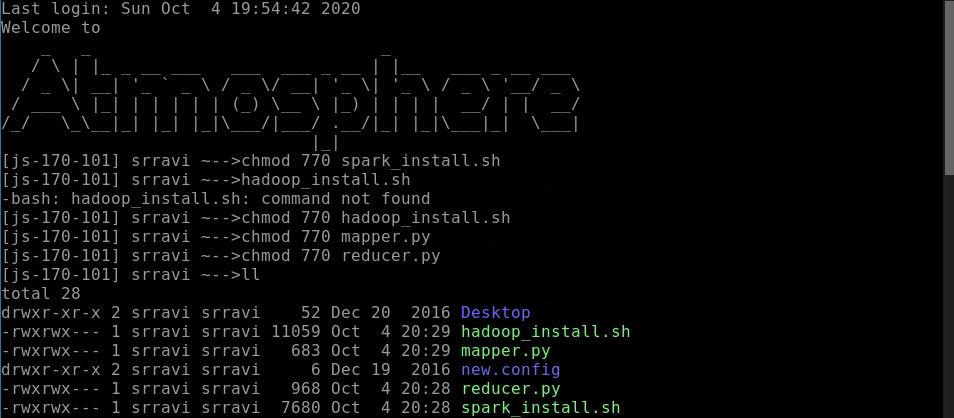
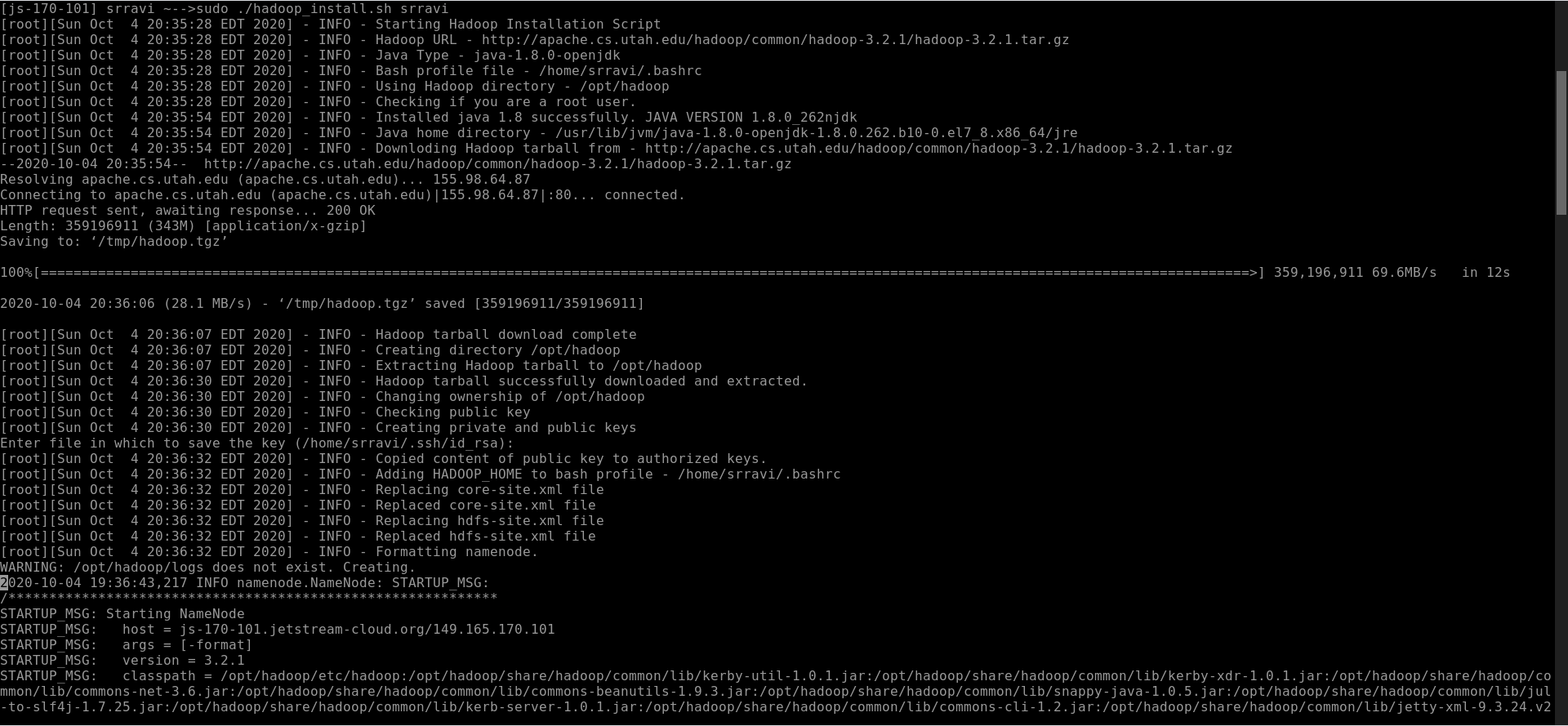
Sreeti Ravi

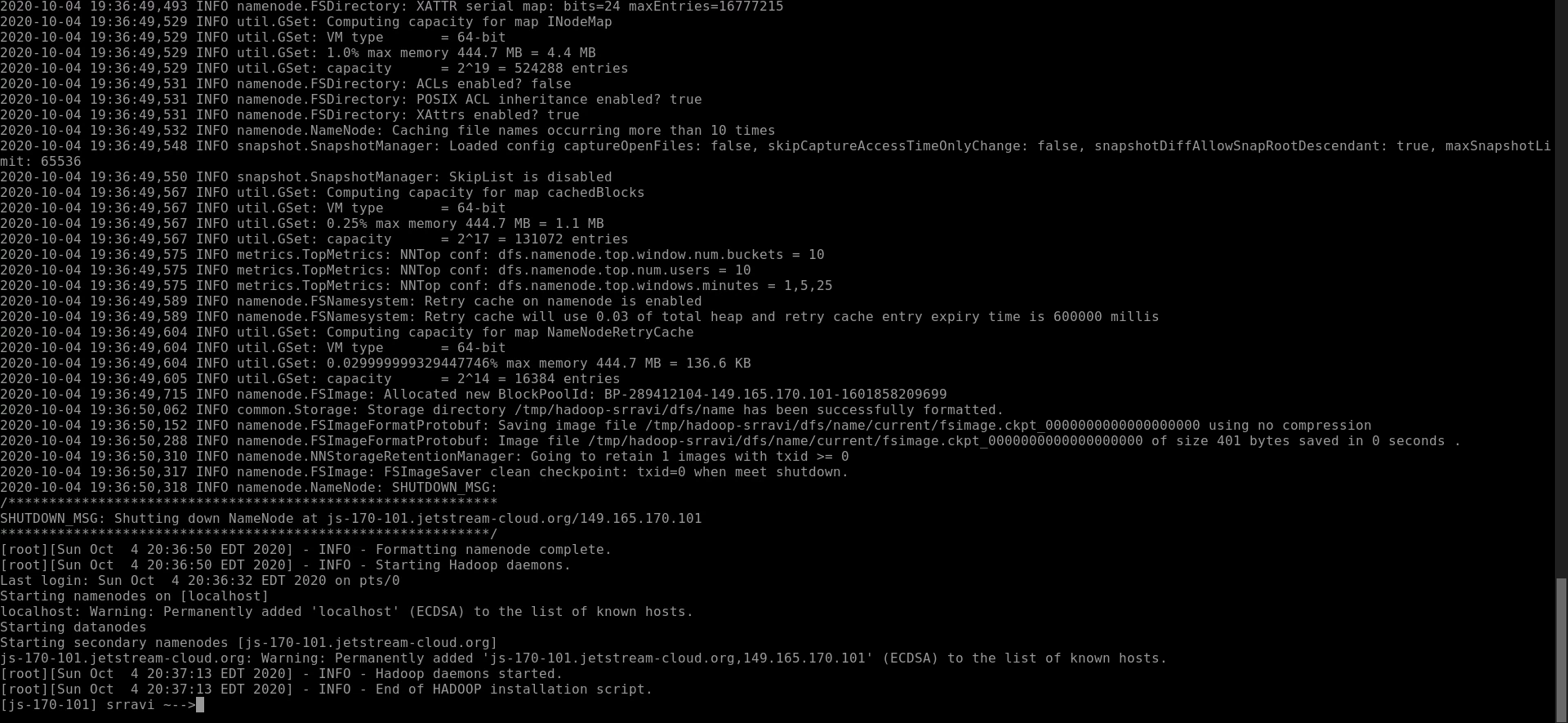
Distributed Computing

1. Permissions

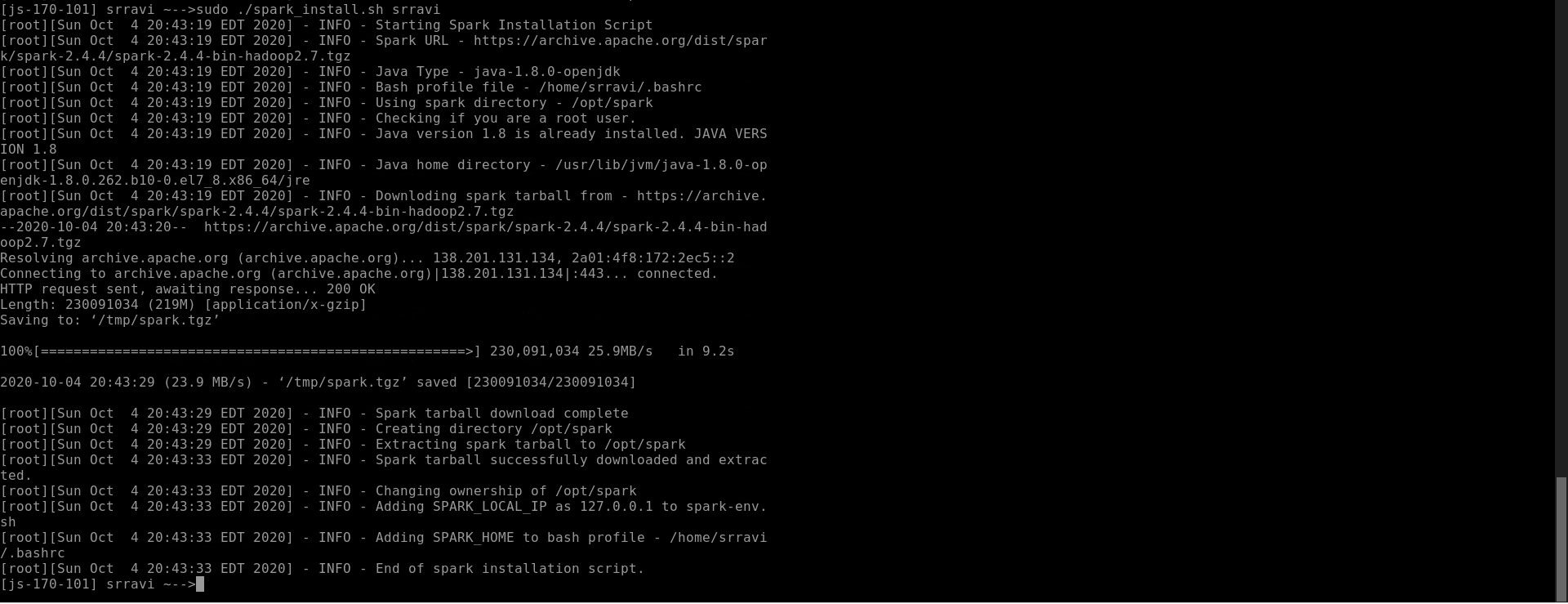


1. Hadoop Install

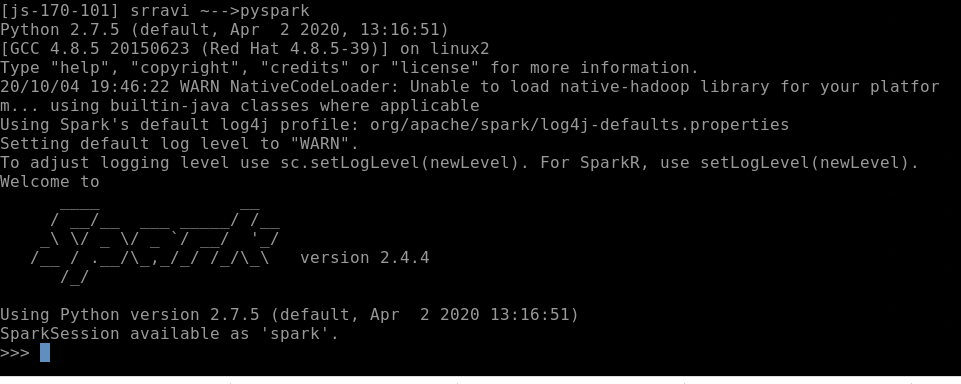




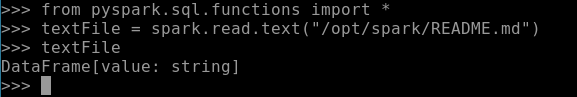
1. Spark Install



1. Pyspark Shell



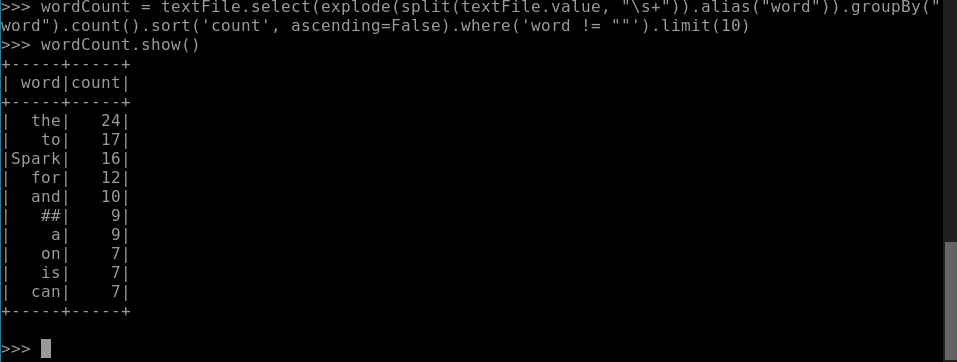
1. Pyspark Read



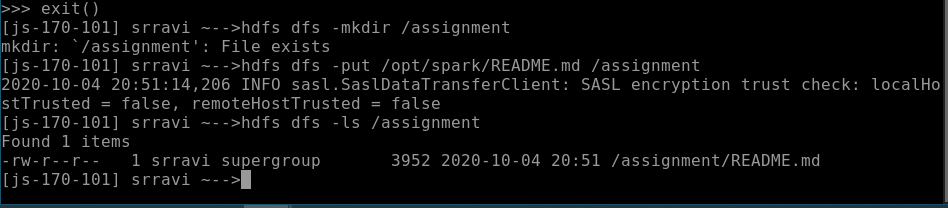
1. Pyspark Count Lines



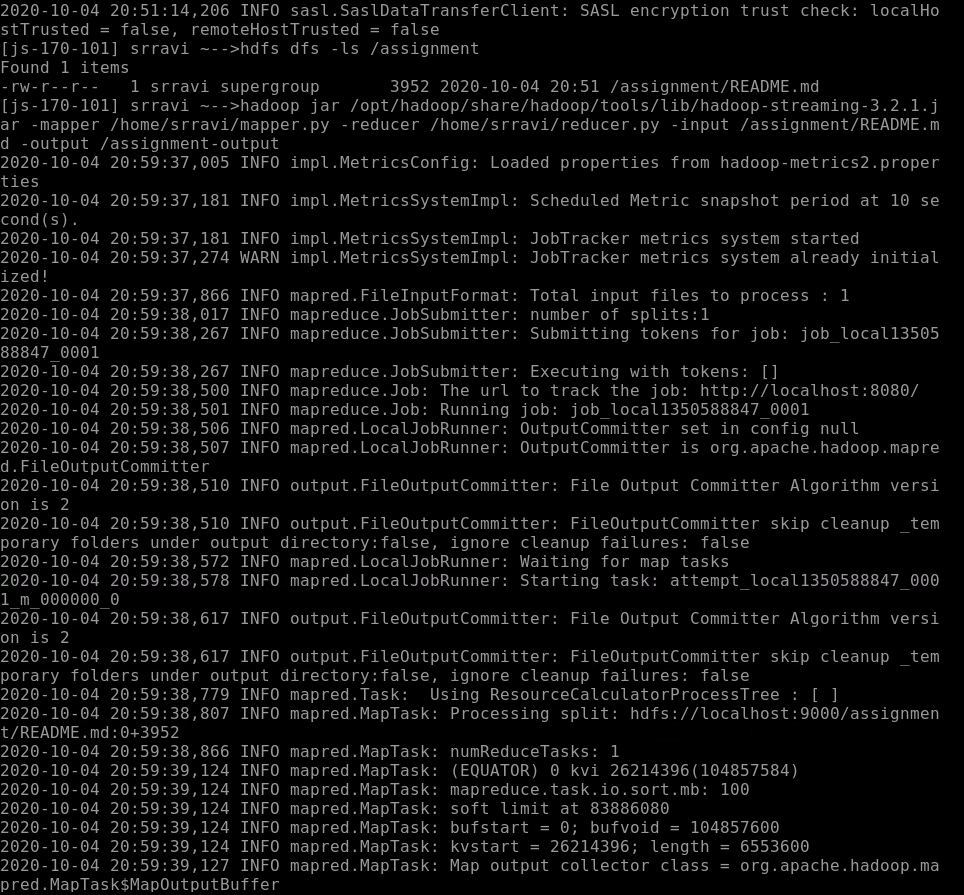
1. Word Count



1. MapReduce hdfs transfer



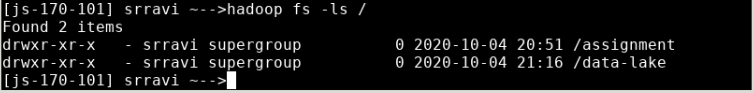
1. Mapreduce Word Count



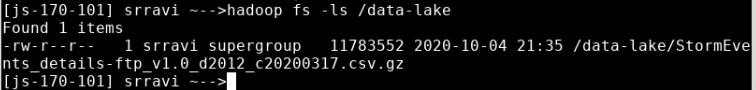
1. Mapreduce Output



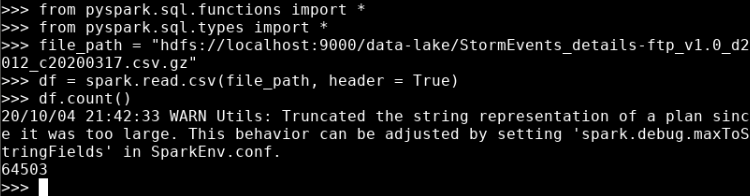
1. Create directory in HDFS



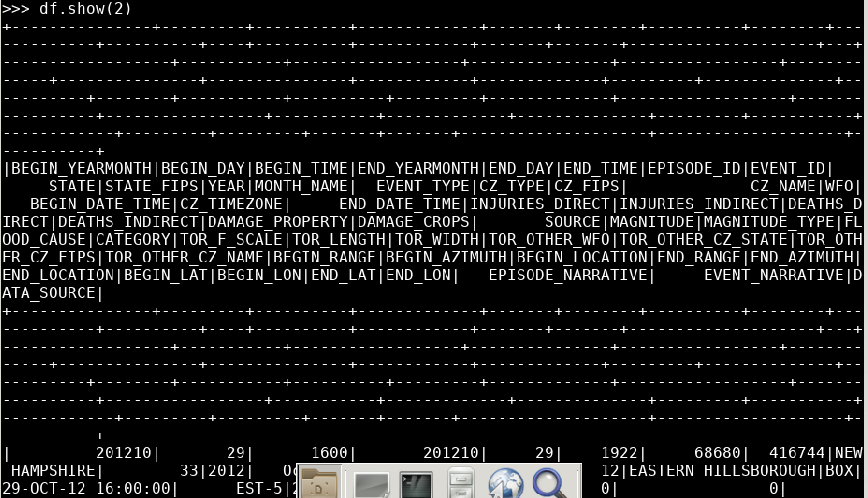
1. Copy file into HDFS



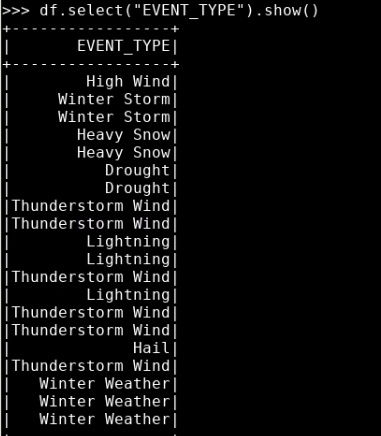
1. Reading the file



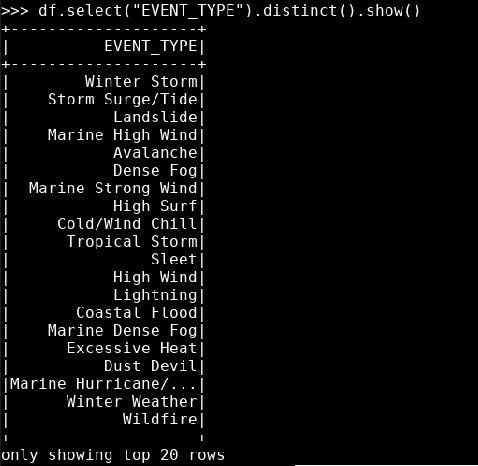
1. Show 2 rows

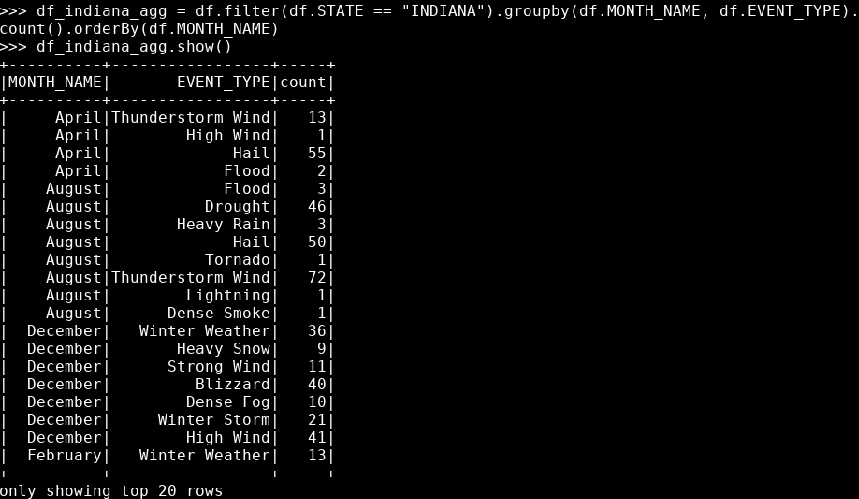


1. Show Event Type

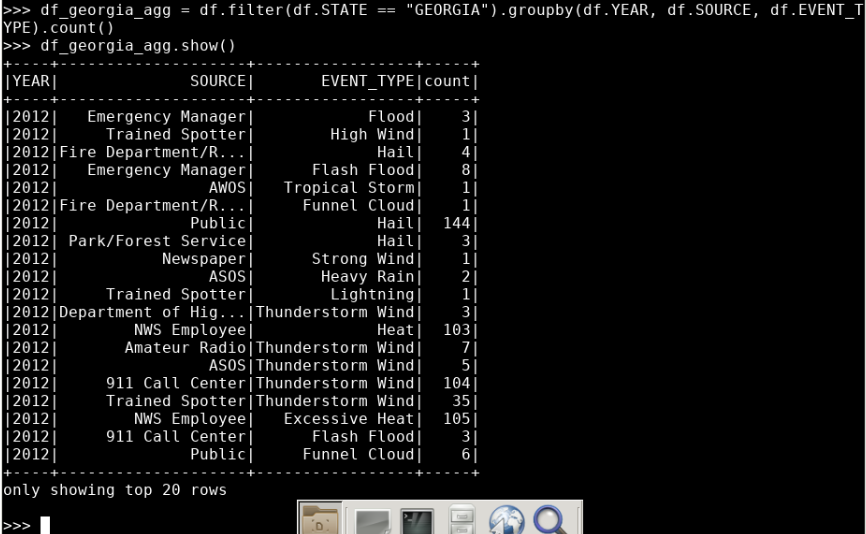


1. Show distinct event type





1. Using filter to count how many events of each type were reported by each source in Georgia for your selected year



This assignment went well for the most part. I had a hard time figuring out how to copy the scripts onto my VM. I downloaded Cyberduck and WinSCP. I created a public SSH key and added it to my VM but couldn’t figure out what to do on Cyberduck or WinSCP. I eventually found the VM web desktop and dragged and dropped the files into the VM. After that the installation of Hadoop and Spark were easy. Between the two methods of word frequencies, I preferred using Spark because I understood the code and the process better. Pyspark was more user friendly than MapReduce in terms of code and readability. They had similar processes in terms of implementation. I just had to import functions, read the file and run the queries against it. The query I wrote filters on Georgia and is grouped by year, source and then event type, which is what the assignment asked for. The assignment said to filter on a year, but there was only one year of data in the file. Even thought I struggled in the beginning, it was probably the easiest thing to struggle on and I really enjoyed the rest of the assignment. I think it helped me understand the material better.