Summer school 2018 Valencia, Spain

Exercises I for Big data databases (28.06.2018)

Big data and Hadoop

General Instruction: There are one or multiple correct choices for each question.

1. Which of the Vs results in increased algorithmic complexity (which can cause analyses to not be able to finish running in reasonable amounts of time)?
   A. Velocity
   B. Volume
   C. Variety
   D. Veracity

2. Which of the Vs results in challenges due to graphs created from varying kinds, formats, sources, and meanings of data?
   A. Volume
   B. Variety
   C. Velocity
   D. Veracity

3. Updating a graph with a stream of posting information on Facebook is an example of which of the Vs?
   A. Velocity
   B. Volume
   C. Variety
   D. Veracity

4. The hoax, propaganda and fake news in Internet are examples of which of the Vs?
   A. Velocity
   B. Volume
   C. Variety
   D. Veracity

5. Following issues may be caused by lot of small files in HDFS:
A. NameNode memory usage increases significantly
B. Network load decreases
C. Number of map tasks need to process the same amount of data will be larger.
D. I/O rate will be faster

6. You are writing a 10GB file to a HDFS filesystem with a default block size of 128MB. How many blocks will the file be broken into?
A.1280
B.160
C.80
D.480

7. You are writing a 10GB file into HDFS with a replication of 2 and block size of 64MB. How much total disk space will this file use?
A.20GB
B.128GB
C.10GB
D.30GB

8. Name of the parameter that controls the replication factor in HDFS?
A.dfs.block.replication
B.dfs.replication.count
C.dfs.replication
D.replication.xml

9. Check answers that apply when replication is lowered.
A.HDFS is less robust
B.Less likely that data will be local to more workers
C. I/O rate will be worse
D.HDFS will have more space available

10. What are the main differences between Hadoop v1 and Hadoop v2?
A. Hadoop v1 has a single Master NameNode server, but Hadoop v2 may have multiple NameNode servers.
B. Hadoop v2 came up with new framework YARN (Yet another Resource Navigator), which provides ability to run Non-MapReduce application. But Hadoop v1 had no YARN framework.
C. Hadoop v1 treated all storage devices on a DataNode as a single uniform pool, but heterogeneous storage is part of Hadoop v2, where the system distinguishes between storage types and also makes the storage type information available to frameworks and applications.
D. Hadoop v1 has the Single-Point-of-Failure (SPOF) for both NameNode and DataNode, but Hadoop v2 has the feature to overcome SPOF with multiple Namenodes and DataNodes.

Summer school 2018 Valencia, Spain

Exercises II for Big data databases (28.06.2018)

NoSQL databases

1.What are the main differences between NoSQL and RDBMS databases?

A. The data format in RDBMS is organized and structured, but the data in NoSQL may be unstructured and unordered.
B. The data in RDBMS are stored in tables, but NoSQL can support different format of storage, such as key-value, XML document and graph.
C. RDBMS support ACID properties, but most of NoSQL databases cannot provide such transaction guarantee
D. RDBMS cannot support distributed or parallel processing, but NoSQL is very good for the scalability.

2. When should we select a NoSQL database instead of a relational database?
A. When we need to process a large scale of data, and those data have different formats.
B. When we need to provide the strong ACID transaction guarantee and the response speed is critical
C. When we need to perform the complicate joins and queries on data
D. When we need to process JSON documents and large graphs.

3. Select the correct statements on NoSQL databases.
A. The database management systems which are highly scalable and flexible are known as NoSQL databases.
B. NoSQL databases allow us to store and process unstructured and semi-structured data which is not possible when we make use of Relational database management system.
C. NoSQL gives an opportunity to the companies to store massive amount of structured and unstructured data in real time.
D. In today's time, big firms such as- Google, Facebook, Amazon, etc. use NoSQL for providing cloud-based services for storing data in real time.

4. How does NoSQL database tackle the challenge of Big Data?

A. Big data applications are generally looked from 4 perspectives: Volume, Velocity, Variety and Veracity. NoSQL databases can solve all problems for these four V's.
B. NoSQL is often used to store big data. NoSQL stores provide simpler scalability and improved performance relative to traditional RDMS.

C. NoSQL help big data moment in a big way by storing unstructured data and providing a means to query them as per requirements.
D. There are different kinds of NoSQL data stores, which are useful for different kinds of Big Data applications.

5. Select the correct statements on different database products:
A. Key-value store databases: Redis, Riak, DynamoDB, Memcache
B. Column family store databases: Cassandra, Big Table
C. Graph store databases: Neo4j, InfiniteGraph,HBase
D. Document store databases: MongoDB, CouchDB, Marklogic

6. What is the difference between Graph Database and RDBMS?

A. Graph database contain vertices and edges. Each vertex or node represent a key value or attribute. In RDBMS, attributes are appended in plain table format.
B. One of the biggest differences between graph databases and relational databases is that the connections between nodes in graph databases directly link in such a way that relating data becomes a simple matter of following connections. But relational databases have no such direct link.
C. This property, only found in graph databases, is known as index-free adjacency, and it allows queries to traverse millions of nodes per second, offering response times that are much faster than with relational databases.
D. Unlike relational databases, graph databases are harder to do summing queries and max queries efficiently.

7. What are the correct statements on operational databases and analytical databases:
A. Analytical database management systems is referred to as OLTP (On Line Transaction Processing) databases.
B. Analytical database management systems focuses on static data with complex analytical tasks.
C. Operational databases focuses in dynamic data with ACID transaction properties.
D. Operational databases are databases with an old style, which cannot tackle big data challenge.

8. What are the different kinds of NoSQL data stores?
A. Key-value stores
B. Column family stores
C. Graph stores
D. Document stores

9. What are the correct statements on Strong (Strict) Consistency and Eventual Consistency?
A. In the strong consistency, all replicas must be in the same state for the next operation to

occur on any value.
B. In the eventual consistency, all read operations always return the value from the last finalized write operation.
C. In the strong consistency, at any given point readers will see some written value, and there is the guarantee that any two readers will see the exact same write.
D. In the eventual consistency, all replicas will eventually have the latest update; it's just a matter of time when that will happen.

10. What are the correct explanations on CAP Theorem?
A. During normal operation (lack of network partition), the CAP theorem still imposes constraints on the availability or consistency of data.
B. "Consistency" in CAP Theorem has the same meaning as "Consistent" for ACID property of database transaction.
C. There are systems that are available and partition tolerant but cannot guarantee consistency.
D. CAP says that, in case of a network partition (a rare occurrence) one needs to choose between availability and partition tolerance.