Post Graduate Diploma in Big Data Analytics (PG-DBDA)

Course Focus

The theoretical and practical mix of the Post Graduate Diploma in Big Data Analytics (PG-DBDA) programme has the following focus:

- To explore the fundamental concepts of big data analytics with in-depth knowledge and understanding of the big data analytics domain
- To understand the various search methods and visualization techniques and to use various techniques for mining data stream
- To analyse and solve problems conceptually and practically from diverse industries, such as government manufacturing, retail, education, banking/ finance, healthcare and pharmaceutical
- To undertake consulting and industrial projects with significant data analysis component for better understanding of the theoretical concepts from statistics, economics and building future solutions data analytics to make an impact in the technological advancement
- To use advanced analytical tools/ decision-making tools/ operation research techniques to analyse the complex problems and get ready to develop such new techniques for the future
- To learn Cloud Computing, accessing resources and services needed to perform functions with dynamically changing needs

Course Outcome

- After completing this course students will be trained in statistics and machine learning using Python.
- They will make data driven decisions which provide them a competitive advantage in the market, technologies like Hadoop, Spark, Hive, Machine Learning provides a spring board for AI which makes them ready for Industry 4.0.
- At the end of the course students will be able to work as Data Analysts, Data Engineers.
- Studying Big Data will broaden their horizon by surpassing market forecast / predictions for Big Data Analytics

Course Contents

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

Cloud Computing

- Cloud Computing Basics
- Understanding Cloud Vendors (AWS:EC2 instance, lambda and Heroku: Heroku platform, Heroku Data services)
- Definition
- Characteristics
- 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 Products and Solutions
- Cloud Pricing
- Compute Products and Services
- Elastic Cloud Compute
- Dashboard

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
- Modules
- Dictionaries
- 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
- Operations Exception
- Exception Handling
- except clause
- Try-finally clause
- User Defined Exceptions
- Data wrangling
- Data cleaning
- Load images and audio files using python libraries(pillow/scikit-learn)
- Creation of python virtual environment

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, Tidyr, etc…
- Queuing Theory
- Case Study

Data Collection and DBMS (Principles, Tools & Platforms)

- 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, Window function, Case statements.
- 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 Architecture comparison with MongoDB
- working with Cassandra
- Connecting DB’s with Python
- Introduction to Data Driven Decisions
- Enterprise Data Management
- data preparation and cleaning techniques

Object Oriented Programming with Java 8:

- Oops Concepts
- Data Types
- Operators and Language
- Constructors
- Inner Classes and Inheritance
- Interface and Package
- Exceptions
- Collections
- Threads
- Java.lang
- Java.util
- Java Virtual Machine
- Reflection in JVM
- JVM’s architecture,
- Lambda Expressions
- Functional Programming and Interfaces
- Introduction to Streams
- Introduction of JDBC API
-

Big Data Technologies:

1.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

2.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 (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

3.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

4.Apache Airflow

- 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

5.Introduction to HIVE

- Programming with Hive: Data warehouse system for Hadoop
- Optimizing with Combiners and Practitioners (lab)
- Bucketing, more common algorithms: sorting
- indexing and searching (lab)
- Relational manipulation: map-side and reduce-side joins (lab)
- Evolution
- purpose and use
- Case Studies on Ingestion and warehousing

6.HBase

- Overview
- comparison and architecture
- java client API
- CRUD operations and security

7.Apache Spark

- APIs for large-scale data processing: Overview
- Linking with Spark
- Initializing Spark
- Resilient Distributed Datasets (RDDs)
- External Datasets
- RDD Operations
- 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
- Mapreduce
- Connecting DB’s with Spark

Advanced Analytics using Statistics

- 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)
- Data Sampling and Estimation
- Statistical Interfaces
- Predictive modeling 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- chi-Square, 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

Practical Machine Learning

- Supervised and Unsupervised Learning
- Uses of Machine learning
- Clustering
- K means
- Hierarchical Clustering
- Decision Trees
- Classification problems
- Bayesian analysis and Naïve Bayes classifier
- Random forest
- Gradient boosting Machines
- Association rules learning
- PCA
- Apriori
- Support vector Machines
- Linear and Non liner classification
- ARIMA
- XG Boost
- CAT Boost
- Neural Networks and its application
- Tensorflow 2.x framework
- Deep learning algorithms
- KNN
- NLP
- Bert in NLP
- NLP transformers
- NLTK
- Introduction to Pytorch framework
- AI and its application

Data Visualization - Analysis and Reporting

- 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 Power BI
- Modern Data Analytic Tools
- Visualization Techniques.

Q. What is the value of the course in the international market?

A. The course has been a trend-setting course due to its unique curriculum and the opportunities that it generates; hence it gives the edge over for the students and gives an international edge.