Project : MapReduce with Hadoop
Phase 1

CS: 417 Distributed Systems

The MapReduce project is divided into two phases. The first phase would be a simple MapReduce task on AWS. This would ensure that the your Hadoop and AWS work environment are properly set up. And then it would be easier for you all to work on phase two of the project.

Phase 1:

The objective of the phase one is to write a very simple Hadoop MapReduce program that counts the number of occurrences of each word in a text file. The input text file will be provided to you.

Sample input: inputfile.txt
[hello everyone I said hello]

Sample output:

hello	2
everyone	1
I	1
Said	1

You need to solve this using AWS. You may sign up for a free account using your rutgers.edu account.

Your S3 bucket should have the MapReduce jar file, input folder with input files in it and output folder for output files.

Also, give me the permissions to list, upload/delete, view, edit your S3 buckets so that I can see your results. My AWS account id is **neelmay.desai@rutgers.edu**

Follow the steps explained in the AWS recitation and you should be fine.

For submission:

For submitting the phase one on SAKAI, you need to submit the following files:

1) Your MapReduce java file
2) Output files generated
3) Jar file you created
4) A small (less than 1 page) PDF report explaining how you solved this problem and any difficulties/challenges you faced. Also write an approximate amount of time you spent on phase 1.

Some Useful Links:

Giving S3 Bucket permissions and adding a grantee
http://docs.aws.amazon.com/AmazonS3/latest/UG/EditingBucketPermissions.html

To create jar file:
http://www.skylit.com/javamethods/faqs/createjar.html

Basic MapReduce tutorial:
https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

AWS related links:
http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-launch-custom-jar-cli.html

MapReduce basic idea:
https://www.cs.rutgers.edu/~pxk/417/notes/content/mapreduce.html

Good Luck.