PILOT HADOOP TOWARDS 2500 NODES AND CLUSTER REDUNDANCY

Stuart Pook (s.pook@criteo.com @StuartPook)
Online advertising
Target the right user
At the right time
With the right message
SIMPLIFIED BUSINESS MODEL

- We buy
  - Ads spaces
- We sell
  - Clicks — that convert — a lot
- We take the risk
HADOOP AT CRITEO BACK IN 2014 (AM5)

- 1200 nodes
- 39 PB raw capacity
- > 100 000 jobs/day
- 10 000 CPUs
- HPE ProLiant Gen8
- 105 TB RAM
- 40 TB imported/day
- Cloudera CDH4
TODAY PRIMARY DATA INPUT

- Kafka
- 500 billion events per day
- up to 4.3 million events/second
- JSON → protobuf
- 72 hour buffers
Paid

Compute and Data Are Essential

- Extract, Transform & Load logs
- Bidding models
- Billing
- Business analysis
HADOOP PROVIDES LOCAL REDUNDANCY

- Failing datanodes (1 or 2)
- Failing racks (1)
- Failing namenodes (1)
- Failing resourcemanager (1)
NO PROTECTION AGAINST

- Data centre disaster
- Multiple datanode failures in a short time
- Multiple rack failures in a day
- Operator error
DATA BACKUP IN THE CLOUD

- Backup long
  - Import 100 TB/day
  - Create 80 TB/day
  - Backup at 50 Gb/s?
- Restore 2 PB too long at 50 Gb/s
- What about compute?
COMPUTE IN THE CLOUD

- 20 000 CPUs requires reservation
- Reservation expensive
- No need for elasticity (batch processing)
- Criteo has data centres
- Criteo likes bare metal
- Cloud > 8 times more expensive
- In-house get us exactly the network & hardware we need
BUILD NEW DATA CENTRE (PA4) IN 9 MONTHS

- Space for 5000 machines
- Non-blocking network
- 10 Gb/s endpoints
- Level 3 routing
- Clos topology
- Power 1 megawatt + option for 1 MW
- It's impossible
NEW HARDWARE

- Had one supplier
- Need competition to keep prices down
- 3 replies to our call for tenders
- 3 similar 2U machines
  - 16 (or 12) 6 TB SATA disks
  - 2 Xeon E5-2650L v3, 24 cores, 48 threads
  - 256 GB RAM
  - Mellanox 10 Gb/s network card
  - 2 different RAID cards
TEST THE HARDWARE

- Three 10 node clusters
- Disk bandwidth
- Disk errors?
- Network bandwidth
- Teragen
  - Zero replication (disks)
  - High replication (network)
- Terasort
HARDWARE IS SIMILAR

- Eliminate the constructor with
  - 4 DOA disks
  - Other failed disks
  - 20% higher power consumption

- Choose the cheapest and most dense
  - Huawei
  - LSI-3008 RAID card
MIX THE HARDWARE

- Operations are more difficult
- Multiple configurations needed
- Some clusters have both hardware types
- We have more choice at each order
- Avoid vendor lock-in
HAVE THE DC, BUILD HADOOP

- Configure using Chef
- Infrastructure is code in git
- Automate to scale (because we don't)
- Test Hadoop with 10 hour petasorts
- Tune Hadoop for this scale
- Namenode machine crashes
  - Upgrade kernel
  - Rolling restart on all machines
  - Rack by rack, not node by node
A MISTAKE

- Made master and datanodes the same
  - Just one type of machine
  - Many available nodes if failure
  - Moving services with kerberos hard
  - Move master nodes to bigger machines
  - Offload DNS & KDC
FASTER MASTER NODES

• Namenode does many sequential operations
  ▪ Long locks
  ▪ Failovers too slow
  ▪ Heartbeats lost
  ▪ → fast CPU

• Two big namenodes
  ▪ 512 GB RAM
  ▪ 2 × Intel Xeon CPU E5-2643 v4 @ 3.40GHz
  ▪ 3 × RAID 1 of 2 SSD
PA4 CLUSTER ONLINE

- More capacity as have 2 clusters
- Users love it
- Users find new uses
- Soon using more capacity than the new cluster
- Impossible to stop the old cluster
SITUATION NOW WORSE

- Two critical clusters
- Two Hadoop versions to support CDH 4 & 5
- Not one but two SPOFs
GROW THE NEW DATACENTRE

- Add hundreds of new nodes
- Soon the new cluster will be big enough
  - 1,370 datanodes
  - +650 in Q3 2017
  - +900 in Q4 2017
  - ~3,000 datanodes end 2017
- Too many blocks for the namenode?
ONE SPOF IS ENOUGH

- Move all jobs and data to the new cluster (PA4)
- Stop the old cluster (AM5)
- Only one SPOF but still no redundancy
- Have file backups on different technology
2018 BUILD ANOTHER CLUSTER

- Human users (development, machine-learning, BI)
- QA & non-regression for service jobs
- All data for service jobs
- ➔ PA4 backup for service jobs
- But RAM is expensive
ANOTHER ROUND OF TENDORS

- Need more CPU
- Denser machines
- 4U, 8 nodes, 16 CPUs, $4 \times 8 \times 2.5''$ disks (8/U)
- 2U, 4 nodes, 8 CPUs, $6 \times 4 \times 2.5''$ disks (12/U)
- Infrastructure validation
- Hadoop tests and benchmarks
THE RAID CONTROLLER STORY
FIRST RAID CONTROLLER

- Historically HPE-Smart Array P420 (PMC-Sierra)
- Only RAID 0 users in Criteo
- OS status = RAID card status
  - Skip volumes flagged as bad by the RAID card
  - Very rare cases of fsck failures
  - Filesystems mounted with errors=remount-ro
  - Very rare cases of unflagged read-only filesystems
  - Access to the volumes blocked
- Assumed to be standard behaviour
- Operations need RAID card error flag
  - Out-of-band status → Jira ticket
  - Identification LED → disk swap
SECOND RAID CONTROLLER LSI 3008

- Disks vanished
- No diagnostic LED
- Used ID LED on other disks
- Later tested OK
- Worked after power cycle
- Change 700 cards on a running cluster
  - All blocks lost on each card change
  - No downtime allowed
  - Rack by rack
  - Many HDFS stability problems
THIRD RAID CONTROLLER LSI-3108

- LSI RAID card
- Now OS flags bad disks before card
  - Failing fsck
  - Read-only filesystems
  - Volume seen as OK by the RAID card
  - O/S can access the volume and get timeouts
  - No error for out of band monitoring
  - No error for in-band monitoring
- We can only handle OK or Failed volumes
2 SOLUTIONS WITH VENDOR LOCK-IN

- Buy all machines from HPE
- Get the supplier to “solve the problem for us
  - Agent running in the controller
  - They develop in China
  - We debug in France (in prod)
2 COMPLICATED SOLUTIONS

- Create RAID0 team to
  - Handle all error conditions
  - Stop access to the volume
  - Reformat (once) volume with read errors
  - Open tickets
  - Set identification LED
- Work with LSI to tweak their controller
STAYING ALIVE

- Automate operations
  - Disk changes
  - Ramp ups
- Infrastructure as code
  - Tests: kitchen, ChefSpec & preprod
  - Merge requests with reviews
- Choreograph operations
  - Restarts (machines or services)
NEED LAKE CLUSTER

- Infrastructure validation
- Test configuration recipes
- Test Hadoop patches
- And lab for hardware tests
• Increase bandwidth and time limit for checkpoint
• 332 GB heap for namenode
  ▪ 180 GB reserved for native code & OS
• Tune GC for namenode
  ▪ Serial → not efficient on multi-thread
  ▪ Parallel → long pauses + high throughput
  ▪ Concurrent Mark Sweep → short pauses + lower throughput
  ▪ G1 → in prod
    o G1NewSizePercent=0
    o G1RSetRegionEntries=4096
    o +ParallelRefProcEnabled & +PerfDisableSharedMem
• Azul not required for 1300 datanodes but 3000?
STAYING ALIVE — FIX BUGS

The cluster crashes, find the bug, if fixed, backport it, else fix

- Fix
  - HDFS-10220 expired leases make namenode unresponsive and failover

- Backport
  - YARN-4041 Slow delegation token renewal prolongs RM recovery
  - HDFS-9305 Delayed heartbeat processing causes storm of heartbeats
  - YARN-4546 ResourceManager crash due to scheduling opportunity overflow
  - HDFS-9906 Remove spammy log spew when a datanode is restarted
STAYING ALIVE — MONITORING

- **HDFS**
  - Namenode: missing blocks, GC, checkpoints, safemode, QPS, live datanode
  - Datanodes: disks, read/write throughput, space

- **YARN**
  - Queue length, memory & CPU usage, job duration (scheduling + run time)
  - ResourceManager: QPS
  - Bad nodes
  - Probes to emulate client behavior with witness jobs

- **Zookeeper**
  - Availability, probes
CLUSTER COLLAPSE

- Lots of blocks → 132 GB namenode (NN) heap full
- User creates 20 million files & 20 PB data on a Friday afternoon
- NN gets stuck doing GC → no throughput
- Increase standby heap size to 85% RAM via restart
- Too many requests during restart (iptables)
- Failover crashed
- Fsimage on active corrupt as too big for transfer
- Copy missing NN edits from journal node
- Restart 1200 datanodes in batches
- 36 hours to recover the cluster
RESOURCES MANAGER SLOWS

- Event Event Type: KILL_CONTAINER sent to absent container
- These messages happen occasionally
- Almost no jobs running (8% capacity used)
- Need to kill the applications
- During NodeManager’s resync with the ResourceManager?
NEED SERVICE-LEVEL AGREEMENT (SLA)

- Define Time for operations
- Job duration
- Request handling
- Measure Monitoring
- Respect Some services are “best effort”
OPERATOR ERROR

- Same operators on both clusters
- One chef server for both clusters
- Single mistake → both clusters

To make error is human. To propagate error to all server in automatic way is #devops.
WE HAVE

- 2 prod clusters
- 2 pre-prod clusters
- 1 infrastructure cluster
- 2 running CDH4
- 3 running CDH5
- 2682 datanodes
- 49,248 cores
- 135 PB disk space
- 842 TB RAM
- > 300,000 jobs/day
- 100 TB imported daily
- 6 PB created or read per day
UPCOMING CHALLENGES

- Optimize and fix Hadoop
- Add hundreds more datanodes
- Create a new bare-metal data-centre
- Make 2 big clusters work together
- Improve scheduling
- We are hiring

- Come and join us in Paris, Palo Alto or Ann Arbor
  - s.pook@criteo.com @StuartPook

Questions?