B.Tech. III (CSE) Semester – V
DATA SCIENCE (CORE ELECTIVE-1)
CS321

	L	T	P	Credit
Scheme	**3**	**0**	**0**	**03**

1. Course Outcomes (COs): **At end of the Course student will be able to**	
CO1	understand types of data and various data science approaches.
CO2	apply various data pre-processing and manipulation techniques including various distributed analysis paradigm using hadoop and other tools and perform advance statistical analysis to solve complex and large dataset problems.
CO3	analyse different large data like text data, stream data, graph data.
CO4	interpret and evaluate various large datasets by applying Data Mining techniques like clustering, filtering, factorization.
CO5	design the solution for the real life applications.

2. Syllabus

- **INTRODUCTION** **(02 HOURS)**

  Examples, Applications and Results Obtained Using Data Science Techniques, Overview of the Data Science Process.

- **MANAGING LARGESCALE DATA** **(02 HOURS)**

  Types of Data and Data Representations, Acquire Data (E.G., Crawling), Process and Parse Data, Data Manipulation, Data Wrangling and Data Cleaning.

- **PARADIGMS FOR DATA MANIPULATION, LARGE SCALE DATA SET** **(08 HOURS)**

  Mapreduce (Hadoop), Query Large Data Sets in Near Real Time with Pig and Hive, Moving from Traditional Warehouses to Map Reduce, Distributed Databases, Distributed Hash Tables.

- **TEXT ANALYSIS** **(10 HOURS)**

  Data Flattening, Filtering and Chunking, Feature Scaling, Dimensionality Reduction, Nonlinear Factorization, Shingling of Documents, Locality Sensitive Hashing for Documents, Distance Measures, LSH Families for Other Distance Measures, Collaborative Filtering.

- **MINING DATA STREAM** **(08 HOURS)**

  Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Moments, Windows, Clustering for Streams.

- **ADVANCE DATA ANALYSIS** **(12 HOURS)**

  Graph Visualization, Data Summaries, Hypothesis Testing, ML Model-Checking and Comparison, Link Analysis, Mining of Graph, Frequent Item Sets Analysis, High Dimensional Clustering, Hierarchical Clustering, Recommendation Systems.

Total Contact Time: 42 Hours

---

3. Books Recommended:

1. Tom White, “Hadoop: The Definitive Guide”, 4th Edition, O’reilly Media, 2015, ISBN: 9781491901687.
2. Anand Rajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, 2nd Edition, Cambridge University Press, 2014, ISBN: 9781107077232.
3. Peter Bruce, Andrew Bruce, “Practical Statistics for Data Scientists: 50” by , 1st Edition, O’reilly publishing house, 2017, ISBN: 9781491952962.
4. Joel Grus, J. “Data science from scratch”, 1st Edition, O'Reilly Media, 2015, ISBN: 9781491901410.
5. Montgomery, Douglas C., and George C. Runger. “Applied statistics and probability for engineers”, John Wiley & Sons, 7th Edition, 2018, ISBN: 9781119400363.