**Comparativa de Herramientas de Big Data: Spark vs Hadoop**

Nombre del autor: Keilor Fallas Prado  
Fecha: 13 de diciembre de 2024

**Introducción al tema:**  
El Big Data se ha convertido en un pilar fundamental para la toma de decisiones en diversas industrias. En este artículo, exploraremos dos herramientas clave en este campo, Apache Spark y Apache Hadoop, comparando sus características, casos de uso y rendimiento. Nuestro objetivo es proporcionar una guía que ayude a los lectores a elegir la solución más adecuada para sus necesidades.

**Desarrollo del tema propuesto:**

Para comenzar a hablar sobre este tema tenemos que entender cada una de las herramientas del big data, una de ellas es Hadoop es una herramienta que comienza en el 2006 como un proyecto de Yahoo, donde se convierte en un código abierto Apache de suma importancia más tarde, Hadoop también conocido por sus siglas HDFS es el mas robusto para almacenar datos y tiene gran poder de escalabilidad. Puede almacenar archivos en el formato nativo de Hadoop y los paraleliza en un clúster, otra característica importante en resaltar de esta herramienta es que está construida en el lenguaje Java y se por medio de Hive se logra realizar consultas similares a la interfaz de SQL.

Ahora vamos a conocer un poco sobre Spark que es otra herramienta del big data muy utilizada actualmente para el manejo de datos, esta herramienta es impulsada en torno a Spark Core el motor que impulsa la programación, las optimizaciones y la abstracción de RDD. Tal que esto se utiliza mucho por los científicos de datos se lo logro utilizar Python y R para hacer consultas, aparte de estos dos programas también se logra usar Java para hacer trabajos con spark. En el momento actual Databricks supervisa el desarrollo de Spark. Una característica importante del spark es que trabaja con procesamiento en memoria por lo que consume RAM y por ende tiende a necesitar mucho espacio en trabajos grandes y como se menciono anteriormente que es un modelo de programación en RDDs, estos son inmutables y distribuidos de objetos que permiten realizar operaciones.

En comparativa de estas dos herramientas sería la mejor es la pregunta, Hadoop al ser una herramienta muy robusta para mayor procesamiento de datos y de alta escalabilidad, con gran tolerancia a fallas y flexibilidad, mientras Spark una herramienta que sobresale por su velocidad, también escalable, pero necesitara más RAM, por ende elevaría los costos, al analizar estas dos herramientas ninguna es mejor que otra si no depende las necesidades que la persona o empresa necesitar satisfacer.

**Aplicaciones:**En Hadoop algunas aplicaciones son:

Telecomunicaciones e investigaciones Científica: estos utilizan almacenamiento y procesamiento Batch

Automotriz: Un ejemplo puede ser en los sensores de los autos para analizarlos a largo plazo

Spark:

Marketing y publicidad: estos datos se necesitan analizar en tiempo real.

Salud y comercio electrónico: en estos casos se necesitan métodos de predicción como machine learning.

**Conclusión:**Si necesitas una solución económica para almacenar y procesar gran cantidad de datos de forma batch y que no requiera análisis muy rápido o tiempo real la mejor opción será Hadoop.

Si necesitas un procesamiento en tiempo real o tareas como machine learning y puedes asumir los costes necesarios la mejor opción es Spark.

Bibliografía:

[Hadoop vs. Spark: una comparación cara a cara | Logz.io](https://logz.io/blog/hadoop-vs-spark/)

[Spark vs Hadoop MapReduce: 5 Key Differences | Integrate.io](https://www.integrate.io/blog/apache-spark-vs-hadoop-mapreduce/)

[Apache Hadoop 3.4.1 – HDFS Architecture](https://hadoop.apache.org/docs/r3.4.1/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html)

[Apache Hadoop 3.4.1 – MapReduce Tutorial](https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html)  
[HDFS Architecture Guide](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html)  
[Overview - Spark 3.5.3 Documentation](https://spark.apache.org/docs/latest/)