1.what is cluster node size

42 nodes

2.waht is the driver program

The spark driver is the program that declares the transformations and actions on RDDs of data and submits such requests to the master.

In practical terms, the driver is the program that creates the SparkContext, connecting to a given Spark Master. In the case of a local cluster, like is your case, the master\_url=spark://<host>:<port>

Its location is independent of the master/slaves. You could co-located with the master or run it from another node. The only requirement is that it must be in a network addressable from the Spark Workers.

* The driver prepares the context and declares the operations on the data using RDD transformations and actions.
* The driver submits the serialized RDD graph to the master. The master creates tasks out of it and submits them to the workers for execution. It coordinates the different job stages.
* The workers is where the tasks are actually executed. They should have the resources and network connectivity required to execute the operations requested on the RDDs.
* Driver program is responsible for launching various parallel operations on the cluster.
* Driver program contains application's *main()* function.
* It is the process which is running the user code which in turn create the SparkContext object, [**create RDDs**](http://data-flair.training/blogs/how-to-create-rdds-in-apache-spark/) and performs [**transformation and action operation on RDD**](http://data-flair.training/blogs/rdd-transformations-actions-apis-apache-spark/).
* Driver program access [**Apache Spark**](http://data-flair.training/blogs/apache-spark-introduction-spark-comprehensive-tutorial/) through a [**SparkContext**](http://data-flair.training/blogs/sparkcontext-in-apache-spark-tutorial/) object which represents a connection to computing cluster (From Spark 2.0 onwards we can access SparkContext object through SparkSession).
* Driver program is responsible for converting user program into the unit of physical execution called task.
* It also defines distributed datasets on the cluster and we can apply different operations on Dataset (transformation and action).
* Spark program creates a logical plan called [**Directed Acyclic graph**](http://data-flair.training/blogs/directed-acyclic-graph-dag-in-apache-spark/) which is converted to physical execution plan by the driver when driver program runs.

3. what is the broad cost variable

A broadcast variable. Broadcast variables allow the programmer to keep a read-only variable cached on each machine rather than shipping a copy of it with tasks. They can be used, for example, to give every node a copy of a large input dataset in an efficient manner. Spark also attempts to distribute broadcast variables using efficient broadcast algorithms to reduce communication cost.

Broadcast variables are created from a variable v by calling [SparkContext.broadcast(T, scala.reflect.ClassTag<T>)](https://spark.apache.org/docs/1.5.1/api/java/org/apache/spark/SparkContext.html#broadcast(T,%20scala.reflect.ClassTag)). The broadcast variable is a wrapper around v, and its value can be accessed by calling the value method. The interpreter session below shows this:

they're variables that we want to share throughout our cluster. However there are a couple of conditions that are important to understand. Broadcast variables have to be able to fit in memory on one machine. That means that they definitely should NOT be anything super large, like a large table or massive vector. Secondly, broadcast variables are immutable, meaning that they cannot be changed later on. This may seem inconvenient but it truly suits their use case. If you need something that can change, I'd certainly point you to accumulators which will be covered in another post. So now we know that broadcast variables are:

* Immutable
* Distributed to the cluster
* Fit in memory

4.what is your cluster manager

YARN

5.spark life cycle

6.how much exp in scala

7.what is the case class

Notice how the keyword new was not used to instantiate the Book case class. This is because case classes have an apply method by default which takes care of object construction.

Its used to create schema for datasets

9.how you get data

from rdbms Oracle Database

10. how you learn scala

through Internal training in the company