SVEUČILIŠTE JURJA DOBRILE U PULI
FAKULTET INFORMATIKE

Tin Pritišanac

Analiza tržišta automobila 1970.-2024.

SEMINARSKI RAD

Pula, rujan, 2025. godine

SVEUČILIŠTE JURJA DOBRILE U PULI
FAKULTET INFORMATIKE

Tin Pritišanac

Analiza tržišta automobila 1970.-2024.

SEMINARSKI RAD

JMBAG: 0171256219, izvanredni student
Studijski smjer: Informatika
Kolegij: Skladišta i rudarenje podataka
Mentor: doc.dr.sc. Goran Oreški

Pula, rujan, 2025. godine

IZJAVA O AKADEMSKOJ ČESTITOSTI

Ja, dolje potpisan Tin Pritišanac, 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, rujan, 2025. godine

Sadržaj

1 Uvod

U današnjem digitalno vođenom poslovnom okruženju, sposobnost efikasnog prikupljanja, obrade i analize velikih količina podataka postala je ključni čimbenik uspjeha za organizacije u gotovo svim industrijskim granama. Automobilska industrija, kao jedna od najkompleksnijih i najkonkurentnijih grana gospodarstva, posebno se oslanja na napredne tehnologije skladištenja i analize podataka kako bi razumjela tržišne trendove, potrebe kupaca i operacijske učinkovitosti [1].

Skladišta podataka (eng. *data warehouses*) predstavljaju temelj modernih sustava za podršku odlučivanju, omogućujući integraciju različitih izvora podataka u jedinstvenu, koherentnu strukturu optimiziranu za analitičke potrebe [2]. U kontekstu automobilske industrije, ovakvi sustavi omogućuju analizu složenih odnosa između cijena vozila, karakteristika proizvođača, tržišnih segmenata i regionalnih specifičnosti.

Ovaj rad predstavlja sveobuhvatan pristup razvoju sustava za skladištenje i rudarenje podataka primjenjenog na analizu automobilskog tržišta. Kroz razvoj kompletnog ETL (Extract, Transform, Load) procesa, projekt demonstrira transformaciju sirovih podataka o automobilima u strukturirani dimenzijski model prilagođen OLAP (Online Analytical Processing) analizama. Korištenjem skupa podataka koji sadrži preko 97.000 zapisa o vozilima različitih proizvođača, modela i karakteristika, razvijen je sustav koji omogućuje dubinsku analizu tržišnih trendova i poslovnih uvida.

Glavni cilj ovog projekta je ilustracija praktične primjene teorijskih koncepata skladišta podataka kroz razvoj funcionalnog sustava koji može poslužiti organizacijama poput autoklubova, analitičkih kuća ili samim proizvođačima automobila u donošenju informiranih poslovnih odluka [3]. Projekt obuhvaća sve ključne faze razvoja sustava - od eksploratorne analize početnih podataka, preko dizajna relacijskog i dimenzijskog modela, do implementacije ETL procesa i prijedloga OLAP analiza.

Struktura rada prati logični tijek razvoja sustava, počevši od analize i pripreme početnog skupa podataka, preko stvaranja normaliziranog relacijskog modela, do konačne implementacije zvjezdastog modela optimiziranog za analitičke potrebe. Svaki korak popraćen je detaljnim objašnjenjima projektnih odluka i praktičnih implementacijskih izazova, čineći ovaj rad korisnim resursom za razumijevanje kompleksnosti razvoja realnih sustava za skladištenje podataka.

Kroz ovaj projekt, nastoji se pokazati kako tehnologije poput Apache Spark-a, MySQL-a i Tableau-a mogu biti integrirane u koherentan sustav koji omogućuje ne samo tehnički ispravan rad, već i stvaranje dodane vrijednosti kroz kvalitetne poslovne uvide [4].

2 Projektni zadatak

2.1 Definicija problema i motivacija

Automobilska industrija je jedna od najkompleksnijih gospodarskih grana koja generira ogromne količine podataka - od osnovnih karakteristika vozila, preko cijena i tržišnih trendova, do regionalnih specifičnosti i preferencija kupaca. Organizacije koje se bave analizom automobilskog tržišta, poput autoklubova, analitičkih kuća ili samih proizvođača, suočavaju se s izazovom pretvaranja ovih velikih količina podataka u korisne poslovne uvide.

Tradicionalni pristup analize podataka korištenjem jednostavnih baza podataka i osnovnih alata za izvještavanje često se pokazuje neadekvatnim za složene analitičke potrebe. Potreba za dubinskim analizama tržišnih trendova, usporednim analizama proizvođača, segmentacijom kupaca i predviđanjem buduće dinamike tržišta zahtijeva sofisticiraniji pristup [1].

Upravo tu se prepoznaje potreba za razvojem naprednog sustava skladištenja i analize podataka koji će omogućiti organizacijama da iz sirovih podataka izvuku maksimalnu vrijednost te donose informirane strateške odluke.

2.2 Cilj i opis projektnog zadatka

Glavni cilj ovog projekta je razvoj potpunog sustava za skladištenje i rudarenje podataka prilagođenog analizi automobilskog tržišta. Projekt obuhvaća kompletan tijek rada - od početnih sirovih podataka do konačnih analitičkih izvještaja koji mogu koristiti stvarnim organizacijama u njihovom poslovanju.

Konkretno, projekt ima za cilj:

- **Dizajnirati i implementirati relacijski model podataka** koji odgovara strukturi realnih podataka o automobilima, uključujući sve važne entitete i njihove međusobne odnose
- **Razviti dimenzijski model (star schema)** optimiziran za OLAP analize, koji omogućuje efikasno izvršavanje složenih analitičkih upita
- **Implementirati ETL proces** koji transformira podatke iz relacijske strukture u dimenzijski model koristeći suvremene tehnologije poput Apache Spark-a
- **Kreirati sustav za OLAP analize** koji demonstrira praktičnu primjenu različitih analitičkih operacija (slice, dice, drill-down, roll-up, pivot)
- **Pokazati poslovnu vrijednost** kroz konkretne scenarije korištenja koji ilustriraju kako ovakav sustav može koristiti organizacijama u donošenju poslovnih odluka

2.3 Opseg i ograničenja projekta

Projekt se fokusira na analizu podataka o automobilima prodavanima na fiktivnom tržištu, koristeći skup od preko 97.000 zapisa koji obuhvaća vozila različitih

proizvođača, modela i karakteristika. Ovaj skup podataka pruža reprezentativan uzorak koji omogućuje demonstraciju svih ključnih koncepata skladišta podataka.

Vremenski okvir podataka pokriva razdoblje od 1970. do 2024. godine, s naglaskom na zadnja dva desetljeća, što omogućuje analizu dugoročnih trendova i promjena na tržištu. Podaci uključuju ključne atribute poput cijene, godine proizvodnje, kilometraže, vrste goriva, veličine motora i drugih karakteristika relevantnih za tržišnu analizu.

Projekt se ograničava na demonstraciju tehničkih mogućnosti i metodologija, a ne pretendira na potpunu komercijalnu implementaciju. Fokus je na edukacijskim aspektima i ilustraciji najboljih praksi u razvoju sustava za skladištenje podataka.

2.4 Očekivani rezultati i doprinosi

Na završetku projekta očekuje se sljedeće:

1. **Funkcionalno skladište podataka** s implementiranim relacijskim i dimenzijskim modelom podataka
2. **Potpuno funkcionalan ETL proces** koji automatizira transformaciju podataka između različitih modela
3. **Demonstracija OLAP mogućnosti** kroz konkretne analitičke scenarije i upite
4. **Dokumentacija procesa** koja može služiti kao vodič za buduće slične projekte
5. **Praktični uvidi u tržište automobila** koji ilustriraju poslovnu vrijednost analitičkih sustava

Ovaj projekt predstavlja praktičnu demonstraciju kako teorijski koncepti skladišta podataka i business intelligence mogu biti primijenjeni u realnom kontekstu, pružajući studentima i praktičarima vrijedan uvid u izazove i mogućnosti moderne analize podataka [2].

2.5 Metodologija rada

Projekt slijedi strukturiran pristup razvoja sustava za skladištenje podataka, koji se može podijeliti u pet glavnih faza:

1. **Eksploratorna analiza podataka** - detaljno istraživanje početnog skupa podataka radi razumijevanja strukture, kvalitete i potencijalnih izazova
2. **Dizajn relacijskog modela** - stvaranje normalizirane strukture podataka koja odražava realne poslovne entitete i njihove odnose
3. **Implementacija dimenzijskog modela** - razvoj star schema arhitekture optimizirane za analitičke potrebe

4. **ETL proces** - implementacija sustava za ekstraktiranje, transformaciju i učitavanje podataka koristeći Apache Spark

5. **OLAP analize** - demonstracija analitičkih mogućnosti kroz praktične scenarije kreiranja grafova i tablica u programu Tableau

Svaka faza dokumentirana je s objašnjenjima projektnih odluka, tehničkih izazova i načina njihova rješavanja, čineći projekt korisnim resursom za razumijevanje praktičnih aspekata razvoja sustava za skladištenje podataka.

3 Odabir i analiza skupa podataka

3.1 Izvor i kriteriji odabira podataka

Za potrebe ovog projekta odabran je javno dostupan skup podataka s Kaggle platforme pod nazivom "90,000+ Cars Data From 1970 to 2024" [5]. Ovaj dataset sadrži informacije o preko 97.000 automobila u razdoblju od 1970. do 2024. godine, što ga čini izvrsnim kandidatom za demonstraciju koncepata skladišta podataka i analitičkih sustava.

Kriteriji koji su vodili odabir ovog specifičnog skupa podataka uključuju:

- **Veličina i kompleksnost** - Dataset s preko 97.000 zapisa omogućuje demonstraciju tehnika rukovanja velikim količinama podataka karakterističnih za stvarne poslovne scenarije
- **Bogatstvo atributa** - Podatak sadrži 10 različitih atributa koji pokrivaju ključne aspekte automobilskog tržišta: model, godinu proizvodnje, cijenu, vrstu mjenjača, kilometražu, tip goriva, porez, potrošnju, veličinu motora i proizvođača
- **Vremenski raspon** - Podaci se protežu kroz razdoblje od više desetljeća, što omogućuje analizu dugoročnih trendova i temporalnih promjena na tržištu
- **Praktična relevantnost** - Automobilska industrija predstavlja kompleksan sektor s jasnima hijerarhijama (proizvođač → model) i različitim kategorijama koje se prirodno mapiraju na dimenzijske modele podataka
- **Dostupnost i licenca** - Dataset je javno dostupan pod MIT licencom, što omogućuje slobodno korištenje u edukacijske svrhe

Odabrani dataset pruža solidan temelj za ilustraciju ključnih koncepata poslovne inteligencije i skladišta podataka, uključujući normalizaciju, denormalizaciju, ETL procese i OLAP analize.

3.2 Eksploratorna analiza podataka

Prije bilo kakve obrade ili transformacije podataka, provedena je detaljna eksploratorna analiza podataka (EDA) [6]. Ovaj korak je kritičan jer omogućuje razumijevanje strukture, kvalitete i karakteristika skupa podataka, što je neophodno za donošenje informiranih odluka o daljnjem pristupu [7].

3.2.1 Osnovna struktura podataka

Početni dataset sadrži sljedeće atribute:

Atribut	Tip podatka	Opis
model	string	Model automobila
year	integer	Godina proizvodnje
price	integer	Cijena u britanskim funtama
transmission	string	Tip mjenjača (Manual, Automatic, Semi-Auto)
mileage	integer	Kilometraža vozila
fuelType	string	Tip goriva (Petrol, Diesel, Hybrid, Electric)
tax	integer	Godišnji porez
mpg	float	Milje po galonu (potrošnja)
engineSize	float	Veličina motora u litrima
Manufacturer	string	Proizvođač automobila

Tablica 1: Struktura originalnog skupa podataka

Analiza je provedena koristeći Python biblioteke pandas i numpy, kako je prikazano u sljedećem kodu:

Listing 1: Segment skripte za analizu - osnovni pregled strukture podataka

```
import pandas as pd

# Ucitavanje podataka iz CSV datoteke
PATH = "../data/cars_data_original.csv"
data = pd.read_csv(PATH, delimiter=',')

# Ispis osnovnih informacija o skupu podataka
print("Dimenzije skupa podataka:", data.shape)
print("Tipovi podataka:")
print(data.dtypes)
print("Nedostaju vrijednosti:")
print(data.isna().sum())
print("Broj jedinstvenih vrijednosti po stupcima:")
print(data.nunique())
...
```

3.2.2 Ključni uvidi iz eksploratorne analize

Analiza je otkrila nekoliko važnih karakteristika skupa podataka:

1. **Kvaliteta podataka** - Dataset ne sadrži nedostajuće vrijednosti (null values), što značajno pojednostavljuje faze čišćenja i pripreme podataka

2. **Problematične vrijednosti** - Identificirano je 268 automobila s veličinom motora od 0.0 litara, što upućuje na greške u podacima ili specifične kategorije vozila (električni automobili) koje zahtijevaju posebnu obradu

3. **Distribucija proizvođača** - Analiza je pokazala prisutnost 9 glavnih proizvođača: Ford, Volkswagen, BMW, Skoda, Toyota, Mercedes-Benz (označen kao "merc"), Vauxhall, Audi i Hyundai (označen kao "hyundi")

4. **Vremenski raspon** - Podaci pokrivaju godine od 1970. do 2024., s najvećom koncentracijom vozila iz 2010-ih godina

5. **Cijenski raspon** - Cijene se kreću od nekoliko stotina do preko 100.000 funti, što reflektira širokospan spektar od budget do luksuznih vozila

3.3 Priprema i obrada podataka

Na temelju uvida dobivenih iz eksploratorne analize, implementiran je proces pripreme podataka koji obuhvaća čišćenje, standardizaciju i proširenje originalnog skupa podataka. Ovaj proces je ključan za stvaranje kvalitetne osnove za kasnije faze razvoja skladišta podataka [8].

3.3.1 Čišćenje i standardizacija podataka

Prvi korak u pripremi podataka bio je proces čišćenja i standardizacije, implementiran u sljedećem kodu:

Listing 2: Čišćenje i standardizacija podataka

```
import pandas as pd

# Ucitavanje originalnog skupa podataka
df = pd.read_csv("data/cars_data_original.csv", delimiter=',')

# Uklanjanje whitespace znakova iz naziva stupaca i podataka
df.columns = df.columns.str.strip()
df = df.apply(lambda x: x.str.strip() if x.dtype == "object" else x)

# Pretvaranje cijena i kilometraze u numericke tipove
df['price'] = pd.to_numeric(df['price'])
df['mileage'] = pd.to_numeric(df['mileage'])

# Standardizacija naziva proizvodjaca
df = df.rename(columns={'Manufacturer': 'manufacturer'})
df['manufacturer'] = df['manufacturer'].str.lower()
df['manufacturer'] = df['manufacturer'].replace({
    'bmw': 'BMW',
    'merc': 'Mercedes-Benz',
    'volkswagen': 'Volkswagen',
    'toyota': 'Toyota',
    'hyundi': 'Hyundai',
    'vauxhall': 'Vauxhall',
    'audi': 'Audi',
    'skoda': 'Skoda',
    'ford': 'Ford'
})

```

```
# Uklanjanje redaka s nedostajucim vrijednostima
df = df.dropna()
```

3.3.2 Proširenje skupa podataka hijerarhijskim atributima

Jedan od ključnih koraka u pripremi podataka za skladište bilo je proširenje originalnog skupa s dodatnim atributima koji omogućuju stvaranje bogatijih hijerarhija i boljih analitičkih mogućnosti. Ovaj proces je implementiran kroz dodavanje sljedećih dimenzija:

Listing 3: Prosirenje skupa podataka novim dimenzijama

```
# Dodavanje desetljeca na temelju godine proizvodnje
df['decade'] = (df['year'] // 10 * 10).astype(str) + 's'

# Mapiranje proizvodjaca na zemlje i regije
manufacturer_mapping = {
    'Hyundai': {'country': 'South Korea', 'region': 'Asia'},
    'Volkswagen': {'country': 'Germany', 'region': 'Europe'},
    'BMW': {'country': 'Germany', 'region': 'Europe'},
    'Skoda': {'country': 'Czech Republic', 'region': 'Europe'},
    'Ford': {'country': 'United States', 'region': 'North America'},
    'Toyota': {'country': 'Japan', 'region': 'Asia'},
    'Mercedes-Benz': {'country': 'Germany', 'region': 'Europe'},
    'Vauxhall': {'country': 'United Kingdom', 'region': 'Europe'},
    'Audi': {'country': 'Germany', 'region': 'Europe'}
}

# Dodavanje geografskih dimenzija
df['country'] = df['manufacturer'].map(
    lambda mfr: manufacturer_mapping.get(mfr, {}).get('country', 'Unknown'))
df['region'] = df['manufacturer'].map(
    lambda mfr: manufacturer_mapping.get(mfr, {}).get('region', 'Unknown'))

# Kategorizacija kilometraze
mileage_ranges = [
    (0, 5000, 'Very Low'),
    (5000, 20000, 'Low'),
    (20000, 50000, 'Medium'),
    (50000, 100000, 'High'),
    (100000, 150000, 'Very High'),
    (150000, float('inf'), 'Extreme')
```

```
]

df['mileageCategory'] = 'Unknown'
for low, high, category in mileage_ranges:
    mask = (df['mileage'] >= low) & (df['mileage'] < high)
    df.loc[mask, 'mileageCategory'] = category

# Klasifikacija velicine motora
engine_size_classes = [
    (0.1, 1.5, 'Small'),
    (1.5, 2.5, 'Medium'),
    (2.5, float('inf'), 'Large')
]

df['engineSizeClass'] = 'Unknown'
for low, high, category in engine_size_classes:
    mask = (df['engineSize'] >= low) & (df['engineSize'] <
       high)
    df.loc[mask, 'engineSizeClass'] = category

# Dodavanje starosti vozila i kategorizacije
current_year = 2025
df['age'] = current_year - df['year']

age_categories = [
    (0, 3, 'New'),
    (3, 7, 'Recent'),
    (7, 12, 'Mature'),
    (12, 20, 'Old'),
    (20, float('inf'), 'Vintage')
]

df['ageCategory'] = 'Unknown'
for low, high, category in age_categories:
    mask = (df['age'] >= low) & (df['age'] < high)
    df.loc[mask, 'ageCategory'] = category
```

3.3.3 Podjela skupa podataka za simulaciju različitih izvora

Za potrebe demonstracije ETL procesa koji može rukovati podacima iz različitih izvora, prošireni skup podataka podijeljen je na dva dijela u omjeru 80:20. Ovaj pristup omogućuje simulaciju realnog scenarija gdje skladište podataka prima informacije iz više nezavisnih sustava.

Listing 4: Podjela skupa podataka na dva dijela

```
import pandas as pd

# Ucitavanje prosirenog skupa podataka
```

```
df = pd.read_csv("processed/cars_data_EXPANDED.csv",
    delimiter=',')

# Nasumicna podjela u omjeru 80:20
df20 = df.sample(frac=0.2, random_state=1)
df80 = df.drop(df20.index)

print(Velicina 80% skupa:", df80.shape)
print("Velicina 20% skupa:", df20.shape)

# Spremanje u zasebne datoteke
df80.to_csv("processed/cars_data_80.csv", index=False)
df20.to_csv("processed/cars_data_20.csv", index=False)
```

3.4 Rezultati pripreme podataka

Na završetku procesa pripreme podataka, od originalnog skupa s 10 atributa stvaren je prošireni dataset s 16 atributa koji uključuje:

- **Originalne atribute** - svih 10 osnovnih karakteristika vozila
- **Temporalne dimenzije** - decade, age, ageCategory
- **Geografske dimenzije** - country, region
- **Kategorizirane atribute** - mileageCategory, engineSizeClass

Ovakav pristup pripreme podataka osigurava da konačni skup sadrži sve potrebne elemente za stvaranje bogatog dimenzijskog modela koji može podržati kompleksne OLAP analize. Dodatne hijerarhije omogućuje implementaciju naprednih analitičkih scenarija poput drill-down operacija od regije prema specifičnim modelima automobila, ili roll-up agregacija od individualnih vozila prema tržišnim segmentima.

Pripravljeni podaci predstavljaju čvrstu osnovu za sljedeću fazu projekta - dizajn i implementaciju relacijskog modela podataka koji će služiti kao izvor za ETL proces prema dimenzijskom skladištu podataka.

4 Relacijski model podataka

Nakon uspješno provedene analize podataka, sljedeći korak u razvoju skladišta podataka predstavlja kreiranje relacijskog modela koji će osigurati strukturirano i normalizirano čuvanje podataka. Relacijski model omogućuje organizaciju podataka u tablice povezane preko stranih ključeva, što omogućava efikasno dohvaćanje i manipulaciju podataka [9].

4.1 Analiza entiteta i atributa

Temeljito razumijevanje strukture podataka o automobilima omogućilo je identifikaciju ključnih entiteta koji će činiti okosnicu relacijskog modela. Kroz analizu originalnog skupa podataka identificirani su sljedeći glavni entiteti:

Osnovni entiteti:

- **Automobil** - središnji entitet koji sadrži osnovne karakteristike vozila
- **Proizvođač** - entitet koji predstavlja tvrtke koje proizvode automobile
- **Model** - specifični model automobila određenog proizvođača
- **Zemlja** - zemlja podrijetla proizvođača
- **Regija** - geografska regija kojoj pripada zemlja

Klasifikacijski entiteti:

- **Tip mjenjača** - klasifikacija prema vrsti transmisije
- **Tip goriva** - kategorije goriva koje automobil koristi
- **Desetljeće** - vremenski period proizvodnje
- **Kategorija starosti** - klasifikacija prema godinama starosti
- **Kategorija kilometraže** - klasifikacija prema prijeđenim kilometrima
- **Klasa veličine motora** - kategorije prema volumenu motora

Ova kategorizacija omogućuje stvaranje normaliziranog modela koji minimizira redundanciju podataka i omogućuje efikasne upite.

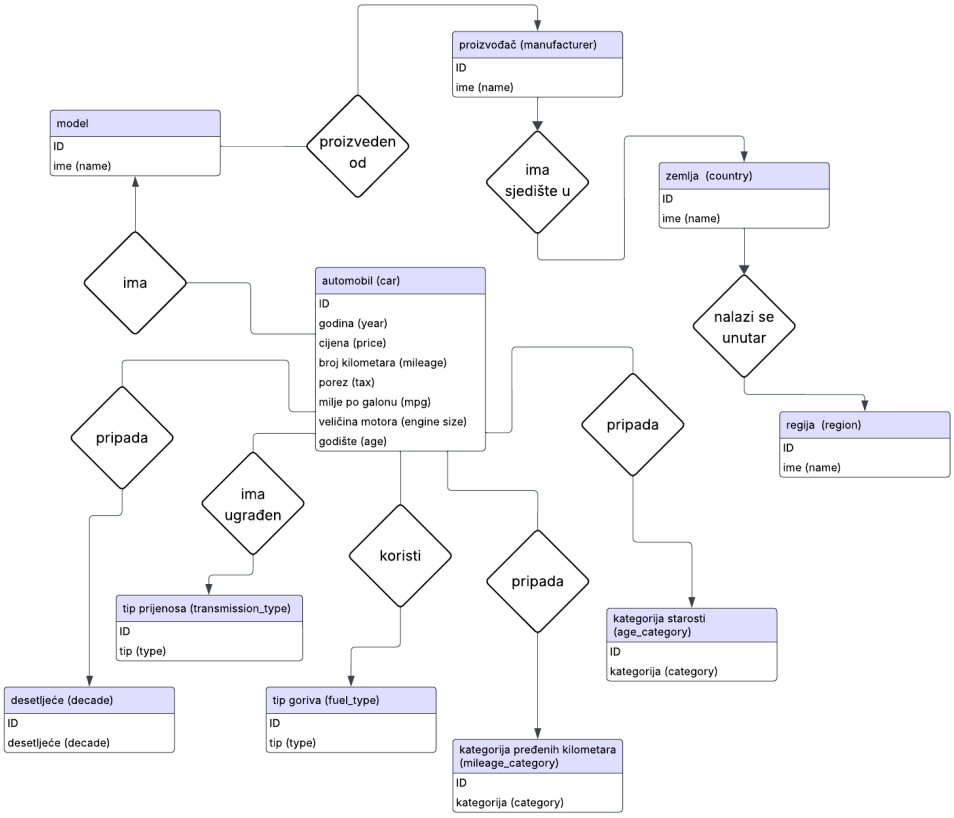
4.2 Konceptualni model podataka

Konceptualni model predstavlja visokorazinsku apstrakciju poslovnih zahtjeva bez ulaska u tehnička ograničenja. Za potrebe ovog projekta definiran je model koji odražava prirodne veze između entiteta u domeni trgovine automobilima.

Ključne veze u modelu uključuju:

- Svaki automobil pripada određenom modelu (1:N)
- Svaki model proizvodi točno jedan proizvođač (1:N)

- Svaki proizvođač dolazi iz jedne zemlje (1:N)
- Svaka zemlja pripada jednoj regiji (1:N)
- Automobil ima jednu kategoriju za svaki klasifikacijski atribut (1:N)

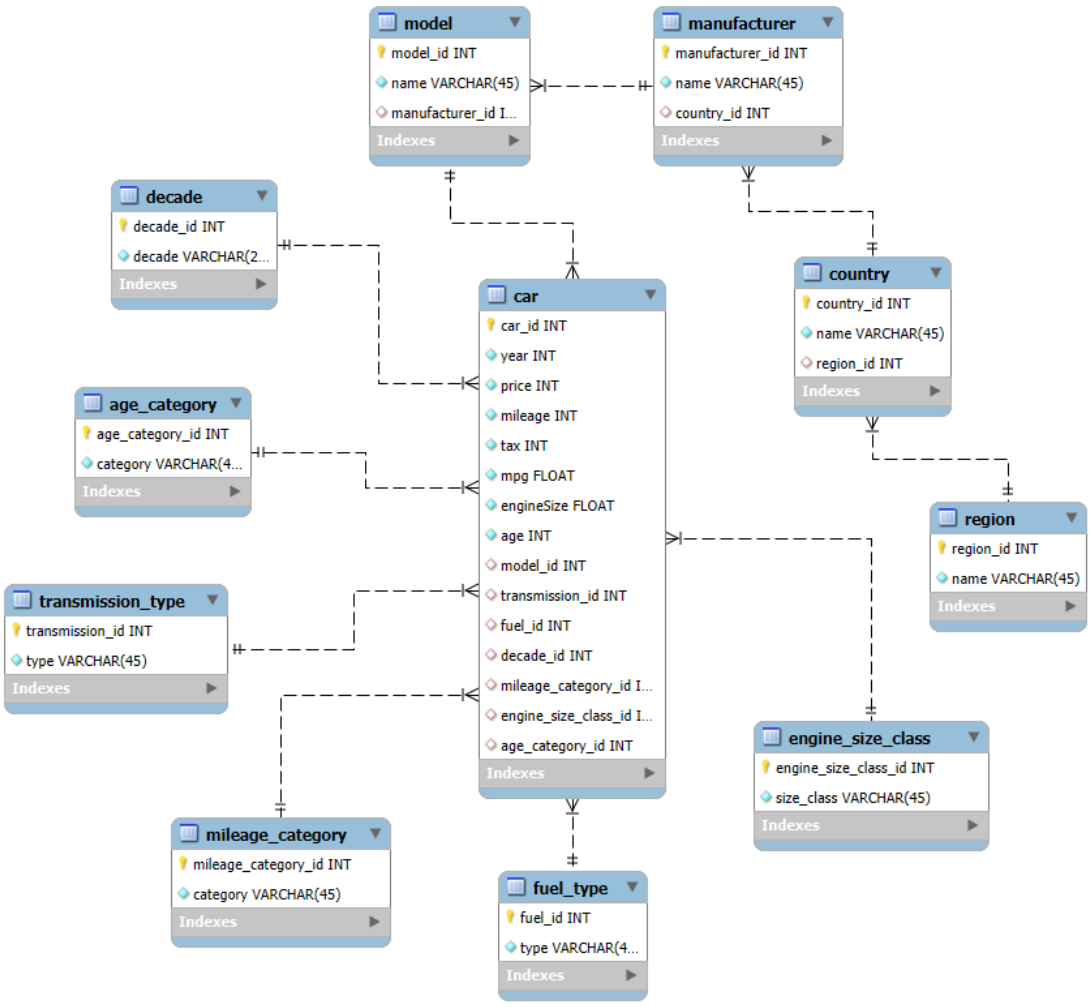


Slika 1: Konceptualni Entity-Relationship dijagram

ER dijagram prikazuje kompletan konceptualni model s entitetima, atributima i vezama između njih. Dijagram je kreiran koristeći standardnu notaciju koja jasno označava kardinalnosti i ograničenja. Na dijagramu je jasno vidljiva hijerarhijska struktura od regije do modela automobila, kao i klasifikacijski entiteti koji omogućuju analitičko izvještavanje. Svaki entitet sadrži primarni ključ i relevantne atribute koji podržavaju analitičke zahtjeve.

4.3 Implementacija relacijskog modela

Logički model podataka implementiran je koristeći MySQL bazu podataka, pri čemu je osobita pozornost posvećena normalnim formama i referencijalnim ograničenjima. Model slijedi treću normalnu formu (3NF) što osigurava minimizaciju redundancije podataka.



Slika 2: Implementirani relacijski model u MySQL bazi podataka

Implementacija uključuje:

- **Referencijalnu cjelovitost** - svi strani ključevi osiguravaju postojanje povezanih zapisa
- **Jedinstvene ograničenja** - sprječavanje dupliciranja entiteta
- **NOT NULL ograničenja** - osiguravanje kompletnosti kritičnih atributa
- **Auto-increment primarni ključevi** - efikasno upravljanje identifikatorima

4.4 Automatizacija kreiranja i popunjavanja baze

Za potrebe reproducibilnosti i održivosti projekta razvijene su Python skripte koje automatiziraju proces kreiranja i popunjavanja baze podataka. Glavna skripta koristi SQLAlchemy ORM za definiranje sheme baze.

Listing 5: Definiranje ORM modela za glavni entitet automobila

```
class Car(Base):
    __tablename__ = 'car'
    car_id = Column(Integer, primary_key=True, autoincrement
       =True)
    year = Column(Integer, nullable=False)
    price = Column(Integer, nullable=False)
    mileage = Column(Integer, nullable=False)
    tax = Column(Integer, nullable=False)
    mpg = Column(Float, nullable=False)
    engineSize = Column(Float, nullable=False)
    age = Column(Integer, nullable=False)

    # Strani kljucevi
    model_id = Column(Integer, ForeignKey('model.model_id'))
    transmission_id = Column(Integer, ForeignKey('
       transmission_type.transmission_id'))
    fuel_id = Column(Integer, ForeignKey('fuel_type.fuel_id'
       ))
    # ... ostali strani kljucevi
```

Proces popunjavanja baze provodi se u kontroliranim fazama:

1. **Kreiranje osnovnih entiteta** - regije, zemlje, proizvođači
2. **Kreiranje klasifikacijskih tabela** - tipovi goriva, mjenjača, kategorije
3. **Kreiranje modela** - povezivanje s proizvođačima
4. **Umetanje automobila** - povezivanje sa svim referentnim entitetima

4.5 Validacija podataka i testiranje

Za osiguravanje ispravnosti procesa migracije podataka u zasebnoj skripti implementiran je sveobuhvatan sustav testiranja koji uspoređuje originalne CSV podatke s podacima u bazi.

Listing 6: SQL upit za rekonstrukciju originalnih podataka

```
query = """
SELECT mfg.name as 'manufacturer', mdl.name as 'model'
, c.year, c.price, c.mileage, c.tax, c.mpg, c.engineSize
, t.type as 'transmission', f.type as 'fuelType'
, d.decade, cnt.name as 'country', r.name as 'region'
, mc.category as 'mileageCategory'
, esc.size_class as 'engineSizeClass'
, c.age, ac.category as 'ageCategory'
FROM car c
JOIN model mdl ON c.model_id = mdl.model_id
JOIN manufacturer mfg ON mdl.manufacturer_id = mfg.
   manufacturer_id
```

```
-- ... ostali JOIN-ovi
ORDER BY c.car_id ASC
"""
```

Test provjerava:

- **Integritet strukture** - postojanje svih stupaca u bazi
- **Kompletnost podataka** - broj zapisa u bazi odgovara CSV datoteci
- **Ispravnost sadržaja** - vrijednosti u bazi identične su originalnim podacima
- **Referencijalnu konzistentnost** - svi strani ključevi ispravno povezani

5 Dimenzijski model podataka

5.1 Izrada star scheme

5.2 Kreiranje dimenzijskih tablica

5.3 Kreiranje tablice činjenica

6 ETL proces

6.1 Izvlačenje podataka

6.2 Transformacija podataka

6.3 Popunjavanje skladišta podataka

7 OLAP analiza

7.1 Definiranje prikaza podataka

7.2 Vizualizacija podataka u Tableau

7.2.1 Graf 1 TODO

8 Zaključak

Literatura

[1] N. Silva, J. Barros, M. Y. Santos, C. Costa, and P. Cortez. Advancing Logistics 4.0 with the Implementation of a Big Data Warehouse: A Demonstration Case for the Automotive Industry. *Electronics*, 10(18):2221, 2021. [Na internetu]. Dostupno: https://www.mdpi.com/2079-9292/10/18/2221 [pristupano 29. kolovoza 2025.].

[2] G. Garani, A. Chernov, and I. Savvas. A Data Warehouse Approach for Business Intelligence. In *2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 70–75, 2019. [Na internetu]. Dostupno: https://ieeexplore.ieee.org/document/8795395 [pristupano 29. kolovoza 2025.].

[3] P. Nima. Data Warehousing and Business Intelligence Project on Car Insights in the United Kingdom. ResearchGate, 2018. [Na internetu]. Dostupno: https://www.researchgate.net/publication/330837638 [pristupano 29. kolovoza 2025.].

[4] B. Leka, D. Leka, and B. Baraku. Driving Operational Excellence: Business Intelligence in the Car Parts Industry. *WSEAS Transactions on Business and Economics*, 22:356–367, 2025. [Na internetu]. Dostupno: https://www.wseas.com/journals/bae/2025/a505118-356.pdf [pristupano 29. kolovoza 2025.].

[5] Meruvu Likith. 90,000+ Cars Data From 1970 to 2024. Kaggle, 2022. [Na internetu]. Dostupno: https://www.kaggle.com/datasets/meruvulikith/90000-cars-data-from-1970-to-2024 [pristupano 29. kolovoza 2025.].

[6] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain. Exploratory Data Analysis. In *Secondary Analysis of Electronic Health Records*, pages 185–203. Springer, 2016. [Na internetu]. Dostupno: https://link.springer.com/chapter/10.1007/978-3-319-43742-2_15 [pristupano 29. kolovoza 2025.].

[7] N. Ekbote and P. Dhanshetti. Techniques of Exploratory Data Analysis. *Madhya Pradesh Journal of Social Sciences*, 28(2):45–58, 2023. [Na internetu]. Dostupno: https://www.researchgate.net/publication/374674185 [pristupano 29. kolovoza 2025.].

[8] Z. Wu and Z. Wu. Exploration, Visualization, and Preprocessing of High-Dimensional Data. In *Statistical Methods in Molecular Biology*, pages 163–177. Springer, 2009. [Na internetu]. Dostupno: https://link.springer.com/protocol/10.1007/978-1-60761-580-4_8 [pristupano 29. kolovoza 2025.].

[9] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Pearson, 7th edition, 2017. [Na internetu]. Dostupno: https://www.pearson.com/us/higher-education/program/Elmasri-Fundamentals-of-Database-Systems-7th-Edition/PGM189052.html [pristupano 29. kolovoza 2025.].

Popis slika

Popis tablica