

Pig

1. Install Pig

- `$ sudo yum install -y pig`

2. Run grunt shell

- `$ pig
grunt>`

3. Load data from musicbrainz paths

```
grunt> tracks = LOAD 'musicbrainz_tracks' AS (trackid:int, releaseid:int,  
trackname:chararray, position:int);
```

```
grunt> releases = LOAD '/user/cloudera/musicbrainz_releases' AS (id:int,  
album:chararray, artist:chararray);
```

4. Join bags releases and tracks

- `grunt> jnd = JOIN tracks BY releaseid, releases BY id;`

5. Filter by position 1 on list

- `grunt> filtro = FILTER jnd BY position == 1;`

6. Store result *filtro* on HDFS

- `grunt> STORE filtro INTO 'music_position_1';`

```
cloudera@elephant:~  
File Edit View Search Terminal Tabs Help  
cloudera@elephant:~  
2014-10-08 02:05:57,164 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://hose:50036/jobdetails.jsp?jobid=job_201410072245_0007  
2014-10-08 02:06:13,305 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 3% complete  
2014-10-08 02:06:14,318 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 11% complete  
2014-10-08 02:06:15,333 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 19% complete  
2014-10-08 02:06:16,341 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 22% complete  
2014-10-08 02:06:17,356 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 30% complete  
2014-10-08 02:06:18,359 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 30% complete  
2014-10-08 02:06:19,368 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 41% complete  
2014-10-08 02:06:22,388 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 48% complete  
2014-10-08 02:06:23,396 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete  
2014-10-08 02:06:25,438 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 90% complete  
2014-10-08 02:06:42,806 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete  
2014-10-08 02:06:42,808 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:  
  
HadoopVersion PigVersion UserId StartedAt FinishedAt Features  
2.0.0-cdh4.7.0 0.11.0-cdh4.7.0 cloudera 2014-10-08 02:05:50 2014-10-08 02:06:42 HASH_JOIN,FILTER  
  
Success!  
  
Job Stats (time in seconds):  
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs  
job_201410072245_0007 3 1 22 16 18 17 13 13 13 13 filtro,jnd,releases,tracks HASH_JOIN hdfs://elephant:8020/user/cloudera/music_position_1,  
  
Input(s):  
Successfully read 3190809 records from: "hdfs://elephant:8020/user/cloudera/musicbrainz_tracks"  
Successfully read 274267 records from: "/user/cloudera/musicbrainz_releases"  
  
Output(s):  
Successfully stored 266717 records (20304977 bytes) in: "hdfs://elephant:8020/user/cloudera/music_position_1"  
  
Counters:  
Total records written : 266717  
Total bytes written : 20304977  
Spillable Memory Manager spill count : 0  
Total bags proactively spilled: 0  
Total records proactively spilled: 0  
  
Job DAG:  
job_201410072245_0007  
  
2014-10-08 02:06:42,133 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD DISCARDED TYPE CONVERSION FAILED 35 time(s).  
2014-10-08 02:06:42,133 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 40 time(s).  
2014-10-08 02:06:42,133 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
grunt>
```

7. Check results in file

- `grunt> quit`
- `$ hadoop fs -tail music_position_1/part-r-00000`