edureka!
Discover Learning

PG Program in Big Data Engineering

About NITR

The National Institute of Technology, Rourkela is one of India’s premier national level institutions for technical education in the country and is ranked 2nd among NITs by MHRD. NITR has also been recognized as an Institute of National Importance by the Government of India. NITR has partnered with Edureka to design this outline Post-Graduate Program in Big Data Engineering in order to develop the next cadre of highly skilled professionals and experts in the field of Big Data.

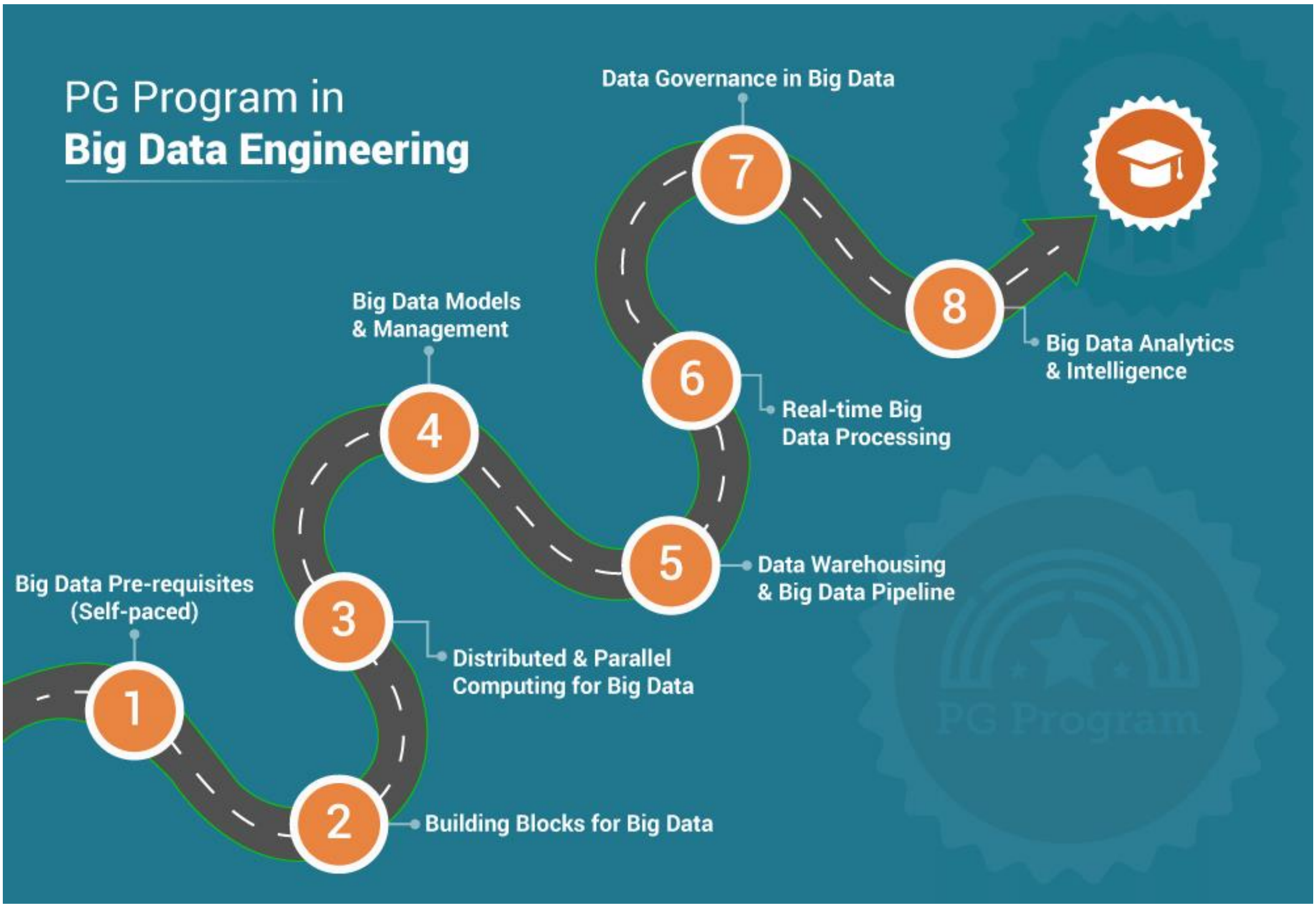
About Edureka

Edureka is an e-learning education platform focused on delivering high-quality learning to Technology professionals. We have the highest course completion rate in the industry, and we strive to create an online ecosystem for our global learners to equip themselves with industry-relevant skills in today’s cutting-edge technologies.

We have an easy and affordable learning solution that is accessible to millions of learners. With our students spread across countries like the US, India, UK, Canada, Singapore, Australia, Middle East, Brazil and many others, we have built a community of over 1 million learners across the globe.

About the Course

The NITR faculty, Edureka Experts, Industry Veterans, a large pool of SMEs have jointly collaborated to create this online PG program in Big Data Engineering. This PGP is a rigorous 9 months program with over 450 hours of intensive learning, designed to make you not just industry ready, but also primed to excel. Also, this program is designed to equip you with the skills you need to rise to the top in a career in the field of Big Data.


PG Program in
Big Data Engineering
Big Data Pre-requisites (Self-paced)
1
2
Building Blocks for Big Data
3
Distributed & Parallel Computing for Big Data
4
Big Data Models & Management
5
Data Warehousing & Big Data Pipeline
6
Real-time Big Data Processing
7
Data Governance in Big Data
8
Big Data Analytics & Intelligence
PG Program

Index

**Depending on industry requirements, Edureka may make changes to the course curriculum*

Big Data Pre-requisites (Self-paced)

Course Curriculum

About the Course

In To become a big data engineer, one should know the fundamentals of operating system, Database and Programming. In this course, you will learn the basics of Linux, SQL and Java which will setup the foundation for this PG program.

Module 1: Linux Fundamentals

Learning Objectives:

In this module, you will learn about Linux fundamentals such as Linux installation, shell scripting, basic and advanced Linux commands, package management, ssh and remote log-in.

Topics:

- What is Linux?
- Linux Distribution
- Ubuntu Installation
- Shell Scripting
- Some basic and advanced Linux commands
- Package Management
- ssh and Remote log-in

Module 2: SQL Essentials

Learning Objectives:

In this Module, you will learn the concepts related to DBMS such as Data modelling, Data import and export, Entity Relationship diagram, Normalization concept. You will also learn on MySQL commands, nested queries, views, triggers, Relational Algebra and Relational Calculus.

Topics:

- MySQL
- Features of MySQL
- MySQL Workbench
- MySQL functionalities
- Three modules of Workbench
- Data Modelling
- Data import and export
- Database and DBMS
- Entity Relationship Diagram
- Normalization
- Different Normal forms
- MySQL commands
- Nested Queries
- Views and Triggers
- Relational Algebra and Calculus
- SQL operations
- Joins in SQL
- Types of Joins

Module 3: Java Essentials for Big Data

Learning Objectives:

In this module, you will learn fundamentals of Java such as data types, operators, control statements, methods, constructors, packages, interfaces, and object-oriented programming concepts. You will also master important concepts such as Exception handling, File input/output operations and Java Collection Framework. Also, you will learn about various parsers such as DOM, SAX, Stax parser in Java.

Topics:

- Introduction to Java
- Java Installation
- Modifiers and Variables in Java
- Java Data Types
- Data Types Conversion in Java
- Java Operators
- Control Statements in Java
- Methods in Java
- Arrays and their Types
- Strings and their Operations
- Classes and Objects
- Constructors and their types
- Static and This Keyword in Java
- Oops Concept in Java
- Interfaces in Java
- Packages in Java
- Exceptions and its types
- Exception Handling in Java
- Throw and throws keyword in Java
- File I/O Operation in Java
- Wrapper Classes in Java
- Java Collection Frameworks
- List and its Classification in Java
- Queue in Java
- Sets and their Classification in Java
- Introduction to XML
- DOM Parser in Java
- SAX Parser
- StAX Parser In Java
- Introduction To XPath
- Introduction To JAXB
- JSON And Its Significance

Building Blocks For big data

Course Curriculum

About the Course

This course covers the introduction to Big Data and various Big Data applications in different domains. In this course, you will learn about some of the important concepts related to Data Structures, algorithms, distributed systems, divide and conquer technique and various computing models such as sequential and parallel computing models. In the end, you will learn about the different logical layers of a Big Data System and you will know about the various tools available in the Big Data ecosystem.

Module 1: Introduction to Big Data, Data Structures and Algorithms

Learning Objectives:

After completing this module, you should be able to:

- ✓ Understand Global Data Explosion
- ✓ Understand 5 V's of Big Data
- ✓ Know how organizations are analyzing Big Data
- ✓ Know who is a Big Data Engineer
- ✓ Learn how to become a Big Data Engineer
- ✓ Understand the basics of Data Structure
- ✓ Know about the Asymptotic Notations
- ✓ Implement LinkedList Operations

Topics:

- ✓ Global Data Explosion
- ✓ Structure of Data
- ✓ What is Big Data?
- ✓ The 5 V's of Big Data
- ✓ Big Data Analytics
- ✓ Big Data Engineer
- ✓ Introduction to Data Structures
- ✓ Algorithms and Asymptotic Notation
- ✓ Linear Data Structures - Linked List

Hands-on:

- ✓ Implement Singly LinkedList and perform various operations such as insertion, deletion and searching

Module 2: Non-linear Data Structures for Big Data

Learning Objectives:

After completing this module, you should be able to:

- ✓ Understand Non-Linear Data Structures
- ✓ Perform operations on trees
- ✓ Know about the types and applications of graphs
- ✓ Write Java Programs using Java Collection Framework
- ✓ Learn how to implement Set and Map Interface

Topics:

- ✓ Non-Linear Data Structures
- ✓ Tree
    - ○ General Trees
    - ○ Binary Trees
    - ○ Binary Search Tree
    - ○ Basic operations on Trees
    - ○ Balancing of Binary search tree
- ✓ AVL Trees
- ✓ Graphs
    - ○ Types of Graphs
    - ○ Applications of Graphs
    - ○ Graphs Traversal Algorithms
- ✓ Hashing
- ✓ Java Collection Framework
- ✓ Set
    - ○ HashSet
    - ○ LinkedHashSet
    - ○ TreeSet
- ✓ Map
    - ○ HashMap
    - ○ LinkedHashMap
    - ○ TreeMap

Hands-on:

- ✓ Implement Binary Search Tree and perform various operations such as insertion, deletion, and searching of a node
- ✓ Implement AVL tree and perform various rotations required to get a balanced BST
- ✓ Implement various classes in the Set interface such as HashSet, LinkedHashSet and TreeSet
- ✓ Implement various classes in Map interface such as HashMap, LinkedHashMap, and TreeMap

Module 3: Designing High Performant Distributed Algorithms

Learning Objectives:

After completing this module, you should be able to:

- Understand what are Distributed Systems?
- Explain the advantages and applications of Distributed System
- Know the characteristics of Distributed Systems
- Explain what are Clusters?
- Understand the Top-Down Approach to algorithm design
- Learn various computing models
- Identify sorting algorithms in a Distributed Environment

Topics:

- What is a Distributed Algorithm?
- What is a Distributed System?
- Advantages of distributed systems
- Applications of Distributed Systems
- Characteristics of Distributed Systems
- Challenges of Distributed Systems
- What is a cluster?
- The top-down approach to algorithm design
- Divide and Conquer algorithms
    - Binary Search
    - Quick Sort
    - Merge Sort
- Sequential Computing Models
    - Random Access Machine Model
- Parallel Computing Models
    - Shared Memory Model
    - Distributed Memory Model
- Sorting in Distributed Systems
    - Sample Sort

Hands-on:

- Implement Binary search algorithm to search an element
- Implement Quicksort to sort an array
- Implement Mergesort to sort an array
- Implement Binary search algorithm with Quicksort

Module 4: Tool Stack Options for Big Data

Learning Objectives:

After completing this module, you should be able to:

- Understand the Hadoop-based Big Data platform
- Learn how to design a logical layers Big Data System
- Know the components of Hadoop Ecosystem
- Learn about various Big Data Technologies
- Analyze the Uber use case of Big Data

Topics:

- ✓ Big Data problems and solutions
- ✓ Hadoop-based Big Data platform
- ✓ Logical layers of Big Data System
- ✓ Hadoop Ecosystem and its component
- ✓ Big Data Technologies
- ✓ Uber Data Platform (Use-Case)

Distributed & Parallel Computing for Big Data

Course Curriculum

About the Course

Distributed and Parallel Computing forms the basis of Big Data. This course, on Distributed & Parallel Computing for Big Data will help you to gain in-depth knowledge on Big Data, Hadoop Distributed File System, MapReduce. You will also learn about Scala programming language, in-memory computation using Spark and RDDs, a fundamental data structure of Spark.

Module 1: Background of Distributed & Parallel Computing Systems

Learning Objectives:

At the end of this module, you should be able to:

- ✓ Understand what is Distributed and Parallel Computing?
- ✓ Know about the properties of a Distributed System
- ✓ Understand Google File System
- ✓ Learn about Remote Procedure Call, Java RMI, Logical Clocks
- ✓ Understand Map and Reduce construct
- ✓ Know about various applications of Map & Reduce construct
- ✓ Evaluate the performance of Map Construct & Reduce Construct
- ✓ Understand Iterative MapReduce
- ✓ Know the difference between Thread vs Process
- ✓ Implement multithreading in Java

Topics:

- ✓ Why Distributed and Parallel Computing
- ✓ Properties of Distributed System
- ✓ Google File System
- ✓ Remote Procedure Call
- ✓ Java RMI
- ✓ Logical Clocks
- ✓ Map Construct and its example
- ✓ Applications of Map & Reduce Construct
- ✓ Map & Reduce Construct use-case
- ✓ Performance Evaluation of Map Construct & Reduce Construct
- ✓ Iterative Map Reduce
- ✓ Thread vs Process
- ✓ Multithreading in Java

Hands-on:

- ✓ Implement Multithreading in Java

Module 2: Hadoop – A Distributed and Highly Scalable Computing Platform

Learning Objective

At the end of this module, you should be able to:

- Understand the Distributed File System (DFS)
- Limitations & Solutions of Big Data Architecture
- Explain what is Hadoop?
- Hadoop & its Features
- Know about various tools in the Hadoop Ecosystem
- Learn about various modes in Hadoop
- Learn about Hadoop daemons
- Analyse a typical Hadoop cluster configuration

Topics:

- Distributed File System (DFS)
- What is Hadoop?
- Hadoop Ecosystem
- Hadoop cluster modes
- Hadoop Daemons
- Typical Hadoop Cluster Configuration

Module 3: Storing Big Data in a Distributed Cluster - HDFS

Learning Objectives:

At the end of this module, you should be able to:

- Explain Hadoop Distributed File System
- Explain core components of Hadoop 2.x
- Explain YARN
- Understand File blocks and Rack Awareness concepts
- Explain the Read & Write mechanism in Hadoop
- Work with various Hadoop terminal commands

Topics:

- HDFS
- Hadoop 2.x core components
- YARN
- File Blocks
- Rack Awareness
- Hadoop read and write mechanism
- Hadoop 2.x cluster architecture
- Hadoop terminal commands

Hands-on:

- Run the HDFS Commands

Module 4: Advanced HDFS Concepts

Learning Objectives:

At the end of this module, you should be able to:

- Understand Hadoop 2.x Cluster Architecture – Federation
- Understand Hadoop 2.x cluster architecture-High availability and resource management
- Configure various Hadoop 2.x configuration files
- Work with Hadoop Web UI parts
- Access the files using HDFS APIs

Topics:

- Hadoop 2.x cluster architecture-Federation
- Hadoop 2.x cluster architecture-High availability and resource management
- Hadoop 2.x configuration files
- Hadoop Web UI parts
- Listing Files Using HDFS APIs
- Access HDFS using JAVA API (Read/Write Operations)

Hands-on:

- Implement various HDFS operations using Java APIs

Module 5: MapReduce Computational Model

Learning Objectives:

At the end of this module, you should be able to:

- Explain MapReduce Computational Model
- Understand the different stages of the MapReduce computational model such as Mapper, Reducer, Combiner
- Know about YARN MR Application Execution Flow
- Know about different Input and Output Formats available in the MapReduce Framework
- Explain the Serialization concept in MapReduce
- Implement MapReduce Writable Interface
- Implement a MapReduce Program and run it

Topics:

- What is MapReduce?
- Where is MapReduce used?
- Difference between traditional and MapReduce way
- Hadoop 2.x MapReduce Architecture
- Stages involved in the MapReduce computation model – Mapper, Sorting & Shuffling, Reducer, Combiner
- YARN MR Application Execution Flow
- MapReduce Paradigm
- Input Splits and HDFS blocks
- MapReduce Job Submission Flow
- Various Input and Output Formats
- Serialization in the MapReduce
- Hadoop Writable Interface
- Write a MapReduce Program

Hands-on:

- Implement MapReduce programs
- Implement Combiner in a MapReduce program

Module 6: Advanced MapReduce Concepts

Learning Objectives:

At the end of this module, you should be able to:

- Understand the role of Partitioner in MapReduce
- Implement Custom Input Format in MapReduce
- Implement Counters in MapReduce
- Know how to make use of Distributed Cache in MapReduce Program
- Perform join operation in MapReduce

Topics:

- Partitioner
- Multiple Inputs
- Custom Input Format
- Counters
- Distributed Cache
- Joins
- Map-side Join
- Reduce-side Join
- Sequence File

Hands-on:

- Analysing the datasets using Map-side, Reduce-Side join, Distributed Cache

Module 7: Scala Essentials for Spark - I

Learning Objectives:

At the end of this module, you should be able to:

- Understand what is Scala?
- Relate why Spark choose Scala?
- Implement basic Scala Operations
- Understand data types in Scala
- Implement variable types in Scala
- Use control structures in Scala
- Implement functions in Scala
- Use Collections in Scala

Topics:

- Introduction to Scala
- Scala for Spark
- Scala framework
- Scala Operators
- Data types
- Variables
- Control Statements
- Different types of loops
- Functions
- Collections in Scala – Array and Array Buffer, Lists, Tuples, Sets, Maps

Hands-on:

- Implement Control Statements in Scala
- Implement different types of loops
- Implement functions in Scala
- Run the Scala Programs

Module 8: Scala Essentials for Spark - II

Learning Objectives:

At the end of this module, you should be able to:

- Create Classes in Scala
- Understand Getters and Setters
- Implement Constructors
- Describe Singletons
- Describe Companion Objects
- Implement Inheritance in Scala
- Implement Traits
- Understand Layered Traits
- Implement Higher-Order Functions and Anonymous Functions

Topics:

- Classes and Objects in Scala
- Getters and Setters
- Singleton Objects
- Companion Objects
- Inheritance in Scala
- Traits in Scala
- Anonymous Functions in Scala
- Higher-order functions in Scala

Hands-on:

- Implement classes and objects in Scala
- Implement Traits in Scala
- Implement the Higher-order and Anonymous functions in Scala
- Implement Scala Programs

Module 9: In-memory Computation for Big Data

Learning Objectives:

At the end of this module, you should be able to:

- The challenges in existing Computing methods
- What is RDD, its functions, Transformations
- Perform data loading and saving through RDDs
- View RDD Lineage and RDD Persistence
- Write Spark Programs using RDD concepts
- RDD Partitioning and how it helps to achieve parallelism

Topics:

- ✓ Computing Models
- ✓ What is RDD?
- ✓ Features of RDDs
- ✓ Ways to create RDDs
- ✓ RDD workflow
- ✓ RDD operations- transformations and actions
- ✓ Data loading and saving using RDDs
- ✓ General RDDs functions

Hands-on:

- ✓ Perform various actions, transformations, and operations on RDDs

Module 10: Advance RDDs Concepts in Spark

Learning Objectives:

At the end of this module, you should be able to:

- ✓ Implement Key-value pair RDD
- ✓ Implement Other Pair RDD functions
- ✓ Perform joins in RDD
- ✓ Pass functions as a parameter in Spark
- ✓ Explain shared variables- broadcast variables and accumulators

Topics:

- ✓ Key-value pair RDD
- ✓ Other Pair RDD functions
- ✓ Double RDD functions
- ✓ RDD joins
- ✓ Passing functions as a parameter in Spark
- ✓ Shared variables- broadcast variables and accumulators

Hands-on:

- ✓ Implement Spark programs using RDDs concepts

Big Data Models & Management

Course Curriculum

About the Course

Big Data Models & Management course will help you master relational as well as non-relational databases. You will start this course by learning Relational Database Management System concepts & querying relational database using SQL. Moving ahead, you will learn Hive Query Language & Spark SQL. At the end of this course, you will master NoSQL databases such as HBase & Cassandra.

Module 1: Introduction to Relational Database Management System

Learning Objectives:

This model will help you to learn about how to design a database, ER diagrams, UML diagrams. You will also learn about various normal forms, files and indexing concept in relational databases.

Topics:

- Relational Databases
- Database Design
- Database Storage and Querying
- Database Schema
- Relational Query Language
- Relational Operations
- Overview of the design process
- Entity-Relationship model
- UML (Sequence Diagram)
- Features of good relational design
- Atomic Domains and Normalization
- Files Concept
- Indexes Concept
- Measuring of Query Cost

Hands-on:

- Designing Entity-Relationship model
- Designing UML (Sequence Diagram)
- Designing Database Schema

Module 2: Storing, Handling & Managing Data in Relational Database

Learning Objectives:

This module will help you to work on advanced MySQL concepts such as join operation, nested query, creating views and procedures. You will also learn about how to load and write data from/into comma-delimited file.

Topics:

- Selection Operations
- Sorting
- Joins Operation
- Evaluation of Expressions
- Functions and their types
- Nested Query
- Views
- Procedures
- Cursors
- Triggers
- Transactions
- Recovery Algorithms
- ACID properties
- Locks
- Loading data from comma-delimited files
- Writing data into comma-delimited files

Hands-on:

- Implement nested queries, views, triggers in SQL
- Loading data from comma-delimited files
- Writing data into comma-delimited files

Module 3: SQL Like Query Processing Engine for Big Data - Hive

Learning Objectives:

This module will help you to understand Hive concepts, Hive Data types, partition and bucketing in Hive, internal and external tables, loading and querying data in Hive, running hive scripts.

Topics:

- Introduction to Apache Hive
- Hive Architecture and Components
- Hive Metastore
- Comparison with Traditional Database
- Hive Data Types and Data Models
- Hive Partition
- Hive Bucketing
- Hive Tables (Managed Tables and External Tables)
- Importing Data
- Querying Data & Managing Outputs

Hands-on:

- Importing data in Hive
- Querying Data and managing outputs

Module 4: Advanced Hive Concepts

Learning Objectives:

This model will help to work on some advanced concepts in Hive such as how to join tables, Hive Indexes and views, Hive UDF, Hive query optimizers and Hive Thrift server.

Topics:

- Hive Script
- Retail use case in Hive
- Hive QL: Joining Tables, Dynamic Partitioning
- Custom MapReduce Scripts
- Hive Indexes and views
- Hive Query Optimizers
- Hive Thrift Server
- Hive UDF

Hands-on:

- ✓ Running Hive scripts
- ✓ Creating Hive UDFs
- ✓ Joining tables in Hive

Module 5: Relational Data Processing with Spark - SparkSQL & DataFrames

Learning Objectives:

This model will help you to learn about SparkSQL which is used to process structured data with SQL queries. You will be learning about various concepts in Spark SQL such as data-frames and datasets along with different kinds of SQL operations performed on the data-frames.

Topics:

- ✓ Need for Spark SQL
- ✓ What is Spark SQL?
- ✓ Spark SQL Architecture
- ✓ SQL Context in Spark SQL
- ✓ User-defined Functions
- ✓ Data Frames & Datasets
- ✓ Interoperating with RDDs
- ✓ JSON and Parquet File Formats
- ✓ Loading Data through Different Sources

Hands-on:

- ✓ Load data through different sources
- ✓ Creating Data Frames and Datasets
- ✓ Implement SQL queries on DataFrames and Datasets

Module 6: Handling Heterogeneous Data with NoSQL

Learning Objectives:

This model will help you to learn about how to handle heterogeneous data using HBase. You will get to know about the fundamentals of NoSQL database, CAP theorem and intro to HBase.

Topics:

- ✓ Database categories: What is NoSQL?
- ✓ Why NoSQL?
- ✓ Benefit over RDBMS
- ✓ Types of NoSQL Database
- ✓ NoSQL vs. SQL Comparison
- ✓ Column-oriented vs Row-oriented
- ✓ ACID & Base Property
- ✓ CAP Theorem, implementing NoSQL
- ✓ How NoSQL helps in handling heterogeneous data?
- ✓ Different NoSQL databases and their comparison
- ✓ Introduction to HBase and its architecture
- ✓ HBase roles (master, server, etc)

Module 7: Exploring Column Based NoSQL Database - HBase

Learning Objectives:

This model will help you to learn about advance Apache HBase concepts. You will find demos on HBase Bulk Loading & HBase Filters. You will also learn what Zookeeper is all about, how it helps in monitoring a cluster & why HBase uses Zookeeper.

Topics:

- HBase v/s RDBMS
- HBase Components
- HBase Architecture
- HBase Run Modes
- HBase Configuration
- HBase Cluster Deployment
- HBase Data Model
- HBase Shell
- HBase Client API
- HBase Data Loading Techniques
- Apache Zookeeper Introduction
- ZooKeeper Data Model
- Zookeeper Service
- HBase Bulk Loading
- Getting and Inserting Data
- HBase Filters

Hands-on:

- Working on HBase Data loading, bulk loading and HBase Filters

Module 8: Working with NoSQL Database Cluster - Apache Cassandra

Learning Objectives:

This model will help you to learn about Database Model and similarities between RDBMS and Cassandra Data Model. You will also understand the key Database Elements of Cassandra and learn about the concept of Primary Key.

Topics:

- NoSQL databases
- Common characteristics of NoSQL databases
- CAP theorem
- How Cassandra solves the Limitations?
- History of Cassandra
- Features of Cassandra
- Introduction to Database Model
- Types of Data Models
- Understand the analogy between RDBMS and Cassandra Data Model
- Understand following Database Elements: Cluster, Keyspace, Column Family/Table, Column
- Column Family Options
- Columns
- Wide Rows, Skinny Rows
- Static and dynamic tables

Hands-on:

- Create Keyspace in Cassandra
- Creating Tables

Module 9: Deep Diving into Apache Cassandra

Learning Objectives:

This model will help you to learn about the complex inner workings of Cassandra such as Gossip Protocol, Read Repairs and so on. In addition, you will learn about Keyspace and its attributes in Cassandra. You will also create Keyspace, learn how to create a Table and perform operations like Inserting, Updating and Deleting data from a table while using CQLSH.

Topics:

- ✓ Cassandra Architecture
- ✓ Different Layers of Cassandra Architecture
- ✓ Gossip Protocol
- ✓ Partitioning and Snitches
- ✓ Vnodes and How Read and Write Path works
- ✓ Understand Compaction
- ✓ Anti-Entropy and Tombstone
- ✓ Repairs in Cassandra
- ✓ Hinted Handoff
- ✓ Replication Factor, Replication Strategy
- ✓ Different Data Types Used in Cassandra
- ✓ Collection Types
- ✓ What are CRUD Operations
- ✓ Insert, Select, Update and Delete of various elements
- ✓ Various Functions Used in Cassandra
- ✓ Importance of Roles and Indexing
- ✓ Tombstones in Cassandra

Hands-on:

- ✓ Check Created Keyspace in System_Schema.Keyspaces
- ✓ Update Replication Factor of Previously Created Keyspace
- ✓ Drop Previously Created Keyspace
- ✓ Create A Table Using cqlsh
- ✓ Create A Table Using UUID & TIMEUUID
- ✓ Create A Table Using Collection & UDT Column
- ✓ Create Secondary Index On a Table
- ✓ Insert Data into Table
- ✓ Insert Data into Table with UUID & TIMEUUID Columns
- ✓ Insert Data Using COPY Command
- ✓ Deleting Data from Table

Data Warehousing & Big Data Pipeline

Course Curriculum

About the Course

Data Warehousing & Big Data Pipeline course helps you understand how to perform ETL on raw data & design a data warehouse. In this course, you'll master Data Warehousing concepts, importing & exporting data using Sqoop, ETL concepts using Hive & Spark SQL. At the end of this course, you'll learn how to create an end-to-end workflow & schedule jobs using Apache Oozie.

Module 1: Introduction to Data Warehousing & Data Mining

Learning Objectives:

This module will help you to understand the fundamentals of data mining & data warehousing. You'll learn how to perform data pre-processing, data cleaning, data integration & data transformation. So, that moving ahead you'll be able to design & build an end-to-end Data Warehouse.

Topics:

- ✓ Why Data Mining?
- ✓ What Is Data Mining?
- ✓ What Kinds of Data Can Be Mined?
- ✓ What Kinds of Patterns Can Be Mined?
- ✓ Which Technologies Are Used?
- ✓ Which Kinds of Applications Are Targeted?
- ✓ Major Issues in Data Mining
- ✓ Data Objects and Attribute Types
- ✓ Basic Statistical Descriptions of Data
- ✓ Data Visualization
- ✓ Measuring Data Similarity and Dissimilarity
- ✓ Data Pre-processing: An Overview
- ✓ Data Cleaning
- ✓ Data Integration
- ✓ Data Reduction
- ✓ Data Transformation and Data Discretization
- ✓ Data Warehouse: Basic Concepts
- ✓ Data Warehouse Modelling: Data Cube and OLAP
- ✓ Data Warehouse Design and Usage
- ✓ Data Warehouse Implementation

Hands-on:

- ✓ Designing the complete ETL Pipeline
- ✓ Designing a data warehouse
- ✓ Creating a data warehouse

Module 2: Large Scale Data Ingestion Using Sqoop

Learning Objectives:

This module will help you to learn how to use Sqoop for importing & exporting data from relational database to Hadoop Distributed File System. It will give you an in-depth knowledge about Sqoop commands & Incremental import.

Topics:

- Why Sqoop?
- What is Hadoop?
- Sqoop Architecture
- How Sqoop Import and Export works?
- Sqoop Commands
- Importing Data
- Incremental Import
- Free-form Query Import
- Generated Code
- Importing Data in Hive
- Importing Large Objects
- Importing Data into HBase
- Export

Hands-on:

- Importing data from Relational Database to HDFS using Sqoop
- Exporting data from HDFS to Relational Database using Sqoop
- Importing data incrementally using Sqoop
- Importing data in Hive from Relational Database using Sqoop
- Importing data from Relational Database to HBase using Sqoop

Module 3: Performing ETL Using Hive

Learning Objectives:

This module will help you to learn how to perform ETL using Hive. It covers in-depth knowledge about ETL concepts such as Data Manipulation, aggregation, sampling, grouping, design optimization & data optimization, and helps you understand how it can be implemented using Hive.

Topics:

- Data Manipulation
- Data Aggregation and Sampling
- Basic aggregation
- Enhanced aggregation
- Grouping sets
- Rollup and Cube
- Aggregation condition
- Window functions
- Sampling
- Performance utilities
- Design optimization
- Use skewed/temporary tables
- Data optimization
- Compression
- Storage & Job optimization
- Parallel execution
- Join optimization
- Common join
- Map join
- Bucket map join
- Sort merge bucket (SMB) join
- Sort merge bucket map (SMBM) join
- Skew join
- Optimizer

Hands-on:

- Performing ETL using Hive
- Aggregating and sampling data using Hive
- Performing Grouping operations using Hive
- Performing bucketing & partitioning in Hive

Module 4: Building Robust ETL Pipeline with Spark

Learning Objectives:

This module will help you to learn how to build a robust ETL pipeline using Apache Spark. This module also covers advanced topics such as partitioning, Columnar Storage, UDFs, UDAFs & interactive analysis.

Topics:

- Data Processing Interface
- Hive Interoperability
- Performance
- Reduced Disk I/O
- Partitioning
- Columnar Storage
- In-Memory Columnar Caching
- Skip Rows
- Predicate Pushdown
- Query Optimization
- ETL (Extract Transform Load)
- Distributed JDBC/ODBC SQL Query Engine
- Data Warehousing using Spark
- Processing Data Programmatically with SQL/HiveQL
- UDFs and UDAFs
- Interactive Analysis
- Interactive Analysis with Spark SQL JDBC Server

Hands-on:

- Performing ETL using Spark
- Perform partitioning in Spark
- Perform in-memory caching in Spark
- Integrating Hive & Spark

Module 5: Workflow and Job Scheduling – I

Learning Objectives:

This module primarily focuses on designing a workflow. It gives you an idea about the considerations that you need to keep in mind while designing a workflow. This module will help you understand the concepts related to workflow & scheduling such as different types of schedulers, resource allocation & pre-emption.

Topics:

- The FIFO Scheduler
- The Capacity Scheduler
- Queues and Sub queues
- How the Cluster Allocates Resources
- Pre-empting Applications
- Enabling the Capacity Scheduler
- A Typical Capacity Scheduler
- The Fair Scheduler
- Queues
- Configuring the Fair Scheduler
- How Jobs Are Placed into Queues
- Application Pre-emption in the Fair Scheduler
- Security and Resource Pools
- A Sample fair-scheduler.xml File
- Submitting Jobs to the Scheduler
- Moving Applications between Queues
- Monitoring the Fair Scheduler
- Comparing the Capacity Scheduler and the Fair Scheduler
- Similarities between the Two Schedulers
- Differences between the Two Schedulers

Module 6: Workflow and Job Scheduling – II

Learning Objectives:

This module will help you to learn how Apache Oozie is used to schedule jobs & create a workflow. This module covers in-depth knowledge about Oozie workflows, Oozie coordinator and monitoring Oozie jobs.

Topics:

- ✓ What is Oozie?
- ✓ Scheduling with Oozie
- ✓ Running Oozie Applications
- ✓ Oozie Workflows
- ✓ Monitoring Oozie Workflow Job
- ✓ Oozie Coordinators
- ✓ Oozie Application Lifecycle
- ✓ Time and Data Trigger Coordinators
- ✓ Monitoring an Oozie Coordinator Job
- ✓ Oozie Bundles Parameters
- ✓ Variables and Functions
- ✓ Application Deployment Model,
- ✓ Oozie Architecture

Hands-on:

- ✓ Scheduling Job with Oozie
- ✓ Creating Oozie workflow
- ✓ Working with the Oozie Co-Ordinator

Real-time Big Data Processing

Course Curriculum

About the Course

Processing data in real-time to get instantaneous results from input data is the major requirement of enterprises in today's world. This course will help you in understanding how you can process the data in real-time by using Spark Streaming APIs such as DStream and Structured Streaming API. In addition, you will also design an end-to-end data pipeline using various tools present in Big Data Ecosystem.

Module 1: Working with Streaming Data

Learning Objectives:

This module will help you to learn how to process streaming data in real-time. This module will cover basic concepts related to stream processing.

Topics:

- What Is Stream Processing?
- Stream Processing Use Cases
- Advantages of Stream Processing
- Challenges of Stream Processing
- Stream Processing Design Points
- Record-at-a-Time Versus Declarative APIs
- Event Time Versus Processing Time
- Continuous Versus Micro-Batch Execution

Module 2: Streaming Data Processing with Spark

Learning Objectives:

This module will help you to work on Spark streaming APIs to process the real-time data coming from various sources. You will get to know about the various concepts such as stateful processing, Windows on Event Time, Tumbling Windows, Handling Late Data with Watermarks, Dropping Duplicates in a Stream and streaming data sources.

Topics:

- Spark Streaming APIs - The DStream API
- Event Time
- Stateful Processing
- Dropping Duplicates in a Stream
- Arbitrary Stateful Processing
- Time-Outs

- Arbitrary Stateful Processing
- Event-Time Basics
- Windows on Event Time
- Tumbling Windows
- Handling Late Data with Watermarks
- Output Modes
- Spark Data Sources
- Flume and Spark Streaming Integration
- Sentiment Analysis using Spark Streaming

Hands-on:

- Implement Flume and Spark Streaming Integration
- Implement Sentiment Analysis using Spark Streaming

Module 3: Structured Streaming in Spark

Learning Objectives:

This module will help you to learn about Structured Streaming, which is a new streaming API, introduced in spark 2.0. You will get to know how you can express your streaming computation the same way you would express a batch computation on static data.

Topics:

- Structured Streaming Basics
- Transformations and Actions
- Input Sources
- Sinks
- Output Modes
- Triggers
- Event-Time Processing
- Structured Streaming in Action
- Transformations on Streams
- Selections and Filtering
- Aggregations
- Joins
- Input and Output
- Where Data Is Read and Written (Sources and Sinks)
- Reading from the Kafka Source
- Writing to the Kafka Sink
- How Data Is Output (Output Modes)
- When Data Is Output (Triggers)
- Streaming Dataset API

Hands-on:

- Process streaming data where Kafka is a data source
- Streaming Dataset API

Module 4: Low Latency and Real Time Data Processing with Kafka – I

Learning Objectives:

This module will help you to understand where Kafka fits in the Big Data space. You will learn about Kafka components, Kafka cluster architecture and how to send and consume messages. At the end, you will learn about Kafka producers which send records to topics.

Topics:

- Publish/Subscribe Messaging
- What is Kafka?
- Topics and Partitions
- Producers and Consumers
- Brokers and Clusters
- Kafka Producers: Writing Messages to Kafka
- Sending a Message to Kafka
- Configuring Producers
- Serializers
- Partitions

Hands-on:

- Sending messages from Producer to Consumer using CLI
- Sending messages from Producer to Consumer using Java APIs

Module 5: Low Latency and Real Time Data Processing with Kafka – II

Learning Objectives:

This module will help you to learn about Kafka consumers and various Kafka APIs such as Kafka Connect & Kafka Stream API.

Topics:

- Kafka Consumers: Reading Data from Kafka
- Consumers and Consumer Groups
- The Poll Loop
- Commits and Offsets
- Deserializers
- Standalone Consumer
- Building Data Pipelines
- Data Formats
- Transformations
- Kafka Connect
- Kafka Stream API

Hands-on:

- Running Kafka Connect and Kafka Stream API

Module 6: Building Streaming Data Pipeline Using Kafka, Spark & NoSQL

Learning Objectives:

This module will help you to create an end-to-end data pipeline using Kafka, Spark & NoSQL.

Topics:

- End-to-end data pipeline using Kafka, Spark & NoSQL

Hands-on:

- Building an end-to-end data pipeline using Kafka, Spark and NoSQL

Data Governance in Big Data

Course Curriculum

About the Course

This course will help you to understand the important concepts of Data Governance: Data Access, Data Integrity & Access Policies. Also, you will learn how to govern data in Big Data Platform.

Module 1: Maintaining Data Integrity in Big Data Platforms

Learning Objectives:

This module will help you to understand what Data Governance is and how to govern data in Big Data Solutions. You'll get to know various important Data Governance concepts such as Metadata management, Data Enrichment & Enhancement, Data Classification, Data Lineage & Impact Analysis.

Topics:

- Data Governance for Big Data
- Big Data Oversight: Five Key Concepts
- Big Data Governance Framework
- Identifying the critical dimensions of data quality
- Cloudera Navigator
- Metadata Management
- Consistency of Metadata and Reference Data for Entity Extraction
- Data Enrichment and Enhancement
- Data Classification
- Data Lineage and Impact Analysis

Module 2: Access Policies in Big Data Platforms

Learning Objectives:

This module will help you to understand data security concepts in Big Data platforms such as Auditing, Access Control, Policy Enforcement. You will also learn how to automate data lifecycle.

Topics:

- Auditing and Access Control
- Policy Enforcement and Data Lifecycle Automation

Module 3: Security Concepts in Hadoop – I

Learning Objectives:

This module will help you to learn different aspects of security such as authentication & authorization. You will also understand how to ensure security in the Hadoop platform using Kerberos. Advancing ahead in this module, you'll master how to setup Kerberos & how to secure a Hadoop cluster using Kerberos.

Topics:

- ✓ Hadoop Security — An Overview
- ✓ Authentication, Authorization, and Accounting
- ✓ Hadoop Authentication with Kerberos Kerberos and How It Works
- ✓ The Kerberos Authentication Process
- ✓ Kerberos Trusts
- ✓ A Special Principal
- ✓ Adding Kerberos Authorization to your Cluster
- ✓ Setting Up Kerberos for Hadoop
- ✓ Securing a Hadoop Cluster with Kerberos

Module 4: Security Concepts in Hadoop – II

Learning Objectives:

This module will help you to understand more ways of ensuring security in Big Data platforms such as HDFS permissions, Service Level Authorization & HDFS Transparent Encryption.

Topics:

- ✓ How Kerberos Authenticates Users and Services
- ✓ Managing a Kerberized Hadoop Cluster
- ✓ Hadoop Authorization
- ✓ HDFS Permissions
- ✓ Service Level Authorization
- ✓ Auditing Hadoop
- ✓ Auditing HDFS Operations
- ✓ Auditing YARN Operations
- ✓ Securing Hadoop Data
- ✓ HDFS Transparent Encryption
- ✓ Encrypting Data in Transition

Big Data Analytics & Intelligence

Course Curriculum

About the Course

In previous courses, we have discussed how to store and process big data. But one cannot make the strategic decisions by processing the data only. So, the final need is to analyze this data for gaining insights that can help in making strategic decisions. In today's world, properly leveraged data can give organizations with all types a competitive advantage. This course will help you understand some of the important concepts of big data analytics.

Module 1: Introduction to Big Data Analytics

Learning Objectives:

This module will help you to understand the complex process of examining large and varied data sets to uncover information including hidden patterns, unknown correlations and market trends. In this module, you will learn about Big Data analytics concepts such as data analytics life cycle, Pre-processing data, Hypothesis and Modeling, Measuring the effectiveness, Making improvements.

Topics:

- Introduction to Big Data Analytics
- Big Data Analytics Project Life Cycle
- Identifying the problem and outcomes
- Data collection
- Pre-processing data and ETL
- Performing analytics
- Visualizing data
- Data Science project life cycle
- Hypothesis and Modeling
- Measuring the effectiveness
- Making improvements
- Communicating the results

Hands-on:

- Pre-processing data and performing ETL

Module 2: Business Intelligence & Visualization for Big Data

Learning Objectives:

This module will help you to get a brief idea of data visualization using Tableau. In addition, you will learn how to transform data into interactive dashboards and analyze things visually.

Topics:

- ✓ What is Data Visualization?
- ✓ Why Data Visualization?
- ✓ Different BI tools in the market
- ✓ Tableau Desktop/Tableau Public UI
- ✓ Basic Charts and Graphs
- ✓ Data blending
- ✓ LOD
- ✓ Functions
- ✓ Reader, Public, Server
- ✓ Creating Basic Graphs in Tableau
- ✓ Understanding the concept of:
- ✓ Sorting
- ✓ Calculation
- ✓ Parameter
- ✓ Filtering
- ✓ Forecasting
- ✓ Trend lines and reference line
- ✓ Creating a Dashboard and Stories
- ✓ Publishing Dashboard

Hands-on:

- ✓ Visualizing data and creating dashboards

Module 3: Advanced Analytics Using ML – I

Learning Objectives:

This module will help you to learn the concept of Machine Learning and its types. In addition, you will get to know about MLlib which is Apache Spark's scalable machine learning library.

Topics:

- ✓ Machine Learning Overview
- ✓ Machine Learning Terminologies
- ✓ Machine Learning Types - Supervised, Unsupervised, Reinforcement Learning
- ✓ Machine Learning Process
- ✓ Spark Machine Learning Library
- ✓ Machine Learning Pipelines
- ✓ Transformers and Estimators

Module 4: Advanced Analytics Using ML – II

Learning Objectives:

This module will help you to work on various machine learning algorithms and utilities using Spark MLlib.

Topics:

- ✓ Working with Vectors
- ✓ Algorithms
- ✓ Feature Extraction
- ✓ Statistics
- ✓ Classification and Regression
- ✓ Clustering
- ✓ Collaborative Filtering and Recommendation

- Dimensionality Reduction
- Model Evaluation
- Tips and Performance Considerations
- Preparing Features
- Configuring Algorithms
- Caching RDDs to Reuse
- Recognizing Sparsity
- Level of Parallelism
- Pipeline API

Hands-on:

- Analyzing data using various ML algorithms such as clustering, classification, regression and decision tree