

Nombre: Pablo Elías Ramírez Escalante		Matrícula: AL02883894
Infraestructura para Big Data		Nombre del profesor: Miguel de Jesús Martínez Felipe
Módulo 1		Actividad 1
Fecha: 20/01/2024		
Bibliografia: <p>DataScience ForBusiness (2020) <i>Big data II: Tecnología Big Data EN 6 minutos, YouTube.</i> Recuperado de: https://www.youtube.com/watch?v=EfOMesB7sMQ</p> <p>Google Cloud (s.f.) <i>Modelo de Programación de Apache beam cloud dataflow google cloud, Google.</i> Recuperado de: https://cloud.google.com/dataflow/docs/concepts/beam-programming-model?hl=es-419</p> <p>MongoDB (s.f.) <i>La Plataforma de Datos Para Aplicaciones, MongoDB.</i> Recuperado de: https://www.mongodb.com/es</p> <p>Normesta (2023) <i>Tutorial: Azure Data Lake Storage Gen2, Azure Databricks y spark - azure storage, Azure Storage Microsoft Learn.</i> Recuperado de: https://learn.microsoft.com/es-es/azure/storage/blobs/data-lake-storage-use-databricks-spark</p> <p>Sosa Olascoaga, J. (2017). Generador de arquitecturas de solución sobre el ecosistema Hadoop para problemas de Big Data. Universidad de los Andes. Recuperado de: https://repositorio.uniandes.edu.co/flip/?pdf=https://repositorio.uniandes.edu.co/server/api/core/bitstreams/7be85e8a-a1b0-4dce-b565-280e5b136e0b/content</p>		



TOP

1. Sentido del Humor
2. Bondad
3. Juicio

Hadoop Distributed File System (HDFS) con Apache Flink:

Definición: Flink es un sistema de procesamiento de datos en tiempo real y por lotes, mientras que HDFS es un sistema de archivos distribuido diseñado para almacenar grandes cantidades de datos de manera confiable.

Antecedentes: HDFS es parte del proyecto Apache Hadoop, mientras que Flink es un proyecto independiente de la Apache Software Foundation.

Flink ofrece procesamiento de datos en tiempo real y por lotes, mientras que HDFS ofrece almacenamiento distribuido y tolerante a fallos.

Arquitectura: HDFS emplea un modelo maestro/nodo, mientras que Flink emplea un modelo de flujo de datos.

Costos: Los costos del hardware pueden estar relacionados con HDFS, que es de código abierto. Flink es de código abierto y los gastos pueden depender de los recursos de la nube o del hardware.

Google Cloud Storage con Apache Beam:

Definición: Google Cloud Storage es un servicio de almacenamiento de objetos en la nube, y Apache Beam es un modelo de procesamiento de datos unificado que permite la ejecución en múltiples motores de ejecución.

Antecedentes: Apache Beam es un proyecto de la Fundación de Software Apache, y Google Cloud Storage es un servicio de Google Cloud Platform.

Características: Apache Beam ofrece un modelo unificado para el procesamiento de datos en batch y streaming, mientras que Google Cloud Storage ofrece escalabilidad y durabilidad.

Arquitectura: Google Cloud Storage emplea una arquitectura de almacenamiento de objetos en la nube, y Apache Beam permite la ejecución en varios motores de procesamiento.

Costos: El almacenamiento y la transferencia de datos son los costos de Google Cloud Storage. Los costos de Apache Beam pueden variar según la infraestructura utilizada y es de código abierto.

Microsoft Azure Data Lake Storage con Databricks:

Definición: Databricks es una plataforma de análisis de datos basada en Apache Spark, y Azure Data Lake Storage es un servicio de almacenamiento de datos en la nube.

Antecedentes: Databricks se originó como un proyecto independiente basado en Apache Spark, pero ambos son productos de Microsoft Azure.

Características: Databricks ofrece capacidades avanzadas de análisis y procesamiento de datos con Apache Spark, mientras que Azure Data Lake Storage ofrece almacenamiento de datos escalable.

Arquitectura: Databricks se basa en la arquitectura de Apache Spark, y Azure Data Lake Storage sigue una arquitectura de almacenamiento en la nube.

Costos: El almacenamiento y el acceso a los datos son los costos de Azure Data Lake Storage. El uso de la plataforma y los recursos de la nube puede generar costos para Databricks.



MongoDB:

Definición: MongoDB es una base de datos NoSQL orientada a documentos. Tiene flexibilidad y escalabilidad horizontal al almacenar datos en formato BSON (Binary JSON).

Antecedentes: MongoDB fue creado por MongoDB Inc. y ha ganado popularidad en entornos que requieren un manejo eficiente de datos no estructurados o semi estructurados.

Características: estructura flexible, consultas basadas en documentos JSON, escalabilidad horizontal y capacidades avanzadas de indexación.

Arquitectura: Permite una mayor flexibilidad en la representación de la información al almacenar los datos en documentos BSON en lugar de tablas convencionales.

Costos: MongoDB solo ofrece una versión de código abierto, pero también hay una versión comercial que incluye más funciones y soporte técnico. Los precios pueden variar según la versión y los servicios asociados.

Apache Spark:

Definición: Apache Spark es un marco de procesamiento de datos en memoria simple y rápido. Permite el análisis distribuido de datos y es tolerante a fallos.

Contexto: Spark, que fue creado en la Universidad de California, Berkeley, se ha convertido en un estándar para el procesamiento de datos a gran escala.

Características: Procesamiento en memoria, soporte para lenguajes como Scala, Python y Java y capacidades de procesamiento de datos en tiempo real.

Arquitectura: DAG (Grafo Acíclico Dirigido) utiliza un modelo de ejecución y ofrece módulos para aprendizaje automático, transmisiones, SQL y procesamiento de lotes.

Costos: Spark es gratuito y de código abierto. Si se utilizan los recursos de hardware y los servicios de nube, los costos asociados aumentarán.