SSTable

Let's learn how Tablets are stored in SSTables.

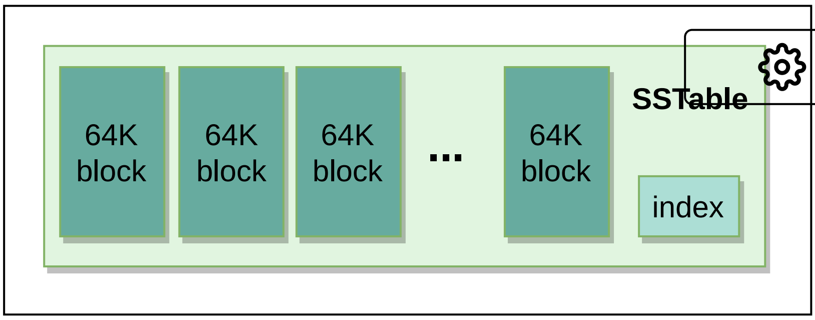
We'll cover the following

- How are Tablets stored in GFS?
  - Table vs. Tablet vs. SSTable

How are Tablets stored in GFS?#

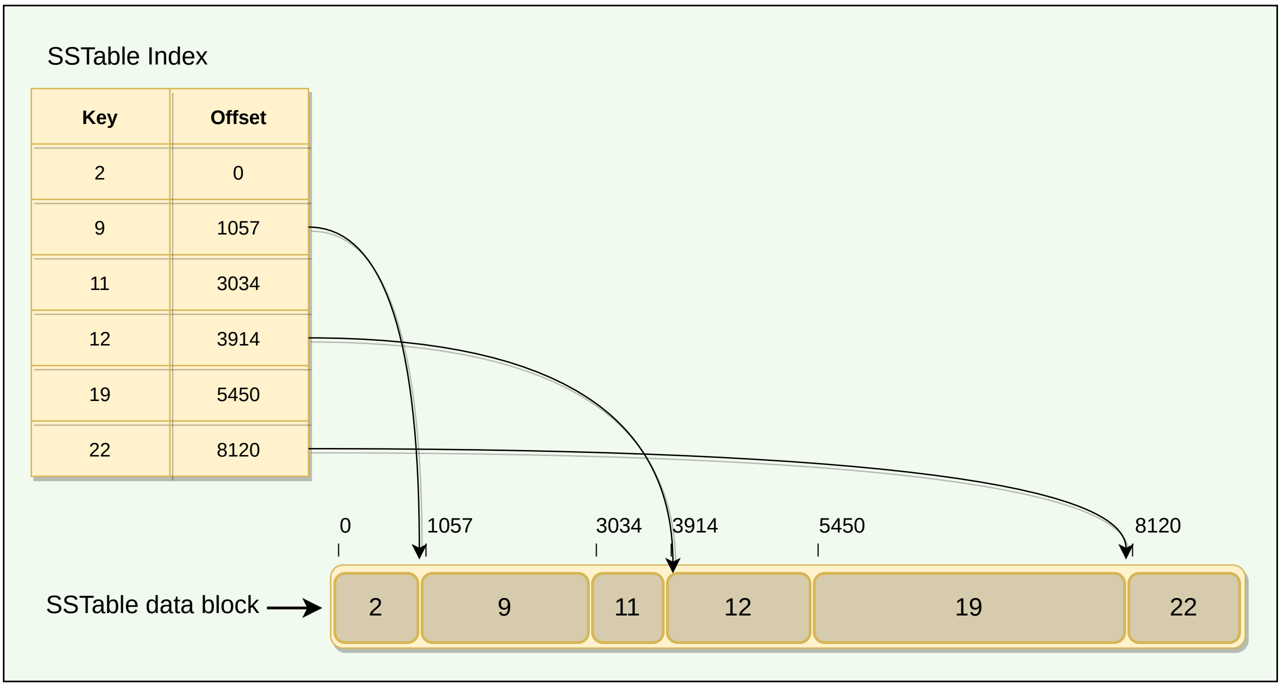
BigTable uses Google File System (GFS), a persistent distributed file storage system to store data as files. The file format used by BigTable to store its files is called SSTable:

- SSTables are persisted, ordered maps of keys to values, where both keys and values are arbitrary byte strings.
- Each Tablet is stored in GFS as a sequence of files called SSTables.
- An SSTable consists of a sequence of data blocks (typically 64KB in size).



SSTable contains multiple blocks

- A block index is used to locate blocks; the index is loaded into memory when the SSTable is opened.


SSTable Index

Key	Offset
2	0
9	1057
11	3034
12	3914
19	5450
22	8120

Reading data from SSTable

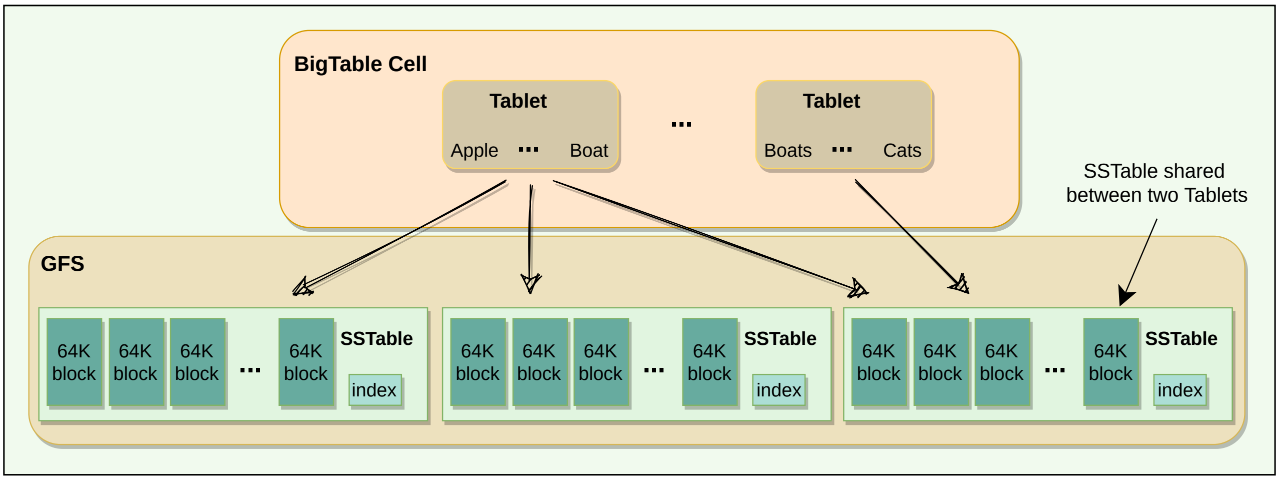
- A lookup can be performed with a single disk seek. We first find the appropriate block by performing a binary search in the in-memory index, and then reading the appropriate block from the disk.
- To read data from an SSTable, it can either be copied from disk to memory as a whole or just the index. The former approach avoids subsequent disk seeks for lookups, while the latter requires a single disk seek for each lookup.
- SSTables provide two operations:

  - Get the value associated with a given key
  - Iterate over a set of values in a given key range
- Each SSTable is immutable (read-only) once written to GFS. If new data is added, a new SSTable is created. Once an old SSTable is no longer needed, it is set out for garbage collection. SSTable immutability is at the core of BigTable's data checkpointing and recovery routines. SSTable's immutability provides following advantages:
  - No synchronization is needed during read operations.
  - This also makes it easier to split Tablets.
  - Garbage collector handles the permanent removal of deleted or stale data.

Table vs. Tablet vs. SSTable#

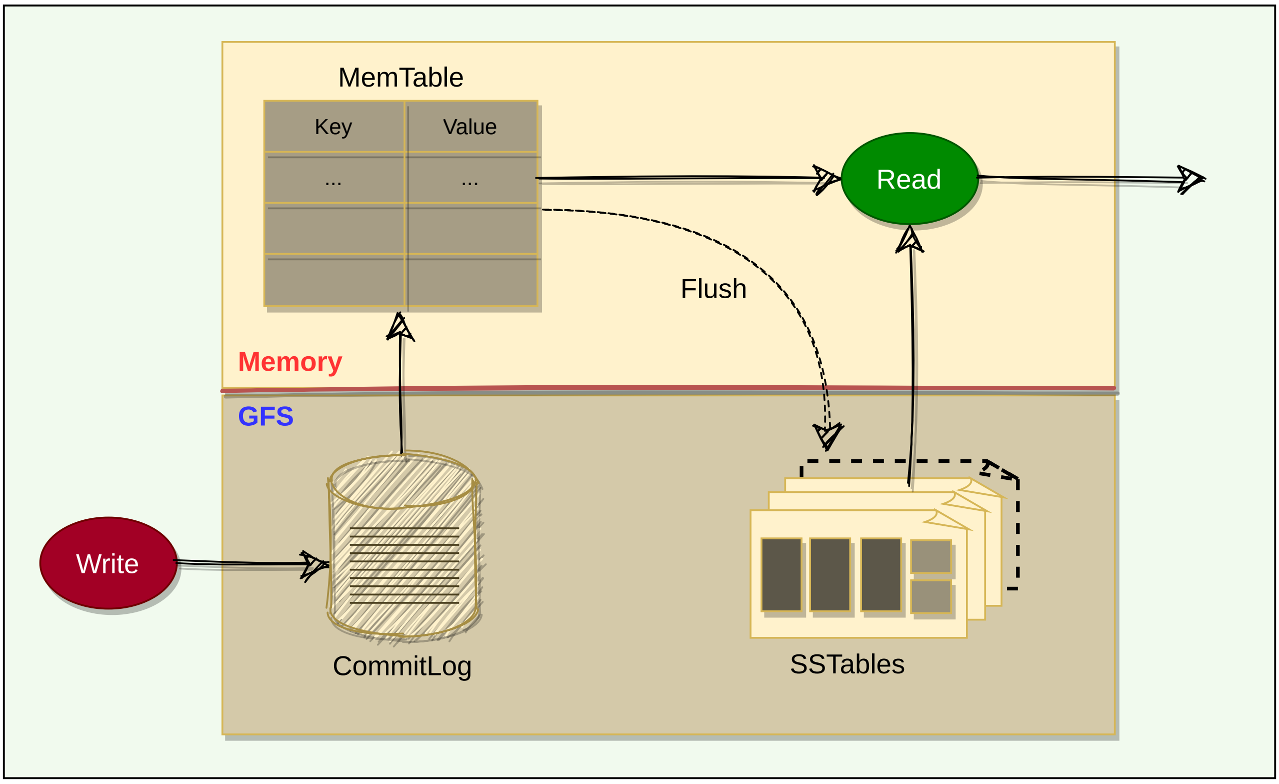
Here is how we can define the relationship between Table, Tablet and SStable:

- Multiple Tablets make up a table.
- SSTables can be shared by multiple Tablets.
- Tablets do not overlap, SSTables can overlap.



Tablet vs SSTable

- To improve write performance, BigTable uses an in-memory, mutable sorted buffer called **MemTable** to store recent updates. As more writes are performed, MemTable size increases, and when it reaches a threshold, the MemTable is frozen, a new MemTable is created, and the frozen MemTable is converted to an SSTable and written to GFS.
- Each data update is also written to a commit-log which is also stored in GFS. This log contains redo records used for recovery if a Tablet server fails before committing a MemTable to SSTable.
- While reading, the data can be in MemTables or SSTables. Since both these tables are sorted, it is easy to find the most recent data.



Read and write workflow

← Back

Partitioning and High-level Architecture

Next →

GFS and Chubby

☑ Mark as Completed

Report an Issue