| **DATA ANALYSIS** (6 HOURS) | Applications | Careers | Types of Data | Data Warehousing | MDM Vs. DW | What makes a Data Scientist? | Statistics |
| **PROGRAMMING** (8 HOURS) | Basics of Java | Creating Java Application Programs | Implementing Loops, Arrays | Basic Commands of Linux for Better Understanding of Hadoop |
| **BIG DATA** (2 HOURS) | Characteristics of Big Data | Challenges | Popular Tools Used to Store, Process, Analyze & Visualize Big Data | Use Cases for Big Data |
| **HADOOP ECO-SYSTEM** (2 HOURS) | Characteristics | Eco-system & Core Components | Where Hadoop Fits? | When to Use & Not Use | Apache Hadoop Distributions | Job Trends |
| **HDFS & YARN** (4 HOURS) | HDFS Architecture and Features | Files and Data Blocks | Classic vs. YARN | Daemons | Containers | Speculative Execution | HDFS Federation |
| **HADOOP SETUP** (6 HOURS) | Deployment Modes | Setting up a Pseudo-distributed Cluster | Hortonworks Sandbox Configuration | HDFS File System Operations | Hadoop Services using Ambari | Web UI | Filesystem & Linux Commands |
| **MAPREDUCE** (9 HOURS) | Architecture and Use Cases | Programming | Packaging MapReduce Jobs in a JAR | Mapper & Reducer Counts | Partitioners & Custom Partitioners |
| **HADOOP STREAMING** (2 HOURS) | Hadoop Streaming using Python | Demo: Writing Python Scripts for Streaming | Testing Python Scripts | Executing YARN Jar on Python Script |
| **PIG** (6 HOURS) | Architecture | Pig Latin Basics | Execution Modes | Pig Processing | Built-in and User Defined Functions & Operators |
| **HIVE** (6 HOURS) | Architecture | Warehouse Directory & Metastore | Query Language | Data Processing | Built-in Functions | Joins and Bucketing | Partitioning Data |
| **HBASE** (4 HOURS) | Hbase Data Model | Row Oriented v/s Column Oriented | Storage | Architecture | Shell Commands | Bulk Load Data into Hbase |
| **SQOOP AND FLUME** (4 HOURS) | Setup MySQL RDBMS & Sqoop | Sqoop Connectors, Commands | Importing Data to HDFS & Hive | Exporting Data to MySQL | Flume Data Ingestion |
| **OOZIE** (6 HOURS) | Features and Challenges | DAG Architecture | Setting up Database & Oozie Configuration | Creating Workflows | Submitting and Managing Oozie Jobs |
| **PROBABILITY THEORY** (7 HOURS) | Events, Probabilities, Rules | Conditional Probability | Distribution | Central Limit Theorem | Expectation & Variance | Naïve Bayes | Design of Experiments |
| **BASIC STATISTICS** (2 HOURS) | Events and their Probabilities | Rules of Probability | Distribution of a Random Variable | Central Limit Theorem | Naïve Bayes |
| **HYPOTHESIS & OTHER TESTS** (6 HOURS) | Hypothesis | Probability | One Sample / Two Samples T-Test | Paired T-test | Proportional, Non Parametric One Sample, Chi Square, Z, F Test |
| **CORRELATION ANALYSIS** (2 HOURS) | Pattern Discovery | Statistics Associated with Cross Tabulations | Chi Square, Phi Coefficient, Contingency Coefficient | Correlation Analysis |
| **LINEAR REGRESSION** (6 HOURS) | Assumptions | Hypothesis | Variable and Model Significance | Regression Table | Anova Table | Multicollinearity | Heteroscedasticity |
| **ANOVA** (4 HOURS) | One Way Analysis of Variance | Assumptions | Statistics | Interpreting Results | Two Way Analysis of Variance | Analysis of Covariance |
| **LOGISTIC REGRESSION** (6 HOURS) | Assumptions | Reason for Logit Transform | Hypothesis | Variable Model Significance | Regression Table | Chi Square Test | ROC Curve |
## MACHINE LEARNING (2 HOURS)
- What is Machine Learning?
- Types of Problems and Tasks
- Features, Models and Design of ML Study

## OTHER MODELS (6 HOURS)
- Distance-based and Non Linear Models
- KNN
- K Means
- SVM
- Bayesian Network Models
- Neural Networks
- Perceptron, MLP, Back Propagation

### SEMESTER 2: SPECIALIZATION IN EITHER R OR PYTHON 85 HOURS

#### INTRODUCTION TO R SOFTWARE (10 HOURS)
- Installation
- Architecture
- Installing Packages
- Setting Directories
- Basic Operations
- Scalars, Vectors

#### LINEAR REGRESSION (8 HOURS)
- Basic Statistics Refresher
- Covariance and Correlation
- Multivariate Analysis
- Assumptions of Linearity
- Hypothesis Testing
- Limitations of Regression
- Case Study For Linear Regression: Case for Prediction Problem

#### LOGISTIC REGRESSION (6 HOURS)
- The Logistic Transform
- Logistic Regression Modelling
- Model Optimisation
- Understanding the ROC Curve
- Case Study

#### OTHER MODELS (6 HOURS)
- SVM
- Tuning the Model
- Case Study

#### DECISION TREE (4 HOURS)
- Classification Trees
- Regression Trees
- Case Study for Decision Tree

#### SEGMENTATION (4 HOURS)
- Clustering
- Kmeans Algorithm
- Cluster Size vs Definition Optimisation
- K- mediod and Fuzzy K means
- Case for Clustering on Bank Customer Data Set

#### ASSOCIATION RULE MINING (4 HOURS)
- Supervised Vs Unsupervised Learning
- Recommendation Engines
- Association Rule Mining
- Case Study For Market Basket Analytics

#### TIME SERIES (4 HOURS)
- Time Series Decomposition
- Moving Average & Exponential Smoothing Methods
- AR, MA, ARIMA, SARIMA, RMSE and MAPE
- Case Study

#### KNN ALGORITHM (6 HOURS)
- K Nearest Neighbours Algorithm for Classification
- Lazy Learning Notion
- Data Transformations
- Evaluation of Model
- Pros and Cons
- Case Study

#### NAÏVE BAYES ALGORITHM (6 HOURS)
- Bayesian Theorem
- Probabilities
- Conditional and Joint Probabilities Notion
- Traditional and Naive Approach
- Model Building
- Case Study

#### ANN & SVM (12 HOURS)
- Neural Networks
- Structure of Network
- The ANN Model
- Training the Model
- Testing and Validation
- SVM
- Tuning the Model
- Case Study

#### ENSEMBLE MODELS (6 HOURS)
- Entropy
- Information Value
- Decision Tree Pruning
- Model Validation & Performance
- Bagging, Boosting Trees
- Random Forests
- Case Study

#### PYTHON BASICS (4 HOURS)
- What is Python?
- Installing Anaconda
- Spyder Integrated Development Environment (IDE)
- Python Basics and String Manipulation

#### DEALING WITH DATA (12 HOURS)
- Data Management
- Lists, Tuples, Dictionaries, Variables
- Crud Operations
- Pydoop, Pymongo, Pyspark
- Data scraping and Collection
- Data Structures in Python Used for Data Analysis: Numpy Arrays, Indexing, Pandas

#### DATA FRAME MANIPULATION (4 HOURS)
- Data Management
- Lists, Tuples, Dictionaries, Variables
- Crud Operations
- Pydoop, Pymongo, Pyspark
- Data scraping and Collection
- Data Structures in Python Used for Data Analysis: Numpy Arrays, Indexing, Pandas

#### NATURAL LANGUAGE PROCESSING (8 HOURS)
- Text Preprocessing
- Stemming
- Bag of Words Approach and Naïve Bayes
- Latent Semantic Analysis
- Tagging, Categorization
- Sentiment Analysis

www.imarticus.org
# CURRICULUM DIPLOMA IN BIG DATA & ANALYTICS

<table>
<thead>
<tr>
<th>VISUALIZATION (6 HOURS)</th>
<th>Image Processing</th>
<th>Extractors for Image Processing</th>
<th>Using Classifiers</th>
<th>Text Extraction from Image</th>
<th>Data Visualization in Python: Charts, Plots etc</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTHER TOOLS (4 HOURS)</td>
<td>Other Predictive Modelling Tools</td>
<td>Machine Learning</td>
<td>Sklearn Library and Statsmodels</td>
<td>Simple Regression Analysis</td>
<td>Multiple Regression</td>
</tr>
<tr>
<td>KNN AND NAÏVE BAYES ALGORITHMS (12 HOURS)</td>
<td>K Nearest Neighbours Algorithm for Classification</td>
<td>Naïve Bayes Algorithm for Multi Class Predictions</td>
<td>Model Building, Testing</td>
<td>Case Studies</td>
<td></td>
</tr>
<tr>
<td>ANN AND SVM (12 HOURS)</td>
<td>Artificial Neural Networks</td>
<td>Structure of Network</td>
<td>Support Vector Machines</td>
<td>Build, Test, Train, Validate Model</td>
<td>Case Studies</td>
</tr>
<tr>
<td>ENSEMBLE MODEL (6 HOURS)</td>
<td>Entropy</td>
<td>Information Value</td>
<td>Decision Tree Pruning</td>
<td>Model Validation &amp; Performance</td>
<td>Bagging, Boosting Trees</td>
</tr>
</tbody>
</table>

**SEMESTER 3: SAS, TABLEAU AND INTERVIEW PREP**  
70 HOURS

<table>
<thead>
<tr>
<th>INTRODUCTION (12 HOURS)</th>
<th>What is SAS?</th>
<th>Submitting a SAS Program</th>
<th>SAS Program Syntax</th>
<th>Accessing Data, Reporting and Formatting Data Values</th>
<th>SAS Datasets</th>
<th>SAS Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA MANIPULATION (12 HOURS)</td>
<td>Reading SAS Datasets, Excel Data, Raw Files, Database Data</td>
<td>Creating Summary Reports</td>
<td>Combining Datasets</td>
<td>Summarizing Data</td>
<td>Observations</td>
<td>Functions in SAS</td>
</tr>
<tr>
<td>BASIC STATISTICS &amp; REGRESSION (12 HOURS)</td>
<td>Measures of Central Tendency</td>
<td>Measures of Dispersion</td>
<td>Skewness and Kurtosis</td>
<td>Linear Regression &amp; Modeling</td>
<td>Logistic Regression &amp; ROC Curve</td>
<td>Model Parameter Significance Evaluation</td>
</tr>
<tr>
<td>TIME SERIES (6 HOURS)</td>
<td>Time Series Decomposition</td>
<td>Simple &amp; Weighted Moving Average Method</td>
<td>Exponential Smoothing Methods</td>
<td>Stationarity of Data</td>
<td>ARIMA Models</td>
<td>Case Studies on Time Series</td>
</tr>
<tr>
<td>TABLEAU (2 HOURS)</td>
<td>Relevance of Visualization</td>
<td>What is Tableau?</td>
<td>Uses</td>
<td>Installation and Architecture</td>
<td>Working with Tableau</td>
<td>Exporting, Connecting and Loading</td>
</tr>
<tr>
<td>DATA AND CHARTS (4 HOURS)</td>
<td>Different Types of Charts</td>
<td>Data Organization</td>
<td>Calculated Metrics</td>
<td>Sorting, Filtering</td>
<td>Totals &amp; Sub Totals</td>
<td>Aggregated Measures</td>
</tr>
<tr>
<td>VISUALIZATION (4 HOURS)</td>
<td>Advanced Visualization</td>
<td>Combination Charts</td>
<td>Reference Lines, Reference Bands</td>
<td>Pareto Analysis</td>
<td>Market Basket Analysis</td>
<td>Mapping</td>
</tr>
<tr>
<td>DATA PRESENTATION (5 HOURS)</td>
<td>Dashboard Layouts and Formatting</td>
<td>Interactivity using Actions</td>
<td>Dashboard Best Practices</td>
<td>Data Summarization</td>
<td>Making Presentations Relevant</td>
<td>Publishing on Web</td>
</tr>
</tbody>
</table>

www.imarticus.org