
venolearn
EXPERIENCE BEYOND IT TRAINING

Big Data with Hadoop

Big Data is the technology that involves the management and analysis of really large chunks of data, impossible to be processed by traditional computers. Data from Social Media, Stock Exchange, Power Grids, Search Engines etc. are some examples of Big Data.

Hadoop is an open-source fault-tolerant framework that allows one to store and process data. In this course, we will learn how to work with big data using one of the most popular frameworks, Hadoop! The course will also cover Hadoop ecosystems models such as MapReduce, Hive, Pig, etc.

What are the benefits of using Hadoop?

Hadoop's design principles based on performance efficiency and developer-friendliness are of great value for anybody working in the Data Science domain. Here are some of its benefits:

- Scalability – Being a highly scalable platform, thousands of Terabytes of data could be effortlessly managed and processed in parallel using Hadoop.
- Self-healing feature – Hadoop is designed to automatically route failure in a transparent fashion. Hence, it's called a fault-tolerant system.
- Cost-effectiveness – Who doesn't want to save money? The cost-effective storage and management solutions offered by Hadoop are unmatched.

Fast and Flexible – Data extraction, processing, data warehousing, data analysis and more could be performed quickly

Why should you learn Hadoop and Big Data?

- Big data and analytics have been proven to have made considerable impact on the strategies and revenues of most businesses that employ it.
- There is an increasing need for IT professionals with Hadoop and Big Data skills.
- Hadoop offers an accelerated career growth, thanks to its wide adoption across different sectors such as social media, advertising, marketing, research, analysis, governance, etc.
- IT Professionals with Hadoop skills are offered excellent salary packages.

Curriculum:

1)Big-Data and Hadoop

- Introduction to big data and Hadoop
- Hadoop Architecture
- Installing Ubuntu with Java 1.8 on VM Workstation 11
- Hadoop Versioning and Configuration
- Single Node Hadoop 1.2.1 installation on Ubuntu 14.4.1
- Multi Node Hadoop 1.2.1 installation on Ubuntu 14.4.1
- Linux commands and Hadoop commands
- Cluster architecture and block placement
- Modes in Hadoop

    1)Local Mode

    2) Pseudo Distributed Mode

    3)Fully Distributed Mode

- Hadoop Daemon
- Master Daemons(Name Node, Secondary Name Node, Job Tracker)
- Slave Daemons(Job tracker, Task tracker)
- Task Instance
- Hadoop HDFS Commands
- Accessing HDFS

    1) CLI Approach

    2)Java Approach

2)Map-Reduce

- Understanding Map Reduce Framework
- Inspiration to Word-Count Example
- Developing Map-Reduce Program using Eclipse Luna
- HDFS Read-Write Process
- Map-Reduce Life Cycle Method
- Serialization(Java)
- Datatypes.
- Comparator and Comparable
- Custom Output File
- Analysing Temperature dataset using Map-Reduce
- Custom Partitioner & Combiner

- Running Map-Reduce in Local and Pseudo Distributed Mode

3)Advanced Map-Reduce

- Enum(Java)
- Custom and Dynamic Counters
- Running Map-Reduce in Multi-node Hadoop Cluster
- Custom Writable
- Site Data Distribution
    1) Using Configuration
    2) Using DistributedCache
    3) Using stringifie
- Input Formatters
    1) NLine Input Formatter
    2) XML Input Formatter
- Sorting
    1) Primary Reverse Sorting
    2) Secondary Sorting
- Compression Technique
- Working with Sequence File Format
- Working with AVRO File Format
- Testing MapReduce with MR Unit
- Working with NYSE DataSets
- Working with Million Song DataSets
- Running Map-Reduce in Cloudera Box

4)HIVE

- Hive Introduction & Installation
- Data Types in Hive
- Commands in Hive
- Exploring Internal and External Table Partitions
- Complex data types
- UDF in Hive
    1) Built-in UDF
    2) Custom UDF
- Thrift Server
- Java to Hive Connection
- Joins in Hive
- Working with HWI
- Bucket Map-side Join
- More commands
    1) View

2) SortBy
3) Distribute By
4) Lateral View

- Running Hive in Cloudera

5)SQOOP

- Sqoop Installations and Basics
- Importing Data from Oracle to HDFS
- Advance Imports
- Real Time UseCase
- Exporting Data from HDFS to Oracle
- Running Sqoop in Cloudera

6)PIG

- Installation and Introduction
- WordCount in Pig
- NYSE in Pig
- Working With Complex Datatypes
- Pig Schema
- Miscellaneous Command
  1) Group
  2) Filter
  3) Order
  4) Distinct
  5) Join
  6) Flatten
  7) Co-group
  8) Union
  9) Illustrate
  10) Explain

- UDFs in Pig
- Parameter Substitution and DryRun
- Pig Macros
- Running Pig in Cloudera

7)HBase

- HBase Introduction & Installation
- Exploring HBase Shell
- HBase Storage Techinique

- HBasing with Java
- CRUD with HBase
- Hive HBase Integration

8)OOZIE

- Installing Oozie
- Running Map-Reduce with Oozie
- Running Pig and Sqoop with Oozie