**Project-3: ETL for Healthcare Data  
  
By Manjunath P**

 **Data Ingestion:**

* Use Azure Databricks to read the diabetes dataset from a cloud storage solution (e.g., Azure Blob Storage or Azure Data Lake Storage).
* Implement secure and efficient data loading practices to ensure data integrity.

 **Data Transformation:**

* Clean the data by handling missing or null values and ensuring data consistency.
* Perform necessary transformations, such as converting data types, creating new columns (e.g., age groups), and applying aggregations or calculations.
* Use Delta Lake in Databricks for efficient data versioning and transaction management.

 **Data Loading:**

* Load the transformed data into a Delta Lake table or a target data warehouse like Azure SQL Data Warehouse (Synapse).
* Ensure the data is properly indexed and optimized for querying.

 **Monitoring and Logging:**

* Implement monitoring and logging to track the ETL pipeline's performance and identify any issues.
* Use Azure Databricks job scheduling to automate the pipeline execution at regular intervals.

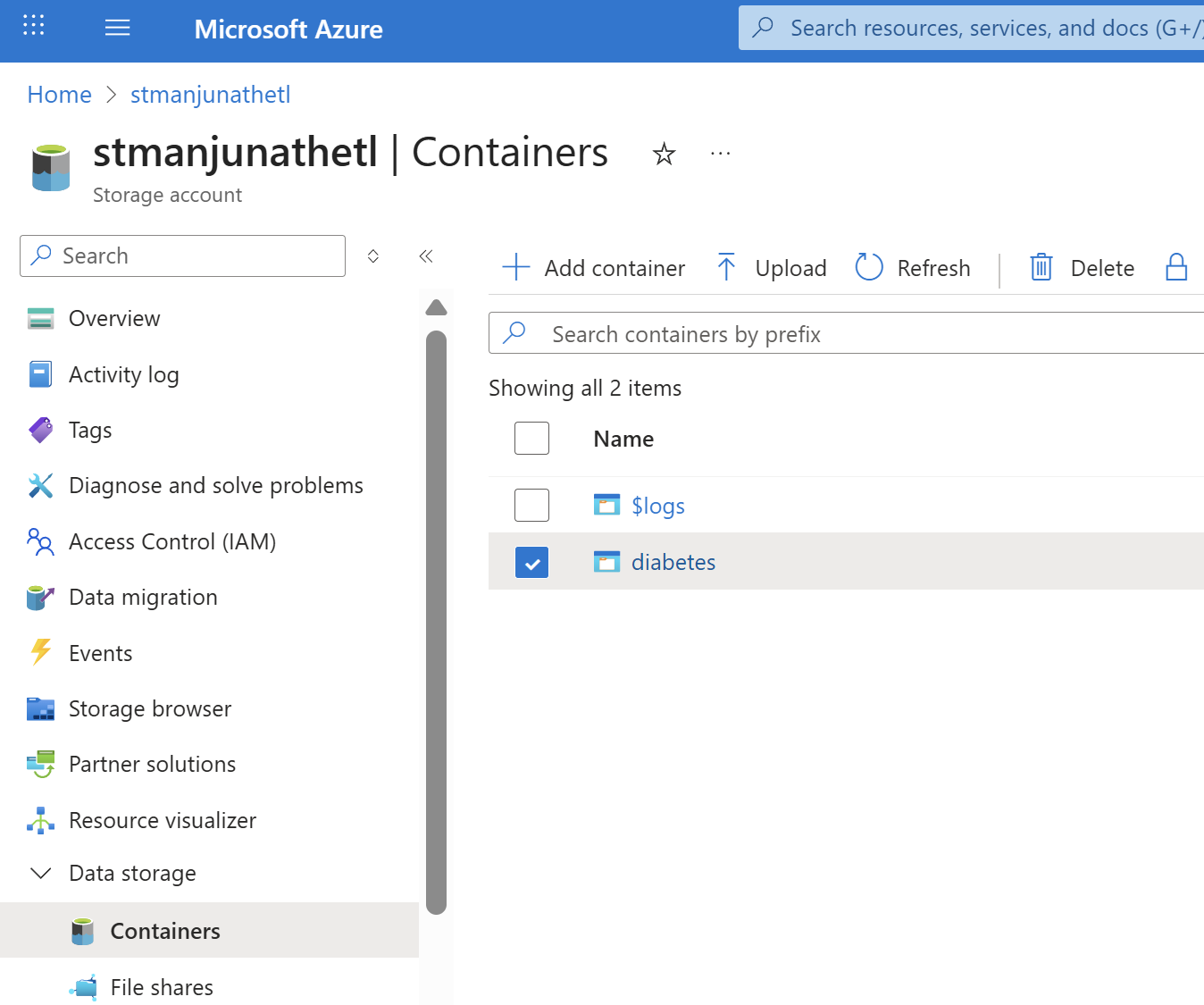
## Dataset

The dataset used in this project consists of the following columns:

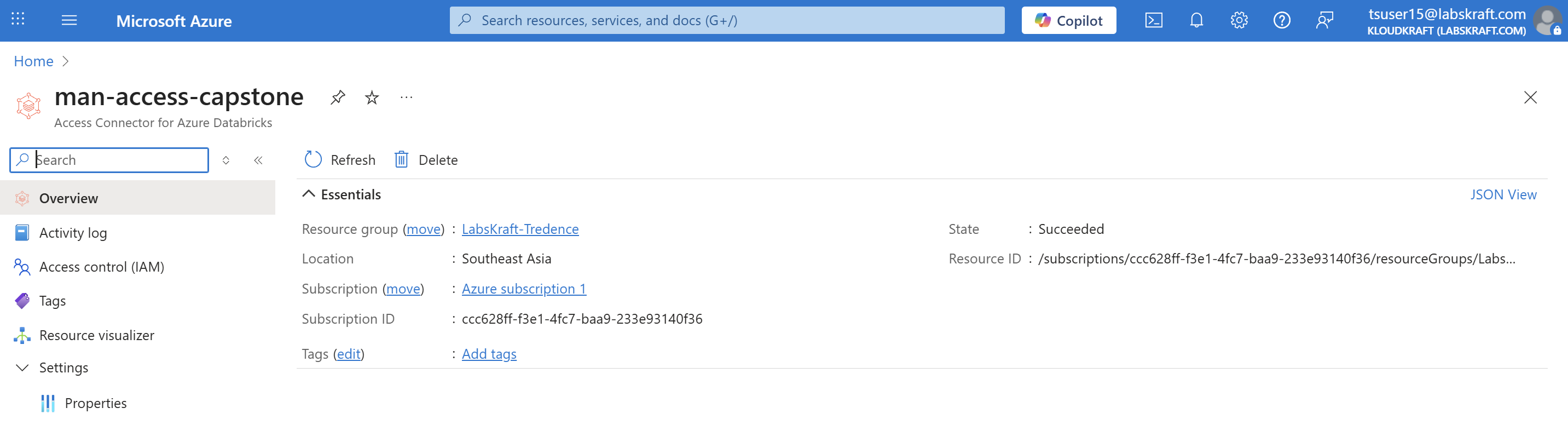
* **Pregnancies**: Number of pregnancies
* **Glucose**: Plasma glucose concentration
* **BloodPressure**: Diastolic blood pressure (mm Hg)
* **SkinThickness**: Triceps skin fold thickness (mm)
* **Insulin**: Serum insulin (mu U/ml)
* **BMI**: Body mass index
* **DiabetesPedigreeFunction**: Diabetes pedigree function
* **Age**: Age of the patient
* **Outcome**: Class label indicating diabetes (1) or no diabetes (0)

**Project Setup**

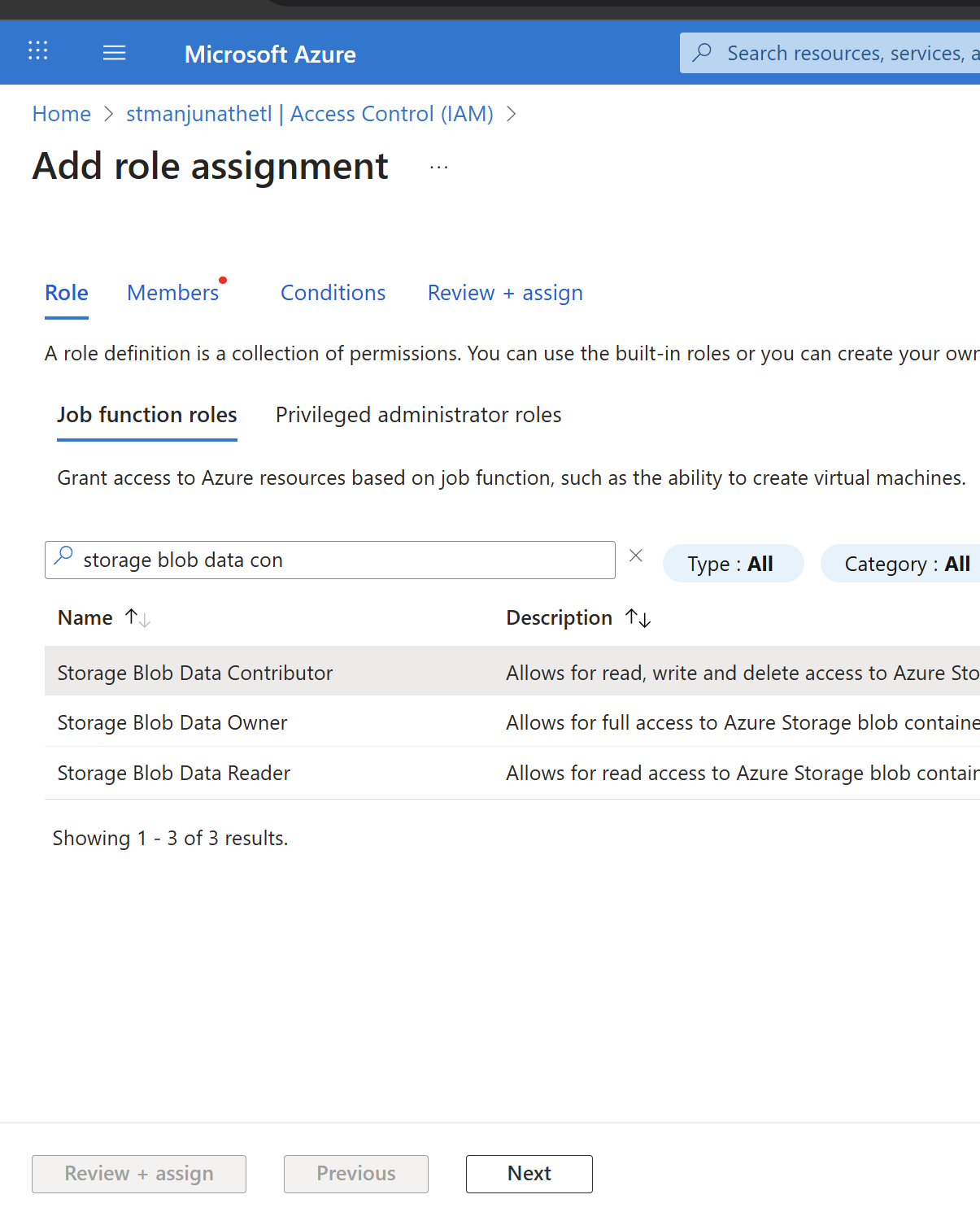
* Create a resource group in your azure portal for the ETL project
* Using it create a storage account and a container inside named ‘diabetes’

****

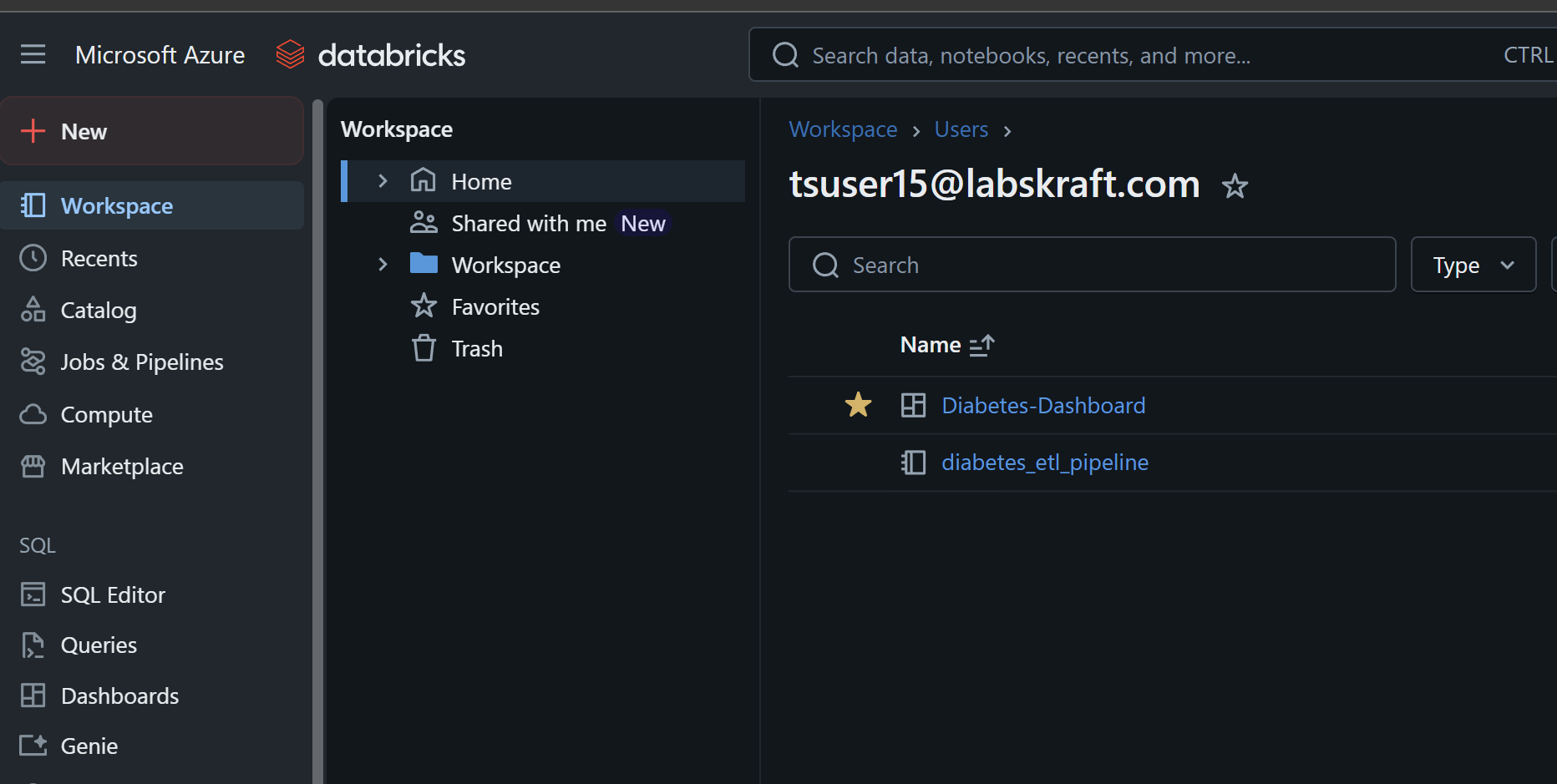
* Create an ACCESS CONNECTOR for Azure Databricks in the same region

****

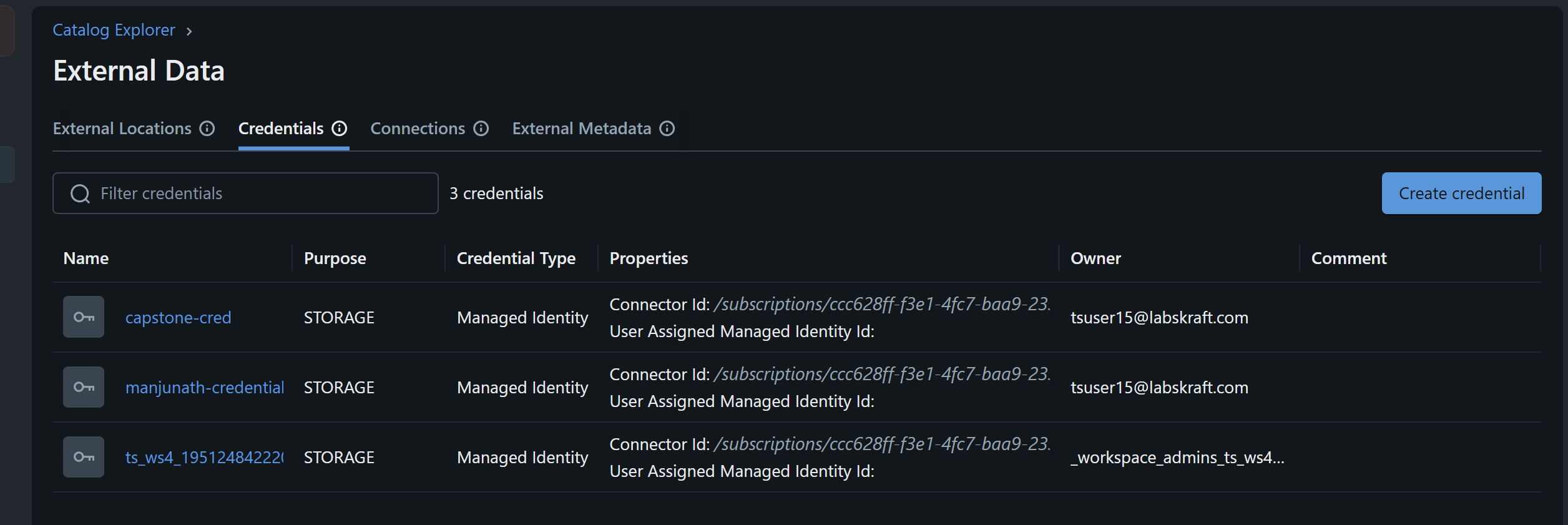
* Assign ‘Storage Blob Data Contributor’ role to the container for the access connector

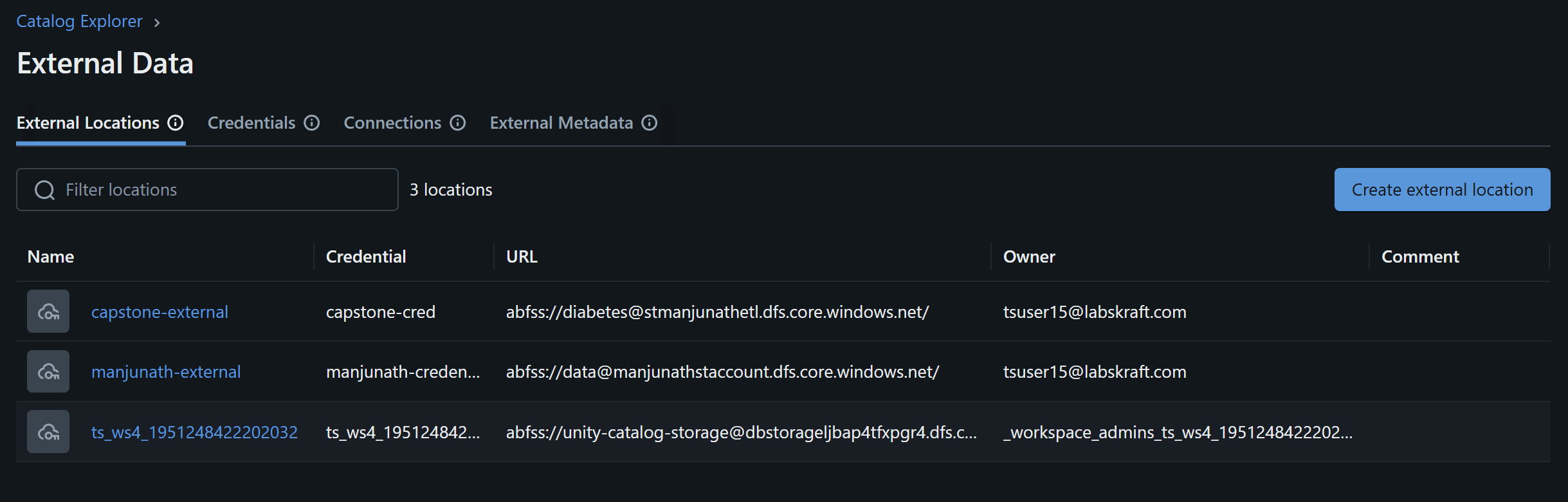
****

* Create Azure Databricks workspace under the same resource group

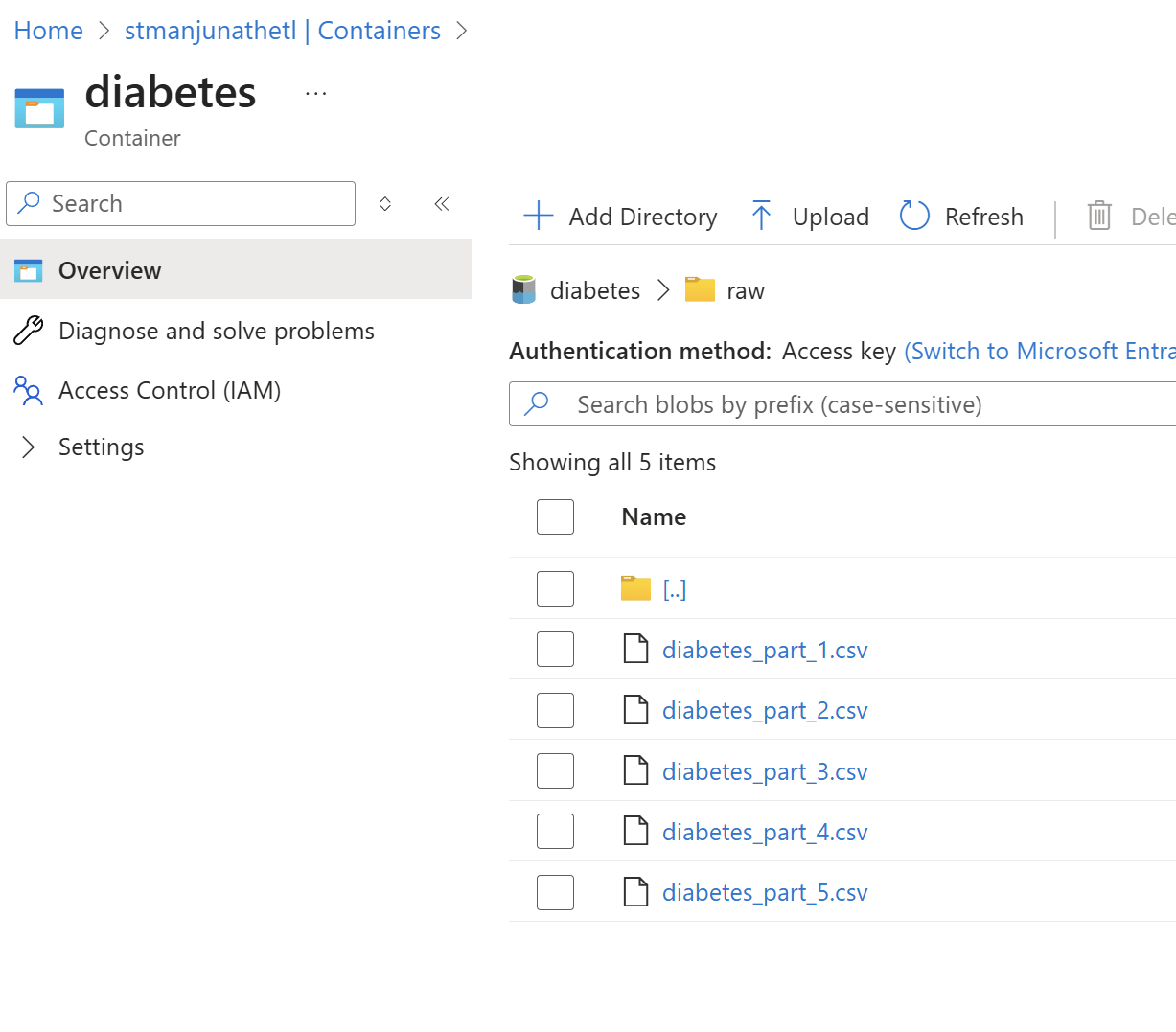
****

* Setup Credentials and External location to access data from the ADLS gen2 container with /raw folder

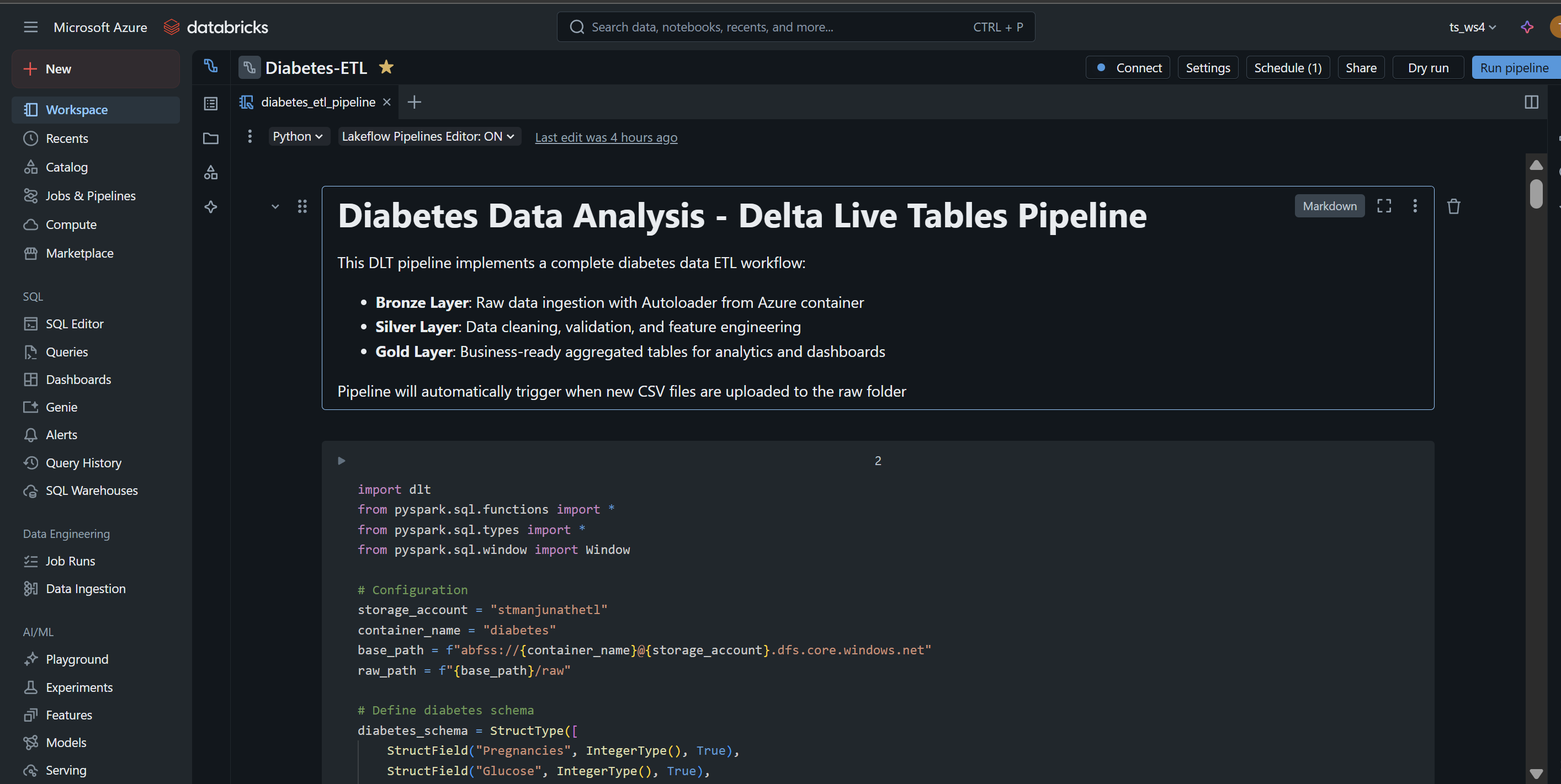
****



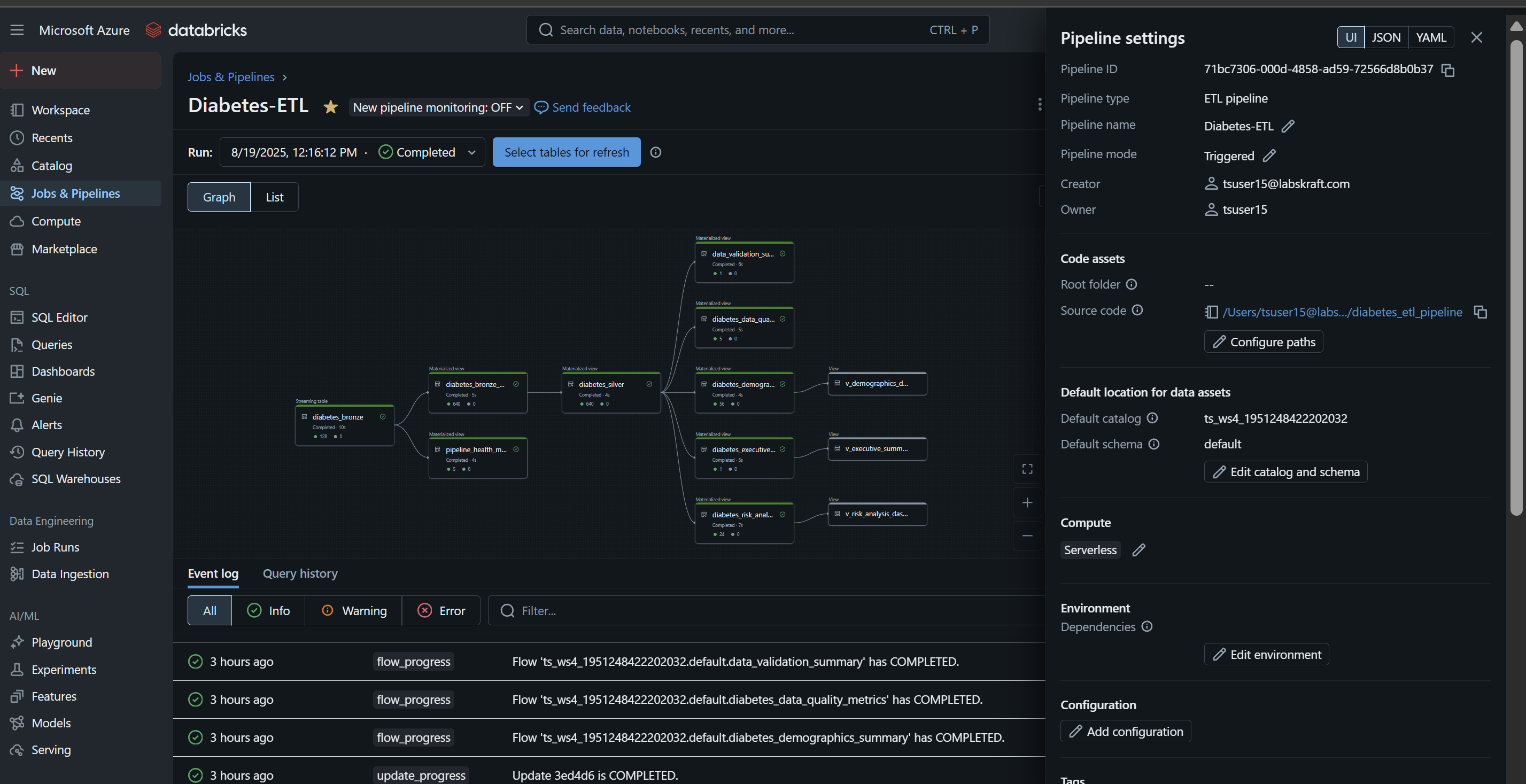
* Upload the csv files to this /raw directory under the diabetes container, from which the pipeline will read and ingest the data

****

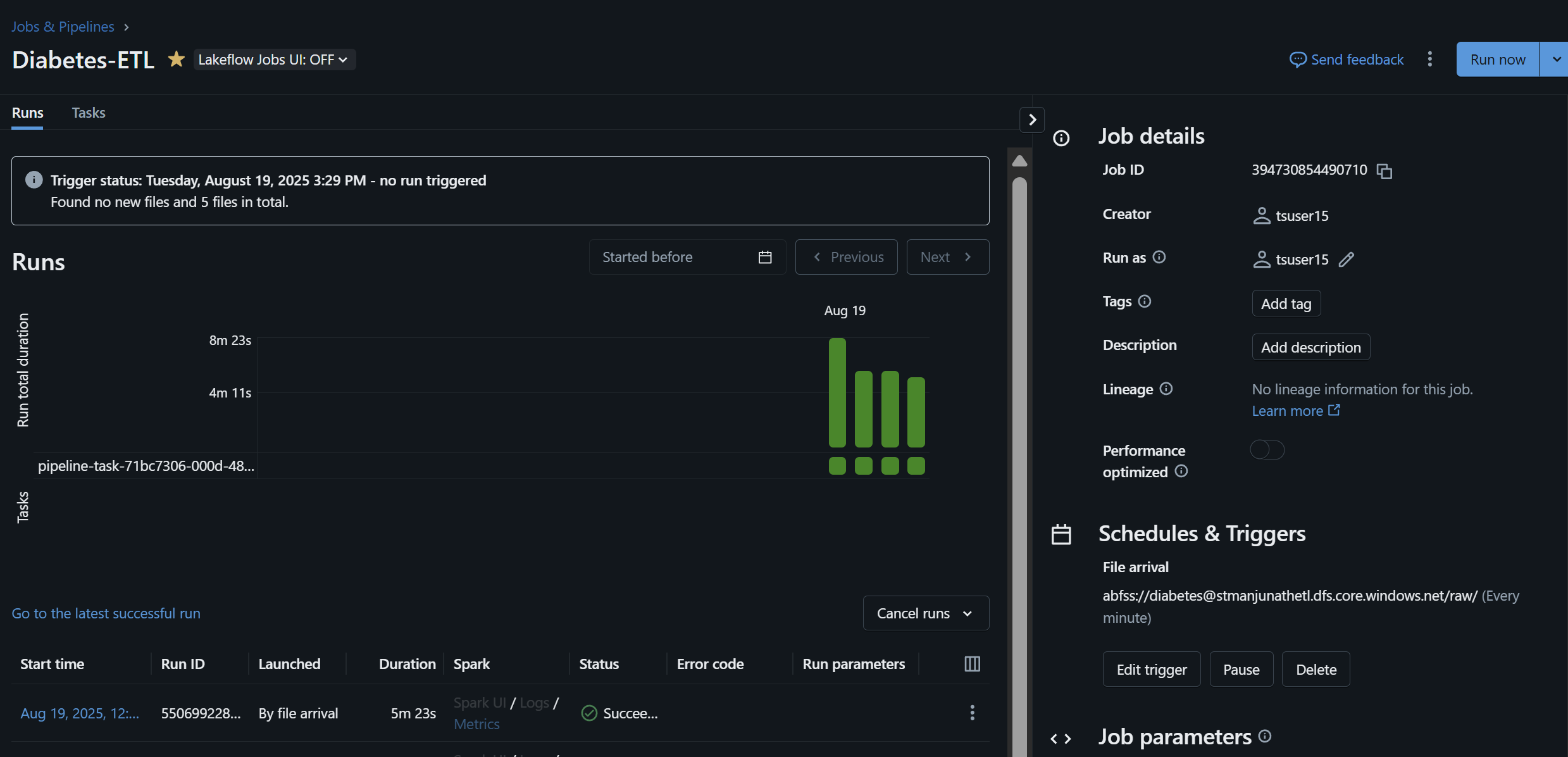
* In the DBX workspace, create a notebook for the pipeline and fill it with code.

****

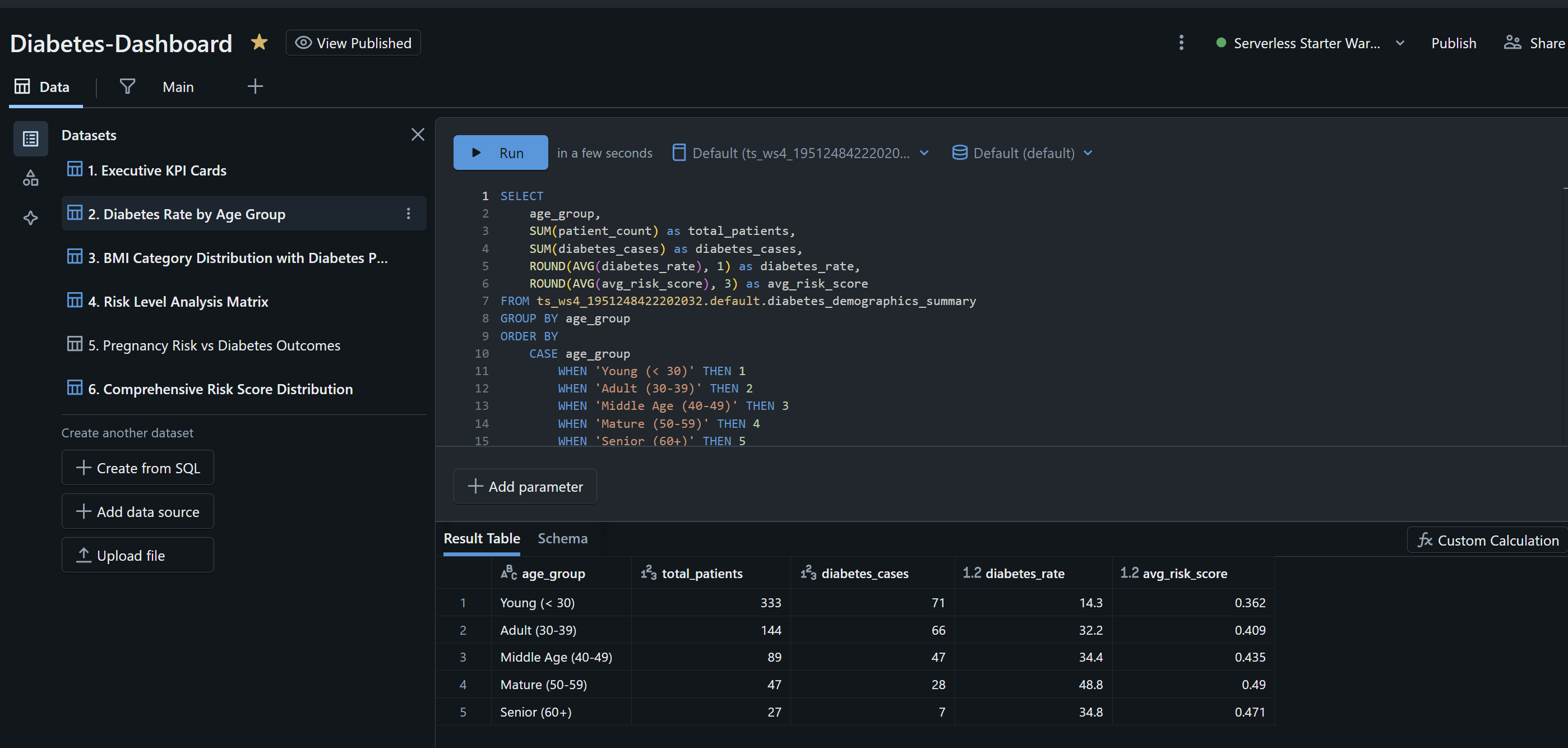
* Then setup an ETL pipeline that uses the above notebook with Serverless Compute

****

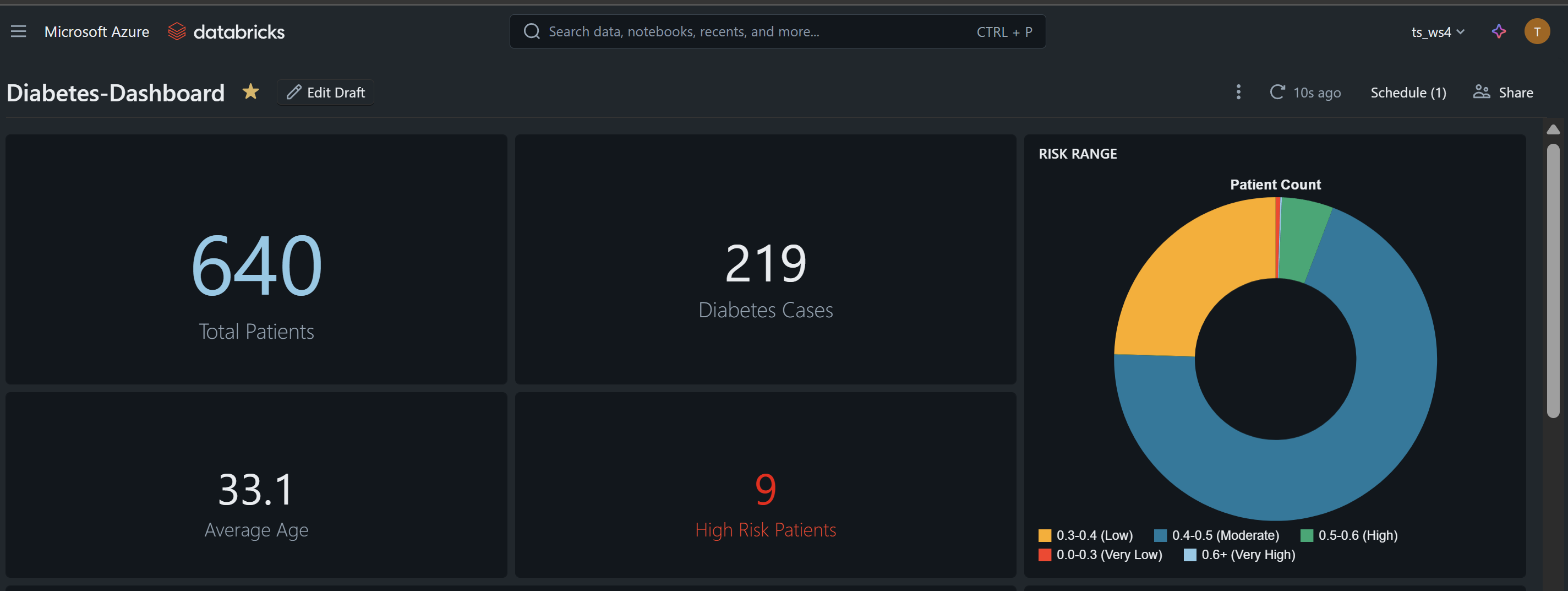
* For this ETL pipeline, orchestrate a job that has file-based trigger wrt the /raw folder

****

* After running the pipeline, build visualization using Databricks Dashboards feature using SQL queries
* Use SQL queries to get the relevant views for the dashboards

****

**Final Dashboard**

****

****