Hadoop DFS (HDFS) – Post OSG 3.5

D. Bradley, S. Dasu, A. Mohapatra, C. Seys, C. Vuosalo
(HEP Computing Group)

Outline

- HDFS in OSG 3.5
- No HDFS in OSG 3.6 (and beyond)
- Explore HDFS with Apache Bigtop
- Looking into the future

HDFS in OSG 3.5

Data storage on Disks (DNs, NN)

+

Interfaces for Data Access (FUSE, gridFTP, and xrootd)

Data Storage on Disks (DNs, NN)

- Hadoop 2.6.0 source rpms were packaged by Cloudera (cdh 5.12.1)
  - hadoop-hdfs-namenode (NN)
  - hadoop-hdfs-datanode (DN)
  - hadoop-hdfs-fuse (FUSE)
  - hadoop-hdfs and more
- NameNode was a single point of failure prior to Hadoop 2.0.0
- HDFS High Availability (HA) support since Hadoop 2.0.0
  - Provides stability and interruption free operation

Interfaces for Data Access

- **FUSE**
  - **hadoop-hdfs-fuse-2.6.0+cdh5.12.1+2540-1.cdh5.12.1.p0.3.8.osg35.el7.x86_64.rpm**
- **GridFTP**
  - **gridftp-hdfs-1.1.1-1.2.osg35.el7.x86_64**
  - **osg-gridftp-hdfs-3.5-4.osg35.el7.x86_64**
- **Xrootd**
  - **xrootd-hdfs-2.1.8-1.osg35.el7.x86_64**

No HDFS in OSG 3.6

What it means for sites ?

OSG 3.6 (and Beyond)

- GridFTP, GSI, and Hadoop are no longer available
- XrootD will be available
- HDFS DSI (gridftp/xrootd related) packages will not be supported
  - gridftp-hdfs-*
  - xrootd-hdfs-*

Options for Sites

- Continue with Hadoop-hdfs
  - Self managed installations/upgrades (includes hdfs-fuse DSI)
- Continue with xrootd-hdfs
  - Self managed (i.e. prepare rpm, test, validate etc.)
- Continue with gridftp-hdfs
  - Fine until the end of OSG 3.5 support (self managed afterwards)
  - CMS is actively working to replace gridftp with WebDAV this year. That would help USCMS T2 sites get rid of gridftp usage before OSG 3.5 end-of-life

Hadoop cluster at Wisconsin T2

- Operational since 2010 (non-HA mode until 2016, HA since 2017)
- 10 PB HDFS distributed across ~350 nodes (shared with htcondor/compute nodes)
- Xrootd and GridFTP services on ~300 nodes for hdfs data access
- Hadoop-2.6.0 (OSG 3.5) until Jan 2021

- Were planning to upgrade to Hadoop-3.x (with erasure coding) for a while

Hadoop Upgrade Plan (2.6 → 3.x)

- Cloudera cdh5.x is still stuck with Hadoop-2.6 base, and cdh6.x enterprise (Hadoop-3.x) is not free.
- Rolling upgrade from 2.6 to 3.x is not possible due to incompatible changes
- Decided to upgrade to a stable version > 2.6 and learned about Apache Bigtop
- Apache Bigtop packages Hadoop rpms. Once a year release cycle so far. Latest release 1.5.0 (Dec 2020) contains Hadoop-2.10.1

HDFS with Apache Bigtop

An opportunity to explore !

Pre-Upgrade Homework

- Verify that Hadoop-2.6.0 from Bigtop repo is compatible with OSG 3.5 and adapt as needed
- Created a small test cluster with a few DNs and experimented with Hadoop-2.6.0 (Bigtop-1.0.0), Hadoop-2.8.5 (Bigtop-1.4.0), and Hadoop-2.10.0 (Bigtop-1.5.0)
- Tested the HA setup and the functionality of hdfs-fuse, xrootd-hdfs and gridftp-hdfs interfaces successfully
- Configured all necessary steps in puppet to prepare for the production cluster upgrade

Upgrade and Finalize

- Performed Rolling Upgrade
- Experienced a few hiccups on the way (details at https://github.com/cwseys/OSG-Bigtop). A few noted ones :
  - Datanodes (running 2.6) were not reporting their blocks to NN (running 2.10.1). A known bug.
  - Datanodes needed more JVM heap during the upgrade
  - xrootd, gridftp, and other daemons had to be restarted to use the new HDFS libraries
- The cluster was fully operational during the upgrade phase

Post Upgrade

- Upgraded to Hadoop-2.10.1 (OSG 3.5 + Bigtop 1.5.0) in late Jan 2021
- Cluster is running stably for over a month
- Datanodes are running fine with the JVM heap size restored to Hadoop-2.6 era values

Future

- Next Bigtop release expected in 2021 will have Hadoop-3.x with erasure coding
- Current plan is to continue with HDFS as long as it's feasible and sustainable with minimal effort
- Hope to collaborate with other CMS T2s in same boat

Shout out

A big "Thank You" to OSG team for supporting Hadoop HDFS and related packages for 10+ years

Questions/Comments ?