

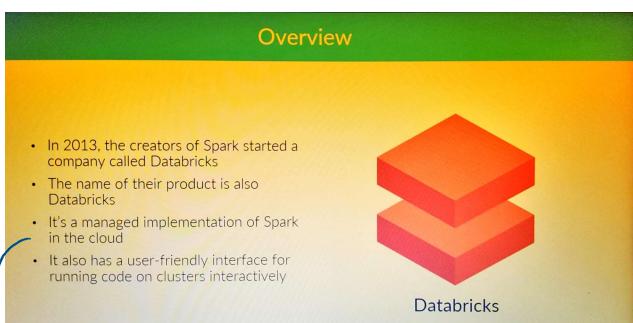
Spark in databricks

Thursday, 5 January 2023 3:47 PM

Overview

Spark.

- It is an open source framework for doing big data processing.
→ It was developed as a replacement for Apache Hadoop's Map Reduce framework.
→ Both Spark and MapReduce process data on compute clusters.
→ Both Spark's big advantage is that it does in-memory processing, which can be orders of magnitude faster than the disk based processing that MapReduce uses.



Set up demo



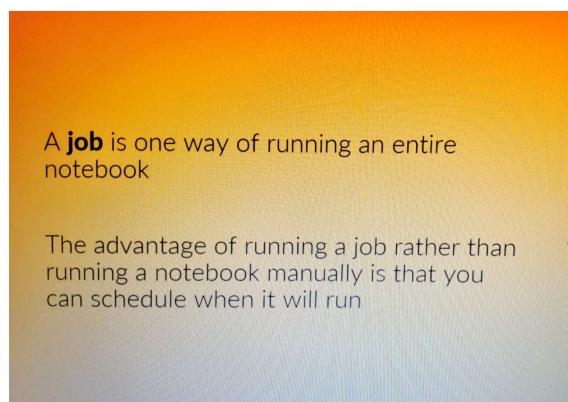
[https://github.com/cloudacademy/azure-databricks.](https://github.com/cloudacademy/azure-databricks)

From <https://cloudacademy.com/course/running-spark-on-azure-databricks/notebooks/?context_resource=lp&context_id=3191>

Jobs

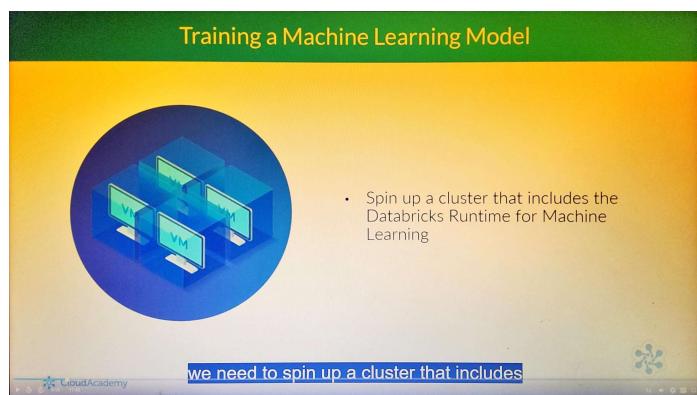
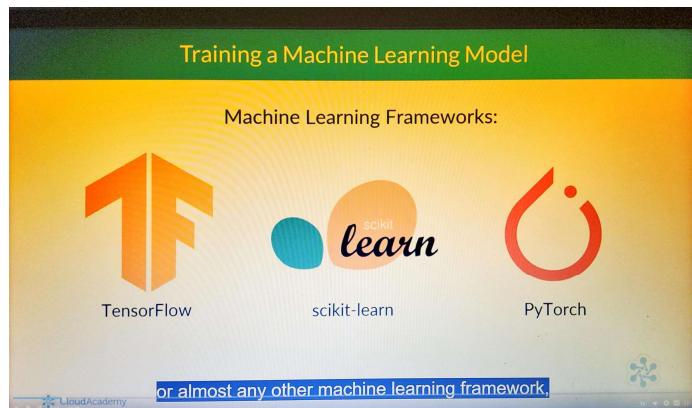
- Notebooks for experimentation.
→ But if you've put together a solid workflow or you want to run a regular schedule. → Answer: create a job.

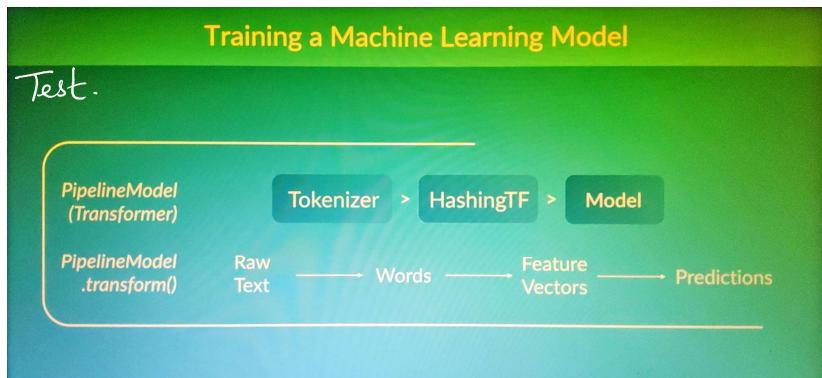
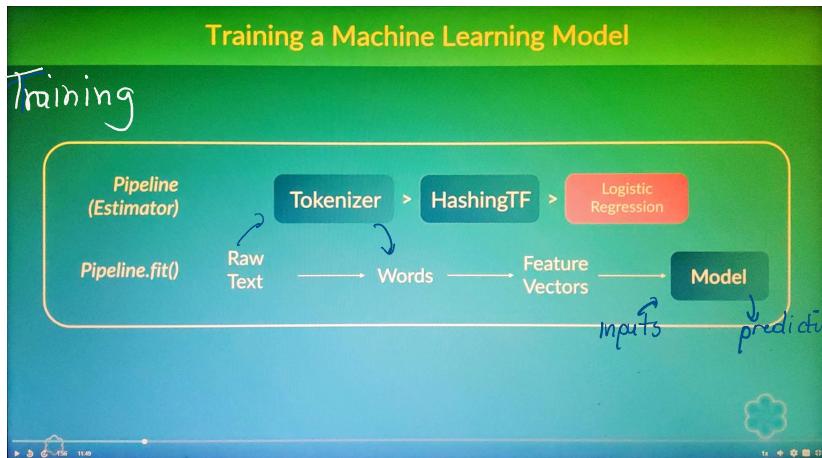
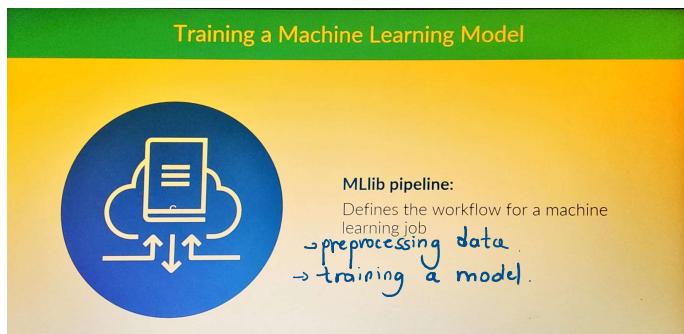
- Notebooks for experimentation.
- But if you've put together a solid workflow or you want to run it on a regular schedule. ⇒ Answer: create a job.



If you run a job on an existing cluster, instead of on a new one, it'll be more expensive. Why? It's because when you run a job on an existing cluster, DB considers it to be an interactive workload, so you get charged the interactive price. If you create a new cluster for the job to run on and that cluster is only active while the job is running, then you get charged the automated workload price, which is less.
auto scaling activates 2 to n (workers)
 max 8.

Training ML Model





Data: handwritten

Decision trees

These notebooks show you how to perform classifications with decision trees.

Decision trees for digit recognition notebook

Decision Trees for handwritten digit recognition

This notebook demonstrates learning a Decision Tree using Spark's distributed implementation. It gives the reader a better understanding of some oftentimes counterintuitive aspects of decision trees, using examples to demonstrate how tuning the hyperparameters can improve accuracy.

Background: To learn more about Decision Trees, check out the resources at the end of this notebook. The visual description of ML and Decision Trees provides nice intuition helpful to understand this notebook, and Intuitively gives lots of details.

Data: We use the classic MNIST handwritten digit recognition dataset.

Goal: Our goal for our data is to learn how to recognize digits 0 - 9 from images of handwriting. However, we will focus on understanding trees, not on this particular learning problem.

Taking it further: There are several hyperparameters which can affect the accuracy of the learned model. There is no one "best" setting for these for all datasets. To get the optimal accuracy, we need to tune these hyperparameters based on our data.

we need to import it to our workspace.

Deploying a trained model

Deploying a Trained Model

Now that you have a trained model, you can use it in a production environment



Save the trained model



Import to the system where you want to run it

Then you can import it into the production system

Deploying a Trained Model

You can save the model as: **MLWriter**, **MLeap**, or **Databricks ML Model Export**



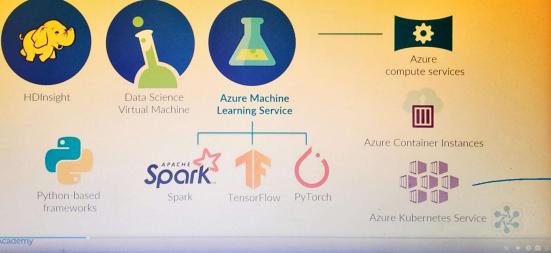
- To save a pipeline, run the write and save methods on it.
`pipeline.write.save("example-pipeline")`
- To load a saved model on another system, run the read and load methods on an empty pipeline.
`pipeline.read.load("example-pipeline")`

where you load the saved pipeline needs

MLWriter → is a Spark Class that can save ML components.
 It's an abstract class, and its methods are inherited by Spark's ML Components

Deploying a Trained Model

There are a number of options for running Spark on Azure:



HDInsight
Data Science Virtual Machine
Azure Machine Learning Service
Python-based frameworks
Apache Spark
TensorFlow
PyTorch
Azure Container Instances
Azure Kubernetes Service

(3 options)

Automatically scales up and down as needed.

Deploying a Trained Model

Steps to access the Azure ML service from Databricks:

1. Install the Azure ML SDK as a library in Databricks
2. Attach that library to one or more of the Databricks clusters
3. Call the Azure ML service from any of the Databricks clusters that have the library attached to them

you'll be able to call the Azure ML service

Deploying a Trained Model



Create a workspace
The workspace is where you keep track of everything related to your machine learning activities
You need to register your trained model in the workspace

Deploying a Trained Model



Create a scoring script
This script will be used to load input data, feed it into the model, and return a prediction
It has to include an init function and a run function

If it has to include an init function and a run function.

Deploying a Trained Model



Create a container image
The image should contain the trained model, the scoring script, and any dependencies that are required by either the model or the script
Dependencies are managed by Conda

Deploying a Trained Model



Azure Container Instances
This service makes it easy to spin up a single container instance
If you expect to have a low volume

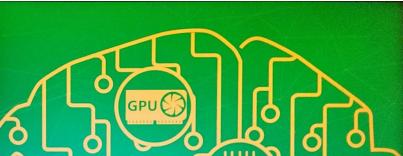
Deploying a Trained Model



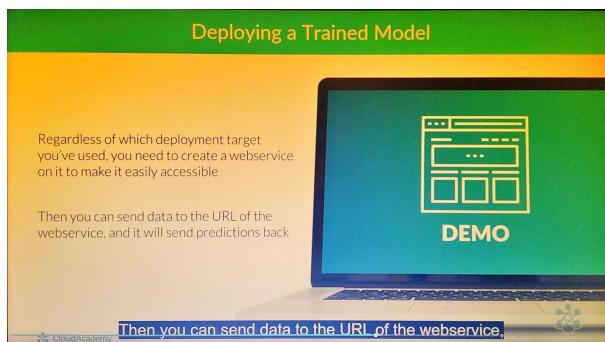
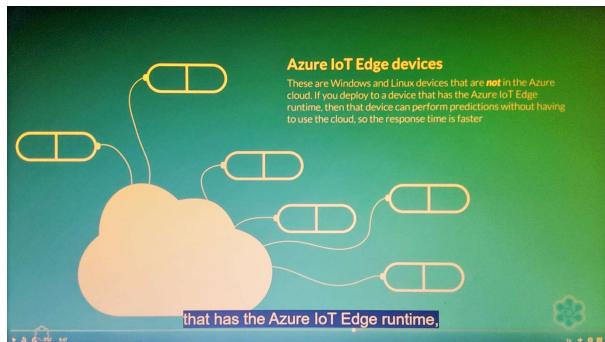
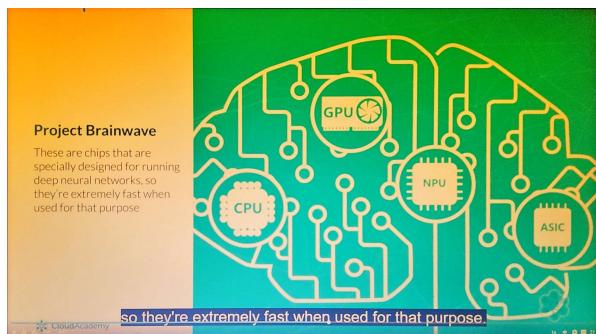
Azure Kubernetes Service
AKS has the advantage of being able to scale up and down based on demand. It's a full container orchestration service, and it's becoming the standard for running containers on Azure.

AKS has the advantage of being able to scale up

other options:



Project Brainwave
These are chips that are specially designed for training

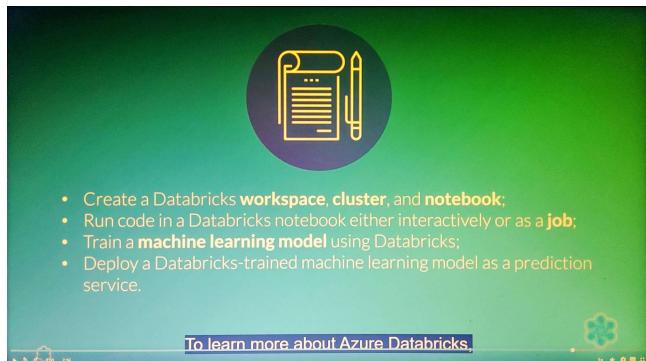
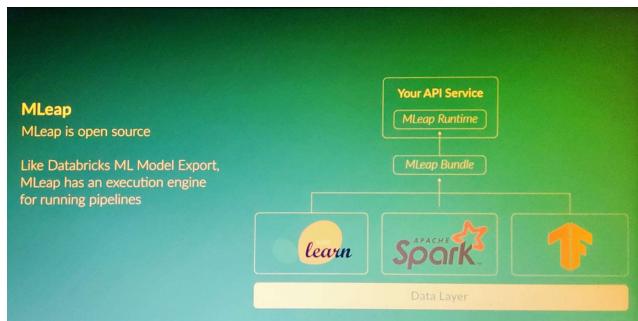
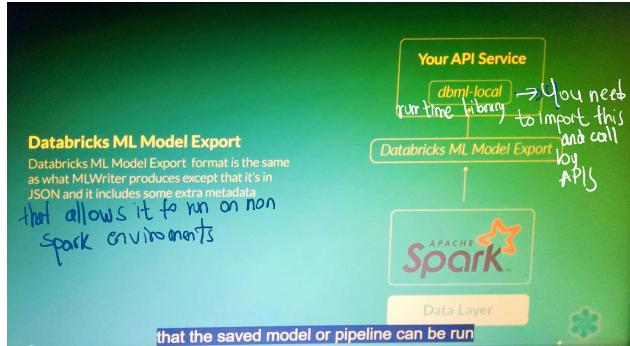


Microsoft Azure

Workspace

- Workspace
 - ? Documentation
 - </> Release Notes
 - Training & Tutorials
 - Shared
 - Users
 - Databricks_AML...
 - decision-trees
 - test





Question 1

CORRECT

Which of the following statements about Apache Spark is correct?

It does memory-based processing.

It cannot process data on compute clusters.

It is slower than MapReduce.

It has no pre-built machine learning algorithms.

I don't know yet

Explanation Both Spark and MapReduce process data on compute clusters, but one of Spark's big advantages is that it does in-memory processing, which can be orders of magnitude faster than the disk-based processing that MapReduce uses. There are plenty of other differences between the two systems as well, but we don't need to go into the details here. Not only does Spark handle data analytics tasks, but it also handles machine learning. It has a library called MLlib that includes a variety of pre-built algorithms, such as logistic regression, naive Bayes, and random forest. At the moment, it doesn't include neural networks. However, you can still create neural networks on Spark using other machine learning frameworks, such as TensorFlow.
[Learn more: /course/running-spark-on-azure-databricks/overview/](#)

Bookmark

Review answers Recommendation Other updated skills

1 2 3 4 5

Question 2

CORRECT

What is a "notebook" in the context of Azure Databricks?

a distributed filesystem that's installed on a Databricks cluster

a document where you can enter some code, run it, and view the results within the document

an open-source framework for doing big data processing

a class that can save machine learning components

I don't know yet

Explanation A notebook is a document where you can enter some code and run it, and the results will be shown in the notebook.
[Learn more: /course/running-spark-on-azure-databricks/notebooks/](#)

Bookmark

Something wrong with this question? Report an issue

Screen clipping taken: 1/6/2023 7:22 PM

Screen clipping taken: 1/6/2023 7:21 PM

Review answers Recommendation Other updated skills 9

1 2 3 4 5

Question 3

Question

Correct

In Azure Databricks, what is the difference in price between running a job on an existing cluster and running a job on a new cluster?

- They cost the same.
- Running a job on a new cluster is more expensive.
- It depends on the size of your job.
- Running a job on an existing cluster is more expensive.
- I don't know yet

Explanation

If you run a job on an existing cluster instead of a new one, it'll be more expensive.

Bookmark

Learn more: [/course/running-spark-on-azure-databricks/jobs/](#)

Something wrong with this question? [Report an issue](#)

Review answers Recommendation Other updated skills 9

1 2 3 4 5

Question 4

Question

Correct

What is Apache Spark?

- an open-source framework for doing big data processing
- a document where you can enter some code, run it, and view the results within the document
- a class that can save machine learning components
- a distributed filesystem that's installed on a Databricks cluster
- I don't know yet

Explanation

Apache Spark is an open-source framework for doing big data processing.

Bookmark

Learn more: [/course/running-spark-on-azure-databricks/overview/](#)

Something wrong with this question? [Report an issue](#)

Screen clipping taken: 1/6/2023 7:22 PM

Screen clipping taken: 1/6/2023 7:22 PM

Review answers Recommendation Other updated skills 9

1 2 3 4 5

Question 5

Question

Correct

Which of the following statements about running notebooks and jobs in Azure Databricks is false?

- You can schedule when a job will run.
- You can schedule when a notebook will run.
- Running a job allows you to keep a record of previous runs.
- A job is simply one way of running an entire notebook.
- I don't know yet

Explanation

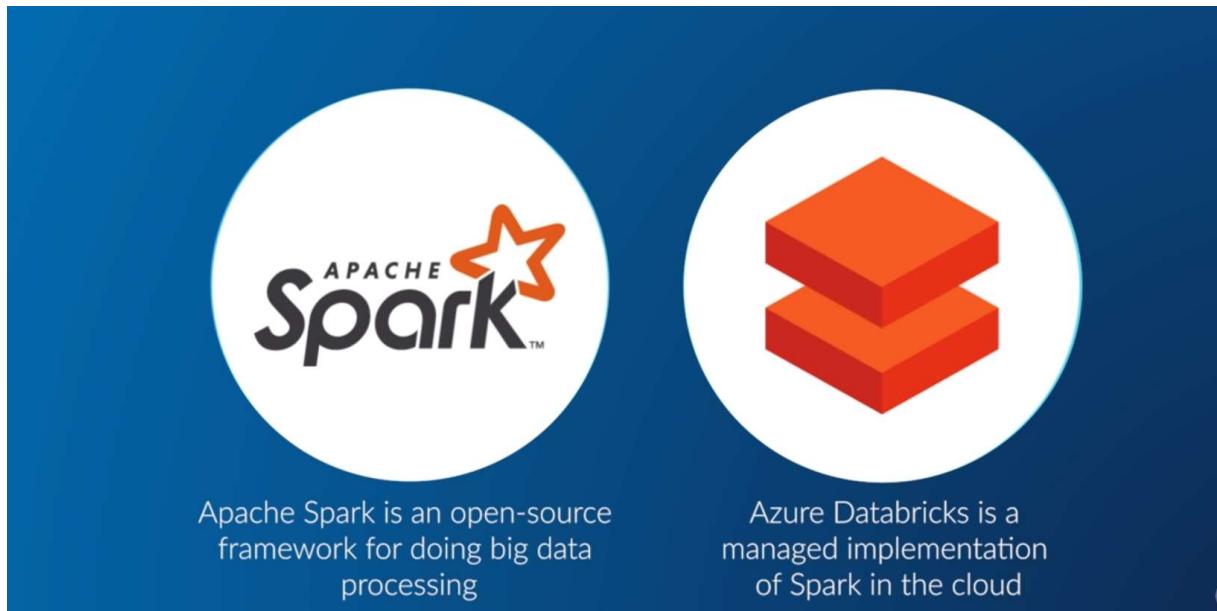
A job is simply one way of running an entire notebook. The advantage of running a job rather than running a notebook manually is that you can schedule when it will run. It also allows you to keep a record of previous runs.

Bookmark

Learn more: [/course/running-spark-on-azure-databricks/jobs/](#)

Something wrong with this question? [Report an issue](#)

Screen clipping taken: 1/6/2023 7:23 PM



Apache Spark is an open-source framework for doing big data processing

Azure Databricks is a managed implementation of Spark in the cloud

Screen clipping taken: 1/6/2023 7:24 PM

Workspace

A Databricks workspace is where you store your notebooks and other related items

Screen clipping taken: 1/6/2023 7:25 PM

Databricks File System

DBFS is a distributed filesystem that's installed on a Databricks cluster and backed by Azure Storage

Screen clipping taken: 1/6/2023 7:25 PM

Job

A job is a way of running an entire notebook at scheduled times

It also keeps a record of previous runs

Screen clipping taken: 1/6/2023 7:25 PM

MLib

MLlib is Spark's own machine learning library, and it comes preinstalled on all of the Databricks runtime versions



Screen clipping taken: 1/6/2023 7:25 PM

Saving

Three options for saving a trained model are
MLWriter, MLeap, and Databricks ML Model
Export

Screen clipping taken: 1/6/2023 7:26 PM

You can implement prediction service by running a Mleap Bundle on any system.
That has the Mleap Runtime installed

MLWriter -> Spark Environments
Mleap -> Non Spark Environments