Big Data Analytics - All Questions with Correct Answers Highlighted

Module 1: Introduction to Big Data Analytics 1. Which of the following correctly describes the 4Vs of Big Data?

A. Volume, Velocity, Variety, Veracity

B. Value, Velocity, Variety, Validity

C. Volume, Variability, Visualization, Velocity

D. Veracity, Volume, Version, Velocity

2. In data preprocessing, normalization is primarily used to:

A. Remove missing values

B. Standardize data ranges

C. Detect outliers

D. Reduce redundancy

3. Which architecture pattern is most suitable for Big Data scalability?

A. Monolithic

B. Client-server

C. Layered

D. Distributed

4. Which statement about Big Data Analytics applications is true?

A. Real-time only

B. Offline only

C. Combines data engineering and ML

D. Requires mainframe

5. In designing data architecture, scalability can best be achieved through:

A. Vertical scaling

B. Horizontal scaling

C. Dynamic caching

D. Thread-based design

6. Which of the following represents unstructured data?

A. CSV file

B. SQL table

C. Image file

D. Excel sheet

7. The term 'veracity' in Big Data refers to:

A. Speed of data

B. Trustworthiness of data

C. Volume of data

D. Type of data

8. Which is an example of data preprocessing?

A. Model training

B. Data cleaning

C. Visualization

D. Prediction

9. Which of the following ensures scalability in Big Data systems?

A. Centralized storage

B. Vertical scaling

C. Distributed processing

D. Sequential execution

10. The best example of semi-structured data is:

A. JSON file

B. Video clip

C. Image folder

D. SQL table

Module 2: Hadoop and its Ecosystem 1. Hadoop's design principle primarily supports:

A. Centralized computing

B. Vertical scalability

C. Distributed data processing

D. Memory-based computation

2. The NameNode in HDFS is responsible for:

A. Storing data blocks

B. Managing metadata

C. Handling replication

D. Both A and C

3. Hadoop YARN acts as:

A. Query engine

B. Resource manager

C. Database manager

D. Data storage tool

4. Sqoop is primarily used for:

A. Import/export between RDBMS and HDFS

B. Statistical analysis

C. Workflow scheduling

D. Managing nodes

5. Flume is designed for:

A. Data ingestion from multiple sources

B. Data analysis

C. Machine learning

D. Workflow automation

6. HDFS is designed for:

A. High latency access

B. Real-time processing

C. Streaming large files

D. Transaction processing

7. YARN in Hadoop stands for:

A. Yet Another Resource Negotiator

B. Yielding Accurate Resource Notation

C. Your Adaptive Runtime Network

D. Yearly Allocation Resource Node

8. Flume is primarily used for:

A. Data ingestion

B. Query execution

C. Task scheduling

D. File replication

9. The default HDFS block size is:

A. 32 MB

B. 64 MB

C. 128 MB

D. 256 MB

10. HDFS achieves fault tolerance through:

A. Data encryption

B. Data compression

C. Data replication

D. Backup scheduling

Module 3: NoSQL, MongoDB, and Cassandra 1. NoSQL systems primarily differ from SQL databases in their:

A. Transaction support

B. Schema flexibility

C. Query language

D. Use of indexes

2. The shared-nothing architecture in NoSQL ensures:

A. Data redundancy

B. Independent node operation

C. Single-point dependency

D. Centralized metadata

3. Which type of NoSQL database does MongoDB represent?

A. Column-family

B. Key-value

C. Document-oriented

D. Graph

4. Cassandra is optimized for:

A. Read-heavy workloads

B. Write-heavy workloads

C. Graph traversal

D. Transaction consistency

5. NoSQL architecture achieves horizontal scalability by:

A. Increasing CPU cores

B. Adding more nodes

C. Increasing threads

D. Using a single server

6. Which data format is used by MongoDB?

A. XML

B. BSON

C. CSV

D. YAML

7. Cassandra follows which architecture?

A. Master-slave

B. Peer-to-peer

C. Centralized

D. Hierarchical

8. In MongoDB, a collection is equivalent to:

A. Table

B. Column

C. Row

D. Record

9. The CAP theorem stands for:

A. Consistency, Availability, Partition tolerance

B. Capacity, Accuracy, Partitioning

C. Connectivity, Access, Performance

D. Communication, Allocation, Partition

10. Cassandra stores data as:

A. Documents

B. Key-value pairs

C. Graphs

D. Tables

Module 4: MapReduce, Hive, and Pig 1. The “map” phase in MapReduce is responsible for:

A. Aggregating results

B. Filtering and sorting data

C. Generating key-value pairs

D. Shuffling results

2. Reducer receives data based on:

A. Value

B. Key

C. DataNode location

D. Replication factor

3. Hive is best suited for:

A. Real-time processing

B. Batch data warehousing

C. Stream analytics

D. Online transactions

4. Pig Latin scripts are translated into:

A. HiveQL

B. MapReduce jobs

C. Spark RDDs

D. Python scripts

5. Hive Metastore stores:

A. HDFS metadata

B. Schema info for Hive tables

C. Hadoop config

D. Query logs

6. MapReduce programming model consists of:

A. One function

B. Map and Reduce

C. Three functions

D. Mappers only

7. The purpose of the shuffle phase in MapReduce is:

A. Filtering

B. Sorting and grouping keys

C. Reducing output

D. Loading data

8. Combiner in MapReduce is used to:

A. Combine reducers

B. Optimize network traffic

C. Split input files

D. Map the data

9. Pig Latin is:

A. Declarative language

B. Procedural language

C. Object-oriented language

D. SQL scripting

10. The job output of MapReduce is stored in:

A. Hive table

B. RDBMS

C. HDFS output directory

D. Pig cache

Module 5: Machine Learning, Web and Social Network Analytics 1. Regression analysis is primarily used to:

A. Identify clusters

B. Predict continuous outcomes

C. Classify categorical values

D. Detect outliers

2. A high variance model in machine learning indicates:

A. Underfitting

B. Overfitting

C. Regularization

D. Bias reduction

3. Which algorithm is most suitable for “finding similar items”?

A. Apriori

B. k-NN

C. PageRank

D. Naïve Bayes

4. Which measure best captures set similarity?

A. Pearson correlation

B. Jaccard coefficient

C. Euclidean distance

D. Manhattan distance

5. PageRank algorithm was originally designed for:

A. Social networks

B. Search engine optimization

C. Web classification

D. Link prediction

6. Outliers in data can significantly affect:

A. Regression models

B. Clustering accuracy

C. Both A and B

D. None

7. Association rule mining focuses on:

A. Classification

B. Regression

C. Discovering relationships

D. Clustering

8. Collaborative filtering is used in:

A. Fraud detection

B. Recommendation systems

C. Sentiment analysis

D. Classification

9. In text mining, tokenization means:

A. Grouping text

B. Splitting text into words

C. Counting characters

D. Encoding words

10. Social networks are often represented as:

A. Tables

B. Graphs

C. Matrices

D. Lists