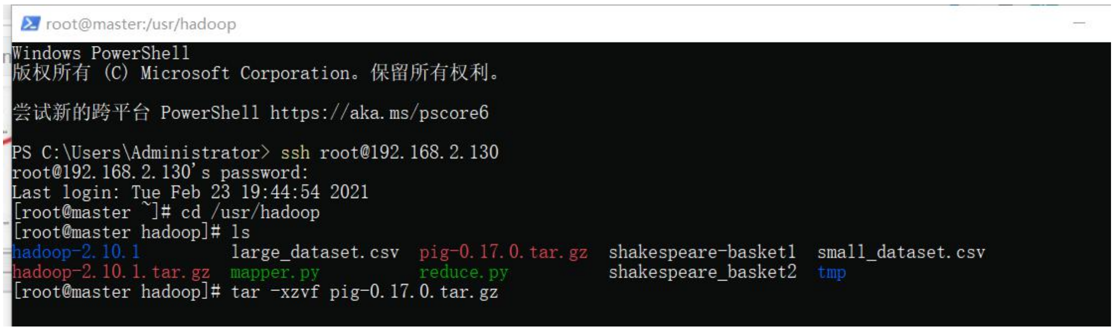
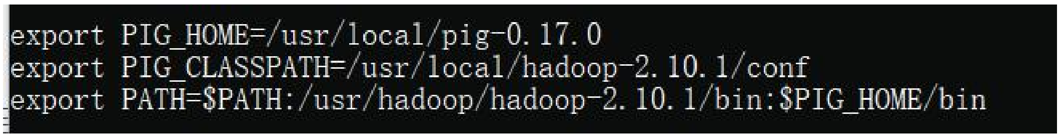
IEMS5730 Homework2

Q1 Pig Setup



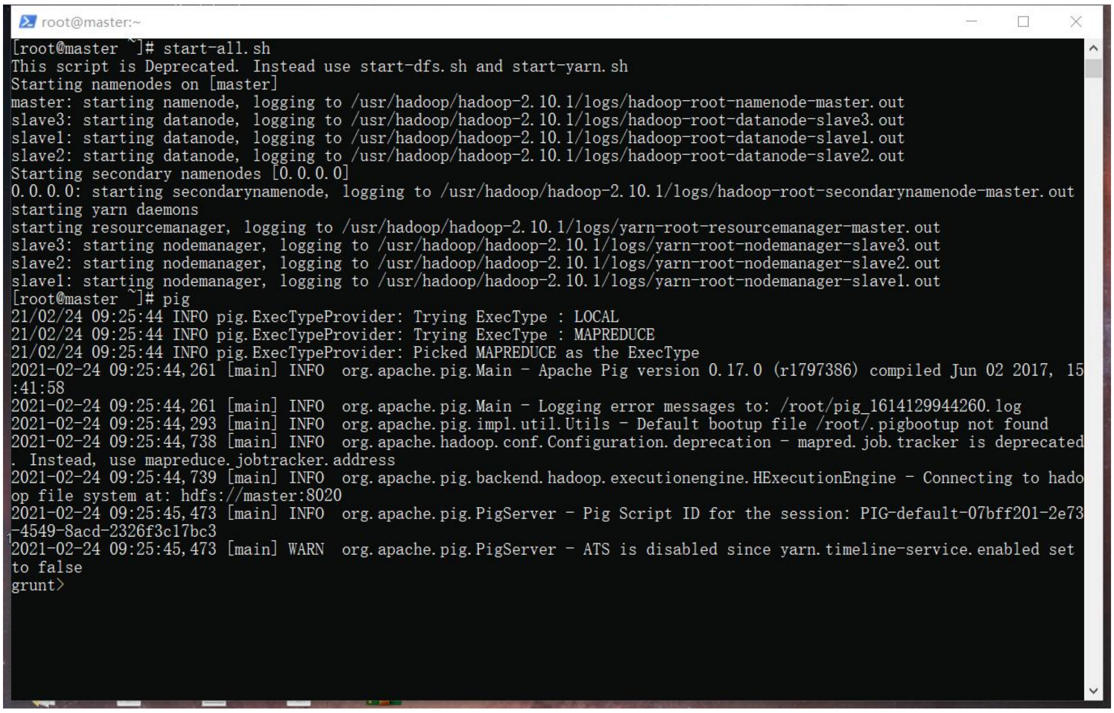
```
Windows PowerShell
版权所有 (C) Microsoft Corporation。保留所有权利。

尝试新的跨平台 PowerShell https://aka.ms/pscore6

PS C:\Users\Administrator> ssh root@192.168.2.130
root@192.168.2.130's password:
Last login: Tue Feb 23 19:44:54 2021
[root@master ~]# cd /usr/hadoop
[root@master hadoop]# ls
hadoop-2.10.1         large_dataset.csv  pig-0.17.0.tar.gz  shakespeare-basket1  small_dataset.csv
hadoop-2.10.1.tar.gz  mapper.py          reduce.py          shakespeare_basket2  tmp
[root@master hadoop]# tar -xzvf pig-0.17.0.tar.gz
```

P1 Pig install



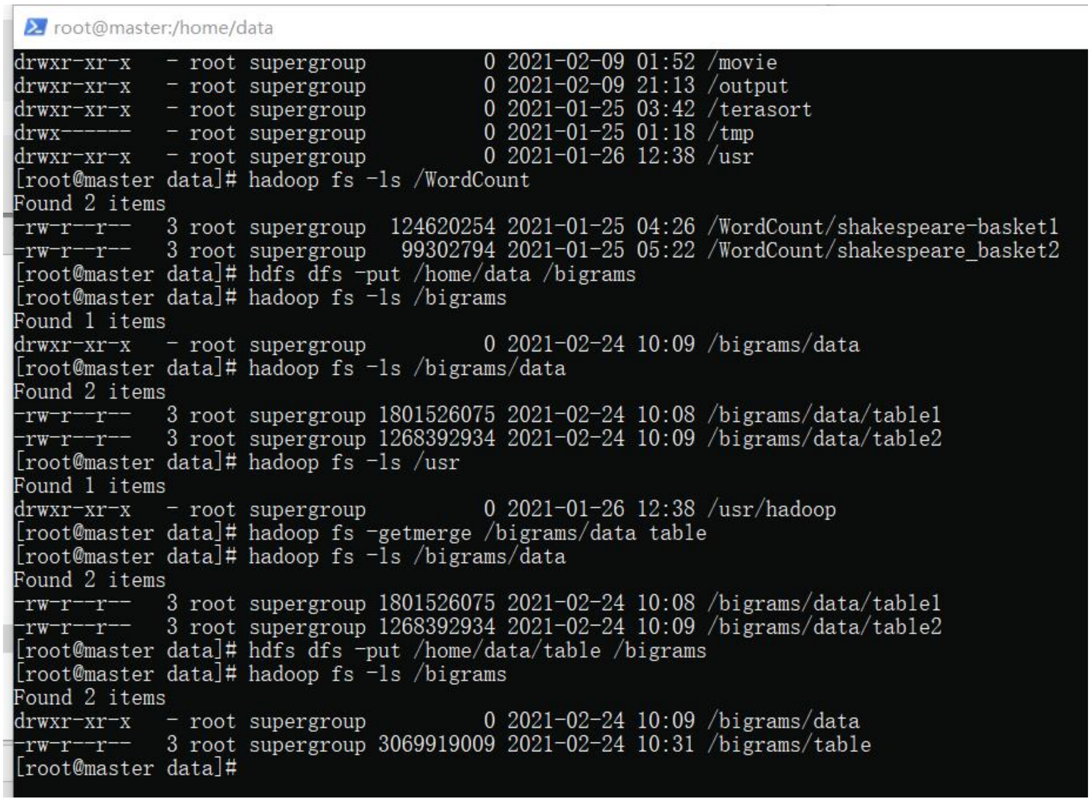
```
export PIG_HOME=/usr/local/pig-0.17.0
export PIG_CLASSPATH=/usr/local/hadoop-2.10.1/conf
export PATH=$PATH:/usr/hadoop/hadoop-2.10.1/bin:$PIG_HOME/bin
```

P2 Pig environment setting



```
[root@master ~]# start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [master]
master: starting namenode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-namenode-master.out
slave3: starting datanode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-datanode-slave3.out
slave1: starting datanode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-datanode-slave1.out
slave2: starting datanode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-datanode-slave2.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-secondarynamenode-master.out
starting yarn daemons
starting resourcemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-resourcemanager-master.out
slave3: starting nodemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-nodemanager-slave3.out
slave2: starting nodemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-nodemanager-slave2.out
slave1: starting nodemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-nodemanager-slave1.out
[root@master ~]# pig
21/02/24 09:25:44 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
21/02/24 09:25:44 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
21/02/24 09:25:44 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-02-24 09:25:44,261 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15
:41:58
2021-02-24 09:25:44,261 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1614129944260.log
2021-02-24 09:25:44,293 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2021-02-24 09:25:44,738 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2021-02-24 09:25:44,739 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://master:8020
2021-02-24 09:25:45,473 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-07bff201-2e73
-4549-8acd-2326f3c17bc3
2021-02-24 09:25:45,473 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set
to false
grunt>
```

P3 Connect to pig

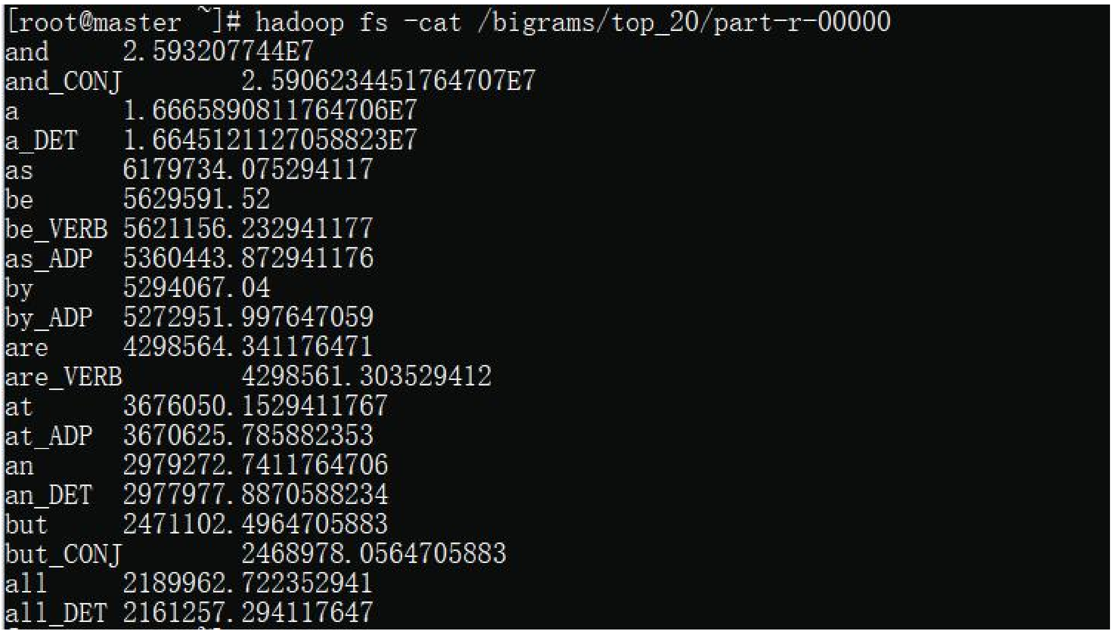


```
drwxr-xr-x   - root supergroup          0 2021-02-09 01:52 /movie
drwxr-xr-x   - root supergroup          0 2021-02-09 21:13 /output
drwxr-xr-x   - root supergroup          0 2021-01-25 03:42 /terasort
drwx------   - root supergroup          0 2021-01-25 01:18 /tmp
drwxr-xr-x   - root supergroup          0 2021-01-26 12:38 /usr
[root@master data]# hadoop fs -ls /WordCount
Found 2 items
-rw-r--r--   3 root supergroup  124620254 2021-01-25 04:26 /WordCount/shakespeare-basket1
-rw-r--r--   3 root supergroup   99302794 2021-01-25 05:22 /WordCount/shakespeare_basket2
[root@master data]# hdfs dfs -put /home/data /bigrams
[root@master data]# hadoop fs -ls /bigrams
Found 1 items
drwxr-xr-x   - root supergroup          0 2021-02-24 10:09 /bigrams/data
[root@master data]# hadoop fs -ls /bigrams/data
Found 2 items
-rw-r--r--   3 root supergroup 1801526075 2021-02-24 10:08 /bigrams/data/table1
-rw-r--r--   3 root supergroup 1268392934 2021-02-24 10:09 /bigrams/data/table2
[root@master data]# hadoop fs -ls /usr
Found 1 items
drwxr-xr-x   - root supergroup          0 2021-01-26 12:38 /usr/hadoop
[root@master data]# hadoop fs -getmerge /bigrams/data table
[root@master data]# hadoop fs -ls /bigrams/data
Found 2 items
-rw-r--r--   3 root supergroup 1801526075 2021-02-24 10:08 /bigrams/data/table1
-rw-r--r--   3 root supergroup 1268392934 2021-02-24 10:09 /bigrams/data/table2
[root@master data]# hdfs dfs -put /home/data/table /bigrams
[root@master data]# hadoop fs -ls /bigrams
Found 2 items
drwxr-xr-x   - root supergroup          0 2021-02-24 10:09 /bigrams/data
-rw-r--r--   3 root supergroup 3069919009 2021-02-24 10:31 /bigrams/table
[root@master data]#
```

P4 Combine two tables into one table and upload to hdfs

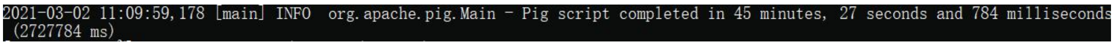
Pig Script:

```
A = load '/bigrams/table' as (bigram: chararray, year: int, match_count: int, volume_count: int);
describe A;
grouped = group A by bigram;
avgoccurence = foreach grouped generate group, AVG(A.match_count) as avgoccur;
sorted = order avgoccurence by avgoccur desc;
top_20 = limit sorted 20;
describe top_20;
dump top_20;
Store top_20 into '/bigrams/top_20';

hadoop fs -cat /bigrams/top_20
```



```
[root@master ~]# hadoop fs -cat /bigrams/top_20/part-r-00000
and     2.593207744E7
and_CONJ        2.5906234451764707E7
a       1.6665890811764706E7
a_DET   1.6645121127058823E7
as      6179734.075294117
be      5629591.52
be_VERB 5621156.232941177
as_ADP  5360443.872941176
by      5294067.04
by_ADP  5272951.997647059
are     4298564.341176471
are_VERB        4298561.303529412
at      3676050.1529411767
at_ADP  3670625.785882353
an      2979272.7411764706
an_DET  2977977.8870588234
but     2471102.4964705883
but_CONJ        2468978.0564705883
all     2189962.722352941
all_DET 2161257.294117647
```

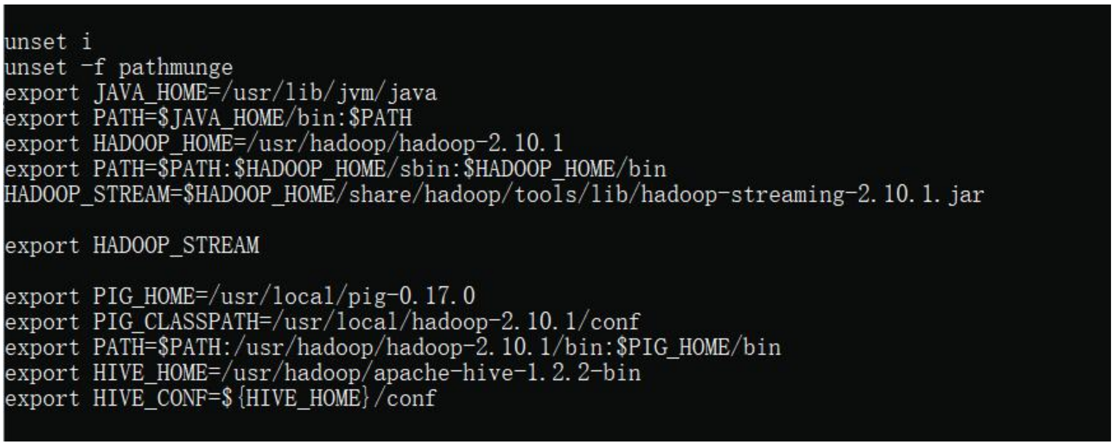
P5 Pig result



```
2021-03-02 11:09:59,178 [main] INFO  org.apache.pig.Main - Pig script completed in 45 minutes, 27 seconds and 784 milliseconds
 (2727784 ms)
```

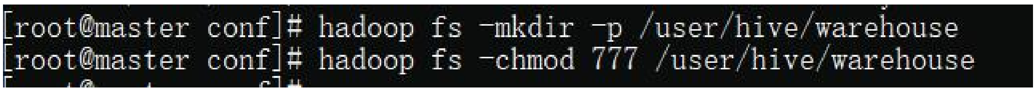
P6 Pig running time(45minutes, 27seconds and 784milliseconds)

Q2 Hive Setup



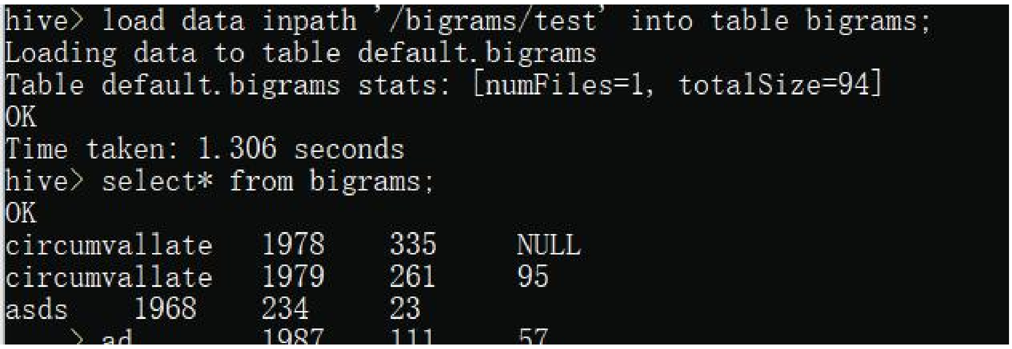
```
unset i
unset -f pathmunge
export JAVA_HOME=/usr/lib/jvm/java
export PATH=$JAVA_HOME/bin:$PATH
export HADOOP_HOME=/usr/hadoop/hadoop-2.10.1
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
HADOOP_STREAM=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-2.10.1.jar

export HADOOP_STREAM

export PIG_HOME=/usr/local/pig-0.17.0
export PIG_CLASSPATH=/usr/local/hadoop-2.10.1/conf
export PATH=$PATH:/usr/hadoop/hadoop-2.10.1/bin:$PIG_HOME/bin
export HIVE_HOME=/usr/hadoop/apache-hive-1.2.2-bin
export HIVE_CONF=${HIVE_HOME}/conf
```

P7 Hive install and environment setting



```
[root@master conf]# hadoop fs -mkdir -p /user/hive/warehouse
[root@master conf]# hadoop fs -chmod 777 /user/hive/warehouse
```

P8 Grant permissions

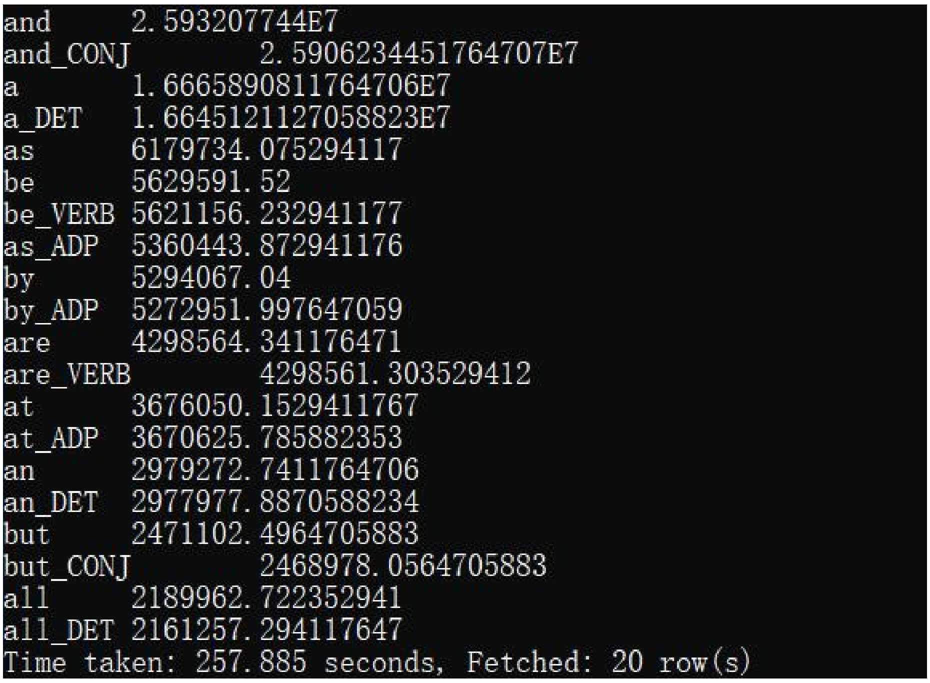


```
hive> load data inpath '/bigrams/test' into table bigrams;
Loading data to table default.bigrams
Table default.bigrams stats: [numFiles=1, totalSize=94]
OK
Time taken: 1.306 seconds
hive> select* from bigrams;
OK
circumvallate   1978    335     NULL
circumvallate   1979    261     95
asds    1968    234     23
    > ad        1987    111     57
```

P9 Hive test

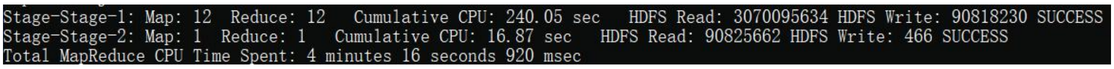
Hive scripts:

create table bigrams(bigram string, year int, match_count int, volume int) row format delimited fields terminated by '\t';
show tables;
LOAD DATA INPATH '/bigrams/table' INTO TABLE bigrams;
SELECT bigram,AVG(match_count) as avg FROM bigrams GROUP BY bigram order by avg desc limit 20;



```
and     2.593207744E7
and_CONJ        2.5906234451764707E7
a       1.6665890811764706E7
a_DET   1.6645121127058823E7
as      6179734.075294117
be      5629591.52
be_VERB 5621156.232941177
as_ADP  5360443.872941176
by      5294067.04
by_ADP  5272951.997647059
are     4298564.341176471
are_VERB        4298561.303529412
at      3676050.1529411767
at_ADP  3670625.785882353
an      2979272.7411764706
an_DET  2977977.8870588234
but     2471102.4964705883
but_CONJ        2468978.0564705883
all     2189962.722352941
all_DET 2161257.294117647
Time taken: 257.885 seconds, Fetched: 20 row(s)
```

P10 Hive result(same as pig)

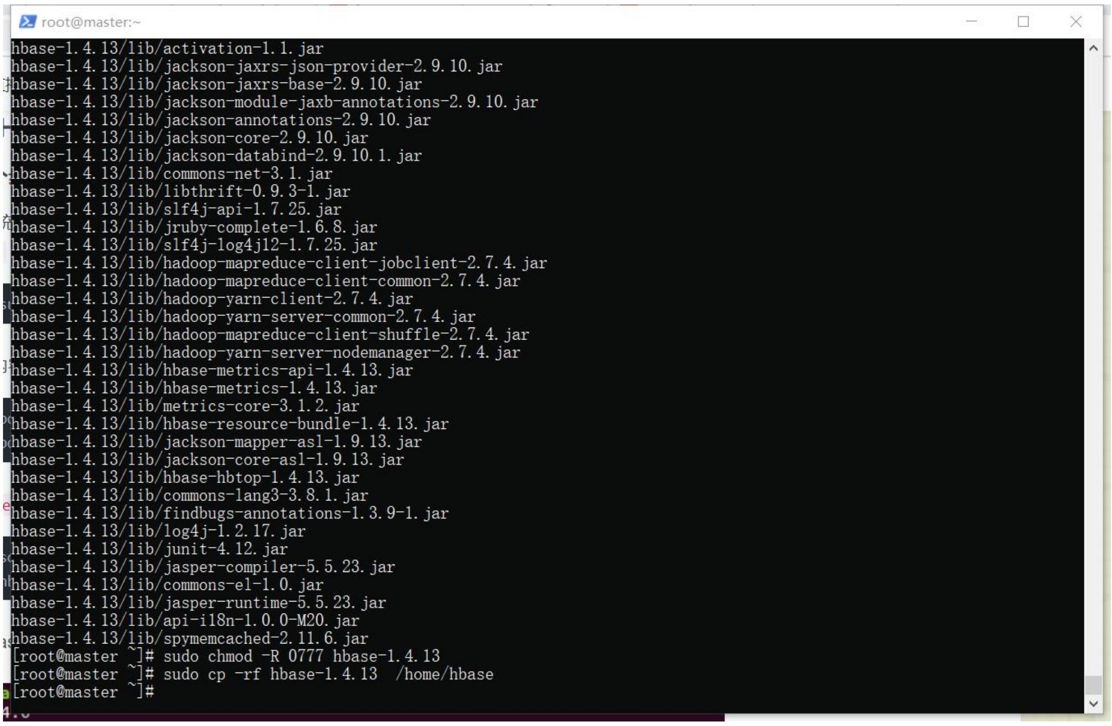


```
Stage-Stage-1: Map: 12  Reduce: 12   Cumulative CPU: 240.05 sec   HDFS Read: 3070095634 HDFS Write: 90818230 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 16.87 sec   HDFS Read: 90825662 HDFS Write: 466 SUCCESS
Total MapReduce CPU Time Spent: 4 minutes 16 seconds 920 msec
```

P11 Hive running time(4 minutes 16 seconds 920 msec)

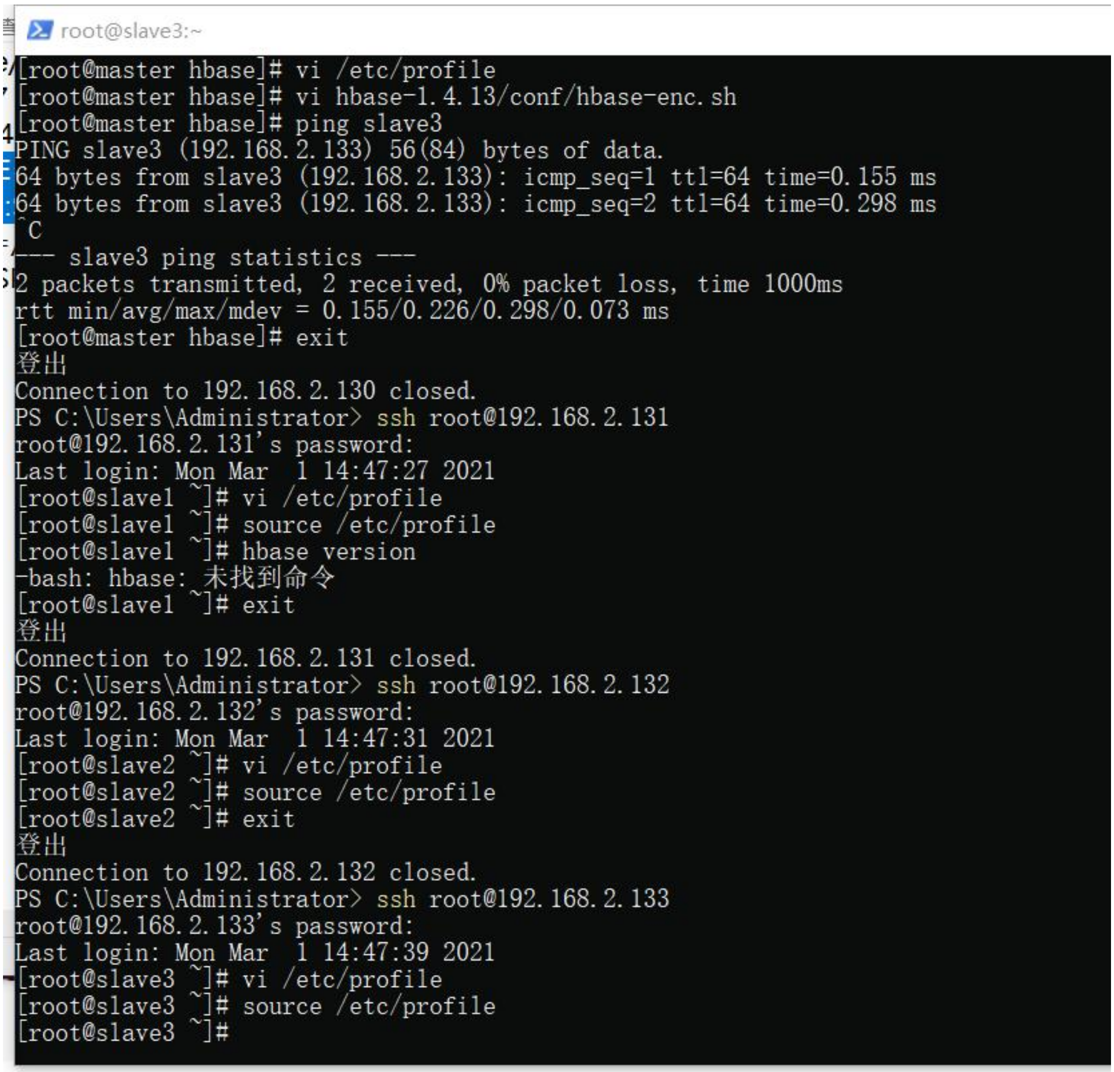
In P6 and P11, we can compare the running time of pig and hive. From my results, the hive running time is much more less than pig's. The pig spent 45 minutes while hive spent only about 4 minutes. Put aside the time complexity of my program and the system delay, hive can run faster than pig to get the same result.

Q3: Hbase Setup



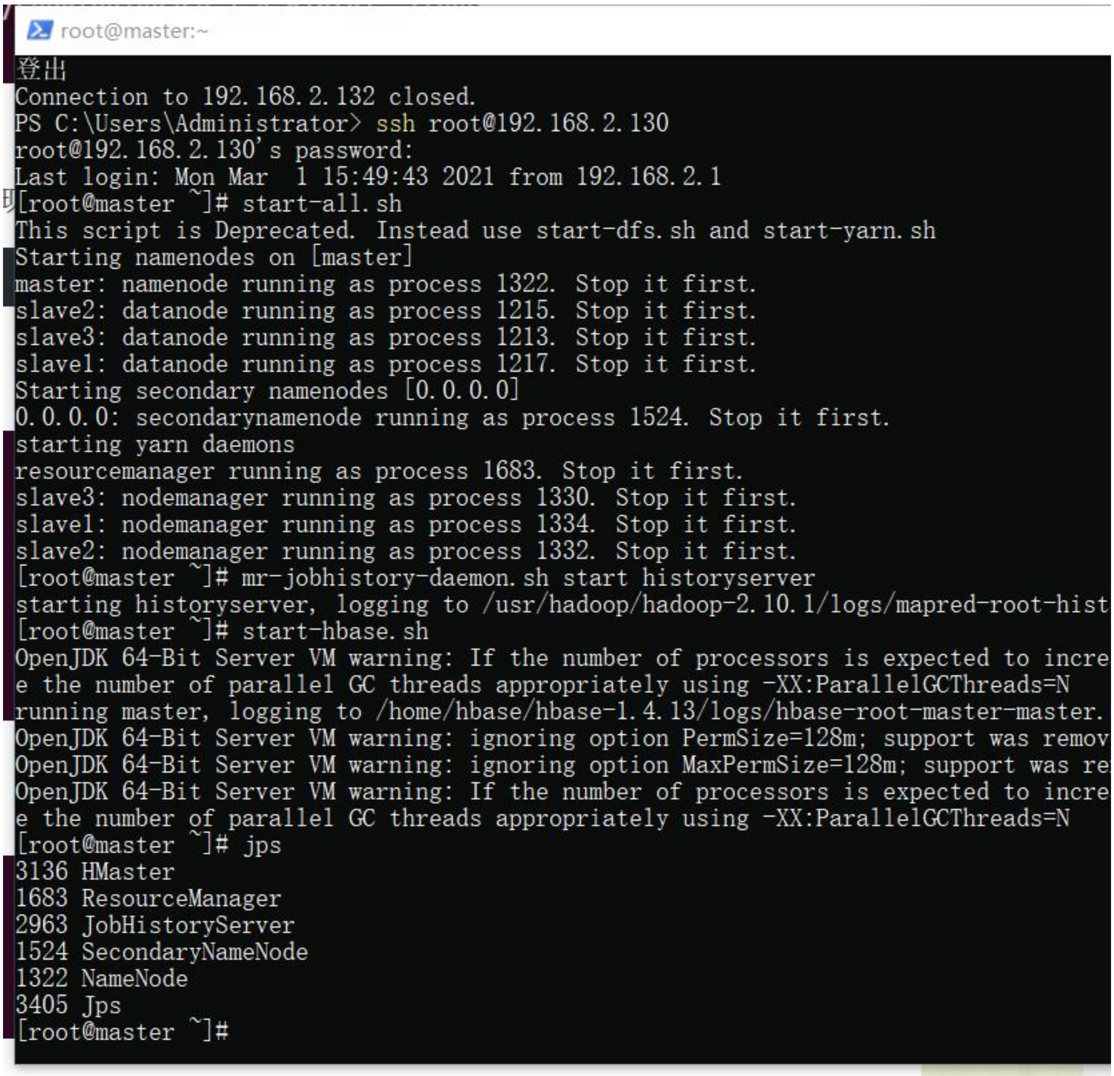
```
hbase-1.4.13/lib/activation-1.1.jar
hbase-1.4.13/lib/jackson-jaxrs-json-provider-2.9.10.jar
hbase-1.4.13/lib/jackson-jaxrs-base-2.9.10.jar
hbase-1.4.13/lib/jackson-module-jaxb-annotations-2.9.10.jar
hbase-1.4.13/lib/jackson-annotations-2.9.10.jar
hbase-1.4.13/lib/jackson-core-2.9.10.jar
hbase-1.4.13/lib/jackson-databind-2.9.10.1.jar
hbase-1.4.13/lib/commons-net-3.1.jar
hbase-1.4.13/lib/libthrift-0.9.3-1.jar
hbase-1.4.13/lib/slf4j-api-1.7.25.jar
hbase-1.4.13/lib/jruby-complete-1.6.8.jar
hbase-1.4.13/lib/slf4j-log4j12-1.7.25.jar
hbase-1.4.13/lib/hadoop-mapreduce-client-jobclient-2.7.4.jar
hbase-1.4.13/lib/hadoop-mapreduce-client-common-2.7.4.jar
hbase-1.4.13/lib/hadoop-yarn-client-2.7.4.jar
hbase-1.4.13/lib/hadoop-yarn-server-common-2.7.4.jar
hbase-1.4.13/lib/hadoop-mapreduce-client-shuffle-2.7.4.jar
hbase-1.4.13/lib/hadoop-yarn-server-nodemanager-2.7.4.jar
hbase-1.4.13/lib/hbase-metrics-api-1.4.13.jar
hbase-1.4.13/lib/hbase-metrics-1.4.13.jar
hbase-1.4.13/lib/metrics-core-3.1.2.jar
hbase-1.4.13/lib/hbase-resource-bundle-1.4.13.jar
hbase-1.4.13/lib/jackson-mapper-asl-1.9.13.jar
hbase-1.4.13/lib/jackson-core-asl-1.9.13.jar
hbase-1.4.13/lib/hbase-hbtop-1.4.13.jar
hbase-1.4.13/lib/commons-lang3-3.8.1.jar
hbase-1.4.13/lib/findbugs-annotations-1.3.9-1.jar
hbase-1.4.13/lib/log4j-1.2.17.jar
hbase-1.4.13/lib/junit-4.12.jar
hbase-1.4.13/lib/jasper-compiler-5.5.23.jar
hbase-1.4.13/lib/commons-el-1.0.jar
hbase-1.4.13/lib/jasper-runtime-5.5.23.jar
hbase-1.4.13/lib/api-i18n-1.0.0-M20.jar
hbase-1.4.13/lib/spymemcached-2.11.6.jar
[root@master ~]# sudo chmod -R 0777 hbase-1.4.13
[root@master ~]# sudo cp -rf hbase-1.4.13  /home/hbase
[root@master ~]#
```

P12 Hbase setup



```
[root@master hbase]# vi /etc/profile
[root@master hbase]# vi hbase-1.4.13/conf/hbase-enc.sh
[root@master hbase]# ping slave3
PING slave3 (192.168.2.133) 56(84) bytes of data.
64 bytes from slave3 (192.168.2.133): icmp_seq=1 ttl=64 time=0.155 ms
64 bytes from slave3 (192.168.2.133): icmp_seq=2 ttl=64 time=0.298 ms
^C
--- slave3 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1000ms
rtt min/avg/max/mdev = 0.155/0.226/0.298/0.073 ms
[root@master hbase]# exit
登出
Connection to 192.168.2.130 closed.
PS C:\Users\Administrator> ssh root@192.168.2.131
root@192.168.2.131's password:
Last login: Mon Mar  1 14:47:27 2021
[root@slave1 ~]# vi /etc/profile
[root@slave1 ~]# source /etc/profile
[root@slave1 ~]# hbase version
-bash: hbase: 未找到命令
[root@slave1 ~]# exit
登出
Connection to 192.168.2.131 closed.
PS C:\Users\Administrator> ssh root@192.168.2.132
root@192.168.2.132's password:
Last login: Mon Mar  1 14:47:31 2021
[root@slave2 ~]# vi /etc/profile
[root@slave2 ~]# source /etc/profile
[root@slave2 ~]# exit
登出
Connection to 192.168.2.132 closed.
PS C:\Users\Administrator> ssh root@192.168.2.133
root@192.168.2.133's password:
Last login: Mon Mar  1 14:47:39 2021
[root@slave3 ~]# vi /etc/profile
[root@slave3 ~]# source /etc/profile
[root@slave3 ~]#
```

P13 Environment configuration in master and slaves

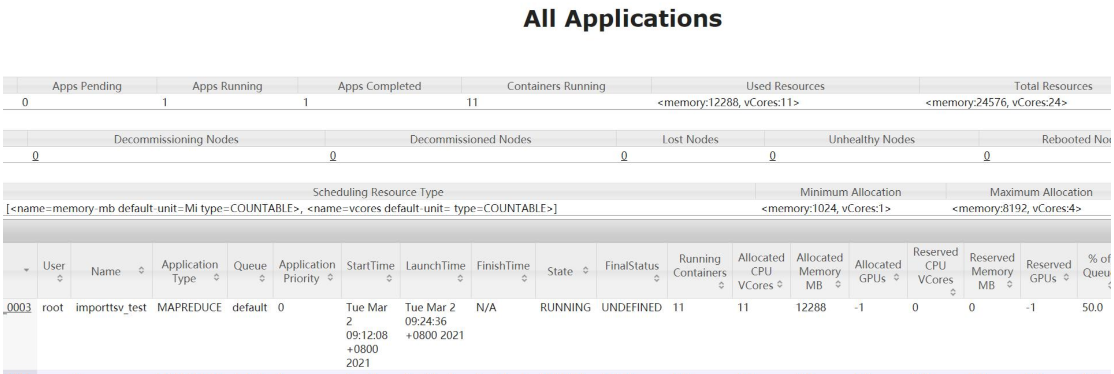


```
登出
Connection to 192.168.2.132 closed.
PS C:\Users\Administrator> ssh root@192.168.2.130
root@192.168.2.130's password:
Last login: Mon Mar  1 15:49:43 2021 from 192.168.2.1
[root@master ~]# start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [master]
master: namenode running as process 1322. Stop it first.
slave2: datanode running as process 1215. Stop it first.
slave3: datanode running as process 1213. Stop it first.
slave1: datanode running as process 1217. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 1524. Stop it first.
starting yarn daemons
resourcemanager running as process 1683. Stop it first.
slave3: nodemanager running as process 1330. Stop it first.
slave1: nodemanager running as process 1334. Stop it first.
slave2: nodemanager running as process 1332. Stop it first.
[root@master ~]# mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /usr/hadoop/hadoop-2.10.1/logs/mapred-root-hist
[root@master ~]# start-hbase.sh
OpenJDK 64-Bit Server VM warning: If the number of processors is expected to incre
e the number of parallel GC threads appropriately using -XX:ParallelGCThreads=N
running master, logging to /home/hbase/hbase-1.4.13/logs/hbase-root-master-master.
OpenJDK 64-Bit Server VM warning: ignoring option PermSize=128m; support was remov
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128m; support was re
OpenJDK 64-Bit Server VM warning: If the number of processors is expected to incre
e the number of parallel GC threads appropriately using -XX:ParallelGCThreads=N
[root@master ~]# jps
3136 HMaster
1683 ResourceManager
2963 JobHistoryServer
1524 SecondaryNameNode
1322 NameNode
3405 Jps
[root@master ~]#
```

P14 Hbase cluster started successfully

ImportTsv script:

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv '-Dimporttsv.separator= '
-Dimporttsv.bulk.output=/output/hfile
-Dimporttsv.columns=HBASE_ROW_KEY,cf:a test      /bigrams/table2
```



All Applications

Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
0	1	1	11	<memory:12288, vCores:11>	<memory:24576, vCores:24>

Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Noc
0	0	0	0	0

Scheduling Resource Type	Minimum Allocation	Maximum Allocation
[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

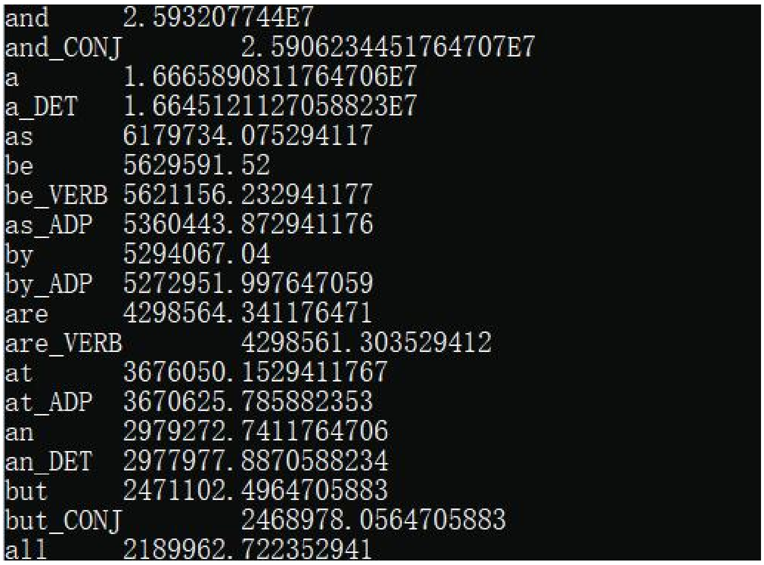
	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queu
_0003	root	importtsv_test	MAPREDUCE	default	0	Tue Mar 2 09:12:08 +0800 2021	Tue Mar 2 09:24:36 +0800 2021	N/A	RUNNING	UNDEFINED	11	11	12288	-1	0	0	-1	50.0

P15 ImportTsv mission on yarn

Hbase shell script:

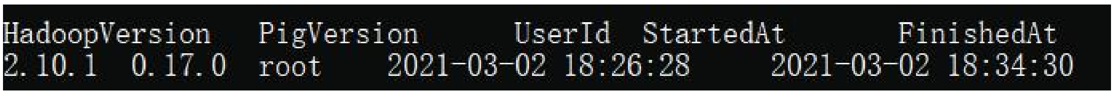
```
create 'table2', {NAME => 'bigram'}, {NAME => 'year'}, {NAME => 'match_count'},{NAME => 'volume_count'}
describe 'table2'
put 'table2', 'ierg4330', '2019', '100','4'
scan 'table2', FILTER=>"ValueFilter(=, 'binary:1671')"
```

Bonus: Setup Pig on hadoop over Kubernetes



```
and       2.593207744E7
and_CONJ          2.5906234451764707E7
a         1.6665890811764706E7
a_DET     1.6645121127058823E7
as        6179734.075294117
be        5629591.52
be_VERB   5621156.232941177
as_ADP    5360443.872941176
by        5294067.04
by_ADP    5272951.997647059
are       4298564.341176471
are_VERB          4298561.303529412
at        3676050.1529411767
at_ADP    3670625.785882353
an        2979272.7411764706
an_DET    2977977.8870588234
but       2471102.4964705883
but_CONJ          2468978.0564705883
all       2189962.722352941
```

P16 Pig result



```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt
2.10.1  0.17.0  root    2021-03-02 18:26:28     2021-03-02 18:34:30
```

P17 Running time

In this part, the running time of the same script is 8 minutes and 2 seconds, compared to the running time in Q1(45 minutes). It’s much less in the pig via Kubernetes.