



SETTING UP SPARK DATABRICKS COMMUNITY EDITION

Ramakrishna Addanki
Data Engineering Lead

AGENDA

- 1) Setting up databricks cluster for session.
- 2) Understanding Databricks Runtime.
- 3) Creating Cluster
- 4) Spark UI, LOGS & METRICS
- 5) Installing Libraries
- 6) Creating Notebooks
- 7) Executing code in notebook
- 8) Git / bitbucket Integration



Step 1: Launch the sign up wizard and select a subscription type

Go to [Databricks](#) and click **Try Databricks** at the top right.

Step 2: Click **GET STARTED to choose 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](#)

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

[GET STARTED](#)

SETTING UP SPARK CLUSTER

Step 3: When you select Community Edition you'll see a registration form.

Fill in the registration form.

Click Sign Up.

Read the Terms of Service and click **Agree**.

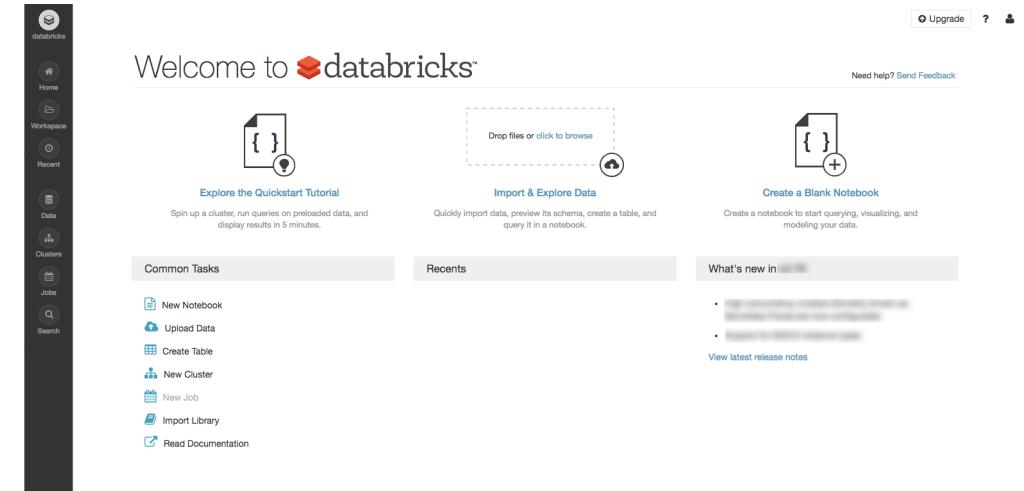
Step 4: When you receive the “Welcome to Databricks” email, click the link to verify your mail address.

Sign Up for Databricks Community Edition

First Name *	Last Name *
<input type="text"/>	<input type="text"/>
Company Name *	Work Email *
<input type="text"/>	<input type="text"/>
Phone Number	What is your intended use case? *
<input type="text"/>	<input type="text" value="- Please Select -"/>
How would you describe your role? *	
<input type="text" value="- Please Select -"/>	

SETTING UP SPARK CLUSTER

Step 5: Log into Databricks using the credentials you supplied when you registered. You'll see the Databricks Community Edition home page.



SETTING UP SPARK CLUSTER

SETTING UP SPARK CLUSTER

- 1) Databricks Community Edition is fully resourced;
- 2) you are not required to supply any compute or storage resources.
- 3) However, several features available in the Databricks Platform Free Trial, such as the [REST API](#), are not available in Databricks Community Edition.

For more info refer:

<https://databricks.com/product/faq/community-edition>

The screenshot shows the Databricks Clusters page. On the left, a vertical sidebar lists navigation options: Home, Workspace, Recents, Data, Clusters (which is selected and highlighted with a red box), Jobs, and Search. A red arrow points from the 'Clusters' option in the sidebar to the 'Create Cluster' button at the top of the main content area. The main content area has two sections: 'Interactive Clusters' and 'Automated Clusters'. Both sections contain a message 'No clusters found'. At the top right of the main area, there are buttons for 'All' (selected), 'Created by me', and a search bar labeled 'Filter'. The status '0 clusters, 0 pinned' is also displayed.

Clusters

+ Create Cluster

Interactive Clusters

No clusters found

All Created by me Filter

0 clusters, 0 pinned

Automated Clusters

No clusters found

Home

Workspace

Recents

Data

Clusters

Jobs

Search

CREATING CLUSTER |



databricks

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



enter cluster name then create cluster button will activate

Databricks Runtime Version

Runtime: 6.2 (Scala 2.11, Spark 2.4.4)

New This Runtime version supports only Python 3.

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](#).

[Instances](#) [Spark](#)

Availability Zone

us-west-2c

CREATING CLUSTER

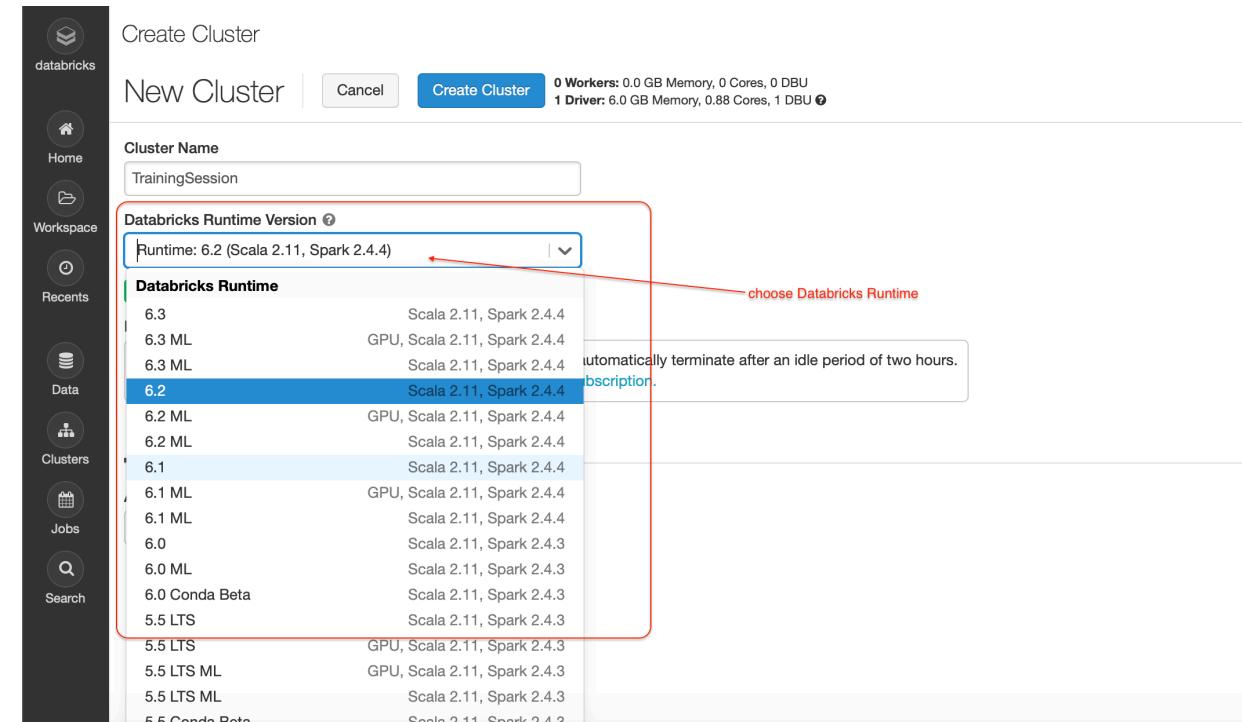
Databricks Runtime includes Apache Spark but also adds a number of components and updates that substantially improve the usability, performance, and security of big data.

Installed Java, Scala, Python, and R libraries.

Ubuntu and its accompanying system libraries

GPU libraries for GPU-enabled clusters.

Databricks services that integrate with other components of the platform, such as notebooks, jobs, and cluster manager



CREATING CLUSTER

Create Cluster

New Cluster | Cancel | Create Cluster | 0 Workers: 0.0 GB Memory | 1 Driver: 6.0 GB Memory

Cluster Name
TrainingSession

Databricks Runtime Version ⓘ
Runtime: 6.3 (Scala 2.11, Spark 2.4.4)

New This Runtime version supports only Python 3.

Instance
Free 6GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period. For [more configuration options](#), please [upgrade your Databricks subscription](#).

Instances: **Spark**

Availability Zone ⓘ
us-west-2c

```
us-west-2c
us-west-2b
us-west-2a
us-west-2d
```

Cluster

Cluster | Cancel | Create Cluster | 0 Workers: 0.0 GB Memory, 0 Cores, 0 DBL | 1 Driver: 6.0 GB Memory, 0.88 Cores, 1 DBL

Memory: As a Community Edition user, your cluster will automatically terminate after an idle period. For [more configuration options](#), please [upgrade your Databricks subscription](#).

Spark

Configuration ⓘ
Spark configuration options here. Provide only one key-value pair per line:
spark.speculation=true
spark.kryo.registrator my.package.MyRegistrar

Environment Variables ⓘ
JAVACONFIGNESS=0JAVA_OPTS="-D... -D... -XX:MaxPermSize=256m
JAVA_OPTS="\$JAVA_OPTS -D... ""

CREATING CLUSTER



databricks

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 Hide empty fields

```
1 {  
2   "num_workers": 0,  
3   "cluster_name": "TrainingSession",  
4   "spark_version": "6.3.x-scala2.11",  
5   "spark_conf": {},  
6   "aws_attributes": {  
7     "first_on_demand": 0,  
8     "availability": "ON_DEMAND",  
9     "zone_id": "us-west-2c",  
10    "spot_bid_price_percent": 100,  
11    "ebs_volume_count": 0  
12  },  
13  "node_type_id": "dev-tier-node",  
14  "ssh_public_keys": [],  
15  "custom_tags": {},  
16  "spark_env_vars": {},  
17  "autotermination_minutes": 120,  
18  "init_scripts": []  
19 }
```

Copy

you can also choose json to provide config's to spark instead of UI

UI | JSON



Home

Workspace

Recents

Data

Clusters

Jobs

Search

CREATING CLUSTER



databricks

Clusters

[+ Create Cluster](#)[All](#)[Created by me](#)[Filter](#)

1 clusters, 0 pinned

[▼ Interactive Clusters](#)

Name	State	Nodes	Driver	Worker	Runtime	Creator	Actions
TrainingSession	Running	1 (0 spot)	Community ..	Community ..	6.3 (includes ...)	addanki.ram..	

[▼ Automated Clusters](#)

cluster got created

you can see spark UI by clicking over here

No clusters found

[Home](#)[Workspace](#)[Recents](#)[Data](#)[Clusters](#)[Jobs](#)[Search](#)

CLUSTER CREATED



TrainingSession

[Edit](#)[Clone](#)[Restart](#)[Terminate](#)[Delete](#)[Configuration](#)[Notebooks \(0\)](#)[Libraries](#)[Event Log](#)[Spark UI](#)[Driver Logs](#)[Metrics](#)[Apps](#)[Spark Cluster UI - Master ▾](#)

Hostname: ec2-54-214-223-181.us-west-2.compute.amazonaws.com Spark Version:6.3.x-scala2.11

[Jobs](#)[Stages](#)[Storage](#)[Environment](#)[Executors](#)[SQL](#)[JDBC/ODBC Server](#)

Spark Jobs [\(?\)](#)

User: root**Total Uptime:** 10 min**Scheduling Mode:** FAIR

▼ Event Timeline

 Enable zooming

Executors

■ Added■ Removed

Executor driver added

Jobs

■ Succeeded■ Failed■ RunningSat 8
February 2020

Sun 9

Mon 10

Tue 11

Wed 12

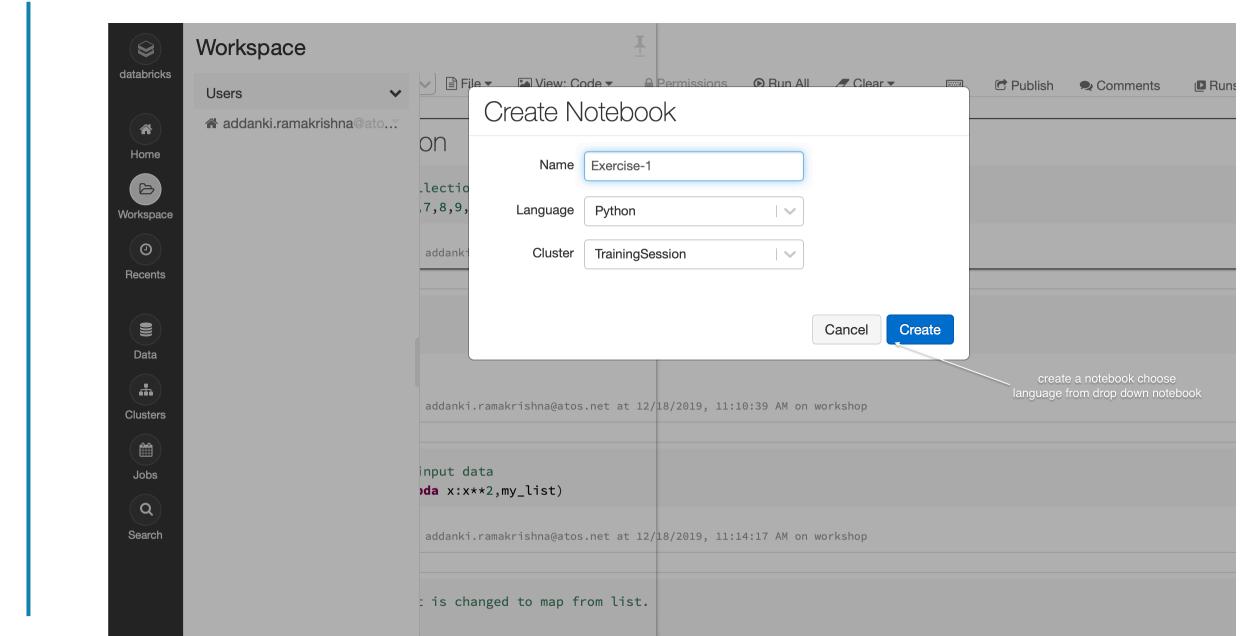
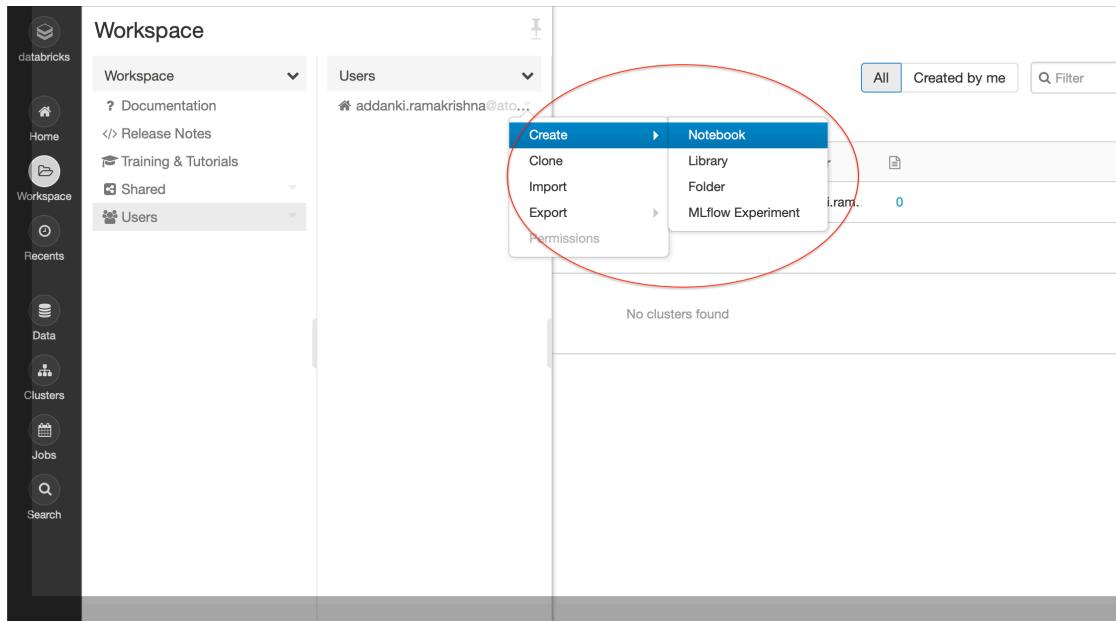
Thu 13

Fri 14

SPARK UI, LOGS & METRICS

The screenshot shows the Databricks interface with a sidebar on the left containing icons for Home, Workspace, Recents, Data, Clusters, Jobs, and Search. The main area is titled 'Clusters / TrainingSession' and shows a 'TrainingSession' configuration. A modal window titled 'Install Library' is open, displaying the 'Library Source' tab ('Upload' is selected) and the 'Library Type' tab ('Python Egg' is selected). A large central area is labeled 'Drop EGG here'. At the bottom right of the modal are 'Cancel' and 'Install' buttons.

INSTALLING LIBRARIES



CREATING NOTEBOOKS

Exercise-1 (Python)

TrainingSession

File View: Code Permissions Run All Clear Publish Comments Runs

Cmd 1

```
1 | 1+1
Out[1]: 2
Command took 0.09 seconds --
```

you can execute your pyspark code over here

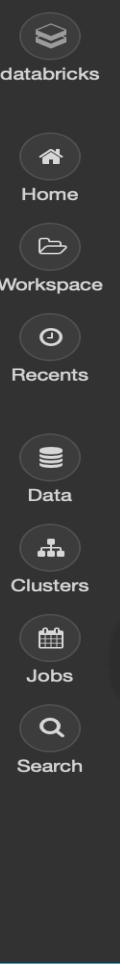
Cmd 2

```
1
```

Shift+Enter to run [shortcuts](#)

The screenshot shows a Databricks notebook interface. On the left is a sidebar with icons for Home, Workspace, Recents, Data, Clusters, Jobs, and Search. The main area is titled 'Exercise-1 (Python)'. It contains two command cells. The first cell, 'Cmd 1', has a code input field containing '1+1' and an output field showing 'Out[1]: 2'. A note says 'you can execute your pyspark code over here'. The second cell, 'Cmd 2', has a code input field containing '1'. At the bottom, it says 'Shift+Enter to run [shortcuts](#)'. The top navigation bar includes options like File, View: Code, Permissions, Run All, Clear, Publish, Comments, and Runs.

EXECUTING CODE IN NOTEBOOK



User Settings

Password Git Integration **Notebook Settings**

Git integration

Databricks supports notebook version control integration with either GitHub or Bitbucket Cloud. To connect to git repositories and make commits, we need a personal access token (for GitHub) or an app password (for Bitbucket Cloud).

Generating tokens

To generate a GitHub personal access token, follow the [GitHub documentation](#). When using GitHub, the token must have the “repo” permission.

To generate a Bitbucket Cloud app password, follow the [Bitbucket Cloud documentation](#). When using Bitbucket Cloud, the app password must have “read” and “write” permission under repository.

For more information on Git integration, see the Databricks documentation on [GitHub integration](#) or [Bitbucket Cloud integration](#).

Git provider

GitHub

Bitbucket Cloud

TOKEN OR APP PASSWORD

Token or app password with repo read/write permis...

Save

choose one as you prefer to use
use token to connect to git or bitbucket

GIT/ BIT BUCKET INTEGRATION OF NOTEBOOK |

QUESTIONS ??



