MBD512 Assignment (Apache Pig)

Samiksha Agarwal

August 21, 2018

1 Make hdfs Directory

We will first create a directory on `/user/curaj/pig` on hdfs and put our dataset there for forther computation.

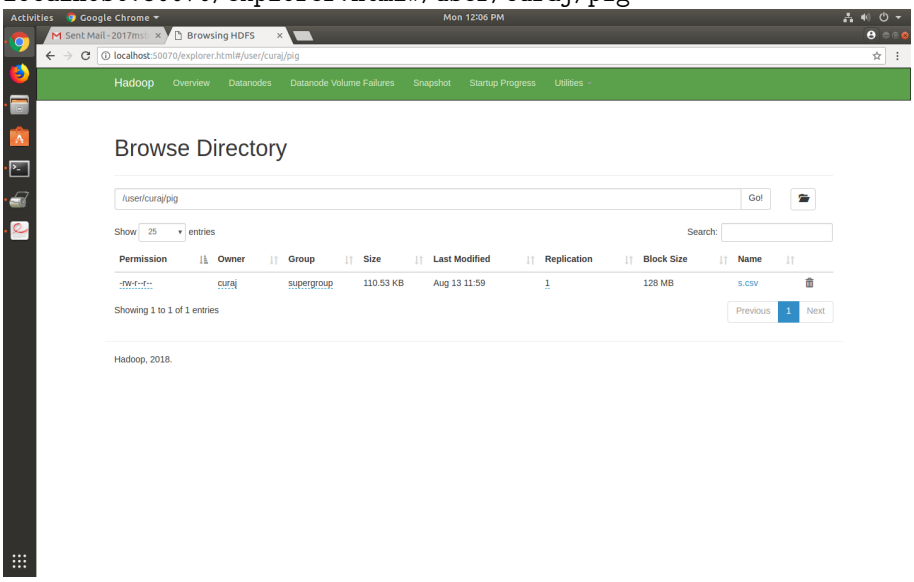
- We will use the following command for the same
  `hadoop fs -mkdir -p /user/curaj/pig`

```
curaj@curaj:~$ hadoop fs -mkdir -p /user/curaj/pig
```

- Now we will put our dataset that is in my home directory in hdfs.
  `hadoop fs -put /home/curaj/s.csv /user/curaj/pig`

```
curaj@curaj:~$ hadoop fs -put /home/curaj/s.csv /user/curaj/pig
```

- Check file in hdfs by using web browser
  `localhost:50070/explorer.html#/user/curaj/pig`



- Start pig in mapreduce
  `pig -x mapreduce`

```
curaj@curaj:~$ pig -x mapreduce
18/08/13 12:00:02 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/08/13 12:00:02 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/08/13 12:00:02 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-08-13 12:00:02,452 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2018-08-13 12:00:02,453 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/curaj/pig_1534141802452.log
2018-08-13 12:00:02,463 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/curaj/.pigbootup not found
2018-08-13 12:00:02,784 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2018-08-13 12:00:02,784 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2018-08-13 12:00:03,068 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-b754baf3-4ed9-46fd-8b50-a7d2f00d958d
```

- Load data using the below command
  `data = LOAD 'pig/s.csv' using PigStorage(',') As (street:chararray,city:chararray,zip:chararray,state:chararray, beds:int,baths:int,sq__ft:int,type:chararray,sale_date:chararray,price:chararray,latitude:chararray,longitude:chararray);`

```
grunt> data = LOAD '/pigdataset/st.csv' using PigStorage(',') As (street:chararr
ay,city:chararray,zip:chararray,state:chararray,beds:chararray,baths:chararray,s
q__ft:chararray,type:chararray,sale_date:chararray,price:chararray,latitude:char
array,longitude:chararray);
```

- Dump data using
  `dump data;`

```
0 EDT 2008,224500,38.69757,-120.995739)
(7540 HICKORY AVE,ORANGEVALE,95662,CA,3,1,1456,Residential,Thu May 15 00:00:00 E
DT 2008,225000,38.703056,-121.235221)
(5024 CHAMBERLIN CIR,ELK GROVE,95757,CA,3,2,1450,Residential,Thu May 15 00:00:00
 EDT 2008,228000,38.389756,-121.446246)
(2400 INVERNESS DR,LINCOLN,95648,CA,3,2,1358,Residential,Thu May 15 00:00:00 EDT
 2008,229027,38.897814,-121.324691)
(5 BISHOPGATE CT,SACRAMENTO,95823,CA,4,2,1329,Residential,Thu May 15 00:00:00 ED
T 2008,229500,38.467936,-121.445477)
(5601 REXLEIGH DR,SACRAMENTO,95823,CA,4,2,1715,Residential,Thu May 15 00:00:00 E
DT 2008,230000,38.445342,-121.441504)
(1909 YARNELL WAY,ELK GROVE,95758,CA,3,2,1262,Residential,Thu May 15 00:00:00 ED
T 2008,230000,38.417382,-121.484325)
(9169 GARLINGTON CT,SACRAMENTO,95829,CA,4,3,2280,Residential,Thu May 15 00:00:00
 EDT 2008,232425,38.457679,-121.35962)
(6932 RUSKUT WAY,SACRAMENTO,95823,CA,3,2,1477,Residential,Thu May 15 00:00:00 ED
T 2008,234000,38.499893,-121.45889)
(7933 DAFFODIL WAY,CITRUS HEIGHTS,95610,CA,3,2,1216,Residential,Thu May 15 00:00
:00 EDT 2008,235000,38.708824,-121.256803)
(8304 RED FOX WAY,ELK GROVE,95758,CA,4,2,1685,Residential,Thu May 15 00:00:00 ED
T 2008,235301,38.417,-121.397424)
(3882 YELLOWSTONE LN,EL DORADO HILLS,95762,CA,3,2,1362,Residential,Thu May 15 00
:00:00 EDT 2008,235738,38.655245,-121.075915)
grunt>
```

- Group by type by using
  `type_group=GROUP datta by type`

```
grunt> type_group=GROUP data by type;
```

- Count groupby by using
  `count=foreach type_group Generate COUNT(data.type)`

```
grunt> count = foreach type_group Generate COUNT(data.type);
```

- To see output we need to use the following command
  `dump count;`

```
curaj@curaj: ~
File  Edit  View  Search  Terminal  Help
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:234)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
2018-08-16 14:07:54,681 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning
aggregation.
2018-08-16 14:07:54,681 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-08-16 14:07:54,684 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2018-08-16 14:07:54,690 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2018-08-16 14:07:54,690 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Residential,917)
(Condo,54)
(Multi-Family,13)
(Unkown,1)
(type,1)
grunt>
```

- For ordering in descending order i used:
  `cd = ORDER count BY cnt DESC;`

```
grunt> cd = ORDER count BY cnt DESC;
```

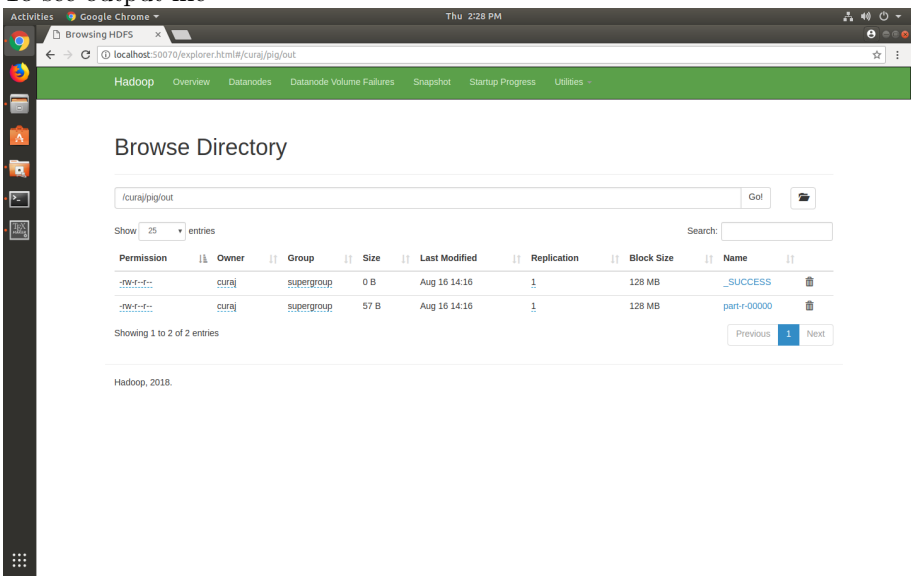
- To see output we need to use the following command
  `dump cd;`

```
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:234)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
2018-08-16 15:45:41,575 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning
aggregation.
2018-08-16 15:45:41,575 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-08-16 15:45:41,578 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2018-08-16 15:45:41,589 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2018-08-16 15:45:41,589 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Residential,917)
(Condo,54)
(Multi-Family,13)
(Unkown,1)
(type,1)
grunt>
```

- To store output file in hdfs i used:
  `store cd into '/curaj/pig/out' using PigStorage(',');`

```
store cd into '/curaj/pig' using PigStorage(',');
```

- To see output file



- Desired output is:

```
Residential,917
Condo,54
Multi-Family,13
Unkown,1
type,1
```

- So, here are the frequencies of each house type in descending order.