GUJARAT TECHNOLOGICAL UNIVERSITY

Subject Code: Big Data Applications(Elective)
Subject Code: 3735503

Semester III

Type of course: ME - Computer Engineering (HIGH PERFORMANCE COMPUTING [HPC])

Prerequisite:

1. Bigdata
2. Hadoop

Rationale: NA

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks				Total Marks
L	T	P	C	Theory Marks		Practical Marks		
				ESE (E)	PA (M)	PA (V) ESE	PA (I)	
3	2#	0	4	70	30	30	20	150

L- Lectures; T- Tutorial/Teacher Guided Student Activity; P- Practical; C- Credit; ESE- End Semester Examination; PA- Progressive Assessment;

Content:

Sr. No.	Content	Total Hrs	% Weightage
1	Introduction of Apache Sqoop, Apache Oozie, Apache Crunch, Apache Flume, Apache Avro, Identifying and Collecting Input Data, Selecting Tools for Data Processing and Analysis, Presenting Results to the User, Defining and Using Data Sets, Metadata Management, Selecting a File Format, Performance Considerations, Fundamental Data Module Concepts Loading, Accessing, and Deleting a Data	6	15
2	Managing Workflows with Apache Oozie, Defining an Oozie Workflow, Validation, Packaging, and Deployment, Running and Tracking Workflows Using the CLI Hue UI for Oozie, Processing Data Pipelines, with Apache Crunch, Understanding the Crunch Pipeline, Comparing Crunch to Java Map Reduce Working with Crunch Projects, Reading and Writing Data in Crunch, Data Collection API, Functions, Utility Classes in the Crunch API, Working with Tables in Apache Hive What is Apache Hive?, Accessing Hive, Basic Query Syntax, Creating and Populating Hive Tables, How Hive Reads Data, Using the Regex in Hive,	6	15
3	Designing and Building Big Data Applications, Developing User-Defined Functions, What are User-Defined Functions?, Implementing a User-Defined Function, Deploying Custom Libraries in Hive, Registering a User-Defined Function in Hive Executing Interactive Queries with Impala, What is Impala?, Comparing Hive to Impala, Running Queries in Impala, Support for User-Defined Functions, Data and Metadata Management	7	20
4	What is Apache Solr Search?, Understanding Apache Solr Search, Search Architecture, Supported Document Formats, Indexing Data with Apache Solr Search, Collection and Schema Management, Morphlines	7	20

	Indexing Data in Batch Mode, Indexing Data in Near Real Time, Presenting Results to Users, Solr Query Syntax, Building a Search UI with Hue, Accessing Impala through JDBC Powering a Custom Web Application with, Impala and Search		

Reference Books:

1. Professional Hadoop Solutions Boris Lublinsky, Kevin T. Smith, AlexeyYakubovich
ISBN: 978-1-118-61193-7
2. Hadoop - The Definitive Guide by Tom White O'Reilly; 3 edition ISBN: 9781449311520
3. Professional Hadoop Solutions by John Wiley & Sons ISBN: 978-1118611937

Course Outcome:

After learning the course the students should be able to:

1. Describes the Data management
2. Design how the Bigdate Application are working in the Data management
3. Demonstrate the Scalable Assessment of infrastructure management with Hadoop
4. Describes how the performance analysis is done

Review Presentation (RP): The concerned faculty member shall provide the list of peer reviewed Journals and Tier-I and Tier-II Conferences relating to the subject (or relating to the area of thesis for seminar) to the students in the beginning of the semester. The same list will be uploaded on GTU website during the first two weeks of the start of the semester. Every student or a group of students shall critically study 2 papers, integrate the details and make presentation in the last two weeks of the semester. The GTU marks entry portal will allow entry of marks only after uploading of the best 3 presentations. A unique id number will be generated only after uploading the presentations. Thereafter the entry of marks will be allowed. The best 3 presentations of each college will be uploaded on GTU website