Adriel Araya Vargas 2019312845 Instituto Tecnológico de Costa Rica Resumen 7 Bases de Datos 2 IC-4302

Google BigQuery

Este documento blanco presenta Google BigQuery, un servicio de consultas interactivas basado en la nube y totalmente administrado para conjuntos de datos masivos. BigQuery es la implementación externa de una de las tecnologías centrales de la compañía, llamada Dremel. Este documento analiza la singularidad de la tecnología como un motor de consultas masivamente paralelo habilitado para la nube, las diferencias entre BigQuery y Dremel, y cómo BigQuery se compara con otras tecnologías como MapReduce/Hadoop y soluciones de almacén de datos existentes.

Dremel es un servicio de consultas que te permite ejecutar consultas similares a SQL en conjuntos de datos muy, muy grandes y obtener resultados precisos en cuestión de segundos. Solo necesitas un conocimiento básico de SQL para consultar conjuntos de datos extremadamente grandes de manera ad hoc. En Google, tanto los ingenieros como los no ingenieros, incluidos los analistas, el personal de soporte técnico y los gerentes de cuentas técnicas, utilizan esta tecnología muchas veces al día. BigQuery es la implementación pública de Dremel y proporciona las mismas características principales de Dremel a desarrolladores externos a través de una API REST, una interfaz de línea de comandos, una interfaz web, control de acceso y más, manteniendo el rendimiento de consultas sin precedentes de Dremel.

Dremel es capaz de escanear 35 mil millones de filas sin un índice en cuestión de segundos. Utiliza la infraestructura de Google para paralelizar cada consulta y ejecutarla en decenas de miles de servidores simultáneamente. Esto permite lograr un rendimiento de consultas extremadamente rápido con una relación costo-valor muy atractiva. Además, no se requiere inversión de capital por parte del usuario para la infraestructura de soporte. Dremel almacena datos en un formato de almacenamiento columnar, lo que permite obtener una alta relación de compresión y rendimiento de escaneo. También utiliza una arquitectura de árbol para enviar consultas y agregar resultados en segundos en miles de máquinas. Dremel aprovecha estas tecnologías para lograr un rendimiento sin precedentes y es entregado como un servicio en la nube.

BigQuery es un servicio de consultas para conjuntos de datos grandes que se compara con tecnologías existentes de Big Data como MapReduce y soluciones de almacenamiento de datos. Mientras que MapReduce es un marco de programación diseñado para procesar lotes de datos grandes, BigQuery es una herramienta de análisis de datos interactiva diseñada para finalizar la mayoría de las consultas en segundos o unos pocos segundos. A diferencia de MapReduce, BigQuery puede ser utilizado por usuarios no programadores y es más adecuado para consultas ad hoc y análisis rápido de datos.

MapReduce es una tecnología de cómputo distribuido que permite implementar funciones "mapper" y "reducer" personalizadas y ejecutar procesos por lotes en cientos o miles de servidores de forma concurrente. Es adecuado para aplicaciones como el análisis de registros, análisis de actividad de usuarios, minería de datos y procesamiento de datos no estructurados. Sin embargo, MapReduce es más lento que BigQuery, ya que puede llevar minutos, horas o incluso días completar el procesamiento de una consulta.

En comparación, BigQuery ofrece respuestas rápidas, especialmente para consultas interactivas y ad hoc. Es fácil de usar para usuarios no programadores y es adecuado para casos de uso como OLAP/BI (procesamiento analítico en línea/business

intelligence) y análisis de datos en tiempo real. BigQuery también es capaz de manejar consultas de unión y actualizaciones de datos, mientras que MapReduce se centra en procesamiento por lotes y no es adecuado para análisis de datos ad hoc.

En cuanto a las soluciones de almacén de datos y las aplicaciones de OLAP/BI, BigQuery ofrece ventajas significativas en términos de costos y rendimiento. A diferencia de las soluciones tradicionales basadas en bases de datos relacionales o multidimensionales, BigQuery no requiere construir índices previos a la consulta y proporciona un rendimiento de escaneo completo más rápido debido a su capacidad de paralelismo masivo y optimización de almacenamiento columnar.

La importación de datos a BigQuery se realiza en dos pasos: primero, se cargan los datos en Google Cloud Storage y luego se importan a BigQuery utilizando herramientas de línea de comandos, interfaz web o API. El uso de Google Cloud Platform y BigQuery ofrece ventajas como costos reducidos, escalabilidad, seguridad administrada y facilidad de integración con otras herramientas y servicios de Google.