CELEBAL SUMMER INTERNSHIP

PROJECT REPORT

Project Title: Change Data Capture (CDC) ETL Pipeline using Azure Data Factory and Databricks

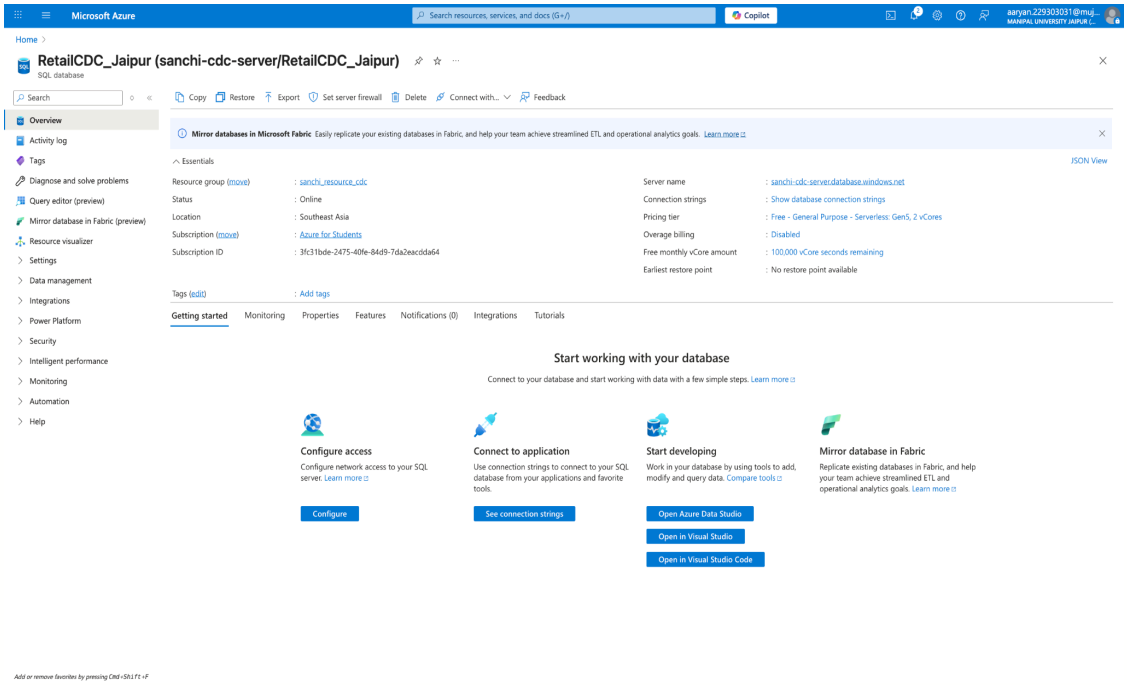
Submitted by: Sanchi Mishra

Domain: Data Engineering
Student ID: CT_CSI_DE_5042
Internship Duration: 2 months
College Name: Manipal University Jaipur

Project Description: This project focuses on implementing Change Data Capture (CDC) using Databricks, a unified data analytics platform. The primary objective is to efficiently identify and process data changes (inserts, updates, deletes) in source systems and reflect them in the target data lake or data warehouse in near real-time. Leveraging Apache Spark, Delta Lake, and streaming capabilities within Databricks, the project aims to build a scalable and reliable CDC pipeline that supports incremental data ingestion, ensures data consistency, and minimizes processing overhead.

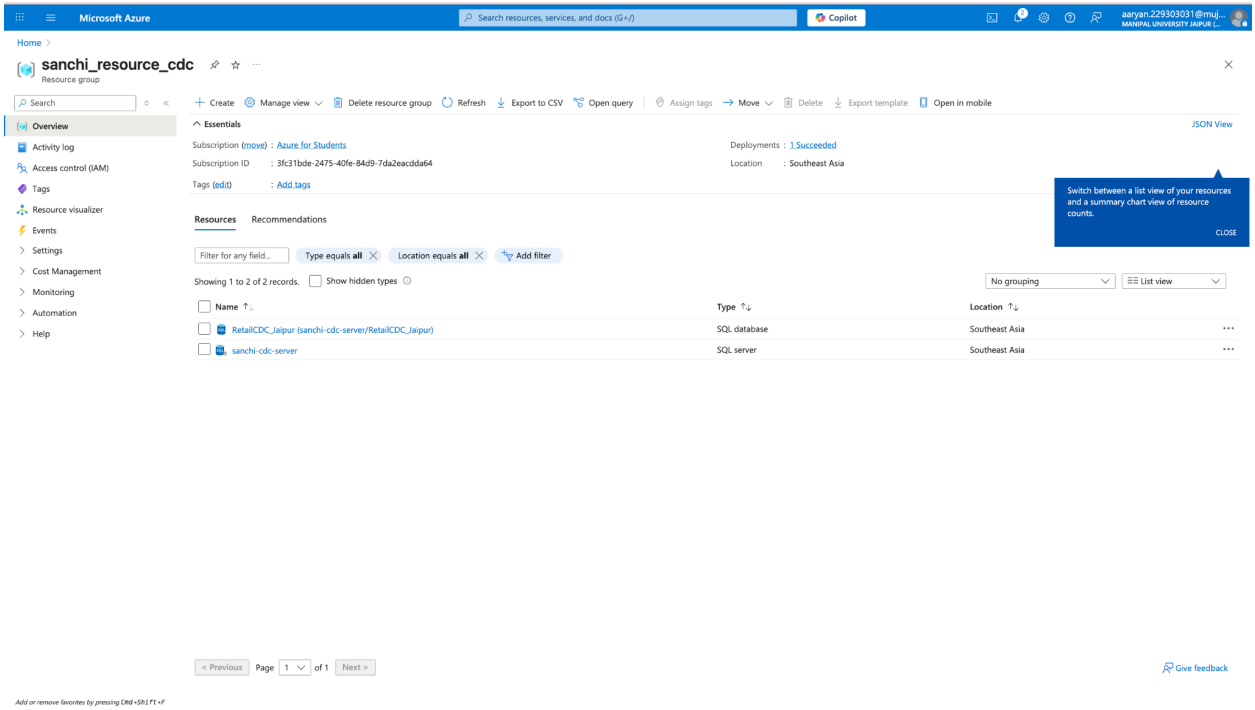
Github Repository: Celebal Project

1. SQL Database Creation: RetailCDC_Jaipur



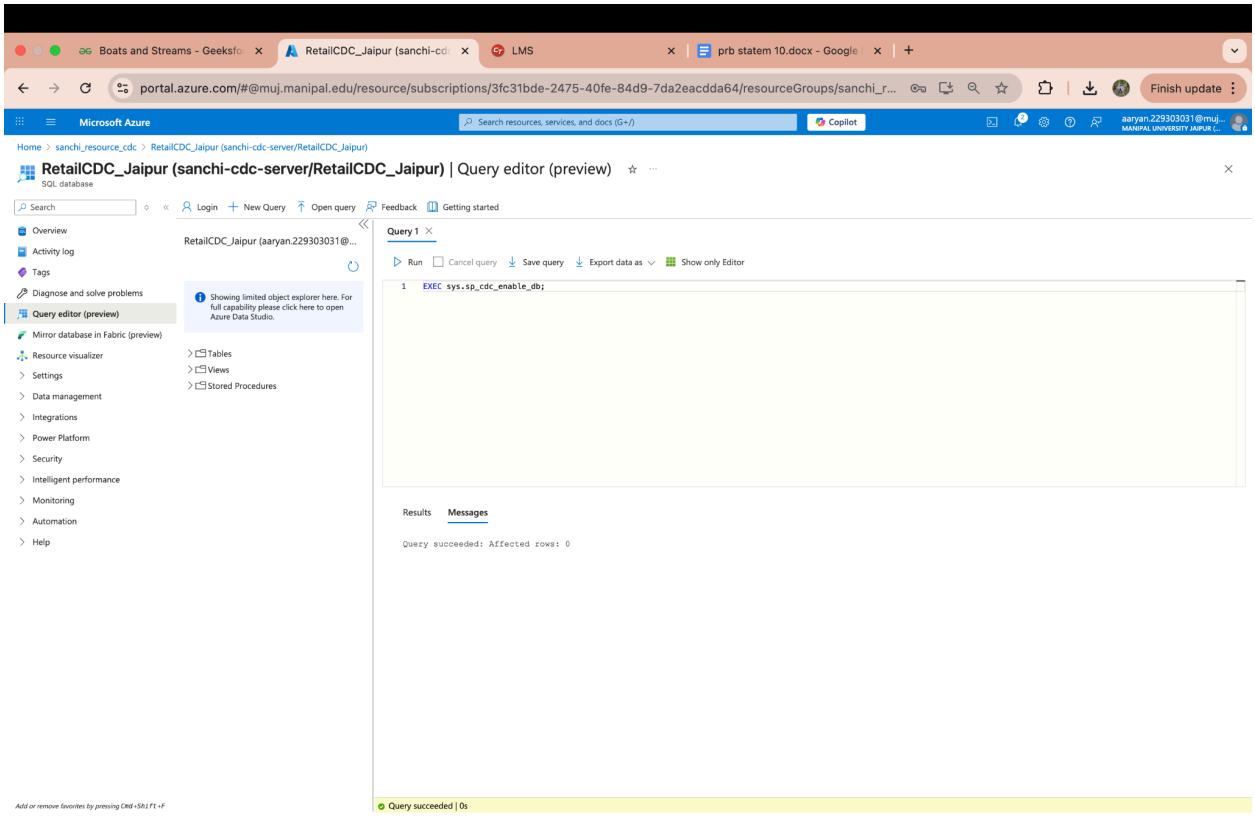
Add or remove favorites by pressing Ctrl+Shift+F

2. Resource Creation: sanchi_resource_cdc

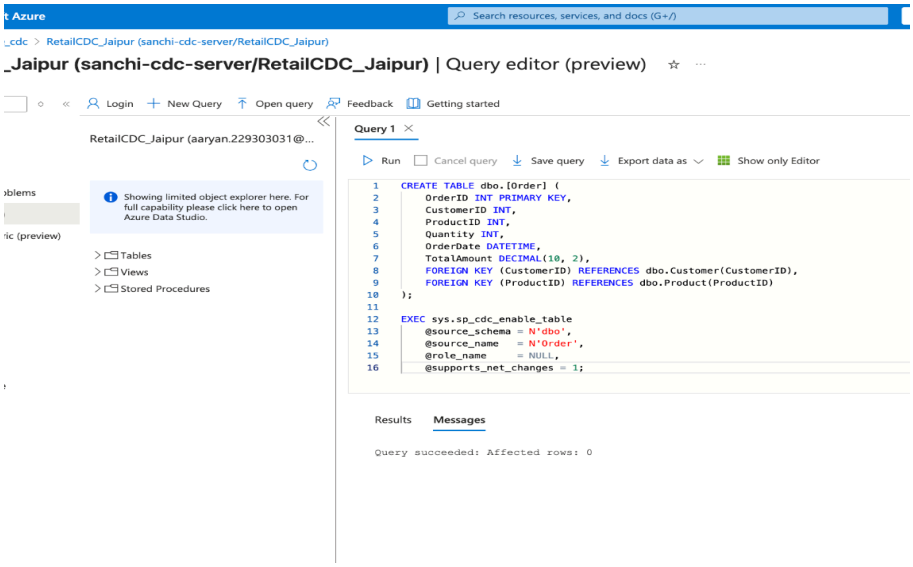


Add or remove favorites by pressing Ctrl+Shift+F

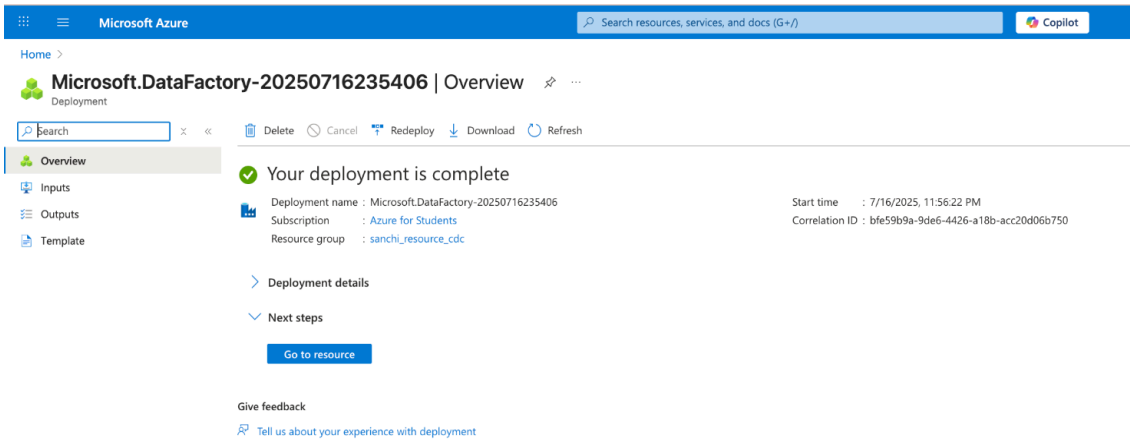
3. Enable CDC at Database Level



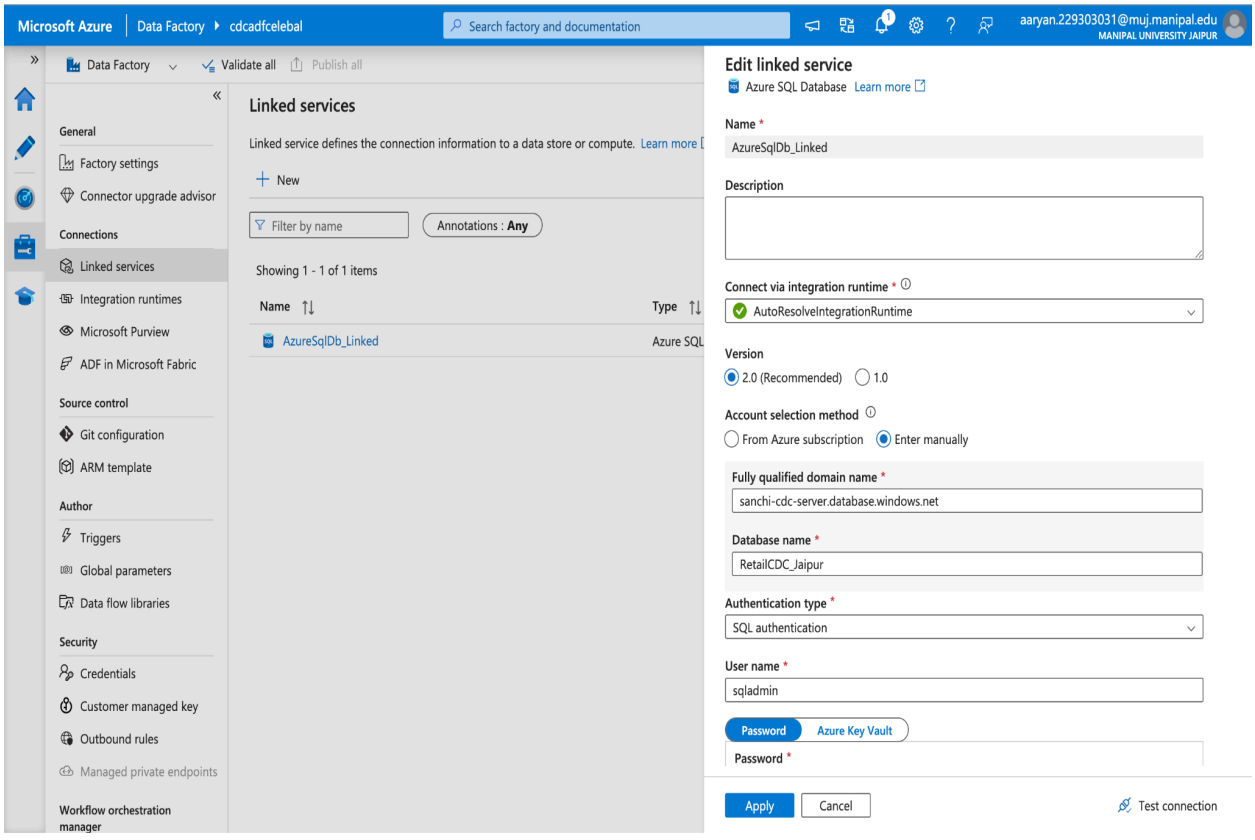
4. Creating Tables (Customer, Order, Product, Inventory) and Enabling CDC (attaching just one eg)



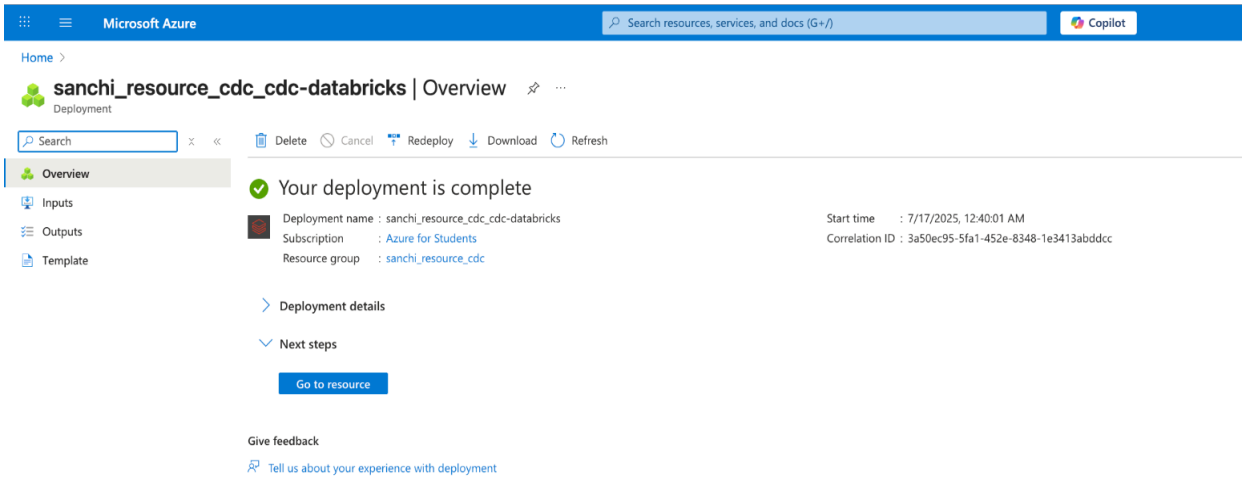
5. Azure Data Factory Creation and Deployment



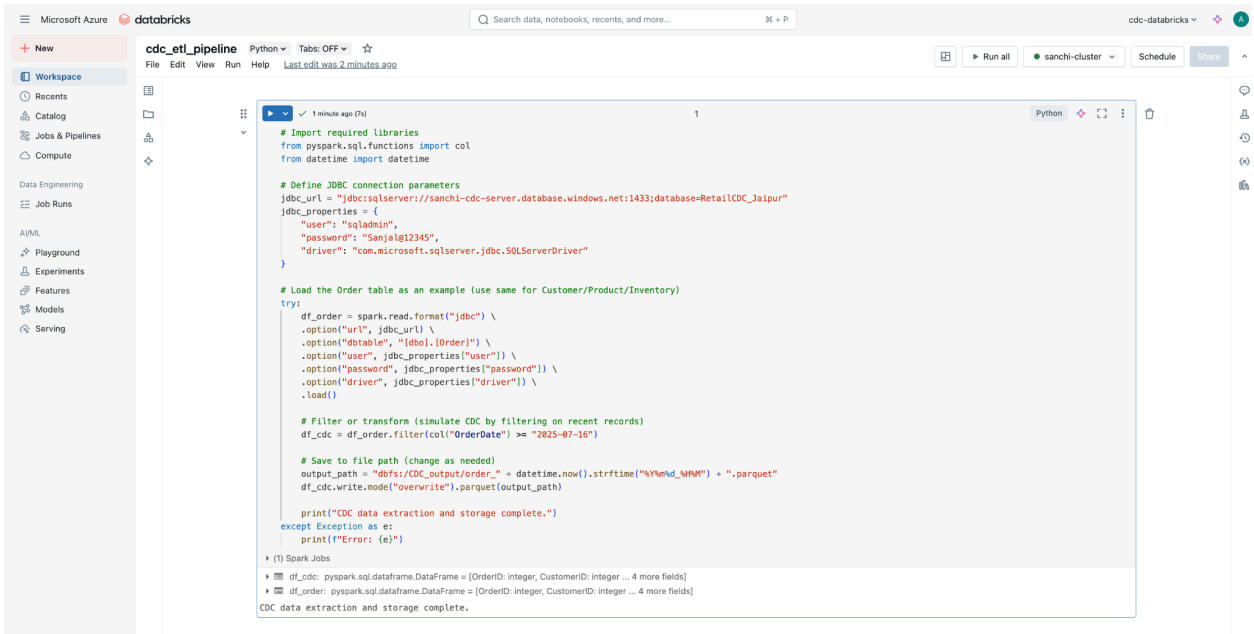
6. Azure Linked Service



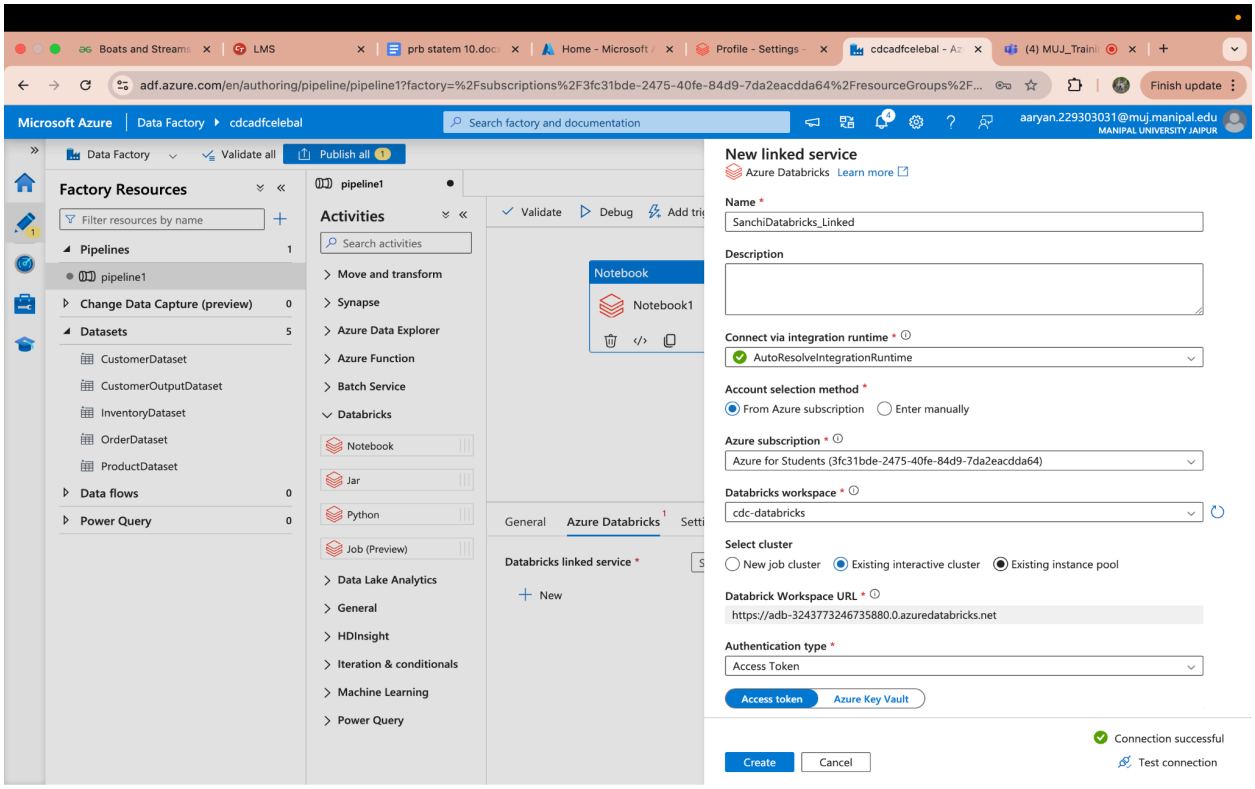
7. Data Bricks WorkSpace Creation



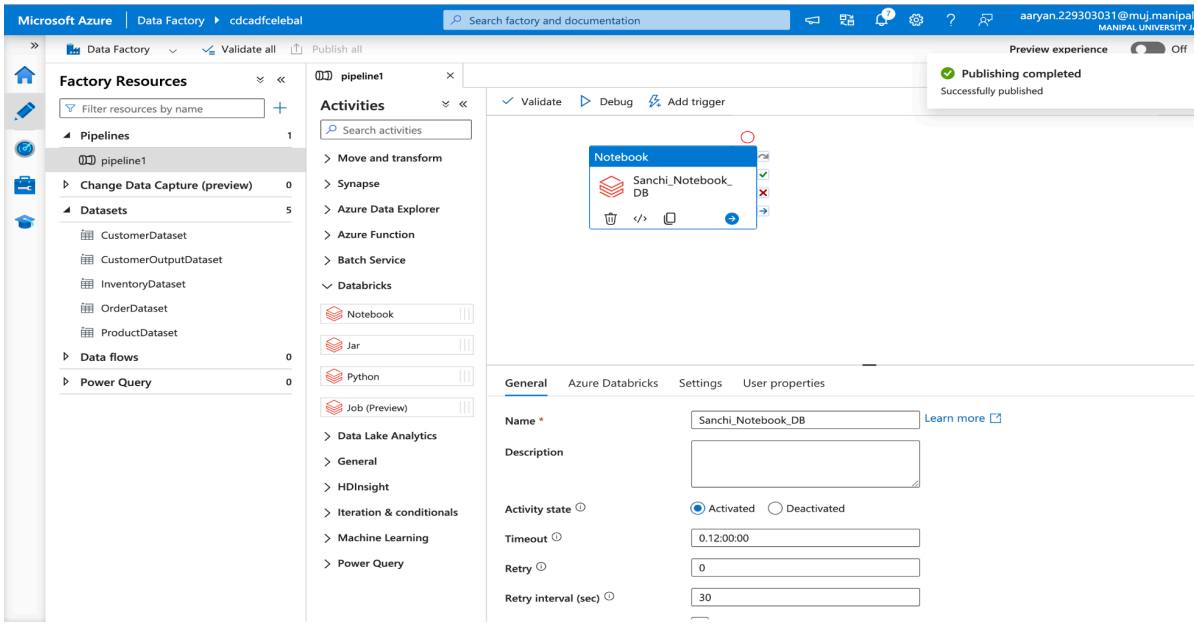
8. Data Bricks Notebook Glimpse (Full Notebook attached in Github Repository)



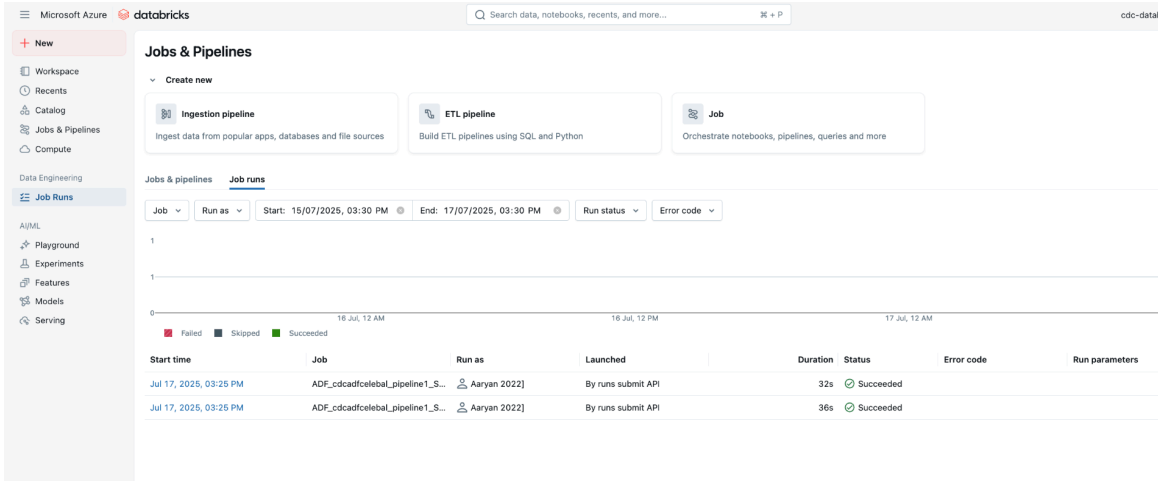
9. Linking Database to Azure



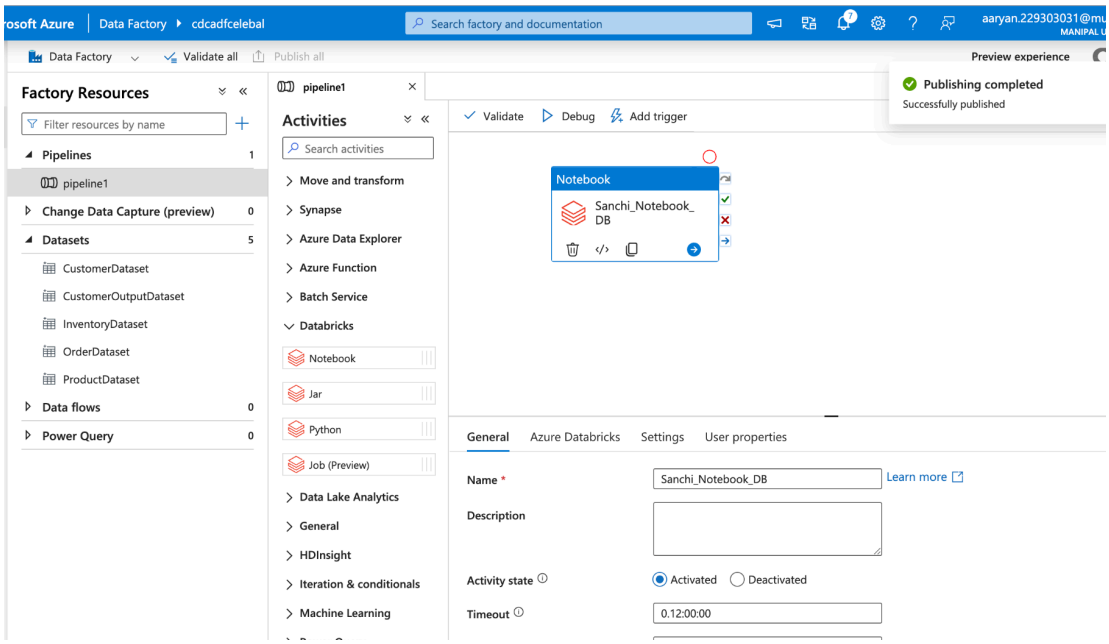
10. Notebook in Azure Published



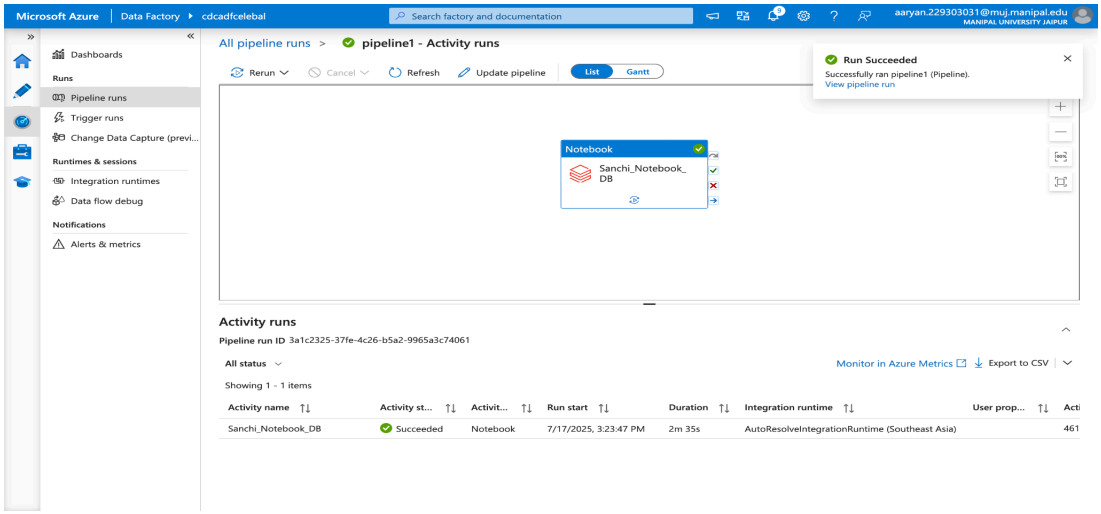
11. Data Bricks Pipeline Ran Successfully



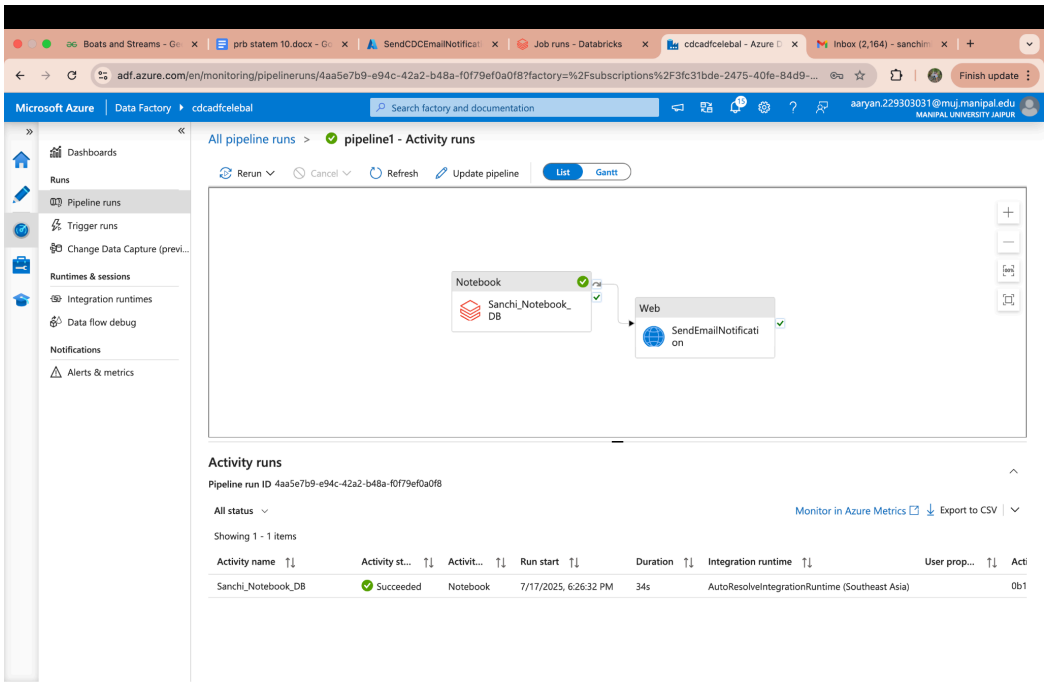
12. Notebook in Azure Linked and Published



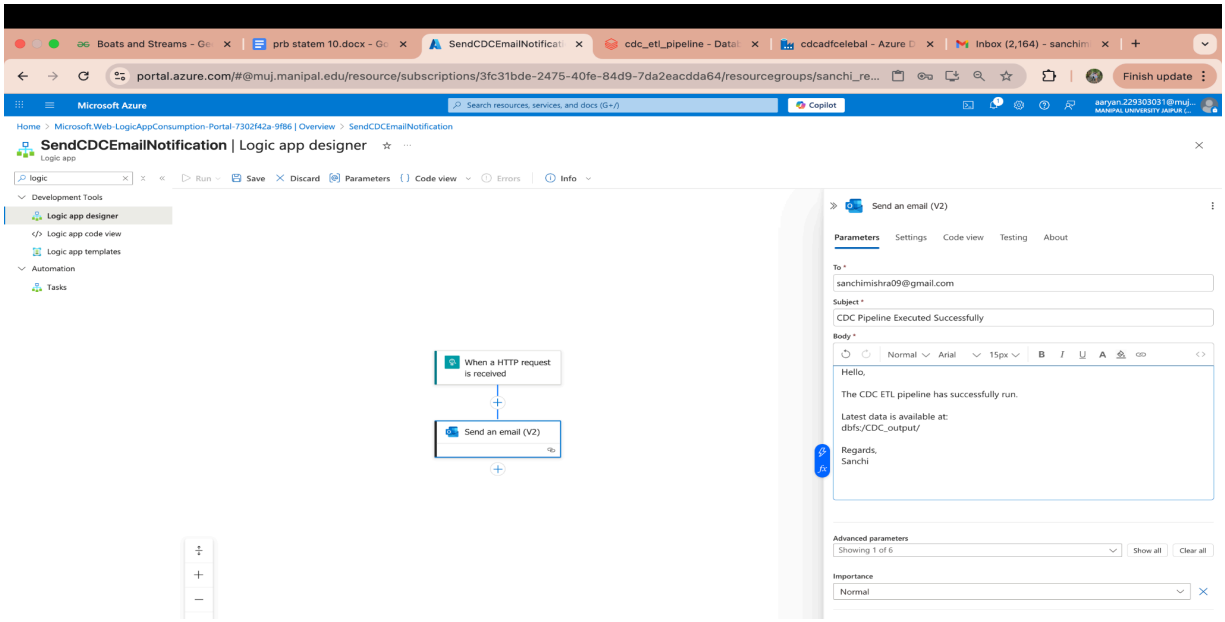
13. Pipeline Ran Successfully



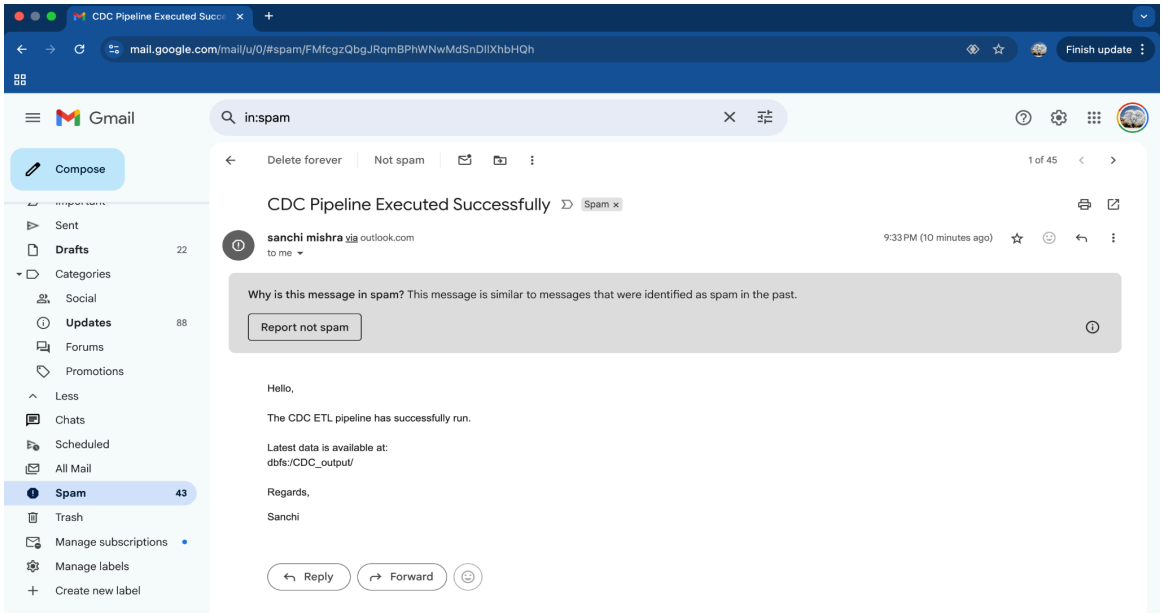
14. Email and Notebook Connected and Ran



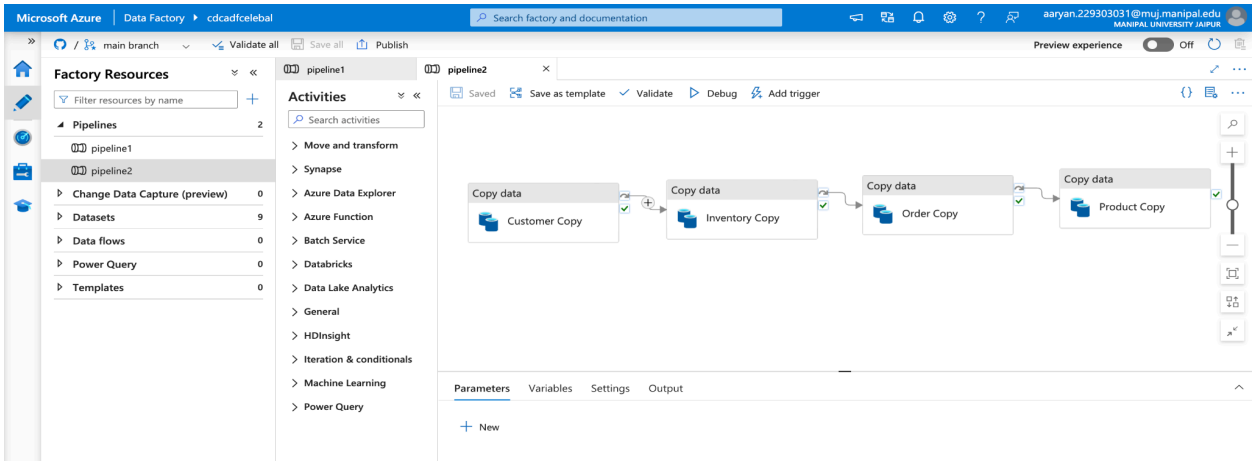
15. Email Setup

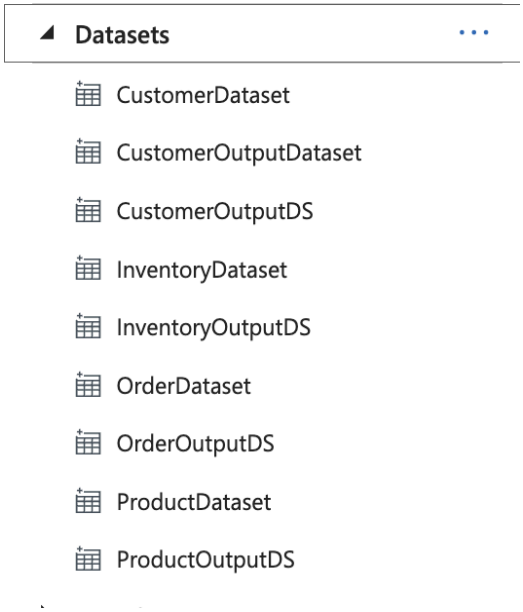


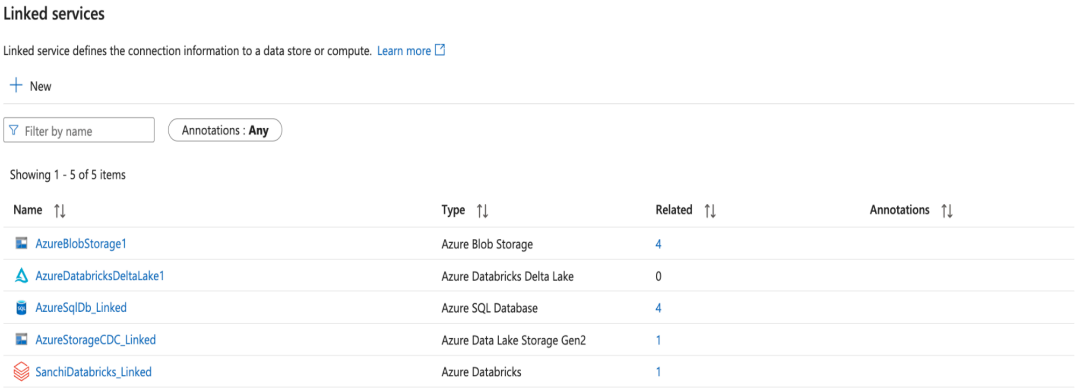
16. Email Received

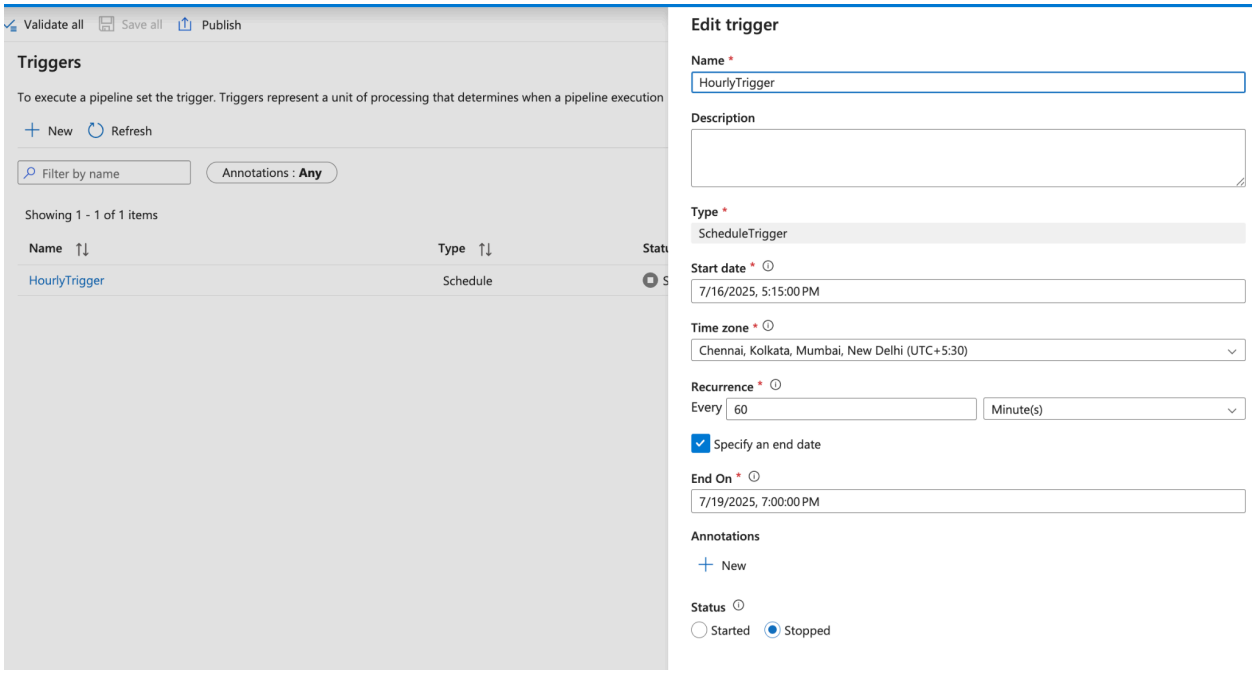


17. A few more screenshots from the pipelines: Datasets, Triggers, Emails, Deployments and Outputs









I am pleased to share that I have successfully completed the **Change Data Capture (CDC) ETL Pipeline Project** using Azure Data Factory and Azure Databricks

. The solution includes:

- CDC-enabled SQL Server tables
- Automated ETL using Databricks notebooks
- ADF pipelines for scheduling and integration
- Email notifications on successful runs
- Export of updated data to Azure Blob Storage

I have also attached a few screenshots as proof of successful pipeline runs, output verification, and final email notifications and a github repository.

I am truly grateful for the opportunity to work on this project and enhance my practical knowledge of modern data engineering pipelines. Looking forward to receiving your feedback and contributing further.

Thank you!

Sanchi Mishra