Sveučilište Jurja Dobrile u Puli

Fakultet informatike u Puli

Ivana Pleše

**Australske prometne nesreće sa smrtnim ishodom 1989-2021**

Seminarski rad

Pula, svibanj 2024.

Sveučilište Jurja Dobrile u Puli

Fakultet informatike u Puli

Ivana Pleše

**Australske prometne nesreće sa smrtnim ishodom 1989-2021**

Seminarski rad

JMBAG: 0303100674

Studijski smjer: Informatika

Kolegij: Skladišta i rudarenje podataka

Mentor: doc.dr.sc. Goran Oreški

Pula, svibanj 2024.



IZJAVA O AKADEMSKOJ ČESTITOSTI

Ja, dolje potpisana Ivana Pleše, ovime izjavljujem da je ovaj seminarski rad rezultat isključivo mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na objavljenu literaturu kao što to pokazuju korištene bilješke i bibliografija. Izjavljujem da niti jedan dio seminarskog rada nije napisan na nedozvoljen način, odnosno da je prepisan iz kojega necitiranog rada, te da ikoji dio rada krši bilo čija autorska prava. Izjavljujem, također, da nijedan dio rada nije iskorišten za koji drugi rad pri bilo kojoj drugoj visokoškolskoj, znanstvenoj ili radnoj ustanovi.

Student

­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Pula, svibanj 2024.

**SADRŽAJ**

[**1.** **POSLOVNA INTELIGENCIJA** 5](#_Toc167574855)

[**2.** **UVOD** 7](#_Toc167574856)

[**3.** **ODABIR SKUPA PODATAKA** 2](#_Toc167574857)

[**3.1.** **OSNOVNA ANALIZA PODATAKA** 2](#_Toc167574858)

[**4.** **RELACIJSKI MODEL PODATAKA** 4](#_Toc167574859)

[**4.1.** **IZRADA BAZE PODATAKA** 4](#_Toc167574860)

[**4.2.** **PUNJENJE BAZE PODATAKA** 4](#_Toc167574861)

[**4.3.** **EER DIJAGRAM** 4](#_Toc167574862)

[**5.** **DIMENZIJSKI MODEL PODATAKA** 6](#_Toc167574863)

[**5.1.** **STAR SCHEMA** 7](#_Toc167574864)

[**5.2.** **ETL PROCES** 7](#_Toc167574865)

OLAP ANALIZA

ZAKLJUČAK

LITERATURA

# **POSLOVNA INTELIGENCIJA**

Poslovna inteligencija (Business Intelligence, BI) predstavlja proces prikupljanja, analiziranja, tumačenja i prikazivanja podataka za podršku donošenju poslovnih odluka. Ovaj koncept integrira tehnologiju, alate i postupke za prikupljanje podataka iz mnogih izvora, njihovu transformaciju u korisne informacije i njihovu prezentaciju u svrhu podrške strateškom, taktičkom i operativnom donošenju odluka u poslovnom okruženju. Cilj poslovne inteligencije je podržati bolje poslovno odlučivanje i omogućiti organizacijama da postignu konkurentsku prednost. Poslovna inteligencija obuhvaća nekoliko ključnih komponenti: Prikupljanje podataka: Proces prikupljanja podataka iz različitih izvora, uključujući interne izvore poput baza podataka, ERP sustava, CRM sustava i vanjske izvore poput tržišnih istraživanja, društvenih mreža i javnih baza podataka. Skladištenje podataka: Prikupljeni podaci se pohranjuju u skladišta podataka (Data Warehouses) ili manje skladišne jedinice poznate kao data marts. Skladišta podataka omogućuju centraliziranu pohranu podataka i olakšavaju pristup i analizu podataka. ETL proces: Extract, Transform, Load (ETL) procesi uključuju izdvajanje podataka iz izvora, njihovu transformaciju u odgovarajući format i učitavanje u skladište podataka. ETL procesi osiguravaju kvalitetu podataka i njihovu pripremljenost za analizu. Analiza podataka: Korištenjem različitih analitičkih tehnika, poput statističke analize, rudarenja podataka (Data Mining), prediktivne analitike i strojnog učenja, organizacije mogu otkriti obrasce, trendove i korelacije u podacima. Izvještavanje i vizualizacija podataka: Izvještaji, nadzorne ploče (dashboards) i vizualizacije podataka predstavljaju rezultate analiza na razumljiv i interaktivan način. Alati poput Power BI, Tableau i Qlik omogućuju korisnicima da stvaraju dinamičke vizualizacije i prilagođene izvještaje. Što se tiče funkcioniranja procesa poslovne inteligencije, prvo je važna integracija i učitavanje podataka: Podaci iz izvornog sustava integrirani su i učitani u skladište podataka ili drugi repozitorij. Priprema podataka za analizu: Skupovi podataka organizirani su u analitičke modele podataka ili OLAP kocke kako bi se pripremili za analizu. Analitički upiti: BI analitičar, ostali analitičari i poslovni korisnici pokreću analitičke upite prema dostupnim podacima. Vizualizacija i izvještavanje: Rezultati upita ugrađeni su u podatke, vizualizaciju, nadzorne ploče, izvješća i online portale.Korištenje informacija: Rukovoditelji i radnici koriste informacije za bolje donošenje odluka i strateško planiranje.

Prednosti poslovne inteligencije i njene primjene donose brojne pogodnosti organizacijama, uključujući poboljšano donošenje odluka: BI omogućuje menadžerima i poslovnim korisnicima donošenje odluka temeljenih na podacima, smanjujući subjektivnost i povećavajući točnost i pouzdanost odluka. Povećanje učinkovitosti: Analiza podataka može identificirati neučinkovitosti i uska grla u poslovnim procesima, omogućujući organizacijama da optimiziraju resurse i povećaju operativnu učinkovitost. Bolje razumijevanje tržišta i kupaca: BI pruža duboke uvide u ponašanje i preferencije kupaca, omogućujući organizacijama da prilagode svoje proizvode i usluge kako bi bolje zadovoljile potrebe kupaca. Povećanje konkurentne prednosti: Organizacije koje uspješno primjenjuju BI mogu brže reagirati na tržišne promjene, identificirati nove poslovne prilike i ostati ispred konkurencije. Unapređenje financijske učinkovitosti: BI omogućuje detaljnu analizu financijskih podataka, što pomaže u identificiranju područja za smanjenje troškova i povećanje prihoda.

Poslovna inteligencija koristi širok spektar alata i tehnologija za prikupljanje, pohranu, analizu i vizualizaciju podataka: Alati za prikupljanje podataka: Alati poput Apache NiFi, Talend i Informatica omogućuju integraciju podataka iz različitih izvora. Sustavi za skladištenje podataka: Oracle, Microsoft SQL Server, Amazon Redshift i Google BigQuery su popularni sustavi za skladištenje podataka koji pružaju skalabilne i pouzdane platforme za pohranu podataka. ETL alati: Alati poput Apache NiFi, Talend, Informatica i Pentaho Data Integration olakšavaju ETL procese, omogućujući učinkovitu transformaciju i prijenos podataka u skladišta podataka. Analitički alati: Alati poput R, Python, SAS i SPSS pružaju napredne analitičke mogućnosti za rudarenje podataka, prediktivnu analitiku i strojno učenje. Alati za izvještavanje i vizualizaciju podataka: Power BI, Tableau, Qlik Sense i Looker omogućuju stvaranje interaktivnih izvještaja i vizualizacija koje olakšavaju razumijevanje složenih podataka.

Implementacija poslovne inteligencije nije bez izazova. Neki od glavnih izazova uključuju: Kvaliteta podataka: Loša kvaliteta podataka može dovesti do netočnih analiza i loših odluka. Osiguravanje točnosti, potpunosti i konzistentnosti podataka ključno je za uspjeh BI projekata. Integracija podataka: Integracija podataka iz različitih izvora može biti složena, posebno kada se radi o velikim količinama podataka ili različitim formatima podataka. Sigurnost i privatnost: Zaštita podataka i osiguravanje privatnosti podataka su kritični aspekti poslovne inteligencije, posebno s obzirom na sve strože regulative o zaštiti podataka. Kultura podataka: Uvođenje BI rješenja zahtijeva promjenu kulture unutar organizacije, gdje svi zaposlenici trebaju biti educirani i motivirani za korištenje podataka u svakodnevnim odlukama. Unatoč izazovima, uspješna primjena BI-a može donijeti značajne koristi, omogućujući organizacijama da budu agilnije, efikasnije i bolje informirane.

# **UVOD**

S obzirom da količina podataka generiranih svakodnevno raste eksponencijalno, upravljanje, analiza i vizualizacija tih podataka postali su ključni aspekti za donošenje informiranih poslovnih odluka. U sklopu kolegija "Skladišta i rudarenje podataka", cilj ovog projekta bio je primijeniti stečena teorijska znanja na samostalni praktičan primjer, koristeći stvarne podatke.

Projekt započinje preuzimanjem skupa podataka s Kaggle platforme, renomirane baze podataka koja nudi raznovrsne skupove podataka za analizu. Nakon preuzimanja, podaci su analizirani i prilagođeni u Jupyter Notebook okruženju, što je omogućilo detaljnu preglednost i pripremu podataka za daljnju obradu. Sljedeći korak uključivao je izradu baze podataka u MySQL-u putem Python skripti. Ova faza je obuhvatila kreiranje baza podataka i tablica te njihovo popunjavanje podacima iz CSV datoteke. Nakon toga, prikazani su Entitetsko-atributski dijagram (EER) i Star schema model, koji su osigurali vizualni prikaz strukture baze podataka i međusobnih odnosa među podacima. U Pentaho Data Integration alatima, kreirane su dimenzijske tablice i tablica činjenica, što je omogućilo strukturirano skladištenje podataka i njihovu pripremu za analitičke procese. Na kraju, korištenjem Microsoft Power BI alata, izrađene su vizualizacije koje omogućuju jednostavno i intuitivno pretraživanje, analizu i interpretaciju podataka.

Ovaj projekt integrira više faza procesa rada s podacima, od ekstrakcije i transformacije podataka, preko skladištenja, do analize i vizualizacije. Kroz ove korake, demonstrirana je važnost skladištenja podataka, pravilne analize te vizualizacije kao završne faze koja omogućuje krajnjim korisnicima donošenje utemeljenih odluka na temelju analiziranih podataka.

# **ODABIR SKUPA PODATAKA**

Podatci za ovaj projekt preuzeti su sa službene web stranice Kaggle, pod nazivom “Australian Fatal Road Accident 1989-2021”,

(link: <https://www.kaggle.com/datasets/deepcontractor/australian-fatal-car-accident-data-19892021>). U opisu podataka stoji: “Australska baza podataka o smrtnim slučajevima u prometnim nesrećama pruža osnovne pojedinosti o smrtnim slučajevima u prometnim nesrećama u Australiji koje policija svaki mjesec prijavljuje državnim i teritorijalnim tijelima za sigurnost na cestama. Smrtni slučajevi na cestama iz posljednjih mjeseci su preliminarni i serija je podložna reviziji.” Nakon detaljnog proučavanja ovih podataka, izabrani su jer sadrže sve potrebno za napraviti ovaj projekt. CSV datoteka sadrži više od 54 tisuće redaka I 23 stupca, što je dovoljno podataka za daljnje korake. Važno je da ima i vremensku dimenziju, od godine do detaljnijih stupaca koji opisuju koji dan je u tjednu, koje vrijeme je, je li noć ili dan i radi li se o blagdanskom periodu. Također, važni su kvantitativni i kvalitativni podatci koji također ovdje postoje. Nedostajućih vrijednosti nema previše, no svakako će biti eliminirane u daljnjoj izradi. Po završetku analize csv datoteka sa kojom će se izraditi projekt sastoji se od idućih stupaca: Crash id, Penalty Price (dodatno izrađen stupac sa random brojevima od 30 do 400 sa korakom od 10), Crash Type, State, Month, Year, Dayweek, Time, Crash Type, Bus Involvement, Heavy Rigid Truck Involvement, Articulated Truck Involvement, Speed Limit, Road User, Gender, Age, Christmas Period, Easter Period, Age Group, Day of week, Time of day. Izbrisani su stupci National Remoteness Areas, SA4 Name 2016, National LGA Name 2017 i National Road Type.

## **OSNOVNA ANALIZA PODATAKA**

Osnovna analiza podataka provedena je u Jupyteru kako bi uopće bilo moguće raditi sa preuzetim podatcima. Prvo su učitani podatci iz csva u dataframe pomoću importanog pandasa. Zamjenjene su prazne vrijednosti sa NaN vrijednosti. Zatim je bitno učitti pregled prvih 5 redaka svakog stupca kako bismo vidjeli je li sve dobro namješteno zasad. Zatim je potrebno učitati sve podatke radi analize kako bi dobili ispis dimenzija skupa podataka. Sljedeći korak bio je ispis imena stupaca i nedostajućih vrijednosti. Slijedi ispis broja jedinstvenih vrijednosti po stupcima, ispis tipova podataka po stupcima odnosno analiza kvantitativnih i kvalitativnih podataka, brisanje tri stupca koji se neće koristiti, dodavanje jednog novog stupca, ponovni pregled svih stupaca te zatim spremanje u novu csv datoteku. Također, skup je podijeljen na dva dijela, jedan od 80% podataka i drugi od 20%. Pregledani su i tipovi svih podataka u pojedinom stupcu te je napokon datoteka bila spremna za raditi sa njom.

# **RELACIJSKI MODEL PODATAKA**

## **IZRADA BAZE PODATAKA**

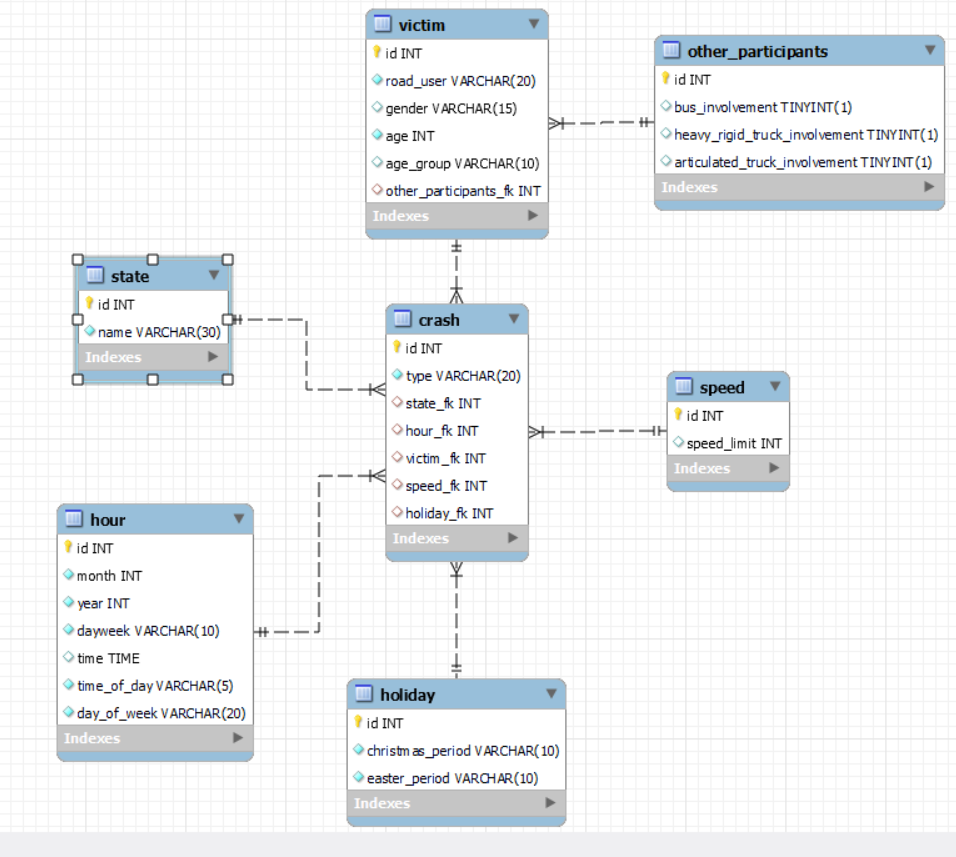
Nakon analize skupa podataka, idući korak bio je stvoriti smisleni relacijski model koji bi omogućio učinkovito skladištenje i obradu podataka. Za stvaranje baze podataka napisana je Python skripta povezana sa MySQL programom. MySQL je relacijski sustav za upravljanje bazama podataka (RDBMS) koji se koristi za pohranu i upravljanje podacima, a izabran je zbog svoje učinkovitosti, pouzdanosti i široke primjene u industriji. Analizom podataka identificirane su ključne relacije i definirane kardinalnosti među njima. Na temelju ovih analiza, stvoren je konceptualni model od osam relacija koje su definirane u skladu s analiziranim podacima. Svaka tablica je dizajnirana tako da podržava specifične atribute i veze. Konkretno, izrađena je baza podataka pod imenom prometne\_nesrece, a u njoj su izrađene tablice: crash, state, speed, victim, hour, other\_participants i holiday. Svi stupci iz csv datoteke poredani su u atribute tih tablica, te su dodana ograničenja na tablicama poput PRIMARY KEY i FOREIGN KEY kako bi se moglo povezati entitete.

## **PUNJENJE BAZE PODATAKA**

Sljedeći korak bio je otvoriti MySQL i pregledati je li baza uspješno stvorena te nalaze li se sve tablice i atributi u njoj. Nadalje se ponovno koristi Python skripta za popunjavanje tablica vrijednostima iz csv datoteke. Prvo povezujemo skriptu sa csv datotekom koja je ranije analizirana i u kojoj se nalazi 80% podataka. Pomoću pd.DataFrame, odabrani su atributi jedan po jedan te povezani sa nazivom stupaca iz kojih se žele izvući podatci.

## **EER DIJAGRAM**

Pomoću MySQL opcije Reverse Engineer kreiran je EER dijagram. Enhanced Entity- Relationship dijagram je grafički alat koji se koristi za modeliranje i dizajn baza podataka i služi za vizualno prikazivanje strukture baze, uključuje entitete, njihove atribute i odnose.



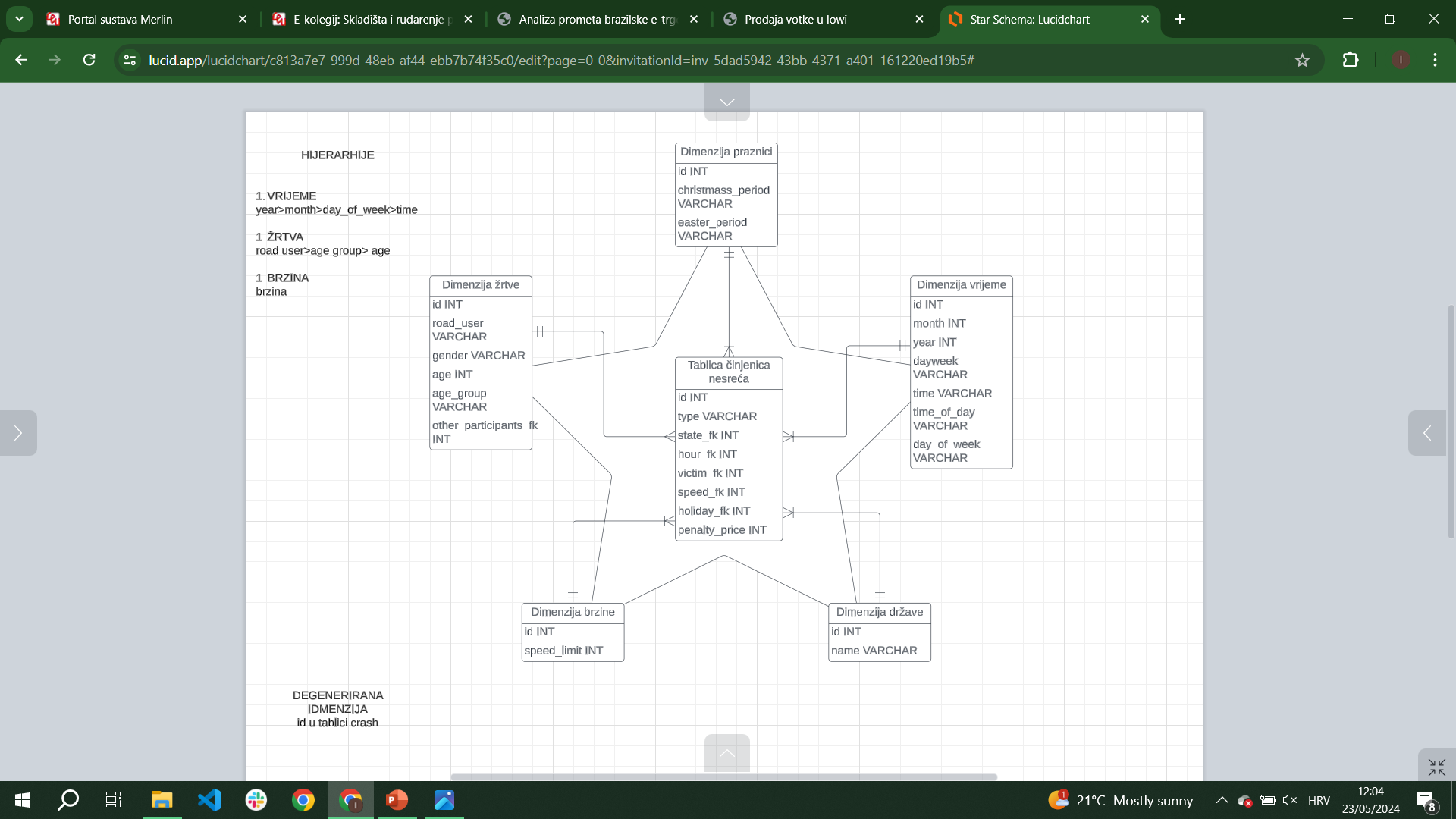
Slika 1: Prikaz EER dijagrama

# **DIMENZIJSKI MODEL PODATAKA**

Kako bi se strukturirali i rasporedili podaci na način koji olakšava analizu i ekstrakciju informacija, ključna tehnika koja se koristi u skladištenju podataka je dimenzijsko modeliranje. Dimenzijsko modeliranje omogućava organizaciju podataka u formatu koji je optimiziran za brzo i efikasno izvršavanje analitičkih upita te je razumljiv krajnjim poslovnim korisnicima.

Dakle dimenzijski model podataka je pristup dizajnu baze podataka koji optimizira pristup podacima za analitičke svrhe. Umjesto normaliziranih struktura koje se koriste u operativnim bazama podataka, dimenzijski model koristi denormalizirane tablice kako bi se poboljšali upiti. Dimenzijski model sastoji se od dvije osnovne vrste tablica: tablice činjenica i dimenzijskih tablica. Tablica činjenica nalazi se u središtu modela i sadrži numeričke podatke ili mjerenja koja želimo analizirati, poput iznosa prodaje, broja narudžbi ili količine proizvoda. U ovom primjeru to će biti penalty\_price atribut. Dimenzijske tablice okružuju tablicu činjenica i sadrže opisne atribute koji daju kontekst numeričkim podacima, kao što su vrijeme, lokacija, proizvod ili kupac. Ovaj raspored poznat je kao shema zvijezde, gdje tablica činjenica ima ključne veze prema dimenzijskim tablicama. Jedna od prednosti dimenzijskog modela je upotreba tehničkih ključeva (surrogate keys), koji djeluju kao jedinstveni identifikatori za retke tablica. Ovi ključevi omogućuju učinkovito povezivanje podataka između tablica, neovisno o prirodnim ključevima iz izvornog sustava. Tehnički ključevi pomažu u održavanju integriteta podataka i pojednostavljuju upravljanje podacima.

## **STAR SCHEMA**

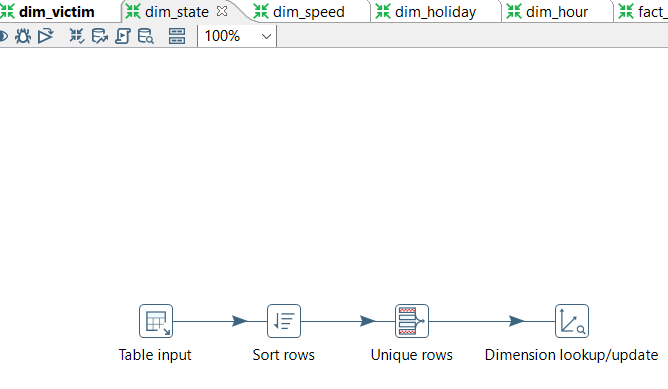


Slika 2: Prikaz star scheme izrađene u Lucidchartu

## **ETL PROCES**

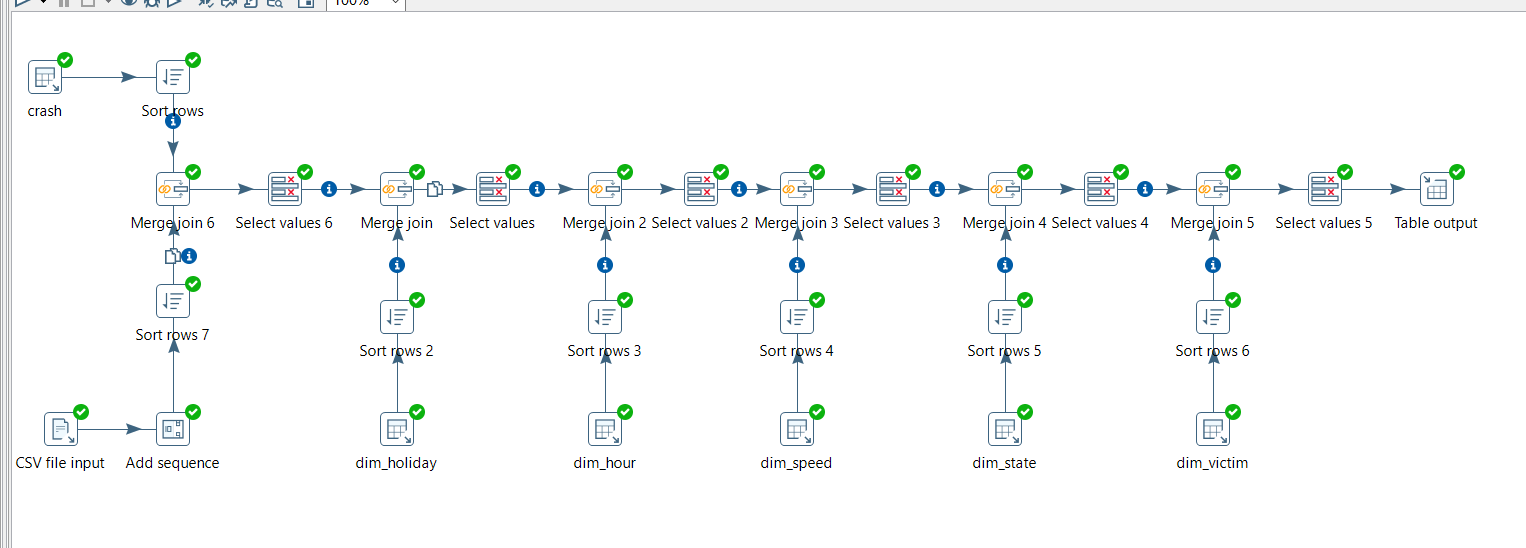
Za kreiranje i punjenje dimenzijskog modela koristio se alat Pentaho s kojim smo se upoznali u sklopu predavanja ovog kolegija. Pentaho pruža grafičko sučelje koje olakšava proces pretvorbe relacijskog modela u dimenzijski model. Prilikom kreiranja dimenzijskog modela, važno je slijediti nekoliko pravila kako bi se osigurala učinkovitost i točnost analitičkih procesa. Prvi korak u stvaranju dimenzijskog modela je odvajanje tablice činjenica od dimenzija. Tablica činjenica sadrži mjere poslovanja koje želimo pratiti, dok dimenzijske tablice pružaju kontekst za te mjere. Moguće je i spajanje sličnih dimenzija kako bi se smanjila kompleksnost modela. U tom slučaju povezane relacije koje opisuju slične entitete mogu se spojiti u jednu dimenziju. Ova redundancija olakšava kasniju analizu podataka i omogućuje brže izvršavanje upita. Sljedeći korak je punjenje tablica podacima. Ovaj proces uključuje izdvajanje podataka iz relacijskih tablica, njihovu transformaciju i unos u dimenzijske i tablicu činjenica. U alatu Pentaho, ovaj proces se provodi kroz ETL (Extract, Transform, Load) procese, koji omogućuju učinkovitu obradu i migraciju podataka iz izvora u dimenzijski model. ETL proces započinje ekstrakcijom podataka iz izvornog sustava, nakon čega slijedi transformacija podataka kako bi se uklonile nepravilnosti, konsolidirali duplikati i pripremili podaci za unos u skladište podataka. Transformirani podaci se zatim unose u odgovarajuće dimenzijske tablice i tablicu činjenica.

Konkretno, stvoren je dimenzijski model od pet dimenzija i jedne tablice činjenica, što je prikazano ranije u Star Schemi. Prvo je bilo potrebno spojiti Pentaho na MySQL bazu podataka na kojoj se nalaze svi podatci, dakle prometne\_nesrece. Također, u MySQL je potrebno i stvoriti novu bazu podataka koja će služiti kao skladište podataka, odnosno u nju će biti smještene sve dimenzijske tablice. Ona je nazvana pentaho\_nesrece. Stvoren je dimenzijski model od pet dimenzija i jedne tablice činjenica, što je prikazano ranije u Star Schemi. Prvo su kreirane dimenzije jedna po jedna ali na vrlo slične načine, pomoću Table Input, Sort Rows, Unique Rows, Select Values i Dimension Lookup/Update opcija. Po redu: u Table input-u bira se tablica iz stvorene konekcije na bazu podataka, te se odabire SQL select statement koji ispisuje odabranu tablicu. U Sort rows koraku sortirani su podatci prema id-u, uzlazno. U dimension lookup/update koraku prvo je potrebno odabrati target schemu, u ovom slučaju pentaho\_nesrece. Zatim target table (dim\_state, dim\_holiday, dim\_speed, dim\_victim, dim\_hour). U keys odabiru biramo ponovno id, a u Fieldsu sve ostale atribute te dimenzije. Slijedi vrlo važan korak, Technical key field, gdje se dodaje novi atribut svim dimenzijama (state\_tk, tk\_holiday, speed\_tk, victim\_tk i tk\_hour).



Slika 3: Prikaz koraka korištenih za izradu dimenzijskih tablica

Slijedi najkompleksniji postupak, stvaranje tablice činjenica fact\_crash.



Napokon se puni dimenzija te su sve dimenzije vidljive u MySQL bazi pentaho\_nesrece.

