GUJARAT TECHNOLOGICAL UNIVERSITY

BRANCH NAME: INFORMATION TECHNOLOGY
SUBJECT NAME: BIG DATA ANALYTICS
SUBJECT CODE: 3712313

Type of course: Elective

Prerequisite: Data Structure, Computer Architecture and Organization

Rationale:
Understand big data for business intelligence. Learn business case studies for big data analytics. Understand NoSQL big data management. Perform map-reduce analytics using Hadoop and related tools.

Teaching and Examination Scheme:

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

Content:

Sr. No.	Content	Total Hrs	% Weightage
1	What is big data, why big data, convergence of key trends, unstructured data, industry examples of big data, web analytics, big data and marketing, fraud and big data, risk and big data, credit risk management, big data and algorithmic trading, big data and healthcare, big data in medicine, advertising and big data, big data technologies, introduction to Hadoop, open source technologies, cloud and big data, mobile business intelligence, Crowd sourcing analytics, inter and trans firewall analytics.	**7**	14%
2	Introduction to NoSQL, aggregate data models, aggregates, key-value and document data models, relationships, graph databases, schemaless databases, materialized views, distribution models, sharding, master-slave replication, peer peer replication, sharding and replication, sharding in MongoDB, consistency, relaxing consistency, version stamps, map-reduce, partitioning and combining, composing map-reduce calculations.	**9**	20%
3	Data format, analyzing data with Hadoop, scaling out, Hadoop streaming, Hadoop pipes, design of Hadoop distributed file system (HDFS), HDFS concepts, Java interface, data flow, Hadoop I/O, data integrity, compression, serialization, Avro, file-based data structures	**9**	18%
4	MapReduce workflows, unit tests with MRUnit, test data and local tests, anatomy of MapReduce job run, classic Map-reduce, YARN, failures in classic Map-reduce and YARN, job scheduling, shuffle and sort, task execution, MapReduce types, input formats, output formats	**10**	20%

5	Hbase, data model and implementations, Hbase clients, Hbase examples, praxis. Cassandra, Cassandra data model, Cassandra examples, Cassandra clients, Hadoop integration.	**7**	14%
6	Pig, Grunt, pig data model, Pig Latin, developing and testing Pig Latin scripts. Hive, data types and file formats, HiveQL data definition, HiveQL data manipulation, HiveQL queries.	**6**	14%
	Total	**48**	**100%**

Reference Books:

1. DT Editorial Services, Big Data Black Book, Dreamtech Press
2. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.
3. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.
4. Eric Sammer, "Hadoop Operations", O'Reilley, 2012.
5. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.
6. Lars George, "HBase: The Definitive Guide", O'Reilley, 2011.
7. Eben Hewitt, "Cassandra: The Definitive Guide", O'Reilley, 2010.
8. Alan Gates, "Programming Pig", O'Reilley, 2011.
9. MongoDB: The definitive Guide, 3rd Edition, O'Reilley, 2016.

Course Outcome:

After completion of course, students would be able to:

- Describe big data and use cases from selected business domains
- Explain NoSQL big data management
- Install, configure, and run Hadoop and HDFS
- Understand use of MongoDB to support deployment with very large data sets and high throughput operations
- Perform map-reduce analytics using Hadoop
- Use Hadoop related tools such as HBase, Cassandra, Pig, and Hive for big data analytics

List of Experiments:

1. To install Hadoop 3.X on Ubuntu.
2. To count words in the input document on Hadoop using Map Reduce Programming.
3. To install MongoDB and perform various commands in MongoDB.
4. To load a local file on to HDFS from local file system.
5. To install Cassandra and perform CRUD operations on Cassandra.
6. To use Pig in Hadoop local mode and Map-reduce mode.
7. To perform relational operations on database using pig.

List of Open Source Software/learning website:

Apache Hadoop, Apache Cassandra, Mongo DB, Pig