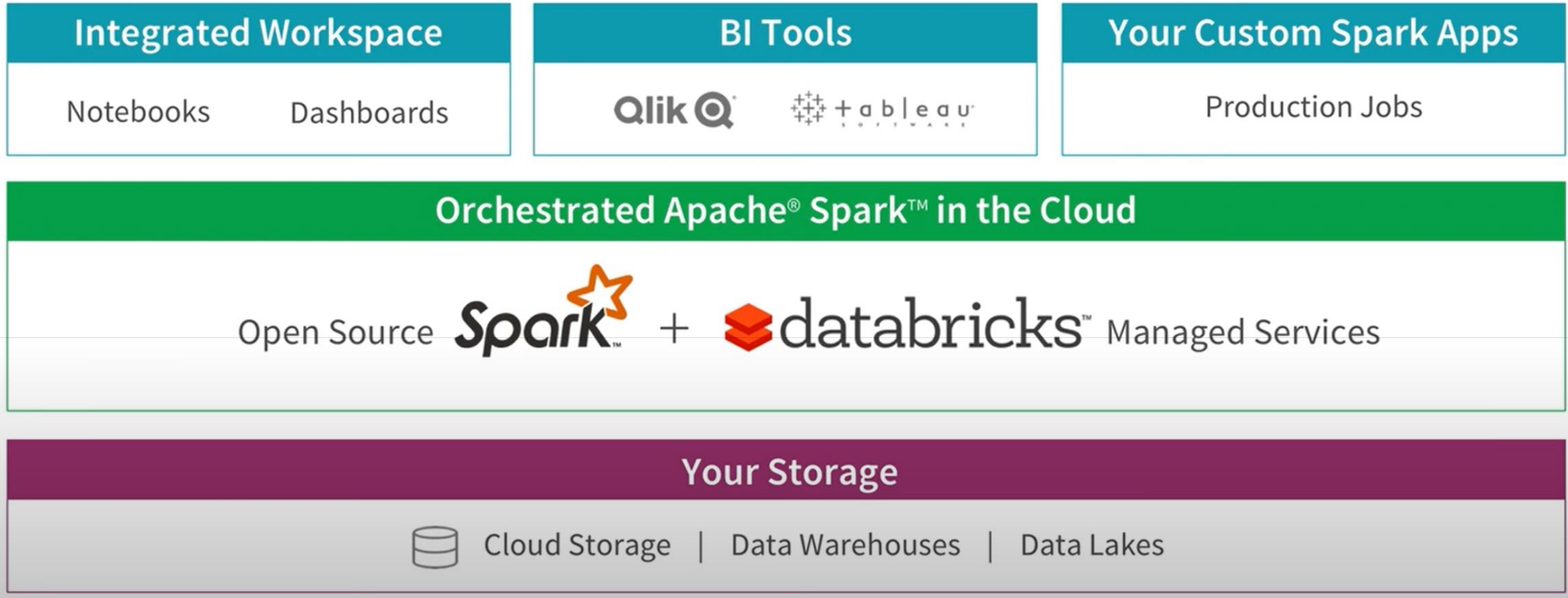
Big Data Tools and Techniques

Chapter 7: Introduction to Spark (Part C)

Introduction to Databricks



Login - Databricks Community Ec

community.cloud.databricks.com/login.html

Important Notice: Acceptable use and unused account termination policy and Terms of Use update.

databricks

Sign In to Databricks
Community Edition

Email / Username

Password

Forgot Password?

Sign In

New to Databricks? Sign Up.

Privacy Policy | Terms of Use

Search for anything

ENG

2:55 PM
6/6/2020

Try Databricks

An open and unified data analytics platform for data engineering, machine learning, and analytics

From the original creators of Apache Spark™, Delta Lake, MLflow, and Koalas

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...

Databricks Community Edition

community.cloud.databricks.com/

Upgrade

databricks

Home

Workspace

Recents

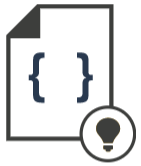
Data

Clusters

Jobs

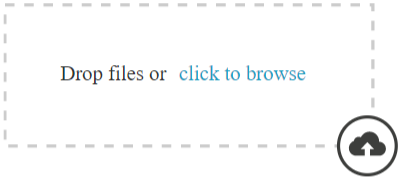
Search

Welcome to databricks



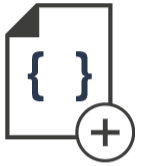
Explore the Quickstart Tutorial

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.

Common Tasks

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

Recents

- Quickstart Notebook

What's new in v3.21

View latest release notes

Search for anything

ENG

3:03 PM

6/6/2020

Databricks Community Edition

community.cloud.databricks.com/

databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Workspace

Users

Quickstart Notebook

Upgrade

atabricks

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.

splay

Recents

Quickstart Notebook

What's new in v3.21

View latest release notes

Search for anything

ENG

3:11 PM

6/6/2020

Databricks Community Edition

community.cloud.databricks.com/

databricks

Home

Workspace

Recents

Data

Clusters

Jobs

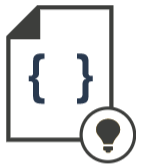
Search

Recent

Quickstart Notebook

Upgrade

ome to databricks



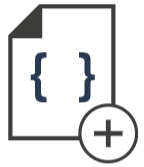
xplore the Quickstart Tutorial

ter, 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.

ks

ook

e

r

Experiment

ry

mentation

Recents

Quickstart Notebook

What's new in v3.21

View latest release notes

Search for anything

ENG

3:12 PM

6/6/2020

Databricks Community Edition

community.cloud.databricks.com/

Search Workspace

databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Upgrade

ome to databricks

xplore the Quickstart Tutorial

ter, 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.

ks

ook

e

r

Experiment

ry

mentation

Recents

Quickstart Notebook

My First Notebook

What's new in v3.21

View latest release notes

Quickstart Notebook.sql

Show all

Search for anything

ENG

3:12 PM

6/6/2020

Create Cluster - Databricks Comm

community.cloud.databricks.com/

databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Create Cluster

New Cluster

Cancel | Create Cluster

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

Cluster Name

My First Spark Cluster

UI | JSON

Databricks Runtime Version

Runtime: 6.5 (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.

Instances | Spark

Availability Zone

us-west-2c

Search for anything

ENG 3:13 PM 6/6/2020

databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search

Clusters

+ Create Cluster

All | Created by me

Filter

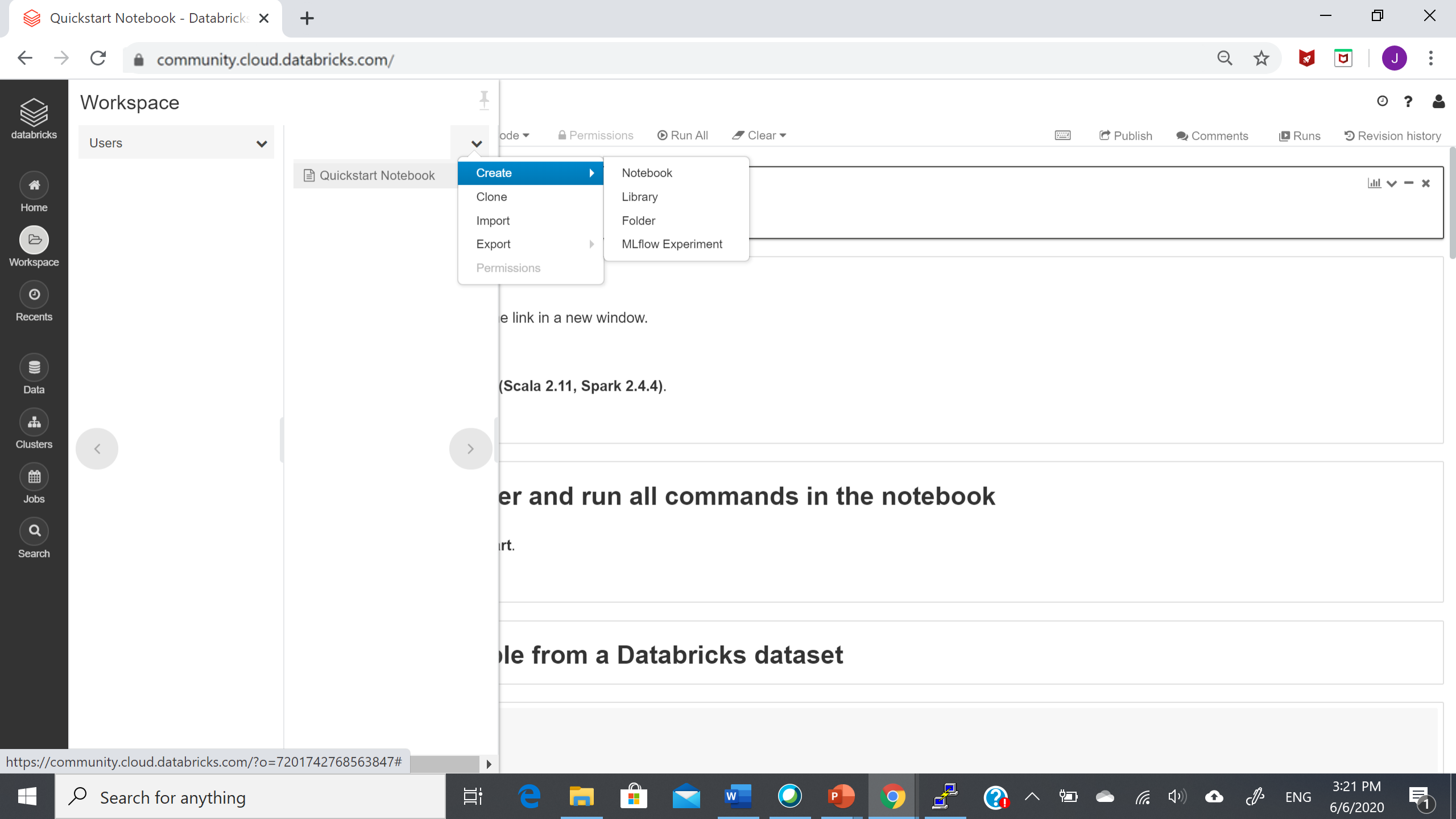
All-Purpose Clusters

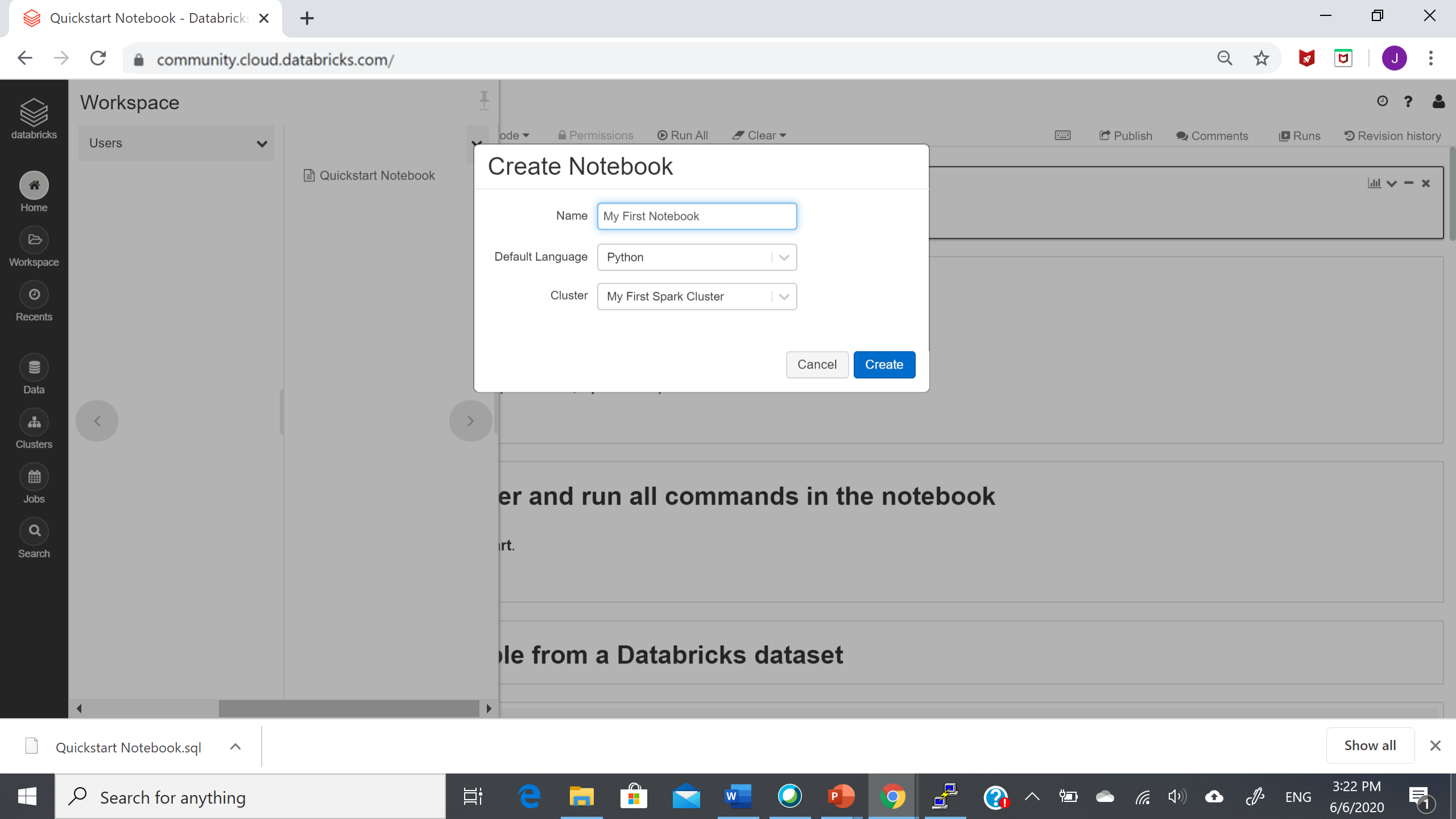
1 clusters, 0 pinned

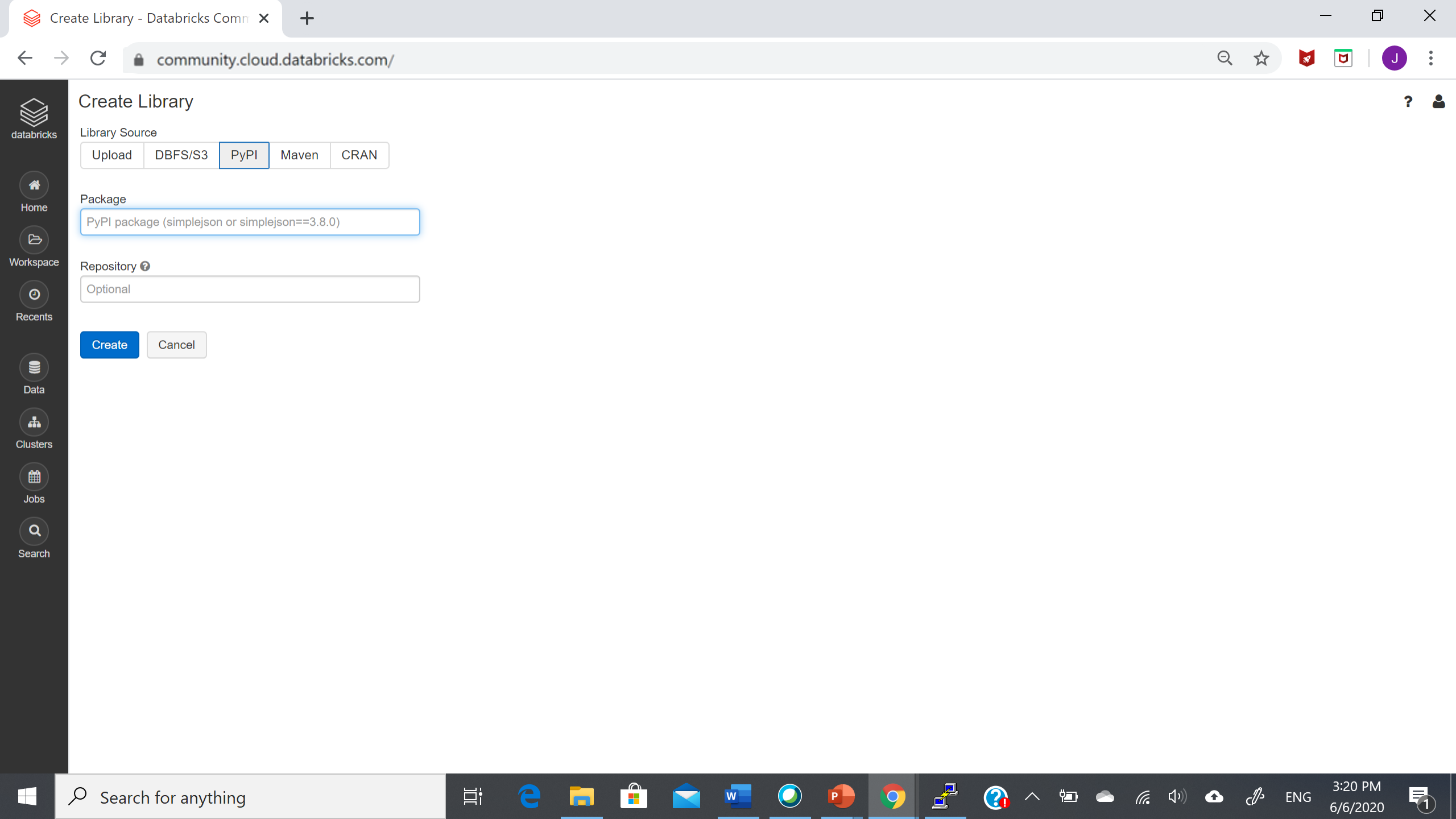
Name	State	Nodes	Driver	Worker	Runtime	Creator		Actions
My First Spark Cluster	Pending	0	Community Optimize	Community Optimize	6.5 (includes Apache ..		0	...

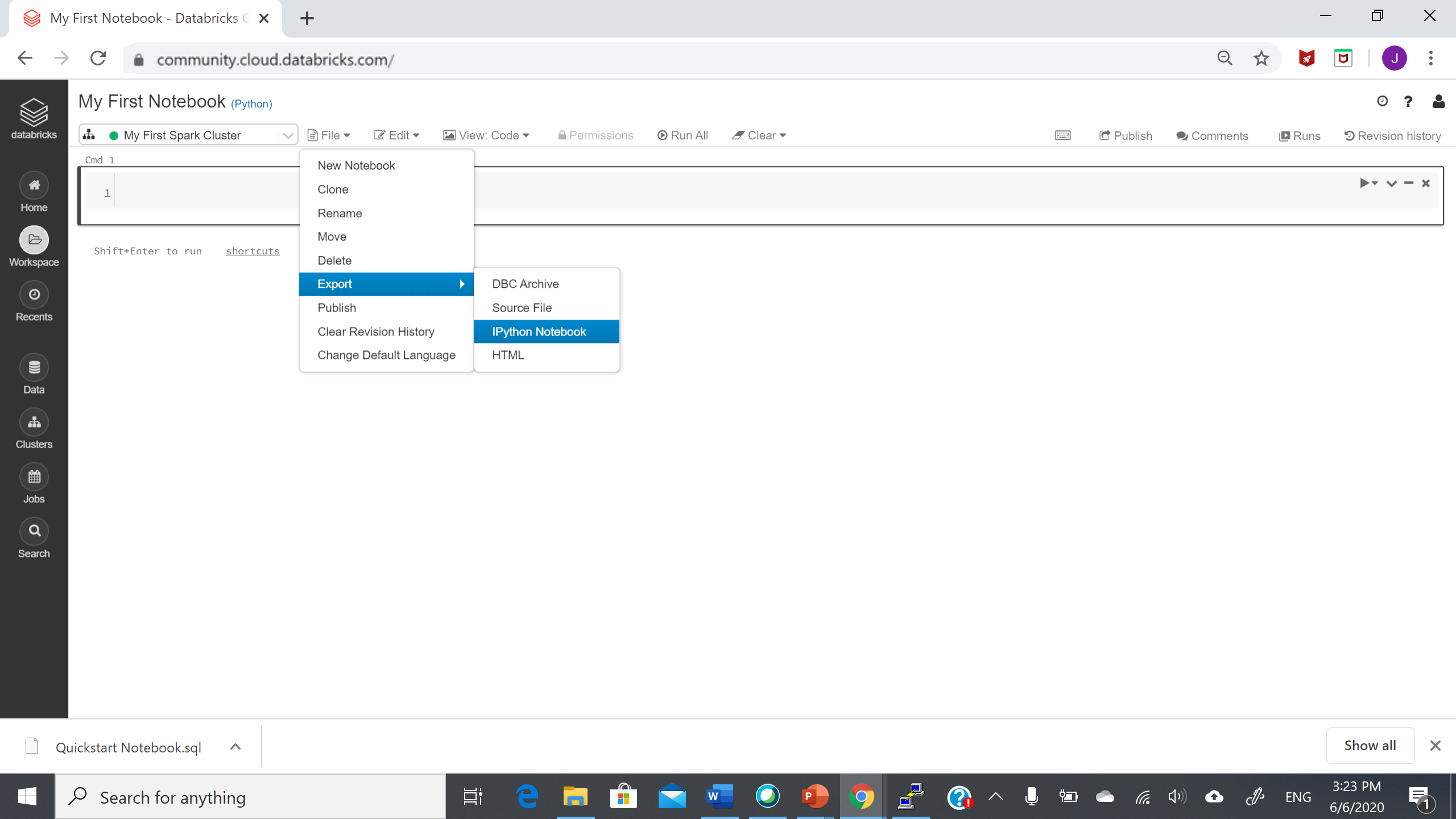
Job Clusters

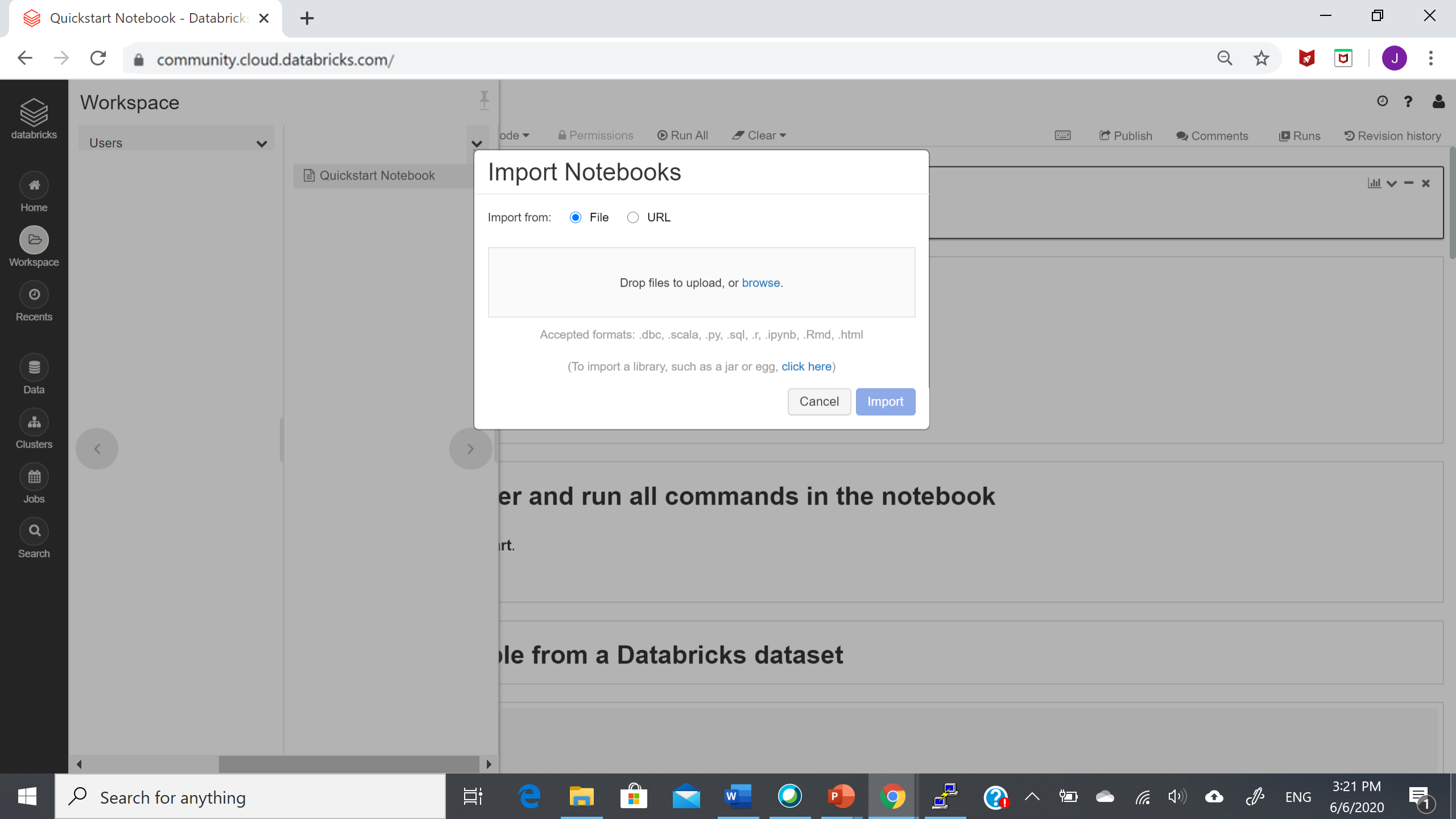
No clusters found


Quickstart Notebook - Databricks
community.cloud.databricks.com/
Workspace
databricks
Users
Quickstart Notebook
Create
Clone
Import
Export
Permissions
Notebook
Library
Folder
MLflow Experiment
Permissions
Run All
Clear
Publish
Comments
Runs
Revision history
Home
Workspace
Recents
Data
Clusters
Jobs
Search
e link in a new window.
(Scala 2.11, Spark 2.4.4).
er and run all commands in the notebook
rt.
le from a Databricks dataset
https://community.cloud.databricks.com/?o=7201742768563847#
Search for anything
ENG
3:21 PM
6/6/2020


Quickstart Notebook - Databricks
community.cloud.databricks.com/
Workspace
databricks
Users
Quickstart Notebook
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Permissions
Run All
Clear
Publish
Comments
Runs
Revision history
Create Notebook
Name
My First Notebook
Default Language
Python
Cluster
My First Spark Cluster
Cancel
Create
er and run all commands in the notebook
rt.
le from a Databricks dataset
Quickstart Notebook.sql
Show all
Search for anything
ENG
3:22 PM
6/6/2020


Create Library - Databricks Comm
community.cloud.databricks.com/
databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Create Library
Library Source
Upload
DBFS/S3
PyPI
Maven
CRAN
Package
PyPI package (simplejson or simplejson==3.8.0)
Repository
Optional
Create
Cancel
Search for anything
ENG
3:20 PM
6/6/2020


My First Notebook - Databricks
community.cloud.databricks.com/
databricks
My First Notebook (Python)
My First Spark Cluster
File
Edit
View: Code
Permissions
Run All
Clear
Publish
Comments
Runs
Revision history
Cmd 1
1
Shift+Enter to run
shortcuts
Home
Workspace
Recents
Data
Clusters
Jobs
Search
New Notebook
Clone
Rename
Move
Delete
Export
Publish
Clear Revision History
Change Default Language
DBC Archive
Source File
IPython Notebook
HTML
Quickstart Notebook.sql
Show all
Search for anything
ENG
3:23 PM
6/6/2020


Quickstart Notebook - Databricks
community.cloud.databricks.com/
databricks
Workspace
Users
Quickstart Notebook
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Permissions
Run All
Clear
Publish
Comments
Runs
Revision history
Import Notebooks
Import from:
File
URL
Drop files to upload, or browse.
Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .Rmd, .html
(To import a library, such as a jar or egg, click here)
Cancel
Import
er and run all commands in the notebook
rt.
le from a Databricks dataset
Search for anything
ENG
3:21 PM
6/6/2020

Quickstart Notebook - Databricks

community.cloud.databricks.com/

databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Quickstart Notebook (SQL)

Detached | File | Edit | View: Code | Permissions | Run All | Clear | Publish | Comments | Runs | Revision history

Cmd 1

Databricks in 5 minutes

Cmd 2

Create a quickstart cluster

1. In the sidebar, right-click the **Clusters** button and open the link in a new window.
2. On the Clusters page, click **Create Cluster**.
3. Name the cluster **Quickstart**.
4. In the Databricks Runtime Version drop-down, select **6.3 (Scala 2.11, Spark 2.4.4)**.
5. Click **Create Cluster**.

Cmd 3

Attach the notebook to the cluster and run all commands in the notebook

1. Return to this notebook.
2. In the notebook menu bar, select Detached > **Quickstart**.
3. When the cluster changes from to ●, click **Run All**.

Cmd 4

The next command creates a table from a Databricks dataset

Cmd 5

```
1 DROP TABLE IF EXISTS diamonds;
2
3 CREATE TABLE diamonds
```

Search for anything

ENG 3:05 PM 6/6/2020

Quickstart Notebook (SQL)

Cmd 4

The next command creates a table from a Databricks dataset

Cmd 5

```
1 DROP TABLE IF EXISTS diamonds;
2
3 CREATE TABLE diamonds
4 USING csv
5 OPTIONS (path "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv", header "true")
6
```

OK

Command took 46.18 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 6

```
1 SELECT * from diamonds
```

_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

Showing the first 1000 rows.

Command took 6.11 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 7

```
1 %python
```

Quickstart Notebook - Databricks

community.cloud.databricks.com/

Quickstart Notebook (SQL)

Detached | File | Edit | View: Code | Permissions | Run All | Clear | Publish | Comments | Runs | Revision history

databricks | Home | Workspace | Recents | Data | Clusters | Jobs | Search

Cmd 7

```
%python
diamonds = spark.read.csv("/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv", header="true", inferSchema="true")
diamonds.write.format("delta").save("/delta/diamonds")
```

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

Command took 13.50 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 8

```
DROP TABLE IF EXISTS diamonds;

CREATE TABLE diamonds USING DELTA LOCATION '/delta/diamonds/'
```

OK

Command took 0.77 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 9

```
SELECT * from diamonds
```

_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

Showing the first 1000 rows.

Command took 6.89 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 10

Search for anything

ENG 3:06 PM 6/6/2020

Quickstart Notebook (SQL)

Command took 6.89 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 10

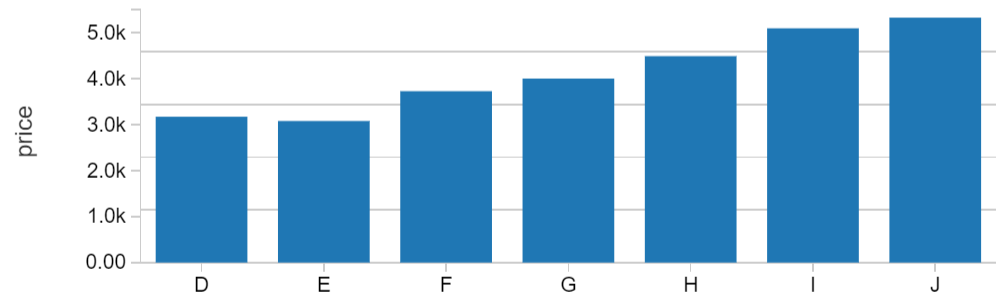
The next command manipulates the data and displays the results

Specifically, the command:

1. Selects color and price columns, averages the price, and groups and orders by color.
2. Displays a table of the results.

Cmd 11

```
1  SELECT color, avg(price) AS price FROM diamonds GROUP BY color ORDER BY color
```



Plot Options...

Command took 2.83 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 12

Convert the table to a chart

Quickstart Notebook - Databricks

community.cloud.databricks.com/ ...067445025/command/1158110067445026

databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Quickstart Notebook (SQL)

Detached | File | Edit | View: Code | Permissions | Run All | Clear | Publish | Comments | Runs | Revision history

Convert the table to a chart

Under the table, click the bar chart icon.

Cmd 13

Repeat the same operations using Python DataFrame API.

This is a SQL notebook; by default command statements are passed to a SQL interpreter. To pass command statements to a Python interpreter, include the `%python` magic command.

Cmd 14

The next command creates a DataFrame from a Databricks dataset

Cmd 15

```
%python
diamonds = spark.read.csv("/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv", header="true", inferSchema="true")
```

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

Command took 1.67 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Cmd 16

The next command manipulates the data and displays the results

Cmd 17

```
%python
from pyspark.sql.functions import avg

display(diamonds.select("color","price").groupBy("color").agg(avg("price")).sort("color"))
```

Search for anything

ENG 3:08 PM 6/6/2020

Quickstart Notebook - Databricks

community.cloud.databricks.com/

Quickstart Notebook (SQL)

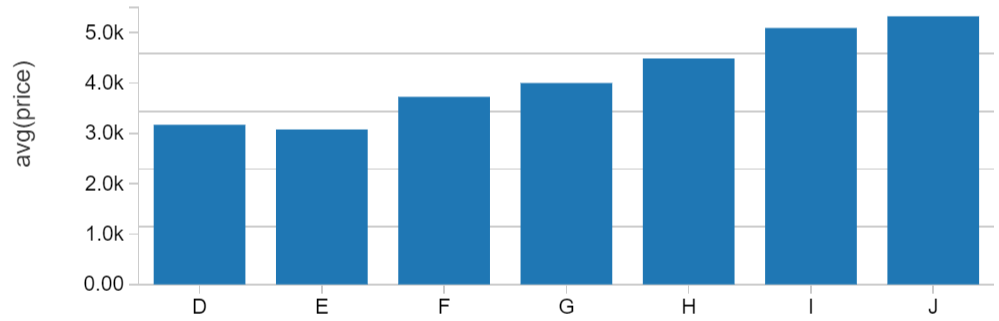
Detached | File | Edit | View: Code | Permissions | Run All | Clear | Publish | Comments | Runs | Revision history

databricks | Home | Workspace | Recents | Data | Clusters | Jobs | Search

The next command manipulates the data and displays the results

Cmd 17

```
1 %python
2 from pyspark.sql.functions import avg
3
4 display(diamonds.select("color","price").groupBy("color").agg(avg("price")).sort("color"))
```



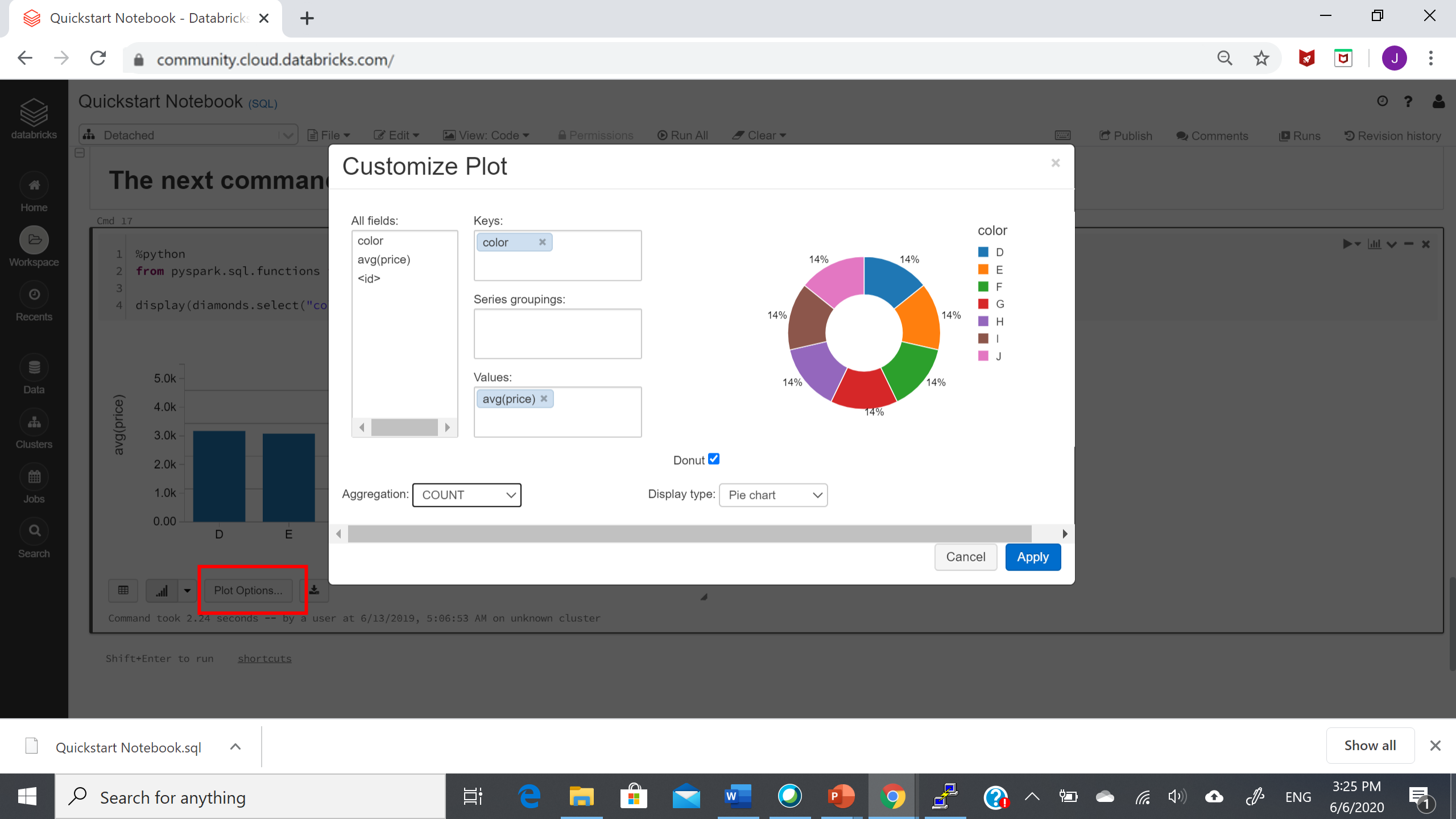
Plot Options...

Command took 2.24 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster

Shift+Enter to run shortcuts

Search for anything

ENG 3:10 PM 6/6/2020


Quickstart Notebook - Databricks
community.cloud.databricks.com/
Quickstart Notebook (SQL)
Detached
File
Edit
View: Code
Permissions
Run All
Clear
Publish
Comments
Runs
Revision history
databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
The next comman
Cmd 17
%python
from pyspark.sql.functions
display(diamonds.select("co
avg(price)
5.0k
4.0k
3.0k
2.0k
1.0k
0.00
D
E
Plot Options...
Command took 2.24 seconds -- by a user at 6/13/2019, 5:06:53 AM on unknown cluster
Shift+Enter to run
shortcuts
Customize Plot
All fields:
color
avg(price)
<id>
Keys:
color
Series groupings:
Values:
avg(price)
14%
14%
14%
14%
14%
14%
14%
color
D
E
F
G
H
I
J
Donut
Aggregation:
COUNT
Display type:
Pie chart
Cancel
Apply
Quickstart Notebook.sql
Show all
Search for anything
ENG
3:25 PM
6/6/2020

databricks

Home

Workspace

Recent

Tables

Clusters

Jobs

Apps

Search

Settings

Mobile Sample (SQL)

Attached: reporting | View: Code | File | Permissions | Run All | Comments | Revision history

```
> select m.ClientID, c.CountryCode3, m.SessionID
  from mobile_sample m
     join countrycodes c
        on m.Country = c.CountryName
```

(4) Spark Jobs

ClientID	CountryCode3	SessionID	DeviceMake
96046	ARG	0	ASUS
96987	ARG	0	ASUS
127767	ARG	0	ASUS
3846	AUS	0	SAMSUNG
8509	AUS	0	Apple
10085	AUS	0	HTC
14719	AUS	0	HTC
15286	AUS	0	Apple
15286	AUS	0	Apple
15286	AUS	0	Apple
15286		0	Apple
15286		0	Apple
15286		0	Apple
15286		0	Apple
15286		0	Apple
15286		0	Apple

Bar | Quantile | Scatter | Histogram | Map | Box plot | Line | Q-Q plot | Area | Pivot | Pie

Command took 2.75s

databricks

Home

Workspace

Recent

Tables

Clusters

Jobs

Apps

Search

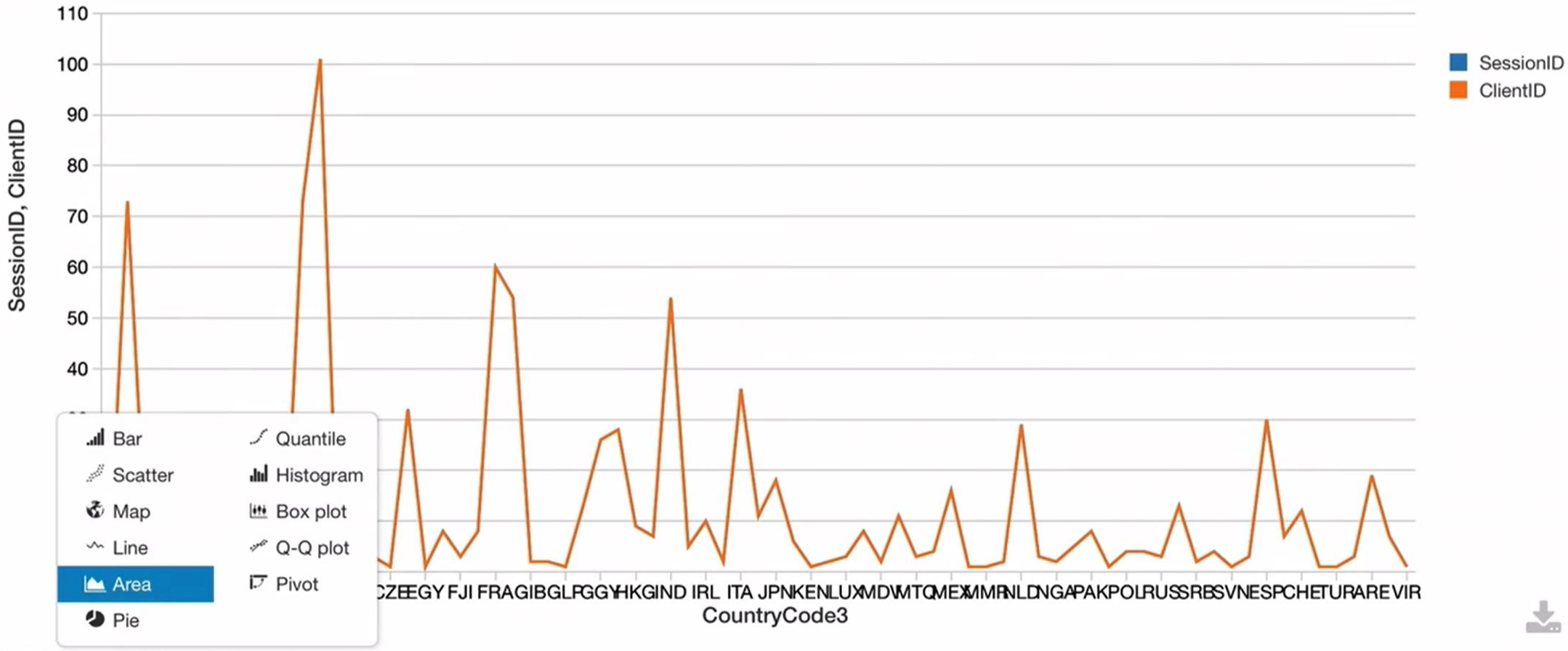
Settings

Mobile Sample (SQL)

Attached: reporting | View: Code | File | Permissions | Run All | Comments | Revision history

```
> select m.ClientID, c.CountryCode3, m.SessionID
  from mobile_sample m
     join countrycodes c
        on m.Country = c.CountryName
```

(4) Spark Jobs



Plot Options...

databricks

Home

Workspace

Recent

Tables

Clusters

Jobs

Apps

Search

Settings

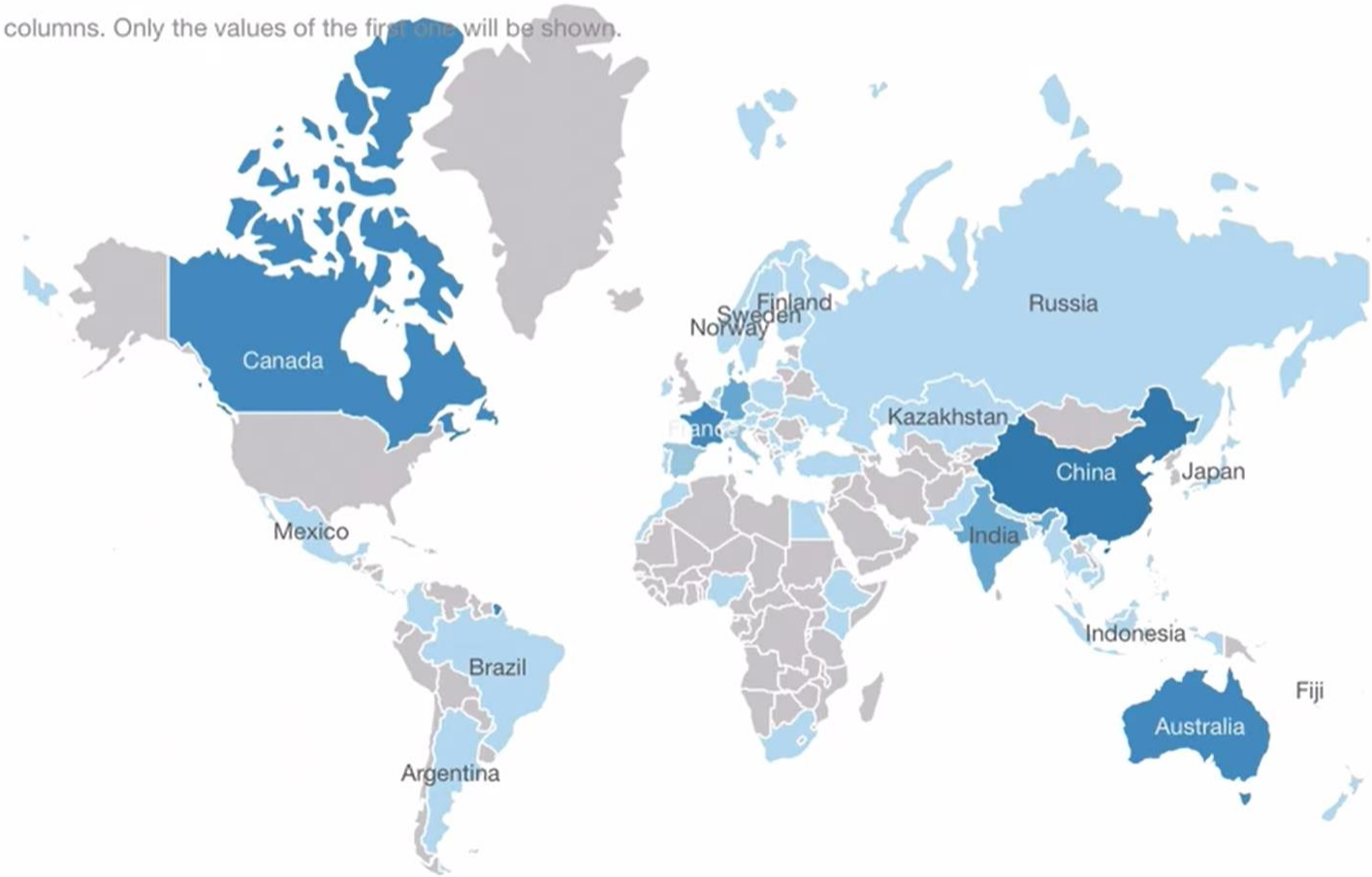
Mobile Sample (SQL)

Attached: reporting | View: Code | File | Permissions | Run All | Comments | Revision history

```
> select m.ClientID, c.CountryCode3, m.SessionID
  from mobile_sample m
     join countrycodes c
        on m.Country = c.CountryName
```

(4) Spark Jobs

More than two numerical columns. Only the values of the first one will be shown.



Plot Options...