ABOUT THIS TRAINING:

The world of **Hadoop** and “**Big Data**" can be intimidating - hundreds of different technologies with cryptic names form the ***Hadoop ecosystem***.

This comprehensive training has been designed by industry experts considering current industry job requirements to provide in-depth learning on big data and Hadoop Modules.

This industry oriented program is a combination of the training courses in Hadoop developer, Hadoop administrator, Hadoop testing, and analytics.

This Hadoop training will also prepare you for the “Big Data Certification of Cloudera- **CCP and CCA**”.

With this course, you'll not only understand what those systems are and how they fit together - but you'll go hands-on and learn how to use them to solve real business problems!

- So it's not just theory…

It's filled with hands-on activities and exercises backed with plenty discussions of use cases of different clients across domains. We believe this will help you to build your confidence on the technology and its usage.

You'll find a range of activities in this course for people at every level. If you're a project manager who just wants to learn the buzzwords, there are web UI's for many of the activities in the course that require no programming knowledge. If you're comfortable with command lines, we'll show you how to work with them too. And if you're a programmer, we'll challenge you with writing real scripts on a Hadoop system using java, Scala, Pig Latin, and Python.

YOUR TAKEAWAY FROM TRAINING:

You'll walk away from this course with a real, deep understanding of Hadoop and its associated distributed systems, and you can apply Hadoop to real-world problems.

Plus a valuable completion certificate is waiting for you at the end!

WHAT YOU WILL LEARN IN THIS BIG DATA HADOOP ONLINE TRAINING COURSE?

1. Detailed understanding of Big Data analytics
2. Master fundamentals of Hadoop 2.8 and YARN for designing distributed systems that manages "big data" using Hadoop and related technologies for storing and analyzing data at scale.
3. Understand the architecture of HDFS and MapReduce for parallel storage and parallel processing.
4. Understand the configuration choices you should make for stability, reliability and optimized task scheduling on your distributed system.
5. Analyze relational data using Hive and MySQL(Connecting Hadoop to Other DBs)
6. Analyze non-relational data using HBase
7. Use Pig and Spark to create scripts to process data on a Hadoop cluster in more complex ways.
8. Understand how Hadoop clusters are managed by YARN, Zookeeper and Hue.
9. Know how to schedule your Hadoop jobs using Oozie.
10. Collect data from a variety of sources to your Hadoop cluster using Sqoop and Flume
11. Master Hadoop administration activities like cluster managing, monitoring, administration and troubleshooting
12. Learn testing Hadoop applications using MR Unit and other automation tools.
13. Know basics of Spark, Spark RDD, Graphx, MLlib and writing Spark applications
14. Practice real-life projects using Hadoop and Apache Spark
15. Discussion on industry use cases of different clients across domains
16. Be equipped to clear Big Data Hadoop Certification. (CCP and CCA)

RECOMMENDED SKILLS PRIOR TO TAKING THIS COURSE

There is no pre-requisite to take this Big data training and to master Hadoop. But basics of UNIX, SQL and java/Python would be good. At GraduIT, we provide complimentary Self-Paced unix and Java course with our Big Data Hadoop training to brush-up the required skills so that you are good on your Hadoop learning path.

WHO SHOULD TAKE THIS BIG DATA HADOOP TRAINING COURSE?

1. Programming Developers and System Administrators
2. Experienced working professionals , Project managers
3. Big Data Hadoop Developers eager to learn other verticals like Testing, Analytics, Administration
4. Business Intelligence, Data warehousing and Analytics Professionals
5. Graduates, undergraduates eager to learn the latest Big Data technology can take this Big Data Hadoop Certification online training

BIT ON HADOOP:

Hadoop enables to BUILD AN INSIGHT-DRIVEN BUSINESS.

To be specific, Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

Almost every large company you might want to work at uses Hadoop in some way, including Amazon, Ebay, Facebook, Google, LinkedIn, IBM, Spotify, Twitter, and Yahoo! And it's not just technology companies that need Hadoop; even the New York Times uses Hadoop for processing images.

CURRICULUM:

Module 1: Understanding Big Data and Hadoop

- Introduction to Big Data And Hadoop
- Discussion on Big Data and its Sources and Challenges related to it
- Understanding attributes of Big Data and different data varieties
- Discussing Uses Cases on “Opportunity for Business” in Big Data
- Discussion on different solutions for problems related to Big Data
- Comparison of Hadoop vs traditional systems Solutions
- Understanding Complete Solution architecture from Data Acquisition to Data Analysis for Business
- Discussion on technologies getting used for End to End solution for **Big Data Analysis driven business**
- Session on PYTHON, UNIX and JAVA(Self-paced learning videos)

Module 2: Understanding Hadoop from Thousand feet View

- Bit on history of Hadoop and its evolution
- Understanding parallel processing and parallel storage architecture
- Relating MAP REDUCE and HDFS Architecture to above
- Overview of all technology stacks related to Hadoop Ecosystem.
- Discussion on Different Distributions of Hadoop, Different vendors of Hadoop
- Installation and set-up of Hadoop Cluster (Cloudera preferably)

Module 3: Understanding Parallel Storage solution with HDFS

- Understanding HDFS Architecture in detail
- Understanding Hadoop Master-Slave Architecture
- Understanding Name Node, Data Node, Secondary Name Node
- Discussion on Blocks and Data Replication
- Learning common HDFS commands and practicing them
- Discussion on Typical Production Cluster and its configurations for stability, reliability and Optimization
- Understanding Anatomy of file Read, Write operations in HDFS
- Discussion on Hadoop 2.x Cluster Architecture - Federation and High Availability

Module 4: Understanding Parallel Process solution with MapReduce

- Understanding Hadoop 2.x MapReduce Architecture (YARN Architecture)
- Understanding Hadoop 2.x MapReduce Components
- Discussion on YARN MR Application Execution Flow
- Discussion on Anatomy of MapReduce Program –Work Flow
- Discussion on basic MapReduce API Concepts
- Writing MapReduce Driver, Mappers, and Reducers using JAVA
- Demo on MapReduce program execution
- Understanding Input Splits and its relationship with HDFS Blocks
- Discuss on Advanced Concepts:
  - Distributed Cache
  - Combiner and Partitioner
  - Counters
  - Map side and Reduce side Joins
  - Use of Compression techniques (Snappy, LZO and Zip)
  - Advanced Data types in Map Reduce(Writable and WritableComparable)
- Hands-on exercises on Map Reduce Program Execution
- Demo on Telecom Data set

Module 5: Learn to analyze relational data using Hive

- Understanding of Hive Framework & Components
- Discussion on relationship between Hive and MapReduce (When to choose What)
- Understanding and hands on
  - Hive Data Types
  - **Hive DDL** – Create/Show/Drop Data Base and Tables
  - Hive DML – Load Files & Insert Data
  - Hive SQL - Select, Filter, Join, Group By
- Understanding of Internal and External Tables
- Understanding of Programming structure in UDF, Partitions and Buckets
- Discussion on Limitations of Hive.
- Discussion on HIVE on SPARK.

Module 6: Learn Data Flow ETL Scripting Language "Pig"

- Introduction to Apache Pig
- Discussion on relationship between Hive ,MapReduce and Pig
- Grunt
- Shell and Utility components
- Different data types in Pig

- Programming Structure in Pig
- Modes Of Execution in Pig
- Experiencing Pig Script by writing Evaluation, Filter, Load and Store functions
- UDFs in Pig
- Understand Integration of HBASE with Pig (After Completion of HBASE Module)

Module 7: Learn to analyze non-relational data using HBase

- Introduction to NoSQL Databases
- Understanding HBase non-relational distributed Data base and its architecture
- When/Why to use HBase
- Understanding HBase Client API
- Know how to load Data into HBase
- Know how to query Data from HBase
- Discussion on relationship between Hive ,MapReduce ,Pig and HBase
- Overview of other non-relational like Cassandra, and MongoDB and its advantages over RDBMS and Career path in NO SQL DBs.

Module 8: Learn how to Collect data from a variety of sources to your Hadoop cluster using Sqoop, Flume

- Understand Why and what is SQOOP
- Understand SQOOP Architecture
- Learn how to Importing Data Using SQOOP to HDFS/HIVE/HBase From RDBMS
- Understand Why and what is Flume
- Understand Flume Architecture
- Learn how to efficiently collect, aggregate, and move large amounts of streaming data into the Hadoop Distributed File System (HDFS)

Module 9: Learn how to schedule Hadoop jobs using Oozie

- Introduction to Work Flow Management and Oozie
- Learn Oozie Components And Workflow
- Ozzie Commands
- Experience scheduling jobs using Oozie

Module 10: Discussion on Emerging Technologies in Big Data Landscape

- Discussion on:
  - Emerging Trends In Big Data Landscape
  - Introduction to **SPARK** – In Memory Parallel Processing Framework for Real Time/Near Real Time (NRT) operations

- Best practices for Hadoop Developer
- Discussion on Interview preparation