Spark Python, desarrollo de aplicaciones para big data

Curso práctico de 3 días - 21h
Ref.: QNC - Precio 2025: 1 330€ sin IVA

Spark es un marco informático distribuido para manipular grandes cantidades de datos. Inicialmente diseñado para acelerar el procesamiento de Hadoop, se ha convertido en un sistema independiente. Se puede programar en cuatro lenguajes, incluido Python, que se ha convertido en el lenguaje dominante. Este curso le introduce a Spark Python.

OBJETIVOS PEDAGÓGICOS

Al término de la formación, el alumno podrá:

Descubra los conceptos fundamentales de Spark

Uso del concepto RDD de Spark

Explotación de datos con Spark SQL

Realice análisis en tiempo real con Spark Streaming

Uso de Spark con Jupyter notebooks, manipulación de datos con Pyspark como con Pandas.

Aprendizaje automático con Spark

MÉTODOS PEDAGÓGICOS

Cada tema se ilustra mediante demostraciones en un clúster en la nube. Los participantes completan ejercicios tras la presentación de los conceptos.

Los temas se ilustran con numerosos ejercicios.

PARTICIPANTES

Cualquier persona familiarizada con Python que quiera aprender más sobre el framework Spark de la Fundación Apache.

REQUISITOS PREVIOS

Buen conocimiento del lenguaje Python.

COMPETENCIAS DEL FORMADOR

Los expertos que imparten la formación son especialistas en las materias tratadas. Han sido validados por nuestros equipos pedagógicos, tanto en el plano de los conocimientos profesionales como en el de la pedagogía, para cada curso que imparten. Cuentan al menos con entre cinco y diez años de experiencia en su área y ocupan o han ocupado puestos de responsabilidad en empresas.

MODALIDADES DE EVALUACIÓN

El formador evalúa los progresos pedagógicos del participante a lo largo de toda la formación mediante preguntas de opción múltiple, escenificaciones de situaciones, trabajos prácticos, etc.
El participante también completará una prueba de posicionamiento previo y posterior para validar las competencias adquiridas.

MEDIOS PEDAGÓGICOS Y TÉCNICOS

• Los medios pedagógicos y los métodos de enseñanza utilizados son principalmente: ayudas audiovisuales, documentación y soporte de cursos, ejercicios prácticos de aplicación y ejercicios corregidos para los cursillos prácticos, estudios de casos o presentación de casos reales para los seminarios de formación.
• Al final de cada cursillo o seminario, ORSYS facilita a los participantes un cuestionario de evaluación del curso que analizarán luego nuestros equipos pedagógicos.
• Al final de la formación se entrega una hoja de presencia por cada media jornada de presencia, así como un certificado de fin de formación si el alumno ha asistido a la totalidad de la sesión.

MODALIDADES Y PLAZOS DE ACCESO

La inscripción debe estar finalizada 24 horas antes del inicio de la formación.

ACCESIBILIDAD DE LAS PERSONAS CON DISCAPACIDAD

¿Tiene alguna necesidad específica de accesibilidad? Póngase en contacto con la Sra. FOSSE, interlocutora sobre discapacidad, en la siguiente dirección psh-accueil@orsys.fr para estudiar de la mejor forma posible su solicitud y su viabilidad.

PROGRAMA

última actualización: 08/2024

1) Introducción a Apache Spark

- Historia del marco.
- Los cuatro componentes principales: Spark SQL, Spark Streaming, MLlib y GraphX.
- Herramientas y bibliotecas de Python para Spark: PySpark, Jupyter notebooks, Koalas.
- Conceptos de programación Spark.
- Ejecución de Spark en un entorno distribuido.

Trabajo práctico : Configuración del entorno Python para Spark. Implementación de scripts que manipulan conceptos de Spark.

2) Uso de Spark con Python: conjuntos de datos distribuidos resilientes (RDD)

- Configuración del entorno Python.
- Conexión a Spark con Python: contextos y sesiones.
- Introducción a los RDDs. Creación, manipulación y reutilización de RDDs.
- Las principales funciones/transformaciones, implementación de algoritmos map/reduce.
- Acumuladores y variables de emisión.
- Utiliza particiones.
- Uso de cuadernos y envío de trabajos de Python.

Trabajo práctico : Manipulación de contextos y sesiones. Creación y reutilización de RDDs. Envío de trabajos.

3) Tratamiento de datos estructurados

- Introducción a Spark SQL y DataFrames y conjuntos de datos.
- Los distintos tipos/formatos de fuentes de datos.

- Interoperabilidad con los RDD.
- Utilice la biblioteca PySpark Pandas.

Ejecución de consultas con Spark SQL. Implementación de DataFrames y conjuntos de datos. Manipulación de DataFrame.

4) Aprendizaje automático con Spark

- Introducción al aprendizaje automático.
- Las diferentes clases de algoritmos.
- Presentación de MLlib.
- Implementación de los distintos algoritmos en MLlib.

Trabajo práctico : Aplicación del aprendizaje supervisado mediante clasificación.

5) Análisis en tiempo real con Spark Streaming

- Comprensión de la arquitectura de streaming.
- Presentación de flujos discretizados (DStreams).
- Los distintos tipos de fuente.
- Manipulación de API (agregación, marca de agua, etc.).
- Aprendizaje automático en tiempo real.

Trabajo práctico : Creación de estadísticas en tiempo real a partir de una fuente de datos y predicciones mediante aprendizaje automático.

6) Teoría de grafos

- Introducción a la teoría de grafos (nodos, aristas, grafos dirigidos, caminos, principales algoritmos).
- Uso de la API.
- Presentación de las bibliotecas GraphX y GraphFrame.

Trabajo práctico : Implementación de un algoritmo de page rank y visualización del gráfico.

FECHAS

Contacto