# Scalable programming with Scala and Spark

**Modality: Self-Paced Learning**

**Duration: 9 Hours**

## About the course:

If you are a data scientist or an analyst, you're accustomed to having various frameworks for working with information. Python, SQL, Java, R, and so forth. With Spark, you have a solitary engine where you can find out and play with a lot of information, run algorithms of machine learning and afterward utilize a similar framework to productionize your code.

**Scala**: Scala is a universally useful programming language - like C++ or Java. Its accessibility of a REPL environment and the practical programming nature make it especially appropriate for a distributed computing system like Spark.

**Analytics**: Using Scala and Spark you can explore and analyze your information in an intelligent situation with quick feedback. The course will tell the best way to use the intensity of Dataframes and RDDs to control information easily.

**Machine Learning and Data Science**: Spark's built-in libraries and core functionality make it simple to actualize complex calculations like Recommendations with not too many lines of code. We'll cover an assortment of datasets and calculations including MapReduce, PageRank, and Graph datasets.

## Course Objective:

Scala Programming Constructs: Traits, Classes, Closures, First Class Functions, Case Classes Currying.

Lots of cool stuff.

- Utilizing Spark Streaming for stream preparing
- Spark SQL and Dataframes to work with Twitter information.
- Recommendations for Music utilizing the Audioscrobbler and Alternating Least Squares dataset
- Utilizing the PageRank calculation with the dataset of Google web graph.
- Working with graph information utilizing the dataset of Marvel Social system.

Spark basic and advanced features:

- Resilient Distributed Datasets, Actions (reduce, aggregate), Transformations (map, filter, flatMap)
- Pair RDDs, combineByKey, reduceByKey,
- Accumulator and Broadcast variables
- Spark for MapReduce.
- The Java API for Spark.

- Spark Streaming, Spark SQL, GraphX and MLlib.

## Audience:

- Specialists who need to utilize a distributed computing engine for stream processing or batch or both
- An analyst who needs to use Spark for breaking down fascinating datasets
- Data Scientists who need a solitary engine for modeling and analyzing information and productionizing it.

## Prerequisite:

All models work without or with Hadoop. If you might want to utilize Spark with Hadoop, you'll require to have Hadoop introduced (either in cluster mode or pseudo-distributed).

## Course Outline:

- You, This Course and Us
- Introduction to Spark
- Resilient Distributed Datasets
- Advanced RDDs: Pair Resilient Distributed Datasets
- Advanced Spark: Accumulators, Spark Submit, MapReduce , Behind The Scenes
- PageRank: Ranking Search Results
- Spark SQL
- MLlib in Spark: Build a recommendations engine
- Spark Streaming
- Graph Libraries
- Scala Language Primer
- Supplementary Installs