UNIVERSITY OF OTAGO EXAMINATIONS 2019

Information Science

INFO 408

Management of Large-Scale Data
Semester One

(TIME ALLOWED: 2 HOURS)

This examination paper comprises 3 pages.

Candidates should answer questions as follows:

Answer THREE of the five questions (total of 60 marks).

The following material is provided:

N/A

Use of calculators:

No calculators permitted.

Candidates are permitted copies of:

N/A

Other Instructions:

N/A

TURN OVER

ANSWER THREE QUESTIONS (TOTAL 60 MARKS).
Read all questions before choosing which to answer. If you answer more than three questions, only the *first three* will be marked, in the order they appear in your exam script. Cross out anything you do not wish to be marked. **The suggested time allocation for answering each question is 40 minutes.**

1. Three of the technical challenges of managing the storage and retrieval of big data are known as "the three Vs". Explain the challenge(s) that each "V" brings to data management, illustrating these challenges with a realistic data management scenario. Discuss developments in database management systems that help to alleviate each challenge, by describing features and restrictions (if any) of relevant database management systems (DBMSs). (**Note:** Some discussions refer to four or five Vs, which also include one or both of the non-technical issues of *veracity* and *value* of data—these are *excluded* from the scope of this question.)
(20 marks)

2. In the context of the CAP theorem for distributed systems, Daniel Abadi wrote "in reality, there are only two types of systems: CP/CA and AP. [So] if there is a partition, does the system give up availability or consistency?"

   (a) Explain the CAP theorem, including a clear description of what the C, A, and P stand for. (6 marks)

   (b) Consider a distributed key-value store that uses the quorum consensus protocol, with N database nodes, a read quorum of R nodes, and a write quorum of W nodes. Explain how the protocol works when a client (i) reads and (ii) writes. (6 marks)

   (c) With reference to the read and write operations you discussed in (b), discuss how the choice of values for R and W affects whether the database will give up availability or consistency when a network partition occurs.
(8 marks)

3. Compare and contrast Apache Spark with older tools such as Hadoop that are based on the Hadoop File System (HDFS) and MapReduce, with particular focus on differences in the way data are managed within each tool, the way(s) in which each tool can query and process data, and the kinds of problems that each tool is best suited to. Use examples to illustrate your answer. (20 marks)

4. Compare and contrast "NewSQL" DBMSs with "traditional" SQL DBMSs, with particular focus on their ability to scale to massive distributed data sets. Discuss the technical and architectural differences between these two approaches, as well as their relative advantages and disadvantages. (20 marks)

5. The performance and scalability of large distributed data management systems can be affected by many factors, such as the system architecture, how DBMSs work internally, and the design of database schemas, applications, and queries. Discuss *four* potential performance bottlenecks for such systems. Explain why each bottleneck impacts performance and scalability, and suggest ways to reduce or eliminate this impact. Distinguish clearly between issues that are a consequence of run-time configuration (i.e., operational) versus those that are inherent in the system design (i.e., structural). (20 marks)

END

INFO408

PLEASE DO NOT

TURN OVER YOUR

EXAMINATION

PAPER UNTIL

INSTRUCTED TO

DO SO