# Implementing a Recommender System

Zach Adair

MSDS 600

12/12/2018

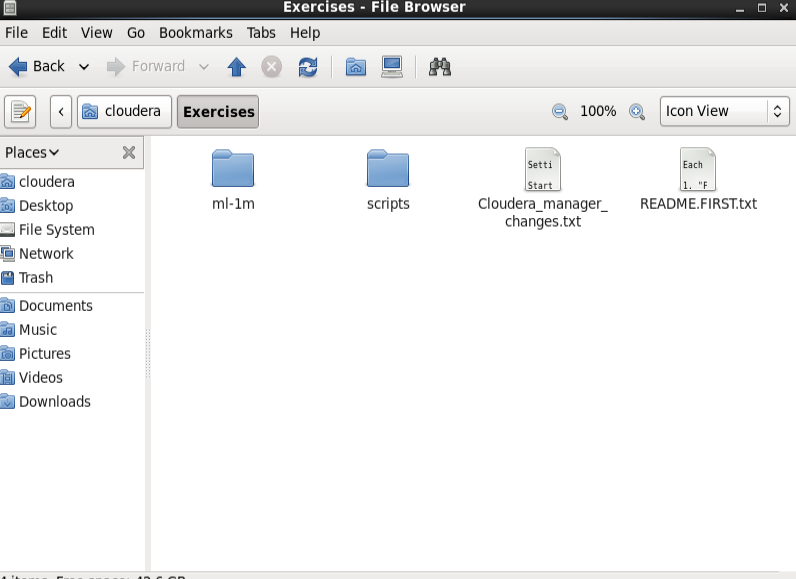
## Objective

Follow the steps laid out to create a Recommender system in Cloudera from the downloaded data.

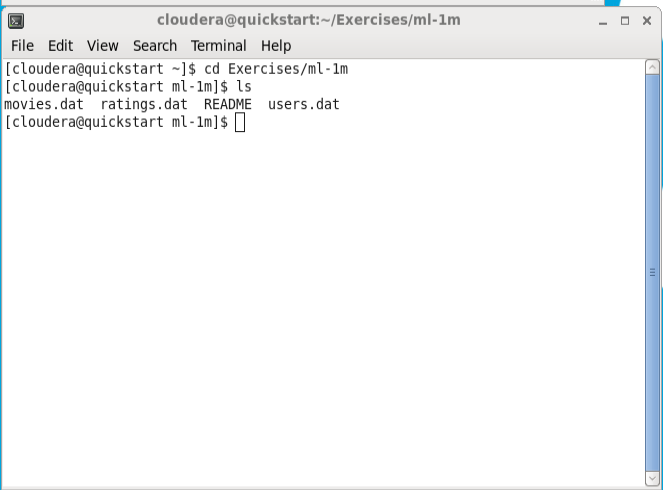
## Process

### **Part 1**

1. Upload the files into Cloudera and create the path way in the terminal. I moved both the ml-1m and scripts folder under cloudera/Exercises.

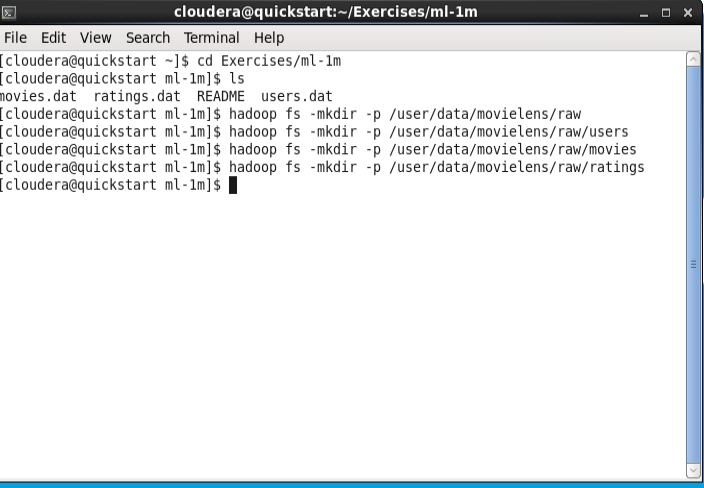


Then also created a file path, below is the file path created in the terminal for ml-1m.



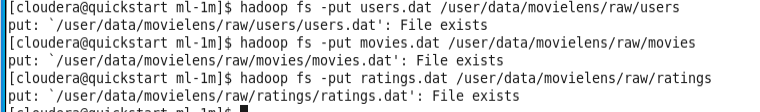
1. Prepare to store data into the HDFS file system by creating directories to it

Using the code provided, created the Hadoop file locations for the three datasets.

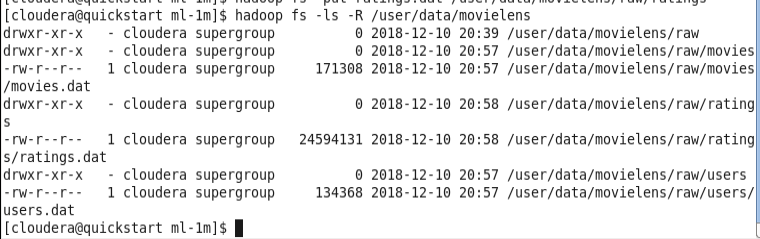


1. Create the HDFS directory structure and place the raw data files there

Using the 3 data files from the ml-1m folder.



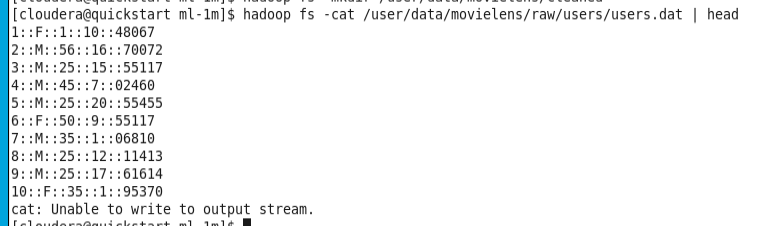
1. Confirm the directors in HDSF



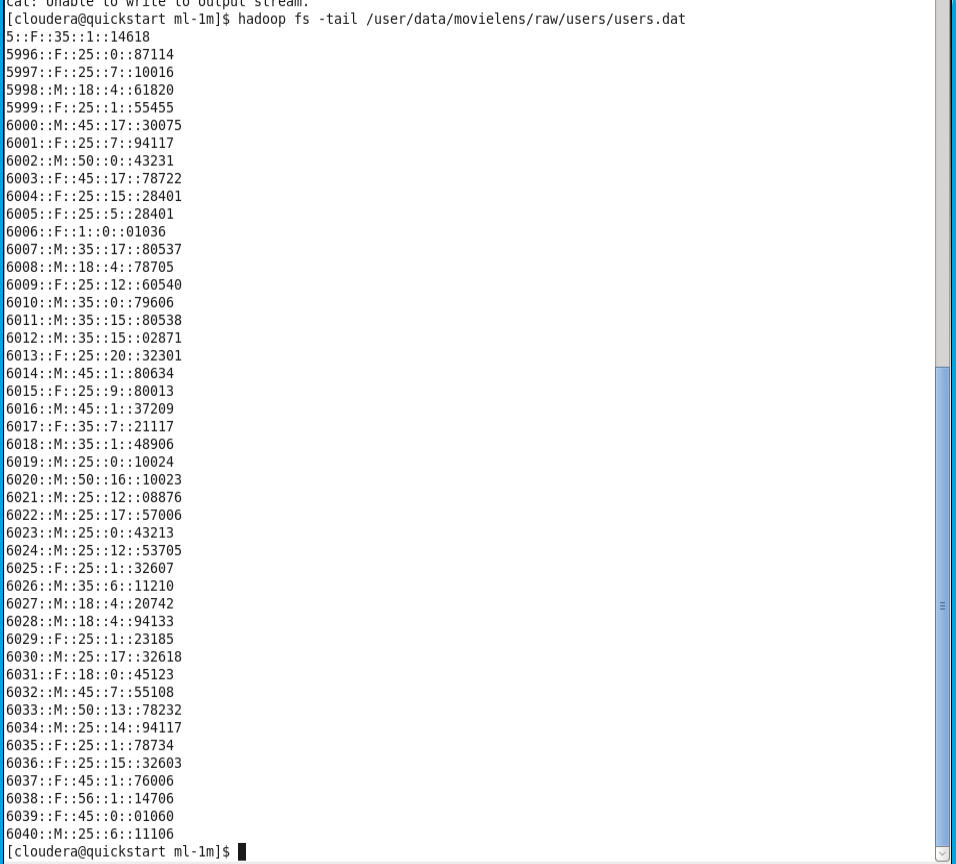
1. Make a subdirectory to contain the “cleaned” data that will be created in Step 2.



1. Using the other commands that is in *From the Expert, from this week’s content section.*



Above is the head function

lsls

Above is the tail functioncd

### **Part 2 – Preprocessing data using Python – Run the data file cleaning scripts:**

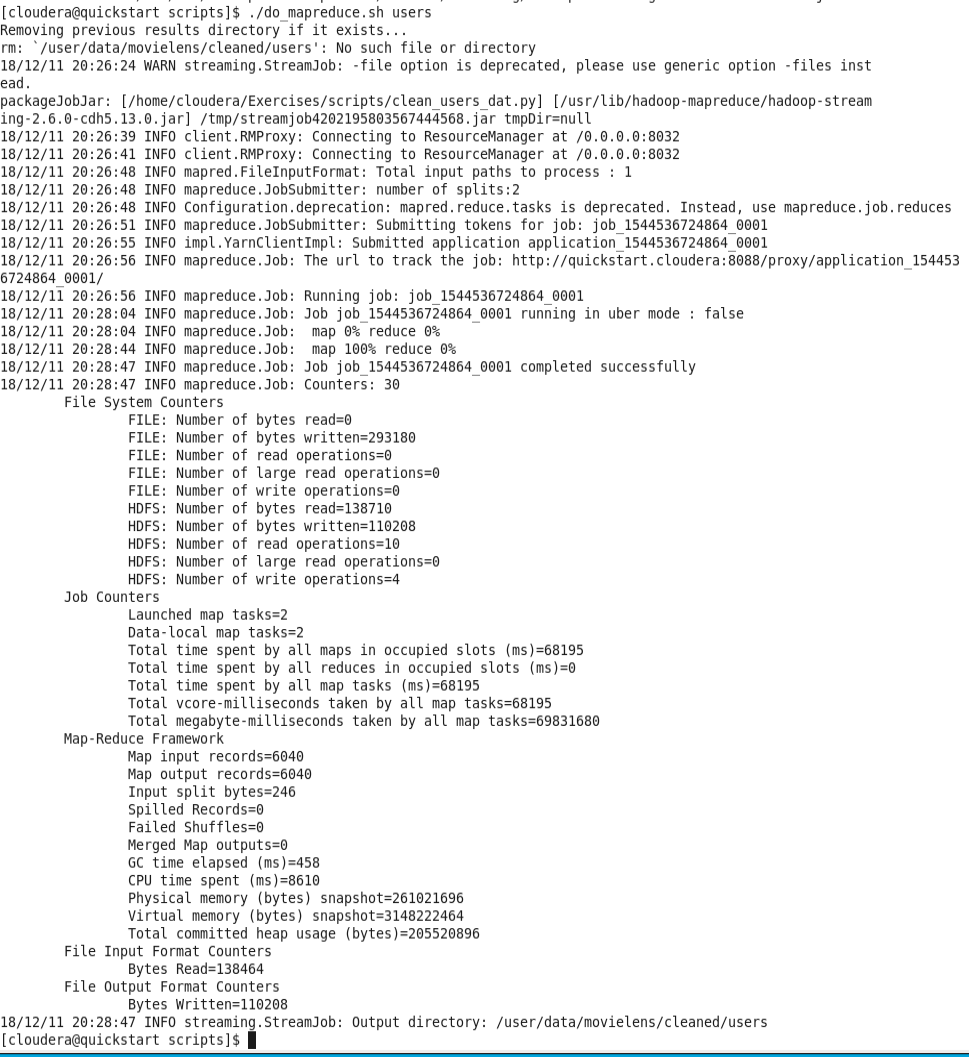
1. Preprocessing data using Python – Run the data file cleaning scripts

Start by changing the directory to the scripts folder

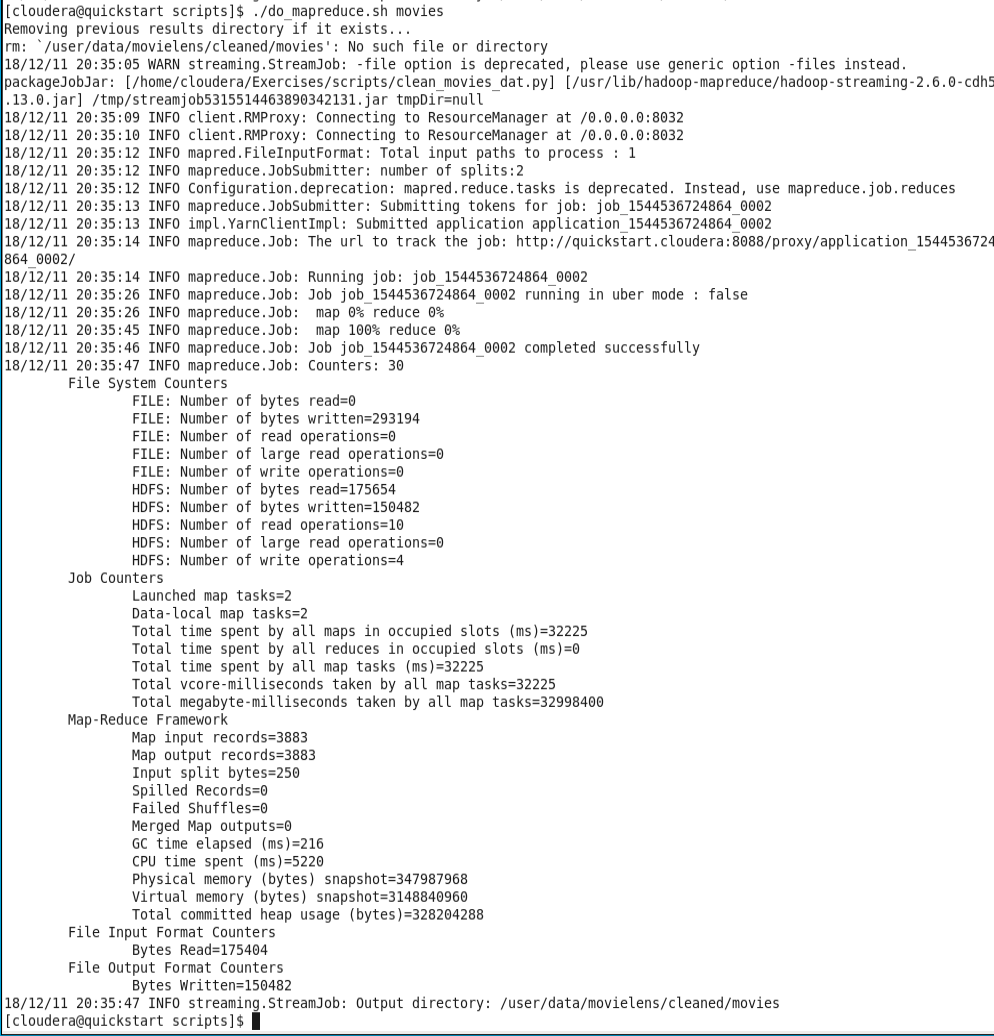


Also ran the ls function to make sure I had the necessary files within the scripts folder.

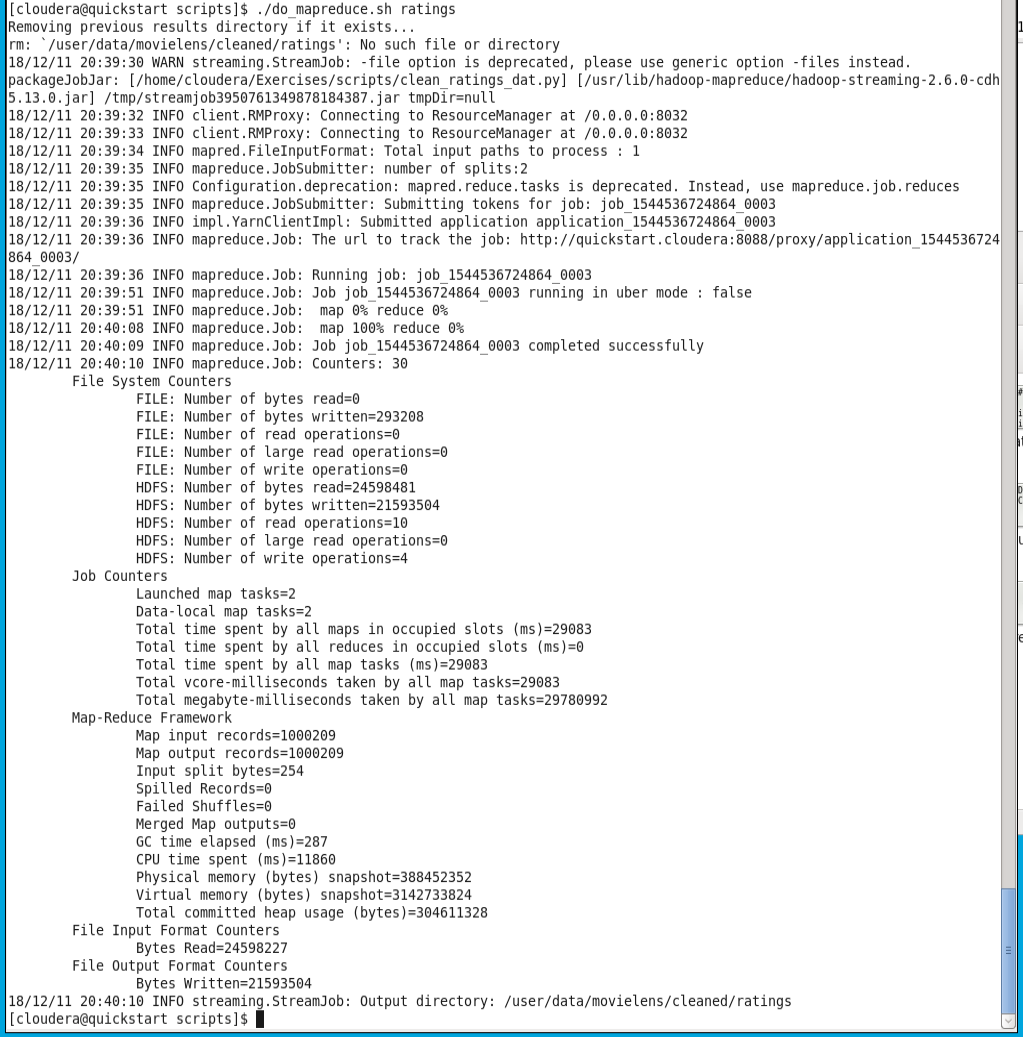
Next run the necessary MapReduce scripts for users, movies and ratings.



Above is the MapReduce for the Users file



Above is from the Movies file

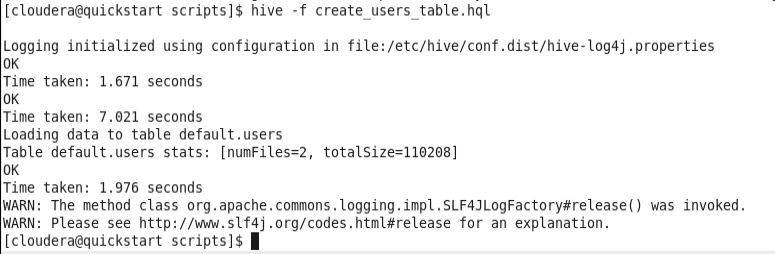


Above is from the ratings file

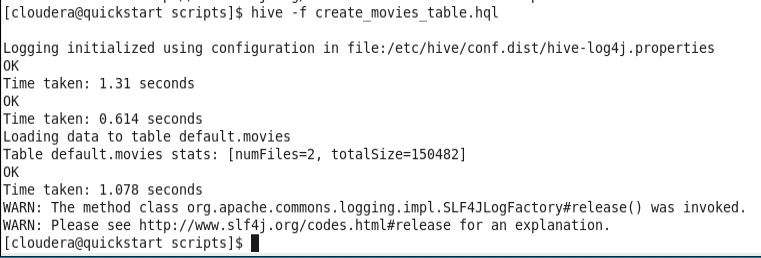
### **Part 3 – Create Hive tables creating Hive tables makes it easy to run queries for filtering**

1. Create Hive tables using -f tells Hive to execute HiveQL code in referred file

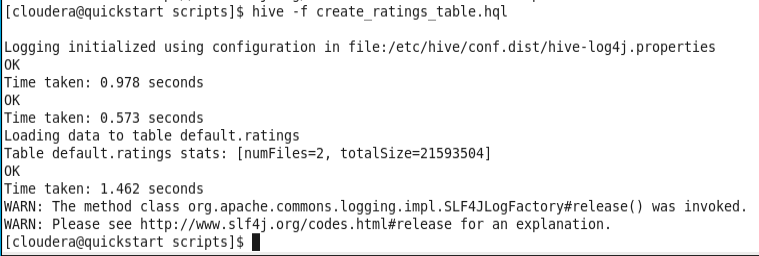
This round of code will create the hive tables for users, movies and ratings.



Users



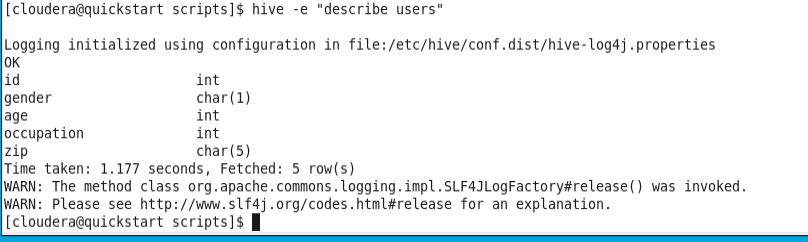
Movies



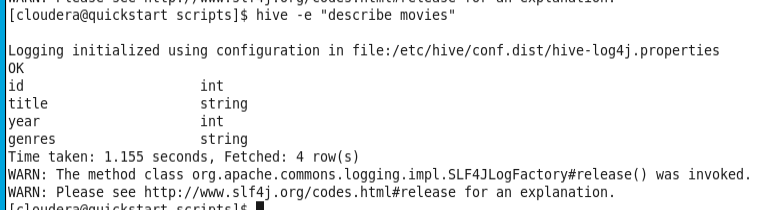
Ratings

1. Examine the data fields

This code can describe the data fields that coincide with the Hive tables that were just created, I’m going to do that for the three tables just made.



Users Description



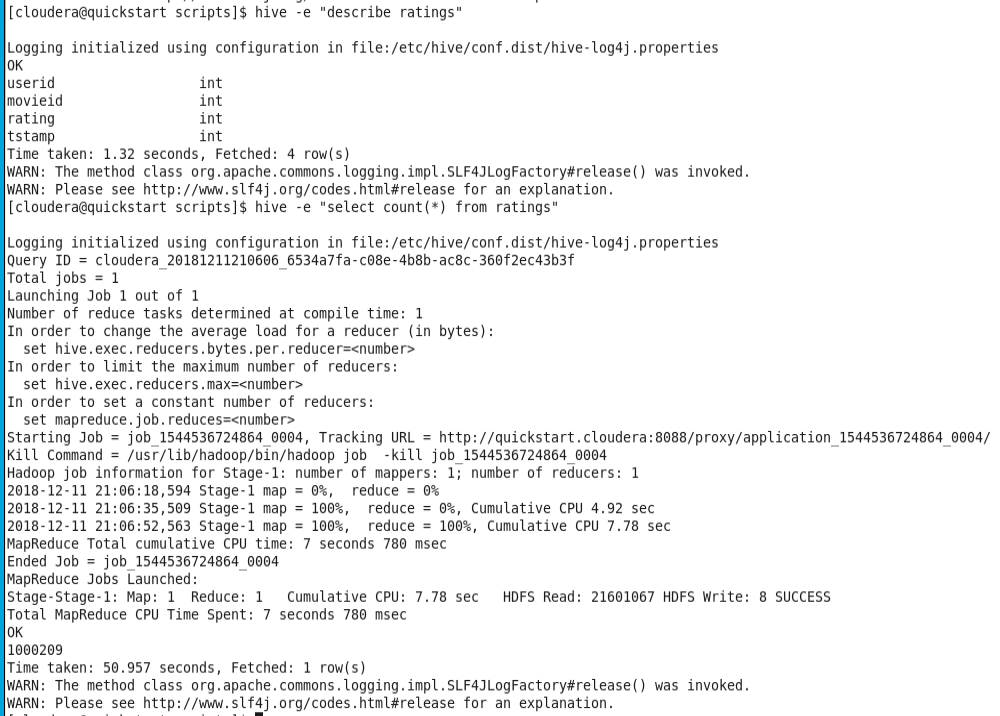
Movies Description



Ratings Description

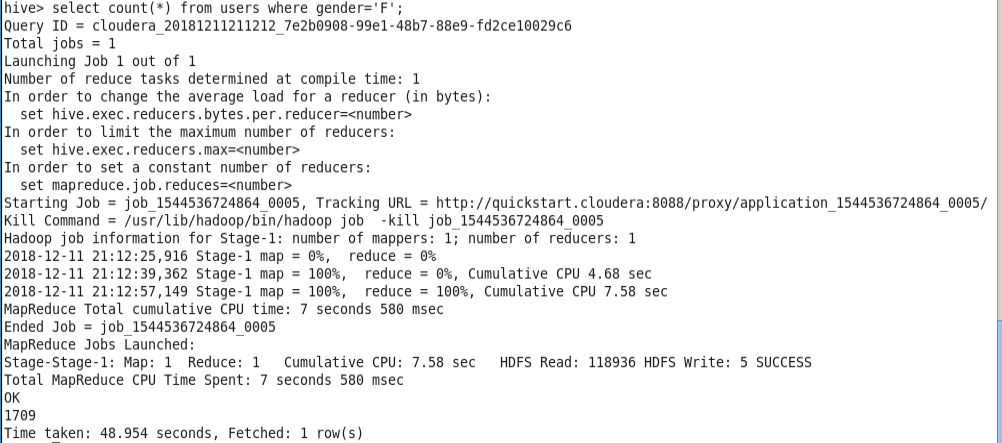
1. Determine # of ratings

Use the code to determine the count on specific ratings



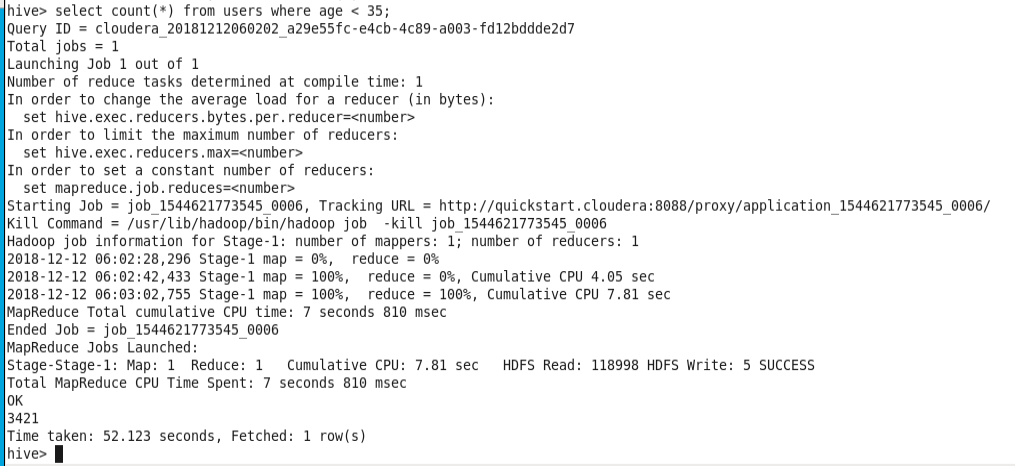
The count is totaled at 1,000,209 ratings.

1. Prompt Hive and run a few queries such as:
   1. How many users are female?



It appears there are a total of 1709 female users.

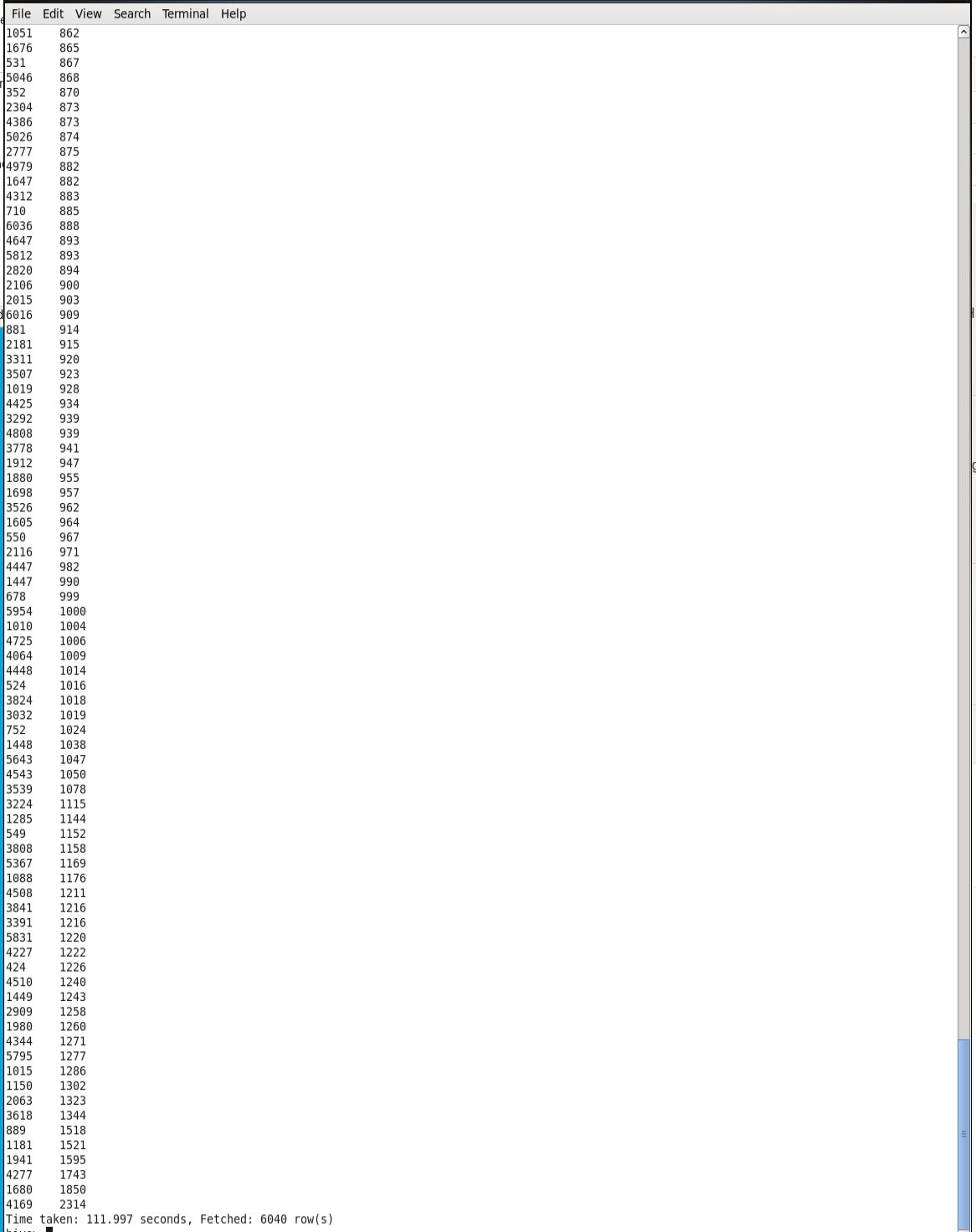
* 1. How many users are under 35 years old?



The total number of users under the age of 35 is 3421.

* 1. Determine the top 3 users with the most ratings?



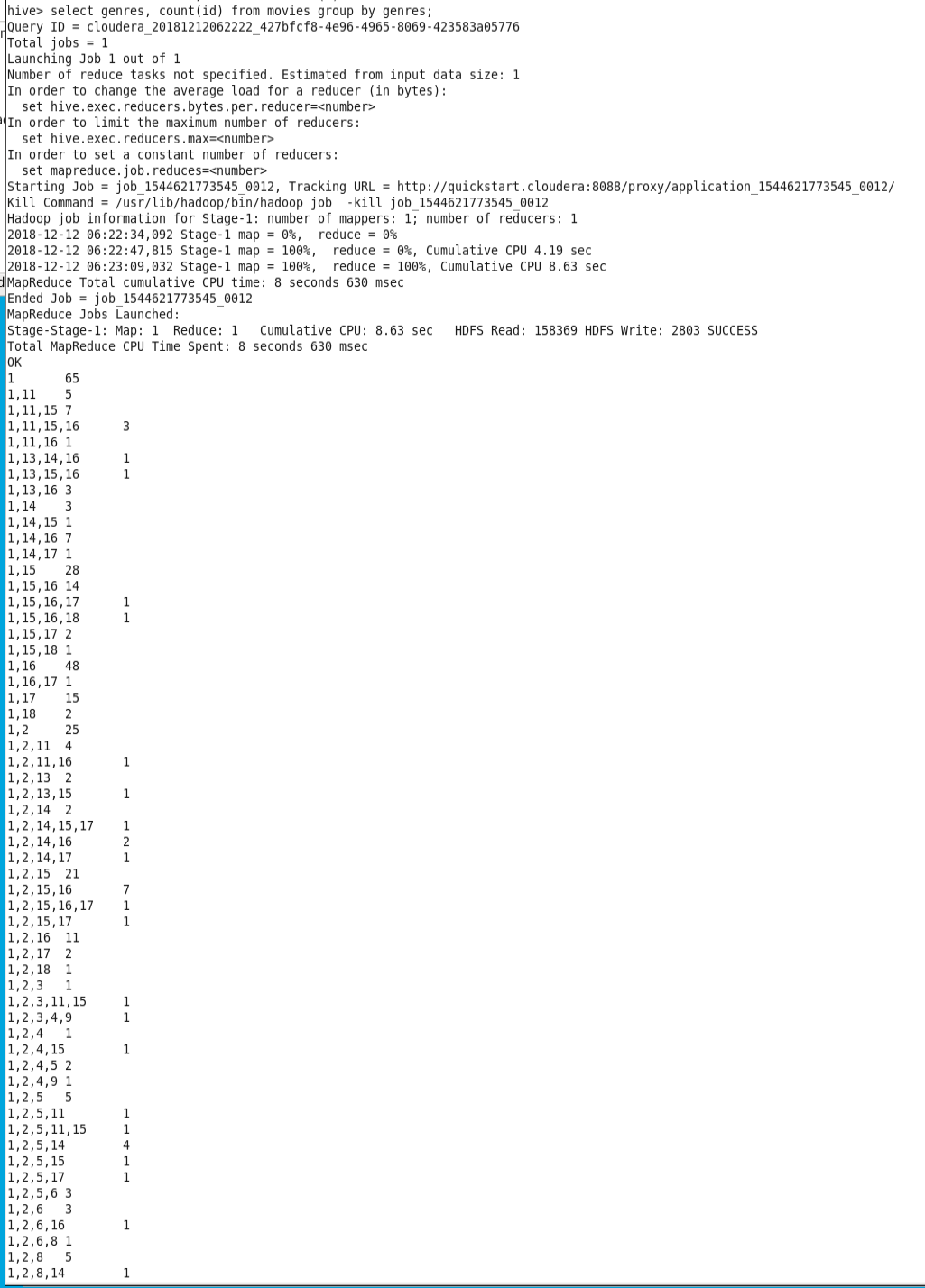


* 1. How many users are in each occupation?



The occupation with the most users is occupation 0. The occupation with the least user is occupation 8.

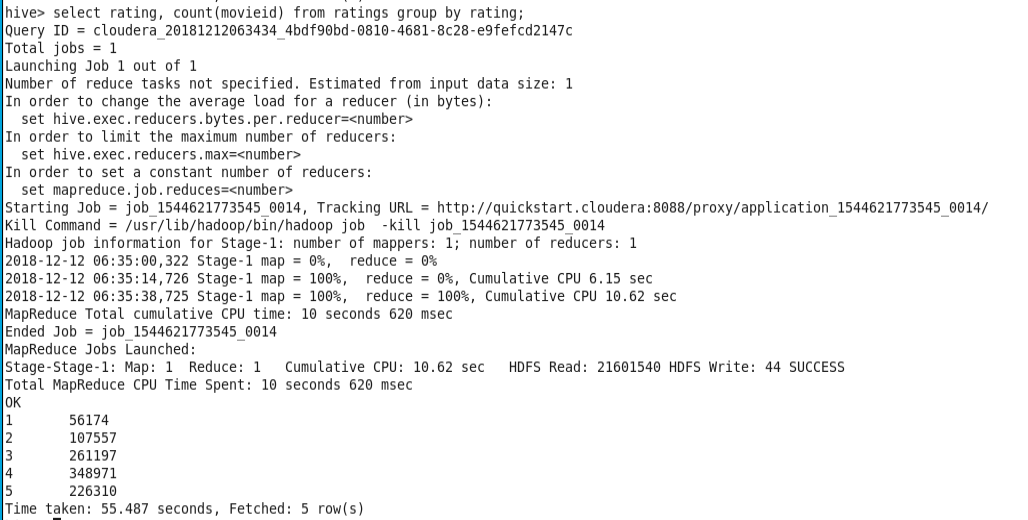
* 1. Create a query on my own – I am creating a query to select the number of movies by movie genre. Since some movies have multiple genres, the data comes back a little sloppy.





Each noted mix of genres is displayed along with the number of movies with that mix of genres.

* 1. I am creating another query, this time I created a query to count the number of movies based on a specific rating.

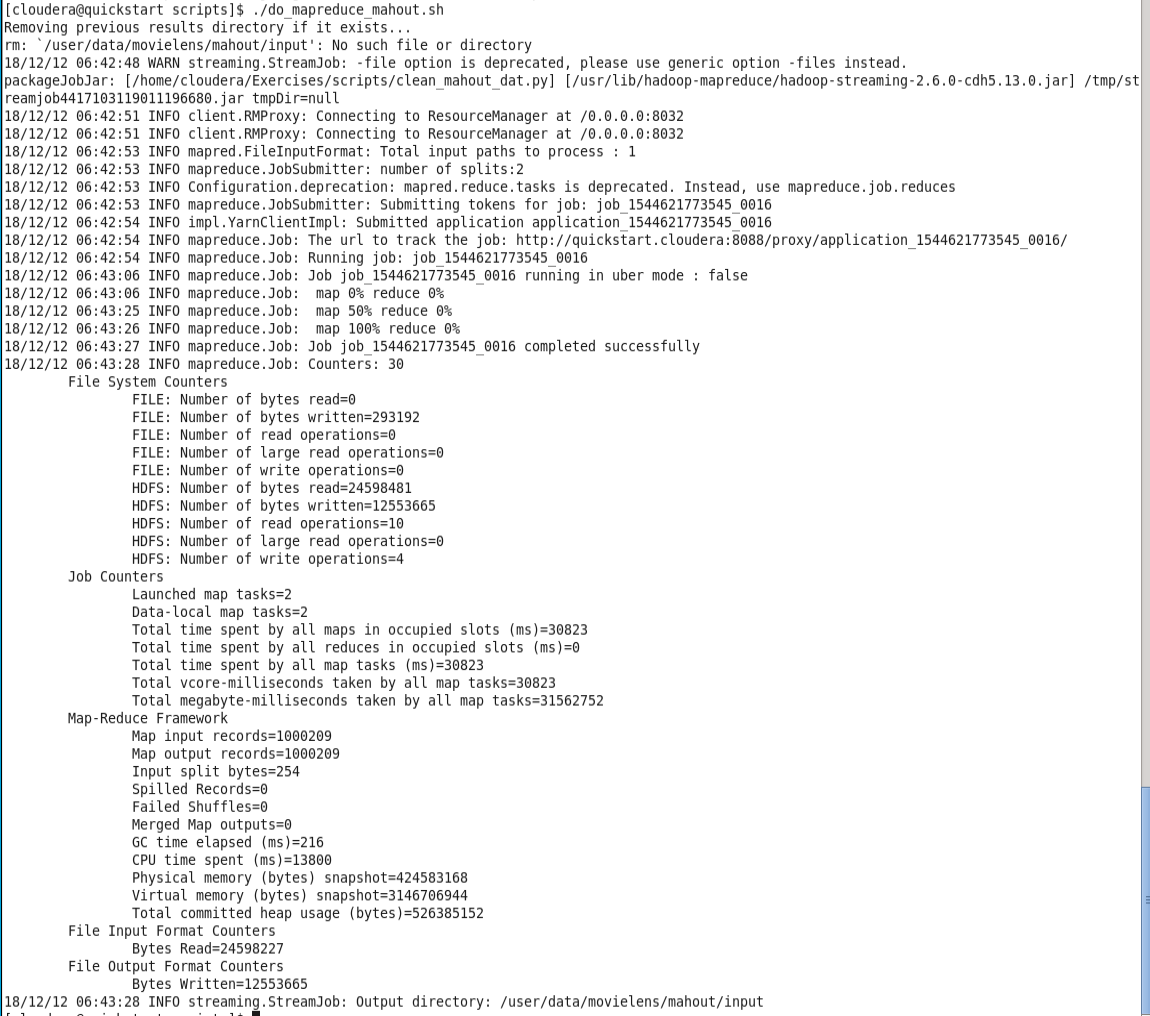


Based on my counts, the majority of movies are rated 4 and 3, followed by 5, 2 and a 1 rating has been given the fewest times.

### **Part 4 – Build the basic recommender system using Mahout**

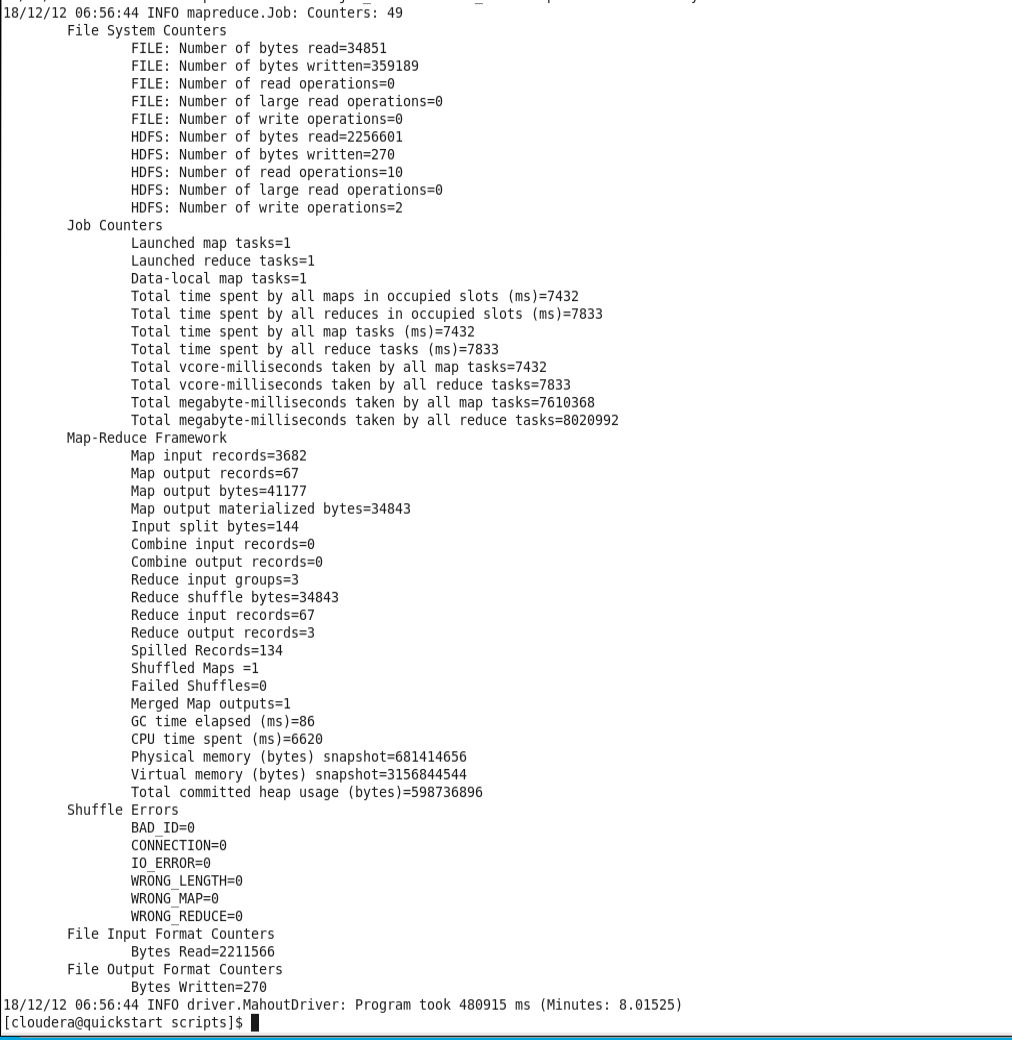
1. This script specifies to Mahout that users 1,2 and 3 will receive recommendations based on their respective movie rating history.

Start by running the mahout.sh code

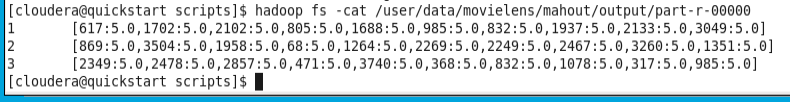




Next run the ./ output for the mahout.sh



Then finally run the last line of code.



The last line of codes output is the recommendations set up from our new recommender system for users 1, 2 and 3.

Reflection  
From this project, I’ve learned more about Cloudera, Hadoop and Hive and how those work through any issues I need to edit my data files for use within the system. Being able to set up recommender systems using Hadoop and being able to query data with Hive are important techniques to learn as a Data Scientist and understanding how to use Cloudera to accomplish these tasks makes me confident of being able to use this tool to analyze any data set I will have put in front of me in the future.