


Robin Jha
1.125
Deliverable #2

For this project I had to gather the data myself for which I used ~50 node Mesos cluster. On the cluster I have agents running to collect the current disk used percentage, which is then written to Elasticsearch as shown in the screenshot below (extra fields have been covered for security reasons):



Once the agents started pushing data, I wrote a module in Java to get this data from Elasticsearch and do some processing on it. I then compiled this module into a jar and submitted this as a job to the Spark cluster that was setup as a part of deliverable #1. Here is the UI for Spark Master (some fields have been covered for security reasons):

 **Spark Master** at spark://[REDACTED]

URL: spark://[REDACTED]
REST URL: spark://[REDACTED] (cluster mode)
Alive Workers: 5
Cores in use: 80 Total, 0 Used
Memory in use: 160.0 GB Total, 0.0 B Used
Applications: 0 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-201611[REDACTED]	[REDACTED]	ALIVE	16 (0 Used)	32.0 GB (0.0 B Used)
worker-201611[REDACTED]	[REDACTED]	ALIVE	16 (0 Used)	32.0 GB (0.0 B Used)
worker-201611[REDACTED]	[REDACTED]	ALIVE	16 (0 Used)	32.0 GB (0.0 B Used)
worker-201611[REDACTED]	[REDACTED]	ALIVE	16 (0 Used)	32.0 GB (0.0 B Used)
worker-201611[REDACTED]	[REDACTED]	ALIVE	16 (0 Used)	32.0 GB (0.0 B Used)

Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----------------	------	-------	-----------------	----------------	------	-------	----------

Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20161114053155-0001	[REDACTED]	80	32.0 GB	2016/11/14 05:31:55	root	FINISHED	1.6 min
app-20161110055121-0000	[REDACTED]	80	32.0 GB	2016/11/10 05:51:21	root	FINISHED	1.6 min

To submit the Spark job, I logged into the container for master and ran the command below:

```
./spark-submit --class com.ml.pie.run.PieExecutor  
/mnt/mesos/sandbox/pie-1.0-SNAPSHOT-jar-with-  
dependencies.jar DiskUsagePredictor
```

Jobs currently in progress is listed under “Running Applications” on the Spark UI.

Currently the output for my Spark job looks as follows. I am able to get the IP and disk usage on the host. The next step is to use Machine Learning techniques on this dataset:

```
Running PIE Job for predicting : Disk Usage

2016-11-17 18:04:48 INFO DiskUsageStats:183 - processing Host IP: 100.200.100.50 Disk Name : /
2016-11-17 18:04:49 INFO DiskUsageStats:122 - ++++++
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr
2016-11-17 18:04:48 INFO DiskUsageStats:183 - processing Host IP: 100.200.101.50 Disk Name : /
2016-11-17 18:04:49 INFO DiskUsageStats:122 - ++++++
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr
2016-11-17 18:04:49 INFO DiskUsageStats:183 - processing Host IP: 100.200.102.50 Disk Name : /
2016-11-17 18:04:49 INFO DiskUsageStats:122 - ++++++
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:49 INFO ESHistoryProcessor:116 - Disk Name is :/usr
2016-11-17 18:04:50 INFO DiskUsageStats:183 - processing Host IP: 100.200.103.50 Disk Name : /
2016-11-17 18:04:50 INFO DiskUsageStats:122 - ++++++
2016-11-17 18:04:50 INFO ESHistoryProcessor:116 - Disk Name is :/
2016-11-17 18:04:50 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:50 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:50 INFO ESHistoryProcessor:116 - Disk Name is :/usr
2016-11-17 18:04:50 INFO ESHistoryProcessor:116 - Disk Name is :/boot
2016-11-17 18:04:50 INFO ESHistoryProcessor:116 - Disk Name is :/usr/share/oem
2016-11-17 18:04:50 INFO ESHistoryProcessor:116 - Disk Name is :/usr
```