

# Spark Python, toepassingen ontwikkelen voor big data

Praktijkcursus van 3 dagen - 21h

Ref : QNC - Prijs 2025 : € 1 910 excl. BTW

Spark is een gedistribueerd rekenraamwerk voor het manipuleren van grote hoeveelheden gegevens. In eerste instantie ontworpen om Hadoop-verwerking te versnellen, is het nu een autonoom systeem geworden. Het kan worden geprogrammeerd in vier talen, waaronder Python, dat de dominante taal is geworden. In deze cursus maakt u kennis met Spark Python.

## PEDAGOGISCHE DOELSTELLINGEN

Na afloop van de opleiding kan de cursist:

- Ontdek de fundamentele concepten van Spark
- Het RDD-concept van Spark gebruiken
- Gegevens exploiteren met Spark SQL
- Real-time analyses uitvoeren met Spark Streaming
- Spark gebruiken met Jupyter notebooks, gegevens manipuleren met Pyspark zoals met Pandas
- Machinaal leren met Spark

## PEDAGOGISCHE METHODEN

Elk onderwerp wordt geïllustreerd met demonstraties op een cluster in de cloud. Deelnemers maken oefeningen nadat de concepten zijn gepresenteerd.

## EXERCISE

Er worden talloze oefeningen gebruikt om de onderwerpen te illustreren.

## HET PROGRAMMA

laatste update: 02/2025

### 1) Inleiding tot Apache Spark

- Geschiedenis van het kader.
- De vier hoofdcomponenten: Spark SQL, Spark Streaming, MLlib en GraphX.
- Python-tools en -bibliotheken voor Spark: PySpark, Jupyter-notebooks, Koala's.
- Spark programmeerconcepten.
- Spark draaien in een gedistribueerde omgeving.

*De Python-omgeving voor Spark opzetten. Scripts implementeren die Spark concepten manipuleren.*

### 2) Spark gebruiken met Python: veerkrachtige gedistribueerde datasets (RDD)

- Uw Python-omgeving configureren.
- Verbinding maken met Spark met Python: contexten en sessies.
- Inleiding tot RDD's. RDD's maken, manipuleren en hergebruiken.
- De belangrijkste functies/transformaties, implementatie van map/reduce algoritmen.
- Accumulatoren en uitzendvariabelen.
- Gebruik partities.
- Notebooks gebruiken en Python-taken indienen.

*Contexten en sessies manipuleren. Aanmaken en hergebruiken van RDDs. Taken indienen.*

### 3) Omgaan met gestructureerde gegevens

- Inleiding tot Spark SQL en DataFrames en datasets.
- De verschillende soorten/formaten gegevensbronnen.

## DEELNEMERS

Iedereen die bekend is met Python en meer wil leren over het Spark framework van de Apache Foundation.

## VOORAFGAANDE VEREISTEN

Goede kennis van de taal Python.

## VAARDIGHEDEN VAN DE CURSUSLEIDER

De deskundigen die de cursus leiden zijn specialisten op het betreffende vakgebied. Zij werden geselecteerd door onze pedagogische teams zowel om hun vakkennis als hun pedagogische vaardigheden voor elke cursus die zij geven. Zij hebben minstens vijf tot tien jaar ervaring in hun vakgebied en oefenen of oefenden verantwoordelijke bedrijfsfuncties uit.

## BEOORDELINGSMODALITEITEN

De cursusleider beoordeelt de pedagogische vooruitgang van de deelnemer gedurende de gehele cursus aan de hand van meerkeuzevragen, praktijksituaties, praktische opdrachten, ... De deelnemer legt ook van tevoren en naderhand een test af ter bevestiging van de verworven kennis.

## PEDAGOGISCHE EN TECHNISCHE MIDDELEN

- De gebruikte pedagogische middelen en cursusmethoden zijn voornamelijk: audiovisuele hulpmiddelen, documentatie en cursusmateriaal, praktische oefeningen en correcties van de oefeningen voor praktijkstages, casestudies of reële voorbeelden voor de seminars.
- Na afloop van de stages of seminars verstrekt ORSYS de deelnemers een evaluatievragenlijst over de cursus die vervolgens door onze pedagogische teams wordt geanalyseerd.
- Na afloop van de cursus wordt een presentielijst per halve dag verstrekt, evenals een verklaring van de afronding van de cursus indien de stagiair alle sessies heeft bijgewoond.

## TOEGANGSMODALITEITEN EN -TERMIJNEN

De inschrijving dient 24 uur voor aanvang van de cursus plaatsgevonden te hebben.

## TOEGANKELIJKHEID VOOR MINDERVALIDEN

Is voor u speciale toegankelijkheid vereist? Neem contact op met mevr. FOSSE, contactpersoon voor mindervaliden, via het adres [psh-accueil@ORSYS.fr](mailto:psh-accueil@ORSYS.fr) om uw verzoek en de haalbaarheid daarvan zo goed mogelijk te bestuderen.

- Interoperabiliteit met RDD's.
  - Gebruik de PySpark Pandalas-bibliotheek.
- Uitvoeren van queries met Spark SQL. Implementeren van DataFrames en datasets. DataFrame manipulatie.*

#### 4) Machinaal leren met Spark

- Inleiding tot machinaal leren.
  - De verschillende klassen van algoritmen.
  - MLlib introduceren.
  - Implementatie van de verschillende algoritmen in MLlib.
- Implementatie van leren onder toezicht door classificatie.*

#### 5) Real-time analyse met Spark Streaming

- De architectuur van streaming begrijpen.
- Presentatie van gediscrètiseerde streams (DStreams).
- De verschillende soorten bronnen.
- API-manipulatie (aggregatie, watermerken, enz.).
- Real-time machinaal leren.

*Real-time statistieken maken van een gegevensbron en voorspellingen doen met behulp van machine learning.*

#### 6) Grafentheorie

- Inleiding tot grafentheorie (knopen, randen, gerichte grafieken, paden, belangrijkste algoritmen).
- De API gebruiken.
- Presentatie van de GraphX en GraphFrame bibliotheken.

*Implementatie van een algoritme voor paginarangschikking en weergave van de grafiek.*

## DATA

---

Neem contact met ons op