BCA(N)-125

Data Science and Big Data

1. Basics of Data Science

Introduction to Data Science: Probability: Conditional Probability and Bayes Theorem, Random Variables and Basic Distributions: Binomial Distribution, Probability Distribution of Continuous Random Variable & The Normal Distribution, Sampling Distribution and the Central Limit Theorem, Statistical Hypothesis Testing: Estimation of Parameters of the Population, Significance Testing of Statistical Hypothesis, Example Using Correlation and Regression & Types of Errors in Hypothesis Testing.

Data Preparation for Analysis: Need for Data Preparation, Data Preprocessing: Data Cleaning, Data Integration, Data Reduction & Data Transformation, Selection and Data Extraction, Data Curation: Steps of Data Curation, Importance of Data Curation, Data Integration: Data Integration Techniques & Data Integration Approaches, Knowledge Discovery.

Analysis of Simple Algorithm: Complexity analysis of Algorithms, Euclid Algorithm for GCD, Polynomial Evaluation Algorithm, Exponent Evaluation, Sorting Algorithm

Data Visualization and Interpretation: Different types of plots, Histograms, Box plots, Scatter plots, Heatmap, Bubble Chart, Bar Chart, Distribution plot, Pair plot, Line graph, Pie Chart, Doughnut Chart, Area Chart.

2. Big Data and its Management

Big Architecture: Big Data and Characteristics and Big Data Applications, Big data Applications, Structured vs Semi-structured and Unstructured Data, Big Data vs Data Warehouse, Distributed File System, HDFS and Map Reduce, Apache Hadoop 1 and 2 (YARN).

Programming Using MapReduce: Map Reduce Operations, Loading data into HDFS: Installing Hadoop and Loading Data, Executing the MapReduce phases: Execution of Map Phase, Shuffling and sorting, Reduce phase execution & Node Failure and MapReduce, Algorithms using MapReduce: Word counting and Matrix-Vector Multiplication.

Other Big data Architectures and Tools: Apache SPARK framework, HIVE: Working of HIVE Queries, Installation of HIVE and Writing Queries in HIVE, HBase: HBase Installation & Working with HBase, Other tools **.**

3. Big Data Analysis

Mining Big Data: Finding Similar Items: Jaccard Similarity of Sets, Documents Similarity & Collaborative Filtering and Set Similarity, Finding Similar Documents: Shingles, Minhashing &

Locality Sensitive Hashing, Distance Measures: Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance & Hamming Distance, Introduction to Other Techniques.

Mining Data Streams: Data Streams: Model for Data Stream Processing, Data Stream Management: Queries of Data stream,

Example of Data Stream & Issues and Challenges of Data Stream, Data Sampling in Data Streams: The Representation Sample, Filtering of Data Streams: Bloom Filter, Algorithm to Count Different Elements in Stream.

Link Analysis: Link analysis, Page Rankin, Different Mechanisms of Finding PageRank: Finding PageRank & Web Structure and Associated Issues, Use of PageRank in Search Engines: Page Rank computation using Map-reduce, Topic sensitive PageRank, Link Spam, Hubs and Authorities.

Web and Social Network Analysis: Introduction to Web Analytics, Advertising on the Web: Issues & Algorithm, Recommendation Systems: The Long Tail, The Model and Content-Based Recommendations, Mining Social Networks: Social Networks as Graphs, Varieties of Social Networks, Distance Measure of Social Network Graphs & Clustering of Social Network Graphs.

4. Programming for Data Analysis

Basics of R Programming: Environment of R, Data types, Variables, Operators, Factors, Decision Making, Loops, Functions, Data Structures in R: Strings and Vectors, Lists & Matrices, Arrays and Frames.

Data Interfacing and Visualisation in R: Reading Data From Files: CSV File, Excel File, Binary Files, XML Files, JSON Files, interfacing with Databases & Web Data, Data Cleaning and Pre-processing, Visualisations in R: Bar Charts, Box Plots, Histograms, Line Graphs & Scatterplots.

Data Analysis and R: Chi-Square Test, Linear Regression, Multiple Regression, Logistic Regression, Time Series Analysis.

Advance Analysis Using R: Decision Trees, Random Forest, Classification, Clustering, Association Rules.