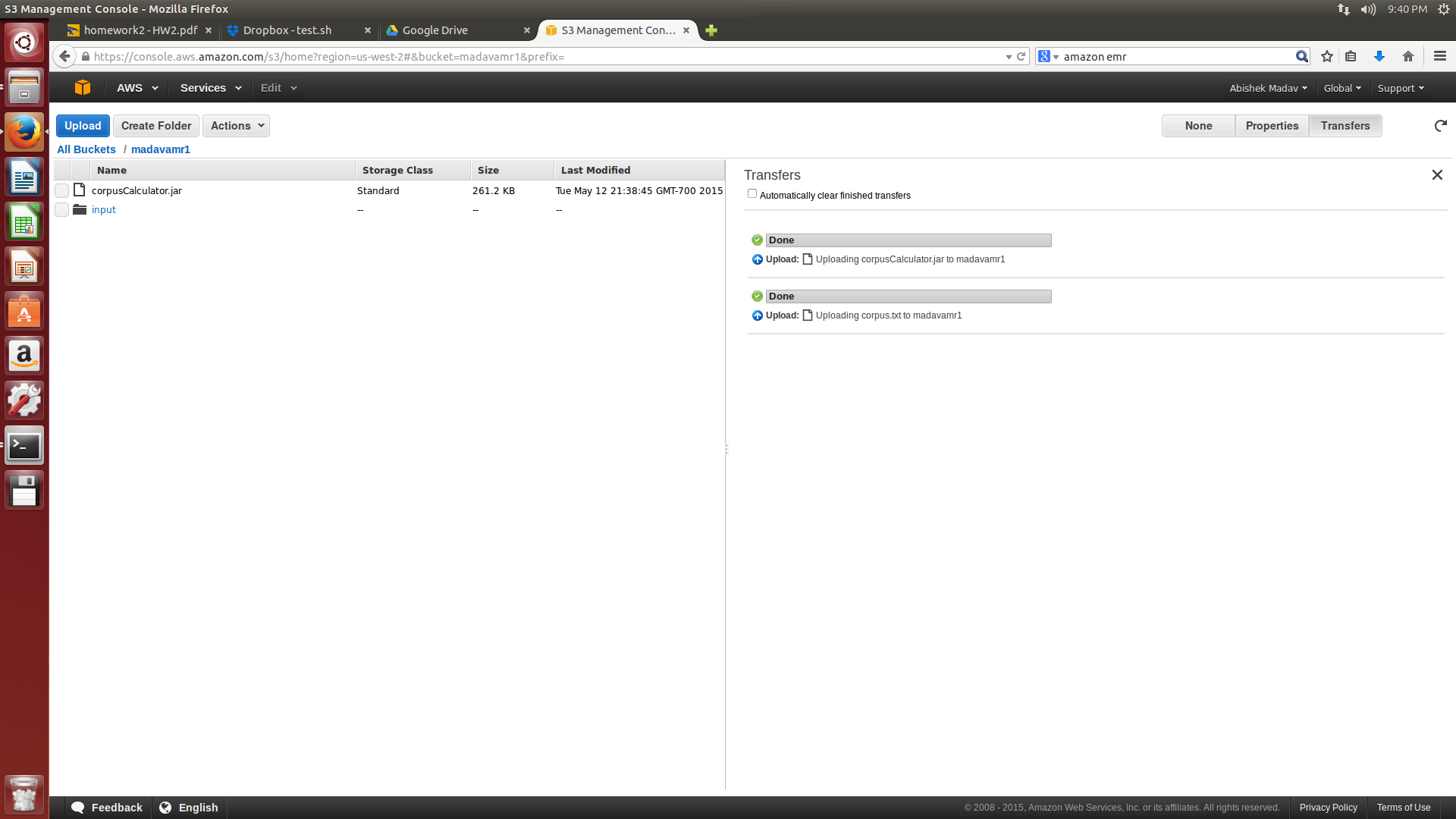
**Experience using EMR for Project 2 – Report**

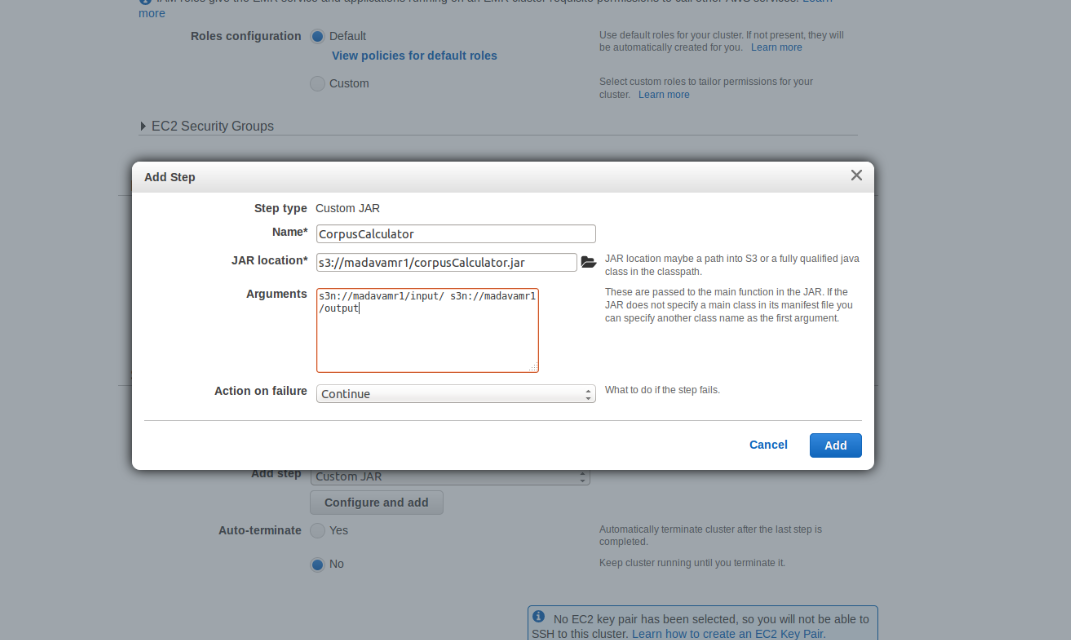
The project aimed to build a program to calculate the probability of sentences from a sample of corpus. The program for the same was developed using three map-reduce jobs. The implementation decision has chosen to keep the corpus file as a distributed cache among the reducers’ node for the third job. Amazon EMR was used to test the implementation in a clustered environment. The stages for the same can be show as below:

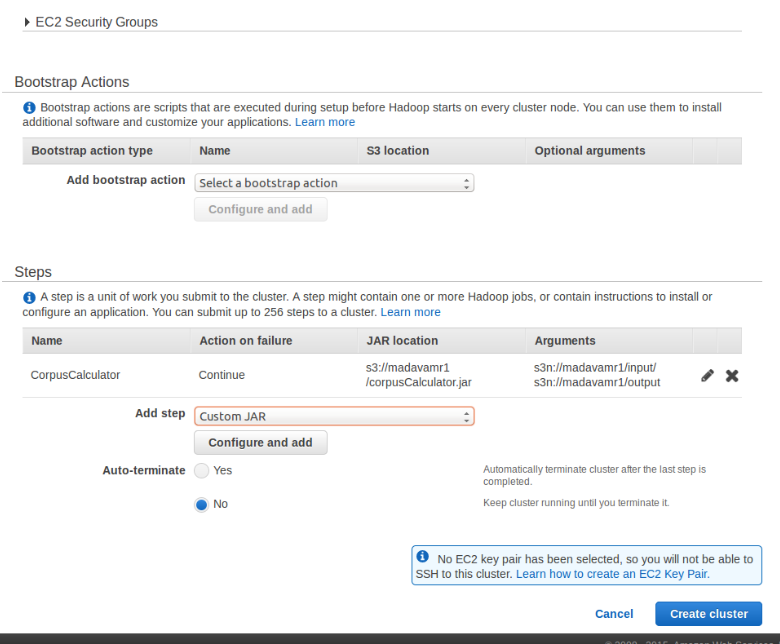
1) Setting up the Input folder in S3.



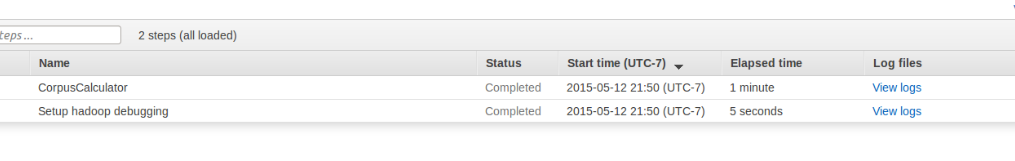
The bucket in S3 maintains the jar file and the required corpus input text file.

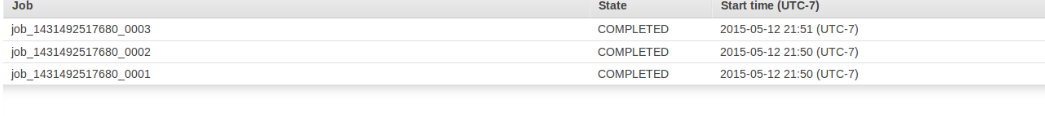
2) After the input path in set in the S3, the cluster is being configured for the execution as:-



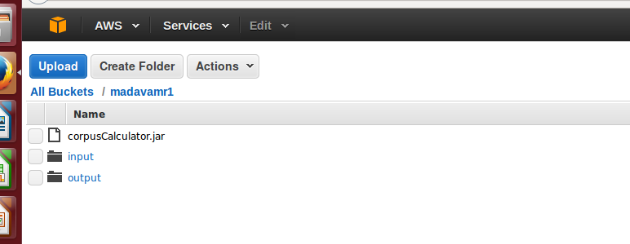


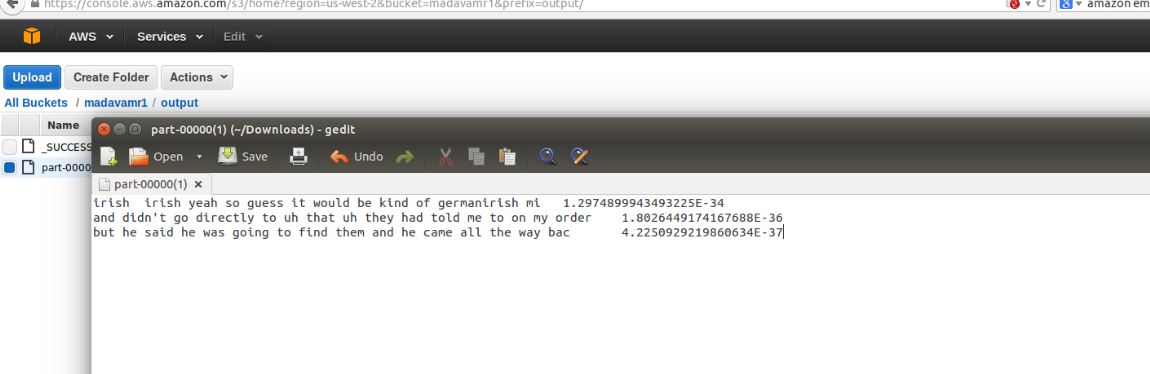
3) The ‘Create Cluster’ starts initializing the cluster and the jar execution on completion was shown as follows:





4) The output folder in the S3 folder contains the output as:





Since the implementation requires the corpus to be presented as a distributed cache, the location to the file from the S3 had been indicated to the reducer node. The EMR has helped me set a thorough understanding of the program execution in the cluster.