* How to execute Spark Programs?

1. Interactive Clients :- spark-shell, Notebook

2. Submit Job :- spark-submit, Databricks Notebook, Rest API.

- A real-life production implementation is all about packaging your spark application and submitting it to the cluster for execution.
- Spark-submit is most commonly used tool for this method.

However, most of the Spark vendors are going to offer you some other alternatives

Example :- Databricks cloud will allow you to submit Notebook itself, and you do not need to package your application and use the spark-submit tool.

Most of the cloud-based Spark Vendors will allow you to use Rest-APIs. And they internally take care of running the job on the spark cluster.

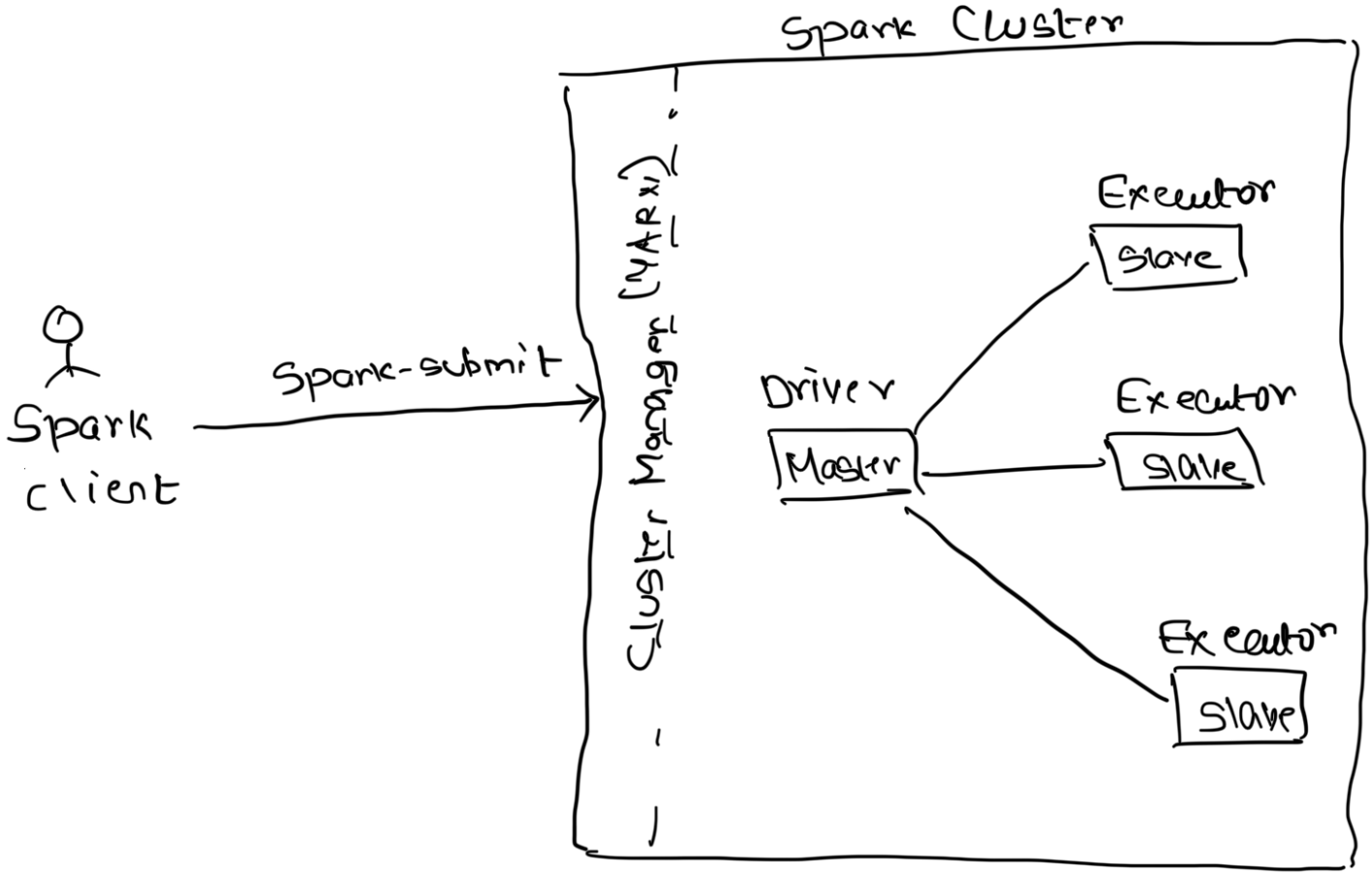
All those method are vendor-specific, but

spark-submit is a universally accepted method and works in almost all cases.

---

* How Spark Distributed Processing Model works?

⇒



-Suppose, you are using spark-submit utility, and you submitted an application A1.
Now, spark engine will request the cluster manager to give a container and start a driver process A1.

Once started, the driver will ask for
some more containers from the cluster
manager and start all executors.

And that's all, Now your driver and executors
are responsible for running your application
code and doing the job that you wanted.

*** Every Spark application applies a
master-slave architecture and runs
independently on the cluster ****

---

* How Spark runs your application on local
machine when we do not have a cluster
and a cluster manager?

⇒ You can execute a spark application on your
local machine without even having a real
cluster.

—In spark.configuration file you state,

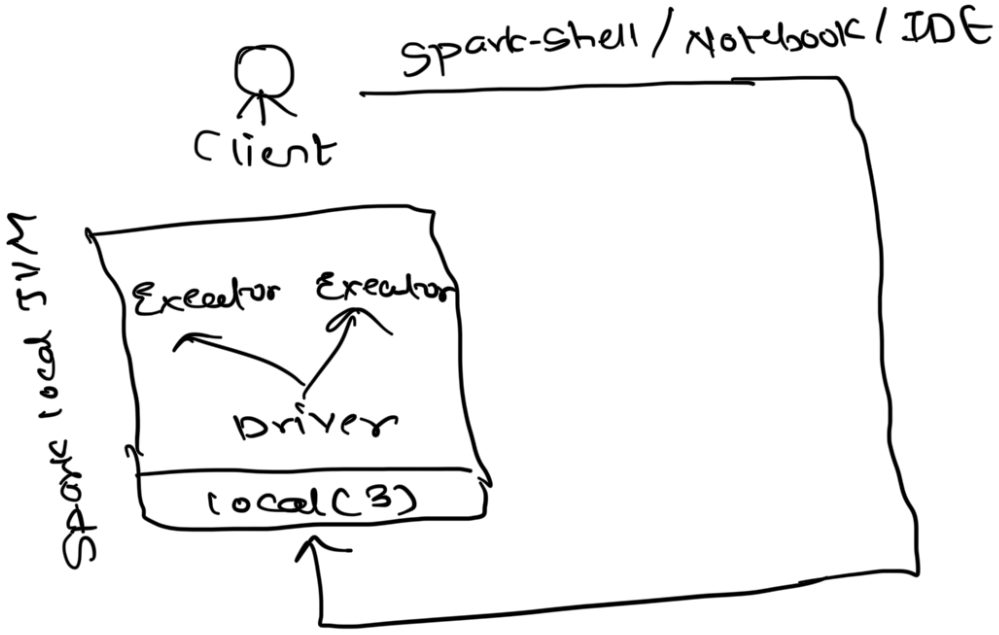
spark.master = local[3]

In this case spark runs locally as a
multi-threaded application. In this case,
I configured it to start 3 threads.

— If you say local and do not give any

number, then it becomes a single-threaded application.

- So when, you configure your application to run with a single local thread, then you will have a driver only and no executors. And, in that case your driver is forced to do everything itself, nothing happens in parallel.
- However, when you run your application with three local threads, then your application will have one driver and two executor.



* How does Spark run with an interactive clients?

⇒ Sparks gives you choice to run your [illegible] of the [illegible] mode:

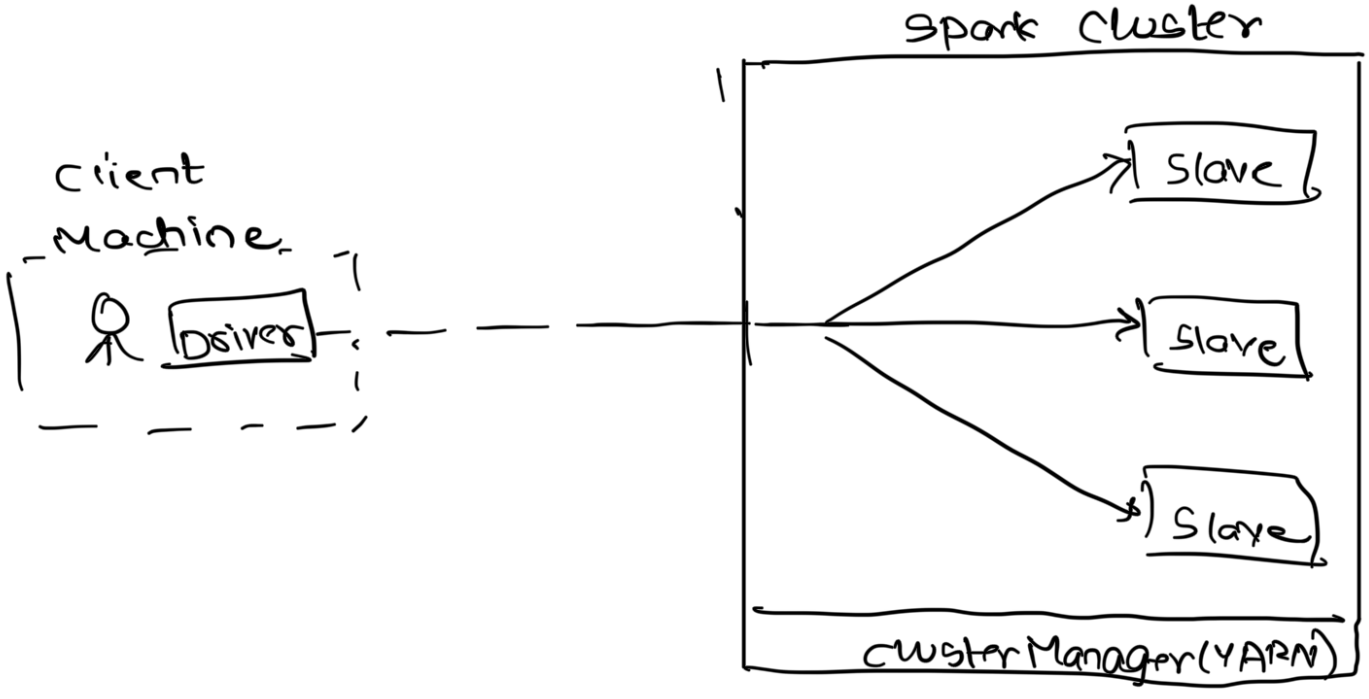
application in one of the

1) Client mode :-

The client mode is designed for interactive clients such as spark-shell and notebooks.

In this mode, the spark driver process runs locally at your client machine.

However, the driver still connects to the cluster manager, and starts all the executors on the cluster.



This is how your spark-shell and Notebooks are working.

However, when your client log-off from your driver dies and

client machine, then your [illegible]
here executors also dies.
So client mode is suitable for interactive work but not for long-running jobs.

2] Cluster mode :-

⇒ The cluster mode is designed to submit your application to the cluster and let it run. In this mode everything runs on the cluster. Your driver as well as the executors.

- It is meant for submitting long-running jobs to the cluster.

---

* When to use what ?

Cluster Manager	Execution Modes	Execution Tools
1. local[n]	Client	IDE, Notebook
2. YARN	Client	Notebook, Shell
3. YARN	Cluster	Spark submit