ZHILIAN TECHNOLOGY CENTER

How Zhaopin built its Event Center using Apache Pulsar

Penghui Li
Sijie Guo

Zhaopin.com is the biggest online recruitment service provider in China

Zhaopin.com provides job seekers a comprehensive resume service, latest employment, and career development related information, as well as in-depth online job search for positions throughout China

Zhaopin.com provides professional HR services to over 2.2 million clients and its average daily page views are over 68 million.

Who are we

Penghui Li

- Tech lead of infrastructure team at zhaopin.com
- 5+ years of experiences developing message queues and microservices
- Apache Pulsar Committer

Who are we

Sijie Guo

- Apache Pulsar Committer & PMC Member
- Apache BookKeeper Committer & PMC Member
- Interested in technologies around Event Streaming
- Worked for Twitter and Yahoo before

1. Why building an Event Center
2. Why Apache Pulsar
3. Apache Pulsar at Zhaopin
4. Streaming Platform
5. Zhaopin’s contributions to Apache Pulsar

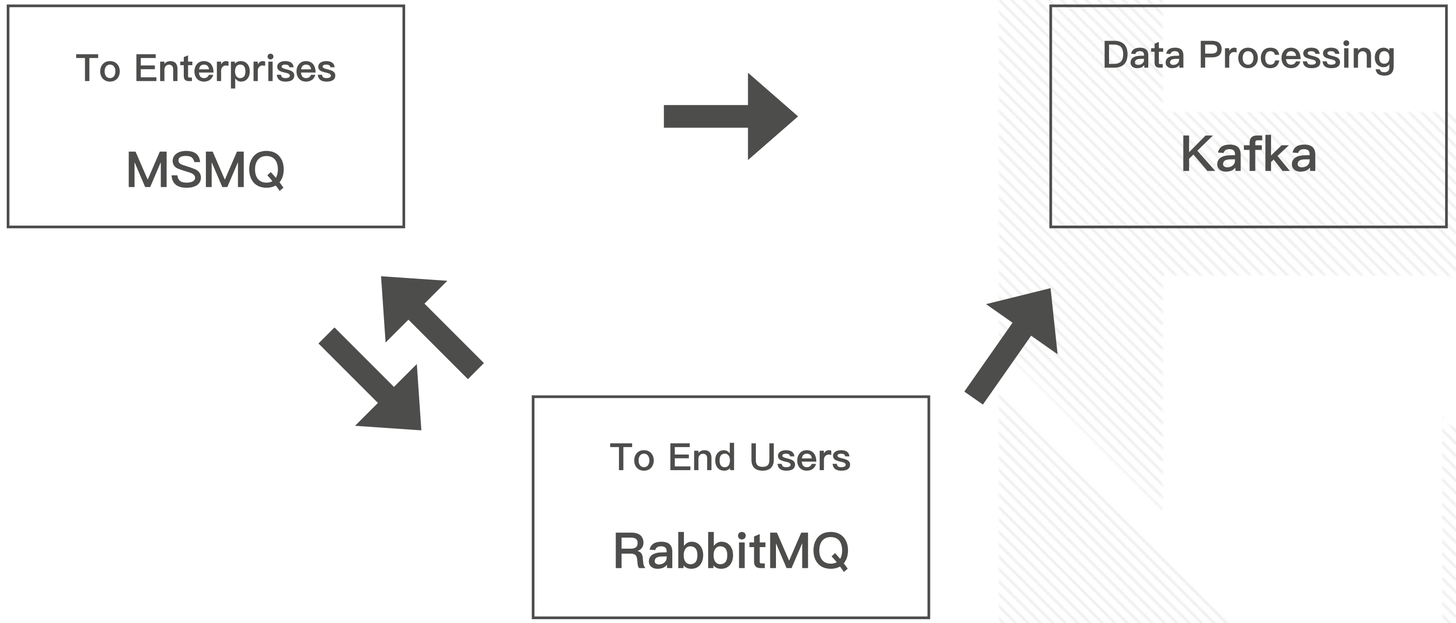
Why building an Event Center

Data Silos –> Unified Platform

Data Silos

Pain Points

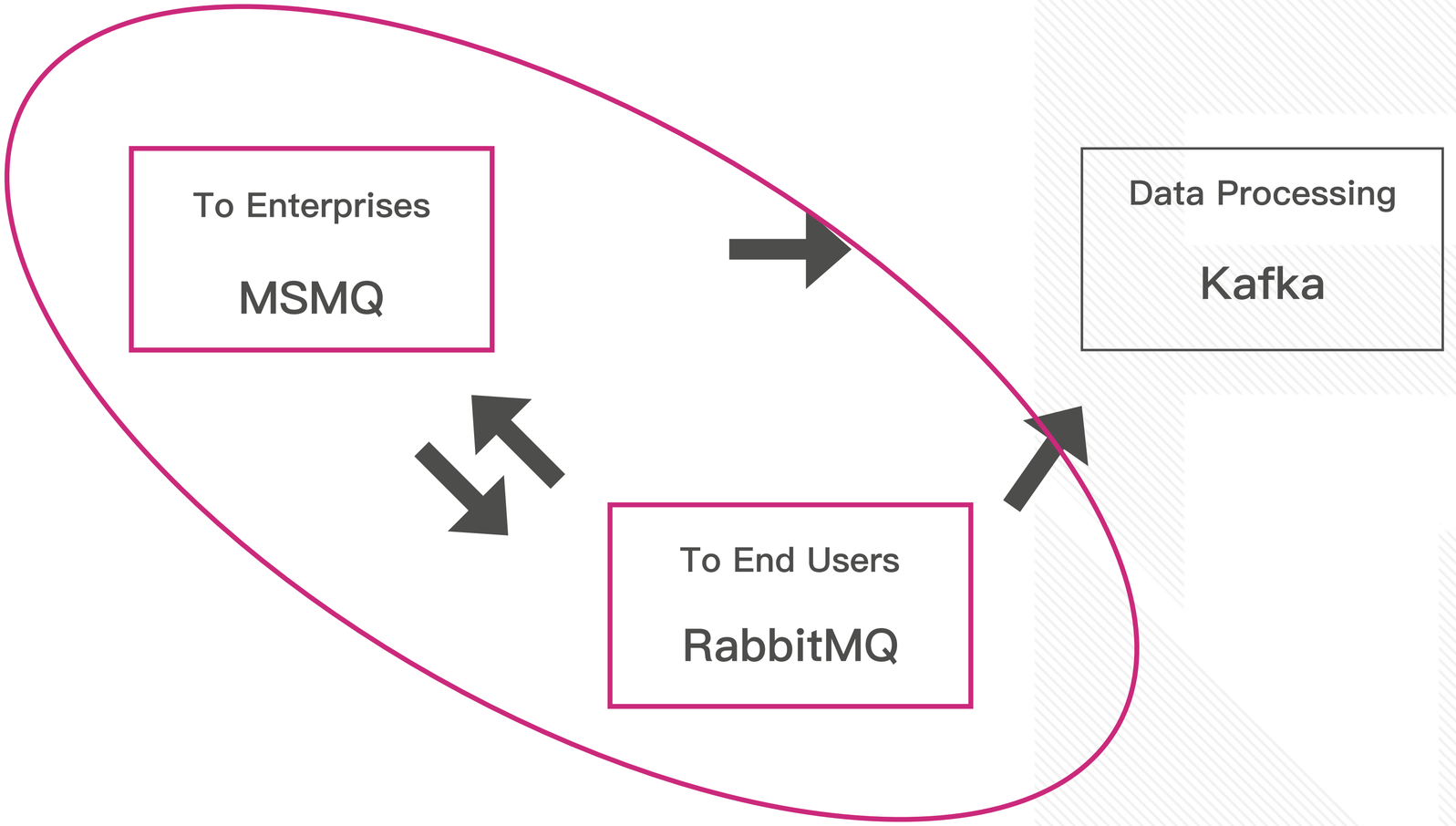
- High Maintenance Cost
- Extremely hard to share data cross teams
- Inconsistency between data silos
- Doesn’t Scale
- No consistent SLA



Data Silos

Pain Points

- High Maintenance Cost
- Extremely hard to share data cross teams
- Inconsistency between data silos
- Doesn’t Scale
- No consistent SLA



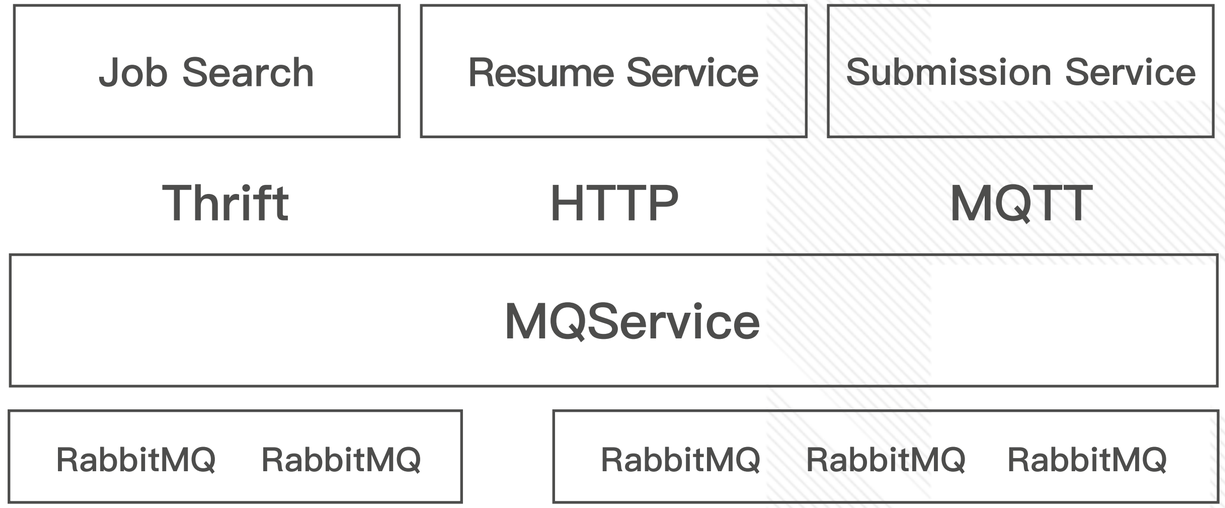
Unification – MQService

Problems Solved:

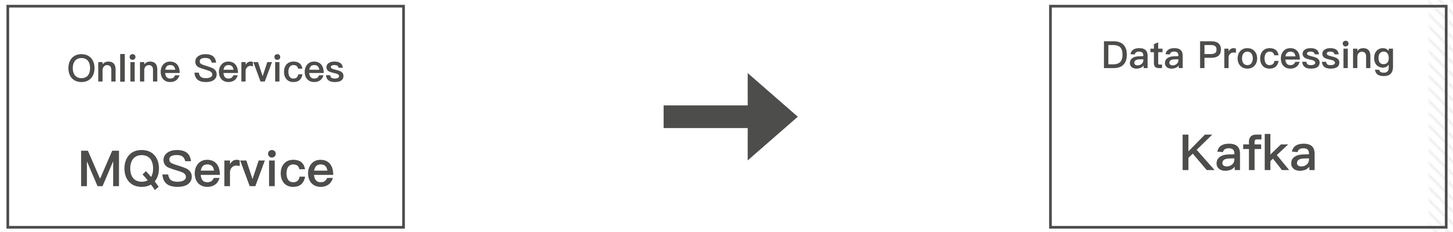
- Simplified Operations
- Scale-out Service
- High availability

Problems Unsolved:

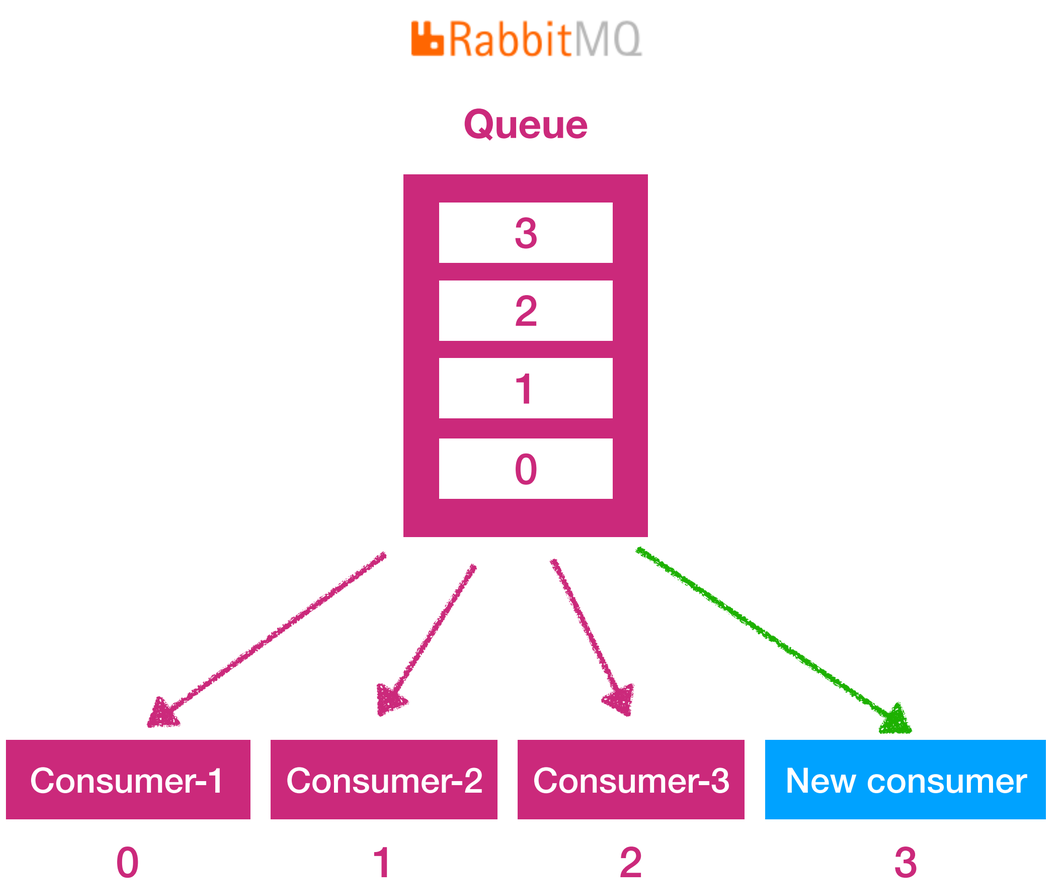
- Keep messages for longer period
- Data rewind
- Order Guarantee



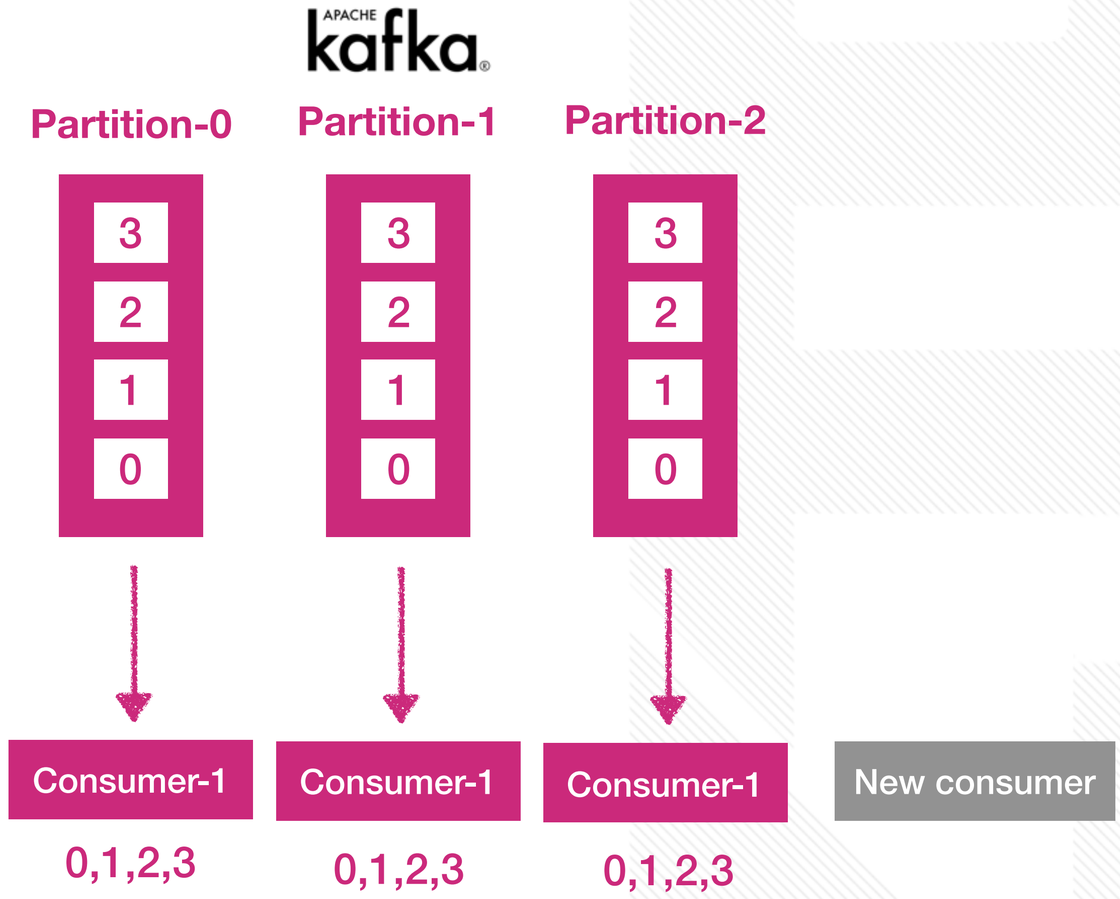
Unification – MQService



Why Building an Event Center



Better consumption parallelism



Better order guarantee

Why Building an Event Center

RabbitMQ is better for work queue use cases, more consumers can increase consumption. Kafka need more partitions to increase consumption.

We used RabbitMQ a lot for work queue use cases.

Why Building an Event Center

Kafka integrates well with the data processing ecosystem (Flink, Spark), and provides high throughput.

We used Kafka a lot for data processing.

Why Building an Event Center

But

The cost of operating two different message systems is high

Data sits at two different silos

We need a unified platform to handle both scenarios

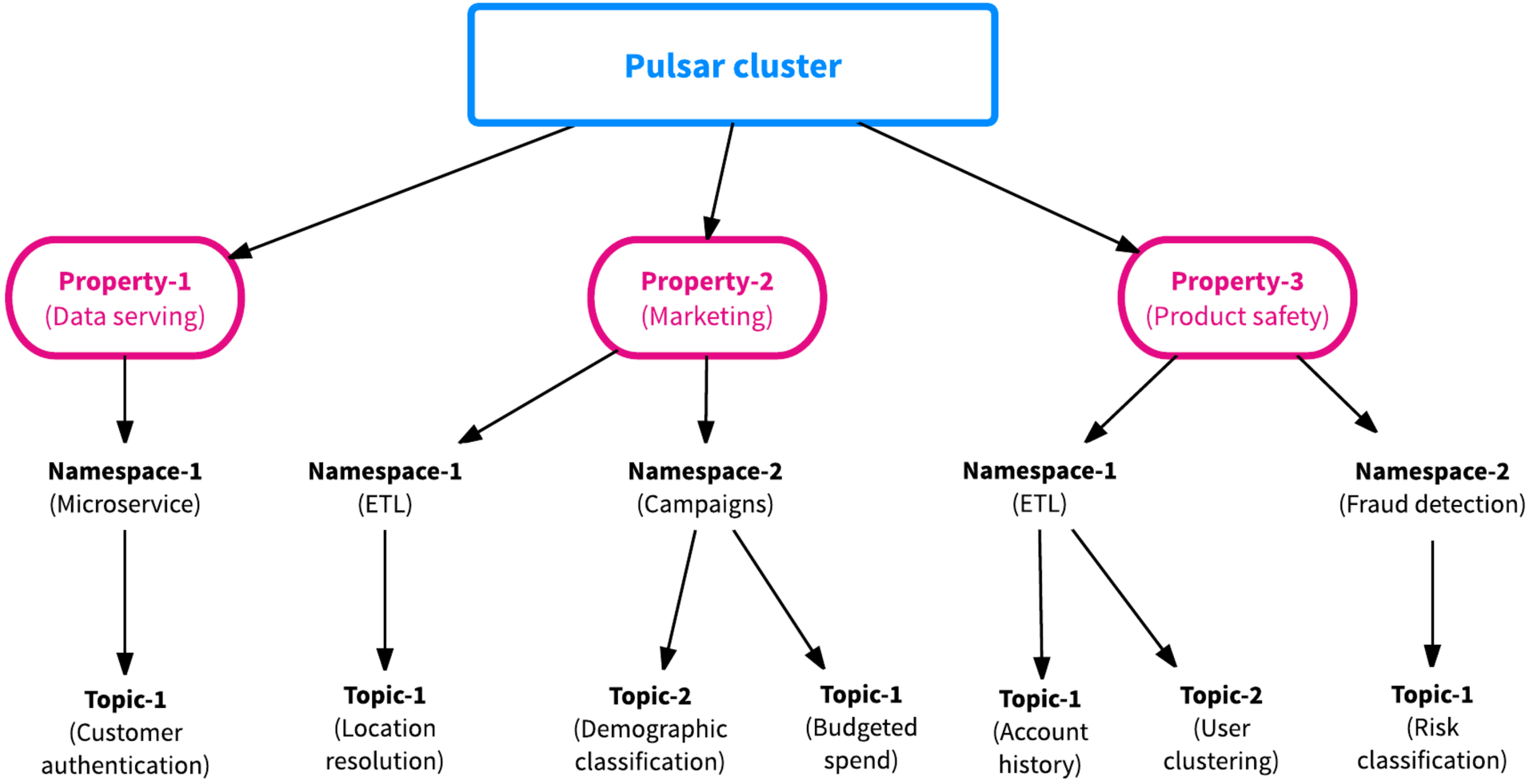
Why Apache Pulsar

Pulsar == Messaging + Storage

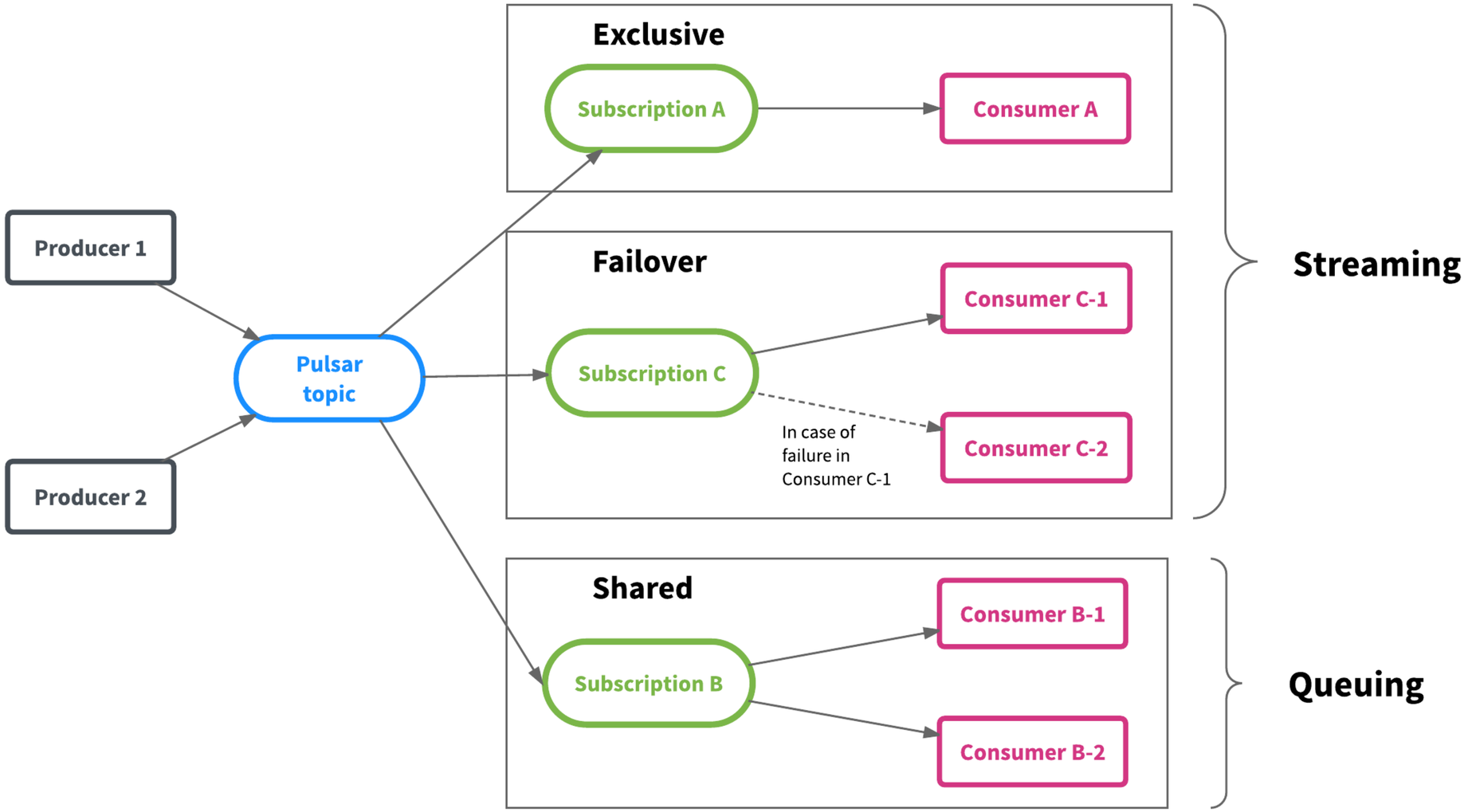
What is Apache Pulsar

“Flexible Pub/Sub **messaging**
backed by durable log/stream **storage**”

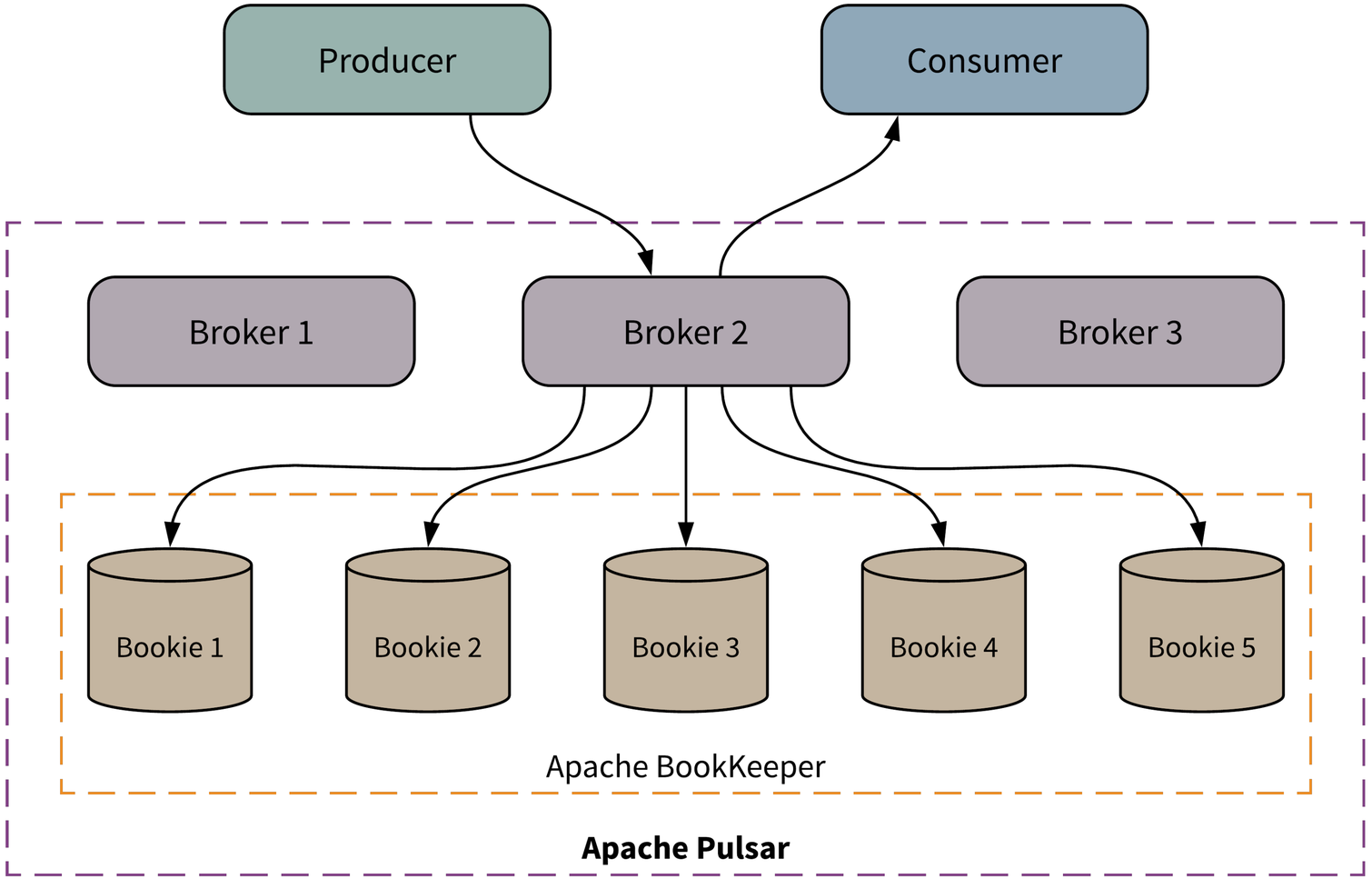
Apache Pulsar – Multi Tenancy



Apache Pulsar – Queue + Streaming



Apache Pulsar – Cloud Native



Layered Architecture

- Independent Scalability
- Instant Failure Recovery
- Balance–free on cluster expansions

Why Apache Pulsar

1. Pulsar provides a better abstraction of consumption patterns
2. Pulsar provides better fault tolerance and consistency options
3. Pulsar uses a scalable storage system (Apache Bookkeeper)
4. Hierarchical topic management and resource isolation

Perfect match with our requirement.

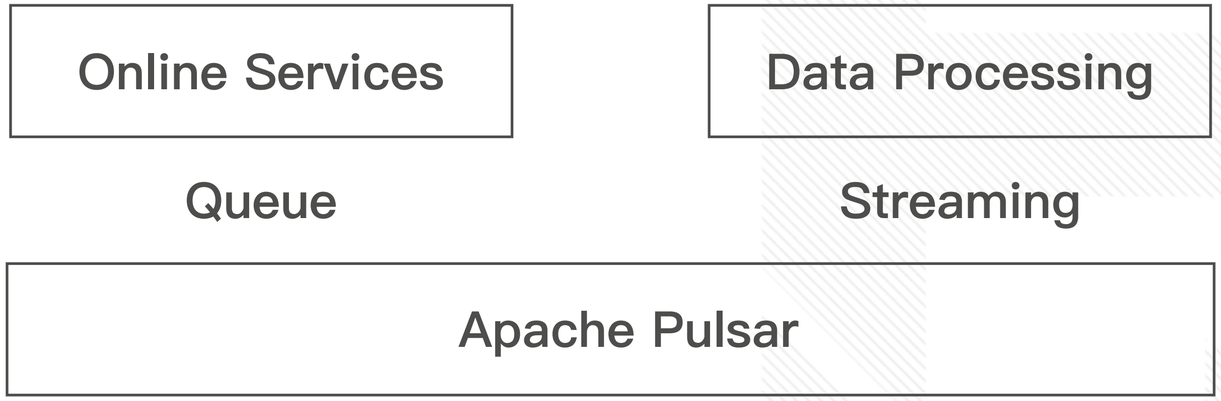
Apache Pulsar at Zhaopin

20+ core services, 6 billions msgs/day

Unification – Apache Pulsar

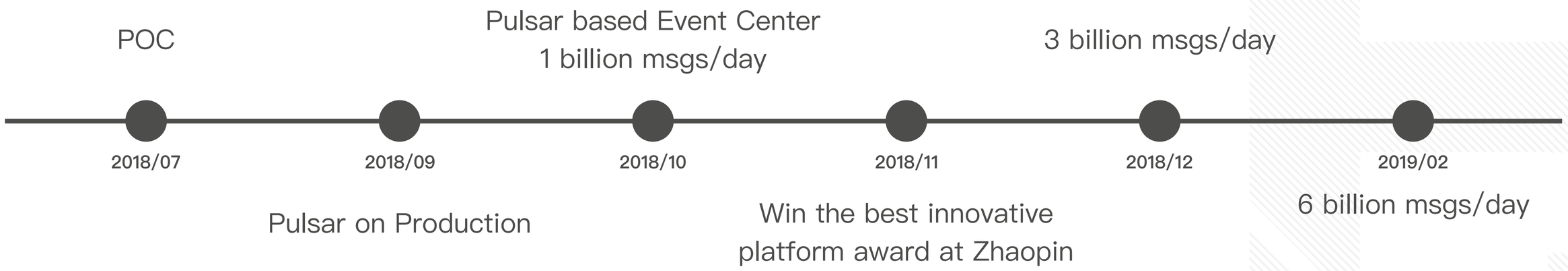
Problem Solved:

- No Data Silos
- Queue + Streaming
- Disaster Recovery
- Infinite Message Storage (via Tiered Storage)
- Data rewinding



Milestones

ZHILIAN TECHNOLOGY CENTER



Core Metrics

ZHILIAN TECHNOLOGY CENTER

50+ Namespaces

3000+ Topics

6+ billion Messages per day

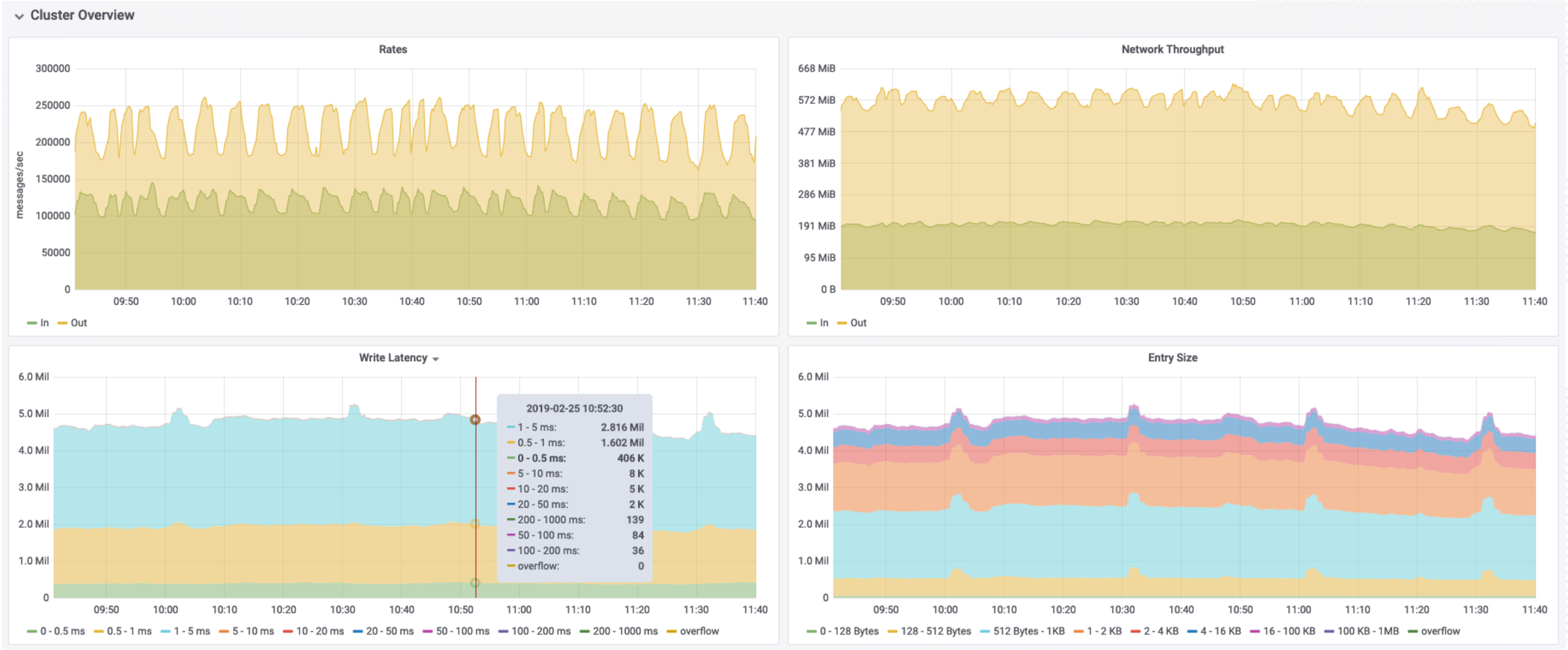
3TB Storage per day

20+ Core Services

System Metrics

ZHILIAN TECHNOLOGY CENTER

Write 100K+/s Read 200K+/s Network In 190MB+/s Network Out 550MB+/s Latency 99.5% < 5ms



Pulsar at Zhaopin

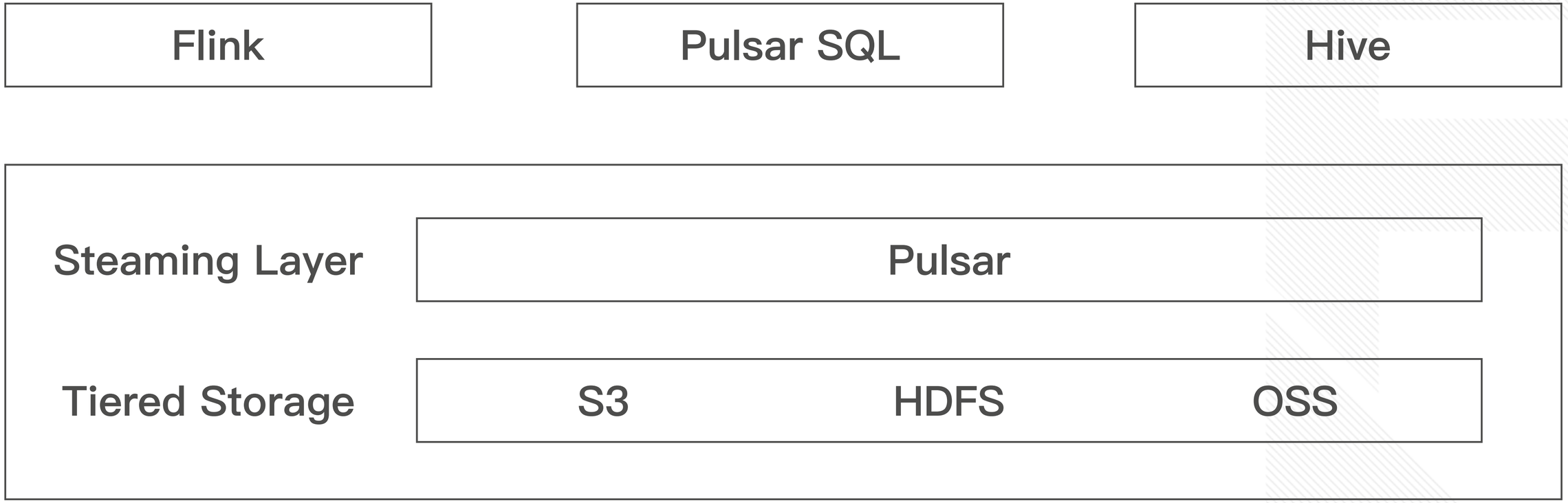
ZHILIAN TECHNOLOGY CENTER

1. One copy of data, single source-of-truth.
2. Don't worry about data consistency between RabbitMQ and Kafka
3. Multi-tenancy makes topic management easier
4. Strong data durability allows us to stop worrying about message loss

Streaming Platform

Beyond an Event Center

Streaming Platform



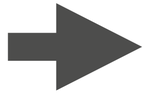
Stream to Stream

Table -> Table

Stream -> Table

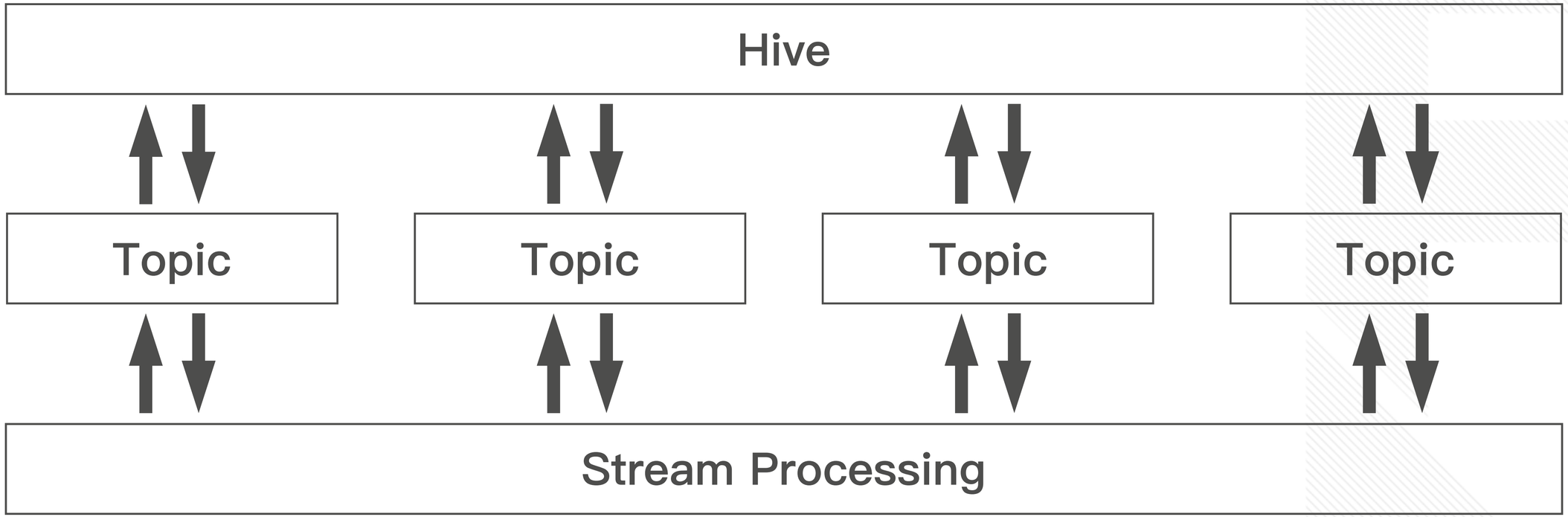
Table -> Stream

Stream -> Stream



Stream -> Stream

Unified Data Processing



Contribute to Apache Pulsar

Zhaopin's Contributions to Pulsar

ZHILIAN TECHNOLOGY CENTER

Client interceptors

We use this feature to track message between producer and consumers

Dead Letter Topic

Time partitioned message tracker

Service url provider

We use this feature to dynamically switching traffic

Hive Pulsar integration

Muti-version Schema and more...

Thank you