Build a Real-time Data Pipeline with Apache Pulsar and Apache Cassandra

Data in motion with Pulsar Functions

Mary Grygleski
Streaming Developer Advocate
@mgrygles

This slide deck can be accessed here:

http://bit.ly/42hpFsf

Mary Grygleski

The Passionate Developer Advocate

@mgrygles

https://www.linkedin.com/in/mary -grygleski/

https://www.twitch.tv/mgrygles

https://discord.gg/RMU4Juw

Who is Mary?

Mary is a Senior Developer Advocate at DataStax, a leading Data Management Company that specializes in Database-as-a-Service, NoSQL, Big Data, Streaming, and the Cloud-Native platform. Previously she was with the Java and WebSphere/Open Source Advocacy team at IBM.

Based out of Chicago, Mary is a Java Champion and President and Executive Board Member of the Chicago Java Users Group (CJUG). She is also co-organizers for the Data, Cloud and AI In Chicago, Chicago Cloud, and IBM Cloud Chicago meetup groups.

She has extensive experience in product and application design, development, integration, and deployment experience, and specializes in Event-driven, Reactive Java, Open Source, and cloud-enabled distributed systems.

Senior Developer Advocate

Passionate Advocate

Java Champion

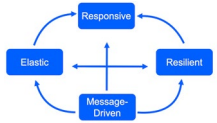
PULSAR

JAKARTA EE

Open Liberty

RED HAT OPENSHIFT

MICROPROFILE

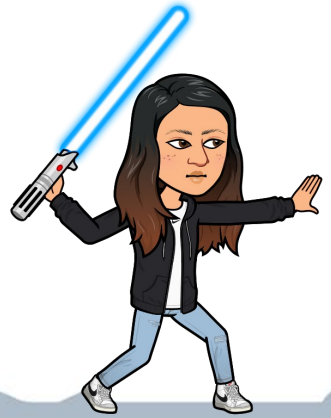


mgrygles

mary-grygleski

mgrygles

mgrygles



- Streaming
- Distributed Systems
- Reactive Systems
- IoT/MQTT

Mary Grygleski

01 PULSAR

Introduction to Apache Pulsar

02

Data Science in Event Streaming

03

Pulsar Functions

04

Introduction to Apache Cassandra

05

Pulsar I/O

06

What's next?
Resources and Links

Agenda

01 PULSAR

Introduction to Apache Pulsar

02

Data Science in Event Streaming

03

Pulsar Functions

04 cassandra

Introduction to Apache Cassandra

05

Pulsar I/O

06

What's next?

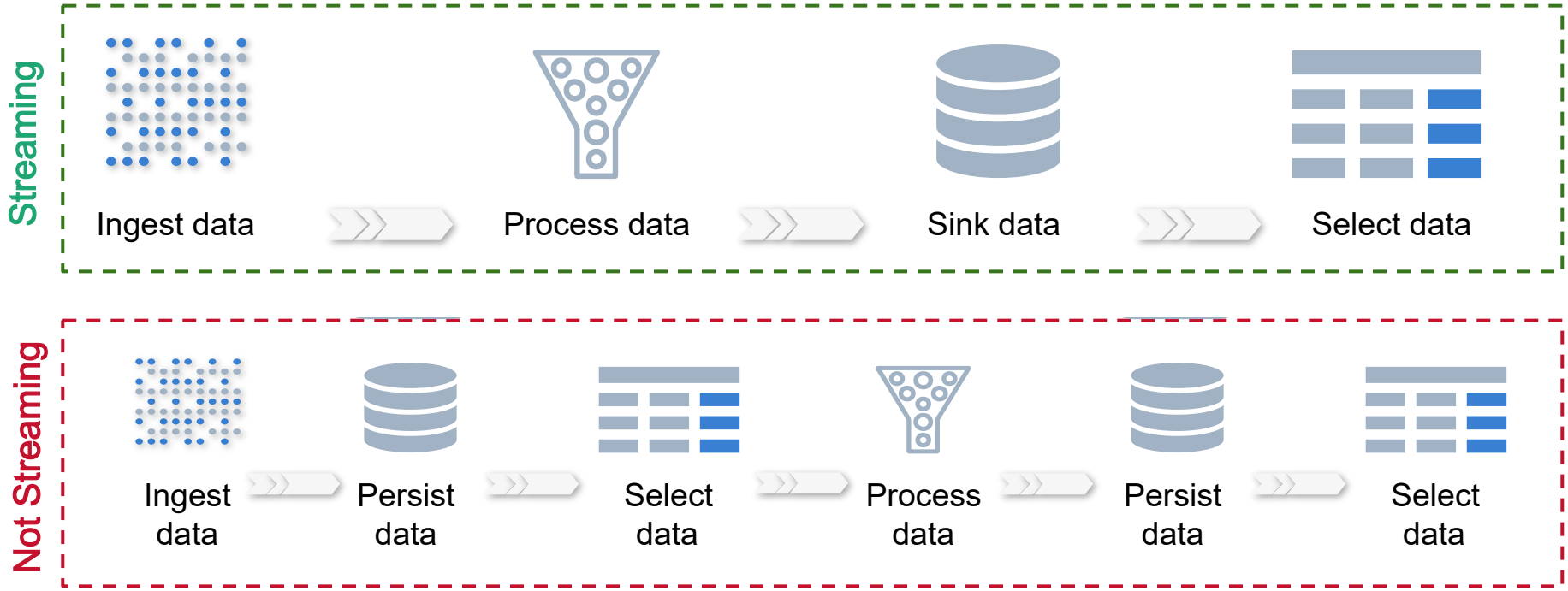
Quiz, Homework, Next week

Agenda

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

Event Streaming == Message Streaming



Streaming vs not streaming

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

Apache Pulsar

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 Delivers in Ways Other Streaming Platforms Can't

Producer

Client application sending messages to topic managed by Broker

Consumer

Client application reading messages from a topic managed by Broker

Broker

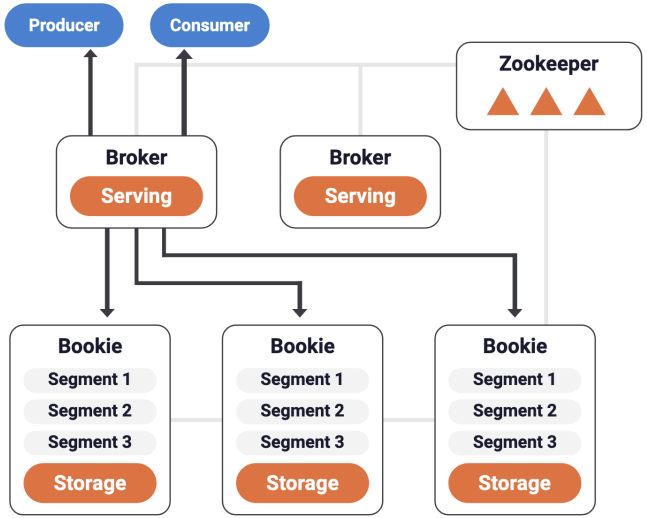
A stateless process that handles incoming message, message dispatching, communicates with the Pulsar configuration store, and stores messages in BookKeeper instances

BookKeeper

Persistent message store

ZooKeeper

Holds cluster metadata, handles coordination tasks between Pulsar clusters



Pulsar Components

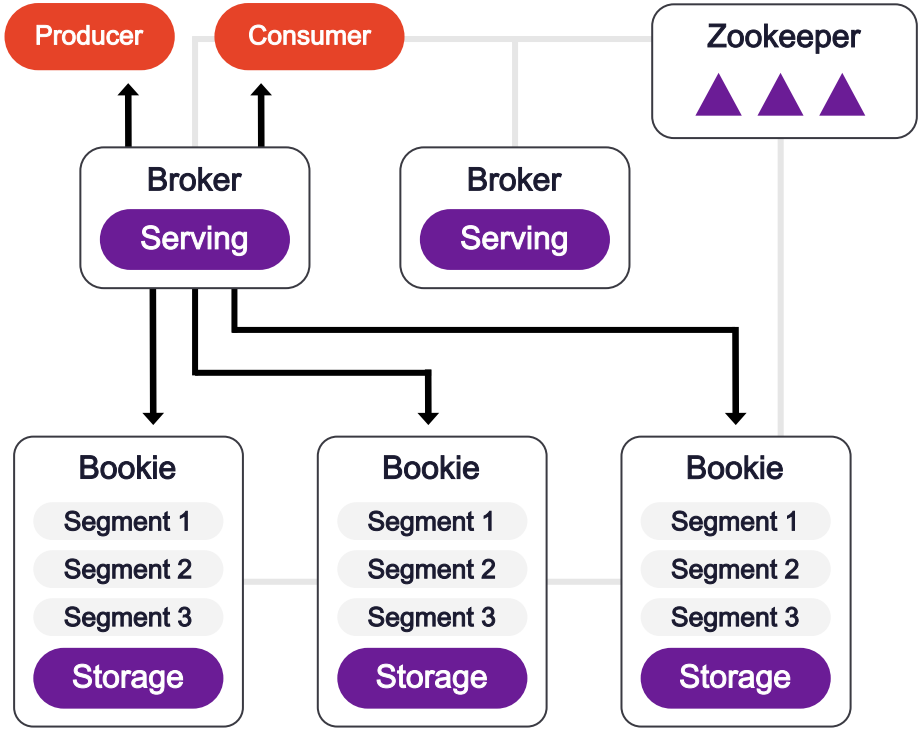
Distributed, tiered architecture

Separated compute from storage

Zookeeper holds metadata for the cluster

Stateless Broker handles producers and consumers

Storage is handled by Apache Bookkeeper



Cloud-Native Architecture

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

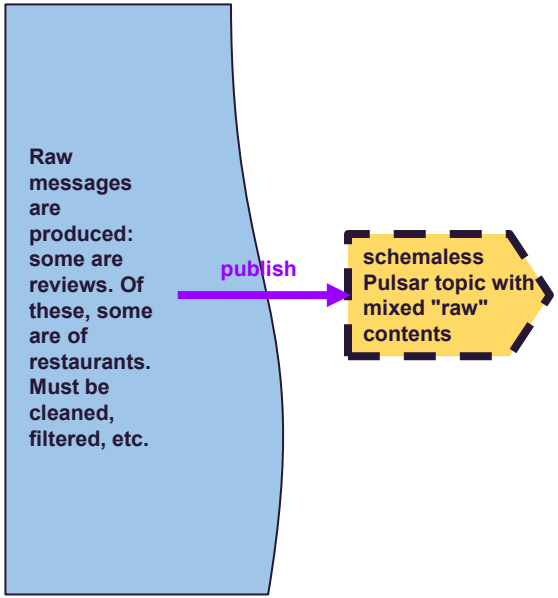
Free monthly credits to help you get started quickly

DataStax Astra: Streaming Made Easy in the Cloud

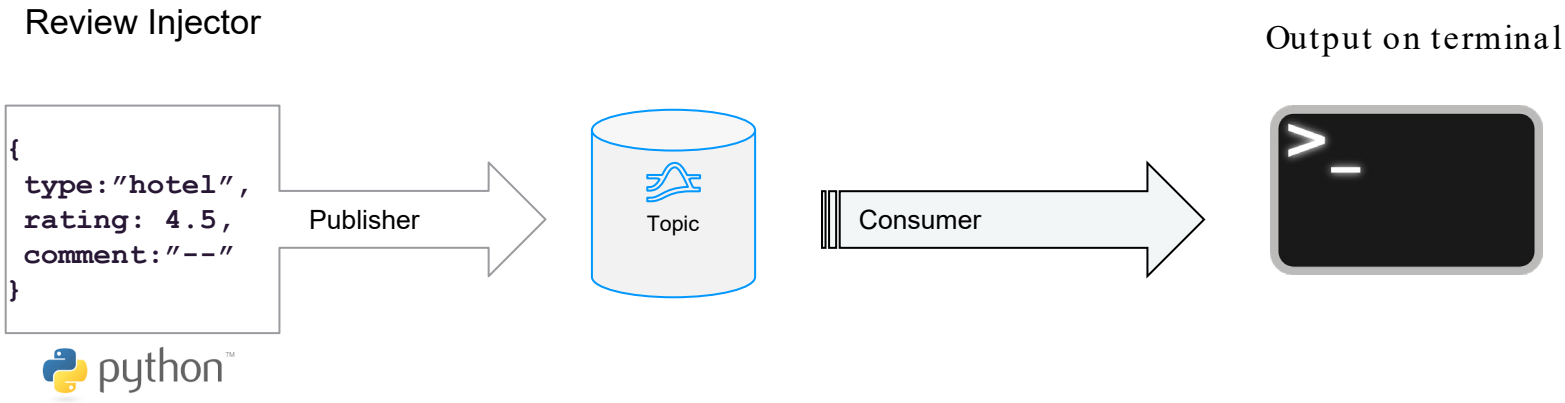
Lab 1

Producer & Consumer

https://github.com/mgrygles -lab/



Business Architecture



Logical Architecture

01 PULSAR

Introduction to Apache Pulsar

02

Data Science in Event Streaming

03

Pulsar Functions

04

Introduction to Apache Cassandra

05

Pulsar I/O

06

What's next?

Quiz, Homework, Next week

Agenda

01

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.

02

Secure Social Media, Protect Customer Privacy

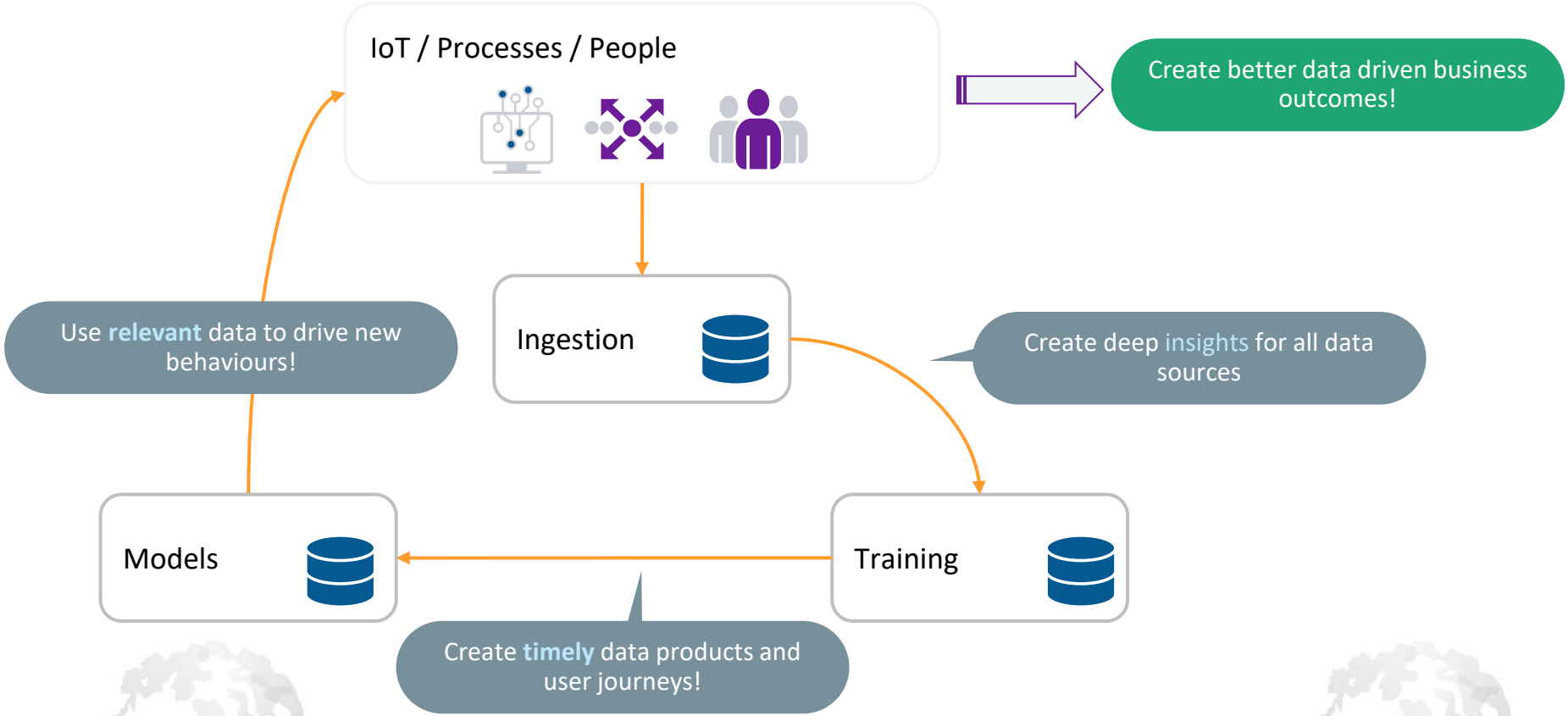
Identify out-of-the-ordinary patterns to prevent malicious attacks on digital and physical assets from unauthorized applications and individuals.

03

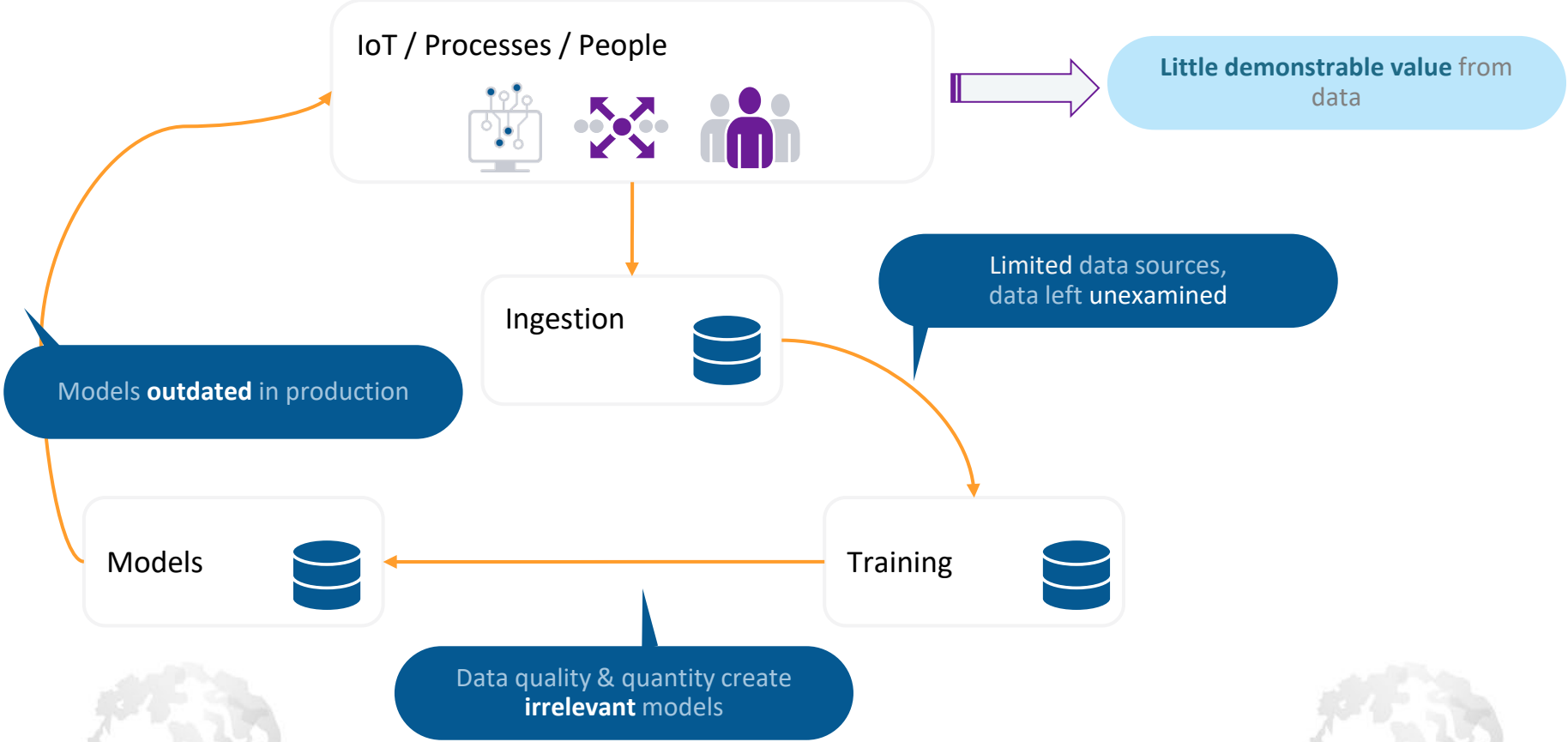
IoT Data Ingestion and Classification

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

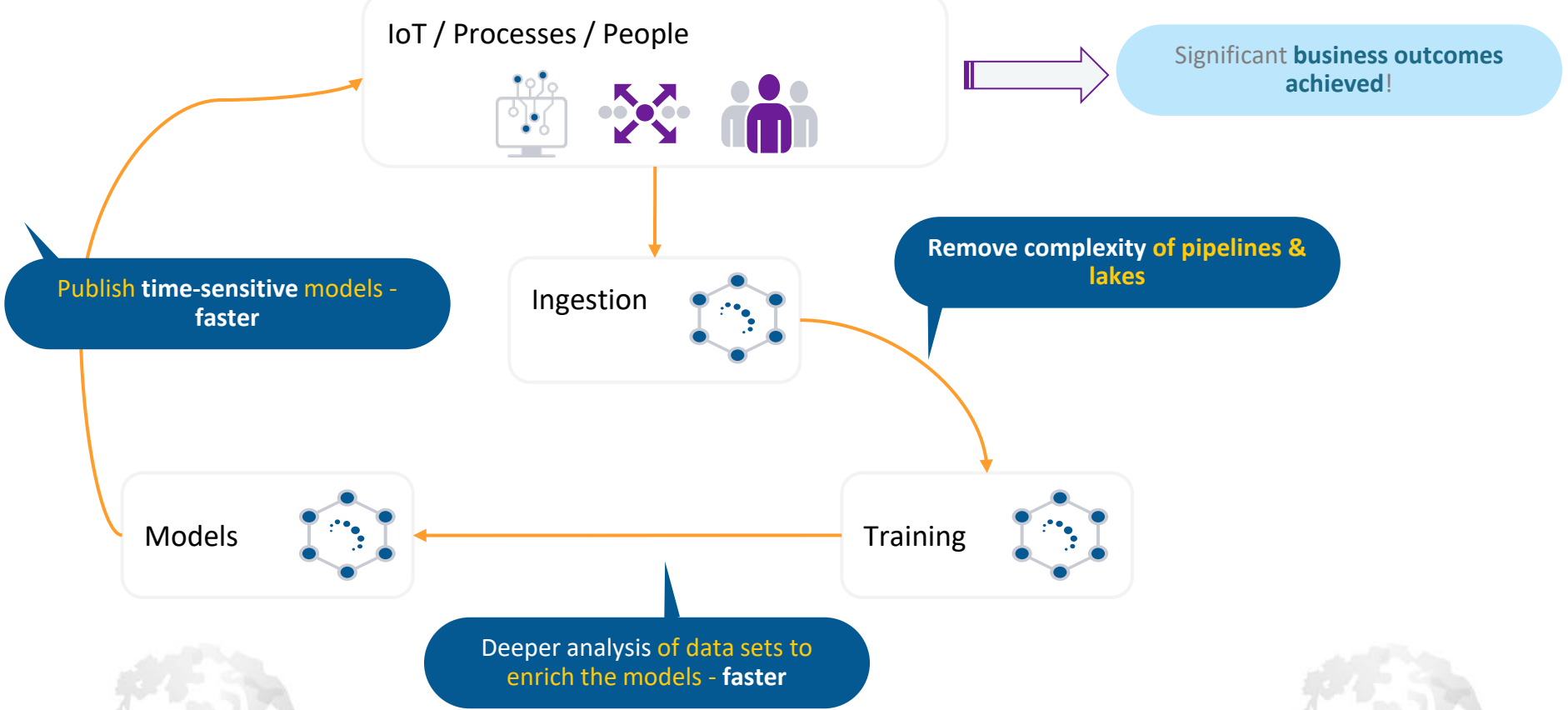
DataScience with events



Expected Real-Time Data Pipeline



What's not working ?



Cassandra and Pulsar to the rescue

Our Data pipeline today

01 PULSAR

Introduction to Apache Pulsar

02

Data Science in Event Streaming

03

Pulsar Functions

04

Introduction to Apache Cassandra

05

Pulsar I/O

06

What's next?

Quiz, Homework, Next week

Agenda

Serverless function platform purpose-built for streaming data pipelines.

Simple Function Architecture

Triggered from input topic

Simple programmatic interface

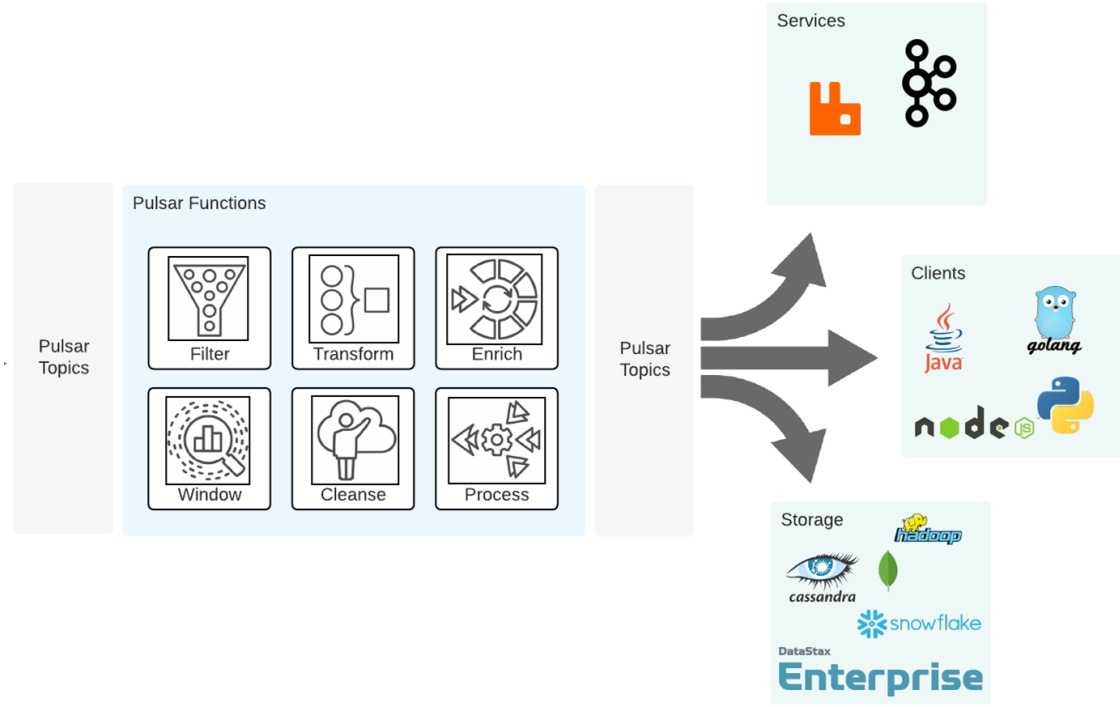
Push function result to output topic

Built for DevOps

Standard Kubernetes based runtime

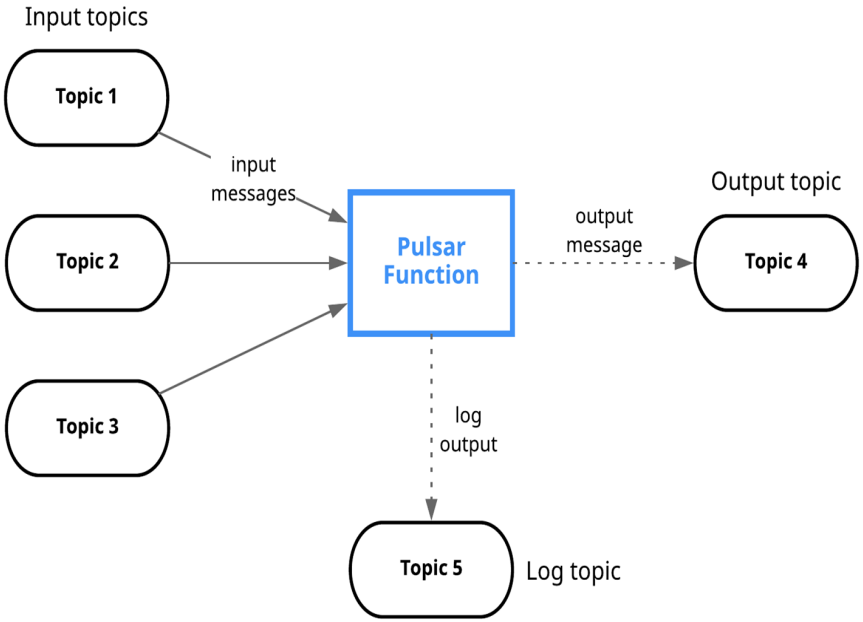
Automated deployments

CI/CD friendly

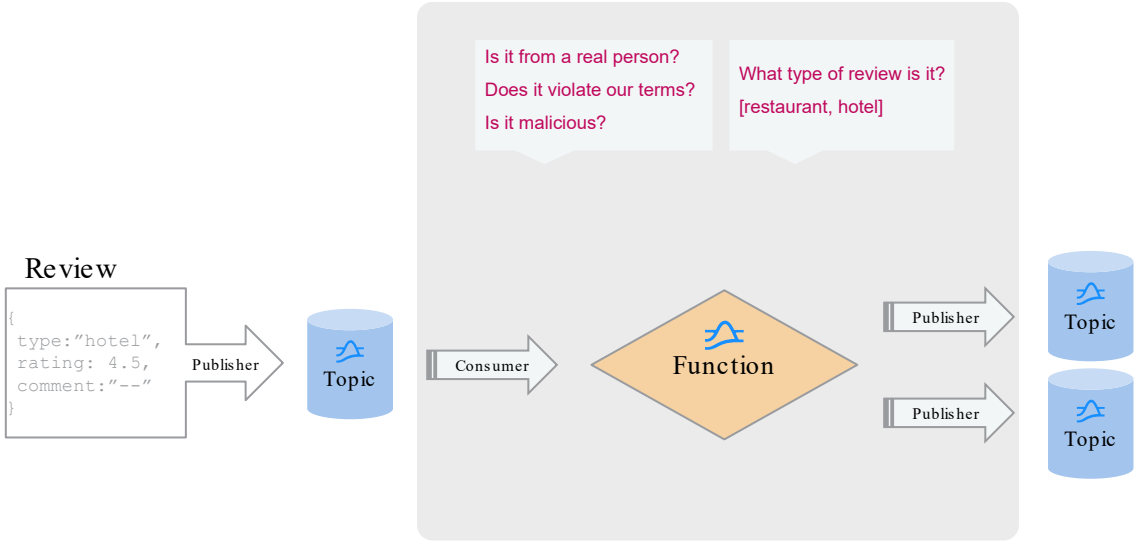


Pulsar Functions

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



Pulsar Functions



Architecture Overview

Lab 2
Pulsar Functions

https://github.com/datastaxdevs/workshop-pulsarfunctions-data-in-motion

01 PULSAR

Introduction to Apache Pulsar-

02 

Data Science in Event Streaming

03 

Pulsar Functions

04 cassandra

Introduction to Apache Cassandra-

05 

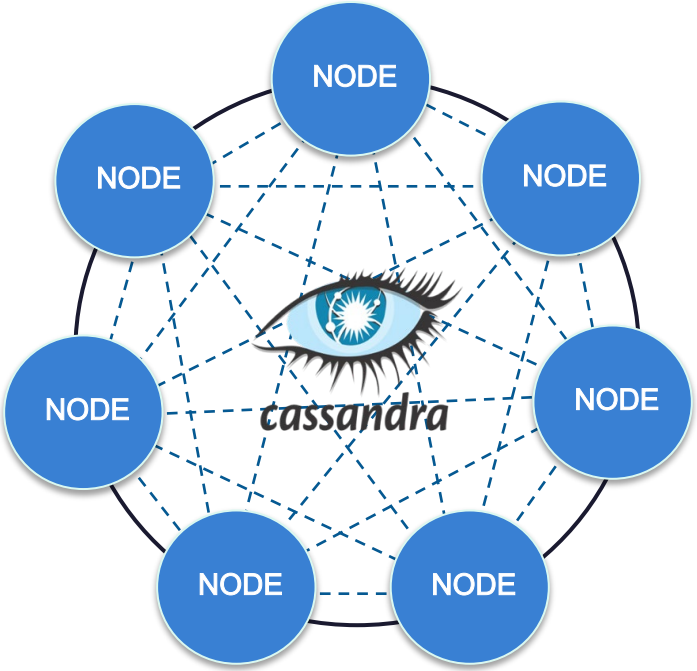
Pulsar I/O

06 

What's next?

Quiz, Homework, Next week

Agenda

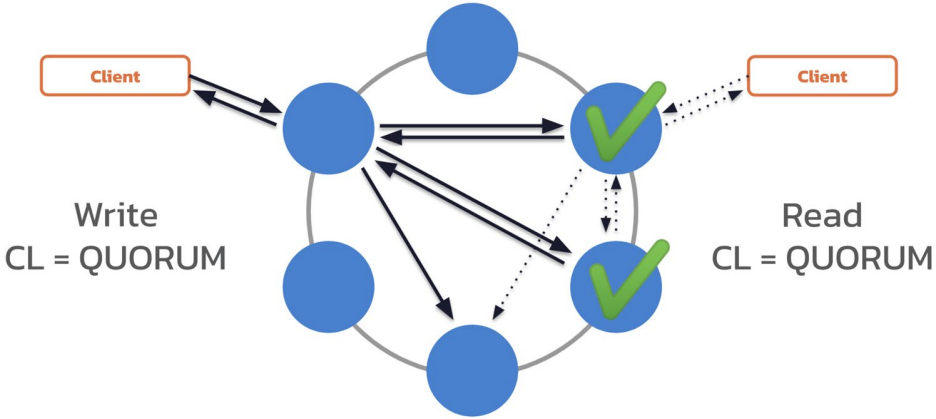


- Big Data Ready
- Read / Write Performance
- Linear Scalability
- Highest Availability
- Self-Healing and Automation
- Geographical Distribution
- Platform Agnostic
- Vendor Independent

Apache Cassandra's Awesomeness

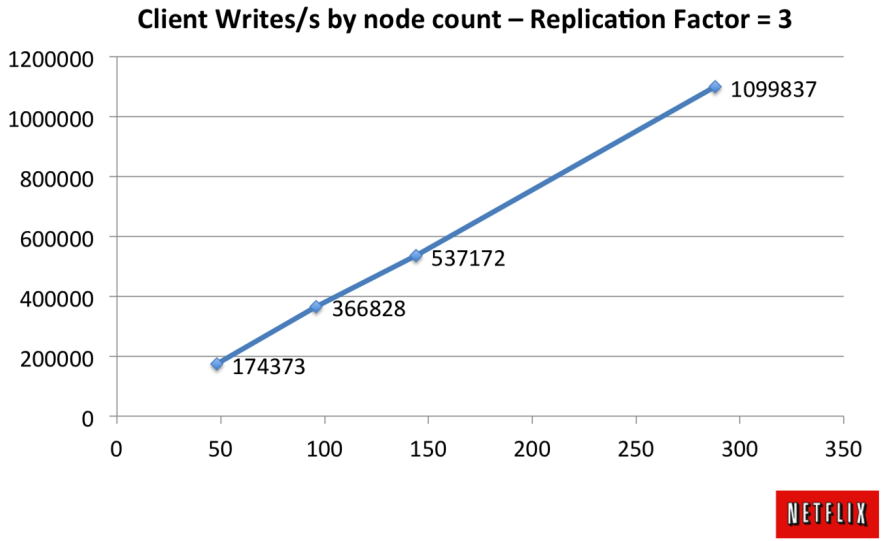
Even a single Cassandra node is very performant but a cluster consisting of multiple nodes and data centers brings throughput to the next level.

Decentralisation (masterless architecture) means that every node is able to deal with any request, read or write.



Read / Write Performance

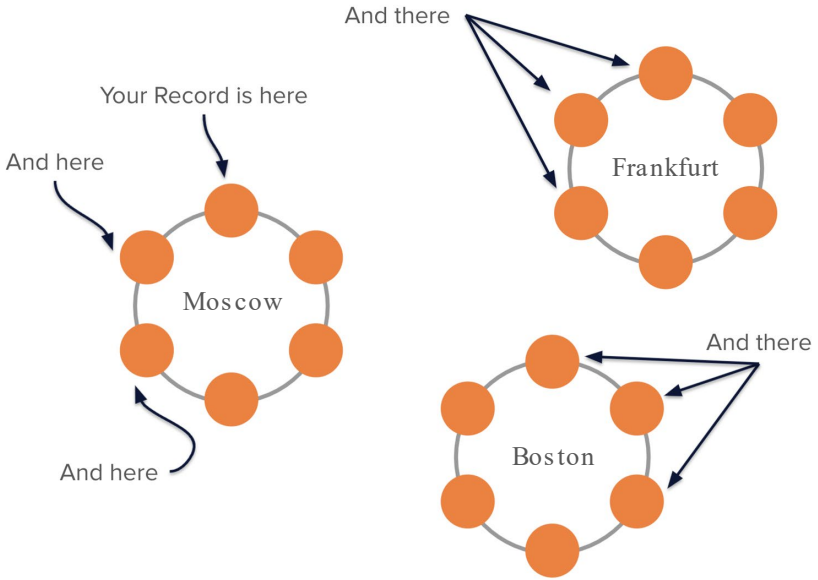
- For volume or velocity, there are no limitations
- Linear - No overhead on new nodes, scales with your needs*



Linear Scalability

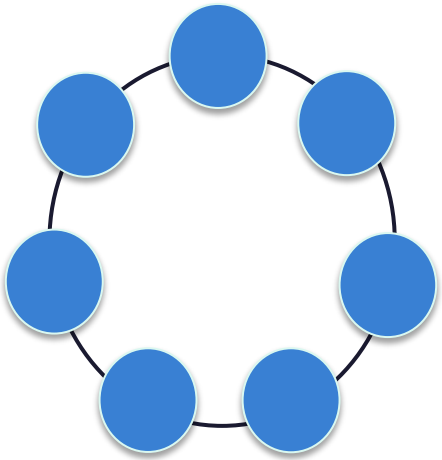
Replication, Decentralisation, and Topology-Aware Placement Strategy take care of possible downtimes:

- Multiple Live Replicas
- No Single Point of Failure
- Network topology-aware data placement
- Client-side Smart Reconnection and Strong Retry Mechanism



Highest Availability

Partitioning over distributed architecture makes the database capable to handle data of any size: we mean petabytes scale. Need more volume? Add more nodes.



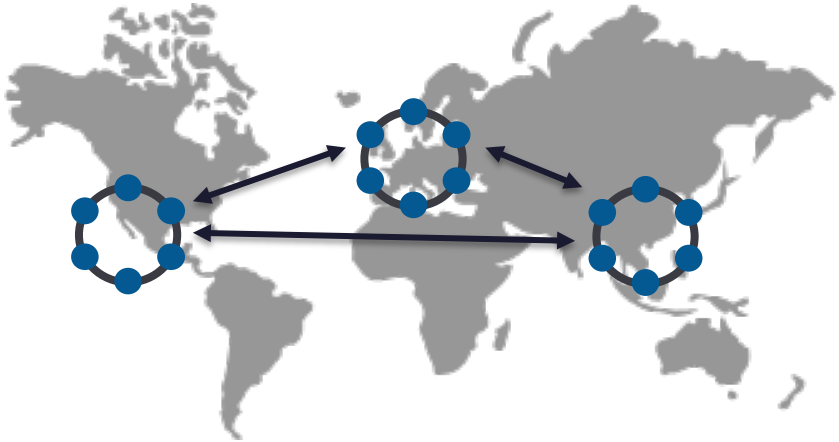
Big Data Ready

Operations for a huge cluster can be exhausting so Apache Cassandra clusters are smart and able to scale, change data placement and recover automagically.

Self-Healing and Automation

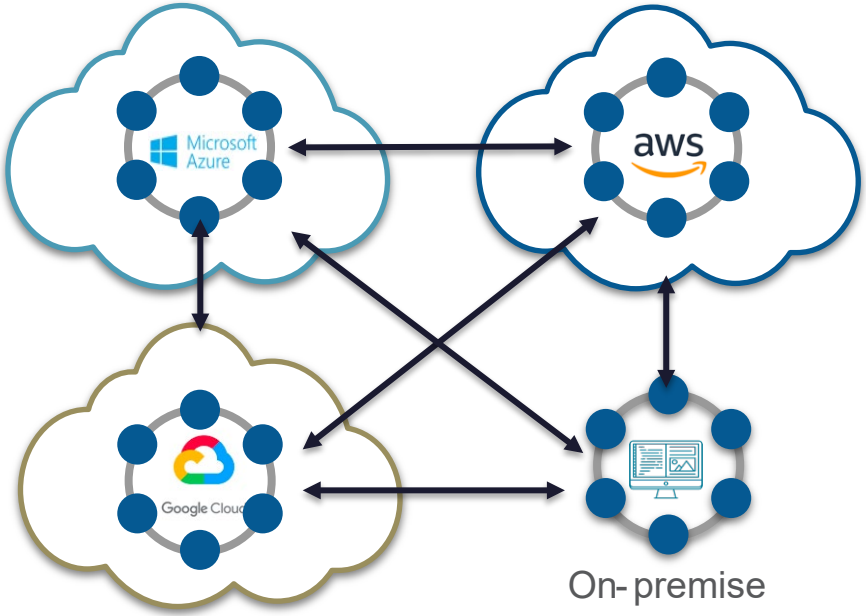
Cassandra's trademark is multi-datacenter deployments, granting you an exceptional capability for disaster tolerance while keeping your data close to your clients - worldwide.

All DCs are active (available for both writes and reads)!



Geographical Distribution

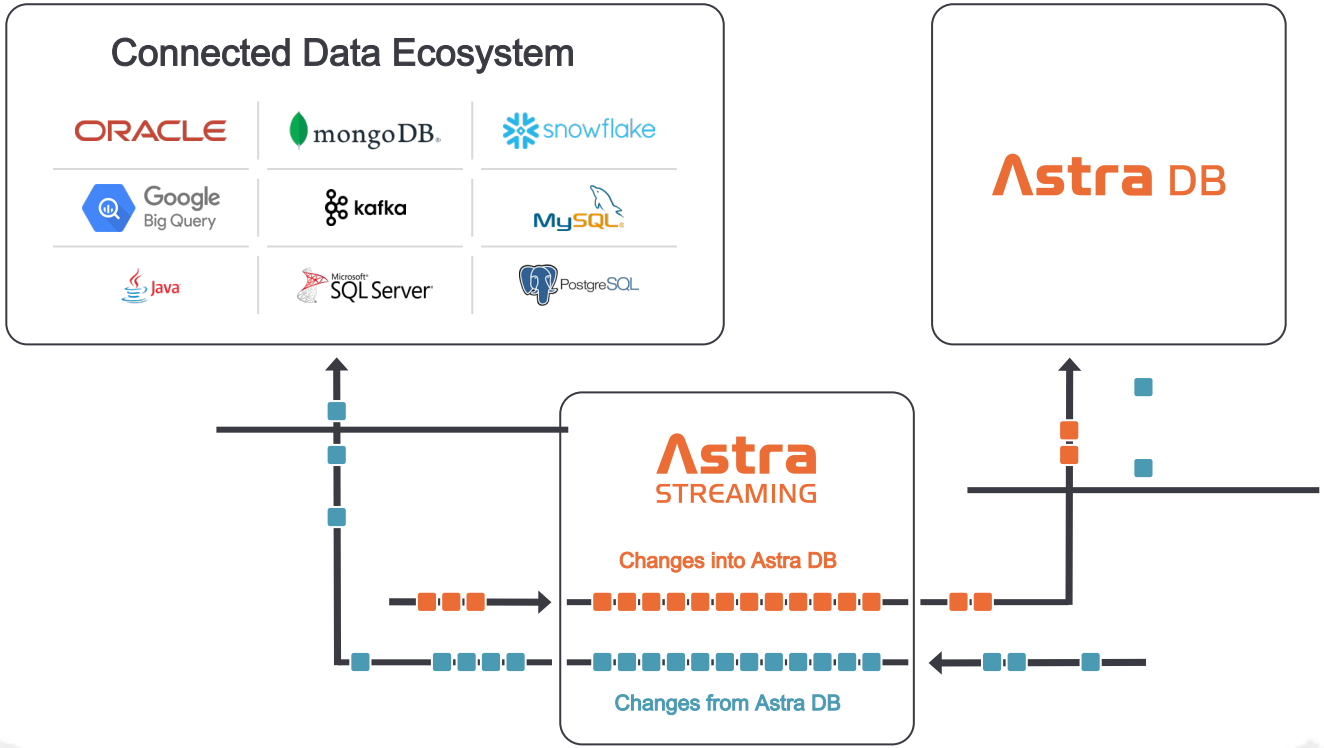
Apache Cassandra is **not bound to any platform** or service provider, helping you build hybrid-cloud and multi-cloud solutions with ease.



Platform Agnostic

Cassandra doesn't belong to any of commercial vendors but controlled by a non-profit Open Source **Apache Software Foundation**, already familiar to you by *Hadoop*, *Spark*, *Kafka*, *Zookeeper*, *Maven* and many other projects.

Vendor Independent



Build Real Time Data Pipelines with Astra

01 PULSAR

Introduction to Apache Pulsar

02

Data Science in Event Streaming

03

Pulsar Functions

04

Introduction to Apache Cassandra

05

Pulsar I/O

06

What's next?

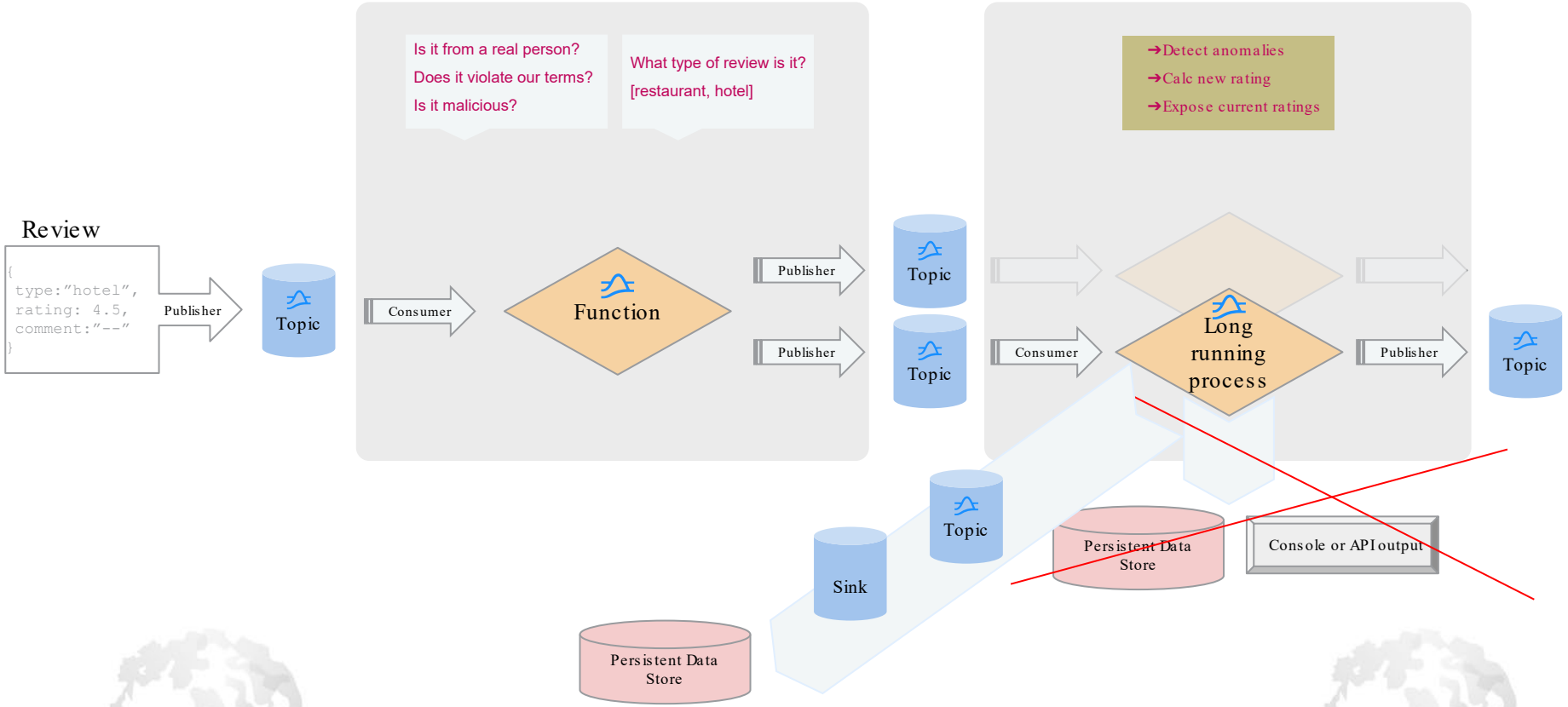
Quiz, Homework, Next week

Agenda

- 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

Pulsar I/O Connectors



Architecture Overview

Lab 3

Pulsar I/O

https://github.com/datastaxdevs/workshop-pulsarfunctions-data-in-motion

01 PULSAR

Introduction to Apache Pulsar

02

Data Science in Event Streaming

03

Pulsar Functions

04 cassandra

Introduction to Apache Cassandra

05

Pulsar I/O

06

What's next?

Resources and Links

Agenda

Resources / Links:

PULSAR https://pulsar.apache.org/

https://bookkeeper.apache.org/

https://zookeeper.apache.org

DATASTAX ASTRA DB https://astra.datastax.com

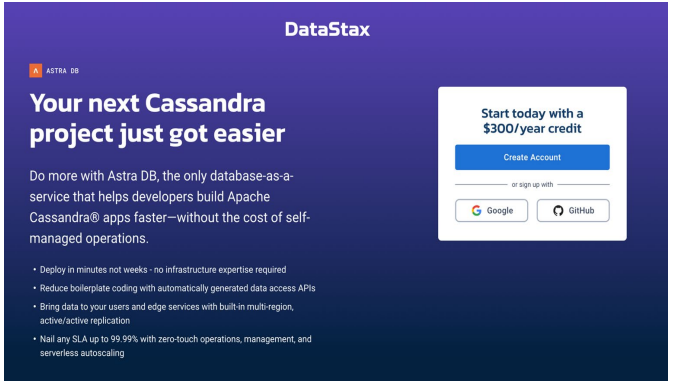
DATASTAX ASTRA STREAMING https://www.datastax.com/products/astra -streaming

DATASTAX LUNA STREAMING https://www.datastax.com/products/luna -streaming

DATASTAX ASTRA

CDC for Astra: https://docs.datastax.com/en/astra/docs/astream-cdc.html

DATASTAX ASTRA



1.- Create an Astra account at

https://www.datastax.com/lp/next -cassandra-project

2.- Add a payment method, enter OpenSource200 for an additional $200 in credits

Check out 5 Minutes About Pulsar on YouTube

https://bit.ly/3bgkRxJ

Follow Mary's Twitch Stream

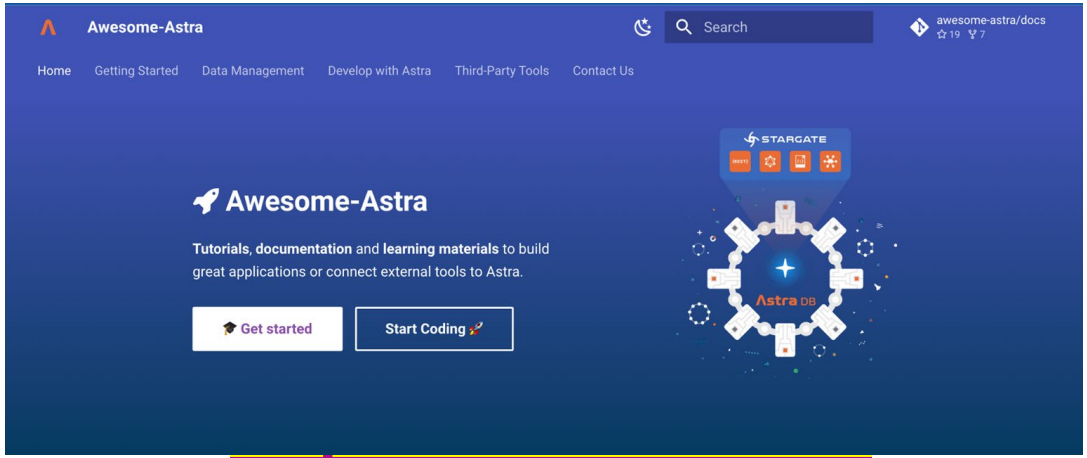
(Different topics: Java, Open Source, Distributed Messaging, Event-Streaming, Cloud, DevOps, etc)

Wednesday at 2pm-US/CST

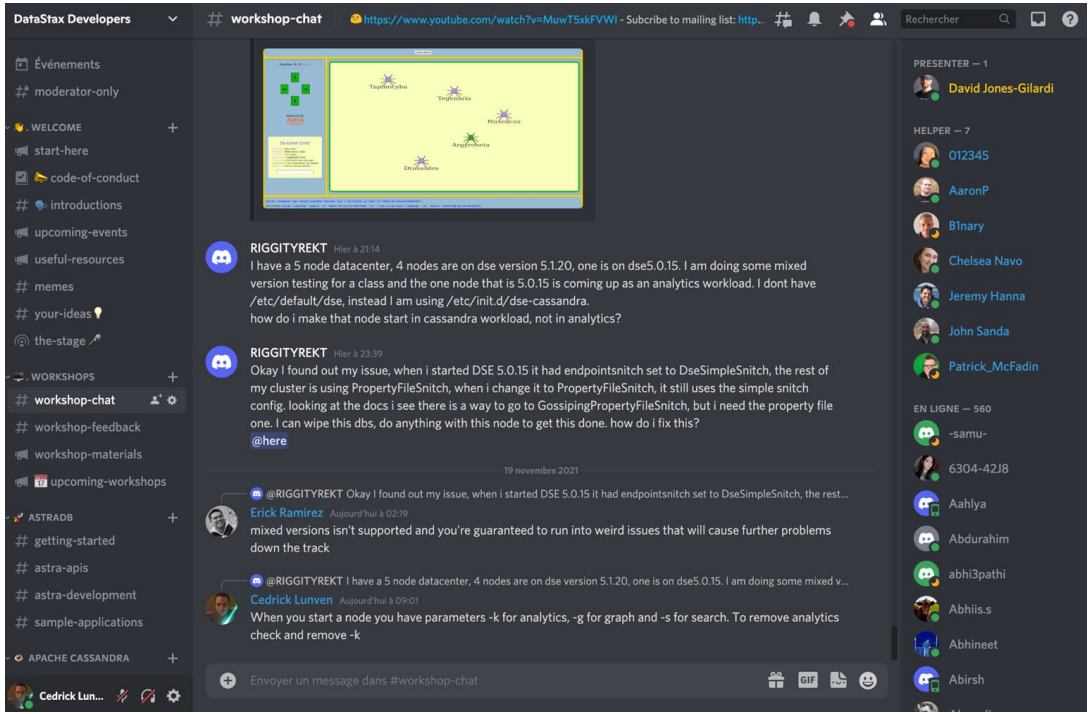
https://twitch.tv/mgrygles

How to start coding all of this?

Check out Awesome-Astra



astra.github.io/docs/



!discord

dtsx.io/discord

Datastax Developers Discord (18k+)

...n to NoSQL
...ses
...k views •
...Streamed 1 week ago

Crash Course | Introduction to Cassandra for Developers

331 views • 1 week ago

Introduction to NoSQL Databases

3.5K views •
Streamed 1 week ago

Introduction to NoSQL Databases

10K views •
Streamed 1 week ago

Build your own NETFLIX clone!

4K views •
Streamed 2 weeks ago

Build your own NETFL...
clone!

7.4K views •
Streamed 2 weeks ago

Astra Streaming Demo

177 views • 2 weeks ago

Kubernetes Ingress Management with Traefik...

496 views •
Streamed 2 weeks ago

Build your own TikTok clone!

1.9K views •
Streamed 3 weeks ago

Build your own TikTok Clone!

4K views •
Streamed 3 weeks ago

How to use the Connect Driver in Astra DB

113 views • 4 weeks ago

How to use the CQL Console in Astra DB

39 views • 4 weeks ago

How to create an Authentication Token in...

37 views • 4 weeks ago

How to use the Data Loader in Astra DB

62 views • 4 weeks ago

Astra DB Sample App Gallery

36 views • 4 weeks ago

How to use Secure Connect in Astra DB

42 views • 4 weeks ago

Cassandra Day India: CL Room (Workshops)

2.4K views •
Streamed 4 weeks ago

Cassandra Day India: RF Room (Talks)

1.3K views •
Streamed 1 month ago

Thank You

https://www.linkedin.com/in/marv -arvaleski/

@mgrygles

https://discord.gg/RMU4Juw