**MSDS 7330 File Organization and Database Management**

**Final Exam Answer Sheet**

Name: Andrew Abbott

Date: December 4, 2016

1. E
   * A, complex data structures require many tables, not contained by a table.
   * B, RDBMS tables do not support arrays.
   * C, Embedded classes are a feature of object oriented databases.
2. D
   * A, Do not require schema and do not require joins.
   * B, Do not require joins or e-r model.
   * C, flexible design, OLAP, Key-Value, Column-Family, Graph, Document.
3. B
   * The Map inputs a key-value pair and outputs an intermediate key-value pair. The reduce takes that intermediate key-value pair and reduces it to a possibly smaller set.
   * Source: “MapReduce: Simplified Data Processing on Large Clusters,” Jeffrey Dean and Sanjay Ghemawat.
4. C
   * A, The MapReduce library in the program splits the input file into multiple partitions.
   * B, The master partition picks idle workers and assigns them map or reduce tasks.
   * C, A reduce worker reads all intermediate keys so that all occurrences of the same key are grouped together.
   * Source: “MapReduce: Simplified Data Processing on Large Clusters,” Jeffrey Dean and Sanjay Ghemawat.
5. A
   * B, the reduce results are appended to a final output file
   * C, The map intermediate results are combined before being input to the reduce workers.
   * Source: “MapReduce: Simplified Data Processing on Large Clusters,” Jeffrey Dean and Sanjay Ghemawat.
6. D
   * All answers come directly from the paper conclusions.
7. E
   * All of the answers come directly from the paper introduction.
   * ” Bigtable: A Distributed Storage System for Structured Data,” Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber
8. C
   * C is a quote from the paper introduction.
9. D
   * Reading, writing and sessions are atomic, using Chubby lock service.
10. B
    * BigTable uses a three-level hierarchy to store tablet location information.
    * From paper section 5.
11. B
    * I choose B because RDBMS is not designed for an unknown number of the same item.
12. D
    * May contain complex data structures, including reference, arrays and embedded documents.
    * Source: MongoDBOverview handout.
13. A
    * The papers refers to these failures as continuous.
    * “Dynamo: Amazon’s Highly Available Key-value Store,” Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels.
14. B
    * “Dynamo provides a simple primary-key only interface to meet the requirememts of these applications,” pg. 1 of paper.
15. D
    * “Dynamo sacrifices consistency under certain failure scenarios.” Paper pg. 1.
16. A
    * B, A tradeoff cannot guarantee performance.
    * C, Nodes added and removed without need for redistribution.
17. D
    * Satisfying all three at the same time is impossible.
18. B
    * Horizontal scaling is very important, basically available and eventually consistent. 11.3 lesson.
19. D
    * B, it doesn’t have SQL, uses other languages.
    * C, Supports atomic updates at the document level.
    * I am less sure of this one but I think both b and c are true so I choose all of the above.
20. A
    * Eventual consistency
21. A
    * A, provides ACID, important for many transactions.
    * B, good for variable data and web content.
    * C, good for social media, relationships
    * D, good for unstructured, banking is structured
22. D
    * “Cassandra can be described as fast and easily scalable with write operations spread across the cluster. The cluster does not have a master node, so any read and write can be handled by any node in the cluster.”
    * Source: NoSQLDatabasesOverview\_2.pdf handout.
23. A
    * Source: Wikipedia:Shard (database architecture)
24. D
    * Avinash Lakshman and Prashant Malik developed Cassandra at Facebook.
    * Source: wikipedia
25. * Advantages
      1. Horizontal scaling
         + Distributed horizontally rather than vertically. Even distributed on processors across the world.
      2. Variable data structures
         + So much unstructured data created continuously. That data is stored in databases. NoSQL databases are schema free and do not require joins.
      3. Data volume
         + Trillions of records captured daily. Sharding and commodity servers handle this cheaply. Supports linear scalability.
      4. Cloud computing
         + Servers in the sky. Do not require in-house servers.
      5. Agile development
         + Combination of above allow for iterative and continuous development and new and emerging apps.
    * Disadvantages
      1. Widespread use of RDBMS.
         + RDBMS is already everywhere. Databases already developed.
      2. Not necessarily ACID
         + Basically available, soft state, eventually consistent. RDBMS is better for high transaction uses.
      3. SQL designed for RDBMS
         + NoSQL has various methods to read/write/query.
      4. Maintenance
         + Harder to find NoSQL admins and experts.
      5. Analytics requirements
         + Common Business intelligence tools may not have connectivity to NoSQL.
    * Sources: Unit 11, http://www.techrepublic.com/blog/10-things/10-things-you-should-know-about-nosql-databases/