FLiP Stack for Cloud Data Lakes

Timothy Spann

Stream Native

Tim Spann, Developer Advocate

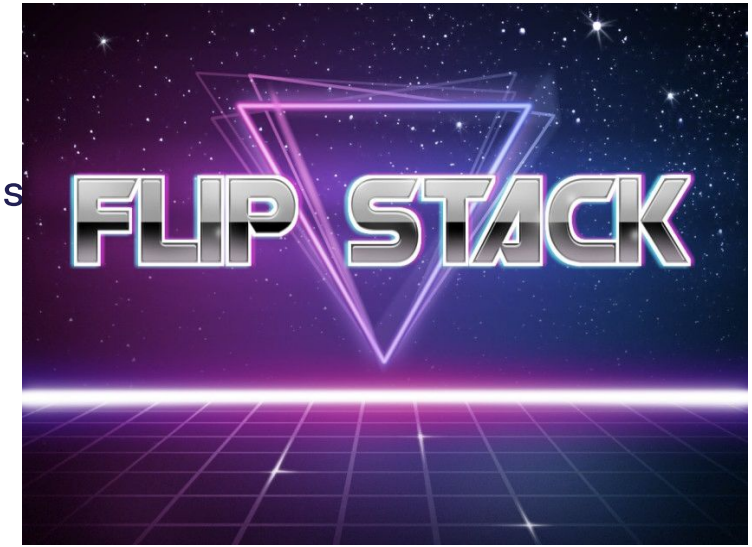
DZone Zone Leader and Big Data MVB Data DJay

StreamNative

Founded by the original developers of Apache Pulsar and Apache BookKeeper, StreamNative builds a cloud-native event streaming platform that enables enterprises to easily access data as real-time event streams.

FLiP(N) Stack

- Apache **Fl**ink
- Apache **P**ulsar
- StreamNative's Flink Connector for Puls
- Apache **N**iFi
- Apache +++



Apache Pulsar

StreamNative

Apache PULSAR is an open source, cloud-native distributed messaging and streaming platform.

What are the Benefits of Pulsar?

Multi-Tenancy

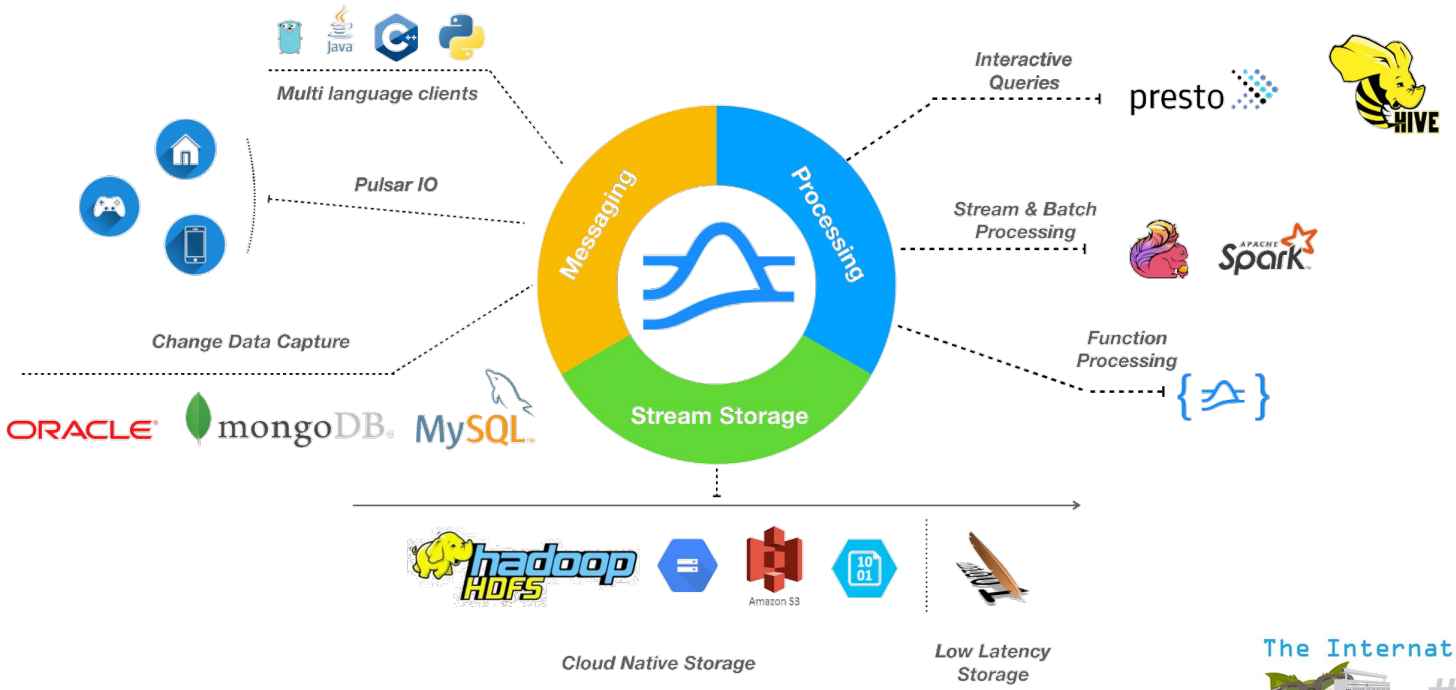
Scalability

Geo-Replication

Unified Messaging Model

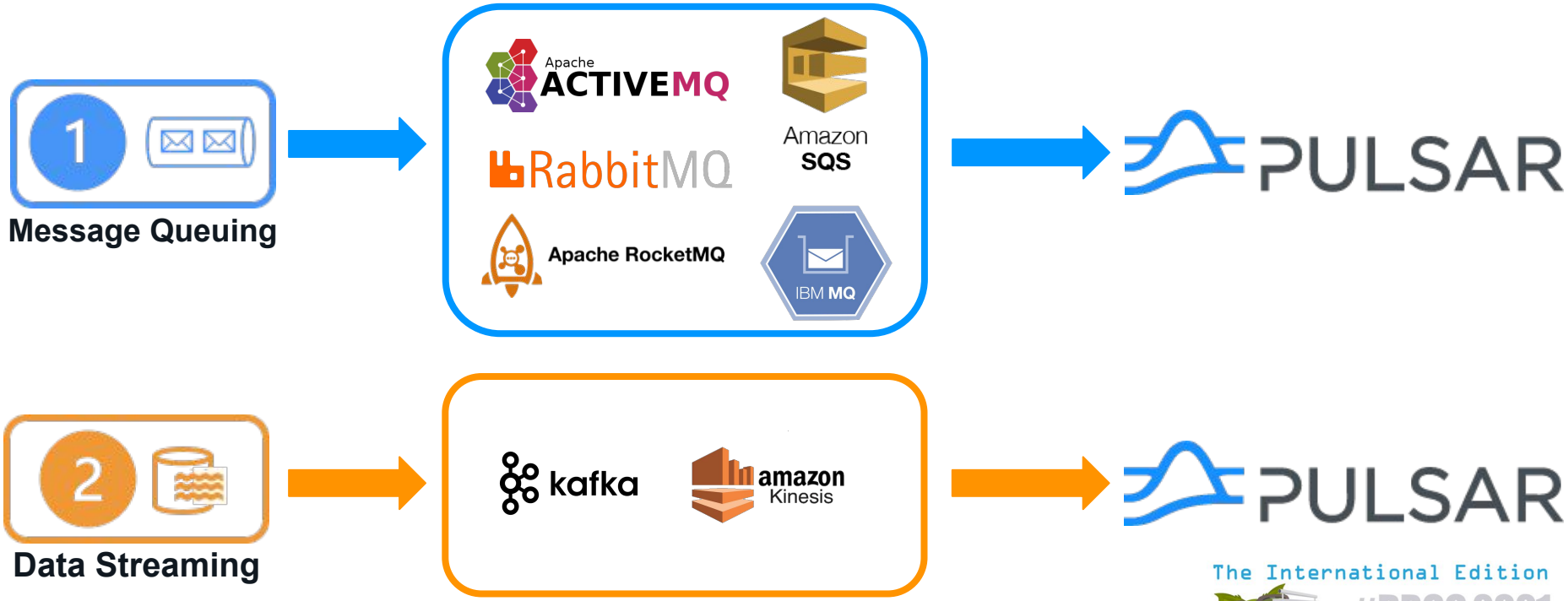
Data Durability

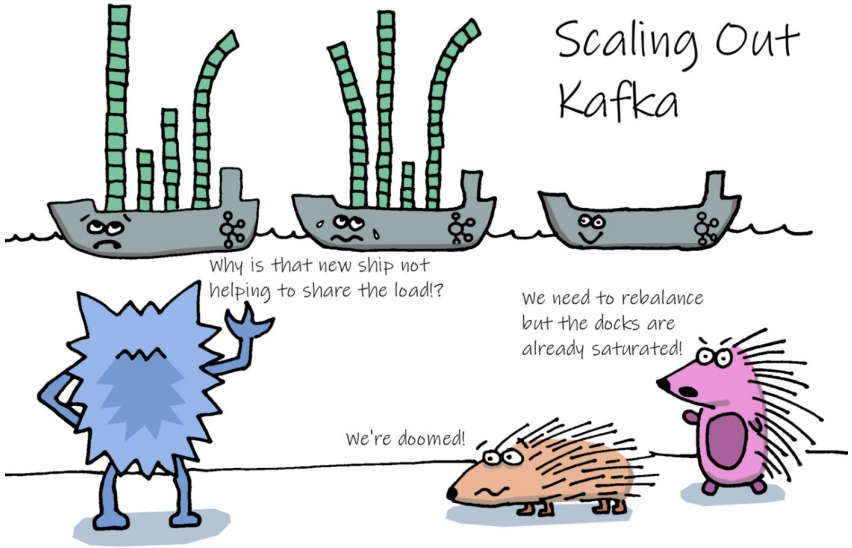
Apache Pulsar

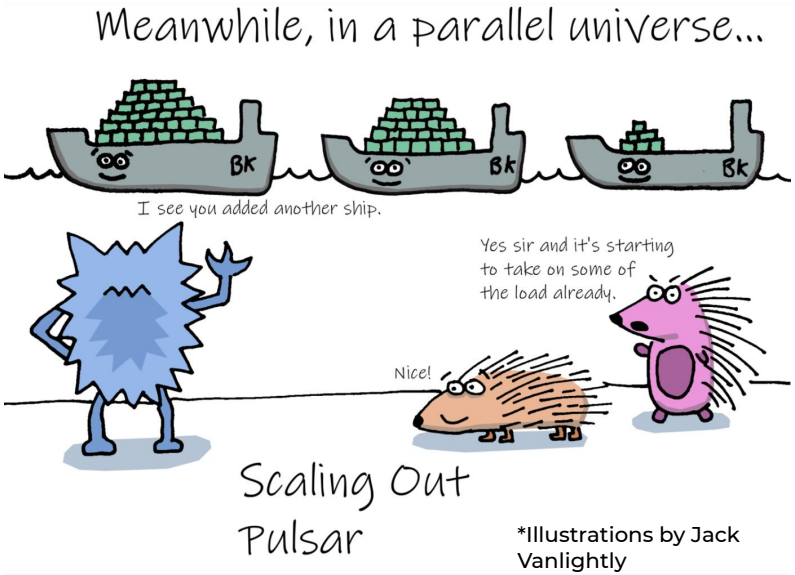


StreamNative

A Unified Messaging Platform





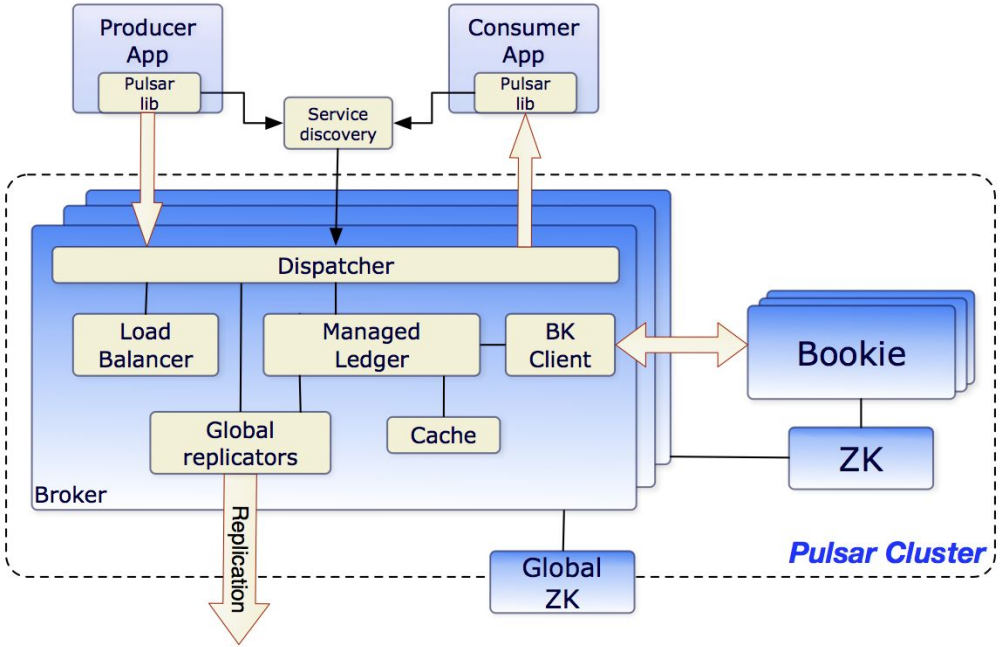


Pulsar is built for easy scale-out.

Apache Pulsar Overview

Enable Geo-Replicated Messaging

- Pub-Sub
- Geo-Replication
- Pulsar Functions
- Horizontal Scalability
- Multi-tenancy
- Tiered Persistent Storage
- Pulsar Connectors
- REST API
- CLI
- Many clients available
- Four Different Subscription Types
- Multi-Protocol Support
  - MQTT
  - AMQP
  - JMS
  - Kafka
  - ...



What is the Pulsar Ecosystem?

- **Functions and Connectors**
  - Functions: Lightweight stream processing
  - Connectors: Part of “Pulsar IO”, includes “Source” and “Sink” APIs
    - Files, Databases, Data tools, Cloud Services, etc
- **Protocol Handlers**
  - Allows Pulsar to handle additional protocols by an extendable API running in the broker
    - AoP (AMQP), KoP (Kafka), MoP (MQTT)

What is the Pulsar Ecosystem? (cont'd)

- **Processing Engines**
  - Supports modern processing engines
    - Flink and Spark, as well as Pulsar SQL (Presto/Trino)
- **Offloaders**
  - Allows data to be offloaded to cloud storage and used with existing Pulsar APIs
    - S3, GCP Cloud Storage, HDFS, File (NFS), Azure Blob Storage (in Pulsar 2.7.0)

Pulsar Functions

Provides a simple API to:

- Receive a message (consume)
- Process the message using your own code
- Send a message (produce)

Takes care of the boilerplate code so there is no need to create producers and consumers.

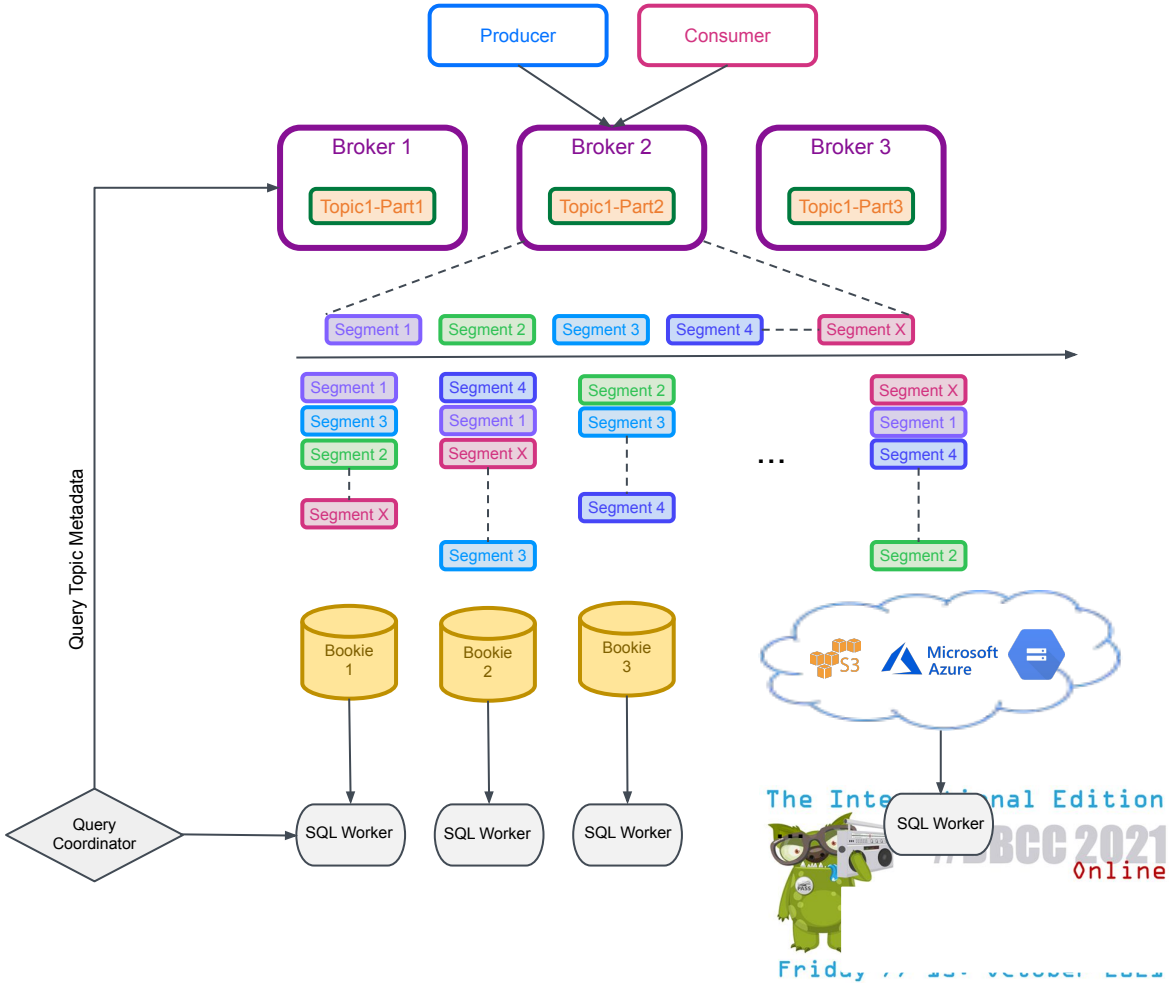
Moving Data In and Out of Pulsar

IO/Connectors are a simple way to integrate with external systems and move data in and out of Pulsar.

- Built on top of Pulsar Functions
- Built-in connectors - hub.streamnative.io

Pulsar SQL

Presto/Trino workers can read segments directly from bookies (or offloaded storage) in parallel.



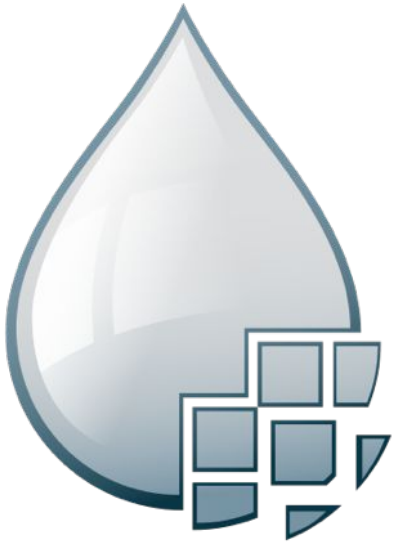
Query Your Topics with Pulsar SQL (Trino)

```
presto> select camera, cpu, cputempf, gputempf, memory, top1, top1pct, uuid, __publish_time__, __message_id__, __key__ from  pulsar."public/default".iotjetsonjson;
  camera    | cpu  | cputempf | gputempf | memory |       top1       |    top1pct     |                uuid                |     __publish_time__     | __message_id__ |
------------+------+----------+----------+--------+------------------+----------------+------------------------------------+--------------------------+----------------+
 /dev/video0 |  8.7 | 82       | 82       |   33.5 | microphone, mike | 18.85986328125 | xav_uuid_video0_lgl_20211001183019 | 2021-10-01 14:30:30.657 | (564,3,0)      |
 /dev/video0 |  8.7 | 82       | 82       |   33.6 | microphone, mike | 19.22607421875 | xav_uuid_video0_kpt_20211001183033 | 2021-10-01 14:30:44.380 | (564,4,0)      |
 /dev/video0 | 12.0 | 80       | 81       |   33.5 | microphone, mike | 12.53662109375 | xav_uuid_video0_gzd_20211001182930 | 2021-10-01 14:29:48.756 | (564,0,0)      |
 /dev/video0 |  8.5 | 82       | 82       |   33.6 | microphone, mike |        14.0625 | xav_uuid_video0_wlw_20211001182951 | 2021-10-01 14:30:02.919 | (564,1,0)      |
 /dev/video0 |  8.5 | 82       | 82       |   33.5 | microphone, mike |     29.8828125 | xav_uuid_video0_ulq_20211001183005 | 2021-10-01 14:30:16.787 | (564,2,0)      |
(5 rows)
(END)
```

```
presto> show tables in pulsar."public/default";
     Table
---------------
 generator_test
 iotjetsonjson
 mqtt-2
(3 rows)

Query 20211001_054538_00008_s8x23, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
0:00 [3 rows, 105B] [14 rows/s, 493B/s]
```

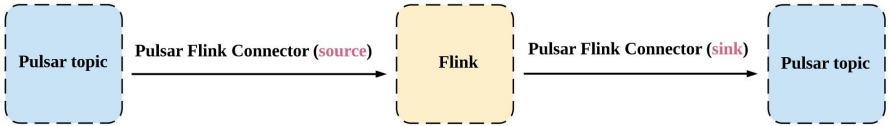
Why Apache NiFi?

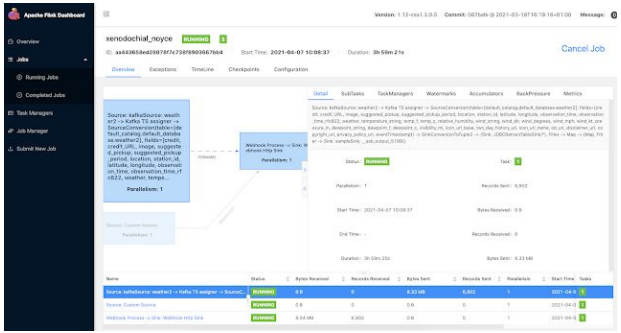


- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models

- Hundreds of processors
- Visual command and control
- Over a 300 sources
- Flow templates
- Pluggable/multi-role security
- Designed for extension
- Clustering
- Version Control

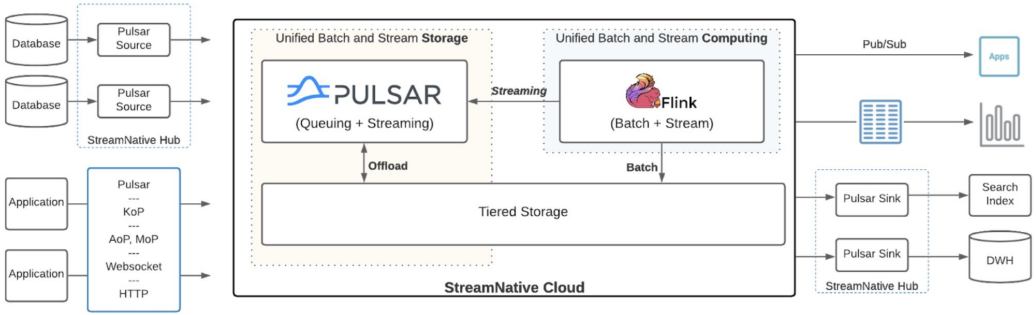
Why Apache Flink?

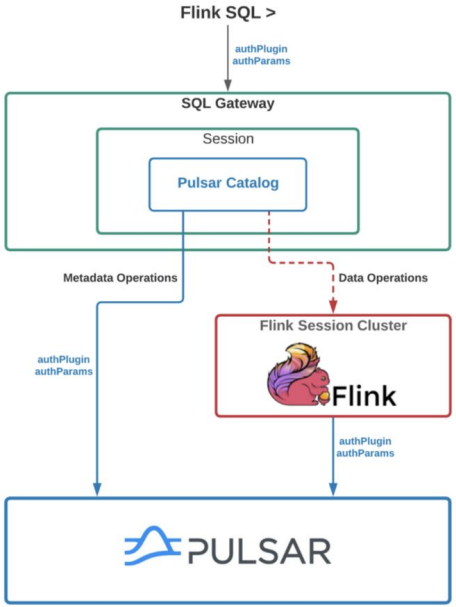


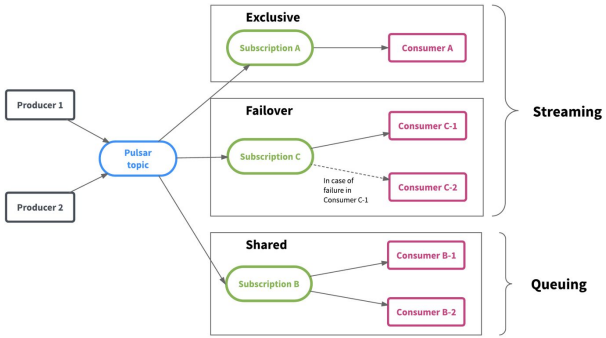


- Unified computing engine
- Batch processing is a special case of stream processing
- Stateful processing
- Massive Scalability
- Flink SQL for queries, inserts against Pulsar Topics
- Streaming Analytics
- Continuous SQL
- Continuous ETL
- Complex Event Processing
- Standard SQL Powered by Apache Calcite

Flink + Pulsar





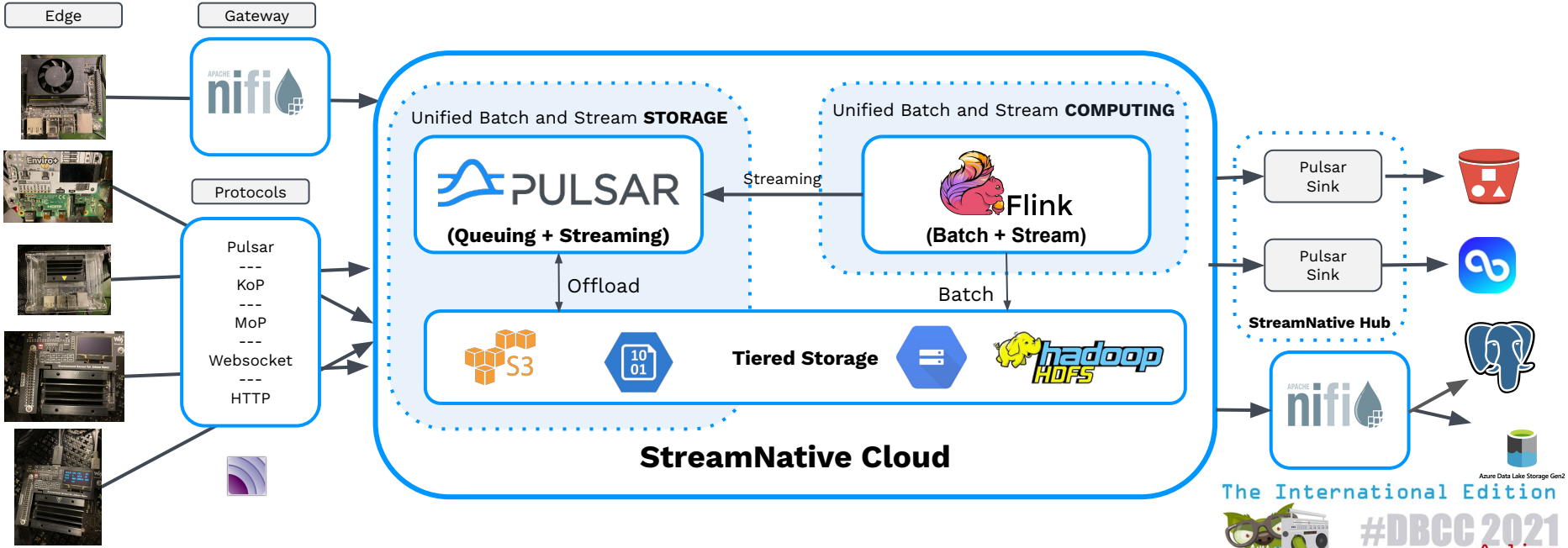


https://flink.apache.org/2019/05/03/pulsar-flink.html
https://github.com/streamnative/pulsar-flink
https://streamnative.io/en/blog/release/2021-04-20-flink-sql-on-streamnative-cloud

StreamNative Cloud

End-to-End Streaming Edge App

Apache Flink - Apache Pulsar - Apache NiFi <-> Devices <-> Cloud Data Lake



StreamNative

The International Edition

#DBCC 2021 Online

DataBlaster CommunityConference

Friday // 15. October 2021

StreamNative Cloud

Powered by Apache Pulsar, StreamNative provides a cloud-native, real-time messaging and streaming platform to support multi-cloud and hybrid cloud strategies.

Cloud Native

Built for Containers

Flink SQL

Demo

StreamNative

Demo Walkthrough

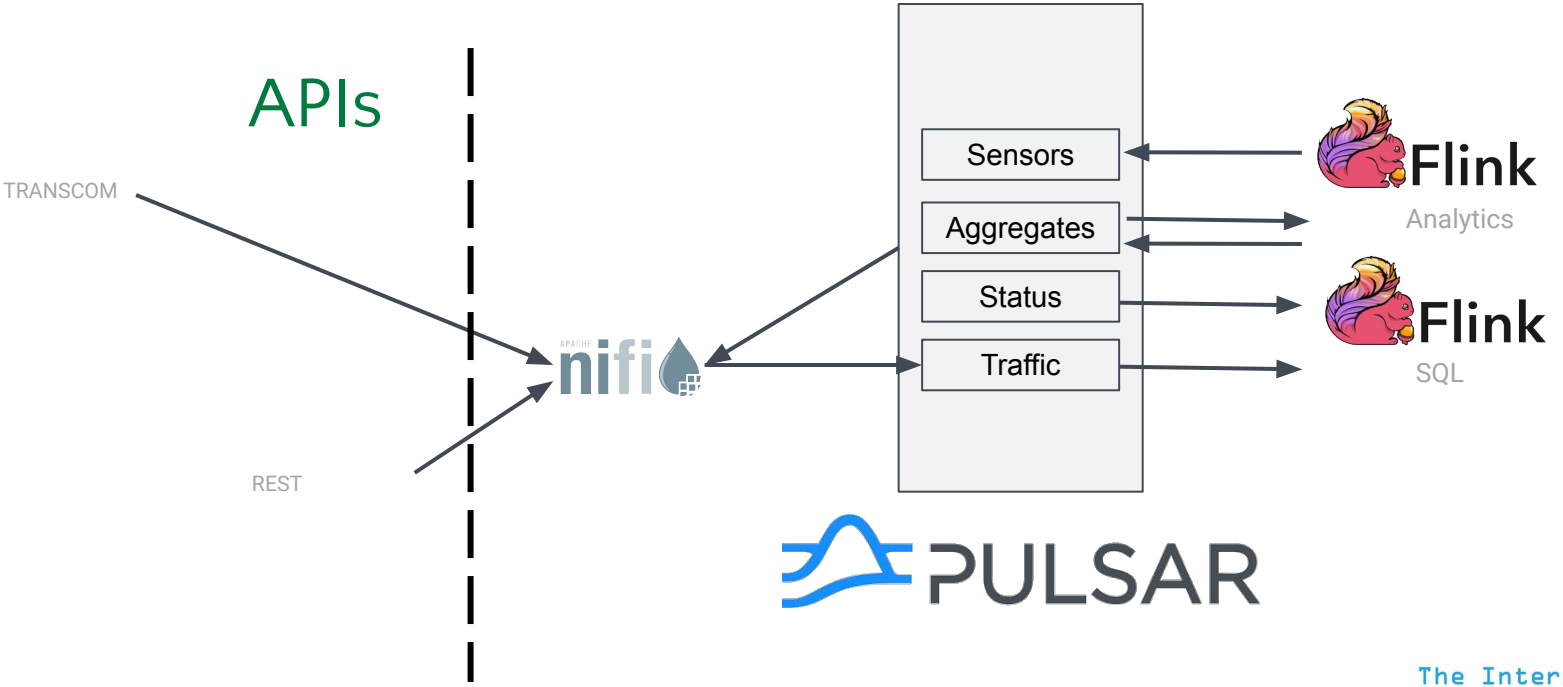
```
Flink SQL> use catalog pulsarcatalog;
[INFO] Execute statement succeed.

Flink SQL> show tables;
+----------------+
|     table name |
+----------------+
|              # |
|   click_events |
|    hello_world |
|        kafka-1 |
|        kafka-2 |
|        kafka-3 |
|        kafka-4 |
|        kafka-5 |
|         mqtt-1 |
|         mqtt-2 |
|         mqtt-3 |
|         mqtt-4 |
|         mqtt-5 |
|        mqtt-go |
|       mqtt-mac |
|      mqtt-nifi |
|    mqtt-nvidia |
|    mqtt-python |
|       mqtt-rp4 |
|       my-topic |
| nvidia-kafka-1 |
|    rp4-kafka-1 |
|           rwar |
+----------------+
23 rows in set

Flink SQL>
```

{"entriesAddedCounter":1,"numberOfEntries":1,"totalSize":651,"currentLedgerEntries":1,"currentLedgerSize":651,"lastLedgerCreatedTimestamp":"2021-09-13T16:13:06.6-04:00","waitingCursorsCount":0,"pendingAddEntriesCount":0,"lastConfirmedEntry":"7076:0","state":"LedgerOpened","ledgers":[{"ledgerId":7076,"entries":0,"size":0,"offloaded":false,"underReplicated":false}],"cursors":{},"schemaLedgers":[],"compactedLedger":{"ledgerId":-1,"entries":-1,"size":-1,"offloaded":false,"underReplicated":false}}

Real-Time Cloud Streaming Pipeline



Wrap-Up

StreamNative

Deeper Content

- https://www.datainmotion.dev/2020/04/building-search-indexes-with-apache.html
- https://github.com/tspannhw/nifi-solr-example
- https://github.com/streamnative/pulsar-flink
- https://www.linkedin.com/pulse/2021-schedule-tim-spann/
- https://github.com/tspannhw/SpeakerProfile/blob/main/2021/talks/20210729_HailHydrate!FromStreamtoLake_TimSpann.pdf
- https://streamnative.io/en/blog/release/2021-04-20-flink-sql-on-streamnative-cloud
- https://docs.streamnative.io/cloud/stable/compute/flink-sql

@PaasDev

timothyspann

https://www.pulsardeveloper.com/

Interested In Learning More?

Resources

[Flink SQL Cookbook]

[The Github Source for Flink SQL Demo]

[The GitHub Source for Demo]

Free eBooks

[Manning's Apache Pulsar in Action]

[O'Reilly Book]

Upcoming Events

[10/21] Trino Summit

PULSAR SUMMIT

Pulsar Summit Asia

November 20–21, 2021

Contact us at partners@pulsar-summit.org to become a sponsor or partner

influx/days

Timothy Spann

Developer Advocate,
StreamNative

Using FLaNK with InfluxDB for EdgeAI IoT at Scale

InfluxDays North America 2021 | October 27, 2021

trino SUMMIT

10:45 - 11:25am EST
7:45 - 8:25am PST
3:45 - 4:25pm BST
8:15 - 8:55pm IST

FLiP Into Trino

Remember the days when you could wait until your batch data load was done and then you could run some simple queries or build stale dashboards? Those days are over, today you need instant analytics as the data is streaming in real-time. You need universal analytics where that data is. I will show you how to do this utilizing the latest cloud native open source tools. In this talk we will utilize Trino, Apache Pulsar, Pulsar SQL and Apache Flink to analyze instantly data from IoT, sensors, transportation systems, Logs, REST endpoints, XML, Images, PDFs, Documents, Text, semistructured data, unstructured data, structured data and a hundred data sources you could never dream of streaming before. I will teach how to use Pulsar SQL to run analytics on live data.

Tim Spann
Developer Advocate
StreamNative

David Kjerrumgaard
Developer Advocate
StreamNative

StreamNative

The International Edition
#DBCC 2021
Online
DataBlaster
CommunityConference
Friday // 15. October 2021

Wednesday, October 27, 2021

9:00 AM - 9:25 AM PDT

AI DEVWORLD | AI OPEN TALKS | AI FOR THE ENTERPRISE

OPEN TALK (AI): Utilizing Apache Kafka, Apache NiFi and MiNiFi for EdgeAI IoT at Scale

Timothy Spann
StreamNative, Developer Advocate

Add

Thursday, October 28, 2021

9:00 AM - 9:25 AM PDT

API INNOVATION | EMERGING APIS

PRO TALK (API): Apache NiFi 101: Introduction and Best Practices

Timothy Spann
StreamNative, Developer Advocate

Add

StreamNative

The International Edition
#DBCC 2021
Online
DataBlaster
CommunityConference
Friday // 15. October 2021

Questions

Let’s Keep in Touch!

Speaker Name

Speaker title

@PassDev

https://www.linkedin.com/in/timothyspann

https://github.com/tspannhw

StreamNative

Thank you!

Please submit your feedback here:

https://bit.ly/dbcc2021-feedback

The International Edition
#DBCC 2021
Online
DataBlaster
CommunityConference
Friday // 15. October 2021