All about data engineering

TOP QUESTIONS FOR DATA ENGINEERING INTERVIEWS

Ujjwal Sontakke
Data Engineer

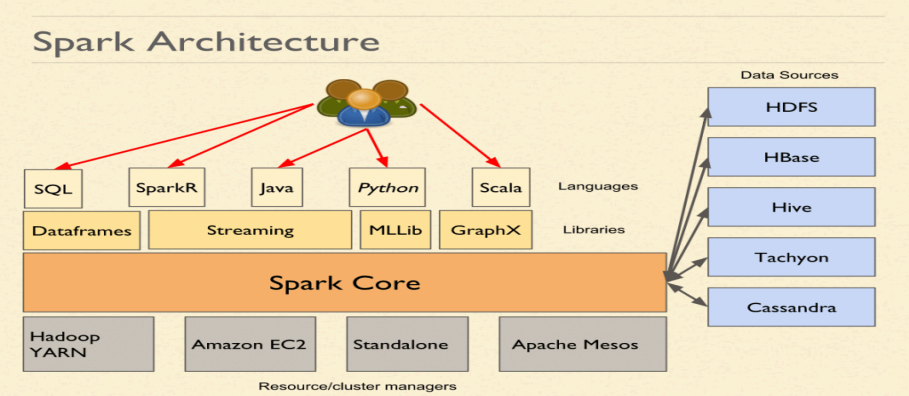
How to decide executors
How to deploy spark cluster on kubernetes

1) How to connect s3 to databricks
2) How to connect aws data catalog to databricks
3) How to submit databrick job
4) How to create job on databricks
5) How to create dependent jobs on databricks
6) How to create pipeline in databricks

1) What is spark?
    - Spark is an open source distributed computing system designed to process large amount of data (structured semi structured and unstructured) in distributed and parallelize manner across the cluster of computers
    - Its ability to process data in memory makes it faster than other big data processing framework which users disc based storage data processing
    - Unified engine for large scale data analytics .
    - Spark also provides a wide range of libraries for Data science ,ML, DL, streaming and graph processing.
2) Difference Between Spark and MR?

points	Spark	MR
Data processing model	In-memory Spark support batch processing,streaming processing and interactive proessing	Disc based storage MR only supports Batch data processing
Speed	Faster than Map reduce(10 to 100x) ,Spark caches the data in memory which makes it faster execution of code and uses DAG execution	Slower in speed . MR used 2 stage processing model for execution (MAP and then Reduce)
Ease of development	Spark provides wide range of high level API for different Language like (python,scala,java,R) this make it easy for development	MR provides only support for JAVA In MR developer needs to write code in Low level for data processing

3) Explain architecture of spark?



4) What is RDD?
   - RDD stands for resilient distributed datasets it is an fundamental Data structure of an apache spark
   - RDD is an distributed collection of objects that can be processed in parallel,RDD can be created from data stored in distributed storage such as HDFS,local file system and NOSQL DBs
   - RDDs are fault tolerant means they can recover from Node failure and this fault tolerance is achieved by Maintaining lineage information .which records the information about the transformation applied to RDD.in the case of node failure the lost information can be received from lineage information.
   - RDD supports two type of operations action and transformations,transformation create new RDD from an existing RDD by applying a function to each element of an RDD
   - On the other hand actions triggers the computation of an RDD and returns a value to the driver program or write the data to external storage

5) Difference Between narrow and wide transformation?
   - Narrow Transformations : narrow transformations are the one which do not require shuffling and data-exchange between the partitions .they can be executed in each partitions independently (MAP FILTER UNION)
   - Wide transformations: wide transformations are the transformations that require data exchange between the partitions and they require shuffling across the network ,which can be expensive operations in terms of time and network bandwidth (group by reduce by join)
   - The main difference between narrow and wide transformations is that narrow transformations can be executed in each partition **independently**, while wide transformations require data exchange between partitions. This means that wide transformations can be slower and more resource-intensive than narrow transformations.
   - It is important to note that the choice of transformation depends on the application requirements and the size of the data being processed. In general, it