Big Data
BigTable

Texto base: Juan Esquivel Rodríguez
Profesor: Luis Alexánder Calvo Valverde

Bigtable

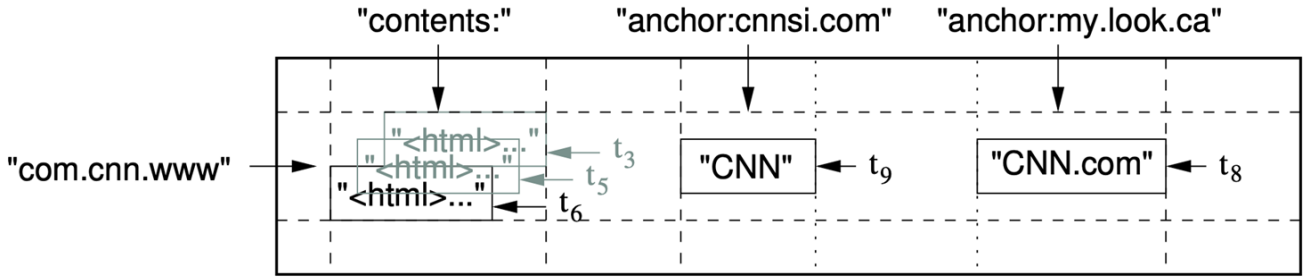
- Híbrido entre una base de datos y un sistema de archivos
- Objetivo primario
  - Almacenar datos estructurados (i.e. objetos)
  - Escalabilidad a petabytes de información
  - Acceso rápido y confiable
  - Distribuido en miles de máquinas
- Componente de modelado
  - Cómo se estructuran los datos
- Implementación
  - Altamente acoplado para cumplir objetivos.

Bigtable: Modelo

- Bigtable es un mapa multidimensional, persistente y distribuído
- Como cualquier tabla estándar posee filas con celdas
  - Tipo string (bytes sin interpretar)
  - Referenciadas por (fila, columna), de tipo string también
- Se deja al usuario la responsabilidad de hacer las conversiones necesarias al insertar o extraer datos.
- Versiones
  - Cada escritura en celda se asocia a estampilla de tiempo.
- Cada elemento está dado por un par llave / valor
  - Llave: (fila:string, columna:string, tiempo:int64)
  - Valor: string

Bigtable: Ejemplo páginas Web

- Indexado por URL de las páginas
  - Se convierte en el identificador de la fila
- Información sobre la página
  - Almacenada en diferentes columnas, dependiendo del tipo de dato y uso.
- Contenido de la página con versiones
  - Reflejando el punto en el tiempo cuando se obtuvieron



Bigtable: Filas

- Tienen un tamaño máximo de 64KB
  - Uso común es entre 10 y 100 bytes.
- Cualquier lectura o escritura sobre una fila es atómica
- Las filas son almacenadas agrupadas
  - Basado en orden lexicográfico
  - Lecturas de rangos eficientes
- Se espera que los clientes aprovechen esta característica al diseñar sus llaves.
  - E.g. Revertir los URLs de páginas permite agruparlas, ya que los prefijos son iguales y contiguos lexicográficamente

Bigtable: Columnas y familias

- Las columnas se agrupan en familias.
  - Control de acceso
  - Contabilidad de memoria
  - Compresión de datos
- Deben crearse familias en el orden las centenas
  - Cantidad de columnas de cada familia pueda ser muy amplia
- Nombre de la familia aparece como prefijo de la columna
  - El nombre de una columna siempre será familia:calificador

Bigtable: Versiones

- Las versiones son manejadas no por consecutivos sino por tiempo
- La asignación de tiempo puede ser hecha por Bigtable o manualmente.
- Versiones más recientes son las leídas primero.
- Se puede activar recolección de basura automática
  - Últimas N versiones
  - Respecto a una fecha (e.g. cantidad de días atrás)

Bigtable: Almacenamiento

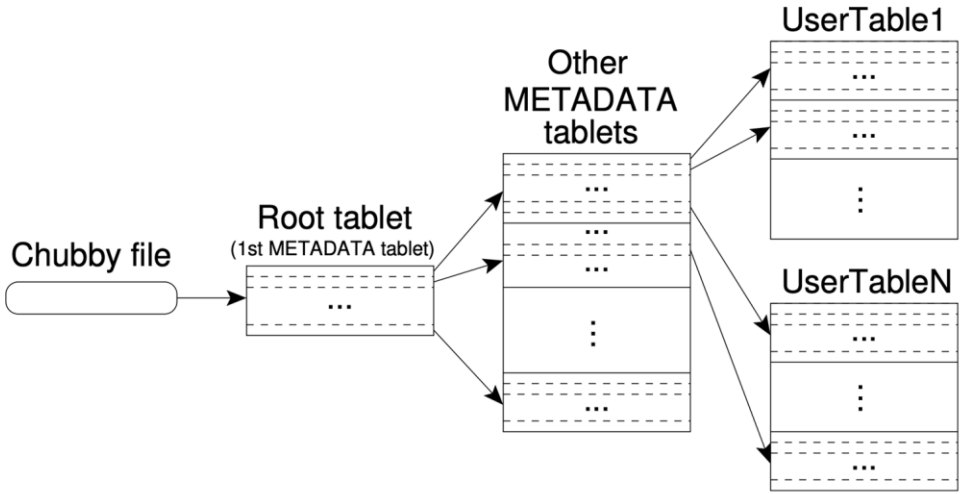
- Capa de almacenamiento de datos es GFS
- SSTables
  - Formato de archivos
  - Cada SSTable es un mapa inmutable ordenado de pares llave/valor
  - Cada tabla contiene bloques (de 64KB por defecto)
  - Se referencia por índices almacenados al final del archivo
  - Pueden ser mapeadas directamente a memoria para mayor rendimiento
- La búsqueda de regiones de objetos se hace mediante búsqueda binaria
  - Como las SSTables tienen los índices inmutables al final, es realizable

Bigtable: Servidores

- Opera con dos tipos de servidores
  - Coordinadores (masters)
  - De datos (tablet servers)
- Cada servidor maneja entre 10 y 1000 tablets con información
  - Particiona cuando el tamaño lo amerita (cada 100-200MB por defecto)
- Funciones de coordinadores
  - Asignar tablets a cada tablet server
  - Balancear la carga
  - Recolección de basura
  - Manejar la estructura de familias de columnas
  - Administrar la creación de tablet servers adicionales
- La comunicación de los clientes con el maestro no es muy frecuente
  - Los clientes se comunican directo con los servidores de tablets.
- Cada cluster de Bigtable puede almacenar múltiples Bigtables

Bigtable: Jerarquía de tablets

- Análoga a un árbol B que
- Por diseño, nunca será de más de 3 niveles
- Escenario modesto de configuración
  - Cada tabla de metadatos tiene un límite de 128MB
  - Es posible direccionar 2^{61} bytes



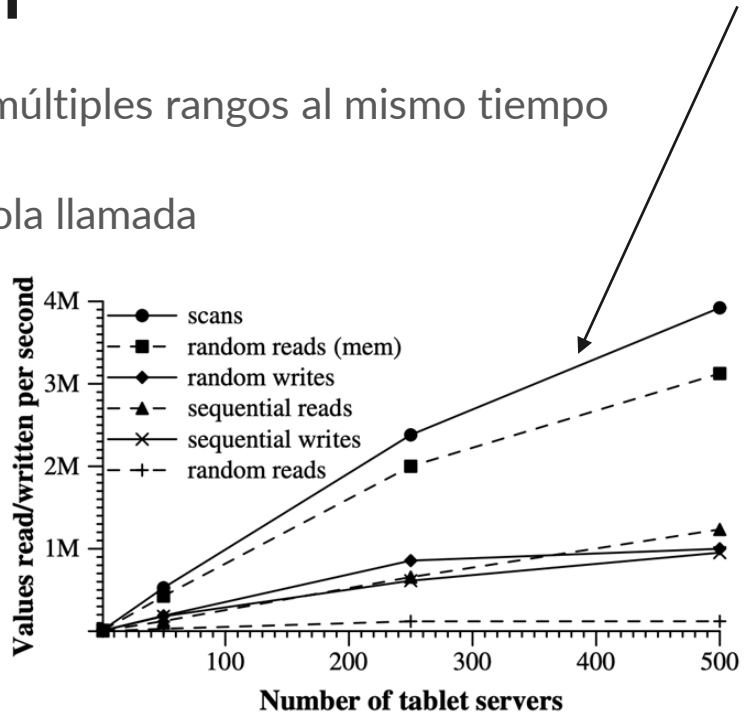
Bigtable: Rendimiento

- Benchmark con N tablet servers
  - 1 GB de memoria
  - Escribían a una celda GFS con 1786 máquinas
  - N clientes generando carga de lectura sobre Bigtable
- Se define R como el número de filas en cada prueba
  - Seleccionada para que cada una escribiera alrededor de 1GB de datos por servidor de tablets.

Bigtable: Rendimiento Scan

- Utilizaba las funcionalidades para obtener múltiples rangos al mismo tiempo
  - Basado en nombre de la fila
- Esto permitía obtener más datos con una sola llamada

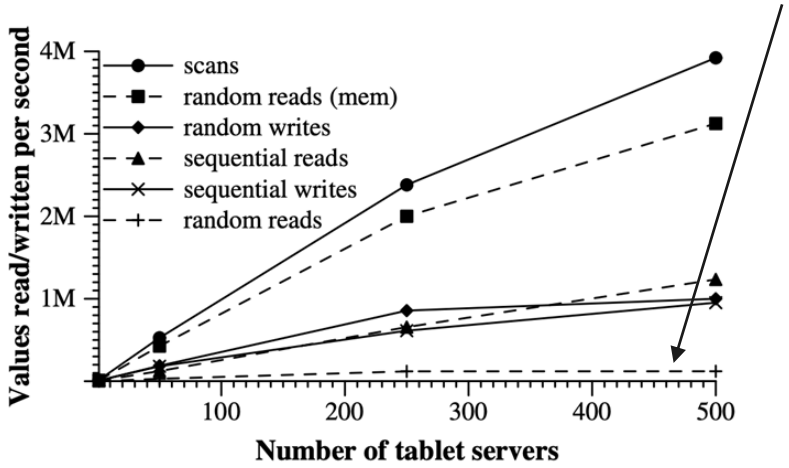
Experiment	# of Tablet Servers			
	1	50	250	500
random reads	1212	593	479	241
random reads (mem)	10811	8511	8000	6250
random writes	8850	3745	3425	2000
sequential reads	4425	2463	2625	2469
sequential writes	8547	3623	2451	1905
scans	15385	10526	9524	7843



Bigtable: Rendimiento Random Reads

- Número de fila generado entre 0 y R-1
- Se aplica el módulo a una función de hash, para tratar de simular una distribución uniforme

Experiment	# of Tablet Servers			
	1	50	250	500
random reads	1212	593	479	241
random reads (mem)	10811	8511	8000	6250
random writes	8850	3745	3425	2000
sequential reads	4425	2463	2625	2469
sequential writes	8547	3623	2451	1905
scans	15385	10526	9524	7843



Bigtable: Configuraciones ejemplo

Project name	Table size (TB)	Compression ratio	# Cells (billions)	# Column Families	# Locality Groups	% in memory	Latency-sensitive?
Crawl	800	11%	1000	16	8	0%	No
Crawl	50	33%	200	2	2	0%	No
Google Analytics	20	29%	10	1	1	0%	Yes
Google Analytics	200	14%	80	1	1	0%	Yes
Google Base	2	31%	10	29	3	15%	Yes
Google Earth	0.5	64%	8	7	2	33%	Yes
Google Earth	70	–	9	8	3	0%	No
Orkut	9	–	0.9	8	5	1%	Yes
Personalized Search	4	47%	6	93	11	5%	Yes

Referencias

- Ghemawat, S; Gobioff, H; Leung, S. The Google File System. https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
- Chang, F; Dean, J; Ghemawat, S; Hsieh, W; Wallach, D; Burrows, M; Chandra, T; Fikes, A; Gruber, R. Bigtable: A Distributed Storage System for Structured Data. https://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf