# Databricks

This is a company that was founded by the original creators of Apache Spark Databricks makes use of Apache Spark to provide a Unified Analytics platform.

# what is cluster in databrick?

In Databricks, a cluster is a set of computation resources. It's essentially virtual machines. If you have a multi node cluster, then you have a driver node and worker nodes that helps you execute tasks.

Databricks makes a distinction between **all-purpose clusters** and **job clusters**.

Use all-purpose clusters to analyse data collaboratively using interactive notebooks, and you use job clusters to run fast and robust automated jobs.

## All-purpose cluster

So, all-purpose clusters can be shared by multiple users. They typically use to run notebooks. They remain active until you terminate them. And this is created manually, as you can see.

## Job cluster

Job clusters are created when you create a job. These clusters are terminated automatically after the job is completed.

# Pricing

# Create Azure databricks service.

Go to Azure portal and in search box type databricks, click on create and fill necessary information, here workspace name should be unique.

## Example

As part of POC, we have created Azure databricks service.

Subscription: -

Resource Group: -

Azure databricks service: -

# Create Cluster in Azure databricks.

Go to the compute section select the cluster type you need to create then click on “Create Cluster” button. Fill in the necessary detail as per your requirement and click on create “Create Cluster” button. It will take a few minutes to spin the cluster.

## Example

As part of POC, we have created cluster Azure databricks.

Subscription: -

Azure databricks cluster: -

# Mounting ADLS to DBFS

* Create the azure AD app for the service principal and create the secret.
* Create Azure Data Lake Storage Gen2.
* Assign and Storage Blob Data Contributor role to azure ad app created in step one.
* Create Azure key vault to store the secrets of service principle in same resource group in which we created the Azure databricks service.
* Go to the Azure databricks URL in URL remove everything after the # in URL and type “secrets/createScope” (case sensitive) and hit enter. You will be taken to create secret scope page, enter the required necessary details and click on create to create secret scope.
* Create the python notebook and use following code to mount the ADLS to Azure databricks.



In this code you need to replace <container\_name>, <account\_name> and <mount\_point>.

# How to create use Databricks notebooks

* To access your databricks workspace, you need to Search the databricks service in azure portal and click on it to Launch Workspace.
* Click on workspace, under shared account and a user account. So, we will create the shared notebook. Right click on Shared and click on folder to create a folder (Folder Name - Test).

A screenshot of a computer

Description automatically generated

* Now right click on previously created folder and create a notebook. Enter the name (Name – Test Notebook) of the notebook, the Language, default language would be python we will keep it same. Select the cluster which we have created. And click on create.