CCC Simulation Exam

Q1. Which of the following is not a term for measuring data quantity?

a) Zettabyte

b) Kilobyte

c) **Alphabyte**

d) Gigabyte

Q2. Name the author of one of the early articles on Big Data discussing volume, velocity, and variety.

a) Doug Laggett

b) **Doug Laney**

c) Doug Langley

d) Derek Laney

Q3. Who should be involved in a Big Data project?

a) **C-Suite members, Domain Experts, Scary Data People, IT professionals**

b) C-Suite members, Domain Experts, Project Managers, Data Administrators

c) Scary Data People, IT professionals, Project Managers, Data Administrator

d) Scary Data People, IT professionals, Project Managers, Domain Experts

Q4. According to McKinsey Global Institute (MGI), what is the potential benefit of Big Data for US Health Care? Big Data?

a) $300 million

b) **$300 billion**

c) $500 million

d) $150 million

Q5. What is the decrease in piracy, according to the International Maritime Bureau, in the first 6 months of 2012, as a result of using Big Data?

a) 12.5%

b) 21%

c) **54%**

d) 80%

CCC Simulation Exam

Q6. What percentage of people in the US can be identified using DOB, ZIP, and sex?

a) 7.6%

b) **85%**

c) 16%

d) 45%

Q7. Which company uses MongoDB as the basis of its content management system?

a) BBC

b) **Forbes**

c) Time Warner

d) Associated Press

Q8. eBay is running a Hadoop cluster consisting of how many nodes?

a) **532** (This was back in 2017)

b) 121

c) 79

d) 812

Q9. The three main characteristics of Big Data are:

a) Validity, velocity, veracity.

b) Volume, veracity, variety.

c) Volume, validity, venality.

d) **Velocity, variety, volume.**

Q10. Some of the main players in the Enterprise Systems space are:

a) PeopleSoft, JD Edwards, Quora.

b) **SAP, Oracle, Microsoft.**

c) Facebook, Instagram, Twitter.

d) MS Access, MySQL, DB2.

CCC Simulation Exam

Q11. In the case of SAP, data is best accessed through:

a) **A connector.**

b) A Flat file.

c) MS Excel.

d) MySQL.

Q12. Data warehouses, as a rule, are based on:

a) Hierarchical database technology.

b) **Relational database technology.**

c) Spreadsheet files.

d) Network database technology.

Q13. When working with metadata, we are interested in:

a) Message content.

b) **Information, such as, date and time, author, and origin.**

c) The time when a database table was created.

d) Recipient's personal details.

Q14. The Public Feed API of Facebook is available to:

a) The general public.

b) Some universities in US and Canada.

c) A select number of Silicon Valley start-ups.

d) **A limited set of media publishers.**

Q15. The Facebook Graph consists of:

a) **Nodes, edges, fields.**

b) Nodes, fields, topologies.

c) Nodes, images, audio.

d) Video, audio, nodes.

CCC Simulation Exam

Q16. Twitter's streaming APIs provide access to public tweets with very low delay. The stream is limited to:

a) 9.5% of all tweets.

b) 5.6% tweets of a select group of users.

c) **1% of all tweets.**

d) 1% of a user's tweets for a specific period of the day.

Q17. Twitter provides the following streaming APIs:

a) **Public streams, user streams, site streams** (User & Site streams are no longer available)

b) Public streams, data streams, user streams

c) Facebook streams, Google streams, user streams

d) Daily streams, monthly streams, weekly streams

Q18. The main providers of commodities-related datasets are:

a) World Gold Council, World Bank, New York Stock Exchange.

b) **World Gold Council, International Monetary Fund, World Bank.**

c) International Monetary Fund, Reserve Bank of China, World Bank.

d) Bank of America, BNP Paribas, Banco Itaú.

Q19. Data mining is the process of:

a) Updating data elements in large datasets.

b) Extracting data from datasets.

c) **Discovering patterns within large datasets.**

d) Maintaining large datasets by removing and adding records.

Q20. Which of the following is not a type of data mining?

a) Classification

b) Clustering

c) **Cubing**

d) Association

CCC Simulation Exam

Q21. The main algorithm used for clustering is:

a) **K-means.**

b) L-means.

c) K-averages.

d) L-averages.

Q22. Weka is a popular data mining application developed at:

a) **The University of Waikato.**

b) The University of Konstanz.

c) The University of Chicago.

d) The University of Queensland.

Q23. Name any two of the main Hadoop modules.

a) RDBMS and NoSQL

b) **HDFS and YARN**

c) MapReduce and NoSQL

d) Pig and SQL

Q24. The failover mechanism of Hadoop is implemented in:

a) Data centers.

b) Hardware.

c) **Software.**

d) Servers.

Q25. Which of the following is true?

a) **DataNode periodically sends a HeartBeat to a NameNode.**

b) DataNode is a server managing access to files and the namespace.

c) NameNode manages the storage on a node.

d) NameNode periodically sends a HeartBeat to a DataNode.

CCC Simulation Exam

Q26. Hbase is:

a) **A high-performance database sitting on top of HDFS.**

b) Based on Oracle’s BigTable model.

c) A data warehouse software that allows accessing of large datasets.

d) Very similar to an RDBMS and applications can be easily ported.

Q27. Which of the following is not a type of Hadoop installation?

a) Local

b) **Hybrid**

c) Pseudo-distributed

d) Fully distributed

Q28. MapReduce:

a) Is used to map relational tables to Hadoop.

b) Takes several datasets and combines them for further processing.

c) Uses similar syntax to SQL.

d) **Takes a big dataset and splits it across several nodes.**

Q29. YARN resource management component works on a:

a) **Master-slave model.**

b) Master-master model.

c) Slave-slave model.

d) Peer-to-peer model.

Q30. The output of both Map and Reduce phases is:

a) A single key-value pair.

b) A CSV file.

c) **Set of key-value pairs.**

d) A database table.

CCC Simulation Exam

Q31. Which of the following is not provided by Hive?

a) Extract/Transform/Load (ETL) functionality

b) **MQL functionality**

c) Query Language (QL) functionality

d) Access to HDFS files

Q32. What type of database is MongoDB?

a) Open-source relational database

b) Relational database

c) **Document database**

d) MapReduce database

Q33. Explain what sharding means in MongoDB context.

a) The process in which the same data set is maintained in several locations.

b) **The process of storing data records across multiple nodes.**

c) The process in which MapReduce also allows large data sets to be condensed into aggregated results.

d) The process in which MongoDB aggregates the initial data set into a summarized one.

Q34. MongoDB is not available for:

a) Mac OS.

b) Linux.

c) **Chrome OS.**

d) Solaris.

Q35. The schema in document databases is:

a) Fixed.

b) **Flexible.**

c) Denormalized.

d) Bi-directional.

CCC Simulation Exam

Q36. In a document database, a collection corresponds to:

a) A record in a relational database.

b) A field in a relational database.

c) Key-value pair in Hadoop.

d) **A table in a relational database.**

Q37. In MongoDB the maximum size of a document is:

a) 8 MB.

b) **16 MB.**

c) 1 GB.

d) 128 MB.

Q38. Names in MongoDB cannot start with:

a) A number.

b) % character.

c) **$ character.**

d) # character.

Q39. Names in MongoDB cannot contain:

a) .

b) _

c) -

d) +

Q40. Embedded documents in MongoDB design should be used when:

a) There are no relationships between entities.

b) We need to model hierarchical sets.

c) **There are one-to-many relationships between entities.**

d) There are many-to-many relationships between entities.