Q1 (a)

1. **Industries where Big Data is used**:
   - Healthcare
   - Banking and Finance
2. **Social Media Platform using Big Data**:
   - Facebook
3. **Distributed Processing Framework introduced by Apache**:
   - Apache Hadoop

Q1 (b) MCQs

1. **Best definition of Big Data**:
   - **(b) Data that exceeds the processing capacity of traditional databases**
2. **How Big Data helps prevent fraud**:
   - **(b) Using predictive analytics and machine learning to detect anomalies**
3. **Technology used for Distributed and Parallel Computing in Big Data**:
   - **(b) Hadoop**
4. **Primary function of In-Memory Computing**:
   - **(b) Processing large datasets using RAM for faster computation**
5. **Core component of Hadoop responsible for distributed storage**:
   - **(b) HDFS (Hadoop Distributed File System)**
6. **Tool used for analyzing data with Hadoop**:
   - **(b) Apache Pig**
7. **IBM's Big Data strategy includes which platform for analysis**:
   - **(a) Infosphere Big Insights and Big Sheets**

Q2 (a)

1. **Difference between HDFS and Traditional Databases**:
    - **HDFS**: Designed for storing and processing large volumes of unstructured data across distributed nodes.
    - **Traditional Databases**: Work with structured data and require schema-defined storage.
2. **Key Components of the Hadoop Ecosystem (2 Marks)**

- **HDFS (Hadoop Distributed File System) – For distributed storage**
- **MapReduce – Processing framework for distributed computing**
- **YARN (Yet Another Resource Negotiator) – Manages resources and task scheduling**
- **HBase – NoSQL database for real-time data storage**
- **Hive – SQL-like querying for Hadoop data**
- **Pig – Data flow language for processing large data sets**
- **Sqoop & Flume – Data ingestion tools**

Q2 (b)

1. Role of NameNode and DataNode in HDFS (3 Marks)

- **NameNode:**
    - Stores metadata (directory structure, file locations).
    - Manages namespace and file system operations.
- **DataNode:**
    - Stores actual data blocks.
    - Sends periodic heartbeats to NameNode.

2. Role of Big Data Analytics in Retail Industry (3 Marks)

- **Customer Personalization** – Analyzes purchase history to provide tailored recommendations.

- **Inventory Management** – Predicts demand trends to optimize stock levels.
- **Fraud Detection –** Identifies suspicious transactions in real-time.
- **Case Study:** Amazon uses Big Data to recommend products based on browsing history and purchase patterns.

Q3. Role of Hadoop in Handling Large-Scale Data (5 Marks)

Hadoop is an open-source framework that efficiently processes and stores large-scale data using distributed computing. It supports data-intensive applications by providing scalability, fault tolerance, and cost-effective storage.

1. **Scalability**
   - Hadoop scales horizontally, allowing organizations to add more nodes as data volume increases.
   - Works across a cluster of machines rather than relying on a single high-power server.
2. **Fault Tolerance**
   - Data is replicated across multiple nodes in **HDFS** (Hadoop Distributed File System).
   - If a node fails, another node provides backup, ensuring no data loss.
3. **Distributed Processing**
   - Uses **MapReduce** to process data in parallel across multiple machines.
   - Speeds up computation by breaking down tasks into smaller chunks.
4. **Cost-Effectiveness**
   - Uses **commodity hardware** instead of expensive high-performance servers.
   - Reduces storage and processing costs for big data applications.

5. **Comparison with Traditional Databases**
   - Unlike traditional databases, which handle structured data with a centralized approach, Hadoop processes **structured, semi-structured, and unstructured data** in a distributed manner.

- Ideal for large datasets where real-time transaction processing is not required.

Designing a Data Processing Solution Using Hadoop (5 Marks)

A business that handles large volumes of **unstructured data** (e.g., social media data, logs, multimedia files) requires an efficient data processing solution. Hadoop provides a scalable and distributed ecosystem to manage such data efficiently.

1. **Hadoop Distributed File System (HDFS) – Storage Layer**
   - Stores large unstructured datasets across multiple machines.
   - Uses replication (default: 3 copies) for fault tolerance and reliability.
2. **MapReduce – Processing Layer**
   - Splits data into chunks and processes them in parallel across different nodes.
   - Efficient for batch processing of large-scale unstructured data.
3. **YARN – Resource Management**
   - Manages cluster resources and job scheduling dynamically.
   - Allows multiple applications to run simultaneously in a Hadoop cluster.
4. **Apache Hive – Query and Analysis**
   - Provides an SQL-like interface for querying large datasets stored in HDFS.
   - Useful for structured analysis of unstructured data (e.g., customer trends, sentiment analysis).
5. **Apache HBase – Real-Time Processing**
   - NoSQL database that enables fast read/write operations on large data volumes.
   - Suitable for real-time analytics on semi-structured or unstructured data.

Q3. Big Data is defined by its unique characteristics, often referred to as the 5 Vs:

1. **Volume**

    - Refers to the massive amount of data generated every second from sources like social media, IoT devices, business transactions, and sensors.
    - Example: Facebook processes over 500 terabytes of new data daily.

2. **Velocity**
    - Represents the speed at which data is generated, collected, and processed.
    - Technologies like real-time analytics and stream processing (e.g., Apache Kafka, Spark Streaming) handle high-velocity data.
    - Example: Stock market data updates in milliseconds.
3. **Variety**
    - Data comes in different formats: structured, semi-structured, and unstructured.
    - Sources include databases (structured), XML/JSON logs (semi-structured), and videos/images (unstructured).
    - Example: Emails, social media posts, satellite images, etc.
4. **Veracity**
    - Ensures the quality and reliability of data by handling inconsistencies, biases, and noise.
    - Techniques like data cleansing and machine learning improve accuracy.
    - Example: Fake news detection in media platforms.
5. **Value**
    - The ultimate goal of Big Data is to extract meaningful insights that drive business decisions.
    - Example: Amazon's recommendation engine uses Big Data analytics to suggest products.

Q4. Steps to Install Hadoop (5 Marks)

- **Install Java –** Install JDK 8 or higher and check using java -version.

- **Download and Extract Hadoop –** Get Hadoop from the Apache website and extract it.
- **Set Environment Variables –** Configure Hadoop path in .bashrc or hadoop-env.sh.
- **Start Hadoop Services** – Format NameNode and run start-dfs.sh and start-yarn.sh.

Q4. MapReduce in Hadoop code in python

```
from mrjob.job import MRJob


class WordCount(MRJob):
    def mapper(self, _, line):
        for word in line.split():
            yield word.lower(), 1


    def reducer(self, word, counts):
        yield word, sum(counts)


if __name__ == '__main__':
    WordCount.run()
```