Data Engineering with Hadoop & Spark

Duration – 5 days

**Description:** Storing, managing, and processing datasets are foundational to both applied computer science and data science. Indeed, successful deployment of data science in any organization is closely tied to how data is stored and processed. This course introduces the fundamentals of data storage, retrieval, and mainly processing

This is hands on course which will introduce and enable the participants to work with different data sets with different file formats.

Engineering techniques using different tools like Hadoop Framework, Spark ,and data import-export utility tool Sqoop.

**Prerequisite:**

* Familiarity with UNIX systems (basic commands)
* Familiarity with vi, vim or gedit text editor preferred
* Basic programming background

**Objectives:** Every Organization need to construct well-curated Data Pipelines that can be instrumental in achieving business goals. Data Engineers can streamline data flow while maintaining information integrity. The domain of Data Engineering facilitates data infrastructures based on Databases and Data Warehouses, ensuring that data is in a usable shape by the time it reaches Data Scientists and Business user as a business support system. It primarily focuses on practical applications of Data Collection and Data Cleaning. Key responsibilities of every Data Engineer is:

* Data Acquisition
* Data Modelling
* Data Preparation
* Building Data Pipeline Systems

**Audience:** Graduates with any programming background

**Table of Content:**

**HDFS Architecture and hands-on:**

* Big Data and Hadoop Overview
* Hadoop Services (NN, DN, SNN, RM and NM)
* Hadoop Setup (Single Node and Multi Node)
* HDFS Commands – hands-on

**Hadoop: Processing Components**

* Map Reduce – Concept
* Understand Word-Count Program
* Map Reduce – Hands on (Sales Wise Sales, Weather Data Analysis)

**Adv. Map Reduce:**

* Combiner,
* Partitioner and
* Distributed Cache

**Apache HIVE and its Components:**

* Hive Architecture (Driver, Shell, Hive Engine, MetaStore)
* Creating HIVE Tables and Loading data
* Querying Hive database
* HIVE UDF
* HIVE Hands-on

**Apache SQOOP**

* Importing data from RDBMS to HDFS
* Importing data from RDBMS to HDFS in incremental append mode
* Importing data from RDBMS to HIVE
* Importing data from RDBMS to HBASE
* Exporting data from HDFS to RDBMS
* Full hands-on for all possible scenarios

**Scala Language:**

* Scala Data Types
* Controls Structures
* Classes and Properties
* Methods
* Traits
* Functional Programming
* Collections
* Collections: List, Array, ArrayBuffer, Map, Tuple, Set
* Files and Exception handling
* Database (working with Mysql)

**Spark Core:**

* Introduction to Data Analysis with Spark
* Downloading Spark and Getting Started
* RDD Basics
* Creating RDDs
* RDD Operations
* Transformations
* Actions
* Lazy Evaluation
* Common Transformations and Actions
* Converting Between RDD Types
* Persistence (Caching)

**Spark: Working with Key/Value Pairs**

* Creating Pair RDDs
* Transformations on Pair RDDs
* Aggregations
* Grouping Data
* Joins
* Sorting Data
* Actions On Pair RDD

**Spark: Loading and Saving Your Data**

* File Formats
* File systems
* Structured Data with Spark SQL
* Databases

**Advanced Spark:**

* Accumulators
* Broadcast Variables
* Working on a Per-Partition Basis

**Spark Job:**

* Creating a jar for Spark Application
* Submitting Spark application to Local
* Submitting Spark application to YARN Cluster
* Submitting Spark application to Standalone Cluster

**Spark SQL**

* Spark Hive Integration
* HIVE Context
* Spark SQL Context
* Accessing HIVE UDFs via Spark
* Spark SQL Performance

**Processing Different File Formats**

* Text Files
* CSV Files
* Json Files
* Parquet Files
* Avro Files

**Spark User Defined Functions**

* Spark’s UDF
* Accessing HIVE UDF from Hive and Spark

**Spark Streaming:**

* Architecture and Abstraction
* Transformations
* Stateless Transformations
* Stateful Transformations
* Windowed transformations
* UpdateStateByKey transformation