CIS 4930 Introduction to Hadoop and Big Data

Write and Run a Spark Application and Configure a Spark Application

Julie Reyes | u76631122

Introduction:

The purpose of this lab is to write a self-contained Python Spark application to submit to the cluster. This lab differs from previous labs, in that an RDD will be created via a python application, rather than utilizing the interactive spark application. The program for this lab is a simple python program that counts the number of JPG requests in a web log file on the Hadoop cluster.

Lab:

A python application was created with the text editor gedit**.** The following is the code used to create the application named CountJPGs.py:

```
import sys
from operator import add

from pyspark import SparkContext

if __name__ == "__main__":
    if len(sys.argv) != 2:
        print >> sys.stderr, "Usage: countJPGs <file>"
        exit(-1)
    sc = SparkContext(appName="CountJPGs")
    logRDD = sc.textFile(sys.argv[1], 1)
    output = logRDD.filter(lambda line: ".jpg" in line).count()
    print output
sc.stop()
```

The CountJPGs.py program performs in a similar manner to the previous assignment as it counts the number of JPGs in the weblog file. In the CountJPGs.py program, a logRDD is created from the file at the path entered in the command line. The output is calculated by filtering the logRDD for lines that contain jpg and counting the total amount of JPG lines, which was completed in the line that states:

```
output = logRDD.filter(lambda line: ".jpg" in line).count()
```

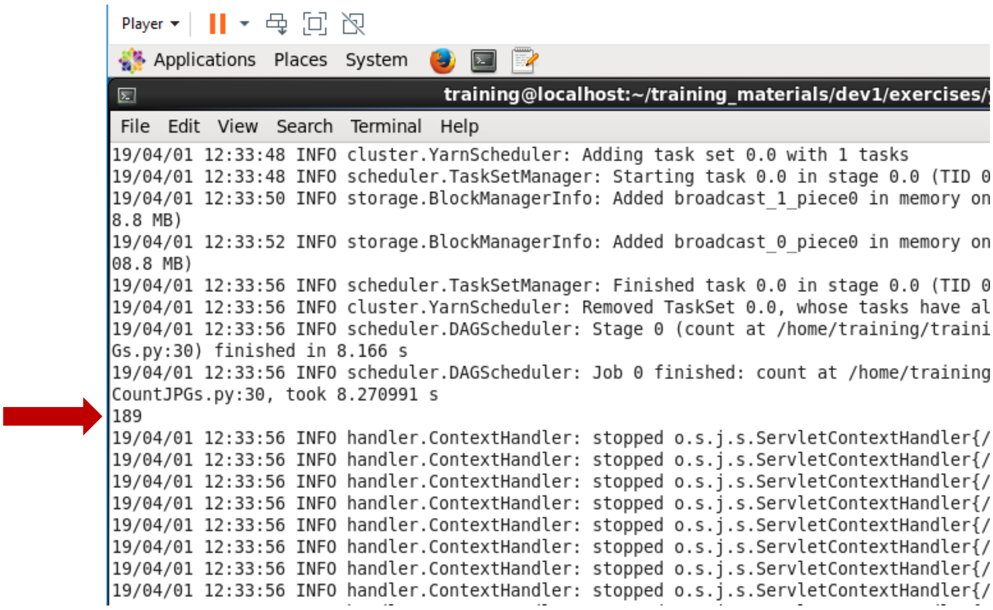
Then, the final total count is printed to the terminal by the line:

```
print output
```

Lastly, the program stops the spark context to terminate the application with the sc.stop().To submit the application the following command line was entered into the terminal:

spark-submit CountJPGs.py /loudacre/weblogs/*

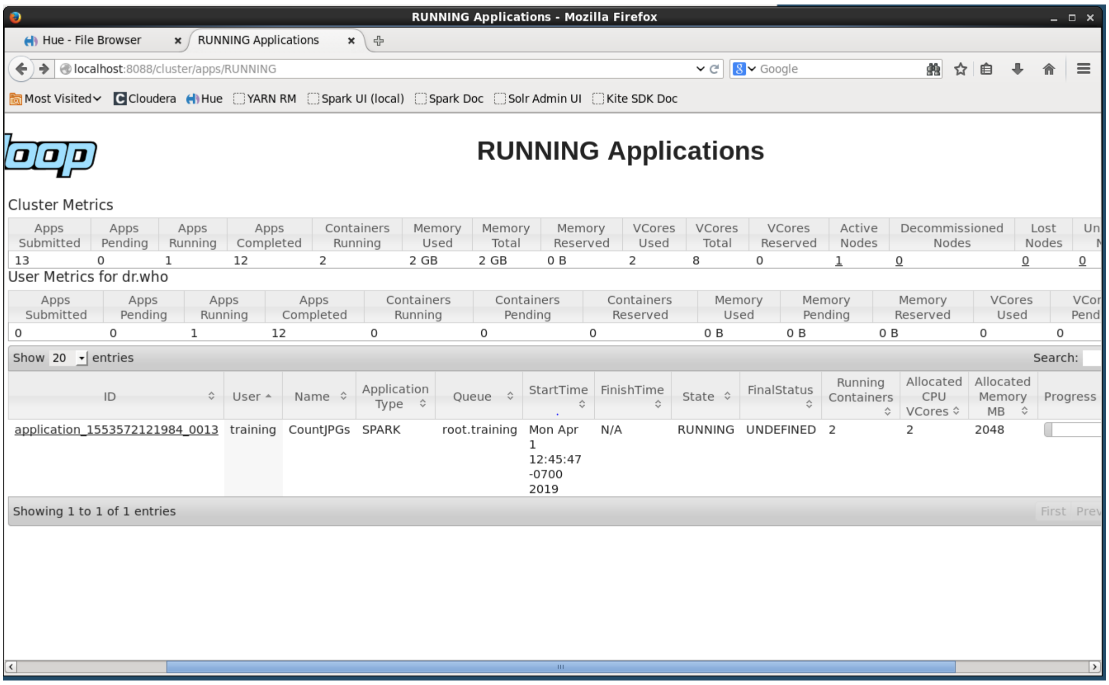
This job ran locally produced the following output:


Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/
File Edit View Search Terminal Help

```
19/04/01 12:33:48 INFO cluster.YarnScheduler: Adding task set 0.0 with 1 tasks
19/04/01 12:33:48 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0
19/04/01 12:33:50 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on
8.8 MB)
19/04/01 12:33:52 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on
08.8 MB)
19/04/01 12:33:56 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0
19/04/01 12:33:56 INFO cluster.YarnScheduler: Removed TaskSet 0.0, whose tasks have al
19/04/01 12:33:56 INFO scheduler.DAGScheduler: Stage 0 (count at /home/training/traini
Gs.py:30) finished in 8.166 s
19/04/01 12:33:56 INFO scheduler.DAGScheduler: Job 0 finished: count at /home/training
CountJPGs.py:30, took 8.270991 s
189
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
19/04/01 12:33:56 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/
```

Therefore, there were a total of 189 webpages that were of type JPG in the weblogs directory.

Next, the job was submitted to the cluster and tracked via the yarn resource manager UI. The following is a snapshot of the job being ran on the cluster and tracked through the Yarn resource manager:



RUNNING Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Un N
13	0	1	12	2	2 GB	2 GB	0 B	2	8	0	1	0	0	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used	VCor Pend
0	0	1	12	0	0	0	0 B	0 B	0 B	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Progress
application_1553572121984_0013	training	CountJPGs	SPARK	root.training	Mon Apr 1 12:45:47 -0700 2019	N/A	RUNNING	UNDEFINED	2	2	2048	

Showing 1 to 1 of 1 entries First Prev

You can see that the application 1553572121984_0013 is running when the job is submitted to the cluster. To submit the application to the cluster the following command was used:

```
spark-submit    --master yarn-client    CountJPGs.py /loudacre/weblogs/*
```