Welcome

Thanks for joining the Hadoop user forum!

Bring together the users of the service
Share our plans with you
Listen to your feedback and ideas
Understand how we can provide a better service
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00</td>
<td>Welcome coffee</td>
</tr>
<tr>
<td>09:30</td>
<td>Hadoop Service</td>
</tr>
<tr>
<td>09:30</td>
<td><strong>Status</strong></td>
</tr>
<tr>
<td></td>
<td>Overview of the service evolution and</td>
</tr>
<tr>
<td></td>
<td>presentation of main changes in the last</td>
</tr>
<tr>
<td></td>
<td>year</td>
</tr>
<tr>
<td></td>
<td>Speaker: Pedro Andrade (CERN)</td>
</tr>
<tr>
<td>09:50</td>
<td><strong>Roadmap</strong></td>
</tr>
<tr>
<td></td>
<td>Operational goals for next year and</td>
</tr>
<tr>
<td></td>
<td>long-term service strategy</td>
</tr>
<tr>
<td></td>
<td>Speaker: Emil Kleszcz (CERN)</td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="20231204-Hadoop" /></td>
</tr>
<tr>
<td>10:20</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>10:30</td>
<td>Hadoop User Communities</td>
</tr>
<tr>
<td>10:30</td>
<td><strong>ATLAS EventIndex Team</strong></td>
</tr>
<tr>
<td></td>
<td>Speaker: Grigori Rybklin (University Paris-</td>
</tr>
<tr>
<td></td>
<td>Saclay (FR))</td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="ATLAS EventIndex" /></td>
</tr>
<tr>
<td>10:45</td>
<td><strong>MONIT Team</strong></td>
</tr>
<tr>
<td></td>
<td>Speaker: Nikolay Tsvetkov (CERN)</td>
</tr>
<tr>
<td>11:00</td>
<td><strong>NXCAL Team</strong></td>
</tr>
<tr>
<td></td>
<td>Speaker: Jakub Wozniak (CERN)</td>
</tr>
<tr>
<td>11:15</td>
<td><strong>SWAN Team</strong></td>
</tr>
<tr>
<td></td>
<td>Speaker: Enric Tejedor Saavedra (CERN)</td>
</tr>
<tr>
<td>11:30</td>
<td><strong>Discussion</strong></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Discussion" /></td>
</tr>
</tbody>
</table>

04 Dec 2023
Pedro Andrade | Hadoop service overview
Hadoop Service Overview
Overview

1. Introduction: hadoop, ecosystem, history
2. People: team, users and projects
3. Infrastructure: resources, architecture, clusters, tech stack
4. User information: tips and tricks, communication
Introduction
Hadoop - Ecosystem - History
Introduction

What is Hadoop?

“An open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.”

Why Hadoop?

- Too much Data (TB per day)
- Vertical scaling doesn’t cut it
- Horizontal scaling is linear

![Diagram of interconnected network with CPU, Memory, and Disks]
Ecosystem

- **HDFS**: Hadoop Distributed File System
- **YARN**: Cluster resource manager
- **Zookeeper**: Coordination of distributed systems
- **Hive**: SQL
- **Pig**: Scripting
- **Sqoop**: Data exchange
- **Spark**: Data processing
- **HBase**: Column store
- **Phenix**: SQL

**MapReduce**

Pedro Andrade | Hadoop service overview
04 Dec 2023
History

2014
Initial prototypes
Apache Spark support
Initial SQL-based tools

2016
Production-ready service
Added HDFS backups
High-availability

2018
NXCALS dedicated cluster
Migration to Hadoop 3

2020
Production-ready NXCALS
Analytix and Lhadoop clusters merge

2022
New service architecture
NXCALS HBase isolation

2024
Why we are here today :)
People
Team - Users - Projects
Users

1721 registered users
Projects

### HDFS files/folders

<table>
<thead>
<tr>
<th>Project</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>NXCALS</td>
<td>8.7 PB</td>
</tr>
<tr>
<td>Security</td>
<td>1.3 PB</td>
</tr>
<tr>
<td>CMS</td>
<td>279.9 TB</td>
</tr>
<tr>
<td>ITMON</td>
<td>987.7 TB</td>
</tr>
<tr>
<td>MONIT</td>
<td>1.3 PB</td>
</tr>
<tr>
<td>ATLAS Rucio</td>
<td>1.6 PB</td>
</tr>
</tbody>
</table>

### HDFS disk space

<table>
<thead>
<tr>
<th>Project</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>NXCALS</td>
<td>33.2 Mil</td>
</tr>
<tr>
<td>Security</td>
<td>3.5 Mil</td>
</tr>
<tr>
<td>CMS</td>
<td>586.5 K</td>
</tr>
<tr>
<td>ITMON</td>
<td>2.3 Mil</td>
</tr>
<tr>
<td>MONIT</td>
<td>4.2 Mil</td>
</tr>
<tr>
<td>ATLAS Rucio</td>
<td>10.7 Mil</td>
</tr>
</tbody>
</table>

04 Dec 2023

Pedro Andrade | Hadoop service overview
Infrastructure

Resources - Architecture - Clusters - Tech Stack
## Resources: bare metal

<table>
<thead>
<tr>
<th>Order</th>
<th>Installation</th>
<th>End of life</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3</td>
<td>2017</td>
<td>2024</td>
<td>IT support</td>
<td>24 nodes and 6 JBODs</td>
</tr>
<tr>
<td>HDP1</td>
<td>2017</td>
<td>2022</td>
<td>retired</td>
<td>8 nodes and 16 JBODs</td>
</tr>
<tr>
<td>HDP2</td>
<td>2018</td>
<td>2023</td>
<td>IT support</td>
<td>42 nodes and 76 JBODs</td>
</tr>
<tr>
<td>HDP3</td>
<td>2019</td>
<td>2024</td>
<td>IT support</td>
<td>8 nodes and 16 JBODs</td>
</tr>
<tr>
<td>HDP4</td>
<td>2020</td>
<td>2025</td>
<td>IT support</td>
<td>16 nodes and 32 JBODs</td>
</tr>
<tr>
<td>HDP5</td>
<td>2021</td>
<td>2026</td>
<td>warranty</td>
<td>60 nodes and 32 JBODs</td>
</tr>
<tr>
<td>HDP6</td>
<td>2022</td>
<td>2027</td>
<td>warranty</td>
<td>20 nodes and 15 JBODs</td>
</tr>
<tr>
<td>HDP7</td>
<td>2023</td>
<td>2028</td>
<td>warranty</td>
<td>24 nodes and 24 JBODs</td>
</tr>
</tbody>
</table>
## Resources: virtual machines

### Production

<table>
<thead>
<tr>
<th>Resource</th>
<th>Quota</th>
<th>Usage</th>
<th>Usage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>38</td>
<td>35</td>
<td>92.11%</td>
</tr>
<tr>
<td>Cores</td>
<td>440</td>
<td>432</td>
<td>98.18%</td>
</tr>
<tr>
<td>RAM</td>
<td>901 GB</td>
<td>810 GB</td>
<td>89.89%</td>
</tr>
</tbody>
</table>

### Development

<table>
<thead>
<tr>
<th>Resource</th>
<th>Quota</th>
<th>Usage</th>
<th>Usage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>50</td>
<td>41</td>
<td>82.00%</td>
</tr>
<tr>
<td>Cores</td>
<td>300</td>
<td>259</td>
<td>86.33%</td>
</tr>
<tr>
<td>RAM</td>
<td>614 GB</td>
<td>486 GB</td>
<td>79.04%</td>
</tr>
</tbody>
</table>
Architecture: general-purpose cluster

**Namnode**
- NameNode
- ResourceManager
- JournalNode
- Zookeeper
- SparkHistory
- HBaseMaster
- HiveServer
- HiveMetastore
- PhoenixQueryServer

**Quorum**

**DataNode 0**
- DataNode
- NodeManager
- JournalNode
- Zookeeper
- HBaseRegionServer

**DataNode 1**
- DataNode
- NodeManager
- JournalNode
- Zookeeper
- HBaseRegionServer

**DataNode n**
- DataNode
- NodeManager
- HBaseRegionServer

04 Dec 2023
Pedro Andrade | Hadoop service overview
## Architecture: NXCALS

Based on same common architecture and components

However, due to NXCALS critically, service is provided via 2 distinct clusters:

<table>
<thead>
<tr>
<th>HDFS cluster</th>
<th>HBase cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>● 3 quorum nodes</td>
<td>● 3 quorum nodes</td>
</tr>
<tr>
<td>○ 2 namenodes</td>
<td>○ 2 namenodes</td>
</tr>
<tr>
<td>○ 2 Resource Managers</td>
<td>○ 2 HBase masters</td>
</tr>
<tr>
<td>○ Zookeeper quorum</td>
<td>○ Zookeeper quorum</td>
</tr>
<tr>
<td>○ Journal quorum</td>
<td>○ Journal quorum</td>
</tr>
<tr>
<td>● multiple datanodes</td>
<td>● multiple datanodes</td>
</tr>
<tr>
<td></td>
<td>○ HBase regionservers</td>
</tr>
</tbody>
</table>
## Clusters

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Environment</th>
<th>Provisioning</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytix</td>
<td>All</td>
<td>PROD</td>
<td>Bare metal</td>
<td>60</td>
</tr>
<tr>
<td>Nxcals Prod</td>
<td>HDFS</td>
<td>PROD</td>
<td>Bare metal</td>
<td>53</td>
</tr>
<tr>
<td>Nxcals Prod Online</td>
<td>HBase</td>
<td>PROD</td>
<td>Bare metal</td>
<td>32</td>
</tr>
<tr>
<td>Nxcals Perftest</td>
<td>HDFS</td>
<td>PROD</td>
<td>Bare metal</td>
<td>10</td>
</tr>
<tr>
<td>Nxcals Perftest Online</td>
<td>HBase</td>
<td>PROD</td>
<td>Bare metal</td>
<td>6</td>
</tr>
<tr>
<td>Nxcals Dev</td>
<td>HDFS</td>
<td>PROD</td>
<td>Virtual machines</td>
<td>10</td>
</tr>
<tr>
<td>Nxcals Dev Online</td>
<td>HBase</td>
<td>PROD</td>
<td>Virtual machines</td>
<td>10</td>
</tr>
<tr>
<td>Hadoop QA</td>
<td>All</td>
<td>QA</td>
<td>Virtual machines</td>
<td>10</td>
</tr>
<tr>
<td>Stager</td>
<td>All</td>
<td>DEV</td>
<td>Virtual machines</td>
<td>5</td>
</tr>
</tbody>
</table>
## Clusters

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Active Datanodes</th>
<th>HDFS Capacity</th>
<th>HDFS Usage</th>
<th>Number of Files</th>
<th>Active Region Servers</th>
<th>Database Size</th>
<th>Number of Regions</th>
<th>Number of Regions Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANALYTIX (HDFS + HBase)</strong></td>
<td>58</td>
<td>24.6 PB</td>
<td>31.6%</td>
<td>39.0 Mil</td>
<td>58</td>
<td>134 TB</td>
<td>14.0 K</td>
<td>23.4%</td>
</tr>
<tr>
<td><strong>NXCALS Prod (HDFS)</strong></td>
<td>51</td>
<td>18.4 PB</td>
<td>54.2%</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
</tr>
<tr>
<td><strong>NXCALS Online (HBase)</strong></td>
<td>30</td>
<td>343 TB</td>
<td>28.2%</td>
<td>591 K</td>
<td>30</td>
<td>14.0 TB</td>
<td>23.6 K</td>
<td>39.3%</td>
</tr>
</tbody>
</table>
Software

- Spark v3.5.0 (latest is v3.5.0)
- Hive v2.3.3 (latest is v3.1.0)
- Phoenix v5.1.0 (latest is v5.1.3)
- Zookeeper v3.6.1 (latest is v3.8.3)
- Hadoop v3.2.1 (latest is v3.3.6)
- HBase v2.3.4 (latest is v2.5.6)
Tech stack

Software Distribution
- Apache Hadoop
- Custom RPMs

Installation and Configuration
- CentOS 7.9/Alma 9
- Puppet: custom modules

Security
- AuthN (Kerberos)
- Fine-grained authZ (LDAP)

High Availability
- Automatic master failover

Changes Management
- Rolling approach
- No downtime
- Transparent most of the cases

Monitoring and SLS
- MONIT, custom Collectd plugin
- OpenSearch, Grafana
Disaster recovery

Inline with IT department BC/DR strategy

- **Protect the service against**
  - Data corruption
  - Data deletion
  - Cluster unavailability
  - Data center issues

- **Caused by**
  - Hardware failures (disk, node, rack)
  - Human errors (logical, accidental)
  - Application errors
  - Data mismanagement
  - Site failure
Disaster recovery

Hadoop scenarios when DR is needed
● One of the cluster nodes (with all the disks) goes down and we cannot recover it.
● The whole cluster goes down and we cannot restore the data.
● Some data is corrupted by a user or the system and needs to be restored back-in-time.

Implementation guidelines
● Make sure users/projects’ data and service metadata is safe
● Use as much as possible native Hadoop functionality
● Run the DR tools (backup and restore) as frequently as possible
● HDFS backups to CTA and HBase snapshot to external HDFS cluster
● Keep DR artifacts in separate DC, for now still in the same DC :(
Tips & tricks

- **Monitoring**
  - To visualize the clusters health and performance
  - [https://monit-grafana.cern.ch/?orgId=23](https://monit-grafana.cern.ch/?orgId=23)

- **HDFS Web UI**
  - To inspect cluster health, datanodes utilization, snapshot info, browse the filesystem
  - [Analytix, NXCALS](https://analytix.org)

- **YARN Web UI**
  - To inspect details of the applications status and history
  - [Analytix, NXCALS](https://analytix.org)

- **Spark Web UI**
  - To inspect details of spark jobs status and history
  - [Analytix, NXCALS](https://analytix.org)
Tips & tricks

- Client edge nodes
  - To connect and interact with the clusters
  - [https://hadoop.docs.cern.ch/getstart/client_edge_machine/](https://hadoop.docs.cern.ch/getstart/client_edge_machine/)
- QA/Dev cluster
  - To test your applications and workflows
- Email notifications
  - To alert you when your cluster is getting full
Communication

- **ServiceNow**
  - Open tickets to report problems or ask new features
  - Request access to clusters and quota modifications
  - [https://cern.service-now.com/service-portal?id=service_element&name=Hadoop-Service](https://cern.service-now.com/service-portal?id=service_element&name=Hadoop-Service)

- **Documentation**
  - To know more about the service
  - [https://hadoop.docs.cern.ch](https://hadoop.docs.cern.ch)

- **Mattermost**
  - For a direct support chat with the team
  - [https://mattermost.web.cern.ch/it-dep/channels/it-hadoop-service](https://mattermost.web.cern.ch/it-dep/channels/it-hadoop-service)

- **User Forum**
  - To discuss the service status and plans
Thank you for your attention
Ecosystem

Online Transactional system?

- Y
- N

Expected system growth < 1GB/day?

- Y
- N

Structured data?

- N
- Y

Structured data?

- Y
- N

Frequent analytics needed?

- Y
- N

Structured data?

- Y
- N

Analytics needed?

- Y
- N

This storage technology decision tree is just for demonstration, should not be used as a definitive guide!

PostgreSQL, MySQL, Oracle

HBase + PHOENIX

HDFS + columnar file format

HYBRID SYSTEM

Sqoop -> HDFS

HBase