Azure Case Study - Swiggy

Comprehensive Data Pipeline with Azure Data Factory, Databricks and Dashboard on Restaurants Dataset

January 7 2025
Vismaya B
Developer 1 – Software Engineering

Objective / Goal

- **Objective:**

Build a comprehensive, scalable data pipeline using Azure services to process and analyze the Swiggy restaurants dataset.

- **Goal:**

Transform raw data into actionable insights through structured data processing (Bronze ⇒ Silver ⇒ Gold), with business-ready visualizations.

Approach

1.Data Ingestion

- **Source**: Raw JSON data from HTTP source.
- **Tool Used:** Azure Data Factory (ADF)
- **Copy Data Activity**: This activity is used to copy raw data from the HTTP source (JSON format) to Azure Data Lake Storage Gen2 (ADLS2).

2. Data Transformation

- **Tool Used:** Databricks
- Transforms raw Bronze Layer data into structured Delta Tables (Silver Layer).
- Cleans data by addressing inconsistencies and missing values.
- Filters the latest records from the Silver Layer and saves them into the Gold Layer for analysis and visualization.

Approach

3. Analysis & Visualization

- **Tools Used:** Databricks
- Queries the Gold Layer Delta Table for analysis using Spark SQL.
- Performs SQL-based analytics to derive meaningful insights.
- Generates visualizations in Databricks to represent the analyzed data effectively.

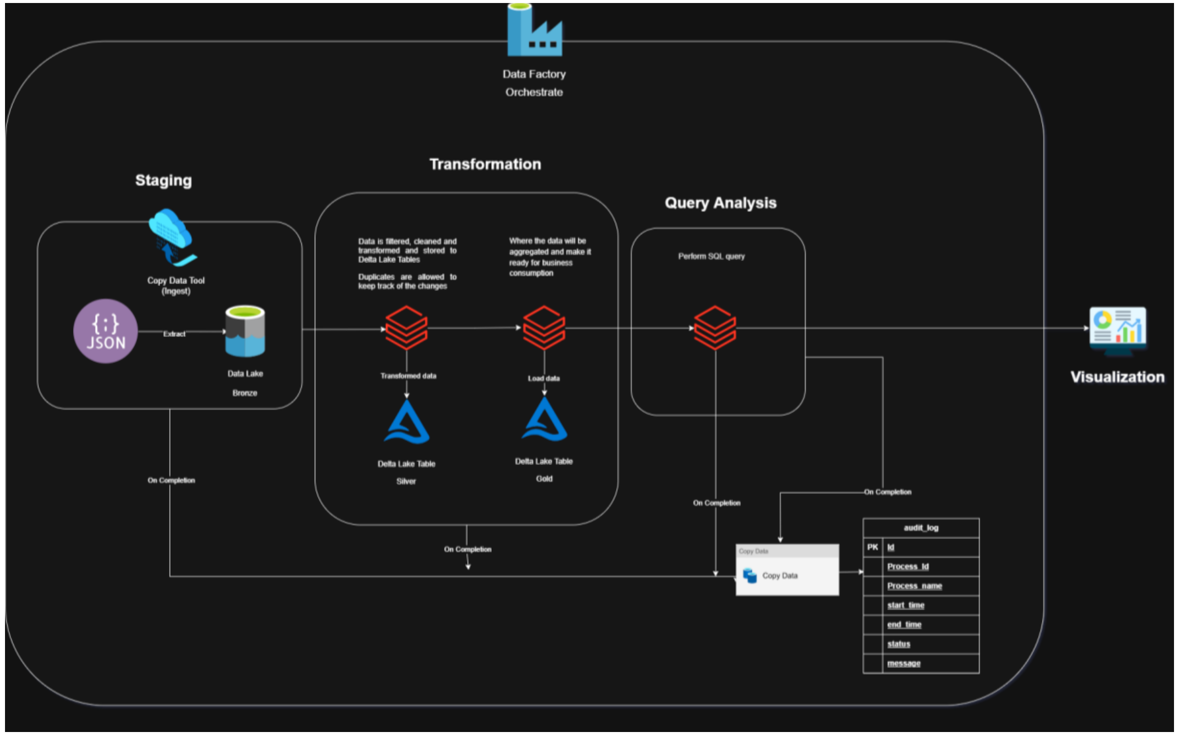
4. Security and Logging

- Used Azure Key Vault with Databricks Secret Scope for secure credential management.
- Pipeline details (e.g., name, status, timestamps) logged into SQL audit tables via ADF Copy Activity.

Approach

1. Architecture:

1. **Medallion Architecture:** Incrementally improve data quality and structure across Bronze, Silver, and Gold layers.



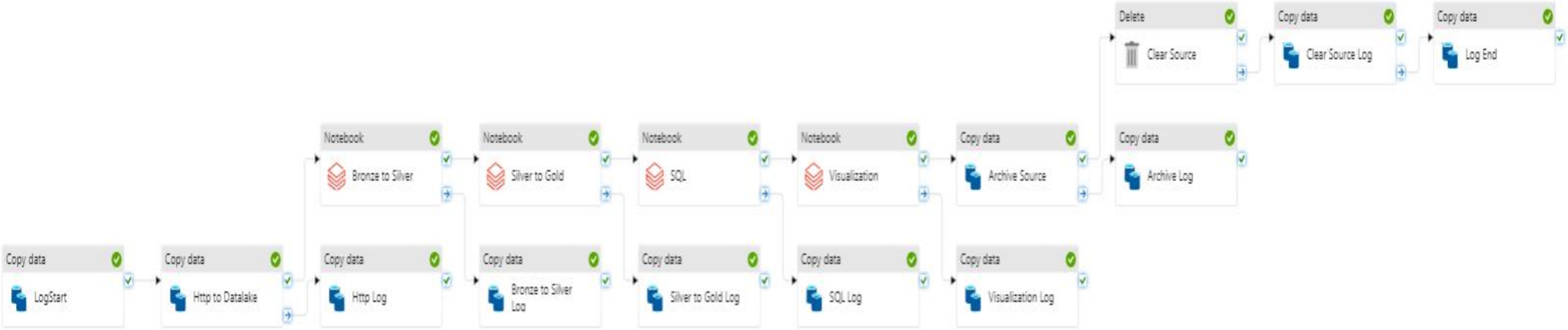
Data is filtered, cleaned and transformed and stored to Delta Lake Tables

Duplicates are allowed to keep track of the changes

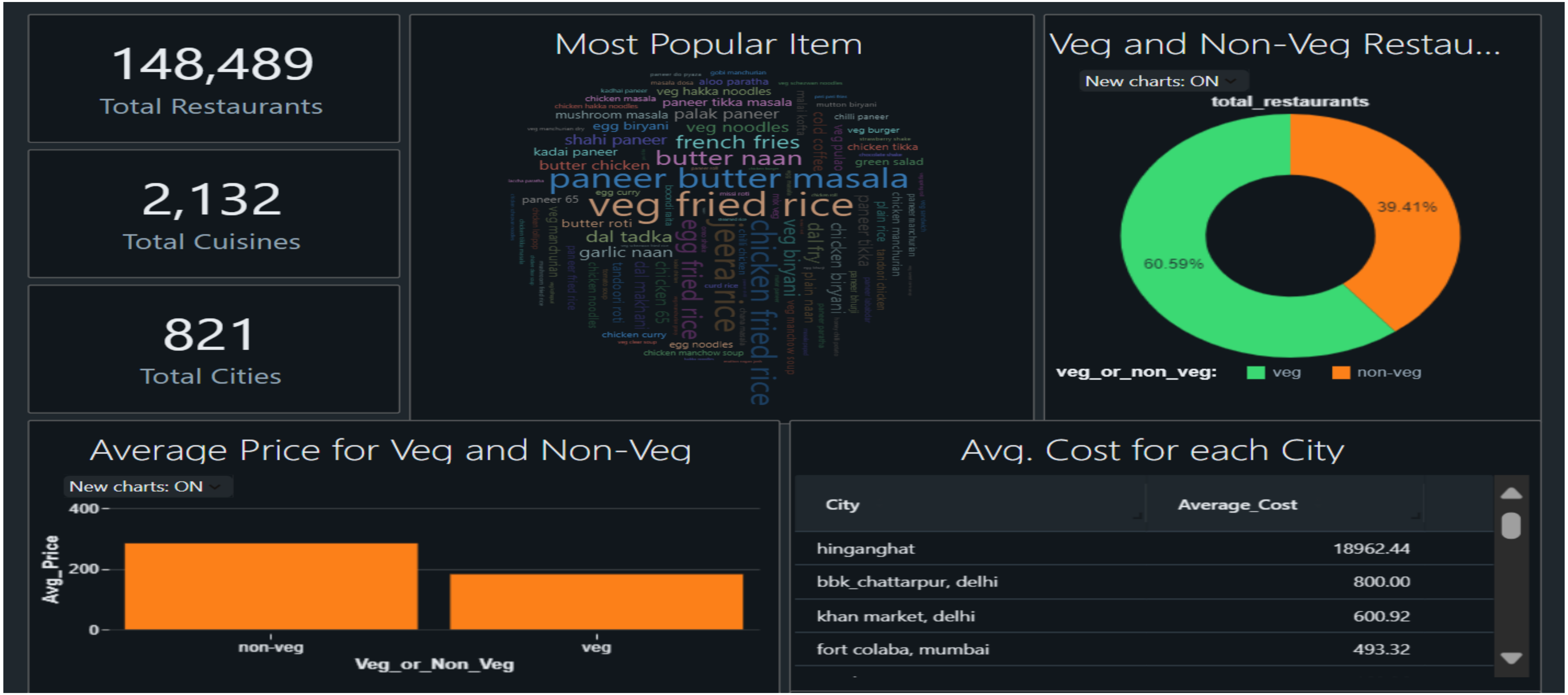
Where the data will be aggregated and make it ready for business consumption

audit_log	
PK	Id
	Process_Id
	Process_name
	start_time
	end_time
	status
	message

ADF pipeline



Output / Visualization



148,489
Total Restaurants

2,132
Total Cuisines

821
Total Cities

Most Popular Item

Veg and Non-Veg Restau...

New charts: ON

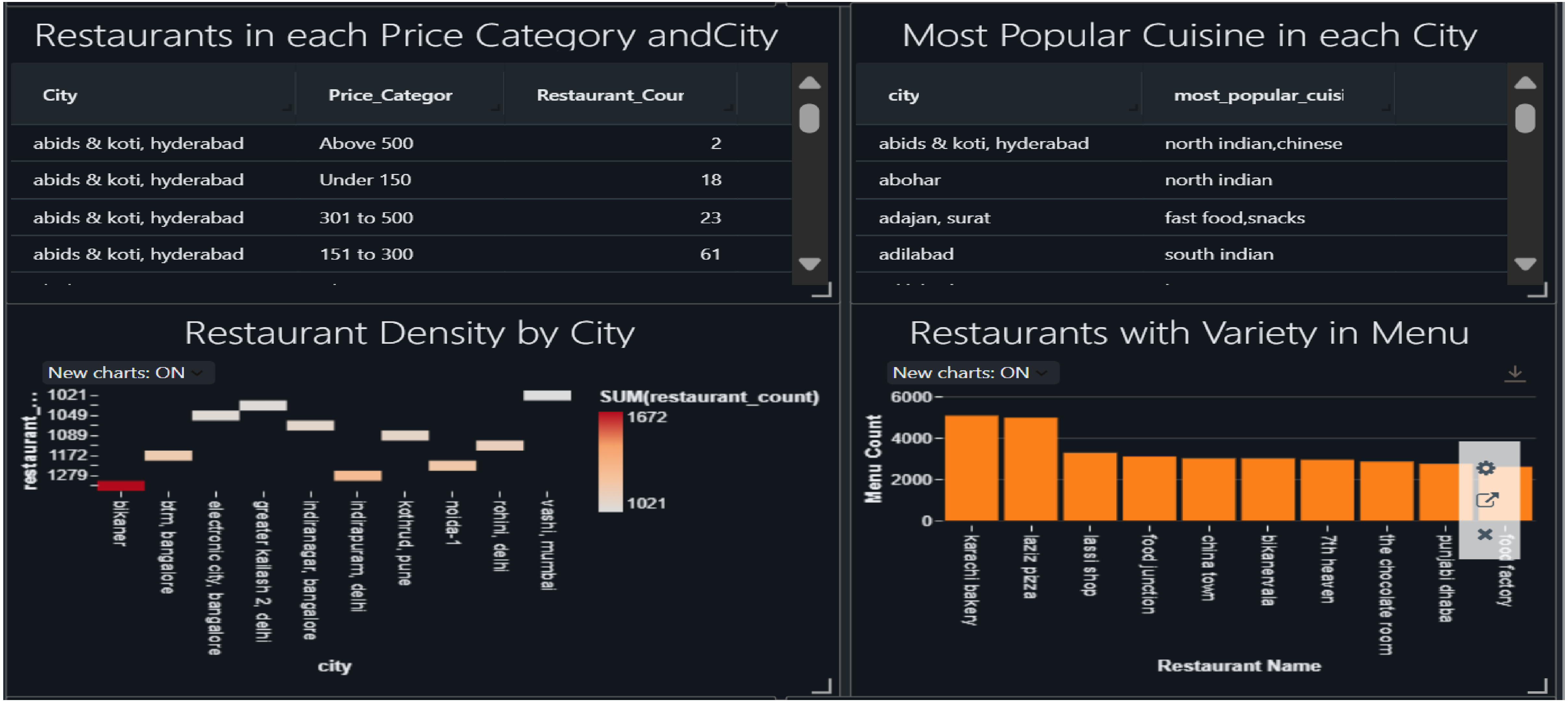
Average Price for Veg and Non-Veg

New charts: ON

Avg. Cost for each City

City	Average_Cost
hinganghat	18962.44
bbk_chattarpur, delhi	800.00
khan market, delhi	600.92
fort colaba, mumbai	493.32

Output / Visualization



Restaurants in each Price Category andCity

City	Price_Categor	Restaurant_Cour
abids & koti, hyderabad	Above 500	2
abids & koti, hyderabad	Under 150	18
abids & koti, hyderabad	301 to 500	23
abids & koti, hyderabad	151 to 300	61

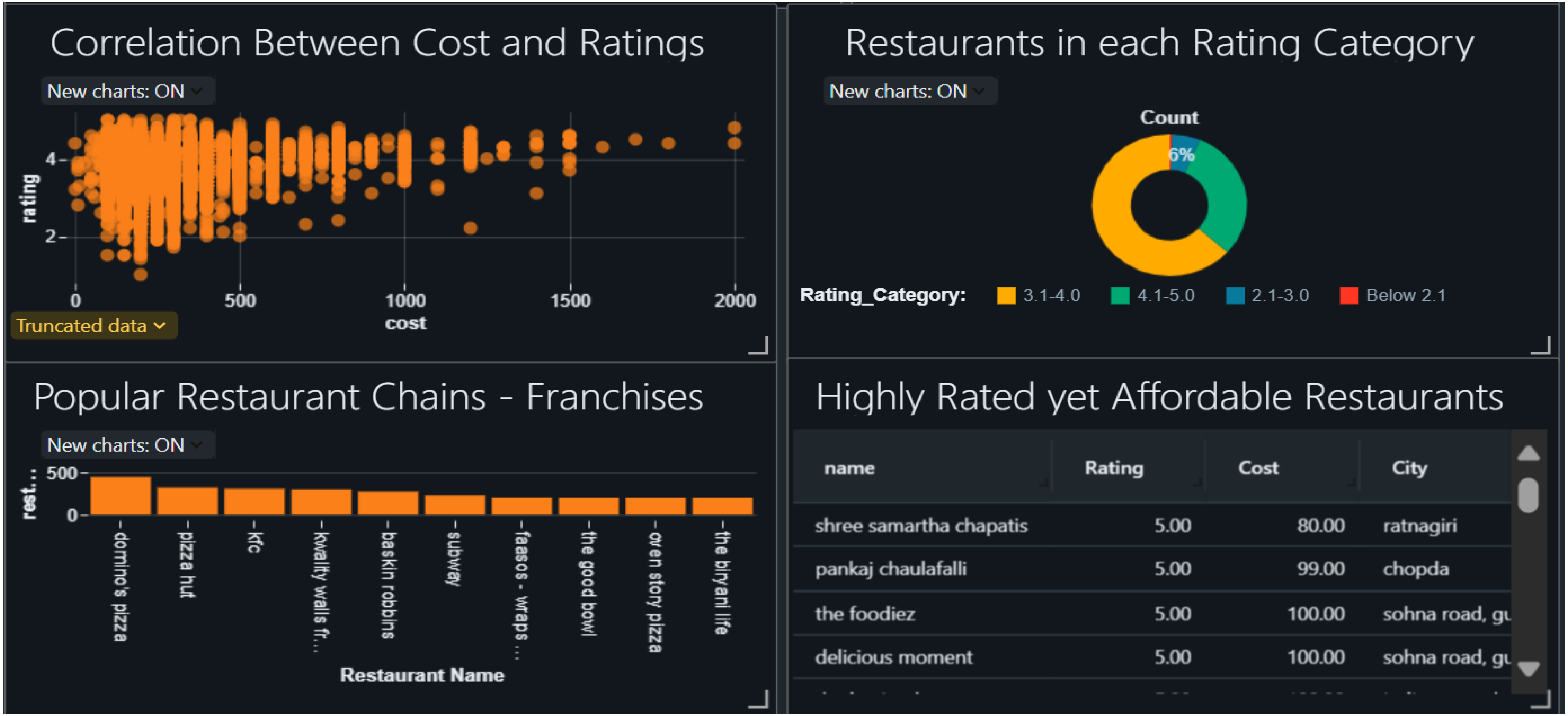
Most Popular Cuisine in each City

city	most_popular_cuisi
abids & koti, hyderabad	north indian,chinese
abohar	north indian
adajan, surat	fast food,snacks
adilabad	south indian

Restaurant Density by City

Restaurants with Variety in Menu

Output / Visualization



name	Rating	Cost	City
shree samartha chapatis	5.00	80.00	ratnagiri
pankaj chaulafalli	5.00	99.00	chopda
the foodiez	5.00	100.00	sohna road, gu
delicious moment	5.00	100.00	sohna road, gu

Insights

1. Popular Food Preferences

•**Insight:** Indian comfort food is highly favored. Promotions featuring these items can attract more customers. Partnering with restaurants to create combo offers for these dishes would likely increase order volumes.

2. Vegetarian vs Non-Vegetarian Restaurants

•**Insight:** Vegetarian restaurants dominate the market. This suggests a strong demand for vegetarian options. Non-veg restaurants could expand their menu to include popular vegetarian dishes to capture a wider audience.

3. Restaurant Density by City

•**Insight:** In high-density cities, competition is intense.
Swiggy can focus on optimizing delivery times and introducing exclusive offers for popular restaurants to differentiate itself.

Insights

4. Most Popular Cuisines by City

- **Insight:** Cuisine preferences vary significantly by city. This insight can guide city-specific promotions or recommendations.

5. Restaurants with Extensive Menus

- **Insight:** Promote restaurants with larger menus in the app as they appeal to customers seeking diverse options.

6. Cost and Rating Correlation

- **Insight:** Affordable restaurants with high ratings are valuable for attracting budget-conscious customers. Highlight these as "Top Budget Picks" in the app.

Insights

7. Rating Distribution

•**Majority of Restaurants:** Fall into the 3.1- 4.0 rating.
•**Highly Rated Restaurants:** Less common but can be promoted as “Premium Picks.”

•**Insight:** Encourage restaurants with lower ratings to improve through loyalty programs and customer feedback.

8. Popular Restaurant Chains

•**Top Chains:** Domino’s, KFC, Pizza Hut.

•**Insight:** Well-established franchises dominate the market. Strengthen partnerships with these brands to secure exclusive discounts or priority listings in the app.

11. Highly Rated Yet Affordable Restaurants

•**Examples:** Shree Samartha Chapatis (5.0 rating, ₹80).
•Pankaj Chaufalalli (5.0 rating, ₹99).

•**Insight:** Feature these hidden gems in-app campaigns to attract quality-conscious but budget-sensitive customers.

Challenges Faced

- **1. Fetching Data from Nested JSON**
- Faced challenges due to inconsistencies in the nested JSON structure.
- Resolved by analyzing the schema and using PySpark to break down and process the JSON data effectively.
- **2. Managing Credentials Securely**
- Ensured secure storage of credentials without exposing them in pipelines or notebooks.
- Utilized Azure Key Vault for secure credential management.
- Integrated with Databricks Secret Scope for seamless and protected access.
- **3. Data Logging**
- Implemented an effective method for monitoring pipeline activity.
- Evaluated two approaches:
  - Using Copy Activity to log details into a SQL audit table.
  - Integrating Azure Data Factory with Log Analytics Workspace for centralized monitoring and logging.
- **4. Data Quality**
- Ensured consistent and accurate results by handling:
  - Duplicates
  - Missing values

Learnings

Robust Logging: Learned the importance of detailed logging for traceability and quick debugging of pipeline issues.

Medallion Architecture Benefits: Realized the advantages of modular pipelines (Bronze, Silver, Gold layers) for progressive data enrichment.

Azure Log Analytics: Learned how to use Azure Log Analytics to capture logs and errors during data ingestion and transformation, aiding in debugging and improving pipeline performance.

Schema Understanding and Transformation: Gained valuable experience in handling complex nested JSON schemas and converting them into usable formats like CSV.

Thank you