Specialised Programme on Big data Analytics
(12 Weeks)

OS basics and Linux

Linux History and Operation ,The Evolution of Linux ,Basics of Linux operating system,Linux Operations as a Server ,The Architecture and Structure of Linux Installing and Configuring Linux (Ubuntu and CentOS) Introduction to Installation and Media Types ,Logging In and Out of a Linux System,Basic Commands (ls, cp, mv, sort, grep, cat,head,tail, man, locate, find, diff, file, rm, mkdir, rmdir, cd, pwd, ln and ln –s, gzip and gunzip, zip and unzip, tar and its variants, touch, echo, who, whoami, ps, kill,makefile,etc.) Getting Acquainted with the Linux Environment, Use various commands in Linux system. As root, create a directory dbda and under it create a directory named test and create 100 files under it with name file1, file2..file100 - all this using a single command: Gaining confidence with Linux o Access control list and chmod command,chown and chgrp commands ,Commands like telnet, ftp, ssh, and sftp o Basic of I/O system with mount and unmount. vi/vim/gedit editior ,Features and different modes of vi editor Editing using vi editor Find and replace commands,cut-copy-paste commands ,The set command ,Other related commands of vi

Python Programming

Installing Python, Introduction to Python Basic Syntax, Data Types,Variables,Operators, Input/output, Strings, Python data structure ,Lists, Tuples, Dictionaries, Sets., If, If- else, Nested if-else,Looping, For, While, Nested loops,Control Structure, Uses of Break & Continue,Functions and methods and Exception Handling,OOPs Concepts,Python classes and objects,Introduction and Installation of Machine learning packages like PANDAS, NUMPY,SKLearn, Matplotlib, Seaborn,Mathematical Computing with Python (NumPy),Data Manipulation with Pandas,Machine Learning with Scikit–Learn.,Introduction to Data Visualization in Python (matplotlib)

Statistical Analysis with R programming

Introduction to statistics

Descriptive Statistics, summary statistics, Plots(Box Plots, Scatter plot, Pie Charts, Bar charts, Histogram), Basic probability theory, uni-variate and multi-variate distribution, Data Exploration & preparation, Concepts of Correlation, Regression, Hypothesis Testing , Parametric Tests: ANOVA ,Non-parametric Tests- chi-Square, Overview of Factor Analysis, Directional Data Analytics , Functional Data Analysis ,Dimensionality issues, Ridge & lasso regression, bias/variance trade off, density, PCA, feature selection, Bagging and boosting, Simulation : Monte carlo

Introduction to R programming

Introduction & Installation of R, Exploring RStudio, Basic Mathematical & Arithmetic operations in R, Data Objects- Data Types & Data Structures (e.g. vectors, lists. Arrays, matrices, data frames) ,Packages in R , Working with Packages ,Handling Data in R Workspace, Reading & Importing data From Text files, Excel files, Multiple databases, Exporting Data from R . Advanced visualization with R.

Introduction to Business Analytics using some case studies, Case studies: Making Right Business Decisions based on data , Visualization and Exploring Data , Descriptive Statistical

Measures, Probability Distribution and Data, Sampling and Estimation, Statistical Interfaces, Regression Analysis, Forecasting Techniques, Simulation and Risk Analysis, Optimization, Linear, Non linear, Integer, Decision Analysis, strategy and Analytics .

Big Data and Hadoop , Map Reduce, Apache sparks, MongoDB

Introduction to Big data and Hadoop

Introduction to big data platform, Structured and unstructured data, Challenges of Conventional Systems, Big data use cases, Introduction to Hadoop, Brief History and Evolution of Hadoop, Comparison with Other Systems. Big Data/Hadoop Platforms, Hadoop Distributions and Vendors.

Hadoop Environment

Hadoop Installation, Setting up a Hadoop Cluster, Cluster specification, Cluster Setup and Installation, Single and Multi-Node Cluster Setup on Virtual Machine, Hadoop Configuration, Security in Hadoop, Administering Hadoop, Hadoop benchmarks.

Hadoop Architecture

Hadoop Architecture, Core components of Hadoop, Common Hadoop Shell commands

HDFS

Distributed File System, What is HDFS, Where does HDFS fit in, Core components of HDFS, HDFS Daemons, Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node, Hadoop Distributed File System, Components of Hadoop, Design of HDFS,HDFS Architecture overview, Hadoop distribution and basic commands, The HDFS command line and web interfaces, Analyzing the Data with Hadoop, Scaling Out, Hadoop event stream processing, high availability and Name Node federation, HDFS – Monitoring & Maintenance.

Introduction to Map Reduce Framework

Hadoop Map Reduce paradigm, Map and Reduce tasks, Map Reduce Execution Framework, Map Reduce Daemons, Anatomy of a Map Reduce Job run, Partitioner and Combiners, Input Formats and Output Formats. Map Reduce program structure, Use of combiner and partitioner.

Introduction to Hive

The Hive Data-ware House, Hive architecture and Installation, Comparison with Traditional Database, Basics of Hive Query Language, Working with Hive QL, Datatypes, Operators and Functions, Hive Tables (Managed Tables and Extended Tables), Partitions and Buckets, Storage Formats, Importing data, Altering and Dropping Tables. Querying with Hive QL, Querying Data-Sorting, Aggregating, Joins and Sub queries, Views, Data manipulation with Hive, User Defined Functions, Appending data into existing Hive table, Writing HQL scripts. Data analysis with Flume and Sqoop.

Introduction to Sparks

Overview, Linking with Spark, Initializing Spark, Resilient Distributed Datasets (RDDs), External Datasets, RDD Operations, Passing Functions to Spark, Working with Key-Value Pairs, Shuffle operations, RDD Persistence, Removing Data, Shared Variables, Deploying to a Cluster , Working with Spark with Hadoop , Working with Spark without Hadoop and their differences , Spark MLlib ,Spark SQL and Spark Streaming.

Introduction to MongoDB

Overview of SQL (DDL, DML, TCL), Introduction to NoSQL, Difference between SQL and NoSQL, working with MongoDB (Installation, CRUD operations, Aggregation pipeline, Indexing, Data Modeling)

Machine Learning

Introduction to Machine Learning and data preprocessing,What is machine learning? Types of learning,Applications of Machine learning,Evaluating ML techniques.,Data cleaning,Scaling of continuous features,Encoding of categorical features,Train and Test Split,Machine Learning Algorithms,Linear Regression,Download Dataset, perform linear and multiple linear regression and check for MAE, MSE, RMSE and also check F1 score and explain with conclusion.Bayesian analysis and Naïve bayes classifier,Assigning probabilities and calculating results, Naïve Bayes case study,K-Nearest Neighbors ,Algorithm and case study,Decision Trees, Decision Trees case study,Ensemble Learning: Concept of model ensembling,Random forest,Gradient boosting Machines,Model Stacking,Support Vector Machines,Logistic regression classification algorithm and check for classification report, f1 score for all algorithms.ML in Real Time,Algorithm Performance Metrics , ROC and AOC,Confusion Matrix , F1 Score , MSE and MAE,Different type of Unsupervised Machine Learning Algorithms,Clustering, K-mean,Agglomerative clustering,Association rule mining,Apriori algorithm