UNIT-2 (DATA ANALYTICS CS 503) Assignment Detailed Answers RUSTAMJI INSTITUTE OF TECHNOLOGY, BSF ACADEMY, TEKANPUR

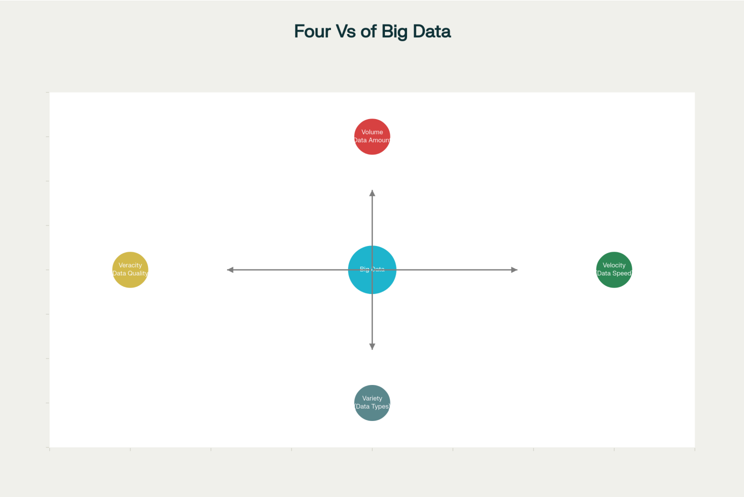
Submission deadline: 29/09/2025

Department: Computer Science & Engineering

1. Explain the Four V's of Big Data and their significance in understanding the challenges and opportunities of Big Data.

Introduction:

Big Data refers to extremely large datasets characterized by four key dimensions—Volume, Velocity, Variety, and Veracity—commonly called the Four Vs. Understanding these is crucial to managing the challenges and leveraging the opportunities presented by Big Data analytics.



Definition of the Four Vs

V Description Real-World Example Volume The sheer amount of data generated and stored Social networks like Facebook generate petabytes daily Velocity The speed of data creation and movement Stock exchange market feeds data at millisecond speeds Variety Diverse types and formats of data Text, video, images, sensor data from IoT devices Veracity Trustworthiness and accuracy of data Data quality concerns in social media misinformation

1

Significance and Challenges

- Volume challenges infrastructure to scale storage and processing.

- Velocity requires real-time processing capabilities to derive timely insights.

- Variety demands flexible data integration frameworks to handle heterogeneous formats.

- Veracity stresses the importance of clean, reliable data for accurate decision-making.

Overall, the Four Vs framework guides organizations in developing effective Big Data strategies by addressing complexity, speed, and quality.

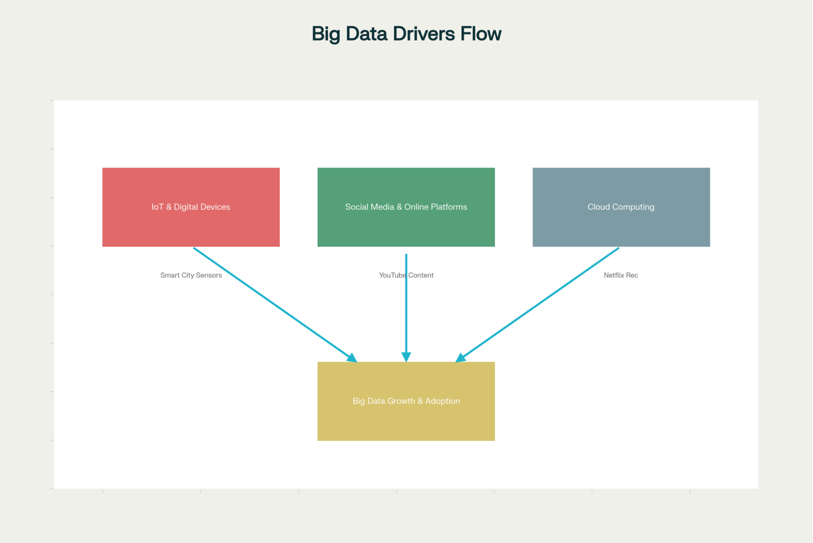
2

2. Discuss the key drivers behind the rise of Big Data.

Highlight at least three major drivers with real-world examples.

Introduction:

Big Data has escalated in importance due to several interrelated technological and societal drivers. These drivers create vast quantities of data and the need to analyze it intelligently.



Major Drivers

Driver Description Example Proliferation of Explosion of sensors and con-
Smart city traffic sen-Digital Devices nected devices generating contin-
sors, wearable health & IoT uous data streams monitors Expansion of So-

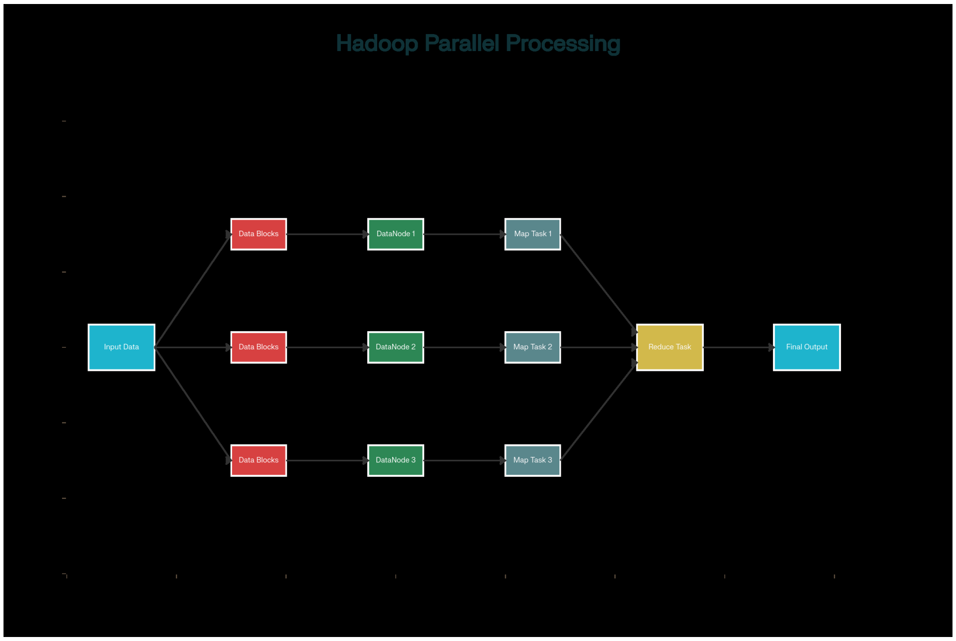
Massive user-generated content in Facebook, Twitter, cial Media and text, images, and videos YouTube content Online Platforms uploads Growth of Cloud Accessible, scalable storage and Netflix leveraging Computing compute power cloud to analyze viewer data in real time

Other Supporting Drivers • Increased adoption of Machine Learning and Artificial Intelligence • Digital transformation and automation of traditional businesses • Regulatory requirements pushing digital record-keeping (e.g., healthcare) Together, these drivers shape the Big Data landscape by increasing the data volume, variety, and the demand for sophisticated analytics platforms.

3

3. What are the core components of the Hadoop Architecture, how do they support scalable data analytics, and how does Hadoop enable parallel processing in Big Data environments?

Introduction: Hadoop is an open-source framework enabling distributed storage and parallel processing over commodity hardware, providing scalability and fault tolerance essential for Big Data analytics.



Core Components Overview

Component Role and Function Key Fea-tures/Subcomponents HDFS Distributed, fault-tolerant NameNode (metadata), storage system splitting DataNodes (storage), block data into blocks replication MapReduce Processing framework for Hadoop jobs split into Map distributed parallel comput- & Reduce tasks ing YARN Resource manager for ResourceManager, NodeM-scheduling and cluster anager, ApplicationMaster management Hadoop Com-Shared utilities, libraries, Java libraries, APIs mon and configuration for Hadoop modules

Support for Scalable Data Analytics

1. HDFS enables storage scale-out by adding commodity nodes, managing petabytes of data.

2. MapReduce parallels computation by distributing workloads across cluster nodes.

3. YARN provides resource scheduling and multi-tenant job management.

4. Hadoop’s design promotes fault tolerance with replicated data and job recovery mechanisms.

Parallel Processing Model

- Data is split into blocks and stored across nodes.
- Map tasks process data in parallel locally at each node.
- Results are shuffled, sorted, and aggregated by Reduce tasks.
- The combined output forms the final analytical result.

This architecture ensures efficient scaling and fast analytics on massive datasets.