# Learning Apache Spark with Python

## CONTENTS

CHAPTER

TWO

# WHY SPARK WITH PYTHON ?

**Chinese proverb**

**Sharpening the knife longer can make it easier to hack the firewood** – old Chinese proverb

I want to answer this question from the following two parts:

## 2.1 Why Spark?

I think the following four main reasons from Apache Spark™ official website are good enough to convince you to use Spark.

1. Speed

   Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

   Apache Spark has an advanced DAG execution engine that supports acyclic data flow and in-memory computing.

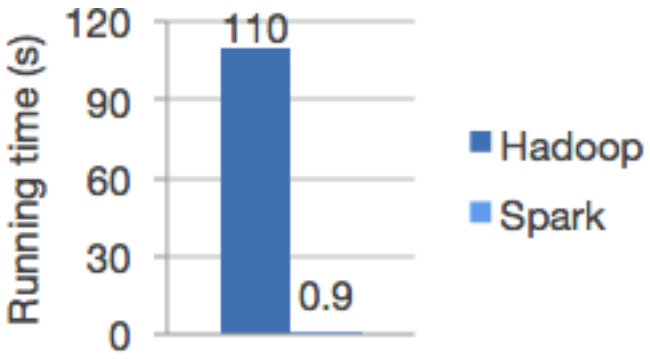



Fig. 1: Logistic regression in Hadoop and Spark

2. Ease of Use

   Write applications quickly in Java, Scala, Python, R.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it interactively from the Scala, Python and R shells.

3. Generality

   Combine SQL, streaming, and complex analytics.

   Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.

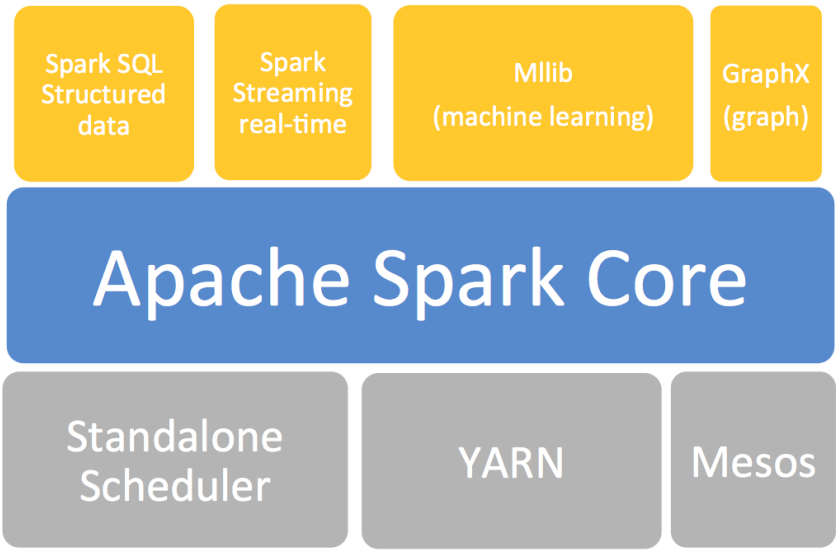



Fig. 2: The Spark stack

4. Runs Everywhere

   Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

## 2.2 Why Spark with Python (PySpark)?

No matter you like it or not, Python has been one of the most popular programming languages.


Fig. 3: The Spark platform

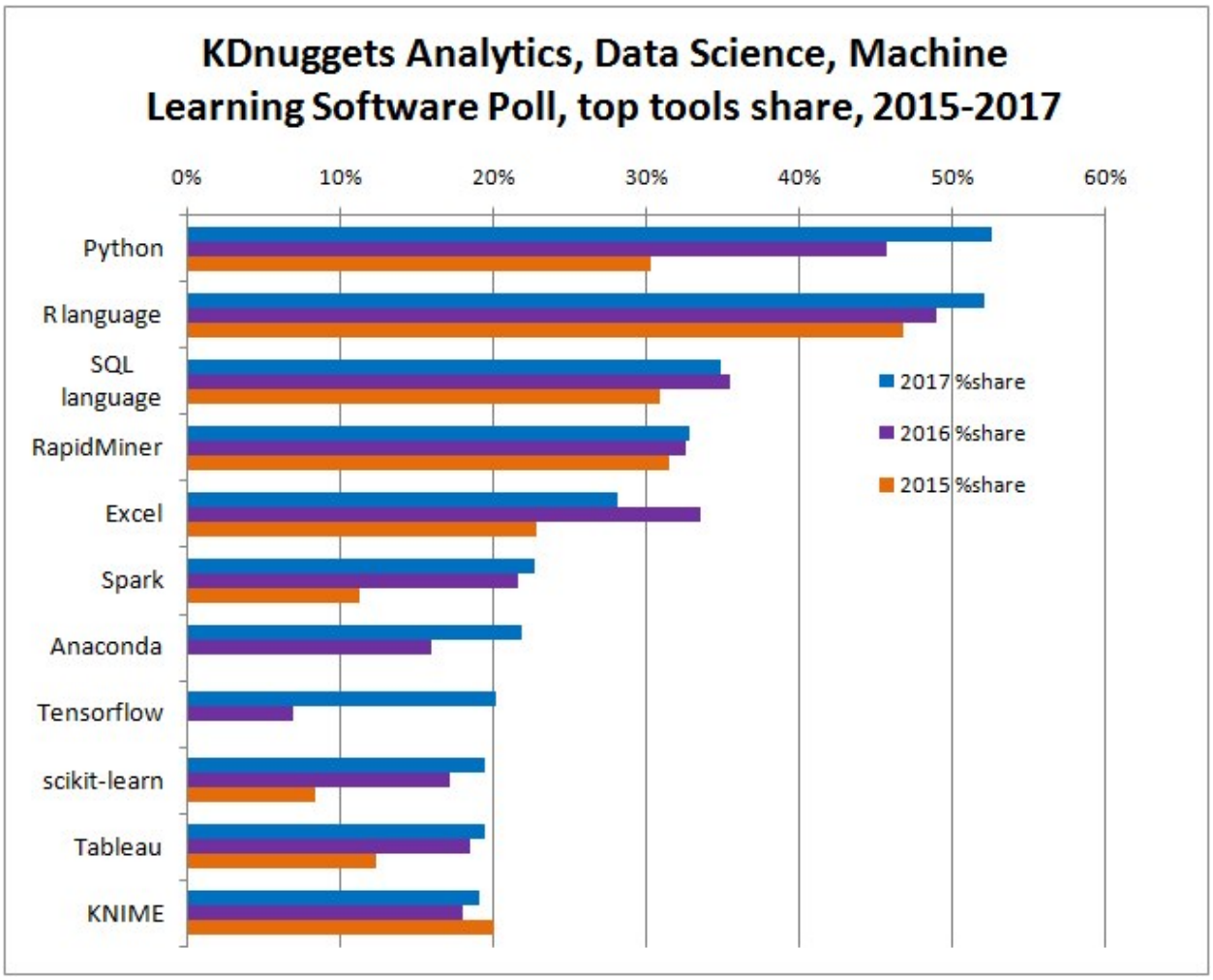



Fig. 4: KDnuggets Analytics/Data Science 2017 Software Poll from kdnuggets.

CHAPTER

THREE

# CONFIGURE RUNNING PLATFORM

**Chinese proverb**

**Good tools are prerequisite to the successful execution of a job.** – old Chinese proverb

A good programming platform can save you lots of troubles and time. Herein I will only present how to install my favorite programming platform and only show the easiest way which I know to set it up on Linux system. If you want to install on the other operator system, you can Google it. In this section, you may learn how to set up Pyspark on the corresponding programming platform and package.

## 3.1 Run on Databricks Community Cloud

If you don't have any experience with Linux or Unix operator system, I would love to recommend you to use Spark on Databricks Community Cloud. Since you do not need to setup the Spark and it's totally **free** for Community Edition. Please follow the steps listed below.

1. Sign up a account at: https://community.cloud.databricks.com/login.html

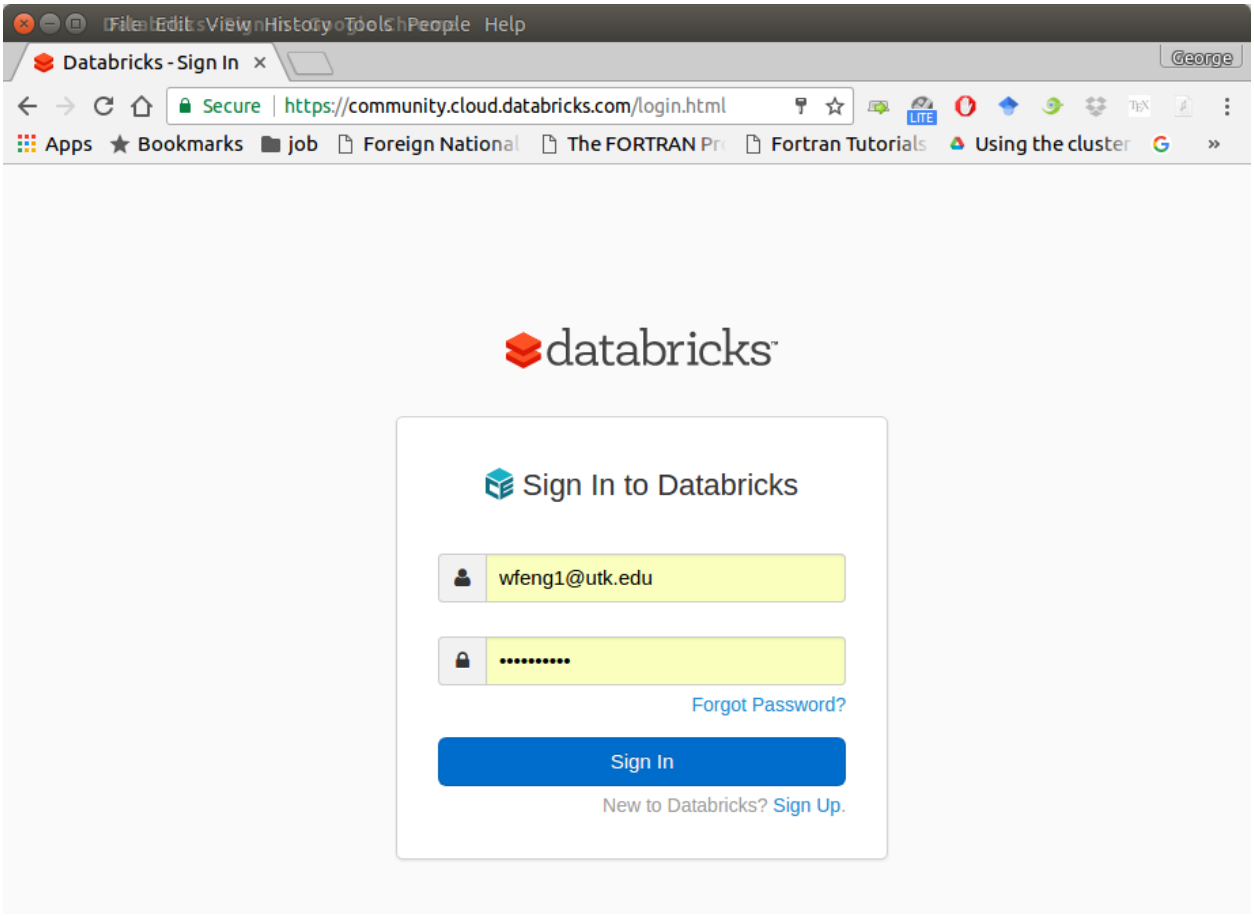
2. Sign in with your account, then you can creat your cluster(machine), table(dataset) and notebook(code).

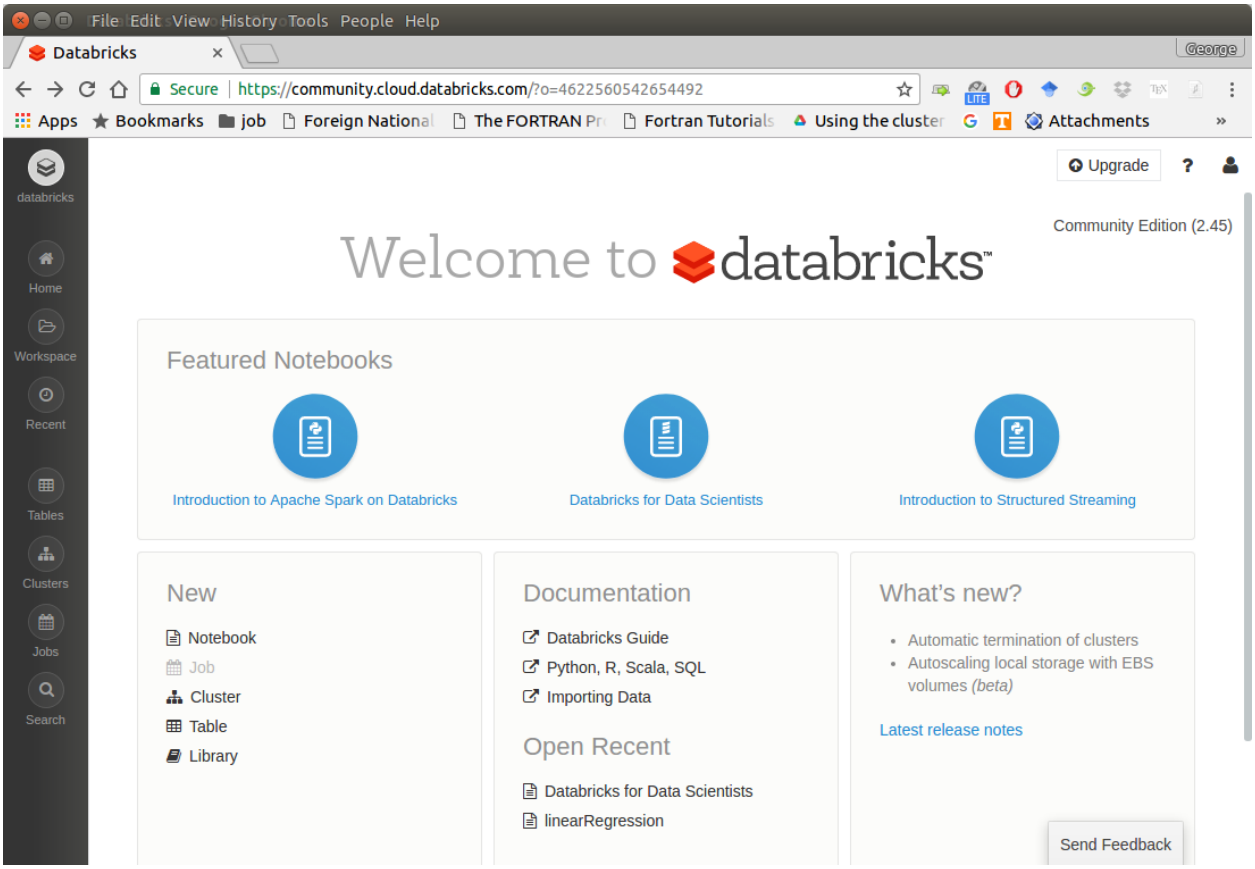
3. Create your cluster where your code will run

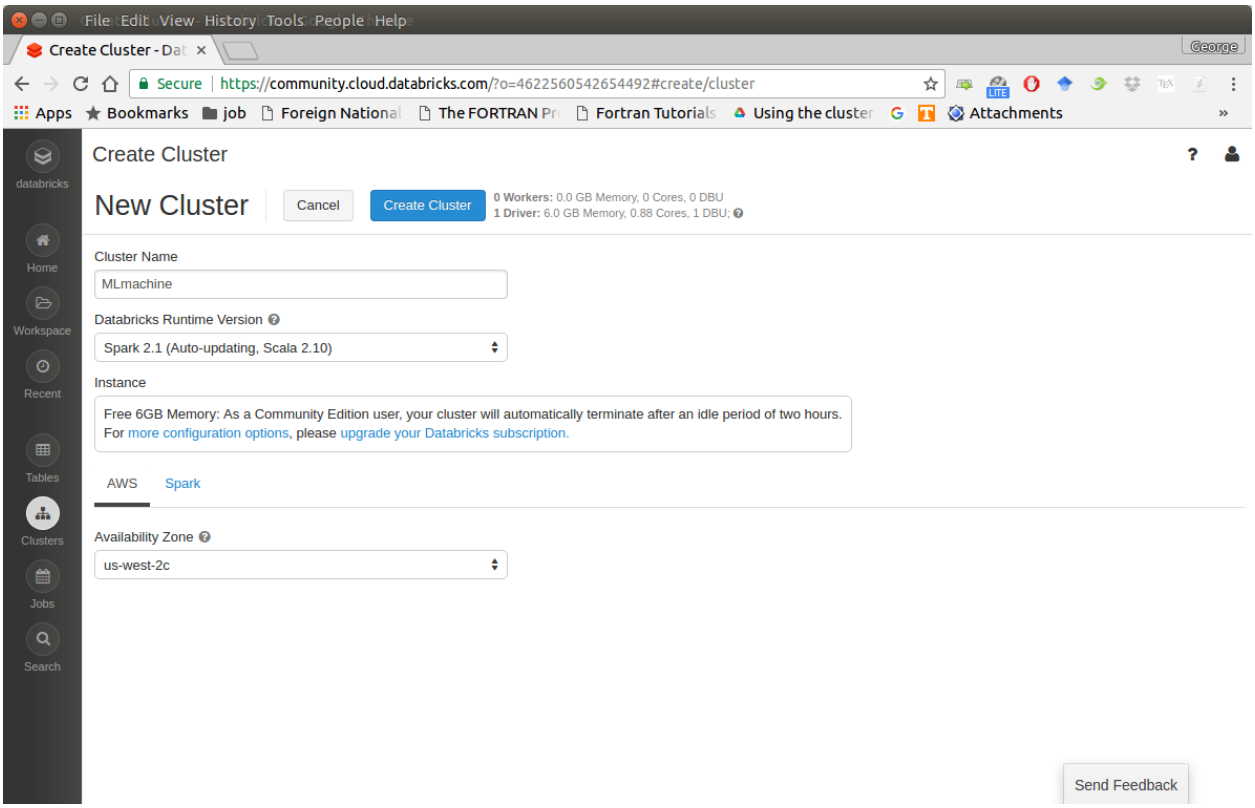
4. Import your dataset

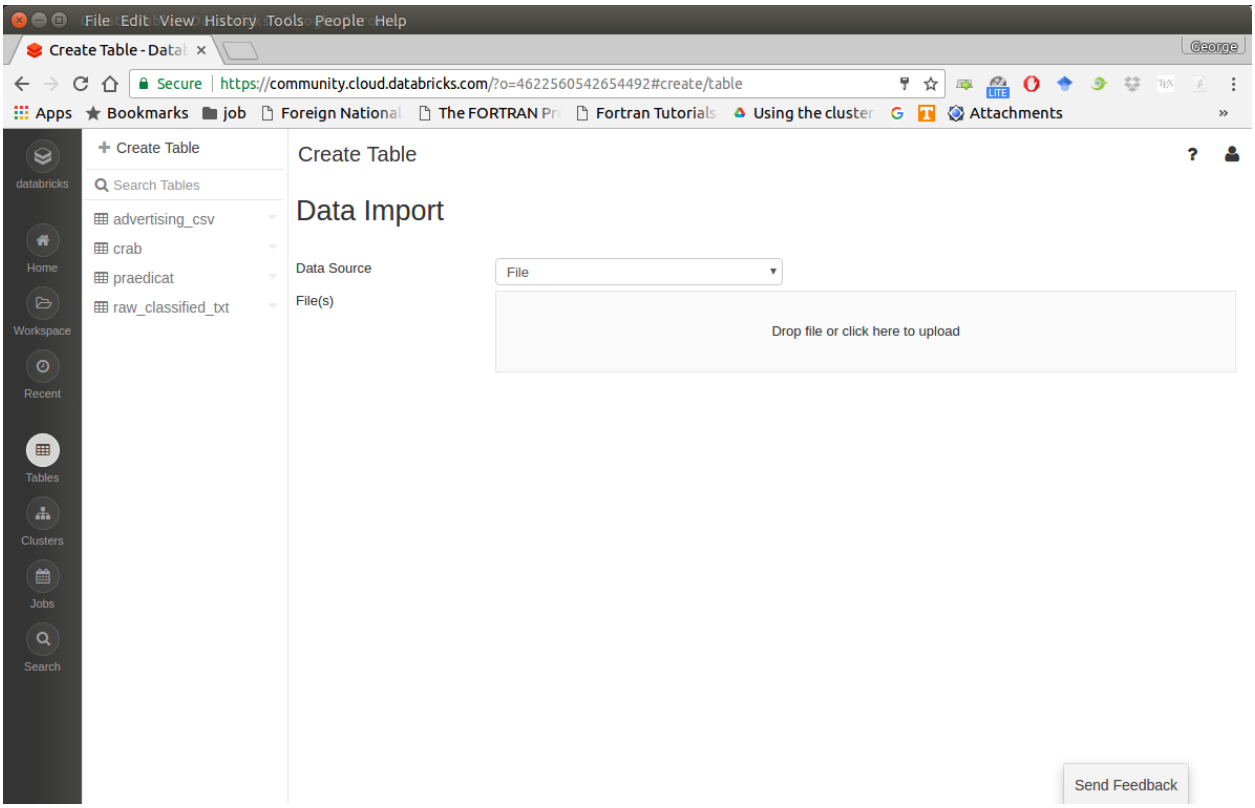
   **Note:** You need to save the path which appears at Uploaded to DBFS: /FileStore/tables/05rmhuqv1489687378010/. Since we will use this path to load the dataset.

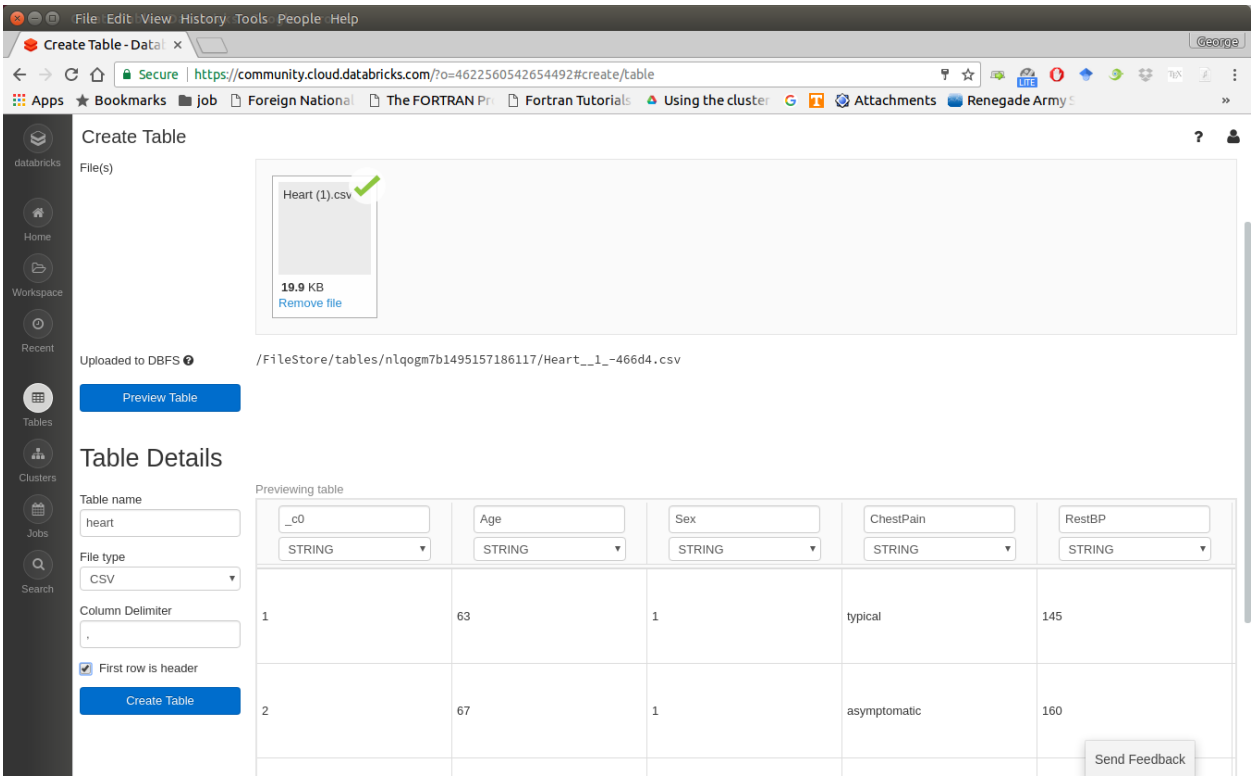
5. Create your notebook


File Edit View History Tools People Help
Databricks - Sign In
George
Secure | https://community.cloud.databricks.com/login.html
Apps Bookmarks job Foreign National The FORTRAN Pr Fortran Tutorials Using the cluster
databricks
Sign In to Databricks
wfeng1@utk.edu
Forgot Password?
Sign In
New to Databricks? Sign Up.


File Edit View History Tools People Help
Databricks
George
Secure | https://community.cloud.databricks.com/?o=4622560542654492
Apps Bookmarks job Foreign National The FORTRAN Pr Fortran Tutorials Using the cluster Attachments
databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Upgrade
Community Edition (2.45)
Welcome to databricks
Featured Notebooks
Introduction to Apache Spark on Databricks
Databricks for Data Scientists
Introduction to Structured Streaming
New
Notebook
Job
Cluster
Table
Library
Documentation
Databricks Guide
Python, R, Scala, SQL
Importing Data
Open Recent
Databricks for Data Scientists
linearRegression
What's new?
Automatic termination of clusters
Autoscaling local storage with EBS volumes (beta)
Latest release notes
Send Feedback


File Edit View History Tools People Help
Create Cluster - Dat
George
Secure | https://community.cloud.databricks.com/?o=4622560542654492#create/cluster
Apps Bookmarks job Foreign National The FORTRAN Pr Fortran Tutorials Using the cluster Attachments
databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Create Cluster
New Cluster
Cancel
Create Cluster
0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 6.0 GB Memory, 0.88 Cores, 1 DBU;
Cluster Name
MLmachine
Databricks Runtime Version
Spark 2.1 (Auto-updating, Scala 2.10)
Instance
Free 6GB 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.
AWS
Spark
Availability Zone
us-west-2c
Send Feedback


File Edit View History Tools People Help
Create Table - Data
George
Secure | https://community.cloud.databricks.com/?o=4622560542654492#create/table
Apps Bookmarks job Foreign National The FORTRAN Pr Fortran Tutorials Using the cluster Attachments
databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Create Table
Search Tables
advertising_csv
crab
praedicat
raw_classified_txt
Create Table
Data Import
Data Source
File
File(s)
Drop file or click here to upload
Send Feedback


File Edit View History Tools People Help
Create Table - Data
George
Secure | https://community.cloud.databricks.com/?o=4622560542654492#create/table
Apps Bookmarks job Foreign National The FORTRAN Pr Fortran Tutorials Using the cluster Attachments Renegade Army
databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Create Table
File(s)
Heart (1).csv
19.9 KB
Remove file
Uploaded to DBFS
/FileStore/tables/nlqogm7b1495157186117/Heart__1_-466d4.csv
Preview Table
Table Details
Table name
heart
File type
CSV
Column Delimiter
,
First row is header
Create Table
Previewing table
_c0
Age
Sex
ChestPain
RestBP
STRING
1
63
1
typical
145
2
67
1
asymptomatic
160
Send Feedback