Data Science and Big Data Analytics

Course 6958 – 32 Hours

Overview

Business success in the information age is predicated on the ability of organizations to convert massive amount of raw data coming from various sources into high-grade business information. Many organizations are overwhelmed by the sheer volume of information they have to process in order to stay competitive. Traditional database systems may become either prohibitively expensive to handle the exponential growth of data volumes or found unsuitable for the job. Data Science and Big Data Analytics represent an emerging discipline that helps get a handle on the situation and capitalize on the wealth of information assets within your organization.

On Completion, Delegates will be able to

This intensive training course provides theoretical and technical aspects of Data Science and Business Analytics. The course covers the fundamental and advanced concepts and methods of deriving business insights from Big Data. The course is supplemented by hands-on labs that help attendees reinforce their theoretical knowledge of the learned material.

Who Should Attend

Business Analysts, IT Architects and Managers

Prerequisites

Participants should have the general knowledge of statistics and programming

Course Contents

Introduction to Big-Data Business Analysis and Logical Architecture

- Organization Requirements
  - Data needed, business logic, data resources
  - Different kind of users and organizations
  - Historical depth
  - Data in rest vs Data in Motion
- What is Big-Data
- Big-Data Characteristics & types
- Challenges and complexity
- Use cases in today's world
- Big-Data Architect roles
- Data scientist

Big-Data - data classification

- Data sources
- Content formats
- Data frequency
- Processing methodology
- Data consumers & data distribution
- Analysis type : batch, near real time (NRT) & real-time (RT)
- Types of data models (E-R, OLAP, Tabular , vector)
- No – SQL data models : Document, Graph and other NoSQL models
- Vector design for statistical model
- Data Serialization, Apache Avro
- Data Security

Introduction to Hadoop Eco-system

- Introduction
- Hadoop distributors (Cloudera, Hortonworks, MapR)
- Building blocks of Hadoop (NameNode, DataNode…)
- Introduction to HDFS
- Map-reduce pattern
- Distribute cache
- Hadoop 2 – YARN (Yet Another Resource Management)
- Introduction to Hive for ad-hoc queries
    - Hive basics
    - Hive data types
    - HiveQL
- Pig:
    - Introduction to Pig as data flow language
    - Pig Latin basic expressions
    - Operators for data processing
- Hbase:
    - Introduction to Hbase for processing huge tables
    - Hbase data model
    - Hbase vs. RDBMS
    - Client API (CRUD, queries and batch operations)
    - Interactive REST clients

Introduction to NoSQL

- RDBMS challenge in Big-data world
- No-SQL vs traditional relational data
- Scaling vs. consistency
- No-SQL DB types
- Transaction in No-SQL
- Applying map-reduce
- No-SQL leading implementations
- Exploring MongoDB

Introduction Statistical programming for Big-Data

- What is statistical programming
- The need for Big-Data
- R statistical language
- Introduction to R
    - installation
    - basic usage & general statistics with R
    - Data mining using R (rattle(
    - Hadoop-R

Introduction to "R " Spark

- Basics Spark
- More on RDD Operations
- Caching
- Modules built on Spark
- Spark Streaming
- Spark SQL