Spark :

### What is spark ?

It is a distributed computing engine which distributes our data to process it.

## Spark Architecture :

Worker Node

Executer Cashe

Task

Task

Executer Cashe

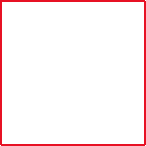
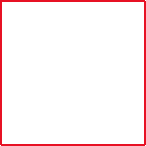
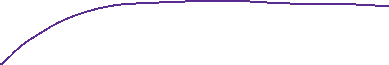
Worker Node

Cluster

Manager

Spark context

Driver Program



# Employee table notes :

Creating employee dataframe from employee csv file.

emp = spark.read.format('csv').option('inferSchema',True).option('header', True).load('/FileStore/tables/employees.csv')

1. Selecting only two columns ,

emp.select('EmployeeID', 'FirstName').display()

1. Selecting two columns with orderby clause

emp.select('EmployeeID', 'FirstName').orderBy(['EmployeeID'], ascending=[False]).display()