Top Grade Curriculum

Digital Hub Tech
IT/SOFTWARE TRAINING SERVICES

Big data Hadoop Spark
On Premises + Cloud
Course Syllabus

Institute: Digital IT Hub, Hyderabad **Website:** www.digitalhubtech.co.in **Mob:** +91 86887 09041 | +91 90596 49041

About Digital IT Hub

Digital IT Hub is one of the Institute which not only equips in Technology Skills but will train for free in Basic Soft Skills, Mock Interviews, Crack Interview Skills, Work Ethics, Corporate Values, etc that need to know to get into IT Industry. Our All Trainings are Very Special given by Expert Real Time Experienced Instructors and we enable each and every student of ours to do their own real time Project Work by the end of the Program.

We do not buy a Job by bribing Companies to secure a place in IT, instead we equip with the Skills needed to get employed in IT and will support you with number of relevant Opportunities so that career in IT becomes assured. Our support will be there before/after get placed in an IT Company as that's our mission too. If want to just have an IT Certification, You can do the Course anywhere. If you aspire to get into an IT Job, Digital IT Hub is the Best Place to choose where your IT dream will definitely come into reality.

Please go through the long list of our Student Reviews to get more about us.

About Bigdata Hadoop Spark Course

The Apache Hadoop Software framework allows for the distributed processing of large data sets with multiple Clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines. It is capable to handle all the complex Data management challenges.

Currently, Big Data Hadoop Technology is Highly demand in IT Sector and Plenty of Career opportunities are there. Work towards making the most out of the rising career opportunities in this leading Big Data domain by getting enrolled for the **Digital IT Hub Technologies** Top listed Hadoop Training in Hyderabad.

This **Real Time Deep Drive Advance Big data Hadoop Spark Training Program** will make the Learners to Expert Level.

Eligibliity/Qualifiication

Highest Degree	: B.E/B.Tech/MCA/MBA/M.Sc/MS/ME/M.Tech/BCA/BSC/BCom/BA
Experienced Candidates	: Interested to build Career in Data Platform
Professional Role	: Developer, Tester, Production Support, Data Engineer
College Students	: Any Stream of Graduates & Post-Graduates, Freshers

Additional Benefit

- ✓ Complete Guidance for Spark and Hadoop Certification
- ✓ Mock Interviews
- ✓ Soft Copy Materials
- ✓ Hard Copy Materials
- ✓ Assistance in RESUME Preparation
- ✓ All the Real time Scenarios will be provided
- ✓ Interview Questions Discussion
- ✓ 4-5 Related Courses will be given Free with Self based Training Kit (Soft Copy, Hard Copy, Recordings, Exercises, FAQ, etc.)
- ✓ Real Time Live Project Hands on
- ✓ College Project Submission
- ✓ Job Assistance

COURSE CONTENT

Part - 0 : Pre-Requisites

- ✧ Operating System : Windows, Unix/Linux
- ✧ Database : DBMS Concepts, SQL
- ✧ Data Warehouse : DWH, ETL & BI Concepts
- ✧ Programming Language : Python, Java
- ✧ Scripting Language : Shell Script, Scala
- ✧ Software Engineering : Concepts

Part - I : Big data Hadoop Eco System [Duration : 30 Hrs]

- ✧ Concepts : About Big data and Hadoop
- ✧ Hadoop Eco System : Components
- ✧ Data Storage : HDFS
- ✧ Data Processing : Map Reduce
- ✧ Memory Management : Yarn
- ✧ Distributed Application : Zoo Keeper
- ✧ Job Scheduling : Oozie
- ✧ Import/Export : SQOOP
- ✧ Relational Database/Warehouse : Hive
- ✧ Non Relational Database : Impala
- ✧ Stream Processing : Kafka
- ✧ ETL : Pig

Part - II : Spark Duration : 30 Hrs]

- ✧ Data Bricks
- ✧ Spark Introduction
- ✧ Data Frame
- ✧ Resilient Distributed Dataset (RDD)
- ✧ Spark SQL
- ✧ Spark Streaming

Part - I : Big data Hadoop Eco Sysem

❖ Big data Concepts

- Introduction to Big data
- Characteristics of Big data
- Relation between Big Data and Hadoop
- Big Data Opportunities
- Challenges with Big data
- Hadoop - Big Data Solutions
- Difference between Hadoop 1.X.X , Hadoop 2.X.X & 3.X.X Version

❖ Hadoop Eco System

- Introduction to Eco System
- Hadoop Architecture
- OLAP Database Limitation
- Uses of connected Components
- Oozie vs Zoo Keeper

❖ HDFS : Data Storage

- Introduction to HDFS
- Apache HDFS Architecture
- Cluster Environment
- How the Data stored in HDFS ?
- What is BLOCK ?
- Replication Factor in HDFS
- HDFS Commands

❖ Map Reduce : Data Processing

- Introduction to Map Reduce
- Difference between Traditional RDBMS and Map Reduce
- Map Reduce essential is in Hadoop
- Hadoop Processing Daemons
- Input Split
- Map Reduce Life Cycle
- Map Reduce Programming Model
- Map Reduce Terminologies
- Combiner and Reducer
- Serialization vs De-Serialization
- Compiling and Verifying Map-Reduce Program
- Word Count Example

❖ Yarn : Memory Management

- What is YARN?
- Difference between Map Reduce & YARN
- When to use YARN
- YARN Process Flow

- YARN Architecture
  - Resource Manager
  - Application Master
  - Node Manager
- YARN Web UI
- Different Configuration Files for YARN

❖ Zoo Keeper : Distributed Application

- What is Zoo Keeper ?
- Why Required ?
- Zoo Keeper Architecture
- Advantages and Disadvantages
- Apache Zoo Keeper Application
- Workflow
- CLI
- API

❖ Oozie : Job Scheduling

- Oozie Introduction
- Oozie Architectrure
- Oozie Configuration Files
- Oozie Job Submission
  - Workflow .xml
  - Coordinator.xml

❖ Hive : Relational Data Base/Data Warehouse

- Introduction to Hive
- Hive Architecture
  - Driver
  - Compiler
  - Executor (Semantic Analyzer)
- Need of Apache Hive in Hadoop
- Collection Data Type
- Work on Hive Database
- Hive Shell
- Meta Store
- Hive Table Operation
- Column Operation
- HIVEQL SELECT
- Views and Index
- Built-in Operators
- Buit-in Functions
- HIVEQL Joins
- Sub Queries
- Partitioning and Bucketing
- Internal vs External Table
- Hive Serializer/Deserializer
- Semi Structured Data Processing Using Hive

- Compressing and Migrating Hive Tables
- Dynamic substation of Hive and Different ways of running Hive
- ACID in HIVE
- How to enable Update in HIVE
- Log Analysis on Hive

❖ Impala : Non Relational Database

- Introduction to Impala
- General Impala Commands
- Query Processing
- Impala Table operation
- Work on Table Data
- User Permission

❖ Kafka : Stream Processing

- Introduction to Kafka
- Installation of Kafka
- Difference between MQ Vs Kafka
- Basic Operation using Kafka

Optional

❖ SQOOP : Import/Export

- Introduction to SQOOP
- SQOOP Import and Export
- SQOOP Job
- Connect to Relational Database using SQOOP
- SQOOP Commands
- Code Gen
- Eval
- Working in Database
- Table Operation

❖ Pig : ETL

- Introduction to Pig
- Pig Data Type
- Pig Execution
- Grunt Shell
- Pig Latin
- Operators
- Grouping
- Join
- Combining
- Spliting
- Filtering & Sorting
- Buit-in Functions
- UDF
- Pig Scripting

Part - II : Spark

❖ Data Bricks

✧ Introduction to Data Bricks

- Basic Concepts
- About UI
- Data Lake
- Cluster Set Up Using Data Bricks

✧ ETL

- Mordern ETL & ELT Process
- Transformation
- Data Loading Using Data Bricks

❖ Spark Introduction

✧ Objective

- Motivation for Spark
- Processing Engine Concept
- Spark Vs Map Reduce Processing
- Advantages of IN_MEMORY Processing over DISK Based
- Where to use Spark
- ROI Comparison of Hadoop Processing over Spark Processing
- Why Spark Processing is Faster than Map Reduce?
- Spark Benefits

✧ Architecture

- Hadoop Vs Spark Architectures
- Spark Master
- Spark Driver
- Spark Worker Node
- Spark Runtime Managers
  - Standalone
  - YARN
  - Apache Mesos
- Start Spark Deamons

✧ Spark Basics

- Creating Spark Context
- Creating Spark Conf, Spark Session
- File Operations in Spark Shell
- Linking and Initializing Spark
- Caching in Spark
- Real time Examples of Spark

❖ Spark Core

✧ Introduction to Resilient Distributed Datasets (RDD)

- How to create a RDD
- RDD Types
- Core Features of RDD
  - Lazily Evaluated
  - Immutable
  - Partitioned

✧ RDD Operations

- Different Transformations in RDD
- Different Actions in RDD
- Loading Data through RDD
- Saving Data

✧ Loading and Saving Data through different File Formats

- Text, CSV, TSV, JSON, PARQUET, ORC, Object files
- As a Hadoop file

✧ Key-Value Pair RDD operations

- Spark Storage Persistence Levels
- Running Spark in a Clustered Mode
- Deploying Application with spark-submit
- Cluster Management

✧ Accumulators

- Introduction to Accumulators
- Practical applicability of accumulators
- Real time examples on Accumulators

✧ Broadcast Variables

- Introduction to Broadcast variables
- Practical applicability of Broadcast variables
- Real time examples on Broadcast variables

❖ Spark SQL

✧ Introduction

- SQL Context
- Hive Vs Spark SQL
- Spark SQL support for Text Files, Parquet and JSON files
- Data Frames
- Data Sets
- Data Frames vs Data Sets – Performance Optimization
- Real Time Examples

✧ Different File Formats Support in Spark SQL

- Text
- JSON

- CSV
- ORC
- TSV
- Parquet

- **Integration with Spark SQL**
  - Data warehouse - Hive
  - RDBMS - SQL : MySQL
  - Non RDBMS - NOSQL : Cassandra

❖ Spark Processing – With different Programming Languages

- **Scala**
  - Installing Scala
  - How to use “Spark-Shell”
  - Examples on Spark with Scala

- **Python**
  - Installing Python
  - How to use “PySpark”
  - Examples on Spark with Python

- **R**
  - Installing R
  - How to use “SparkR”
  - Examples on Spark with R Language

❖ Spark Streaming

- **Introduction to Spark Streaming**
  - Architecture of Spark Streaming
  - Streamings : DStreams, SSC, Kafka, Flume

- **DStreams**
  - RDD vs Discretized Streams (DStreams)
  - DStream Operations
    - Window Operations
    - Transform Operations

Optional

❖ Spark MLib

- Introduction to Machine Learning
- Vector Class in MLib
- Spark MLib Algorithms introduction
- Classification and Regression Algorithms
- Naïve Bayes Classification Algorithm
- Decision Trees Algorithm Overview

❖ Spark Project

- Real Time Projects On Spark with Hadoop Integration
- Proof Of Concepts (POCs)

****