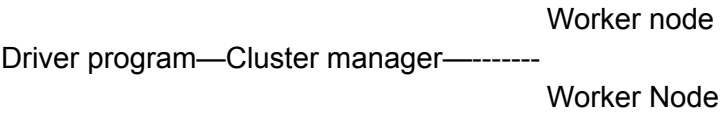
1.. Explain Architecture of Spark ?

Spark is master slave architecture consists of single master and multiple slave.
It depends on 2 abstraction

- Directed acyclic graph
- Resilient distributed dataset



Driver program runs main() function and creates sparkcontext object which coordinate spark application and runs independent set of process on cluster.

Cluster manager consists of various types of cluster managers like hadoop yarn apache etc, it allocates resource of application.

Worker node is salve node, it runs the apps code in cluster.

2. Difference between Hadoop and spark

Compared to spark, hadoop is cheap. Spark has built in ML libraries whereas hadoop integrates with external libraries.
Hadoop stores and process the data in external storage, in spark it stores in internal storage.

3. Difference between RDD, Dataframe, Dataset

RDD is slower compare to other two. Dataset is faster than RDD and little slower than datarames..
In RDD schemas need to defines manually, dataframe and dataset have auto discovery of file schemas.

4. Explain the similarities in all API of Spark

5. What is Transformation? Explain in detail

It is also know as feature engineering which creates new feature rom existing that may help in improving the model performance
It can also be used for feature reduction. It can be done in many ways by linear combination of original features or by using non linear function..

6. What is Actions in spark? Explain in detail

Action in spark is any operation that does not return RDD. their operations which trigger a spark to compute and return a result to the spark driver program or write data to external storage..

7. What is the Wide Transformation ?, explain with example

No need of shuffling data across nodes. Transformation involves data movement between partitions are known as wide transformation. Function such as groupBykey, aggregateBykey, join, repartition are some example of wide transformation.

8. What is Narrow Transformation? Explain with example

We need to shuffle data across nodes. Operations where each input partition of an RDD is used to compute only one output partition of resulting RDD. It includes map, filter, union.

9. Write down the query of wide n narrow transformation with example?

```
df_sal = df.filter("Salary > 6000")
df_ordered = df.orderBy("Salary")
```

10. Explain Kerberos Architecture

Main components of kerberos are

- Authentication server - performs the initial authentication and ticket granting service
- Database - verifies the access rights of users in the database.
- Ticket grating server - issues ticket for server.