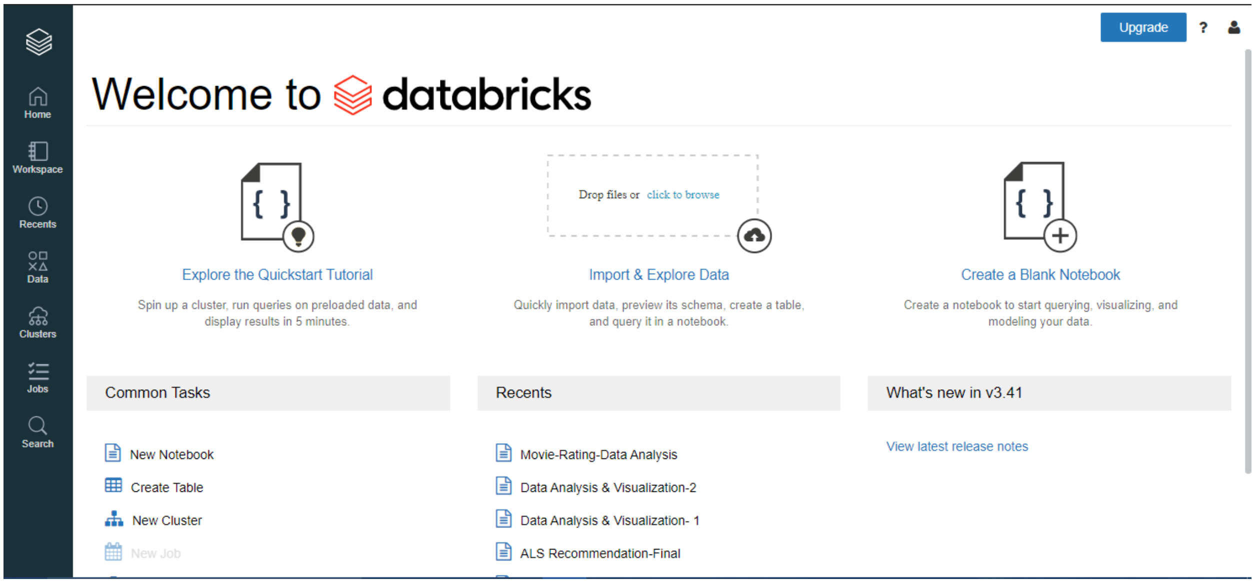
DataBricks Cloud Community Edition - Documentation.

Step 1: Create a new account in DataBricks Cloud Community Edition.

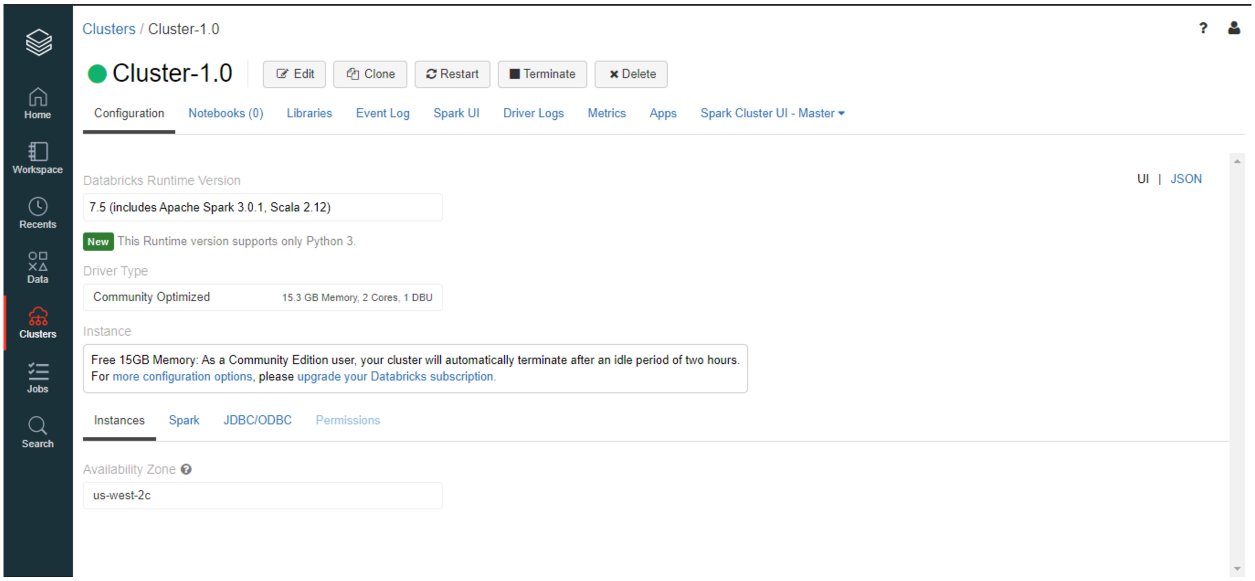


Explore the Quickstart Tutorial: Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

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.

Step 2: Create a new Spark Cluster



Databricks Runtime Version: 7.5 (includes Apache Spark 3.0.1, Scala 2.12)

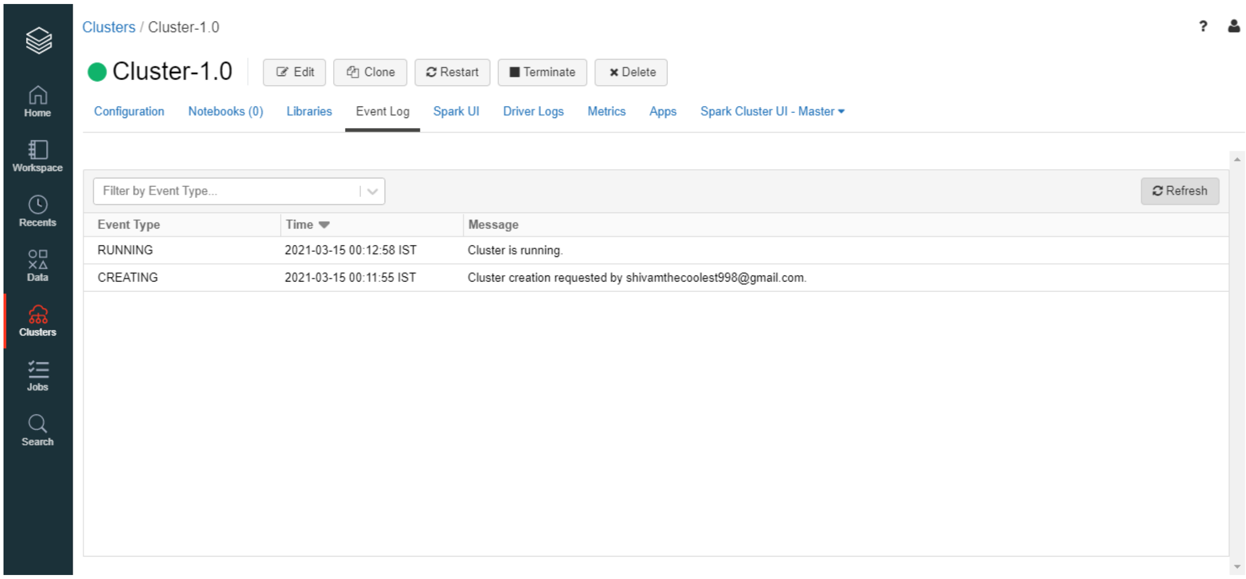
New This Runtime version supports only Python 3.

Driver Type: Community Optimized 15.3 GB Memory, 2 Cores, 1 DBU

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.

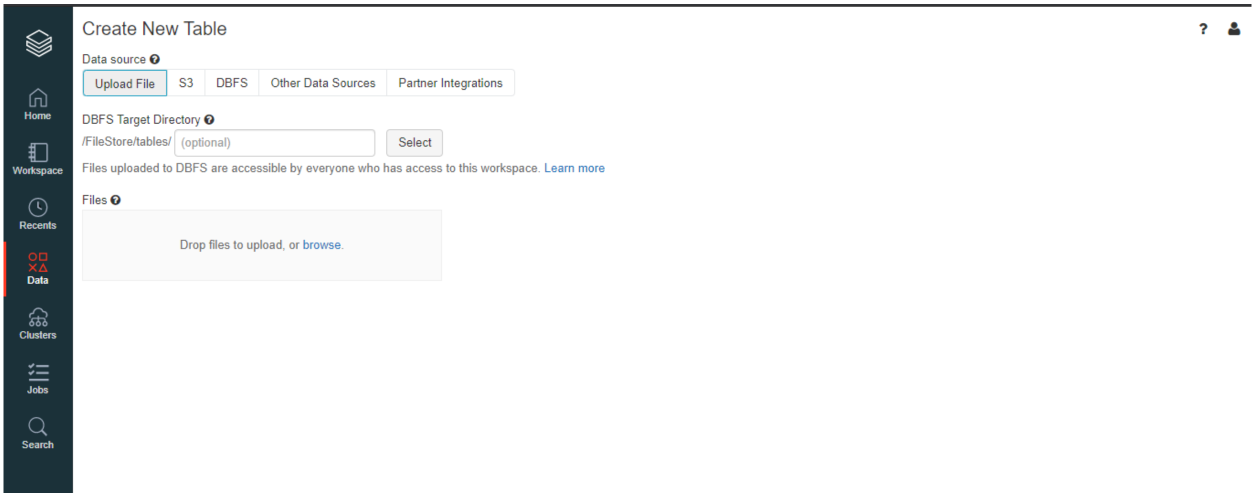
Availability Zone: us-west-2c

Cluster is up and running.



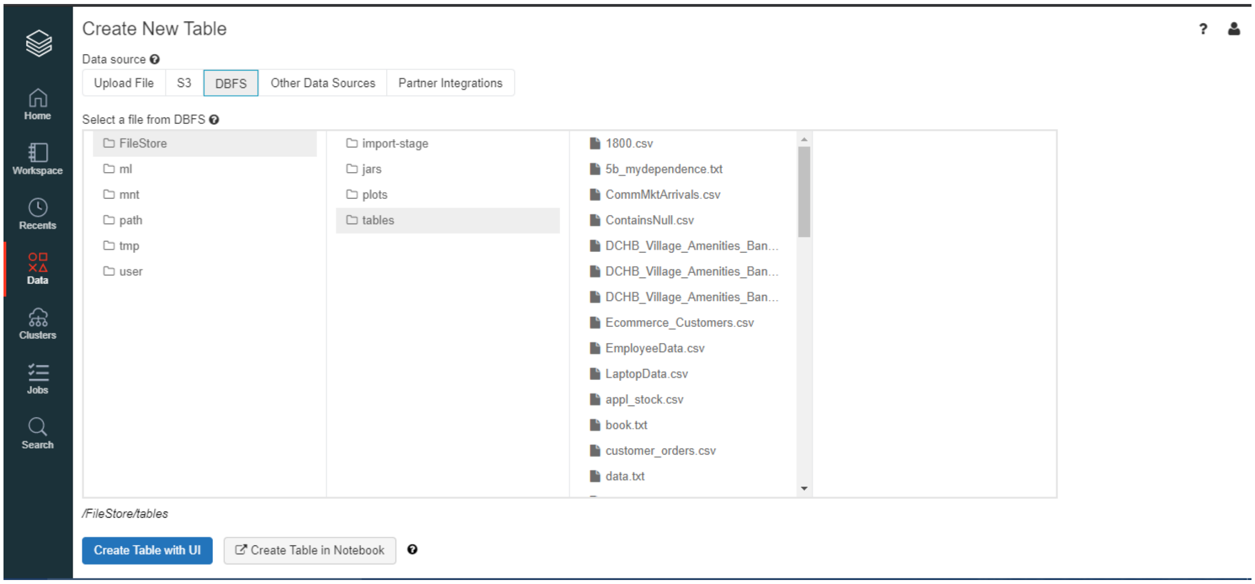
Event Type	Time	Message
RUNNING	2021-03-15 00:12:58 IST	Cluster is running.
CREATING	2021-03-15 00:11:55 IST	Cluster creation requested by shivamthecoolest998@gmail.com.

Step 3: Upload files from Local System to DBFS Databricks File System or connect from AWS S3 bucket by mounting it on Databricks.

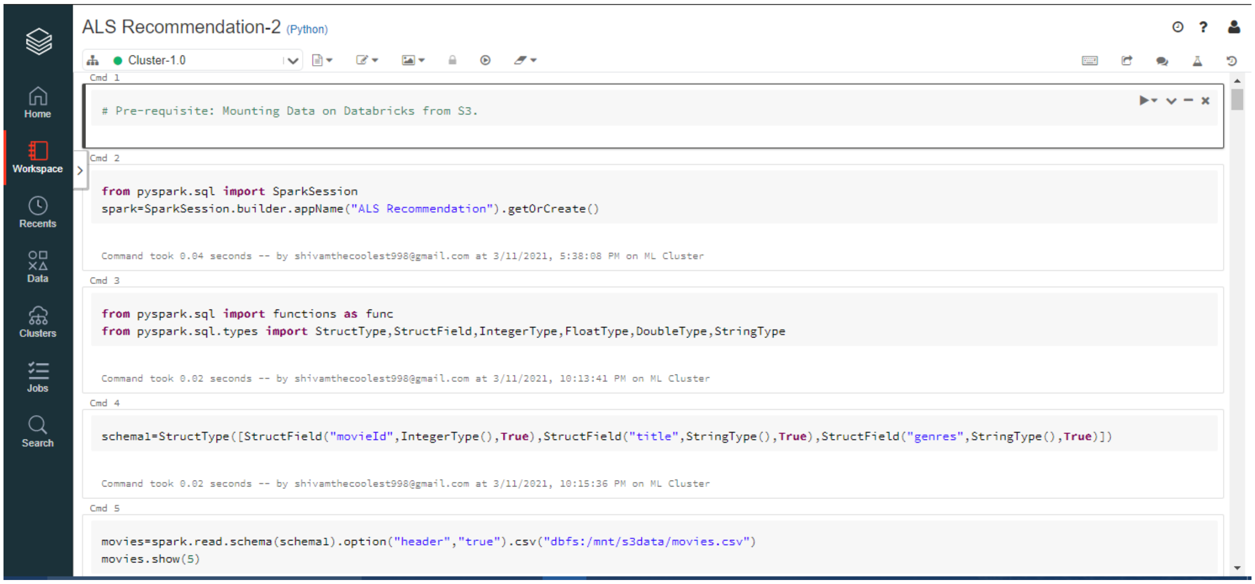


Files uploaded to DBFS are accessible by everyone who has access to this workspace. Learn more

By default, the uploaded Files will stored at dbfs:/FileStore/tables/filename location.



Step 4: Create a new Notebook (Python - Pyspark) and attach it to running cluster and start writing & executing Spark Codes.



Cmd 1

```
# Pre-requisite: Mounting Data on Databricks from S3.
```

Cmd 2

```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("ALS Recommendation").getOrCreate()

Command took 0.04 seconds -- by shivamthecoolest998@gmail.com at 3/11/2021, 5:38:08 PM on ML Cluster
```

Cmd 3

```
from pyspark.sql import functions as func
from pyspark.sql.types import StructType,StructField,IntegerType,FloatType,DoubleType,StringType

Command took 0.02 seconds -- by shivamthecoolest998@gmail.com at 3/11/2021, 10:13:41 PM on ML Cluster
```

Cmd 4

```
schema1=StructType([StructField("movieId",IntegerType(),True),StructField("title",StringType(),True),StructField("genres",StringType(),True)])

Command took 0.02 seconds -- by shivamthecoolest998@gmail.com at 3/11/2021, 10:15:36 PM on ML Cluster
```

Cmd 5

```
movies=spark.read.schema(schema1).option("header","true").csv("dbfs:/mnt/s3data/movies.csv")
movies.show(5)
```