नेशनल इंस्टीट्यूट ऑफ इलेक्ट्रॉननक्ट्स 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

Certified Data Analyst (Big Data)

NSQF Aligned

Mode: ONLINE (Blended)

Index

Course Prospectus

Course Name: Certified Data Analyst (Big Data)

Course Code: DS 200

NSQF Level: 05

Duration: 240 Hours, 3 Months

Last Date of Registration: **08-02-2023**

Date of publishing Provisional Selection List: **10-02-2023**

Payment of first installment fee: **10-02-2023 to 15-02-2023**

Course Start Date: **15-02-2023**

Preamble:

The explosion of social media and the computerization of every aspect of social and economic activity resulted in creation of large volumes of mostly unstructured data: web logs, videos, speech recordings, photographs, e-mails and similar. In a parallel development, computers keep getting ever more powerful and storage ever cheaper. Today, we have the ability to reliably and cheaply store huge volumes of data, efficiently analyse them, and extract business and socially relevant information. Data Analytics is a field that dissects, efficiently extricates data from, or in any case manages informational indexes that are excessively huge or complex to be managed by customary information preparing application programming.

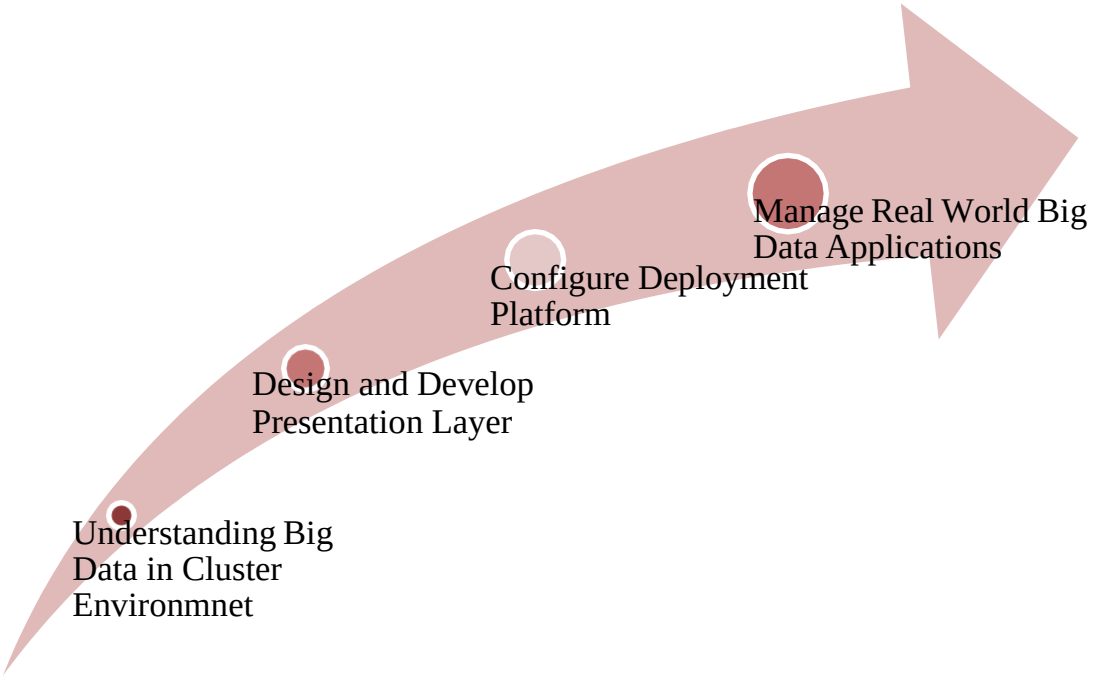
Objective of the Course:

The course progresses through Configure Deployment Platform, Concepts of Data Base Management System, Design and Development of Presentation, Big Data Technology and its implementation. The objective of this program is to provide skills to participants to analyse large volume of data using various tools and techniques. 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 various Big Data Technologies and its usage as well as provide hands-on experience with solving Big Data issues.

Outcome of the Course:

On completion of the Course, the Participants will would be able to analyze massive amount of data using various Big Data Techniques.

Flow of the Course



Course Structure

This course consists of following modules:

DS 200	Module Name	Duration (in Hours)
Module 1	Configuring Platform for Data Engineering	75
Module 2	Hadoop Eco System	90
Module 3	Enhancing Communication Skill	15
Module 4	Project	60
Total Duration		240

Course Fees

Training Fee: Rs. 14,000/- including GST (Can be paid as a single installment of Rs. 14,000 or in 2 instalments as given below)

Total Fee: Rs. 14,000/-

Registration Fee	**Rs. 1000/-** **Adjustable with Training fee**	
Instalment No.	**Fee including GST in Rs.**	Last Date
1	6000.00	15-02-2023
2	7000.00	15-04-2023
Total	**14,000.00**	

**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

- ✓ Final year student of B.E./B.Tech/ /MCA/MBA
  Or
- ✓ Three years polytechnic Diploma in Computer Science/IT/Computer Applications/Computer Engineering/Electronics with 2 Years of Experience in Data Science Domain.
  Or
- ✓ NSQF aligned course at Level 4 in Data Science domain with 2 Years of Experience in IT-ITeS Sector
  Or
- ✓ Any graduate with one-year Diploma/PG Diploma in Computer Science/IT/Computer Applications/Computer Engineering /Electronics with 1 Year of Experience in IT-ITeS Sector.

Number of Seats: 40 (Forty) – Total

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: 08th February, 2023

Registration Procedure

All interested candidates are required to fill the Registration form online with registration fees before **08th February, 2023** with all the necessary information.

Selection Criteria of candidates

Selection of candidates will be based on first come first serve basis.

- ✓ The first list of Provisionally Selected Candidates will be published on NIELIT Chennai website (www.nielit.gov.in/chennai) on **10-02-2023** 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:
- ✓ **Required Documents:**
  - 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**

- ✓ **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 installment of Rs. 6000/- on or before **15-02-2023** by payment mode mentionedabove.
- ✓ 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 get a welcome message in his login id provided during registration. The credential and url for online portal will be shared through WhatsApp or email.

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 candidates are not uploading consecutive 3 assignments within assigned time, then their candidature will be cancelled without any notice 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 classes and labs are online cloud based from **4 pm to 6 pm**

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:

1. **Dr. Sanjeev Kumar Jha**, **Mobile: 7765803105**

Course enquiries

Students can enquire about the various courses either on telephone or by personal contact between 9.15 A.M. to 5.15 P.M. (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 counselling will be given to crack their interviews.

Important Dates

- **Last Date of Registration: 08-02-2023**
- **Display of Provisional Selection List: 10-02-2023**
- **Payment of first installment fee: 10-02-2023 to 15-02-2023**
- **Course Start Date: 15-02-2023**
- **Payment of second installment fee: 15-04-2023**

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 / Assignment(Marks)**	**Major Project/ Dissertation**		**Total**
Papers	**No. of Questions/ Paper**	**Papers**	**Duration (in Min)**	**Marks/Paper**			**No. Of Projects**	**Marks**	
2	200	1	90	90	15	15	1	30	350

Theory Papers

- **Theory 1** – Configuring Platform for Data Engineering
- **Theory 2** – Big Data Analytics

Practical Papers

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

	Theory	Practical	Internal Assessment	Project/ Presentation/ Assignment
No of Papers	02	01	01	01
Marks Each Paper	100	90	30	30
Total Marks	200	90	30	30
Grand Total Marks				350

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), MTech (Digital Electronics), MBA (Marketing) and PhD (Strategic Management). He has More than 29 years‟ 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.

Detailed Curriculum

Module 1: Configuring Platform for Data Engineering

Introduction to Virtual Machine:

- Creating and configuring Virtual Machine, Installing Ubuntu Operating System on virtual machine
- Introducing Ubuntu, Installing Ubuntu: Starting Up, Logging in, Exploring the Desktop, Ubuntu Basics.

Understanding Linux Environment

- Operating System Concepts: Linux History, Benefits of Linux, Different Flavors of Linux, Browsing the File System, Understanding File System Concept, Managing Files, Real and Virtual Files, Mounting, File Searches, File Size, File Space Understanding Linux Files/Directories: Viewing Text Files, Using a Command Line Text Editor, Creating Files, Searching through Files, Comparing Text Files, Copying, Moving, Managing Files.

Basic Commands:

- Ubuntu Commands, Running Basic commands, Piping and Filtering Commands, Directory and File handling commands, 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 otherbasic 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 migration, migration types and process.

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 & StringBuffer
- Create and Manipulate Calendar Data
- Declare and Use of Arraylist

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

Application Layer:

HTML5:

- Basic Structure of HTML, Head Section,Formatting Tags, Tables,Attributes,Lists,Frames, HTML5 Introduction, HTML5 New Elements. Form validations.

JavaScript:

- Introduction to Client Side Scripting Language, Variables in Java Script, Operators in JS, Conditional Statements, JS Popup Boxes, JS Events, Basid Form Validation inJava script.

Database Management Systems:

- Introduction to data, data analysis and data analyst. Introduction to database, characteristics of data in database, DBMS, advantages of DBMS, file- oriented approach versus Database-oriented approach to Data Management, disadvantages of file- oriented approach. Fundamental integrity rules: entity integrity, referential integrity.

Datawarehousing using MySQL

Datawarehousing 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.

Module 2: 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)
- HBase Constraints
- Case Study - Log Data and Timeseries Data
- Case Study - Customer/Order
- HBase and MapReduce
- HBase 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,ArrayBuffer,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.