Tipo : Lectura
Capítulo : Big Data 2.0 - Spark

I. OBJETIVO

Ampliar sus conocimientos sobre Spark.

II. LECTURAS COMPLEMENTARIAS

Se sugiere revisar los siguientes enlaces para profundizar en los conceptos tratados en el presente capítulo.

Spark vs Hadoop, ¿quién saldrá vencedor?

Apache **Spark vs Hadoop** son dos de los productos más importantes y conocidos de la familia de Big Data.

Aunque hay quienes ven estos dos frameworks como competidores en el espacio de big data, no es tan fácil hacer una comparación Spark vs Hadoop. Hacen muchas cosas igual, pero hay algunas áreas donde ambos no se superponen. Por ejemplo, **Apache Spark no tiene sistema de archivos y, por lo tanto, depende del sistema de archivos distribuido de Hadoop.**

Si revisas Google Trends, podrás ver que Hadoop tiene más popularidad en comparación con Apache Spark. Pero a pesar de esto, empresas como Yahoo, Intel, Baidu, Trend Micro y Groupon ya están utilizando Apache Spark.

Apache Spark vs Hadoop son comparables en diferentes parámetros. ¿Quieres saber cuáles son los campos que marcan la diferencia?

Spark vs Hadoop. La batalla está servida

La resolución de enigma Spark vs Hadoop está servida en tres claves:

a) **Usabilidad.** Una de las cuestiones más habituales al contrastar ambos frameworks está relacionada con su facilidad de uso. **¿Cuál es más user friendly? ¿Spark vs Hadoop?** En este caso **Apache Spark superaría a su contrincante puesto que viene equipado con APIs realmente sencillas para Scala, Python, Java y Spark SQL.**Además, aporta feedback en formato REPL sobre los comandos. Por su parte, si bien es verdad que MapReduce tiene complementos como Pig y Hive que lo hacen algo más fácil de usar, al final lo que sucede es que **la lógica simple necesita más programación (los programas deben estar escritos en Java), por lo que lo que se gana en usabilidad por una parte quedaría perdido por otra.**

b) **Rendimiento.** Este punto quizás sea el más complicado de resolver en cualquier comparativa Spark vs Hadoop. La cuestión es que, **como ambos procesan los datos de manera diferente, no es nada fácil determinar quién logra un mayor desempeño.** Para tomar una decisión habría que tener en cuenta que:

En lo que respecta a **Spark:**

- **Trabaja in memory** y, por lo tanto, todos los procesos se aceleran.
- Pero necesita más memoria para el almacenamiento.
- **Su rendimiento puede verse mermado debido a la necesidad de utilizar aplicaciones pesadas.**

En el caso de **Hadoop:**

- **Los datos están en disco y eso hace que todo resulte más lento.**
- La ventaja es que, en comparación con la otra alternativa, las necesidades de almacenamiento son inferiores.
- **Al ocuparse de eliminar los datos cuando no son ya necesarios, no produce pérdidas de rendimiento significativas para aplicaciones pesadas.**

c) **Seguridad**. Si en usabilidad Spark vencía a Hadoop, en este caso no tiene nada que hacer. **Hadoop no tiene rivales** ya que:

- **Proporciona a sus usuarios todos los beneficios de los avances obtenidos en los proyectos de seguridad de Hadoop** (Knox Gateway o Sentry son algunos ejemplos).
- HDFS admite la autorización de nivel de servicio, que garantiza los permisos adecuados para los clientes a nivel de archivo
- Y, además.... tiene **Hadoop YARN**

Por su parte, **Spark necesita ejecutarse en HDFS** para acceder a permisos de nivel de archivo y, además, **para obtener beneficios de seguridad ha de recurrir a Hadoop YARN.**

Pero entonces, **¿quién puede considerarse vencedor de la competición Spark vs Hadoop?** Cada uno domina al otro en distintas áreas. Por ejemplo, **Hadoop sería la elección acertada cuando el tamaño de la memoria es significativamente menor que el tamaño de los datos; pero, si se busca rapidez, no cabría plantearse otra opción que Spark. ¿Con cuál te quedas?** ¿Crees que Spark podría terminar sustituyendo a MapReduce? ¿Te parece más factible que Hadoop continúe disfrutando de su hegemonía?

Bibliografía

- **Spark vs Hadoop, ¿quién saldrá vencedor?**
  https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/spark-vs-hadoop-quien-saldra-vencedor
  Consulta: 01 de diciembre del 2017