**HADOOP DEVELOPMENT COURSE CONTENT**

**Course Duration : 60 Hrs**

|  |  |
| --- | --- |
| **Introduction to Distributed systems**   * High Availability * Scaling * Advantages * Challenges   **Introduction to Big Data**   * What is Big data * Big Data opportunities * Big Data Challenges   **Introduction to HADOOP**   * HADOOP Distributed File System * HADOOP Architecture * Map Reduce & HDFS   **HADOOP Administration**   * HADOOP Installation & Configuration * Setting up Standalone system * Setting up pseudo distributed cluster * Setting up distributed cluster * SSH Configuration * Monitoring :  Logging, Metrics * Working with admin tools * Optimization * Cloudera Manager Installation procedure in AWS Cloud (Optional) – CDH4   **The HADOOP Distributed File System (HDFS)**   * HDFS Design & Concepts * Blocks, Name nodes and Data nodes * HDFS Federation * HDFS High-Availability * Hadoop DFS  The Command-Line Interface * Basic File System Operations * Anatomy of File Read * Anatomy of File Write * Reading Data from a Hadoop URL * Reading Data Using the File System API   **Map Reduce**   * Map and Reduce Basics. * How Map Reduce Works * Anatomy of a Map Reduce Job Run * Job Submission, Job Initialization, Task Assignment, Task Execution * Progress and Status Updates * Job Completion, Failures * Shuffling and Sorting. * Partition & Combiner * Hadoop Streaming * Use of Apache Oozie   **Map/Reduce Programming – Java Programming**   * Hands on "Word Count" in Map/Reduce in Eclipse * Sorting files using Hadoop Configuration API discussion * Emulating "grep" for searching inside a file in Hadoop * Chain Mapping API discussion * Job Dependency API discussion and Hands on * Input Format API discussion and hands on * Input Split API discussion and hands on * Custom Data type creation in Hadoop     What is Hadoop in Big Data?  Hadoop is an open-source framework for distributed storage and processing of large data sets. It uses the Hadoop Distributed File System (HDFS) and MapReduce programming model to enable scalable and parallel data processing.  Introduction to Hadoop.  This introductory level training on Big Data Hadoop gives an overview of the MapReduce programming model with the use of a simple word counting mechanism with existing tools which highlights the challenges around processing data at a big scale. Thus, this will help you learn the use of Hadoop in different real-life situations to gain a deeper appreciation of its simplicity. Following are some other important details for you to know before applying for the program. | **Hive**   * Installation * Hive Services   **Hive Shell**  **Hive Server**  **Hive Web Interface (HWI)**   * Meta store * Hive QL * OLTP vs OLTA * Working with Tables * Working with Partitions * Hive JDBC programming * User Defined Functions * Hands on Exercises   **Pig**   * Installation * Execution Types * Grunt Shell * Pig Latin * Data Processing   **Loading and Storing**  **Filtering**  **Grouping & Joining**   * Working with Functions * User Defined Functions * Hands on Exercises   **SQOOP**   * Installation * Import data from RDBMS * Export data to RDBMS * Hands on Exercises   **H-Base**   * HBase Installation in Cloud – optional (if cloud infra exists) * HBase concepts * HBase vs RDBMS * Master & Region Servers   **Introduction and overview of other Echo systems**   * Zoo Keeper * Oozie * Hue * Flume * NoSQL * Casandra * Mahout – Machine Learning Techniques Discussion and Sample Prediction Generation     **Discussion on some business use cases** |