CLOUDERA

Partner Skill Up:
Enable a Streaming CDC Solution

Tim Spann
Principal Developer Advocate in Data In Motion for Cloudera, Global
tspann@cloudera.com

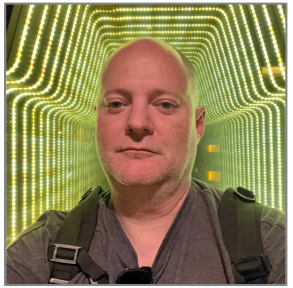
Salvador Almazan
Partner Solutions Engineer, US
salmazan@cloudera.com

CLOUDERA TEAM

Meet the Cloudera team

Today's Leads

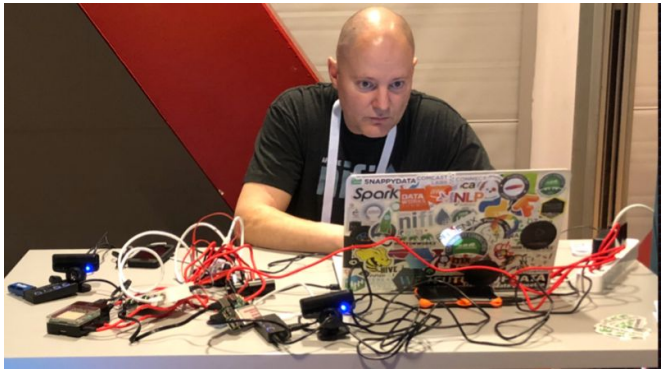
Salvador Almazan
Sr. Partner Solutions Engineer
salmazan@cloudera.com



Tim Spann
DIM Developer Advocate
tspann@cloudera.com

TODAY'S LEAD

Who am I?



Principal Data-in-Motion Developer Advocate

DZone Zone Leader and Big Data MVB

Princeton and NYC Future of Data Meetups
ex-Pivotal Field Engineer ex-StreamNative ex-PwC

https://github.com/tspannhw https://twitter.com/PaaSDev
https://www.datainmotion.dev/
https://medium.com/@tspann

Future of Data - NYC / Princeton + Virtual

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY

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

From Big Data to AI to Streaming to LLM to Cloud to Analytics to NLP to Fast Data to Machine Learning to Microservices to ...

ICEBERG

@PaasDev

Streaming Change Data Capture (CDC) 3+ Unique Ways

In this next session,

learn how to use Debezium with Flink, Kafka, and NiFi for Change Data Capture using two different mechanisms: Kafka Connect and Flink SQL.

With the virtual nature of today's world, streaming data is more critical than ever. Join Cloudera Chief Data-In-Motion Principal, Tim Spann, and Partner Solution Engineer, Salvador Alamazan as they look closely at key CDC use cases, discuss why Debezium is the best option for handling CDC and use examples to show you how to demonstrate value.

This is a must-attend experience!

https://medium.com/cloudera-inc/cdc-not-cat-data-capture-e43713879c03

https://dzone.com/articles/streaming-change-data-capture-data-two-ways

WHAT IS IT? WHY DO I NEED IT?

What is Change Data Capture

Full Fidelity vs Point in Time

Point in Time

JDBC Queries only get the point in time of the query.

1st Query - John as True, and Jane as false

2nd Query - John as True and Jane as False even thought at one point in the day John set his marketing opt-in to false

3rd Query - John as False, and Jane as False. This query missed that Jane had ever set the opt-in to True

Query

Query

Query

ROW	First	Email	Marketing Opt-In	Last_Updated
1	John	Doe	TRUE	1/1/2022
2	Jane	Doe	FALSE	1/1/2022
1	John	Doe	FALSE	1/2/2022
1	John	Doe	TRUE	1/2/2022
2	Jane	Doe	TRUE	1/2/2022
2	Jane	Doe	FALSE	1/2/2022
1	John	Doe	FALSE	1/3/2022

Full Data Fidelity

CDC Gets all the changes affecting the row
CDC Data would collect every Marketing Opt-In value change

Why Change Data Capture?

Most Common Use Cases - Distribution or Synchronization

Analytics {Distribution}

Full Fidelity Analytics Offline, Look at any point in time

Operational Data Store {Distribution}

Prevent load on operational databases by replicating the database for reporting queries to run in isolation

Migrations {Synchronization}

Keeping two databases in synchronization, existing and new systems need to coexist for some time

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?

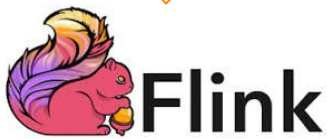
Want Low Latency?

Want Advanced Windowing and State?

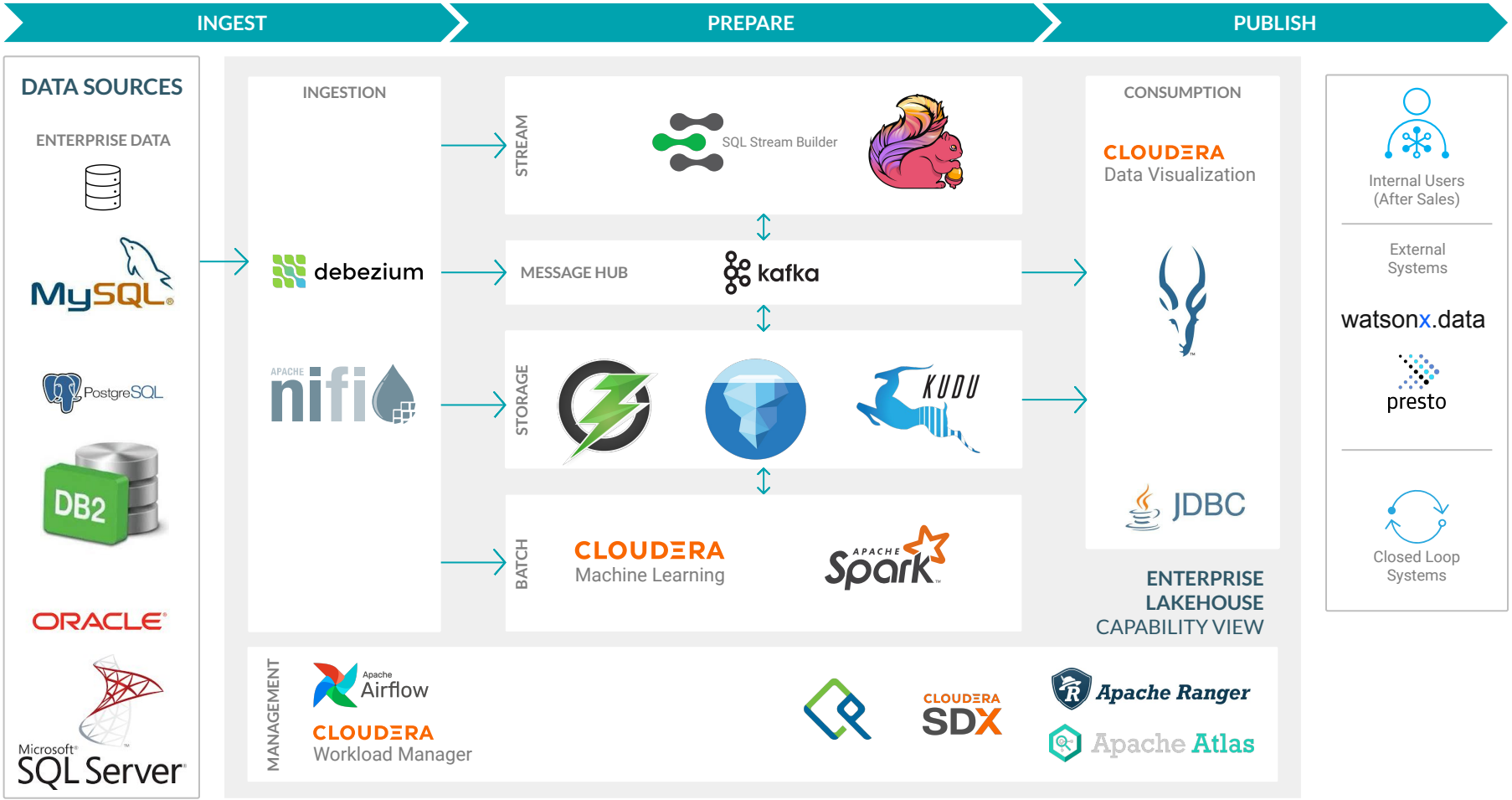
Automatic records immediately

Pure SQL

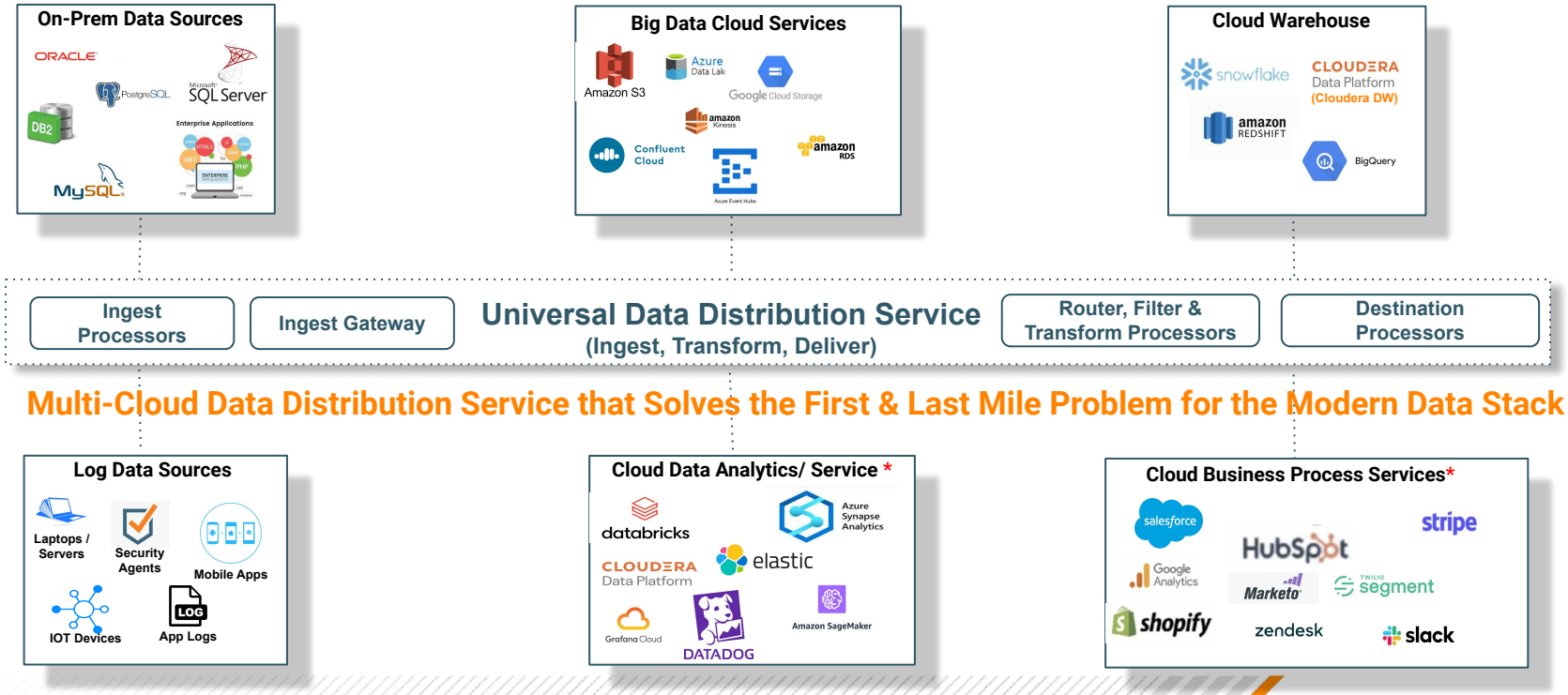
Debezium



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

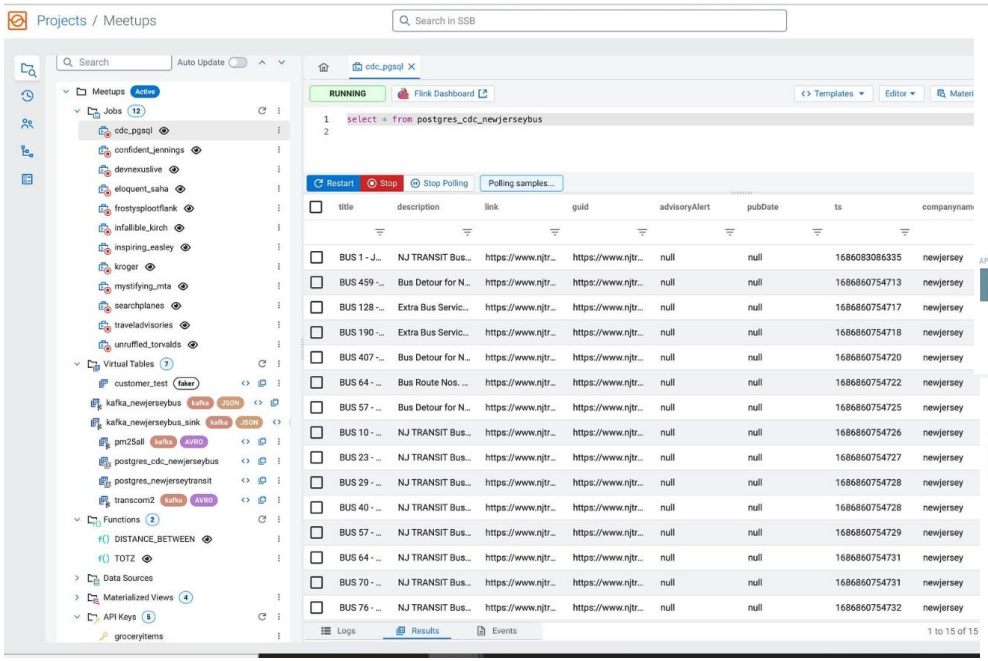


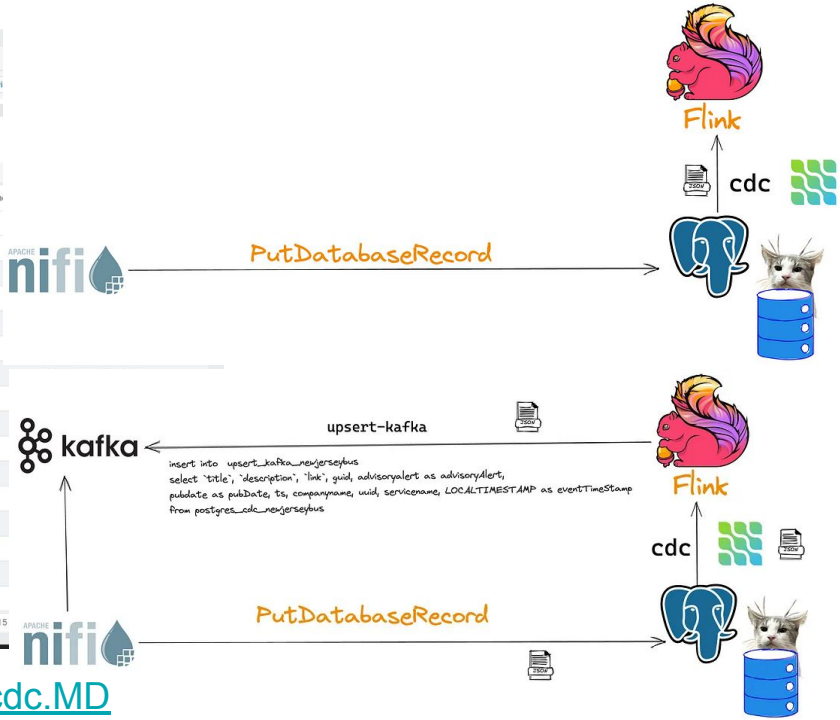
Data Distribution as a Universal, Hybrid, Multi-Cloud Data Service



CDC with SQL Stream Builder (Flink SQL)

Streaming CDC with Cloudera SQL Stream Builder (Flink SQL)

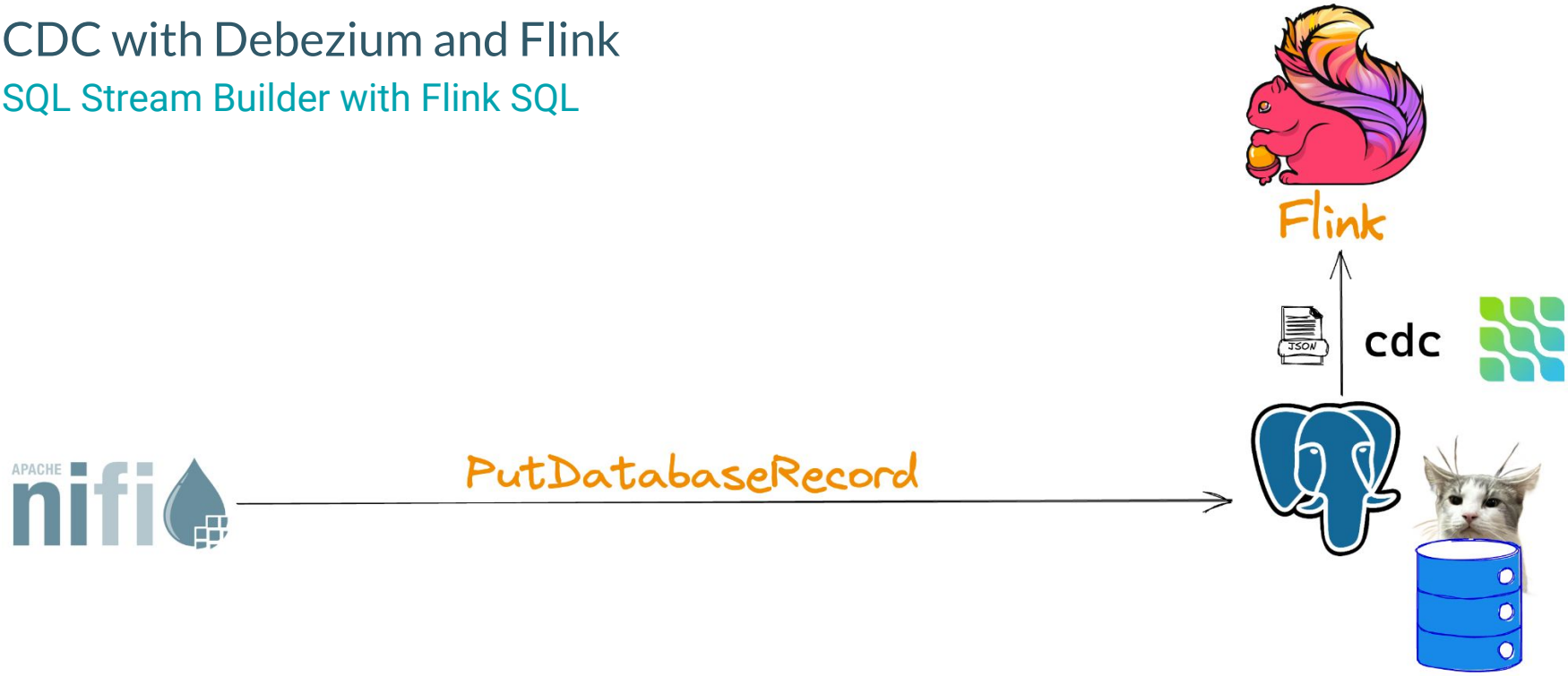




https://github.com/tspannhw/FLaNK-CDC/blob/main/flinkcdc.MD

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

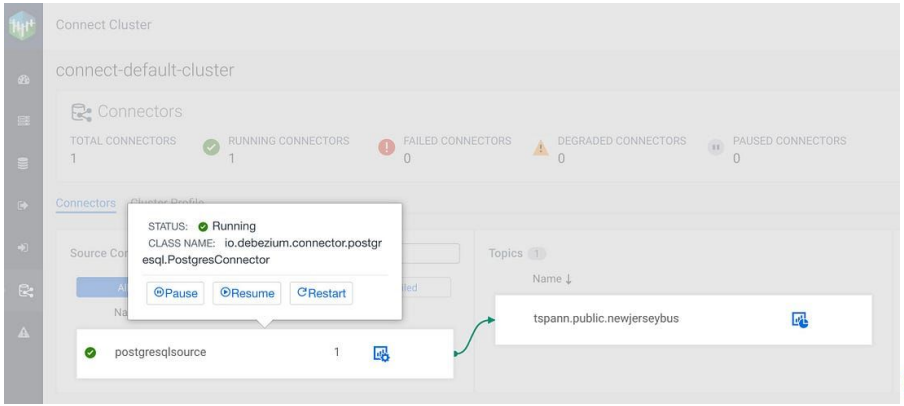
```
CREATE TABLE `postgres_cdc_newjerseybus` (
    `title` STRING,
    `description` STRING,
    `link` STRING,
    `guid` STRING,
    `advisoryAlert` STRING,
    `pubDate` STRING,
    `ts` STRING,
    `companyname` STRING,
    `uuid` STRING,
    `servicename` STRING
) WITH (
  'connector' = 'postgres-cdc',
  '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'
  );
```

CDC with Kafka Connect

Streaming CDC with Cloudera Streams Messaging Manager (Kafka)



https://github.com/tspannhw/FLaNK-CDC/blob/main/kafkacdc.md

CDC with Debezium and Kafka

Kafka Connect

What is it?

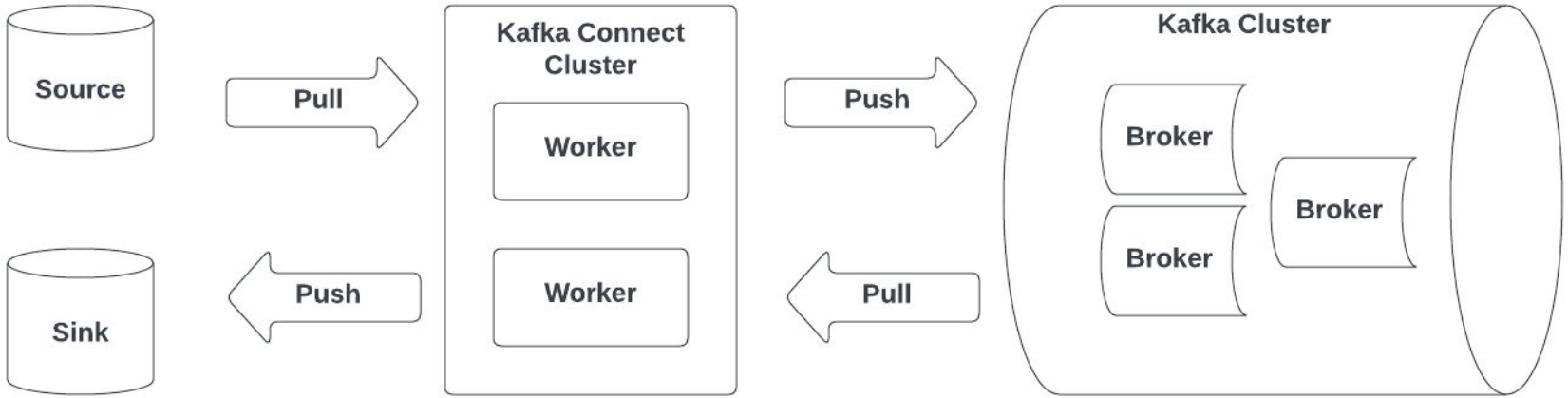
Kafka Connect

- Connect data sources (or sinks) with Kafka
  - Ideal for moving large amounts of data into or out of Kafka
  - In a reliable manner
  - In a performant way
- A tool built upon Apache Kafka
- Present since end of 2015
- Makes easier to implement the most common use case, earlier (presumably) implemented with Producer/Consumer API

Basic Concepts

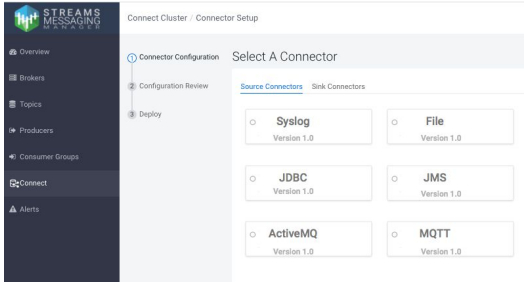
Kafka Connect

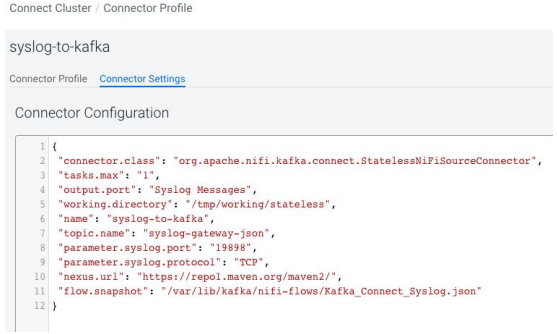
- Connect framework uses “Workers” (Java processes) to do data processing.
- Workers Run Connectors
- Workers can be organized into a Connect cluster.



KConnect Connectors

- New Connectors
- CDC Debezium Connectors
- SMM UI Integration
- Reuse your Kafka Infrastructure
- Enterprise Security for Secrets, Authentication and Authorization





Sources

- ActiveMQ (via JMS)
- MQTT
- Syslog over TCP
- JDBC
- JMS
- HTTP

Sinks

- AWS S3
- ADLS
- Kudu
- HDFS
- HTTP

CDC Connectors

- MySQL
- PostgreSQL
- Oracle
- MS SQL Server
- IBM DB2

- Connector Selection
- Connector Configuration
- 3 Configuration Review
- 4 Deploy

Configuration Review

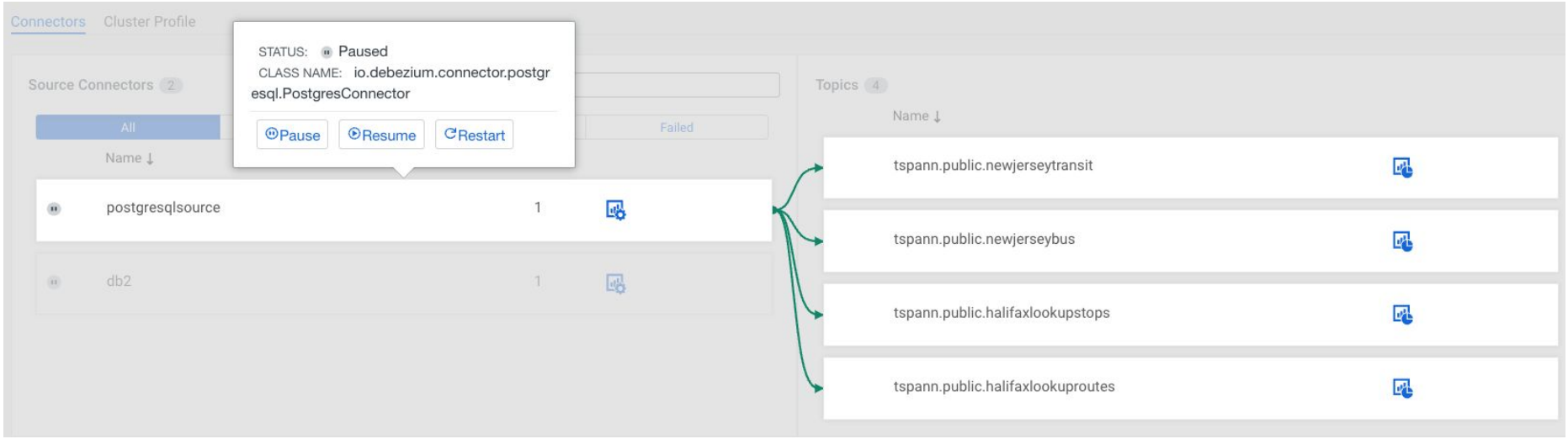
Please take a moment to review your configuration.

Search

Properties

postgresqlsource	
connector.class	io.debezium.connector.postgresql.PostgresConnector
database.dbname	tspann
database.history.kafka.bootstrap.servers	${cm-agent:ENV:KAFKA_BOOTSTRAP_SERVERS}
database.history.kafka.topic	schema-changes.bus-postgres
database.hostname	192.168.1.153
database.password	••••••
database.port	5432
database.server.id	184055
database.server.name	tspann
database.user	tspann
plugin.name	pgoutput
tasks.max	1

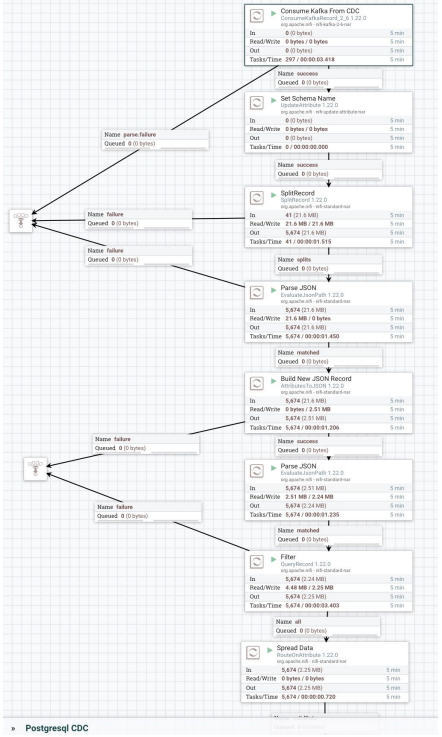
Cancel ← Back Deploy →



CDC with NiFi

Apache NiFi + Apache Kafka

Process Debezium Kafka Formats From Kafka Connect and others



Configure Processor | ConsumeKafkaRecord_2_6 1.22.0

Stopped

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

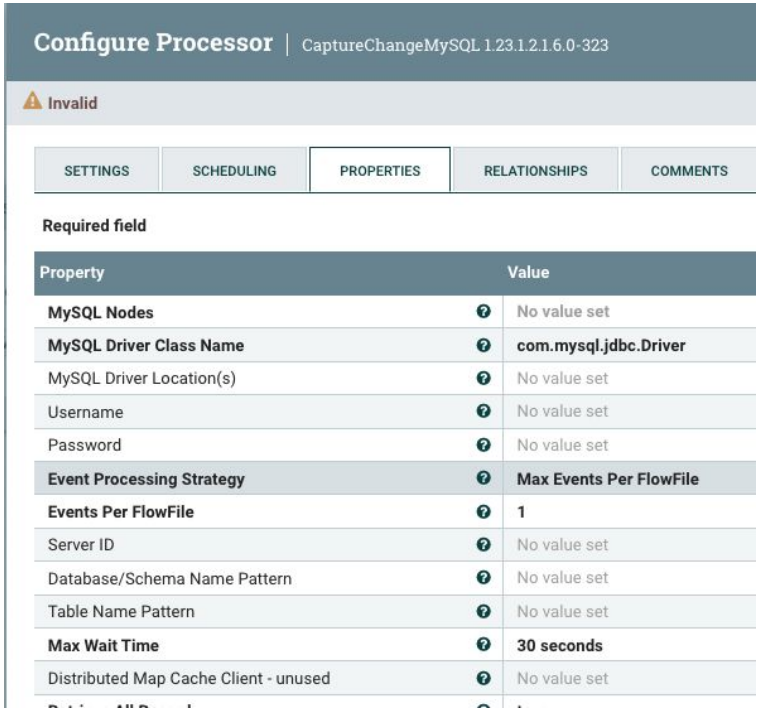
Property	Value
Kafka Brokers	kafka:9092
Topic Name(s)	tspann.public.newjerseybus
Topic Name Format	names
Value Record Reader	CDP Infer JsonTreeReader
Record Value Writer	Standard Inherit JsonRecordSetWriter
Group ID	kafka-nifi-cdc-reader
Output Strategy	Use Content as Value
Headers to Add as Attributes (Regex)	No value set
Key Attribute Encoding	UTF-8 Encoded
Commit Offsets	true
Max Uncommitted Time	1 secs
Honor Transactions	true

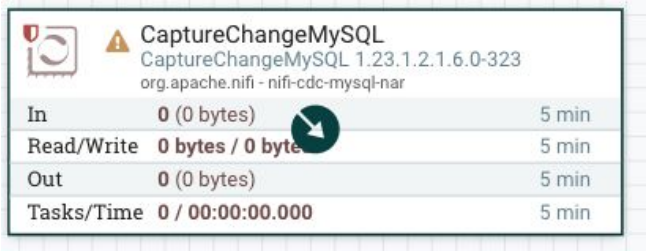
Property	Value
Destination	flowfile-attribute
Return Type	json
Path Not Found Behavior	ignore
Null Value Representation	empty string
after	$.payload.after
before	$.payload.before
sourceconnector	$.payload.source.connector
sourcedb	$.payload.source.db
sourcelsn	$.payload.source.lsn
sourcename	$.payload.source.name
sourceschema	$.payload.source.schema
sourcesequence	$.payload.source.sequence

Property	Value
after	$.payload.after
before	$.payload.before
sourceconnector	$.payload.source.connector
sourcedb	$.payload.source.db
sourcelsn	$.payload.source.lsn
sourcename	$.payload.source.name
sourceschema	$.payload.source.schema
sourcesequence	$.payload.source.sequence
sourcesnapshot	$.payload.source.snapshot
sourcetable	$.payload.source.table
sourcetsms	$.payload.source.ts_ms
sourcetxid	$.payload.source.txId

Apache NiFi to Databases

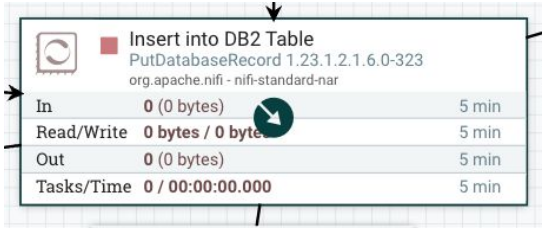
SQL and MySQL CDC





Apache NiFi to Databases

SQL Query Access



Processor Details

Running — STOP & CONFIGURE

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Record Reader	**CDP Infer JsonTreeReader**
Database Type	Oracle 12+
Statement Type	**INSERT**
Data Record Path	No value set
Database Connection Pooling Service	**Oracle12DBCPConnectionPool**
Catalog Name	No value set
Schema Name	TSPANN
Table Name	**NEWJERSEYBUS**
Translate Field Names	true
Unmatched Field Behavior	Ignore Unmatched Fields
Unmatched Column Behavior	Ignore Unmatched Columns
Quote Column Identifiers	false

Record Reader	**JsonTreeReaderSchema**
Database Type	Generic
Statement Type	**INSERT**
Data Record Path	No value set
Database Connection Pooling Service	**ibmdb2cluster**
Catalog Name	No value set
Schema Name	DB2INST1
Table Name	**TRAVELADVISORY**
Translate Field Names	true
Unmatched Field Behavior	Ignore Unmatched Fields
Unmatched Column Behavior	Ignore Unmatched Columns
Quote Column Identifiers	false

Database Connection Pooling Service	**ibmdb2cluster**
Database Type	**Generic**
Table Name	**TRAVELADVISORY**
Columns to Return	No value set
Additional WHERE clause	No value set
Custom Query	No value set
Record Writer	**JsonRecordSetWriter**
Maximum-value Columns	pubdate
Initial Load Strategy	**Start at Beginning**
Max Wait Time	**0 seconds**
Fetch Size	**0**
Max Rows Per Flow File	**0**

DEMO AND Q&A

EVOLVE NYC

NOV 2, 2023
THE GLASSHOUSE
660 12TH AVE, NEW YORK, NY 10019

presented by
CLOUDERA IBM intel aws

Want to elevate your business with data and AI? Join us at Evolve, a free global live event! Connect with innovators for a day of insights, and leave with actionable solutions that inspire and energize.

FREE LEARNING ENVIRONMENT

Cloudera Streams Processing - Community Edition

- Kafka, KConnect, SMM, SR, Flink, and SSB in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

- Docker compose file of CSP to run from command line w/o any dependencies, including Flink, SQL Stream Builder, Kafka, Kafka Connect, Streams Messaging Manager and Schema Registry
  - `$> docker compose up`
- Licensed under the Cloudera Community License
- **Unsupported**
- Community Group Hub for CSP
- Find it on docs.cloudera.com under Applications

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

Open Source Edition

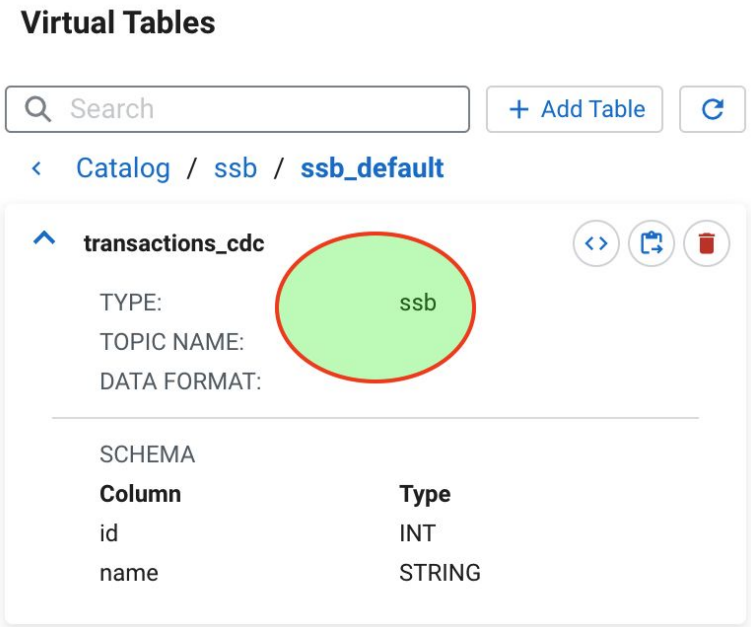
- Apache NiFi in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

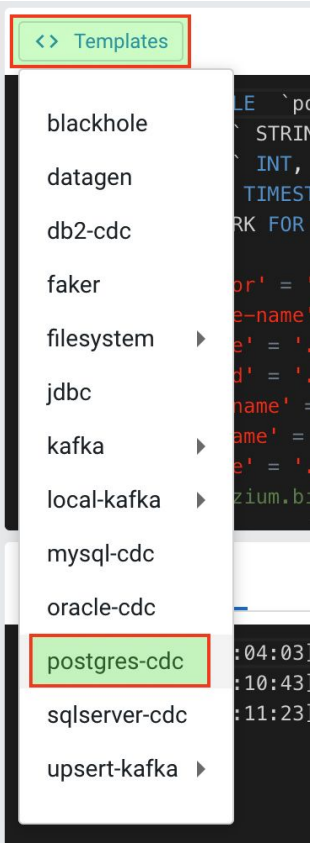
- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

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

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

Cloudera Edge2AI Workshop - CDC - Debezium





https://github.com/asdaraujo/edge2ai-workshop/blob/trunk/workshop_cdc.adoc

RESOURCES AND WRAP-UP

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

References

https://medium.com/@tspann/ingesting-events-into-dockerized-ibm-db2-jdbc-with-apache-nifi-f0ca452d1351

https://community.cloudera.com/t5/Community-Articles/RDBMS-to-Hive-using-NiFi-small-medium-tables/ta-p/244677

https://community.cloudera.com/t5/Community-Articles/MySQL-CDC-with-Kafka-Connect-Debezium-in-CDP-Public-Cloud/ta-p/345321

https://docs.cloudera.com/cdp-private-cloud-base/7.1.9/kafka-connect/topics/kafka-connect-connector-debezium-db2.html

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

https://medium.com/cloudera-inc/building-a-stateful-streaming-intrusion-detection-system-with-sql-stream-builder-4667c87f347f

Streaming Resources

- https://dzone.com/articles/real-time-stream-processing-with-hazelcast-and-streamnative
- https://flipstackweekly.com/
- https://www.datainmotion.dev/
- https://www.flankstack.dev/
- https://github.com/tspannhw
- https://medium.com/@tspann
- https://medium.com/@tspann/predictions-for-streaming-in-2023-ad4d7395d714
- https://www.apachecon.com/acna2022/slides/04_Spann_Tim_Citizen_Streaming_Engineer.pdf

CDC

https://community.cloudera.com/t5/Community-Articles/Incrementally-Streaming-RDBMS-Data-to-Your-Hadoop-DataLake/ta-p/247927

https://community.cloudera.com/t5/Community-Articles/QADCDC-Our-how-to-ingest-some-database-tables-to-Hadoop-Very/ta-p/245229

https://community.cloudera.com/t5/Community-Articles/Ingesting-RDBMS-Data-As-New-Tables-Arrive-Automagically-into/ta-p/246214

https://community.cloudera.com/t5/Community-Articles/Ingesting-Golden-Gate-Records-From-Apache-Kafka-and/ta-p/247557

https://community.cloudera.com/t5/Community-Articles/Simple-Change-Data-Capture-CDC-with-SQL-Selects-via-Apache/ta-p/308390

CDC

https://community.cloudera.com/t5/Community-Articles/RDBMS-to-Hive-using-NiFi-small-medium-tables/ta-p/244677

https://community.cloudera.com/t5/Community-Articles/Simple-Change-Data-Capture-CDC-with-SQL-Selects-via-Apache/ta-p/308376

Cloudera Data Flow / Apache NiFi

https://community.cloudera.com/t5/Community-Articles/Incrementally-Streaming-RDBMS-Data-to-Your-Hadoop-DataLake/ta-p/247927

https://community.cloudera.com/t5/Community-Articles/Ingesting-RDBMS-Data-As-New-Tables-Arrive-Automagically-into/ta-p/246214

https://community.cloudera.com/t5/Community-Articles/Incremental-Fetch-in-NiFi-with-QueryDatabaseTable/ta-p/247073

https://community.cloudera.com/t5/Community-Articles/Simple-Change-Data-Capture-CDC-with-SQL-Selects-via-Apache/ta-p/308390

Cloudera Data Flow / Apache NiFi

https://community.cloudera.com/t5/Community-Articles/Change-Data-Capture-CDC-with-Apache-NiFi-Part-1-of-3/ta-p/246623

https://community.cloudera.com/t5/Community-Articles/Change-Data-Capture-CDC-with-Apache-NiFi-Part-2-of-3/ta-p/246519

https://community.cloudera.com/t5/Community-Articles/Change-Data-Capture-CDC-with-Apache-NiFi-Part-3-of-3/ta-p/246482

Cloudera Streams Messaging / Apache Kafka Connect

https://community.cloudera.com/t5/Community-Articles/MySQL-CDC-with-Kafka-Connect-Debezium-in-CDP-Public-Cloud/ta-p/345321

CDC Debezium KConnectors for PostgreSQL, MySQL, SQL Server, DB2, and Oracle


THANK YOU