|  |  |
| --- | --- |
|  |  |
|  | BigData Lab  Databricks  using Microsoft |

Contents

[Overview 3](#_Toc513494328)

[What is Azure DataBricks ? 4](#_Toc513494329)

[Create a Spark cluster 6](#_Toc513494330)

[Running Spark Jobs 8](#_Toc513494331)

[Terms of Use 12](#_Toc513494332)

|  |
| --- |
| Overview |

### Summary

This lab explores Azure Databricks and shows how to use it for various purposes and products.

### Learning Objectives

Upon completing this lab, you will have hands-on experience with the following functions and concepts related to Big Data:

* Setting up your Azure Databricks Account
* Creating A Cluster
* Working With Spark API
* Working With SQL Api
* Working with Streaming Data
* Working With Dashboards

### Lab Requirements/Prerequisites

* A Microsoft account is required to obtain a Databricks Cluster. Since you will need minimum 12 cores, free and azurepass sıbscriptions will not work. If you don’t already have a Microsoft account, you can obtain one for free from the following link:

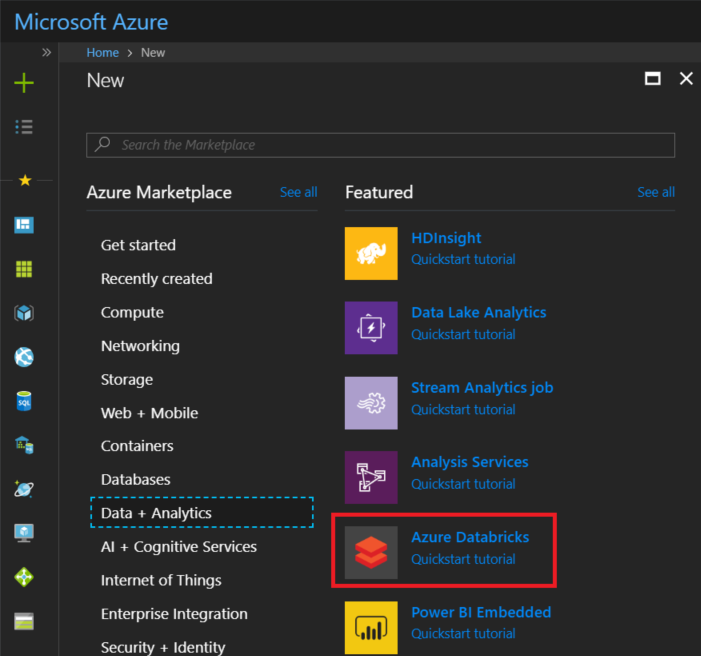
<https://www.microsoft.com/en-us/account/default.aspx>

|  |
| --- |
| What is Azure Databricks ? |

### What is Azure Databricks?

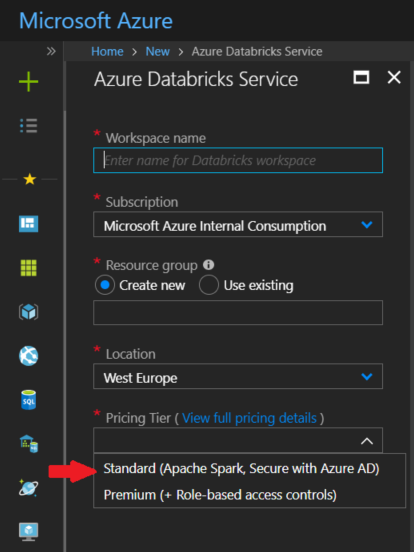
Azure Databricks is a fast, easy, and collaborative Apache Spark-based analytics platform. The diagram below explains its core pillars and key differentiators. In this lab we’ll explore Data Engineering capabilities of Azure Databricks:.

1. Open Google Chrome Browser (Azure Databricks visualizations work better with Chrome)
2. Log in to the Azure portal at http://portal.azure.com using your @microsoft.com address. It is important that your account has an AAD identity as Azure Databricks has built-in security, and needs the AAD token. o Do not use the Ready subscriptions, if you want to persist the assets.
3. In the Azure portal, click +, click Data + Analytics, and then click Azure Databricks (Preview). Once the **Cognitive Services** homepage is loaded, select the **APIs** drop-down me



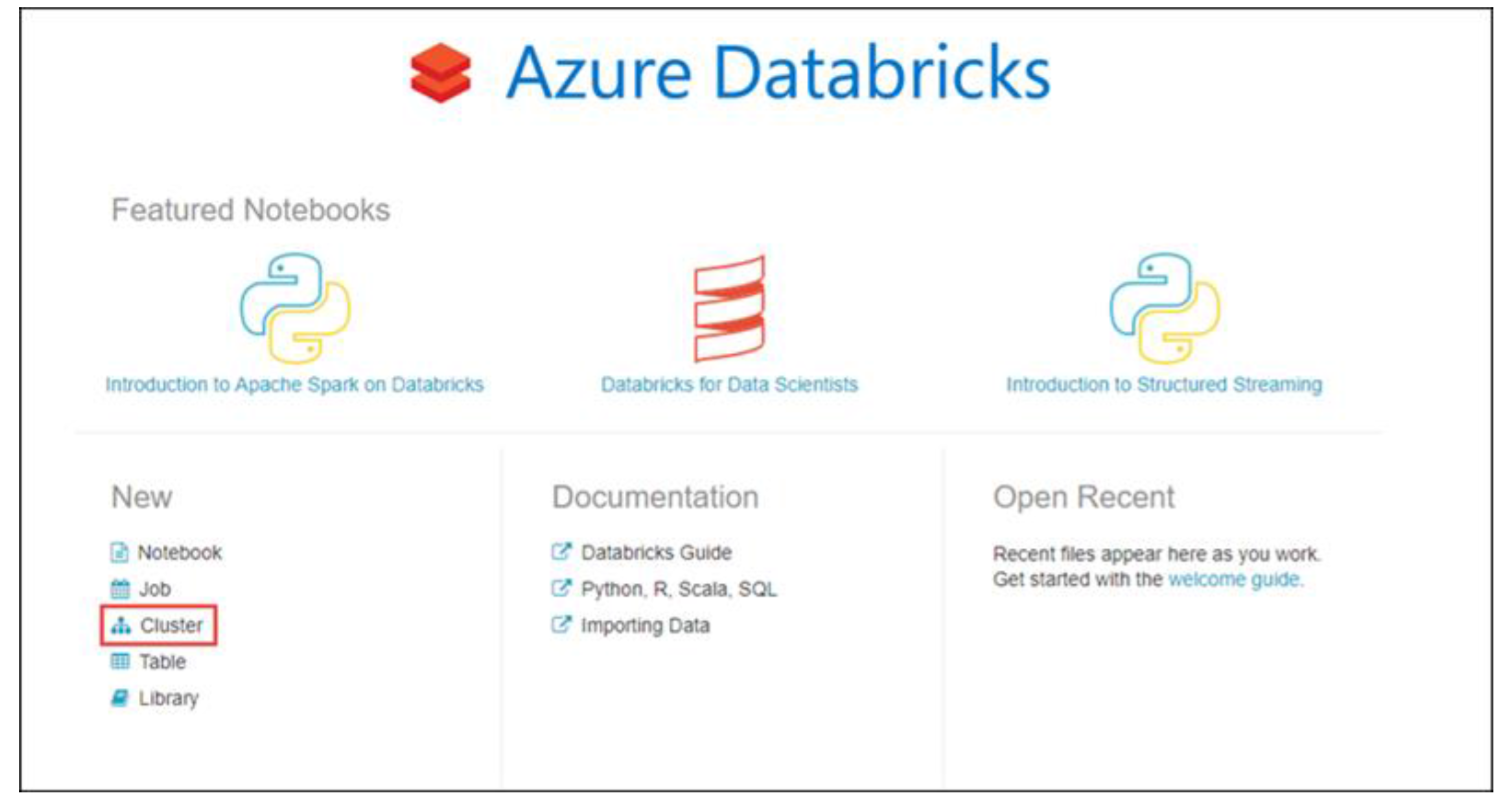
1. Under Azure Databricks, click Create.
2. Under Azure Databricks Service, provide the following values:
   * For Workspace name, provide a name for your Databricks workspace.
   * For Subscription, from the drop-down, select your Azure subscription.
   * For Resource group, specify whether you want to create a new resource group or use an existing one. A resource group is a container that holds related resources for an Azure solution. And it is a convenient way to remove the resources with its all peripherals.
   * For Location, select the closest region. For other available regions, see Azure services available by region.

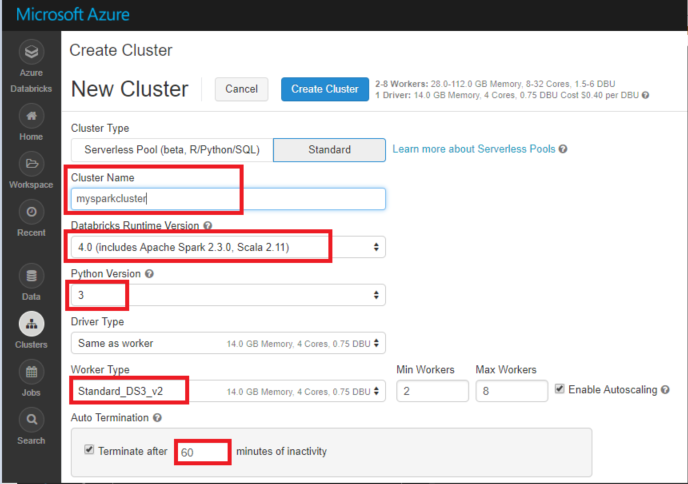
6. Click Create.



|  |
| --- |
| Create a Spark cluster |

**Create a Spark cluster in Databricks**

1. In the Azure portal, go to the Databricks workspace that you created, and then click Initialize Workspace.
2. You are redirected to the Azure Databricks portal. From the portal, click Cluster. 
3. In the New cluster page, provide the values to create a cluster.



* Enter a name for the cluster.
* Make sure you select the Terminate after \_ minutes of activity checkbox. Provide a duration (in minutes) to terminate the cluster, if the cluster is not being used.
* Accept all other default values.
* Click Create cluster.

Once the cluster is in running state, you can attach notebooks to the cluster and run Spark jobs.

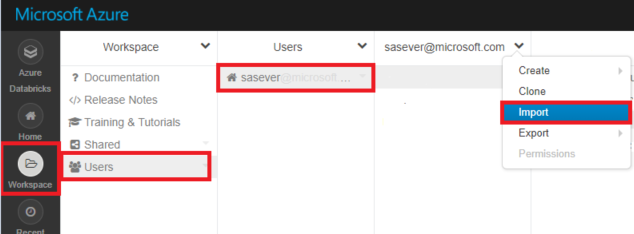
While the cluster is creating (it takes about 5 mins.), feel free to read into this clog post for a technical overview of Azure Databricks: https://azure.microsoft.com/en-us/blog/a-technical-overview-of-azure-databricks/

|  |
| --- |
| Running Spark Jobs |

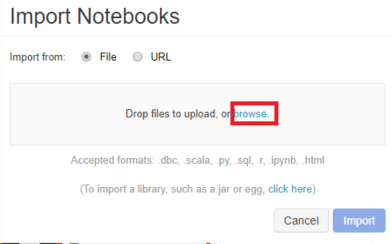
**Running Spark Jobs**

Perform the following steps to create a notebook in Databricks, configure the notebook to read data from an Azure Blob storage account, and then run a Spark SQL job on the data.

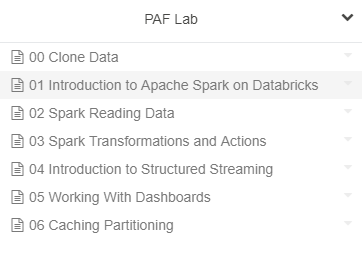
1. In the left pane, click Workspace. From the Workspace drop-down, click Users, select your username, and then click on the drop down arrow next to your name, and select ‘Import’..



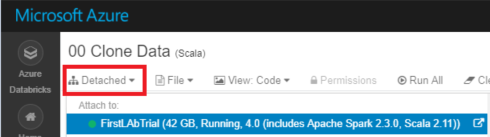
1. In the Import Notebooks window, select ‘browse’.



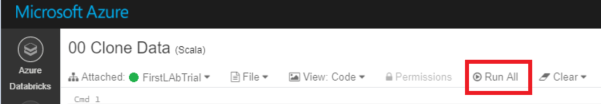
1. In the browse window, navigate to the directory you exported the Workshop Documents and select the PAF Lab.dbc archive file.
2. Back on the Import Notebooks window, click on ‘Import’.
3. Once imported, you should be able to see the following notebooks:



1. Click on the ’00 Clone Data’ notebook to open it.
2. To attach the notebook to a cluster, select the ‘Detached’ icon on the top bar.

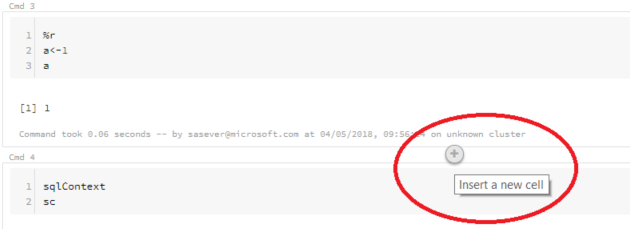


1. Select the cluster created in the earlier steps, in the ‘Attach to’ options.
2. Click on ‘Run all’ option from the top bar:



1. This notebook, prepares data for exercises further. Please look into the notebook to understand the copy syntax. This is optional, as you can connect to multiple data sources, in a similar way.
2. Once this notebook executes successfully, go back to the imported ‘PAF Lab’ folder and open the ’01 Introduction to Apache Spark on Databricks’ Notebook.

* All the exercises and steps are now part of the notebook itself.
* Although when there is a question the answers are usually given within the notebook, try to write them by yourself first.
* To try that you can always add a new code block by clicking insert new cell, which appears when your cursor comes between cells.



* You can execute cell by cell by using ‘Shift + Enter’ after selecting the cell in the notebook.
* Keep following the instructions in the notebook to proceed with the exercise.

12. Once this section of exercises are completed successfully, go back to the imported ‘PAF Lab’ folder and open the ’02 Spark Reading Data’ notebook, and use the same instructions.

13. Go back to the imported ‘PAF Lab’ folder and open the ’03 Spark Transformations and Actions’ notebook, and use the same instructions.

14. Go back to the imported ‘PAF Lab’ folder and open the ’04 Introduction to Structured Streaming’ notebook, and use the same instructions.

15. Go back to the imported ‘PAF Lab’ folder and open the ’05 Working With Dashboards notebook, and use the same instructions.

16. Go back to the imported ‘PAF Lab’ folder and open the ’06 Caching and Partitioning’ notebook, and use the same instructions.

|  |
| --- |
| Terms of Use |

© 2018 Microsoft Corporation. All rights reserved.

By using this Hands-on Lab, you agree to the following terms:

The technology/functionality described in this Hands-on Lab is provided by Microsoft Corporation in a “sandbox” testing environment for purposes of obtaining your feedback and to provide you with a learning experience. You may only use the Hands-on Lab to evaluate such technology features and functionality and provide feedback to Microsoft.  You may not use it for any other purpose. You may not modify, copy, distribute, transmit, display, perform, reproduce, publish, license, create derivative works from, transfer, or sell this Hands-on Lab or any portion thereof.

COPYING OR REPRODUCTION OF THE HANDS-ON LAB (OR ANY PORTION OF IT) TO ANY OTHER SERVER OR LOCATION FOR FURTHER REPRODUCTION OR REDISTRIBUTION IS EXPRESSLY PROHIBITED.

THIS HANDS-ON LAB PROVIDES CERTAIN SOFTWARE TECHNOLOGY/PRODUCT FEATURES AND FUNCTIONALITY, INCLUDING POTENTIAL NEW FEATURES AND CONCEPTS, IN A SIMULATED ENVIRONMENT WITHOUT COMPLEX SET-UP OR INSTALLATION FOR THE PURPOSE DESCRIBED ABOVE.  THE TECHNOLOGY/CONCEPTS REPRESENTED IN THIS HANDS-ON LAB MAY NOT REPRESENT FULL FEATURE FUNCTIONALITY AND MAY NOT WORK THE WAY A FINAL VERSION MAY WORK.  WE ALSO MAY NOT RELEASE A FINAL VERSION OF SUCH FEATURES OR CONCEPTS.  YOUR EXPERIENCE WITH USING SUCH FEATURES AND FUNCITONALITY IN A PHYSICAL ENVIRONMENT MAY ALSO BE DIFFERENT.

**FEEDBACK**.  If you give feedback about the technology features, functionality and/or concepts described in this Hands-on Lab to Microsoft, you give to Microsoft, without charge, the right to use, share and commercialize your feedback in any way and for any purpose.  You also give to third parties, without charge, any patent rights needed for their products, technologies and services to use or interface with any specific parts of a Microsoft software or service that includes the feedback.  You will not give feedback that is subject to a license that requires Microsoft to license its software or documentation to third parties because we include your feedback in them.  These rights survive this agreement.

MICROSOFT CORPORATION HEREBY DISCLAIMS ALL WARRANTIES AND CONDITIONS WITH REGARD TO THE HANDS-ON LAB, INCLUDING ALL WARRANTIES AND CONDITIONS OF MERCHANTABILITY, WHETHER EXPRESS, IMPLIED OR STATUTORY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT.  MICROSOFT DOES NOT MAKE ANY ASSURANCES OR REPRESENTATIONS WITH REGARD TO THE ACCURACY OF THE RESULTS, OUTPUT THAT DERIVES FROM USE OF THE VIRTUAL LAB, OR SUITABILITY OF THE INFORMATION CONTAINED IN THE VIRTUAL LAB FOR ANY PURPOSE.

DISCLAIMER

This lab contains only a portion of the features and enhancements in Microsoft Cognitive Services. Some of the features might change in future releases of the product.