
KubeCon
CloudNativeCon
Europe 2023
TiKV

KubeCon | CloudNativeCon

Europe 2023

Building Apache druid on Kubernetes How Dailymotion Serves Partner Data

druid

Cyril Corbon & Alex Triquet

Bonjour! 💙

Alex Triquet

Resident Apache Foundation Fan

Fluent in Data-Engineer

Joined Dailymotion in 2019

Cyril Corbon

Opensource enthusiast / Techpodcast Addict

Technical debt killer

Joined Dailymotion in 2021

Our Team: DevOps Craft

Supports Teams from Dev to Production

Development

- Helps all data and software engineers
- Provides tooling and components to develop faster
- Supports them in the event of issues

Build and Integration

- Two Release Platforms: one for Adtech and one for the Video platform
- Dozens of builds and tests per day
- Multiple deployments to production every day

Production

- Manages the cloud infrastructure
- Geo-distributed platform both on the cloud and on premises
- Helps teams to build, deploy and run resilient applications
- Level 1 on-duty to ensure the availability of each component of the platform

Dailymotion Ecosystem

B2C

GLOBAL DESTINATION PLATFORM DAILYMOTION.COM

REBUILT

- **390 millions** active users
- **4 billions** views per month
- **92%** of audience on quality content
- **145** countries

Technology Enablers

WORLD-CLASS SAAS VIDEO TECHNOLOGY

REBUILT

- State of the art video technology for a flawless experience
- Include Player, CMS, content delivery and insights
- Distribution across all devices
- Highly scalable hybrid API architecture

END-TO-END ADVERTISING SOLUTION

BUILT

- Best in class brand safe proprietary ad platform
- Integrated with all major demand sources (DV360, The Trade Desk, Amazon…)
- Innovative ad formats and actionable audience segments
- Video performance benchmark above industry
- Independent from Google, Facebook and AWS

B2B

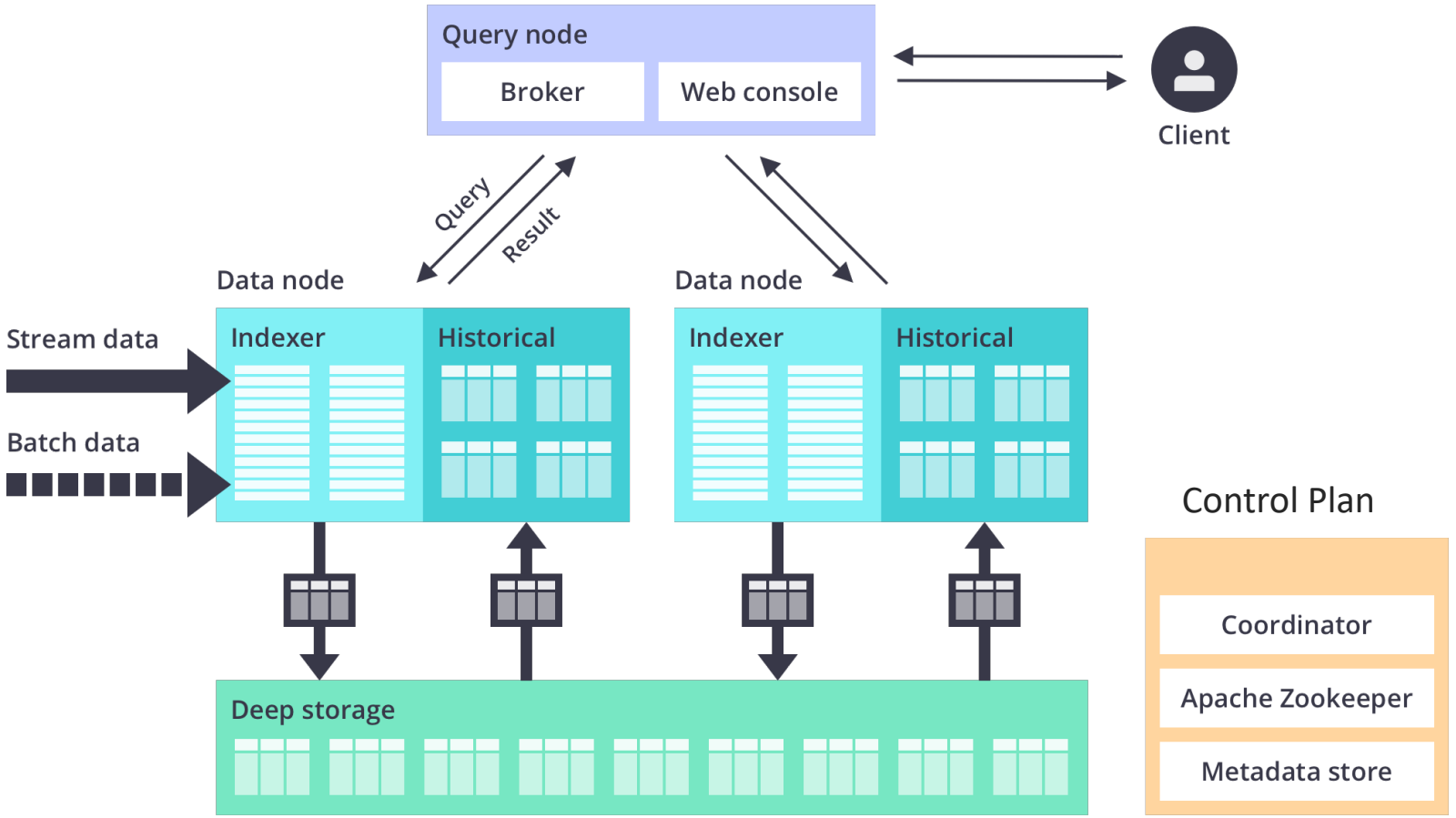
GLOBAL NETWORK OF PREMIUM PUBLISHERS

EXPANDED

- Unique network of **7,000+** publishers
- First-class content, cataloguing millions of videos worldwide
- Constantly refreshed


KubeCon
CloudNativeCon
Europe 2023
What is
Apache Druid?

Druid : Architecture



Druid Architecture - Schema copyright ©2020 Apache Software Foundation

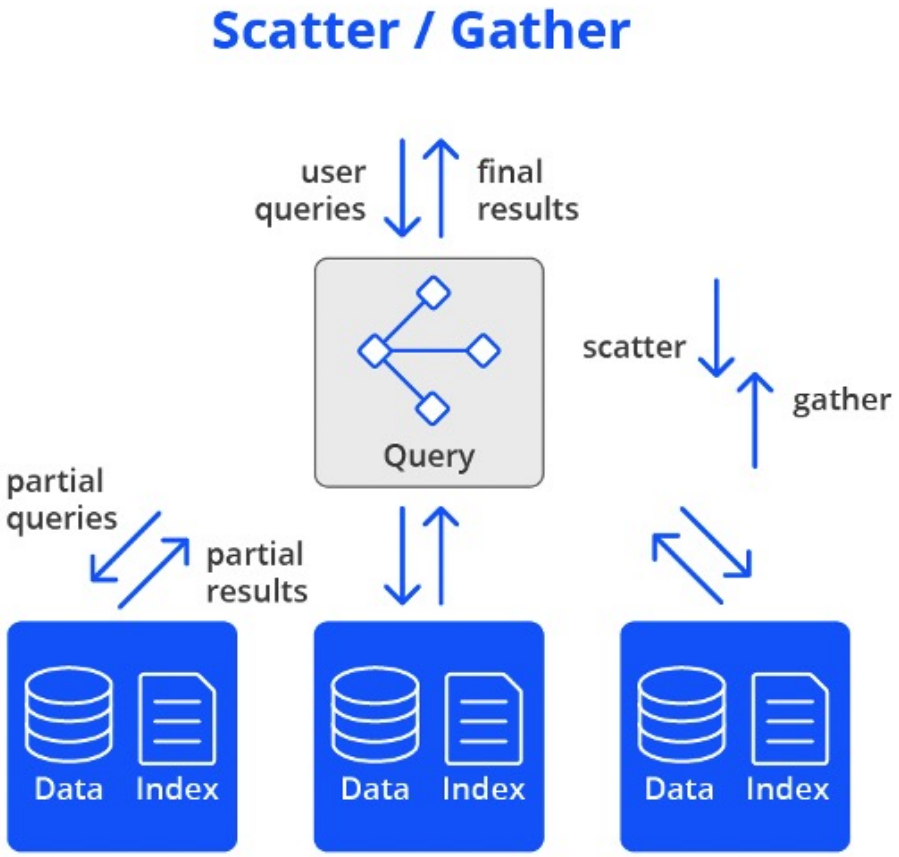
Druid : Query Flow

We use **Apache Druid** to perform aggregation on data that is already pre-aggregated and partitioned by time. This ensures best performances and full analytics capacities

Those time partitions are called segments and contains dimensions (strings) and metrics (floats)

QUERY EXECUTION STEPS

- A Broker node will receive the user queries and scatter it to the Historical nodes
- Historical nodes each execute the partial query on their segment
- Broker node finally gather the partial results and perform the final aggregation of results



TYPICAL QUERY FLOW

Druid : Value Proposition

PROS

- MultiDimensional Queries
- Very Fast
- Scales easily for both data/requests by taking advantage of both horizontal and vertical scaling
- ColumnarDB, a boon for data-engineers
- Documentation is great

CONS

- Setup is complex (separate metadata DB, requires a Zookeeper, a deepstorage and other components)
- Some big and stateful Nodes (Historicals)
- Configuration can be headache inducing, especially if you want to make the most of your resources
- Only suits specific (time-based analytics) needs

In short, the cost of this unique blend of capacities, scaling and performances is a high complexity

Druid : Our Use Case

Druid hosts data related to the performances of videos from partners (meaning anyone with videos on **Dailymotion**) in terms of ads (Through various GO API which handle the requests)

- This data is pre-aggregated as is standard practice for OLAP (and especially crucial for Druid)
- The DataSource used as examples has 52 columns (Numerical formats, strings, no boolean though)
- Data freshness : No later than two hours

```
1  SELECT __time,account_username, video_title, visitor_embed_domain visitor_device_type, outcome
2  FROM ads_agg_v2
3  WHERE account_username = 'lachainelequipe'
4
5
```

Run ••• Auto limit Live query: Auto

__time	account_username	video_title	visitor_device_type	outcome
2022-02-01T00:00:00.000Z	lachainelequipe	Football - : Le replay de la 1/2 finale de Supercoup	www.dailymotion.com	noad
2022-02-01T00:00:00.000Z	lachainelequipe	La Petite Lucarne du 1er février - La petite lucarne	www.dailymotion.com	noad
2022-02-01T00:00:00.000Z	lachainelequipe	Le replay de l'individuel d'Antholz Anterselva - Bia	www.dailymotion.com	noad

Data Ingestion

Standard ingestion tasks

(Hourly and Daily and small backfills) take place through Airflow DAGs.

Heavy ingestion tasks

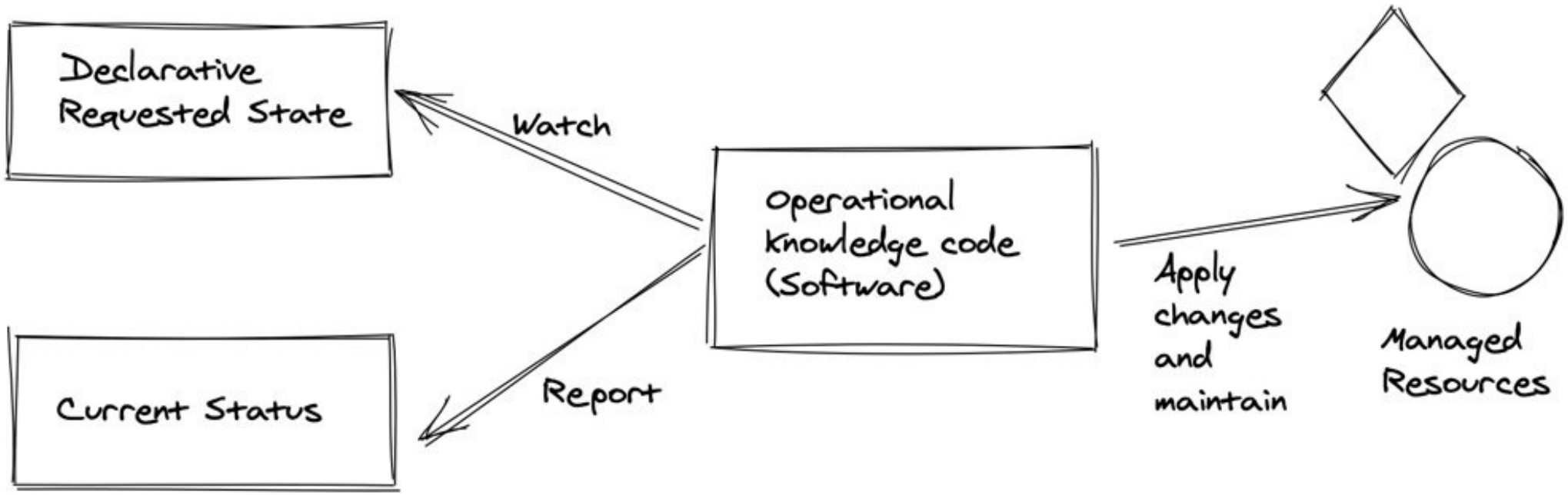
(New DataSources and backfills longer than a month) also uses airflow DAGs but we use a managed Hadoop Cluster to process the heavy amount of data.

In both cases we export Data from a Data Warehouse


KubeCon
CloudNativeCon
Europe 2023
Druid on Kubernetes

Custom resource & Operator Pattern

- Use **Declarative state** to define infrastructure.
- **Manage the lifecycle** of a services (upgrade / leader election, etc)
- **Transfer the knowledge** of maintaining the platform to the operator
- **Multiple frameworks** to facilitate development (kubebuilder)



Operators use cases

Flink

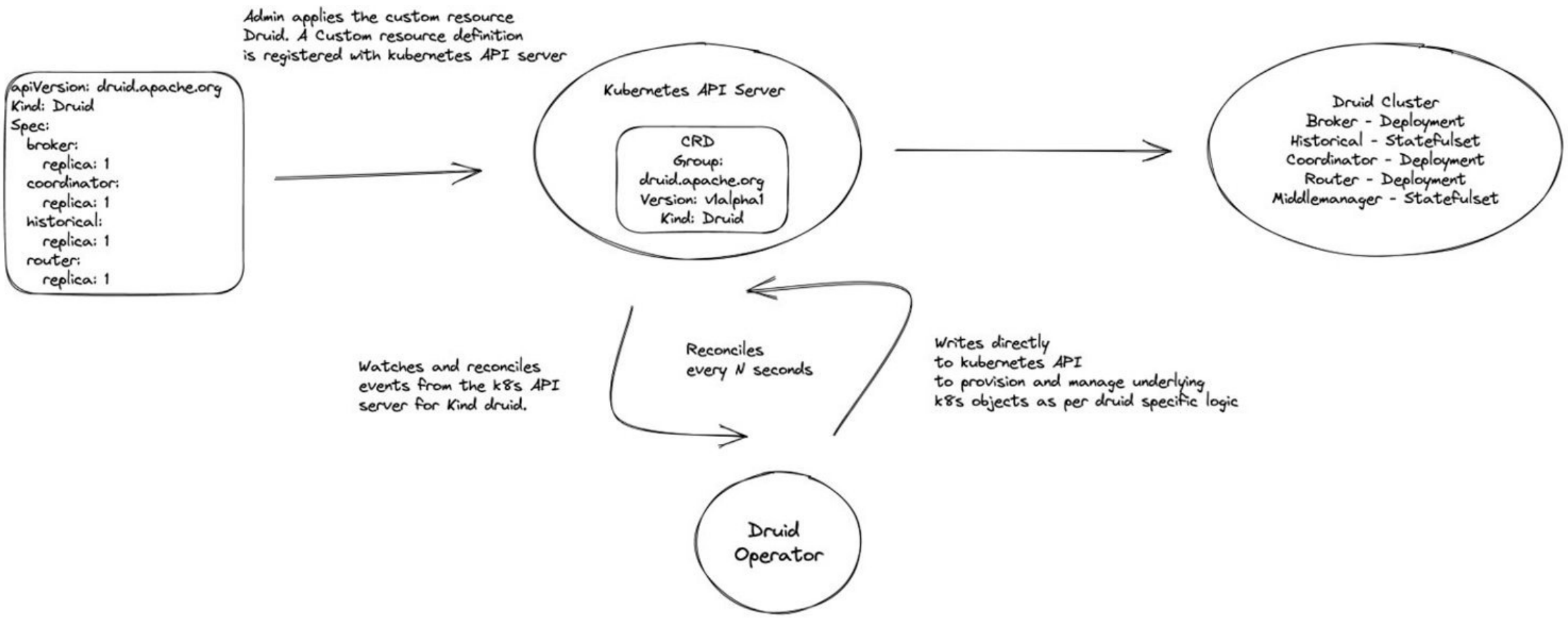
MINIO OPERATOR

Kubeflow

logging operator
BY BANZAICLOUD

- Database management
- Application deployment
- Monitoring and logging
- Machine learning workflows
- Infrastructure management

Druid Operator



Github: *https://github.com/datainfrahq/druid-operator*

Features provided by Druid Operator

- Rolling deploy
- Autoscaling on druid components
- Volumes Expansion
- Management of orphan PVC
- Sidecar injection
- Comprehensive setup management (probe / pdb …)
- Tiering management


KubeCon
CloudNativeCon
Europe 2023
Architecture
& Refactoring

Legacy

- Druid Cluster Step up in 2019 as an experiment at first
- Needs were unclear at the time, and had evolved a lot since
- Some technical choices made then no longer made sense (**TiDB**)

FluxCD to the rescue !

Pain points

- Maintenance (28 hours for a full deployment) due to relying on local SSD
- Set up no longer in phase with our needs
- Costs of scaling was high both in terms of cloud spendings and engineering time
- Several outages made hard to prevent because of the age of the set-up
- Deployment through mostly manual setup ; painful and error prone

Machines and Cluster Sizing

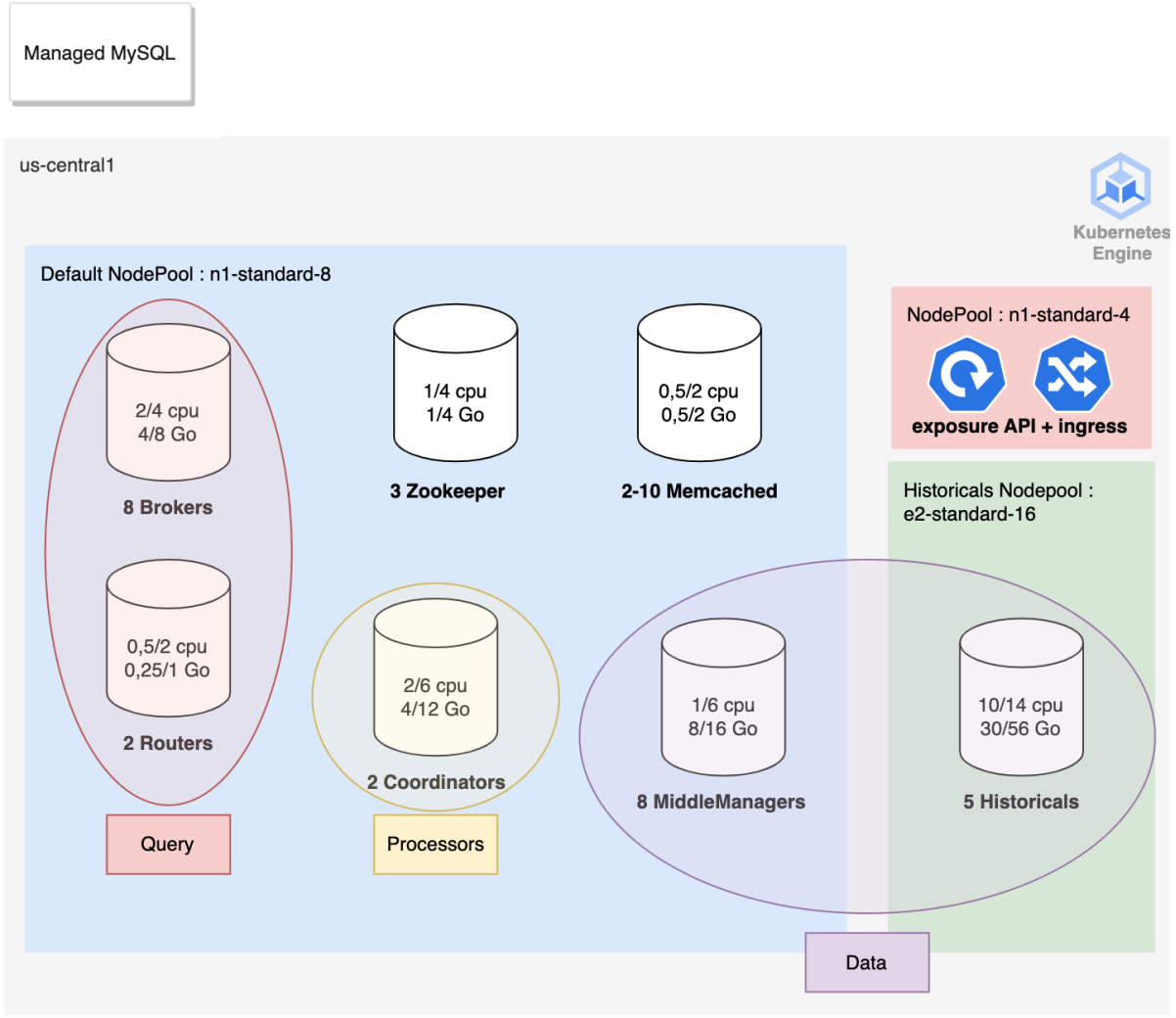
At Dailymotion we leverage Kubernetes on-prem and CSP, our Druid cluster is in the cloud

MACHINES FOR HISTORICALS

- Performances heavily RAM Bound
- We settled for 16vCPU / 64Go RAM
- Mix of spot nodes and standard ones

STORAGE CONFIGURATION

- 4*375Go / PVC per Historical
- 6 Historicals (Satisfying performances and enough for our current storage needs)

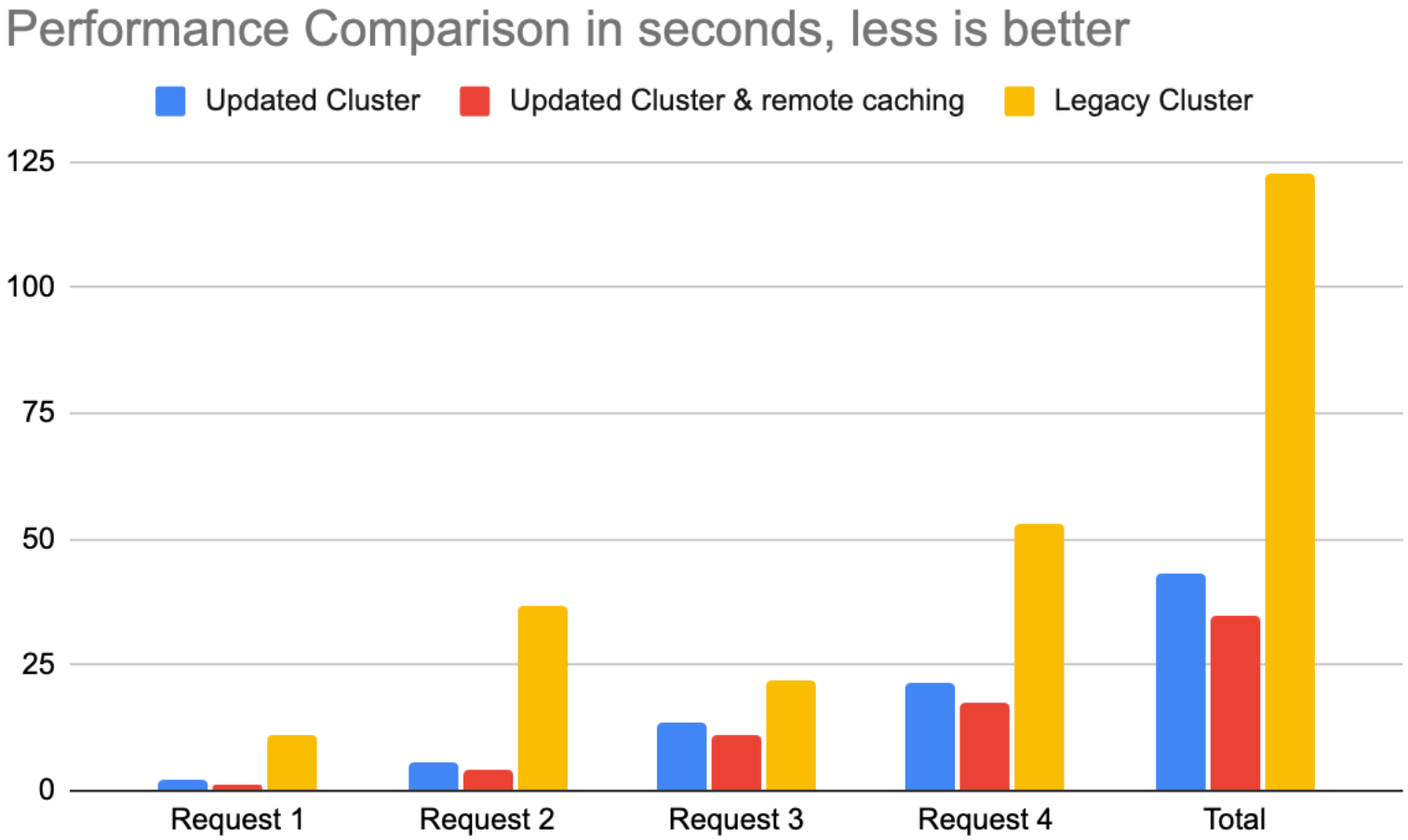


Caching and Performances

Our Legacy Druid cluster used almost no caching

We benchmarked local caching (local to the druid nodes, Druid uses Caffeine) and remote caching (Druid proposes a Memcached integration).

We also benchmarked various storage scenarios

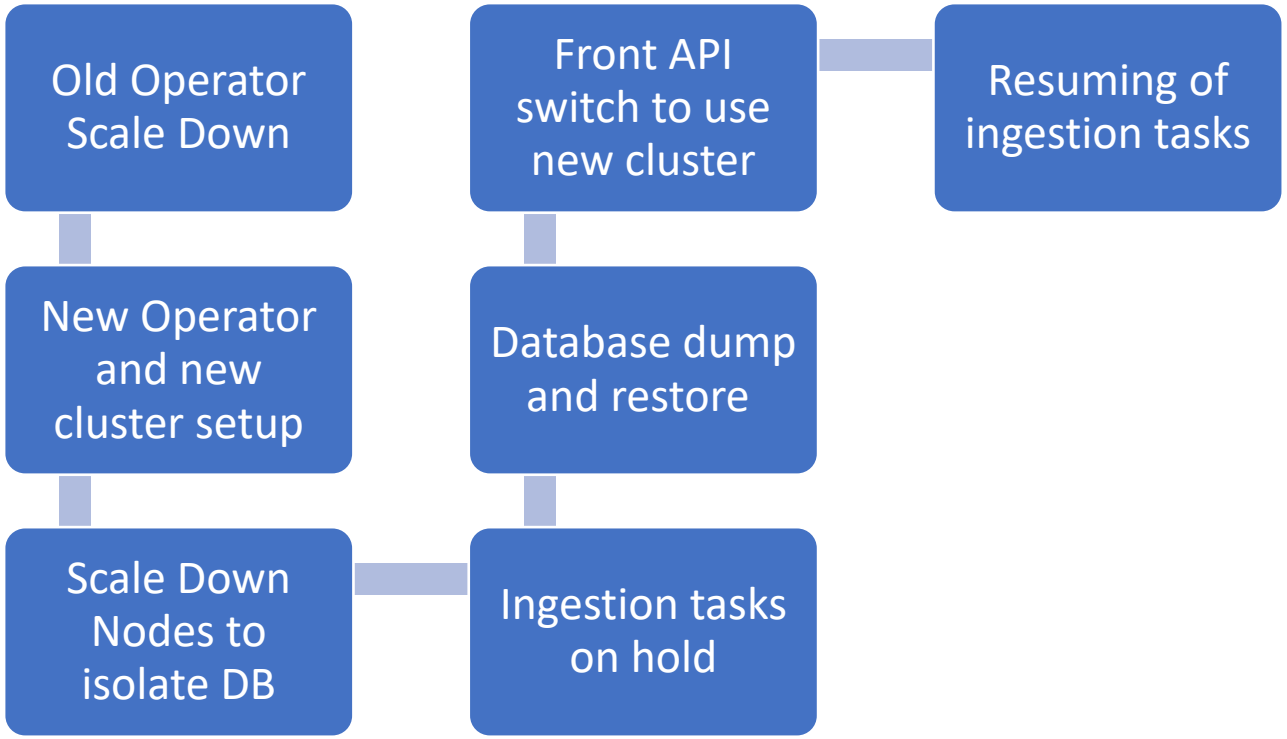


Druid Migration Process

DB Migration was necessary to avoid re-ingesting data
Druid Lacks tooling for prod migration

We created scripts to dump / modify / restore the DB and we set-up the 2 druids in a sort of double run

This process allowed us to migrate **13To** of data in **1 hour** without any downtime for end users



Monitoring and Exposition

Several plugins to monitor druid are available:

- Prometheus-adapters
- Statsd-adapters
- Openmetrics-adapters
- ...

In our case we are using statsd-adapters and we filter the metrics we need

Two ingresses are exposed:

- One created by the Operator for the console
- Another is for our API (protected by mTLS auth).

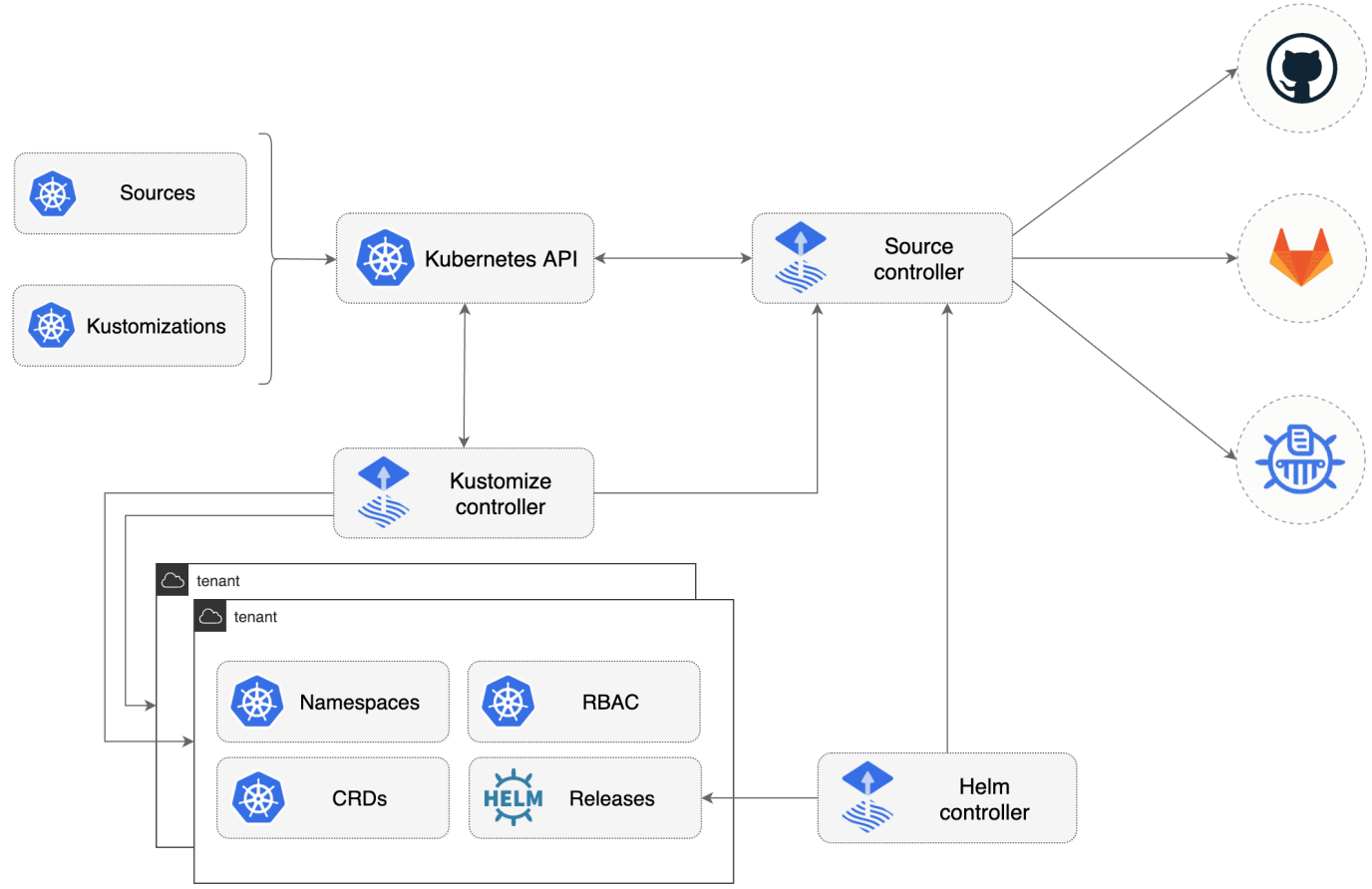
We use oauth2-proxy to authenticate the users on our console ingress.

Our certificates are issued by cert-manager


KubeCon
CloudNativeCon
Europe 2023
Druid
Gitopsed

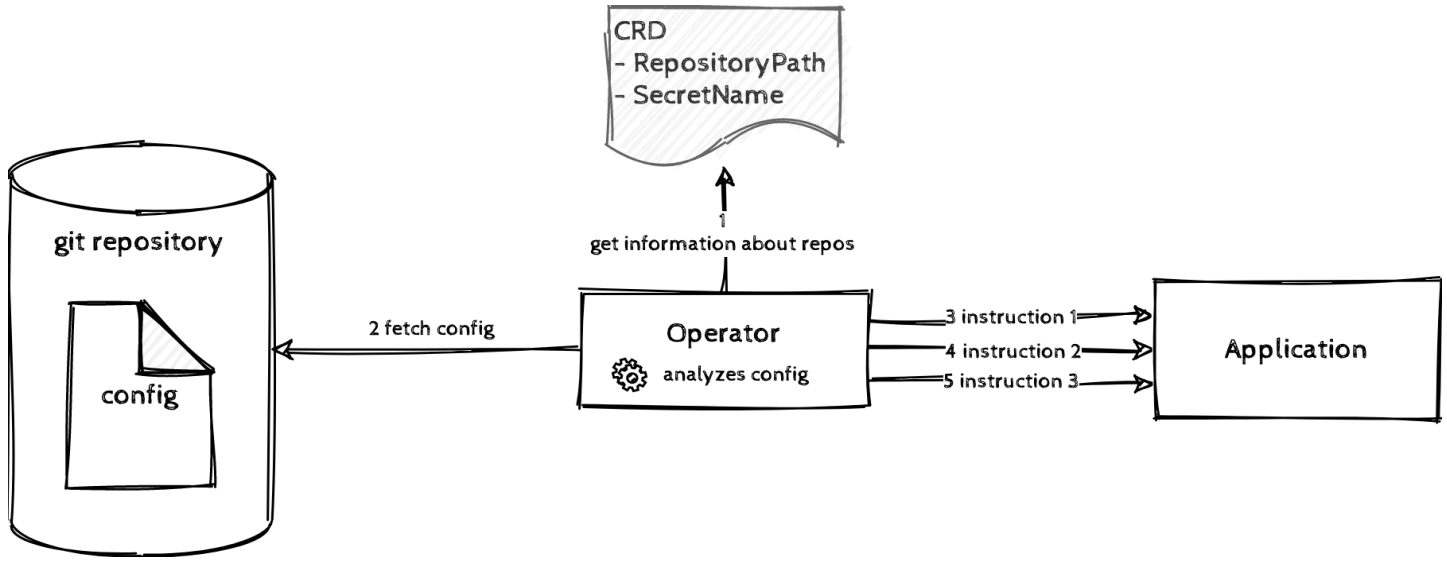
Gitops with FluxCD

GitOps is a way of managing your infrastructure and applications so that whole system is described declaratively, and version controlled (most likely in a Git repository) and having an automated process that ensures that the deployed environment matches the state specified in a repository.



(Druid) Operator and Gitops

- **Github** repository as source
- **Helmrelease** to deploy the operator
- **Kustomization** to deploy the druid cluster
- **Tiering kustomization** to setup zookeeper/Memcached before the druid cluster



Common Operations

AUTOMATIC ROLLOUTS

Having a proper set-up for rollouts makes operations like Kubernetes node upgrades 10 times easier and quicker. The nodes rollout in the following order :

STORAGE UPDATES

We had to migrate our storage classes recently and this proved to be both cumbersome and time consuming (took us about a day) mostly because of the way druid rebalances the data between the Historical nodes
This is the only problematic operation we encountered. Druid operator allows both easy horizontal scaling through the increase of the number of PVC and vertical scaling through volume expansion. (using cascade = orphan deletion strategy)


KubeCon
CloudNativeCon
Europe 2023
Feedback

Our Mistakes :)

PERFORMANCES : NATIVE QUERIES VS SQL

- Translation isn't perfect (improving version to version)
- Much less control on what gets executed
- Rewriting of everything from SQL to Native is in progress
- Tweaking of druid configuration allowed to mitigate those issues

DB MIGRATION AND DEEPSTORAGE

We overlooked how the location (bucket) of the segments in deep storage was specified in the database.

This ended with us having to do a second DB migration

A built-in DB migration tool would be a huge help

RAM CONSUMPTION MONITORING

- We hit some strange behaviors on RAM consumption, took us time to catch those
- Performance critical since historical are IO reliant
- Made more complex to investigate by using Java8 (then) on Kubernetes

Druid and Kubernetes : Next Steps

KubeCon CloudNativeCon Europe 2023

Short Term

- Migration to java 17
- Migration on ARM nodes
- Proxysql to access metadata database
- Zookeeper less thanks to druid new version

Long Term

- Ingestion triggered as Kubernetes Jobs
- Setup realtime ingestion
- Execute ingestion via ingestion controller (available in next druid-operator release)
- Multi-tenancy (several druid clusters in the same k8s)
- Geo-localisation of druid clusters (EMEA / NA / LATAM / APAC)
- Tiering based on data Source


KubeCon
CloudNativeCon
Europe 2023
Special thanks to the druid and druid-operator community ! :)


KubeCon
CloudNativeCon
Europe 2023
Any Kuestion ?

KubeCon CloudNativeCon
Europe 2023
North America 2021

Please scan the QR Code above
to leave feedback on this session