**120 Hour Data Science Course (Big Data)**

**By**

**GRRAS SOLUTIONS PVT. LTD**

1. Tools and Languages for Data Science: Python, SQL, Hadoop, Spark & ML.

2. Total Duration: 120 Hours (approx. 2 months according to 3 hours daily)

3. Overview:-

|  |  |  |
| --- | --- | --- |
| Sl.No. | Topic | Approx. Duration |
|  | Programming with Python | 25 Hours |
|  | DBMS (SQL & No-SQL) | 25 Hours |
|  | Linux | 10 Hours |
|  | Big Data & Hadoop | 50 Hours |
| 5. | Machine Learning | 10 Hours |
|  | **TOTAL** | **120 Hours** |

**a) Core & Advanced Python: About 25 hours**

**Core Python**: Introduction to Python, Basic Syntax, Flow of Control (Modules, Branching), If, If- else, Nested if-else, Looping, For, While, Nested loops, Control Structure, Break, Continue, Pass, Strings and Tuples, Basic Operations, Working with Lists, Introduction, Accessing list, Introducing Tuples, Accessing tuples Operations, Function and Methods, Files, Modules, Dictionaries, Advance Dictionaries, Functions and Functional Programming, Special Functions in python lambda, map and reduce. Advance functions in python, Namespace and Generator and Iterators.

**Advanced Python:** Object Oriented, OOPs concept, Class and object, Attributes, Inheritance, Overloading, Overriding, Data hiding, Meta Classes, Shared Memory concepts, Exception Handling, User Defined Exceptions, Debugging modules, Python Libraries (SCIPY, Scikit-Learn, bs4 etc.)

**b) DBMS (SQL & No SQL): About 25 Hours**

Database Concepts (File System and DBMS), Database Storage Structures (Tablespace, Control files, Data files), Structured and Unstructured data, SQL Commands (DDL, DML & DCL), Data ware Housing concept, No-SQL, Data Models - XML, working with MongoDB). SQL Programing (Download the dataset, Shortcut to upload the data, SELECT \* Statement, Using the WHERE clause to filter data, using Wildcards / Regular Expressions in SQL (% and \_), Order By, Data Types in SQL, implicit Data Conversion in SQL, Using Cast() vs Convert(), Working with NULLs, LEFT, RIGHT, INNER, and OUTER joins, Joins with duplicate values, Joining on multiple fields.

**c) Linux: About 10 Hours**

**Basics of Linux –** File Systems, Users, Groups, Permissions, Mounting, Shell Scripts, Ip & Networking, Firewall, Services and Daemons, Apache Server, MariaDB Server, CGI scripts

**d) Big Data & Hadoop: About 50 Hours**

**Introduction to Big Data:** Big data definition, enterprise / structured data, social / unstructured data, unstructured data needs for analytics, What is Big Data, Big Deal about Big Data, Big Data Sources, Industries using Big Data, Big Data challenges.

**Hadoop:** Introduction of Big data programming-Hadoop, History of Hadoop, The ecosystem and stack, The Hadoop Distributed File System (HDFS), Components of Hadoop, Design of HDFS, Java interfaces to HDFS, Architecture overview, Development Environment, Hadoop distribution and basic commands, Eclipse development, The HDFS command line and web interfaces, The HDFS Java API (lab), Analysing the Data with Hadoop, Scaling Out, Hadoop event stream processing, complex event processing, MapReduce Introduction, Developing a Map Reduce Application, How Map Reduce Works, The MapReduce Anatomy of a Map Reduce Job run, Failures, Job Scheduling, Shuffle and Sort, Task execution, Map Reduce Types and Formats, Map Reduce Features, Real-World MapReduce.

**Hadoop ETL:** Hadoop ETL Development, ETL Process in Hadoop, Discussion of ETL functions, Data Extractions, Need of ETL tools, Advantages of ETL tools.

**Pig and HIVE**- Programming Pig: Engine for executing data flows in parallel on Hadoop, Programming with Hive: Data warehouse system for Hadoop, Optimizing with Combiners and Practitioners (lab), More common algorithms: sorting, indexing and searching (lab), Relational manipulation: map-side and reduce-side joins (lab), evolution, purpose and use, HDFS – Overview and concepts, data flow (read and write), interface to HDFS (HTTP, CLI and Java API), high availability and Name Node federation, Map Reduce developing and deploying programs, optimization techniques, Map Reduce Anatomy, Data flow framework programming Map Reduce best practices and debugging, Introduction to Hadoop ecosystem, integration R with Hadoop.

**Hadoop Environment:** Setting up a Hadoop Cluster, Cluster specification, Cluster Setup and Installation, Hadoop Configuration, Security in Hadoop, Administering Hadoop, HDFS – Monitoring & Maintenance, Hadoop benchmarks, Hadoop in the cloud.

**Apache Spark APIs for large-scale data processing:** Overview, Linking with Spark, Initializing Spark, Resilient Distributed Datasets (RDDs), External Datasets, RDD Operations, Passing Functions to Spark, Working with Key-Value Pairs, Shuffle operations, RDD Persistence, Removing Data, Shared Variables, Deploying to a Cluster.

**e) Introduction to Machine Learning: About 10 Hours**

**Supervised Machine Learning**: Introduction to machine learning, Logistic Regression, Decision Trees, Support Vector Machine, and Recommender Systems.

**Unsupervised Machine Learning:** Introduction to Un-Supervised machine Learning, Clustering, K-Means, and Principle Component Analysis (PCA).