Curso Core Networks Analista de datos SQL con Hive & Hbase

Duración: 5 días - 25 horas

Descripción del curso:

El curso trata de proporcionar los conceptos y habilidades necesarias para que los alumnos puedan desarrollar aplicaciones con Hive y HBase. El alumno conocerá como realizar sentencias SQL contra archivos almacenados en HDFS como si se trataran de tablas. Además conocerá como trabajar con datos complejos, de gran volumen o estructurados en lenguajes como JSON. En el curso se introduce al alumno a trabajar con HBase y llevar a cabo un desarrollo en torno a la misma.

Dirigido a:

Principalmente a desarrolladores, sobre todo para aquellos que tengan conocimientos y experiencia con SQL y que deseen adentrarse en el mundo de análisis de Big Data.

Requisitos:

Estar familiarizado con el lenguaje SQL y haber con bases de datos tradicionales.

Contenido del curso:

1. **Ingesta de datos**

    - Herramienta de integración con RDBMS
    - Características de Sqoop
    - Operaciones con Sqoop

- Formatos de serialización: Avro y Parquet
- Compresión y particionado

2. Apache Hive

- Arquitectura de Hive
- Hcatalog y HiveServer2
- El Metastore de Hive
- HiveQL, databases y tablas
- MapReduce y Spark para Hive
- Hive no es una RDBMS
- Introducción a HiveQL
- Tipos de datos simples y complejos
- Casting de datos y fuera de rango
- Operadores
- Tratamiento de valores null

3. Trabajando con tablas y datos en Hive

- Particionado de tablas y bucketing
- Tablas temporales y Vistas
- Modificando tablas, particiones
- Create table as select
- Create table like
- Carga de datos: Flume
- Uso de HUE y Beeline

4. Funciones con Hive

- Funciones incorporadas
- Uso de las funciones incluidas
- Funciones matemáticas y de fechas
- Funciones con Strings y URL's
- Funciones de agregación

5. Manipulación de datos

- Agregar registros a tabla existente
- Crear nueva tabla a partir existente
- Cruzar datos con Joins: tipos
- Hive y los datos estructurados y complejos

- Almacenar los resultados
- Uso de expresiones regulares

6. Optimización de Hive

- Particionado: estático y dinámico. Bucketing
- Uso adecuado de Joins
- Formatos de serialización
- Best practices con compresión
- Escenarios a evitar
- Selección del motor distribuido: MR o Spark
- Selección configuración adecuada

7. Apache HBase

- Arquitectura de HBase y casos de uso
- Características de HBase
- HBase y HDFS
- El shell de HBase
- Hive como cliente de HBase
- Apache Phoenix

8. Trabajando con tablas en HBase

- Conceptos de Column Family y Column
- Operaciones CRUD
- Propiedades de Column Family
- División de las tablas en regiones
- RegionServers en HBase
- Estructura de una HRegion
- Hfiles y MemStore
- Zookeeper, HBase Master y hbase:meta
- El WAL y tolerancia a fallos

9. Diseño de tablas en HBase

- Guias en el diseño de tablas
- Desnormalización vs normalizaión
- Diferencias con una RDBMS
- Descubrir el patrón de acceso
- Selección de estrategía para la RowKey

- RowKey’s compuestas
- Versions, Time-To-Live y Min-Versions
- Compactaciones y Region splits

10. Optimización de HBase

- Diseño de la RowKey
- Diseño de la Column Family
- BlockSize y Compression
- Bloom filters e indices secundarios
- Particionado de datos, indices
- Estrategías de caching
- HotSpot