University of BRISTOL

Hadoop @ Bristol

A brief overview

University of BRISTOL

Disclaimer

- For in-depth reviews of HDFS for grid sites look at OSG
  - https://digitalcommons.unl.edu/cseconfwork/157/
  - https://arxiv.org/pdf/1508.01443.pdf
- Middleware & documentation
  - https://github.com/opensciencegrid/gridftp-hdfs
  - https://opensciencegrid.github.io/docs/data/hadoop-overview/
- Our experience with DPM + HDFS: December 2015 review
- The following is a view from the only (?) non-OSG site with HDFS storage

University of BRISTOL

What is Hadoop?

- http://hadoop.apache.org/

"The Apache Hadoop software library is a framework that allows for the **distributed processing of large data sets** across clusters of computers using simple programming models. It is designed to **scale up from single servers to thousands of machines, each offering local computation and storage.** Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures."

- In short
  - Worker nodes are both Data (HDFS) and Compute (YARN)
  - Self-healing system, data-locality preferred by work scheduler (YARN)
    - (HDFS is rack, node, and disk-aware)

A brief history

2010

Simon Metson experimenting with first version of the system

- Adhoc configuration, gathering data

2012

CSE, geology, and physics come together to create first version of the DICE

- Hadoop + Bigcouch (distributed CouchDB) are configured
- Mostly used for Smart Metre Data

2013

Reconfiguration for WLCG (compute) with HTCondor

2015

Reconfiguration for WLCG (storage) with DMLite (GPFS maintenance running out)

Data Intensive Computing Environment (DICE)

- Using Cloudera Hadoop Distribution
  - Once they waved the node limit for the ‘free license’ looked like the best way forward
- No longer support BigCouch – no more users
- Added HTCondor as batch system, disabled YARN
- Mounting HDFS (“POSIX”) with hadoop-hdfs-fuse on all WNs as read-only

Today

- DICE provides all of the site's storage and most of the compute
  - 1.1 PB of storage (net 0.55 PB)
  - Approx. 1100 cores (out of ~1300 total)
- Latest 2 generation of nodes with 10 Gbit/s NICs
  - Recommendation from Nebraska T2 (big HDFS site)

Summary

Security is off.

Safemode is off.

18,063,524 files and directories, 14,692,342 blocks = 32,755,866 total filesystem object(s).

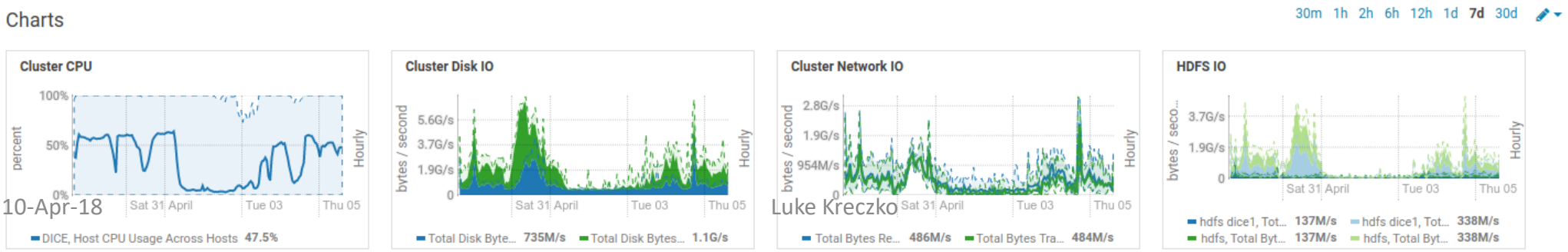
Heap Memory used 19.02 GB of 39.85 GB Heap Memory. Max Heap Memory is 39.85 GB.

Non Heap Memory used 66.1 MB of 97.5 MB Commited Non Heap Memory. Max Non Heap Memory is 130 MB.

Configured Capacity:	1.06 PB
DFS Used:	969.33 TB (89.13%)
Non DFS Used:	21.62 TB
DFS Remaining:	96.56 TB (8.88%)
Block Pool Used:	969.33 TB (89.13%)
DataNodes usages% (Min/Median/Max/stdDev):	67.21% / 89.64% / 94.93% / 4.10%
Live Nodes	58 (Decommissioned: 0)
Dead Nodes	4 (Decommissioned: 0)
Decommissioning Nodes	0
Total Datanode Volume Failures	1 (2.68 TB)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Mon Jan 15 11:15:33 +0000 2018

The good

- Hadoop is an industry standard (we need more of those in WLCG)
- WNs are both data & storage: only need one type of machine
- Setup, upgrade & monitoring provided with Cloudera's Manager
- Files are split into chunks and copied N times (default 3, here 2) across racks, nodes, disks
  - Crashes & corruption are rare (even if entire rack goes!)
  - Easy to replace disks & WNs in this setup



The bad

- Either GUI or command line setup
  - we chose GUI for simplicity, but cmd line is faster for maintenance
- Needs tuning for ROOT files (ROOT is not a good format for HDFS)
  - Increased default chunk size (used to lead to bad behaviour in the past)
- Replacing YARN with HTCondor breaks data locality
  - More network usage than in default Hadoop scenarios
- Data nodes are also compute notes
  - Large IO can use 1-2 cores per node

University of BRISTOL

The ugly

- If you make a mistake during the upgrade process you risk losing all data
  - Read and follow documentation to the dot
- “Not invented here” syndrome guided us towards DPM – no other site has this setup!
  - Also: there is no such thing like a lightweight storage element
- Users expect POSIX compliance
  - And they will break your FUSE mount (read-only access helps with worst cases)

The wishlist

- A stateless storage element
  - Does only path mapping & auth, rest using HDFS API
- HTCondor plugin for data locality
- More sites using similar setup
  - Easier to debug problems and thus improve middleware

University of BRISTOL

Summary

- Hadoop@Bristol developed from a collaboration project across disciplines
  - Now our main compute & storage resource for grid and local users
- Scaling & Maintenance are relatively easy
  - But providing access to the grid is a whole other story
- Hadoop 3 is finally out
  - Brings erasure coding among other things