CompSci 557

Kayla Goetzke

Phil Sauvey

Chapter 14

2. The 3 V’s of big data are volume, velocity, and variety. Volume is the quantity of the data to be stored. Velocity is the speed at which data is entering the system. Variety is the variations in the structure of the data to be stored.

3. Google and Amazon had to address the issue of big data after the dot com bubble burst and the companies that survived experienced significant growth as web commerce was consolidated into a smaller number of companies, and they experienced the pressure of big data.

4. Scaling up happens when you move to faster and larger system. While scaling out is when the workload exceeds the capacity, so you distribute the workload among a number of servers.

5. Stream processing is the process of analyzing the data stream as it enters the system. It decides what data to keep and what to discard before storing it.

6. Feedback loop processing is more focused on the outputs; whereas, stream processing is focused on the inputs.

9. The first key assumption made by Hadoop Distributed File System(HDFS) is that the files will be extremely large. Next, the write-once, read-many model is when a file is created, written and then closed. After it is closed, it cannot be changed. Advancements in HDFS all data to be appended to the end of the file. Another assumption is streaming access. Hadoop is optimized for batch processing of entire files as a continuous stream. The last assumption is fault tolerance. Fault tolerance is when there are multiple copies of the data on multiple devices, so if one fails you can find it on another device.

10. Name nodes store meta data for the file system. The metadata contains the name and block numbers for each file. Data nodes store blocks that contain actual file data within the HDFS.

11.The steps of the MapReduce process are as follows:

1. A Client Node submits a MapReduce job to the job tracker.
2. The job tracker communicates with the name node to determine which data nodes contain the blocks that should be processed for this job
3. The job tracker determines which task trackers are available for work.
4. The job tracker then contacts the task trackers on each of those nodes to begin mappers and reducers to complete that node's portion of the task.
5. The task tracker creates a new JVM to run the map and reduce functions.
6. The task tracker sends heartbeat messages to the job tracker to let the job tracker know the task tracker is still working on the job.
7. The job tracker monitors the heartbeat messages to determine if a task manager has failed. If so the job tracker can reassign that portion of the task to another node.
8. When the entire job is finished, the job tracker changes status to indicate that the job is completed.
9. The client node periodically queries the job tracker until the job status is completed.

12. The reason MapReduce and HDFS are complementary to each other is the implementation of MapReduce compliments the structure of the HDFS. Just as the HDFS structure is composed of a name node and several data nodes, MapReduce uses a job tracker and several task trackers.

13. The four basic categories of NoSQL are key-value database, document databases, column-oriented databases, and graph databases.

14. In a Key Value database the value can contain any type of data, however a document database always stores a document in the value component.

15. In row-centric storage, the data is stored such that a block contains all of the data for a small number of entries in the table. In column-centric storage, each block contains data for one attribute in the table.