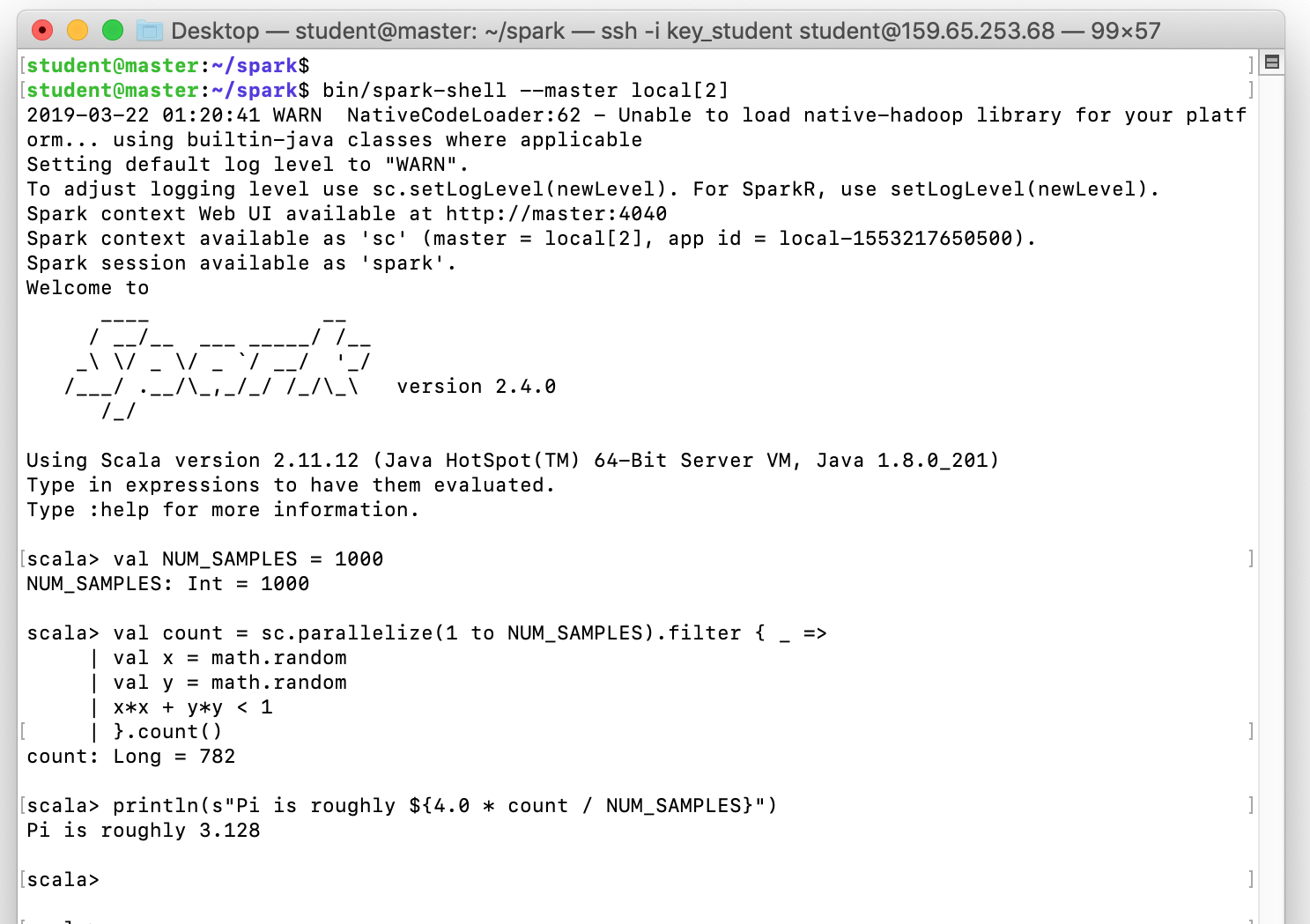
**INFSCI 2750 Cloud Computing**

**Mini Project 2**

Jing Pang [jip45@pitt.edu](mailto:jip45@pitt.edu)   
Haoyang Qian [haq13@pitt.edu](mailto:haq13@pitt.edu)   
Tian Xue [tix20@pitt.edu](mailto:tix20@pitt.edu)

## **Part 1: Spark Setup**

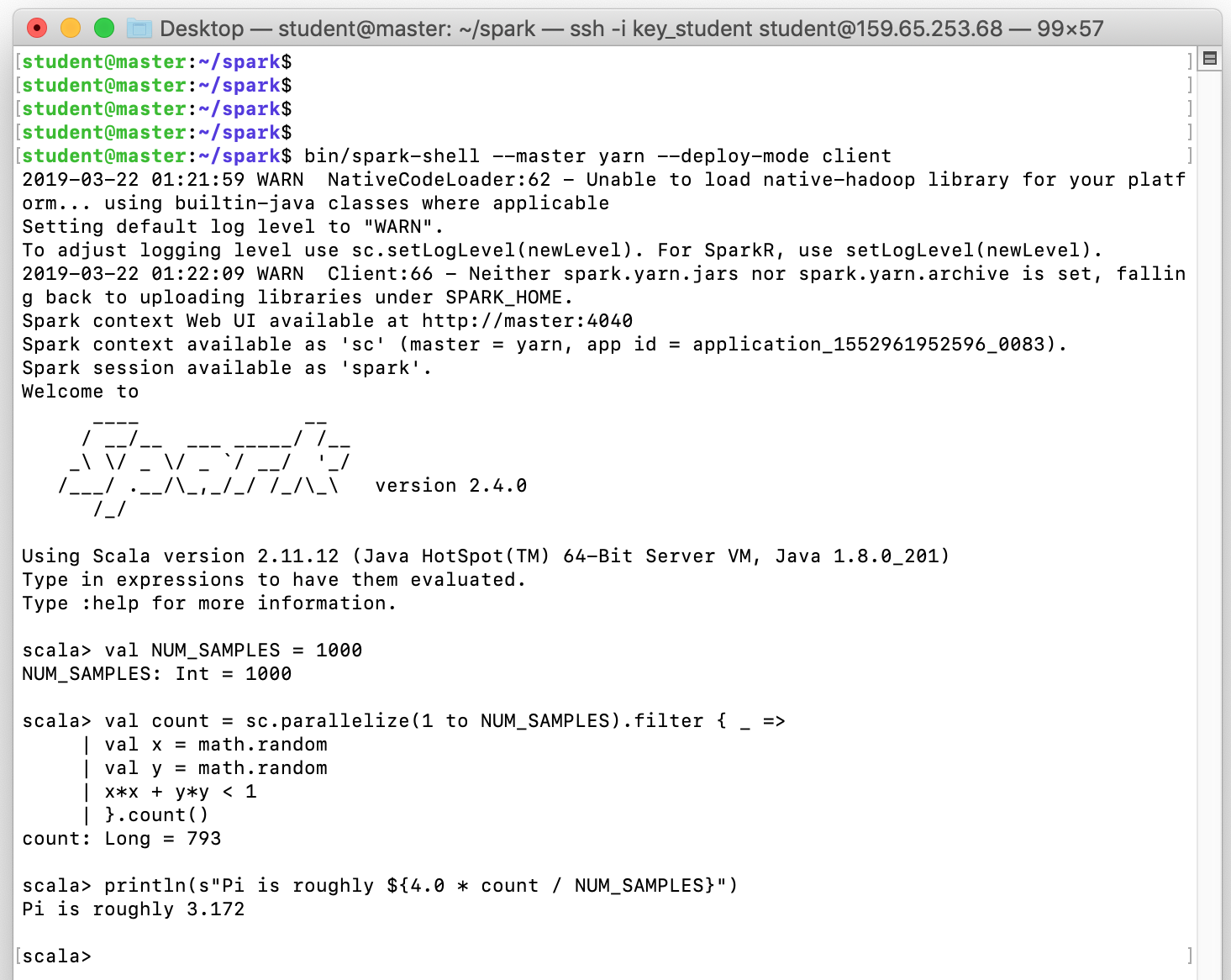
In the beginning, we will setup Spark in our master node.



Then, to fulfill the goal of running Spark on YARN. We will setup HDFS and YARN services and environment configuration. The following picture shows that the configuration setup properly.



So we can test previous example with this environment.



## **Part 2: Developing Spark program 1**

In this case, we will implement with the dataset ‘hetrec2011-lastfm-2k’. In particular, we will be working on the ‘user\_artists.dat’ data file.

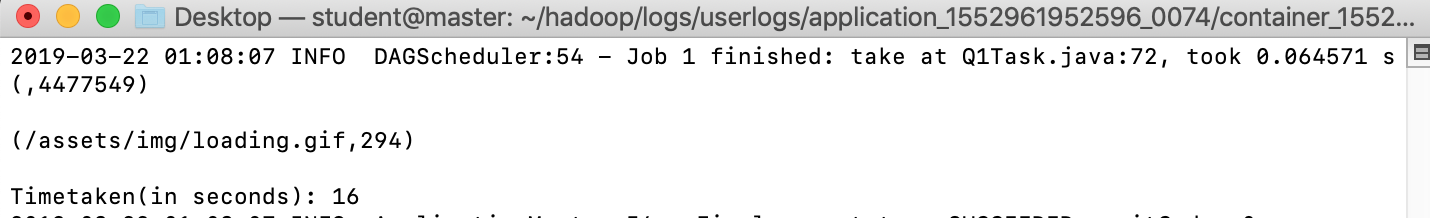
The task is to printout the total listening counts of each artist. The following picture only shows a part of result in the whole file.



## **Part 3: Developing Spark program 2**

1. How many hits were made to the website item “/assets/img/loading.gif”?

Hits: 294

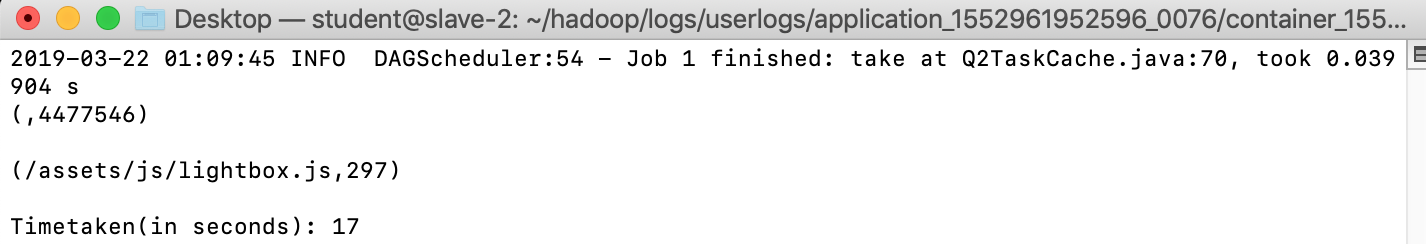
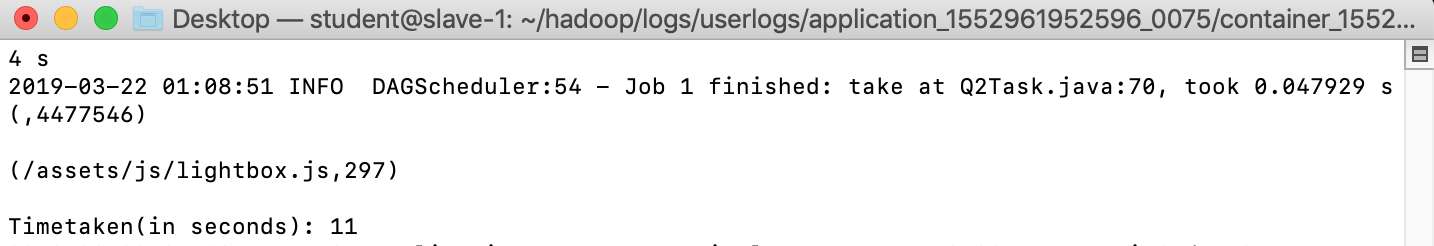


1. How many hits were made to the website item “/assets/js/lightbox.js”?

Hits: 297

Time taken without cache: 11

Time taken with cache: 17



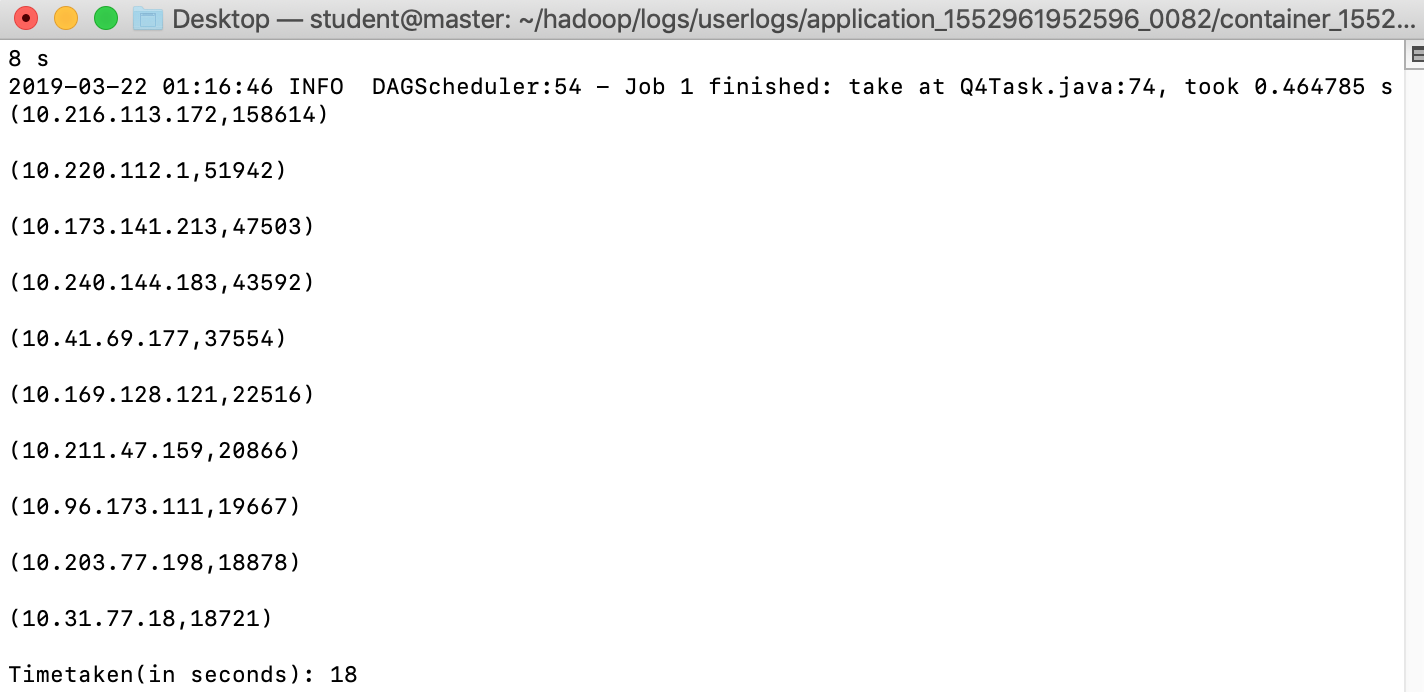
1. Which path in the website has been hit most? How many hits were made to the path?

Most hits of the website: /assets/css/combined.css; hits are 117348.



1. Which IP accesses the website most? How many accesses were made by it?

Most hits of the IP access is 10.216.113.172; hits are 158614.



**RDD Cache Efficiency Analysis:**

Our total running time with cache is longer than that without cache, which is an exception. We speculated that this was because the size of dataset is not big enough. Furthermore, different nodes running the code may be allocated different resources, which might lead to different results of running time. We will continuously look into it to dig out the truth behind it.