



BIG DATA

Big Data Overview

Learning Objective

- ◆ Learning Linux

Topics Covered

- Introduction/Installation of Virtual Box and the Big Data VM
- Introduction to Linux
- Why Linux?
- Windows and the Linux equivalents
- Different flavors of Linux - Unity Shell (Ubuntu UI)
- Linux Commands
- Shell Scripting

Core Java

Learning Objective

- ◆ Learning core java for big data

Topics Covered:

- Java programming fundamentals:**
 - Data types and Operations

- ❑ if conditions, Loops – for, while and do while
- ❑ **Data Handling and Functions:**
 - ❑ Arrays - Single Dimensional and Multidimensional Arrays
 - ❑ Functions, Function with Arguments, Function Overloading
 - ❑ The concept of Static Polymorphism
 - ❑ String Handling - String, StringBuffer Classes
- ❑ **Object Oriented Programming in Java:**
 - ❑ Concept of Object Orientation
 - ❑ Attributes and Methods
 - ❑ Classes and Objects
 - ❑ Methods and Constructors
 - ❑ Default Constructors and Constructors with Arguments
 - ❑ Inheritance, Abstract, Final, Static
 - ❑ Packages

SQL

Learning Objective

- ❖ **Learn the fundamentals of databases and extract information from RDBMS using the structured query language.**

Topics Covered

- ❑ Introduction to RDBMS
- ❑ Installing mysql
- ❑ Retrieving
- ❑ Updating
- ❑ Inserting
- ❑ Deleting
- ❑ Sorting AND Filtering
- ❑ Summarizing AND Grouping

- ❑ Using Subqueries
- ❑ Joining Tables
- ❑ Views
- ❑ Stored Procedure

Python for Big data

Learning Objective

- ❖ **Learning and Building** a foundation for the most in-demand programming language of the 21st century

Topics Covered

- ❑ Python programming
- ❑ Environment Setup
- ❑ Jupyter Notebook Overview
- ❑ Data types:Numbers,Strings,Printing,Lists,Dictionaryes
- ❑ Booleans,Tuples ,Sets
- ❑ Comparison Operators
- ❑ if, elif, else Statements
- ❑ Loops: for Loops, while Loops
- ❑ range()
- ❑ list comprehension
- ❑ functions
- ❑ lambda expressions
- ❑ map and filter
- ❑ methods
- ❑ Programming Exercises
- ❑ Modules and packages
- ❑ Errors and Exception Handling
- ❑ Python for Exploratory Data Analysis:Pandas

- ❑ Introduction to Pandas
- ❑ Series
- ❑ Data Frames
- ❑ Missing Data
- ❑ GroupBy
- ❑ Merging, Joining and Concatenating
- ❑ Operations
- ❑ Data Input and Output

Note :All topics are delivered as Hands-On sessions.

HADOOP

Learning Objective

- ❖ **Learning Hadoop and its Architecture**

Topics Covered

- ❑ Introduction to Big data and Hadoop
 - ❑ What is Big Data?
 - ❑ What are the challenges for processing big data?
 - ❑ What is Hadoop?
 - ❑ Why Hadoop?
 - ❑ History of Hadoop
 - ❑ Hadoop ecosystem
 - ❑ HDFS
 - ❑ MapReduce
- ❑ Understanding the Cluster
 - ❑ Hadoop 2.x Architecture
 - ❑ Typical workflow
 - ❑ HDFS Commands
 - ❑ Writing files to HDFS
 - ❑ Reading files from HDFS
 - ❑ Rack awareness
 - ❑ Hadoop daemons

- ❑ MapReduce
 - ❑ MapReduce overview
 - ❑ Word count problem
 - ❑ Word count flow and solution
 - ❑ MapReduce flow Typical workflow

- ❑ Developing MapReduce Application
 - ❑ Data Types
 - ❑ File Formats
 - ❑ Explain the Driver, Mapper and Reducer code
 - ❑ Configuring development environment - Eclipse
 - ❑ Writing unit test
 - ❑ Running locally
 - ❑ Running on cluster
 - ❑ Hands on exercises Word count problem
 - ❑ MapReduce combiner
 - ❑ MapReduce partitioner
 - ❑ MapReduce distributed cache

Pig

Learning Objective

- ❖ Learning Pig
- ❖ Using Pig for Big data analysis

Topics Covered

- ❑ Introduction and Architecture
- ❑ Different Modes of executing Pig constructs Data Types
- ❑ Dynamic invokers Pig streaming Macros
- ❑ Pig Latin language Constructs (LOAD, STORE, DUMP, SPLIT etc)
- ❑ User Defined Functions
- ❑ Use Cases

Hive

Learning Objective

- ❖ Learning Hive
- ❖ Using Hive for Big data analysis

Topics Covered

- Introduction and Architecture
- Different Modes of executing Hive queries
- Metastore Implementations
- HiveQL(DDL & DML Operations)
- External vs Managed Tables Views
- Partitions & Buckets User Defined Functions
- Transformations using Non Java Use Cases
- Comparison of Pig and Hive.

Note :All topics are delivered as Hands-On sessions.

SQOOP

Learning Objective

- ❖ Learning Sqoop.

Topics Covered

- Sqoop Architecture
- Sqoop Import Command Arguments, Incremental Import
- Sqoop Export
- Sqoop Jobs

- ❑ Hands-on exercises

Flume

Learning Objective

- ❖ Learning flume.

Topics Covered

- ❑ Flume Agent Setup
- ❑ Flume Architecture
- ❑ Types of sources, channels, sinks Multi Agent Flow
- ❑ Hands-on exercises

NoSQL Database:HBase

Learning Objective

- ❖ Learning Big Data Database

Topics Covered

- ❑ NoSQL Concepts
- ❑ Review of RDBMS
- ❑ Need for NoSQL
- ❑ Brewers CAP Theorem
- ❑ ACID vs BASE
- ❑ Schema on Read vs. Schema on Write
- ❑ Different levels of consistency
- ❑ Bloom filters
- ❑ HBase architecture and concepts
- ❑ Hbase Data Model
- ❑ Hbase Shell Interface
- ❑ Hbase Java API

- ❑ Different types of NoSQL databases:
- ❑ Key Value
- ❑ Columnar
- ❑ Document
- ❑ Graph

Note: All topics are delivered as Hands-On sessions.

Oozie

Learning Objective

- ❖ Writing oozie workflow

Topics Covered

- ❑ Oozie workflow creations
- ❑ Oozie Job submission, monitoring, debugging
- ❑ Concepts on Coordinators and Bundles
- ❑ Hands-on exercises

Note: All topics are delivered as Hands-On sessions.

Hadoop Use Cases

Spark

Learning Objective

- ❖ Learning Spark basics and architecture

Topics Covered

- Introduction, and installing IntelliJ, and Scala
- Introduction to Apache Spark
- Spark Basics
- What's New in Spark 3?

Note: All topics are delivered as Hands-On sessions.

Scala

Learning Objective

- ❖ Learning Scala for spark

Topics Covered

- Why Scala
- Scala Installation
- Get deep insights into the functioning of Scala
- Execute Pattern Matching in Scala
- Functional Programming in Scala – Closures, Currying, Expressions, Anonymous Functions
- Know the concepts of classes in Scala
- Object Orientation in Scala – Primary, Auxiliary Constructors, Singleton & Companion Objects
- Traits and Abstract classes in Scala

RDD

Learning Objective

- ❖ Understanding RDD

Topics Covered

- ❑ The Resilient Distributed Dataset
- ❑ Ratings Histogram Example
- ❑ Spark Internals Key / Value RDD's,
- ❑ Example
- ❑ Filtering RDD's, and the Minimum Temperature by Location Example
- ❑ Counting Word Occurrences using Flatmap()
- ❑ Improving the Word Count Script with Regular Expressions
- ❑ Sorting the Word Count Results
- ❑ Find the Total Amount Spent by Customer

SPARK SQL, DATAFRAMES & DATASETS

Learning Objective

- ❖ Understanding SPARK SQL, DATAFRAMES & DATASETS

Topics Covered

- ❑ Introduction to SparkSQL
- ❑ ExampleUsing SparkSQL
- ❑ Using DataSets
- ❑ example using DataSets
- ❑ Ratings Histogram Example

Running Spark on cluster

Learning Objective

- ❖ Understanding running of spark on cluster

Topics Covered

- Using spark-submit to run Spark driver scripts
- Packaging driver scripts with SBT
- Package a Script with SBT and Run it Locally with spark-submit
- Using SBT and spark-submit
- Introducing Amazon Elastic MapReduce
- Creating Similar Movies from One Million Ratings on EMR
- Partitioning
- Best Practices for Running on a Cluster
- Troubleshooting, and Managing
- Dependencies

Machine Learning with Spark ML

Learning Objective

- ❖ Learning Machine Learning on Big data

Topics Covered

- Introducing MLLib
- Using MLLib to Produce Movie Recommendations
- Linear Regression with MLLib
- Running a Linear Regression with Spark
- Predict Real Estate Values with Decision Trees in Spark

Spark Streaming

Learning Objective

- ❖ Learning Spark Streaming

Topics Covered

- The DStream API for Spark Streaming
- Real-time Monitoring of the Most
- Structured Streaming

Kafka

Learning Objective

- ❖ Learning Kafka

Topics Covered

- Kafka Theory and Architecture
- Starting Kafka
- Command Line interface
- Kafka java programming
- Kafka twitter producer and configurations
- Kafka elastic search consumer and configurations