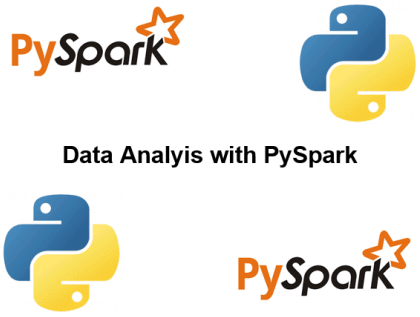
PySpark for Big Data

Duration: 2 days **Price: € 1499**

Open Schedule



Audience PySpark for Big Data

The course PySpark for Big Data is intended for developers and upcoming Data Analysts who want to learn how to use Apache Spark from Python.

Prerequisites training PySpark for Big Data

To participate in this course, some experience with programming is beneficial for understanding. Prior knowledge of Python or big data handling with Apache Spark is not required.

Realization course PySpark for Big Data

The theory is treated on the basis of presentations. Illustrative demos are used to clarify the concepts discussed. There is ample opportunity to practice and alternate theory and practice. The course times are from 9.30 am to 4.30 pm.

Certification course PySpark for Big Data

Participants receive an official certificate PySpark for Big Data after successful completion of the course.

Content Course PySpark for Big Data

In the course PySpark for Big Data participants learn to use Apache Spark from Python. Apache Spark is a Framework for parallel processing of big data. With PySpark, Apache Spark is integrated with the Python language.

Spark Architecture

The course PySpark for Big Data discusses the architecture of Spark, the Spark Cluster Manager and the difference between Batch and Stream Processing.

Hadoop

After a discussion of the Hadoop Distributed File System, parallel operations and working with RDDs, Resilient Distributed Datasets are discussed in the course PySpark for Big Data. The configuration of PySpark applications via SparkConf and SparkContext is also explained.

MapReduce en SQL

Extensive consideration is given to the possible operations on RDDs, including map and reduce. The use of SQL in Spark is also discussed. The GraphX library is discussed and DataFrames is discussed. Iterative algorithms are also treated.

Mlib library

Finally the course PySpark for Big Data pays attention to machine learning with the Mlib library.

Modules Course PySpark for Big Data

Module 1 : Python Primer	Module 2 : Spark Intro	Module 3 : HDFS
Python Syntax	What is Apache Spark?	Hadoop Environment
Python Data Types	Spark and Python	Environment Setup
List, Tuples, Dictionaries	PySpark	Hadoop Stack
Python Control Flow	Py4j Library	Hadoop Yarn
Functions and Parameters	Data Driven Documents	Hadoop Distributed File System
Modules and Packages	RDD's	HDFS Architecture
Comprehensions	Real Time Processing	Parallel Operations
Iterators and Generators	Apache Hadoop MapReduce	Working with Partitions
Python Classes	Cluster Manager	RDD Partitions
Anaconda Environment	Batch versus Stream Processing	HDFS Data Locality
Jupyter Notebooks	PySpark Shell	DAG (Direct Acyclic Graph)
Module 4 : SparkConf	**Module 5 : SparkContext**	**Module 6 : RDD's**
SparkConf Object	Main Entry Point	Resilient Distributed Datasets
Setting Configuration Properties	Executor	Key-Value pair RDDs
Uploading Files	Worker Nodes	Parallel Processing
SparkContext.addFile	LocalFS	Immutability and Fault Tolerance
Logging Configuration	SparkContext Parameters	Transformation Operations
Storage Levels	Master	Filter, groupBy and Map
Serialize RDD	RDD serializer	Action Operations
Replicate RDD partitions	batchSize	Caching and persistence
DISK_ONLY	Gateway	PySpark RDD Class
MEMORY_AND_DISK	JavaSparkContext instance	count, collect, foreach,filter
MEMORY_ONLY	Profiler	map, reduce, join, cache
Module 7 : Spark Processing	**Module 8 : Broadcast and Accumulator**	**Module 9 : Algorithms**
SQL support in Spark	Performance Tuning	Iterative Algorithms
Spark 2.0 Dataframes	Serialization	Graph Analysis
Defining tables	Network Traffic	Machine Learning API
Importing datasets	Disk Persistence	mllib.classification
Querying data frames using SQL	MarshalSerializer	Random Forest
Storage formats	Data Type Support	Naive Bayes
JSON / Parquet	Python's Pickle Serializer	Decision Tree
GraphX	DStreams	mllib.clustering
GraphX library overview	Sliding Window Operations	mllib.linalg
GraphX APIs	Multi Batch and State Operations	mllib.regression

SpiralTrain BV
Standerdmolen 10, 2e verdieping
3995 AA Houten

info@spiraltrain.nl
www.spiraltrain.nl
Tel.: +31 (0) 30 – 737 0661

Locations
Houten, Amsterdam, Rotterdam, Eindhoven, Zwolle, Online