PySpark

Mini Project

Swiggy Restaurant Rating Analysis.





Table of Contents

[1 Introduction 2](#_Toc96721826)

[1.1 Setup Checklist for Mini Project 3](#_Toc96721827)

[1.2 Instructions 3](#_Toc96721828)

[2 Problem Statement 4](#_Toc96721829)

[2.1 Objective 4](#_Toc96721830)

[2.2 Abstract of the project 4](#_Toc96721831)

[2.3 Technology used: 4](#_Toc96721832)

[3 Implementation 5](#_Toc96721833)

[3.1 Summary of the functionality to be built: 5](#_Toc96721834)

[3.2 Guidelines on the functionality to be built: 6](#_Toc96721835)

[3.3 3.3 Data Model: 11](#_Toc96721836)

[4 Reports to be built 12](#_Toc96721837)

[5 Appendix (source Files) 15](#_Toc96721838)

# **Introduction**

This document outlines a mini project. The project is to develop Swiggy Restaurants Reporting System.

This document contains the work flow of the system and gives guidelines on how to build the functionality gradually.

## Setup Checklist for Mini Project

* Hardware:
  + - Intel Pentium 90 or higher (P166 recommended)
    - Microsoft Windows 2010 or above.
    - Memory: 4GB of RAM (4GB or more recommended)
* Software:
  + - Anaconda
    - Python
    - IDE – Pycharm/Jupyter Notebook
    - Pyspark
    - Internet Explorer 10.0 or higher

NOTE: Anaconda will install Python. Other tools like Jupyter notebook, Spyder can be installed through Anaconda as well.

## 1.2 Instructions

* The code modules in the mini project should follow all the coding standards.
* Create a directory by your name in drive **<drive>**. In this directory, create a subdirectory **MiniProject**. Store your Project here.
* You can refer to your course material.
* You may also look up the help provided in the PYSPARK docs and documentation provided with respective tools.

# **Problem Statement**

## 2.1 Objective

Development of a **Swiggy Restaurant Rating Analysis.**

## 2.2 Abstract of the project

The culinary in India is as varied as its culture. On every 10 kilometers of distance we find the variety in food. With increasing technology we order food online. Amongst all the food applications Swiggy is one which is very widely used.

The purpose of this project is to analyze which restaurant is popular in a particular locality of a particular city. It collects the data on various attributes like City, Cuisines, rating in number, rating in text, Average cost of two person, availability of table and online booking. There are 13 prime cities of India, 226 restaurants with 28 different cuisines spread across 137 different localities of these cities.

## 2.3 Technology used:

* Anaconda
* Python
* IDE – Pycharm/Jupyter Notebook
* Pyspark

# **Implementation**

## Summary of the functionality to be built:

The participants need to develop the Swiggy Rating Analysis by building the functionality incrementally in each of the course modules of PYSPARK LOT using one of the ETL and reporting tool.

## Guidelines on the functionality to be built:

**Project flow**

Heterogeneous Sources i.e. operational data (Flat files)

EXTRACT

Load

Transform

In-memory

Dimension Model using PySpark

REPORTS

(CSV)

|  |  |
| --- | --- |
| **FINAL\_FACT** | |
| Fact\_ID | Number(p,s) |
| City\_ID | Number(p,s) |
| Locality\_ID | Number(p,s) |
| Rest\_ID | Number(p,s) |
| Cuisine\_ID | Number(p,s) |
| Average\_cost\_for\_two | Number(p,s) |
| Rating\_ID | Number(p,s) |
| Delivery\_ID | Number(p,s) |
| Table\_booking\_ID | Number(p,s) |
| Votes | Number(p,s) |
| Price\_range | Number(p,s) |

Schema Design:

The project follows the start schema approach.

* **The Dimension tables are** :
* City
* Locality
* Restaurant
* Cuisines
* Delivery
* Table booking
* Rating
* **Fact Table is** :
* Fact\_Swiggy

|  |  |
| --- | --- |
| **LOCALITY\_DIM** | |
| Locality\_ID | Number(p,s) |
| Locality\_name | Number(p,s) |

|  |  |
| --- | --- |
| **CITY\_DIM** | |
| City\_ID | Number(p,s) |
| City\_name | Number(p,s) |

|  |  |
| --- | --- |
| **RESTAURANT\_NAME\_DIM** | |
| Rest\_ID | Number(p,s) |
| Rest\_name | Number(p,s) |

|  |  |
| --- | --- |
| **CUISINE\_DIM** | |
| Cuisines\_ID | Number(p,s) |
| Cuisine\_name | Number(p,s) |

|  |  |
| --- | --- |
| **CUISINE\_DIM** | |
| Cuisines\_ID | Number(p,s) |
| Cuisine\_name | Number(p,s) |

|  |  |
| --- | --- |
| **TABLE\_BOOKING\_DIM** | |
| Table\_booking\_ID | Number(p,s) |
| Table\_booking\_avail | Varchar |

|  |  |
| --- | --- |
| **RATING\_DIM** | |
| Rating\_id | Number(p,s) |
| Rating\_in\_star | Number(p,s) |
| Ratin\_in\_text | Varchar |

* **City:**

This dimension contains information about the City on which the analysis will be carried out.

This CSV file comprises of the following Fields:

* City\_ID
* City\_Name

There are 13 Cities namely Bangalore, Chandigarh, Chennai, Goa, Guwahati, Hyderabad, Jaipur, Kolkata, Lucknow, Mumbai, Patna, Pune and Surat.

* **Locality:**

This dimension contains information about the locality on which the analysis will be carried out.

This CSV file comprises of the following Fields:

* Locality\_ID
* Locality\_Name

There are 137 localities in 13 cities.

* **Restaurant**

This dimension contains Restaurants with its name and ID:

* Restaurant\_ID
* Restaurant\_Name

There are 226 restaurants in 13 cities with various sub-branches and cuisines.

* **Cuisine**

This dimension contains Cuisines available in various restaurants with its name and ID :

* Cuisine\_ID
* Cuisine\_Name

There are 28 different cuisines in the dataset to carry out the analysis. The Cuisines are:

* American
* Asian
* Bakery
* Bengali
* Biryanis
* Café
* Continental
* Desserts
* European
* Fast Food
* Healthy Food
* Hyderabadi
* Italian
* Japanese
* Mediterranean
* Mexican
* Modern-Indian
* North Indian
* Rajasthani
* Sea Food
* South Indian
* Street Food
* **Table\_booking:**

This dimension contains information about the Table booking availability on which the analysis will be carried out.

* Table\_booking\_id
* Table\_booking\_avail
* **Delivery:**

This dimension contains information about the Delivery on which the analysis will be carried out.

* Delivery\_ID
* Delivery\_avail
* **Rating:**

This dimension contains information about the Rating on which the analysis will be carried out.

* Rating\_ID
* Rating\_in\_stars
* Rating\_in\_text

The Fact Table is:

* Fact\_Swiggy

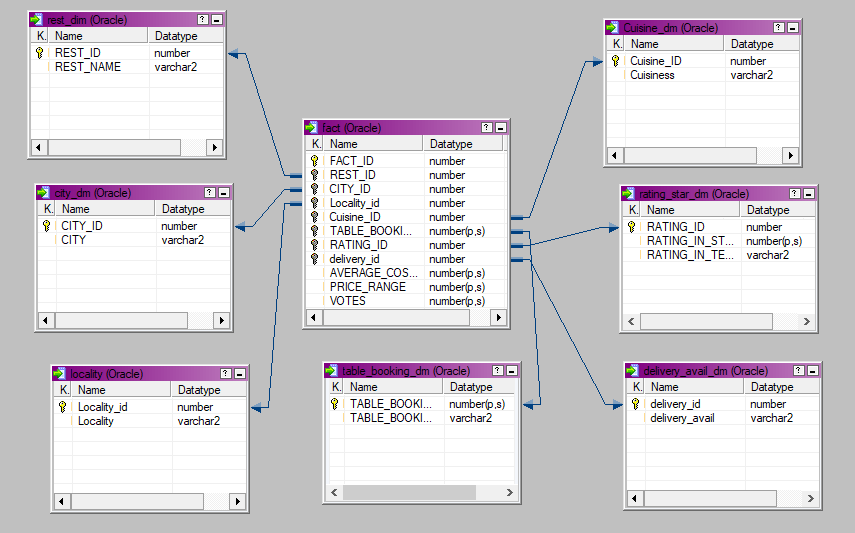
This is a Fact table that contains the IDs of all the cities, restaurants, locality and cuisine along with the cost for two people, ratings and availability of tables and online delivery

* Fact\_ID
* City\_ID
* Locality\_ID
* Rest\_ID
* Cuisine\_ID
* Average\_cost\_for\_two
* Rating\_ID
* Delivery\_ID
* Table\_booking\_ID
* Votes
* Price\_range

## 3.3 Data Model:

This is for reference purpose.

%3CmxGraphModel%3E%3Croot%3E%3CmxCell%20id%3D%220%22%2F%3E%3CmxCell%20id%3D%221%22%20parent%3D%220%22%2F%3E%3CmxCell%20id%3D%222%22%20value%3D%22CITY\_DIM%22%20style%3D%22swimlane%3BfontStyle%3D1%3BchildLayout%3DstackLayout%3Bhorizontal%3D1%3BstartSize%3D26%3BfillColor%3D%23C3ABD0%3BhorizontalStack%3D0%3BresizeParent%3D1%3BresizeParentMax%3D0%3BresizeLast%3D0%3Bcollapsible%3D1%3BmarginBottom%3D0%3BswimlaneFillColor%3D%23ffffff%3Balign%3Dcenter%3BfontSize%3D14%3BlabelBackgroundColor%3Dnone%3B%22%20vertex%3D%221%22%20parent%3D%221%22%3E%3CmxGeometry%20x%3D%22120%22%20y%3D%2246%22%20width%3D%22160%22%20height%3D%2278%22%20as%3D%22geometry%22%2F%3E%3C%2FmxCell%3E%3CmxCell%20id%3D%223%22%20value%3D%22CITY\_ID%22%20style%3D%22text%3BstrokeColor%3Dnone%3BfillColor%3Dnone%3BspacingLeft%3D4%3BspacingRight%3D4%3Boverflow%3Dhidden%3Brotatable%3D0%3Bpoints%3D%5B%5B0%2C0.5%5D%2C%5B1%2C0.5%5D%5D%3BportConstraint%3Deastwest%3BfontSize%3D12%3B%22%20vertex%3D%221%22%20parent%3D%222%22%3E%3CmxGeometry%20y%3D%2226%22%20width%3D%22160%22%20height%3D%2226%22%20as%3D%22geometry%22%2F%3E%3C%2FmxCell%3E%3CmxCell%20id%3D%224%22%20value%3D%22CITY\_NAME%22%20style%3D%22text%3BstrokeColor%3Dnone%3BfillColor%3Dnone%3BspacingLeft%3D4%3BspacingRight%3D4%3Boverflow%3Dhidden%3Brotatable%3D0%3Bpoints%3D%5B%5B0%2C0.5%5D%2C%5B1%2C0.5%5D%5D%3BportConstraint%3Deastwest%3BfontSize%3D12%3B%22%20vertex%3D%221%22%20parent%3D%222%22%3E%3CmxGeometry%20y%3D%2252%22%20width%3D%22160%22%20height%3D%2226%22%20as%3D%22geometry%22%2F%3E%3C%2FmxCell%3E%3C%2Froot%3E%3C%2FmxGraphModel%3E



**Note:** The datatype/length for the Dimension/fact table attributes can be changed as required.Additional fields can be added,if required.

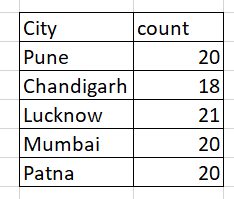
**Data Transformation for data warehouse:**

1. Load the Data into Dimension tables using the CSV files provided.
2. Load the Data into Fact tables using the CSV files provided.
3. Populate Dimension tables before fact table
4. Dimension table should have unique values with unique system generated IDs.. Same for Fact ID
5. Other ID columns in Fact table example in fact\_Swiggy table – rest\_id, city\_id etc. column should be populated with id columns values from corresponding dimension table

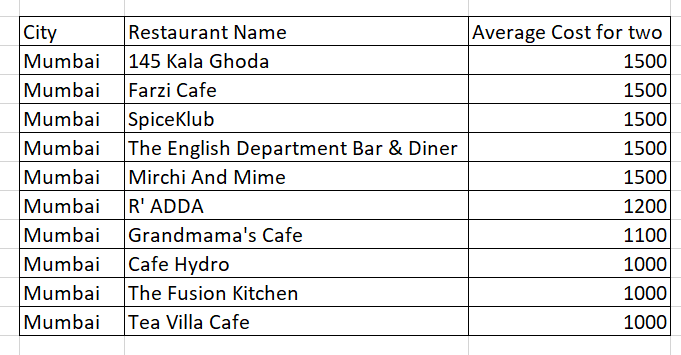
# **Reports to be built**

Following reports to be created. Create CSV file output for each report. Reports to be generated from the Fact and Dimension tables

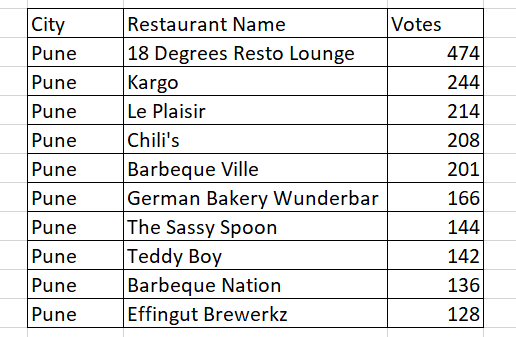
* Count of Restaurant city wise



* City-wise Top 10 restaurant based on Avg cost for 2



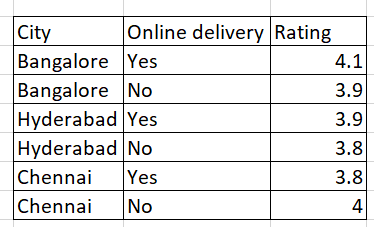
* Top 10 restaurant based on Avg Votes



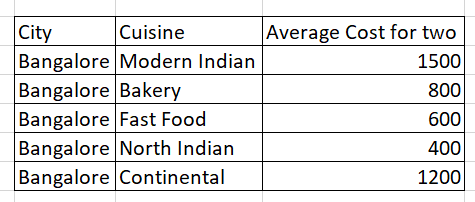
* Top 10 restaurant based on Rating City wise



* Rating based on delivery availability City wise



* Avg cost for 2 based on cuisine city wise



# 

# **Appendix (source Files)**

