Unit-2: Data Analytics CS503

Medium-Length Solutions with Diagrams

1 Q1: The Four V's of Big Data

Volume, **Velocity**, **Variety**, **Veracity** are the defining challenges.

Table 1: The Four V's with meaning and examples

V	Significance	Example
Volume	Need scalable storage	Web logs, genome data
Velocity	Real-time ingestion	Stock trades, IoT sensors
Variety	Flexible schema	Images, text, JSON
Veracity	Data quality management	Surveys, noisy signals

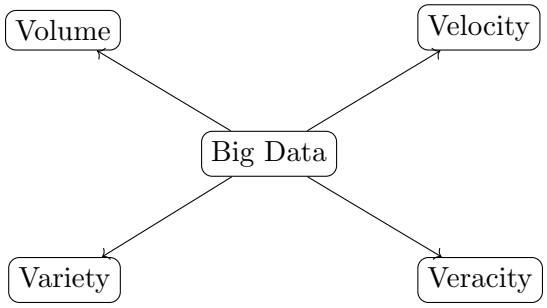

Figure 1: The Four V's visual

2 Q2: Drivers of Big Data

Main drivers: data growth, cloud affordability, and distributed frameworks.

Table 2: Key Drivers Behind Big Data Growth

Driver	Example	Impact
Data explosion	IoT, social media	Continuous huge feeds
Cheap storage	AWS S3, HDD/SSD	Store years of history
Distributed compute	Hadoop, Spark	Parallel batch + streaming

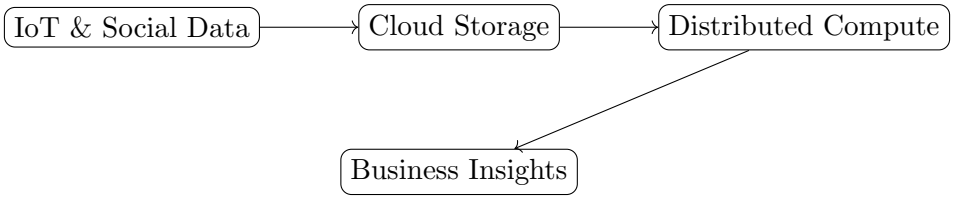

Figure 2: Drivers leading to Big Data insights

3 Q3: Hadoop Architecture

Core components: HDFS (NameNode + DataNodes), YARN (resource manager), MapReduce/Spark (parallel compute).

Parallelism: Data is split into blocks → processed locally by mappers → shuffled → reducers aggregate results.

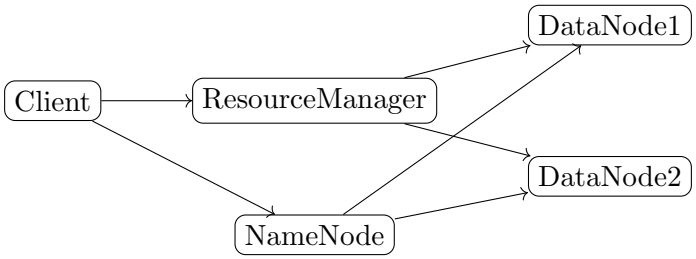

Figure 3: Simplified Hadoop cluster

4 Q4: Big Data Applications (Healthcare Case Study)

Use case: Predicting hospital readmissions.

Pipeline:

1. **Data ingestion:** EHR, labs, claims, wearables.
2. **Storage:** HDFS / cloud lake.
3. **Processing:** Feature engineering.
4. **Model:** XGBoost / Random Forest.
5. **Deployment:** Risk score in EHR dashboard.

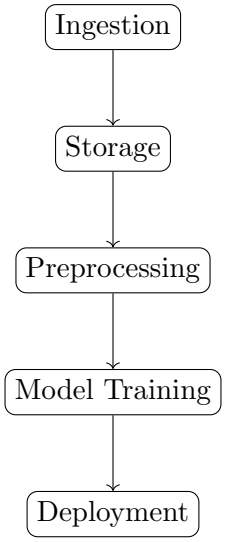

Figure 4: Healthcare analytics pipeline

Table 3: Example model results

Model	AUC	Precision	Recall
Logistic Regression	0.68	0.32	0.41
Random Forest	0.74	0.44	0.53
XGBoost	0.79	0.57	0.61

5 Q5: Short Notes

(a) Crowd Sourcing Analytics

- Definition: use distributed crowd to collect/label data.
- Uses: dataset labeling (image/NLP), citizen science.
- Pros: scalable, fast. Cons: variable quality, bias risk.

(b) Inter-/Trans-Firewall Analytics

- Definition: analytics across network firewalls or org boundaries.
- Uses: threat detection, supply-chain monitoring.
- Approaches: centralized SIEM, federated analytics.
- Challenges: privacy, heterogeneity, compliance.

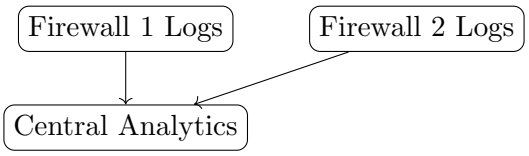

Figure 5: Inter-firewall analytics concept