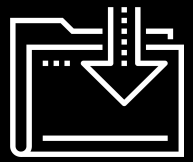
Databricks

Data Boot Camp

Lesson 22.4



© 2022 edX Boot Camps LLC. Confidential and Proprietary. All Rights Reserved.

Congratulations! You've reached the final day of formal in-class instruction. After today, you will consolidate your data analytics skills in your final projects.

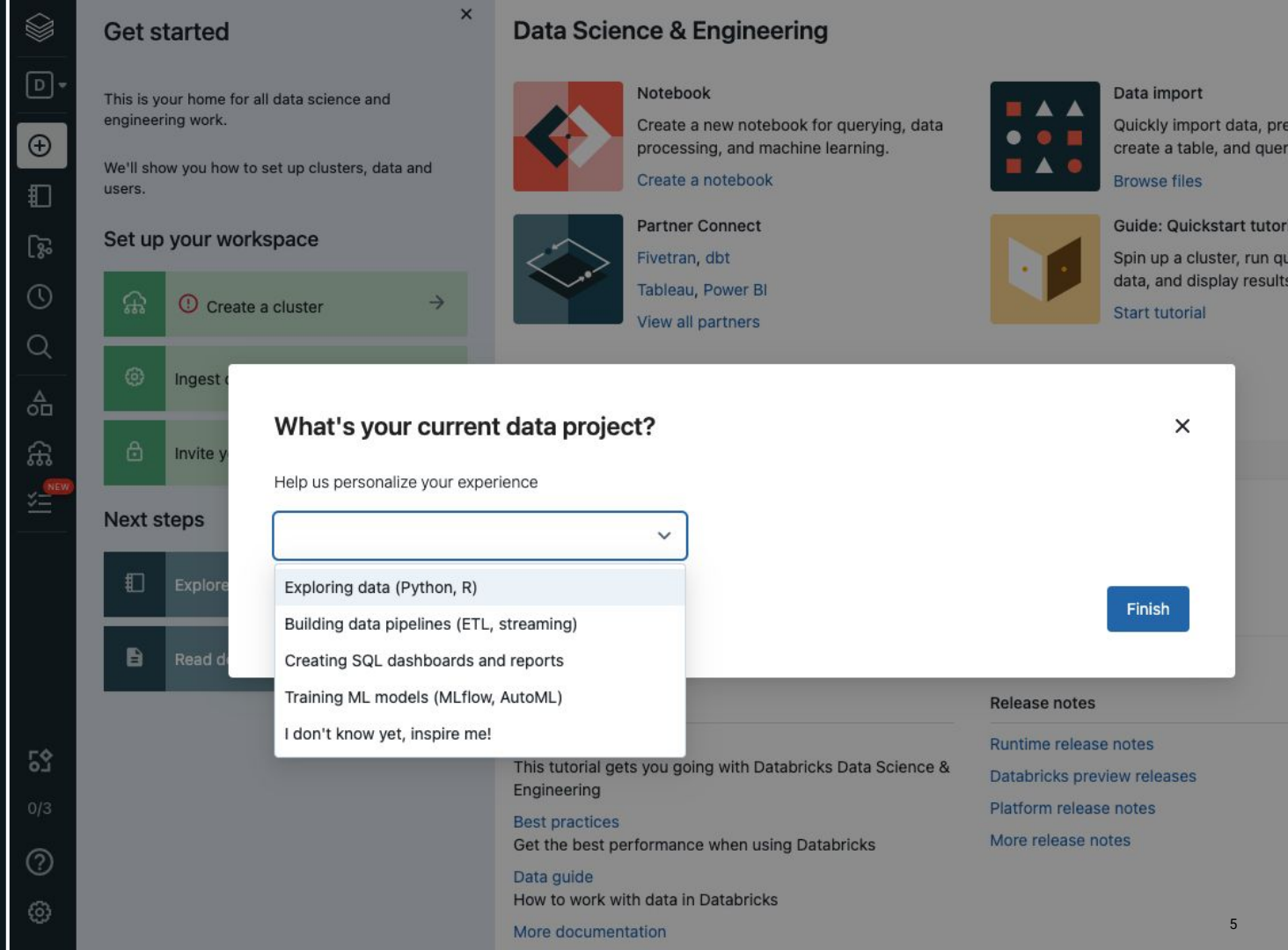
Class Objectives

By the end of this lesson, you will be able to:

- Explain the purpose, key features, and applications of Databricks.
- Set up a Databricks environment.
- Identify the key components of a Databricks environment.
- Navigate the Databricks workspace using dbutils.
- Import data into a new notebook by using Parquet files, CSV files, and S3.
- Explain the advantage of Parquet as a big-data storage format.
- Perform complex data analysis, including joins, using the Python and SQL interfaces.
- Identify two advantages of using Databricks over PySpark for data analysis.

Introduction to Databricks and Signing Up

Today you'll get a chance to practice your SQL skills by using the Databricks interface.



Introduction to Databricks

Companies look for SQL expertise because it is the industry standard for querying databases, including filtering data and working across multiple tables.

Introduction to Databricks

This class will build on your knowledge of Spark and SQL, including filtering data.

- There is significant work required to set up, manage, and use a cluster of on-site CPUs running Spark.
- This includes managing the hardware and scheduling any work to be performed, depending on demand.
- This overhead cost leads to many companies looking for a managed, cloud-based solution.

Introduction to Databricks

Databricks:

- Is a cloud platform for Spark.
- Enables us to run Apache Spark on cloud servers provided by Amazon's AWS or Microsoft's Azure, for example.
- Provides a robust system to manage and optimize clusters of computers for large-scale data analysis.

https://databricks.com/

What are some possible advantages of running Spark on the cloud?

Databrick Advantages

Ease of use:	Databricks can automatically scale its activities up or down based on what's needed. Managing Spark clusters takes a lot of time, and there are many opportunities to make mistakes. Databricks eliminates much of the complexity involved in the process.
Enables collaboration:	Multiple team members can work on the same notebook. For example, after a data engineer prepares the data, one person can work on data analysis while another person works on data visualization.
Cost effective:	Spark can often significantly reduce the time and money spent on managing clusters of computers for data analysis. For a data team with budgetary constraints, running a Spark cluster in the cloud will be cheaper than managing an on-site cluster, which will have administrative costs.
Versatile:	You can use Python, R, SQL, and Scala in the same Databricks notebook.

Databrick Disadvantages

Other cloud platforms can run Spark, and Databricks is more expensive than many of them.

One heavily used cluster can cost thousands of dollars per month or more.

Spark is designed to work with multiple sources of data.

(We'll work with different sources of data, including CSV and Parquet files.)

What is a possible advantage of using a format like Parquet over CSVs?

Parquet Advantage

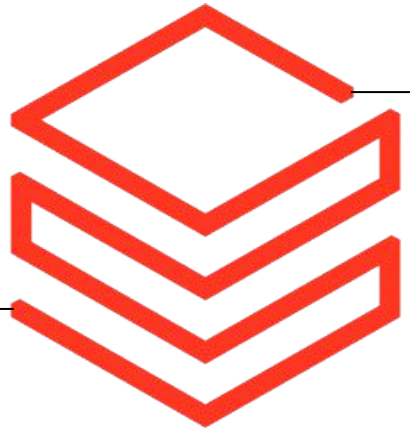
Parquet loads data by specific columns instead of rows.

Traditional data formats store data by row. If Spark uses multiple nodes to perform queries, each node would need to load a copy of all rows of the dataset.

When working with large amounts of data, loading only the specified columns can save time and computing resources.

Key Benefits of Using Databricks

Data analytics teams can spend more time analyzing data and less time configuring and managing clusters.



Teams can also create visualizations quickly to explore data and can create dashboards for an audience.

databricks

Activity: Sign Up for Databricks

In this activity, you will sign up for a free Databricks Community Edition account.

Suggested Time:

10 Minutes

Time's Up! Let's Review.

Questions?

Databricks Basics

Instructor Demonstration

Databricks Basics

Questions?

Activity: Databricks Basics

In this activity, you will will perform basic navigation and data analysis tasks in Databricks.

Suggested Time:

15 Minutes

Time’s Up! Let’s Review.

Questions?

Joins

Instructor Demonstration

Joins

Activity: Joining Animal Species

In this activity, you will perform queries on datasets using both PySpark and SQL interfaces in Databricks.

Suggested Time:

15 Minutes

Time’s Up! Let’s Review.

Questions?


Break

Group Activity

Time to divide into teams!

Group Programming Activity:

In this activity, you will work in groups of three or four to gather data insights from a fictional company's database.

You will query the database in SQL, and many of the queries will require working across multiple tables.

Based on your findings, the groups will create a brief report with recommendations.

Suggested Time:

60 Minutes

Group Presentations

Time’s Up! Let’s Review.

Questions?


The End