Unit 2 Notes: Technical Details of Big Data Components

1. Data Storage Technologies

a. Hadoop

- Based on Map-Reduce architecture.
- Used for batch processing of large datasets.
- Stores and processes data distributedly on commodity hardware.
- Introduced by Apache Software Foundation in December 2011.
- Written in Java.

b. MongoDB

- NoSQL database (not relational like RDBMS).
- Stores data in schema documents similar to JSON.
- Suitable for large distributed architectures.
- Introduced in February 2009 by MongoDB Inc.
- Written in C++, Python, JavaScript, and Go.

c. RainStor

- Database management system using deduplication techniques.
- Helps organizations manage huge data volumes.
- Introduced in 2004 by RainStor Software Company.
- SQL-like operations.

d. Hunk

- Accesses remote Hadoop clusters with virtual indexes.
- Allows Splunk search processing on Hadoop/NoSQL data.
- Introduced in 2013 by Splunk Inc.
- Written in Java.

e. Cassandra

- Open-source, distributed NoSQL database.
- Features: Fault-tolerance, scalability, MapReduce support, etc.

Unit 2 Notes: Technical Details of Big Data Components

- Developed in 2008 by Apache Software Foundation (for Facebook).

- Written in Java.

2. Data Mining Technologies

a. Presto

- Distributed SQL query engine.

- Supports Cassandra, Hive, RDBMS, proprietary sources.

- Developed in 2013 by Apache Software Foundation.

- Written in Java.

b. RapidMiner

- Data science platform with graphical UI.

- Developed in 2001, initially called YALE.

- Written in Java.

c. ElasticSearch

- Search engine based on Lucene library.

- Distributed, real-time search.

- Developed in 2010 by Shay Banon.

- Written in Java.

3. Data Analytics Technologies

a. Apache Kafka

- Distributed streaming platform.

- Handles real-time messaging.

- Developed by Apache Software Foundation in 2011.

- Written in Java.

b. Splunk

- Indexing, searching, analyzing real-time data.

- Introduced in 2014.

- Written in AJAX, Python, C++, XML.

c. KNIME

- Platform for visual workflows.
- Built on Eclipse.
- Developed in 2008.

d. Apache Spark

- In-memory computing and batch processing.
- Developed in 2009 by Apache Software Foundation.
- Written in Java, Scala, Python, R.

e. R Language

- Statistical computing and graphics.
- Introduced in February 2000.
- Written in Fortran.

f. Blockchain

- Secures transactions, data sharing.
- Practical launch in Bitcoin 2009.
- Languages: Python, C++, JavaScript.

4. Data Visualization Technologies

a. Tableau

- Data visualization tool.
- Developed in May 2013.
- Written in Python, C, C++, Java.

b. Plotly

- Graphing library and tool.
- Introduced in 2012.

- Based on JavaScript.

5. Emerging Big Data Technologies

TensorFlow, Apache Beam, Docker, Apache Airflow, Kubernetes.

- Focus on machine learning, pipeline building, containerization, workflow automation, container management.

6. Text Analytics and Streams in Big Data

- Text Analytics: Preprocessing, NLP, Sentiment Analysis.
- Streams: Real-time processing via Kafka, Spark Streaming, Flink, Storm.
- Applications: Fraud detection, IoT analytics, stock analysis.

7. Intelligent Data Analysis (IDA) in Big Data

- Techniques: Machine Learning, Statistical Analysis, Data Mining, Predictive Analytics.
- Tools: Spark, H2O.ai, TensorFlow, KNIME, Python, R.
- Applications: Healthcare, Finance, Retail, Manufacturing, Marketing.
- Challenges: Data quality, scalability, interpretability.

8. Analytic Processes and Tools in Big Data

- Steps: Data Collection -> Storage -> Processing -> Analysis -> Visualization -> Decision Making.
- Tools: Spark, Hive, Talend, Tableau, Power BI.

9. Modern Data Analytic Tools

- Categories: Storage, Processing, Machine Learning, Integration, Visualization, Streaming.
- Examples: HDFS, BigQuery, TensorFlow, Talend, Tableau.

10. Cloud and Big Data

- Benefits: Scalability, cost-efficiency, accessibility.
- Components: Storage (S3, Azure Blob), Processing (EMR, Dataflow), ML (SageMaker, AI Platform).

11. High-Value Big Data Use Cases

- Customer Insights, Predictive Maintenance, Fraud Detection, Healthcare, Smart Cities, Cybersecurity, etc.

Unit 2 Notes: Technical Details of Big Data Components

12. Big Data Technical Components

- Storage: HDFS, S3, Data Lakes.

- Processing: Hadoop, Spark, Flink.

- Integration: ETL tools (NiFi, Talend).

- Security: Kerberos, Ranger.

13. Data Science in Big Data

- Tools: Pandas, scikit-learn, TensorFlow, Spark MLlib.

- Techniques: Machine learning, visualization, integration.

14. Big Data Exploration

- Objectives: Understand data structure, trends, anomalies.

- Tools: Spark, Hive, Tableau.

15. Security and Intelligence in Big Data

- Data Protection: Encryption, Masking.

- Real-time Monitoring: ELK Stack, Apache Metron.

- Threat Detection: ML models.

16. Operations Analysis in Big Data

- Use cases: Predictive maintenance, supply chain optimization, workforce management.

- Tools: Spark, Flink, Tableau, Databricks.