BINAIA

BCIT in

# Big Data Engineering

binaia.es

# Objetivos

El objetivo general de la formación consiste en preparar a los alumnos de forma intensiva en alguna de las principales herramientas utilizadas en el área de Big Data, llegando a conocerlas y a ser capaces de desenvolverse en estas de forma fluida consiguiendo ser productivos de forma inmediata tan pronto ingresan en un equipo "data driven" o también denominados orientado al dato.

En concreto, los objetivos específicos de la formación son:

- Conocer el origen del framework Apache Hadoop y sus componentes.
- Entender el concepto de Map Reduce y procesamiento distribuido.
- Aprender a realizar scripts de Apache Pig y su utilidad.
- Conocer Apache Sqoop y su utilidad.
- Entender Apache Flume.
- Aprender a utilizar Apache Hive.
- Aprender a desarrollar pipelines de ingesta y transformación de datos en Spark.
- Conocer los conceptos básicos de Apache Kafka.
- Realizar procesamiento de datos en streaming con Apache Kafka, Prestodb y Apache Spark.
- Conocer la biblioteca Delta y aprender a crear un Delta Lake.
- Conocer herramientas de creación de flujos de datos y orquestación como Apache Nifi y Apache Airflow.
- Aprender a identificar la herramienta idónea para cada caso de uso.
- Conocer las principales distribuciones y entornos de desarrollo Big Data.
- Aprender a utilizar Git como herramienta de control de versiones.

# Requisitos

Es necesario tener nociones de básicas de programación, conocimiento del lenguaje de consultas SQL. Deseable tener conocimientos en alguna herramienta de control de versiones.

# Metodología

En esta formación, el aprendizaje se llevará a cabo de una forma dinámica con una enseñanza teórica y práctica compaginando ambas mediante la realización de ejercicios que permitan hacer uso de todo lo aprendido buscando asentar el conocimiento.

# Programa

El programa se estructura en los bloques que se describen a continuación:

## Bloque 1: Fundamentos básicos de Big Data

- Apache Hadoop y sus componentes (HDFS, Map Reduce, Yarn).
- Apache Sqoop.
- Apache Pig.
- Apache Flume.
- Apache Hive y formatos de archivo (CSV, Avro, Parquet).
- Spark: RDDs y DStreams.
- Distribución Cloudera.

## Bloque 2: Apache Spark Básico

- Spark-shell, Pyspark.
- DataFrames.
- Jupyter notebooks.
- Databricks.

## Bloque 3: Apache Spark Avanzado

- DataFrames y Spark SQL.
- Lectura/Escritura de datos en ficheros: CSV, JSON, Parquet, Avro.
- Lectura/Escritura de datos desde Bases de Datos relacionales y Apache Hive.

- Joins, Window Operations y Cache.

## Bloque 4: Procesamiento de Streams

- Conceptos básicos de Apache Kafka.
- Consulta de topics de Kafka utilizando Prestodb.
- Procesamiento de streams con PySpark (DStreams y Structured Streaming) y Apache Kafka.

## Bloque 5: Delta Lake

- Qué es un Delta Lake, características y las diferencias con un Data Lake.
- Entender:
- Transaction Log (Delta Log).
- Time Travel.
- Change Data Capture (CDC).
- VACUUM.
- Merge command.

## Bloque 6: Orquestación

- Aprender Nifi.
- Aprender Airflow.

## Bloque 7: Clúster Big Data

- Aprender a crear tu propio clúster y distribución Big Data

## Bloque 8: Especialización

En este bloque se llevará a cabo una especialización en una de las siguientes herramientas:

- AWS
- Databricks
- ELK
- Snowflake
- Terraform