# Curso Core Networks Analista de datos SQL con Hive & Impala

**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 e Impala. 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 o con archivos serializados como Parquet o Avro. El curso está orientado a que el alumno pueda aplicar sus conocimientos SQL al mundo Big data haciendo enfasis en las diferencias con las RDBMS clásicas.

## Dirigido a:

Principalmente a desarrolladores, sobre todo a aquellos que tengan conocimientos y experiencia con SQL y que deseen adentrarse en el mundo de análisis SQL con 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 y operaciones con Sqoop

- Características y operaciones con Flume
- Formatos de serialización: Avro y Parquet
- Compresión y particionado
- HDFS como elemento de almacenamiento
- YARN como gestor de recursos
- Los containers y el ApplicationMaster

**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 y UDF’s
- Utilizando macros: ejemplos
- Uso de las funciones incluidas
- Funciones matemáticas y de fechas
- Funciones con Strings y URL’s
- Funciones de agregación

## 5. Manipulación y carga de datos

- Agregar registros a tabla existente
- Crear nueva tabla a partir existente
- Modos de insercción de datos: flume...
- 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
- El síndrome del archivo pequeño: solución
- Uso adecuado de Joins
- Formatos de serialización: cuando usarlos
- Best practices con compresión
- Escenarios a evitar
- Selección del motor distribuido: MR o Spark
- Selección configuración adecuada
- Uso de ventanas y rangos
- Uso de SerDes adicionales
- Hive sobre Spark: implicaciones

## 7. Apache Impala

- Arquitectura de Impala
- Impala y el metastore de Hive
- Best practices a aplicar con Hive
- Diferencias entre Hive/MRv2 e Impala
- Consideraciones de RAM al usar Impala
- Impala y Admission Control
- Análisis de consultas de Impala
- Interpretar métricas de Impala

## 8. Clientes de HBase

- Consultas sobre HBase: consideraciones
- Configuración como cliente de HBase
- Tipos soportados de HBase

- Carga de datos en HBase usando Impala
- Optimización del rendimiento
- Apache Phoenix como capa SQL de HBase
- Instalación y configuración de Phoenix

**9. Seguridad**

- Autenticación y autorización
- Herramientas de autorización
- Planificación de los roles de acceso
- Aplicar seguridad a diferentes niveles