Aakrist Goyal
2201331540002

Page No.
Date

1. What is Big Data and where do we need to apply Big Data?
→ Big Data refers to datasets characterized by the 5Vs:
Volume: Extremely large data sizes (eg. Petabytes)
Velocity: High speed generation (e.g. real time IOT data).
Variety: Diverse formats (Structured, Semi-structured, Unstructured).
Veracity: Challenges in data accuracy and quality
Value: Insights derived for decision making

Applications:
→ Healthcare: Predictive analytics for disease trends
→ Retail: Customer behaviour analysis for personalized marketing
→ Finance: Fraud detection using transaction patterns
→ Smart cities: Traffic optimization via sensor data

2. Define "Data Locality Optimization"?
→ Data Locality Optimization prioritizes processing data on the node where it is stored (e.g., in Hadoop clusters). This minimizes network transfer overhead, reduces latency, and improves efficiency.

3. List down the frameworks associated in Hadoop?
→ Hadoop's ecosystem includes multiple frameworks for storage, processing, and coordination:
- HDFS: Distributed storage system for handling large files across clusters.
- Map Reduce: Batch Processing model for parallel computation.
- YARN: Resource manager for scheduling tasks and managing cluster resources.
- Hive: SQL like interface (HiveQL) for querying data stored in HDFS.
- Pig: High level scripting language for data transformation.
- HBase: NoSQL database for real-time read/write access to large datasets.

- Spark: In-memory processing engine for faster analytics (not part of Hadoop but often integrated).
- Zookeeper: Coordination service for managing distributed systems.

4. (a) What are the big data characteristics?

→ The 5Vs define Big Data:
Volume: Massive Scale (eg. Terabytes to Zettabytes).
Velocity: High-speed data streams (eg. Social Media, IOT Sensors).
Variety: Mixed data types (eg. text, videos, logs).
Veracity: Data quality challenges (eg incomplete or noisy data)
Value: Business insights derived through analytics

5. What is Map Reduce Programming Model?

→ MapReduce is a distributed processing framework for parallel computation across clusters.
It involves two phases:-

(i) Map Phase:-
- Splits input data into chunks.
- Processes each chunk to produce intermediate key-value pairs

(ii) Reduce Phase:-
- Aggregates intermediate results by key.
- Produces final output

6. Write the difference between Operational and Analytical system with reference to Big Data?

→ Aspect	Operational Systems	Analytical Systems
• Purpose	Real-time transaction processing (OLTP)	Historical data analysis (OLAP)
• Data Type	Structured, normalized data	Aggregated, denormalized data
• Query	Simple, frequent transactions	Complex queries with joins
• Example	Banking transactions	Business intelligence dashboard

• Big Data Role	Limited due to latency constraints	Core focus (eg - Hadoop, Spark)

7. Discuss and compare NO SQL Relational Databases & produce some database names.

→

Feature	Relational (SQL) Databases	NO SQL Databases
Data Structure	Tables with rows and columns. Structured data.	Flexible structures: documents, key-value Pairs
Schema	Fixed schema	Schema-less or dynamic
Scalability	Vertical scaling	Horizontal Scaling
Consistency	ACID compliance	BASE principles
Use case	Complex queries	Unstructured data

Examples of Relational Databases
↳ MySQL, PostgreSQL, Oracle, Microsoft SQL server
Examples of NO SQL Databases
↳ MongoDB (document), Cassandra (Wide-column), Redis (Key-value), Neo4j (Graph), Amazon DynamoDB.

8. Write down any 4 industry examples of Big Data?

→ i) Healthcare:
Use case: Predictive analytics for patient readmission risks using EHR (Electronic Health Records).

→ ii) E-commerce:
Use case: Recommendation engines (e.g. Amazon's "Customers who bought this also bought").

→ iii) Transportations:
Use case: Route optimization for logistics using GPS and traffic data (e.g., Uber)

→ iv) Entertainment:
Use case: Content personalization on Netflix based on viewing history.

Aakrist Goyal
2201331540002

Page No.
Date

9. Write down the disadvantages of aggregate oriented Database and how these can be overcome?

→ Disadvantages:

- Limited Joins: Poor support for complex relationship (eg - Social Netwo
- Eventual Consistency: Data may be temporarily inconsistent across nodes.
- Redundancy: Data duplication due to denormalization.

Solutions:

- Denormalization: Pre-join data during writes to optimize reads.
- Hybrid Architecture: Combine NoSQL with relational databases for transactions (eg. CQRS pattern).
- Application-Level Joins: Handle joins in code instead of database.

10. What are different types of digital data and explore big data use in reference to cloud computing?

→ Digital Data Types:-

i) Structured: Tabular data (eg. SQL databases, CSV files).
ii) Semi-Structured: Self-describing formats (eg. JSON, XML, logs)
iii) Unstructured: No Predefined formats (eg. emails, videos, social media posts).

Big Data in Cloud Computing:

- Storage: Scalable solutions like AWS S3, Google Cloud Storage.
- Processing: Managed services like AWS EMR (Elastic MapReduce), Azure Databricks.
- Analytics: Tools like Google BigQuery (Serverless data warehousing).
- Cost Efficiency: Pay-as-you-go models reduce upfront infrastructure costs.