Apache Big-Data

Cloud-native

USABILITY

+COLLABORATION

Agenda

- Border area
- [DEMO] Recycling YAML files
- Cars and washing machines
- [DEMO] Collaboration and Usability

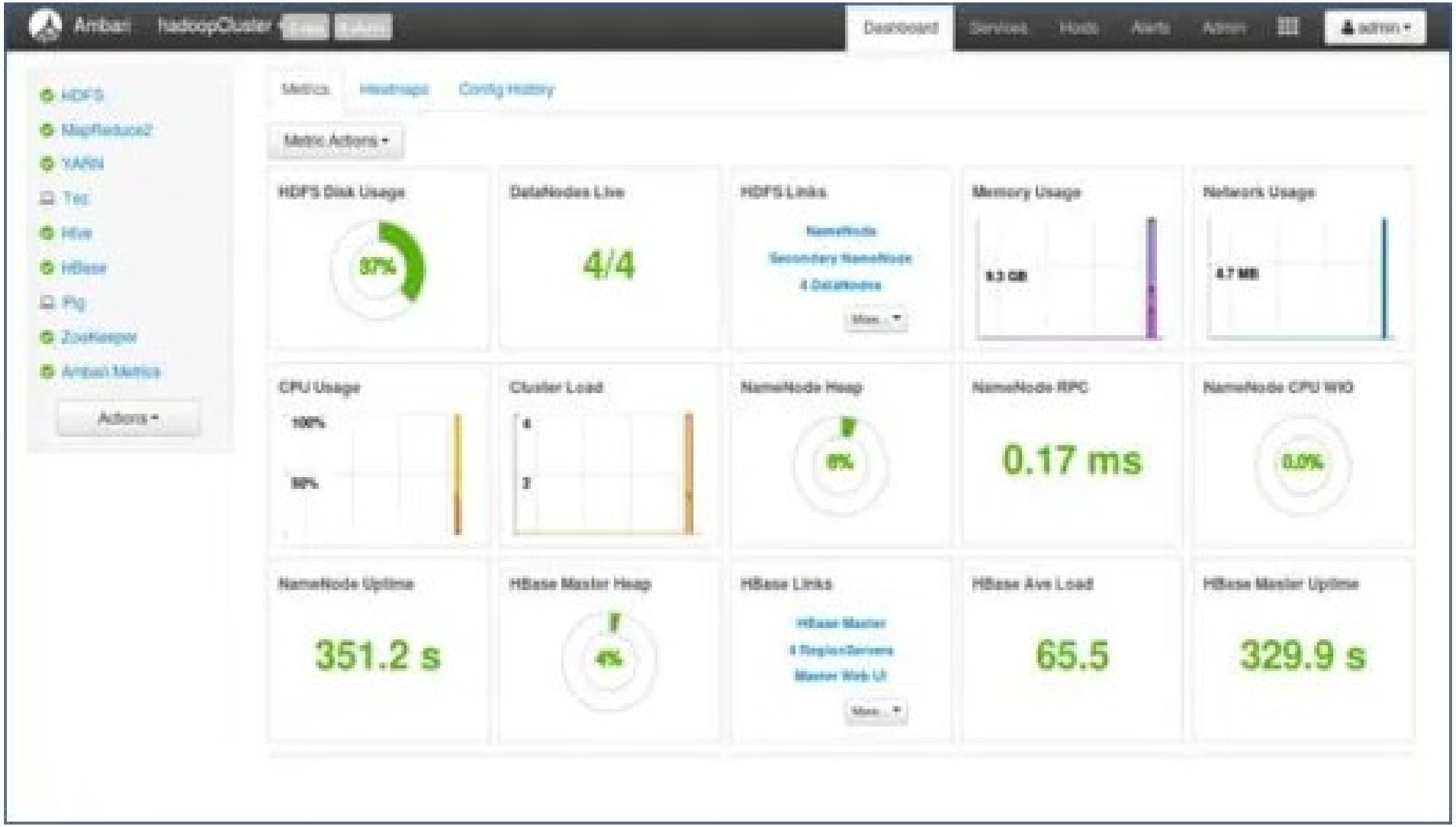
Márton Elek

- Apache Hadoop/Ratis PMC
  - elek@apache.org
- Cloudera
  - Principal software engineer
- twitter.com/@anzix
- Ozone:
  - Ozone Explained (Youtube)
- Kubernetes + Apache Bigdata:
  - **github.com/elek**/flekszible
  - flokkr.github.io

Márton Elek

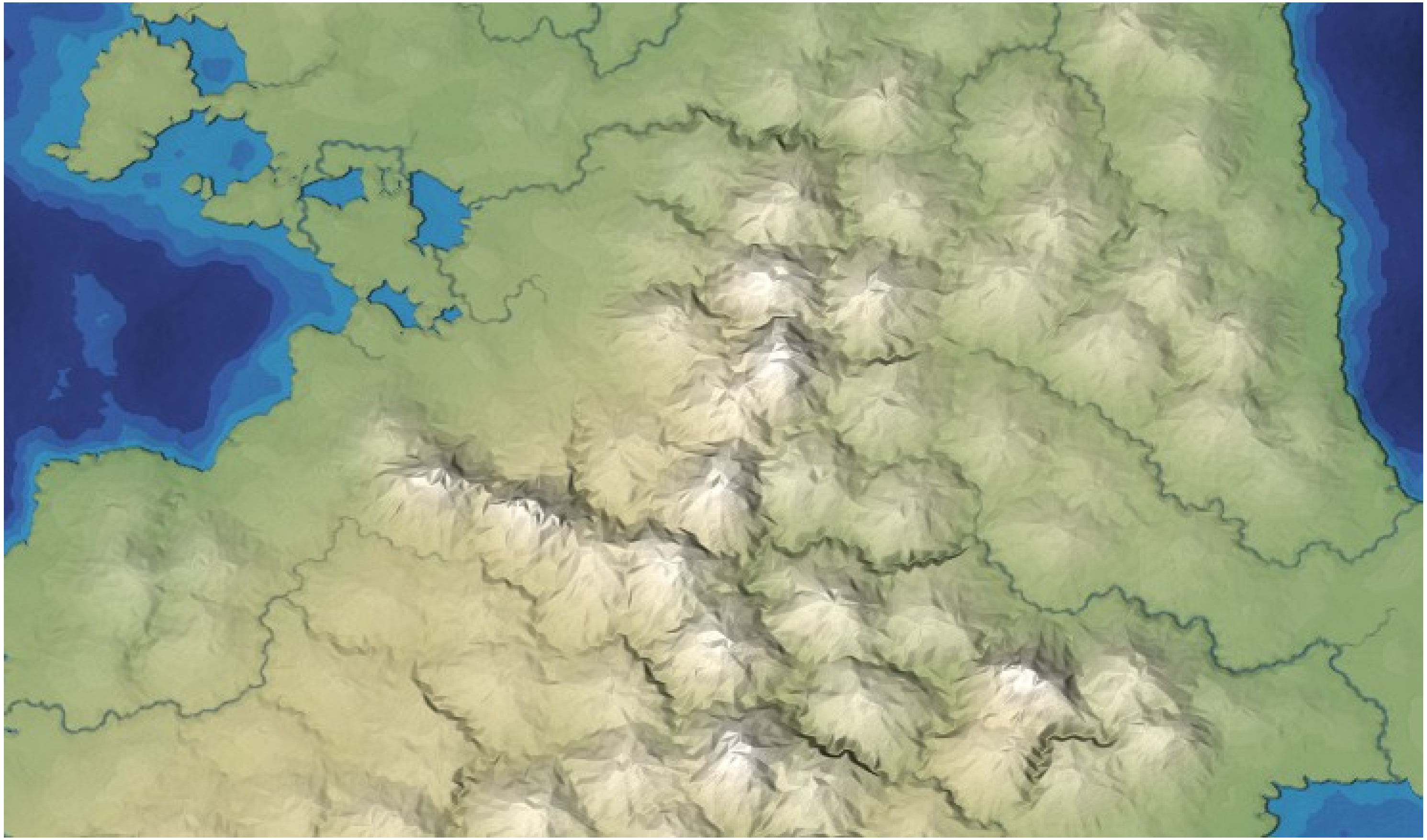
- Apache Hadoop/Ratis PMC
  - elek@apache.org
- Cloudera
  - Principal software engineer
- twitter.com/@anzix
- Ozone:
  - Ozone Explained (Youtube)
- Kubernetes + Apache Bigdata:
  - **github.com/elek**/flekszible
  - flokkr.github.io

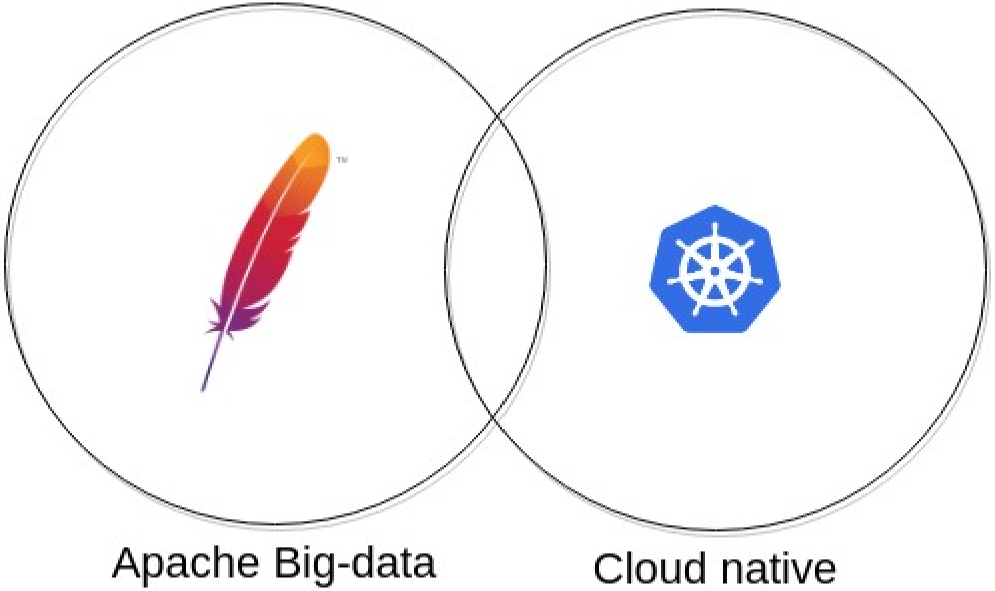
How to manage cluster?



Borders

Nobody’s land



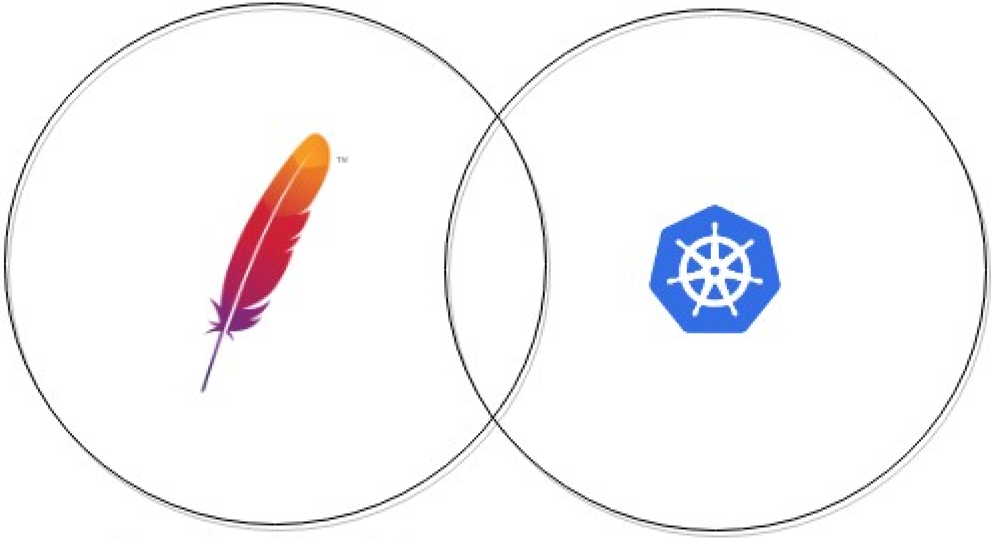


Apache Big-data

Cloud native

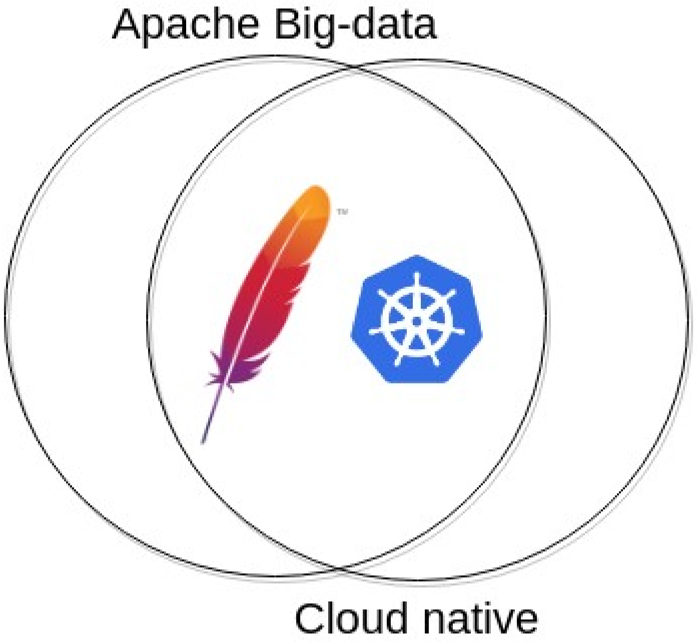
Apache Big-data

Cloud native



Apache Big-data

Cloud native

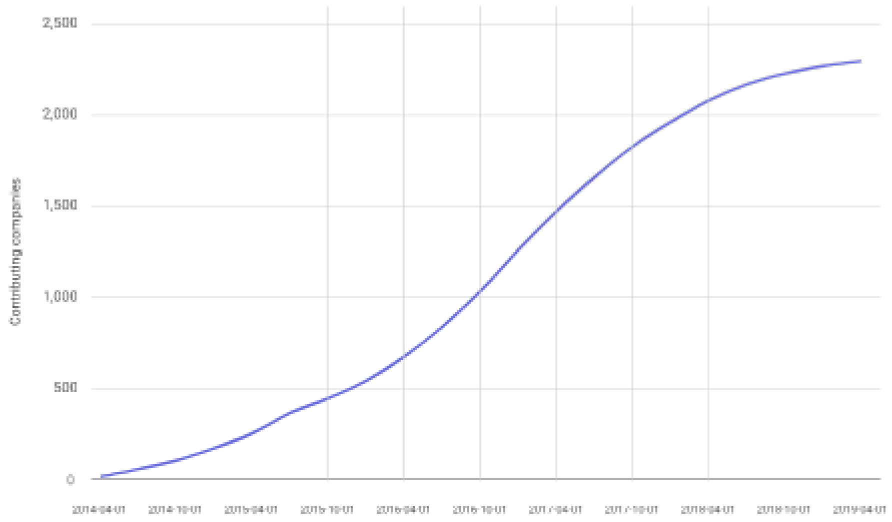


Apache Big-data
Cloud native

Cloud-native

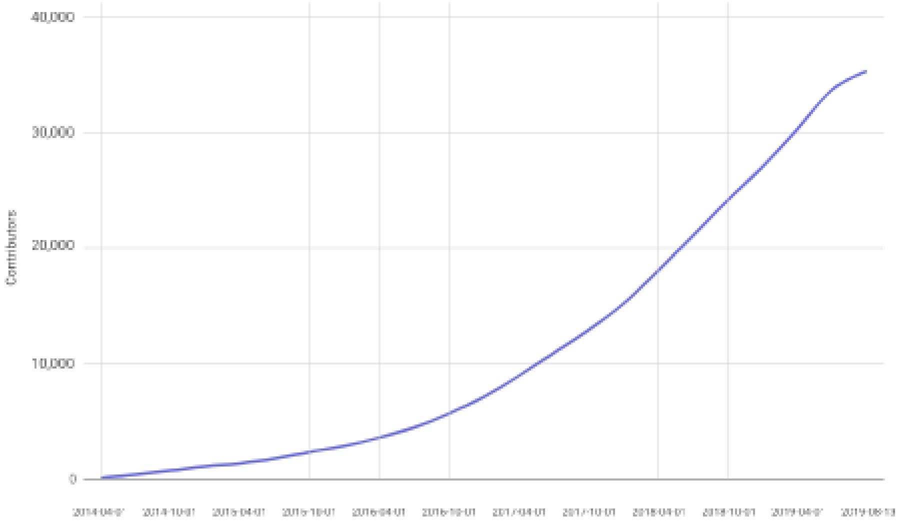
Why is it so popular?

Growth: Contributing Companies and Contributors



Cumulative growth of number of contributing companies

`



Cumulative growth of contributors

Why is it so popular?

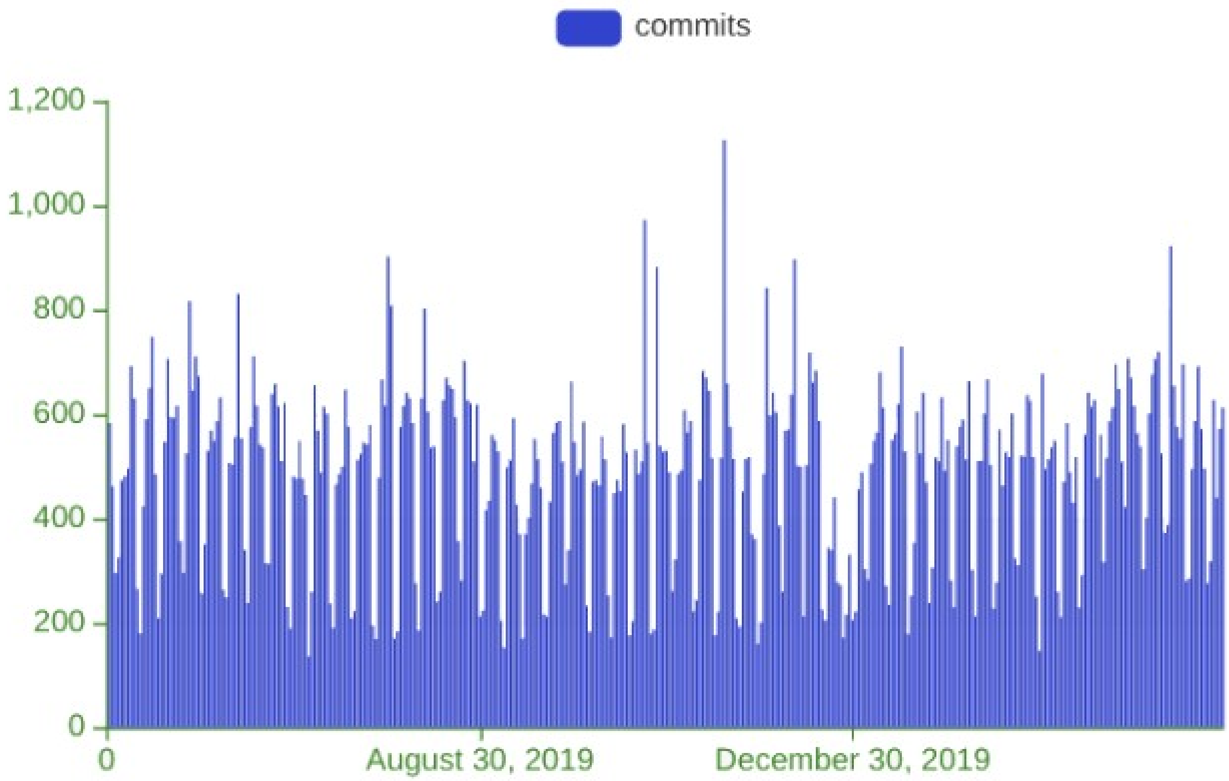
- Good low level design
- Flexible and generic api
  - apiVersion/kind → CRUD
- YAML files everywhere
- → Very popular

COLLABORATION

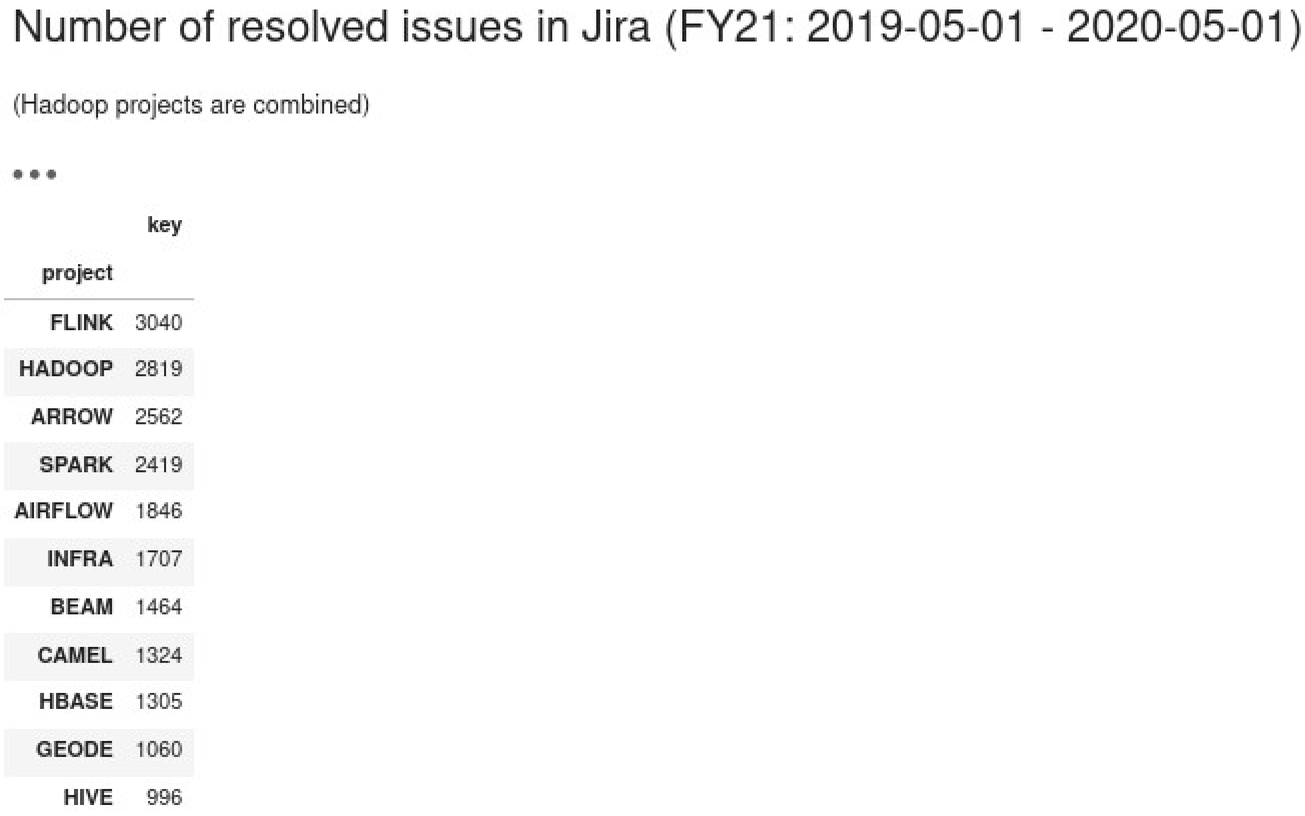
Apache Big-data

Apache ...

Commit History



Apache … Big-Data



Number of resolved issues in Jira (FY21: 2019-05-01 - 2020-05-01)

(Hadoop projects are combined)

•••

	key
project	
FLINK	3040
HADOOP	2819
ARROW	2562
SPARK	2419
AIRFLOW	1846
INFRA	1707
BEAM	1464
CAMEL	1324
HBASE	1305
GEODE	1060
HIVE	996

Apache Big-data ecosystem

- It’s an ecosystem
  - Collaboration is in the genes
- Hadoop compatible file system
  - Used by Spark, Flink, Hive, Hbase…
- Hadoop RPC/Configuration
  - Used by many projects

COLLABORATION

Nobody is an island

- Most of big-data projects used **together** with others
  - Kafka alone?, Spark alone?
- There are only a few meta-project
  - Apache Ambari: cluster management
  - Apache Bigtop: packaging in the same way

COLLABORATION?

Collaboration

- Why don’t we have more collaborative, meta-projects?
- Projects are complex and hard to configure them
  - Complexity is increasing with each project
- Can K8s help in the collaboration?

USABILITY?

Recycling

K8s resource management?

- Helm
  - No real collaboration, not flexible
- Kustomize
  - Transformations are not re-usable
- Any new tool?
  - Reusable components + transformations

github.com/elek/flekszible

- Kubernetes Yaml file manager
  - Reusable k8s files
  - **Reusable transformations (!!!)**
- → Final cluster definition

DEMO

flokkr.github.io

- Docker images for Apache big-data projects
- Reusable k8s files + **transformations!!**
- Did we solve all the problems:
  - COLLABORATION: DONE
  - USABILITY: ???

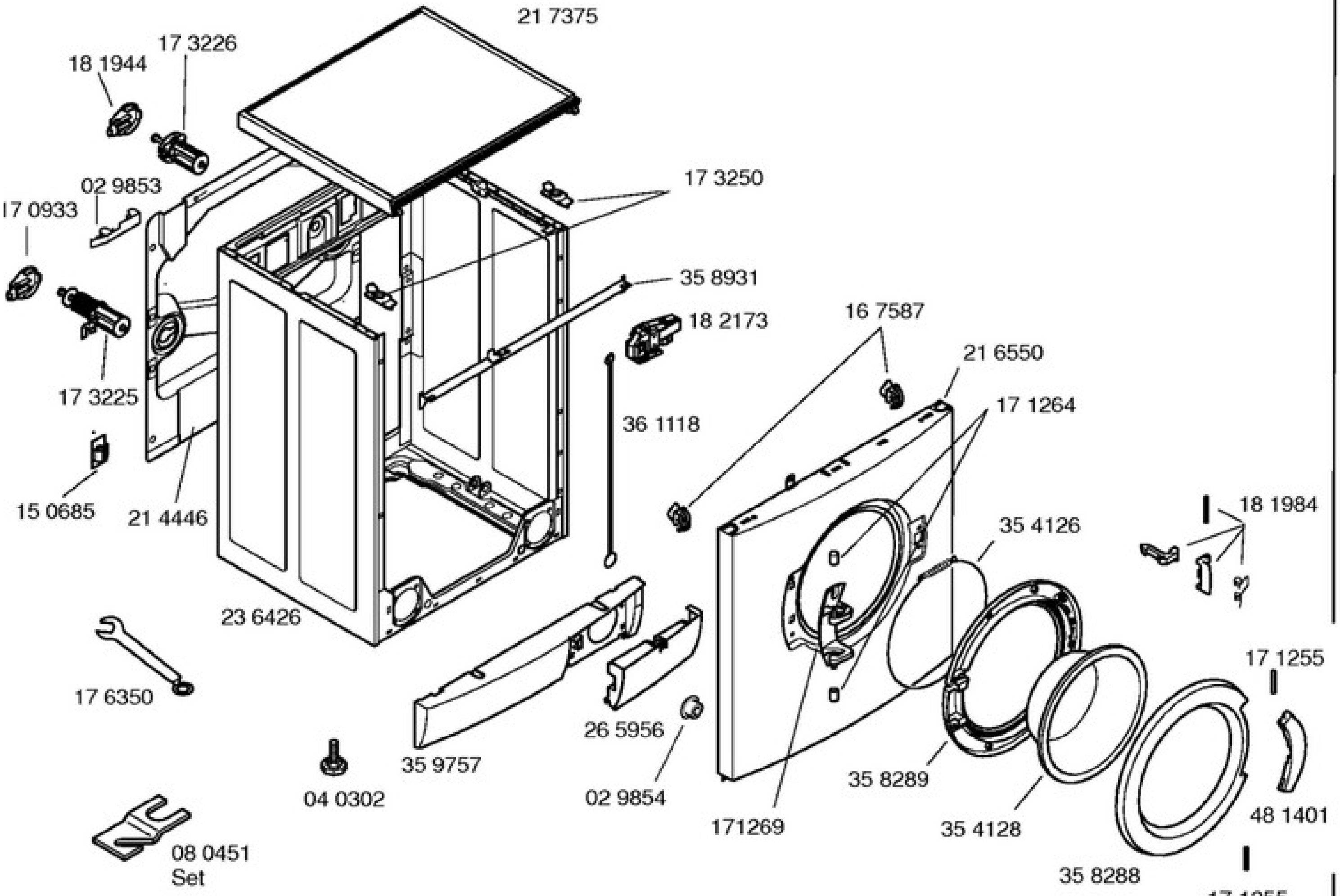
Config granularity

H [redacted]

BOSCH Maxx comfort
Reduced Ironing
Rinse Plus
Intensive Wash
Menu
Change
Additional Options
Start/ Pause
Easy-care
Cottons+Linens
Off
30
40
60
Delicates/Silk 30
cold
Woollens 30
30°
40
60
60 Stains
90
30
Mix
Drain
Rinse
Spin
WFR 2830
Water Conservation Rating
AAA
BOSCH
2
YEAR
WARRANTY

Mat. – Nr. – Konstante
3740 . . . . .

Dämmatten–Set
21 6336



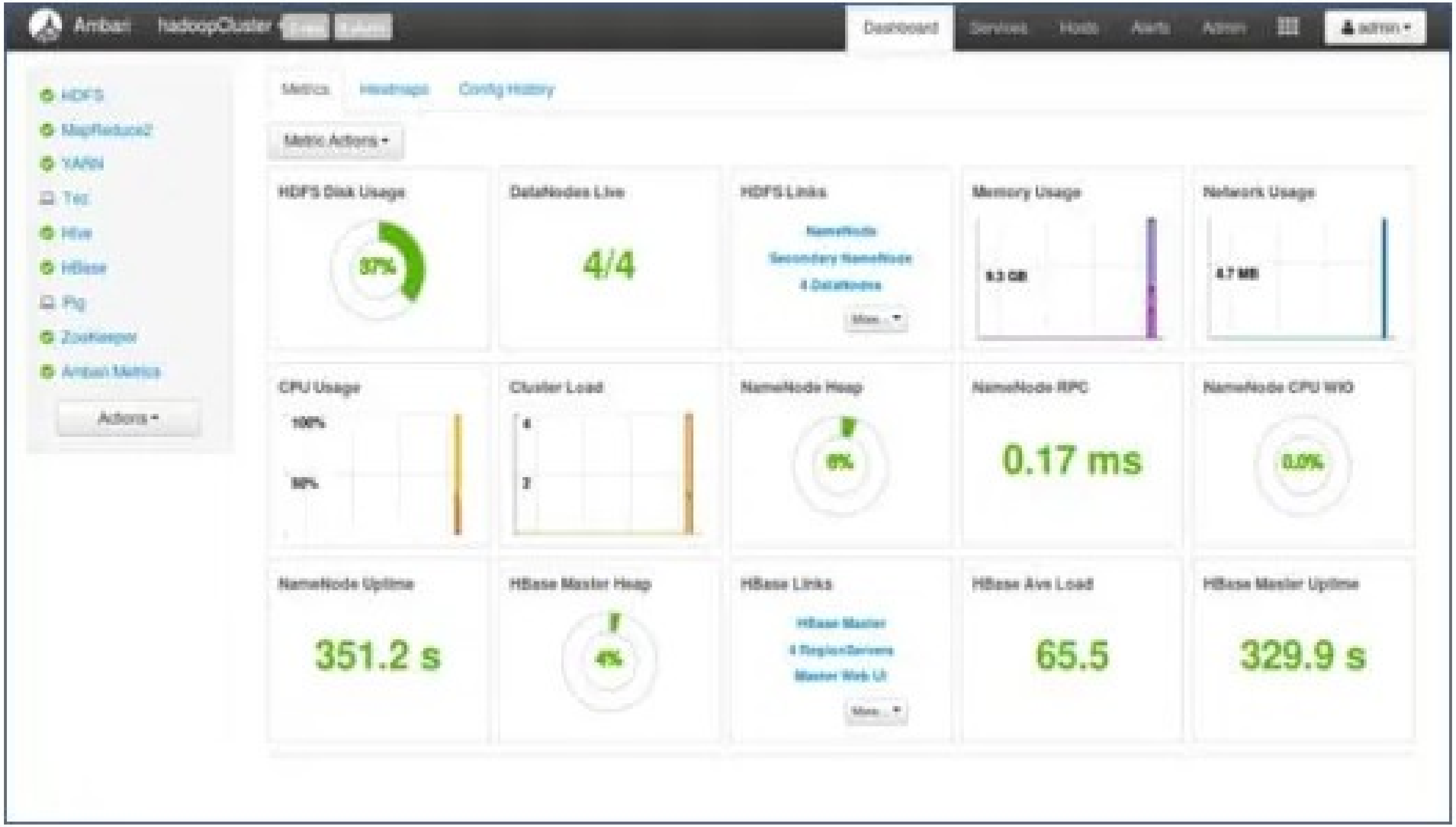
121921-5/2

Usability

- The challenge is to find the right abstraction level
  - [Top] Give me a secure HDFS+Kafka
  - ???
  - [Bottom]: HDFS-SITE.XML_dfs.checksum.type: "NULL"

Cluster management

How to manage cluster?



COLLABORATION

+USABILITY

DEMO

Summary

- Kubernetes is a good tool to improve **Collaboration** with increased **Usability**
- But we need good tool with the good abstraction
- New Apache “meta-projects” can be useful
  - → Incubator? (if interested → elek@apache.org)

Márton Elek

- elek@apache.org
- twitter.com/anzix
- Ozone:
  - Ozone Explained (Youtube)
- Kubernetes + Apache Bigdata:
  - **github.com/elek**/flekszible
  - flokkr.github.io