# Principales herramientas Big Data en el 2019

## Ingesta o adquisición de datos

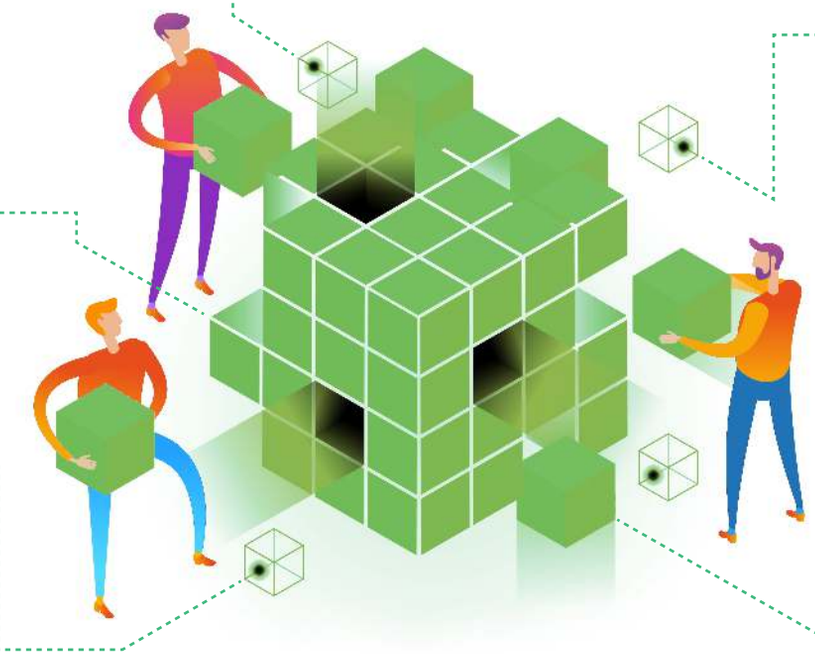
- **Kafka** produce y consume eventos en tiempo real.
- **Flume** permite la ingesta de logs o eventos para enviarlos a sistemas de almacenamiento distribuidos.
- **Sqoop** Importa y expota bb.dd relacionales a HDFS o Hive.

## Procesamiento batch

- **Spark** framework de procesamiento mediante APIs de alto nivel como Dataframes o SQL.
- **Hive** almacena datos sobre HDFS para procesar y analizar datos con un interfaz como SQL.

## Procesamiento en Streaming

- **Spark Structured** aporta capacidades para procesar eventos mediante el motor de Spark SQL.
- **Kafka Streams**, es una librería para la construcción de aplicaciones en Streaming.
- **Flink** una tecnología de real time que hace frente a Spark.



## Almacenamiento NoSQL

Bases de datos clave-valor:

- **Redis** en memoria.
- **HBase**, de tipo column family integrada con el ecosistema de Hadoop.
- **Cassandra** de tipo column family con una arquitectura en anillo.
- **MongoDB** orientada a documentos sin esquema.
- **Elasticsearch** motor de optimizado para búsqueda libre y analítica a escala.

## Analítica de datos

- **Python** lenguaje de programación estadístico y analítico optimizado para modelización avanzada.
- **R entorno** software para computación estadística y analítica para el desarrollo de modelos.
- **Keras** y **PyTorch** APIs de alto nivel para desarrollo de redes neuronales

Firmado por: **Francisco Javier Lahoz**

Head of Data Engineering en Orange
Director del Master en Big Data Management ICEMD