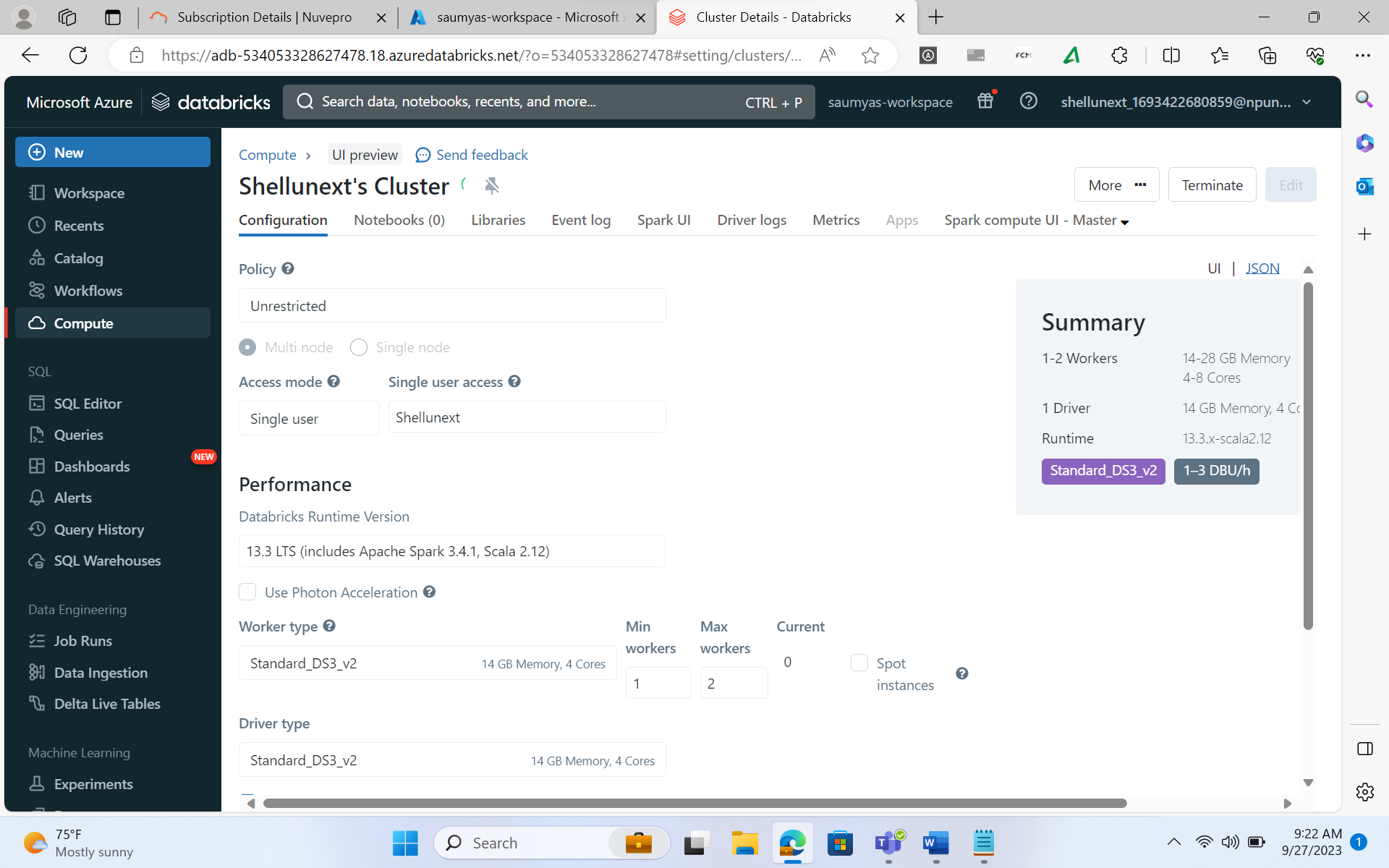
**Custom Training – September 27, 2023**

**Databricks Day 02**

1. **Creation of Databricks workspace & cluster**

****

**Lakehouse vs Datalake**

* **Datalake + Data Warehouse = Lakehouse** (stores data in delta format, advanced version of parquet)
* Put all data in Lakehouse in delta format.
* **Datalake has some limitations:**
  + No ACID transactions
  + Structured, semi structured, unstructured data is supported.
  + No capability to handle streaming data.
* **Datawarehouse:**
  + Very good acid transactions
  + Semi structured & unstructured data cannot be stored here.
  + No capability to handle streaming data.

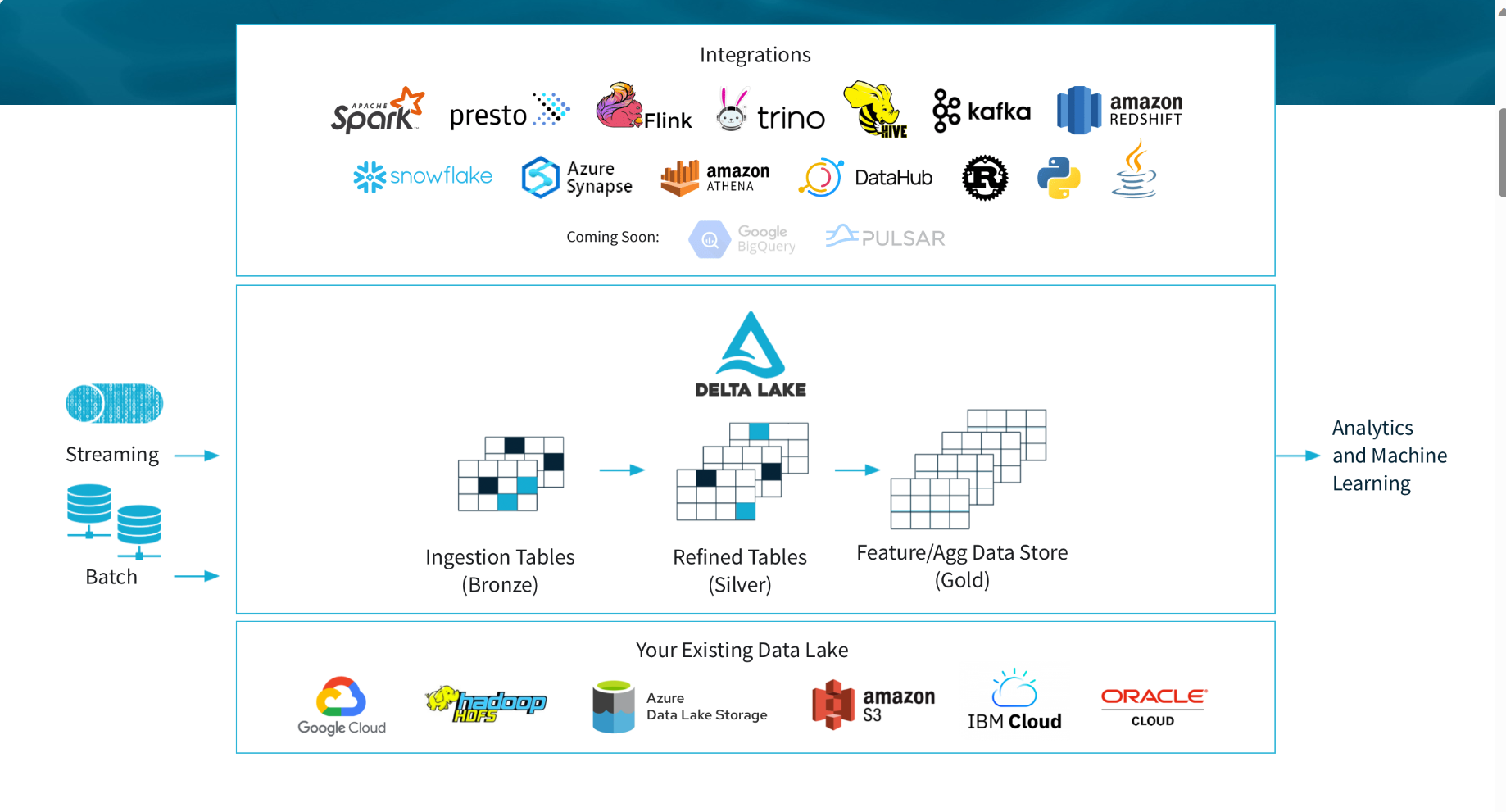
**Medallion/Multihop Architecture:**

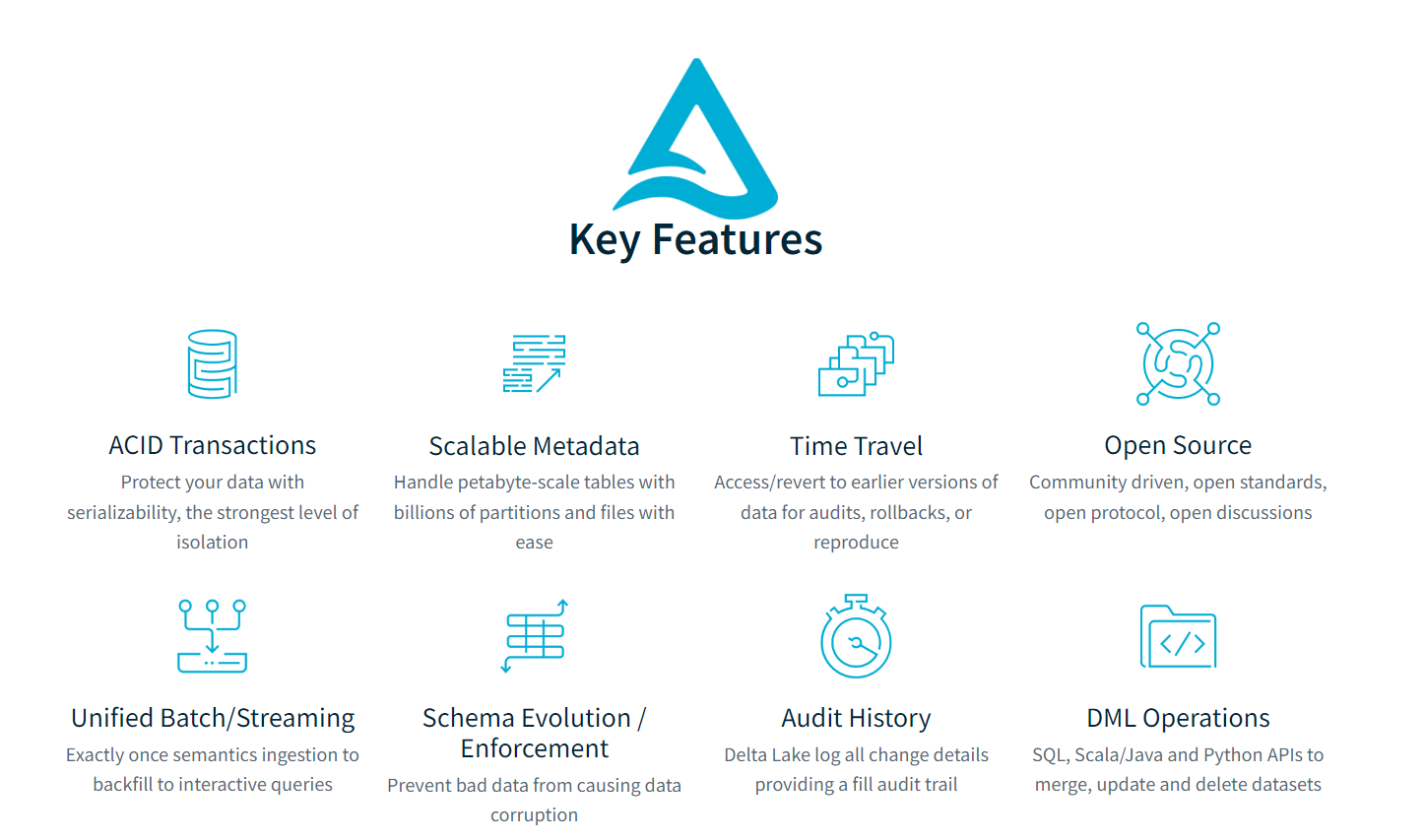
[What is the medallion lakehouse architecture? - Azure Databricks | Microsoft Learn](https://learn.microsoft.com/en-us/azure/databricks/lakehouse/medallion)

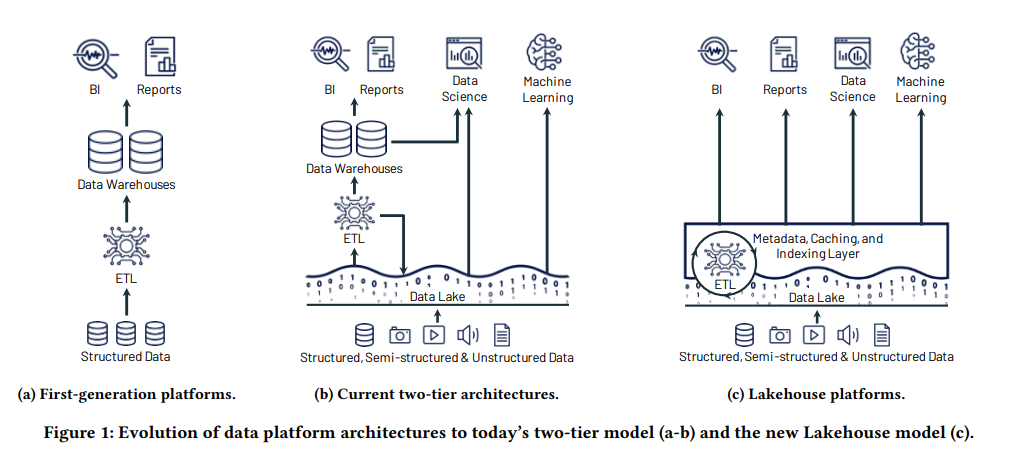
**Bronze, silver, and gold architecture:** Azure Databricks works well with a [medallion architecture](https://techcommunity.microsoft.com/t5/analytics-on-azure/how-to-reduce-infrastructure-costs-by-up-to-80-with-azure/ba-p/1820280) that organizes data into layers:

* Bronze: Holds raw data.
* Silver: Contains cleaned, filtered data.
* Gold: Stores aggregated data that's useful for business analytics.

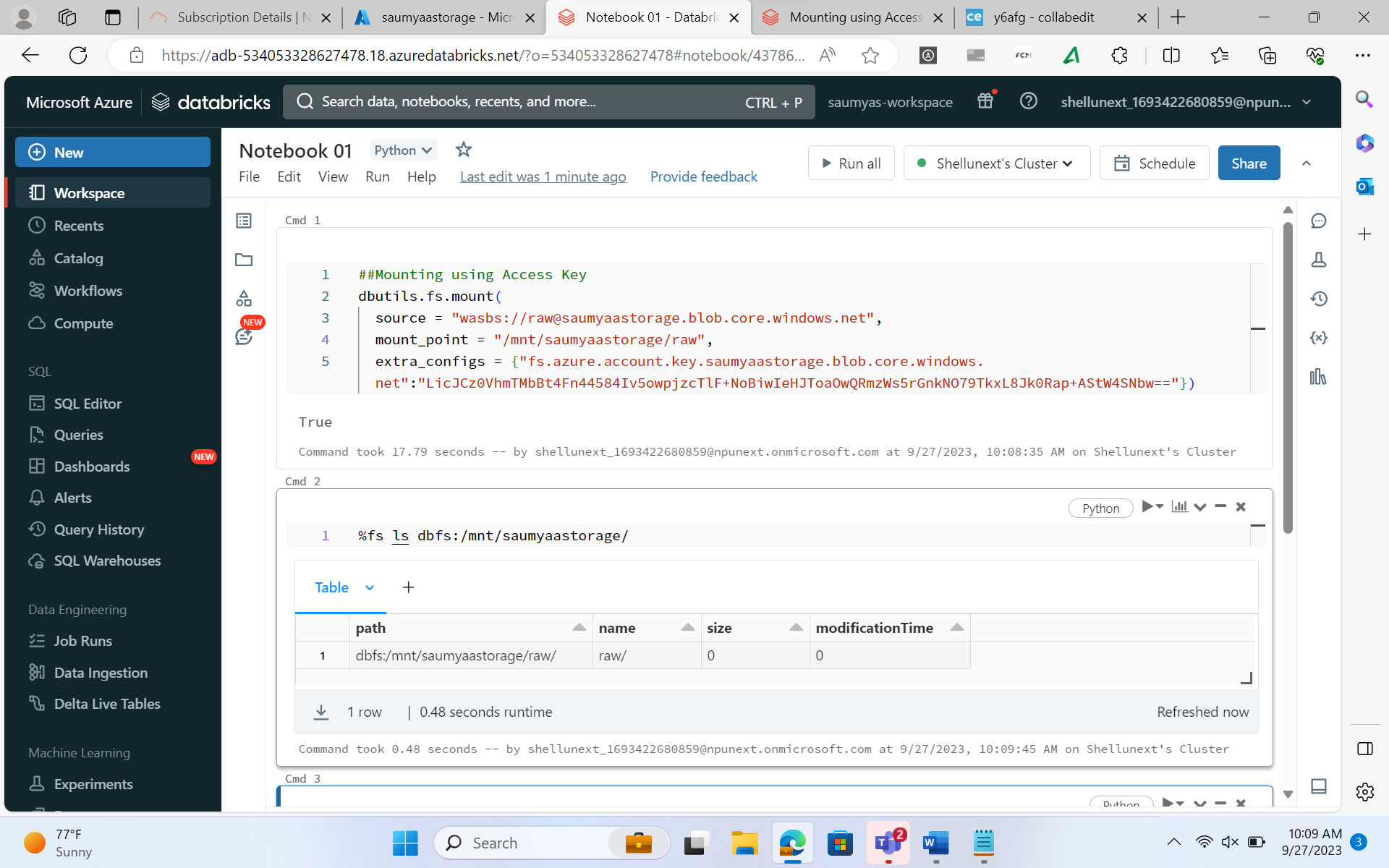
[Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics (cidrdb.org)](https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf)

****

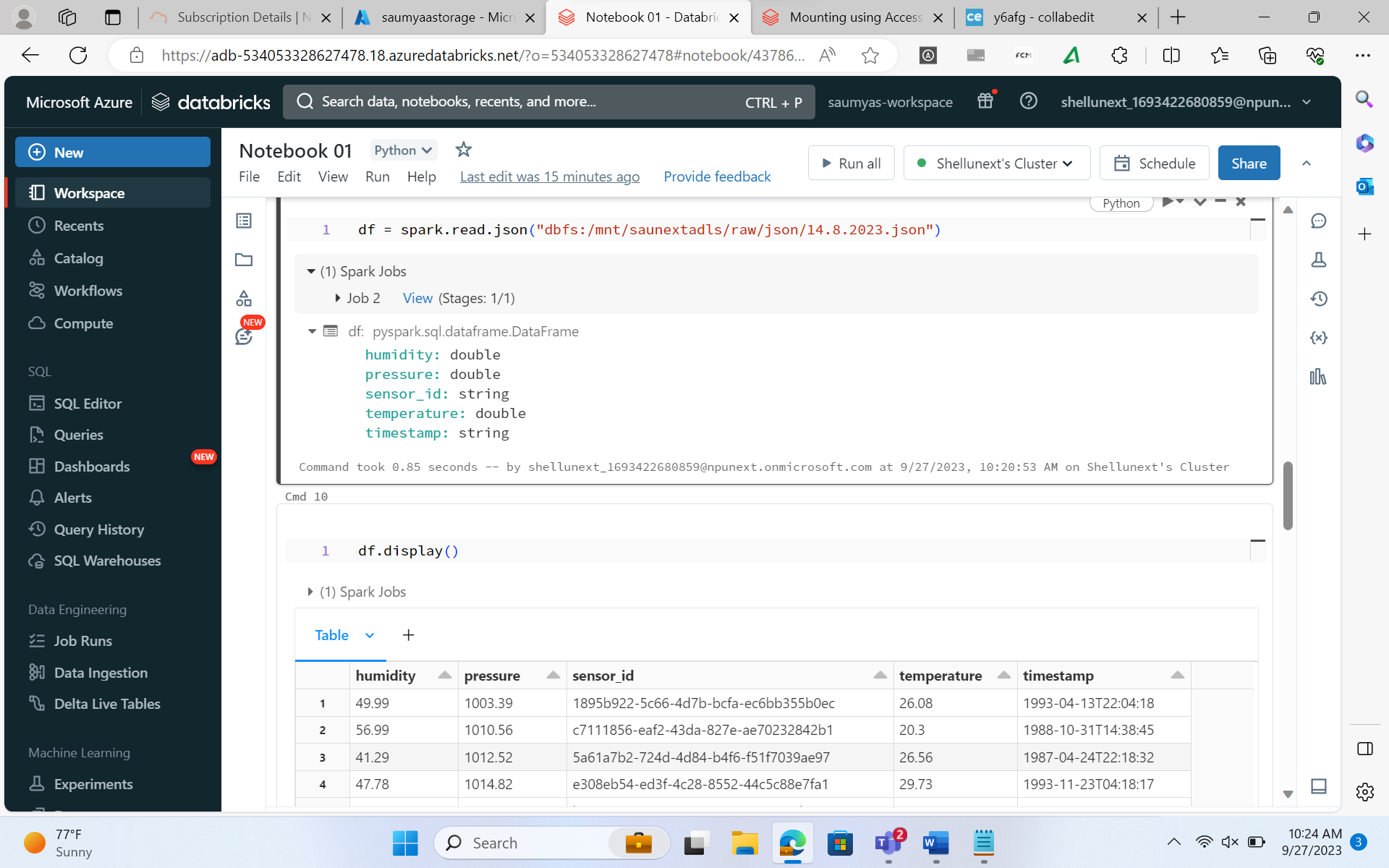
****

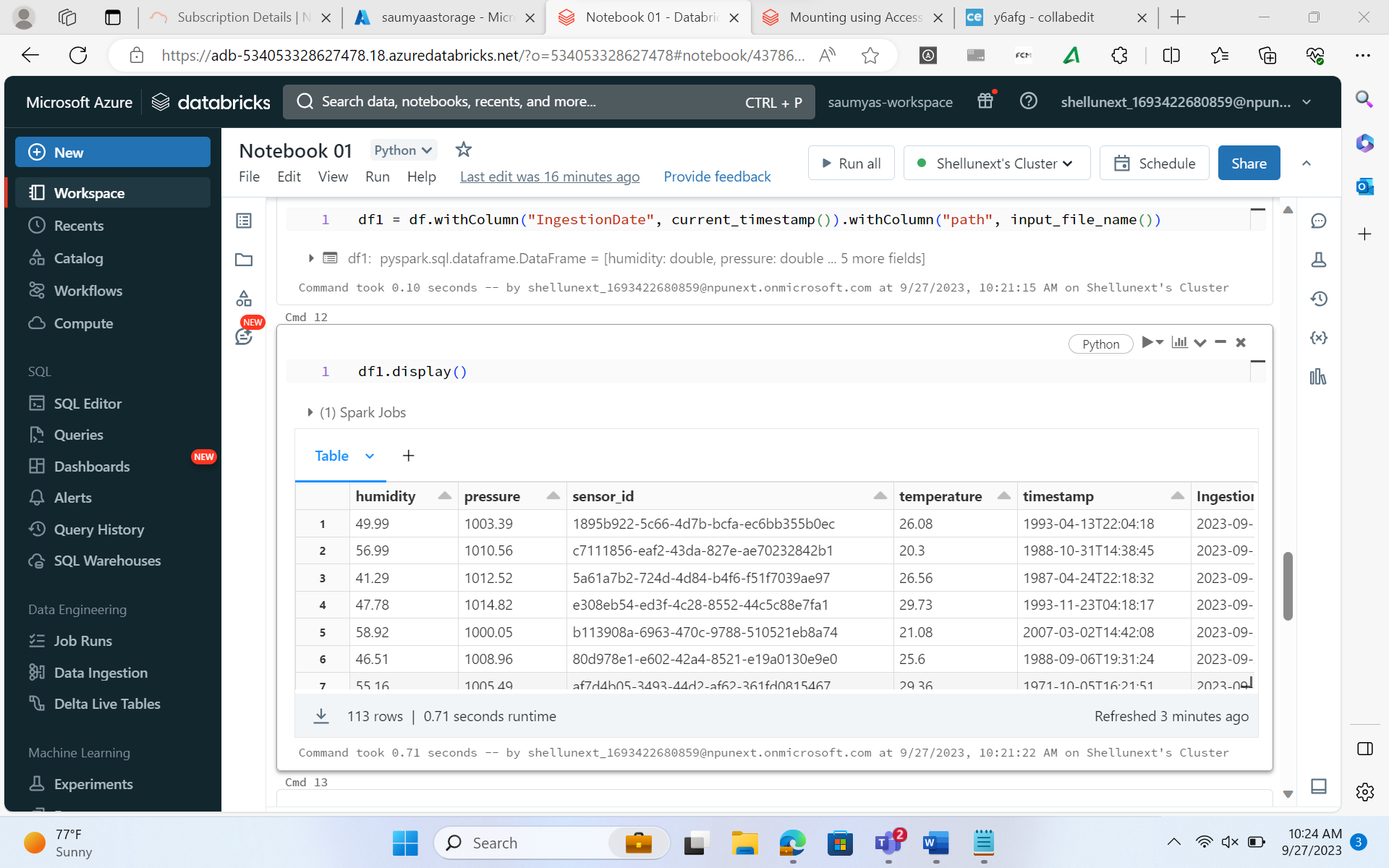
****

1. **Mounting Storage Account using Storage Account**

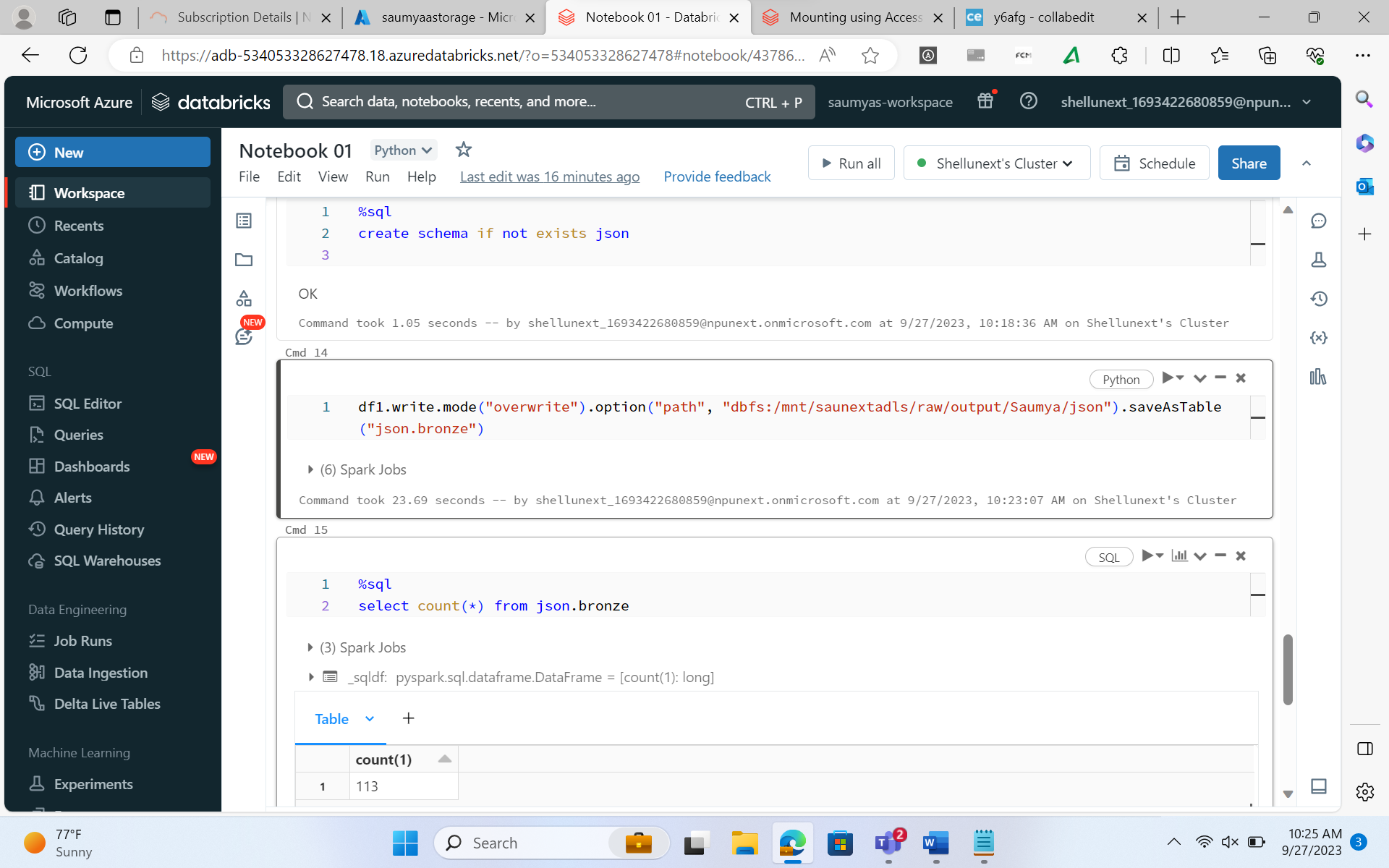
****

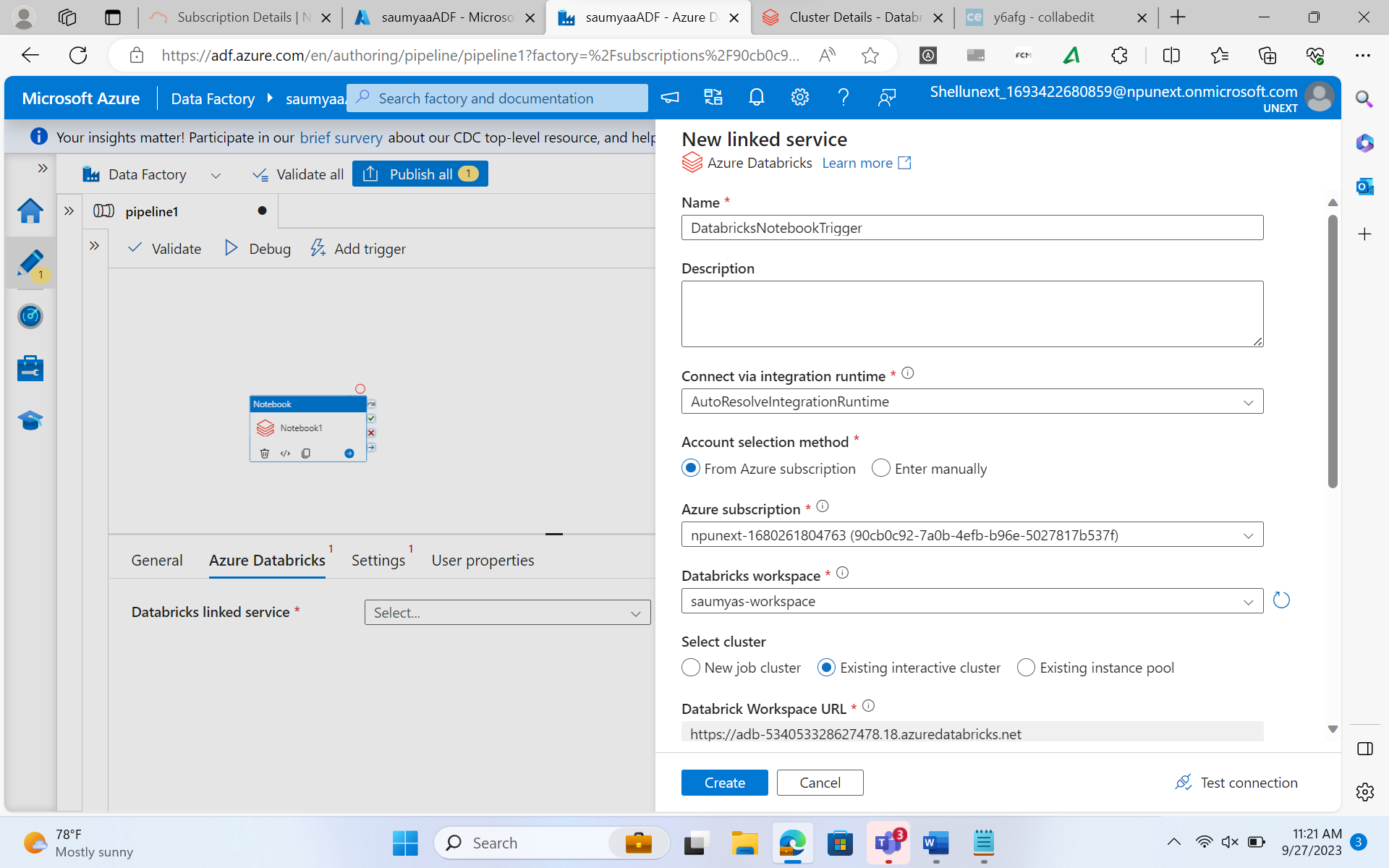
1. **Creating a DataFrame**

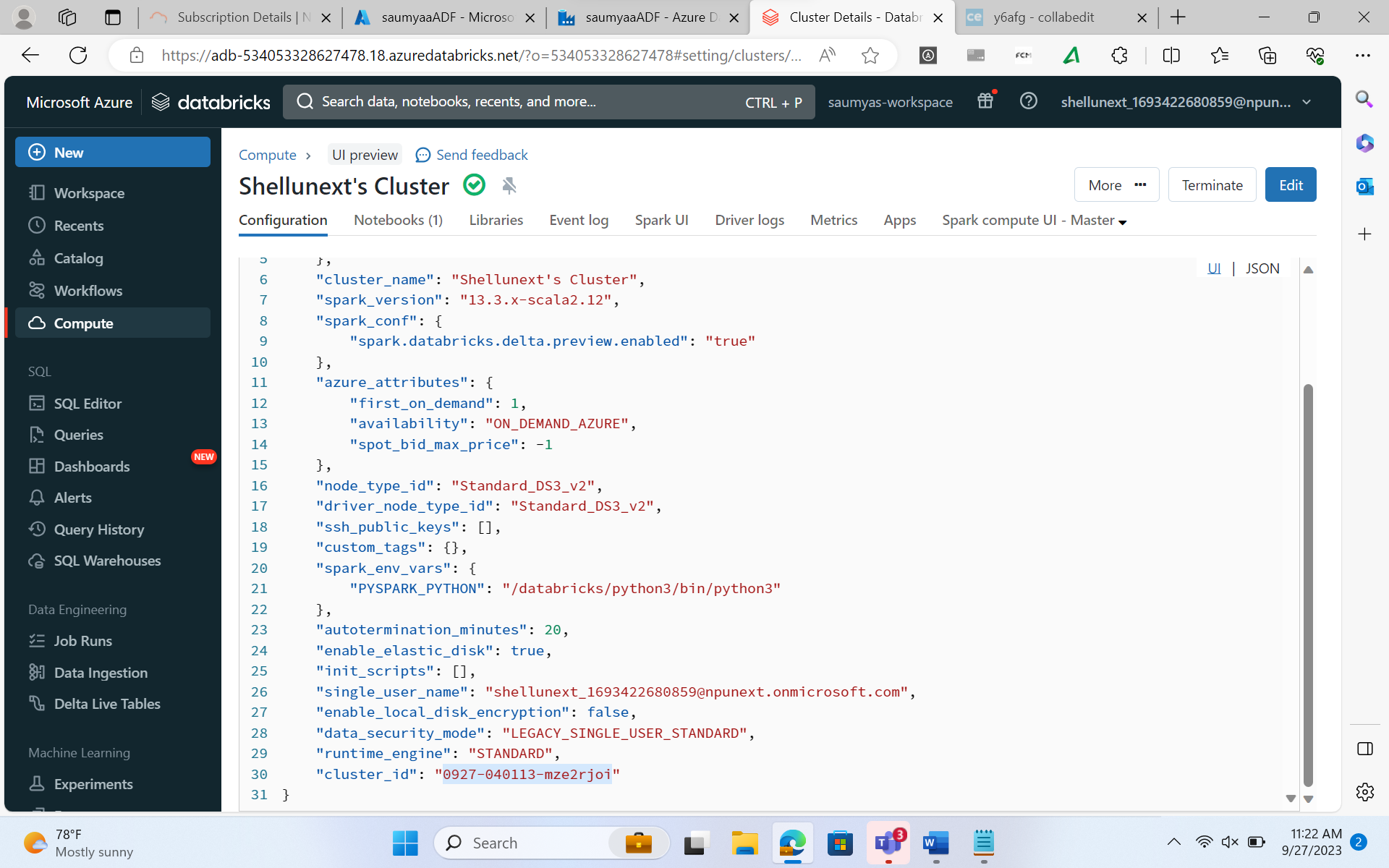
****

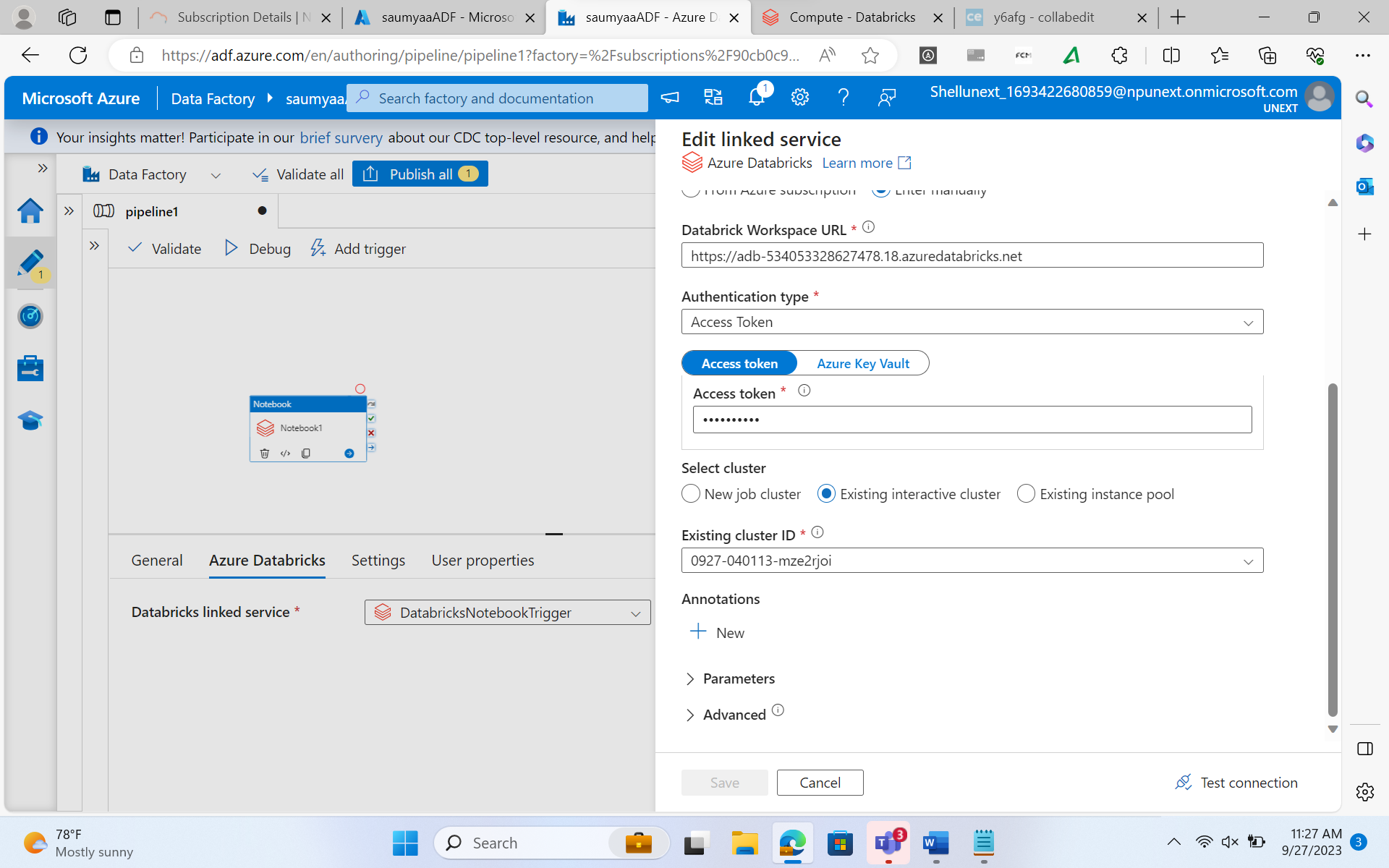
****

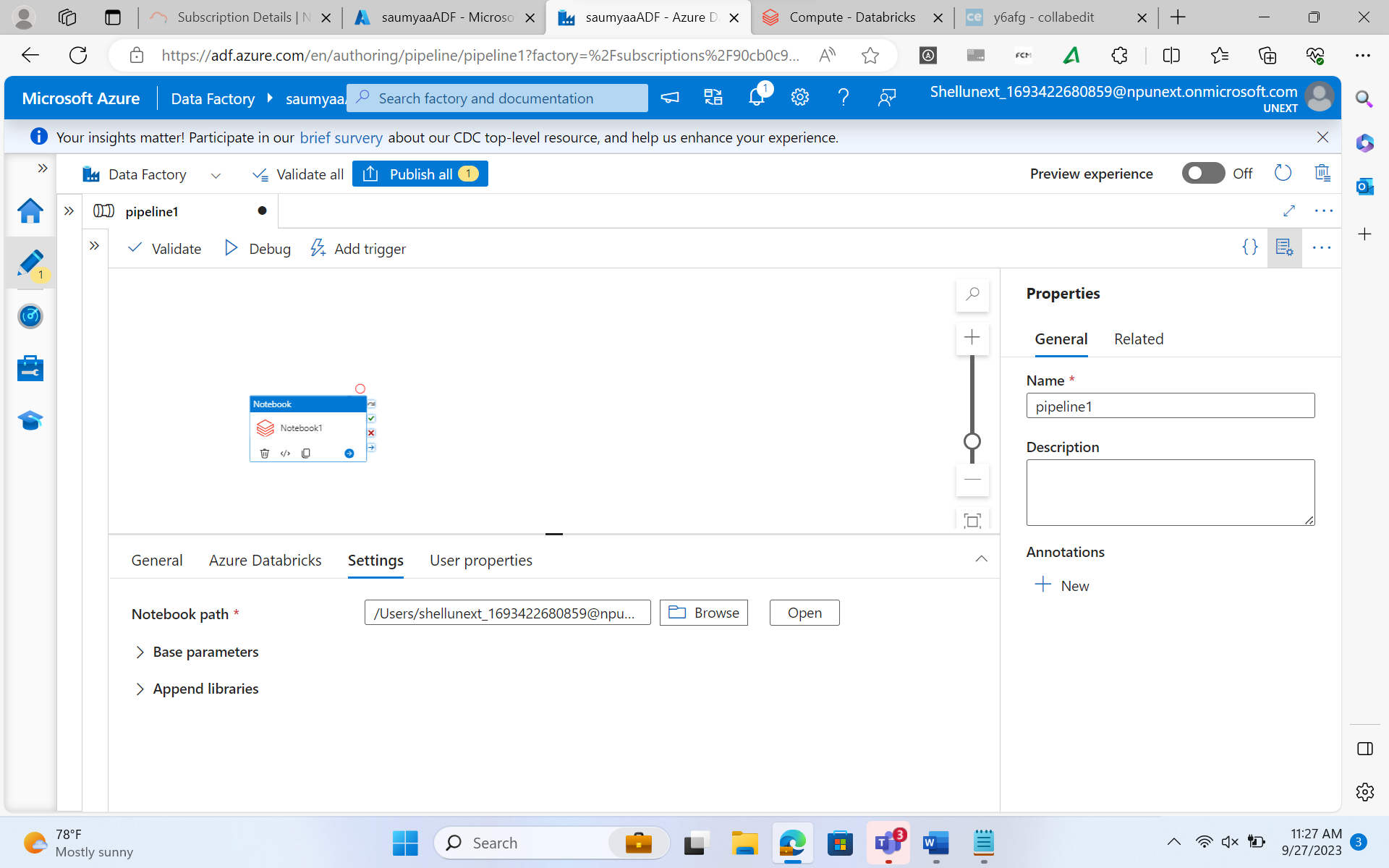
1. **Creating table from DataFrame**

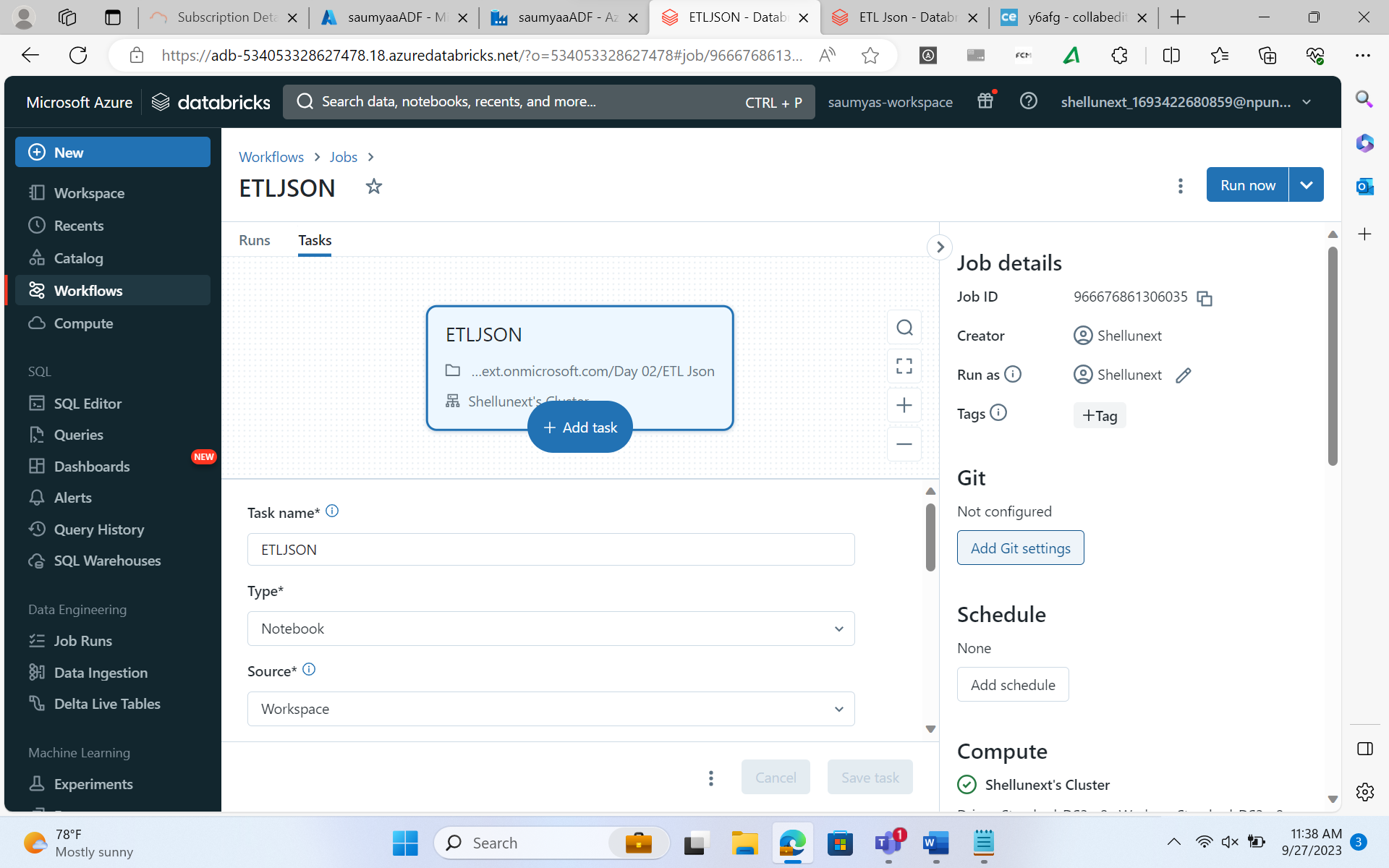
****

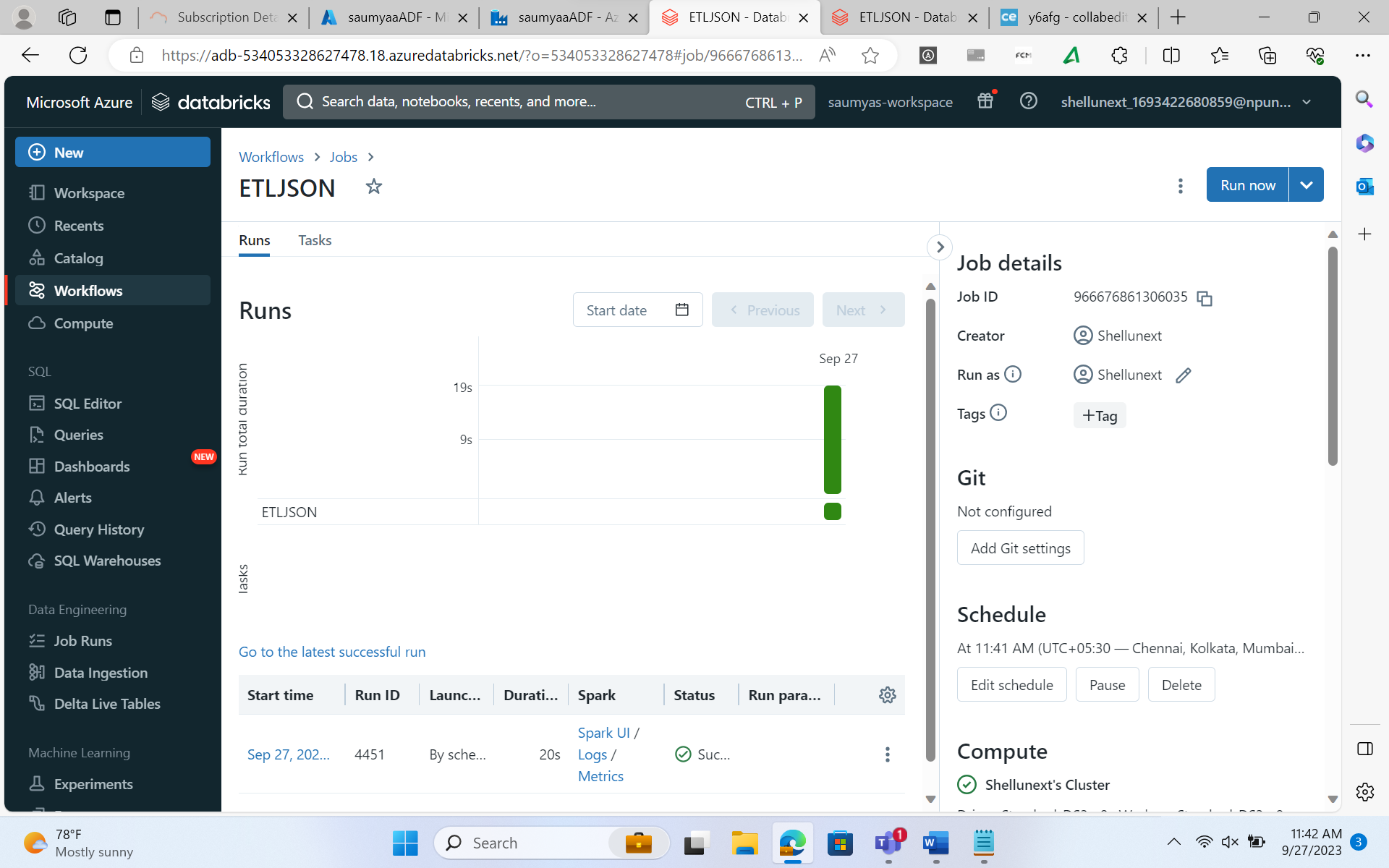
****

****

****

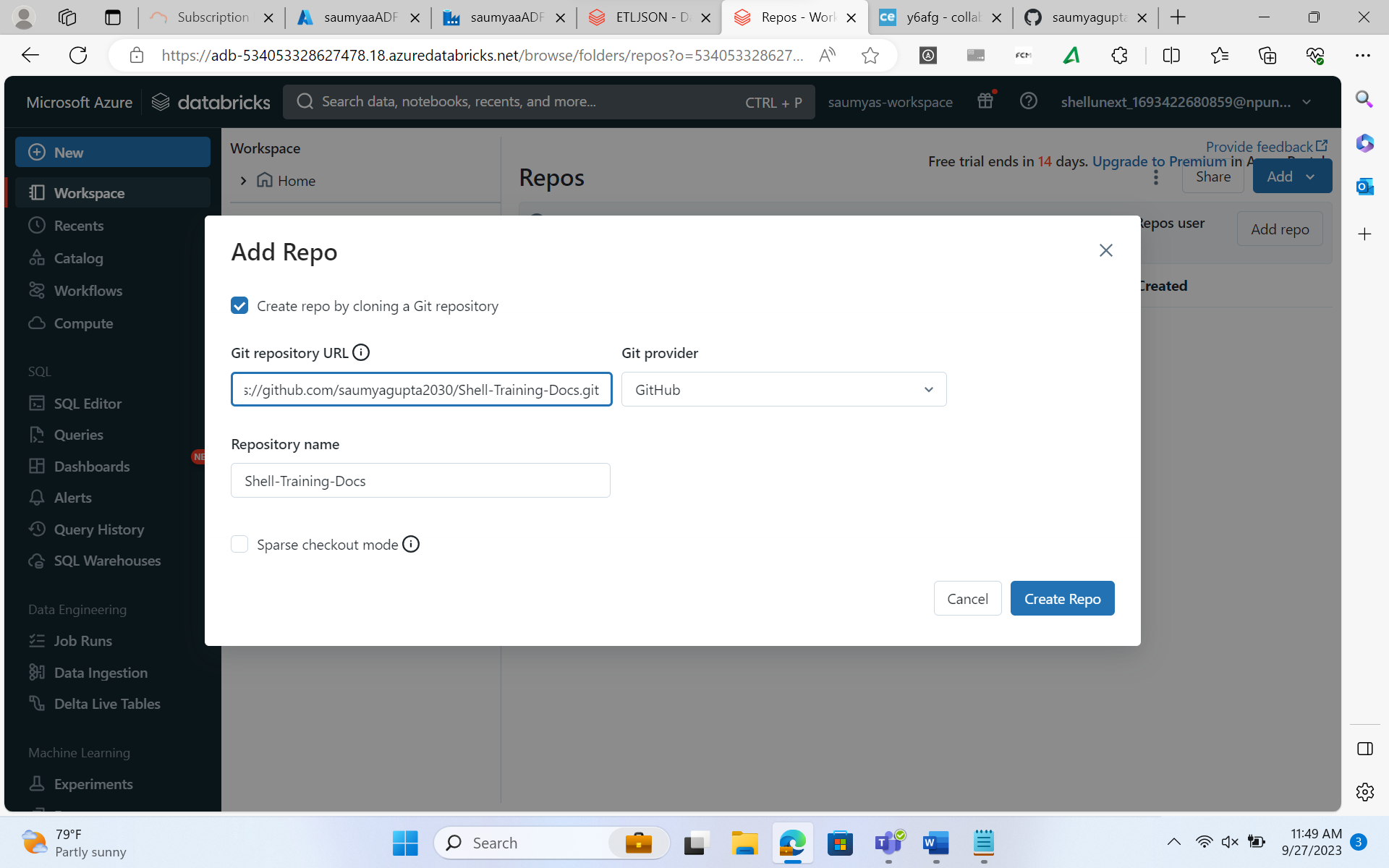
****

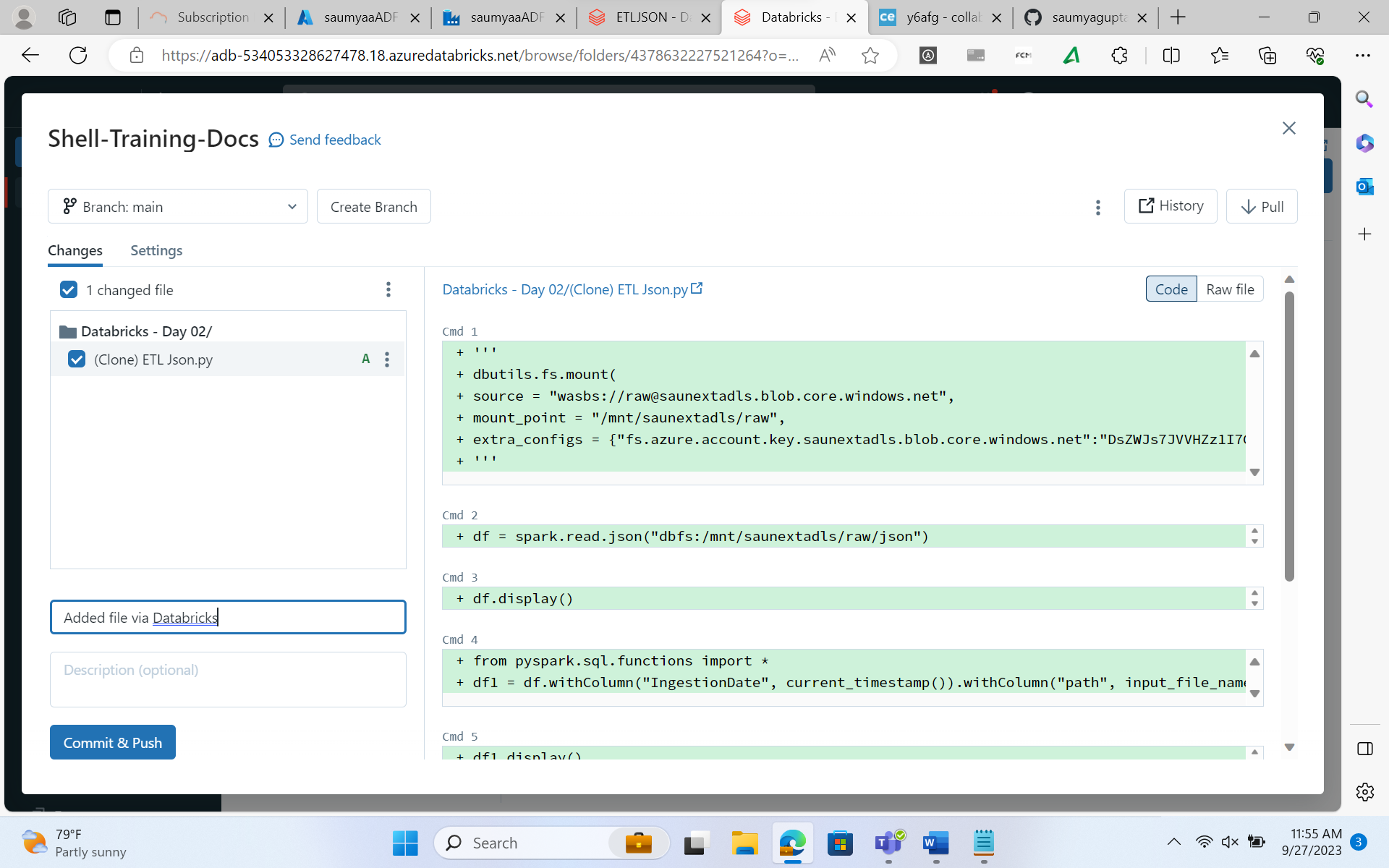
****

****

**Run databricks notebook in datafactory.**

**Export the notebook to github using workflows in datafactory.**

****

****

**Donot go to commit**

**Go to user settings**

* saveAsTable saves everything in a delta format by default, only in databricks. History is only available for delta tables in databricks, as delta\_logs in file types crc & json. CRC = cyclic redundant check. With these json files we can do time travel.