Apache Hadoop 3.x State of The Union and Upgrade Guidance

Wei-Chiu Chuang  
@Cloudera,  
Apache Hadoop PMC

Wangda Tan  
Sr. Manager, Compute Platform  
@Cloudera,  
Apache Hadoop PMC
Agenda

- Hadoop Community Updates & Overview
- Updates from YARN, Submarine, HDFS, Ozone
- Upcoming releases
- Upgrade guidance
Community Updates
is hadoop dead

Is Hadoop Dead? How Kubernetes and Cloud-Native Could ...
https://www.bmc.com › blogs › hadoop-cloud-native-kubernetes

Is Hadoop Officially Dead? - Datanami
https://www.datanami.com

The Death of Hadoop? | Transforming Data with Intelligence
https://tdwi.org › articles › 2019/02/26 › ta-all-death-of-hadoop

Why is Hadoop dying? | Packt Hub
https://hub.packtpub.com › why-is-hadoop-dying

Is Hadoop dead and is it time to move to Spark? - Quora
Jun 24, 2015 - ATA I think this is like asking: is Linux dead and should we move to Docker? or
Resolved Issues by Top 10 ASF Projects

<table>
<thead>
<tr>
<th>project</th>
<th>issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 HADOOP</td>
<td>1930</td>
</tr>
<tr>
<td>2 ARROW</td>
<td>1693</td>
</tr>
<tr>
<td>3 FLINK</td>
<td>1600</td>
</tr>
<tr>
<td>4 SPARK</td>
<td>1364</td>
</tr>
<tr>
<td>5 BEAM</td>
<td>1074</td>
</tr>
<tr>
<td>6 AIRFLOW</td>
<td>1073</td>
</tr>
<tr>
<td>7 HBASE</td>
<td>982</td>
</tr>
<tr>
<td>8 IGNITE</td>
<td>817</td>
</tr>
<tr>
<td>9 CAMEL</td>
<td>779</td>
</tr>
<tr>
<td>10 GEODE</td>
<td>708</td>
</tr>
</tbody>
</table>
Resolved Issues within Hadoop by Subproject (Monthly)
Resolved issue in Hadoop (Monthly)
Number of Unique #Contributors of Hadoop (Monthly)

(All pictures credits to Marton Elek)
Hadoop 3.x Overview
Big Data/Long Running Services With Hadoop 3

BATCH WORKLOADS

SERVICES

DEEP LEARNING APPS

Hadoop Ozone

HIVE on LLAP

TensorFlow

mxnet

docker

PUBLIC CLOUD STORAGE

COMPUTE (on-prem/on-cloud)

STORAGE
Themes of Hadoop 3.x

- Scalability
- Containerization
- Cost-efficiency
- Cloud-native
- Machine Learning
YARN
Containerization

- Production-ready Docker container support on YARN.
  - Containerized Spark
    - Package/Dependency Isolation
  - Interactive Docker Shell support (YARN-8762)
- OCI/squashfs (Like runc) container runtime.
YARN in a cloud-native environment  YARN-9548

- Autoscaling
  - Scaling recommendations
- Smarter scheduling
  - **Bin-packing**  Pack containers as opposed to spreading them around to downscale nodes better
  - Account for speculative nodes like spot instances
- Downscaling nodes
  - Improved Decommissioning
  - Consider shuffle/auxiliary services data

**Ongoing Effort**
Global Scheduling Framework  YARN-5139

Scheduler Capabilities enhancements

- Look at several nodes at one time.
- Fine grained locks.
- Multiple allocation threads.
- 5-10x allocation throughput gains.

Available since 3.0.0
Other Enhancements

- Node Attributes: Tagging node with attribute and schedule containers based on that. (3.2.0)
- Placement Constraint: Affinity, Anti-Affinity, etc. (3.1.0)
- Dynamic Auto Queue Creation (Capacity Scheduler) (3.1.0)
- Scheduling Activity Troubleshooter. (3.3.0)

<table>
<thead>
<tr>
<th>Priority</th>
<th>AllocationRequestId</th>
<th>ResourceName</th>
<th>Capability</th>
<th>NumContainers</th>
<th>RelaxLocality</th>
<th>ExecutionType</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>-1</td>
<td>*</td>
<td>&lt;memory:40960, vCores:30&gt;</td>
<td>42</td>
<td>true</td>
<td>GUARANTEED</td>
</tr>
</tbody>
</table>

Showing 1 to 1 of 1 entries

(Update via Application Activities REST API)

<table>
<thead>
<tr>
<th>Priority</th>
<th>AllocationRequestId</th>
<th>AllocationState</th>
<th>Diagnostics</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>-1</td>
<td>SKIPPED</td>
<td>1 node: insufficient resources=[vcores], 3 nodes: insufficient resources=[memory=mb, vcores]</td>
</tr>
</tbody>
</table>
Submarine
Machine Learning – Hadoop Submarine

- Started since Aug 2018.
- Benefit from Hadoop’s feature like GPU/Docker on YARN support.
- Enables Infra engineers / data scientists to run deep learning apps
  - Tensorflow, Pytorch, MXNet.. on YARN/K8s
  - Supports Hadoop 2.7+.
- LinkedIn TonY joined Submarine family
Machine Learning – Hadoop Submarine

- Lots of new stuff in upcoming releases (0.3.0).
  - Mini-submarine for easy trying Submarine from single node.
  - Brand-new Submarine web interface for end-to-end user Experiences.
  - Tensorflow/PyTorch on K8s.
- 15+ Contributors and community is fast growing.
Machine Learning – Hadoop Submarine Prod Use cases

**NetEase:**
- One of the largest online game/news/music provider in China.
- 245 GPU Cluster runs Submarine.
- One of the model built is music recommendation model which invoked 1B+/days.

**LinkedIn:**
- 250+ GPU machines
- 500+ TensorFlow trainings/day.
- Serves applications in recommendation systems and NLP.
- Collaboration on Submarine/TonY runtime and SDK development.

**Ke.com:**
- 50+ GPU machines (includes 19 multi-v100 GPU machines), based on Hadoop trunk (3.3.0).
- Serves applications like image/voice recognition, etc.

And many users are evaluating Submarine…
Machine Learning – Submarine new UI demo
Storage
HDFS Updates - Consistent Read from Standby

- Offload reads to non-active NameNodes to improve overall file system performance.
- Consistency: if a client can report the last transaction ID seen by it, then a standby can allow a read if it has caught up to that transaction ID seen by the client.
- Used in production at Uber and LinkedIn.
HDFS Updates - Router Based Federation

- Router based Federation Supports Security.
- Lots of work on scalability and the ability to handle slower sub-clusters.
- We are seeing usage across the industry
And many more HDFS features

- Selective Wire Encryption
- Cost based Fair call queue
- Dynamometer
- Storage Policy Satisfier
- Support Non-volatile storage class memory in HDFS cache directives

Ongoing development

- RPC support for TLS
- KMSv2
- OpenTracing integration
- JDK11 support
Cloud Connector Updates - S3A/S3Guard

S3A File system supports Delegation Tokens.

- Full user + secret + encryption keys: simplest, but secrets do not leave your system.
- Generated session tokens + encryption keys: keeps the long lived secrets locally; life of non-renewable tokens limited

S3Guard is no longer considered experimental

- Maintain consistency through corner cases involving partial failure of rename/delete operations.
- Out of band support - detecting and adapting to other applications overwriting files.
- Tracking of etag and version Ids for stricter consistency when you want to defend against OOB changes.
- “authoritative mode” improves performance dramatically.
ABFS: “Azure Datalake Gen 2” Connector

- A high performance cloud store & filesystem for Azure
- Added in Hadoop 3.2.0;
- Stabilization in trunk with all fixes backported to 3.2.1
- Has a similar extension point for Delegation Token plugins as S3A. (though implementing DTs is “left as an exercise”. Contributions welcome)

Credit to Thomas Marquardt and Da Zhou @Microsoft for their work—and welcome to the Hadoop Committer Team!
Hadoop Common
On going Effort

- RPC support for TLS
- KMSv2
- OpenTracing integration
- JDK11 support
Ozone
Ozone

- Object Store made for Big Data workloads.
  - A long term successor of HDFS.
  - In-place upgrade from HDFS (roadmap)
- Contribution from Hortonworks/Cloudera/Tencent ...
- Tremendous progress over past year
Ozone Upcoming releases

- Three Alpha Releases so far.
  - 0.2: basic object store.
  - 0.3: s3 protocol.
  - 0.4: Security and Ranger support.
- 0.4.1 release (Native ACLs) coming out soon (December-ish).
- 0.5.0 will be the beta release.
  - Reliability and performance improvement.
  - HA
Releases
# Release Plan - Core Hadoop

<table>
<thead>
<tr>
<th>Year</th>
<th>Version</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>2.6.5</td>
<td>Stabilization, Maintenance, Bug fix</td>
</tr>
<tr>
<td></td>
<td>2.7.5 - 2.7.7</td>
<td>Stabilization, Maintenance, Bug fix</td>
</tr>
<tr>
<td></td>
<td>2.8.3 - 2.8.5</td>
<td>Stabilization, Maintenance, Bug fix</td>
</tr>
<tr>
<td>2019</td>
<td>3.1.2</td>
<td>Stabilization, Maintenance, Bug fix</td>
</tr>
<tr>
<td></td>
<td>3.2.0</td>
<td>Node Attributes, Submarine, Storage Policy Satisfier, ABFS connector</td>
</tr>
<tr>
<td></td>
<td>2.10 (Planned)</td>
<td>YARN resource types/GPU support (YARN-8200)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selective wire encryption (HDFS-13541)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HDFS Rolling upgrades from 2.x to 3.x(HDFS-14509)</td>
</tr>
<tr>
<td>2019</td>
<td>3.1.3 (RC0, Target Sep 2019)</td>
<td>Stabilization, Maintenance, Bug fix</td>
</tr>
<tr>
<td></td>
<td>3.2.1 (Sep 2019, released).</td>
<td>Stabilization, Maintenance, Bug fix (GA of 3.2)</td>
</tr>
<tr>
<td></td>
<td>3.3.0 (Planned)</td>
<td>Stabilization, Maintenance, Bug fix (GA of 3.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stabilization, Maintenance, Bug fix (GA of 3.2)</td>
</tr>
</tbody>
</table>
# Release Plan - Submarine

- Voted to become a separate Apache project
- No longer part of Core Hadoop releases

<table>
<thead>
<tr>
<th>Year</th>
<th>Version</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>0.1.0</td>
<td>YARN</td>
</tr>
</tbody>
</table>
|      |         | - Distributed Tensorflow
|      |         | - MXNet   |
| 2019 | 0.2.0   | Support for other runtimes |
|      |         | - Pytorch
|      |         | - Linkedin’s TonY
|      |         | - Zeppelin Notebook support |
|      | 0.3.0 (Planned) | Support K8s runtimes |
|      |         | - Mini-submarine
|      |         | - Submarine-workbench
|      |         | - Submarine SDK |
End of Life Policy

- EOL of Releases with no maintenance release in long term (1.5+ yrs)
- Security-only releases on EOL versions if requested.
- EOLed Versions
  - Hadoop 2.7.x (and lower)
  - Hadoop 3.0.x
Upgrades
(Hadoop 2 -> Hadoop 3)
## Express/Rolling Upgrades

### Express Upgrades
- Stop the world Upgrades
- Cluster downtime
- Less stringent prerequisites
- Process
  - Upgrade masters and workers in one shot

### Rolling Upgrades
- Preserve cluster operation
- Minimizes Service impact and downtime
- Can take longer to complete
- Process
  - Upgrades masters and workers in batches
Recommendation - Express or Rolling?

- **Major version upgrade**
  - Challenges and issues in supporting Rolling Upgrades

- **Technical challenges with rolling upgrade**
  - Lot of work done/WIP by Hadoop community to support upgrades without Downtime. Should be part of releases soon.
  - Backward incompatible changes blocks rolling upgrade.

- **Recommended**
  - *Express Upgrade* from Hadoop 2 to 3
Compatibility

Wire compatibility

- Preserves compatibility with Hadoop 2 clients
- Distcp/WebHDFS compatibility preserved

API compatibility

Not fully, but minimal impact.

- Dependency version bumps
- Removal of deprecated APIs and tools
- Shell script rewrite, rework of Hadoop tools/scripts.
Source & Target Versions

Upgrades Validated with

<table>
<thead>
<tr>
<th>Hadoop 2 Base version</th>
<th>Hadoop 3 Base version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache Hadoop 2.8.x</td>
<td>Apache Hadoop 3.1.x</td>
</tr>
</tbody>
</table>

Why 2.8.x release?
● Most of production deployments are close to 2.8.x

What should users of 2.6.x and 2.7.x do?
● Do more validations before upgrading, we do see some users directly upgrade from 2.7.x to 3.x.
Upgrade Process/Details

Refer to our earlier talk for further details

Migrating Hadoop cluster and workloads from Hadoop 2 to Hadoop 3
Many successful use cases for Hadoop 3.x (New And Upgrade) in Production
Summary of upgrade

- **Hadoop 3**
  Eagerly awaited release with lots of new features and optimizations!
- Lots of large clusters already on Hadoop 3 at enterprises
- **Express Upgrades** are recommended
- If you haven't upgraded yet, NOW is the best time!
Questions?
Rate today’s session