Assignment 2 – Hadoop and Hbase – Distributed Systems

Assessment Summary

Weighting: 15%

Due Date: 11pm Sun 4 May (End of Week 12)

Group Assignment

Submission

One word document containing all your answers

Assignment Overview

In this assignment, you are asked to write a report to include the following parts:

(1) A description on how JSON objects like the ones given the file 20objects.txt can be stored in Hadoop. The stored objects are called data.

(2) A query that you identify for a certain application such that the result of the query can be calculated from the stored data

(3) A description of how MapReduce can be used to answer the query

(4) A logic design of Hbase table schemas for storing the same objects in 20objects.txt in HBase

(5) A description of how Hbase tables can be accessed and processed to answer the same query in Part (2)

(6) A comparison of Hadoop and Hbase for their cons and pros in answering the query in your described storage design.

This type of report supplies critical information to help an organization to decide whether a new system like Hadoop or Hbase is suitable for its business. Note that it is not reasonable to put a new system on trial in business operations without pre-analysis. This assignment serves to be this type of critical analysis of a software system before its employment.

The materials that you use include lecture slides, the recommended videos and tutorials (week 8 folder), and other internet resources and/or books that you can find. Note that you must NOT copy these materials; otherwise, you commit plagiarism and the university formal plagiarism procedure will be used.

You complete the assignment together with two other students whom you choose yourself (in groups of three). If you do the assignment by your own or do it in a group of two, you must complete all components properly to get full marks. You must complete the parts in order. Otherwise, your discussion and references will be ambiguous.

If you do the assignment with other students, only one of you submits. Make sure that at the very **beginning** of the word document, you list all the IDs and names.

Application and requirements

Before you do the assignment, you should work on the examples in the lecture slides to ensure that you have proper understanding. If necessary, you may read additional references as suggested in the slides.

You are given a word document as a template with sample samples, and you put your answers in the respective sections. Following are some specific information/requirements for the parts.

(1) In Task1, you assume that you will have a Hadoop system that has one master node with id m1 and three slave nodes with ids v1, v2, and v3 respectively. Then you partition the dataset 20objects.txt into 3 fragments (also called blocks or partitions) and decide the nodes to store the fragments. Each fragment is replicated such that 2 copies for each fragment are stored. You include a table to show the fragments, the membership of the JSON objects in the fragments, and the nodes that store fragments. A JSON object is represented by its id in the table. You also plot a diagram to show how you distribute the fragments and their replications on the nodes.

(2) Identify a query for an application such that the answer to this query can be derived from the stored data/objects. The simplest query can be counting the number of words in all tweets. If you choose a more complicated query, you have more space to develop your answers in the following steps and consequently you will get better marks. On the

contrary, if you choose a simple query, you will have short answers for the following parts and you will get less marks.

(3) Describe how MapReduce is used to answer the above query. You need to describe how map() works, how the shuffler works, and how reduce() works. The discussion must refer to your fragments and objects, and specific nodes as shown. Some tables are given in the template for you do use. If you are not happy with the structure of these tables, you can design your own, but make sure that you supply all necessary information.

(4) Design logical schemas in Hbase to store the twitter objects. Show the schemas together with sample data (2 objects with their attributes and values).

Each schema would include a row key, and some column families which you design. The sample data would be presented as attribute-value pairs in each column family. You do not need to consider timestamps and regions.

You write no more than 2 dot points about the reasons why you choose such row keys and column families.

(5) For the same query used in Part (2), describe how Hbase data should be accessed and processed to answer the query based on your design. Your description should refer to your design and your sample data in Step (4). Ambiguous description values very little.

(6) Summarize the pros and cons of Hadoop and MapReduce respectively and justify them referring to your examples (your data, query, and descriptions). Without examples, you receive only very little marks.

Marking Criteria

You must complete the parts in order. Full marks of the parts are below.

Full marks	Part1	Part2	Part3	Part4	Part 5	Part 6
	2	2	3	3	3	2

The marking will consider the following factors

- Correctness and quality of the description.
- The descriptions refer to specific nodes, data fragments or tables in the analysis
- Correct concepts and logics
- The writing as a whole makes sense and is correct.

Plagiarism

If your solution, or part of it, is not written by you (your idea and your own typing), you commit plagiarism. The investigation will involve an oral or written test. If you commit plagiarism, you will be penalized and a record will be kept in your file.

Extensions

Extensions for assignments are available under the following conditions

- permanent or temporary disability, or
- compassionate grounds

In all cases, documentary evidence (e.g. medical certificate, road accident report, obituary) must be presented to the Course Coordinator. A medical certificate produced on or after the due date will not be accepted unless you are hospitalized.

If you apply for extension within 24 hours before the deadline, you must see the course coordinator in person unless you are in an emergency situation like being admitted in a hospital.

Late Penalties

Unless you have an extension, late submission will incur a penalty of 30% deduction per day (or part of it) of lateness.