SystemRequirementsSpecification Index

For

Data Bricks Netflix shows data ingestion and analysis using pyspark .

Version 1.0

IIHT Pvt. Ltd.
fullstack@iiht.com

Problem Statement : **DataBricks Netflix shows data ingestion and analysis using pyspark .**

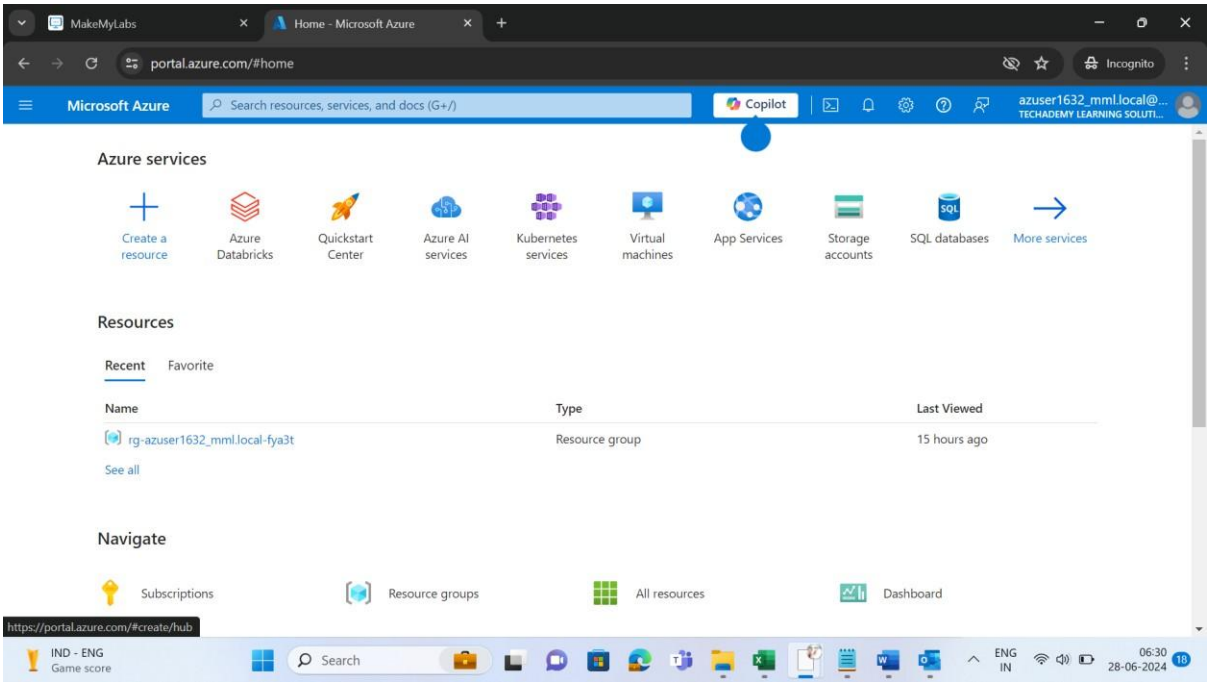
.

Description : Use relevant methods operations toperform specified activities which are given in the instructions.

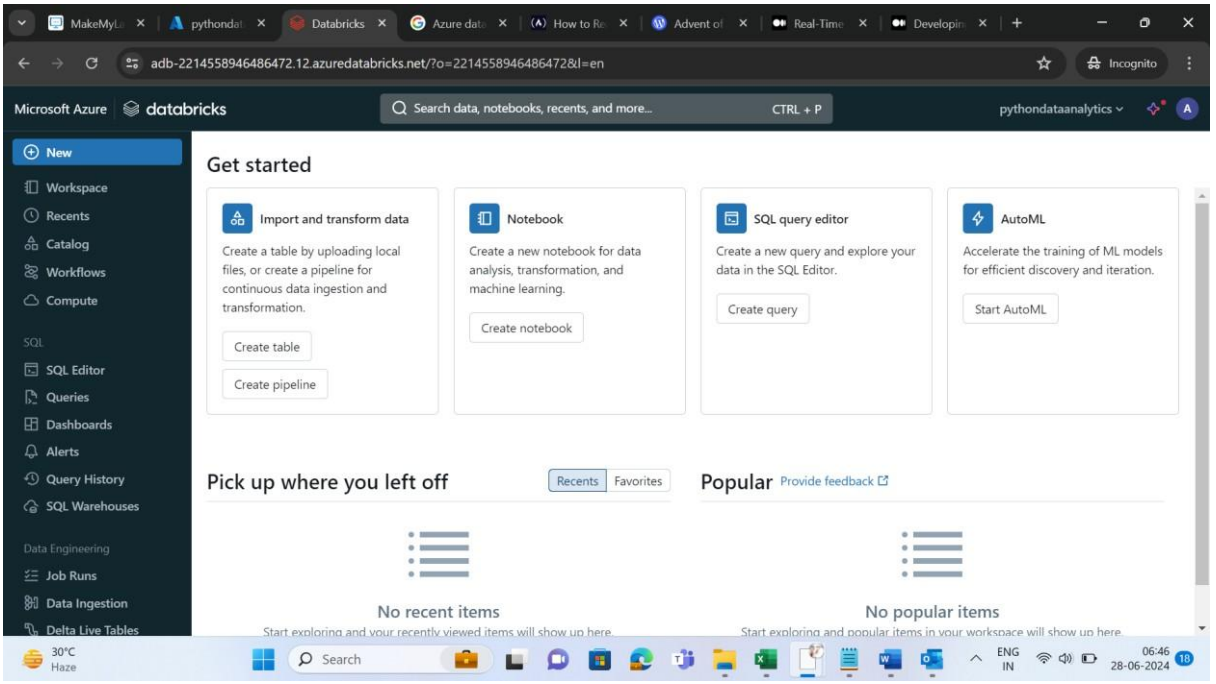
ABC Media Inc. is a global media company that operates a streaming platform similar to Netflix, offering a diverse catalog of movies and TV shows to millions of subscribers worldwide. To enhance user experience and optimize content offerings, ABC Media Inc. aims to leverage advanced data analytics to understand preferences, and content performance, most watched . By leveraging Databricks with PySpark, ABC Media Inc. transforms complex data into actionable insights, gaining a competitive edge in the global streaming industry

- Steps to login in the Azure account → Azure Databricks→ create workspace → Create cluster →

- Use the given dataset to create a table in the workspace



- Click on new and create a new notebook



Once you have imported the table check for respective rows and columns with their respective datatypes are imported correctly.

Table +

	ABC _1	ABC _2
1	show_id	string
2	type	string
3	title	string
4	director	string
5	cast	string
6	country	string
7	date_added	string
8	release_year	bigint
9	rating	string
10	duration	string
11	listed_in	string
12	description	string

Note

- You should be able to import the data github
- You should be able to write the code in pyspark code
- Create the workspace name as **PYTHONDATAANALYTICS** .

Obtain the Dataset from GitHub

Step 1

1. **Find the Dataset on GitHub:**
   - Go to the GitHub repository that contains the dataset.
   - Navigate to the file you want to import.
2. **Copy the Raw URL:**
   - Click on the dataset file in the GitHub repository.
   - Click the "Raw" button to open the raw file content.
   - Copy the URL from the browser's address bar. This URL will be used to download the data directly.

Step 2:

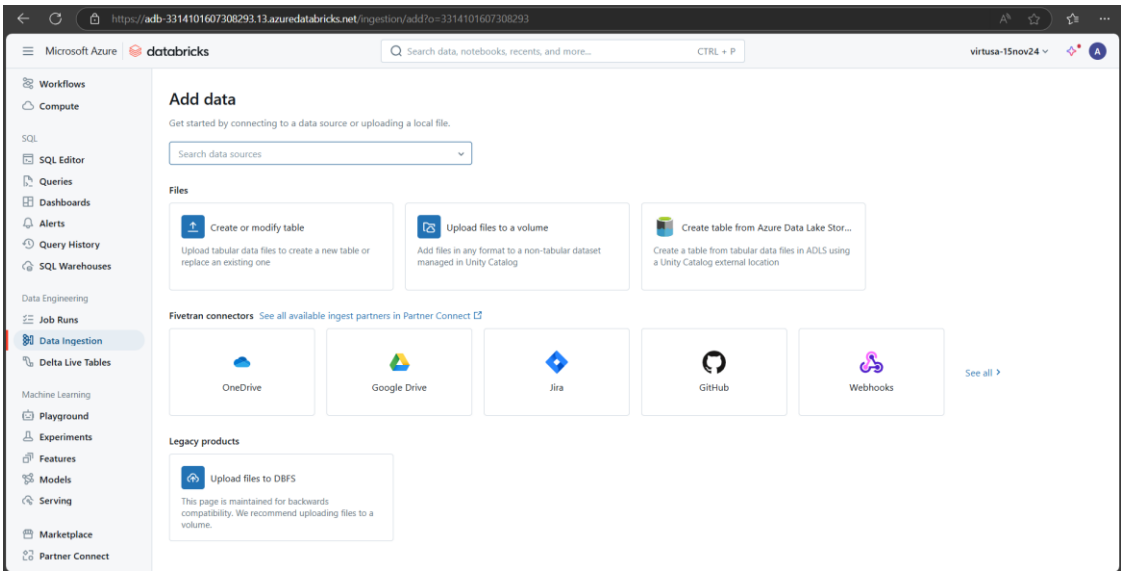
Import Dataset into Azure Databricks

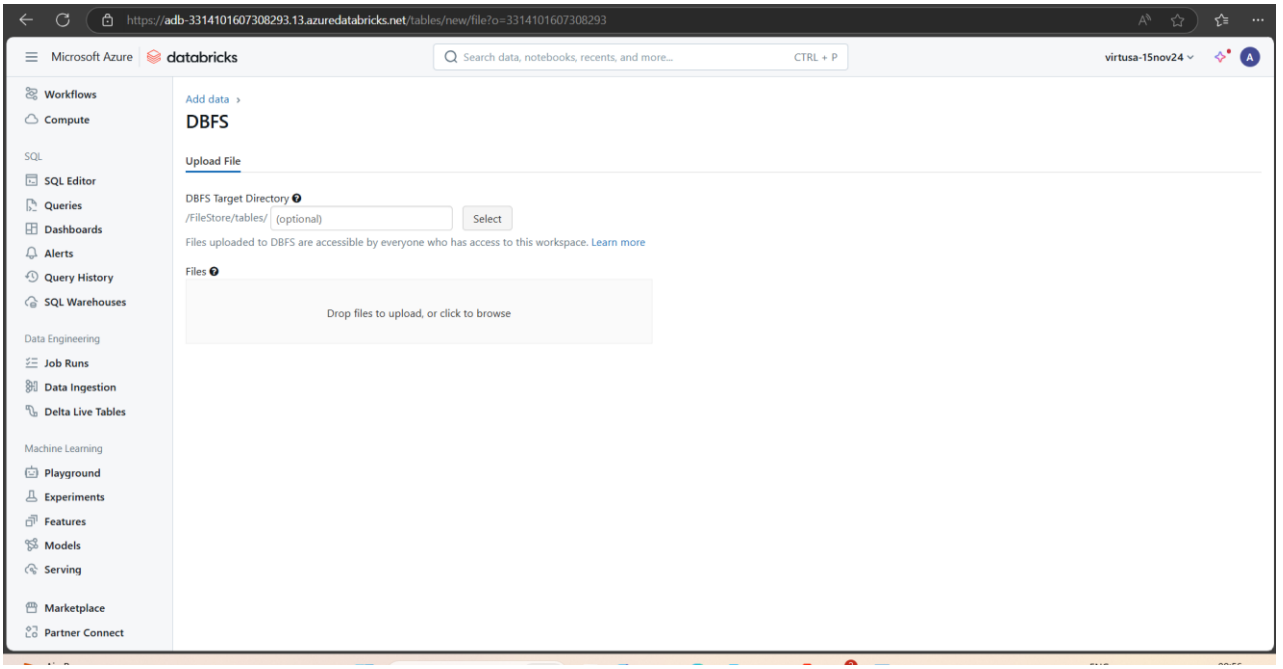
Method 1: Import Using Databricks Notebook

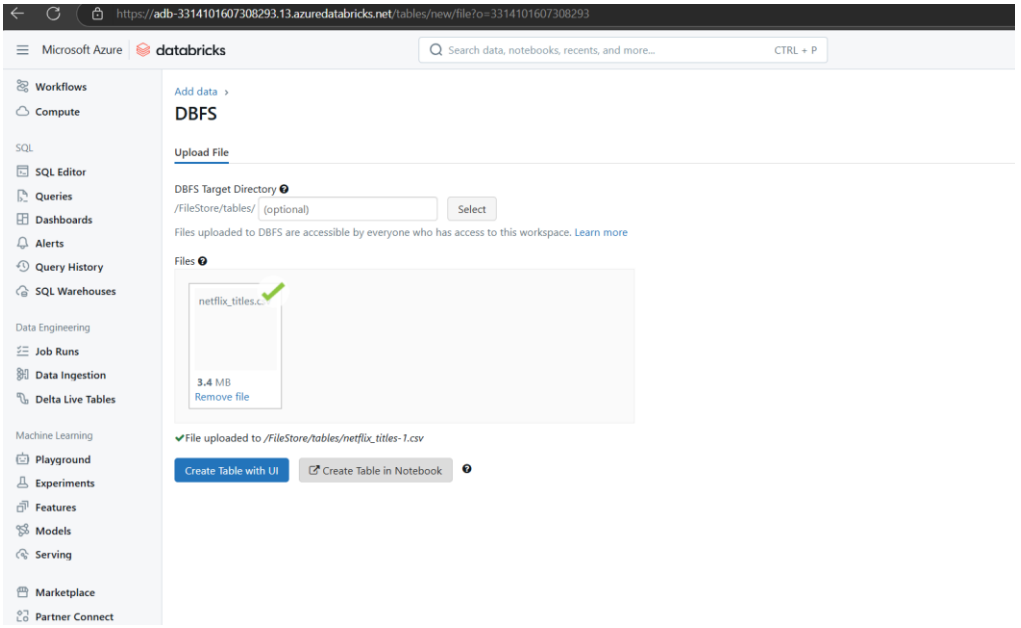
1. **Open a Notebook:**
   - In Azure Databricks, go to the "Workspace" tab.
   - Click on the "Create" button and select "Notebook."
   - Name your notebook and choose the language (e.g., Python).
   - Attach the notebook to your running cluster.
2. **Use PySpark to Load the Dataset from GitHub URL:**

Here's an example using PySpark to directly read a CSV file from GitHub: Pyspark

First upload the code from the git to DBFS by clicking on data ingestion





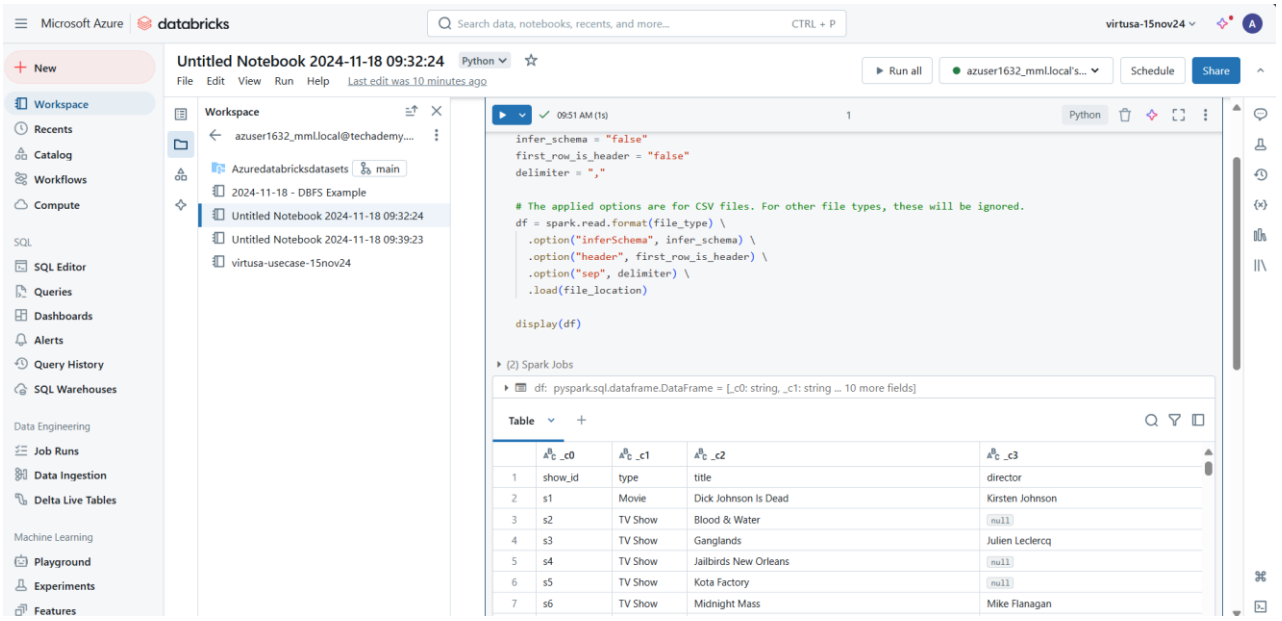



```
file_location = "/FileStore/tables/netflix_titles.csv"
file_type = "csv"

# CSV options
infer_schema = "false"
first_row_is_header = "false"
delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.
df = spark.read.format(file_type) \
  .option("inferSchema", infer_schema) \
  .option("header", first_row_is_header) \
  .option("sep", delimiter) \
  .load(file_location)

display(df)
```



Dataset link

https://github.com/IIHTDevelopers/Azuredatabricksdatasets.git

Problems

1. How can one access the table according to the instructions given take screenshot whenever necessary ? (PySpark code)
2. How would you retrieve and display the counts of movies and TV shows from the dataset? (PySpark code)

3. How would you handle null values in the database, specifically using "NA" as a replacement? (PySpark code)
4. How would you filter the database to count the number of movies and TV shows? (PySpark code)
5. How would you query the database to display the types of genres available? (PySpark code)
6. How could you create a bar graph using the genre data extracted from the table? (Use matplotlib)

Execution Steps to Follow:

1. Open the Azure dashboard console search for Databricks
2. Import the dataset from the public repository
3. Perform all the query respective to the question provided
4. Take screenshots of the query execution
5. Upload the code to the Github

____x____