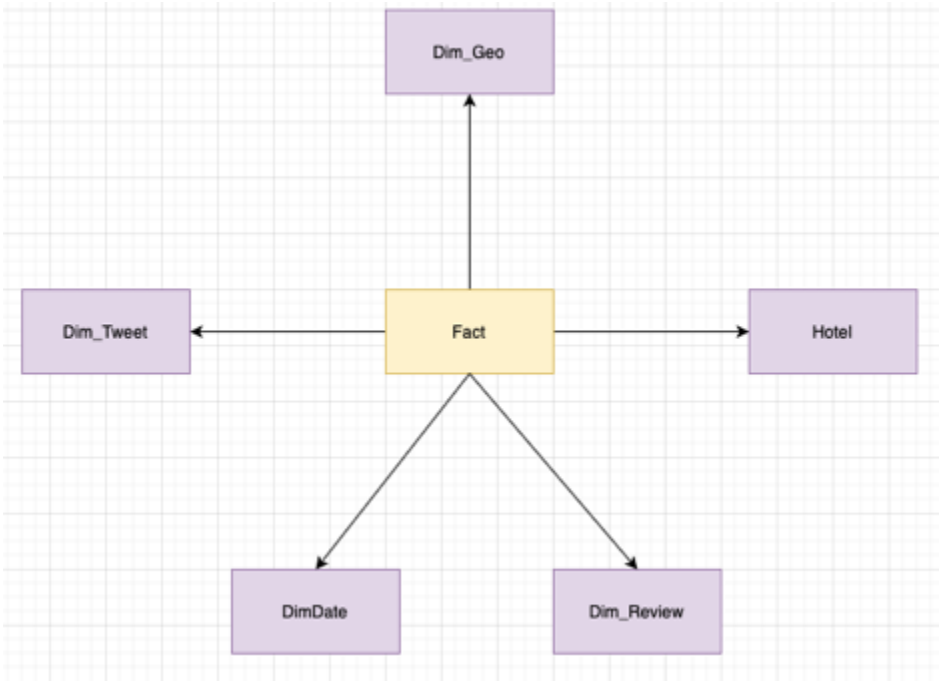
Project Report

Type of schema model used - Star schema model



Define Business Process

For the purpose of the project, the business process is to enhance the effectiveness in the working of the hotels and to predict the popularity of a given hotel, the positive and negative keywords that are being generated on social media which would help the hotel management to perform a daily analysis on the review of hotels based on their social media websites, make changes in their working style based on feedback from the analysis and improve upon the areas they are lacking in. Main purpose of this evaluation is to daily analyze social media records and compare it with the historic reviews and make improvements, if required

The service sectors today like the hotels need to be active in terms of both customer relationship and customer satisfaction thus efficiency is needed in providing best services and analyzing those services based on customer review on social media or data available with the hotel management relating to the same. We attempted to provide a Review

Analysis Engine which can be used by all hotels in general to analyze their progress on daily basis

Define grain of the Data Warehouse

Analysis can be achieved on a weekly basis which means that data would be provided on day by day assessment and evaluation may be completed on granular degree and take effective measure right away

Create Dimensions:

Identify the attributes from Data Tables and create separate length table for every of them

Fact table:

Create a fact table with all the measures left after creating facts. These measures can further be used for the purpose of analysis

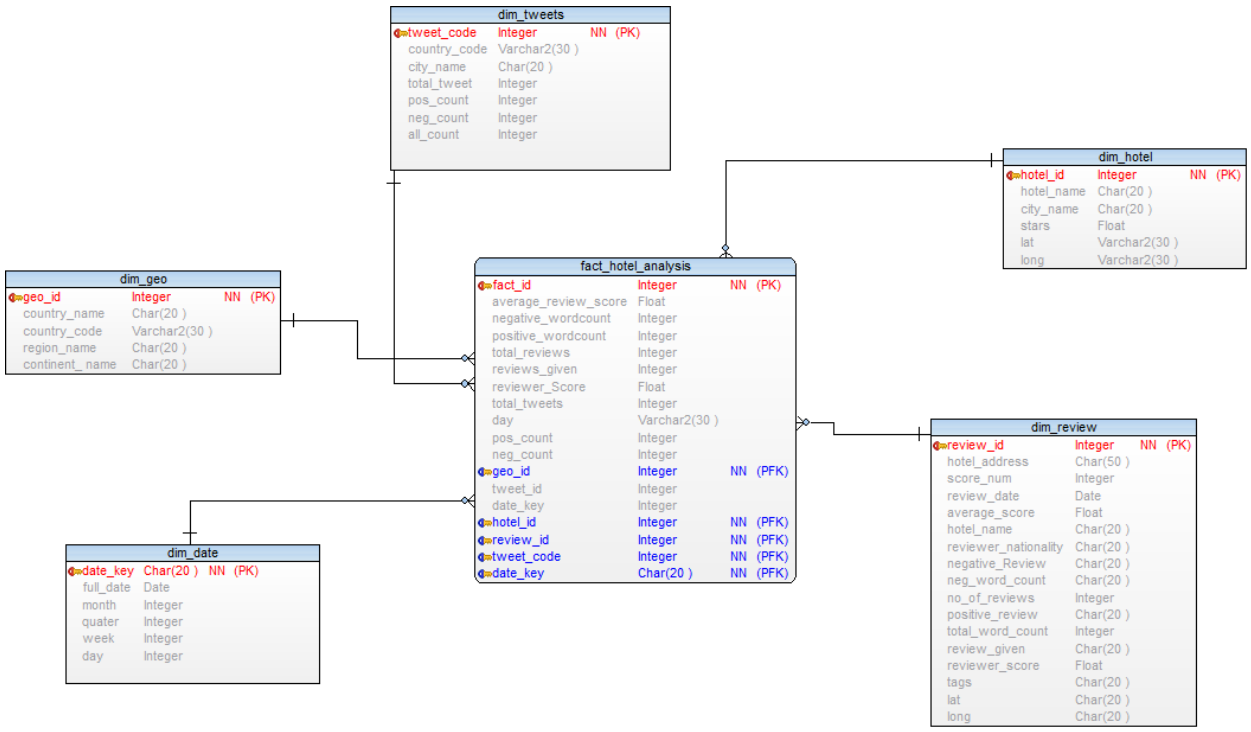
Report consists of following parameters given below:

1. Methodologies and Architecture used to build Data Warehouse
2. Data Modelling – Properly documenting the schema, usage of Datasets and Drill down approach usage in the model
3. Extract, Transform and Load(ETL) – Information about complexity of ETL, usage of Emerging Technologies in ETL, Automated ETL, describing methodologies used for ETL
4. Business Intelligence – Number of Business Queries will be explained in this document with methodologies used and how all datasets have been used in the process of building, critical evaluation of Business Queries using appropriate academics

1. Different Data Sources used in the project are: -
    - Country – Geonames http://www.geonames.org/countries/
    - Hotels – Github https://github.com/lucasmonteiro001/free-world-hotel- database/blob/master/hotels.csv.zip
    - Sentimental City Hotel Review Data: Twitter https://developer.twitter.com/

- Hotel Reviews Data- Kaggle https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe
- Create and Populate Date Table https://www.codeproject.com/Articles/647950/Create-and-Populate-Date- Dimension-for-Data-Warehouse

ER diagram:



Overview of Extract Transform and Load

ETL is process of Extracting Data from sources and Loading it into Data Warehouse.

Extracting:

Sources which are used in Data Warehouse, Sources can be any type Structured, Semi structured, Unstructured. In this project different sources of Data being used like

Structured Data (CSV Files) downloaded from Websites, some data was scraped from website and some unstructured data is Extracted from twitter using R Code which then converted to csv.

Two Types of Extractions methods used: Logical:

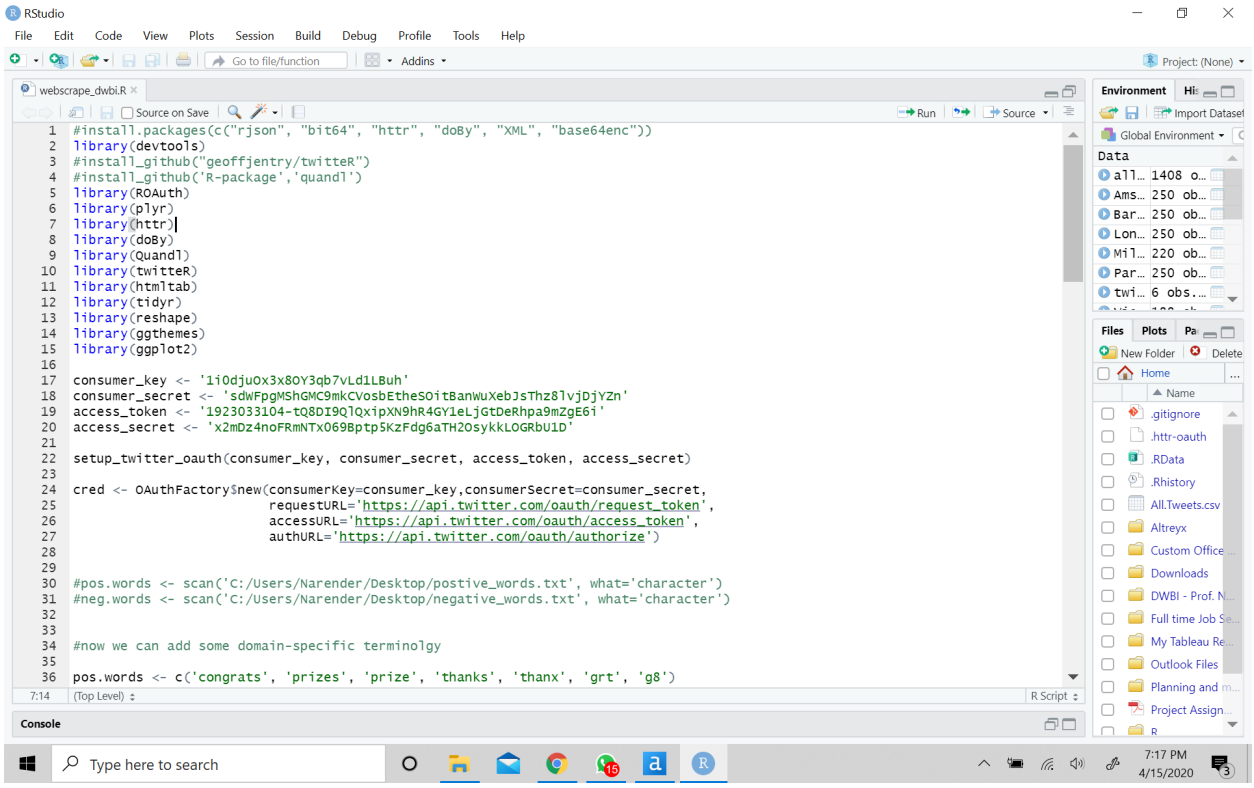
I. Full Extraction is extraction of data one time and no timestamps required in this extraction
II. Incremental Extraction is used when only changed data being extracted

Physical:

I. Online Extraction is done directly from source
II. Offline Extraction is done from Flat File, Dump File

In this project online Extraction done through R Code like Twitter Analysis and Web Scraping from Geonames using R Code. Offline Extraction of Datasets Done using R Code, Some of Screenshots for extractions used in this project are given below:

- R Code used

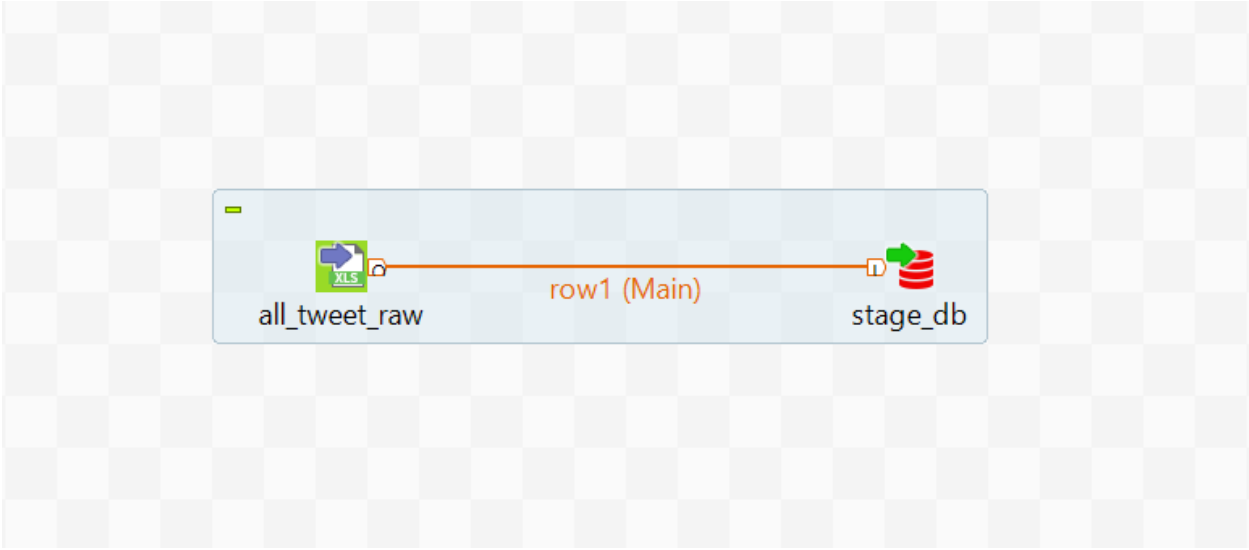


- Some CSV also downloaded from websites like Kaggle and GitHub.
- Data extracted is stored in SQL Server 2017 Database – Dimension Database for

  Dimensions and Facts, Stage Dimension is used as staging Area.

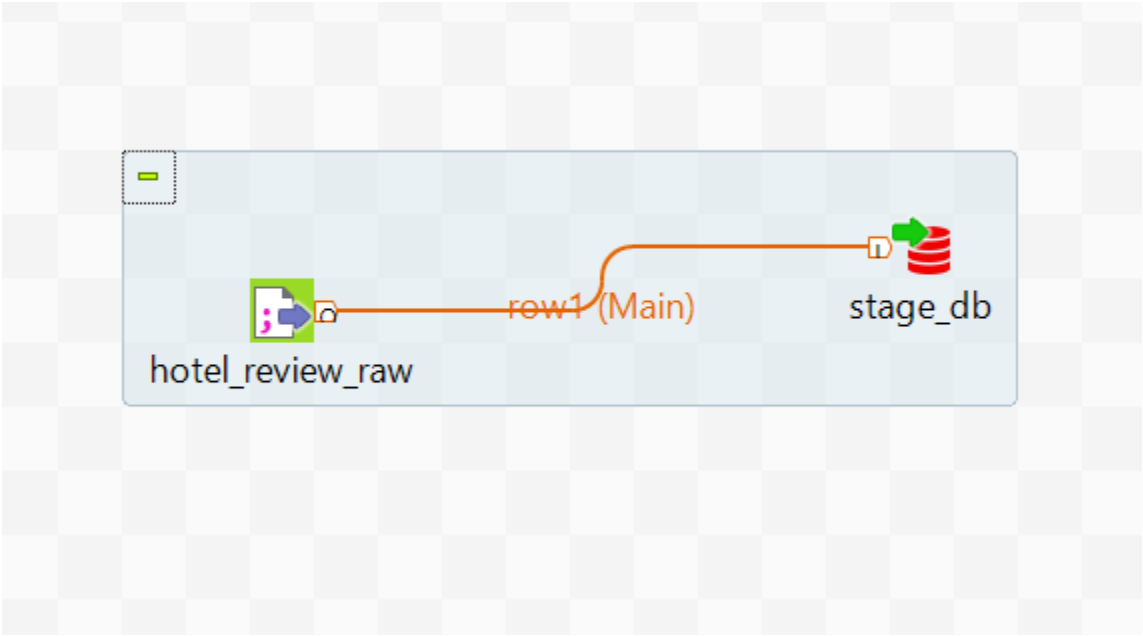
- Some of the Screenshot followed for extractions given below:

1) Staging:
   - All tweets: The twitter text data is loaded in staging

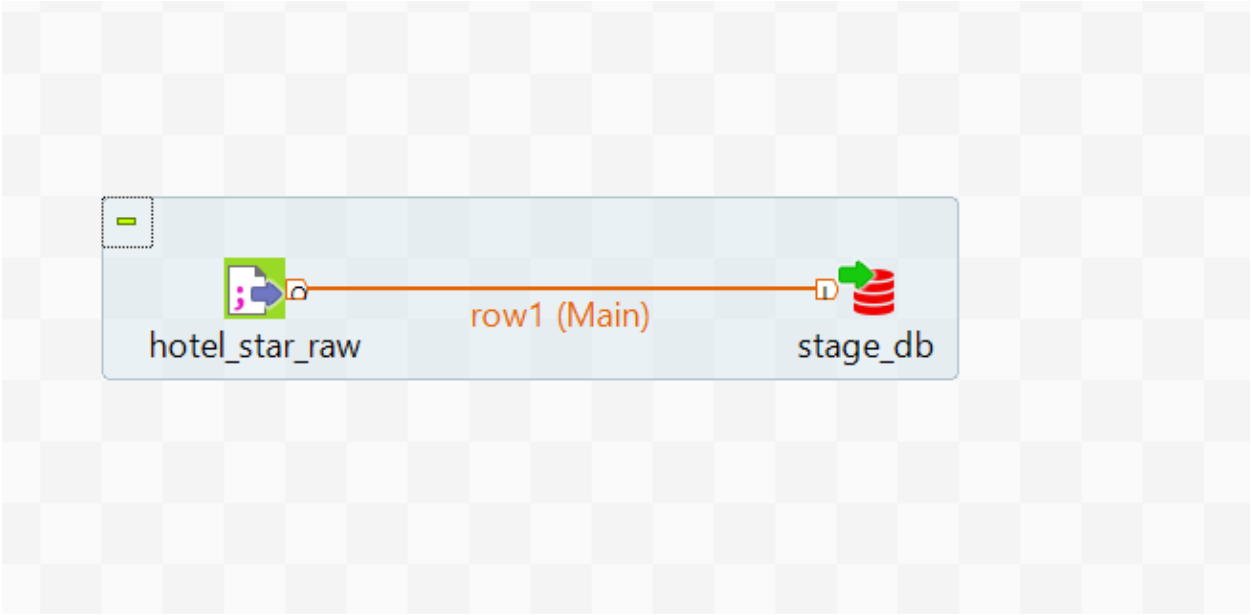


   - Country code: The Geo data is loaded in staging

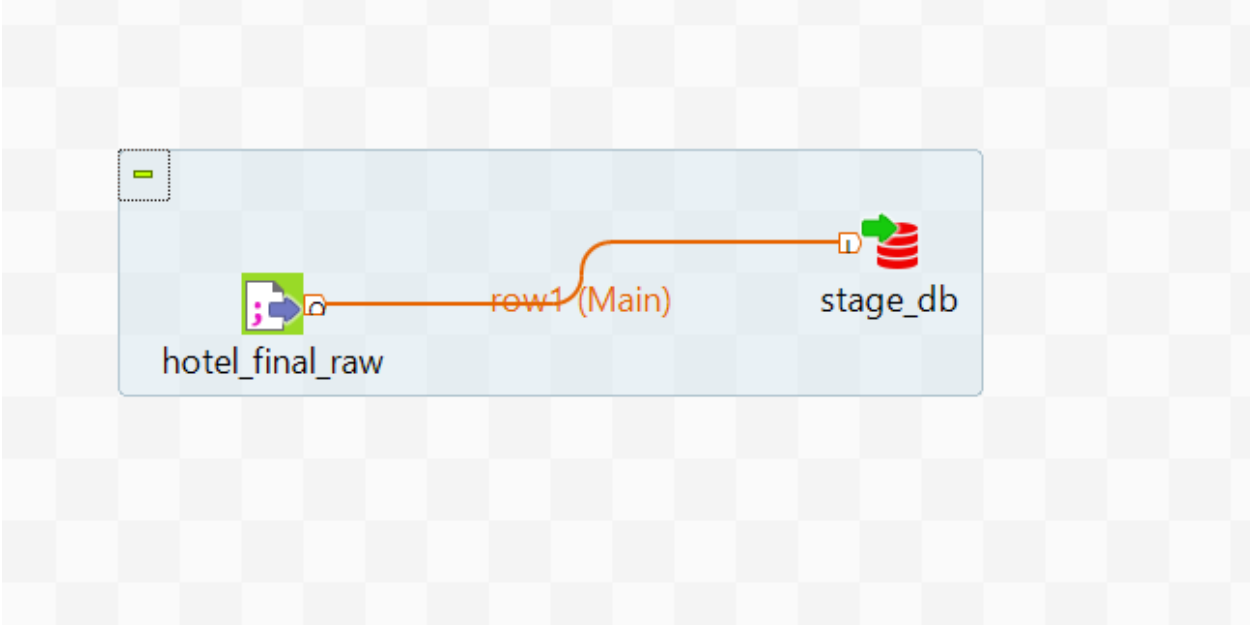
- Hotel review: The hotel reviews having the count, average score , reviewer score and rest data is loaded



- Hotel star:

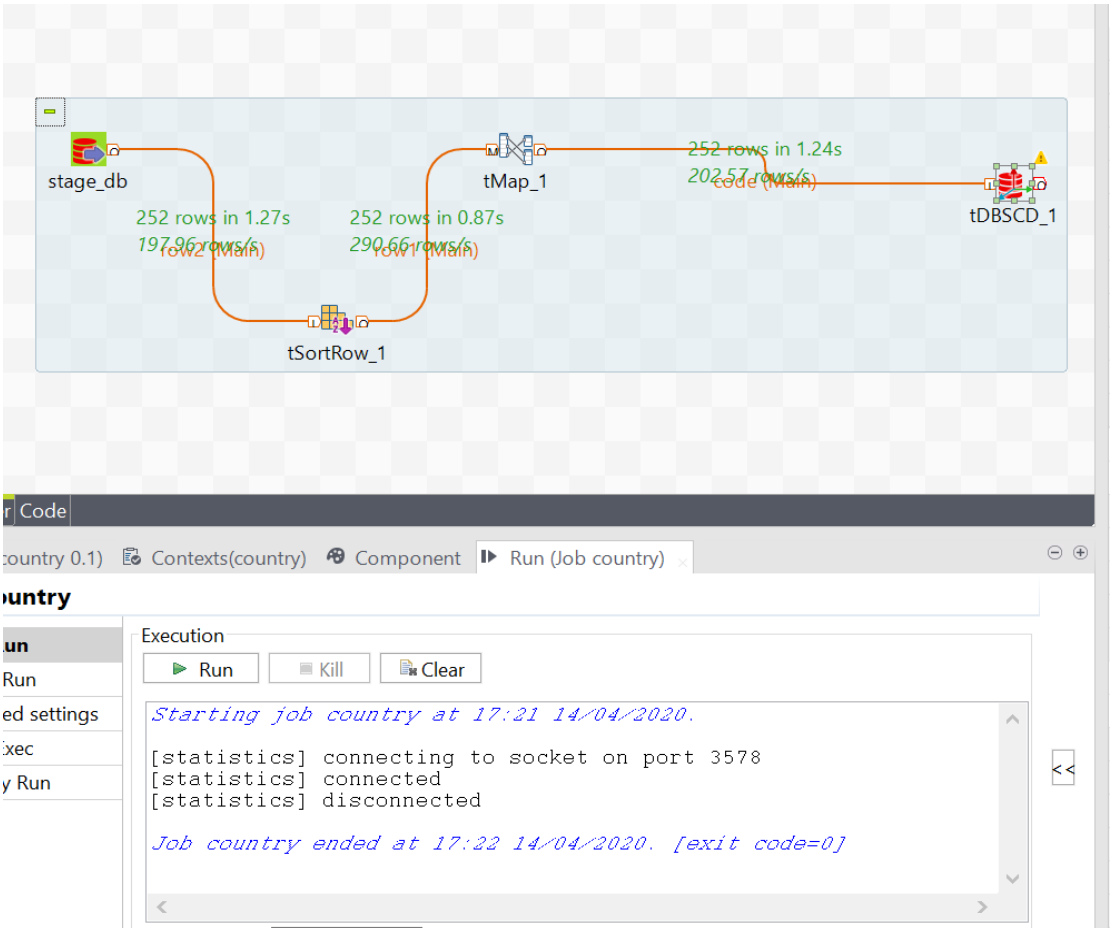


- Hotel final:

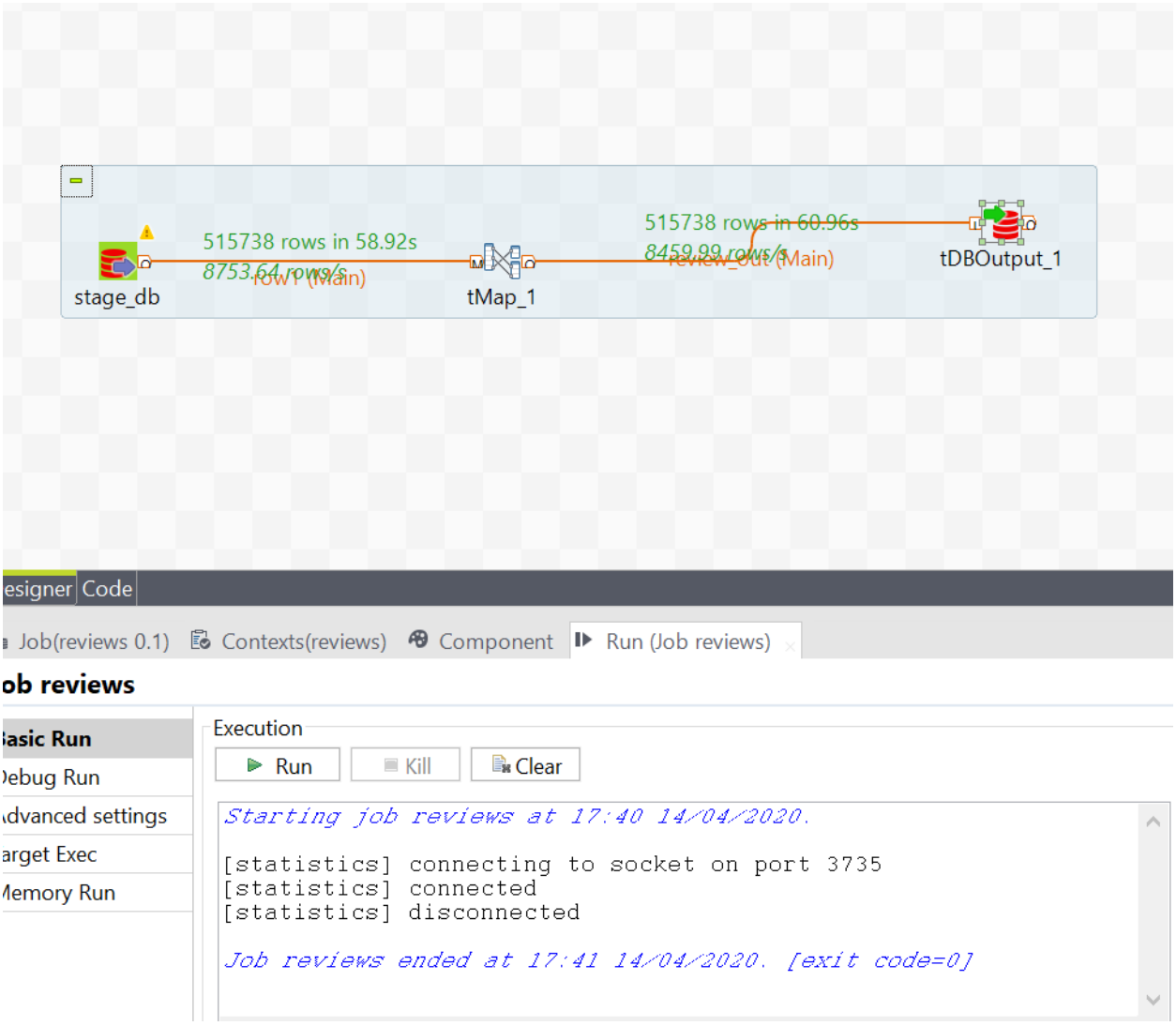


2)Final db:

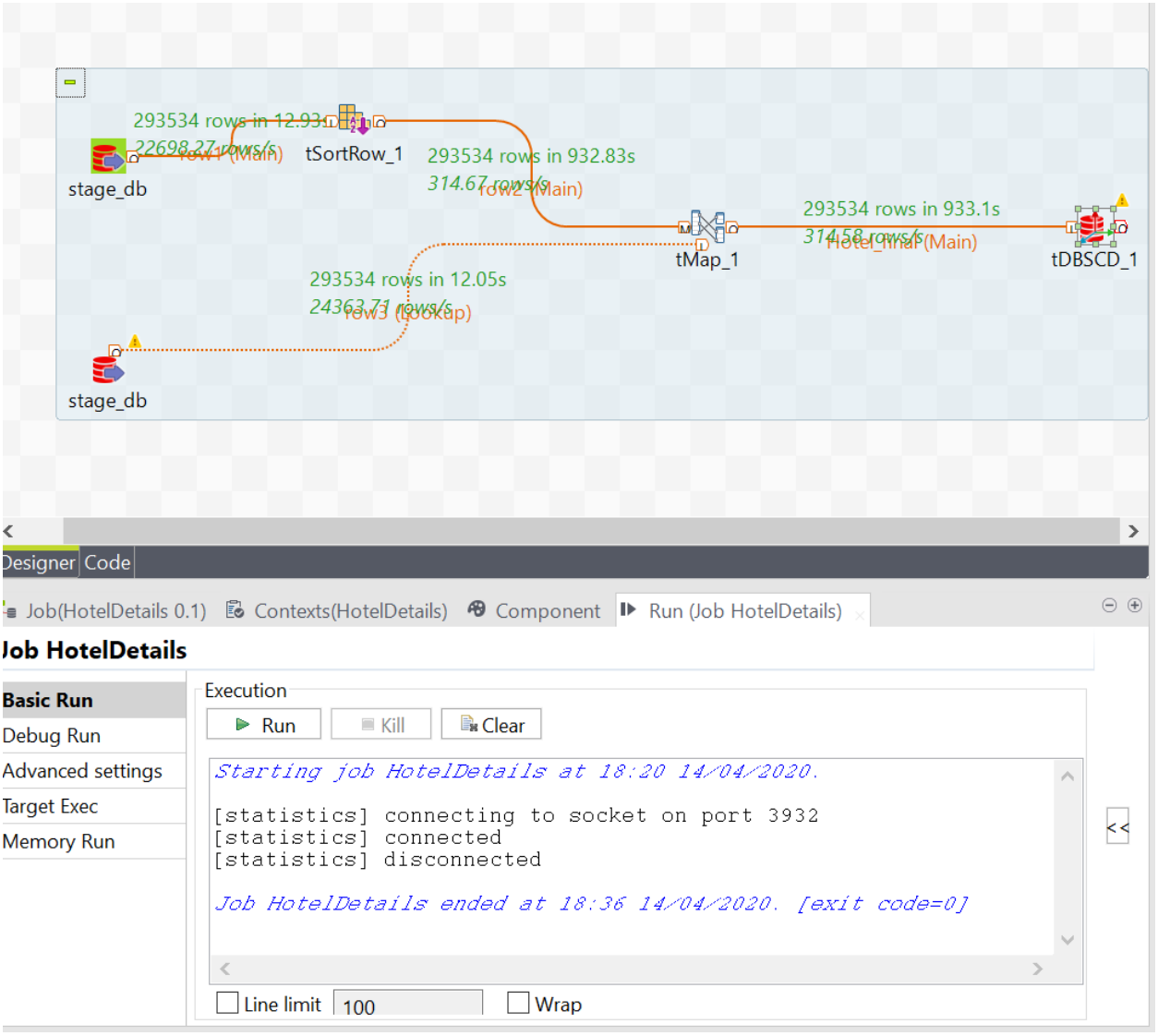
- Final_countrycode:



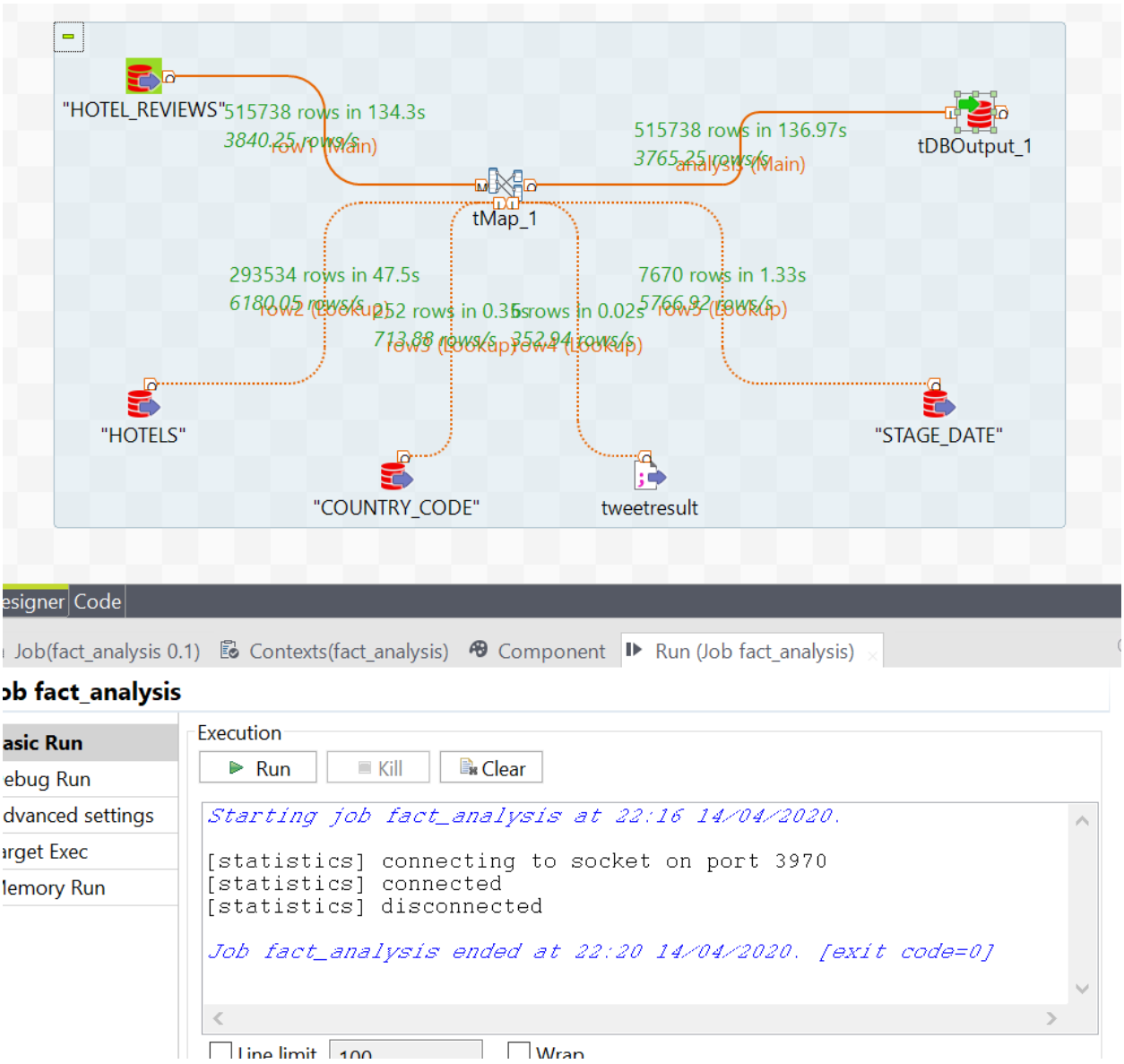
- Hotel review:



- Hotel details:



- Fact table:


"HOTEL_REVIEWS"
515738 rows in 134.3s
3840.25 rows/s
row1 (Main)
515738 rows in 136.97s
3765.25 rows/s
analysis (Main)
tDBOutput_1
tMap_1
293534 rows in 47.5s
6180.05 rows/s
row2 (Lookup)
252 rows in 0.35s
713.88 rows/s
row3 (Lookup)
5 rows in 0.025s
352.94 rows/s
row4 (Lookup)
7670 rows in 1.33s
5766.92 rows/s
row5 (Lookup)
"HOTELS"
"COUNTRY_CODE"
tweetresult
"STAGE_DATE"
esigner
Code
Job(fact_analysis 0.1)
Contexts(fact_analysis)
Component
Run (Job fact_analysis)
ob fact_analysis
asic Run
ebug Run
dvanced settings
arget Exec
lemory Run
Execution
Run
Kill
Clear
Starting job fact_analysis at 22:16 14/04/2020.
[statistics] connecting to socket on port 3970
[statistics] connected
[statistics] disconnected
Job fact_analysis ended at 22:20 14/04/2020. [exit code=0]
Line limit
100
Wrap

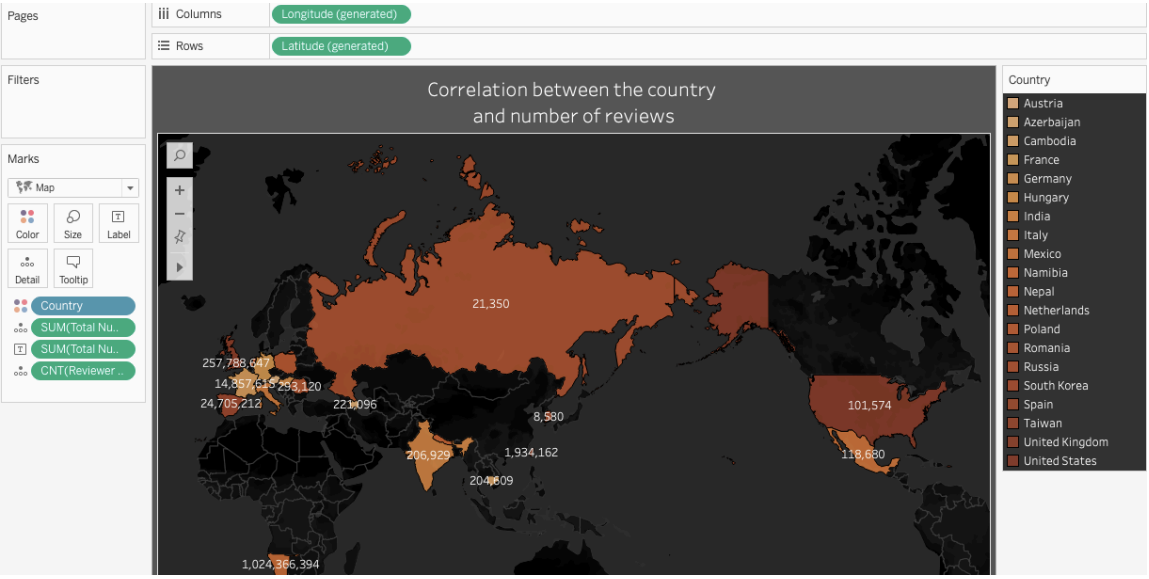
Business Intelligence

Here in this project got number of opportunities to slice and Dice Cube and find answer to different Business queries, but in this project, we have represented four main Business Queries that we found most attractive and provide insight in terms of choosing hotel before booking on basis of daily and historical data present for analysis.

Business Query 1: Are hotel really using power of reviews for business, which category of hotels more into this strategy, did it really effect the activeness of hotel customer to review?

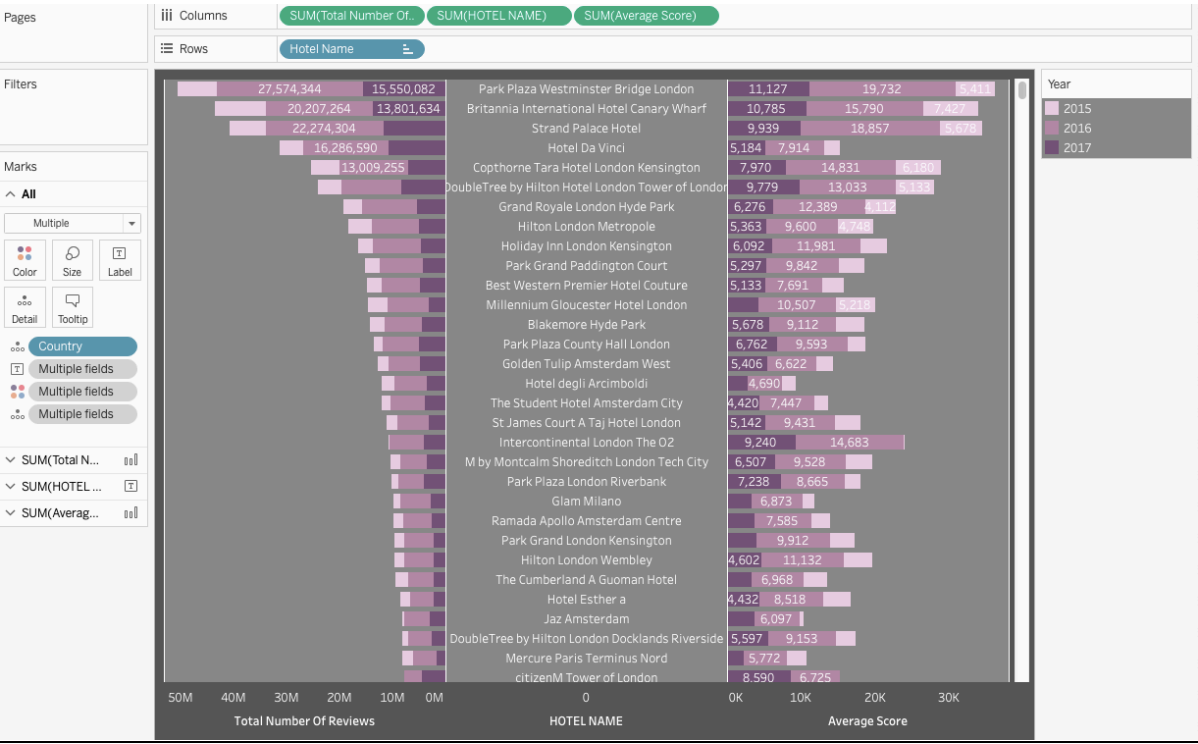
Hotel of 4-star category strengthened their relationship with customer more compared to others, so did they got more reviews and as we all know most of people go for holidays on weekends, from last several years there is change found in activeness of people to review a hotel.

Business Query 2: Find correlation between reviewer country and number of reviews given by country, is there any relation with nationality of country and overall score, day of maximum reviews?

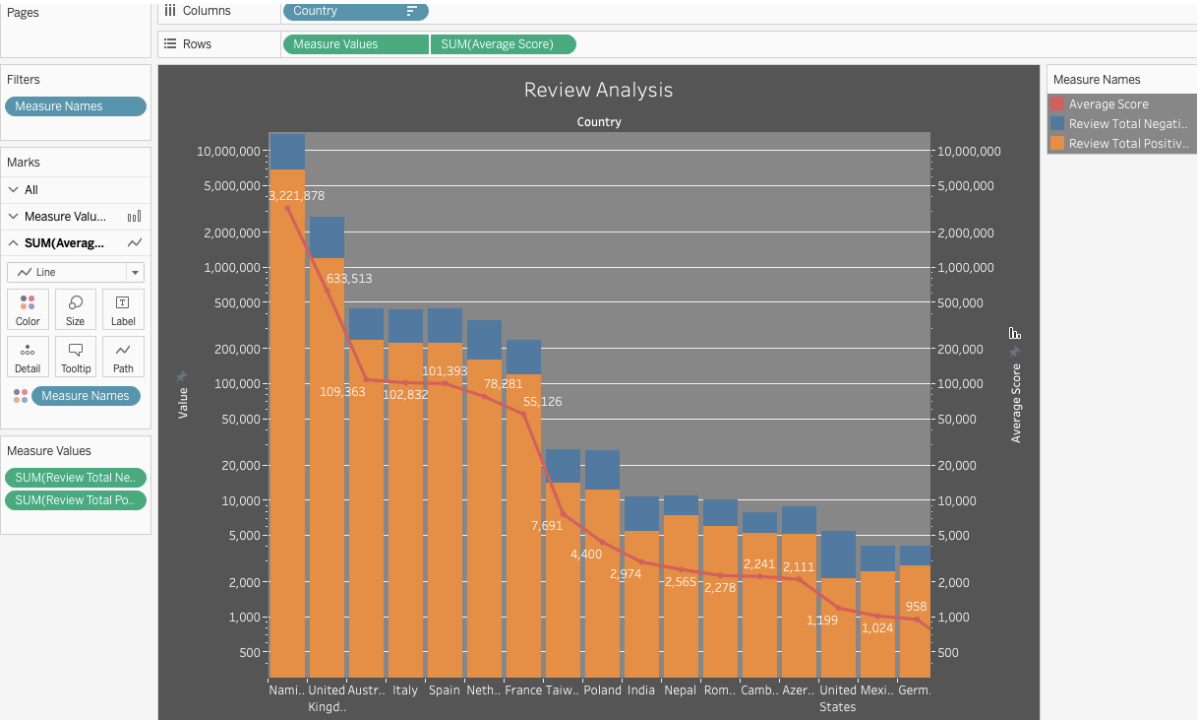


Business Query 3: Is getting maximum review is sign that services and overall experience of customer is good in that hotel?

After analyzing overall performance on the basis of popularity and average score, it's found that getting more review is sign of popularity but not better services, services of hotel can be understood by average score of reviewers



Business Query 4: Provide sentimental analysis from reviews to analyze Daily score of cities and compare it with historical score of hotels.



Final Dashboard