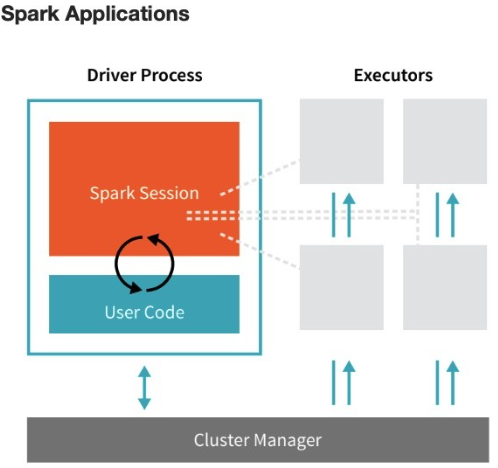
**What is Sparks:**

Spark is a program that runs on distributed systems to combine their resources for processing big data.

**Cluster Manager:**

The cluster of machines that Spark will leverage to execute tasks will be managed by a cluster manager like Spark’s Standalone cluster manager, YARN - Yet Another Resource Negotiator, Kubernetes.

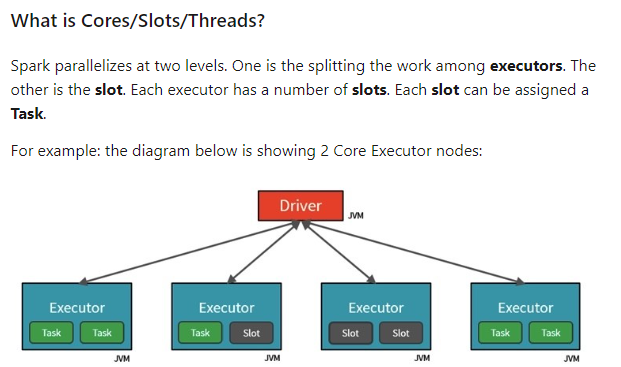


**Driver:**

* Analyse and distribute the work between the executors
* Handles the user input

**Executor:**

* Handles the work driver assigned to them



Then the driver also decides how to partition the data suck that each task gets the partition on which it must work.

**What is a Partition?**

In order to allow every executor to perform work in parallel, Spark breaks up the data into chunks, called partitions.



**DAG Scheduler** is the scheduling layer of Apache Spark that implements stage-oriented scheduling. It transforms a logical execution plan to a physical execution plan (using stages).

**Executor**

Works on JVM

Task:

A diagram of a work flow

Description automatically generatedStage:

A diagram of a stage

Description automatically generated

Job:

A white background with black text

Description automatically generated

Partitions:  
A diagram of a diagram of a different way

Description automatically generated with medium confidence

Narrow partitions start with the same number of partitions and ends with the same number of partitions

In wide partitions, a new stage is generated every time a shuffling happens.

With each shuffle new stage is created below on the LHS each core counts the colour of the partition assigned to them and in the next stage a new core combines the result to get the final output.

Lookup Tables VS Dimension Tables:  
Key Differences:

* Lookup Table:
  + Function: Primarily used for quick value lookups based on a single key, often replacing complex calculations with simple table lookups.
  + Data Structure: Typically small and contains a limited set of key-value pairs.
  + Example: A table mapping product codes to product names.
* Dimension Table:
  + Function: Provides descriptive attributes about data points in a fact table, allowing for filtering and analysis based on characteristics like time, location, or customer details.
  + Data Structure: Contains detailed information about a specific dimension, often with multiple attributes related to that dimension.