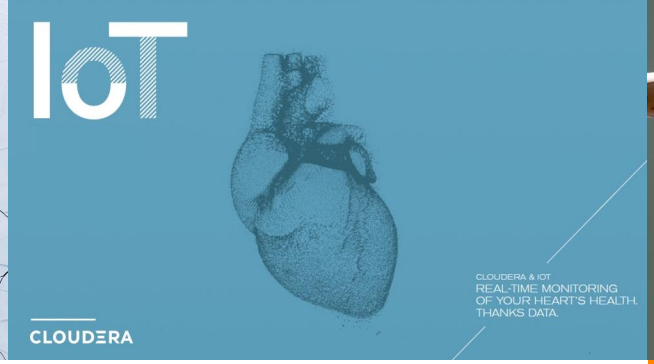
CLOUDERA

Continuous SQL with Kafka and Flink

Tim Spann
Principal Developer Advocate

February 20, 2024

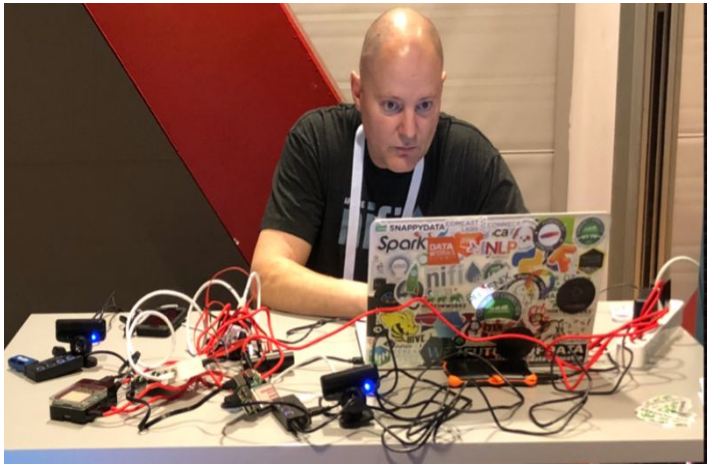
https://www.meetup.com/dba-fundamentals-group/events/296855261/



CLOUDERA

EDGE
2AI
CLOUDERA

ENTERPRISE
DATA
CLOUD
CLOUDERA



Tim Spann

Twitter: @PaasDev // Blog: **datainmotion.dev**

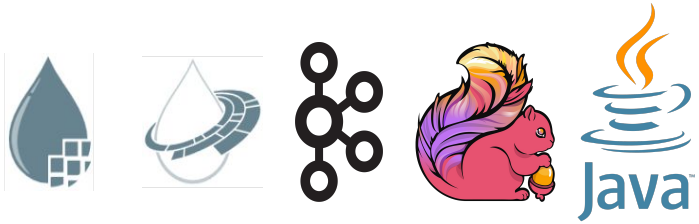
Principal Developer Advocate.

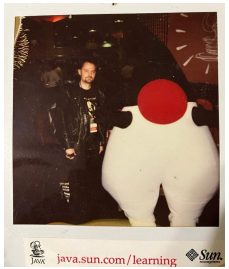
Princeton Future of Data Meetup.

ex-Pivotal, ex-Hortonworks, ex-StreamNative, ex-PwC, ex-HPE

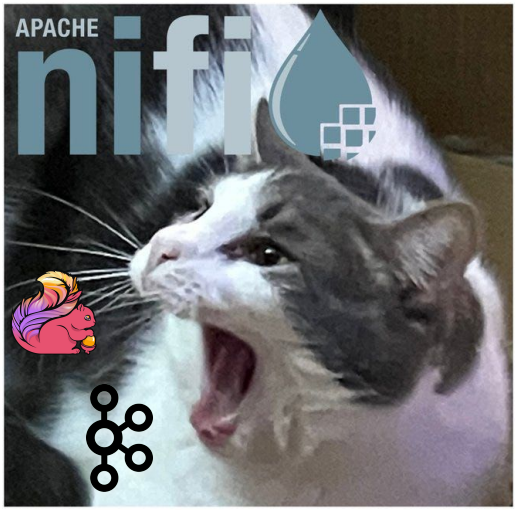
https://medium.com/@tspann

https://github.com/tspannhw





FLaNK Stack Weekly by Tim Spann

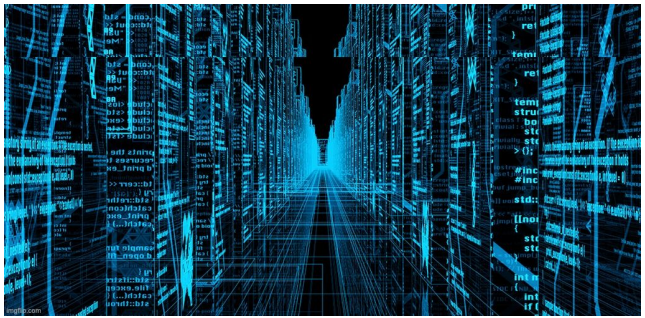


https://bit.ly/32dAJft

https://www.meetup.com/futureofdata-princeton/

This week in Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java and Open Source friends.

Future of Data - NYC + NJ + Philly + Virtual



https://www.meetup.com/futureofdata-princeton/

From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...

ICEBERG

@PaasDev

Introduction

Overview

Streaming Projects

Streaming Analytics

Demos

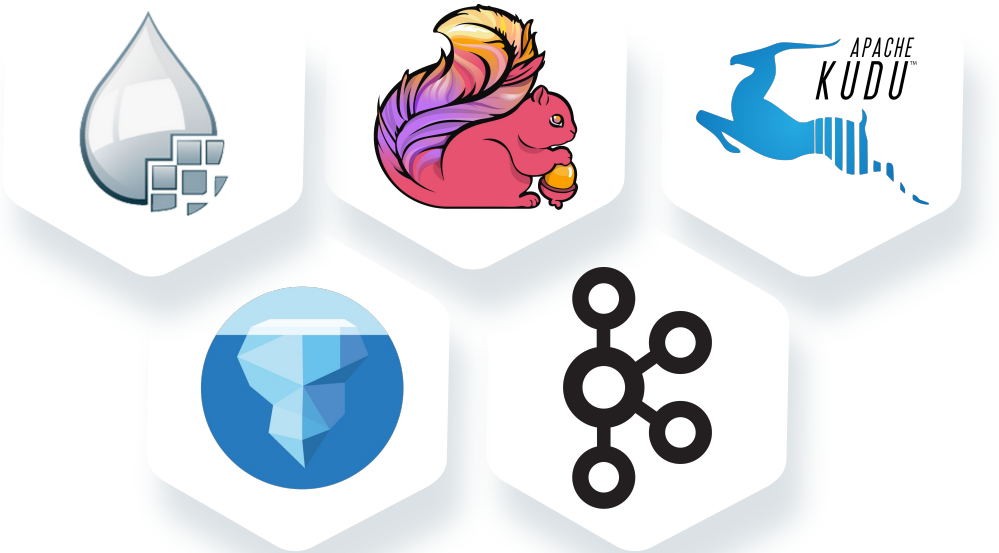
Resources

Q&A

FLANK


DATA ENGINEER
CODER
JAVA DEVELOPER
STREAMING ENGINEER
imgflip.com

BUILDING REAL-TIME REQUIRES A TEAM



Spark, NiFi, Flink? Which engine to choose?

Already using Spark?

Want unified Batch/Stream?

Want highest Throughput?

Don't need low latency?

Large files?

Scheduled batches?

Replacing Sqoop, ETL

Need NiFi?

Simple JDBC queries?

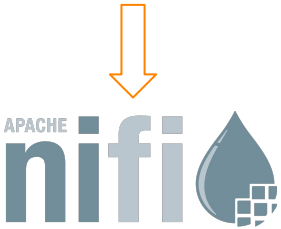
Transform individual records?

Want easy development?

Lots of small files, events, records, rows?

Continuous stream of rows

Support many different sources



Need Flink?

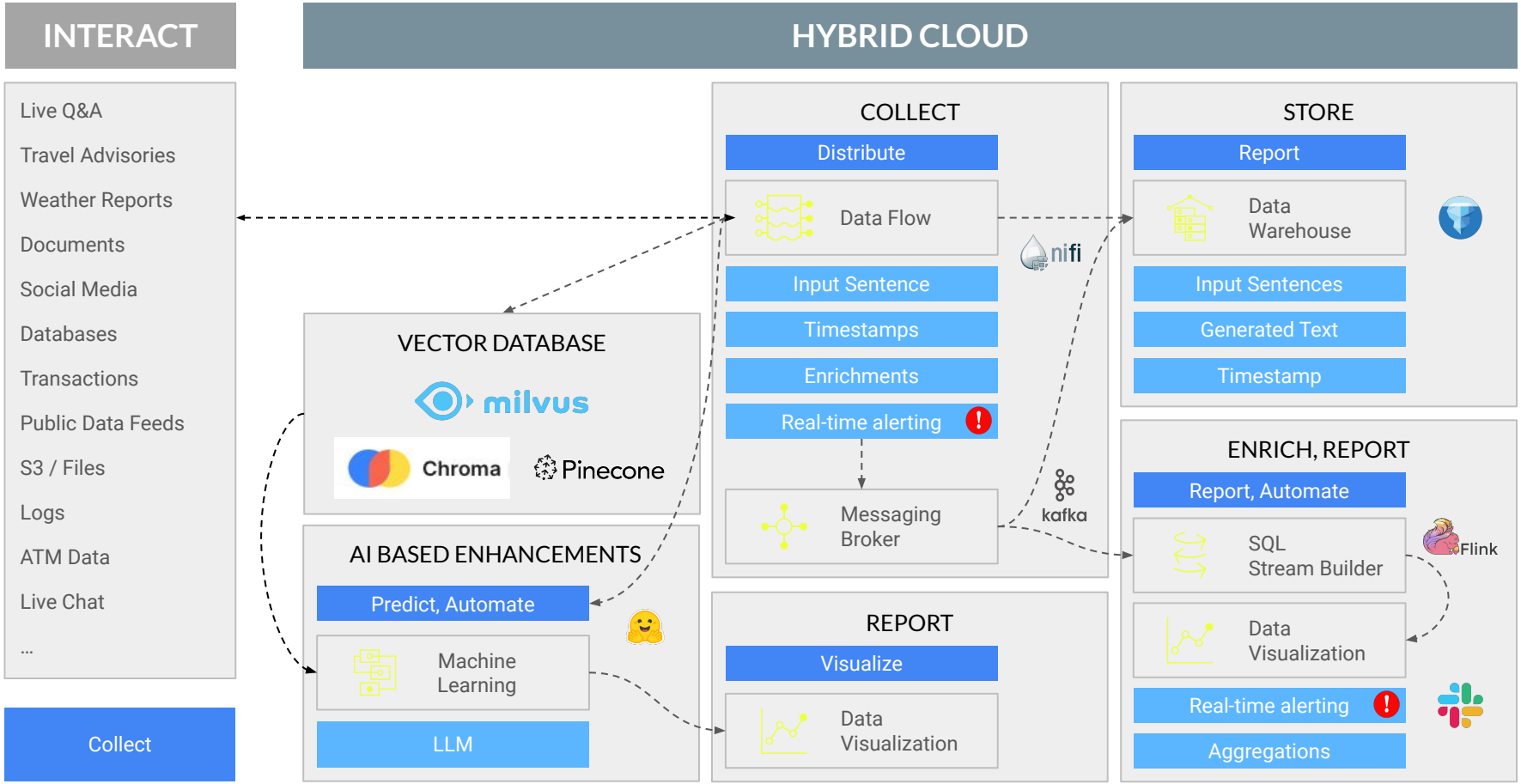
Need Microservices, Batch and Stream?

Want high Throughput?

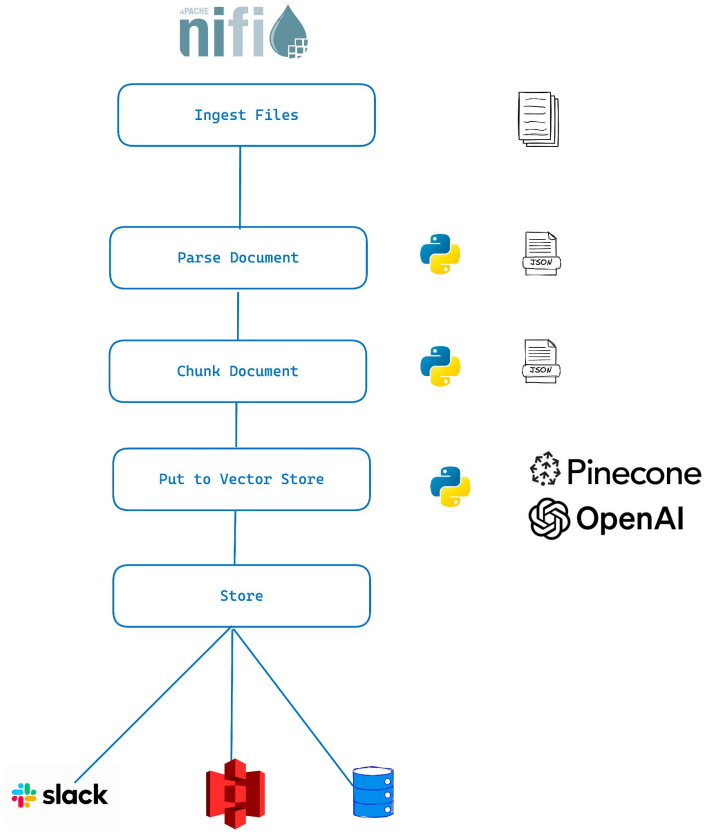
Want Low Latency?

Want Advanced Windowing and State?

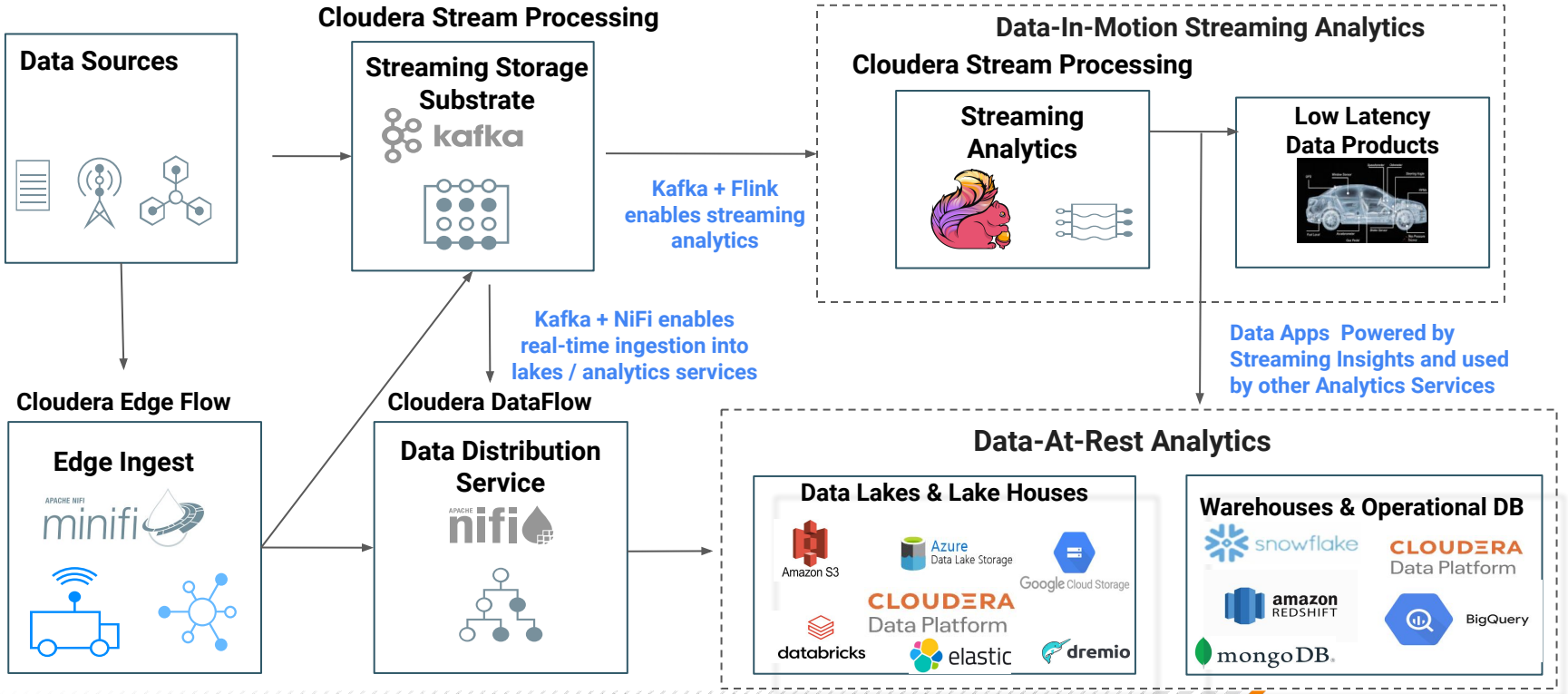
Happy with a New Solution that is best-in-class?


INTERACT
HYBRID CLOUD
Live Q&A
Travel Advisories
Weather Reports
Documents
Social Media
Databases
Transactions
Public Data Feeds
S3 / Files
Logs
ATM Data
Live Chat
...
Collect
COLLECT
Distribute
Data Flow
nifi
Input Sentence
Timestamps
Enrichments
Real-time alerting
Messaging Broker
kafka
STORE
Report
Data Warehouse
Input Sentences
Generated Text
Timestamp
VECTOR DATABASE
milvus
Chroma
Pinecone
AI BASED ENHANCEMENTS
Predict, Automate
Machine Learning
LLM
REPORT
Visualize
Data Visualization
ENRICH, REPORT
Report, Automate
SQL Stream Builder
Flink
Data Visualization
Real-time alerting
Aggregations

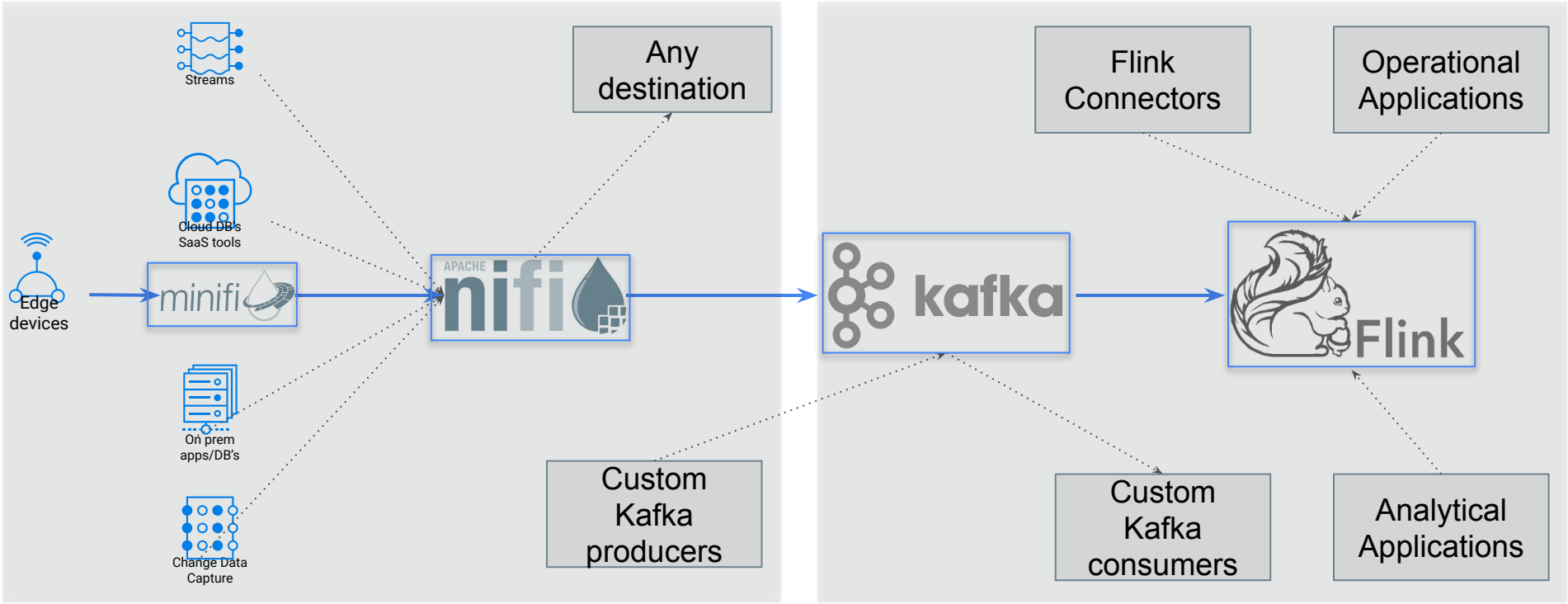
EXAMPLES


Ingest Files
Parse Document
Chunk Document
Put to Vector Store
Store
JSON
JSON
Pinecone
OpenAI
slack

Analytics-in-Stream



Streaming Data Pipelines with Cloudera Data Platform



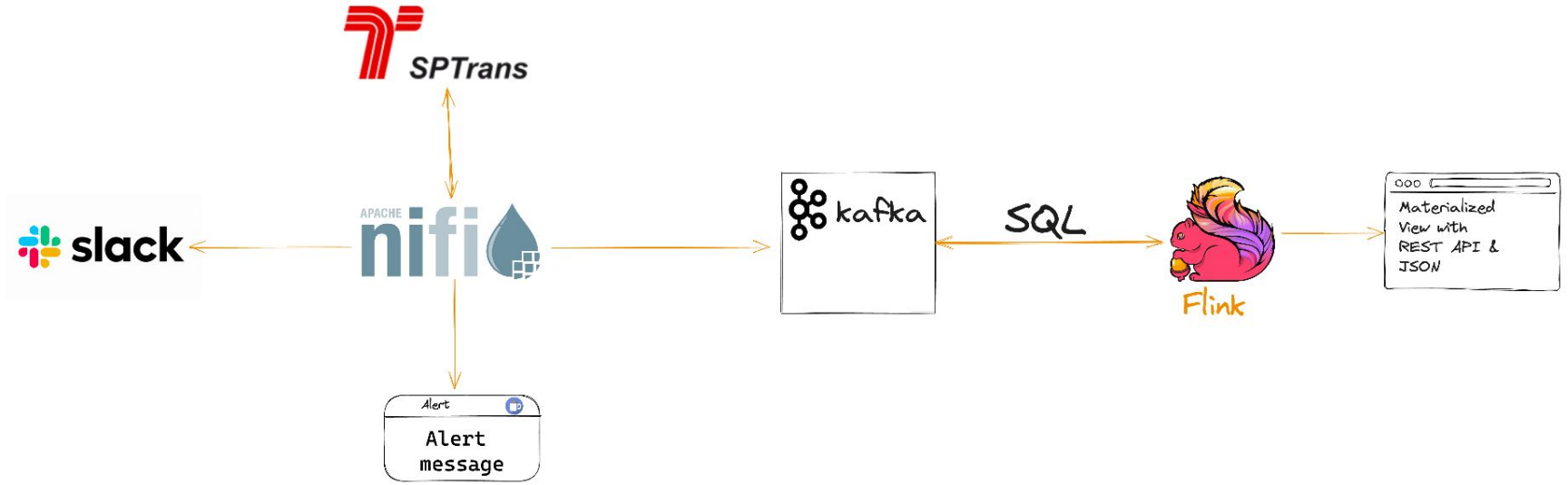
STREAMING DATA MOVEMENT & PROCESSING

STREAMING DATA PROCESSING & ANALYTICS

NiFi/Kafka/Flink - Data Tables - Brazil SPTrans

Show 10 entries

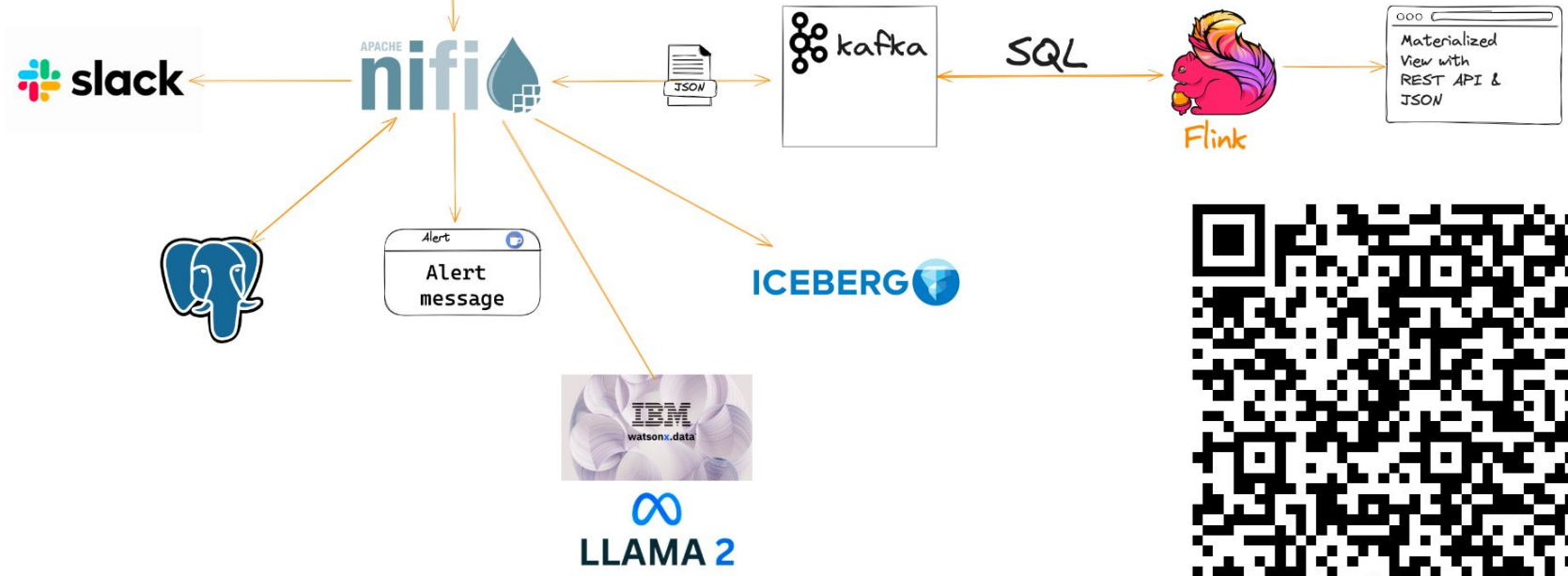
HR	Vehicle	Line ID	Line Origin	Line Destination	Lat/Long
17:08	21434	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.537837,-46.63
17:08	21243	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.529571,-46.59
17:08	21242	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.4884085,-46.5
17:08	21399	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.47630525,-46.5


SPTrans
slack
APACHE
nifi
kafka
SQL
Flink
Materialized
View with
REST API &
JSON
Alert
Alert
message

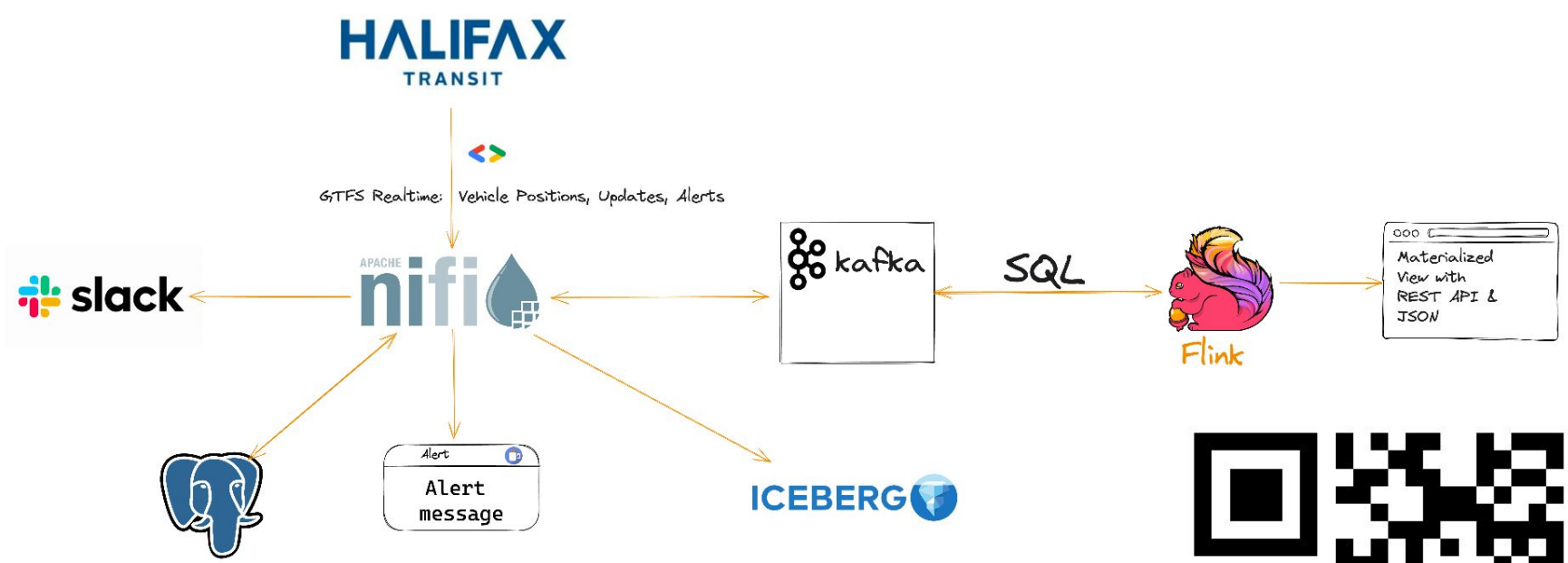
https://github.com/MobilityData/mobility-database-catalogs/

Every Transit System

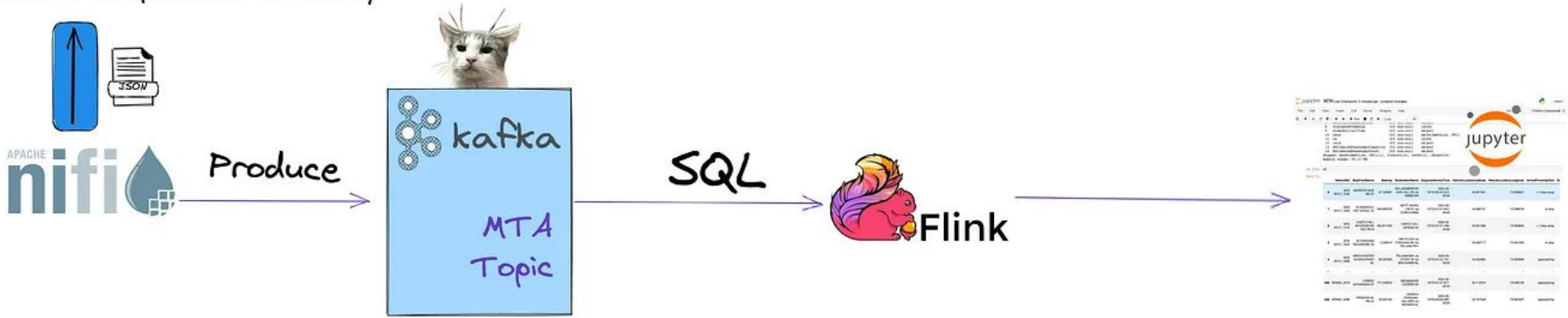
GTFS Realtime: Vehicle Positions, Updates, Alerts



CLOUDERA


HALIFAX
TRANSIT
GTFS Realtime: Vehicle Positions, Updates, Alerts
slack
APACHE
nifi
kafka
SQL
Flink
Materialized
View with
REST API &
JSON
Alert
Alert
message
ICEBERG

Metropolitan Transportation Authority



CLOUDERA

APACHE KAFKA

What is Can You Do With Apache Kafka?

Web site activity: track page views, searches, etc. in real time

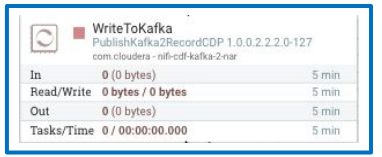
Events & log aggregation: particularly in distributed systems where messages come from multiple sources

Monitoring and metrics: aggregate statistics from distributed applications and build a dashboard application

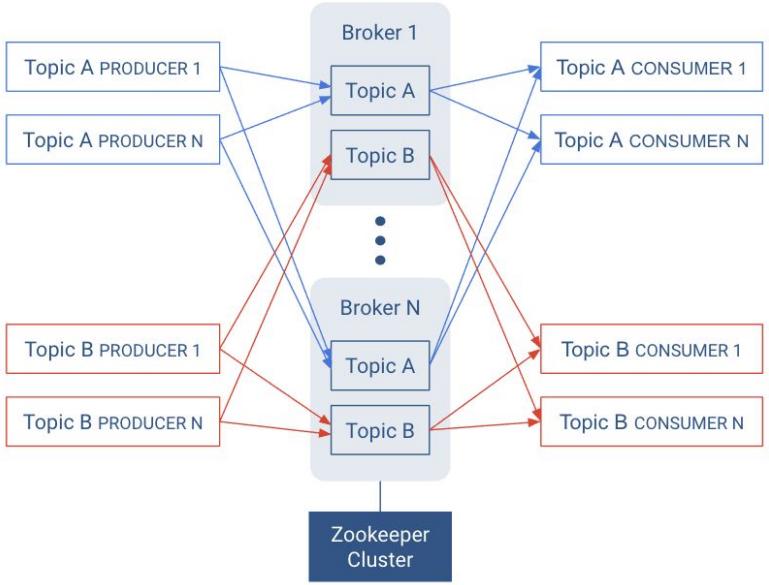
Stream processing: process raw data, clean it up, and forward it on to another topic or messaging system

Real-time data ingestion: fast processing of a very large volume of messages

STREAMS MESSAGING WITH KAFKA



- Highly reliable distributed messaging system.
- Decouple applications, enables many-to-many patterns.
- Publish-Subscribe semantics.
- Horizontal scalability.
- Efficient implementation to operate at speed with big data volumes.
- Organized by topic to support several use cases.

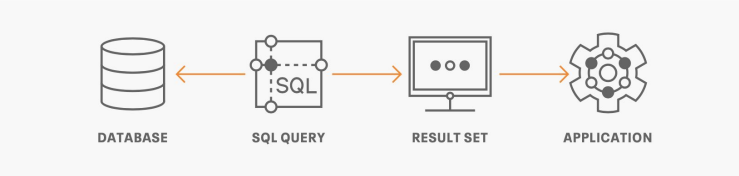


APACHE FLINK

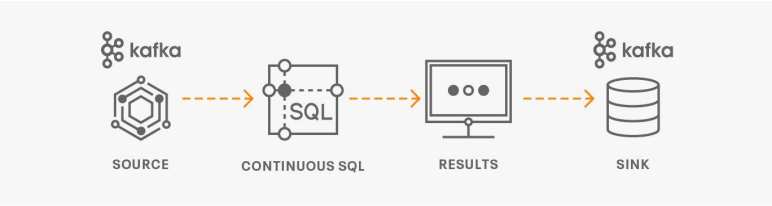
CONTINUOUS SQL

- SSB is a Continuous SQL engine
- It's SQL, but a slightly different mental model, but with big implications

Traditional Parse/Execute/Fetch model



Continuous SQL Model



Hint: The query is boundless and never finishes, and time matters

AKA: `SELECT * FROM foo WHERE 1=0 -- will run forever`

Flink SQL

Key Takeaway: Rich SQL grammar with advanced time and aggregation tools

```
-- specify Kafka partition key on output
SELECT foo AS _eventKey FROM sensors

-- use event time timestamp from kafka
-- exactly once compatible
SELECT eventTimestamp FROM sensors

-- nested structures access
SELECT foo.'bar' FROM table; -- must quote nested
column

-- timestamps
SELECT * FROM payments
WHERE eventTimestamp > CURRENT_TIMESTAMP-interval
'10' second;

-- unnest
SELECT b.*, u.*
FROM bgp_avro b,
UNNEST(b.path) AS u(pathitem)
```

```
-- aggregations and windows
SELECT card,
MAX(amount) as theamount,
TUMBLE_END(eventTimestamp, interval '5' minute) as
ts
FROM payments
WHERE lat IS NOT NULL
AND lon IS NOT NULL
GROUP BY card,
TUMBLE(eventTimestamp, interval '5' minute)
HAVING COUNT(*) > 4  -- >4==fraud

-- try to do this ksql!
SELECT us_west.user_score+ap_south.user_score
FROM kafka_in_zone_us_west us_west
FULL OUTER JOIN kafka_in_zone_ap_south ap_south
ON us_west.user_id = ap_south.user_id;
```

SQL STREAM BUILDER (SSB)

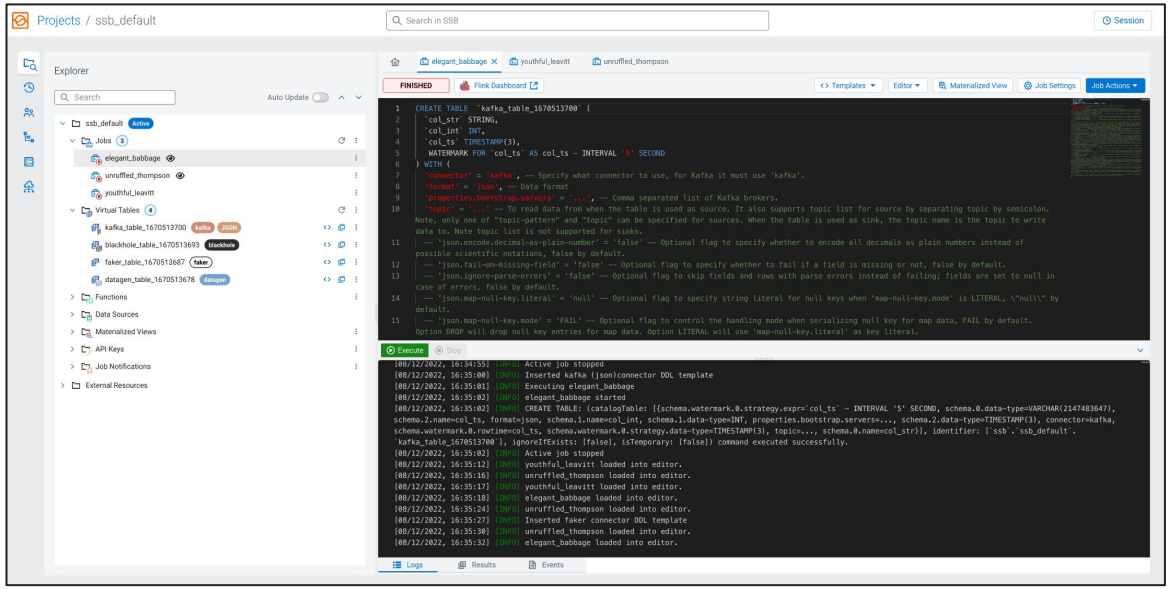
Democratize access to real-time data with just SQL

SQL STREAM BUILDER allows developers, analysts, and data scientists to **write streaming applications** with industry standard **SQL**.

No Java or Scala code development required.

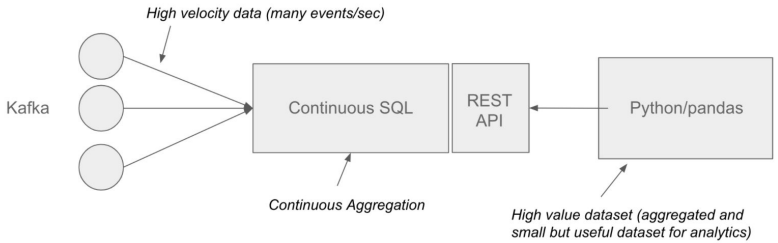
Simplifies access to data in Kafka & Flink. Connectors to batch data in HDFS, Kudu, Hive, S3, JDBC, CDC and more

Enrich streaming data with batch data in a single tool



SSB MATERIALIZED VIEWS

Key Takeaway; MV's allow data scientist, analyst and developers consume data from the firehose




```
SELECT userid,
max(amount) as max_amount,
sum(amount) as sum_amount,
count(*) as thecount,
tumble_end(eventTimestamp, interval '5' second) as ts
FROM authorizations
GROUP BY userid, tumble(eventTimestamp, interval '5' second)
HAVING count(*) > 1
```

```
[90]: import pandas as pd
[91]: mv = "https://xxxyyyyzzz"
[92]: df = pd.read_json(mv)
[93]: len(df.keys())
[93]: 5
[95]: df['ts'] = pd.to_datetime(df['ts'])
[97]: df.dtypes
[97]: max_amount              int64
      sum_amount              int64
      thecount                int64
      ts             datetime64[ns]
      userid                  int64
      dtype: object
[98]: df.set_index('userid').sort_values(by=['thecount'], ascending=False).head()
```

[98]:

userid	max_amount	sum_amount	thecount	ts
787	34911	57304	10	2020-06-16 19:52:15
744	77407	95407	9	2020-06-16 19:52:15
78	86761	330397	9	2020-06-16 19:52:15
541	78762	282682	8	2020-06-16 19:52:15
926	85636	129728	8	2020-06-16 19:52:15

```
[ ]:
```

ICEBERG INTEGRATION

Robust Next Generation Architecture for Data Driven Business

Unified Processing Engine

Massive Open table format

- Maximally open
- Maximally flexible
- Ultra high performance for MASSIVE data

Iceberg Support for Flink APIs through SSB

Feature support	Flink	Notes
SQL create catalog	✔	
SQL create database	✔	
SQL create table	✔	
SQL create table like	✔	
SQL alter table	✔	Only support altering table properties, column and partition changes are not supported
SQL drop_table	✔	
SQL select	✔	Support both streaming and batch mode
SQL insert into	✔	Support both streaming and batch mode
SQL insert overwrite	✔	
DataStream read	✔	
DataStream append	✔	
DataStream overwrite	✔	
Metadata tables		Support Java API but does not support Flink SQL
Rewrite files action	✔	

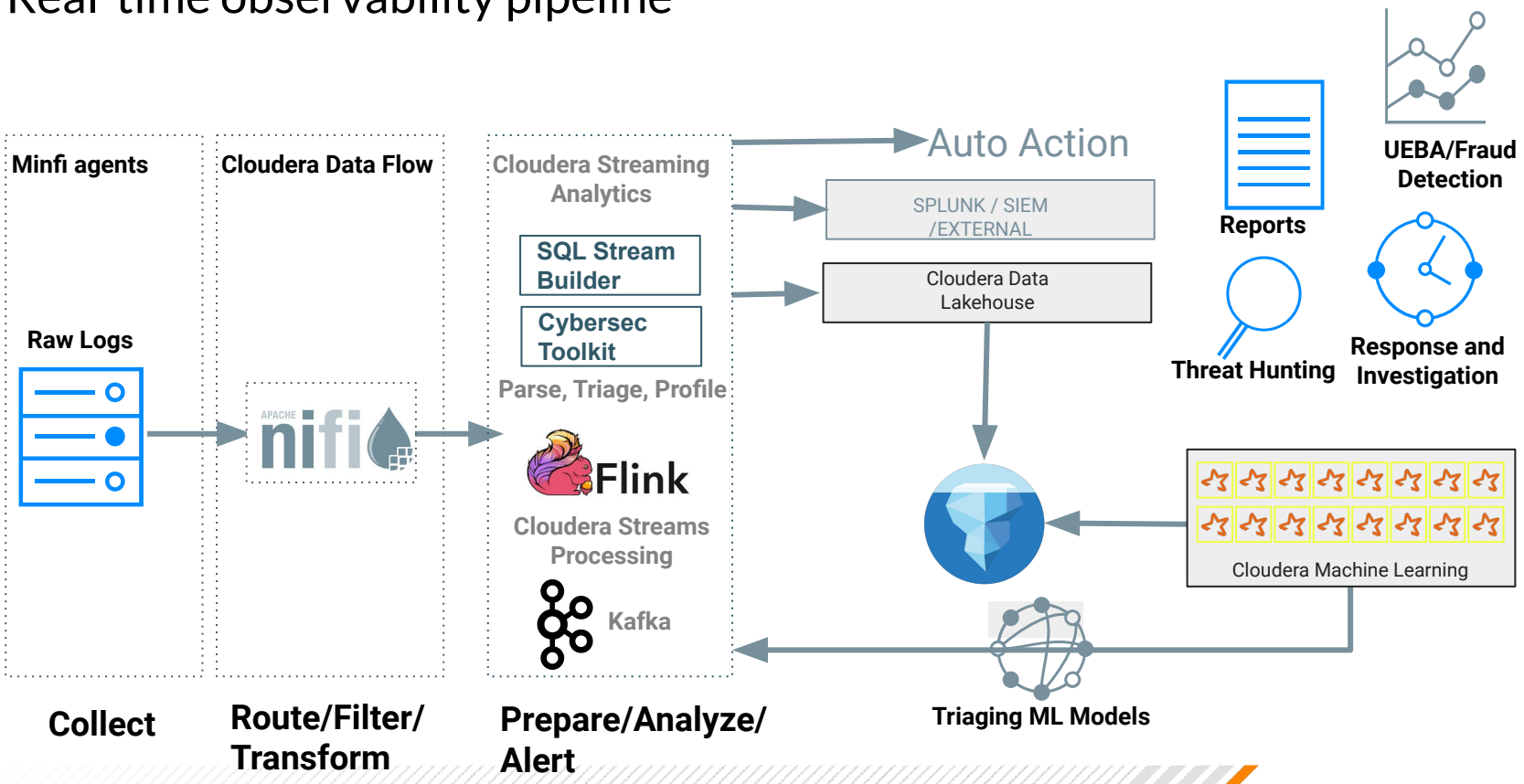
DEMO AND CODE

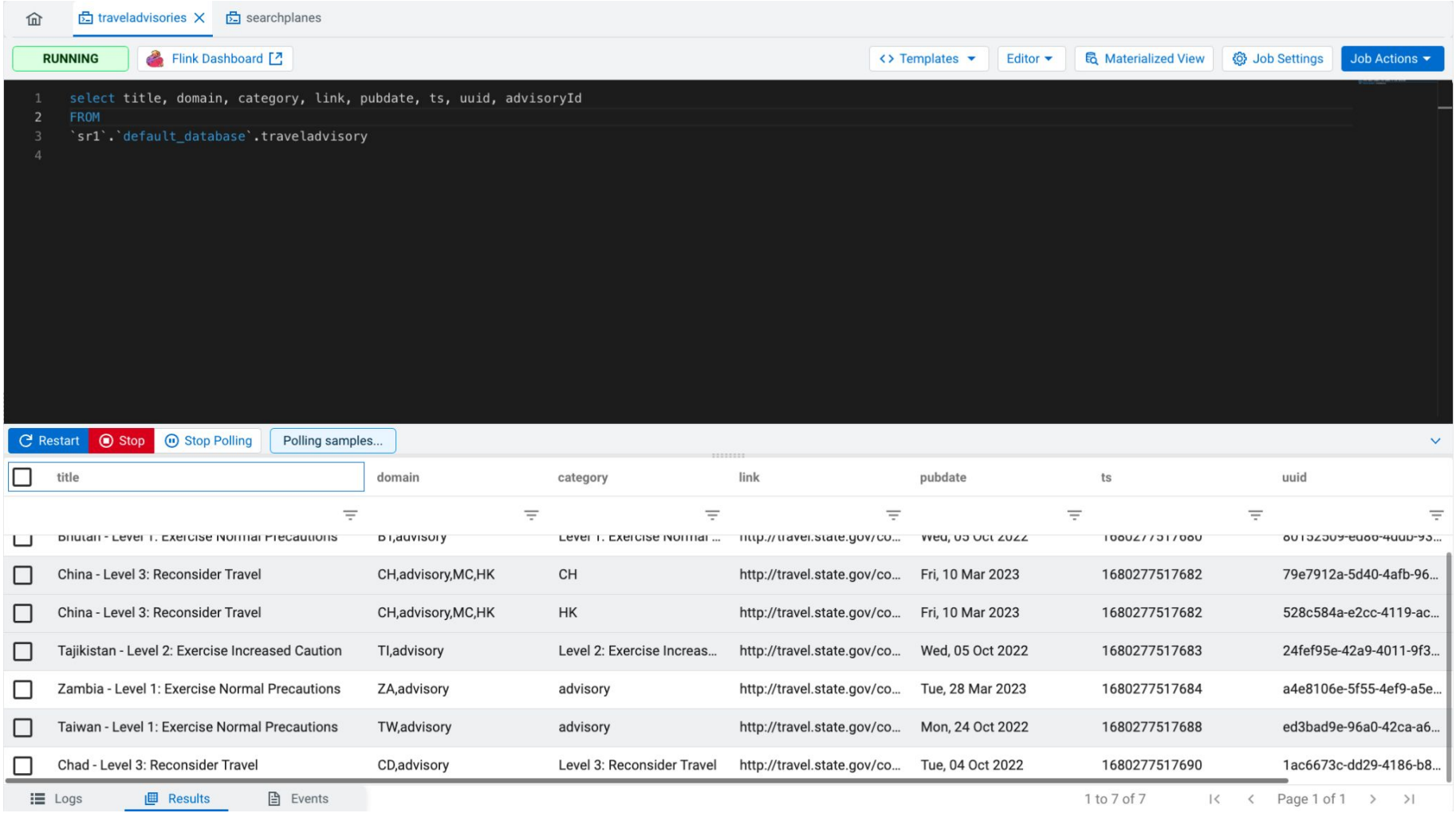
Continuous SQL

```
select max(alt_baro) as MaxAltitudeFeet, min(alt_baro) as MinAltitudeFeet, avg(alt_baro) as AvgAltitudeFeet,
       max(alt_geom) as MaxGAltitudeFeet, min(alt_geom) as MinGAltitudeFeet, avg(alt_geom) as AvgGAltitudeFeet,
       max(gs) as MaxGroundSpeed, min(gs) as MinGroundSpeed, avg(gs) as AvgGroundSpeed,
       count(alt_baro) as RowCount,
       hex as ICAO, flight as IDENT
from `sr1`.`default_database`.`adsb`
group by flight, hex;
```

```
select transcom.title, transcom.description,  mta.VehicleRef,
DISTANCE_BETWEEN(CAST(transcom.latitude as STRING), CAST(transcom.latitude as STRING), mta.VehicleLocationLatitude, mta.VehicleLocationLongitude) as miles,
mta.StopPointName, mta.Bearing, mta.DestinationName, mta.ExpectedArrivalTime, mta.VehicleLocationLatitude, mta.VehicleLocationLongitude,
mta.ArrivalProximityText, mta.DistanceFromStop, mta.AimedArrivalTime, mta.`Date`, mta.ts, mta.uuid, mta.EstimatedPassengerCapacity, mta.EstimatedPassengerCount
from  `schemareg1`.`default_database`.`mta`  /*+ OPTIONS('scan.startup.mode' = 'earliest-offset') */ mta
FULL OUTER JOIN `schemareg1`.`default_database`.`transcom`  /*+ OPTIONS('scan.startup.mode' = 'earliest-offset') */ transcom
ON (transcom.latitude >= CAST(mta.VehicleLocationLatitude as float) - 0.3)
AND (transcom.longitude >= CAST(mta.VehicleLocationLongitude as float) - 0.3)
AND (transcom.latitude <= CAST(mta.VehicleLocationLatitude as float) + 0.3)
AND (transcom.longitude <= CAST(mta.VehicleLocationLongitude as float) + 0.3)
WHERE  mta.VehicleRef is not null
AND transcom.title is not null
AND DISTANCE_BETWEEN(CAST(transcom.latitude as STRING), CAST(transcom.latitude as STRING), mta.VehicleLocationLatitude, mta.VehicleLocationLongitude) <= 120
```

Real-time observability pipeline




traveladvisories
searchplanes
RUNNING
Flink Dashboard
Templates
Editor
Materialized View
Job Settings
Job Actions
select title, domain, category, link, pubdate, ts, uuid, advisoryId
FROM
`sr1`.`default_database`.traveladvisory
Restart
Stop
Stop Polling
Polling samples...
title
domain
category
link
pubdate
ts
uuid
China - Level 3: Reconsider Travel
CH,advisory,MC,HK
CH
http://travel.state.gov/co...
Fri, 10 Mar 2023
1680277517682
79e7912a-5d40-4afb-96...
China - Level 3: Reconsider Travel
CH,advisory,MC,HK
HK
http://travel.state.gov/co...
Fri, 10 Mar 2023
1680277517682
528c584a-e2cc-4119-ac...
Tajikistan - Level 2: Exercise Increased Caution
TI,advisory
Level 2: Exercise Increas...
http://travel.state.gov/co...
Wed, 05 Oct 2022
1680277517683
24fef95e-42a9-4011-9f3...
Zambia - Level 1: Exercise Normal Precautions
ZA,advisory
advisory
http://travel.state.gov/co...
Tue, 28 Mar 2023
1680277517684
a4e8106e-5f55-4ef9-a5e...
Taiwan - Level 1: Exercise Normal Precautions
TW,advisory
advisory
http://travel.state.gov/co...
Mon, 24 Oct 2022
1680277517688
ed3bad9e-96a0-42ca-a6...
Chad - Level 3: Reconsider Travel
CD,advisory
Level 3: Reconsider Travel
http://travel.state.gov/co...
Tue, 04 Oct 2022
1680277517690
1ac6673c-dd29-4186-b8...
Logs
Results
Events
1 to 7 of 7
Page 1 of 1

Materialized View

Configuration

Primary Key

uuid

Retention (Seconds)

Min Row Retention Count

10000

API Key

traveladvisory1

Enable MV

Recreate on Job Start

Ignore NULLs

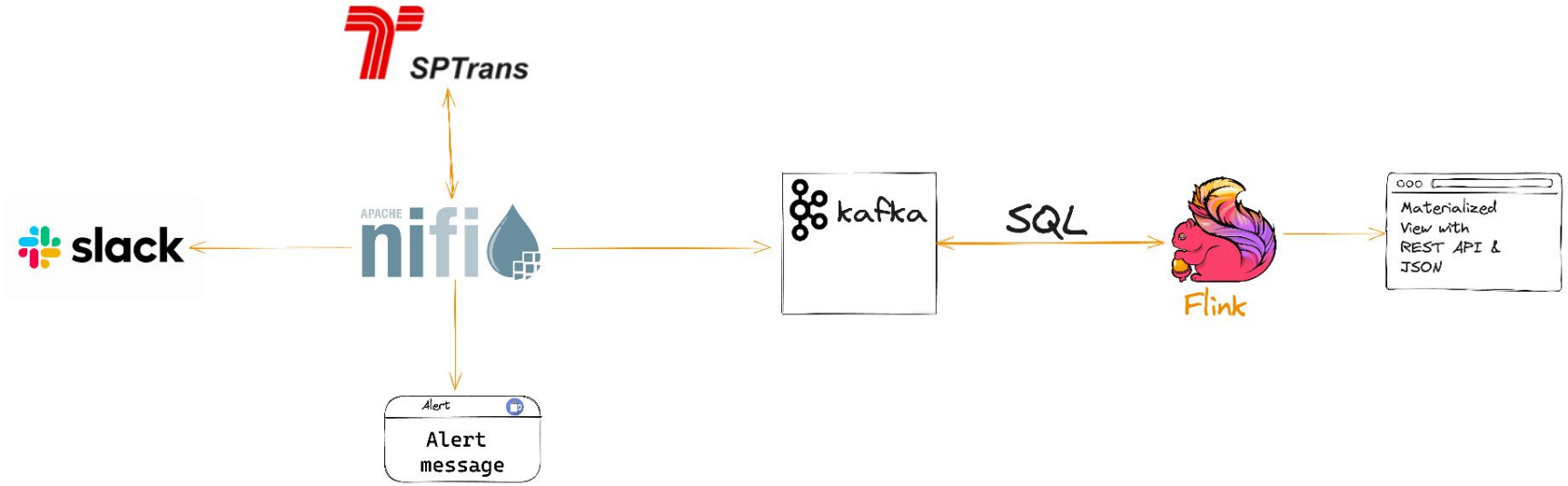
Queries

Add New Query

/api/v1/query/5201/travel?key=66ba91a9-507f-422c-bbb4-86250a9f7bb1&limit=100

Data in Motion: Overview e Novidades do NiFi, Kafka e Flink

Apresentador: Tim Spann - Principal DIM Specialist and Developer Advocate



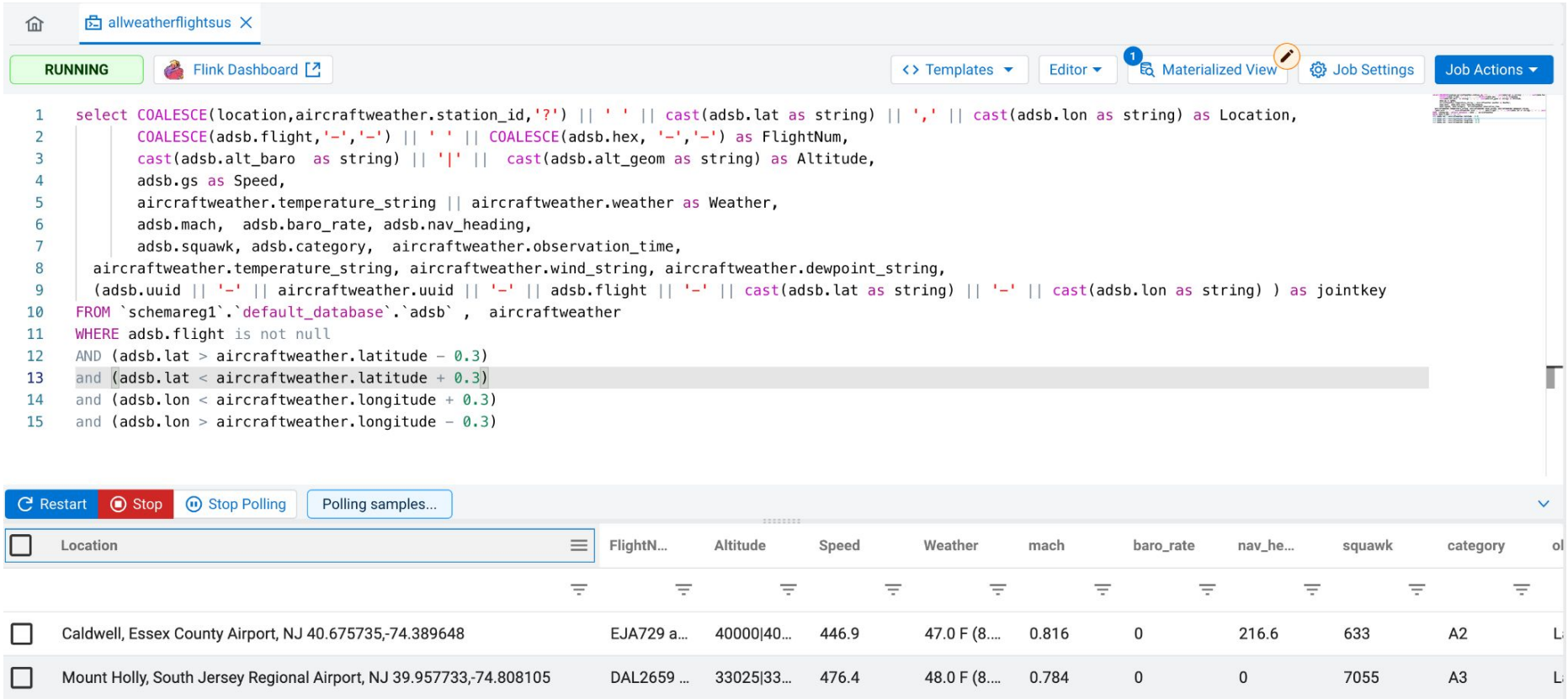
https://medium.com/cloudera-inc/transit-in-sao-paulo-brasil-flank-style-eaec6753cc63

NiFi/Kafka/Flink - Data Tables - Brazil SPTrans

Show 10 entries

Search:

HR	Vehicle	Line ID	Line Origin	Line Destination	Lat/Long	Date/Time
17:08	21434	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.537837,-46.6328475	2023-09-08T20:07:30Z
17:08	21243	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.529571,-46.5984615	2023-09-08T20:07:31Z
17:08	61677	32840	PQ. RES. COCAIA	PQ. IBIRAPUERA	-23.6532785,-46.7017075	2023-09-08T20:07:35Z
17:08	61683	32840	PQ. RES. COCAIA	PQ. IBIRAPUERA	-23.718092,-46.699059	2023-09-08T20:07:20Z
17:08	61517	32840	PQ. RES. COCAIA	PQ. IBIRAPUERA	-23.58114725,-46.6574995	2023-09-08T20:07:28Z
17:08	41014	33514	VL. DALILA	TERM. PQ. D. PEDRO II	-23.5383225,-46.563772	2023-09-08T20:08:04Z
17:08	41019	33514	VL. DALILA	TERM. PQ. D. PEDRO II	-23.5443805,-46.5217695	2023-09-08T20:07:45Z


allweatherflightsus
RUNNING
Flink Dashboard
Templates
Editor
Materialized View
Job Settings
Job Actions
select COALESCE(location,aircraftweather.station_id,'?') || ' ' || cast(adsb.lat as string) || ',' || cast(adsb.lon as string) as Location,
COALESCE(adsb.flight,'-','-') || ' ' || COALESCE(adsb.hex, '-','-') as FlightNum,
cast(adsb.alt_baro as string) || '|' || cast(adsb.alt_geom as string) as Altitude,
adsb.gs as Speed,
aircraftweather.temperature_string || aircraftweather.weather as Weather,
adsb.mach, adsb.baro_rate, adsb.nav_heading,
adsb.squawk, adsb.category, aircraftweather.observation_time,
aircraftweather.temperature_string, aircraftweather.wind_string, aircraftweather.dewpoint_string,
(adsb.uuid || '-' || aircraftweather.uuid || '-' || adsb.flight || '-' || cast(adsb.lat as string) || '-' || cast(adsb.lon as string)) as jointkey
FROM `schemareg1`.`default_database`.`adsb` , aircraftweather
WHERE adsb.flight is not null
AND (adsb.lat > aircraftweather.latitude - 0.3)
and (adsb.lat < aircraftweather.latitude + 0.3)
and (adsb.lon < aircraftweather.longitude + 0.3)
and (adsb.lon > aircraftweather.longitude - 0.3)
Restart
Stop
Stop Polling
Polling samples...
Location
FlightN...
Altitude
Speed
Weather
mach
baro_rate
nav_he...
squawk
category
Caldwell, Essex County Airport, NJ 40.675735,-74.389648
EJA729 a...
40000|40...
446.9
47.0 F (8....
0.816
0
216.6
633
A2
Mount Holly, South Jersey Regional Airport, NJ 39.957733,-74.808105
DAL2659 ...
33025|33...
476.4
48.0 F (8....
0.784
0
0
7055
A3

CDC ENGINE SELECTION HOW TO DO IT?

Kafka Connect, NiFi, Flink? Which engine to choose? Or All 3?

Already using Kafka?

Simple setup for many tables

Want metadata augmented data

Don't need low latency?

Visual monitoring

Easy manual scaling

Easy to combine with NiFi

Debezium

Already using NiFi?

Simple JDBC queries?

Transform individual records?

Want easy development with UI?

Lots of small files, events, records, rows?

Continuous stream of rows

Support many different sources

Debezium coming

Need for Fast Flink?

Strong control of table and joins

Want high Throughput?

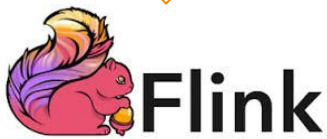
Want Low Latency?

Want Advanced Windowing and State?

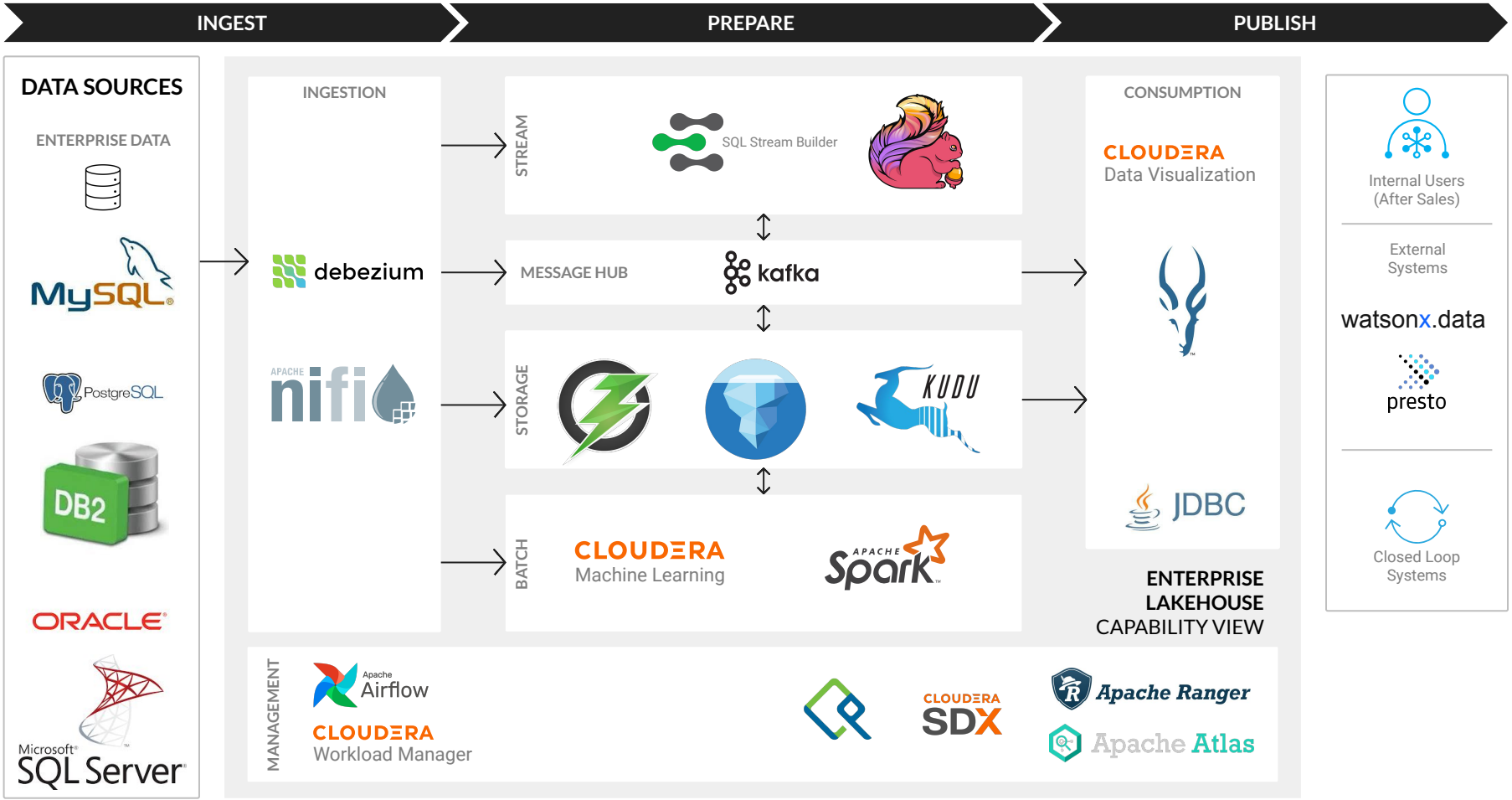
Automatic records immediately

Pure SQL

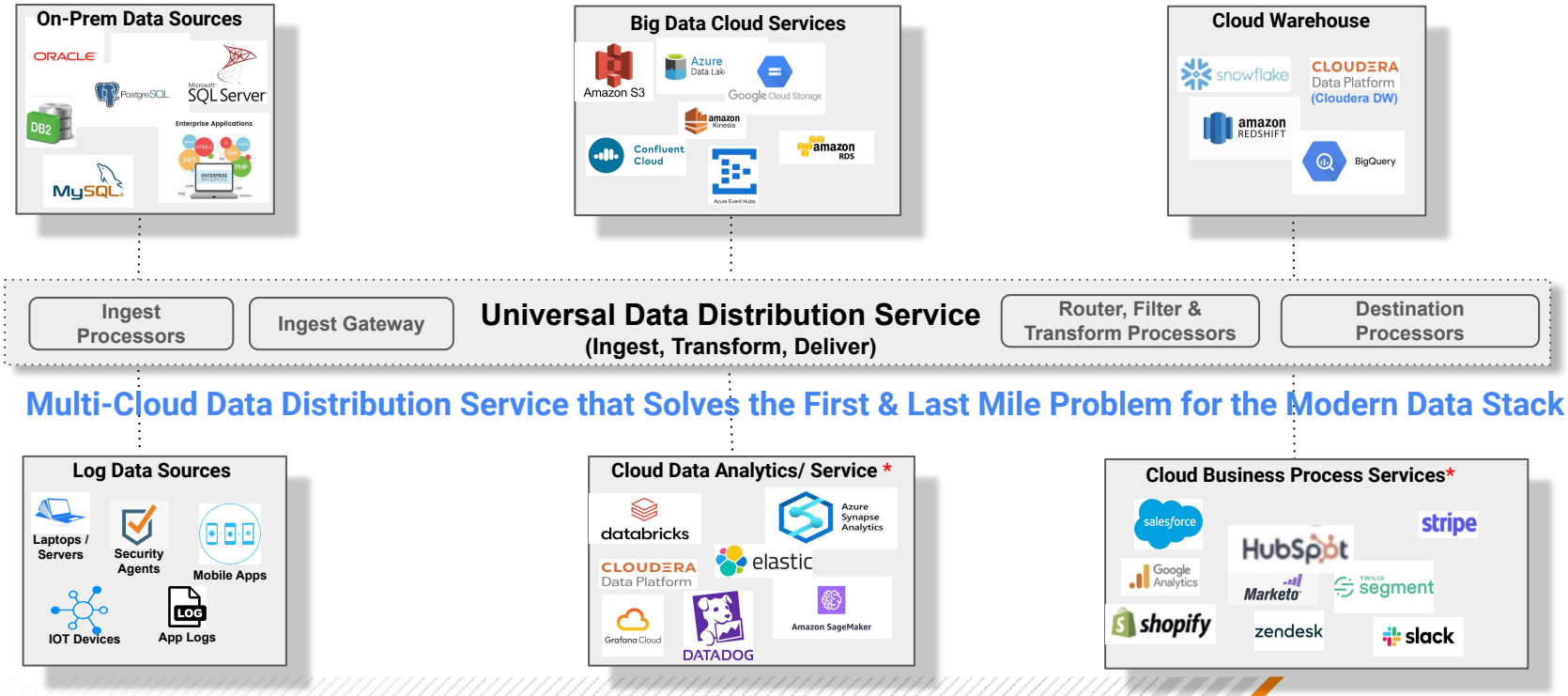
Debezium



CDC ARCHITECTURE - Using FLaNK to pull the data out of anything in near-real time

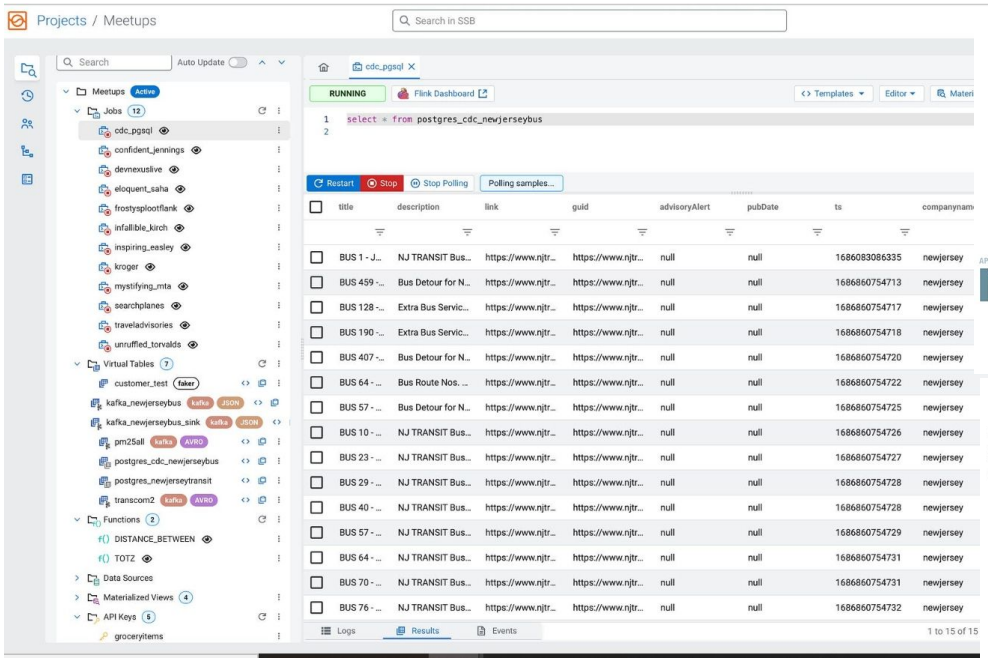


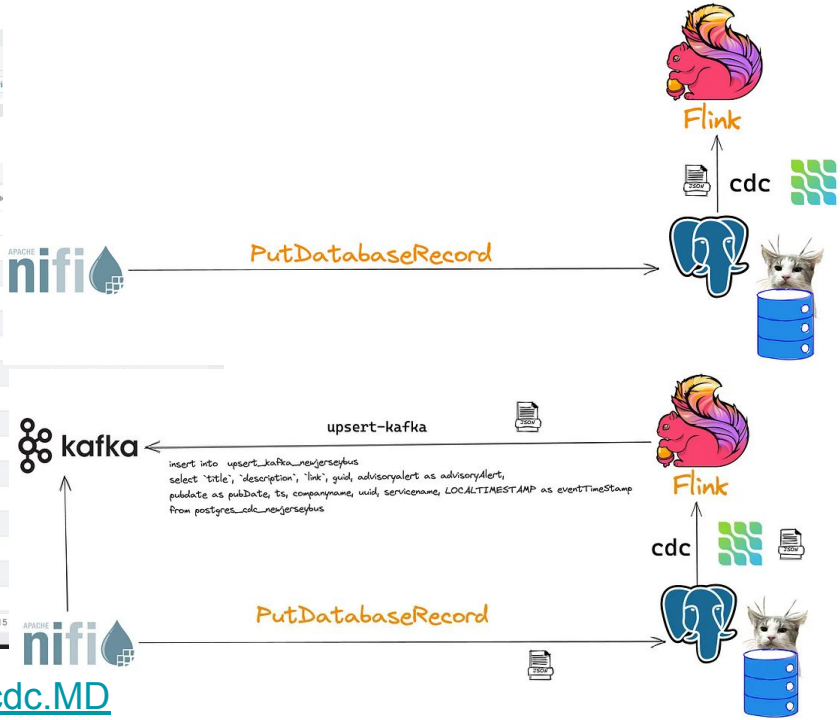
Data Distribution as a Universal, Hybrid, Multi-Cloud Data Service



CDC with SQL Stream Builder (Flink SQL)

Streaming CDC with Cloudera SQL Stream Builder (Flink SQL)

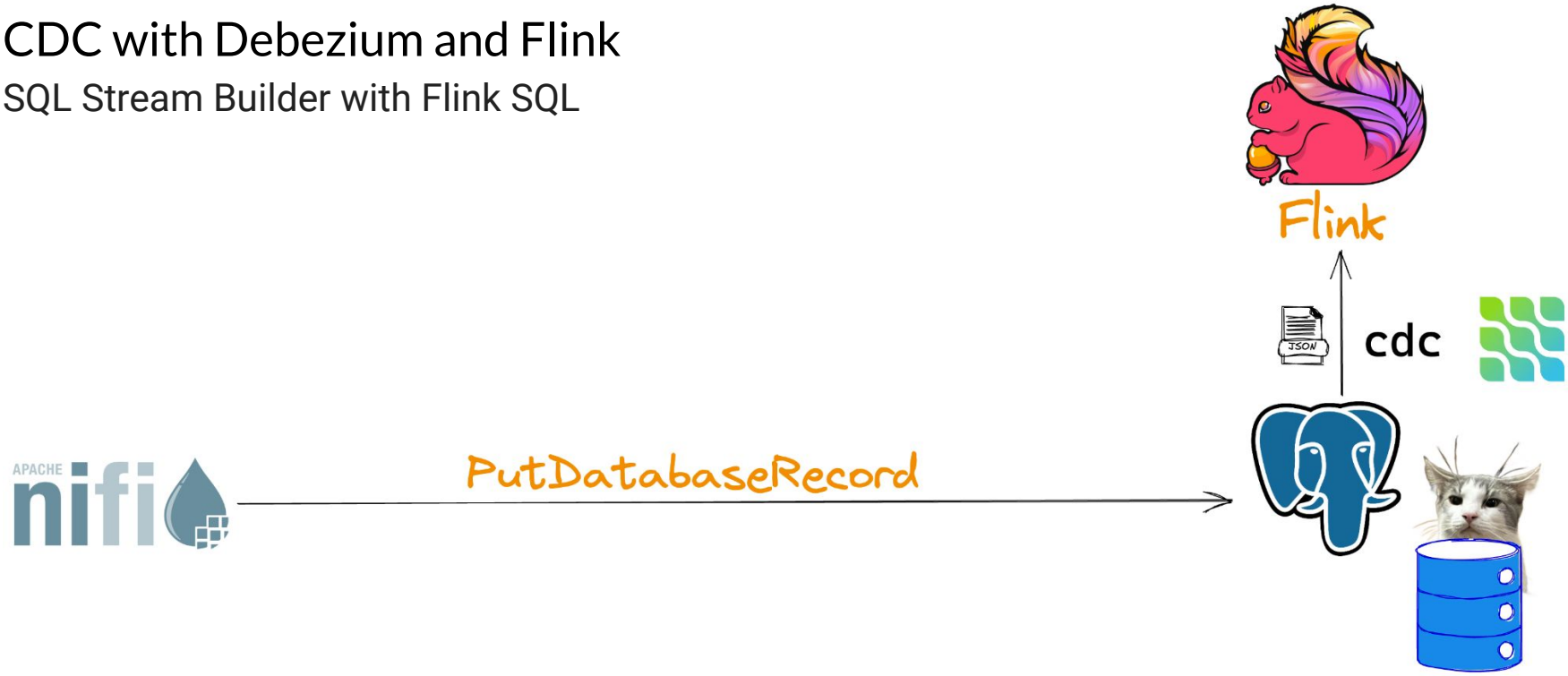




https://github.com/tspannhw/FLaNK-CDC/blob/main/flinkcdc.MD

CDC with Debezium and Flink

SQL Stream Builder with Flink SQL



https://docs.cloudera.com/csa/1.10.0/how-to-ssb/topics/csa-ssb-cdc-connectors.html

CDC with Debezium and Flink

SQL Stream Builder with Flink SQL

```
1  select * from postgres_cdc_newjerseybus
```

Execute | Stop

	title	description	link	guid	advisoryalert	pubdate	ts	companyname
☐	BUS 707 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074228	newjersey
☐	BUS 755 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074229	newjersey
☐	BUS 804 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074231	newjersey
☐	BUS 834 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074234	newjersey
☐	BUS 127 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074208	newjersey
☐	BUS 148 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074211	newjersey
☐	BUS 196 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074215	newjersey
☐	BUS 346 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074217	newjersey
☐	BUS 409 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074220	newjersey
☐	BUS 455 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074221	newjersey
☐	BUS 606 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074225	newjersey
☐	BUS 709 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074228	newjersey
☐	BUS 803 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074231	newjersey
☐	BUS 822 - Aug 16, 2023 03:39:35 PM	New NJ TRANSIT Bus Schedules – Effective Saturday, September 2, 2023	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 16, 2023 03:39:35 P...	1694185074234	newjersey
☐	BUS 873 - Aug 08, 2023 11:15:54 AM	NJ TRANSIT to Pilot Contactless Tap to Pay – Effective Immediately	https://www.njtransit.co...	https://www.njtransit.co...	null	Aug 08, 2023 11:15:54 A...	1694185074237	newjersey

postgres_cdc_newjerseybus
Virtual Table

Details

TYPE:	ssb

Schema

Column	Type
title	STRING
description	STRING
link	STRING
guid	STRING
advisoryalert	STRING
pubdate	STRING
ts	STRING
companyname	STRING
uuid	STRING
servicename	STRING

DDL

```
CREATE TABLE `ssb`.`Meetups`.`postgres_cdc_newjerseybus` (
  `title` VARCHAR(2147483647),
  `description` VARCHAR(2147483647),
  `link` VARCHAR(2147483647),
  `guid` VARCHAR(2147483647),
  `advisoryalert` VARCHAR(2147483647),
  `pubdate` VARCHAR(2147483647),
  `ts` VARCHAR(2147483647),
  `companyname` VARCHAR(2147483647),
  `uuid` VARCHAR(2147483647),
  `servicename` VARCHAR(2147483647)
) WITH (
  'hostname' = '192.168.1.153',
  'password' = '******',
  'decoding.plugin.name' = 'pgoutput',
  'connector' = 'postgres-cdc',
  'port' = '5432',
  'database-name' = 'tspann',
  'schema-name' = 'public',
  'table-name' = 'newjerseybus',
  'username' = 'tspann'
)

```

Flink SQL Tables - Debezium CDC From Database Tables

```
CREATE TABLE `postgres_cdc_newjerseybus` (
    `title` STRING,
    `description` STRING,
    `link` STRING,
    `guid` STRING,
    `advisoryAlert` STRING,
    `pubDate` STRING,
    `ts` STRING,
    `companyname` STRING,
    `uuid` STRING,
    `servicename` STRING
) WITH (
  'connector' = 'postgres-cdc',
  'database-name' = 'tspann',
  'hostname' = '192.168.1.153',
  'password' = 'tspann',
  'decoding.plugin.name' = 'pgoutput',
  'schema-name' = 'public',
  'table-name' = 'newjerseybus',
  'username' = 'tspann',
  'port' = '5432'
  );
```

Flink SQL Tables - Upsert to Kafka Topics

```
CREATE TABLE  `upsert_kafka_newjerseybus` (
  `title` String,
  `description` String,
  `link` String,
  `guid` String,
  `advisoryAlert` String,
  `pubDate` String,
  `ts` String,
  `companyname` String,
  `uuid` String,
  `servicename` String,
   `eventTimestamp` TIMESTAMP(3),
   WATERMARK FOR `eventTimestamp` AS `eventTimestamp` - INTERVAL '5' SECOND,
   PRIMARY KEY (uuid) NOT ENFORCED
) WITH (
  'connector' = 'upsert-kafka',
  'topic' = 'kafka newjerseybus',
  'properties.bootstrap.servers' = 'kafka:9092',
  'key.format' = 'json',
  'value.format' = 'json'
  );
```

RESOURCES/WRAP-UP

```
SELECT n.speed, n.travel time, n.borough, n.link name, n.link points,
       n.latitude, n.longitude, DISTANCE BETWEEN(CAST(t.latitude as STRING),
                 CAST(t.latitude as STRING),
                 m.VehicleLocationLatitude, m.VehicleLocationLongitude) as miles,
       t.title, t.`description`, t.pubDate, t.latitude, t.longitude,
       m.VehicleLocationLatitude, m.VehicleLocationLongitude,
       m.StopPointRef, m.VehicleRef,
       m.ProgressRate, m.ExpectedDepartureTime, m.StopPoint,
       m.VisitNumber, m.DataFrameRef, m.StopPointName,
       m.Bearing, m.OriginAimedDepartureTime, m.OperatorRef,
       m.DestinationName, m.ExpectedArrivalTime, m.BlockRef,
       m.LineRef, m.DirectionRef, m.ArrivalProximityText,
       m.DistanceFromStop, m.EstimatedPassengerCapacity,
       m.AimedArrivalTime, m.PublishedLineName,
       m.ProgressStatus, m.DestinationRef, m.EstimatedPassengerCount,
       m.OriginRef, m.NumberOfStopsAway, m.ts
FROM  jsonmta /*+ OPTIONS('scan.startup.mode' = 'earliest-offset') */  m
      FULL OUTER JOIN jsontranscom /*+ OPTIONS('scan.startup.mode' = 'earliest-offset') */  t
       ON (t.latitude >= CAST(m.VehicleLocationLatitude as float) - 0.3)
       AND (t.longitude >= CAST(m.VehicleLocationLongitude as float) - 0.3)
       AND (t.latitude <= CAST(m.VehicleLocationLatitude as float) + 0.3)
       AND (t.longitude <= CAST(m.VehicleLocationLongitude as float) + 0.3)
     FULL OUTER JOIN nytrafficspeed /*+ OPTIONS('scan.startup.mode' = 'earliest-offset') */  n
     ON (n.latitude >= CAST(m.VehicleLocationLatitude as float) - 0.3)
     AND (n.longitude >= CAST(m.VehicleLocationLongitude as float) - 0.3)
     AND (n.latitude <= CAST(m.VehicleLocationLatitude as float) + 0.3)
     AND (n.longitude <= CAST(m.VehicleLocationLongitude as float) + 0.3)
WHERE m.VehicleRef is not null
  AND   t.title is not null
```

https://github.com/tspannhw/FLaNK-Transit

https://medium.com/@tspann/cdc-not-cat-data-capture-e43713879c03

FLaNK for Halifax Canada Transit — NiFi, Kafka, Flink, SQL, GTFS-RT | by Tim Spann | Cloudera | Dec, 2023 | Medium

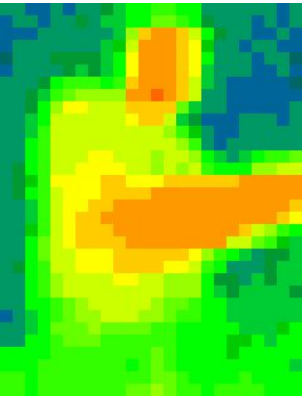
Never Get Lost in the Stream. NiFi-Kafka-Flink for getting to work… | by Tim Spann | Cloudera | Dec, 2023 | Medium

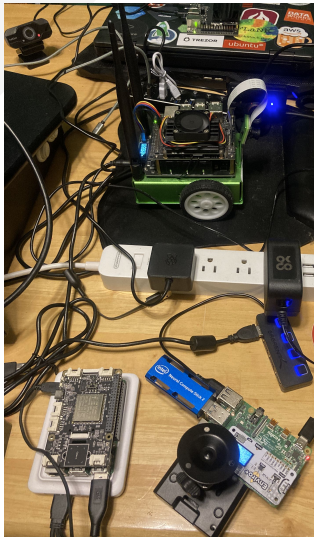
Iteration 1: Building a System to Consume All the Real-Time Transit Data in the World At Once | by Tim Spann | Cloudera | Medium

Watching Airport Traffic in Real-Time | by Tim Spann | Cloudera | Medium

Resources

https://medium.com/@tspann/llm-pipelines-with-pinecone-and-huggingface-with-python-and-apache-nifi-a96c20be93b7





THANK YOU