BU MET CS755 Assignment 2 Group 3

02.26.2018

Group 3

Daniel Budris, Don Yoo, Kanchan Mohite, Corey Drees

Overview

As part of Assignment 2 taxi trip details are provided in two sizes one is small dataset of size 94MB and other big dataset of size 8.83GB.

The assignment was to create:

1. Hadoop MapReduce job that computes a list of (**hour of day, number of errors**). GPS error is defined by missing GPS position (Longitude/Latitude)
2. Hadoop MapReduce job to compute the five worst taxis. First a set of (taxi, percent of errors) is derived and then second MapReduce job is implemented to find 5 worst taxis
3. MapReduce job to compute set of (driver, money per minute) using which top 10 best drivers can be found

Github project urls:

https://github.com/danbudris/BU_MET_CS755/assignment2

https://github.com/DonYoo/Cloud-Computing/tree/master/Assignment2

https://github.com/kanchan06mohite/cc_assignment2/tree/master/aasignment2_proj

https://github.com/corey-guy/cc-assignment2

Task1 will return the hours of the day and the number of errors per hour.

Task2 will return an unsorted list of <Medallion Number, Error Rate>.
Task 2 Part 2 returns the highest error rates among the computed error rates; 5 per map.

Task3 will return an unsorted list of <Medallion Number, Money Per Minute>.
Task 3 Part 2 returns the top earning cabs per minute among the output of Task 3; see Task 3 Subsection for top 10 results.

Commands to Execute:

There is a build script in the root of the submitted assignment. Execute the build script in the desired task directory to generate the jar file for that task. Adjust the command below to target the desired jar, if executing from locally. Otherwise, upload the jar to S3 and specify in EMR console.

hadoop jar target/assign2-0.1-SNAPSHOT-jar-with-dependencies.jar ./input ./output

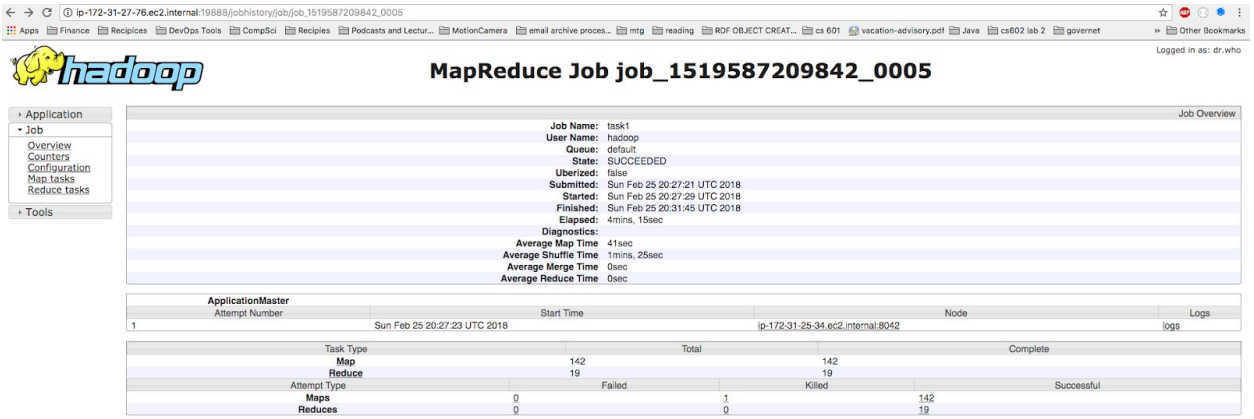
Format:

<hadoop binary> jar <path to jar file> <path to directory containing input files> <path to output; must not exist>

Commands to Compile and Package:

1. sudo mvn -e clean compile assembly:single
2. sudo mvn package

Task 1 Execution (calculate errors per hour):



ip-172-31-27-76.ec2.internal:19888/jobhistory/job/job_1519587209842_0005

Apps | Finance | Recipices | DevOps Tools | CompSci | Recipies | Podcasts and Lectur... | MotionCamera | email archive proces... | mtg | reading | RDF OBJECT CREAT... | cs 601 | vacation-advisory.pdf | Java | cs602 lab 2 | governet | Other Bookmarks

Logged in as: dr.who

MapReduce Job job_1519587209842_0005

- Application
- Job
  - Overview
  - Counters
  - Configuration
  - Map tasks
  - Reduce tasks
- Tools

Job Overview	
Job Name:	task1
User Name:	hadoop
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Feb 25 20:27:21 UTC 2018
Started:	Sun Feb 25 20:27:29 UTC 2018
Finished:	Sun Feb 25 20:31:45 UTC 2018
Elapsed:	4mins, 15sec
Diagnostics:	
Average Map Time	41sec
Average Shuffle Time	1mins, 25sec
Average Merge Time	0sec
Average Reduce Time	0sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sun Feb 25 20:27:23 UTC 2018	ip-172-31-25-34.ec2.internal:8042	logs

Task Type	Total	Complete
Map	142	142
Reduce	19	19

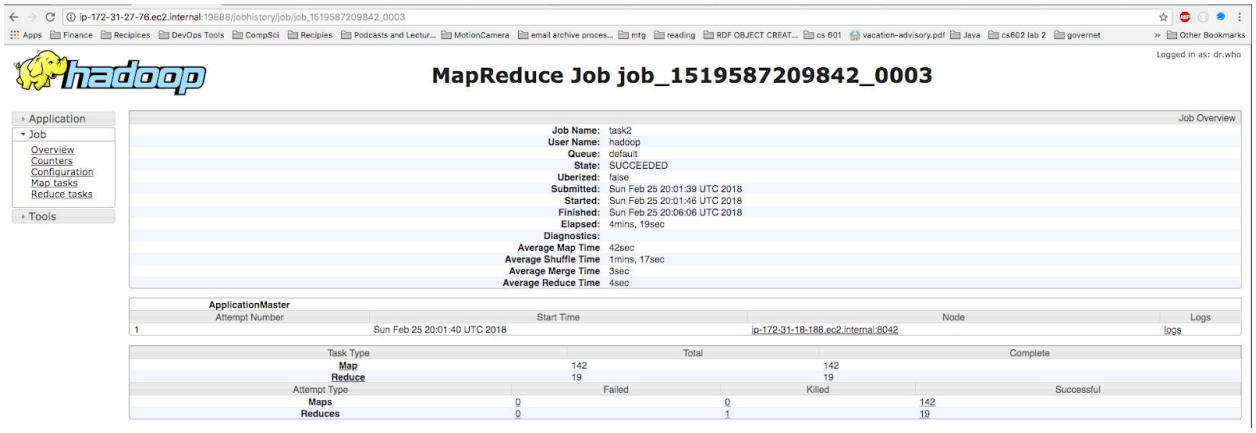
Attempt Type	Failed	Killed	Successful
Maps	0	1	142
Reduces	0	0	19

Task 1 Final Result

The following page contains a list representing the hours of the day from 0 -23, with a value equal to the number of GPS errors which occurred during that time period.

00	116055
01	89526
02	69277
03	54444
04	42140
05	36306
06	67436
07	108479
08	127918
09	130800
10	127126
11	131993
12	139394
13	140292
14	147635
15	145712
16	127443
17	149571
18	174809
19	179546
20	168344
21	167194
22	160924
23	142821

Task 2 Step 1 Execution (compute error rate per taxi):



MapReduce Job job_1519587209842_0003

Job Name:	task2
User Name:	hadoop
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Feb 25 20:01:39 UTC 2018
Started:	Sun Feb 25 20:01:46 UTC 2018
Finished:	Sun Feb 25 20:06:06 UTC 2018
Elapsed:	4mins, 19sec
Diagnostics:	
Average Map Time	42sec
Average Shuffle Time	1mins, 17sec
Average Merge Time	3sec
Average Reduce Time	4sec

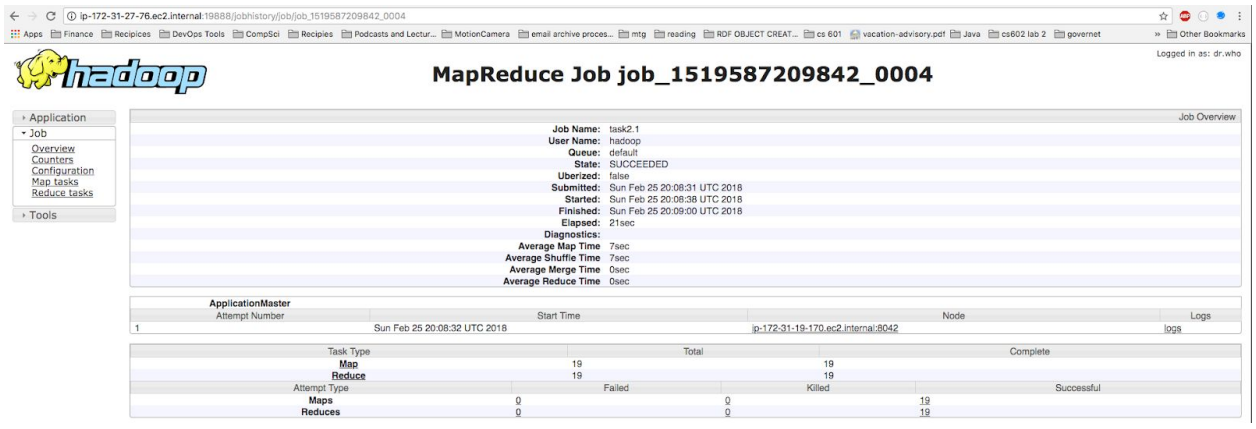
ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sun Feb 25 20:01:40 UTC 2018	ip-172-31-18-188.ec2.internal:8042	logs

Task Type	Total	Complete
Map	142	142
Reduce	19	19

Attempt Type	Failed	Killed	Successful
Maps	0	0	142
Reduces	0	1	19

Task 2 Step 2 Execution (compute worst taxis by error rate):



MapReduce Job job_1519587209842_0004

Job Name:	task2.1
User Name:	hadoop
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Feb 25 20:08:31 UTC 2018
Started:	Sun Feb 25 20:08:38 UTC 2018
Finished:	Sun Feb 25 20:09:00 UTC 2018
Elapsed:	21sec
Diagnostics:	
Average Map Time	7sec
Average Shuffle Time	7sec
Average Merge Time	0sec
Average Reduce Time	0sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sun Feb 25 20:08:32 UTC 2018	ip-172-31-19-170.ec2.internal:8042	logs

Task Type	Total	Complete
Map	19	19
Reduce	19	19

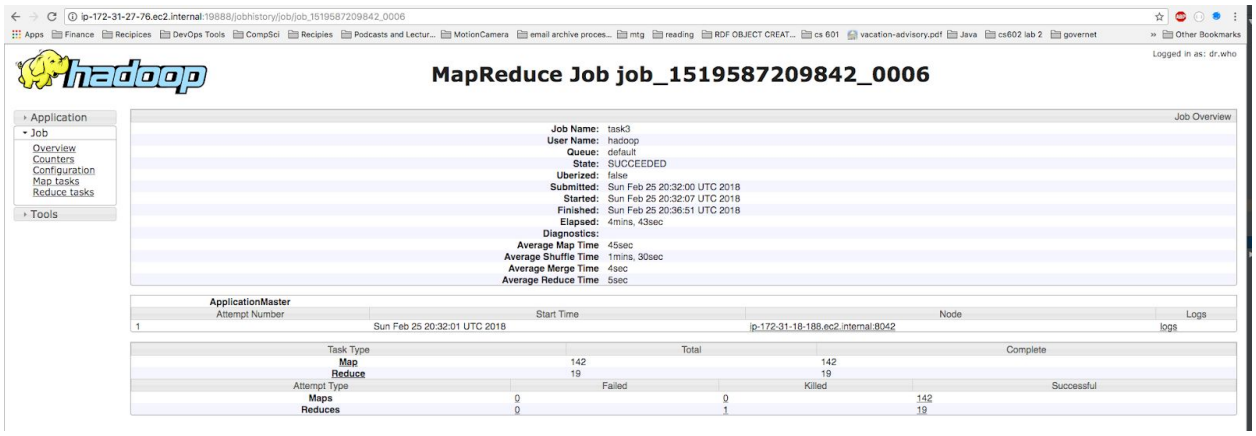
Attempt Type	Failed	Killed	Successful
Maps	0	0	19
Reduces	0	0	19

Task 2 Final Results:

The following page contains the results from Task 2 Execution 2. It is the top 95 worst taxi by error rates. This list was generated by executing the Task 2 Step 2 map/reduce on the results of Task 2 Step 1, then downloading the resulting 19 files from S3 and concatenating them.

1.0	1A1B65DCF008F4B204AC65A5AD89ED41
1.0	D34156D38E2BB0802CFE0FC22577188D
1.0	189C1B13B5EF665B1E6F51AB2C5E369F
1.0	0219EB9A4C74AAA118104359E5A5914C
1.0	FF96A951C04FBCEDE5BCB473CF5CBDBF
1.0	0EE3FFCBDFD8B2979E87F38369A28FD9
1.0	D1F0ED6BAB91522CFF5CAD4F4C13206A
1.0	022B8DF4D6D7C4DCF11233DD74C9E189
1.0	029B453F97625ED1100B553674437545
1.0	F40C6B8673036B010620FE31F194814C
1.0	242E3880050FF5CD9F468081FDEB2A77
1.0	E2ABD9D79BB36E566D39CD751E012D16
1.0	21ACA83090608E3531592A205424CF01
1.0	0AC8D7FEC3D5D41536FE12A8F85B9ED4
1.0	FADBF451645457CDED082F3B97E5064B
1.0	36F5EE6F1888BBA3D85DF33B25EC3EBC
1.0	F91BEA8C011C2874561B308997C4924A
1.0	10FEE41A73CB8A2250DE12C631069A04
1.0	2BB26418292BF4AD3389E36B88C2502B
1.0	EA45F3D3D1E15388F162F3166345030A
1.0	291C64FAAE76CECAD37A3C99D3A6EEBD
1.0	E200FEFDCCEFF516BE2057949E64E4D8
1.0	04CD21118F47FA3B2359C65AC063CF0B
1.0	24901F55D7339A250E42466B97BCF1EB
1.0	BEE2B289BC49B76F36482073D4EF87B7
1.0	0972E3ED97DA223496CEDB7E304CB1CA
1.0	FE394A936F110ADBEF3CF7BD167EB12B
1.0	09D734FEF89D285B673B369B47600EC4
1.0	063620286468AC8945F25F9E53DC093B
1.0	F8D5E1C3CE80253626B37058F67BEA65
1.0	19427F3EA180AEF220CD4E568E2C476E
1.0	DB82B96AE988894E60B8EB3BDA4222AE
1.0	1938FB24FCA72E6F88DD0D32A36BA2D9
1.0	1CD95D7378F2999FE588A598A466ABC8
1.0	FC7074BC59332F362E535C13EBA5339D
1.0	2427250EE6AAEB600E662EBCCA51A431
1.0	FDFE7E83D6F57D6D18C901E3924E7534
1.0	087BF626C3F075B5963B60C62CDB2085
1.0	2B4463611B7D160AD26A04E5E5762AAB
1.0	CF613B8747B135B52F9E36C13AF79769
1.0	2DEA61EED4BCEEC564A00115C4D21334
1.0	93F80CB2B6E17471405933418FC309C0
1.0	12CE65C3876AAB540925B368E8A0E181
1.0	0A0C3F3F29F62642A6DD9D9A087BFBBF
1.0	F23C8C815496EBDF6DBC2A4AA47B03E0
1.0	27CA0F96AC9F521A7BB85E25AE01BF77
1.0	F9A719F1D27466C8EFC81BDE0606573C
1.0	1533072F784BC7BB342B4BD2935938D6
1.0	0F1C915E984249892B078896F227F46D
1.0	F7F210DA15DBB4283299B2AFFD9BB238
1.0	067D656CDED3EEFD77F33BB8F67DC655
1.0	F427DF008ECC3F56914BB27286FAC660
1.0	00AC8ED3B4327BDD4EBBEBCB2BA10A00
1.0	03161C41A5C96BA272C80994F196DDC0
1.0	ED8F3855D3FFAFEEB9EE45EB6EDEB4DE
1.0	6009157C175AAFFCBACB125B9C0D6837
1.0	FE513AE22CB4F54547E14748E469BD38
1.0	165F05ACA6203A8F38C306AD114E2C05
1.0	3C44D0C5AF81AFA5D4A877BBD69B8E27
1.0	EFD901F046773D229CCBED39788EA5B7
1.0	4CF6AD8FD807A215E74DDAA82775553B
1.0	B29C3B3C3A99B96F5379D8D70B88B444
1.0	1805C5DDC14A8CD06174238AEFEEC9E8
1.0	433AB61374DBC1729C252F951C64AB59
1.0	F294908EAA4954F1DD8B22E922733640
1.0	66160ECF58CF25FBB46F84BEDE0C5243
1.0	F29D71F3B25F5BCCBC35445F895294B5
1.0	54ADBF2188B5FB5080702EC7B281D210
1.0	0B555EC534B208DDD8211150204151D8
1.0	FE757A29F1129533CD6D4A0EC6034106
1.0	1D569B1E7EB6CD3987C411E3717E4E35
1.0	FC221309746013AC554571FBD180E1C8
1.0	0CAEEA5D95C687B4F7A683D162830BE4
1.0	00DC83118CA675B9A2876C35E3398AF5
1.0	C32146C961ED76668D0E911D4DB8C513
1.0	3317B831A6D8A0FC01BCB27B2E178E1A
1.0	F4AE8CBC91AC2004AB333370F3616793
1.0	1F5AC15C6E163FE538AE57734789E7F6
1.0	20468543B4287CC6DA3E9E8098E4E930
1.0	FDCECB96F38C465AE9C593F13E9163AB
1.0	1ECE9AB1BC7E02671C8526541172399D
1.0	EEC4CECD06C7F0950D5FA04E5084B2F8
1.0	0A1CDBC8EEE4A0A1F990D511697DD877
1.0	02510B3B0E797E51AF73361185F62D0B
1.0	F194698DD90363E76D5B310381702C88
1.0	56DBBF8AE7305DDA3C30BC96C0D2220E
1.0	E123303EBE96A8D4736B405D46CEA902
1.0	233EB0028FD0C8B9B002B96E45AA5182
1.0	2211BD68FD108E048BF084A3E9E17689
1.0	DB455A37668B3BD9014A4134CBE2294A
1.0	14C5001FBF4706F49E6D436FA1EC8428
1.0	B5A1CC78FB1CE28CD81EC2273703FF1E
1.0	0FCBC61ACD0479DC77E3CCCC0F5FFCA7
1.0	05177950A30C6A5820365D911170C4FB
1.0	F13E1C7DFED00FF4F562C7E86569C36E

Task 3 Step 1 Execution (compute money per minute):



ip-172-31-27-76.ec2.internal:19888/jobhistory/job/job_1519587209842_0006

MapReduce Job job_1519587209842_0006

Job Overview

Job Name:	task3
User Name:	hadoop
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Feb 25 20:32:00 UTC 2018
Started:	Sun Feb 25 20:32:07 UTC 2018
Finished:	Sun Feb 25 20:36:51 UTC 2018
Elapsed:	4mins, 43sec
Diagnostics:	
Average Map Time	45sec
Average Shuffle Time	1mins, 30sec
Average Merge Time	4sec
Average Reduce Time	5sec

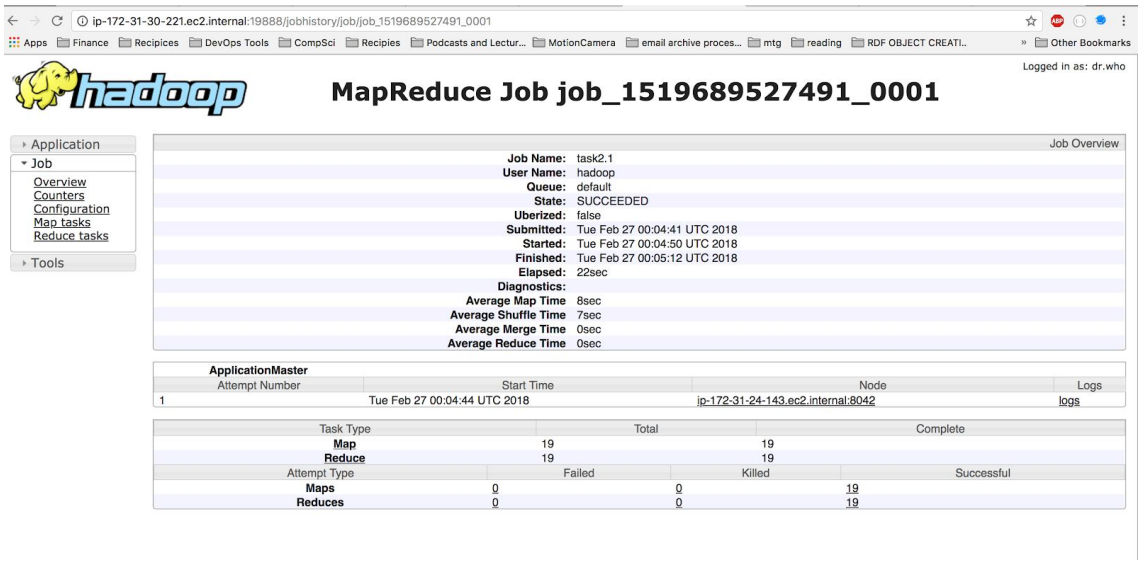
ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sun Feb 25 20:32:01 UTC 2018	ip-172-31-18-188.ec2.internal:8042	logs

Task Type	Total	Complete
Map	142	142
Reduce	19	19

Attempt Type	Failed	Killed	Successful
Maps	0	0	142
Reduces	0	1	19

Task 3 Step 2 Execution (sort highest rate of money per minute):



ip-172-31-30-221.ec2.internal:19888/jobhistory/job/job_1519689527491_0001

MapReduce Job job_1519689527491_0001

Job Overview

Job Name:	task2.1
User Name:	hadoop
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Tue Feb 27 00:04:41 UTC 2018
Started:	Tue Feb 27 00:04:50 UTC 2018
Finished:	Tue Feb 27 00:05:12 UTC 2018
Elapsed:	22sec
Diagnostics:	
Average Map Time	8sec
Average Shuffle Time	7sec
Average Merge Time	0sec
Average Reduce Time	0sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Tue Feb 27 00:04:44 UTC 2018	ip-172-31-24-143.ec2.internal:8042	logs

Task Type	Total	Complete
Map	19	19
Reduce	19	19

Attempt Type	Failed	Killed	Successful
Maps	0	0	19
Reduces	0	0	19

Task 3 Final Results (top ten taxis by money per minute):

The following page contains the concatenated, sorted results from task 3.2, which is the top ten taxis by money earned per minute.

52.34853135320577	E0B315ABFE103A63E919FEB6760209E9
55.77605728692445	BBC3E143398F546F06B964C6FA21BD07
80.3637632371067	69D959302A72045B426E936BB0BB2723
87.18989690721648	9DAAD7BFA53C91605104DD1874EF97E4
104.66472412703968	FA8C6BEC76883E9E5080EDD69D66B411
105.25148315250308	E8B6C24CA6EE4ED994B038DCD633B58D
117.0	C46F08489A5517D480B4DB06F691FD08
125.29339542036914	E77A964307CF49B32AD77E298A4951D0
133.77006756756754	2CB4FE05D307D6294A6E31C00E5F2755
138.2	8CBAFBD97A86B7CC787401F7BBCA3F9D