Annexure - I
PG-Diploma in Big Data Analytics (PG-DBDA) Aug 16

Course Modules:

Sr. No.	Module Name	Hours
1.	Statistical Analysis with R	100
2.	Programming with Python	50
3.	Fundamentals of Linux Programming	40
4.	Java with Scala	80
5.	Cloud Computing & Operations	30
6.	Data Collection and DBMS (Principles, Tools & Platforms)	80
7.	Big Data Technologies	130
8	Data Visualization - Analysis and Reporting	40
9.	Business Decisions and Analytics	50
10.	High Performance Computing Solution & Applications	20
11.	Practical Machine Learning	60
12.	Aptitude	50
13.	Effective Communication	50
14.	Project	120
	Total	**900**

Course Contents:

Module 1: Statistical Analysis with R (100 Hours)

Probability & Statistics: Introduction to Statistics- Descriptive Statistics, Summary Statistics Basic probability theory, Statistical Concepts (uni-variate and bi-variate sampling, distributions, re-sampling, statistical Inference, prediction error), Probability Distribution (Continuous and discrete- Normal, Bernoulli, Binomial, Negative Binomial, Geometric and Poisson distribution) , Bayes' Theorem, Central Limit theorem, Data Exploration & preparation, Concepts of Correlation, Regression, Covariance, Outliers etc.

R Programming: Introduction & Installation of R, R Basics, Finding Help, Code Editors for R, Command Packages, Manipulating and Processing Data in R, 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, Building R Packages, Running and Manipulating Packages, Non parametric Tests- ANOVA, chi-Square, t-Test, U-Test, Introduction to Graphical Analysis, Using Plots(Box Plots, Scatter plot, Pie Charts, Bar charts, Line Chart), Plotting variables, Designing Special Plots, Simple Liner Regression, Multiple Regression

Module 2: Programming with Python (50 Hours)

Introduction to Python, Basic Syntax, Data Types, Variables, Operators, Input/output, Flow of Control (Modules, Branching), 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, Introduction, Accessing list, Operations, Function and Methods, Files, Modules, Dictionaries, Functions and Functional Programming, Declaring and calling Functions, Declare, assign and retrieve values from Lists, Introducing Tuples, Accessing tuples

Advanced Python: Object Oriented, OOPs concept, Class and object, Attributes, Inheritance, Overloading, Overriding, Data hiding, Operations Exception, Exception Handling, Except clause, Try finally clause, User Defined Exceptions

Python Libraries

Introduction to Machine learning packages like NUMPY, SCIPY, PANDAS etc.

Module 3: Fundamentals of Linux Programming (40 Hours)

Linux History and Operation, Installing and Configuring Linux, Shells, Commands, and Navigation, Common Text Editors, Administering Linux, Introduction to Users and Groups, Linux shell scripting

Module 4: Java with Scala (80 Hours)

Data Types, Operators and Language, Constructs, Inner Classes and Inheritance, Interface and Package, Exceptions, Threads

Introduction, Unified Types, Classes, Traits, Mixin Class Composition, Anonymous Function Syntax, Higher-order Functions, Nested Functions, Currying, Case Classes, Pattern Matching, Singleton Objects, XML Processing, Regular Expression Patterns, Extractor Objects, Sequence Comprehensions, Generic Classes

Module 5: Cloud Computing & Operations (30 Hours)

Introduction to Cloud Computing: Definition, Characteristics, Components, Cloud provider, SAAS, PAAS, IAAS and other Organizational scenarios of clouds, Administering & Monitoring cloud services, benefits and limitations, Deploy application over cloud. Comparison among SAAS, PAAS, IAAS, Cloud computing platforms: Infrastructure as service: Amazon EC2, Platform as Service: Google App Engine, Microsoft Azure Utility Computing, Elastic Computing, SLA, clusters, cloud analytics, challenges of cloud environment, HPC in the cloud

Module 6: Data Collection and DBMS (Principles, Tools & Platforms) (80 Hours)

Database Concepts (File System and DBMS), Database Storage Structures (Tablespace, Control files, Data files), Structured and Unstructured data, SQL Commands (DDL, DML & DCL), Dataware Housing concept , No-SQL, Data Models - XML, working with MongoDB),

Tools - OLTP and OLAP, data preparation and cleaning techniques

Module 7: Big Data Technologies (130 Hours)

Introduction to Big Data- Big data definition, enterprise / structured data, social / unstructured data, unstructured data needs for analytics, What is Big Data, Big Deal about Big Data, Big Data Sources, Industries using Big Data, Big Data challenges.

Hadoop: Introduction of Big data programming-Hadoop, History of 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 (lab), 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 ETL: Hadoop ETL Development, ETL Process in Hadoop, Discussion of ETL functions, Data Extractions, Need of ETL tools, Advantages of ETL tools.

Hadoop Reporting Tools: Jaspersoft (reporting and analytics server), Pentaho (data integration and business analytics), Splunk (platform for IT analytics), Talend (big data integration, data management and application integration)

Introduction to Pig and HIVE- Programming Pig: Engine for executing data flows in parallel on Hadoop, Programming with Hive: Data warehouse system for Hadoop, Optimizing with Combiners and Partitioners (lab), More common algorithms: sorting, indexing and searching (lab), Relational manipulation: map-side and reduce-side joins (lab), evolution, purpose and use, HDFS – Overview and concepts, data flow (read and write), interface to HDFS (HTTP, CLI and Java API), high availability and Name Node federation, Map Reduce developing and deploying programs, optimization techniques, Map Reduce Anatomy, Data flow framework programming Map Reduce best practices and debugging, Introduction to Hadoop ecosystem, integration R with Hadoop

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, Hadoop in the cloud.

Introduction to Apache Spark and Use Cases

Apache Spark APIs for large-scale data processing: 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

Apache Phoenix: Apache Phoenix Overview, Need of Phoenix, Features, Installation and Configurations, Views and Multi Tenancy, Secondary Indexes, Joins, Query Optimizations, Roadmap of Phoenix.

Module 8: Data Visualization - Analysis and Reporting (40 Hours)

Information Visualization, Data analytics Life Cycle, Analytic Processes and Tools, Analysis vs. Reporting, Modern Data Analytic Tools, Visualization Techniques, Visual Encodings, Visualization algorithms, Data collection and binding, Cognitive issues, Interactive visualization,

Visualizing big data – structured vs unstructured, Visual Analytics, Geomapping, Dashboard Design,

Module 9: Business Decisions and Analytics (50 Hours)
Introduction to Business Analytics using some case studies, Making Right Business Decisions based on data, Exploratory Data Analysis - Visualization and Exploring Data, Descriptive Statistical Measures, Probability Distribution and Data, Sampling and Estimation, Statistical Interfaces, Predictive modeling and analysis, Regression Analysis, Forecasting Techniques, Simulation and Risk Analysis, Optimization, Linear, Non linear, Integer, Decision Analysis, Strategy and Analytics
Overview of Factor Analysis, Directional Data Analytics, Functional Data Analysis

Module 10: High Performance Computing Solution & Applications (20 Hours)
Parallel Processing Concepts: Physical Organization and building blocks of High Performance Computing Systems, Processors and Multi-Core Architectures, Vector processing, Super-scalar, In-order execution, Instruction-Level Parallelism etc., FMA, 32 and 64 bit types, ISA, Accelerators such as GPGPUs and Xeon Phi. Threads and Processes, Multi-processing OS, Parallel I/O, General concepts
Parallel Programming Models and Parallel Algorithms Design: Application domains of HPC, Decomposition Techniques: Data parallelism, Functional parallelism, Divide and Conquer etc., Characteristics of Tasks and Interactions, Mapping Techniques for Load Balancing, Methods for Containing Interaction Overheads, Granularity of parallelism, Programming OpenMP
Application porting, execution and scalability analysis: Compiler flags, vectorization, memory alignment of data, Porting of scientific & engineering parallel applications on Linux, Measurement of Application execution time and memory consumption with small, medium and large datasets, Performance metrics and Scalability analysis of codes, Amdahl's law, identification of performance bottlenecks, Profiling of applications to find opportunities for performance optimization, Use of existing mathematical libraries and performance analysis tools, **Optimizations:** Compiler optimization techniques, Explicit optimizations, Addition of directives, Restructuring of code for performance optimization, Communication optimization through configuration of MPI calls of the underlying MPI implementation

Module 11: Practical Machine Learning (60 Hours)
Supervised and Unsupervised Learning , Uses of Machine learning , Clustering, K means, Hierarchical Clustering, Decision Trees, Oblique trees, Classification problems, Bayesian analysis and Naïve bayes classifier, Random forest, Gradient boosting Machines, Association rules learning, Apriori and FP-growth algorithms, Support vector Machines, Linear and Non liner classification, ARIMA, ML in real time, Neural Networks and its application, Neural Net & its applications

Effective Communication and Aptitude & English (100 hours)
Project (120 hours)