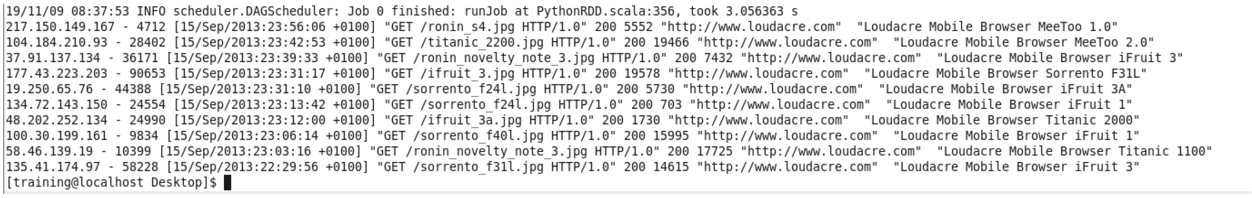
Benjamin Morrison
Assignment 6
Hadoop
Nov. 10, 2019

Write and run a Spark application

This lab is a continuation of the previous lab where we learned how to create and manipulate RDDs. Here, we are taking those principals, putting them into a singular file (in this case, python) and running them on the cluster. Below is an explanation of my .python file and what it does.

The purpose of this assignment is to write a file that counts the number of JPG requests in the weblog files in the hadoop filesystem from the previous lab and print those results. First, we import '*sys*' which allows us to pass in the file path arguments that the lab instructs us to do. Next, we set up spark context by using '*sc = SparkContext()*". Using string concatenation, we add argv[1] to the string "hdfs:" and assign it to the variable '*filepath*'. From here we create an RDD from the logfile which has been assignmed the filepath from the variable mentioned above. From here, we create a new RDD from the results of searching for instances of ".jpg" in the previous RDD. Lastly, we can use a simple for loop to print out the first 10 instances of the results from the most recent RDD.

Below is a screenshot of the final output:



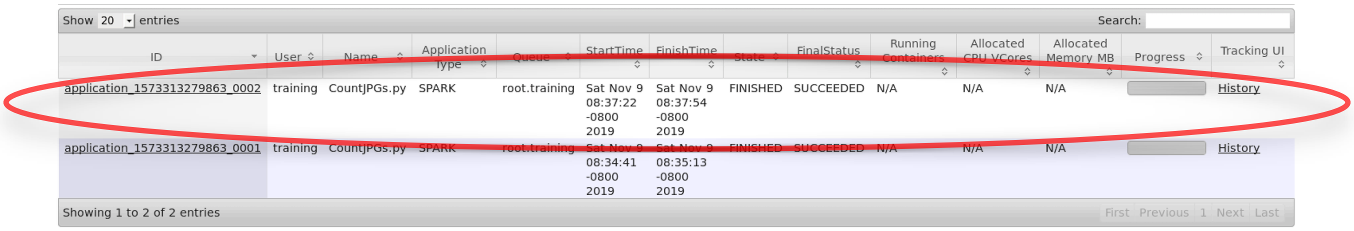
```
19/11/09 08:37:53 INFO scheduler.DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:356, took 3.056363 s
217.150.149.167 - 4712 [15/Sep/2013:23:56:06 +0100] "GET /ronin_s4.jpg HTTP/1.0" 200 5552 "http://www.loudacre.com"  "Loudacre Mobile Browser MeeToo 1.0"
104.184.210.93 - 28402 [15/Sep/2013:23:42:53 +0100] "GET /titanic_2200.jpg HTTP/1.0" 200 19466 "http://www.loudacre.com"  "Loudacre Mobile Browser MeeToo 2.0"
37.91.137.134 - 36171 [15/Sep/2013:23:39:33 +0100] "GET /ronin_novelty_note_3.jpg HTTP/1.0" 200 7432 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 3"
177.43.223.203 - 90653 [15/Sep/2013:23:31:17 +0100] "GET /ifruit_3.jpg HTTP/1.0" 200 19578 "http://www.loudacre.com"  "Loudacre Mobile Browser Sorrento F31L"
19.250.65.76 - 44388 [15/Sep/2013:23:31:10 +0100] "GET /sorrento_f24l.jpg HTTP/1.0" 200 5730 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 3A"
134.72.143.150 - 24554 [15/Sep/2013:23:13:42 +0100] "GET /sorrento_f24l.jpg HTTP/1.0" 200 703 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 1"
48.202.252.134 - 24990 [15/Sep/2013:23:12:00 +0100] "GET /ifruit_3a.jpg HTTP/1.0" 200 1730 "http://www.loudacre.com"  "Loudacre Mobile Browser Titanic 2000"
100.30.199.161 - 9834 [15/Sep/2013:23:06:14 +0100] "GET /sorrento_f40l.jpg HTTP/1.0" 200 15995 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 1"
58.46.139.19 - 10399 [15/Sep/2013:23:03:16 +0100] "GET /ronin_novelty_note_3.jpg HTTP/1.0" 200 17725 "http://www.loudacre.com"  "Loudacre Mobile Browser Titanic 1100"
135.41.174.97 - 58228 [15/Sep/2013:22:29:56 +0100] "GET /sorrento_f31l.jpg HTTP/1.0" 200 14615 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 3"
[training@localhost Desktop]$ 
```

This is a list of the first ten JPG requests pulled from the RDD.

This result was given after submitting the application to the cluster using the command below:

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

We can track the progress of the job created from the above command by navigating to the URL http://localhost:8088 from the VM. Below is a screenshot from this address that shows the process from the above screenshot after it has finished.



ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Progress	Tracking UI
application_1573313279863_0002	training	CountJPGs.py	SPARK	root.training	Sat Nov 9 08:37:22 -0800 2019	Sat Nov 9 08:37:54 -0800 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1573313279863_0001	training	CountJPGs.py	SPARK	root.training	Sat Nov 9 08:34:41 -0800 2019	Sat Nov 9 08:35:13 -0800 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A		History

Showing 1 to 2 of 2 entries