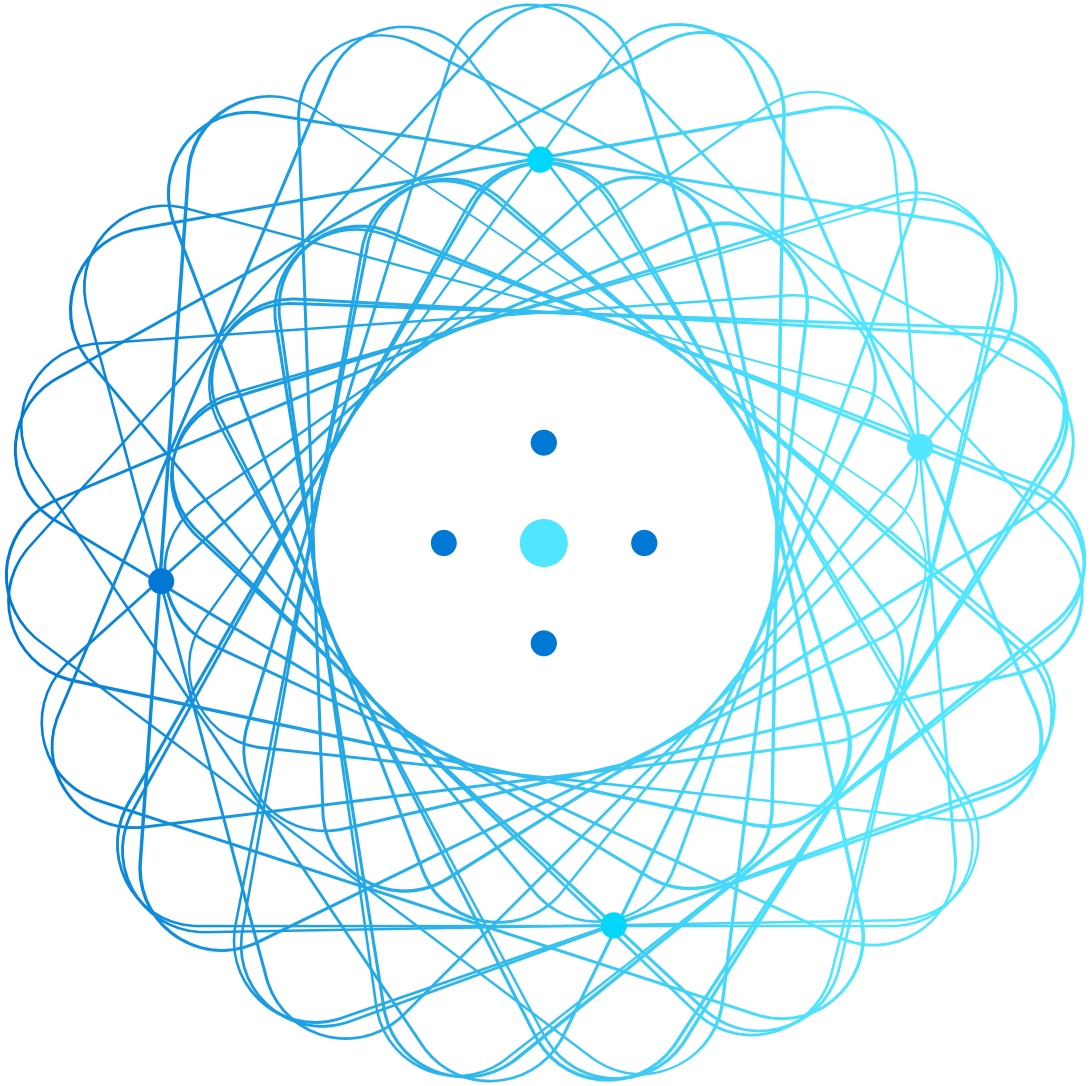
Microsoft Azure

Data engineering with Azure Databricks



Agenda

Explore Azure Databricks

Use Apache Spark in Azure Databricks

Run Azure Databricks notebooks in Azure Data Factory

Explore Azure Databricks

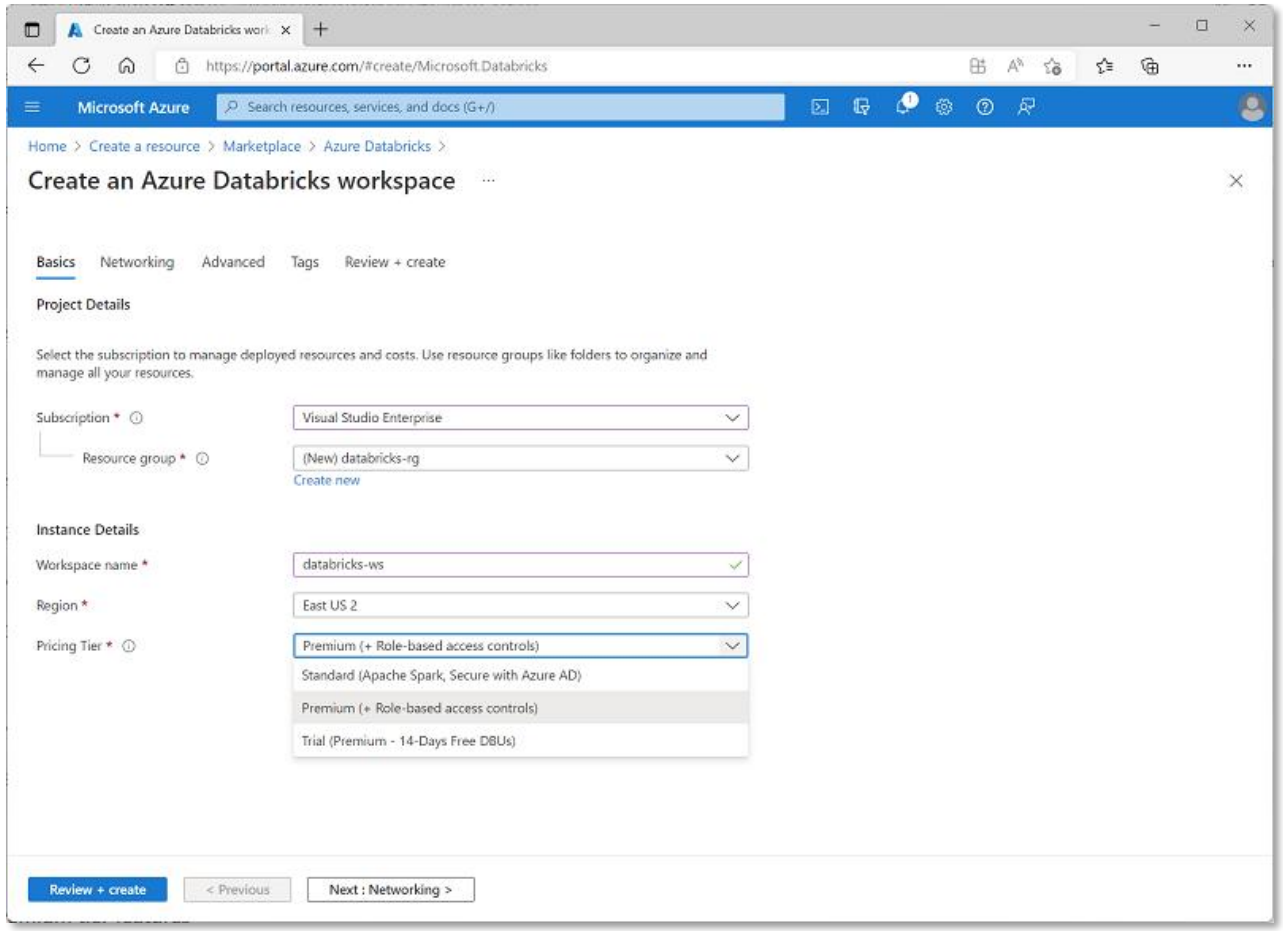
What is Azure Databricks?

Fully managed, cloud-based data analytics platform

- Built on Apache Spark
- Web-based portal

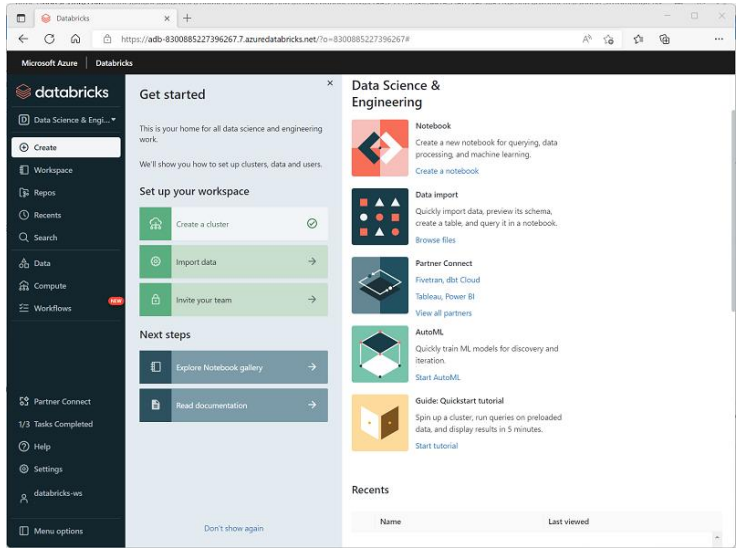
Provisioned as an Azure resource

- Standard tier
- Premium tier
- Trial



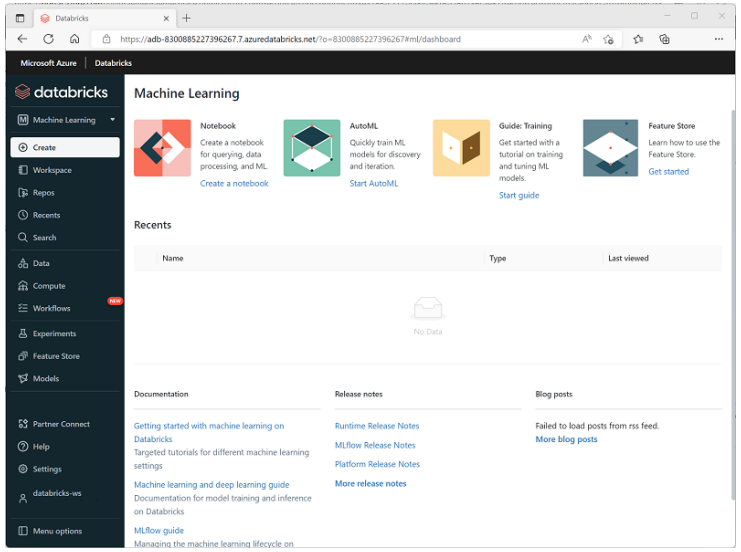
Azure Databricks workloads

Data Science and Engineering



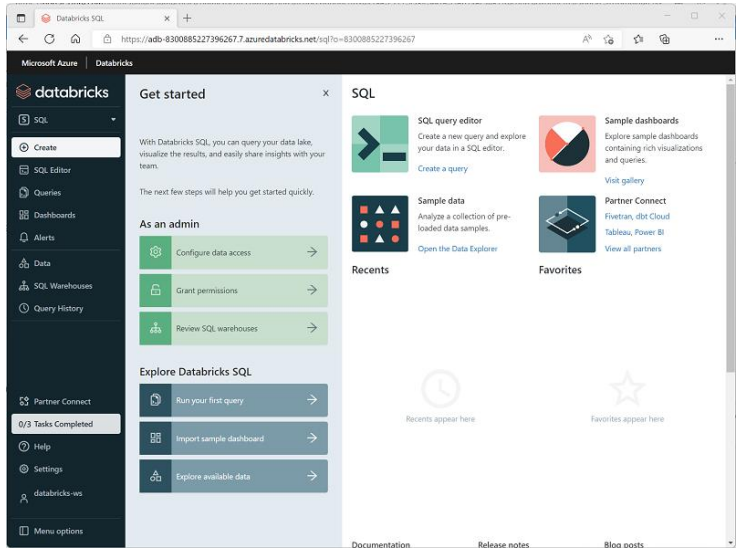
Use notebooks to run Apache Spark code to manipulate and explore data

Machine Learning



Train predictive models using SparkML and other machine learning frameworks

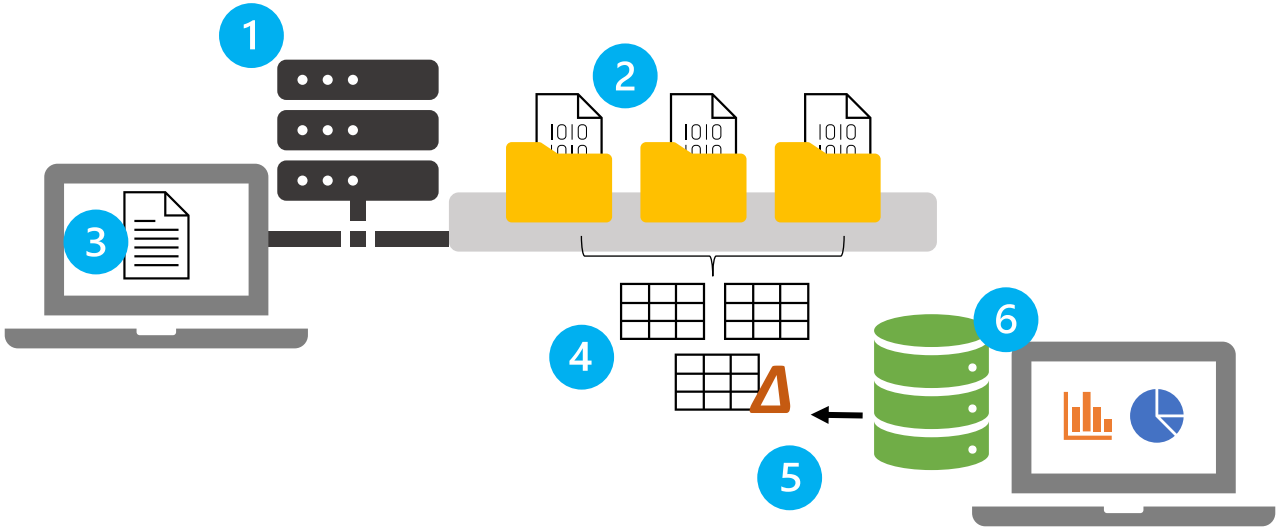
SQL



Store and query data in relational tables using SQL

Only available in *Premium* tier workspaces

Key concepts



1. **Apache Spark clusters** provide highly scalable parallel compute for distributed data processing
2. **Databricks File System (DBFS)** provides distributed shared storage for data lakes
3. **Notebooks** provide an interactive environment for combining code, notes, and images
4. **Metastore** provides a relational abstraction layer, enabling you to define tables based on data in files
5. **Delta Lake** builds on the metastore to enable common relational database capabilities
6. **SQL Warehouses** provide relational compute endpoints for querying data in tables

Demo: Explore Azure Databricks

You can try this for yourself later by following the instructions at the link below:

https://aka.ms/mslearn-adb

Knowledge check

You plan to create an Azure Databricks workspace and use the SQL persona view in the Azure Databricks portal. Which of the following pricing tiers can you select?

- ❑ Enterprise
- ❑ Standard
- ☑ Premium

You need to use Spark to process data in files, preparing it for analysis. Which persona view should you use in the Azure Databricks portal?

- ☑ Data Science and Engineering
- ❑ Machine Learning
- ❑ SQL

You've created an Azure Databricks workspace in which you plan to use code in notebooks to process data. What must you create in the workspace?

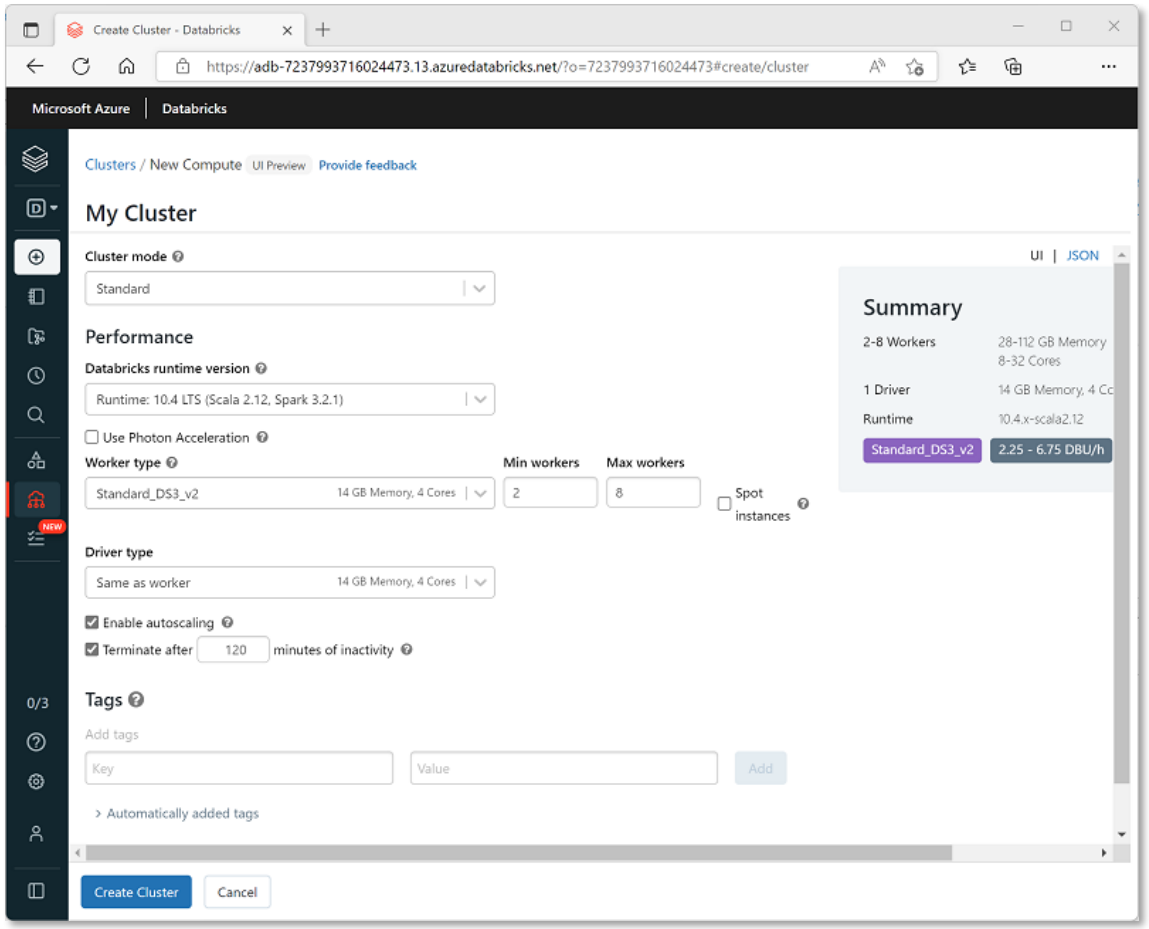
- ❑ A SQL Warehouse
- ☑ A Spark cluster
- ❑ A Windows Server virtual machine

Use Apache Spark in Azure Databricks

Create a Spark cluster

Create a cluster in the Azure Databricks portal, specifying:

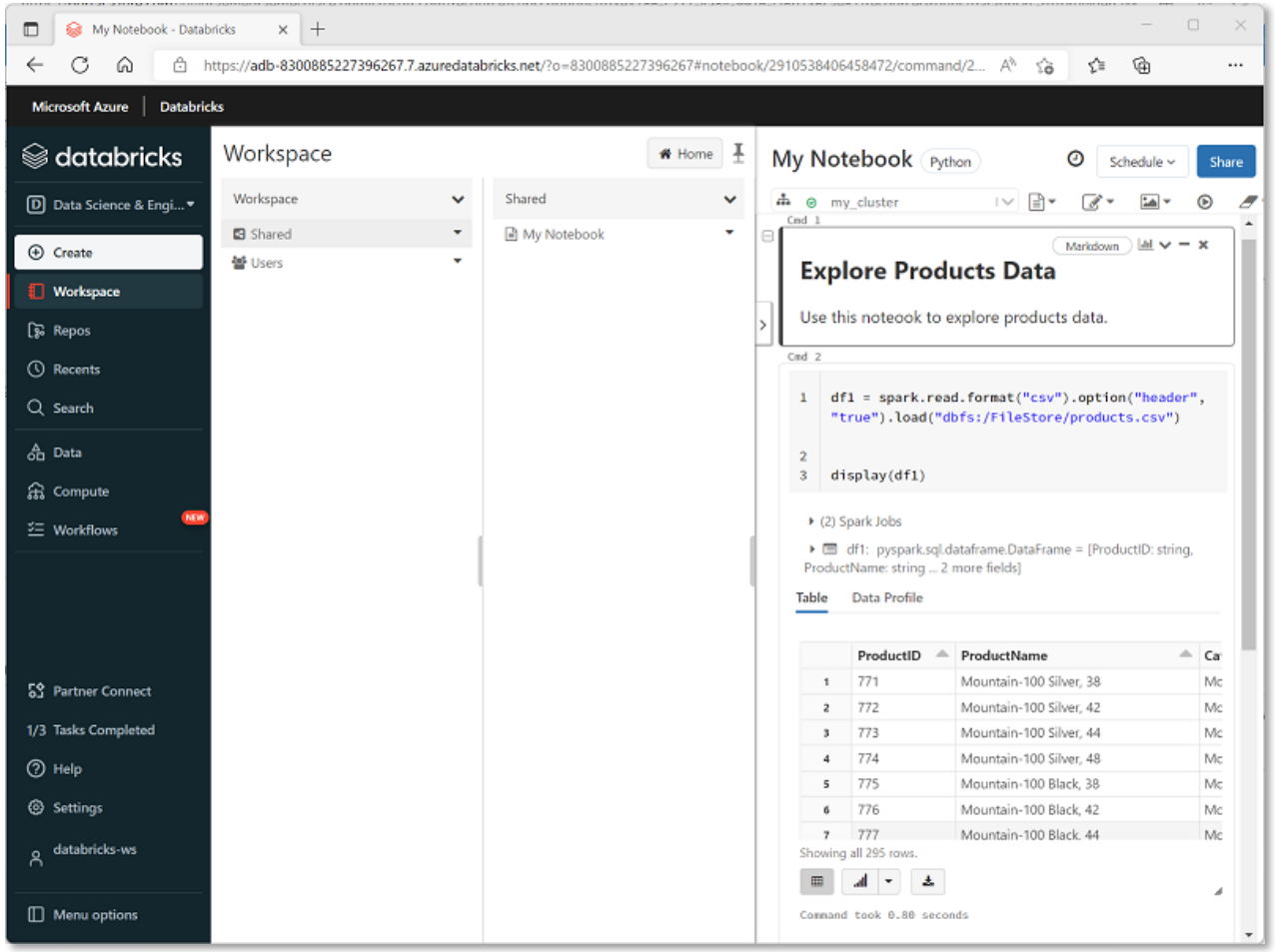
- Cluster name
- Cluster mode (standard, high-concurrency, or single-node)
- Databricks Runtime version
- Worker and driver node VM configuration
- Autoscaling and automatic shutdown



Use Spark in notebooks

Interactive notebooks

- Syntax highlighting and error support
- Code auto-completion
- Interactive data visualizations
- The ability to export results



Use Spark to work with data files

Dataframe API

```
%pyspark
df=spark.read.load('/data/products.csv',
    format='csv',
    header=True
)
display(df.limit(10))
```

Spark SQL API

```
%pyspark
df.createOrReplaceTempView("products")
```

```
%sql
SELECT Category, COUNT(ProductID) AS ProductCount
FROM products
GROUP BY Category
ORDER BY Category
```

Visualize data

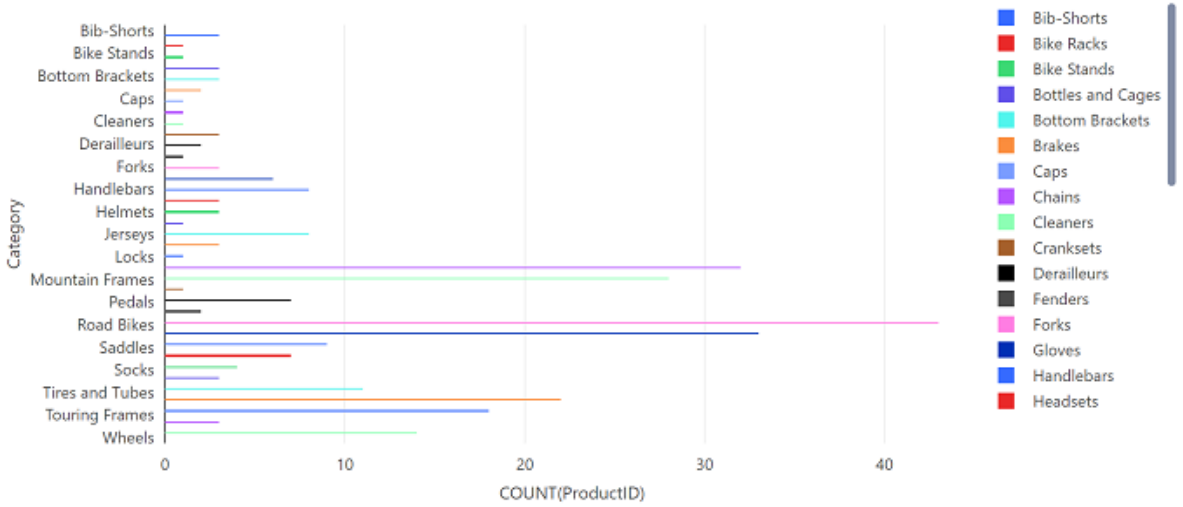
Built-in charts

```
%sql

SELECT Category, ProductID
FROM products;
```

(1) Spark Jobs

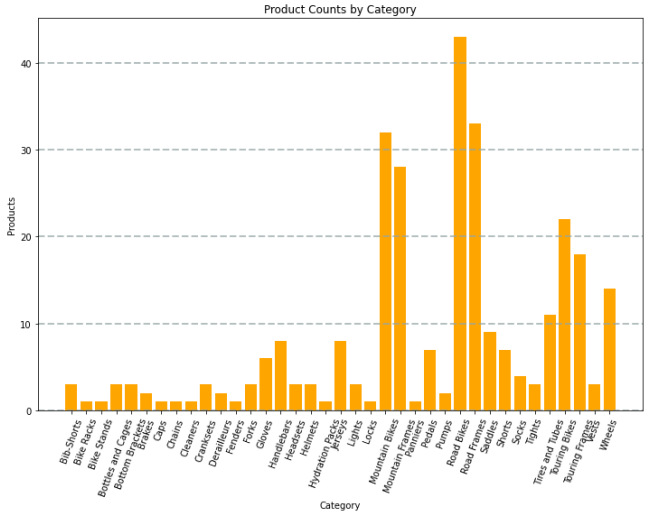
_sqldf: pyspark.sql.dataframe.DataFrame = [Category: string, ProductID: string]

Table | Visualization 1 | +



Graphics packages

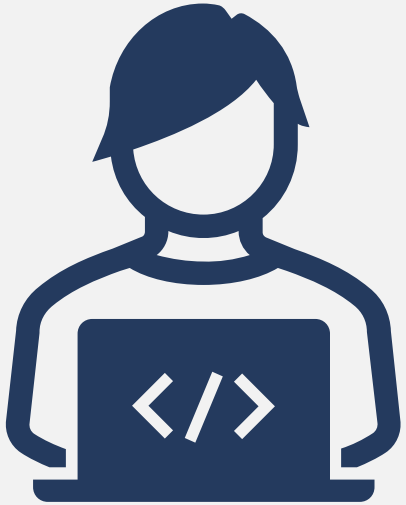
```
from matplotlib import pyplot as plt

fig = plt.figure(figsize=(12,8))
plt.bar(x=data['Category'],
        height=data['ProductCount'],
        color='orange')
plt.show()
```



Exercise: Use Spark in Azure Databricks

Use the hosted lab environment provided, or view the lab instructions at the link below:

https://aka.ms/mslearn-databricks-spark



Knowledge check

Which definition best describes Apache Spark?

- ❑ A highly scalable relational database management system
- ❑ A virtual server with a Python runtime
- ☑ A distributed platform for parallel data processing using multiple languages

You need to use Spark to analyze data in a parquet file. What should you do?

- ☑ Load the parquet file into a dataframe
- ❑ Import the data into a table in a serverless SQL pool
- ❑ Convert the data to CSV format

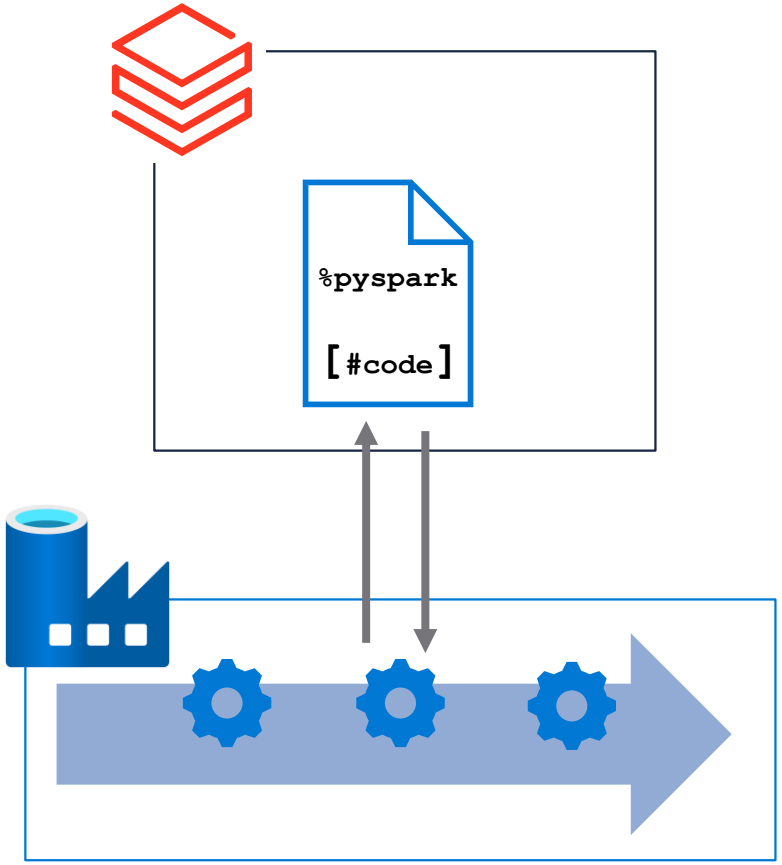
You want to write code in a notebook cell that uses a SQL query to retrieve data from a view in the Spark catalog. Which magic should you use?

- ❑ %spark
- ❑ %pyspark
- ☑ %sql

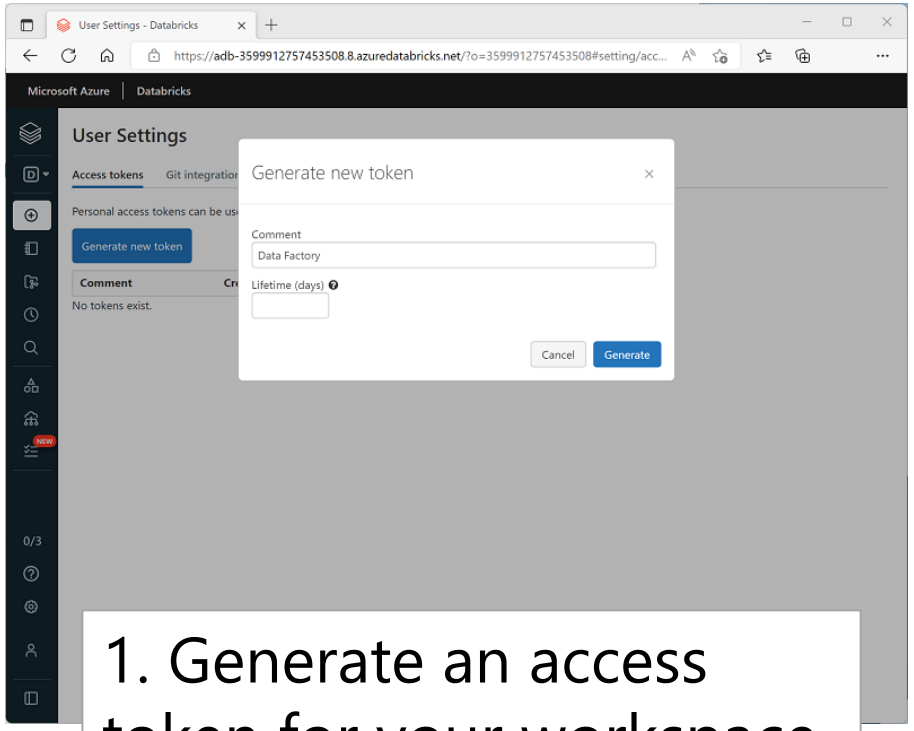
Run Azure Databricks notebooks in Azure Data Factory

Azure Databricks notebooks and pipelines

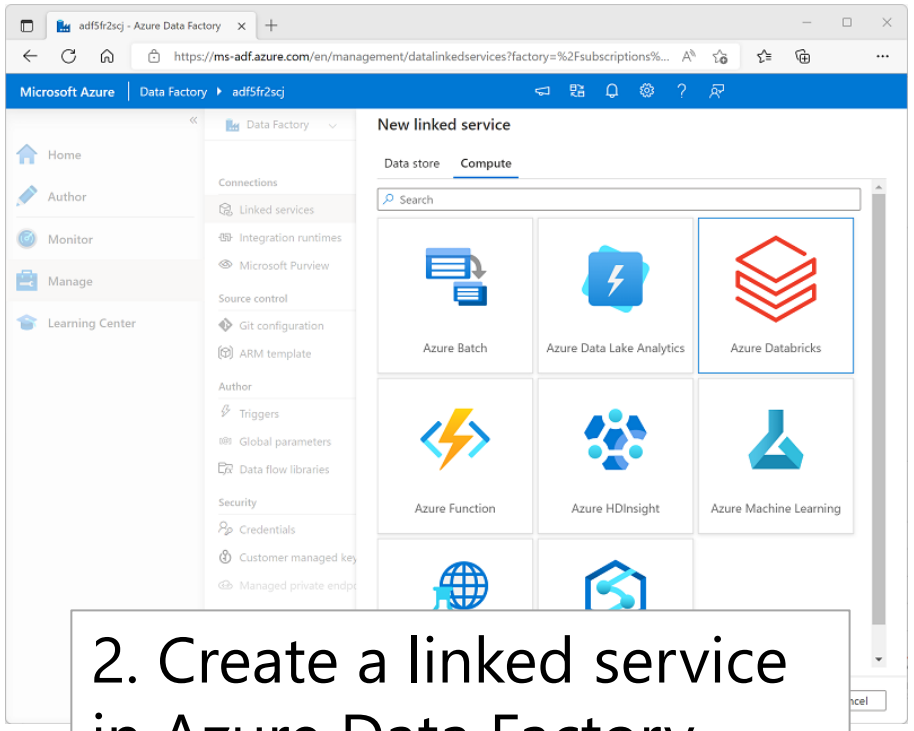
- Use notebooks to develop and test data transformation code
- Incorporate notebooks into data ingestion and transformation pipelines using Azure Data Factory or Azure Synapse Analytics
- Run pipelines on-demand, at scheduled times, or in response to events



Create a linked service for Azure Databricks



1. Generate an access token for your workspace

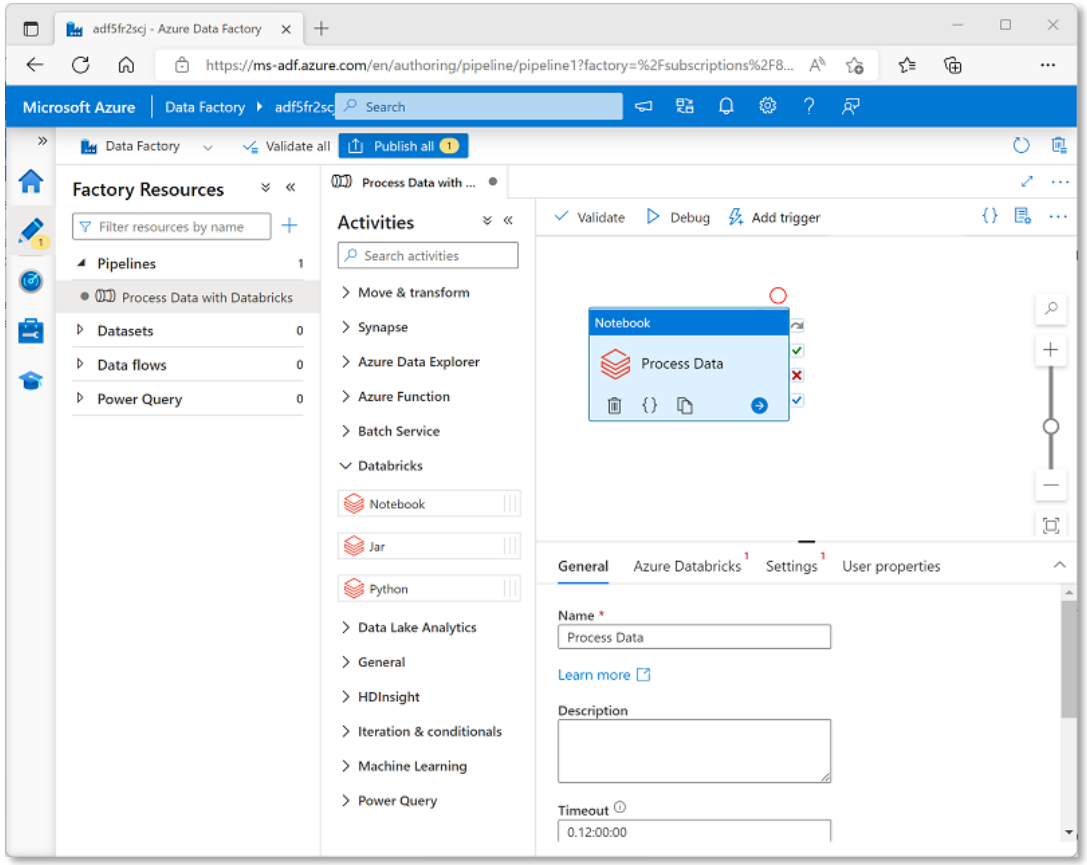


2. Create a linked service in Azure Data Factory

Use a Notebook activity in a pipeline

Add a *Notebook* activity to a pipeline, specifying:

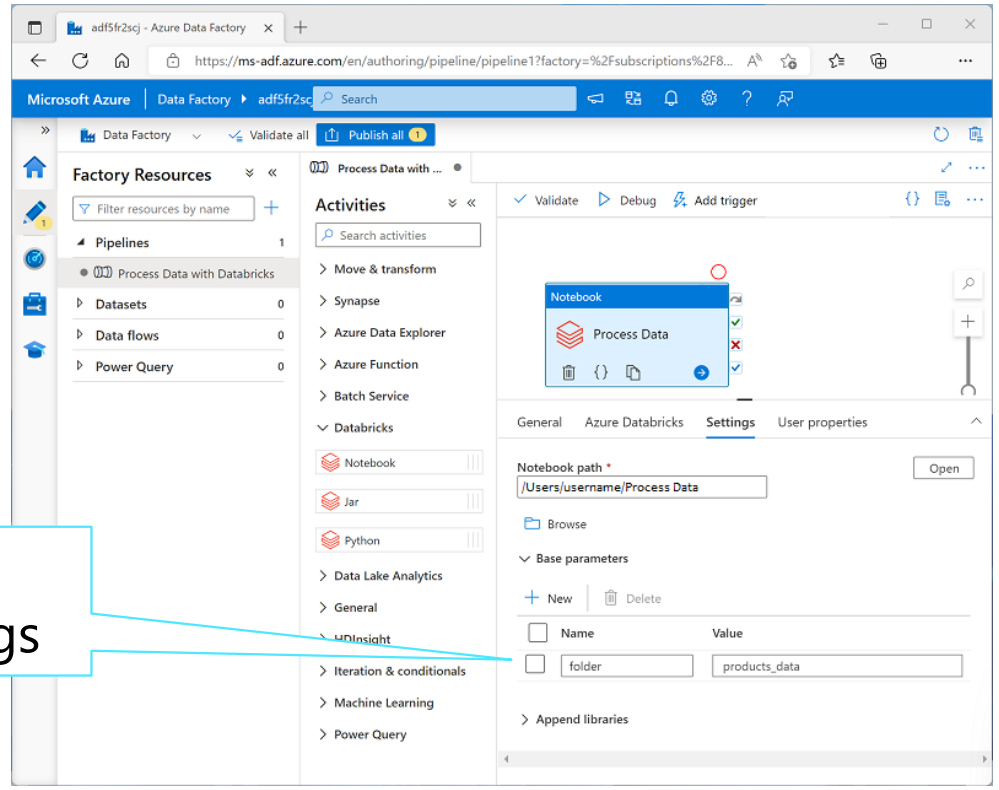
- **General** properties such as name, timeout, and number of retries
- **Azure Databricks** properties - the linked service for your workspace
- **Settings**, such as the notebook path and parameter details
- **User properties** to define custom configuration values



Use parameters in a notebook

Use the ***dbutils*** library to work with parameters

```
# define parameter with default value
dbutils.widgets.text("folder", "data")

# Get input parameter value if passed
folder = dbutils.widgets.get("folder")
...

# Return output parameter
path = "dbfs:/{0}/products.csv".format(folder)
dbutils.notebook.exit(path)
```



Set parameter values in notebook activity settings

Demo: Run an Azure Databricks Notebook with Azure Data Factory

You can try this for yourself later by following the instructions at the link below:

https://aka.ms/mslearn-databricks-factory

Knowledge check

You want to connect to an Azure Databricks workspace from Azure Data Factory. What must you define in Azure Data Factory?

- ❑ A global parameter
- ☑ A linked service
- ❑ A customer managed key

You need to run a notebook in the Azure Databricks workspace referenced by a linked service. What type of activity should you add to a pipeline?

- ☑ Notebook
- ❑ Python
- ❑ Jar

You need to use a parameter in a notebook. Which library should you use to define parameters with default values and get parameter values that are passed to the notebook?

- ❑ notebook
- ❑ argparse
- ☑ dbutils.widget

Further reading

Data engineering with Azure Databricks
https://aka.ms/mslearn-azure-databricks