Magic Commands

======================

%run

* to import a notebook into another notebook. Use %run ./<notebook\_path>

%md

* markdown is used for heading and general statements.
* Used as %md

#Markdown

%lang

* language is used to change the programming language for a cell

Note: try to keep less than 100 cells in a single notebook

Structured Streaming

====================

Use of checkpointing during writeStream

* checkpoint
  + it stores the data and metadata
  + it helps restore the original state if the action process fails midway

Output Modes

* append
  + adds new data to the same file and stores in the destination
* complete
  + used mostly for aggregation transformations
* update
  + it checks if the records have been changed or not
  + if they have, then it updates the latest values and stores in the destination

Delta Table

* the default provider while creating tables
* it provides timestamping features
  + it allows users to get back to the previous versions as all the changes are stores as different versions
  + either the table can rollback to the previous versions, or records can be fetched from the previous versions without affecting the current version
  + all these information are stored in a delta\_log folder
  + delta\_log folder stores data as json files. Crc file is also present inside delta\_log older
  + there is a vacuum feature that allows removal of json files that are more than 7 days old
* the default provider is parquet for tables
* if delta is mentioned, then the table created is a Hive table

%sql

VACUUM e\_zip\_tbl retain 169 Hours

%sql

OPTIMIZE e\_zip\_tbl ZORDER BY (RecordNumber)

For restore operation

%sql

restore idazip version as of 0

Medallion/Multi Hop Architecture

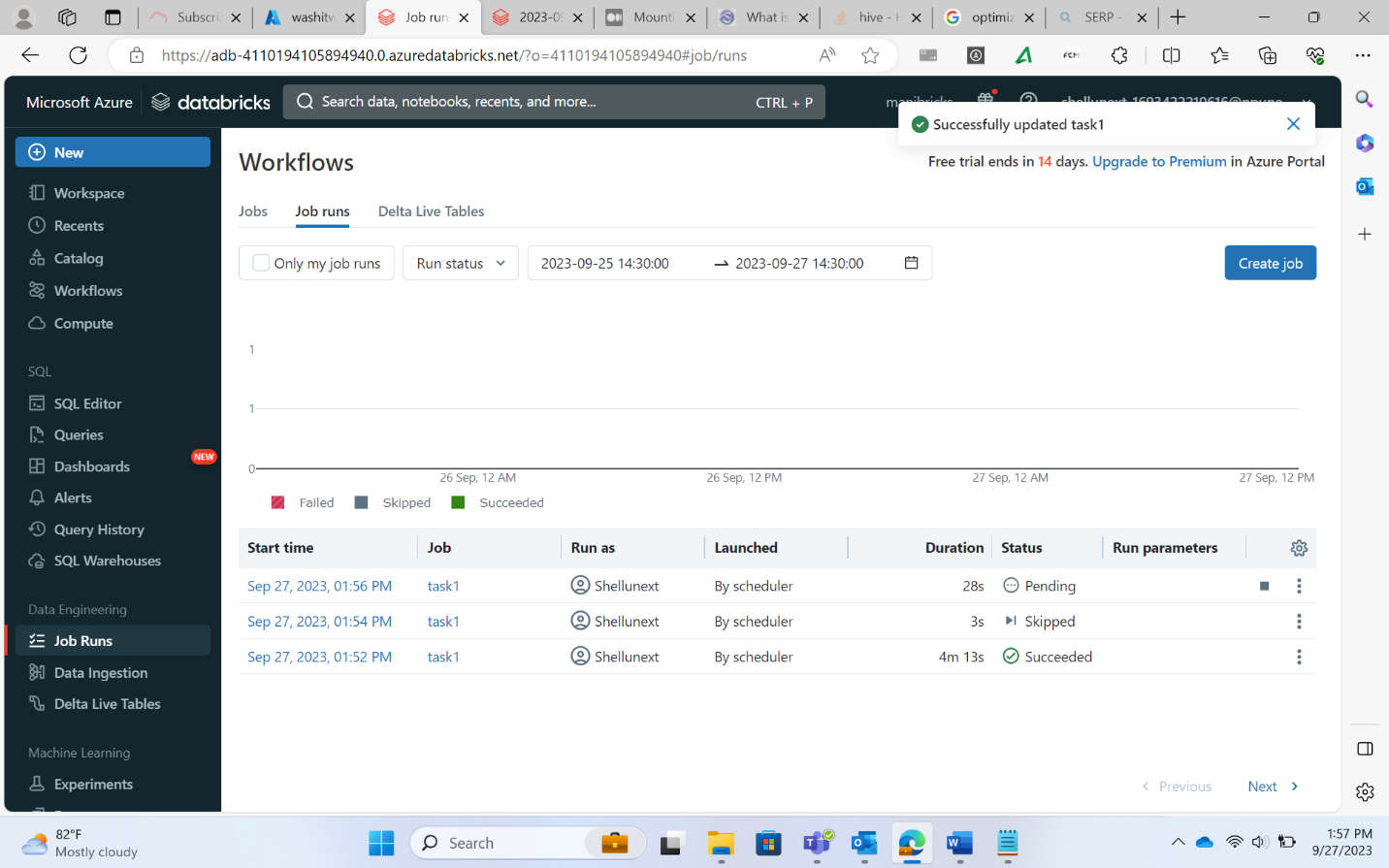
* single source of truth
  + all the data is present in one place
  + ADLS Gen2
  + In multi hop architecture Raw/Bronze, Silver, Gold zones are present
  + Bronze zone:
    - Acts as the single source of truth
    - All the data is dumped here
    - The data is curated and cleansed here and sent to the Silver Zone
  + Silver zone:
    - Based on the business requirement, further operations are done here
    - Aggregations, joins, and any operations that need to be done on the cleansed data are done here in the Silver Zone
  + Gold Zone:
    - Here the final results are produced that need to be handed over to the concerned parties
    - It produces the files that are needed at the end of the project

Jobs

* Using Workflows jobs can be created
* These jobs automate the execution of a notebook at specific intervals as specified by the user
* For jobs, tasks need to be created first
* At first only one task creation option is present
* Relationships can be created between the tasks by a process called Orchestration

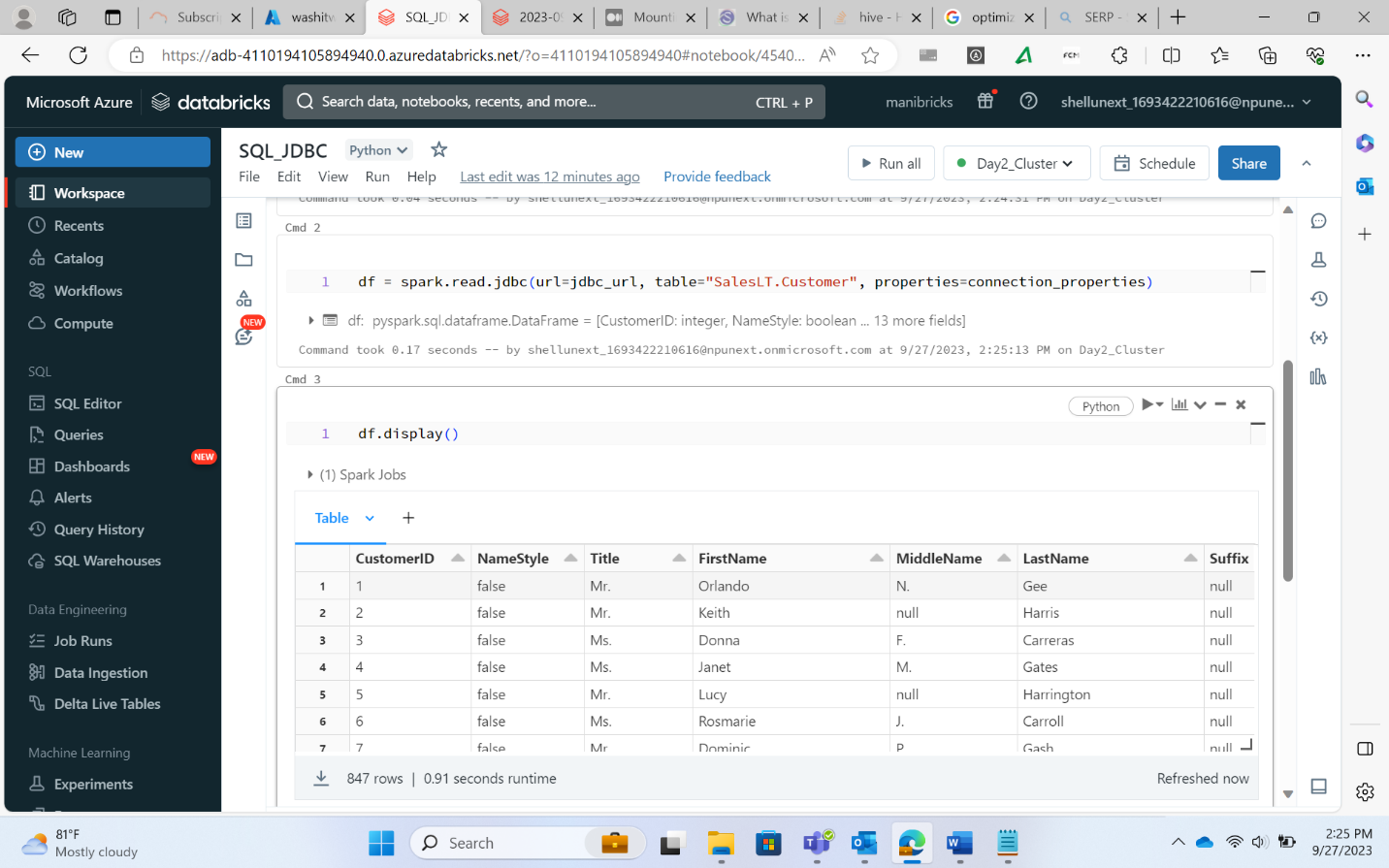
\*Note: No unnecessary actions like show, display, count etc, should be placed in the Production environment.

Schedules and Triggers



JDBC Connection to Databricks

* Create one Azure SQL DB - client
* The connection string is given to everyone
* Data is read from this



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