Capstone Project

by Elias Joy

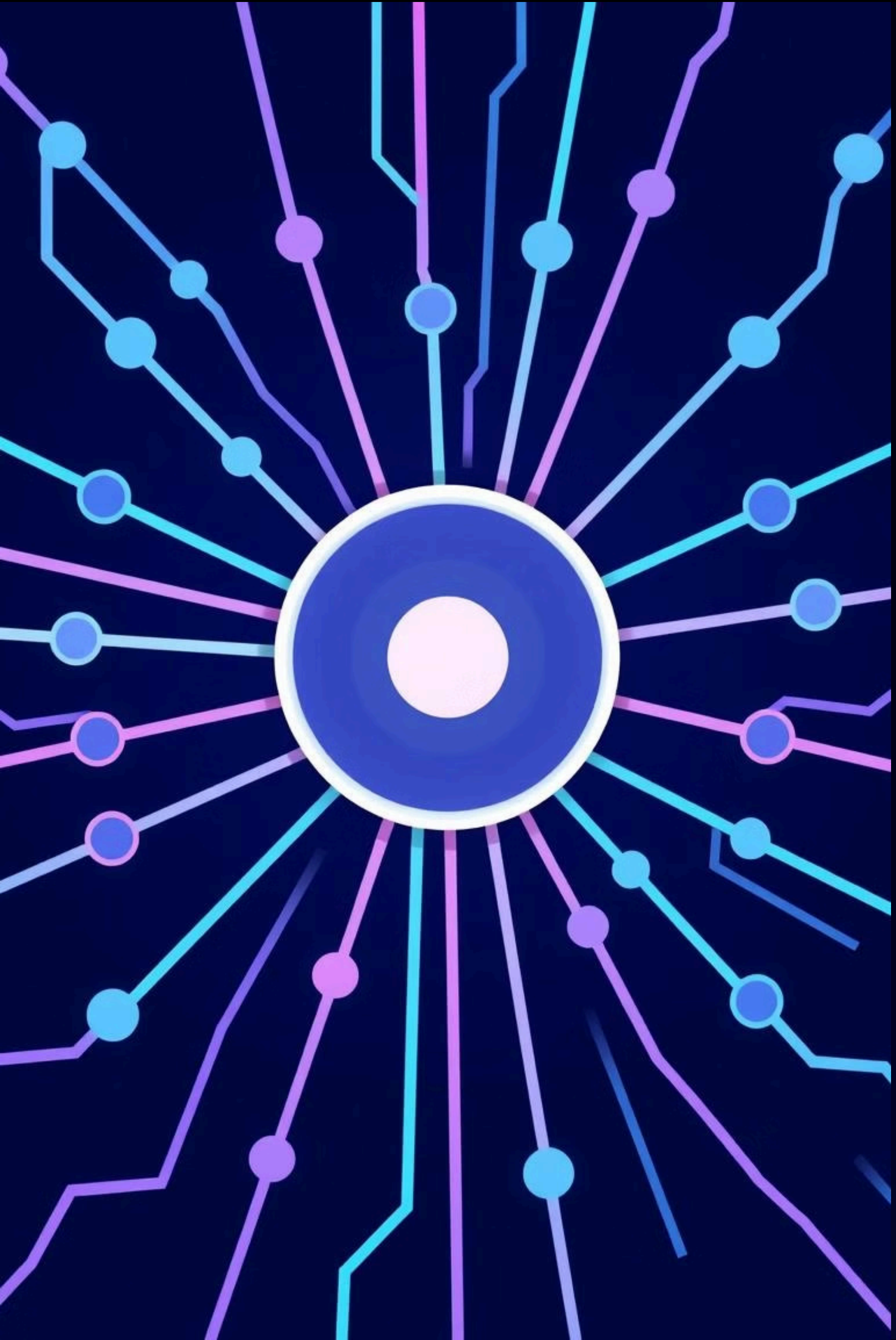
Problem Overview

Comprehensive Data Pipeline with Azure Data Factory, Databricks and Dashboard for Insightful analysis on Uber/Lyft Cab Prices and Weather Impact on Surcharge

The problem statement of this project is to design and implement a scalable data pipeline using Azure Data Factory and Databricks to ingest, transform, and analyze Uber/Lyft cab pricing data and weather conditions.

The aim is to develop a reliable data model and visualization layer that provide actionable insights into the impact of weather on cab surcharges.

This will assist stakeholders in identifying patterns, forecasting demand, and optimizing pricing strategies under varying conditions, thereby enhancing the overall decision-making process for pricing and operational efficiency.



Input Data Sources

1 **Weather Data**

Pulled from a CSV file hosted in a Git repository. Containing detailed weather conditions for Boston during the period of cab data collection.

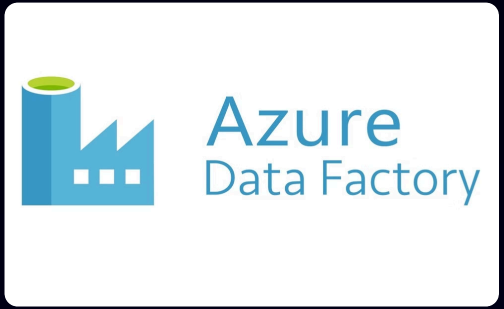
2 **Uber Data**

Stored in Azure Blob Storage as CSV file, containing trip details and pricing information of Uber Cab Rides.

3 **Lyft Cab Data**

Sourced from an Azure SQL Database, containing trip details and pricing information of Lyft Cab Rides

Tech Stacks Used



Data Ingestion

Azure Data Factory orchestrates the data ingestion process, pulling data from various sources.

Data Storage

Data is stored in various locations, including **Azure Blob Storage, Azure SQL Database**, and **Cosmos DB.**

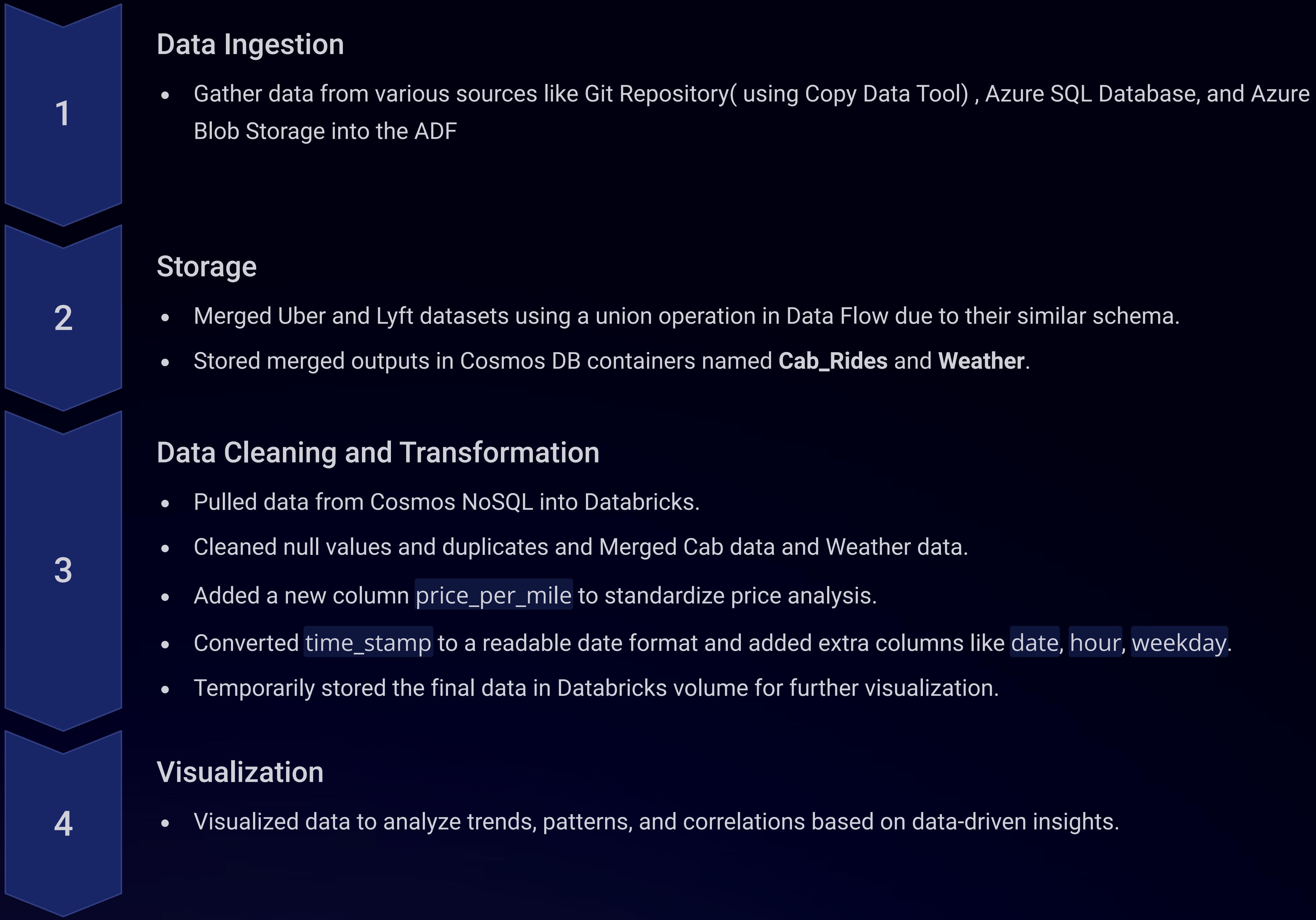
Data Processing & Transformation

Azure **Databricks** performs data cleansing, transformation, and feature engineering using **PySpark** and **Spark-SQL**.

Data Visualization

Databricks dashboard presents visualization of key insights from the processed data

Solution Flow Diagram



1

Data Ingestion

- Gather data from various sources like Git Repository(using Copy Data Tool) , Azure SQL Database, and Azure Blob Storage into the ADF

2

Storage

- Merged Uber and Lyft datasets using a union operation in Data Flow due to their similar schema.
- Stored merged outputs in Cosmos DB containers named **Cab_Rides** and **Weather**.

3

Data Cleaning and Transformation

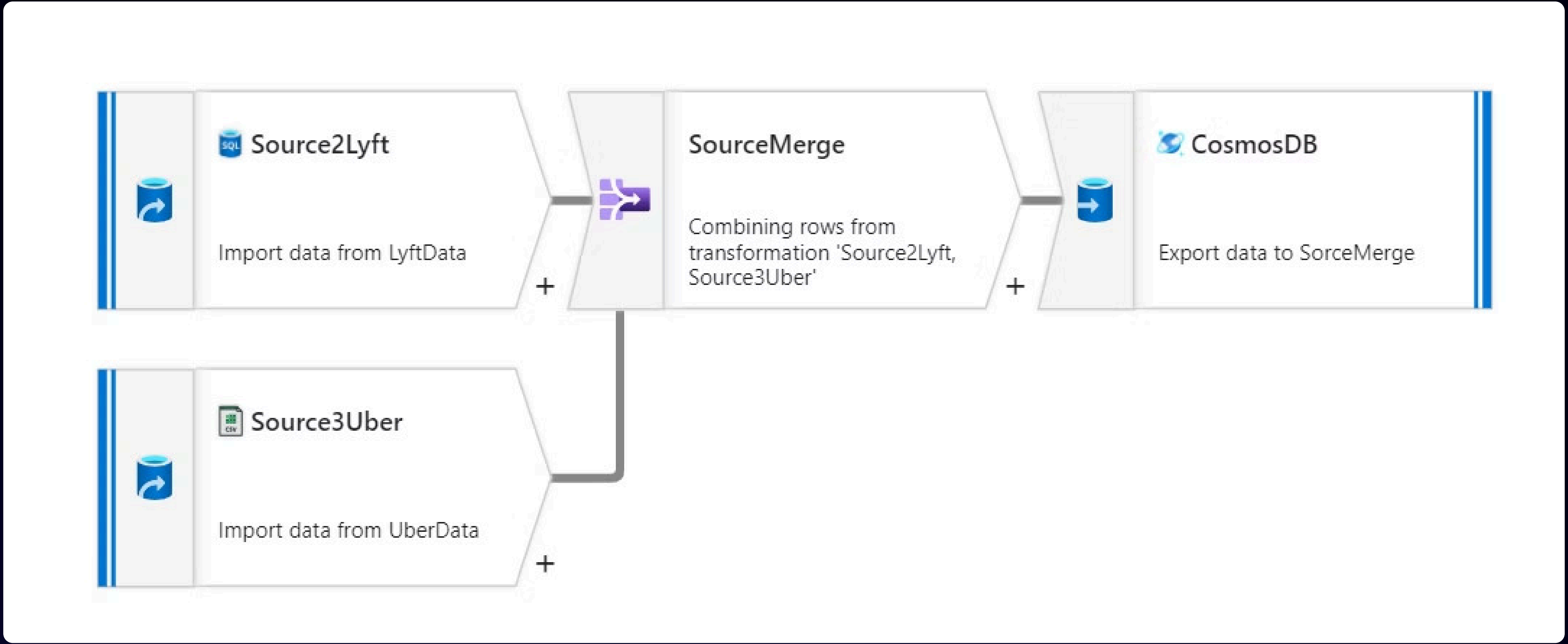
- Pulled data from Cosmos NoSQL into Databricks.
- Cleaned null values and duplicates and Merged Cab data and Weather data.
- Added a new column `price_per_mile` to standardize price analysis.
- Converted `time_stamp` to a readable date format and added extra columns like `date`, `hour`, `weekday`.
- Temporarily stored the final data in Databricks volume for further visualization.

4

Visualization

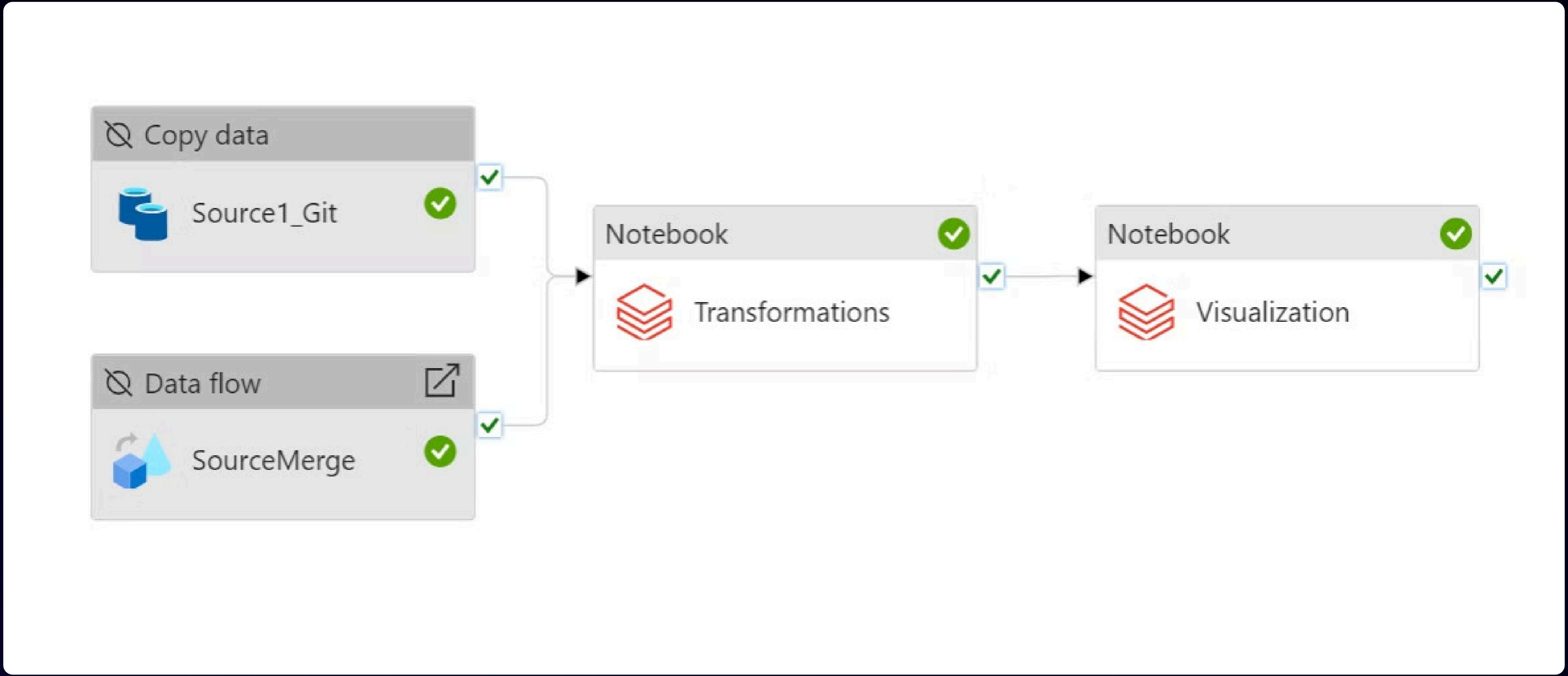
- Visualized data to analyze trends, patterns, and correlations based on data-driven insights.

Data Flow : Merge Similar Sources



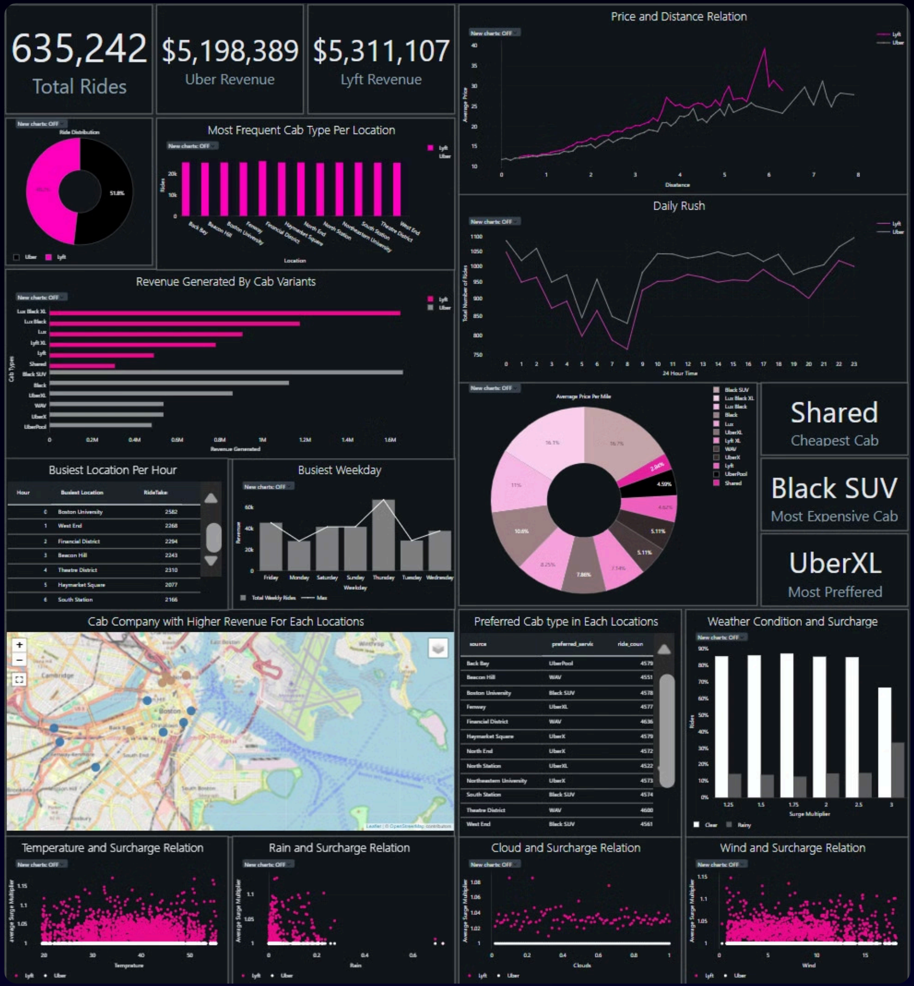
Source2Lyft
Import data from LyftData
SourceMerge
Combining rows from transformation 'Source2Lyft, Source3Uber'
CosmosDB
Export data to SorceMerge
Source3Uber
Import data from UberData

ADF Pipeline



Dashboard

https://adb-3249565684546148.8.azuredatabricks.net/?o=3249565684546148#notebook/566598230449711



635,242
Total Rides

$5,198,389
Uber Revenue

$5,311,107
Lyft Revenue

Shared
Cheapest Cab

Black SUV
Most Expensive Cab

UberXL
Most Preffered

Busiest Location Per Hour

Hour	Busiest Location	RideTake
0	Boston University	2582
1	West End	2268
2	Financial District	2294
3	Beacon Hill	2243
4	Theatre District	2310
5	Haymarket Square	2077
6	South Station	2166

Preferred Cab type in Each Locations

source	preferred_servic	ride_coun
Back Bay	UberPool	4579
Beacon Hill	WAV	4551
Boston University	Black SUV	4578
Fenway	UberXL	4577
Financial District	WAV	4636
Haymarket Square	UberX	4579
North End	UberX	4572
North Station	UberXL	4522
Northeastern University	UberX	4573
South Station	Black SUV	4574
Theatre District	WAV	4600
West End	Black SUV	4561

Most Frequent Cab Type Per Location
Revenue Generated By Cab Variants
Price and Distance Relation
Daily Rush
Busiest Weekday
Cab Company with Higher Revenue For Each Locations
Weather Condition and Surcharge
Temperature and Surcharge Relation
Rain and Surcharge Relation
Cloud and Surcharge Relation
Wind and Surcharge Relation

Solution Benefits

Business Impact

Enables ride-sharing companies to understand how weather affects demand, optimize pricing models, and better serve customers.

Data Insights

Clear correlations can be derived between weather conditions (e.g., rain, snow) and surcharges, helping to predict peak demand periods.

Enhanced Business Insights

Gain valuable insights into demand patterns, identify ride trends, and optimize resource allocation.

Key Insights:

- **Customer Preference**: Uber was the most preferred service, despite Lyft generating higher revenue.
- **Revenue Analysis**: Lyft's revenue advantage was due to its surge multiplier, while Uber maintained a constant surge multiplier of 1.
- **Surge Multiplier Insights**: The surge multiplier for Lyft was not influenced by weather conditions. Instead, it was determined by specific Lyft variants operating in particular locations.
- **Day-wise Trends**: Thursday emerged as the busiest weekday for both services.
- **Service Comparisons**:
  - **Most Popular Service**: UberXL had the highest usage.
  - **Cheapest Option**: Shared Lyft was the least expensive.
  - **Most Expensive Option**: Black SUV was the most premium offering.

Challenges Faced

Data Accuracy	Integrating data from multiple disparate sources into a single unified pipeline
Handling merge of two different datasets	Handling and cleaning null values and duplicates in merged datasets.
Data transformations	Data formatting and type conversion issues during transformation (e.g., `time_stamp` column) due to multiple format mismatches
System Implementation and Maintenance	Implement and maintain the system effectively to ensure seamless operation.
Data inconsistency	There existed many data that were inconsistent throughout the dataset, which were cleaned out before analysis.