नेशनल इंस्टीट्यूट ऑफ इलेक्ट्रॉननक्स 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

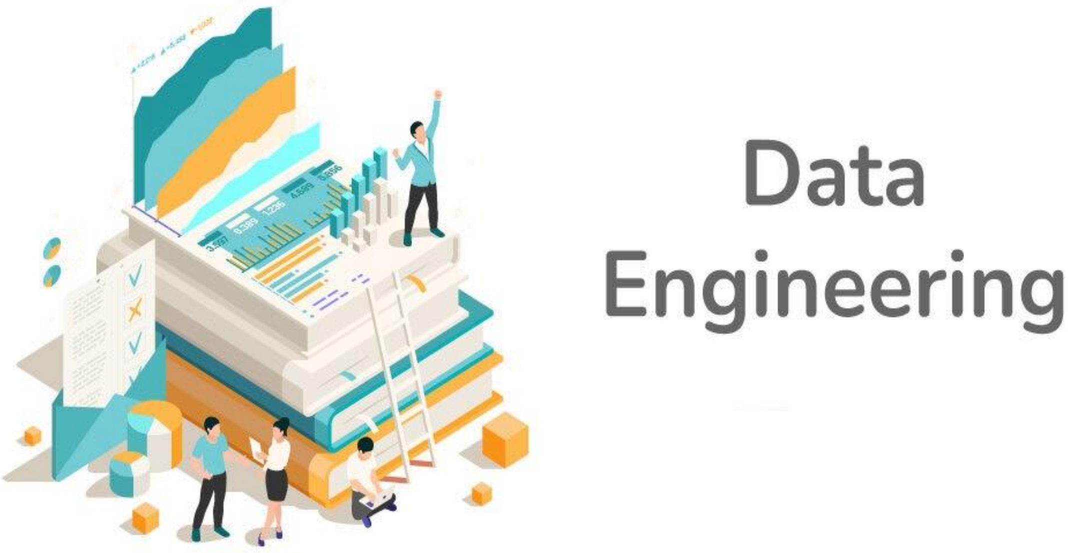
ISTE Complex, 25, Gandhi Mandapam Road, Chennai - 600025

Course Prospectus

PG Program in Data Engineering

NSQF Aligned

Mode: Classroom



Index

Course Prospectus

Course Name: PG Program in Data Engineering

Course Code: DS 500

NSQF Level: 06

Duration: 840 Hours, 6 Months

Last Date of Registration: 14-10-2022

Date of publishing Provisional Selection List: 17-10-2022

Payment of first installment fee: 17-10-2022 to 20-10-2022

Course Start Date: 21-10-2022

Preamble:

Data Science refers to extraction of knowledge from large volumes of data that are structured or unstructured, which is continuation of data mining and predictive analytics. It involves different categories of analytical approaches for modeling various types of business scenarios and arriving at solution and strategies for optimal decision-making in marketing, finance, operations, organizational behaviour and other managerial aspects. This new field of study breaks down into a number of different areas, from constructing big data infrastructure and configuring the various server tools that sit on top of the hardware, to performing the analysis and developing the right transformations to generate useful results.

Objective of the Course:

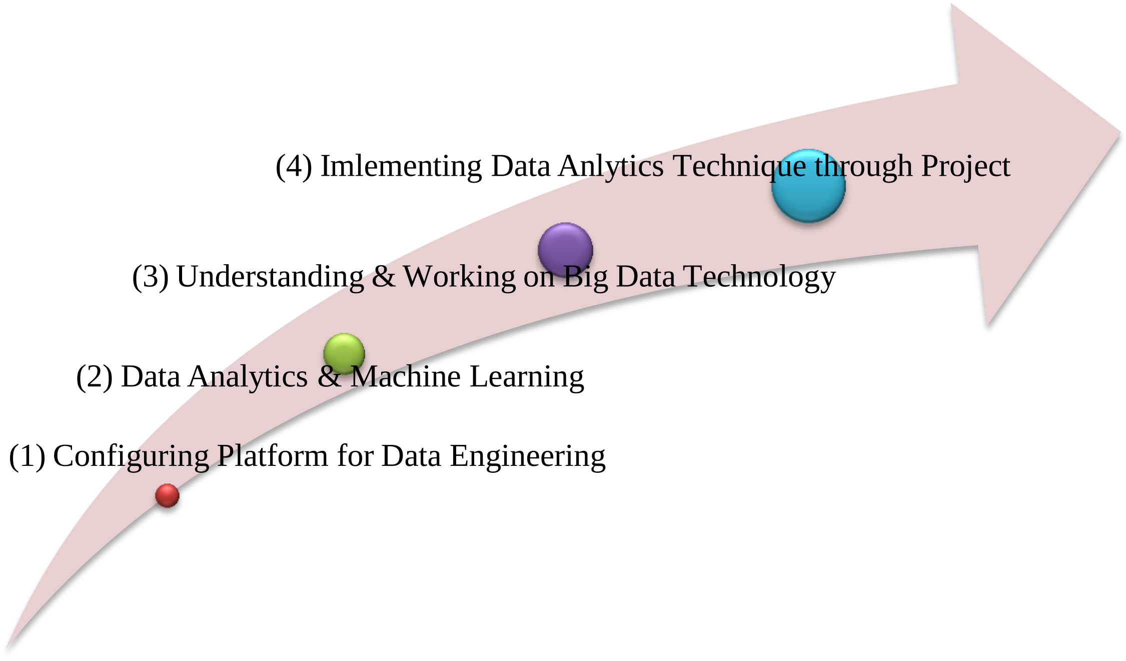
The **PG Program in Data Engineering** a unique 6-month (840 Hours) program offered by NIELIT Chennai is an excellent blend of knowledge and practice in the field of Data Science and its industrial applications. The program is targeted for creating qualified Data Science Engineers. The course progresses through the Operating System, concepts of Data and its storage, programming for data science, Big Data Technology and its implementation. Various advanced tools such as R and Python, along with MySQL, Apache Cassandra, Java Programming and Hadoop Framework are used for achieving the goal of solving critical business and Analytic problems. The Program also offers six weeks of hands-on real – life analytical projects for participants to get equipped with strong analytical and programming which makes them highly demanding and employable on completion of the program. The course has been designed after proper industry survey and consultation with multiple industry leaders to ensure that participants learn exactly what employers need.

The objective of this program is to make Statistical Analysts, Data Scientists, Data Analysts, Big Data Engineer, Hadoop Developer. There is a huge demand for resources skilled in Data Science. There is huge shortage of Data Science Professionals world-wide. So, it is quite obvious that existing resources along with new candidates who are interested in perusing career in this field needs to be trained. Our objective is to create a pool of talent who can meet this demand. This course is meant to sensitize students for computational statistics applications and usage as well as provide hands-on experience with solving real world data science issues.

Outcome of the Course:

On completion of the Course, the Participants will learn the concept of Data Analytics using open source statistical tools like R, Python and some very good visualization tools and techniques. They will be able to implement industry-oriented Data Analytics Project.

Full Flow of Course


(4) Imlementing Data Anlytics Technique through Project
(3) Understanding & Working on Big Data Technology
(2) Data Analytics & Machine Learning
(1) Configuring Platform for Data Engineering

Course Structure

This course contains total three modules. After completing the three modules, the students have to do a 120 Hours project using any of the topics studied in the course.

DS 600	Module Name	Duration (in Hours)
Module 1	Configuring Platform for Data Engineering	240
Module 2	Data Analytics & Machine Learning	240
Module 3	Big Data Analytics	240
Module 4	Mini Project (Implementation of Data Analytics)	120
Total Duration		840

Course Fees

Total Course fee is Rs.52,000/- including GST (Can be paid as a single instalment of Rs. 52,000/- or in 2 instalments as given below)

Registration Fee	**Rs. 1000/- for SC-ST (Refundable*)**	**Rs. 1000/- for others Adjustable with Training fee**	
Instalment No.	SC-ST Candidates (Fee including GST in Rs.)	General Candidates (Fee including GST in Rs.)	Last Date
1	*	25,000.00	20-10-2022
2		26,000.00	20-12-2022
Total	**1,000.00***	**52,000.00**	
***Tuition Fees are waived for eligible SC/ST students who are successfully completing the course with NSQF certification under SCSP/TSP Scheme**			

**GST is Applicable as per Govt. Norms GST (currently it is 18%).*

Registration Fee- Refund Policy:

(Non-Refundable if candidate is selected for admission but did not join and if a candidate has applied but not eligible.)

However, the registration fee shall be refunded on few special cases as given below:

- ✓ Candidates are eligible but not selected for admission.
- ✓ Course postponed and new date is not convenient for the student.
- ✓ Course cancelled.

Eligibility

- ✓ B.E./B.Tech/M.S./M.C.A./M.C.S./DOEACC „B" Level/M.Sc./Master Degree in Mathematics or Statistics or Operations Research/Economics or Econometrics/Applied Mathematics/Applied Statistics/M.B.A.
- ✓ Minimum Age: 20 Years

Number of Seats: 25 (Twenty-Five) - Total

Category	No. of Seats
SC (15%)	04
ST (7.5%)	02
OTHERS	19
Total	**25**

Note: Seats are allocated based on the merit basis

How to Apply?

Candidates can apply online in our website http://reg.nielitchennai.edu.in. Payment towards non-refundable registration fee can be paid through any of the following modes:

- ✓ Online transaction: Account No: 31185720641 Branch: Kottur (Chennai), IFS Code: SBIN0001669.
- ✓ GPAY/any UPI

Note: *The Institute will not be responsible for any mistakes done by either the bank concerned or by the depositor while remitting the amount into our account*

Last date of Registration: 14th October, 2022

Registration Procedure

All interested candidates are required to fill the Registration form online with registration fees before **14th October, 2022** with all the necessary information.

Selection Criteria of candidates

Selection of candidates will be based on their marks in the qualifying examination subject to eligibility and availability of seats.

- ✓ The first list of Provisionally Selected Candidates will be published on NIELIT Chennai website (www.nielit.gov.in/chennai) on **17-10-2022** by **5:00 PM**. In case of vacancy, an additional selection list will be prepared and the selection will be intimated by email only.
- ✓ Provisionally selected candidate has to upload following document on registration portal for online verification:
- ✓ **For SC/ST:**
  - Original Copies of Proof of Age, Qualifying Degree (Consolidated Mark sheet & Degree Certificate/Course Completion Certificate), 10th and 12th mark sheets.
  - Self-attested copy of community certificate.
  - One passport size photograph.
  - **AADHAR Identity proof must for SC/ST Candidates** (For availing concession).
  - Candidates under fee-waver category (SC-ST) have to give undertaking (notarized) through post indicating that they will not discontinue the course in between.
- ✓ **For Others (General, OBC, EWS) :**
  - Original Copies of Proof of Age, Qualifying Degree (Consolidated Mark sheet & Degree Certificate/Course Completion Certificate), 10th and 12th mark sheet.
  - One passport size photograph.
  - Self-attested copy of Govt. issued photo ID card.
- ✓ After document verification selected candidates have to pay first instalment of Rs. ***25,000/-*** on or before **20-10-2022 by** payment mode mentioned above.
- ✓ Selected candidates are requested to upload the proof of remittance of fee on registration portal and also send the proof of remittance of fee as email to skjha@nielit.gov.in/trng.chennai@nielit.gov.in.

Admission: All provisionally selected candidates whose documents are verified and paid the fees (full or first installment) and verified by accounts section of NIELIT Chennai will be notified about course timing through e-mail/Whatsapp.

Discontinuing the course

- ✓ No fees (including the security deposit) under any circumstances, shall be refunded in the event of a student who have completed the process of admission or discontinuing the course in between. No certificate shall be issued for the classes attended.
- ✓ If candidate's attendance is less than 75%, then they will not be allowed to appear in final examination and all fees paid will be forfeited.
- ✓ If candidates are not appearing for any internal examinations/practical their candidature will be cancelled without any notice and all fees paid will be forfeited.

Course Timings: This program is a practical oriented one and hence there shall be more lab than theory classes. The Class timing is from 10:00 am to 5:00 pm and Monday to Friday and offered at NIELIT Chennai in classroom mode.

Address:

National Institute of Electronics and Information Technology
ISTE Complex, No. 25, Gandhi Mandapam Road, Chennai – 600025
E-mail: trng.chennai@nielit.gov.in/Phone: 044-24421445

Contact Person: Dr. Sanjeev Kumar Jha, Mobile: 7765803105

Course enquiries:

Students can enquire about the various courses either on telephone or by personal contact between 9.15AM to 5.15PM. (Lunch time 1.00 PM to 1.30 PM) Monday to Friday.

Placement:

Students who have completed the course successfully and qualified, placement guidance and career counseling will be given to crack the interviews.

Important Dates

Last Date of Registration: 14-10-2022

Display of Provisional Selection List: 17-10-2022

Payment of first installment fee: 17-10-2022 to 20-10-2022

Course Start Date: 21-10-2022

Payment of second installment fee: 20-12-2022

Examination & Certification

- ✓ Final Certificates will be issued after successful completion of all the modules including mini project. For getting certificate a candidate has to pass each module individually with minimum required marks of 50%.

NSQF Examination Pattern:

Examination scheme is as follows:

Theory **(Each Question will carry 1 mark)** **Duration (in Min): 90**		**Practical**			**Internal Assessment (Marks)**	**Project / Presentation / Assignment** **(Marks)**	**Major Project/** **Dissertation**		**Total**
Papers	**No. of Questions/** **Paper**	**Papers**	**Duration (in Min)**	**Marks/Paper**			**No. Of Projects**	**Marks**	
3	100	2	180	90	60	60	1	100	700

Theory Papers

- **Theory 1** – Configuring Platform for Data Engineering
- **Theory 2** – Data Analytics and Machine Learning
- **Theory 3** – Big Data Analytics

Practical Papers

- **Practical 1 –** Configuring Platform for Data Engineering& Machine Learning
- **Practical 2** – Big Data Analytics

Means of Assessment						
	Theory	Practical	Internal Assessment	Project/ Presentation/ Assignment	Major Project/ Dissertation	
No of Papers	03	02	01	01	01	
Marks Each Paper	100	90	60	60	Presentation	Projects
					40	60
Total Marks	300	180	60	60	100	
					700	

Grading Scheme

Following Grading Scheme (on the basis of total marks) will be followed:

Grade	S	A	B	C	D	Fail
Marks Range (in %)	85 to 100	75 to 84	65 to 74	55 to 64	50 to 54	Below 50

Final Grading as per above grading scheme will be given on the basis of total marks obtained in all modules.

Director, NIELIT Chennai

Dr. Pratap Kumar S

Director

Dr. Pratap Kumar S, is BTech (Electrical Engineering), M Tech (Digital Electronics), MBA (Marketing) and PhD (Strategic Management). He has More than 29 year‟s experience in planning and execution of industrial consultancy projects, and capacity building projects funded by both industry and central & state ministries. Executed 7 major industrial consultancy projects and associated with the development of more than 50 product technologies, empowered more than 10,000 candidates through various capacity building programs and facilitated more than 40,000 job seekers through various job fairs and outreach programs. He has expertise in Strategy, Product Development, Automotive Electronics, Embedded Systems, and Power Electronics.

Programme Co-Ordinator

Dr. Sanjeev Kumar Jha

Joint Director and Head (Academics)

Dr. Sanjeev Kumar Jha, is Masters in Statistics and Ph.D. in Computer Science and Engineering. He has extensive experience of more than two decades as an educator and researcher. He has published various research papers. Currently, he is Joint Director (Technical) at National Institute of Electronics and Information Technology (NIELIT), Chennai. He has expertise in various domains like Data Science, Big Data, Power BI and Software Development (Open Source). He has more than 24 years of experience in planning and execution of various training Programs and capacity building projects funded by both industry and central & state ministries.

Faculties

Gayathri V, is Masters in Computer Science and Engineering. She is a Resource Person (IT) at National Institute of Electronics and Information Technology (NIELIT), Chennai. She has an experience as an educator in various domains like Networks and security, Artificial Intelligence, DBMS etc. for 4 years. Her area of interest is Data Science & Networks

Jayakodi R

Resource Person

Jayakodi R, is Masters of Engineering specialization in Software systems. She is a Resource Person (IT) at National Institute of Electronics and Information Technology (NIELIT), Chennai. She has an experience in various domains like Telecommunication domain and hands on experience in Java technologies. Her areas of interests include Web Technologies, Databases and Software Development.

Vignesh M

Resource Person

Vignesh M, completed Computer Science & Engineering. He is a Resource person (Data Science) at National Institute of Electronics and Information Technology (NIELIT), Chennai.His area of interest is R, Python and Machine Learning.

Ragesh Varma T

Resource Person

Ragesh Varma T, completed Computer Science and Engineering and Post Graduate program in Data Science and Engineering. He is a Resource Person (IT) at National Institute of Electronics and Information Technology (NIELIT), Chennai. his area of interest in Machine Learning and DBMS.

Detailed Curriculum

Module 1: Configuring Platform for Data Engineering

Understanding Linux Environment & Basic Commands:

Understanding Linux Environment:

- Introduction, Linux Architecture, Boot Process, Kernel, System Initialization, GUI, and CLI (Access a shell prompt and issue commands with correct syntax.
- **Commands**:
- file handling commands, sort, tr, cut, find, grep, egrep, using filters, cat, mkdir, who and other basic commands. vi editor

Linux Package management and Process Monitoring:

- su login,sudo,apt-get,ps command, kill command and other related commands, single and multi-user mode of Ubuntu.
- Important Files and Directories.

BASH Scripting:

- Introduction to BASH, Variables(System & User defined),Exporting Variables, Special Shell Variables, Control Structures, Understanding execution mode of BASH script, Array, functions, BASH debugging

Case Study:

- BASH Script for removing missing value
- 2.BASH Script for generating Frequency Distribution Table from given data (consisting of 10000 records).
- BASH Script for removing blank lines from a file.
- BASH Script to find frequency of a word from several files.
- BASH Script to Merge files based on some fields.

Configuring Secure Shell & LAN

LAN

- Introduction, Configure LAN on Ubuntu

Secure Shell:

- Understanding & Configuring Secure Shell, Access remote systems using ssh, SCP, Passwordless SSH, Configure key-based authentication for SSH

User Administration

User Management:

- Adding/Modifying/Deleting new users, Understanding User Id and other related fields. Understanding /etc/passwd and /etc/shadow, Password Aging Policies**,** Switching Accounts, sudo access

Group Management:

- User Private Groups, Group Administration.
- **Understanding SUID and SGID Executable, Sticky Bit, Default File**
- **chmod and chown command**

Virtualization

- Introduction to Virtualization
- Virtual Machine installation, Configuring Virtual Machines, Install Ubuntu/Centos systems as virtual guests, configure systems to launch virtual machines at boot. , Creating Clone of a Virtual Machine and its restoration, virtual LAN, Memory addressing, Paging, Memory mapping, virtual memory, complexities and solutions of memory virtualization, VM configurations, VM migrations, Migration types and process.

- **Basics of Information Security & Cloud**

Java for Hadoop

Java Introduction:

- Concept of OOPs
- Introduction to Java
- Configure JAVA PATHs in PATH variable and other related places in Linux.
- Features of Java
- Working with Java Variables
- Declaring and Initializing Variables
- Primitive Data Types
- Class & Object Fundamentals
- Object Lifecycle
- Read and Write Java Object Fields

Understanding JAR file and its working

Java Operators and Decision Constructs

Using loop Constructs in Java

- while, for, switch case etc.

Array & String:

- Creating and using One-Dimensional Array
- Creating and using Multi-Dimensional Array
- String Class and related functions.

Methods and Encapsulation:

- Java Method
- Static and Final Keyword
- Constructors and Access Modifiers in Java
- Encapsulation

Inheritance:

- Polymorphism Casting and Super
- Abstract Class and Interfaces

Exception Handling:

- Types of Exceptions and Try-catch Statement
- Throws Statement and Finally Block
- Exception Classes
- Creating Custom Exception Classes

Work with Selected classes

- String & String Buffer
- Create and Manipulate Calendar Data
- Declare and Use of Array list

Collection Framework

- Introduction to Collection Framework
- Core Collection in Java

- Core Collection framework
- Types of Collection,
- Hierarchy of Collection Framework
- Commonly used methods of Collection interface
- Iterator Interface
- Methods of Iterator interface

File Handling and Serialization

- The Classes for Input and Output
- The Standard, Streams
- Working with File Object
  - File I/O Basics, Reading and Writing to Files
  - Buffer and Buffer Management
- Read/Write Operations with File Channel.
- Serialization

JDBC

- JDBC and its Architecture
- Drivers in JDBC
- JDBC API and Examples
- Transaction Management in JDBC

Data warehousing using MySQL

Data warehousing concept

Data Base Design using MySQL:

- Concept of RDBMS, Storage Engine, Structure of MySQL
- Creating Database, Data Types, Table etc.

Relational Model and SQL:

- Relation Model, MySQL Query, Creating and Using a Database, Select, Operators, group by, order by, Primary Key, etc.

Database Design using the Relational Model:

- Making Relation between tables, Foreign Key, joins etc.

Export & Import Data

- Export and Import of External Data, Interacting with different tables, Backup and Recovery.

Basics of NoSQL and Apache Cassandra

Introduction to NoSQL and Cassandra:

- Understanding NoSQL, Types of NoSQL databases, Usage of NoSQL databases, NoSQL EcoSystem,Overview of Cassandra, Features of Cassandra, Cassandra Vs. MongoDB

Architecture of Apache Cassandra:

- Understanding high level Cassandra architecture,
- Peer-to-Peer design, Network topology, Virtual Node, Components of Cassandra, Partitioner and Replication, Memtables and SSTables, Bloom Filters, Managers and Services, Cassandra read and write process, Failure scenario.

Apache Cassandra:

Installation &Configuration

- Versions of Apache Cassandra
- Understanding Pre-requisite for Installation
- Installing Cassandra
- Linux Commands to auto start Apache Cassandra
- Logging setup in Cassandra
- Understanding Replication Factor
- Cassandra Cluster
- Miscellaneous setting

Understanding Apache Cassandra Data model:

- Introduction to Data Model
- Design between RDBMS and Cassandra
- Understanding Cassandra API:CQL-API and thrift API)

Cassandra Monitoring Tools:

- Introduction of Monitoring Tools
- Cluster Statistics

i. nodetool

ii. JConsole

iii. Table Statistics

- Table Statistics
- Thread Pool
- Compaction Metrics

Cassandra Cluster:

- Introduction to Cluster
- Layers of Cassandra Cluster
  - Node Cluster
  - Keyspace
  - Column Families
  - Rows
  - Column
- Cluster Builder

Cassandra CQLSH

- Introduction to CQL

Documented Shell Commands:

- Help,Version,Color,No Color
- DEBUG,Execute,File,U,P
- Exit,Describe,Expand etc.

CQL: Data Definition Commands:

- Create Keyspace
- Use Keyspace
- Alter Keyspace
- Drop Keyspace
- Create Table
- CRUD Operation
- Alter table
- Add Column to a table
- Drop a Column

- Truncate Table
- Drop Table

CQL: Data Manipulation Commands:

- Insert Command
- Update Command
- Delete Command
- Batch Command

CQL Clauses:

- Select
- where
- Order by

Cassandra Data types

- Build-in
  - (Boolean,blob,ascii,bigint,counter,decimal,double,float,inet,int,text,varchar,timestamp,varint etc.)
- Collection data Type
  - List (Create,Insert,Update,Verify)
  - Map((Create,Insert,Update,Verify)
  - Set(Create,Insert,Update,Verify)
- User Defined Data type
  - Create
  - Alter
  - Add
  - Drop
  - Describe
- Database User and Roles
- Control Commands
- Complex query
- Built-in and User defined Function
- Run CQL Scripts from the command line
- JSON support

Indexes and Composite Columns:

Overview of Index and benefit:

- Understanding Index
- Create Index
- Drop Index
- Index on Distributed Database
- Clustered Indexes vs Non-Clustered Indexes
- Secondary Index
- Composite Columns
- Data Partitioning
- Data Colocation

Cassandra Interfaces:

- Java interfaces to connect Cassandra
- ODBC interface to connect Cassandra

Module2: Data Analytics & Machine Learning

Basic Concept of Data Analytics & Data Manipulation in R

- Introduction to Data Analytics
- Basic Features of R, Installation of R Studio and method of accessing through URL.
- Basic Data Sets: Vector, Matrices, List, Array, Factors
- Data Frames, Data Types, Operators, Basic Constructs, R Functions, String Handling, R Packages
- Data Reshaping, Data Pipelines, and Data Manipulation.

Python Basics

- Features of Python.
- Basic Syntax, Variable and Data Types, Operators.
- Conditional Statement, Loops, Functions, File Handling in Python.

OOPs concept & Exception Handling in Python

- Concept of class, object and instances, Constructor, Inheritance.
- Programming using Oops support, Exception Handling.

Understanding Data Frame in Python

- Working with Pandas data structures.
- Series and Data Frames,
- Accessing data: indexing, slicing, Boolean indexing, dropping
- Import from and Export to .csv Files, Output to and Input from EXCEL Files, selecting, creating, and combining rows and columns
- Pandas: XLS
- Pandas: JSON
- Missing Value
- Data Aggregation, group by etc.
- Reshaping and transforming data.

Data Visualization using Python

- Understanding on Data Visualization,
- Using Python Library for visualization: MatplotLib,Seaborn,plotly,ggplot,GeoPlotlib.gleam etc.
- Pie Chart, Histogram, Box Plot and other visualisation techniques.

Inferential Statistics in Python

- Introduction to Inferential statistics.
- Random Variable
- Measure of Central Tendency SciPy package
- Understanding Mathematical Expectation.
- Distribution Functions (Discrete and Continuous)
  - Binomial, Poisson
  - Normal Distribution
- Constructing a Statistical Model.
- Fitting Model to given data.
- Testing of Hypothesis:
  - Introduction to ToH
  - Understanding Null and Alternative hypothesis.
  - Critical Region
  - Level of Significance, P Value
  - T Test

  - Z Test
  - Goodness of fit Test
  - Chi Square Test

Time Series Analysis using Python

- Time Series Introduction, Understanding Time Series Data.
- Importing Time Series Data in Python.
- Working with Time Series Libraries like autots,tsfresh,dart,atspy etc.
- Understanding Panel Data.
- Visualisation of Time Series Data
- Patterns in a Time Series
- Additive and Multiplicative Time Series
- Decomposing a time series into its components.
- Working with Seasonal & Nonseasonal Time Series.
- Working with Stationary and Non-Stationary Time series.
- Test for Stationarity of Time Series Data.
  - Augmented Dickey Fuller(ADH) Test
  - Kwiatkoski-Phillips-Schmidt-Shin(KPSS) Test
  - Philips Peron(PP) Test
- Understanding noise in Time Series Data.
  - Understanding white noise and stationary series.
- De-trending a Time series data.
- De-seasonalise a Time Series Data
- Test for Seasonality of Data
- Handling Missing Value in Time Series Data.
  - Backward Fill
  - Linear Interpolation
  - Quadratic Interpolation
  - Mean of Nearest Neighbours
  - Mean of seasonal counterparts
- Smoothening a Time Series data.
- Autocorrelation & Partial autocorrelation function.
- Lag Plots
- Forecasting a Time Series Data.
- Causality Test for Time Series Data.

Machine Learning

Introduction to Machine Learning:

- Basic Concepts of Machine Learning
- End-to-end Process of Investigating Data through a Machine Learning Lens.
- Application of Machine Learning.

Types of Machine Learning

Supervised

- Classification
- Regression:
  - Linear Regression
  - Generalized Linear Regression

- Logistic Regression
- Multiple regression
- Poisson Regression

Unsupervised

Clustering

- The k-Means Clustering
- The k-Medoids Clustering
- Hierarchical Clustering
- Density-based Clustering

Dimensionality Reduction

- Principle Component Analysis
- K-nearest neighbour
- Discriminant Analysis

Anomaly Detection

- KNN

Outlier Detection, Association Rules

- Basics of Association Rules
- Association Rule

Mining

- Text Mining

Tree, Decision Tree, Splits, Entropy etc.

Neural Networks

Introduction

- Understanding Neural Networks
- Building simple Neural Network in Python
- Multiple Input & Outputs
- Use of NumPy to build Neural Network

Updating Weights in Simplest Neural Network

- Simple Error analysis
- Working with 1 attribute
- Small Steps
- Extending Simplest Neural Network to Multiple Inputs
- Extending to Multiple Outputs
- Combining Multiple Input and Outputs

Extending Neural network to Completed Data Sets

- Extending Neural Network to Use Multiple Samples
- Goodness of Fit Parameters

- Perceptron Learning & Binary Classification
- Back Propagation Learning
- Learning Feature Vectors for Words & Object Recognition.

Deep Learning Applications

- Artificial Neural Networks
- Introduction to KERAS for Classification and Regression in Typical Data Science Problems

- Creating a Neural Network Training Models and Monitoring
- Introducing TensorFlow,
- Neural Networks using TensorFlow, Debugging and Monitoring, Convolutional Neural Networks
- CNN using TensorFlow
- Unsupervised Learning
- Working with PyTorch
- Case Studies
- Cassandra Python Connectivity

Module 3: Big Data Analytics

Introduction of Big Data Analytics

Introduction to Big Data:

- Big Data for Data Engineering
- Big Data Introduction
- Attributes of Big Data
- Other technologies vs Big Data
- Big Data & Data Science
- Processing Big Data

Introduction to Hadoop:

- Introduction to Hadoop Ecosystem
- Compare Hadoop vs. traditional systems
- Hadoop Architecture
- Understanding HDFS

Configuring Hadoop:

- Installing Hadoop
- Standalone mode
- Pseudo Distributed Mode
- Fully Distributed
- Understanding Hadoop Cluster
- Monitoring the Cluster Health
- Starting and Stopping the Nodes

HDFS Architecture

- Distributing Processing System
- Core Components of Hadoop
- HDFS Architecture, HDFS Design
- HDFS role in Hadoop
- Features of HDFS
- Daemons of Hadoop and its functionality Name node, Data node
- Secondary Name Node
- Job Tracker, Task Tracker
- Anatomy of File Write & File Read
- Network Topology
- Heartbeat Signal
- How to Store the Data into HDFS
- How to Read the Data from HDFS
- CLI commands (Hadoop FS shell)
- Hadoop Administration & Admin Commands

Hadoop MapReduce using Python

- Concepts of HDFS Java API

- Overview of MapReduce Framework,
- MapReduce Architecture & Daemons
- Job tracker and Task tracker
- YARN and its Processing Application
- YARN MR Application Execution Flow
- Data Flow In MapReduce
- Introduction to Hadoop Streaming
- Streaming Command Options
- Use cases of MapReduce, Anatomy of MapReduce Program
- Basic MapReduce API Concepts.
- Writing MapReduce Driver, Mappers, and Reducers, Unit Testing MapReduce Programs etc,, Case Studies
- Generic Command Options
- Basic MapReduce Sample Program-1
- Basic MapReduce Sample Program-2
- Chaining of MR Jobs
- Custom Combiner
- Generic OptionParser
- Analysis of IRIS dataset
- Built-in and Custom Counters in Hadoop
- Custom Partititioner
- Hadoop Sequence File Format
- Read Write Sequence File
- Hadoop Data Types
- Processing of XML File
- Data Compression with Hadoop
- Data Serialization
- Use Cases
- Integration of Cassandra and Hadoop
- Hadoop Map Reduce using Cassandra

Working with Pig and HIVE

Pig

- Installation & configuration of Pig
- Architecture
- Datatypes (scalar, complex)
- Running Pig (interactive, Batch)
- Pig Operators – Local, Store, Dump, Distinct
- Filter, For Each, generate, Limit
- Union, join, order by, Describe, Group by, Avg Default UDFs available (Built in function)
- REG EX, EXPLAIN, Parallel processing,
- Custom UDF

Hive

- Installation of Hive and its configuration
- Understanding Hive Services
- Architecture of Hive
- Comparing Hive to traditional Databases
- Relational Data Analysis – (data types (primitive, complex) databases tables, create, alter, delete Hive Schema & Data storage
- Loading data into Hive
- Views
- Storing query results (store)

- Apply Statistical functions .
- Text processing - Built in functions, string functions, regular expressions
- Managed vs External Tables.

Apache HBase

- HBase Introduction
- HBase vs RDBMS (fixed Vs flexible schema)
- Understanding HBase Configuration files and its configuration in Hadoop Eco System
  - HBase run modes: Standalone and Distributed
  - Understanding and working with Zookeeper, Master and Region servers in fully distributed Mode.
- HBase Architecture and Components,
- HBase Data Model
- Understanding Conceptual View, Physical View, Namespace,Table,Row,Column Family, Cells,Data Model Operations,Versions,Sort Order,Column Metadata,Joins,ACID etc.
- HBase commands
  - *General Commands:* Status,Version,Table_help(scan, drop, get, put, disable, etc.), whoami etc.
  - *Table Management Commands:* Create,List,Describe,Disable,Disable_all,Enable,Enable_all,Drop,Drop_all,Show_filters,Alter,Alter_status etc.
  - *Data manipulation commands:* count,put,get,delete,delete all,truncate,scan etc.
  - *Cluster Replication Commands:* add_peer,remove_peer,start_replication,stop_replication etc.
  - Understanding TTL(Time to Live)
- H Base Constraints
- Case Study - Log Data and Time series Data
- Case Study - Customer/Order
- H Base and MapReduce
- H Base Backup and Restore

Scala, Apache Spark, Kafka & Flume

- Introduction to Scala
- Scala REPL (“Read-Evaluate-Print-Loop”)
- Basic Scala Operations
- Variable Types in Scala
- Control Structures in Scala
- Functions and Procedures
- Collections in Scala
- Array, Array Buffer, Map, Tuple, Lists etc.
- Object Oriented Programming and Functional Programming Concepts in Scala
- Methods, classes, and objects in Scala
- Packages and package objects
- Traits and trait linearization
- Java Interoperability
- Introduction to functional programming
- Functional Scala for data science
- Importance of functional programming and Scala for learning Spark
- Pure functions and higher-order functions
- Using higher-order functions
- Error handling in functional Scala

- Functional programming and data mutability
- Scala Collection API & Scala Implicits

Introducing Apache Spark

- Introduction of Apache Spark: Need of Apache Spark, Feature of Apache Spark.
- Understanding concept of Spark Cluster Modes on YARN.
- Apache Spark Installation and Configuration
- Understanding Spark Cluster Modes on YARN
- Spark Applications
- The back bone of Spark - RDD (Resilient Distributed Dataset)
- Loading Data
- What is Lambda
- Using the Spark shell
- Actions and Transformations
- Associative Property
- Implant on Data
- Persistence
- Caching
- Loading and Saving data
- Operations of RDD
  - Challenges in Existing Computing Methods
  - Probable Solution & How RDD Solves the Problem
  - Introduction to RDD, its operations, Transformations & Actions
  - Data Loading and Saving Through RDDs
  - Key-Value Pair RDDs, Other Pair RDDs, Two Pair RDDs
  - RDD Lineage & RDD Persistence
  - WordCount Program Using RDD Concepts
  - RDD Partitioning achieve Parallelization.
  - Using Accumulators
  - Creating custom Accumulators
  - Using Broadcast variables
  - Passing Functions to Spark
- Data Frames and Spark SQL
  - Introduction to Spark SQL & its architecture
  - SQL Context in Spark SQL
  - User Defined Functions
  - Data Frames & Datasets
  - Interoperating with RDDs
  - JSON and Parquet File Formats
  - Loading Data
  - Spark-Hive Integration
- Writing and deploying spark applications
  - Creating the SparkContext
  - Building a Spark Application using Scala.
  - The Spark Application Web UI.
  - Configuring Spark Properties
  - Running Spark on Cluster
  - Executing Parallel Operations
  - Understanding Stages and Tasks
  - Case Study

- Machine Learning using MLLib
  - Introduction to MLlib
  - Features of MLlib and MLlib Tool
  - Various ML algorithms supported by MLlib
  - Optimization Techniques
  - Supervised Learning
    - Linear Regression
    - Logistic Regression
    - Decision Tree
    - Random Forest
  - Unsupervised Learning
    - K-Means Clustering
  - Spark Algorithms for Machine Learning
  - Pyspark MLLib.
  - Integration with Hadoop PySpark-Environment and Configuration.
- **Working with Apache Kafka**
  - Introduction of Kafka & its architecture.
  - Kafka Components
  - Configuring Kafka Cluster.
  - **Kafka Producer.**
    - Constructing Kafka Producer.
    - Understanding Kafka Topics.
    - Sending Message to Kafka.
    - Producing Keyed and Non-Keyed Messages.
    - Sending Message.
    - Configuring Kafka Producer.
    - Serializing using Apache Avro.
    - Partitions.
  - **Kafka Consumer.**
  - **Kafka Internals**
    - Cluster Membership,
    - Controller, replication
    - Broker configuration etc.
  - **Kafka Stream Processing.**
- **Stream Processing using Spark and Kafka.**
- **Introduction of Apache Flume & its architecture.**
  - Flume Sources & Sinks.
  - Flume Channels & Flume Configuration.
  - Integrating Apache Flume and Apache Kafka.
- **Spark GraphX Programming**
  - A brief introduction to graph theory.
  - GraphX.
  - VertexRDD and EdgeRDD.
  - Graph operators.
  - Pregel API.
  - PageRank.