NATIONAL COLLEGE
OF
IRELAND

2018

Data Warehouse & Business Intelligence Project

Hotel Review Analysis Engine

TC

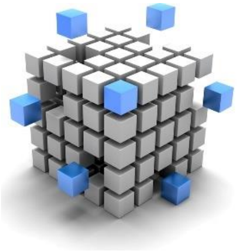
NARENDER KATARIA
X17168716

Data Warehouse and Business Intelligence Project

This project will provide insight to Hotel Industry, different data sources used for purpose of generating knowledge in respect of **Hotel Review Analysis Engine** (Paris, Amsterdam, Barcelona, Vienna, Milan, London (1500 Hotels)) as my topic with the help of different tools.

Data Sources and Initial Cleaning

Technologies



Business Analysis

Table of Contents

Introduction

Objective

Hotel Industry is growing at exponential rate so is the data increasing with tourism industry, tourism industry effecting both people and economy of country as whole, so in this project of Data Warehouse tried to fetch out some useful knowledge from data available on internet using some datasets, web scraping and using API's in general.

Data warehouse project is project which give insight or confidence in decision making with relevant all-round information present with Company/Industry. It is an optimized version of operational database provide only relevant information for and also provide fast access to Data, Data Warehouse is subject oriented, Integrated, Time-variant, Non-Volatile Platform

This report will cover architecture used and overall implementation approach followed for this exhaustive data warehouse project. This Project will Analyse key relationship Hotels their category and reviews. Project will also demonstrate about approach and platform used to deliver this project.

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.

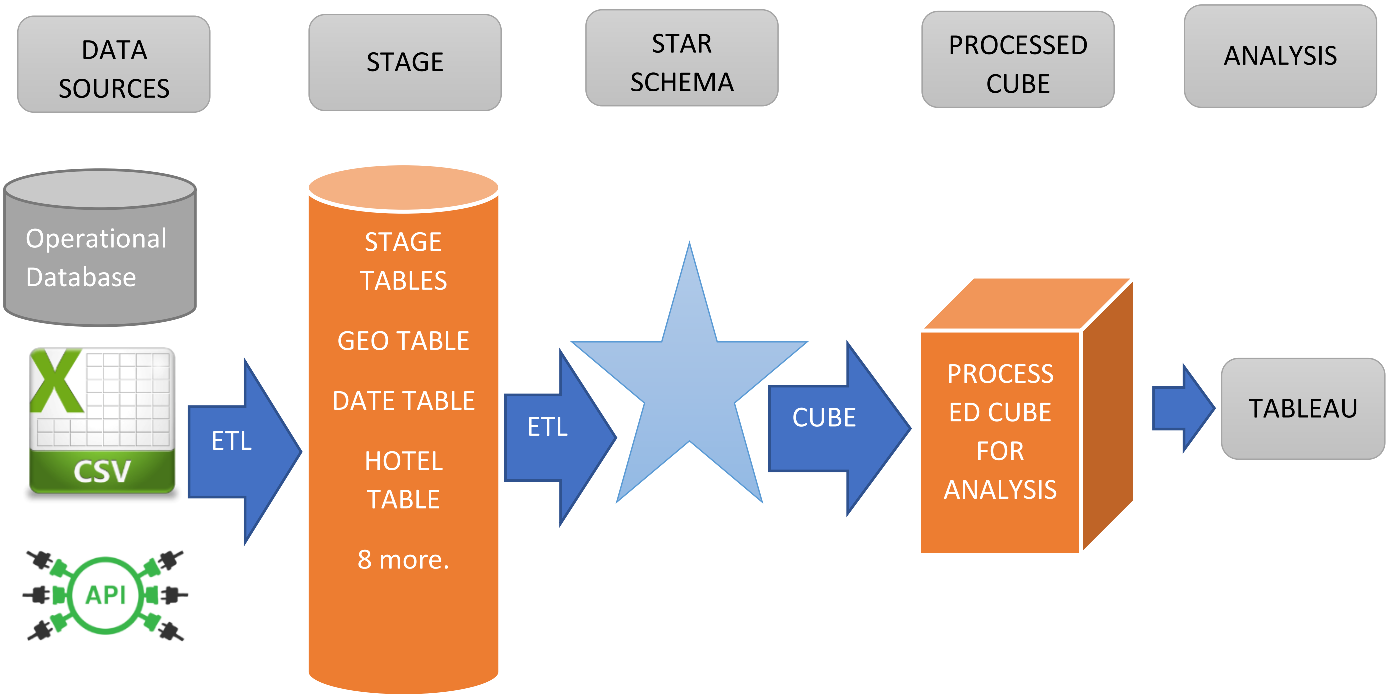
Methodology

Two methodologies are famous one is Bill Inmon and Other is Ralph Kimball, For this project Ralph Kimball methodologies was used because it suits the requirement of the this small scale project , here I will be discussing several comparison between the two which made me choose Ralph Kimball Approach , one of the main reason behind is Bill Inmon is suitable for large enterprise with Big Budget, Larger Time Scale, Corporate Environment, Changeable Sources with Top Down Approach and Relational 3NF
But I choose for Ralph Kimball because its suitability to Small Scale Projects, Small Business Area, User based, small budget with Bottom-up approach through informative bus architecture, Multi-Dimensional and Structure which can be easily understood by actual user.

My requirement for this project being minimum with not stable sources being used, easy to understand architecture and modelling to create database with perspective to user in mind, can be easily understood by the user, Star Schema and Snowflakes Schema structure are famous in Ralph Kimball approach, ill be using Star Schema because tables granularity is same for all the tables used in the implementation of Data Warehouse.

Data Warehouse-Architecture

Architecture used can be explained by using diagram given below: -



Architecture of Data Warehouse consists of three-layer Data Source Layer, Operational Data Store Layer, Data Warehouse Layer.

1. *Data Sources Layer*: Different Data Sources are used for building a Data Warehouse like relational and non-relational databases, structured or semi structured data sources, legacy flat files consist of historical data, OLTP operational Data sources consists of small business events and data which can also be used for building up of data warehouse.

   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-Wareho

2. *Operational Data Store Layer*: also called Data Staging area where all transformation and cleaning of data done to make it suitable for usage of Data warehouse Schema, this is

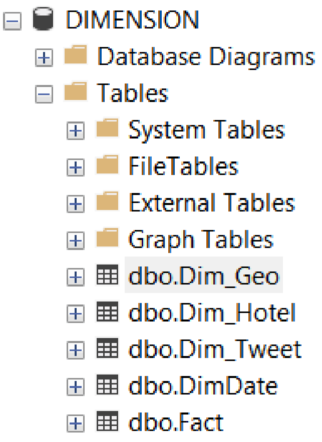
place where all operations on data done, it will be discussed in detail later in this Document.

3. *Data warehouse*: warehouse keep all records of historical data after proper transformation and integration which can be accessed by users using different analytical tools like OLAP to be processed into cube and using which analysis done by users.

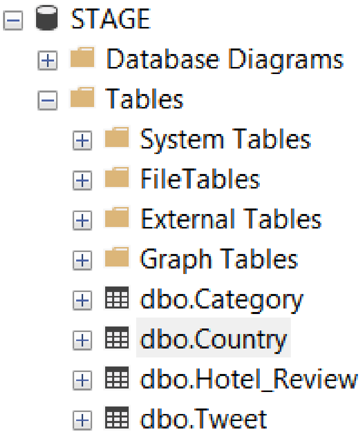
ETL – Extract-transform and Load tools are for purpose of cleaning and transforming data and making it suitable for schema. ETL is done on data sources coming from outside and make it adaptable for Schema, ETL is done for finding duplicates, errors and inconsistency in data. SSIS tool is used for ETL process and deploying OLAP Cube.

SQL Server 2017 used as database for storing and retrieving tables as required: -

Database engine to store fact and Dimensions in Database and Analysis Services is used to store processed Dimensions and Cube.



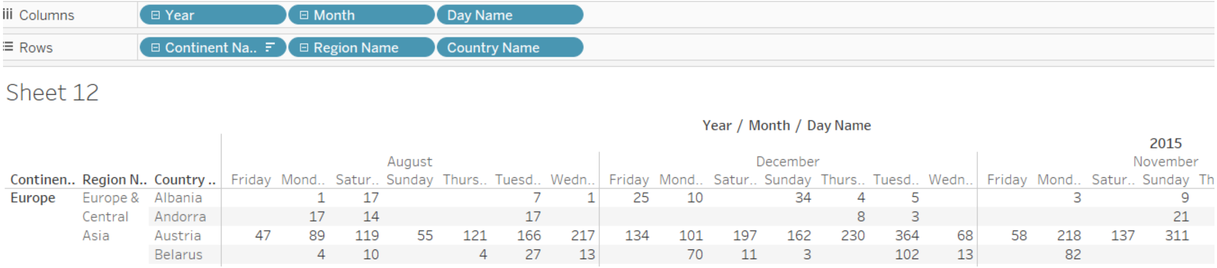
- DIMENSION
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - External Tables
    - Graph Tables
    - dbo.Dim_Geo
    - dbo.Dim_Hotel
    - dbo.Dim_Tweet
    - dbo.DimDate
    - dbo.Fact



- STAGE
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - External Tables
    - Graph Tables
    - dbo.Category
    - dbo.Country
    - dbo.Hotel_Review
    - dbo.Tweet

Drill Down Approach Snapshots

Multidimensional Drill down



Sheet 12

Year / Month / Day Name

																		2015			
			August							December							November				
Continen..	Region N..	Country ..	Friday	Mond..	Satur..	Sunday	Thurs..	Tuesd..	Wedn..	Friday	Mond..	Satur..	Sunday	Thurs..	Tuesd..	Wedn..	Friday	Mond..	Satur..	Sunday	Th
Europe	Europe & Central Asia	Albania		1	17			7	1	25	10		34	4	5			3		9	
		Andorra		17	14			17						8	3					21	
		Austria	47	89	119	55	121	166	217	134	101	197	162	230	364	68	58	218	137	311	
		Belarus		4	10		4	27	13		70	11	3		102	13		82			

Dimensional/Data Modelling

So, for this project I used bottom up Approach of Ralph Kimball, according to Kimball Approach we can have either start or snowflake schema which can have one or more fact tables and some dimension tables, fact will give you all measures whereas dimensions contains information, dimensions contains a primary which is used by fact table as foreign key.

Four Step of Data Modelling used in this project:

Define Business Process: - Business process used for this project is to provide effectiveness in work of hotels and to find in advance about the popularity of Hotel if going down, how many positive and negative words are generating on social media, hotel management to daily do daily analysis of review on social networking websites and according change the working style and improve in lacking area. Main motive of this analysis is to analyses social media data on daily basis comparing it with historical reviews and improve as required.

Today service Sector like hotel require to be active in terms of customer relationship and satisfaction of customer as whole. So, it is required to be efficient in terms of providing best services as whole and analysing the services given on time using reviews of people on social media or any data available with Hotel Management. So, I tried to provide a general *Review Analysis Engine* which can be used by all hotels in general to analyse their progress on daily basis.

Define grain of the Data Warehouse: Analysis will be done on daily basis, so it will provide us information on daily review and analysis can be done on granular level and take effective measure instantly.

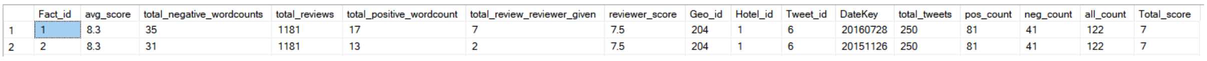
Create Dimensions: Identify the attributes from Data Tables and create separate dimension table for each of them.

Fact table: After creating facts create a fact table with all measures left, these measures further be used for purpose of analysis.

Data Modelling

Fact Table

Will store most granular measurement of business process, in this project review we are going to analysis will be of daily basis.



	Fact_id	avg_score	total_negative_wordcounts	total_reviews	total_positive_wordcount	total_review_reviewer_given	reviewer_score	Geo_id	Hotel_id	Tweet_id	DateKey	total_tweets	pos_count	neg_count	all_count	Total_score
1	1	8.3	35	1181	17	7	7.5	204	1	6	20160728	250	81	41	122	7
2	2	8.3	31	1181	13	2	7.5	204	1	6	20151126	250	81	41	122	7

Dimension Tables

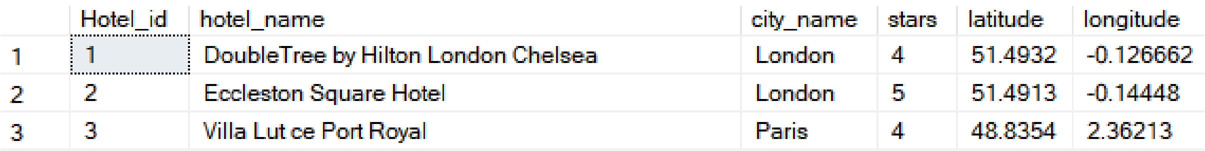
Will provide information about who, what, where and whop about business process, in this project we are using Four-dimension tables (Hotel, Tweet, Geography, Date).

Date Dimension will be a Role-Playing Dimension used for many dates.

Degenerated Dimensions also used with no Dimension Table of their own like Hotel Category Table, City Table, Country Table, Continent Table.

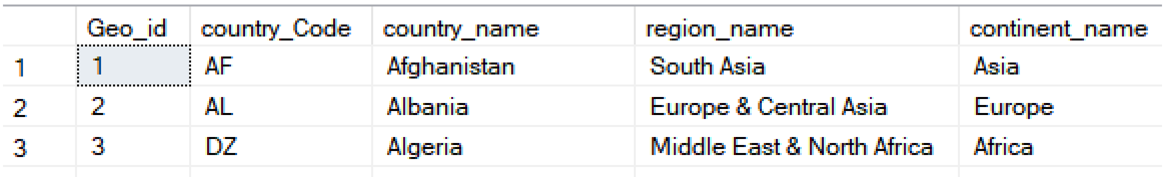
SCD 2 Dimension used to add new row to dimension tables if any.

Hotel Dimension



	Hotel_id	hotel_name	city_name	stars	latitude	longitude
1	1	DoubleTree by Hilton London Chelsea	London	4	51.4932	-0.126662
2	2	Eccleston Square Hotel	London	5	51.4913	-0.14448
3	3	Villa Lut ce Port Royal	Paris	4	48.8354	2.36213

Geography Dimension



	Geo_id	country_Code	country_name	region_name	continent_name
1	1	AF	Afghanistan	South Asia	Asia
2	2	AL	Albania	Europe & Central Asia	Europe
3	3	DZ	Algeria	Middle East & North Africa	Africa

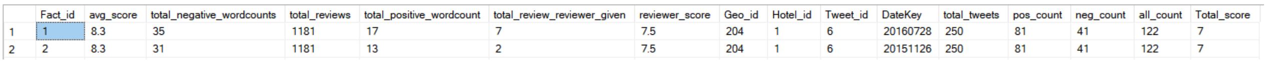
Primary Key

Used by dimension table to uniquely identify the row in the table. Hotel_id, Geo_id are primary keys for dimension tables shown in above tables.

Foreign Key

Foreign Key is a Column in Fact Table to reference another to create a Join between Fact and Dimension Table.

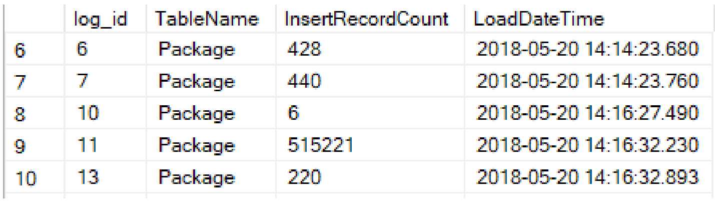
Fact Table with Foreign Keys like Hotel_id, Geo_td, Tweet_id,Datekey



	Fact_id	avg_score	total_negative_wordcounts	total_reviews	total_positive_wordcount	total_review_reviewer_given	reviewer_score	Geo_id	Hotel_id	Tweet_id	DateKey	total_tweets	pos_count	neg_count	all_count	Total_score
1	1	8.3	35	1181	17	7	7.5	204	1	6	20160728	250	81	41	122	7
2	2	8.3	31	1181	13	2	7.5	204	1	6	20151126	250	81	41	122	7

Auditing and Lineage

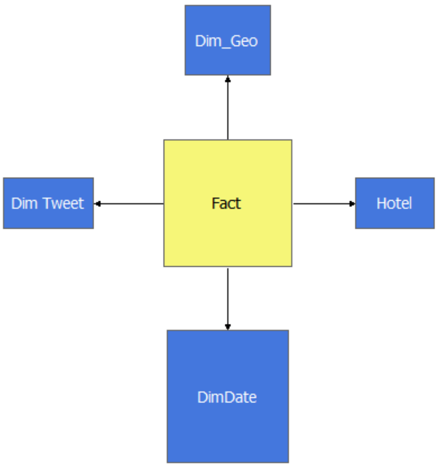
Auditing and lineage used to store logs into database for purpose of auditing who made update, when it's done, how many rows are updated.



	log_id	TableName	InsertRecordCount	LoadDateTime
6	6	Package	428	2018-05-20 14:14:23.680
7	7	Package	440	2018-05-20 14:14:23.760
8	10	Package	6	2018-05-20 14:16:27.490
9	11	Package	515221	2018-05-20 14:16:32.230
10	13	Package	220	2018-05-20 14:16:32.893

Star Schema Design

Star Schema has a single table for each dimension, each table contains all attributes for that dimension, particularly a demoralized form.



Star Join in Dimensional modelling is used to join both fact and Dimension tables, in start join facts are contained in fact table like average review score, total reviews score etc and information in Dimension table like geography dimension, time dimension, Tweet, Hotel dimension with category of hotel like Star of hotel.

Star join use both fact and dimension tables to answer any query related to dimension and facts and up to most granular level of join.

For example:

1. Which days in a week Hotel must consider strategically important to gain average score.
2. Is there any relationship between Star category of hotel with no. of reviews received?
3. Find out and compare Average Score of Historical data with Twitter sentimental Score?

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.

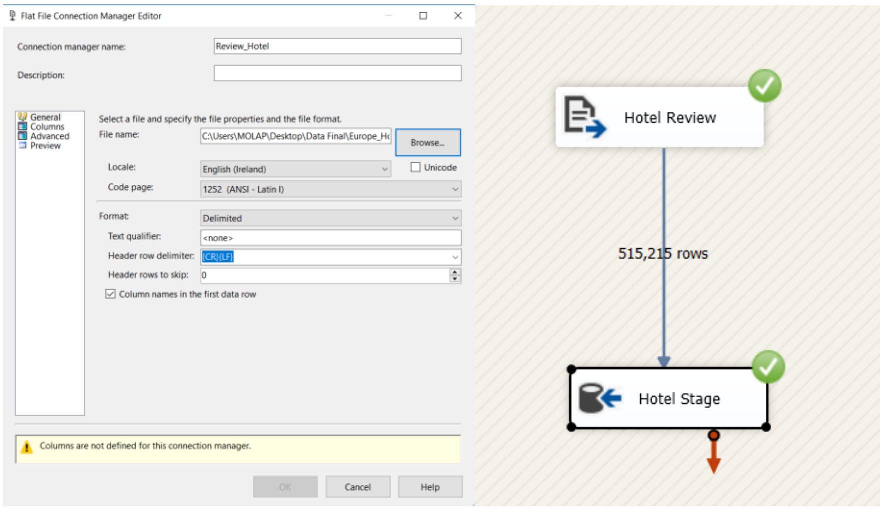
Physical:

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: given in Appendices I*
- 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:

Connection Manager for creating Connections connection with Flat files and Other Sources.



Flat File Connection Manager Editor

Connection manager name: Review_Hotel

Description:

Select a file and specify the file properties and the file format.

File name: C:\Users\MOLAP\Desktop\Data Final\Europe_Hc

Locale: English (Ireland)

□ Unicode

Code page: 1252 (ANSI - Latin I)

Format: Delimited

Text qualifier: <none>

Header row delimiter: {CR}{LF}

Header rows to skip: 0

☑ Column names in the first data row

Columns are not defined for this connection manager.

Transformation:

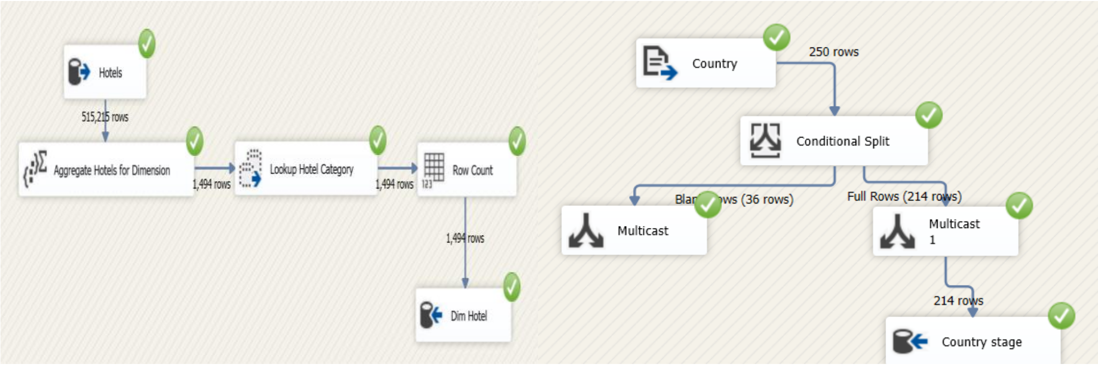
Is most complex in terms of processing time, here we do simple data conversion to complex data aggregation, scrubbing.

Number of transformations techniques are present in *SQL Server Integration Services*, but for this project tried to use minimum transformations as required for the project. Some Transformation done before staging Area and some before loading Data into dimensions and fact table.

Different Transformations used are:

- Multistage Data transfer.
- Pipelined Transfer.
- Create Table using SQL.
- Use of Merge, Sort, Multicast, Aggregate, Conditional Split.
- Multistage insert of Data.

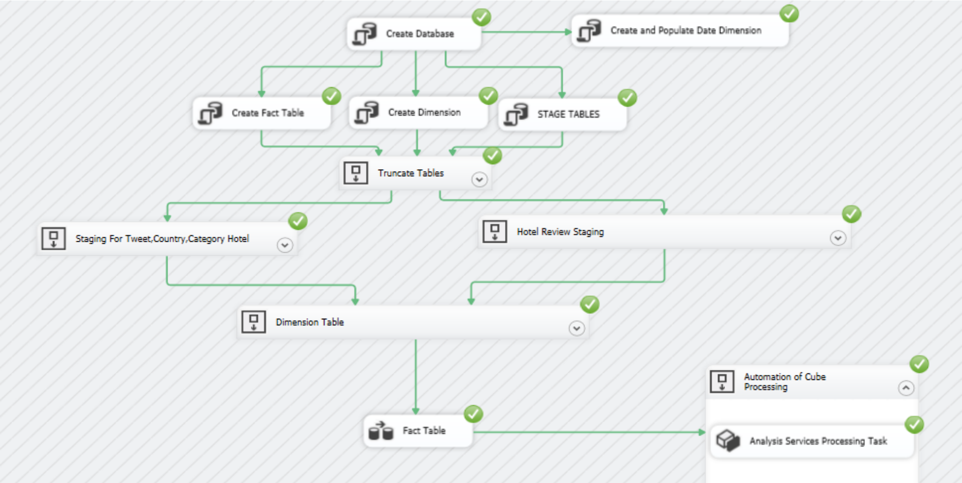
Some of the transformations used are given below:



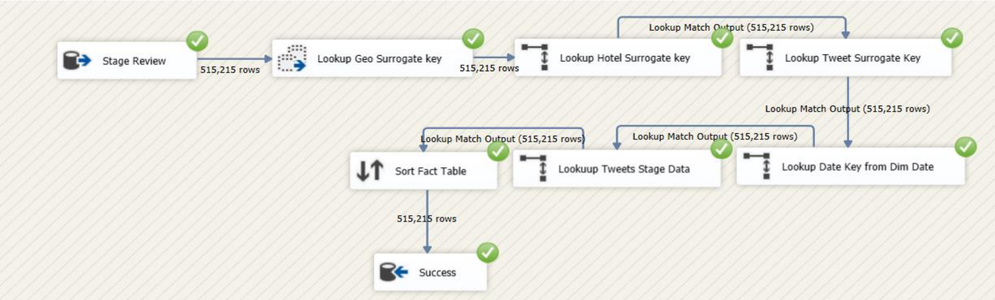
Loading Data into Dimensions and Fact:

- Using SQL, used SQL code to Populate Date Dimension (Code Taken online)
- Using pipelined multistage insert of Data.

Fully Automated Control Flow:

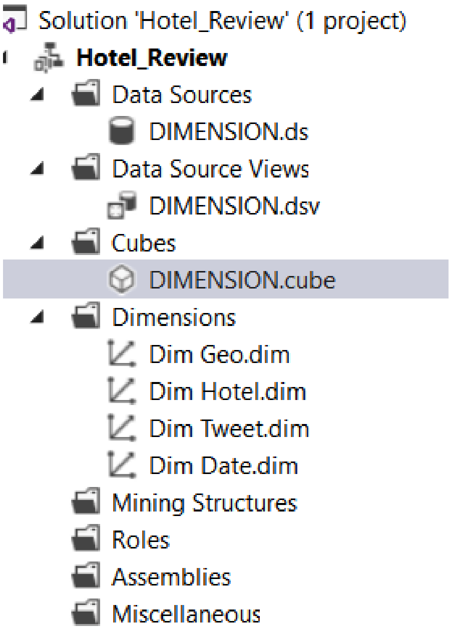

Create Database
Create and Populate Date Dimension
Create Fact Table
Create Dimension
STAGE TABLES
Truncate Tables
Staging For Tweet,Country,Category Hotel
Hotel Review Staging
Dimension Table
Automation of Cube Processing
Fact Table
Analysis Services Processing Task

Data Flow for Fact Table:

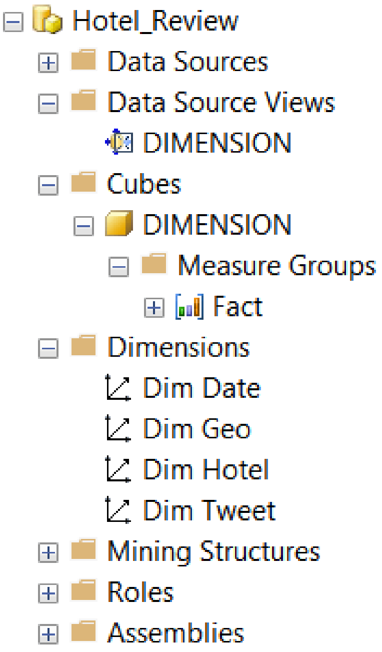

Stage Review
515,215 rows
Lookup Geo Surrogate key
515,215 rows
Lookup Hotel Surrogate key
Lookup Match Output (515,215 rows)
Lookup Tweet Surrogate Key
Lookup Match Output (515,215 rows)
Lookup Date Key from Dim Date
Lookup Match Output (515,215 rows)
Lookuup Tweets Stage Data
Lookup Match Output (515,215 rows)
Sort Fact Table
515,215 rows
Success

OLAP CUBE

View of Cube in SQL Server Analysis Service and Analysis Server in SQL Server Management Studio

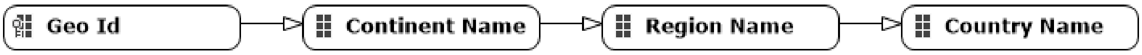


- Solution 'Hotel_Review' (1 project)
  - **Hotel_Review**
    - Data Sources
      - DIMENSION.ds
    - Data Source Views
      - DIMENSION.dsv
    - Cubes
      - DIMENSION.cube
    - Dimensions
      - Dim Geo.dim
      - Dim Hotel.dim
      - Dim Tweet.dim
      - Dim Date.dim
    - Mining Structures
    - Roles
    - Assemblies
    - Miscellaneous

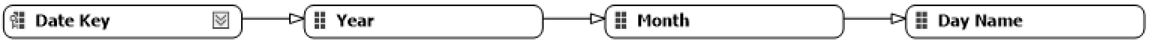


- Hotel_Review
  - Data Sources
  - Data Source Views
    - DIMENSION
  - Cubes
    - DIMENSION
      - Measure Groups
        - Fact
  - Dimensions
    - Dim Date
    - Dim Geo
    - Dim Hotel
    - Dim Tweet
  - Mining Structures
  - Roles
  - Assemblies

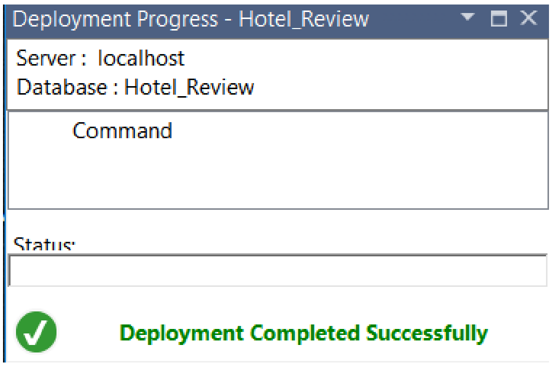
Attribute Relationships for Geography Dimension:



Attribute Relationships for Date Dimension:



OLAP Cube Deployment



Deployment Progress - Hotel_Review

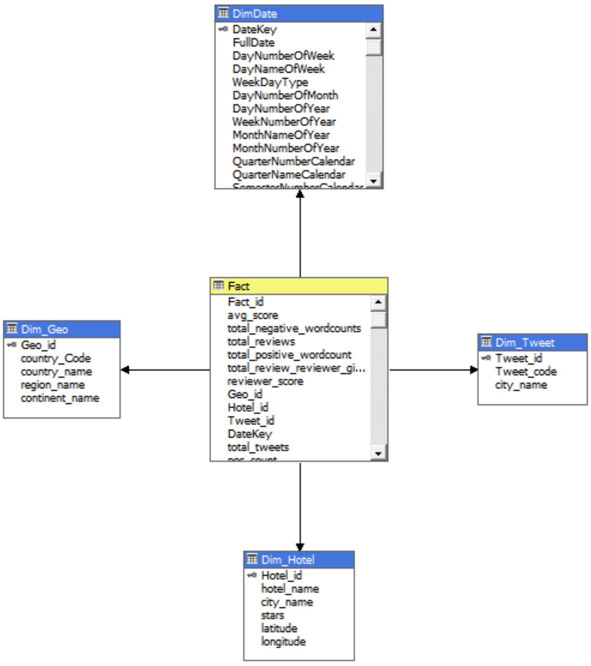
Server : localhost
Database : Hotel_Review

Command

Status:

Deployment Completed Successfully

OLAP Cube Star Schema

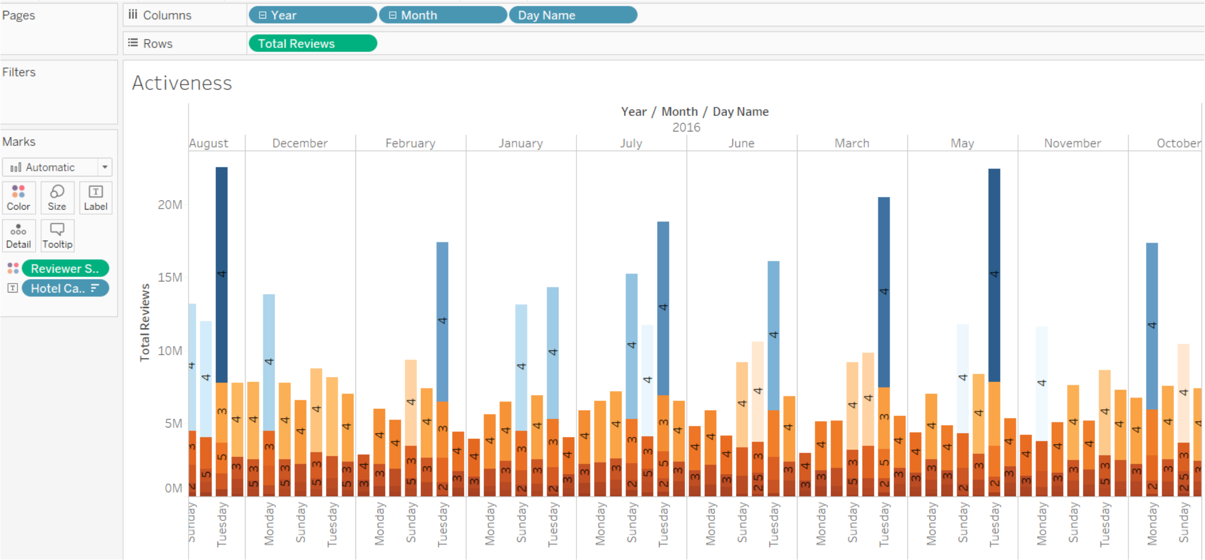


Business Intelligence

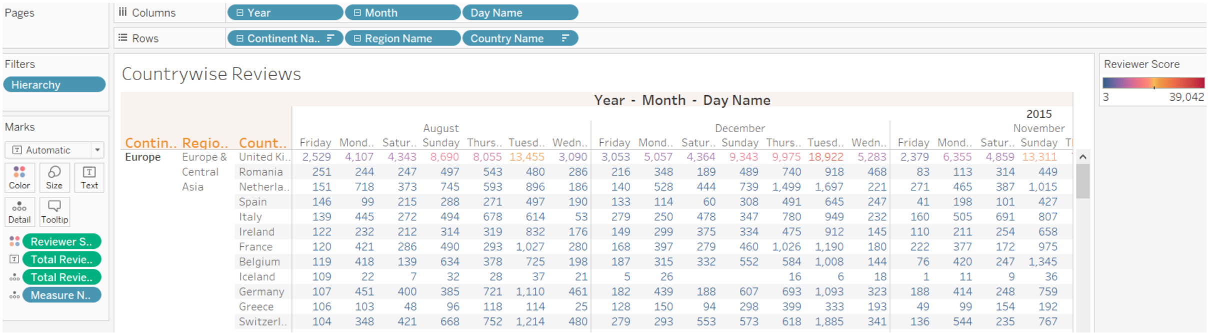
Here in this project got number of opportunity to slice and Dice Cube and find answer to different Business queries, but in this project ill be representing four main Business Queries I found most attractive and provide insight in terms of choosing hotel before booking on basis of daily and historical 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?



Columns: Year, Month, Day Name
Rows: Continent Na.., Region Name, Country Name

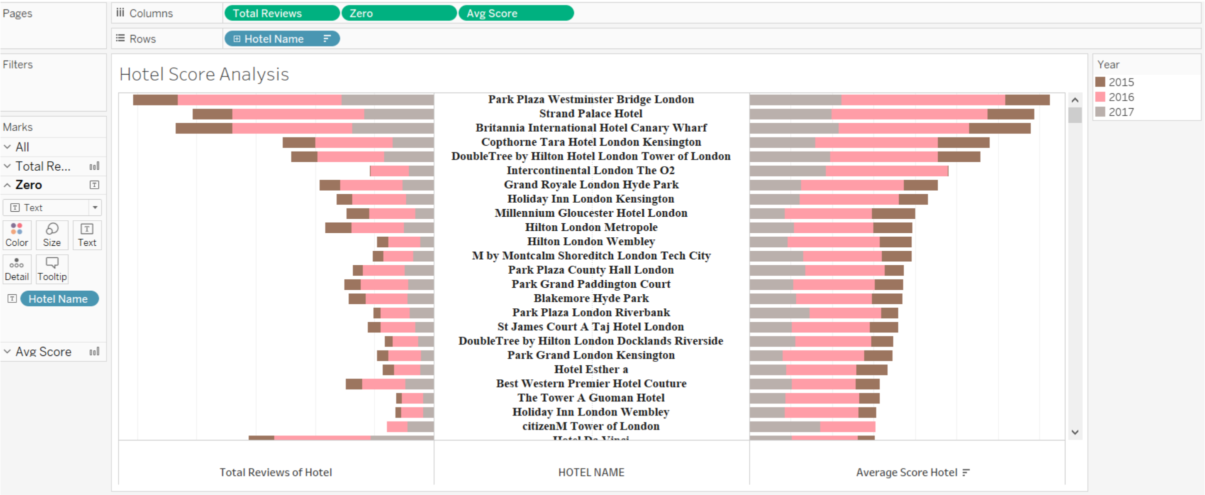
Countrywise Reviews

Year - Month - Day Name

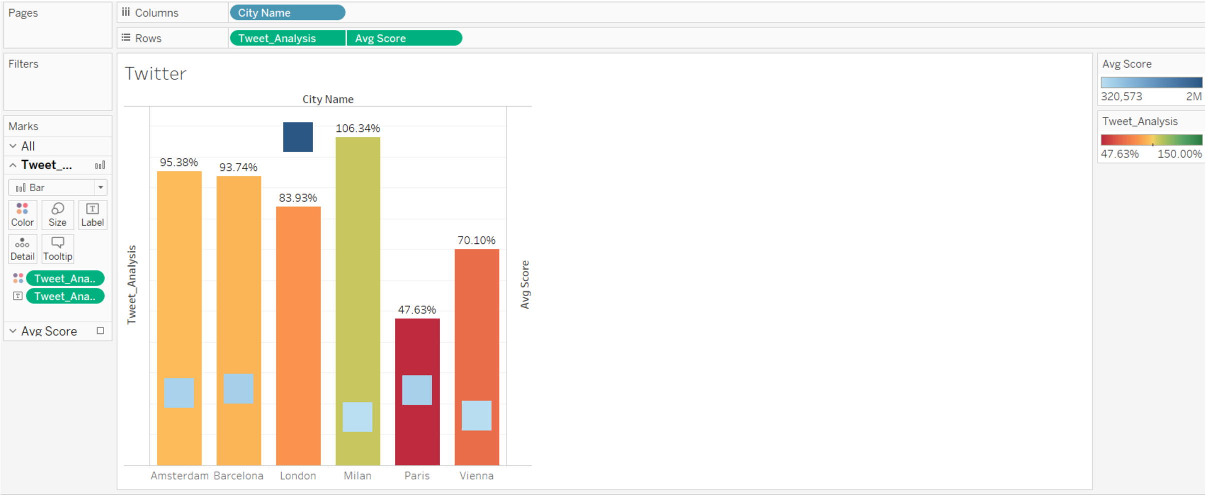
																			2015	
						August						December							November	
Contin..	Regio..	Count..	Friday	Mond..	Satur..	Sunday	Thurs..	Tuesd..	Wedn..	Friday	Mond..	Satur..	Sunday	Thurs..	Tuesd..	Wedn..	Friday	Mond..	Satur..	Sunday
Europe	Europe & Central Asia	United Ki..	2,529	4,107	4,343	8,690	8,055	13,455	3,090	3,053	5,057	4,364	9,343	9,975	18,922	5,283	2,379	6,355	4,859	13,311
		Romania	251	244	247	497	543	480	286	216	348	189	489	740	918	468	83	113	314	449
		Netherla..	151	718	373	745	593	896	186	140	528	444	739	1,499	1,697	221	271	465	387	1,015
		Spain	146	99	215	288	271	497	190	133	114	60	308	491	645	247	41	198	101	427
		Italy	139	445	272	494	678	614	53	279	250	478	347	780	949	232	160	505	691	807
		Ireland	122	232	212	314	319	832	176	149	299	375	334	475	912	145	110	211	254	658
		France	120	421	286	490	293	1,027	280	168	397	279	460	1,026	1,190	180	222	377	172	975
		Belgium	119	418	139	634	378	725	198	187	315	332	552	584	1,008	144	76	420	247	1,345
		Iceland	109	22	7	32	28	37	21	5	26			16	6	18	1	11	9	36
		Germany	107	451	400	385	721	1,110	461	182	439	188	607	693	1,093	323	188	414	248	759
		Greece	106	103	48	96	118	114	25	128	150	94	298	399	333	193	49	99	154	192
		Switzerl..	104	348	421	668	752	1,214	480	279	293	553	573	618	1,885	341	136	544	235	767

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

After analysing 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 reviewer.


Pages
Columns
Total Reviews
Zero
Avg Score
Rows
Hotel Name
Filters
Marks
All
Total Re...
Zero
Text
Color
Size
Text
Detail
Tooltip
Hotel Name
Avg Score
Hotel Score Analysis
Park Plaza Westminster Bridge London
Strand Palace Hotel
Britannia International Hotel Canary Wharf
Copthorne Tara Hotel London Kensington
DoubleTree by Hilton Hotel London Tower of London
Intercontinental London The O2
Grand Royale London Hyde Park
Holiday Inn London Kensington
Millennium Gloucester Hotel London
Hilton London Metropole
Hilton London Wembley
M by Montcalm Shoreditch London Tech City
Park Plaza County Hall London
Park Grand Paddington Court
Blakemore Hyde Park
Park Plaza London Riverbank
St James Court A Taj Hotel London
DoubleTree by Hilton London Docklands Riverside
Park Grand London Kensington
Hotel Esther a
Best Western Premier Hotel Couture
The Tower A Guoman Hotel
Holiday Inn London Wembley
citizenM Tower of London
Hotel Da Vinci
Total Reviews of Hotel
HOTEL NAME
Average Score Hotel
Year
2015
2016
2017

Business Query 4: Provide sentimental analysis from twitter to analyse Daily score of 6 cities (#Hotel+CityName) and compare it with historical score of hotels.



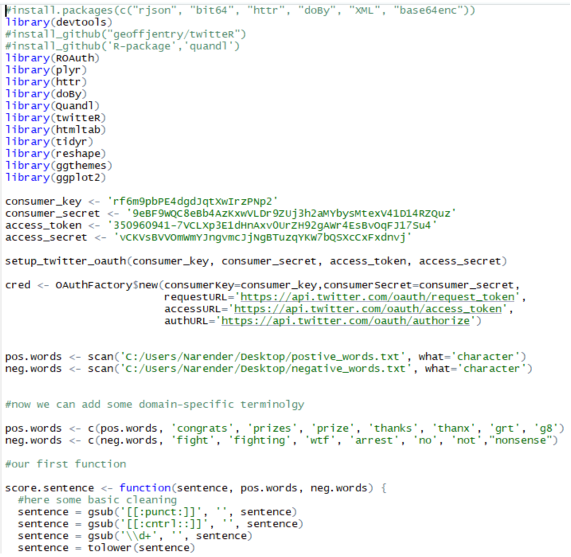
Analysis of Project in general:

This project can further be extended by automating the R Code and connecting it with SSIS to check daily analysis of Hotels of city.

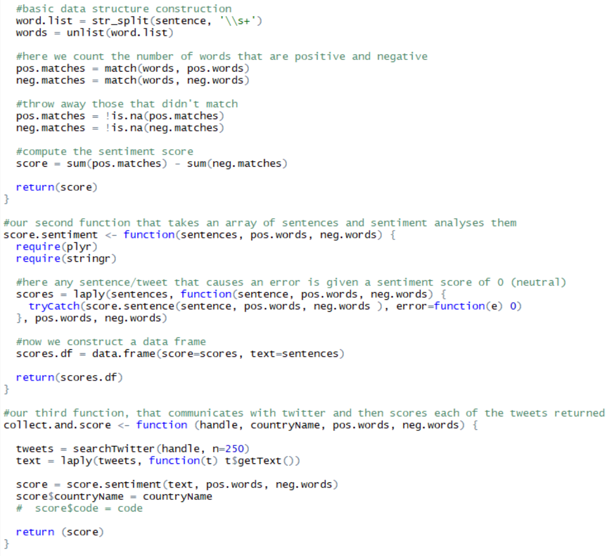
We can also further automated process to find analysis score for all 1500 hotels data for 6 different most famous cities in terms of Tourism and Analyse using this Daily analysis Engine with comparing it with historical Data.

Appendices I

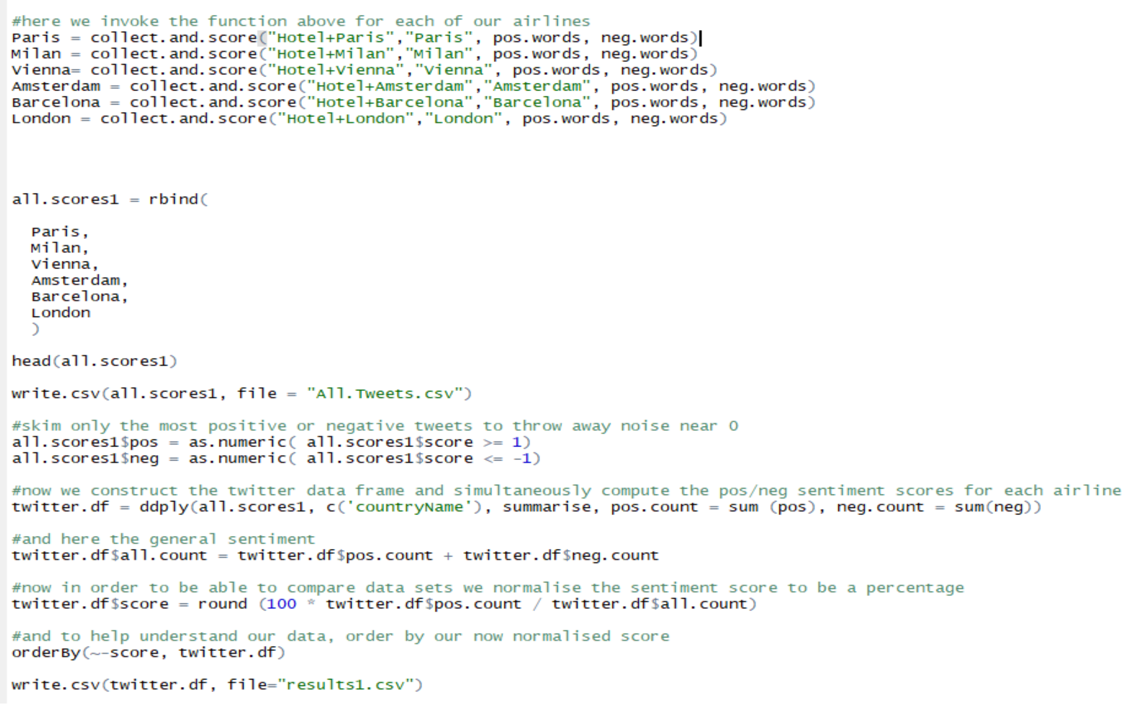
R Code



```
#install.packages(c("rjson", "bit64", "httr", "doBy", "XML", "base64enc"))
library(devtools)
#install_github("geoffjentry/twitteR")
#install_github('R-package','quandl')
library(ROAuth)
library(plyr)
library(httr)
library(doBy)
library(Quandl)
library(twitteR)
library(htmltab)
library(tidyr)
library(reshape)
library(ggthemes)
library(ggplot2)

consumer_key <- 'rf6m9pbPE4dgdJqtXwIrzPNp2'
consumer_secret <- '9eBF9WQC8eBb4AzKxwVLDr9ZUj3h2aMYbysMtexV41D14RZQuz'
access_token <- '350960941-7VCLXp3E1dHnAxvOUrZH92gAWr4EsBvOqFJ17Su4'
access_secret <- 'vCKVsBVVOmWmYJngvmcJjNgBTuzqYKW7bQSXcCxFxdnvj'

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

cred <- OAuthFactory$new(consumerKey=consumer_key,consumerSecret=consumer_secret,
                         requestURL='https://api.twitter.com/oauth/request_token',
                         accessURL='https://api.twitter.com/oauth/access_token',
                         authURL='https://api.twitter.com/oauth/authorize')


pos.words <- scan('C:/Users/Narender/Desktop/postive_words.txt', what='character')
neg.words <- scan('C:/Users/Narender/Desktop/negative_words.txt', what='character')


#now we can add some domain-specific terminolgy

pos.words <- c(pos.words, 'congrats', 'prizes', 'prize', 'thanks', 'thanx', 'grt', 'g8')
neg.words <- c(neg.words, 'fight', 'fighting', 'wtf', 'arrest', 'no', 'not',"nonsense")

#our first function

score.sentence <- function(sentence, pos.words, neg.words) {
  #here some basic cleaning
  sentence = gsub('[[:punct:]]', '', sentence)
  sentence = gsub('[[:cntrl::]]', '', sentence)
  sentence = gsub('\\d+', '', sentence)
  sentence = tolower(sentence)
```



```
  #basic data structure construction
  word.list = str_split(sentence, '\\s+')
  words = unlist(word.list)

  #here we count the number of words that are positive and negative
  pos.matches = match(words, pos.words)
  neg.matches = match(words, neg.words)

  #throw away those that didn't match
  pos.matches = !is.na(pos.matches)
  neg.matches = !is.na(neg.matches)

  #compute the sentiment score
  score = sum(pos.matches) - sum(neg.matches)

  return(score)
}

#our second function that takes an array of sentences and sentiment analyses them
score.sentiment <- function(sentences, pos.words, neg.words) {
  require(plyr)
  require(stringr)

  #here any sentence/tweet that causes an error is given a sentiment score of 0 (neutral)
  scores = laply(sentences, function(sentence, pos.words, neg.words) {
    tryCatch(score.sentence(sentence, pos.words, neg.words ), error=function(e) 0)
  }, pos.words, neg.words)

  #now we construct a data frame
  scores.df = data.frame(score=scores, text=sentences)

  return(scores.df)
}

#our third function, that communicates with twitter and then scores each of the tweets returned
collect.and.score <- function (handle, countryName, pos.words, neg.words) {

  tweets = searchTwitter(handle, n=250)
  text = laply(tweets, function(t) t$getText())

  score = score.sentiment(text, pos.words, neg.words)
  score$countryName = countryName
  #  score$code = code

  return (score)
}
```



```
#here we invoke the function above for each of our airlines
Paris = collect.and.score("Hotel+Paris","Paris", pos.words, neg.words)
Milan = collect.and.score("Hotel+Milan","Milan", pos.words, neg.words)
Vienna= collect.and.score("Hotel+Vienna","Vienna", pos.words, neg.words)
Amsterdam = collect.and.score("Hotel+Amsterdam","Amsterdam", pos.words, neg.words)
Barcelona = collect.and.score("Hotel+Barcelona","Barcelona", pos.words, neg.words)
London = collect.and.score("Hotel+London","London", pos.words, neg.words)



all.scores1 = rbind(

  Paris,
  Milan,
  Vienna,
  Amsterdam,
  Barcelona,
  London
  )

head(all.scores1)

write.csv(all.scores1, file = "All.Tweets.csv")

#skim only the most positive or negative tweets to throw away noise near 0
all.scores1$pos = as.numeric( all.scores1$score >= 1)
all.scores1$neg = as.numeric( all.scores1$score <= -1)

#now we construct the twitter data frame and simultaneously compute the pos/neg sentiment scores for each airline
twitter.df = ddply(all.scores1, c('countryName'), summarise, pos.count = sum (pos), neg.count = sum(neg))

#and here the general sentiment
twitter.df$all.count = twitter.df$pos.count + twitter.df$neg.count

#now in order to be able to compare data sets we normalise the sentiment score to be a percentage
twitter.df$score = round (100 * twitter.df$pos.count / twitter.df$all.count)

#and to help understand our data, order by our now normalised score
orderBy(~-score, twitter.df)

write.csv(twitter.df, file="results1.csv")
```