**Expertise Training for Big Data (Hadoop Eco System + Spark + Cloud)**

* Kick Start your Big Career with BigData using our training

We offer training for Big Data Technologies such as Hadoop Eco System, NoSQL DataStores, Apache Spark, Apache Storm, Search Technologies and Cloud Services like Azure and AWS. We also provide customized, private trainings on any Big Data Technologies.

**Who should go for this Course?**

This course is a foundation to anyone who aspires to embark into the field of big data and keep abreast of the latest developments around fast and efficient processing of ever-growing data using Hadoop, HBase, Spark and related projects. The course is ideal for:

1. Big Data enthusiasts

2. Student, Researcher, Software architects, engineers and developers

3. Data Scientists and analytics professionals

|  |  |  |
| --- | --- | --- |
| **Training Plan** | | |
| **Date/Duration** | **Title** | **Sub Topic** |
| **Day1 -3 Hours** | **Big Data Knowledge** | Introduction to Big Data Analytics, Linux and Hadoop Installation, Hive, Pig Installation |
| **Hadoop: Basic Concepts** | Problems with traditional large-scale systems |
| An Overview of Hadoop - Big Data (What, Why, Who) |
| The Motivation For Hadoop |
| Requirements for a new approach |
| Anatomy of a Hadoop Cluster |
| Other Hadoop Ecosystem Components |
| **Day2 - 3 Hours** | **MapReduce** | Basic MapReduce API Concepts |
| How MapReduce Works |
| The MapReduce Flow |
| Map-Only MapReduce Jobs |
| Splittable File Formats |
| Determining the Optimal Number of Reducers/Mappers |
| **Day3 -1 Hours Lab - 2 Hours** | **MapReduce  - More depth** | Examining a Sample MapReduce Program |
| Writing a MapReduce Program |
| Common MapReduce Algorithms |
| Sorting and Searching/ The Secondary Sort |
| ToolRunner – Debugging(LocalJobRunner) |
| Indexing, Sorting |
| Joining Data Sets in MapReduce |
| CustomInputFormat – Distributed Cache – GenericOptionsParser |
| Map-Side Joins/Reduce-Side Joins |
| **Day4 - 1 Hours Lab - 2 Hours** | **The Hadoop Distributed File System** | HDFS Concepts |
| HDFS Design - Architecture |
| HDFS File System/Interface and CLI |
| Delving Deeper Into The Hadoop API |
| **Day5 - 1 Hours Lab - 2 Hours** | **Hive** | Hive Basics, Hive QL, Hive vs RDBMS - |
| Architecture – Installation –Configuration |
| Tables – DDL – DML – UDF – UDAF – Partitioning – Bucketing – MetaStore |
| Hive-Hbase Integration – Hive Web Interface – Hive Server(JDBC,ODBC, Thrift) – File Formats (RCFile - ORCFile) – Other SQL on Hadoop |
| **Day6 - 1 Hours Lab - 2 Hours** | **PIG** | Architecture –Installation |
| Pig Basics, Pig grunt/editor, Hive vs Pig, Pig Latin Syntax |
| Data Types –Functions (Eval, Load/Store, String, DateTime), Pig UDF and Data processing |
| Joins - Pig Server –Macros- UDFs Performance - Troubleshooting – Commonly Used Functions |
| Creating InputFormats and OutputFormats |
| **Day7 - 1 Hours Lab - 2 Hours** | **Hbase** | Hbasics, HbBase Vs RDBMS, Introduction to NoSQL – CAP Theorem – Classification of NoSQ |
| Hadoop Schema/Loading Data |
| HBASE and HDFS- Architecture (Read Path, Write Path, Compactions, Splits) |
| Installation – Configuration - Role of Zookeeper |
| HBase Shell - Java Based APIs (Scan, Get, other advanced APIs )– Introduction to FiltersRowKey Design |
| Map reduce Integration – Performance Tuning –What’s New in HBase 0.98 – Backup and Disaster Recovery - Hands On |
| **Day8 - 1 Hours Lab - 2 Hours** | **SQOOP  - Integrating Hadoop Into The Workflow** | Relational Database Management Systems |
| Storage Systems, Exporting Data from HDFS to RDBMS |
| Importing Data from RDBMSs With Sqoop |
| Architecture , Installation |
| Commands(Import , Hive-Import, EVal, Hbase Import, Import All tables, Export) – Connectors to Existing DBs and DW |
| **Day9 - 1 Hours Lab - 2 Hours** | **Apache Oozie** | Actions, Workflow, Trigger and Bundels, Coordinator, Action (Mapreduce, Hive, Pig, Sqoop) – Introduction to Bundle – Mail Notifications |
| Architecture, Installation |
| **Day10 - 1 Hours Lab - 2 Hours** | **Flume** | Why Flume ? - Architecture, Configuration (Agents), Sources(Exec-Avro-NetCat) |
| Channels(File,Memory,JDBC, HBase), Sinks(Logger, Avro, HDFS, Hbase, FileRoll), Contextual Routing (Interceptors, Channel Selectors) - Introduction to other aggregation frameworks |
| **Day11 - 1 Hours Lab - 2 Hours** | **Practical Development Tips and Techniques** | Testing with MRUnit |
| Debugging MapReduce Code |
| Using LocalJobRunner Mode For Easier Debugging |
| Reducing Intermediate Data With Combiners |
| Retrieving Job Information with Counters |
| **Day12 - 3 Hours Lab** | **Hadoop Case Study - Assignments** |  |
| **Day 13 - 1 Hours Lab - 2 Hours** | **Administration** | Multi Node Cluster Setup using AWS Cloud Machines –Hardware Considerations |
| Software Considerations - Commands (fsck, job, dfsadmin) |
| Schedulers in Job Tracker - RackAwareness Policy |
| Balancing - NameNode Failure and Recovery - commissioning and Decommissioning a Node – Compression Codecs |
| **Day 14 - 3 Hour** | **Apache Spark & Scala** | Understand Scala and its implementation |
| Apply Control Structures, Loops, Collection, and more |
| Master the concepts of Traits and OOPS in Scala |
| Understand functional programming in Scala |
| Get an insight into the big data challenges |
| Learn how Spark acts as a solution to these challenges |
| Install Spark and implement Spark operations on Spark Shell |
| **Day 15 - 1 Hours Lab - 2 Hours** | **Apache Spark & Scala - More depth** | Understand the role of RDDs in Spark |
| Implement Spark applications on YARN (Hadoop) |
| Stream data using Spark Streaming API |
| Implement machine learning algorithms in Spark using MLlib API |
| Analyze Hive and Spark SQL architecture |
| Implement SparkSQL queries to perform several computations |
| Understand GraphX API and implement graph algorithms |
| Implement Broadcast variable and Accumulators for performance tuning |