DATA SCIENCE

By Dr. Vishwanath Rao

Introduction to Data Science

- Need for Data Scientists
- Foundation of Data Science
- What is Business Intelligence
- What is Data Analysis, Data Mining, and Machine Learning
- Analytics vs Data Science
- Value Chain
- Types of Analytics
- Lifecycle Probability
- Analytics Project Lifecycle

Data

- Basis of Data Categorization
- Types of Data
- Data Collection Types
- Forms of Data and Sources
- Data Quality, Changes and Data Quality Issues, Quality Story
- What is Data Architecture
- Components of Data Architecture
- OLTP vs OLAP
- How is Data Stored?

Big Data

- What is Big Data?
- 5 Vs of Big Data
- Big Data Architecture, Technologies, Challenge and Big Data Requirements
- Big Data Distributed Computing and Complexity
- Hadoop
- Map Reduce Framework
- Hadoop Ecosystem

Data Science Deep Dive

- What is Data Science?
- Why are Data Scientists in demand?
- What is a Data Product
- The growing need for Data Science
- Large-Scale Analysis Cost vs Storage
- Data Science Skills
- Data Science Use Cases and Data Science Project Life Cycle & Stages
- Map-Reduce Framework

- Hadoop Ecosystem
- Data Acquisition
- Where to source data
- Techniques
- Evaluating input data
- Data formats, Quantity and Data Quality
- Resolution Techniques
- Data Transformation
- File Format Conversions
- Anonymization

Apache Spark

- Introduction to Apache Spark
- Why Spark
- Batch Vs. Real-Time Big Data Analytics
- Batch Analytics – Hadoop Ecosystem Overview
- Real-Time Analytics Options
- Streaming Data – Storm
- In Memory Data – Spark, What is Spark?
- Spark benefits to Professionals
- Limitations of MR in Hadoop
- Components of Spark
- Spark Execution Architecture
- Benefits of Apache Spark
- Hadoop vs Spark

Spark Core Architecture

- Spark & Distributed Systems
- Spark for Scalable Systems
- Spark Execution Context
- What is RDD
- RDD Deep Dive and Dependencies
- RDD Lineage
- Spark Application In Depth and Spark Deployment
- Parallelism in Spark
- Caching in Spark

Machine Learning Using Python

- Introduction to Machine Learning
- Areas of Implementation of Machine Learning
- Why Python
- Major Classes of Learning Algorithms

- Supervised vs Unsupervised Learning
- Learning NumPy
- Learning Scipy
- Basic plotting using Matplotlib
- Machine Learning application

Supervised and Unsupervised learning

- Classification Problem
- Classifying with k-Nearest Neighbours (kNN)

Algorithm

- General Approach to kNN
- Building the Classifier from Scratch
- Testing the Classifier
- Measuring the Performance of the Classifier
- Clustering Problem
- What is K-Means Clustering
- Clustering with k-Means in Python and an

Application Example

- Introduction to Pandas
- Creating Data Frames
- GroupingSorting
- Plotting Data
- Creating Functions
- Converting Different Formats
- Combining Data from Various Formats
- Slicing/Dicing Operations.

Scikit and Introduction to Hadoop

- Introduction to Scikit-Learn
- Inbuilt Algorithms for Use
- What is Hadoop and why it is popular
- Distributed Computation and Functional Programming
- Understanding MapReduce Framework Sample MapReduce Job Run

Hadoop and Python

- PIG and HIVE Basics
- Streaming Feature in Hadoop
- Map Reduce Job Run using Python
- Writing a PIG UDF in Python
- Writing a HIVE UDF in Python

- Pydoop and MRjob Basics