An Inside Look at Google BigQuery

Cómo maneja Google las operaciones diarias del Big Data, servicios como Search, YouTube, Gmail y Google Drive, es demasiada la información que transita por estos medios cada día, la tecnología empleada para este trabajo es Dremel.

Dremel

Es un servicio de consultas que le permite ejecutar consultas similares a SQL contra conjuntos de datos muy, muy grandes y obtener resultados precisos en cuestión de segundos. Solo necesita un conocimiento básico de SQL para consultar conjuntos de datos extremadamente grandes de una manera ad hoc.

BigQuery: Externalización del Dremel

BigQuery es la implementación pública de Dremel que se lanzó recientemente a disponibilidad general. BigQuery proporciona el conjunto básico de características disponibles en Dremel a desarrolladores de terceros.

El Dremel tiene el *super poder* de una escalabilidad super alta y la mayoría de veces devuelve resultados en segundos o decenas de segundos sin importar cuán grande sea el conjunto de datos consultados.

Columnar Storage and Tree Architecture of Dremel

Columnar Storage

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 volumen.

Ventajas del almacenamiento columnar:

- Minimización del tráfico.
- Mayor relación de compresión.

Tree Architecture

La arquitectura forma un árbol distribuido masivamente paralelo para empujar una consulta al árbol y luego agregar los resultados de las hojas a una velocidad increíblemente rápida.

Dremel: Key to Run Business at “Google Speed”

Este ha sido utilizado por Google desde 2006 y ha estado evolucionando en los últimos 6 años y se ha incluido en las siguientes aplicaciones:

- Análisis de documentos web rastreados.
- Seguimiento de datos de instalación para aplicaciones en el mercado de Android.
- Informes de bloqueo para los productos de Google.
- Resultados de OCR de Google Books.
- Análisis de spam.
- Depuración de azulejos del mapa en Google Maps.
- Migración de tabletas en instancias Bigtable administradas.

- Resultados de las pruebas realizadas en el sistema de compilación distribuido de Google.
- Estadísticas de E/S de disco para cientos de miles de discos.
- Monitoreo de recursos para trabajos ejecutados en los centros de datos de Google.
- Símbolos y dependencias en la base de códigos de Google.

BigQuery versus MapReduce

Estas son las diferencias entre las dos:

- Dremel está diseñado como una herramienta interactiva de análisis de datos para grandes conjuntos de datos.
- MapReduce está diseñado como un marco de programación para procesar por lotes grandes conjuntos de datos.

Comparing BigQuery and MapReduce

MapReduce es una tecnología de computación distribuida que le permite implementar funciones personalizadas de "mapeador" y "reductor" programáticamente y ejecutar procesos por lotes con ellos en cientos o miles de servidores simultáneamente.

Cuando utilizar BigQuery o MapReduce según estos criterios: Usar BigQuery

- Encontrar registros particulares con condiciones especificadas.
- Rápida agregación de estadísticas con condiciones que cambian dinámicamente.
- Análisis de datos de ensayo y error.

Usar MapReduce

- Ejecución de una minería de datos compleja en Big Data que requiere múltiples iteraciones y rutas de procesamiento de datos con algoritmos programados.
- Ejecución de operaciones de unión grandes en conjuntos de datos enormes.
- Exportación de gran cantidad de datos después del procesamiento.
- Utilice MapReduce para grandes operaciones de combinación y conversiones de datos, luego utilice BigQuery para una rápida agregación y análisis de datos ad-hoc en el conjunto de datos de resultados.
- Utilice BigQuery para una comprobación previa mediante un análisis rápido de datos, luego escriba y ejecute el código MapReduce para ejecutar un procesamiento de datos de producción o minería de datos.

Data Warehouse Solutions and Appliances for OLAP/BI

Muchas empresas han estado utilizando soluciones de almacenamiento de datos o dispositivos para sus casos de uso de OLAP/ BI durante muchos años. En OLAP/BI, aproximadamente tiene las siguientes tres alternativas para aumentar el rendimiento del manejo de Big Data:

1. OLAP relacional (ROLAP)
2. OLAP multidimensional (MOLAP)
3. Análisis completo