Resumen 7

Edgar André Araya Vargas 2020142856

BigQuery Technical Whitepaper

Google maneja Big Data diariamente para sus servicios como Búsqueda, YouTube, Gmail y Google Docs. Utilizan una tecnología llamada Dremel, que es un servicio de consultas que permite realizar consultas similares a SQL en conjuntos de datos muy grandes y obtener resultados precisos en segundos. Tanto los ingenieros como otros roles en Google utilizan Dremel regularmente.

Dremel es una tecnología clave utilizada por Google desde 2006, que ha evolucionado continuamente en los últimos 6 años. Ha sido utilizada en diversas aplicaciones, como el análisis de documentos web, seguimiento de datos de instalación de aplicaciones en el mercado de Android, informes de errores, resultados de OCR de Google Books, análisis de spam, depuración de mapas en Google Maps, migraciones de tabletas en instancias de Bigtable, monitoreo de recursos, entre otros.

Google ha lanzado recientemente BigQuery como un servicio disponible para cualquier negocio o desarrollador, permitiendo a terceros aprovechar la potencia de Dremel para el procesamiento de grandes volúmenes de datos. BigQuery proporciona características similares a Dremel, como acceso a través de una API REST, interfaz de línea de comandos, gestión de esquemas de datos y la integración con Google Cloud Storage. BigQuery y Dremel comparten la misma arquitectura y características de rendimiento, lo que permite a los usuarios aprovechar la infraestructura computacional masiva de Google. Además, BigQuery ofrece beneficios como la replicación en múltiples regiones y alta escalabilidad de centros de datos, sin necesidad de administración por parte del desarrollador.

Comparado con MapReduce, Dremel está diseñado como una herramienta de análisis de datos interactiva para conjuntos de datos grandes, mientras que MapReduce es un marco de programación para procesar lotes de datos grandes. Dremel puede finalizar la mayoría de las consultas en segundos o segundos, mientras que MapReduce puede tomar mucho más tiempo. MapReduce es una tecnología de cómputo distribuido que permite procesar datos en paralelo, pero no es adecuada para análisis ad hoc o tareas de análisis iterativo o de una sola vez debido a su lenta velocidad de respuesta.

BigQuery y MapReduce son tecnologías fundamentalesmente diferentes y se utilizan en diferentes casos de uso. BigQuery es adecuado para consultas interactivas y de prueba y error en conjuntos de datos grandes, mientras que MapReduce se utiliza para procesamiento por lotes de conjuntos de datos grandes y conversiones o agregaciones de datos que consumen mucho tiempo.

Dremel puede escanear 35 mil millones de filas sin un índice en cuestión de segundos. Dremel, el servicio de consultas en paralelo masivamente distribuido en la nube, comparte la infraestructura de Google, por lo que puede paralelizar cada consulta y ejecutarla en decenas de miles de servidores simultáneamente. Puedes ver las economías de escala inherentes a Dremel. La plataforma en la nube de Google permite lograr un rendimiento de consulta súper rápido con una relación costo-valor muy atractiva. Además, no se requiere una inversión de capital por parte del usuario para la infraestructura de soporte.

Dremel almacena datos en su almacenamiento columnar, lo que significa que separa un registro en valores de columna y almacena cada valor en un volumen de almacenamiento

diferente, mientras que las bases de datos tradicionales normalmente almacenan todo el registro en un solo volumen. Esto se llama almacenamiento columnar y se ha utilizado en soluciones tradicionales de almacén de datos. El almacenamiento columnar tiene las siguientes ventajas: minimización de tráfico y mayor relación de compresión.

En resumen, BigQuery ofrece las siguientes ventajas:

Alto rendimiento de escaneo completo: Permite realizar consultas ad hoc rápidas en grandes volúmenes de datos sin necesidad de índices o estructuras de datos predefinidas.

Costo-efectividad: El precio de BigQuery es significativamente más bajo en comparación con las soluciones tradicionales de almacén de datos, lo que permite a las empresas ahorrar costos significativos.

Almacenamiento columnar: Utiliza un almacenamiento columnar que ofrece una alta eficiencia de compresión y rendimiento de E/S en comparación con el almacenamiento basado en filas.

Paralelismo masivo: Aprovecha la economía de escala de la plataforma en la nube de Google para lograr un alto rendimiento de E/S en consultas masivamente paralelas.

Fácil integración y administración: BigQuery es un servicio completamente administrado que no requiere planificación de capacidad, monitorización 24/7 ni actualizaciones manuales de seguridad. Además, ofrece una API REST que permite construir paneles de control y aplicaciones móviles de forma sencilla.

En resumen, BigQuery es un servicio de consultas de Google que permite ejecutar consultas SQL en grandes volúmenes de datos en segundos. Es altamente eficiente, escalable y rentable, proporcionando un rendimiento excepcional para consultas ad hoc de OLAP/BI. Es una solución de base de datos en la nube que supera a los sistemas tradicionales de almacén de datos y dispositivos especializados. BigQuery acelera el procesamiento de Big Data y permite a las empresas obtener resultados rápidos para el análisis de datos críticos.