

Jeff Feng, administrador
de productos de Tableau Software

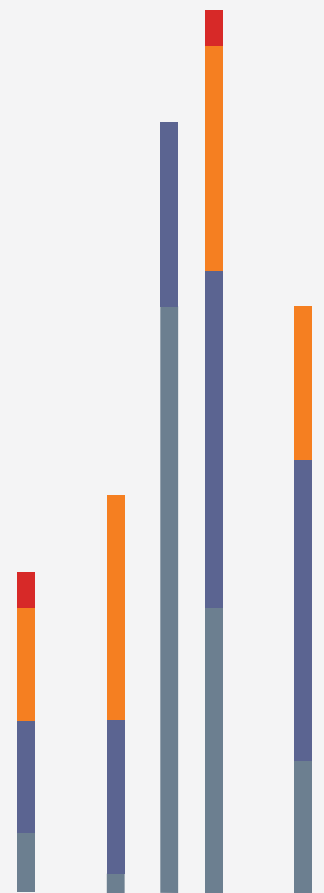
5 prácticas recomendadas para Tableau y Hadoop

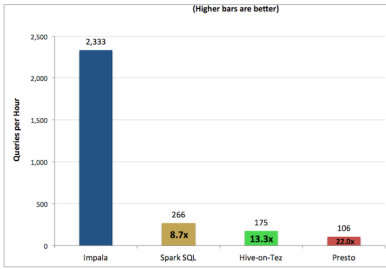
Tableau está diseñado para favorecer las conversaciones con datos en tiempo real y en múltiples plataformas. Los usuarios corporativos que alguna vez sintieron que las herramientas tradicionales obstaculizaban su trabajo adoptaron este modus operandi. Pero ¿qué sucede cuando las consultas se responden después de horas o minutos en lugar de unos segundos? ¿Puede mantenerse su “flujo”?

Vivimos en una era en la que las personas pueden analizar millones o incluso miles de millones de filas de datos y los usuarios esperan obtener resultados casi al instante ([consulte el estudio](#) sobre la regla de los 2 segundos para la recuperación de información). Cuando los tiempos de respuesta y las interacciones del usuario tardan más de 2-3 segundos, este se distrae del “flujo del análisis visual”. Por lo tanto, es fundamental proporcionar rápidas velocidades de consulta para que los usuarios se mantengan interesados y obtengan más información de sus implementaciones de big data.

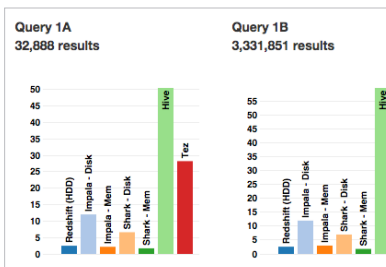
Los usuarios pueden aplicar distintas prácticas recomendadas para maximizar el rendimiento de sus visualizaciones y dashboards de Tableau en plataformas de big data. La gran mayoría de las prácticas recomendadas se relacionan con las cinco actividades siguientes:

1. Aprovechamiento de un rápido motor de consulta interactiva
2. Uso estratégico de conexiones en vivo y extracciones
3. Ajuste de los datos procedentes del mar de datos
4. Optimización de las extracciones
5. Personalización del rendimiento de la conexión





- *New Benchmarks for SQL-on-Hadoop: Impala 1.4 Widens the Performance Gap (Nuevas pruebas comparativas de SQL en Hadoop: Impala 1.4 amplía la brecha de rendimiento)*



- *Una prueba comparativa de big data ofrece comparaciones cuantitativas y cualitativas de cinco sistemas.*

Aprovechamiento de un rápido motor de consulta interactiva

Las consultas de Hive ejecutadas en Hadoop con MapReduce son intrínsecamente lentas debido a la sobrecarga asociada con el mapeo de las consultas SQL en los trabajos de MapReduce. Hive en MapReduce es magnífico para el procesamiento en lotes, como el de aplicaciones de extracción, transformación y carga (ETL), porque presenta una gran tolerancia a errores; sin embargo, su rendimiento no es muy satisfactorio. Las mejoras de Hive introducen nuevos marcos de trabajo para las aplicaciones, como Tez (posibilita las consultas interactivas) y Spark (permite el procesamiento en memoria), que incrementan significativamente las velocidades de consulta.

Además de Hive en Hadoop, existen muchas opciones magníficas para acelerar sus consultas. Según pruebas comparativas recientes, Impala es ampliamente considerado como el motor más rápido en Hadoop. Y, aunque se encuentra en las primeras etapas de desarrollo, Spark SQL mostró un gran potencial para convertirse en un rápido motor de procesamiento de datos. Puede procesar datos almacenados en Hadoop o SchemaRDD de Spark a los que se hace referencia mediante un almacén de metadatos de Hive. Tanto Impala como Spark SQL son conectores reconocidos y admitidos por Tableau. Pivotal HAWQ, Presto y Apache Drill son otras tecnologías que suelen mencionarse en debates acerca del rendimiento en Hadoop.

Otra alternativa es considerar opciones fuera de Hadoop. Bases de datos de análisis rápidas, como Actian Vector, HP Vertica, Teradata Aster Data, SAP Hana, ParAccel, Pivotal Greenplum y otras, pueden ser muy útiles a fin de hospedar datos para consultas de baja latencia de los usuarios corporativos de Tableau después del procesamiento en Hadoop. Asimismo, los servicios de infraestructura hospedados en la nube son cada vez más populares. Google BigQuery saca partido de la infraestructura masiva de Google, que destaca en el procesamiento de datos y la habilitación de consultas rápidas, especialmente en grandes conjuntos de datos. Por otro lado, Amazon Redshift es un almacén de datos en columnas completamente administrado que se centra en el acceso rápido a los datos. Finalmente, existe un grupo de tecnologías emergentes de proyectos nuevos y proyectos de código abierto que usan cubos OLAP (AtScale, eBay Kylin) o motores de indexación (JehroData) para Hadoop y proporcionan la capacidad de hacer consultas a mil millones de filas o más con baja latencia.

2.

Uso estratégico de conexiones en vivo y extracciones

La arquitectura de datos híbrida de Tableau para la conexión a una fuente de datos en tiempo real o al motor de datos de Tableau mediante una extracción en memoria proporciona a los usuarios una gran flexibilidad para trabajar con big data. Las extracciones son ideales para situaciones en las que los motores de consulta rápidos no están disponibles, los conjuntos de datos son pequeños o medianos (cientos de millones de filas o menos), o el análisis sin conexión es necesario. Para conjuntos de datos más grandes, Hadoop Hive y otros motores de consulta presentan una mejor escalabilidad que Tableau debido a su ejecución distribuida. Además, cuando hay un motor de base de datos rápido o se requiere un análisis en tiempo real, una conexión en vivo es la mejor opción. La lista completa de escenarios recomendados se presenta a continuación, en la Figura 1.



 Extracciones recomendadas	 Conexiones en vivo recomendadas
Ejecución lenta de consultas a base de datos	Cuando hay un rápido motor de consulta de base de datos
Cuando se usan conjuntos de datos más pequeños (p. ej., cientos de millones de filas o menos)	Cuando se requieren conjuntos de datos grandes
Cuando se requiere análisis sin conexión	Cuando se requiere análisis en tiempo real
Cuando se usa SQL personalizado	Cuando un libro de trabajo usa funciones SQL en bruto para transferencia
Cuando se requieren funcionalidades de análisis adicionales (configuración, clasificación, distinción, mediana)	Cuando se requiere seguridad eficaz en el nivel del usuario (excepto para extracciones publicadas en Tableau Server)

Figura 1: Condiciones recomendadas para extracciones frente a conexiones en vivo.

3.

Ajuste de los datos procedentes del mar de datos

Uno de los tantos beneficios de Hadoop es que su escalabilidad, rentabilidad y capacidad de manejar datos sin estructura lo convierten en el mar de datos ideal: un repositorio que contiene todos sus datos en formato nativo. Tableau es una herramienta eficaz para explorar sus datos en el mar de datos. Sin embargo, si desea que sus trabajadores del conocimiento logren el máximo rendimiento de sus visualizaciones con Hadoop, lo primero que debe hacer es ajustar el conjunto de datos.

Como administrador de TI, existen varias técnicas que puede aplicar para incrementar la eficacia de su clúster de Hadoop:

Diseño de partición: organizar una tabla de Hive en archivos individuales (cada uno con muchos bloques de datos), en un sistema distribuido con uno o más campos de partición, puede acelerar significativamente las consultas mediante una consulta filtrada por un campo no particionado.

Tamaño del conjunto de datos: cuando sabe qué dimensiones y medidas desea observar en un conjunto de análisis y conoce el rango de los registros, limitar el conjunto de datos final que se expone a sus trabajadores del conocimiento garantiza la mejora del rendimiento.

Campos agrupados en clústeres como campos agrupados y claves de unión: los campos agrupados en clústeres pueden dictar la forma en que los datos de la tabla se separan en el disco. Las funciones JOIN y GROUP BY aplicadas a los campos agrupados en clústeres mejoran el rendimiento.

Formato de archivos de almacenamiento: el formato de los archivos desempeña un papel clave en la ejecución eficaz de las consultas. Emplee el formato de archivos que mejor se ajuste al motor de consulta que esté usando. Para Hive, el mejor formato es el de filas y columnas optimizadas (ORC), y, para Impala, el mejor es Parquet.

Diseño del modelo de datos:

- Tipos de datos: use tipos numéricos siempre que sea posible, ya que son mucho más rápidos que las cadenas.
- Uniones: evite las uniones innecesarias, ya que se implementan inadecuadamente en muchos sistemas de big data. Si usa uniones, ejecute primero una declaración COMPUTE STAT para ayudar al motor de procesamiento a optimizar automáticamente el rendimiento de las consultas de uniones.
- Fórmulas: evite el uso de fórmulas que no se puedan evaluar eficazmente.

4.

Optimización de las extracciones

El motor de datos de Tableau es una base de datos de análisis en memoria que saca partido de toda la jerarquía de la memoria, desde el disco hasta la caché en L1. Puede ser una herramienta eficaz para acelerar su análisis. Aunque no está diseñado para la misma escala que Hadoop, el motor de datos de Tableau puede proporcionar resultados de baja latencia basados en extracciones de datos con una cardinalidad de cientos de millones de filas y una gran cantidad de columnas. Si bien el aprovechamiento de extracciones en el motor de datos de Tableau suele incrementar el rendimiento al instante, existen diferentes oportunidades para acelerar las consultas mediante la compresión de sus datos:

Definición de filtros: cree filtros para concentrarse en los datos que le interesan.

Ocultamiento de campos en desuso: oculte los campos que no sean necesarios para el análisis a fin de que sus extracciones sean compactas y concisas.

Agregación de dimensiones visibles: cuando no necesite datos específicos, haga agregaciones previas de estos para obtener una vista más general y ver la misma información con consultas más rápidas.

Fechas agrupadas: agrupe las fechas de acuerdo a escalas de tiempo más generales siempre que sea posible.

Muestreo: para las bases de datos que lo admiten, el muestreo puede compactar los datos en gran medida y seguir representando las tendencias generales de estos.

N principales: si solo busca los valores más altos de un conjunto de datos, este es un método eficaz para reducir el tamaño del conjunto de datos.

5.

Personalización del rendimiento de la conexión

Como usuario de Tableau, tiene numerosas oportunidades de optimizar el rendimiento de su conexión para las consultas en vivo:

SQL personalizado: el SQL personalizado le permite usar expresiones SQL como base para una conexión de Tableau. Además, puede ser especialmente eficaz para restringir el tamaño del conjunto de datos (usando la cláusula LIMIT) a fin de que usted pueda explorar un nuevo conjunto de datos o trazar un perfil de él.

SQL inicial: el SQL inicial otorga la capacidad de definir parámetros de configuración y realizar el trabajo cuando establece una conexión. Puede hacer cosas como estas:

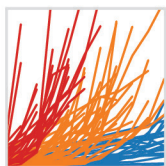
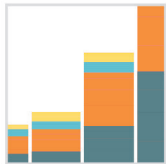
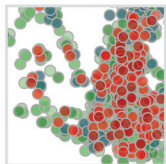
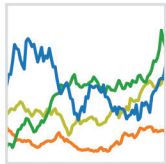
- Incrementar el paralelismo para el análisis de datos mediante la reducción del tamaño predeterminado de bloques para funciones Map y Reduce.
- Optimizar el rendimiento de las uniones mediante la activación de campos agrupados en clústeres.
- Ajustar configuraciones de distribuciones inconsistentes mediante la activación de ajustes que indican a Hive que aplique un enfoque diferente para los trabajos de MapReduce.

Resumen

Llegó la era de los big data: los volúmenes de datos se incrementan y las organizaciones llevan sus infraestructuras de datos a Hadoop, Spark y NoSQL para admitir los nuevos datos estándares. Gracias a Tableau y su capacidad de desarrollar el potencial de usuarios corporativos comunes, la información visual de los big data está llegando a todas las personas. Mediante la aplicación de las prácticas recomendadas y la personalización de estas para que se ajusten a sus aplicaciones, podrá maximizar el valor agregado de sus inversiones en big data.

Acerca de Tableau

Tableau ayuda a las personas a ver y comprender datos. Tableau ayuda a todas las personas a analizar, visualizar y compartir información rápidamente. Más de 26 000 cuentas de clientes obtienen resultados rápidos con Tableau, en la oficina o en cualquier otro lugar. Además, decenas de miles de personas usan Tableau Public para compartir datos en sus blogs y sitios web. Para ver la forma en que Tableau puede ayudarlo, descargue la versión de prueba gratuita en www.tableau.com/es-es/trial.



Recursos adicionales

[Descargar versión de prueba gratuita](#)

[Demostraciones de productos](#)

[Capacitación y tutoriales](#)

[Comunidad y soporte](#)

[Historias de clientes](#)

[Soluciones](#)

Informes relacionados

[La visión de Tableau sobre los big data](#)

[7 sugerencias para tener éxito con los big data](#)

[Fostering a Data-Driven Culture: A Special Report from the Economist Intelligence Unit and Tableau \(Creación de una cultura impulsada por los datos: un informe especial de Economist Intelligence Unit y Tableau\)](#)

[Big Data: The Next Industrial Revolution \(Big data: la próxima revolución industrial\)](#)

[Tableau Software and Big Data \(Tableau Software y los big data\)](#)

[Aberdeen Group: Maximizing the Value of Analytics and Big Data \(Aberdeen Group: cómo maximizar el valor del análisis y de los big data\)](#)

[Ver todos los informes](#)