Streaming Aggregation of Cloud Scale Telemetry

Shay Qiuxuan Lin
Staff Software Engineer, Confluent

Telemetry

Telemetry is the unit measurements in distributed systems, the foundation for Chaos Engineering

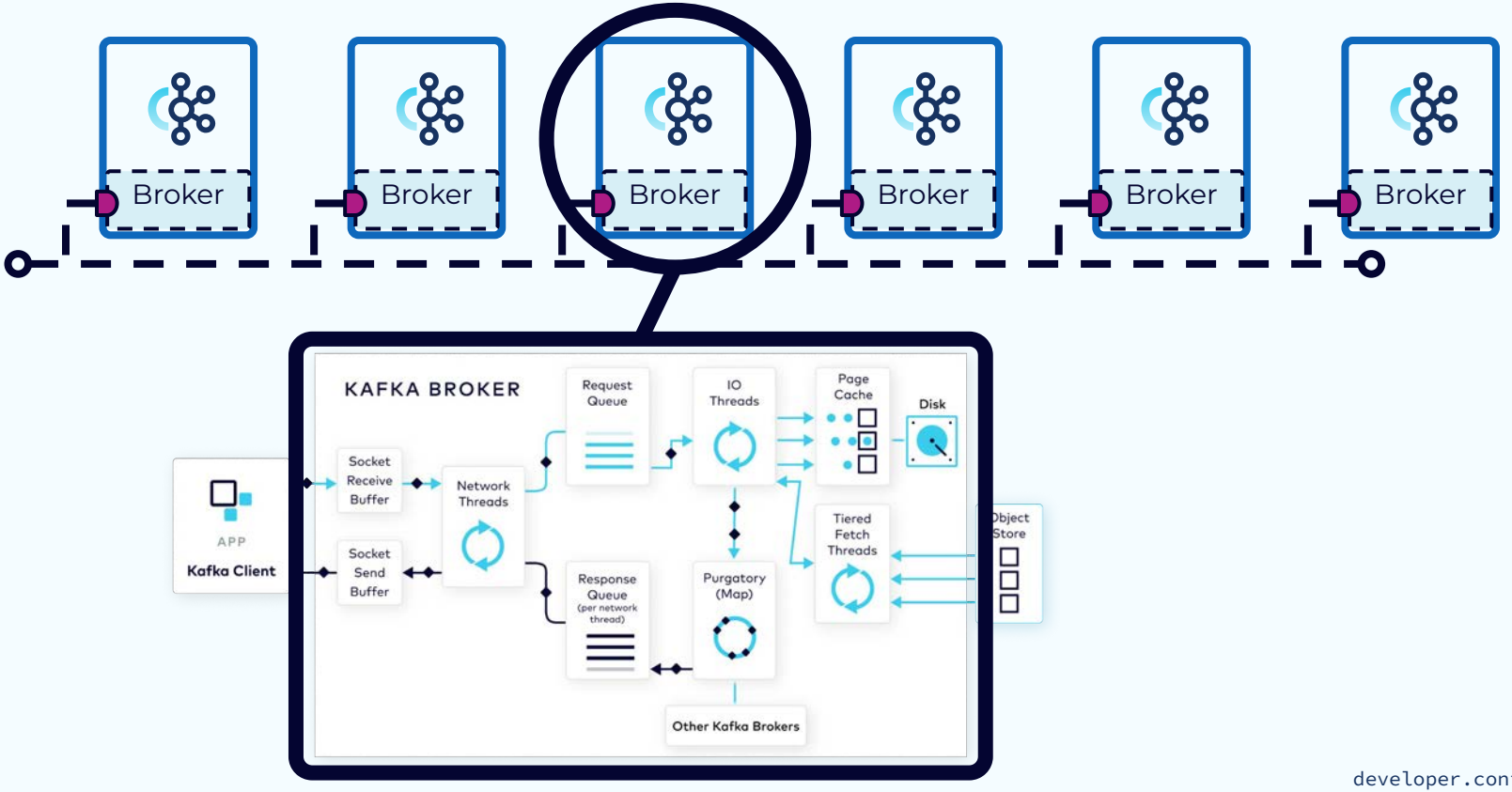
Telemetry

- Establish measurements for Knowns
- Discover Unknown Unknowns
- Experiment continuously

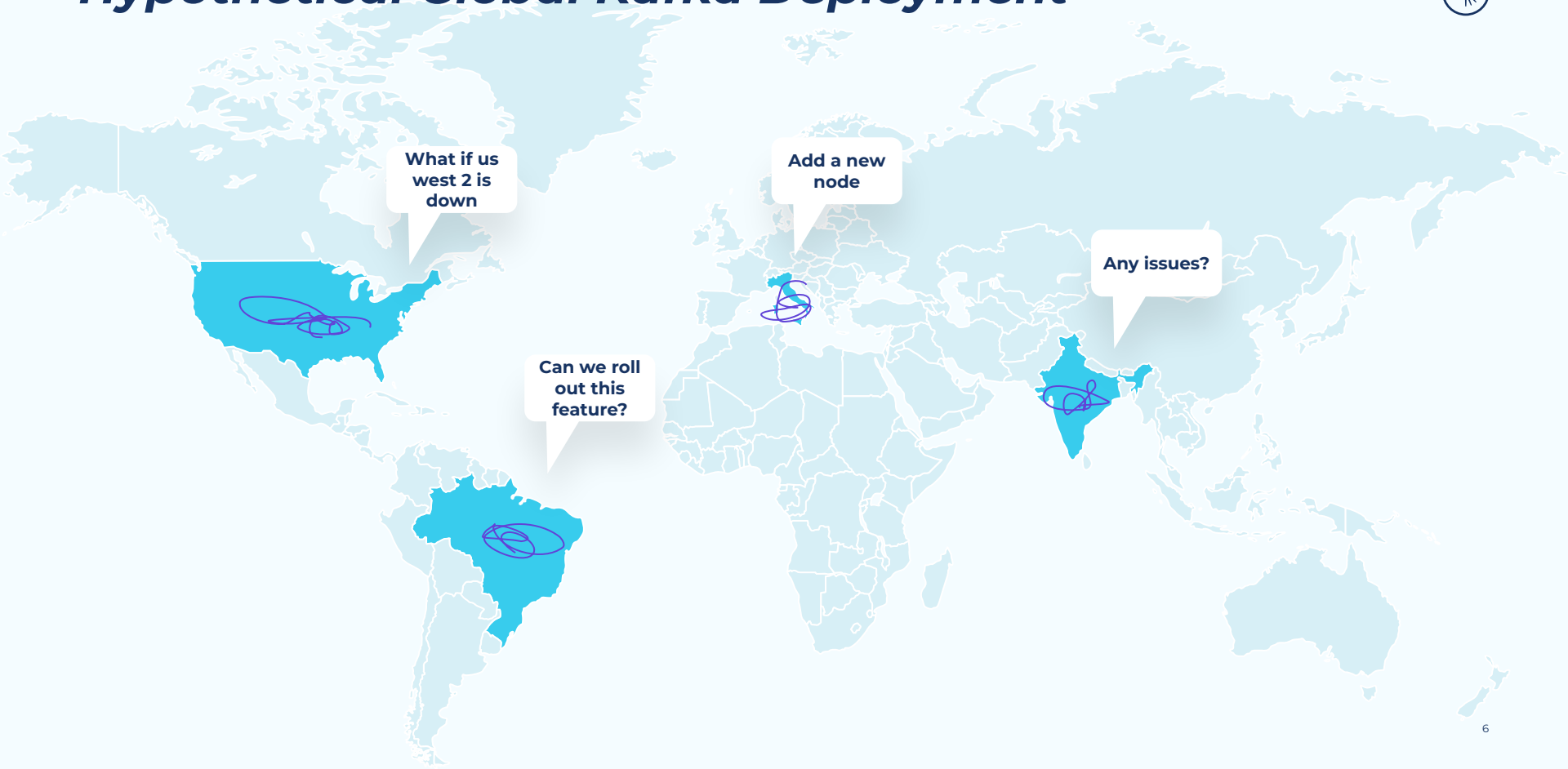
Kafka Reliability and Chaos Engineering

Inside The Apache Kafka Broker

CONFLUENT Developer

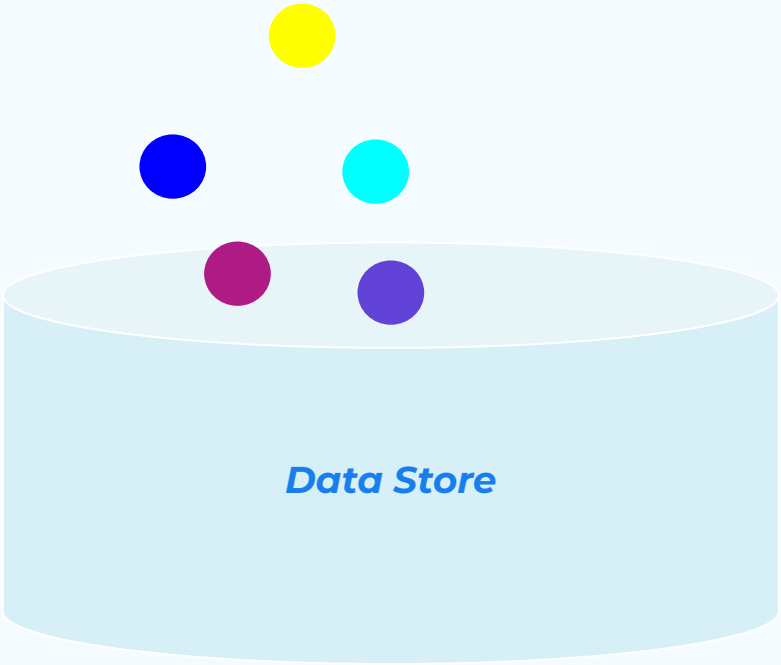


Hypothetical Global Kafka Deployment



Serving Telemetry: The evolution

Kafka Telemetry Serving

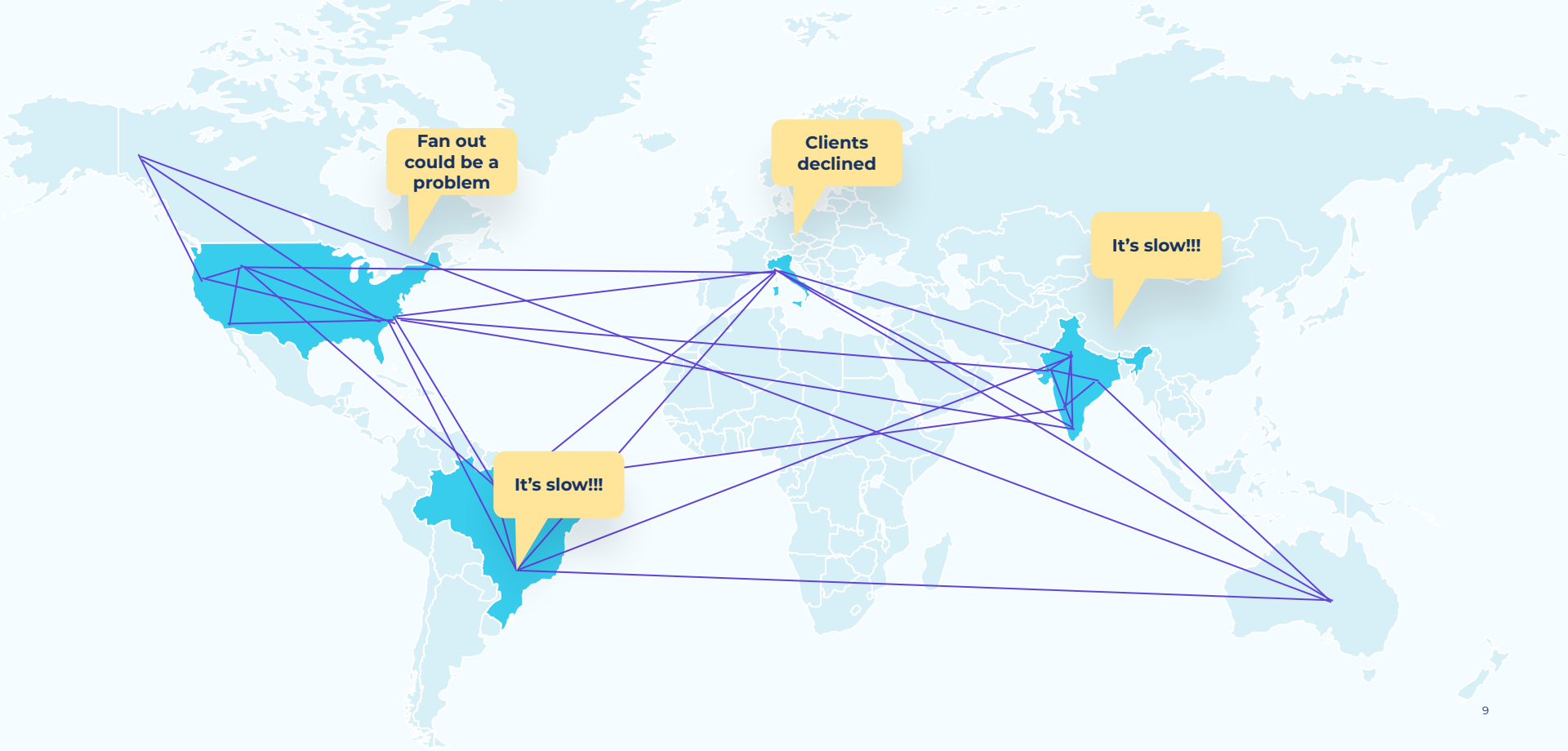


Usage Patterns

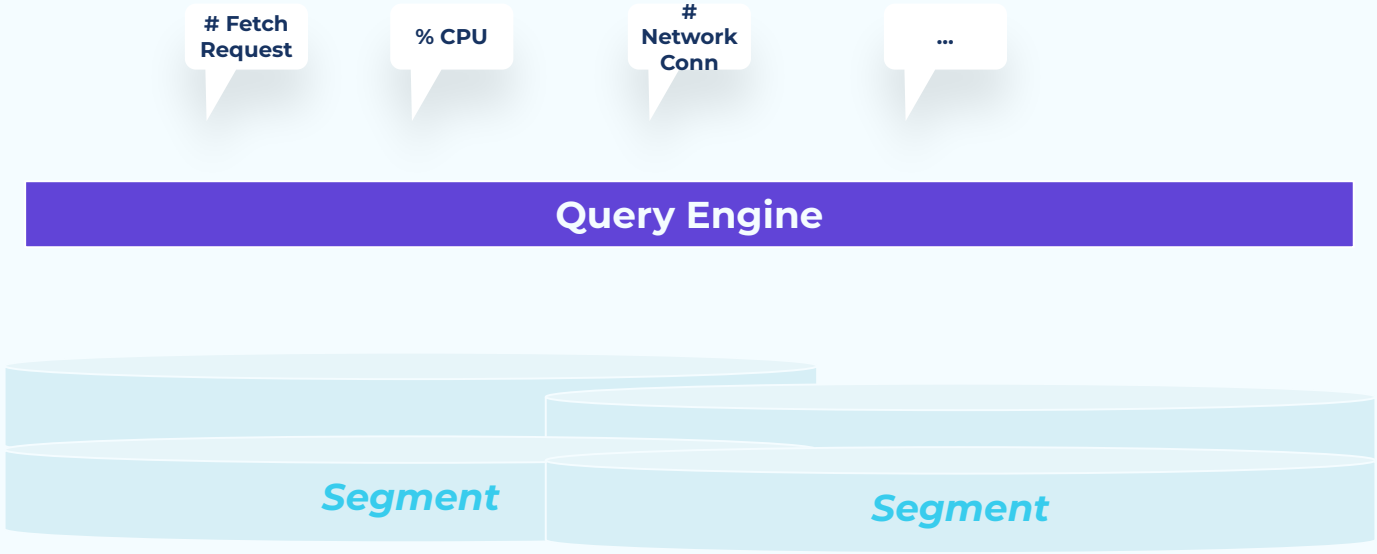
- (Facts) Total network in bytes in an hour
- (Trends) CPU Usage Trend in last 7 days
- (Diagnose) Point in time value for storage
- (Attribution) Compare CPU trend and fetch request in last 7 days

...

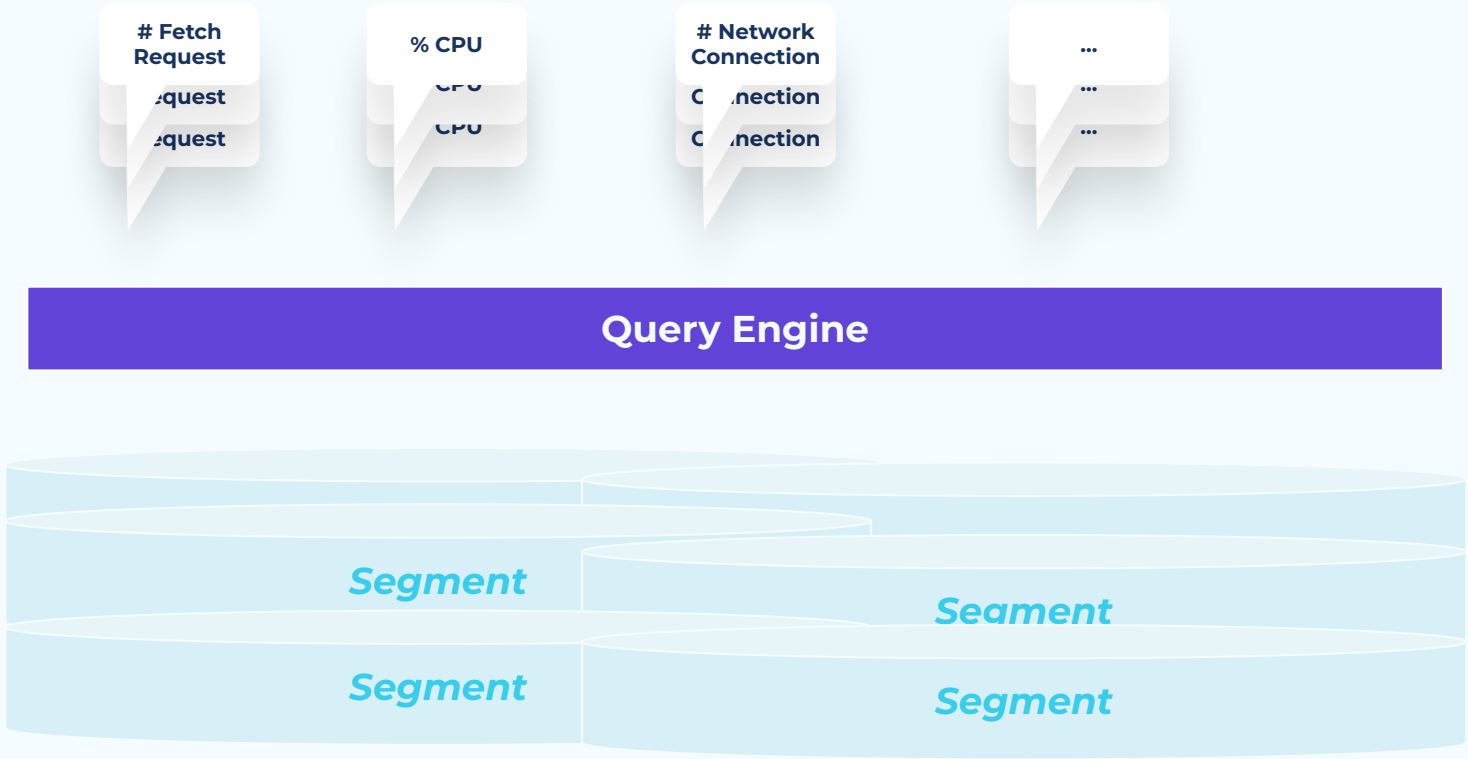
Hypothetical Global Kafka Deployment



Time Series Optimized(OLAP)

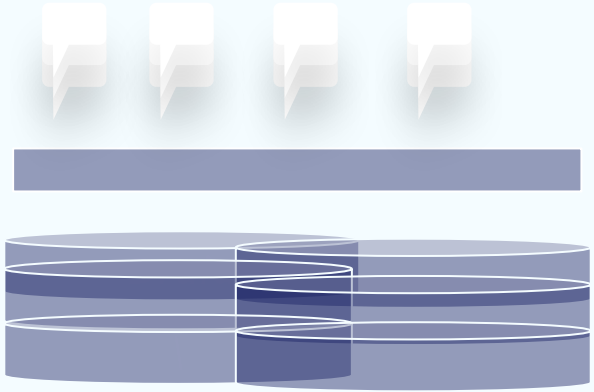


Highly concurrent ingestion and queries



Scalability Concerns

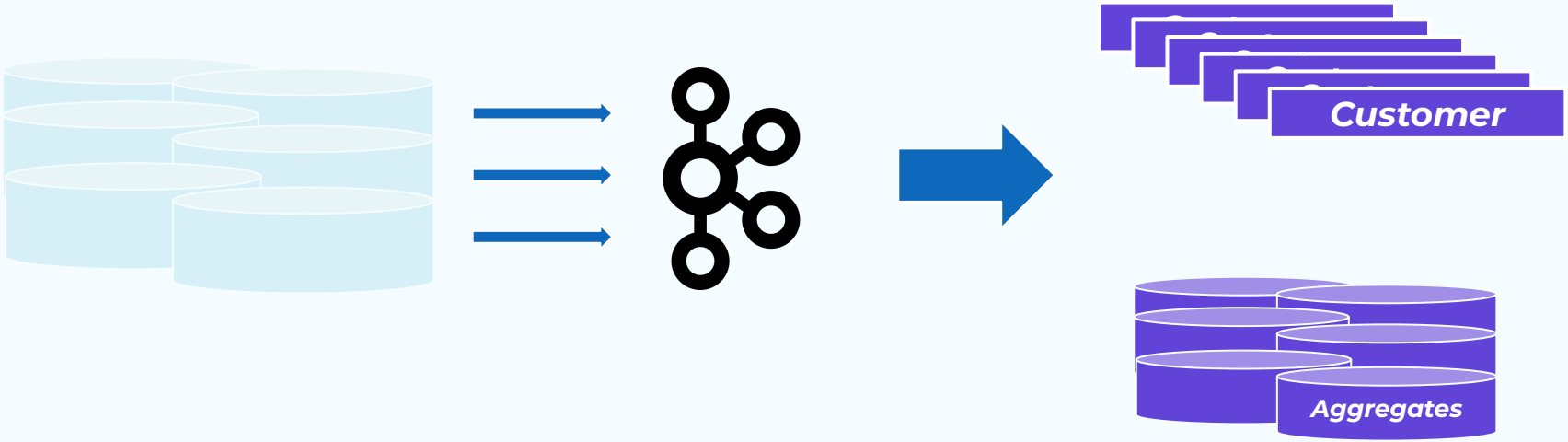
Highly concurrent ingestion and queries



- Repetitive queries on hot entities
- Rising compute cost and serving cost
- Metric definition needs to be consistent
- Increasing customers depend on the reliability

Architecture Deep Dive

Solution: Pushing Asynchronous Aggregation



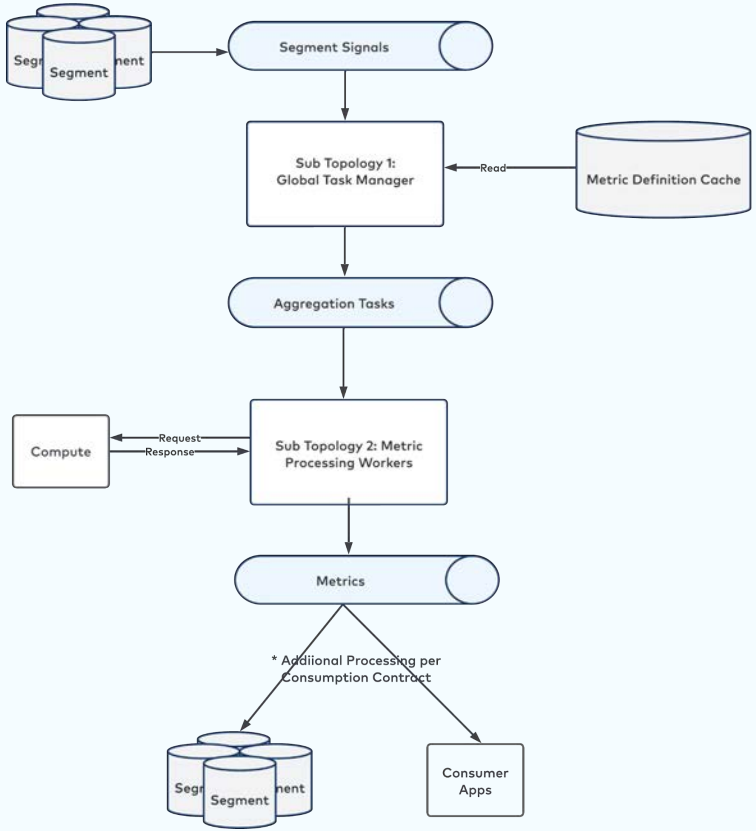
Raw Telemetry

Kafka Streams
Predefined aggregations

Consumers

Kafka Streams

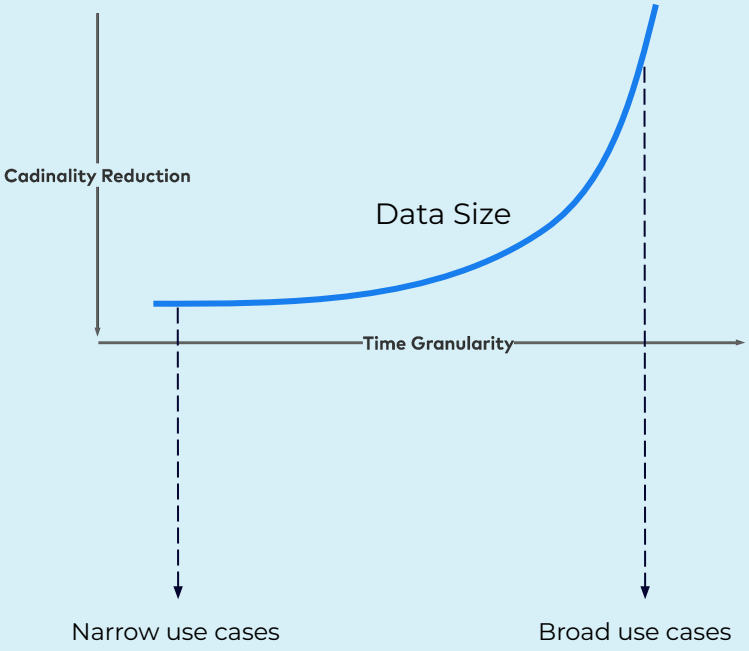
Topology



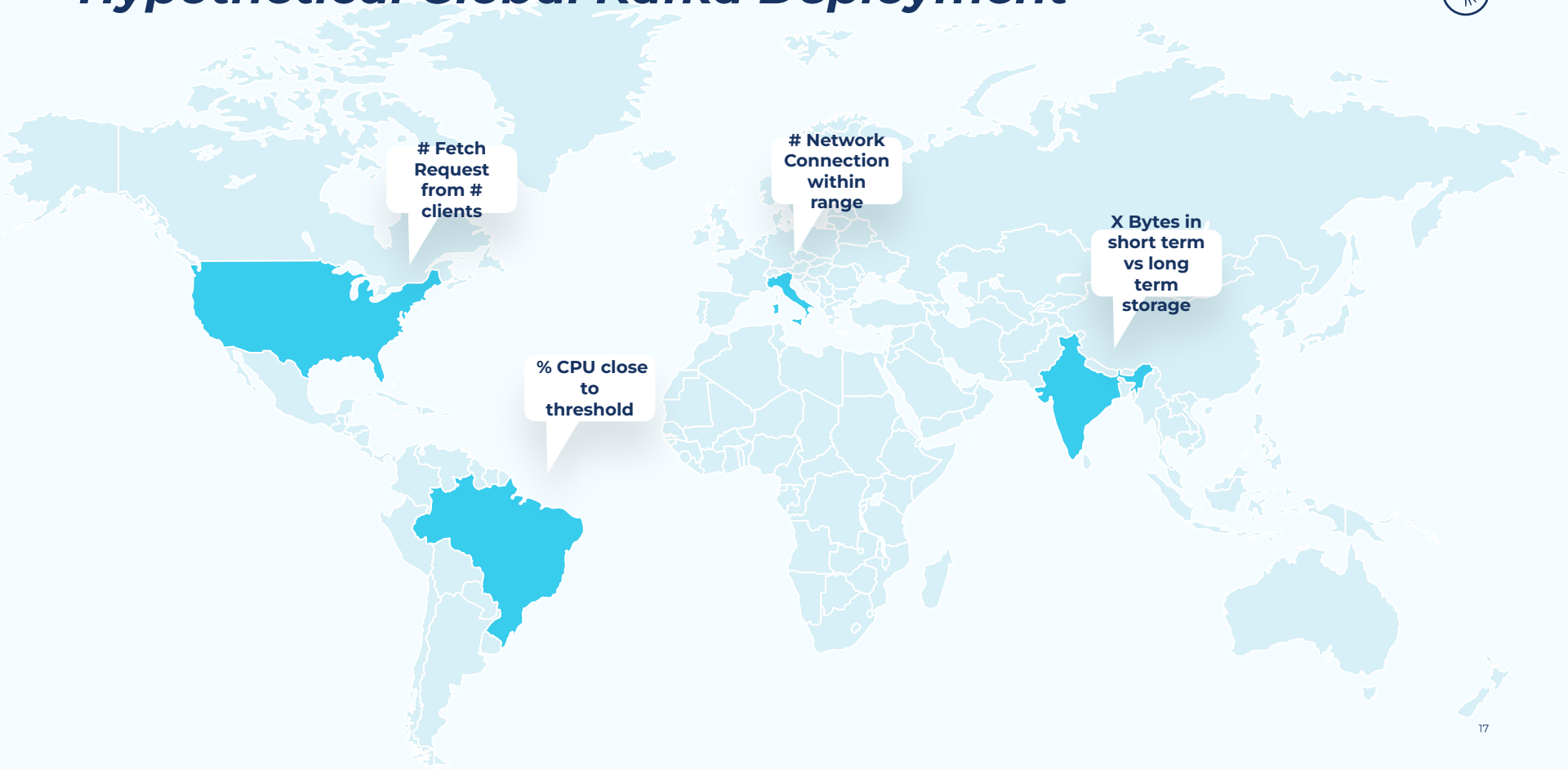
Alternative Architectures

Offline Custom rollup tasks, e.g. Apache Pinot: Star Tree Index, Apache Druid

Online Raw telemetry stream processing (Flink, KStreams)



Hypothetical Global Kafka Deployment



Takeaways

- Data platform evolution for chaos engineering

CONFLUENT

Twitter: @QiuxuanL

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