Report For Spark Lab:

In this lab we need to count the JPGs request in the weblog files.There are many ways to write down the code of this project,I have used scala and python to write down the code of this problem.

For Python:

The python file is named as CountJPGs.py,in this python file I have written the code for counting the number of JPGs request.

Steps:

- For running spark in python mode we need to run a command as **pyspark**
- Now we need to setup the spark context by running the command as sc=SparkContext()
- Now at the end we need to submit the code file by using the command by using this command:
  spark-submit CountJPGs.py

For Scala:

Most of the spark code is always written in scala and the spark itself is written in scala.

Steps:

- We need to setup the sparkcontext for running into scala mode.

  val sc = new SparkContext()

- We need to create the package by using the command as mvn package and it will make a jar file named as countjpgs.jar.We need to run the following command to submit the jar file.

For submitting spark application to the yarn cluster we need to run the command and these command will be different for python and scala.

For python:

spark-submit
--master yarn –client
CountJPGs.py

For Scala:

spark-submit
--class CountJPGs.py
--master yarn-client

- We also need to run the url http://localhost:8088 for visiting the yarn and notedown the application id.

Configure a Spark Application:

- We need to rerun the code files which was written in python or scala by running the same commands of spark submit.
- We need to visit the resource manager UI again and need to notedown the application name listed in one specified command line
- We will make a configuration file named as myspark.conf which will contains all the setting for the configuration.
- After all this process again we need to rerun the code and open the YARN UI where application name will be displayed
- At the end we will setup the logging levels.