Hadoop Frameworks Training Curriculum

STRUCTURE

www.cromacampus.com | helpdesk@cromacampus.com | +91-120-4155255

Hadoop Frameworks Training Curriculum

"Master various popular Hadoop Frameworks and make yourself stand out in the IT and the Big Data space."

List of Popular Hadoop Frameworks covered in the course:

1. Apache PIG
2. Apache Spark with Scala
3. Apache HIVE
4. Apache SQOOP
5. Apache HBase
6. Apache Flume
7. Apache Drill
8. Apache Kafka
9. Apache Storm

Apache PIG Training

Course Objectives:

Pig is a high-level platform for creating MapReduce programs used with Hadoop. The language for this platform is called Pig Latin. In this course we will go through the PIG data flow platform and the language used by PIG tool. The concepts which are covered in this course are:

- Writing complex MapReduce transformations using a simple scripting language.
- Basics of Big Data, Hadoop and MapReduce Framework.
- PIG Data Model and Different type of operators to operate on datasets.
- Built-in Functions as well as User Defined Functions for performing a specific task.
- Running PIG Script, Unit Testing and Compression.
- Many more advance topics such as Embedding PIG in Java, PIG Macros etc.

Course Content:

Module 1: Introduction

- Big Data Overview
- Apache Hadoop Overview
- Hadoop Distribution File System
- Hadoop MapReduce Overview
- Introduction to PIG
- Prerequisites for Apache PIG
- Exploring use cases for PIG
- History of Apache PIG
- Why you need PIG?
- Significance of PIG
- PIG over MapReduce
- When PIG suits the most?
- When to avoid PIG?

Module 2: PIG Architecture

- PIG Latin Language
- Running PIG in Different Modes
- PIG Architecture
- GRUNT Shell
- PIG Latin Statements
- Running Pig Scripts
- Utility Commands

Module 3: Data Models, Operators, and Streaming in PIG

- PIG Data Model- Scalar Data type
- PIG Data Model - Complex Data Type
- Arithmetic Operators
- Comparison Operators
- Cast Operators
- Type Construction Operator
- Relational Operators
- Loading and Storing Operators
- Filtering Operators
- Filtering Operators-Pig Streaming with Python
- Grouping and Joining Operators-
- Sorting Operator
- Combining and Splitting Operators
- Diagnostic Operators

Module 4: Functions in PIG

- Eval Functions
- Load and Store Functions
- Tuple and Bag Functions
- String Functions
- Math Function

Module 5: Advanced Concepts in PIG

- File compression in PIG
- Intermediate Compression
- Pig Unit Testing
- Embedded PIG in JAVA
- Pig Macros
- Import Macros
- Parameter Substitutions

Apache Spark with Scala Training

Course Objectives:

- Frame big data analysis problems as Apache Spark scripts
- Develop distributed code using the Scala programming language
- Optimize Spark jobs through partitioning, caching, and other techniques
- Build, deploy, and run Spark scripts on Hadoop clusters
- Process continual streams of data with Spark Streaming
- Transform structured data using Spark SQL, Data Sets, and Data Frames
- Traverse and analyze graph structures using Graph X
- Analyze massive data set with Machine Learning on Spark

Course Content:

Module 1: Introduction

- Big Data Overview
- Apache Hadoop Overview
- Hadoop Distribution File System
- Hadoop MapReduce Overview
- Introduction to IntelliJ and Scala
- Installing IntelliJ and Scala
- Apache Spark Overview
- What's new in Apache Spark 3?

Module 2: Scala

- Scala Basics
- Flow control in Scala
- Functions in Scala
- Data Structures in Scala

Module 3: Using Resilient Distributed Datasets

- The Resilient Distributed Dataset
- Ratings Histogram Example
- Key / Value RDD's, and the Average Friends by Age example
- Filtering RDD's, and the Minimum Temperature by Location Example
- Check Your Results and Implementation Against Mine

Module 4: Spark SQL, Data Frames, and Data Sets

- Introduction to Spark SQL
- What are Data Frames?
- What are Data Sets?
- Item-Based Collaborative Filtering in Spark, cache (), and persist ()

Module 5: Running Spark on a cluster

- What is a Cluster?

- Cluster management in Hadoop
- Introducing Amazing Elastic MapReduce
- Partitioning Concepts
- Troubleshooting and managing dependencies

Module 6: Machine Learning with Spark ML

- Introducing MLLib
- Using MLLib
- Linear Regression with MLLib

Module 7: Spark Streaming

- Spark Streaming
- The DStream API for Spark Streaming
- Structured Streaming

Module 8: Graph X

- What is Graph X?
- About Pregel
- Breadth-First-Search with Pregel
- Using Pregel API with Spark API

Apache HIVE Training

Course Objectives:

- Install and Work on Hive
- Troubleshoot Hive Issues
- Partition and bucket data

Course Content:

Module 1: Introduction

- Big Data Overview
- Hadoop Overview
- What is a Hadoop Framework?
- Types of Hadoop Frameworks
- What is Hive?
- Motivation Behind the Tool
- Hive use cases
- Hive Architecture
- Different Modes of HIVE

Module 2: Installing and Configuring HIVE

- Downloading, installing, and configuring HIVE
- Hive Shell Commands
- Different configuration properties in HIVE
- Beeswax
- Installing and configuring MySQL Database
- Installing Hive Server

Module 3: Working on HIVE

- Databases in Hive
- Datatypes in Hive
- Schema on Read
- Schema on Write
- Download Datasets
- Internal Tables
- External Tables
- Partition in HIVE
- Bucketing in HIVE

Module 4: HIVE Implementation

- Hive in Real Time Projects
- Auditing in Hive
- Troubleshooting Infra issues in Hive
- Troubleshooting User issues in Hive

Apache Sqoop Training Curriculum

Course Objectives:

- Understand lifecycle of Sqoop command.
- Use Sqoop import command to migrate data from MySQL to HDFS.
- Use Sqoop import command to migrate data from MySQL to Hive.
- Use various file formats, compressions, file delimiter, where clause and queries while importing the data.
- Understand split-by and boundary queries.
- Use incremental mode to migrate the data from MySQL to HDFS.
- What is Sqoop export?
- Using Sqoop export, migrate data from HDFS to MySQL.
- Using Sqoop export, migrate data from Hive to MySQL

Course Content:

Module 1: Hadoop Overview

- Course overview
- Big Data Overview
- Hadoop Overview
- HDFS
- YARN Cluster Overview
- Cluster Setup on Google Cloud
- Environment Update

Module 2: Sqoop Overview

- Sqoop Introduction
- Why Sqoop?
- Sqoop Features
- Flume vs Sqoop
- Sqoop Architecture & Working
- Sqoop Commands.

Module 3: Sqoop Import

- Managing Target Directories
- Working with Parquet File Format
- Working with Avro File Format
- Working with Different Compressions
- Conditional Imports
- Split-by and Boundary Queries
- Field delimiters
- Incremental Appends
- Sqoop Hive Import
- Sqoop List Tables/Database

Module 4: Sqoop Export

- Export from HDFS to MySQL
- Export from Hive to MySQL
- Export Avro Compressed to MySQL
- Sqoop with Airflow

Module 5: Career Guidance and Roadmap

Apache HBase Training Curriculum

Course Objectives:

- Understand the fundamentals of HBase
- Grab data from various RDBMS/Flat files into the HBASE systems
- Understand the prerequisites necessary to get started with HBase
- Understand table design and perform CRUD operations
- Install and configure a new HBase cluster
- Find out how the communication between the client and server happens in HBase
- Optimize an HBase cluster using different Hadoop and HBase parameters
- Get to know the concepts of scaling with HBase through practical examples

Module 1: Hadoop Overview

- Course overview
- Big Data Overview
- Hadoop Overview
- HDFS
- Hadoop Ecosystem
- What is a Hadoop Framework?
- Types of Hadoop frameworks

Module 2: No SQL Databases HBase

- NoSQL Databases HBase
- NoSQL Introduction
- HBase Overview
- HBase Architecture
- Data Model
- Connecting to HBase
- HBase Shell

Module 3: Administration in HBASE

- Introduction
- Learn and Understand HBase Fault Tolerance
- Hardware Recommendations
- Software Recommendations
- HBase Deployment at Scale
- Installation with Cloudera Manager
- Basic Static Configuration
- Rolling Restarts and Upgrades
- Interacting with HBase

Module 4: Troubleshooting in HBASE

- Introduction
- Troubleshooting Distributed Clusters

- Learn How to Use the HBase UI
- Learn How to Use the Metrics
- Learn How to Use the Logs

Module 5: Tuning in HBASE

- Introduction
- Generating Load & Load Test Tool
- Generating With YCSB
- Region Tuning
- Table Storage Tuning
- Memory Tuning
- Tuning with Failures
- Tuning for Modern Hardware

Module 6: Apache HBase Operations Continuity

- Introduction
- Corruption: hbck
- Corruption: Other Tools
- Security
- Security Demo
- Snapshots
- Import, Export and Copy Table
- Cluster Replication

Module 7: Apache HBASE Ecosystem

- Introduction
- Hue
- HBase With Apache Phoenix'

Module 8: Career Guidance and Roadmap

Apache Flume Training Curriculum

Course Objectives:

- Learn Flume fundamentals, Architecture, Data flow mode, Reliability and Recoverability.
- Get hands-on training on aggregating, streaming data flows into HDFS, Apache Flume Data Transfer in Hadoop, process & analyse data., etc.
- You will also get an exposure to industry based Real-time projects in various verticals.
- Get certified in Apache Flume and start applying for Jobs

Course Content:

Module 1: Overview

- Course overview
- Big Data Overview
- Hadoop Overview
- HDFS
- Hadoop Ecosystem
- What is a Hadoop Framework?
- Types of Hadoop frameworks
- Flume Overview
- Architecture
- Data flow mode
- Reliability and Recoverability

Module 2: Setting up Agents

- Setting up an individual agent
- Configuring individual components
- Wiring the pieces together
- Data ingestion
- Executing Commands
- Network streams
- Setting Multi-Agent Flow
- Consolidation
- Multiplexing the flow
- Configuration
- Defining the flow
- Configuring individual components
- Adding multiple flows in an agent

Module 3: Configuring A Multi Agent Flow

- Fan out flow
- Flume Sources
- Avro Source, Exec Source
- NetCat Source

- Sequence Generator Source
- Syslog Sources
- Syslog TCP Source
- Syslog UDP Source
- Legacy Sources
- Avro Legacy Source
- Thrift Legacy Source
- Custom Source

Module 4: Flume Sinks

- HDFS Sink
- Logger Sink
- Avro Sink
- IRC Sink
- File Roll Sink
- Null Sink
- HbaseSinks
- HbaseSink
- AsyncHBaseSink
- Custom Sink

Module 5: Flume Channels

- Memory Channel
- JDBC Channel
- Recoverable Memory Channel
- File Channel, Pseudo Transaction Channel
- Custom Channel
- Flume Channel Selectors
- Replicating Channel Selector
- Multiplexing Channel Selector
- Custom Channel Selector

Module 6: Flume Sink Processors

- Default Sink Processor
- Failover Sink Processor
- Load balancing Sink Processor
- Custom Sink Processor

Module 7: Flume Interceptors

- Timestamp Interceptor
- Host Interceptor
- Flume Properties
- Property

Module 8: Security

- Monitoring
- Troubleshooting
- Handling agent failures
- Compatibility
- HDFS
- AVRO

Module 9: Career Guidance and Roadmap

Apache Drill course Curriculum

Course Objectives:

- Learn how to quickly cleanse, manipulate, and analyze data using Drill
- Access Drill programmatically using Python or R
- Use Drill to connect to and query multiple data sources

Course Content:

Module 1: Overview

- Course overview
- Big Data Overview
- Hadoop Overview
- HDFS
- Hadoop Ecosystem
- What is a Hadoop Framework?
- Types of Hadoop frameworks
- Drill Overview
- What does Drill do?
- How does Drill work?
- Kinds of data which can be queried with Drill

Module 2: Installing & Configuring Drill

- Comparison of embedded and distributed modes
- Introducing and configuring workspaces
- Demonstrate Drill's various interfaces

Module 3: Querying Simple Delimited Data

- SQL fundamentals
- Querying a simple CSV file
- Arrays in Drill
- Accessing columns in Arrays

Module 4: Configuration Options

- Extracting headers from csv files
- Changing delimiter characters
- Specifying options in a query

Module 5: Understanding Data Types and Functions in Drill

- Overview of Drill Data Types
- Converting Strings to Numeric Data Types
- Complex Conversions
- Windowing functions

Module 6: Working with Dates and Times in Drill

- Understanding dates and times in Drill
- Converting strings to dates
- Reformatting dates
- Intervals and date/time arithmetic in Drill

Module 7: Analyzing Nested Data with Drill

- Issues querying nested data with Drill
- Maps and Arrays in Drill
- Querying deeply nested data in Drill

Module 8: Other Data Types

- Log files
- HTTPD

Module 9: Connecting Multiple Data Sources and programming languages

- MySQL
- Hadoop
- MongoDB
- Python
- R programming language
- Issues when using other data sources'

Module 10: Career Guidance and Roadmap

Apache Kafka Training Curriculum

Course Objectives:

- Learn Kafka and its components
- Set up an end to end Kafka cluster along with Hadoop and YARN cluster
- Integrate Kafka with real time streaming systems like Spark & Storm
- Describe the basic and advanced features involved in designing and developing a high throughput messaging system
- Use Kafka to produce and consume messages from various sources including real time streaming sources like Twitter
- Get an insight of Kafka API and Understand Kafka Stream APIs

Course Content:

Module 1: Overview

- Introduction to Big Data
- Big Data Analytics
- Need for Kafka
- What is Kafka?
- Kafka Features
- Kafka Concepts
- Kafka Architecture
- Kafka Components
- Zookeeper
- Where is Kafka Used?
- Kafka Installation
- Kafka Cluster
- Types of Kafka Clusters
- Configuring Single Node Single Broker Cluster

Module 2: Kafka Producer

- Configuring Single Node Multi Broker Cluster
- Constructing a Kafka Producer
- Sending a Message to Kafka
- Producing Keyed and Non-Keyed Messages
- Sending a Message Synchronously & Asynchronously
- Configuring Producers
- Serializers
- Serializing Using Apache Avro
- Partitions

Module 3: Kafka Consumers

- Consumers and Consumer Groups
- Standalone Consumer
- Consumer Groups and Partition Rebalance

- Creating a Kafka Consumer
- Subscribing to Topics
- The Poll Loop
- Configuring Consumers
- Commits and Offsets
- Rebalance Listeners
- Consuming Records with Specific Offsets
- De-serializers

Module 4: Kafka Internals

- Cluster Membership
- The Controller
- Replication
- Request Processing
- Physical Storage
- Reliability
- Broker Configuration
- Using Producers in a Reliable System
- Using Consumers in a Reliable System
- Validating System Reliability
- Performance Tuning in Kafka

Module 5: Cluster Architecture and Administering Kafka

- Use Cases - Cross-Cluster Mirroring
- Multi-Cluster Architectures
- Apache Kafka's Mirror Maker
- Other Cross-Cluster Mirroring Solutions
- Topic Operations
- Consumer Groups
- Dynamic Configuration Changes
- Partition Management
- Consuming and Producing
- Unsafe Operations

Module 6: Kafka Monitoring & Kafka Connect

- Considerations When Building Data Pipelines
- Metric Basics
- Kafka Broker Metrics
- Client Monitoring
- Lag Monitoring
- End-to-End Monitoring
- Kafka Connect
- When to Use Kafka Connect?
- Kafka Connect Properties

Module 7: Kafka Stream Processing

- Stream Processing
- Stream-Processing Concepts
- Stream-Processing Design Patterns
- Kafka Streams by Example
- Kafka Streams: Architecture Overview

Module 8: Kafka Integration with Hadoop, Storm, and Spark

- Apache Hadoop Basics
- Hadoop Configuration
- Kafka Integration with Hadoop
- Apache Storm Basics
- Configuration of Storm
- Integration of Kafka with Storm
- Apache Spark Basics
- Spark Configuration
- Kafka Integration with Spark

Module 9: Kafka Integration with Flume, Talend and Cassandra

- Flume Basics
- Integration of Kafka with Flume
- Cassandra Basics such as and Key Space and Table Creation
- Integration of Kafka with Cassandra
- Talend Basics
- Integration of Kafka with Talend

Module 10: Career Guidance and Roadmap

Apache Storm Training Curriculum

Course Objectives:

- Apache Storm is an open-source and distributed stream processing computation framework used for processing large volumes of high-velocity data.
- This training will help you learn reliable real-time data processing capabilities of Storm and, how Storm is different from Hadoop & Kafka.
- You can smartly use Apache Storm at various place such as Ecommerce, Supply chain, Streaming etc.

Course Content:

Module 1: Introduction to Big Data and Real Time Big data processing

- Big Data
- Hadoop
- Batch Processing
- Real-time analytics
- Storm origin
- Architecture
- Comparison with Hadoop and Spark

Module 2: Storm installation and groupings

- Installation of Storm
- Nimbus Node
- Supervisor Nodes
- Worker Nodes
- Running Modes
- Local Mode
- Remote Mode
- Stream Grouping
- Shuffle Grouping
- Fields Grouping
- All Grouping
- Custom Grouping
- Direct Grouping
- Global Grouping
- None Grouping

Module 3: Storm Spouts & Bolts

- Basic components of Apache Storm
- Spout
- Bolts
- Running Mode in Storm
- Reliable and unreliable messaging

- Spouts
- Introduction
- Data fetching techniques
- Direct Connection
- Enqueued message
- DRPC
- How to create custom Spouts
- Introduction to Kafka Spouts
- Bolts
- Bolt Lifecycle
- Bolt Structure
- Reliable and Unreliable Bolts
- Basic topology example using Spout and bolts
- Storm UI

Module 4: Apache Storm and Kafka

- What is Apache Kafka?
- Setting up Standalone Kafka
- How to use Kafka Producer
- How to use Kafka Consumer
- Hand on Kafka
- How Kafka Spout works in Apache Storm and its configuration?

Module 5: Trident Topology

- Trident Design
- Trident in Storm
- RQ Class, Coordinator, Emitter bolt
- Committer Bolts, Partitioned Transactional Spouts
- Transaction Topologies

Module 6: Career Guidance and Roadmap