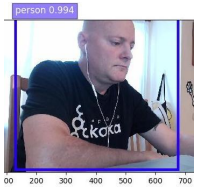
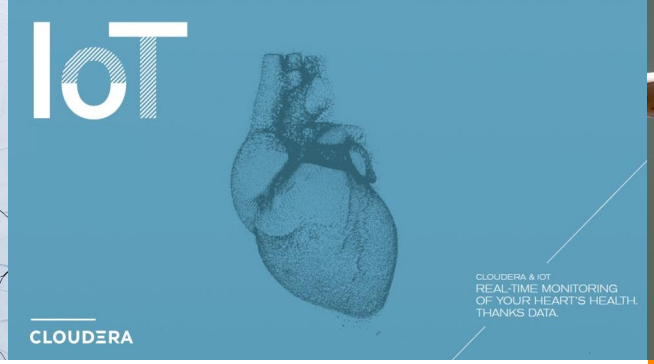
CLOUDERA

Unlocking Financial Data with Real-Time Pipelines

Tim Spann
Principal Developer Advocate

14-December-2023



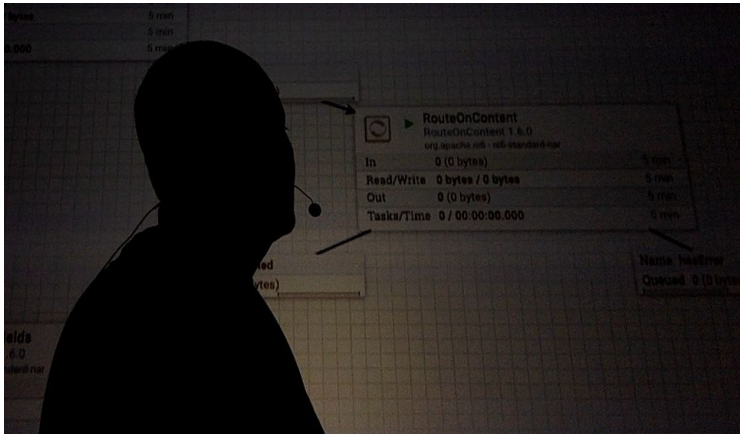


CLOUDERA

EDGE
2AI
CLOUDERA

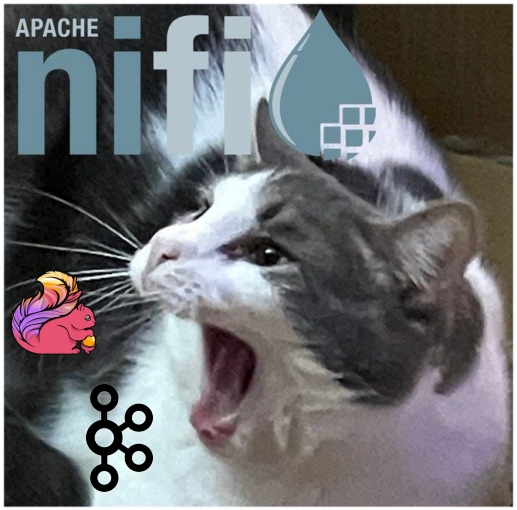
ENTERPRISE
DATA
CLOUD
CLOUDERA

Agenda (30 minutes)



- Introduction
- Overview
- Apache Kafka
- Apache Flink
- Apache NiFi
- Use Cases
- Demos

FLaNK Stack Weekly by Tim Spann

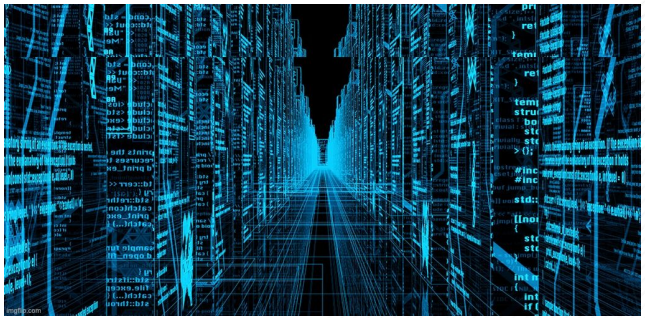


https://bit.ly/32dAJft

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

Tim Spann

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

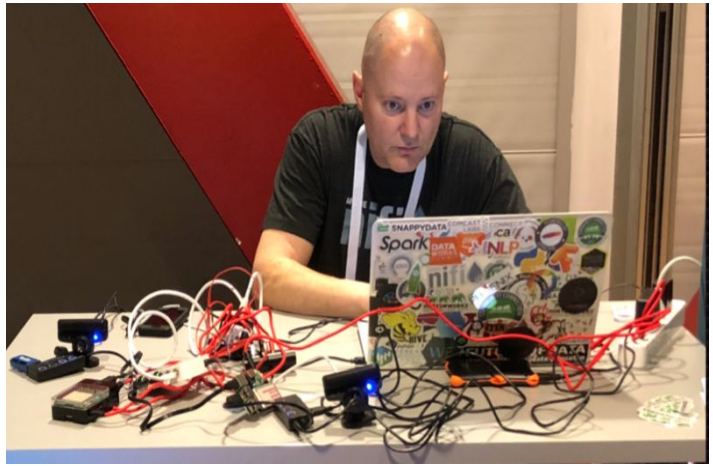
Principal Developer Advocate.

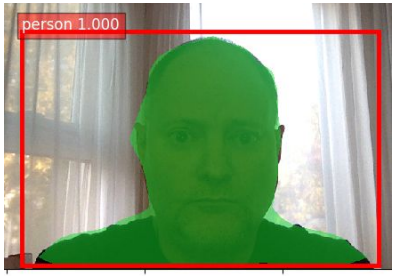
Princeton Future of Data Meetup.

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

https://medium.com/@tspann

https://github.com/tspannhw





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 th 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 processi 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.

Key Points to be Covered:

Introduction to Real-Time Data Pipelines: a. The limitations of traditional batch processing in the financial domain. b. Understanding the need for real-time data processing.

Apache Flink: Powering Real-Time Stream Processing: a. Overview of Apache Flink and its role in real-time stream processing. b. Use cases for Apache Flink in the financial industry. c. How Flink enables fast, scalable, and fault-tolerant processing of streaming financial data.

Apache Kafka: Building Resilient Event Streaming Platforms: a. Introduction to Apache Kafka and its role as a distributed streaming platform. b. Kafka's capabiliti in handling high-throughput, fault-tolerant, and real-time data streaming. c. Integration of Kafka with financial data sources and consumers.

Apache NiFi: Data Ingestion and Flow Management: a. Overview of Apache NiFi and its role in data ingestion and flow management. b. Data integration and transformation capabilities of NiFi for financial data. c. Utilizing NiFi to collect and process financial data from diverse sources.

Iceberg: Efficient Data Lake Management: a. Understanding Iceberg and its role in managing large-scale data lakes. b. Iceberg's schema evolution and table-leve metadata capabilities. c. How Iceberg simplifies data lake management in financial institutions.

Real-World Use Cases: a. Real-time fraud detection using Flink, Kafka, and NiFi. b. Portfolio risk analysis with Iceberg and Flink. c. Streamlined regulatory reporting leveraging all four technologies.

Best Practices and Considerations: a. Architectural considerations when building real-time financial data pipelines. b. Ensuring data integrity, security, and compliance in real-time pipelines. c. Scalability and performance optimization techniques.

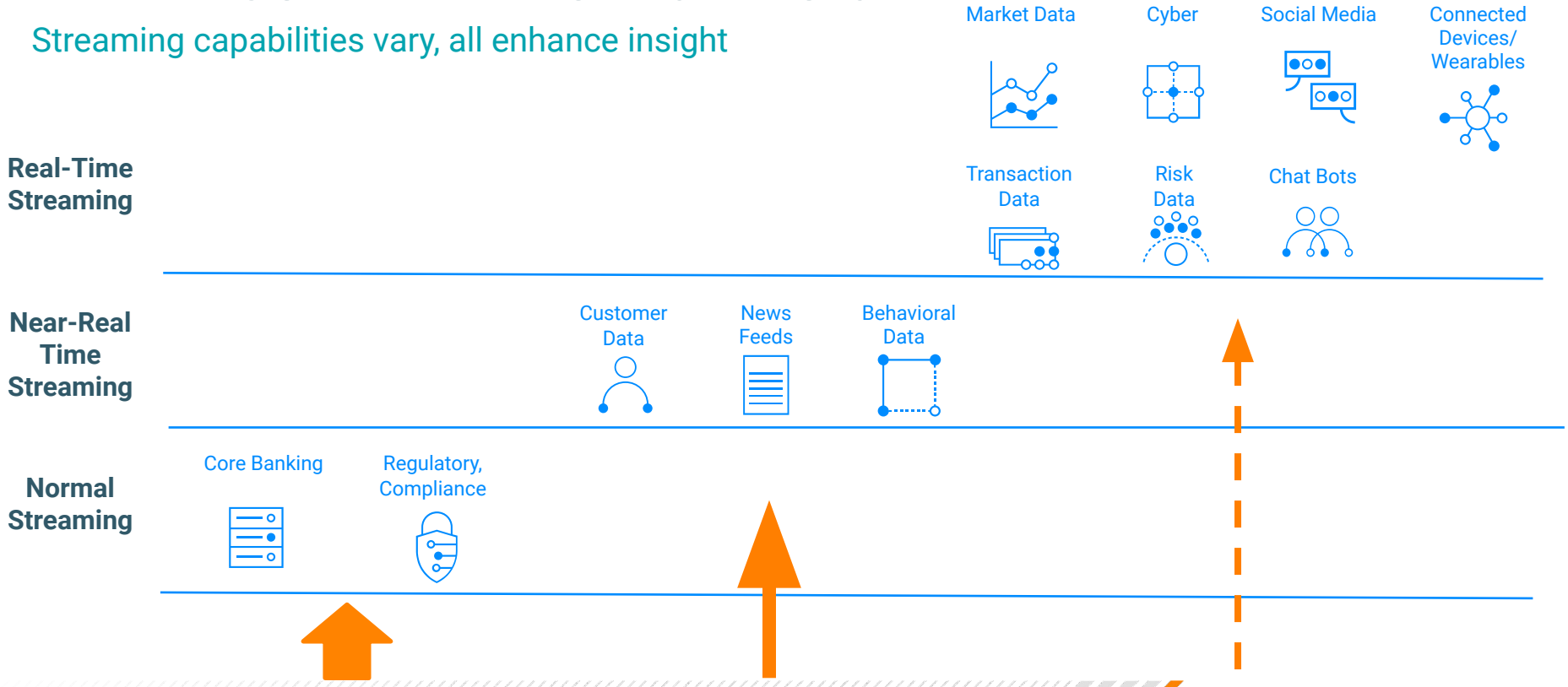
Conclusion: In this talk, we will demonstrate the power of combining Apache Flink, Apache NiFi, Apache Kafka, and Iceberg to unlock financial data's true potentia Attendees will gain insights into how these technologies can empower financial institutions to make informed decisions, respond to market changes swiftly, and comply with regulations effectively. Join us to explore the world of real-time data pipelines and revolutionize financial data management.

OVERVIEW

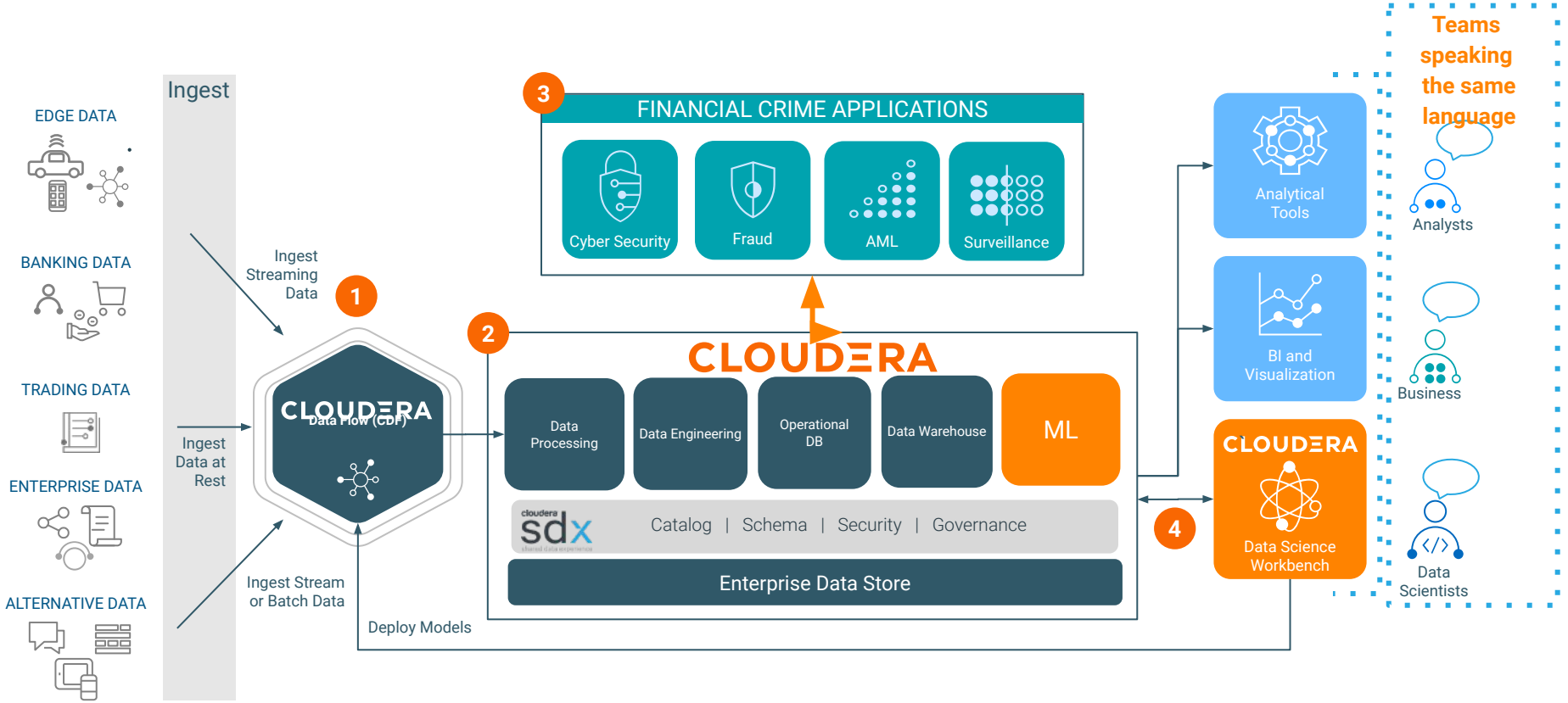

CUNK ON FLANK
THEY SAY IT'S FOR FAST DATA
IS IT SO FAST THAT IT ARRIVES BEFORE I ASKED FOR IT
imgflip.com

DATA VELOCITY in FINANCIAL SERVICES

Streaming capabilities vary, all enhance insight

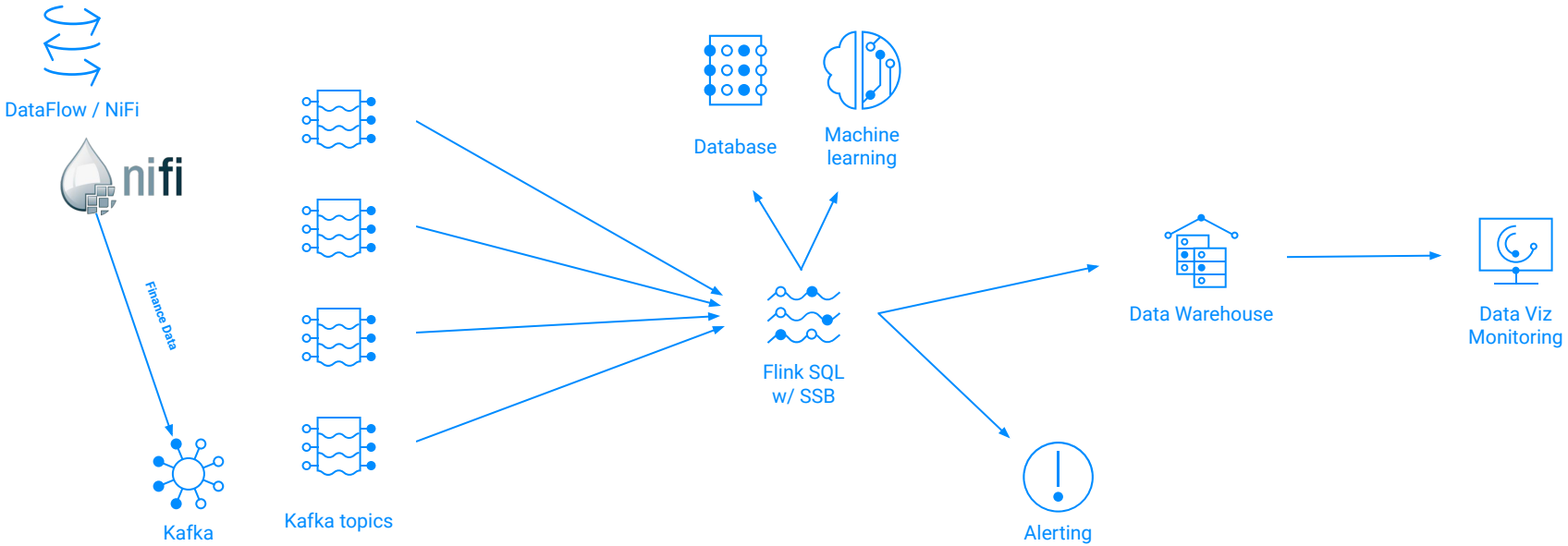


NEXT GEN PLATFORM FOR TACKLING FINANCIAL CRIME



Architecture in the context of Financial Use Cases

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



NIFI MEETS AI

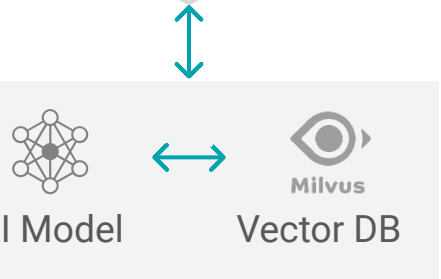
Unstructured file types

Structured Sources

Other enterprise data

Data in Motion With Cloudera

Capture, process & distribute any data, anywhere



AI Model

Vector DB

Applications/API's

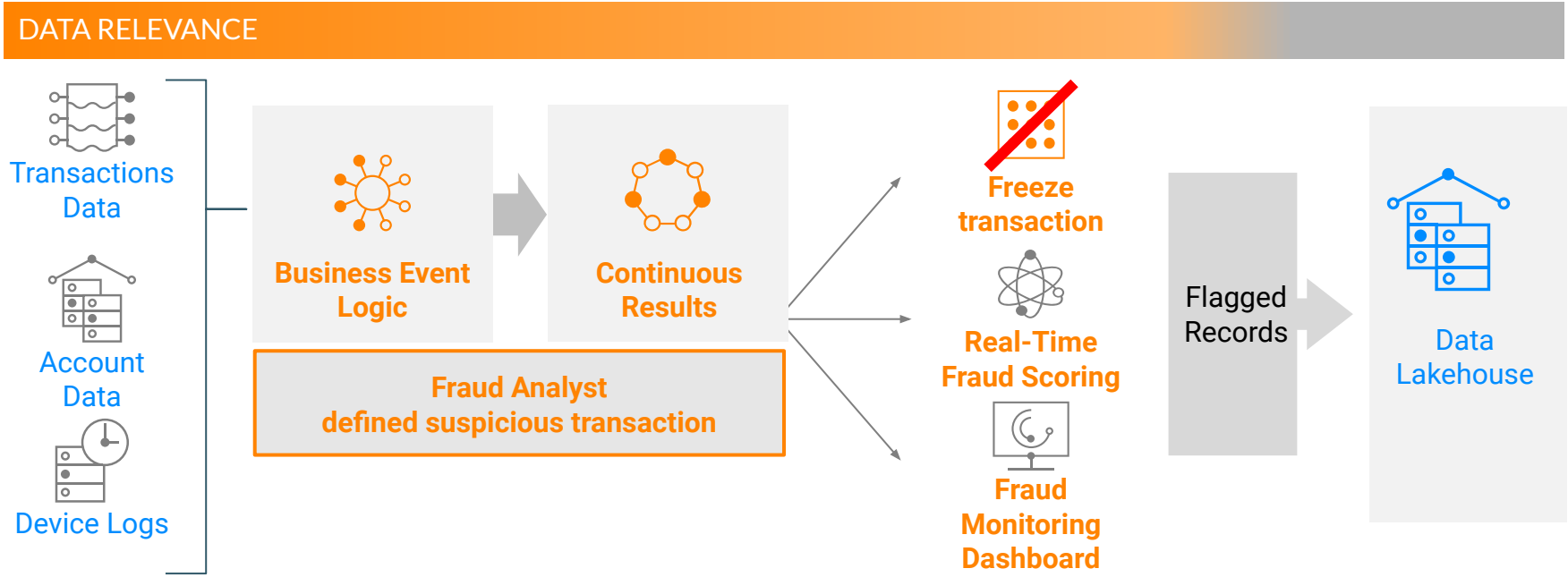
Streams

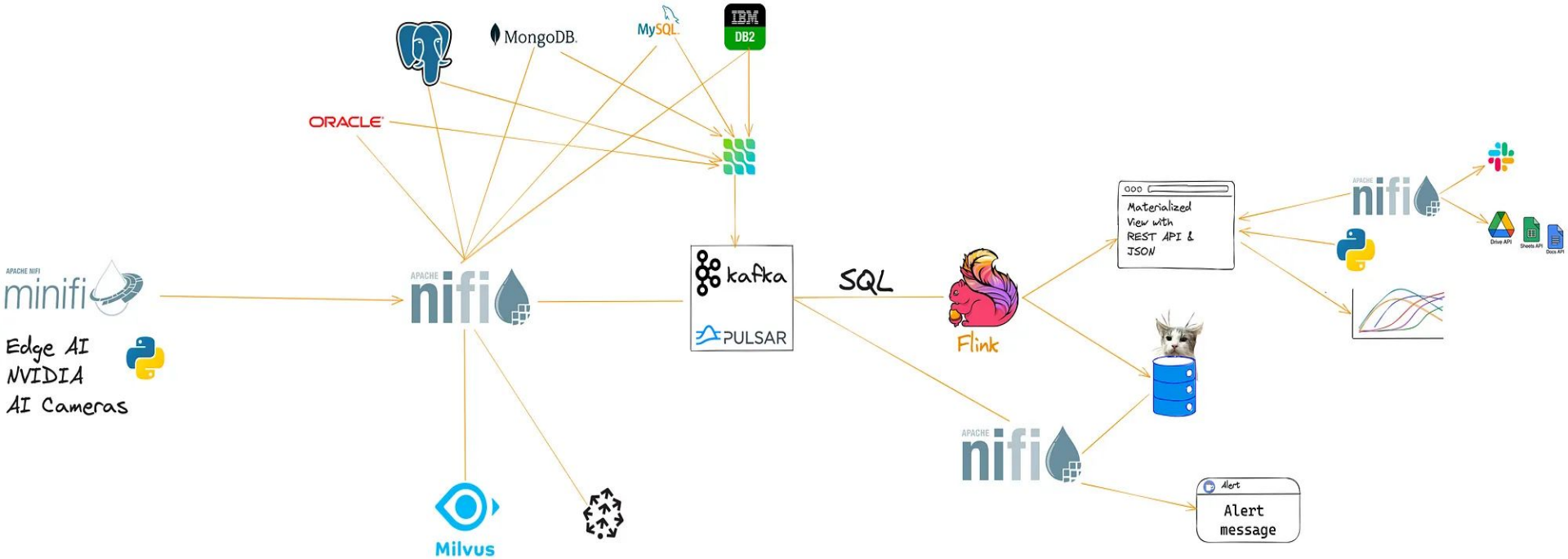
Materialized Views

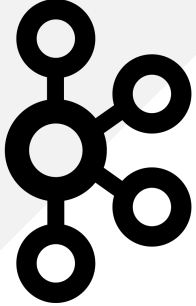
Open Data Lakehouse

Stop Fraud When It Happens—Real Life Example

Simplified example of deployed use case

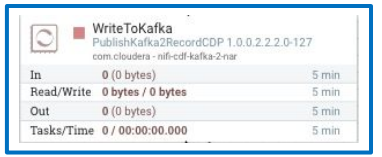



MongoDB.
MySQL
IBM
DB2
ORACLE
APACHE NIFI
minifi
Edge AI
NVIDIA
AI Cameras
APACHE
nifi
kafka
PULSAR
SQL
Flink
Materialized
View with
REST API &
JSON
APACHE
nifi
Milvus
APACHE
nifi
Alert
Alert
message

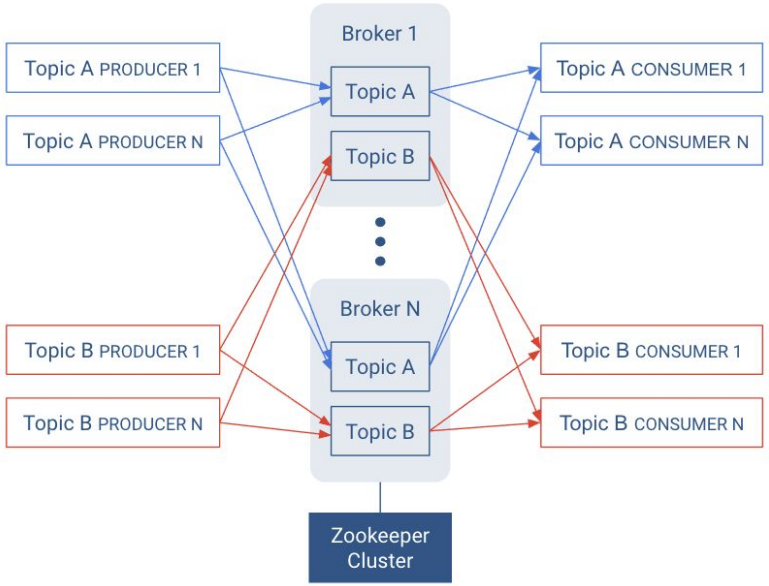


APACHE KAFKA

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.

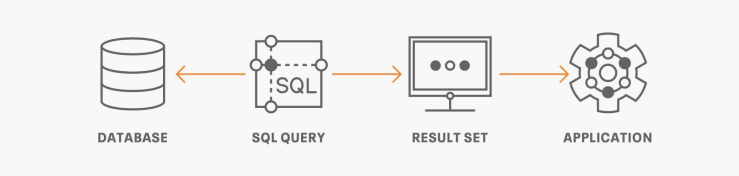


APACHE FLINK

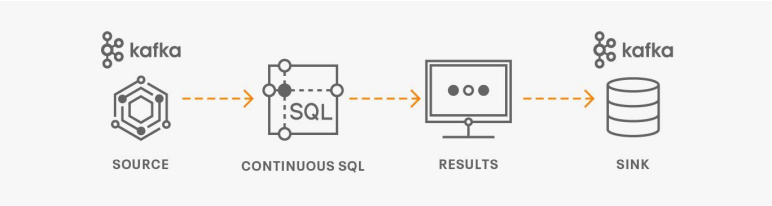
CONTINUOUS SQL

- 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`

SQL STREAM BUILDER (SSB)

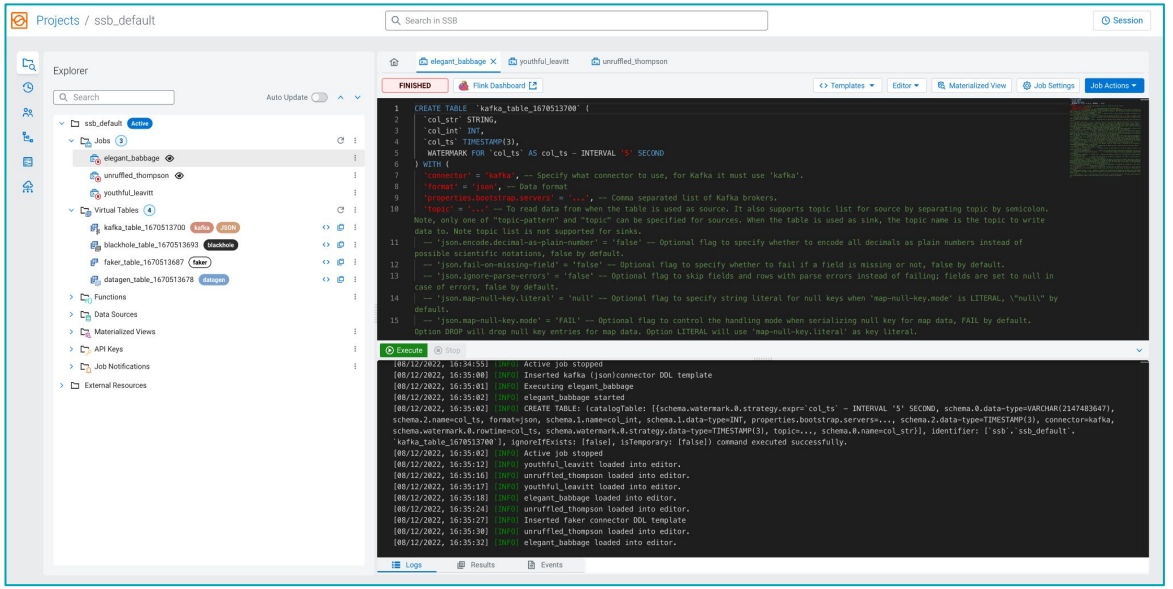
Democratize access to real-time data with just SQL

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

No Java or Scala code development required.

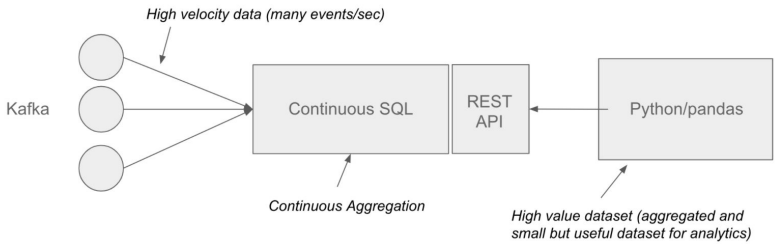
Simplifies access to data in Kafka & Flink. Connectors to batch data in HDFS, Kudu, Hive, S3, JDBC, CDC and more

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]:

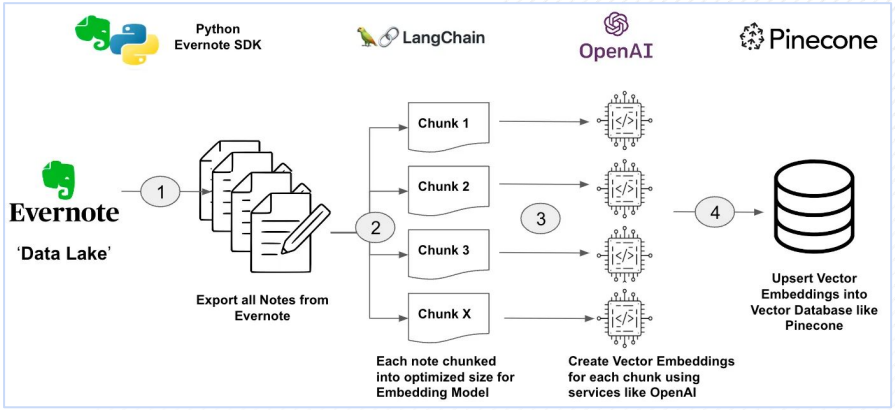
userid	max_amount	sum_amount	thecount	ts
787	34911	57304	10	2020-06-16 19:52:15
744	77407	95407	9	2020-06-16 19:52:15
78	86761	330397	9	2020-06-16 19:52:15
541	78762	282682	8	2020-06-16 19:52:15
926	85636	129728	8	2020-06-16 19:52:15

```
[ ]:
```

APACHE NIFI

NiFi 2.0.0-M1 is here… https://medium.com/cloudera-inc/getting-ready-for-apache-nifi-2-0-5a5e6a67f450

- **First-class citizen Python API**
- **Rules Engine**
- **NiFi Stateless at Process Group level**
- **Java 21 (virtual threads, perf improvements, etc)**

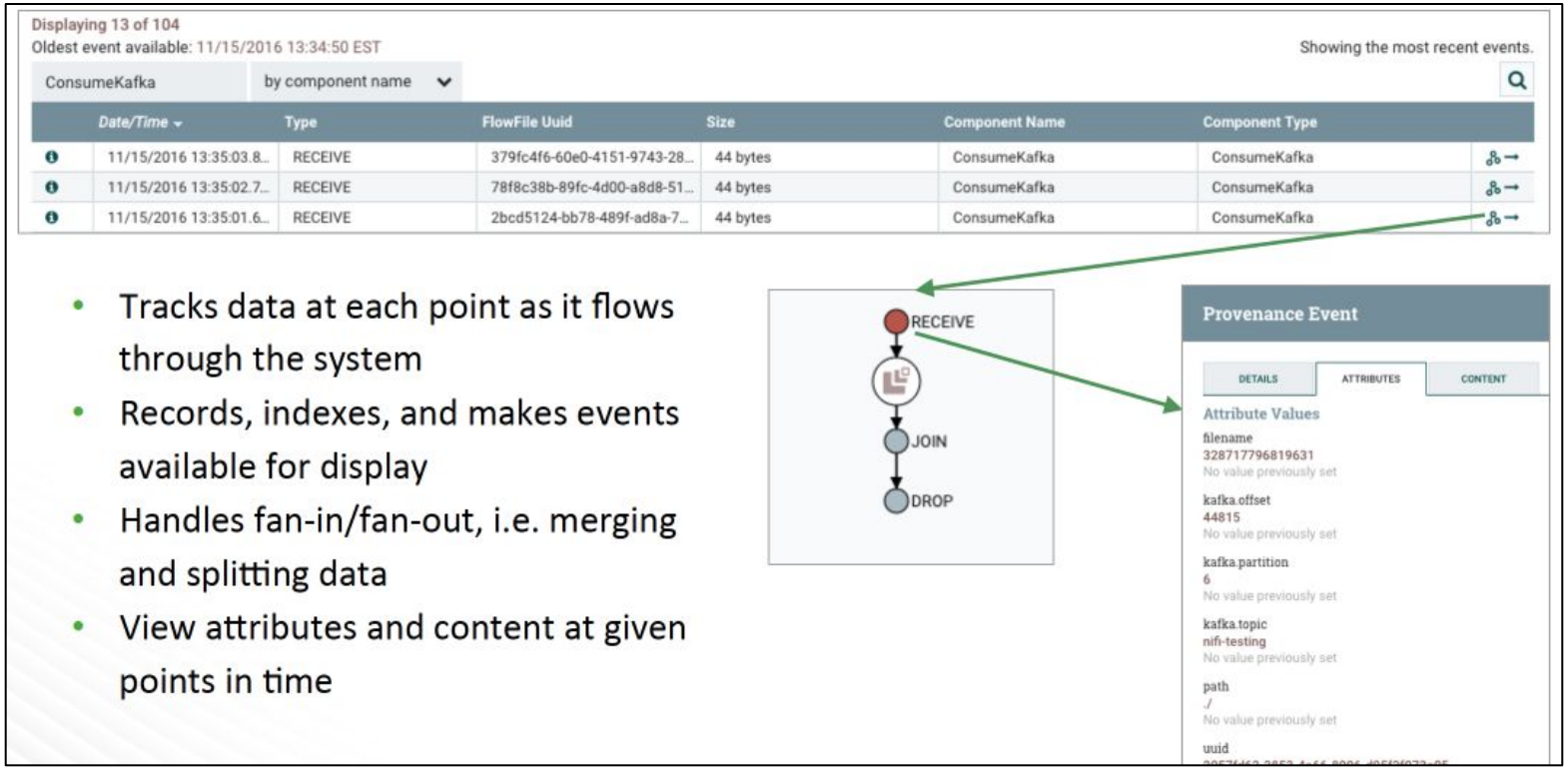


Closing the gap between data engineers and data scientists…

- Export documentation (Sharepoint, OCR) to build the knowledge base powering your chatbot
- Scrape the internet (Sitemap) to build the knowledge base powering your chatbot
- Real-time streaming ingest of Slack to build the knowledge base powering your chatbot

https://medium.com/@george.vetticaden/accelerating-ai-data-pipelines-building-an-evernote-chatbot-with-apache-nifi-2-0-and-generative-ai-9d977466ff4c

PROVENANCE



DEMO

Cloudera Stream Processing Community Edition

CSP Community Edition

A readily available, dockerized deployment of Apache Kafka and Apache Flink that allows you to test the features and capabilities of Cloudera Stream Processing.

Learn More

- **Zero to Flink in less than an hour**
  - Experiment with features
  - Develop apps locally
- One docker compose file of CSP which includes:
  - All dependencies required to run
  - Kafka, Kafka Connect and Flink
  - Streams Messaging Manager
  - Schema Registry
  - **SQL Stream Builder**
- Licensed under the Cloudera Community License
- Community Group Hub (Discussion Forum) for CSP
- Find it on docs.cloudera.com under Applications

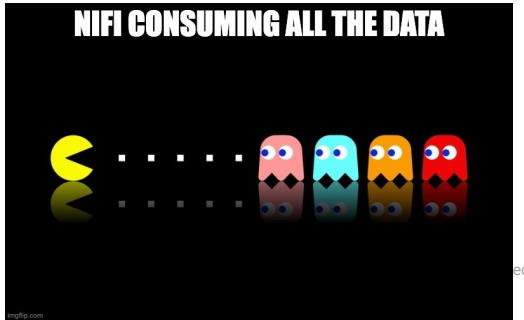
Open Source Edition

- Apache NiFi in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

- Licensed under the ASF License
- **Unsupported**

https://hub.docker.com/r/apache/nifi



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

https://events.dzone.com/dzone/Data-Pipelines-Investigating-the-Modern-Day-Stack


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

THANK YOU


GEN AI NEEDS STREAMING