adform

HOW WE MANAGE OUR KAFKA CLUSTERS

Robert Fabisiak
Tomasz Gintowt

21.11.2019 Warszawa

FORWARD. TOGETHER.

Robert Fabisiak

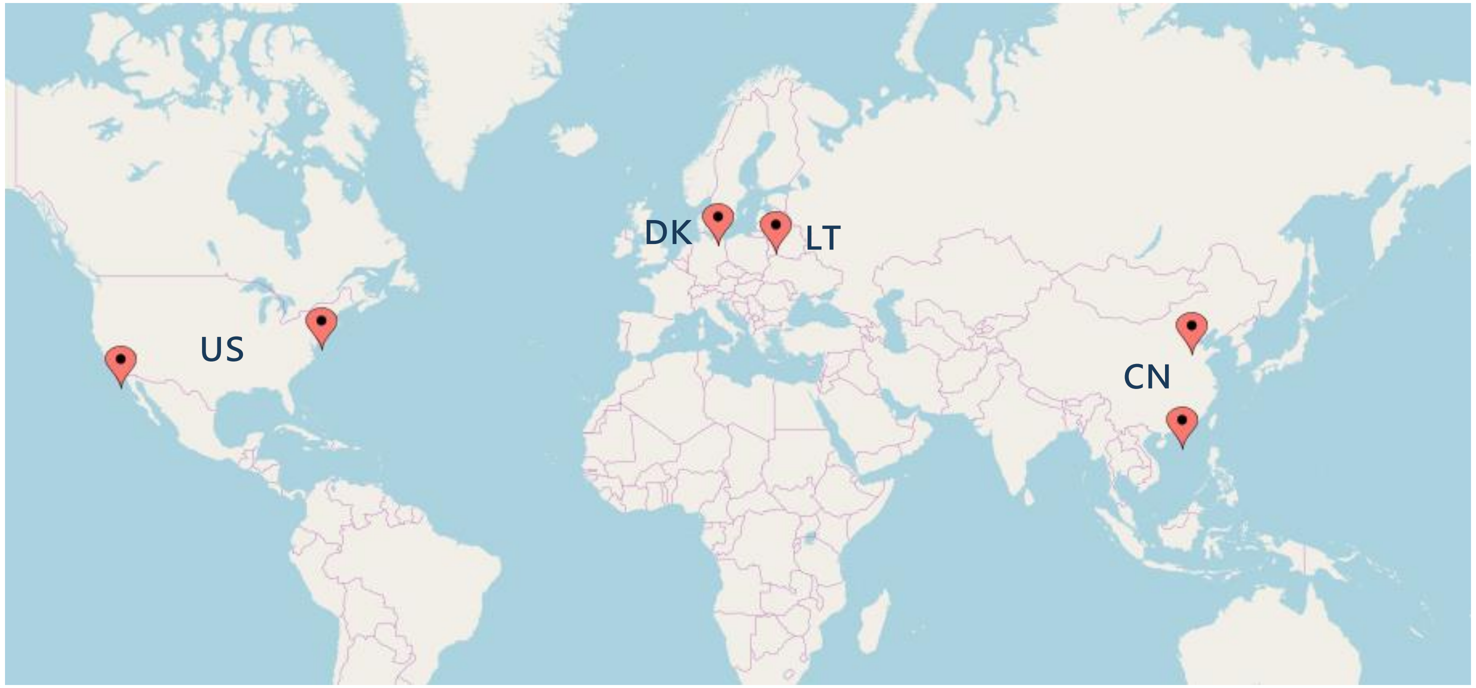
Tomasz Gintowt

Agenda

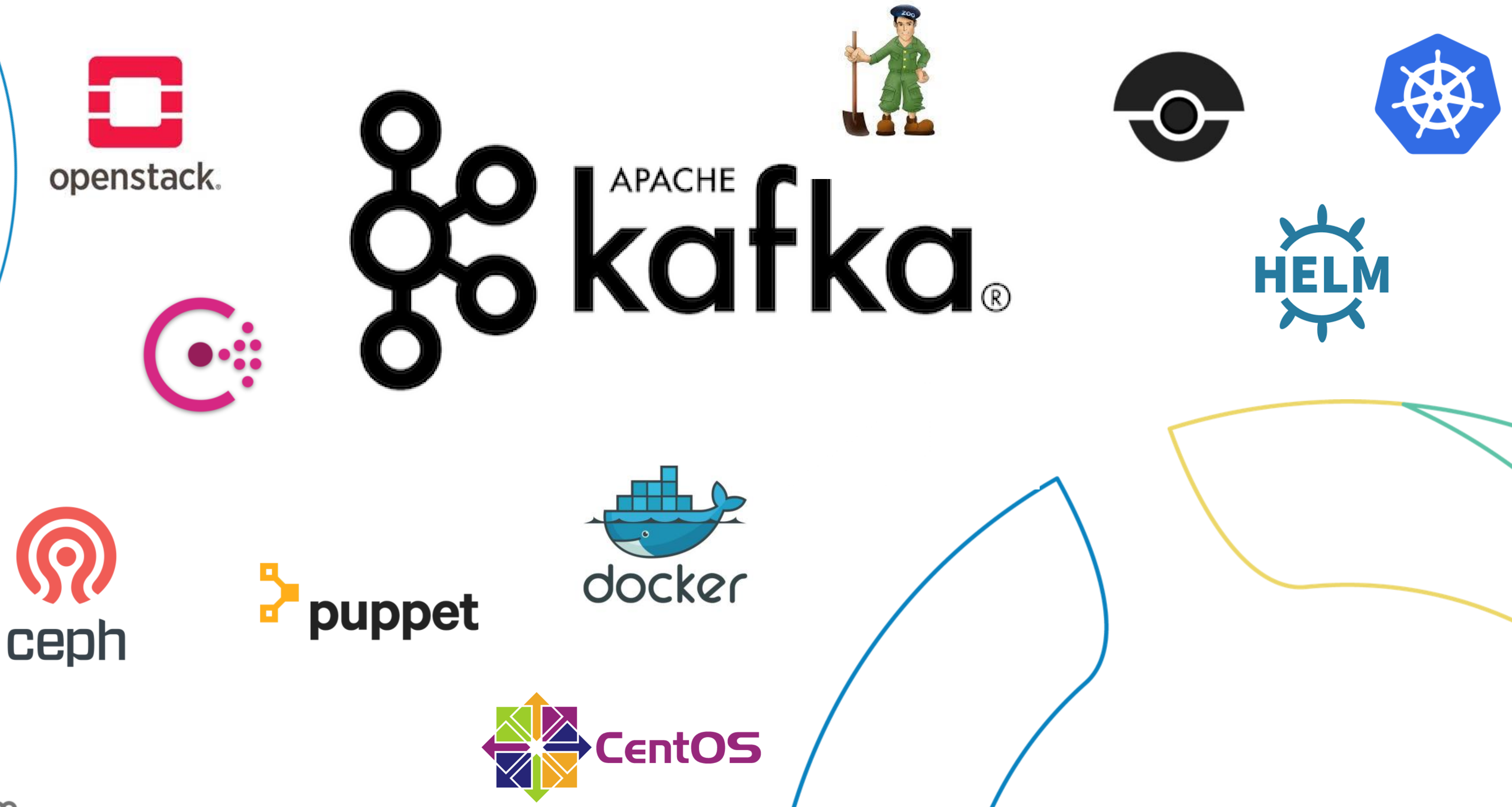
- Our clusters
- Manage clusters with Ansible
- Tools and Metrics
- Problems and Solutions

Kafka Clusters

- More than 100 servers (50% Bare Metal, 50% VM)
- Biggest cluster
  - 51 HW nodes
  - 460TiB storage
  - 16 TiB RAM
  - 2400 VCores
- **181 000 000 000 messages per day**
- Mirror data from remote clusters



Technologies



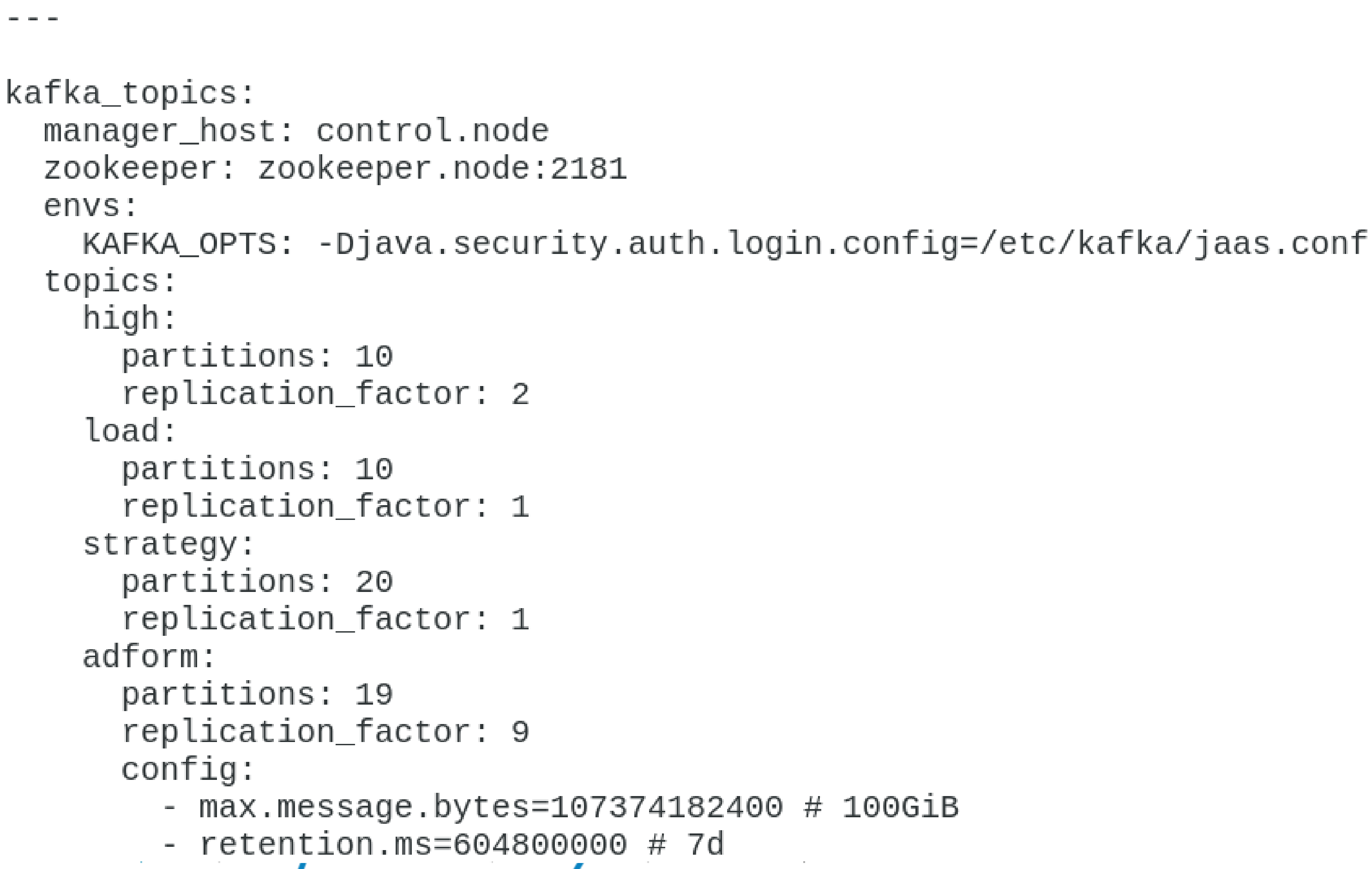
Automate upgrades

ANSIBLE

- **Follow official guidelines**
- Rolling upgrade with serial=1
  - stop and upgrade package
  - change configuration
  - start service
  - verify cluster health
- **Upgrade time reduced from weeks to days**

Automate topic management

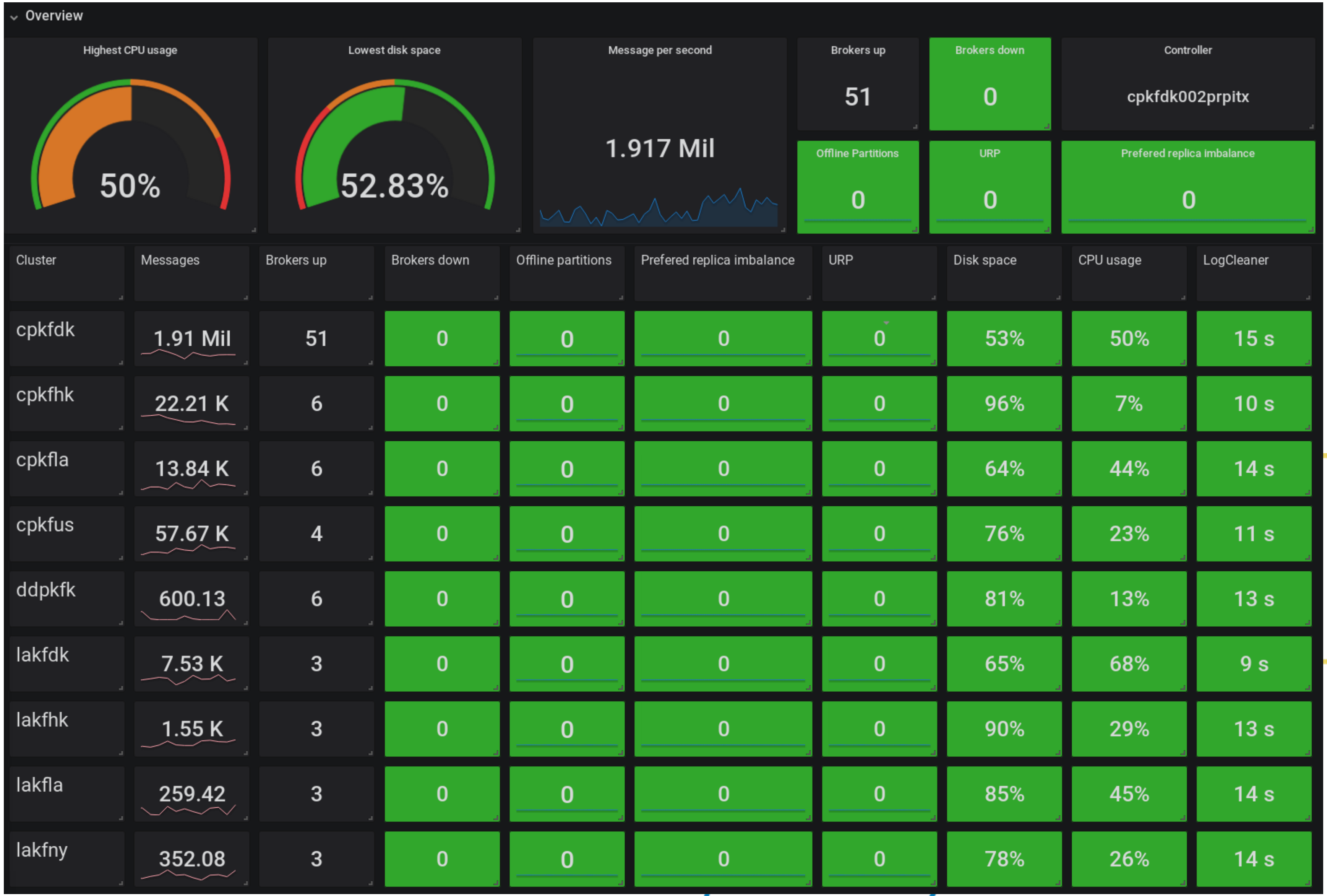
- deploy request
- CI build job
  - verify and validate YAML
  - check cluster utilization
  - check user quota
  - send notify



```
---

kafka_topics:
  manager_host: control.node
  zookeeper: zookeeper.node:2181
  envs:
    KAFKA_OPTS: -Djava.security.auth.login.config=/etc/kafka/jaas.conf
  topics:
    high:
      partitions: 10
      replication_factor: 2
    load:
      partitions: 10
      replication_factor: 1
    strategy:
      partitions: 20
      replication_factor: 1
    adform:
      partitions: 19
      replication_factor: 9
      config:
        - max.message.bytes=107374182400 # 100GiB
        - retention.ms=604800000 # 7d
```

Monitoring - Grafana



Overview

Highest CPU usage: 50%

Lowest disk space: 52.83%

Message per second: 1.917 Mil

Brokers up: 51

Brokers down: 0

Controller: cpkfdk002prpitx

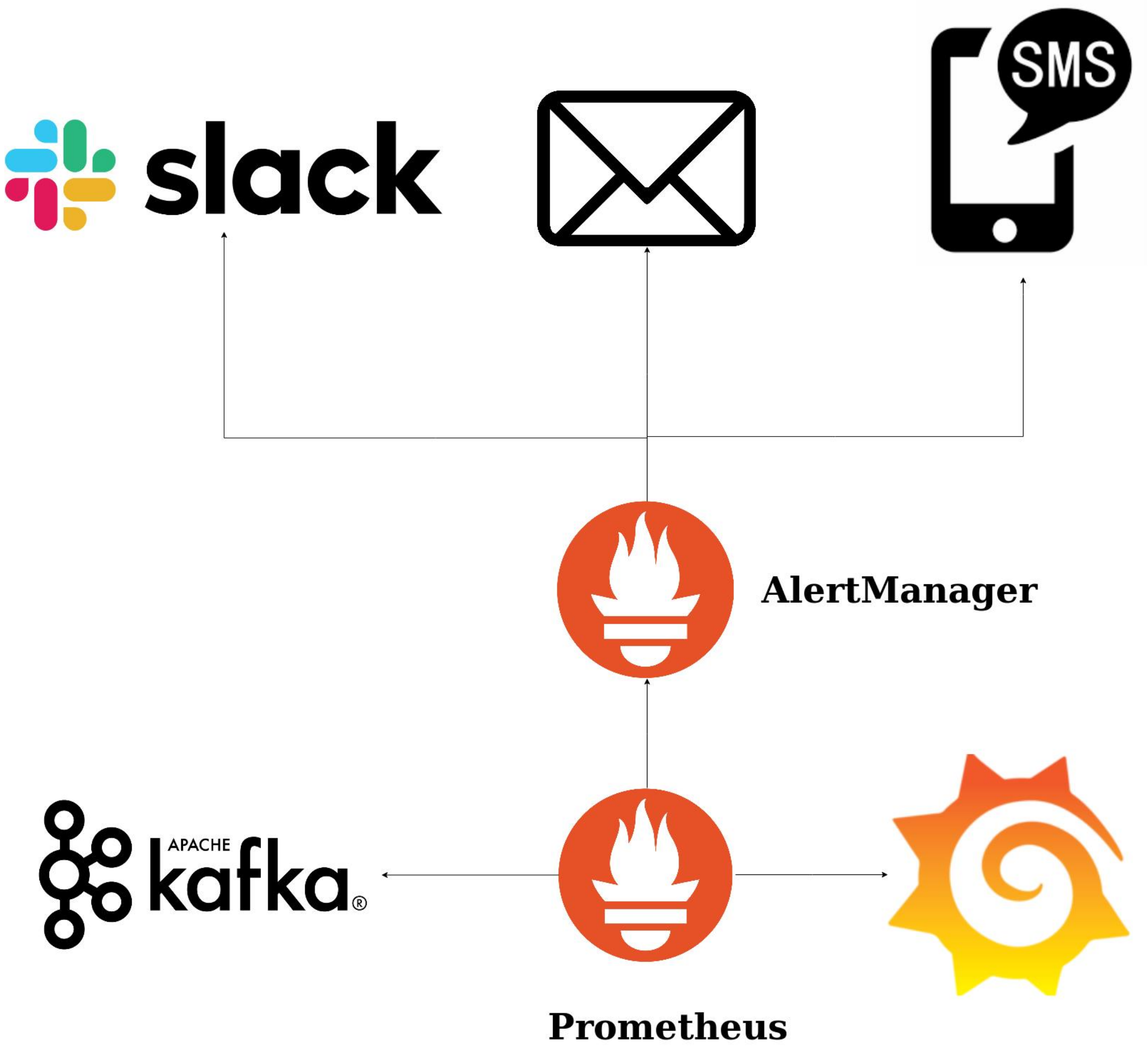
Offline Partitions: 0

URP: 0

Prefered replica imbalance: 0

Cluster	Messages	Brokers up	Brokers down	Offline partitions	Prefered replica imbalance	URP	Disk space	CPU usage	LogCleaner
cpkfdk	1.91 Mil	51	0	0	0	0	53%	50%	15 s
cpkfhk	22.21 K	6	0	0	0	0	96%	7%	10 s
cpkfla	13.84 K	6	0	0	0	0	64%	44%	14 s
cpkfus	57.67 K	4	0	0	0	0	76%	23%	11 s
ddpkfk	600.13	6	0	0	0	0	81%	13%	13 s
lakfdk	7.53 K	3	0	0	0	0	65%	68%	9 s
lakfhk	1.55 K	3	0	0	0	0	90%	29%	13 s
lakfla	259.42	3	0	0	0	0	85%	45%	14 s
lakfny	352.08	3	0	0	0	0	78%	26%	14 s

Monitoring - alerting



Monitoring - metrics

Kafka

- Messages in/out
- Offline partitions
- Broker up/down
- Partition Leader per broker
- Prefered replica imbalance
- URP
- LogCleaner
- Consumer Lag

JVM

- Heap Memory used
- GC Process time

Zookeeper

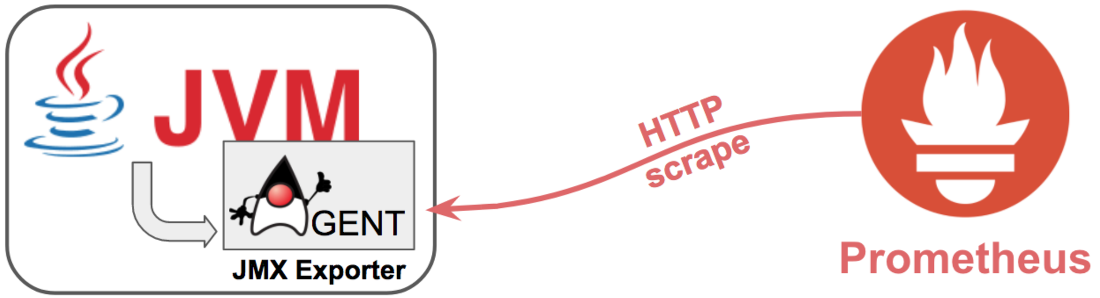
- Nodes
- Followers
- In-Sync Followers
- Pending Syncs

JMX exporter

- Installed on all nodes
- ~4k metrics per node
- Whitelisted MBeans objects:
  - Java
  - Cluster
  - Controller
  - Log
  - Network
  - Server

Kafka exporter

- Single exporter per cluster
- Most important metrics:
  - kafka_consumergroup_lag
  - kafka_topic_partition_leader
  - under_replicated_partition



Kafka-tools

Partition reassignment and preferred replica election.

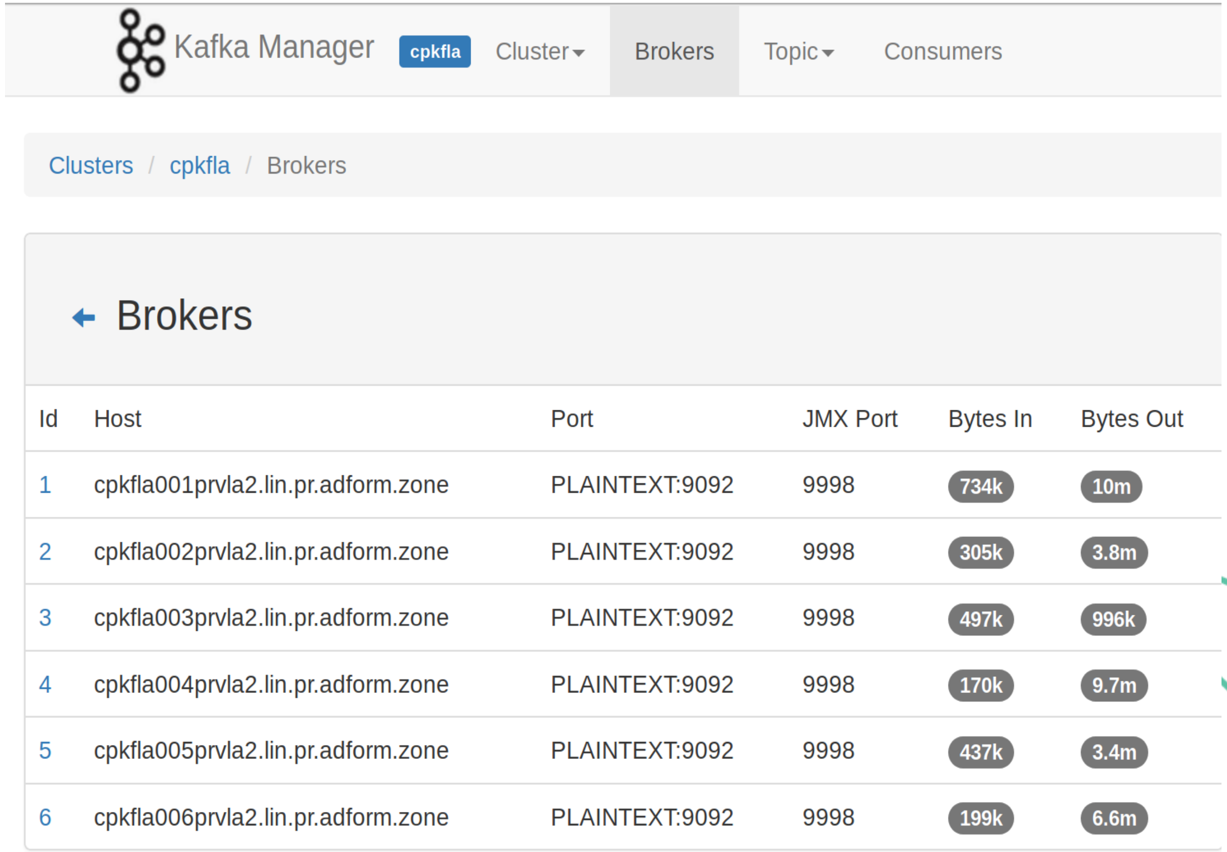
Modules:

- **Balance**
- Clone
- Elect
- Remove
- Trim

Type of balance:

- Count
- **Size**
- Even
- Leader

Kafka Manager



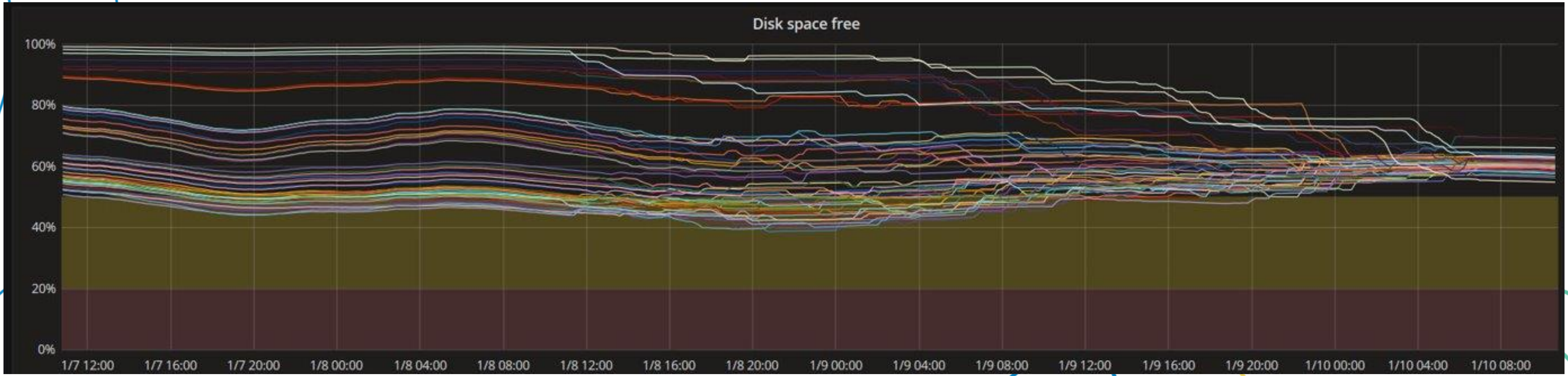
Clusters / cpkfla / Brokers

Brokers

Id	Host	Port	JMX Port	Bytes In	Bytes Out
1	cpkfla001prvla2.lin.pr.adform.zone	PLAINTEXT:9092	9998	734k	10m
2	cpkfla002prvla2.lin.pr.adform.zone	PLAINTEXT:9092	9998	305k	3.8m
3	cpkfla003prvla2.lin.pr.adform.zone	PLAINTEXT:9092	9998	497k	996k
4	cpkfla004prvla2.lin.pr.adform.zone	PLAINTEXT:9092	9998	170k	9.7m
5	cpkfla005prvla2.lin.pr.adform.zone	PLAINTEXT:9092	9998	437k	3.4m
6	cpkfla006prvla2.lin.pr.adform.zone	PLAINTEXT:9092	9998	199k	6.6m

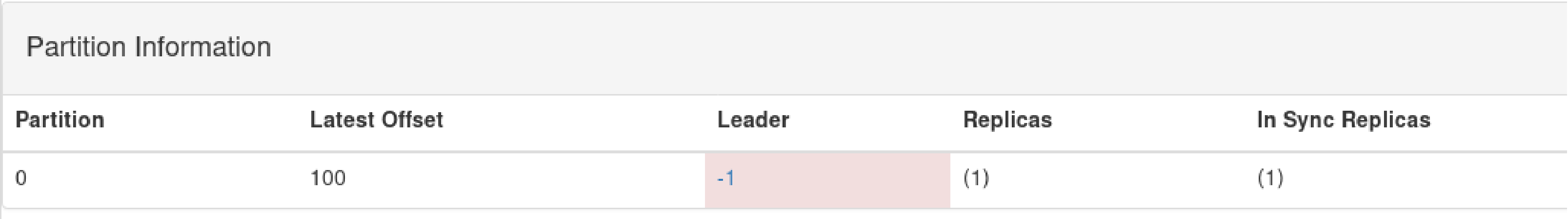
Problems and Solutions

Data rebalance results



Problems and Solutions

Data rebalance with offline partitions



Partition Information

Partition	Latest Offset	Leader	Replicas	In Sync Replicas
0	100	-1	(1)	(1)

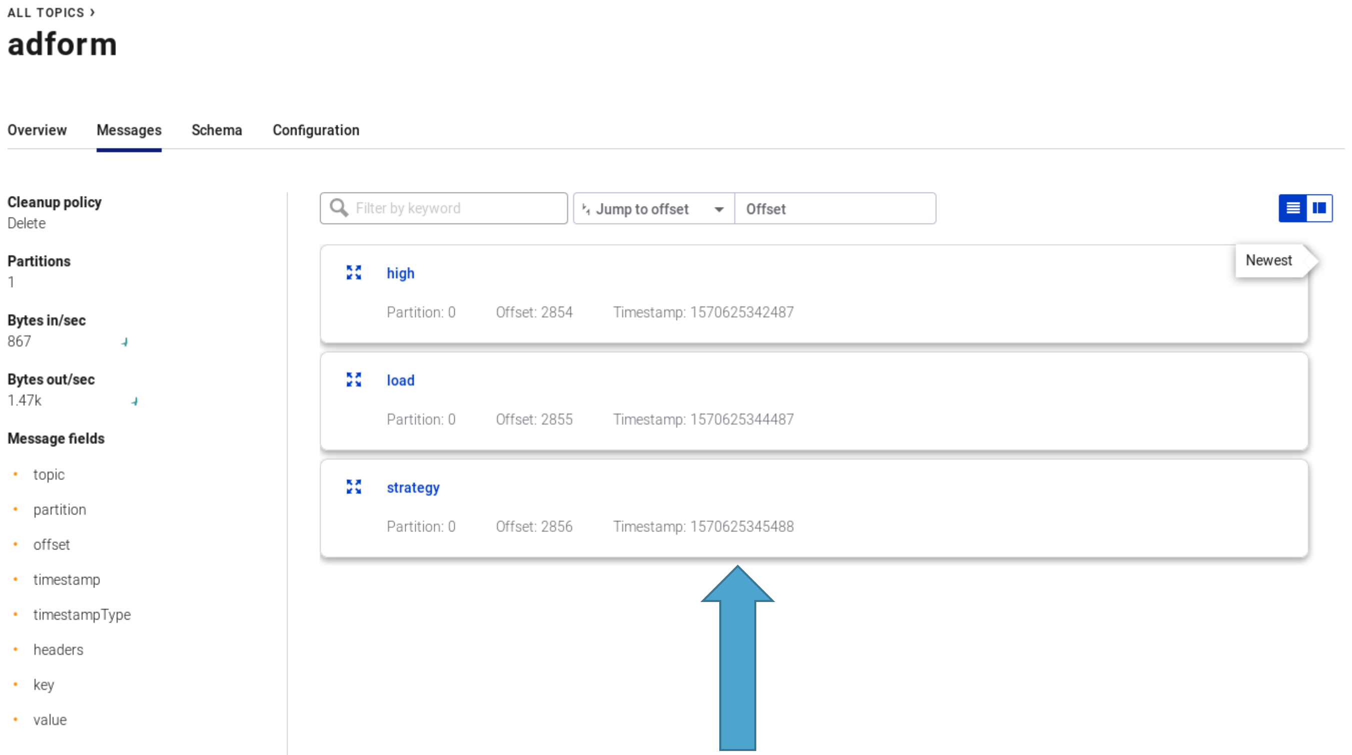
```
$ kafka-reassign-partitions --zookeeper zookeeper.service:2181/kafka --reassignment-json-file /tmp/a.json --verify
Status of partition reassignment:
Reassignment of partition high_load_strategy_adform-0 is still in progress
```

Solution

- reelect kafka controller with zookeper CLI
- always configure replicas for topics

Problems and Solutions

Wrong record timestamp



ALL TOPICS ›

adform

Overview | Messages | Schema | Configuration

Cleanup policy
Delete

Partitions
1

Bytes in/sec
867

Bytes out/sec
1.47k

Message fields

- topic
- partition
- offset
- timestamp
- timestampType
- headers
- key
- value

high
Partition: 0 Offset: 2854 Timestamp: 1570625342487

load
Partition: 0 Offset: 2855 Timestamp: 1570625344487

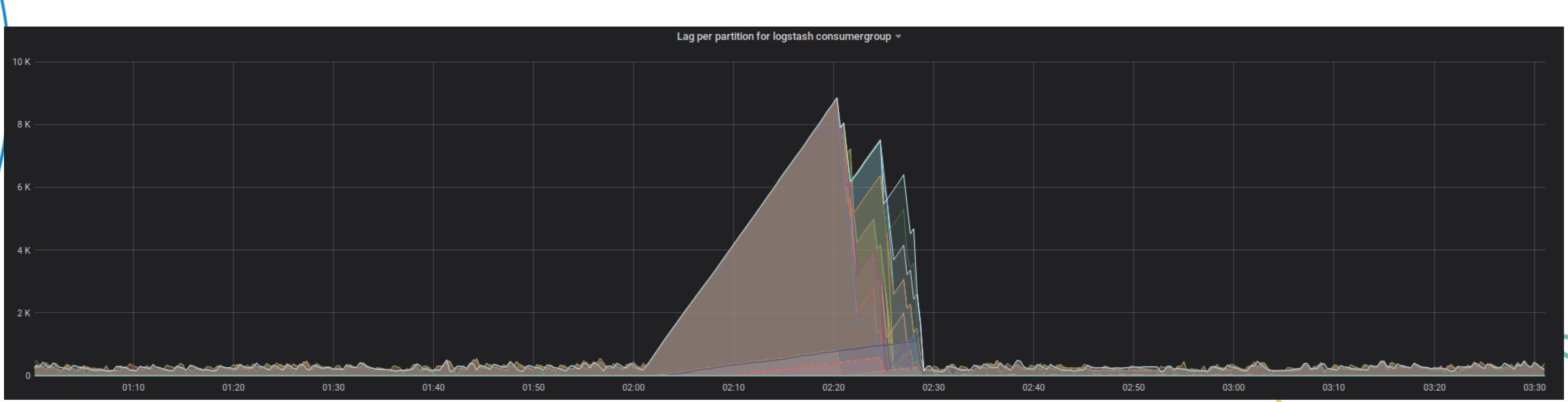
strategy
Partition: 0 Offset: 2856 Timestamp: 1570625345488

Solution

- Configure `log.message.timestamp.difference.max.ms`

Problems and Solutions

Consumer group constant rebalance

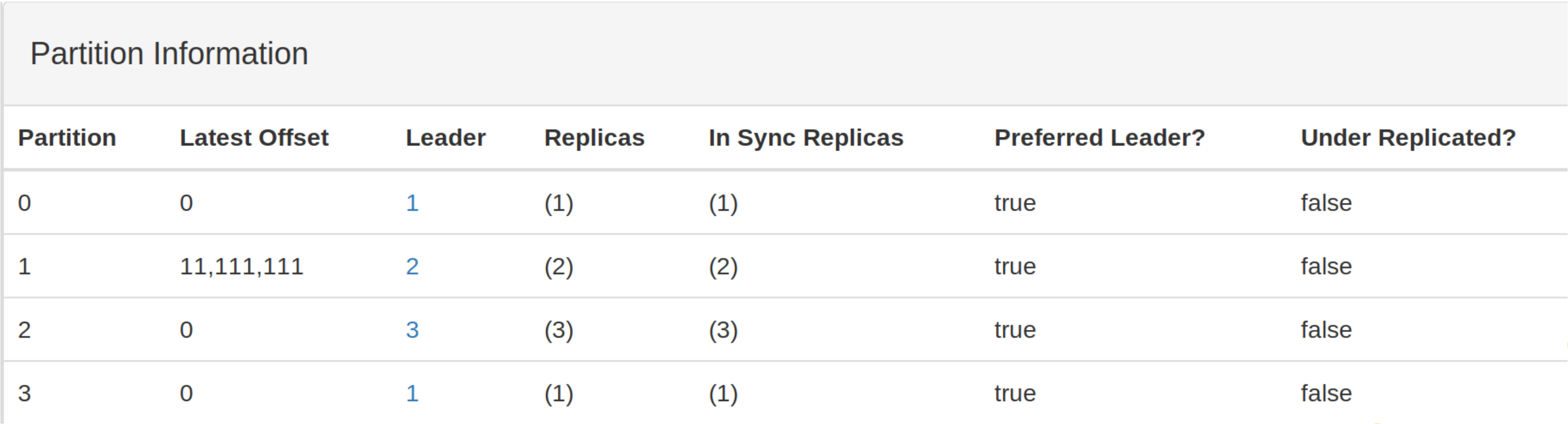


Solution

- Monitor consumer performance
- session.timeout.ms
- max.poll.interval.ms
- heartbeat.interval.ms

Problems and Solutions

NULL key is not null



Partition Information

Partition	Latest Offset	Leader	Replicas	In Sync Replicas	Preferred Leader?	Under Replicated?
0	0	1	(1)	(1)	true	false
1	11,111,111	2	(2)	(2)	true	false
2	0	3	(3)	(3)	true	false
3	0	1	(1)	(1)	true	false

Solution:

- `Not always NULL mean null, know your libs.`

adform

THANK YOU!

FORWARD. TOGETHER.