Big Data - Chapter 1 MCQ Practice

50 multiple-choice questions based on Chapter 1: Overview of Big Data

Instructions

Choose the single best answer for each question. Each question has four choices: A, B, C, and D. The answer key is provided at the end of the document.

Core Concepts and 5Vs

1. Big Data refers to data that is mainly difficult to process because it is:

A. Only numerical and stored in tables
B. Small, simple, and slow-moving
C. Large, fast, and complex
D. Always manually collected

2. Which of the following is not one of the 5Vs of Big Data?

A. Volume
B. Variety
C. Verification
D. Velocity

3. The Volume characteristic of Big Data refers to:

A. The speed at which data is generated
B. The size or amount of data
C. The trustworthiness of data
D. The business usefulness of data

4. The Variety characteristic means that Big Data includes:

A. Only structured data
B. Only relational database records
C. Different formats such as structured, semi-structured, and unstructured data
D. Only real-time sensor data

5. Which of the following is an example of structured data?

A. A video file
B. A tweet
C. A table in a SQL database
D. An audio recording

6. Which data type includes formats such as HTML, XML, and JSON?

A. Structured data
B. Semi-structured data
C. Unstructured data
D. Encrypted data

7. Which of the following is an example of unstructured data?

A. MySQL table
B. Oracle database record
C. Microsoft SQL Server table
D. Video stream

8. The Velocity characteristic of Big Data is mainly concerned with:

A. How useful the data is
B. How fast data is generated and processed
C. How clean the data is
D. How many formats the data has

9. Which Big Data characteristic deals with bias, noise, uncertainty, and reliability?

A. Volume
B. Variety
C. Veracity
D. Value

10. In Big Data, Value refers to:

A. The ability to extract meaningful insights and business benefit
B. The number of files stored
C. The physical cost of storage devices
D. The speed of internet access

History and Drivers

11. According to the chapter, the term Big Data emerged in the:

A. 1960s
B. 1970s
C. 1990s
D. 2020s

12. Google introduced MapReduce in:

A. 1990
B. 2004
C. 2006
D. 2010

13. Hadoop was officially unveiled in:

A. 1995
B. 2000
C. 2006
D. 2015

14. MapReduce helped enable:

A. Manual data entry
B. Large-scale data processing across distributed systems
C. Small spreadsheet calculations
D. Traditional centralized database processing only

15. Which technology was strongly influenced by the principles of MapReduce?

A. Hadoop
B. Excel
C. PowerPoint
D. Microsoft Word

16. Which of the following is listed as a key driver of Big Data?

A. Decrease in digital devices
B. Growth of IoT and sensors
C. Reduction in internet usage
D. Elimination of cloud computing

Traditional Data vs. Big Data

17. Traditional data approaches mainly focus on:

A. Massive, fast-moving, diverse datasets
B. Distributed storage and flexible data models
C. Structured, moderate-size datasets using centralized relational databases
D. Unstructured data only

18. Big Data approaches usually use:

A. Distributed storage and parallel processing
B. Only one centralized server
C. Only fixed schemas
D. Only manual analysis

19. In traditional data systems, scalability is usually:

A. Horizontal, or scale out
B. Vertical, or scale up
C. Impossible
D. Based only on cloud storage

20. In Big Data systems, scalability is usually:

A. Vertical only
B. Horizontal, or scale out
C. Not supported
D. Limited to one database

21. Traditional data systems usually use which query model?

A. SQL
B. NoSQL only
C. MapReduce only
D. Spark only

22. Big Data systems may use:

A. SQL, NoSQL, MapReduce, and Spark
B. Only SQL
C. Only spreadsheets
D. Only paper-based records

Architecture, Ingestion, Storage, and Governance

23. A Big Data architecture is designed to handle:

A. Only small business reports
B. Data ingestion, processing, and analysis
C. Only data deletion
D. Only user interface design

24. Data ingestion is the process of:

A. Deleting old records
B. Importing or loading data into the architecture
C. Visualizing data in charts only
D. Encrypting data for storage only

25. Batch ingestion usually means data is imported:

A. Immediately as it is generated
B. On a scheduled or interval-based basis
C. Only through sensors
D. Without any processing

26. Real-time ingestion is especially important for:

A. Time-sensitive applications such as fraud detection
B. Old archived documents only
C. Weekly manual reports only
D. Static spreadsheets only

27. Which tool is mentioned as popular for real-time data ingestion?

A. Apache Kafka
B. Microsoft Excel
C. PowerPoint
D. Oracle Forms

28. Distributed file systems such as HDFS or Amazon S3 help by:

A. Keeping all data on one small machine
B. Spreading data across multiple machines
C. Preventing any form of parallel processing
D. Removing the need for storage

29. Which process removes or corrects inaccurate records and inconsistencies?

A. Data cleaning
B. Data visualization
C. Data deletion
D. Data ownership

30. Data enrichment means:

A. Reducing all data into one column
B. Adding extra information or context to increase value
C. Removing all unstructured data
D. Encrypting data permanently

31. Analytical sandboxes are used for:

A. Isolated data exploration, machine learning, and predictive modeling
B. Deleting enterprise data warehouses
C. Replacing all security systems
D. Preventing analysis

32. The main goal of most Big Data solutions is to:

A. Store data without using it
B. Provide insights through analysis and reporting
C. Replace all business employees
D. Convert all data into images

33. Big Data governance is concerned with:

A. Rules, policies, data access, quality, and responsible use
B. Only increasing data volume
C. Only designing dashboards
D. Removing security protocols

34. Which tool is mentioned as an example for adding a governance layer?

A. Apache Atlas
B. Apache Spark
C. Apache Kafka
D. MongoDB

35. Apache Hadoop includes which major components?

A. HDFS and MapReduce
B. Power BI and Excel
C. HTML and CSS
D. MongoDB and Cassandra only

Tools and Use Cases

36. Apache Spark is known for:

A. In-memory processing
B. Manual data entry
C. Relational tables only
D. Removing the need for analytics

37. Compared to Hadoop MapReduce, Spark can accelerate processing because it supports:

A. Paper-based storage
B. In-memory processing
C. Manual indexing
D. Fixed-schema reporting only

38. NoSQL databases are designed mainly for handling:

A. Only structured financial tables
B. Unstructured and semi-structured data
C. Printed documents only
D. Small relational databases only

39. Which of the following is an example of a NoSQL database mentioned in the chapter?

A. MongoDB
B. Microsoft Word
C. Power BI
D. Excel

40. Which use case involves personalized movie and TV recommendations?

A. Flight companies
B. Streaming apps such as Netflix
C. Smart cities
D. Scientific research

Tricky Review Questions

41. Which option best distinguishes Volume from Velocity?

A. Volume is about data quality; Velocity is about data usefulness
B. Volume is about data size; Velocity is about speed of generation and processing
C. Volume is about structured data; Velocity is about unstructured data
D. Volume is about real-time analytics; Velocity is about storage cost

42. A company stores customer transactions in relational tables, analyzes them weekly, and uses a fixed schema before inserting data. This is closest to:

A. Big Data approach with schema-on-read
B. Traditional data approach with schema-on-write
C. Real-time streaming architecture
D. NoSQL-based horizontal scaling

43. Which situation best represents Veracity?

A. A company collects petabytes of sensor readings
B. A platform receives user clicks every millisecond
C. A dataset contains duplicate, noisy, biased, or unreliable records
D. A business extracts profit-generating insights from customer data

44. Which pairing is most accurate?

A. Apache Kafka - batch ingestion
B. Apache NiFi - real-time ingestion
C. Apache Kafka - real-time ingestion
D. Apache Atlas - in-memory processing

45. In the Big Data architecture, which layer is mainly responsible for importing data into the system before further processing?

A. Data consumption
B. Data ingestion
C. Analytics and servicing
D. Governance

46. Which statement best explains why distributed file systems improve Big Data processing?

A. They convert all data into structured tables
B. They eliminate the need for data cleaning
C. They spread data across multiple machines, improving performance and fault tolerance
D. They guarantee that all data is accurate and meaningful

47. Which of the following is the best example of semi-structured data?

A. A table in Microsoft SQL Server
B. A JSON file containing customer activity logs
C. A video from a social media platform
D. A raw audio recording from a call center

48. Which comparison between traditional data and Big Data is correct?

A. Traditional data uses horizontal scaling; Big Data uses vertical scaling
B. Traditional data mainly supports AI and predictive analytics; Big Data mainly supports BI reporting
C. Traditional data usually has limited fault tolerance; Big Data often uses redundancy and replication
D. Traditional data uses flexible schema-on-read; Big Data uses fixed schema-on-write

49. A data science team uses an isolated environment to test machine learning models without affecting the main data infrastructure. This is an example of:

A. Data governance
B. Analytical sandboxing
C. Data ingestion
D. Batch scheduling

50. Which statement about Hadoop and Spark is most accurate?

A. Hadoop uses HDFS for storage and MapReduce for parallel processing, while Spark supports in-memory processing
B. Hadoop is mainly a NoSQL database, while Spark is mainly a governance tool
C. Hadoop is used only for real-time streaming, while Spark is used only for storage
D. Hadoop and Spark are both commercial cloud providers

Answer Sheet

Correct answers for the 50 multiple-choice questions

1. C	**11**. C	**21**. A	**31**. A	**41**. B
2. C	**12**. B	**22**. A	**32**. B	**42**. B
3. B	**13**. C	**23**. B	**33**. A	**43**. C
4. C	**14**. B	**24**. B	**34**. A	**44**. C
5. C	**15**. A	**25**. B	**35**. A	**45**. B
6. B	**16**. B	**26**. A	**36**. A	**46**. C
7. D	**17**. C	**27**. A	**37**. B	**47**. B
8. B	**18**. A	**28**. B	**38**. B	**48**. C
9. C	**19**. B	**29**. A	**39**. A	**49**. B
10. A	**20**. B	**30**. B	**40**. B	**50**. A

Topic coverage: Big Data definitions, 5Vs, history, drivers, traditional vs. Big Data approaches, architecture layers, data sources, ingestion, storage and processing, analytics, governance, Hadoop, Spark, NoSQL databases, cloud providers, and use cases.