Autumn 2019

Homework 2

Big Data Management

1. Read Chapter 25.4.2 in the book “Fundamentals of Database Systems (7th Edition)”. The electrical version is available online in the library of our university. Use your words to describe the MapReduce procedure for Sort-Merge Join and Map-side Hash Join. (2 marks)

2. Download the text to Alice's Adventures from http://www.gutenberg.org/files/11/11-0.txt (If it redirects you to a page with a welcome popup, click on the "Plain Text UTF-8" option on that page or just download the attachment below) and run wordcount on it. This can be done by using hadoop commands. How many times does the word Cheshire occur? (Do not include the word 'Cheshire with an apostrophe. The string -->'Cheshire<-- does not count) Paste the screenshot after you ran the wordcount command. (2 marks)

3. The set of example MapReduce applications includes wordmedian, which computes the median length of words in a text file. If you run wordmedian using words.txt (the Shakespeare text) as input, what is the median word length? Paste the screenshot after you ran the Mapreduce codes.

   Note that wordmedian prints the median length to the terminal at the end of the MapReduce job; the output file does not contain the median length. (2 marks)

4. Which are the typical application scenarios for a MapReduce program? (Select 2 choices) (0.5 mark)

A. Perform the matrix multiplication and other complicated operations
B. Run machine learning algorithms with many iterations
C. Compute the inverted indices
D. Summarize the number of pages crawled per host

5. MapReduce is an abstraction to hide the following messy details of parallelization, including (Select 3 choices) (0.5 mark)

   A. fault-tolerance
   B. data distribution
   C. high performance
   D. load balancing

6. Which are the correct statements on the functions of Mapper and Reducer? including (Select 2 choices) (0.5 mark)

A. Each Mapper can do something to each individual key-value pair.
B. Each Mapper can look at key-value pairs of other mappers.
C. Each Reducer can aggregate data.
D. Each Reduce can look at multiple values from other reducers.

7. You are writing a 10GB file to a HDFS filesystem with a default block size of 128MB. How many blocks will the file be broken into? (Select 1 choice) (0.5 mark)

A.1280
B.160
C.80
D.480