Data Warehouse Project

This Data warehouse will serve as a tool to analyze correlation between Yelp ratings and weather, with focus on finding dependencies between certain weather conditions and trends in places rating. This Data project will cover analysis of temperature, conditions and e.g. humidity impact and how it affects user’s tendency to give recommendations or stars on Yelp rating. Other aspect of analysis will contain checking in what weather circumstances grades and reviews were more frequent, therefore if they have impact on customers’ decisions to visit places. These will be vital for businesses in making decisions.

GROUP WORK

Project Topic:

Weather analysis in relation to how it affects customers review of business, and patronization frequency.

Characteristics of the problem area

Customers or users of business give reviews and recommendation on Yelp, after they must have used a service. There are factors that affects customers perceptions, and weather is one of them. For example, customers might give a poor review when it is raining and a good review when it is sunny, this might be as a result of their mood or facilities are not put in place to satisfy customers needs.

Description of Area of Analysis

Data available from Yelp includes; Review, Users, Business, and Check-in data. We also have corresponding weather condition data for selected locations used in this Yelp dataset at specific time. These data will help use analyze which weather condition is most suitable for specific business based on customers reaction.

Purpose of the Project:

Expectations:

Hourly weather conditions at specific cities, for specific business and corresponding reviews.

The scope of the analysis:

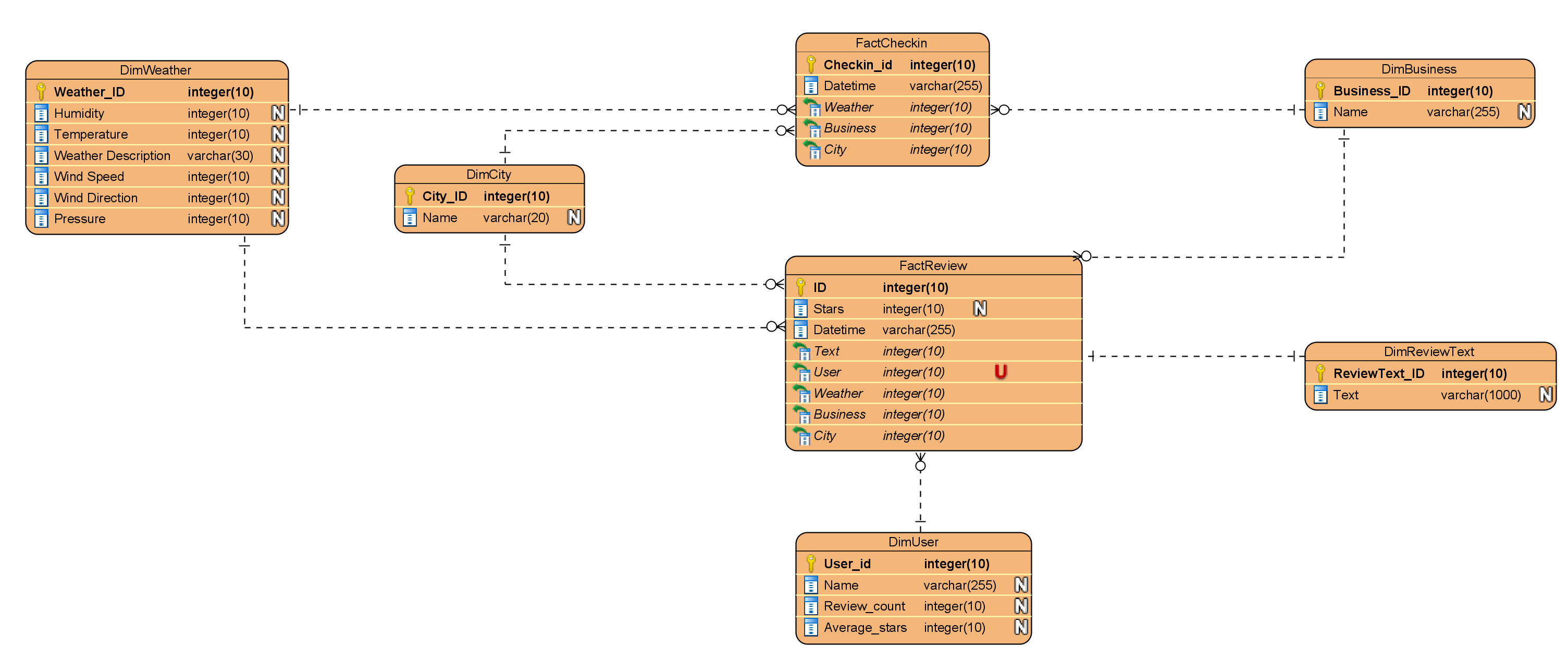
1. Weather Dimensions contains Humidity, Temperature, Weather Description, Wind speed, Wind Direction, and Pressure
2. City Dimensions; City name
3. Business Dimensions; Business name
4. Review Text Dimensions; Text
5. User Dimensions; Name Review Counts and average star rating
6. Review Facts; Stars and Datetime, with text, user, weather and business
7. Check-in Facts; datetime with weather business and City.

1.6. Data sources (location, format, availability)

The Yelp Business Dataset, Check-in Dataset, Review Dataset and User Dataset are all academic dataset made available for use on Kaggle. The data comes from the website [Yelp Dataset | Kaggle](https://www.kaggle.com/yelp-dataset/yelp-dataset). These are on JSON format and can be downloaded directly with no data license information on the specified page.

Datasets for Historical weather Data spanning from 2012 to 2017 was also gotten from Kaggle, the dataset contains approximately 5 years of high temporal resolution (hourly measurements) data of various weather attributes, such as temperature, humidity, air pressure. This data is available for 30 US and Canadian Cities, as well as 6 Israeli cities. Available here [Historical Hourly Weather Data 2012-2017 | Kaggle](https://www.kaggle.com/selfishgene/historical-hourly-weather-data). These are on JSON format and can be downloaded directly with no data license information on the specified page.

Entity Relationship (ER) Diagram:



Here we have five (5) dimensions and 2 facts table, leading us to a fact constellation schema.

There are also three (3) common dimensions between these facts.

The facts were chosen since they contain appropriate measures which are needed to get facts where the dimensions answers question of “where”, “who”, and “what”.

DimWeather: This Entity contains atmospheric condition of a place at a given time, it is made up of Humidity, temperature, weather description, wind speed, wind direction and pressure, all gotten from the same weather dataset with a unique combination.

DimCity: This Entity covers selected cities with unique IDs. The selected cities are also common in both weather dataset and Dataset made available on Yelp.

DimBusiness: DimBusiness Entity covers all business reviewed buy the customers and also used by customers in corresponding cities listed on DimCity entity, these data are from Yelp. It contains both the business ID and Name of the business as well.

DimReview: The DimReview entity contains text which are comments made by customers, these comments are vital as one can connote customers feeling from documents, which must have led to how each customer rates the business.

DimUser: Here we see each business user, entity contains name, Review counts, average stars and user identity. These are ones providing the comments and the ratings under each weather condition.

FactCheckin: This is a fact entity used in this model, it is linked with three dimension entities and tends to be used to determine the frequency of usage for each business, it also shows how popular businesses are. These entities are DimWeather, DimBusiness and DimCity. Asides from the three foreign keys, it contains datetime measures and Checkin\_id.

FactReview: This FactReview Entity contains all entities on the list with three common entities with the FactCheckin entity. Other measures include Star Ratings for each business by each user, and the Datetime. This entity has to do with everything associated with user’s review.

ETL PROCESS

In this stage dataset from data sources where transformed using Kettle Pentaho.

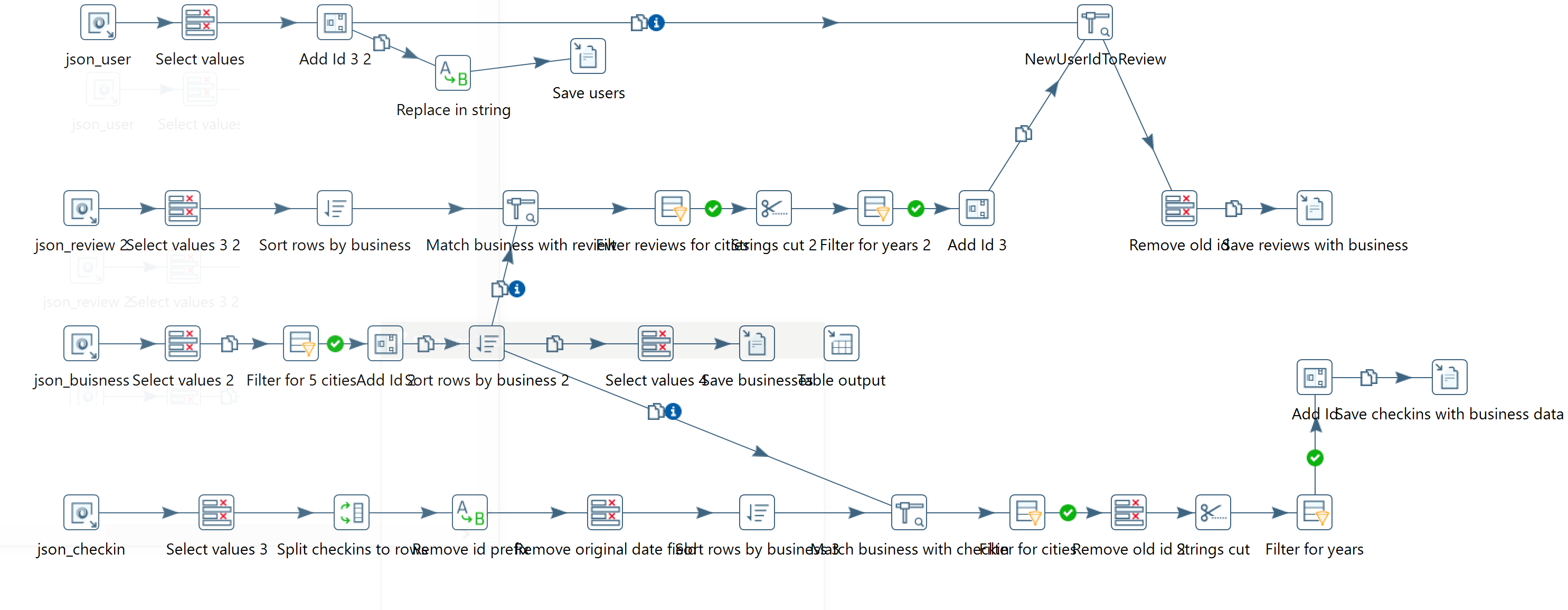
First major Extraction Loading and transformation involved the user data, the review data, the business data and checkin datasets. Essential values where selected from the User datasets, surrogate key was added, and texts where transformed to strings the exported to csv file which will be used in the database.

Same was also repeated for the Review file, where specific values form the dataset where selected and sorted by businesses, then matched also by business, reviews were also separated for each city, then tailored to get specific time range (2 years), finally surrogate keys were added as primary key for each rows.

The business datasets, some attributes were selected, it was then filtered for 5 cities and surrogate keys were added, then each of these businesses were matched with corresponding review, finally format was changed to csv for database use.

The Checkin dataset, needed attributes where selected, prefix attached to each ID was removed, rows where sorted by the business and then these checkin were filtered according to city. Primary keys were added and then exported to csv file format.

Below are processes in detail:

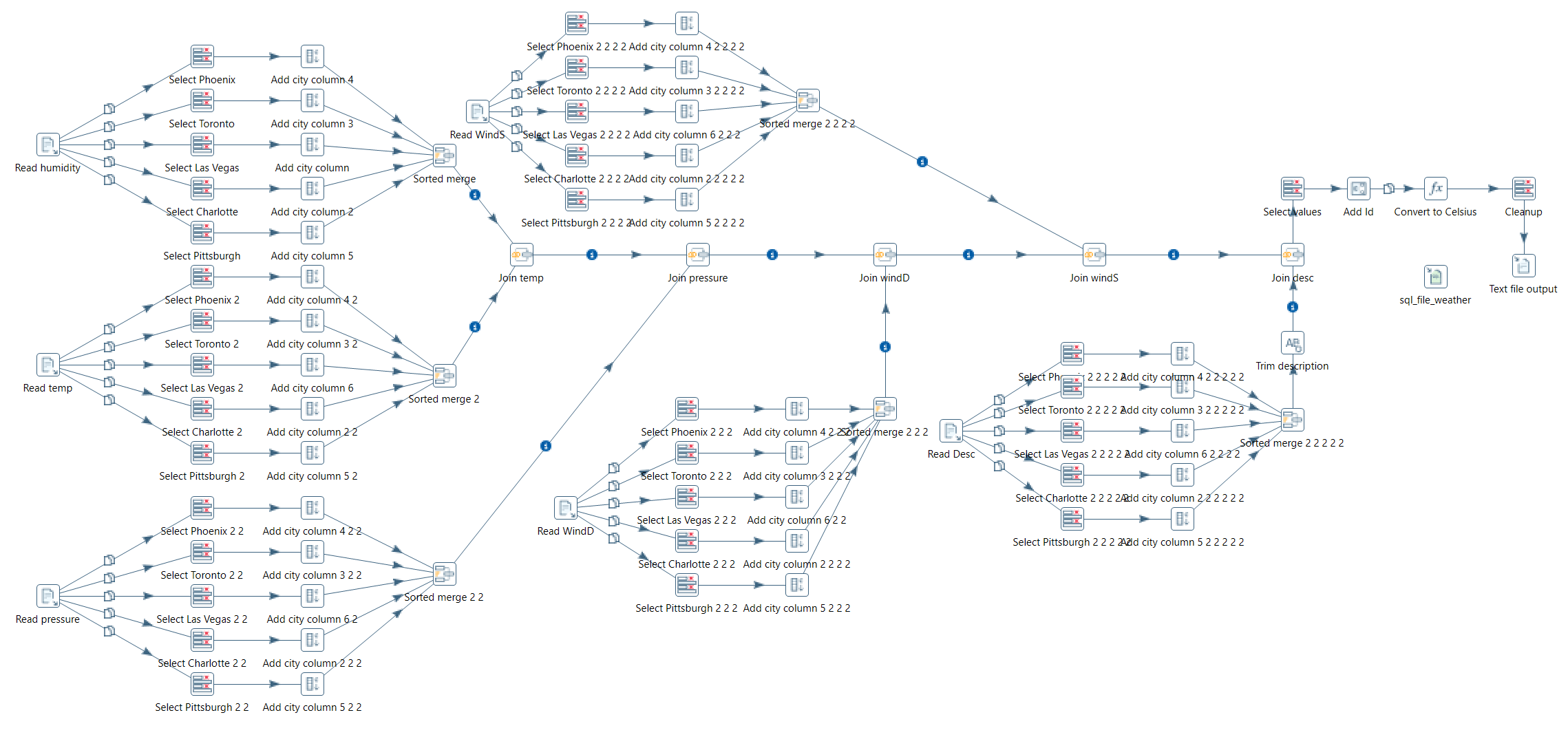


ETL Process for Weather Datasets:

Weather datasets are already in csv form, and then each atmospheric condition was needed to be extracted per city. Phoenix, Toronto, Pittsburgh, Las-Vegas and Charlotte were selected.

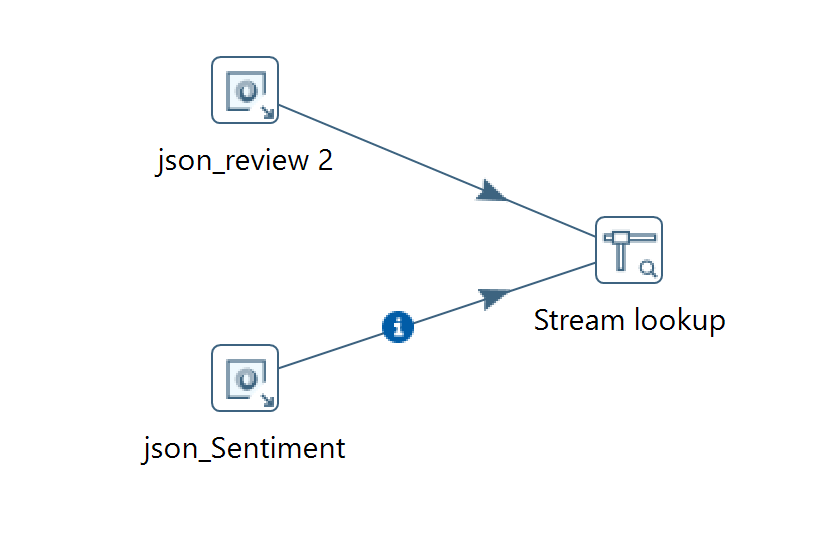
Weather attributes humidity, temperature, pressure, wind speed and direction, then weather description were selected.

In the ETL process each weather attribute where first selected per city, then rows were transformed to columns, these was finally sorted and merged, then merged with every other weather attribute. these values where selected and primary keys added. Later on, it was converted to Celsius to ensure uniformity.



Review Sentiment Analysis ETL process:

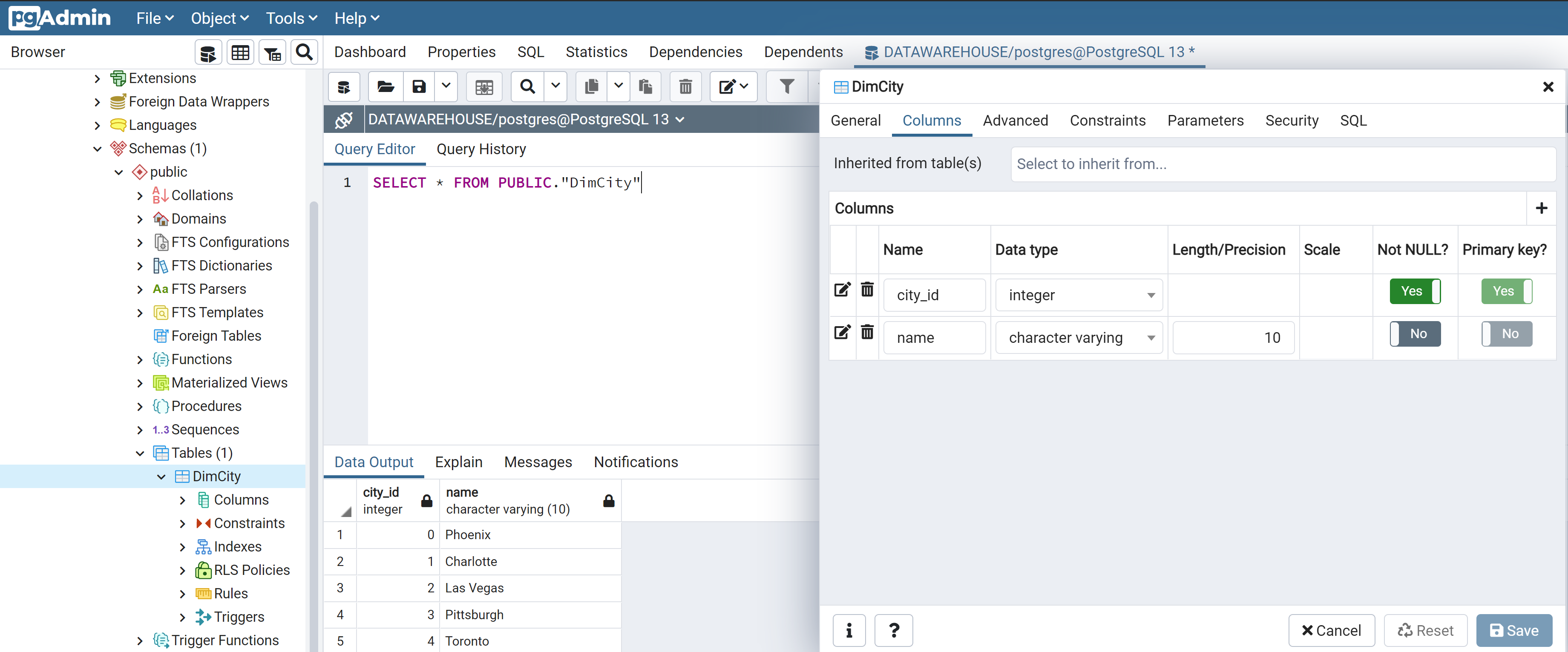
Comments made by users were needed to be reviewed, this will need some text extraction sentiment program. This program was used for this process [CoreNLP](https://stanfordnlp.github.io/CoreNLP/index.html?fbclid=IwAR2U8y5s3zCcptoBT6mH7X7SoqtCIz8imk82RRrJpHYFjSYLnFA58P_oYEQ). This was then merged with JSON review file alongside sentiment file. Here Rows are being sent to step [Stream lookup] so that they can be used as an additional information. That target step specificially reads from step [json\_sentiment] and it is treaded as a special case.



DATABASE

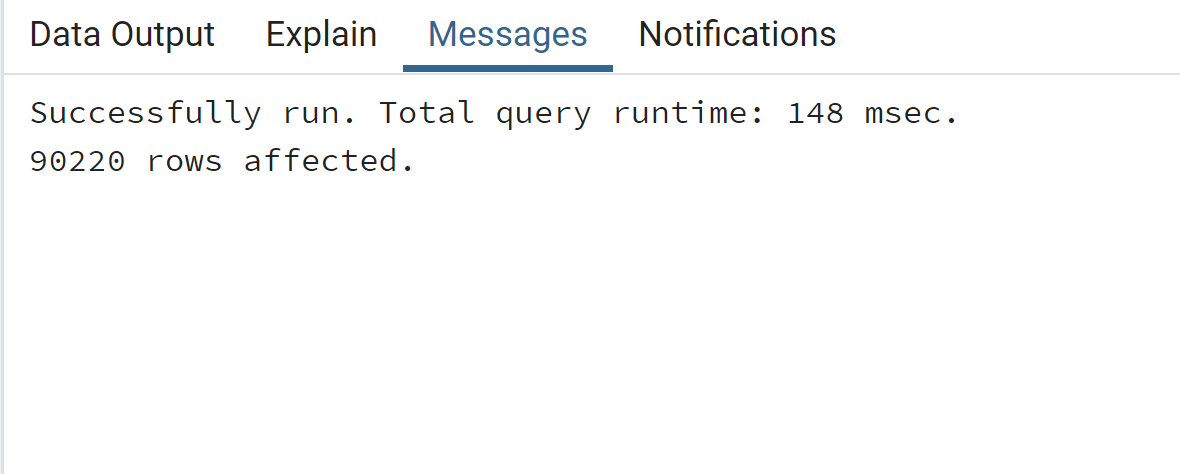
Here Postgres is used to query and add constraints to data. Each of the tabled previously described will be available for query.

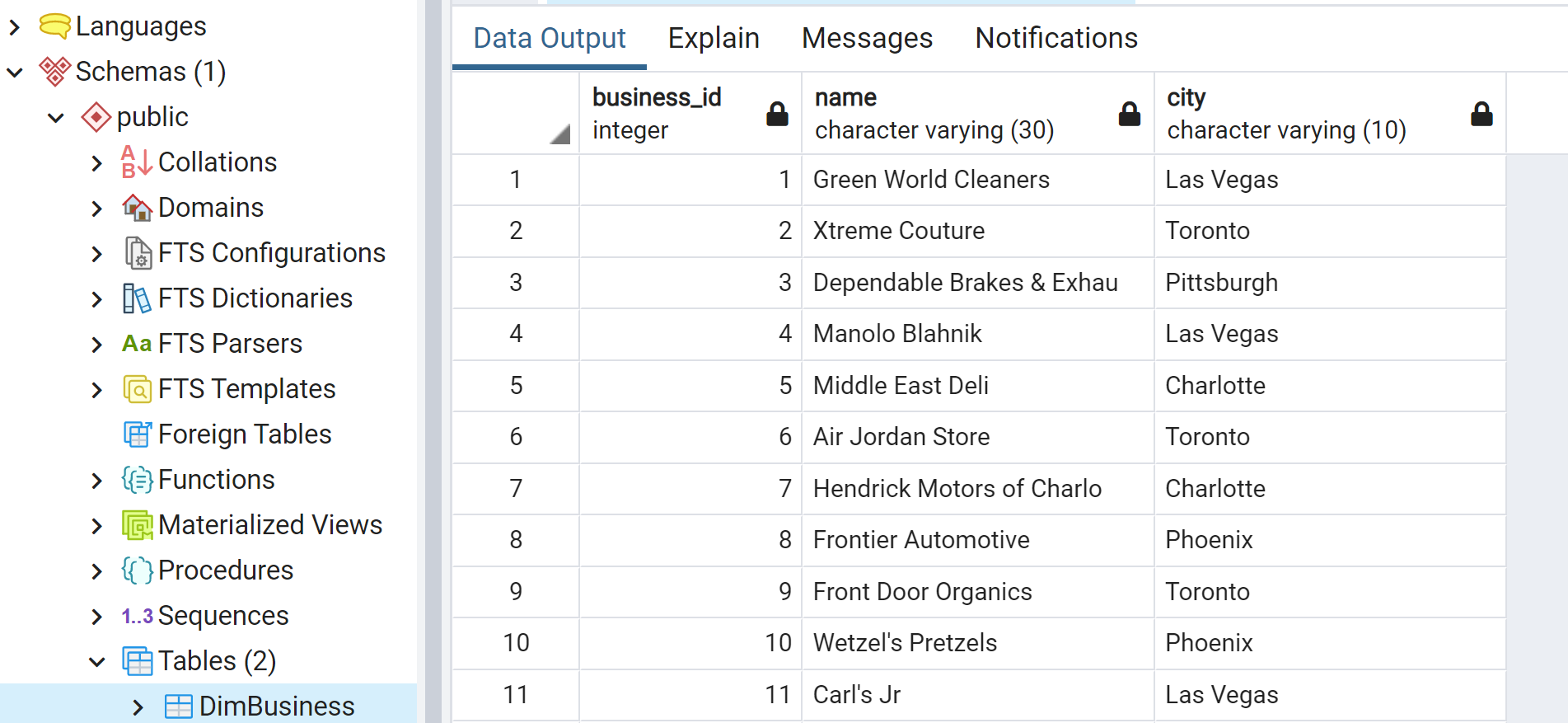
DimCity Table: Primary Key here is the City\_id, Name is the City name, we have two attributes (Columns), in Public Schema. This table contains 5 rows.

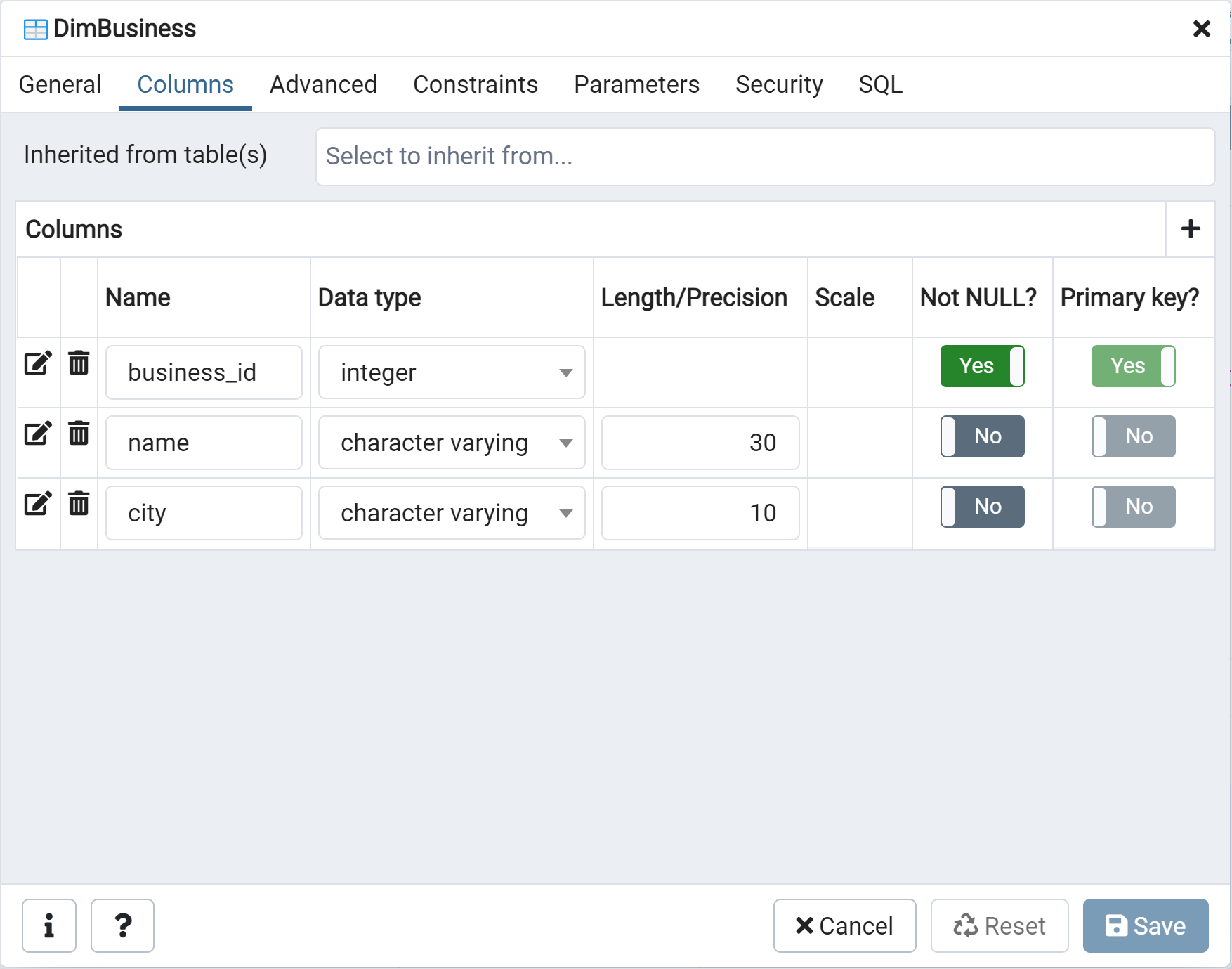


DimBusiness Table:

Primary Key for DimBusiness Table is the Business\_id. There are three attributes (Columns), table is also in Public Schema. This table contains 90220 rows. Other columns include name (Business Name); city (Business Location).







DimWeather:

DimWeather table has Weather\_ID attribute as its primary column and every other column can be null too. Following are column data types (Weather\_ID int, Temperature float, Humidity float ,Pressure float ,WindDirection float, WindSpeed float, Description text, City varchar(10))

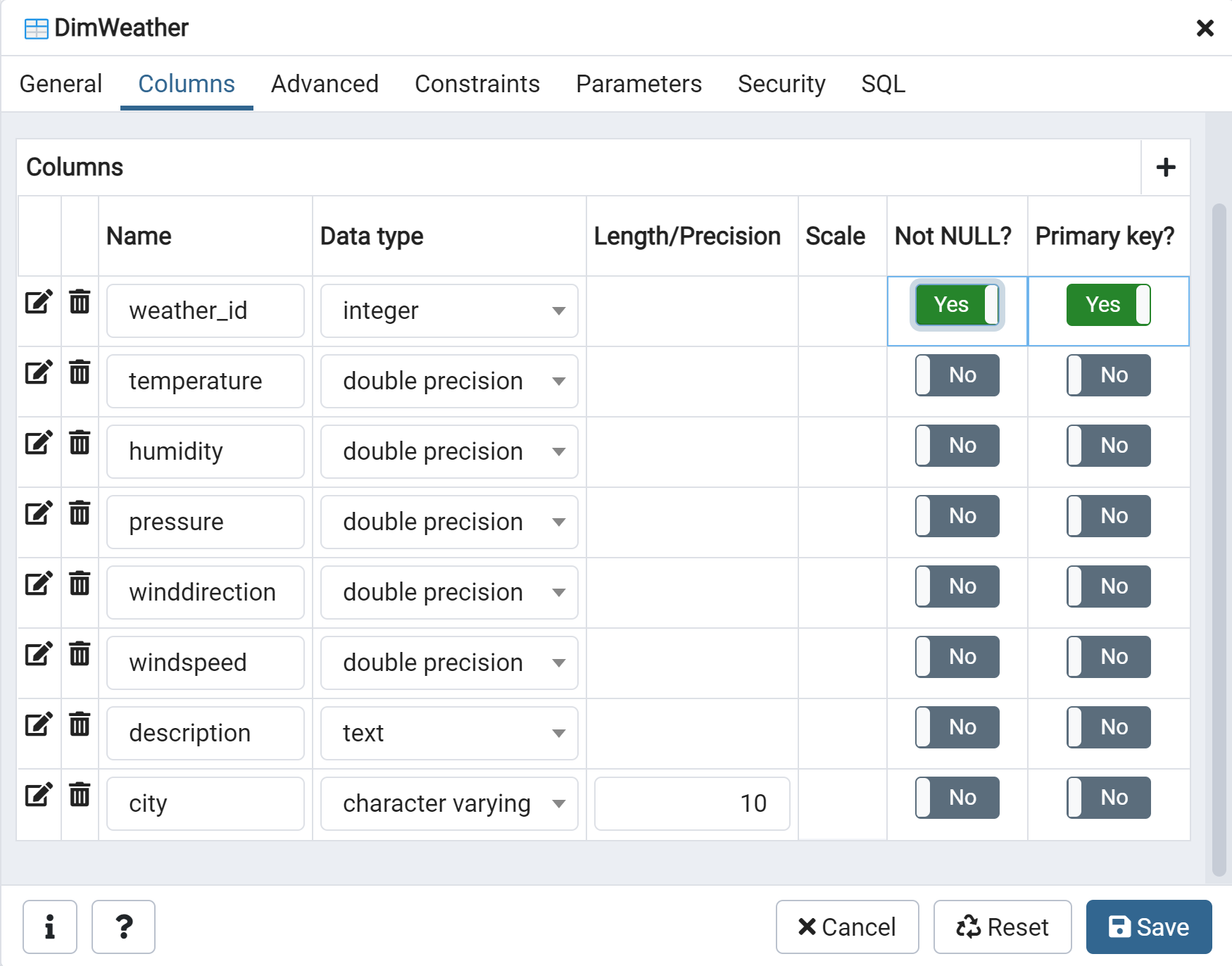
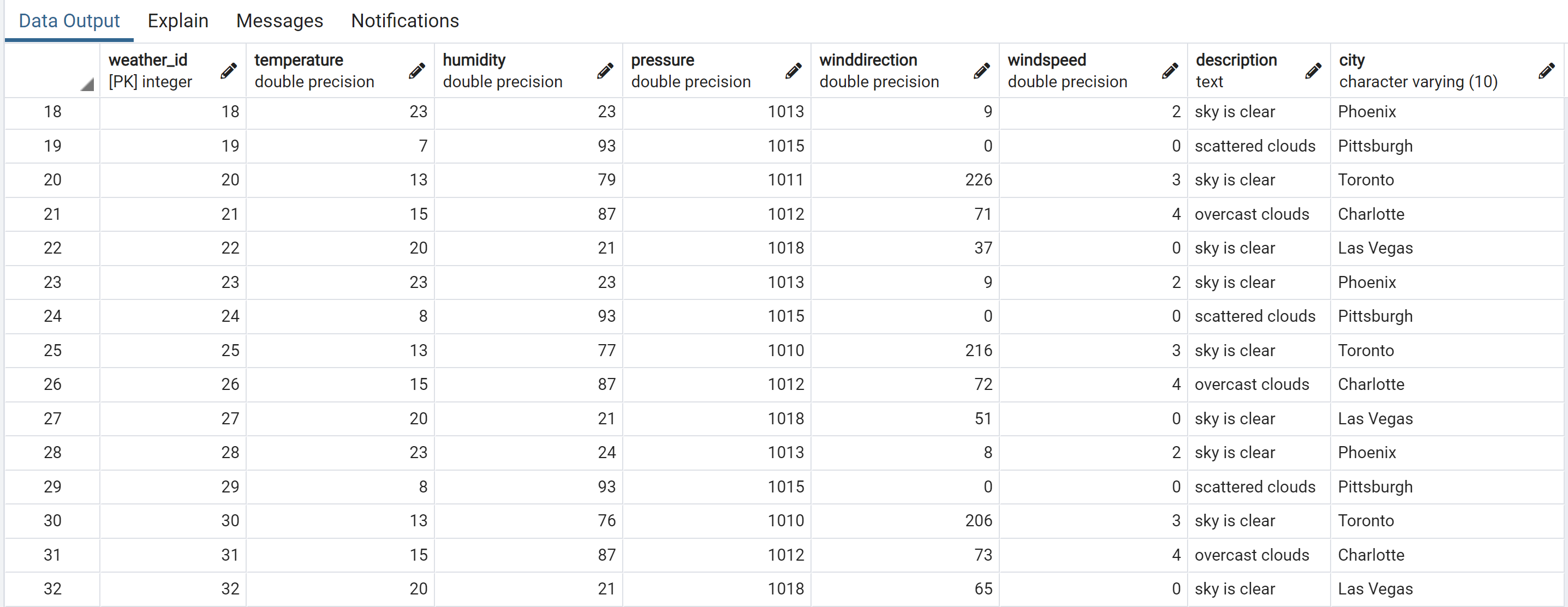


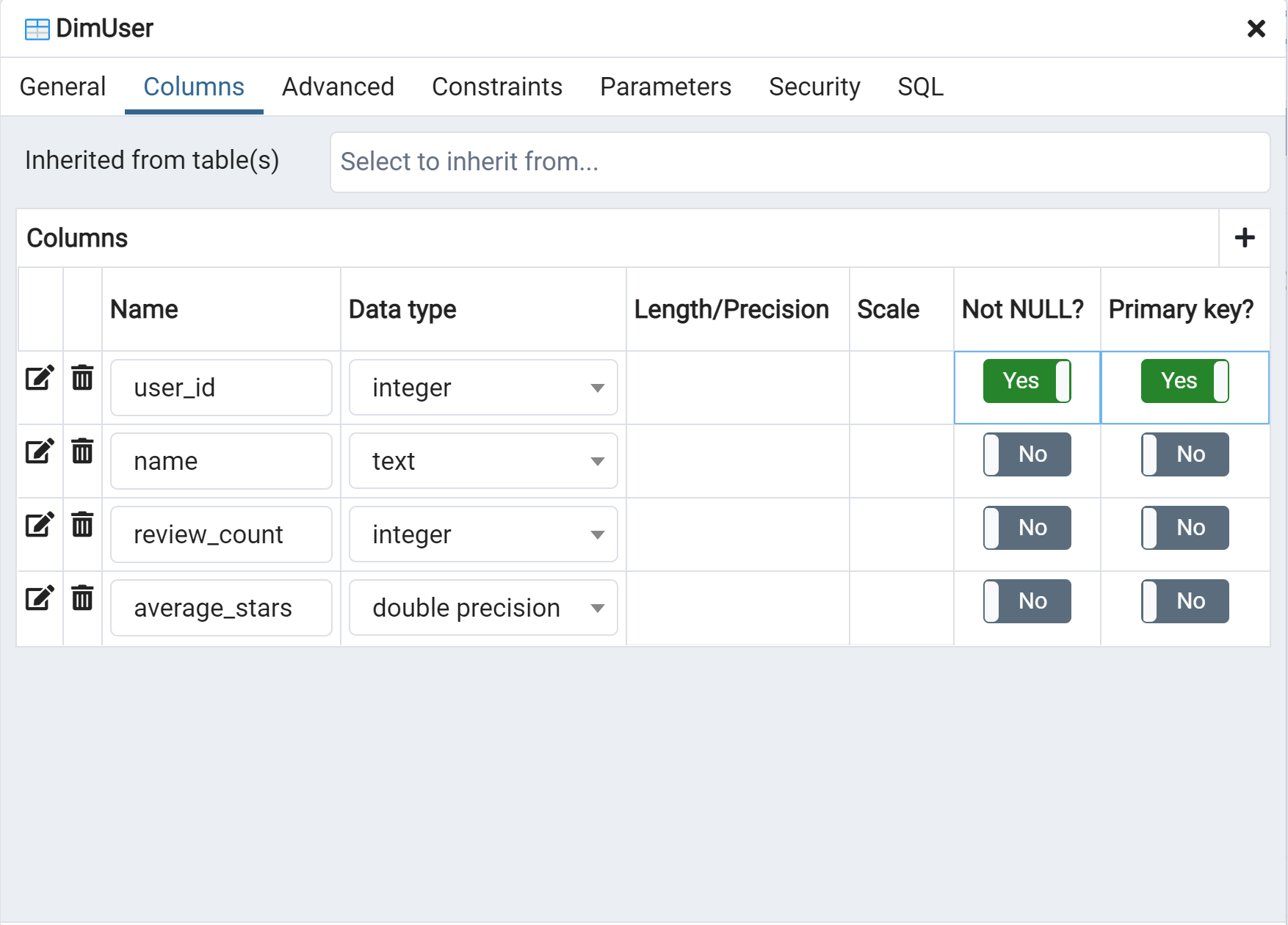
Table contains 227815 rows, excerpts below:

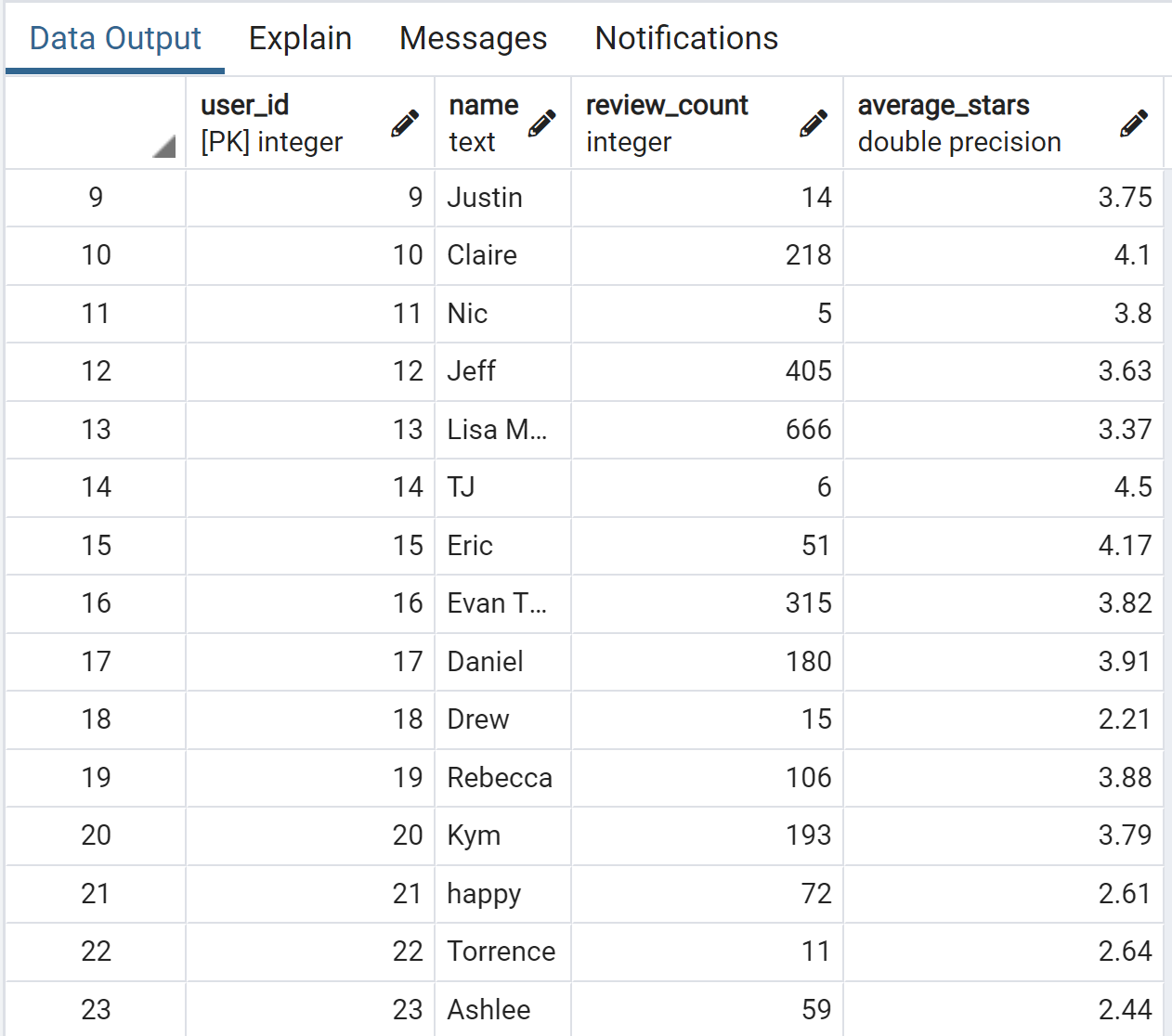


DimUser Table:

We have 1968703 distinct users, with User\_ID as its Primary Key, it contains other attributes;

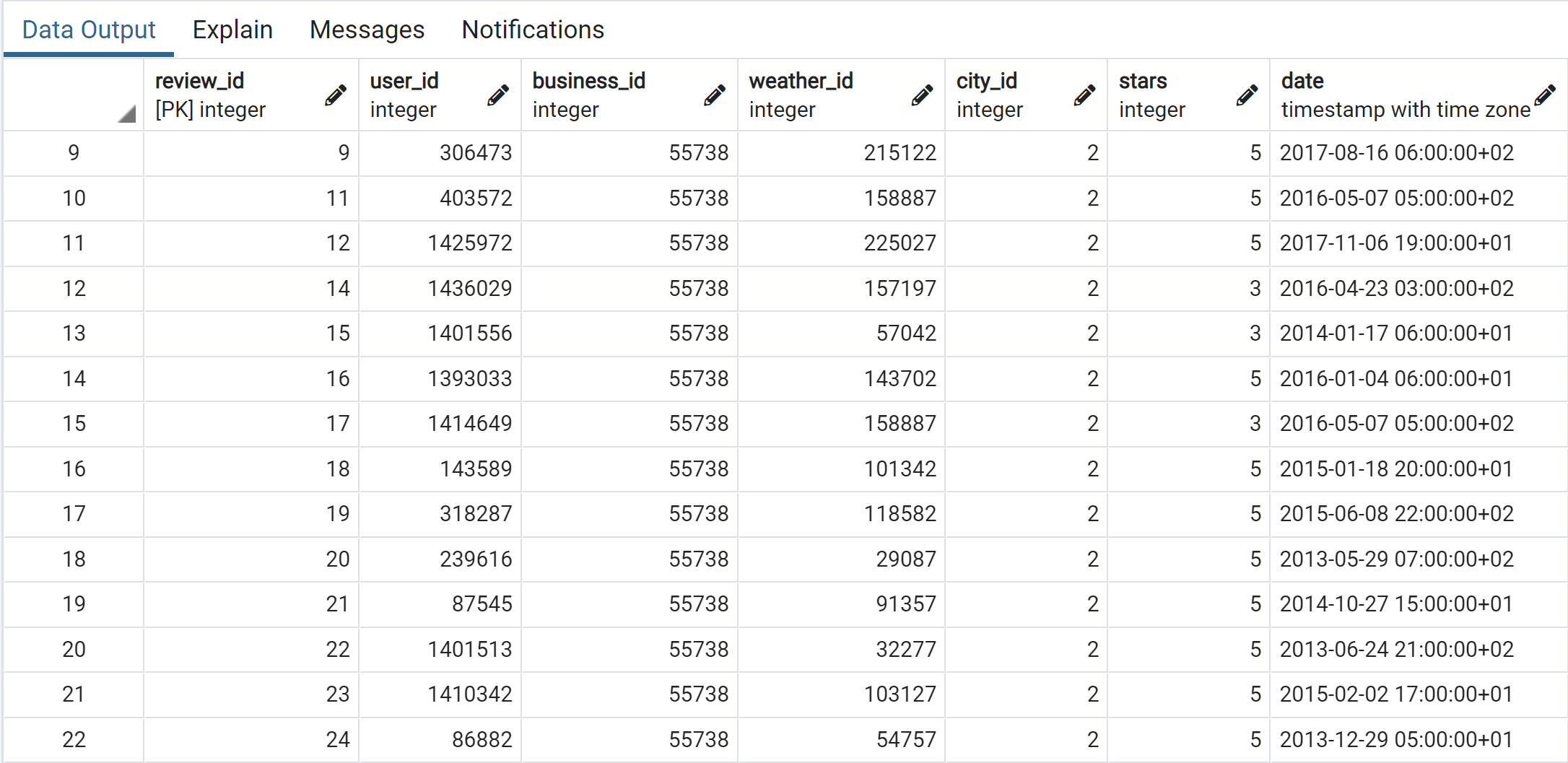
Name of User, Review\_Count, and Average Stars.





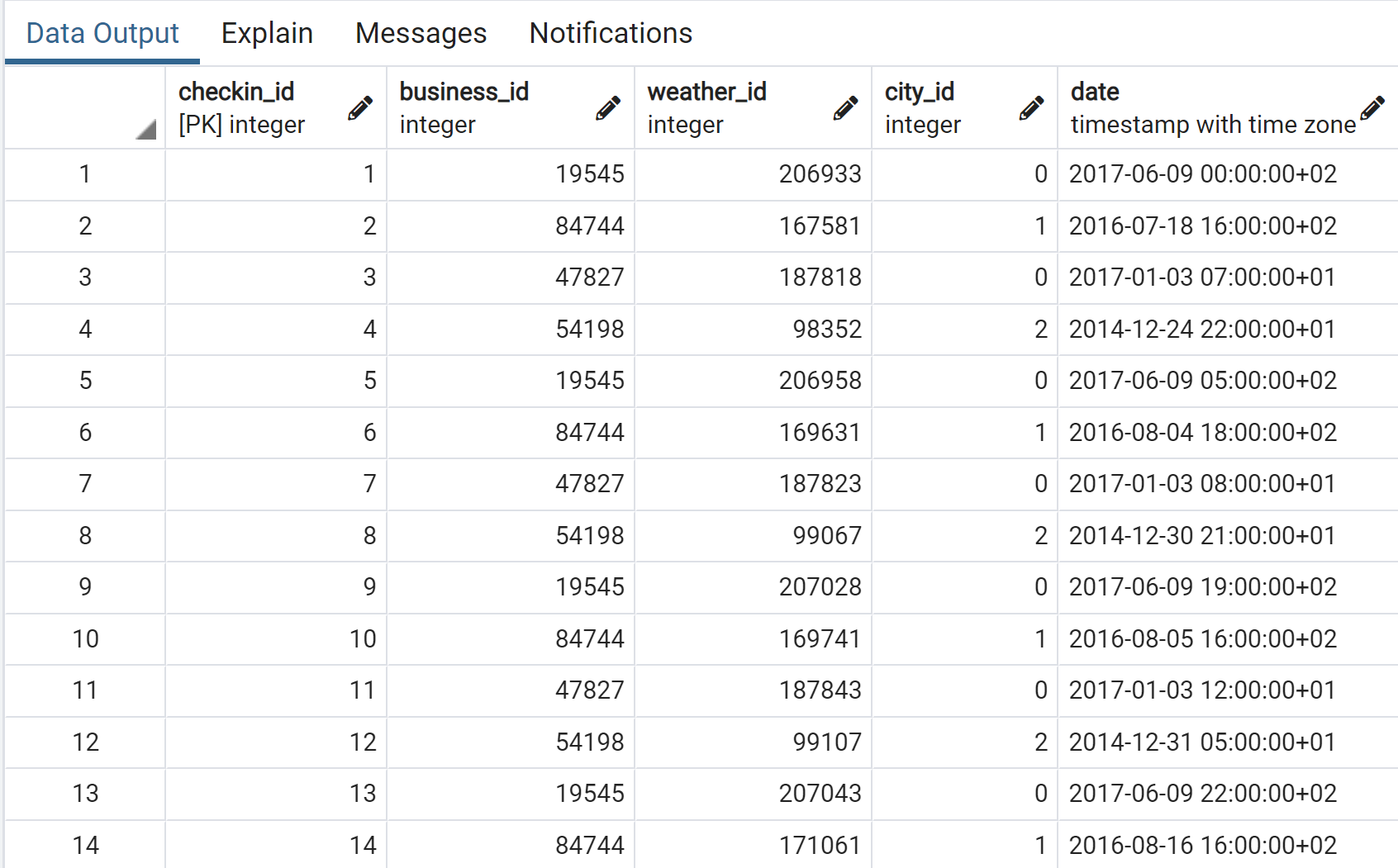
FactReview:

This is a fact table, contains the following attributes; review\_id int,user\_id int,business\_id int,Weather\_ID int,City\_id int,stars int,date date. It contains 2516617 rows. Below is the excerpt from PgAdmin4.



FactCheckin:

Another fact table (Fact constellation) is the FactCheckin, this contains the following attributes; Checkin\_ID int, business\_id int,Weather\_ID int,City\_id int,date date and contains 8099629 rows. Below is the excerpt from PgAdmin4.



Results and User Environment:

For this stage, we are to use Tableau