DATASTAX
DEVELOPERS

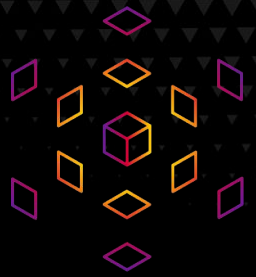
CASSANDRA SUMMIT
MARCH 13-14, 2023 • SAN JOSE, CA

TRAINING DAY

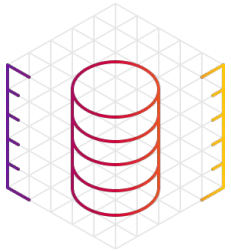
CASSANDRA SUMMIT
MARCH 13-14, 2023 • SAN JOSE, CA

APACHE CASSANDRA® FOR ARCHITECTS AND DATA ENGINEERS:

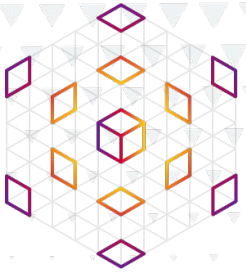
3 - Event Streaming with Pulsar



#1
Introduction to Apache Pulsar™



Event Streaming <===> Message Streaming

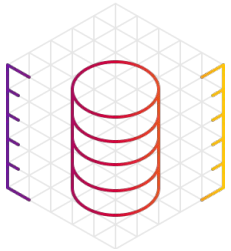


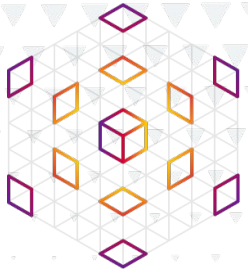
Event streaming

Message streaming

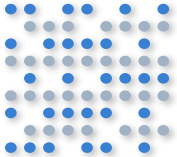
- Watch for events with “the system” or application
- Publish messages and receive events
- Make decisions on data in real time
- Ingest high frequency of messages with very low latency and consume at a different rate

Streaming vs Not Streaming





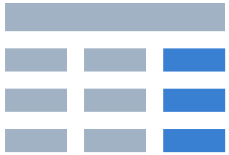
Streaming



Ingest data

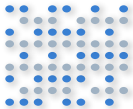
Process data

Sink data



Select data

Not Streaming



Ingest data

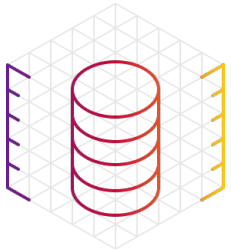
Persist data

Select data

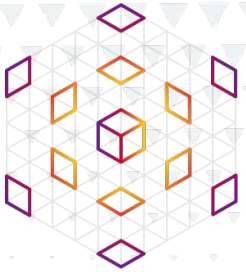
Process data

Persist data

Select data



Apache Pulsar™



Open source

Created by Yahoo
Contributed to the Apache Software Foundation 2016
Top-level project 2018

Cloud-native design

Cluster based
Multi-tenant
Simple client APIs (Java, C#, Python, Go, Node, ...)
Separate compute and storage!

Guaranteed message delivery

If a message successfully reaches a Pulsar broker, it will be delivered to its intended target.

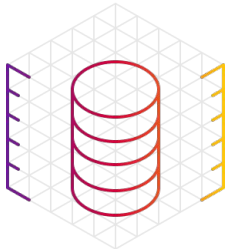
Light-weight serverless functions framework

Create complex processing logic within a Pulsar cluster (aka: data pipeline)

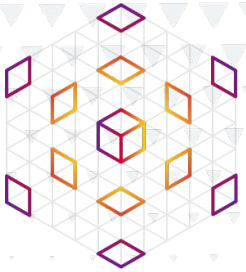
Tiered storage offloads

Offload data from hot/warm storage to cold/long-term storage when the data is aging out

PULSAR



Pulsar Delivers in Ways Other Streaming Platforms Can't



Distributed Architecture

Pulsar separates processing, storage, and platform management to provide improved operations, scalability, and high availability.

Geo-Replication

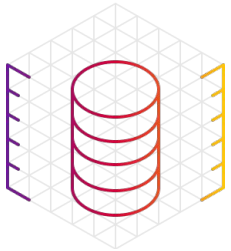
Out-of-the-box support for message replication across data centers. Producers and consumers can interact with topics regardless of their location.

Multi-tenancy

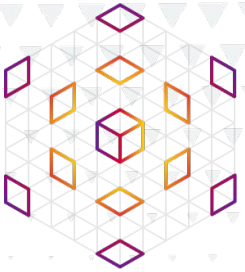
Consolidated messaging/streaming platform which provides effective permission control within business domain context, and better IT resource utilization reducing Total Cost of Ownership (TCO)

Message Delivery

Pulsar supports four subscription types giving consumers control and providing queuing, guaranteed ordering, and guaranteed delivery.



Pulsar Components



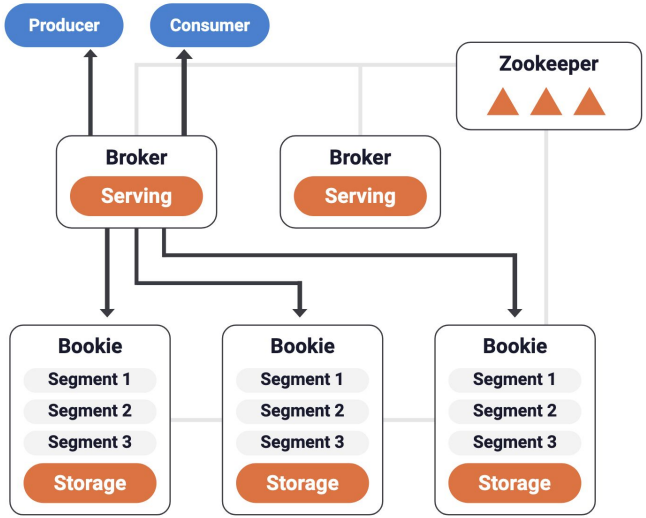
Producer
Client application sending messages to topic managed by Broker

Consumer
Client application reading messages from a topic managed by Broker

Broker
A stateless process that handles incoming message, message dispatching, communicates with the Pulsar configuration store, and stores messages in BookKeeper instances

BookKeeper
Persistent message store

ZooKeeper
Holds cluster metadata, handles coordination tasks between Pulsar clusters



Cloud Native Architecture

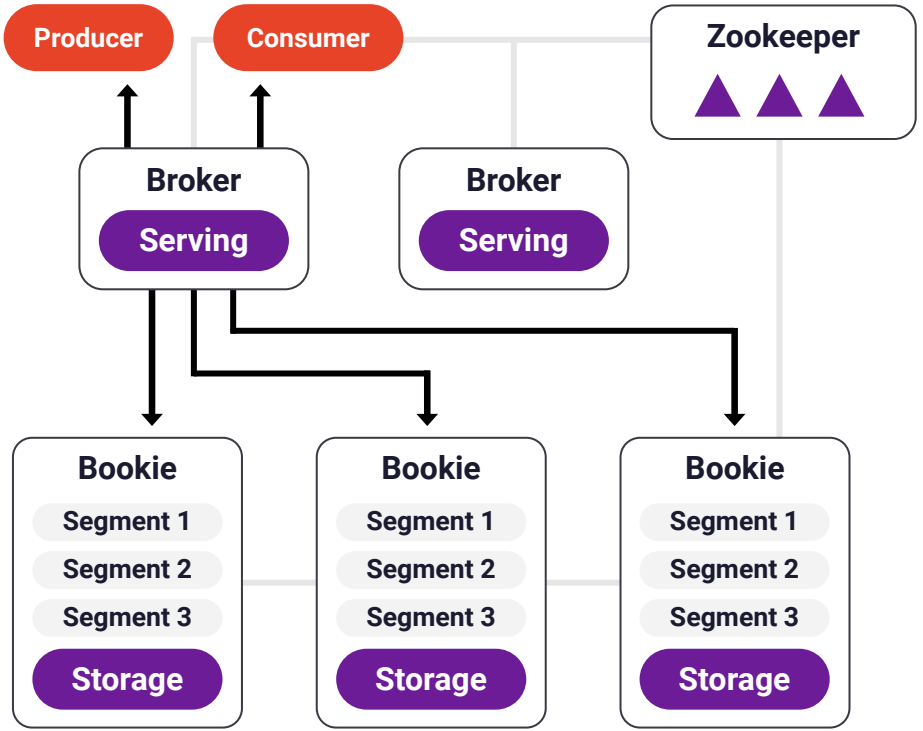
Distributed, tiered architecture

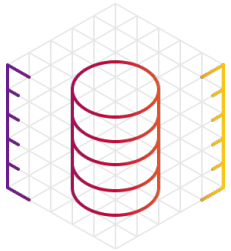
Separated compute from storage

Zookeeper holds metadata for the cluster

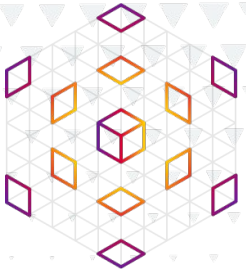
Stateless Broker handles producers and consumers

Storage is handled by Apache Bookkeeper





Astra Streaming



Pulsar-as-a-Service

Streaming-as-a-Service built on Apache Pulsar

No Operations

Eliminate the overhead to install, operate, and scale Pulsar

Powerful Tools and APIs

Leverage the same tools used to interact with Pulsar on prem

Cloud Native

Built to run on any cloud

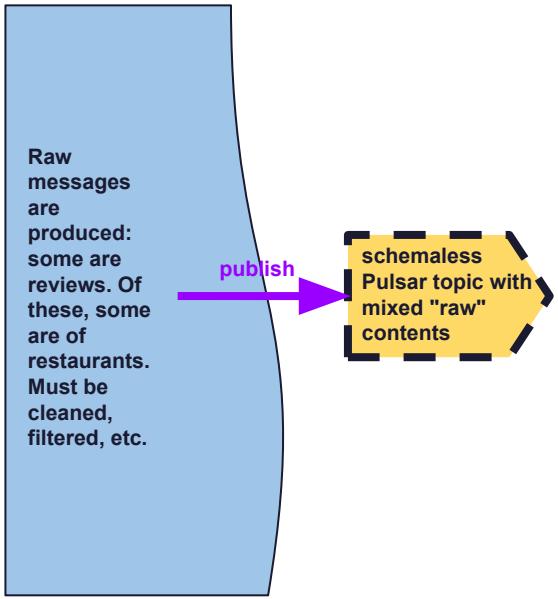
Zero Lock-in

Leverage Pulsar's built in integration with existing developer tools

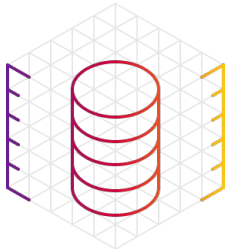
Start for Free

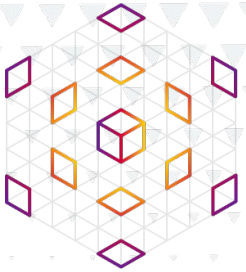
Free monthly credits to help you get started quickly

Business Architecture



Logical Architecture





Review Injector

```
{
 type:"hotel",
 rating: 4.5,
 comment:"--"
}
```

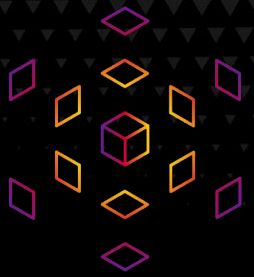
python

Publisher

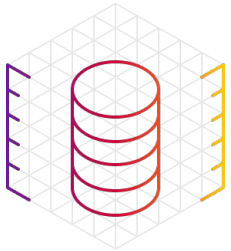
Topic

Consumer

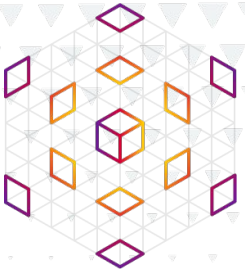
Output on terminal



#2
Datascience in Event Streaming



Data Science with events



Fraud Detection

Needed to ingest high-speed writes of customer event traffic for real-time fraud detection and analysis. Geo-replication must have little to no latency.

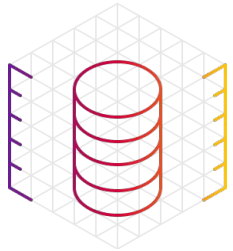
Secure Social Media, Protect Customer Privacy

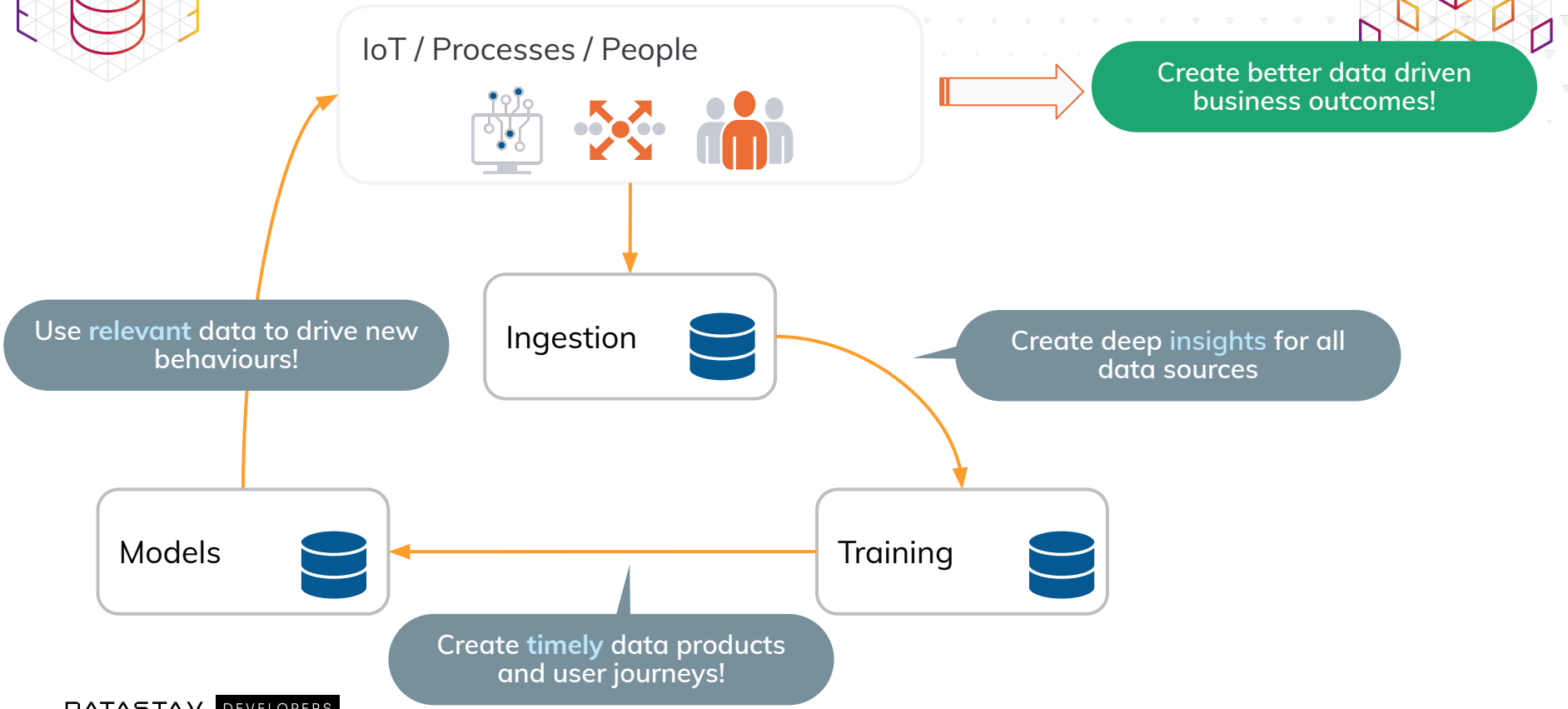
Identify out-of-the-ordinary patterns to prevent malicious attacks on digital and physical assets from unauthorized applications and individuals.

IoT Data Ingestion and Classification

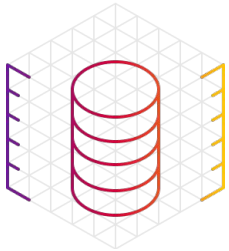
Take in high speed data with very little latency, while processing at a different [slower] speed internally.

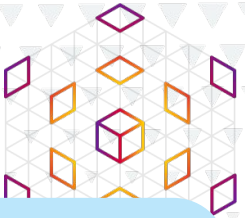
Real-time data pipelines

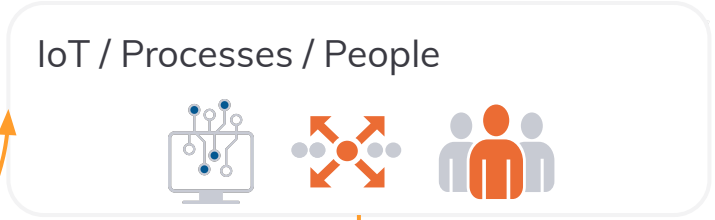




What is not working ?







Little demonstrable value from data

Limited data sources, data left unexamined

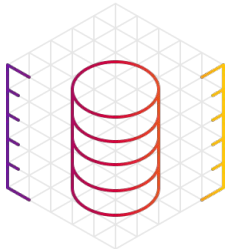
Ingestion

Models outdated in production

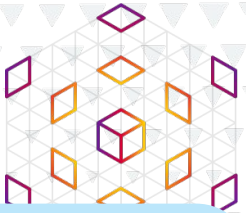
Models

Training

Data quality & quantity create irrelevant models



Cassandra and Pulsar



IoT / Processes / People

Significant **business outcomes achieved!**

Publish **time-sensitive** models - **faster**

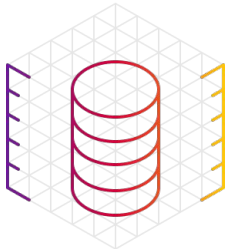
Ingestion

Remove complexity of pipelines & lakes

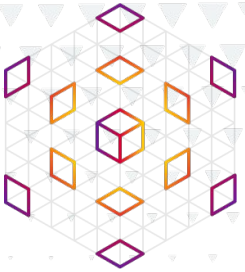
Models

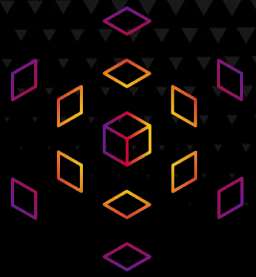
Training

Deeper analysis of data sets to enrich the models - **faster**

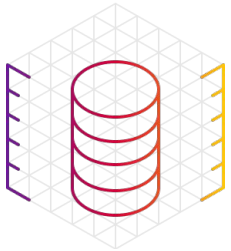


Our Data pipeline today

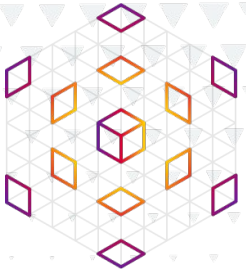




#3
Cassandra CDC & Pulsar Functions



❯ Pulsar Functions



Serverless function platform purpose-built for streaming data pipelines.

Simple Function Architecture

Triggered from input topic

Simple programmatic interface

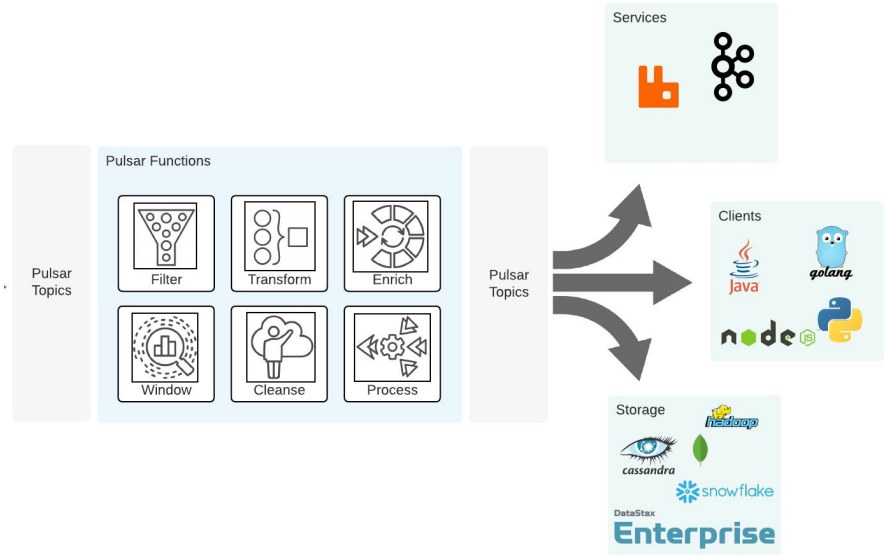
Push function result to output topic

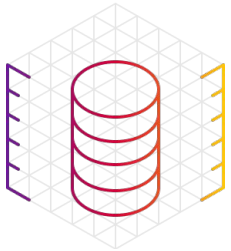
Built for DevOps

Standard Kubernetes based runtime

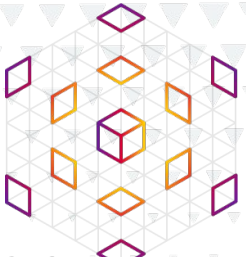
Automated deployments

CI/CD friendly

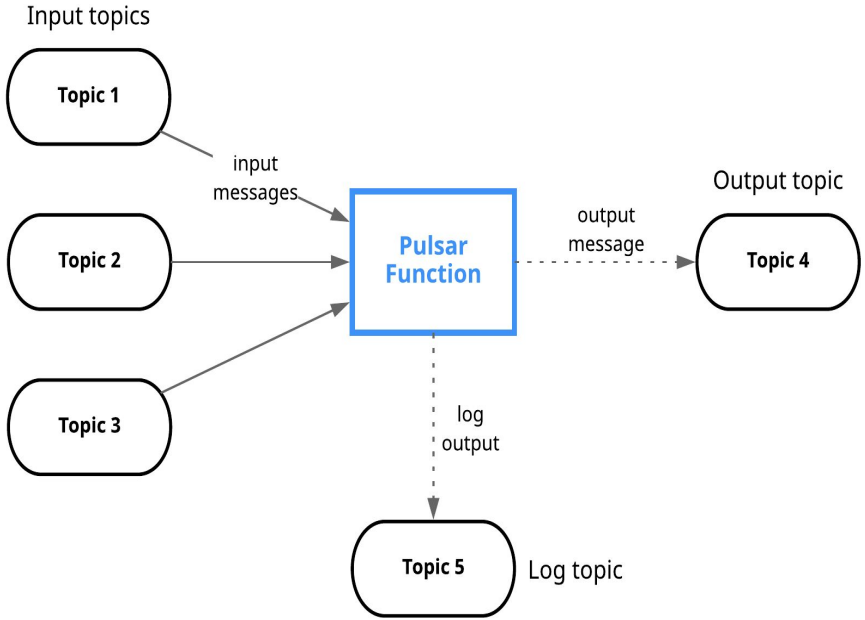




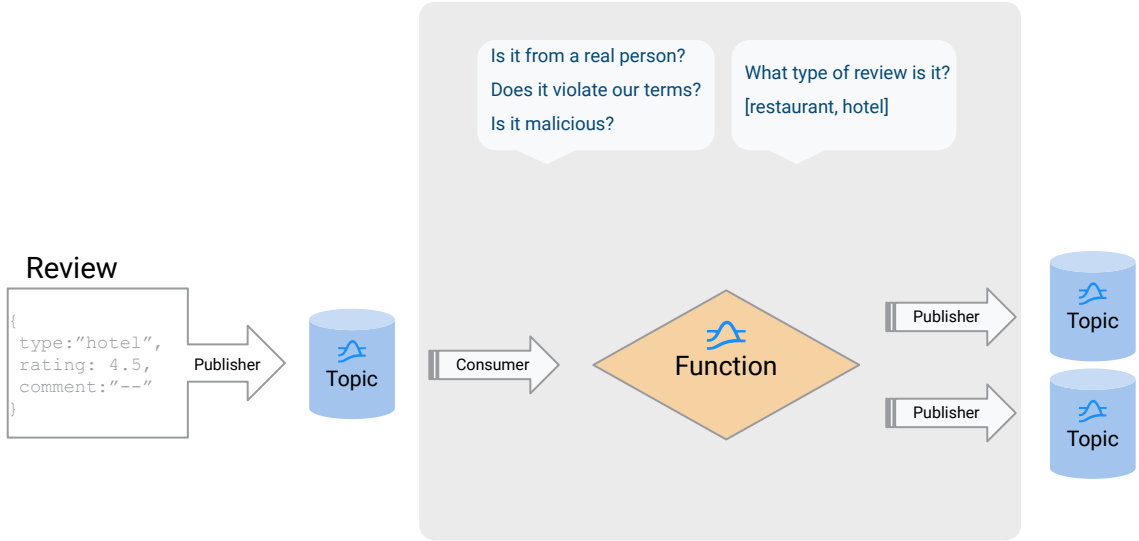
❯ Pulsar Functions

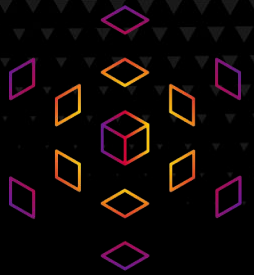


- Allows complex streaming processing
- Light-weight
- Function-as-a-service (AWS Lambda, Google Function, …)
- Main languages:
  - Java
  - Python
  - **Go**



Architecture Overview





#3
Pulsar I/O

Pulsar IO

- **Pulsar I/O**
  - Source Connectors
  - Sink Connectors
- **Built-in Source Connector**
  - RDBMS
  - Kafka (**DataStax Enhanced version**)
  - Kinesis
  - … …
- **Built-in Sink Connector**
  - ElasticSearch
  - Cassandra (**DataStax Enhanced Version**)
  - MongoDB
  - RDBMS
- **CDC Connector**
  - Canal
  - Debezium (MySQL, PostgreSQL, MongoDB)
- **Custom I/O Connector through API**

Architecture Overview