Univerzitet u Novom Sadu

Fakultet Tehničkih Nauka

Inženjerstvo Informacionih Sistema

Projektovanje skladišta podataka

*Analiza restorana sa korisničkim kritikama*

Student: Pavle Šupeta Predmetni profesor: Sonja Ristić

IT61/2017 Predmetni asistent: Dajana Narandžić

**Sadržaj:**

# Zadatak i ciljevi projekta

Da bi kompanija bila uspešna u budućnosti, moraju da se donose dobre odluke. A za donošenje takvih odluka potrebno je da se sve relevantne informacije uzmu u razmatranje. Najbolji izvor za sve informacije je dobro konstruisan Data Warehouse.

DataWarehouse je najbolja opcija danas jer integriše razne izvore podataka, pruža ogroman prostor, bezbedan je i omogućava pretraživanje u real-time vremenu.

Cilj projekta je da na osnovu analize podataka i izveštaja o podacima mogu doneti odluke koje bi bile od pomoći za poslovanje.

# Opis postupaka projektovanja DataWarehouse sistema



Slika 1 Postupak razvoja DW Sistema

Istraživanje podataka – prilikom istraživanja podataka tražio sam odgovarajući dataset koji bi bio pogodan za izradu projekta. Dataset koji sam izabrao je baziran na velikoj količini podataka koja je skupljena na Fakultetu Prirodnih nauka u Meksiku, dataset je baziran na podacima o restoranima i kritikama od strane korisnika

Obrada podataka – prilikom obrade podataka morao sam da filtriram podatke, uredim kolone, kao i da izbacim neke nepotrebne podatke koji nisu bili u skladu sa temom projekta.

Definisanje ciljeva – uzeta su u obzir pitanja od strane korisnika, i naspram njih su ustanovljeni ciljevi za izradu projekta.

Izrada ETL procesa – Koristeći SSIS alat, uz pomoć ETL procesa kreirao sam 9 tabela dimenzija iz izvornih podataka kao i tabelu činjenica koja upotpunjava zvezdastu šemu samog DW-a

Izrada izveštaja – Za izradu izveštaja koristio sam SSRS dodatak za VS, uz pomoću kojeg sam korisnički definisana pitanja preveo u vizualno prikazane odgovre sa detaljnom reprezentacijom podataka i tokova promena istih

Analiza rezultata – Tokom same analize rezultata ustanovljava se uspešnost pri davanju odgovora na korisničke zahteve, kao i moguća ekspanzija i nadogradnja samog sistema.

# Specifikacija zahteva korisnika

Korisnik je kreirao zahteve i pitanja na koja treba dati odgovore putem izveštaja. Na sva korisnikova pitanja uspešno su dati odgovori.

1. Kako izgleda odnos restorana prema visini cena?
2. Koliko ima gostiju koji su posetili porodične restorane?

-Koja je omiljena kuhinja gostiju koji posećuju porodične restorane?

1. Kako su korisnici ocenili restorane meksičke kuhinje?
2. Analiza korisnika:

-da li konzumiraju alkohol?

-da li su pušači?

-koliko imaju godina?

-njihovi najposećeniji restorani i ocene za te restorane.

1. Koliko restorani dobijaju kritika u odnosu na vreme?

# Specifikacija modela

## Specifikacija izvora podataka

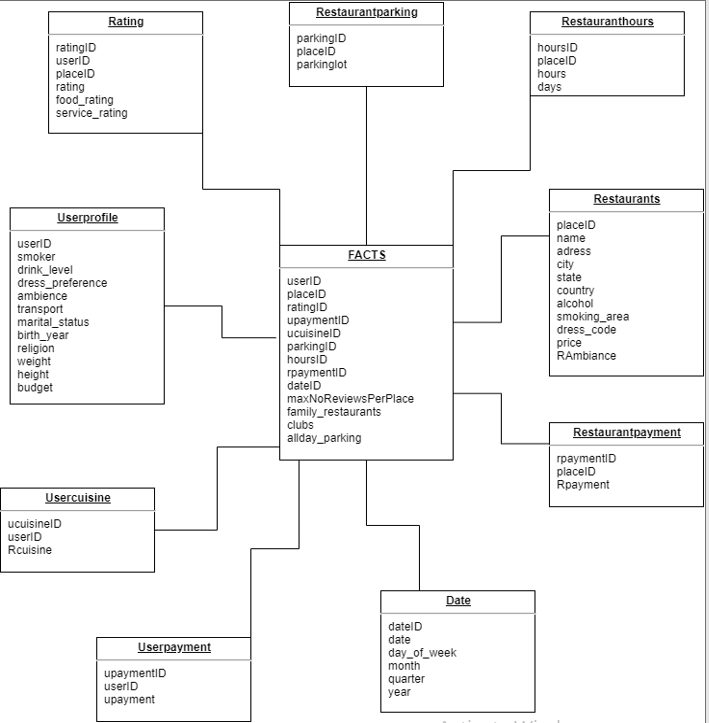
Podaci su preuzeti sa sajta [Kaggle](https://www.kaggle.com/uciml/restaurant-data-with-consumer-ratings). Podaci su sačuvani u 9 csv fajlova, koji su nakon obrade i analize pretvoreni u 9 tabela dimenzija i tabelu činjenica.

Obeležja od značaja koja se nalaze u csv fajlovima su sledeća:

* UserID - jedinstveni identifikator korisnika
* Smoker - bool vrednost na pitanje da li je korisnik pušač
* Drink\_level – nivo korisnikovog konzumiranja alkohola
* Dress\_preference – preferirani način oblačenja korisnika
* Ambience – preferirani ambijent korisnika
* Transport – prevozno sredstvo koje korisnik ima/koristi
* Marital\_status – korisničko bračno pitanje
* Birth\_year - godina rodjenja
* Weight - težina
* Height – visina
* UcuisineID - jedinstveni identifikator kuhinje vezane za korisnika
* Rcuisine(user) - omiljena kuhinja korisnika
* UpaymentID – jedinstveni identifikator načina plaćanja vezanog za korisnika
* Upayment - način plaćanja
* RatingID – jedinstveni identifikator rejtinga
* Rating – celokupna ocena(0 1 2)
* Food-rating – ocena hrane
* Service-rating – ocena usluge
* PlaceID – jedinstveni identifikator restorana
* Name – naziv restorana
* Adress – adresa
* City – grad
* Country – država
* Alcohol – mogućnost prodaje alkohola
* Smoking\_area – da li postoji prostor za pušače (bool)
* Price – nivo cena ( low medium high)
* HoursID – jedinstveni identifikator radnog vremena
* Days – radni dani
* Hours – radni sati
* Rpayment – načini plaćanja koje restoran prima
* RcuisineID – jedinstveni identifikator kuhinje koju restoran sluzi
* Rcuisine - kuhinje koju restoran sluzi

## Specifikacija ciljnog DataWarehouse sistema

Na osnovu svih podataka koji su priloženi sklopljena je sledeća OLAP šema



Slika 2 OLAP šema

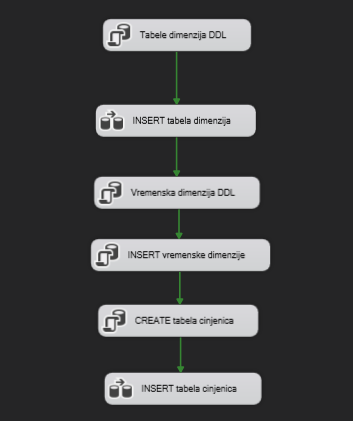
### Specifikacija zahtevanih dimenzija

Identifikovo je 9 tabela dimenzija uz dodatnu vremensku tabelu dimenzija i one su:

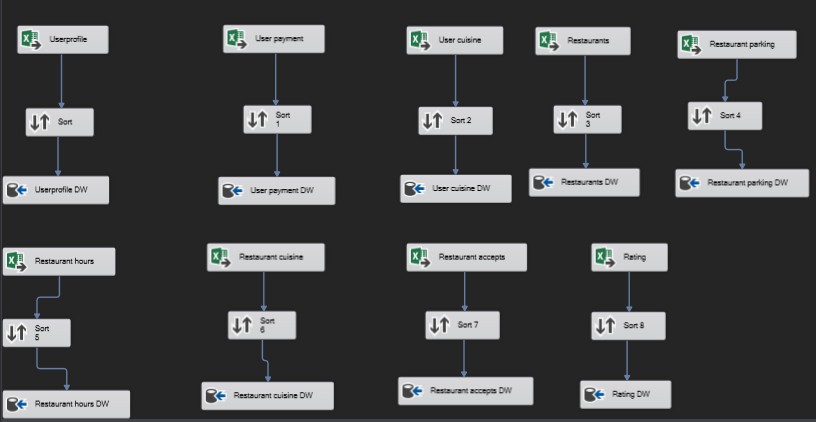
1. **Userprofile -** predstavlja tabelu koja sadrži sve informacije o korisniku
2. **Usercuisine –** predstavlja informacije o tipovima kuhinje koje odredjeni korisnik preferira
3. **Userpayment -** predstavlja podatke o nacinu placanja za odredjenog korisnika
4. **Date –** vremenska dimenzija koja sadrzi podatke o datumima, danima, kvartalima I godinama
5. **Restaurants** - tabela koja sadrzi portrebne informacije o restoranima
6. **Restaurantcuisine –** tabela koja sadrzi podatke o stilovima spremanja u odredjenom restoranu
7. **Restaurantparking -** informacije o tipu I postojanju parking mesta
8. **Restauranthours –** podaci o radnom vremenu restorana
9. **Restourantpayment –** dozvoljeni nacini placanja u restoranu
10. **Rating –** sadrzi infromacije o ocenama I datumu ocenjivanja
11. Specifikacija zahtevanih mera

# Opis ETL procesa

Pod ETL procesima spada kreiranje svih tabela dimenzija i kreiranje tabele činjenica.



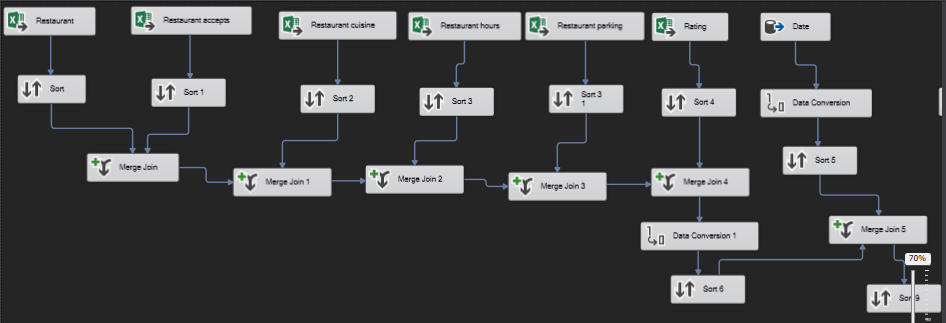
Slika 3 ETL procesi



\*PRAVLJENJE TABELE DIMENZIJA\*



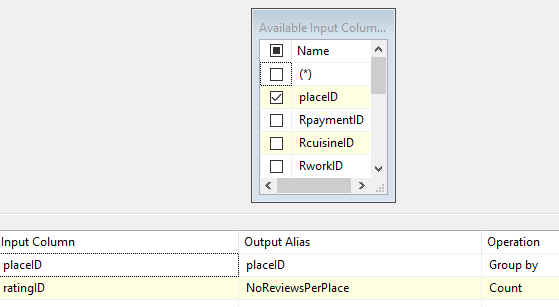
\*kreiranje vremenske dimenzije\*



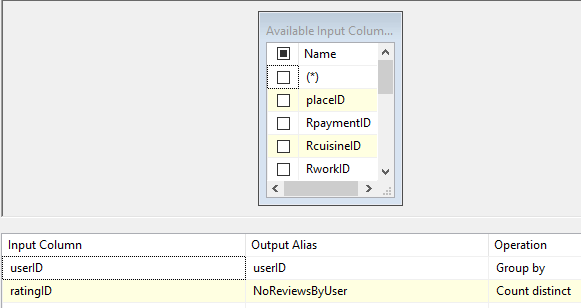


U tabeli cinjenica izradjene su dve dodatne mere za olaksavanje izrade izvestaja a to su:

1) NoReviewsPerPlace – Broj kritika za svaki restoran



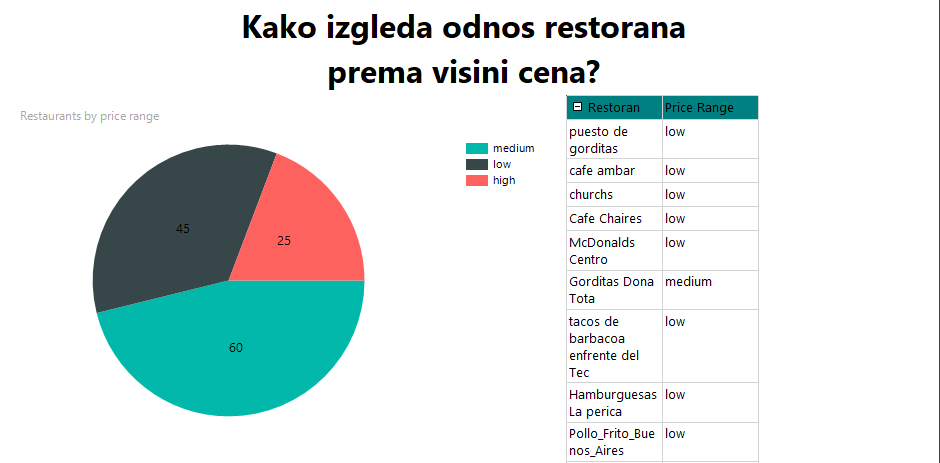
2) NoReviewsByUser – broj kritika ostavljen od strane svakog korisnika



Celokupna integracija ETL procesa, dodavanje dimenzija kao i mera je prošla bez većih problema osim jednog. Naime zbog obimnosti podataka, upis u bazu se ne može izvršiti metodom FAST LOAD. Ovaj problem se može rešiti jednostavnom implementacijom Row Sampling alata, ali to nije primenjen zbog prikaza maksimalne količine podataka.

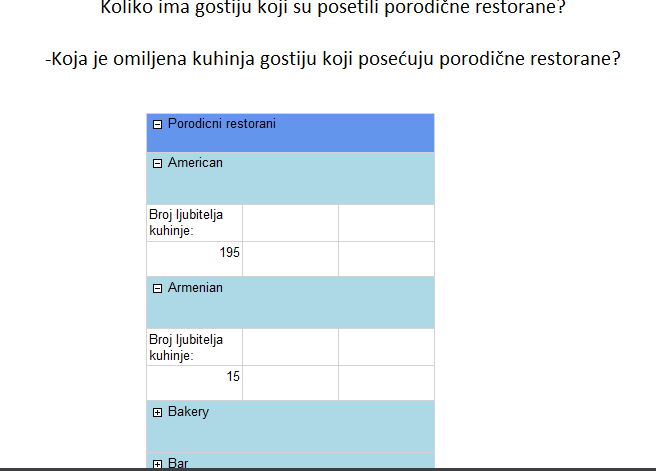
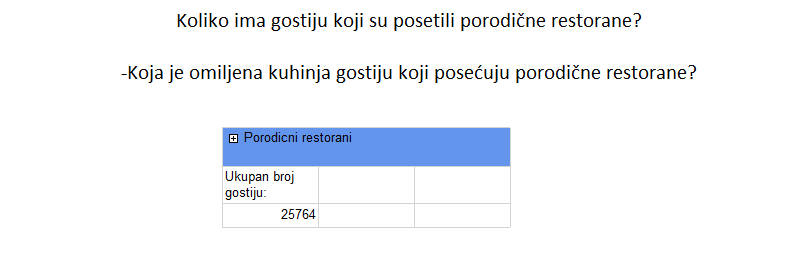
Prikaz izveštaja kao odgovor na zahteve koji su postavljeni od strane korisnika

Kako izgleda odnos restorana prema visini cena?

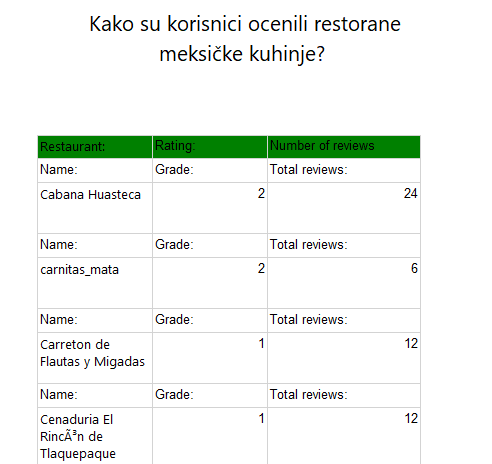


Koliko ima gostiju koji su posetili porodične restorane?

-Koja je omiljena kuhinja gostiju koji posećuju porodične restorane?



Kako su korisnici ocenili restorane meksičke kuhinje?



Analiza korisnika:

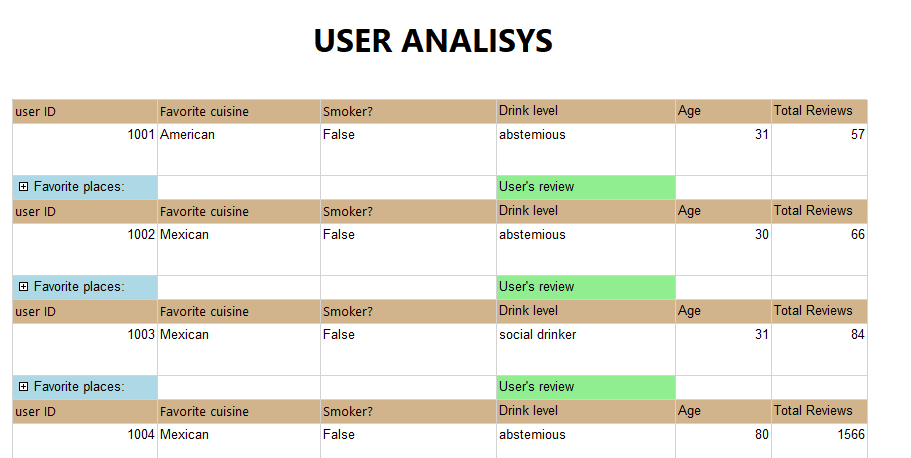
-da li konzumiraju alkohol?

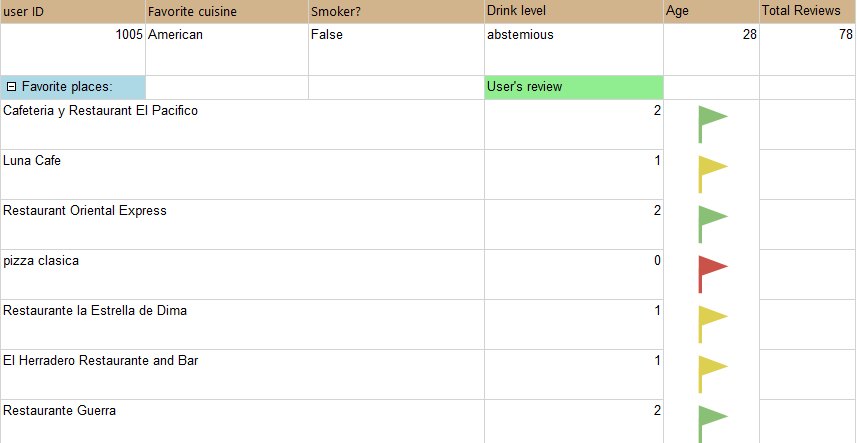
-da li su pušači?

-koliko imaju godina?

-njihovi najposećeniji restorani i ocene za te restorane.

Koliko restorani dobijaju kritika u odnosu na vreme?





# Zaključak

Celokupna izrada projekta je bila jako interesantna, činjenica da smo sami birali svoje podatke kao da smo i sami morali da ih razumemo i adaptiramo svojim potrebama je bila u isto vreme olakšavajuća koliko i otežavajuća. Najveći problem sa kojim sam se ja našao jeste nepotpuna kombinacija podataka, uz to nedostatak nekih možda ključnijih podataka kao što su imena korisnika. Uz sve to, proširio sam svoje znanje u Excelu!   
Kad uzmemo u obzir proširenje sistema, obavezno je naglasiti da je potrebno filtriranje podataka, kao i odgovarajuća segmentacija istih. Na primer, potrebno je jasnije definisati korisnike, njihove atribute i veze sa ostalim tabelama. Potrebno je uvesti vremensko ograničenje za tabelu rejting(u ovom slučaju sam ja ručno to radio). Možda proširiti bazu podataka dodanim podacima o restoranima i korisnicima van centralne Amerike. Ove mere, za početak, mogu olakšati ali i proširiti opseg podataka što će povećati zadovoljstvo korisnika, kao i samih developera.