CHAPTER 14: BIG DATA ANALYTICS AND NOSQL

1. Much ambiguity exists in defining Big Data.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.649

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. For a data set to be considered Big Data, it must display all the “3 Vs” – volume, velocity and variety.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.650

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. Scaling out is keeping the same number of systems, but migrating each system to a larger one.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.651

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. In many ways, the issues of associated with volume and velocity are the same.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Moderate REF: p.652

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Big Data

1. The analysis of data to produce actionable results is feedback loop processing.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.653

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. Relational databases rely on unstructured data.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Moderate REF: p.653

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Big Data

1. One tenet of Big Data is that all data that is capable of being captured should be.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Moderate REF: p.654

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Big Data

1. The ability to graphically data in a way that makes it understandable is the concept of value.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.654

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. Characteristics that are important in working with data in the relational database model also apply to Big Data.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Moderate REF: p.655

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Big Data

1. Hadoop is a database that has become the de facto standard for most Big Data storage and processing.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.655

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Under the HDFS system, using a write-one, ready-many model simplifies concurrency issues.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.656

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. A block report is used to let the name node know that the data mode is still available.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.657

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. A reduce function takes a collection of key-value pairs with the same key value and summarizes them into a single result.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.658

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Hive is a good choice for jobs that require a small subset of data to be returned very quickly.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.660

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Hadoop is a high-level tool that requires little effort to create, manage and use.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.660

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Flume is a tool for converting data back and forth between a relational database and the HDFS.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Easy REF: p.661

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Most NoSQL products run only in a Linux or Unix environment.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.662

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. Key-value and document databases are structurally similar.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.663-664

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. A column-family database is a NoSQL database model that organizes data in key-value pairs with keys mapped to a set of columns in the value component.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.666

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. Interest in graph databases can be tied to the area of social networks.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.668

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. Explanatory analytics uses predictive analytics as a stepping stone to create explanatory models.
   1. True
   2. False

*ANSWER:* False

PTS: 1 DIF: Difficulty: Moderate REF: p.670

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Data Analytics

1. Data mining focuses on the discovery and explanation stages of knowledge acquisition.
   1. True
   2. False

*ANSWER:* True

PTS: 1 DIF: Difficulty: Easy REF: p.671

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. \_\_is NOT one of the “3 Vs” of Big Data.
   1. Volume b. Velocity

c. Validation d. Variety

*ANSWER:* c

PTS: 1 DIF: Difficulty: Easy REF: p.649

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. \_\_\_\_ is keeping the same number of systems, but migrating each system to a larger system.
   1. Clustering b. Scaling up

c. Streaming d. Scaling out

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.651

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. \_\_ \_\_\_ focuses on filtering data as it enters the system to determine which data to keep and which to discard.
   1. Scaling up b. Feedback loop processing

c. Stream processing d. Scaling out

*ANSWER:* C

PTS: 1 DIF: Difficulty: Easy REF: p.652

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. A(n) \_\_ is a process or set of operations in a calculation.
   1. algorithm b. feedback loop

c. stream d. structure

*ANSWER:* a

PTS: 1 DIF: Difficulty: Easy REF: p.653

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. Big Data:
   1. relies on the use of structured data b. captures data in whatever format it naturally exists

c. relies on the use of unstructured data d. imposes a structure on data when it is captured

*ANSWER:* b

PTS: 1 DIF: Difficulty: Moderate REF: p.654

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Big Data

1. In the context of Big Data, \_\_\_\_\_ relates to differences in meaning.
   1. variety b. variability

c. veracity d. viability

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.654

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. In the context of Big Data, \_\_\_\_\_ refers to the trustworthiness of a set of data.
   1. value b. variability

c. veracity d. viability

*ANSWER:* c

PTS: 1 DIF: Difficulty: Easy REF: p.654

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. By default, Hadoop uses a replication factor of:
   1. one b. two

c. three d. four

*ANSWER:* c

PTS: 1 DIF: Difficulty: Easy REF: p.656

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Which of the following is NOT a key assumption of the Hadoop Distributed File System?
   1. High volume b. Write-many, read-once

c. Streaming access d. Fault-tolerance

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.655-656

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. When using a HDFS, the \_\_\_\_\_ node creates new files by communicating with the \_\_\_\_ node.
   1. client, name b. name, client

c. client, data d. data, client

*ANSWER:* a

PTS: 1 DIF: Difficulty: Easy REF: p.657

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. When using a HDFS, a heartbeat is sent every \_\_\_\_\_ to notify the name node that the data mode is still available.
   1. 3 hours b. 3 seconds

c. 6 hours d. 6 seconds

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.657

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. When using MapReduce, a \_\_\_\_\_\_\_ function takes a collection and data and sorts and filters it into a set of key-value pairs.
   1. reduce b. map

c. data d. block

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.658

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. When using MapReduce, best practices suggest that the number of mappers on a given node should be:
   1. 100 or more b. 100 or less

c. 50 or less d. at least 300

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.659

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. processing occurs when a program runs from beginning to end without any user interaction.
   1. Hadoop b. Block

c. Hive d. Batch

*ANSWER:* d

PTS: 1 DIF: Difficulty: Easy REF: p.660

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Two of the most popular applications to simplify the process of creating MapReduce jobs are Hive and
   1. Flume b. Pig

c. Sqoop d. Impala

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.660

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. \_\_\_ is a tool for converting data back and forth between a relational database and the HDFS.
   1. Flume b. Pig

c. Sqoop d. Impala

*ANSWER:* c

PTS: 1 DIF: Difficulty: Easy REF: p.661

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. \_\_\_ was the first SQL-on-Hadoop application.
   1. Flume b. Pig

c. Sqoop d. Impala

*ANSWER:* d

PTS: 1 DIF: Difficulty: Easy REF: p.662

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Which of the following is NOT one of the standard NoSQL categories?
   1. document databases b. column-oriented databases

c. graph databases d. chart databases

*ANSWER:* d

PTS: 1 DIF: Difficulty: Easy REF: p.662

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. To query the value component of the pair when using a key-value database, use get or:
   1. store b. fetch

c. retrieve d. gather

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.663

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. Document databases group documents into logical groups called:
   1. buckets b. sets

c. collections d. blocks

*ANSWER:* c

PTS: 1 DIF: Difficulty: Easy REF: p.664

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. \_\_\_\_\_\_minimizes the number of disk reads necessary to retrieve a row of data.
   1. Column-oriented database b. Row-centric storage

c. Column-family database d. Column-centric storage

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.665

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. Modeling and storing data about relationships is the focus of:
   1. key-value databases b. column-oriented databases

c. document databases d. graph databases

*ANSWER:* d

PTS: 1 DIF: Difficulty: Easy REF: p.668

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. uses statistical analysis to answer questions about the how and why of relationships.
   1. Explanatory analytics b. Data mining

c. Predictive analytics d. Knowledge acquisition

*ANSWER:* a

PTS: 1 DIF: Difficulty: Easy REF: p.670

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. uses statistical tools to answer questions about future data occurrences.
   1. Explanatory analytics b. Data mining

c. Predictive analytics d. Knowledge acquisition

*ANSWER:* c

PTS: 1 DIF: Difficulty: Easy REF: p.670

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. The goal of the \_\_\_\_\_ phase of data mining is to identify common data characteristics or patterns.
   1. data preparation b. data analysis and classification

c. knowledge acquisition d. prognosis

*ANSWER:* b

PTS: 1 DIF: Difficulty: Easy REF: p.672

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. The end user decides what techniques to apply to the data when using the \_\_\_\_\_ mode of data mining
   1. guided b. prognosis

c. directed d. automated

*ANSWER:* a

PTS: 1 DIF: Difficulty: Easy REF: p.673

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. Most BI vendors are dropping the term “data mining” and replacing it with the term:
   1. explanatory analytics b. data analytics

c. predictive analytics d. knowledge acquisition

*ANSWER:* c

PTS: 1 DIF: Difficulty: Easy REF: p.674

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. is the Big Data “3 V” that relates to the speed at which data is entering the system.

*ANSWER:* Velocity

PTS: 1 DIF: Difficulty: Easy REF: p.649

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. Scaling out is also referred to as .

*ANSWER:* clustering

PTS: 1 DIF: Difficulty: Moderate REF: p.649

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Big Data

1. \_\_\_\_\_\_ refers to the analysis of the data to produce actionable results.

*ANSWER:* Feedback loop processing

PTS: 1 DIF: Difficulty: Easy REF: p.653

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. A method of text analysis that attempts to determine if a statement conveys a positive, negative, or neutral attitude is referred to as \_\_\_analysis.

*ANSWER:* sentimental

PTS: 1 DIF: Difficulty: Easy REF: p.654

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. isthe coexistence of a variety of data storage and data management technologies within an organization’s infrastructure.

*ANSWER:* Polyglot persistence

PTS: 1 DIF: Difficulty: Easy REF: p.655

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Big Data

1. Within MapReduce, a runs maps and reduces functions.

*ANSWER:* task tracker

PTS: 1 DIF: Difficulty: Easy REF: p.659

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Most organizations that use Hadoop also use a set of other related products that interact and complement each other to produce an entire \_\_\_\_\_\_ of applications and tools.

*ANSWER:* ecosystem

PTS: 1 DIF: Difficulty: Easy REF: p.660

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. languages allow the user to specify what they want, not how to get it which is very useful for query processing.

*ANSWER:* Declarative

PTS: 1 DIF: Difficulty: Easy REF: p.661

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Within Hadoop, is used for producing data pipeline tasks that transform data in a series of steps.

*ANSWER:* Pig

PTS: 1 DIF: Difficulty: Easy REF: p.661

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. Within Hadoop, can transfer data in both directions - into and out of HDFS.

*ANSWER:* Sqoop

PTS: 1 DIF: Difficulty: Easy REF: p.661

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Hadoop

1. databases simply store data with no attempt to understand the contents of the value component or its meaning.

*ANSWER:* Key-value KV

PTS: 1 DIF: Difficulty: Easy REF: p.663

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. is a human-readable text format for data interchange that defines attributes and values in a document.

*ANSWER:* JavaScript Object Notation

JSON

PTS: 1 DIF: Difficulty: Easy REF: p.664

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. do not store relationships as perceived in the relational model and generally have no support for join operations.

*ANSWER:* Document databases

PTS: 1 DIF: Difficulty: Easy REF: p.665

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. refers to traditional, relational database technologies that use column-centric, not row-centric storage.

*ANSWER:* Column-oriented database

Columnar database

PTS: 1 DIF: Difficulty: Easy REF: p.665

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. In a column family database, a column that is composed of a group of other related columns is called a(n) \_\_\_\_\_\_.

*ANSWER:* super column

PTS: 1 DIF: Difficulty: Easy REF: p.667

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. In a graph database, the representation of a relationship between nodes is called a(n)\_\_\_\_\_.

*ANSWER:* edge

PTS: 1 DIF: Difficulty: Easy REF: p.668

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. A query in a graph database is called a(n)\_\_\_\_\_.

*ANSWER:* traversal

PTS: 1 DIF: Difficulty: Easy REF: p.668

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. A database model that attempts to provide ACID-compliant transactions across a highly distributed infrastructure is \_\_\_\_\_.

*ANSWER:* NewSQL

PTS: 1 DIF: Difficulty: Easy REF: p.669

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: NoSQL

1. is a continuous spectrum of knowledge acquisition that goes from discovery to explanation to prediction..

*ANSWER:* Data analytics

PTS: 1 DIF: Difficulty: Easy REF: p.670

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. In the \_\_\_\_ phase of data mining, findings are used to predict future behavior and forecast business outcomes.

*ANSWER:* prognosis

PTS: 1 DIF: Difficulty: Easy REF: p.672

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. The origins of \_\_\_\_ can be traced back to the banking and credit card industries.

*ANSWER:* predictive analytics

PTS: 1 DIF: Difficulty: Easy REF: p.674

NAT: BUSPROG: Technology STATE: DISC: Information Technologies

KEY: Bloom's: Knowledge TOP: Data Analytics

1. Discuss the “3 Vs” of Big Data. How has the definition of Big Data regarding these items changed over time?

*ANSWER:* The three V’s are Volume, Velocity and Variety

Volume is the quantity of data to be stored and a key characteristic of Big Data. The storage capacities associated with Big Data are very large. As storage needs increase, they can be handled by scaling up or scaling out. Scaling up is keeping the same number of systems but migrating each to a larger system. Scaling out involves distributing data storage structures across a cluster of commodity servers.

Velocity is the speed at which data enters the system and is another key characteristic. In many ways, the issues of velocity mirror those of volume. The velocity of processing can be broken down into two categories: stream and feedback loop.

Variety refers to the vast array of formats and structures in which the data may be captured. Big Data requires that the data be captured in whatever format it naturally exists.

The lack of specific values associated with these characteristics is what leads to ambiguity in defining Big Data. What is considered Big Data changes over time, but the key is the characteristics are present to an extent that the current relational database technology struggles with managing the data.

There is also some disagreement about which of the 3 Vs must be present for a data set to be considered Big Data. Originally it was conceived as a combination of the 3 Vs. Recent changes in technology have led to Big Data being redefined as involving any, but not necessarily all of the 3 Vs.

PTS: 1 DIF: Difficulty: Moderate REF: p.649-654

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Big Data

1. Define the four key assumptions of the Hadoop Distributed File System (HDFS).

*ANSWER: High volume*: The volume of data in Big Data applications is expected to be in terabytes, petabytes or larger. Hadoop assumes HDFS files will be extremely large

*Write-once, ready-many*: This model simplifies concurrent issues and improves overall data throughput. Using this model, a file is created, written to the file system and then closed. Once the file is closed, changes cannot be made to its contents which improves overall system performance and works well for the types of tasks performed by many Big Data applications.

*Streaming access*: Unlike transaction processing systems, Big Data applications typically process entire files. Hadoop is optimized for batch processing of entire files as continuous streams of data.

*Fault tolerance*: Hadoop is designed to be distributed across thousands of low-cost, commodity computers. The HDFS is designed to replicate data across many devices so that, when one fails, the data is still available from another device. By default, Hadoop uses a replication factor of three, meaning that each block of data is stored on three devices.

PTS: 1 DIF: Difficulty: Moderate REF: p.655-656

NAT: BUSPROG: Analytic STATE: DISC: Information Technologies

KEY: Bloom's: Comprehension TOP: Hadoop

1. Discuss the need for a Hadoop ecosystem and identify the key components.

*ANSWER:* Because Hadoop is a very low-level tool requiring considerable effort to create, manage, and use,

it presents quite a few obstacles. This has resulted in a host of related applications that attempt to make Hadoop easier to use and more accessible to users who are not skilled at complex Java programming. Most organizations that use Hadoop also use a set of other related products that interact and complement each other to produce an entire ecosystem of applications and tools.

MapReduce simplification applications have been developed to simplify the process of creating MapReduce jobs. Two of the most popular are Hive and Pig.

Data ingestion applications help to “ingest” or gather data into Hadoop from existing systems and include Flume. Sqoop is a tool for converting data back and forth between a relational database and HDFS.

Direct query applications attempt to provide faster query access than is possible through MapReduce and include HBase and Impala.

PTS: 1 DIF: Difficulty: Moderate REF: p.660-662

NAT: BUSPROG: Analytic STATE: DISC: Information Technology

KEY: Bloom's Comprehension TOP: Hadoop

1. What is NoSQL and what are the major NoSQL approaches (categories)?

*ANSWER:* NoSQL is the unfortunate name given to a broad array of nonrelational database technologies that have developed to address Big Data challenges. The name is unfortunate because it does not describe what the NoSQL technologies are, but rather what they are not. Even that explanation is poor. Literally hundreds of products can be considered as NoSQL. Most of them fit into one of four categories: key-value data stores, document databases, column-oriented databases and graph databases.

PTS: 1 DIF: Difficulty: Moderate REF: p.

NAT: BUSPROG: Analytic STATE: DISC: Information Technology

KEY: Bloom's Comprehension TOP: NoSQL

1. Discuss NewSQL and what does it attempts to do.

*ANSWER:* NewSQL is a database model that attempts to provide ACID-compliant transactions across a highly distributed infrastructure and are the latest technologies to appear to appear in the data management arena to address Big Data problems. As a new category of data management products, NewSQL databases have not yet developed a track record of success and have been adopted by relatively few organizations.

Because no technology can perfectly provide the advantages of both RDBMS and NoSQL, NewSQL has disadvantages, principally centered around its heavy use of in-memory storage.

PTS: 1 DIF: Difficulty: Moderate REF: p.669-670

NAT: BUSPROG: Analytic STATE: DISC: Information Technology

KEY: Bloom's Comprehension TOP: NoSQL

1. Explain the concept of data analytics. What are the various tools of data analytics?

*ANSWER:* Data analytics is a subset of business intelligence (BI) functionality that encompasses a wide range ofmathematical, statistical, and modeling techniques with the purpose of extracting knowledge from data.Data analytics is used at all levels within the BI framework, including queries and reporting, monitoringand alerting, and data visualization. Hence, data analytics is a “shared” service that is crucial to what BIadds to an organization. Data analytics represents what business managers really want from BI: theability to extract actionable business insight from current events and foresee future problems oropportunities. Data analytics tools can be grouped into two separate (but closely related and oftenoverlapping) areas:

* Explanatory analytics focuses on discovering and explaining data characteristics and relationships based on existing data. Explanatory analytics uses statistical tools to formulate hypotheses, test them, andanswer the how and why of such relationships.
* Predictive analytics focuses on predicting future data outcomes with a high degree of accuracy.Predictive analytics uses sophisticated statistical tools to help the end user create advanced models thatanswer questions about future data occurrences.

PTS: 1 DIF: Difficulty: Moderate REF: p.669-670

NAT: BUSPROG: Analytic STATE: DISC: Information Technology

KEY: Bloom's Comprehension TOP: Data Analytics