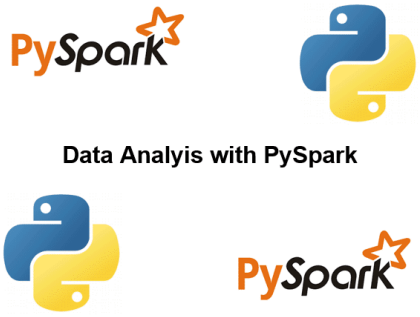
PySpark voor Big Data

Duur: 2 dagen Prijs: € 1499

Open Rooster



Doelgroep Cursus PySpark voor Big Data

De cursus PySpark voor Big Data is bedoeld voor developers en aankomende Data Analisten die Apache Spark willen leren gebruiken vanuit Python.

Voorkennis training PySpark voor Big Data

Om aan deze cursus deel te nemen is kennis enige ervaring met programmeren bevorderlijk voor de begripsvorming. Voorafgaande kennis van Python of big data handling met Apache Spark is niet nodig.

Uitvoering cursus PySpark voor Big Data

De theorie wordt behandeld aan de hand van presentaties. Illustratieve demo's worden gebruikt om de behandelde concepten te verduidelijken. Er is voldoende gelegenheid om te oefenen en afwisseling van theorie en praktijk. De cursustijden zijn van 9.30 tot 16.30.

Certificering cursus PySpark voor Big Data

De deelnemers krijgen na het goed doorlopen van de cursus een officieel certificaat PySpark voor Big Data.

Inhoud Cursus PySpark voor Big Data

In de cursus PySpark voor Big Data leren de deelnemers Apache Spark vanuit Python te gebruiken. Apache Spark is een Framework voor parallelle processing van big data. Met PySpark wordt Apache Spark geïntegreerd met de Python taal.

Spark Architectuur

In de cursus PySpark voor Big Data komt aan de orde komt de architectuur van Spark, de Spark Cluster Manager en het verschil tussen Batch en Stream Processing.

Hadoop

Na een bespreking van het Hadoop Distributed File System wordt ingegaan op parallelle operaties and het werken met RDD's, Resilient Distributed Datasets. De configuratie van PySpark applicaties via SparkConf en SparkContext komt eveneens aan bod in de cursus PySpark voor Big Data.

MapReduce en SQL

Uitgebreid wordt ingegaan op de mogelijke operaties op RDD's waaronder map en reduce. Ook komt het gebruik van SQL in Spark aan de orde. De GraphX library wordt besproken en er wordt ingegaan op DataFrames. Verder komen iteratieve algorithmen aan de orde.

Mlib library

Tenslotte wordt in de cursus PySpark voor Big Data aandacht besteed aan machine learning met de Mlib library.

SpiralTrain BV
Standerdmolen 10, 2e verdieping
3995 AA Houten

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

Locaties
Houten, Amsterdam, Rotterdam, Eindhoven, Zwolle, Online

Modules Cursus PySpark voor 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

Locaties
Houten, Amsterdam, Rotterdam, Eindhoven, Zwolle, Online