

DSM361 Big Data Fundamentals

Semester: VI	Course Title: Big Data Fundamentals	Credit: 4
Course Code: DSM361		(3 T + 1 P)

Course Outcomes: On successful completion of the course the learner will be able to

CO	COGNITIVE ABILITIES	COURSE OUTCOMES
CO1	REMEMBERING	Recall the characteristics, ecosystem components, and evolution of Big Data technologies.
CO2	UNDERSTANDING	Explain the architecture and working of Hadoop, Spark, and other Big Data frameworks.
CO3	APPLYING	Implement data processing workflows using Hadoop, Spark, Hive, and Pig.
CO4	ANALYSING	Compare various Big Data storage, processing, and analytics technologies.
CO5	EVALUATING	Evaluate Big Data tools for batch and real-time analytics based on requirements.
CO 6	CREATING	Design and implement Big Data solutions for real-world applications.

	PSO 1	PSO 2	PSO 3	PSO 4	PSO 5
CO 1	3	3	2	2	1
CO 2	2	3	3	2	1
CO 3	2	3	2	3	2
CO 4	2	3	3	2	2
CO 5	3	3	3	2	1
CO 6	3	3	3	3	2

Unit No.	Detailed Syllabus	Teaching Hours
I	Introduction to Big Data & Ecosystem Definition of Big Data, Characteristics – Volume, Velocity, Variety, Veracity, Value, Evolution of Big Data technologies, Big Data vs Traditional Data, Structured data, Semi-structured data, Unstructured data, Limitations of RDBMS, Applications – Business, Healthcare, Social media, IoT, Scientific research, Big Data ecosystem overview, Hadoop ecosystem introduction, Spark overview, Kafka overview, Hive overview, Pig overview	15
II	Hadoop Framework & Data Processing Hadoop architecture, HDFS, YARN, MapReduce components, Hadoop cluster setup concepts, HDFS data blocks, Replication, Fault tolerance, File read/write process, MapReduce programming model, Mapper, Reducer, Combiner, Partitioner, Use cases, Workflow, Data ingestion tools, Sqoop basics, Flume basics	15
III	Advanced Big Data Technologies & Analytics	15

	Apache Spark architecture, RDD concepts, Transformations, Actions, Hive architecture, Hive data types, Hive queries, Pig Latin scripting basics, HBase overview, Cassandra overview, Batch analytics, Real-time analytics, MLlib introduction, Streaming analytics concepts	
IV	Practical Applications • Setting up Hadoop and running basic HDFS commands • Implementing MapReduce programs for sample datasets • Data ingestion using Sqoop and Flume • Spark RDD transformations and actions • Hive queries and Pig scripts execution • NoSQL operations in HBase or Cassandra or MongoDB • Mini projects: - o Log analysis using Hadoop/Spark - o Real-time streaming analytics demo - o Social media sentiment analysis using Spark MLlib	15

Suggested Reference Books:

1. Big Data: Principles and Best Practices of Scalable Real-Time Data Systems – Nathan Marz, James Warren
2. Hadoop: The Definitive Guide – Tom White
3. Learning Spark: Lightning-Fast Big Data Analysis – Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia
4. Programming Hive – Edward Capriolo, Dean Wampler, Jason Rutherglen
5. HBase: The Definitive Guide – Lars George
6. Cassandra: The Definitive Guide – Eben Hewitt


H.O.D
Dept of DATA SCIENCE & ANALYTICS
M.G. Science Institute, Ahmedabad-9.