

Azure Databricks

What is Azure Databricks?

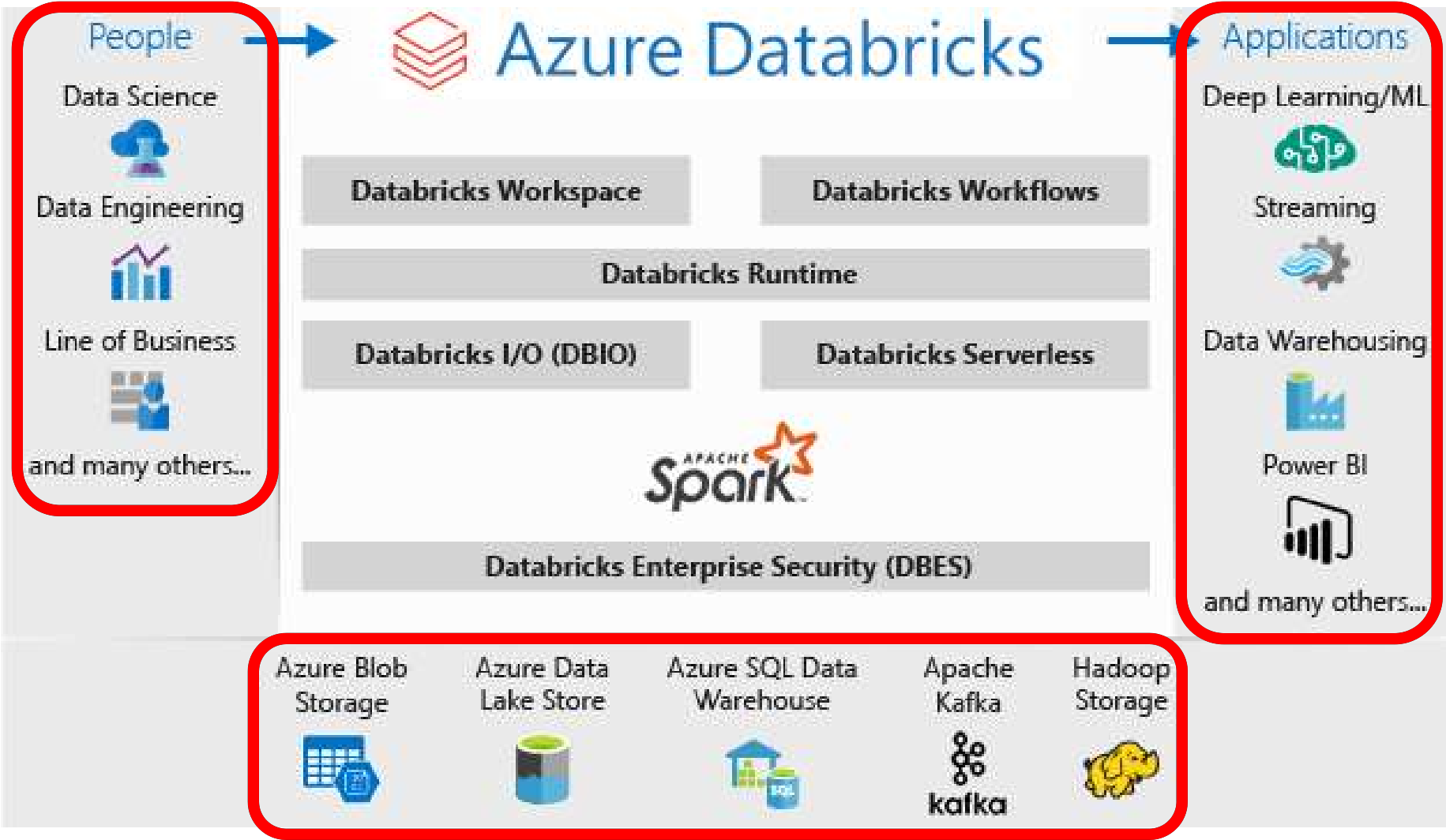
Apache Spark-based

Analytics platform

Provides

- One-click setup
- Streamlined workflows and
- An interactive workspace
- Enables collaboration between data scientists, data engineers, and business analysts.

Azure Databrick



Azure Databrick

For a big data pipeline, the data is ingested into Azure

This data lands in

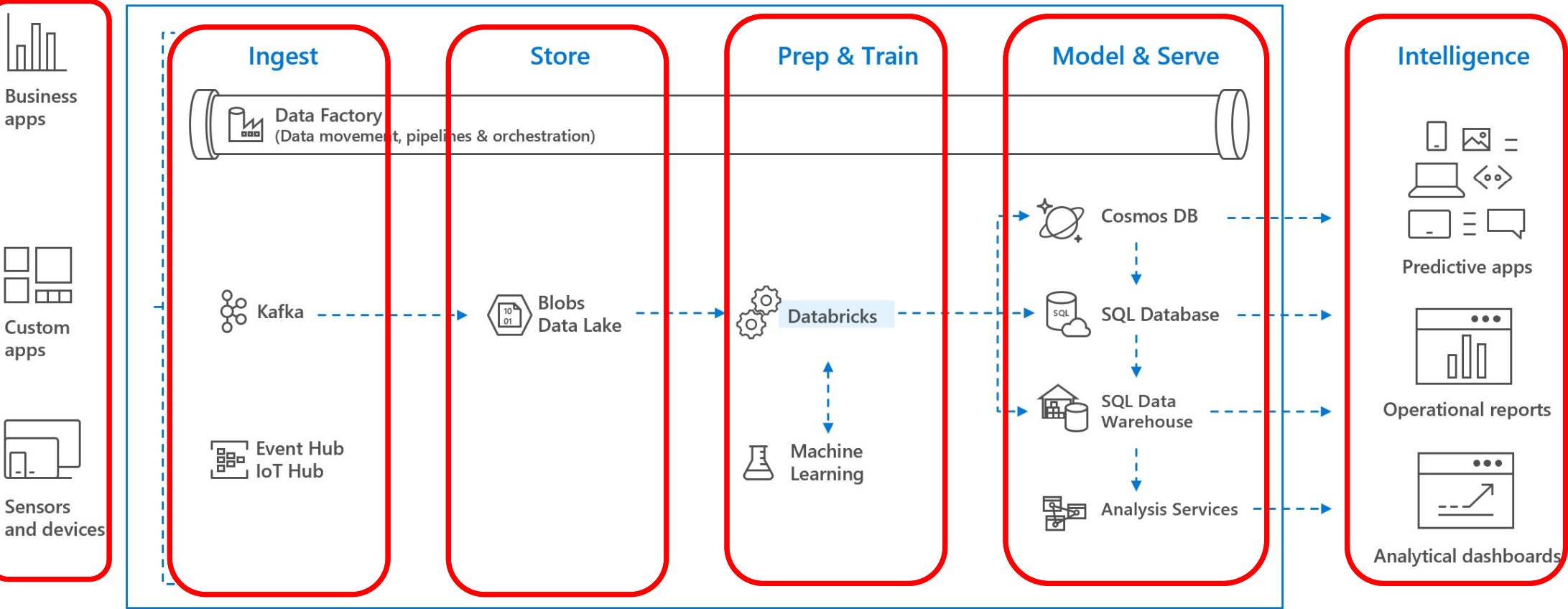
- Azure Blob Storage or
- Azure Data Lake Storage

Use Azure Databricks to read data from multiple data sources

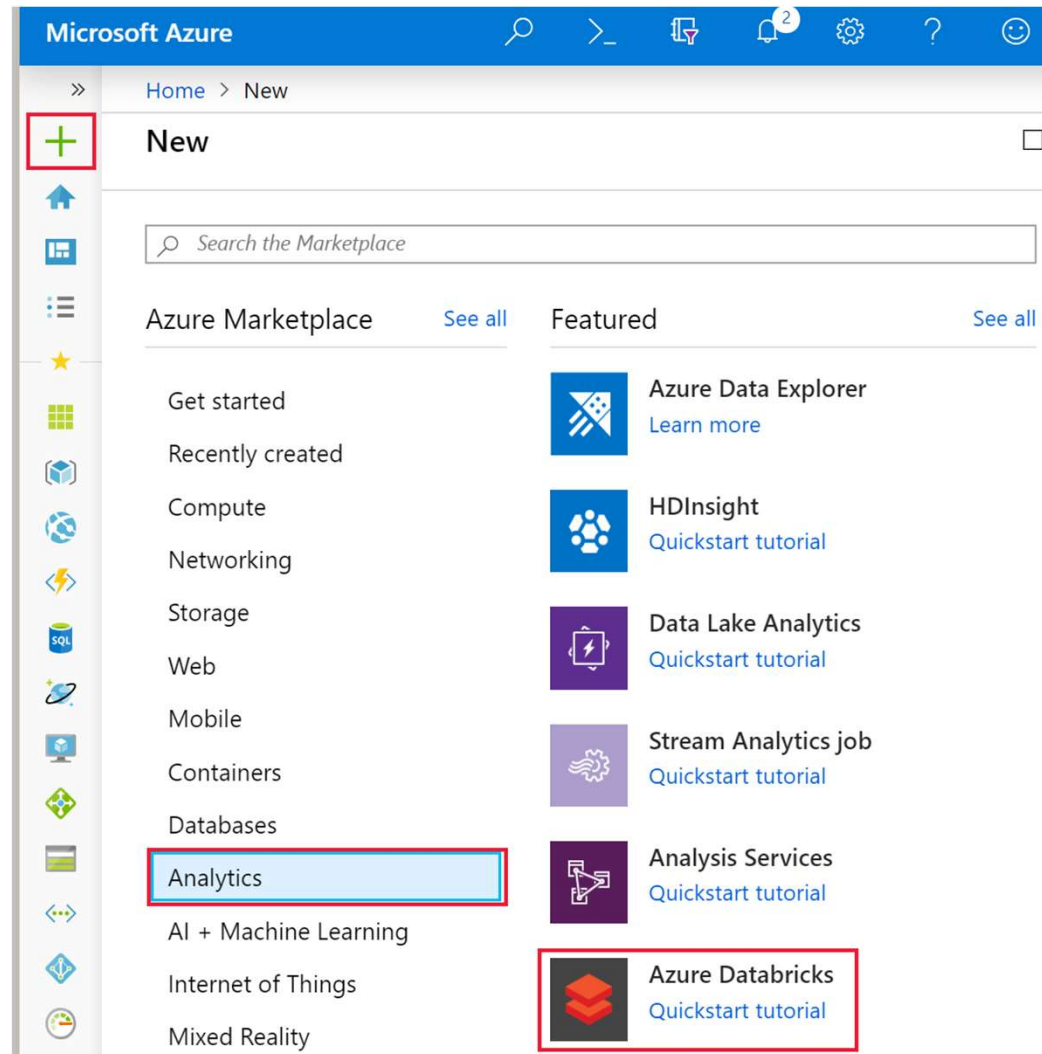
- Azure Blob Storage
- Azure Data Lake Storage
- Azure Cosmos DB, or
- Azure SQL Data Warehouse

Using Databricks, turn it into breakthrough insights

Azure Databrick



Hands-On: Create Databricks Workspace



Hands-On: Create Databricks Workspace

[Basics *](#)[Networking](#)[Tags](#)[Review + Create](#)

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ

<your subscription> ▼

Resource group * ⓘ

(New) databricks-quickstart ▼

[Create new](#)

Instance Details

Workspace name *

mydatabricksws ✓

Location *

West US 2 ▼

Pricing Tier * ⓘ

Standard (Apache Spark, Secure with Azure AD) ^

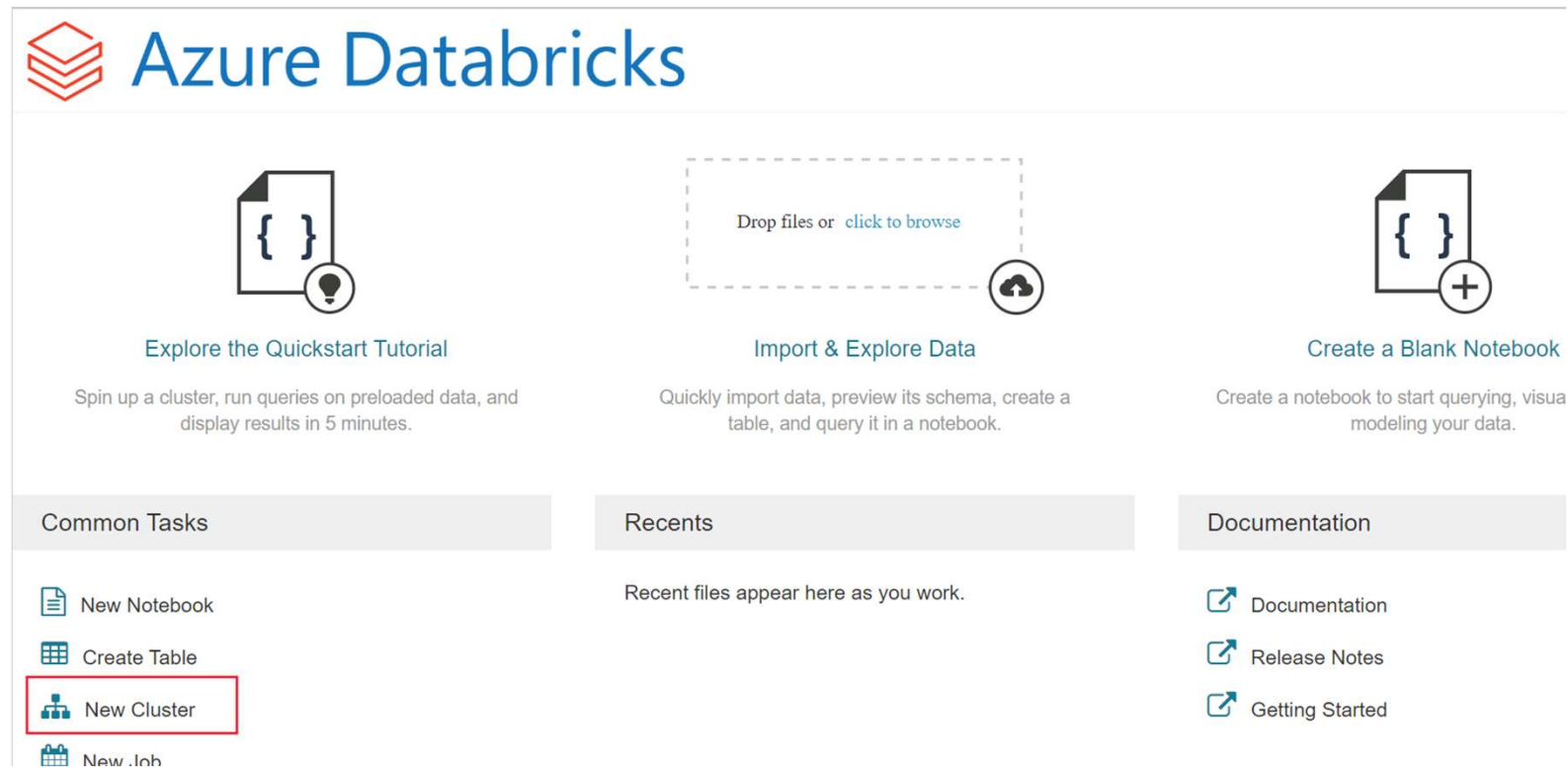
Standard (Apache Spark, Secure with Azure AD)

Premium (+ Role-based access controls)

Trial (Premium - 14-Days Free DBUs)

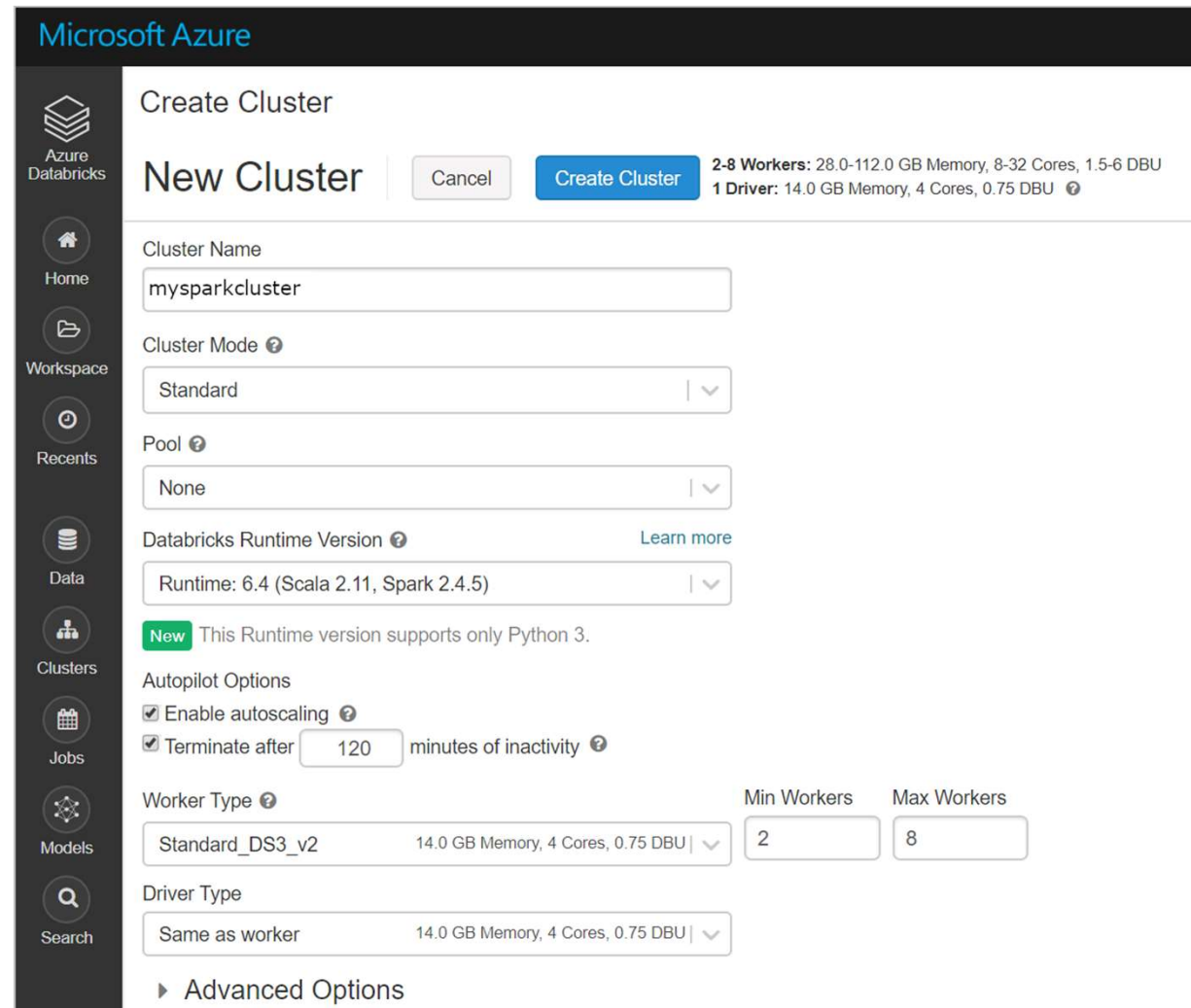
Hands-On: Create a Spark cluster in Databricks

- Go to the Databricks workspace that you created, and then click Launch Workspace.
- You are redirected to the Azure Databricks portal.
- Click New Cluster



Hands-On: Create a Spark cluster in Databricks

- Make sure you select the Terminate after ___ minutes of inactivity checkbox
- Provide a duration (in minutes) to terminate the cluster, if the cluster is not being used.



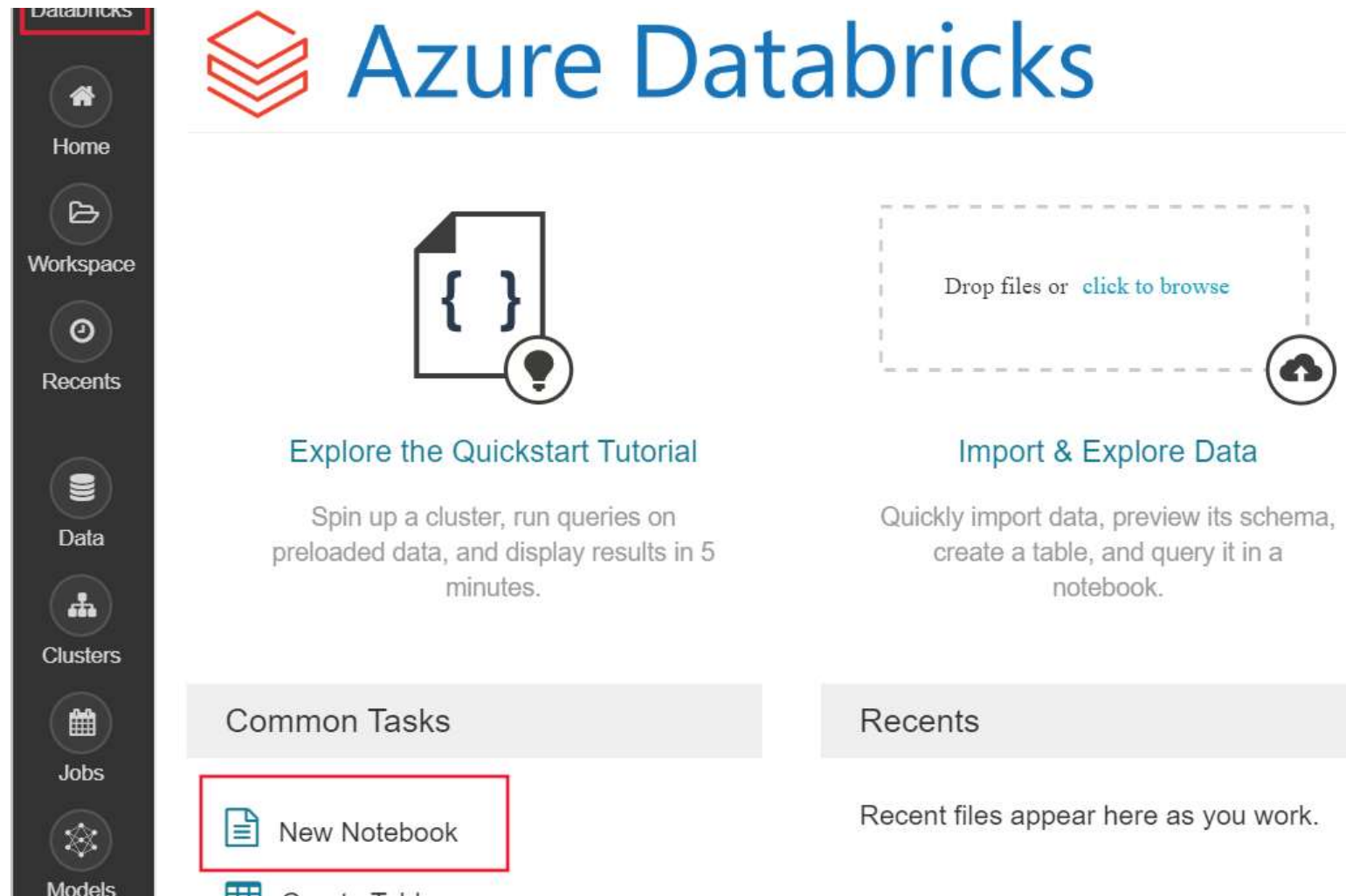
The screenshot shows the 'Create Cluster' page in the Microsoft Azure Databricks portal. The left sidebar contains navigation icons for Home, Workspace, Recents, Data, Clusters, Jobs, and Models, along with a Search icon. The main content area is titled 'Create Cluster' and 'New Cluster'. It includes a 'Cancel' button and a 'Create Cluster' button. The cluster configuration is as follows:

- Cluster Name:** mysparkcluster
- Cluster Mode:** Standard
- Pool:** None
- Databricks Runtime Version:** Runtime: 6.4 (Scala 2.11, Spark 2.4.5). A 'Learn more' link is available.
- Autopilot Options:**
 - ☒ Enable autoscaling
 - ☒ Terminate after 120 minutes of inactivity
- Worker Type:** Standard_DS3_v2 (14.0 GB Memory, 4 Cores, 0.75 DBU). Min Workers: 2, Max Workers: 8.
- Driver Type:** Same as worker (14.0 GB Memory, 4 Cores, 0.75 DBU).

At the bottom, there is a link to 'Advanced Options'.

Run a Spark SQL job

Hands-On: Run a Spark SQL job



The screenshot displays the Azure Databricks web interface. On the left is a dark sidebar with navigation icons and labels: Home, Workspace, Recents, Data, Clusters, Jobs, and Models. The 'Databricks' label at the top of the sidebar is highlighted with a red box. The main workspace area features the 'Azure Databricks' logo at the top. Below the logo, there's a section for file upload with a dashed box containing the text 'Drop files or [click to browse](#)'. To the right of this is a circular icon with an upward arrow. Below the upload section, there are two main cards: 'Explore the Quickstart Tutorial' and 'Import & Explore Data'. The 'Explore the Quickstart Tutorial' card includes a document icon with curly braces and a lightbulb, and text describing a 5-minute tutorial. The 'Import & Explore Data' card includes text about importing data and querying it. At the bottom, there are two sections: 'Common Tasks' and 'Recents'. The 'Common Tasks' section has a red box around the 'New Notebook' button, which is represented by a document icon. The 'Recents' section has text indicating that recent files appear here as you work.

Databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Models

Azure Databricks

Drop files or [click to browse](#)

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.

Common Tasks

New Notebook

Recents

Recent files appear here as you work.

Hands-On: Run a Spark SQL job

- The following command sets the Azure storage access information.
 - `blob_account_name = "azureopendatastorage"`
 - `blob_container_name = "citydatacontainer"`
 - `blob_relative_path = "Safety/Release/city=Boston"`
 - `blob_sas_token = r"?st=2019-02-26T02%3A34%3A32Z&se=2119-02-27T02%3A34%3A00Z&sp=rl&sv=2018-03-28&sr=c&sig=XIJVWA7fMXCSxCKqJm8psM0h0W4h7cSYO28coRqF2fs%3D"`

Hands-On: Run a Spark SQL job

- The following command allows Spark to read from Blob storage remotely
 - `wasbs_path = 'wasbs://%s@%s.blob.core.windows.net/%s' % (blob_container_name, blob_account_name, blob_relative_path)`
 - `spark.conf.set('fs.azure.sas.%s.%s.blob.core.windows.net' % (blob_container_name, blob_account_name), blob_sas_token)`
 - `print('Remote blob path: ' + wasbs_path)`

Hands-On: Run a Spark SQL job

- The following command creates a DataFrame
 - `df = spark.read.parquet(wasbs_path)`
 - `print('Register the DataFrame as a SQL temporary view: source')`
 - `df.createOrReplaceTempView('source')`





Hands-On: Run a Spark SQL job

- Run a SQL statement return the top 10 rows of data
 - `print('Displaying top 10 rows: ')`
 - `display(spark.sql('SELECT * FROM source LIMIT 10'))`

```
1 print('Displaying top 10 rows: ')
2 display(spark.sql('SELECT * FROM source LIMIT 10'))
```

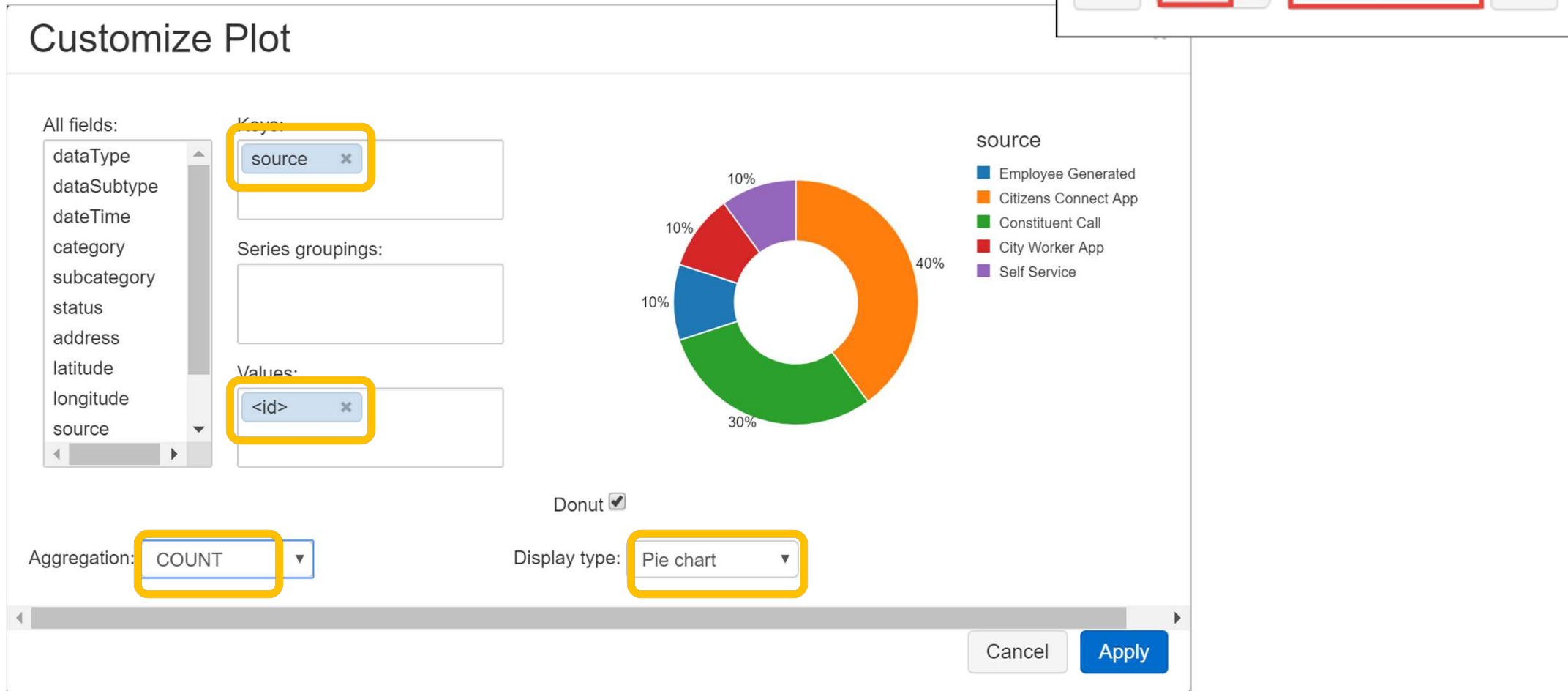
▶ (1) Spark Jobs

| data Type ▼ | data Subtype ▼ | date Time ▼ | category ▼ | sub category ▼ | status ▼ | address ▼ | latitude ▼ | longitude ▼ | source ▼ | extended Properties ▼ |
|-------------|----------------|------------------------------|----------------------------------|----------------------------|----------|-------------------------------------|------------|-------------|----------------------|-----------------------|
| Safety | 311_All | 2011-08-11T11:02:16.000+0000 | Recycling | Request for Recycling Cart | Closed | 43 Howell St Dorchester MA 02125 | 42.3255 | -71.0587 | Employee Generated | null |
| Safety | 311_All | 2016-12-15T09:08:21.000+0000 | Street Cleaning | Pick up Dead Animal | Closed | 74 Aldie St Allston MA 02134 | 42.3588 | -71.1335 | Citizens Connect App | null |
| Safety | 311_All | 2017-01-26T18:45:00.000+0000 | Enforcement & Abandoned Vehicles | Parking Enforcement | Closed | 98 Waltham St Roxbury MA 02118 | 42.3436 | -71.0713 | Constituent Call | null |

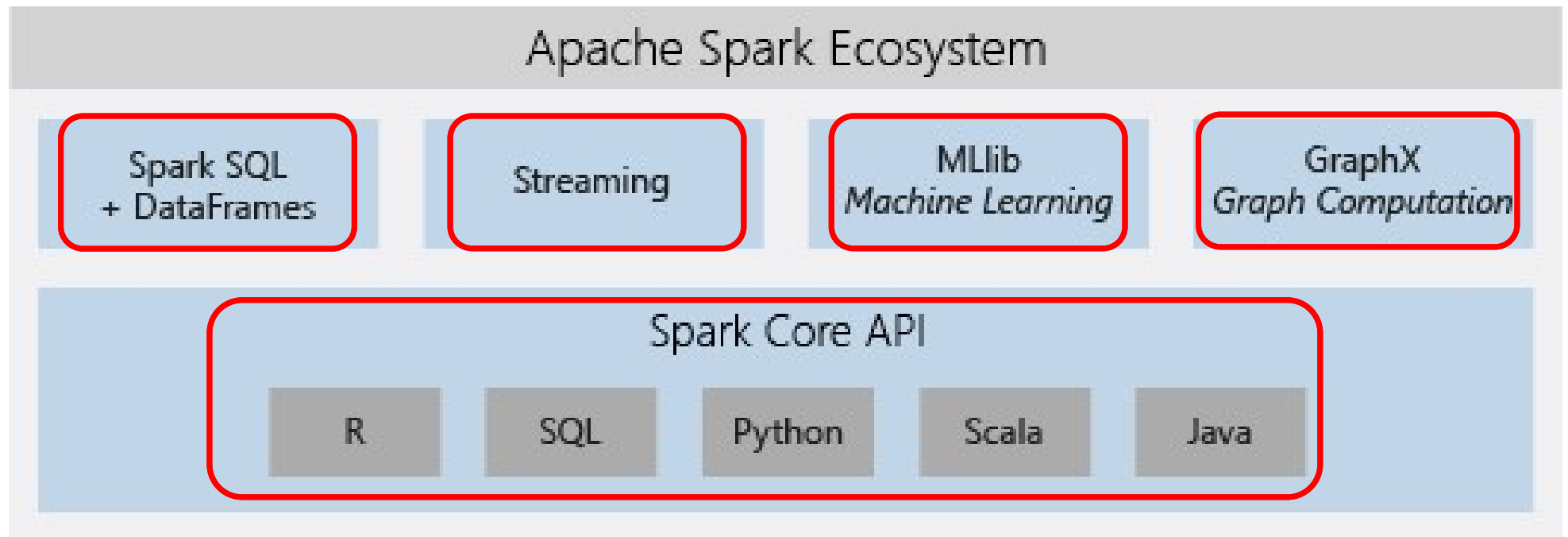


Hands-On: Run a Spark SQL job

- You now create a visual representation of this data



Apache Spark-based analytics platform



Thanks