Bigtable

David Wyrobnik, MEng

Overview

- What is Bigtable?
- Data Model
- API
- Building Blocks
- Implementation

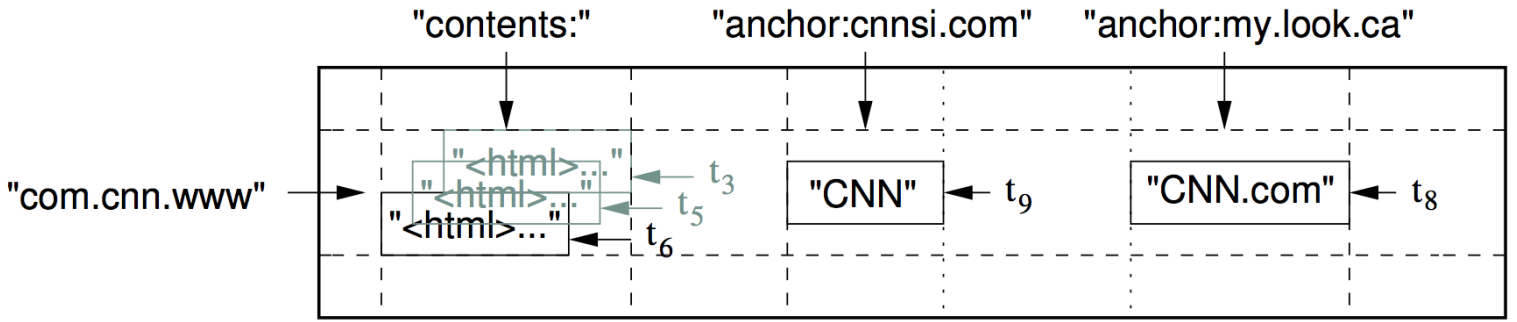
What is Bigtable (high level)

- “Distributed storage system for structured data” - title of paper
- “BigTable is a compressed, high performance, and proprietary data storage system built on Google File System, Chubby Lock Service, SSTable (log-structured storage like LevelDB) and a few other Google technologies.” - wikipedia
- “A Bigtable is a sparse, distributed, persistent multidimensional sorted map” - paper

Data Model

Data Model

- (row:string, column:string, time:int64) ➔ array of bytes



Data Model continued

- Timestamps can be assigned automatically (“real time”) or by client
- Versioned data management, two per-column-family settings for garbage-collection
  - last *n* versions of a cell should be kept
  - only new-enough versions kept (e.g. only values that were written in the last seven days)

API

API

- Functions for creating and deleting
  - tables and column families
- Functions for changing
  - clusters, table, and column family metadata (such as control rights)
- Write, delete, and lookup values in individual rows
- Iterate over subset of data in table
- Single-row transactions ➔ perform atomic read-modify-write sequences
- No general transactions across rows, but supports batching writes across rows
- Bigtable can be used with MapReduce (common use case)

Building Blocks and Implementation

Building Blocks

- Google-File-System (GFS) to store log and data files.
- SSTable file format.
- Chubby as a lock service
- Bigtable uses Chubby
  - to ensure at most one active master exists
  - to store bootstrap location of Bigtable data
  - to discover tablet servers
  - to store Bigtable schema information (column family info for each table)
  - to store access control lists

Implementation

- Three major components:
  - library that is linked into every client
  - one master server
  - many tablet servers
- Master mainly responsible for assigning tablets to tablet servers
- Tablet servers can be added or removed dynamically
- Tablet server store typically 10-1000 tablets
- Tablet server handle read and writes and splitting of tablets that are too large
- Client data does not move through master.

Tablet Location

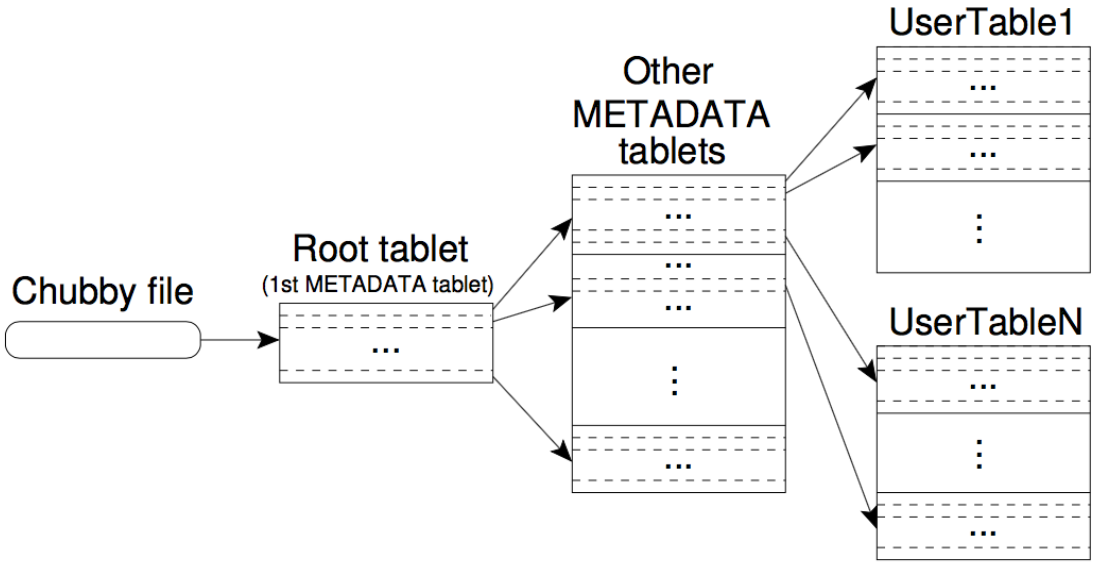
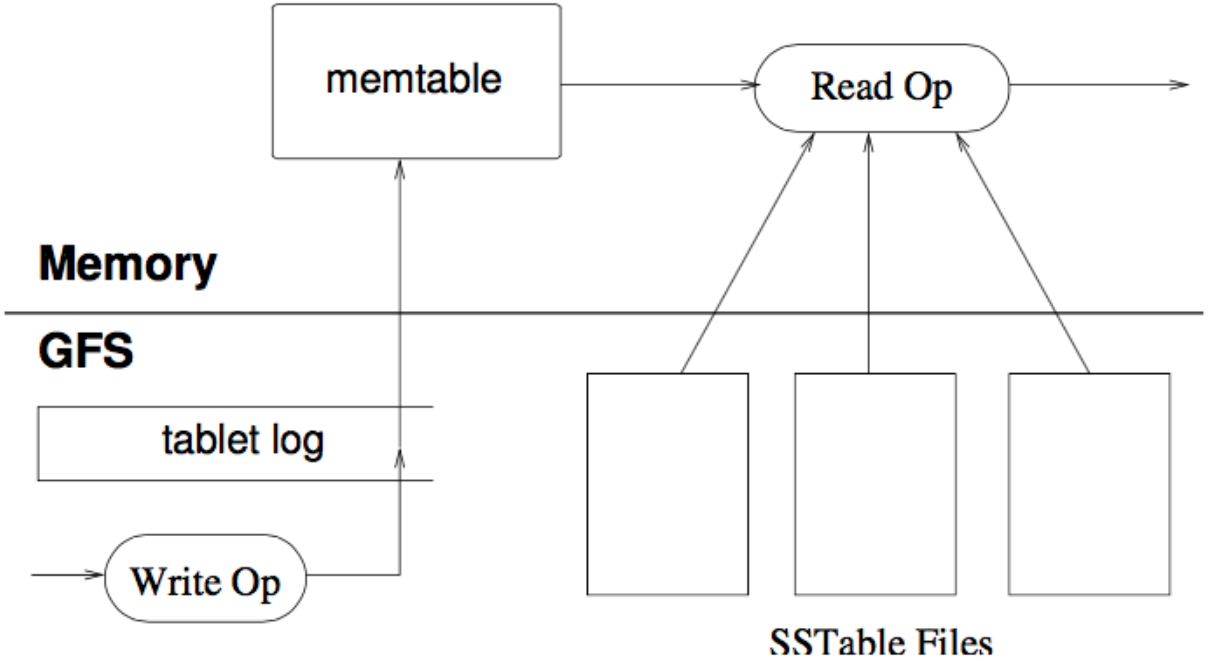


Figure 4: Tablet location hierarchy.

Tablet Assignment

- Master keeps track of live tablet servers, current assignments, and of unassigned tablets
- Master assigns unassigned tablets to tablet servers by sending a tablet load request
- Tablet servers are linked to files in Chubby directory (servers directory)
- When new master starts:
  - Acquires unique master lock in Chubby
  - Scans live tablet servers
  - Gets list of tablets from each tablet server, to learn which tablets are assigned
  - Scans METADATA table to learn set of existing tablets ➜ adds unassigned tablets to list

Tablet Serving



Consistency

- Bigtable has a strong consistency model, since operations on rows are atomic and tablets are only served by one tablet server at a time

Discussion