skillsoft global knowledge™

HADOOP DEVELOPER FOUNDATIONS

Course Code: 100682

Learn about the Hadoop ecosystem and how to process large data streams.

Apache Hadoop is a framework for processing Big Data, and Spark is a new in-memory processing engine. This course will introduce you to the Hadoop ecosystem and Spark.

This course explores processing large data streams in the Hadoop ecosystem. Working in a hands-on learning environment, you'll learn techniques and tools for ingesting, transforming, and exporting data to and from the Hadoop ecosystem for processing. You'll also process data using Map/Reduce and other critical tools, including Hive and Pig. Towards the end of the course, we'll review other useful tools such as Oozie and discuss security in the ecosystem.

What You'll Learn

Join an engaging hands-on learning environment, where you'll explore:

- Introduction to Hadoop
- HDFS
- YARN
- Data Ingestion
- HBase
- Oozie
- Working with Hive
- Hive advanced
- Hive in Cloudera/Hortonworks Distribution (or tools of choice)
- Working with Spark
- Spark Basics
- Spark Shell
- RDDs
- Spark Dataframes and Datasets
- Spark SQL
- Spark API programming
- Spark and Hadoop
- Machine Learning (ML/MLlib)
- GraphX

- Spark Streaming

This course has a 50% hands-on labs to 50% lecture ratio with engaging instruction, demos, group discussions, labs, and project work.

Who Needs to Attend

Experienced Developers and Architects seeking to be proficient in Hadoop, Hive, and Spark within an enterprise data environment.

Prerequisites

Before attending this course, you should be:

- Familiar with a programming language
- Comfortable in Linux environment (be able to navigate Linux command line, edit files using vi or nano)

HADOOP DEVELOPER FOUNDATIONS

Course Code: 100682

VIRTUAL CLASSROOM LIVE	$3,295 CAD	4 Day

Virtual Classroom Live Outline

Introduction to Hadoop

- Hadoop history, concepts
- Ecosystem
- Distributions
- High-level architecture
- Hadoop myths
- Hadoop challenges
- Hardware and software

HDFS

- Design and architecture
- Concepts (horizontal scaling, replication, data locality, and rack awareness)
- Daemons: Namenode, Secondary Namenode, and Datanode
- Communications and heart-beats
- Data integrity
- Read and write path
- Namenode High Availability (HA), Federation

YARN

- YARN Concepts and architecture
- Evolution from MapReduce to YARN
- Data Ingestion
- Flume for logs and other data ingestion into HDFS
- Sqoop for importing from SQL databases to HDFS, as well as exporting back to SQL
- Copying data between clusters (distcp)
- Using S3 as complementary to HDFS
- Data ingestion best practices and architectures
- Oozie for scheduling events on Hadoop

HBase

- Concepts and architecture
- HBase vs RDBMS vs Cassandra
- HBase Java API
- Time series data on HBase
- Schema design

Oozie

- Introduction to Oozie
- Features of Oozie
- Oozie Workflow
- Creating a MapReduce Workflow
- Start, End, and Error Nodes
- Parallel Fork and Join Nodes
- Workflow Jobs Lifecycle
- Workflow Notifications
- Workflow Manager
- Creating and Running a Workflow
- Oozie Coordinator Sub-groups
- Oozie Coordinator Components, Variables, and Parameters

Working with Hive

- Architecture and design
- Data types
- SQL support in Hive
- Creating Hive tables and querying
- Partitions
- Joins
- Text processing
- Labs: various labs on processing data with Hive

Hive advanced

- Transformation and Aggregation
- Working with Dates, Timestamps, and Arrays
- Converting Strings to Date, Time, and Numbers
- Create new Attributes, Mathematical Calculations, and Windowing Functions
- Use Character and String Functions
- Binning and Smoothing
- Processing JSON Data
- Execution Engines (Tez, MR, Spark)

Hive in Cloudera or HortonWorks Distribution (or tools of choice)

- Impala architecture
- Impala joins and other SQL specifics

Spark Basics

- Big Data, Hadoop, and Spark
- What’s new in Spark v2

- Spark concepts and architecture
- Spark ecosystem (core, spark sql, mlib, and streaming)

Spark Shell

- Spark web UIs
- Analyzing dataset

RDDs

- RDDs concepts
- RDD Operations/transformations
- Labs: Unstructured data analytics using RDDs
- Data model concepts
- Partitions
- Distributed processing
- Failure handling
- Caching and persistence
- **Spark Dataframes and Datasets**
- Intro to Dataframe/Dataset
- Programming in Dataframe/Dataset API
- Loading structured data using Dataframes

Spark SQL

- Spark SQL concepts and overview
- Defining tables and importing datasets
- Querying data using SQL
- Handling various storage formats: JSON/Parquet/ORC

Spark API programming (Scala and Python)

- Introduction to Spark API
- Submitting the first program to Spark
- Debugging/logging
- Configuration properties

Spark and Hadoop

- Hadoop Primer: HDFS/YARN
- Hadoop + Spark architecture
- Running Spark on YARN
- Processing HDFS files using Spark
- Spark and Hive

Machine Learning (ML/MLlib)

- Machine Learning primer
- Machine Learning in Spark: MLlib/ML
- Spark ML overview (newer Spark2 version)
- Algorithms: Clustering, Classifications, and Recommendations

GraphX

- GraphX library overview
- GraphX APIs

Spark Streaming

- Streaming concepts
- Evaluating Streaming platforms
- Spark streaming library overview
- Streaming operations
- Sliding window operations
- Structured Streaming
- Continuous streaming
- Spark and Kafka streaming

skillsoft global knowledge™

HADOOP DEVELOPER FOUNDATIONS

Course Code: 100682

PRIVATE GROUP TRAINING	4 Day

Visit us at www.globalknowledge.com or call us at 1-866-716-6688.

Date created: 7/30/2025 6:20:30 PM