HBase
Operations on EC2

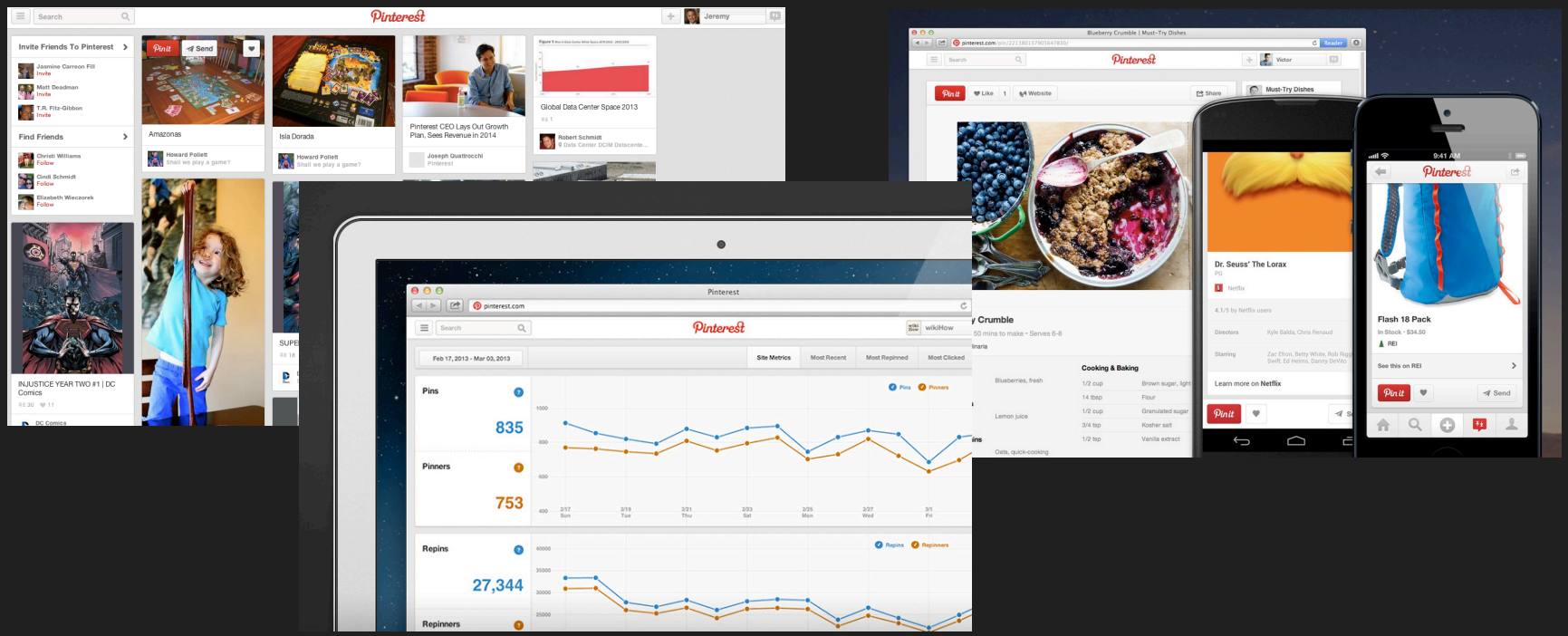
Jeremy Carroll
Big Data Gurus

Pinterest **Engineering**

Overview

- Deployment Strategies for EC2
- Validating Design
- Production Support

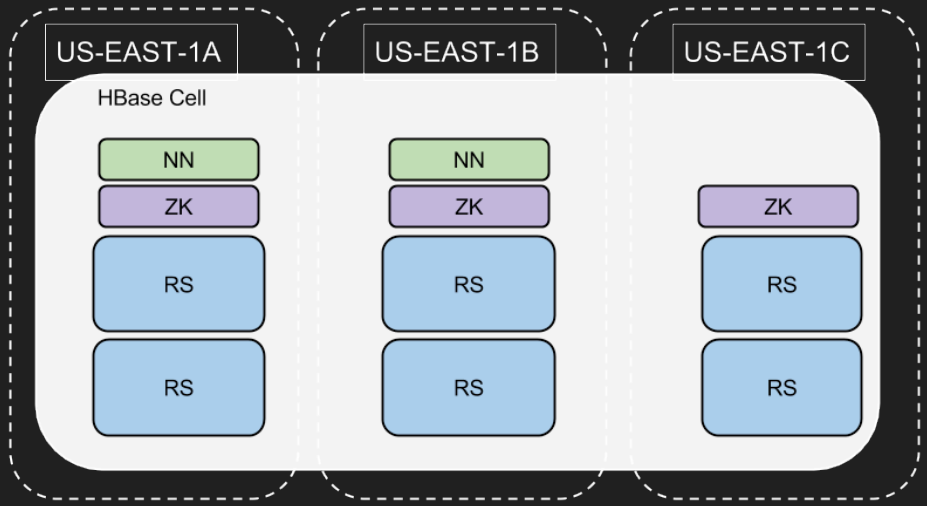
Powered by HBase



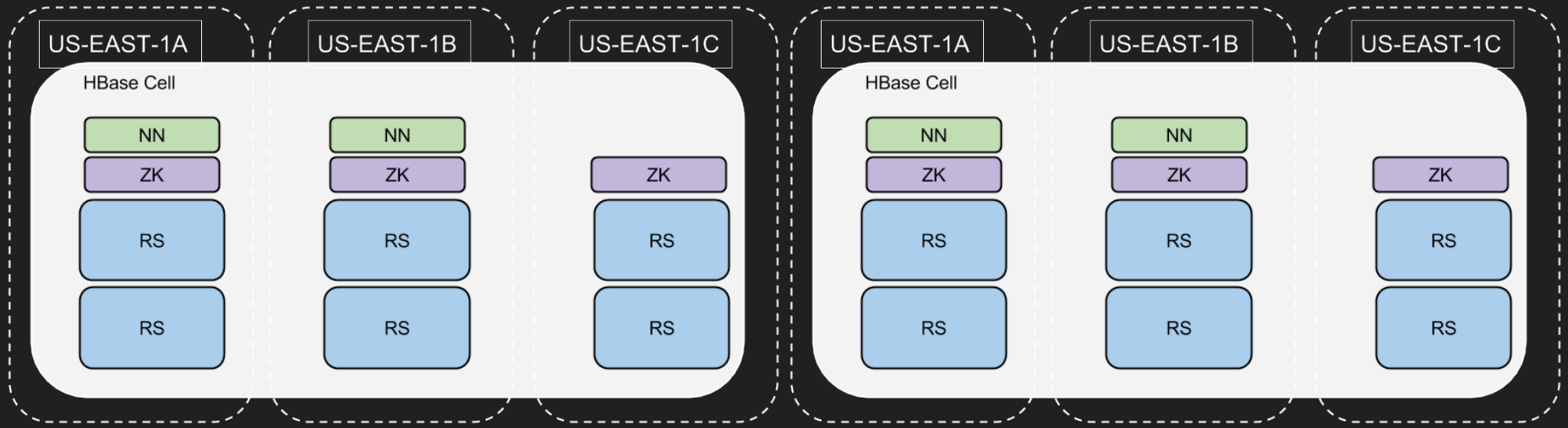
Lets Deploy

- First Question Asked
- Rack Locality?
- Cloud Concepts

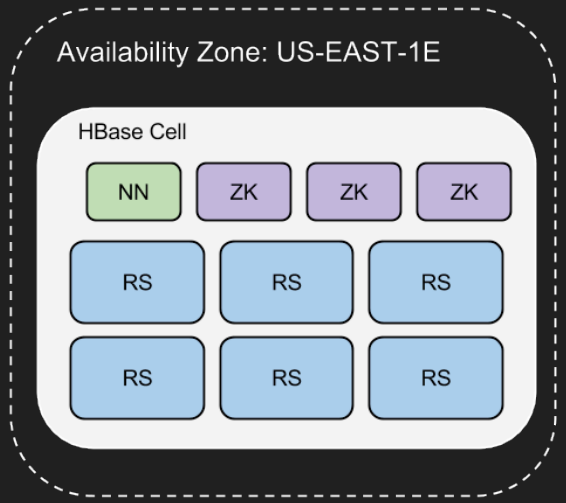
High Availability



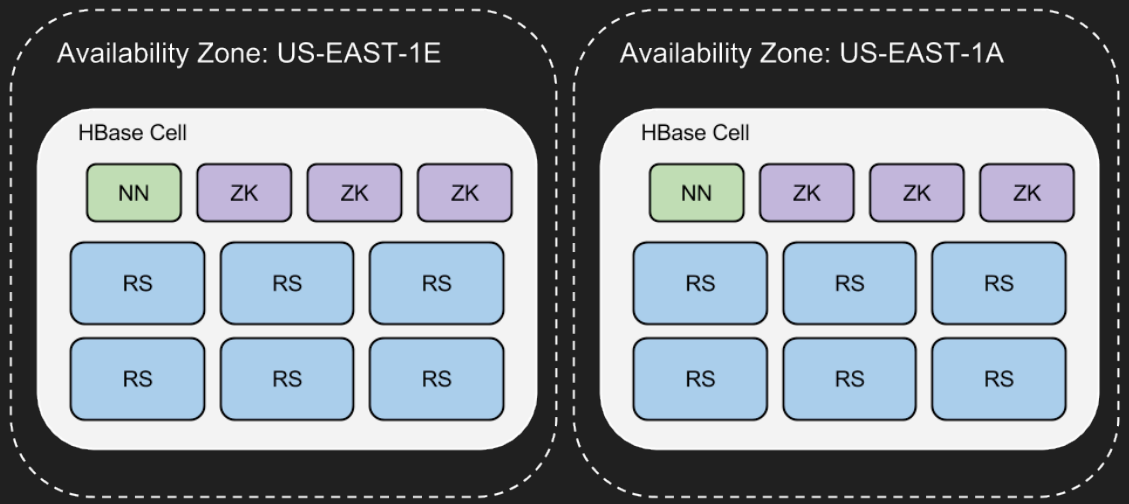
Logical Separation



Cell Based


Availability Zone: US-EAST-1E
HBase Cell
NN
ZK
ZK
ZK
RS
RS
RS
RS
RS
RS

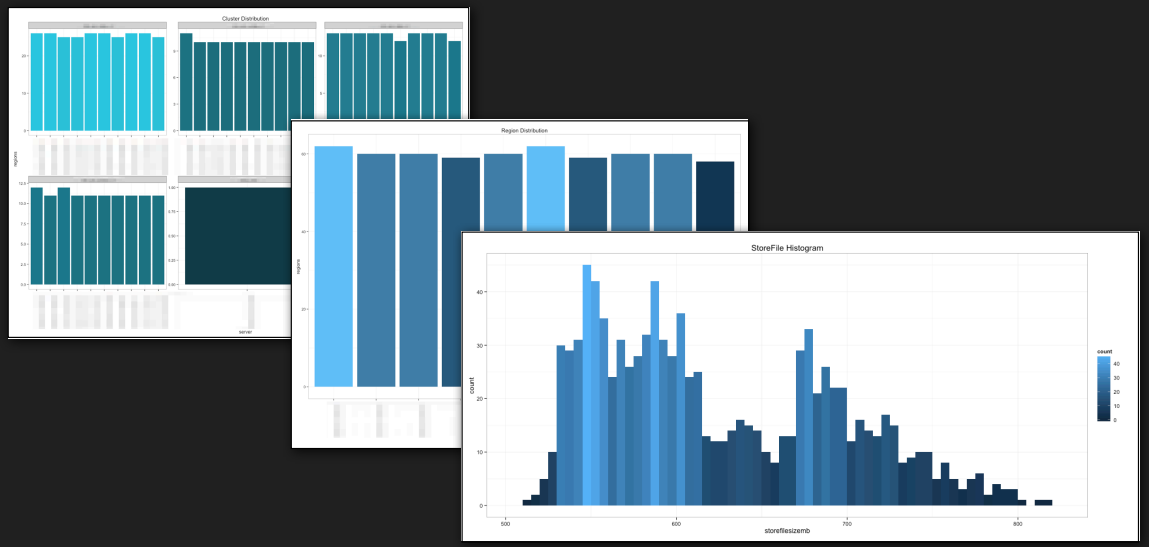
Logical Separation



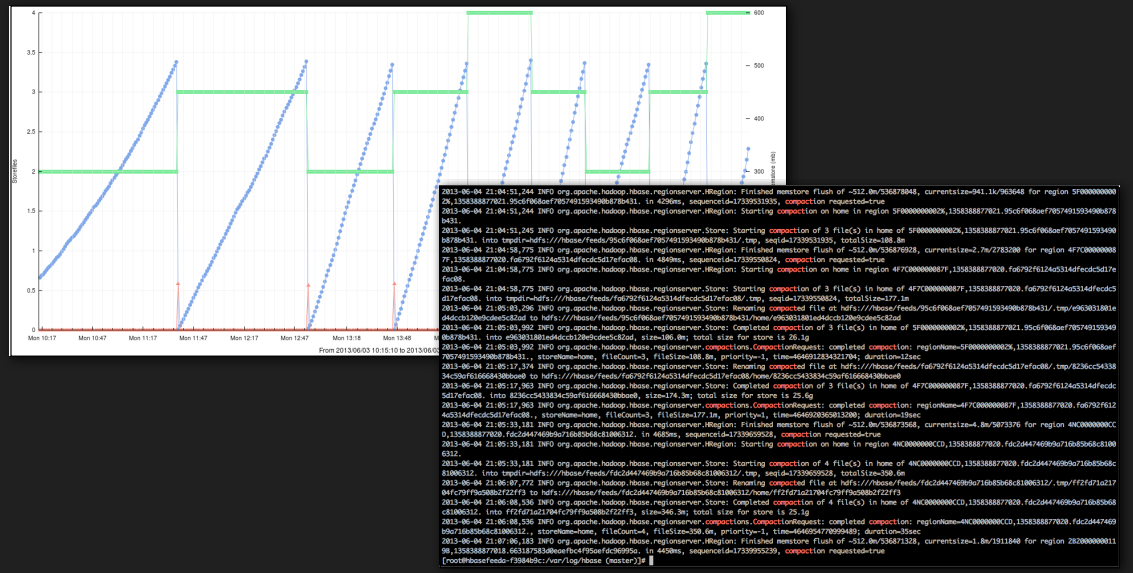
Does This Work?

- Schema Design
- Hot Spots
- Load Testing
- Tools

Does This Work?



Compaction


Mon 10:17
Mon 10:47
Mon 11:17
Mon 11:47
Mon 12:17
Mon 12:47
Mon 13:18
Mon 13:48
From 2013/06/03 10:15:10 to 2013/06/03
2013-06-04 21:04:51,244 INFO org.apache.hadoop.hbase.regionserver.HRegion: Finished memstore flush of ~512.0m/536878048, currentsize=941.1k/963648 for region 5F0000000002K,1358388877021.95c6f068aef7057491593490b878b431. in 4296ms, sequenceid=17339531935, compaction requested=true
2013-06-04 21:04:51,244 INFO org.apache.hadoop.hbase.regionserver.HRegion: Starting compaction on home in region 5F0000000002K,1358388877021.95c6f068aef7057491593490b878b431.
2013-06-04 21:04:51,245 INFO org.apache.hadoop.hbase.regionserver.Store: Starting compaction of 3 file(s) in home of 5F0000000002K,1358388877021.95c6f068aef7057491593490b878b431. into tmpdir=hdfs:///hbase/feeds/95c6f068aef7057491593490b878b431/.tmp, seqid=17339531935, totalSize=188.8m
2013-06-04 21:04:58,775 INFO org.apache.hadoop.hbase.regionserver.HRegion: Finished memstore flush of ~512.0m/536876928, currentsize=2.7m/2783200 for region 4F7C000000087F,1358388877020.fa6792f6124a5314dfecdc5d17efac08. in 4849ms, sequenceid=17339550824, compaction requested=true
2013-06-04 21:04:58,775 INFO org.apache.hadoop.hbase.regionserver.HRegion: Starting compaction on home in region 4F7C000000087F,1358388877020.fa6792f6124a5314dfecdc5d17efac08.
2013-06-04 21:04:58,775 INFO org.apache.hadoop.hbase.regionserver.Store: Starting compaction of 3 file(s) in home of 4F7C000000087F,1358388877020.fa6792f6124a5314dfecdc5d17efac08. into tmpdir=hdfs:///hbase/feeds/fa6792f6124a5314dfecdc5d17efac08/.tmp, seqid=17339550824, totalSize=177.1m
2013-06-04 21:05:03,296 INFO org.apache.hadoop.hbase.regionserver.Store: Renaming compacted file at hdfs:///hbase/feeds/95c6f068aef7057491593490b878b431/.tmp/e963031801ed4dccb120e9cdee5c82ad to hdfs:///hbase/feeds/95c6f068aef7057491593490b878b431/home/e963031801ed4dccb120e9cdee5c82ad
2013-06-04 21:05:03,992 INFO org.apache.hadoop.hbase.regionserver.Store: Completed compaction of 3 file(s) in home of 5F0000000002K,1358388877021.95c6f068aef7057491593490b878b431. into e963031801ed4dccb120e9cdee5c82ad, size=186.0m; total size for store is 26.1g
2013-06-04 21:05:03,992 INFO org.apache.hadoop.hbase.regionserver.compactions.CompactionRequest: completed compaction: regionName=5F0000000002K,1358388877021.95c6f068aef7057491593490b878b431., storeName=home, fileCount=3, fileSize=188.8m, priority=-1, time=4646912834321704; duration=12sec
2013-06-04 21:05:17,174 INFO org.apache.hadoop.hbase.regionserver.Store: Renaming compacted file at hdfs:///hbase/feeds/fa6792f6124a5314dfecdc5d17efac08/.tmp/8236cc5433834c59af616668430bbae0 to hdfs:///hbase/feeds/fa6792f6124a5314dfecdc5d17efac08/home/8236cc5433834c59af616668430bbae0
2013-06-04 21:05:17,963 INFO org.apache.hadoop.hbase.regionserver.Store: Completed compaction of 3 file(s) in home of 4F7C000000087F,1358388877020.fa6792f6124a5314dfecdc5d17efac08. into 8236cc5433834c59af616668430bbae0, size=174.3m; total size for store is 25.6g
2013-06-04 21:05:17,963 INFO org.apache.hadoop.hbase.regionserver.compactions.CompactionRequest: completed compaction: regionName=4F7C000000087F,1358388877020.fa6792f6124a5314dfecdc5d17efac08., storeName=home, fileCount=3, fileSize=177.1m, priority=1, time=4646920365013200; duration=19sec
2013-06-04 21:05:33,181 INFO org.apache.hadoop.hbase.regionserver.HRegion: Finished memstore flush of ~512.0m/536873568, currentsize=4.8m/5073376 for region 4NC0000000CCD,1358388877020.fdc2d447469b9a716b85b68c81006312. in 4685ms, sequenceid=17339659528, compaction requested=true
2013-06-04 21:05:33,181 INFO org.apache.hadoop.hbase.regionserver.HRegion: Starting compaction on home in region 4NC0000000CCD,1358388877020.fdc2d447469b9a716b85b68c81006312.
2013-06-04 21:05:33,181 INFO org.apache.hadoop.hbase.regionserver.Store: Starting compaction of 4 file(s) in home of 4NC0000000CCD,1358388877020.fdc2d447469b9a716b85b68c81006312. into tmpdir=hdfs:///hbase/feeds/fdc2d447469b9a716b85b68c81006312/.tmp, seqid=17339659528, totalSize=350.6m
2013-06-04 21:06:07,772 INFO org.apache.hadoop.hbase.regionserver.Store: Renaming compacted file at hdfs:///hbase/feeds/fdc2d447469b9a716b85b68c81006312/.tmp/ff2fd71a2174f4fc79ff9a588b2f22ff3 to hdfs:///hbase/feeds/fdc2d447469b9a716b85b68c81006312/home/ff2fd71a2174f4fc79ff9a588b2f22ff3
2013-06-04 21:06:08,536 INFO org.apache.hadoop.hbase.regionserver.Store: Completed compaction of 4 file(s) in home of 4NC0000000CCD,1358388877020.fdc2d447469b9a716b85b68c81006312. into ff2fd71a2174f4fc79ff9a588b2f22ff3, size=346.3m; total size for store is 25.1g
2013-06-04 21:06:08,536 INFO org.apache.hadoop.hbase.regionserver.compactions.CompactionRequest: completed compaction: regionName=4NC0000000CCD,1358388877020.fdc2d447469b9a716b85b68c81006312., storeName=home, fileCount=4, fileSize=350.6m, priority=-1, time=4646954479999489; duration=35sec
2013-06-04 21:07:06,183 INFO org.apache.hadoop.hbase.regionserver.HRegion: Finished memstore flush of ~512.0m/536871328, currentsize=1.8m/1911840 for region 2820000000119 8,1358388877018.663187583d0eaefbc4f95aefdc96995a. in 4450ms, sequenceid=17339955239, compaction requested=true
[root@hbasefeeds-i-f3884b3c:/var/log/hbase (master)]#

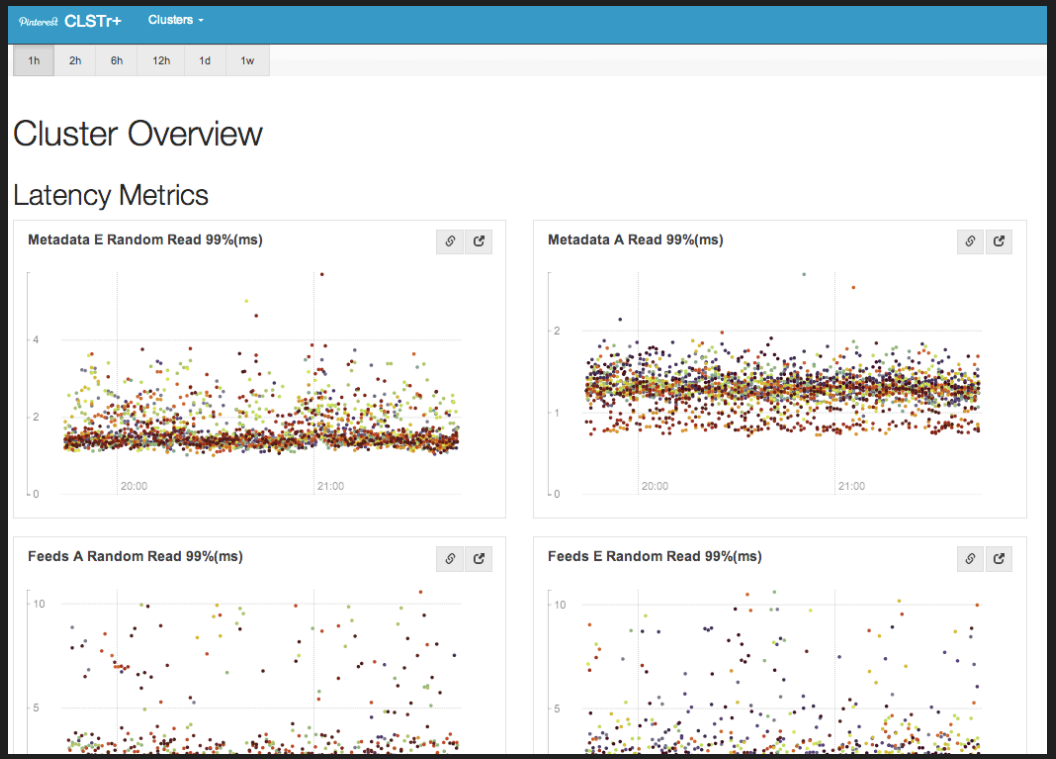
OpenTSDB


TSD
Time Series Database
Graph
Stats
Logs
Version
4057346 points retrieved, 196252 points plotted
907707 points retrieved, 63846 points plotted
907707 points retrieved, 817005 points plotted

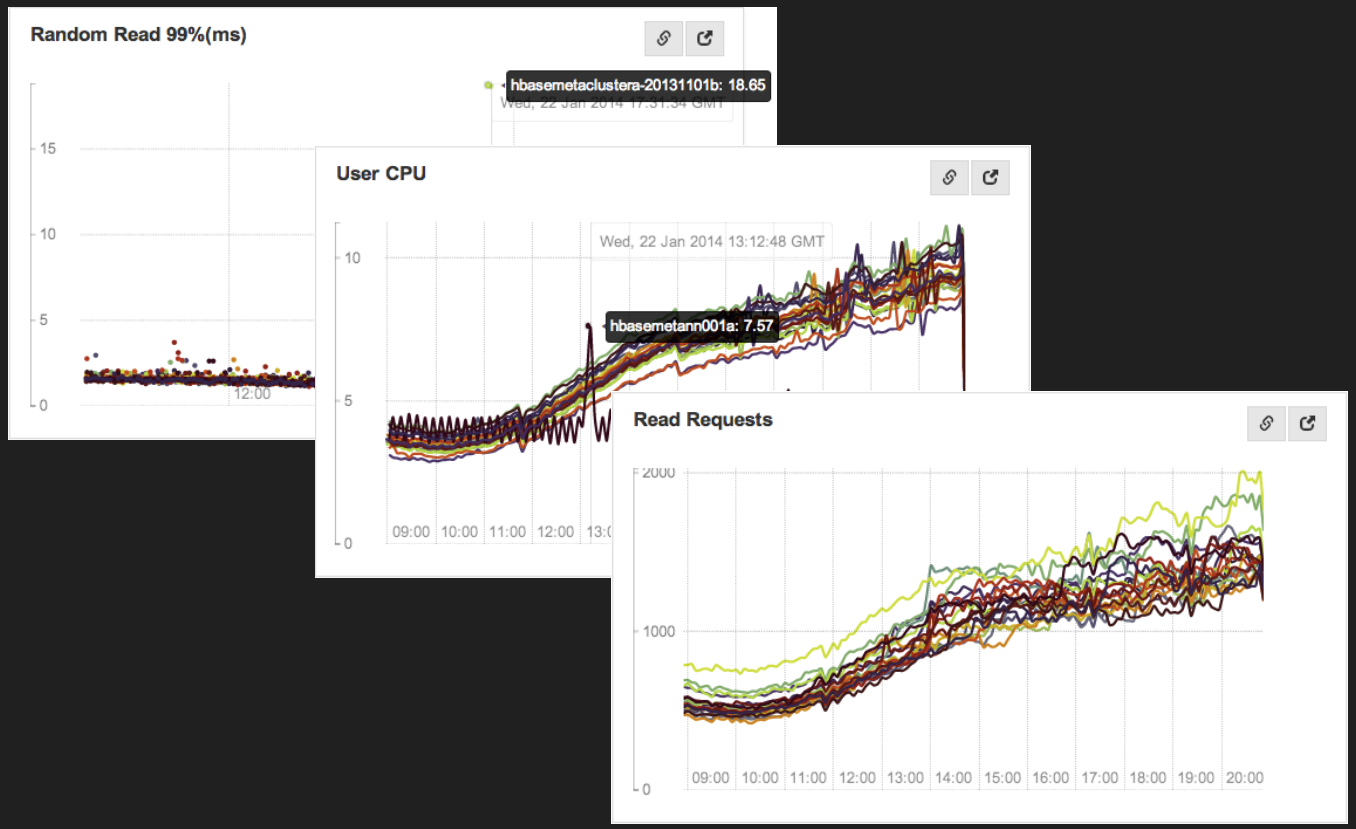
Production

- Monitoring
- Alerting
- Health

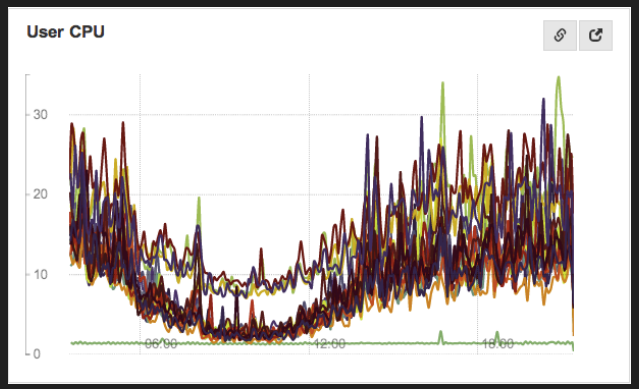
Monitoring

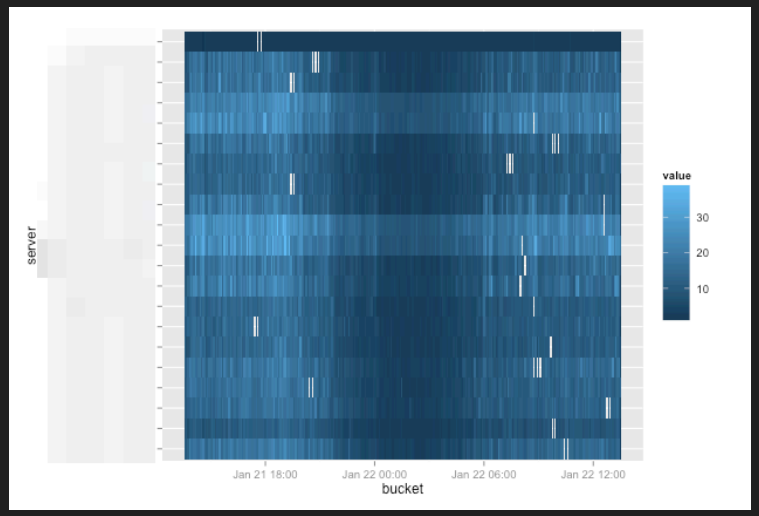

Pinterest CLSTr+
Clusters
1h
2h
6h
12h
1d
1w
Cluster Overview
Latency Metrics
Metadata E Random Read 99%(ms)
4
2
0
20:00
21:00
Metadata A Read 99%(ms)
2
1
0
20:00
21:00
Feeds A Random Read 99%(ms)
10
5
Feeds E Random Read 99%(ms)
10
5

Baselines

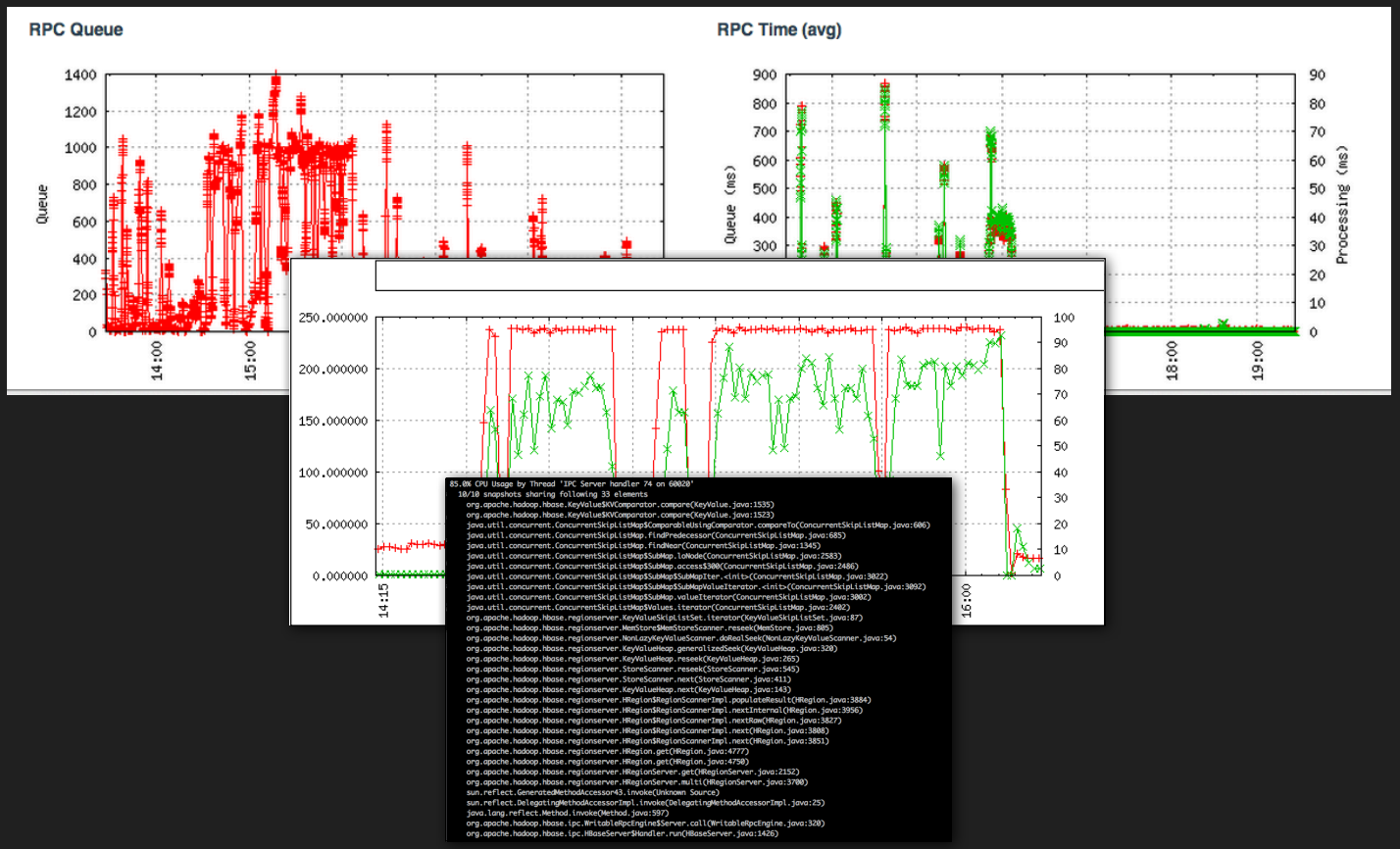

Random Read 99%(ms)
hbasemetaclustera-20131101b: 18.65
15
10
5
0
12:00
User CPU
Wed, 22 Jan 2014 13:12:48 GMT
hbasemetann001a: 7.57
10
5
0
09:00
10:00
11:00
12:00
Read Requests
2000
1000
0
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00

Visualization

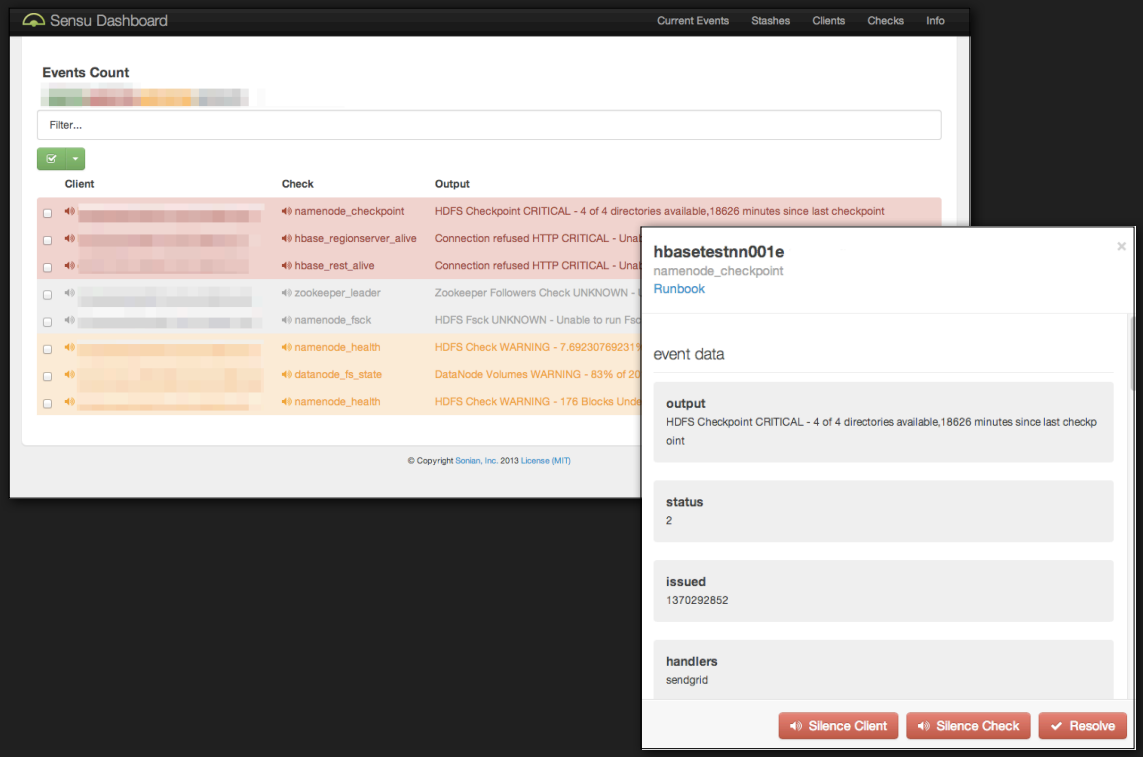




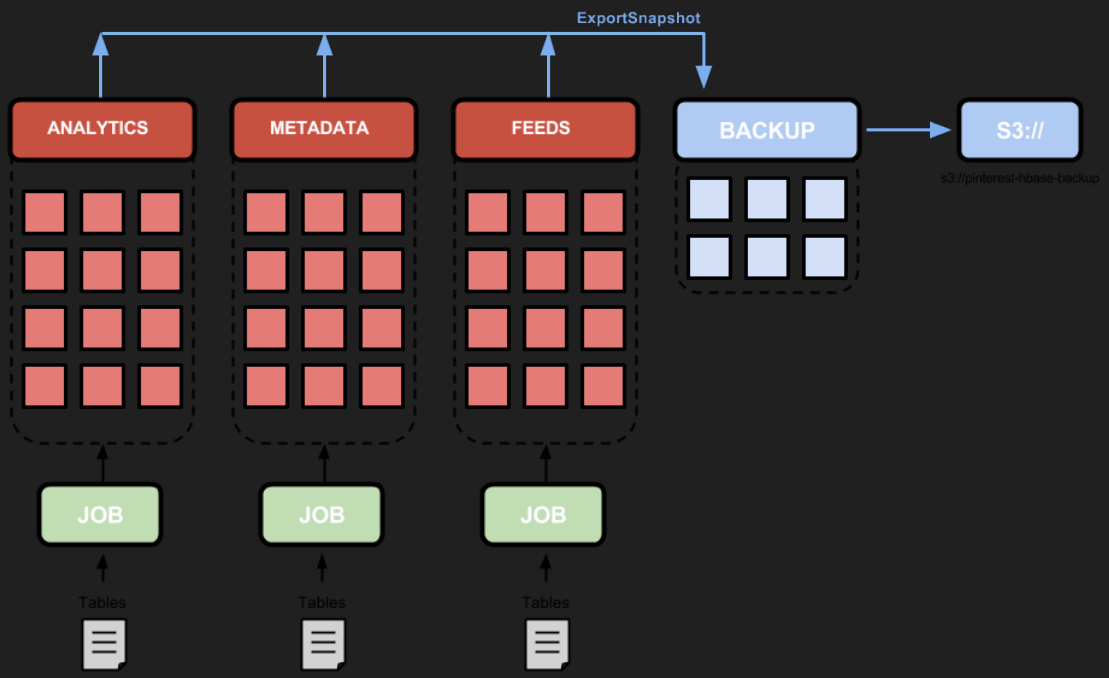
Problems


RPC Queue
Queue
1400
1200
1000
800
600
400
200
0
14:00
15:00
RPC Time (avg)
Queue (ms)
Processing (ms)
900
800
700
600
500
400
300
90
80
70
60
50
40
30
20
10
0
18:00
19:00
250.000000
200.000000
150.000000
100.000000
50.000000
0.000000
14:15
16:00
85.0% CPU Usage by Thread 'IPC Server handler 74 on 60020'
10/10 snapshots sharing following 33 elements
org.apache.hadoop.hbase.KeyValue$KVComparator.compare(KeyValue.java:1535)
org.apache.hadoop.hbase.KeyValue$KVComparator.compare(KeyValue.java:1523)
java.util.concurrent.ConcurrentSkipListMap$ComparableUsingComparator.compareTo(ConcurrentSkipListMap.java:606)
java.util.concurrent.ConcurrentSkipListMap.findPredecessor(ConcurrentSkipListMap.java:685)
java.util.concurrent.ConcurrentSkipListMap.findNear(ConcurrentSkipListMap.java:1345)
java.util.concurrent.ConcurrentSkipListMap$SubMap.loNode(ConcurrentSkipListMap.java:2583)
java.util.concurrent.ConcurrentSkipListMap$SubMap.access$300(ConcurrentSkipListMap.java:2486)
java.util.concurrent.ConcurrentSkipListMap$SubMap$SubMapIter.<init>(ConcurrentSkipListMap.java:3022)
java.util.concurrent.ConcurrentSkipListMap$SubMap$SubMapValueIterator.<init>(ConcurrentSkipListMap.java:3092)
java.util.concurrent.ConcurrentSkipListMap$SubMap.valueIterator(ConcurrentSkipListMap.java:3082)
java.util.concurrent.ConcurrentSkipListMap$Values.iterator(ConcurrentSkipListMap.java:2402)
org.apache.hadoop.hbase.regionserver.KeyValueSkipListSet.iterator(KeyValueSkipListSet.java:87)
org.apache.hadoop.hbase.regionserver.MemStore$MemStoreScanner.reseek(MemStore.java:805)
org.apache.hadoop.hbase.regionserver.NonLazyKeyValueScanner.doRealSeek(NonLazyKeyValueScanner.java:54)
org.apache.hadoop.hbase.regionserver.KeyValueHeap.generalizedSeek(KeyValueHeap.java:320)
org.apache.hadoop.hbase.regionserver.KeyValueHeap.reseek(KeyValueHeap.java:265)
org.apache.hadoop.hbase.regionserver.StoreScanner.reseek(StoreScanner.java:545)
org.apache.hadoop.hbase.regionserver.StoreScanner.next(StoreScanner.java:411)
org.apache.hadoop.hbase.regionserver.KeyValueHeap.next(KeyValueHeap.java:143)
org.apache.hadoop.hbase.regionserver.HRegion$RegionScannerImpl.populateResult(HRegion.java:3884)
org.apache.hadoop.hbase.regionserver.HRegion$RegionScannerImpl.nextInternal(HRegion.java:3956)
org.apache.hadoop.hbase.regionserver.HRegion$RegionScannerImpl.nextRaw(HRegion.java:3827)
org.apache.hadoop.hbase.regionserver.HRegion$RegionScannerImpl.next(HRegion.java:3808)
org.apache.hadoop.hbase.regionserver.HRegion$RegionScannerImpl.next(HRegion.java:3851)
org.apache.hadoop.hbase.regionserver.HRegion.get(HRegion.java:4777)
org.apache.hadoop.hbase.regionserver.HRegion.get(HRegion.java:4750)
org.apache.hadoop.hbase.regionserver.HRegionServer.get(HRegionServer.java:2152)
org.apache.hadoop.hbase.regionserver.HRegionServer.multi(HRegionServer.java:3700)
sun.reflect.GeneratedMethodAccessor43.invoke(Unknown Source)
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:25)
java.lang.reflect.Method.invoke(Method.java:597)
org.apache.hadoop.hbase.ipc.WritableRpcEngine$Server.call(WritableRpcEngine.java:320)
org.apache.hadoop.hbase.ipc.HBaseServer$Handler.run(HBaseServer.java:1426)

Alerting

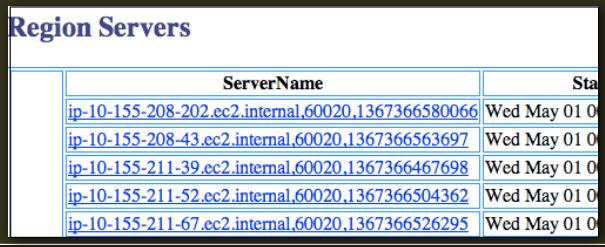

Sensu Dashboard
Current Events
Stashes
Clients
Checks
Info
Events Count
Filter...
Client
Check
Output
namenode_checkpoint
HDFS Checkpoint CRITICAL - 4 of 4 directories available,18626 minutes since last checkpoint
hbase_regionserver_alive
hbase_rest_alive
zookeeper_leader
namenode_fsck
namenode_health
datanode_fs_state
© Copyright Sonian, Inc. 2013 License (MIT)
hbasetestnn001e
namenode_checkpoint
Runbook
event data
output
HDFS Checkpoint CRITICAL - 4 of 4 directories available,18626 minutes since last checkpoint
status
2
issued
1370292852
handlers
sendgrid
Silence Client
Silence Check
Resolve

Baselines


ExportSnapshot
ANALYTICS
METADATA
FEEDS
BACKUP
S3://
s3://pinterest-hbase-backup
JOB
JOB
JOB
Tables
Tables
Tables

Snapshots & DNS
HBASE-8473

17:10 <jeremy_carroll> jmhsieh: I think I found the root cuase. All my region servers reach the barrier, but it does not continue.
17:11 <jeremy_carroll> jmhsieh: All RS have this in their logs:
DEBUG org.apache.hadoop.hbase.procedure.Subprocedure: Subprocedure 'backup1' coordinator notified of 'acquire', waiting on 'reached' or 'abort' from coordinator.
17:11 <jeremy_carroll> jmhsieh: Then the coordinator (Master) never sends anything. They just sit until the timeout.
17:12 <jeremy_carroll> jmhsieh: So basically 'reached' is never obtained. Then abort it set, and it fails.
...
17:24 <jeremy_carroll> jmhsieh: Found the bug. The hostnames dont match the master due to DNS resolution
17:25 <jeremy_carroll> jmhsieh: The barrier aquired is putting in the local hostname from the regionservers. In EC2 (Where reverse DNS does not work well), the master hands the internal name to the client.
17:26 <jeremy_carroll> jmhsieh: So it's waiting for something like 'ip-10-155-208-202.ec2.internal, 60020,1367366580066'
zNode to show up, but instead 'hbasemetaclustera-d1b0a484,60020,1367366580066,' is being inserted. Barrier is not reached
17:27 <jeremy_carroll> jmhsieh: Reason being in our environment the master does not have a reverse DNS entry. So we get stuff like this on RegionServer startup in our logs.
17:27 <jeremy_carroll> jmhsieh: 2013-05-01 00:03:00,614 INFO org.apache.hadoop.hbase.regionserver.HRegionServer: Master passed us hostname to use. Was=hbasemetaclustera-d1b0a484, Now=ip-10-155-208-202.ec2.internal
17:54 <jeremy_carroll> jmhsieh: That was it. Verified. Now that Reverse DNS is working, snapshots are working. Now how to figure out how to get Reverse DNS working on Route53. I wished there was something like 'slave.host.name' inside of Hadoop for this. Looking at source code.



Region Servers

	ServerName	Sta
	ip-10-155-208-202.ec2.internal,60020,1367366580066	Wed May 01 0
	ip-10-155-208-43.ec2.internal,60020,1367366563697	Wed May 01 0
	ip-10-155-211-39.ec2.internal,60020,1367366467698	Wed May 01 0
	ip-10-155-211-52.ec2.internal,60020,1367366504362	Wed May 01 0
	ip-10-155-211-67.ec2.internal,60020,1367366526295	Wed May 01 0

Thanks!