

CLOUDERA

CLOUDERA STREAM PROCESSING (CSP)

Purnima Kuchikulla

Tim Spann

Abdelkrim Hadjidj

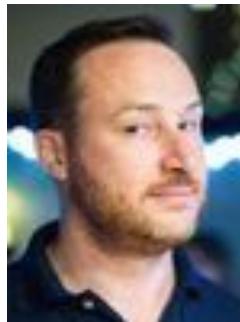
Andre Sousa Dantas De Araujo

Daniel Chaffelson

Welcome to our Edge Management Lab

Who are we?

Cloudera Data in Motion Field Team



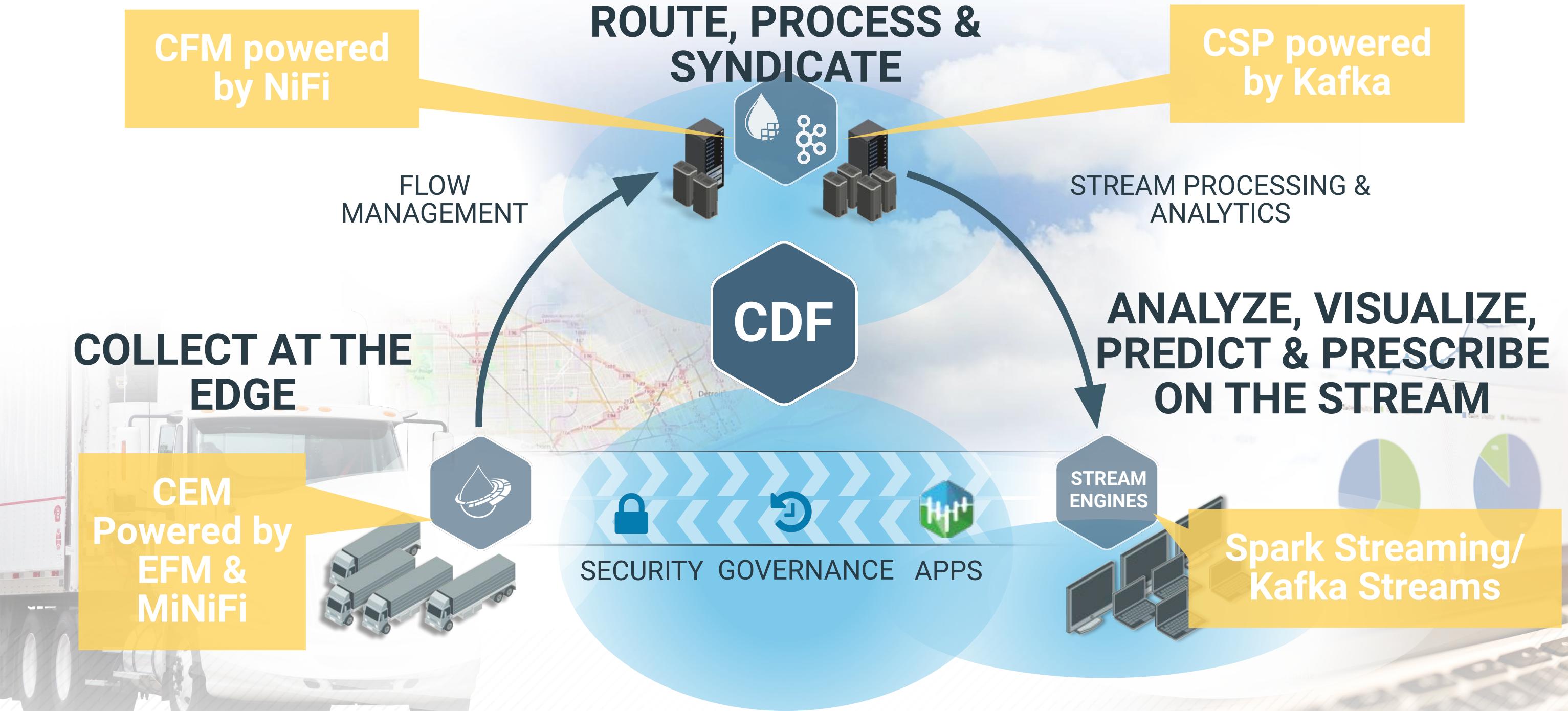
@PaasDev



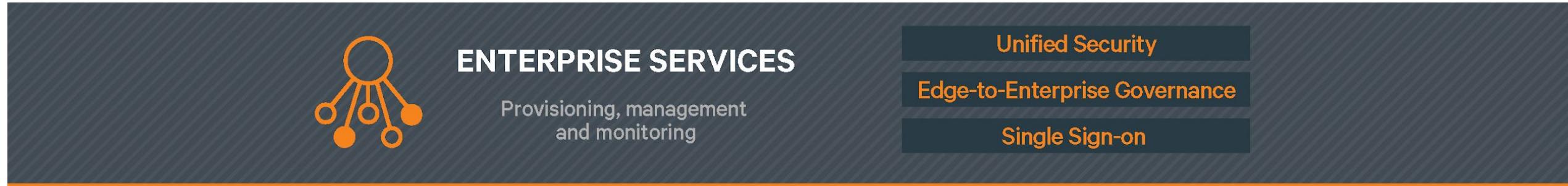
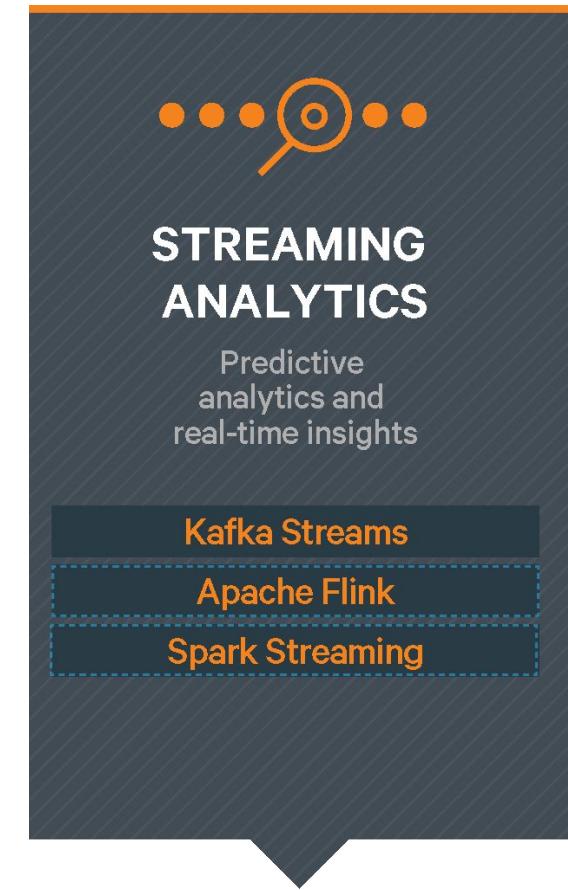
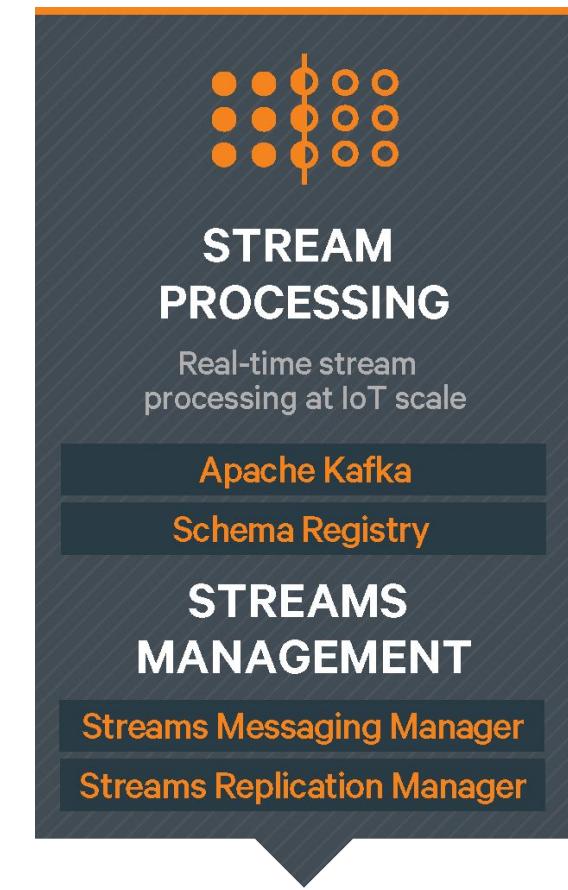
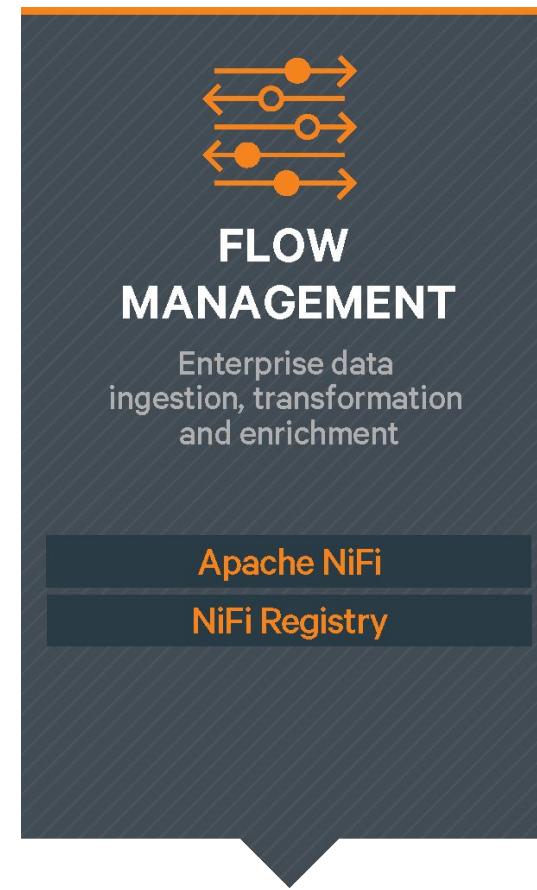
Kafka



Cloudera DataFlow (CDF)



CLOUDERA DATAFLOW DATA-IN-MOTION PLATFORM

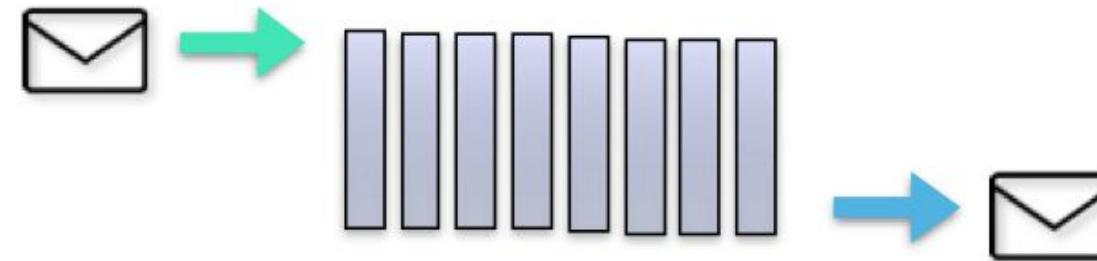


What is Kafka

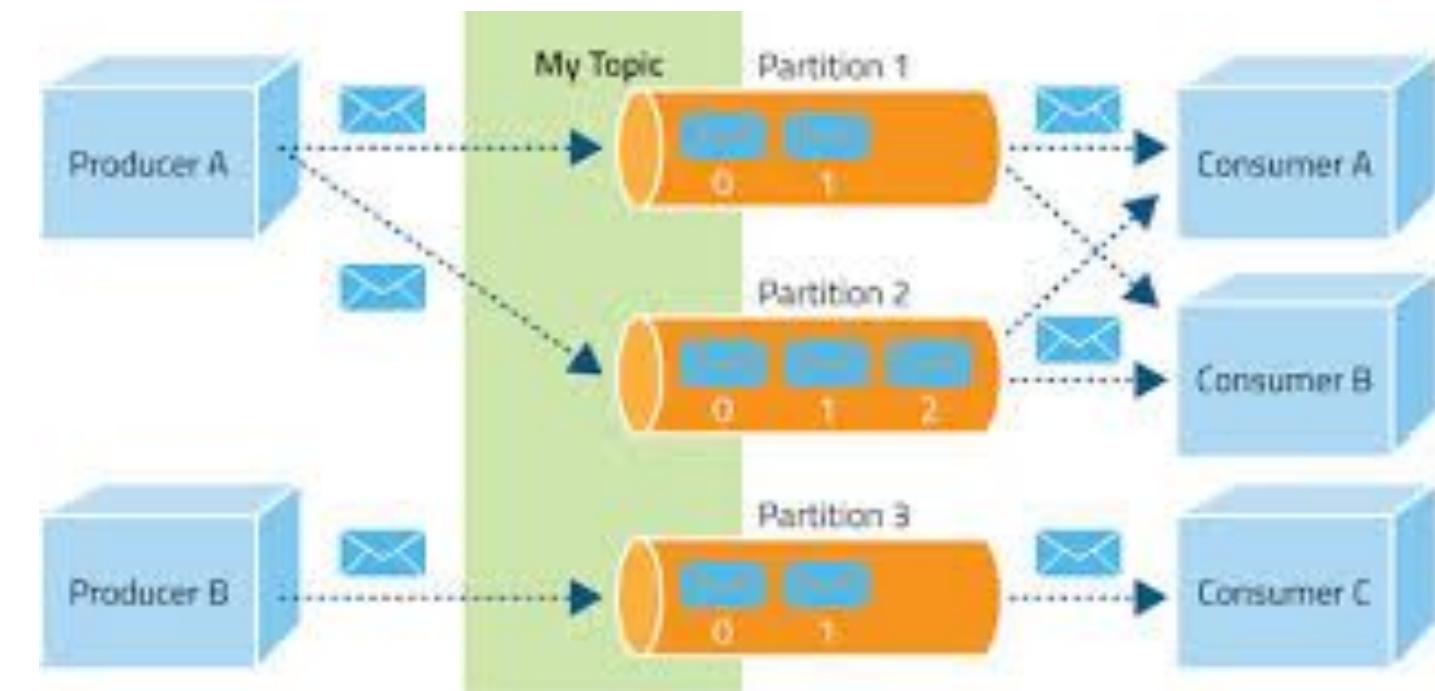


- ➊ Who is Kafka?
- ➋ What is it
 - A messaging system
 - It can handle lots of messages
 - Messages cannot get lost
 - Message order is preserved
 - Message will get delivered
 - A *producer* sends the message and one or more *consumers* can read the message
- ➌ Kafka was built at LinkedIn in 2011
- ➍ Open sourced as an Apache project

Inside Kafka



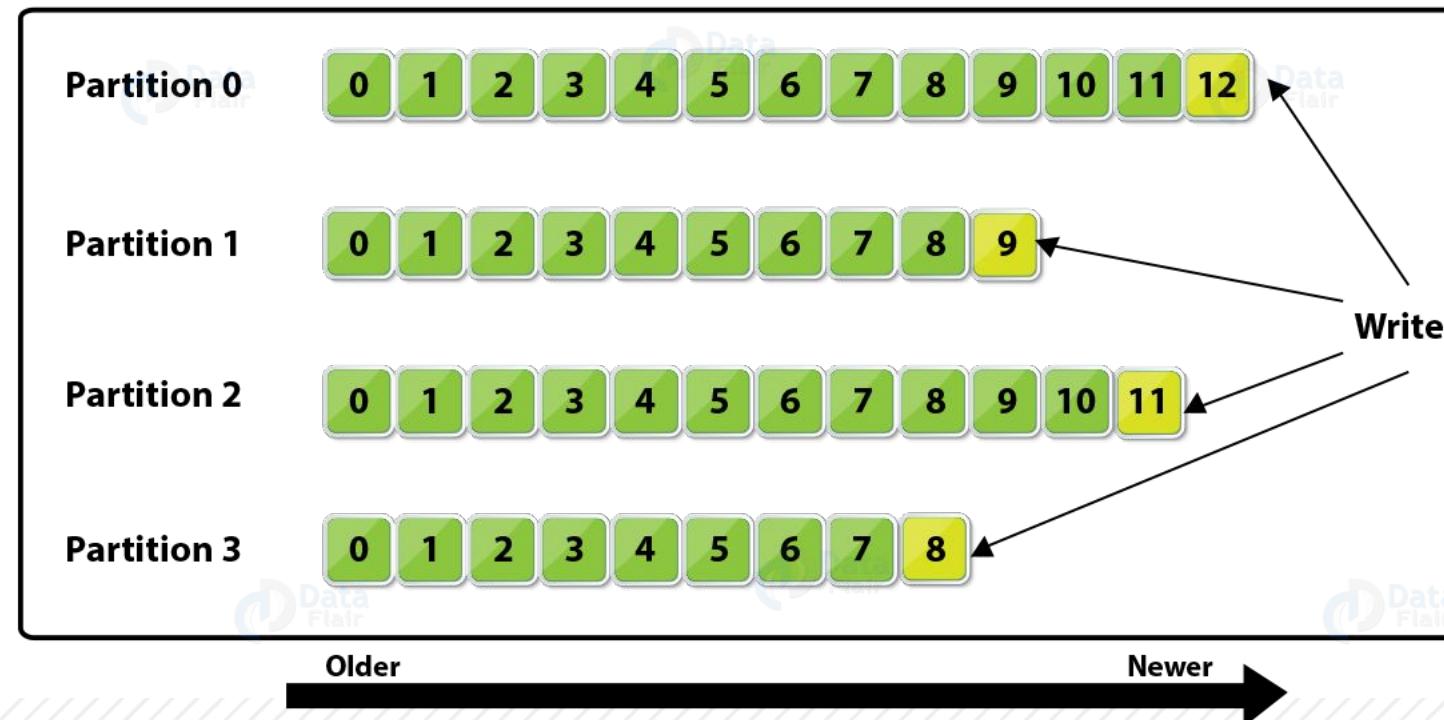
- Kafka topic and topic partition
- Kafka Broker
- Producers
- Consumers



How does Kafka preserve message order

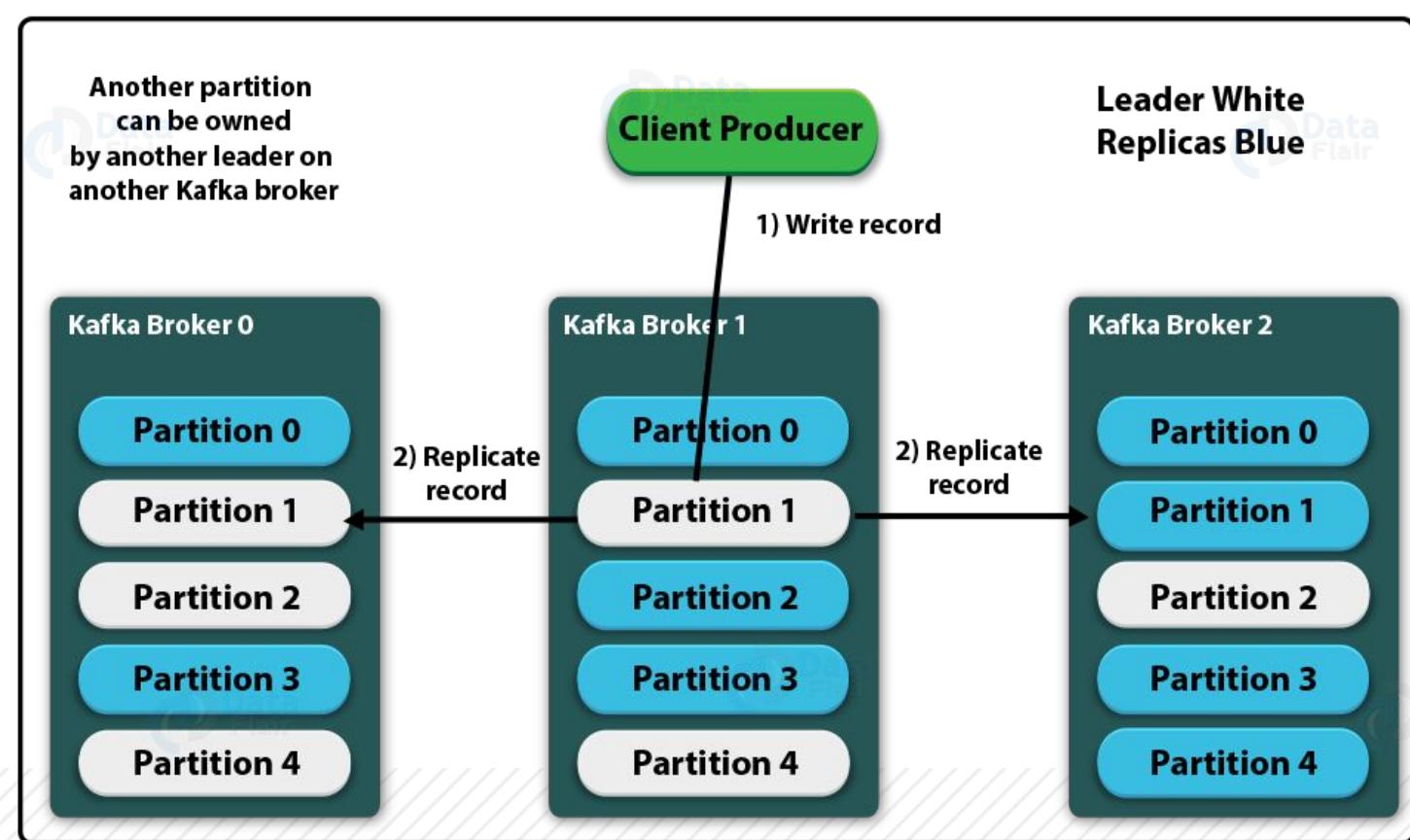
- Partition algorithm is fixed (hash on key)
- Stored as a log sequential write to a file
- Consume in order based on offset

Kafka Topic Partitions Layout

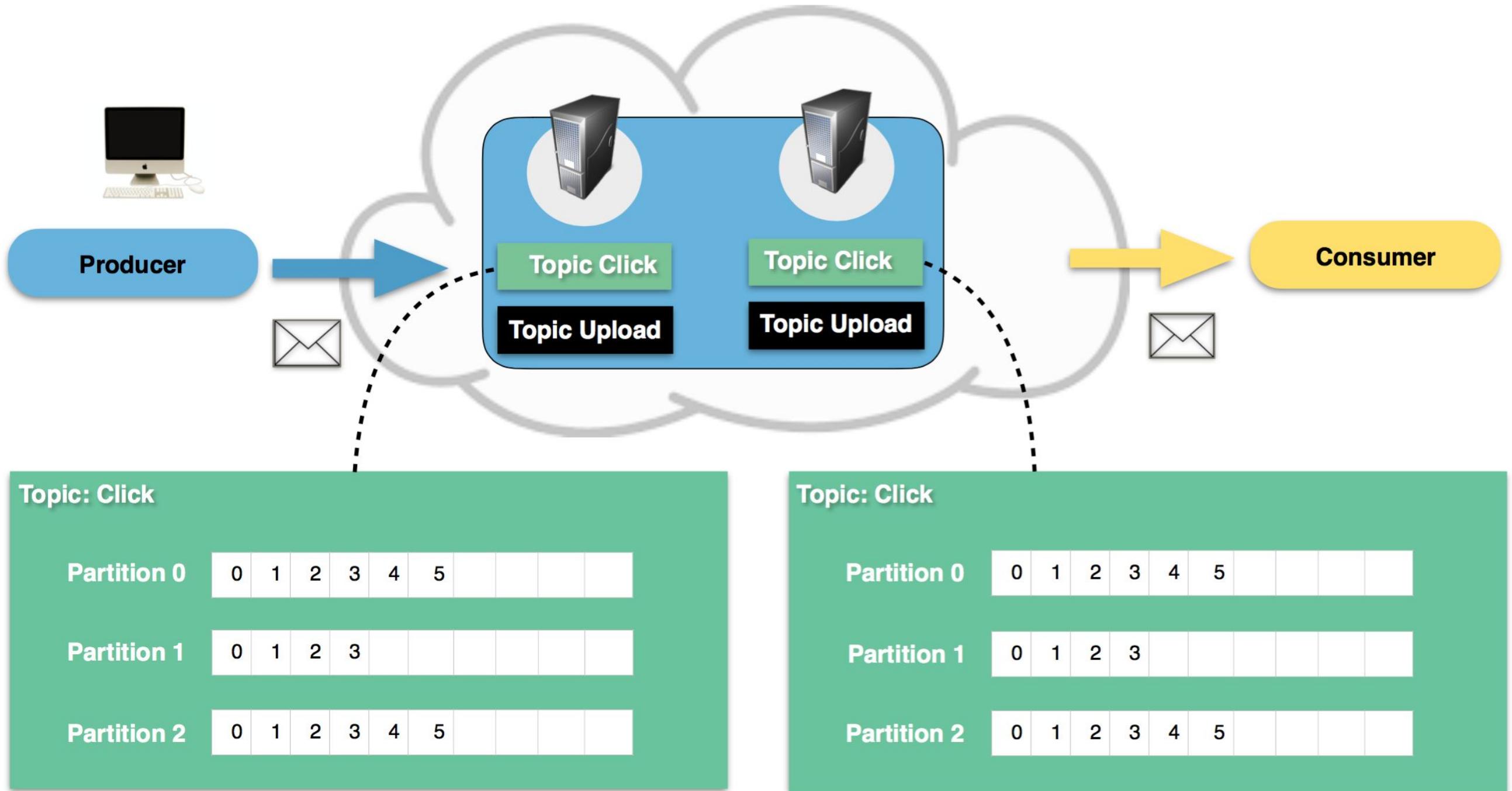


How does Kafka prevent data loss

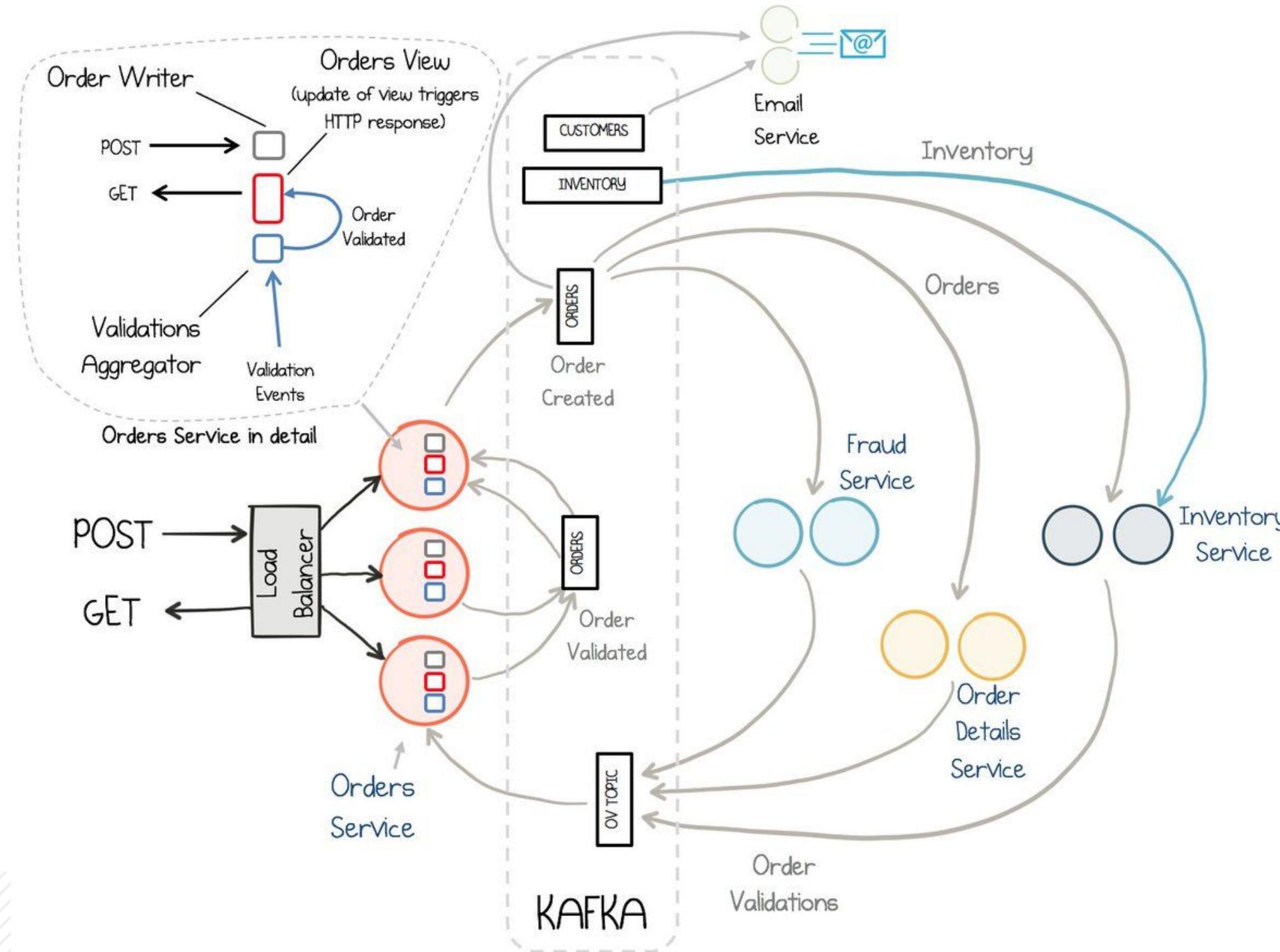
- Replicate, replicate, replicate
- Acknowledge you got the message
- Keep it even after it is consumed



Kafka Use Case: Clickstream Data



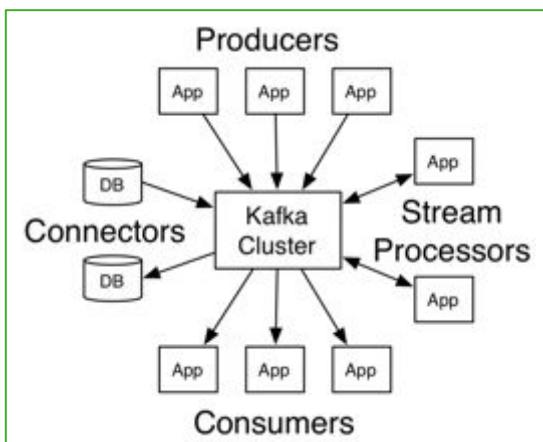
Kafka Use Case: Online ordering system



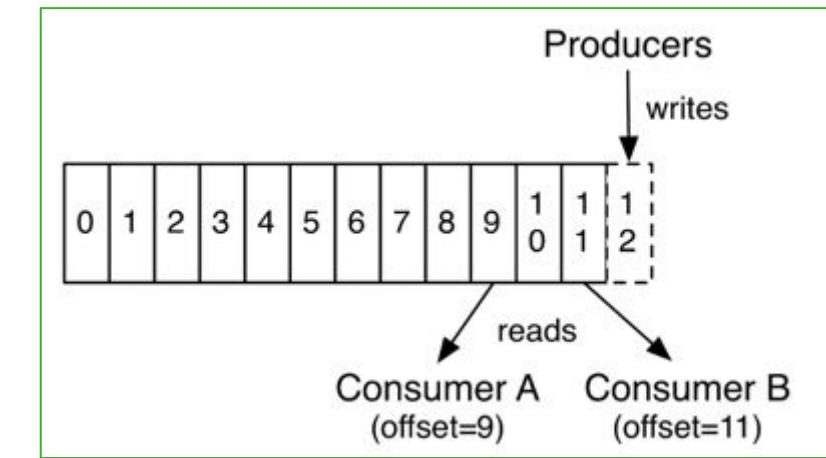
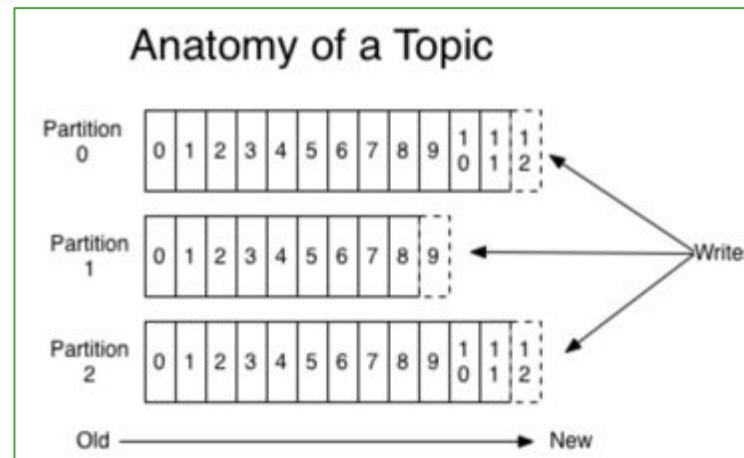
Kafka Basics - Summary

Kafka has 4 core APIs

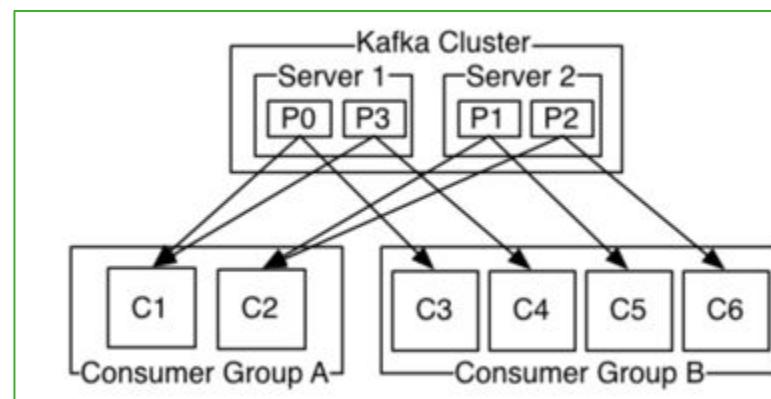
1. [Producer API](#)
2. [Consumer API](#)
3. [Streams API](#)
4. [Connector API](#)



Anatomy of a Kafka Topic



Kafka Consumers



Kafka's Omnipresence Has Led to the Onset of “Kafka Blindness”

- ◆ What is “Kafka Blindness”?
 - Customers who use Kafka today struggle with monitoring / “seeing”/troubleshooting what is happening in their clusters
- ◆ Who is Affected?
 - Platform Operation Teams
 - Developers / DevOps Teams
 - Security / Governance Teams
- ◆ What are the Symptoms?
 - Difficulty seeing who is producing and consuming data
 - Difficulty understanding the flow of data from producers -> topics consumers
 - Difficulty troubleshooting/monitoring.

Streams Messaging Manager (SMM)

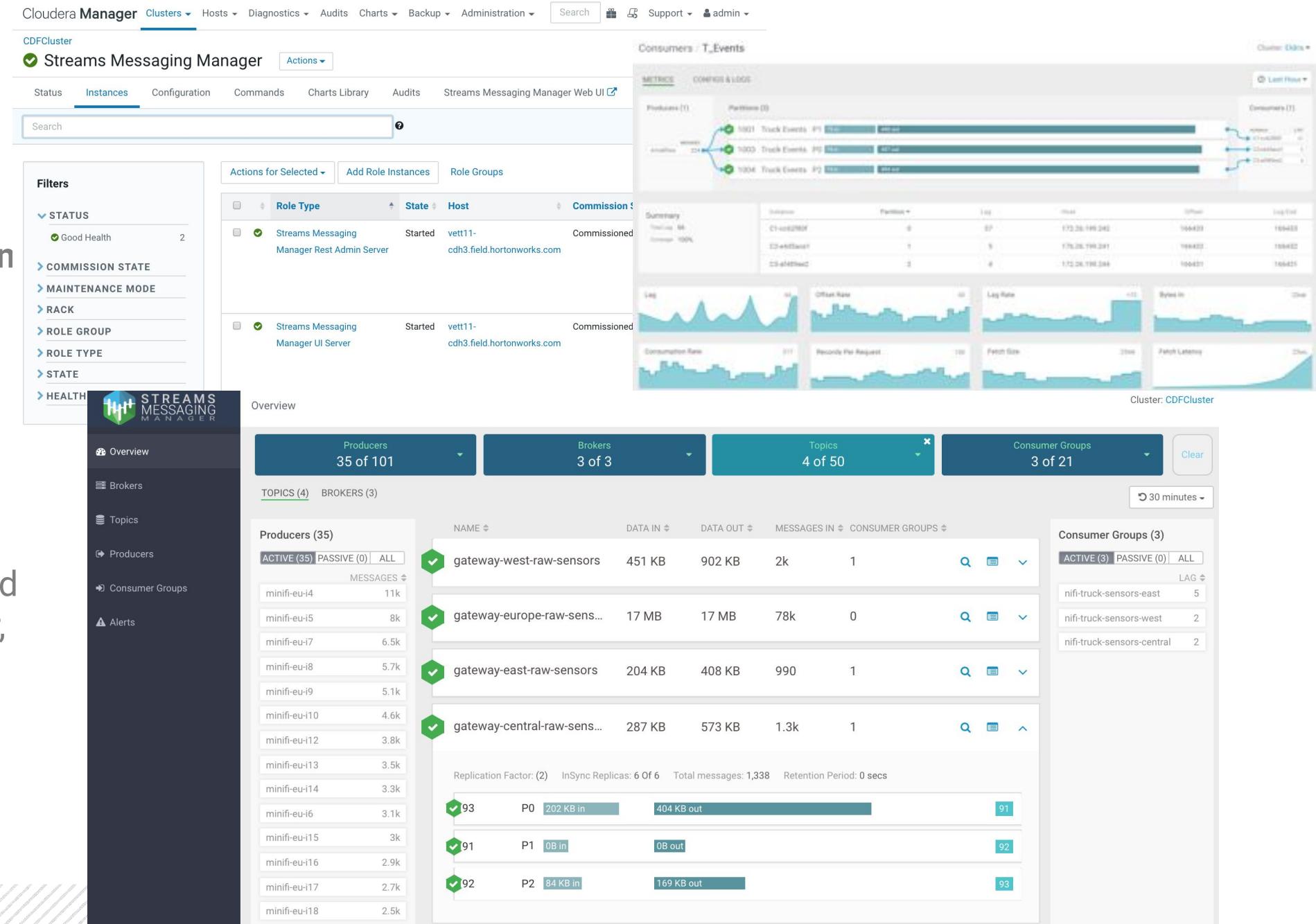
Curing the Kafka Blindness

Problem Statement / Requirements

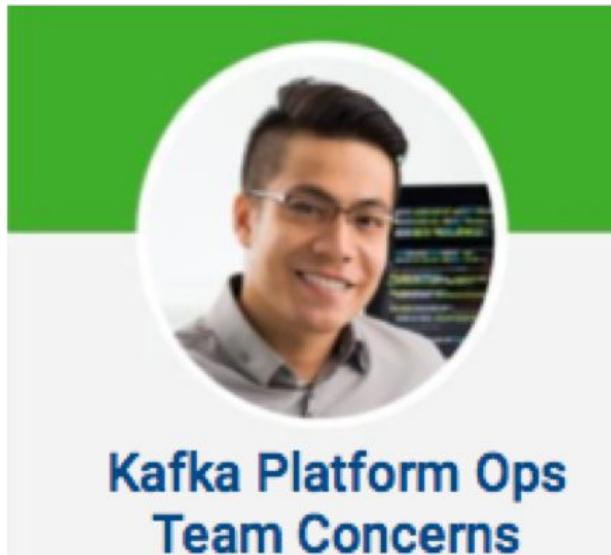
- Customers struggle with monitoring / “seeing”/ troubleshooting what is happening in their Kafka clusters
- They need an enterprise Kafka monitoring solution full integrated with platform services including CM and Sentry

Solution

- SMM provides single monitoring Dashboard for Kafka Clusters across 4 entities: Broker, Producer, Topic, Consumer
- SMM integrated with CM Log Search
- Kerberized based authentication and rich access Control via Sentry
- DataPlane Platform NOT required

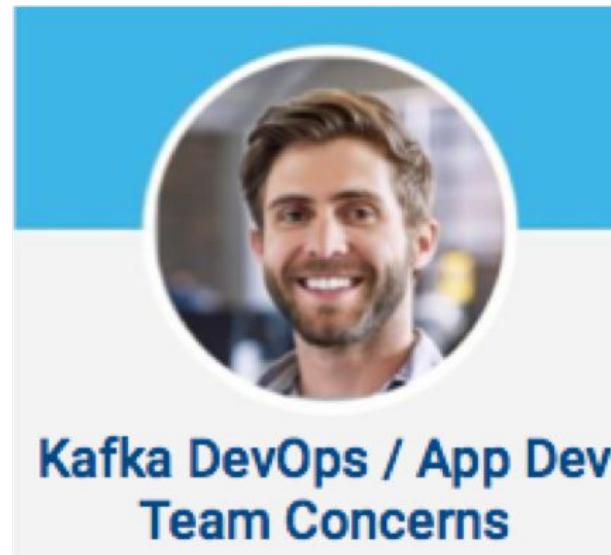


SMM Addresses the Distinct Needs of 3 Personas/Teams



**Kafka Platform Ops
Team Concerns**

**Concerned with monitoring
the overall health of the
cluster and the infrastructure
it runs on**



**Kafka DevOps / App Dev
Team Concerns**

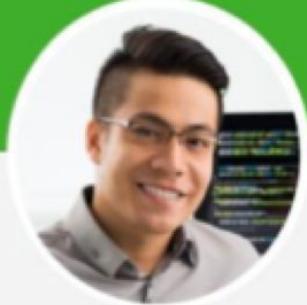
**Concerned with monitoring
the Kafka entities associated
with their apps**



**Governance / Security
Team Concerns**

**Concerned with audit,
compliance, access control &
chain of custody requirements**

SMM for the Platform Ops Team



Kafka Platform Ops
Team Concerns

Do I have any offline topic partitions?

Which consumer group is falling behind the most?

Are any of my brokers down?

Which producers are generating the most data right now?

What is the throughput in/out for a given partition on that broker?

What hosts are my brokers located on?

Are all my replicas in my topic in-synch?

How many total active producers/consumers is there currently?

Which producers are generating the most data right now?

Are there any skewed partitions for a broker?

How many total active producers and consumers exist now?

Are any of my brokers running hot? Which broker has the highest throughput in and out rates?

How many total topics does my Kafka cluster have?

Which producers are generating the most data right now?

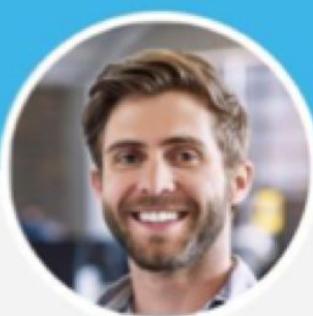
Are any of my brokers running out of disk space?

How of the cluster is being used, how much capacity do I have available per broker / cluster?

Which partitions are located on each broker?

Which of my topics has produced/consumers the most messages over the last N minutes/hours?

SMM for the DevOps/AppDev Teams



Kafka DevOps / App Dev
Team Concerns

Find all entities (producer, consumers, topics) associated with my app.

What brokers holds the partitions for my app topics?

What is the total number of messages into my topic over the last N minutes/hours?

Are there consumers in a consumer-group for a given topic slow/falling behind?

What topic(s) are the consumer group consuming messages from?

What is the retention rate for my app topics?

Who are all the producers and consumers connected to my app topics?

Did a consumer rebalance occur for a given topic?

How many active consumers instances are in a given consumer group?

Are any of my consumers/consumer-groups that are under-consuming?

What is the replication factor for my app topics?

What type of events are in my application topics? What does the event look like?

Are any of my consumers/consumer-groups that are over-consuming?

SMM for the Security and Governance Team



Governance / Security
Team Concerns

When was a topic created?

How has the schema evolved for a given topic?

Which consumers have consumed from a topic?

Which users/groups/service accounts have read from a given topic?

When was the topic configuration last modified?

What is the schema for a given topic?

How does data flow across multiple kafka hops?

What are the ACL policies for a given topic?

Which users/groups/service accounts have sent data to a topic?

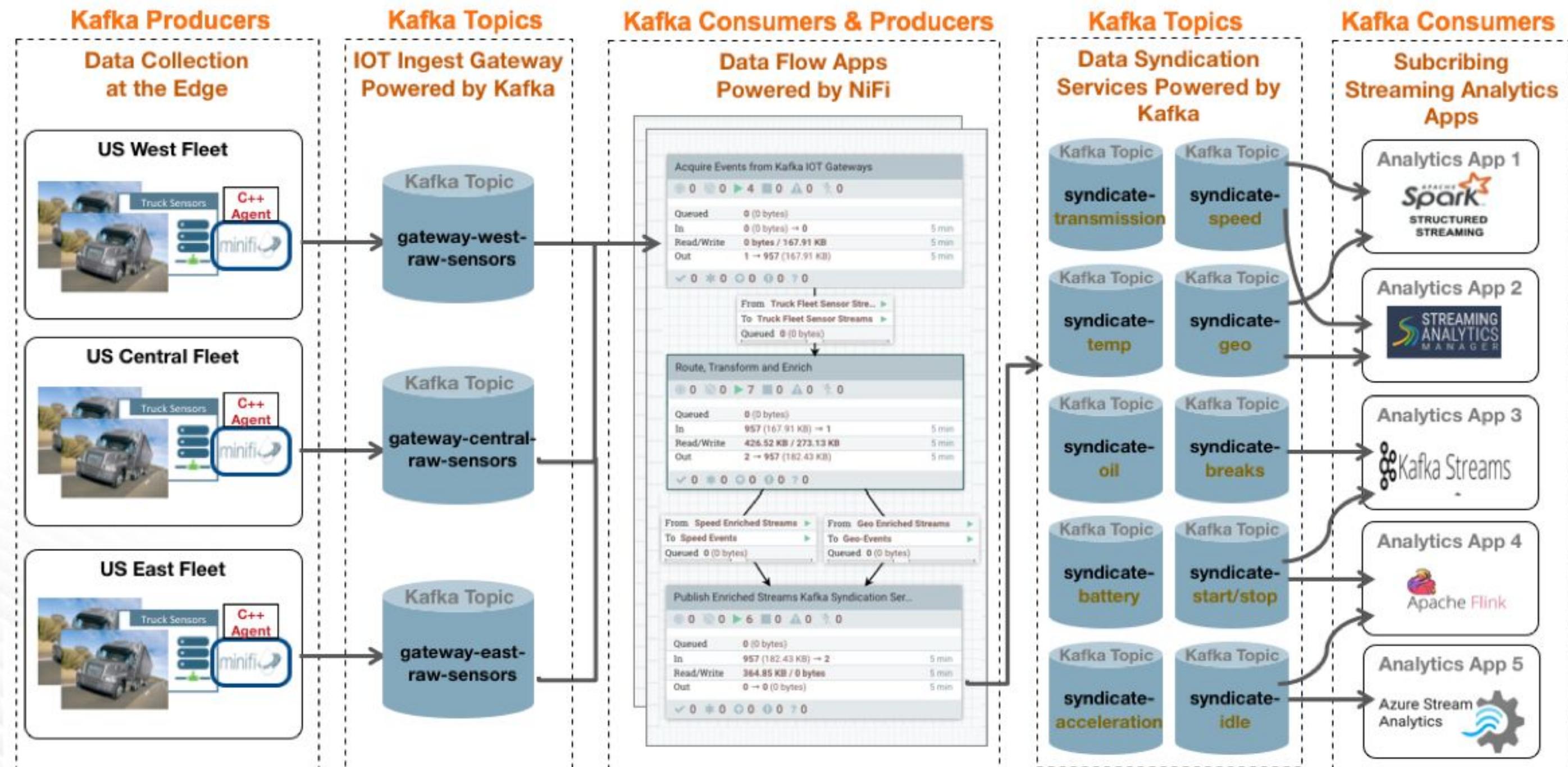
Which producers have sent data to the topic?

What is the lineage of a kafka topic?

Who has edited the ACL policies of a given topic?

When were additional brokers added to the topic?

Demo Setup: Dev Ops / App Dev Persona – Monitoring the Streaming Truck App



Kafka Challenges

Shared Schema Registry

How do I associate schemas for messages in Kafka topics

Agility and Self-Service

How do I develop, deploy & manage Kafka producers and consumers without code in self service manner

Kafka Blindness

How do I manage/monitor the different kafka clusters, topics, consumers, and producers

Data Collection

Kafka as a Service

I want to provide my users self service capabilities for Kafka: creation of clusters, monitoring, management



Kafka Topics

IOT Ingest Gateway Powered by Kafka

Kafka Topic gateway-west-raw-sensors

US Central Fleet



Balanced Kafka at Scale

As my clusters grow larger, I need the system to detect and automate balancing policies

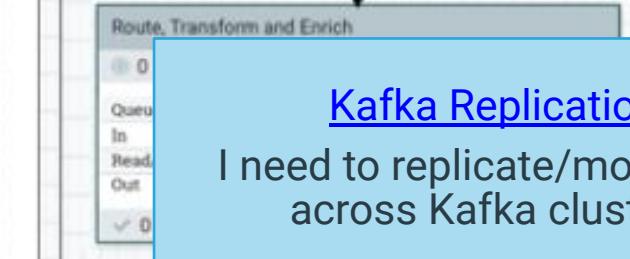


Kafka Topic gateway-central-raw-sensors

Kafka Topic gateway-east-raw-sensors

Consumers & Producers

Data Flow App Powered by NiF



Kafka Replication

I need to replicate/move data across Kafka clusters

Kafka Topics

Data Syndication Services Powered by Kafka

Kafka Topic syndicate-transmission

Kafka Topic syndicate-speed

Kafka Topic syndicate-temp

Kafka Topic syndicate-geo

Kafka Topic syndicate-oil

Kafka Topic syndicate-breaks

Kafka Topic syndicate-battery

Kafka Topic syndicate-start/stop

Kafka Topic syndicate-acceleration

Kafka Topic syndicate-idle

Kafka Consumers

Subscribing Streaming Analytics Apps

Analytics App 1



Analytics App 2



Analytics App 3



Analytics App 4



Analytics App 5



SQL/Access Patterns on Kafka

I want to treat Kafka as a table and execute SQL on it for ETL and analytics

Tutorial - Log In Instructions

EDGE2AI Workshop Home

Go to this website: **http://**

Logout



Cloudera Manager <http://54.190.22.232:7180/>

Edge Flow <http://54.190.22.232:10080/efm/ui/>

NiFi <http://54.190.22.232:8080/nifi/>

NiFi Registry <http://54.190.22.232:18080/nifi-registry/>

Schema Registry <http://54.190.22.232:7788/>

SMM <http://54.190.22.232:9991/>

Hue <http://54.190.22.232:8888/>

Cloudera Data Science Workbench <http://cdsw.54.190.22.232.nip.io/>

SSH Connection

Download key:

Download SSH Key

And then run:

`chmod 400 workshop.pem`

`ssh -i workshop.pem centos@54.190.22.232`

Main Dashboard View

STREAMS MESSAGING MANAGER

Overview

Cluster: OneNodeCluster

Producers 29 Brokers 1 Topics 40 Consumer Groups 16 Clear

TOPICS (40) BROKERS (1)

The Kafka cluster called OneNodeCluster is selected

30 minutes

Consumer Groups (16)

NAME	DATA IN	DATA OUT	MESSAGES	CONSUMER GROUPS
syndicate-tra...	783 KB	0B	3.5k	0
syndicate-sp...	0B	0B	0	0
syndicate-sp...	0B	0B	0	0
syndicate-oil	940 KB	0B	4.3k	0
syndicate-ge...	0B	0B	0	0
syndicate-ge...	0B	0B	0	3
syndicate-bat...	854 KB	0B	3.9k	0
load-optimiz...	1.3m			
fuel-micro-se...	0.6m			
supply-chain...	0.4m			
predictive-mi...	0.3m			
energy-micro...	0.3m			
audit-micro-...	0.2m			
compliance-...	0.2m			
adjudication-...	0.2m			
approval-mic...	0.1m			
nifi-truck-sen...	2			
nifi-truck-sen...	2			
flink-analytic...	0			
kafka-strea...	0			
spark-strea...	0			

Producers (29)

ACTIVE ...	PASSIV ...	ALL
MESSAGES		
geo-critical-e...	89k	
geo-critical-e...	45k	
minifi-eu-i1	45k	
load-optimiz...	42k	
geo-critical-e...	30k	
minifi-eu-i2	23k	
geo-critical-e...	23k	
fuel-apps	21k	
geo-critical-e...	18k	
minifi-eu-i3	15k	
supply-chain...	14k	
predictive-ap...	11k	
energy-apps	8.5k	
audit-apps	7.1k	
compliance-...	6.1k	

Find the Most Active Producer in my Cluster

The screenshot shows the Streams Messaging Manager interface with the following details:

- Overview:** The main dashboard displays cluster statistics: 1 Broker, 40 Topics, and 16 Consumer Groups.
- Producer Activity:** A callout box highlights the "MESSAGES" sorting option in the Producers section, stating: "Click on Messages to sort on messages sent by all producers in the last 30 mins".
- Most Active Producer:** A callout box on the left side of the Producers table highlights the "geo-critical-event-collector-i1" producer, noting: "A Kafka producer called geo-critical-event-collector-i1 is the most active producer sending 89K messages in the last 30 mins".
- Producer Data:** The table lists 29 producers, including:
 - syndicate-tra... (783 KB, 0B, 3.5k, 0)
 - syndicate-sp... (0B, 0B, 0, 0)
 - syndicate-sp... (0B, 0B, 0, 0)
 - syndicate-oil (940 KB, 0B, 4.3k, 0)
 - syndicate-ge... (0B, 0B, 0, 0)
 - syndicate-ge... (0B, 0B, 0, 3)
 - syndicate-bat... (854 KB, 0B, 3.9k, 0)
- Consumer Groups:** A sidebar shows 16 consumer groups with their respective lag times.

Find the Consumer Who Has Fallen Behind the Most

The screenshot shows the Streams Messaging Manager interface with the following details:

- Overview** section with counts: Producers (29), Brokers (1), Topics (40).
- Topics** tab selected.
- Producers** table:
 - syndicate-tra... (783 KB, 0B, 3.5k, 0)
 - syndicate-sp... (0B, 0B, 0, 0)
 - syndicate-sp... (0B, 0B, 0, 0)
 - syndicate-oil (940 KB, 0B, 4.3k, 0)
 - syndicate-ge... (0B, 0B, 0, 0)
 - syndicate-ge... (0B, 0B, 0, 3)
 - syndicate-bat... (854 KB, 0B, 3.9k, 0)
- Consumer Groups** table:
 - load-optimiz... (1.3m lag)
 - fuel-micro-se... (0.6m lag)
 - supply-chain... (0.4m lag)
 - predictive-mi... (0.3m lag)
 - energy-micro... (0.3m lag)
 - audit-micro... (0.2m lag)
 - compliance-... (0.2m lag)
 - adjudication-... (0.2m lag)
 - approval-mic... (0.1m lag)
 - nifi-truck-sen... (2 lags)
 - nifi-truck-sen... (2 lags)
 - flink-analytic... (0 lags)
 - kafka-strea... (0 lags)
 - spark-strea... (0 lags)
- A callout box points to the **Consumer Groups** table with the text: "Click on LAG to sort on consumer lag across all consumers in the last 30 mins".
- A callout box points to the **load-optimiz...** entry in the **Consumer Groups** table with the text: "Consumer group named load-optimizer-micro-service has significantly more lag (1.3m) than any other consumer in the cluster."

Broker Centric View – View Details of the Brokers in My Cluster

The screenshot shows the Streams Messaging Manager Overview dashboard. At the top, there are four main metrics: Producers (29), Brokers (1), Topics (40), and Consumer Groups (16). A green callout box highlights the 'Brokers' tab, which is currently selected. Below the metrics, there are sections for Producers, Topics, and Consumer Groups. The 'Producers' section lists 29 active producers across various topics like geo-critical-e... and minifi-eu-i1. The 'Topics' section shows 40 topics with their respective partitions and replicas. The 'Consumer Groups' section lists 16 consumer groups with their lag times. On the left sidebar, there are navigation links for Overview, Brokers, Topics, Producers, Consumer Groups, and Alerts.

Cluster: OneNodeCluster

Overview

Cluster: OneNodeCluster

Producers 29

TOPICS (40) BROKERS (1)

Click on the Brokers tab to see a broker centric view of the Dashboard

Brokers 1

Topics 40

Consumer Groups 16

Clear

30 minutes

Producers (29)

ACTIVE PASSIV... ALL

MESSAGES

geo-critical-e... 89k

geo-critical-e... 45k

minifi-eu-i1 45k

load-optimiz... 42k

geo-critical-e... 30k

minifi-eu-i2 23k

geo-critical-e... 23k

fuel-apps 21k

geo-critical-e... 18k

minifi-eu-i3 15k

supply-chain... 14k

predictive-ap... 11k

energy-apps 8.5k

audit-apps 7.1k

compliance-... 6.1k

IP INPUTS MESSAGES PARTITIONS REPLICAS

8 ip-10-0-1-248.us-west-... 93 MB 0.4m 194 194 L C

Consumer Groups (16)

ACTIVE PASSIV... ALL

LAG

load-optimiz... 1.3m

fuel-micro-se... 0.6m

supply-chain... 0.4m

predictive-mi... 0.3m

energy-micro... 0.3m

audit-micro... 0.2m

compliance-... 0.2m

adjudication-... 0.2m

approval-mic... 0.1m

nifi-truck-sen... 2

nifi-truck-sen... 2

flink-analytic... 0

kafka-strea... 0

spark-strea... 0

Broker Centric View: Find my Hottest Broker – Broker with Highest Throughput In

Step 1
Click on the Brokers tab to see a broker centric view of the Dashboard

Step 2
Click on Throughput to sort on data in across all brokers

NAME	THROUGHPUT	MESSAGES IN	PARTITIONS	REPLICAS
1001 c-dps-connected-dp13.field.hortonwor...	17MB	80k	21	36
1002 c-dps-connected-dp12.field.hortonwor...	16MB	75k	16	34
1005 c-dps-connected-dp11.field.hortonwor...	14MB	63k	15	31
1003 c-dps-connected-dp14.field.hortonwor...	10MB	42k	16	30
1004 c-dps-connected-dp15.field.hortonwor...	7MB	33k	14	30
geo-critical-event-collec...	7.7k			
geo-critical-event-collec...	7.1k			
geo-critical-event-collec...	6.5k			
minifi-eu-i6	6.5k			
audit-apps	6k			
geo-critical-event-collec...	6k			
geo-critical-event-collec...	5.6k			
minifi-eu-i7	5.5k			
compliance-apps	5.2k			
minifi-eu-i8	4.8k			
geo-critical-event-collec...	4.8k			
geo-critical-event-collec...	4.6k			
adjudication-apps	4.5k			
minifi-eu-i9	4.3k			
approval-apps	4k			
minifi-eu-i10	3.9k			

Consumer Groups (26)

ACTIVE (18) PASSIVE (8) ALL

route-micro-service	0
load-optimizer-micro-se...	5
fuel-micro-service	2
supply-chain-micro-serv...	1
predictive-micro-service	1.3k
energy-micro-service	984
audit-micro-service	812
compliance-micro-servi...	698
adjudication-micro-servi...	599
approval-micro-service	542
flink-analytics-geo-event	224
kafka-streams-analytics...	224
spark-streaming-analyti...	224
nifi-truck-sensors-east	4
nifi-truck-sensors-west	2
nifi-truck-sensors-central	1
ranger_entities_consum...	1
atlas	0

This is a multi-node cluster. What you are working with is a single node cluster.

Find Partitions on a given Broker and Understand flow of data flow from Producer to selected Broker Partition to Consumer

The screenshot shows the Streams Messaging Manager interface with the following details:

- Overview:** Producers 1 of 83, Brokers 3 of 5, Topics 1 of 27.
- Producers:** 1 active producer named "nifi-syndicate-speed-avro" with 1.6k messages.
- Brokers:** 3 brokers listed, with broker 1001 highlighted.
- Topics:** 1 topic named "syndicate-speed-event-avro" with 21 partitions and 36 replicas.
- Consumer Groups:** 3 consumer groups: "sam-speed-stream-consum...", "sam-speed-stream-consum...", and "sam-speed-stream-consum...".

Three callouts provide instructions:

- Step 1:** Click on Panel expand to get more details on the broker like all partitions that are stored on the broker.
- Step 2 - Analysis:** Note that partition 0 of topic syndicate-speed has high throughput-out on that partition.
- Step 3:** Click on the partition and see who are all the producers and consumers sending/consuming from that partition. There is 1 producer and 3 consumer groups explaining why the high throughput out vs in.

Detailed description of the highlighted broker panel:

- Topic:** syndicate-speed-event-avro - P0
- Data In:** 93385
- Data Out:** 14260475
- Profile Filter:** EXPLORE

Partition	Throughput	Messages In	Messages Out
P0	14MB out	91KB in	14MB out
P1	28KB in	28KB in	0B out
P2	72KB in	153KB in	0B out

Analyze Detailed Broker Host Metrics – Cloudera Manager Integration

Overview

Cluster: OneNodeCluster

TOPICS (5) BROKERS (1)

Producers (12)

ACTIVE (1) PASSIVE (0) ALL

MESSAGES

	minifi-eu-i1	minifi-eu-i2	minifi-eu-i3	minifi-truck-w1	minifi-truck-w2	minifi-truck-w3	minifi-truck-c1	minifi-truck-c2
45k	23k	15k	798	660	568	496	442	

Brokers (1 of 1)

Topics 5 of 40

Consumer Groups 3 of 16

30 minutes

NAME THROUGHPUT MESSAGES IN PARTITIONS REPLICAS

8 ip-10-0-1-248.us-west-2.compute... 93 MB 0.4m 194 194

Consumer Groups ACTIVE (2) PASSIVE (0) ALL LAG

nifi-truck-sensors- 2

nifi-truck-sensors- 2

FREE MEMORY FREE DISK CPU IDLE 14.98 LOAD AVERAGE 2.07

DISK I/O 1059780.20

syndicate-tr... P0 202 KB in 0B out

syndicate-tr... P1 68 KB in 0B out

syndicate-tr... P2 271 KB in 0B out

Cloudera Manager Clusters Hosts Diagnostics Audits Charts Backup Administration

OneNodeCluster / Kafka / ip-10-0-1-248

! Kafka Broker (id: 8) (Active Controller)

Actions

30 minutes preceding S

Status Configuration Processes Commands Charts Library Audits Log Files Stacks Logs Quick Links

Health Tests Create Trigger

Host Health Suppress...
The health of this role's host is bad. The following health tests are bad:
agent parcel directory.

Show 7 Good

Log Directory Free Space Suppress...
This role has no Log Directory configured.

Charts

30m 1h 2

Messages Received @

messages / second

KAFKA_BROKER (ip-10-0-1-248.us-west-2.compute...) 247

Bytes Received @

bytes / second

KAFKA_BROKER (ip-10-0-1-248.us-west-2.compute...) 54.1K/s

Bytes Fetched @

bytes / second

KAFKA_BROKER (ip-10-0-1-248.us-west-2.compute...) 427b/s

Partitions @

partitions

KAFKA_BROKER (ip-10-0-1-248.us-west-2.compute...) 194

Health History

> ! Host Health Bad 10:41 PM

> ● Host Health Concerning 9:40 PM

> ● 3 Became Good Sep 20 11:14 PM

> ● 3 Became Good Sep 20 11:11 PM

Leader Replicas @

Offline Partitions @

Click on the CM icon on the broker panel and the CM host detail view for that broker is displayed providing host level metrics and a view of other services running on that host

The screenshot shows the Cloudera Manager interface for a Kafka broker. At the top, there are navigation tabs for Producers, Brokers, Topics, and Consumer Groups. The 'Brokers' tab is selected, showing one broker named '8' (ip-10-0-1-248.us-west-2.compute...). Below the broker list are resource usage metrics: Free Memory, Free Disk, CPU Idle, and Load Average. A green callout box points to the broker row, stating: 'Click on the CM icon on the broker panel and the CM host detail view for that broker is displayed providing host level metrics and a view of other services running on that host'. The main content area displays the 'Kafka Broker (id: 8) (Active Controller)' page. It includes sections for Health Tests (with one critical error about host health), Charts (for messages received, bytes received, bytes fetched, partitions, and leader replicas), and a Health History log. The bottom left corner contains a copyright notice for Cloudera Inc. 2011–2016.

Keyword Search via Log Search

Overview

The screenshot shows the Cloudera Manager interface for a cluster named "OneNodeCluster". The top navigation bar includes tabs for "Clusters", "Hosts", "Diagnostics", "Audits", "Charts", "Backup", and "Administration". The "Diagnostics" tab is currently selected.

The main dashboard displays metrics for Producers (12 of 29), Brokers (1 of 1), Topics (5 of 40), and Consumer Groups (3 of 16). A green callout box points to the "Log Search" icon on the Broker panel, which is highlighted with a green border. The callout text reads: "Click on the Log Search icon on the broker panel and the Log Search detail view is displayed. This enables you to search for specific keywords and to filter for specific log levels, components, and time ranges."

The "Logs" section on the right allows users to search for keywords, filter by sources (Services, Cloudera Manager Agent, Cloudera Manager Server), services (Kafka), hosts (ip-10-0-1-248.us-west-2.compute.internal), role types (Kafka Broker), minimum log level (WARN), and timeout (sec). The search results show three log entries from the Kafka API:

Hosts	Log Level	Time	Source	Message
ip-10-0-1-248.us-west-2.compute.internal	ERROR	September 22, 2019 12:14 AM	KafkaApis	[KafkaApi-8] Number of alive brokers '1' does not meet the required replication factor '3' for the transactions state topic (configured via 'transaction.state.log.replication.factor'). This error can be ignored if the cluster is starting up and not all brokers are up yet. View Log File
ip-10-0-1-248.us-west-2.compute.internal	ERROR	September 22, 2019 12:14 AM	KafkaApis	[KafkaApi-8] Number of alive brokers '1' does not meet the required replication factor '3' for the transactions state topic (configured via 'transaction.state.log.replication.factor'). This error can be ignored if the cluster is starting up and not all brokers are up yet. View Log File
ip-10-0-1-248.us-west-2.compute.internal	ERROR	September 22, 2019 12:14 AM	KafkaApis	[KafkaApi-8] Number of alive brokers '1' does not meet the required replication factor '3' for the transactions state topic (configured via 'transaction.state.log.replication.factor'). This error can be ignored if the cluster is starting up and not all brokers are up yet. View Log File

At the bottom left, there are copyright notices: "© Cloudera Inc. 2011 – 2016. All Rights Reserved" and page numbers "2" and "9".

SMM DevOps/App Dev Use Cases

Topic Centric Dashboard View: Filter on Topics associated with my Topic

STREAMS MESSAGING MANAGER

Overview Cluster: orlandostreamcluster

Producers: 83 | Brokers: 5 | Topics: 27 | Consumer Groups: 0 minutes

TOPICS (27) BROKERS (5)

Producers (83) ACTIVE (83) PASSIVE (0) ALL

	NAME	DATA IN	DATA OUT	MESSAGES IN	CONSUMER GROUPS
<input checked="" type="checkbox"/>	syndicate-all-geo-critical-eve...	26MB	0B		
<input checked="" type="checkbox"/>	route-planning	26MB	28KB		
<input checked="" type="checkbox"/>	gateway-europe-raw-sensors	18MB	0B	88k	0
<input checked="" type="checkbox"/>	load-optimization	5MB	28KB	23k	1
<input checked="" type="checkbox"/>	fuel-logistics	1MB	28KB	6.5k	1
<input checked="" type="checkbox"/>	supply-chain	863KB	28KB	4.3k	1
<input checked="" type="checkbox"/>	audit-events	804KB	27KB	3.9k	1
<input checked="" type="checkbox"/>	compliance	697KB	29KB	3.4k	1
<input checked="" type="checkbox"/>	predictive-alerts	648KB	28KB	3.3k	1
	route-apps	0.1m			
	minifi-eu-i1	25k			
	load-optimizer-apps	23k			
	geo-critical-event-collector-i1	23k			
	minifi-eu-i2	12k			
	geo-critical-event-collector-i2	11k			
	geo-critical-event-collector-i3	7.7k			
	minifi-eu-i4	6.3k			
	geo-critical-event-collector-i8	6.3k			
	geo-critical-event-collector-i4	5.8k			
	fuel-apps	5.3k			
	geo-critical-event-collector-i10	5k			
	geo-critical-event-collector-i5	4.6k			
	geo-critical-event-collector-i11	4.6k			
	minifi-eu-i6	4.2k			
	geo-critical-event-collector-i12	4.2k			
	audit-apps	3.9k			
	geo-critical-event-collector-i13	3.9k			
	minifi-eu-i3	3.8k			
	geo-critical-event-collector-i6	3.8k			

Consumer Groups ACTIVE (18) PASSIVE (8) ALL

	NAME	LAG
	route-micro-service	0.2m
	load-optimizer-micro-service	12k
	fuel-micro-service	6k
	supply-chain-micro-service	4k
	predictive-micro-service	3k
	energy-micro-service	2.3k
	audit-micro-service	1.9k
	compliance-micro-service	1.7k
	adjudication-micro-service	1.4k
	approval-micro-service	1.3k
	flink-analytics-geo-event	525
	kafka-streams-analytics-geo...	525
	spark-streaming-analytics-g...	525
	nifi-truck-sensors-west	2
	nifi-truck-sensors-east	2
	nifi-truck-sensors-central	1
	ranger_entities_consumer	1
	atlas	0

gatew

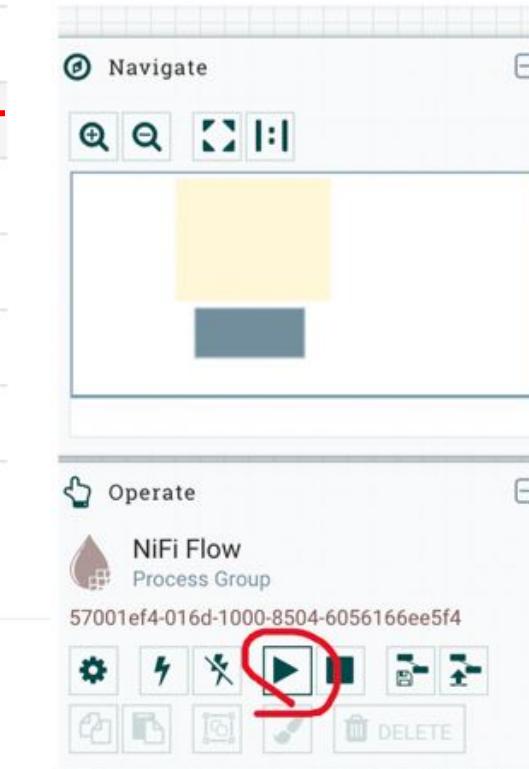
Use the Filter to filter on topics and select all the IOT gateway topics



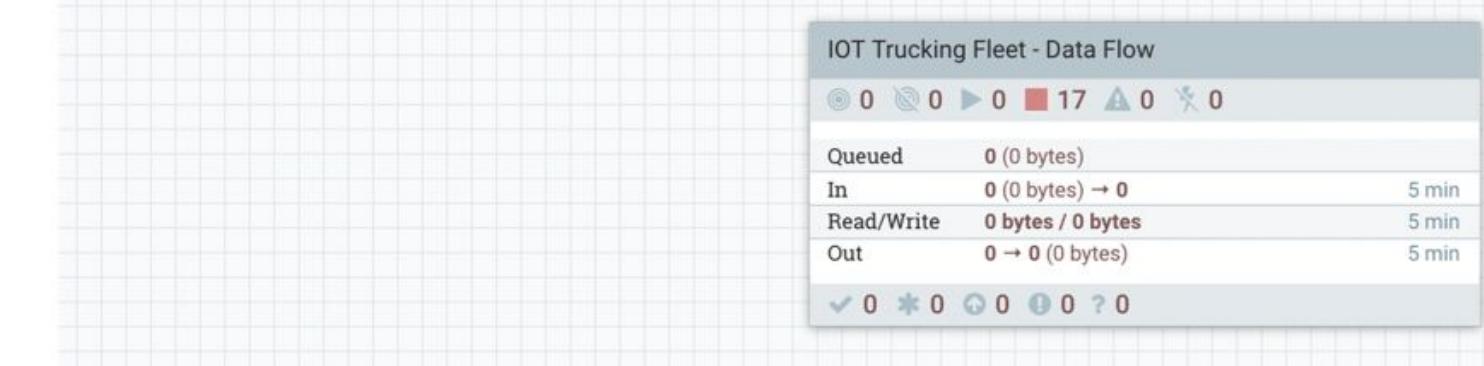
pk-strata

Cloudera Manager	http://54.185.190.193:7180/
Edge Flow	http://54.185.190.193:10080/efm/ui/
NiFi	http://54.185.190.193:8080/nifi/
NiFi Registry	http://54.185.190.193:18080/nifi-registry/
Schema Registry	http://54.185.190.193:7788/
SMM	http://54.185.190.193:9991/
Hue	http://54.185.190.193:8888/
Cloudera Data Science Workbench	http://cdsw.54.185.190.193.nip.io/

Start NiFi Flow



1. Nifi ingest Trucking CSV Events with kafka headers that contains schema name which is a pointer to schema in HWX Schema Registry..
2. Nifi extracts the schema name from kafka header and fetches schema from the HWX Schema Registry to perform record based processing including filtering, routing and enrichment
3. These CSV records are then enriched and converted into Avro Records and send to a Kafka Topic. When publishing to Kafka Topic, Nifi will look up the schema associated with the kafka topic in HWX SR and encode the Avro binary with HWX schema encoding so that SAM can work with the data.



Intelligent Filtering – Selected Topics causes Producers / Consumer to be Intelligently Filtered

Overview

Cluster: OneNodeCluster

The screenshot shows the Apache NiFi interface with several panels and annotations:

- Producers:** 12 of 29 (Topics: 5). A callout box labeled "Intelligent Filtering" states: "SMM automatically filters the producers associated with the selected topics. 12 of the 29 producers have been identified as sending data to the 5 topics selected".
- Brokers:** 1 of 1.
- Topics:** 5 of 10. A callout box labeled "User Action" states: "5 IOT Gateway topics have been selected". A modal window lists the selected topics:
 - gateway-west-raw-sensors
 - gateway-europe-raw-sensor
 - gateway-east-raw-sensors
 - gateway-east-raw-sensor
 - gateway-central-raw-sensors
- Consumer Groups:** 3 of 17. A callout box labeled "Intelligent Filtering" states: "SMM automatically filters the consumers associated with the selected topics. 3 of the 17 consumers have been identified as consuming data from the 5 topics selected". A table shows consumer groups:

Consumer Groups (3)	ACTIVE (3)	PASSIVE (0)	ALL
nifi-truck-sensors-west	2	0	LAG: 2
nifi-truck-sensors-east	0	0	LAG: 0
nifi-truck-sensors-central	0	0	LAG: 0

Find the Hottest Topic – Topic With Highest Throughput-In

Overview

Cluster: OneNodeCluster

Producers
12 of 29

Brokers
1 of 1

Topics
5 of 40

Consumer Groups
3 of 17

Clear

TOPICS (5) BROKERS (1)

Analysis

Kafka topic called gateway-europe-raw-sensors has more data being sent to it than any other topic: 88K messages totaling 16 MB in the last 30 mins

Name	Data In	Data Out
gateway-west-raw-sensors	411 KB	0B
minifi-eu-i2	23k	0B
minifi-eu-i3	15k	0B
minifi-truck-w1	788	0B
minifi-truck-w2	662	0B
minifi-truck-w3	562	0B
minifi-truck-c1	496	0B
minifi-truck-c2	442	0B
minifi-truck-c3	398	0B
minifi-truck-e1	364	0B
minifi-truck-e2	332	0B
minifi-truck-e3	304	0B

Step 1
Click on DATA IN to sort on data-in across all topics in the last 30 mins

Topics
5 of 40

Consumer Groups
3 of 17

30 minutes

Consumer Groups (3)

ACTIVE (3) PASSIVE (0) ALL LAG

Consumer Group	LAG
nifi-truck-sensors-west	2
nifi-truck-sensors-east	0
nifi-truck-sensors-central	0

How are the Partitions Laid out for the Topic? Who are the Producers and Consumers? Are there any Partition Skews?

Overview

Producers 12 of 30

TOPICS (5) BROKERS (1)

Step 2
Click on the Topic to see who are all the producers sending data to the topic

Brokers 1 of 1

Topics 5 of 40

Consumers 3 of 30

30 minutes

Consumer Groups (3)

ACTIVE (3) PASSIVE (0) ALL

LAG

nifi-truck-sensors-west 2

nifi-truck-sensors-east 0

nifi-truck-sensors-ce... 0

NAME DATA IN DATA OUT MESSAGES IN CONSUMER GROUPS

gateways-europe-raw-sensors 96 KB 0B 486 1

Topic: gateway-europe-raw-sensors - P1
DATA IN 351606
DATA OUT 0
PROFILE FILTER EXPLORE

Replication Factor: (1) InSync Replicator

minifi-eu-i1 11k

minifi-eu-i2 5.5k

minifi-eu-i3 3.7k

minifi-truck-w1 188

minifi-truck-w2 162

minifi-truck-w3 136

minifi-truck-c1 120

minifi-truck-c2 104

minifi-truck-c3 96

minifi-truck-e1 88

minifi-truck-e2 80

minifi-truck-e3 74

ALL PARTITIONS

Analysis
Note that for each partition there is no data going out (0B) and we see no data going to any consumer groups. This means that while the topic has lots of producers, there is no consumers which could indicate a problem

NAME DATA IN DATA OUT MESSAGES IN CONSUMER GROUPS

gateways-east-raw-sensors 49 KB 0B 242 1

How does Data Flow between Producers to Topics to Consumers?

The screenshot shows the Apache Kafka Cluster Overview page. The top navigation bar includes 'Overview' and 'Cluster: chicagostreamcluster'. The main interface displays 'Topics (4)' and 'Brokers 3 of 3'. A large green box labeled 'Step 1' with the instruction 'Expand details of a topic that has consumers' highlights the 'Topics' section. Another green box labeled 'Step 2' with the instruction 'Click on the topic to see all producers sending data to it and all consumers consuming from it' highlights the 'Producers' section. A central green box labeled 'Analysis 2' states: 'Note that there is no data in 2 of the 4 partitions. This could be a partition/event key skew issue'. A final green box labeled 'Analysis 1' states: 'We have 3 truck producers from the west fleet sending data to gateway-west topic and a NiFi consumer called truck-sensors-west consuming from it'.

Step 1
Expand details of a topic that has consumers

Step 2
Click on the topic to see all producers sending data to it and all consumers consuming from it

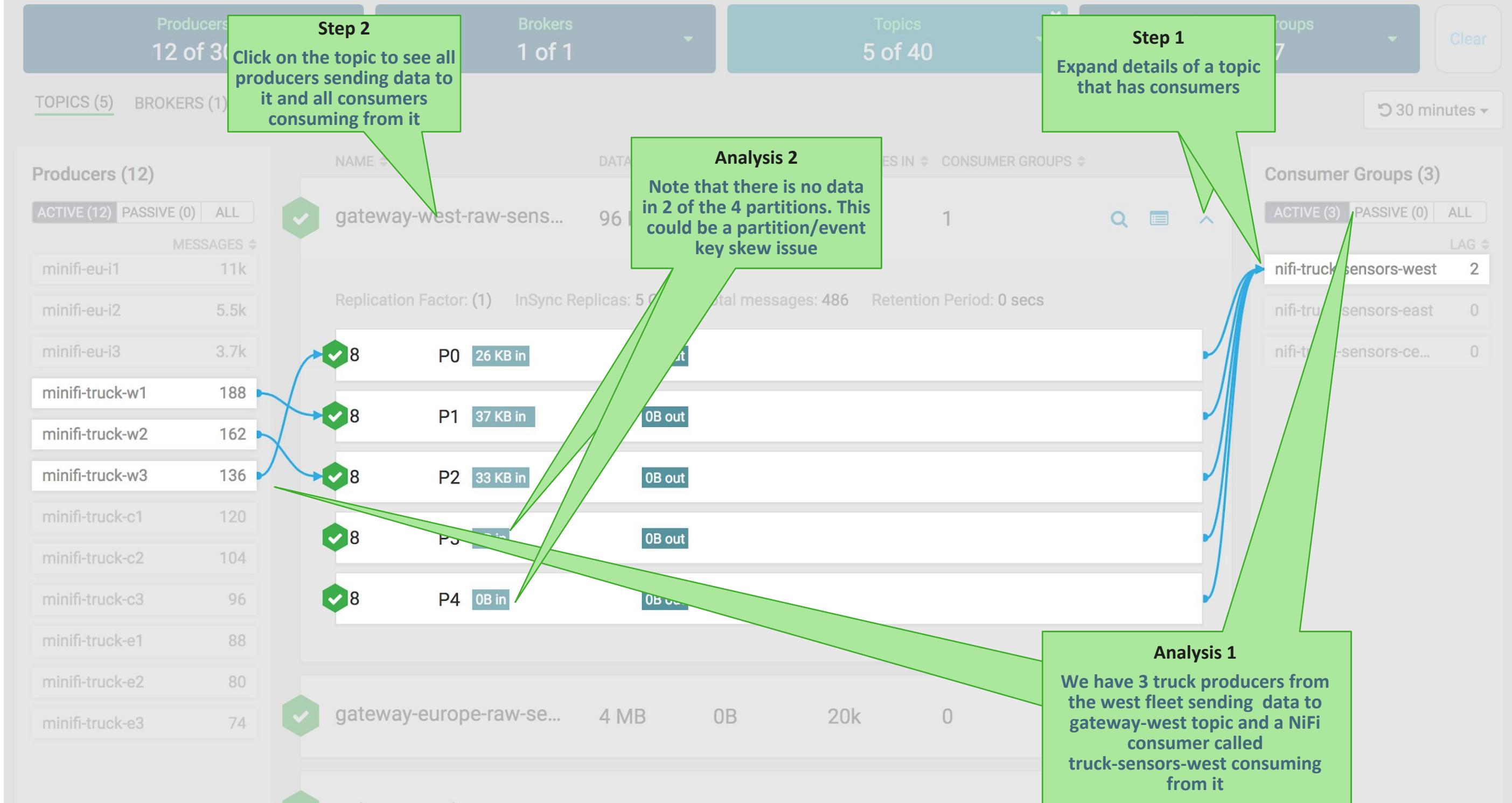
Analysis 2
Note that there is no data in 2 of the 4 partitions. This could be a partition/event key skew issue

Analysis 1
We have 3 truck producers from the west fleet sending data to gateway-west topic and a NiFi consumer called truck-sensors-west consuming from it

How does Data Flow between Producers to Topics to Consumers?

Overview

Cluster: OneNodeCluster



Explore/Search Messages in the Kafka Topic

Overview

Producers
12 of 30

TOPICS (5) BROKERS (1)

Producers (12)

ACTIVE (12) PASSIVE (0) ALL

minifi-truck-w1 188

NAME ◆ DATA IN ◆ DATA OUT ◆ MESSAGES IN ◆ CONSUMER

gateway-west-raw-sens... 96 KR 0R 486 1

Replication Factor: (1) InSync Replic

8 P0 26 KB in

8 P1 37 KB in 0B out

8 P2 33 KB in 0B out

8 P3 0B in 0B out

8 P4 0B in 0B out

Topic: gateway-west-raw-sensors - P1

DATA IN 37824

DATA OUT 0

PROFILE FILTER EXPLORE

Topics / gateway-west-raw-sensors

METRICS DATA EXPLORER CONFIG

FROM

Partition 1

TO

409

Click on the explorer icon to search for events in the Kafka Topic

Topics / gateway-west-raw-sensor

METRICS DATA EXPLORER CONFIGS LATENCY

Cluster: OneNodeCluster

DESERIALIZER:

Keys: String

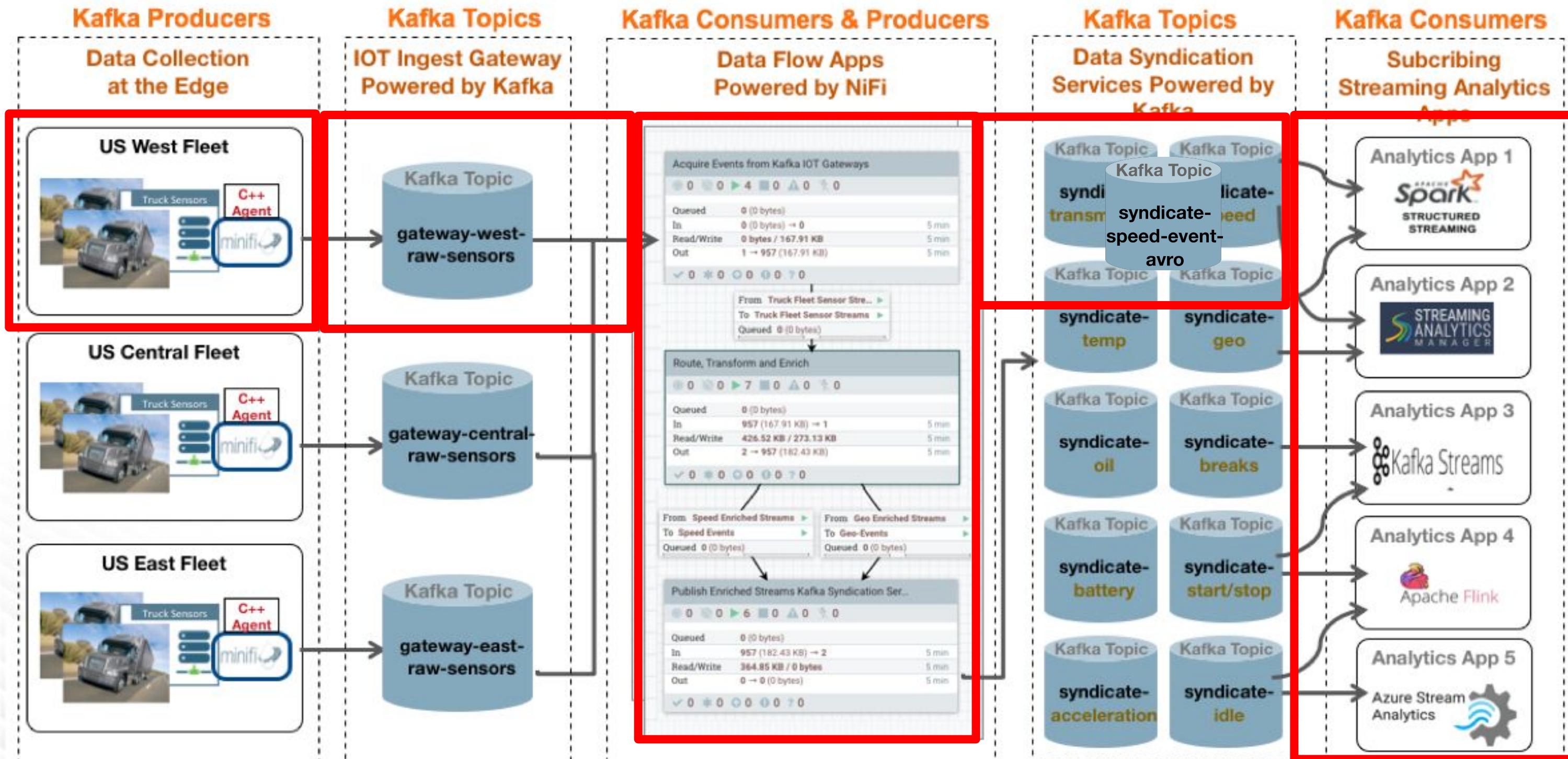
Values: String


TO OFFSET

1
424
0

Offset	Timestamp	Key	Value
409	Mon, Sep 23 2019, 9:20:35	10	2019-09-23 16:20:35.482 1569255635482 truck_speed_event 718 10 George Vetticaden 8 Saint Louis to Tulsa 66
410	Mon, Sep 23 2019, 9:20:39	10	2019-09-23 16:20:39.711 1569255639711 truck_geo_event 718 10 George Vetticaden 8 Saint Louis to Tulsa Normal 38.14 -91.3 1
411	Mon, Sep 23 2019, 9:20:39	10	2019-09-23 16:20:39.712 1569255639712 truck_speed_event 718 10 George Vetticaden 8 Saint Louis to Tulsa 58
412	Mon, Sep 23 2019, 9:20:44	10	2019-09-23 16:20:44.73 1569255644730 truck_geo_event 718 10 George Vetticaden 8 Saint Louis to Tulsa Normal 38.09 -91.44 1
413	Mon, Sep 23 2019, 9:20:44	10	2019-09-23 16:20:44.731 1569255644731 truck_speed_event 718 10 George Vetticaden 8 Saint Louis to Tulsa 72
414	Mon, Sep 23 2019, 9:20:49	10	2019-09-23 16:20:49.172 1569255649172 truck_geo_event 718 10 George Vetticaden 8 Saint Louis to Tulsa Normal 38.04 -91.55 1
415	Mon, Sep 23 2019, 9:20:49	10	2019-09-23 16:20:49.172 1569255649172 truck_speed_event 718 10 George Vetticaden 8 Saint Louis to Tulsa 69
416	Mon, Sep 23 2019, 9:20:53	10	2019-09-23 16:20:53.14 1569255653140 truck_geo_event 718 10 George Vetticaden 8 Saint Louis to Tulsa Normal 37.99 -91.69 1

Recap: What Did We Just see? Tracking the flow of data across multiple Kafka Hops with SMM & Atlas Integration-Powerful



SMM 1.2 New Features

New Features

Topic Lifecycle Management

- Create
- Update
- Delete

Alerting

- Alert Notifier
- Alert Policy

Schema Registry Integration

- Data Governance
 - Provide reusable schema (centralized registry)
 - Define relationship between schemas (version management)
 - Enable generic format conversion, and generic routing (schema validation)
- Operational Efficiency
 - To avoid attaching schema to every piece of data (centralized registry)
 - Consumers and producers can evolve at different rates (version management)
 - Data quality (schema validation)

LESS OF THIS

```
[cloudbreak@ip-10-0-1-199 ~]$ kafka-topics.sh --list --zookeeper ip-10-0-1-115.eu-west-1.compute.internal:2181
ATLAS_ENTITIES
ATLAS_HOOK
.consumer_offsets
.smm_alert_notifications
.transaction_state
adjudication
alerts-speeding-drivers
approval
audit-events
compliance
energy-mgmt
fleet-supply-chain
fuel-logistics
gateway-central-raw-sensors
gateway-east-raw-sensor
gateway-east-raw-sensors
gateway-europe-raw-sensors
gateway-west-raw-sensors
gdeleon-test
load-optimization
myFirstTopic
nifi-kafka-demo
predictive-alerts
route-planning
supply-chain
syndicate-all-geo-critical-events
syndicate-battery
syndicate-geo-event-avro
syndicate-geo-event-json
syndicate-oil
syndicate-speed-event-avro
syndicate-speed-event-json
syndicate-transmission
[cloudbreak@ip-10-0-1-199 ~]$ pwd
"/tmp/cloudbreak"
```

```
Last login: Tue Mar 19 00:42:54 2019 from 195.53.52.170
cat /etc/motd
=====
* : 

[cloudbreak@ip-10-0-1-199 ~]$ export PATH=$PATH:/usr/hdp/current/kafka-broker/bin/
[cloudbreak@ip-10-0-1-199 ~]$ kafka-configs.sh --zookeeper ip-10-0-1-136.eu-west-1.compute.internal:2181
escribe --entity-name testKnoxSetUp
Configs for topic 'testKnoxSetUp' are cleanup.policy=compact,delete
```

```
[cloudbreak@ip-10-0-1-199 ~]$ kafka-topics.sh --zookeeper ip-10-0-1-136.eu-west-1.compute.internal:2181
tion-factor 1 --topic topic-command-line
Created topic "topic-command-line".
[cloudbreak@ip-10-0-1-199 ~]$
```

Topic Lifecycle Management

The screenshot shows the Streams Messaging Manager interface. On the left, there's a sidebar with navigation links: Overview, Brokers, Topics (which is selected and highlighted in grey), Producers, Consumer Groups, Alerts, and Replication. The main area has a title bar 'Topics' and a sub-header 'Cluster: orlandostreamcluster'. Below this is a summary section with metrics: Under Replicated Offline Partitions (0/0), a search bar, and a button labeled 'Add New'. The main content area is titled 'Add Topic' and contains fields for 'TOPIC NAME' (set to 'syndicate-geo-event-json-2') and 'PARTITIONS' (set to '3'). It also includes sections for 'Availability' (with five options: Maximum, High, Moderate, Low, Custom, where Maximum is selected), 'Replication Factor' (with four rows: Factor 3, Factor 3, Factor 2, Factor 1), and 'Limits' (with a dropdown set to 'compact'). At the bottom are 'Advanced' and 'Save' buttons.

Add Topic
User friendly UI to create new topics. Simple and Advance features are available

Topic Update

The screenshot shows the Cloudera Manager interface for managing topics. On the left, there's a sidebar with various icons. The main area displays a summary of cluster metrics (Total Bytes In: 64MB, Total Bytes Out: 2MB, Produced Per Sec: 243, Fetched Per Sec: 705) and a list of topics (35 total). A green callout box points to the search bar at the top of the topic list, containing the text: "Search Topic Search the Topic you would like to update. Then click on profile". Another green callout box points to the "CONFIGS" tab in the detailed topic configuration dialog, containing the text: "Update Topic Click on Config and you can change Cleanup Policy or click Advanced to modify the configuration parameters." The detailed configuration dialog is open for the topic "testKnoxSetUp". It includes tabs for METRICS, DATA EXPLORER, and CONFIGS. The CONFIGS tab shows various configuration parameters:

Name	Value
compression.type	producer
min.insync.replicas	1
segment.jitter.ms	0
cleanup.policy	delete
flush.ms	9223372036854775
segment.bytes	1073741824
retention.hours	168
message.format.version	2.0-IV1
file.delete.delay.ms	60000
max.message.bytes	1000000
min.compaction.lag.ms	0
message.timestamp.type	CreateTime
preallocate	false
min.cleanable.dirty.ratio	0.5
index.interval.bytes	4096
unclean.leader.election.enable	false
retention.bytes	-1
delete.retention.hours	24
segment.ms	604800000
message.timestamp.difference.max.ms	9223372036854775
segment.index.bytes	10485760

At the bottom right of the configuration dialog, there are "simple" and "Save" buttons.

Alerting – Create Notifier

Notifier

NAME
email_notifier

DESCRIPTION
Notifies via Email

PROVIDER
Email

FROM ADDRESS
smm.barcelona1@gmail.com

TO ADDRESS
smm.barcelona1@gmail.com

USERNAME
smm.barcelona1@gmail.com

PASSWORD

1st Key Construct of Alerts

Alert Notifier

1. Email
2. http end point
3. Kafka topic

PASSWORD
.....

SMTP HOSTNAME
smtp.gmail.com

SMTP PORT
587

ENABLE AUTH
 ENABLE SSL ENABLE STARTTLS

PROTOCOL
smtp

ENABLE DEBUG

NOTIFIER RATE LIMIT

COUNT	DURATION
2	HOUR

Alerting – Create Alert

Alert Policy

NAME
High Lag - Kafka Streams Truck Join Micro Service

DESCRIPTION
High Lag - Kafka Streams Truck joining the speed and geo streams

EXECUTION INTERVAL IN SECONDS
60

EXECUTION DELAY IN SECONDS
300

ENABLE

Policy

COMPONENT TYPE
IF... Consumer

TARGET NAME
nifi-truck-sensors-west

+

ATTRIBUTE	CONDITION	VALUE
CONSUMER GROUP LAG	<	50

+

Action

NOTIFICATION
x email_notifier

Preview

IF CONSUMER: nifi-truck-sensors-west has CONSUMER GROUP LAG > 50 THEN notify by email_notifier

Cancel Save

2nd Key Construct of Alerts

Alert Policy

1. Defined for 6 key entities (cluster, broker, topic, producer, consumer, latency, cluster replication)
2. Metrics defined on entities
3. Complex alerts
4. Includes notifier when triggered

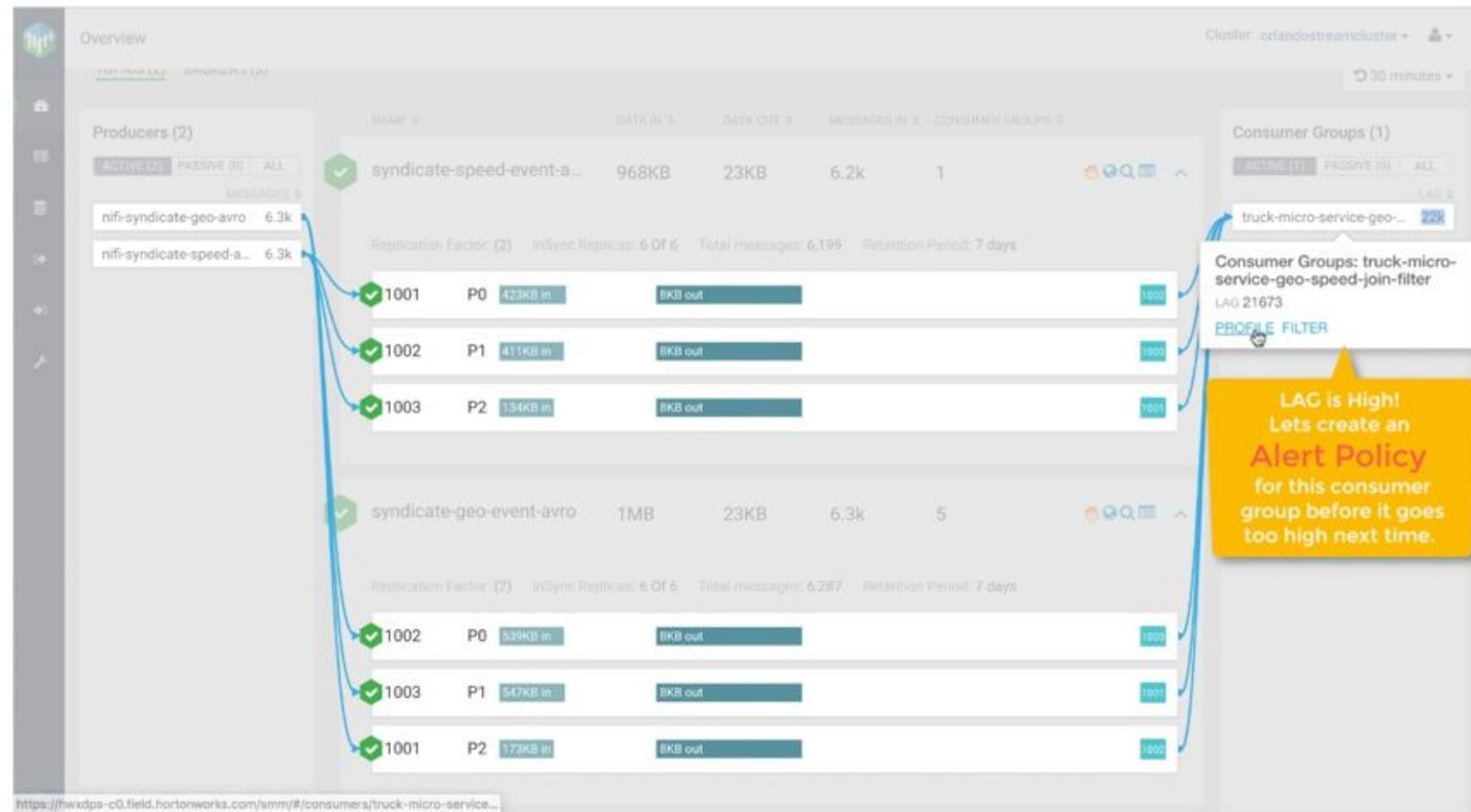
Alert History

HISTORY	ALERT POLICIES	NOTIFIERS			
Title	Timestamp	Component name	Type	State	Payload
Alert: Lag_Alert	2m 32s ago	load-optimizer-micro-service	CONSUMER	RAISED	Alert policy : 'Lag_Alert' For CONSUMER=load
Alert: Lag_Alert	7m 32s ago	load-optimizer-micro-service	CONSUMER	RAISED	Alert policy : 'Lag_Alert' For CONSUMER=load

Disable Alert

HISTORY	ALERT POLICIES	NOTIFIERS	ADD NEW
NAME	CONDITION	DESCRIPTION	ENABLE
gdeleon-alert	IF Topic: gdeleon-test has BYTES IN PER SEC >= 10	my alert	<input checked="" type="checkbox"/>
Lag_Alert	IF Consumer: load-optimizer-micro-service has CONSUMER GROUP LAG >= 10		<input type="checkbox"/>

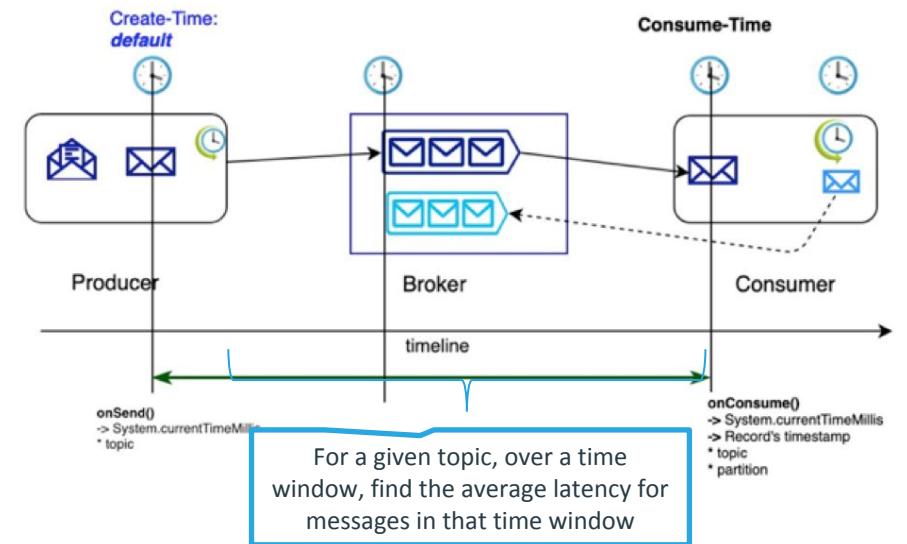
Example: Alerting on Micro-Service Consumer Group with High Lag



New DevOps Monitoring Capability with End to End Latency View

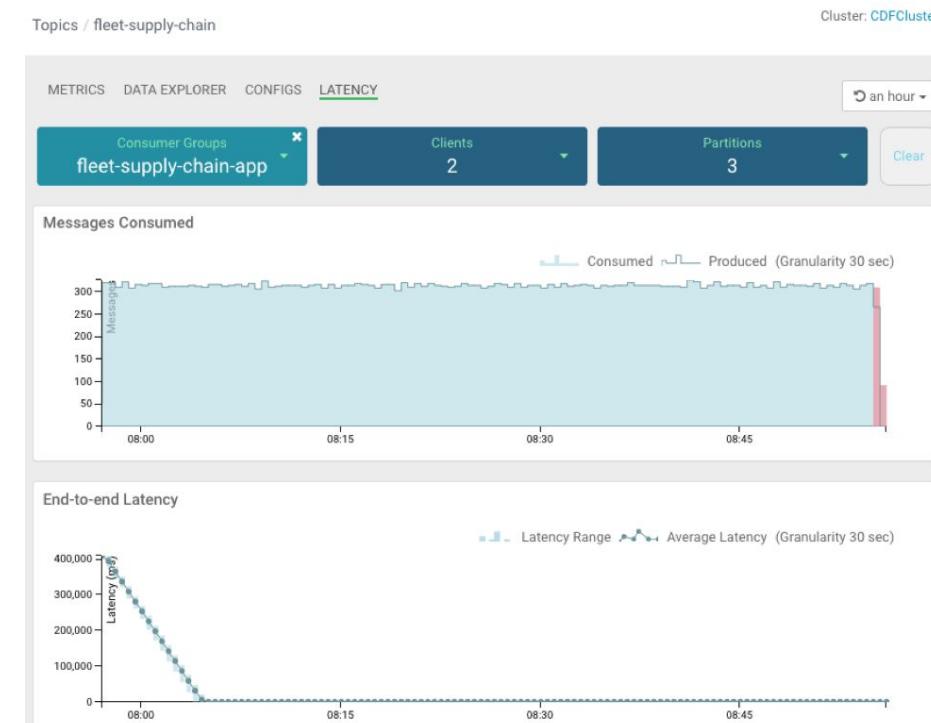
Problem Statement / Requirements

- Kafka DevOps teams need the ability to see an end to end latency view of messages produced and consumed across producers to topics/brokers to consumers
- Common DevOps questions include:
 - For the last five minutes, what is the average latency of messages consumed after being produced.
 - For the last hour, which topics have under consumption?



Solution / Benefit

- SMM 2.0 provides new monitoring view for end to end Latency
- End to End Latency view is powered by new embedded Kafka streams application that calculates end to end latencies and new interceptors that can be used to instrument producer and consumer clients.



Select a Topic of Interest

Overview Cluster: hdf_mpack

Producers: 100 | Brokers: 3 | Topics: 13 | Consumer Groups: 2 | Clear | 30 minutes

TOPICS (13) BROKERS (3)

Producers (100) Consumer Groups (2)

NAME	DATA IN	DATA OUT	MESSAGES IN	CONSUMER GROUPS
TruckEvents2	616B	2KB	1.6k	0
TruckEvents	1KB	928B	997	0

ACTIVE (100) PASSIVE (0) ALL

MESSAGES

8-5-producer-23	24k
8-5-producer-99	24k
8-5-producer-37	24k
8-5-producer-0	24k
8-5-producer-45	24k
8-5-producer-63	24k
8-5-producer-96	24k
8-5-producer-39	23k
8-5-producer-20	23k
8-5-producer-14	23k
8-5-producer-81	23k
8-5-producer-85	23k

ACTIVE (0) PASSIVE (2) ALL

LAG

ExampleTestGroup2	0
ExampleTestGroup	0

Replication Factor: (2) InSync Replicas: 8 Of 8 Total messages: 1,620 Retention Period: 0 secs

P0 P1 P2 P3

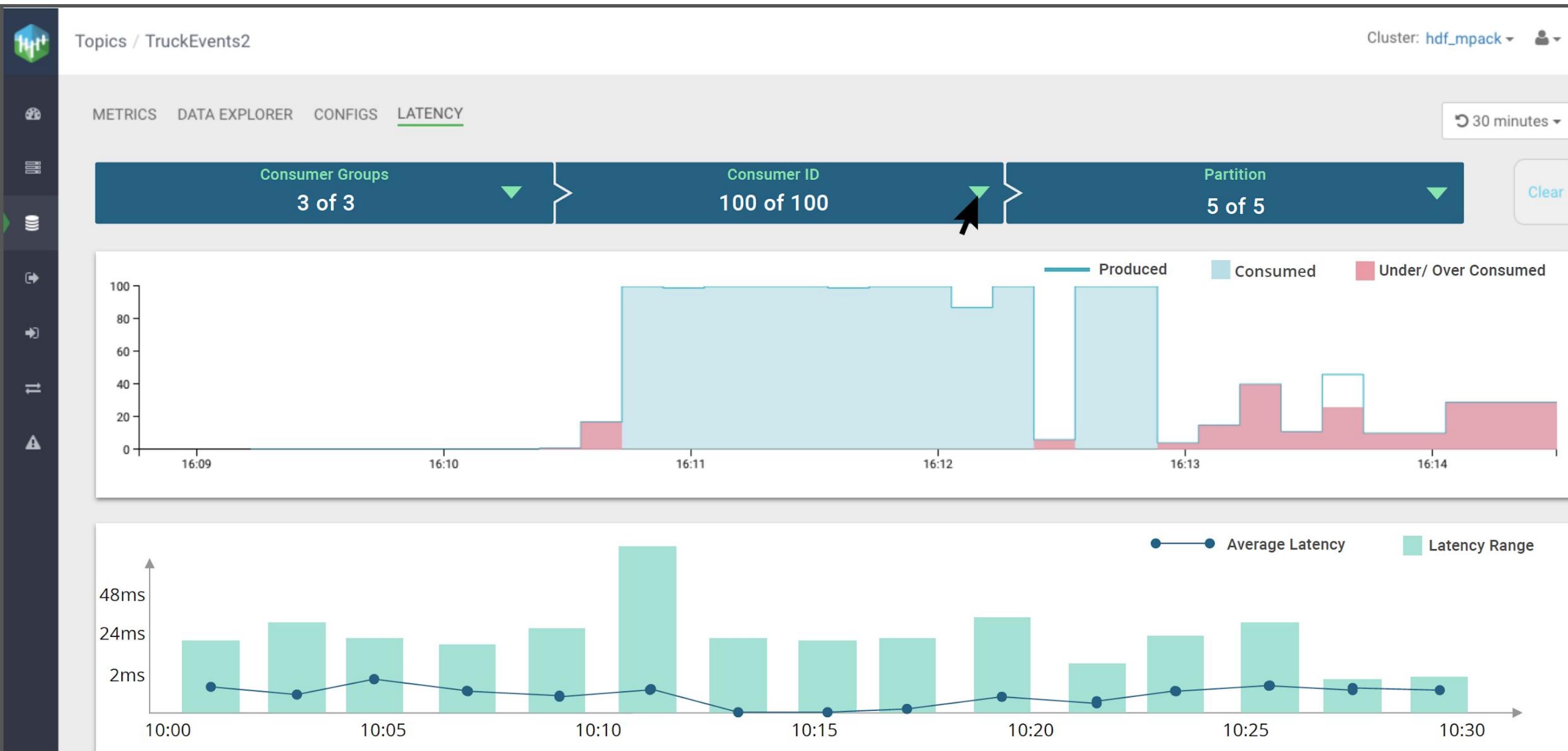
2KB in 2KB out 2KB in 1KB out 2KB in 2KB out 2KB in 1KB out

0 1 2 1

TruckEvents2

TruckEvents

Select the End to End Latency Tab



SMM Atlas Integration

Explore Metadata about the Topic in Atlas

Topics / gateway-west-raw-sensors

METRICS DATA EXPLORER CONFIGS

Producers (3)

Replication Factor: (2) InSync Replicas: 8 Of 8 Total messages: 805 Retention Period: 7 days

		MESSAGES	
	1001	P0	0B in 0B out
	1002	P1	0B in 0B out
	1003	P2	62KB in 62KB out
	1004	P3	98KB in 98KB out

Summary

Number of Replicas	2
Number of Partitions	4
Total number of Brokers for Topic	5
Preferred Replication %	100
Under Replicated %	0

Atlas Integration is not available on the workshop cluster.

Cluster: orlandostreamcluster

The screenshot shows the Apache Atlas interface with the following details:

- Basic Search:** Type: kafka_topic, Query: name=gateway-west-raw-sensors
- Properties:** avgMessageSizeInBytes: 0, avroSchema, contactInfo, description, desiredRetentionInHrs: 0, keyClassName, maxThroughputPerSec: 0, name: gateway-west-raw-sensors, numberOfEventsPerDay: 0, owner, partitionCount: 0, partitionCountLocal: 0, partitionCountNational: 0, qualifiedName: gateway-west-raw-sensors@orlandostreamcluster

Click on Atlas Link to see the metadata of the topic gateway-west-raw-sensors in Atlas

If Atlas does not come up then use this link and then pick Type as Kafka_Topic for search
<https://99.80.132.89:8443/pkuc-wv-smm-m/dp-proxy/atlas/>

Traverse the flow of data across multiple Kafka Topics using SMM and Atlas Integration

Topics / gateway-west-raw-sensors

METRICS DATA EXPLORER CONFIGS

Producers (3)

Replication Factor: (2) InSync Replicas: 8 Of 8 Total messages: 805 Retention Pe...

MESSAGES 1001 P0 0B in 0B out

minifi-truck-w1 298

minifi-truck-w3 230

minifi-truck-w2 263

1002 P1 0B in 0B out

1003 P2 62KB in 62KB out

1004 P3 98KB in 98KB out

Question

The topic has one active consumer which is a NiFi consumer. Which Kafka topic if any is this NiFi Flow consumer publishing events to?

Summary

Bytes In Count 65753969 Bytes Out Count 65762482 Messages In Count 300798

Number of Replicas 2

Number of Partitions 4

Total number of Brokers for Topic 5

Preferred Replication % 100

Under Replicated % 0

Atlas Integration is not available on the workshop cluster.

Cluster: orlandostreamcluster

30 minutes

Consumer Groups (1)



30 minutes

nifi-truck-sensors-west LAG 2

Step 1

Click on Atlas Icon to see lineage of the the topic gateway-west-raw-sensors

Apache Atlas

SEARCH CLASSIFICATION GLOSSARY

Basic Advanced

Search By Type: kafka_topic

Search By Query: name=gateway-west-raw-sensors

Clear Search

Favorite Searches Save Save As

You don't have any favorite search.

gateway-west-raw-sensors (kafka_topic)

Classifications:

Term:

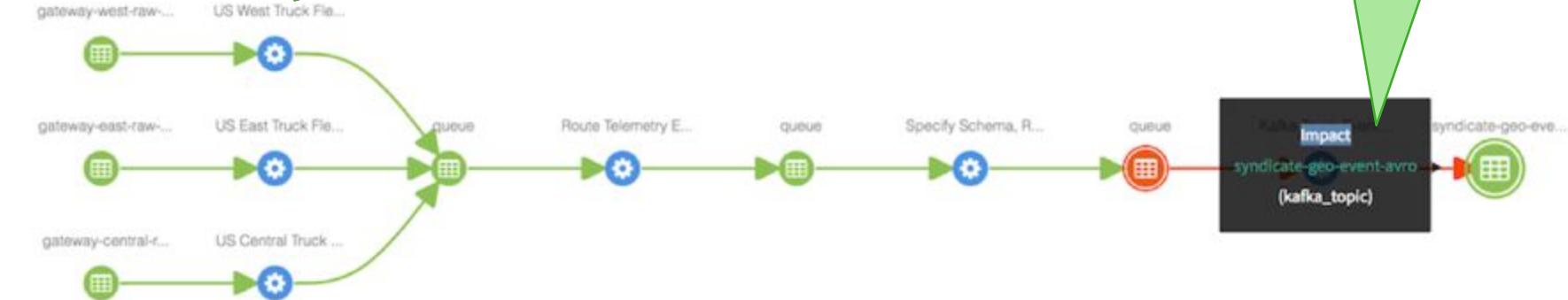
Properties

Lineage

Relationships

Classifications

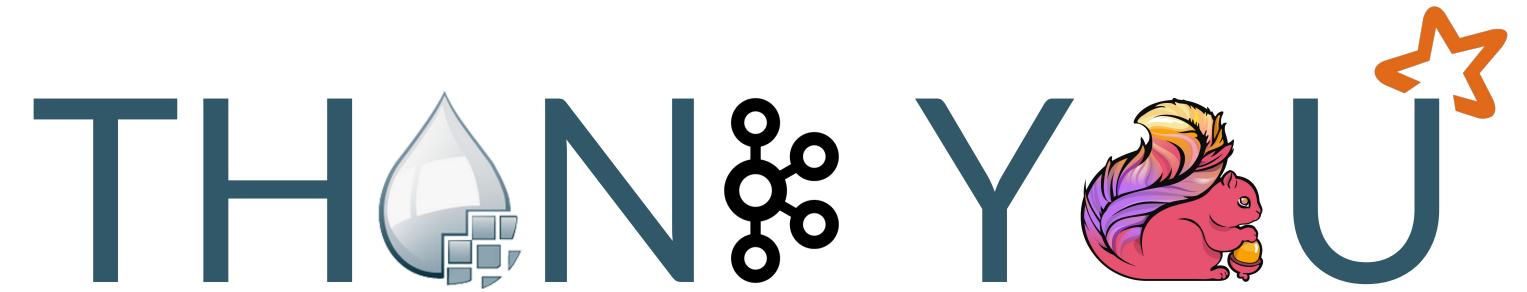
Audits



Analysis

NiFi App consumes from the gateway-west-raw-sensors topic and publishes events to downstream Kafka topic called syndicate-geo-event-avro

TH^ON^G Y^OU[★]

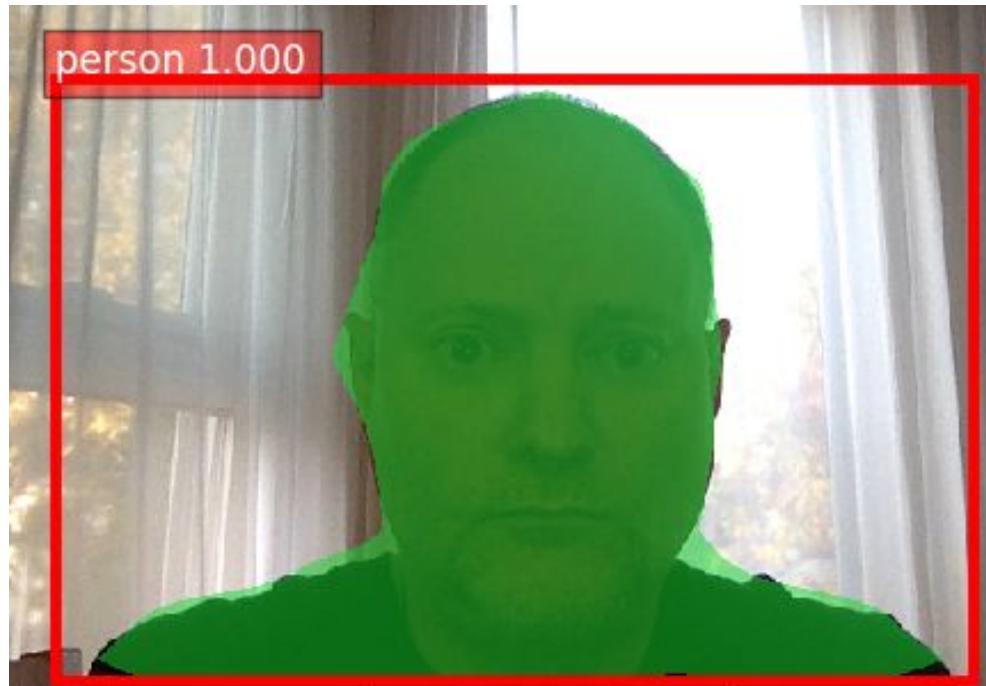


CLOUDERA

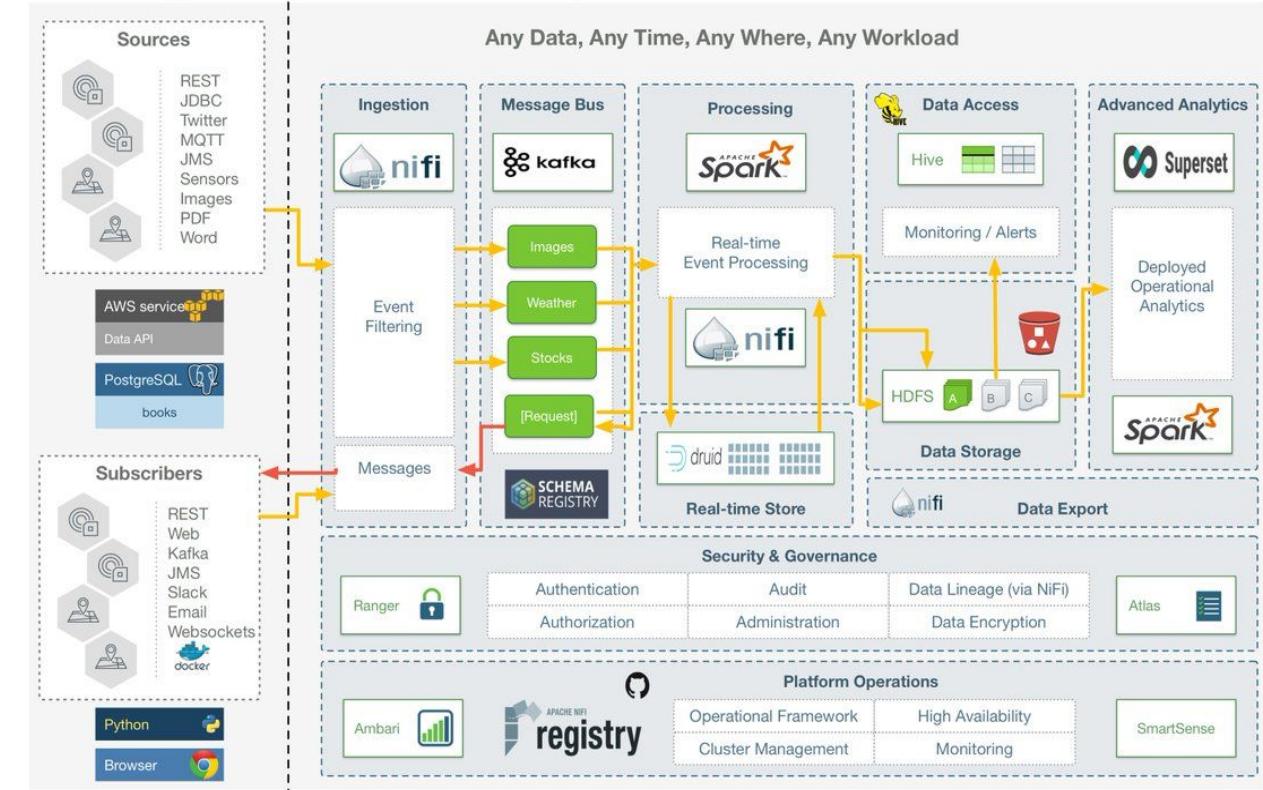
SMM Workshop - Kafka Streams

Timothy Spann

DZone Zone Leader and Big Data MVB;
Princeton Future of Data Meetup;
ex-Pivotal Field Engineer;
Author of Apache Kafka RefCard
<https://github.com/tspannhw>



Streaming with Apache Kafka and Apache NiFi



<https://community.hortonworks.com/articles/227560/real-time-stock-processing-with-apache-nifi-and-ap.html>

<https://community.hortonworks.com/users/9304/tspann.html>

<https://dzone.com/users/297029/bunkertor.html>

<https://www.datainmotion.dev/>

Contents

Kafka CRUD

Kafka Streams Overview

Kafka Streams Walk Through

Kafka Topics and Testing Your Microservice

Kafka Streams Architecture

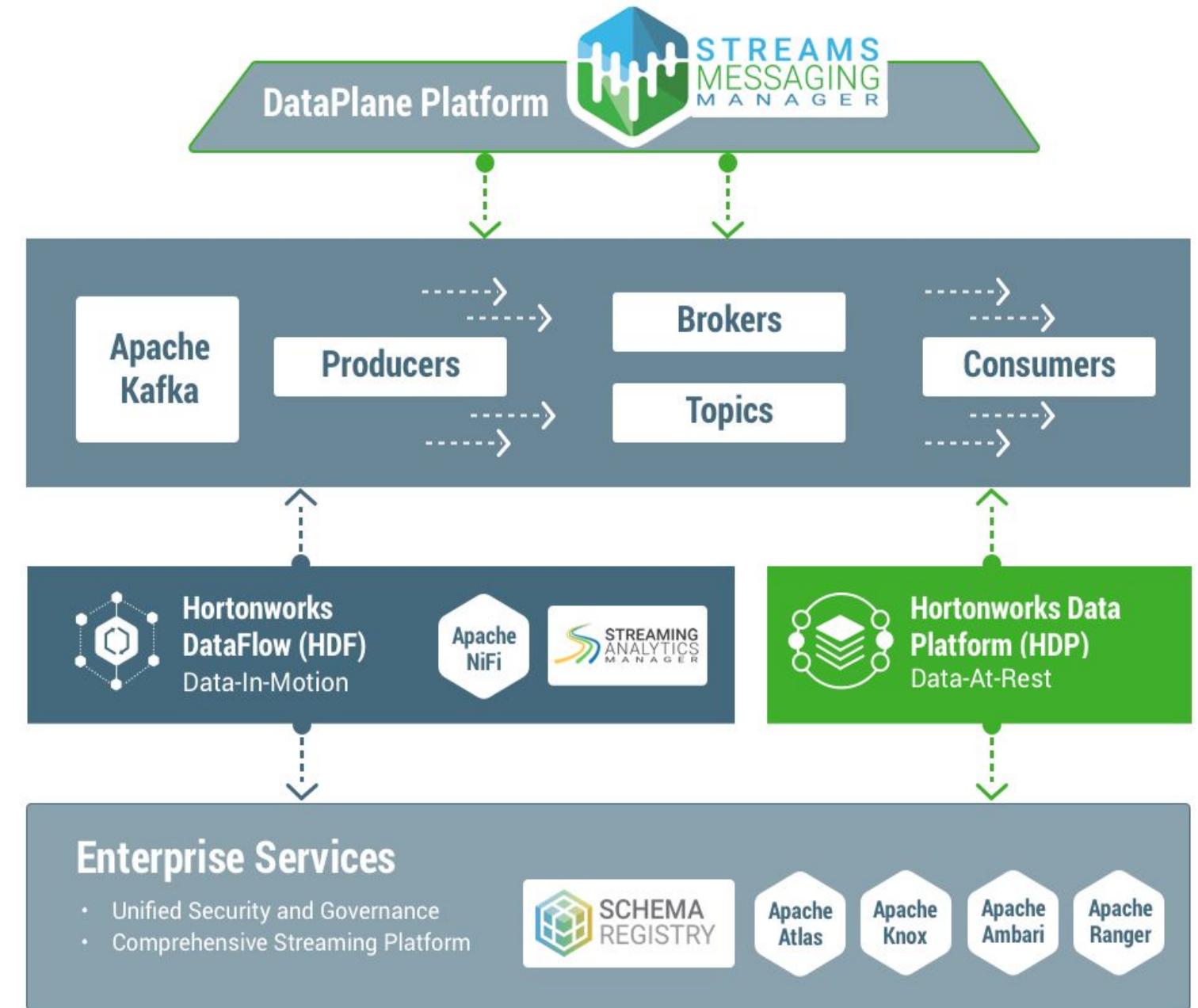
Kafka Streams Example

<https://github.com/purn1mak/SMM-NewYork>

KAFKA CRUD

Streams Messaging Manager (SMM)

- Open Source tool to Cure the “Kafka Blindness”
- Single Monitoring Dashboard for all your Kafka Clusters across 4 entities
- Notification on available metrics
- Designed for the Enterprise
 - Support for Secure Kafka cluster
 - Rich Access Control Policies (ACLS)
 - Supports multiple Kafka Clusters
- REST as a First Class Citizen



Create a Kafka Topic

The screenshot shows the Cloudera Streams Messaging interface. On the left, there's a sidebar with icons for topics, brokers, and logs. The main area displays a list of topics with their sizes: Total bytes in 129 KB and Total bytes out 153 KB. Below this is a table with columns: Name, Last modified, Last offset, and Last partition. A search bar at the top right is set to 'Last 30 minutes'. A modal window titled 'Add Topic' is open in the center. It has fields for 'TOPIC NAME' (cleandata) and 'PARTITIONS' (1). Under 'Availability', there are five options: MAXIMUM (selected), HIGH, MODERATE, LOW, and CUSTOM. For each, there are fields for 'REPLICATION FACTOR 2', 'MIN IN SYNC', and 'REPLICA 2'. Under 'Limits', there's a dropdown for 'CLEANUP POLICY' with options: 'Delete' (selected), 'compact', and 'delete'. At the bottom of the modal are 'Advanced', 'Cancel', and 'Save' buttons.

Create a Kafka Topic

Add Topic

TOPIC NAME cleandata PARTITIONS 1

Availability

MAXIMUM HIGH MODERATE LOW CUSTOM

REPLICATION FACTOR 3 MIN INSYNC REPLICA 2
REPLICATION FACTOR 3 MIN INSYNC REPLICA 1
REPLICATION FACTOR 2 MIN INSYNC REPLICA 1
REPLICATION FACTOR 1 MIN INSYNC REPLICA 1

Limits

CLEANUP.POLICY

Select...
compact
delete
compact,delete

Advanced Cancel Save

0

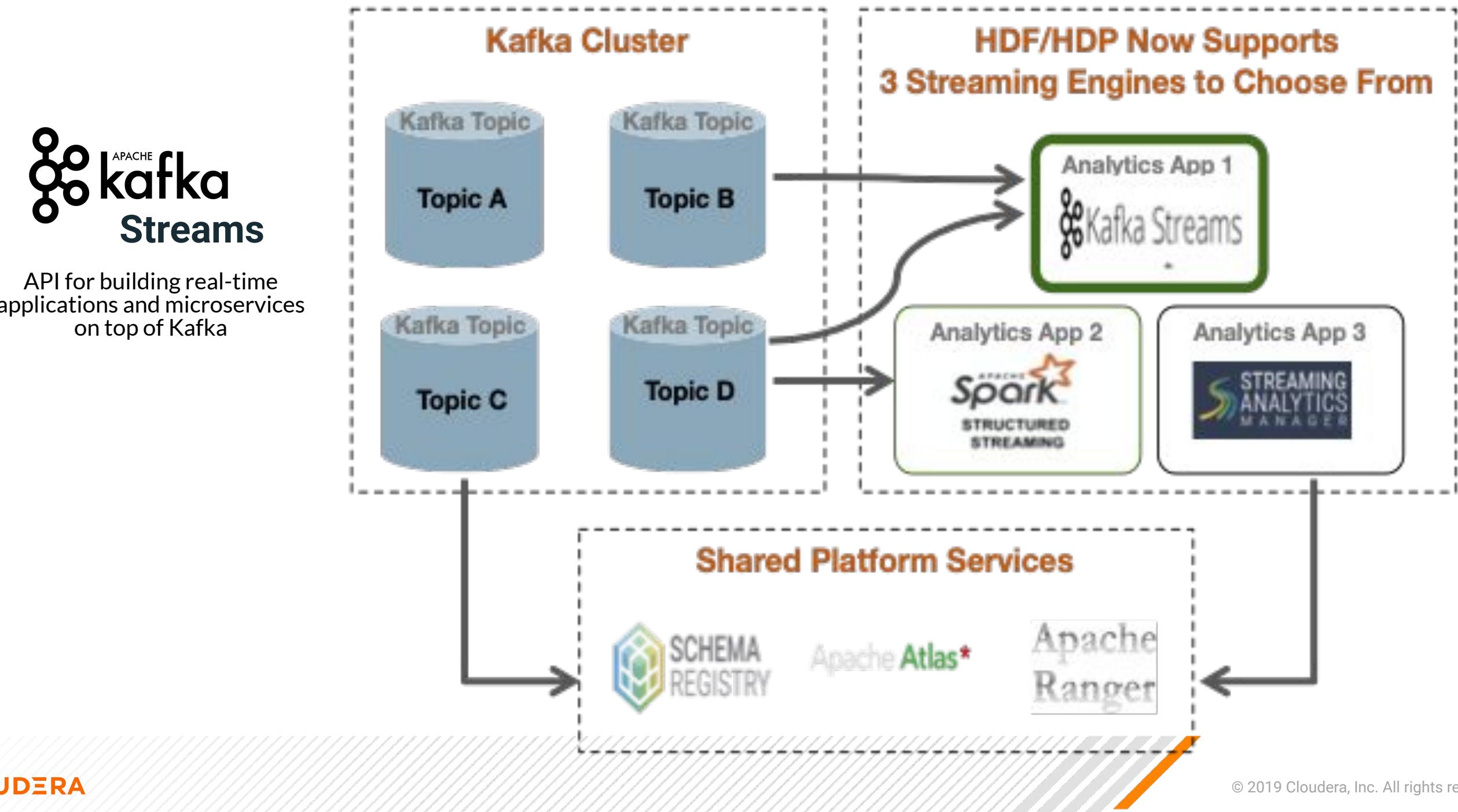
- Topic Name
- Partitions (1 or more)
- Pick a Template with replicas and replication factor
- Carefully select a Cleanup Policy **DELETE**
 - (your messages need a key)

Create Kafka Topics For Kafka Streams Application

- Topic Name: **streams-plaintext-input**
 - Partitions: 1
 - Replicas: 1
 - Delete
- Topic Name: **streams-wordcount-output**
 - Partitions: 1
 - Replicas: 1
 - Delete

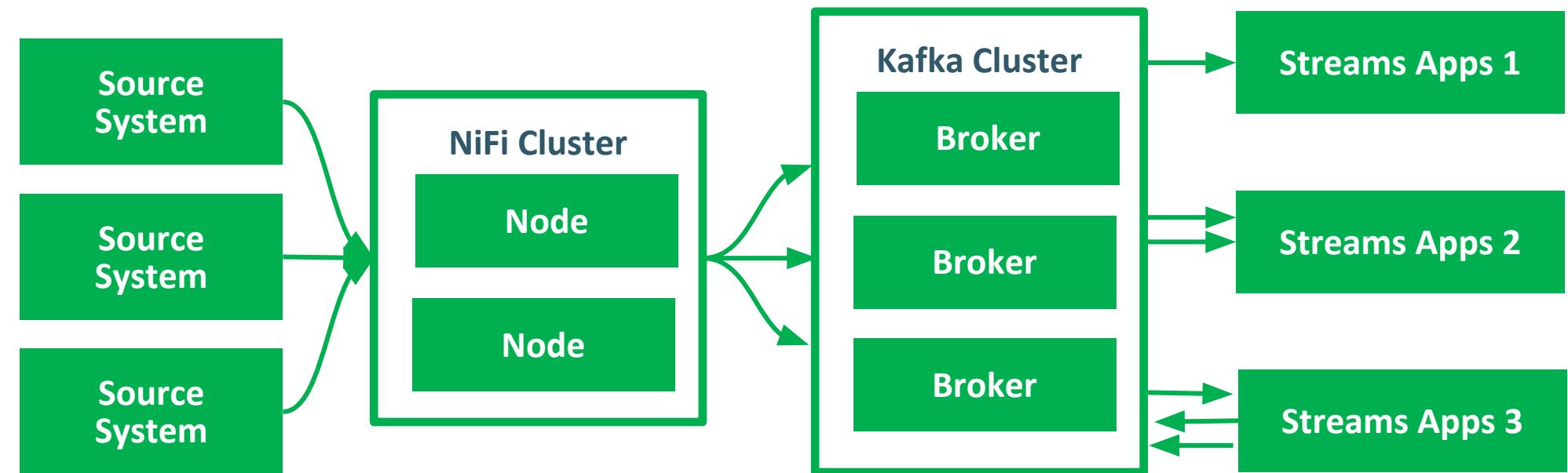
KAFKA STREAMS

Secure and Governed Microservices with Kafka Streams Support



Kafka Streams

- A client library for applications and microservices on top of Kafka
- Written in Java or Scala
- Elastic, highly scalable, fault-tolerant
- Supports At-Least-Once and Exactly-once semantics inside Kafka
- Deploy to containers, VMs, bare metal, cloud, K8, devices.



Kafka Streams Library

Stream Processing Style	Event a time
Delivery Guarantee	Exactly Once (within Kafka)
State Management	Implicit support with RocksDB
Fault Tolerance	Implicit support with internal Kafka topics / ZK
Advanced feature: Joins/ Aggregations/ Windowing	Stream/Stream joins, Stream/Table joins, joins with only message key, no OOO aggregation processors
Advanced Feature: Watermarking (late arriving data)	Not fully supported
Latency	Low
Throughput	Medium
APIs	Simple APIs (Compositional)

Kafka Streams Library

SQL DSL	Nothing in Apache Kafka.
Lambda Architecture	No Supported
GUI Tooling / Code-Less Approach	Not Supported
Maturity	Relatively New
Use Cases	Lightweight eventing based microservices, ETL
Cluster Requirement	No

KAFKA STREAMS WALK THROUGH

Tutorial - Log In Instructions



Cloudera Manager <http://54.190.22.232:7180/>

Edge Flow <http://54.190.22.232:10080/efm/ui/>

NiFi <http://54.190.22.232:8080/nifi/>

NiFi Registry <http://54.190.22.232:18080/nifi-registry/>

Schema Registry <http://54.190.22.232:7788/>

SMM <http://54.190.22.232:9991/>

Hue <http://54.190.22.232:8888/>

Cloudera Data Science Workbench <http://cdsw.54.190.22.232.nip.io/>

SSH Connection

Download key:

Download SSH Key



And then run:

`chmod 400 workshop.pem`

`ssh -i workshop.pem centos@54.190.22.232`

Topic - Create

Add Topic

TOPIC NAME

PARTITIONS

Availability



MAXIMUM



HIGH



MODERATE



LOW



CUSTOM

REPLICATION

FACTOR 3

MIN INSYNC

REPLICA 2

REPLICATION

FACTOR 3

MIN INSYNC

REPLICA 1

REPLICATION

FACTOR 2

MIN INSYNC

REPLICA 1

REPLICATION

FACTOR 1

MIN INSYNC

REPLICA 1

Limits

CLEANUP.POLICY

Advanced

Cancel

Save

Topics							
Total Bytes In	Total Bytes Out	Produced Per Sec	Fetched Per Sec	In Sync Replicas	Out Of Sync	Under Replicated	Offline Partitions
88 MB	2 MB	212	169	196	0	0	0
Topics (42)							
NAME	DATA IN	DATA OUT	MESSAGES IN	CONSUMER GROUPS			
streams-plaintext-input	0B	0B	0	0			
streams-wordcount-output	0B	0B	0	0			

Cluster: OneNodeCluster
Topic added successfully

Setup Your Development Environment

```
sudo su
```

```
yum install maven curl wget unzip zip git -y
```

```
mkdir /opt/demo
```

```
chmod -R 777 /opt/demo
```

```
cd /opt/demo
```

Generate a Kafka Streams Project

```
mvn archetype:generate \  
  -DarchetypeGroupId=org.apache.kafka \  
  -DarchetypeArtifactId=streams-quickstart-java \  
  -DarchetypeVersion=2.2.0 \  
  -DgroupId=streams.examples \  
  -DartifactId=streams.examples \  
  -Dversion=0.1 \  
  -Dpackage=myapps
```

Update Word Count Kafka Connection

Edit the file

/opt/demostreams.examples/src/main/java/myapps/**WordCount.java**

```
Properties props = new Properties();
props.put(StreamsConfig.APPLICATION_ID_CONFIG,
"streams-pipe");
props.put(StreamsConfig.BOOTSTRAP_SERVERS_CONFIG,
"<YOURAWSINTERNALIPNAME>:9092");
```

Build a Kafka Streams Project

```
rm /opt/demostreams.examples/src/main/java/myapps/Pipe.java  
rm /opt/demostreams.examples/src/main/java/myapps/LineSplit.java  
mvn clean package
```

KAFKA TOPIC INTERACTION

Produce Kafka Messages

```
/opt/cloudera/parcels/CDH/lib/kafka/bin/kafka-console-producer.sh --broker-list  
<YOURAWSINTERNALIPNAME>:9092 --topic streams-plaintext-input
```

Instructions: Then type messages with a return.

Run Kafka Streams Java Application

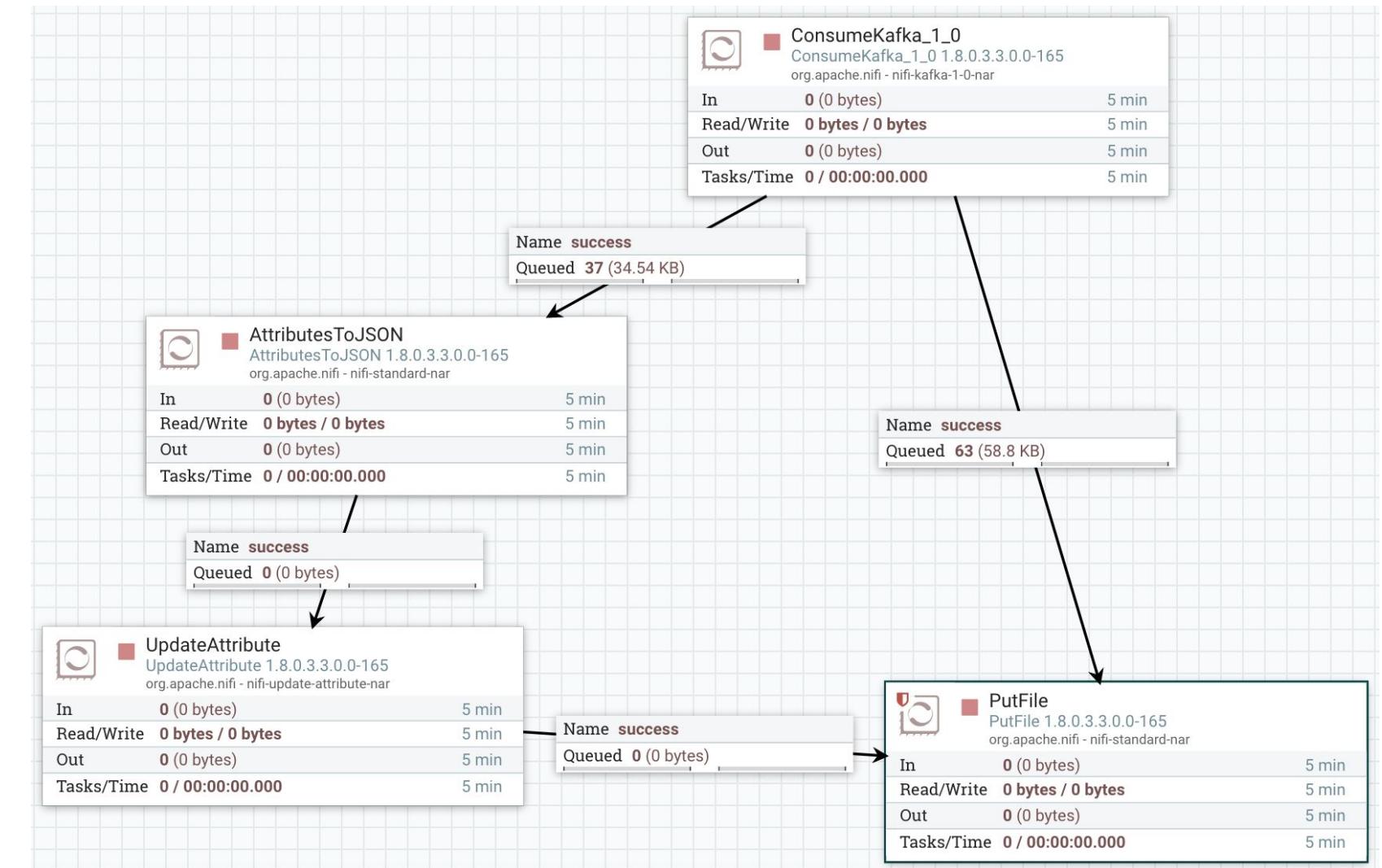
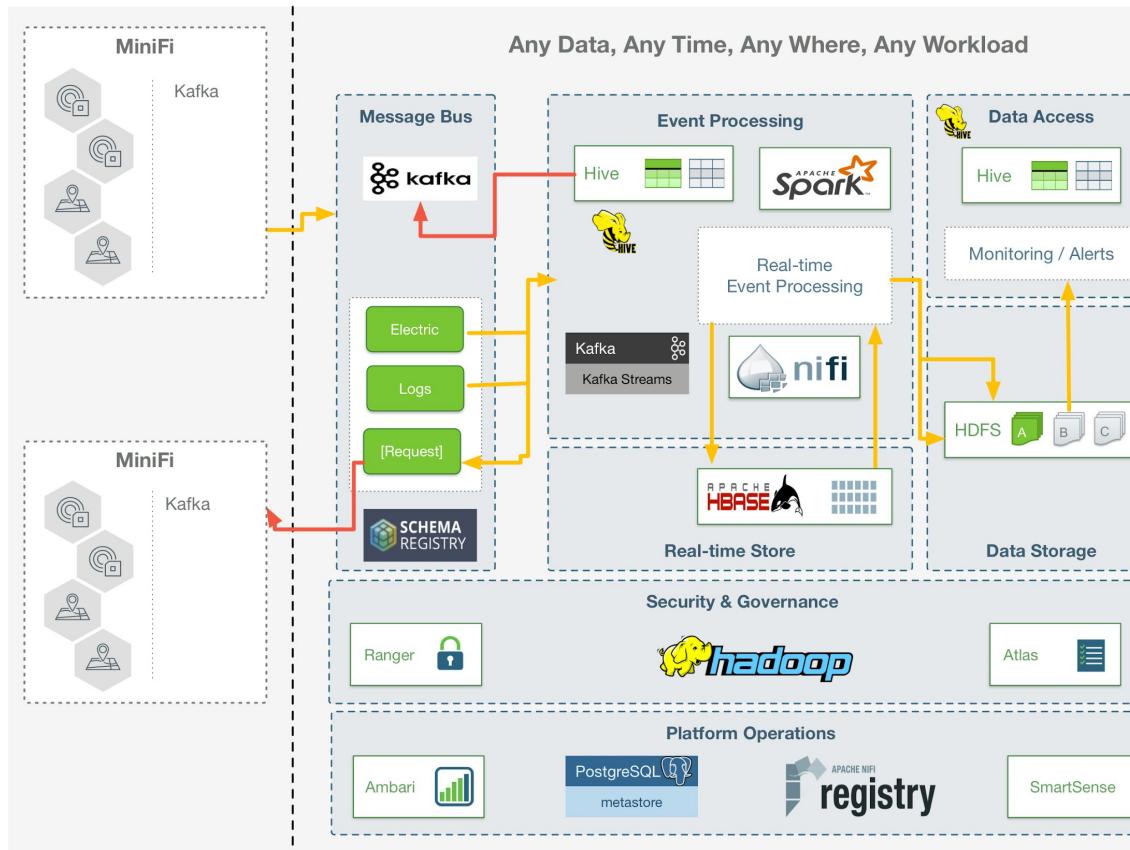
```
mvn exec:java -Dexec.mainClass=myapps.WordCount
```

Consume Kafka Messages

```
/opt/cloudera/parcels/CDH-6.3.0-1.cdh6.3.0.p0.1279813/lib/kafka/bin/kafka-console-consumer.sh --bootstrap-server <YOURAWSINTERNALIPNAME>:9092 \
--topic streams-wordcount-output \
--from-beginning \
--formatter kafka.tools.DefaultMessageFormatter \
--property print.key=true \
--property print.value=true \
--property key.deserializer=org.apache.kafka.common.serialization.StringDeserializer \
--property value.deserializer=org.apache.kafka.common.serialization.LongDeserializer
```

CONSUME AND PRODUCE KAFKA MESSAGES

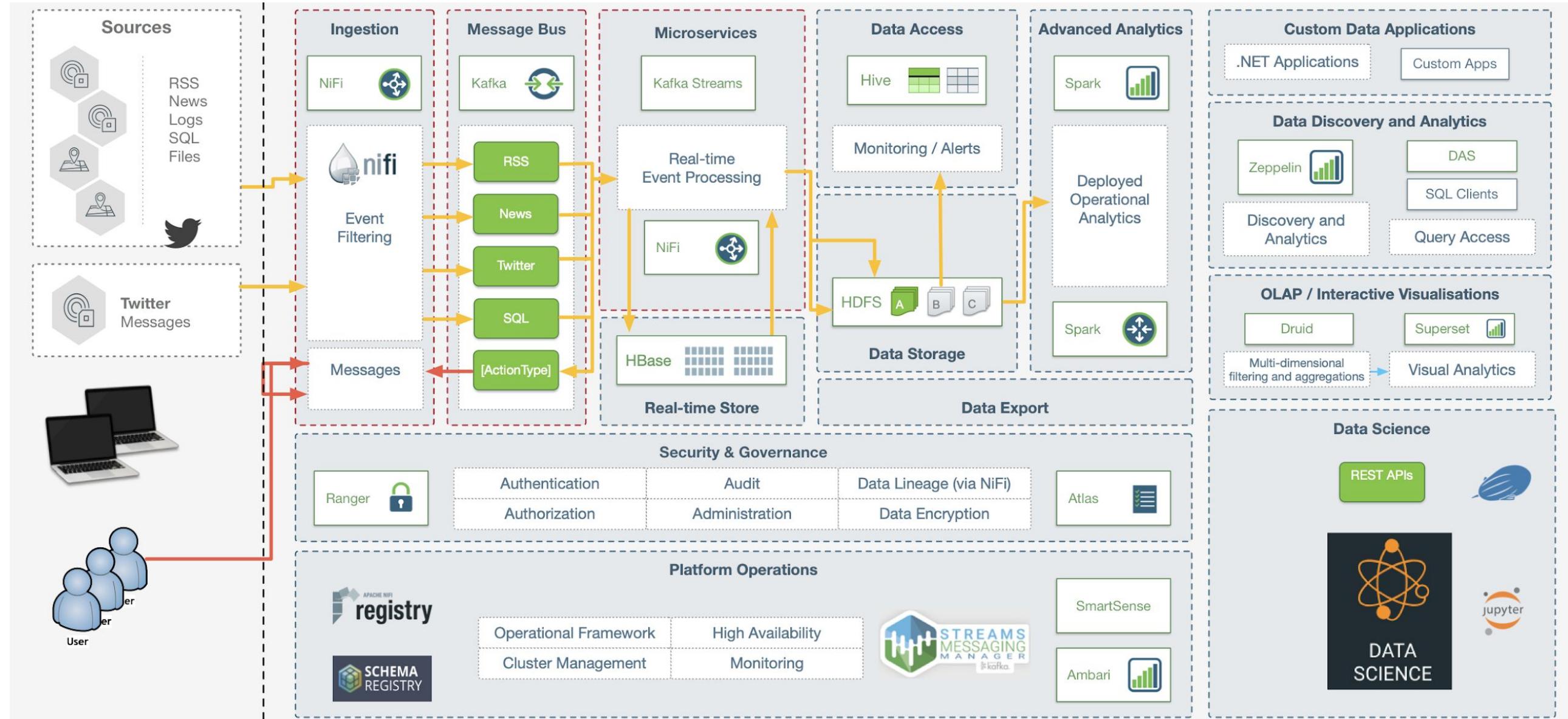
IoT Edge Use Cases with Apache Kafka and Apache NiFi - MiniFi



<https://community.cloudera.com/t5/Community-Articles/IoT-Edge-Use-Cases-with-Apache-Kafka-and-Apache-NiFi-MiniFi/ta-p/249232>

KAFKA STREAMS ADDITIONAL EXAMPLE

Kafka Streams Example Architecture



Kafka Streams

```
Timer timer = new Timer();  
timer.schedule(new DisplayStatus(), 0, 120000);  
  
final StreamsBuilder builder = new StreamsBuilder();  
KStream<String, String> source = builder.stream("bme680");  
  
source.foreach((key, value) -> processValues(key, value));  
source.to("bme680out");  
  
final Topology topology = builder.build();  
final KafkaStreams streams = new KafkaStreams(topology, props);  
final CountDownLatch latch = new CountDownLatch(1);
```

Kafka Streams

```
Runtime.getRuntime().addShutdownHook(new Thread(STREAMS_SHUTDOWN_HOOK) {  
    @Override  
    public void run() {  
        log.error(SHUTDOWN);  
        streams.close();  
        latch.countDown();  
    }  
});  
  
try {  
    streams.start();  
    latch.await();  
} catch (Throwable e) {  
    System.exit(1);  
}
```

Kafka Streams Example 2

The screenshot shows an IDE interface with the following details:

- Project Structure:** The project is named "kstreams" located at "/Volumes/TSPANN/projects/kstreams". It contains a ".idea" folder, a "src" directory with "main" and "java" sub-directories, and a "com.dataflowdeveloper.kstream" package containing a class "BME680". Inside "resources", there are "log4j2.properties" and "log4j2_p.swp" files.
- Code Editor:** The "BME680.java" file is open. The code initializes MQTT connections. A specific line of code, `options.setCleanSession(true);`, is highlighted with a yellow background.
- Run Tab:** The "Run" tab shows a single configuration named "BME680" with the command: `/Library/Java/JavaVirtualMachines/jdk1.8.0_121.jdk/Contents/Home/bin/java ...`
- Output Tab:** The output window displays the following log messages:

```
objc[80616]: Class JavaLaunchHelper is implemented in both /Library/Java/JavaVirtualMachines/jdk1.8.0_121.jdk/Contents/Home/bin/java ...
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
```

Cloudera Managed Kafka

Cloudera Manager Clusters ▾ Hosts ▾ Diagnostics ▾ Audits Charts ▾ Backup ▾ Administration ▾ Search Support ▾ admin ▾

nyc KAFKA-2 Actions ▾ ← 24 hours preceding Sep 12, 7:54 PM UTC → 🔍

Status Instances Configuration Commands Charts Library Audits Quick Links ▾ Hide Descriptions 30m 1h 2h 6h 12h 1d 7d 30d

Status Page Charts

Topics Filter

- __smm_consumer_metrics
- __smm_producer_metrics
- global-retail-pos
- global-retail-store**
- global-retail-web
- heartbeats
- local-retail
- mm2-configs.paris.internal
- mm2-offset-syncs.paris.internal
- mm2-offsets.paris.internal
- Events

Total Bytes Received Across Kafka Broker Topics
The sum of the **Bytes Received** metric computed across all this entity's descendant Kafka Broker Topic entities.

bytes / second

Thu 12 06 AM 12 PM 06 PM

CD-KAFKA-plvdjsn:global-retail-store, total_kafka_b... 0

Total Bytes Fetched Across Kafka Broker Topics
The sum of the **Bytes Fetched** metric computed across all this entity's descendant Kafka Broker Topic entities.

bytes / second

Thu 12 06 AM 12 PM 06 PM

CD-KAFKA-plvdjsn:global-retail-store, total_kafka_b... 0

Total Bytes Rejected Across Kafka Broker Topics
The sum of the **Bytes Rejected** metric computed across all this entity's descendant Kafka Broker Topic entities.

bytes / second

Thu 12 06 AM 12 PM 06 PM

CD-KAFKA-plvdjsn:global-retail-store, total_kafka_b... 0

Total Messages Received Across Kafka Broker Topics
The sum of the **Messages Received** metric computed across all this entity's descendant Kafka Broker Topic entities.

messages / sec...

Thu 12 06 AM 12 PM 06 PM

CD-KAFKA-plvdjsn:global-retail-store, total_kafka_b... 0

Total Rejected Message Batches Across Kafka Broker Topics
The sum of the **Rejected Message Batches** metric computed across all this entity's descendant Kafka Broker Topic entities.

message_batches...

Thu 12 06 AM 12 PM 06 PM

CD-KAFKA-plvdjsn:global-retail-store, total_kafka_b... 0

Total Fetch Request Failures Across Kafka Broker Topics
The sum of the **Fetch Request Failures** metric computed across all this entity's descendant Kafka Broker Topic entities.

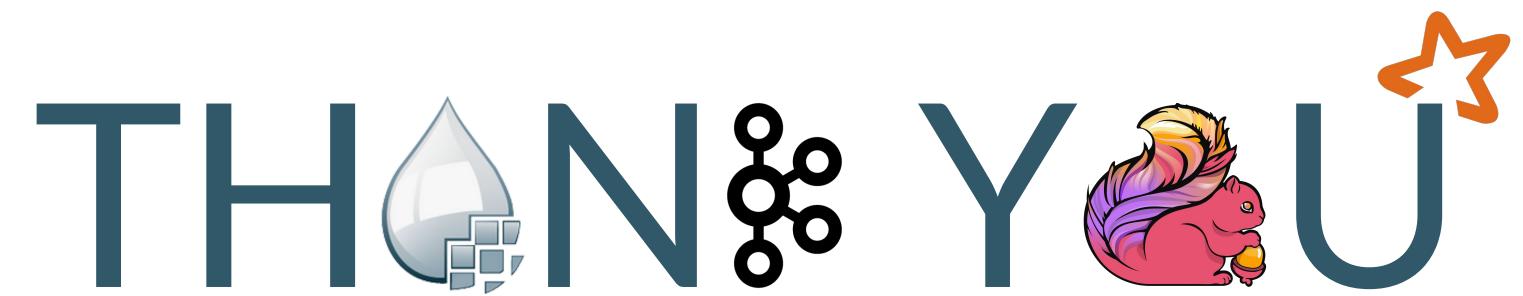
fetch_requests ...

Thu 12 06 AM 12 PM 06 PM

CD-KAFKA-plvdjsn:global-retail-store, total_kafka_f... 0

Feedback

TH^ON^G Y^OU[★]



CLOUDERA

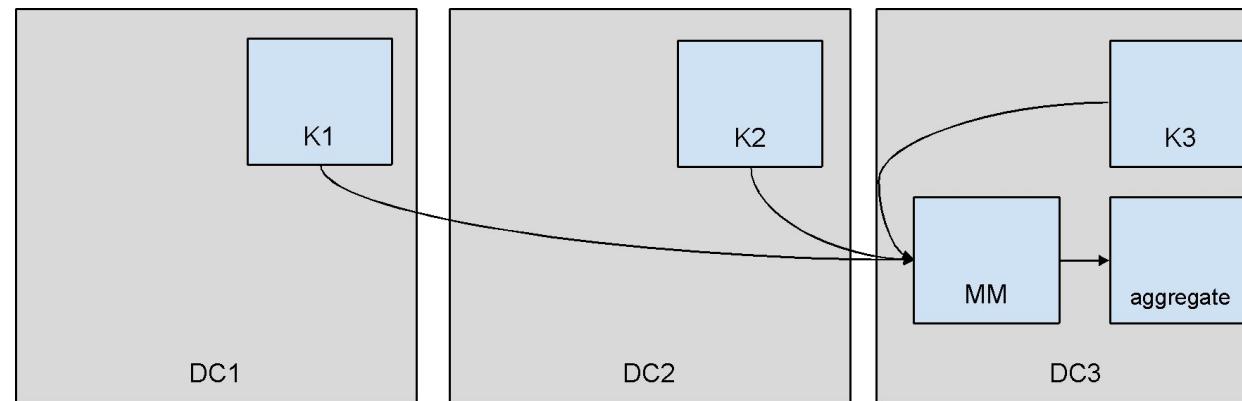
Kafka Replication

Kafka Replication Use Cases

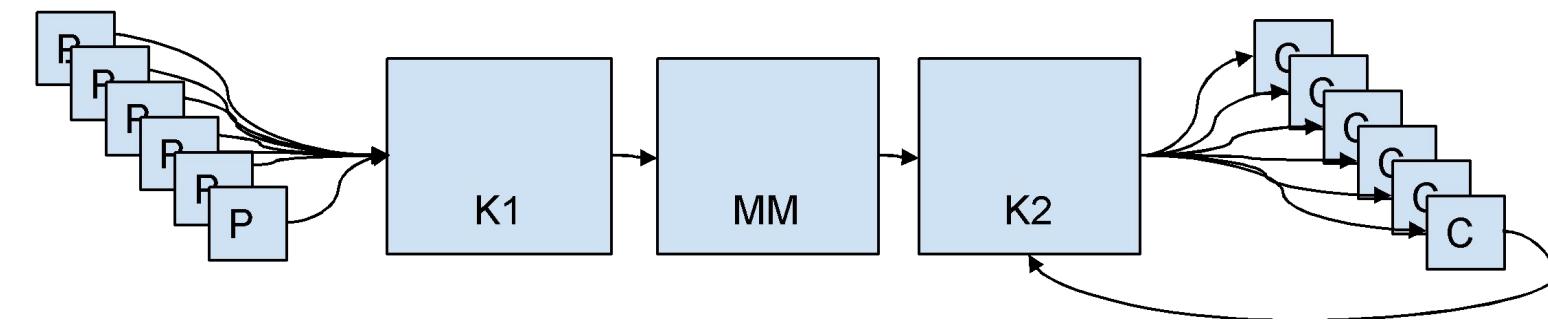
Disaster Recovery In an event of a partial or complete datacenter disaster, providing failover/fallback to a secondary cluster in a different region / DC	Geo-Locality Active-active geo-localized deployments allows users to access a near-by data center to optimize their architecture for low latency and high performance.	Data Movement / Deployment Use Kafka to synchronize data between on-prem applications and cloud deployments
Centralized Analytics Aggregate data from multiple Kafka clusters into one location for organization-wide analytics	Workload Isolation Creation of different envs for SDLC: Dev, Test, Prod. Clusters for specific use case cases (ETL, ingestion, analytics, etc)	Legal / Compliance Since different regions have different data storage and security requirements, clusters need to be created in region but data still needs to be shared.

Various Replication Deployments Based on Use Cases

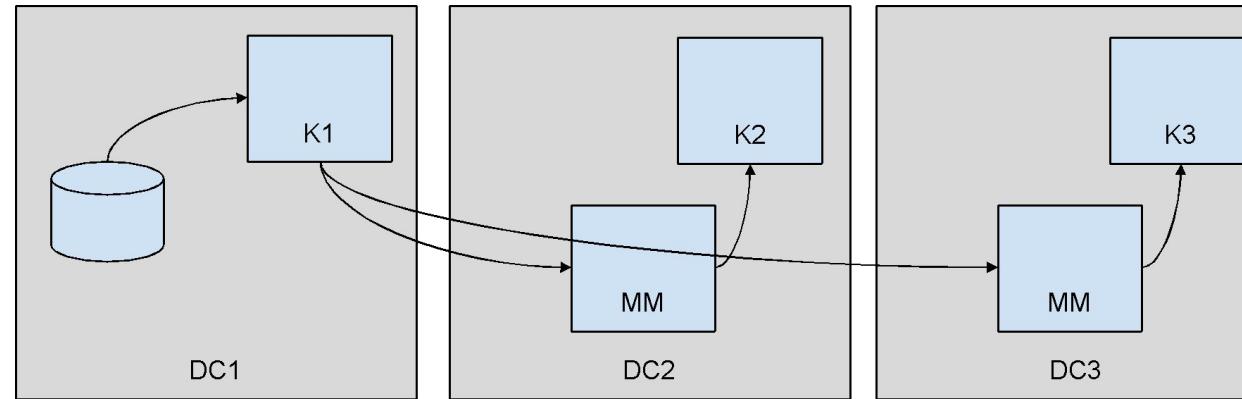
Aggregation



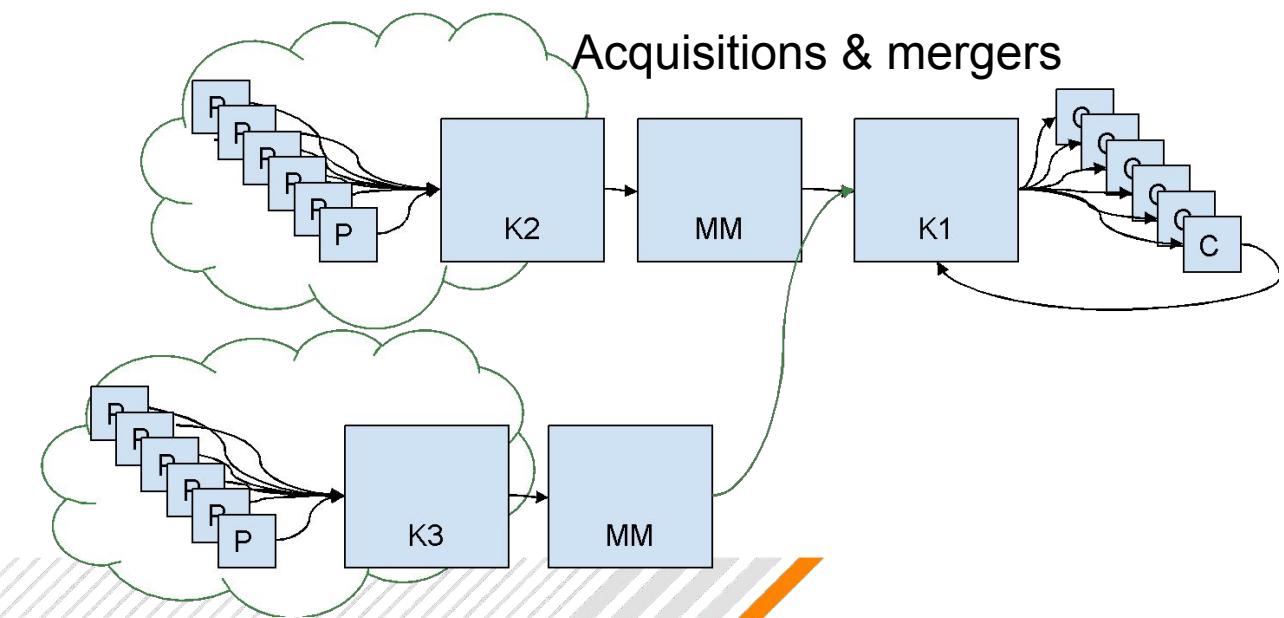
Segmentation



Data Deployment



Acquisitions & mergers



Legacy MirrorMaker

Limitations with Legacy MirrorMaker(MM) for Replication Use Cases

Challenge / Limitation	Description
Static Whitelists and Blacklists	Changes to replication jobs require restart of MirrorMaker cluster.
No Configuration/Metadata Synch	Topic configuration changes in the origin cluster are not detected and propagated to the destination cluster. New Topics are not detected
Scalability and Throughput Limitations due to Rebalances	Cannot scale replication processes as Kafka traffic increases
Lack of Monitoring and Operational Support	No tooling to install and manage replication cluster. Difficult to monitor replication lag, consumer and producer metrics for replication workflows
No Disaster Recovery, Migration, Failover	MM doesn't support active/active without capabilities such as topic renaming, etc. No support for client failover/failover out of the box.