Keep Kafka Reliable

FREEWHEEL
A COMCAST COMPANY

FEBRUARY 29, 2024

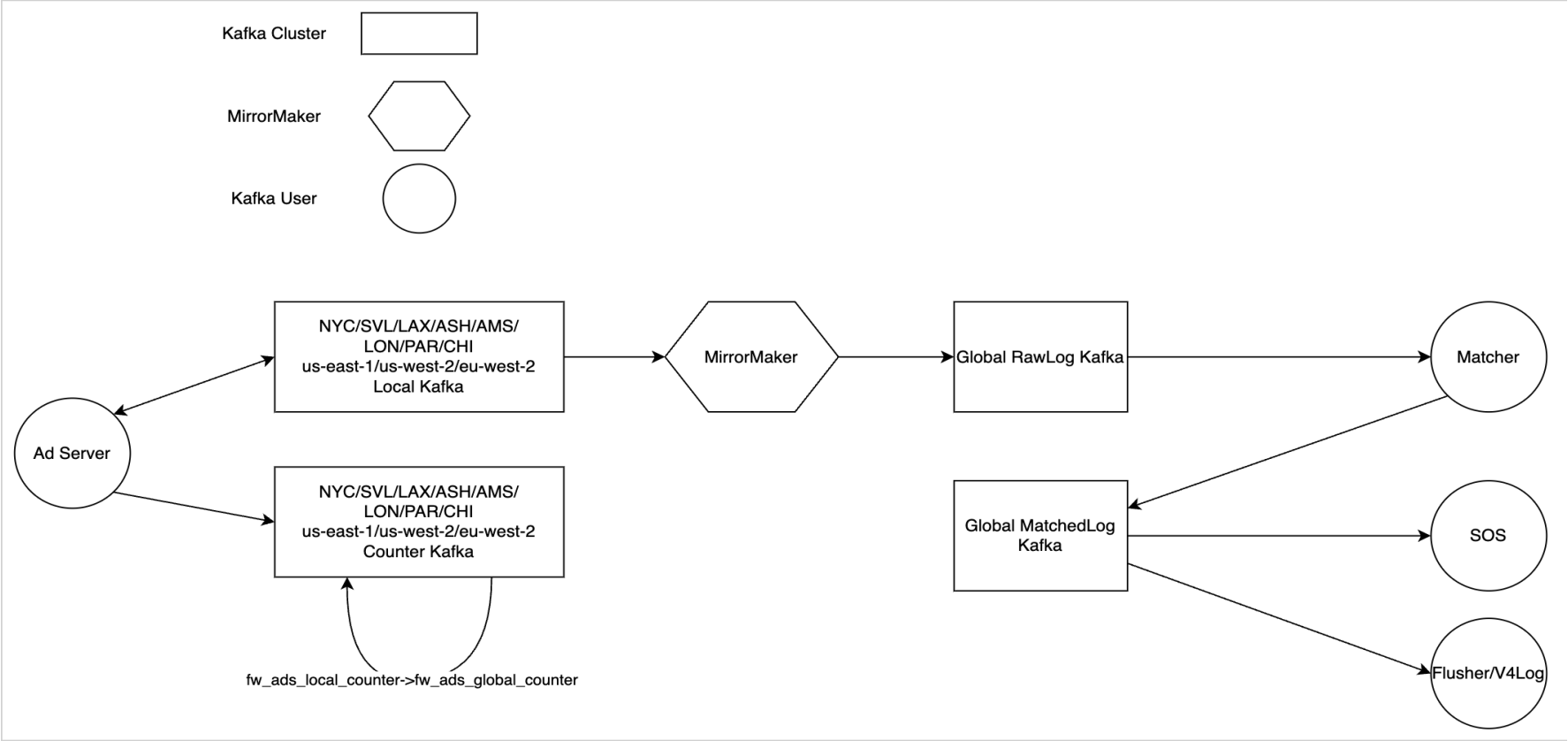
TABLE OF CONTENTS

- Introduction
- ZSTD Cutover
- Connect Cluster Upgrade & Tuning
- Kafka Partition
- Usage of AI

© 2024 FreeWheel. Confidential.

Data Flow

- **Realtime Data Flow**

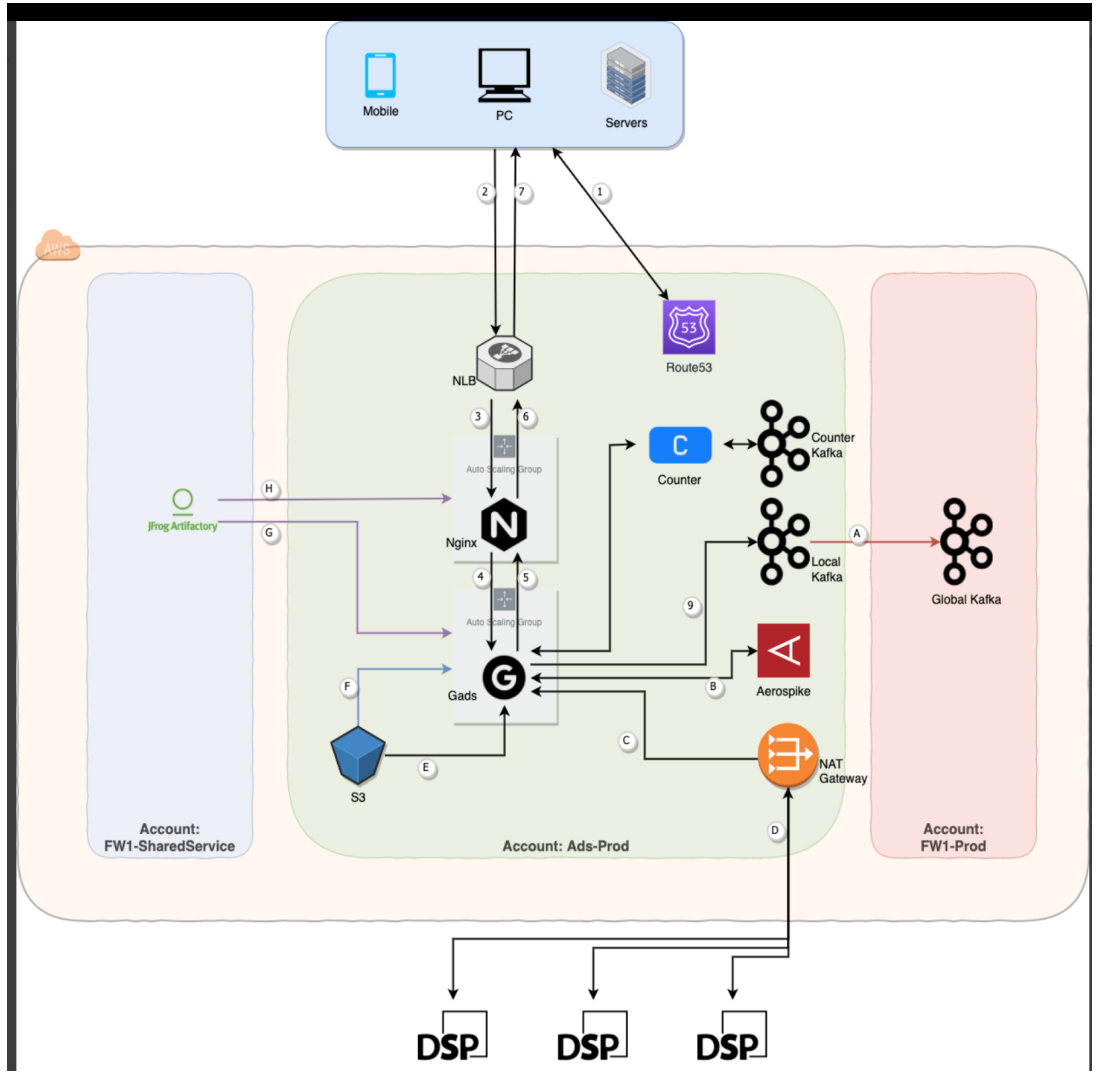


© 2024 FreeWheel. Confidential.

FREEWHEEL

Data Flow

- **ADS Data Flow**

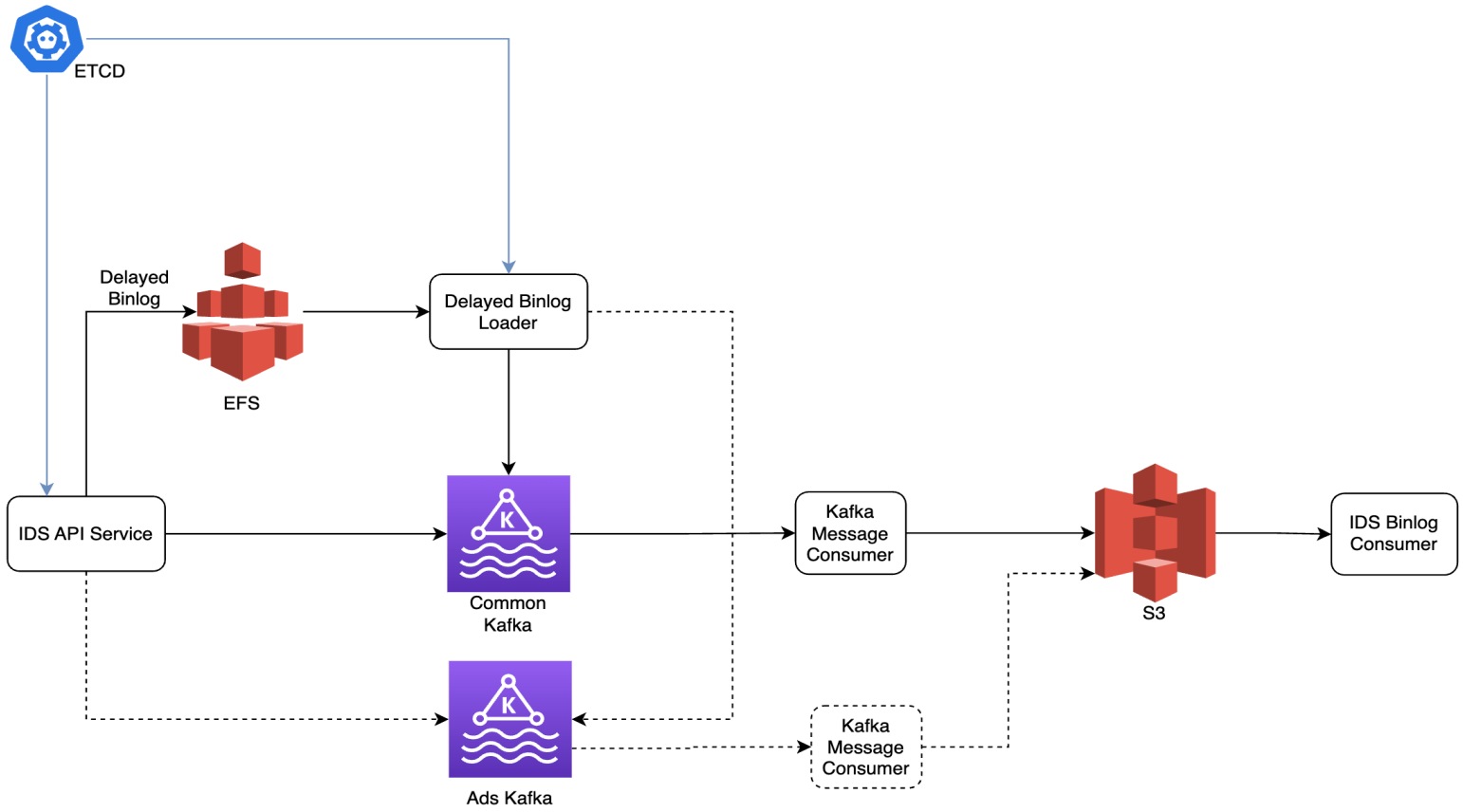


© 2024 FreeWheel. Confidential.

FREEWHEEL

Data Flow

- **AIM Team Data Flow**



© 2024 FreeWheel. Confidential.

FREEWHEEL

Apply ZSTD compression on ADS Topics

FREEWHEEL
A COMCAST COMPANY

FEBRUARY 29, 2024

1.Motivation

The disk space requirements for storing Kafka messages take into account these factors:

- Average message size
- Messages per day
- Retention period (days)
- Replication factor

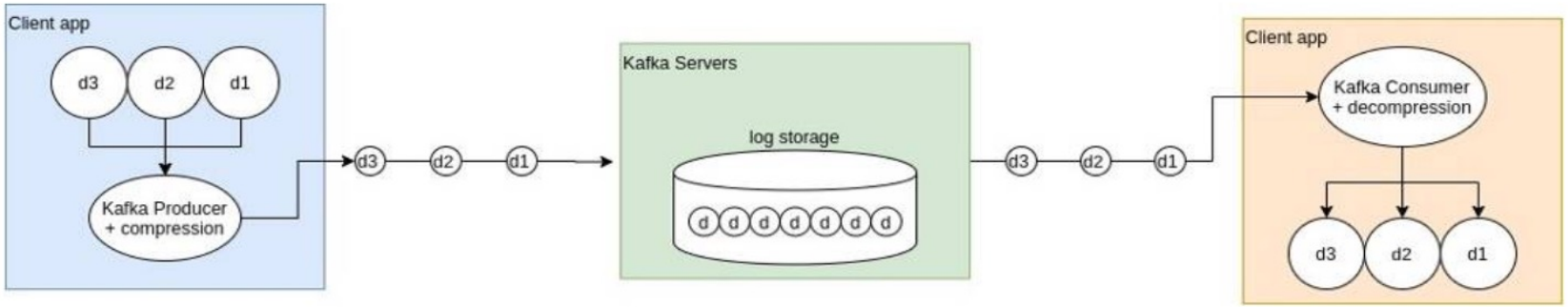
You can compute your disk space requirement as:
(avg-msg-size) x (msgs-per-day) x (retention-period-days) x (replication-factor)

Example:
Average message size is 10kb
Messages per day is 1,000,000
Retention period is 5 days
Replication factor is 3
Using our disk space utilization formula:
10 x 1000000 x 5 x 3 = 150,000,000 kb = 146484 MB = 143 GB

In that , we need to implement some compression on the data, so that you minimize the disk space utilization

© 2024 FreeWheel. Confidential.

2.Architecture about Message Compression in Kafka



Eng Team:

Take sarama library as an example:

1. adopt zstd compression in producer
   a. need set compression.type = zstd
   b. need set clientVersion = V_2_1_0
2. adopt zstd compression in consumer
   a. need set clientVersion = V_2_1_0

SRE Team:

Enable the topics with zstd compression type on Kafka brokers for six local clusters.

use1-local euw2-local

use2-local euc1-local

usw1-local

usw2-local

© 2024 FreeWheel. Confidential.

FREEWHEEL

3. Cooperate with ENG Team

- **Notify the topic owner(most are ads team in this case)**
- **Change the compression config for each topic on each Kafka topics one by one.**

```
use1-local
for i in `cat adslist`; do ./kafka-configs.sh --zookeeper zookeeper5.us-east-1.prd.ads.aws.fwmrm.net:2181/kafka --entity-type
topics --entity-name $i --alter --add-config compression.type=zstd; done
```

- **Inform the in the slack channel for transparency**

Monday, July 24th

Monday, July 31st

Lihua Wu 9:35 AM

按计划 kafka topics zstd cutover plan 今天准备将 `AWS-VIR` local kafka cluster的以下topic msg压缩格式改为zstd。预计观察一天没问题后，再将以下topic所有local kafka cluster的topic msg压缩格式改为zstd

@hwcao @dsli @Geng Jinhe 请知晓。FYI @Fei Li @xqzhang @Nick Yang @Chenlei Zhao @Fachang Yang

```
fw_ads_linear_bind_info
mrm-linear-p1-psns-384777
mrm-linear-p1-psns-384776
mrm-linear-p1-psns-505334
mrm-linear-p1-psns-505333
mrm-linear-p1-decisions
mrm-linear-p1-errors
mrm-linear-p1-parsing
mrm-linear-p1-unknown-devices
```

2

23 replies Last reply 1 month ago

Xiaomei Liu 3:48 PM

Message #zstd-exploring

Lihua Wu 1 month ago

https://wiki.freewheel.tv/display/viqa/Post+check+wiki%3A+zstd+compression+in+ads+local+kafka+cutover+in+IT4 @dsli

1 1

Chenlei Zhao 1 month ago
mrm-linear-p1-psns-384777 changed to zstd compression type at 7:15 UTC time on Cluster use1

Chenlei Zhao 1 month ago
mrm-linear-p1-psns-384776 changed to zstd compression type at 7:21 UTC time on Cluster use1

Chenlei Zhao 1 month ago
mrm-linear-p1-psns-505334 changed to zstd compression type at 7:22 UTC time on Cluster use1

Chenlei Zhao 1 month ago
mrm-linear-p1-psns-505333 changed to zstd compression type at 7:28 UTC time on Cluster use1

Chenlei Zhao 1 month ago
mrm-linear-p1-decisions changed to zstd compression type at 7:30 UTC time on Cluster use1

Chenlei Zhao 1 month ago
mrm-linear-p1-errors changed to zstd compression type at 7:35 UTC time on Cluster use1

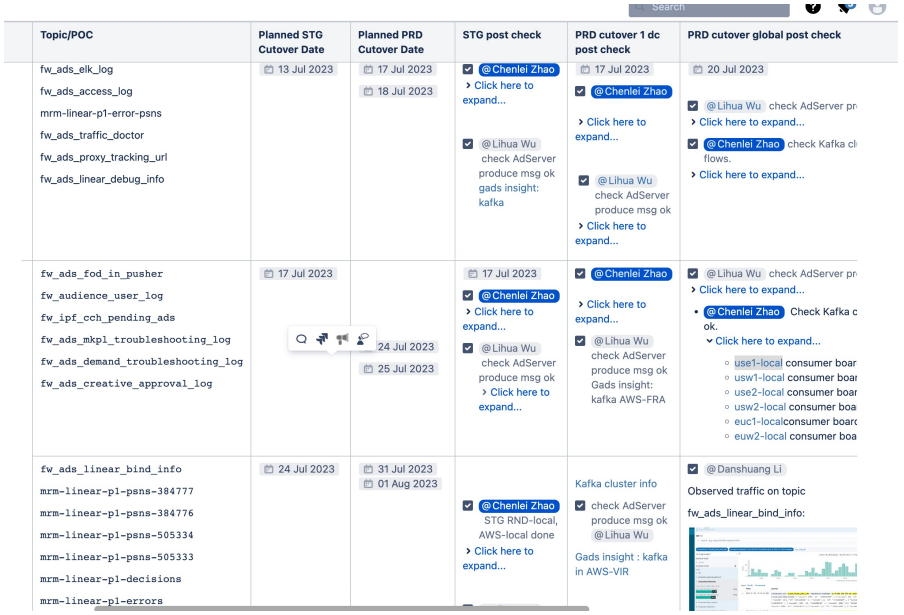
Chenlei Zhao 1 month ago
mrm-linear-p1-parsing changed to zstd compression type at 7:40 UTC time on Cluster use1

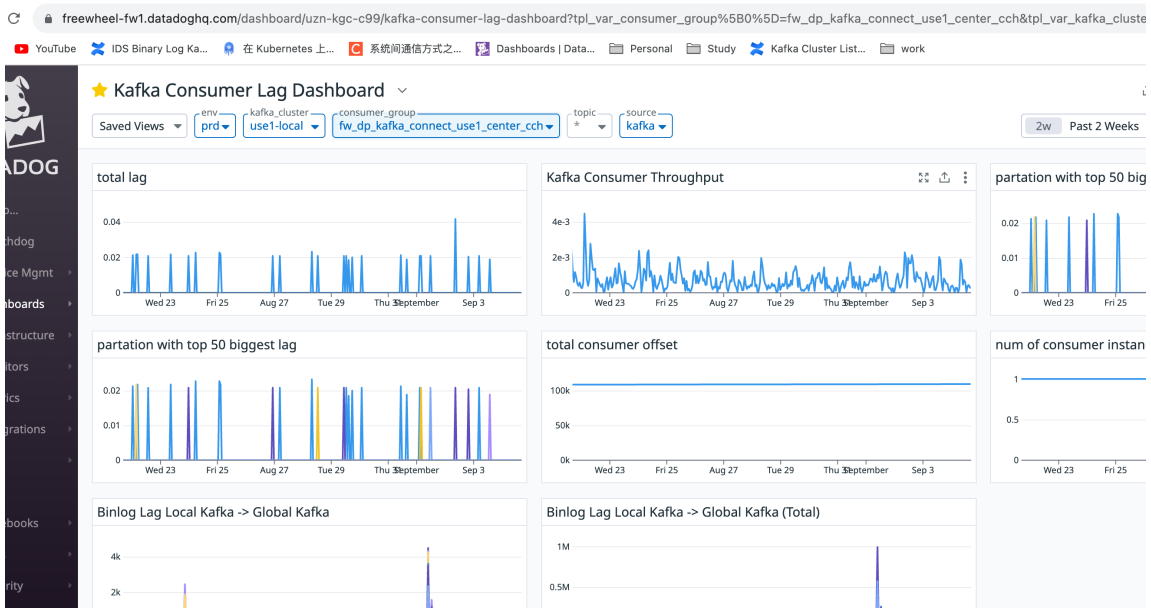
© 2024 FreeWheel. Confidential.

FREEWHEEL

Cooperate with ENG Team

- **Complete the post-check wiki with monitor status of each topic**
- **[post check wiki]**



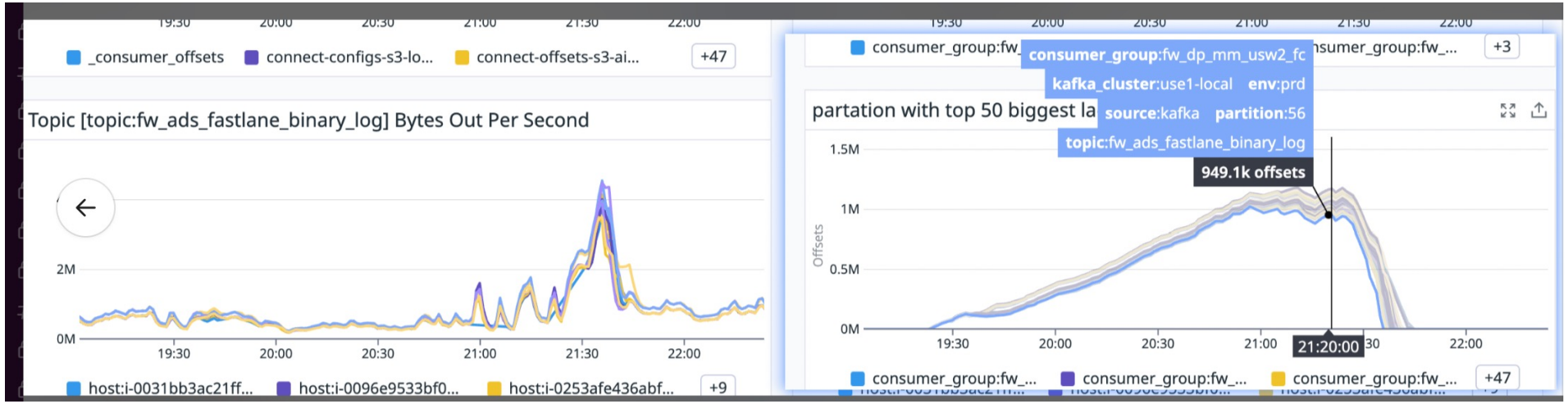


© 2024 FreeWheel. Confidential.

FREEWHEEL

4. Issue happens

- **There is OOM error of mirror maker which cause the lag for consumer fw_dp_mm_usw2_fc**



- **Rolled back the operations and increased the memory to 512M of the MM pod then restarted the consumer.**

© 2024 FreeWheel. Confidential.

FREEWHEEL

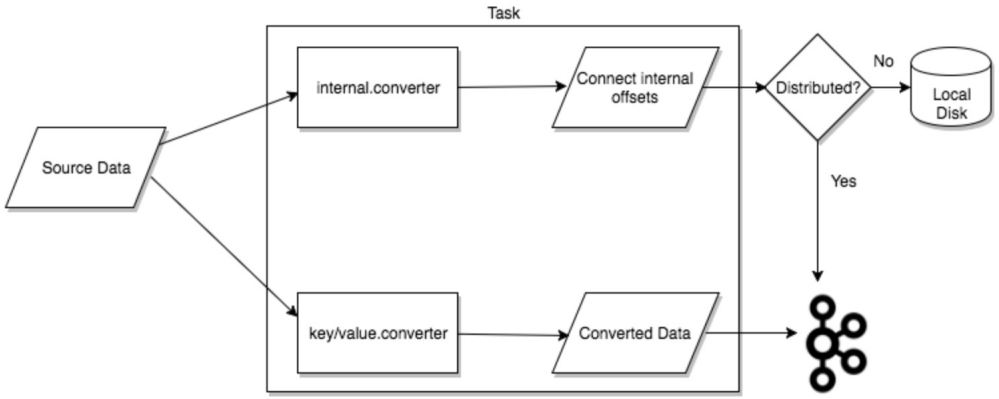
Kafka Connector

FREEWHEEL
A COMCAST COMPANY

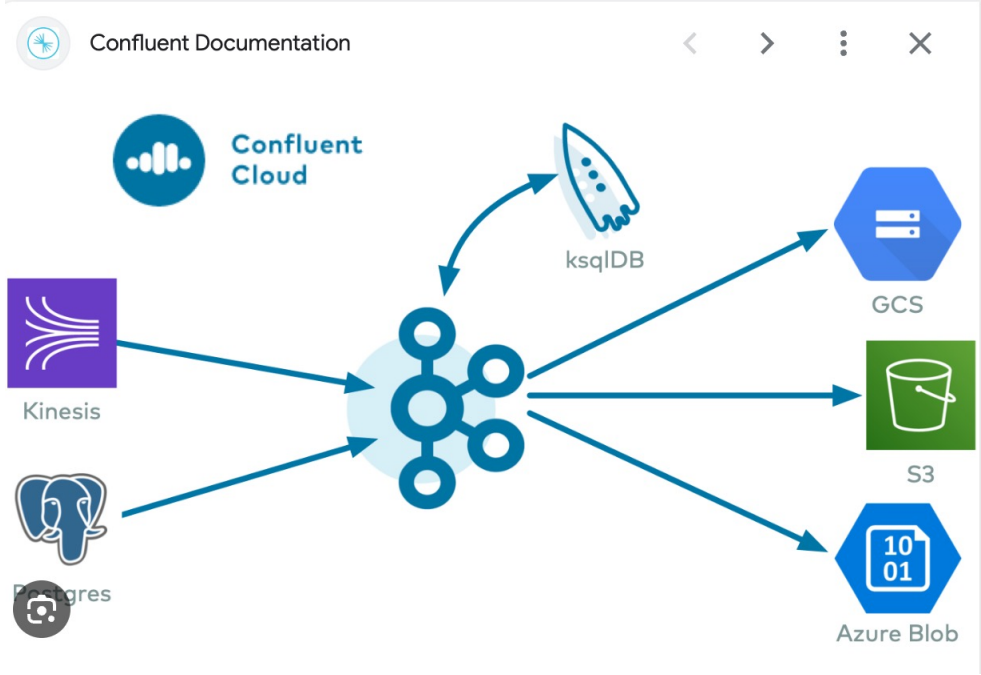
FEBRUARY 29, 2024

1. Background

Source Connector



Sink Connector

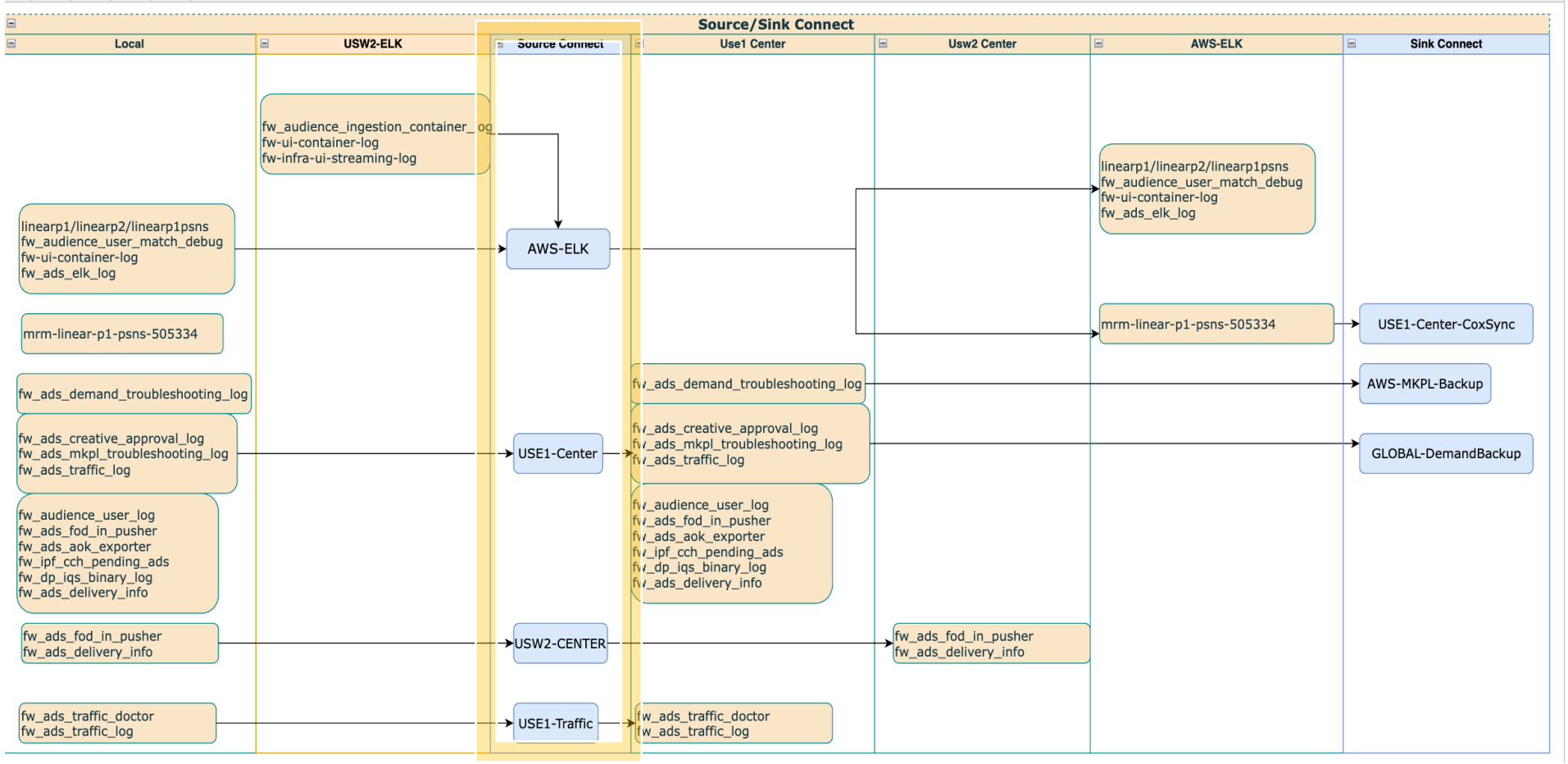


© 2024 FreeWheel. Confidential.

FREEWHEEL

2. Source Connectors

- **Source Connectors**

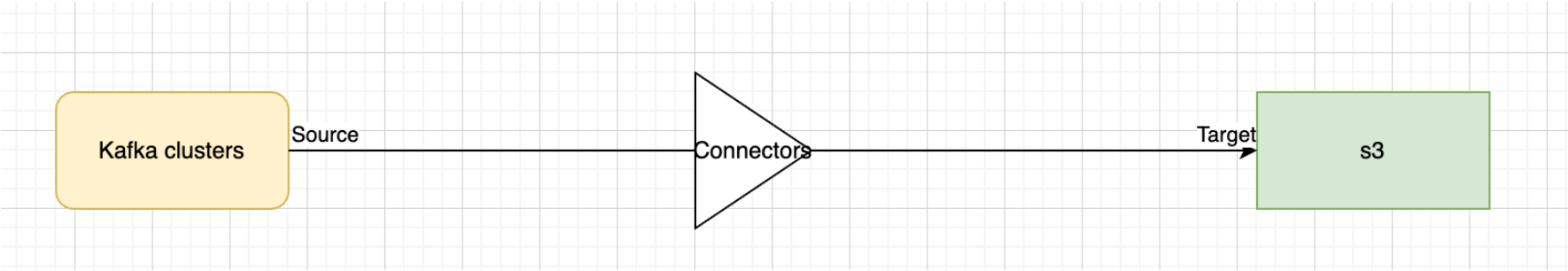


© 2024 FreeWheel. Confidential.

FREEWHEEL

3. Sink Connectors

- **Sink Connector**



Topic list

- fw_ads_binary_log
- fw_ads_rtb_auction_log
- fw_ads_large_msg
- fw_ads_demand_troubleshooting_log
- fw_ads_mkpl_troubleshooting_log
- Etc…

S3 list

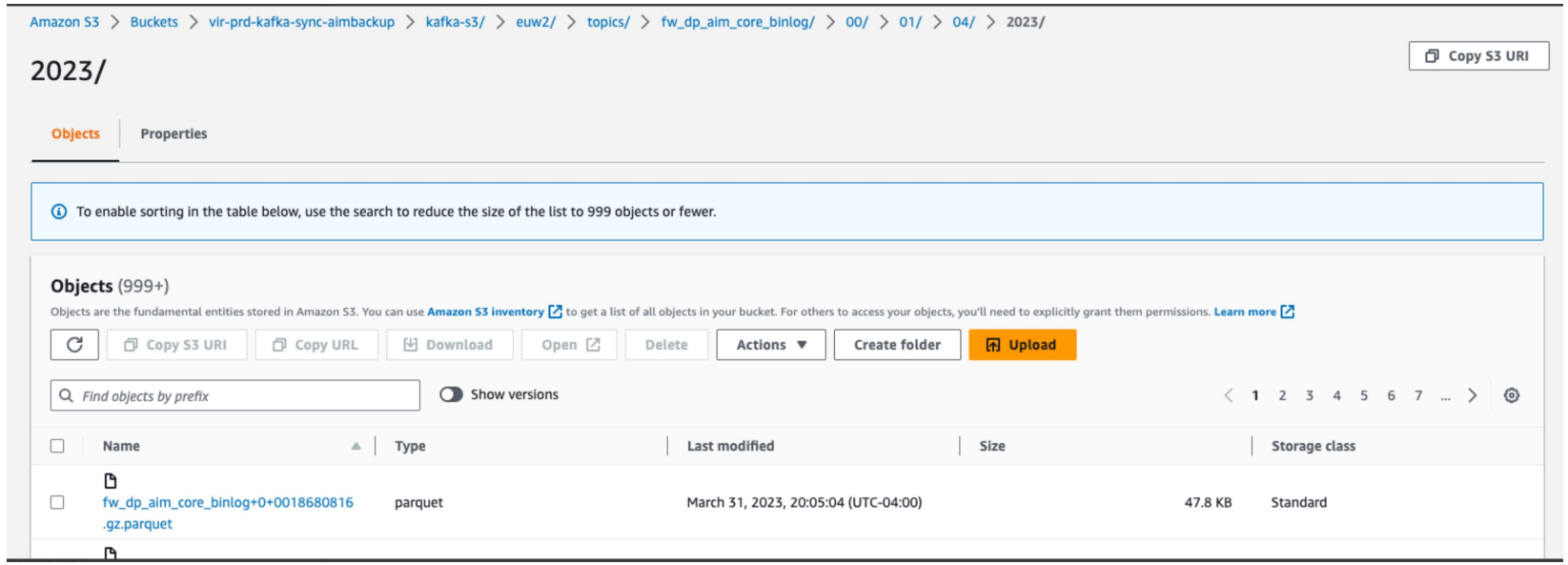
- vir-prd-kafka-sync-s3backup
- fw_dp_kafka_connect_aim_log_s3_backup
- vir-prd-kafka-sync-rtbackup
- vir-prd-kafka-sync-mkplbackup
- Etc…

© 2024 FreeWheel. Confidential.

FREEWHEEL

4. Data in S3

Time Based Partitioner:
<prefix>/<topic>/year=YYYY/month=MM/day=dd/hour=HH/<topic>+<kafkaPartition>+<startOffset>.<format>.HH/dd/MM/YYYY



© 2024 FreeWheel. Confidential.

FREEWHEEL

5. Config of Connectors

Source Connector

```yaml
- name: source-kafka_use1_local-use1_center-market_ingestion
  config: {
    "connector.class": "com.fw.kafka.connect.source.kafka.KafkaSourceConnector",
    "source.bootstrap.servers": "rawlog-local.kafka.us-east-1.prd.ads.aws.fwmrm.net:9092",
    "connector.consumer.group.id": "fw_dp_kafka_connect_use1_center_market_ingestion",
    "topic.mapping": "",
    "tasks.max": "2",
    "name": "source-kafka_use1_local-use1_center-market_ingestion",
    "topic": "fw_ads_market_ad_ingestion",
    "producer.override.compression.type": "zstd",      Chenlei Zhao, 3 weeks ago • OPS-16743
    "producer.override.max.request.size": "2097152",
    "connector.consumer.client.id": "fw_dp_kafka_connect_use1_center",
    "value.converter": "org.apache.kafka.connect.converters.ByteArrayConverter",
    "key.converter": "org.apache.kafka.connect.converters.ByteArrayConverter"
  }
```

Sink Connector

```yaml
- name: sink_kafka_local-s3_optimization_engine_log
  config: {
    "connector.class": "io.confluent.connect.s3.S3SinkConnector",
    "tasks.max": "8",
    "topics": "fw_ads_optimization_engine",
    "s3.region": "us-east-1",
    "s3.bucket.name": "vir-prd-bidding-strategy",
    "s3.part.size": "5242880",
    "flush.size": "100000",
    "rotate.schedule.interval.ms": "300000",
    "storage.class": "io.confluent.connect.s3.storage.S3Storage",
    "format.class": "io.confluent.connect.s3.format.parquet.ParquetFormat",
    "value.converter": "com.fw.kafka.connect.converters.ProtobufConverter",
    "key.converter": "org.apache.kafka.connect.converters.ByteArrayConverter",
    "value.converter.protobuf.schema.registry.url": "http://schema-registry.kafka.fw1.aws.fwmrm.net:8091",
    "value.converter.protobuf.schema.registry.subject": "af_optimization_engine_log_pb_schema",      Nick
    "schema.compatibility": "NONE",
    "partitioner.class":"io.confluent.connect.storage.partitioner.TimeBasedPartitioner",
    "partition.duration.ms":3600000,
    "path.format": "dd/MM/YYYY",
    "locale":"en-US",
    "timezone":"UTC",
    "timestamp.extractor":"Record",
    "topics.dir":"kafka-sync/fw_ads_optimization_engine_log/euw2",
    "consumer.override.group.id": "fw_dp_kafka_connect_aws_local_oe_s3_backup",
    "parquet.codec":"zstd",
    "prefix.wrapper.enable":"False"
  }
clusterName: EUC1-S3-Backup
```

© 2024 FreeWheel. Confidential.

FREEWHEEL

6. Sink Connectors Tuning

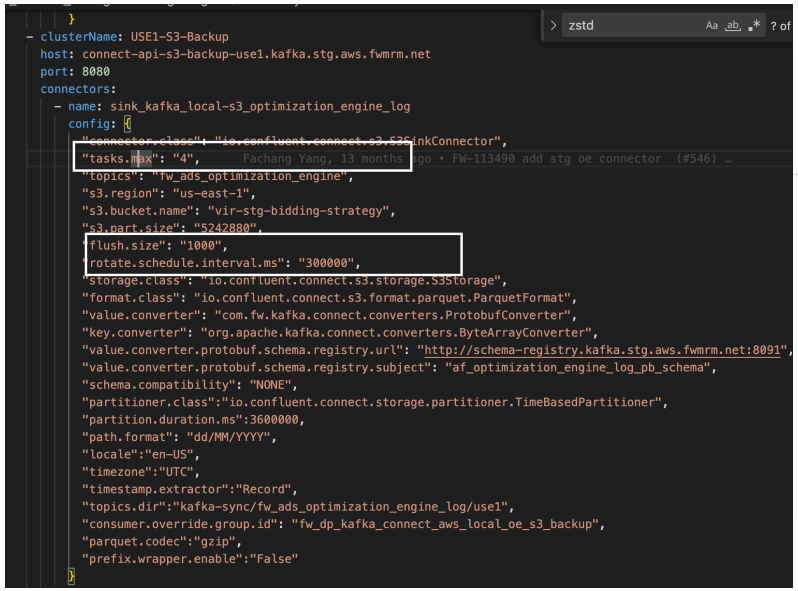
When there is lag happens, we need to tune the connector configuration.

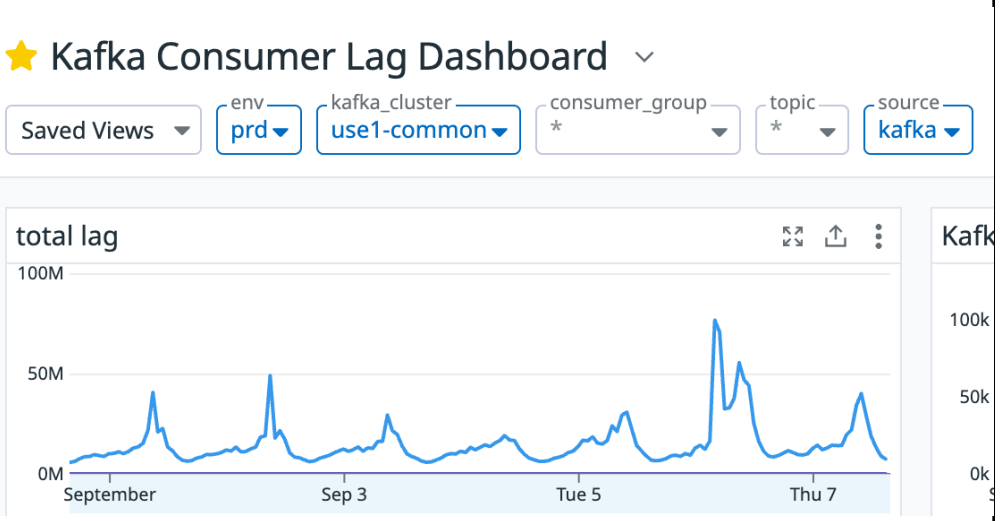
1. **Check the CPU first of the Connector cluster, if the CPU runs high, add more resources.**

2. **Task.max**

3. **Flush.size**

4. **Rotate.schedule.interval**





© 2024 FreeWheel. Confidential.

FREEWHEEL

5. BAU works of related configs

- **Request**

- **Answer**

一个kafka的问题，通过butress的kafka包消费topic fw_ads_delivery_info 的message时，发现header里的pool和group字段读取到的是base64编码后的数据
"group":"YXdzLW52LXN0Zw=="","pool":"YWRzLWRlZmF1bHQ="
http://manager.kafka.stg.aws.fwmrm.net:8088/ui/stg-center-us-east/topic?search=fw_ads_delivery_info&topicListView=HIDE_INTERNAL

但在regression中手动用buttress produce的message没有出现这个问题，这里直观感觉区别就是stg中的message是gads produce的，没有用buttress，但目前没有确认到相关代码，所以现在想确认：
1 header的value是否确实有base64编码?
2 如果有，是gads/kafka broker/ buttress customer 哪部分做的?

```
  }
- name: source-kafka_aws_local-use1_center-Dr_Strange
  config: {
    "connector.class": "com.fw.kafka.connect.source.kafka.KafkaSourceConnector",
    "source.bootstrap.servers": "b0.local.kafka.stg.aws.fwmrm.net:9092,b1.local.kafka.stg.aws.fwmrm.net:9092,b2.local.kafka.stg.av
    "connector.consumer.group.id": "fw_dp_kafka_connect_use1_center",
    "topic.mapping": "",
    "producer.override.max.request.size": "2097152",
    "tasks.max": "2",
    "name": "source-kafka_aws_local-use1_center-Dr_Strange",
    "topic": "fw_ads_delivery_info",
    "producer.override.compression.type": "gzip",
    "connector.consumer.client.id": "fw_dp_kafka_connect_use1_center",
    "value.converter": "org.apache.kafka.connect.converters.ByteArrayConverter",
    "header.converter": "org.apache.kafka.connect.converters.ByteArrayConverter",      Nick Yang, 2 weeks ago • OPS-168878 config
    "key.converter": "org.apache.kafka.connect.converters.ByteArrayConverter"
  }
```

© 2024 FreeWheel. Confidential.

FREEWHEEL

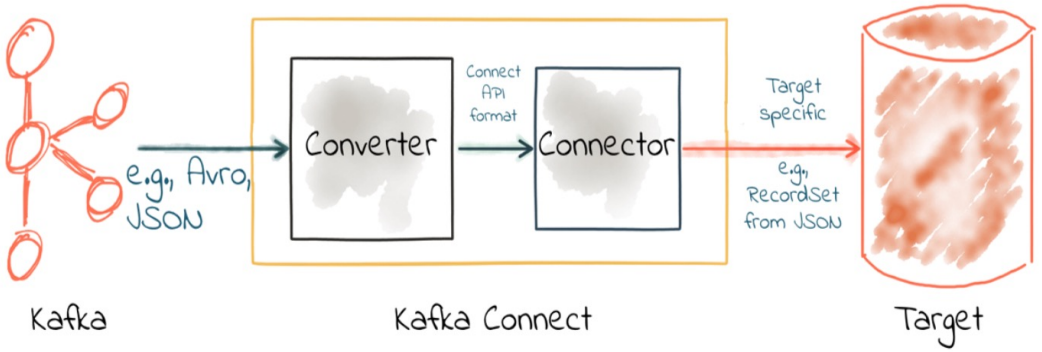
6. Serialization in Kafka

Kafka messages are *just pairs of key/value bytes* on brokers.

Serialization is a general term that covers *deserializing* and *serializing*.

When you push an array of bytes through a *deserializer*, it gives you an object on the other end:

A *serializer* is just the opposite—you give it an object, and it returns an array of bytes:



© 2024 FreeWheel. Confidential.

FREEWHEEL

Kafka Partition

FREEWHEEL
A COMCAST COMPANY

FEBRUARY 29, 2024

1. Background

0	0	2	(2,1,0)	(1,0,2)	true	false
1	0	1	(1,0,2)	(1,0,2)	true	false
2	0	0	(0,2,1)	(0,1,2)	true	false
3	0	2	(2,0,1)	(0,1,2)	true	false
4	0	1	(1,2,0)	(1,0,2)	true	false
5	0	0	(0,1,2)	(0,1,2)	true	false
6	0	2	(2,1,0)	(1,0,2)	true	false
7	0	1	(1,0,2)	(1,0,2)	true	false
8	0	0	(0,2,1)	(0,1,2)	true	false
9	0	2	(2,0,1)	(0,1,2)	true	false
10	0	1	(1,2,0)	(1,0,2)	true	false
11	0	0	(0,1,2)	(0,1,2)	true	false
12	0	2	(2,1,0)	(1,0,2)	true	false
13	0	1	(1,0,2)	(1,0,2)	true	false
14	0	0	(0,2,1)	(0,1,2)	true	false
15	57,773,686	2	(2,0,1)	(2,1)	true	true

All message wrote to the partition 15 in short time with

Two under replicated partitions

Yunliang He 4:34 PM
为啥以前就一个 partition，也不报错。

这次也写一个，就不行呢。

Chenlei Zhao 4:35 PM
以前是一个分区一个副本都在一个机器上，不需要备份

你现在是三个备份，写的太快了，其他两个来不及fetch

而且下一步你这个写均匀了，就可以考虑限速了

不可能一下写这么大量

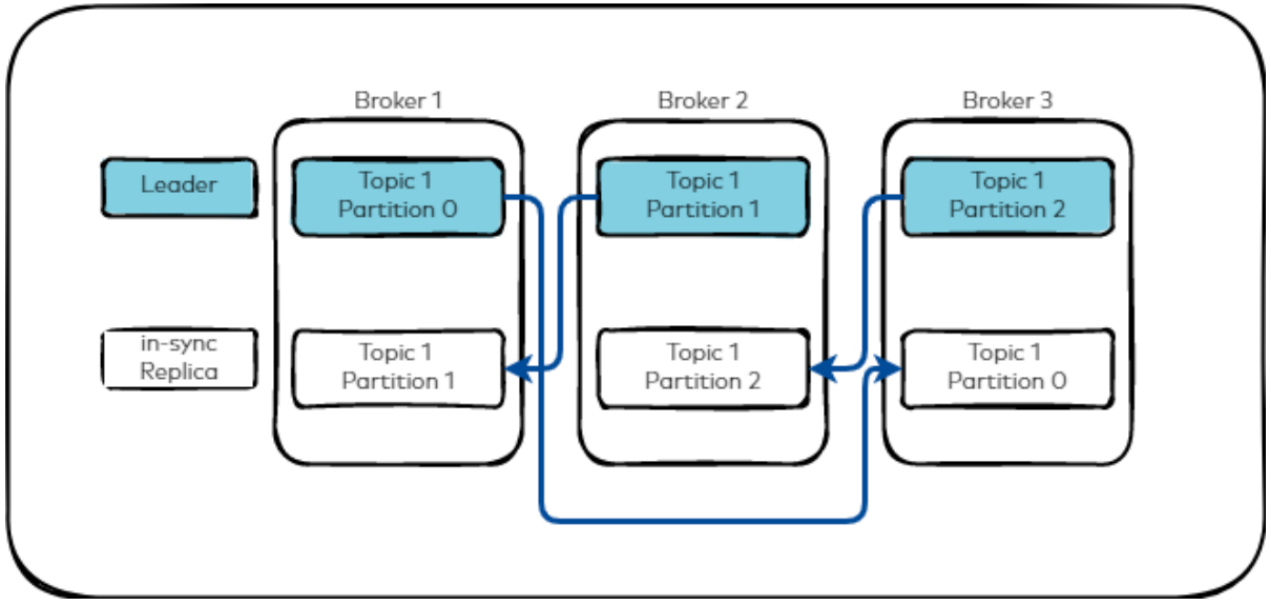
Yunliang He 4:37 PM
哦哦

我先均匀写一下

© 2024 FreeWheel. Confidential.

FREEWHEEL

1.Replications



ISR (In-Sync Replication)

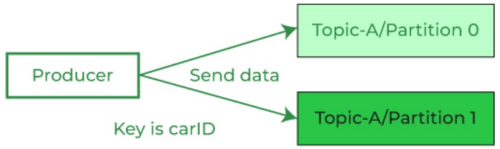
1.A node must be able to maintain its session with the controller

2.If it is a follower it must replicate the writes happening on the leader and not fall “too far” behind

© 2024 FreeWheel. Confidential.

FREEWHEEL

2. Kafka Message Key



Example:

carID_123 data will always be in partition 0
carID_234 data will always be in partition 0

carID_345 data will always be in partition 1
carID_456 data will always be in partition 1

Message Key generated

- Round-Robin
- **All** the messages that share the same key will always go to the same partition in **time order**.

© 2024 FreeWheel. Confidential.

FREEWHEEL

Take Advantage of AI

- **Capacity Planning**
- **Anomaly Detection**

FREEWHEEL
A COMCAST COMPANY

FEBRUARY 29, 2024

Capacity Planning

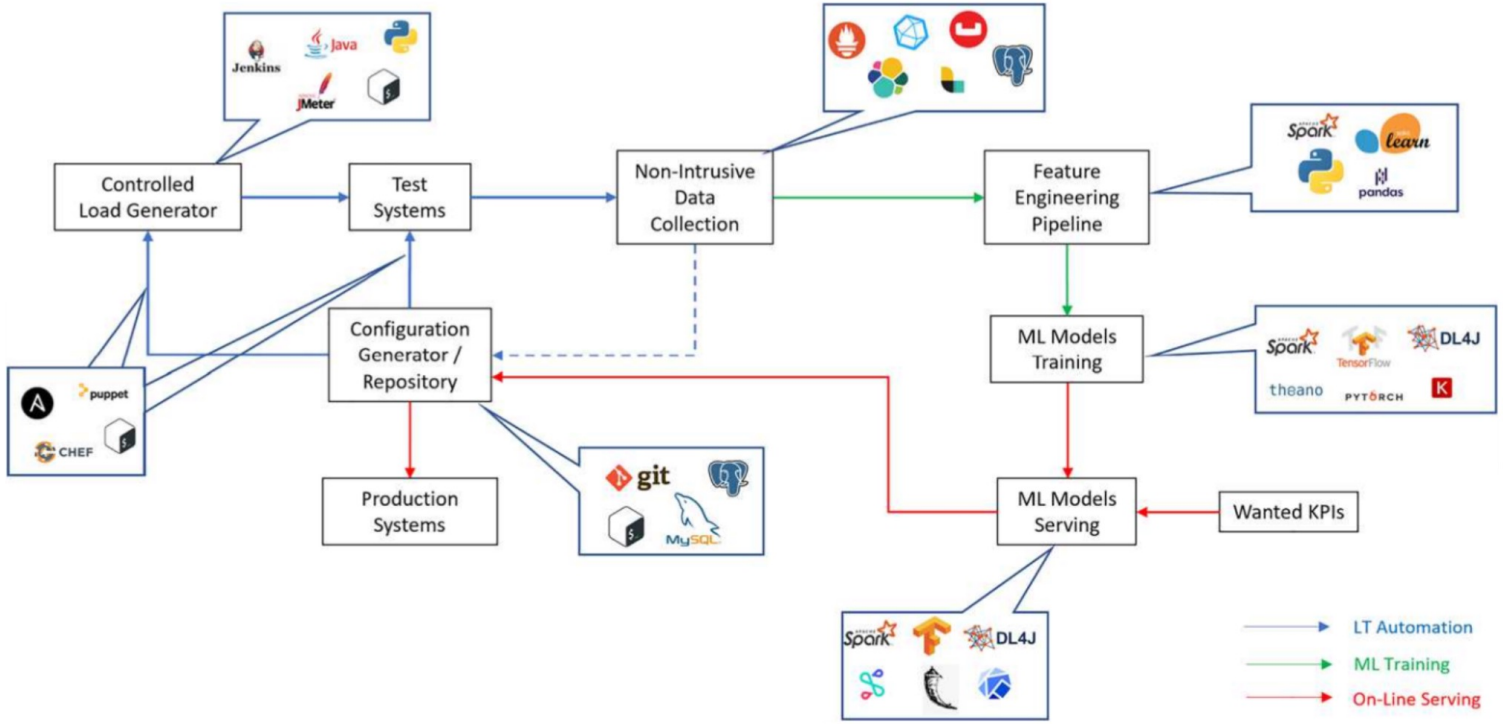
Motivation

- Due to the complexity of modern large-scale systems, there may be hundreds of combinations of configuration parameter values, and environment and deployment settings.
- Estimating the system throughput given a set of configuration parameter values is important for capacity planning and reducing costs.

© 2024 FreeWheel. Confidential.

Capacity Planning

Approach



Complete Process with its three stages:

- Automated load testing
- Feature Engineering and ML Training
- Model On-Line Serving

© 2024 FreeWheel. Confidential.

FREEWHEEL

Capacity Planning

Approach

Feature Engineering and ML Training

- background.threads: The number of threads to use for various background processing tasks. The range of values we used in our experiments for this parameter was [5-30]
- num.network.threads: The number of threads that the server uses for receiving requests from the network and sending responses to the network. The range of values we used in our experiments for this parameter was [3-6].
- num.network.threads: The number of threads that the server uses for receiving requests from the network and sending responses to the network. The range of values we used in our experiments for this parameter was [3-6].
- num.replica.fetchers: The number of fetcher threads used to replicate messages from a source broker. Increasing this value can increase the degree of I/O parallelism in the follower broker. The range of values we used in our experiments for this parameter was [1-2].

1	LoopId	LoopStartTime	BackgroundThreads	LogCleanerThreads	NumIoThreads	NumNetworkThreads	NumPartitions	NumNodes	NumReplicaFetchers	ThreadsClient	MessageSize	TotalMessages
2	1	2019-07-06_03-02-41	10	1	4	3	1	1	1	10	10240	333614
3	2	2019-07-06_03-37-18	10	1	4	3	1	1	1	10	102400	74311
4	3	2019-07-06_04-11-56	10	1	4	3	1	1	1	15	10240	485829
5	4	2019-07-06_04-46-36	10	1	4	3	1	1	1	15	102400	75750
6	5	2019-07-06_05-20-39	10	1	4	3	1	1	1	20	10240	613018
7	6	2019-07-06_05-55-22	10	1	4	3	1	1	1	20	102400	76619
8	7	2019-07-06_06-30-01	10	1	4	3	1	1	1	25	10240	682651
9	8	2019-07-06_07-04-46	10	1	4	3	1	1	1	25	102400	76257
0	9	2019-07-06_07-39-25	10	1	4	3	1	1	1	50	10240	726962
1	10	2019-07-06_08-14-07	10	1	4	3	1	1	1	50	102400	76106
2	11	2019-07-06_08-48-48	10	1	4	3	1	1	2	10	10240	333584
3	12	2019-07-06_09-23-24	10	1	4	3	1	1	2	10	102400	75049

© 2024 FreeWheel. Confidential.

FREEWHEEL

Capacity Planning

Outcome

Feature Engineering and ML Training

System Type	Model	R2 Score	Median (%) Deviation	MAPE	MAE	MSE	RMSE
Open-Src.	XGBoost	0.99964	0.0865 %	0.8060 %	2,193	2.4028e7	4,901
	Random F.	0.99958	–0.1828 %	0.6617 %	2,266	2.8253e7	5,315
	MLP NN	0.99969	0.1648 %	0.6822 %	2,110	2.0594e7	4,538
	CNN	0.99949	0.1590 %	1.1940 %	3,555	3.4482e7	5,872
	LSTM NN	0.99902	0.2583 %	1.4341 %	5,586	6.6280e7	8,141
	Linear Regr.	0.90930	7.3951 %	41.3498 %	66,759	6.1744e9	78,579
Entprz. 1	XGBoost	0.99324	0.6323 %	2.3596 %		–	
	Random F.	0.95078	-0.1450 %	4.5820 %			
	MLP NN	0.99512	2.2929 %	2.3435 %			
	CNN	0.98317	–1.0807 %	3.5182 %			
	LSTM NN	0.92908	–16.0805 %	8.0132 %			
	Linear Regr.	0.98830	4.6561 %	21.3450 %			
Entprz. 2	XGBoost	0.99973	0.8091 %	4.5485 %		–	
	Random F.	0.99647	1.5526 %	67.5510 %			
	MLP NN	0.99991	1.0868 %	3.3082 %			
	CNN	0.99067	–8.1808 %	23.6534 %			
	LSTM NN	0.97589	–5.9882 %	62.2091 %			
	Linear Regr.	0.92218	4.3187 %	10.4637 %			

They found that the MAPE （Mean absolute percentage error ）values are low (0.68% to 4.5%) for models such as XGBoost and MLP Neural Network, suggesting that the prediction results have a high accuracy.

Mean Absolute Percentage Error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method and measures the size of the error in percentage terms. It is calculated by the formula: $MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{actual_i - predicted_i}{actual_i} \right|$, given n points in the testing set.

© 2024 FreeWheel. Confidential.

FREEWHEEL

Capacity Planning

Results

- The ML models can predict the throughput with a high accuracy. In particular, XGBoost and MLP Neural Network achieve the best prediction result and the findings are consistent across the studied systems.

- Engineers could discover the large set combinations of configuration parameter values and environment settings

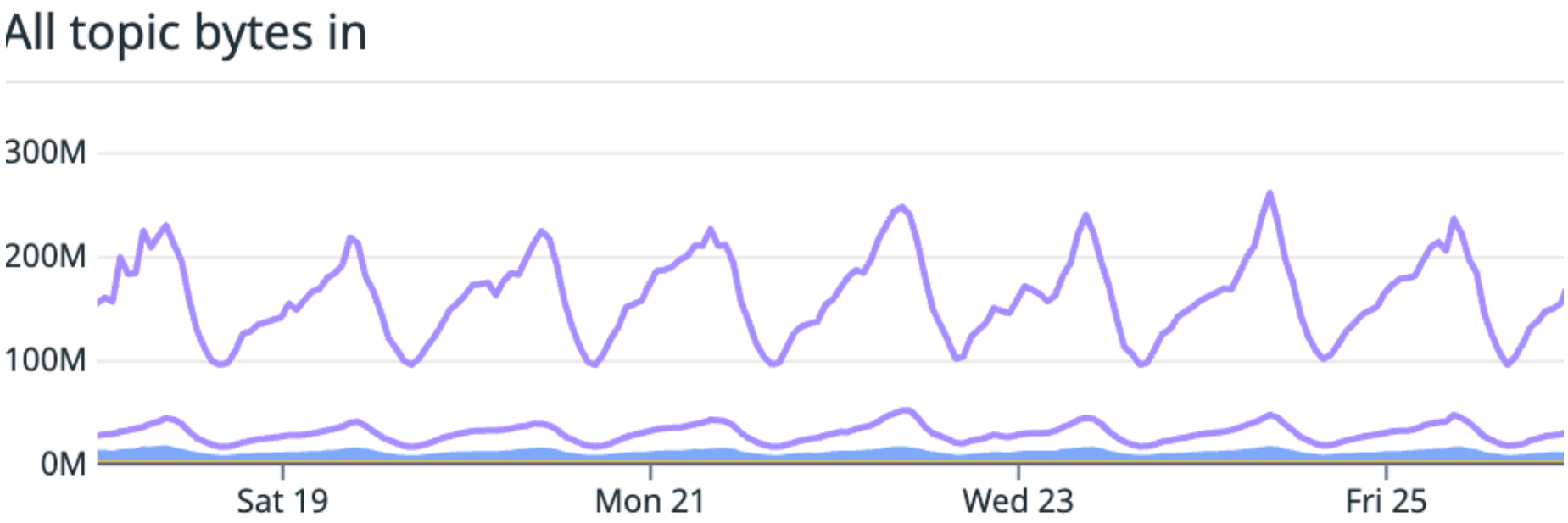
- After being applied to the industry, The results from these simulations pointed out the relationship between different parameters (underlying infrastructure and software configurations) and provided important insight to development teams on where improvement efforts could be focused in the next iterations.

© 2024 FreeWheel. Confidential.

Anomaly Detection

Motivation

- **When received the alert can not tell the root cause by eyes at immediately.**



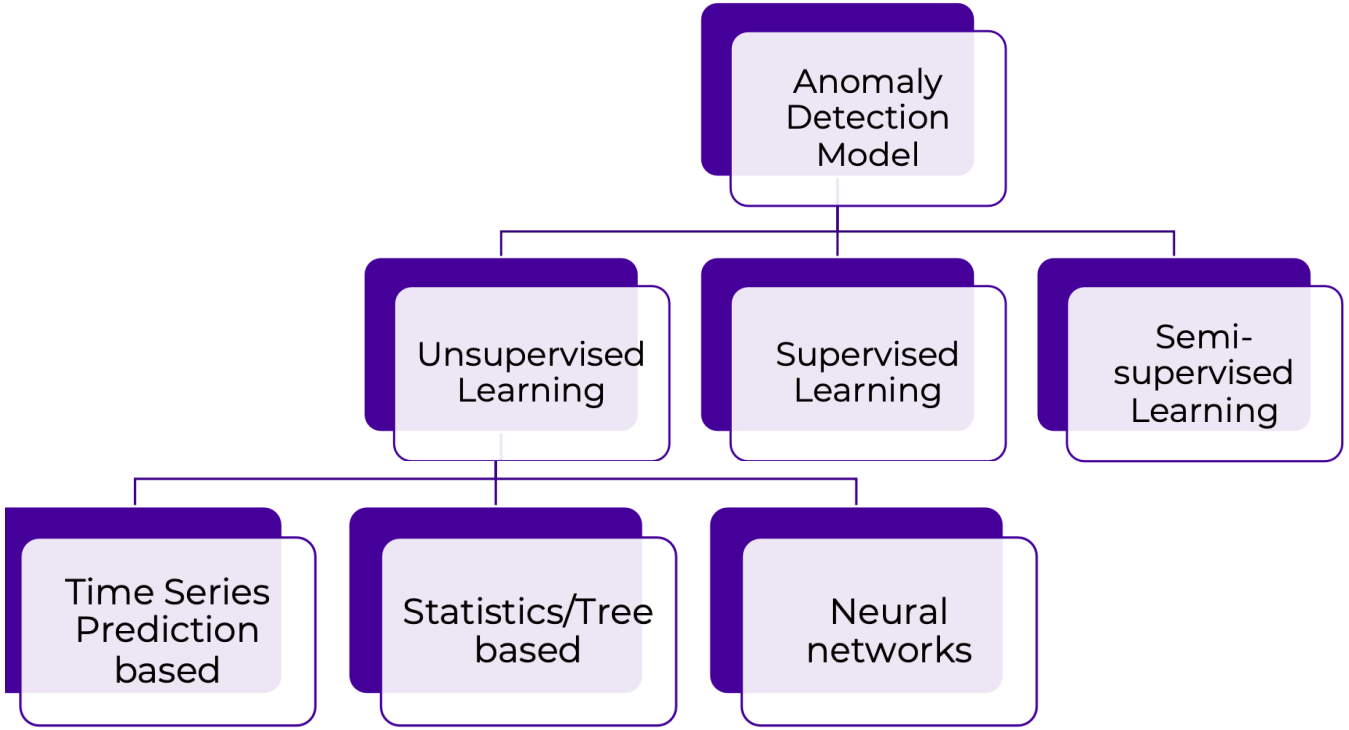
© 2024 FreeWheel. Confidential.

Anomaly Detection

Approach

topic_byte_in

Time	{topic="fw_ads_binary_log"}
2023-08-30 12:36:15	111 MiB/s
2023-08-30 12:36:30	111 MiB/s
2023-08-30 12:36:45	110 MiB/s
2023-08-30 12:37:00	111 MiB/s
2023-08-30 12:37:15	116 MiB/s
2023-08-30 12:37:30	119 MiB/s
2023-08-30 12:37:45	120 MiB/s
2023-08-30 12:38:00	117 MiB/s
2023-08-30 12:38:15	116 MiB/s
2023-08-30 12:38:30	113 MiB/s
2023-08-30 12:38:45	110 MiB/s
2023-08-30 12:39:00	107 MiB/s
2023-08-30 12:39:15	106 MiB/s
2023-08-30 12:39:30	111 MiB/s
2023-08-30 12:39:45	112 MiB/s
2023-08-30 12:40:00	110 MiB/s
2023-08-30 12:40:15	111 MiB/s
2023-08-30 12:40:30	120 MiB/s
2023-08-30 12:40:45	118 MiB/s
2023-08-30 12:41:00	112 MiB/s
2023-08-30 12:41:15	108 MiB/s
2023-08-30 12:41:30	106 MiB/s
2023-08-30 12:41:45	105 MiB/s



© 2024 FreeWheel. Confidential.

FREEWHEEL

Anomaly Detection

Approach

Threshold Based Alert

Trigger alert when impression fall below that predefined threshold

Disadvantages

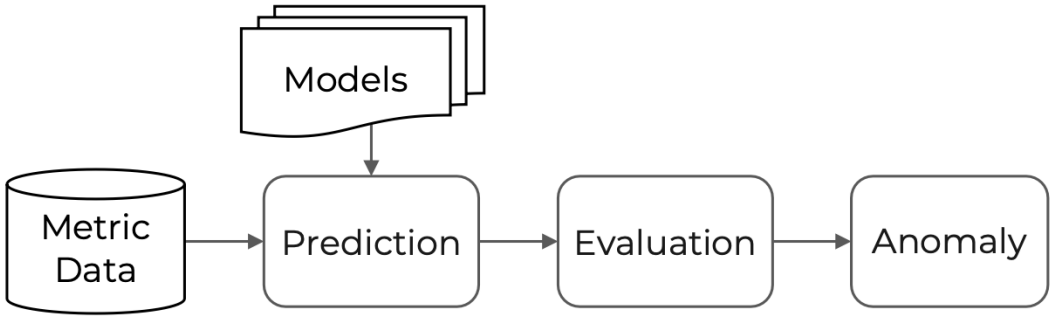
- Unable to detect abnormal drop during high traffic period.
- False alerts are caused by traffic's uncertainty and noise

© 2024 FreeWheel. Confidential.

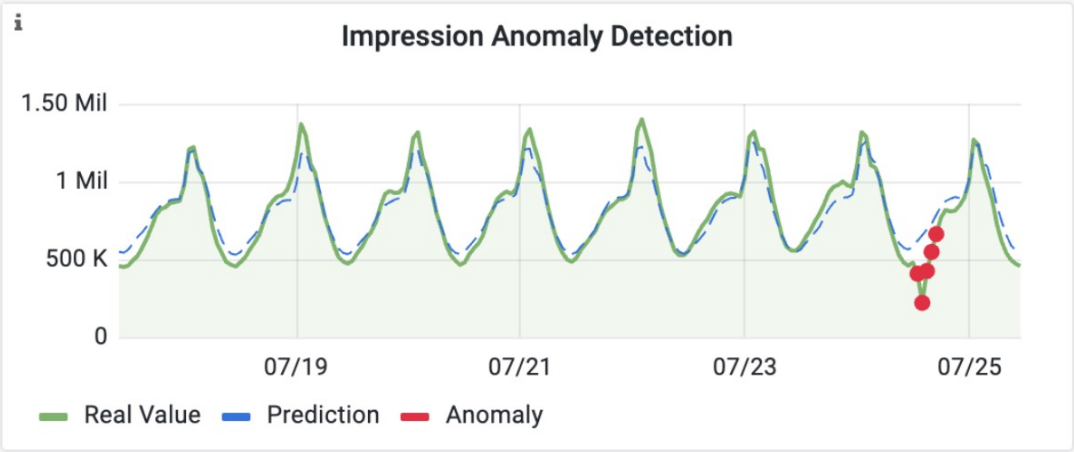
Anomaly Detection

Approach

TIME SERIES ANOMALY DETECTION & MODEL



Result



Real Value Predicted Value Anomaly Candidates

Advantages

- Find anomalies with high severity
- High precision (90%+)
- Low latency

© 2024 FreeWheel. Confidential.

FREEWHEEL

Anomaly Detection

Approach

Wave

Alert rule

Metrics

m1 : MIN_OVER_TIME | <compute> | fw1_data-Cortex::kafka_server_produce_client_id_throttle_time

Check Level

Check each combinations of: client_dsh_id ×

Send individual alert per: Please select the extra tags you want to add in the email title

☐ Include time field in the alert title ⓘ

Global Filter

f1	type	MATCH	^.*broker.*$
f2	env	IN	prd ×
f3	kafka_cluster	IN	use2-local ×

Expr: ⓘ f1 AND f2 AND f3

Metrics Expressions

e1 : <offset_op> | m1

Schedule & TimeRange & Granularity

Metric(s) data delay is 15s ⚠, data interval is 15s ⚠.

Check last 2m data every 2m .

Aggregate 2m data with granularity 2m , align data window by granularity ⓘ .

Thresholds

Trigger WARN alert, when: e1 anomaly raise with 100% confidence

Trigger CRIT alert, when: e1 anomaly raise with 100% confidence

© 2024 FreeWheel. Confidential.

FREEWHEEL

One-Sentence Summary

Object	What I Did	What I will Do
Infrastructure	• Kafka on-call rotation • Scale in/out Kafka clusters • Scale up/down Kafka clusters • ZSTD enablement for topics written by ADS in Local ADS Kafka • Golden Image Adoption of Connect Cluster • Kafka DR	• Kafka on-call rotation • Scale in/out Kafka clusters • Scale up/down Kafka clusters
Security(CGA)	• EBS encryption (DEV)	• EBS encryption (PRD)
Cost & Efficiency	• BAU+ review	• BAU+ review for each major changes of Infra

© 2024 FreeWheel. Confidential.