Desarrollador Apache Hadoop

Duración: 88 h

Módulo nº 1

PROGRAMACIÓN JAVA SE 7 (30 h)

Objetivo:

Realización de múltiples operaciones sobre tablas de bases de datos, incluyendo la creación, lectura, actualización y borrado de la tecnología JDBC. Procesar cadenas usando expresiones regulares. Crear aplicaciones multi-hilo de alto rendimiento que evitan deadlocks. Internacionalizar aplicaciones Java. Crear aplicaciones que usen el framework de Java Collections. Implementar la funcionalidad de entrada/salida (E/S) para leer y escribir a ficheros de datos y texto y comprender streams de E/S avanzados. Manipular ficheros, directorios y sistemas de ficheros usando la especificación del JDK7 NIO.2. Aplicar patrones de diseño comunes y mejores prácticas. Crear aplicaciones de tecnología Java que hagan un uso correcto de las características de orientación a objetos del lenguaje Java, como la encapsulación, herencia y polimorfismo. Ejecutar una aplicación Java desde la línea de comandos.

Contenidos teórico-prácticos:

- Introducción a la Plataforma Java
- Sintaxis Java y Revisión de Clases
- Encapsulación y Polimorfismo
- Diseño de Clases en Java
- Diseño Avanzado de Clases
- Herencia con Interfaces Java
- Genéricos y Colecciones
- Procesamiento de Cadenas
- Excepciones y Assertions
- Fundamentos de E/S
- E/S con NIO 2
- Hilos
- Concurrencia
- Aplicación de Base de Datos con JDBC
- Localización

Módulo nº 2

DESARROLLADOR CLOUDERA PARA APACHE HADOOP (58 h)

Objetivo:

Tecnologías clave de Hadoop. Cómo funciona HDFS MapReduce. Cómo desarrollar aplicaciones MapReduce. Cómo crear unidades de testeo (unit tests) para aplicaciones MapReduce. Cómo usar los combiners, partitioners, y la cache distribuida de un MapReduce. Mejores prácticas para el desarrollo y depuración de aplicaciones MapReduce. Cómo implementar la entrada y salida de datos de aplicaciones MapReduce. Algoritmos para tareas comunes de MapReduce. Cómo unir conjuntos de datos en MapReduce. Cómo se integra Hadoop en el CPD. Cómo usar los algoritmos Machine Learning de Mahout. Cómo utilizar Hive y Pig para el desarrollo rápido de aplicaciones. Cómo crear grandes workflows utilizando Oozie.

Contenidos teórico-prácticos:

- Introducción
- La Motivación de Hadoop
  - Problemas con los sistemas tradicionales a gran escala
  - Necesidades para una nueva aproximación
- Hadoop: Conceptos Básicos
  - El Proyecto Hadoop y Componentes Hadoop
  - El Sistema de Ficheros Distribuido Hadoop (HDFS)
  - Ejercicios prácticos: Uso de HDFS
  - Cómo funciona MapReduce
  - Ejercicio práctico: Ejecutando un Job MapReduce
  - Cómo Operan los Cluster Hadoop
  - Otros Proyectos del Ecosistema Hadoop
  - Otros Componentes del Ecosistema Hadoop
- Escribiendo una aplicación MapReduce
  - El Flujo MapReduce
  - Conceptos Básicos de la API MapReduce
  - Escribiendo Drivers, Mappers, y Reducers de MapReduce en Java
  - Escribiendo Mappers y Reducers en Otros Lenguajes utilizando la Streaming API
  - Acelerando el Desarrollo en Hadoop utilizando Eclipse
  - Ejercicio Práctico: Escribiendo un Programa MapReduce
  - Diferencias entre la antigua API de MapReduce y la nueva
- Programas Unit Testing de MapReduce
  - Unit testing

  - Los frameworks JUnit y MRUnit
  - Escribir Unit Tests con MRUnit
  - Ejercicios Prácticos: Escribiendo Unit Tests con el Framework MRUnit
- Profundizando en la API de Hadoop
  - Uso de la Clase ToolRunner
  - Reducción de los Datos Intermedios con los Combiners
  - Ejercicios Prácticos: Escribiendo e Implementando un Combiner
  - Configurando y Ajustando los Mappers y Reducers utilizando los Métodos configure y close
  - Escritura de Custom Partitioners para mejor Balanceo de Carga
  - Práctica Opcional: Escribiendo un Partitioner
  - Accediendo a HDFS desde el código de la aplicación
  - Usando la Cache Distribuida
  - Uso de la Librería de APIs de Mappers, Reducers, y Partitioners
- Consejos Prácticos de Desarrollo y Técnicas
  - Estrategias para depuración de Código MapReduce
  - Probando código MapReduce localmente utilizando el LocalJobRunner
  - Escribiendo y Visualizando Ficheros de Log
  - Obteniendo Información de los Jobs con los Counters
  - Determinando el número Óptimo de Reducers para un Job
  - Creando Jobs MapReduce Map-Only
  - Ejercicio práctico: Usando Contadores y un Job Map-Only
- Entrada y Salida de Datos
  - Creando Implementaciones Custom Writable y WritableComparable
  - Guardando Datos Binarios usando SequenceFile y Ficheros de Datos Avro
  - Implementando Input Formats y Output Formats personalizados
  - Problemas a tener en cuenta cuando se utiliza la Compresión de Ficheros
  - Ejercicios prácticos: Utilizando SequenceFiles y File Compression
- Algoritmos Típicos de MapReduce
  - Ordenando y Buscando en Grandes Conjuntos de Datos
  - Realizando una segunda ordenación
  - Indexando Datos
  - Ejercicios Prácticos: Creando un Índice Invertido
  - Computing Term Frequency –Inverse Document Frequency
  - Calculando la Co-Ocurrencia de Palabras (repetición de palabras)
  - Ejercicio práctico: Calculando la Co-Ocurrencia de Palabras (opcional)
  - Ejercicio práctico: Implementando la Co-Ocurrencia con un WritableComparable propio
- Uniendo Conjuntos de Datos en MapReduce
  - Escribiendo un Map-Side Join
  - Escribiendo un Reduce-Side Join
- Integrando Hadoop en el Workflow
  - Integrando Integrating Hadoop into an Existing Enterprise
  - Loading Data from an RDBMS into HDFS by Using Sqoop

  - Hands-On Exercise: Importing Data With Sqoop
  - Managing Real-Time Data Using Flume
  - Accessing HDFS from Legacy Systems with FuseDFS and HttpFS
- Machine Learning and Mahout
  - Introducción a Machine Learning
  - Utilizando Mahout
  - Ejercicio práctico: Usando Mahout Recommender
- Introducción a Hive y Pig
  - La Motivación de Hive y Pig
  - Fundamentos de Hive
  - Ejercicio práctico: Manipulación de datos con Hive
  - Fundamentos de Pig
  - Ejercicio práctico: Utilizar Pig para obtener los nombres de películas de nuestro Recommender
  - Eligiendo entre Hive y Pig
- Introducción a Oozie
  - Introducción a Oozie
  - Creando Workflows en Oozie
  - Ejercicio práctico: Ejecutando un Workflow en Oozie