Cloudera Developer Training for SPARK and HADOOP

Duración

Días: 5 Días
Horas: 28 horas

Descripción

Este curso práctico de cuatro días ofrece los conceptos clave y la experiencia que necesitan los desarrolladores para desarrollar aplicaciones paralelas de alto rendimiento con Apache Spark 2.
Con esta actualización del curso, simplificamos la agenda para ayudarles a ser productivos rápidamente con las tecnologías más importantes, incluido Spark 2.

Una experiencia práctica

Los ejercicios prácticos tienen lugar en un clúster en vivo, que se ejecuta en la nube. Se construirá un clúster privado para que cada estudiante use durante la clase.
A través de una discusión guiada por un instructor y ejercicios interactivos y prácticos, los participantes navegarán el ecosistema Hadoop, aprendiendo a :

- Distribuir, almacenar y procesar datos en un clúster de Hadoop
- Escribir, configurar e implementar aplicaciones Spark en un clúster
- Utilizar el shell Spark para el análisis interactivo de datos
- Procesar y consultar datos estructurados utilizando Spark SQL
- Usar Spark Streaming para procesar una secuencia de datos en vivo

A quién se dirige

Este curso está orientado a desarrolladores e ingenieros de software con experiencia en programación.

Objetivos

Los participantes aprenderán cómo utilizar Spark SQL para hacer consultas de datos estructurados y Spark Streaming para realizar procesamiento en tiempo real sobre datos en transmisión desde una variedad de fuentes. También practicarán la escritura de aplicaciones que usan core Spark para realizar el procesamiento de ETL y algoritmos iterativos. El curso cubre cómo trabajar con grandes conjuntos de agrupaciones de datos almacenados en un sistema de archivos distribuido y ejecutar aplicaciones Spark en un cluster Hadoop. Después de tomar este curso, los participantes estarán preparados para enfrentar los desafíos del mundo real y construir aplicaciones para ejecutar decisiones más rápidas y mejores y análisis interactivos, aplicables a una amplia variedad de casos de uso, arquitecturas e industrias.

Requisitos

Este curso está diseñado para desarrolladores e ingenieros que tienen experiencia en programación, pero no se requiere conocimiento previo de Hadoop y/o Spark.

- Los ejemplos de Apache Spark y los ejercicios prácticos se presentan en Scala y Python. Se requiere la capacidad de programar en uno de esos idiomas.
- Se asume una familiaridad básica con la línea de comandos de Linux
- Conocimiento básico de SQL es útil

Certificación

Al finalizar el curso, se anima a los asistentes a continuar su estudio y registrarse en el examen CCA Spark y Hadoop Developer.

La certificación es un gran diferenciador ya que le ayuda profesionalmente a distinguirse en la materia, proporcionando a las empresas y clientes una evidencia tangible de su conocimiento, habilidades y experiencia.

Contenido

1. Introducción a Apache Hadoop y el Ecosistema de Hadoop
    - Descripción general de Apache Hadoop
    - Ingestión y almacenamiento de datos
    - Procesamiento de datos
    - Análisis de datos y exploración
    - Otras herramientas del ecosistema
    - Introducción a los ejercicios prácticos

2. Almacenamiento de archivos Apache Hadoop
    - Componentes del clúster Apache Hadoop
    - Arquitectura HDFS
    - Usando HDFS

3. Procesamiento distribuido en un clúster Apache Hadoop
    - Arquitectura YARN
    - Trabajando con YARN

4. Conceptos básicos de Apache Spark
    - ¿Qué es Apache Spark?
    - Arranque del Spark Shell
    - Uso del Spark Shell
    - Primeros pasos con DataSets y DataFrames
    - Operaciones con DataFrames

5. Trabajando con DataFrames y Esquemas
    - Crear DataFrames a partir de orígenes de datos
    - Guardar DataFrames en orígenes de datos
    - Esquemas de DataFrames
    - Ejecución apremiante y demorada (eager y lazy)

6. Analizar datos con consultas de DataFrame
    - Consultar DataFrames usando expresiones de columna
    - Consultas de agrupación y agrecación
    - Uniendo DataFrames

7. Descripción general de RDD
    - Visión general del RDD
    - Fuentes de datos RDD
    - Crear y guardar RDD
    - Operaciones con RDD

8. Transformando datos con RDD
    - Escribir y paso de funciones de transformación
    - Ejecución de transformación
    - Convertir entre RDD y DataFrames

9. Agregando datos con RDDs de pares
    - RDD de pares clave-valor (Pair RDD)
    - Map Reduce
    - Otras operaciones con Pair RDD

10. Consulta de tablas y vistas con Apache Spark SQL
    - Consulta de tablas en Spark con SQL
    - Consulta de archivos y vistas
    - La API de catálogo
    - Comparación de Spark SQL, Apache Impala y Apache Hive-on-Spark

11. Trabajando con conjuntos de datos en Scala
    - DataSets y DataFrames
    - Crear DataSets
    - Cargando y guardando DataSets
    - Operaciones del DataSets

12. Escribir, configurar y ejecutar Apache Spark Applications
    - Escribir una aplicación Spark
    - Crear y ejecutar una aplicación

- Modo de despliegue de aplicaciones
- La interfaz de usuario web de la aplicación Spark
- Configuración de propiedades de la aplicación

13. Procesamiento distribuido
    - Revisión: Apache Spark en un clúster
    - Particiones RDD
    - Ejemplo: particionamiento en consultas
    - Etapas y tareas: Planificación de ejecución de trabajo
    - Ejemplo: Plan de ejecución del Catalyst
    - Ejemplo: plan de ejecución RDD

14. Persistencia de datos distribuidos
    - Persistencia de DataFrame y Dataset
    - Niveles de almacenamiento de persistencia (Storage Level)
    - Visualización de RDD persistentes

15. Patrones comunes en Apache Spark Data Processing
    - Casos de uso de Apache Spark comunes
    - Algoritmos iterativos en Apache Spark
    - Aprendizaje automático (Machine Learning)
    - Ejemplo: k-means

16. Apache Spark Streaming: Introducción a DStreams
    - Descripción de Apache Spark Streaming
    - Ejemplo: recuento de solicitudes de transmisión
    - DStreams
    - Desarrollo de aplicaciones de Streaming

17. Apache Spark Streaming: procesamiento de lotes múltiples
    - Operaciones de lotes múltiples
    - Time-Slicing
    - Operaciones con Estado
    - Operaciones de desplazamiento de ventanas
    - Vista previa: Streaming estructurado

18. Apache Spark Streaming: fuentes de datos
    - Visión general de la fuente de transmisión de datos
    - Fuentes de datos Apache Flume y Apache Kafka
    - Ejemplo: utilizando un origen de datos directo de Kafka