ANALYSIS ON UBER/LYFT CAB PRICES AND WEATHER IMPACT ON SURCHARGE

CAPSTONE PROJECT

PROBLEM OVERVIEW

•**Context:**

In the dynamic ride-hailing market, surge pricing is a critical element that impacts customer satisfaction and driver availability. Surge prices are influenced by factors like demand, location, and weather conditions.

•**Goal:**

The objective is to create a comprehensive data pipeline that analyzes Uber/Lyft cab prices and explores how weather impacts surcharges. By implementing a robust system, businesses can gain valuable insights into price fluctuations and external factors affecting the ride-hailing ecosystem.

Details of Input Data

- **Sources of Data:**

  - **Uber Dataset:** Stored in Azure Blob Storage, containing cab ride data for various Uber cab types and their prices for specific locations.

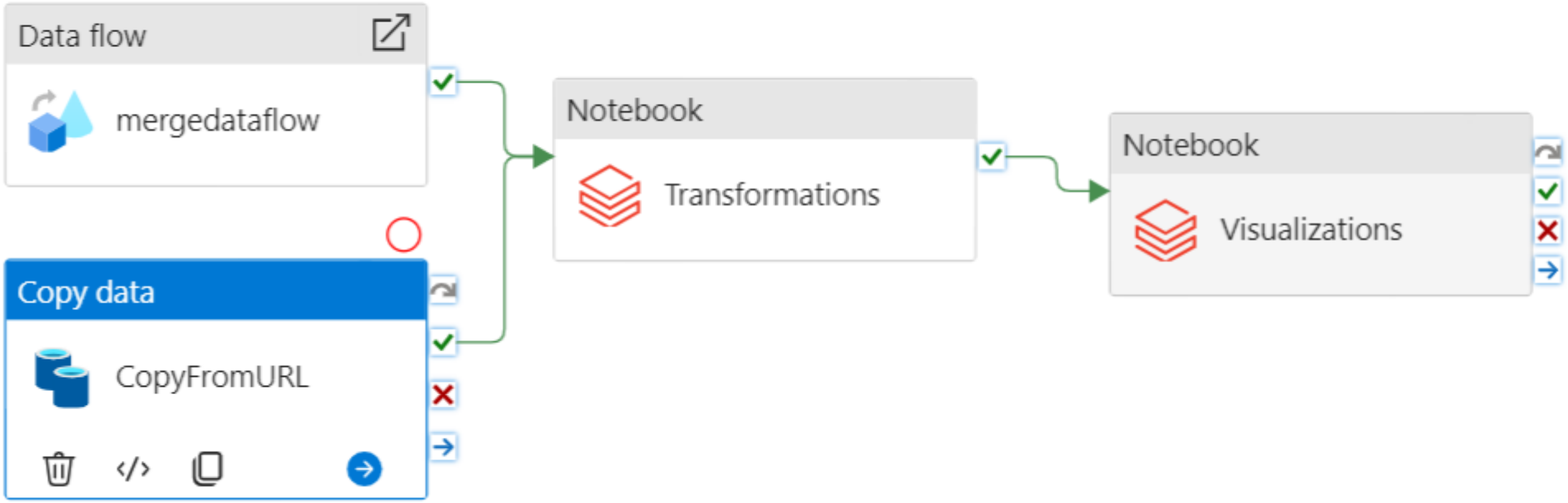
  - **Lyft Dataset:** Stored in an Azure SQL database, covering various Lyft cab types and prices for specific locations.

  - **Weather Dataset:** Available in HTTP format, containing weather attributes such as temperature, rain, and cloud coverage for all locations in the dataset.

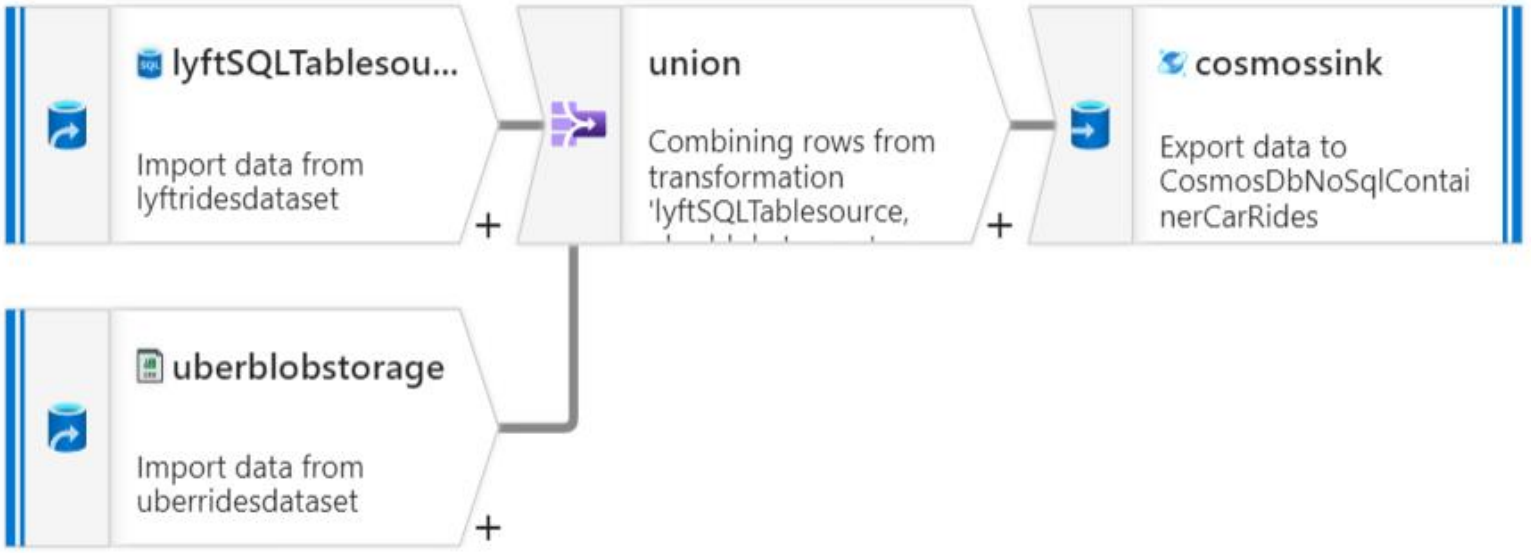
Tech Stack:

- **Data Ingestion:** Azure Data Factory
- **Landing Zone:** Azure Cosmos DB
- **Data Processing & Transformation:** PySpark in Azure Databricks
- **Materialized Views:** Azure SQL Database
- **Visualization Layer:** Databricks Dashboard

Solution Flow Diagram - ADF Pipeline



Solution Flow diagram



Steps Overview:

Azure Environment Setup: Successfully configured Azure Data Factory (ADF), Databricks, Cosmos DB, and SQL Database.

Data Ingestion: Established data pipelines using ADF to ingest data from Blob Storage (Uber data), Azure SQL Database (Lyft data), and HTTP (weather data).

Data Transformation: Performed data cleaning and enrichment in Databricks with PySpark.

Materialized Views: Created materialized views for optimized querying.

Data Visualization: Developed interactive dashboards in Databricks for insightful analysis.

Solution Benefits

- **Revenue Insights**: Although Uber had a higher number of rides, Lyft generated more revenue due to surge pricing.

- **Surge-Weather Analysis:** There was no significant correlation between weather conditions and surge pricing.

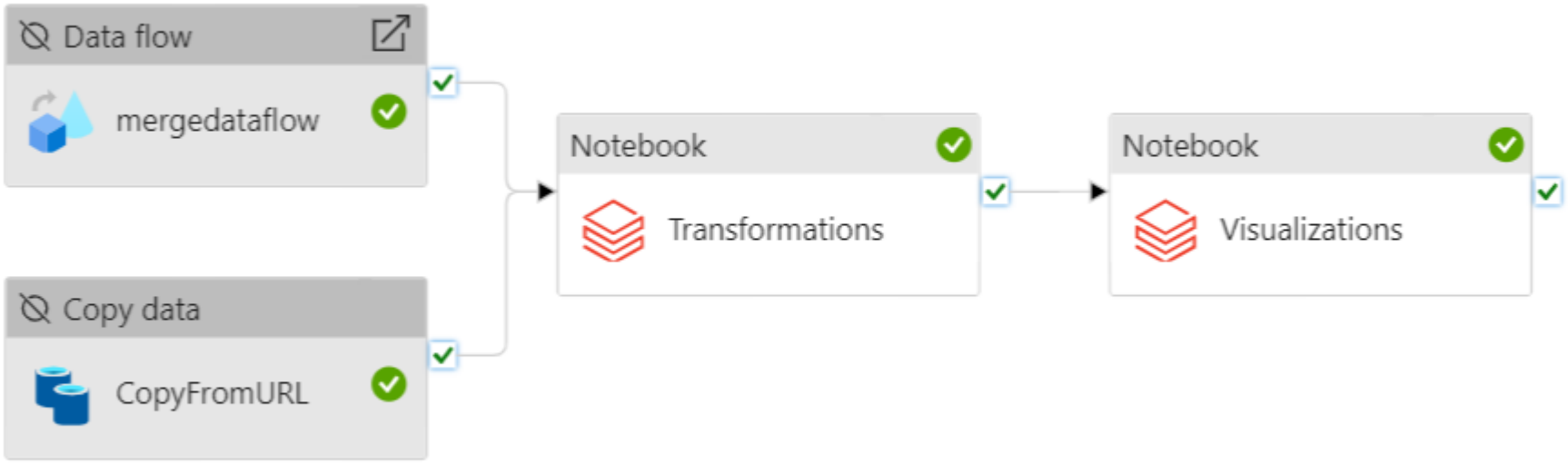
- **Enhanced Decision Making:** The insights help stakeholders better understand pricing trends and improve pricing strategies.

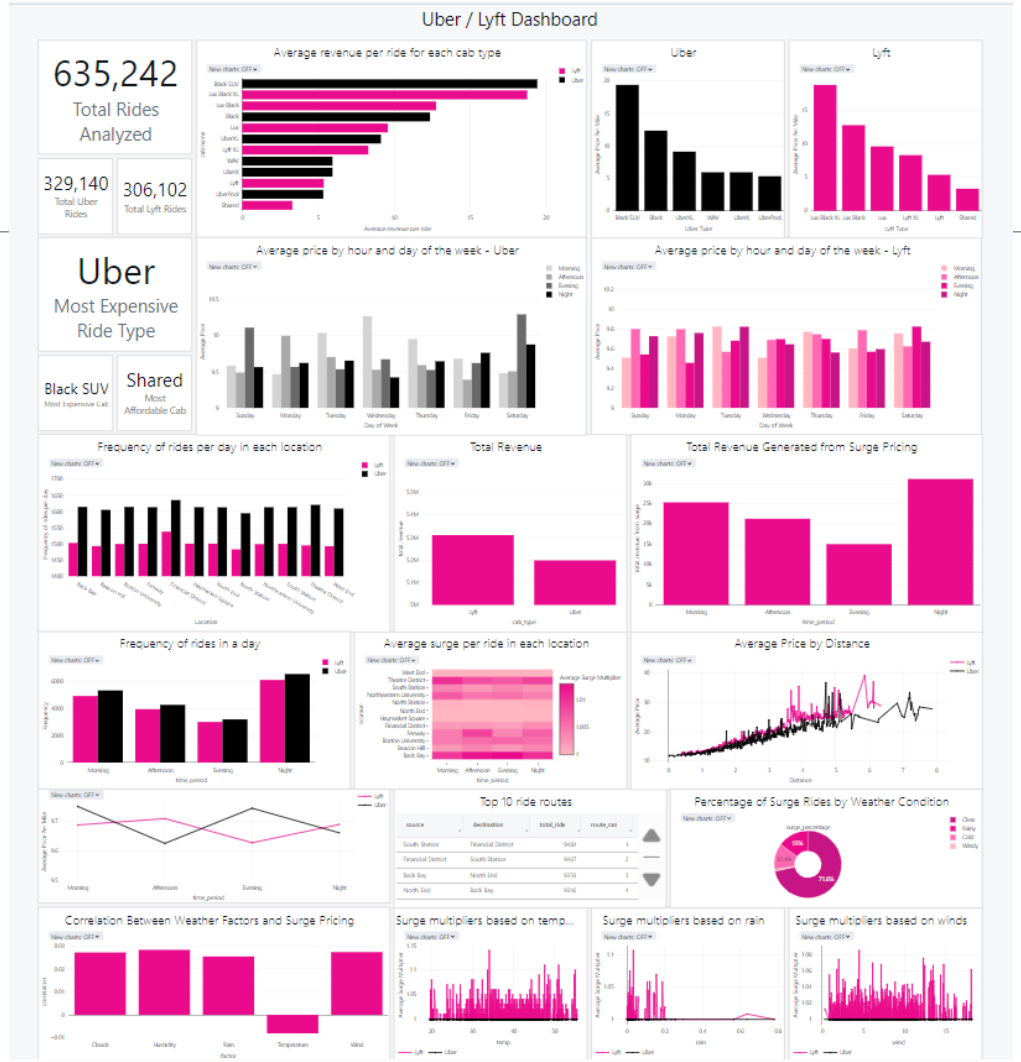
- **Operational Efficiency:** Automated data pipelines using ADF and Databricks streamlined data ingestion and transformation.

Testing

- **Test Data:** Used the Kaggle dataset alongside live data collected from APIs for testing.
- **Testing Steps:**
  - Verified the dashboards against SQL query results for accuracy.
  - Validated data ingestion pipelines using small batches of data.

Results


Data flow
mergedataflow
Copy data
CopyFromURL
Notebook
Transformations
Notebook
Visualizations



Results - Dashboard

Challenges faced

- **Pipeline Delays:** Pipelines took a long time to run.

- **Data Consistency:** Misaligned timestamps across sources needed careful handling.

- **Scaling Issues:** Performance challenges with large datasets were mitigated through optimization in Databricks.