MSDS 7346
Cloud Computing
Mini Project 3 – Log Analysis

Name:

This is a mini project for MSDS 7346, Cloud Computing. For this assignment, turn in a single pdf file containing all of your answers. The file should be named ¡yourLastName¿MiniProject-Number.pdf. For example, the file name for my mini project 1 would be 'RafiqiMiniProject-Number.pdf'.

Collaboration is expected and encouraged; however, each student must hand in their own homework assignment. To the greatest extent possible, answers should not be copied but, instead, should be written in your own words. Copying answers from anywhere is plagiarism, this includes copying text directly from the textbook. Do not copy answers. Always use your own words and your own code. Directly under each question list all persons with whom you collaborated and list all resources used in arriving at your answer. Resources include but are not limited to the textbook used for this course, papers read on the topic, and Google search results. Don't forget to place your name on the first page of the pdf document.

Log Analysis Using EMR

We have now successfully built the building blocks of the MapReduce application in our previous session. You can use those program to build a log analysis application as discussed in the class. The application should take syslog-formatted log records as input and determine the frequency of log events using Amazon EMR to count the number of records per second. I have provided a sample syslog-formatted file as an input.

Although we are keeping this application primarily focused on log analysis, but counting and frequency analysis has many known uses in other data analysis situations. The MapReduce application is performing what is considered a summarization design pattern by simply summing up the values of a common key. Other real-world applications of this technique are:

Load or usage analysis

Many times it is useful to know how many users access a server or a website throughout a time period. Web access logs or application logs that include the time stamps of user events could be imported and processed with a similar MapReduce application to determine usage frequency. At this time we will not be doing load or usage analysis in this lab.

Question 1 : The objective of this lab is to get hands-on experience with developing and deploying MapReduce applications on AWS Elastic Map Reduce. This is simply the extension of what we did in the class. The only new item is enhancing the logic, you all should be familiar with steps of deploying an application in EMR.

In this lab you are asked to develop two simple MapReduce applications using Pig and Python:

1) Develop MapReduce Application to count number of events per second in the the sample log file (provided) using Python.
2) Develop MapReduce Application to count number of events per second in the the sample log file (provided) using Pig.
3) Deploy both applications in AWS EMR cluster
4) Submit screenshots

Submission: Submit different screen shots to show completion of each steps

Collaborators:

Resources: