**NIELIT, CHENNAI**
**ISTE COMPLEX, 25, GANDHI MANDAPAM ROAD, CHENNAI**

# Course Syllabus

**Name of the Group:** Data Science

**Course Name:** PG Diploma in Data Science & Analytics

**Course Code:** DS 500

**Start Date:** 22-10-2018

**Duration:** 880 Hours**,** 24 Weeks (8 Hours per day)

## Course Structure

This course contains total six modules. After completing the first five modules, the students have to do a 120 Hours project using any of the topics studied to earn the PG Diploma.

| DS 500 | Module Name | Duration (in Hours) |
|---|---|---|
| DS 501 | System Administration under Linux Operating System & Advance Shell Programming | 120 |
| DS 502 | Data Analytics using R | 120 |
| DS 503 | Data Storage Technique & Data Warehousing using MySQL | 120 |
| DS 504 | Object Oriented Programming : Java & Python | 200 |
| DS 505 | Big Data Technology using Hadoop and Spark | 200 |
| DS 506 | Mini Project (Implementation of Data Analytics) | 120 |
| **Total Duration** | | 880 |

## Modularization

DS 501**:** System Administration under Linux Operating System & Advance Shell Programming

**Module Objective**

This module makes the participant completely conversant in Linux System Administration and Shell Programming. The course is an in-depth coverage on Linux system fundamentals (the essentials of Linux) as well as advanced administration including monitoring and troubleshooting. It starts with Linux environment and then jumps to Advance Bash Shell scripting/programming which is an essential component of Linux Operating System. The course will be focusing primarily on CLI commands as opposed to GUIs so that the participant will have a significantly high learning curve.

**Module Duration**: 120 Hours

**Pre-Requisite:** BE/B.Tech/MCA/M.Sc.(CS/IT)/DOEACC B Level/Master Degree in Statistics with Knowledge of Statistics and Computer Programming

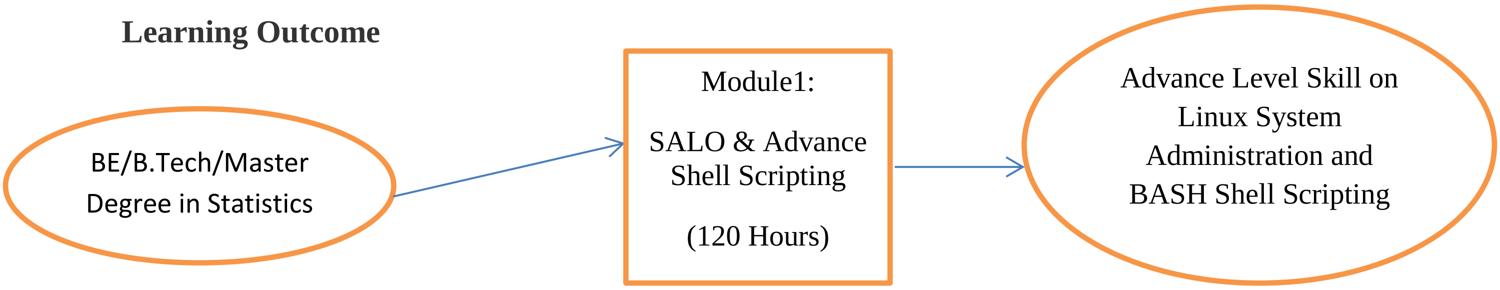
## DS 501 Syllabus

| Section | Topics to be covered | Duration<br>( In Hours) |
|---|---|---|
| DS 501.1 | Installation and Initialization | **04** |
| DS 501.2 | Basic Linux Commands | **04** |
| DS 501.3 | Package Management and process Monitoring | **08** |
| DS 501.4 | Important Files, Directories and Utilities | **04** |
| DS 501.5 | Advance Shell Programming | **44** |
| DS 501.6 | System Services | **08** |
| DS 501.7 | User Administration | **08** |
| DS 501.8 | File System Security & Advanced File System Management | **16** |
| DS 501.9 | Server Configuration & Virtualization | **08** |
| DS 501.10 | Samba and Mail Services Virtualization | **08** |
| DS 501.11 | Advance Security & Networking Concepts | **08** |
| **Total Duration** | | **120** |

**Tools to be used**

1. Ubuntu Operating System
2. VMware

**Learning Outcome**



BE/B.Tech/Master Degree in Statistics

Module1:

SALO & Advance Shell Scripting

(120 Hours)

Advance Level Skill on Linux System Administration and BASH Shell Scripting

Upon successful completion of this module, the student will have the ability to:

- Comprehend Ubuntu Linux & Install Ubuntu Linux
- Comprehend Basic Linux Commands
- Comprehend Software Management
- Comprehend complete file system architecture of Linux
- Comprehend advanced level skills to build the requisite expertise through shell scripting to manage, operate and maintain an enterprise network using Linux/Unix.
- Comprehend the Linux daemons and other processes.
- Comprehend User Administration.
- Comprehend File System Security and Management
- Comprehend Virtualization
- Comprehend Samba and Mail Services
- Comprehend Network Security Management& Remote Administration

**Recommended Books**

**Text Books**

1. Linux Shell Scripting Cookbook by Sarath
2. Lakshman Linux System Administration by Roderick W Smith, Vicki Stanfield Hunt Smith Stanfield

**Reference Books**

1. Shell Scripting: Expert Recipes for Linux, Bash, and more by Steve Parker
2. Linux System Administrator's Guide Version by Lars Wirzenius
3. Linux Bible by Christopher Negus
4. Effective AWK Programming: Universal Text Processing and Pattern Matching by O' Reilly
5. Mastering Unix Shell programming by Randal K Michael
6. Shell Scripting: Expert Recipes for Linux, Bash, and More by Steve Parker

**DS 502:** Data Analytics using R

**Module Objective**

This module makes the participant conversant with the concept of Data Science and techniques to be used for data analytics including the construction of different statistical Models used for Data Analytics. The module is an in-depth coverage on various Statistical Techniques and goodness of fit tests used for data analytics. The module is practical oriented. For Analysis R software is used. What makes this course unique is that participant will continuously practice their newly acquired skills through R Studio. In the final section, participant will dive deeper into the graphical capabilities of R, and create their own stunning data visualizations.

**Module Duration**: 120 Hours

**Pre-Requisite:** BE/B.Tech/MCA/M.Sc/DOEACC B Level with Knowledge of Statistics and Computer Programming.

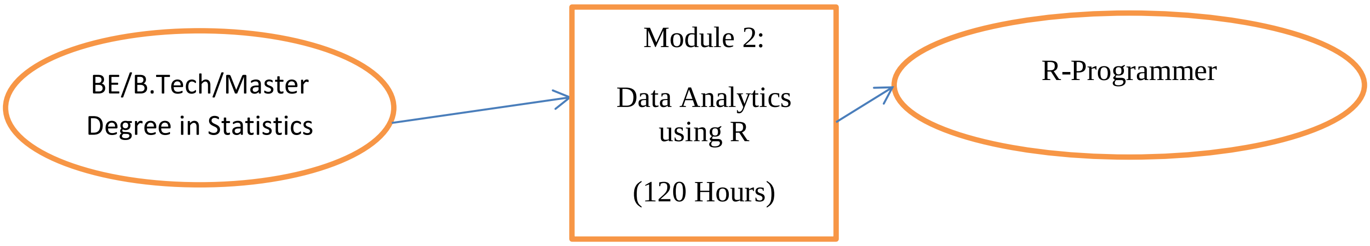
**DS 502 Syllabus**

| Section | Topics to be covered | Duration<br>( In Hours) |
|---|---|---|
| **DS 502**.1 | Concept of Data Analytics & R | **08** |
| **DS 502**.2 | Data Manipulation in R | **08** |
| **DS 502**.3 | Data Import Techniques | **04** |
| **DS 502**.4 | Exploratory Data Analysis | **12** |
| **DS 502**.5 | Data Visualization | **16** |
| **DS 502**.6 | Data Mining: Clustering Technique | **20** |
| **DS 502**.7 | Data Mining: Association Rule Mining and Sentiment Analysis | **12** |
| **DS 502**.8 | Regression | **08** |
| **DS 502**.9 | Anova | **04** |
| **DS 502**.10 | Predictive Analysis & Simulation | **12** |
| **DS 502**.11 | Implementation of Decision tree | **16** |
| **Total Duration** | | **120** |

**Tools to be used**

1. Ubuntu Operating System
2. VMware

**Learning Outcome**



Upon successful completion of this module, the student will have the ability to:

- Learn Data Science concepts of R and functioning of R
- Understand Exploratory Data Analytics
- Learn to create various graphics
- Understand Data Mining
- Learn Regression Analysis
- Fit a Statistical Model
- Learn Predictive Analysis
- Implement Decision Tree

**Recommended Books**

**Text Books**

1. R for Data Analysis in Easy Steps by Mike Mc Grath
2. Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist by Thomas Mailund

**Reference Books**

1. Advanced R: Data Programming and the Cloud by by**:** Matt Wiley,Joshua F. Wiley
2. Statistical Analysis with R For Dummies by**:** Joseph Schmuller

**DS 503: Data Storage–** Data Storage technique & Data warehousing using MySQL

**Module Objective**

This module makes the participant conversant with the concept of Data Storage and techniques to be used for fetching data from database. Participants will learn exciting concepts and skills for designing data warehouses and creating data integration workflows. Participants will have hands-on experience for data warehouse design and use open source products for manipulating pivot tables and creating data integration workflows. After successful completion of the module participant will be able to perform various activities of data-warehousing using MySQL.

**Module Duration**: 120 Hours

**Pre-Requisite:** BE/B.Tech/MCA/M.Sc.(CS/IT)/DOEACC B Level/Master Degree in Statistics with good knowledge of computer.

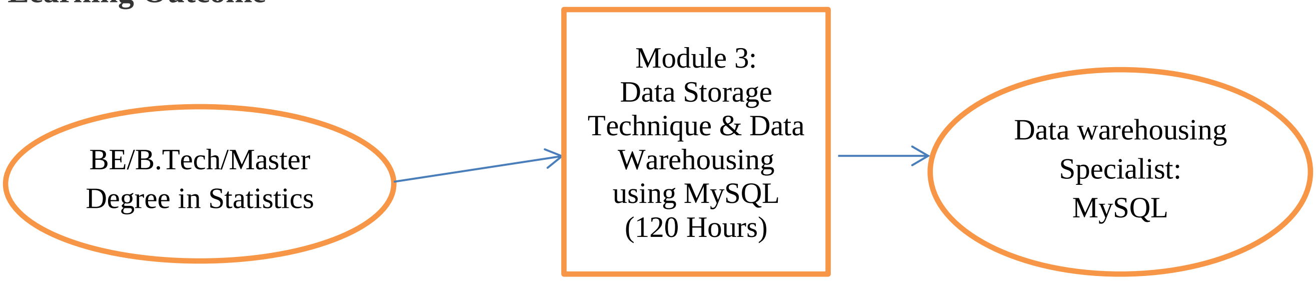
**DS 503: Syllabus**

| **Section** | **Topics to be covered** | **Duration ( In Hours)** |
|---|---|---|
| **DS 503**.1 | Introductory Concepts | **04** |
| **DS 503**.2 | Database Design | **08** |
| **DS 503**.3 | Relational Model and SQL | **08** |
| **DS 503**.4 | Database design using the relational model | **12** |
| **DS 503**.5 | Storage and Indexing Structures | **12** |
| **DS 503**.6 | Transaction Processing and Concurrency Control (OLTP & OLAP) | **20** |
| **DS 503**.7 | Database recovery techniques | **12** |
| **DS 503**.8 | Query Processing and Optimization | **08** |
| **DS 503**.9 | Database Security and Authorization | **04** |
| **DS 503**.10 | Enhanced Data Models for specific applications | **12** |
| **DS 503**.11 | Enhanced Data Models for specific applications | **16** |
| **DS 503**.12 | Distributed databases and issues | **08** |
| **Total Duration** | | **120** |

## Tools to be used

1. Ubuntu Operating System
2. VMware
3. MySQL

## Learning Outcome


BE/B.Tech/Master Degree in Statistics
Module 3: Data Storage Technique & Data Warehousing using MySQL (120 Hours)
Data warehousing Specialist: MySQL

Upon successful completion of this module, the student will have the ability to:

- Design a Database
- Understand Database Relational Models
- Learn to design and execute various SQL and Store Procedures
- Understand OLAP and OLTP
- Learn Data Models
- Understand Distributed Systems

## Recommended Books

### Text Books

1. SQL for MySQL: A Beginner's Tutorial by Bjoni Darmawikarta
2. Open Source Data Warehousing and Business Intelligence by Lakshman Bulusu

### Reference Books

1. Agile Data Warehousing for the Enterprise: A Guide for Solution Architects and Project Leaders by Ralph Hughes
2. Data Warehousing in the Age of Big Data by Krish Krishnan

**DS 504: Programming for Data Science –** Basic Java & Python Programming for Data Science

**Module Objective**

This module is specially designed for improving basic concepts of Java. This module makes the participant conversant with the concept of Java to be used in Hadoop and Advance Python programming for Data Science. After successful completion of the module participants will be capable of understanding the concepts used in Map Reduce , Pig Hive etc. Participants will learn exciting concepts and skills for advance analysis using Python.

**Module Duration**: 200 Hours

**Pre-Requisite:** BE/B.Tech (CS/IT/ECE)/MCA/M.Sc.(CS/IT)/DOEACC B Level/Master Degree in Statistics with good knowledge of computer.

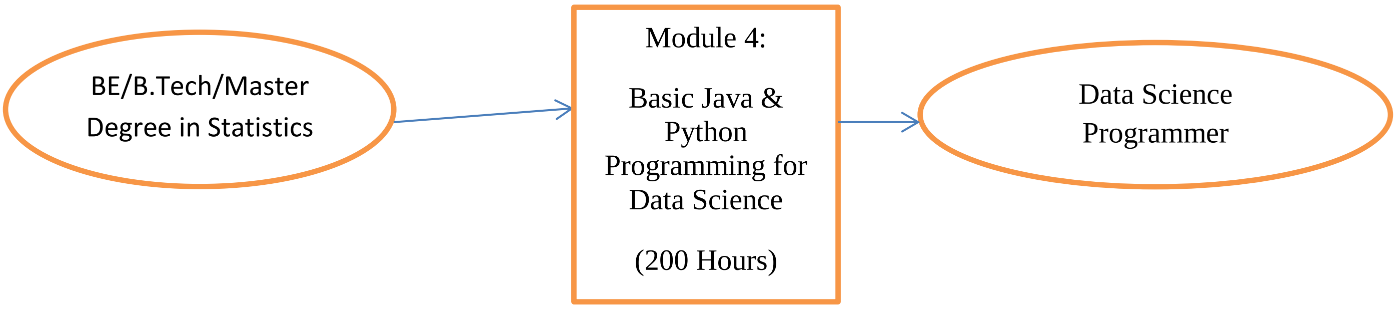
**DS 504: Syllabus**

| **Section** | **Topics to be covered** | **Duration ( In Hours)** |
|---|---|---|
| **DS 504**.1 | Basic Java | **04** |
| **DS 504**.2 | Arrays, Objects and Classes | **08** |
| **DS 504**.3 | Control Flow Statements | **08** |
| **DS 504**.4 | Inheritance and Interfaces | **08** |
| **DS 504**.5 | Exception Handling & Serialization | **12** |
| **DS 504**.6 | Collections | **20** |
| **DS 504**.7 | Reading and Writing files | **20** |
| **DS 504**.8 | Python Basics | **20** |
| **DS 504**.9 | OOPs concept in Python | **12** |
| **DS 504**.10 | Exception Handling in Python | **08** |
| **DS 504**.11 | Python for Data Science an Introduction | **20** |
| **DS 504**.12 | Pre Processing of Data | **08** |
| **DS 504**.13 | Visualising the Data | **08** |
| **DS 504**.14 | Exploratory Data Analysis, Clustering and identification of Outliers using Python | **12** |
| **DS 504**.15 | Performing Cross-Validation, Selection, and Optimization using Python | **20** |
| **DS 504**.16 | Learning from Data using Python | **12** |
| **Total Duration** | | **200** |

**Tools to be used**

1. Ubuntu Operating System
2. VMware
3. MySQL
4. Python

**Learning Outcome**


BE/B.Tech/Master Degree in Statistics
Module 4:
Basic Java & Python Programming for Data Science
(200 Hours)
Data Science Programmer

Upon successful completion of this module, the student will have the ability to:

- Understand the basic concepts of Java.
- Understand Python Programming.
- Learn Exploratory Data Analysis, Clustering and identification of Outliers using Python.

**Recommended Books**

**Text Books**

1. Pro Java  Programming by Brett Spell
2. Python for Data Science For Dummies by Luca Massaron, John Paul Mueller

**Reference Books**

1. Exploring Java : Build Modularized Applications in Java by Fu Cheng
2. Learn to Program with Python by Irv Kalb
3. Fundamentals of Python: Data Structures by Kenneth A. Lambert
4. Professional Python by **:** Luke Sneeringer

**DS 505:** Big Data Technology using Hadoop and Spark

**Module Objective**

This Module is proposed to give participant all around learning of the Big Data framework using Hadoop and Spark, including YARN, HDFS and Map Reduce. Participant will be able to learn how to use Pig, Hive etc. to practice and examine tremendous datasets stored in the HDFS and use various tools for data ingestion. After completion of the module participant will have complete knowledge of Data Analytics.

**Module Duration**: 200 Hours

**Pre-Requisite:** BE/B.Tech (CS/IT/ECE)/MCA/M.Sc.(CS/IT)/DOEACC B Level/Master Degree in Statistics with good knowledge of computer having good knowledge of Java and Python.

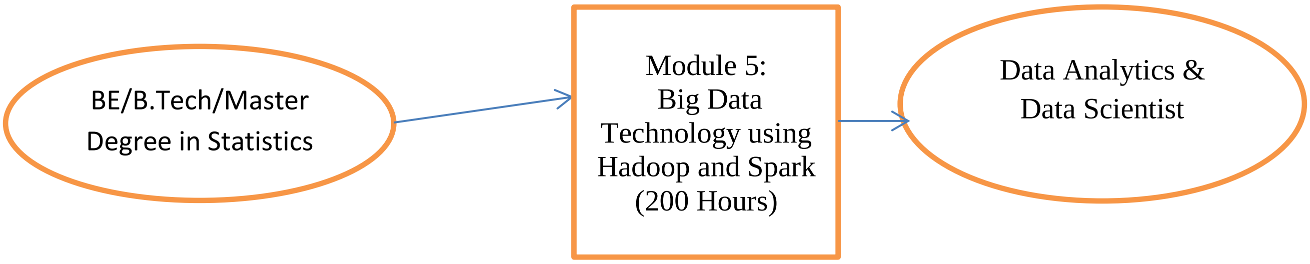
**DS 505: Syllabus**

| **Section** | **Topics to be covered** | **Duration ( In Hours)** |
|---|---|---|
| **DS 505**.1 | Introduction to Big Data and Hadoop Eco System | **04** |
| **DS 505**.2 | Hadoop: Eco System | **16** |
| **DS 505**.3 | HDFS Architecture | **20** |
| **DS 505**.4 | HDFS JAVA API | **20** |
| **DS 505**.5 | Map reduce | **20** |
| **DS 505**.6 | Hadoop ETL | **20** |
| **DS 505**.7 | Hadoop Reporting Tools | **20** |
| **DS 505**.8 | Hadoop Environment: Setting up Hadoop Cluster and HDFS Monitoring | **20** |
| **DS 505**.9 | Pig and HIVE | **20** |
| **DS 505**.10 | Apache Spark | **20** |
| **DS 505**.11 | Apache Spark API | **20** |
| **Total Duration** | | **200** |

**Tools to be used**

1. Ubuntu Operating System
2. VMware
3. MySQL
4. Python
5. Hadoop

**Learning Outcome**


BE/B.Tech/Master Degree in Statistics
Module 5: Big Data Technology using Hadoop and Spark (200 Hours)
Data Analytics & Data Scientist

Upon successful completion of this module, the student will have the ability to:

- Understand the various parts of Hadoop
- Learn Hadoop Distributed File System (HDFS) and YARN building, and make sense of how to function with them for limit and resource organization
- Understand MapReduce and its qualities and retain advanced MapReduce thoughts
- Ingest data using Sqoop and Flume
- Get a working learning of Pig and its parts
- Do functional programming in Spark, and execute and create Spark applications
- Make database and tables in Hive .
- Grasp and work with HBase, its outline and data accumulating, and take in the difference among HBase and RDBMS
- Understand the typical use occasions of Spark and distinctive natural estimations
- Learn Spark SQL, making, changing, and addressing data diagrams

**Recommended Books**

**Text Books**

1. Hadoop for Dummies by  Dirk deRoos,et al.
2. Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools by Deepak Vohra

**Reference Books**

1. Big Data and Hadoop: Learn by Example by Mayank Bhushan

**DS 506:** Mini Project (Implementation of Data Analytics)

**Module Objective**

The main objective of this module is for development of a mini project by implementing all Data Analytics Concepts.

**Module Duration**: 120 Hours

**Pre-Requisite:** BE/B.Tech (CS/IT/ECE)/MCA/M.Sc. (CS/IT)/DOEACC B Level/Master Degree in Statistics with good knowledge of Data Science.