Datawarehousing And Business Intelligence

Topic: Analysis of Top Restaurents

SUBMITTED BY:

Student ID: X17126339
Student Name: Alankrita Khadtare
Course Name : MSc in Data Aanalytics

SUBMITTED TO:

Prof. Noel Cosgrave
Associate Faculty
School of Computing

Contents

1 Introduction

Lodging and hoteling has become inseparable part of working people around the world due to global spread of distributed businesses. People find the perfectly suited places for comfortable staying at restaurant via ratings from Social Medias like Facebook, Twitter, micro-blogs, and so on. Restaurants, on the hand, use the same sources to reach the customers by making them provide reviews for further decision making and growing their business. Professional bodies like AAA (America Auto Association) publish Diamond Rating Guidelines to evaluate quality of restaurants [Deck Weight (2009)]. Thus, Online reviews of the customers and their ratings help customers like dinners as well as organizations like restaurants. The purpose of this project is to acquire customer review data from social media and build the environment for effective analysis useful for restaurant business.

Social media contains structured, semi-structured and unstructured data. It is complex data huge in size and needs proper storage structure, efficient computing environment and appropriate technology tools to gain insights from data. Current platforms for processing such big data are: i) Data warehouse building and business analytics, ii) Machine Learning based modelling of tasks on high performance, parallel-distributed computing environment like Compute Unified Device Architecture (CUDA), iii) Hadoop-MapReduce and SPARK environments. In our project, the focus is on building data warehouse for restaurant's online customer reviews data collected from different social media sources. We employ natural language processing concepts via ETL tools and build data warehouse and perform multi-dimensional data analytics and also visualize sample query results.

The report proceeds as follows. We start with description of data sources and their requirement for data warehouse building and querying. Next section follows with review of related literature wherein we describe role of social media for customer knowledge management and business strategies planning, computing platforms for social media big data analytics, and methodologies of sentiment analysis. Next section follows about data warehouse architecture and description about selecting specific architecture for restaurant review data. We list technologies and tools required for implementation. Next we proceed with different steps of modelling data warehouse with appropriate ETL strategy from designing schema to developing populated deployable data cube. Finally, we present results of various case studies with visualization tools.

2 Data Sources

As per the requirements of the project and building the data warehouse, data is gathered from five different sources, four structured and one unstructured. The data was processed, combined and manipulated to obtain value for our analysis.

2.1 Limitations

- As restaurants is a huge domain, it is not possible to retrieve data for all top restaurants worldwide. Therefore, we initially fetched data for top 250 restaurants amongst which we focussed on restaurants in USA which counted to 27.
- Also, we fetched the data for one year because of huge size of data.

1. Dataset Source – 1 (Structured Data):

Our first structured data is downloaded from the statista website[1] for top 250 restaurant companies worldwide. It was published in July 2018. We then filtered the downloaded data in R concentrating on top restaurants in USA which counts to 27 to conduct our analysis. The attributes of the dataset include Company name, Listing ID, Exchange, ISIC code, Street, City, Region, Zip Code, Phone and Website. We included this dataset because we want to analyse the business insights in the distribution of sales for different restaurant companies at different places in USA.

2. Dataset Source – 2 (Structured Data):

One of the data files were taken from the Yahoo finance website [2]. For 27 restaurants in USA, we obtained the stock details on monthly basis from August 2017 to July 2018(12 months*27 restaurants = 324 rows) in R. There is no publish date for the data downloaded from this website as data is updated every second. This dataset was included to find insights about the varying stock value prices for different restaurant companies during this period. It was taken to analyse if the stock details parameters such as Average, High, Low, Close and Adjacent close values affected the Sentiment score rating.

3. Dataset Source – 3 (Structured Data):

Our third source of structured data was fabricated in excel for building the data warehouse. This dataset comprises of fields such as Company name, Date, Revenue and Employees. It includes 324 records for top 27 restaurants in USA which are same as those obtained from statista for 12 months period. This dataset was included in order to know the total sales revenue for the restaurant companies during a time period. Also, it was included to find if there were any other factors impacting Sales data.

[1] https://www.statista.com/study/44764/top-100-restaurants/
[2] https://finance.yahoo.com/quote/USFD/history?p=USFD

4. Dataset Source – 4 (Structured Data):

This is our second source of structured data fabricated in excel. The dataset includes the fields Company, Listing ID, Cuisines, Rating(5-pointer scale), Service, Food, Value and Atmosphere for 27 restaurants. The motivation behind including this dataset was to analyse if customers were inclined to specific cuisines at specific restaurants. Also, it served in the form of feedback rating and the restaurant service.

5. Dataset Source – 5(Unstructured Data)

We obtained the reviews based on tweets for top 27 restaurants in USA which is our unstructured source fetched from the website yelp.ie [3]. To achieve this, firstly, tweets were fetched in R language by calling API web services for these top 27 restaurants. 3704 tweets were obtained for each of these 27 restaurants from the period August 2017 to July 2018. That means, we obtained 308 tweets*12 months*27 restaurants = 99,792 tweets. We then stored these tweets in a CSV file which is now converted to a structured file format which includes Date, Sentiment score, Positive Sentiment, Negative Sentiment and Company Name. We obtained reviews for building our warehouse as this is the best way to know if there was any relation between sentiment scores and stock details.

3 Literature Review

Social media has become a part of everyone's life on a daily basis sharing all possible opinions, views about everything, whether they liked it or not. This is called sentiment, which can be positive, negative or neutral. Analysing sentiments is an issue because of extensive data getting generated every second on different social media platforms. Traditional methods are not possible to process huge amount of structured, semi-structured and unstructured data getting generated and stored. High performance computing platforms like Data warehouses, Hadoop-MapReduce and CUDA from NVedia are appropriate to deal with big data getting generated from social medias like Facebook, Twitter , micro-blogs, videos, TV channel news and so on. Globally distributed businesses cannot survive without taking cognizance of customer knowledge via social media. Popularity and consistent growth of organizations like Starbucks are examples where social media adaptation have paved the way for their growth [Shirdastian et al. (2017)]. The following subsections reviews literature on analysis of online customer reviews with needful computing environments.

[3] https://www.yelp.ie/biz/mcdonalds-new-york-100?start=40

3.1 Online Restaurents Customer Reviews and Sentiment Analytics

The authors, Chua & Banerjee (2013), Gallaugher & Ransbotham (2010), analyzed the extent to which the use of social media can support the Customer Knowledge Management (CKM) in organizations relying on traditional business models. Data was retrieved from varied sources such as newspapers, magazines, scholarly publications and social media services. The methodology was focused on use of combination of qualitative case study and netnography on Starbucks. The paper elaborated comparative study of social media supported four CKM frameworks incorporated by Starbucks: i) Micro blogging services (twitter), ii) social networking services (Facebook), iii) local aware mobile services (foursquare) and iv) corporate discussion forum services (mystarbucksidea). The study showed that dynamic customer centric approach for making social media tools for CKM served as effective branding and marketing instrument for the Starbucks Corporation. The branding and customer marketing are possible business solutions obtained from analysis of customer knowledge management data by raising complex business intelligence (BI) queries. The authors, Shirdastian, Laroche & Richard (2017), have investigated customer sentiment analysis of i) brand sentiments and ii) brand authenticity in social media on twitter data from Starbucks. They used qualitative as well as quantitative methods for sentiment analysis and brand authenticity prediction. Social media constituents are social networks (Facebook), Micro-blogs (Twitter), Blogs (Blogger and WordPress), social book marking and review sites like yelp. Companies like Starbucks strengthen customer relationship and consequently increase firm's revenue and profit by analyzing social media big data and business intelligence.

Multiple restaurants can be rated according to three main aspects of AAA, Restaurants Diamond Rating Guideline [Deck Weight (2009)], which focuses on the quality of food provided, the service received and the ambience around. The authors carried out sentiment analysis on restaurants online customers reviews with five hypotheses; more positive customer's sentiment about food, service, ambience, pricing and special contexts, more will be customer's review. For this, data was collected from Yelp Dataset Challenge provided by Yelp Inc. Sentiment analysis was performed using semi-supervised machine learning approach. This paper is without limitations and these five aspects are equally important in rating a restaurant. The future work of this paper focused on generating "data-driven"list of aspects to rate restaurants, which are derived by purely mining review texts.

The authors, Joorabchi & Mahdi (2008), aimed at gathering restaurant parameters from user's reviews, structure it and feed the recommendation module- system that identifies and provides recommended digital or content items for users. The paper mainly focused on the extraction of parameters like service quality, food quality, cuisine type, price level, noise level, etc. The authors used bootstrapping for trigger words dictionary learning as part of restaurant in-

formation extraction and simple semi-supervised scheme for the identification of new trigger words. To study the languages of "real-word" text, Corpus analysis based on non-contiguous bigrams was used. It incuded tokenization, lemmatization, sentence splitting, trigger words-short, precise- dictionaries construction using bootstrapping method and pattern development, i.e. part-of speech (POS) distribution analysis. Future work of this paper focused on VP pattern construction, object oriented sentiment analysis with restaurant.

3.2 Social Media Computing Environments and Technologies

The authors, Verma et al. (2015), processed the data effectively and efficiently using Hadoop framework. The paper provides an approach that mines unstructured sentiment information, refinement can be made through iterative process of cleaning /pre-processing, which would ultimately eliminate the outliers and noise from the data source. It was a challenge to extract the real-time data especially in networking architecture. So the authors of this paper have described a cloud-centred platform with software designed network (SDN). The framework designed in this paper extracted and analysed the opinions for social customer relationship management (CRM) which gave general opinion about the client fulfilment [Dasgupta et al. (2015)]. A recommendation system was designed describing ranks and rating of various items. It was based on user-based collaborative filtering algorithm in which key-words extracted were from passive users' review and used to indicate user preferences. To increase the scalability and efficiency the proposed recommendation system was built on top of Hadoop. Future work of this paper was the implementation of the technologies that they mentioned to improve the performance of Hadoop MapReduce Framework.

The data are growing massively in terabytes or even petabytes, and even some are hidden. Time spent in these data performing tasks is under estimated. In order to crawl huge amount of data, there could be any mishaps like power failure, and later crawling of same data wouldn't be possible because of no URL or link. So, data needs to get mined in order to build data warehouse for easy processing. The authors, Bansal & Kagemann (2015), focused on data warehouse building for data mining in library also called as Biblio-Mining. The first step was to identify the data sources, i.e. internal data (present in the library) and external data (not present in the library) sources. This paper focused on external data sources. For this purpose, ETL (Extract, Transform, Load) tool were used. Extraction was time consuming because of removing or taking out necessary data from different sources. Data extraction contained files in BibTex format which had standard entry type and format. The author has said that keywords and linkscan get crawled. Keywords and links gets crawled from the link provided in the BibTex file. Transformation is the next step in ETL. It ensures data consistency, data integrating, and mapping external data into the DW. For this, a flexible data model was created that supported XML (provides links to download BibTex information and convert into XML). Cleansing of noisy data,

missing values were also performed. XML simplifies, validate, transform data and generate desired data. Lastly, data was loaded into the data warehouse. Here, the data needed to get updated to maintain its consistency. Tables were created and data retrieved using necessary SQL commands. Fast recovery of data, random addition of file, new entry log, failure and duplications were removed. Iterative use of ETL enhanced crawling and cleansing process in order to achieve the consistency and guaranteed an updated data warehouse.
A data warehouse architectures proposed by Ralph Kimball's has bottom up approach and Bill Inmon's top down approach [Han et al. (2011), Sen & Sinha (2005)]. The authors, Jukic (2006), elaborated on how Kimball's approach is faster than the Inmon's approach and doesn't need the data model to be in normalized form. Kimaball's approach is easy to manage as it occupies less space in the database.

In summery, social media has an impact on growth of businesses. Social media generated big data analytics platforms and services are atmost important to both customers and organizations alike. Enterprise data warehousing and data analytics plays vital role among many other approaches. Our project deals with building data warehouse on online customer reviews collected from restaurant websites and analysis of their business issues.

4 Data Warehouse Architecture

Keeping in mind the two common approaches derived from Ralph Kimball's bottom-up approach and Bill Inmon's top-down approach to build a data warehouse, this Data Warehouse is being built on the design introduced by Ralph Kimball [Han & Kamber (n.d.), Sen & Sinha (2005)]. The author Jukic (2006) elaborated on how Kimball's approach is faster than Inmon's approach and doesn't need the data to be in normalized form. We chose to design our data warehouse using Kimball's approach for the following reasons:

- Data is collected from different sources which is not normalized. This helps in optimizing the process and achieve quick-win[Sansu (n.d.)].
- Kimball's process works in stages- initializing the repository of data(data mart) and then adding up other data marts with the data flowing from source to data marts and finally to the data warehouse, thus forming a bus architecture[Ariyachandra & Watson (2010)].
- Data warehouse can be built in a short time period and requires less maintenance. Changes can be made in the design in a slowly changing time-frame[Sansu (n.d.)]
- Performance is much better as there is no complexity.

- The BI tools can easily drill within the schema due to the bottom-up approach. This makes navigation of data much more efficient.

The following is the architecture diagram for our Restaurants Data Warehouse:

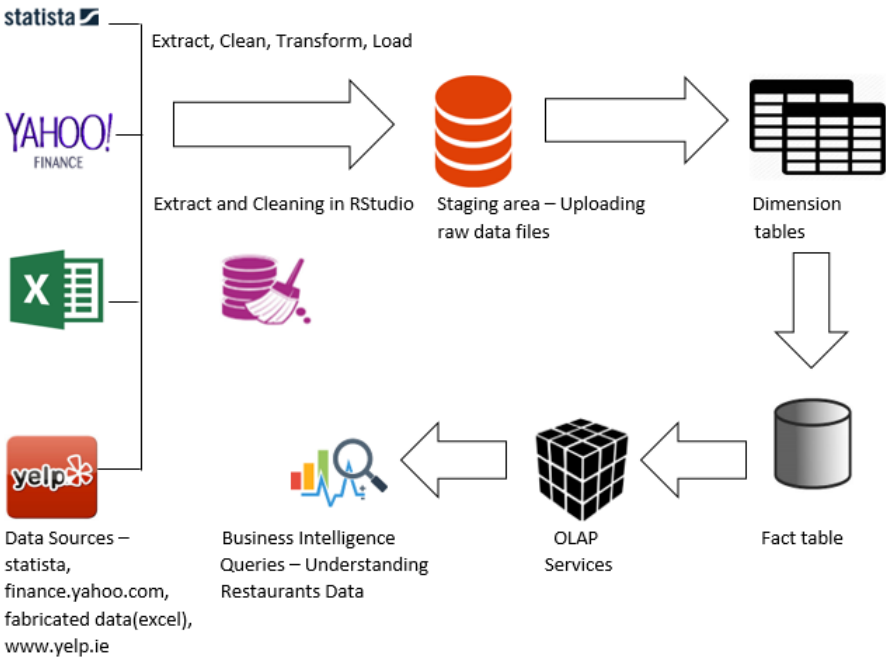

Figure 1: Data Warehouse Architecture for Restaurants Data.

5 Technologies Used in the Implementation of Data Warehouse Project:

- MS Excel For cleaning the structured datasets like filtering rows, removing unwanted data and deleting duplicates.
- R Studio to fetch the reviews based on tweets by calling API web services according to top restaurants in USA and extracting them in CSV file. Also, R was used for cleaning purpose.
- MS SQL Server Management Studio is used for creating the database. In our case, Restaurant_chain. Also, for creating the required tables in the database.
- SSIS - For loading the raw data tables into the staging area, then loading dimension and fact tables and finally loading them into the database server.
- SSAS is used for creating and deploying the OLAP cube.
- Tableau is used to run BI queries and make analyses from visualizations.

6 Data Warehouse Data Model

In this section, we will be discussing the data model for our Restaurants data and the dimensions that we have identified and the fact table.

6.1 Star Schema

We will be using Star schema to design our data warehouse. We are using star schema because

- It is easier to analyse and report BI needs using star schema relevant to businesses- in our case, Restaurants data.
- Large number of Data Marts, the subsets of data warehouses, usually created for different departments, are not being built in our case. Thus, saving storage space is not our priority[Kimball et al. (1998)].
- Star schema provides better processing efficiency as there is no query complexity and all the dimension tables are directly connected to the fact table.
- ETL Processing is easy in Star Schema [Kimball et al. (1998)].
- It simplifies future actions to the business users.

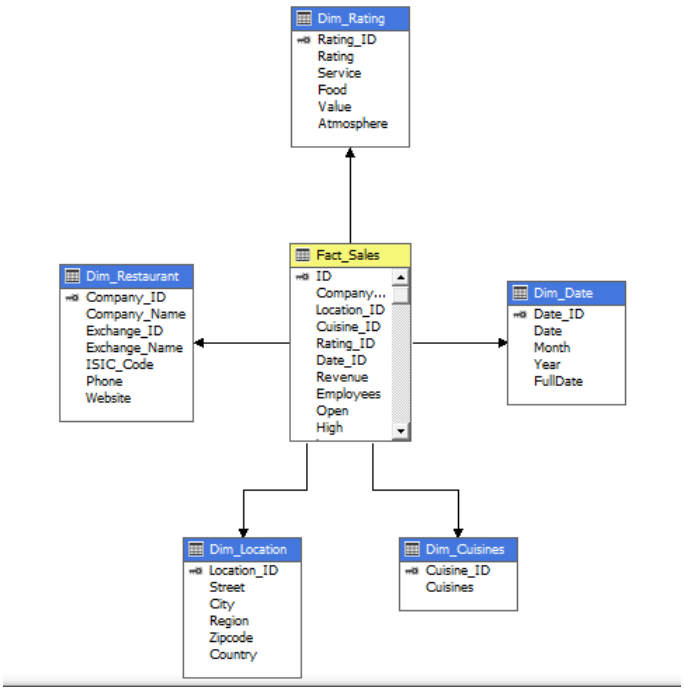

Figure 2: Data Modeling-RestaurantStarSchema.

The Fact_Sales at the centre stores aggregated data created from mentioned sources. It is connected to dimension tables Dim_Date, Dim_Restaurant, Dim_Rating, Dim_Cuisine and

Dim_Location with primary key and foreign key relationships between them. Above star schema diagram illustrates the relationship between the fact and dimension tables.

6.2 Dimension Tables

Here, we discuss our dimension tables in detail and the fields used from ourdimension and staging tablesto populate the fact table.

1. Dim_Location:
   As the name suggests, this dimension table contains the location attributes such as Street, City, Region, Zipcode and Country. This dimension is included because it gives us information about the restaurants data in different regions of USA. As shown in the above Star schema, our dimension table Dim_Location includes Location_ID, Street, City, Region, Zipcode and Country. This hierarchy facilitates drill-down operation in our cube[Gorla (n.d.)]

2. Dim_Restaurant:
   This dimension includes Company_ID, Company_Name, Exchange_ID, Exchange_Name, ISIC_Code, Phone and Website. This again helps in drill-down functionality in our cube[Gorla (n.d.)] This dimension is included to find insights in the business-related terms and patterns. It is included to know how restaurants businesses can be strengthened by analysing the stock details, number of employees, ratings and sentiment reviews.

3. Dim_Date:
   This dimension table contains all the date attributes such as Date_ID, Date, Month, FullDate and Year. This is important as it helps to analyse different parameters related to restaurants during a time period.

4. Dim_Rating:
   This dimension table contains Rating_ID, Rating, Service, Food, Value and Atmosphere. This dimension is included to find if there exists any relationship between these attributes and other restaurants data.

5. Dim_Cuisine:
   This dimension table includes Cuisine_ID and Cuisines. It is included to find if any specific restaurant served as a part of attraction for favourite cuisine at particular location. Also, it helps to find insights with other restaurant data such as stock details, ratings and sentiment score for particular restaurant.

6.3 Fact Table

Fact_Sales is our fact table which includes the ID which is unique identifier to identify each row and dimension IDs from dimension tables and measures from our staging tables. These dimension IDs include Company_ID, Location_ID, Cuisine_ID, Rating_ID and Date_ID and measures such as Open, High, Low, Close, Adj_Close, Volume, Revenue, Employees, Positive Sentiment and Negative Sentiment. The analysis for our business intelligence queries will be carried out via Fact_Sales.

7 Logical Data Map

The logical data map describes a table including the actual mapping from source to target tables(dimensions and fact table)[*Inside the Logical Data Map in Data Warehouse ETL Toolkit - Inside the Logical Data Map in Data Warehouse ETL Toolkit (8158)* (n.d.)]. It also includes transformations specific to cleaning and the exact manipulation of the data required to transform it from its original format to that of its final destination[*Inside the Logical Data Map in Data Warehouse ETL Toolkit - Inside the Logical Data Map in Data Warehouse ETL Toolkit (8158)* (n.d.)].

8 ETL Strategy

This is the most important part for building the data warehouse[8]. As the name suggests, the ETL strategy includes the processes for extracting the data from multiple sources, cleaning the unwanted data, transforming it into desired format to make it logical and ultimately loading it into the warehouse[8]. Microsoft's SSIS tool is used in our project for performing ETL process. The subsections below explain in detail the ETL strategy used to build our data warehouse.

8.1 Extraction

In the extraction process, we collected the data from multiple sources. Structured data was downloaded from two sources, the third and fourth structured sources were fabricated as already mentioned and for unstructured source, we collected the tweets from yelp.ie via R code. The data source links are described in detail in the data sources section above and the extraction code for tweets is documented in the appendix section. Also, we normalised the negative sentiment score(-1) for unstructured data to lie between 0 to 1 in R as database doesn't support negative values.

Target				Source			
Table Name	**Column Name**	**Data Type**	**Table Type**	**Table Name**	**Column Name**	**Data Type**	**Transformations**
Dim_Date	Date	int	Dimension	Stock_details	Date	DateTime	Get Day from Date
				Sentiment_Score	Date	DateTime	Get Day from Date
	Month	int	Dimension	Stock_details	Date	DateTime	Get Month from Date
				Sentiment_Score	Date	DateTime	Get Month from Date
	Year	int	Dimension	Stock_details	Date	DateTime	Get Year from Date
				Sentiment_Score	Date	DateTime	Get Year from Date
Dim_Restaurant	Company_Name	varchar	Dimension	Restaurant	Company Name	varchar	
	Exchange_ID		Dimension	Restaurant	Listing_ID	varchar	Case converted for consistency
	Exchange_Name	varchar	Dimension	Restaurant	Exchange	varchar	
	ISIC_Code	int	Dimension	Restaurant	Primary ISIC code	int	
	Phone	int	Dimension	Restaurant	Phone	int	Trimmed to have same length
	Website	varchar	Dimension	Restaurant	Website	varchar	
Dim_Location	Street	varchar	Dimension	Restaurant	Street	varchar	
	City	varchar	Dimension	Restaurant	City	varchar	
	Region	varchar	Dimension	Restaurant	Region	varchar	
	Zipcode	int	Dimension	Restaurant	Zipcode(if applicable)	int	
	Country	varchar	Dimension	Restaurant	Country	varchar	
Dim_Rating	Rating	float	Dimension	Ratings	Rating	int	
	Service	int	Dimension	Ratings	Service	int	
	Food	int	Dimension	Ratings	Food	int	
	Value	int	Dimension	Ratings	Value	int	
	Atmosphere	int	Dimension	Ratings	Atmosphere	int	
Dim_Cuisines	Cuisines	varchar	Dimension	Ratings	Cuisines	varchar	

Target				Source			
Table Name	**Column Name**	**Data Type**	**Table Type**	**Table Name**	**Column Name**	**Data Type**	**Transformations**
Fact_Sales	Company_ID	int	Fact	Dim_Restaurant	Company_ID	int	
	Location_ID	int	Fact	Dim_Location	Location_ID	int	
	Date_ID	int	Fact	Dim_Date	Date_ID	int	
	Open	int	Fact	Stock_details	Open	int	
	High	int	Fact	Stock_details	High	int	
	Low	int	Fact	Stock_details	Low	int	
	Close	int	Fact	Stock_details	Close	int	
	Adj_Close	int	Fact	Stock_details	Adj_Close	int	
	Volume	int	Fact	Stock_details	Volume	int	
	Revenue	int	Fact	Sales_Restaurant	Revenue	int	Rounded to million
	Employees	int	Fact	Sales_Restaurant	Employees	int	

8.2 Cleaning

The downloaded data files(CSV) contained large number of records for top 250 restaurants worldwide. It was then filtered using excel to focus on top restaurants in USA as already mentioned in the data sources. Additional columns in the dataset which were not required were removed using excel. Missing values occurred for the columns but were ignored as they were not mandatory. Most of the cleaning was done using R using Execute Process Task component provided by SSIS. The Execute Process Task calls R file which does initial extraction of data from data sources such as getting tweets from yelp using Rest API, dataset cleaning, removing punctuation, making transformations to fit into warehouse and finally store it in a CSV file. Following is the screenshot for extracting and cleaning using R:

Figure 3: ETL - Cleaning Using Script Task

8.3 Transformations

SSIS has components for transforming the data into a desired format as data is moved further. We have used data conversion component to change the data type from regular string to Unicode string while loading the staging tables. Date was transformed into its appropriate format using derived column component to get date, month and year. Sort and Merge transformations are used for combining two sorted datasets based on values in their key columns.

8.4 Loading

In this step, we start creating the required tables in our database and then ultimately populate the fact table. For our data warehouse in the project, we have performed loading at the following stages:

1. Loading data into Staging area:

Here, we first added the Flat File source components for loading our data files(CSV) in a Sequence Container. We then loaded these files using the Excel Source component provided by the SSIS. Before loading the tables in the staging area, we emptied the data from the tables by executing 'Truncate Staging Area' to avoid data duplication once the data is reloaded. We have used the concept of 'bulk load' for loading our staging tables. The following are the screenshots for loading our structured and unstructured data into the staging area:

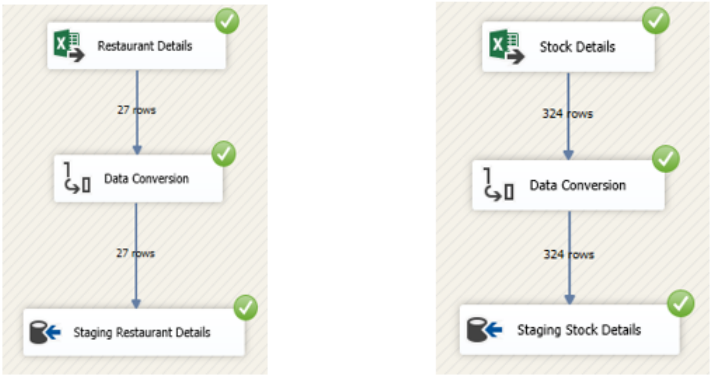

Figure 4: Loading Structured Data (Restaurant & Stock Details)

Figure 5: Loading Structured Data (Rating & Sales Details)

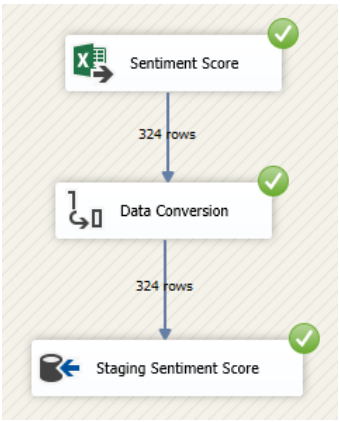

Figure 6: Loading Unstructured Data (Sentiment Score)

2. Loading data to the Dimension tables:

After loading the data to the Staging area successfully, we start creating the dimension tables for our warehouse. We have created the dimension tables- Dim_Location and Dim_Restaurant from the Restaurant table, Dim_Date from Sentiment_Score and Stock_Details and Dim_Rating and Dim_Cuisine from Rating table. Here also, we execute 'Truncate Warehouse Tables' to ensure there is no repetitive data before loading the dimension and fact tables.

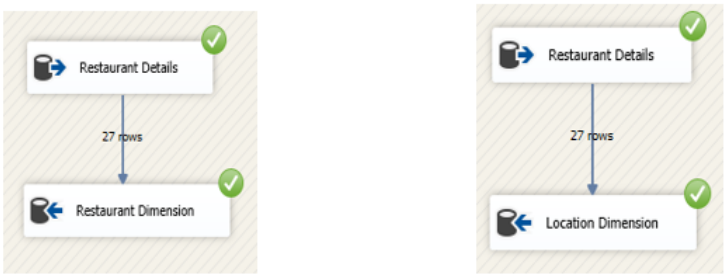

Figure 7: Loading Dimension Data (Restaurant & Location)

Figure 8: Loading Dimension Data (Cuisine & Rating Details)

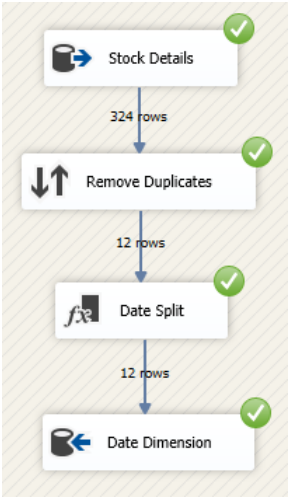

Figure 9: Loading Dimension Data (Date)

3. Populating data into the Fact table:

Once the data is loaded into the dimension tables, we used Sort and Merge join components by joining the data from staging and dimension tables to populate the Fact table. As shown in the below screenshot, we have first merged the Dimension_IDs of all the dimension tables and measures from our staging tables. Finally, we combined them together using merge join to load them into the fact table.

Figure 10: Loading Fact Data

Following screenshot shows the complete ETL execution process:

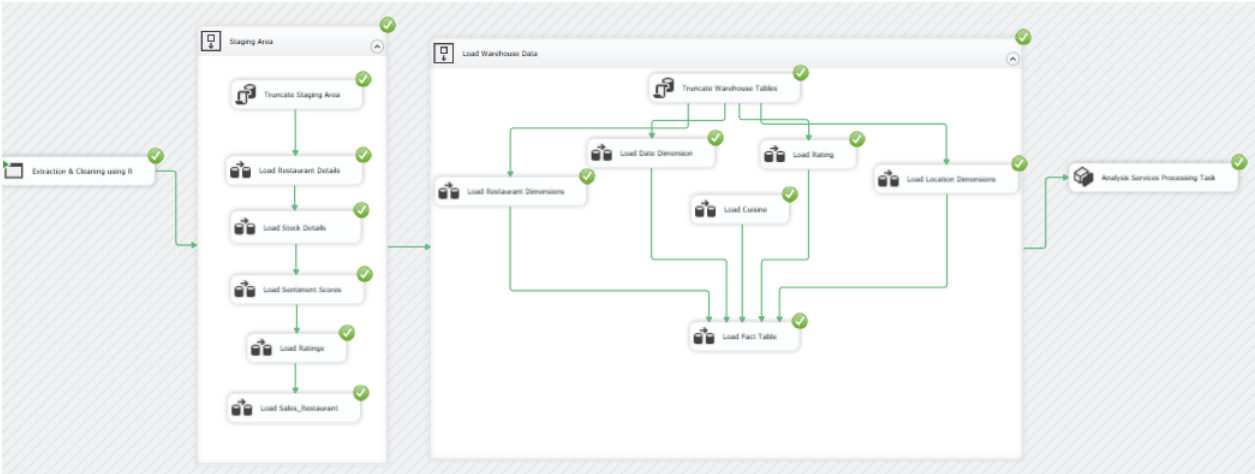

Figure 11: ETL Execution Process Flow

9 Cube Deployment

After loading all the tables, we go ahead in SSAS for creating and deploying the cube. As we have already deployed the cube using automation, we simply go and browse it:

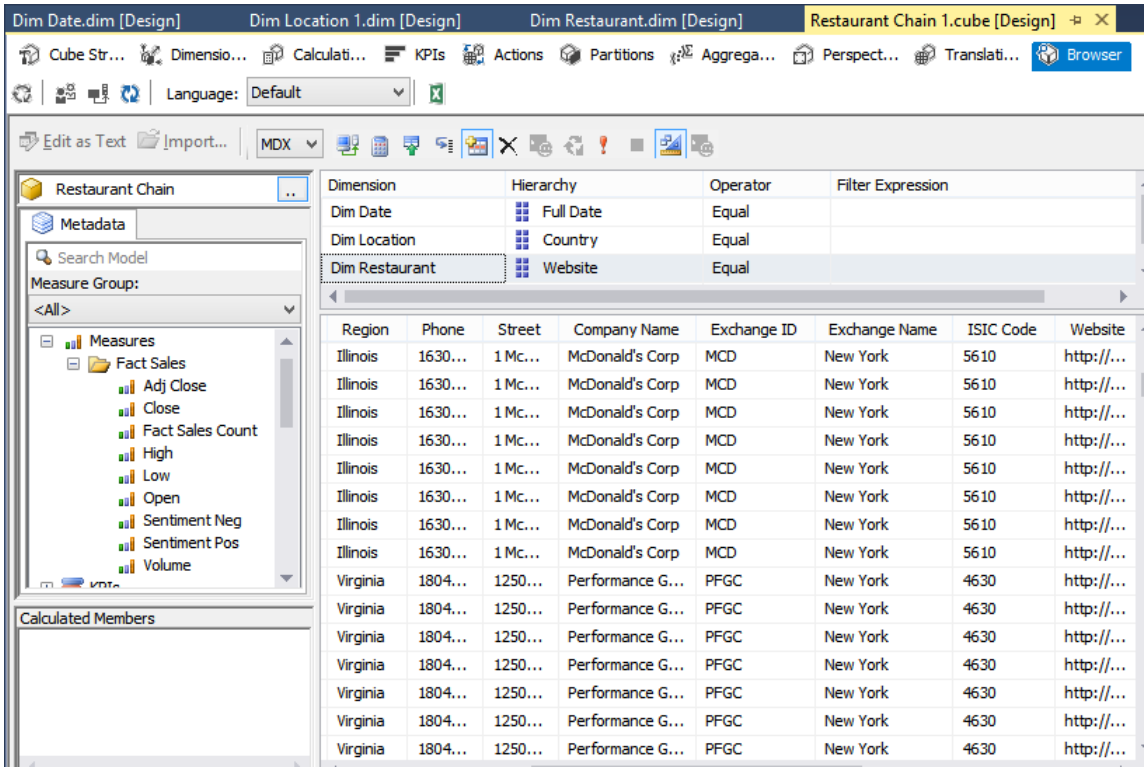

Figure 12: Automated Cube.

The above screenshot demonstrates that our cube has data in it and can be used for analysis.

10 Applications of DataWarehouse- Case Studies

Following are the Business Intelligence queries for analysing Restaurants data.We used tableau for analysing our case studies and visualizing data.

Case Study 1:

Is sentiment affecting the stock price ?

For conducting this analysis, we compared the fields positive and negative sentiment scores for open and close stock prices for 12 months period. The graph below demonstrates if sentiment scores had any effect on stock prices:

Data Sources:

1. positive, negative – Sentiment_Score.csv(Dataset Source - 5)
2. open, close – Stock_details.csv (Dataset Source - 2)

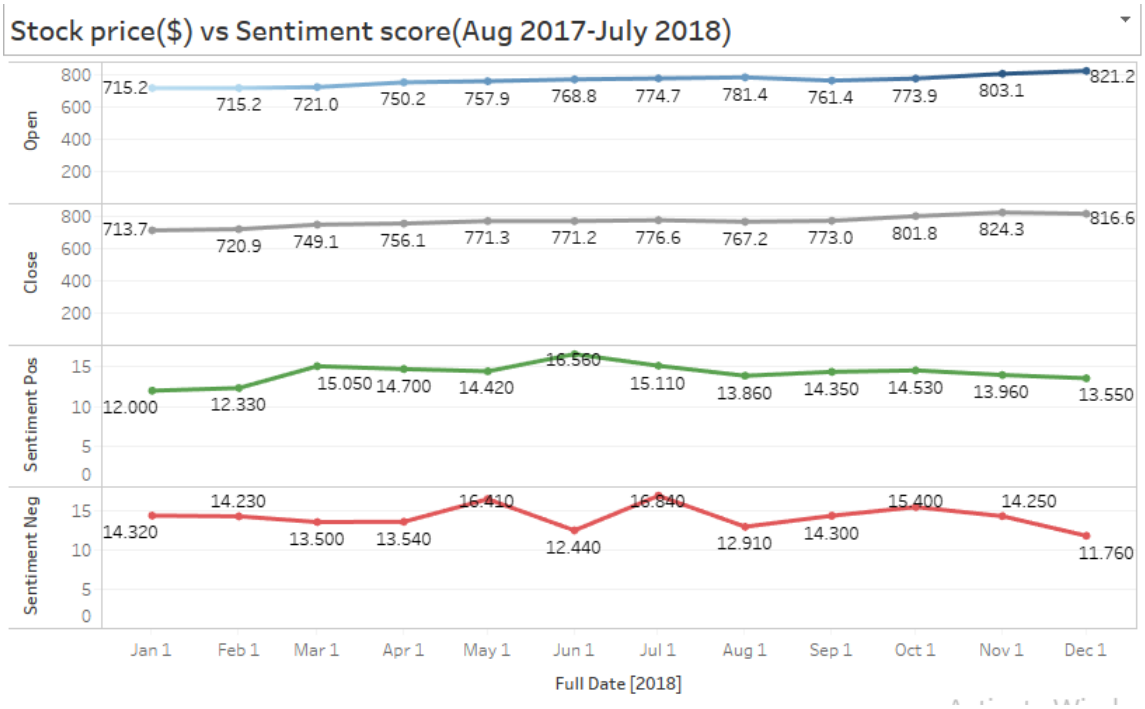

Figure 13: Case Study 1.

From the above graph, we can analyse that the overall stock prices are not much impacted by the sentiment scores although the sentiment scores show fluctuating records. The graph represents a consistent rise in the distribution of sales for both open and close values from January to December. Clearly, highest stock prices were recorded in December with open share prices crossing over 820 USD and close stock prices making about 820 USD. There was a slight increase in the positive sentiment score in the month of March and remained consistent in the same range between 14-15 until the month of May. Maximum positive reviews were recorded in June with the score crossing over 16 which also indirectly impacts the stock prices. The graph highlights a slight decrease in the positive reviews from people and remained consistent until December. Finally, we can analyse that restaurants consistently recorded negative reviews until April and thereafter fluctuated by the end of December. A sudden rise in the scores of negative sentiments were recorded in the months of May and July of over 16 with scores falling in June. Finally, we can analyse that the negative sentiment graph shows a slight rise for the months of September and October and falls for the following months.

Case Study 2:

How are the company share values scattered amongst all other companies?

Here, we used the fields Open stock value prices and Company Name for restaurants in a treemap to conduct our analysis. We also used high stock values for comparison purpose.

Data Sources:

1. Open, High – Stock_details.csv(Dataset Source - 2)

2. Company_Name – Restaurant.csv(Dataset Source - 1)

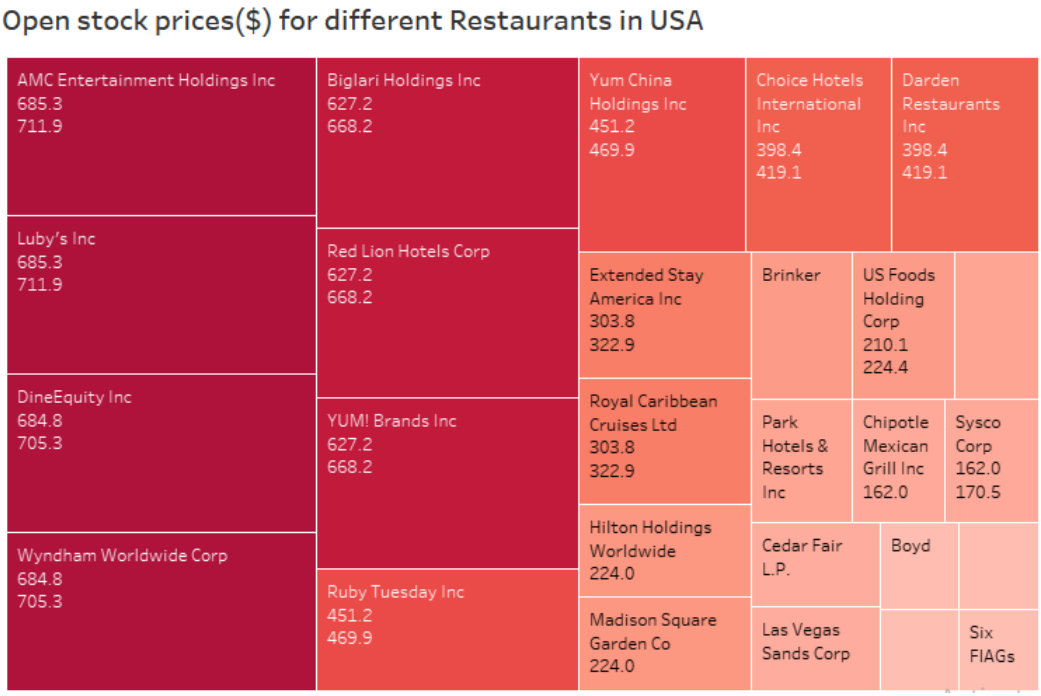

Figure 14: Case Study 2.

The above treemap clearly indicates that the Open stock value prices were highest for the restaurants AMC Entertainment Holdings Inc. and Luby's Inc. both making about 685 USD. Almost the same estimates of about 685 USD were observed by DineEquity Inc. and Wyndham Worldwide Corp. Same sales prices were recorded by Biglari Holdings Inc., Red Lion Hotels Corp. and YUM! Brands Inc. Ruby Tuesday and YUM! China Holdings crossed over 450 USD. Also, Choice Hotels and Darden Restaurants had similar estimates of about 420 USD. We can also analyse from the above treemap that a majority of restaurants had their sales in the range of 200-300 USD including the US Food Holding and Chipotle Mexican Grill. The Open stock prices were lowest for the restaurants namely, Boyd Gaming Corp., Performance Group, Aramark and Six FIAGs Entertainment Corp.

Case Study 3:

By considering the company share values which are the places where companies with higher and lower stock values are located?

For conducting this analysis, we used the fields Zipcode and the Latitude and Longitude co-ordinates. We also dragged companies' high stock value prices to analyse the distribution.

Data Sources:

1. Zipcode, Latitude, Longitude – Restaurant.csv(Dataset Source - 1)
2. High – Stock_details.csv(Dataset Source - 2)

The map below explains the high and low sales distribution for different restaurants at different places in USA.

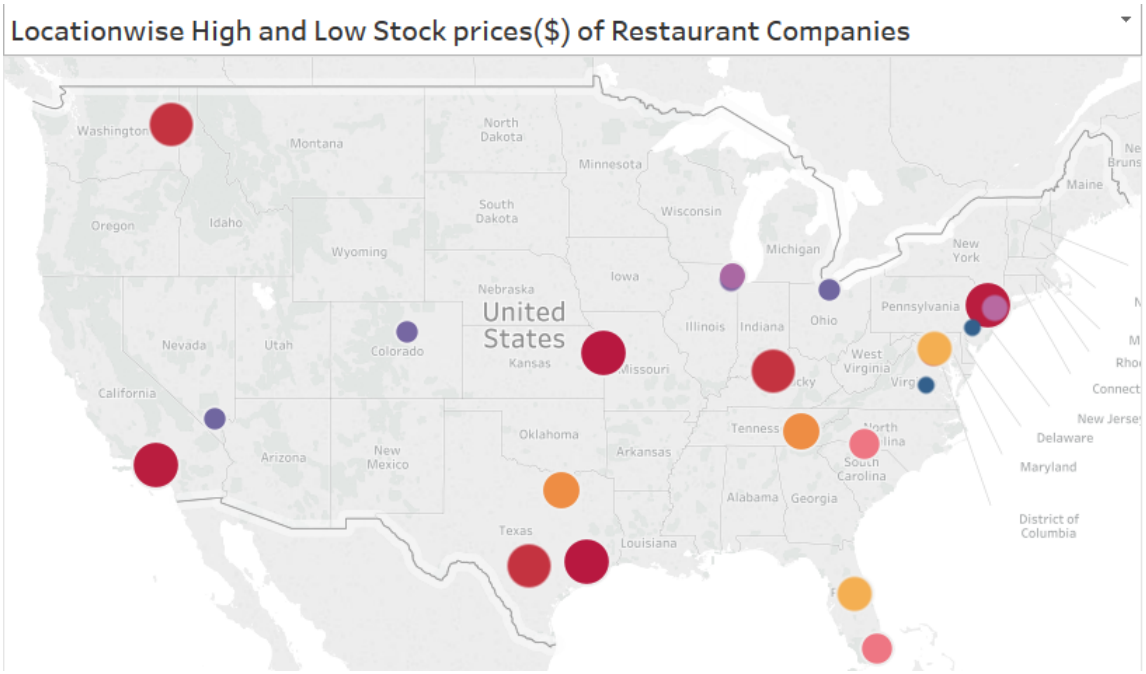

Figure 15: Case Study 3.

The above map clearly shows that the highest sales were observed by the restaurants AMC Entertainment Holdings Inc. and Luby's Inc. situated in Missouri and Louisiana both making 711.9 USD. They were then followed by Wyndham Worldwide Corp. in Pennsylvania, Dine Equity Inc in California, Red Lion Hotels Corp. in Washington and YUM! Brands Inc. and Kentucky. We can also analyse from the above graph that a large number of restaurants has its headquarters originating from areas between Pennsylvania and New York to Virginia in the south. Restaurants are sparsely located with fair share prices in the regions surrounding to Kentucky and South Georgia extending over to the coast. Restaurants with its highest market share were moderately distributed around Texas. There were very few restaurants in Washington and California although their stock prices were high.

11 Conclusion

The data warehouse built in our project was mainly focused on analysis of online customer reviews and stock details for Restaurants data. Identifying and extracting social media data from relevant data sources of websites was a difficult task. We used R code for extracting and preprocessing of complex unstructured data. Microsoft's tools were used for intermediate operations during modeling star schema and processing cube for our data warehouse.Finally, we formulated business queries and presented results using visualization tools such as tableau.

12 Appendix

R Code Used for Extracting Yelp Reviews

```
#Sentiment
#require(RSentiment)
install.packages('futile.logger', repos='http://cran.us.r-project.org')

#DOM and HTTP Request Processing
require(rvest)
require(httr)
require(httpuv)
require(XML)
require(jsonlite)

#String Manipulation
require(stringr)

#Database
library(RODBC)
#library(RSQLServer)

#Logger
require(futile.logger)

#Other
library(magrittr)

#Variables
api_base_url <- 'https://api.yelp.com/'
access_token <- NULL
restaurant_id <- NULL
restaurant_list <- c()

#Login Credentials
credentials <- list(
'grant_type' = 'client_credentials',
'client_id' = '2feX-HoiFo0XAKemgKxlTw',
'client_secret' = '
    shADXwgL8hMQkqYFr5udQFXfnwcYZuGXwIdpap0pdEdHXd7BLNQeofBRYmstnYtxZRZDllJKuRWHpDA
    -prjpCEXUByFPhY-GHfz9tqywGrCCk9P9mri2adnK3bvrW3Yx'
)

#get Access token
```

```
print('Getting access token')
access_token <- content(POST(paste(api_base_url, "oauth2/token", sep = ""),
    body = credentials))$access_token

#get Access token
csv_restaurant_list <-  read.csv('C:\\Users\\Alankrita\\Documents\\Data
    Warehouse\\Stock_2005-2017.csv', stringsAsFactors = FALSE)
recordCount <- length(csv_restaurant_list$Company)

#for each restaurant in the list get the Latitude and Longitude using API
    call
for (i in 1:recordCount) {
flog.info(paste("Getting Latitude and Longitude for restaurant ",'\n'))
#Framing request URL
Search_url <- modify_url(
api_base_url,
path = c("v3", "businesses", "search"),
query = list(
term = as.character('starbucks'),
location =as.character( csv_store_list\$location[i]),
radius = "100",
limit = 1
)
)

#Send Request and Get the search result
restaurant_detail <- GET(Search_url, add_headers('Authorization' = paste("
    bearer", access_token)))
restaurant_Name <- as.character(csv_restaurant_list$Company[i])
tryCatch({
reviews <- content(restaurant_detail)$businesses[[1]]$reviews  #get Review
    Count
}, error = function(e) {
flog.error('Error occured while searching for values')
review_Count <- 'NULL'
})
#Consolidate all the Column  into one consolidated list
restaurant_list <- rbind(restaurant_list, restaurant)
flog.info(paste('Processed', i, '....\n'))
}
#Assign column names to the list
colnames(restaurant_list) <-  c("Name", "id", "Latitude", "Longitude", "
    Review")
flog.info('Done with retriving Restaurant parameters')
#write.csv(restaurant_list, paste(EnvProp["output_file_location", 1], "
```

```
    Restaurant_Location_Details.csv", sep = ""))
flog.info('Stored data in csv file')

}


write.csv(restaurant_list, file="restaurant.csv")

restaurant_list <- read.csv("restaurant.csv")
restaurant_list\$review = gsub('https://','',restaurant_list\$review)
restaurant_list\$review = gsub('http://','',restaurant_list\$review)
restaurant_list\$review=gsub('[^[:graph:]]', ' ',restaurant_list\$review)
restaurant_list\$review = gsub('[[:punct:]]', '', restaurant_list\$review)
restaurant_list\$review = gsub('[[:cntrl:]]', '', restaurant_list\$review)
restaurant_list\$review = gsub('\\d+', '', restaurant_list\$review)
restaurant_list\$review=str_replace_all(restaurant_list\$review,"[^[:graph
    :]]", " ")

restaurant_list$review<- calculate_score(restaurant_list$review,1)
write.csv(restaurant_list, file="sentiment.csv")
```

References

Ariyachandra, T. & Watson, H. (2010), 'Key organizational factors in data warehouse architecture selection', *Decision support systems* **49**(2), 200–212.

Bansal, S. K. & Kagemann, S. (2015), 'Integrating big data: a semantic extract-transform-load framework', *Computer* **48**(3), 42–50.

Chua, A. Y. & Banerjee, S. (2013), 'Customer knowledge management via social media: the case of starbucks', *Journal of Knowledge Management* **17**(2), 237–249.

Dasgupta, S. S., Natarajan, S., Kaipa, K. K., Bhattacherjee, S. K. & Viswanathan, A. (2015), Sentiment analysis of facebook data using hadoop based open source technologies, *in* 'Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on', IEEE, pp. 1–3.

Deck Weight, J. E. (2009), 'Aaa five diamond designation: Is it worthwhile to convert?'.

Gallaugher, J. & Ransbotham, S. (2010), 'Social media and customer dialog management at starbucks.', *MIS Quarterly Executive* **9**(4).

Gorla, N. (n.d.), 'Features to consider in a data warehousing system', **46**(11), 111–115.
URL: *http://doi.acm.org/10.1145/948383.948389*

Han, J. & Kamber, M. (n.d.), *Data mining: concepts and techniques*, 3rd ed edn, Elsevier.

Han, J., Pei, J. & Kamber, M. (2011), *Data mining: concepts and techniques*, Elsevier.

Inside the Logical Data Map in Data Warehouse ETL Toolkit - Inside the Logical Data Map in Data Warehouse ETL Toolkit (8158) (n.d.).
URL: *https://www.wisdomjobs.com//e-university/data-warehouse-etl-toolkit-tutorial-201/inside-the-logical-data-map-8158.html*

Joorabchi, A. & Mahdi, A. E. (2008), A new method for bootstrapping an automatic text classification system utilizing public library resources, *in* 'Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science, Cork, Ireland (August 2008)'.

Jukic, N. (2006), 'Modeling strategies and alternatives for data warehousing projects', *Communications of the ACM* **49**(4), 83–88.

Kimball, R., Reeves, L., Ross, M. & Thornthwaite, W. (1998), *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*, John Wiley & Sons.

Sansu, G. (n.d.), 'Inmon or kimball: Which approach is suitable for your data warehouse?'. **URL:** *https://www.computerweekly.com/tip/Inmon-or-Kimball-Which-approach-is-suitable-for-your-data-warehouse*

Sen, A. & Sinha, A. P. (2005), 'A comparison of data warehousing methodologies', *Communications of the ACM* **48**(3), 79–84.

Shirdastian, H., Laroche, M. & Richard, M.-O. (2017), 'Using big data analytics to study brand authenticity sentiments: The case of starbucks on twitter', *International Journal of Information Management* .

Verma, J. P., Patel, B. & Patel, A. (2015), Big data analysis: recommendation system with hadoop framework, *in* 'Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on', IEEE, pp. 92–97.