Dreifð gagnasöfn

Gagnasafnsfræði 2011

Hallgrímur H. Gunnarsson

Háskóli Íslands

2011-11-17

Stór gagnasöfn

Til íhugunar

- Hvað er til ráða þegar gagnasafn rúmast ekki lengur fyrir á einni vél eða þegar ein vél ræður ekki við útreikninga?
- Hvaða afleiðingar hefur það í för með sér að skipta gagnasafni á margar vélar?
- Hvaða eiginleika hreyfinga er hægt að tryggja þegar gagnasafni er skipt á margar vélar?
- Hvernig getum við beislað afl úr mörg hundruð eða þúsund vélum?
- Hvernig er hægt að byggja áreiðanlegt dreift kerfi ofan á vélbúnað sem getur (og mun á endanum) hrynja?

Google File System

GFS

Notkun GFS hjá Google:

- 200+ clusters
- Mörg þúsund vélar í hverjum cluster
- Mörg þúsund biðlarar að nota hvern cluster
- 4+ PB skráarkerfi
- 40 GB/s I/O

Google File System

GFS

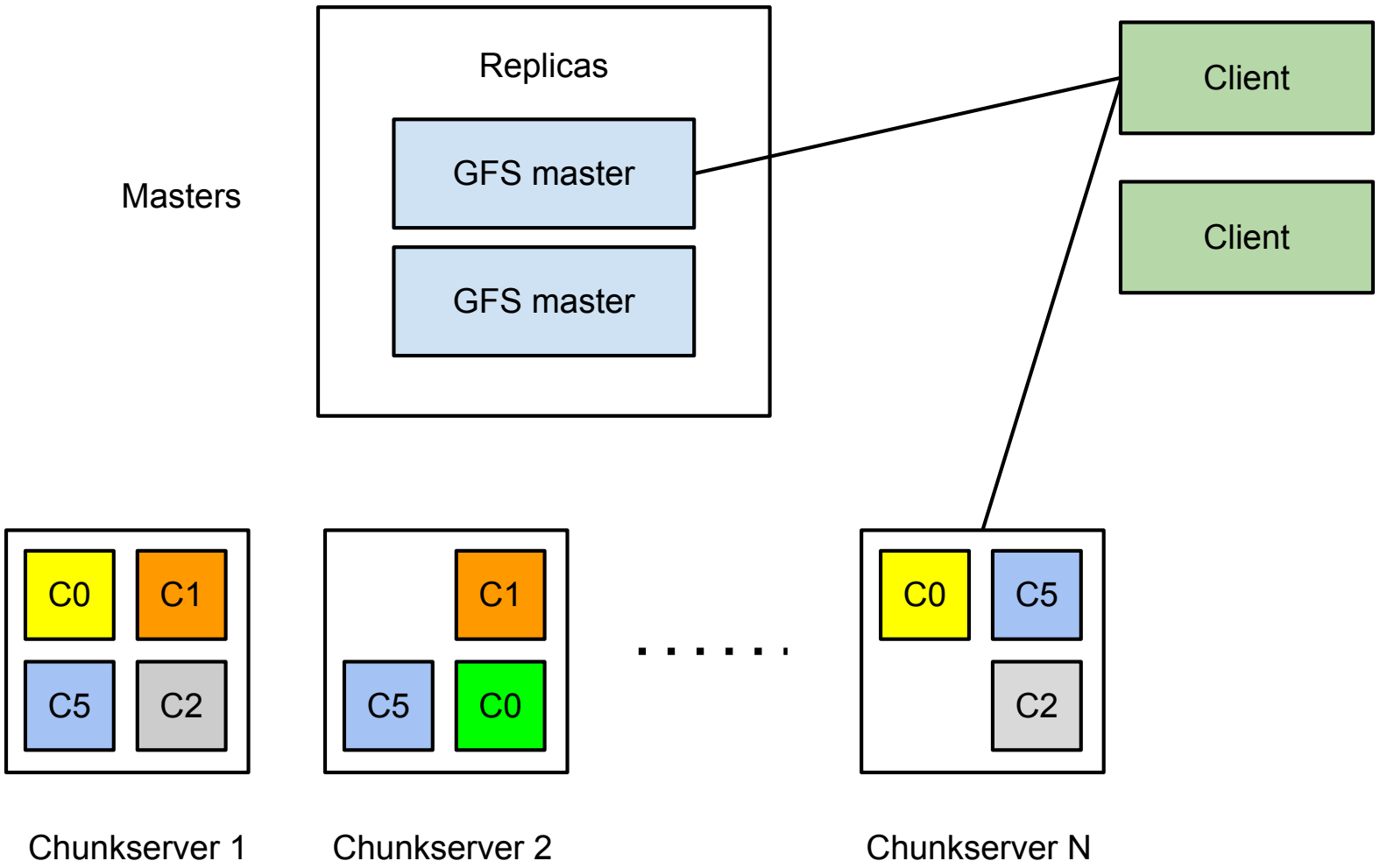
Hönnun:

- Skrár brotnar upp í 64 MB blokkir, s.k. "chunks"
- Biðlari (client) flettir upp í master þjóni, tengist svo gagnaþjóni beint (millilíðalaust) og sækir gögn
- Master þjónn geymir metadata upplýsingar

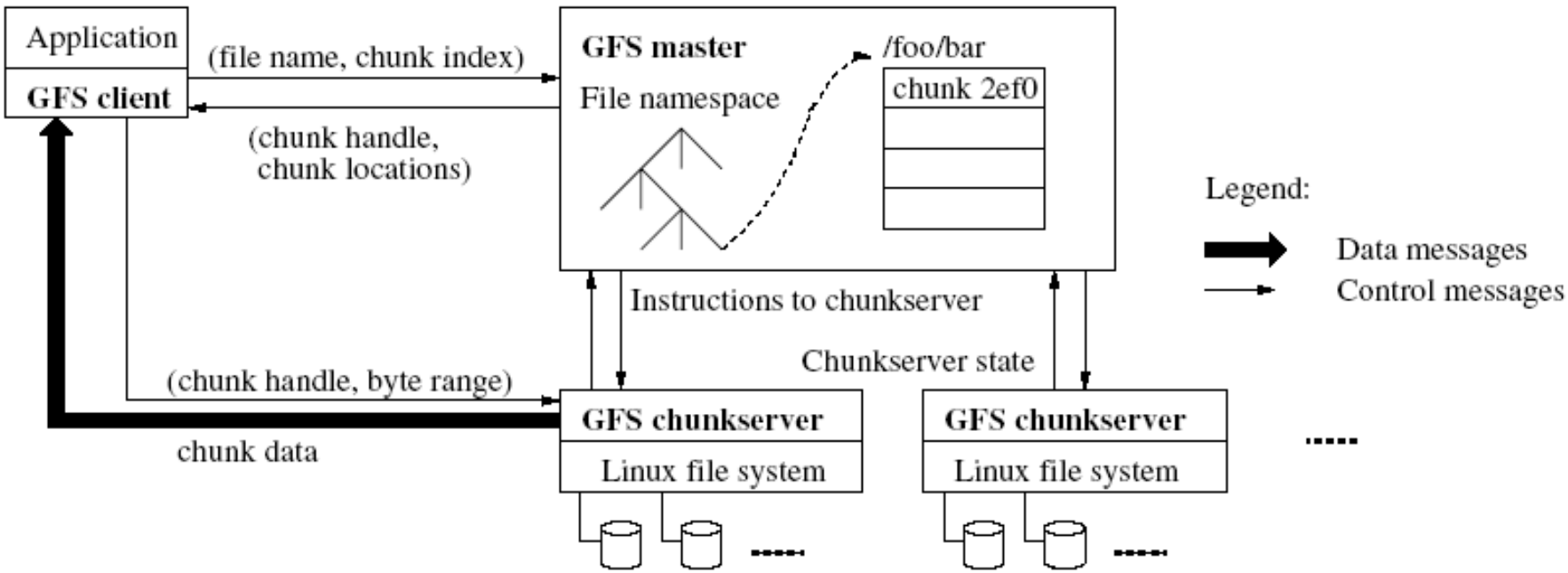
Metadata upplýsingar í master:

- Nafnatré skráarkerfisins (möppustrúktúr o.s.frv.)
- inode fyrir hverja skrá sem inniheldur lista yfir chunks
- Vörpun: (filename, position) $\rightarrow$ (chunk ID, chunk locations)

Google File System



Google File System



BigTable

Semi-structured gögn

Google vinnur með mikið af semi-structured gögnum:

- URLs: vefsíður, hlekkir, pagerank, crawl metadata
- Notendagögn: notendastillingar, search history
- Kortagögn: staðir (búðir, o.s.frv.), götur, gervihnattamyndir

Gagnamagn

Mikið magn af gögnum:

- Milljarðar vefsíðna, margar útgáfur, um 20 KB per útgáfa
- Hundruðir milljóna notendur, þúsundir fyrirspurna per sek
- 100TB+ af gervihnattamyndum

BigTable

BigTable

Sérsmíðaður gagnagrunnur sem Google notar fyrir sínar eigin þjónustur

- Yfir 100 verkefni hjá Google nota BigTable
- Um 500 BigTable clusters í gangi út um allan heim, hver cluster hefur mismunandi hlutverk
- Stærsti clusterinn: 70+ PB, 10M ops/sek, 30 GB/s I/O

Þjónustur

Þjónustur sem nota BigTable hjá Google eru m.a.

- Google Maps og Google Earth
- YouTube og Gmail
- Crawling/indexing pipeline

BigTable

Gagnalíkan

Gögn eru geymd í töflum þar sem:

- Tafla samanstendur af röðum. Röð hefur streng sem lykil.
- Röð inniheldur dálka. Dálkur hefur tvíþættan lykil (lýst síðar)
- Dálkur samanstendur af tímasettum gildum.
- Gildi er blob (ótúlkað fylki af bætum)

Vörpun

Töflunni mætti lýsa sem vörpun:

$$(row: string, column: string, time: int64) \rightarrow blob$$

BigTable

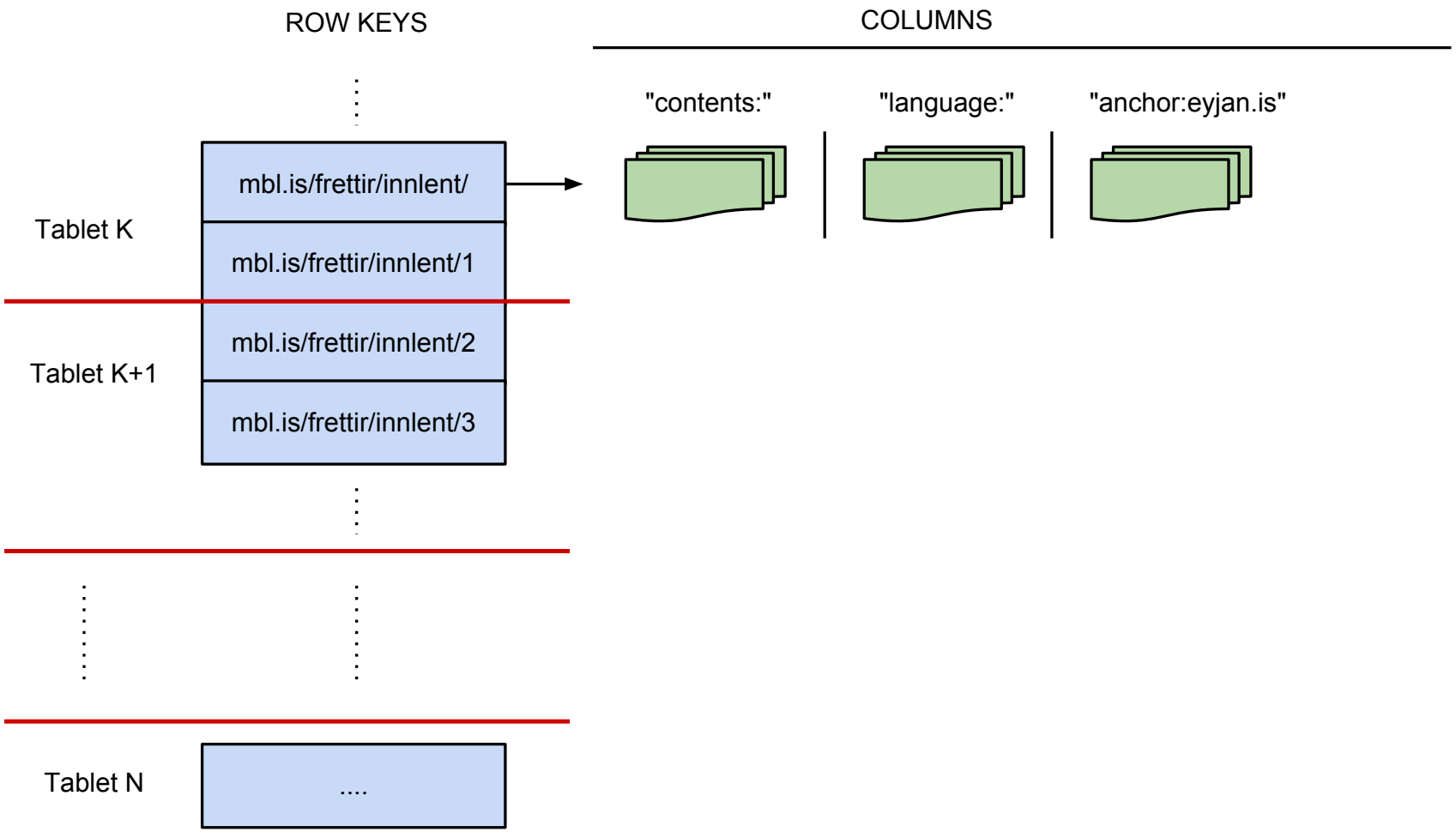
Töflur

- Raðir í töflu þurfa ekki að hafa sömu dálka (semi-structured)
- Raðir eru geymdar í stafrófsröð (samkvæmt row key)
- Hægt að lesa/skrifa staka röð með óskiptum (atomic) hætti
- Tímasett gildi, hægt að geyma margar útgáfur af gögnum

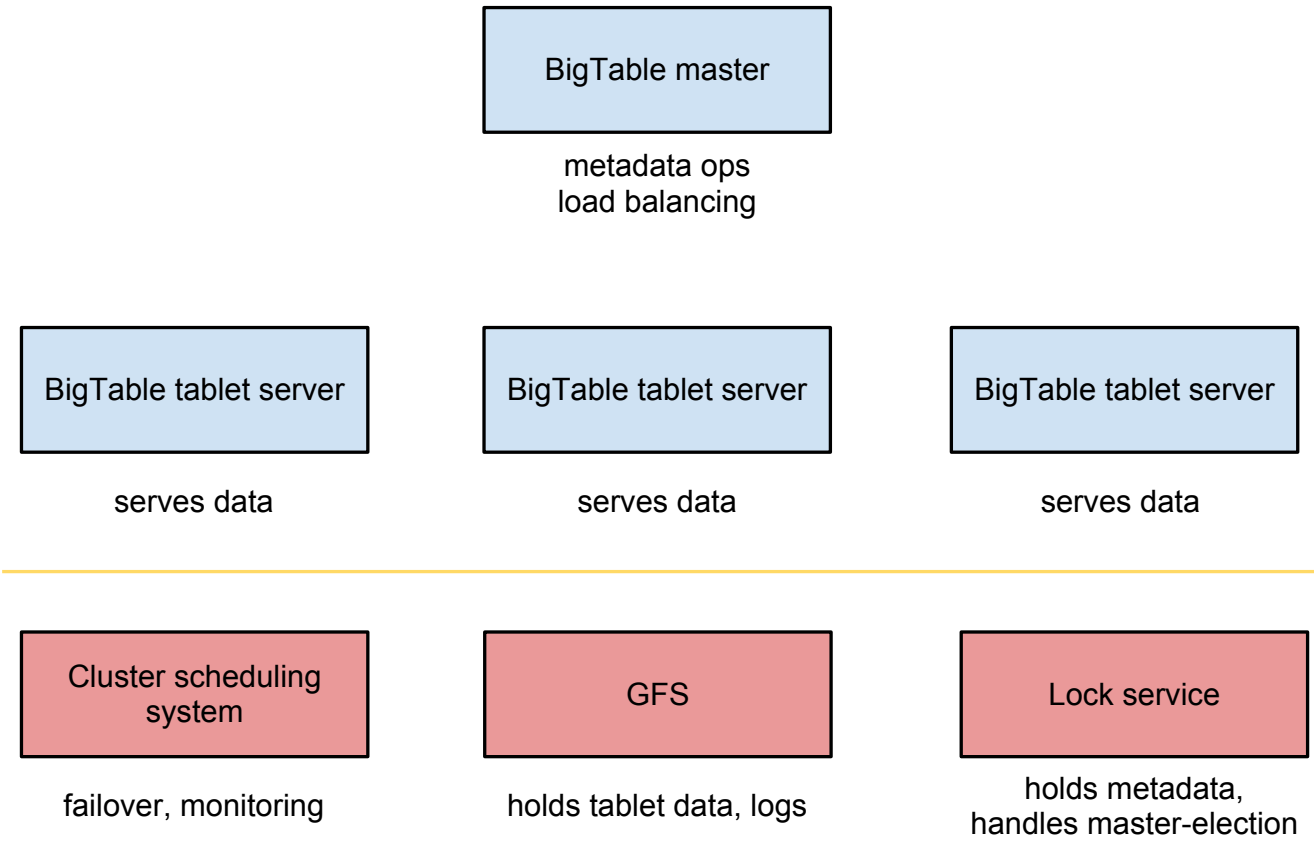
Tablets

- Tafla er hlutuð í svokölluð tablets þ.a. hvert tablet geymir ákveðið bil af röðum úr töflunni
- Hvert tablet er 100-200 MB
- Ef tablet verður of stórt þá er því splittað í tvö tablets
- Tablets er dreift á mismunandi netþjóna

BigTable


ROW KEYS
COLUMNS
"contents:"
"language:"
"anchor:eyjan.is"
mbl.is/frettir/innlent/
mbl.is/frettir/innlent/1
mbl.is/frettir/innlent/2
mbl.is/frettir/innlent/3
....
Tablet K
Tablet K+1
Tablet N

BigTable cluster



BigTable

BigTable Master

Hlutverk:

- Úthlutar tablets til tablet þjóna
- Bregst við þegar tablet þjónar koma og fara úr clusternum
- Passar að enginn hafi of mikið að gera (load balancing)
- Sér um metadata aðgerðir: stofna töflu, stofna dálkafjölskyldu

Lásaþjónusta (Chubby)

Hlutverk:

- Kosning á master þjóni
- Bootstrap fyrir staðsetningu á rótar tablet (METADATA)
- Geymir schema upplýsingar fyrir töflur
- Discovery/keepalive þjónusta fyrir tablet þjóna

BigTable

Tablet þjónn

Hlutverk:

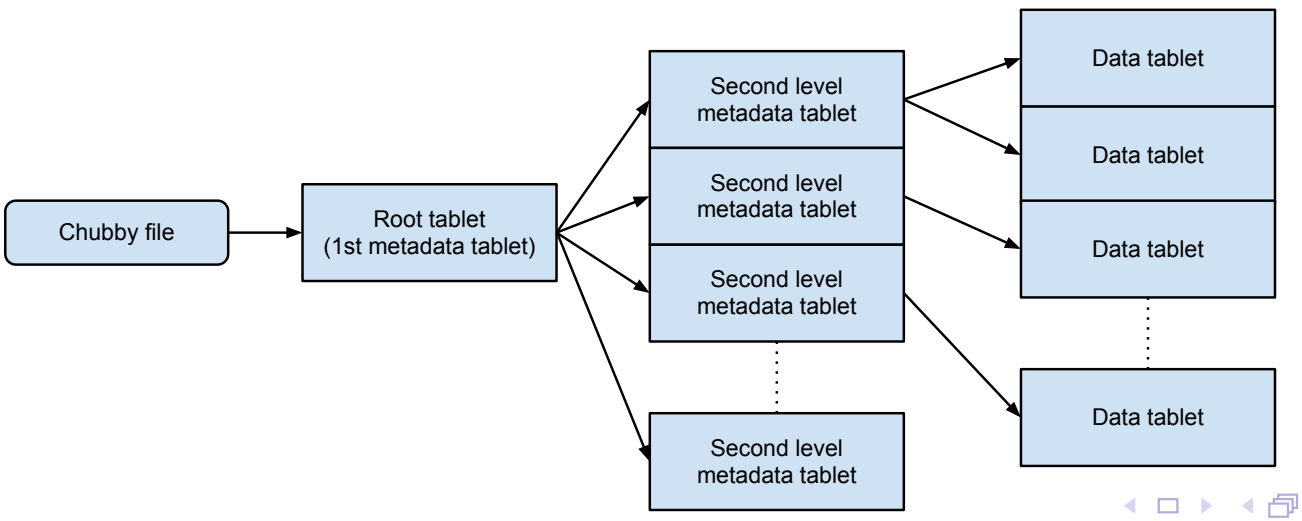
- Geymir eina eða fleiri tablets. Allt geymt í GFS.
- Miðlar gögnum úr tablets til biðlara (BigTable clients)
- Biðlari sækir gögn millliliðalaust beint frá tablet þjóni
- Lætur vita af sér í gegnum Chubby með því að búa til og taka lás á skrá í ákveðinni möppu
- Tablet þjónn viðheldur opinni setu (e. session) við Chubby lásaþjónustuna.
- Ef tablet þjónn missir setuna sína þá missir hann alla lása og er þá talinn dauður

BigTable

Uppflettingartré

Svipað og B+ tré:

- Uppfletting: tablename og key $\rightarrow$ tablet server
- Chubby inniheldur staðsetningu á rótarhnútnum
- Tablet í rótinni inniheldur vörpun tablename:interval $\rightarrow$ 2nd level tablet
- Annars stigs töflur innihalda síðan enn nákvæmari vörpun yfir tablets með sjálfum gögnunum



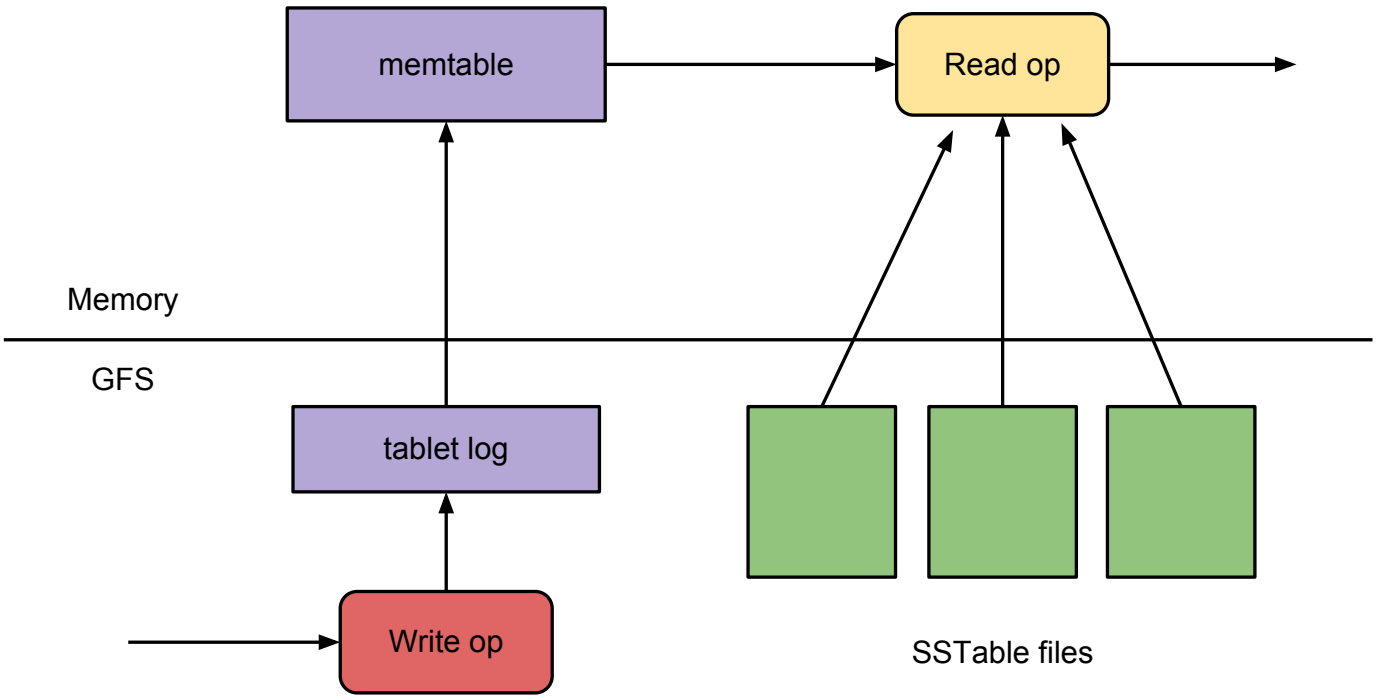
BigTable

Gagnageymsla

Tablets eru geymd í sérstökum read-only SSTable skrám:

- Nýjar uppfærslur eru skrifaðar í dagbók (WAL) og geymdar í minni í svokölluðu memtable
- Eldri uppfærslur eru geymdar í óbreytanlegum (immutable) SSTable skrám sem innihalda sorted key-value vörpun geymda í 64 KB blokkum. Aftast í SSTable er flýtivísir fyrir blokkir.
- Reglulega eru nýjar færslur skrifaðar út úr minni sem nýtt SSTable
- Lestur á gögnum byggist á samröðun á nýjum uppfærslum í minni og gögnum úr SSTable skrám
- Litlar SSTable eru reglulega sameinaðar í nýja stærri SSTable
- Kostur: runubundin skrif á disk, ekki random I/O

BigTable


memtable
Read op
Memory
GFS
tablet log
Write op
SSTable files

Heimildir

- *BigTable: A Distributed Storage System for Structured Data*, Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber, OSDI '06
- *Designs, Lessons and Advice from Building Large Scale Distributed Systems*, Jeff Dean, LADIS '09
- *The Google File System*, Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, SOSP '03