Resumen 8

Edgar André Araya Vargas 2020142856

Bigtable: A Distributed Storage System for Structured Data

Bigtable, desarrollado por Google, es un sistema de almacenamiento distribuido diseñado específicamente para gestionar grandes volúmenes de datos estructurados. Destaca por su escalabilidad, alto rendimiento y disponibilidad. Su amplia implementación abarca más de sesenta productos y proyectos de Google, entre los que se incluyen Google Analytics, Google Finance y Google Earth. A diferencia de los sistemas de base de datos tradicionales, Bigtable presenta un modelo de datos sencillo que otorga a los clientes el control total sobre la organización y el formato de los datos almacenados. Estos datos se indexan mediante nombres de filas y columnas, y se tratan como cadenas sin interpretar. Una de las características destacadas de Bigtable es su capacidad para almacenar múltiples versiones de los datos, las cuales se indexan según su marca de tiempo. El sistema se basa en unidades llamadas "tabletas", que permiten una distribución y un equilibrio de carga eficientes. Además, Bigtable se utiliza en diversas cargas de trabajo, desde el procesamiento por lotes hasta la entrega de datos a los usuarios finales. A lo largo del tiempo, el sistema ha sido refinado para mejorar su rendimiento, y se han extraído lecciones valiosas de su diseño y soporte. En resumen, Bigtable ha demostrado ser una solución eficiente y versátil para el manejo de grandes volúmenes de datos estructurados dentro del entorno de Google.

Su API ofrece funciones para crear, eliminar y modificar tablas y familias de columnas, así como para controlar los metadatos. Los clientes pueden escribir, buscar y procesar datos en Bigtable utilizando abstracciones como RowMutation y Scanner.

Bigtable utiliza el formato de archivo interno Google SSTable para almacenar los datos de manera persistente y ordenada. Además, integra el sistema de archivos distribuido de Google (GFS) para el almacenamiento de archivos y depende de Chubby, un servicio de bloqueo distribuido, para garantizar la consistencia y disponibilidad. Chubby utiliza el algoritmo de Paxos y desempeña diversas funciones, como la asignación de un maestro, el almacenamiento de información de esquema y control de acceso, y el descubrimiento de servidores de tabletas.

Entre las características adicionales de Bigtable se encuentran las transacciones a nivel de fila, los contadores de enteros en las celdas y la capacidad de ejecutar scripts en los servidores. Bigtable se ha demostrado eficiente y confiable, con un porcentaje muy bajo de tiempo de indisponibilidad debido a problemas con Chubby, según mediciones realizadas en múltiples clústeres.

Este servicio de Google se compone de tres elementos principales: una biblioteca que se enlaza con cada cliente, un servidor principal y múltiples servidores de tabletas. Los servidores de tabletas pueden ser agregados o eliminados de forma dinámica en un clúster para adaptarse a los cambios en las cargas de trabajo. El servidor principal se encarga de asignar tabletas a los servidores de tabletas, detectar la adición y eliminación de servidores de tabletas, equilibrar la carga de las tabletas y gestionar la limpieza de archivos en el sistema de archivos GFS. Además, el servidor principal maneja cambios en el esquema, como la creación de tablas y familias de columnas.

Cada servidor de tableta administra un conjunto de tabletas. El servidor de tableta se encarga de procesar las solicitudes de lectura y escritura a las tabletas que ha cargado, y también divide las tabletas que han crecido demasiado en tamaño. Los clientes de Bigtable se comunican directamente con los servidores de tabletas para

leer y escribir datos, sin depender del servidor principal para obtener información acerca de la ubicación de las tabletas.

La ubicación de las tabletas se almacena en una jerarquía de tres niveles, similar a un árbol B+. La información de ubicación se guarda en Chubby, el servicio de nombres distribuido utilizado internamente por Google mencionado anteriormente. Cada clúster de Bigtable puede almacenar múltiples tablasy cada tabla está compuesta por un conjunto de tabletas, y cada tableta contiene todos los datos asociados con un rango de filas. A medida que una tabla crece, se divide automáticamente en múltiples tabletas y la información de ubicación de las tabletas se almacena en memoria y se guarda en caché en las bibliotecas de cliente para reducir los costos de acceso.

La asignación de tabletas se lleva a cabo mediante el servidor principal. El servidor principal realiza un seguimiento de los servidores de tabletas activos y la asignación actual de tabletas a servidores de tabletas. Cuando una tableta no está asignada y hay un servidor de tableta disponible con suficiente espacio, el servidor principal asigna la tableta enviando una solicitud de carga al servidor de tableta correspondiente.

El estado persistente de una tableta se almacena en GFS. Los servidores de tabletas leen los metadatos de las tabletas desde una tabla de metadatos que contiene la ubicación de las SSTables (Sorted String Table) y puntos de registro que apuntan a registros de confirmación en un registro de operaciones. Las operaciones de escritura se registran en el registro de operaciones, mientras que las operaciones de lectura se ejecutan en una vista combinada de las SSTables y la tabla en memoria (memtable).

A medida que se realizan operaciones de escritura, el tamaño de la memtable aumenta. Cuando la memtable alcanza un umbral determinado, se congela y se convierte en una SSTable, lo que reduce el uso de memoria y la cantidad de datos que deben leerse del registro de operaciones durante la recuperación. Además, se llevan a cabo compactaciones menores y mayores para fusionar las SSTables y mantener un número manejable de archivos. Las compactaciones mayores permiten que Bigtable recupere recursos utilizados por los datos almacenados.

El documento proporciona información sobre la evaluación de rendimiento de Bigtable de Google. A continuación, se presenta un resumen de la sección mencionada:

En un experimento realizado se observó que las lecturas aleatorias eran más lentas que las demás operaciones debido a la transferencia de bloques de SSTable de 64 KB desde GFS a los servidores de tabletas. Las escrituras aleatorias y secuenciales tenían un rendimiento mejorado, ya que los servidores de tabletas registraban las escrituras en un registro de confirmación y utilizaban la confirmación de grupo para transmitir eficientemente las escrituras a GFS. Las lecturas secuenciales y las exploraciones tenían un rendimiento aún mejor, ya que los servidores de tabletas podían devolver una gran cantidad de valores en respuesta a una única solicitud del cliente. A medida que se incrementaba el número de servidores de tabletas, el rendimiento agregado aumentaba, pero no de manera lineal debido a la carga desequilibrada y las limitaciones de ancho de banda. Todo esto demostrando superioridad y utilidad en este sistema distribuido de almacenamiento.

En la sección final del documento, se mencionan tres casos de uso de Bigtable en productos de Google: Google Analytics, Google Earth y Personalized Search. Estos productos utilizan Bigtable para almacenar y procesar datos a gran escala, adaptados a sus respectivos requisitos y patrones de acceso.