Spark on Yarn Architecture

Spark on Yarn Architecture

1. How to execute the spark programs on spark cluster.

Ans:

Interactive mode - spark-shell/pyspark/notebooks
Submitting a job - spark-submit

2. How does spark execute our programs on the cluster.

Ans: Master/Slave architecture

each application has driver which is the master process.

each application has a bunch of executors which are the slave process.

Driver - is responsible for analysing the work, divides the work in many tasks, distributes the task, schedules the task and monitors.

Executor - is responsible to execute the code locally on that JVM (Executor machine)

3. who executes where?

The executors are always launched on the cluster machines (worker nodes)

However, for the driver we have the flexibility to launch it on the client machine or the cluster machines.

client mode
cluster mode

whenever the driver runs on client machine we say as client mode.

whenever the driver runs on cluster or the executor we say as cluster mode.

The preferred approch for production is the cluster mode.

Client mode is not preffered because if the client machine goes down or is shut down then the driver stops.

4. Who controls the cluster & how spark gets the driver and executor

Cluster manager:

- YARN
- Mesos
- Kubernetes
- Spark Standalone

What is a spark session

Is like a datastructure where driver maintains all the information including executor location and status.

This is the entrypoint for any spark application.

Spark on Yarn architecture in client mode

1. when we launch spark-shell automatically spark session is created
2. as soon as spark session is created request goes to the Yarn resource manager.
3. Yarn RM will create a container on one of the node manager and will launch an Application master for this spark application.
4. This application Master will negotiate for resources from the Yarn Resource manager in for of containers.
5. The Yarn RM will create containers on the node managers.
6. Now the APP master will launch the executors in these containers.
7. Now the drivers and executors can communicate directly without the involvement of containers.

Spark on yarn architecture in cluster mode

the only difference here is that the spark driver runs on the application master.