

Big Data Analytics Homework1 uni: xw2401 name: Xucan Wang

Part 1: PIG Usage

1) Start Hadoop

```
wangxucan — ssh localhost — 85x28
Starting namenodes on [localhost]
Password:
localhost: starting namenode, logging to /usr/local/Cellar/hadoop/2.7.3/libexec/logs/hadoop-wangxucan-namenode-dyn-209-2-209-185.dyn.columbia.edu.out
Password:
localhost: starting datanode, logging to /usr/local/Cellar/hadoop/2.7.3/libexec/logs/hadoop-wangxucan-datanode-dyn-209-2-209-185.dyn.columbia.edu.out
Starting secondary namenodes [0.0.0.0]
Password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cellar/hadoop/2.7.3/libexec/logs/hadoop-wangxucan-secondarynamenode-dyn-209-2-209-185.dyn.columbia.edu.out
16/09/27 13:32:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
[starting resourcemanager, logging to /usr/local/Cellar/hadoop/2.7.3/libexec/logs/yarn-wangxucan-resourcemanager-dyn-209-2-209-185.dyn.columbia.edu.out]
Password:
localhost: starting nodemanager, logging to /usr/local/Cellar/hadoop/2.7.3/libexec/logs/yarn-wangxucan-nodemanager-dyn-209-2-209-185.dyn.columbia.edu.out
wangxucan (master) 2.7.3 $ jps
688 NameNode
1016 ResourceManager
1114 NodeManager
779 DataNode
1150 Jps
895 SecondaryNameNode
[wangxucan (master) 2.7.3 $ ]
```

1C) Download Sample Dataset

wget https://github.com/hortonworks/tutorials/raw/hdp-2.5/driver_data.zip

unzip driver_data.zip

```
2.7.3 — -bash — 80x24
[wangxucan (master) 2.7.3 $ ls
INSTALL_RECEIPT.json  driver_data  pig_1474917874955.log
LICENSE.txt           hillary.txt  pig_1474919547395.log
NOTICE.txt            libexec      puzzle1.dta
README.txt            pig-0.16.0   sbin
__MACOSX              pig-0.16.0.tar.gz
bin                   pig_1474915113546.log
```

```
2.7.3 — ssh localhost — 80x24
[wangxucan (master) 2.7.3 $ ./bin/hadoop fs -mkdir /user/pig_example
16/09/27 13:42:45 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
```

```
2.7.3 — ssh localhost — 80x24
[wangxucan (master) 2.7.3 $ ./bin/hadoop fs -put driver_data/truck_event_text_par
tition.csv /user/pig_example
16/09/27 13:43:38 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
```

2a) Load truck data and define a schema to use in PIG

```
truck_events = LOAD '/user/pig_example/truck_event_text_partition.csv' USING PigStorage(',')
AS (driverId:int, truckId:int, eventTime:chararray,
eventTime:chararray, longitude:double, latitude:double,
eventKey:chararray, correlationId:long, driverName:chararray,
routeId:long,routeName:chararray,eventDate:chararray);
DESCRIBE truck_events;
```

```
2.7.3 — ssh localhost — 88x26
2016-09-27 13:46:40,889 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to
load native-hadoop library for your platform... using builtin-java classes where applic
able
2016-09-27 13:46:40,911 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-09-27 13:46:40,911 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
fs.default.name is deprecated. Instead, use fs.defaultFS
2016-09-27 13:46:40,911 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExec
utionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2016-09-27 13:46:41,675 [main] INFO org.apache.pig.PigServer - Pig Script ID for the se
ssion: PIG-default-91eafdf5-13ac-4714-af90-4c81a073d178
2016-09-27 13:46:41,675 [main] WARN org.apache.pig.PigServer - ATS is disabled since ya
rn.timeline-service.enabled set to false
grunt> truck_events = LOAD '/user/pig_example/truck_event_text_partition.csv' USING PigS
torage(',')
>> AS (driverId:int, truckId:int, eventTime:chararray,
>> eventTime:chararray, longitude:double, latitude:double,
>> eventKey:chararray, correlationId:long, driverName:chararray,
>> routeId:long,routeName:chararray,eventDate:chararray);
2016-09-27 13:46:58,952 [main] INFO org.apache.hadoop.conf.Configuration.deprecation -
fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> DESCRIBE truck_events;
[truck_events: {driverId: int,truckId: int,eventTime: chararray,eventType: chararray,long
itude: double,latitude: double,eventKey: chararray,correlationId: long,driverName: chara
rray,routeId: long,routeName: chararray,eventDate: chararray}]
```

2b) Load truck data and define a schema to use in PIG

```
truck_events_subset = LIMIT truck_events 100;
DESCRIBE truck_events_subset;
```

```
[grunt> truck_events_subset = LIMIT truck_events 100;
[grunt> DESCRIBE truck_events_subset;
truck_events_subset: {driverId: int,truckId: int,eventTime: chararray,eventType: chararr
ay,longitude: double,latitude: double,eventKey: chararray,correlationId: long,driverName
: chararray,routeId: long,routeName: chararray,eventDate: chararray}
grunt> ]
```

2c) Describe a subset of the data (specific columns)

```
specific_columns = FOREACH truck_events_subset GENERATE driverId, eventTime, eventType;
```

```
DESCRIBE specific_columns;
```

```
[grunt> specific_columns = FOREACH truck_events_subset GENERATE driverId, eventTime, eventType;
[grunt> DESCRIBE specific_columns;
specific_columns: {driverId: int,eventTime: chararray,eventType: chararray}
grunt> ]
```

2d) Store a subset into an HDFS location

```
STORE specific_columns INTO '/user/pig_example' USING PigStorage(',');
```

2e) Perform a join using multiple tables

```
grunt> truck_events = LOAD '/user/pig_example/truck_event_text_partition.csv' USING PigStorage(',')
>> AS (driverId:int, truckId:int, eventTime:chararray,
>> eventType:chararray, longitude:double, latitude:double,
>> eventKey:chararray, correlationId:long, driverName:chararray,
>> routeId:long,routeName:chararray,eventDate:chararray);
2016-09-27 14:03:28,676 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> drivers = LOAD '/user/pig_example/drivers.csv' USING PigStorage(',')
>> AS (driverId:int, name:chararray, ssn:chararray,
>> location:chararray, certified:chararray, wage_plan:chararray);
2016-09-27 14:03:39,997 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> join_data = JOIN truck_events BY (driverId), drivers BY (driverId);
grunt> DESCRIBE join_data;
join_data: {truck_events::driverId: int,truck_events::truckId: int,truck_events::eventTime: chararray,truck_events::eventType: chararray,truck_events::longitude: double,truck_events::latitude: double,truck_events::eventKey: chararray,truck_events::correlationId: long,truck_events::driverName: chararray,truck_events::routeId: long,truck_events::routeName: chararray,truck_events::eventDate: chararray,drivers::driverId: int,drivers::name: chararray,drivers::ssn: chararray,drivers::location: chararray,drivers::certified: chararray,drivers::wage_plan: chararray}
```

2f) Perform a calculation using GROUP BY

```
grunt> truck_events = LOAD '/user/pig_example/truck_event_text_partition.csv' USING PigStorage(',')
>> AS (driverId:int, truckId:int, eventTime:chararray,
>> eventType:chararray, longitude:double, latitude:double,
>> eventKey:chararray, correlationId:long, driverName:chararray,
>> routeId:long,routeName:chararray,eventDate:chararray);
2016-09-27 14:04:56,495 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> filtered_events = FILTER truck_events BY NOT (eventType MATCHES 'Normal');
grunt> grouped_events = GROUP filtered_events BY driverId;
grunt> DESCRIBE grouped_events;
grouped_events: {group: int,filtered_events: {(driverId: int,truckId: int,eventTime: chararray,eventType: chararray,longitude: double,latitude: double,eventKey: chararray,correlationId: long,driverName: chararray,routeId: long,routeName: chararray,eventDate: chararray)}}
```

```
DUMP grouped_events;
```



```
2.7.3 — ssh localhost — 88x26
, FILTER

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime M
axReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime  Alias  Feature0
utputs
job_local382486062_0001 1 1 n/a n/a n/a n/a n/a n/a n
/a n/a filtered_events,grouped_events,truck_events GROUP_BY hdfs://l
ocalhost:9000/tmp/temp-1258169098/tmp31616167,

Input(s):
Successfully read 17076 records (15300568 bytes) from: "/user/pig_example/truck_event_t
xt_partition.csv"

Output(s):
Successfully stored 24 records (10762057 bytes) in: "hdfs://localhost:9000/tmp/temp-1258
169098/tmp31616167"

Counters:
Total records written : 24
Total bytes written : 10762057
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
2.7.3 — ssh localhost — 167x52
ized
2016-09-27 14:08:09,070 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSI
ON_FAILED 6 time(s).
2016-09-27 14:08:09,070 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-09-27 14:08:09,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-09-27 14:08:09,073 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2016-09-27 14:08:09,079 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-09-27 14:08:09,080 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10,{{(10,85,00:13.1,Unsafe tail distance,-91.18,38.22,10|85|9223370572464762694,3660000000000000000,George Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22),(1
0,85,00:39.7,Overspeed,-94.23,37.09,10|85|9223370572464736126,3660000000000000000,George Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22),(10,85,59:46.9,Overs
peed,-95.5,36.37,10|85|9223370572464788896,3660000000000000000,George Vetticaden,1390372503,Saint Louis to Tulsa,2016-05-27-22)}}))
(11,{{(11,74,00:14.1,Lane Departure,-88.77,40.76,11|74|9223370572464761716,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(1
1,74,00:41.0,Lane Departure,-90.07,35.68,11|74|9223370572464734786,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,00
:85.4,Unsafe following distance,-89.74,39.1,11|74|9223370572464770396,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,59:56
.4,Lane Departure,-87.67,41.87,11|74|9223370572464779456,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,59:38.0,Unsaf
e tail distance,-89.17,40.38,11|74|9223370572464797796,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,59:47.3,Unsaf
e tail distance,-89.63,39.84,11|74|9223370572464788546,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,59:29.1,Overspe
ed,-88.07,41.48,11|74|9223370572464886746,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,00:32.0,Unsafe tail distanc
e,-90.2,38.65,11|74|9223370572464743846,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22),(11,74,00:23.1,Unsafe tail distance,
-88.42,41.11,11|74|9223370572464752715,3660000000000000000,Jamie Engesser,1567254452,Saint Louis to Memphis Route2,2016-05-27-22)}}))
(12,{{(12,104,00:47.6,Unsafe following distance,-90.0,37.72,12|104|9223370572464728186,3660000000000000000,Paul Coddling,24929475,Peoria to Cedar Rapids,2016-05-27-22)}})
(13,{{(13,89,00:47.7,Lane Departure,-89.03,41.92,13|89|9223370572464728156,3660000000000000000,Joe Niemiec,927636994,Des Moines to Chicago.kml,2016-05-27-22)}})
(14,{{(14,25,00:48.4,Unsafe following distance,-91.63,41.72,14|25|9223370572464727394,3660000000000000000,Adis Cesir,160405074,Joplin to Kansas City Route 2,2016-05-27-22)}})
(15,{{(15,51,00:48.8,Lane Departure,-90.04,35.19,15|51|9223370572464727025,3660000000000000000,Rohit Bakshi,1384345811,Joplin to Kansas City,2016-05-27-22)}})
(16,{{(16,12,00:48.9,Lane Departure,-89.52,40.7,16|12|9223370572464726925,3660000000000000000,Tom McCuch,1961634315,Saint Louis to Memphis,2016-05-27-22)}})
(17,{{(17,15,00:48.4,Lane Departure,-90.79,38.83,17|15|9223370572464727374,3660000000000000000,Eric Mizell,1927624662,Springfield to KC Via Columbia,2016-05-27-22)}})
(18,{{(18,16,00:47.2,Overspeed,-94.28,39.53,18|16|9223370572464728575,3660000000000000000,Grant Liu,1565885487,Springfield to KC Via Hanibal,2016-05-27-22)}})
(19,{{(19,26,00:48.6,Unsafe following distance,-94.57,35.37,19|26|9223370572464727224,3660000000000000000,Ajay Singh,1962261785,Wichita to Little Rock.kml,2016-05-27-22)}})
(20,{{(20,41,00:46.9,Overspeed,-89.03,41.92,20|41|9223370572464728915,3660000000000000000,Chris Harris,160779139,Des Moines to Chicago Route 2,2016-05-27-22)}})
(21,{{(21,109,00:46.8,Unsafe tail distance,-88.07,41.48,21|109|9223370572464729016,3660000000000000000,Jeff Markham,1594289134,Memphis to Little Rock Route 2,2016-05-27-22)}})
(22,{{(22,87,00:46.5,Unsafe tail distance,-90.04,35.19,22|87|9223370572464729286,3660000000000000000,Nadeem Asghar,1198242881,Saint Louis to Chicago Route2,2016-05-27-22)}})
(23,{{(23,68,00:47.8,Lane Departure,-89.52,40.7,23|68|9223370572464727994,3660000000000000000,Adam Diaz,160405074,Joplin to Kansas City Route 2,2016-05-27-22)}})
(24,{{(24,97,00:48.6,Lane Departure,-89.17,40.38,24|97|9223370572464727226,3660000000000000000,Don Hilborn,1090292248,Peoria to Cedar Rapids Route 2,2016-05-27-22)}})
(25,{{(25,96,00:40.1,Overspeed,-89.54,36.84,25|96|9223370572464735726,3660000000000000000,Jean-Philippe Player,371182829,Memphis to Little Rock,2016-05-27-22)}})
(26,{{(26,57,00:48.8,Overspeed,-95.99,36.17,26|57|9223370572464727046,3660000000000000000,Michael Aube,1325712174,Saint Louis to Tulsa Route2,2016-05-27-22)}})
(27,{{(27,105,00:48.0,Unsafe following distance,-90.79,38.83,27|105|9223370572464727846,3660000000000000000,Mark Lochbihler,1325562373,Springfield to KC Via Columbia Ro
ute 2,2016-05-27-22)}})
(28,{{(28,39,00:47.5,Overspeed,-94.28,39.53,28|39|9223370572464728273,3660000000000000000,Olivier Renault,137128276,Springfield to KC Via Hanibal Route 2,2016-05-27-22)}})
(29,{{(29,66,00:47.8,Overspeed,-94.57,35.37,29|66|9223370572464728016,3660000000000000000,Teddy Choi,803014426,Wichita to Little Rock Route 2,2016-05-27-22)}})
(30,{{(30,58,00:49.3,Unsafe following distance,-89.03,41.92,30|58|9223370572464726546,3660000000000000000,Dan Rice,160779139,Des Moines to Chicago Route 2,2016-05-27-22)}})
(31,{{(31,18,00:47.6,Lane Departure,-88.07,41.48,31|18|9223370572464728166,3660000000000000000,Rommel Garcia,1594289134,Memphis to Little Rock Route 2,2016-05-27-22)}})
(32,{{(32,42,00:48.7,Unsafe following distance,-90.04,35.19,32|42|9223370572464727106,3660000000000000000,Ryan Templeton,1090292248,Peoria to Cedar Rapids Route 2,2016-05-27-22)}})
( // eventTime eventTune eventKwv driverName routeName eventDate))
```

Part 2: Hive

1) Using "SHOW TABLES" to see if any tables exist:

```
hive> SHOW TABLES;
```

2) Create a table:

```
hive> CREATE TABLE test (name STRING, gender STRING, year INT, month INT);
```

```
hive> SHOW TABLES;
```

```
hive> SELECT * FROM test;
```

```
hive> quit;
```



The screenshot shows a terminal window titled "wangxucan — ssh localhost — 80x24". The output of the Hive commands is as follows:

```
adoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/Cellar/hive/2.1.0/
/libexec/lib/hive-common-2.1.0.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
[hive> SHOW TABLES;
OK
Time taken: 2.016 seconds
[hive> CREATE TABLE test(name STRING,gender STRING,year INT,month INT);
OK
Time taken: 0.634 seconds
[hive> SHOW TABLES;
OK
test
Time taken: 0.051 seconds, Fetched: 1 row(s)
[hive> SELECT * FROM test;
OK
Time taken: 3.17 seconds
hive> █
```

3) Practical Example

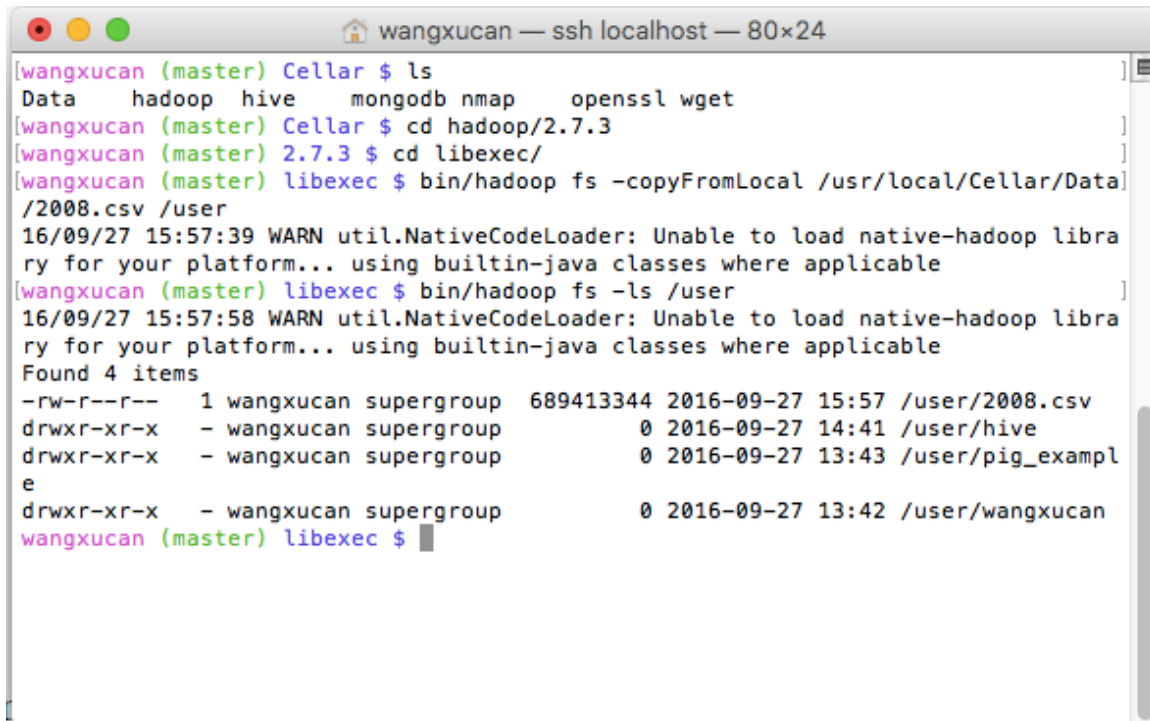
According to the data in the 2008.csv, we create a table with the following attributes

```
hive> create table airline (
  > Year int,
  > Month int,
  > DayofMonth int,
  > DayOfWeek int,
  > DepTime int,
  > CRSDepTime int,
  > ArrTime int,
  > CRSArrTime int,
  > UniqueCarrier varchar(5),
  > FlightNum int,
```

```

> TailNum varchar(8),
> ActualElapsedTime int,
> CRSElapsedTime int,
> AirTime int,
> ArrDelay int,
> DepDelay int,
> Origin varchar(3),
> Dest varchar(3),
> Distance int,
> TaxiIn int,
> TaxiOut int,
> Cancelled int,
> CancellationCode varchar(1),
> Diverted varchar(1),
> CarrierDelay int,
> WeatherDelay int,
> NASDelay int,
> SecurityDelay int,
> LateAircraftDelay int
> )
> row format delimited fields terminated by ',' stored as
textfile;
Upload dataset to HDFS

```



A terminal window titled 'wangxucan — ssh localhost — 80x24' showing the following commands and output:

```

[wangxucan (master) Cellar] $ ls
Data  hadoop  hive    mongodb  nmap    openssl  wget
[wangxucan (master) Cellar] $ cd hadoop/2.7.3
[wangxucan (master) 2.7.3] $ cd libexec/
[wangxucan (master) libexec] $ bin/hadoop fs -copyFromLocal /usr/local/Cellar/Data/
/2008.csv /user
16/09/27 15:57:39 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[wangxucan (master) libexec] $ bin/hadoop fs -ls /user
16/09/27 15:57:58 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r--  1 wangxucan supergroup  689413344 2016-09-27 15:57 /user/2008.csv
drwxr-xr-x  - wangxucan supergroup           0 2016-09-27 14:41 /user/hive
drwxr-xr-x  - wangxucan supergroup           0 2016-09-27 13:43 /user/pig_exempl
e
drwxr-xr-x  - wangxucan supergroup           0 2016-09-27 13:42 /user/wangxucan
[wangxucan (master) libexec] $

```

```
wangxucan — ssh localhost — 110x33
hive> DESCRIBE airline;
OK
year                int
month               int
dayofmonth           int
dayofweek            int
deptime             int
crsdeptime           int
arrtime             int
crsarrrtime         int
uniquecarrier        varchar(5)
flightnum            int
tailnum              varchar(8)
actualelapsedtime    int
crselapsedtime       int
airtime             int
arrdelay            int
depdelay            int
origin              varchar(3)
dest                varchar(3)
distance            int
taxiin              int
taxiout             int
cancelled            int
cancellationcode     varchar(1)
diverted             varchar(1)
carrierdelay         int
weatherdelay         int
nasdelay            int
securitydelay        int
lateaircraftdelay    int
Time taken: 0.061 seconds, Fetched: 29 row(s)
```

Load data into the database

```
hive> LOAD DATA INPATH '/user/2008.csv' INTO TABLE airline;
Loading data to table default.airline
OK
Time taken: 0.374 seconds
```

Queries:

1) **SELECT * FROM airline LIMIT 10;**

```
wangxucan — ssh localhost — 171x33
weatherdelay        int
nasdelay            int
securitydelay        int
lateaircraftdelay    int
Time taken: 0.061 seconds, Fetched: 29 row(s)
hive> LOAD DATA INPATH '/user/2008.csv' INTO TABLE airline;
Loading data to table default.airline
OK
Time taken: 0.374 seconds
hive> SELECT * FROM airline LIMIT 10;
OK


| year | month | dayofmonth | dayofweek | deptime | crsdeptime | arrtime | crsarrrtime | uniquecarrier | flightnum | tailnum | actualelapsedtime | crselapsedtime | airtime | arrdelay | depdelay | origin | dest | distance | taxiin | taxiout | cancelled | cancellationcode | diverted | carrierdelay | weatherdelay | nasdelay | securitydelay | lateaircraftdelay |
|------|-------|------------|-----------|---------|------------|---------|-------------|---------------|-----------|---------|-------------------|----------------|---------|----------|----------|--------|------|----------|--------|---------|-----------|------------------|----------|--------------|--------------|----------|---------------|-------------------|
| 2008 | 1     | 3          | 4         | 2003    | 1955       | 2211    | 2225        | WN            | 335       | N712SW  | 128               | 150            | 116     | -14      | 8        | IAD    | TPA  | 810      | 4      | 8       | 0         |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 754     | 735        | 1002    | 1000        | WN            | 3231      | N772SW  | 128               | 145            | 113     | 19       | IAD      | TPA    | 810  | 5        | 10     | 0       |           |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 628     | 620        | 804     | 750         | WN            | 448       | N428WN  | 96                | 90             | 76      | 14       | 8        | IND    | BWI  | 515      | 3      | 17      | 0         |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 926     | 930        | 1054    | 1100        | WN            | 1746      | N612SW  | 88                | 90             | 78      | -6       | -4       | IND    | BWI  | 515      | 3      | 7       | 0         |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 1829    | 1755       | 1959    | 1925        | WN            | 3920      | N464WN  | 90                | 90             | 77      | 34       | 34       | IND    | BWI  | 515      | 3      | 10      | 0         |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 1940    | 1915       | 2121    | 2110        | WN            | 378       | N726SW  | 101               | 115            | 87      | 11       | 25       | IND    | JAX  | 680      | 4      | 10      | 0         |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 1937    | 1830       | 2037    | 1940        | WN            | 509       | N763SW  | 240               | 250            | 230     | 57       | 67       | IND    | LAS  | 1591     | 3      | 7       | 0         |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 1039    | 1040       | 1132    | 1150        | WN            | 535       | N428WN  | 233               | 250            | 219     | -18      | -1       | IND    | LAS  | 1591     | 7      | 7       | 0         |                  |          |              |              |          |               |                   |
| 2008 | 1     | 3          | 4         | 617     | 615        | 652     | 650         | WN            | 11        | N689SW  | 95                | 95             | 70      | IND      | MCI      | 451    | 6    | 19       | 0      |         |           |                  |          |              |              |          |               |                   |


Time taken: 0.259 seconds, Fetched: 10 row(s)
hive>
```

2) SELECT * FROM airline WHERE dest='TPA'&&distance>1000;
 SELECT * FROM airline WHERE dest='TPA' AND distance>1000;

wangxucan — ssh localhost — 171x33

9	16	0	0																		
2008	12	12	5	705	710	1048	1034	DL	1449	N973DL	223	204	193	14	-5	BOS	TPA	1185	4	26	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	12	5	1507	1445	1853	1807	DL	1451	N913DL	226	202	196	46	22	BOS	TPA	1185	4	26	0 2
4	0	14																			
2008	12	12	5	1051	1040	1800	1817	DL	1478	N371DA	249	277	225	-17	11	LAX	TPA	2158	5	19	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	12	5	2217	2220	529	555	DL	1480	N371DA	252	275	225	-26	-3	LAX	TPA	2158	4	23	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	12	5	730	735	1051	1031	DL	1503	N965DL	201	176	165	20	-5	LGA	TPA	1011	3	33	0 2
0	0	0																			
2008	12	12	5	1138	1135	1448	1442	DL	1505	N907DL	190	187	158	6	3	LGA	TPA	1011	4	28	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	12	5	1432	1430	1735	1735	DL	1507	N951DL	183	185	163	0	2	LGA	TPA	1011	4	16	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	12	5	1948	1930	2243	2237	DL	1509	N907DL	175	187	153	6	18	LGA	TPA	1011	4	18	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	13	6	938	935	1544	1545	DL	1162	N608DA	246	250	198	-1	3	SLC	TPA	1887	6	42	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	13	6	734	740	1024	1054	DL	1417	N968DL	170	194	157	-30	-6	BDL	TPA	1111	3	10	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	13	6	707	715	1018	1047	DL	1449	N999DN	191	212	168	-29	-8	BOS	TPA	1185	3	20	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	13	6	1030	1045	1743	1815	DL	1478	N3767	245	270	226	-32	-7	LAX	TPA	2158	5	14	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	13	6	753	735	1054	1043	DL	1503	N951DL	181	188	143	11	18	LGA	TPA	1011	5	33	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	13	6	1105	1115	1347	1431	DL	1505	N943DL	162	196	146	-44	-10	LGA	TPA	1011	4	12	0 N
ULL	NULL	NULL	NULL	NULL	NULL																
2008	12	13	6	1423	1430	1705	1744	DL	1507	N916DL	162	194	146	-39	-7	LGA	TPA	1011	5	11	0 N
ULL	NULL	NULL	NULL	NULL	NULL																

Time taken: 0.235 seconds, Fetched: 22476 row(s)

3) hive> SELECT avg(DepTime) FROM airline WHERE Month=6;

```

[hive> SELECT avg(DepTime) FROM airline WHERE Month=6;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = wangxucan_20160927212336_e1e324f4-f8aa-4006-9ca3-e85d4fb132bd
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2016-09-27 21:23:40,553 Stage-1 map = 0%, reduce = 0%
2016-09-27 21:23:43,577 Stage-1 map = 100%, reduce = 0%
2016-09-27 21:23:48,637 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local780972079_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 7700439632 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1338.6650198917914
Time taken: 12.491 seconds, Fetched: 1 row(s)

```

Part3: Hbase

1) Install Hbase
 brew install hbase


```
wangxucan ~ $ brew install hbase
sed: .git/GITHUB_HEADERS: No such file or directory
==> Auto-updated Homebrew!
Updated 1 tap (homebrew/core).
==> New Formulae
compose2kube
==> Updated Formulae
awscli      gssh      launch     redis
bind        harfbuzz  lean-cli   rtv
bmake       httpstat  libphonenum tbb
certigo     jetty     mpfr       varnish
devd        jruby     node       wellington
ejabberd    jsonschema2pojo nsd        youtube-dl
fobis       kapacitor ponyc
==> Deleted Formulae
caudec

==> Migrating HOMEBREW_REPOSITORY (please wait)...
==> Migrated HOMEBREW_REPOSITORY to /usr/local/Homebrew!
Homebrew no longer needs to have ownership of /usr/local. If you wish you can
return /usr/local to its default ownership with:
  sudo chown root:wheel /usr/local
==> Installing dependencies for hbase: lzo
==> Installing hbase dependency: lzo
==> Downloading https://homebrew.bintray.com/bottles/lzo-2.09.el_capitan.bottle.
##### 100.0%
==> Pouring lzo-2.09.el_capitan.bottle.tar.gz
📦 /usr/local/Cellar/lzo/2.09: 29 files, 565.0K
==> Installing hbase
==> Downloading https://homebrew.bintray.com/bottles/hbase-1.2.2.el_capitan.bott
##### 100.0%
==> Pouring hbase-1.2.2.el_capitan.bottle.tar.gz
==> Using the sandbox
==> Caveats
To have launchd start hbase now and restart at login:
  brew services start hbase
Or, if you don't want/need a background service you can just run:
  /usr/local/opt/hbase/bin/start-hbase.sh
==> Summary
📦 /usr/local/Cellar/hbase/1.2.2: 9,806 files, 332.2M
wangxucan ~ $
```

b) Configuring Hbase by modifying 2 files

hbase-env.sh

```
export JAVA_HOME="/usr/libexec/java_home"
```

```
change      HBASE_OPTS="$HBASE_OPTS-Djava.net.preferIPv4Stack=true      —
```

```
Djava.security.krb5.realm= -Djava.security.krb5.kdc=
```

hbase-site.xml

```
<configuration>
```

```
  <property>
```

```
    <name>hbase.rootdir</name>
```

```
    <value>hdfs://localhost:9000/hbase</value>
```

```
  </property>
```

```
</configuration>
```

c) run Hbase

```
./start-hbase.sh
```

```
wangxucan bin $ ./start-hbase.sh
starting master, logging to /usr/local/var/log/hbase/hbase-wangxucan-master-WangXucandeMacBook-Air.local.out
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option PermSize=128m; support was removed in 8.0
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=128m; support was removed in 8.0
wangxucan bin $ jps
1168 NodeManager
1312 HMaster
821 DataNode
727 NameNode
1368 Jps
1066 ResourceManager
941 SecondaryNameNode
wangxucan bin $
```

c) Launch the Hbase shell

./hbase shell

```
bin — java -Dproc_shell -XX:OnOutOfMemoryError=kill -9 %p -Djava.net.preferIPv4Stack=true -Djava.security.krb5.realm=-...
[wangxucan 1.2.2 $ cd bin
[wangxucan bin $ ./start-hbase.sh
starting master, logging to /usr/local/var/log/hbase/hbase-wangxucan-master-WangXucandeMacBook-Air.local.out
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option PermSize=128m; support was removed in 8.0
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=128m; support was removed in 8.0
[wangxucan bin $ jps
1168 NodeManager
1312 HMaster
821 DataNode
727 NameNode
1368 Jps
1066 ResourceManager
941 SecondaryNameNode
[wangxucan bin $ ./hbase shell
2016-09-28 20:39:51,253 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform..
. using builtin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/Cellar/hbase/1.2.2/libexec/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/
StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/Cellar/hadoop/2.7.3/libexec/share/hadoop/common/lib/slf4j-log4j12-1.7.1
0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.2, r3f671c1ead70d249ea4598f1bbcc5151322b3a13, Fri Jul 1 08:28:55 CDT 2016

hbase(main):001:0>
```

d) Query Practice

create a table named "product"

```
[hbase(main):007:0> create 'product','name','price'
0 row(s) in 1.2540 seconds

=> Hbase::Table - product
[hbase(main):008:0> list
TABLE
product
1 row(s) in 0.0100 seconds

=> ["product"]
```

add some data into the table and scan the table

```
bin — java -Dproc_shell -XX:OnOutOfMemoryError=kill -9 %p -Djava.net.preferIPv4S...
[
=> ["product"]
hbase(main):009:0> put 'product','row1','name:bottle','price:15'
0 row(s) in 0.0210 seconds

hbase(main):010:0> put 'product','row2','name:lamp','price:20'
0 row(s) in 0.0130 seconds

hbase(main):011:0> put 'product','row3','name:book','price:10'
0 row(s) in 0.0130 seconds

hbase(main):012:0> put 'product','row4','name:mirror','price:5'
0 row(s) in 0.0140 seconds

hbase(main):013:0> scan 'product'
[ROW          COLUMN+CELL
  row1         column=name:bottle, timestamp=1475112167406, value=price:15
  row2         column=name:lamp, timestamp=1475112185525, value=price:20
  row3         column=name:book, timestamp=1475112204240, value=price:10
  row4         column=name:mirror, timestamp=1475112237558, value=price:5
4 row(s) in 0.0390 seconds]
```

retrieve the specific row in the table

```
[hbase(main):014:0> get 'product','row2'
COLUMN      CELL
  name:lamp   timestamp=1475112185525, value=price:20
1 row(s) in 0.0970 seconds]
```

disable the table

```
[hbase(main):015:0> disable 'product'
0 row(s) in 2.2850 seconds

[hbase(main):016:0> scan 'product'
ROW          COLUMN+CELL

ERROR: product is disabled.

Here is some help for this command:
```

delete the table

```
[hbase(main):017:0> drop 'product'
0 row(s) in 1.2570 seconds

[hbase(main):018:0> list
TABLE
0 row(s) in 0.0040 seconds

=> []
hbase(main):019:0> █
```