

## **PG Diploma in Big Data Analytics**

### **Linux Programming and Cloud Computing - 50 Hours**

**Linux Programming:** Installation (Ubuntu and CentOS), Basics of Linux, Configuring Linux, Shells, Commands, and Navigation, Common Text Editors, Administering Linux, Introduction to Users and Groups, Linux shell scripting, shell computing, Introduction to enterprise computing, Remote access.

**Introduction to Git/GitHub/Gitlab:** Introduction to Version control systems, Creating GitHub repository, Using Git – Introduction to get commands, Creating projects on Github/gitlab and managing code repos.

**Introduction to Cloud Computing:** Cloud Computing Basics, Understanding Cloud Vendors (AWS:EC2 instance, lambda and Azure: Azure virtual machines, Azure data factory), Definition, Characteristics, Components, Cloud provider, SAAS, PAAS, IAAS and other Organizational scenarios of clouds, benefits and limitations, Deploy application over cloud. Comparison among SAAS, PAAS, IAAS, Cloud Products and Solutions, Compute Products and Services, Elastic Cloud Compute, Dashboard, deploy AI and analytics workloads in Cloud environments with sample mini project.

### **Python and R Programming - 80 Hours**

**Python Programming:** Python basics, If, If- else, Nested if-else, Looping, For, While, Nested loops, Control Structure, Break, Continue, Pass, Strings and Tuples, Accessing Strings, Basic Operations, String slices, Working with Lists, Accessing list, Operations, Function and Methods, Files, Pickling, Modules, Dictionaries, Dictionary Comprehension, Functions and Functional Programming, Declaring and calling Functions, Declare, assign and retrieve values from Lists, Introducing Tuples, Accessing tuples, Visualizing using Matplotlib, Seaborn, OOPs concept, Class and object, Attributes, Inheritance, Overloading, Overriding, Data hiding, Generators, Decorators, Operations Exception, Exception Handling, except clause, Tryfinally clause, User Defined Exceptions, Data wrangling, Data cleaning, Load images and audio files using python libraries(pillow/scikit-learn), Creation of python virtual environment, Logging in Python.

**R Programming:** Reading and Getting Data into R, Exporting Data from R, Data Objects- Data Types & Data Structure. Viewing Named Objects, Structure of Data Items, Manipulating and Processing Data in R (Creating, Accessing, sorting data frames,

Extracting, Combining, Merging, reshaping data frames), Control Structures, Functions in R (numeric, character, statistical), working with objects, Viewing Objects within Objects, Constructing Data Objects, Packages – Tidyverse, Dplyr, Tidy etc., Queuing Theory, Case Study.

### **Java Programming - 70 Hours**

Introduction to Java Virtual Machine, Data Types, Operators and Language, OOPs Concepts, Constructs, Inner Classes and Inheritance, Interface and Package, Exceptions, Collections, Threads, Java.lang,, Java.util, Lambda Expressions, Introduction to Streams, Introduction of JDBC API.

### **Advanced Analytics using Statistics - 90 Hours**

Introduction to Business Analytics using some case studies, Summary Statistics, Making Right Business Decisions based on data, Statistical Concepts, Descriptive Statistics and its measures, Probability theory, Probability Distributions (Continuous and discrete- Normal, Binomial and Poisson distribution) and Data, Sampling and Estimation, Statistical Interfaces, Predictive modelling and analysis, Bayes' Theorem, Central Limit theorem, Statistical Inference Terminology (types of errors, tails of test, confidence intervals etc.), Hypothesis Testing, Parametric Tests: ANOVA, t-test, Non parametric Tests- chiSquare, U-Test Data Exploration & preparation, Concepts of Correlation, Covariance, Outliers, Simulation and Risk Analysis, Optimization, Linear, Integer, Overview of Factor Analysis, Directional Data Analytics, Functional Data Analysis, Predictive Modelling (From Correlation To Supervised Segmentation): Identifying Informative Attributes, Segmenting Data By Progressive Attributive, Models, Induction And Prediction, Supervised Segmentation, Visualizing Segmentations, Trees As Set Of Rules, Probability Estimation; Decision Analytics: Evaluating Classifiers, Analytical Framework, Evaluation, Baseline, Performance And Implications For Investments In Data; Evidence And Probabilities: Explicit Evidence Combination With Bayes Rule, Probabilistic Reasoning; Business Strategy: Achieving Competitive Advantages, Sustaining Competitive Advantages

**Python Libraries** – Pandas, Numpy, Scrapy, Plotly, Beautiful soup

### **Data Collection and DBMS (Principles, Tools & Platforms) - 90 Hours**

Database Concepts (File System and DBMS), OLAP vs OLTP, Database Storage Structures (Tablespace, Control files, Data files), Structured and Unstructured data, SQL Commands

(DDL, DML & DCL), Stored functions and procedures in SQL, Conditional Constructs in SQL, data collection, Designing Database schema, Normal Forms and ER Diagram, Relational Database modelling, Stored Procedures, Triggers. The tools and how data can be gathered in a systematic fashion, Data ware Housing concept, No-SQL, Data Models - XML, working with MongoDB, Cassandra- overview, comparison with MongoDB, working with Cassandra, Connecting DB's with Python, Introduction to Data Driven Decisions, Enterprise Data Management, data preparation and cleaning techniques. Understanding Data Lakes – concepts, architecture and components, Data Lake vs. Data Warehouse vs. Lakehouse, data storage management, processing and transformation, workflow orchestration, analytics in Data Lake, case study using Delta Lake with analytics and AI.

### **Big Data Technologies - 150 Hours**

**Introduction to Big Data-Big Data** - Beyond The Hype, Big Data Skills And Sources Of Big Data, Big Data Adoption, Research And Changing Nature Of Data Repositories, Data Sharing And Reuse Practices And Their Implications For Repository Data Curation,

**Hadoop:** Introduction of Big Data Programming-Hadoop, The ecosystem and stack, The Hadoop Distributed File System (HDFS), Components of Hadoop, Design of HDFS, Java interfaces to HDFS, Architecture overview, Development Environment, Hadoop distribution and basic commands, Eclipse development, The HDFS command line and web interfaces, The HDFS Java API, Analyzing the Data with Hadoop, Scaling Out, Hadoop event stream processing, complex event processing, MapReduce Introduction, Developing a Map Reduce Application, How Map Reduce Works, The MapReduce Anatomy of a Map Reduce Job run, Failures, Job Scheduling, Shuffle and Sort, Task execution, Map Reduce Types and Formats, Map Reduce Features, Real-World MapReduce,

**Hadoop Environment:** Setting up a Hadoop Cluster, Cluster specification, Cluster Setup and Installation, Hadoop Configuration, Security in Hadoop, Administering Hadoop, HDFS – Monitoring & Maintenance, Hadoop benchmarks,

**Apache Airflow/ETL Informatica:** Introduction to Data warehousing and Data lakes, Designing Data warehousing for an ETL Data Pipeline, Designing Data Lakes for ETL Data Pipeline, ETL vs ELT

**Introduction to HIVE:** Programming with Hive: Data warehouse system for Hadoop, Optimizing with Combiners and Practitioners, Bucketing, more common algorithms: sorting, indexing and searching, Relational manipulation: map-side and reduce-side joins, evolution, purpose and use, Case Studies on Ingestion and warehousing

**HBase:** Overview, comparison and architecture, java client API, CRUD operations and security

**Apache Spark:** Overview, APIs for large-scale data processing, Linking with Spark, Initializing Spark, Resilient Distributed Datasets (RDDs), External Datasets, RDD Operations, Passing

Functions to Spark, Job optimization, Working with Key-Value Pairs, Shuffle operations, RDD Persistence, Removing Data, Shared Variables, EDA using PySpark, Deploying to a Cluster Spark Streaming, Spark MLlib and ML APIs, Spark Data Frames/Spark SQL, Integration of Spark and Kafka, Setting up Kafka Producer and Consumer, Kafka Connect API, Map reduce, Connecting DB's with Spark

### **Data Visualization - Analysis and Reporting - 50 Hours**

Business Intelligence- requirements, content and managements, information Visualization, Data analytics Life Cycle, Analytic Processes and Tools, Analysis vs. Reporting, MS Excel: Functions, Formula, charts, Pivots and Lookups, Data Analysis Tool pack: Descriptive Summaries, Correlation, Regression, Introduction to Tableau, Data sources in Tableau, Taxonomy of data visualization, Numeric, String, Date Calculations, LOD (Level of Detail) Expressions, Modern Data Analytic Tools, Visualization Techniques.

### **Practical Machine Learning - 140 Hours**

#### **Machine Learning:**

Introduction to machine learning, Formal learning model – PAC learning, Bias complexity trade off, The VC Dimension, Nonuniform learnability (Structural risk minimization and Occam's Razor and No Free Lunch Theorem), Regularization and Stability, Model Selection and Validation, Machine Learning taxonomy – Supervised, Unsupervised and Semi-supervised Learning, practical use cases of Machine learning, Unsupervised Learning – Clustering (K-Means and its variants), Hierarchical Clustering, Dimension Reduction (PCA, Kernel PCA, LDA, Random Projections), Fundamentals of information theory, Supervised Learning with simple and ensemble learning – Classification and Regression (KNN, Decision Trees, Bayesian analysis and Naïve Bayes classifier, Random forest, Gradient boosting Machines, SVM, XGBoost, CatBoost, Linear and Non-linear regression), Time series Forecasting.

#### **Deep Learning:**

Introduction to neural networks (Neurons, construction of networks, backpropagation),

Introducing Modern Practical Deep Networks (Deep Feedforward Networks, Regularization for Deep Learning, Optimization for Training Deep Models), Convolutional Neural Networks, Sequence modelling using recurrent neural networks, Transfer Learning, Autoencoders, Object Detection, Object Segmentation and Tracking, Concepts of NLP.

**Generative AI:**

Introduction to transformers, Difference between encoder, decoder and encoder-decoder architectures, Attention Mechanisms, Overview of BERT, Application of transformers, Introduction to LARGE LANGUAGE MODELS, Understanding and handling TEXT DATA, Understand the concept of fine-tuning pre-trained model, Reward Models and Alignment Strategies, Practical case studies using SLMs and LLMs, Deployment of LLMs.

**Aptitude & Effective Communication - 90 Hours**

**Aptitude:** Percentage, Profit & Loss, Ratio & Proportion, Average, Mixture & Allegation, Simple Interest & Compound Interest, Number Systems, Series, Cyclicity & Remainders, Data Interpretation, Syllogism, Coding & Decoding, Blood Relations, Seating Arrangements (Linear & Circular), Ages, Puzzles, Time, Speed & Distance, Trains, Boats & Streams, Time & Work, Wages (Man days), Pipes & Cisterns, Clocks, Permutations & Combinations, Probability, Calendar.

**Effective Communication:** Fundamentals of Communication, The Art of Communication, Personality Development, English Grammar, Correct Usage of English, Common Mistakes in English Communication, Listening Skills, Reading Skills, Writing Skills, Public Speaking, Presentation Skills, Group Discussions, Interpersonal Skills, Personal Interviews

**Project - 90 Hours**

Clustering and filtering approach in big data using Machine Learning Models, Energy efficient in big data gathering, Dynamic Big Data Storage on Cloud & Fine-Grained Updates.