Lab 3: Map/Reduce and Google Cloud

CS 660 Data Science at Scale

Jeremy Johnson

Lab 3 Problems (GFS)

1. Read Google File System Paper
2. Read Chapter 3 of “Hadoop the Definitive Guide” on HDFS
3. Prepare summary of the paper including
    1. 10-15 slides with notes
    2. Contribution
    3. Problem and Assumptions
    4. Overview (Design and Implementation)
    5. Performance Summary

Lab 3 Problems (Google Cloud)

4. Get Google Cloud coupon, logon and explore [https://cloud.google.com]
   4. Create a VM (default machine with Debian GNU/Linux) – allow http. ssh into the machine and verify python3 works
   5. Install pip and mrjob
   6. upload mrjob program and input and verify that it works. Time the execution and compare to equivalent python program
   7. Create storage bucket in Google storage and upload input file from (6), note time compared to uploading in 6.
   8. Use gsutil from VM to copy file from Google storage
   9. Create python script to create a large input for (6) by appending a bunch of copies of the input you used in (6). Time MRJob on the larger input.

Lab 3 Problems (Google Cloud)

5. Create a Hadoop cluster using dataproc – see instructions in MRJob documentation
    1. Create python script to create a large input for (4.6) by appending a bunch of copies of the input you used in (4.6). Time MRJob on the larger input.
    2. Explore different parameter settings and see how they affect the runtime