CS 425 Distributed Systems, Fall 2010, University of Illinois at Urbana-Champaign

Homework 1 [Introduction and Cloud Computing]: Solution

Out: Aug 31, 2010. Due Date: Sep 9, 2010 (Thursday).

Note: (1) Please hand in **hardcopy solutions that are typed** (you may use your favorite word processor). We will not accept handwritten solutions. Figures and equations (if any) may be drawn by hand. (2) Please **start each problem on a fresh sheet**, and **type your name at the top of each sheet**. (3) Homeworks will be **due at the beginning of class on the day of the deadline.** (4) Each problem has the same grade value as the others. (5) Unless otherwise specified, the only resources you can avail of in your HWs are the provided course materials (slides, textbooks, etc.), and communication with instructor/TA via newsgroup and e-mail.

Relevant Lectures for this Homework: 1-3.

1. What is the main difference between an HTTP GET request and an HTTP PUT request? Hint: For this question, you can use the World Wide Web as a resource.

   Solution:

   HTTP GET: Request to read a Web page
   HTTP PUT: Request to store a Web page

2. A medium-sized organization wishes to run a service for 2048 CPUs and 600 Terabytes, lasting M months. The service will run atop the Amazon EC2/S3 cloud infrastructure. Amazon S3 has told you that the storage will cost 12 cents per GB-month. What is the breakeven time for storage, assuming 100% utilization (i.e., find the value of M)? You are given that the total cost for buying this storage is $ 400 K, and the additional per-month cost of operating it is $ 8 K. [Note: you do not need to include the cost of power, networking, etc., as they are already a part of the operating cost above.]

   Solution:

   Monthly cost for outsourcing via Amazon S3: $600\,TB \times 1024\,GB/TB \times \$0.12/GB$ = $\$73728$
   Monthly cost of owning: $\$400K/M + \$8K$

   More preferable to own if:

   $400K/M + 8K < 73728$
   $\Rightarrow M > 6.086$

3. You are given a MapReduce implementation where you only have to write the Map and Reduce functions. The Map function you will write takes as input a (key, value) record and returns either a (key, value) record or nothing. The Reduce function you will write takes as input (key, list of all values for that key) and returns either a record or nothing. The framework already takes care of iterating the Map function over all the records in the input file, key-based intermediate data transfer between Map and Reduce, and storing the returned value of Reduce – you do not have to worry about these.

   You are now given an input file which contains comprehensive information about a social network that has asymmetrical (directed) links, i.e., a network where users 'follow' other users but not necessarily vice-versa (e.g., twitter). Each record in this input file is **(userid-a, userid-b)**, where **userid-a** 'follows' **userid-b** (i.e., points to it). Note that this record tells you nothing about whether or not **userid-b** follows **userid-a**.

   Write a MapReduce program (i.e., Map function and Reduce function) that outputs all pairs of userids who follow each other. Pseudocode is OK.

   Solution:

Function 1 MAP ($userid\text{-}a,\ userid\text{-}b$)

1: **if** $userid\text{-}a < userid\text{-}b$ **then** //lexicographic ordering
2: $string \leftarrow$ "$userid\text{-}a,\ userid\text{-}b$"
3: **else**
4: $string \leftarrow$ "$userid\text{-}b,\ userid\text{-}a$"
5: **end if**
6: **return** $(string,\ 1)$

Function 2 REDUCE ($key,\ list\ of\ values$)

1: **if** $sum(list\ of\ values) = 2$ **then**
2: **return** key
3: **else**
4: **return**
5: **end if**

4. A MapReduce program you have written to process wikipedia pages has one Reduce task but can have multiple Map tasks. Each Map task can process an average of 10 wikipedia pages per second. You have a dataset with 3.0 Million wikipedia pages. You would like to process this dataset within 10 hours. How many Map tasks would you need in the datacenter? You may assume that the datacenter network has infinite bandwidth, that the datacenter is infinitely large, that the Reduce task takes negligible additional time, and that there are no failures or other applications running in the datacenter.

   Solution:

   Let, N = the number of Map tasks

   $\frac{3\times10^6\ pages}{N\times10\ pages/sec} = (10 \times 3600)\ sec$

   N = 8.33

   Therefore, 9 Map tasks will be needed.

5. Which companies invented each of the following software systems: (i) MapReduce, (ii) Hadoop, (iii) Pig Latin, (iv) DryadLINQ. Hint: For this question, you can use the World Wide Web as a resource.

   Solution:

   (i) MapReduce: Google
   (ii) Hadoop: Apache Software Foundation
   (iii) Pig Latin: Yahoo!
   (iv) DryadLINQ: Microsoft