[What are RDD Operations in Spark?](https://www.zeolearn.com/interview-questions/spark#collapse-beginner-1408)

There are two types of operations in spark, a transformation operation and an Action. Transformations create new RDDs by transforming old RDDs in some way. An action is when you work with the actual RDD and apply some sort of function.

What do you understand by Lazy Evaluation?

Lazy evaluation in Spark means that the execution will not start until an action is triggered.

What is a DAG in spark?

DAG stands for Directed Acyclic Graph. It is a set of Vertices and Edges, where vertices represent the RDDs and the edges represent the Operation to be applied on RDD.

What is the role of a spark Driver?

In Spark DAG, every edge directs from earlier to later in the sequence. On the calling of Action, the created DAG submits to DAG Scheduler which further splits the graph into the stages of the task. Apache Spark DAG allows the user to dive into the stage and expand on detail on any stage.

What is Shuffling in Spark?

Shuffling is simply the process of redistributing data across partitions that may or may not cause the data to move around nodes. Any join in Spark will create shuffling.

What are the deploy modes in Spark?

Deploy modes in spark simply specify where the program will be run first. It can run on a worker node inside the cluster or on an external client.

What is the difference between RDD and a data frame?

RDD are Resilient Distributed Datasets are read only immutable partitions of data. It’s the fundamental data structure in spark.

Data frames are an abstraction on RDD. They are organized into named columns and rows equivalent to a table. This allows the user to impose a structure onto the RDDs. RDD have no optimization engine and they cannot take advantage of sparks optimizers whereas Data frames are able to take full advantage of these.

How does spark achieve fault tolerance?

Fault refers to failure of a data node, thus fault tolerance in Spark is the capability to operate and to recover loss after a fault occurs. Spark achieves fault tolerance by operating on data in fault-tolerant file systems like HDFS. HDFS creates copies on blocks of data in different data nodes so if one data node fails, all the data is in different blocks on different data nodes. Since Apache Spark RDD is an immutable dataset, each Spark RDD remembers the lineage of the deterministic operation that was used on fault-tolerant input dataset to create it. If due to a worker node failure any partition of an RDD is lost, then that partition can be re-computed from the original fault-tolerant dataset using the lineage of operations.