IEMS 5709 Spring 2016 Homework 1

Release date: Jan 27, 2016

Due date: Feb 22, 2016 23:59

The solution will be posted right after the deadline, so no late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student__________________________) Date:_______________________________

Name___________________________________ SID_________________________________

Submission notice:

- Submit your homework via the elearning system

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

Q1 [30 marks]: Multi-node Hadoop cluster setup

In this problem, you are required to setup a Hadoop cluster using Amazon EC2, or Google Compute Engine, or Windows Azure, or the OpenStack deployed in IE department Data Intensive Cluster (DIC). The tradeoff is that OpenStack is not very stable and sometimes may crash, while other platforms may charge you if the usage exceeds the quota of the free tier subscription. Ref [1] - [4] provide the tutorial for each platform.

In order to setup a Hadoop cluster with multiple VMs, you need to setup a single-node Hadoop for each VM. Then change the configurations of Hadoop master and slaves. Ref [5] - [6] provide the tutorial of installing single-node Hadoop and Hadoop cluster. Notice that in Ref [5] - [6], the OS and Hadoop version may be different, but the configuration is similar.

Grading Scheme:

(a) Install a Single-Node Hadoop (Hadoop version: 2.7.1). [10 marks]

(b) Install and setup a Multi-Node Hadoop cluster **with 4 VMs (1 master and 3 slaves)**. [20 marks]

Hints:

1. Launching instances with Ubuntu 14.04 LTS is recommended.
2. After installing a single-node Hadoop, you can save the system image and launch VM with that image, so that you can save the effort of installing the single-node Hadoop on each VM.
3. After finishing step (a), you are suggested to run some sample hadoop jobs to verify if the installation is successful.

Submission Reuqirements:

1. Use command *'jps'* to make sure all the process are running.
2. Include all the keysteps, together with screenshots, into the report.

Q2 [70 marks]: Basic text processing with Hadoop

We will accomplish some simple tasks of text proessing in this question.

(a) [10 marks] Setup

- We have downloaded the top 100 popular ebooks from Gutenberg project which offers over 45,000 free ebooks. We also add the works of Shakespeare into the data set.
- Please download the data set (93M in total) from the following URL.
https://github.com/YangRonghai/100ebooks/blob/master/data.tar.gz?raw=true
- Upload the "data/" directory to HDFS as input for following tasks.

For each part of the problem below, record the system time it takes for your Hadoop programme to finish the task.

(b) [10 marks] Basic word counting

What is the total number of words in all these books? By "word", we mean English words in the text, i.e. excluding punctuations, page numbers and special symbols.

(c) [10 marks] File counting

How many works are there?

Hint: No need to write code for this subproblem. Check the log of last problem and argue based on your knowledge of Hadoop scheduling.

(d) [10 marks] Distinct word counting

How many distinct words appear in these books? For example, even though "The" may appear many times, it should only be counted once.

(e) [10 marks] Frequent words

Find out the Top100 most frequently used words in these books and their corresponding relative frequencies (appearance percentage in the works). Illustrate your results using a histogram.

(f) [10 marks] Frequent words excluding stop words

In the last subproblem, you may find some of the Top100 most frequently used words not quite interesting (because they are so common that they are widely used in any writings). Those are called "stopwords" (Ref [7]) in the context of Natural Language Processing (NLP). What are the Top10 most frequently used words excluding stopwords? Please use the sample stopwords list in Ref [8].

Requirement:

- Use “common-english-words.txt” as an individual file. In other words, do not directly include the content in your code; load it in your mapper and perform the filtering.

Hint: use “file” option to specify more supplementary files for the Hadoop job(s).

(g) [10 marks] Run with different number of mappers and reducers

Rerun your programme for (f) multiple times while modifying the number of mappers and reducers for your MapReduce job(s) each time. You need to examine and report the performance of your programme for at least 4 different runs. Each run should use a different combination of number of mappers and reducers. For each run, performance statistics to be reported should include: (i) the time consumed by the entire MapReduce job(s) and the maximum, minimum and average time consumed by (ii) mapper tasks and (iii) reducer tasks. State clearly the machine type and the computing platform you use, e.g. using the Azure account or OpenStack (with specified model) or something else?

Tabulate the time consumption for each MapReduce job and its tasks. One example is given in the following table. Explain your observations.

Maximum mapper time	Minimum mapper time	Average mapper time	Maximum reducer time	Minimum reducer time	Average reducer time	Total job
60s	40s	50s	60s	40s	50s	2.5 min

Hints:

1. If you are using Java, you can specify the number of reducers with the following code:

   ```
   job.setNumReduceTasks(20)
   ```

   If you use Hadoop streaming with Python, you can specify it via the following command option:

   ```
   -D mapred.reduce.tasks=20
   ```

2. You can set the number of mappers in a similar way (`job.setNumMapTasks / -D mapred.map.tasks`). If this does not work, you may need to modify the split size in the $HADOOP_HOME/etc/hadoop/mapred-site.xml:

   ```
   mapred.min.split.size = 268435456 (256M)
   ```

   You can go to Ref[9] for more information

3. You DO NOT need to write a separate program for each sub-questoin. You can write one program giving out all the outputs.

Submission Requirements:

1. Include all the codes and commands.
2. Report results (statistics/figures) as required in each sub-question

References:

[1] Microsoft Azure Tutorial
https://azure.microsoft.com/en-us/get-started/
[2] Google Compute Engine Tutorial:
https://cloud.google.com/compute/docs/quickstart
[3] AWS Tutorial:
https://aws.amazon.com/getting-started
[4] OpenStack in DIC Manual:
http://mobitec.ie.cuhk.edu.hk/iems5709Spring2016/tutorial/launch_VMs.pdf
[5] Single-Node Hadoop setup in Ubuntu:
http://pingax.com/install-hadoop2-6-0-on-ubuntu/
[6] Hadoop cluster setup in Ubuntu:
http://pingax.com/install-apache-hadoop-ubuntu-cluster-setup/
[7] What are stop words?
http://en.wikipedia.org/wiki/Stop_words
[8] Stop Words List
http://www.textfixer.com/resources/common-english-words.txt
[9] How many Mappers and Reducers?
http://wiki.apache.org/hadoop/HowManyMapsAndReduces