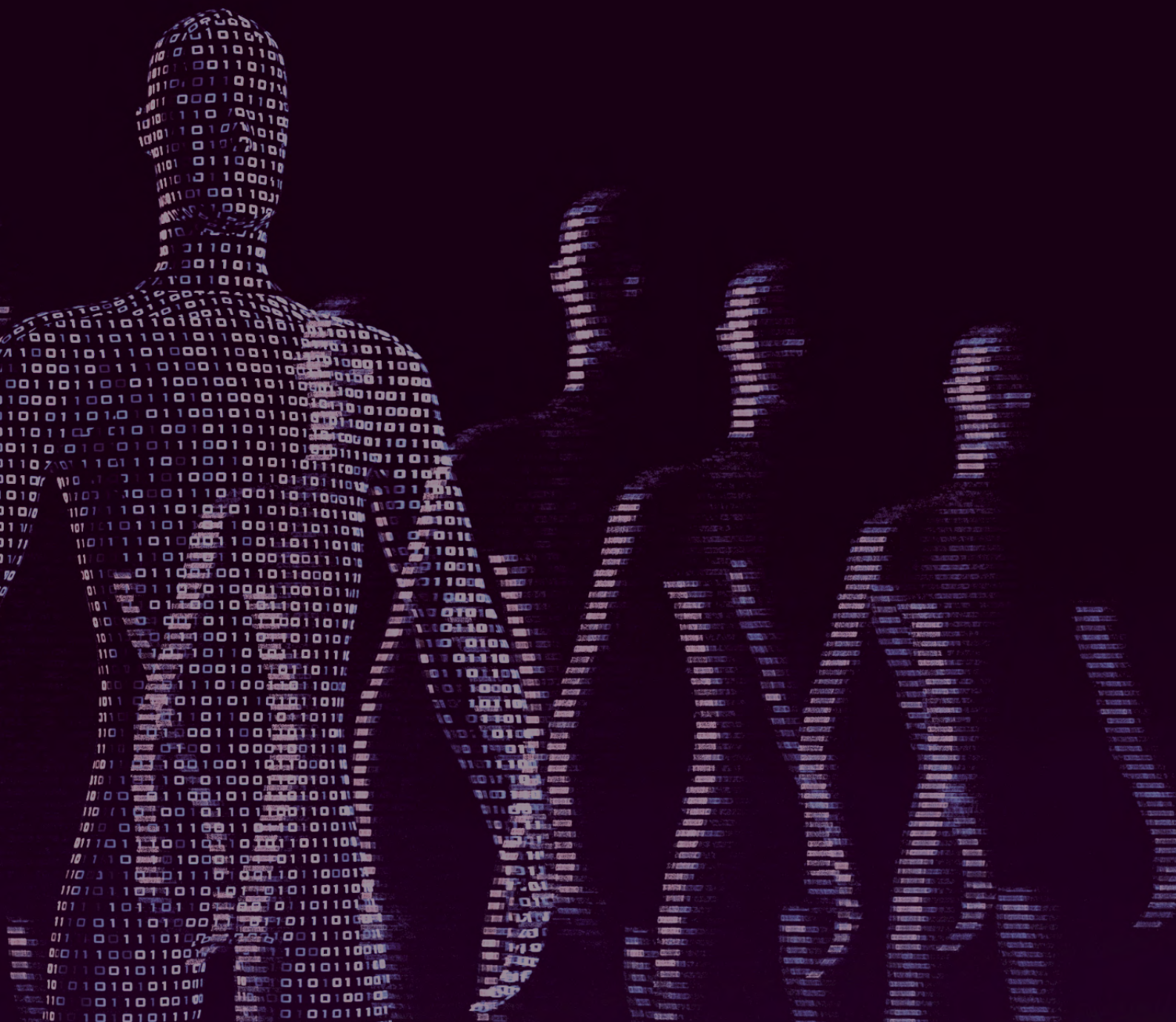
pue | talent

1ª Edición Becas PUE Talent

Curso oficial Cloudera Data Engineering: Developing Applications with Apache Spark



PUE Talent es una iniciativa diseñada para el reclutamiento de talento que quiera incorporarse a la empresa de referencia internacional en Big Data, reconocida recientemente como Cloudera Platinum Partner. Enmarcada en PUE Talent, te ofrecemos esta beca de formación y certificación oficial valorada en 2.000€ con la que podrás realizar de forma gratuita el curso y certificación Cloudera Data Engineering: Developing Applications with Apache Spark

Las plazas son limitadas y se asignarán tras un proceso de selección en donde se tendrán en cuenta los requisitos de acceso que encontrarás en la descripción de las becas.

Este curso forma parte de la 1ª edición de las Becas PUE Talent. Una iniciativa diseñada para el reclutamiento de talento que quiera incorporarse a la empresa de referencia internacional en Big Data, reconocida recientemente como Cloudera Platinum Partner.

REQUISITOS

1. Experiencia en desarrollo de software.
2. Conocimiento de Java, Python, Scala o Spark.
3. Nivel de inglés (mínimo B2).

ACERCA DE ESTE CURSO

Este curso te proporcionará los conceptos clave y el conocimiento necesario para desarrollar aplicaciones paralelas de alto rendimiento en Cloudera Data Platform (CDP) con Apache Spark .

Practicarás la escritura de aplicaciones Spark que se integran con los componentes principales de CDP, como Hive y Kafka, aprenderás a usar Spark SQL para consultar datos estructurados, Spark Streaming para procesar datos de transmisión en tiempo real y a trabajar con "big data" almacenado en un sistema de archivos distribuido.

En definitiva, serás capaz de crear aplicaciones para tomar mejores decisiones y más rápidas, además de saber ejecutar un análisis interactivo aplicado a una amplia variedad de casos de uso, arquitecturas e industrias.

OBJETIVOS DEL CURSO

1. Distribuir, almacenar y procesar datos en un clúster CDP
2. Escribir, configurar e implementar aplicaciones Apache Spark
3. Usar los intérpretes de Spark y las aplicaciones de Spark para explorar, procesar y analizar datos distribuidos
4. Consultar datos con tablas Spark SQL, DataFrames y Hive
5. Usar Spark Streaming junto con Kafka para procesar un flujo de datos

CONTENIDOS

Módulo 1: Introducción a Zeppelin
- ¿Por qué Notebooks?
- Notas de Zeppelin
- Demo: Apache Spark en 5 minutos

Módulo 2: Introducción a HDFS
- Descripción general de HDFS
- Componentes e interacciones de HDFS
- Interacciones HDFS adicionales
- Descripción general de Ozone
- Ejercicio: Trabajar con HDFS

Módulo 3: Introducción a YARN
- Descripción general de YARN
- Componentes e interacción de YARN
- Trabajar con YARN
- Ejercicio: Trabajar con YARN

Módulo 4: Historial de procesamiento distribuido
- Los Años del Disco: 2000 -> 2010
- Los Años de la Memoria: 2010 -> 2020
- Los años de la GPU: 2020 ->

Módulo 5: Trabajar con RDDs
- Conjuntos de datos distribuidos resilientes (RDDs)
- Ejercicio: Trabajar con RDDs

Módulo 6: Trabajar con DataFrames
- Introducción a DataFrames
- Ejercicio: Introducción a DataFrames
- Ejercicio: Lectura y escritura de DataFrames
- Ejercicio: Trabajar con Columns
- Ejercicio: Trabajando con Complex Types
- Ejercicio: Combinar y dividir DataFrames
- Ejercicio: Resumir y agrupar DataFrames
- Ejercicio: Trabajar con UDFs
- Ejercicio: Trabajar con Windows

Módulo 7: Introducción a Apache Hive
- Acerca de Apache Hive

Módulo 8: Integración de Hive y Spark
- Integración de Hive y Spark
- Ejercicio: Integración de Spark con Hive

Módulo 9: Visualización de datos con Zeppelin
- Introducción a la visualización de datos con Zeppelin
- Análisis de Zeppelin
- Colaboración Zeppelin
- Ejercicio: AdventureWorks

Módulo 10: Desafíos del procesamiento distribuido
- Shuffle
- Skrew
- Order

Módulo 11: Procesamiento distribuido Spark
- Procesamiento distribuido Spark
- Ejercicio: Explorar el orden de ejecución de consultas

Módulo 12: Persistencia distribuida de Spark
- Persistencia de DataFrame y Dataset
- Niveles de almacenamiento de persistencia
- Visualización de RDDs persistentes
- Ejercicio: Dataframes persistentes

Módulo 13: Escribir, configurar y ejecutar aplicaciones Spark
- Escribir una aplicación Spark
- Creación y ejecución de una aplicación
- Modo de despliegue de aplicaciones
- La interfaz de usuario (UI) web de la aplicación Spark
- Configuración de las propiedades de la aplicación
- Ejercicio: Escribir, configurar y ejecutar una aplicación Spark

Módulo 14: Introducción a Structured Streaming
- Introducción a Structured Streaming
- Ejercicio: Procesamiento de datos en Streaming

Módulo 15: Procesamiento de mensajes con Apache Kafka
- ¿Qué es Apache Kafka?
- Descripción general de Apache Kafka
- Escalado de Apache Kafka
- Arquitectura de un clúster de Apache Kafka
- Herramientas de líneas de comandos de Apache Kafka

Módulo 16: Structured Streaming con Apache Kafka
- Recibir mensajes de Kafka
- Envío de mensajes Kafka
- Ejercicio: Trabajar con mensajes Streaming de Kafka

Módulo 17: Agregar y unir Streaming DataFrames
- Agregar Streaming
- Unir Streaming DataFrames
- Ejercicio: Agregar y unir Streaming DataFrames

Apéndice: Trabajar con conjuntos de datos en Scala
- Trabajar con conjuntos de datos en Scala
- Ejercicio: Uso de conjuntos de datos en Scala

Plazas limitadas. Si te interesa, **inscríbete ya** para que podamos evaluar tu candidatura.

ⓘ ¿TIENES DUDAS?

Para cualquier consulta no dudes en contactarnos a través del siguiente correo electrónico: **puetalent@pue.es**

pue | talent