# Project Report

## Cloud-Based Retail Data Analytics Pipeline Using Microsoft Azure

Author: Pooja Arumugam

MS in Data Analytics Engineering

Northeastern University

June 2025

### Abstract

This report outlines the development of an end-to-end retail analytics pipeline using Microsoft Azure. The pipeline is designed to handle ingestion, processing, storage, and reporting of retail data in a scalable and efficient manner. Key technologies include Azure Data Factory for orchestration, Data Lake Gen2 for storage, Databricks for transformation, Synapse Analytics for querying, and Power BI for visualization.

### Introduction

In the current data-driven landscape, retail companies require advanced tools to analyze customer behavior, sales performance, and product trends. Traditional ETL systems often lack the flexibility and scalability needed to process large datasets in real time. This project leverages cloud-native technologies to build a robust, modular data pipeline capable of transforming raw data into actionable insights.

### Objectives

The main goals of this project are:

- To build a scalable, modular cloud data pipeline

- To ingest raw retail data from an external source into a data lake

- To process and clean data using PySpark in Azure Databricks

- To expose cleaned data through serverless SQL views in Synapse

- To visualize key metrics and KPIs using Power BI dashboards

### Architecture Overview

The architecture follows a bronze-silver-gold model. Data is ingested using Azure Data Factory into a Bronze zone in ADLS Gen2. Databricks notebooks are then used to transform the data and store it in a Silver zone. Serverless SQL views in Synapse Analytics read the Silver zone data to create Gold views. These views serve as the data source for Power BI dashboards.

## Implementation

1. Data Ingestion: ADF pipelines extract raw Parquet files and store them in the Bronze zone of ADLS.

2. Data Transformation: Azure Databricks is used to clean data, handle null values, perform joins, and standardize schema. The output is written back to the Silver zone.

3. Serving Layer: Synapse uses `OPENROWSET` to read Parquet files and expose data via SQL views.

4. Reporting: Power BI connects to Synapse and visualizes sales trends, product performance, and customer behavior.

## SQL Code Sample

```
CREATE VIEW gold.customers AS

SELECT * FROM OPENROWSET(

  BULK
'https://<storage_account>.blob.core.windows.net/silver/AdventureWorks_Customers/',

  FORMAT = 'PARQUET'

) AS customers;
```

## Challenges

- Mounting issues in Databricks due to incorrect storage key: Resolved by updating workspace credentials.

- OPENROWSET not recognizing schema: Ensured consistent schema and proper folder path structure.

- ADF failure in parameterized pipelines: Fixed by correcting dataset reference and runtime settings.

- Power BI failed to refresh: Configured DirectQuery and enabled gateway refresh.

## Results

- Efficient data ingestion from raw files using ADF

- Clean, analysis-ready datasets processed via Databricks

- Serverless querying through Synapse reduced infrastructure cost

- Real-time Power BI dashboards enable actionable business insights

## Future Work

- Implement Great Expectations for data validation and profiling

- Automate notebook execution using ADF triggers

- Include multiple data domains (e.g., sales, marketing)

- Apply machine learning for sales forecasting and segmentation

- Implement CI/CD using GitHub Actions or Azure DevOps

## Conclusion

This project demonstrates the power and flexibility of Azure for building modern data pipelines. The use of serverless and cloud-native tools enables efficient processing, seamless integration, and cost-effective business intelligence solutions. The modular nature of this pipeline also makes it easy to scale and adapt for future enterprise use cases.

## References

- Microsoft Learn: Azure Data Factory, Synapse, and Databricks

- Databricks Documentation

- Power BI Official Documentation

- Parquet Format Specification

- GitHub Azure Samples