Cassandra Day

2022

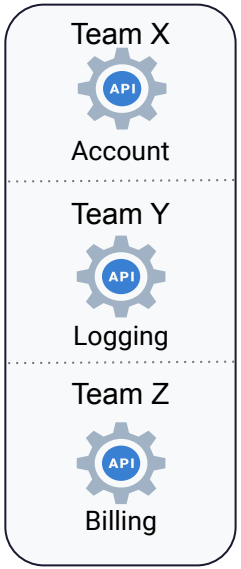
cdc = true

Leveraging Change Data Capture for Apache Cassandra

Sponsored by DataStax

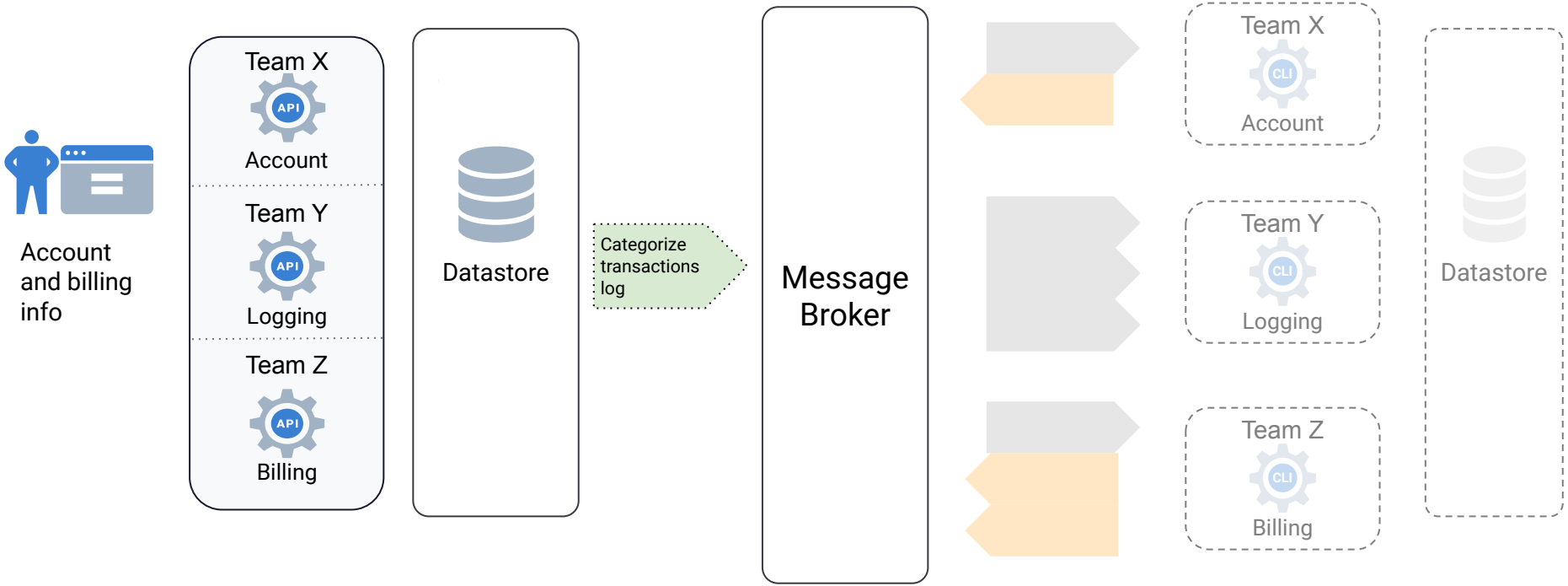
Monolith design

Account and billing info



We need real time updates but there's no time to refactor!

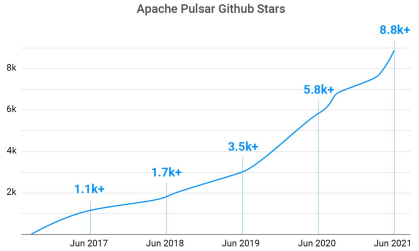
Change data capture (CDC)

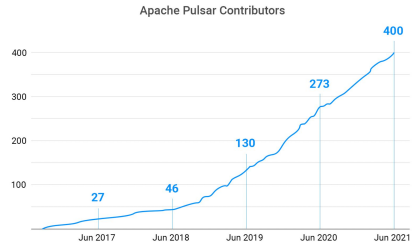


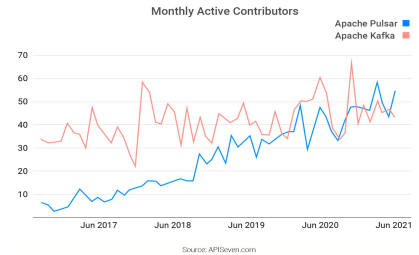
What is Apache Pulsar

- Unified, distributed messaging and streaming platform
- Open source
  - Originally developed at Yahoo!
  - Contributed to the Apache Software Foundation (ASF) in 2016
  - Top 10 Apache project (2021) https://thestack.technology/top-apache-projects-in-2021-from-superset-to-nuttx/
- Cloud Native
  - K8s
  - Multi-cloud and hybrid-cloud

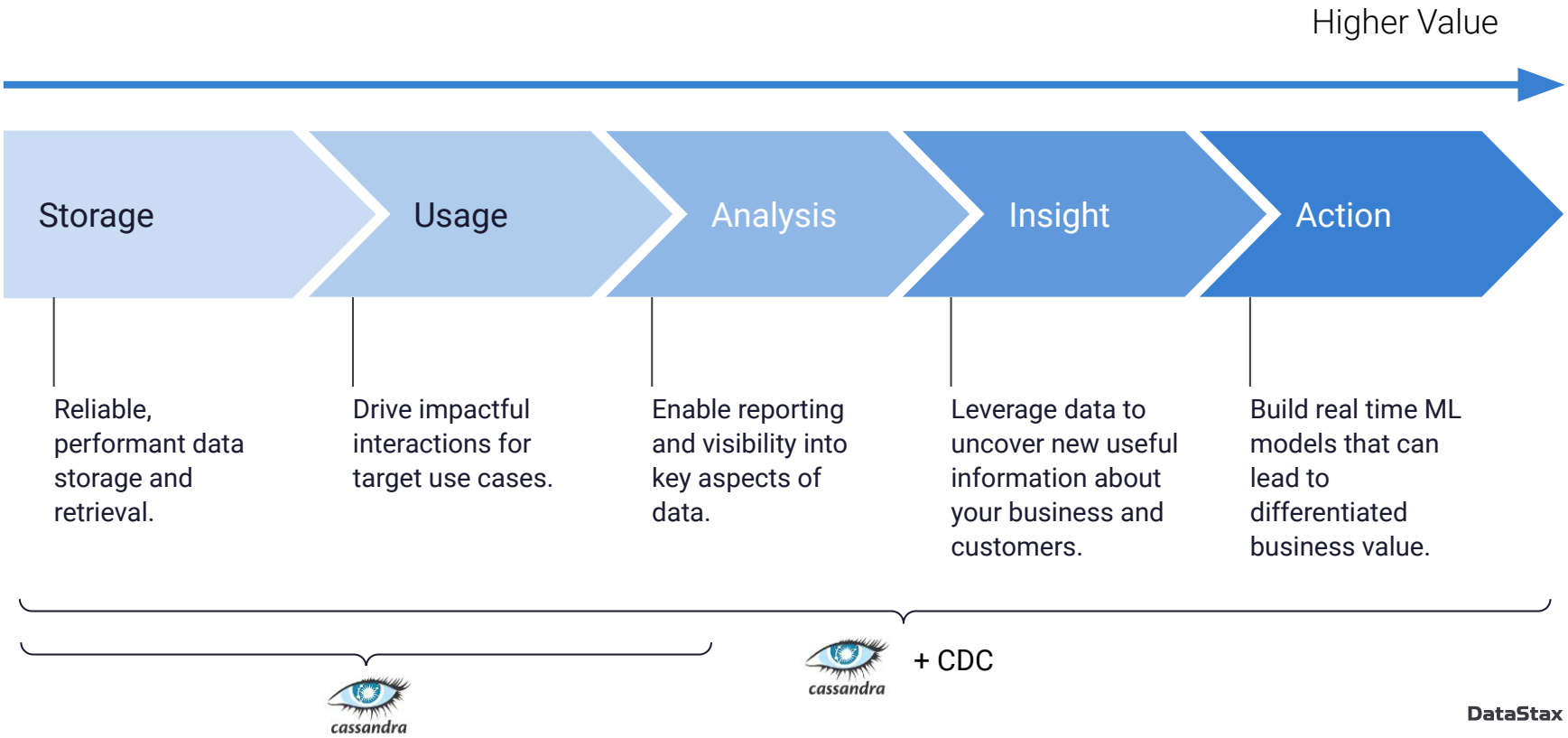
Four Reasons Why Apache Pulsar is Essential to the Modern Data Stack







Deliver more value from your Cassandra investment



Common Use Cases for CDC

- **Data Integration:** Most common use cases, push data into BigQuery, Snowflake or other data platforms.
- **Event Driven Architectures:** Take action by notifying applications to execute business logic when a change is made on the database.
- **Data Science:** Capture time series of database changes to train machine learning models.
- **Full Text Search:** Keep search indexes up to date using systems like Elastic or OpenSearch.
- **Real Time Applications:** Stream changes to applications to update user experiences and drive notifications.

Who can leverage CDC for Cassandra?

Data Engineering & Data Ops

- Provide real time access to data throughout your data ecosystem.

App Developers and Architects

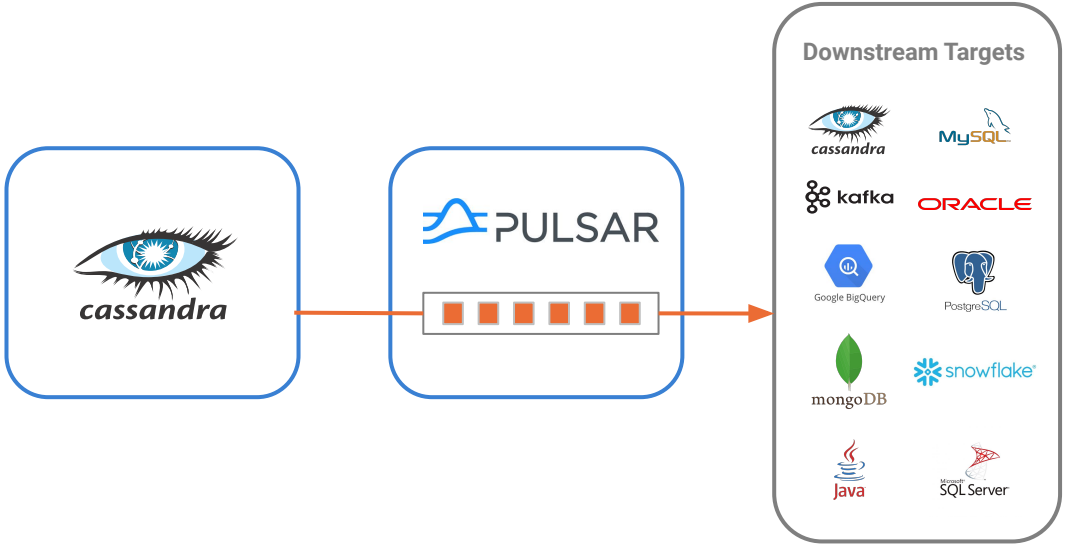
- Make applications more responsive by triggering application logic in response to data change events in Cassandra.

Data Science

- Access deeper insights by extracting a time series of database events and push them to your data lake for advanced processing.

Data Architects

- Improve data quality and standardization by automating transformation and cleansing into your pipelines.



Real Time Data Pipelines with CDC for Cassandra & Pulsar Functions

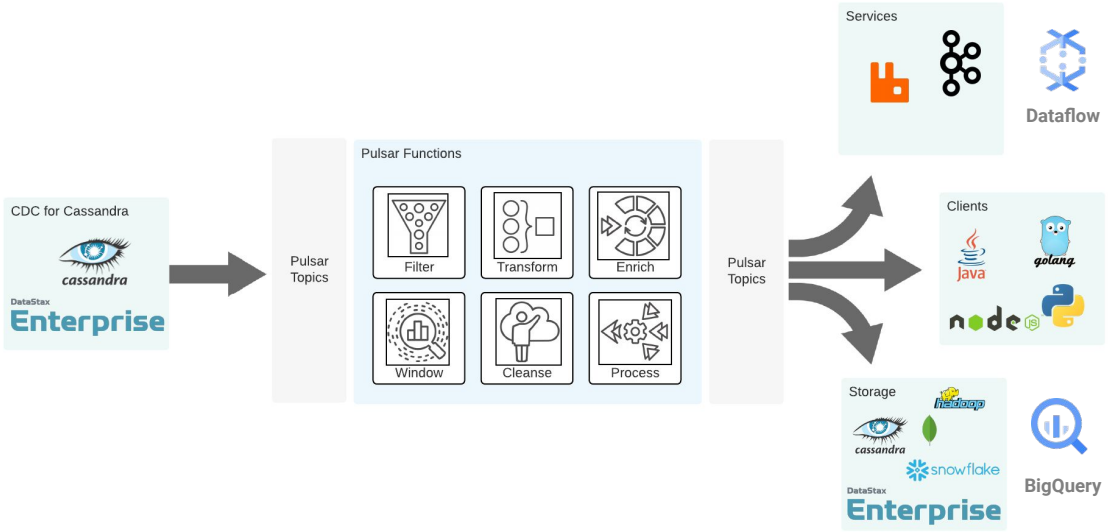
Serverless function platform purpose-built for streaming data pipelines.

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



CDC for Cassandra Architecture

Cassandra Change Agent

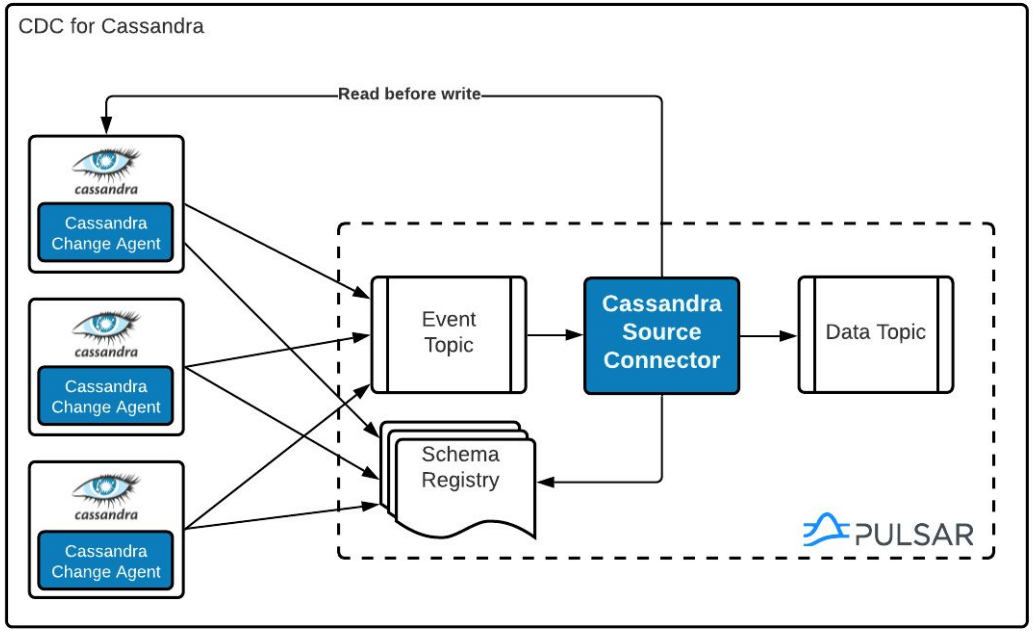
- Local Java agent that runs on each Cassandra node.

Apache Pulsar

- Highly scalable event streaming platform receives incoming messages from the Cassandra Change Agent.

Cassandra Source Connector

- A connector that runs inside Apache Pulsar and provides deduplication logic for the stream of data changes.



Benefits of Pulsar as a Modern Streaming Platform

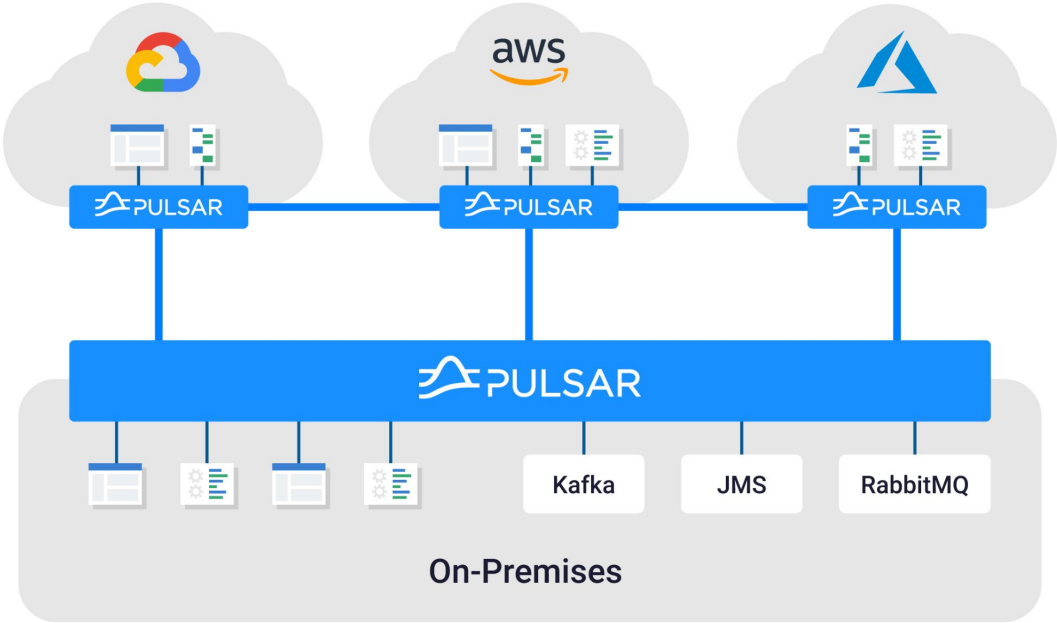
Apache Pulsar represents the **Next Generation of Enterprise Messaging**

Unified Solution for

- Pub/Sub
- Queuing
- Streaming
- Message mediation & enrichment

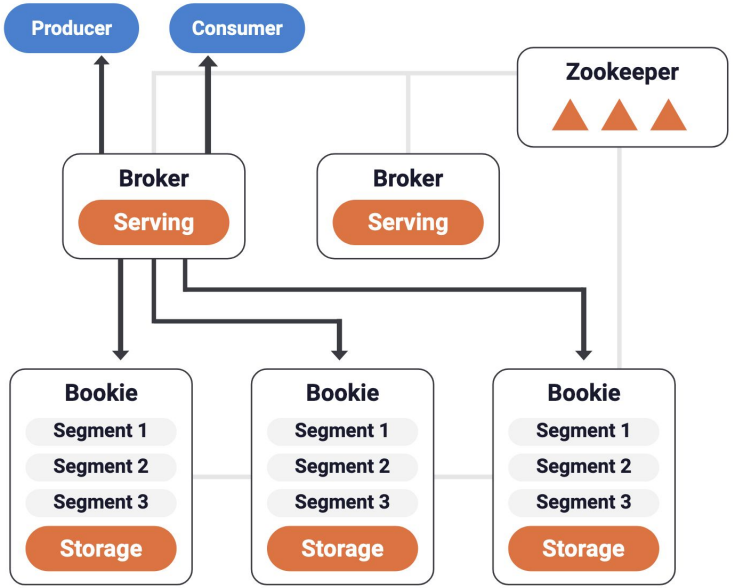
Delivering Important Outcomes

- Modern and secure platform
- Simplified operations
- Clear understanding of cost
- Productive and current tools for developers



Key Differentiators

- Tiered Architecture and Segment-based Storage
  - Fast, Low impact, horizontal scaling
- Native Geo-Replication
  - Simple to enable/modify/disable any time
- Unified and Flexible Message Processing Model
  - Queuing and Pub/Sub
- Robust and Powerful Multi-Tenancy
  - Effective resource segregation mechanisms



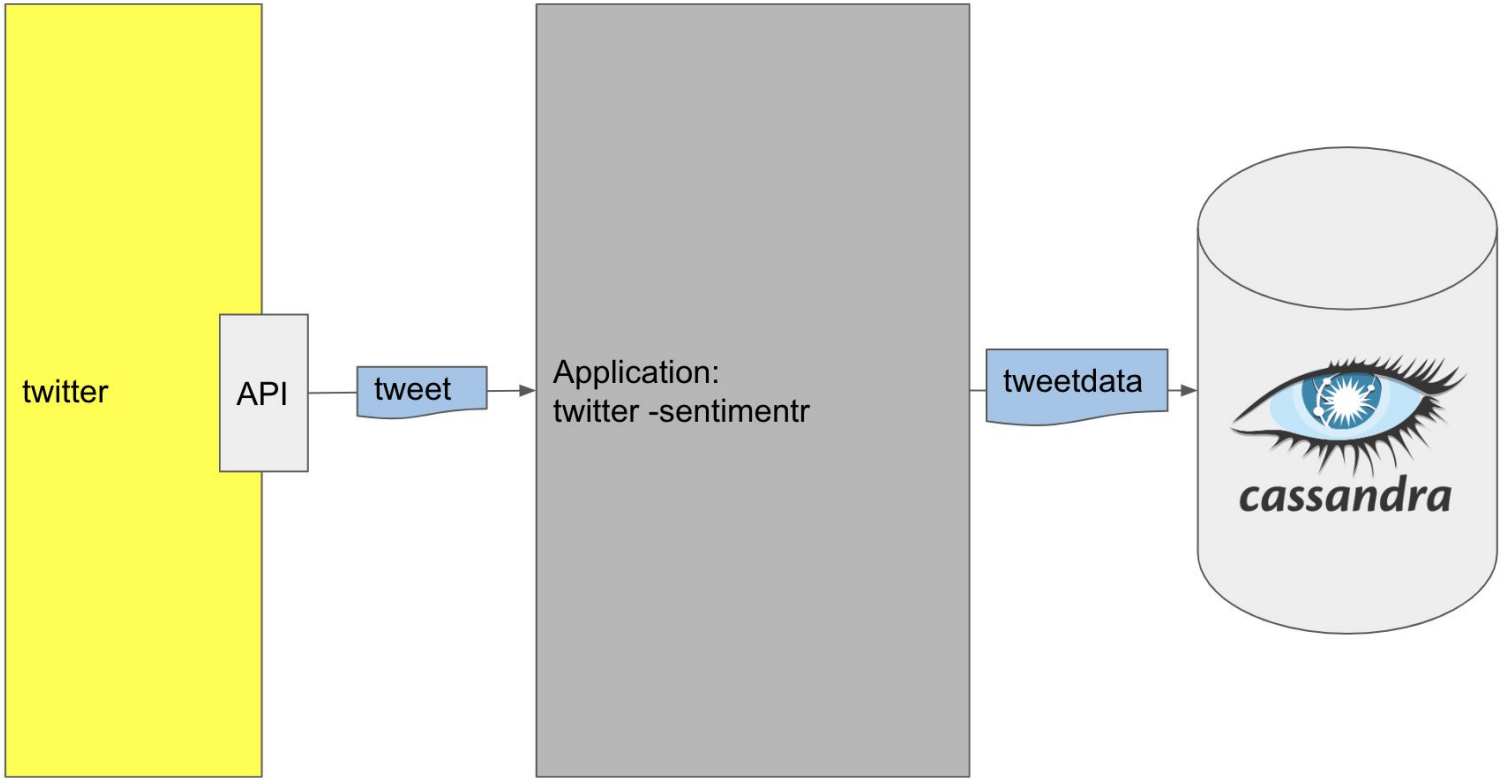
Demo:

Use Case:

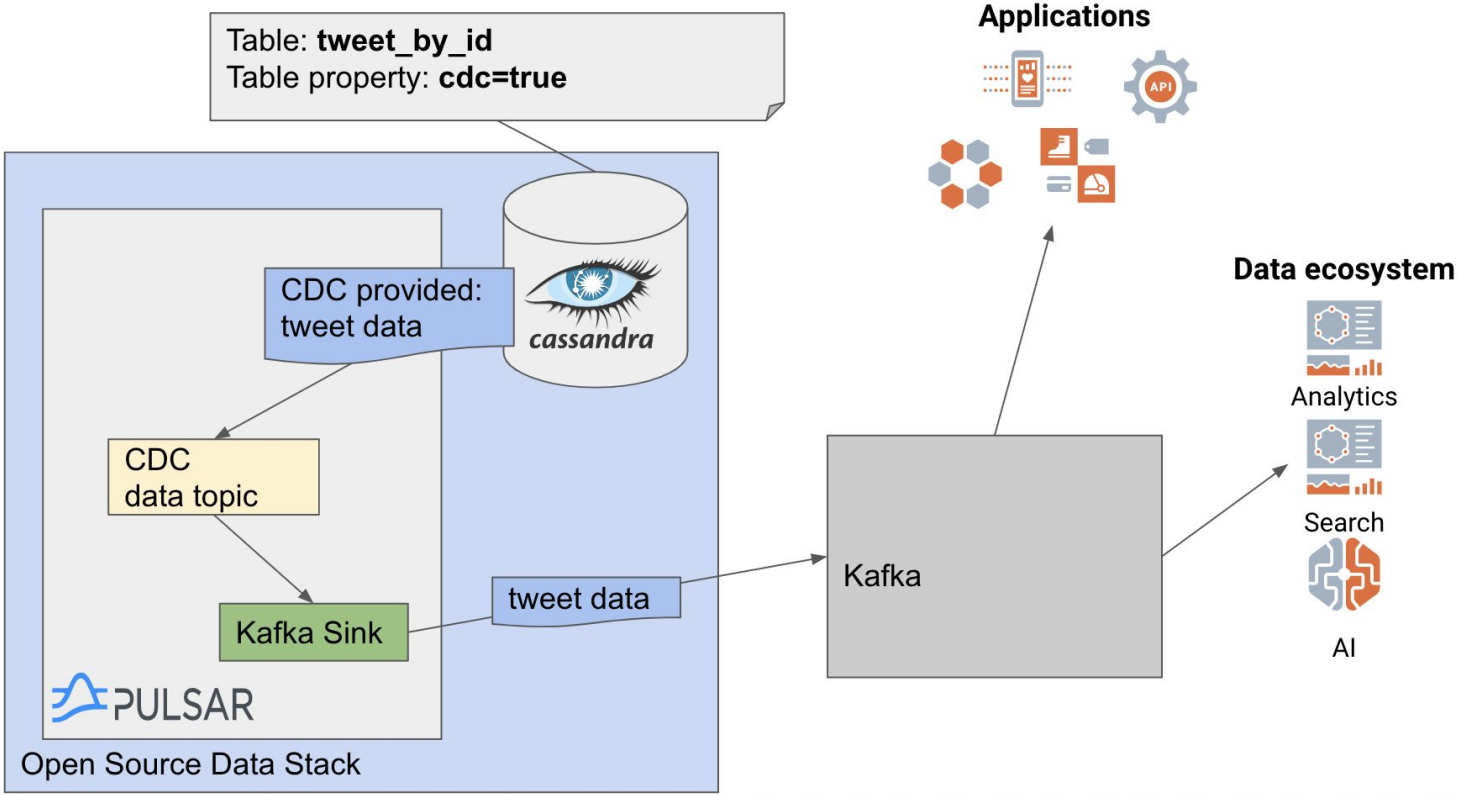
- Real-time Twitter Data
- Search Pattern (ex. brands, products, services,…)
- Insights (ex. customer experience, product and service adoption, …)
- Learn
- Take action (ex. improve products, services continuously
- GitHub Repo: https://github.com/difli/cdc-to-kafka-for-twitter-sentimentr-up

- Microservice Architecture
- Real-time Twitter Data
- Cassandra, CDC, Pulsar, Prometheus, Grafana, Kafka, …
- Pulsar
  - Functions
  - Sources, Sinks
- Change Data Capture (CDC)
  - Stream Realtime Data to Kafka

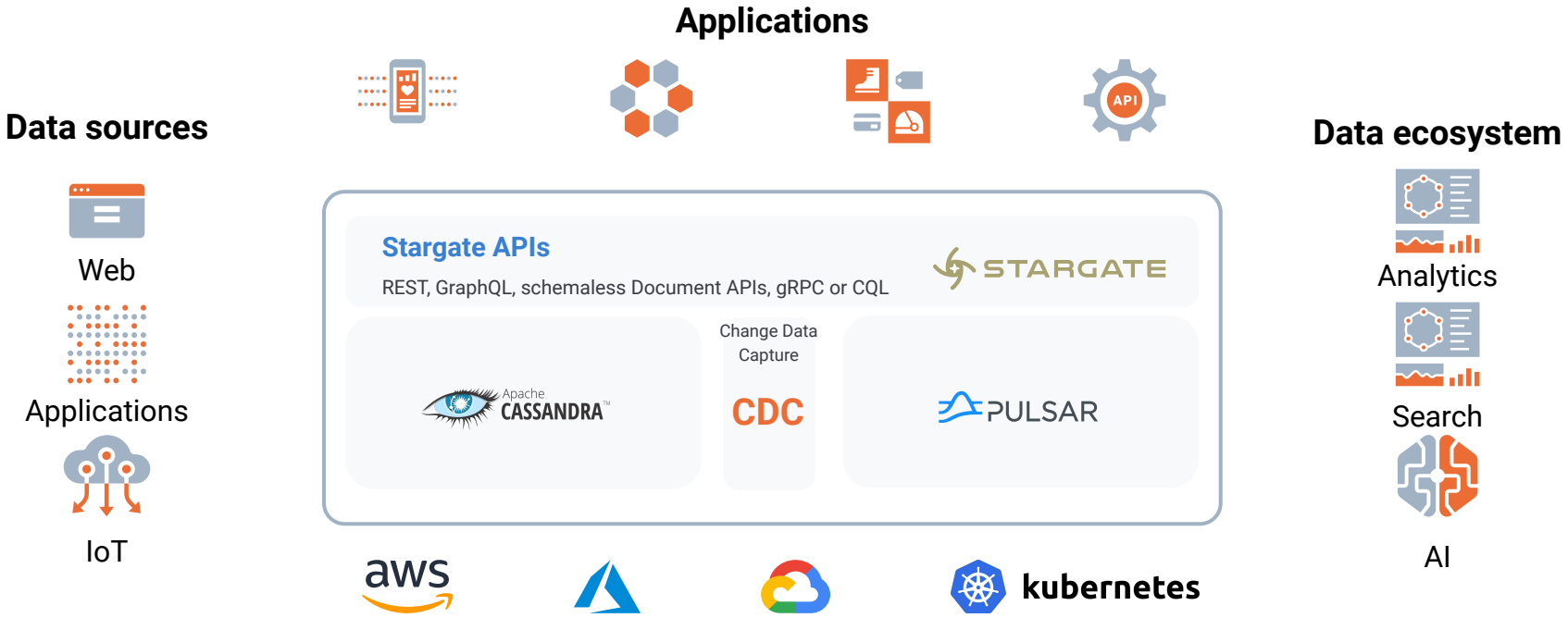
Demo: The Application



Demo: Change Data Capture in Action



Open Source Data Stack



Thank you!

Sponsored by DataStax