HBase Internals and Operations

Bloomberg Engineering

SRECon19 Asia/Pacific
June 13, 2019

Biju Nair
Software Engineer
bnair10@bloomberg.net

TechAtBloomberg.com

Agenda

- Introduction to HBase
- Operating HBase
- Questions

Bloomberg in a nutshell

The ***Bloomberg Terminal*** delivers a diverse array of information on a single platform to facilitate financial decision-making.

Bloomberg technology by the numbers

- **5,000+** software engineers
- **150+** technologists and data scientists devoted to machine learning
- One of the largest private networks in the world
- **120 billion** pieces of data from the financial markets each day, with a peak of more than 10 million messages/second
- **2 million** news stories ingested / published each day (500+ news stories ingested/second)
- News content from **125K+ sources**
- Over **1 billion** messages and Instant Bloomberg (IB) chats handled daily

TechAtBloomberg.com

Bloomberg

Engineering

HBase at Bloomberg

- Started with **v0.94.6**
- **>2 billion** reads per day
- **>1 billion** writes per day
- **51+ TB** of compressed data stored in HBase

HBase Principles

- Ordered Key Value Store

- Distributed shared nothing

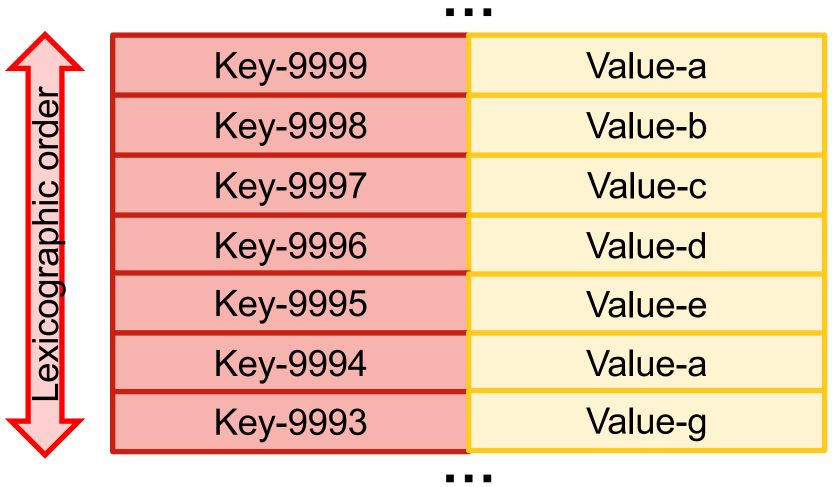
Key Value

...

Key-9999	Value-a
Key-9998	Value-b
Key-9997	Value-c
Key-9996	Value-d
Key-9995	Value-e
Key-9994	Value-f

...

Ordered Key Value

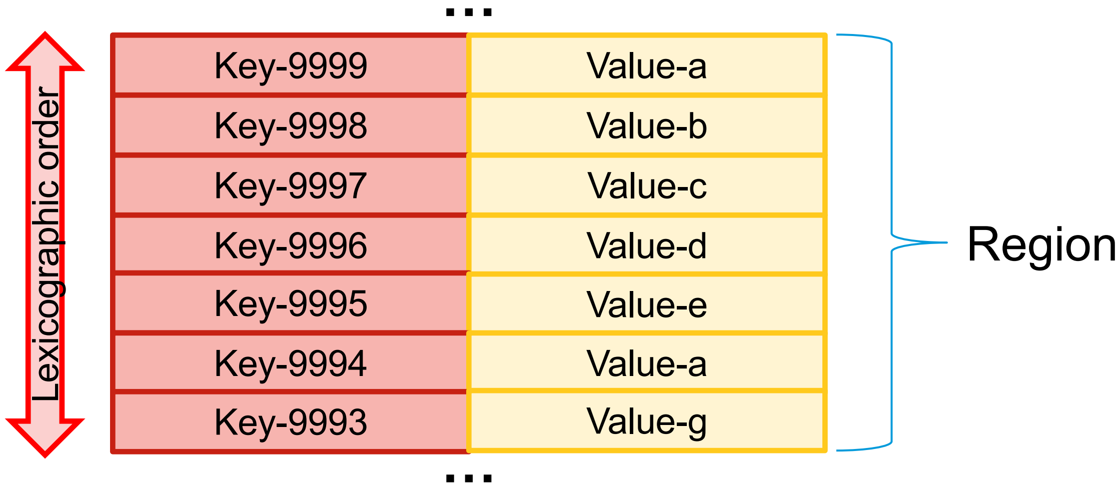


...

Key-9999	Value-a
Key-9998	Value-b
Key-9997	Value-c
Key-9996	Value-d
Key-9995	Value-e
Key-9994	Value-a
Key-9993	Value-g

...

Ordered Key Value

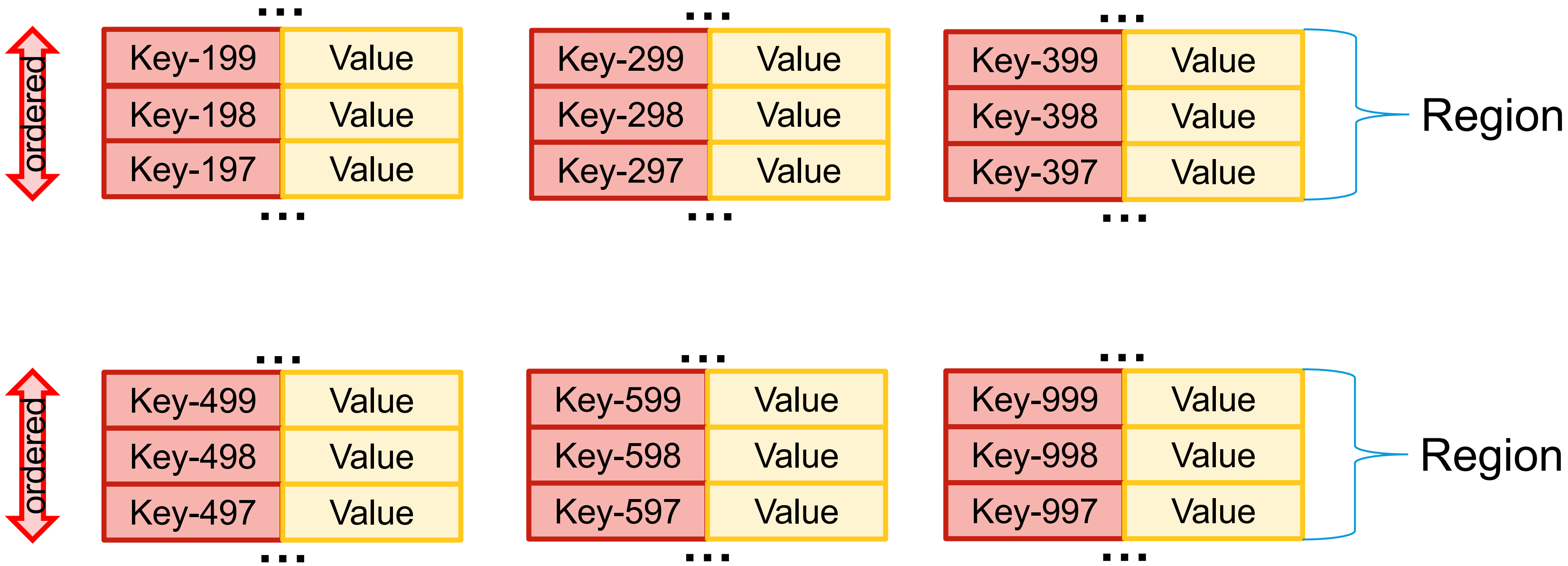


…

Key-9999	Value-a
Key-9998	Value-b
Key-9997	Value-c
Key-9996	Value-d
Key-9995	Value-e
Key-9994	Value-a
Key-9993	Value-g

…

Distributed Ordered Key Value



ordered

...

Key-199	Value	Key-299	Value	Key-399	Value
Key-198	Value	Key-298	Value	Key-398	Value
Key-197	Value	Key-297	Value	Key-397	Value

...

Region

ordered

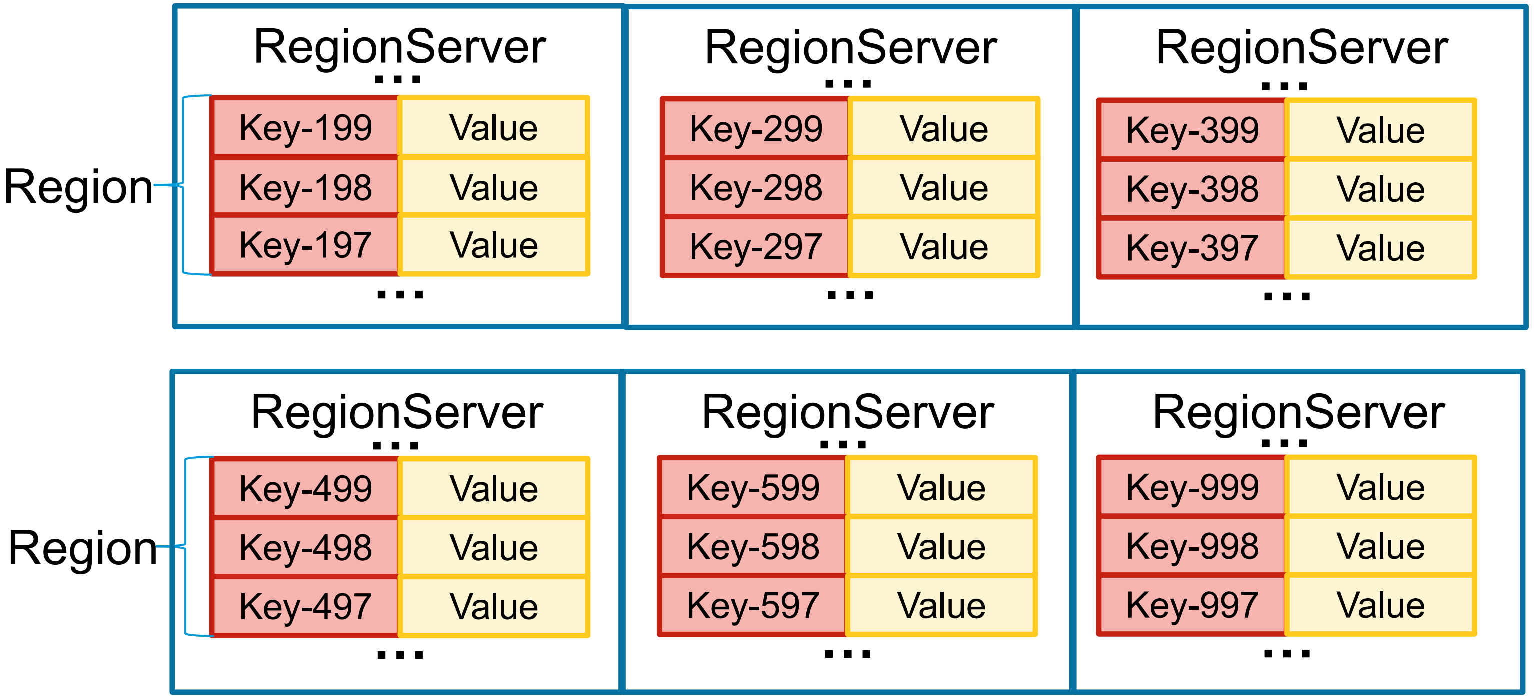
...

Key-499	Value	Key-599	Value	Key-999	Value
Key-498	Value	Key-598	Value	Key-998	Value
Key-497	Value	Key-597	Value	Key-997	Value

...

Region

Distributed Ordered Key Value



Region

Region

RegionServer
…
Key-199 Value
Key-198 Value
Key-197 Value
…

RegionServer
…
Key-299 Value
Key-298 Value
Key-297 Value
…

RegionServer
…
Key-399 Value
Key-398 Value
Key-397 Value
…

RegionServer
…
Key-499 Value
Key-498 Value
Key-497 Value
…

RegionServer
…
Key-599 Value
Key-598 Value
Key-597 Value
…

RegionServer
…
Key-999 Value
Key-998 Value
Key-997 Value
…

Table Row View

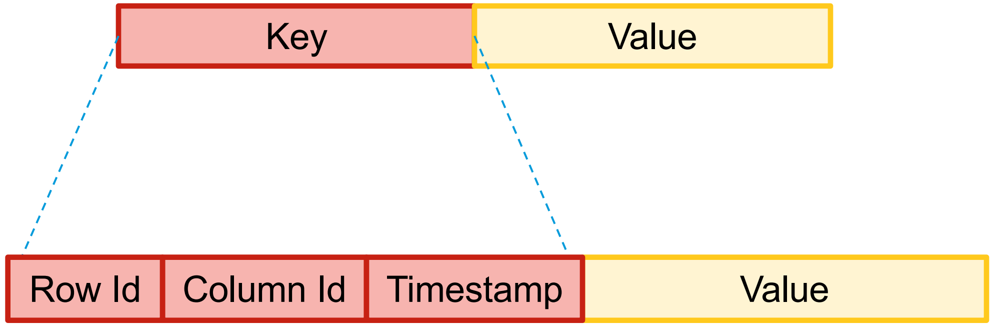
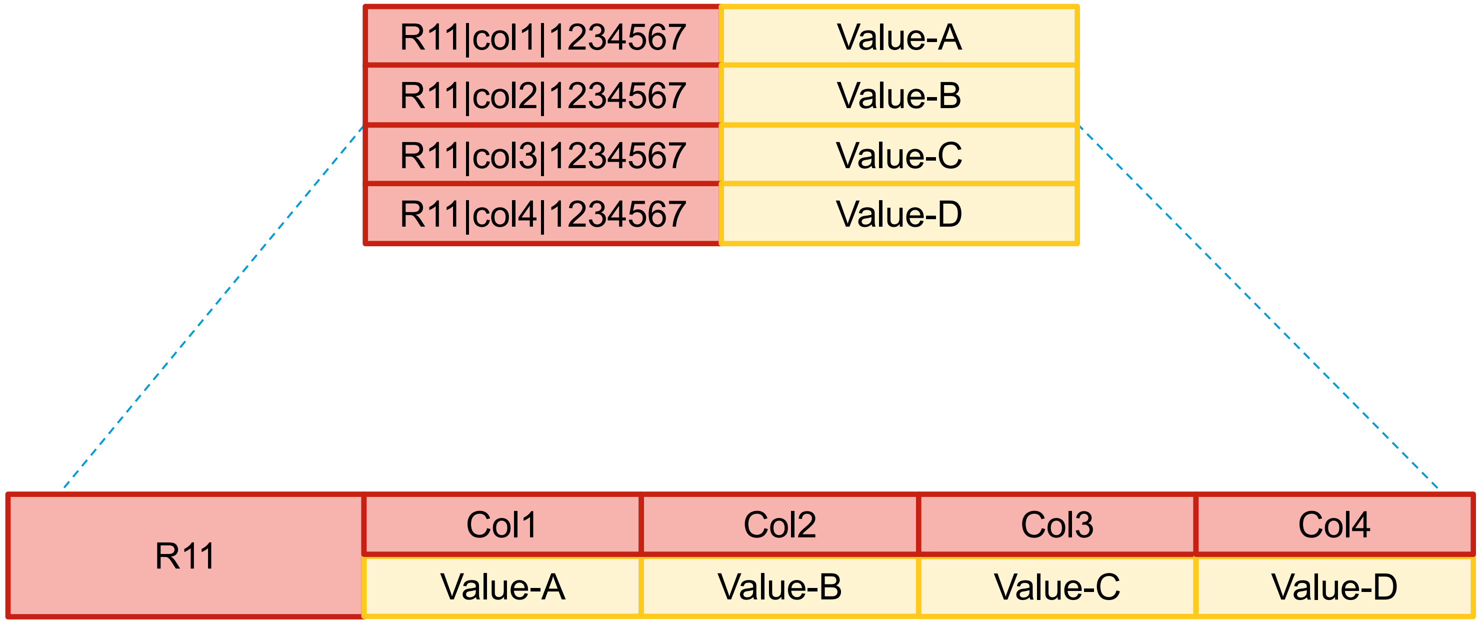


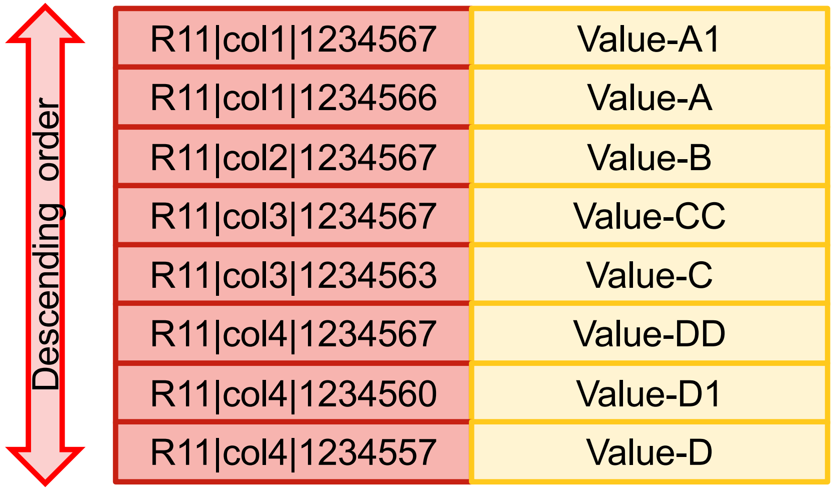
Table Row View



R11|col1|1234567	Value-A
R11|col2|1234567	Value-B
R11|col3|1234567	Value-C
R11|col4|1234567	Value-D

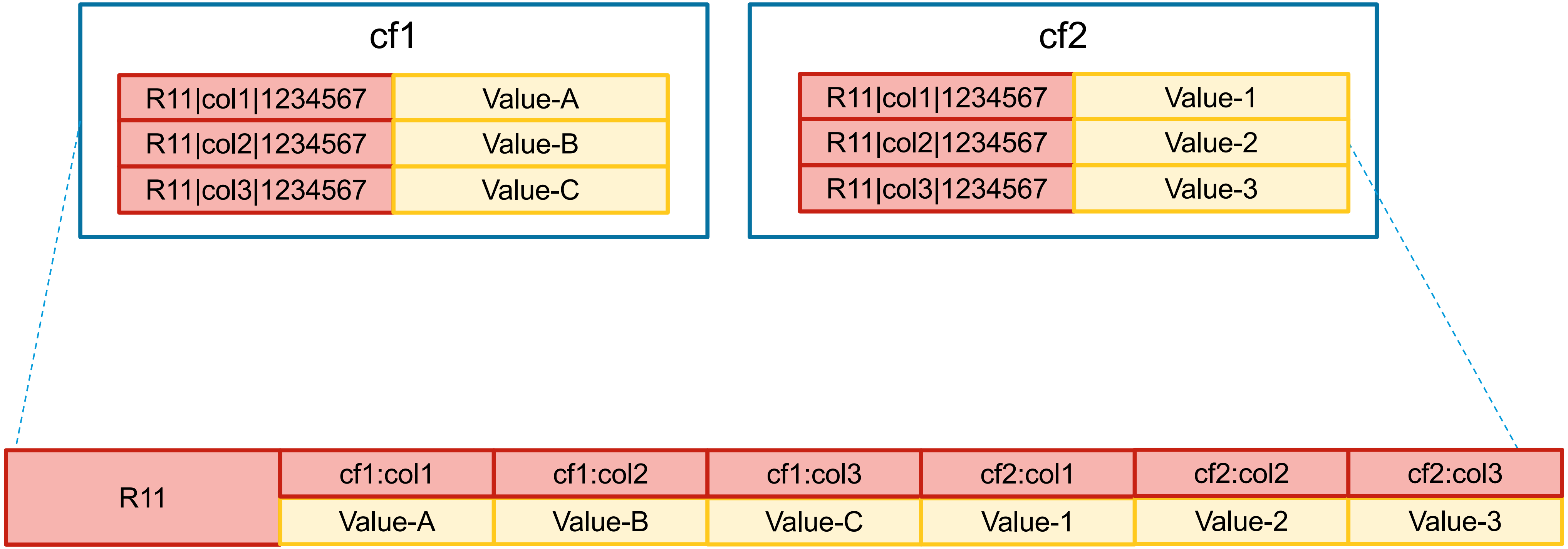
R11	Col1	Col2	Col3	Col4
	Value-A	Value-B	Value-C	Value-D

Versioning



R11|col1|1234567	Value-A1
R11|col1|1234566	Value-A
R11|col2|1234567	Value-B
R11|col3|1234567	Value-CC
R11|col3|1234563	Value-C
R11|col4|1234567	Value-DD
R11|col4|1234560	Value-D1
R11|col4|1234557	Value-D

Column Family



cf1

R11|col1|1234567	Value-A
R11|col2|1234567	Value-B
R11|col3|1234567	Value-C

cf2

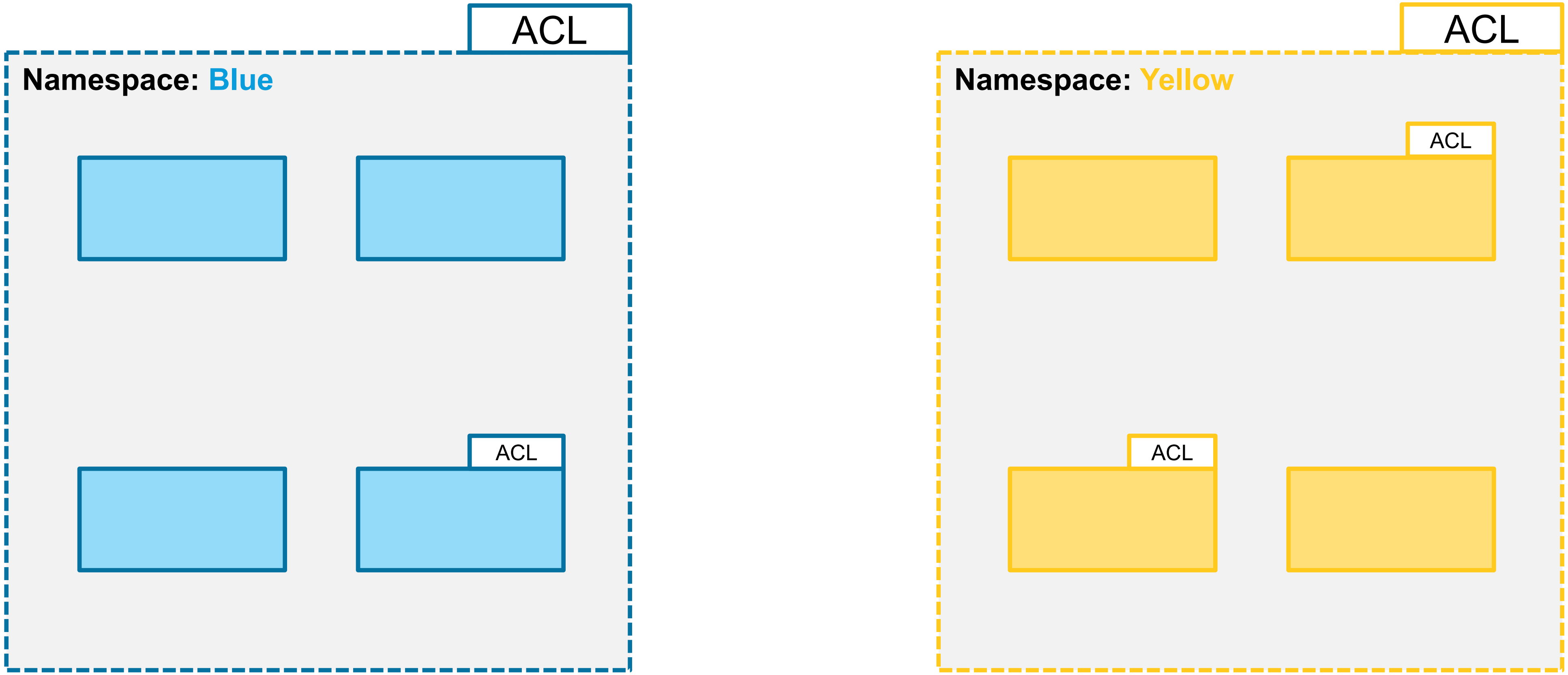
R11|col1|1234567	Value-1
R11|col2|1234567	Value-2
R11|col3|1234567	Value-3

R11	cf1:col1	cf1:col2	cf1:col3	cf2:col1	cf2:col2	cf2:col3
	Value-A	Value-B	Value-C	Value-1	Value-2	Value-3

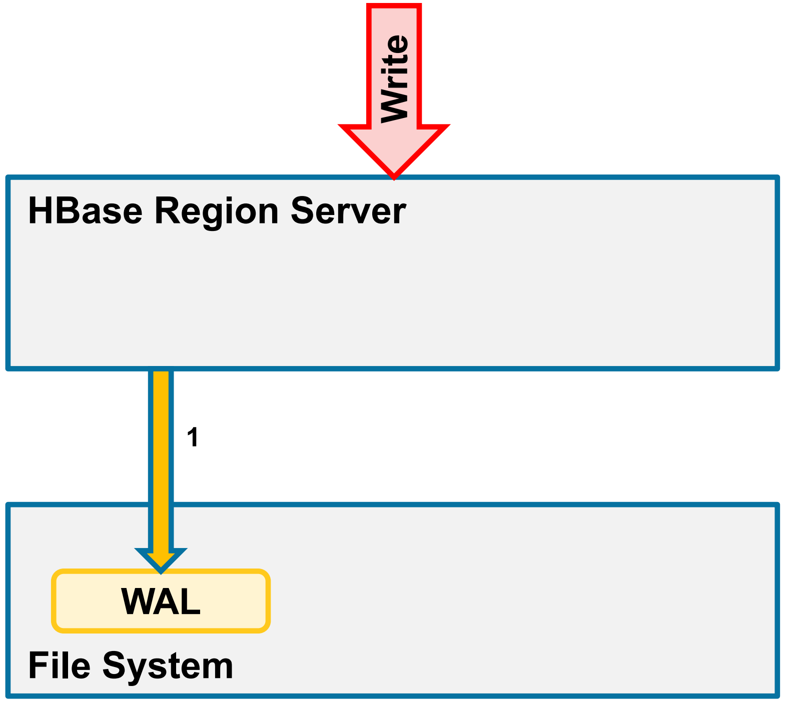
ACIDity

- **A**tomic at row level
- **C**onsistent to a point in time before the request
- **I**solation through MVCC (reads) and row locks (mutations)
- **D**urability is guaranteed for all successful mutations

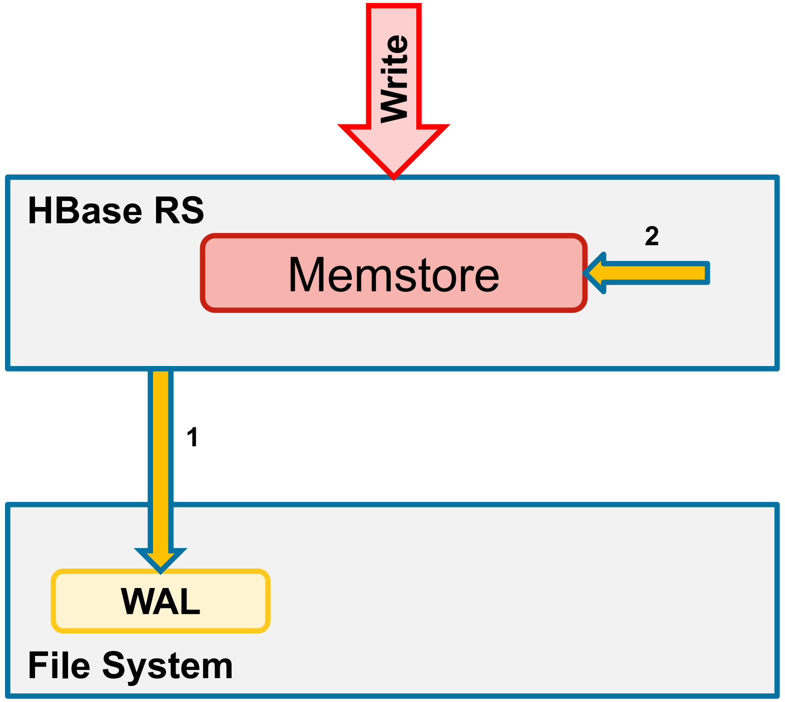
Namespace



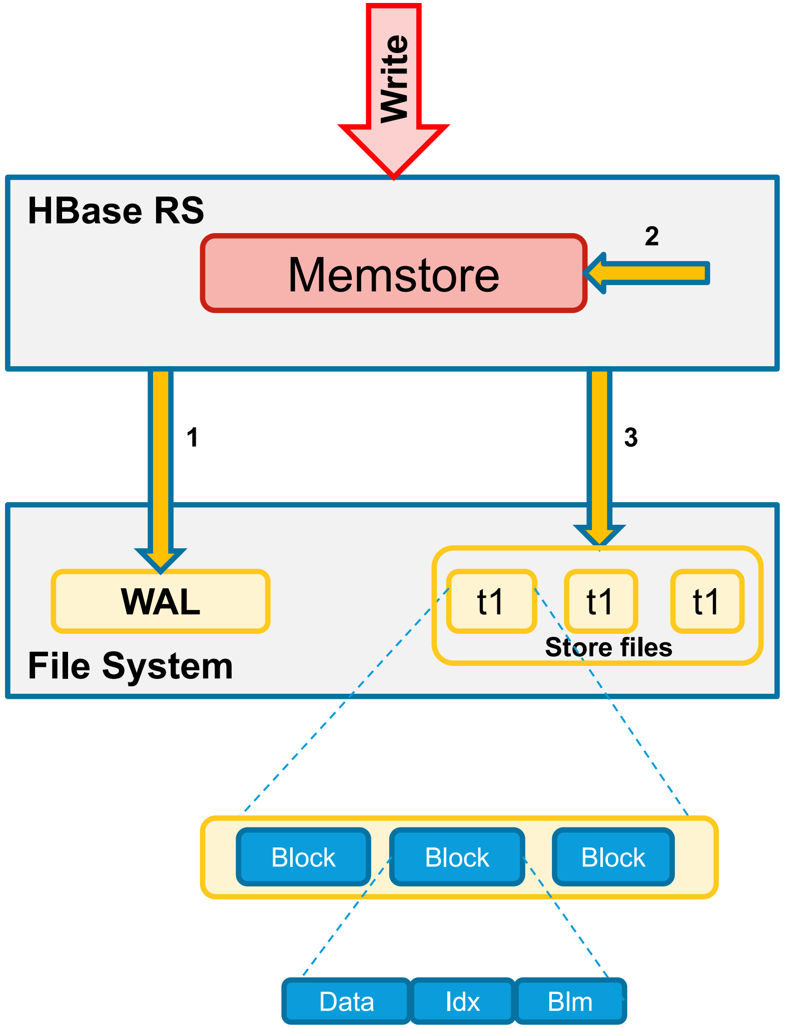
HBase Write



HBase Write



HBase Write

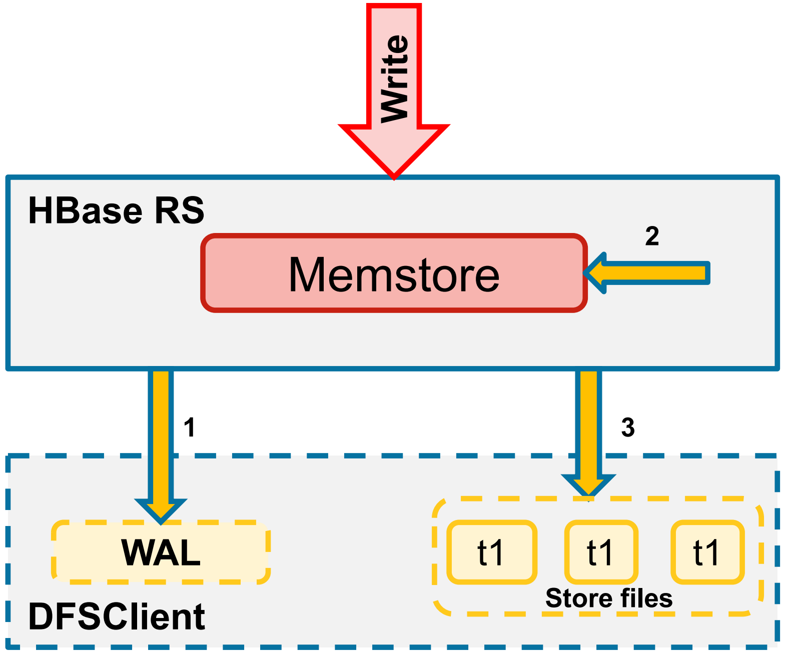


TechAtBloomberg.com

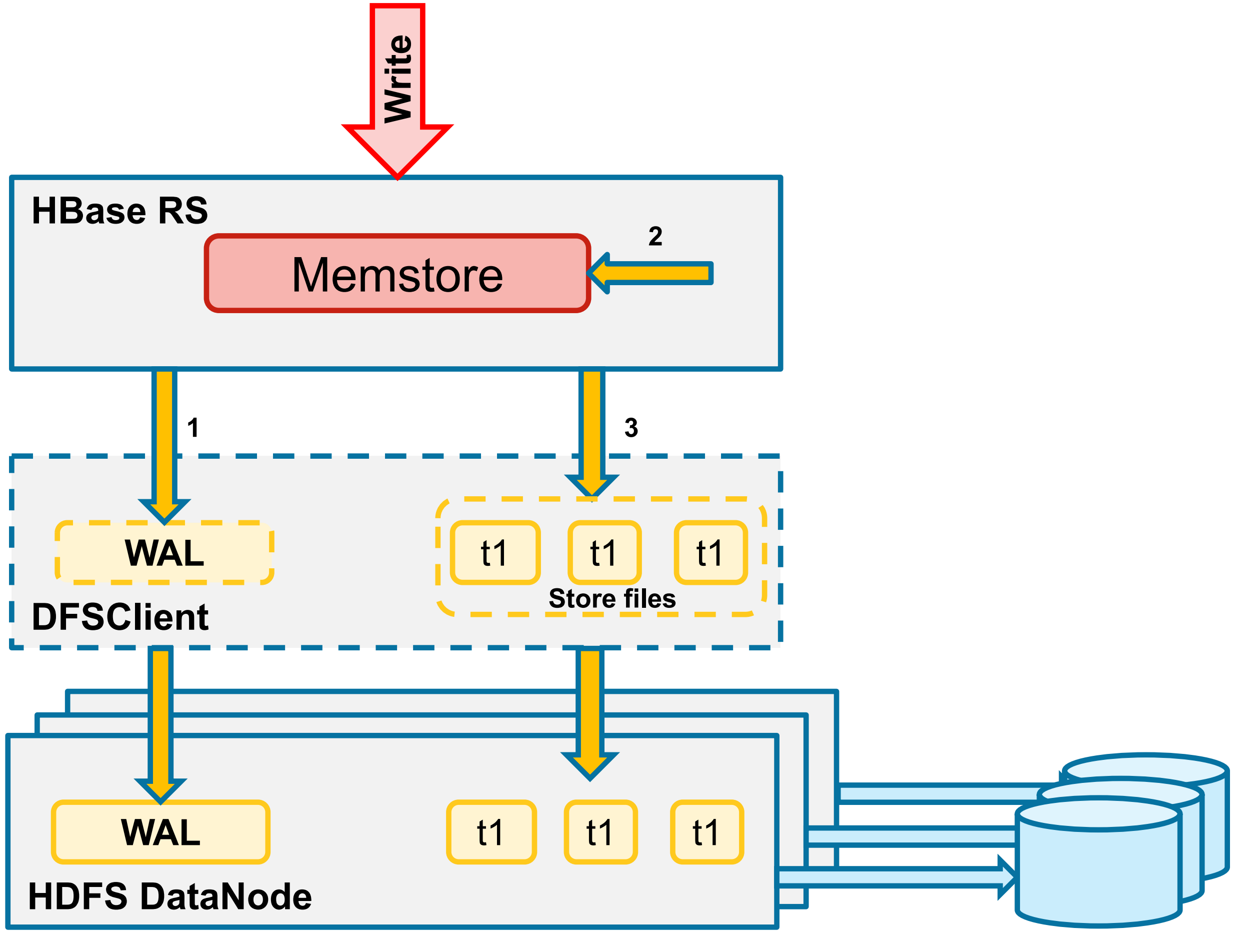
Bloomberg

Engineering

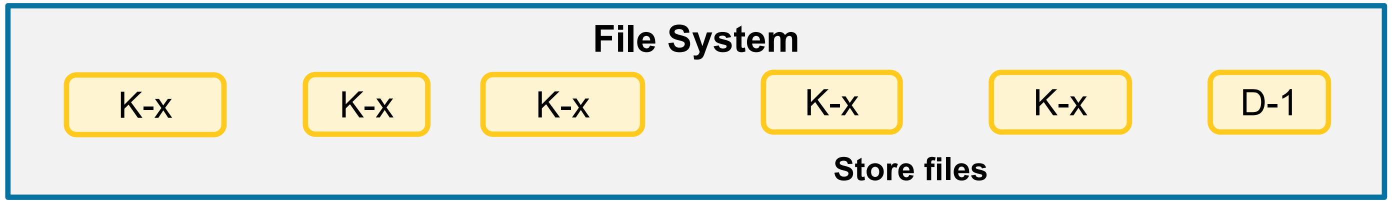
HBase Write


Write
HBase RS
Memstore
2
1
3
WAL
DFSClient
t1
t1
t1
Store files

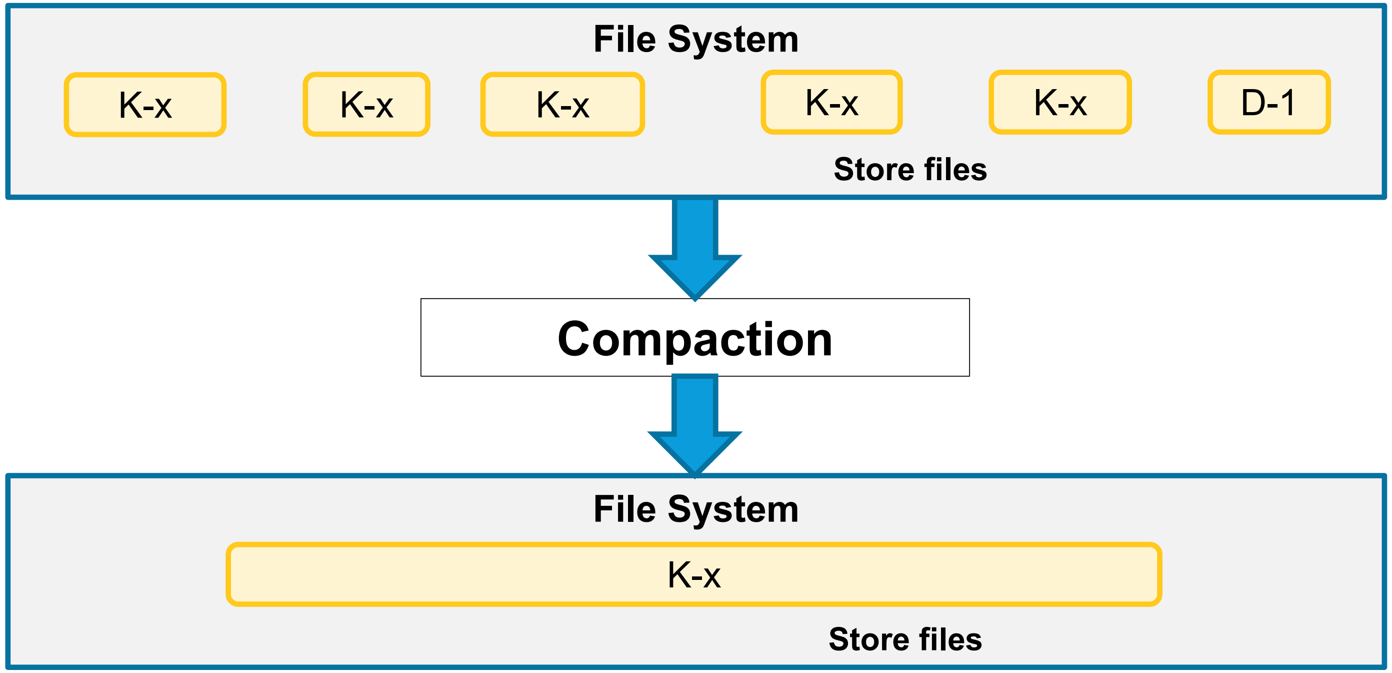
HBase Write


Write
HBase RS
Memstore
2
1
3
WAL
DFSClient
t1
t1
t1
Store files
WAL
t1
t1
t1
HDFS DataNode

Compaction



Compaction



File System

K-x

K-x

K-x

K-x

K-x

D-1

Store files

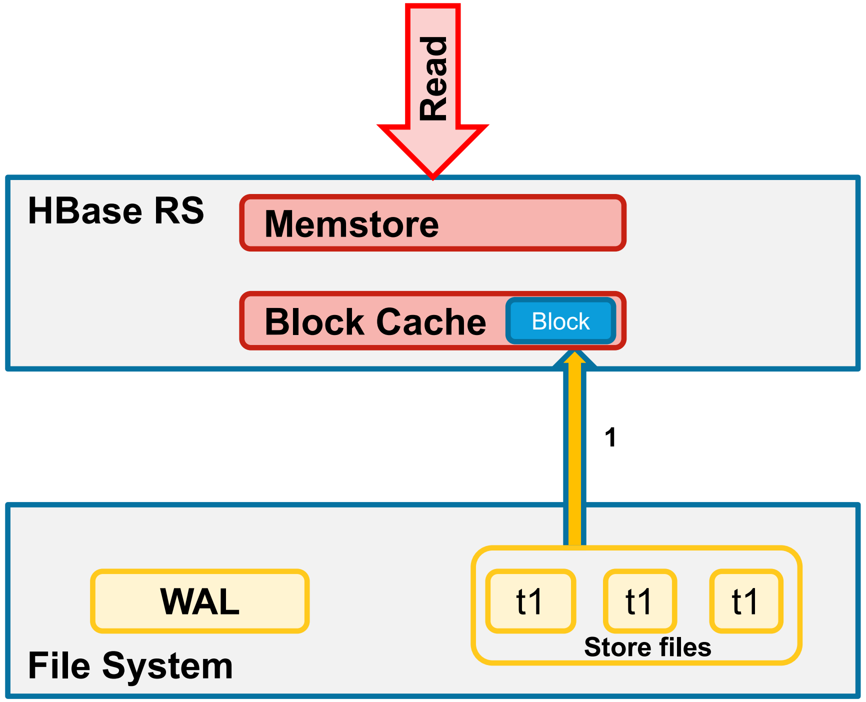
Compaction

File System

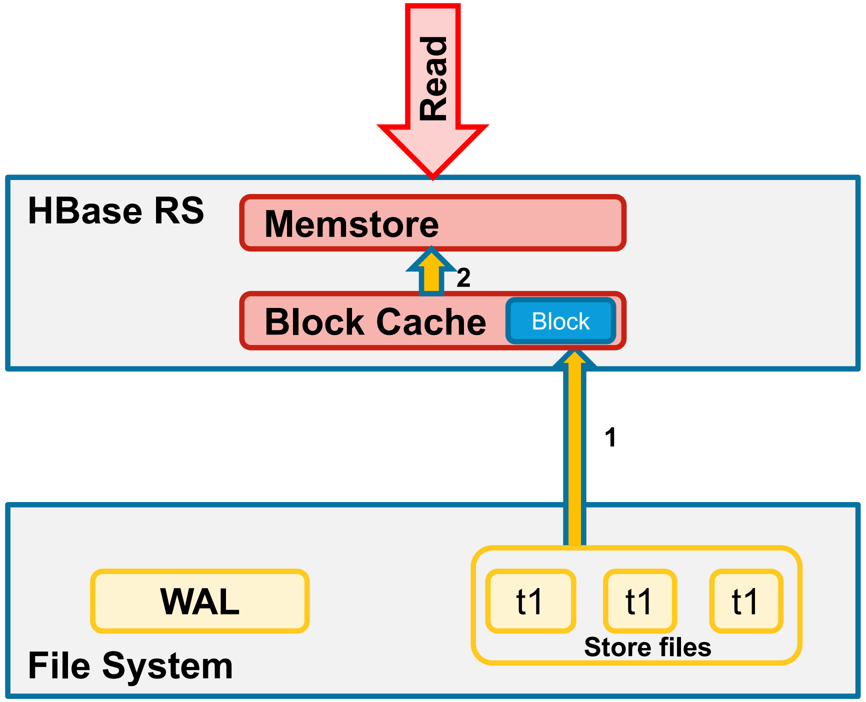
K-x

Store files

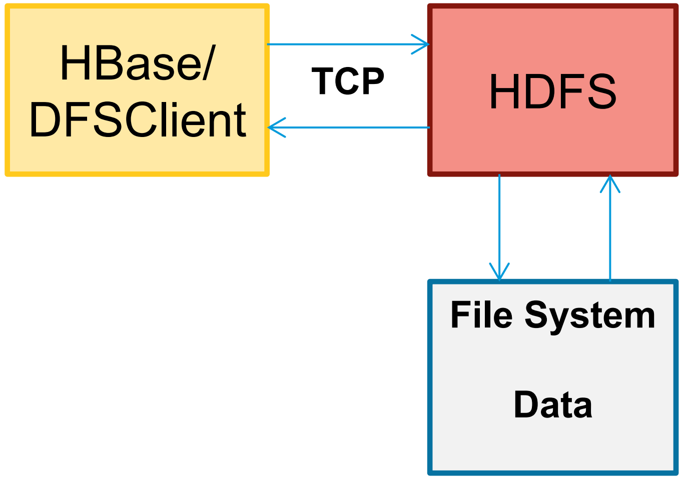
HBase Read


Read
HBase RS
Memstore
Block Cache
Block
1
WAL
t1
t1
t1
Store files
File System

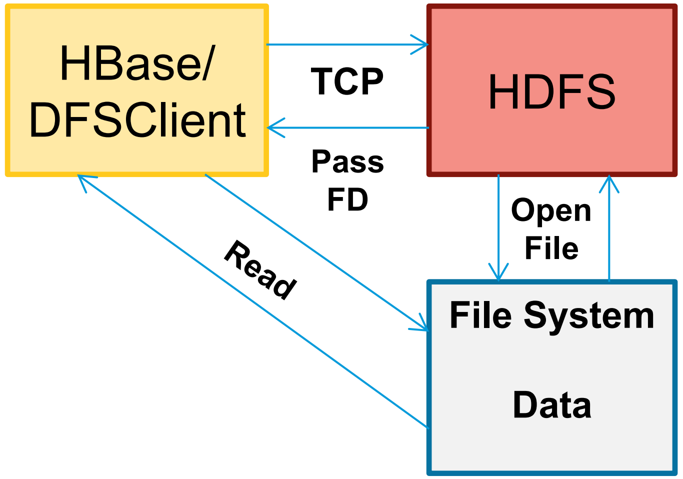
HBase Read


Read
HBase RS
Memstore
2
Block Cache
Block
1
WAL
t1
t1
t1
Store files
File System

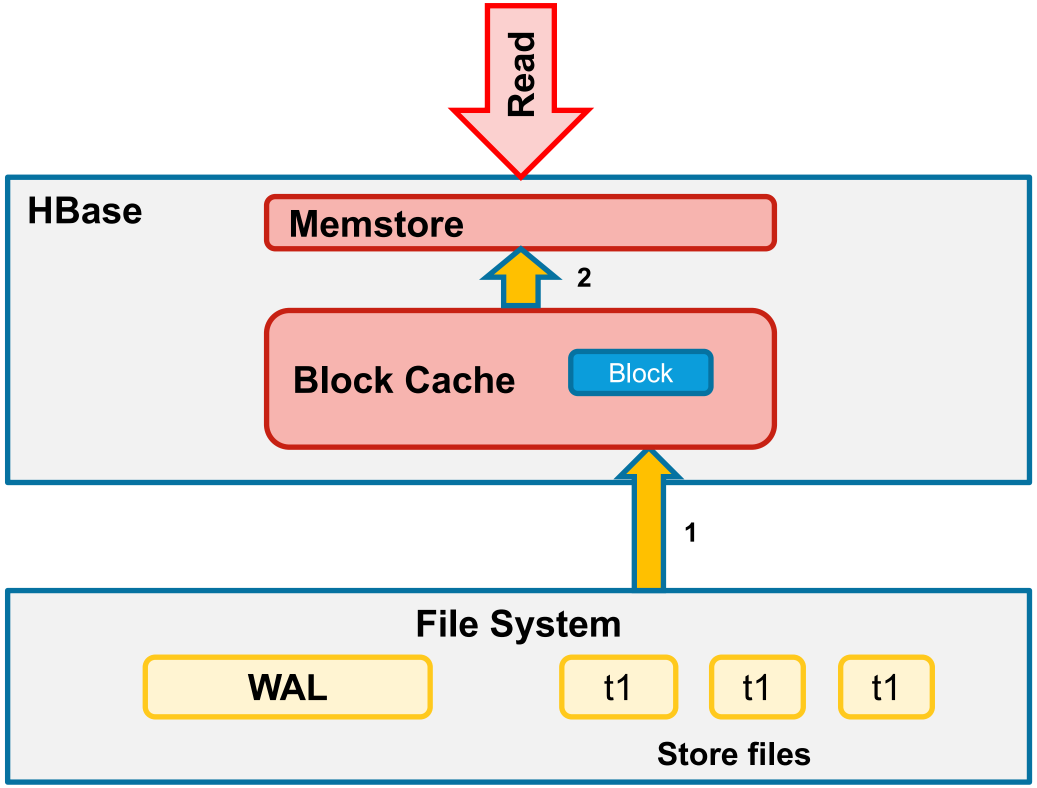
HDFS Read



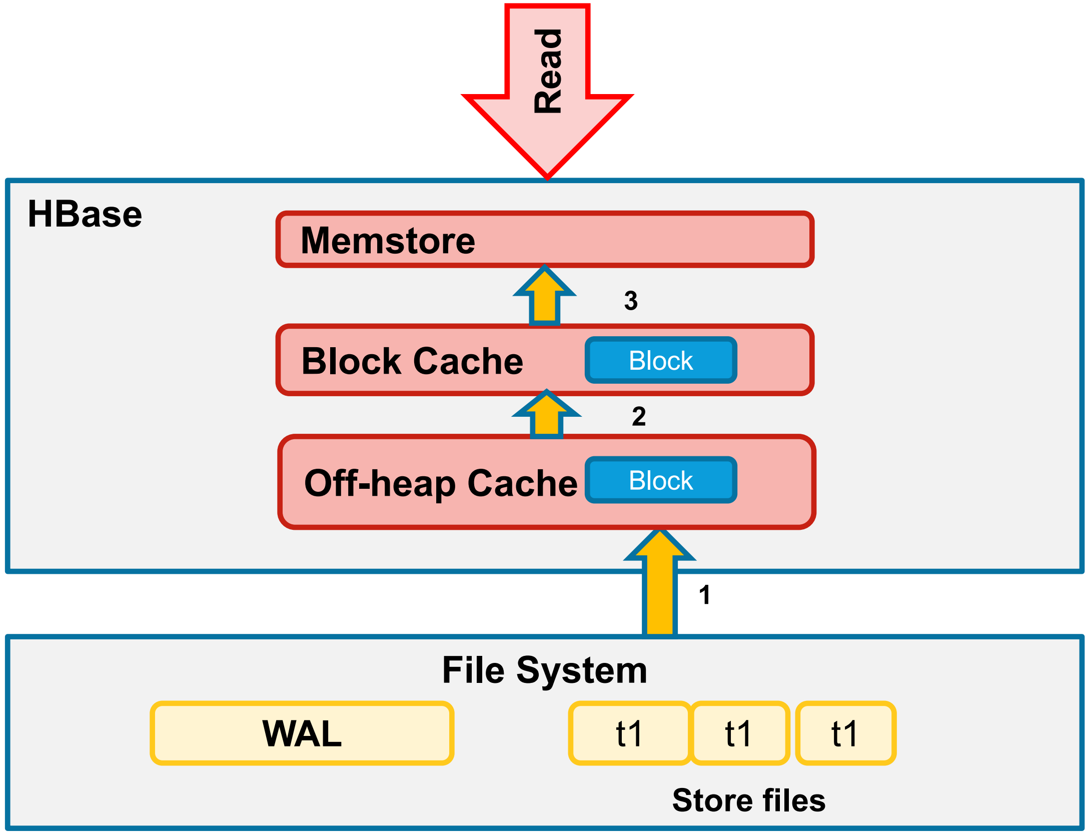
HDFS Short-Circuit Read



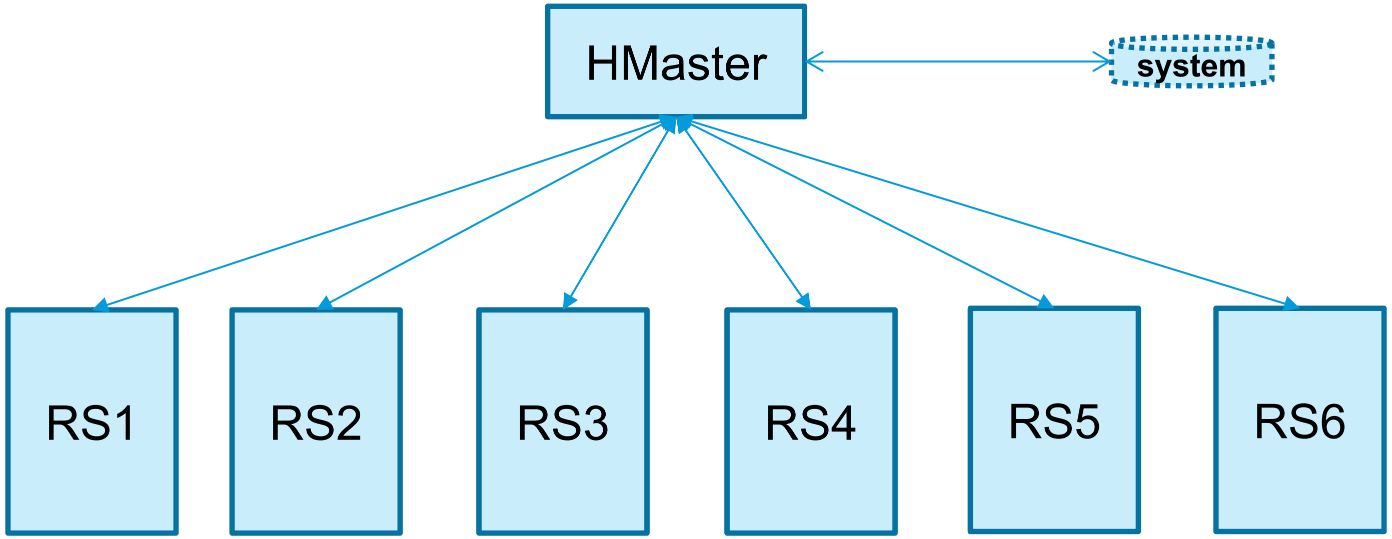
Large Read Cache



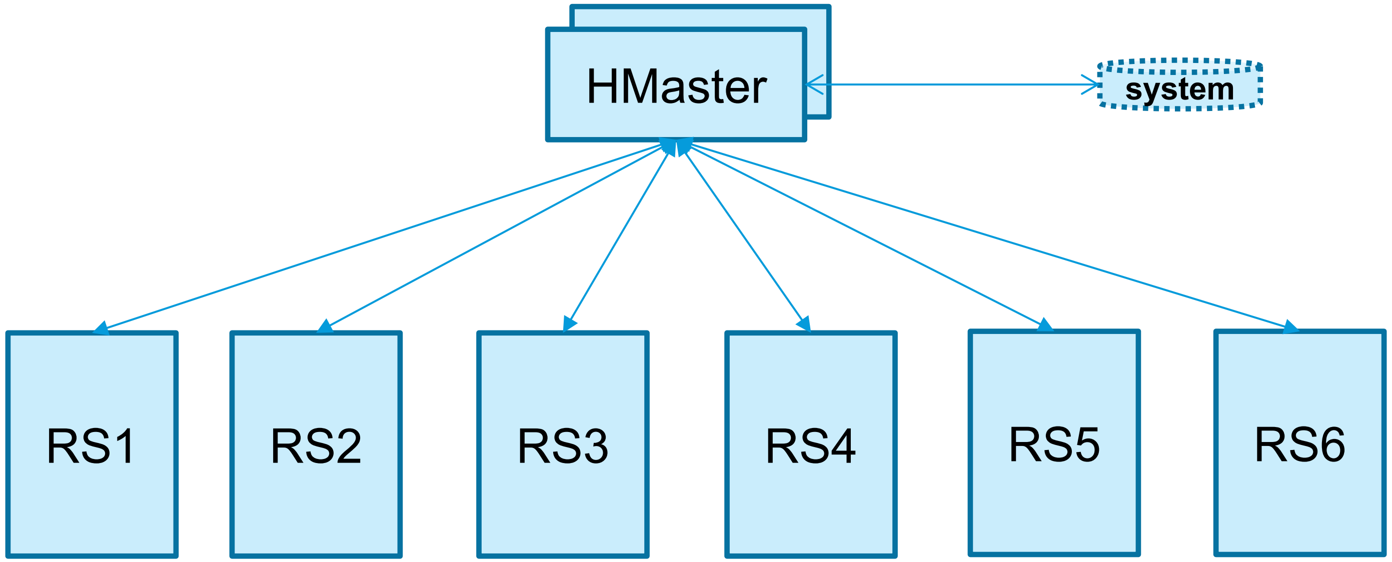
Large Read Cache



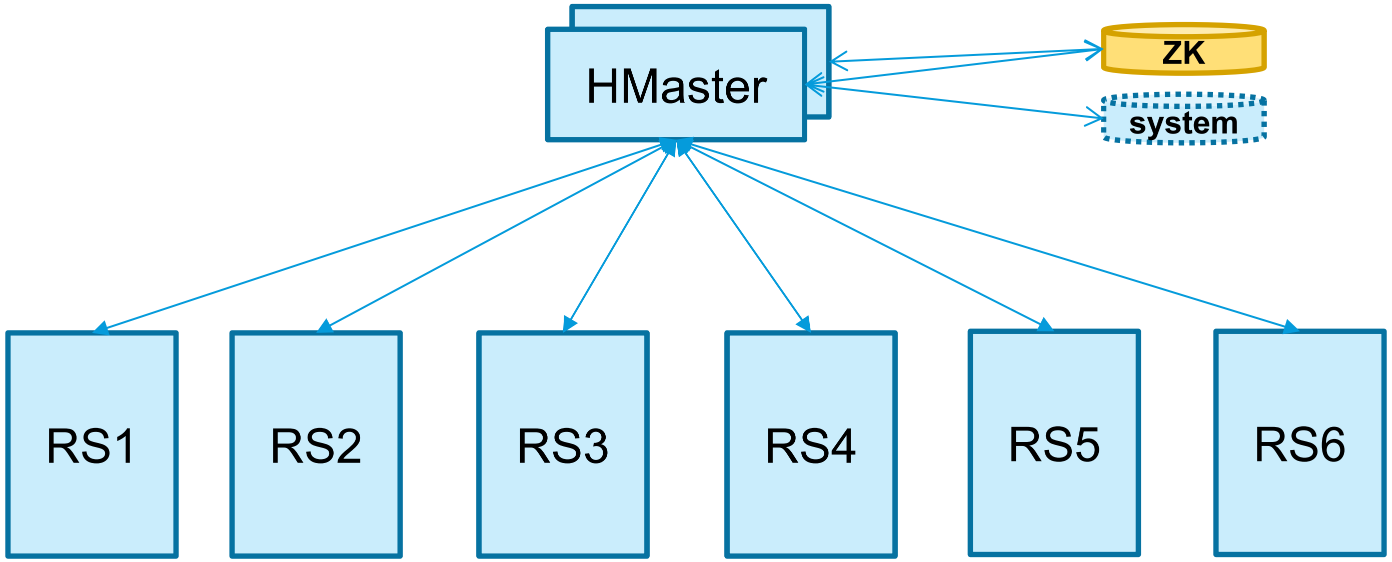
HBase Complete



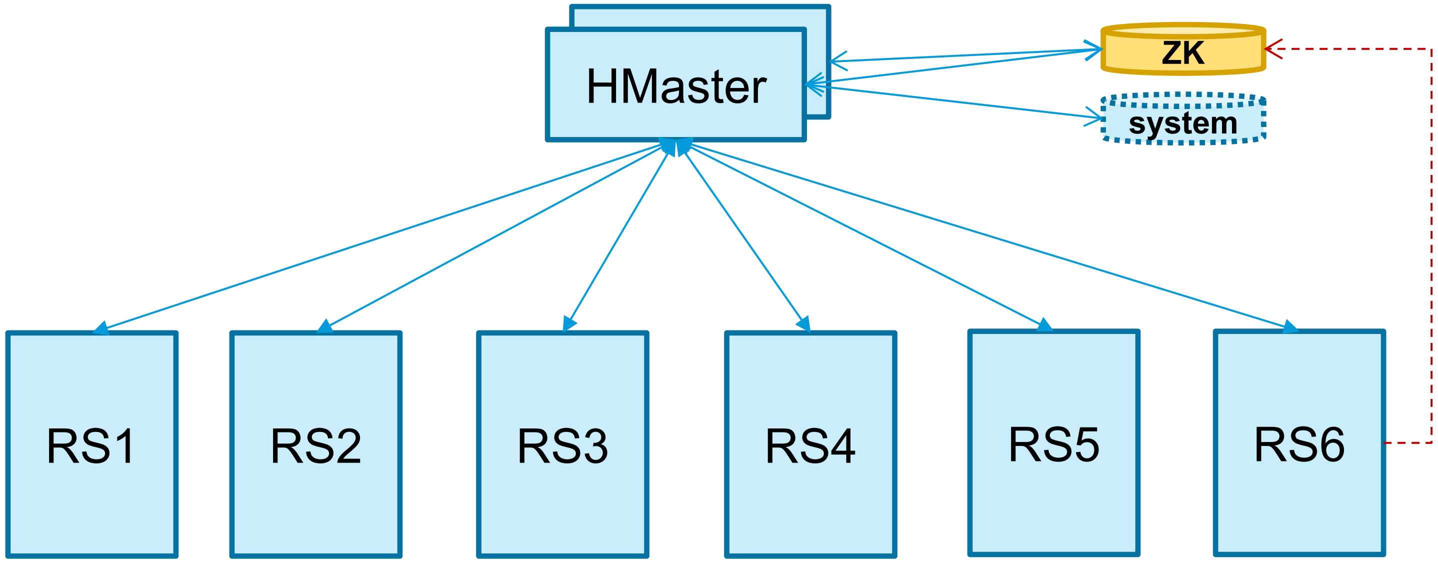
HBase Complete



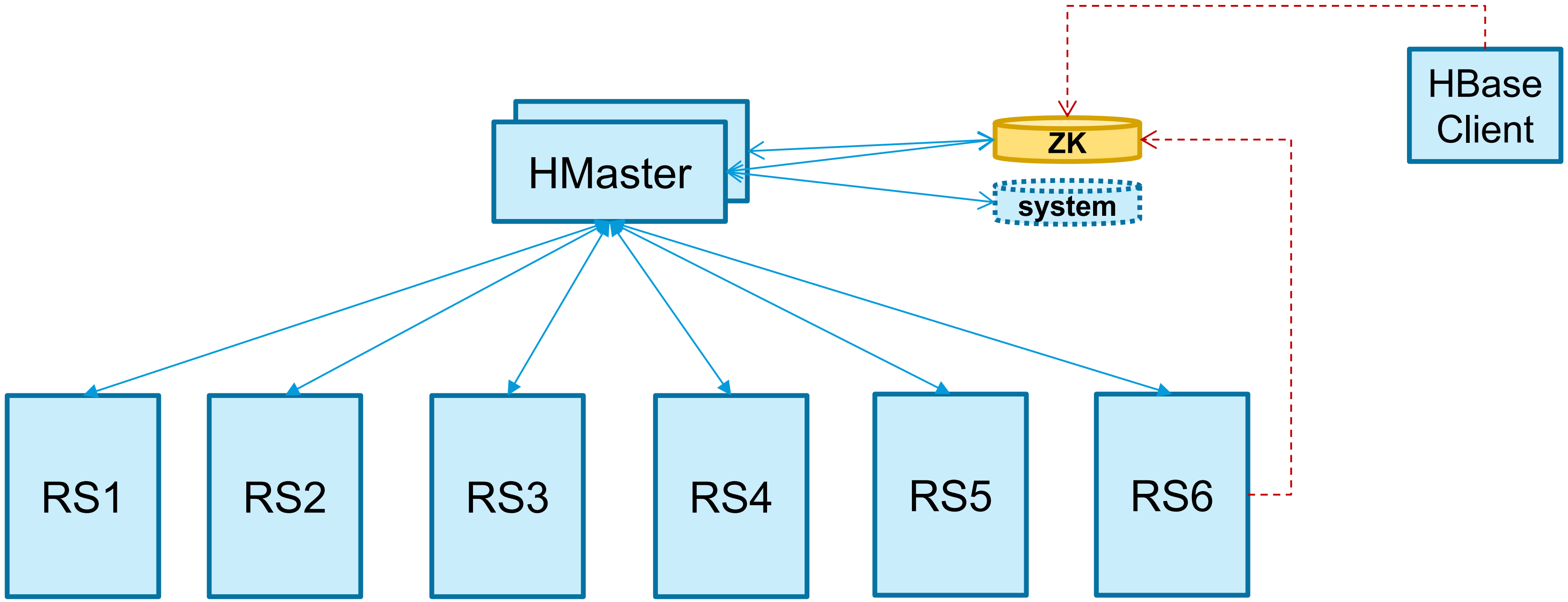
HBase Complete



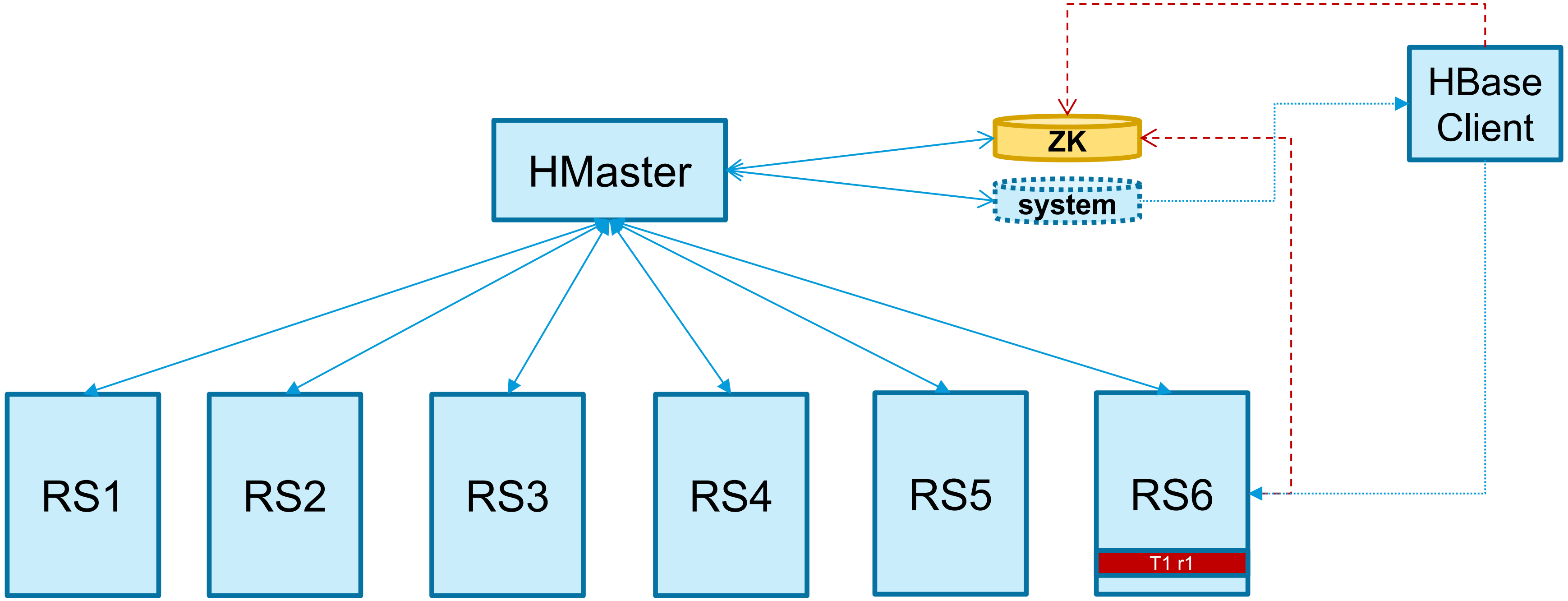
HBase Complete



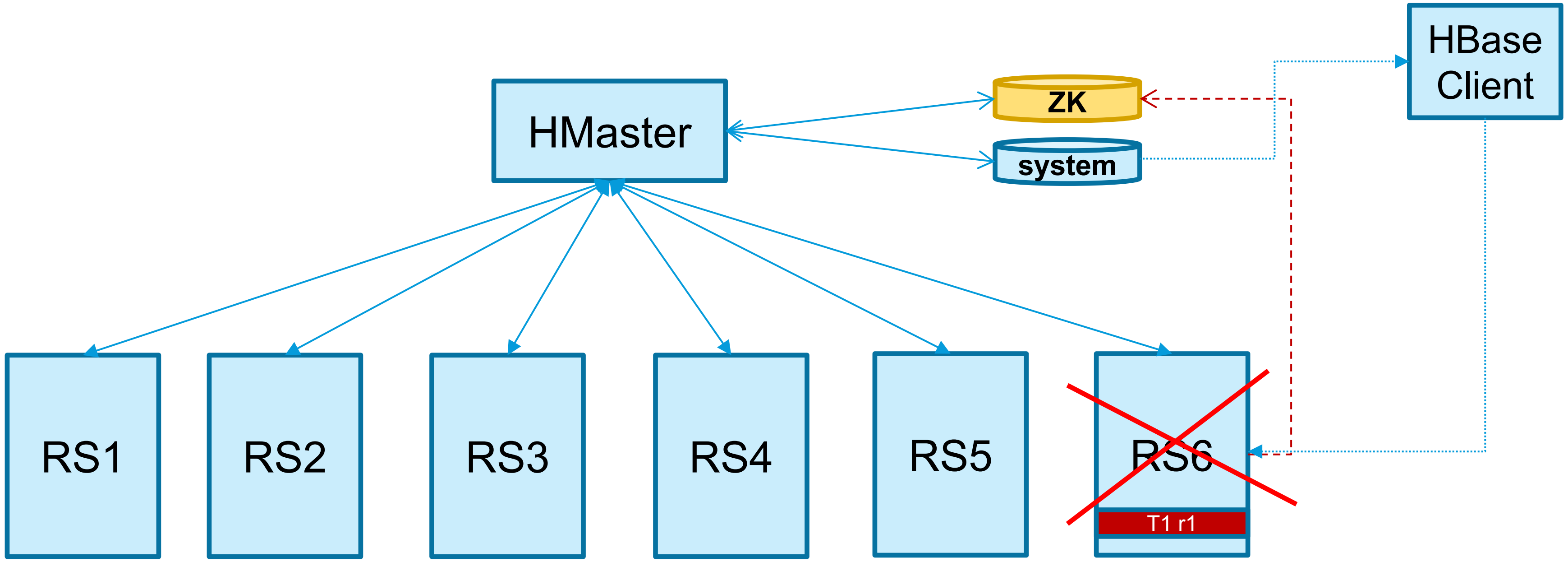
HBase Complete



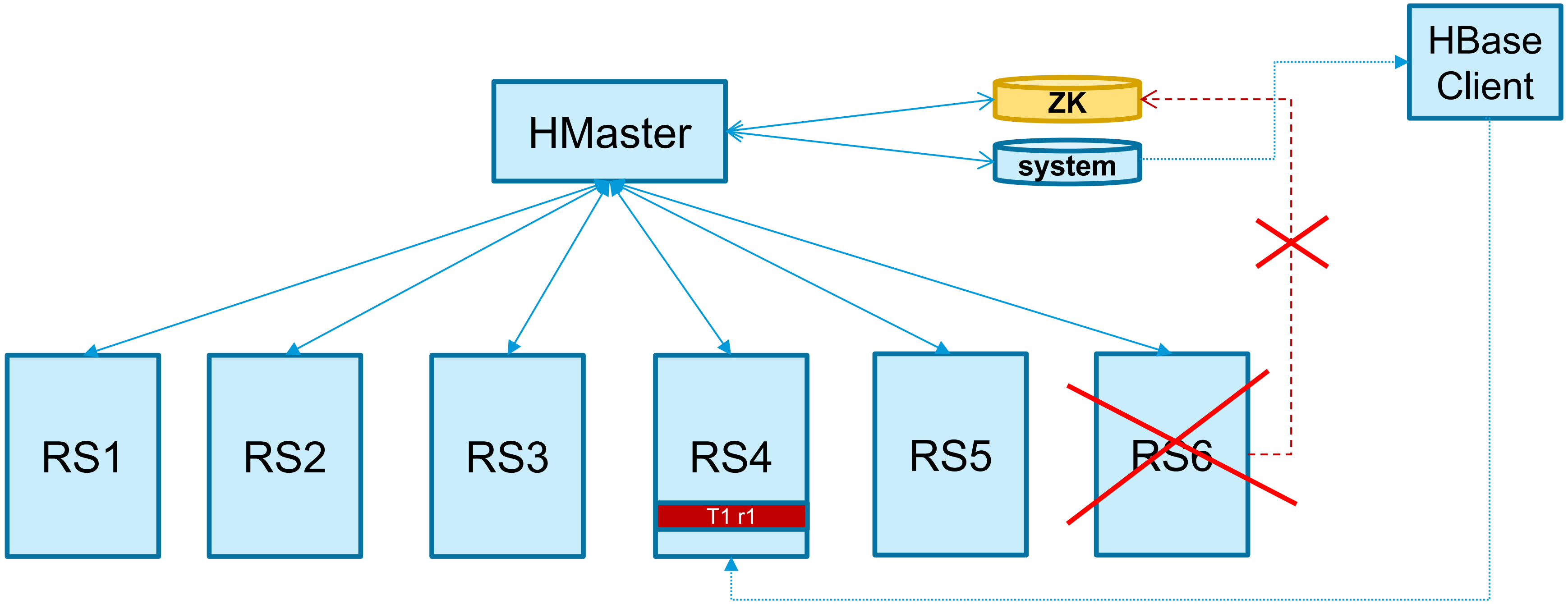
Region Server Failure



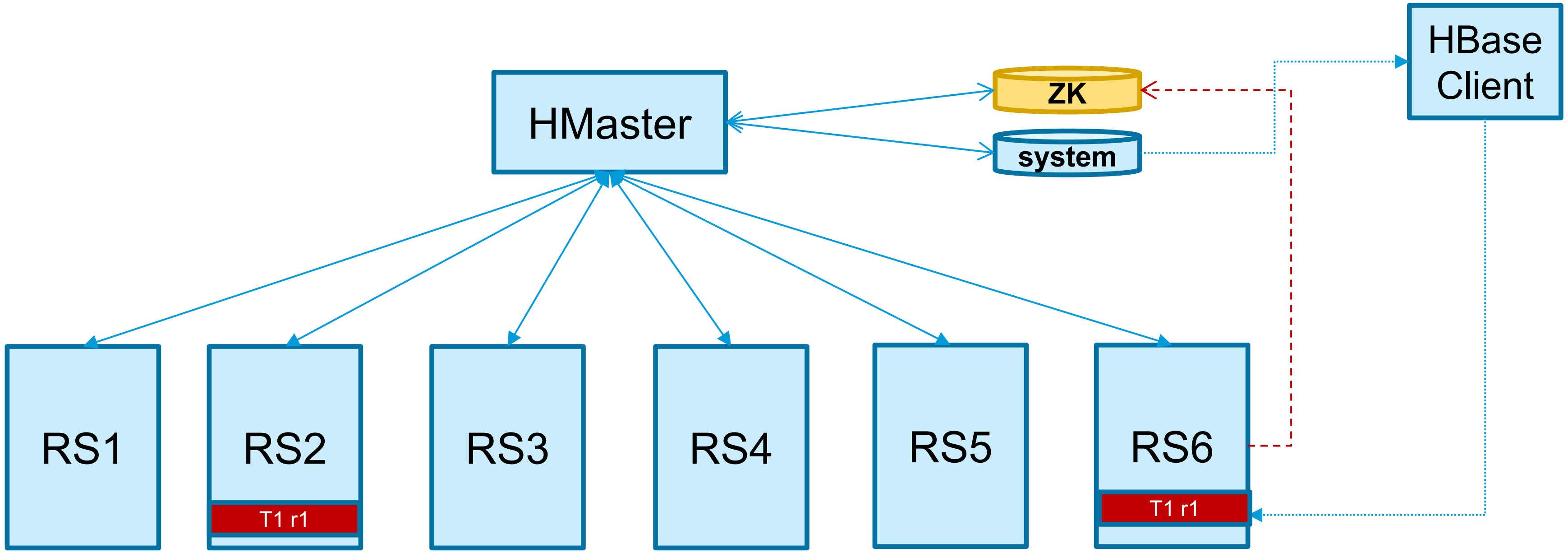
Region Server Failure



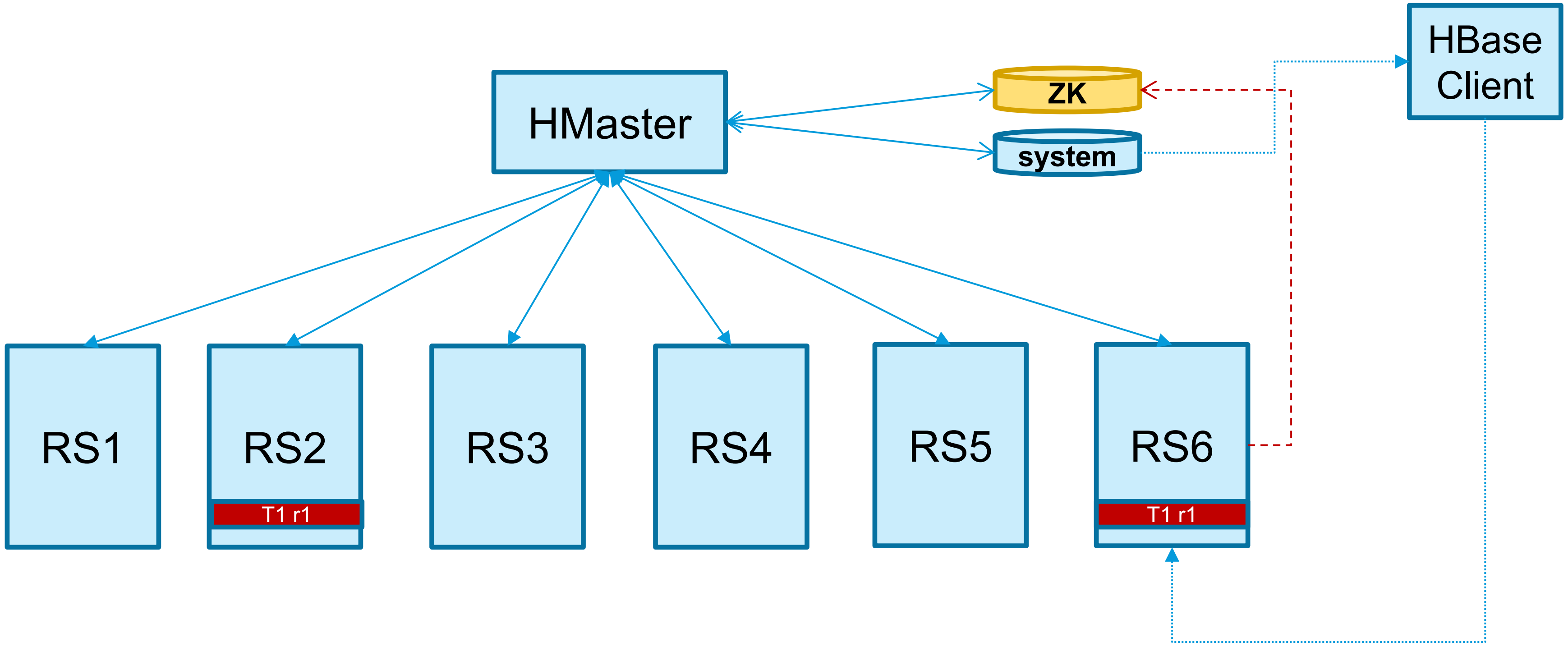
Region Server Failure



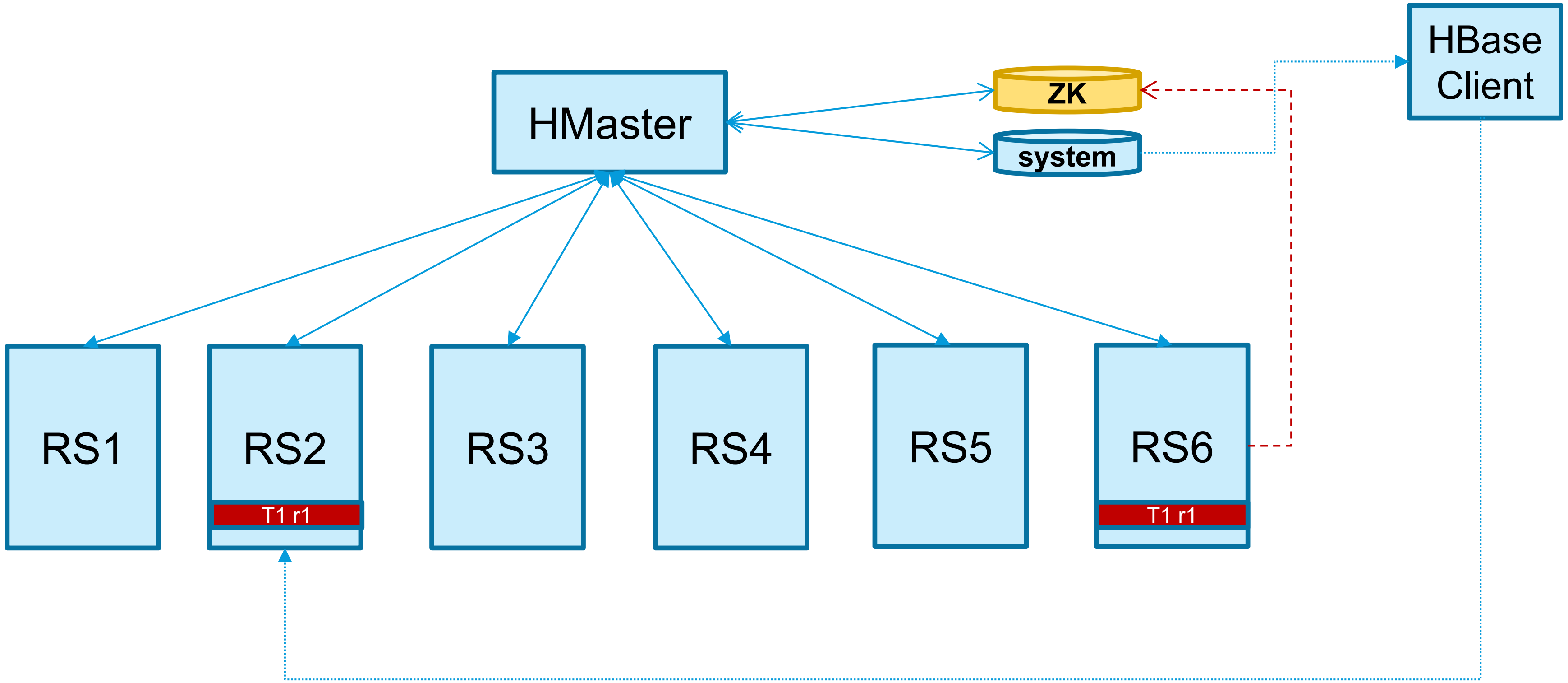
Region Replication



Region Replication

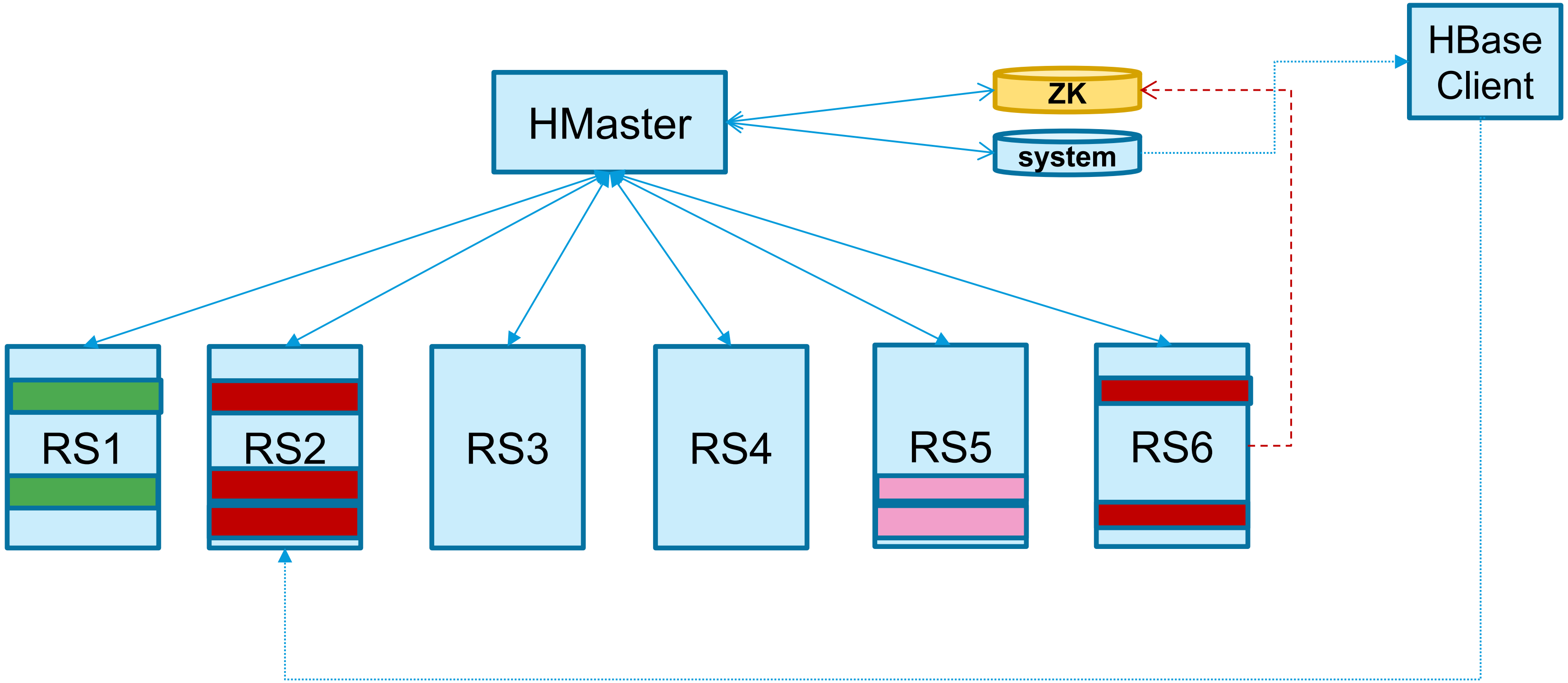


Region Replication

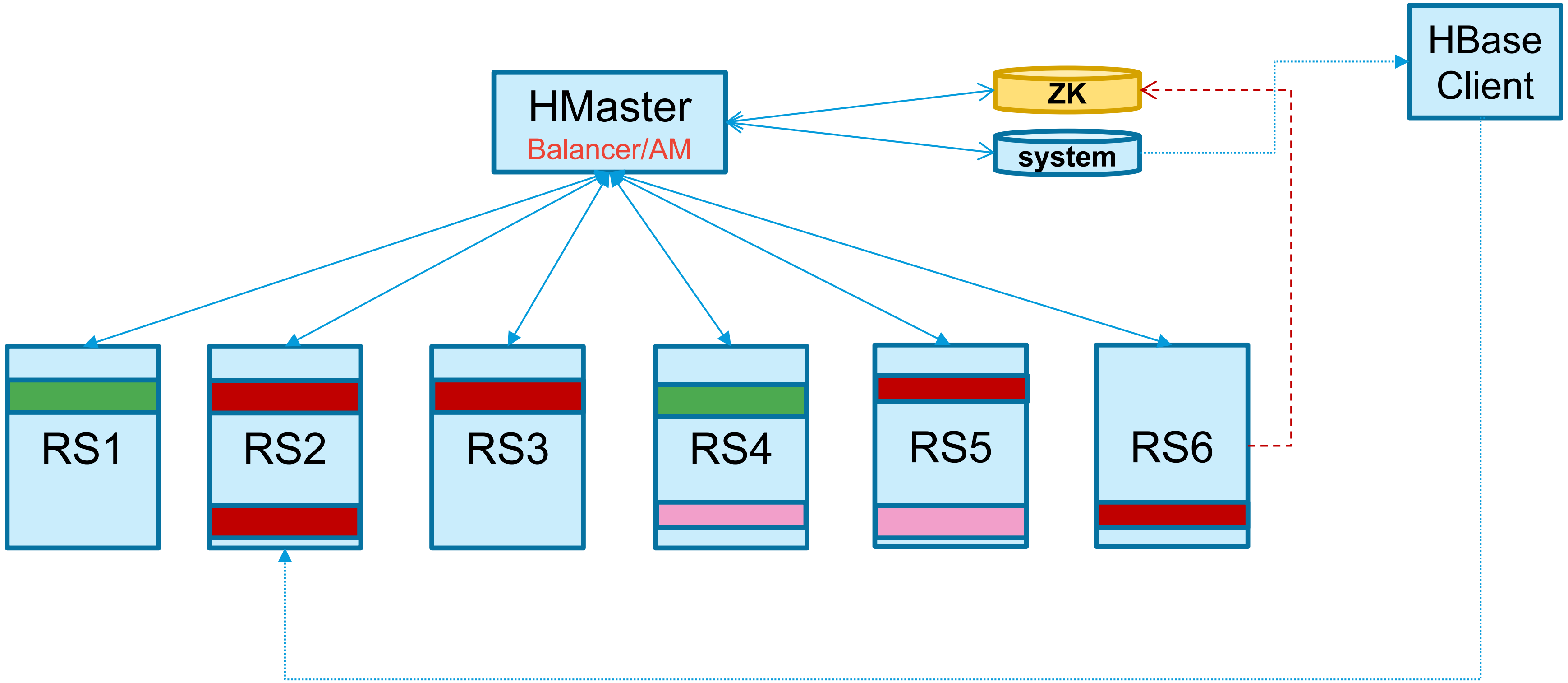


https://www.youtube.com/watch?v=I6S-Vbs9WsU

Load Balancing



Load Balancing



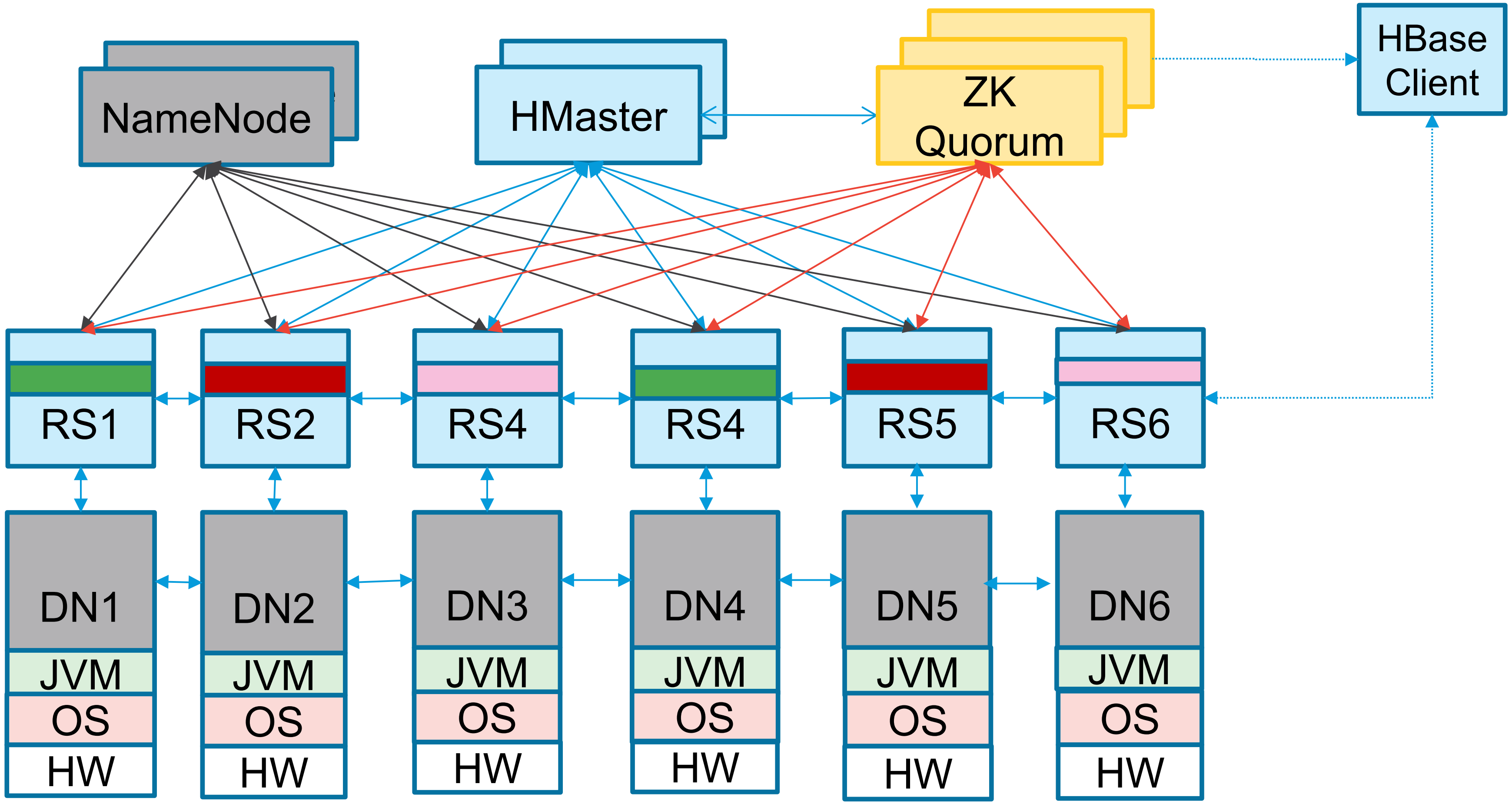
Balancer

- Region Count Cost
- Primary Region Count Cost
- Table Skew Cost
- Locality Cost
- Rack Locality Cost
- Region Replica Host Cost
- Region Replica Rack Cost
- Read Request Cost
- Write Request Cost
- Memstore Size Cost
- Storefile Size Cost
- Move Cost

Other Features

- HBase Replication
- HBase multi-tenancy support
  — https://www.youtube.com/watch?v=bZjz2G38Ju0
- HBase Co-processors and Filters
  — https://www.slideshare.net/Hadoop_Summit/hbase-coprocessors-uses-abuses-solutions

Operator View



ZooKeeper Availability

- ZK Quorum

- One leader and remaining followers
  — stat
  — ruok
  — mntr

- Test for availability
  — e.g., List children of a znode

HBase Availability

- Master Availability - http://hmaster-node:16010/jmx
  — "name" : "Hadoop:service=HBase,name=Master,sub=Server","tag.isActiveMaster" : "true"

- Dead RegionServers
  — "name" : "Hadoop:service=HBase,name=Master,sub=Server", "numDeadRegionServers" : 0

- Region In Transition
  — "name" : "Hadoop:service=HBase,name=Master,sub=AssignmentManger","ritCount" : 0

- Test for availability
  — e.g., Query system table by listing tables

HDFS Availability

- Namenode Availability - http://namenode-host:50070/jmx
  —"name" : "Hadoop:service=NameNode,name=FSNamesystem", "tag.HAState" : "active"
- Dead Datanodes
  — "name" : "Hadoop:service=NameNode,name=NameNodeInfo", "DeadNodes" : "{}“
- Missing Blocks
  — "name" : "Hadoop:service=NameNode,name=NameNodeInfo", "NumberOfMissingBlocks" : 0
- Percentage Used
  — "name" : "Hadoop:service=NameNode,name=NameNodeInfo", "PercentUsed" : 59
- Under replicated blocks
  — "name" : "Hadoop:service=NameNode,name=FSNamesystemState","UnderReplicatedBlocks":0
- Test for availability
  — e.g., Append data to a test file

HBase Performance

- RegionServer JMX metrics - http://rs-node:60300/jmx
  - "name“:"Hadoop:service=HBase,name=RegionServer,sub=Server“
    - Blockcache hit ratio
    - Request counts
    - Request response time
    - Compaction related metrics
    - Region count
    - Flush related metrics
    - Percentage of files local
    - Split related metrics
  - "name" : "Hadoop:service=HBase,name=RegionServer,sub=Tables",
    - Table level metrics

https://www.slideshare.net/MichaelStack4/hbaseconasia2018-track31-serving-billions-of-queries-in-millisecond-latencies

JVM

- GC – JMX Metrics
  - "name" : "java.lang:type=GarbageCollector,name=ParNew",
  - "name" : "java.lang:type=GarbageCollector,name=ConcurrentMarkSweep",
- GC Logging
  - -verbose:gc
  - -XX:+PrintHeapAtGC
  - -XX:+PrintGCDetails
  - -XX:+PrintGCTimeStamps
  - -XX:+PrintGCDateStamps
  - -XX:+PrintGCApplicationStoppedTime
  - -XX:+PrintClassHistogram
  - -XX:+PrintGCApplicationConcurrentTime
  - -XX:+PrintTenuringDistribution
  - -Xloggc:

OS/HW

- Memory
- CPU
- Disk
- Networking

Logs

- ZooKeeper Log
- HDFS
  - Namenode log
  - Datanode log
- HBase
  - Master log
  - RegionServer log
- OS
  - Syslog

Interacting with HBase

- HBase shell
  — DDL: create namespace/table, alter
  — Security: grant, revoke
  — DML: get, put, scan
  — Tools: assign, compact, balance
  — General: status
- HBase admin API
- HBase client API

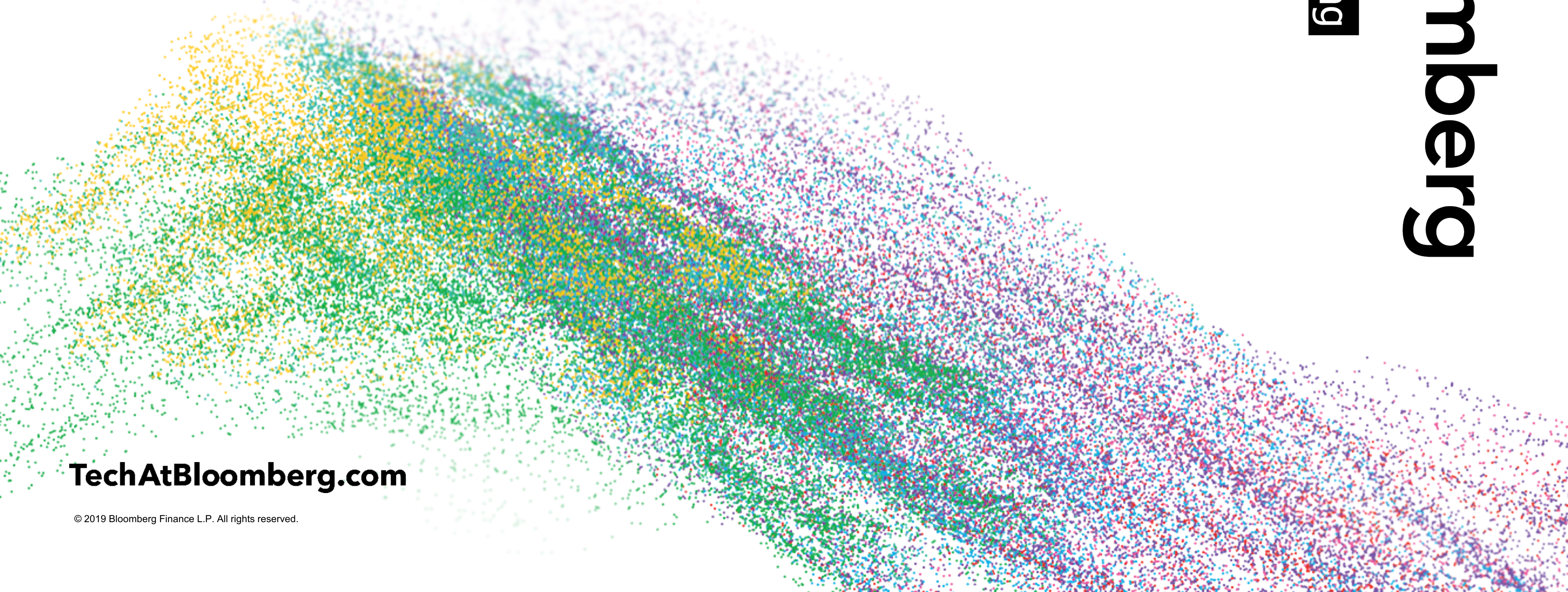
Data Backup / Restore

- Snapshot
  — hbase shell > snapshot 'table', 'table_mmddyy’
- Restore from snapshot
  — hbase shell > restore_snapshot 'table_mmddyy'
- Export Snapshot
  — $ hbase org.apache.hadoop.hbase.snapshot.ExportSnapshot
- CopyTable
  — hbase org.apache.hadoop.hbase.mapreduce.CopyTable

Thank You!

Acknowledgement: Apache HBase Community
Reference: http://hbase.apache.org
Connect with Hadoop Team: hadoop@bloomberg.net

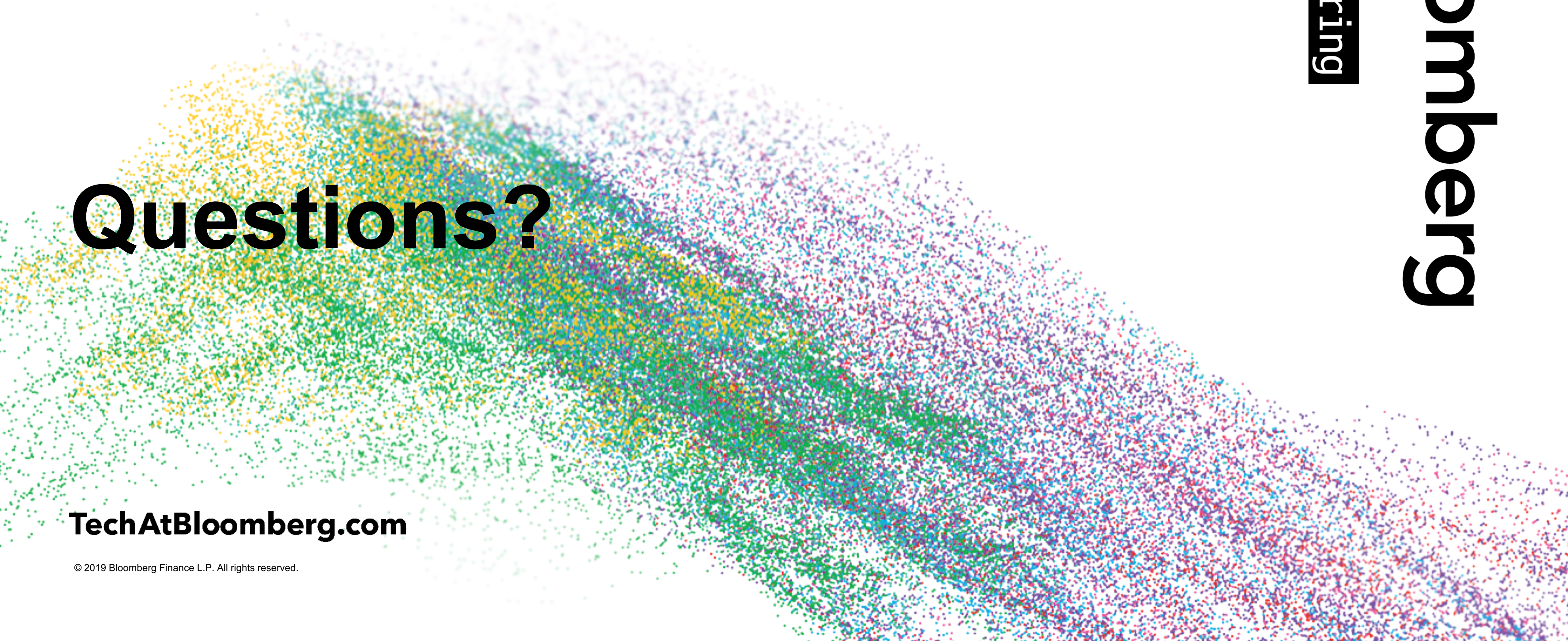
Bloomberg
Engineering



TechAtBloomberg.com

We are hiring!

https://www.bloomberg.com/careers



Questions?

Bloomberg Engineering

TechAtBloomberg.com