Big Data Hadoop Architect Online Training (Big Data Hadoop + Apache Spark & Scala + MongoDB Developer And Administrator + Apache Cassandra + Impala Training + Apache Kafka + Apache Storm)

1 Big Data Hadoop

1. Introduction
   - About this Course
   - About Big Data
   - Course Logistics
   - Introductions

2. Hadoop Fundamentals
   - The Motivation for Hadoop
   - Hadoop Overview
   - HDFS
   - MapReduce
   - The Hadoop Ecosystem
   - Lab Scenario Explanation
   - Hands-On Exercise: Data Ingest with Hadoop Tools

3. Introduction to Pig
   - What Is Pig?
   - Pig’s Features
   - Pig Use Cases
   - Interacting with Pig

4. Basic Data Analysis with Pig
   - Pig Latin Syntax
   - Loading Data
   - Simple Data Types
   - Field Definitions
   - Data Output
   - Viewing the Schema
   - Filtering and Sorting Data
5. Processing Complex Data with Pig
- Storage Formats
- Complex/Nested Data Types
- Grouping
- Built-in Functions for Complex Data
- Iterating Grouped Data
- Hands-On Exercise: Analyzing Ad Campaign Data with Pig

6. Multi-Dataset Operations with Pig
- Techniques for Combining Data Sets
- Joining Data Sets in Pig
- Set Operations
- Splitting Data Sets
- Hands-On Exercise: Analyzing Disparate Data Sets with Pig

7. Extending Pig
- Adding Flexibility with Parameters
- Macros and Imports
- UDFs
- Contributed Functions
- Using Other Languages to Process Data with Pig
- Hands-On Exercise: Extending Pig with Streaming and UDFs

8. Pig Troubleshooting and Optimization
- Troubleshooting Pig
- Logging
- Using Hadoop’s Web UI
- Optional Demo: Troubleshooting a Failed Job with the Web UI
- Data Sampling and Debugging
- Performance Overview
- Understanding the Execution Plan
- Tips for Improving the Performance of Your Pig Jobs

9. Introduction to Hive
- What Is Hive?
- Hive Schema and Data Storage
10. Relational Data Analysis with Hive
- Hive Databases and Tables
- Basic HiveQL Syntax
- Data Types
- Joining Data Sets
- Common Built-in Functions
- Hands-on Exercise: Running Hive Queries on the Shell, Scripts, and Hue

11. Hive Data Management
- Hive Data Formats
- Creating Databases and Hive-Managed Tables
- Loading Data into Hive
- Altering Databases and Tables
- Self-Managed Tables
- Simplifying Queries with Views
- Storing Query Results
- Controlling Access to Data
- Hands-on Exercise: Data Management with Hive

12. Text Processing with Hive
- Overview of Text Processing
- Important String Functions
- Using Regular Expressions in Hive
- Sentiment Analysis and N-Grams
- Hands-on Exercise (Optional): Gaining Insight with Sentiment Analysis

13. Hive Optimization
- Understanding Query Performance
- Controlling Job Execution Plan
- Partitioning
- Bucketing
- Indexing Data

14. Extending Hive
Big Data Hadoop Architect Online Training

- SerDes
- Data Transformation with Custom Scripts
- User-Defined Functions
- Parameterized Queries
- Hands-on Exercise: Data Transformation with Hive

15. Introduction to Impala
- What is Impala?
- How Impala Differs from Hive and Pig
- How Impala Differs from Relational Databases
- Limitations and Future Directions
- Using the Impala Shell

2 Apache Spark & Scala

1 Introduction to Spark
- Limitations of MapReduce in Hadoop Objectives
- Batch vs. Real-time analytics
- Application of stream processing
- How to install Spark
- Spark vs. Hadoop Eco-system

2 Introduction to Programming in Scala
- Features of Scala
- Basic data types and literals used
- List the operators and methods used in Scala
- Concepts of Scala

3 Using RDD for Creating Applications in Spark
- Features of RDDs
- How to create RDDs
- RDD operations and methods
- How to run a Spark project with SBT
- Explain RDD functions and describe how to write different codes in Scala

4 Running SQL queries Using SparkSQL
- Explain the importance and features of SparkSQL

Web - www.multisoftvirtualacademy.com  Email - info@multisoftvirtualacademy.com
5 Spark Streaming
- Explain the concepts of Spark Streaming
- Describe basic and advanced sources
- Explain how stateful operations work
- Explain window and join operations

6 Spark ML Programming
- Explain the use cases and techniques of Machine Learning (ML)
- Describe the key concepts of Spark ML
- Explain the concept of an ML Dataset, and ML algorithm, model selection via cross validation

7 Spark GraphX Programming
- Explain the key concepts of Spark GraphX programming
- Limitations of the Graph Parallel system
- Describe the operations with a graph
- Graph system optimizations

3 MongoDB Developer And Administrator

1 An Overview of the Course
- Introduction to the course
- Table of Contents
- Course Objectives
- Course Overview
- Value to Professionals and Organizations

2 MongoDB A Database for the Modern Web
- MongoDB-A Database for the Modern Web
- Objectives
- What is MongoDB?
- JSON
- JSON Structure
- BSON
- MongoDB Structure
3 CRUD Operations in MongoDB

- CRUD Operations in MongoDB
- Objectives
- Data Modification in MongoDB
- Batch Insert in MongoDB
- Ordered Bulk Insert
- Performing Ordered Bulk Insert
- Unordered Bulk Insert
- Performing Un-ordered Bulk Insert
- Inserts: Internals and Implications
Performing an Insert Operation
Retrieving the documents
Specify Equality Condition
Retrieving Documents by Find Query
$in, $or, and “AND” Conditions
$or Operator
Specify AND/OR Conditions
Retrieving Documents by Using FindOne, AND/OR Conditions
Regular Expression
Array Exact Match
Array Projection Operators
Retrieving Documents for Array Fields
$Where Query
Cursor
Retrieving Documents Using Cursor
Pagination
Pagination: Avoiding Larger Skips
Advance query option
Update Operation
Updating Documents in MongoDB
$SET
Updating Embedded Documents in MongoDB
Updating Multiple Documents in MongoDB
$Unset and $inc Modifiers
$inc modifier to increment and decrement
Replacing Existing Document with New Document
$Push and $addToSet
Positional Array Modification
Adding Elements into Array Fields
Adding Elements to Array Fields Using AddToSet
Performing AddToSet
Upsert
Removing Documents
Performing Upsert and Remove Operation

4 Indexing and Aggregation
Indexing and Aggregation
Objectives
Introduction to Indexing
- Types of Index
- Properties of Index
- Single Field Index
- Single Field Index on Embedded Document
- Compound Indexes
- Index Prefixes
- Sort Order
- Ensure Indexes Fit RAM
- Multi-Key Indexes
- Compound Multi-Key Indexes
- Hashed Indexes
- TTL Indexes
- Unique Indexes
- Sparse Indexes
- Demo — Create Compound, Sparse, and Unique Indexes
- Text Indexes
- Demo — Create Single Field and Text Index
- Text Search
- Index Creation
- Index Creation on Replica Set
- Remove Indexes
- Modify Indexes
- Demo Drop and Index from a Collection
- Rebuild Indexes
- Listing Indexes
- Demo Retrieve Indexes for a Collection and Database
- Measure Index Use
- Demo — Use Mongo Shell Methods to Monitor Indexes
- Control Index Use
- Demo — Use the Explain, $Hint and $Natural Operators to Creation
- Index Use Reporting
- Geospatial Index
- Demo Create Geospatial Index
- MongoDBs Geospatial Query Operators
- Demo — Use Geospatial Index in a Query
- $GeoWithin Operator
- Proximity Queries in MongoDB Aggregation
- Pipeline Operators and Indexes
- Aggregate Pipeline Stages
The Aggregation Example
- Use Aggregate Function
- MapReduce
- Use MapReduce in MongoDB

Aggregation Operations
- Use Distinct and Count Methods
- Use the Group Function

Aggregation Operations (contd.)
- Use the Group Function

5 Replication and Sharding
- Replication and Sharding
- Objectives
- Introduction to Replication
- Master-Slave Replication
- Replica Set in MongoDB
- Replica Set in MongoDB (contd.)
- Automatic Failover
- Replica Set Members
- Priority 0 Replica Set Members
- Hidden Replica Set Members
- Delayed Replica Set Members
- Delayed Replica Set Members (contd.)
- Demo-Start a Replica Set

- Write Concern
- Write Concern (contd.)
- Write Concern Levels
- Write Concern for a Replica Set
- Modify Default Write Concern

- Read Preference
- Read Preference Modes
- Blocking for Replication

- Tag Set
- Configure Tag Sets for Replica set
- Replica Set Deployment Strategies
Big Data Hadoop Architect Online Training

- Replica Set Deployment Strategies (contd.)
- Replica Set Deployment Patterns

- Opglog File
- Replication State and Local Database
- Replication Administration
- Demo—Check a Replica Set Status
- Sharding
- When to Use Sharding?
- What is a Shard?
- What is a Shard Key
- Choosing a Shard Key
- Ideal Shard Key
- Range-Based Shard Key
- Hash-Based Sharding
- Impact of Shard Keys on Cluster Operation

- Production Cluster Architecture
- Config Server Availability
- Production Cluster Deployment
- Deploy a Sharded Cluster
- Add Shards to a Cluster
- Demo—Create a Sharded Cluster
- Enable Sharding for Database
- Enable Sharding for Collection
- Enable Sharding for Collection (contd.)
- Maintaining a Balanced Data Distribution
- Splitting

- Chunk Size
- Special Chunk Type
- Shard Balancing
- Shard Balancing (contd.)
- Customized Data Distribution with Tag Aware Sharding
- Tag Aware Sharding
- Add Shard Tags
- Remove Shard Tags

6 Developing Java and Node JS Application with MongoDB
7 Administration of MongoDB Cluster Operations

- Administration of MongoDB Cluster Operations
- Objectives
- Capped Collection
4 Apache Cassandra
1. Introduction to Cassandra Enterprise

- Comparison of RDBMS and NOSQL
- What is Cassandra?
- Why Cassandra?

2. Cassandra Enterprise Operations and Performance Tuning

- Environment Setup Introduction
- Vnodes and Single-Token Nodes
- Managing Cassandra and Adding Nodes
- Best Practices for Adding Nodes
- Maintaining Cassandra
- Performance tuning
- Environment Tuning
- Cassandra Tunning
- Disk Tuning

3. Cassandra Enterprise Search with Apache Solr

- Introduction to Cassandra Enterprise Search
- Functional Use cases
- Text Search
- Solr Search Queries
- Solr Schema
- Solr Core

4. Cassandra Core Concepts

- Prepare the Operating System
- Install a Cassandra Distribution
- Configure
- Start and Stop Cassandra
- How to use Nodetool
- CQLSH command
- CCM Overview
- Internal Architecture: Request Coordination
- Replication
- Consistency Level
Big Data Hadoop Architect Online Training

- Anti Entropy Operation
- How Nodes Communicate
- System Key-Space
- Introduction to Data Model and CQL
- Column Families
- CF to CQL
- CQL to DDL
- Clustering Order by
- Secondary Index
- UUID, Timestamp, Counter
- Collections, UDT, Tuple
- DevCenter
- Insert, Update, DELETE, TTL, LWT, Upsert
- Data Model Functions

5. Data Modeling with Cassandra Enterprise

- Introduction to KillrVideo
- Quick Wins
- Conceptual Data Modelling
- Relationship Keys
- Hierarchy
- Modelling Methodologies
- Chebotko Diagrams
- Mapping Rules
- Mapping Patterns
- Data Duplication
- Data Consistency
- Transactions
- Data Aggregations

6. Cassandra Enterprise Analytics with Apache Spark

- Spark Architecture
- Spark Shell
- Web UI
- Resilient Distributed Datasets
- Ways to create an RDD
- RDD Transformations
5 Impala Training

1 An Introduction to Impala

- An overview to the Impala
- What is Impala?
- The benefits of Impala
- Exploratory Business Intelligence
- The Impala Installation
- Starting and Stopping Impala
- Data Storage
- Managing Metadata
- Controlling Access to Data
- Impala Shell Commands and Interface

2 Querying with Hive and Impala
Big Data Hadoop Architect Online Training

- Querying with Hive and Impala
- SQL Language Statements
- DDL Statements
- CREATE the DATABASE
- CREATE the TABLE
- Internal and External Tables
- Loading Data in Impala Table
- The ALTER TABLE
- The DROP TABLE
- What is DROP DATABASE?
- Describing the Statement
- Explaining the Statement
- SHOW the TABLE Statement
- INSERT Statement
- SELECT Statement
- Data Type
- The Operators
- About the Functions
- The CREATE VIEW in Impala
- Hive and Impala Query Syntax

3 Data Storage and File Format

- About the Data Storage and File Format
- The Partitioning Tables
- SQL Statements for Partitioned Tables
- File Format and Performance Considerations
- Choosing the File Type and Compression Technique

4 Working with the Impala

- Working with the Impala
- Know Impala Architecture
- What is Impala Daemon?
- About the Impala Statestore
- Impala Catalog Service
- Query Execution Flow in Impala
- User - Defined Functions
6 Apache Kafka

1 Big Data Overview

- Big Data—Introduction
- Three Vs of Big Data
- The Data Volume
- The Data Sizes
- The Data Velocity
- The Data Variety
- The Data Evolution
- The Features of Big data
- About the Industries with Examples
- What is the Big Data Analysis?
- The Technology Comparison
- The Stream
- Apache Hadoop
- Hadoop Distributed File System
- MapReduce
- About the Real-Time Big Data Tools
- Apache Kafka
- Apache Storm
- Apache Spark
- Apache Cassandra
- Apache Hbase
- The Real-Time Big Data Tools—Uses
- The Real-Time Big Data—Use Cases

2 An Introduction to the Zookeeper

- ZooKeeper—Introduction
- Distributed Applications
- Challenges for Distributed Applications
- Partial Failures
- Race Conditions
3 Introduction to the Kafka

- Apache Kafka—Introduction
- Kafka History
- Kafka Use Cases
- Aggregating User Activity Using Kafka—Example
- Kafka Data Model
- Topics
- Partitions
- Partition Distribution
- Producers
- Consumers
- Kafka Architecture
- Types of Messaging Systems
- Queue System—Example
- Publish-Subscribe System—Example
- Brokers
- Kafka Guarantees
- Kafka at LinkedIn
- Replication in Kafka
- Persistence in Kafka
4 About Installation and Configuration

- Installation and Configuration
- Kafka Versions
- OS Selection
- Machine Selection
- Preparing for Installation
- Stop the Kafka Server

5 About the Kafka Interfaces

- An Introduction to the Kafka Interfaces—
- How to Create a Topic?
- How to Modify a Topic?
- What are Kafka-topics.sh Options?
- Creating a Message
- What are the kafka-console-producer.sh Options?
- Reading a Message
- The kafka-console-consumer.sh Options
- Reading a Message—
- Java Interface to Kafka
- Producer Side API
- The Consumer Side API
- Compiling a Java Program
- Running the Java Program
- Java Interface Observations

7 Apache Storm

1 Overview of Big Data

- Big Data: An Overview
- The objective of learning Big Data
- What are Big data?
- 3 Vs of Big Data
- About Data Volume
- The Data Sizes
- The Velocity of Data
Big Data Hadoop Architect Online Training

- About the Variety of Data
- The Data Evolution
- Features of Big data
- The Industry Examples
- The Big data Analysis
- What is Technology Comparison
- Know the Apache Hadoop
- HDFS
- MapReduce
- The Real Time Big data with examples
- The Real Time Big data Tools
- Zookeeper

2 Introduction to Storm

- What is Apache Storm
- What are the Uses of Storm?
- What is a Stream?
- The Industry use cases for STORM
- About STORM Data Model
- Describing the Storm Architecture
- About the Storm Processes
- The Sample Program
- About the Storm Components
- What is Storm Spout:
- What is Storm Bolt?
- The Storm Topology
- The Examples of Storm
- Serialization-Deserialization
- How to Submit a Job to Storm
- The Types of Topologies

3 Installation and Configuration

- What are Storm Versions?
- About the OS selection
- Describing the Machine Selection
How to Prepare for Installation
- Download Kafka
- Download Storm
- Installing Kafka Demonstrations
- Setting Up Multi-node Storm Cluster

4 Storm Advanced Concepts
- Defining the Storm Advanced Concepts
- The Types of Spouts
- About the Structure of Spout
- The Structure of Bolt
- The Stream Groupings
- Reliable Processing in Storm
- What are Ack and Fail?
- Defining Ack Timeout
- What is Anchoring?
- The Topology Lifecycle
- The Data Ingestion in Storm
- Data Ingestion in Storm Example
- Data Ingestion in Storm Check Output
- Screen Shots for the Real Time Data Ingestion
- Spout Definition
- The definition of Bolt
- Topology–Connecting Spout and Bolt
- What is Wrapper Class?

5 Storm Interfaces
- Introduction to the Storm Interfaces
- The Java Interface to Storm
- Compiling and running the Java interface to Storm (Demonstration)
- About the Spout Interface
- About the IRichSpout Methods
- About the BaseRichSpout Methods
- What is OutputFieldsDeclarer Interface?
- Spout Definition Examples
Big Data Hadoop Architect Online Training

- About the Bolt Interface
- The Irichbolt Methods
- What are Baserichbolt Methods?
- The Ibasicbolt Methods
- Bolt Interface Examples 1 & 2
- The Topology Interface
- About the TopologyBuilder Methods
- What are BoltDeclarer Methods?
- What are StormSubmitter Methods?
- Examples of Topology Builder
- Some Facts About Apache Kafka
- The Kafka Data Mode
- Facts About Apache Cassandra
- The Real Time Data Analysis Platform
- Kafka Interface to Storm
- About Kafka Spout
- The Kafka Spout Configuration
- About Kafka Spout Schemes
- Using Kafka Spout in Storm
- Storm Interface to Cassandra
- How to Insert or Update Cassandra
- How to Set Up Cassandra Session
- How to Insert or Update Data into Cassandra from Bolt?
- The Kafka Storm Cassandra

6 Storm Trident

- Storm Trident: An Overview
- Introduction to Trident
- About the Trident Data Model
- The Stateful Processing uses the Trident
- What are the Operations in Trident?
- About Trident State
- About the Trident Topology
- The Trident Spouts
- Compiling and running the Trident spout (Demonstration)
Big Data Hadoop Architect Online Training

- The Fault-tolerance Levels
- Pipelining
- Exactly Once processing
- The Spout Definition Example
- The Trident Operation Example
- About the Storing the Output Example
- Topology Connecting Spout and Bolt
- Topology Main Function
- About the Wrapper class
- The Trident Advantages