ScyllaDB Summit 2023

Sink Your Teeth into Streaming at Any Scale

David Kjerrumgaard, Developer Advocate at Stream Native

Tim Spann, Developer Advocate at Stream Native

StreamNative

David Kjerrumgaard

Stream Native

- Apache Pulsar Committer | Author of Pulsar In Action
- Former Principal Software Engineer on Splunk's messaging team that is responsible for Splunk's internal Pulsar-as-a-Service platform.
- Former Director of Solution Architecture at Streamlio.

HORTONWORKS FedEx LOCKHEED MARTIN splunk>

Tim Spann

StreamNative

- FLiP(N) Stack = Flink, Pulsar and NiFi Stack
- Streaming Systems & Data Architecture Expert
- DZone Iot and ML Top Expert
- 15+ years of experience with streaming technologies including Pulsar, Flink, Spark, NiFi, Big Data, Cloud, MXNet, IoT, Python and more.
- Today, he helps to grow the Pulsar community sharing rich technical knowledge and experience at both global conferences and through individual conversations.

CLOUDERA Pivotal Hewlett Packard Enterprise HORTONWORKS

Presentation Agenda

- The Team to Stream
- What is Apache Pulsar?
- Join The Streams with Flink SQL
- Fast Sink to ScyllaDB
- Reference Architecture
- Demo

Stream Team

Building Real-Time Requires a Team

Apache Pulsar

Apache Pulsar is a Cloud-Native Messaging and Event-Streaming Platform.

Apache Pulsar Timeline

2012

CREATED

Originally developed inside Yahoo! as Cloud Messaging Service

2016

OPEN SOURCE

Pulsar committed to open source.

2018

APACHE TLP

Pulsar becomes Apache top level project.

TODAY

GROWTH

10x Contributors

10MM+ Downloads

Ecosystem Expands

Kafka on Pulsar

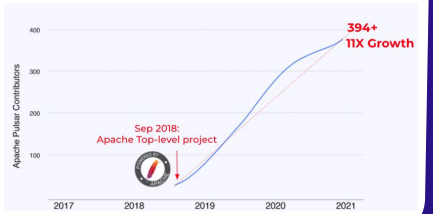
AMQ on Pulsar

Functions

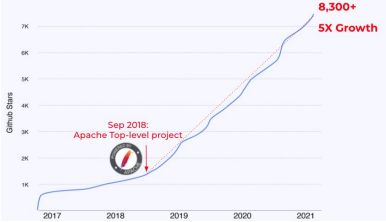
...

Pulsar Growth

11X Growth



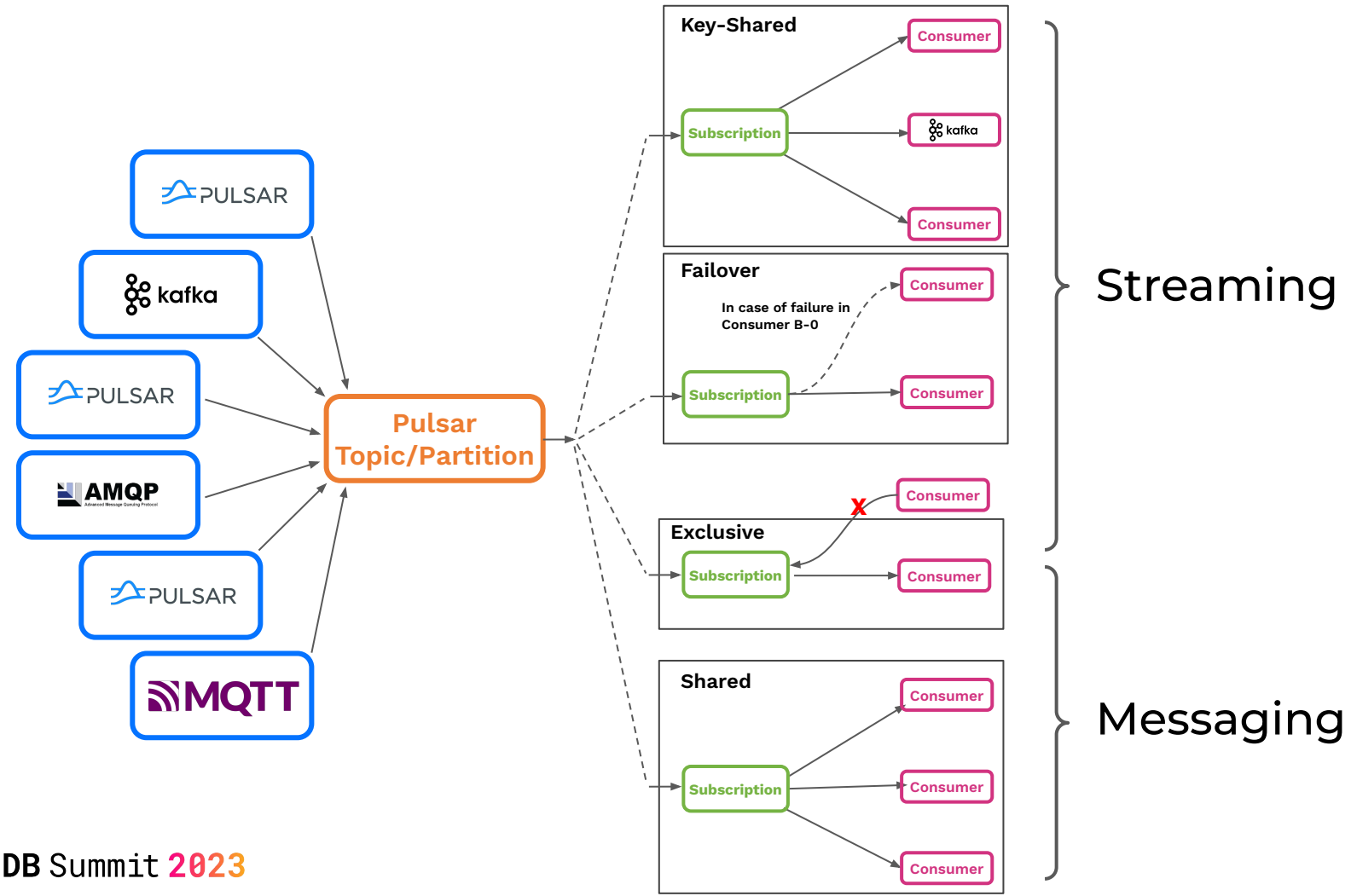
5X Growth



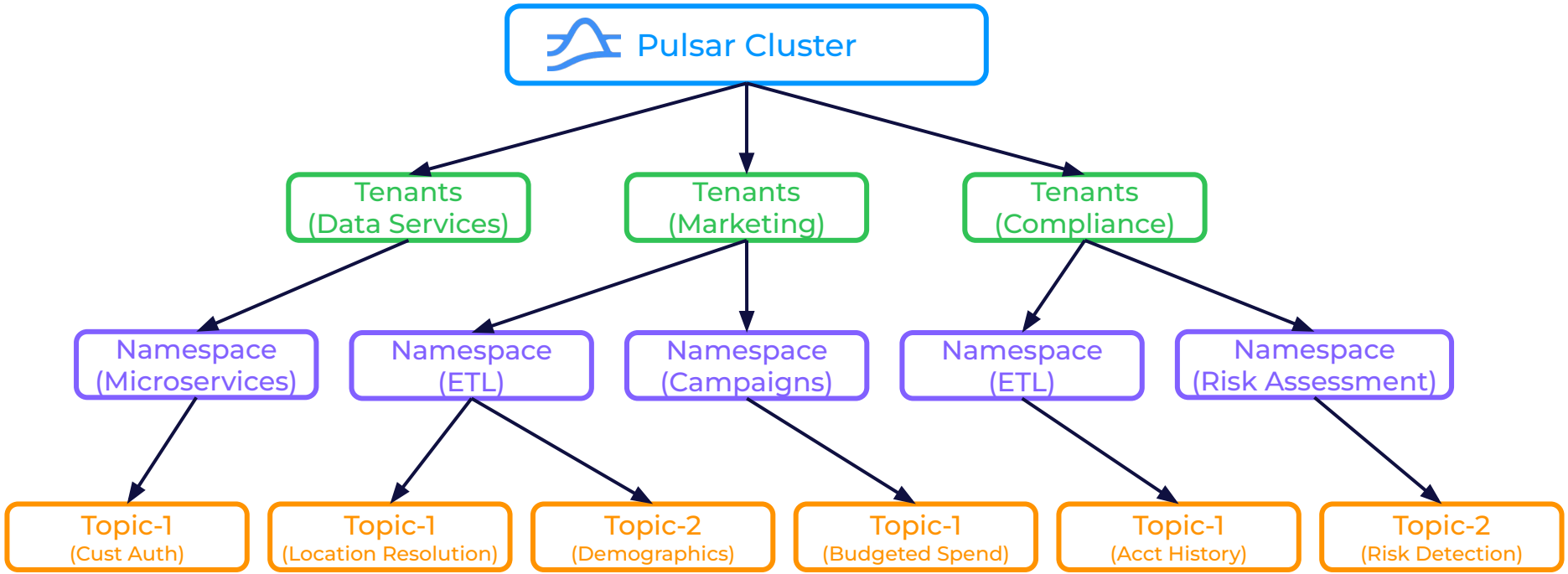
10,000,000+

Docker Images

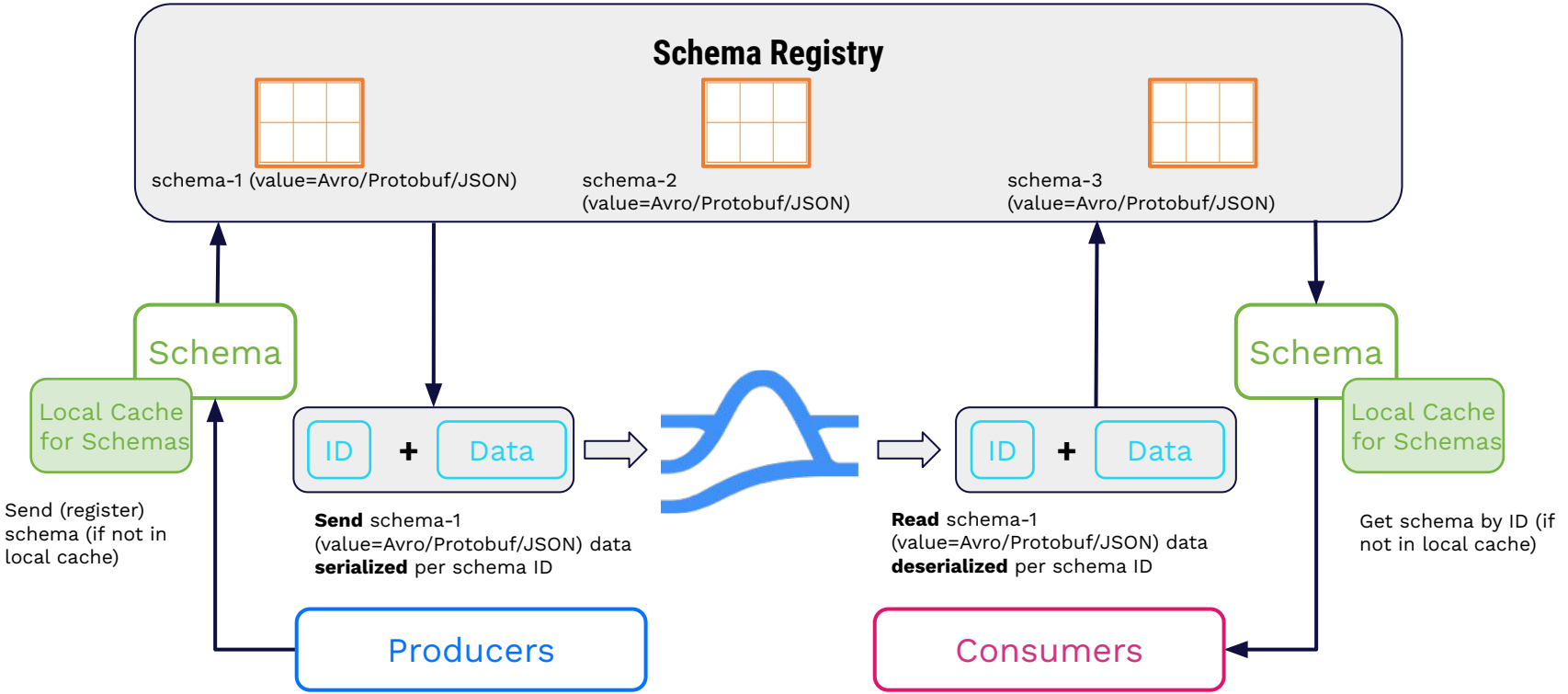
Contributors | **Github Stars** | **Downloads**


PULSAR
kafka
PULSAR
AMQP
PULSAR
MQTT
Pulsar Topic/Partition
Key-Shared
Subscription
Consumer
kafka
Consumer
Failover
In case of failure in Consumer B-0
Subscription
Consumer
Consumer
Exclusive
X
Consumer
Subscription
Consumer
Shared
Subscription
Consumer
Consumer
Consumer
Streaming
Messaging

Multi-Tenancy Model

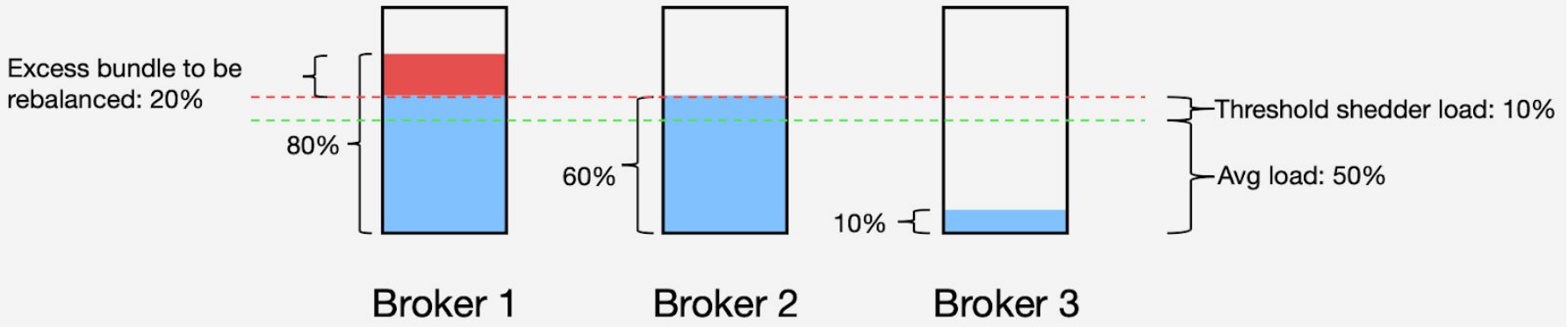


Integrated Schema Registry



Automatic Load Balancing

Pulsar supports automatic load shedding whenever the system recognizes a particular broker is overloaded, the system forces some traffic to be reassigned to less-loaded brokers.

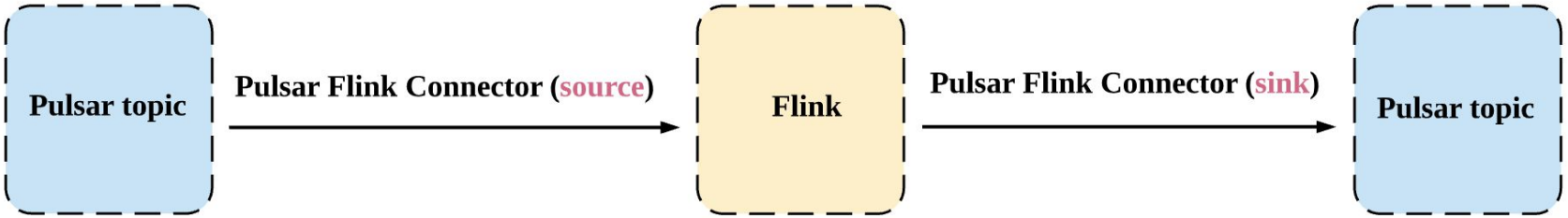


Key Features	Apache Pulsar	Apache Kafka
Message retention (time-based)	✔	✔
Message replay	✔	✔
Message retention (acknowledge-based)	✔	✘
Built-in tiered storage	✔	✘
Processing capabilities (fully managed)	Pulsar Functions	KStream
Queue semantics (round robin)	✔	✘
Queue semantics (key based)	✔	✘
Dead letter queue	✔	✘

Key Features	Apache Pulsar	Apache Kafka
Scheduled and delayed delivery	✔	✔
Rebalance-free scaling	✔	✔
Elastically scalable	✔	✘
Maximum number of topics	Millions	Up to 100k
Built-in multi-tenancy	✔	✘
Built-in geo-replication	✔	✘
Built-in schema management	✔	✘
End-to-end encryption	✔	✘

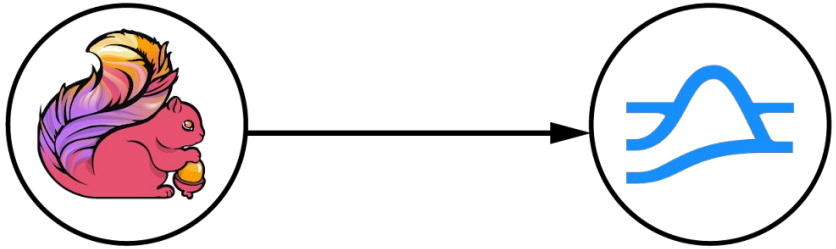
Flink SQL

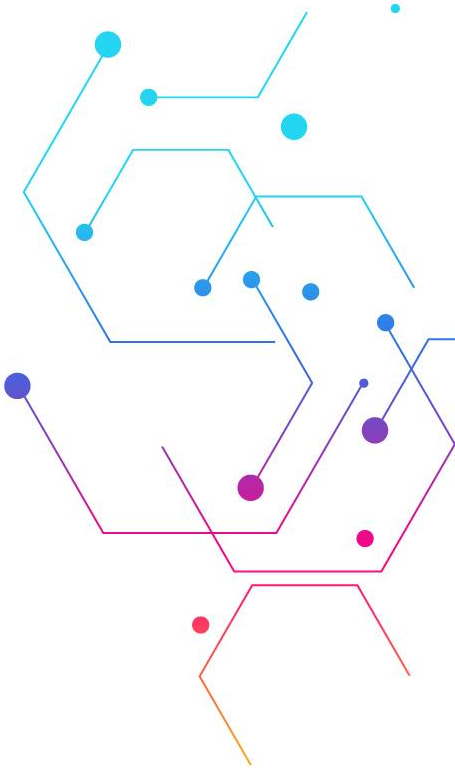
Pulsar-Flink Connector



- Pulsar provides a connector that enables Flink to read & write data directly from/to Pulsar topics.
- This makes Pulsar Topic data available for Flink SQL queries.
- https://hub.streamnative.io/data-processing/pulsar-flink/1.15.1.1/

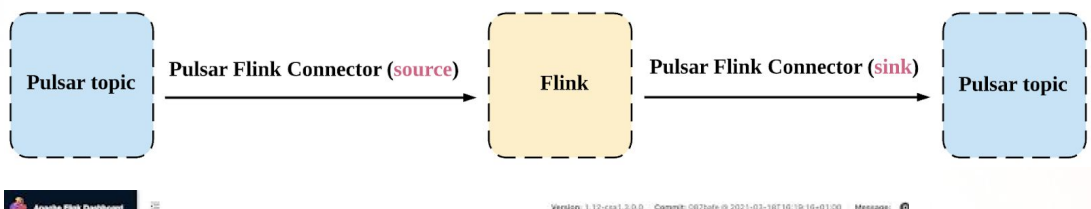
Flink SQL

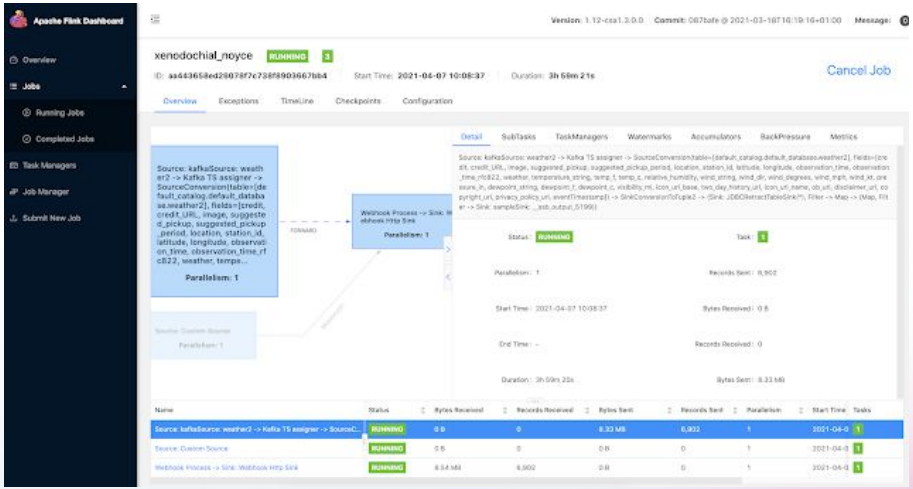




Apache Flink

- 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



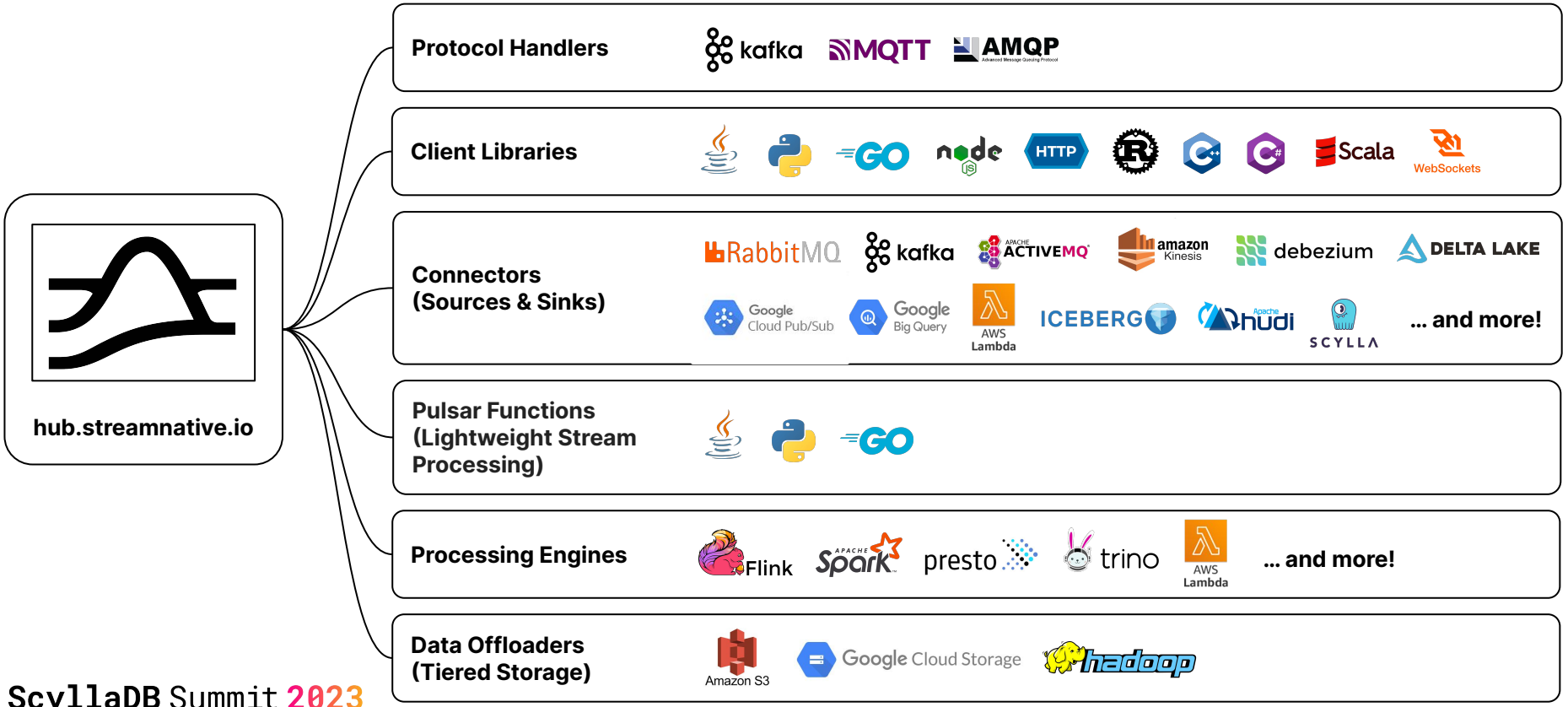


Streaming Tables

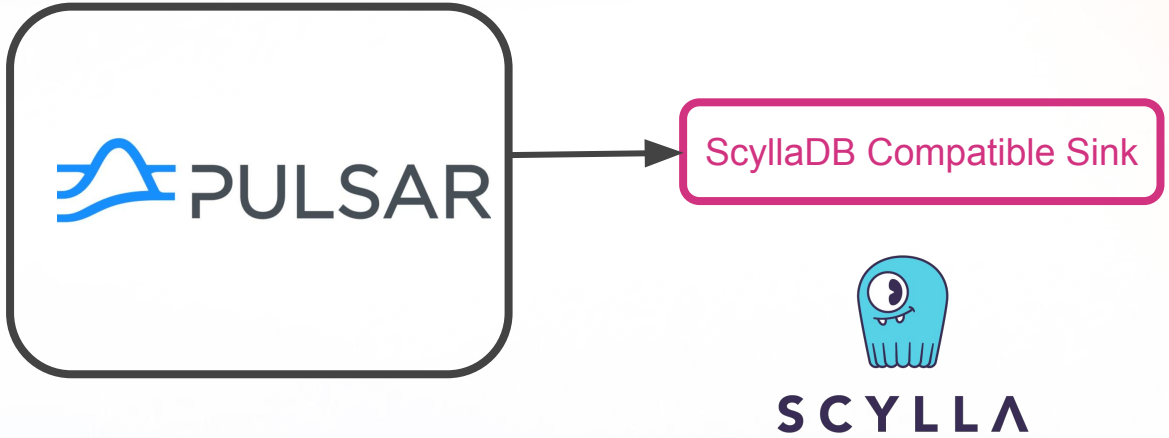
- A streaming table represents an unbounded sequence of structured data, i.e. “facts”
- Facts in a streaming table are ***immutable***, which means new events can be inserted into a table, but existing events can never be updated or deleted.
- All the topics within a Pulsar namespace will automatically be mapped to streaming tables in a catalog configured to use a pulsar connector.
- Streaming tables can also be created or deleted via DDL queries, where the underlying Pulsar topics will be created or deleted.

Sink to ScyllaDB

StreamNative Pulsar ecosystem



ScyllaDB Compatible Sink Connector



pulsar.apache.org/docs/en/io-quickstart/

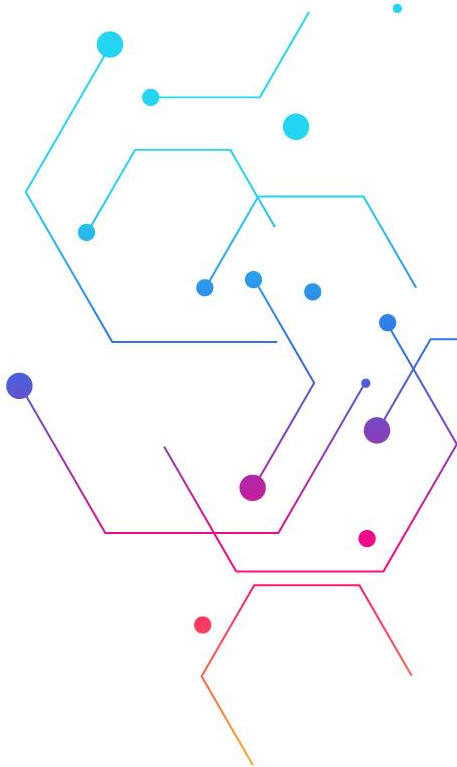
David's ScyllaDB Sink Update

Connector automatically interrogates the database to determine the schema type definition at runtime.

Includes a framework that extracts the values from the generic schema types (GenericRecord, and JSON String)

Performance improvements & Bug Fixes

PR: https://github.com/apache/pulsar/pull/16179



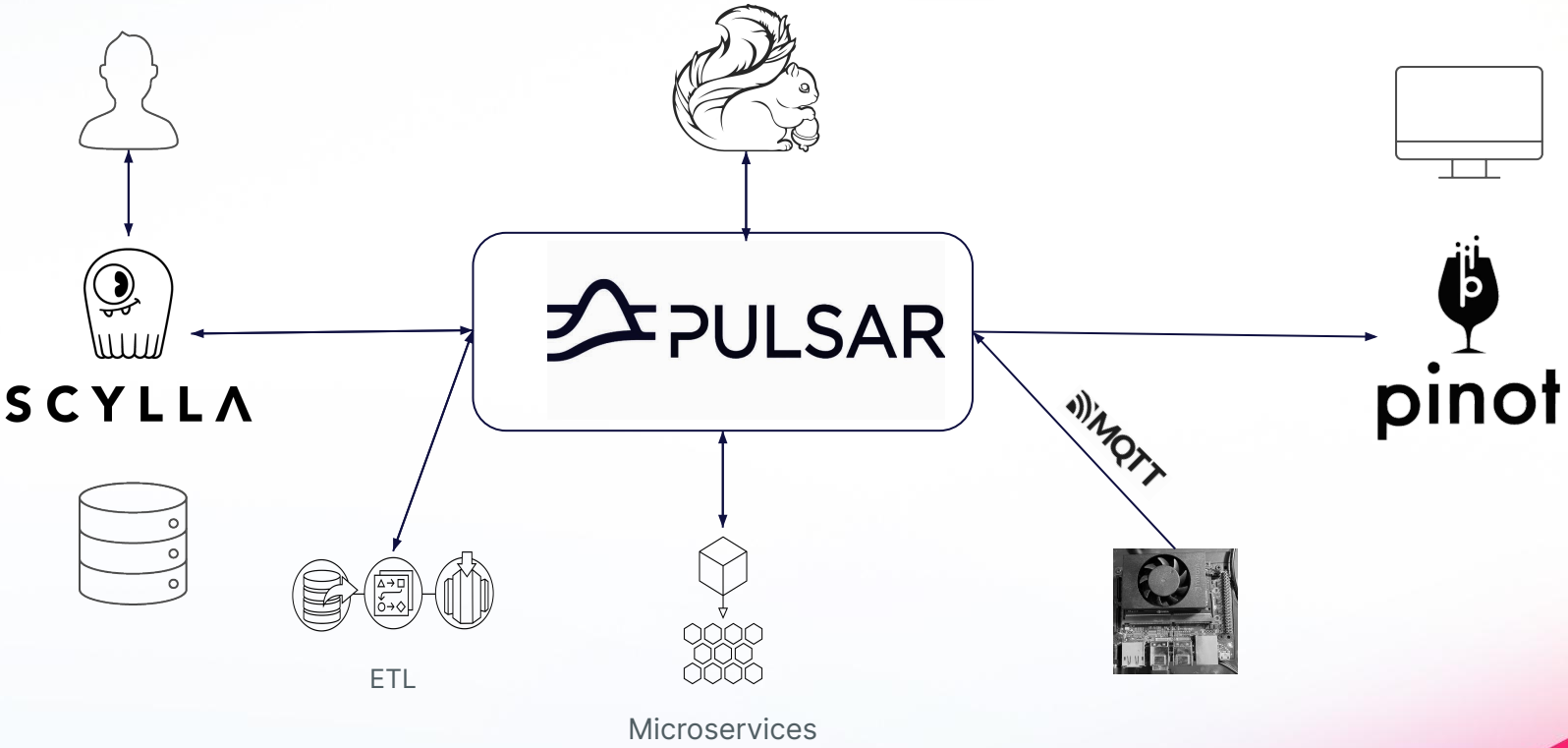
Building Real-Time Together

Easily integrate ScyllaDB's fast distributed database system and the StreamNative streaming platform with Pulsar sources and sinks via various protocols and libraries.

Create high-speed streaming pipelines for log analysis, IoT sensing, and more.

Reference Architecture

Reference Architecture



Reference Architecture

ScyllaDB <-> Apache Pulsar -> Flink SQL -> Apache Pulsar -> Apache Pinot

Flink SQL for Continuous Analytics, Ingest, Real-time Joins, Real-Time Analytics

Apache Pulsar for routing, transformation, central messaging hub, data channels

ScyllaDB for instant applications and massive data feeds

Apache Pinot for low latency instant result queries

Thank You

Stay in Touch

David Kjerrumgaard

david@streamnative.io

@DavidKjerrumga1

https://github.com/david-streamlio

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

Timothy Spann

tim@streamnative.io

@PaaSDev

https://github.com/tspannhw

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