L	T	P	C
3	0	0	3

Course Code: INT404R01
Semester: VII

BIG DATA ANALYTICS

Course Objective:

This course will help the learner to discuss the fundamental techniques of data science and data analysis suitable for big data and solve it using Hadoop.

UNIT - I **12 Periods**

Getting Started with Big Data: Grasping the Fundamentals of Big Data. **An Operating System for Big Data:** Basic Concepts - Hadoop Architecture - Working with a Distributed File System - Working with Distributed Computation - Submitting a MapReduce Job to YARN. **A Framework for Python and Hadoop Streaming:** Hadoop Streaming - A Framework for MapReduce with Python - Advanced MapReduce

UNIT - II **11 Periods**

MapReduce and the New Software Stack: Map Reduce Algorithms - Communication Cost Model. **Mining Data Streams:** Stream Data Model - Sampling Data in a Stream - Filtering Streams - Counting Distinct elements in a Stream - Estimating moments - Counting ones in a Window - Decaying Windows. **Link Analysis:** Pagerank Algorithm - Efficient Computation of Pagerank - Topic-Sensitive PageRank - Link Spam.

UNIT - III **11 Periods**

Recommendation Systems: A Model for Recommendation System - Content Based Recommendations - Collaborative Filtering - Dimensionality Reduction. **Mining Social Network Graphs:** Social Networks as Graphs - Clustering of Social Network Graphs - Direct Discovery of Communities - Partitioning of Graphs - Finding overlapping communities - Simrank - Counting Triangles.

UNIT - IV **11 Periods**

Structured Data Queries with Hive: The Hive Command-Line Interface (CLI), Hive Query Language, Data Analysis with Hive, HBase. **Data Ingestion:** Relational Data with Sqoop. **Analytics with Higher-Level APIs:** Pig, Spark's Higher-Level APIs. **Machine Learning:** Scalable Machine Learning with Spark

TEXTBOOKS

1. Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman, *Mining of Massive Data Sets*, 2019.
2. Benjamin Bengfort, and Jenny Kim, *Data Analytics with Hadoop*, O'Reilly, 2016.
3. Judith Hurwitz, Alan Nugent, Dr. Fern Halper, and Marcia Kaufman, *Big Data for Dummies*, John - Wiley and Sons, 2013.

REFERENCES

1. Jimmy Lin, Chris Dyer, *Data-Intensive Text Processing with MapReduce*, Morgan & Claypool Publishers, 2010.
2. Paul C. Zikopoulos, Chris Eaton, and Dirk deRoos, Thomas Deutsch, George Lapis, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, The McGraw-Hill Companies, 2012.

3. Srinath Perera, and Thilina Gunarathne, *Hadoop MapReduce Cookbook*, Packt Publishing, 2013.
4. Tim McGovern, *Big Data Now: 2014 Edition*, O'Reilly Media, Inc., 2015.
5. Jason Venner, *Pro Hadoop*, Apress, 2009.

UNITWISE LEARNING OUTCOMES

Upon successful completion of each unit, the learner will be able to

Unit I	• Discuss the fundamental concepts of big data and Hadoop - MapReduce framework • Demonstrate the MapReduce algorithms using python and Hadoop Streaming
Unit II	• Explain the methods of processing and filtering of data streams • Apply link analysis for page rank computation
Unit III	• Describe the methods for developing a model for recommendation system. • Explain dimensionality reduction and its implementation • Solve community detection problems in social network graphs by applying algorithms for clustering and partitioning
Unit IV	• Illustrate the different programming languages in Hadoop such as Hive, HBase, Pig, Spark

COURSE LEARNING OUTCOMES

Upon successful completion of this course, the learner will be able to

CO No.	Course Outcomes	Knowledge Level
1	Discuss the fundamental concepts of big data and Hadoop - MapReduce framework	K3
2	Demonstrate the MapReduce algorithms using python and Hadoop Streaming	K4
3	Explain the methods of processing and filtering of data streams and page rank algorithms	K3
4	Describe the methods for developing a model for recommendation system with dimensionality reduction	K3
5	Solve community detection problems in social network graphs by employing algorithms for clustering and partitioning	K3
6	Illustrate the different programming languages in Hadoop such as Hive, HBase, Pig, Spark	K3

L	T	P	C
0	0	2	1

Course Code: INT435
Semester: VII

BIG DATA ANALYTICS & APPLICATIONS LABORATORY

Course Objectives

This course will help the learner to practice the fundamental techniques of data analysis suitable for big data using Hadoop framework and R programming language.

LIST OF EXERCISES

1. Design MapReduce technique for word counting using python on Hadoop cluster
2. Develop MapReduce algorithm for finding the coolest year from the available Weather data using java program on Hadoop cluster
3. Design a bloom filter to remove the duplicate users from the Log file and analyse the filter with different cases.
4. Implement the Flajolet-Martin algorithm to extract the distinct twitter users from the twitter data set.
5. Demonstrate the significance Page rank algorithm in the Hadoop platform with available data set using MapReduce based Matrix vector multiplication algorithm.
6. Design a friend of friend's network using Girvan Newman algorithms from the social network data.
7. Demonstrate the relational algebra operations such as sort, group, join, project, and filter using Hive and Pig.
8. Load the unstructured data into the Hadoop and convert it into the structured data using Hive. Develop a Hive and HBase Databases, Tables, Views, Functions and Indexes and perform the some perform basic query operations.
9. Implement a Pig Latin scripts to sort, group, join, project, and filter your data.
10. Implement the collaborative filtering system using PySpark
11. Perform the Logistic regression classification, SVM and Decision tree classifier algorithms using PySpark and display the result with graph and compare the accuracy of an algorithms using Precision, Recall and F-Measure.
12. Implement the KMean clustering algorithm using PySpark.

COURSE LEARNING OUTCOMES

Upon successful completion of this course, the learner will be able to:

CO No.	Course Outcome	Knowledge Level
1	Demonstrate MapReduce algorithms using python and Java on Hadoop cluster	K3
2	Experiment Bloom filter and Flajolet-Martin algorithms on the streaming data	K4
3	Solve community detection problems in social network graphs by employing algorithms for clustering and partitioning	K3
4	Experiment different relational algebra operations using Hive and Pig	K4
5	Demonstrate collaborative filtering based recommendation system using PySpark	K3
6	Analyze linear regression, SVM and Decision tree and K-Mean algorithms using PySpark	K4