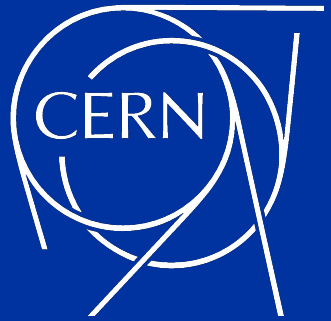


Hadoop User Forum #1

04.12.2023

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

Agenda

09:00 → 09:30 **Welcome coffee** 30m

09:30 → 10:20 **Hadoop Service**

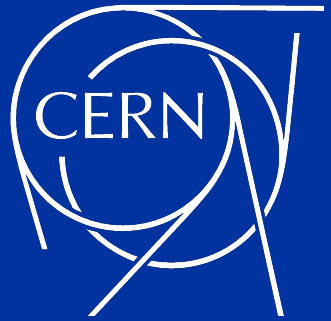
- **09:30** **Status** 20m
  Overview of the service evolution and presentation of main changes in the last year
  Speaker: Pedro Andrade (CERN)
- **09:50** **Roadmap** 30m
  Operational goals for next year and long-term service strategy
  Speaker: Emil Kleszcz (CERN)
  20231204 - Hadoop...

10:20 → 10:30 **Break** 10m

10:30 → 11:30 **Hadoop User Communities**

- **10:30** **ATLAS EventIndex Team** 15m
  Speaker: Grigori Rybkin (Université Paris-Saclay (FR))
  ATLAS EventIndex i...
- **10:45** **MONIT Team** 15m
  Speaker: Nikolay Tsvetkov (CERN)
- **11:00** **NXCALS Team** 15m
  Speaker: Jakub Wozniak (CERN)
- **11:15** **SWAN Team** 15m
  Speaker: Enric Tejedor Saavedra (CERN)
  SWAN @ Hadoop u...

11:30 → 12:00 **Discussion** 30m



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

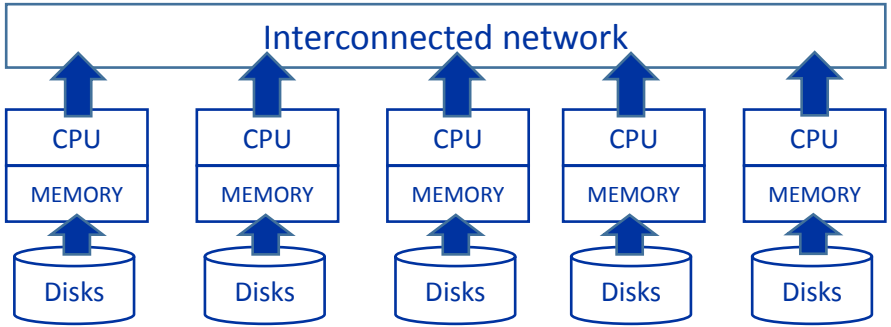
Introduction

What is Hadoop ?

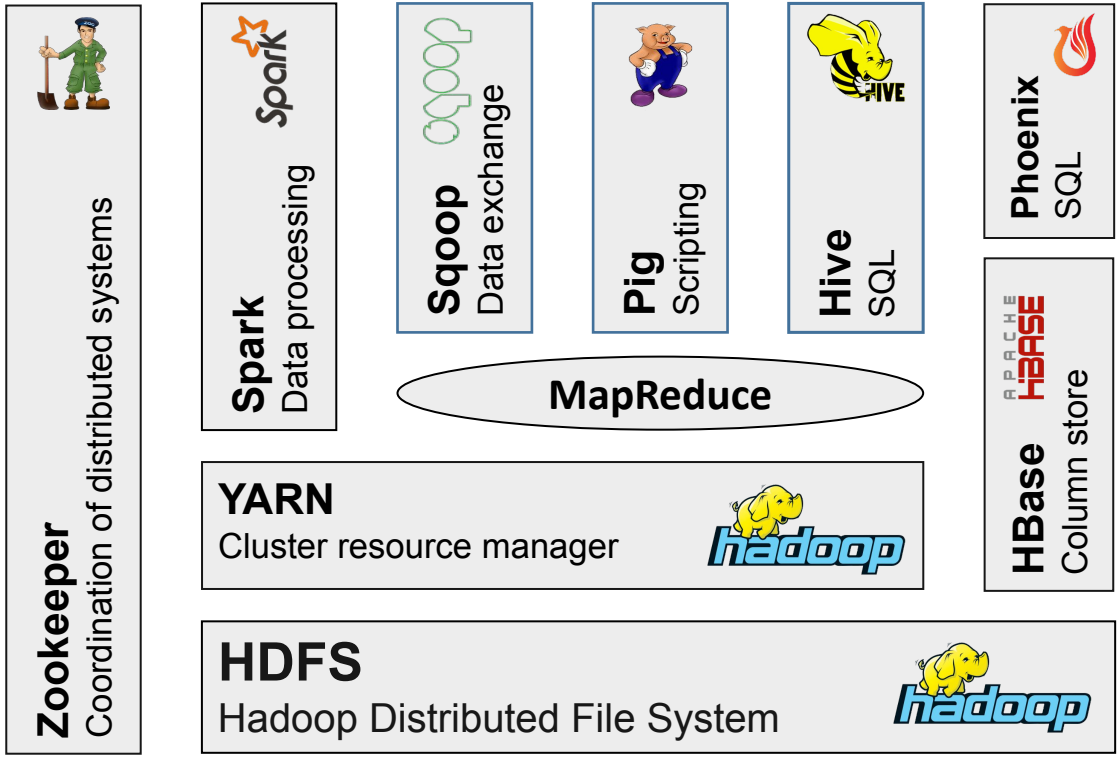
“An open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.”

Why Hadoop ?

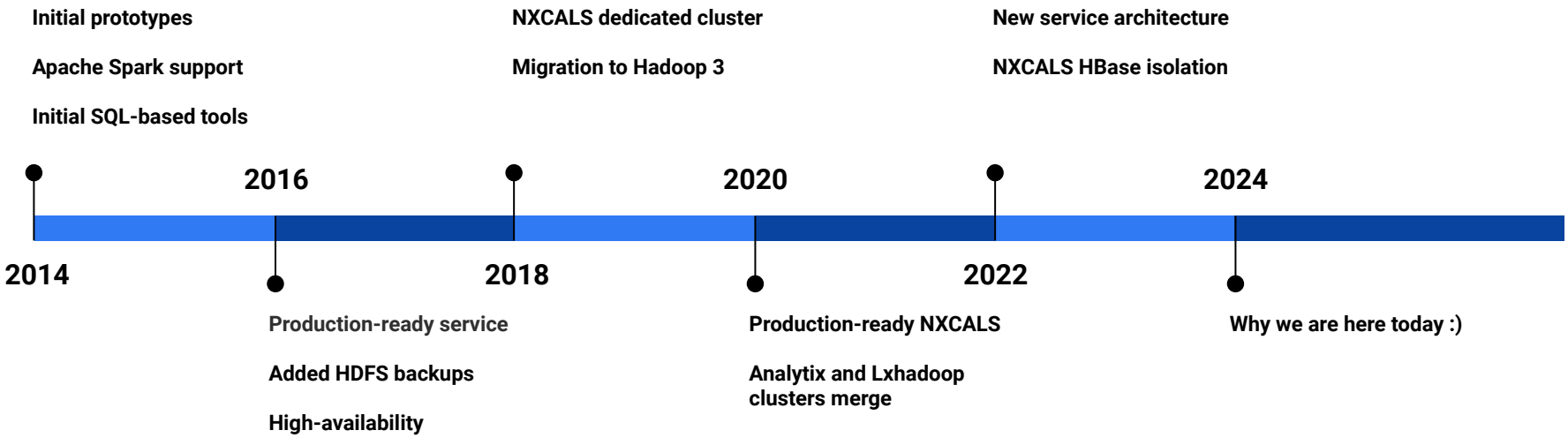
- Too much Data (TB per day)
- Vertical scaling doesn’t cut it
- Horizontal scaling is linear



Ecosystem


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

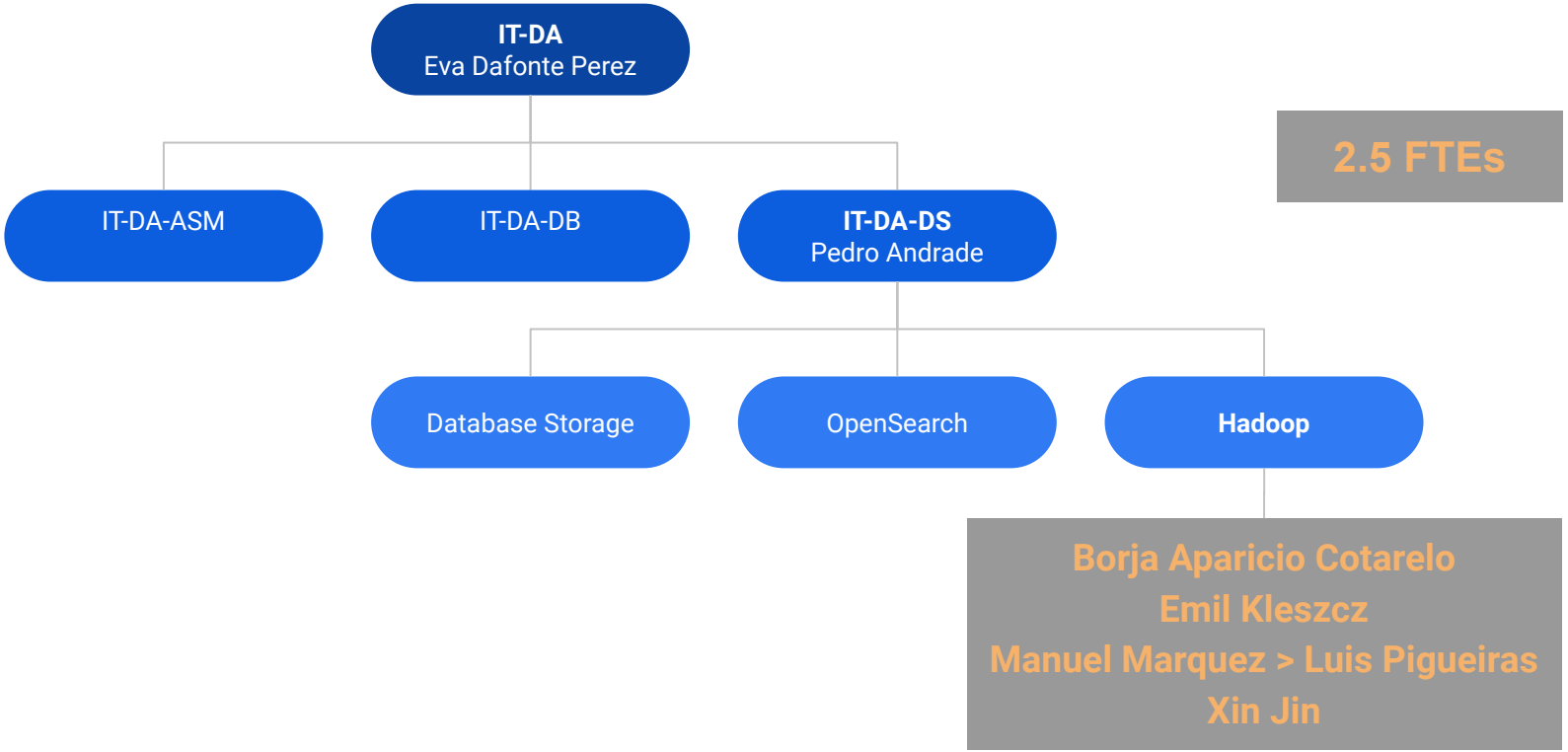
History


Initial prototypes
Apache Spark support
Initial SQL-based tools
NXCALS dedicated cluster
Migration to Hadoop 3
New service architecture
NXCALS HBase isolation
2014
2016
2018
2020
2022
2024
Production-ready service
Added HDFS backups
High-availability
Production-ready NXCALS
Analytix and Lxhadoop clusters merge
Why we are here today :)

People

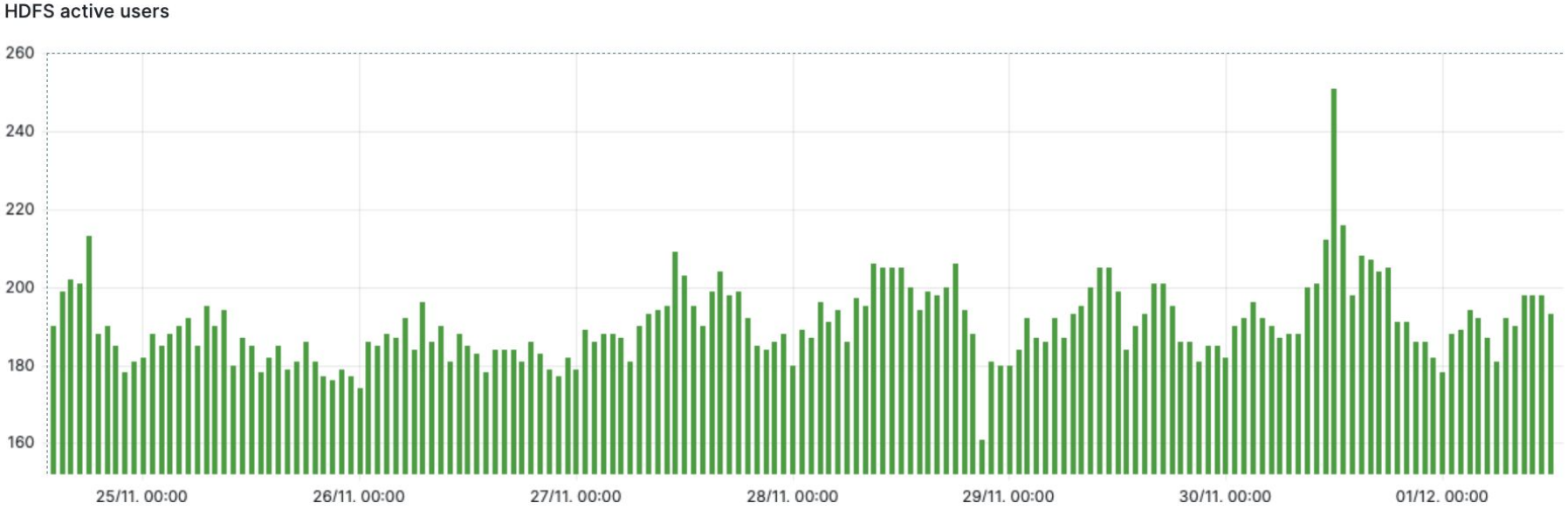
Team - Users - Projects

Team

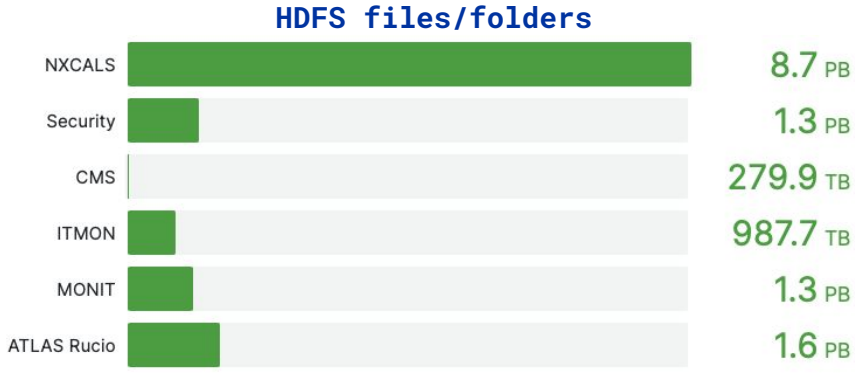

IT-DA
Eva Dafonte Perez
IT-DA-ASM
IT-DA-DB
IT-DA-DS
Pedro Andrade
Database Storage
OpenSearch
Hadoop
2.5 FTEs
Borja Aparicio Cotarelo
Emil Kleszcz
Manuel Marquez > Luis Pigueiras
Xin Jin

Users

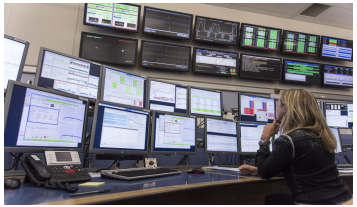
1721 registered users



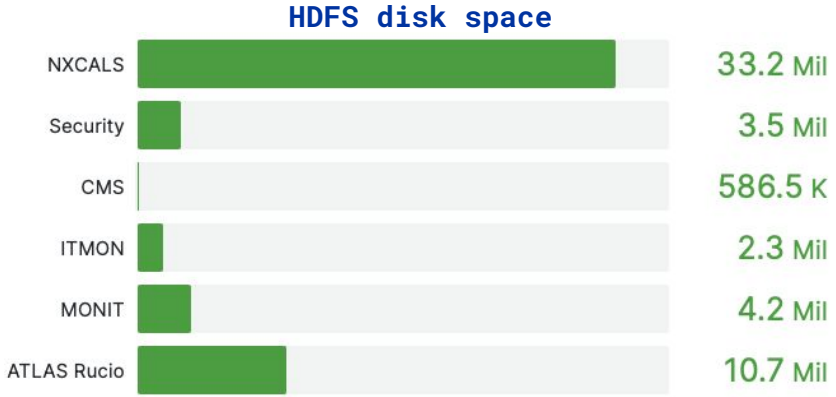
Projects



ATLAS EXPERIMENT



CMS



Infrastructure

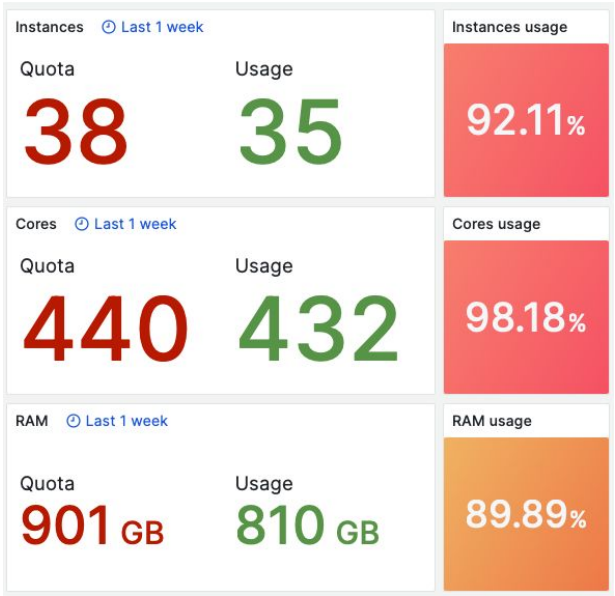
Resources - Architecture - Clusters - Tech Stack

Resources: bare metal

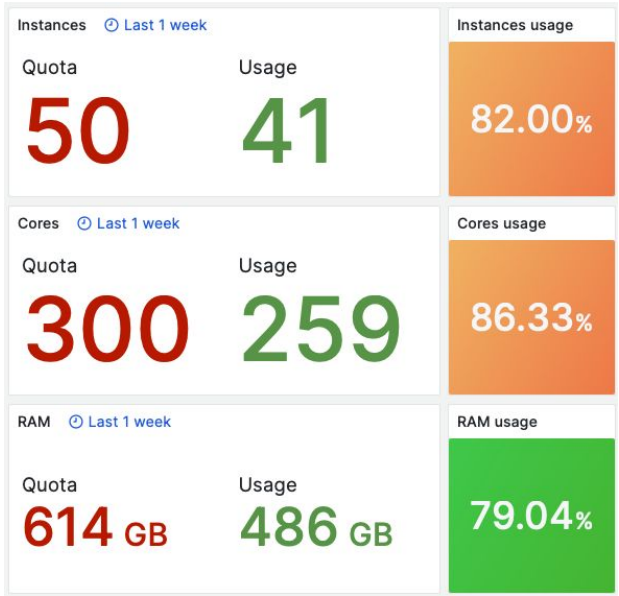
Order	Installation	End of life	Status	Description
L3	2017	2024	IT support	24 nodes and 6 JBODs
HDP1	2017	2022	retired	8 nodes and 16 JBODs
HDP2	2018	2023	IT support	42 nodes and 76 JBODs
HDP3	2019	2024	IT support	8 nodes and 16 JBODs
HDP4	2020	2025	IT support	16 nodes and 32 JBODs
HDP5	2021	2026	warranty	60 nodes and 32 JBODs
HDP6	2022	2027	warranty	20 nodes and 15 JBODs
HDP7	2023	2028	warranty	24 nodes and 24 JBODs

Resources: virtual machines

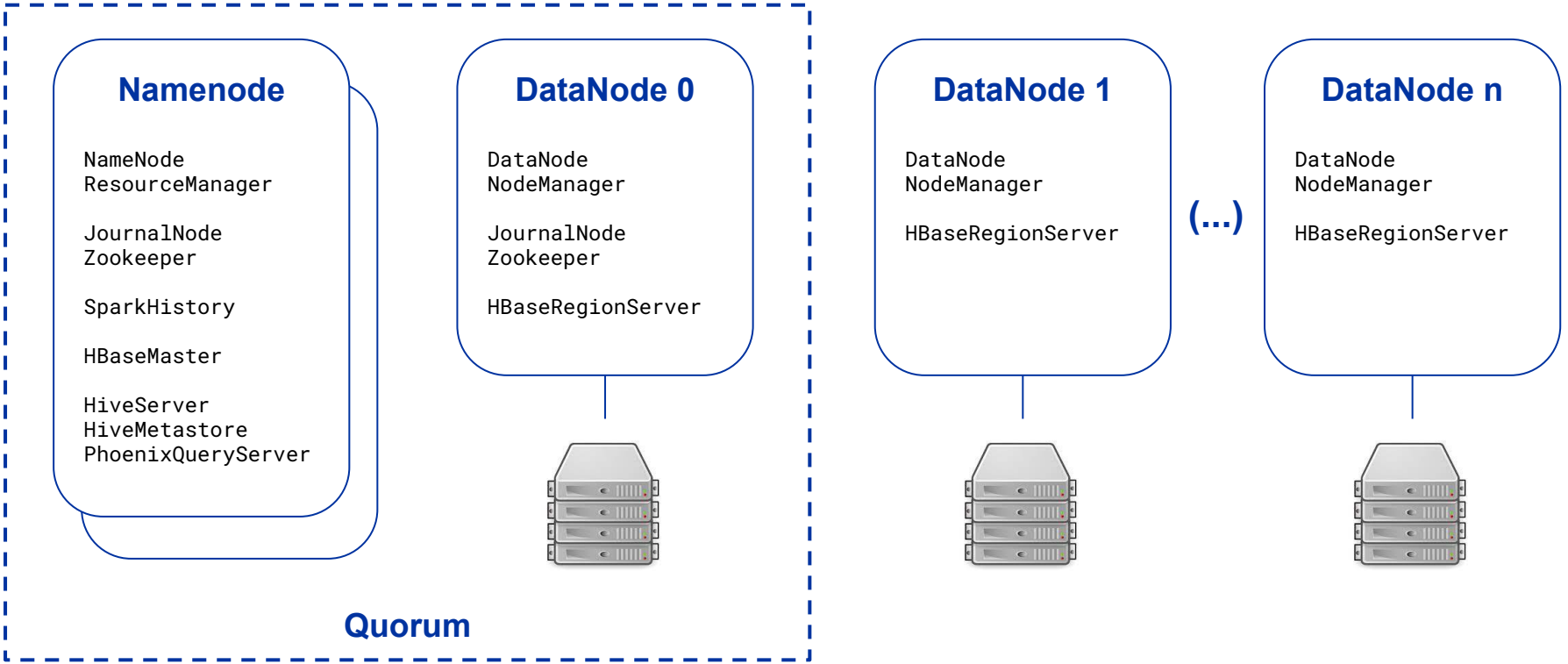
Production



Development



Architecture: general-purpose cluster



Architecture: NXCALS

Based on same common architecture and components

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

HDFS cluster

- 3 quorum nodes
  - 2 namenodes
  - 2 Resource Managers
  - Zookeeper quorum
  - Journal quorum
- multiple datanodes

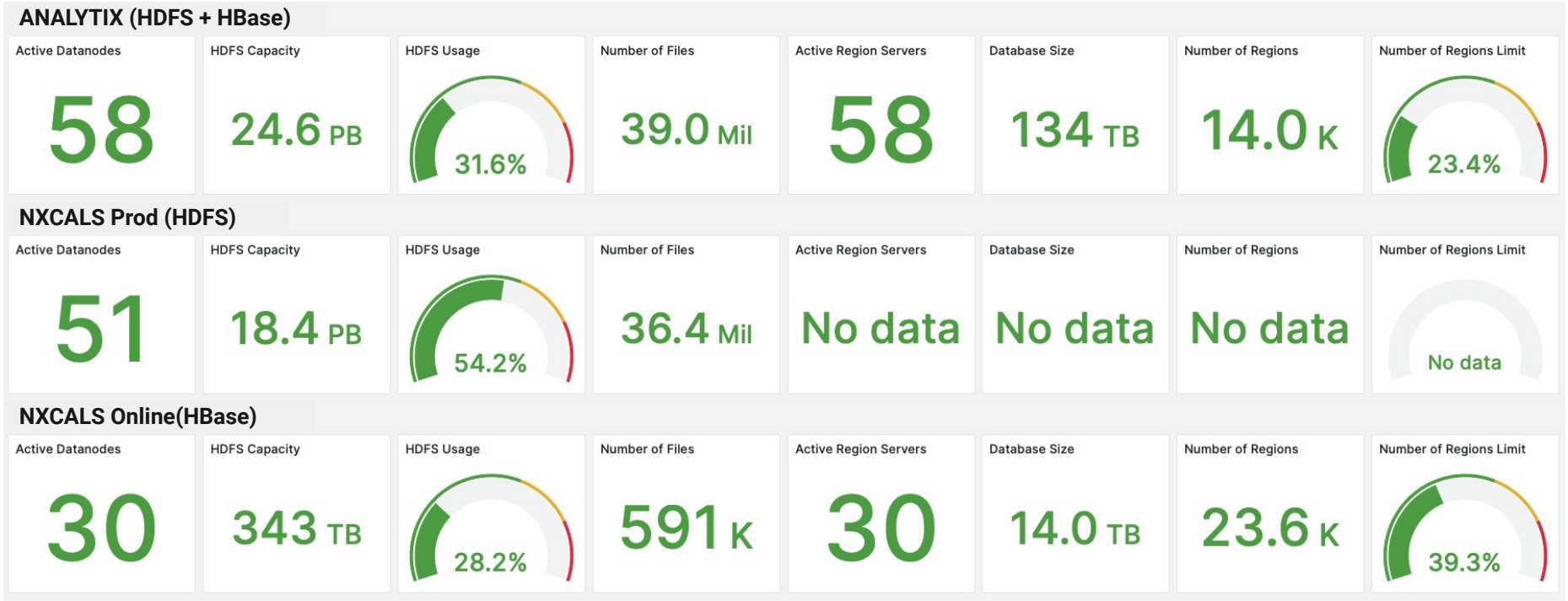
HBase cluster

- 3 quorum nodes
  - 2 namenodes
  - 2 HBase masters
  - Zookeeper quorum
  - Journal quorum
- multiple datanodes
  - HBase regionservers

Clusters

Name	Type	Environment	Provisioning	Nodes
Analytix	All	PROD	Bare metal	60
Nxcals Prod	HDFS	PROD	Bare metal	53
Nxcals Prod Online	HBase	PROD	Bare metal	32
Nxcals Perftest	HDFS	PROD	Bare metal	10
Nxcals Perftest Online	HBase	PROD	Bare metal	6
Nxcals Dev	HDFS	PROD	Virtual machines	10
Nxcals Dev Online	HBase	PROD	Virtual machines	10
Hadoop QA	All	QA	Virtual machines	10
Stager	All	DEV	Virtual machines	5

Clusters


ANALYTIX (HDFS + HBase)
Active Datanodes
58
HDFS Capacity
24.6 PB
HDFS Usage
31.6%
Number of Files
39.0 Mil
Active Region Servers
58
Database Size
134 TB
Number of Regions
14.0 K
Number of Regions Limit
23.4%
NXCALS Prod (HDFS)
Active Datanodes
51
HDFS Capacity
18.4 PB
HDFS Usage
54.2%
Number of Files
36.4 Mil
Active Region Servers
No data
Database Size
No data
Number of Regions
No data
Number of Regions Limit
No data
NXCALS Online(HBase)
Active Datanodes
30
HDFS Capacity
343 TB
HDFS Usage
28.2%
Number of Files
591 K
Active Region Servers
30
Database Size
14.0 TB
Number of Regions
23.6 K
Number of Regions Limit
39.3%

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

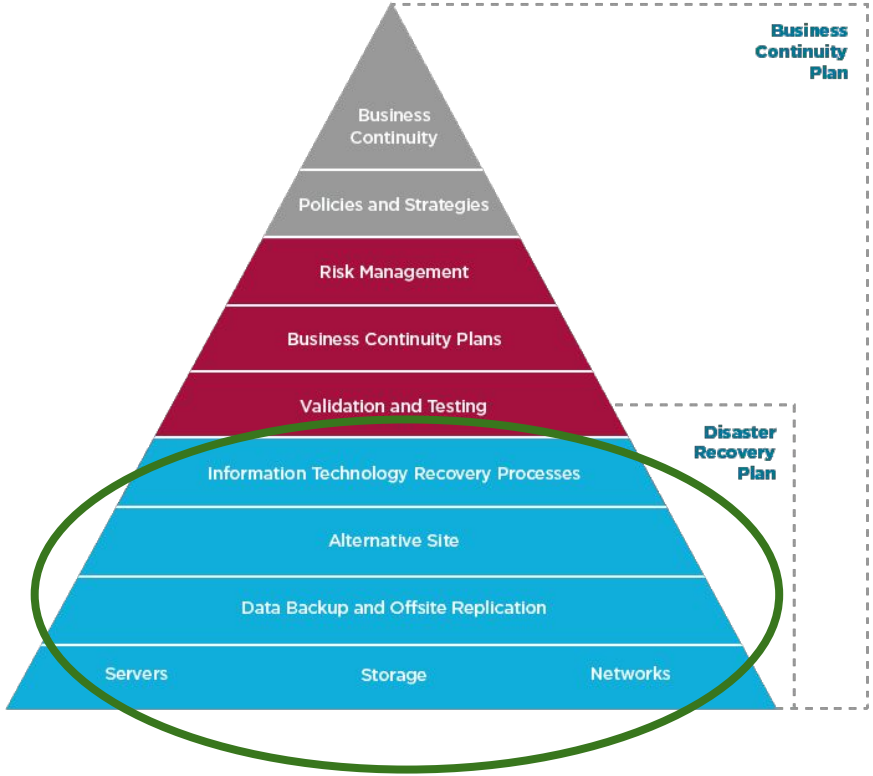
Monitoring and SLS

- 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 :(

User information

Tips & Tricks - Communication

Tips & tricks

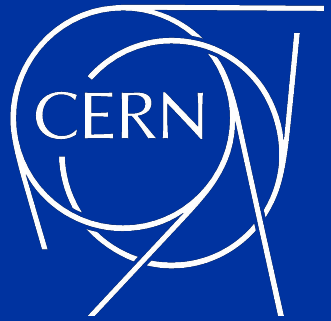
- **Monitoring**
  - To visualize the clusters health and performance
  - https://monit-grafana.cern.ch/?orgId=23
- **HDFS Web UI**
  - To inspect cluster health, datanodes utilization, snapshot info, browse the filesystem
  - Analytix, NXCALS
- **YARN Web UI**
  - To inspect details of the applications status and history
  - Analytix, NXCALS
- **Spark Web UI**
  - To inspect details of spark jobs status and history
  - Analytix, NXCALS

Tips & tricks

- **Client edge nodes**
  - To connect and interact with the clusters
  - https://hadoop.docs.cern.ch/getstart/client_edge_machine/
- **QA/Dev cluster**
  - To test your applications and workflows
- **Email notifications**
  - To alert you when you 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
- **Documentation**
  - To know more about the service
  - 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
- **User Forum**
  - To discuss the service status and plans

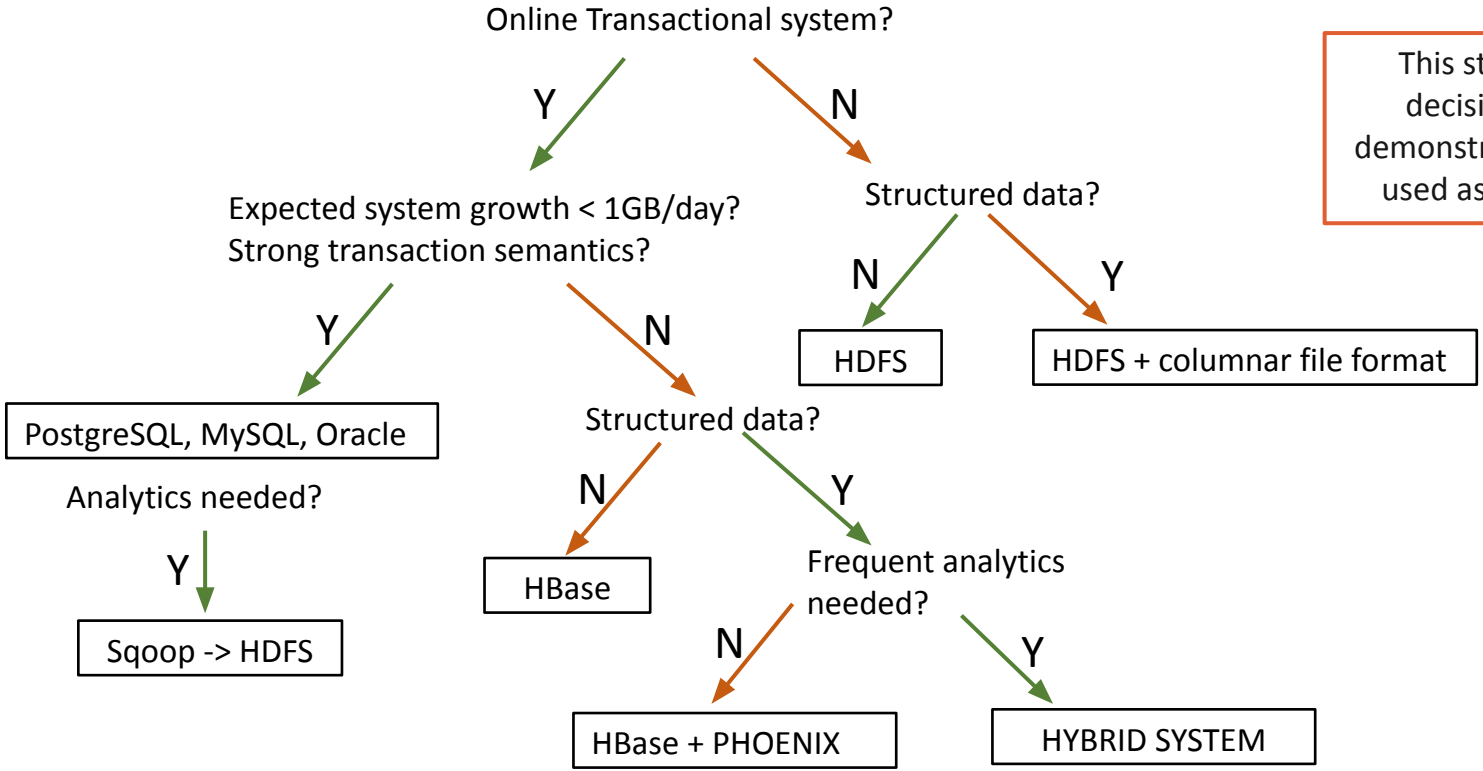


Thank you for your attention

Pedro Andrade @ CERN

04.12.2023

Ecosystem



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