Spark in databricks

Thursday, 5 January 2023 3:47 PM

Overview

Spark.

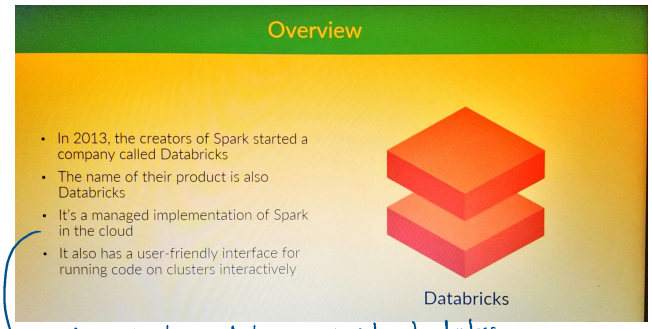
It is an open source framework for doing big data processing

→ It was developed as a replacement for Apache Hadoop's Map Reduce Framework.

→ Both Spark and MapReduce process data on compute clusters.

but one of Spark's big advantage is that is does in-memory processing, which can be orders of magnitude faster than the disk based processing that MapReduce uses.

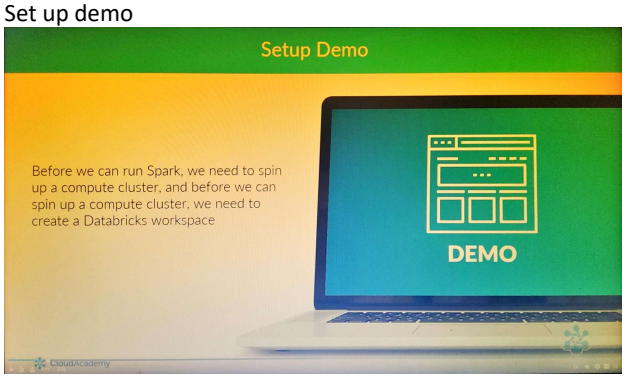
It doesn't include N.N. but you can use another framework on it, such as Tensorflow.



→ You don't need to worry about clusters.

Microsoft is partner of Databricks, to integrate it with Azure Services.

Set up demo



https://github.com/cloudacademy/azure-databricks.

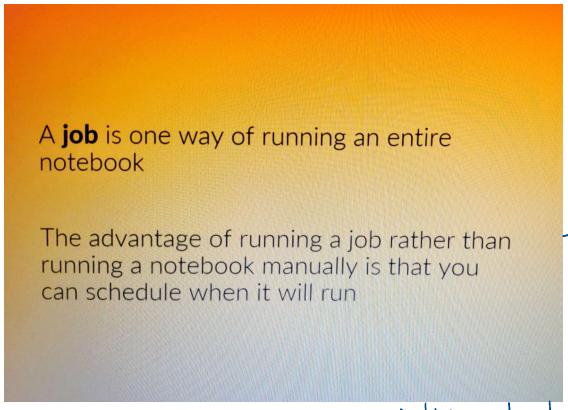
From <https://cloudacademy.com/course/running-spark-on-azure-databricks/notebooks/?context_resource=lp&context_id=3191>

Jobs

→ Notebooks for experimentation.

→ But if you've put together a solid workflow or you want to run ... schedule. => Answer: create a job.

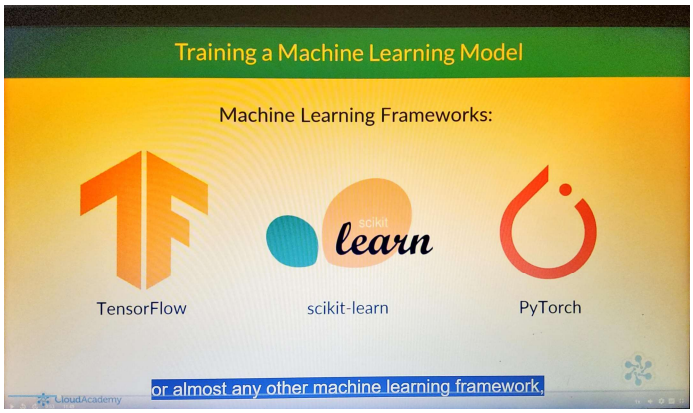
→ Notebooks for experimentation.
→ But if you've put together a solid workflow or you want to run it on a regular schedule. ⇒ Answer: create a job.

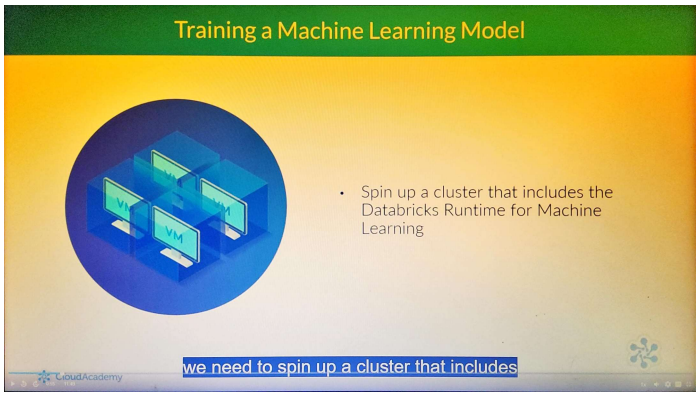


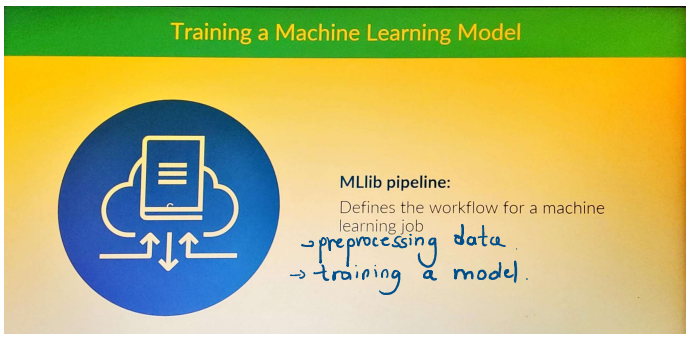
→ It also allows you to keep a record of previous runs!

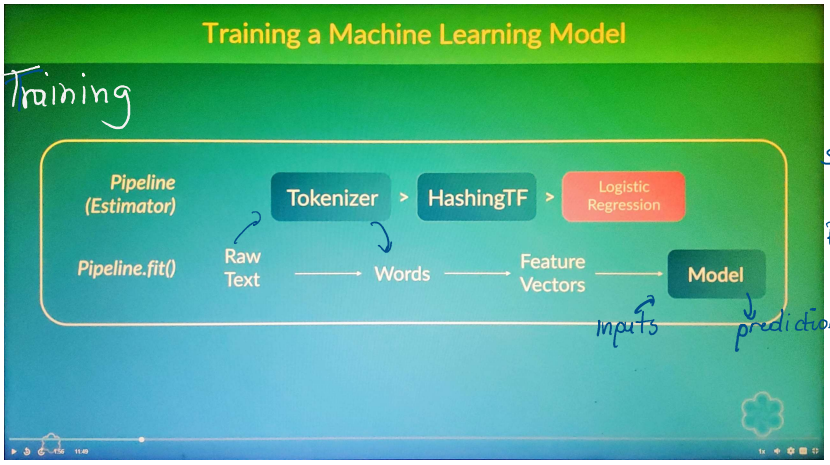
If you run a job on an existing cluster, instead of on a new one, it'll be more expensive. why? It's because when you run a job on an existing cluster, DB considers it to be an interactive workload, so you get charged the interactive price.
If you create a new cluster for the job to run on and that cluster 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

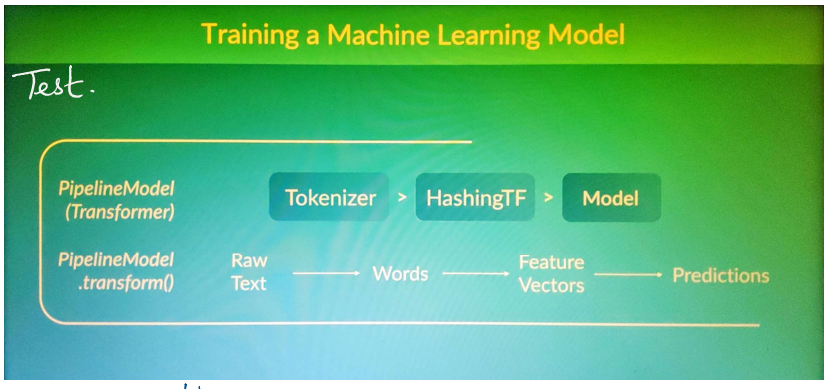




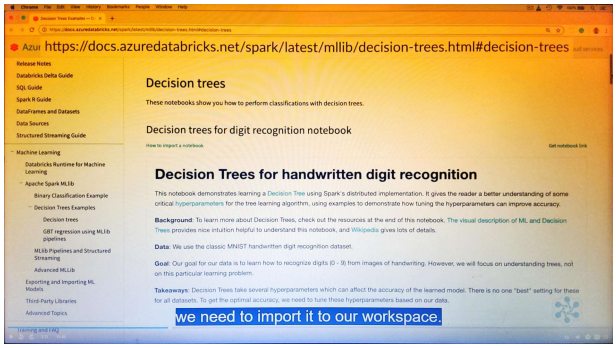




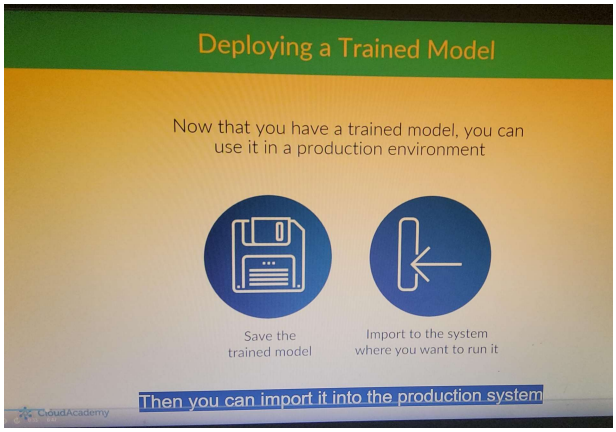
stages in blue → transformers (they transform the df into another)
Red is an estimator
It's an algorithm that takes a df and produces a model

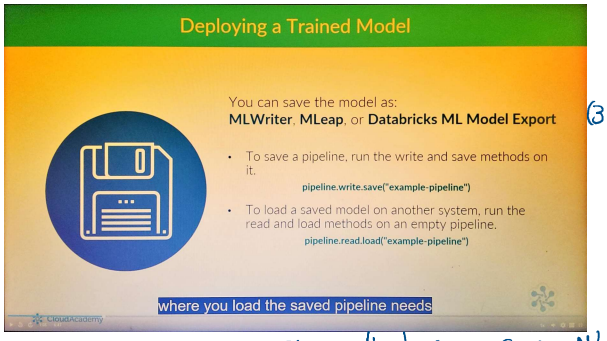


Data: handwritten



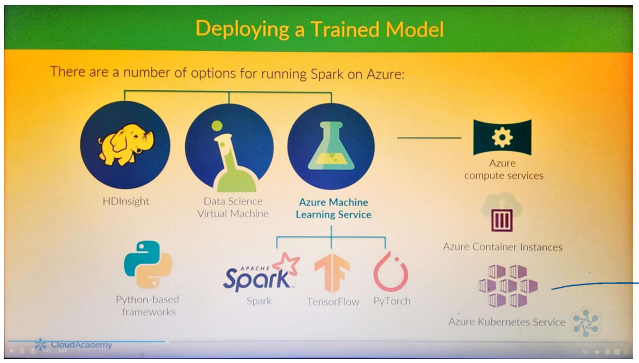
Deploying a trained model



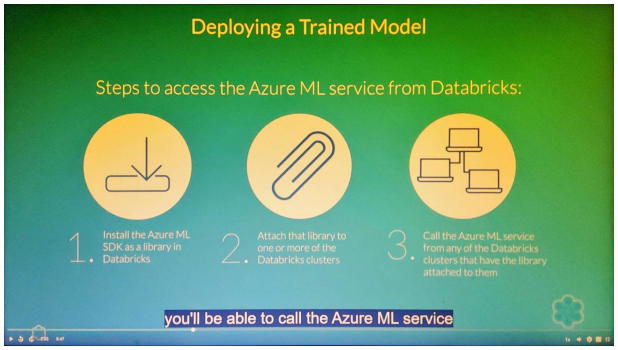


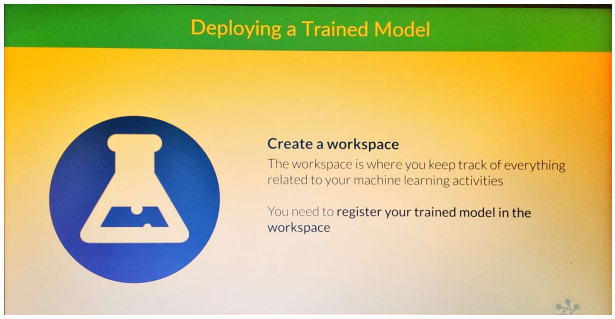
(3 options)

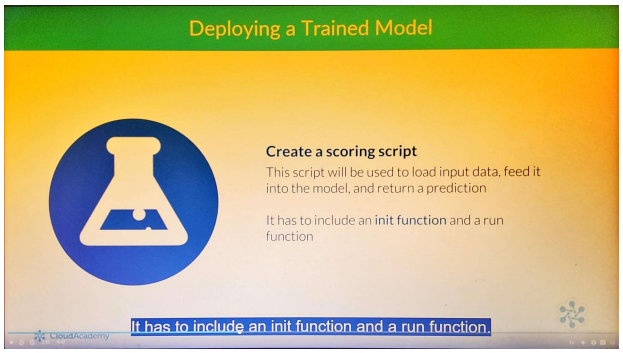
MLWriter → is a Spark Class that can save ML components. It's an abstract class, and its methods are inherited by Spark's ML Components

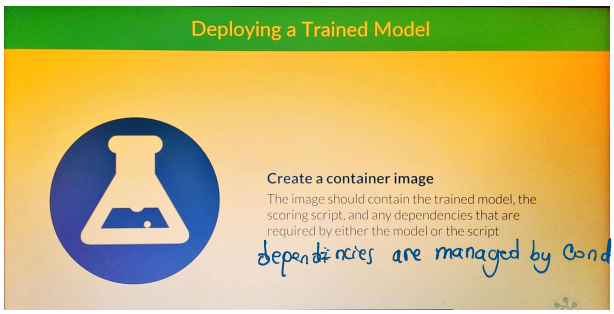


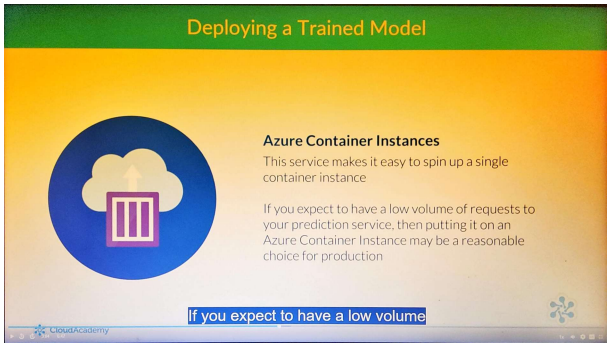
Automatically scales up and down ~~the~~ as needed.

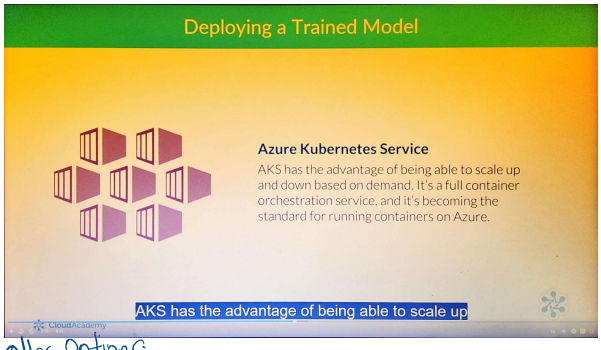




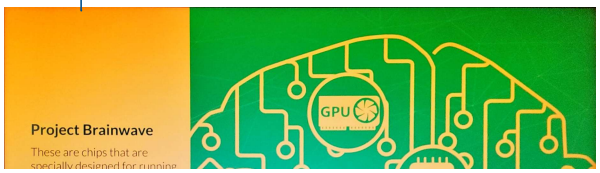


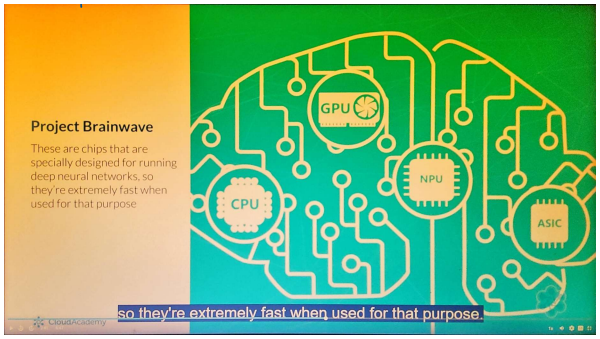


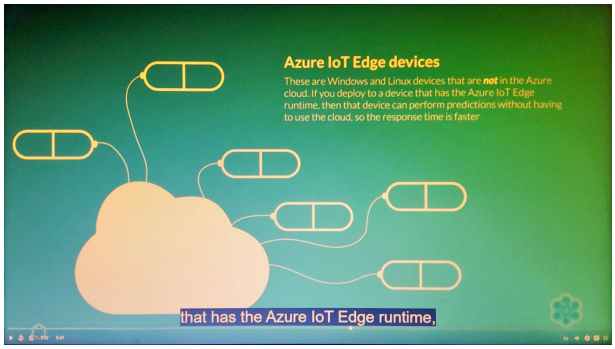


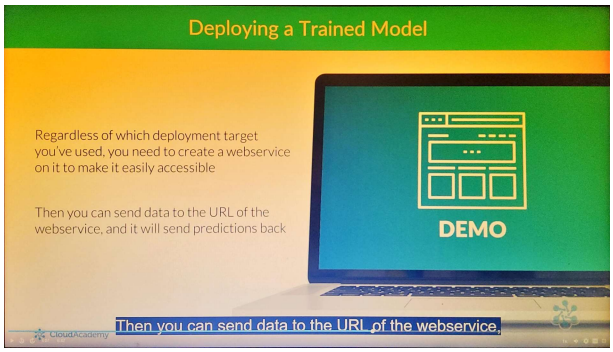


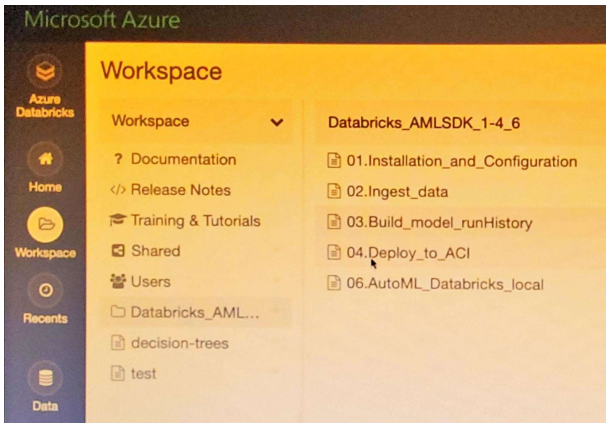
other options:

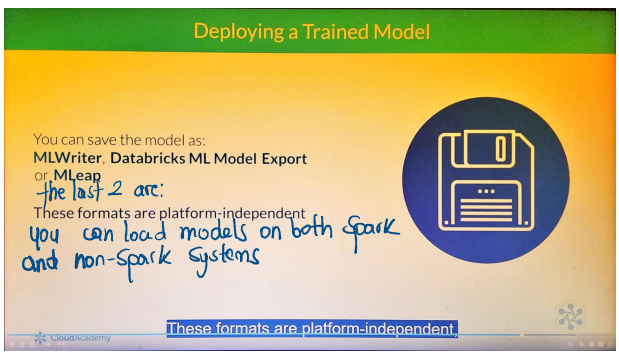


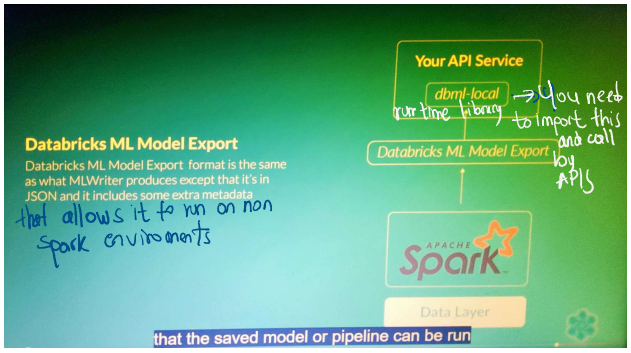

Project Brainwave
These are chips that are specially designed for running deep neural networks, so they're extremely fast when used for that purpose
GPU
NPU
CPU
ASIC
so they're extremely fast when used for that purpose.

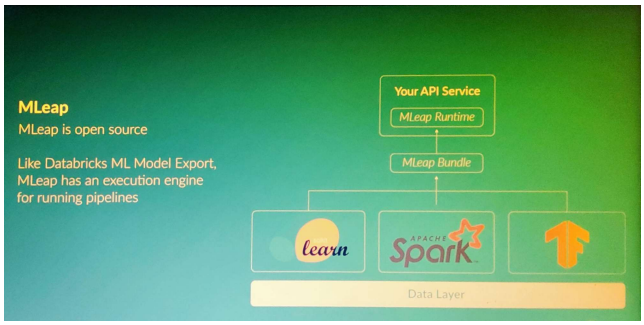

Azure IoT Edge devices
These are Windows and Linux devices that are not in the Azure cloud. If you deploy to a device that has the Azure IoT Edge runtime, then that device can perform predictions without having to use the cloud, so the response time is faster
that has the Azure IoT Edge runtime,

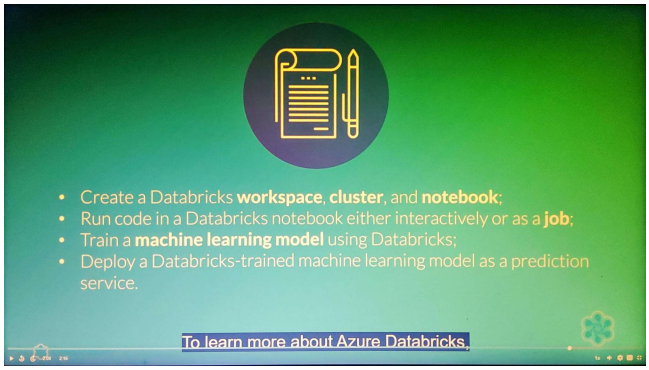

Deploying a Trained Model
Regardless of which deployment target you've used, you need to create a webservice on it to make it easily accessible
Then you can send data to the URL of the webservice, and it will send predictions back
DEMO
Then you can send data to the URL of the webservice.

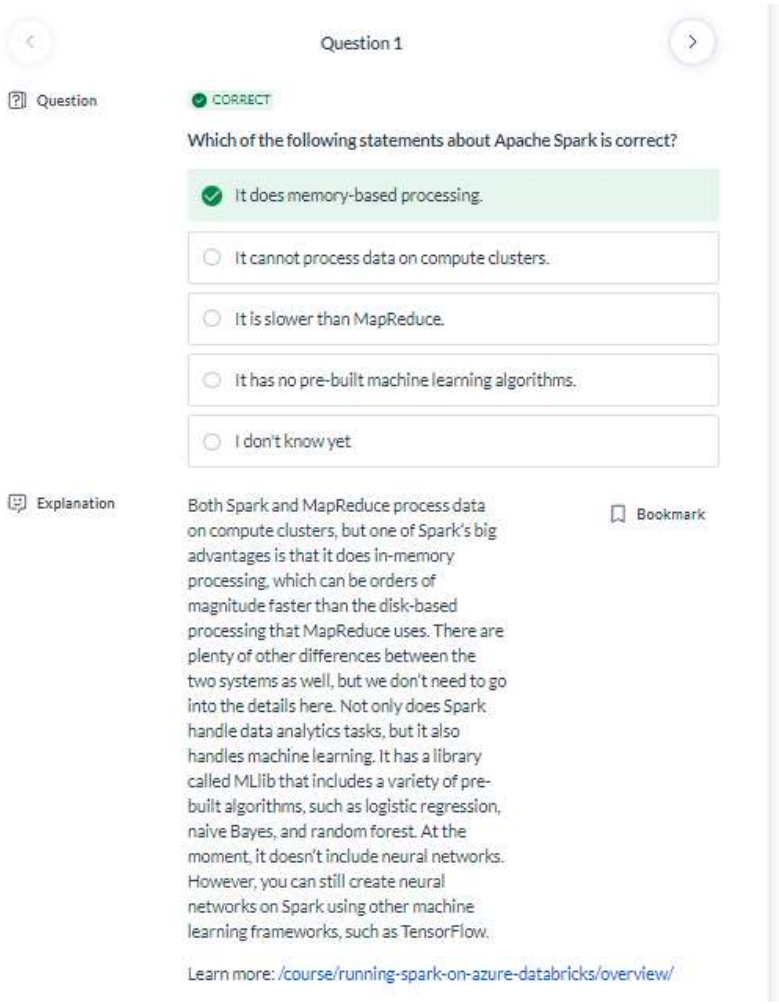

Microsoft Azure
Azure Databricks
Home
Workspace
Recents
Data
Workspace
Workspace
Documentation
Release Notes
Training & Tutorials
Shared
Users
Databricks_AML...
decision-trees
test
Databricks_AMLSDK_1-4_6
01.Installation_and_Configuration
02.Ingest_data
03.Build_model_runHistory
04.Deploy_to_ACI
06.AutoML_Databricks_local


Deploying a Trained Model
You can save the model as:
MLWriter, Databricks ML Model Export
or MLeap
the last 2 are:
These formats are platform-independent
you can load models on both Spark and non-Spark systems
These formats are platform-independent.

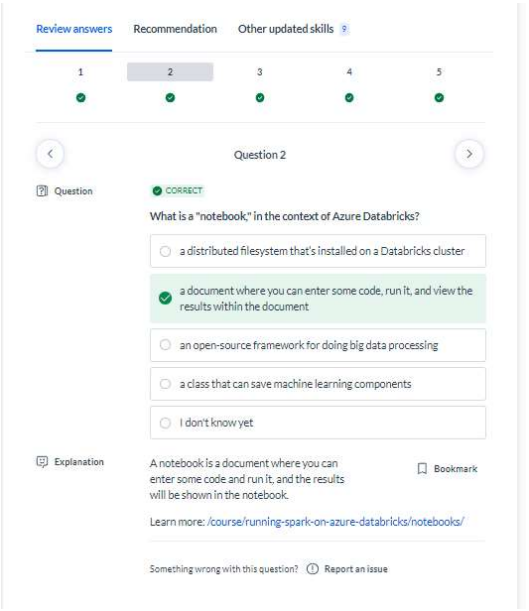

Your API Service
dbml-local
Databricks ML Model Export
Databricks ML Model Export
Data Layer
that the saved model or pipeline can be run


MLeap
MLeap is open source
Like Databricks ML Model Export, MLeap has an execution engine for running pipelines
Your API Service
MLeap Runtime
MLeap Bundle
Data Layer

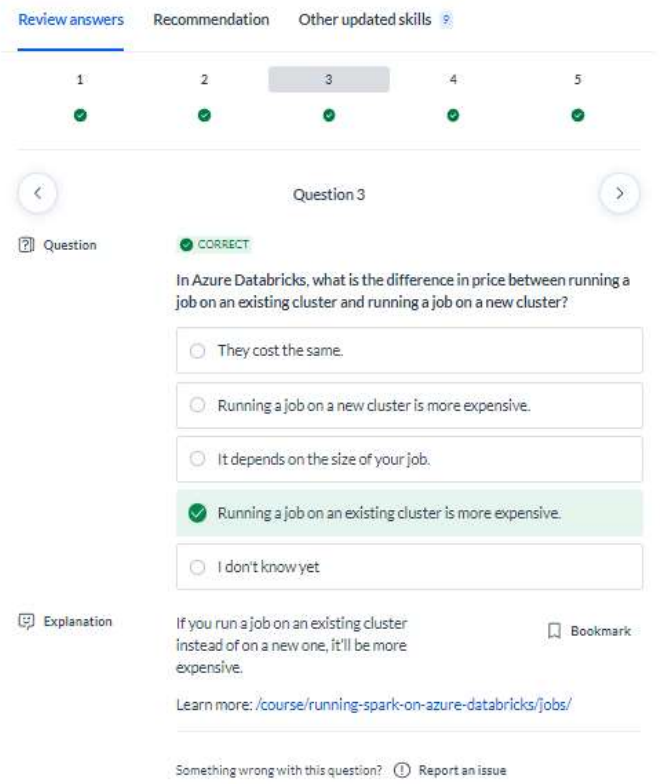

Create a Databricks workspace, cluster, and notebook;
Run code in a Databricks notebook either interactively or as a job;
Train a machine learning model using Databricks;
Deploy a Databricks-trained machine learning model as a prediction service.
To learn more about Azure Databricks,



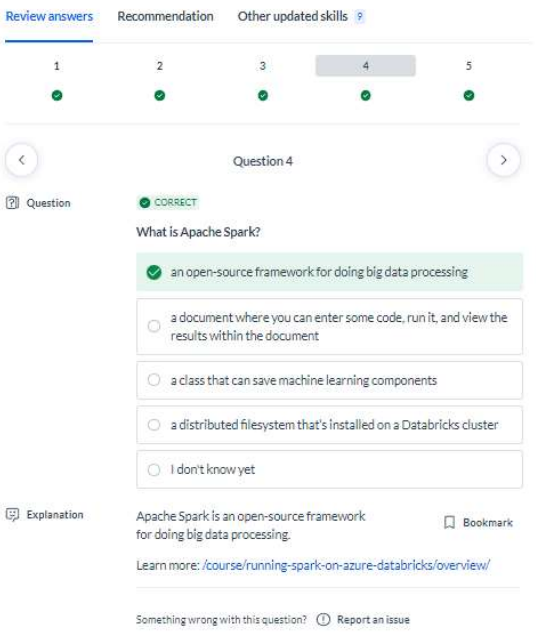
Screen clipping taken: 1/6/2023 7:21 PM



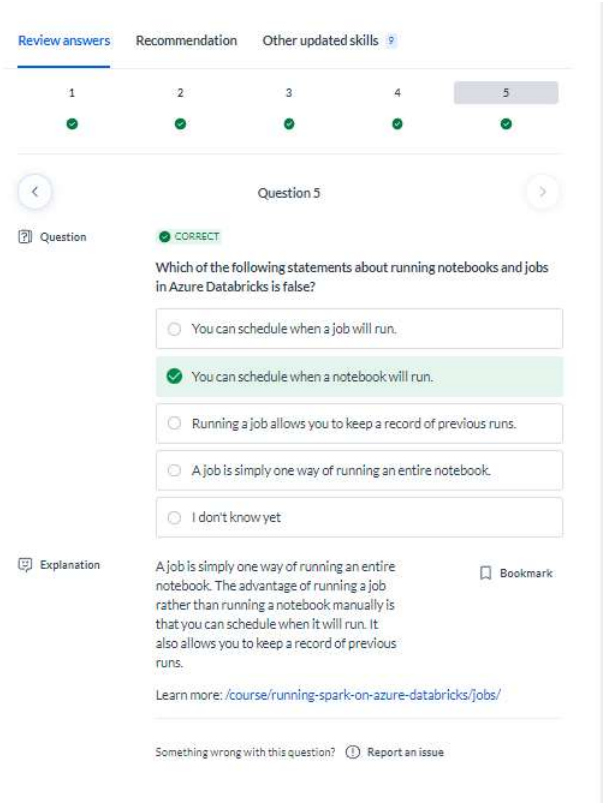
Screen clipping taken: 1/6/2023 7:22 PM



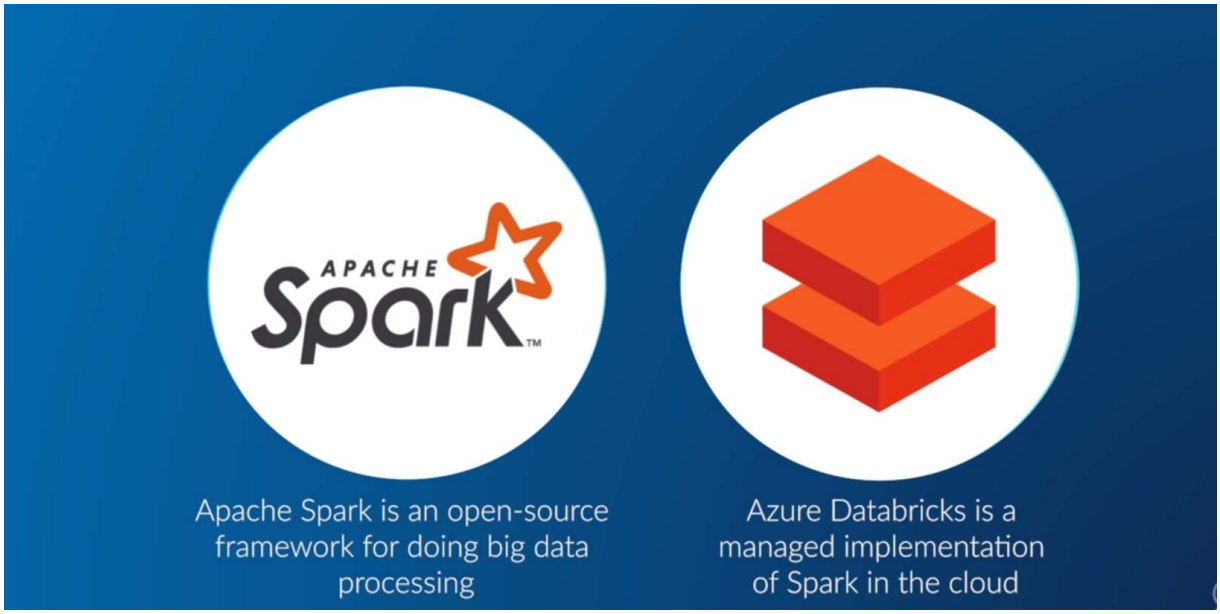
Screen clipping taken: 1/6/2023 7:22 PM



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



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



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 predicition service by running a Mleap Bundle on any system.
That has the Mleap Runtime installed

MLWriter -> Spark Enviroments
Mleap -> Non Spark Enviroments