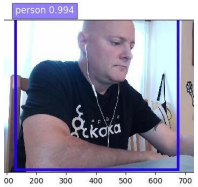
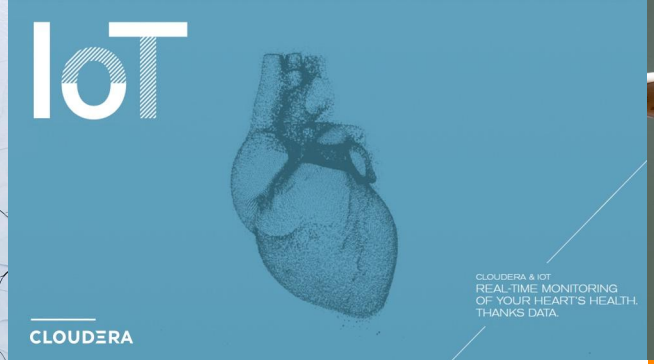
CLOUDERA

Unlocking Financial Data with Real-Time Pipelines (Flink Analytics on Stocks with SQL)

Tim Spann
Principal Developer Advocate

28-February-2024



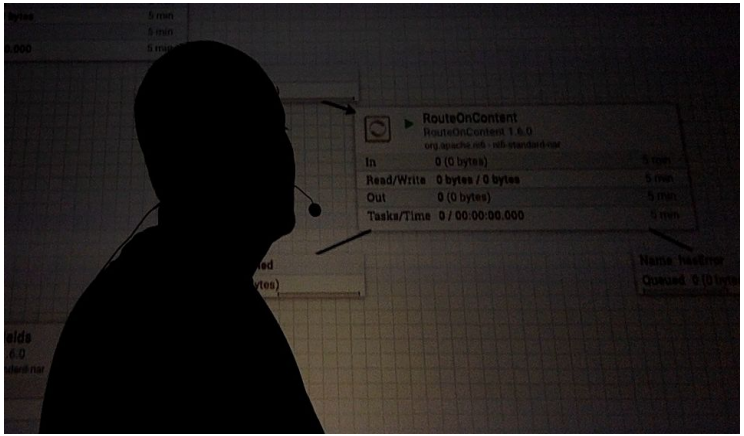


CLOUDERA

EDGE
2AI
CLOUDERA

ENTERPRISE
DATA
CLOUD
CLOUDERA

Agenda (45 minutes)

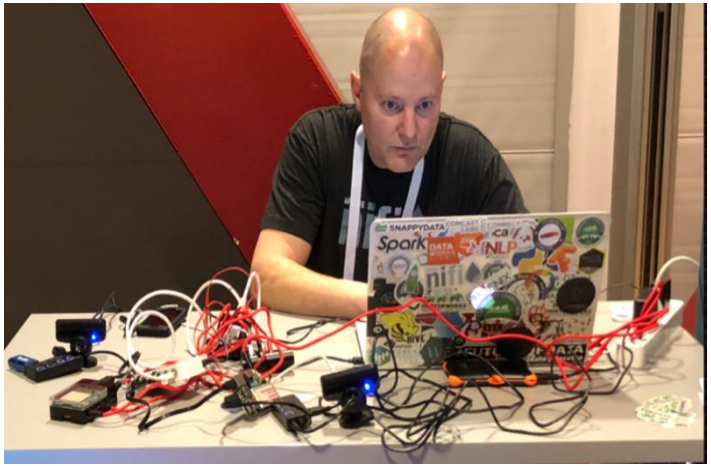


- Introduction
- Overview
- Finance Data
- Apache Kafka and Apache Flink
- Demos

FLaNK
flankstack.dev

Financial institutions thrive on accurate and timely data to drive critical decision-making processes, risk assessments, and regulatory compliance. However, managing and processing vast amounts of financial data in real-time can be a daunting task. To overcome this challenge, modern data engineering solutions have emerged, combining powerful technologies like Apache Flink, Apache NiFi, Apache Kafka, and Iceberg to create efficient and reliable real-time data pipelines. In this talk, we will explore how this technology stack can unlock the full potential of financial data, enabling organizations to make data-driven decisions swiftly and with confidence.

Introduction: Financial institutions operate in a fast-paced environment where real-time access to accurate and reliable data is crucial. Traditional batch processing falls short when it comes to handling rapidly changing financial markets and responding to customer demands promptly. In this talk, we will delve into the power of real-time data pipelines, utilizing the strengths of Apache Flink, Apache NiFi, Apache Kafka, and Iceberg, to unlock the potential of financial data. I will be utilizing NiFi 2.0 with Python and Vector Databases.



Tim Spann

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

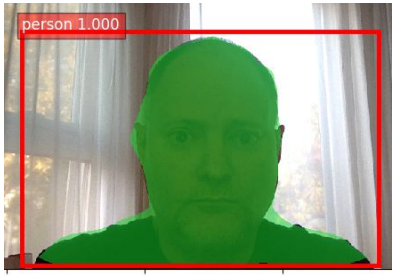
Principal Developer Advocate.

Princeton Future of Data Meetup.

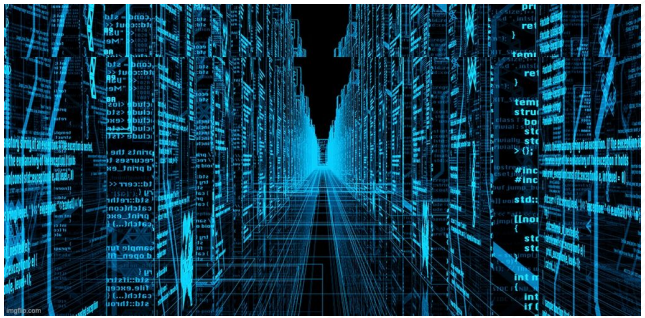
ex-Pivotal, ex-Hortonworks, ex-StreamNative, ex-PwC

https://medium.com/@tspann

https://github.com/tspannhw



Future of Data - NYC + NJ + Philly + Virtual



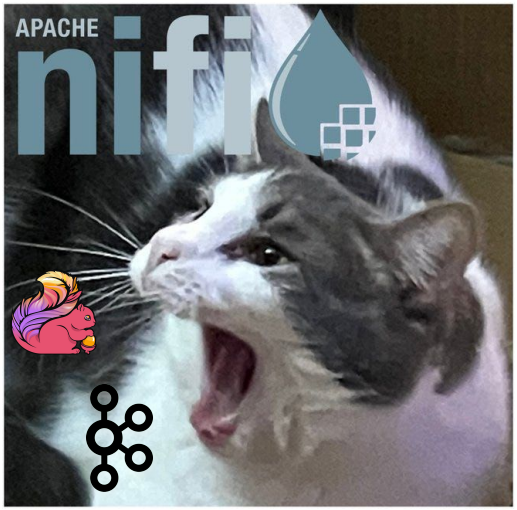
https://www.meetup.com/futureofdata-princeton/

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

FLaNK Stack Weekly by Tim Spann

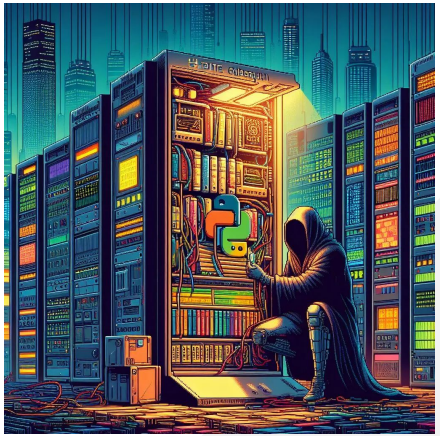
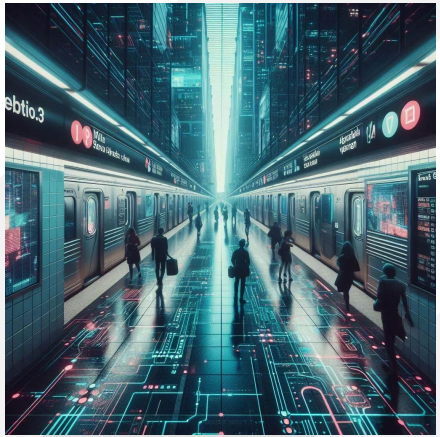


https://bit.ly/32dAJft

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

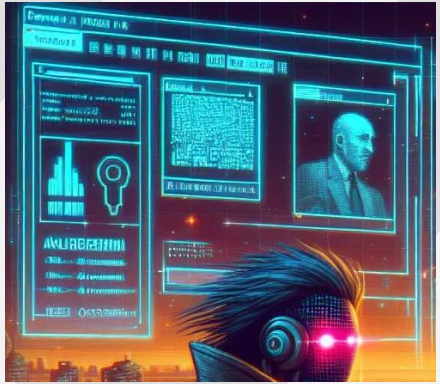
This week in Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java, LLM, GenAI, Vector DB and Open Source friends.



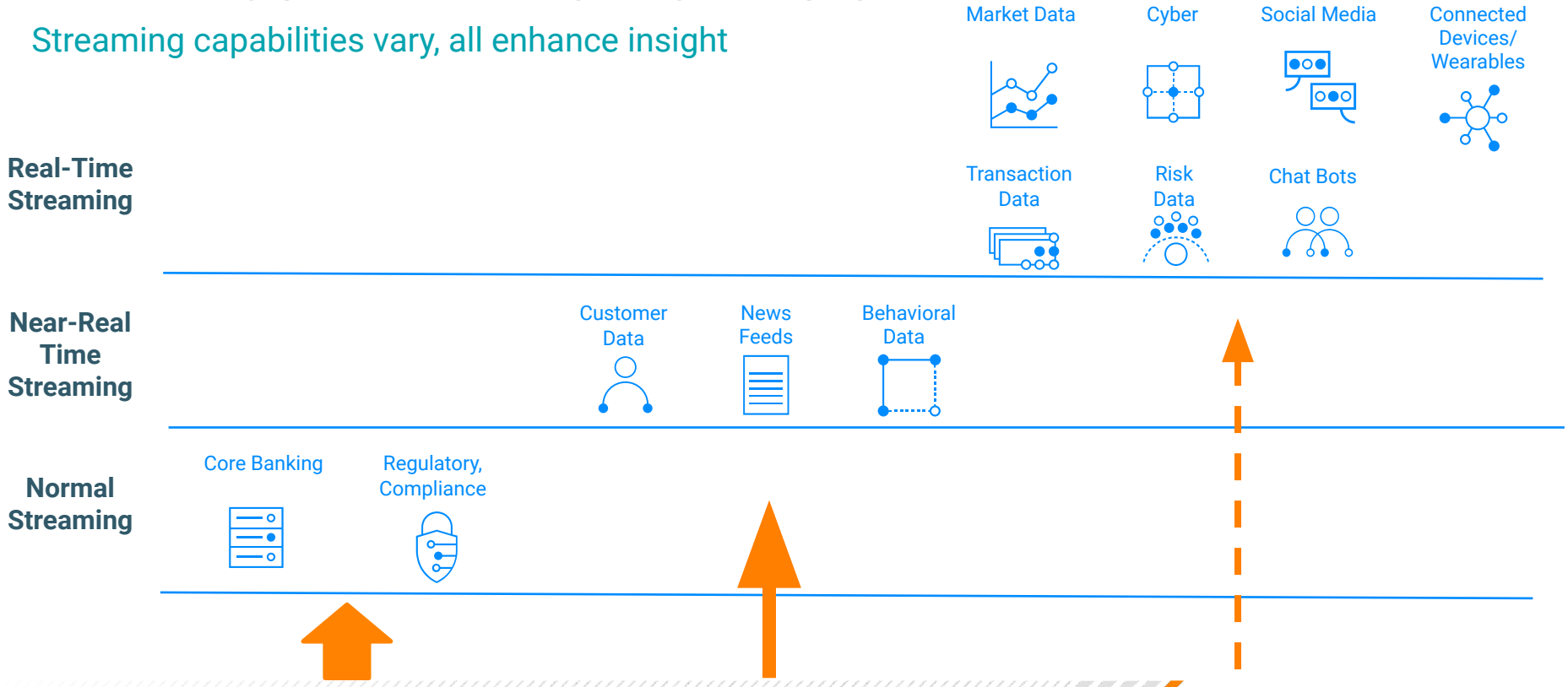
Overview



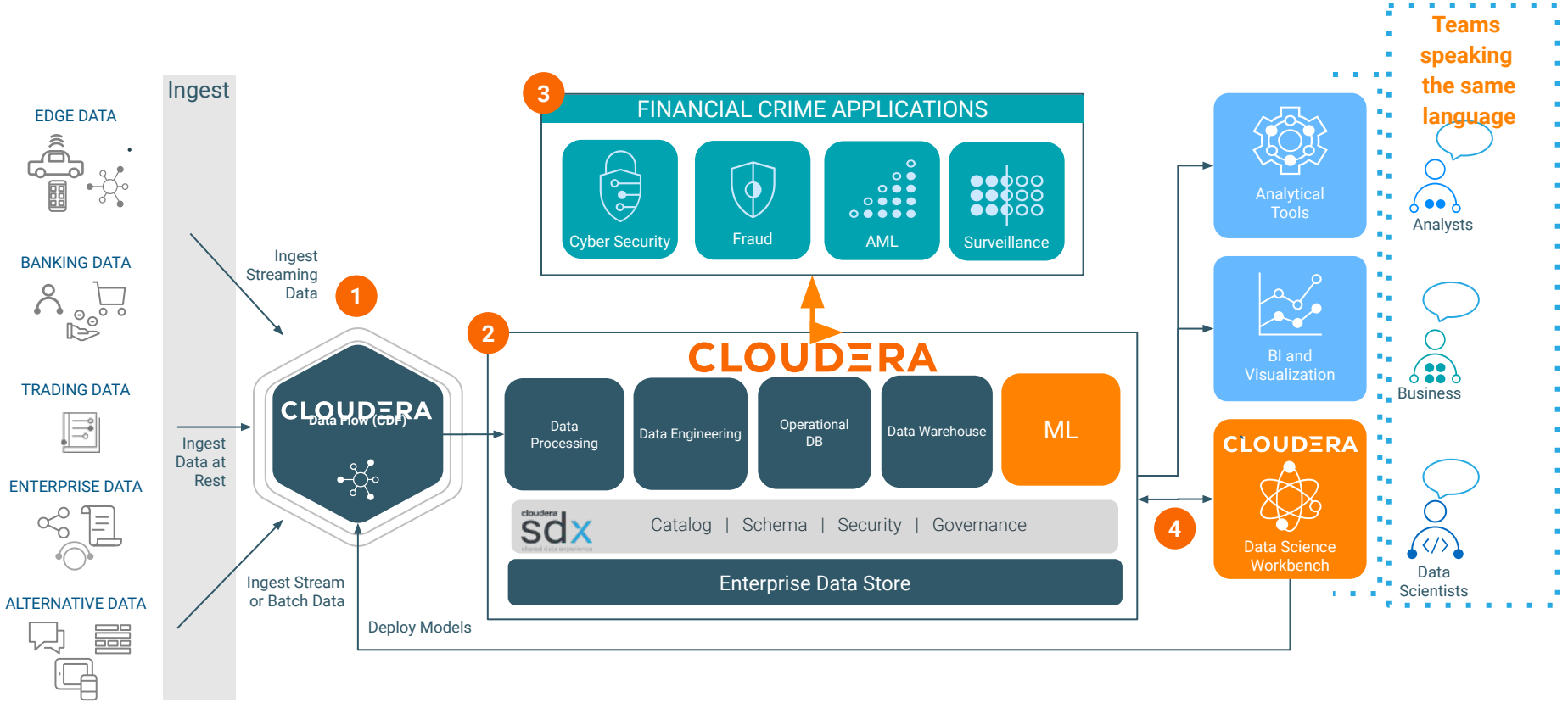


DATA VELOCITY in FINANCIAL SERVICES

Streaming capabilities vary, all enhance insight

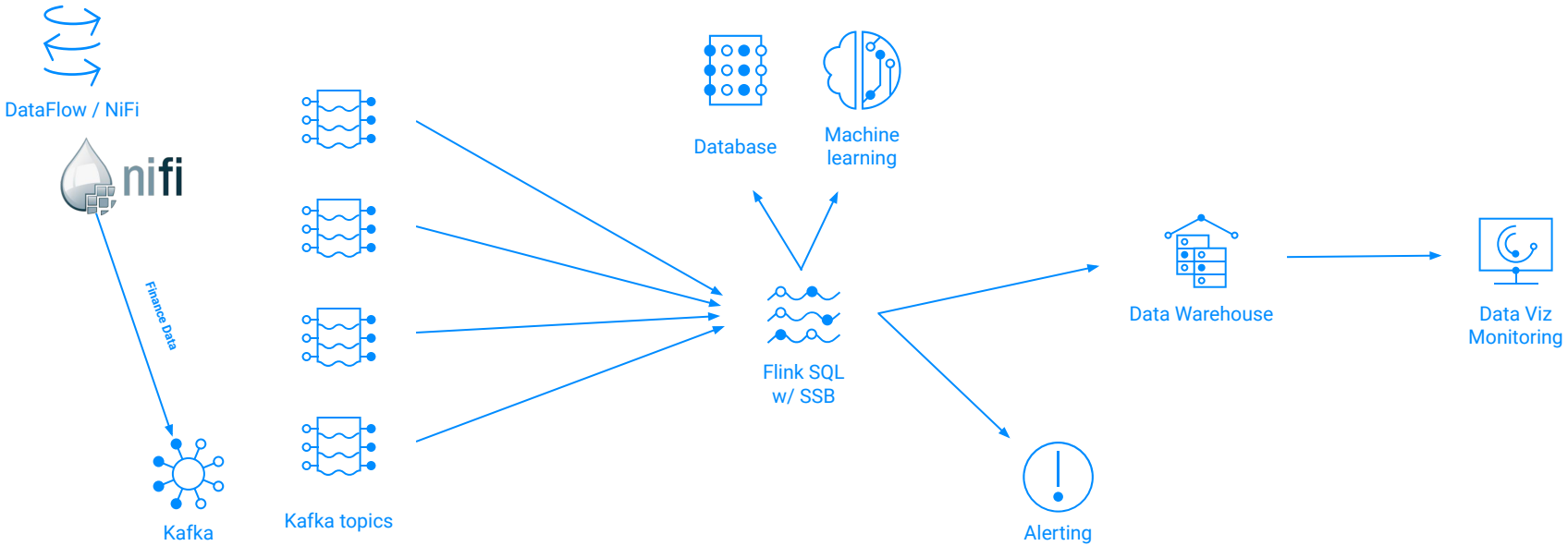


NEXT GEN PLATFORM FOR TACKLING FINANCIAL CRIME



Architecture in the context of Financial Use Cases

Kafka & Flink (Flink SQL with Stream SQL Builder) for real time analytics



NIFI MEETS AI

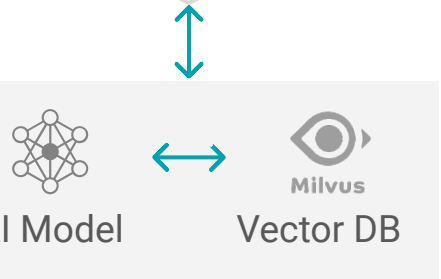
Unstructured file types

Structured Sources

Other enterprise data

Data in Motion With Cloudera

Capture, process & distribute any data, anywhere



AI Model

Vector DB

Applications/API's

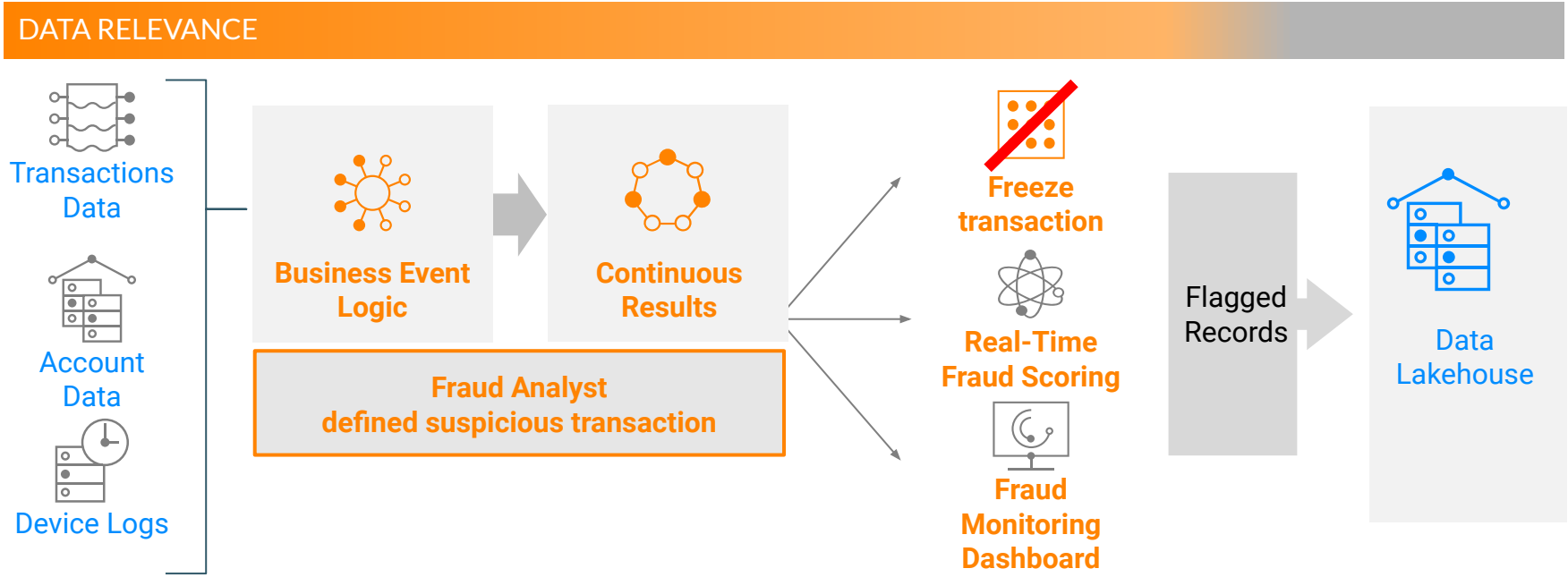
Streams

Materialized Views

Open Data Lakehouse

Stop Fraud When It Happens—Real Life Example

Simplified example of deployed use case



FINANCE DATA

Finnhub Stock API

Configure Processor | QueryRecord 1.21.0

Stopped

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Record Reader	SchemaJsonTreeReader
Record Writer	SchemaJsonRecordSetWriter
Include Zero Record FlowFiles	false
Cache Schema	true
Default Decimal Precision	10
Default Decimal Scale	0
all	SELECT description as companyName, symbol ...

Extract Company Name from User Query via NLP

Convert Company Name to Stock Symbol via Finnhub REST

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

INGEST | PREPARE | PUBLISH

DATA SOURCES

MySQL

PostgreSQL

DB2

ORACLE

Microsoft SQL Server

INGESTION

APACHE nifi

STREAM

SQL Stream Builder

MESSAGE HUB

kafka

STORAGE

KUDU

BATCH

CLOUDERA Machine Learning

APACHE Spark

CONSUMPTION

CLOUDERA Data Visualization

JDBC

ENTERPRISE LAKEHOUSE CAPABILITY VIEW

MANAGEMENT

Apache Airflow

CLOUDERA Workload Manager

CLOUDERA SDX

Apache Ranger

Apache Atlas

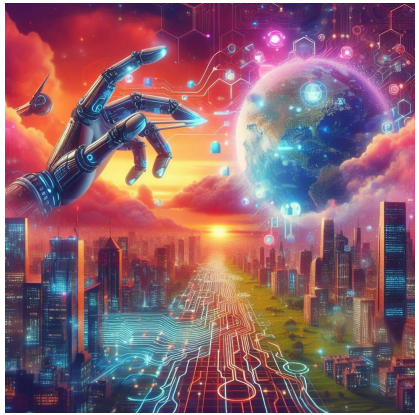
Internal Users (After Sales)

External Systems

watsonx.data

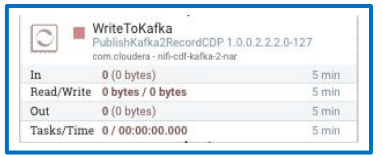
presto

Closed Loop Systems

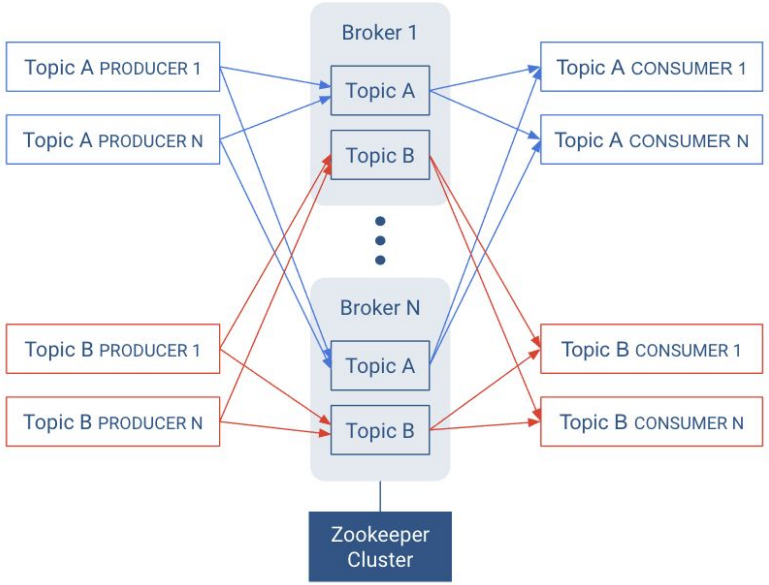


KAFKA and FLINK

STREAMS MESSAGING WITH KAFKA



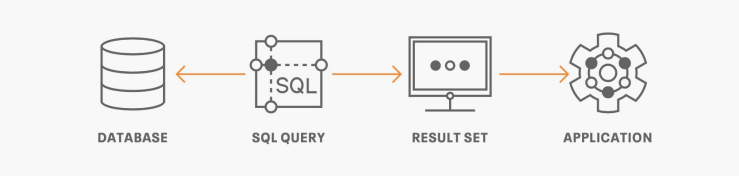
- Highly reliable distributed messaging system.
- Decouple applications, enables many-to-many patterns.
- Publish-Subscribe semantics.
- Horizontal scalability.
- Efficient implementation to operate at speed with big data volumes.
- Organized by topic to support several use cases.



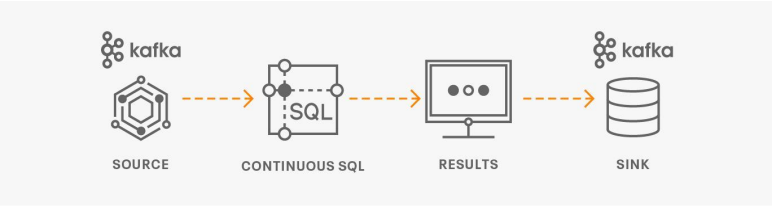
CONTINUOUS SQL

- 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`

SQL STREAM BUILDER (SSB)

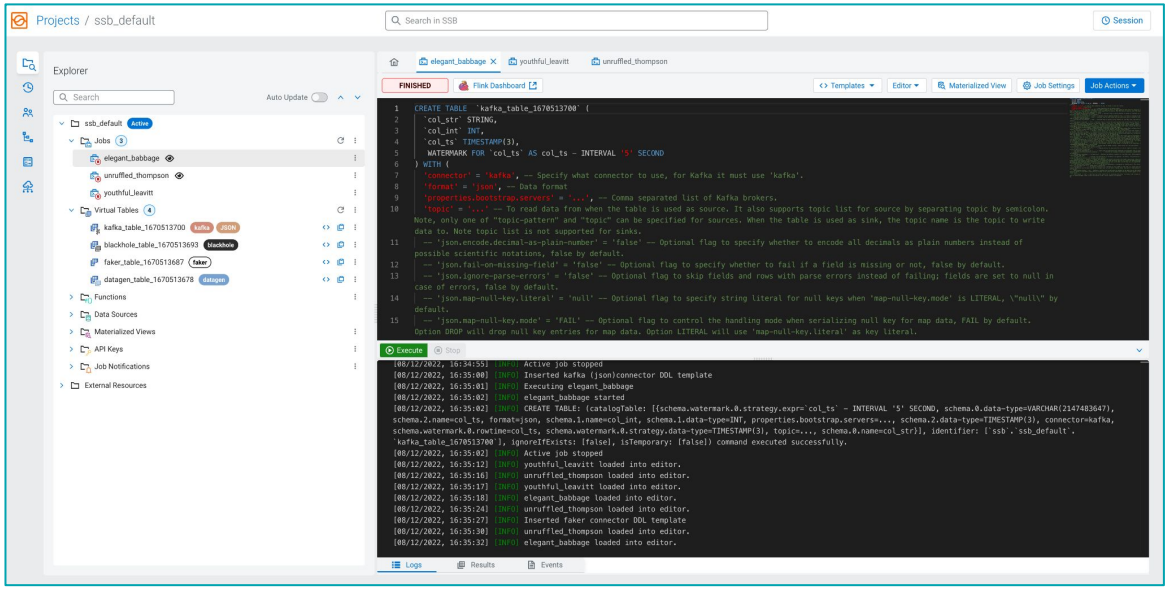
Democratize access to real-time data with just SQL

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

No Java or Scala code development required.

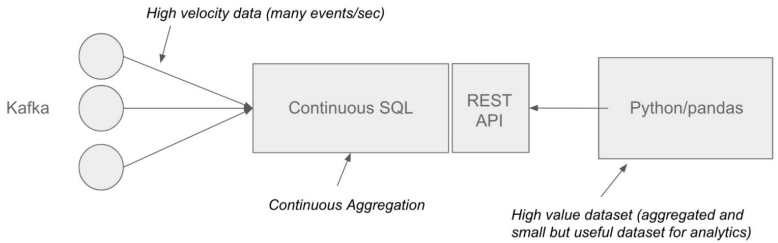
Simplifies access to data in Kafka & Flink. Connectors to batch data in HDFS, Kudu, Hive, S3, JDBC, CDC and more

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

```
[ ]:
```

SQL STREAM BUILDER (SSB)

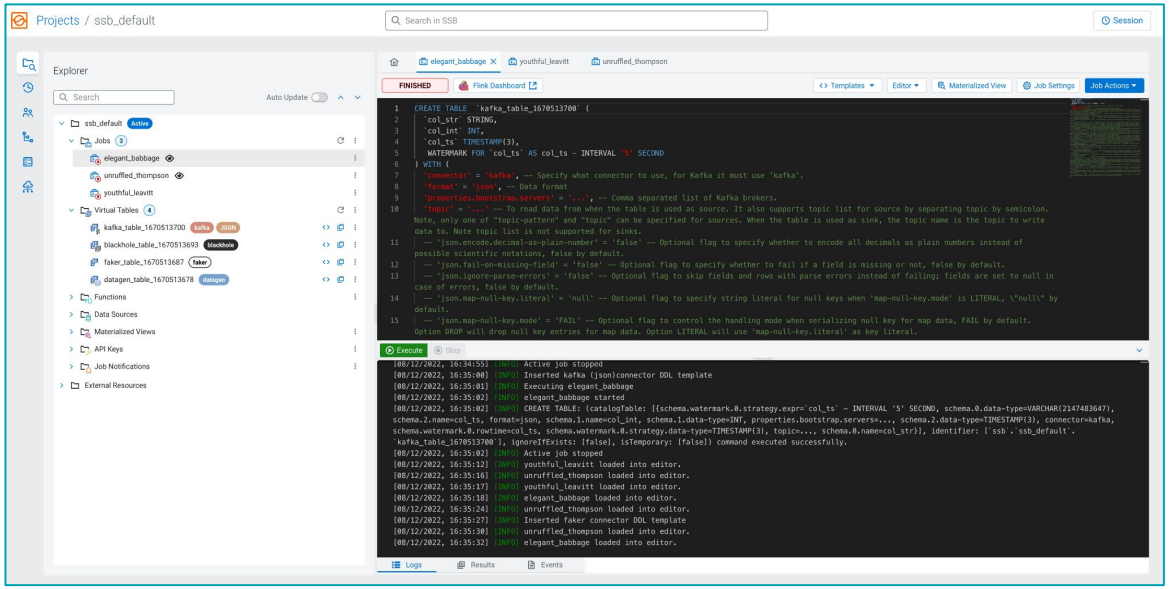
Democratize access to real-time data with just SQL

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



ICEBERG INTEGRATION

Robust Next Generation Architecture for Data Driven Business

Unified Processing Engine

Massive Open table format

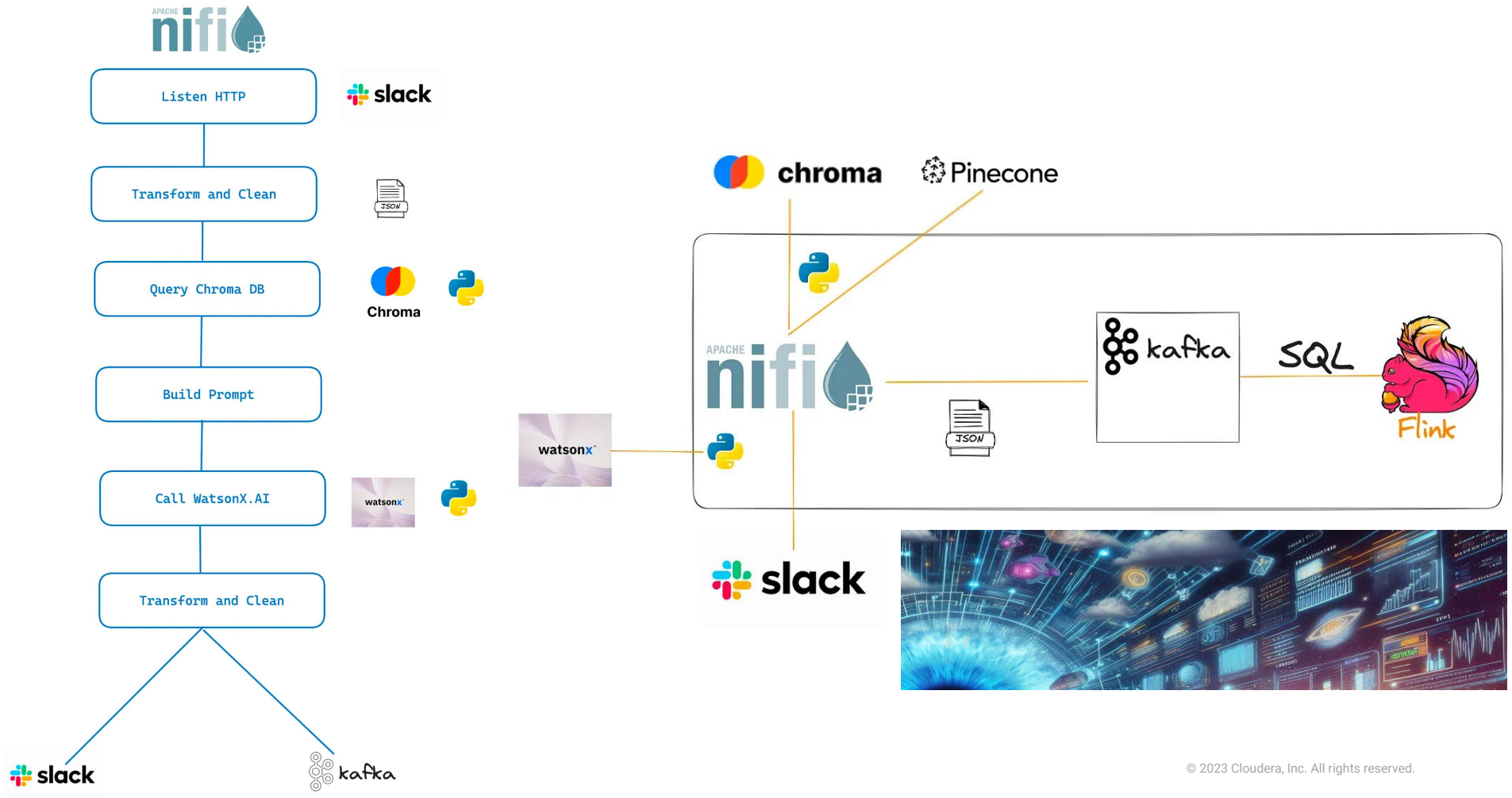
- Maximally open
- Maximally flexible
- Ultra high performance for MASSIVE data

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

https://github.com/tspannhw/FLaNK-Py-Stocks
https://github.com/tspannhw/PaK-Stocks



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
```

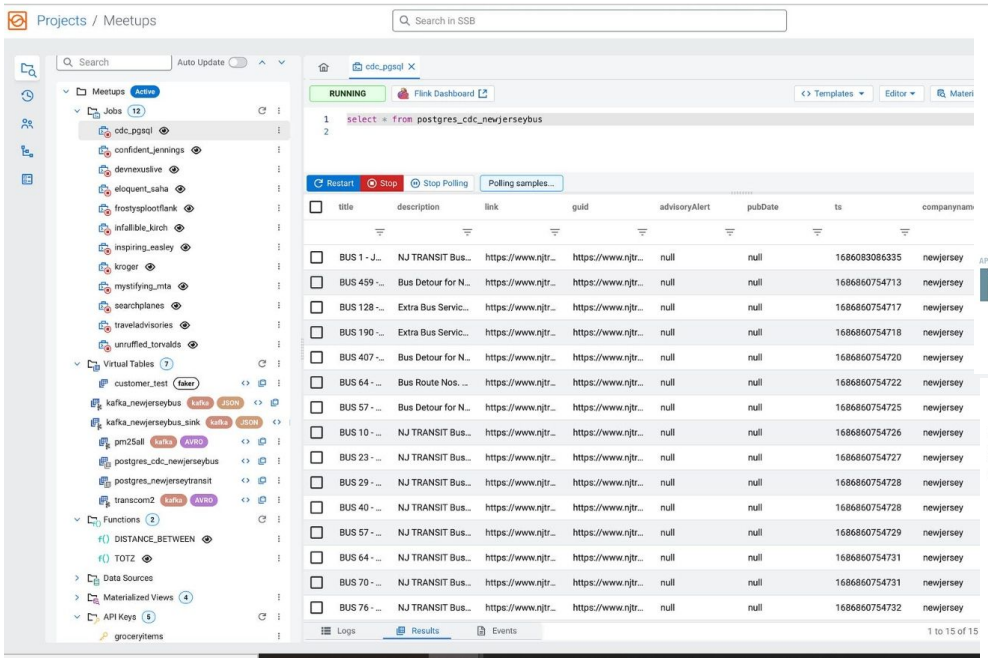
DEMO

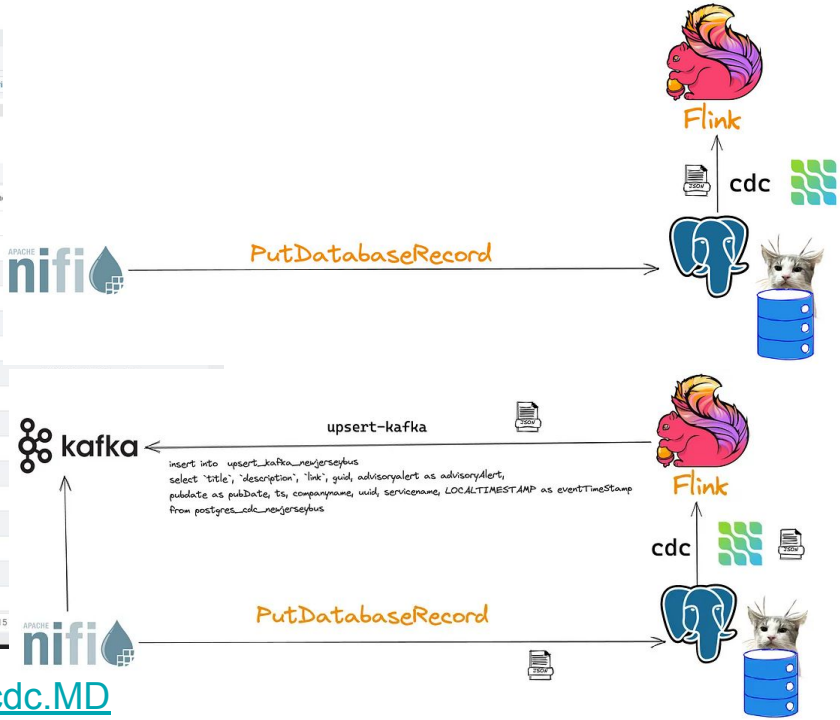

CYBERPUNK
CYBEERPUNK CODELLESS DITATRASE NIFI STREAM


kafka
nifi
FLaNK

CDC with Flink SQL (SSB)

Streaming CDC with Cloudera SQL Stream Builder (Flink SQL)

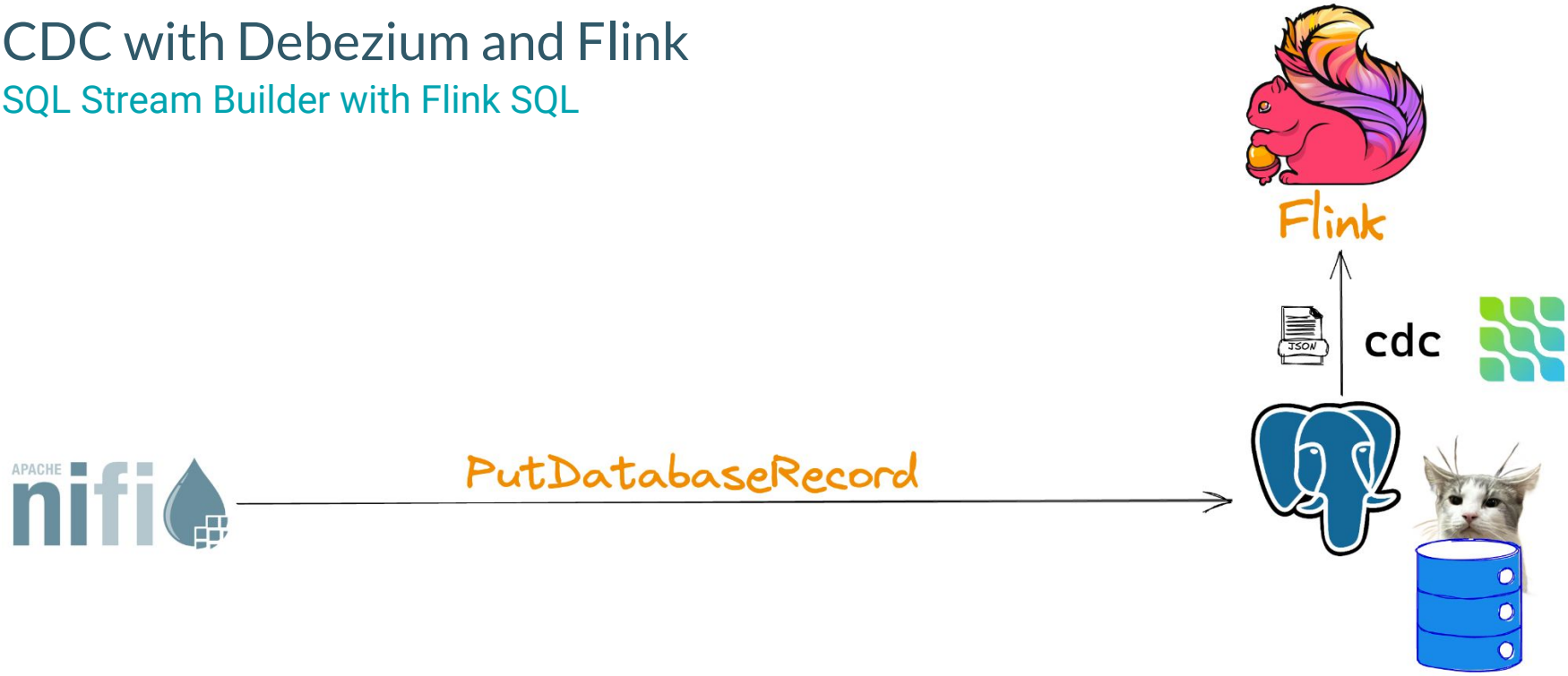




https://github.com/tspannhw/FLaNK-CDC/blob/main/flinkcdc.MD

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

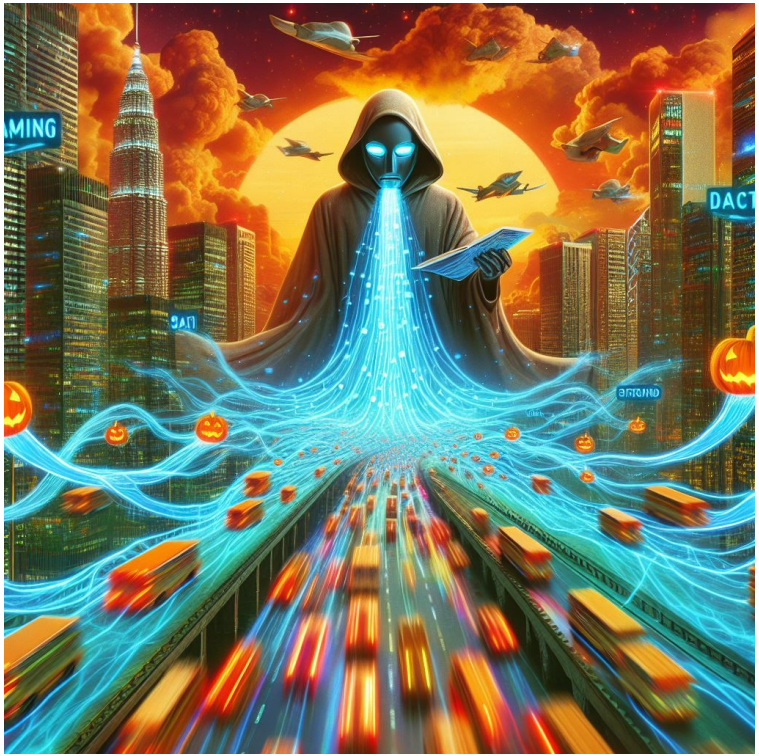
DDL

```
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'
  );
```

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

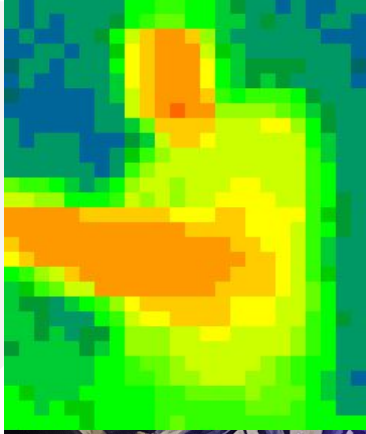
CONF42 PYTHON 2024

DATA SUMMIT
UNLEASH THE POWER OF YOUR DATA
MAY 8–9
BOSTON, MA


CYBERPUN AI
Concuning Real
tinetine data
Kinkat Kavka
Pipeslines


THE FUTURE OF THE WORLD
KAFKA, FLINK, NIFI, ICEBERG


LED TIM
FLYBMPK & FLINKISIAAFKA
HARNESSTHE INFINITE DATASTREAM



THANK YOU


GEN AI NEEDS STREAMING