Big Data – Tricky but simple MCQ Quiz

Q1. Which of the following is a correct "V" of Big Data?
A. Vanishability
B. Variability
C. Visibility
D. Virtuality

Q2. Which option is NOT a core reason Big Data systems use distributed systems?
A. Fault tolerance
B. Scalability
C. Reduced data size
D. Parallel processing

Q3. A company wants full control over hardware and data, with no dependency on the internet. Which deployment suits this best?
A. Cloud
B. Hybrid
C. On-premise
D. Serverless

Q4. In ETL, data is transformed:
A. After loading into storage
B. Before extraction
C. Before loading into storage
D. Only during analysis

Q5. Which statement best describes ELT?
A. Transforming data before storing
B. Storing raw data first, then transforming
C. Deleting unused data early
D. Avoiding transformation entirely

Q6. Which task is MOST aligned with a data engineer's role?

A. Designing UI dashboards

B. Training machine learning models

C. Building data pipelines

D. Writing business reports

Q7. Why is cloud computing popular for Big Data systems?

A. Data never fails

B. Hardware is owned by the user

C. Easy scalability and flexibility

D. No need for system design

Q8. A Big Data system that works today but crashes when data grows is missing which design quality?

A. Speed

B. Scalability

C. Visualization

D. Compression

Q9. Which of the following is a common misconception about Big Data?

A. It involves large datasets

B. It requires distributed systems

C. It is only about storing data

D. It needs careful system design

Q10. Which option sounds correct but is technically wrong?

A. Big Data systems handle large volumes

B. Data engineers manage data flow

C. ETL and ELT are identical processes

D. Cloud helps handle Big Data

Answer Key

Q1: B

Q2: C

Q3: C

Q4: C

Q5: B

Q6: C

Q7: C

Q8: B

Q9: C

Q10: C

Big Data Class Review Quiz (10 Questions)

Based on your class PDF. Do not write the answers on this sheet.

1. Lets talk of a 'myth' about one of the 5 V's of Big Data. Which statement about the 'Volume' of Big Data is most accurate according to the class notes?
    A. Volume is the only factor defining Big Data and has a specified limit (e.g., 1 petabyte).
    B. Volume refers to the speed at which data is generated, not the size.
    C. Volume is the size of data getting generated, but it is not the only factor, and there is no specified limit to its size.
    D. Volume primarily relates to the trustworthiness and quality of data.

2. A system processes a 'live feed' of data continuously as it comes in. According to the lectures' classification of data processing speeds, this is an example of which type of Velocity?
    A. Batch
    B. Near Real Time
    C. Delayed Time
    D. Real Time

3. Which of the following data formats is explicitly listed in an example of **Semi-structured** data?
    A. Text Files
    B. Relational Tables (Rows and Columns)
    C. XML
    D. Audio files

4. What is the term used for the practice of adding more machines (nodes) to a Distributed System to increase resources like storage and performance?
    A. Vertical Scaling
    B. Monolithic Scaling
    C. Horizontal Scaling
    D. Resource Partitioning

5. According to the table comparing Database, Data Warehouse, and Data Lake, which of the following is the primary use case for a **Data Warehouse**?
    A. Real-time transactions and Operational systems (e.g., OLTP, POS)
    B. Advanced analytics and machine learning on raw data
    C. Preparing data for a monthly sales report on structured data
    D. Business intelligence and historical data analysis

6. Which of the following describes the key characteristic of **Reliability and Fault Tolerance** in a Big Data System design?
    A. The system must balance performance and scalability with cost.
    B. The system must have built-in disaster recovery and redundancy offered by a cloud provider.
    C. The system should continue to work/operate even if some component fails.
    D. The system can process larger datasets faster by adding more compute machines.

7. What is the primary factor that causes the 'High upfront costs' associated with an **On-Premise** Big Data deployment?
    A. The pay-as-you-go pricing model for compute and storage.
    B. The high cost of maintaining built-in disaster recovery and redundancy.

C. The high upfront costs for hardware, software, maintenance, and IT staff.

D. The lack of flexibility to dynamically scale resources up or down.

8. Which statement correctly describes a key difference between **ETL** (Extract, Transform, Load) and **ELT** (Extract, Load, Transform) pipelines ?

   A. ETL handles Structured, Semi-structured, and Unstructured data, while ELT is limited to Structured data.

   B. ELT is typically slower because transformations occur beforehand, whereas ETL is faster as transformation happens after loading.

   C. ETL targets Modern data lakes/cloud platforms, while ELT targets Traditional data warehouses.

   D. In ETL, data is transformed before loading into the target system; in ELT, data is transformed after loading.

9. The Data Lake's Data Structure is classified as being able to handle which combination of data types?

   A. Only Structured data (rows and columns).

   B. Only Semi-structured (JSON, XML, CSV) and Unstructured data (Text, Video).

   C. Structured, Semi-structured, and Unstructured.

   D. Only data that has been cleaned and transformed by an ETL process.

10. A Data Engineer's focus is on 'Structured and manageable data.' In contrast, what is the primary focus of a **Big Data Engineer**?

    A. Building ETL pipelines for structured data.

    B. Working with traditional, monolithic systems.

    C. Handling massive datasets (Big Data).

    D. Analyzing social media trends for monthly reports.

Answer Key