Benjamin Goodwin

DS7330 Final Exam

4/12/2021

1) D- All of the above

We can store complex data structures, arrays, and embedded classes all inside of a table.

2) C- flexibility in the database design.

Within the asynconous material, 7.1, you mention, “the main characteristic of thing is that it gives you flexibility in the data model”

3) B Key/Value pairs, Key/Value pairs.

* The paper states, “The computation takes a set of input key/value pairs and produces a set of output key/value pairs. The user of the MapReduce library expresses the computation as two functions: Map and Reduce.”

4) C- Each instance of the reduce function receives all of the intermediate data from every map function.

5) D-All of the above

* On page 3 of the paper on figure one we can see that the map intermediate files are stored on the local disks
* The reduce worker results are stored on local machines, e.g. “local disk on mappers”
* The map intermediate results are never combined or manipulated before being input to the reduce worker function. The combiner is an optimization and a local reducer.

6) D-All of the above

* The text states, “it allows programmers who have no experience with distributed and/or parallel systems to exploit large amounts of resources easily.”
* The text states, “MapReduce has been so successful because it makes it possible to write a simple program and run it efficiently on a thousand machines in the course of half an hour, greatly speeding up the development and prototyping cycle”
* Finally, the text gives several examples of problems that have been expressible as MapReduce problems such as large-scale machine learning problems, clustering problems for the Google News and Froogle products, extraction of data used to produce reports of popular queries, extraction of properties of web pages for new experiments and products
* Large-scale graph computations

7) E- All of the above.

* From Google’s own research, their stated objectives are: Wide applicability, scalability, high performance, and high availability.
* Wide applicability-Big table is used by more than 60 Google products and projects, including Google Analytics, Google Finance, Orkut, personalized search, Writely, and Google Earth.
* Scalability- The products mentioned above use Bigtable for a variety of demanding workloads, which range from throughput-oriented batch-processing jobs to latency-sensitive serving of data to end users.
* High availability-The Bigtable clusters used by these products span a wide range of configurations, from a handful to thousands of services, and store up to several hundred terabytes of data.
* High performance-Bigtable resembles a database and shares many implementation strategies with databases including their high performance.

8) C- Big table provides a simple data model that supports dynamic control over data layout and format.

* This is verbatim from Google’s own research paper.

9) D- All of the above

* Atomic write operations “Bigtable supports single-row transactions, which can be used to perform atomic read-modify-write sequences on data stored under a single row key”
* Access control lists “Bigtable uses Chubby to store Bigtable schema information”
* Atomic read operations “Bigtable supports single-row transactions, which can be used to perform atomic read-modify-write sequences on data stored under a single row key”

10) B-3

* Bigtable uses a three-level hierarchy analogous to that of a B+ tree to store tablet location information.
  + First level is a file stored in Chubby that contains the location of the root tablet
  + Second level is “other metadata tablets” and stores the location of a tablet under a row key that in an encoding of the tablet’s table identifier and its end row.
  + Third level exists as user tables

11) B-Multiple authors, second choice would be D

I first chose B, because multiple authors would require multiple selects and joins, and answer D can be remedied by using varchar(max)

12) D- all of the above

13) A-The normal case

* “As such Amazon’s software systems need to be constructed in a manner that treats failure handling as the normal case without impacting availability or performance.”

14) B- Primary key access

* “There are many services on Amazon’s platform that only need primary-key access to a data store”

15) D-Consistency

* “To achieve this level of availability, Dynamo sacrifices consistency under certain failure scenarios.”

16) D- All of the above

* “Client requests are uniformly assigned to nodes in the ring by a load balancer. Any Dynamo note can act as coordinator for a read request.”
* “Dynamo uses a synthesis of well known techniques to achieve scalability and availability: Data is partitioned and replicated using consistent hashing.”
* “The tradeoffs are in performance, cost efficiency, availability, and durability guarantees.”

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

18) 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.

19) E- None of the above, this one was tough, but it seems like none of this is behavior of MongoDB.

20) D- All of the above

21) C-Graph type

* The graph theory aspect is excellent for monitoring banking transactions and exploring nodes.
* Monitoring key relationships between nodes to obtain information
* I also think that banking information can be argued for almost any of these types
  + However, given the specialization of graph type and its ability for fraud monitoring (mentioned in the videos)

22) B- Cassandra works best under heavy write loads by nature.

23) A-Sharding

24) D-Facebook

* Apache Cassandra was developed by Facebook to power their inbox search feature.

Question 25 on next page

25)

Disadvantages to NoSQL

1. Less mature- RDBMS have been around a lot longer than NoSQL databases. There are many important features of NoSQL databases that have yet to be implemented.
2. Less support- Each NoSQL database in contrast tends to be open-source, with just one or two firms handling the support angle.
3. Business intelligence and analytics-Most features of NoSQL databases are aimed at web 2.0 applications, and typically these databases offer few features for analysis and query ad-hoc.
4. Administration-The end goal of NoSQL was to offer a solution that would require no administration, but the reality on the ground is different, NoSQL databases still demand a lot of technical skill.
5. No advanced expertise- Since NoSQL databases are still new, every NoSQL developer is still learning the ropes and organizations that want to implement NoSQL need to procced with caution.

Advantages to NoSQL

1. Elastic Scalability- RDBMS is not as easy to scale out on commodity clusters, but NoSQL databases are made for transparent expansion, taking advantage of new nodes.
2. Big data applications- Given that transaction rates are growing from recognition, there is need to store massive amounts of data. Volumes are easily handled by NoSQL.
3. Database administration- NoSQL databases require much less hands-on management, with data distribution and auto repair capabilities.
4. Economy- NoSQL databases can be easily installed in cheap commodity hardware clusters as transaction and data volumes increase.
5. Popularity- NoSQL databases are growing in popularity as new big implementations like Apache Cassandra, MongoDB, and others have emerged.