**Capstone Project 6. Comprehensive Data Pipeline with Azure Data Factory, Databricks and Dashboard on Restaurants Dataset | Swiggy**

# 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.

# Expectations

Go through the below case study and come up with a detailed design document with the steps needed to implement it, transformations for ensuring data quality and steps to validate the data after ingestion.

## Approach

1. **Pre-Migration Activities**:
   * **Assessment and Planning**: List out all data sources, assess data quality, and prioritize data sets for migration.
   * **Data Profiling**: Analyze data quality and identify any inconsistencies or redundancies.
   * **Schema Design**: Design the schema for Data Lake tables based on existing data sources.
2. **Input Data Sources (Multiple):**
   * Data sources such as CSV files, JSON, APIs, and relational databases.
3. **Data Extraction**: Use Azure Data Factory to ingest data into cloud.
4. **Data Transformation**: Use PySpark in Databricks for data cleansing, transformation, and enrichment. Ensure data quality in the target after data migration (check for junk data, special characters, duplicates, nulls etc).
5. **Landing Zone:**
6. (NoSQL/ Azure Delta Lake Tables)
7. **Data Ingestion**:
   * Load data into Azure Data Lake tables.
8. **Testing and Validation**:
   * Validate data integrity post-loading to ensure accuracy. Capture all the activities done with screenshots.
9. **SQL:**
   * Write queries for the given requirements. Please see section **SQL Requirements**

## Dataset

Link: <https://www.kaggle.com/datasets/ashishjangra27/swiggy-restaurants-dataset>

🆔 id: Unique identifier for each restaurant.

🍽️ name: Name of the restaurant.

🌆 city: City where the restaurant is located.

⭐ rating: Rating of the restaurant.

👥 rating\_count: Number of people who have given the rating.

💵 cost: Cost of eating at the restaurant.

🍲 cuisine: Types of cuisines the restaurant serves.

📝 lic\_no: License number of the restaurant.

🔗 link: Restaurant's link on the Swiggy website.

📍 address: Full address of the restaurant.

## Environment Setup - Azure

First, we'll set up our Azure environment. This involves creating accounts and setting up services like Azure Data Factory, Databricks, Cosmos DB, and Azure SQL Database.

* **Azure Account**: If you don't already have one, you'll need to create an Azure account/ check with your mentor.
* **Resource Group**: Organize your resources by creating a new resource group in Azure. (Use project/dataset name for better understanding)
* **Azure Data Factory**: Set up Azure Data Factory to create our data workflows and data pipeline
* **Databricks**: Set up a Databricks workspace for data processing and analytics.

# SQL 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 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 |

Load the data as follows:

Pull the records with city as Electronic City,Bangalore with a rating above 4 and rating count of at least 50+

Split the value into Locality and City using code (Do not hardcode the values).

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.