Uka Tarsadia University

B. Tech.

CSE / CSE (CC) / CE (SE)

Semester VII

Program Elective - V
BIG DATA WITH CLOUD COMPUTING
IT6014

EFFECTIVE FROM July-2024

Syllabus version: 1.00

Subject Code	Subject Title
IT6014	**Big Data with Cloud Computing**

Teaching Scheme				**Examination Scheme**				
Hours		**Credits**		**Theory Marks**		**Practical Marks**		**Total Marks**
Theory	Practical	Theory	Practical	Internal	External	Internal	External	
3	2	3	1	40	60	20	30	150

Objectives of the course:

- To understand the concept of big data and apply it through cloud programming.
- To understand the programming of Hadoop, Spark and AWS, Azure in big data.

Course outcomes:

Upon completion of the course, the student shall be able to,

CO1: Understand the background on big data, the source and value of big data.
CO2: Be proficient in Hadoop programming.
CO3: Install and use big data platforms such as Hadoop - HDFS, YARN, Spark.
CO4: Use Spark Streaming, and Spark SQL for practical use cases.
CO5: Deploy applications to cloud AWS, Azure, or Cloudera.
CO6: Understand data governance in big data with the cloud.

Sr. No.	Topics	Hours
Unit – I		
1	**Introduction to Big Data and Cloud Computing:** Big data characteristics and benefits, Data storage technologies and storing data on Amazon S3, Cloud infrastructure and management, Cloud security, Prominent cloud vendors and their services.	6
Unit – II		
2	**Data Extraction, Data Processing and Knowledge Extraction:** Data extraction process and tools, Change Data Capture (CDC), Distributed data processing, Different data processing engines, ETL using Amazon Glue, Data processing using Amazon, Data querying using Amazon Athena.	9
Unit – III		

3	**Cloud for Big Data Storage :** Introduction to storage systems, Cloud storage concepts, Distributed file systems (HDFS, Ceph FS), Cloud databases (HBase, MongoDB, DynamoDB), Cloud Object Storage (Amazon S3, OpenStack Swift).	8
Unit – IV		
4	**Programming Models:** Distributed programming for the cloud, Data-parallel analytics with Hadoop MapReduce (YARN), Data-parallel iterative analytics (Spark), and Graph-parallel analytics with graph Lab 2.0 (PowerGraph).	8
Unit – V		
5	**Big Data Deployment and Real Time Data Processing:** Spark SQL, Big data collection, Data lake, Spark streaming, Big data deployment: Hadoop solution, Cloud solution, Real-time vs batch processing, Advantages and tools, Use cases, Using Amazon Kinesis and Spark streaming.	8
Unit – VI		
6	**Data Governance:** Purpose of data governance, Metadata management, Data quality, Privacy, and security, Data orchestration using AWS data pipeline.	6

Sr.No.	**Big Data with Cloud Computing(Practicals)**	**Hours**
1	Implement cloud-based data warehouses like Amazon Redshift, Google BigQuery, or Snowflake to store and analyze large volumes of structured and unstructured data.	4
2	Implement a data lake architecture using cloud storage services such as Amazon S3 or Azure Data Lake Storage to store vast amounts of raw data in its native format for later processing and analysis.	4
3	Utilize cloud-based batch processing frameworks like Apache Hadoop or Apache Spark on platforms such as Amazon EMR or Google Dataproc to perform large-scale data processing tasks efficiently.	4
4	Implement real-time data processing pipelines using cloud-based streaming platforms like Apache Kafka or Amazon Kinesis for ingesting, processing, and analyzing streaming data in real-time.	4
5	Explore serverless computing services like AWS Lambda, Google Cloud Functions, or Azure Functions to build event-driven data processing applications without managing infrastructure, suitable for sporadic or unpredictable workloads.	4
6	Implement cloud-based data integration and ETL (Extract, Transform, Load) pipelines using tools like Apache NiFi, AWS Glue, or Azure Data Factory to ingest, clean, and transform data.	4
7	Implement cloud-based data integration and ETL (Extract, Transform, Load) pipelines using tools like Apache NiFi, AWS Glue, or Azure Data	6

	Factory to ingest, clean, and transform data from various sources for analysis.	

Text book:

1. Rudy Lai and B. Potaczek, "Hands-On Big Data Analytics with PySpark", Packt publication LTD.

Reference books:

1. "Learning Apache Spark with Python", https://runawayhorse001.githubio/LearningApacheSpark/

2. Danda B. Rawat, Lalit K Awasthi, Valentina Emilia Balas, Mohit Kumar and Jitendra Kumar Samriya "Convergence of Cloud with AI for Big Data Analytics", Wiley Publication.

Course objectives and Course outcomes mapping:

- To understand the concept of big data and apply it through cloud programming: CO1,CO2,CO3.
- To understand the programming of Hadoop, Spark and AWS, Azure in big data: CO4,CO5,CO6.

Course units and Course outcomes mapping:

Unit No.	Unit Name	Course Outcomes					
		CO1	**CO2**	**CO3**	**CO4**	**CO5**	**CO6**
1	Introduction to Big Data and Cloud Computing	✓					
2	Data Extraction, Data Processing and Knowledge Extraction		✓				
3	Cloud for Big Data Storage			✓			
4	Programming Models				✓		
5	Big Data Deployment and Real Time Data Processing					✓	
6	Data Governance						✓

Programme outcomes:

PO 1: Engineering knowledge: An ability to apply knowledge of mathematics, science, and engineering.

PO 2: Problem analysis: An ability to identify, formulates, and solves engineering problems.

PO 3: Design/development of solutions: An ability to design a system, component, or process to meet desired needs within realistic constraints.

PO 4: Conduct investigations of complex problems: An ability to use the techniques, skills, and modern engineering tools necessary for solving engineering problems.

PO 5: Modern tool usage: The broad education and understanding of new engineering techniques necessary to solve engineering problems.

PO 6: The engineer and society: Achieve professional success with an understanding and appreciation of ethical behavior, social responsibility, and diversity, both as individuals and in team environments.

PO 7: Environment and sustainability: Articulate a comprehensive world view that integrates diverse approaches to sustainability.

PO 8: Ethics: Identify and demonstrate knowledge of ethical values in non-classroom activities, such as service learning, internships, and field work.

PO 9: Individual and team work: An ability to function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO 10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give/receive clear instructions.

PO 11: Project management and finance: An ability to demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO 12: Life-long learning: A recognition of the need for, and an ability to engage in life-long learning.

Programme outcomes and Course outcomes mapping:

Programme Outcomes	Course Outcomes					
	CO1	CO2	CO3	CO4	CO5	CO6
PO1	✓	✓	✓			
PO2				✓	✓	
PO3	✓	✓	✓			
PO4				✓		
PO5						✓
PO6				✓		✓

PO7	✓	✓	✓			
PO8						✓
PO9					✓	✓
PO10			✓	✓		
PO11		✓			✓	
PO12						✓