

ECOSISTEMAS ALTERNATIVOS A HADOOP



Apache Spark

- Definición: Plataforma de código abierto para procesamiento de datos a gran escala.
- Antecedentes: Desarrollado en el AMPLab de la Universidad de California en 2009.
- Características y beneficios: Interfaz para programación de clusters completos, paralelismo de datos implícito, tolerancia a fallos, hasta 100 veces más rápido que Hadoop MapReduce para ciertas aplicaciones.
- Arquitectura: Basada en RDD y DAG.
- Costos: Código abierto y gratuito.

GOOGLE CLOUD DATAFLOW

- Definición: Servicio de procesamiento de datos totalmente administrado en la nube.
- Antecedentes: Evolución de los sistemas internos de Google, como MapReduce, FlumeJava y Millwheel.
- Características y beneficios: Programación simplificada, procesamiento por lotes y en tiempo real.
- Arquitectura: Basada en la nube, utiliza un modelo de programación basado en colecciones paralelas de datos.
- Costos: Servicio de pago, con costos variables según el uso.



AMAZON EMR

- Definición: Plataforma de big data en la nube para procesamiento de grandes cantidades de datos.
- Antecedentes: Desarrollado por Amazon para procesar grandes cantidades de datos eficientemente.
- Características y beneficios: Entorno seguro, escalable y flexible, compatible con marcos de trabajo como Apache Spark y Hadoop.
- Arquitectura: Basada en la nube, utiliza tecnologías de big data como Hadoop.
- Costos: Servicio de pago, con costos variables según el uso.



DATABRICKS:

- Definición: Plataforma de análisis unificada basada en la nube.
- Antecedentes: Fundada por los creadores de Apache Spark.
- Características y beneficios: Espacio de trabajo colaborativo, análisis en tiempo real, aprendizaje automático.
- Arquitectura: Basada en la nube, utiliza Apache Spark.
- Costos: Servicio de pago, con costos variables según el uso.



Apache Flink

- Definición: Marco de procesamiento de flujo de código abierto.
- Antecedentes: Creado por la comunidad de Apache Software Foundation.
- Características y beneficios: Procesamiento en tiempo real, tolerancia a fallos.
- Arquitectura: Utiliza un motor de streaming.

