Data Scientist Syllabus

Data Scientist/Analyst, now a day's the most buzzing work in IT world. Businesses are generating so much of the data and the need to analyze the data is top most priority.

Keeping in line with the market requirements, as per the job description for data scientist role, we have designed a new course:

1. **Complete R Programming**: R is a Data Analytical Language
2. **SAS**: For each module of R, we will cover SAS also(Optional)
3. **Python**: Python is an Data Analytical Language (Optional)
4. **Hadoop**: Basic to intermediate aspects of Hadoop
5. **Spark**: Hadoop combined with Spark makes a great combination.(Optional)
6. **Tableau**: It's the Visualization tool, which helps in presenting the reports and graph's to business.
7. **Excel/SQL**: It's very vital for a Data Scientist to work on excel files and Databases.(Optional)

R Programming

What is R?

- Birth and Rise of R
- Links for the necessary software
- GUI of R: IDE and Statistical Analysis Interfaces
- R Workspace
- GUI of RStudio

Data Scientist Syllabus

Basic Operations in R

- Expressions: Basic Idea
- Constant Values: Numeric & Non-numeric
- Arithmetic: Operations and BODMAS
- Conditions: Equality, Greater Than, Less Than, etc.
- Function Calls: Introduction to R Functions
- Symbols & Assignment
- Keywords: NA, Inf, NaN, NULL, TRUE, FALSE
- Naming a Variable: Generally accepted conventions

Data Types & Data Structures in R

- Basic data types
- Basic data structures: Vector, Factor, Matrices, Data Frame, List

Subsetting in R

- Vector Subsetting
- c() function: Creation of Vectors
- Using rep() and seq() functions
- Using factor() to covert vectors to factors
- Using data.frame() to create data frames
- Meta data access: dimnames(), rownames(), colnames()
- Using matrix() to create matrices
- Using array() to create arrays
- Subsetting data frames: row subset, column subset, using subset()
- function
- Assigning to a subset

Data Scientist Syllabus

- Using is.na() to detect NA
- Subsetting factors

Additional Topics on Data structures

- The recycling rule: Uneven arithmetic operation on vectors
- Type coercion: Character to Numeric
- Automatic Type coercion
- Coercing factors: Using as.factor() function
- Changing factor levels
- Attributes:
  - attribute() functions
  - attr()functions
  - names() functions
- Classes: Idea of OOP in R
- Dates: As a special class
- Formulas: As a special class
- Exploring Objects:
  - summary(),
  - str(),
  - dim() functions
- Generic functions

Data Import & Export

- Text formats: Reading Delimited Files
- read.table() function
- Using read.fwf() function for fixed width files
- Using readLines() for reading lines

Data Scientist Syllabus

- Using write. csv() function to store data as CSV files
- Reading Excel file: Package XLConnect
- Reading SPSS file: Package Foreign
- Reading SAS data file: Package sas7bdat
- Database connection: The ideas of ODBC connecting in Windows
- RODBC package: Create and Query database from R
- Basic SQL

Control Structures & User defined Functions

- Conditional Statements
- If statement: The Structure
- If Else statement: The Structure
- ifelse() function
- Iteration & Looping
- The for loop
- The while loop
- The repeat statement
- lapply() function
- sapply() function
- apply() function
- User defined function
- Variable scooping: Global and Local Variables
- Using user defined functions inside function definition

Charting with R

- The plot function
- plot.new() function: Generating new plot object

Data Scientist Syllabus

- plot.window() function: Creating window
- points() function: Plotting points
- axis() function: Generating Axis
- box() function: Creating enclosure
- title() function: Assigning title
- par() function: Fixing plotting parameters
- lines() function: Adding connector lines
- Multi figure layout: Creating multiple charts in the same window
- hist() function: Plotting histograms
- Kernel Density Plot: The non-parametric probability distribution
- Comparing Groups via Kernel Density: Comparing two different
- probability distributions
- Simple Bar Plot: Visualizing categorical data
- Staked Bar Plot: Understating category composition
- Grouped Bar Plot
- Line Charts
- Pie Charts
- Boxplots: Understanding data distributions and outliers
- Geo Charts
- Motion Charts

Analytics and Statistical Analysis Using R

- Summary statistics for data
- t tests: Comparing means
- Anova: Comparing means and causal relations
- Factor Analysis: Dimension Reduction technique
- Cluster Analysis: Segmentation and Homogeneous groups of data

Learn Well Technocraft

Website: www.dw-learnwell.com Contact: +91 8411002339/+91 7709292162
Email: info@dw-learnwell.com
Classroom | Corporate | Online

Data Scientist Syllabus

Analytics & Data Mining Using R

- Linear Regression: Predicting from uni-linear causality
- Logistic Regression: Predicting the probability in a binary outcome
- Situations.
- Time series Analysis: Automated ARIMA
- Decision Trees: Conditional inference trees for classification and
- Profiling

Analytics: Association Rule Mining Using R (Market Basket Analysis)

- Introduction to Association learning
- Different types of association algorithms
- Apriori Algorithm: Support, Confidence and Lift
- Market basket Analysis

Text Mining Using R

- Introduction to Text Mining
- Keyword search
- Word cloud
- Sentiment Analysis
- Twitter Data Analysis – Case Study.

SAS

Note: For each module for R programming SAS topics will be covered including Regression, Machine Learning and Sentiment Analysis etc.

Data Scientist Syllabus

Hadoop

- Hadoop 1.0 overview and enhancements in Hadoop 2.0
- Hadoop installation and setup using Virtual Box and Hortonworks distro.
- Typical Cluster architecture in Hadoop 1.0 vs Hadoop 2.0 (optional)
- HDFS architecture
- MR architecture
- Java MR example (optional for interested students)
- Scoop hands on with example
- Pig hands on with example
- Hive hands on with example

Note: We have a separate 5-6 weekends detailed course for Hadoop also.

Spark

Introduction to Spark

- Installation & Overview.
- Reading data from text files
- Basics of Spark and core concepts like RDD, caching etc.
- Understand few famous programs like word count and additionally trying few more.
- Trying out various APIs offered by Spark Core libraries.

Spark SQL

- Overview of SparkSQL
- Using Hive meta data with Spark SQL

Data Scientist Syllabus

- SchemaRDDs
- Using various File formats like Parquet and JSON
- Using Spark SQL and Hive UDFs

Spark ML

- Overview of Spark ML
- Understanding Vectors
- Understanding Linear regression and running with Spark ML
- Understanding Logistic regression and running with Spark ML
- Running Clustering example with Spark ML
- Dimensionality reduction in Spark using Principal Component Analysis

Tableau

Overview:

- User interface basics
- Connecting to data
- Dimensions vs. measures
- Show Me
- Marks card
- Simple formatting
- Building views
- Building a dashboard

Connecting to Excel, CSV and Text Files:

- Connecting to single or multiple tables
- Connecting live versus importing the data

Data Scientist Syllabus

- Editing data connections after initial connection
- Data source filtering

Working with Data:

- Flat Files (Excel, CSV, Access DB)
- Relational Databases
- ODBC Drivers
- Live or Import Data Connection
- Metadata Management
- Multiple Data Connections
- Creating and Refreshing an Extract

Analysis:

- Hierarchies
- Sorting
- Grouping
- Filtering
- Aggregations
- Trend lines
- Page shelf
- Forecasting

Formatting:

- Row-banding
- Number formatting
- Text formatting
- Shading
- Labels

Data Scientist Syllabus

- Annotations
- Tooltips

Calculation:

- Aggregate Calculations
- Row-Level Calculations
- Quick Table Calculations

Dashboard:

- Dashboard Objects
- Filter Actions
- URL Actions
- Sizing
- Tiled and Floating Sheets
- Dynamic Sheet Titles

Tableau Server(Provided availability of License):

- Publishing the Workbook
- Scheduling Refresh Extract
- Managing Authentication and Authorization
- Monitoring Background Tasks
- Automation of Reports

Note: We have a separate detailed 5-6 weekends course for Tableau also.