

Chapter 2 – The Grunt shell

Exercise

1. This exercise uses the MovieLens data set. This is a free data set provided by the University of Minnesota's GroupLens Research Group, and contains over one million movie ratings. For more information, please see the license information at the end of this exercise.

If you haven't already done so, extract the data from the MovieLens data set:

```
$ tar xvfz ml-data.tar.gz
```

Look at the README file in the ml-data directory.

2. In a terminal, invoke the Grunt shell:

```
$ pig
```

3. In the Grunt shell, ask for help:

```
grunt> help
```

4. Use `copyFromLocal` to copy these files into Hadoop's Distributed File System (put them in `/user/training` with their original file names):

```
/home/training/ml-data/u.data
```

```
/home/training/ml-data/u.item
```

```
/home/training/ml-data/u.user
```

```
/home/training/ml-data/u.info
```

```
grunt> copyFromLocal /home/training/ml-data/u.data u.data
```

```
grunt> copyFromLocal /home/training/ml-data/u.item u.item
```

```
grunt> copyFromLocal /home/training/ml-data/u.user u.user
```

```
grunt> copyFromLocal /home/training/ml-data/u.info u.info
```

5. Use `ls` to verify the file is in HDFS.

```
grunt> ls
```

6. Use `cat` to look at `u.info`

```
grunt> cat u.info
```

7. Use `quit` to exit the Grunt shell.

```
grunt> quit
```

8. In the terminal, use `pig -e` to list the files in HDFS.

```
$ pig -e "ls"
```

This exercise uses the MovieLens data set, or subsets thereof. This data is freely available for academic purposes, and used and distributed by Cloudera with the express permission of the UMN GroupLens Research Group. If you would like to use this data for your own research purposes, you are free to do so, so long as you cite the GroupLens Research Group in any resulting publications. If you would like to use this data for commercial purposes, you must obtain explicit permission. You may find the full dataset, as well as detailed license terms at <http://www.grouplens.org/node/73>