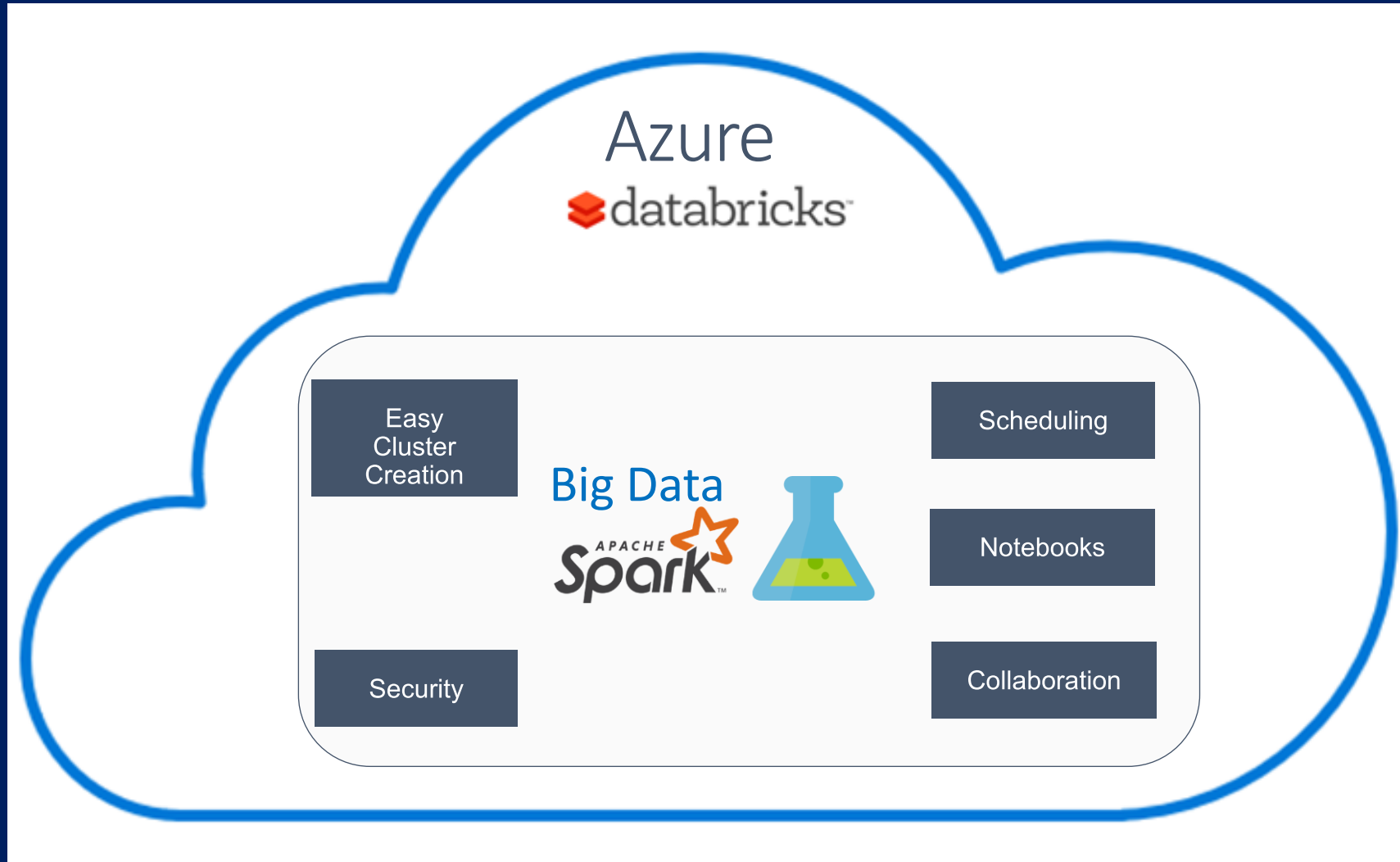


Getting Started



Getting Started with Databricks

- Log Into Databricks.
- Create a Databricks Spark Cluster.
- Upload a File to a Table. **Slides below explain this.**
- Define the Table. Slides below explain this.
- Create or Import a Notebook.
- Using Notebooks. Slides below explain.

Creating a Table from the Lab Files

Creating a Table

The screenshot shows the Microsoft Azure Databricks interface. On the left is a sidebar with navigation icons for Azure Databricks, Home, Workspace, Recents, Data, Clusters, Jobs, and Search. The main area is titled 'Data' and contains two panels: 'Databases' and 'Tables'. The 'Databases' panel shows a list of databases including 'chronic_disease_...', 'default', and 'mydb'. The 'Tables' panel shows a list of tables including 'address_table...', 'birthwt', 'birthwt2', 'birthwt256', 'birthwt_csv', 'birthwt_csv2', 'birthwt_dq', 'birthwtdf', 'bleedingdata', 'bleedingdata12', 'bleedingdata123', and 'bleedingdatatest'. An 'Add Data' button is located at the top right of the 'Tables' panel. A blue line points from the 'Data' icon in the sidebar to the 'Data' section header. Another blue line points from the 'Add Data' button to a purple callout box on the right.

Microsoft Azure

Azure Databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Data

Databases

Filter Databases

chronic_disease_...

default

mydb

Tables

Filter Tables

address_table...

birthwt

birthwt2

birthwt256

birthwt_csv

birthwt_csv2

birthwt_dq

birthwtdf

bleedingdata

bleedingdata12

bleedingdata123

bleedingdatatest

Add Data

Permissions

Run All

SQL DB...

bc.SQLServerDriver")

er.jdbc.SQLServerDriver

m at 6/23/2018, 7:24:59 PM on

ase.windows.net"

the user and password pa

Click Add Data

Click on Data

Creating a Table

Microsoft Azure

Azure Databricks

Home

Workspace

Recents

Data

Create New Table (Python)

Data source ?

Upload File DBFS Other Data Sources

Upload to DBFS ?

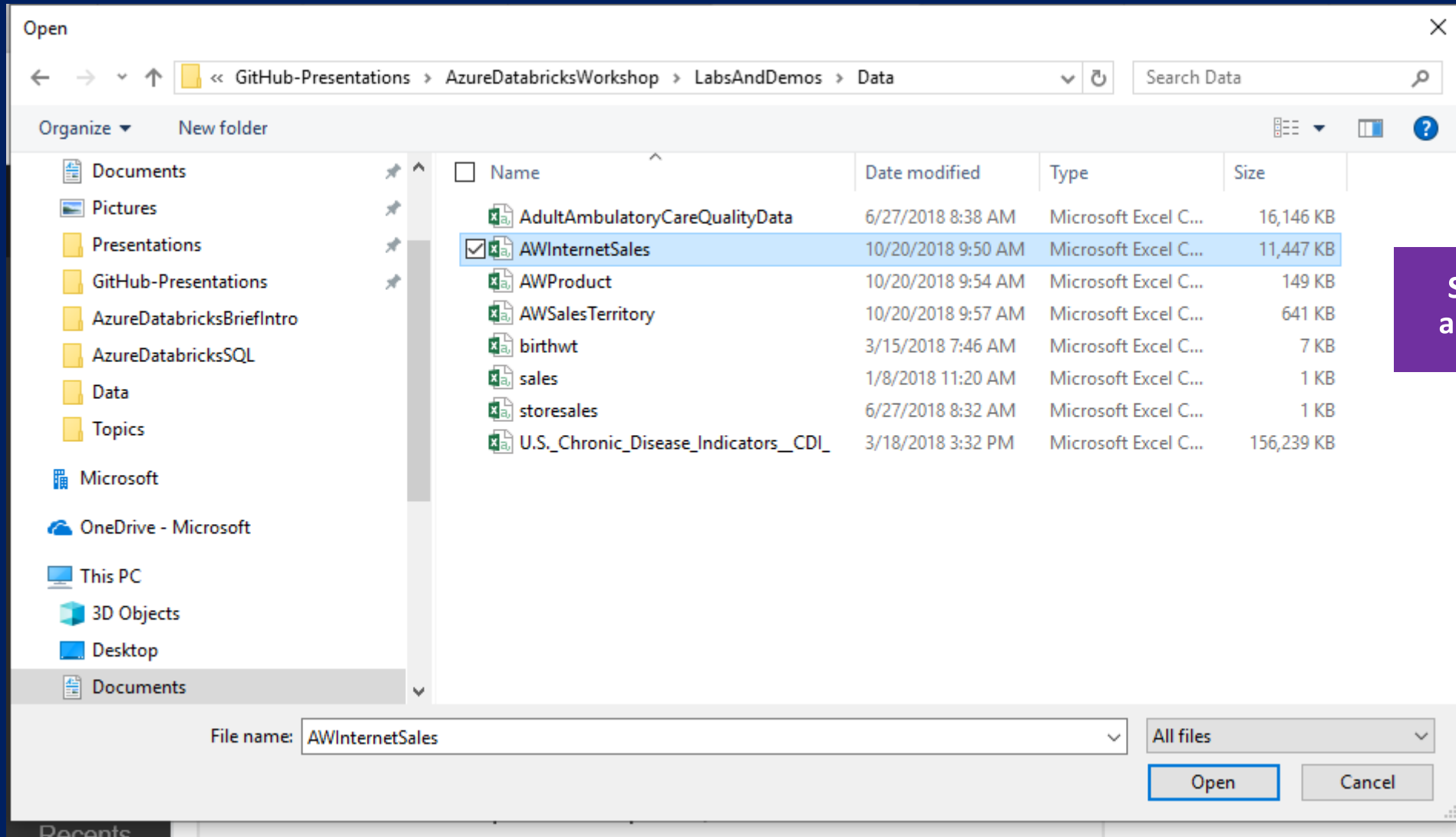
/FileStore/tables/ (optional) Select

File ?

Drop files to upload, or [browse](#).

Click to Upload

Uploading Data



Select the file
and Click Open

Creating a Table

Microsoft Azure

Azure Databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Create New Table (Python)

Data source ?

Upload File DBFS Other Data Sources

Upload to DBFS ?

/FileStore/tables/ (optional) Select

File ?

AWInternetSales

11.7 MB

[Remove file](#)

✓ File uploaded to /FileStore/tables/AWInternetSales.csv

Create Table with UI Create Table in Notebook ?

Click Create Table
with UI

Creating a Table

11.7 MB
[Remove file](#)

✓ File uploaded to `/FileStore/tables/AWInternetSales.csv`

Create Table with UI↗ Create Table in Notebook ?

Select a Cluster to Preview the Table

Choose a cluster with which you will read and preview the data.

Cluster ?

-

✓ -
democlustermedium (42 GB, Running, 4.3 (includes Apache S...

Select a Cluster and
then Click Preview

Creating a Table

Rename Table by
Removing the _csv

Check First Row is
Header and Infer
Schema

Click Create Table

Microsoft Azure

Create New Table

[Preview Table](#)

Specify Table Attributes

Specify the Table Name, Database and Schema to add this to the data UI for other users to access

Table Name [?] awinternetsales

Create in Database [?] default

File Type [?] CSV

Column Delimiter [?] ,

☒ First row is header [?]

☒ Infer schema [?]

☐ Multi-line [?]

[Create Table](#)


Table Preview

| ProductKey | OrderDateKey |
|------------|--------------|
| INT | INT |
| 310 | 20101229 |
| 346 | 20101229 |
| 346 | 20101229 |
| 336 | 20101229 |
| 346 | 20101229 |
| 311 | 20101230 |


Table Ready

Your Table is
Ready!


Microsoft Azure




Azure Databricks




Home




Workspace



Recents




Data




Clusters

Table: awinternetsales

awinternetsales

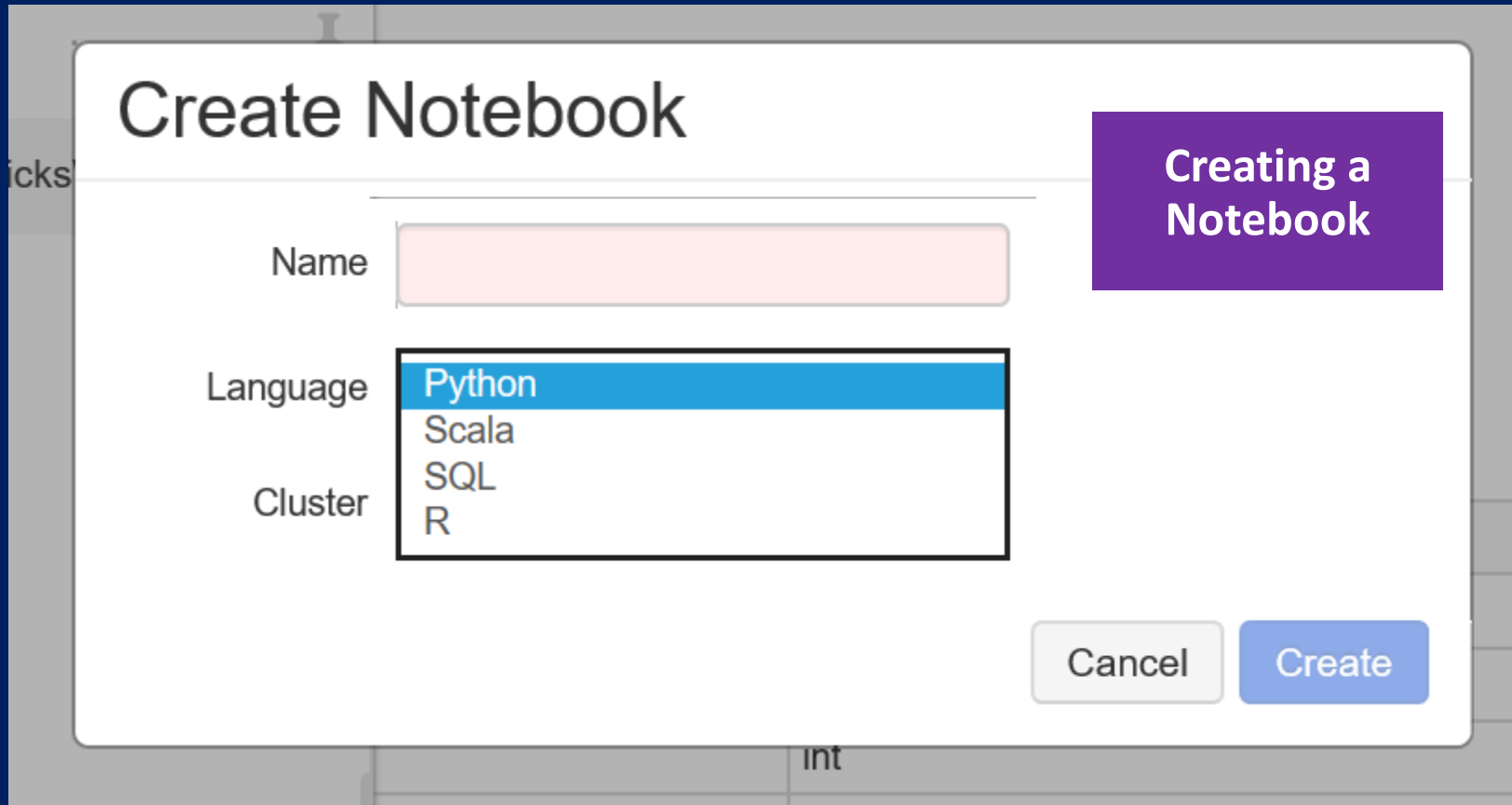
 Refresh

democlustermedium (42 GB, Running, 4.3 (includes Apache ... 

Schema:

| col_name | data_type |
|--------------|-----------|
| ProductKey | int |
| OrderDateKey | int |
| DueDateKey | int |
| ShipDateKey | int |
| CustomerKey | int |
| PromotionKey | int |
| CurrencyKey | int |

Create a Notebook



A screenshot of a 'Create Notebook' dialog box. The dialog has a title bar 'Create Notebook'. It contains three input fields: 'Name' (a text box), 'Language' (a dropdown menu with 'Python' selected), and 'Cluster' (a dropdown menu with 'SQL' selected). There are 'Cancel' and 'Create' buttons at the bottom right. A purple callout box with the text 'Creating a Notebook' is positioned to the right of the dialog.

icks

Create Notebook

Name

Language Python

Cluster SQL

Cancel Create

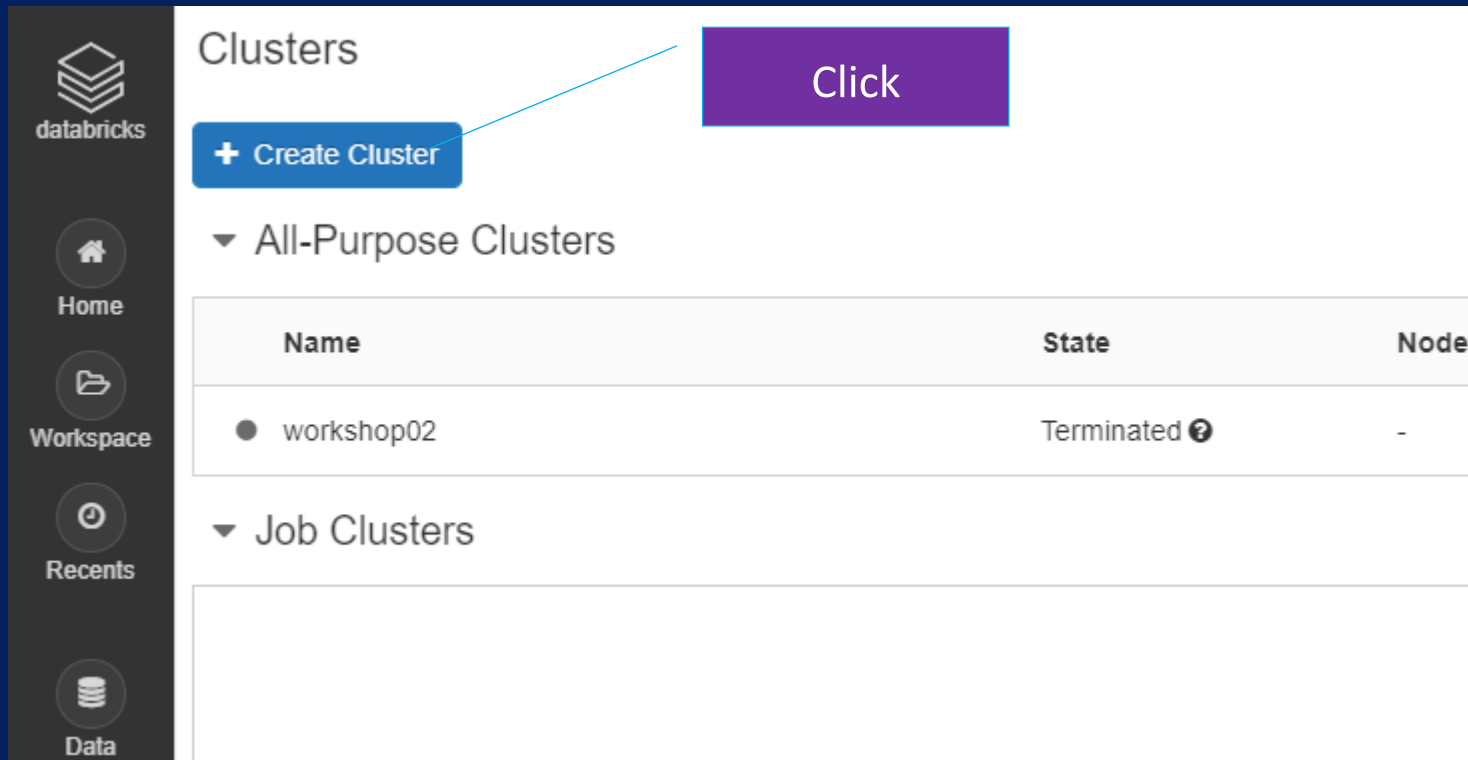
int

Creating a Notebook

Create a Databricks Cluster

Create a Databricks Cluster

First, Log into Databricks





The screenshot shows the Databricks Clusters management interface. On the left is a dark sidebar with navigation icons and labels: 'databricks' (logo), 'Home' (house icon), 'Workspace' (folder icon), 'Recents' (clock icon), and 'Data' (database icon). The main content area is titled 'Clusters'. Below the title is a blue button labeled '+ Create Cluster'. A red arrow points from a red box containing the word 'Click' to this button. Below the button is a section titled '▼ All-Purpose Clusters'. This section contains a table with the following data:


| Name | State | Node |
|--------------|--------------|------|
| ● workshop02 | Terminated ? | - |


Below the table is a section titled '▼ Job Clusters', which is currently empty.


Create a Databricks Cluster



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

mycluster01

Databricks Runtime Version ?

Runtime: 6.5 (Scala 2.11, Spark 2.4.5) | v

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 | v

Click

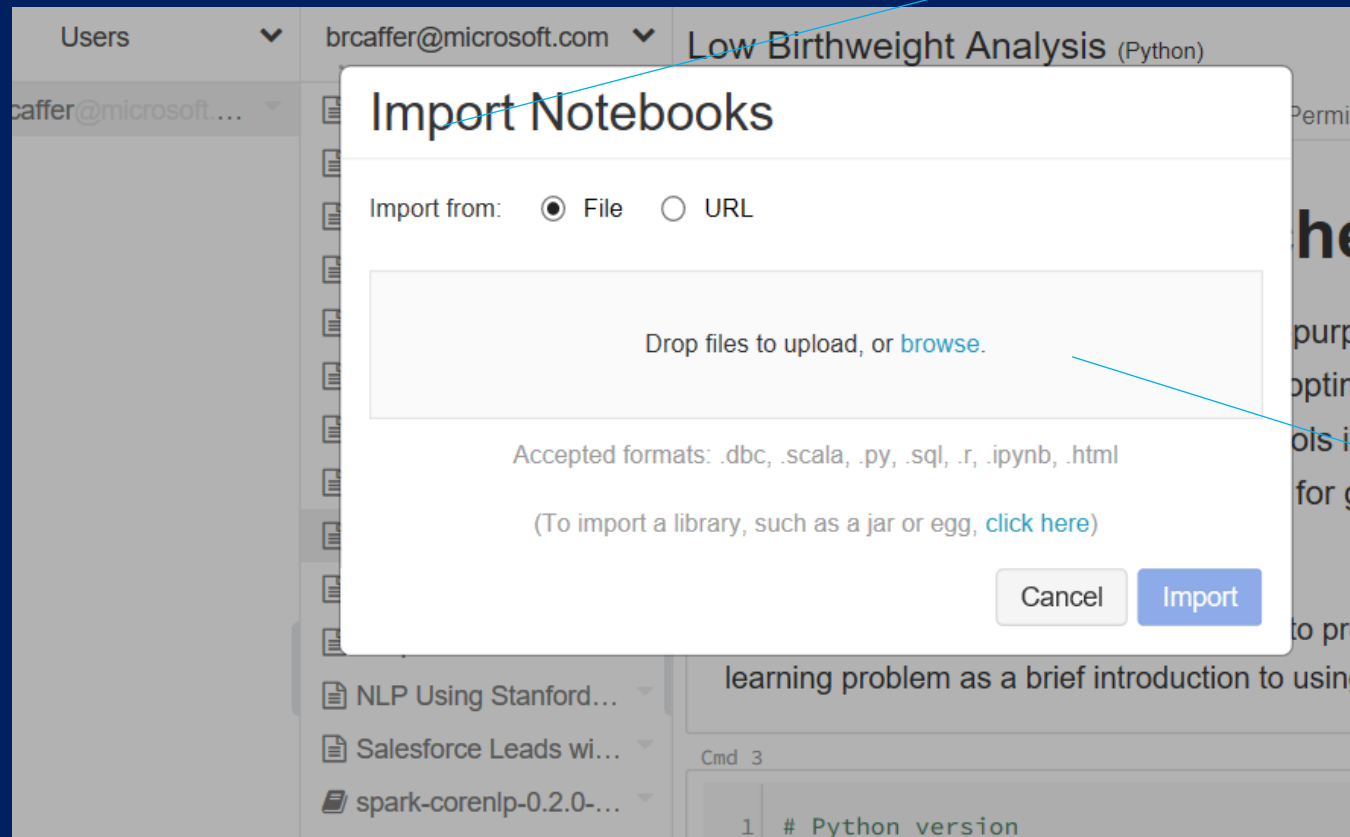
Importing Notebooks

Importing Notebooks

Under Workspace/Users/User
– Select Import from the
Dropdown

The screenshot displays the Databricks Workspace interface. On the left is a dark sidebar with navigation icons and labels: 'databricks', 'Home', 'Workspace', 'Recents', and 'Data'. The main content area is titled 'Workspace'. It features a 'Users' dropdown menu with two entries: 'bryan256@msn.com' and 'brcaffer@microsoft.com'. The 'brcaffer@microsoft.com' entry is selected, and its dropdown menu is open, showing options: 'Create', 'Clone', 'Import' (highlighted in blue), 'Export', and 'Permissions'. Below the user selection, a list of notebooks is visible, including 'Accessing_AW_SQLDB', 'cameron', 'DatabricksWorkshop', 'debug1', 'Dmeo1', 'FinalProject_solution', 'FinalProject_solution_final', and 'pysqldf'. To the right of the notebooks, a table lists workspace configurations with columns for memory, cores, and DBUs. A blue line points from the 'Import' option in the dropdown menu to the text box in the top right corner.

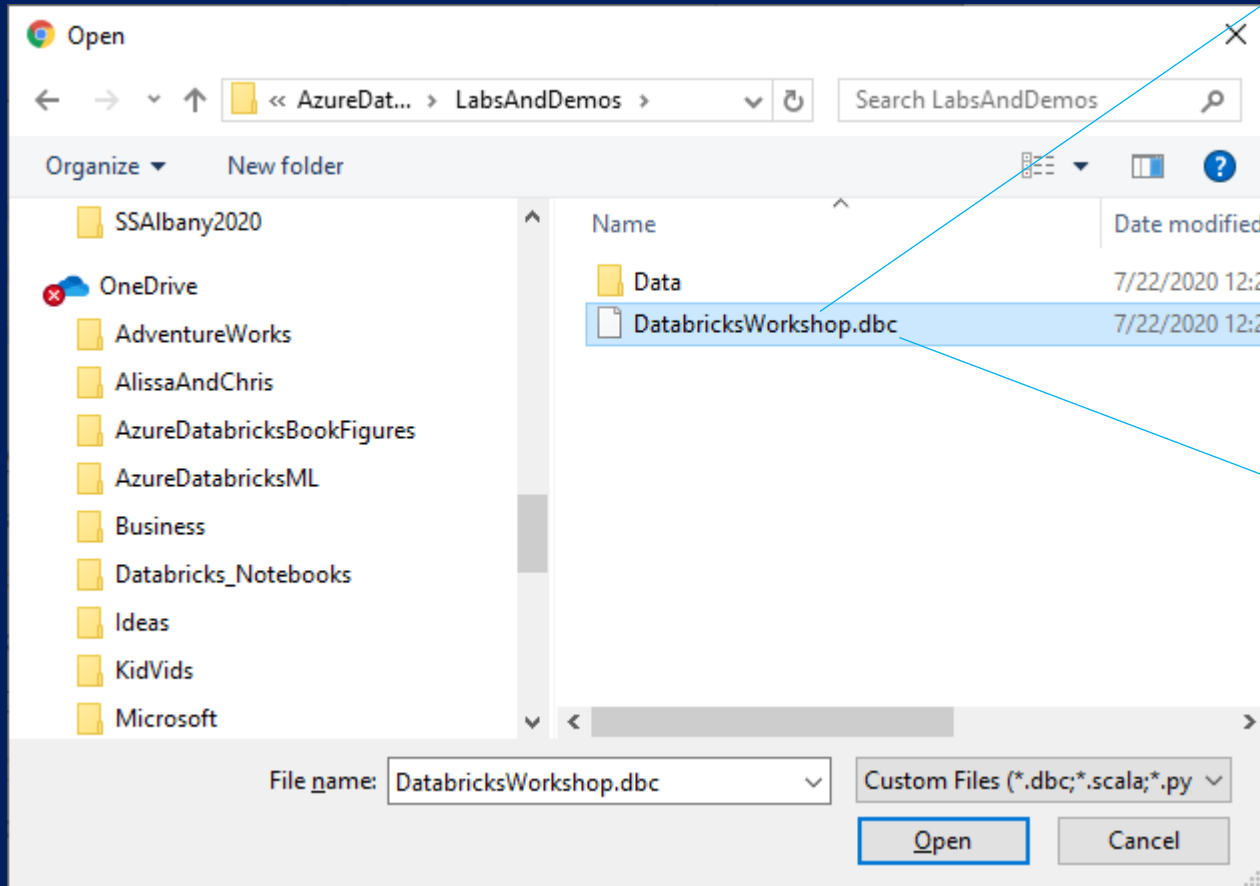
Importing Notebooks



From file or URL

Click to upload
a notebook

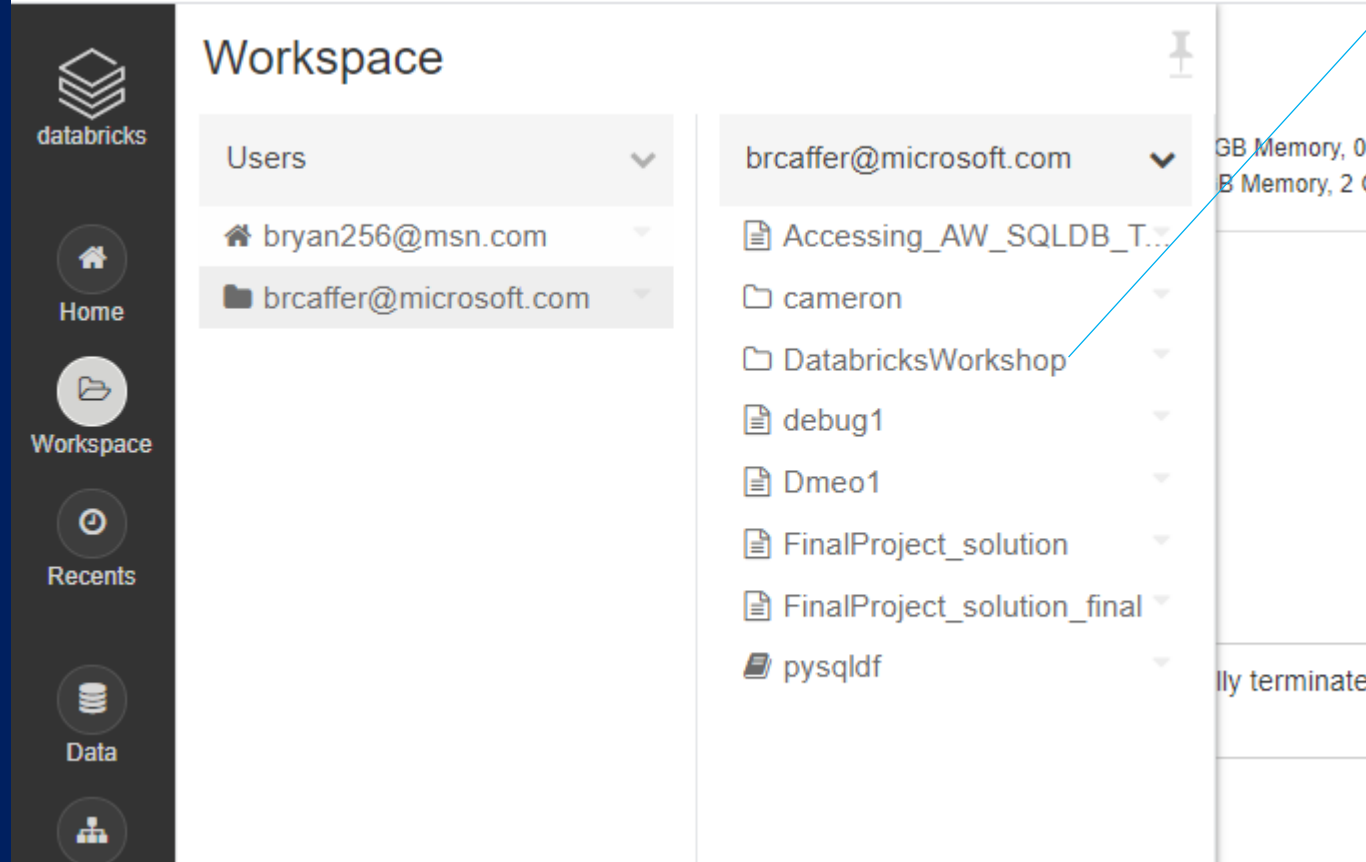
Importing Notebooks



Select file and
Click Open.

DBC extension is
a Databricks
Compressed
Notebook file.

Importing Notebooks



Imported folder.

Importing Notebooks

Azure Databricks Notebooks - Exploring



Robert F. Balazik

Azure Databrick Notebooks

Toolbar

The screenshot displays the Azure Databricks Notebook interface. At the top, there's a browser window with the URL `https://eastus.azuredatabricks.net/?o=1405530884703666#notebook/3648354168357616/command/3648354168357617`. Below the browser, the Microsoft Azure portal header is visible, showing the user `brcaffer@microsoft.com`. The notebook title is "Low Birthweight Analysis (Python)". The interface includes a left sidebar with navigation options like Home, Workspace, Recent, Data, Clusters, Jobs, and Search. The main content area shows a "Welcome to Apache Spark with Python" message, followed by a paragraph about Apache Spark and a link to `http://spark.apache.org/`. Below this, there are two code cells. The first code cell, labeled "Cmd 3", contains Python code to print the version:

```
1 # Python version
2 import sys
3 print('Python: {}'.format(sys.version))
```

. The output of this cell shows "Python: 3.5.2 (default, Nov 23 2017, 16:37:01) [GCC 5.4.0 20160609]" and a timestamp. The second code cell, labeled "Cmd 4", contains SQL code:

```
1 %sql
2
3 show tables
```

. At the bottom of the notebook, there are input fields for "database" and "tableName", and a checkbox for "isTemporary". The Windows taskbar is visible at the very bottom of the image.

Annotations

Code Cells

Azure Databricks Notebooks - Exploring

View: Code ▾ Permissions Run All Clear ▾ Schedule Comments Revisions

Edit mode

- `<Esc>` : Switch to Command Mode
- `<Ctrl> <Alt> F` : Find and Replace
- `<Shift> + <Enter>` : Run command and move to next cell
- `<Alt> + <Enter>` : Run command and insert new cell below
- `<Ctrl> + <Enter>` : Run command
- `<Shift> + <Alt> + <Up>` : Run all above commands (exclusive)
- `<Shift> + <Alt> + <Down>` : Run all below commands (inclusive)
- `<Alt> + <Up> / <Down>` : Move to previous/next cell
- `<Ctrl> + <Alt> + P` : Insert a cell above
- `<Ctrl> + <Alt> + N` : Insert a cell below
- `<Ctrl> + <Alt> + -` : Split a cell at cursor
- `<Ctrl> + <Alt> + <Up>` : Move a cell up
- `<Ctrl> + <Alt> + <Down>` : Move a cell down
- `<Ctrl> + <Alt> + M` : Toggle comments panel
- `<Ctrl> + <Alt> + C` : Copy current cell
- `<Ctrl> + <Alt> + X` : Cut current cell
- `<Ctrl> + <Alt> + V` : Paste cell below
- `<Ctrl> + <Alt> + D` : Delete current cell
- `<Up>` : Move up or to previous cell
- `<Down>` : Move down or to next cell
- `<Tab>` : Autocomplete, indent selection
- `<Shift> + <Tab>` : Unindent selection

Command mode

- `<Enter>` : Switch to Edit Mode
- `<Ctrl> <Alt> F` : Find and Replace
- `<Shift> + <Enter>` : Run command and move to next cell
- `<Ctrl> + <Enter>` : Run command
- `<Shift> + <Alt> + <Up>` : Run all above commands (exclusive)
- `<Shift> + <Alt> + <Down>` : Run all below commands (inclusive)
- `D D` : Delete current cell
- `<Shift> + D D` : Delete current cell(skip prompt)
- `X` : Cut current cell
- `C` : Copy current cell
- `V` : Paste cell below
- `<Shift> + V` : Paste cell above
- `A` : Insert a cell above
- `B` : Insert a cell below
- `O` : Toggle cell output
- `<Space>` : Scroll down
- `<Shift> + <Space>` : Scroll up
- `H` : Toggle keyboard shortcuts menu
- `<Shift> + M` : Merge with cell below
- `<Up> / P / K` : Move to previous cell
- `<Down> / N / J` : Move to next cell
- `<Ctrl> + <Click>` : Select multiple cells

Show
Keys

Azure Databricks Notebooks – Cell Actions

Insert Cell - Displays on Mouse Over

OK

Command took 0.13 seconds -- by brcaffer@microsoft.com at 6/24/2018 3:07:32 PM on clusterdemo10

Cmd 18

+

Insert a new cell

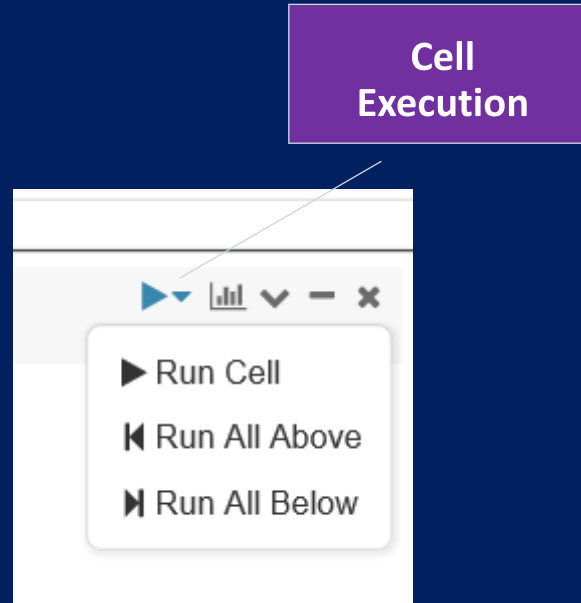
Notice: Underscores are automatically replaced on viewing with spaces...

1 %sql **select** * **from** temp_state_pop

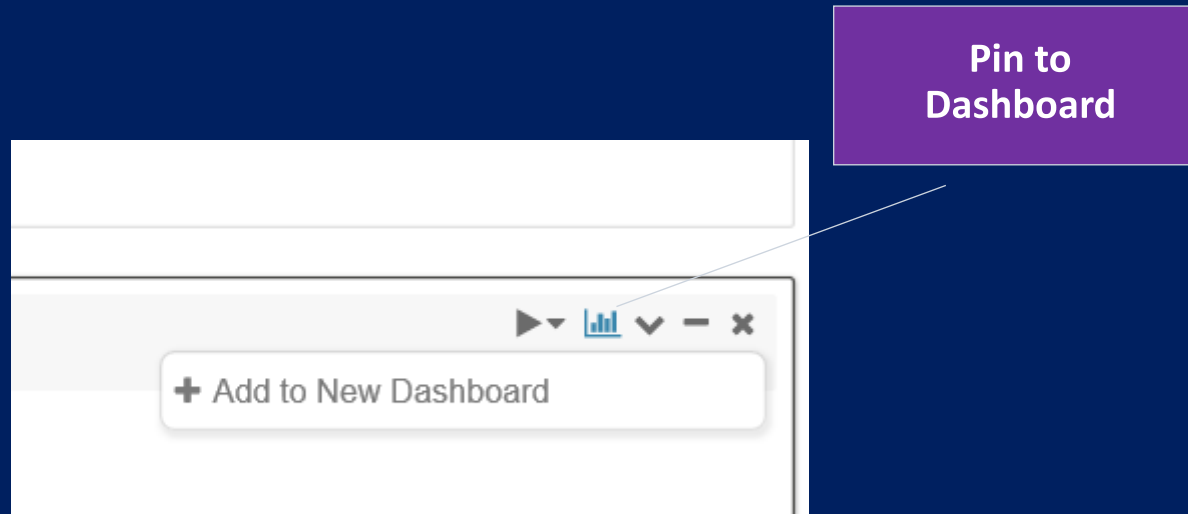
▶ (1) Spark Jobs

| StateRank | City | State | StateCode | Population |
|-----------|------------|---------|------------|--------------------------|
| 2014 rank | Birmingham | Alabama | State Code | 2014 Population estimate |

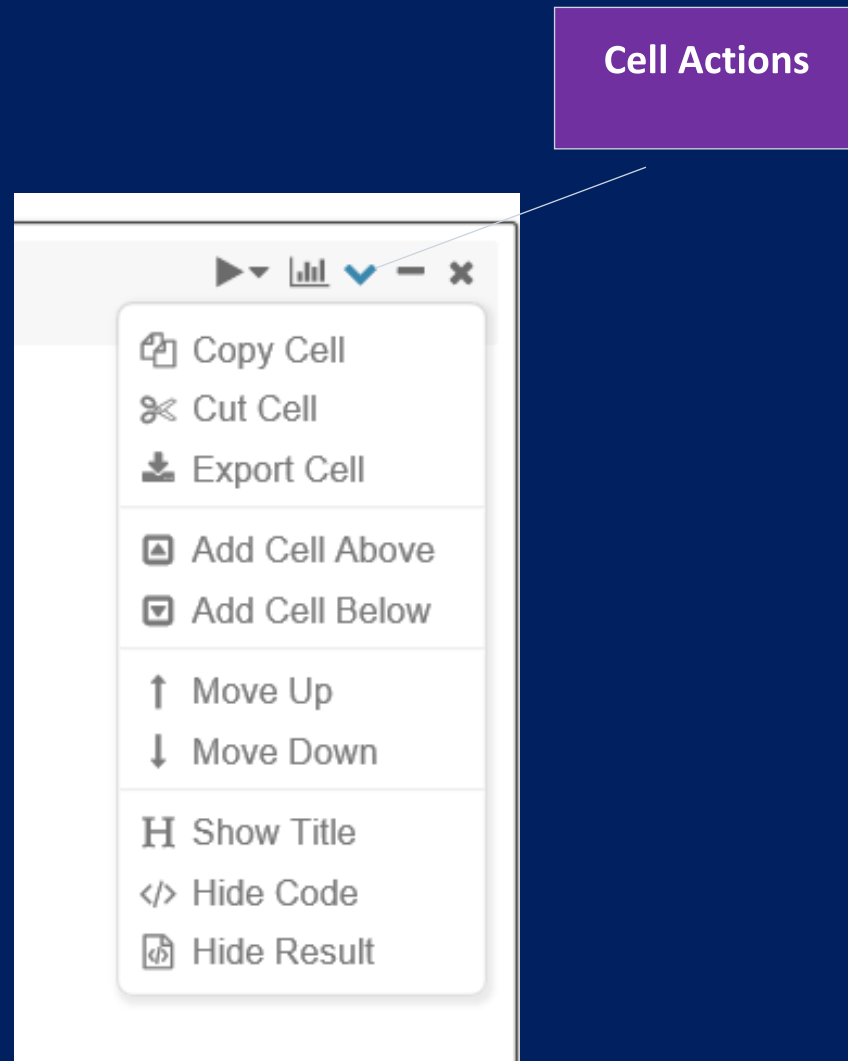
Azure Databricks Notebooks – Cell Actions



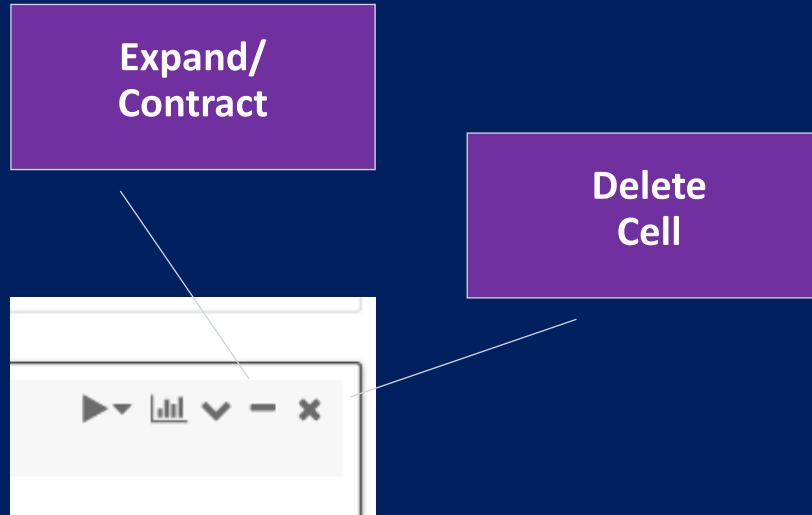
Azure Databricks Notebooks – Cell Actions



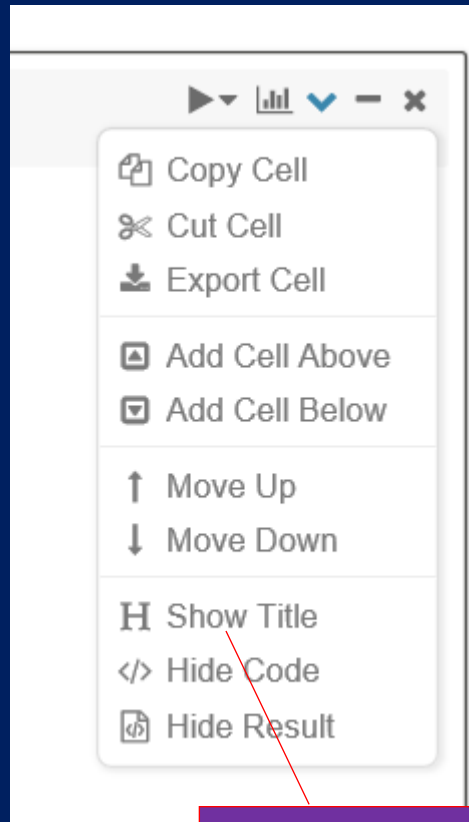
Azure Databricks Notebooks – Cell Actions Menu



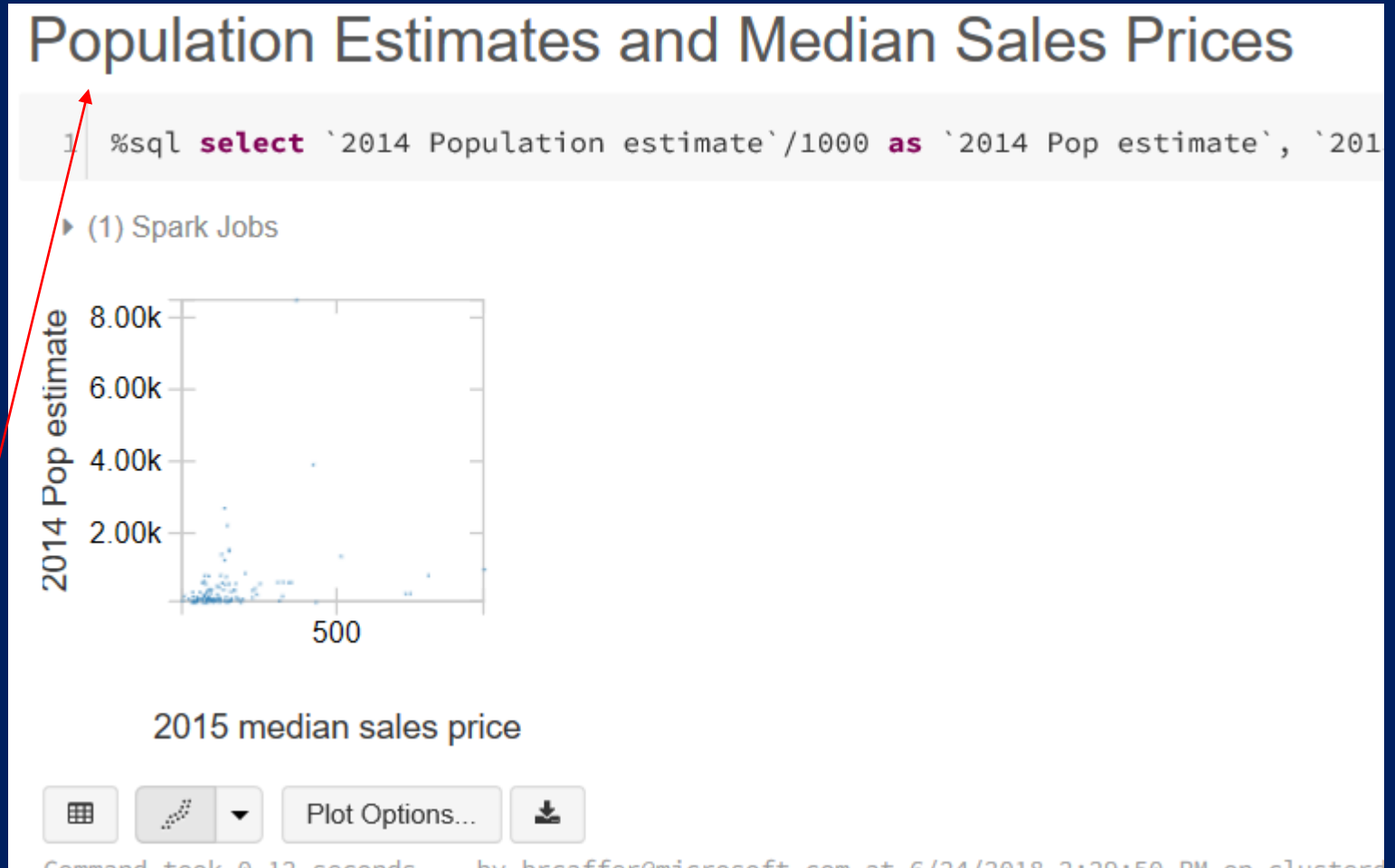
Azure Databricks Notebooks – Cell Actions



Azure Databricks Notebooks – Cell Actions



Show/Hide
Title



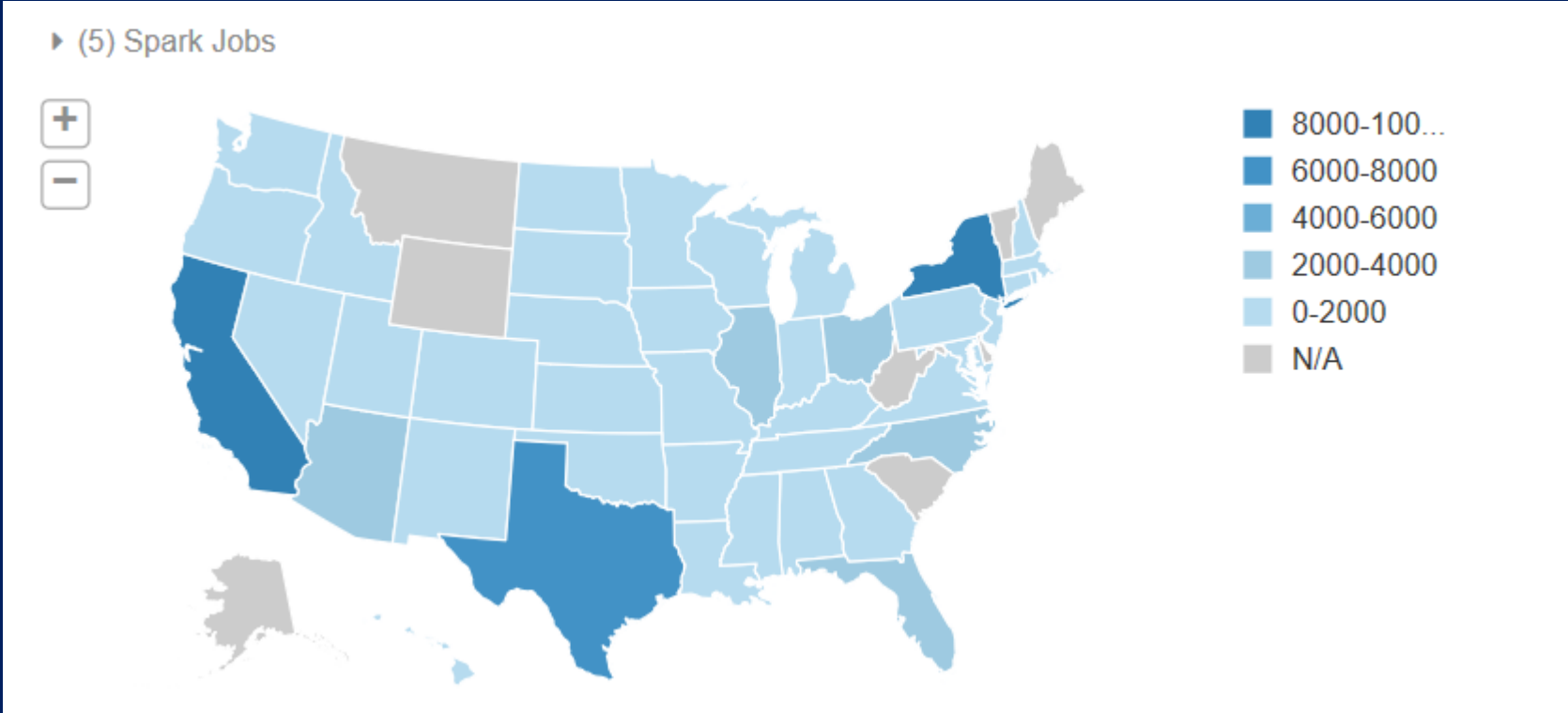
Azure Databricks Notebooks

You Did It!

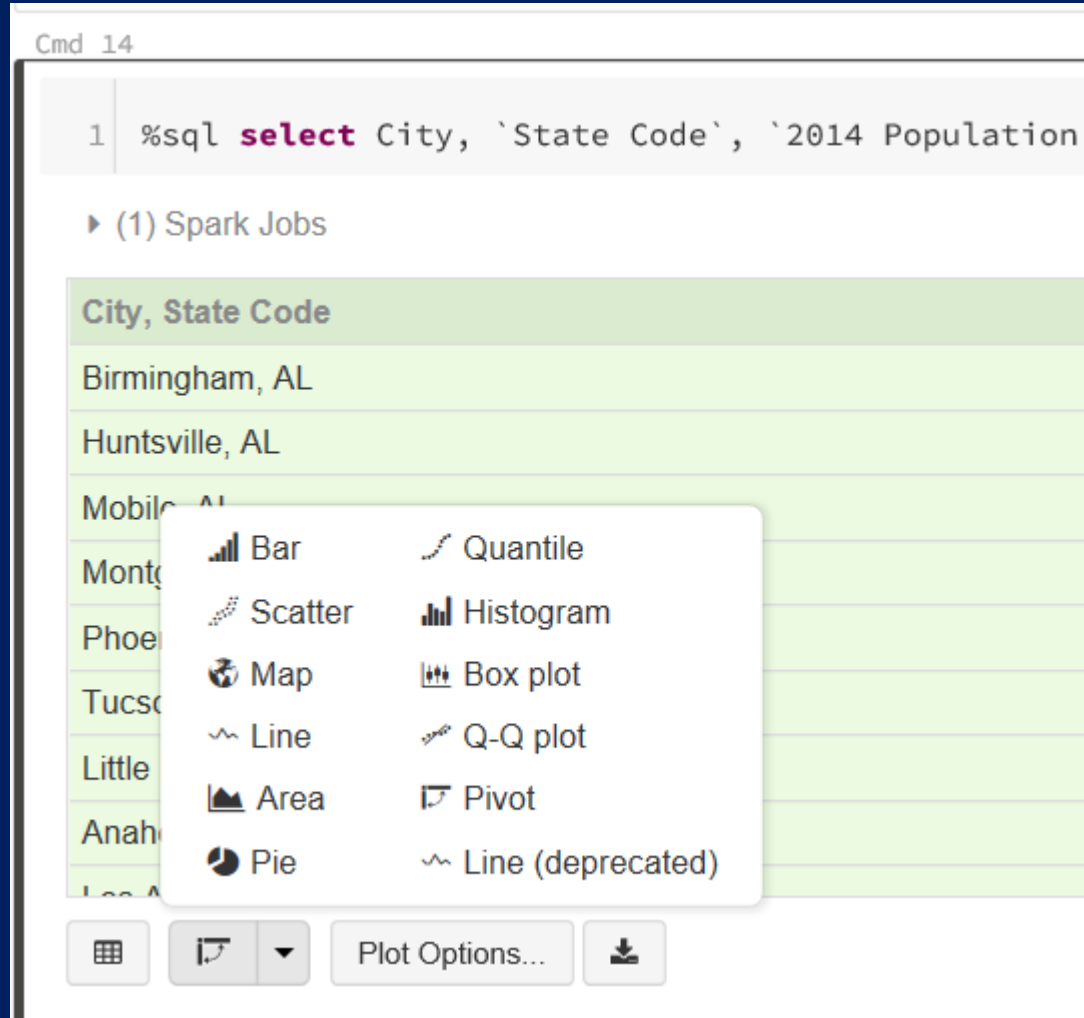
The screenshot displays the Microsoft Azure Databricks web interface. On the left is a dark sidebar with navigation icons for Home, Workspace, Recent, Data, Clusters, Jobs, and Search. The main content area shows a notebook titled "Pop. vs. Price LR 2.0 (1) (Python)". Below the title bar, which includes "Permissions", "Run All", and "Clear" buttons, is a large heading "Population vs. Median Home Prices" with the subtitle "Linear Regression with Single Variable". The notebook content is organized into command blocks:

- Cmd 2:** Contains the text "Note, this notebook requires Spark 2.0+".
- Cmd 3:** Contains a Scala code snippet: `%scala if (org.apache.spark.BuildInfo.sparkBranch < "2.0") sys.error("Attach this notebook")`. Below the code, it states "Command took 5.45 seconds -- by a user at 10/25/2016 5:32:42 AM on unknown cluster".
- Cmd 4:** Contains the heading "Load and parse the data".
- Cmd 5:** Contains the start of a comment: `# Use the Spark CSV datasource with options specifying:`.

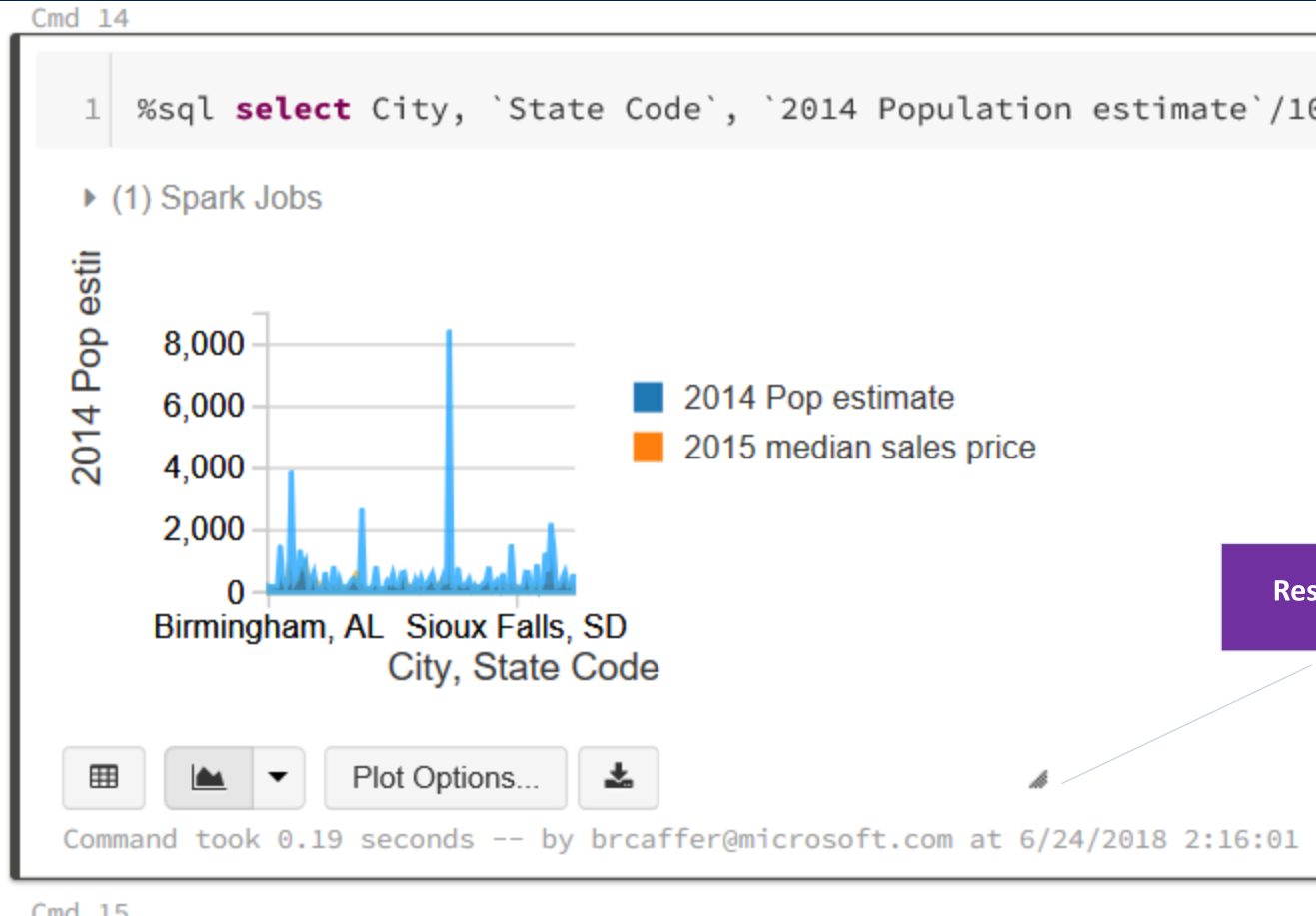
Azure Databricks Notebooks Output View



Azure Databricks Notebooks Output View

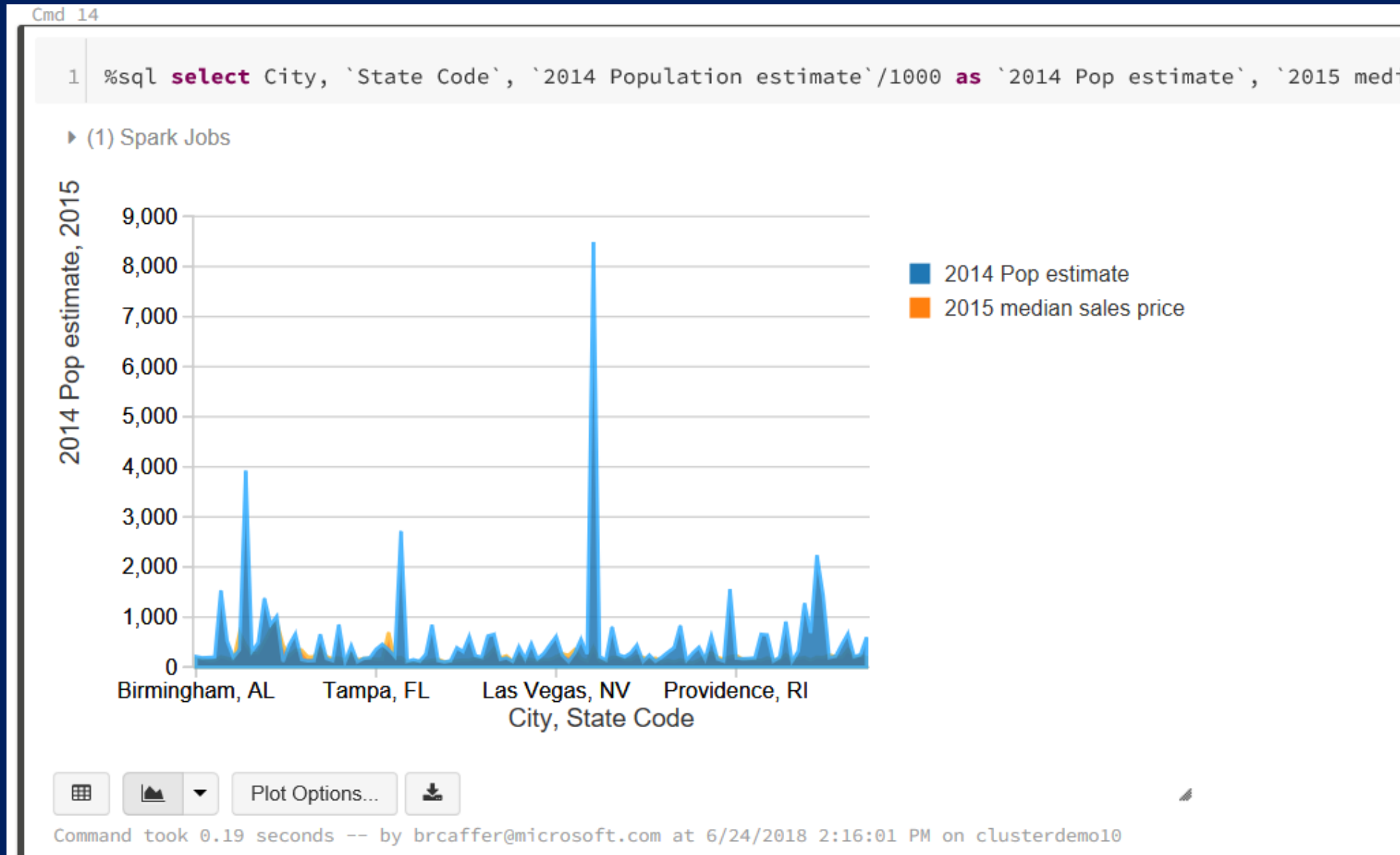


Azure Databricks Notebooks Output View

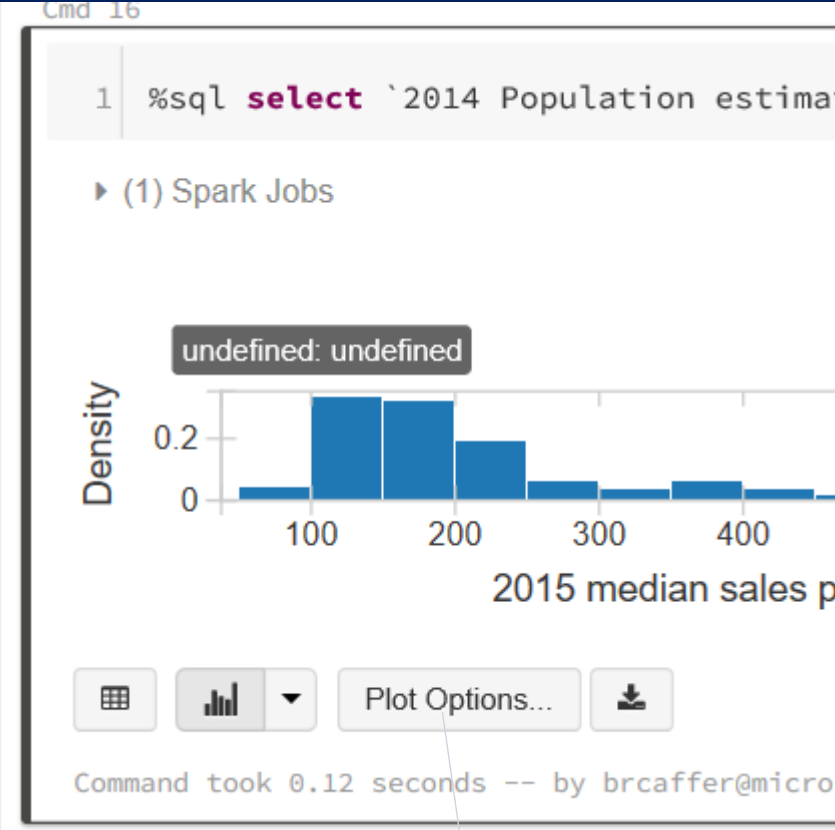


Resize Grabber

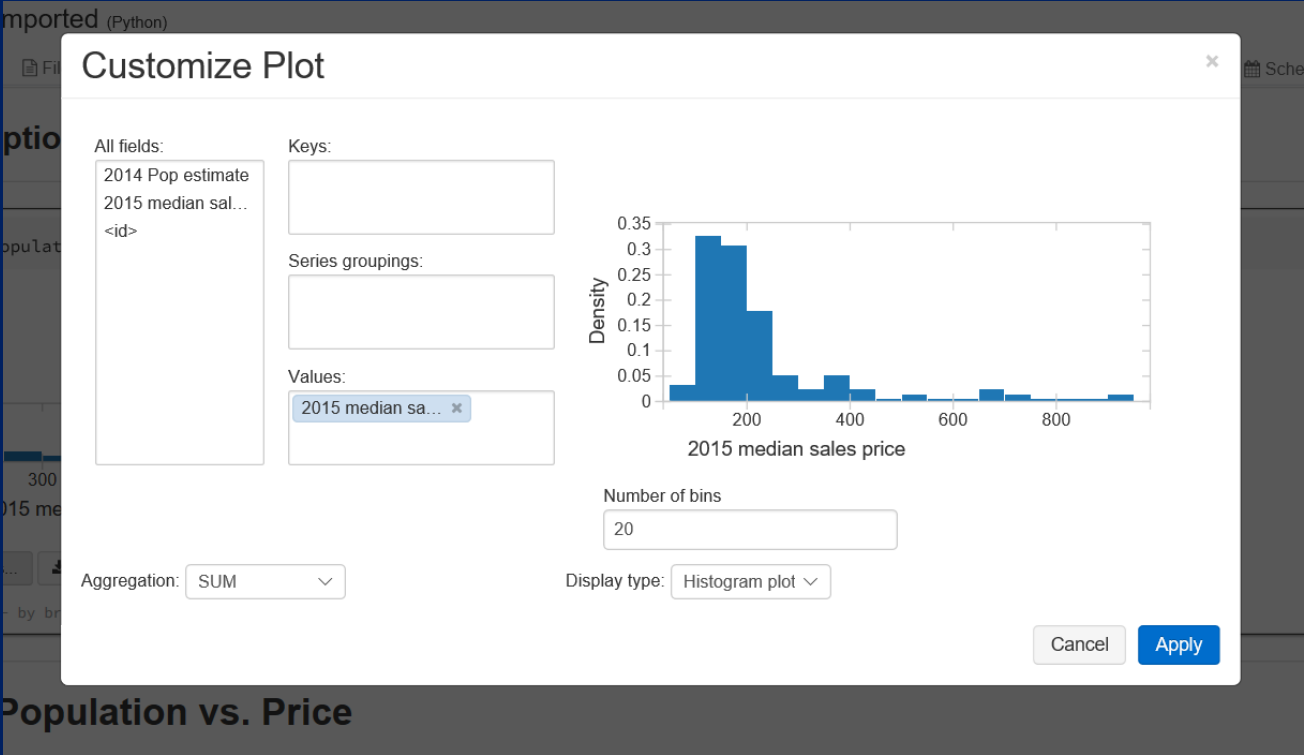
Azure Databricks Notebooks Output View



Azure Databricks Notebooks Output View



Visualization
Panel



Azure Databricks Notebooks Menu Bar

Attach to Cluster

Notebook File Managment

Notebook Cell Filters

Permissions

Schedule Notebook as a Job

Toggle Comments

Pop. vs. Price LR 2.0 Imported (Python)

Attached: clusterdemo10 File View: Code Permissions Run All Clear

Schedule Comments Revision history

Population Estimates and Median Sales Prices

1 %sql select `2014 Population estimate`/1000 as `2014 Pop estimate`, `2015 median sales price` from data_geo

(1) Spark Jobs

4 Pop estimate

Toggle Revision History

Account

Azure Databricks Notebooks Output View

Tabular

Visualize

Download Data

(1) Spark Jobs

| City | State Code |
|-------------|------------|
| Birmingham | AL |
| Huntsville | AL |
| Mobile | AL |
| Montgomery | AL |
| Phoenix | AZ |
| Tucson | AZ |
| Little Rock | AR |
| Anaheim | CA |
| Los Angeles | CA |

Command took 0.19 seconds == by brcaffer@microsoft.com at