TDT4305 2021 - Assignment 1 solutions

Intro

1) For each of the three Vs, answer the following:
   a) Why is this property showing up in the Big Data era and not previously?
      - *See [ElmasriNavathe] section 25.1.*
   b) What challenges does this property give for traditional RDBMSs?
      - *See [ElmasriNavathe] section 25.1.*
2) Given Big Data from a field of your choice, how could you use it to create value for yourself or others? E.g. Big Data in medicine can be used for diagnosing diseases based on patients' health history.
   - *See [ElmasriNavathe] section 25.1 for inspiration.*
3) What are the challenges in ensuring the trustworthiness of Big Data?
   - *See [ElmasriNavathe] section 25.1.*

MapReduce and HDFS

1) Solve the following exercises from the main course book MMDS:

   2.2.1

   a) We expect there to be significant skew, since there is often a big variation in the lengths of the value lists for different keys, so different reducers take different amounts of time.
   b) We expect the impact of skew to be less if we combine the reducers into 10 Reduce tasks. If keys are sent randomly to Reduce tasks, we can expect there will be some averaging of the total time required by the different Reduce tasks. On the other hand, we may not benefit from using 10,000 Reduce tasks, since there is an overhead associated with each task we create, and in general, the number of Reduce tasks should be lower than the number of keys.

   2.3.1

   a) The Map function simply outputs each number. The key can be ignored. A single Reducer scans all received integers and outputs the largest. Since the *max* function is associative and commutative, a combiner can be used after the Map task.
   b) As above, the Map function simply outputs each number and keys can be ignored. A single Reducer sums and counts all numbers and computes the average. Since the *average* function is not associative and commutative, a Combiner cannot be used.

c) We use the same strategy as word count, but ignore the values, i.e. the number of times a word occurred. The Map function again simply outputs each number. A single Reducer is here an identity function. Since this corresponds to the *unique* function, which is associative and commutative, a combiner can be used after the Map task.

2.3.2

- The algorithm does not need generalization to work on non-square matrices.

2) What is the role of the DFS (GFS or HDFS) in a MapReduce system?
   - *See [DeanGhemawat] section 3.4.*
3) What is the difference between a NameNode and a DataNode in HDFS?
   - *See [HDFS] section 2.*
4) What is a data block in HDFS and how is it replicated across data nodes?
   - *See [HDFS] section 3.*