



INFORMATION TECHNOLOGY
ENGINEERING

IT ENGINEERING SEM VIII

BIG DATA ANALYTICS

Programming & development

Course Curriculum



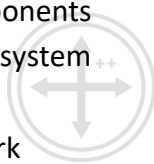
BIG DATA ANALYTICS SEM VIII

Module 1 : Introduction to Big Data

- Introduction to Big Data
- Big Data characteristics
- types of Big Data
- Traditional vs. Big Data business approach
- Big Data Challenges
- Examples of Big Data in Real Life
- Big Data Applications

Module 2: Data Frameworks : Hadoop, NoSQL

- What is Hadoop?
- Core Hadoop Components
 - Hadoop Ecosystem
- Overview of :
 - Apache Spark
 - Pig
 - Hive
 - Hbase
 - Sqoop
- What is NoSQL?
 - NoSQL data architecture patterns:
 - Key-value stores
 - Graph stores
 - Column family (Bigtable) stores
 - Document stores
- Mongo DB



Module 3: MapReduce Paradigm

- MapReduce
 - The Map Tasks
 - Grouping by Key



- The Reduce Tasks
- Combiners
- Details of MapReduce Execution
- Coping With Node Failures
- Algorithms Using MapReduce
 - Matrix-Vector Multiplication by MapReduce
 - Relational-Algebra Operations
 - Computing Selections by MapReduce
 - Computing Projections by MapReduce
 - Union
 - Intersection
 - Difference by MapReduce
 - Computing Natural Join by MapReduce
 - Grouping and Aggregation by MapReduce
 - Matrix Multiplication
 - Matrix Multiplication with One MapReduce Step
- Illustrating use of MapReduce with use of real life databases and applications

Module 4: Mining Big Data Streams

- The Stream Data Model
 - A Data-Stream-Management System
 - Examples of Stream Sources
 - Stream Queries
 - Issues in Stream Processing
- Sampling Data in a Stream
 - Sampling Techniques
- Filtering Streams
 - The Bloom Filter
- Counting Distinct Elements in a Stream
 - The Count-Distinct Problem
 - The Flajolet-Martin Algorithm
 - Combining Estimates
 - Space Requirements
- Counting Ones in a Window
 - The Cost of Exact Counts
 - The Datar-Gionis-Indyk-Motwani Algorithm
 - Query Answering in the DGIM Algorithm



Module 5: Big Data Mining Algorithms

- Frequent Pattern Mining
 - Handling Larger Datasets in Main Memory Basic Algorithm of Park
 - Chen and Yu
 - The SON Algorithm and MapReduce
- Clustering Algorithms
 - CURE Algorithm
 - Canopy Clustering
 - Clustering with MapReduce
- Classification Algorithms
 - Parallel Decision trees
 - Overview SVM classifiers
 - Parallel SVM
 - K-Nearest Neighbor classifications for Big Data
 - One Nearest Neighbour

Module 6: Big Data Analytics Applications

- Link Analysis
 - PageRank Definition
 - Structure of the web
 - dead ends
 - Using Page rank in a search engine
 - Efficient computation of Page Rank
 - PageRank Iteration Using MapReduce
 - Topic sensitive Page Rank
 - link Spam
 - Hubs and Authorities
 - HITS Algorithm
- Mining Social- Network Graphs
 - Social Networks as Graphs
 - Types
 - Clustering of Social Network Graphs
 - Direct Discovery of Communities
 - Counting triangles using Map-Reduce
- Recommendation Engines
 - A Model for Recommendation Systems
 - Content-Based Recommendations
 - Collaborative Filtering

