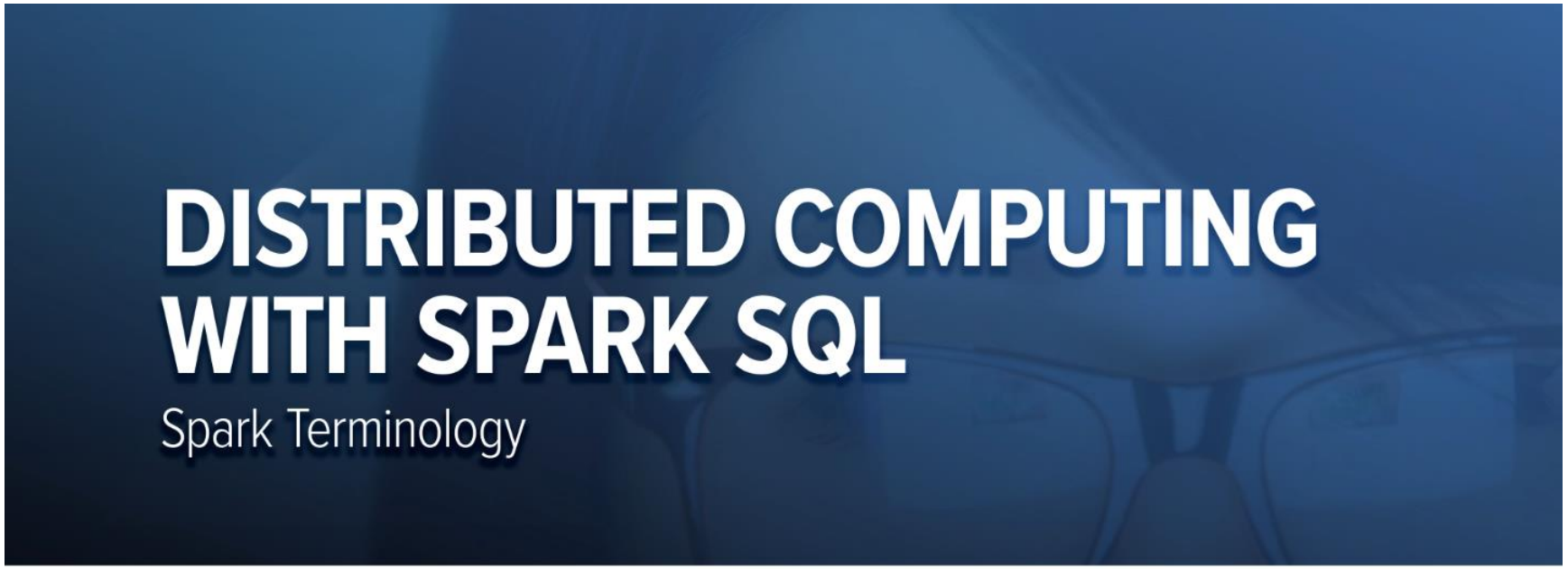
Lesson 2.2: Spark Terminology



Brooke Wenig
Machine Learning Practice Lead
Databricks

UCDAVIS
Continuing and Professional Education

Slide 2: Welcome Back!

Slide 3: Learning Objectives

Learning Objectives

Identify and define Spark terms:

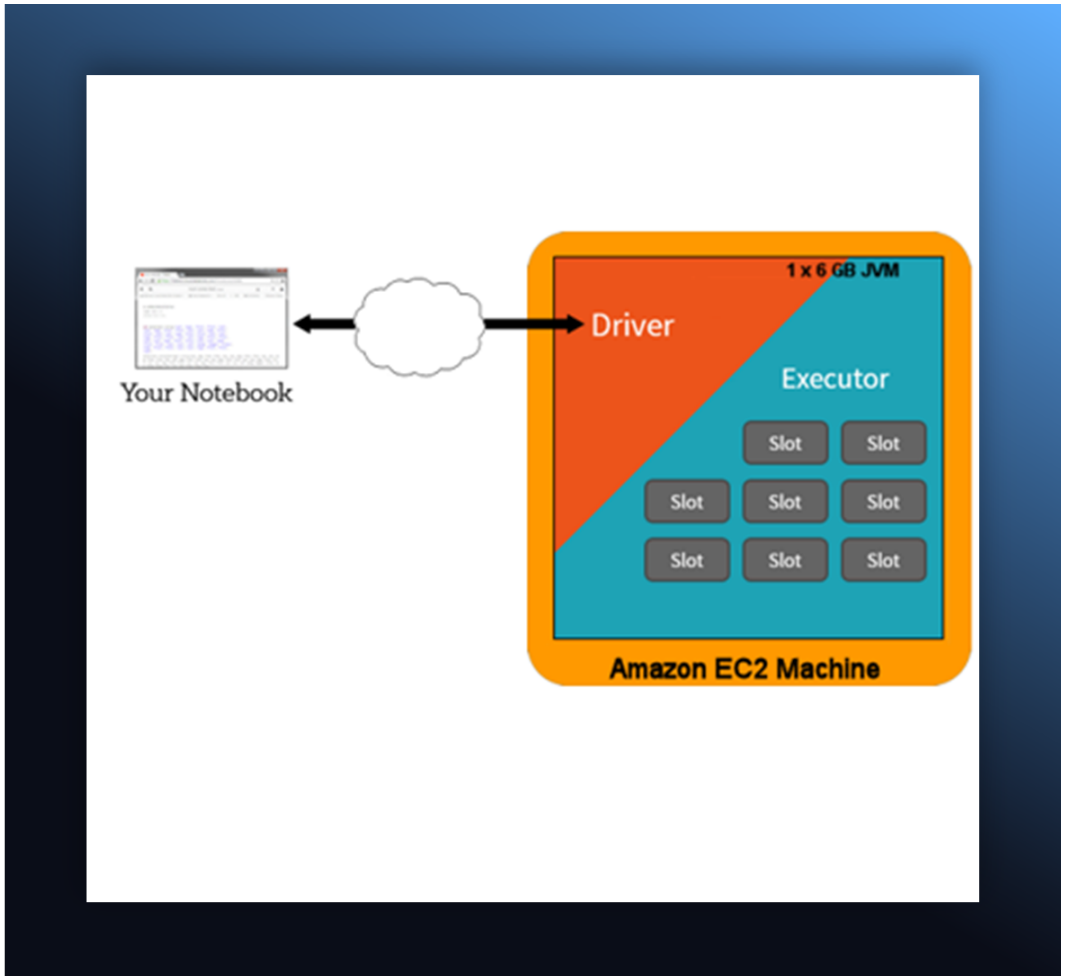
- Partitions
- Slots

Slide 4: Spark Local Mode

Spark Local Mode

Driver and executor are on the same physical machine

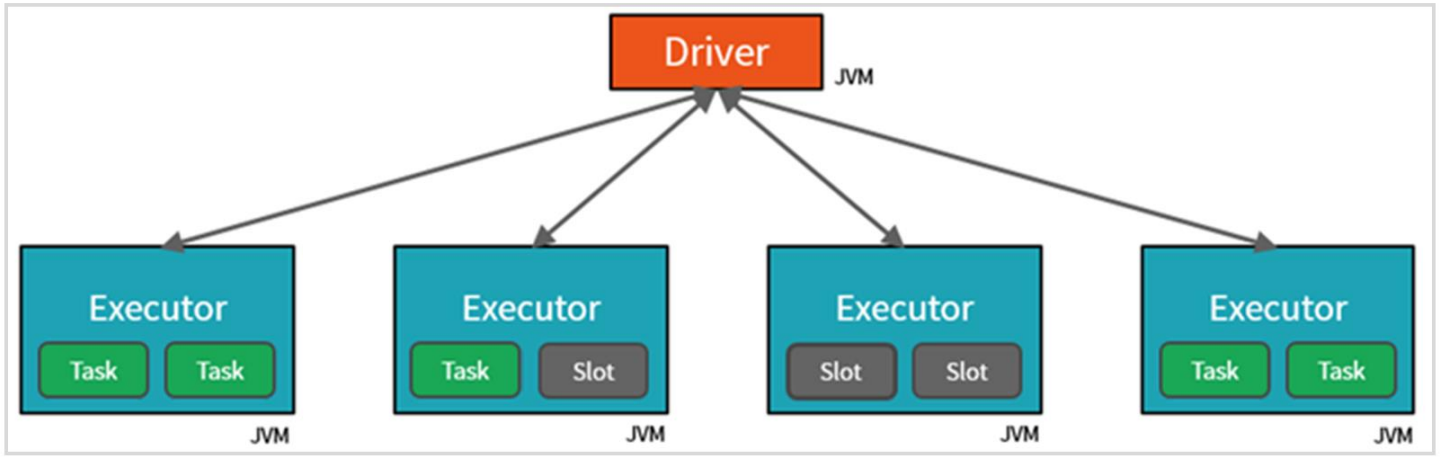
Same architecture for Databricks community edition



Slide 5: Units of Parallelism Within a Spark Cluster

Units of Parallelism within a Spark Cluster

4 executors * 2 slots = 8 units of parallelism



Slide 6: Units of Parallelism

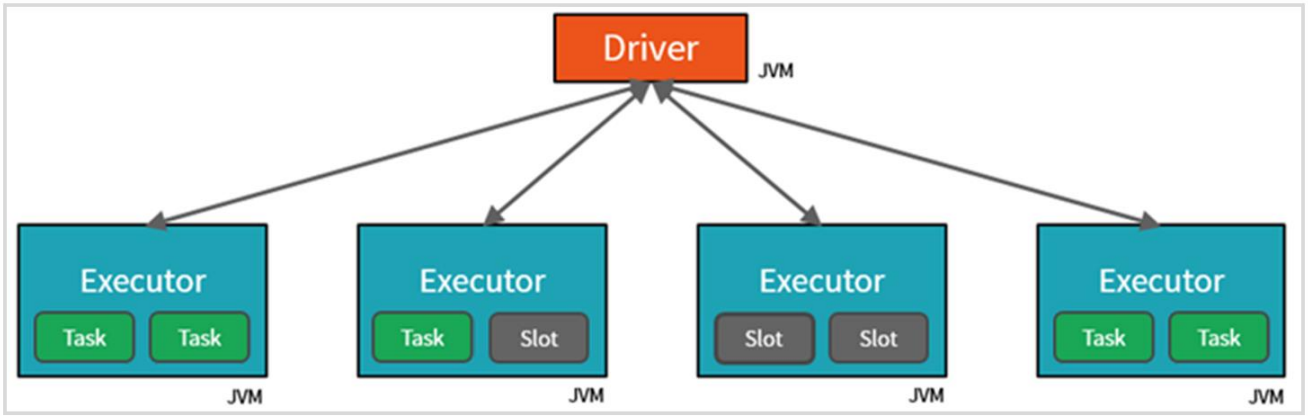
Units of Parallelism

Units of Parallelism Cluster Configuration

MCT: machines * cores * threads

Units of Parallelism for Data are called Partitions

Partitions are part of a large distributed dataset



Slide 7: Determining the Number of Partitions

Determining the Number of Partitions

Size of dataset

- The larger the dataset the more partitions

Underlying partitioning of data by some other feature

Cluster configurations – what happens if I add more partitions?

Slide 8: Grocery Example

Grocery Example

10 friends are to go to the store and pick up 10 items each

Vs.

10 friends are to make 10 trips each to the store and pick up only 1 item each trip

Seeking balance between computation and communication

Slide 9: Coming Up

Coming Up

Caching partitions