MSDS 7330
File Organization and Database Management
Midterm Exam

Final Exam: Directions

This is a final exam for DS7330, File Organization and Database Management. This document contains the questions for the exam. For your answers, create a Word document that clearly identifies every question number and your answer to that question. Name the Word file containing your answers 'yourLastNameDS7330-20150421.doc'. For example, my Word file would have the name RafiqiDS7330-20150421.doc.

Answer each question fully and completely. Show all of your work and state your assumptions where appropriate. Each answer is worth an equal number of points. The questions may have hints embedded within them regarding the answer. Follow these hints as appropriate for full points. Collaboration is expected and encouraged; however, each student must hand in their own exam. To the greatest extent possible, answers should not be copied but, instead, should be written in your own words. Copying answers from anywhere is plagiarism, this includes copying text directly from the textbook. Do not copy answers. Always use your own words. Directly under each question list all persons with whom you collaborated and list all resources used in arriving at your answer. Resources include but are not limited to the textbook used for this course, papers read on the topic, and Google search results. Don't forget to place your name in the Word document itself.

Final Exam: Questions

1) In traditional RDBMS, a table may contain which of the following?
    - a) Complex data structures.
    - b) Arrays.
    - c) Embedded classes.
    - d) All of the above.
    - e) None of the above.

2) Which of the following is a potential driver for the development and/or adoption of a NoSQL database?

a) Simplicity in the database design (e.g., the NoSQL database is simpler than the corresponding SQL database).

b) Minimize the number of translations from how the data is stored to how the data is consumed.

c) Flexibility in the database design.

d) All of the above. [Hint: a description for this answer would explain the benefit for each of the above possible answers.]

e) None of the above. [Hint: a description for this answer would explain how SQL provides each of the above possible answers.]

3) In the paper "MapReduce: Simplified Data Processing on Large Clusters," by Jeffrey Dean and Sanjay Ghemawat, published in the Communications of the ACM, Jan 2008, the concept of performing a map function followed by a reduce function is introduced. The computations take *what* as input and produce *what* as output?

a) Tables, Tables.

b) Key/Value pairs, Key/Value pairs.

c) Documents, Tables.

d) XML, XML.

4) In the paper "MapReduce: Simplified Data Processing on Large Clusters," by Jeffrey Dean and Sanjay Ghemawat, published in the Communications of the ACM, Jan 2008, the concept of performing a map function followed by a reduce function is introduced. Which of the following is true of the way in which the MapReduce operations work?

a) Each instance of the map function receives the complete set of data.

b) Each instance of the MapReduce functionality randomly chooses whether to perform the map function or the reduce function.

c) Each instance of the reduce function receives all of the intermediate data from every map function.

d) All of the above. [Hint: a description for this answer would explain the basic operation for each of the above possible answers.]

e) None of the above. [Hint: a description for this answer would provide the correct operation and explain the basic of this operation for each of the above possible answers.]

5) In the paper “MapReduce: Simplified Data Processing on Large Clusters,” by Jeffrey Dean and Sanjay Ghemawat, published in the Communications of the ACM, Jan 2008, the concept of performing a map function followed by a reduce function is introduced. Which of the following statements is true about the intermediate data generated by the map worker functions and the reduce worker functions? [Hint: If a single answer (a), (b) or (c) is chosen, explain why the chosen answer is correct and the other two are incorrect.]

a) The map intermediate results are stored on local machines.
b) The reduce worker results are stored on local machines.
c) The map intermediate results are never combined or manipulated before being input to the reduce worker function. d) All of the above. [Hint: a description for this answer would explain the basic operation for the map and reduce worker functions.]
e) None of the above. [Hint: a description for this answer would provide the correct operation and explain the basic operation for the map and reduce worker functions.]

6) In the paper “MapReduce: Simplified Data Processing on Large Clusters,” by Jeffrey Dean and Sanjay Ghemawat, published in the Communications of the ACM, Jan 2008, the concept of performing a map function followed by a reduce function is introduced. Which of the following are identified as advantages of MapReduce? [Hint: If a single answer (a), (b) or (c) is chosen, explain why the chosen answer is correct and the other two are incorrect.]

a) MapReduce allows programmers with no parallel system experience to exploit large amounts of resources easily.

b) MapReduce makes it possible to write simple programs that run efficiently on large numbers of machines.

c) MapReduce allows for a large variety of problems to be easily expressible as MapReduce problems.

d) All of the above. [Hint: a description for this answer would explain why each of the answers above provides value.]

e) None of the above. [Hint: a description for this answer would explain why each of the answers above is not a value provided by MapReduce.]

7) In the paper “BigTable: A Distributed Storage System for Structured Data,” by Chang *et al.*, published in OSDI 2006, a distributed storage system for managing structured data that scales to petabytes is presented. The goals for BigTable include which of the following?

a) Scalability
b) High performance
c) High availability
d) Wide applicability
e) All of the above. [Hint: a description of this answer would explain each of the answers a-d above.]
f) None of the above. [Hint: a description for this answer would provide an explanation of the goals for BigTable.]

8) In the paper “BigTable: A Distributed Storage System for Structured Data,” by Chang *et al.*, published in OSDI 2006, a distributed storage system for managing structured data that scales to petabytes is presented. Which of the following best describes BigTable?

a) BigTable provides a scalable implementation of a traditional relational database.
b) BigTable provides a complex data model that supports dynamic control over data layout and format.
c) BigTable provides a simple data model that supports dynamic control over data layout and format.
d) BigTable provides a simple interface to a traditional relational database that supports dynamic control over data layout and format.

9) In the paper “BigTable: A Distributed Storage System for Structured Data,” by Chang *et al.*, published in OSDI 2006, a distributed storage system for managing structured data that scales to petabytes is presented. BigTable provides which of the following traditional RDBMS functionalities?

a) Atomic write operations.
b) Access control rights.
c) Atomic read operations.
d) All of the above. [Hint: a description for this answer would explain each of the answers a-d above.] e) None of the above. [Hint: a description for this answer would provide two examples, and explanations of what functionalities are provided by BigTable.]

10) In the paper “BigTable: A Distributed Storage System for Structured Data,” by Chang *et al.*, published in OSDI 2006, a distributed storage system for managing structured data that scales to petabytes is presented. BigTable utilizes a hierarchy analogous to that of a B^+ tree. How many levels are in this hierarchy? [Hint: explain each level in the hierarchy.]

a) 2
b) 3
c) 4
d) 5

11) Which of the following data is least suited to a traditional RDBMS database?

a) Book title.
b) Multiple authors.
c) Publication date.
d) An abstract spanning several paragraphs.

12) In MongoDB, a table (called a collection) may contain which of the following?

a) Complex data structures.
b) Arrays.
c) Embedded classes.
d) All of the above.
e) None of the above.

13) The paper “Dynamo: Amazon’s Highly Available Key-value Store,” by DeCandia *et al.*, published in SOSP 2007, presents a highly available key-value storage system used by Amazon.com. Scalability and reliability are the primary drivers for the design and implementation of Dynamo. To this end, hardware and network failures are treated as which of the following?

a) The normal case.
b) Rarely occurring special case.
c) Often occurring special case.
d) They are not explicitly considered in the design.

14) The paper “Dynamo: Amazon’s Highly Available Key-value Store,” by DeCandia *et al.*, published in SOSP 2007, presents a highly available key-value storage system used by Amazon.com. Dynamo provides only what kind of access to the data store?

a) Candidate key
b) Primary key
c) Superkey
d) Foreign key

15) The paper “Dynamo: Amazon’s Highly Available Key-value Store,” by DeCandia *et al.*, published in SOSP 2007, presents a highly available key-value storage system used by Amazon.com. Dynamo relaxes which database property to maintain high availability?

a) Isolation
b) Durability
c) Atomicity
d) Consistency

16) The paper “Dynamo: Amazon’s Highly Available Key-value Store,” by DeCandia *et al.*, published in SOSP 2007, presents a highly available key-value storage system used by Amazon.com. Dynamo achieves its performance by utilizing which of the following techniques?

a) Uniform data distribution through consistent hashing.
b) Trade off durability guarantees for performance.
c) Uniformly assigning requests to nodes.
d) All of the above.
e) None of the above.

17) Brewer’s CAP Conjecture states:

a) It is possible for a distributed database to achieve consistency, availability, and partition tolerance simultaneously.

b) It is possible for a distributed database to achieve at most one of consistency, availability, and partition tolerance simultaneously.

c) It is impossible for a distributed database to achieve consistency, availability, and partition tolerance, or any combination of the three, simultaneously.

d) It is impossible for a distributed database to achieve consistency, availability and partition tolerance simultaneously.

18) NoSQL databases are architecturally different from relational database because:

a) they represent relational data in a different manner.
b) they are designed to reap the read and write performance benefits of partition tolerance (horizontal scaling) while leaving either consistency or availability up for negotiation.
c) they are designed to reap the certainty benefits of consistency while leaving either partition tolerance or availability up for negotiation.
d) they are designed to reap the access benefits of availability while leaving either partition tolerance or consistency up for negotiation.

19) MongoDB:

a) doesn't provide ACID guarantees over a series of operations.
b) doesn't have the equivalent of an RDBMs' BEGIN, COMMIT
and ROLLBACK semantics. c) supports atomic, durable
updates on individual documents and consistent reads.
d) All of the above.
e) None of the above.

20) Which of the following statements is true of MongoDB operations?

a) if you have a multi-operation transaction that decrements from one property and increments another (on a single document), then MongoDB can ensure atomicity and durability in this case.
b) if you have a multi-operation transaction that updates properties on separate or multiple documents, then MongoDB cannot ensure atomicity or durability in this case.
c) if you have a multi-operation transaction on a single document, locking strategies are used to ensure strong consistency at the potential cost of high availability.
d) All of the above.
e) None of the above.

21) Which of the following types of databases is most suitable for banking transactions? [Hint: in your explanation describe an example scenario that illustrates why your chosen type works.]

a) Relational type
b) Document type
c) Graph type
d) Key-Value type

22) The Cassandra database operates best under what relative read-write occurrences?

a) Equal numbers of reads and writes.
b) More writes than reads.
c) More reads than writes.
d) All of the above.
e) None of the above.

23) MongoDB is designed to operate with data stored across multiple servers. The process of distributing different data across multiple servers is referred to as?

a) Sharding
b) Replication
c) Copying
d) Splitting

24) Which of the following companies developed Cassandra?

a) Twitter
b) Google
c) LinkedIn
d) Facebook

25) List five 1-2 word advantages of NoSQL databases (generally speaking) over relational databases, and list five 1-2 word disadvantages of NoSQL databases (generally speaking) compared to relational databases. Provide a one to two sentences explanation for each advantage and for each disadvantage.