Data Engineering

## About this course:

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 processing systems in the context of common data analytics processing needs.

This is hands on course which will introduce and enable the participants to workWith differentdata engineering techniques.

## Duration :10 calendar days

## Prerequisites :

Familiarity with unix systems (installing packages, file handling.. etc)

Familiarity with vi text editor (optional)

Basic deployment environment awareness

Lab Software

      Will be provided the day on/ before the training

      Machines with reasonable configuration  - 16GB RAM

      Windows with Virtualization enabled /Ubuntu

      Participants need to have Open  access to the Internet

# Contents

**Additional Topics**

**Please include the below topics for Data engineering:**

* Introduction to data warehouse
  + Schema design
  + ETL process
  + Concept of materialized views
  + Concept of Fact and Dimension tables (Star Schema)

## Day 1

Introduction and overview

Data Engineering

Differences with Data Science

Different Data models – datawarehouses , no sql

Big Data

Commonly used Data Engineering Jargons

Traditional versus modernized Big Data and Cloud Computing Platforms using use cases

HDFS

Introduction and Motivation of Hadoop

What is Big Data

Challenges in Big Data

Challenges in Traditional Application

New Requirements

What is Hadoop

Brief history of Hadoop

Features of Hadoop

Hadoop v/s RDBMS

Overview of HDFS and MapReduce

Directory Structure of Hadoop

HDFS Shell commands

Hands on

## Day 2

HDFS Components

Various process running in Hadoop

Working of NameNode

Working f DataNode

Working f Secondary NameNode

Writing a file in HDFS

Reading a file from HDFS

Understanding HDFS Java classes and methods

Hands on Exercise

Sqoop:

Sqoop Architecture

Import and Export

Sqoop Practicals

Introduction to Hive

Motivation and Understanding Hive

Using Hive Command line Interface

## Day 3

Hive Data types and File Formats

Basic DDL and DML Operations

Schema Design

handson

Overview of BigData and Spark

MapReduce limitations

Spark History

Spark Architecture

Spark and Hadoop Advantages

Benefits of Spark + Hadoop

Introduction to Spark Eco-system

Spark Shell

Basic operations on Shell

Spark Context and Spark Properties

Persistence in Spark

HDFS data from Spark

## Day 4

Understanding RDD

Loading data into RDD

Scala RDD, Paired RDD, Double RDD & General RDD Functions

Transformations, Actions and Shared Variables

Spark Operations

Introduction to Spark SQL

Querying Files as Tables

Exploring the data

Text file Format

JSON file Format

Imputations

Introduction to juyper notebooks

Cells

Notebooks terminology

## Day 5

Python lambdas

Python : Functions

Python oop concepts

Introduction to Dataframes

data transformation operations

Exploring the data

Data Imputation Overview

Data Imputation techinques

Checking dimensions of Data

Statistical Summary of Data

Correlation between attributes

Overview of visualization

Data preparation and Bar Chart

Variable encoding techniques

Data Wrangling techniques

Data Cleanup: Investigation, Matching, and Formatting

Why Clean Data?

Data Cleanup Basics

Identifying Values for Data Cleanup

## Day 6

Formatting Data

Handling Missing Data

Filtering Out Missing Data

Filling In Missing Data

Variable transformations

Finding Outliers and Bad Data

Finding Duplicates

RegEx Matching

Data Transformation

Removing Duplicates

Transforming Data Using a Function or Mapping

Replacing Values

What to Do with Duplicate Records

Data Loading, Storage, and File Formats

Reading and Writing Data in Text Format

Reading Text Files in Pieces

## Day 7

Using HDF5 Format , Parque, Avro

Interacting with Web APIs

Interacting with Databases

Writing Data to Text Format

Working with Delimited Formats

JSON Data

XML and HTML: Web Scraping

Binary Data Formats

Introduction to NOSQL Databases.

NOSql Landscapes

MongoDB architecture

## Day 8

 Basics Of MongoDB

 Installation

Basic commands on mongo shell

Data Modelling Concepts

MongoDB CRUD Introduction, MongoDB CRUD Concept

Index Introduction

Index Concepts

Index Types

Index Properties

Index Creation and Indexing Reference

## Day 9

 What is Apache Kafka

Kafka Features and terminologies

High level Kafka Architecture

Real life Kafka Case Studies

Internals of architecture and core concepts

Kafka components - Broker, Producer, Consumer, Topics, Partitions

Different versions of Kafka

Working of Broker

Broker Deployment

Multiple brokers on single machine

Decommissioning Brokers

## Day 10

Understanding Producer

Basics of producer

Producer Architecture

Producer partition- Custom, Round Robin, Field Based Partition

Producer Java API

Types of Producer - sync,async

Different Producer Configurations

Sync and async producer hands on

Understanding Consumer

Basics of Consumer

Consumer Queuing, Consumer Group

Consumer Java API

Producer and Consumer Hands On

Reading the data from twitter , store it in hdfs

Please feel free to suggest the changes.