Report

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student Cao Yuhang)
Name: Cao Yuhang

Date: 3/17/2018
SID: 1155092180

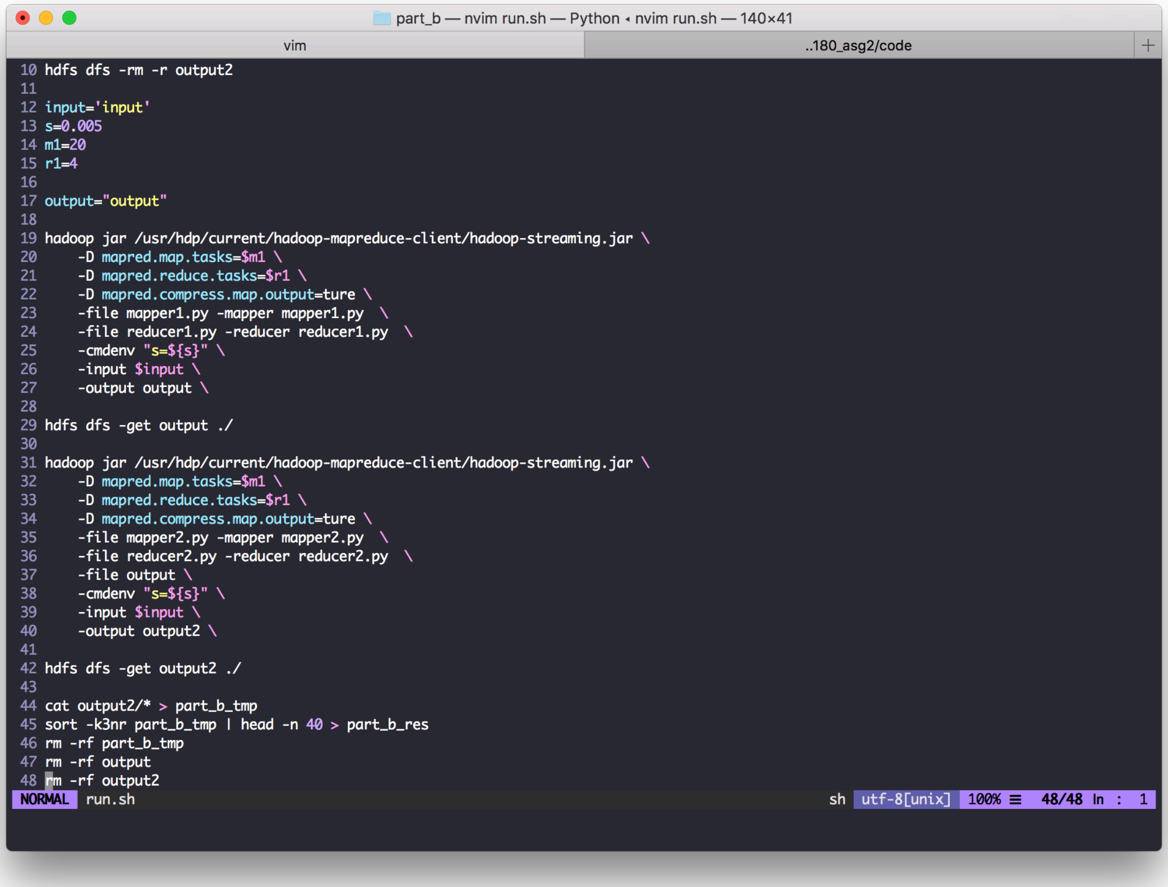
	part (a)	part (b)	part (d)
Job1 time		31s	25
Job2 time		27s	28
Total time	2m50s	58s	53s

* For part (a) and part (d), #mappers is 20, #reducers is 4
* I run part (a) in my own mac book, my aws has no credit
* All time are fast running time for that experiment

Theoretically, running time should be d < b < a, since in part (d) we use PCY algorithm to further filter the false positive compared with part (b); and by using map reduce, parallelization should beat local machine.

According to the form, we can find running time d < b < a, it prove our guess.

All command can be found in “run.sh” in corresponding part directory. Here is the screenshot of “run.sh” for part (b):


part_b — nvim run.sh — Python ◂ nvim run.sh — 140×41
vim
..180_asg2/code
NORMAL run.sh
sh utf-8[unix] 100% ≡ 48/48 ln : 1

```
hdfs dfs -rm -r output2

input='input'
s=0.005
m1=20
r1=4

output="output"

hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \
    -D mapred.map.tasks=$m1 \
    -D mapred.reduce.tasks=$r1 \
    -D mapred.compress.map.output=ture \
    -file mapper1.py -mapper mapper1.py  \
    -file reducer1.py -reducer reducer1.py  \
    -cmdenv "s=${s}" \
    -input $input \
    -output output \

hdfs dfs -get output ./

hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \
    -D mapred.map.tasks=$m1 \
    -D mapred.reduce.tasks=$r1 \
    -D mapred.compress.map.output=ture \
    -file mapper2.py -mapper mapper2.py  \
    -file reducer2.py -reducer reducer2.py  \
    -file output \
    -cmdenv "s=${s}" \
    -input $input \
    -output output2 \

hdfs dfs -get output2 ./

cat output2/* > part_b_tmp
sort -k3nr part_b_tmp | head -n 40 > part_b_res
rm -rf part_b_tmp
rm -rf output
rm -rf output2
```

All the results can be found in corresponding part directory. For example, the result of part (b) can be found in “part_b/part_b_res.txt”

Q2

(a)

$$\begin{cases} 1-\left(1-s^{r}\right)^{B} \geq P1 & s \geq T1 \\ 1-\left(1-s^{r}\right)^{B} \geq P2 & s \leq T2 \end{cases}$$

(b)

$$\begin{cases} 1-\left(1-s^{r}\right)^{20} \geq 0.99 & s \geq 0.9 \\ 1-\left(1-s^{r}\right)^{20} \leq 0.05 & s \leq 0.7 \end{cases}$$

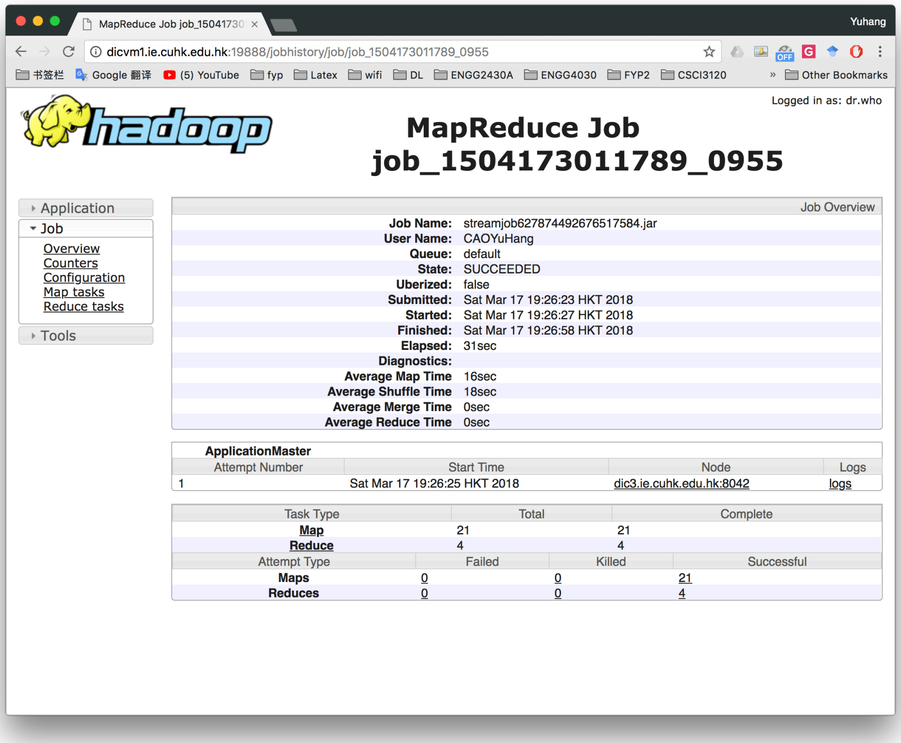
$$\begin{cases} r \geq \log_{0.9}\left(1-\left(1-0.99\right)^{\frac{1}{20}}\right) = 15.01 \\ r \leq \log_{0.7}\left(1-\left(1-0.05\right)^{\frac{1}{20}}\right) = 16.73 \end{cases}$$

$$\begin{cases} B = 20 \\ r = 16 \end{cases}$$

Screen shot of running time of part (b) and part (d):

Part (b):

Job 1:



MapReduce Job job_1504173011789_0955

	Job Overview
Job Name:	streamjob627874492676517584.jar
User Name:	CAOYuHang
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sat Mar 17 19:26:23 HKT 2018
Started:	Sat Mar 17 19:26:27 HKT 2018
Finished:	Sat Mar 17 19:26:58 HKT 2018
Elapsed:	31sec
Diagnostics:	
Average Map Time	16sec
Average Shuffle Time	18sec
Average Merge Time	0sec
Average Reduce Time	0sec

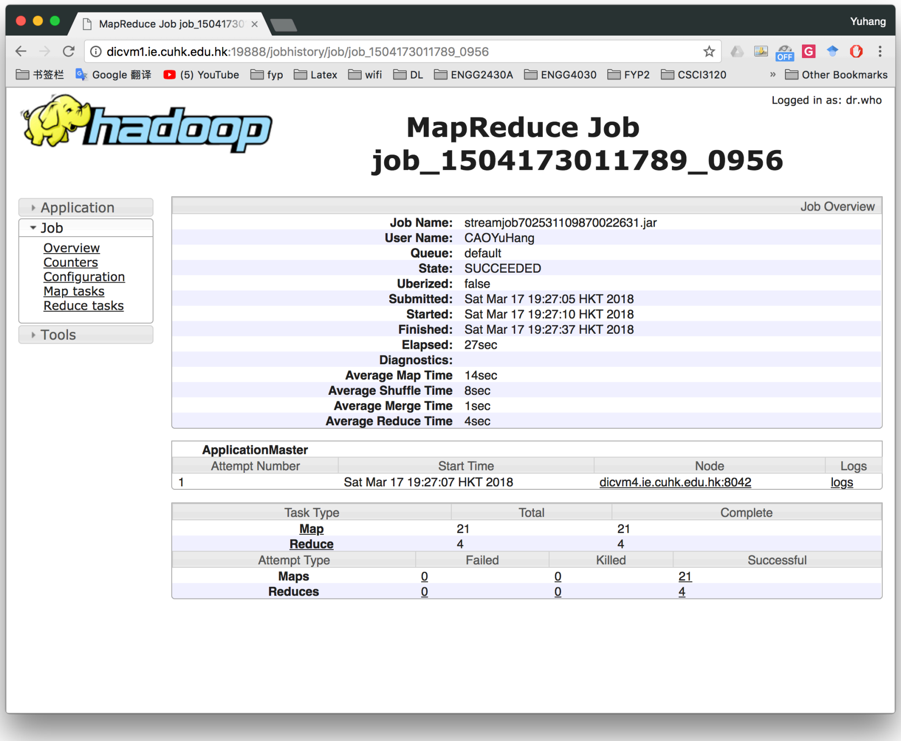
ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sat Mar 17 19:26:25 HKT 2018	dic3.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	21	21
Reduce	4	4

Attempt Type	Failed	Killed	Successful
Maps	0	0	21
Reduces	0	0	4

Job 2:



MapReduce Job job_1504173011789_0956

	Job Overview
Job Name:	streamjob702531109870022631.jar
User Name:	CAOYuHang
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sat Mar 17 19:27:05 HKT 2018
Started:	Sat Mar 17 19:27:10 HKT 2018
Finished:	Sat Mar 17 19:27:37 HKT 2018
Elapsed:	27sec
Diagnostics:	
Average Map Time	14sec
Average Shuffle Time	8sec
Average Merge Time	1sec
Average Reduce Time	4sec

ApplicationMaster

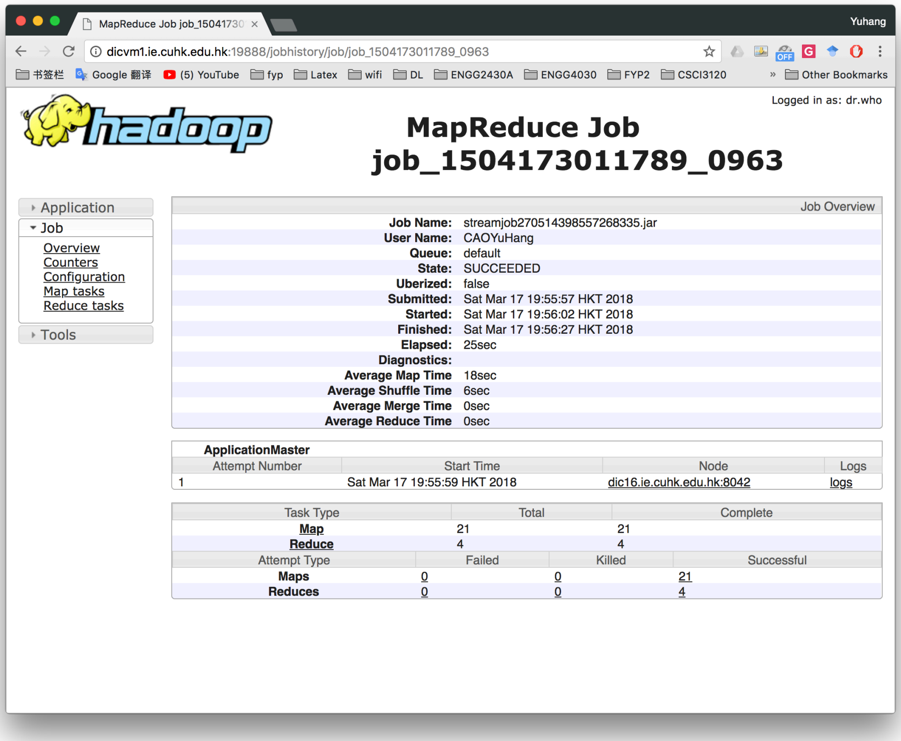
Attempt Number	Start Time	Node	Logs
1	Sat Mar 17 19:27:07 HKT 2018	dicvm4.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	21	21
Reduce	4	4

Attempt Type	Failed	Killed	Successful
Maps	0	0	21
Reduces	0	0	4

Part (d)

Job 1:



Logged in as: dr.who

MapReduce Job job_1504173011789_0963

	Job Overview
Job Name:	streamjob270514398557268335.jar
User Name:	CAOYuHang
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sat Mar 17 19:55:57 HKT 2018
Started:	Sat Mar 17 19:56:02 HKT 2018
Finished:	Sat Mar 17 19:56:27 HKT 2018
Elapsed:	25sec
Diagnostics:	
Average Map Time	18sec
Average Shuffle Time	6sec
Average Merge Time	0sec
Average Reduce Time	0sec

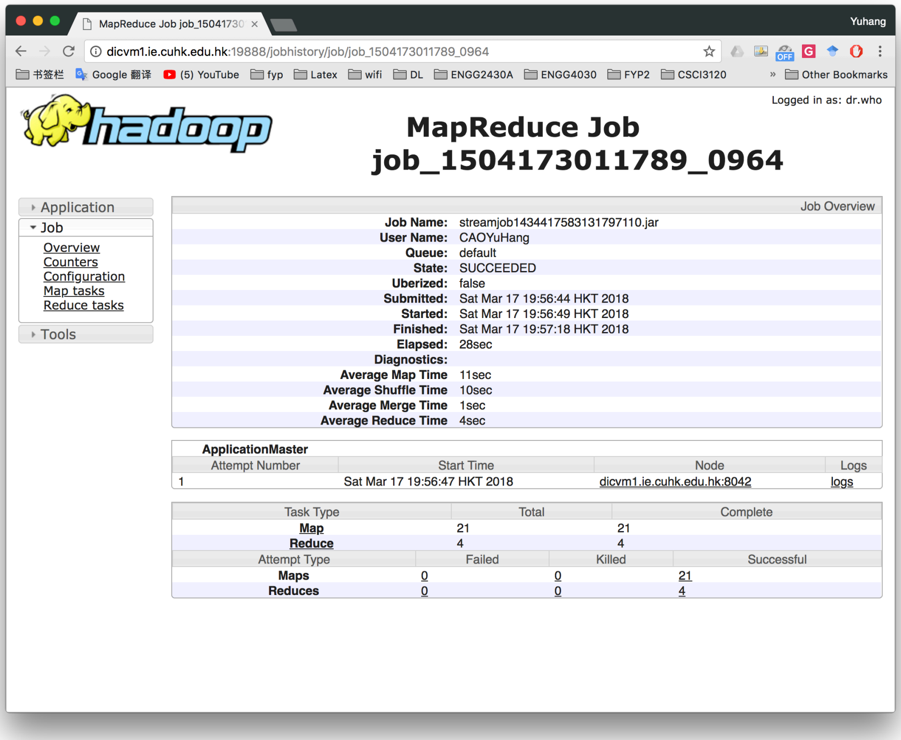
ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sat Mar 17 19:55:59 HKT 2018	dic16.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	21	21
Reduce	4	4

Attempt Type	Failed	Killed	Successful
Maps	0	0	21
Reduces	0	0	4

Job2:



Logged in as: dr.who

MapReduce Job job_1504173011789_0964

	Job Overview
Job Name:	streamjob1434417583131797110.jar
User Name:	CAOYuHang
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sat Mar 17 19:56:44 HKT 2018
Started:	Sat Mar 17 19:56:49 HKT 2018
Finished:	Sat Mar 17 19:57:18 HKT 2018
Elapsed:	28sec
Diagnostics:	
Average Map Time	11sec
Average Shuffle Time	10sec
Average Merge Time	1sec
Average Reduce Time	4sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sat Mar 17 19:56:47 HKT 2018	dicvm1.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	21	21
Reduce	4	4

Attempt Type	Failed	Killed	Successful
Maps	0	0	21
Reduces	0	0	4