TFIDF Report

Big Data Platforms Lab 1

Hippolyte Jacomet
Karl Bou Abboud

OUR APPROACH

We followed the same approach as suggested, with three successive mapreduce jobs. The file TfIdf.java contains the main method for the mapreduce. SortResults is separate and allows to process and sort the results.
For the tfidf, we used a trick found on stackoverflow to pass the number of input documents to the final reducer through the name of the job. This is why the name of job3 ends up being "2", it is an ugly hack but unfortunately the only solution we found so far.

RESULTS

Using only 1 mapper and 1 reducer.

Using several reducers would require concatenating the different outputs in a single file to sort the results.

Top 20 words:

buck, callwild : 0.007727154027237492
dogs, callwild : 0.002532789375594511
thornton, callwild : 0.002189360307717289
myself, defoe-robinson-103.txt : 0.0016590896005008772
spitz, callwild : 0.0013951805882512138
sled, callwild : 0.001352251954766561
francois, callwild : 0.001287859004539582
friday, defoe-robinson-103.txt : 0.001051135534563912
trail, callwild : 8.80036986435381E-4
john, callwild : 8.585726696930545E-4
perrault, callwild : 8.371083529507283E-4
hal, callwild : 7.941797194660755E-4
team, callwild : 7.297867692390963E-4
thoughts, defoe-robinson-103.txt : 6.306813207383471E-4
sol, callwild : 6.224651855274646E-4
leks, callwild : 6.010008687851382E-4
traces, callwild : 6.010008687851382E-4
ice, callwild : 6.010008687851382E-4
around, callwild : 5.580722353004855E-4
dave, callwild : 5.366079185581591E-4

Monitoring

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	Unhealthy Nodes	Rebooted Nodes
114	0	0	114	0	0 B	8 GB	0 B	0	8	0	1	0	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	VCores Pending	VCores Reserved
0	0	0	0	0	0	0	0 B	0 B	0 B	0	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	Tracking UI
application_1492766493795_0116	cloudera	2	MAPREDUCE	root.cloudera	Wed May 17 03:05:53 -0700 2017	Wed May 17 03:06:15 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1492766493795_0115	cloudera	Step 2 : WordCount + wordperdoc	MAPREDUCE	root.cloudera	Wed May 17 03:05:28 -0700 2017	Wed May 17 03:05:50 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1492766493795_0114	cloudera	Step 1 : simple WordCount	MAPREDUCE	root.cloudera	Wed May 17 03:04:56 -0700 2017	Wed May 17 03:05:26 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History

Job Overview

Job Name:	Step 1 : simple WordCount
User Name:	cloudera
Queue:	root.cloudera
State:	SUCCEEDED
Uberized:	false
Submitted:	Wed May 17 03:04:56 PDT 2017
Started:	Wed May 17 03:05:03 PDT 2017
Finished:	Wed May 17 03:05:26 PDT 2017
Elapsed:	22sec
Diagnostics:	
Average Map Time	9sec
Average Shuffle Time	5sec
Average Merge Time	0sec
Average Reduce Time	1sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Wed May 17 03:04:59 PDT 2017	quickstart.cloudera:8042	logs

Task Type	Total	Complete
Map	2	2
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	2
Reduces	0	0	1

Job Overview

Job Name:	Step 2 : WordCount + wordperdoc
User Name:	cloudera
Queue:	root.cloudera
State:	SUCCEEDED
Uberized:	false
Submitted:	Wed May 17 03:05:28 PDT 2017
Started:	Wed May 17 03:05:35 PDT 2017
Finished:	Wed May 17 03:05:50 PDT 2017
Elapsed:	14sec
Diagnostics:	
Average Map Time	4sec
Average Shuffle Time	3sec
Average Merge Time	0sec
Average Reduce Time	1sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Wed May 17 03:05:31 PDT 2017	quickstart.cloudera:8042	logs

Task Type	Total	Complete
Map	1	1
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	1
Reduces	0	0	1

Job Overview

Job Name:	2
User Name:	cloudera
Queue:	root.cloudera
State:	SUCCEEDED
Uberized:	false
Submitted:	Wed May 17 03:05:53 PDT 2017
Started:	Wed May 17 03:05:59 PDT 2017
Finished:	Wed May 17 03:06:14 PDT 2017
Elapsed:	15sec
Diagnostics:	
Average Map Time	4sec
Average Shuffle Time	3sec
Average Merge Time	0sec
Average Reduce Time	1sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Wed May 17 03:05:55 PDT 2017	quickstart.cloudera:8042	logs

Task Type	Total	Complete
Map	1	1
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	1
Reduces	0	0	1