1. What is a spark

Spark is a general purpose, open source framework. It is used yo solve all bigdata problems

It is used to process batch data, streaming data, machine learning data, graphical data.

1. Alternatives to spark:
   1. Apache flink: 90% streaming data, 10% batch data

(like ola/uber…..any iot….use flink)

* 1. Apache ignite: more powerful than spark and ignite

Not recommended for small amount of data.

* 1. Apache beam: only supports java and Python not sql
  2. Apache spark: 90% batch data ,10% streaming data

(share market last 6 months, today live data use spark)

Flink and ignite are not general purpose technologies.

1. Spark architecture:
   1. Driver node: a server holding /contains (pyspark code) drive program is called driver node.
   2. Cluster manager(resource manager) : Scheduling tasks, allocating resources.
2. How does spark architecture work:
   1. First it will launch driver node and cluster manager
   2. Code is written in driver node.
   3. Driver node first request the cluster manager to allocate resources.
   4. Cluster manager allocate the resources in the form of executers
   5. After launch of executors cluster manager acknowledge the allocated resources
   6. Now driver node Copies Pyspark code to all working nodes(executors)
   7. Spark engine process data within executors
   8. After processing data(result) is transferred to the driver node.
   9. Now driver node consolidates the result and send it to the client node.
   10. Where you are doing spark submit that node is called client node.
3. What is spark context and spark session
   1. Java objects+spark context= rdd api
   2. Java objects+sql context=dataframes api
   3. Java objects+ spark streaming context= dstream api
   4. Java object+unify(sparkcontext, sql context, spark streaming context)=sparksession=dataset api

Sparksession is unified context used to process any kind of data.

Spark contet is used to create rdd api.

Spark context only communicates with cluster manager to allocate the resources, only spark context can do this. Spark context is available with in driver program.

That’s why Spark context is mandatory

1. Why you are using spark:
   1. Spark auto optimize the performance
   2. Easily integrates with other technologies
   3. It is 10-100 times faster.
   4. It minimize the cost as it is open source
   5. Spark is general purpose
   6. And used to process batch as well as streaming data.
2. What is rdd: