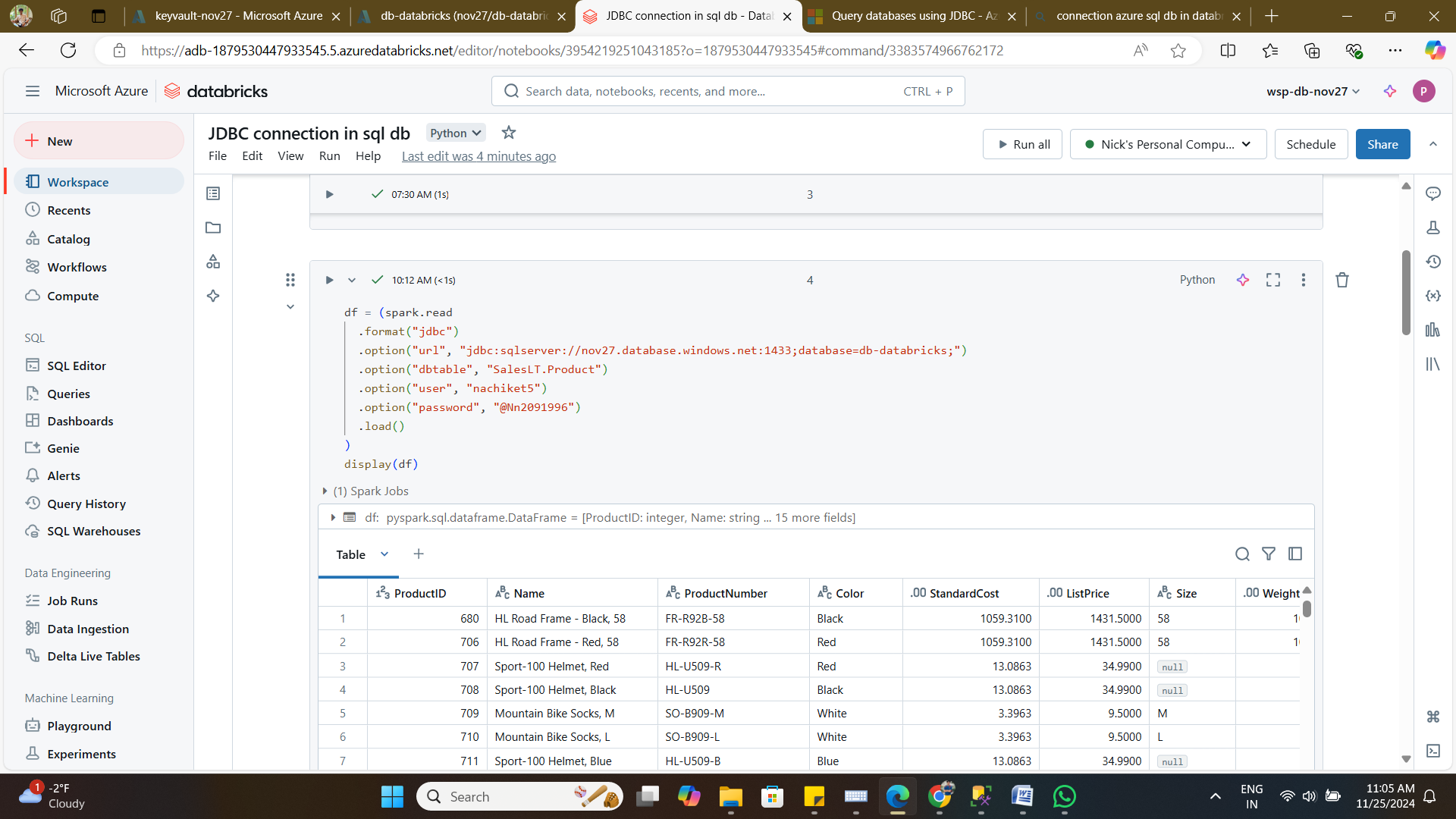
**Data Handling with SQL Database, ADLS Gen2, and Databricks**

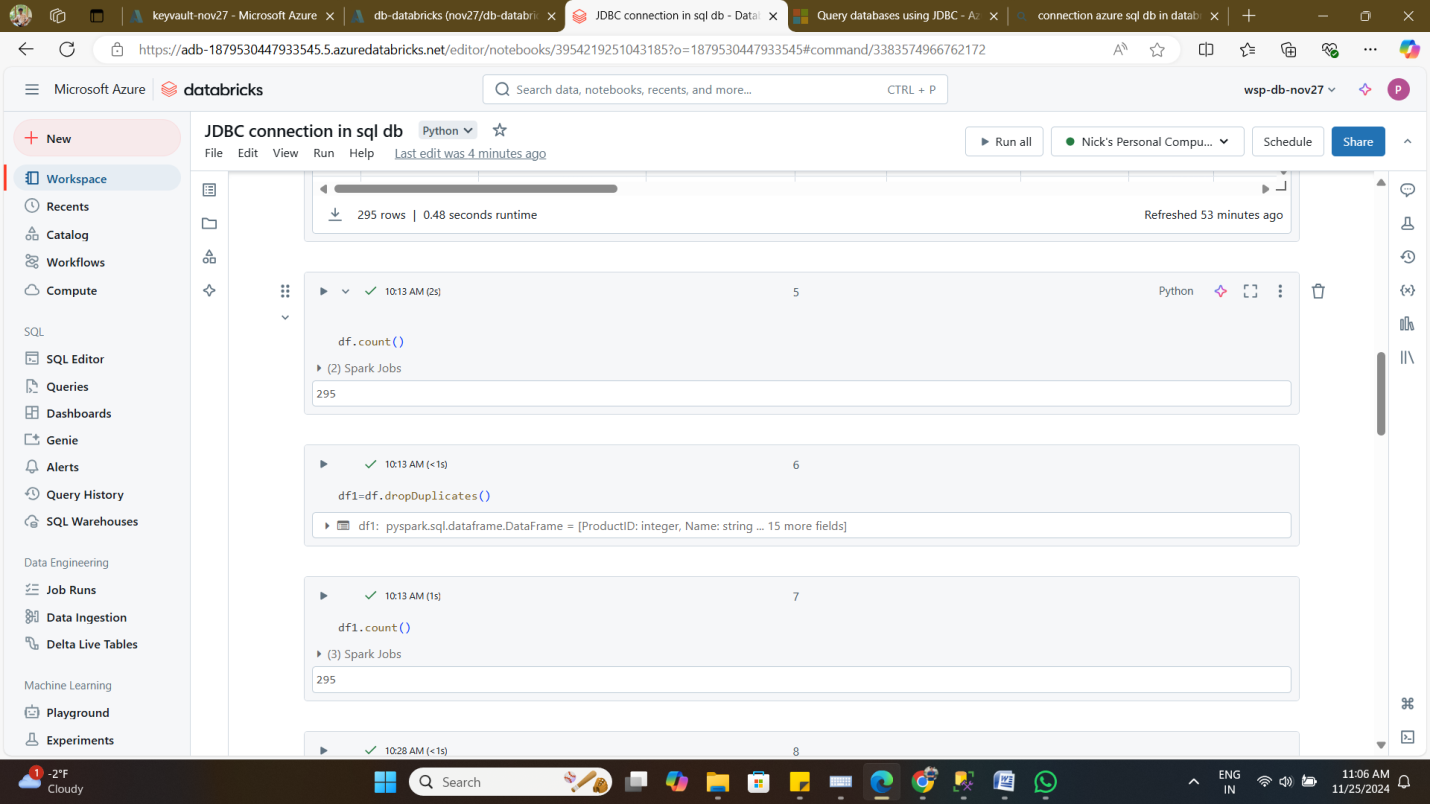
**Nachiket Prajapati**

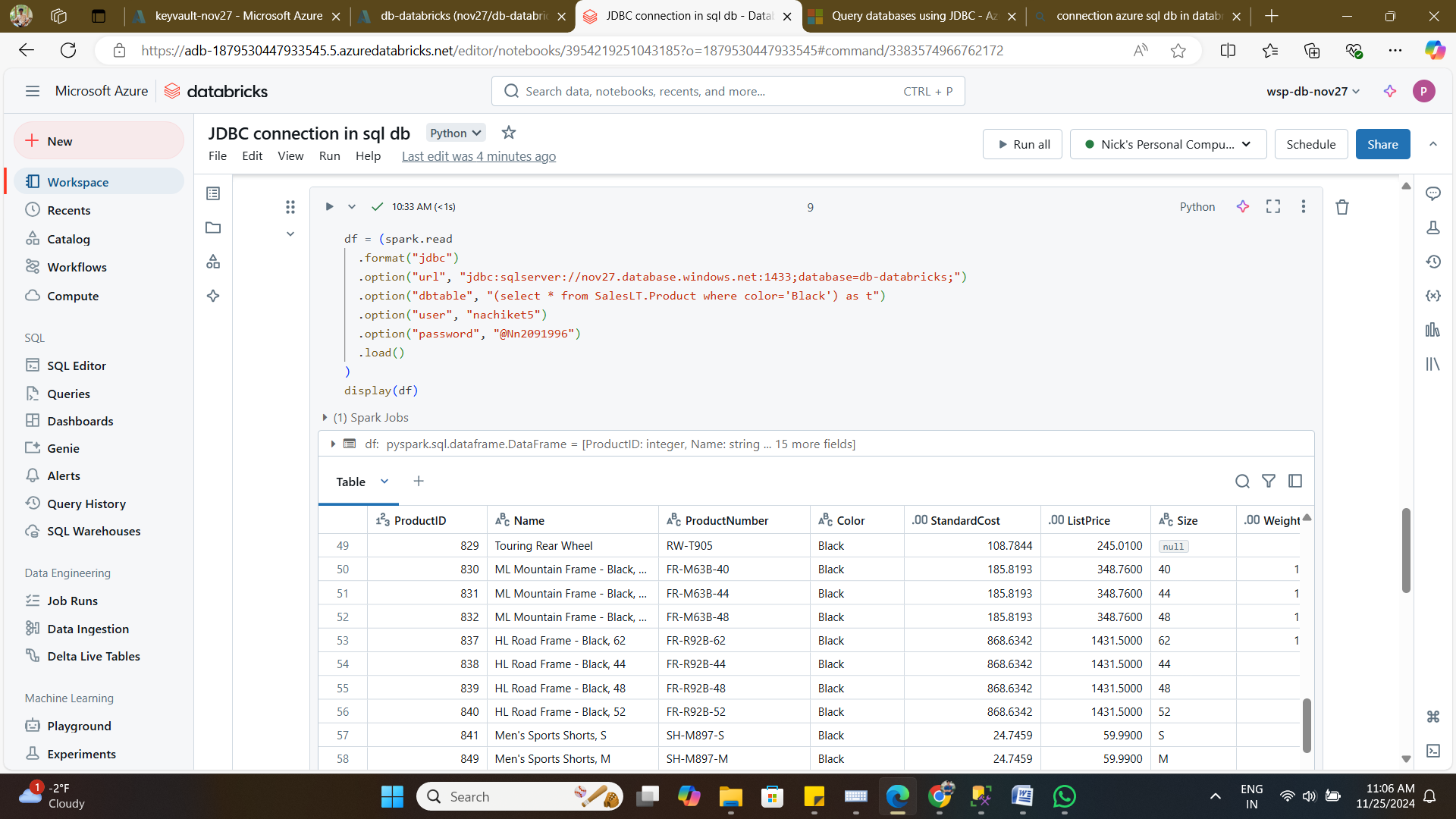
**Objective:**

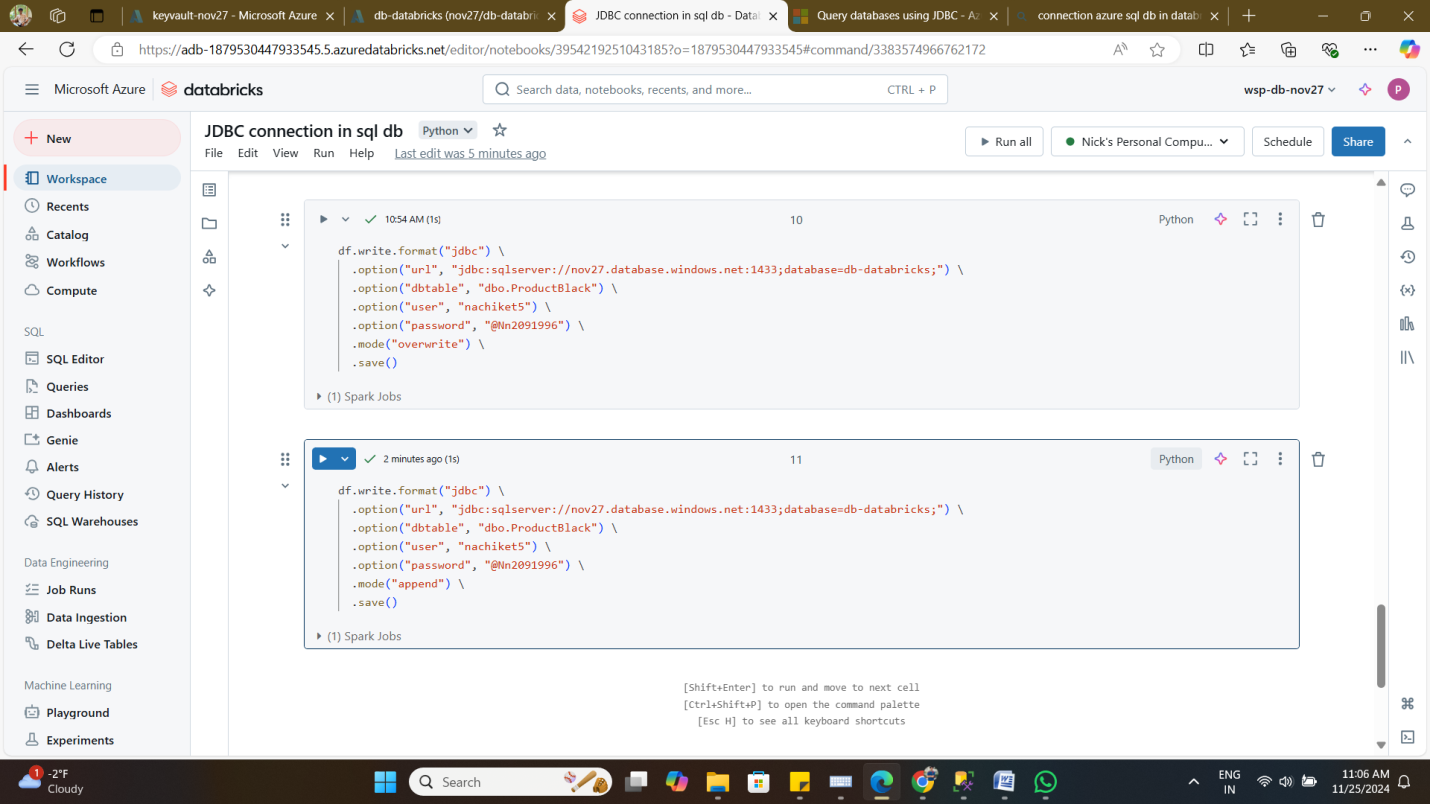
Build a data pipeline to extract data from an SQL database, clean it by removing duplicates, and store it in Azure Data Lake Storage (ADLS) Gen2 using Databricks.

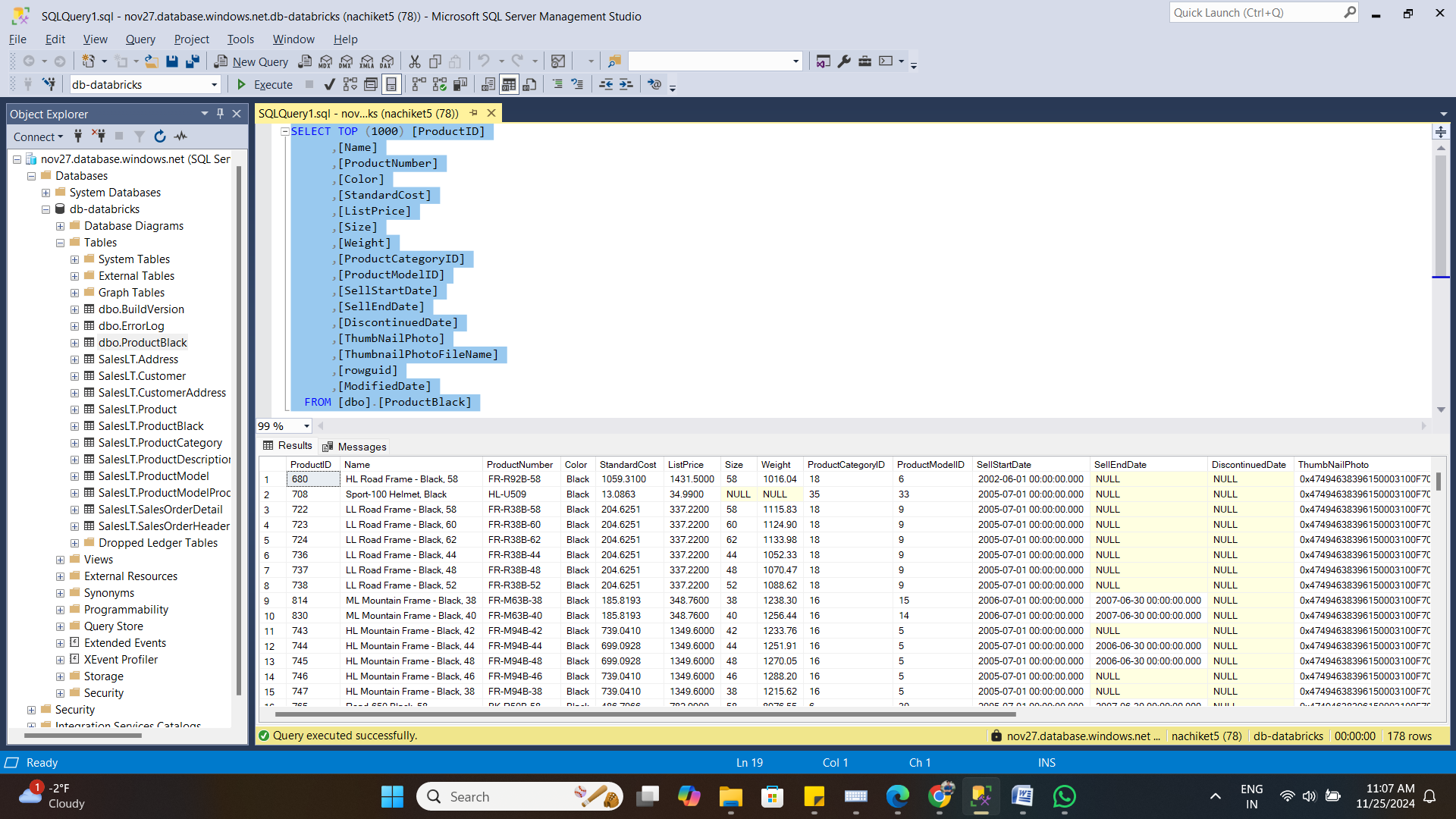
### **Tasks:**

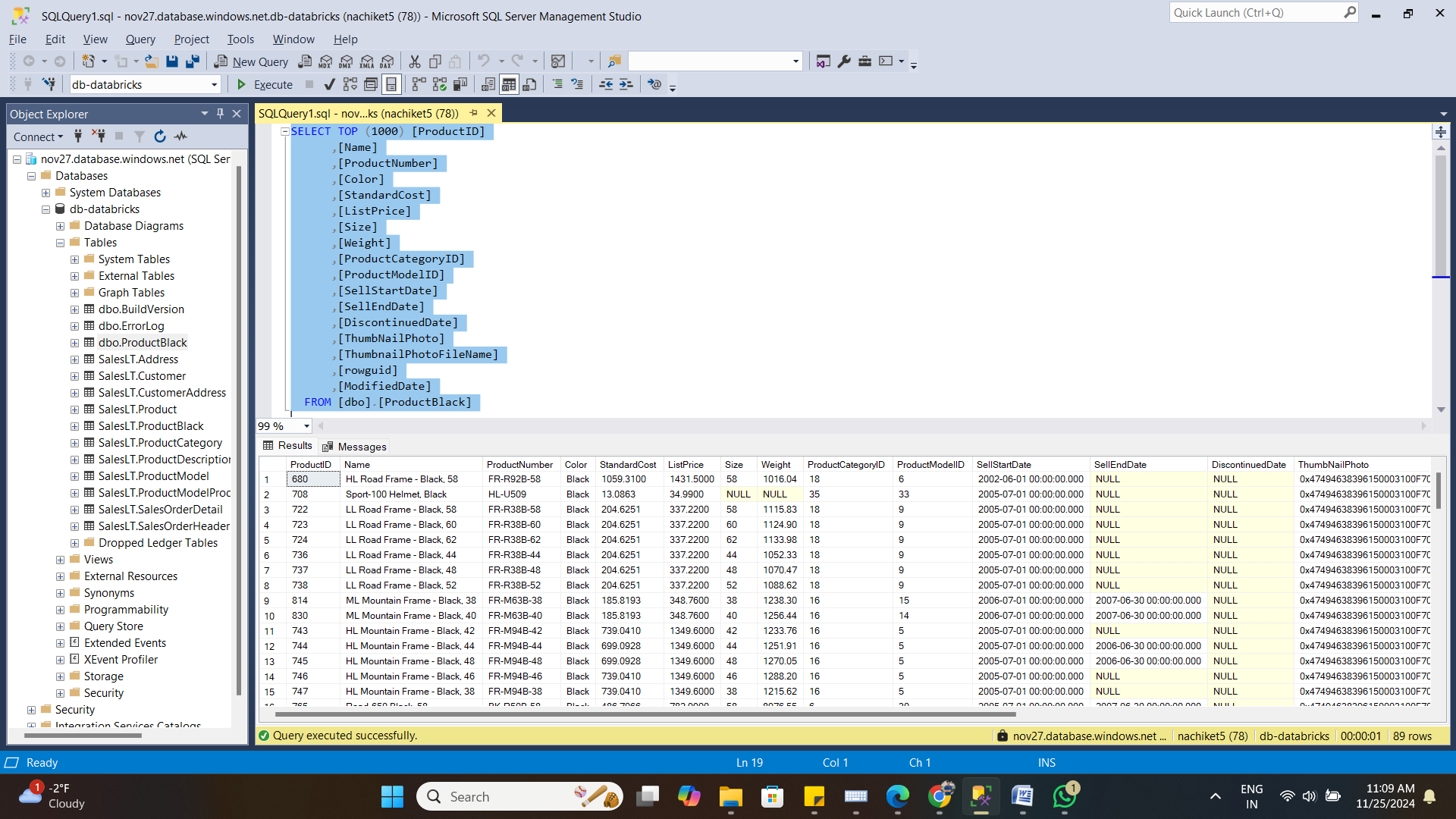
1. **Extract Data from SQL Database**
   * Establish a connection to the SQL database using JDBC.
   * 

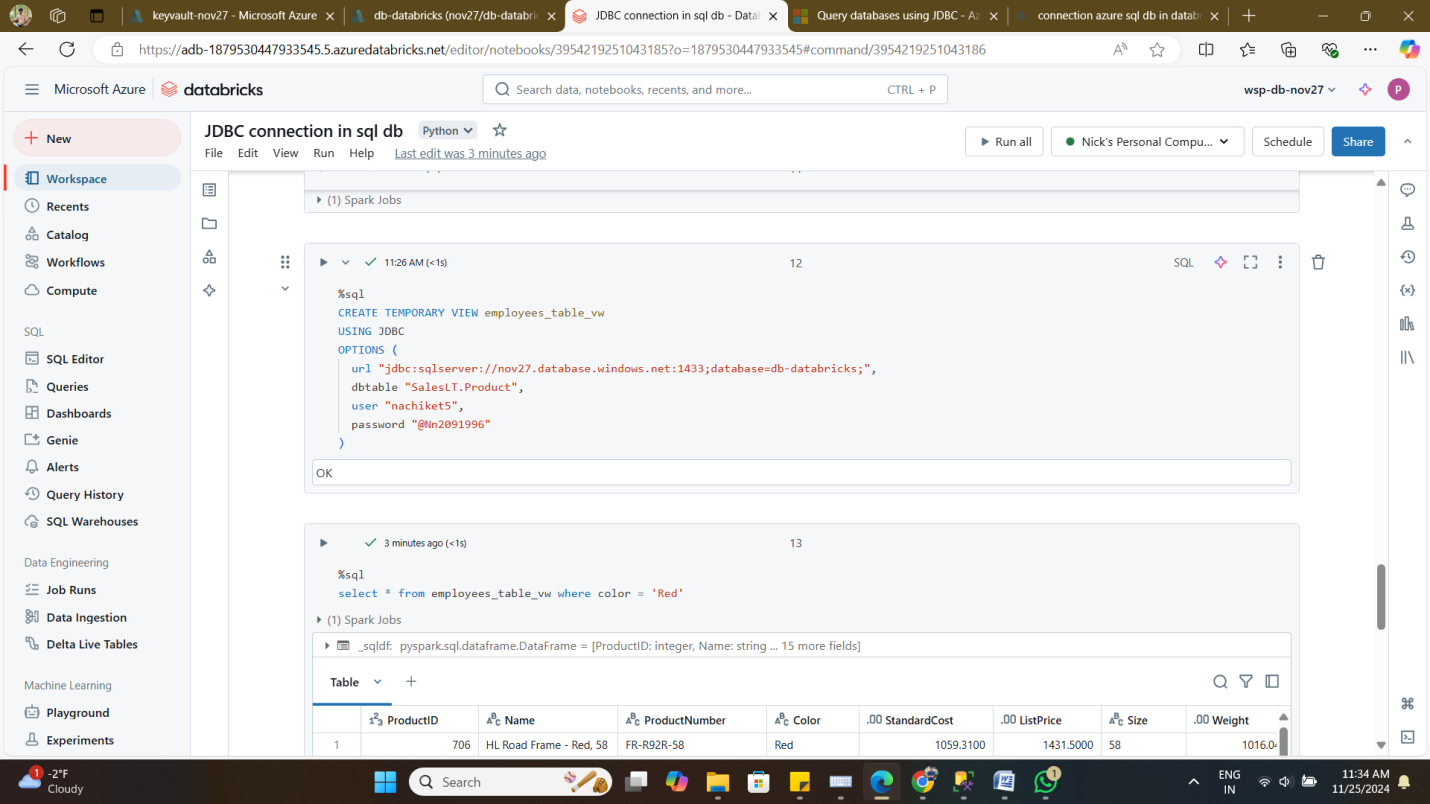


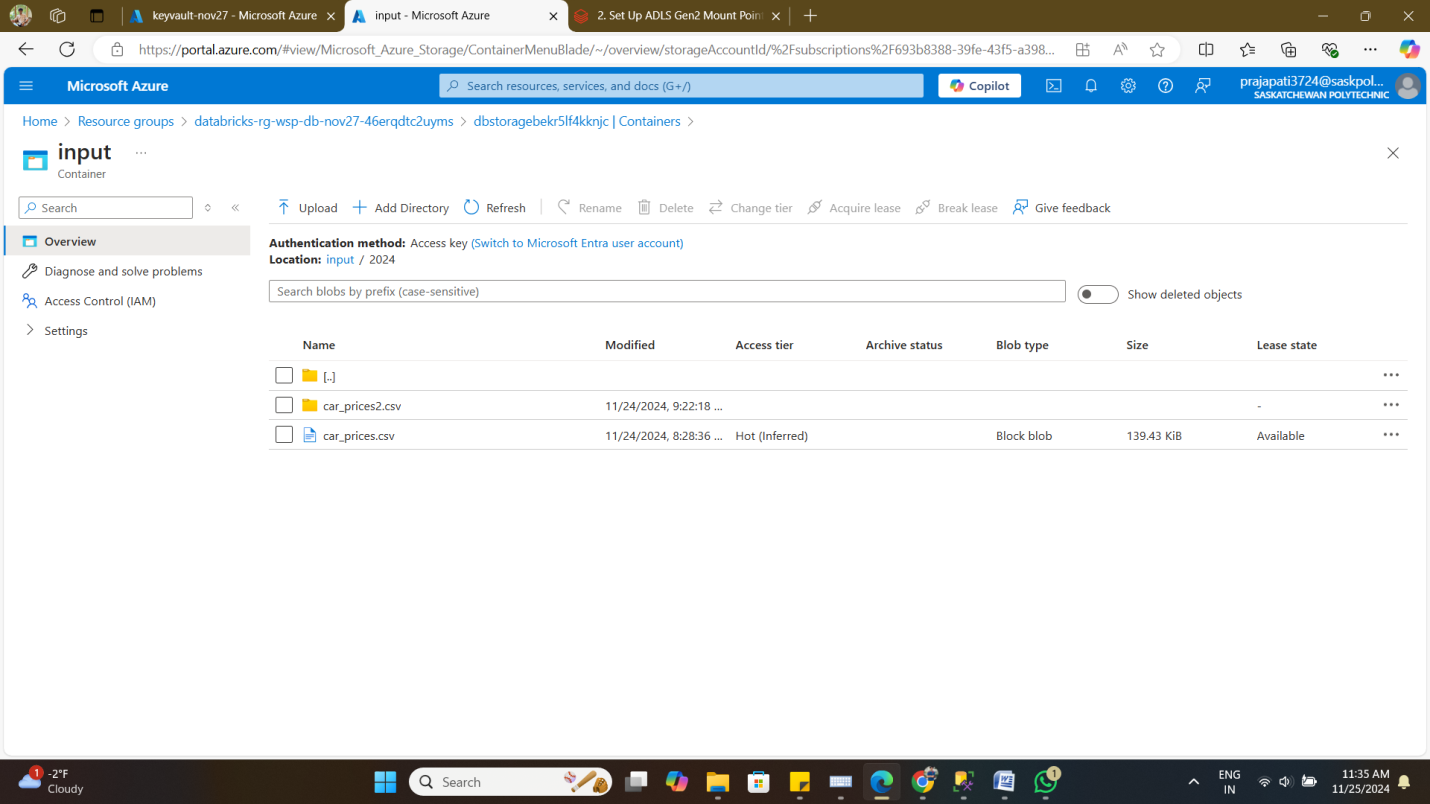


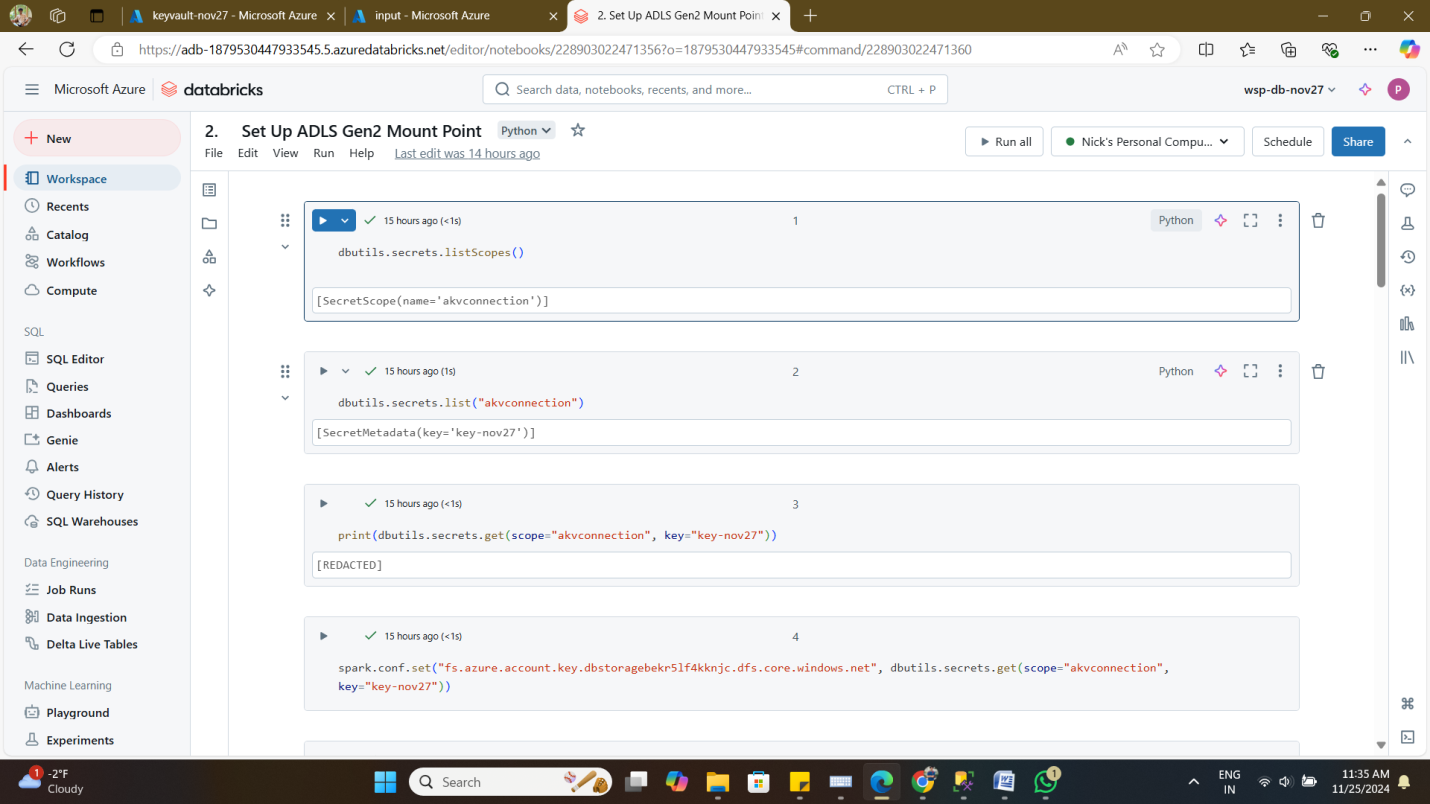


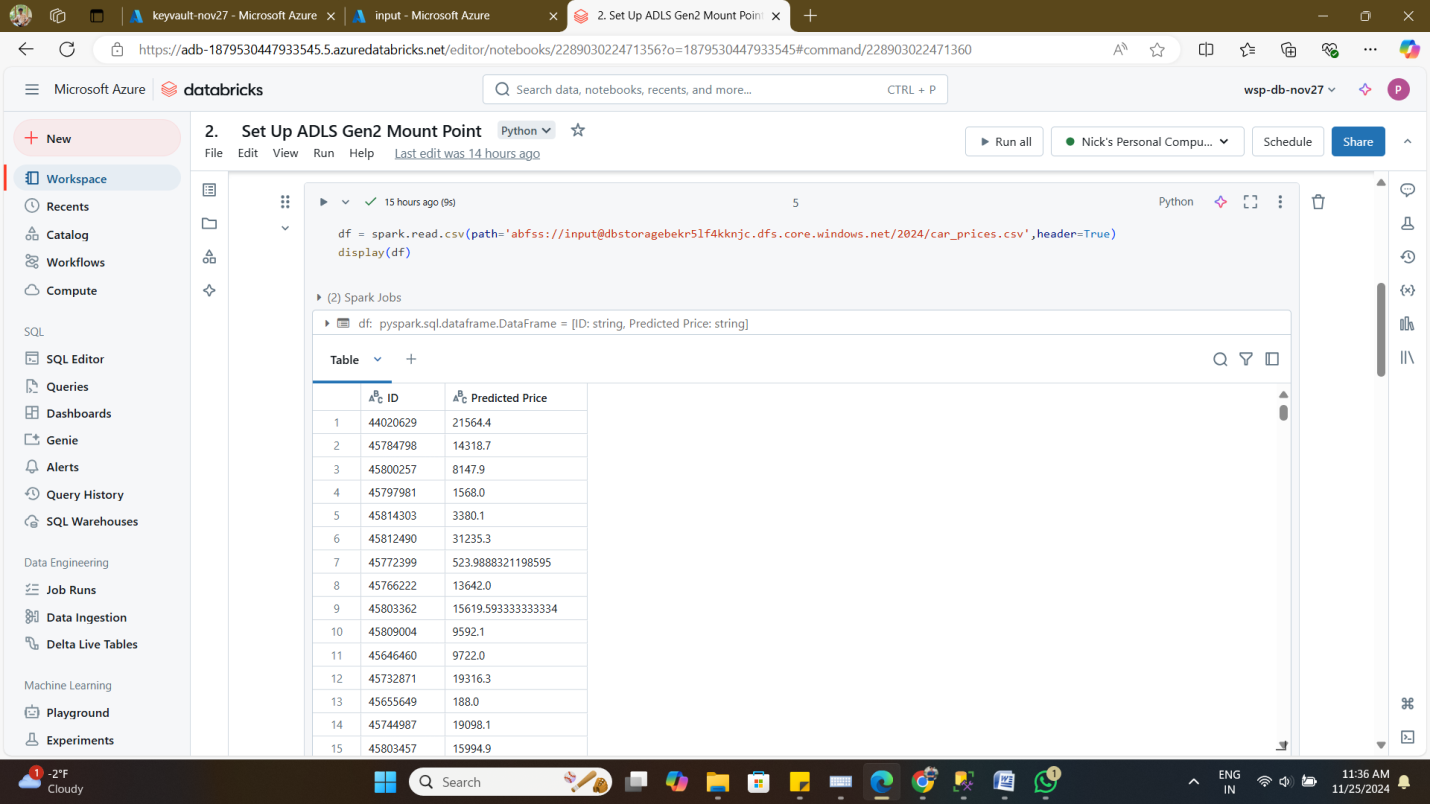




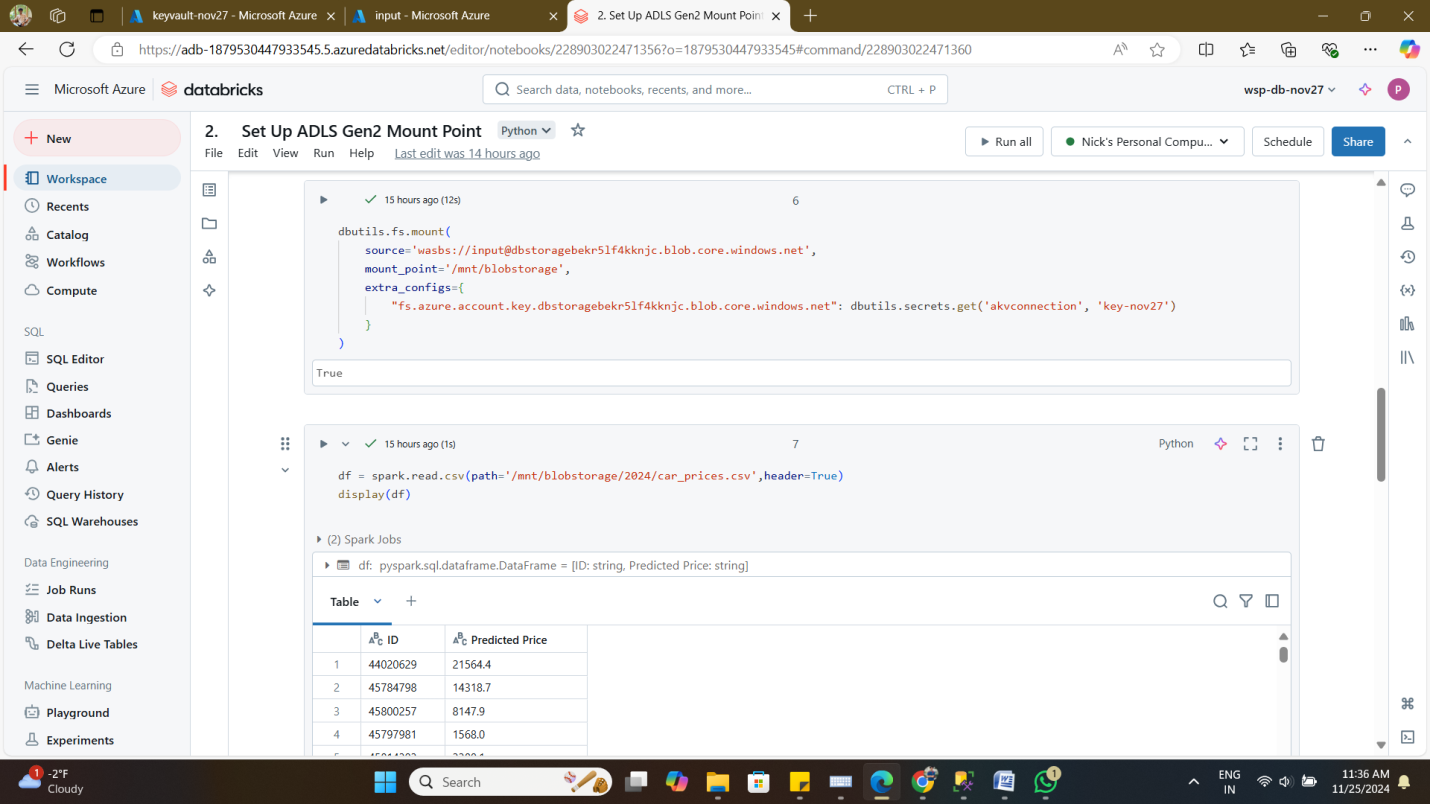


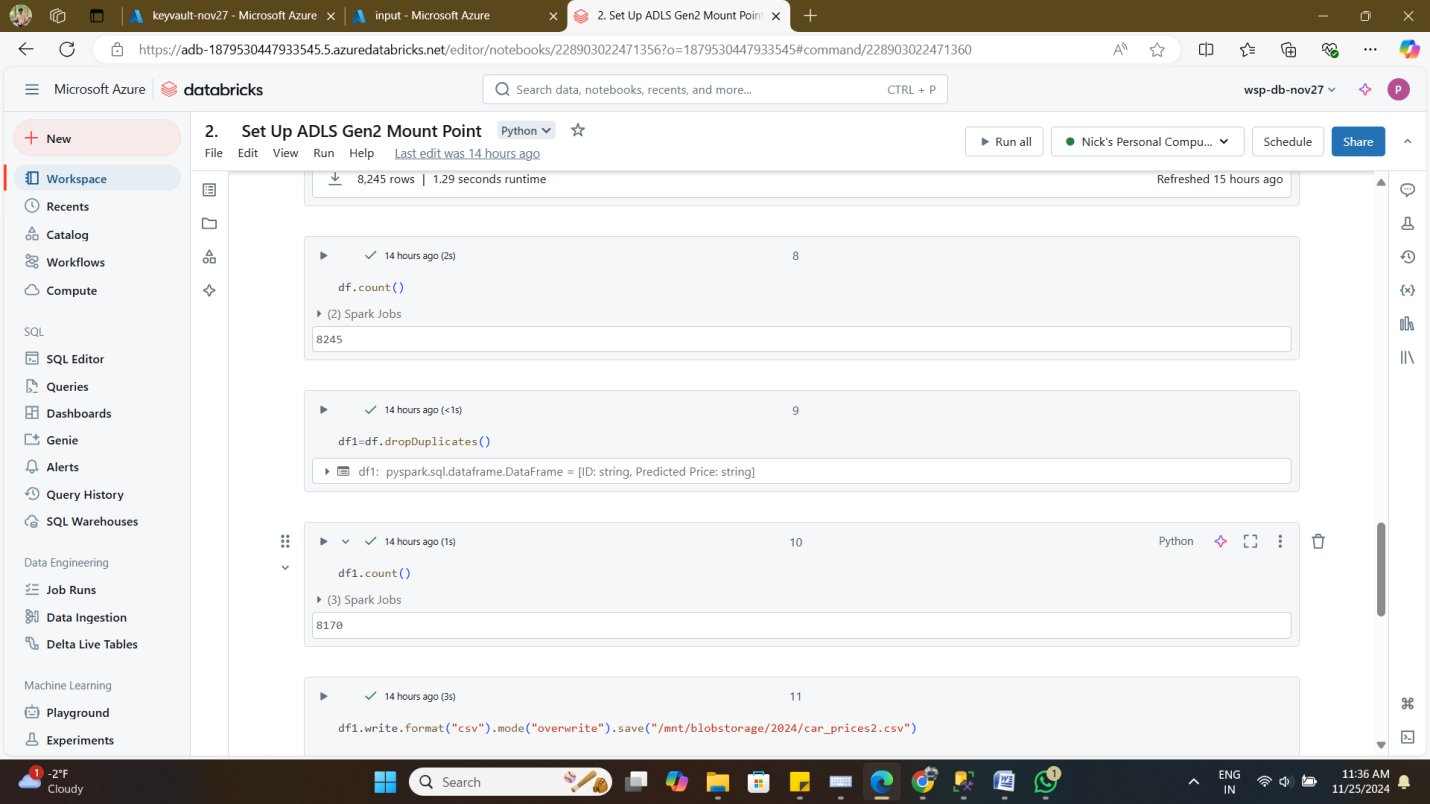
1. **Set Up ADLS Gen2 Mount Point**
   * Create an ADLS Gen2 container.
   * 



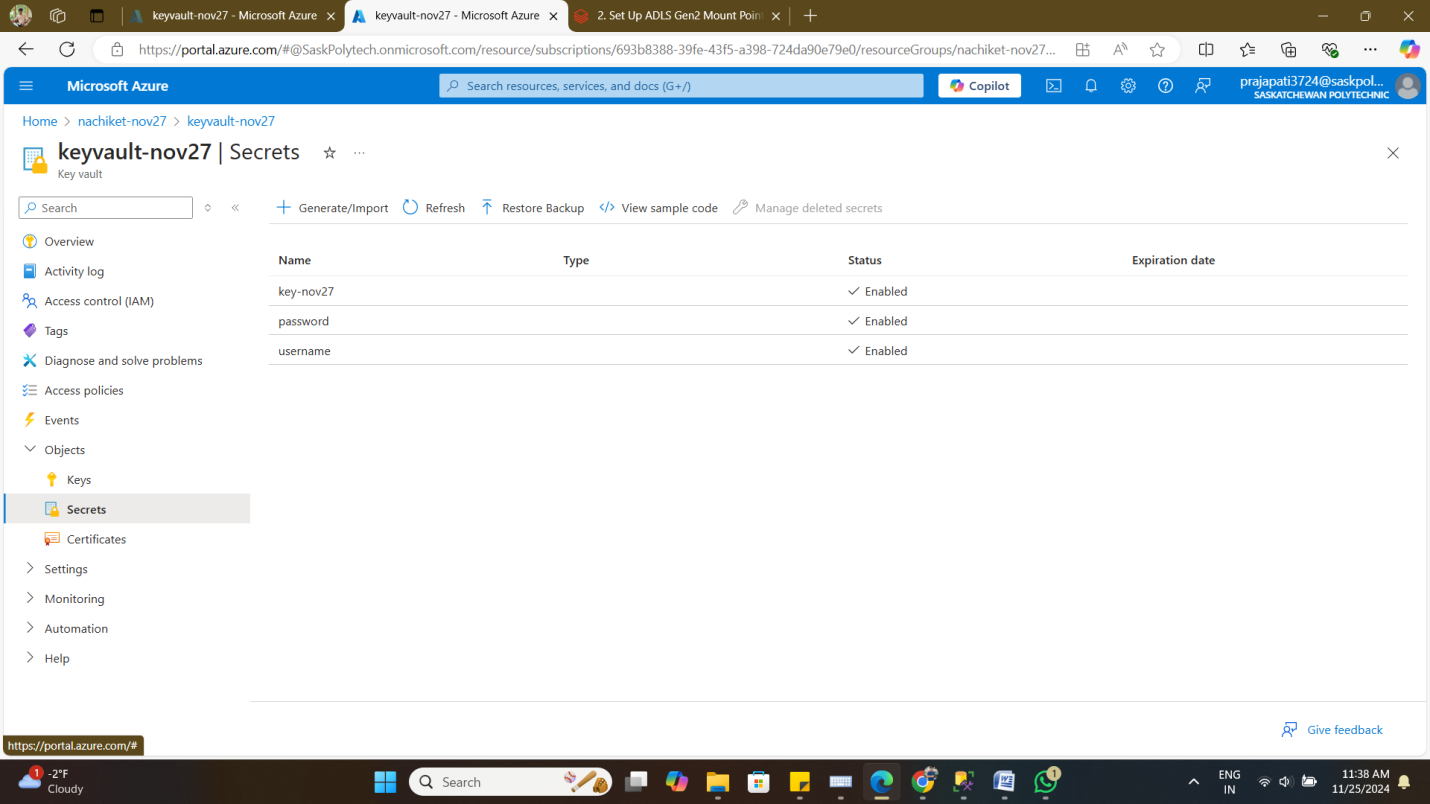


* + Configure a Databricks mount point for ADLS Gen2 access.

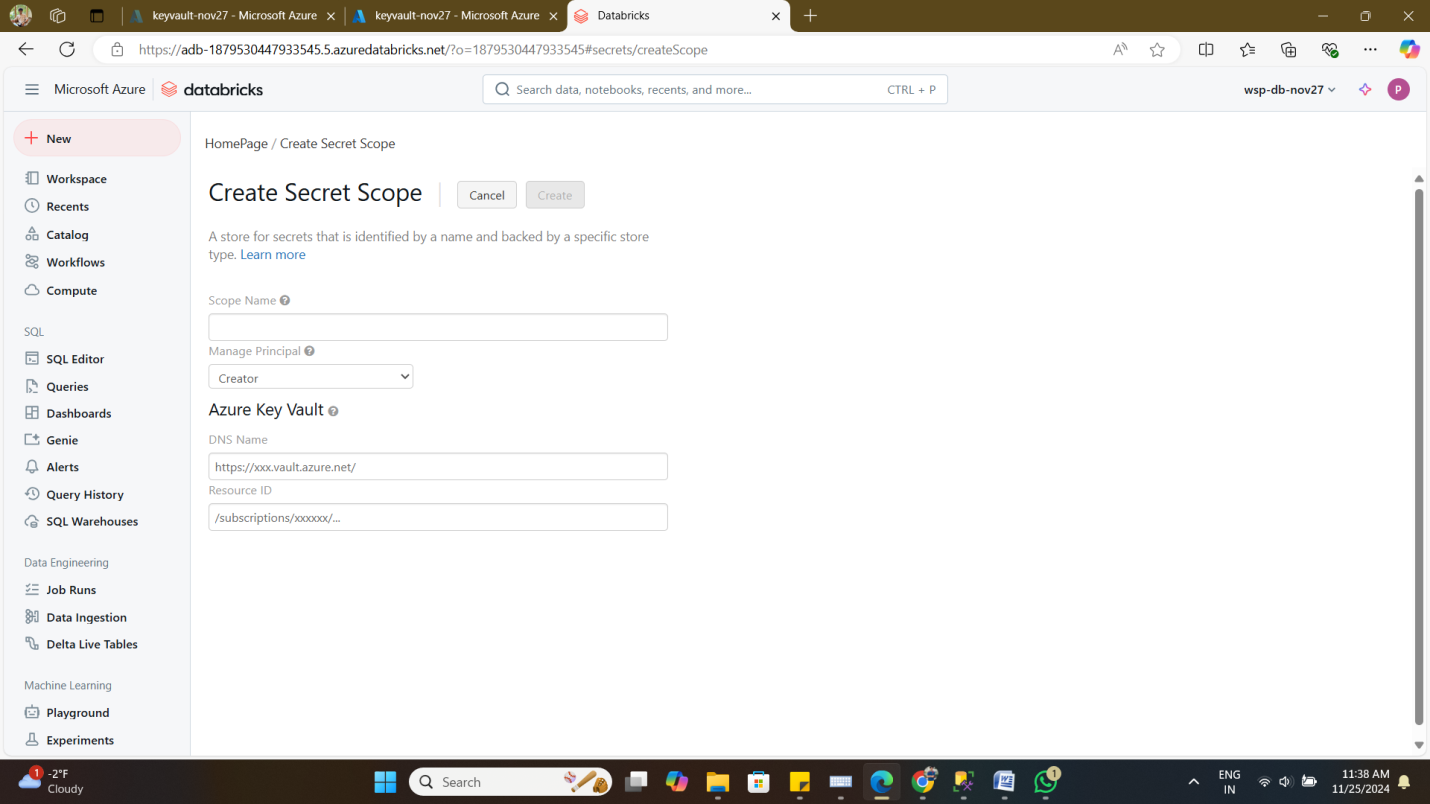




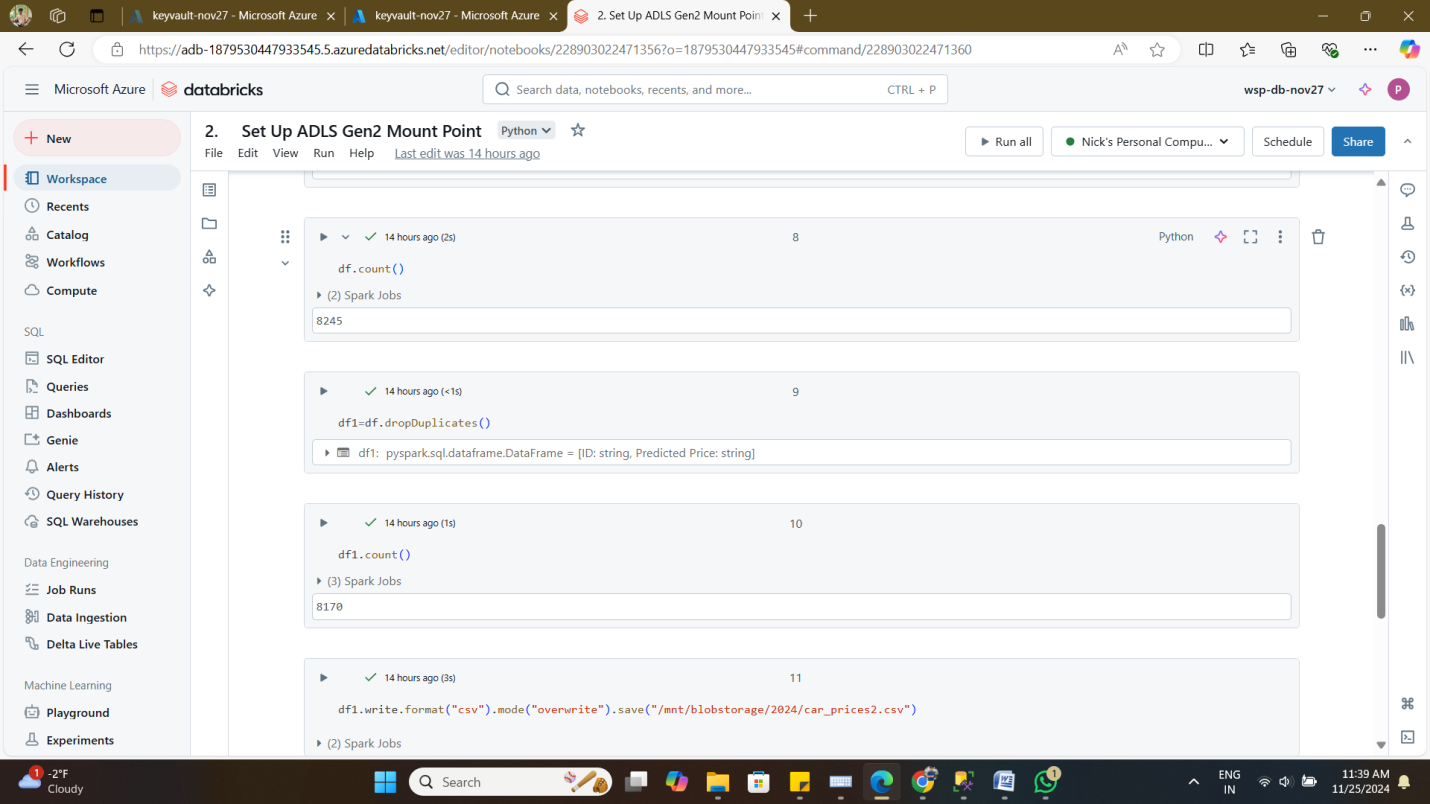
1. **Configure Azure Key Vault and Databricks Scop**e
   * Create an Azure Key Vault and add secrets.



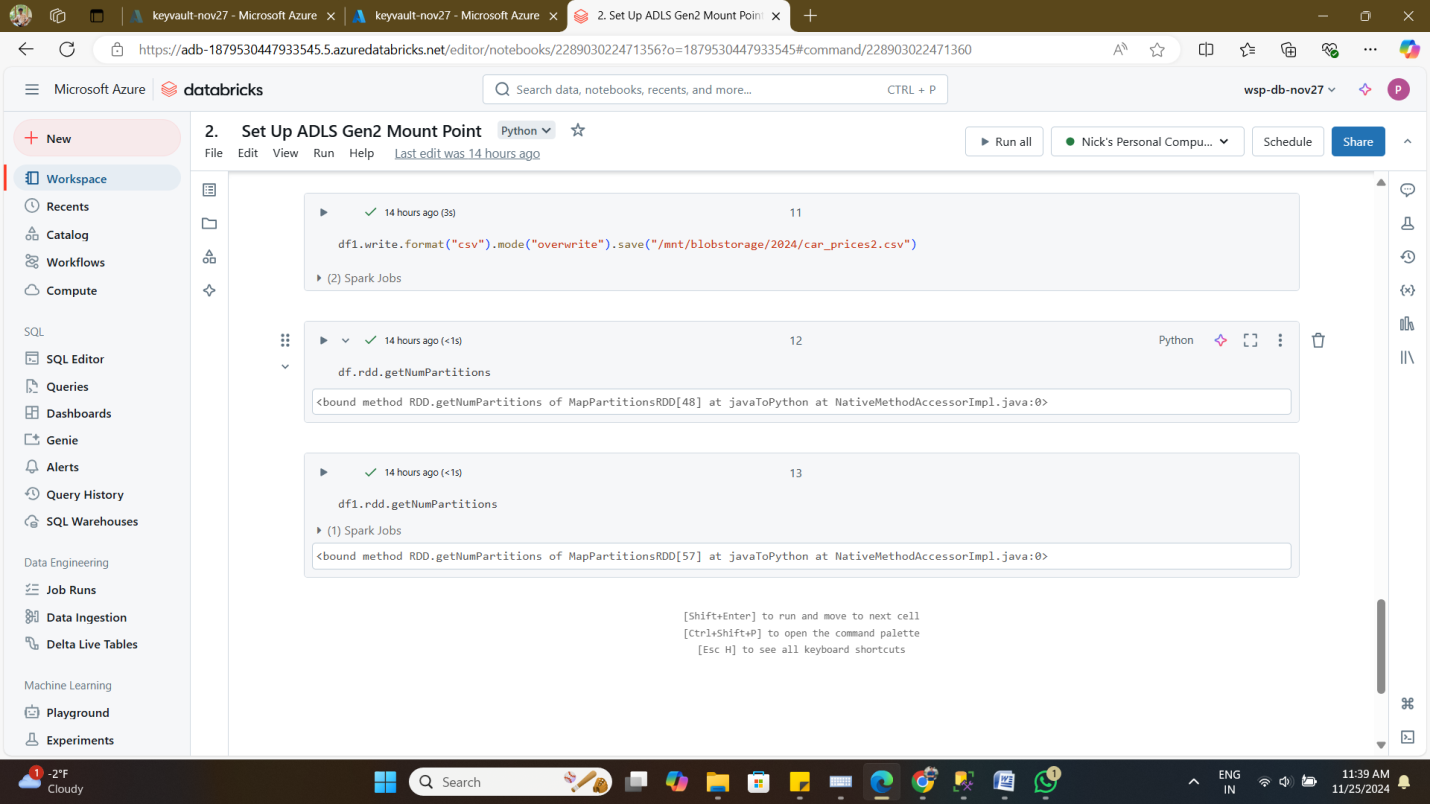
* + Set up a Databricks secret scope for secure credentials management.



1. **Clean Data (Remove Duplicates)**
   * Identify and remove duplicate records.



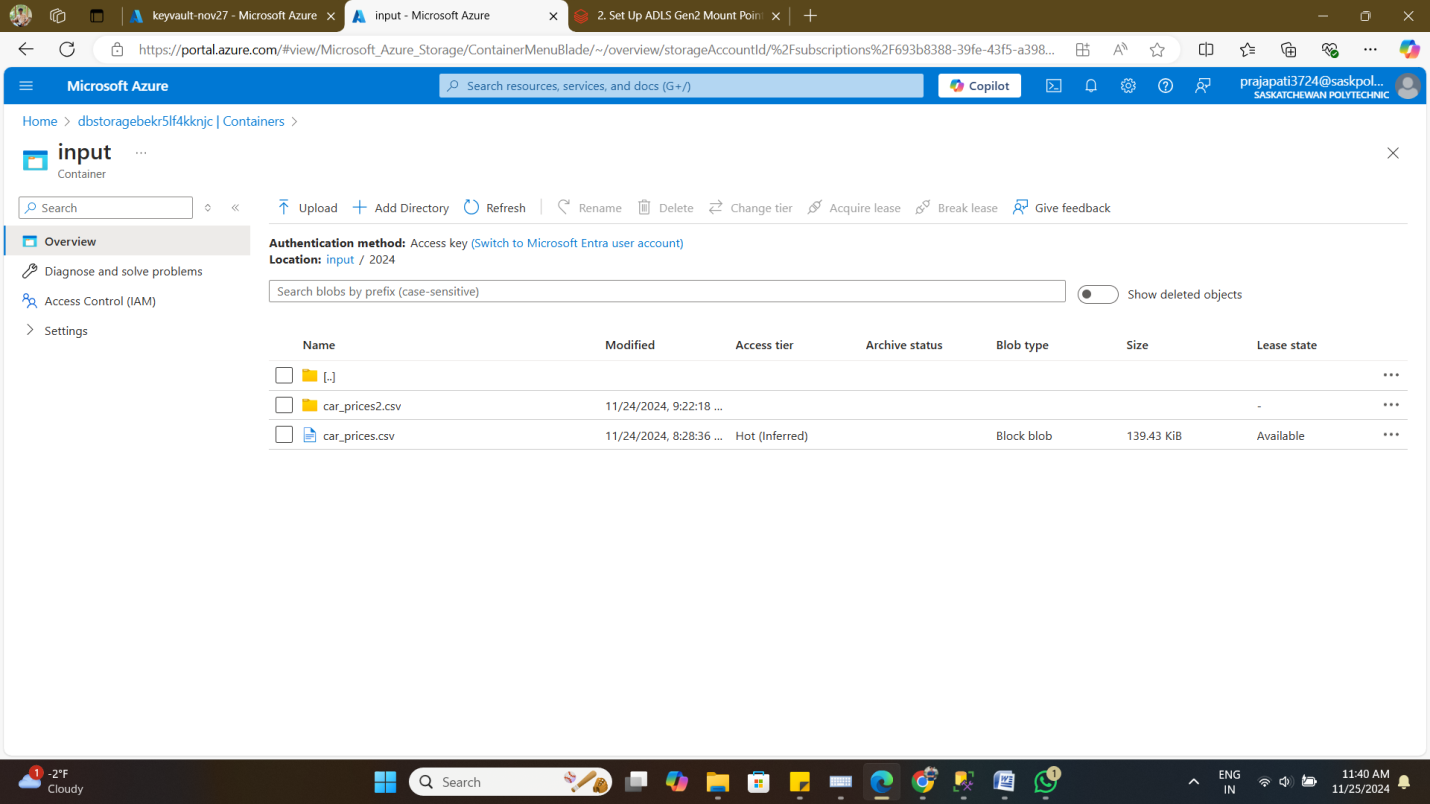
* + Validate the cleaned data.



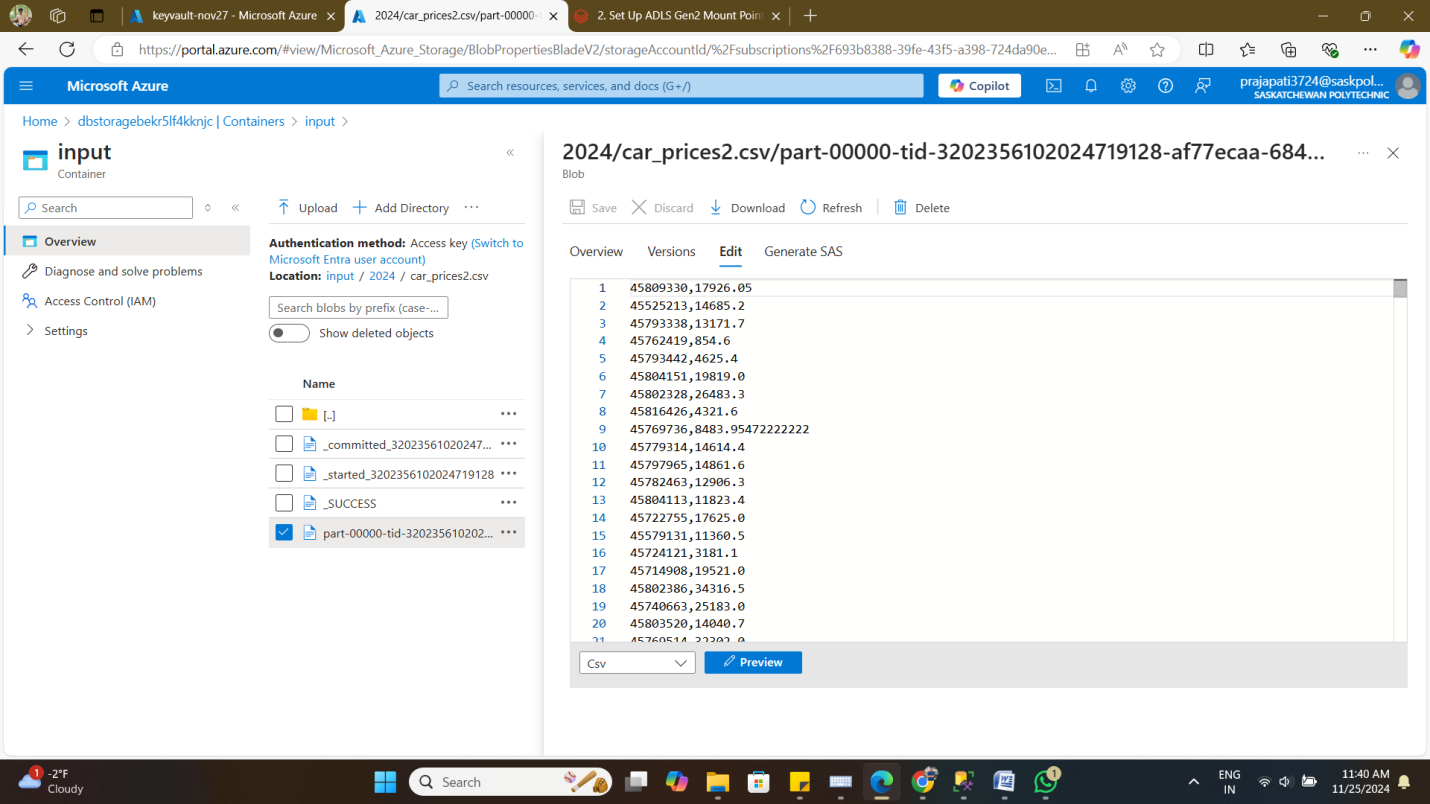
1. **Save Cleaned Data to ADLS Gen2**
   * Save processed data in the designated container.



* + Verify the saved data format.







**Challenges:**

* Establishing Secure Connectivity: Connecting to the SQL database and ADLS Gen2 securely required proper configuration of JDBC, Azure Key Vault, and Databricks scopes.
* Data Duplication: Identifying and removing duplicate records from the dataset was critical to ensure data integrity.
* Secure Credentials Management: Safeguarding sensitive credentials like database login information and Azure storage keys was essential to prevent unauthorized access.
* Data Validation and Storage: Ensuring the processed data was correctly validated and stored in the designated format within the ADLS Gen2 container.

**Solutions:**

* Secure Configurations: Implemented JDBC for SQL connectivity and used Azure Key Vault with Databricks scopes to manage credentials securely.
* Efficient Data Cleaning: Developed and validated scripts in Databricks to identify and eliminate duplicate records effectively.
* Secret Management: Utilized Databricks’ secret scope feature to securely store and retrieve sensitive credentials during the pipeline process.
* Data Verification: Conducted thorough validation checks on the cleaned data before saving it in the desired format in ADLS Gen2.