

Apache Spark 3 Fundamentals

Mohit Batra ([linkedin.com/in/mohitbatra/](https://www.linkedin.com/in/mohitbatra/))

Table of Contents

Apache Spark 3 Fundamentals.....	1
A. GitHub Repo	2
B. Included Files	2
C. Install Spark on macOS.....	2
D. Prerequisites for Cloud Setup	2
E. Azure Synapse Analytics: Steps to create workspace	3
F. Azure Databricks: Steps to create workspace	5

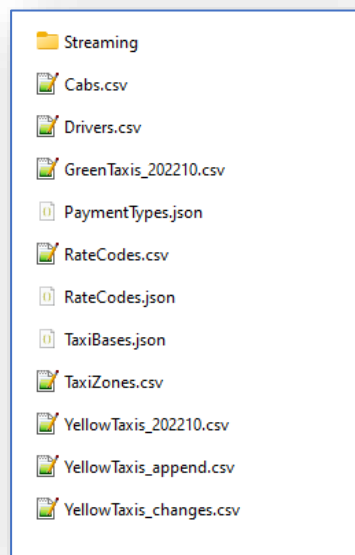
A. GitHub Repo

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:

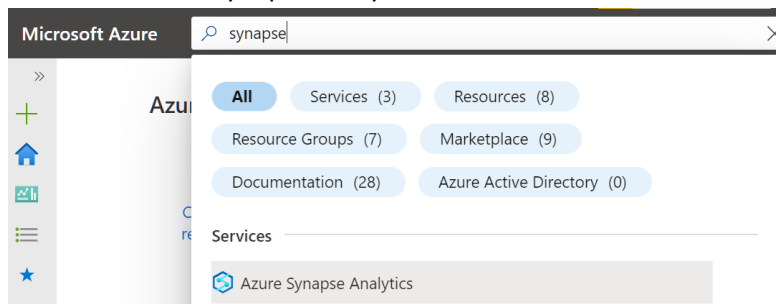
<https://github.com/Crystal-Talks/ApacheSpark3Fundamentals>

D. Prerequisites for Cloud Setup

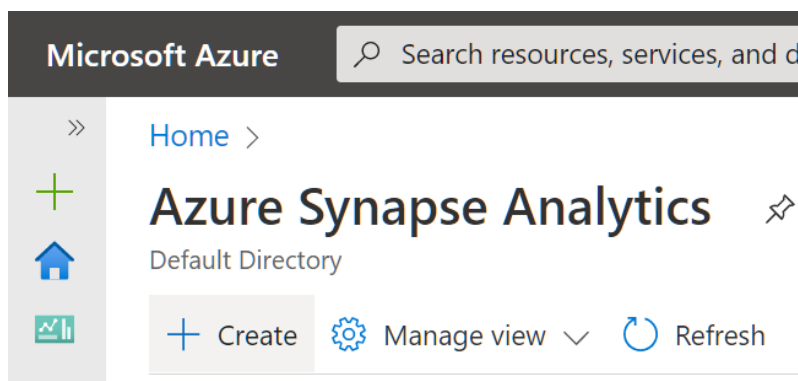
1. Azure subscription
<https://azure.microsoft.com/en-in/free/>

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”

Create Synapse workspace ...

*** Basics** * Security Networking Tags Review + create

Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all of your resources.

Subscription *	MSDN Platforms
Resource group *	(New) PSSparkFundamentalsRG
	Create new
Managed resource group	Enter managed resource group name

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

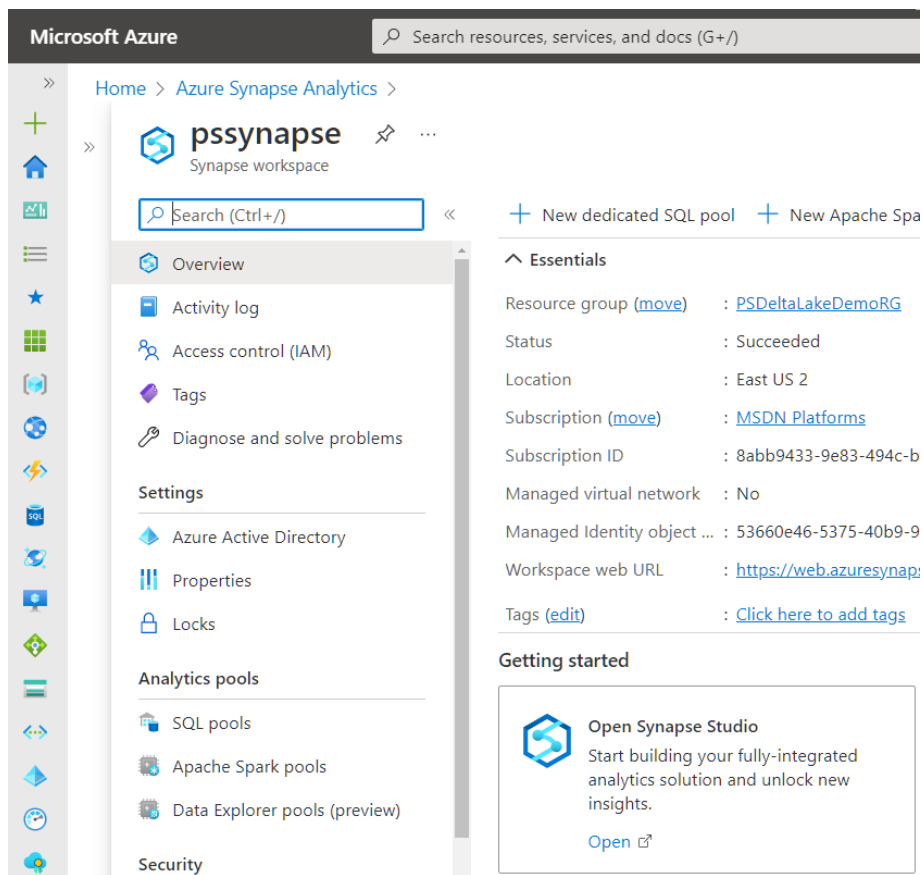
Workspace name *	pssynapse
Region *	East US 2
Select Data Lake Storage Gen2 *	<input checked="" type="radio"/> From subscription <input type="radio"/> Manually via URL
Account name *	(New) pssparkdatalake
	Create new
File system name *	(New) synapsecontainer
	Create new

☒ Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.

i We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the [Storage Blob Data Contributor](#) role. To enable other users to use this storage account after you

[Review + create](#) [< Previous](#) [Next: Security >](#)

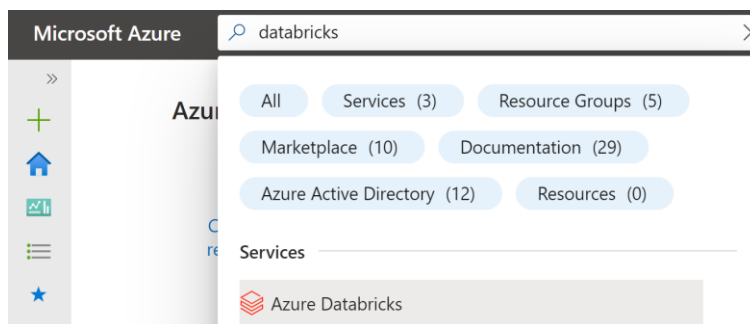
5. Click **Review + Create**.
6. Click **Create**.
7. Once Synapse instance is created, open the instance.



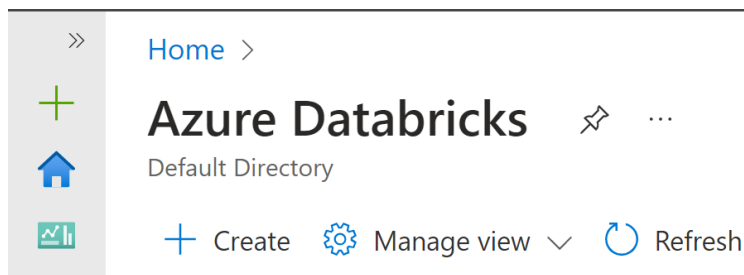
8. Click **Open** to launch Synapse studio.
9. Follow instructions in video to work with Synapse.

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

Home > Azure Databricks >

Create an Azure Databricks workspace

Basics Networking Encryption Tags Review + create

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ MSDN Platforms

Resource group * ⓘ (New) PSSparkFundamentalsRG
[Create new](#)

Instance Details

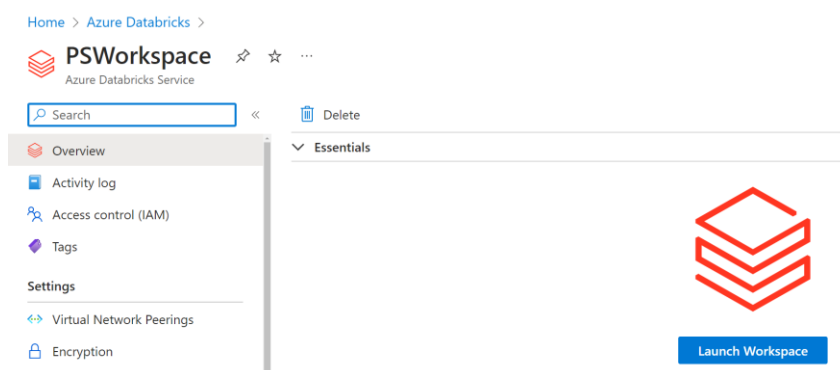
Workspace name * PSWorkspace ✓

Region * East US 2

Pricing Tier * ⓘ Standard (Apache Spark, Secure with Azure AD)

[Review + create](#) < Previous Next : Networking >

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.