HBase at Xiaomi

Phil Yang & Guanghao Zhang
{yangzhe1991, zghao}@apache.org

About Xiaomi

- Founded in 2010
- $45 billion valuation
- 200+ million global users
- More than 20 independent Apps with DAU 10M+
- Products: smart phone, TV, router, smart band…

Since HBaseCon 2016

- 1 new PMC member
- 2 new committers (5 committers now)
- Resolved 180+ issues

Agenda

- HBase at Xiaomi
- Replication Improvements
- Confusing Behaviors
- Scan Improvements
- Async Client

Clusters and Scenarios

- IDC

  4 data centers (China), 30+ online clusters / 2 offline clusters

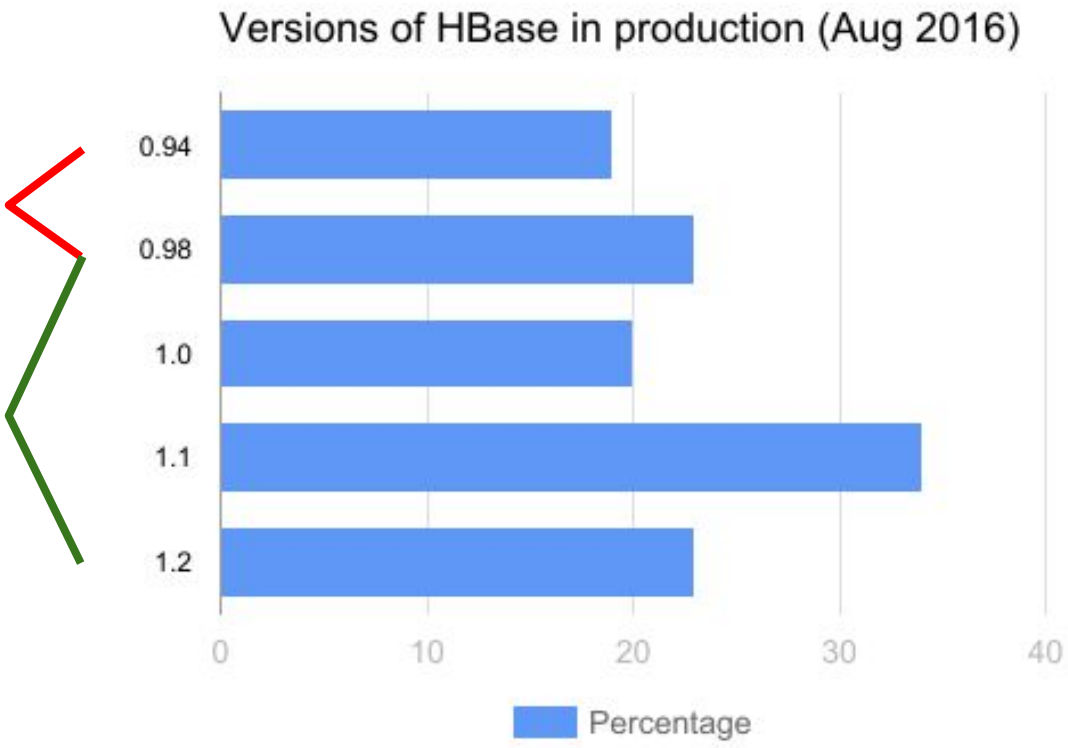
- AWS / Alibaba Cloud

  7 clusters (China/Singapore/US/Europe)

Upgrade 0.94 ⇒ 0.98

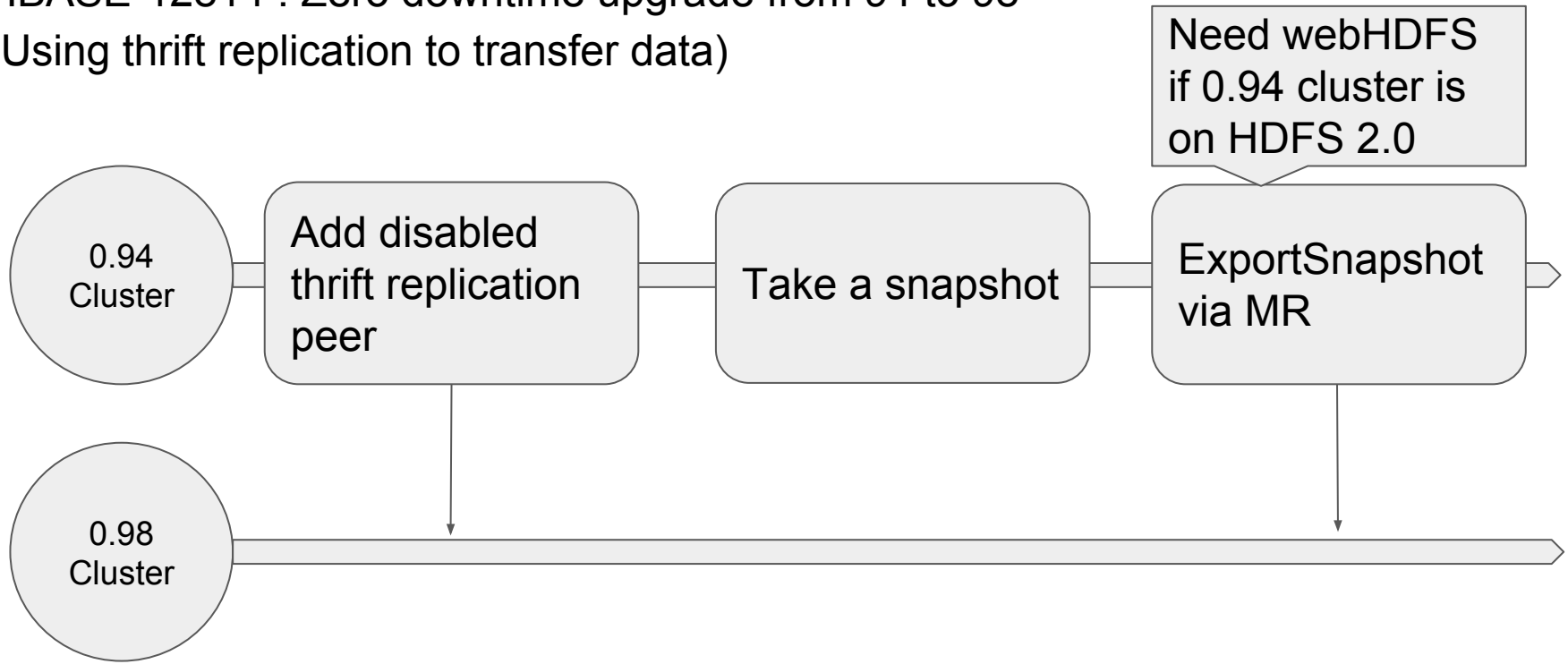
Not compatible

Compatible

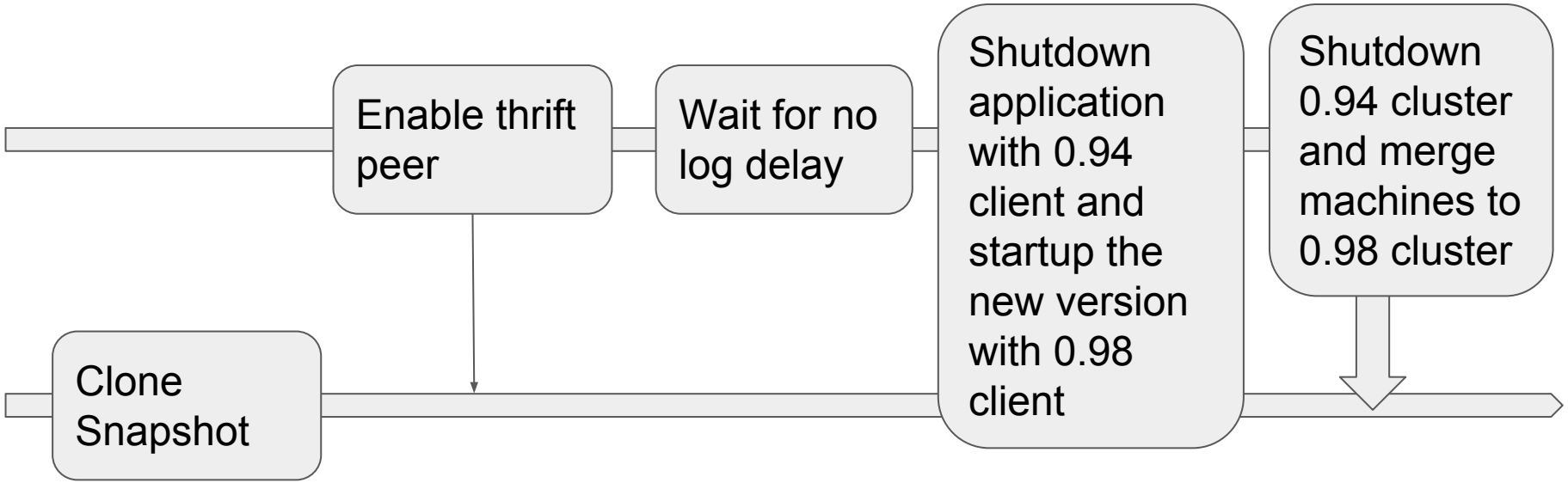


Upgrade 0.94 ⇒ 0.98

HBASE-12814 : Zero downtime upgrade from 94 to 98
(Using thrift replication to transfer data)



Upgrade 0.94 ⇒ 0.98

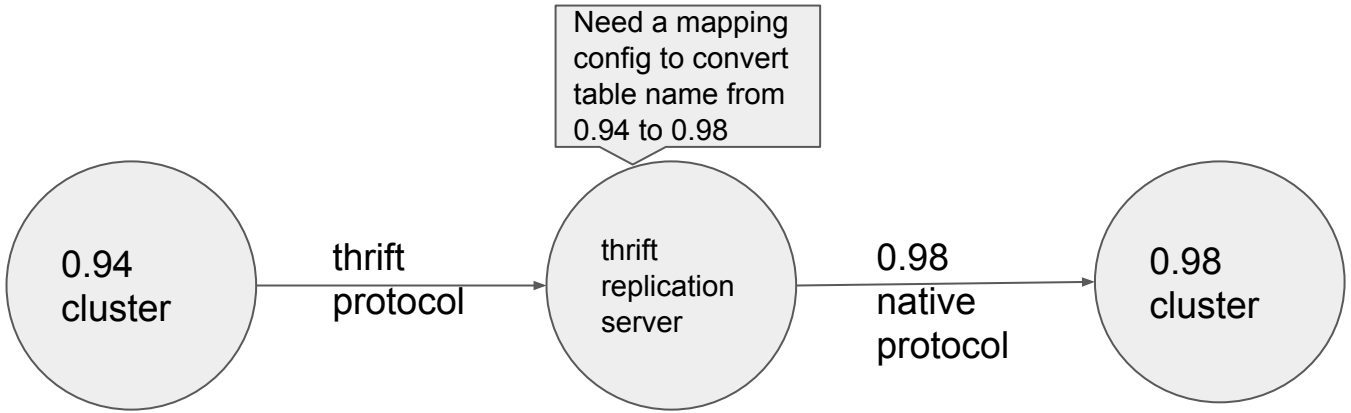


Use a Separate Thrift Server for Replication

Purpose:

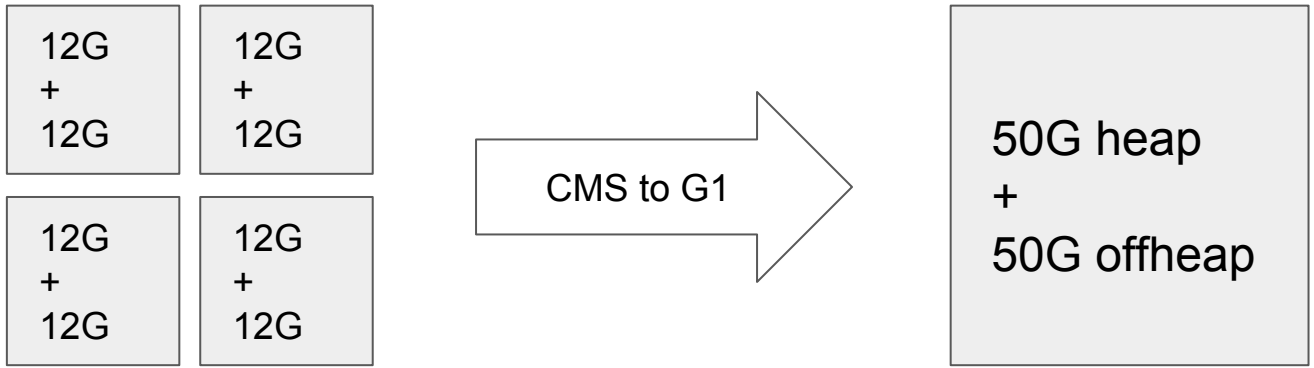
Reduce GC of RS

No need to restart RS while changing table name mapping config



Use G1 GC

24 core CPU / 128G memory



Why: Reduce Full GC, reduce number of instances

Must use latest jdk8 to prevent crashing in heavy load

Use G1 GC

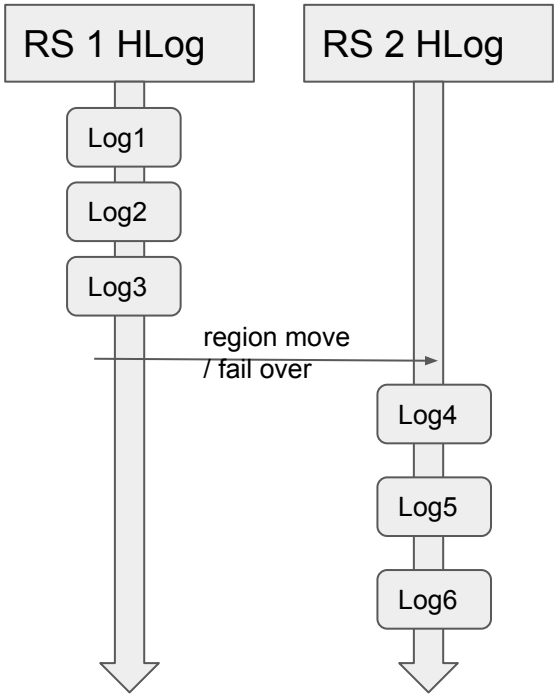
-XX:+UnlockExperimentalVMOptions
-XX:MaxGCPauseMillis={50/90/500} for SSD/HDD/offline cluster
-XX:G1NewSizePercent={2/5} for normal/heavy load cluster
-XX:InitiatingHeapOccupancyPercent=65
-XX:+ParallelRefProcEnabled
-XX:ConcGCThreads=4
-XX:ParallelGCThreads=16
-XX:MaxTenuringThreshold=1
-XX:G1HeapRegionSize=32m
-XX:G1MixedGCCountTarget=64
-XX:G1OldCSetRegionThresholdPercent=5

Replication Improvements

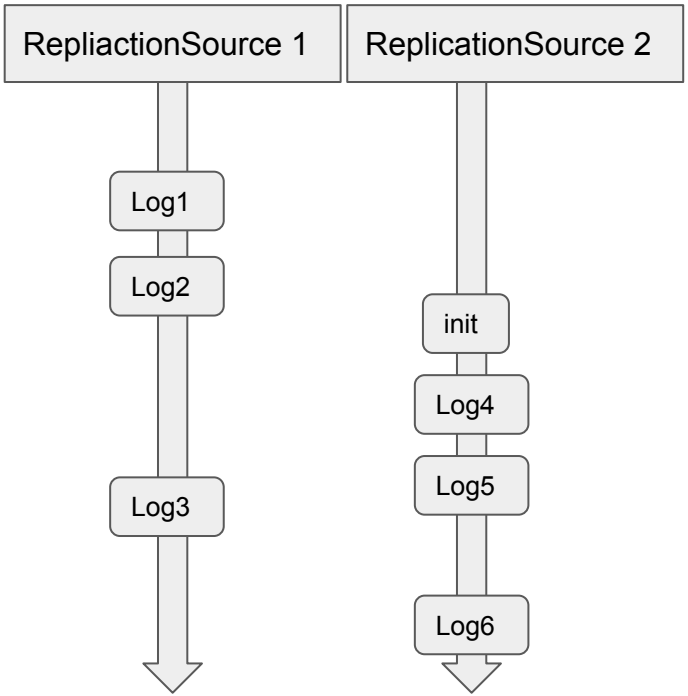
- HBASE-16447 Replication by namespaces config in peer
- HBASE-17296 Provide per peer throttling for replication
- HBASE-17314 Limit total buffered size for all replication sources
- HBASE-12770 Don't transfer all the queued hlogs of a dead server to the same alive server
- HBASE-9465 Push entries to peer clusters serially

HBASE-9465 Serial Replication

Before HBASE-9465(<= 1.3.x):

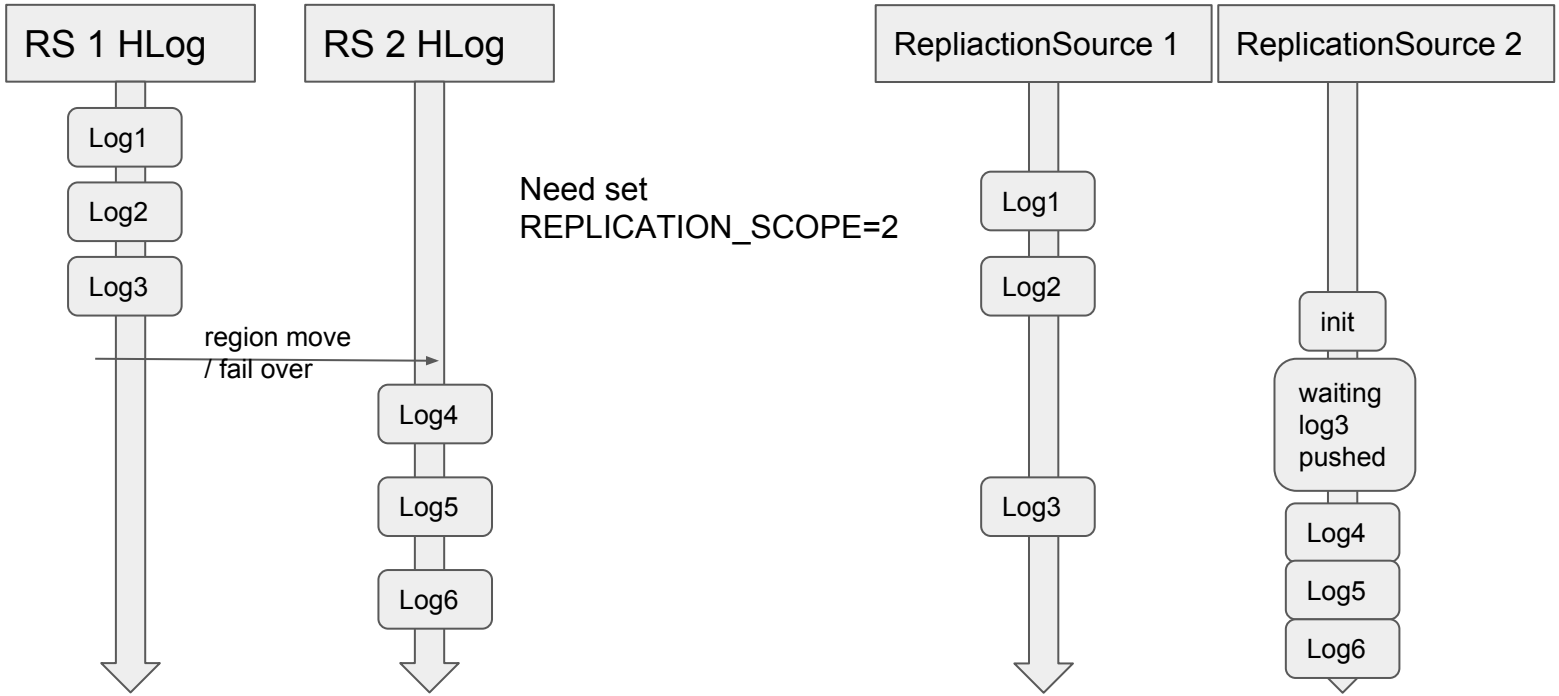


Log4/5 is pushed before log3



HBASE-9465 Serial Replication

After HBASE-9465(>= 1.4.0/2.0.0):



Confusing Behaviors

Inconsistent results between source cluster and peer cluster

- HBASE-9465

Deletes mask puts, even puts that happened after the delete was entered

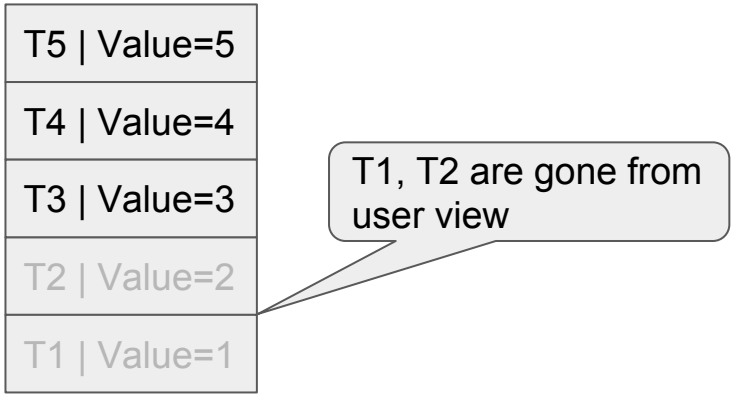
- HBASE-15968

Inconsistent Results when Using Filter to Read Data

Column family's MaxVersions = 3

1. put T1, T2, T3, T4, T5

2. scan with a ValueFilter(value<=3)

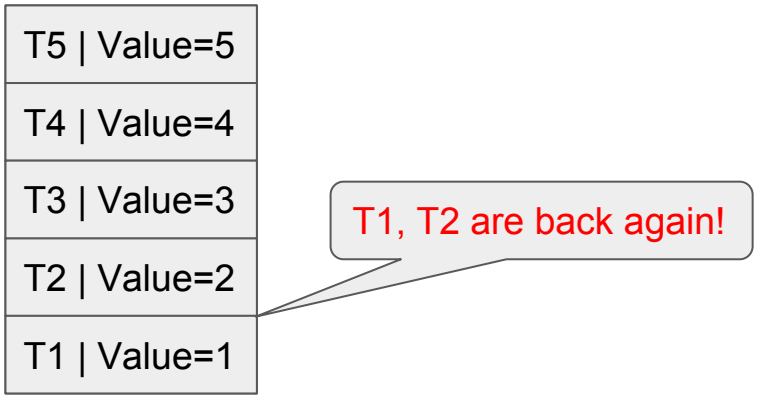


Inconsistent Results when Using Filter to Read Data

Column family's MaxVersions = 3

1. put T1, T2, T3, T4, T5

2. scan with a ValueFilter(value<=3)



Inconsistent Results when Using Filter to Read Data

Solution：Adjust the execution order in ScanQueryMatcher.matchColumn

- check column
- **check by filter**
- check versions

Inconsistent Results when Using Filter to Read Data

Solution：Adjust the execution order in ScanQueryMatcher.matchColumn

- check column
- **check by filter**
- check versions

- check column
- check versions
- **check by filter**

Inconsistent Results when Using Filter to Read Data

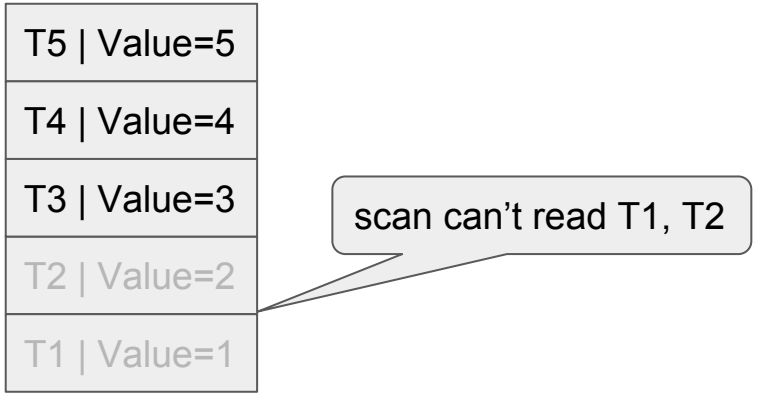
Solution：Adjust the execution order in ScanQueryMatcher.matchColumn

- check column
- **check by filter**
- check versions

- check column
- check versions
- **check by filter**

1. put T1, T2, T3, T4, T5

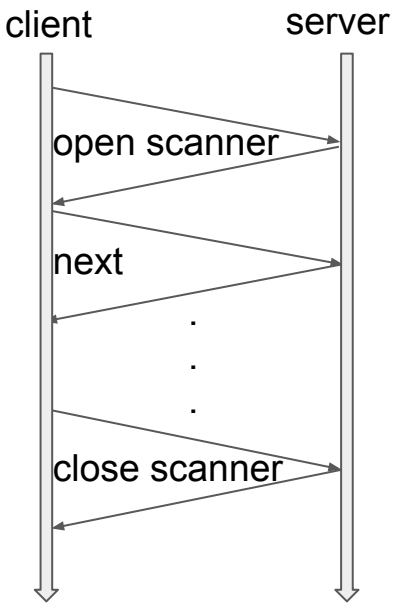
2. scan with a ValueFilter(value<=3)



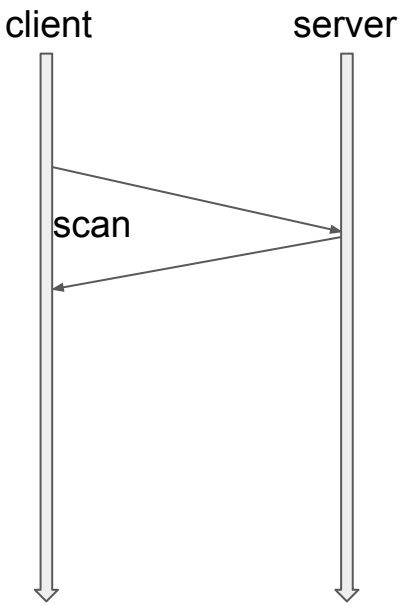
Scan Improvements

- Add inclusive/exclusive support for startRow and stopRow
- Add Scan.setLimit(int) to limit the number of rows for Scan
- Unify the implementation of small scan and regular scan
- Pass mvcc to client when scan

Unify the Implementation of Small Scan and Regular Scan



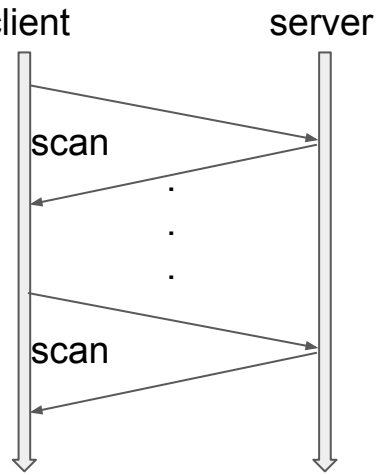
regular scan



small scan

Unify the Implementation of Small Scan and Regular Scan

- All scan rpc requests return results
- Use pread by default and switch to streaming read if needed



unified scan

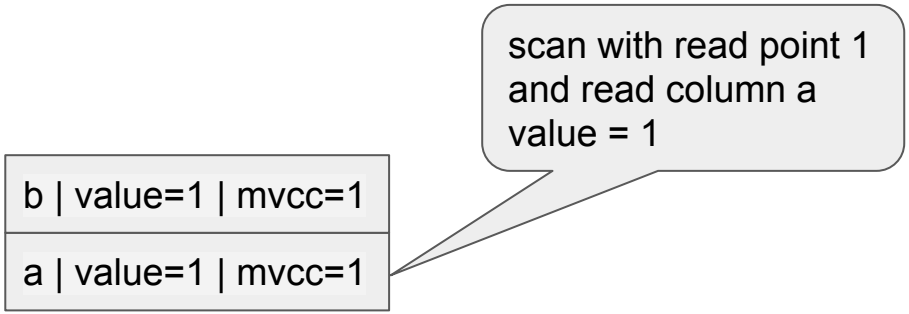
Scan May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and restrat scan

b | value=1 | mvcc=1
a | value=1 | mvcc=1

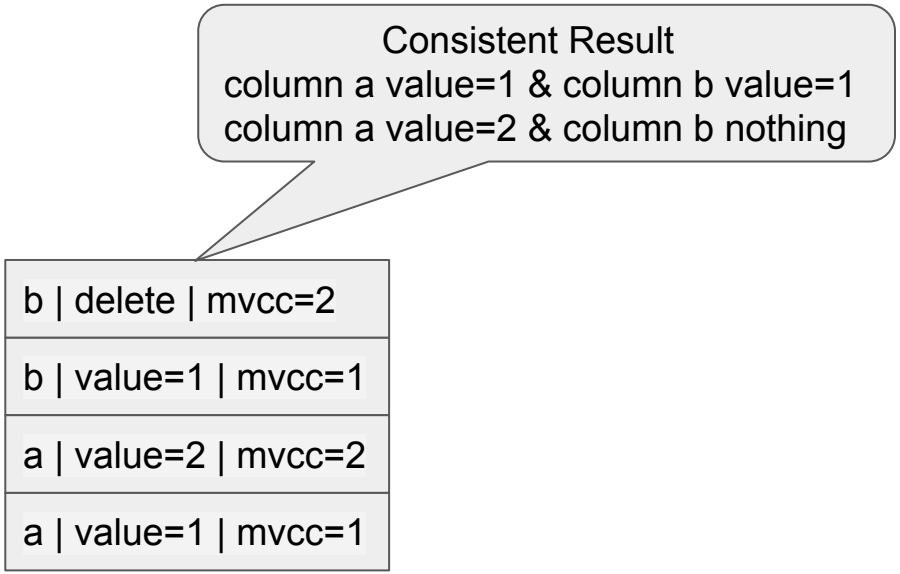
Scan May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and restart scan



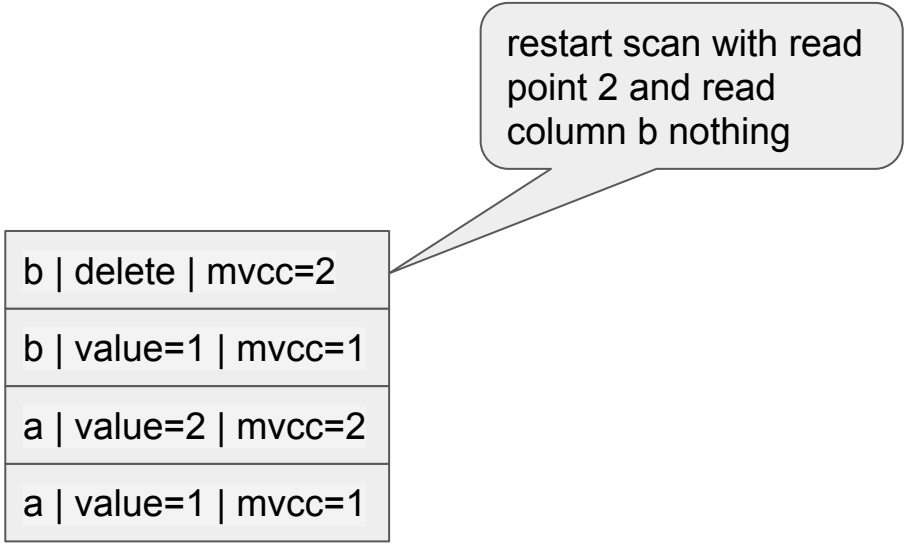
Scan May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and restart scan



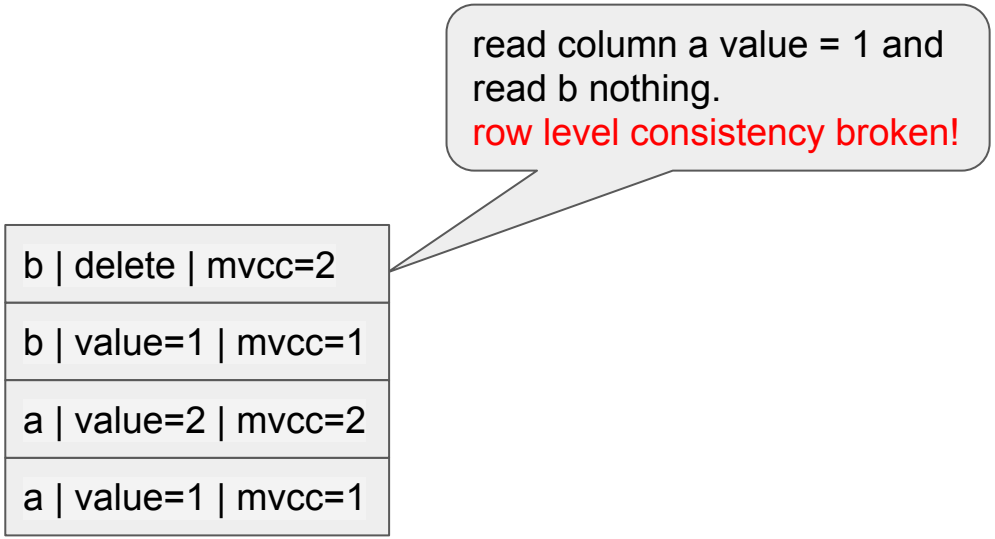
Scan May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and restart scan



Scan May Break Row Level Consistency

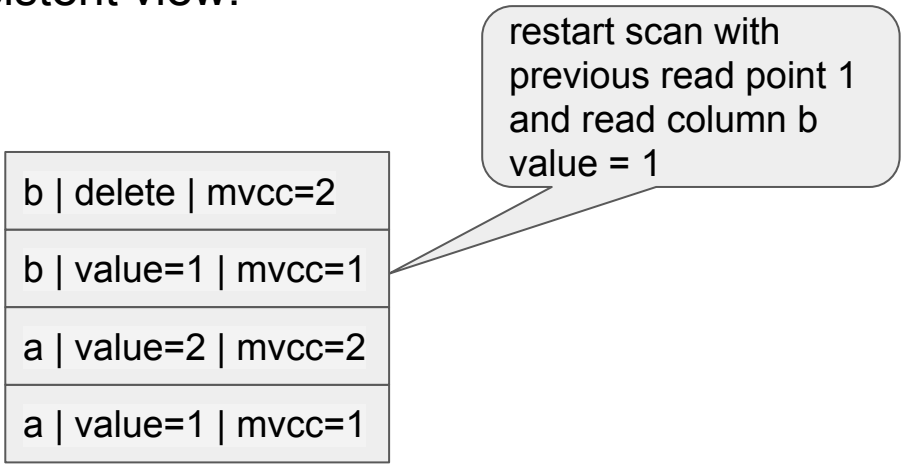
1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and restart scan



HBASE-17167 Pass mvcc to client when scan

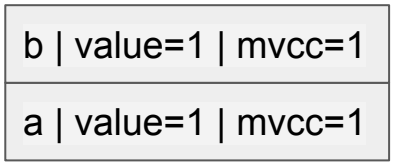
Solution: pass read point to client. When region is moved, use the previous read point to restart a scan to get a consistent view.

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and restart scan



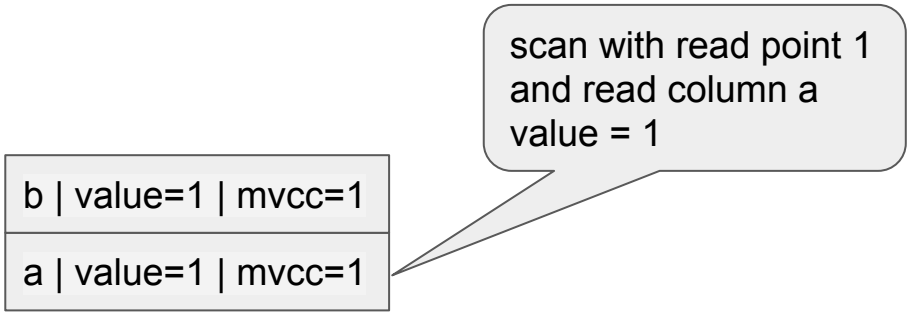
Compaction May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and compact
5. region move and restrat scan



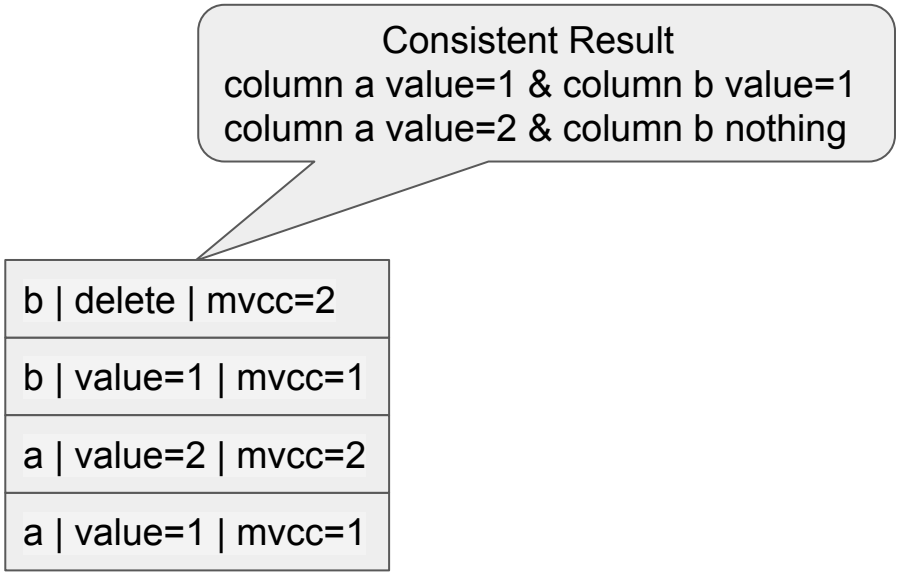
Compaction May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and compact
5. region move and restart scan



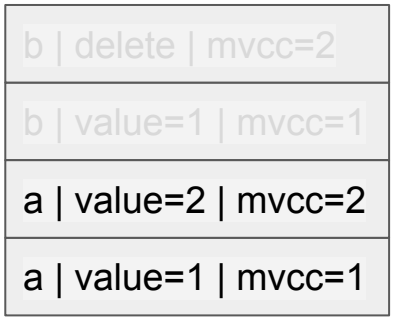
Compaction May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and compact
5. region move and restart scan



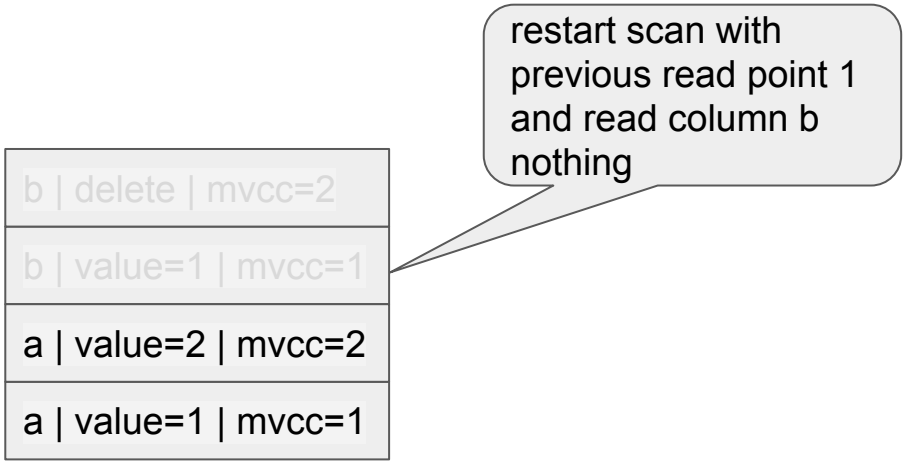
Compaction May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and compact
5. restart scan



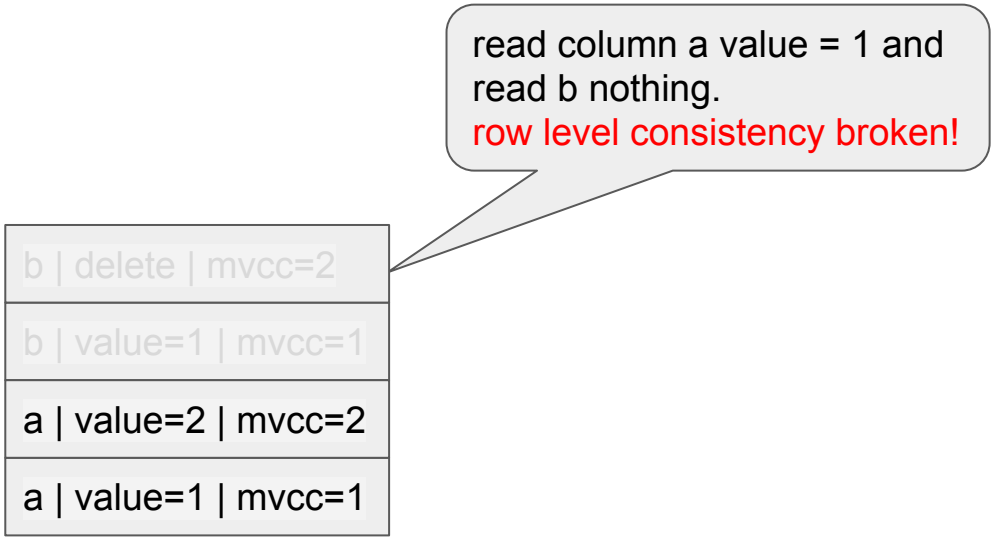
Compaction May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and compact
5. restart scan



Compaction May Break Row Level Consistency

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and compact
5. restart scan

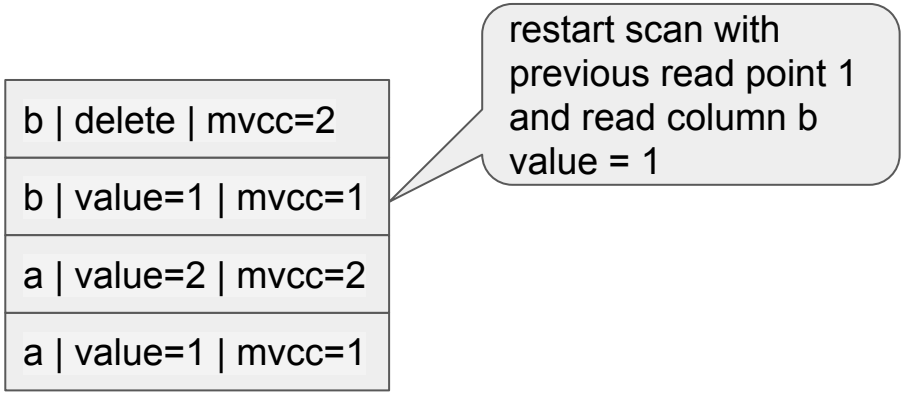


HBASE-17177

Status: OPEN

Candidate solution: Disable compaction for a while when open region

1. put column a, put column b
2. scan
3. put column a, delete column b
4. region move and restart scan



Async HBase Client

Implementation

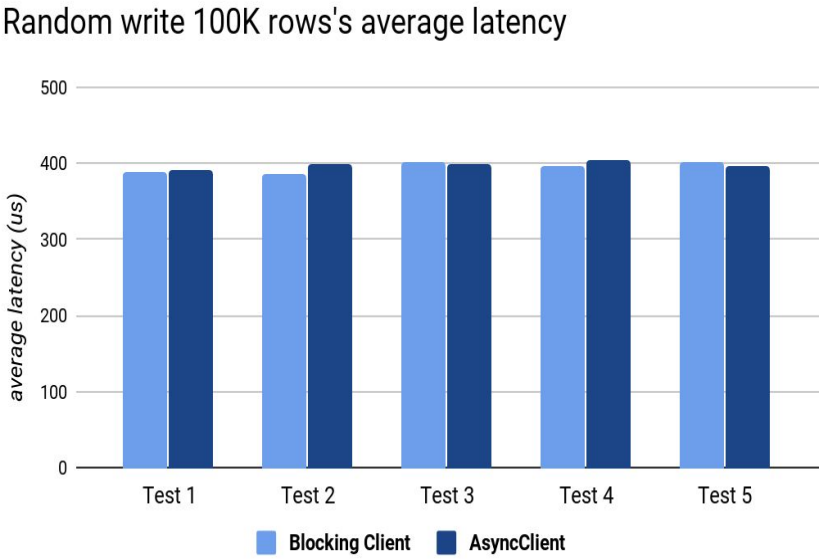
- Use the asynchronous protobuf stub
- Use CompletableFuture (Must use latest jdk8 for performance issue)

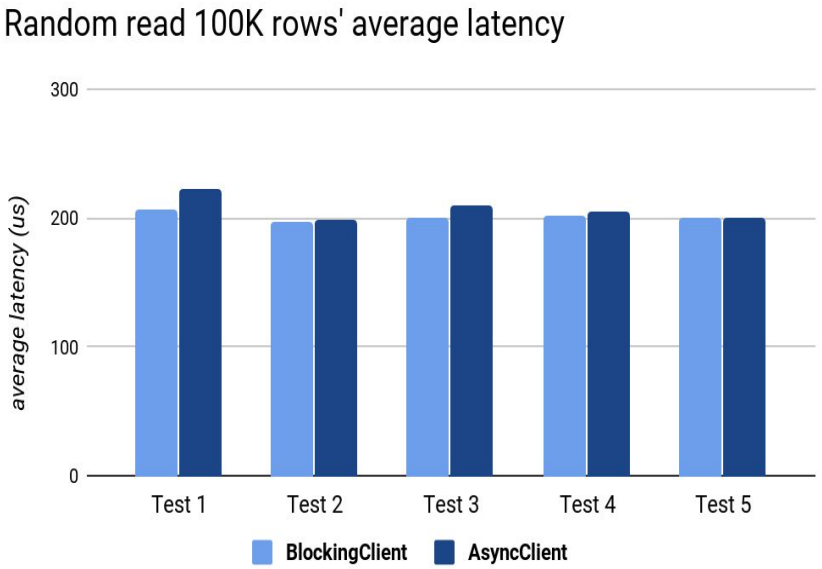
Async HBase Client

User API

- AsyncTable using users' ExecutorService
- RawAsyncTable for experts
- ResultScanner in old style
- (Raw)ScanResultConsumer using observer pattern

PeformanceEvaluation Test





Thank you!

Phil Yang & Guanghao Zhang
{yangzhe1991, zghao}@apache.org