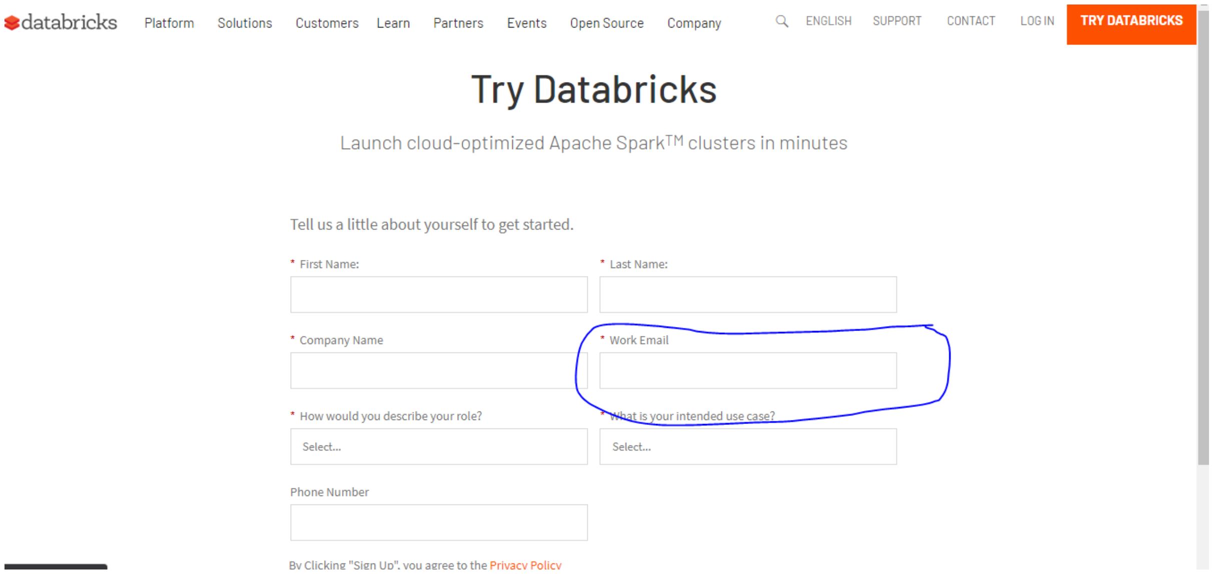
DataBricks Exercise Manual

Exercise 1

Creating your DataBricks Community Workspace

1. Sign up DataBricks community using the below URL
   https://databricks.com/try-databricks?_ga=2.231253720.636689393.1585744376-1730487577.1573900075

2. Please provide your Work Email ID while signing up



Try Databricks

Launch cloud-optimized Apache Spark™ clusters in minutes

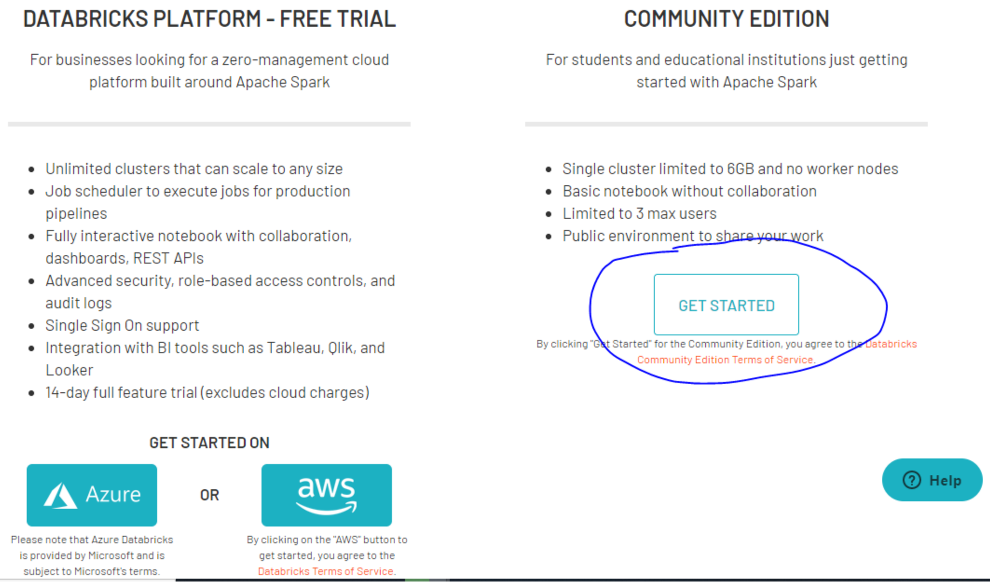
Tell us a little about yourself to get started.

* First Name:
* Last Name:
* Company Name
* Work Email
* How would you describe your role? Select...
* What is your intended use case? Select...

Phone Number

By Clicking "Sign Up", you agree to the Privacy Policy

3. Once you sign up, you will be taken to the page in which you have to select Free Trial or Community Edition. Please select "Get Started" option under the community Edition.



DATABRICKS PLATFORM - FREE TRIAL

For businesses looking for a zero-management cloud platform built around Apache Spark

- Unlimited clusters that can scale to any size
- Job scheduler to execute jobs for production pipelines
- Fully interactive notebook with collaboration, dashboards, REST APIs
- Advanced security, role-based access controls, and audit logs
- Single Sign On support
- Integration with BI tools such as Tableau, Qlik, and Looker
- 14-day full feature trial (excludes cloud charges)

GET STARTED ON

Please note that Azure Databricks is provided by Microsoft and is subject to Microsoft's terms.

By clicking on the "AWS" button to get started, you agree to the Databricks Terms of Service.

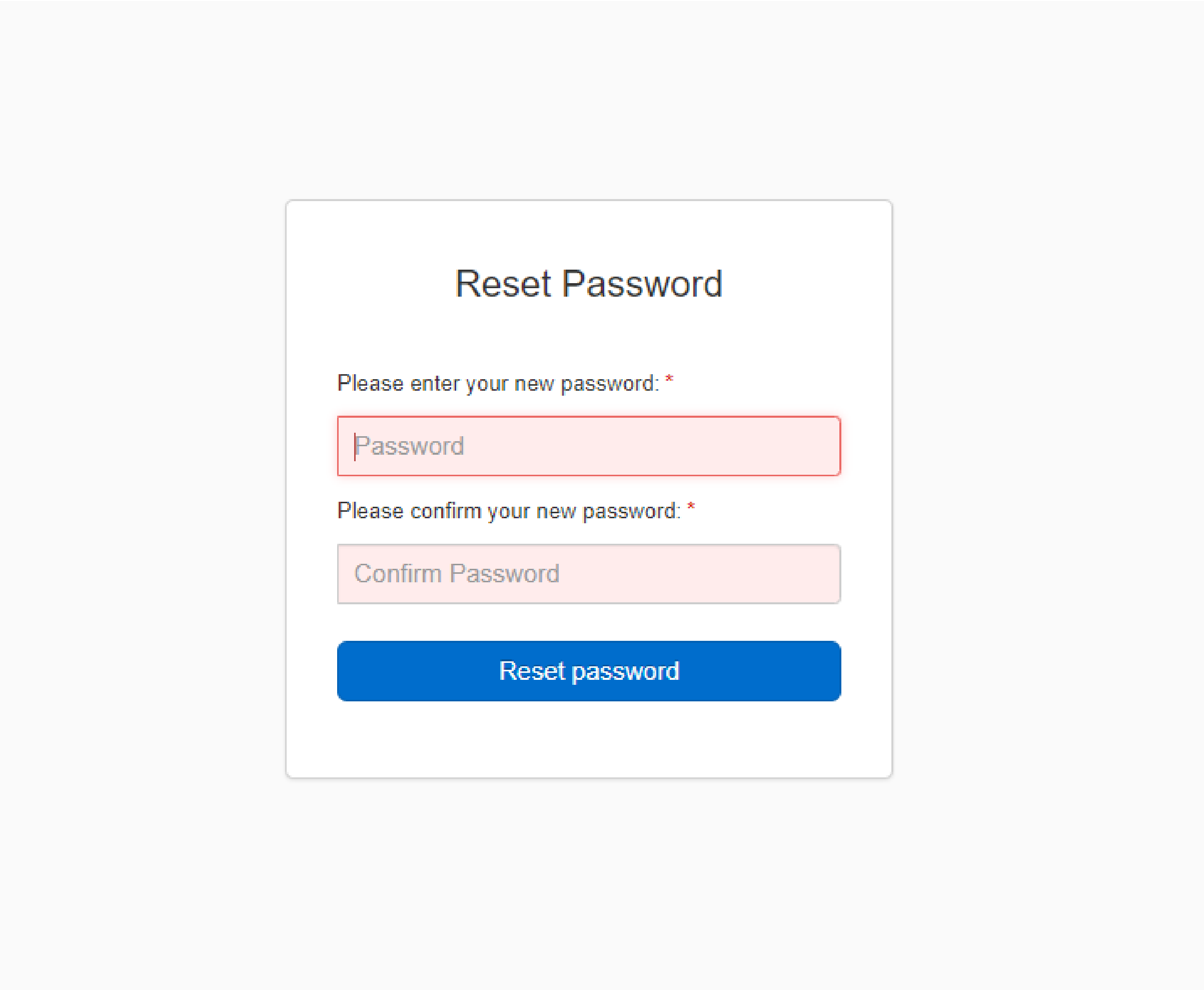
COMMUNITY EDITION

For students and educational institutions just getting started with Apache Spark

- Single cluster limited to 6GB and no worker nodes
- Basic notebook without collaboration
- Limited to 3 max users
- Public environment to share your work

By clicking "Get Started" for the Community Edition, you agree to the Databricks Community Edition Terms of Service.

4. After you clicked “Get Started” within few minutes, you would be receiving an email in your Inbox for verification. Please verify the link provided in the Email. You will be directed to a page in which you have to provide your password for login.



Reset Password

Please enter your new password: *

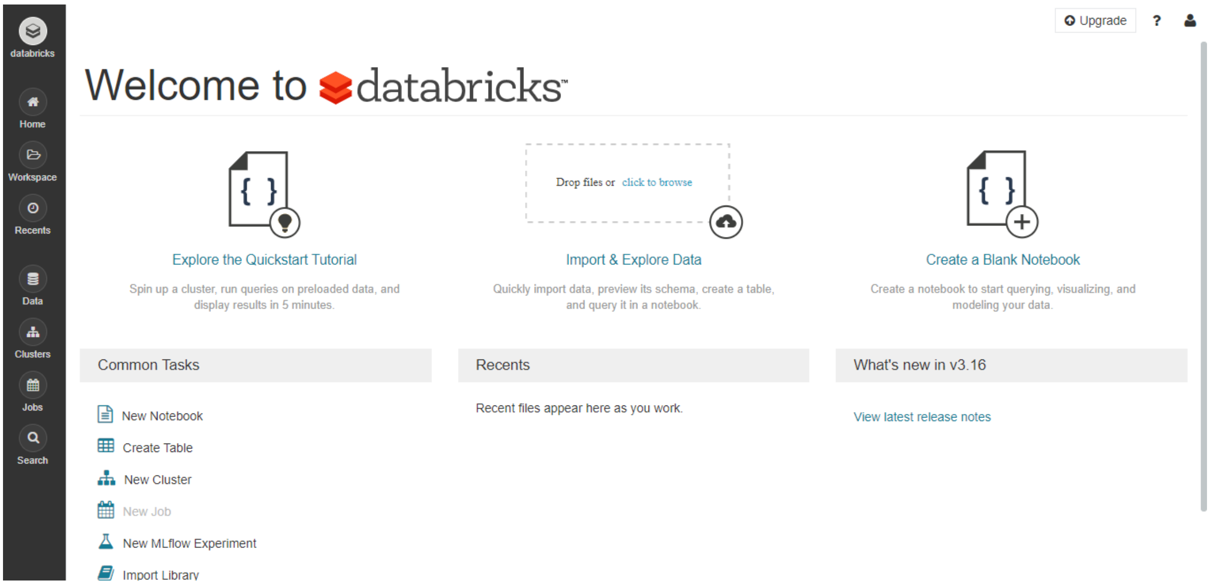
Password

Please confirm your new password: *

Confirm Password

Reset password

5. After providing your password, you will be taken to your DataBricks Workspace



Welcome to databricks

Explore the Quickstart Tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Drop files or click to browse

Import & Explore Data

Quickly import data, preview its schema, create a table, and query it in a notebook.

Create a Blank Notebook

Create a notebook to start querying, visualizing, and modeling your data.

Common Tasks

- New Notebook
- Create Table
- New Cluster
- New Job
- New MLflow Experiment
- Import Library

Recents

Recent files appear here as you work.

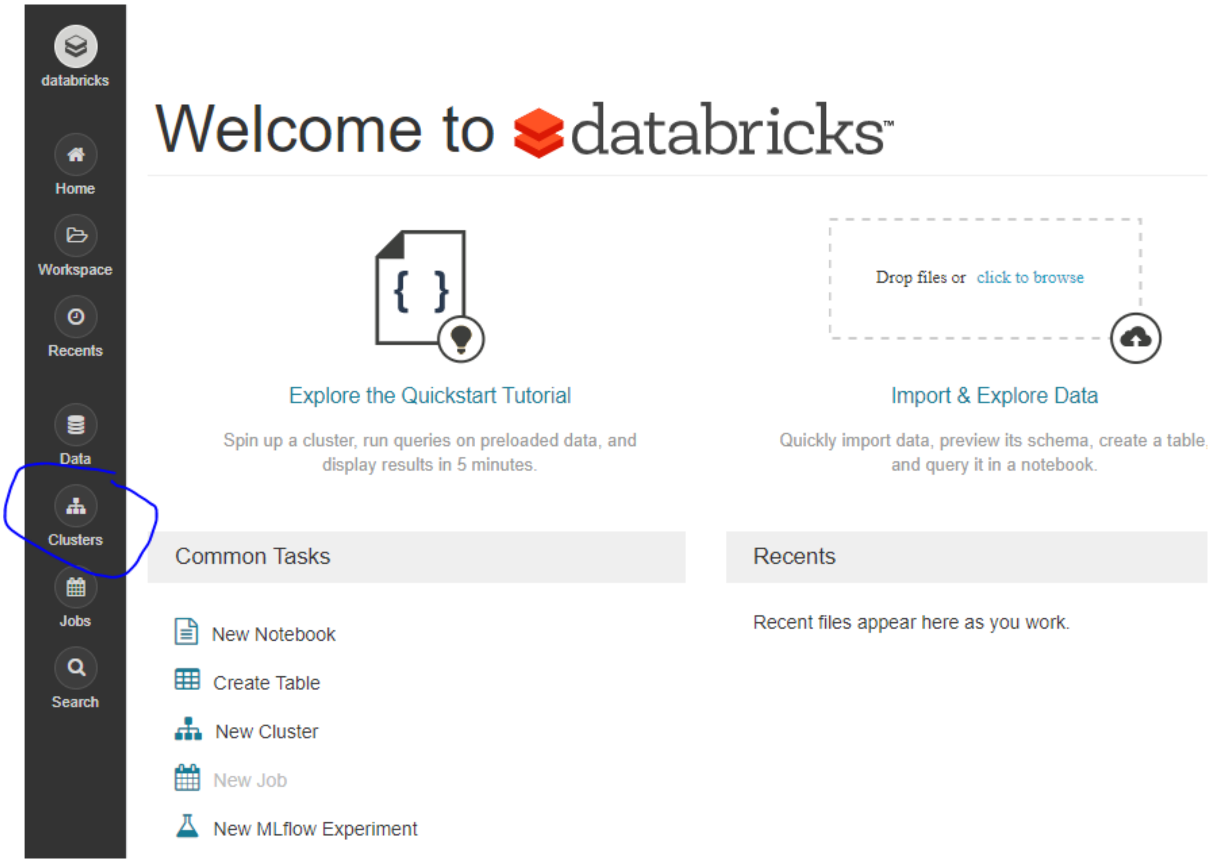
What's new in v3.16

View latest release notes

Exercise 2

Creating a cluster in your DataBricks Workspace

1. Click on Cluster option at the left side of your DataBricks Workspace



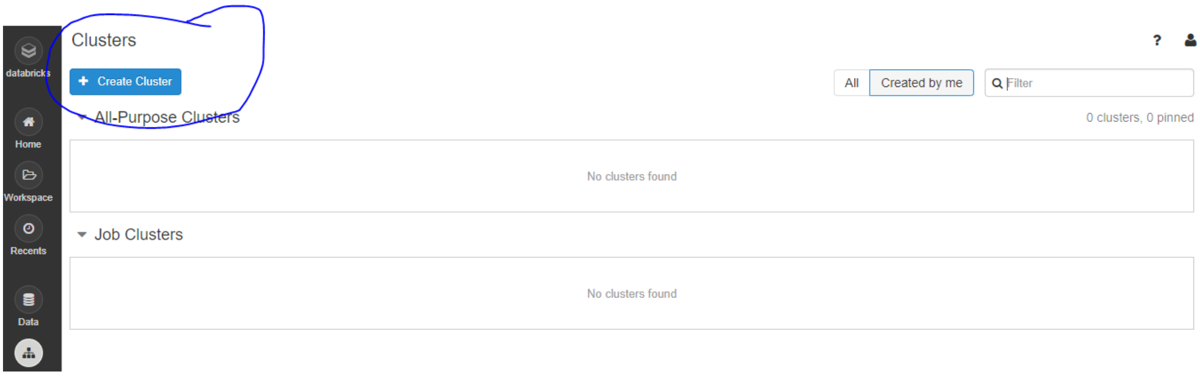
Welcome to databricks

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Quickly import data, preview its schema, create a table, and query it in a notebook.

Recent files appear here as you work.

2. You will find the list of available cluster. Currently you will have no cluster available, Please click on “Create Cluster” option available at the top.

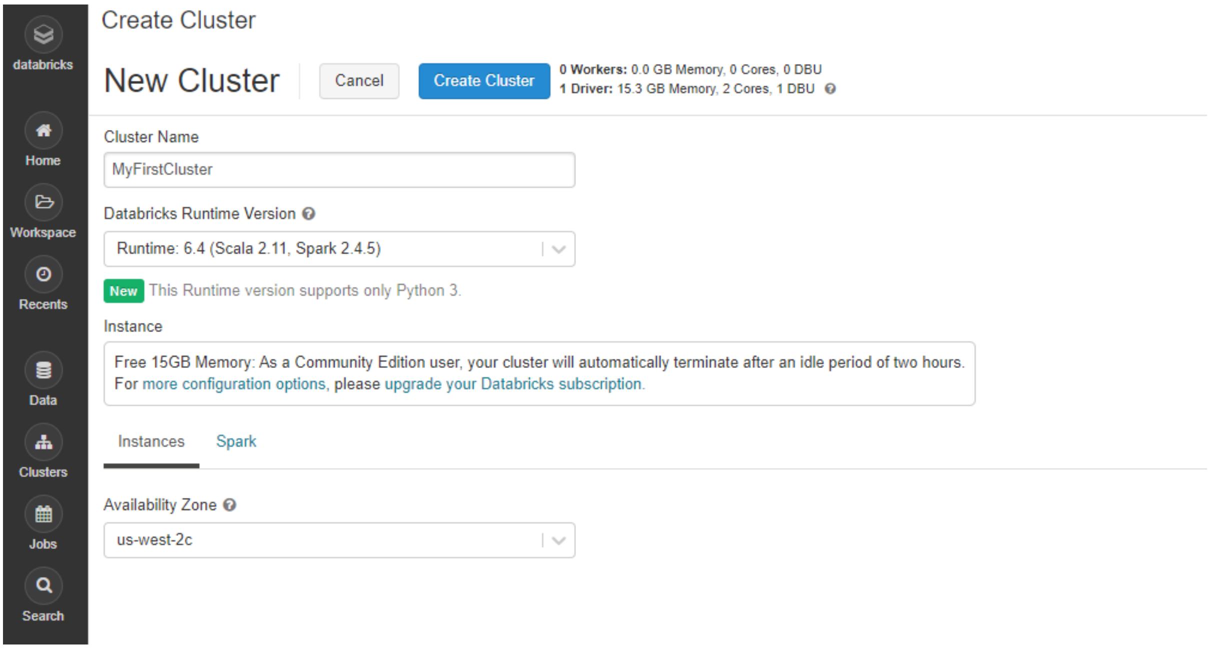


No clusters found

No clusters found

0 clusters, 0 pinned

3. Then you need to provide the cluster name. Make sure the Runtime is 6.4 and you will be noticing the cluster size is 1 driver with 15.3GB of RAM, 2 Cores and 1 DataBricks Unit. Finally click on create cluster button.



Create Cluster

New Cluster

0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU

Cluster Name: MyFirstCluster

Databricks Runtime Version: Runtime: 6.4 (Scala 2.11, Spark 2.4.5)

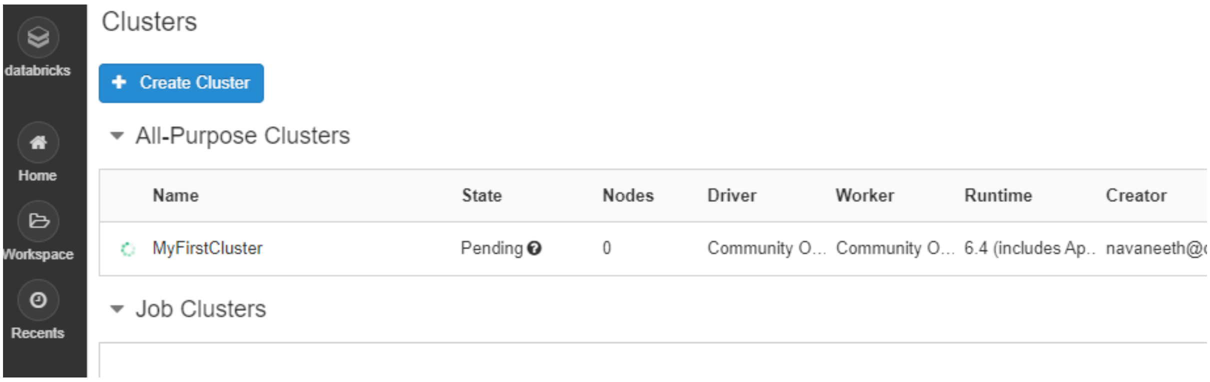
New This Runtime version supports only Python 3.

Instance

Free 15GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours.
For more configuration options, please upgrade your Databricks subscription.

Availability Zone: us-west-2c

4. Your cluster will be in pending state for some time and then it will get running status

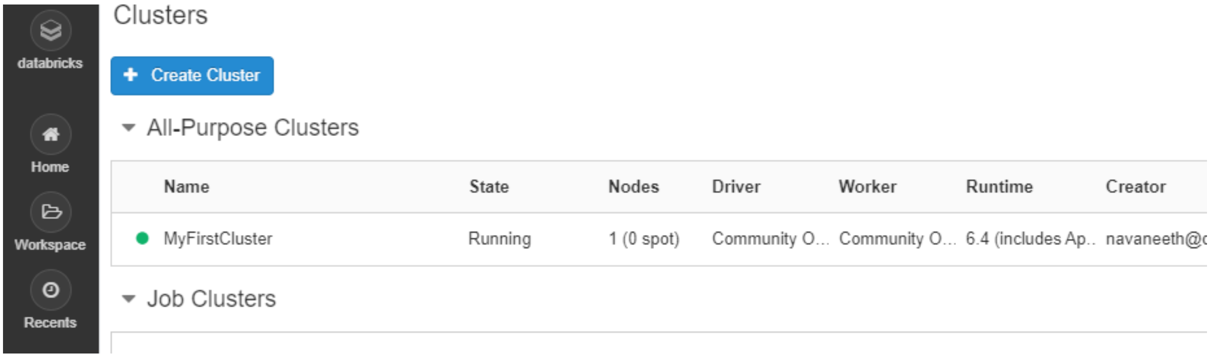


Clusters

All-Purpose Clusters

Name	State	Nodes	Driver	Worker	Runtime	Creator
MyFirstCluster	Pending	0	Community O...	Community O...	6.4 (includes Ap..	navaneeth@c

Job Clusters



Clusters

All-Purpose Clusters

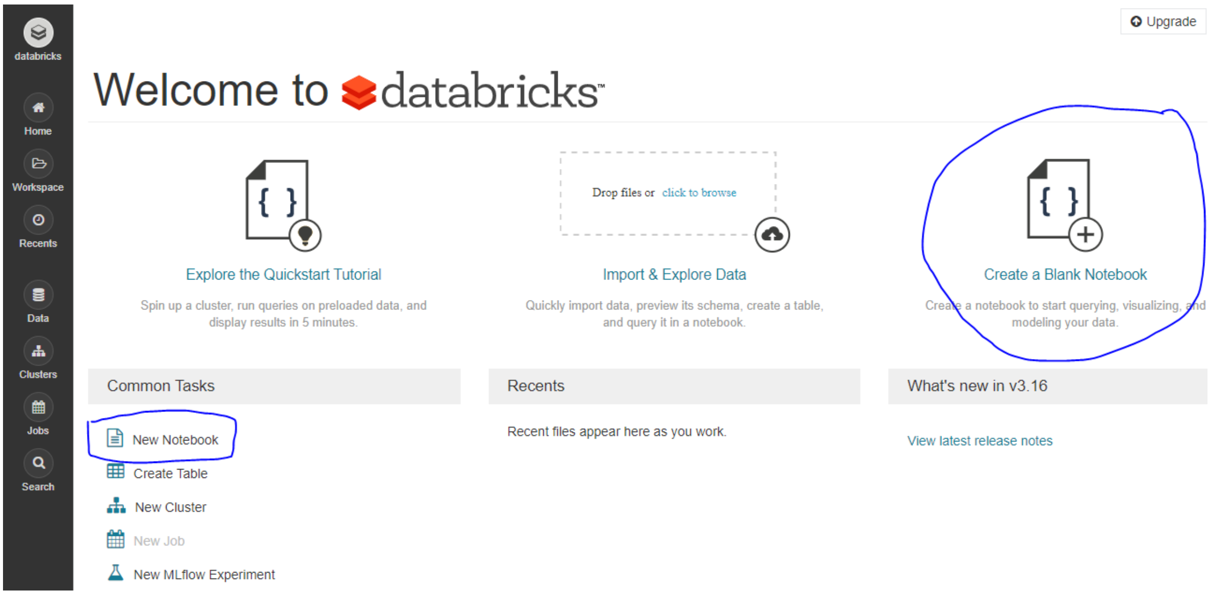
Name	State	Nodes	Driver	Worker	Runtime	Creator
● MyFirstCluster	Running	1 (0 spot)	Community O...	Community O...	6.4 (includes Ap..	navaneeth@c

Job Clusters

Exercise 3

Creating a notebook and fundamentals of notebook in DataBricks Workspace

1. On the DataBricks home page, you can use either, New Notebook options or Create a Blank Notebook option to create a notebook



Welcome to databricks

Explore the Quickstart Tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Import & Explore Data

Quickly import data, preview its schema, create a table, and query it in a notebook.

Create a Blank Notebook

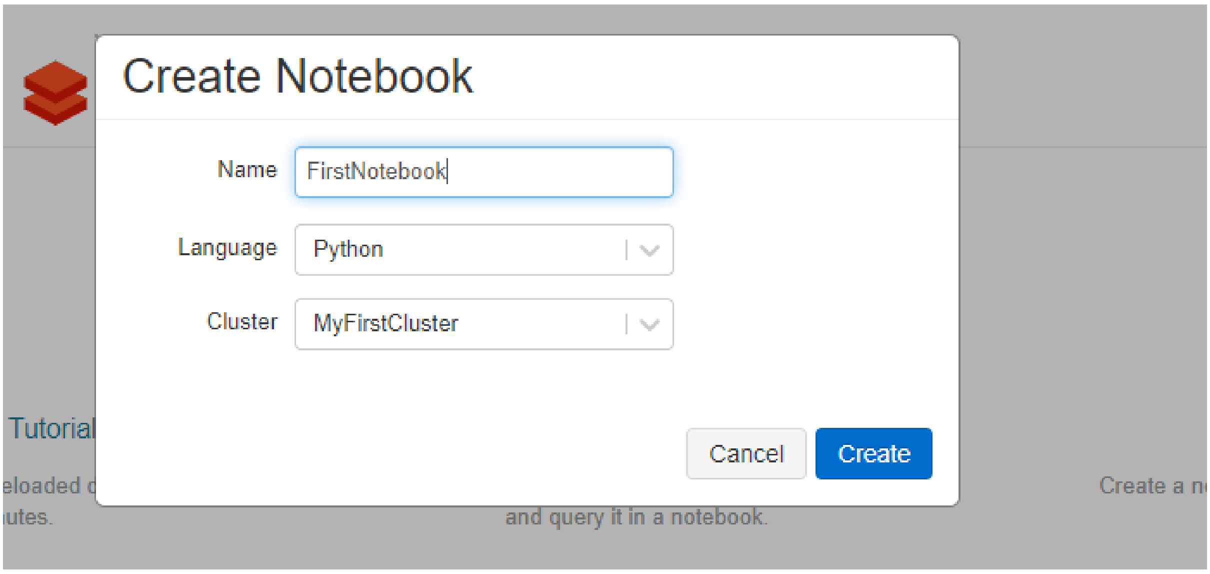
Create a notebook to start querying, visualizing, and modeling your data.

Recent files appear here as you work.

2. Provide a name for the notebook, Select Python, Scala, SQL or R based upon the execution environment and also select eh cluster.

   For current practice select “Python” and the cluster you have created recently.

   Then click on “Create” Option.



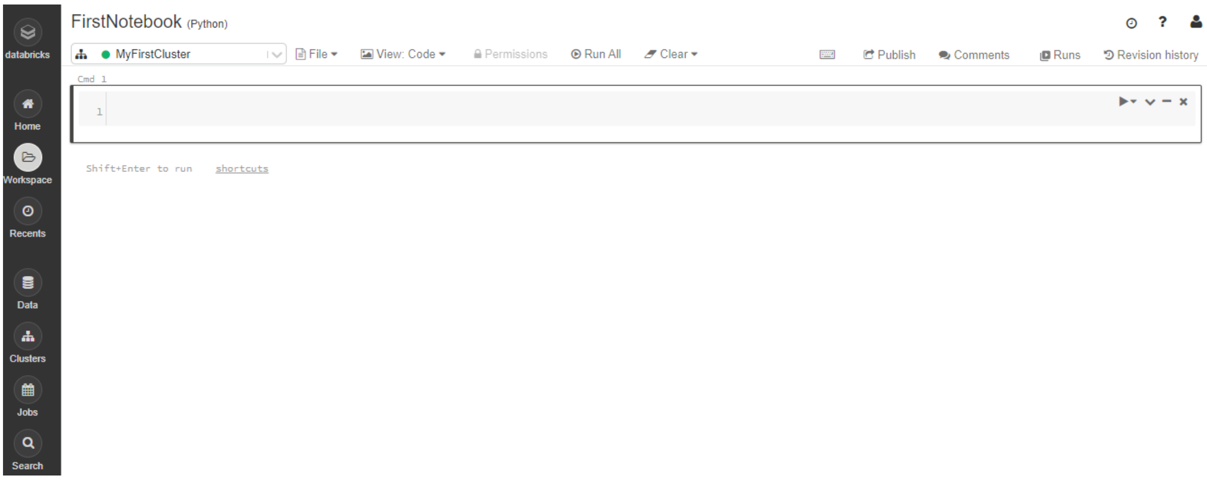
Create Notebook

Name: FirstNotebook

Language: Python

Cluster: MyFirstCluster

3. An empty notebook will be created



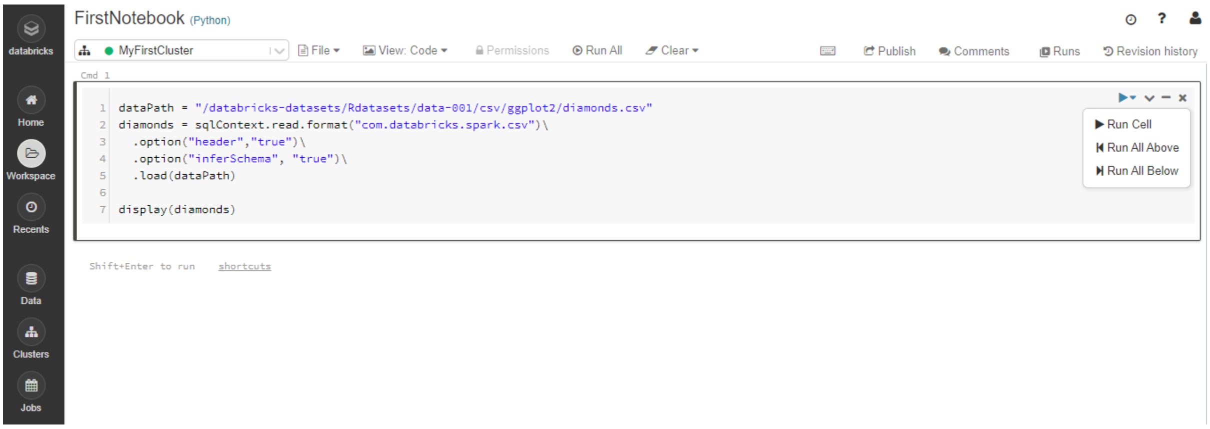
4. Add the below lines of code in the cell provided in the databrick

```
dataPath = "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv"
diamonds = sqlContext.read.format("com.databricks.spark.csv")\
  .option("header","true")\
  .option("inferSchema", "true")\
  .load(dataPath)

display(diamonds)
```

We are using the sample data provided by the databricks for practice

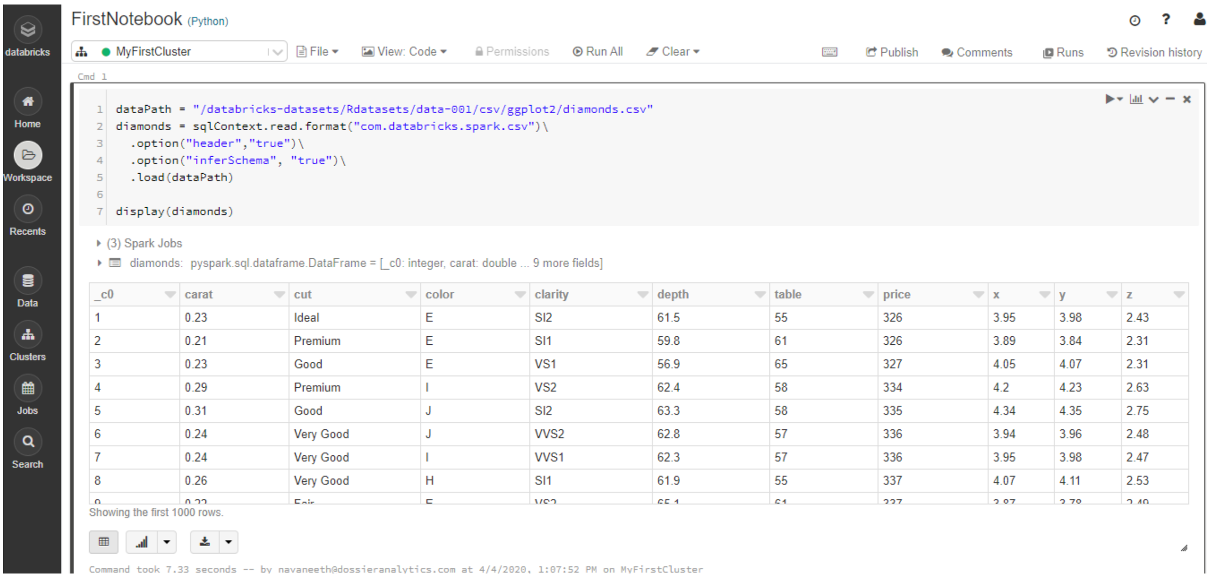
In the above code, we are parsing a CSV, identifying the schema and displaying the data.

We will be learning RDD, DataFrame, SparkSQL in next topics, this is just for demo purpose of notebook, we will be learning more clear explanations in upcoming topic.



```
dataPath = "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv"
diamonds = sqlContext.read.format("com.databricks.spark.csv")\
  .option("header","true")\
  .option("inferSchema", "true")\
  .load(dataPath)

display(diamonds)
```

5. Click on Run Cell option to execute the code. You will see the spark cluster running the job. Once the job execution is done, you will get the result like below



FirstNotebook (Python)

MyFirstCluster · File · View: Code · Permissions · Run All · Clear · Publish · Comments · Runs · Revision history

Cmd 1

```
dataPath = "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv"
diamonds = sqlContext.read.format("com.databricks.spark.csv")\
  .option("header","true")\
  .option("inferSchema", "true")\
  .load(dataPath)

display(diamonds)
```

(3) Spark Jobs

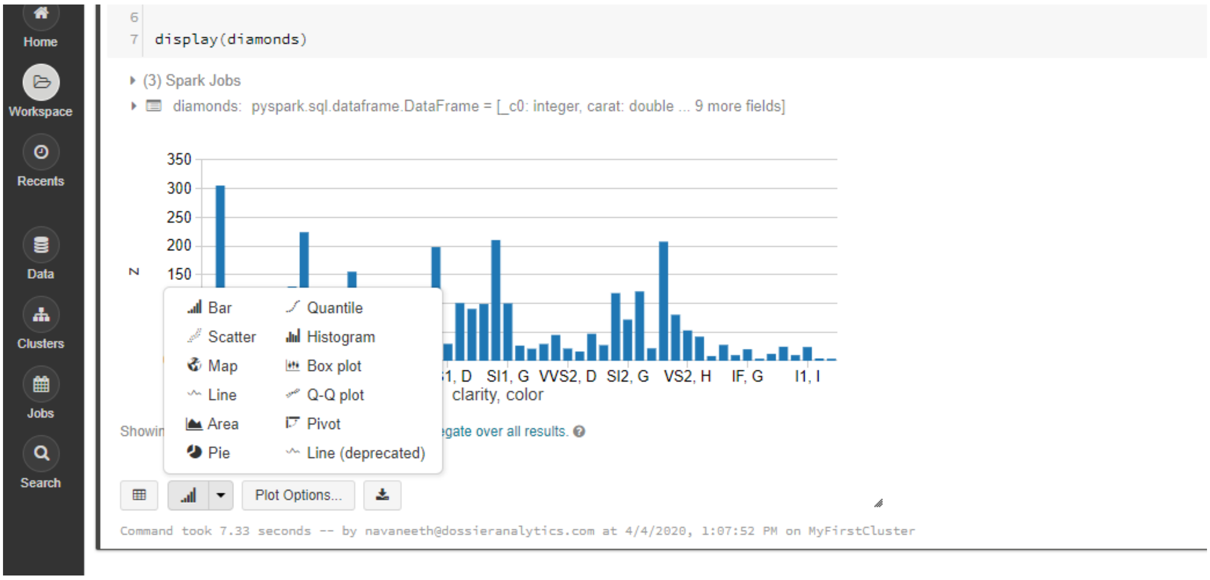
diamonds: pyspark.sql.dataframe.DataFrame = [_c0: integer, carat: double ... 9 more fields]

_c0	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53

Showing the first 1000 rows.

```
Command took 7.33 seconds -- by navaneeth@dossieranalytics.com at 4/4/2020, 1:07:52 PM on MyFirstCluster
```

6. At the bottom of the notebook, we have options for different data visualization options, you can select any one of that now, you may witness meaning less graph. We can see graph with meaningful data while we practice spark.



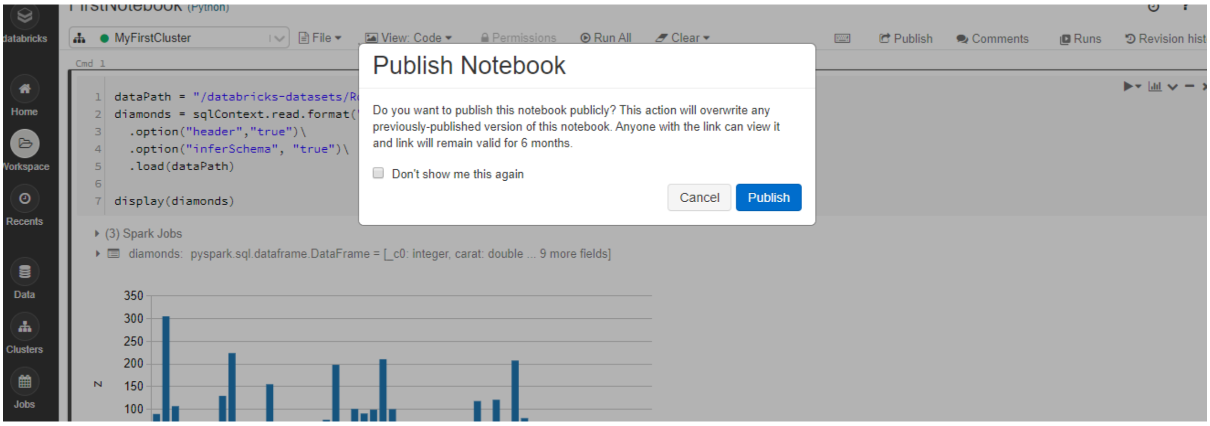
```
display(diamonds)
```

(3) Spark Jobs

diamonds: pyspark.sql.dataframe.DataFrame = [_c0: integer, carat: double ... 9 more fields]

```
Command took 7.33 seconds -- by navaneeth@dossieranalytics.com at 4/4/2020, 1:07:52 PM on MyFirstCluster
```

7. At the top, you have publish option. You can publish the notebook for future reference.



Publish Notebook

Do you want to publish this notebook publicly? This action will overwrite any previously-published version of this notebook. Anyone with the link can view it and link will remain valid for 6 months.

☐ Don't show me this again

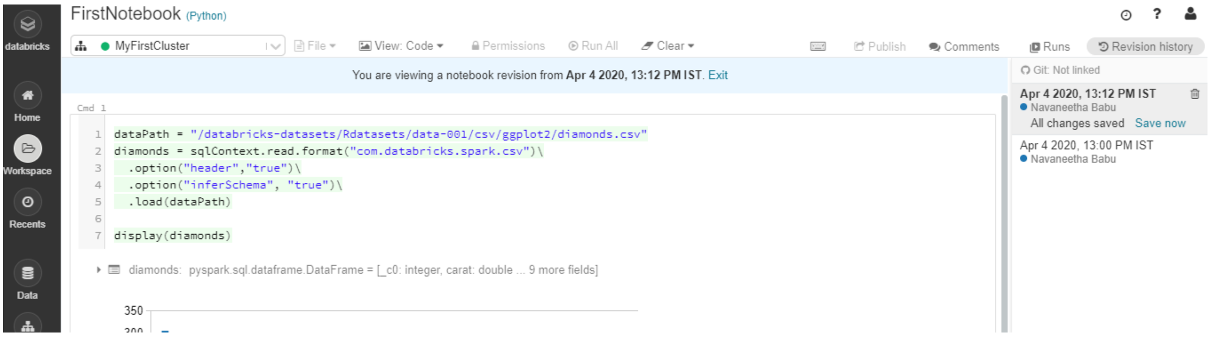
Cancel Publish

```
dataPath = "/databricks-datasets/R
diamonds = sqlContext.read.format(
  .option("header","true")\
  .option("inferSchema", "true")\
  .load(dataPath)

display(diamonds)
```

(3) Spark Jobs

diamonds: pyspark.sql.dataframe.DataFrame = [_c0: integer, carat: double ... 9 more fields]

8. You can also see the revision history in notebook about the previous cell executions



You are viewing a notebook revision from Apr 4 2020, 13:12 PM IST. Exit

Git: Not linked

Apr 4 2020, 13:12 PM IST
Navaneetha Babu
All changes saved Save now

Apr 4 2020, 13:00 PM IST
Navaneetha Babu

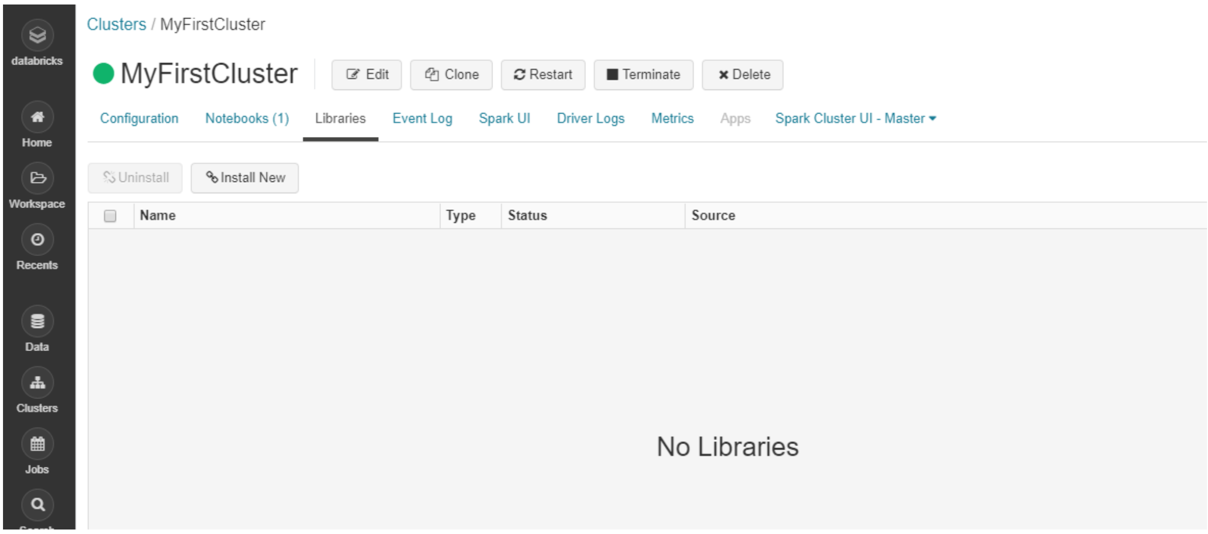
```
dataPath = "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv"
diamonds = sqlContext.read.format("com.databricks.spark.csv")\
  .option("header","true")\
  .option("inferSchema", "true")\
  .load(dataPath)

display(diamonds)
```

diamonds: pyspark.sql.dataframe.DataFrame = [_c0: integer, carat: double ... 9 more fields]

Exercise 4

Adding a library in the cluster

1. Click on the cluster option from the left side bar. Click on the libraries option available when you scroll over.



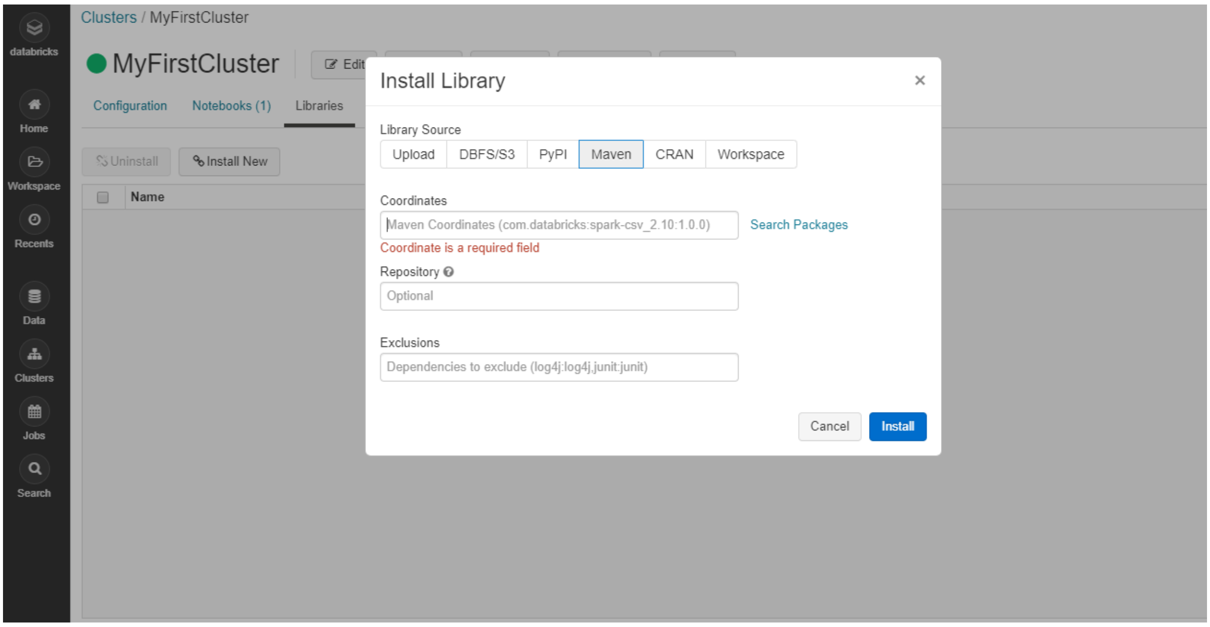
Clusters / MyFirstCluster

MyFirstCluster

☐	Name	Type	Status	Source

No Libraries

2. Click on “Install New” option available and choose maven



Install Library

Library Source: Upload | DBFS/S3 | PyPI | Maven | CRAN | Workspace

Coordinates: Maven Coordinates (com.databricks:spark-csv_2.10:1.0.0) — Search Packages

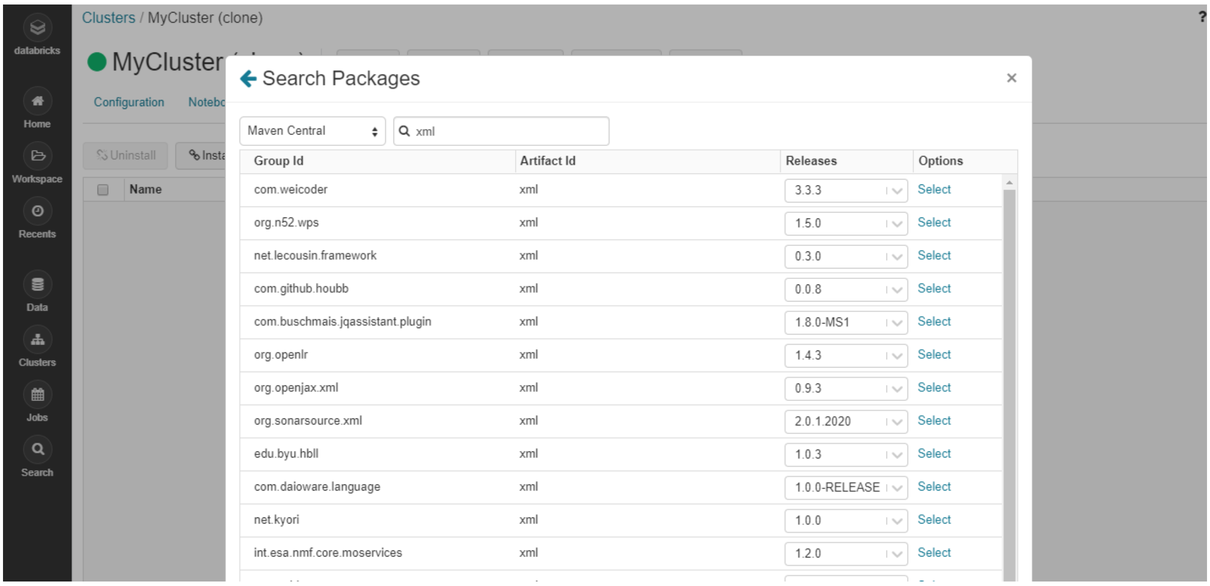
Coordinate is a required field

Repository: Optional

Exclusions: Dependencies to exclude (log4j:log4j,junit:junit)

Cancel | Install

3. Click on “Search Packages” and wait for some time. The libraries will be retrieved for display. Then you can select the library.

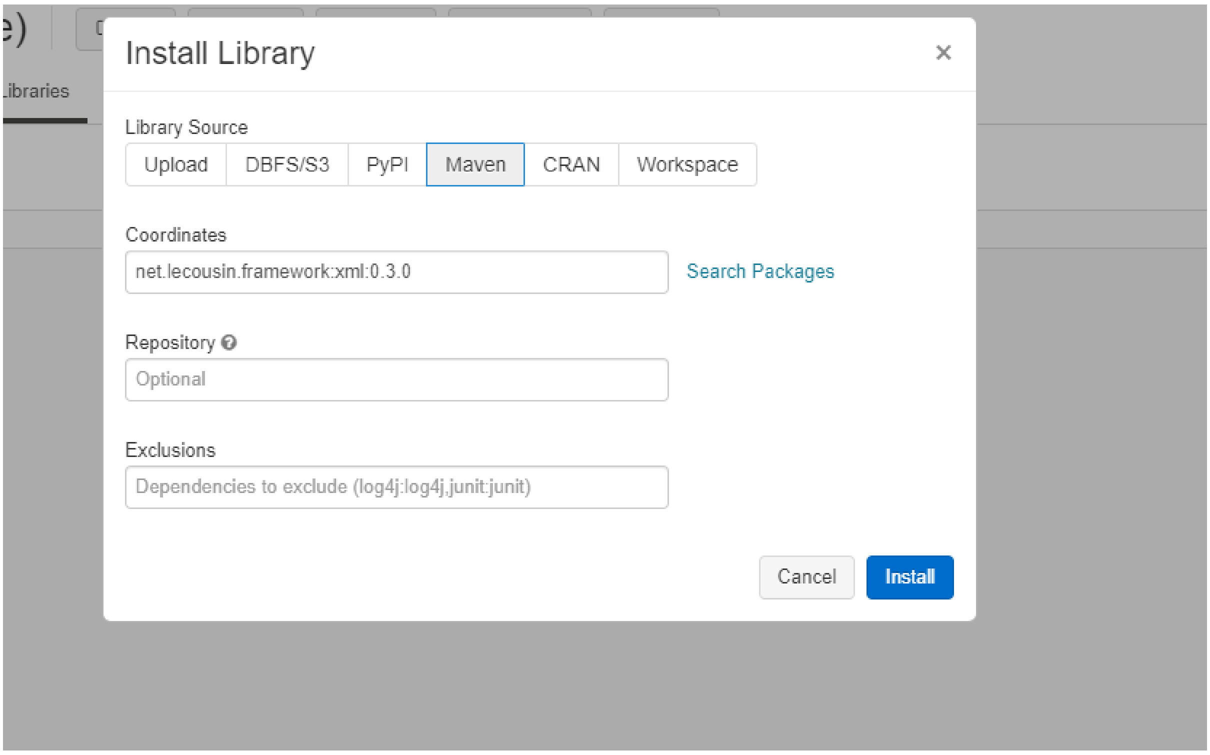


Search Packages

Maven Central | xml

Group Id	Artifact Id	Releases	Options
com.weicoder	xml	3.3.3	Select
org.n52.wps	xml	1.5.0	Select
net.lecousin.framework	xml	0.3.0	Select
com.github.houbb	xml	0.0.8	Select
com.buschmais.jqassistant.plugin	xml	1.8.0-MS1	Select
org.openlr	xml	1.4.3	Select
org.openjax.xml	xml	0.9.3	Select
org.sonarsource.xml	xml	2.0.1.2020	Select
edu.byu.hbll	xml	1.0.3	Select
com.daioware.language	xml	1.0.0-RELEASE	Select
net.kyori	xml	1.0.0	Select
int.esa.nmf.core.moservices	xml	1.2.0	Select

4. Once the library is selected, click on install option



Install Library

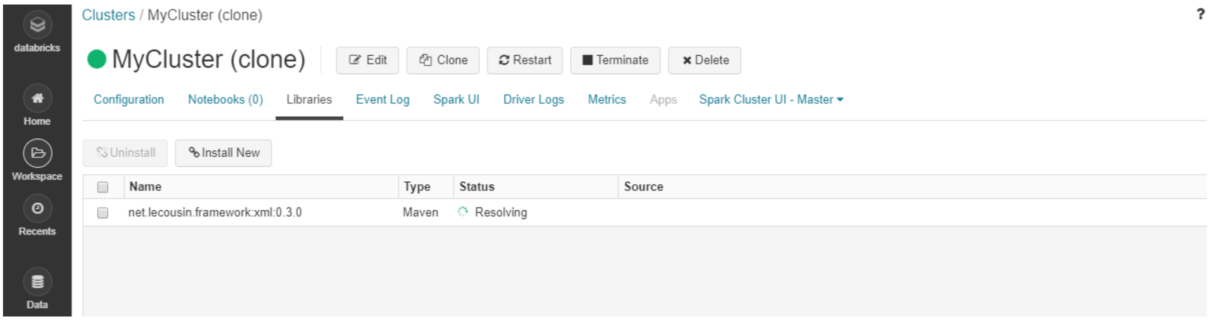
Library Source

Coordinates: net.lecousin.framework:xml:0.3.0

Repository: Optional

Exclusions: Dependencies to exclude (log4j:log4j,junit:junit)

5. The library will be available in the cluster in sometime



Clusters / MyCluster (clone)

MyCluster (clone)

	Name	Type	Status	Source
□	net.lecousin.framework:xml:0.3.0	Maven	Resolving	