Question 1 : Explain Spark Clusters Architecture And Process?

Spark Cluster mainly consist of 2 things that is driver node and executor node
Drvier node that comminuate with executor nodes.
Each of the executor node having slots (execution cores)
Driver node sends the task to the slots on the executor when work has to be done

Driver program access apache spark through sparksession.
In databricks the notebook interface is the driver program.
This driver program creates distributed datasets on the cluster and perform the opertion (transformation & action) to those datasets.

Question 2 : Spark Job Execution process and involved steps?

when any action is called on the RDD, Spark creates the DAG and submits it to the DAG scheduler.
The DAG scheduler splits graph into stages of tasks.
A stage is comprised of tasks based on partitions of the input data.
The Stages are passed on to the Task Scheduler.The task scheduler launches tasks via cluster manager (Spark Standalone/Yarn/Mesos).
The Worker executes the tasks on the Slave.

Job :- Job is a piece of code which takes an input from the source. like hdfs and perform some computation on data and produce the result as a output.
Stage :- Jobs are divided into various stages, stages are classified as a map or reduce stages.
task :- Each stages has some task. one task per partition

Question 3 : Differences between Hadoop Map reduce And Spark?

Hadoop Map Reduce	Spark
Data is stored in disk	Data is stored in Memory
computation is based on disk	Computation is relies on RAM
Fault tolerance is done through replication	Fault Tolerance is done through RDD
Hard to work with real time data	Easy to work with real time data
In comparsion to spark it is less costly	Costly
Only for batch processing	used for interactive query

Question 4 : What are the components of Spark?
Spark is having spark core API in which you can process your data
i) Spark SQL + Dataframes or Structured Data: Spark

SQL
ii) Streaming Analytics: Spark Streaming
iii) Machine Learning: MLlib
iv) Graph Computation: GraphX
v) General Execution: Spark Core

Question 5 : What is Single Node Cluster (Local Mode) in Spark?
Single node cluster is a cluster which is having 1 driver node and no worker node.
In contrast standered cluseter is having alteast 1 worker node.
Single node cluster support all the spark jobs and all the spark data source , including deleta lake

Question 6 : Why RDD resilient?
RDD stands for resilient distributed dataset, resilient means self - auto recover from any failure.
This is also called fault tolerant.
RDD is also having this fault tolerant property which mean it can recover by itself.
Reduntant data plays an important role for data recovery.

RDD lineage: is a graph of all the parent's RDD of RDD.

Question 7 : Difference between persist and cache?

Both the methods are used for performance improvement.
These method save the intermidiate result to reuse it in subsiquent stages.
The only difference is
Cache() -- Save the intermidiate result into memory only.
whereas Persist() - Can Save the result into 5 various storage level
(MEMORY_ONLY, MEMORY_AND_DISK, MEMORY_ONLY_SER, MEMORY_AND_DISK_SER, DISK_ONLY)

Question 8 : What is narrow and wide transformation?
There are 2 type of transformation available which can be applied on RDD.
namely called , narrow and wide transformation.
Narrow transformation does not required data suffle across the partition. example map, filter,
whereas wide transformation required data suffle across the partition. example reduceByKey

Question 9 : Differences between RDD , Dataframe And DataSet?

RDD --
i) Distributed collection of JVM objects,
ii) Funtional Operation (transformatoin and actions)
DataFrame --
i) Distributed collection of row objects
ii) Expression based operation and UDF(user defined funtion)
DataSet --
i) Internally row and externally JVM Object
ii) It takes best out of RDD and dataframe

Question 10 : What are shared variables and it uses?
There are 2 kind of shared variable we have
i) Accumulator variable (we have only single copy on driver machine)
There is a shared copy kept on driver machine.
each executor will update value into this shared variable.
However none of the executor can read the value of accumulator, they can just update/change the value
note:- accumulator is similar to counters in mapreduce

ii) Broadcast variable (we have seperate copy of the variable on each machine)
BV allows the programmer to keeps a read only variable cached on each machine

plays a same role as map side join in hive

Question 11 : how to create UDF in Pyspark?

Question 12 : Explain Stages and Tasks creation in Spark?
Whenever we submit a job , DAG is build, and Spark scheduler create a physical execution plan.
DAG scheduler splits a graph into multiple stages, the stage created based on the transformation.
DAG scheduler then submit the stages into task scheduler.
The number of Task submitted depends on the number of partition available in the textfile.

Question 13 : difference between Coalesce and repartition?
Coalesce :
i) Coalesce reduce the number of

partition in a dataframe.
ii) Coalesce avoid full shuffle, instead of creating a partition it adjust into existing partition.
iii) Coalesce will not trigger partitio

Repartition :
i) Repartition method can be used to increase or decrease the partition in a dataframe.
ii) Repartition is full shuffle operation. which means whole data is taken out from the existing partition and move it to newly created partition.
iii) Repartition triggers shuffling

Question 16 : What happens when use collect() action on DF or RDD?
Don't use collect() on large datasets
if we use collect on large datasets it will collect all the data from all the worker nodes and send it to driver node. it may cause out of memory exception.
Alternate option collect we have is 'take' and 'head'

Question 18 : What is Shuffling?
A shuffle occurs when data is rearranged between partitions.
This is required when a transformation requires information from other partitions,

Question 20 : What types of file format using in big data and those differences?
i) ORC (Optimize row columnar)
ii) Parquet
iii) Avro
iv) csv
v) JSON

Question 22 : Difference between reduceByKey()and groupByKey()?
these transformations operate on pair RDDs.
The pair RDD is an RDD where each element is a pair tuble (key, value).
reduceByKey() transformation is something like grouping + aggregation , OR reduceBykey() equivalent to dataset.group(...).reduce(...).
groupByKey() is just to group your dataset based on a key and send it to other executor for data shuffling.

ReduceByKey is faster in performance compare to groupByKey
ReduceByKey is more efficient than groupByKey

Question 23 : Lazy evaluation in Spark and its benefits?
Lazy Evaluation:
1. Laziness means not computing transformation till it’s need
2. Once, any action is performed then the actual computation starts
3. A DAG (Directed acyclic graph) will be created for the tasks
4. Catalyst Engine is used to optimize the tasks & queries
5. It helps reduce the number of passes

Question 24 : What is Catalyst Optimizer And Explain End to End Process?
Catalyst optimizer is a program which will help in generating equivalent RDD code for dataframe programing.
inorder to enhance the performance of the application

****** NC

Question 25 : Difference between ShuffledHashJoin And BroadcastHashjoin?

Question 26 : How many modes are there for spark execution?
i) Local
-- the _master & _worker run on same machine.
ii) Standalone Scheduler
-- As the signifies standalone cluster is a cluster with only spark specific components ,
-- It doesn't have any dependency on hadoop components
-- Spark driver act as a cluster manager
iii) YARN or iv) Mesos
-- Apache Spark runs on Mesos or YARN without any root-access or pre-installation. It integrates Spark on top Hadoop stack that is already present on the system

Project explanation
utilities
1) Library
2) Funtions
3) pySpark Scehma
4) Setup
5) database ddl scripts
6) Validation
7) NB_sales_landing_staging
8) NB_sales_staging_curated

9) NB_sales_curation_dwh
9) NB_Final_manual_run (community eddition)
11) NB_sales_load_dynamic (used to call all the notebook