रा.इ.सू.प्रौ.सं NIELIT | राष्ट्रीय इलेक्ट्रॉनिकी एवं सूचना प्रौद्योगिकी संस्थान, चेन्नई
National Institute of Electronics & Information Technology, Chennai

Ministry of Electronics & Information Technology
Government of India

नेशनल इंस्टीट्यूट ऑफ इलेक्ट्रॉनिक्स एंड इंफॉर्मेशन टेक्नोलॉजी, चेन्नई

National Institute of Electronics and Information Technology, Chennai

Autonomous Scientific Society of Ministry of Electronics & Information Technology (MeitY), Govt. of India

ISTE Complex, 25, Gandhi Mandapam Road, Chennai - 600025

Syllabus

PG Diploma in Data Science

&

Analytics

Index

Topic **Page No.**

Course Syllabus

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	Basics of Linux Operating System & Cloud	120
DS 502	Data Warehousing using MySQL and MongoDB	120
DS 503	Data Analytics using R & Python	120
DS 504	Fundamentals of Java for Hadoop Framework	120
DS 505	Hadoop Eco System	240
DS 506	Mini Project (Implementation of Data Analytics)	120
Total Duration		840

Modularization

DS 501: Basics of Linux Operating System & Cloud

Module Objective

This module makes the participant completely conversant in Linux System and Shell Programming. The course is an in-depth coverage on Linux as well as basic concepts of virtualization and cloud. It starts with Linux environment and then jumps to Bash Shell scripting/programming which is an essential component of Linux Operating System. It also covers visualization technique, basics of Information Security & Cloud. 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: M.E./M.Tech/B.E./B.Tech/DOEACC B Level/Any Master Degree with Knowledge of Mathematics/Statistics and Computer Programming.

DS 501 Syllabus

Module 501:

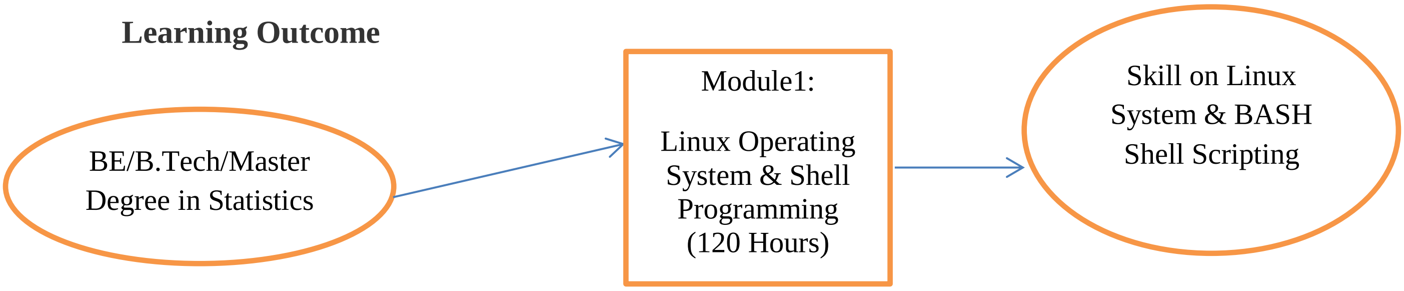
Basics of Linux Operating System & Cloud

Section	Topics to be covered	Theory	Practical	Total Duration (In Hours)
DS 501.1	Installation and Initialization	01	03	04
DS 501.2	Basic Linux Commands	01	03	04
DS 501.3	Package Management and process Monitoring	04	04	08
DS 501.4	Important Files, Directories and Utilities	01	03	04
DS 501.5	Shell Programming	22	26	48
DS 501.6	System Services	02	02	04
DS 501.7	User Administration	03	05	08
DS 501.8	Virtualization	08	08	16
DS 501.9	Basics of Information Security & Cloud	06	18	24
Total Duration		**48**	**72**	**120**

Tools to be used

1. Ubuntu Operating System
2. Virtual Box

Learning Outcome


BE/B.Tech/Master Degree in Statistics
Module1:
Linux Operating System & Shell Programming (120 Hours)
Skill on Linux System & 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 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 Virtualization
- Comprehend Skills on Cloud and Information Security

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 Warehousing using MySQL and MongoDB

Module Objective

This module makes the participant conversant with the concept of Data Storage and techniques to be used for fetching data from database (structured and unstructured). 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 & MongoDB. They will be able to configure replica server and implement the concept of Sharding.

Module Duration: 120 Hours

Pre-Requisite: M.E./M.Tech/B.E./B.Tech/DOEACC B Level/Any Master Degree with Knowledge of Mathematics/Statistics and Computer Programming.

DS 502: Syllabus

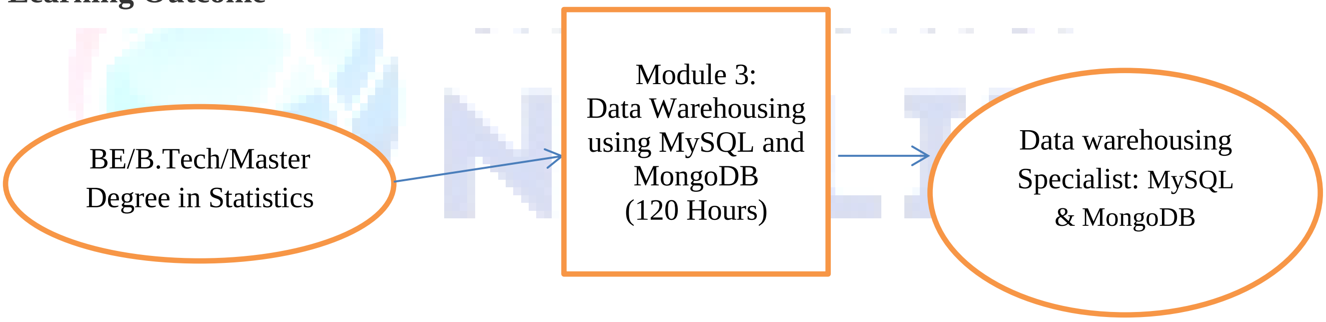
Module 502: Data Warehousing using MySQL and MongoDB				
Section	**Topics to be covered**	**Theory**	**Practical**	**Total Duration (In Hours)**
	MySQL			
DS 502.1	Database Design using MySQL	03	05	08
DS 502.2	Relational Model and SQL	03	05	08
DS 502.3	Database design using the relational model	03	05	08
DS 502.4	Transaction Processing and Concurrency Control in MySQL	06	10	16
	MongoDB			
DS 502.5	Introduction to NoSQL and MongoDB	02	02	04
DS 502.6	Creating, Updating, and Deleting Documents in MongoDB	08	12	20
DS 502.7	MongoDB Query	04	08	12

DS 502.8	Index, Special Index and Collection Types	04	04	08
DS 502.9	Aggregation	02	02	04
DS 502.10	Replication	04	08	12
DS 502.11	Connecting to a Replica Set from Your Application	02	02	04
DS 502.12	Sharding	02	02	04
DS 502.13	Backups	02	02	04
DS 502.14	Deploying MongoDB	04	04	08
Total Duration		**49**	**71**	**120**

Tools to be used

1. Ubuntu Operating System
2. Virtual box
3. MySQL
4. MongoDB

Learning Outcome



BE/B.Tech/Master Degree in Statistics

Module 3: Data Warehousing using MySQL and MongoDB (120 Hours)

Data warehousing Specialist: MySQL & MongoDB

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 NOSQL
- Learn Replication and Sharding
- 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
3. MongoDB The Definitive Guide, O' Reilly by Christina Chodrow

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
3. Mastering MongoDB by Alex Giamas, Publisher: Packt

DS 503: Data Analytics using R & Python

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 & Python is used. What makes this course unique is that participant will continuously practice their newly acquired skills through R Studio and Python both. In the final section, participant will dive deeper into the MongoDB-Python Interaction and Prediction using Time Series Analysis.

Module Duration: 120 Hours

Pre-Requisite: M.E./M.Tech/B.E./B.Tech/DOEACC B Level/Any Master Degree with Knowledge of Mathematics/Statistics and Computer Programming.

DS 503 Syllabus

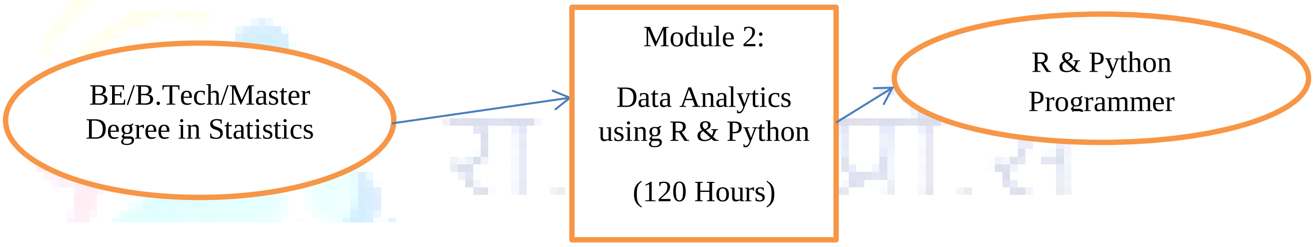
Module 3: Data Analytics using R & Python				
Section	**Topics to be covered**	**Theory**	**Practical**	**Total Duration (In Hours)**
	R			
DS 503.1	Basic Concept of Data Analytics & Data Manipulation in R	06	10	16
DS 503.2	Statistical Distribution using R	08	08	16
DS 503.3	Testing of Hypothesis and Goodness of Fit Test using R	04	04	08
DS 503.4	Data Mining using R	16	20	36
DS 503.5	Bayesian Analysis in R	02	02	04
	Python			
DS 503.6	Python Basics	02	02	04
DS 503.7	OOPs concept & Exception Handling in Python	02	04	06
DS 503.8	Data Analysis in Python	04	04	08

DS 503.9	Inferential Statistics in Python	04	04	08
DS 503.10	Data Visualisation using Python	02	04	06
DS 503.11	MongoDB - Python Interaction	02	02	04
DS 503.12	Introduction to Time Series Analysis	01	01	02
DS 503.13	Time Series Analysis using Python	01	01	02
	Total	**54**	**66**	**120**

Tools to be used

1. Ubuntu Operating System
2. Virtual Box
3. R Studio
4. Python

Learning Outcome


BE/B.Tech/Master Degree in Statistics
Module 2:
Data Analytics using R & Python
(120 Hours)
R & Python Programmer

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 Python-MongoDB Connectivity.

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
3. Fundamentals of Python: Data Structures by Kenneth A. Lambert
4. Python for Data Science for Dummies by Luca Massaron, John Paul Mueller

Reference Books

1. Advanced R: Data Programming and the Cloud by **:** Matt Wiley,Joshua F. Wiley
2. Statistical Analysis with R for Dummies by**:** Joseph Schmuller
3. R and Data Mining -- Examples and Case Studies, Author: Yanchang Zhao,Publisher: Academic Press, Elsevier, ISBN: 978-0-123-96963-7
4. Learn to Program with Python by Irv Kalb
5. Professional Python by: Luke Sneeringer

DS 504: Fundamentals of Java for Hadoop Framework

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: 120 Hours

Pre-Requisite: M.E./M.Tech/B.E./B.Tech/DOEACC B Level/Any Master Degree with Knowledge of Mathematics/Statistics and Computer Programming.

DS 504: Syllabus

Module4:

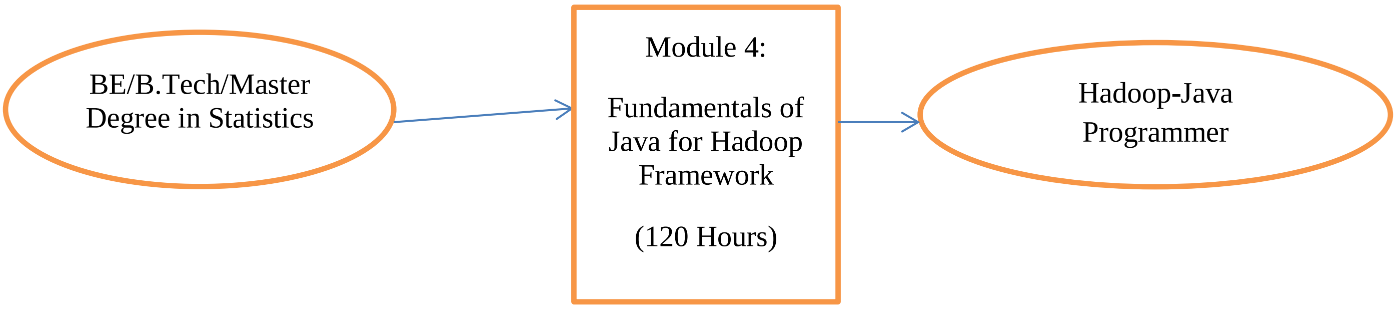
Fundamentals of Java for Hadoop Framework

Section	**Topics to be covered**	**Theory**	**Practical**	**Total Duration (In Hours)**
DS 504.1	Basic Java	04	04	08
DS 504.2	Arrays, Objects and Classes	06	10	16
DS 504.3	Control Flow Statements	04	04	08
DS 504.4	Inheritance and Interfaces	08	08	16
DS 504.5	Exception Handling & Serialization	08	08	16
DS 504.6	Multithreading in Java	12	12	24
DS 504.7	Collections	08	08	16
DS 504.8	Reading and Writing files	08	08	16
Total Duration		**58**	**62**	**120**

Tools to be used

1. Ubuntu Operating System
2. Virtual Box
3. Java
4. NetBeans

Learning Outcome


BE/B.Tech/Master Degree in Statistics

Module 4:

Fundamentals of Java for Hadoop Framework

(120 Hours)

Hadoop-Java Programmer

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

- Understand the basic concepts of Java.
- Understanding MapReduce
- Understanding Multithreading and Serialization.

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: Hadoop Eco System

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: 240 Hours

Pre-Requisite: M.E./M.Tech/B.E./B.Tech/DOEACC B Level/Any Master Degree with Knowledge of Mathematics/Statistics and Computer Programming.

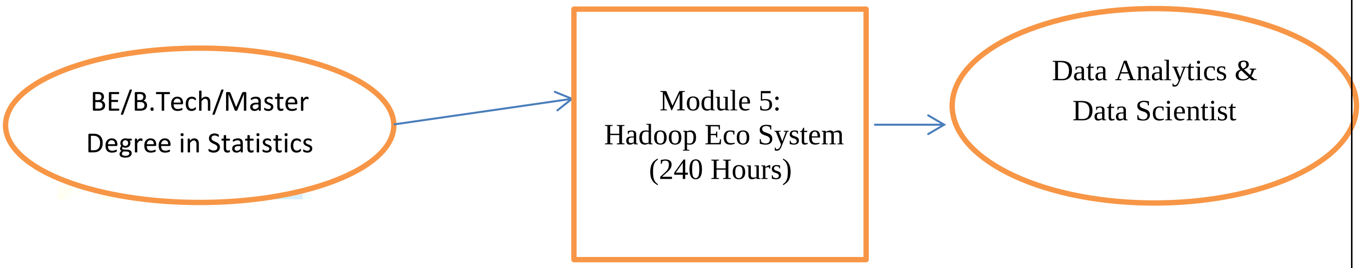
DS 505: Syllabus

Section	Topics to be covered	Theory	Practical	Total Duration (In Hours)
DS 505.1	Introduction to Big Data and Hadoop Eco System	04	-	04
DS 505.2	Configuring Hadoop	08	08	16
DS 505.3	HDFS Architecture	04	08	12
DS 505.4	Hadoop MapReduce	08	16	24
DS 505.5	Working with Sqoop	08	16	24
DS 505.6	Working with Pig and HIVE	08	16	24
DS505.7	Configuring HBase	16	24	40
DS505.8	Machine Learning using Python for Big Data	32	48	80
DS 505.9	Apache Spark	06	10	16
Total		**94**	**146**	**240**

Softwares to be used

1. Ubuntu Operating System
2. Virtual Box
3. MySQL
4. Python
5. R
6. MongoDB
7. Hadoop

Learning Outcome



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
- Implementation of Machine Learning for Big Data.
- 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: M.E./M.Tech/B.E./B.Tech/DOEACC B Level/Any Master Degree with Knowledge of Mathematics/Statistics and Computer Programming and good knowledge of data Analytics.