
venolearn
EXPERIENCE BEYOND IT TRAINING

Big Data Analytics with Spark and Scala

Managing and processing big data is now a core task for a lot of companies,not restricted to the IT sector alone. With popular tools and technology such as MapReduce and Hadoop, analysing and working with big data has never been easier. Recently, Apache Spark has been added to this list and is soon becoming quite popular among the Data Scientists.

Apache Spark is an in-memory distributed framework written in the Scala programming language. The course will cover Spark's programming model complete with practical examples in Spark and Scala.

What uses Spark for Big Data Analytics?

- Advanced Analytics with Spark is a breeze as it provides tools for accelerated queries, a machine learning library, a graph processing and a streamlining analytics engine.
- All major Hadoop distributions are compatible with Spark.
- Spark has seen impressive adoption, thanks to its features that are readily helping the businesses meet their Big Data requirements.

Why must you learn Spark and Scala?

- Spark can be used with Java, Python, and Scala, which makes it a highly versatile and compatible tool.

- Spark and Scala programming are among the top 10 IT skills that are in demand.
- Apache Spark, when used with Scala, results in non-complex programs making it a beginner-friendly tool.
- Learning Apache Spark increases your job prospects considerably.
- Spark Developers take home salaries that are among the highest in the IT industry.

Course Objectives:

After completing the Apache Spark training, you will be able to:

1) Understand Scala and its implementation

2)Master the concepts of traits and OOPS in Scala

3) Install Spark and implement Spark operations on Spark Shell

4) Understand the role of Spark RDD

5) Implement Spark applications on YARN (Hadoop)

6) Learn Spark Streaming API

7) Implement machine learning algorithms in Spark MLlib API

8) Analyze Hive and Spark SQL architecture

9)Understand Spark GraphX API and implement graph algorithm

10) Implement Broadcast variable and Accumulators for performance

11)Project

What are the Pre-requisites for this course?

A basic understanding of functional programming and object oriented programming will help. Knowledge of Java/J2EE will definitely be a plus.

Curriculum:

1)Introduction to scala for Apache Spark

Learning Objectives: -

In this module, you will understand the basics of Scala that are required for programming Spark applications. You can learn about the basic constructs of Scala such as variable types, control structures, collections, and more.

Topics

What is Scala? Why Scala for Spark?Scala in other frameworks, introduction to Scala REPL, basic Scala operations, Variable Types in Scala, Control Structures in Scala, Foreach loop, Functions, Procedures, Collections in Scala- Array, ArrayBuffer, Map, Tuples, Lists, and more.

2)OOPS and Functional Programming in Scala

Learning Objectives –

In this module, you will learn about object oriented programming and functional programming techniques in Scala.

Topics:

Classes in Scala and Setters, Custom Getters and Setters, Properties with only Getters, Auxiliary Constructor, Primary Constructor, Singletons, Companion Objects, Extending a Class, Overriding Methods, Traits as Interfaces, Layered Traits, Functional Programming, Higher Order Functions, Anonymous GettersFunctions.

3)Introduction to Big Data and Apache Spark

Learning Objectives

In this module, you will understand what is big data, challenges associated with it and the different frameworks available. The module also includes a first-hand introduction to Spark.

Topics:

Introduction to big data, challenges with big data, Batch Vs. Real Time big data analytics, Batch Analytics - Hadoop Ecosystem Overview, Real-time Analytics Options, Streaming Data - Spark, In-memory data - Spark, What is Spark?, Spark Ecosystem, modes of Spark, Spark installation demo, an overview of Spark on a cluster, Spark Standalone cluster, Spark Web UI.

4)Spark Common Operations

Learning Objectives

In this module, you will learn how to invoke Spark Shell and use it for various common operations.

Topics

Invoking Spark Shell, creating the Spark Context, loading a file in Shell, performing basic Operations on files in Spark Shell, Overview of SBT, building a Spark project with SBT, running Spark project with SBT, local mode, Spark mode, caching overview, Distributed Persistence.

5)Playing with RDDs

Learning Objectives

In this module, you will learn one of the fundamental building blocks of Spark - RDDs and related manipulations for implementing business logics.

Topics

RDDs, transformations in RDD, actions in RDD, loading data in RDD, saving data through RDD, Key-Value Pair RDD, MapReduce and Pair RDD Operations, Spark and Hadoop Integration-HDFS, Spark and Hadoop Integration-Yarn, Handling Sequence Files, Partitioner.

6)Spark Straming and MLlib

Learning Objectives

In this module, you will learn about the major APIs that Spark offers. You will get an opportunity to work on Spark streaming which makes it easy to build scalable fault-tolerant streaming applications, MLlib which is Spark's machine learning library.

Topics

Spark Streaming Architecture, first Spark Streaming Program,transformations in Spark Streaming, fault tolerance in Spark Streaming, checkpointing, parallelism level, machine learning with Spark, data types, algorithms statistics, classification and regression, clustering, collaborative filtering.

7)GraphX, SparkSQL and Performance Tuning in Spark

Learning Objectives

In this module, you will learn about Spark SQL that is used to process structured data with SQL queries, graph analysis with Spark, GraphX for graphs and graph-parallel computation. You will also get a chance to learn the various ways to optimize performance in Spark.

Topics

Analyze Hive and Spark SQL architecture, SQL Context in Spark SQL, working with Data Frames, implementing an example for Spark SQL, integrating hive and Spark SQL, support for JSON and Parquet File Formats, implement data visualization in Spark, loading of data, Hive queries through Spark, testing tips in Scala, performance tuning tips in Spark, shared variables: Broadcast Variables, Shared Variables: Accumulators.