Custom Training

Day 20

**Databricks**

Structured Streaming: Modes:

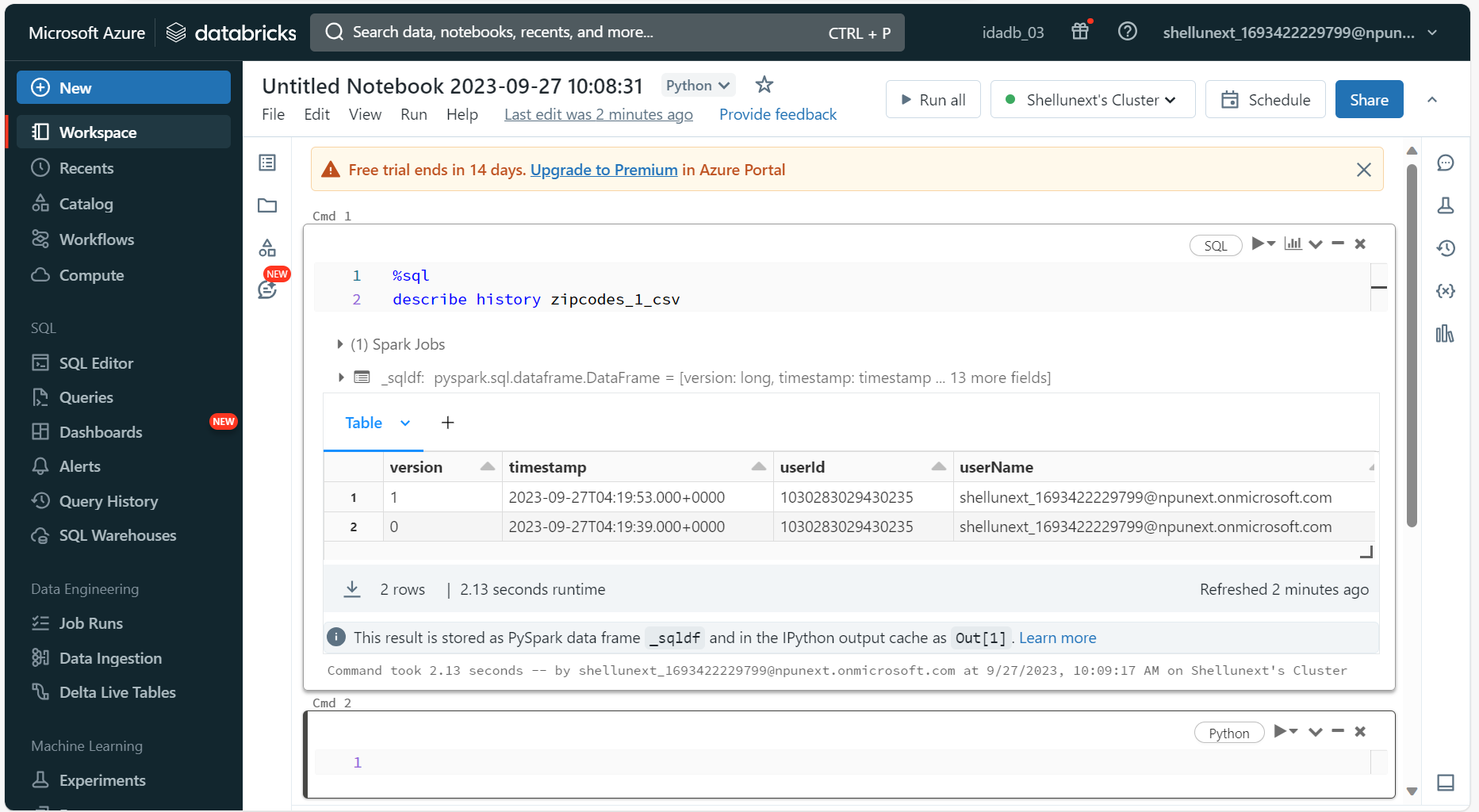
1. Append: Optional when not using aggregation.
2. Complete: Mandatory in case of aggregation.
3. Update: On the go will check for the updation if required it will do.

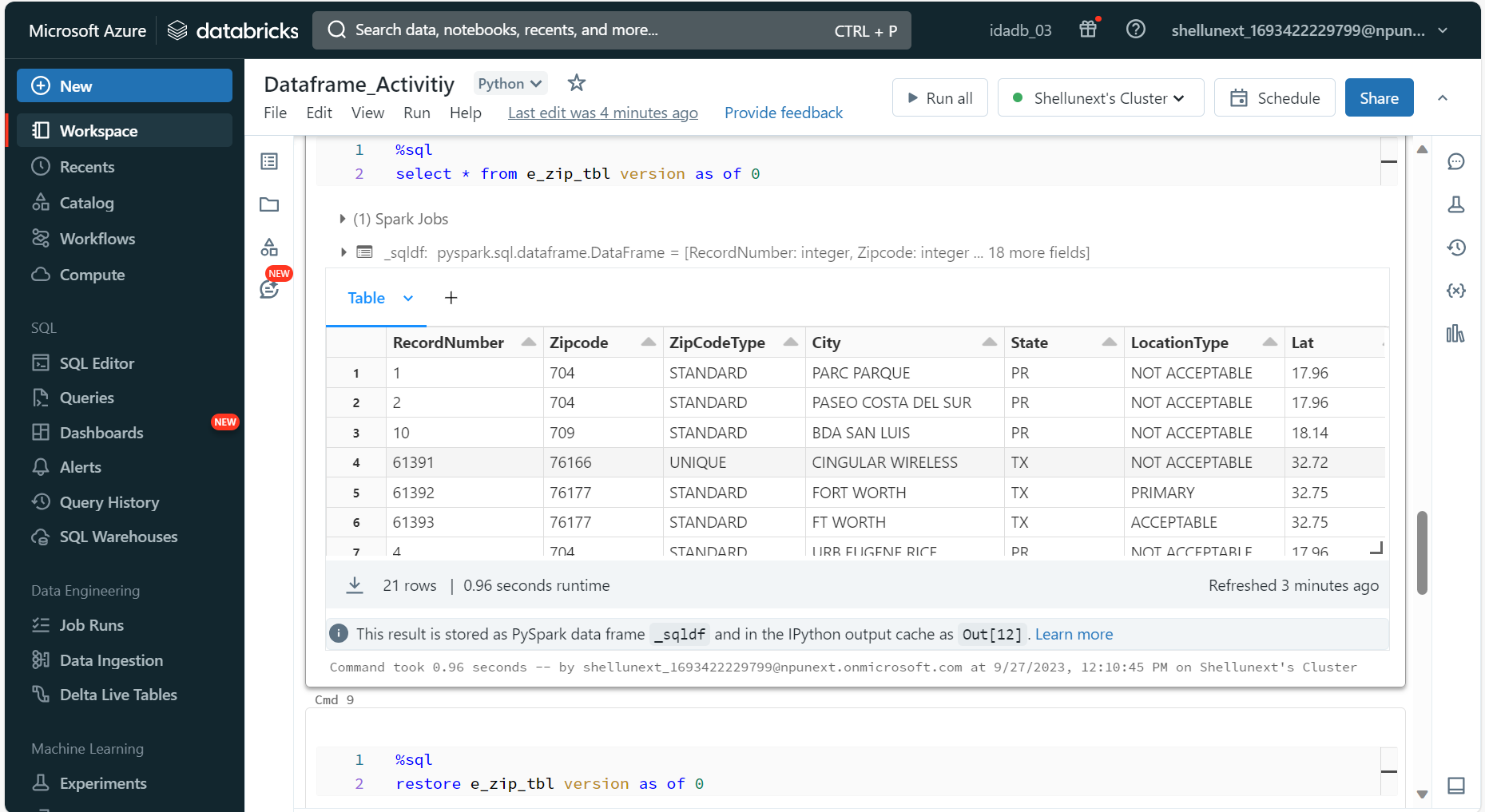
This involves more cost than ADF.

Parquet is columnar format which goes well with big data. But it doesn’t support deletion operation.

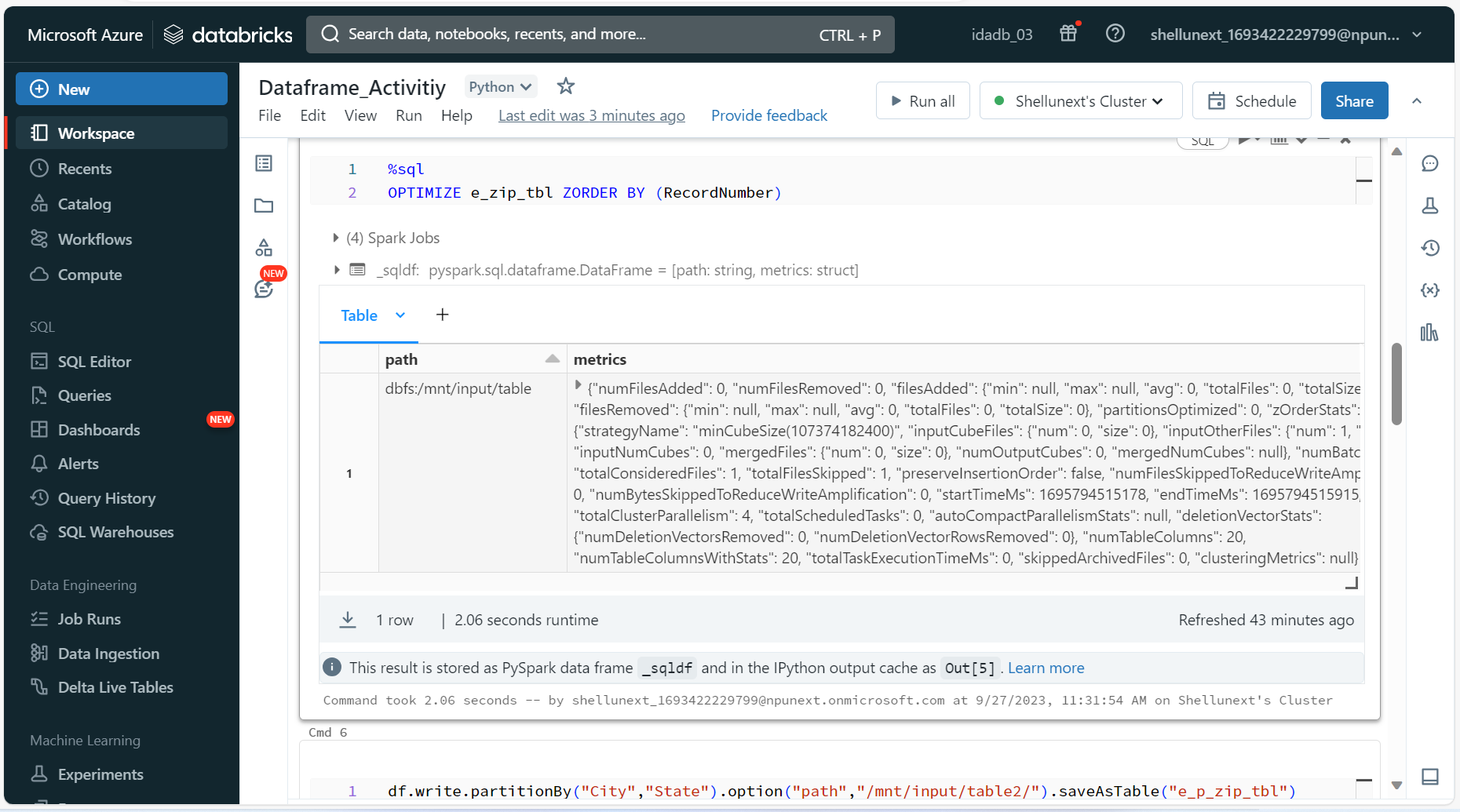
Delta Table: Whenever we create a table in databricks, delta table is created.

1. We can do time-travelling in delta table. We can either mimic the table for that time period or rollback to that particular version. Inside the delta.log folder, the json file will store these versions. Vacuum command is used to clean the versions in periodic intervals.





1. We optimize the partitions so that each partitioned file consists of the same no. of records.

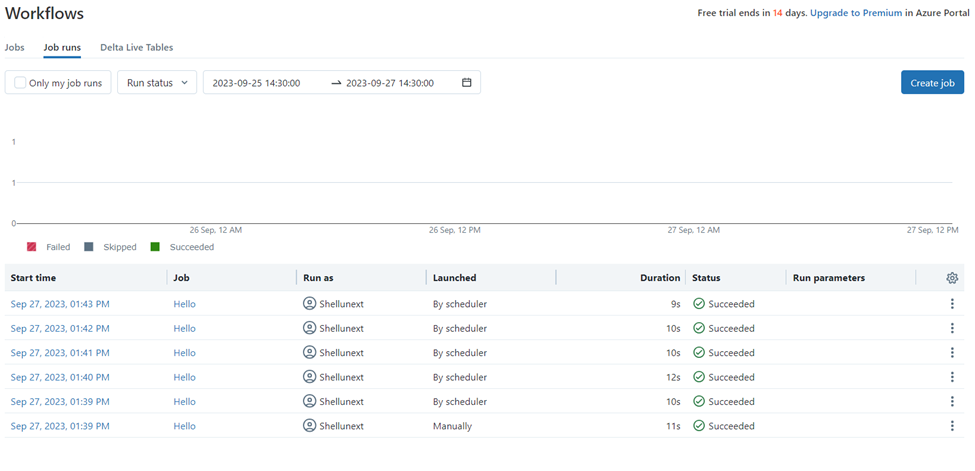


Medallion/ Multi Hop Architecture

We follow three folder architecture:

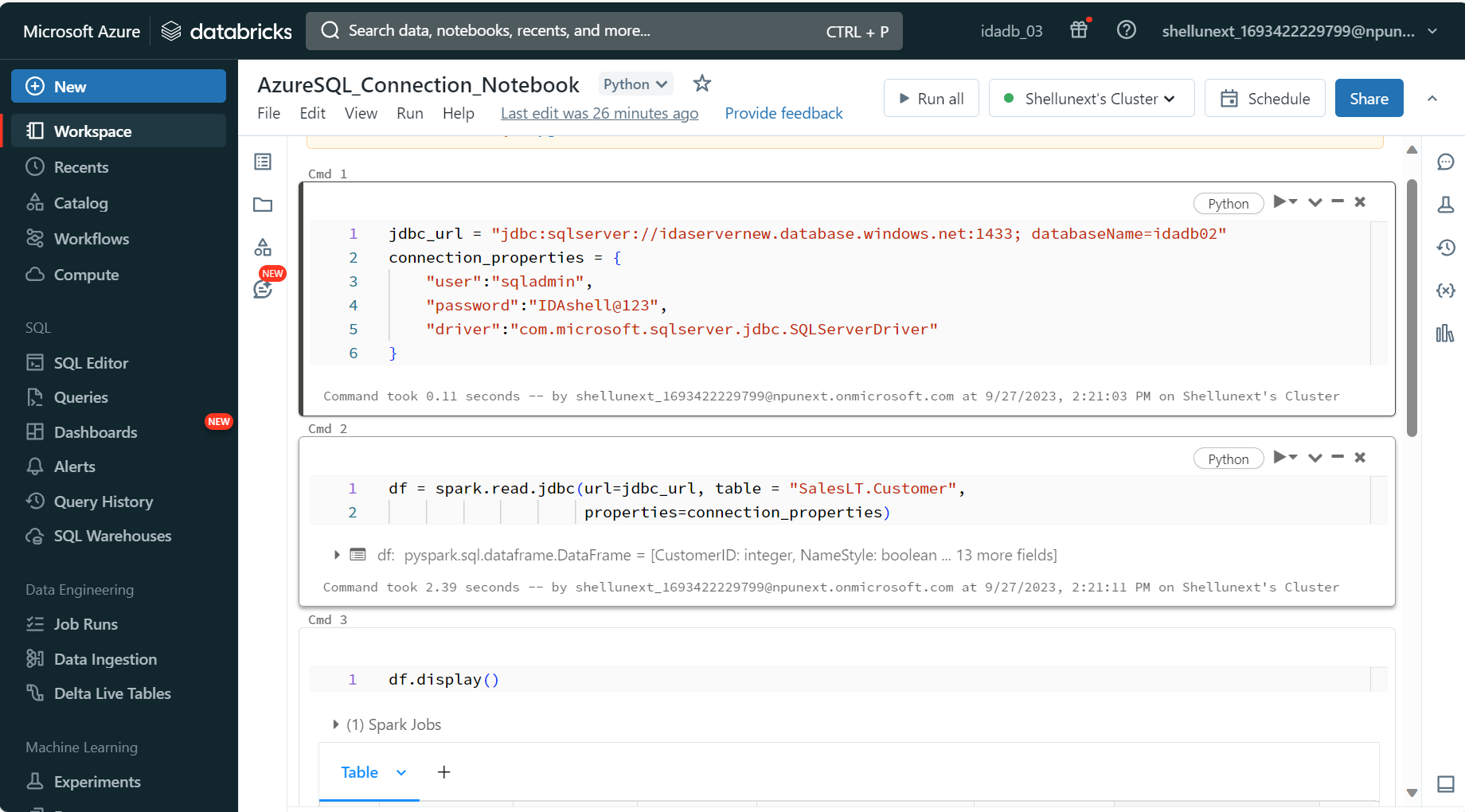
1. Raw/Bronze: All raw data from all the sources are kept in this folder.
2. Silver: Cleansing and formatting the data files into one particular file format.
3. Gold: Performing some aggregations and joins and keeping the data files needed according to the requirements.

Schedule Jobs:



To run our notebooks on a scheduled basis we schedule jobs.

Unity Catalog:



* For the same workspace, there can be multiple users that work on different notebooks.
* For development, production, UAT env there are different workspaces.
* To share notebooks present under different users in different workspaces, the admin must provide access separately to every user for any file present in different workspaces.
* This is a very inefficient and time-consuming process.
* With unity catalog, workspace sharing can be enabled.
* This provides data access without needing admin permission.
* Azure Access Connector is used to provide connection to different workspace files.