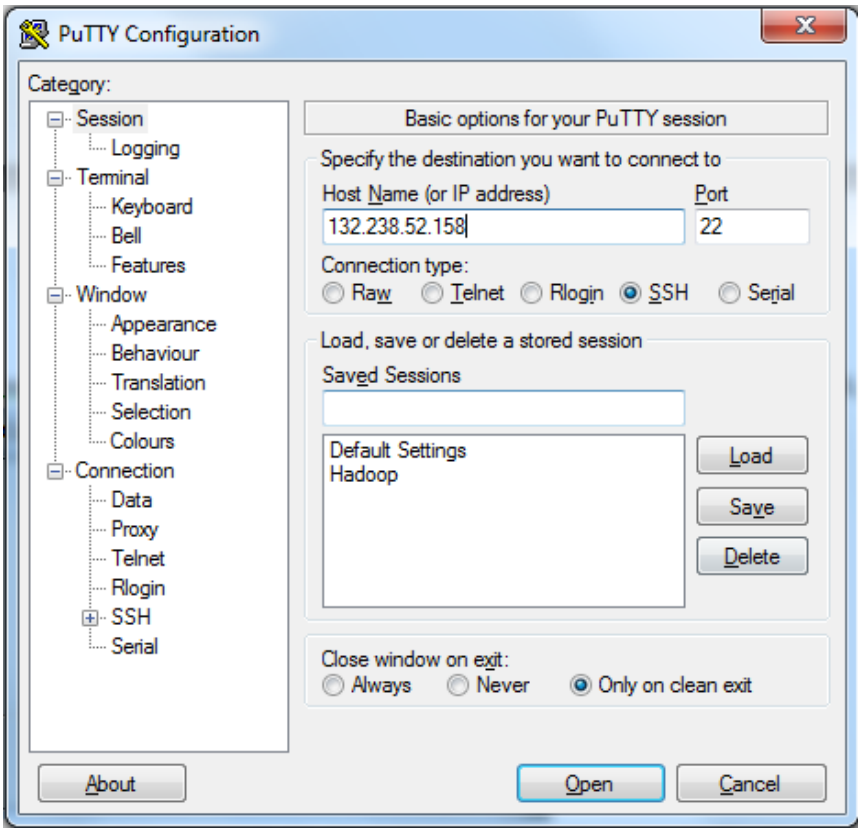
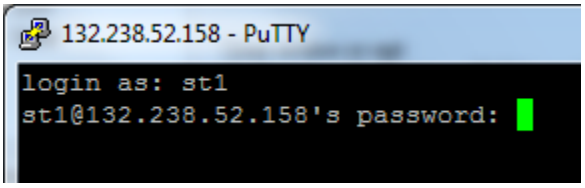
Assignment 4. Pig

1. Start Putty and connect to the name node of the FDU Hadoop cluster. Use IP address = 132.238.52.158 (see the screenshot below). Please double check that SSH is selected as it's shown on the screenshot. Then click the Open button.



2. Enter username and password sent to you by graduate assistant Rajan.



3. After logging in, make a new file movies.txt by typing the following command:

```
cat > movies.txt
```

4. Now, copy and paste the text from below to the Putty shell.

```
1,The Nightmare Before Christmas,1993,3.9,4568
2,The Mummy,1932,3.5,4388
3,Orphans of the Storm,1921,3.2,9062
4,The Object of Beauty,1991,2.8,6150
5,Night Tide,1963,2.8,5126
6,One Magic Christmas,1985,3.8,5333
```

```
7,Muriel's Wedding,1994,3.5,6323
8,Mother's Boys,1994,3.4,5733
9,Nosferatu: Original Version,1929,3.5,5651
10,Nick of Time,1995,3.4,5333
```

5. To exit the text file saving changes, press [Ctrl] and [D] button combination on your keyboard (you need to press and hold the [Ctrl] button and then press [D] button).
6. Now, just to make sure that movies.txt was saved successfully, execute the following command, which will display the contents of the file on the screen:

```
cat movies.txt
```

7. Load movies.txt to HDFS into your directory, which is the same as your username. In the code snippet below,*st**X*** should be substituted with your username (every student received a username in the range from st1 to st70).

```
hadoop fs -put movies.txt /user/stX
```

8. Start Pig by typing the following command:

```
pig
```

9. Load the content of "movies.txt" into a variable named "Movies". Note that st*X* in the code below must be substituted with your username again.

```
Movies = LOAD '/user/stX/movies.txt' USING PigStorage(',') as
(id,name,year,rating,duration);
```

10. To see the content of the variable "Movies", use the following command:

```
DUMP Movies;
```

11. To check the format of the variable "Movies", use the following command:

```
Describe Movies;
```

12. Now that the data is loaded into the variable "Movies", let's filter the data for movies with a rating of greater than 3.5 using the following command:

```
movies_greater_than_three_point_five = FILTER Movies BY rating>3.5;
```

13. From the variable 'movies_greater_than_three_point_five', let's' extract the values for 'year','rating', and 'moviename' and save them in another variable named 'foreachexample'.

```
foreachexample= foreach movies_greater_than_three_point_five generate
year,rating,name;
```

14. Display the contents of the variable foreachexample on the screen:

```
dump foreachexample;
```

15. Let's store the values of variable 'movies_greater_than_three_point_five' into HDFS. Note that st*X* in the code below must be substituted with your username again.

```
STORE movies_greater_than_three_point_five INTO
'/user/stX/movies_greater_than_three_point_five' USING PigStorage (',');
```

16. The command above would store the result in the file called **part-m-00000** placed in the subfolder called **movies_greater_than_three_point_five**. Now that we have the data in HDFS, use the 'cat' command to open the processed file. Note that st*X* in the code below must be substituted with your username again.

```
cat /user/stX/movies_greater_than_three_point_five/part-m-00000;
```

17. Let's copy the file with results from HDFS to the local filesystem by executing the following command. Note that st*X* in the code below must be substituted with your username again.

```
copyToLocal /user/stX/movies_greater_than_three_point_five/part-m-00000
/home/stX/;
```

18. Exit Pig by typing:

```
quit
```

19. Now we are back to Linux shell. List the files in your home directory by executing this command:

```
ls
```

20. Now you can display the contents of the file produced in Pig in the Linux shell:

```
cat part-m-00000
```

Assignment for extra credit. Export the results from HDFS to local file system of a linux machine you are connecting to using Putty. Then retrieve the results from a Linux machine to your personal computer using an FTP client (for example, you can download a program WinSCP, a free secure FTP client).