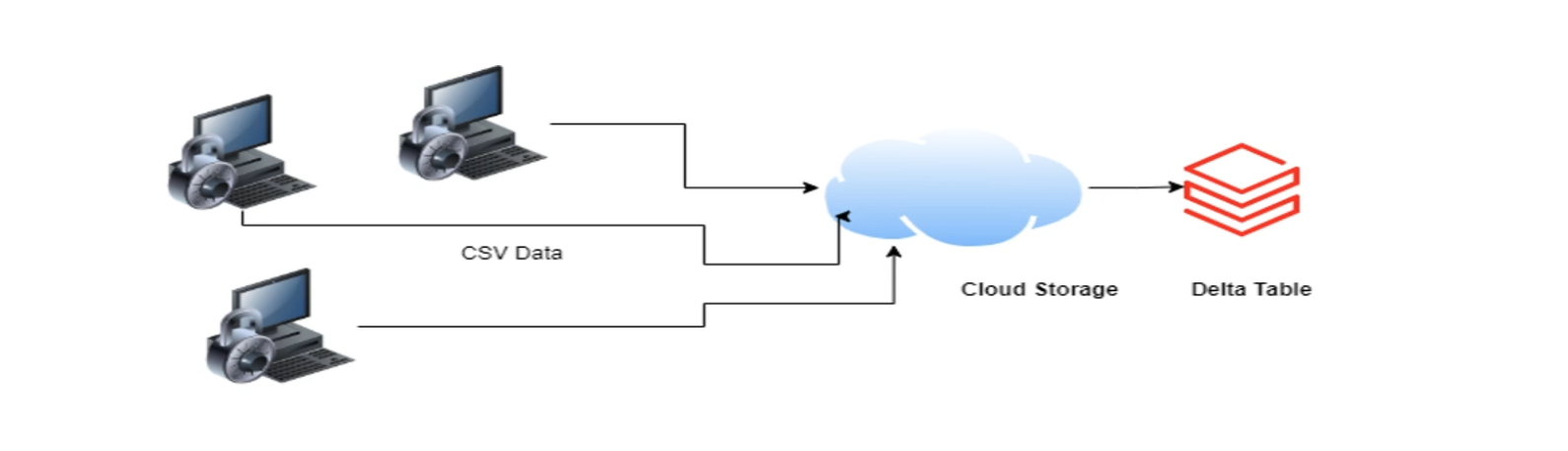
**Project – 2**

**Preview:**

Multiple Internal applications sends the data(huge size) in CSV format on daily basis in the cloud storage location. There are couple of Data/schema validation needed to be performed on this incoming data. Once everything is passed data to be persisted as Delta table in Databricks for downstream system.



**In-Depth of Requirements:**

* Internal Application sends CSV filein Azure data lake storage.
* Validation needed to apply on this follows:
* Check for duplicate rows if it contains duplicate rows, file need to be rejected.
* Need to validate the data format for all the data fileds. Data column names and desired date format is stored in a Azure SQL server. If validation fails file will be rejected.
* Move all the rejected files to Reject folder.
* Move all the passed files to Staging folder.
* Write the passed files as the Delta table in the Azure Databricks.

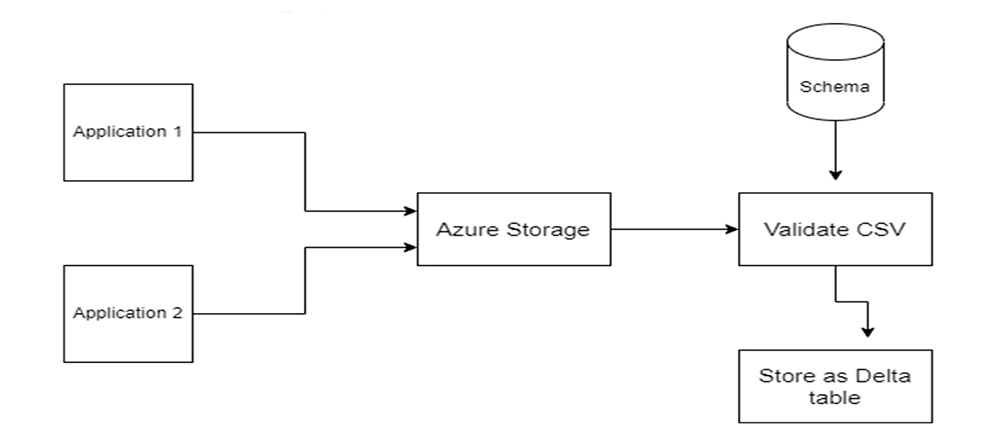


Fig: Diagramatic view of Process

**Architecture:**

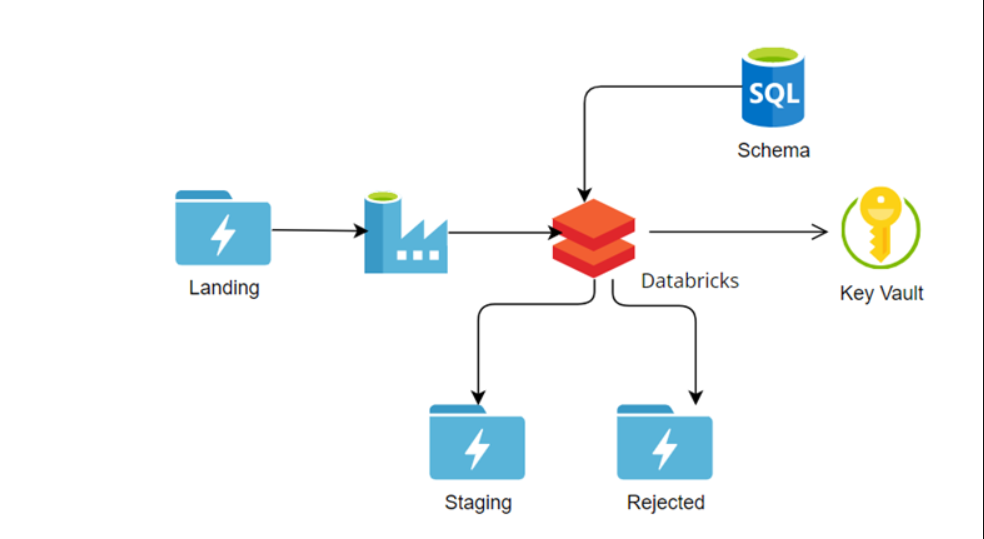
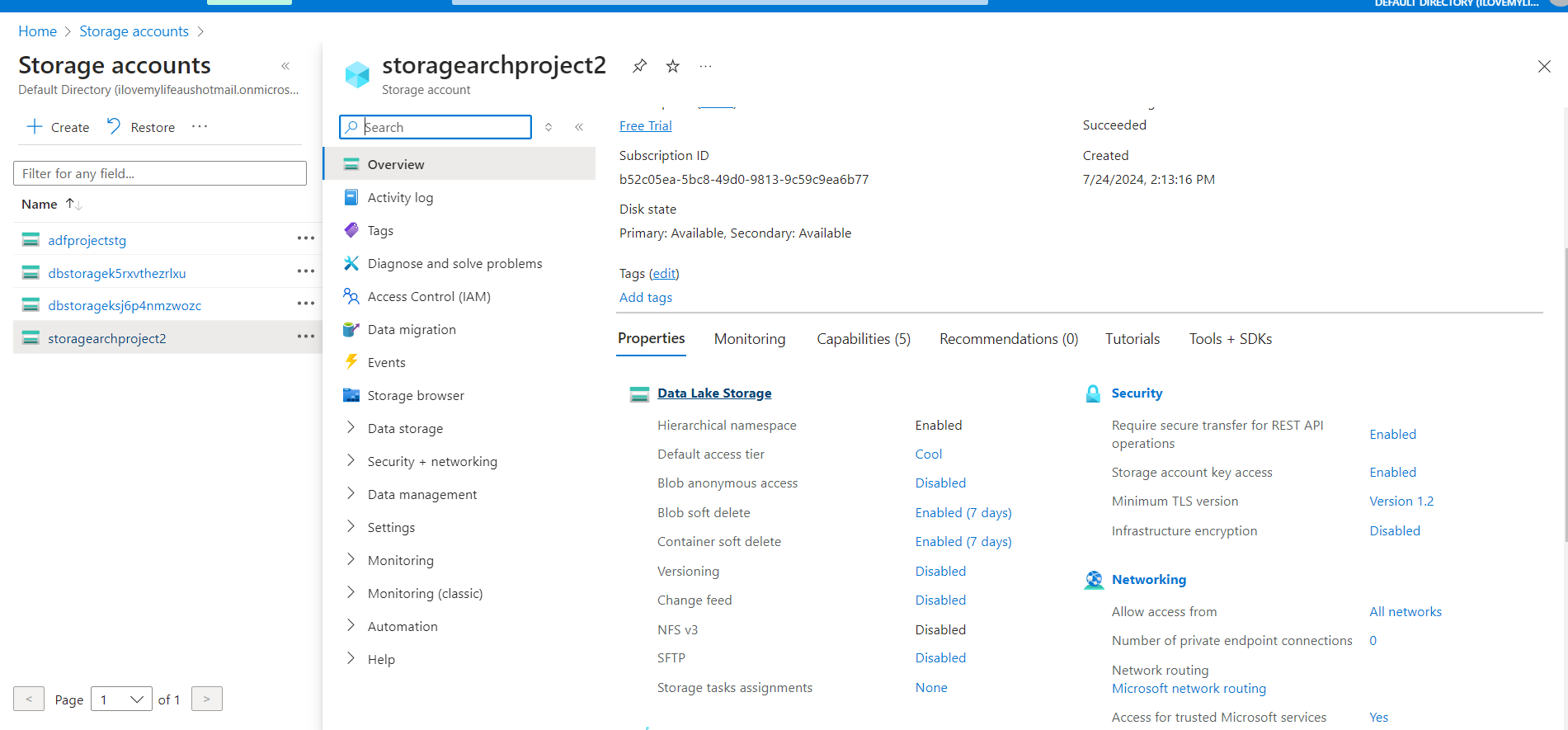


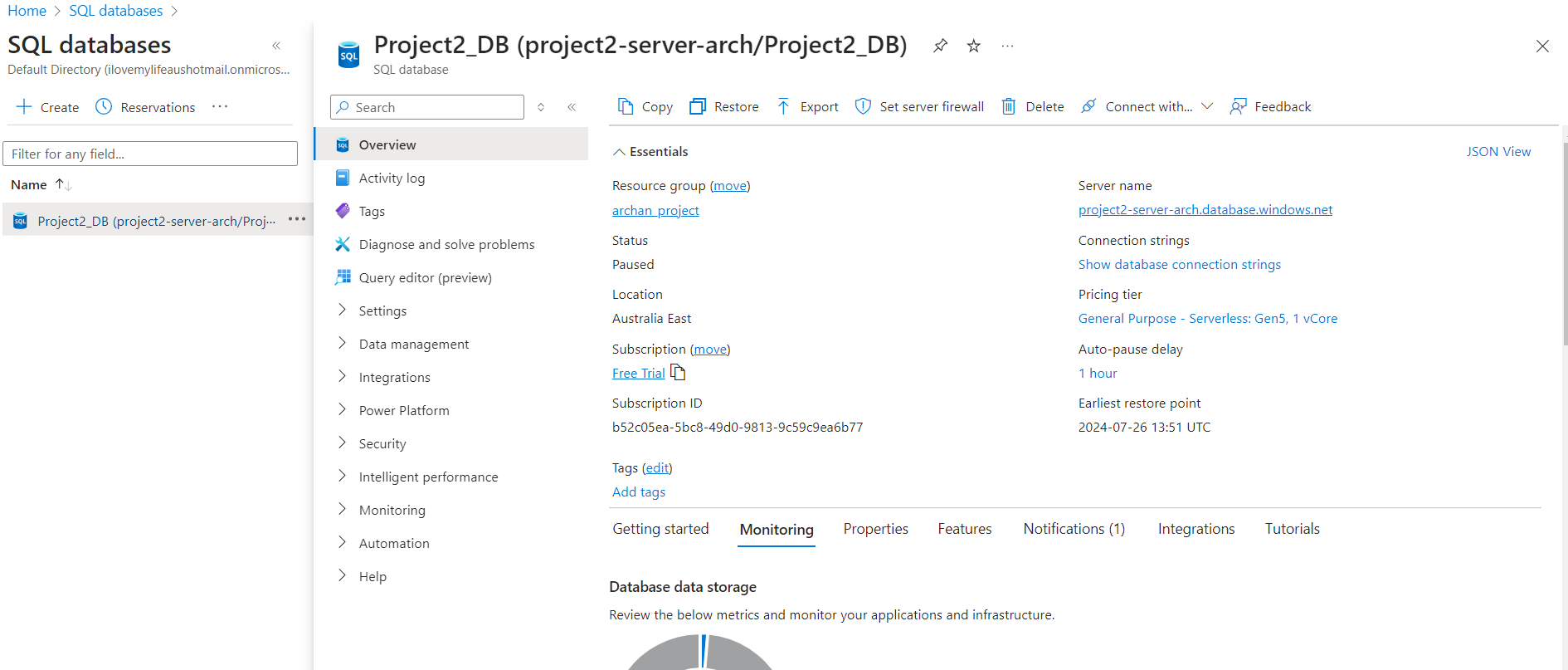
Fig: Architecture Diagram

* Application will send input data file into ADLS location (Landing Folder) and the file need to be validated based on some rules according to the requirements and this task is need to be done daily. So this task need to schedule using ADF.
* ADF gets data from Landing folder and it calls the Databricks.
* Spark notebookfetch the data and to apply the validation rule, data schema need to fetch from SQL Server.
* Based on the validation result, file is moved to Staging folder or Rejection folder by creating Delta lake.
* For connection from ADF to Databricks and from SQL Server to Databricks we will use Key Vault.

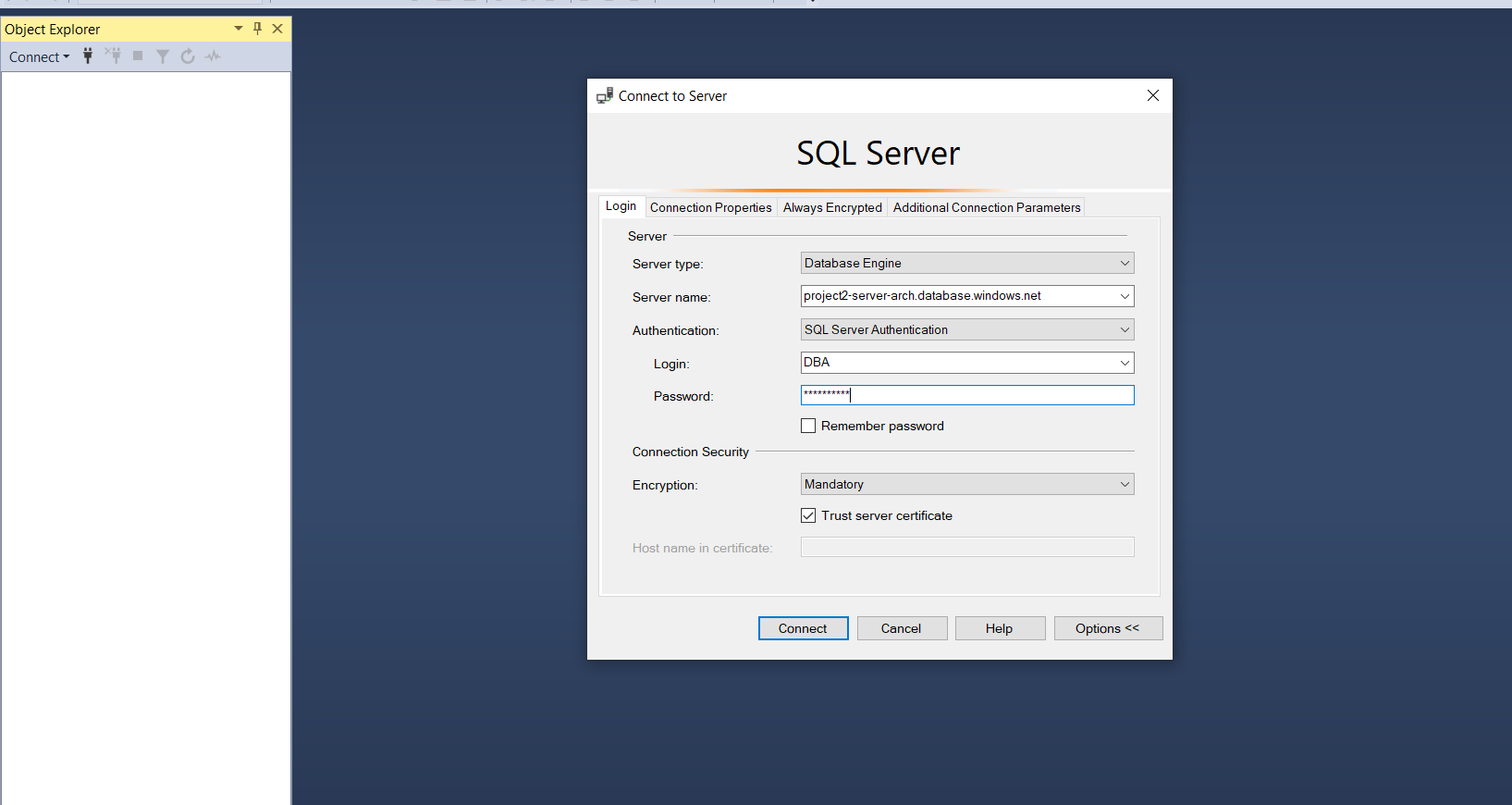
**Step 1: Create a ADLS storage account (Landing Folder)**

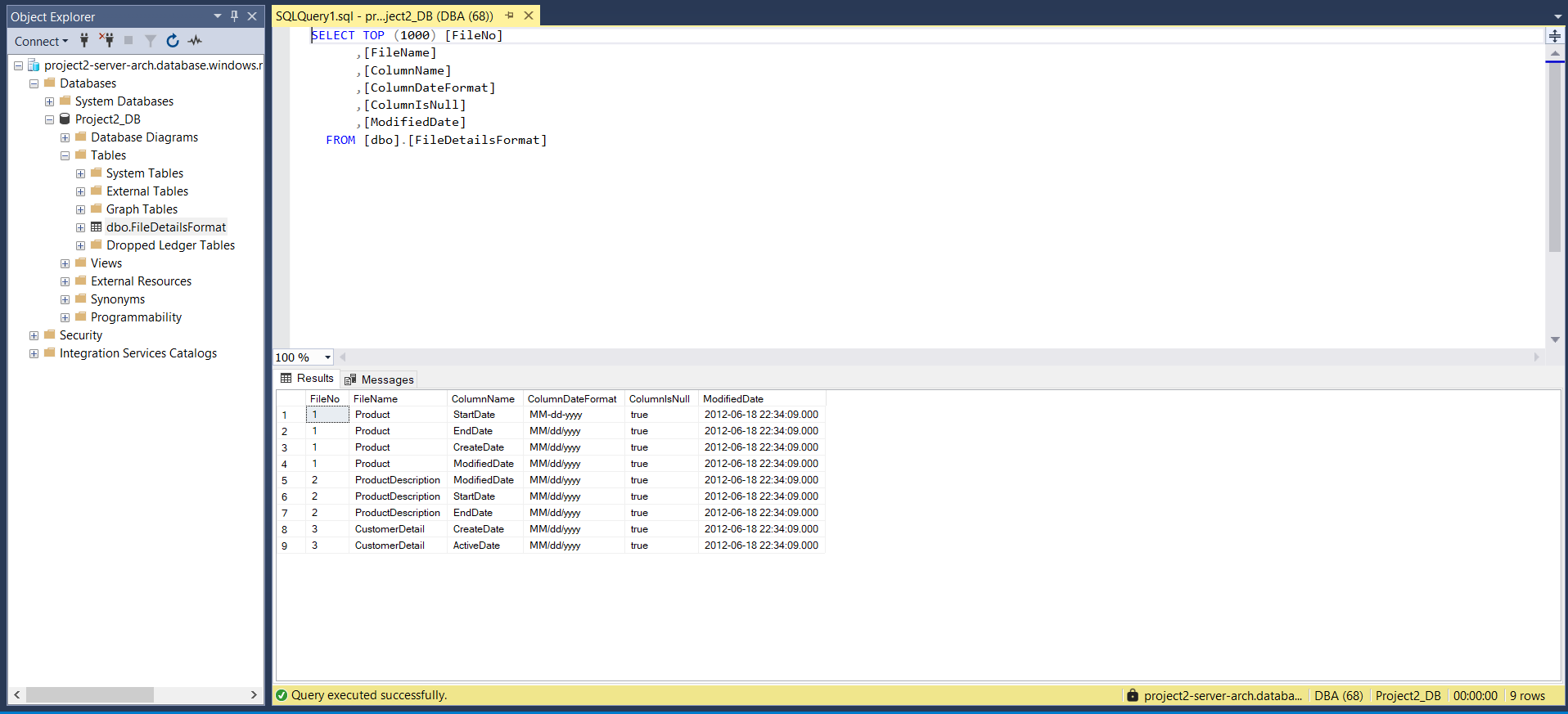


**Step2: Create Azure SQL Server**

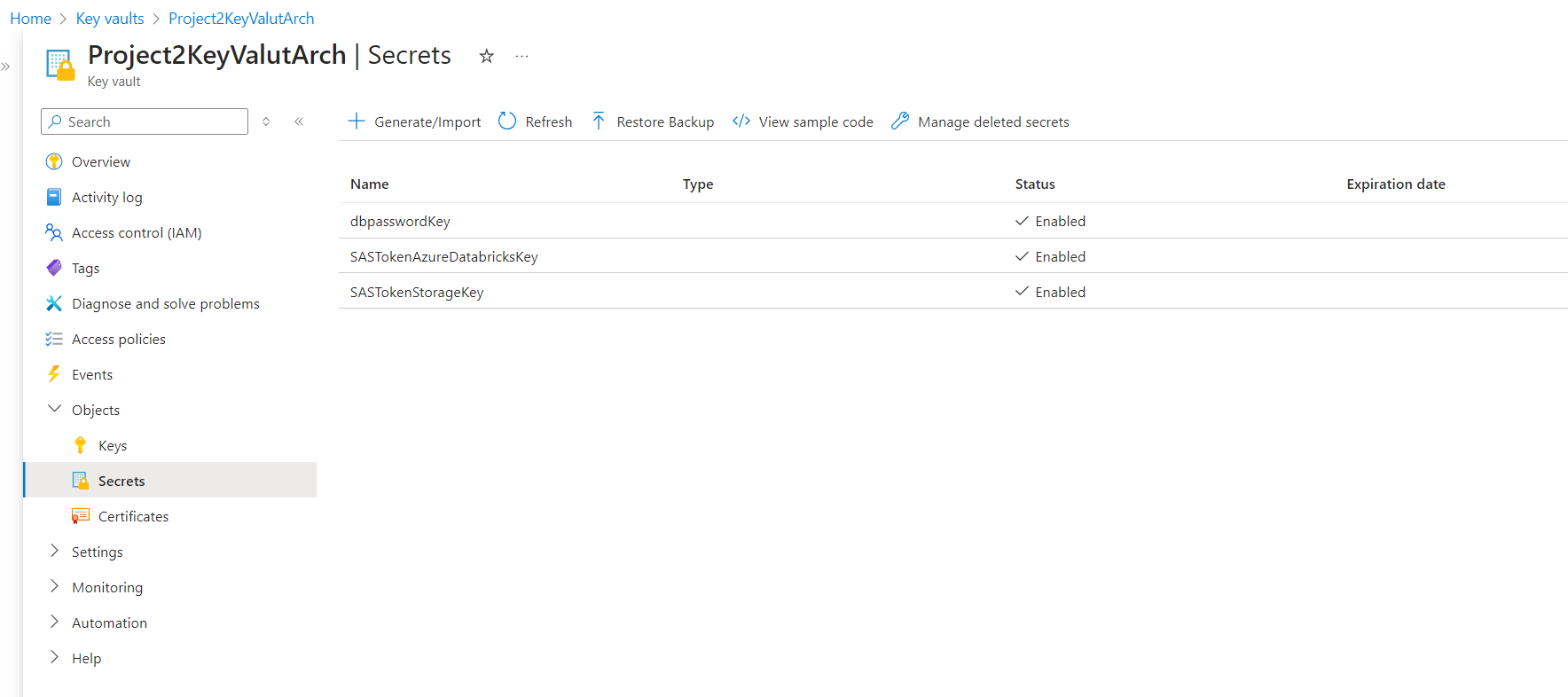


**Step3: Connect SSMS and Data Load**

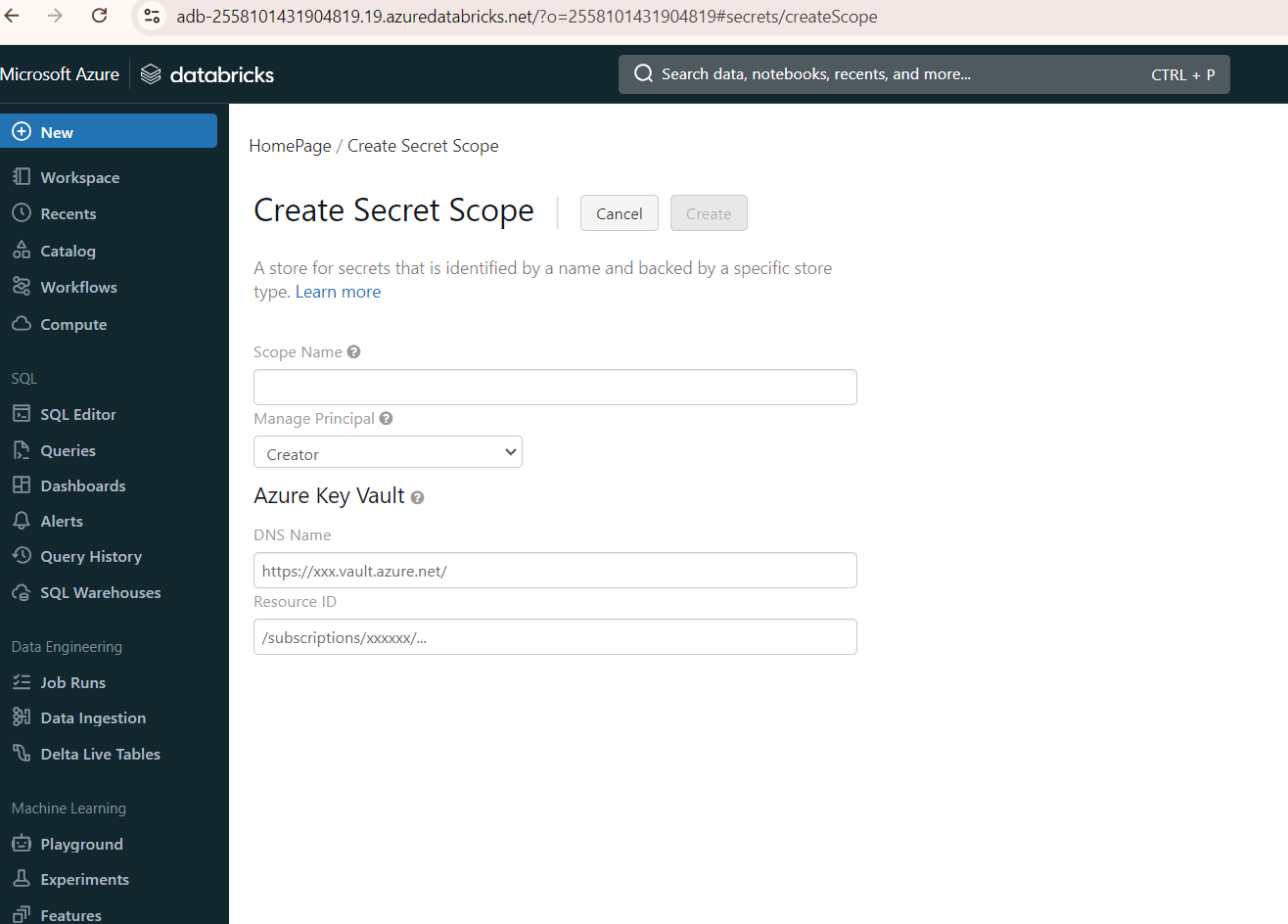




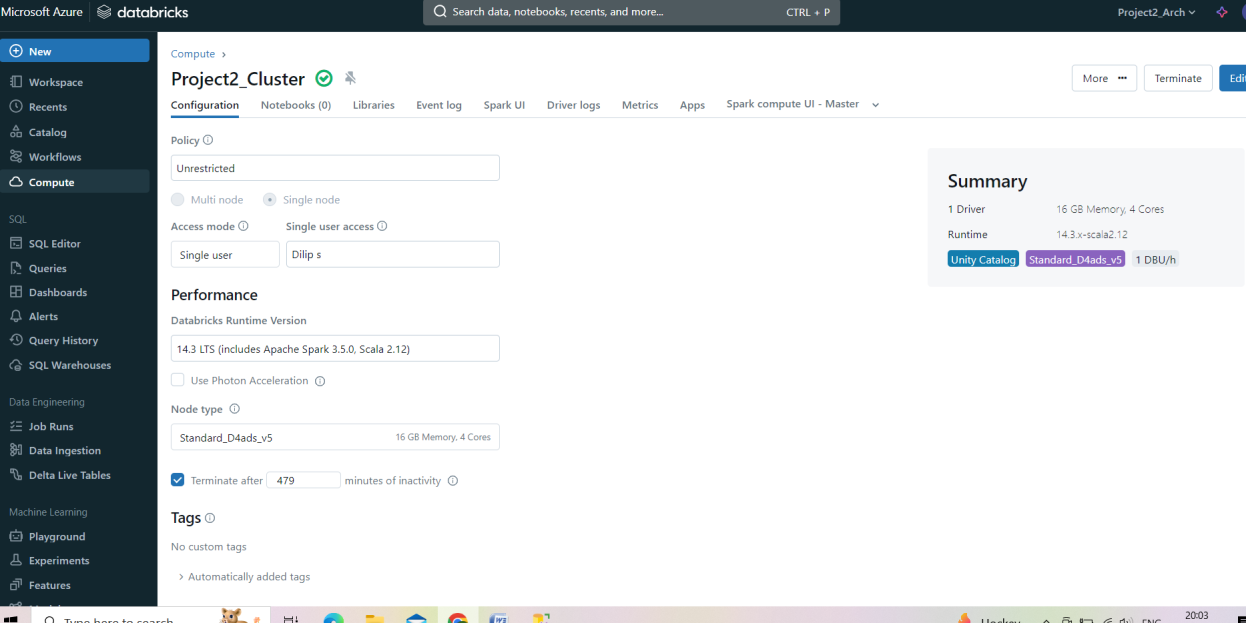
**Step4: Create a Key Vault and Store Secret**



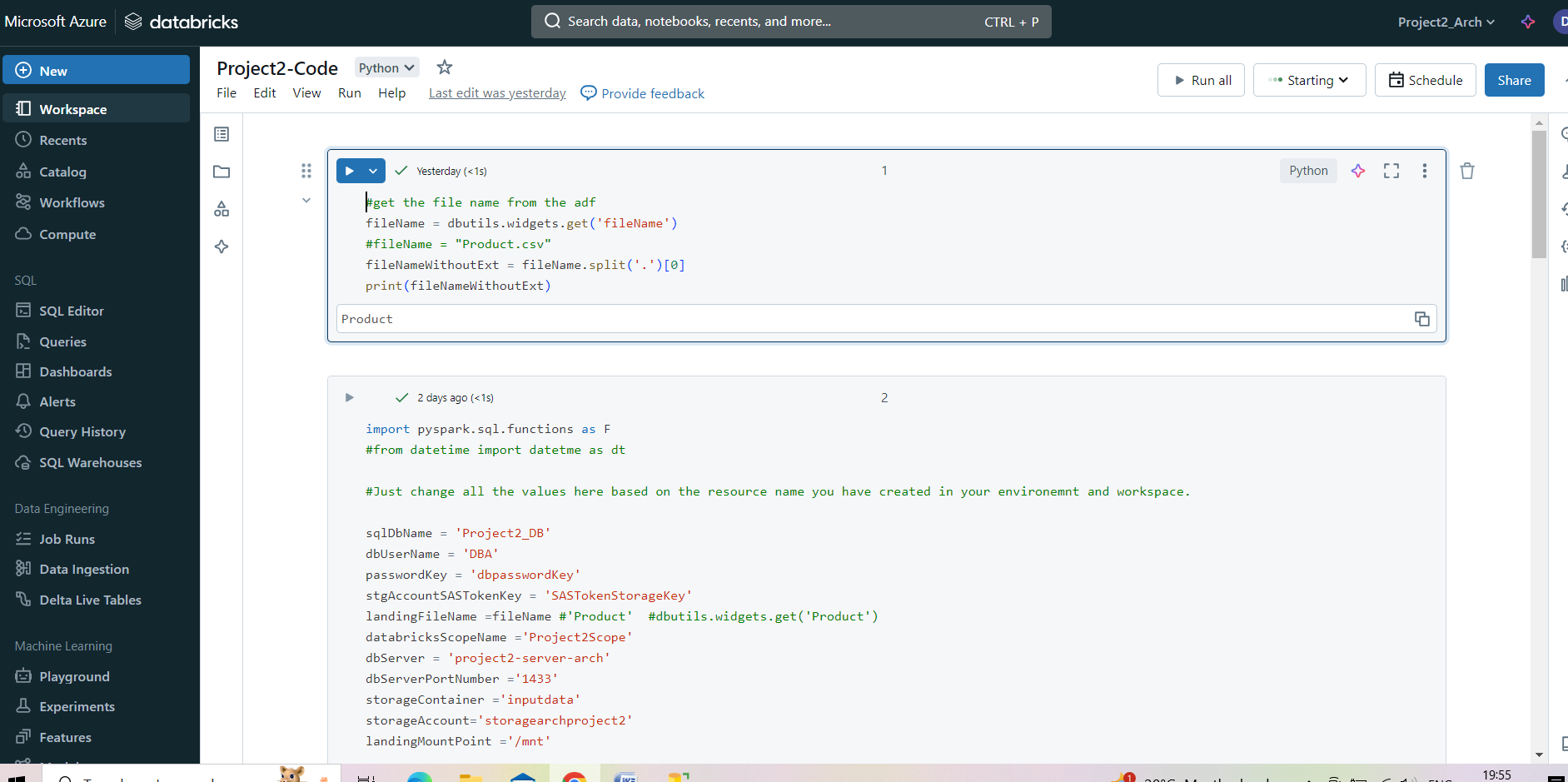
**Step5: Create a Secrete Scope in Databricks**



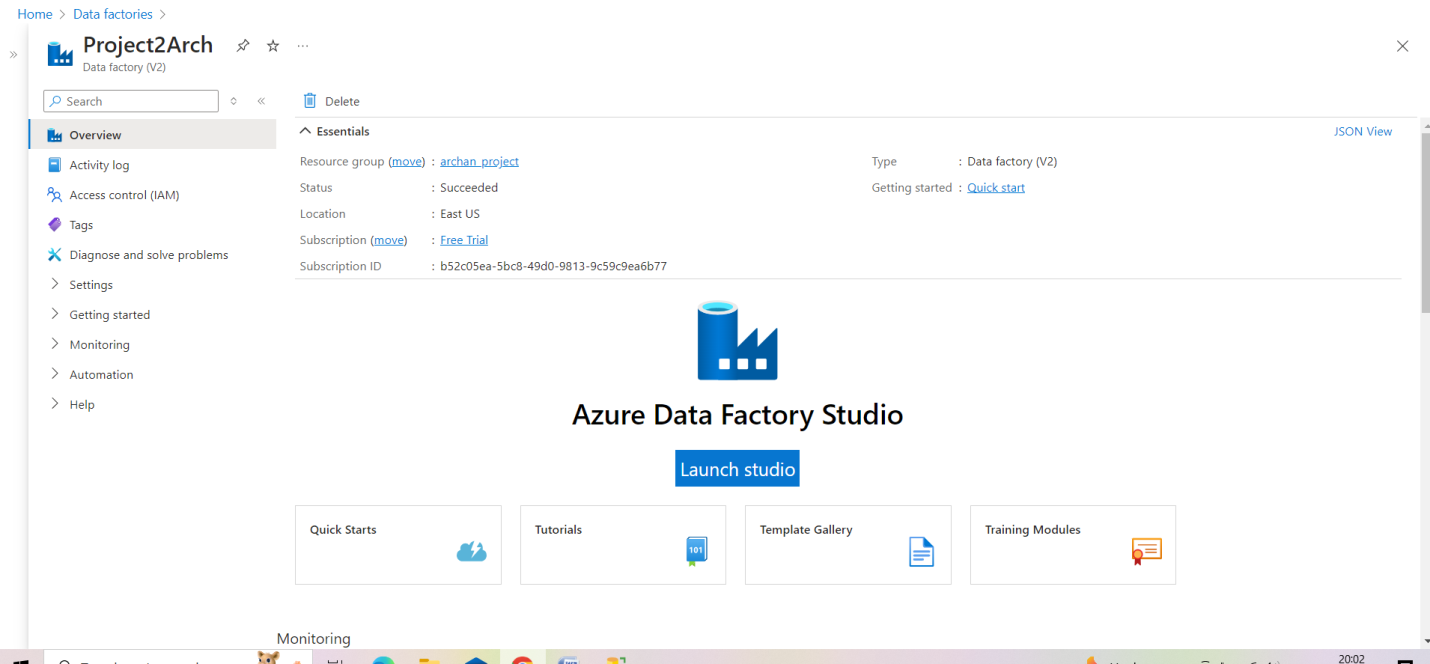
**Step6: Create Cluster in Databricks**



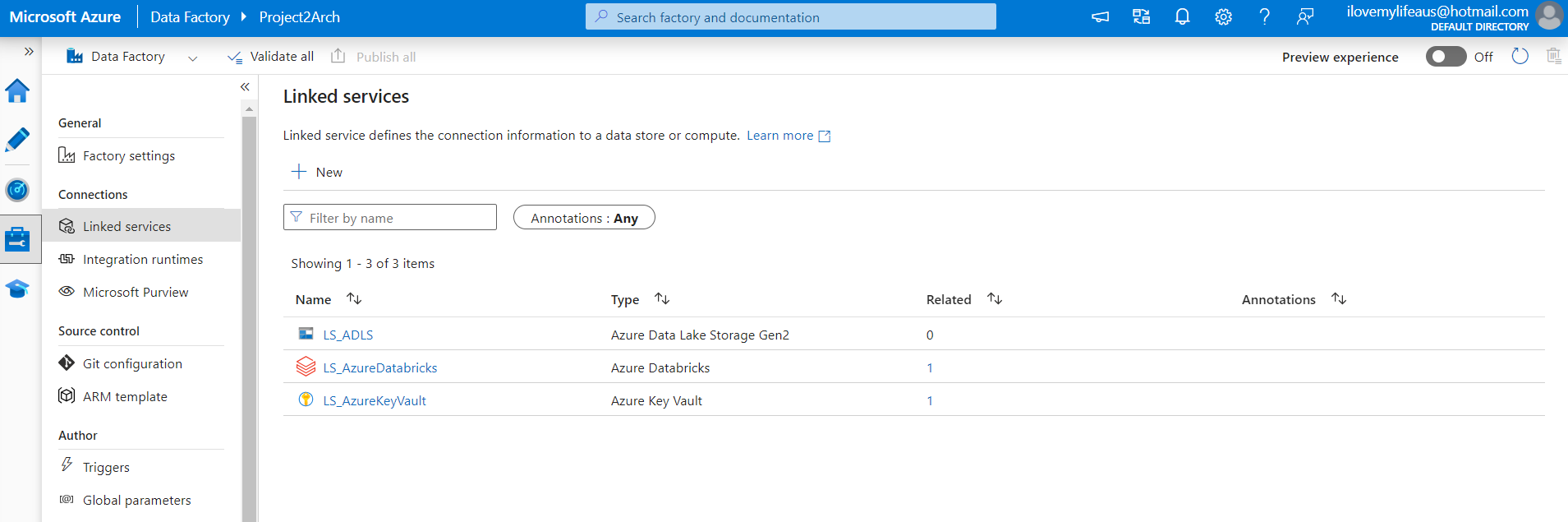
**Step7: Develop a Notebook and logic**



**Step8: Create Azure Data Factory Account**



**Step9: Create Databricks Linked Service**



**Step 10: Create ADF Pipeline with Trigger and Test End to End Flow**

