EEL6935: 022A , Fall 2013: Assignments

Assignment - In progress

Complete the form, then choose the appropriate button at the bottom.

Title	Programming Assignment 1
Due	Sep 20, 2013 11:55 pm
Status	Not Started
Grade Scale	Points (max 100.0)

Instructions

Programming Assignment 1

Due by 11:55pm, Sep. 20, 2012.

Given the following data input (all words separated by tab or space, line by line)

Data sources (make **10 or more copies** of the following file into your input directory)
https://s3.amazonaws.com/uf-eel6935/data/bible.gz

develop a program using Amazon Elastic MapReduce (or on FutureGrid using MyHadoop submission or other cluster-based MapReduce on EC2 or FutureGrid) to achieve **the following objectives:**

Task 1 (10 points). Count one-word frequency as in the wordcount example. (This is required. Copy/paste source code is allowed.)
Task 2 (50 points). Count double-word frequency. For example, "I am given an opportunity to use EC2 so I am very happy and very happy and very happy." The output would be something as follows (in any reasonable order):

I am 2
am given 1
given an 1
an opportunity 1
opportunity to 1
to use 1
use EC2 1
EC2 so 1
so I 1
am very 1
very happy 3
happy and 2
andy very 2

Task 3 (40 points). Using DistributedCache. Find frequency of one-words in another given list (attached).
It is to use the given list to find one-word frequency in bible.gz. For example, small list has "hello world gator". Then you need to find frequency of "hello", "world", and "gator" in bible.gz

Requirements:

* Please use non-streaming style programs (either Java, or C++ via Hadoop Pipes)
* Please use no less than 3 VM instances

Some references:
http://developer.yahoo.com/hadoop/tutorial/index.html
http://hadoop.apache.org/common/docs/r0.20.2/mapred_tutorial.html

http://wiki.apache.org/hadoop/EclipseEnvironment

http://blog.cloudera.com/blog/2013/05/how-to-configure-eclipse-for-hadoop-contributions/

Submissions:
1. Your program, well docoumented in code and in Readme.doc (the document should include a link to AWS outputdata)
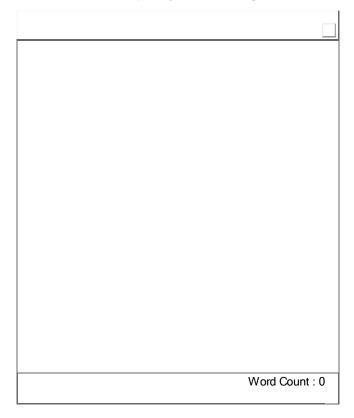2. Results
3. Screenshots

Additional resources for assignment

word-patterns.txt (1 KB; Sep 3, 2013 11:12 pm)

Submission

Assignment Text

This assignment allows submissions using both the text box below and attached documents. Type your comments in the box below and use the Add Attachments button to include other documents. Save frequently while working.

Word Count : 0

Attachments