Lab 8 - Pig

1. Enter the following URL in Google Chrome:
   https://classroom-vc.gre.ac.uk/vsphere-client/

   Launch the Hadoop VM image, following the instructions from Lab 5.
2. Once the Hadoop desktop has loaded, right click on the desktop and start a terminal
3. Ensure Hadoop is shutdown cleanly from previous sessions by typing:

   stop-yarn.sh

   stop-dfs.sh

4. At the command prompt type:

   start-dfs.sh

   start-yarn.sh

   to restart Hadoop.

5. Change directory into your workspace directory

   cd ../workspace

6. Copy the Lab8 directory from `Datasets` into your workspace directory
7. Change current directory to the **workspace/Lab8** directory
8. Copy the `movies.tab` onto the HDFS

   hdfs dfs –put movies.tab

9. Start the grunt shell, type:

   pig

10. Load the movies data into Pig:

```
grunt> m = LOAD 'movies.tab' using PigStorage('\t');
```

11. Use the **DESCRIBE** command to view the **m** relation:

```
DESCRIBE m;
```

12. Because we have not defined a schema for **m**, you will get a message “Schema for m unknown.”

13. Next define a schema:

```
m = LOAD 'movies.tab' using PigStorage('\t') AS
(title:chararray, year:int, length:int, budget:long,
```

```
rating:float, vote:float, r1:float, r2:float, r3:float,
r4:float, r5:float, r6:float, r7:float, r8:float, r9:float,
r10:float, mpaa:chararray, Action:int, Animation:int,
Comedy:int, Drama:int, Documentary:int, Romance:int,
   Short:int);
```

Then

```
DESCRIBE m;
```

14. Define the following relation **A**, which is a collection of 100 entries (arbitrarily selected) from the **m** relation:

```
A = LIMIT m 100;
```

```
DESCRIBE A;
```

15. Notice A has the same schema as **m**, because **A** is a subset of the **m** relation.

16. To view the data of a relation, use the DUMP command. The DUMP operator is used to run Pig Latin statements and display the results on the screen:

```
Dump A;
```

17. Filter is used to get rows matching the expression criteria. List the movies that have a rating greater than 4

```
mgtf = FILTER m BY rating>4.0;
DUMP mgtf;
```

18. You can use multiple conditions with filters and Boolean operators (AND, OR, NOT): List the movies that were released between 1965 and 1985:

```
mb6585 = FILTER m by year>1965 and year<1985;
 DUMP mb6585;
```

19. To store the results use the STORE command to output a relation into a new file in HDFS. Enter the following command to output the **mb6585** relation to a folder named output/dates:

```
 STORE mb6585 INTO 'output/dates';
```

20. Use the **get** command to get the directory **output** from HDFS, see the first lab.

21. The **ILLUSTRATE** operator generates a concise sample dataset, illustrating how your data has been transformed by your Pig Latin statements:

```
ILLUSTRATE mb6585;
```

22. **FOREACH** gives a simple way to apply transformations based on columns. List the movie names its duration in minutes:

```
md = FOREACH m GENERATE title, length;
```

23. Check the results using the DUMP command.

24. The GROUP keyword is used to group fields in a relation. Group By relations are used to work with the aggregation functions on the grouped data.

List the years and the number of movies released each year:

```
gy = group m by year;
```

Check the results using the DUMP command. Can you make sense of what is happening?

(If you see an error message relating to the Job History Server, exit the grunt shell (quit) and type

```
mr-jobhistory-daemon.sh start historyserver
```

at the command prompt.)

Try the following to see what is happening:

```
p = LIMIT gy 3;
DUMP p;
```

Only one film corresponds to the year 1893. There are six films for 1894 etc.

Try:

```
grd = group m by (rating, length);
```

check the results using the DUMP command.

25. Pig provides a bunch of aggregation functions such as:

1 AVG
2 COUNT
3 COUNT_STAR
4 SUM
5 MAX

6 MIN

7

For example, the following gives the count by years:

```
cy = FOREACH gy GENERATE group, COUNT(m);
```

Check the results by typing :

```
DUMP cy;
```

26. To quit the grunt shell type **quit**
27. Shutdown Hadoop gracefully. At the command prompt type:

stop-yarn.sh

stop-dfs.sh