Hortonworks Tutorial 6

Optimization on Pig

In this tutorial, we'll focus on how to optimize and improve the performance of Pig on **TEZ**.

Task 1. Dataset

We will use the same tables used in Tutorial 5: basketball_master.csv, basketball_players.csv and basketball_teams.csv. Upload these files to HDFS, once checking their existence.

Recall that **TEZ**aims to optimise the performance and speed. By default, Pig uses the MapReduce framework, but we can configure Pig to use the **TEZ**.

Task 2. Speed Improvements in Pig

1. Create a new Pig script named "t5". Then copy the pig script below and paste it in editor:

```
players_raw = LOAD '/tmp/FIT5148/basketball_players.csv'
USING PigStorage(',');

players = FILTER players_raw BY $1 > 0;

player_points = FOREACH players GENERATE $0
AS playerID, $1 AS year, $8 AS points;

grp_data = GROUP player_points BY (year);

max_points = FOREACH grp_data GENERATE group as grp,
MAX(player_points.points) AS max_points;

join_max_points = JOIN max_points
BY ($0, max_points), player_points BY (year, points);

join_data = FOREACH join_max_points GENERATE $0
AS year, $2 AS playerID, $1 AS points;

data_order = ORDER join_data BY year desc;

DUMP data_order;
```

2. Execute the pig script using the MapReduce (the default). It'll take few minutes to finish the job. After the job is completed, examine the "ExecType" in Logs where the default mode is mentioned as "MAPREDUCE":
3. This time, we run the same Pig script using "TEZ". Check the "Execute on Tez" box, and Execute it. Observe the response time (much faster or slower?). Also, check the difference in "Logs" (see Fig. 2)
4. You can always check the executed scripts from the "History" tab and the time taken to execute/complete them.

Task 3. Using the Command Line

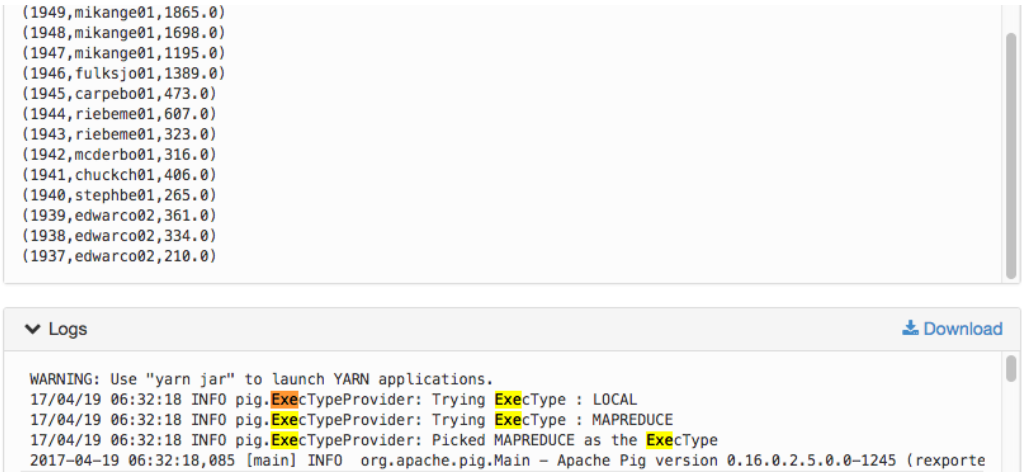

Figure 1: The default ExecType of a Pig query

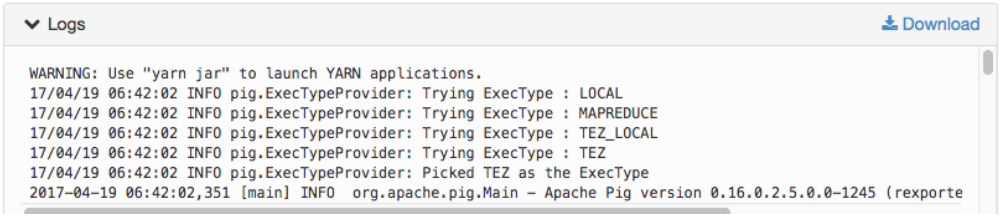

Figure 2: The ExecType: Tez

To use "Tez" with "Pig" we have to use "SSH" via the **VM** instead of **Ambari**. Open your terminal (mac and linux) or puTTY (windows). Type the following command to access the Sandbox through SSH:

```
ssh root@127.0.0.1 -p 2222;
```

If you use the Windows Command, run "putty -ssh root@127.0.0.1 2222" on it. If you access this URL for the first time, there is an prompt box, "PuTTY Security Alert". Read it and click "Yes".

If you login to the sandbox, you will see the following message (see Fig. 3): **"Are you sure you want to continue connecting (yes/no)?"** Enter "yes" to proceed. You will be immediately asked to change the password. Enter current password as "hadoop", and change the password. Make sure you remember the new password because you will be using it the next time you login.

Figure 3: The ExecType: Tez

Let's check current files available on HDFS with the commands:

```
hadoop fs-ls/
hadoop fs-ls/tmp/FIT5148
```

The results will be as Fig. 4

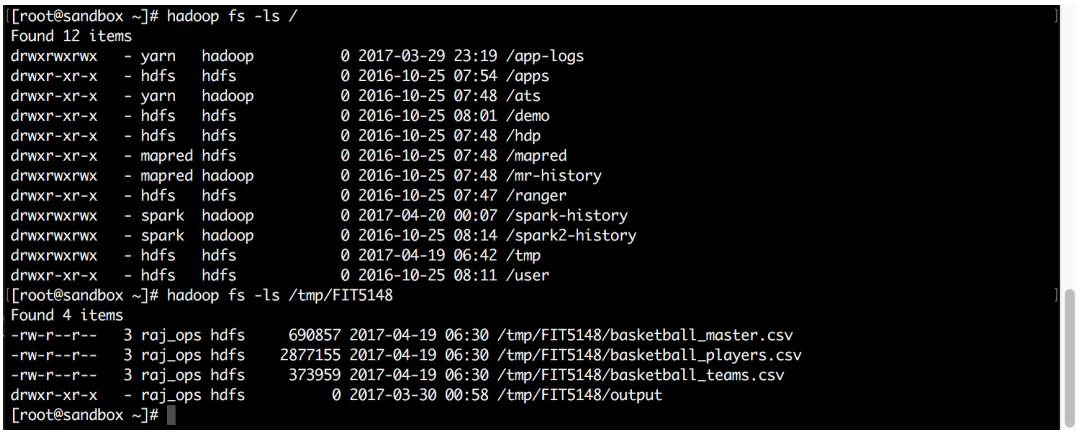

Figure 4: Sandbox data files

First we will run Pig without using Tez. You need to save the pig script below in Word Notepad or Mac TextEdit with the name "script1.pig":

```
players_raw = LOAD '/tmp/FIT5148/basketball_players.csv'
USING PigStorage(',');
players = FILTER players_raw BY $1 > 0;
player_points = FOREACH players
GENERATE $0 AS playerID, $1 AS year, $8 AS points;
grp_data = GROUP player_points BY (year);
max_points = FOREACH grp_data
GENERATE group as grp, MAX(player_points.points) AS max_points;
join_max_points = JOIN max_points
BY ($0, max_points), player_points BY (year, points);
join_data = FOREACH join_max_points
GENERATE $0 AS year, $2 AS playerID, $1 AS points;
data_order = ORDER join_data BY year desc;
DUMP data_order;
```

After creating the "script1.pig" file, copy and paste this file to your Dropbox (or any cloud based storage space that you have such that it can be downloaded later using its URL). If you are using Dropbox, make sure the URL for the file ends with ?dl=1 Then type this command in SSH:

```
wgethttps://www.dropbox.com/YOURFOLDERPATH/script1.pig?dl=1-Oscript1.pig
```

If you put the file in your OneDrive folder, go to the web interface of OneDrive (e.g. `https://onedrive.live.com/`). Right-click on the file and choose "Embed". Then, click the "Generate" button on the right-bottom frame of the page. After that, copy the url tagged by "src" in the "<iframe>". Make changes in the URL by replacing 'embed' with 'download'. After making changes, you URL will be look like:

https://onedrive.live.com/download?cid=2C4B95628EC34191&resid=2C4B95628EC34191%21112&authkey=AGGrdQ8QWjKvzMk.

Note that the "cid", "resid" and "authkey" should be differently shown on your page.

Now, You have done. The above modified link is the direct link of your file.

Then, type the command in SSH:

```
wget fileURL -O script1.pig
```

The "fileURL" must be replaced with the actual data URL that you can download.

Execute the Pig scripts in the file with the MapReduce (the default) in the command line by typing the following command in SSH:

```
pig script1.pig
```

You will see the result of the job with the time taken (see Fig. 5).

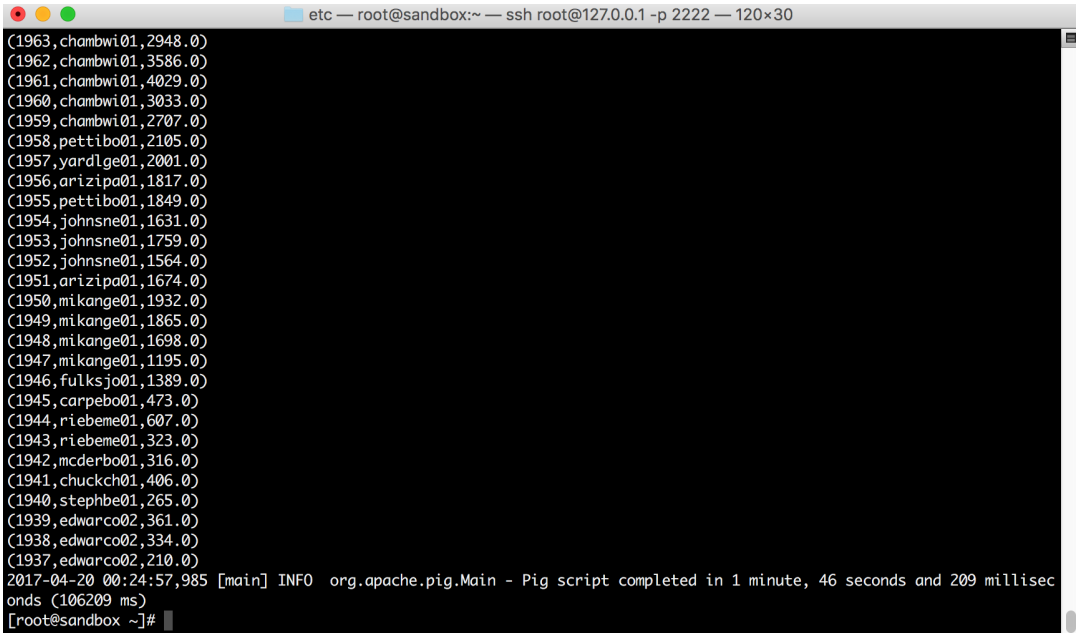

Figure 5: The result of the job, script1.pig

Task 4. Exercise: Optimize a Pig script on Tez

Now, we try to perform optimization and run the same Pig scripts with Tez. Take a look at this URL to know how to optimize a Pig script, and then execute "script1.pig" on Tez: `https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.4.2/bk_dataintegration/content/ch_running-pig-tez.html`

The completion time should be reduced when using Tez compared to MapReduce (Note: your time can be slightly different from ours but running Pig with Tez should show some speed improvement). Find the execution time in the log in SSH and compare it with the time when the script was executed with MapReduce.