## BIG DATA ANALYTICS

| Course Code | 23CS4602 | Year | III | Semester | II |
|---|---|---|---|---|---|
| Course Category | Program Elective-III | Branch | CSE | Course Type | Theory |
| Credits | 3 | L – T – P | 3-0-0 | Prerequisites | Machine Learning |
| Continuous Evaluation: | 30 | Semester End Evaluation: | 70 | Total Marks: | 100 |

| Course Outcomes | | |
|---|---|---|
| Upon successful completion of the course, the student will be able to: | | |
| CO1 | Understand the fundamental concepts of big data | L2 |
| CO2 | Apply NoSQL for storing Big data | L3 |
| CO3 | Apply the concepts of Map reduce and Hive for processing the Big Data | L3 |
| CO4 | Apply the SPARK techniques in achieving data analytics with scalability and streaming capability. | L3 |

## Syllabus

| Unit No. | CONTENTS | Mapped CO |
|---|---|---|
| I | 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. | CO1 |
| II | Introduction to NoSQL, aggregate data models, aggregates, key-value and document data models, relationships, graph databases, schema less databases, materialized views, distribution models, sharding, master-slave replication, peer- peer replication, sharding and replication, consistency, relaxing consistency, version stamps. | CO1,CO2 |

| | | |
|---|---|---|
| **III** | A Weather Dataset: Data format, Analyzing data with Hadoop, scaling out, The Design of Hadoop distributed file system (HDFS), HDFS Concepts: HDFS Federation, HDFS- High availability, Data locality, Java interface, data flow.<br><br>Introduction to Hive, data types, HiveQL data definition, HiveQL data manipulation, Table partitioning, Bucketing, Join strategies. | **CO1, CO3** |
| **IV** | Apache spark- Advantages over Hadoop, lazy evaluation, In memory processing, DAG, Spark context, Spark Session, RDD, Transformations- Narrow and Wide, Actions, Data frames ,RDD to Data frames, Catalyst optimizer, Data Frame Transformations, Working with Dates and Timestamps, Working with Nulls in Data, Joins, Broadcast Variables, Accumulators. | **CO1,CO4** |
| **V** | Spark- Stream Processing Fundamentals, Structured Streaming Basics - Core Concepts, Structured Streaming in Action, Transformations on Streams, Input sources and Output modes, Event-Time and State full Processing - Event Time, State full Processing. | **CO1,CO4** |

**Learning Resources**

**Text Books**


1. Big Data, Big Analytics: Emerging, Michael Minnelli, Michelle Chambers, and AmbigaDhiraj, 1st edition ,2013
2. SPARK: The Definitive Guide, Bill Chambers &MateiZaharia, O'Reilley, 2018-first Edition.
3. Business Intelligence and Analytic Trends for Today's Businesses", Wiley, First edition-2013.
4. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World Polyglot Persistence", Addison-Wesley Professional, 2012
5. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012


**Reference Books**


1. "Hadoop Operations", O'Reilley, Eric Sammer, First Edition -2012.
2. "Programming Hive", O'Reilley, E. Capriolo, D. Wampler, and J. Rutherglen, 2012.
3. "HBase: The Definitive Guide", O'Reilley, Lars George, September 2011: First Edition.


**E-Resources & other digital material**

1. https://nptel.ac.in/courses/106/104/106104189/
2. https://www.coursera.org/specializations/big-data
3. https://www.edx.org/course/big-data-fundamentals