

PySpark Data-Engineering

Creating New Cluster

The screenshot shows the Databricks 'New compute' page for a cluster named 'Jatin J's Cluster'. The left sidebar contains navigation links: New, Workspace, Recents, Search, Catalog, Workflows, Compute, Machine Learning, and Experiments. The 'Compute' section is active. The main content area is titled 'Compute > New compute' and 'Jatin J's Cluster'. It includes a 'Compute name' field with the value 'Jatin J's Cluster', a 'Databricks runtime version' dropdown set to 'Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)', and an 'Instance' section with a warning: 'Free 15 GB Memory: As a Community Edition user, your compute will automatically terminate after an idle period of one or two hours. For more configuration options, please upgrade your Databricks subscription.' Below this is a 'Spark' section with a 'Spark config' field containing 'spark.databricks.rocksDB.fileManager.useCommitService false' and an 'Environment variables' field containing 'PYSPARK_PYTHON=/databricks/python3/bin/python3'. At the bottom are 'Create compute' and 'Cancel' buttons.

Compute > New compute

Jatin J's Cluster

Compute name

Jatin J's Cluster

Databricks runtime version

Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)

Instance

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

Spark

Spark config

spark.databricks.rocksDB.fileManager.useCommitService false

Environment variables

PYSPARK_PYTHON=/databricks/python3/bin/python3

Create compute Cancel

To Create A New Notebook

The screenshot shows the Databricks 'New' dropdown menu. The left sidebar is the same as in the previous image. The 'New' button is highlighted, and a dropdown menu is open showing options: Notebook, Table, Compute, Cluster, Machine Learning, and Experiment. The 'Cluster' option is highlighted. The background shows the 'New compute' page with the 'Compute name' field and 'Databricks runtime version' dropdown.

New

Notebook

Table

Compute

Cluster

Machine Learning

Experiment

Compute name

Jatin J's Cluster

Databricks runtime version

Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)

Instance

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

Spark

Spark config

spark.databricks.rocksDB.fileManager.useCommitService false

Writing Code in Notebook

The screenshot shows the Databricks Spark Programs notebook interface. The left sidebar contains navigation options: New, Workspace, Recents, Search, Catalog, Workflows, Compute, Machine Learning, and Experiments. The main workspace area displays two code cells. The first cell, titled '02:28 PM (24)', contains Python code for creating a SparkContext and a SparkSession, and then creating an RDD. The second cell, titled '02:28 PM (206)', contains code to read a CSV file and display its contents. The output of the second cell shows a Spark DataFrame with 3 rows and 11 columns.

```
from pyspark import SparkContext
from pyspark.sql import SparkSession

sc=SparkContext.getOrCreate()
spark=SparkSession.builder.appName('Pyspark first Program').getOrCreate()

rdd = sc.parallelize([('C',85,76,87,91), ('B',85,76,87,91), ('A', 85,78,96,92), ('A', 92,76,89,96)])
print(type(rdd))

<class 'pyspark.rdd.RDD'>
```

```
data = spark.read.csv("/FileStore/tables/loanData.csv")
data.show()
display(data)
```

3) Spark Jobs

data: pyspark.sql.dataframe.DataFrame = [c0: string, c1: string, ... 11 more fields]										
360	1	Urban	Y							
[LP001027]	Male	Yes	2	Graduate	null	2500	1840	109		
360	1	Urban	Y							
[LP001028]	Male	Yes	2	Graduate	No	3073	8186	200		
360	1	Urban	Y							

To see all the clusters

The screenshot shows the Databricks Compute page. The left sidebar contains navigation options: New, Workspace, Recents, Search, Catalog, Workflows, Compute, Machine Learning, and Experiments. The main workspace area displays a table of clusters. The table has columns for State, Name, Runtime, Active memory, Active cores, Active DBU / h, Source, Creator, Notebooks, and a settings icon. There are two clusters listed: 'Practice' and 'My Cluster'.

State	Name	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	Notebooks	
●	Practice	12.2	15 GB	2 cores	1	UI	jatinji305@gmail.com	1	⋮
⌂	My Cluster	12.2	-	-	-	UI	jatinji305@gmail.com	-	▶ ⋮