GTyde Technologies

Your Success Is Our Success

Syllabus of Data Science with R

Module 1: Introduction to Data Science Methodologies

- Data Types
- Introduction to Data Science Tools
- Statistics
- Approach to Business Problems
- Numerical Categorical
- R, Python, WEKA, RapidMiner

Module 2: Correlation / AssociationRegressionCategorical variables

- Introduction to Correlation Spearman Rank Correlation
- OLS Regression – Simple and Multiple Dummy variables
- Multiple regression
- Assumptions violation – MLE estimates
- Using UCI ML repository dataset or Built-in R dataset

Module 3: Data Preparation

- Data preparation & Variable identification
- Advanced regression
- Parameter Estimation / Interpretation
- Robust Regression
- Accuracy in Parameter Estimation
- Using UCI ML repository dataset or Built-in R dataset

Module 4: Logistic Regression

- Introduction to Logistic Regression
- Logit Function
- Training-Validation approach
- Lift charts
- Decile Analysis

- Using UCI ML repository dataset or Built-in R dataset

Module 5: Cluster AnalysisClassification Models

- Introduction to Cluster Techniques
- Distance Methodologies
- Hierarchical and Non-Hierarchical Procedure
- K-Means clustering
- Introduction to decision trees/segmentation with Case Study
- Using UCI ML repository dataset or Built-in R dataset

Module 6: Introduction and to Forecasting Techniques

- Introduction to Time Series
- Data and Analysis
- Decomposition of Time Series
- Trend and Seasonality detection and forecasting
- Exponential Smoothing
- Building R Dataset
- Sales forecasting Case Study

Module 7: Advanced Time Series Modeling

- Box – Jenkins Methodology
- Introduction to Auto Regression and Moving Averages, ACF, PACF
- Detecting order of ARIMA processes
- Seasonal ARIMA Models (P,D,Q)(p,d,q)
- Introduction to Multivariate Time-series Analysis
- Using built-in R datasets

Module 8: File Input/Output

- Live example/ live project
- Using client given stock prices / taking stock price data

Module 9: Pharmaceuticals

- Box – Jenkins Methodology
- Case Study with the Data
- Based on open set data

Module 10: Market Research

- Case Study with the Data
- Based on open set data

Module 11: Machine Learning

- Supervised Learning Techniques
- Conceptual Overview
- Unsupervised Learning Techniques
- Association Rule Mining Segmentation

Module 12: Fraud Analytics

- Fraud Identification Process in Parts procuring
- Sample data from online
- Text Analytics

Module 13: Text Analytics

- Sample text from online

Module 14: Social Media Analytics

- Social Media Analytics
- Sample text from online

Syllabus of Data Science with Python Course

Module 1: Introduction to Data Science

- What is Data Science?
- What is Machine Learning?
- What is Deep Learning?
- What is AI?
- Data Analytics & it’s types

Module 2: Introduction to Python

- What is Python?
- Why Python?

- Installing Python
- Python IDEs
- Jupyter Notebook Overview

Module 3: Python Basics

- Python Basic Data types
- Lists
- Slicing
- IF statements
- Loops
- Dictionaries
- Tuples
- Functions
- Array
- Selection by position & Labels

Module 4: Python Packages

- Pandas
- Numpy
- Sci-kit Learn
- Mat-plot library

Module 5: Importing data

- Reading CSV files
- Saving in Python data
- Loading Python data objects
- Writing data to csv file

Module 6: Manipulating Data

- Selecting rows/observations
- Rounding Number
- Selecting columns/fields
- Merging data
- Data aggregation
- Data munging techniques

Module 7: Statistics Basics

- Central Tendency
- Mean
- Median
- Mode
- Skewness
- Normal Distribution
- Probability Basics
- What does mean by probability?
- Types of Probability
- ODDS Ratio?
- Standard Deviation
- Data deviation & distribution
- Variance
- Bias variance Trade off
- Underfitting
- Overfitting
- Distance metrics
- Euclidean Distance
- Manhattan Distance
- Outlier analysis
- What is an Outlier?
- Inter Quartile Range
- Box & whisker plot
- Upper Whisker
- Lower Whisker
- catter plot
- Cook's Distance
- Missing Value treatments
- What is a NA?
- Central Imputation
- KNN imputation
- Dummification
- Correlation
- Pearson correlation
- Positive & Negative correlation
- Error Metrics

- Classification
- Confusion Matrix
- Precision
- Recall
- Specificity
- F1 Score
- Regression
- MSE
- RMSE
- MAPE

Module 8: Machine Learning

Module 9: Supervised Learning

- Linear Regression
- Linear Equation
- Slope<
- Intercept
- R square value
- Logistic regression
- ODDS ratio
- Probability of success
- Probability of failure
- ROC curve
- Bias Variance Tradeoff

Module 10: Unsupervised Learning

- K-Means
- K-Means ++
- Hierarchical Clustering

Module 11: Other Machine Learning algorithms

- Title
- Base
- Link
- Style s
- Script

Syllabus of Tableau Course

Module 1: Tableau Course Material

- Start Page
- Show Me
- Connecting to Excel Files
- Connecting to Text Files
- Connect to Microsoft SQL Server
- Connecting to Microsoft Analysis Services
- Creating and Removing Hierarchies
- Bins
- Joining Tables
- Data Blending

Module 2: Learn Tableau Basic Reports

- Parameters
- Grouping Example 1
- Grouping Example 2
- Edit Groups
- Set
- Combined Sets
- Creating a First Report
- Data Labels
- Create Folders
- Sorting Data
- Add Totals, Sub Totals and Grand Totals to Report

Module 3: Learn Tableau Charts

- Area Chart
- Bar Chart
- Box Plot
- Bubble Chart
- Bump Chart
- Bullet Graph
- Circle Views

- Dual Combination Chart
- Dual Lines Chart
- Funnel Chart
- Traditional Funnel Charts
- Gantt Chart
- Grouped Bar or Side by Side Bars Chart
- Heatmap
- Highlight Table
- Histogram
- Cumulative Histogram
- Line Chart
- Lollipop Chart
- Pareto Chart
- Pie Chart
- Scatter Plot
- Stacked Bar Chart
- Text Label
- Tree Map
- Word Cloud
- Waterfall Chart

Module 4: Learn Tableau Advanced Reports

- Dual Axis Reports
- Blended Axis
- Individual Axis
- Add Reference Lines
- Reference Bands
- Reference Distributions
- Basic Maps
- Symbol Map
- Use Google Maps
- Mapbox Maps as a Background Map
- WMS Server Map as a Background Map

Module 5: Learn Tableau Calculations & Filters

- Calculated Fields

- Basic Approach to Calculate Rank
- Advanced Approach to Calculate Ra
- Calculating Running Total
- Filters Introduction
- Quick Filters
- Filters on Dimensions
- Conditional Filters
- Top and Bottom Filters
- Filters on Measures
- Context Filters
- Slicing Fliters
- Data Source Filters
- Extract Filters

Module 6: Learn Tableau Dashboards

- Create a Dashboard
- Format Dashboard Layou
- Create a Device Preview of a Dashboard
- Create Filters on Dashboard
- Dashboard Objects
- Create a Story

Module 7: Server

- Tableau online.
- Overview of Tableau Server.
- Publishing Tableau objects and scheduling/subscription.

Syllabus of Big Data Hadoop with Spark Developer

Module 1: The big picture of Big Data

- Necessity of Big Data and Hadoop in the industry
- Paradigm shift - why the industry is shifting to Big Data tools
- Different dimensions of Big Data
- Data explosion in the Big Data industry

- Various implementations of Big Data
- Different technologies to handle Big Data
- Traditional systems and associated problems
- Future of Big Data in the IT industry

Module 2: Demystifying Hadoop

- Why Hadoop is at the heart of every Big Data solution
- Introduction to the Big Data Hadoop framework
- Hadoop architecture and design principles
- Ingredients of Hadoop
- Hadoop characteristics and data-flow
- Components of the Hadoop ecosystem
- Hadoop Flavors – Apache, Cloudera, Hortonworks, and more

Module 3: Setup and Installation of Hadoop

- Hadoop environment setup and pre-requisites
- Hadoop Installation and configuration
- Working with Hadoop in pseudo-distributed mode
- Troubleshooting encountered problems
- Hadoop environment setup on the cloud (Amazon cloud)
- Installation of Hadoop pre-requisites on all nodes
- Configuration of masters and slaves on the cluster
- Playing with Hadoop in distributed mode

Module 4: HDFS – The Storage Layer

Module 5: HDFS – The Storage Layer

- The need for a distributed processing framework
- Issues before MapReduce and its evolution
- List processing concepts
- Components of MapReduce – Mapper and Reducer
- MapReduce terminologies- keys, values, lists, and more
- Hadoop MapReduce execution flow
- Mapping and reducing data based on keys
- MapReduce word-count example to understand the flow
- Execution of Map and Reduce together
- Controlling the flow of mappers and reducers

- Optimization of MapReduce Jobs
- Fault-tolerance and data locality
- Working with map-only jobs
- Introduction to Combiners in MapReduce
- How MR jobs can be optimized using combiners

Module 6: MapReduce - Advanced Concepts

- Anatomy of MapReduce
- Hadoop MapReduce data types
- Developing custom data types using Writable & WritableComparable
- InputFormats in MapReduce
- InputSplit as a unit of work
- How Partitioners partition data
- Customization of RecordReader
- Moving data from mapper to reducer – shuffling & sorting
- Distributed cache and job chaining
- Different Hadoop case-studies to customize each component
- Job scheduling in MapReduce

Module 7: Hive – Data Analysis Tool

- The need for an adhoc SQL based solution – Apache Hive
- Introduction to and architecture of Hadoop Hive
- Playing with the Hive shell and running HQL queries
- Hive DDL and DML operations
- Hive execution flow
- Schema design and other Hive operations
- Schema-on-Read vs Schema-on-Write in Hive
- Meta-store management and the need for RDBMS
- Limitations of the default meta-store
- Using SerDe to handle different types of data
- Optimization of performance using partitioning
- Different Hive applications and use cases

Module 8: Pig - Data Analysis Tool

- The need for a high level query language - Apache Pig
- How Pig complements Hadoop with a scripting language

- What is Pig
- Pig execution flow
- Different Pig operations like filter and join
- Compilation of Pig code into MapReduce
- Comparison - Pig vs MapReduce

Module 9: NoSQL Database - HBase

- NoSQL databases and their need in the industry
- Introduction to Apache HBase
- Internals of the HBase architecture
- The HBase Master and Slave Model
- Column-oriented, 3-dimensional, schema-less datastores
- Data modeling in Hadoop HBase
- Storing multiple versions of data
- Data high-availability and reliability
- Comparison - HBase vs HDFS
- Comparison - HBase vs RDBMS
- Data access mechanisms
- Work with HBase using the shell

Module 10: Data Collection using Sqoop

- The need for Apache Sqoop
- Introduction and working of Sqoop
- Importing data from RDBMS to HDFS
- Exporting data to RDBMS from HDFS
- Conversion of data import/export queries into MapReduce jobs

Module 11: Data Collection using Flume

- What is Apache Flume
- Flume architecture and aggregation flow
- Understanding Flume components like data Sources and Sinks
- Flume channels to buffer events
- Reliable & scalable data collection tools
- Aggregating streams using Fan-in
- Separating streams using Fan-out
- Internals of the agent architecture

- Production architecture of Flume
- Collecting data from different sources to Hadoop HDFS
- Multi-tier Flume flow for collection of volumes of data using AVRO

Module 12: Apache YARN & advanced concepts in the latest version

- The need for and the evolution of YARN
- YARN and its eco-system
- YARN daemon architecture
- Master of YARN – Resource Manager
- Slave of YARN – Node Manager
- Requesting resources from the application master
- Dynamic slots (containers)
- Application execution flow
- MapReduce version 2 application over Yarn
- Hadoop Federation and Namenode HA

Module 13: Exploring Scala

- Introducing Scala
- Installation and configuration of Scala
- Developing, debugging, and running basic Scala programs
- Various Scala operations
- Functions and procedures in Scala
- Scala APIs for common operations
- Loops and collections- Array, Map, List, Tuple
- Pattern-matching and Regex
- Eclipse with Scala plugin

Module 14: Object-Oriented and Functional Programming

- Introduction to OOP - object oriented programming
- Different oops concepts
- Constructors, getters, setters, singletons; overloading and overriding
- Nested Classes and visibility Rules
- Functional Structures
- Functional programming constructs
- Call by Name, Call by Value

Module 15: Big Data and the need for Spark

- Problems with older Big Data solutions
- Batch vs Real-time vs in-Memory processing
- Limitations of MapReduce
- Apache Storm introduction and its limitations
- Need for Apache Spark

Module 16: A deep dive into Apache Spark

- Introduction to Apache Spark
- Architecture and design principles of Apache Spark
- Spark features and characteristics
- Apache Spark Ecosystem components and their insights

Module 17: Deploying Spark in local mode

- Spark environment setup
- Installing and configuring prerequisites
- Installation of Spark in local mode
- Troubleshooting encountered problems

Module 18: Apache Spark deployment in different modes

- Spark installation and configuration in standalone mode
- Installation and configuration of Spark in YARN mode
- Installation and configuration of Spark on a real cluster
- Best practices for Spark deployment

Module 19: Demystifying Apache Spark

- Working on the Spark shell
- Executing Scala and Java statements in the shell
- Understanding SparkContext and the driver
- Reading data from local file-system and HDFS
- Caching data in memory for further use
- Distributed persistence
- Spark streaming
- Testing and troubleshooting

Module 20: Learning RDDs in Spark

- Introduction to Spark RDDs

- How RDDs make Spark a feature rich framework
- Transformations in Spark RDDs
- Spark RDDs action and persistence
- Lazy operations and fault tolerance in Spark
- Loading data and how to create RDD in Spark
- Persisting RDD in memory or disk
- Pairing operations and key-value in Spark
- Hadoop integration with Spark
- Apache Spark practicals and workshops

Module 21: Spark Streaming

- The need for stream analytics
- Comparison with Storm and S4
- Real-time data processing using streaming
- Fault tolerance and checkpointing in Spark
- Stateful Stream Processing
- DStream and window operations in Spark
- Spark Stream execution flow
- Connection to various source systems
- Performance optimizations in Spark

Module 22: Spark MLlib and Spark GraphX

- Introducing Scala
- Installation and configuration of Scala
- Developing, debugging, and running basic Scala programs
- Various Scala operations
- Functions and procedures in Scala
- Scala APIs for common operations
- Loops and collections- Array, Map, List, Tuple
- Pattern-matching and Regex
- Eclipse with Scala plugin

Module 23: Spark SQL

- Introduction to Spark SQL
- Apache Spark SQL Features and Data flow
- Architecture and components of Spark SQL

- Hive and Spark together
- Data frames and loading data
- Hive Queries through Spark
- Various Spark DDL and DML operations
- Performance tuning in Spark

Module 24: Real Life Hadoop & Spark Project

Live Apache Spark & Hadoop project using Spark & Hadoop components to solve real-world Big Data problems in Hadoop & Spark.

Syllabus of Data Science Capstone Course

Module 1

- Ignite Talk
- Statement of work

Module 2

- Milestone #1 Presentation
- Summary Report + technical report
- Self-/peer- evaluation
- Review another group's reports
- Code (runs as advertised)

Module 3

- Milestone #2 Presentation ("Midterm")
- Summary Report + technical report
- Self-/peer- evaluation
- Review another group's reports
- Code (runs as advertised)

Module 4

- Milestone #3 Presentation
- Summary Report + technical report

- Self-/peer- evaluation
- Review another group's reports
- Code (runs as advertised)

Module 5

- Final Presentation to class
- Final write-up via blog
- Poster and video recording
- Self-/peer- evaluation
- Code (runs, is organized and readable)