**Swiggy Restaurant Dataset Analysis**

**Objective**

This capstone project focuses on building a comprehensive data pipeline for Data Engineering and Analytics. We'll primarily aim to use Azure Data Factory for data ingestion, then will switch to Databricks with PySpark and Spark-SQL for transformation and analytics, and for visualization trainees are independent to use any other options like Dashboard feature of Databricks or Power BI (optional). The architecture includes multiple data sources, a NoSQL landing zone, SQL-based materialized views, and a visualization layer. By the end of this project, we'll have a functional pipeline that provides valuable insights.

**Business Requirements**

1. List the names of tables used.
2. Capture the SQL queries/ PySpark code snippets used for data cleansing.
3. Capture the SQL queries/ PySpark code snippets used for data transformations.
4. Create a master table named t\_rating\_desc as follows:

Structure:

|  |  |
| --- | --- |
| **COLUMN** | **DATATYPE** |
| RATING | NUMBER |
| RATING\_DESC | STRING |

Data:

|  |  |
| --- | --- |
| **RATING** | **RATING\_DESC** |
| 0 | No Rating |
| 1 | Poor |
| 2 | Average |
| 3 | Good |
| 4 | Excellent |
| 5 | Must Try |

1. Write a query that retrieves Restaurant Name, Rating and the corresponding Rating Description. Use the next whole number to get rating description. For the restaurants with blank rating, display the one corresponding to 0.

Example: if the rating of a restaurant is 4.4, display the rating as Must Try.

1. Write a query to get Restaurant Name, City and Rating for Domino's Pizza so that the highest rated is displayed on top. If there are 2 outlets with the same rating, the first on the list should be based on the city in alphabetical order.

Example: If there are 2 Domino's Pizza outlets with rating 3, one in Adityapur and another in Alappuzha then the results should be as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Restaurant\_Name** | **Rating** | **Rating\_Description** | **City** |
| Domino's Pizza | 3 | Good | Adityapur |
| Domino's Pizza | 3 | Good | Alappuzha |
| Domino's Pizza | 2 | Average | Adityapur |

1. Find the number of restaurants with rating above 4 for cusine 'North Indian'.
2. Find the number of restaurants with a rating below the 3.
3. Find the number of people who have rated restaurants with cusine 'Fast Food,American'
4. Find the restaurant in LuluMall Kochi that has the highest number of people rated.
5. Identify the city with maximum number of restaurants.
6. Identify the number of Baskin Robbins outlets with a rating of 4.5 and above.
7. Identify the restaurants that sell least expensive Ice cream yet with a rating of 4 and above in Bangalore.
8. Find the top 10 cities as per the number of restaurants listed on Swiggy.
9. Identify the cuisine that has the top rating in this dataset
10. What is the total number of restaurants listed in swiggy in the city Kannur with a rating count of 100+ ratings?
11. Identify the restaurants that serve more than 1 cuisine. Display the Restaurant Name and the Cuisine

Example: Beverages,Pizzas

1. Write a query to list the restaurant names with the letter ‘d’ in the name and also the position of the letter d.
2. Write a query to display the restaurants that have two or more ‘o’ in the name.

Example: Fresh Food Café

1. Write a query to show the
2. restaurant name and cuisine as follows.

Example:

|  |  |
| --- | --- |
| **AB FOODS POINT** | **Beverages | Pizzas** |
| Janta Sweet House | Sweets | Bakery |
| theka coffee desi | Beverages |
| Singh Hut | Fast Food | Indian |

1. Write a query to display the restaurant\_name and total cost of restaurants that start with S. The total cost should be based on each city. The list should be sorted based on total cost (least total cost should be on top)
2. Create a new table with the structure:

|  |  |
| --- | --- |
| **RESTAURANT\_NAME** | **STRING** |
| LOCALITY | STRING |
| CITY | STRING |
| RATING | DOUBLE |
| RATING\_COUNT | STRING |

1. Load the data as follows:
2. Pull the records with city as Electronic City,Bangalore with a rating above 4 and rating count of at least 50+
3. Split the value into Locality and City using code (Do not hardcode the values).
4. Example:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RESTAURANT\_NAME** | **LOCALITY** | **CITY** | **Rating** | **rating\_count** |
| Grameen Kulfi | Electronic City | Bangalore | 4.6 | 50+ ratings |

1. Write a query to list the restaurant and city that serve a single cuisine.

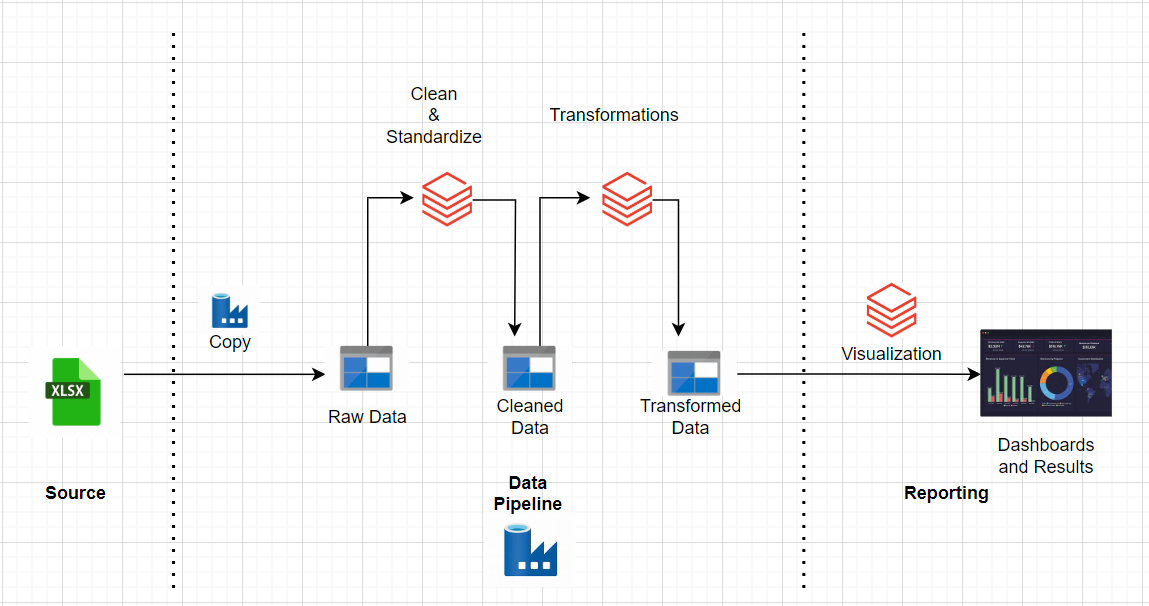
Example:

|  |  |  |
| --- | --- | --- |
| Shri Balaji Vaishno Dhaba | Abohar | North Indian |

1. Identify restaurants with anything other than alphabets in the restaurant name. Display distinct values in the output.
2. Analyze the entire dataset and detail one interesting finding.

**Architecture**

**Proposed System Architecture**



**Services**

**1. Data Ingestion**

* **Azure Data Factory (ADF)**:
  + Used to automate data ingestion from the source (e.g., CSV files) into the **Bronze Layer** in Azure Data Lake Storage.
  + Provides scheduling, monitoring, and orchestration of the data ingestion process.

**2. Data Storage**

* **Azure Data Lake Storage (ADLS)**:

**3. Data Transformation**

* **Azure Databricks**:
  + The main service for data cleaning, enrichment, and transformation.
  + Uses **PySpark** or **SQL** for processing data at scale.
  + Integrates with Delta Lake to handle schema validation, deduplication, and lineage tracking.

**4. Data Visualization**

* **Databricks SQL**:
  + A dedicated tool for querying the Gold Layer (Delta tables).
  + Enables building **SQL-based dashboards** and interactive visualizations directly within the Databricks workspace.
  + Supports visualization types such as bar charts, line graphs, and pie charts.

**5. Orchestration and Workflow Management**

* **Azure Data Factory (ADF)** :
  + Can orchestrate the entire pipeline, including data ingestion, triggering Databricks jobs, and monitoring data workflows.

**6. Monitoring and Logging**

* **Azure Monitor**:
  + Used to track the performance of Azure services (e.g., Databricks clusters, Data Lake) and resource utilization.