Sveučilište Jurja Dobrile u Puli

Fakultet Informatike u Puli

**LAURA LONČARIĆ**

**Prodaja proizvoda diljem svijeta u periodu 2011-2014 godine**

Seminarski rad

Sveučilište Jurja Dobrile u Puli

Fakultet Informatike u Puli

**LAURA LONČARIĆ**

**Prodaja proizvoda diljem svijeta u periodu 2011-2014 godine**

Seminarski rad

**JMBAG : 0303092470, redovni student**

**Studijski smjer: Informatika**

**Kolegij: Sustavi poslovne inteligencije**

**Profesor: doc.dr.sc Goran Oreški**

Sadržaj

[1. Uvod 1](#_Toc73908984)

[2. Odabir skupa podataka 2](#_Toc73908985)

[2.1 Analiza skupa podataka 2](#_Toc73908986)

[3. Relacijski model podataka 3](#_Toc73908987)

[3.1 Entity/Relationship model 3](#_Toc73908988)

[3.2 Popunjavanje baze podataka podacima 6](#_Toc73908989)

[4. Dimenzijski model podataka 9](#_Toc73908990)

[5. ETL proces 11](#_Toc73908991)

[5.1. Stvaranje i punjenje dimenzije „dim\_customer“ 11](#_Toc73908992)

[5.2. Stvaranje i punjenje dimenzije „dim\_location“ 13](#_Toc73908993)

[5.3. Stvaranje i punjenje dimenzije „market\_place“ 14](#_Toc73908994)

[5.4. Stvaranje i punjenje dimenzije „dim\_time\_of\_delivery“ 16](#_Toc73908995)

[5.5. Stvaranje i punjenje dimenzije „dim\_product“ 17](#_Toc73908996)

[5.6. Punjenje tablice činjenica 18](#_Toc73908997)

[6. Vizualizacija podataka 20](#_Toc73908998)

[7. Zaključak 24](#_Toc73908999)

[8. Literatura 25](#_Toc73909000)

# Uvod

Najvažnija komponenta svakog projekta poslovne inteligencije su podaci koji su bitni za donošenje poslovnih odluka. Poslovna inteligencija je rezultat razvoja sustava podrške odlučivanju. Oni kombiniraju operativne podatke s analitičkim alatima kako bi donositeljima odluka predstavili složene i kontinuirane informacije. Objedinjuje skup metodologija i alata kojima se omogućuje obrada podataka iz različitih izvora. Cilj je pretvaranje tih podataka u informaciju potrebnu za donošenje poslovnih odluka te poboljšati pravodobnost i kvalitetu podataka o procesu donošenja odluka.

U ovom projektu prikazan je proces obrade skupa podataka kako bi pružili brz i lak pristup uvidima u trenutno stanje organizacije na temelju dostupnih podataka. Prvo odabiremo skup podataka i analiziramo ga kako bi definirali elemente po kojima ćemo stvoriti potrebne modele. Podatke prvo pohranjujemo u relacijsku bazu podataka što predstavlja izvor skladišta podataka. Zatim stvaramo dimenzijski model kojeg punimo s podacima iz transakcijskog sustava pomoću alata Pentaho. Na kraju stvaramo vizualizaciju podataka pomoću Microsoft Power BI.

Cilj projekta je napraviti skladište podataka koji je jedinstven, kompletan repozitorij podataka prikupljen iz raznih izvora i predstavljen krajnjem korisniku na razumljiv način.

# Odabir skupa podataka

Prvo je potrebno odabrati kvalitetan skup podataka koji zadovoljava određene zahtjeve. Skup podataka pronađen je preko web-a koji sadrži već gotove skupove podataka. Podaci su preuzeti u obliku CSV datoteke na linku: <https://www.kaggle.com/apoorvaappz/global-super-store-dataset>

Podaci predstavljaju detalje narudžbi koje su izvšili ljudi širom svijeta u vremenskom periodu od 1. siječnja 2011. do 31. prosinca 2014. godine.

## Analiza skupa podataka

Podaci su najvažniji element pri odabiru skupa podataka. Najbitnije karakteristike koje određuju spremnost podataka za izradu modela su: pouzdanost izvora, točnost sadržaja, potpunost, konzistentnost, itd. Podaci trebaju biti dovoljno veliki, raznoliki i različitih tipova kako bi mogli provesti kvalitetnu analizu. Skup podataka mora sadržavati dovoljno dimenzija te jednu vremensku dimenziju. Također skup podataka ne smije imati malo jedinstvenih vrijednosti i previše nepostojećih vrijednosti.

Podaci su analizirani pomoću web alata otvorenog koda Jupyter. Utvrdili smo da podaci imaju dovoljan broj podataka 51290 i sastoje se od 24 stupca. Podaci su različitih tipova, sastoje se od brojeva i stringova. Raznoliki su, imaju dimenzije od kojih je jedna vremenska dimenzija, pomoću kojih se promatra određena mjera poslovanja, npr. lokacije, vrijeme, troškovi, kategorije proizvoda, itd.

# Relacijski model podataka

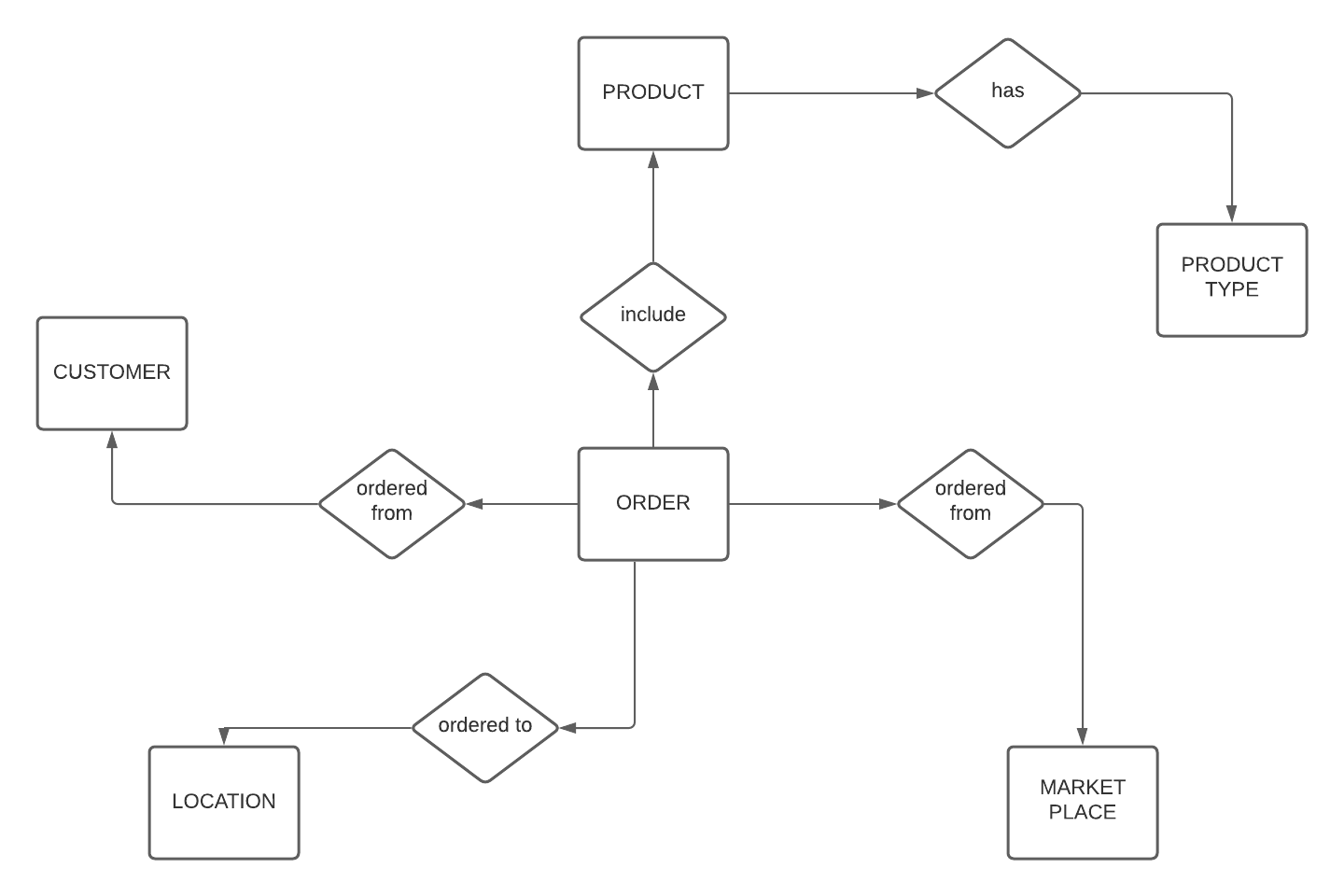
Nakon odabira i analize skupa podataka potrebno je napraviti relacijski model podataka koji je zapravo transakcijski model koji podržava dnevno poslovanje organizacije i odgovoran je za punjenje baze podataka sustava s podacima. On sadrži trenutne, detaljne, promjenjive podatke,

Prije korištenja baze podataka potrebno ju je dizajnirati da efikasno podržava poslovanje organizacije, identificirati tablice, atribute, ograničenja, itd. Rezultat dizajna je apstraktna reprezentacija stvarnosti koja isključuje mnoštvo detalja iz stvarnog svijeta kao što je smanjivanje kompleksnosti, fokusiranje na samo bitne detalje.

## Entity/Relationship model

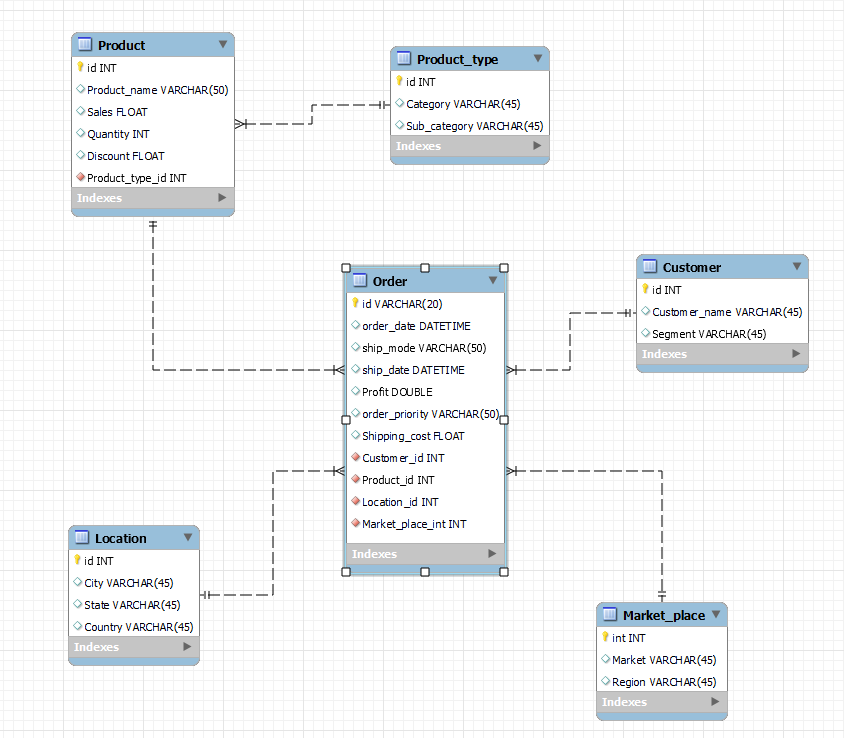
Entity/Relationship model se koristi za konceptualni dizajn. On daje konceptualni pregled nad bazama podataka. Potrebno je odrediti entitete koji predstavljaju objekte od interesa i na ER dijagramu su prikazani kao pravokutnici sa zaobljenim vrhovima, atribute koji predstavljaju činjenice i svojstva o nekom entitetu i prikazani su kao ovalni oblici, veze koje definiraju odnos između dva ili više entiteta koji su prikazani u rombu te kardinalnost veze.

Skup podataka podijeljen je u 6 relacija : order, customer, product, product\_type, location i market\_place. Središnja relacija će predstavljati tablicu činjenica u dimenzijskom modelu dim\_order i sadržava podatke o svim narudžbama, profit, poštarinu. Relacija customer sadrži podatke o kupcima, location podatke o lokacijama narudžbi, market\_place o lokacijama prodajnih mjesta, product sadrži podatke o proizvodima, njihovom nazivu, cijeni, naručenoj količini i product\_type podatke o kategorijama i pod kategorijama kojem pripadaju proizvodi.



*Slika 1. Prikaz jednostavnog ER modela*

Nakon izrade ER dijagrama dobivamo sliku cilja projekta. Pretvorit ćemo ga u oblik koji pruža pregledan cilj projekta.



*Slika 2. Prikaz ER modela*

Iz ER modela možemo vidjeti sve entitete, atribute i veze. Entitet „Order“ se sastoji od id-a, order\_date, ship\_date, ship\_mode, shipping\_cost, profit i order\_priority, detaljima narudžbe te od stranih ključeva: customer\_id, product\_id, location\_id, market\_place\_id. Vidimo da se narudžbe odvijaju s različitih prodajnih mjesta na različite lokacije diljem svijeta. Narudžbe se sastoje od proizvoda, njihove cijene, količine i određenog popusta te ostvarenog profita. Svaki proizvod ima svoju kategoriju i pod kategoriju kojoj pripada. Svaka narudžba ima svog kupca te određenu cijenu poštarine.

## Popunjavanje baze podataka podacima

U sljedećem koraku potrebno je popuniti bazu podataka na temelju izrađenog ER dijagrama. Koristit ću MySQL bazu podataka, besplatan sustav za upravljanje bazom podataka otvorenog koda. Kako bi napunili bazu podataka potrebno je napisati skriptu koja će puniti bazu s podacima. Koristila sam programski jezik Python preko kojeg sam punila bazu podataka s podacima iz CSV datoteke. Potrebno je uspostaviti konekciju na MySQL bazu preko Pythona pomoću pandas funkcije te učitati podatke iz CSV-a.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

*Slika 3. Konekcija na MySQL bazu i učitavanje CSV-a*

Nakon toga potrebno je kreirati tablice pomoću naredbe CREATE TABLE. Izradila sam 6 tablica u koje ću spremati podatke učitane iz CSV-a.

*Slika na kojoj se prikazuje tekst

Opis je automatski generiran*

*Slika 4. Kreiranje tablica*

Prikaz popunjenih tablica u MySQL-u:

Slika na kojoj se prikazuje stol

Opis je automatski generiran Slika na kojoj se prikazuje stol

Opis je automatski generiran

*Slika 5. Prikaz tablice customer Slika 6. Prikaz tablice location*

*Slika na kojoj se prikazuje tekst

Opis je automatski generiran*

*Slika 7. Prikaz tablice product*

Slika na kojoj se prikazuje stol

Opis je automatski generiran Slika na kojoj se prikazuje stol

Opis je automatski generiran

*Slika 8. Prikaz tablice product\_type Slika 9. Prikaz tablice market\_place*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 10. Prikaz tablice order*

# Dimenzijski model podataka

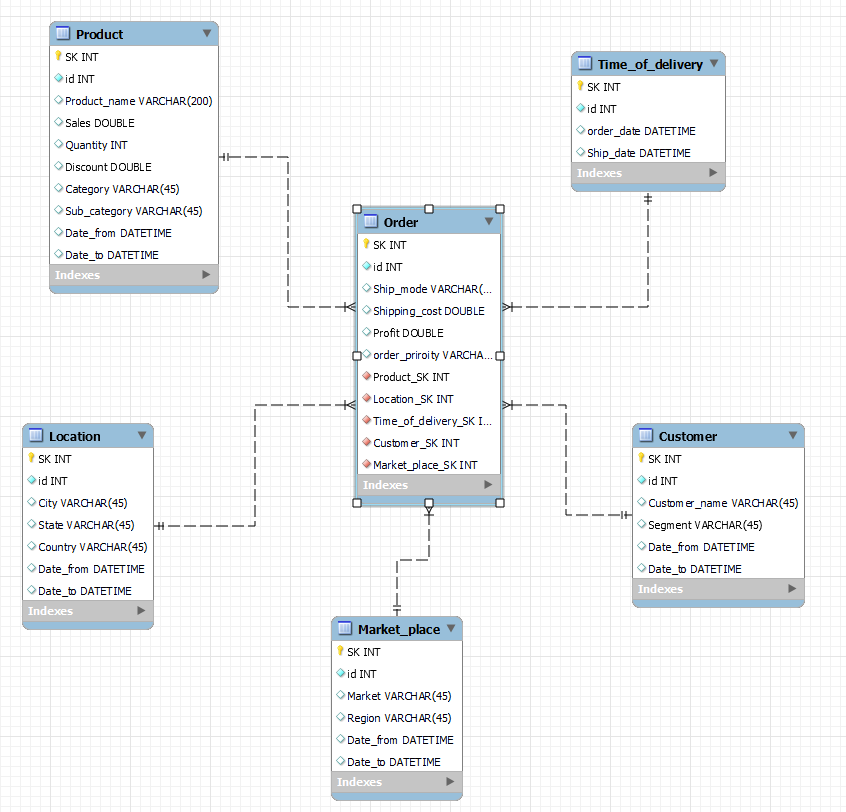
Nakon stvaranja i popunjavanja relacijskog modela potrebno je stvoriti i napuniti dimenzijski model. Dimenzijsko modeliranje omogućava razumljivost podataka poslovnim korisnicima i brze performanse upita. Sastoji se od tablice činjenica i dimenzija.

Tablica činjenica pohranjuje mjere poslovnog procesa koji modeliramo i nalazi se u središtu dimenzijskog modela. Sastoji se od velikog broja zapisa i ima cilj da sadrži najnižu razinu granularnosti. Sadrži surogat ključ koji predstavlja primarni ključ tablice. Često je normalizirana ili gotovo normalizirana te vrlo slična tablici u transakcijskom sustavu.

Dimenzijska tablica sadrži kvalitativan tip podataka. Manje su veličine od tablice činjenica. One su de-normalizirane zbog lakše obrade zahtjevnih i analitičkih upita i čitljivosti modela. Pridružene su tablici činjenica vezom 1:M. sadrži surogat ključ koji je primarni ključ tablice.

Na temelju izrađenog ER modela i popunjene baze podataka izradila sam dimenzijski model podataka. Temelji se na star shemi. Model se sastoji od tablice činjenica i 5 dimenzija. Tablica činjenica je „dim\_order“, a dimenzijske tablice su: „dim\_product“, „dim\_customer“, „dim\_location“, „dim\_market\_place“, „dim\_time\_of\_delivery“. Tablica činjenica nalazi se u sredini i povezana je sa dimenzijskim tablicama preko surogat ključeva. Tablica činjenica sadrži promatranu mjeru profit koja će se promatrati kroz dimenzije.

Koristila sam sporo mijenjajuće dimenzije tipa 2. One dodavaju atribute date\_from i date\_to kako bi se pratila aktivnost pojedinih zapisa. Omogućavaju praćenje povijesti promjena podataka i dodavanje novog zapisa.



*Slika 11. Prikaz dimenzijskog modela*

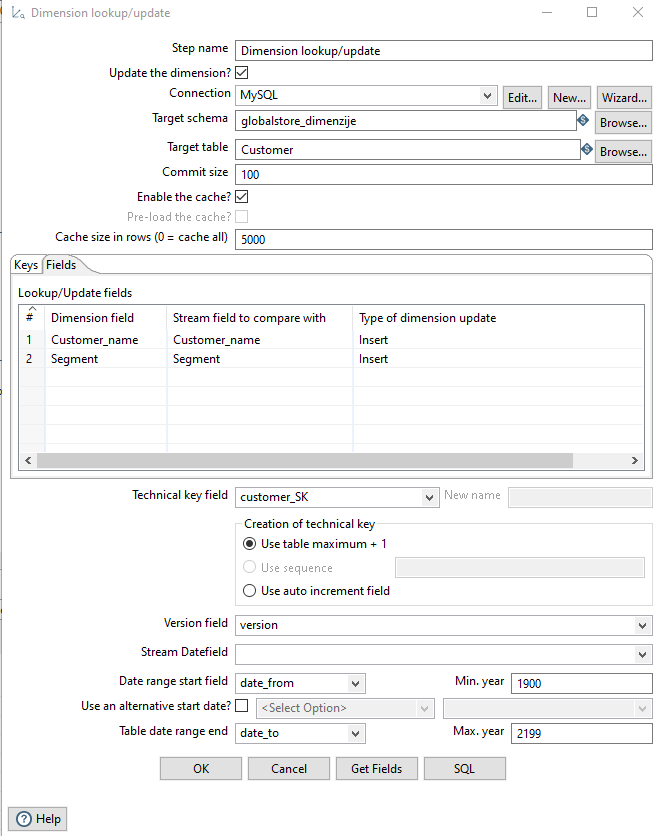
# ETL proces

Nakon izrade dimenzijskog modela podataka slijedi punjenje skladišta podataka sa podacima iz transakcijskog sustava. ETL proces odgovoran je za izdvajanje podataka iz transakcijskog sustava u pripremno područje, transformaciju podataka u strukturu za dimenzijski model i učitavanje podataka u skladište podataka.

Za ovaj proces koristit ćemo alat Pentaho Data Integration, softver za poslovnu inteligenciju koja pruža integraciju podataka, OLAP usluge, izvještavanje, mogućnost rudarenja podataka, izdvajanje, transformaciju i učitavanje.

## Stvaranje i punjenje dimenzije „dim\_customer“

Prva dimenzija je „dim\_customer“. Prvi korak je bio konekcija na MySQL bazu. Pomoću naredbe Table input dohvatila sam podatke iz baze. Nakon toga naredbom Sort rows sortirala sam podatke prema nazivu kupaca abecednim redoslijedom. Zatim sam učitala podatke iz CSV-a naredbom CSV file input te sam naredbom Select values uzela potrebne podatke. Također sam ih sortirala prem nazivu kupaca. Sljedeće sam koristila Unique rows kako bi eliminirala duplikate iz CSV-a. Naredbom Add sequence dodala sam id-eve za podatke koje dolaze iz CSV-a. naredbom Select values 2 sam provjerila jesu li podaci jednaki kao oni iz baze. Kako bi spojila podatke iz MySQL-a i CSV-a koristila sam naredbu Sorted merge. Uzela sam sve podatke kako bi prvo dobila podatke iz MySQL-a, a zatim iz CSV-a. Nakon spajanja podatka potrebno ih je još jednom sortirati po nazivu kupaca. Nakon toga, kako bi eliminirali duplikate koristila sam Unique rows 2 i tada su podaci spremni za unos u dimenzijsku tablicu, pomoću naredbe Dimension lookup/update stvorila sam dimenzijsku tablicu i napunila je podacima. Kreirala sam surogat ključ customer\_SK koji je ujedno i primarni ključ dimenzijske tablice.

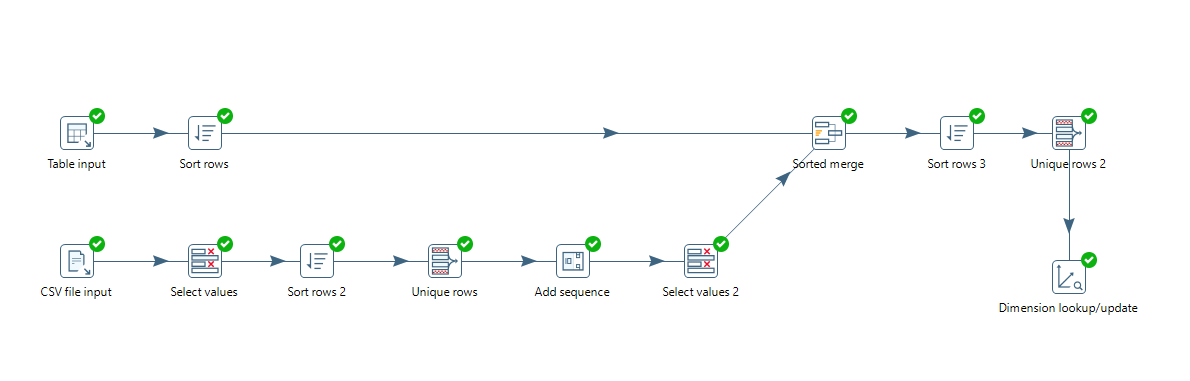


*Slika 12. Prikaz dimension lookup/update*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 13. Prikaz dimenzijske tablice dim\_customer*

**

*Slika14. Proces punjenja dimenzijske tablice dim\_customer*

## Stvaranje i punjenje dimenzije „dim\_location“

Nakon punjenja tablice „dim\_customer“ potrebno je napuniti dimenziju „dim\_location“. Postupak stvaranja i punjenja je jednak kao i prethodni proces. Potrebno je dohvatiti podatke iz MySQL-a naredbom Table input te ih sortirati prema nazivu grada pomoću naredbe Sort rows. Zatim je potrebno učitati podatke iz CSV-a naredbom CSV file input te odabrati potrebne podatke naredbom Select values. Sortiramo ih prema nazivu grada te uklanjamo duplikate naredbom Unique rows. Naredbom Add sequence dodajemo id-eve za podatke koji dolaze iz CSV-a. Podatke iz MySQL-a i CSV-a spajamo naredbom Sorted merge te ih sortiramo prema nazivu grada te uklanjamo duplikate naredbom Unique rows. Nakon toga podaci su spremni za unos u dimenzijsku tablicu naredbom Dimension lookup/update.

Slika na kojoj se prikazuje tekst

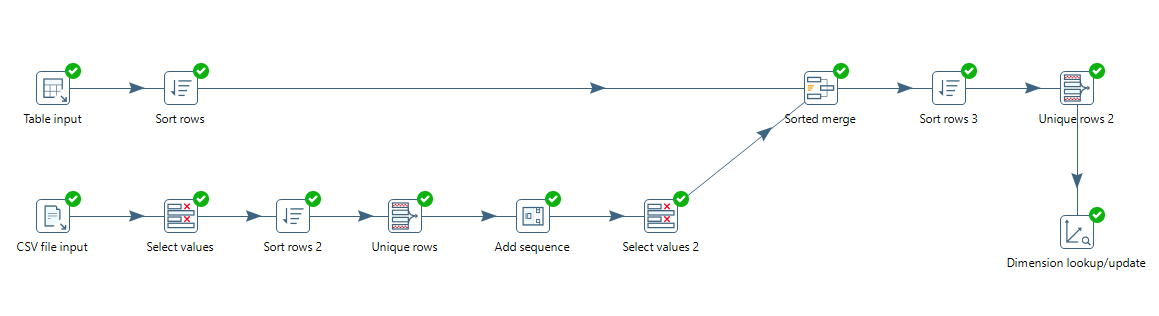
Opis je automatski generiran

*Slika 15. Prikaz dimension lookup/update*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

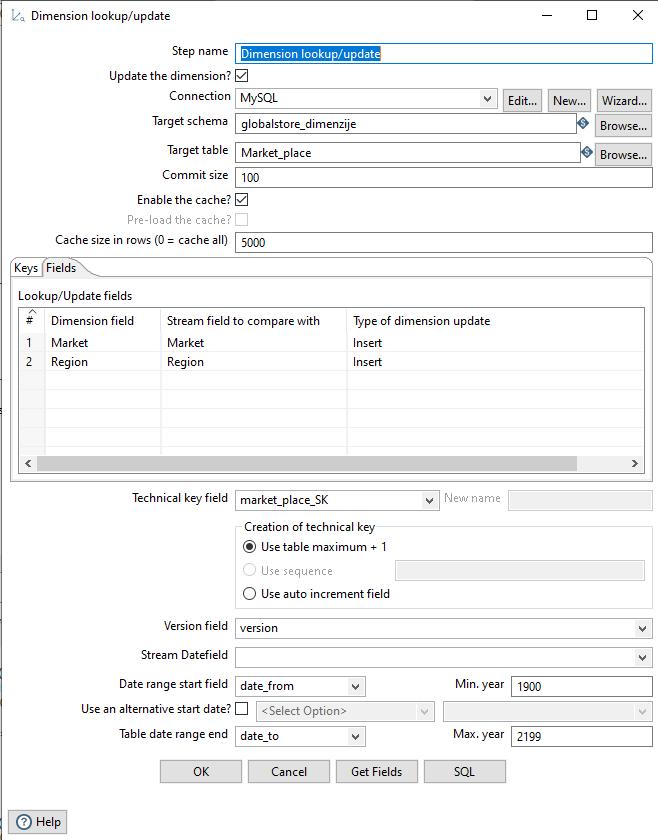
*Slika 16. Prikaz dimenzijske tablice sdim\_location*

**

*Slika 17. Proces punjenja dimenzijske tablice dim\_location*

## Stvaranje i punjenje dimenzije „market\_place“

Sljedeće je punjenje dimeznije market\_place s podacima iz MySQL-a i CSV-a. Postupak je jednak kao i za ostale dimenzijske tablice. Prvo dohvaćamo podatke iz baze podataka koji su nam potrebni i sortiramo ih po marketu. Zatim učitavamo podatke iz CSV-a te odabiremo potrebne podatke naredbom Select values i sortiramo ih po marketu. Potrebno je ukloniti podatke naredbom Unique rows te dodati sekvencu iz CSV-a. Zatim spajamo podatke iz CSV-a i MySQL-a naredbom Sorted merge. Potrebno je ponovno sortirati podatke te eliminirati duplikate koji su se pojavili tijekom spajanja. Na kraju učitavamo podatke u dimenzijsku tablicu te kreiramo surogat ključ „market\_place\_SK“.

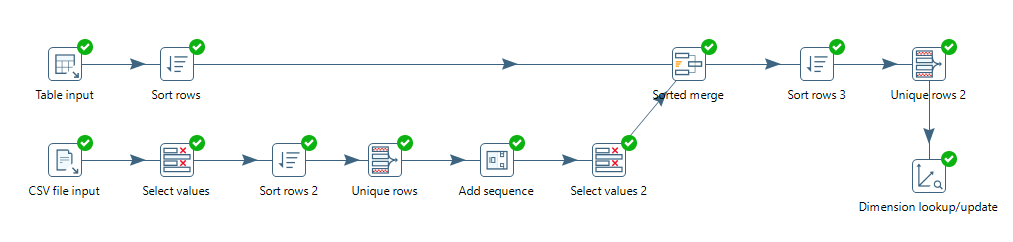


*Slika 18. Prikaz dimension lookup/update*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 19. Prikaz dimenzijske tablice dim\_market\_place*

**

*Slika 20. Proces punjenja dimenzijske tablice dim\_market\_place*

## Stvaranje i punjenje dimenzije „dim\_time\_of\_delivery“

Zatim popunjavamo vremensku dimenzijsku tablicu „dim\_time\_of\_delivery“. Proces punjenja i izdvajanja podataka je jednak kao i prošli. Prvo naredbom Table input dohvaćamo potrebne podatke iz baze te ih sortiramo prema datumima narudžbe. Učitavamo podatke iz CSV-a, uzimamo potrebne podatke te ih sortiramo prema datumima narudžbe. Dodajemo sekvencu id iz CSV-a. Podatke iz dva izvora spajamo u jedno naredbom Sorted merge nakon čega ih ponovno sortiramo te eliminiramo duplikate naredbom Unique rows 3. Zatim podatke učitavamo u bazu naredbom Table output.

Slika na kojoj se prikazuje stol

Opis je automatski generiran

*Slika 21. Prikaz dimenzijske tablice dim\_time\_of\_delivery*

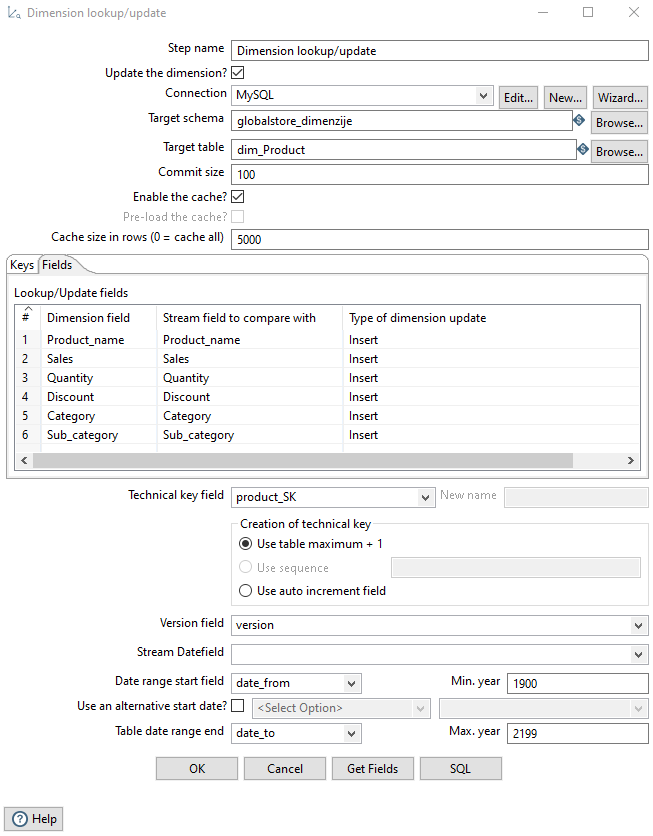
*Slika na kojoj se prikazuje tekst, nebo, karta, svijetlo

Opis je automatski generiran*

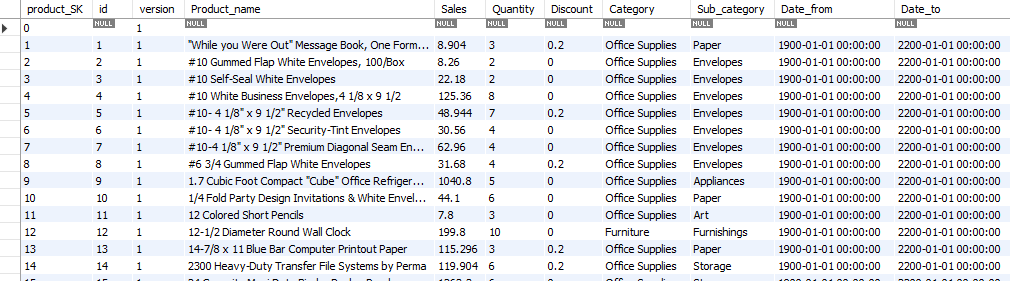
*Slika 22. Proces punjenja dimenzijske tablice dim\_time\_of\_delivery*

## Stvaranje i punjenje dimenzije „dim\_product“

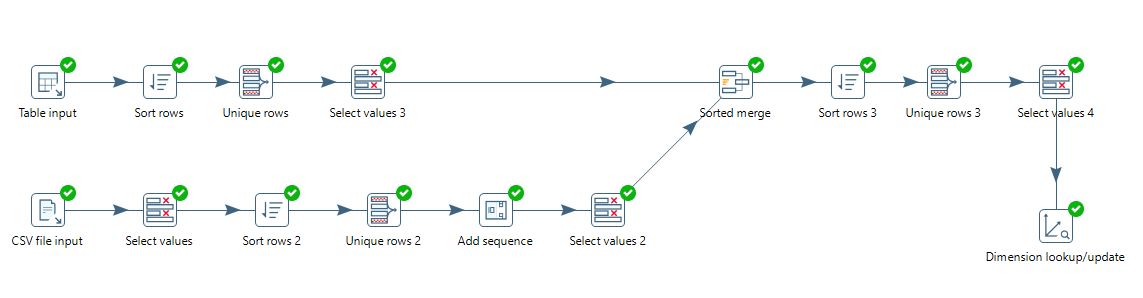
Posljednja dimenzija koju je potrebno puniti podacima je „dim\_product“. Radimo gotovo jednak proces kao i za ostale dimenzije. Prvo uzimamo podatke iz baze podataka naredbom Table input. Potom ih sortiramo prema nazivu proizvoda te uklanjamo duplikate. Učitavamo podatke iz CSV-a naredbom CSV file input te ih sortiramo prema nazivu proizvoda, eliminiramo duplikate te dodajemo id-eve iz CSV-a. Oba izvora podataka spajamo naredbom Sorted merge. Potrebno je ponovno sortirati podatke prema nazivu proizvoda te eliminirati duplikate koji su se pojavili tijekom spajanja podataka. Kada su podaci spremni za punjenje dimenzijske tablice koristimo naredbu Dimnesion lookup/update te kreiramo surogat ključ product\_SK.



*Slika 23. Prikaz dimension lookup/update*

**

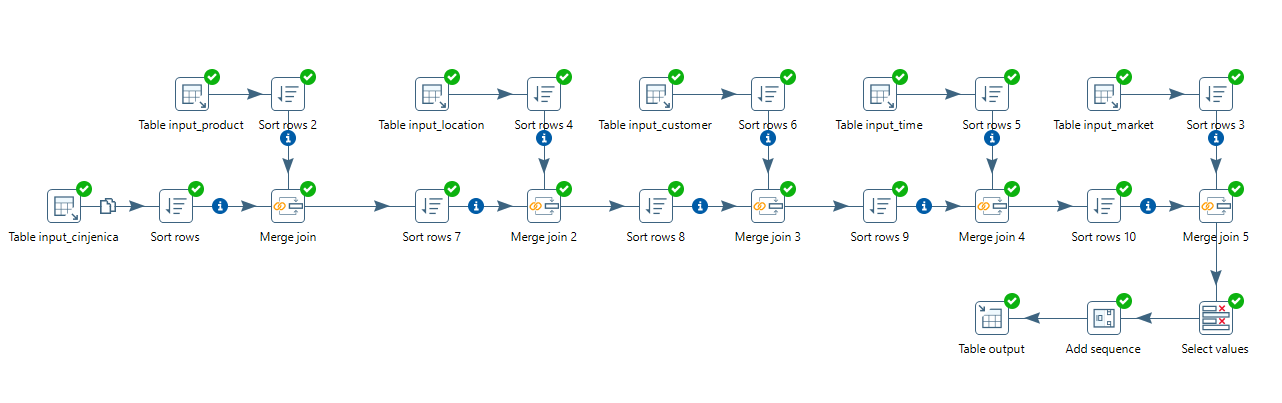
*Slika 24. Prikaz dimenzijske tablice dim\_product*

**

*Slika 25. Proces punjenja dimenzijske tablice dim\_product*

## Punjenje tablice činjenica

Najvažniji dio ETL procesa je punjenje tablice činjenica s podacima koji se nalaze u bazi podataka dimenzijskog modela tako što spaja stare strane ključeve iz relacijskog modela i uvodi novi strani ključ, surogat ključ iz dimenzija. Prvi korak je sortiranje relacije order po stranom ključu product\_SK, kao i relacija product koja se također sortira po primarnom ključu. Naredbom Merge join spajaju se u jednu tablicu. Taj proces ponavljamo za svaku drugu relaciju koji dodajemo u tablicu činjenica i spajamo ih Merge join-om. Na kraju naredom Select values uzimamo podatke koji su nam potrebni za tablicu činjenica te naredbom Table output punimo tablicu činjenica određenim podacima.



*Slika 26. Proces punjenja tablice činjenica dim\_order*

*Slika na kojoj se prikazuje stol

Opis je automatski generiran*

*Slika 27. Prikaz tablice činjenica dim\_order*

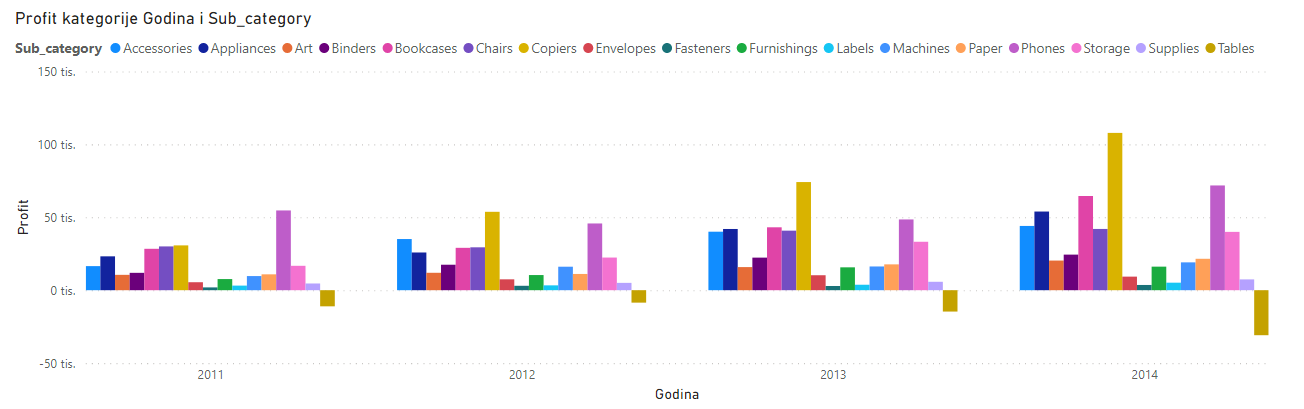
# Vizualizacija podataka

OLAP (OnLine Analytic Processing) je metoda koja podržava i pruža korisniku jednostavnu analizu multi-dimenzijskih podataka. Rezultati su dani u grafičkom obliku. Za definiranje OLAP izvješća potrebno je odabrati mjeru, dimenzije, definirati filtre i izgled.

Nakon stvaranja i punjenja dimenzijskog modela potrebno je instalirati alate za vizualizaciju podataka. U projektu je korišten alat Microsoft Power BI, usluga poslovne analitike tvrtke Microsoft. Cilj mu je pružiti interaktivne vizualizacije i mogućnost poslovne inteligencije s dovoljno jednostavnim sučeljem da krajnji korisnici mogu stvoriti vlastita izvješća.

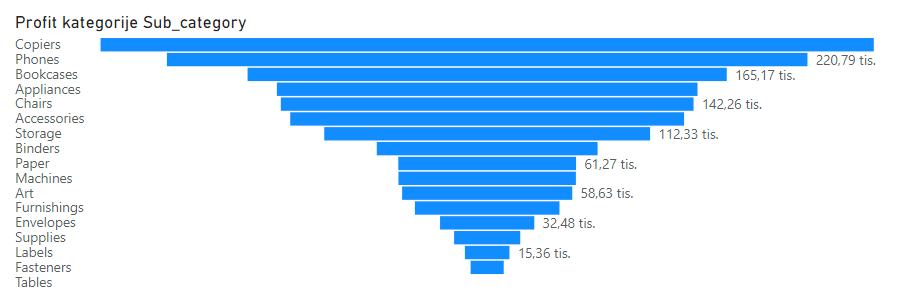
Nakon instalacije programa potrebno je uspostaviti konekciju MySQL-a i MicrosoftPower BI te odabrati sve tablice dimenzijskog modela.

Prvo sam prikazala podatke za zaradu od prodaje proizvoda po godinama u grupiranom stupčastom grafikonu. Kao os sam stavila godine, za legendu sam stavila pod kategorije proizvoda te kao vrijednost profit kao mjera iz tablice činjenica. Prikazuje ukupan profit po godinama po naručenim pod kategorijama proizvoda. Vidimo da 2011. godini mobiteli ostvaruju najveći profit, dok u 2014. godini kopirni strojevi.



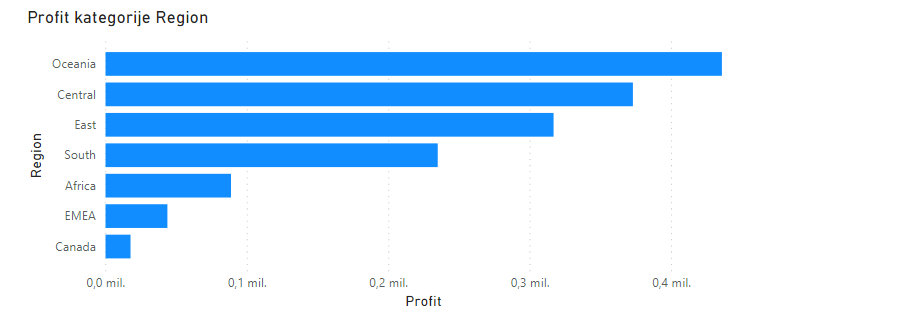
*Slika 28. Prikaz ukupno ostvarenog profita po pod kategorijama proizvoda po godinama*

U drugom grafu uzela sam vrijednost profit kao mjeru iz tablice činjenica te pod kategorije proizvoda. Vidljivo je da su kopirni strojevi ostvarili najveći profit, a stolovi najmanji. Ovaj graf prikazuje ukupan profit po naručenim proizvodima u vremenu od 2011-2014. godine.

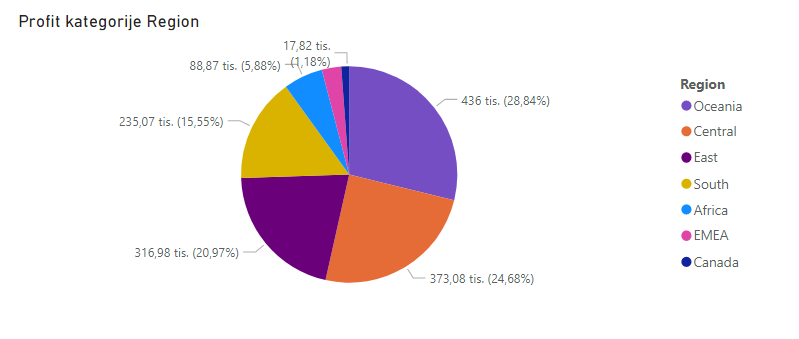


*Slika 29. Prikaz ukupno ostvarenog profita po pod kategorijama proizvoda*

Sljedeći graf prikazuje ukupan profit po prodajnim mjestima. Uzela sam profit kao vrijednost, promatranu mjeru iz tablice činjenica te regije prodajnih mjesta. Graf na slici 30. prikazuje isto što i graf na slici 29. Vidljivo je da prodajna mjesta u Oceaniji ostvaruju najveći profit, a Canada namjanji.

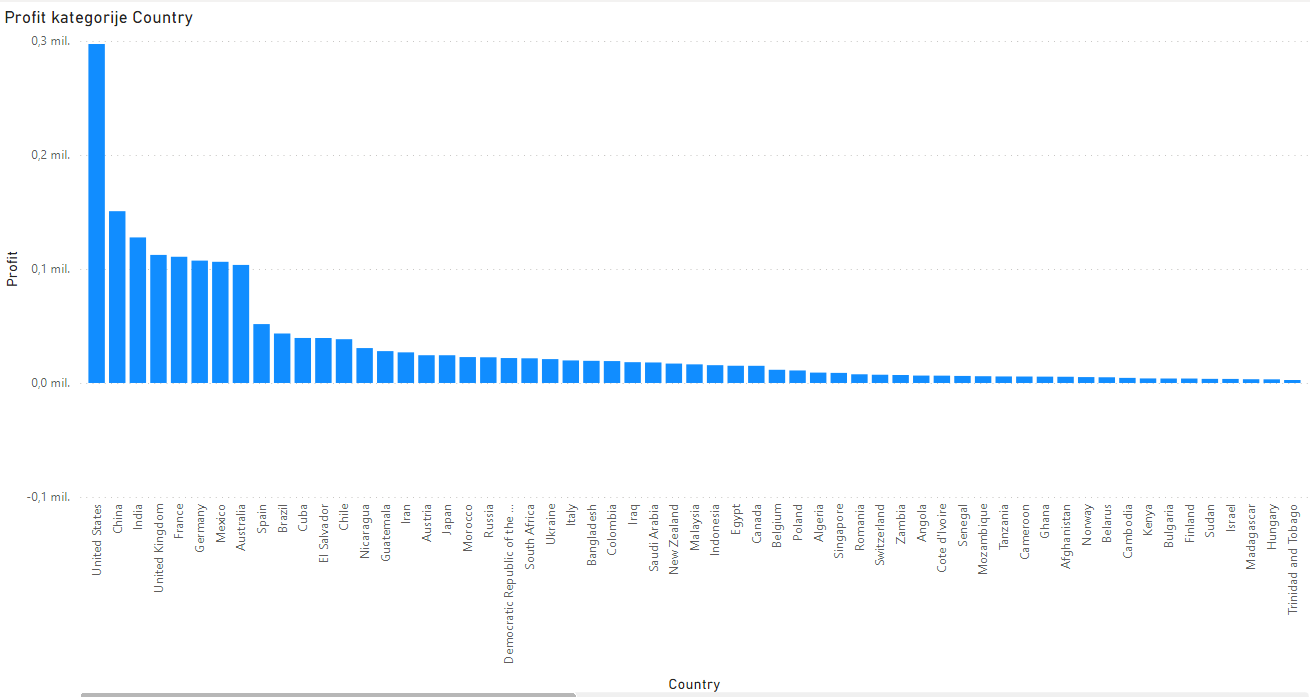


*Slika 30. Prikaz ukupno ostvarenog profita po prodajnim mjestima*

**

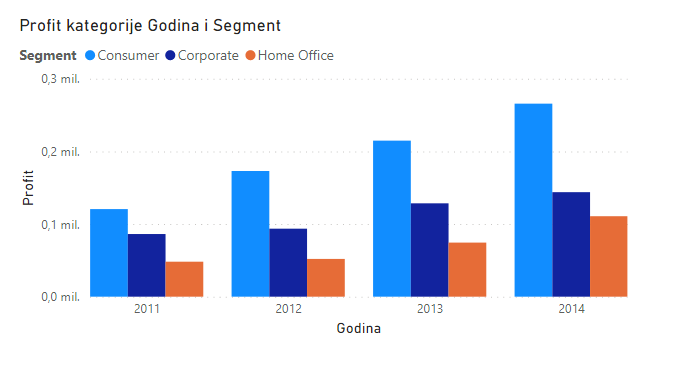
*Slika 31. Prikaz ukupno ostvarenog profita po prodajnim mjestima*

Graf prikazuje ostvareni profit po lokacijama narudžbi kupaca. Kao vrijednost uzet je profit iz tablice činjenica te države kupaca kao os. U grafu je vidljivo da United states stvara najveći profit.



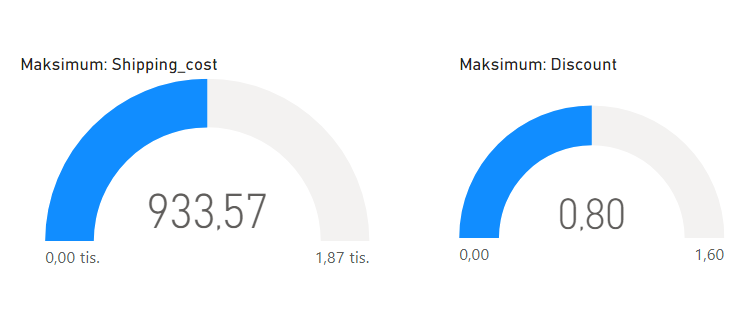
*Slika 32. Prikaz ostvarenog profita po lokacijama narudžbi kupaca*

Sljedeći graf prikazuje ostvareni profit po segmentima kupaca po godinama. Kao mjera uzet je profit iz tablice činjenica te segmenti kupaca. Vidljivo je kupci segmenta consumer su najisplativiji u svakoj godini.



*Slika 33. Prikaz ostvarenog profita po segmentima kupaca po godinama*

Na kraju sam napravila dva mjerila koji prikazuju najveću cijenu poštarine i najveći ostvareni popust.



*Slika 34. Prikaz najveće cijene poštarine i najveći ostvareni popust*

# Zaključak

Ovaj projekt temelji se na stvaranju skladišta podataka koji je jedinstven, kompletan repozitorij podataka prikupljen iz raznih izvora i predstavljen krajnjem korisniku na razumljiv način.

Proces stvaranja je prilično zahtjevan. Započeo je s odabirom skupa podataka s interneta koji su detaljno analizirani. Od analiziranih podatka stvoren je relacijski model podataka. Pomoću programskog jezika Python podaci su filtrirani kako bi popunili relacijsku bazu. Relacijski model pretvoren je u dimenzijski model koristeći ETL proces, izdvajanjem, transformiranjem i učitavanjem podataka. Alatom Pentaho Data Integration podaci su premljeni u dimenzijski model koji se sastoji od dimenzija i tablice činjenica. Na kraju, alatom Microsoft Power BI napravljena je vizualizacija podataka za potrebne analize na temelju kojih se donose važne poslovne odluke.

Rezultat mora biti jednostavan za korištenje poslodavcu te pružiti brz, lak pristup uvidima u trenutno stanje organizacije na temelju dostupnih podataka te lakše donošenje poslovnih odluka.

# Literatura

Pentaho 9.0, Online dokumentacija:

<https://help.pentaho.com/Documentation/9.0>

Kimball, Ralph, i Margy Ross. The Data Warehouse Toolkit. 3. izdanje. 2013

MIT software: „Općenito o Business Ineteligence sustavima“ <http://www.mit-software.hr/usluge/bi/bi1/>