zelo nagnjen k napakam. Z leti se je proces razvil na račun sodobnih orodij, ki nudijo različne funkcije na področju avtomatizacije in nadzora (Seenivasan, 2022).

Proces ETL, je ključen postopek pri obdelovanju podatkov. Odgovoren je za pridobivanje podatkov iz različnih virov, njihovo preoblikovanje v uporabno obliko, do končnega nalaganja teh podatkov v sistem. Proces pomaga podjetjem razumeti podatke s pomočjo njihovega združevanja iz različnih virov in priprave podatkov za analizo (Seenivasan, 2023). Poleg tega z njim zagotavljamo kakovost podatkov.

Če želimo v podatkovnem skladišču preoblikovati podatke ali pa želimo razbremeniti podatkovno skladišče, lahko pridobljene podatke najprej izvozimo v področje priprave podatkov (angl. staging). Tu se podatki preoblikujejo in naknadno naložijo v podatkovno skladišče (Seenivasan, 2023).

Avtomatizacija ETL procesa lahko prihrani čas in zmanjša število napak. Z načrtovanjem izvajanja ETL procesa ob določenih časovnih intervalih lahko zagotovimo, da so podatki ažurni ter da je obremenitev sistema enakomerno porazdeljena (Seenivasan, 2023).

Poleg procesa ETL, obstaja tudi sodobnejši proces ELT oziroma procesa pridobivanja, nalaganja in preoblikovanja podatkov. Uporaba ELT se je razširila z razvojem zmogljivih platform za obdelavo podatkov v sodobnih podatkovnih skladiščih. To je organizacijam omogočilo izkoriščanje računalniške moči podatkovnih skladišč za izvajanje transformacij, kar je privedlo do širše uvedbe ELT. Razvoj ELT je pospešila tudi uporaba računalništva v oblaku in tehnologij masovnih podatkov (angl. big data) (Seenivasan, 2022).

Procesa ETL in ELT imata svoje prednosti in slabosti, kar je prikazano v tabeli 2. Izbira ustreznejšega procesa temelji predvsem na podatkovni arhitekturi, stopnji kompleksnosti transformacij in zmogljivost podatkovnega skladišča. Pred dokončno odločitvijo bi morala podjetja skrbno preučiti svoj pristop k upravljanju podatkov, pri čemer je ključnega pomena količina podatkov, s katerimi delajo, hitrost njihove obdelave ter potreba po analitiki v realnem času. Z izbiro najprimernejšega procesa lahko izboljšajo podatkovne tokove, optimizirajo zmogljivost in pridobijo kakovostnejše vpoglede v podatke, kar jim omogoča večjo konkurenčno prednost (Seenivasan, 2022).

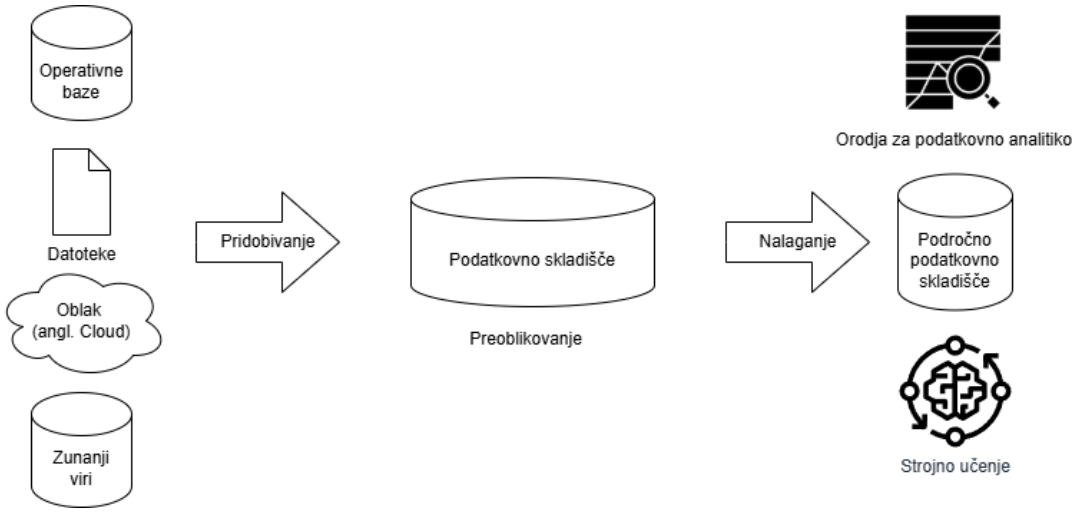
Tabela 2: Primerjava procesov za pridobivanje, obdelavo in nalaganje podatkov

Lastnosti	**ETL**	**ELT**
Zmogljivost	Počasnejši zaradi transformacij podatkov pred nalaganjem	Hitrejši zaradi neposrednega nalaganja surovih podatkov
Skalabilnost	Omejena zaradi kapacitet preoblikovalnega pogona	Odvisna od zmogljivosti podatkovnega skladišča
Stroški	Višji začetni stroški vzpostavitve in vzdrževanja	Nižji začetni stroški vzpostavitve, lahko višji operativni stroški
Okolje	Tradicionalna orodja, relacijske podatkovne baze	Oblačna orodja, namenjena masovnim podatkom
Struktura podatkov	Strukturirani podatki	Podatki so lahko strukturirani, nestrukturirani ali delno strukturirani

Vir: lastno delo na podlagi Seenivasan (2022) in Tayade (2019).

Kot že omenjeno, je proces ETL, prikazan na sliki 8 sestavljen iz treh ključnih korakov: pridobivanja, preoblikovanja in nalaganja podatkov, ki se lahko delijo še naprej. Vsak od teh korakov ima svojo vlogo pri zagotavljanju kakovostnih in uporabnih podatkov za nadaljnjo analizo.

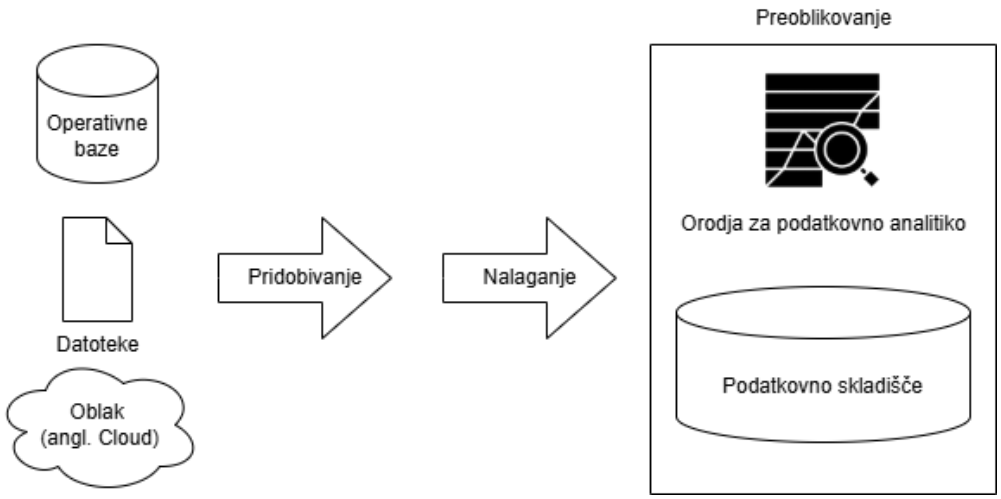
Slika 8: Proces pridobivanja, preoblikovanja in nalaganja podatkov



Vir: lastno delo na podlagi Seenivasan (2023).

V primerih, ko želimo imeti podatke pri roki in jih preoblikovati po potrebi, oziroma na zahtevo, namesto procesa ETL uporabimo proces ELT, prikazan na sliki 9. V sklopu procesa se podatki najprej naložijo v ciljni sistem, kjer se izvaja preoblikovanje glede na zahteve uporabnikov ali analitičnih orodij.

Slika 9: Proces pridobivanja, nalaganja in preoblikovanja podatkov



Vir: lastno delo na podlagi Singhal in Aggarwal (2022).

3.5.1 Pridobivanje podatkov

V koraku pridobivanja se identificirajo vsi ustrezni viri, nato pa se podatki izvlečejo na čim učinkovitejši način. Postopek vključuje pregledovanje datotek ali zbirk podatkov, uporabo različnih kriterijev za izbiro podatkov, iskanje ustreznih podatkov in njihov prenos v datoteko ali drugo zbirko podatkov (Wijaya in Pudjoatmodjo, 2015).

3.5.2 Preoblikovanje podatkov

Preoblikovanje je postopek manipulacije s podatki iz izvornega sistema v drug format v področnem podatkovnem skladišču ali v glavnem podatkovnem skladišču, z namenom pridobivanja smiselnih informacij (Wijaya in Pudjoatmodjo, 2015). V okviru preoblikovanja obstaja več postopkov, ki omogočajo zagotavljanje kakovosti podatkov kot jih navajajo Sreemanthy in drugi (2020), ter spletno mesto Keboola (2022):

- **Čiščenje** je proces sestavljen iz več korakov, zato je podrobneje opisan v podpoglavju.
- **Filtriranje** je ena najenostavnejših oblik preoblikovanja, kjer se posamezni stolpci ali vrstice izločijo pred nalaganjem v podatkovno skladišče.
- **Agregacija** je oblika povzemanja podatkov. Izberemo si dimenzijo in metriko za manipulacijo te dimenzije. Najpogostejši primer je datumska dimenzija: imamo časovno določene podatke vseh dogodkov, zanima nas, koliko dogodkov se je zgodilo vsak dan. Dogodke seštejemo po datumu in dobimo agregirane vrednosti.
- **Združevanje oziroma integracija** je obsežen proces in je zato opisan v podpoglavju.
- **Ločevanje** se najpogosteje uporablja pri nestrukturiranih ali delno strukturiranih podatkih, denimo pri besedilih, ločenih z vejicami, kjer vsak del predstavlja samostojen podatek.
- **Razvrščanje** je proces, ki olajša iskanje podatkov. Podatke lahko pred nalaganjem v podatkovno skladišče razvrstimo po datumu.
- **Odstranjevanje podvojenih vrednosti** pomeni odstranjevanje vrstic zaradi ponavljanja. Pri podatkovnih virih, kjer so vrednosti unikatne, je napake lažje odkriti.

3.5.2.1 Čiščenje podatkov

Čiščenje podatkov je proces, ki se ukvarja z odkrivanjem in odstranjevanjem napak ter neskladij v podatkih, z namenom izboljšanja njihove kakovosti. Težave s kakovostjo podatkov so lahko prisotne v vseh podatkovnih virih. Razlogi za nizko kakovost podatkov so različni, npr. napake pri vnosu podatkov, manjkajoče informacije ali drugi neveljavni podatki. Ko je treba združiti več virov podatkov, se potreba po čiščenju podatkov znatno poveča, ker viri pogosto vsebujejo odvečne podatke v različnih formatih. Da bi zagotovili dostop do natančnih in doslednih podatkov, je nujno združiti različne predstavitve podatkov ter odpraviti podvojene informacije. Proces čiščenja podatkov je običajno sestavljen iz več korakov (Rahm in Do, 2000):

- **Analiza podatkov**: Da bi ugotovili, katere vrste napak in neskladij je treba odstraniti, je potrebna podrobna analiza podatkov. Poleg ročnega pregleda podatkov ali vzorcev podatkov je priporočljivo uporabiti analitična orodja. Omogočijo nam pridobivanje metapodatkov o lastnostih podatkov in odkrivanje težav s kakovostjo podatkov.
- **Definiranje poteka preoblikovanja in preslikave**: Glede na število virov podatkov, njihovo stopnjo heterogenosti in "umazanosti" je treba izvesti določeno število korakov

za pretvorbo in čiščenje. Včasih se uporabljajo sheme za preslikave virov v skupni model podatkov. Sheme določajo, kateri stolpci se medsebojno povežejo v primerih, ko so poimenovani različno. Za podatkovna skladišča se običajno uporablja relacijska predstavitev, ki poveže posamezne tabele; ta predstavitev določa kateri ključ v eni tabeli se poveže z enakim ključem v drugi tabeli. Zgodnji koraki čiščenja podatkov lahko odpravijo težave z enojnimi viri in pripravijo podatke za integracijo. Kasnejši koraki zajemajo shematsko integracijo podatkov in čiščenje ter definiranje težav z več viri, npr. podvojitvami. Za podatkovno skladiščenje bi morala biti kontrola in tok podatkov za te korake preoblikovanja in čiščenja določena s potekom dela (angl. workflow), ki opredeljuje ETL postopek. Transformacije, povezane s shemo, kot tudi koraki čiščenja, bi morali biti določeni s pomočjo deklarativnega jezika za poizvedbe in preslikave, da omogočijo avtomatsko generiranje transformacijske kode. Poleg tega bi moralo biti med pretvorbami podatkov mogoče vključiti uporabniško napisano kodo za čiščenje in posebna orodja. Koraki transformacije lahko zahtevajo povratne informacije uporabnika o posameznih primerih, za katere nimajo vgrajene logike čiščenja.

- **Preverjanje**: Pravilnost in učinkovitost poteka preoblikovanja in definicij preoblikovanja, je treba preizkusiti in ovrednotiti, npr. na vzorcu ali kopiji izvornih podatkov, da se po potrebi izboljšajo definicije. V nekaterih primerih so potrebne večkratne iteracije analize, načrtovanja in preverjanja, ker nekatere napake postanejo očitne šele po uporabi specifičnih preoblikovanj.
- **Preoblikovanje**: Izvedba preoblikovalnih korakov, bodisi z zagonom ETL procesa za nalaganje in osveževanje podatkovnega skladišča bodisi s pomočjo poizvedb iz več virov.
- **Povratni tok očiščenih podatkov**: Ko so napake odstranjene, bi morali očiščeni podatki nadomestiti »umazane« podatke v izvirnih virih. To bi aplikacijam omogočilo uporabo izboljšanih podatkov in preprečilo ponovno čiščenje pri prihodnjih pridobivanjih podatkov. Pri podatkovnem skladiščenju so očiščeni podatki na voljo bodisi iz področja priprave podatkov bodisi direktno iz podatkovnega skladišča.

3.5.2.2 Združevanje oziroma integracija podatkov

Integracija podatkov, kot jo predstavi Lenzerini (2002), je izziv združevanja podatkov, ki se nahajajo v različnih podatkovnih virih in pa zagotavljanje enotne predstavitve podatkov vsem uporabnikom.

Integracija podatkov je ključna za npr.: velika podjetja s številnimi viri podatkov, napredek v obsežnih znanstvenih projektih, kjer nabore podatkov neodvisno prispeva več raziskovalcev, boljše sodelovanje med vladnimi agencijami z individualnimi viri podatkov in zagotavljanje dobre kakovosti iskanja preko milijonov strukturiranih virov podatkov na svetovnem spletu (Halevy in drugi, 2006).

Referenčni modeli, semantična mreža (angl. semantic web), standardi, ontologije in druge tehnologije omogočajo hitro in učinkovito združevanje heterogenih podatkov, pri čemer je zanesljivost pridobljenih informacij v veliki meri odvisna od tega, kako dobro podatki predstavljajo realnost (Brazhnik in Jones, 2007).

Integracija podatkov se lahko izvaja na dva načina. Ko natančno vemo, na katera vprašanja želimo odgovoriti in kateri podatki so na voljo, se osredotočimo na iskanje zanesljivih virov ter pridobivanje potrebnih podatkovnih polj za vnos v podatkovno bazo. Drugi način integracije je namenjen razumevanju moči podatkov na splošno, ko je na voljo več virov podatkov, z različnimi stopnjami zanesljivosti. Pri projektih te vrste je cilj lahko opredeljen kot iskanje korelacije ali odkrivanje novih informacij prek različnih tehnik rudarjenja podatkov. Ti projekti so usmerjeni v ugotavljanje vsebine obstoječih virov, oceno njihove zanesljivosti in določanje, kateri dodatni podatki so potrebni za omogočanje odgovarjanja na kompleksna interdisciplinarna vprašanja (Brazhnik in Jones, 2007).

Pri integraciji se zaradi različnih virov podatkov pogosto pojavljajo izzivi. Za nekatere že obstajajo rešitve, drugi ostajajo nerešeni. Foote (2023) predstavi 4 najpogostejše izzive in njihove rešitve:

- **Podatki niso tam, kjer bi morali biti**. Ta pogosta težava se pojavi, ko podatki niso shranjeni na centralni lokaciji ampak razpršeni po različnih oddelkih organizacije. To povečuje tveganje, da med raziskovanjem pogrešimo ključne informacije. Enostavna rešitev je shranjevanje vseh podatkov na enem mestu (ali morda dveh - v glavni zbirki podatkov in podatkovnem skladišču). Izjema so osebni podatki, ki jih ščiti zakon.
- **Zakasnitve pri zbiranju podatkov**. Pogosto je treba podatke obdelati v realnem času, da zagotovimo natančne in smiselne vpoglede. Če morajo podatkovni tehniki izvesti ročni postopek integracije podatkov, obdelava v realnem času ni možna. To vodi do zakasnitev pri obdelavi strank in analitiki. Rešitev je uporaba avtomatiziranih orodij za integracijo podatkov. Razvili so jih posebej za obdelavo podatkov v realnem času, kar spodbuja učinkovitost in zadovoljstvo strank.
- **Težave z oblikovanjem nestrukturiranih podatkov**. Pogost izziv pri podatkovni integraciji je obdelava nestrukturiranih podatkov (fotografij, video in avdio posnetkov, družbenih medijev). Podjetja vse pogosteje ustvarjajo in zbirajo nestrukturirane podatke, ki pogosto vsebujejo za poslovne odločitve uporabne informacije. Na žalost so nestrukturirani podatki zahtevni za računalniško analizo. Za pomoč pri njihovi analizi so na voljo nova programska orodja, kot sta MonkeyLearn, ki uporablja strojno učenje za iskanje vzorcev in Cogito, ki uporablja obdelavo naravnega jezika (angl. natural language processing).
- **Podatki različne kakovosti**. Podatki slabe kakovosti negativno vplivajo na analize in lahko vodijo do napačnih odločitev. Težavni so npr. zastareli podatki, ki niso več relevantni ali pa so v neposrednem konfliktu z aktualnimi informacijami. Tudi podvojeni ali delno podvojeni podatki lahko ponujajo napačno sliko stanja. Rešitev za te izzive je

izboljšanje procesa ETL, s posebnim poudarkom na izboljšanju procesa čiščenja podatkov. Možnost je tudi vpeljava naprednih orodij ETL za avtomatizacijo teh korakov.

3.5.3 Nalaganje podatkov

Proces nalaganja pomeni shranjevanje podatkov v končno bazo podatkov oziroma podatkovno skladišče ali ciljno destinacijo. Nalaganje je zaključna faza procesa ETL in vključuje nalaganje velikih količin podatkov v ciljno destinacijo (Sreemanthy in drugi, 2020).

3.6 Orodja za pridobivanje, preoblikovanje in nalaganje podatkov

Funkcije procesa ETL lahko vključimo v ustrezna orodja, kar omogoča podjetjem enostavno in učinkovito poslovanje. Pri razvoju ETL orodij je ključnega pomena sposobnost pridobivanja podatkov iz heterogenih podatkovnih virov. Ta orodja omogočajo avtomatizacijo celotnega ETL procesa. Med najbolj prepoznavna orodja spadajo: Informatica, IBM Infosphere Datastage, Ab Initio, SQL Server Integration Services (SSIS), Oracle Data Integrator (Mukherjee in Kar, 2017), pa tudi Power BI. Slednji sicer ni primarno ETL orodje, a ponuja tudi vse korake ETL procesa.

Z leti so se ETL orodja močno razvila. Večina sodobnih orodij omogoča grafični uporabniški vmesnik za oblikovanje poteka dela in upravljanje ETL procesov. Poleg tega ta orodja vključujejo dodatne funkcije za avtomatizacijo, kot so nadzor napak, časovno načrtovanje in optimizacija procesa. Med ključne napredke na tem področju štejemo (Seenivasan, 2022):

- uvedbo orodij za profiliranje podatkov, ki omogočajo uporabnikom načrtovanje ETL sistemov s pomočjo grafičnih razvojnih vmesnikov;
- izboljšane procese preoblikovanja podatkov, posploševanje podatkovnih transformacij in napredno skriptiranje;
- izboljšano odkrivanje napak podatkov in njihovo beleženje za zagotavljanje višje kakovosti podatkov ter celovitih revizijskih sledi.

4 ZAGOTAVLJANJE KAKOVOSTI PODATKOV

Da bi lahko vodilni v podjetjih sprejemali pravilne poslovne odločitve na podlagi analiz, kljub heterogenim virom podatkov, je treba v podjetjih zagotavljati visoko raven kakovosti podatkov.

Izraz kakovost se običajno uporablja za označevanje proizvedenih izdelkov, visoke stopnje spretnosti ali umetniških izražanj. V proizvodnji se kakovost obravnava kot želen cilj, ki ga je mogoče doseči z upravljanjem proizvodnih procesov. Statistično vodenje kakovosti se tu že relativno dolgo uporablja za zagotavljanje skladnosti izdelkov z napovedovanjem

učinkovitosti proizvodnih procesov. Kakovost podatkov je precej težje opredeliti. Podatki namreč nimajo fizičnih značilnosti, ki bi omogočale enostavno oceno kakovosti. Kakovost podatkov je torej odvisna od neopredmetenih lastnosti, kot sta na primer popolnost in doslednost (Veregin, 1999).

Zaradi razširjenosti podatkov je njihova kakovost ključna v vseh poslovnih in vladnih aplikacijah. Rezultira v uspešnosti delovnih procesov, odločanju ter uspešnih medorganizacijskih sodelovanjih (Batini in drugi, 2009). Z pojavom masovnih podatkov so raziskovalci in odločevalci postopoma spoznali, da ima ta ogromna količina informacij številne prednosti – od boljšega razumevanja potreb strank do izboljšanja kakovosti storitev ter napovedovanja in preprečevanja tveganj. Le če so masovni podatki natančni in visoke kakovosti, lahko njihova uporaba in analiza prinesejo resnično dodano vrednost (Cai in Zhu, 2015).

Poleg procesa ETL, namenjenega tudi zagotavljanju kakovosti podatkov, so bile za pomoč pri določanju in zagotavljanju kakovosti podatkov razvite različne dimenzije kakovosti podatkov in metodologije za zagotavljanje kakovosti podatkov.

Tipičen proces za zagotavljanje kakovosti podatkov je po mnenju Askhama in drugih (2013):

- ugotavljanje, katere podatkovne elemente je treba oceniti glede na kakovost podatkov; običajno gre za tiste elemente, ki so ključni za poslovne procese in upravna poročila;
- določanje dimenzij kakovosti podatkov, ki jih je treba uporabiti, ter uteži, ki jim pripadajo;
- določanje vrednosti ali obsegov za posamezne dimenzije, ki predstavljajo dobro oziroma slabo kakovost podatkov; pri tem je treba upoštevati, da lahko en podatkovni nabor podpira več zahtev, zato je pogosto potrebnih več ocen kakovosti;
- uporaba meril ocenjevanja na izbranih podatkovnih elementih;
- pregled rezultatov in ugotavljanje, ali je kakovost podatkov sprejemljiva;
- izvedba korektivnih ukrepov, kjer je to potrebno, na primer čiščenje podatkov in izboljšanje postopkov ravnanja s podatki, da se preprečijo prihodnje težave;
- ponavljanje navedenih korakov na redni osnovi za spremljanje sprememb in trendov kakovosti podatkov.

4.1 Dimenzije kakovosti podatkov

Dimenzija kakovosti podatkov je priznan izraz, ki ga strokovnjaki za upravljanje podatkov uporabljajo za opis lastnosti podatkov, ki jo je mogoče meriti ali oceniti glede na določene standarde, z namenom ugotavljanja kakovosti podatkov. Dimenzije se lahko razlikujejo glede na poslovne potrebe in okolje v katerem podjetje posluje (Askham, in drugi, 2013).

V literaturi se v modelih pogosto pojavi 6 dimenzij, v posameznih modelih pa je teh dimenzij lahko veliko več. Šest dimenzij, ki jih definirajo Askham in drugi (2013), Scarisbrick-Hauser in Rouse (2007), Batini in drugi (2009) in Boufares in Salem (2012) je:

- **Celovitost** (angl. completness) opredeljuje stopnjo do katere so vrednosti prisotne v atributih, kjer so potrebne. Na primer, koliko odstotkov podatkov vsebuje vpisane vrednosti v teh atributih.
- **Konsistentnost** (angl. consistency) opisuje v kakšni meri so podatki zapisani v ustreznem formatu.
- **Edinstvenost** (angl. uniqueness) oziroma **podvajanje** (angl. duplication) pomeni, da se podatki podvajajo. Zaradi upravljanja in združevanja različnih virov pride do različnih pogledov na podatke, zato podvajanja podatkov ne želimo.
- **Veljavnost** (angl. validity) pomeni, da se podatki ujemajo s sintakso (oblika, tip, razpon). Na primer, atribut leto ima vrednost 9999. To je verjetno napačen podatek. Tudi atribut leto z navedbo imena učenca je verjetno napačen podatek.
- **Točnost** (angl. accuracy) je stopnja, do katere podatki pravilno opisujejo objekt ali dogodek v resničnem svetu, ki ga opisujejo.
- **Pravočasnost** (angl. timeliness) se nanaša na pričakovanje, da bo določeni podatkovni element ali več elementov na voljo ob zahtevanem ali določenem času.

Dimenzije, ki jih poleg zgoraj navedenih predstavijo Pipino in drugi (2002), so še: dostopnost, ustrezna količina podatkov, verodostojnost, varnost in dodana vrednost.

4.2 Metodologije za zagotavljanje kakovosti

Obstaja širok spekter tehnik za ocenjevanje in izboljšanje kakovosti podatkov kot so: povezovanje zapisov, poslovna pravila in merila podobnosti. Sčasoma so se te tehnike razvijale v spopadu z naraščajočo kompleksnostjo podatkov v omrežnih informacijskih sistemih. Zaradi raznolikosti in kompleksnosti teh tehnik se je raziskovanje v zadnjem času osredotočilo na opredelitev metodologij, ki pomagajo pri izbiri, prilagajanju in uporabi tehnik ocenjevanja in izboljšanja kakovosti podatkov (Batini in drugi, 2009).

Obstajajo številni okvirji in metodologije za zagotavljanje kakovosti podatkov, podatkovnega skladiščenja in informacij. Razvile so se tudi metodologije za posebne primere zagotavljanja podatkovne kakovosti, nekaj jih je prikazanih v tabeli 3. V nadaljevanju so podrobneje opisane tiste, ki so relevantne za to delo.

Tabela 3: Primeri metodologij za zagotavljanje kakovosti

Krajšava metodologije	**Celotni naziv metodologije**
DQA	Določilo kakovosti podatkov (angl. Data Quality Assessment)

se nadaljuje

Tabela 3: Primeri metodologij za zagotavljanje kakovosti (nad.)

Krajšava metodologije	**Celotni naziv metodologije**
TDQM	Celovito upravljanje kakovosti podatkov (angl. Total Data Quality Management)
DQAF	Okvir za oceno kakovosti podatkov (angl. Data Quality Assessment Framework)
HDQM	Metodologija za zagotavljanje kakovosti heterogenih podatkov (angl. Heterogeneous Data Quality Methodology)
AIMQ	Metodologija za zagotavljanje kakovosti informacij (angl. Methodology for Information Quality Assessment)
TIQM	Celovito upravljanje kakovosti informacij (angl. Total Information Quality Management)
DWQ	Metodologija za zagotavljanje kakovosti podatkovnega skladišča (angl. The Datawarehouse Quality Methodology)

Vir: lastno delo na podlagi Pipino in drugi (2002), Wang (1998), Sebastian-Coleman (2012), Batini in drugi (2011), Lee in drugi (2002), English (1999) in Jeusfeld in drugi (1998).

Pri posameznih metodologijah so upoštevane različne dimenzije za zagotavljanje kakovosti. Nabor dimenzij v metodologijah je razviden iz tabele 4.

Tabela 4: Upoštevane dimenzije za zagotavljanje kakovosti

Krajšava metodologije	**Dimenzije za zagotavljanje kakovosti**	**Fleksibilnost**
DQA	Dostopnost, pravilna količina podatkov, verodostojnost, celovitost, brez napak, konsistentnost, relevantnost, varnost, pravočasnost, razumljivost, dodana vrednost, reprezentativnost, lahka manipulacija	Ja
TDQM	Dostopnost, verodostojnost, celovitost, konsistentnost, dodana vrednost, lahka manipulacija, pravočasnost, razumljivost, brez napak, varnost, relevantnost, ustreznost	Ne
HDQM	Točnost, aktualnost	Ne
AIMQ	Dostopnost, ustreznost, verodostojnost, celovitost, konsistentnost, lahka manipulacija, brez napak, objektivnost, relevantnost, varnost, pravočasnost, razumljivost	Ne

se nadaljuje

Tabela 4: Upoštevane dimenzije za zagotavljanje kakovosti (nad.)

Krajšava metodologije	**Dimenzije za zagotavljanje kakovosti**	**Fleksibilnost**
TIQM	Konsistentnost, celovitost, točnost, preciznost, brez podvajanja, pravočasnost, uporabnost, prisotnost prekomernih podatkov	Ne
DWQ	Pravilnost, celovitost, minimalnost, sledljivost, razumljivost, dostopnost, pravočasnost, odzivnost, točnost, konsistentnost	Ne

Vir: lastno delo na podlagi Batini in drugi (2009) in Cichy in Stefan (2019).

4.2.1 Metodologija za določilo kakovosti podatkov

Metodologija DQA je bila zasnovana za zagotavljanje splošnih načel za opredelitev meril kakovosti podatkov. V literaturi so merila kakovosti podatkov večinoma opredeljena glede na specifične potrebe, za reševanje specifičnih problemov, in so tako odvisna od obravnavanega scenarija. Metodologija DQA je namenjena prepoznavanju splošnih načel merjenja kakovosti, ki so skupna prejšnjim raziskavam. Metodologija ločuje med subjektivnimi in objektivnimi merili kakovosti. Subjektivne metrike merijo zaznave, potrebe in izkušnje deležnikov. Objektivne metrike pa so razvrščene v od nalog neodvisne in odvisne. Prve ocenjujejo kakovost podatkov brez kontekstualnega znanja o aplikaciji, medtem ko so druge opredeljene za specifične kontekste aplikacij in vključujejo poslovna pravila, predpise podjetij in vlad, ter omejitve, ki jih zagotavlja administracija podatkovne baze. Obe metriki sta razdeljeni v tri razrede: Preprosto razmerje (število pričakovanih rezultatov proti številu vseh rezultatov), minimalna ali maksimalna vrednost (minimalna oziroma maksimalna normalizirana vrednost indikatorjev kakovosti podatkov) ter uteženo povprečje (določanje uteži, ki predstavljajo pomembnost dimenzij) (Batini in drugi, 2009).

4.2.2 Metodologija za celovito upravljanje kakovosti podatkov

Metodologija TDQM je bila prva splošna metodologija, objavljena v literaturi o kakovosti podatkov. TDQM je rezultat akademske raziskave, a je bila obsežno uporabljena kot vodilo pri organizacijskih pobudah za prenovitev podatkov (Batini in drugi, 2009).

Osnovni cilj TDQM je podatkovni kakovosti dodati načela celovitega upravljanja kakovosti (angl. Total Quality Management, v nadaljevanju TQM). TQM je preusmeril osredotočenost dejavnosti preoblikovanja iz učinkovitosti na uspešnost, s ponujanjem metodoloških smernic, namenjenih odpravi neskladij med izhodom operativnih procesov in zahtevami strank (Batini in drugi, 2009).

Skladno s temi načeli, TDQM za opis procesov proizvodnje informacij (angl. Information Production, v nadaljevanju IP) predlaga jezik imenovan IP-MAP. IP-MAP je bil prilagojen

po enovitem modelirnem jeziku (angl. Unified Modeling Language - UML) in tudi za podporo organizacijskemu oblikovanju. IP-MAP je edini jezik za modeliranje informacijskih procesov in predstavlja standard (Batini in drugi, 2009).

Cilj TDQM je podpreti celoten proces izboljšanja kakovosti od začetka do konca, od analize zahtev do izvedbe. Cikel TDQM sestavljajo štiri faze, ki izvajajo neprekinjen proces izboljšanja kakovosti: opredelitev, merjenje, analiza in izboljšanje (Batini in drugi, 2009).

V TDQM so opredeljene tudi vloge, odgovorne za različne faze procesa izboljšanja kakovosti. Obstajajo štiri vloge: dobavitelji informacij, ki ustvarjajo ali zbirajo podatke za IP, proizvajalci informacij, ki načrtujejo, razvijajo ali vzdržujejo podatke in pripadajočo sistemsko infrastrukturo, potrošniki informacij, ki uporabljajo podatke v svojem delu, in upravljavci informacijskih procesov, ki so odgovorni za upravljanje celotnega procesa proizvodnje informacij skozi življenjski cikel informacij (Batini in drugi, 2009).

TDQM je celovit tudi z vidika izvajanja, saj ponuja smernice za uporabo metodologije. Pri uporabi TDQM mora organizacija (Batini in drugi, 2009):

- jasno razumeti IP;
- vzpostaviti ekipo IP, ki jo sestavljajo višji izvršni direktor kot prvak TDQM, inženir IP, ki je seznanjen z metodologijo TDQM, in člani, ki so dobavitelji informacij, proizvajalci, potrošniki in upravljavci IP;
- vse deležnike IP poučiti o ocenjevanju kakovosti informacij (angl. Information Quality, v nadaljevanju IQ) in upravljanju IQ;
- institucionalizirati neprekinjeno izboljševanje IP.

Metodologija TDQM se opira na obstoječo literaturo o kakovosti informacij, iz katere črpa kriterije za ocenjevanje IQ ter tehnike za njeno izboljšanje. Povezuje vprašanja kakovosti z ustreznimi postopki in pristopi za izboljšave. Kljub temu pa novejša literatura ne navaja konkretnih industrijsko specifičnih tehnik niti ne ponuja podpore za prilagajanje splošnih pristopov izboljševanja kakovosti specifičnim kontekstom (Batini in drugi, 2009).

4.2.3 Metodologija za zagotavljanje kakovosti heterogenih podatkov

HDQM je metodologija visoke ravni in splošne domene, predlagana kot okvir, v skladu s katerim artikuliramo in uporabljamo znane tehnike in metode. Ni opredeljena kot zbirka novih dejavnosti, s katerimi določamo ali izboljšujemo podatke organizacije. Metodologija je namenjena določanju in izboljšanju kakovosti podatkov vseh oblik (Batini in drugi, 2011).

Implementacija HDQM v kompleksnih primerih je zelo zahtevna. Kvantitativne metrike zahtevajo precejšen napor, saj zahtevajo pristop temelječ na vrsticah in ne izkoriščajo statističnega pristopa. Za razliko od njih, se kvalitativne metrike zanašajo na razpoložljivost in predvsem strokovnost domenskih strokovnjakov, ki se odločajo o pomembnosti

posameznih virov in kako ti vplivajo na celotni informacijski sistem. Poleg tega določajo tudi pragove, ki razvrščajo posamezne procese in določajo prioritete. Za HDQM je potrebna rekonstrukcija trenutnega stanja v podjetju, ki pa je časovno potratna in draga. V primerih, ko viri niso dokumentirani ali dobro poznani, se to pozna še toliko bolj. Je pa HDQM zelo modularna pri upoštevanju posameznih virov in jo je moč uporabiti zgolj za dele celotnega nabora virov, ob predpostavki, da bi bil celoten nabor prevelik strošek za podjetje (Batini in drugi, 2011).

5 ANALIZA PRIMERA

Integracijo podatkovnih virov v BI okolje za analizo digitalnega trženja bom predstavil na poslovnem primeru iz podjetja B2BI d.o.o., v katerem sem zaposlen kot podatkovni analitik. V podjetju želimo sprejemati poslovne odločitve tudi na podlagi podatkov digitalnega trženja. Zaenkrat še nimamo rešitve, ki bi omogočila spremljanje teh podatkov na enem mestu.

Podatki, povezani z digitalnim trženjem, se nahajajo na različnih lokacijah, pri čemer ima vsaka lokacija svoje analitično področje za obdelavo in analitiko podatkov. Podatki, povezani s spletno stranjo podjetja in podatki o oglaševanju na Googlu, so shranjeni v analitičnem orodju GA4. Podatki o oglaševanju na LinkedInu so shranjeni neposredno na tej platformi, podatki o e-poštnih kampanjah pa se nahajajo v CRM sistemu podjetja.

Na osnovi analize opisanega poslovnega primera v nadaljevanju podajam priporočila za zagotavljanje kakovosti podatkov pri integraciji iz več podatkovnih virov. Izdelal sem tudi nadzorno ploščo, katere ključna sestavina so kakovostni podatki in je namenjena pomoči pri sprejemanju odločitev na področju digitalnega trženja.

5.1 Podatkovni viri

Podatke za izdelavo rešitve sem pridobil iz štirih različnih podatkovnih virov, prikazanih na sliki 10: analitičnega orodja GA4, ki vsebuje podatke iz spletne strani podjetja B2BI integrirane s podatki orodja za oglase Google Ads, sistema CRM podjetja B2BI in družbenega omrežja LinkedIn. Vsi viri, z izjemo CRM-sistema, predstavljajo zunanje vire, kar pomeni, da nanje podjetje nima neposrednega vpliva, saj so podatki dostopni v obliki, ki jo določijo ponudniki storitev. Viri so tudi heterogeni, saj izhajajo iz različnih platform in tehnoloških okolij. Čeprav se lahko na prvi pogled zdijo podatki podobni, so pogosto definirani na različne načine, kar lahko povzroči neskladja pri nadaljnji obdelavi ali integraciji.

Slika 10: Uporabljeni podatkovni viri

Vir: lastno delo.

5.1.1 Google Analytics in Google Ads

Google Analytics je glavna analitična platforma podjetja Google. Platforma Google Analytics Universal (GAU) je 1. julija 2023 prenehala obdelovati podatke, zamenjala pa jo je novejša platforma, Google Analytics 4 (GA4). Z zbiranjem podatkov s spletnih strani in aplikacij omogoča uporabnikom boljše razumevanje strank. Podatki se generirajo na nivoju dogodkov namesto na nivoju sej, kar pomeni, da stara analitika ni primerljiva z novejšo (Google, brez datuma d). GA4 na podlagi strojnega učenja nudi tudi napovedno analitiko v obliki metrik kot so verjetnost nakupa, verjetnost izpada in predviden prihodek (McGuirk, 2023).

Google Ads je spletni oglaševalski program podjetja Google. Omogoča ustvarjanje plačljivih spletnih oglasov. Oglasi se uporabljajo za promocijo spletne strani podjetja, izdelkov ali storitev (Google, brez datuma č).

GA4 omogoča povezavo Google Ads računa z računom GA4, kar omogoča avtomatsko integracijo podatkov iz okolja Google Ads v okolje GA4. Za povezavo potrebujemo GA4 račun z ustreznimi pravicami urejevalca in Google Ads račun s pravicami administratorja. Edina omejitev pri povezavi računov je, da lahko na en GA4 račun povežemo največ 400 Google Ads računov, kar v večini podjetij ne bi smelo predstavljati omejitve (Google, brez datuma b).

GA4 omogoča izvoz podatkov v obliki statičnega analitičnega poročila, ki ga lahko uporabnik oblikuje po lastnih željah ali pa uporabi vnaprej pripravljeno poročilo. Poročila so omejena z atributi, kar pomeni, da lahko npr. na poročilu za spremljanje pridobivanja strank spremljamo le medij, s pomočjo katerega smo pridobili nove stranke, stopnjo angažiranosti (angl. engagement rate) strank, konverzije ipd. Ne moremo pa spremljati dogodkov na spletni strani podjetja, ki so na voljo v drugem poročilu. Poročilo se lahko omeji po datumu in izvozi bodisi v formatu, ki je neodvisen od računalniškega okolja in medoperacijsko prenosljiv (angl. Portable Document Format - PDF) bodisi kot besedilna datoteka, ki vsebuje z vejico ločene vrednosti (angl. Comma-seperated values, v nadaljevanju CSV) (Google, brez datuma c).

GA4 omogoča tudi avtomatski dnevni izvoz vseh dogodkov s pomočjo povezave z BigQuery računom, kar pomeni, da lahko uporabnik izven platforme generira poljubna poročila. Prav zaradi avtomatizacije izvoza, ki ga omogoča BigQuery, in kreiranja dinamičnih GA4 poročil izven platforme, sem se v sklopu primera odločil za izvažanje s pomočjo povezave GA4 z oblačnim DWH-jem BigQuery. Primer izvoza tabele v BigQuery je prikazan v prilogi, struktura, ki sem jo uporabil za analizo pa v tabeli 5.

Tabela 5: Struktura GA4 omejene tabele izvožene v BigQuery

Ime stolpca	**Podatkovni tip**	**Opis vrednosti**
event_date	Besedilo	Datum, ko je bil dogodek zabeležen (format LLLLMMDD v registriranem časovnem pasu aplikacije).
event_name	Besedilo	Ime dogodka (prvi obisk, začetek seje, ogled strani).
event_params.key	Besedilo	Identifikator parametra dogodka.
event_params.int_value	Celo število	Če je parameter dogodka predstavljen s celim številom, je to polje izpolnjeno.
event_params.float_value	Decimalno število	Če je parameter dogodka predstavljen z decimalnim številom, je to polje izpolnjeno.
event_params.double_value	Decimalno število	Če je parameter dogodka predstavljen z decimalnim številom z dvojno preciznostjo, je to polje izpolnjeno.
user_id	Besedilo	Edinstven identifikator, dodeljen uporabniku.
user_pseudo_id	Besedilo	Psevdonimni identifikator (npr. identifikator instance aplikacije) za uporabnika.
collected_traffic_source.manual_campaign_id	Besedilo	Ročno določena identifikacijska številka kampanje, ki je bila zabeležena z dogodkom.
collected_traffic_source.manual_medium	Besedilo	Ročno določen medij kampanje, ki je bil zabeležen z dogodkom

se nadaljuje

Tabela 5: Struktura GA4 omejene tabele izvožene v BigQuery (nad.)

Ime stolpca	**Podatkovni tip**	**Opis vrednosti**
collected_traffic_source.manual_source	Besedilo	Ročno določen vir kampanje, ki je bil zabeležen z dogodkom

Vir: lastno delo.

5.1.2 LinkedIn

LinkedIn je družbeni medij ustvarjen z namenom povezovanja uporabnikov in oblikovanja kariernih mrež. Kot trženjsko orodje je LinkedIn odličen za velika B2B podjetja, ki želijo promovirati svoje blagovne znamke in doseči strokovnjake preko oglaševalskih kampanj. Spletna stran poslovnežem omogoča opravljanje različnih nalog kot so: objavljanje promocijskih videoposnetkov in objav, prilagajanje podjetniškega profila, sodelovanje v aktivnih klepetih s strankami in ustvarjanje razpisov za delo (Jeniferin drugi, 2023).

LinkedIn omogoča ročni izvoz podatkov skrbniku strani, torej LinkedIn profilu, ki je LinkedIn stran ustvaril. V izbranem primeru je bila uporabljena LinkedIn stran podjetja B2BI, ki se uporablja za promoviranje video vsebin, dogodkov, zaposlovanje kadra … Podatke za analizo sem pridobil na zavihku »Analytics« in izbral kategorijo »Content« (LinkedIn, brez datuma). Pri izvozu je možno izbirati datumsko obdobje do enega leta. Izvoz podatkov je mogoč v obliki Excel datoteke, ki vsebuje dva lista. List »Metrics« je namenjen podrobnejši analizi, podatki so za izbrano obdobje agregirani na dan natančno, struktura je vidna v tabeli 6. List »All posts« je namenjen analizi objav, podatki so agregirani glede na izbrano obdobje. Za dani primer sem uporabil samo podatke iz lista »Metrics«. Za analize po časovnih obdobjih je priporočljivo shranjevanje podatkov v DWH, več o tem kasneje v delu. LinkedIn ponuja tudi možnost avtomatiziranega prenosa podatkov s pomočjo ključa programskega vmesnika (angl. Application Programming Interface, v nadaljevanju API) (Lambert, 2023).

Tabela 6: Struktura LinkedIn izvoza objav, lista »Metrics«

Ime stolpca	**Podatkovni tip**	**Opis vrednosti**
Date	Datum	Datumsko polje v formatu MM/DD/LLLL
Impressions (organic)	Celo število	Število organskih ogledov objav na datum
Impressions (total)	Celo število	Vsota plačanih in organskih ogledov objav na datum

se nadaljuje

Tabela 6: Struktura LinkedIn izvoza objav, lista »Metrics« (nad.)

Ime stolpca	**Podatkovni tip**	**Opis vrednosti**
Impressions (sponsored)	Celo število	Število ogledov objav, ki so bili posledica plačanega oglaševanja na datum
Unique impressions (organic)	Celo število	Število unikatnih organskih ogledov objav na datum
Unique impressions (sponsored)	Celo število	Število unikatnih ogledov objav, ki so bili posledica plačanega oglaševanja na datum
Unique impressions (total)	Celo število	Vsota unikatnih plačanih in unikatnih organskih ogledov objav na datum
Clicks (organic)	Celo število	Število organskih klikov na objave
Clicks (sponsored)	Celo število	Število klikov na objave, ki so bili posledica plačanega oglaševanja na datum
Clicks (total)	Celo število	Vsota plačanih in organskih klikov na objave na datum
Reactions (organic)	Celo število	Število organskih reakcij na objave na datum
Reactions (sponsored)	Celo število	Število reakcij na objave, ki so bili posledica plačanega oglaševanja na datum
Reactions (total)	Celo število	Vsota plačanih in organskih reakcij na objave na datum
Comments (organic)	Celo število	Število organskih komentarjev na objave na datum
Comments (total)	Celo število	Vsota plačanih in organskih komentarjev na objave na datum
Reposts (organic)	Celo število	Število organskih deljenj objav na datum
Reposts (sponsored)	Celo število	Število deljenj objav, ki so bili posledica plačanega oglaševanja na datum
Reposts (total)	Celo število	Vsota plačanih in organskih deljenj objav na datum
Enagagement rate (organic)	Decimalno število	Stopnja organske angažiransti (vsota organskih klikov, komentarjev in delitev deljenih z številom organskih ogledov)

se nadaljuje

Tabela 6: Struktura LinkedIn izvoza objav, lista »Metrics« (nad.)

Ime stolpca	**Podatkovni tip**	**Opis vrednosti**
Enagagement rate (sponsored)	Decimalno število	Vsota sponzorirane angažiranosti (vsota klikov, komentarjev in delitev, ki so bili posledica plačanega oglaševanja deljenih z številom ogledov, ki so bili posledica plačanega oglaševanja)
Enagagement rate (total)	Decimalno število	Vsota stopnje organske angažiranosti in sponzorirane stopnje angažiranosti

Vir: lastno delo.

5.1.3 Sistem za upravljanje odnosov s strankami

Microsoft Dynamics 365 je nabor inteligentnih poslovnih aplikacij, ki zagotavlja vrhunsko operativno učinkovitost in izjemne uporabniške izkušnje. Podjetjem omogoča, da postanejo agilnejša in zmanjšajo kompleksnost brez povečanja stroškov (Microsoft, brez datuma).

V naboru aplikacij obstaja tudi CRM oziroma sistem za upravljanje odnosov s strankami. CRM je niz integriranih, podatkovno vodenih programskih rešitev, ki pomagajo upravljati, slediti in shranjevati informacije, povezane z obstoječimi in potencialnimi strankami podjetja. Znotraj sistema CRM obstaja več modulov: prodaja, storitve za stranke, terenske storitve, trženje, avtomatizacija projektnih storitev in vpogledi strank. V sklopu magistrskega dela je relevanten modul za trženje oziroma, po novem, Dynamics 365 Customer Insights – Journeys (v nadaljevanju modul za trženje). Modul je zelo napreden in omogoča ustvarjanje poti strank (angl. customer journey), izboljšanje komunikacije in avtomatizacijo poslovnih procesov. Vključuje ključne funkcije, kot so: potencialne stranke, računi, avtomatizacija kampanj, e-pošta, obrazci, ankete, SMS, integracija družbenih medijev, prilagojeni delovni tokovi, avtomatizacije in druge (Mohammadi, 2023).

Modul za trženje razpolaga z različnimi načini izvoza podatkov. Najpogostejši in najenostavnejši je izvoz podatkov v obliki CSV datoteke. Možen je tudi izvoz s pomočjo storitve Microsoft Dynamics 365-Data Export Service, ki sicer omogoča avtomatski izvoz v Microsoft Azure okolje, a zahteva veljavno Azure licenco (365blog, 2017). Izbral sem izvoz CSV datoteke, ker je enostaven in ne zahteva Azure licence. Podatke lahko izvozimo, če imamo dostop do trženjskega modula. Znotraj modula se podatki nahajajo na podstrani vpogledi (angl. insights). Lahko izbiramo med različnimi kategorijami, izbral sem e-poštne podatke. Po izbiri kategorije se nam pokaže gumb za izvoz. CSV struktura izvoženih podatkov je prikazana v tabeli 7.

Tabela 7: Struktura izvoza tabele e-pošte iz modula za trženje

Ime stolpca	**Podatkovni tip**	**Opis vrednosti**
Ime stika	Besedilo	Ime in priimek stranke
Uporabljeni e-poštni naslov	Besedilo	Kontaktni e-naslov
Kategorija nesprejetja	Besedilo	Kategorija nesprejetja e-pošte
Razlog za nezmožnost dostave e-pošte	Besedilo	Razlog, ki je povzročil nezmožnost dostave e-pošte
Časovni žig	Datum in čas	Datum in pošiljanja e-pošte
Časovni žig odprtja	Datum in čas	Datum in čas odprtja e-pošte
Naslov	Besedilo	Poimenovanje e-pošte (novica, vabilo)

Vir: lastno delo.

5.2 Podatkovno skladišče

Zaradi lažje transformacije, hranjenja in dostopnosti podatkov, sem v sklopu naloge podatke naložil v podatkovno skladišče. Čeprav Foote (2023) trdi, da je priporočljivo podatke shranjevati na eno mesto oziroma DWH, sem zaradi avtomatične integracije podatkov iz GA4, ki jo ponuja BigQuery, uporabil dva DWH-ja. Drugi razlog za uporabo dveh DWH-jev pa je pomanjkanje pravic na strežniku, ki bi mi omogočale integracijo podatkov iz enega v drugega. En DWH je tradicionalen drugi pa DWH v oblaku – BigQuery.

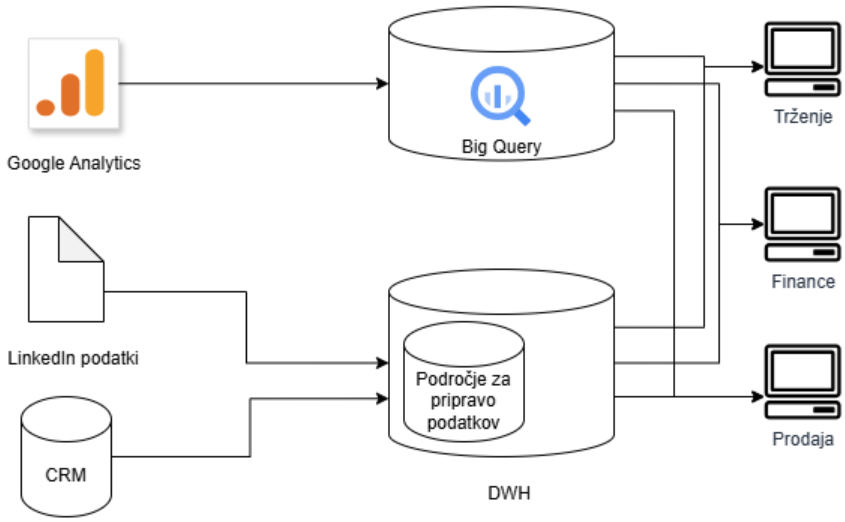
Tradicionalni DWH je postavljen na lokalnem strežniku in je namenjen shranjevanju, transformaciji in združevanju podatkov. V njem so poleg drugih podatkov shranjeni tudi podatki iz CRM sistema in LinkedIn-a. V sklopu DWH obstaja tudi področje za pripravo podatkov, namenjeno transformaciji podatkov. Podatki iz različnih virov v obliki tabel se najprej prečrpajo v področje za pripravo podatkov, kjer se preoblikujejo, nato pa se shranijo v obliko pogledov (angl. view), kar omogoča prihranek fizičnega prostora na strežniku.

Drugi DWH se nahaja v oblaku Google Cloud, katerega lastnik je podjetje Google. Google BigQuery je oblačna DWH storitev, znana po svoji preprostosti in funkcionalnosti. Odličen je za podjetja, ki si ne morejo privoščiti naložb v infrastrukturo za obdelavo velikih količin podatkov. Ta platforma omogoča shranjevanje in pridobivanje velikih količin podatkov v skoraj realnem času, pri čemer je osredotočena na analizo podatkov (Sharma in Kumar, 2022). BigQuery sem uporabil za hranjenje podatkov iz GA4, ker omogoča avtomatičen prenos podatkov iz GA4. V platformi nisem kreiral področja za pripravo podatkov in jih preoblikoval, ker je preoblikovanje podatkov zahtevnejše. Za transformacijo sem uporabil orodje Power BI, ki omogoča grafični vmesnik za preoblikovanje podatkov in povezavo do BigQuery-ja. Platforma BigQuery je lahko brezplačna storitev, če ne naložimo več kot 10 gigabajtov podatkov na mesec in ne procesiramo več kot 1 terabajta podatkov mesečno.

Dodaten razlog za izbiro platforme BigQuery je bilo dejstvo, da sem jo uporabil samo za shranjevanje podatkov digitalnega trženja, ki niso obsežni.

Uporabljena arhitektura DWH, prikazana na sliki 11, tako ni standardna. Najbolj je podobna arhitekturi organizacijskega podatkovnega skladišča, ki ga omenjajo Hajmoosaei in drugi (2011), z dodatkom, da se uporabljata dva DWH-ja in tudi področje za pripravo podatkov.

Slika 11: Izbrana arhitektura podatkovnega skladišča v primeru



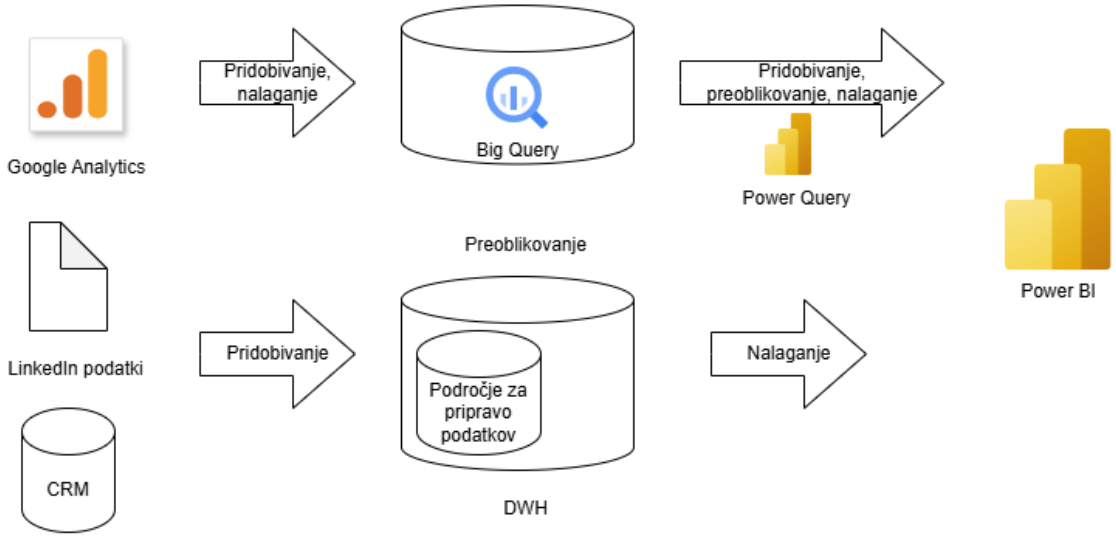
Vir: lastno delo.

5.3 Proces pridobivanja, preoblikovanja in nalaganja podatkov

V sklopu naloge sem za pridobivanje, preoblikovanje in nalaganje podatkov ter zagotavljanje njihove kakovosti uporabil proces ETL. Za podatkovna vira CRM in LinkedIn sem uporabil klasičen proces, za podatke GA4 pa prirejen EL-ETL postopek. Dvojni ETL oziroma EL-ETL sem izbral, ker je preoblikovanje podatkov v storitvi BigQuery lahko plačljivo, če se izvaja v večjih količinah. Drugi razlog za izbor je bilo dejstvo, da orodje Power BI vsebuje prilagojen priključek (angl. connector) za nalaganje podatkov iz BigQuery-ja in brezplačno orodje za preoblikovanje in integracijo podatkov PowerQuery. PowerQuery ponuja grafični vmesnik, ki je prijazen za začetnike in napredne uporabnike.

Uporabnik ne potrebuje naprednega znanja strukturiranega povpraševalnega jezika (angl. Structured Query Language, v nadaljevanju SQL) za preoblikovanje in integracijo podatkov. Opisani proces ETL in EL-ETL je prikazan na sliki 12. Poleg orodja Power BI, sem v sklopu procesa ETL uporabil tudi druga orodja ETL, ki jih bom predstavil v posameznih korakih procesa.

Slika 12: Uporabljena procesa pridobivanja, preoblikovanja in nalaganja podatkov



Vir: lastno delo.

5.3.1 Pridobivanje podatkov

Pridobivanje podatkov je že predstavljeno v podpoglavju 5.1, zato je fokus tega podpoglavja prenos podatkov v DWH oziroma Power BI, kar je prav tako del procesa pridobivanja podatkov.

5.3.1.1 Google Analytics 4 in Google Ads

Integrirani podatki GA4 in Google Ads se v BigQuery dnevno prenašajo v obliki tabel, za vsak dan se generira nova tabela s podatki za tisti dan in imenom, ki ustreza datumu nastanka. GA4 ponuja tudi možnost prenosa takoj, ko se zgodi dogodek na spletni strani, če imamo na projektu v GA4 omogočeno plačilno sredstvo. Podatkovne vrednosti, ki se prenesejo v BigQuery, se lahko v 2 dneh še spremenijo, na dan pa lahko zastonj v oblak prenesemo do milijon vrstic podatkov (Google, brez datuma a). Na BigQuery se nato povežemo s pomočjo orodja Power BI, ki ga bom predstavil v nadaljevanju. Znotraj orodja je vtičnik, ki omogoča branje oziroma pridobivanje podatkov direktno iz BigQuery-ja.

5.3.1.2 LinkedIn

Po pridobitvi Excel datoteke s podatki, lahko podatke v podatkovno skladišče uvozimo na več načinov. Za primer sem izbral najenostavnejši, ne-avtomatiziran način ročnega uvoza. Ker je baza podatkov oziroma DWH postavljen na SQL strežniku, sem za uvoz podatkov in preoblikovanje podatkov uporabil integrirano okolje SQL Server Management Studio (v nadaljevanju SSMS). To okolje nudi orodja za konfiguracijo, nadzor in administracijo

primerkov SQL strežnika in baz podatkov (Microsoft, 2023a). Za uvoz podatkov je na voljo čarovnik s katerim sem si pomagal pri uvozu. Najprej sem izbral vir podatkov, v mojem primeru Excel datoteko, ki sem jo shranil lokalno. Sledila je destinacija, v mojem primeru področje priprave podatkov. Nazadnje sem preveril mapiranje izvornih in destinacijskih stolpcev in podatkovne tipe stolpcev. Nalaganje podatkov se lahko tudi avtomatizira s pomočjo SSIS paketov ali API-ja, v kombinaciji z Python programskim jezikom.

5.3.1.3 Sistem za upravljanje odnosa s strankami

Proces in načini uvoza CSV datoteke iz CRM sistema so enaki kot pri uvozu Excel datoteke oziroma LinkedIn podatkov. Ponovno sem izbral najenostavnejši način, torej uvoz s pomočjo orodja SSMS. Datoteko CSV v DWH sem uvozil s pomočjo čarovnika za uvoz nepovezanih datotek (angl. flat file). Izbral sem mesto, kjer je datoteka shranjena ter poimenovanje tabele in sheme v DWH. Zatem je bil na voljo predogled podatkov in strukture, ki jo je čarovnik prepozna avtomatsko. V naslednjem koraku sem lahko preimenoval stolpce in določil tip podatkov za posamezne stolpce. V zadnjem koraku sem samo potrdil vse izbire in podatki so se prenesli v DWH.

5.3.2 Preoblikovanje podatkov

Preoblikovanje oziroma transformacija podatkov, ki je ključni korak ETL za zagotavljanje kakovosti podatkov, sem izvedel s pomočjo že predstavljenega orodja SSMS in orodjem Power BI, ki ga predstavljam v naslednjem podpoglavju. Podatke, ki se nahajajo v DWH, sem preoblikoval s pomočjo orodja SSMS, podatke iz BigQuery-ja pa z orodjem Power BI. Dve različni orodji sem izbral zaradi poenostavitve samega procesa. Podatke v DWH sem preoblikoval s pomočjo ustvarjanja pogledov. Kot vodilo za preoblikovanje podatkov sem uporabil dimenzije za zagotavljanje kakovosti podatkov.

V sklopu preoblikovanja podatkov sem podatke najprej analiziral. V primeru GA4 sem to storil s pomočjo dokumentacije, s katero razpolaga spletno mesto GA4 in vsebuje metapodatke o pridobljenih podatkov. Tako sem se lahko odločil katere podatke pravzaprav potrebujem za analitično poročilo in katere lahko odstranim. Podatki iz LinkedIna in CRM-ja so bili sami po sebi dovolj govoreči, zato dodatne dokumentacije nisem potreboval. CRM sicer v sklopu modula ponuja tudi orodje za metapodatke, ki ga v tem primeru nisem potreboval.

5.3.2.1 Orodje Microsoft Power BI

Power BI je zbirka programske opreme, aplikacij in povezav, ki skupaj pretvorijo neusklajene vire podatkov v koherentne, vizualno privlačne in interaktivne vpoglede. Podatki so lahko v Excelovi preglednici, zbirki podatkovnih skladišč v oblaku, lahko so tradicionalna podatkovna skladišča (angl. on-premises), itd. Power BI omogoča enostavno

povezavo z viri podatkov, integracijo in preoblikovanje podatkov ter vizualizacijo in analizo. Omogoča tudi deljenje vpogledov z vsakim in vsemi, ki bi jih to zanimalo (Microsoft, 2023b).

Čeprav orodje Microsoft Power BI ni uvrščeno na Brinkerjev zemljevid, ga je Gartner tudi leta 2023 uvrstil v sam vrh, kot je razvidno iz slike 13.

Slika 13: Gartnerjev magični kvadrant analitičnih in poslovnointeligenčnih platform

Vir: lastno delo na podlagi Schlegel in Sun (2023).

5.3.2.2 Google Analytics 4 in Google Ads

Podatke GA4 sem preoblikoval s pomočjo orodja Power Query, ki se nahaja znotraj orodja Power BI. Za orodje sem se odločil zaradi osebnih preferenc, grafičnega vmesnika za preoblikovanje in profiliranje podatkov in zmožnosti povezave na BigQuery. Orodje omogoča tudi ponovitev preoblikovanj znotraj posameznih virov in podpisovanje preoblikovanih podatkov, kar je v tem primeru zelo koristno. Podatki v BigQuery-ju na prvi pogled izgledajo, kot bi bili strukturirani v klasične tabele. Šele v orodju Power BI ugotovimo, da tabele niso klasične in vsebujejo stolpce, v katerih se nahajajo podatki v standardnem formatu za pomnjenje in izmenjavo podatkov v tekstovni obliki (angl. JavaScript Object Notation - JSON) oziroma seznami podatkov.

Prvi korak preoblikovanja podatkov je bil zato strukturiranje tabel v primerno obliko. Uporabil sem tri preoblikovalne funkcije, s pomočjo katerih sem vrstice razdružil in

spremenil v posamezne stolpce. S pomočjo filtriranja stolpca, ki vsebuje informacijo o tipu dogodka, sem odstranil vse prazne vrednosti. Dodal sem stolpec z indeksi, ki mi bo kasneje služil kot pomoč pri odkrivanju duplikatov. Odstranil sem odvečne stolpce in obdržal stolpce, ki bodo služili analizi. Stolpcem sem določil pravilne podatkovne tipe. Odstranil sem duplikate. Ustrezno število podatkov sem potrdil s pomočjo analitične platforme GA4, kjer sem primerjal število dogodkov v izbranem časovnem obdobju.

Za ustrezno integracijo z drugima viroma sem moral podatke dodatno preoblikovati. Filtriral sem jih samo na vire, ki vsebujejo vrednost »Linkedin« ali »lnkd.in« ali »newsletter« in »dyncrm«, da sem dobil samo podatke za primerjavo s preostalima viroma. Za primerjavo podatkov z LinkedInom in CRM-jem sem ustvaril nov pogojni stolpec, ki je vsem vrsticam, ki so vsebovale vrednost »Linkedin« ali »lnkd.in«, priredil vrednost »LinkedIn«, drugim pa »email«. Odstranil sem prvotni stolpec za vir. Tabelo sem grupiral po datumu in viru. Ostale stolpce sem ustrezno agregiral, da so služili pripravi izračunov.

5.3.2.3 LinkedIn

Podatke z družbenega omrežja LinkedIn sem preoblikoval s pomočjo orodja SSMS. Preoblikovanje sem izvedel s pomočjo kreacije pogledov, se pravi kode, ki navidezno preoblikuje prvotne podatke in vrne preoblikovane. Prvo preoblikovanje se je izvedlo že ob nalaganju podatkov v DWH, kjer čarovnik za nalaganje podatkov iz Excel datotek omogoča izbiro podatkovnega tipa posameznih stolpcev. Pravilno je določil tip vseh stolpcev razen datumskega, saj oblike »mesec/dan/leto« orodje ni prepoznalo kot datuma ampak kot besedilo. Tip stolpca sem spremenil s pomočjo funkcije CONVERT, ki pretvori en tip stolpca v drugega s pomočjo zapisa ustreznih parametrov. Za odstranjevanje duplikatov sem uporabil funkcijo DISTINCT, ki izbere samo unikatne vrstice. Ker so podatki agregirani na nivoju dneva, je bila za vsak dan izbrana samo ena vrstica. S pomočjo klavzule WHERE, ki omeji izbor podatkov na neničelne vrednosti, sem dodal tudi pogoj, da vrstica z datumom ni prazna. Za potrebe integracije podatkov sem ustvaril stolpec »vir« in mu določil vrednost »LinkedIn«. Ustreznost podatkov sem preveril s pomočjo analitičnega okolja, ki ga ponuja LinkedIn.

5.3.2.4 Sistem za upravljanje odnosa s strankami

Tudi podatke iz CRM-ja sem preoblikoval s pomočjo orodja SSMS. Ponovno je bil prvi korak preoblikovanja opravljen v čarovniku, ki je v tem primeru pravilno prepoznal vse podatkovne tipe stolpcev. Ponovno sem za odstranjevanje duplikatov uporabil funkcijo DISTINCT. Prazne vrstice sem odstranil z dodatnim pogojem, da vrstica s časovnim žigom (angl. timestamp) ni prazna in klavzulo WHERE. Stolpca časovni žig in časovni žig odprtja sem pretvoril v datum. Dodal sem tudi štiri nove stolpce. Prvega z vrednostjo 1, ki je služil štetju poslanih e-poštnih sporočil, drugi in tretji stolpec sta bila pogojna in sta služila štetju odprtih in neuspešno poslanih e-poštnih sporočil, četrtega pa z vrednostjo »email« in je služil

integraciji podatkov. V drugem stolpcu se je vrstici določila vrednost 1, če je bil v vrstici stolpca »razlog za nezmožnost dostave e-pošte« katerakoli vrednost, v nasprotnem primeru pa se je določila vrednost 0. V tretjem stolpcu se je vrstici določila vrednost 1, če vrstica v stolpcu »časovni žig odprtja« ni bil prazen, drugače se je določila vrednost 0. Ustreznost podatkov sem preveril s pomočjo analitičnega okolja, ki ga ponuja CRM.

Za ustrezno integracijo z GA4 podatki, sem moral podatke še agregirati. To sem naredil tako, da sem iz poizvedbe iz prejšnjega odstavka grupiral datumska stolpca in seštel vrednosti v pogojnih stolpcih.

5.3.3 Integracija podatkov

Integracija podatko**v** je proces, v katerem se podatki iz različnih virov združijo v eno tabelo ali več povezanih tabel, kar omogoča primerjavo podatkov iz različnih področij na osnovi skupnih lastnosti. Za izvedbo integracije podatkov sem uporabil orodje PowerQuery, saj omogoča enotno obdelavo podatkov pred nalaganjem v končno destinacijo. Pred začetkom integracije sem v PowerQuery naložil podatke iz CRM-ja in LinkedIn-a. Podatke sem pridobil preko SQL strežnik priključka, ki omogoča povezavo s SQL strežnikom in izbor želenih tabel za prenos. Za CRM podatke sem ustvaril dve poizvedbi – eno za datum pošiljanja in drugo za datum odprtja. V vsaki poizvedbi sem ohranil en datum, po katerem sem grupiral podatke, pri tem pa sem agregiral stolpce, kot so poslana, zavrnjena in odprta e-poštna sporočila. Zaradi različne količine podatkov v posameznih tabelah sem moral ustvariti začasno datumsko tabelo. To sem storil s poizvedbo, ki sem jo nato povezoval z ostalimi tabelami preko levega zunanjega stika (angl. left outer join). Ohranil sem datumski stolpec in stolpec »vir«, ki sem ju uporabil za ustvarjanje novega stolpca ključ, ki je združeval datum in vir. Nato sem odstranil podvojene zapise v tem stolpcu. Na kreirano poizvedbo sem pridružil poizvedbo z GA4 podatki s pomočjo enake funkcije za združevanje, na osnovi stolpcev »datum« in »vir« iz obeh tabel. Enako sem storil tudi za obe CRM poizvedbi in LinkedIn poizvedbo. Na koncu sem izbral vse relevantne stolpce iz združevalnega procesa.

5.3.4 Nalaganje podatkov

Nalaganje podatkov v ciljno okolje je zadnji korak procesa ETL. Preoblikovane podatke sem naložil v orodje Power BI. Preoblikovani podatki se po potrditvi in shranitvi preoblikovalnega procesa v PowerQuery-ju, naložijo v orodje. Nalaganje podatkov lahko traja dalj časa, če so funkcije za preoblikovanje podatkov zahtevne ali je podatkov veliko.

5.4 Analitično poročilo

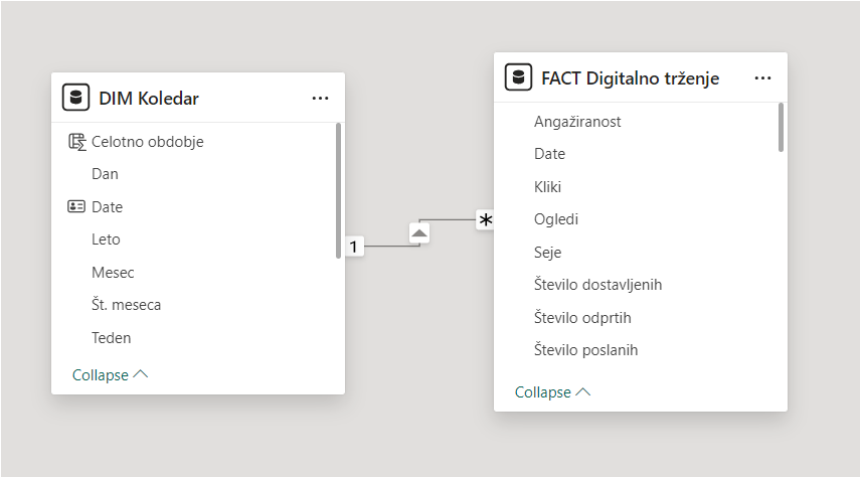
Analitično poročilo je vizualni prikaz podatkov, ki temelji na predhodno očiščenih in obdelanih podatkih. Njegov namen je podpora pri sprejemanju poslovnih odločitev na

podlagi analitičnih vpogledov. Poročilo je sestavljeno iz modelirnega dela, ki vključuje pripravo in analizo podatkov, ter vizualnega dela, kjer so rezultati prikazani v obliki tabel, grafikonov in drugih vizualizacij.

V modelirnem delu se nahajajo tabele s podatki in povezave med posameznimi tabelami. V mojem primeru sta to tabeli s podatki o digitalnem trženju in koledarska tabela, prikazani na sliki 14. Povezave med tabelami omogočajo filtracijo ali grupacijo podatkov na vizualizacijah. Povezal sem koledarsko tabelo s tabelo, ki vsebuje integrirane podatke o digitalnem trženju, kar mi je omogočilo prikaz podatkov po različnih časovnih obdobjih in izbor časovnih obdobij.

Model vključuje tudi metrike oziroma izračune, ki se uporabljajo na vizualizacijah za prikaz ključnih vrednosti. Priprava metrik za merjenje uspešnosti oziroma vpliva je temeljila na integriranih podatkih iz uporabljenih virov. Na podlagi teh sem lahko analiziral vpliv e-poštnega trženja in objav na platformi LinkedIn na obisk spletne strani podjetja. Metrike sem zasnoval s pomočjo dokumentacije iz GA4 in LinkedIn ter na podlagi specifičnih potreb uporabnikov poročila. Poleg osnovnih izračunov za izbrano časovno obdobje sem pripravil tudi primerjalne metrike za pretekla obdobja, kar omogoča vpogled v trende in spremembe skozi čas.

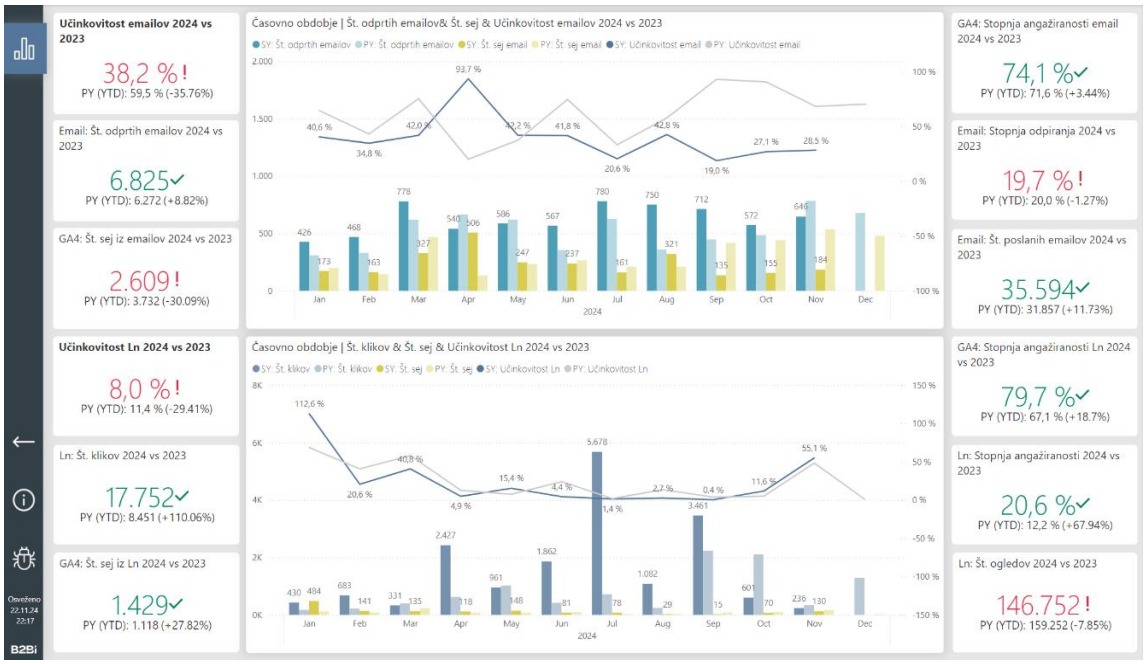
Slika 14: Model analitičnega poročila



Vir: lastno delo.

Na vizualnem delu se nahajajo različni tipi vizualizacij kot so stolpčni grafikoni, palični grafikoni, kartice, KPI-ji in druge vizualizacije. Na razpolago so tudi filtri, ki jih lahko določimo vnaprej ali prepustimo uporabniku, da jih nastavi sam. Za dani primer sem izbral KPI vizualizacije in stolpčni grafikon v kombinaciji s črtnim grafikonom. V vse vizualizacije sem dodal izračune za izbrano obdobje v primerjavi z preteklim obdobjem in preteklim obdobjem do danes. Izdelan vizualni del poročila je prikazan na sliki 15.

Slika 15: Primer analitičnega poročila



Vir: lastno delo.

6 DISKUSIJA

V tem poglavju so interpretirani in analizirani ključni koraki za zagotavljanje kakovosti podatkov pri integraciji podatkov iz področja digitalnega trženja. Delo je bilo zasnovano z namenom oblikovanja predlogov za doseganje kakovosti podatkov pri integraciji podatkov iz več virov za uporabo v BI okolju. Zastavljeni cilji so vključevali pregled literature, analizo obstoječih procesov ETL, identifikacijo ključnih dimenzij kakovosti podatkov ter izvedbo analize na podlagi praktičnega primera.

Izbira preverjenih podatkovnih virov in njihovo poznavanje močno poenostavi integracijo podatkov. Čeprav so v delu uporabljeni trije viri podatkov eni izmed najpogostejših na področju digitalnega trženja, lahko na račun dejstva, da nanje nimamo vpliva (so zunanji viri), predstavljajo izziv z vidika zagotavljanja kakovosti podatkov. Kakovost sem izboljšal z uporabo dimenzij kakovosti podatkov, ki so se v mojem primeru izkazale za primernejše od različnih metodologij za zagotavljanje kakovosti.

Shranjevanje podatkov v DWH, ki omogoča pridobivanje enotnega pregleda nad delom ali celoto digitalnega trženja ali drugih področij, sočasno omogoča podporo pri integraciji podatkov iz različnih podatkovnih virov. Pri zasnovi arhitekture podatkovnega skladišča je bilo izhodišče organizacijsko podatkovno skladišče, ki omogoča centralizacijo oziroma integracijo podatkov na enem mestu. Zaradi avtomatizacije črpanja GA4 podatkov in želje po povečanju pravočasnosti podatkov, sem dodatno uporabil še en DWH. Uporaba dveh DWH-jev je sicer povečala kakovost podatkov, onemogočila pa je integracijo podatkov na

nivoju enega DWH-ja, oziroma uporabo arhitekture organizacijskega podatkovnega skladišča kot ga predstavijo Rifaie in drugi (2008). Za uspešno integracijo podatkov je bila pomembna tudi izbira naprednega BI orodja. Sodobna BI orodja, npr. Power BI, namreč poleg vizualiziranja, preoblikovanja in modeliranja podatkov na nivoju analitičnih poročil oz. nadzornih plošč, omogočajo tudi njihovo integracijo iz več podatkovnih virov. Fleksibilnost arhitekture, ki jo podpirajo sodobna BI orodja, je v skladu s priporočili Tvrdíková (2007) in Yasar (2023), ki poudarjata pomembnost prilagodljivosti arhitektur in orodij za učinkovito integracijo podatkov.

Pri izbiri med procesoma ETL in ELT sem se odločil za ETL oziroma njegovo prilagojeno različico. Izbira je temeljila na infrastrukturi oziroma tradicionalni obliki podatkovnega skladišča, ki sem ga za proces uporabil. Prav tako se podatki, ki sem jih uporabil niso šteli med masovne podatke, zato je bila izbira ETL procesa ustreznejša. Čeprav je ETL proces v osnovi počasnejši, to pri zagotavljanju kakovosti podatkov nima vpliva.

Za podatke iz rešitve CRM in LinkedIna sem uporabil klasičen ETL proces, kjer sem podatke pred nalaganjem v DWH očistil in preoblikoval. Ker sta CRM in LinkedIn razmeroma majhna in strukturirana vira podatkov, dodatna obdelava podatkov po nalaganju, ki se izvede pri ELT procesu, ne bi pomenila večje prednosti. Čiščenje podatkov pred nalaganjem pomeni, da so podatki v DWH že v konsistentni obliki, kar zmanjša potrebo po nadaljnjih prilagoditvah. Veliko vlogo pri izvedbi ETL procesa je imelo tudi SSMS okolje, ki je omogočilo natančen nadzor nad strukturo podatkov ter njihovo preoblikovanje.

Pri podatkih iz GA4 in Google Ads sem uporabil neklasičen EL-ETL proces, kar pomeni, da so bili podatki najprej preneseni v BigQuery, nato pa obdelani v Power BI. Čeprav EL faza procesa ni neposredno izboljšala kakovosti podatkov ampak jih le premaknila, je bila koristna. Omogočila je namreč zmanjšanje stroškov obdelave v oblaku, povečala nabor razpoložljivih podatkov za poročanje in delno avtomatizirala proces. Za ETL fazo procesa sem izbral napredno analitično orodje Power BI, ker omogoča neposredno povezavo z BigQuery prek prilagojenega priključka. Poleg tega ponuja uporabniku prijazen grafični vmesnik, ki olajša integracijo in preoblikovanje podatkov brez potrebe po naprednem poznavanju SQL.

Po identifikaciji virov podatkov je ključno določiti najustreznejši način njihovega pridobivanja. Kot navaja Wijaya in Pudjoatmodjo (2015), proces pridobivanja vključuje izbiro ustreznih kriterijev, prenos podatkov in njihovo shranjevanje v podatkovni sistem. V nalogi je bila uporabljena kombinacija ročnega in avtomatiziranega pridobivanja podatkov, kar ima tako prednosti kot slabosti.

Zunanji viri pogosto omogočajo več načinov dostopa do podatkov. Za podatke iz LinkedIna in rešitve CRM, sem v tem primeru izbral način dostopa prek izvoza Excel in CSV datotek. Takšen način spada med najpogostejše in ne zahteva posebnega tehničnega znanja. Excel in CSV datoteke lahko obdela skoraj vsako napredno ETL orodje, kar pomeni, da pri

shranjevanju podatkov običajno ne pride do težav. Glavna prednost izbranega pristopa je njegova enostavnost. Procesa v tem primeru nisem avtomatiziral, kar povečuje tveganje človeških napak in podvajanja podatkov. Posledično sem moral v sklopu preoblikovanja podatkov zagotoviti, da do podvajanja podatkov ne prihaja. Avtomatizacija s pomočjo Python skript ali SSIS paketov, bi v tem in podobnih primerih dodatno izboljšala natančnost podatkov.

Tudi GA4 omogoča več načinov dostopa do podatkov, vključno z izvozom v Excel, vendar je integracija z okoljem BigQuery bistveno učinkovitejša, saj omogoča avtomatizacijo prenosa in zagotavlja pravočasnost podatkov. Ker se podatki v GA4 lahko spreminjajo še nekaj dni po prvotnem zajemu, je pomembno, da sistem omogoča tudi kasnejše posodobitve podatkov. Izbira okolja BigQuery je bila v tem primeru prava odločitev, saj omogoča dnevno posodabljanje podatkov brez ročnega posredovanja.

Preoblikovanje podatkov je ključni korak za zagotavljanje kakovosti podatkov. Na začetku sem izvedel analizo oziroma profiliranje podatkov, s katero sem preveril začetno kakovost podatkov ter transformacije potrebne za povečanje njihove kakovosti. Podatki digitalnega trženja so lahko različno strukturirani, njihova pretvorba v strukturirano oziroma tabelarično obliko pa olajša delo in zagotavlja **veljavnost podatkov**. V danem primeru je bila večina podatkov pravilno strukturiranih, kar pomeni, da so bili podani v obliki tabel, pri čemer je imel vsak vnos v tabeli enako strukturo. Le podatki GA4 niso bili pravilno strukturirani in njihova pretvorba v pravilno, tabelarično strukturo je povečala njihovo veljavnost.

Zaradi spreminjanja strukture podatkov pri izvozih iz zunanjih virov je prav tako pomembno, da se ustrezno določijo formati podatkov v posameznih stolpcih. Zunanji podatkovni viri pogosto izhajajo iz različnih tujih držav, kar pomeni, da se lahko formati oziroma oblike zapisa podatkov razlikujejo; npr. datum, valuta, cela in decimalna števila. Sodobna ETL in BI orodja omogočajo avtomatsko določanje formatov, kar nam olajša delo. Z določitvijo formatov oziroma oblike podatkov sem dosegel **konsistentnost podatkov**.

Pri uvozu Excel oziroma CSV datotek se pogosto uvozijo tudi prazne vrstice brez podatkov. Te vrstice sem odstranil, saj odstranitev zagotavlja **celovitost podatkov**.

Za podatke digitalnega trženja je značilno tudi, da se ponavadi integrirajo iz več podatkovnih virov, zato se močno poveča verjetnost podvajanja podatkov. Podvajanje je pogostejše tudi v primeru ročnih izvozov podatkov, saj prihaja do človeških napak; npr. ponovnega uvoza podatkov, ki v DWH že obstajajo. S tem, da sem podvojene podatke odkril in odstranil, sem zagotovil **edinstvenost podatkov**.

Ko zagotovimo kakovost po vseh zgoraj omenjenih dimenzijah, je smiselno preveriti tudi ustreznost oziroma **točnost podatkov**. Če zunanji viri ponujajo tudi svojo analitiko, se točnost podatkov najlažje zagotovi s primerjavo izvoženih podatkov s podatki iz analitike. Drugače se lahko uporabi dokumentacija, ki jo ponujajo zunanji podatkovni viri. S pomočjo izdelave lastnega analitičnega poročila je proces preverjanja točnosti še hitrejši.

V danem primeru se je izkazalo, da je zagotavljanje **pravočasnosti podatkov** iz zunanjih virov včasih nemogoče. Nekateri zunanji viri, v tem primeru LinkedIn in GA4, ne omogočata izvoza aktualnih podatkov. Izvozi podatkov potekajo z zamikom nekaj dni ali pa se po nekaj dnevih še lahko spremenijo. Pravočasnost podatkov se lahko izboljša z avtomatizacijo procesa pridobivanja podatkov, kjer je to mogoče.

Izdelava analitičnega poročila s pomočjo naprednega BI orodja ne omogoča le sprejemanja pomembnih odločitev na področju digitalnega trženja in drugih področjih, omogoča tudi uporabo vizualizacij, ki olajšajo odkrivanje napak v podatkih in s tem zagotavljanje kakovosti podatkov.

Celovita analiza primera kaže, da lahko s prilagoditvijo arhitektur, integracijo podatkov, spremljanjem kakovosti podatkov in učinkovito uporabo BI orodij podjetja dosežejo visoko raven kakovosti podatkov tudi v zahtevnih in spremenljivih okoliščinah.

7 SKLEP

V magistrskem delu sem se osredotočil na integracijo raznolikih podatkovnih virov v BI okolje za analizo digitalnega trženja. Glavni namen dela je podati priporočila za zagotavljanje kakovosti podatkov digitalnega trženja in s tem pomagati podjetjem pri integraciji in analitiki teh podatkov.

Prvi cilj je bil pripraviti pregled obstoječe literature na temo poslovne inteligence in BI orodij na področju digitalnega trženja. Ta cilj sem dosegel s preučitvijo širokega spektra znanstvenih virov, ki obravnavajo osnovne pojme poslovne inteligence in poslovne inteligence v digitalnem trženju, torej njeno arhitekturo, aplikacije in koristi za podjetja. Poseben poudarek je bil na orodjih kot so Power BI, Microsoft Analysis Service in BigQuery, ki omogočajo avtomatizacijo analitičnih procesov. Pregled literature je zagotovil trdno teoretično osnovo za nadaljnje raziskave.

Drugi cilj je bil pripraviti pregled načinov integracije podatkov iz različnih kanalov digitalnega trženja in možnost avtomatizacije procesa ETL. Tudi ta cilj je bil dosežen. V delu sem analiziral procesa ETL in ELT ter raziskal ali njuna avtomatizacija omogoča učinkovitejšo integracijo podatkov iz različnih virov, kot so Google Analytics, LinkedIn in CRM sistemi. Ugotovitve so pokazale, da avtomatizirani procesi zmanjšujejo možnost napak in izboljšujejo doslednost podatkov.

Tretji cilj, pregled obstoječe literature o doseganju kakovosti podatkov, je bil dosežen z identifikacijo in analizo ključnih dimenzij kakovosti podatkov - celovitosti, konsistentnosti, edinstvenosti, veljavnosti, točnosti in pravočasnosti. V delu sem opisal tudi metodologije, ki podpirajo zagotavljanje kakovosti, ter predstavil, kako jih je mogoče uporabiti za zagotavljanje kakovosti podatkov.

Četrti cilj je bil izvesti analizo procesa integracije podatkov, ki izhajajo iz več podatkovnih virov v BI okolje na podlagi implementacije konkretnega primera. Ta cilj je bil dosežen delno, saj so bile analize osredotočene na določene podatkovne vire in orodja. Kljub temu so rezultati pokazali, kako lahko pravilno načrtovana integracija podatkov izboljša kakovost informacij, kar podjetjem omogoča boljše razumevanje tržnih dejavnosti in uspešnejše odločanje.

Peti cilj, podati priporočila za zagotavljanje kakovosti podatkov pri integraciji iz več podatkovnih virov, je bil dosežen. Na podlagi ugotovitev iz teoretičnega in empiričnega dela sem oblikoval praktične smernice, ki vključujejo uporabo ETL procesov za zagotavljanje kakovosti podatkov ter integracijo podatkov iz več podatkovnih virov, uporabo naprednih orodij za analitiko in vizualizacijo. Priporočila so uporabna za podjetja, ki želijo izboljšati svoje analitične procese in bolje izkoristiti potencial svojih podatkov.

Kljub doseženim ciljem magistrsko delo vključuje nekaj omejitev. Pristop je bil preizkušen na omejenem številu podatkovnih virov in orodij, kar omejuje njegovo širšo aplikacijo. Poleg tega je bilo težišče obravnave na strukturiranih podatkih, medtem ko so ostali nestrukturirani podatki manj obdelani. Za bolj poglobljeno analizo bi morala biti empirična implementacija razširjena na več primerov in več tehnologij.

Magistrsko delo prinaša dodano vrednost z več vidikov. Združuje teorijo in prakso na področju zagotavljanja kakovosti podatkov digitalnega trženja in integracije. Ugotovitve iz empiričnega dela prinašajo tudi nova spoznanja o izzivih in rešitvah, ki jih podjetja lahko upoštevajo pri delu s heterogenimi podatkovnimi viri. Poleg tega delo ponuja smernice, ki so praktično uporabne, a tudi prilagodljive za različne poslovne potrebe.

Obstaja več možnosti za nadaljnje raziskave. Ena od njih je razširitev metodologije na nestrukturirane podatke, ki pridobivajo vse večji pomen v digitalnem trženju. Druga je uporaba ene ali več naštetih metodologij oziroma okvirjev za zagotavljanje kakovosti podatkov. Prav tako bi bilo zanimivo preučiti implementacijo na drugih poslovnih področjih ali z uporabo dodatnih BI orodij, kar bi razširilo uporabnost pridobljenih ugotovitev.

S tem delom podjetjem ponujam tako teoretične kot praktične smernice za pomoč pri obvladovanju izzivov na področju zagotavljanja kakovosti in integracije podatkov za analizo digitalnega trženja.

LITERATURA IN VIRI

1. 365blog. (2017, 11. april). *Microsoft Dynamics 365* [objava na blogu]. https://www.microsoft.com/en-us/dynamics-365/blog/it-professional/2017/04/11/introduction-to-dynamics-365-data-export-service/
2. Ariyachandra, T. in Watson, H. (2010). Key organizational factors in data warehouse architecture selection. *Decision Support Systems*, *49*(2), 200–212.

3. Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer G. in Schwarzenbach, J. (2013). *The six primary dimensions for data quality assessment.* DAMA UK working group.
4. Batini, C., Barone, D., Canitza, F. in Grega, S. (2011). A data quality methodology for heterogeneous data. *International Journal of Database Management Systems, 3*(1), 60–79.
5. Batini, C., Cappiello, C., Francalanci, C. in Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys (CSUR)*, 1–52.
6. Bhosale, S. S., Sharma, Y. K. in Kurupkar, F. (2020). Role of business intelligence in digital marketing. *International Journal of Advance & Innovative Research, 7*(1), 2–7.
7. Boufares, F. in Salem, A. B. (2012). Heterogeneous data-integration and data quality: Overview of conflicts. V *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)* (str. 867–874). IEEE.
8. Brazhnik, O. in Jones, J. F. (2007). Anatomy of data integration. *Journal of biomedical informatics, 40*(3), 252–269.
9. Brinker, S. (2023, maj). *Chiefmartec*. Pridobljeno 11. decembra 2023 s https://chiefmartec.com/2023/05/2023-marketing-technology-landscape-supergraphic-11038-solutions-searchable-on-martechmap-com/
10. Cahffey, D. in Ellis-Chadwick, F. (2019). *Digital marketing.* Pearson uk.
11. Cai, L. in Zhu, Y. (2015). The Challenges of Data Quality and Data Quality. *Data science journal, 14*(2), 1–10.
12. Chanial, C., Dziri, R., Galhardas, H., Leblay, J., Nguyen, M.-H. L. in Manolescu, I. (2018). ConnectionLens: Finding connections across heterogeneous data sources. *Proceedings of the VLDB Endowment (PVLDB), 11*, 1–4.
13. Cichy, C. in Stefan, R. (2019). An overview of data quality frameworks. *Ieee Access, 7*, 24634–24648.
14. Dewdney, A. in Ride, P. (2014). *The digital media handbook* (2. izd.). Routledge.
15. English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits.* John Wiley & Sons, Inc.
16. Foot, C. (2024, 18. marec). *On-premises vs. cloud data warehouses: Pros and cons* [objava na blogu]. Pridobljeno 1. decembra 2024 s https://www.techtarget.com/searchdatamanagement/tip/On-premises-vs-cloud-data-warehouses-Pros-and-cons
17. Foote, K. D. (2023). *4 Common Data Integration Challenges* [objava na blogu]. Pridobljeno 14. oktobra 2023 s https://www.dataversity.net/4-common-data-integration-challenges/
18. Frankenfield, J. (2024, 17. oktober). *What Is Business Intelligence (BI)? Types, Benefits, and Examples* [objava na blogu]. Pridobljeno 2. decembra 2024 s https://www.investopedia.com/terms/b/business-intelligence-bi.asp
19. FutureLearn. (brez datuma). *Types and sources of data* [objava na blogu]. Pridobljeno 4. oktobra 2023 s https://www.futurelearn.com/info/courses/data-analytics-python-statistics-and-analytics-fundamentals/0/steps/186574

20. Gardner, S. R. (1998). Building the data warehouse. *Communications of the ACM, 41*(9), 52–60.
21. Gehra, B., Gentsch, P. in Hess, T. (2005). Business intelligence for the masses: Datenaufbereitung und Datenanalyse für den Controller im Wandel. *Controlling & Management Review, 49*, 236–242.
22. Google. (brez datuma a). *[GA4] BigQuery Export* [objava na blogu]. Pridobljeno 15. novembra 2023 s https://support.google.com/analytics/answer/9358801?hl=en
23. Google. (brez datuma b). *[GA4] Link Google Ads and Analytics* [objava na blogu]. Pridobljeno 14. novembra 2023 s https://support.google.com/analytics/answer/9379420?hl=en#zippy=%2Cin-this-article
24. Google. (brez datuma c). *[GA4] Share & export reports* [objava na blogu]. Pridobljeno 15. novembra 2023 s https://support.google.com/analytics/answer/9317657?hl=en
25. Google. (brez datuma č). *Google Ads: Definition* [objava na blogu]. Pridobljeno 14. novembra 2023 s https://support.google.com/google-ads/answer/6319?hl=en
26. Google. (brez datuma d). *Migrate from UA to GA4* [objava na blogu]. Pridobljeno 14. novembra 2023 s https://support.google.com/analytics/answer/10089681?hl=en
27. Grabbe, A. (2022, 29. september). *10 Best Self-Service BI Tools for Everyday Marketers* [objava na blogu]. Pridobljeno 13. oktobra 2023 s https://inbound.human.marketing/10-best-self-service-bi-tools-POfor-everyday-marketers
28. Hajmoosaei, A., Kashfi, M. in Kailasam, P. (2011). Comparison plan for data warehouse system architectures. V *The 3rd International Conference on Data Mining and Intelligent Information Technology Applications* (str. 290–293). IEEE.
29. Halevy, A., Rajaraman, A. in Ordille, J. (2006). Data integration: The teenage years. V *Proceedings of the 32nd international conference on Very large data bases* (str. 9–16).
30. Huang, S. (2020). *Classifying homogeneous data, what you need to know!* [objava na blogu]. Pridobljeno 5. oktobra 2023 s https://towardsdatascience.com/classifying-homogeneous-data-what-you-need-to-know-7b3a86e5f855
31. Jenifer, V., Vishmita, S. in Sowmiya, K. (2023). Growth of content marketing through LinkedIn channel. *Journal of Science and Research Archive, 8*(2), 166–171.
32. Jeusfeld, M. A., Quix, C. in Jarke, M. (1998). Design and Analysis of Quality Information for Data Warehouses. V *International Conference on Conceptual Modeling* (str. 349–362). Springer Berlin Heidelberg.
33. Keboola. (2022, 14. september). *Data Transformation in ETL Process Explained* [objava na blogu]. Pridobljeno 9. oktobra 2023 s https://www.keboola.com/blog/etl-transformation#:~:text=Data%20transformation%20is%20part%20of,computing%20new%20dimensions%20and%20metrics.
34. Kellerher, J. D. in Tierney, B. (2018). *Data science.* MIT Press.
35. Lambert, A. (2023, 14. januar). *How to extract data from the LinkedIn API* [objava na blogu]. Pridobljeno 16. novembra 2023 s https://lix-it.com/blog/how-to-extract-data-from-the-linkedin-api/

36. Langseth, J. in Nithi, V. (2003). Why proactive business intelligence is a hallmark of the real-time enterprise: outward bound. *Intelligent enterprise, 5*(18), 34–41.
37. Lee, Y. W., Strong, D. M., Kahn, B. K. in Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*(2), 133–146.
38. Lenzerini, M. (2002). Data integration: A theoretical perspective. V *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (str. 233–246). Association for Computing Machinery.
39. LinkedIn. (brez datuma). *Export your LinkedIn Page analytics report* [objava na blogu]. Pridobljeno 16. novembra 2023 s https://www.linkedin.com/help/linkedin/answer/a551206
40. McGuirk, M. (2023). Performing web analytics with Google Analytics 4: a platform review. *Journal of Marketing Analytics, 11*(4), 854–868.
41. Microsoft. (2023a, 2. feburar). *Learn* [objava na blogu]. Pridobljeno 18. decembra 2023 s https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16
42. Microsoft. (2023b, 22. februar). *What is Power BI?* [objava na blogu]. Pridobljeno 13. oktobra 2023 s https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview
43. Microsoft. (brez datuma). *Microsoft Dynamics 365* [objava na blogu]. Pridobljeno 13. oktobra 2024 s https://www.microsoft.com/en-us/dynamics-365/what-is-dynamics-365
44. Mohammadi, S. (2023). *Microsoft Tech Community* [objava na blogu]. Pridobljeno 19. oktobra 2024 s https://techcommunity.microsoft.com/t5/nonprofit-techies/microsoft-dynamics-365-marketing-crm/ba-p/3954398
45. Mukherjee, R. in Kar, P. (2017). A comparative review of data warehousing ETL tools with new trends and industry insight. V *2017 IEEE 7th International Advance Computing Conference (IACC*) (str. 943–948). IEEE.
46. Negash, S. (2004). Business intelligence. *Communications of the association for information systems, 13*, 176–196.
47. Pipino, L. L., Lee, Y. W. in Wang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM, 45*(4), 211–218.
48. Rahm, E. in Do, H. H. (2000). *Data cleaning: Problems and current approaches*. https://www.betterevaluation.org/sites/default/files/data_cleaning.pdf.
49. Rifaie, M., Kianmehr, K., Alhajj, R. in Ridley, M. J. (2008). *Data Warehouse Architecture and Design.* IEEE.
50. Riley, J. (2017). *Understanding metadata.* National Information Standards Organization.
51. Salah, A. H. in Alzghoul, A. (2024). Assessing the Moderating Role of Customer Orientation on the Impact of Business Intelligence Tools on Digital Marketing Strategy Optimization. *International Review of Management and Marketing, 3*(14), 18–25.
52. Saura, J. R., Palos-Sánchez, P. in Cerdá Suárez, L. M. (2017). Understanding the digital marketing environment with KPIs and web analytics. *Future internet, 4*(9), 1–13.
53. Sawicki, A. (2016). Digital marketing. *World Scientific News* (48), 82–88.

54. Scarisbrick-Hauser, A. in Rouse, C. (2007). The whole truth and nothing but the truth? The role of data quality today. *Direct Marketing: An International Journal*, *1*(3), 161–171.
55. Schlegel, K. in Sun, J. (2023). *Magic Quadrant for Analytics and Business Intelligence Platforms* [objava na blogu]. Pridobljeno 13. oktobra 2023 s https://www.gartner.com/doc/reprints?id=1-2955ETOT&ct=220215&st=sb?ocid=lp_pg398450_gdc_comm_az
56. Sebastian-Coleman, L. (2012). *Measuring data quality for ongoing improvement: a data quality assessment framework.* Morgan Kaufmann.
57. Seenivasan, D. (2022). ETL vs ELT: Choosing the right approach for your data warehouse. *International Journal for Research Trends and Innovation*, *2*(7), 110–122.
58. Seenivasan, D. (2023). ETL (Extract, Transform, Load) Best Practices. *International Journal of Computer Trends and Technology*, *71*(1), 40–44.
59. Sharma, V. in Kumar, P. (2022). Big query at a glance. *International Research Journal of Modernization in Engineering Technology and Science*, 2684–2686.
60. Singhal, B. in Aggarwal, A. (2022). ETL, ELT and Reverse ETL: A business case Study. V *2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)* (str. 1–4). IEEE.
61. Sreemanthy, J., Infant Joseph, V., Nisha, S., Chaaru Prabha, I. in Gokula Priya, R. (2020). Data Integration in ETL Using TALEND. V *2020 6th international conference on advanced computing and communication systems (ICACCS)* (str. 1444–1448). IEEE.
62. Strand, M. in Benkt, W. (2004). Incorporating External Data into Data Warehouses – Problems. V *Proceedings of the 7th International conference on information fusion (Fusion'04)* (str. 288–294). International Society of Information Fusion (ISIF).
63. Suciu, D. (1998). An overview of semistructured data. *ACM SIGACT News*, *29*(4), 28–38.
64. Talend. (brez datuma). *What is a Data Source?* [objava na blogu]. Pridobljeno 4. oktobra 2023 s https://www.talend.com/resources/data-source/#:~:text=A%20data%20source%20is%20the,process%20accesses%20and%20utilizes%20it.
65. Tayade, D. M. (2019). Comparative Study of ETL and E-LT in Data Warehousing. *International Research Journal of Engineering and Technology (IRJET)*, *6*, 2803–2807.
66. Thomsen, E. (2003). BI's Promised Land. *Intelligent enterprise*, *6*(4), 20–25.
67. Tvrdíková, M. (2007). Support of Decision Making by Business Intelligence Tools. V *6th International Conference on Computer Information Systems and Industrial Management Applications (CISIM'07)* (str. 364–368). IEEE.
68. Ur Rehman, K. U., Ahmad, U. in Mahmood, S. (2018). A comparative analysis of traditional and cloud data warehouse. *VAWKUM Transactions on Computer Sciences*, *6*, 34–40.
69. Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177–189.

70. Wang, L. (2017). Heterogeneous data and big data analytics. *Automatic Control and Information Sciences, 3*(1), 8–15.
71. Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM, 41*(2), 58–65.
72. Wijaya, R. in Pudjoatmodjo, B. (2015). An Overview and Implementation of Extraction-Transformation-Loading (ETL) Process in Data Warehouse (Case study: Department of agriculture). V *2015 3rd International Conference on Information and Communication Technology (ICoICT)* (str. 70–74). IEEE.
73. Yasar, K. (2023). *Business intelligence architecture* [objava na blogu]. Pridobljeno 15. oktobra 2023 s https://www.techtarget.com/searchbusinessanalytics/definition/business-intelligence-architecture

PRILOGA

Priloga 1: Struktura GA4 tabele izvožene v BigQuery

Ime stolpca	Podatkovni tip	Opis vrednosti
event_date	Besedilo	Datum, ko je bil dogodek zabeležen (format LLLLMMDD v registriranem časovnem pasu aplikacije).
event_timestamp	Celo število	Čas (v mikrosekundah), ko je bil dogodek zabeležen na odjemalcu.
event_name	Besedilo	Ime dogodka (prvi obisk, začetek seje, ogled strani).
event_params.key	Besedilo	Identifikator parametra dogodka.
event_params.string_value	Besedilo	Če je parameter dogodka predstavljen z nizom, kot je enolični krajevnik vira (angl. Uniform Resource Locator - URL) ali ime kampanje, je to polje izpolnjeno.
event_params.int_value	Celo število	Če je parameter dogodka predstavljen z celim številom, je to polje izpolnjeno.
event_params.float_value	Decimalno število	Če je parameter dogodka predstavljen z decimalnim številom, je to polje izpolnjeno.
event_params.double_value	Decimalno število	Če je parameter dogodka predstavljen z decimalnim številom z dvojno preciznostjo, je to polje izpolnjeno.
event_previous_timestamp	Celo število	Čas (v mikrosekundah), ko je bil dogodek predhodno zabeležen na odjemalcu.
event_value_in_usd	Decimalno število	Pretvorjena vrednost dogodka (v valuto ameriški dolar) za parameter vrednost dogodka.
event_bundle_sequence_id	Celo število	Zaporedna identifikacijska številka svežnja, v katerem so bili ti dogodki naloženi.
event_server_timestamp_offset	Celo število	Časovni zamik med časom zbiranja in časom nalaganja v mikrosekundah.
user_id	Besedilo	Edinstven identifikator, dodeljen uporabniku.
user_pseudo_id	Besedilo	Psevdonimni identifikator (npr. identifikator instance aplikacije) za uporabnika.
privacy_info.analytics_storage	Besedilo	Ali je ciljanje oglasov omogočeno za uporabnika.
privacy_info.ads_storage	Besedilo	Ali je shranjevanje podatkov v Analytics omogočeno za uporabnika.

se nadaljuje

Priloga 1: Struktura GA4 tabele izvožene v BigQuery (nad.)

Ime stolpca	Podatkovni tip	Opis vrednosti
privacy_info.user_transient_token	Besedilo	Ali je spletni uporabnik zavrnil shranjevanje podatkov v Analytics in je razvijalec omogočil merjenje brez piškotkov na osnovi prehodnih žetonov v podatkih strežnika.
user_properties.key	Besedilo	Identifikator lastnosti uporabnika.
user_properties.value.string_value	Besedilo	Tekstovna vrednost lastnosti uporabnika.
user_properties.value.int_value	Celo število	Celoštevilčna vrednost lastnosti uporabnika.
user_properties.value.float_value	Decimalno število	Dvojna vrednost lastnosti uporabnika.
user_properties.value.double_value	Decimalno število	To polje trenutno ni v uporabi.
user_properties.value.set_timestamp_micros	Celo število	Čas (v mikrosekundah), ko je bila lastnost uporabnika nazadnje nastavljena.
user_first_touch_timestamp	Celo število	Čas (v mikrosekundah), ko je uporabnik prvič odprl aplikacijo ali obiskal spletno mesto.
user_properties.value.float_value	Decimalno število	Dvojna vrednost lastnosti uporabnika.
user_ltv.revenue	Decimalno število	Življenjska vrednost (prihodek) uporabnika.

se nadaljuje

Priloga 1: Struktura GA4 tabele izvožene v BigQuery (nad.)

Ime stolpca	Podatkovni tip	Opis vrednosti
user_ltv.currency	Besedilo	Življenjska vrednost (valuta) uporabnika.
device.category	Besedilo	Kategorija naprave (mobilni telefon, tablica, namizni računalnik).
device.mobile_brand_name	Besedilo	Blagovna znamka naprave.
device.mobile_model_name	Besedilo	Ime modela naprave.
device.mobile_marketing_name	Besedilo	Tržno ime naprave.
device.mobile_os_hardware_model	Besedilo	Informacije o modelu naprave, pridobljene neposredno iz operacijskega sistema.
device.operating_system	Besedilo	Operacijski sistem naprave.
device.operating_system_version	Besedilo	Različica operacijskega sistema.
device.vendor_id	Besedilo	Identifikacijska številka prodajalca.
device.advertising_id	Besedilo	Oglasna identifikacijska številka.
device.language	Besedilo	Jezik operacijskega sistema.
device.is_limited_ad_tracking	Besedilo	Nastavitev naprave za omejevanje sledenja oglasov.
device.operating_system	Besedilo	Operacijski sistem naprave.
device.operating_system_version	Besedilo	Različica operacijskega sistema.
device.time_zone_offset_seconds	Celo število	Časovna razlika med časom aplikacije in naprave (sekunde).

se nadaljuje

Priloga 1: Struktura GA4 tabele izvožene v BigQuery (nad.)

Ime stolpca	Podatkovni tip	Opis vrednosti
device.browser	Besedilo	Brskalnik, v katerem je uporabnik gledal vsebino.
device.browser_version	Besedilo	Različica brskalnika, v katerem je uporabnik gledal vsebino.
device.web_info.browser	Besedilo	Brskalnik, v katerem je uporabnik gledal vsebino.
device.web_info.browser_version	Besedilo	Različica brskalnika, v katerem je uporabnik gledal vsebino.
device.web_info.hostname	Besedilo	Gostiteljsko ime, povezano z zabeleženim dogodkom.
geo.city	Besedilo	Mesto, s katerega so bili poročani dogodki, na podlagi naslova internetnega protokola (angl. Internet Protocol, v nadaljevanju IP).
geo.country	Besedilo	Država, s katere so bili poročani dogodki, na podlagi IP naslova.
geo.continent	Besedilo	Celina, s katere so bili poročani dogodki, na podlagi IP naslova.
geo.region	Besedilo	Regija, s katere so bili poročani dogodki, na podlagi IP naslova.

se nadaljuje

Priloga 1: Struktura GA4 tabele izvožene v BigQuery (nad.)

Ime stolpca	Podatkovni tip	Opis vrednosti
geo.metro	Besedilo	Metropolitansko območje, s katerega so bili poročani dogodki, na podlagi IP naslova.
app_info.id	Besedilo	Ime paketa ali identifikator paketa aplikacije.
app_info.version	Besedilo	Različica aplikacije ali kratka različica paketa.
app_info.install_store	Besedilo	Trgovina, ki je namestila aplikacijo.
app_info.firebase_app_id	Besedilo	Firebase App identifikator, povezan z aplikacijo.
app_info.install_source	Besedilo	Trgovina, ki je namestila aplikacijo.
traffic_source.name	Besedilo	Ime trženjske kampanje, ki je prvič pridobila uporabnika.
traffic_source.medium	Besedilo	Ime medija (plačano iskanje, organsko iskanje, e-pošta itd.), ki je prvič pridobil uporabnika.
traffic_source.source	Besedilo	Ime omrežja, ki je prvič pridobilo uporabnika.
stream_id	Besedilo	Številčni identifikator podatkovnega toka, iz katerega izvira dogodek.
platform	Besedilo	Platforma podatkovnega toka (splet, IOS ali Android), iz katere izvira dogodek.
event_dimensions.hostname	Besedilo	Ime gostitelja spletne strani, na kateri se je zgodil dogodek
ecommerce.total_item_quantity	Celo število	Skupno število artiklov v tem dogodku, kar je vsota količin artiklov.
ecommerce. purchase_revenue_in_usd	Decimalno število	Prihodek od nakupa tega dogodka, predstavljen v ameriških dolarjih (USD) z običajno enoto. Prikazano samo za dogodek nakupa.
ecommerce. purchase_revenue	Decimalno število	Prihodek od nakupa tega dogodka, predstavljen v lokalni valuti z običajno enoto. Prikazano samo za dogodek nakupa.
ecommerce. purchase_revenue_in_usd	Decimalno število	Prihodek od nakupa tega dogodka, predstavljen v ameriških dolarjih (USD) z običajno enoto. Prikazano samo za dogodek nakupa.
ecommerce. refund_value_in_usd	Decimalno število	Znesek povračila v tem dogodku, predstavljen v ameriških dolarjih (USD) z običajno enoto. Prikazano samo za dogodek povračila.

se nadaljuje

Priloga 1: Struktura GA4 tabele izvožene v BigQuery (nad.)

Ime stolpca	Podatkovni tip	Opis vrednosti
ecommerce. refund_value	Decimalno število	Znesek povračila v tem dogodku, predstavljen v lokalni valuti z običajno enoto. Prikazano samo za dogodek povračila.
ecommerce. shipping_value_in_usd	Decimalno število	Stroški pošiljanja v tem dogodku, predstavljeni v ameriških dolarjih (USD) z običajno enoto.
ecommerce. tax_value_in_usd	Decimalno število	Stroški pošiljanja v tem dogodku, predstavljeni v lokalni valuti.
ecommerce. tax_value	Decimalno število	Vrednost davka v tem dogodku, predstavljena v ameriških dolarjih (USD) z običajno enoto.
ecommerce. unique_items	Celo število	Število edinstvenih artiklov v tem dogodku, na podlagi ID artikla, imena artikla in blagovne znamke artikla.
ecommerce. transaction_id	Besedilo	Identifikator transakcije.
items.item_id	Besedilo	Identifikator artikla.
items.item_name	Besedilo	Ime artikla.
items.item_brand	Besedilo	Blagovna znamka artikla.
items.item_variant	Besedilo	Različica artikla.
items.item_category	Besedilo	Kategorija artikla.
items.item_category2	Besedilo	Podkategorija artikla.
items.item_category3	Besedilo	Podkategorija artikla.
items.item_category4	Besedilo	Podkategorija artikla.
items.item_category5	Besedilo	Podkategorija artikla.
items.price_in_usd	Decimalno število	Cena artikla v ameriških dolarjih (USD) z običajno enoto.
items.price	Decimalno število	Cena artikla v lokalni valuti.
items.quantity	Celo število	Količina artikla. Če ni določeno, je nastavljena na 1.
items.item_revenue_in_usd	Decimalno število	Prihodek tega artikla, izračunan kot cena v USD*količina. Prikazano samo za dogodke nakupa, v USD z običajno enoto.
items.item_revenue	Decimalno število	Prihodek tega artikla, izračunan kot cena*količina. Prikazano samo za dogodke nakupa, v lokalni valuti z običajno enoto.

se nadaljuje

Priloga 1: Struktura GA4 tabele izvožene v BigQuery (nad.)

Ime stolpca	Podatkovni tip	Opis vrednosti
items.item_refund_in_usd	Decimalno število	Vrednost povračila tega artikla, izračunana kot cena v USD*količina. Prikazano samo za dogodke povračila, v USD z običajno enoto.
items.item_refund	Decimalno število	Vrednost povračila tega artikla, izračunana kot cena*količina. Prikazano samo za dogodke povračila, v lokalni valuti z običajno enoto.
items.coupon	Besedilo	Uporabljena koda kupona za ta artikel.
items.affiliation	Besedilo	Pripadnost izdelka, ki označuje dobaviteljsko podjetje ali lokacijo fizične trgovine.
items.location_id	Besedilo	Lokacija, povezana z artiklom.
items.item_location_id	Besedilo	Identifikator seznama, v katerem je bil artikel predstavljen uporabniku.
items.item_list_id	Besedilo	Ime seznama, v katerem je bil artikel predstavljen uporabniku.
items.item_list_name	Besedilo	Položaj artikla na seznamu.
items.item_list_index	Besedilo	Identifikator promocije izdelka.
items.promotion_id	Besedilo	Ime promocije izdelka.
items.promotion_name	Besedilo	Ime kreativnega dela, uporabljenega v promocijskem mestu.
items.creative_name	Besedilo	Ime mesta za kreativno delo.
items.creative_slot	Besedilo	Identifikator artikla.
items.item_params.key	Besedilo	Identifikator parametra artikla.
items.item_params.value.string_value	Besedilo	Če je parameter artikla predstavljen z besedilom, je to polje izpolnjeno.
items.item_params.value.int_value	Celo število	Če je parameter artikla predstavljen s celim številom, je to polje izpolnjeno.
items.item_params.value.float_value	Decimalno število	Če je parameter artikla predstavljen z decimalnim številom, je to polje izpolnjeno.
items.item_params.value.double_value	Decimalno število	Če je parameter artikla predstavljen z decimalnim številom z dvojno preciznostjo, je to polje izpolnjeno.
collected_traffic_source.manual_source	Besedilo	Ročno določen vir kampanje, ki je bil zabeležen z dogodkom.

se nadaljuje

Priloga 1: Struktura GA4 tabele izvožene v BigQuery (nad.)

Ime stolpca	Podatkovni tip	Opis vrednosti
collected_traffic_source.manual_campaign_id	Besedilo	Ročno določena identifikacijska številka kampanje, ki je bila zabeležena z dogodkom.
collected_traffic_source.manual_campaign_name	Besedilo	Ročno določeno ime kampanje, ki je bilo zabeleženo z dogodkom.
collected_traffic_source.manual_source	Besedilo	Ročno določen vir kampanje, ki je bil zabeležen z dogodkom.
collected_traffic_source.manual_term	Besedilo	Ročno določena ključna beseda/izraz kampanje, ki je bila zabeležena z dogodkom.
collected_traffic_source.manual_content	Besedilo	Dodatni ročno določeni meta podatki kampanje, ki so bili zabeleženi z dogodkom.
collected_traffic_source.gclid	Besedilo	Google identifikator klika (gclid), ki je bil zabeležen z dogodkom.
collected_traffic_source.dclid	Besedilo	Google Marketing Platform (GMP) identifikator (dclid), ki je bil zabeležen z dogodkom.
collected_traffic_source.srsltid	Besedilo	Google Merchant Center identifikator (srsltid), ki je bil zabeležen z dogodkom.
is_active_user	Logična vrednost	Ali je bil uporabnik aktiven ali neaktiven v katerem koli trenutku koledarskega dne.

Vir: lastno delo.