Apache Spark 3 Fundamentals

Mohit Batra (linkedin.com/in/mohitbatra/)

Table of Contents

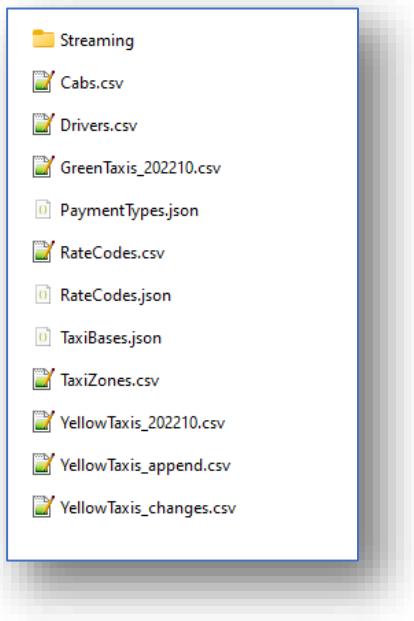
A. GitHub Repo

All code and data files are available in Exercise Files. Refer to following GitHub repo for additional instructions and information.

https://github.com/Crystal-Talks/ApacheSpark3Fundamentals

B. Included Files

1. Code folder
    a. Contains following:
        i. Shell commands
        ii. Jupyter Notebooks
        iii. PyCharm Project
        iv. Cloud Notebooks
2. DataFiles folder



C. Install Spark on macOS

Refer to installation instructions at following GitHub repo:

https://github.com/Crystal-Talks/ApacheSpark3Fundamentals

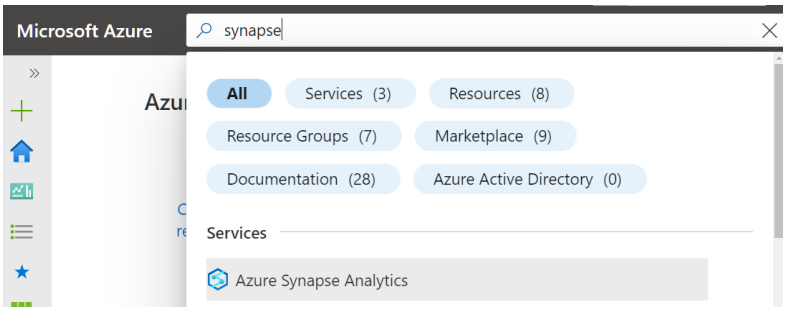
D. Prerequisites for Cloud Setup

1. Azure subscription
    https://azure.microsoft.com/en-in/free/

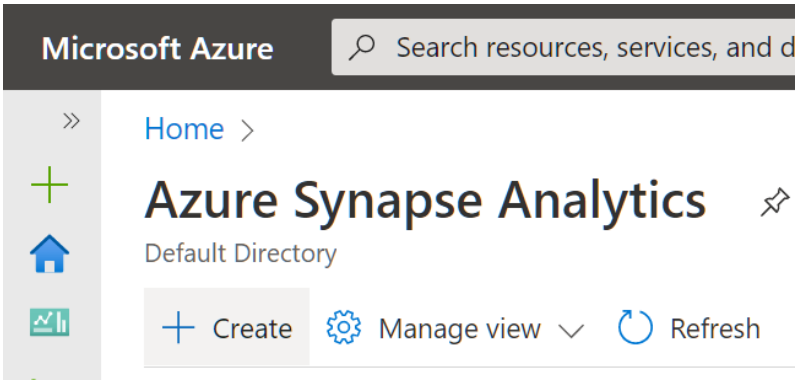
E. Azure Synapse Analytics: Steps to create workspace

1. Go to Azure portal (portal.azure.com)

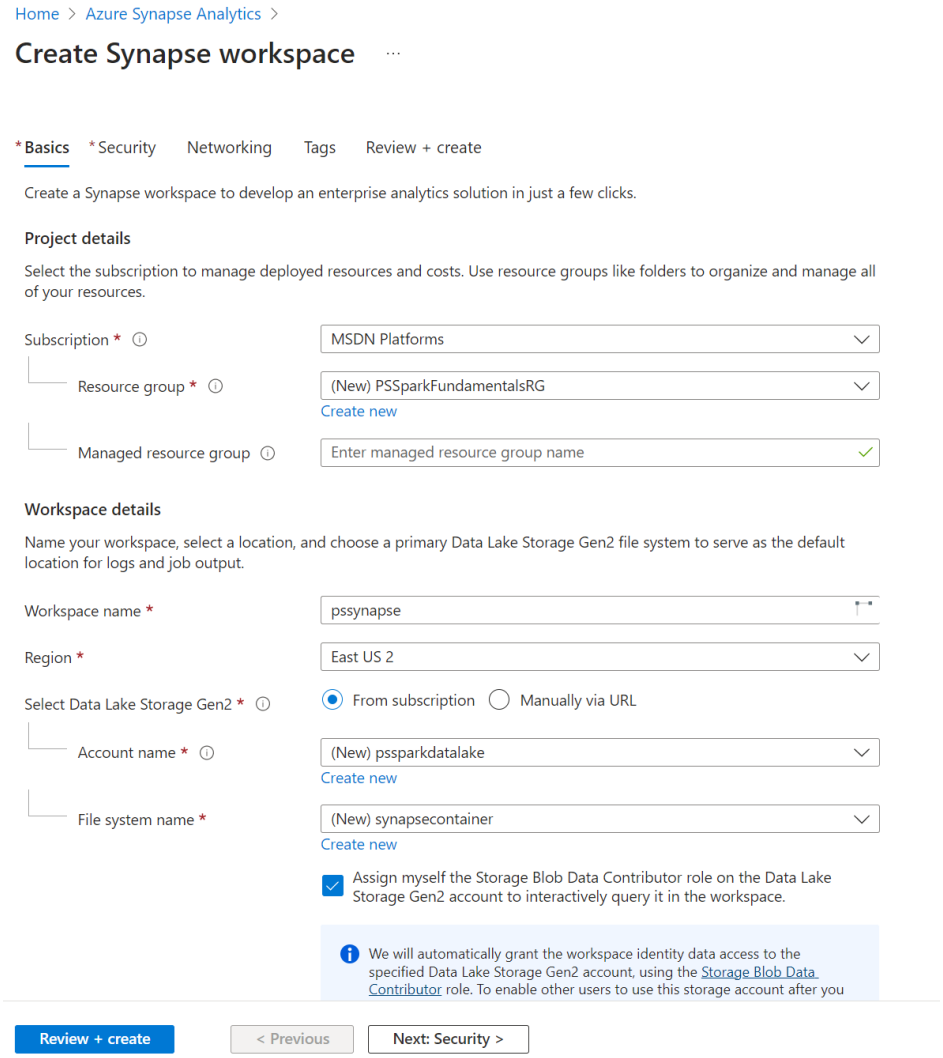
2. Search for Azure Synapse Analytics. Select it.



3. Create a new one.



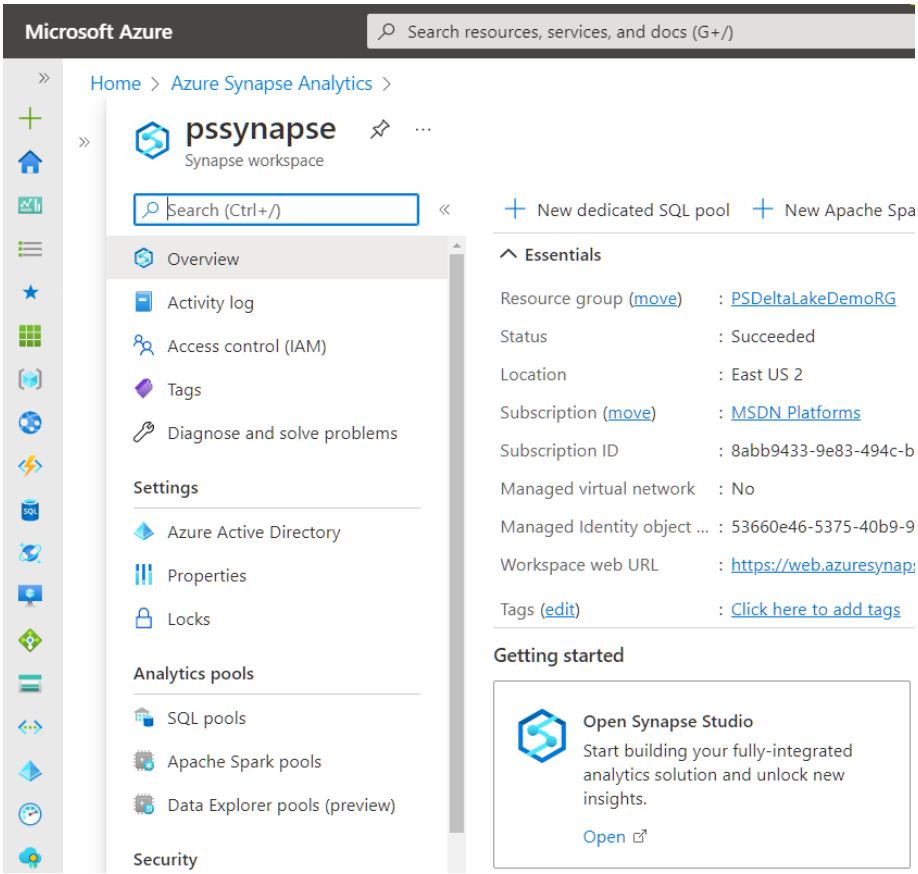
4. In **Basics** tab, provide property values:
    a. Select subscription and resource group (or create new)
    b. **Workspace name:** *Any unique value*
    c. **Region:** *Choose any*
    d. **Data Lake Gen2 account:** Create a new one by providing a unique name
    e. **File system name:** Create a new one – "synapsecontainer"



5. Click **Review + Create**.

6. Click **Create**.

7. Once Synapse instance is created, open the instance.



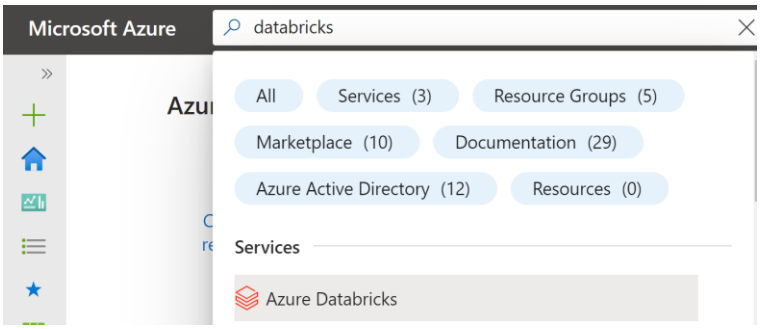
8. Click **Open** to launch Synapse studio.

9. Follow instructions in video to work with Synapse.

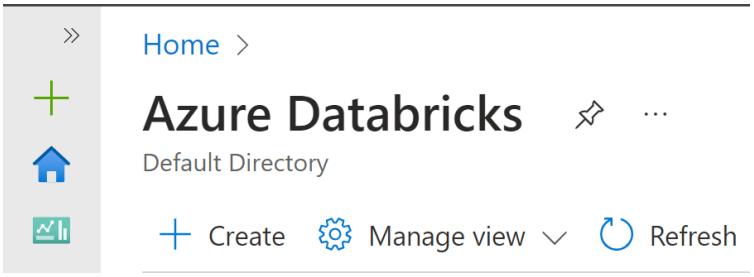
F. Azure Databricks: Steps to create workspace

1. Go to Azure portal (portal.azure.com)

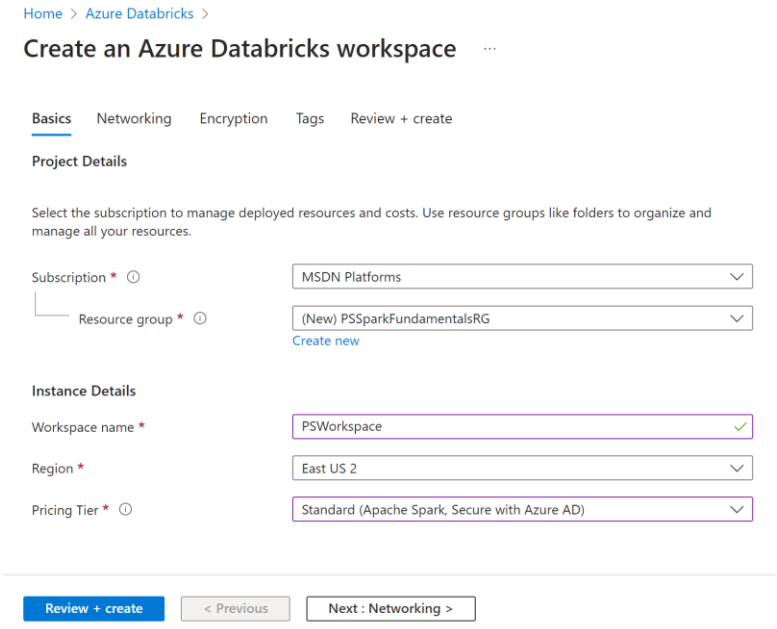
2. Search for Databricks. Select it.



3. Create a new one.



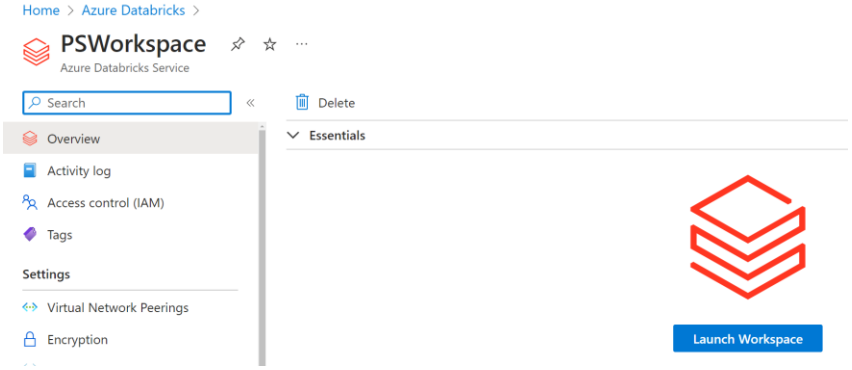
4. In **Basics** tab, provide property values:
    a. Select subscription and resource group (or create new)
    b. **Workspace name:** *Any unique value*
    c. **Region:** *Choose any*
    d. **Pricing Tier:** Standard



5. Click **Review + Create**.

6. Click **Create**.

7. Once Databricks instance is created, open the instance.



8. Click **Open** to launch Databricks workspace.

9. Follow instructions in video to work with Databricks.