

# 大数据流式技术及实践

TalkingData研发副总裁 阎志涛



**SFDC**

SegmentFault  
Developer Conference



# About Me

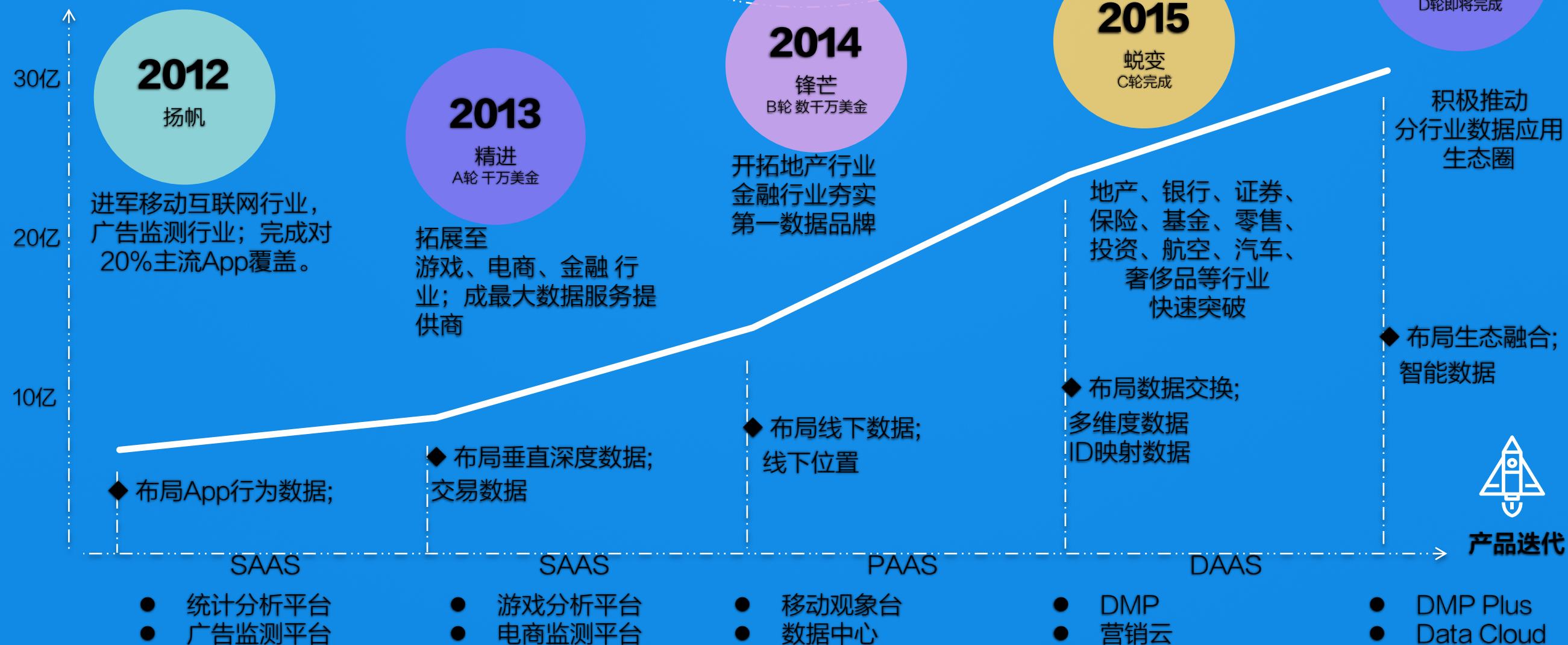
- Vice President of TalkingData
- 15+ years work experience in IT
- 4+ years of big data experience
- High technology fans
- Science fiction lover







数据积累



# About TalkingData



SFDC

SegmentFault  
Developer Conference



# Agenda

- What is stream processing?
- Why stream processing?
- Key technologies for stream processing
- Stream processing reference architecture
- Stream processing at TalkingData



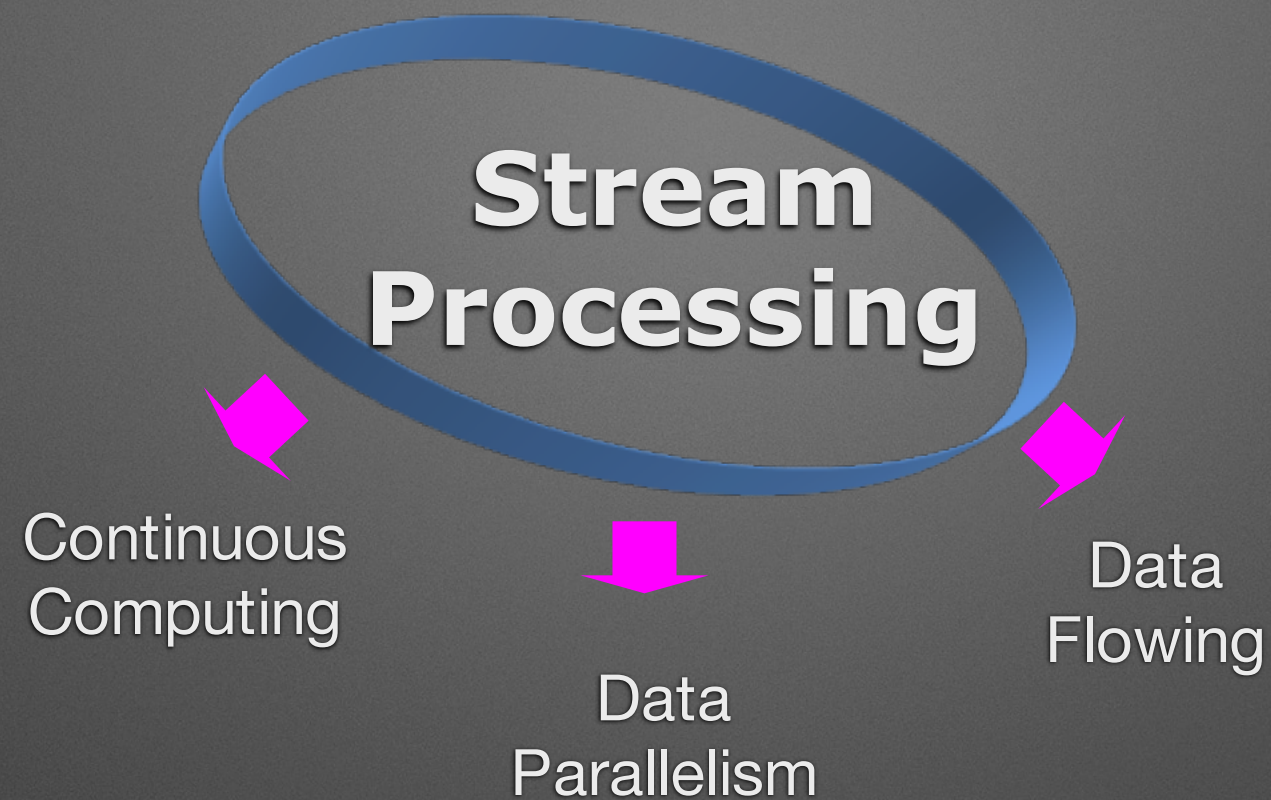


# What Is Stream Processing

- Stream processing is a computer programming paradigm, equivalent to [dataflow programming](#), [event stream processing](#), and [reactive programming](#), that allows some applications to more easily exploit a limited form of [parallel processing](#) -- wikipedia.







# What Is Stream Processing?





Stream processing means

fast and low latency





# Why Stream Processing?

- **Batch Processing**

- **Hourly error logs** - what's wrong in last hour
- **Daily click-install reports** - how many new app users come from ads yesterday
- **Daily fraud reports** - fraud information yesterday

- **Stream Processing**

- **Real-time error metrics** - what's going wrong now
- **Real-time click-install attribution** - get real-time attribution result
- **Real-time antifraud** - block fraudulent on time





# Stream Is New Buzzword!



Batch Processing



Stream Processing





# Key Technologies

- Event Log
- Message Queue
- Stream Computing





# Event Log

- Event log is **Core Abstraction** for data system





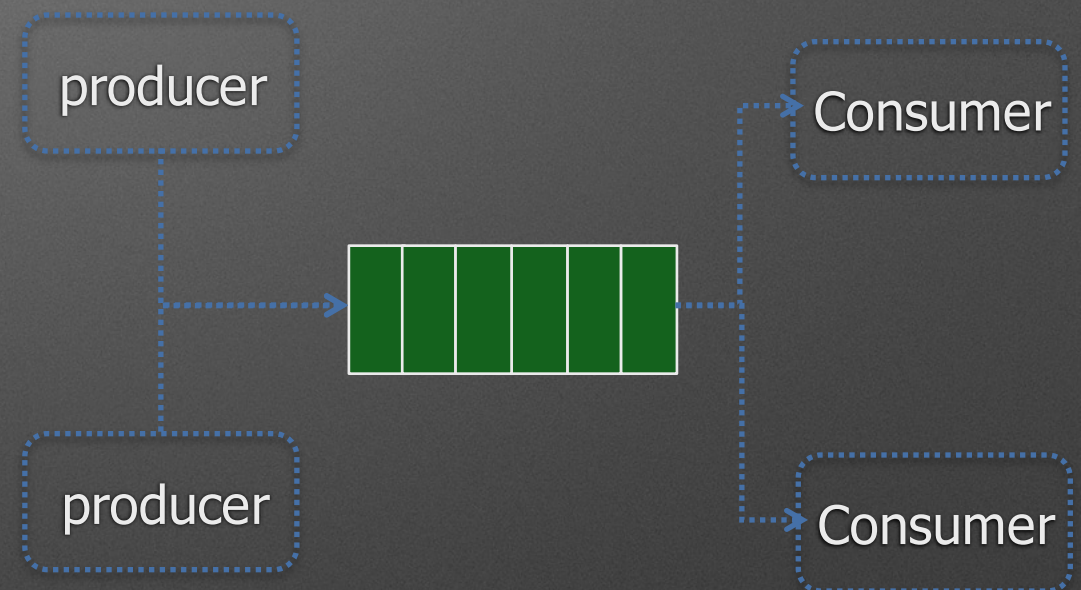
# Event Log





# Message Queue

- Message queues are the **Core Integration** tools
  - Decoupling
  - Parallelism
  - Reliability





# Message Queue

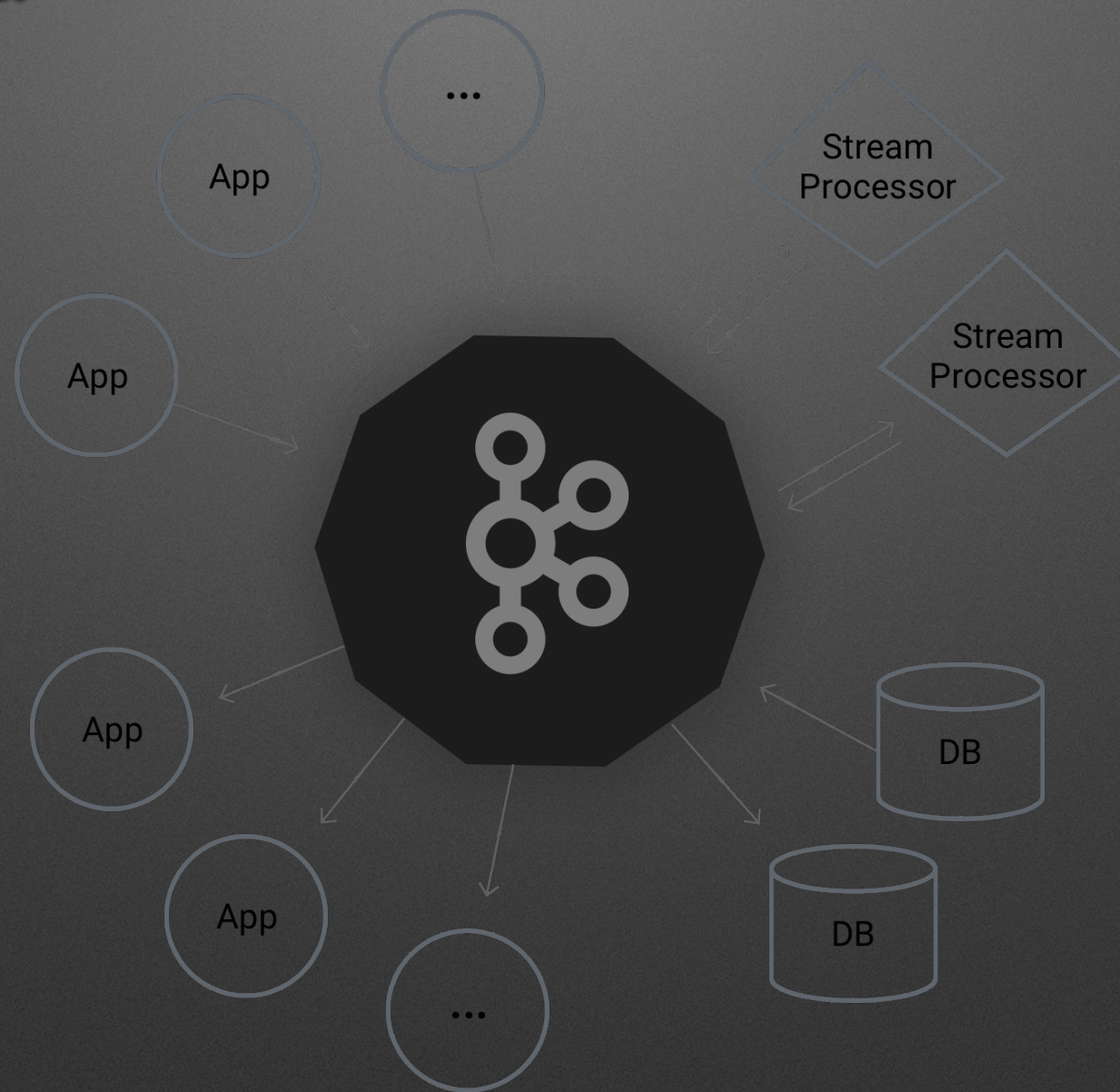
- At most once
- At least once
- Exactly once





# Message Queue

- Apache Kafka

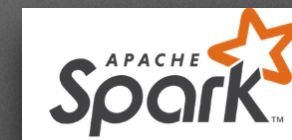




# Stream Computing

- Stream computing platforms are **Core Data Processing** tools

- Apache Spark



- Apache Flink



- Apache Storm



- Apache Apex



- Apache Beam





# Stream Computing

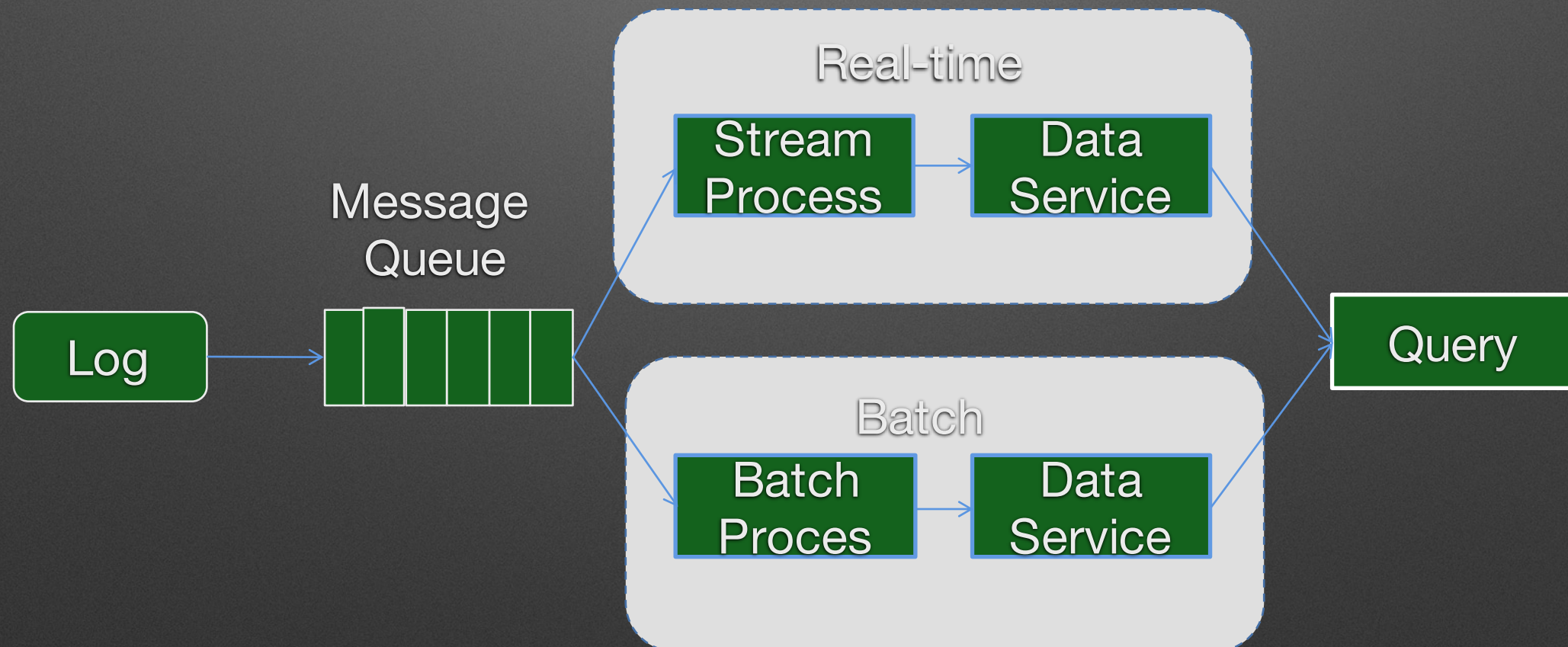
	Spark Streaming	Flink	Storm	Apex
throuput	high	high	low	high
latency	high	low	low	high
exactly-once	yes	yes	yes (through Trident)	yes
community activiness	high	medium	high	low





# Reference Architecture

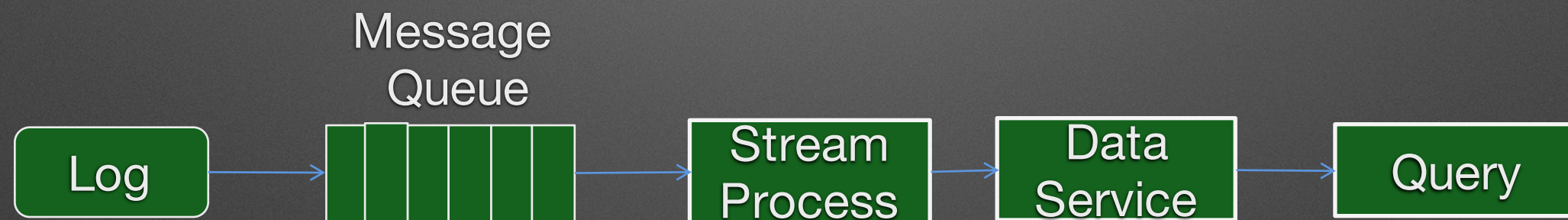
- Lambda Architecture





# Reference Architecture

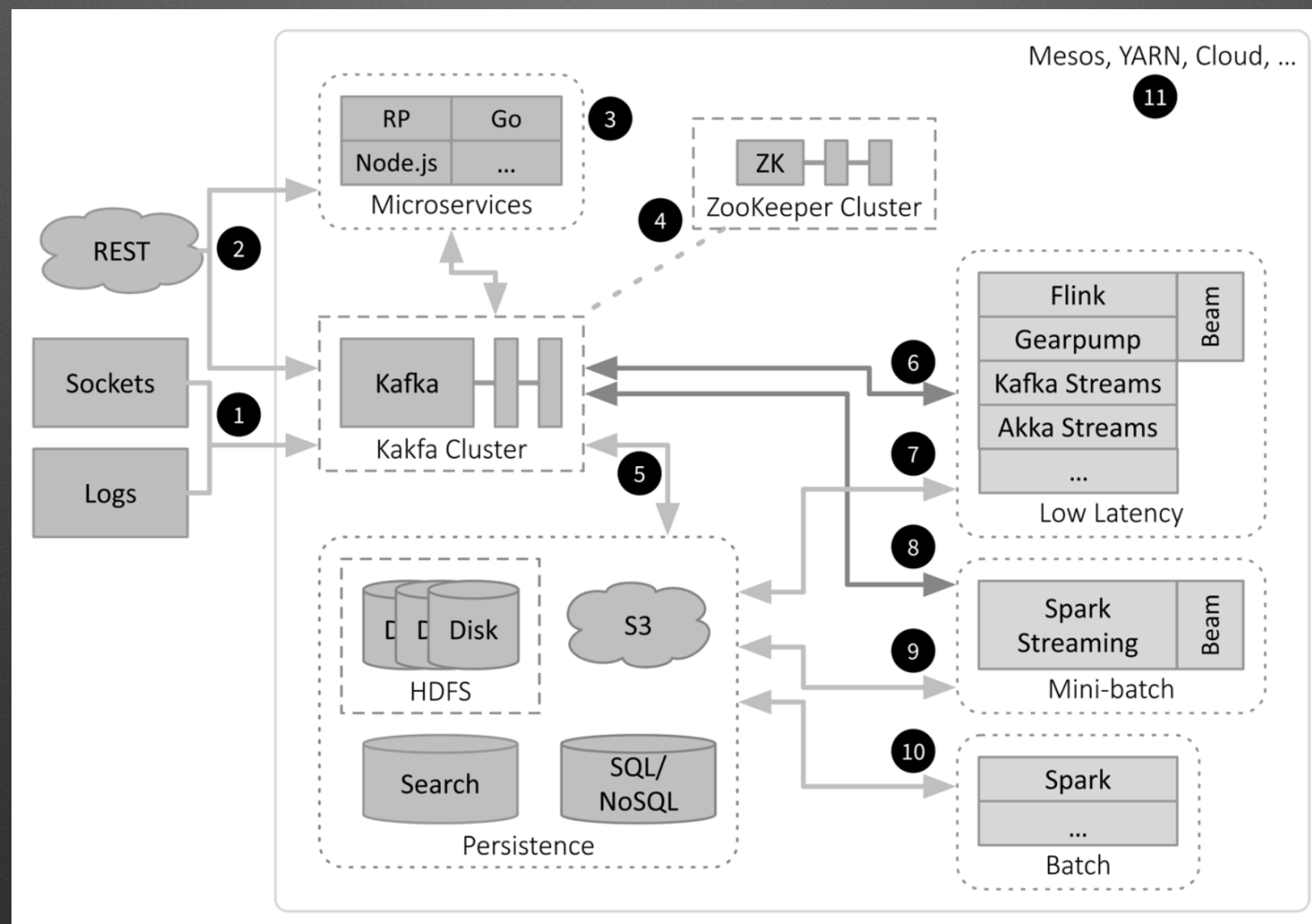
- Kappa Architecture





# Reference Architecture

- Comprehensive Architecture





# Stream Processing At TalkingData

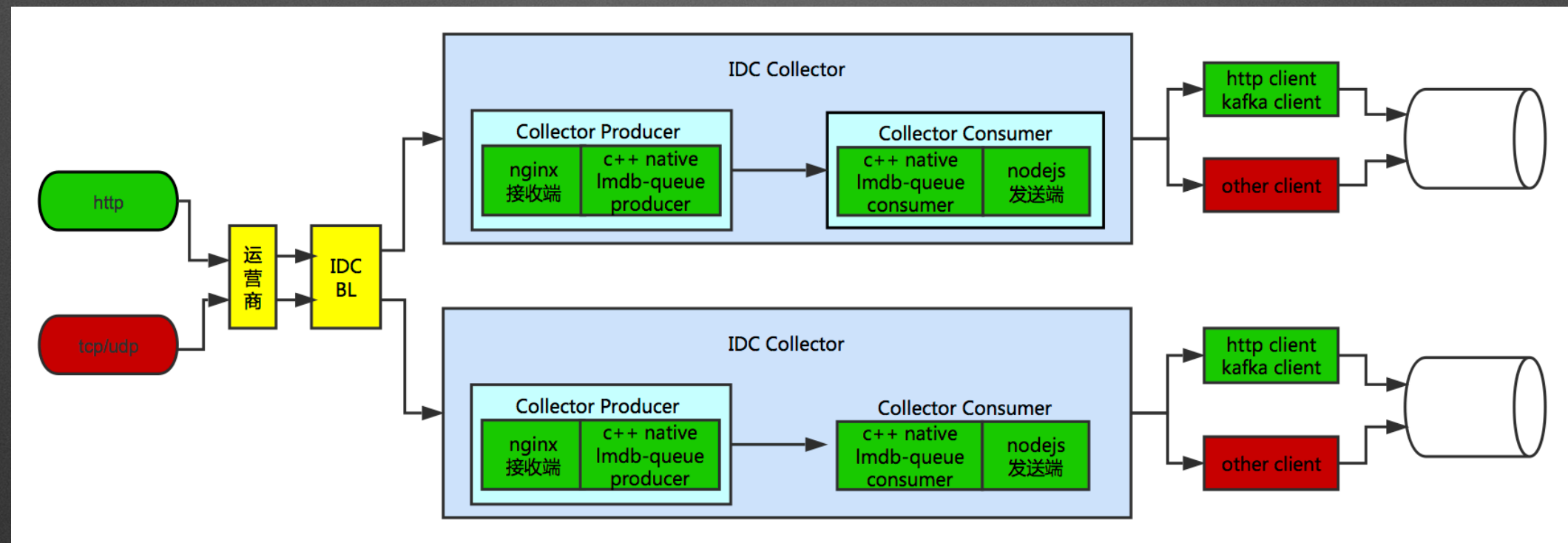
- Stream Data collection
- Stream analytics
- Stream attribution
- Stream fraud detection





# Stream Data Collection

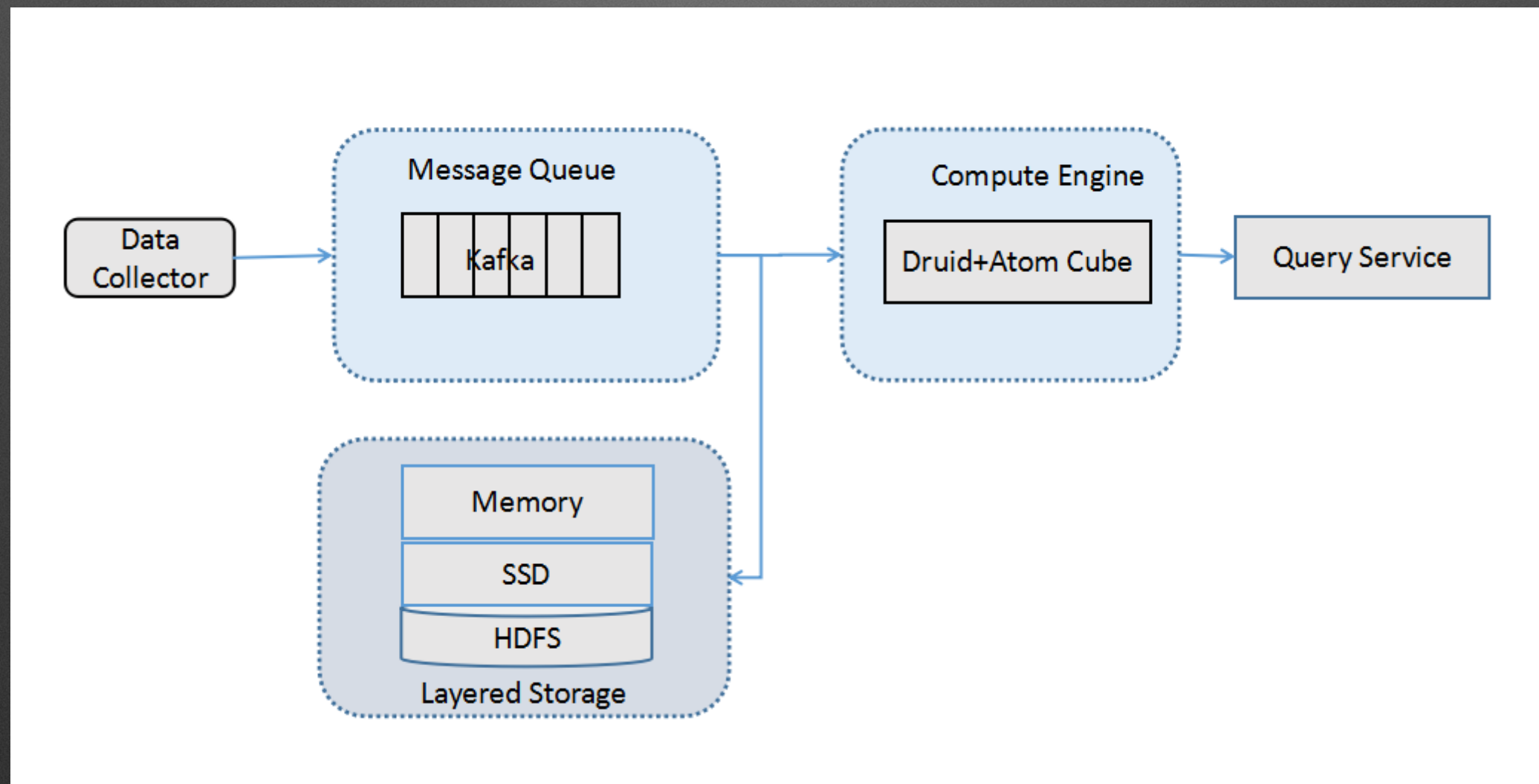
- Data flows as stream





# Stream Analytics

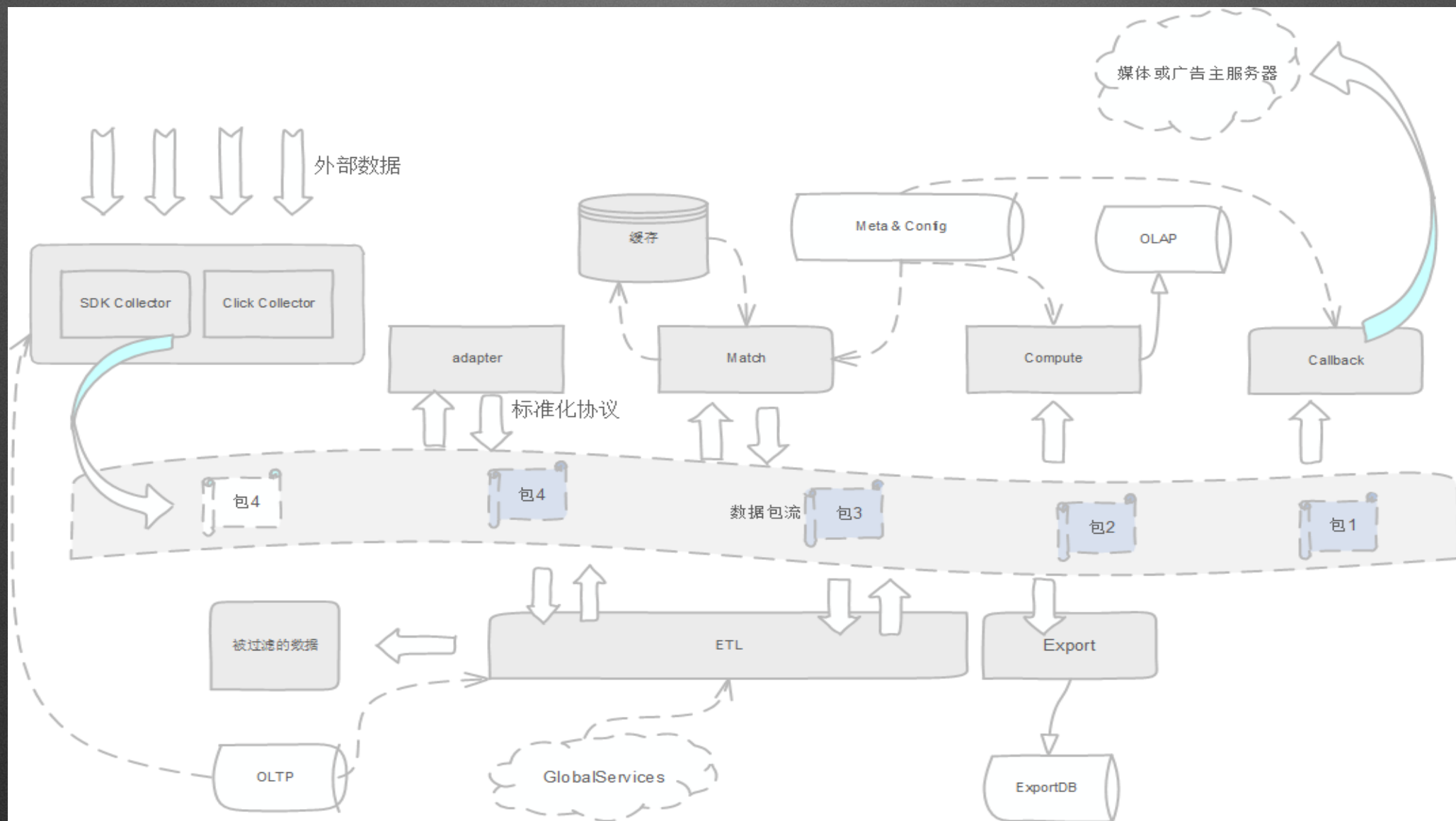
- TalkingData analytics architecture





# Stream Attribution

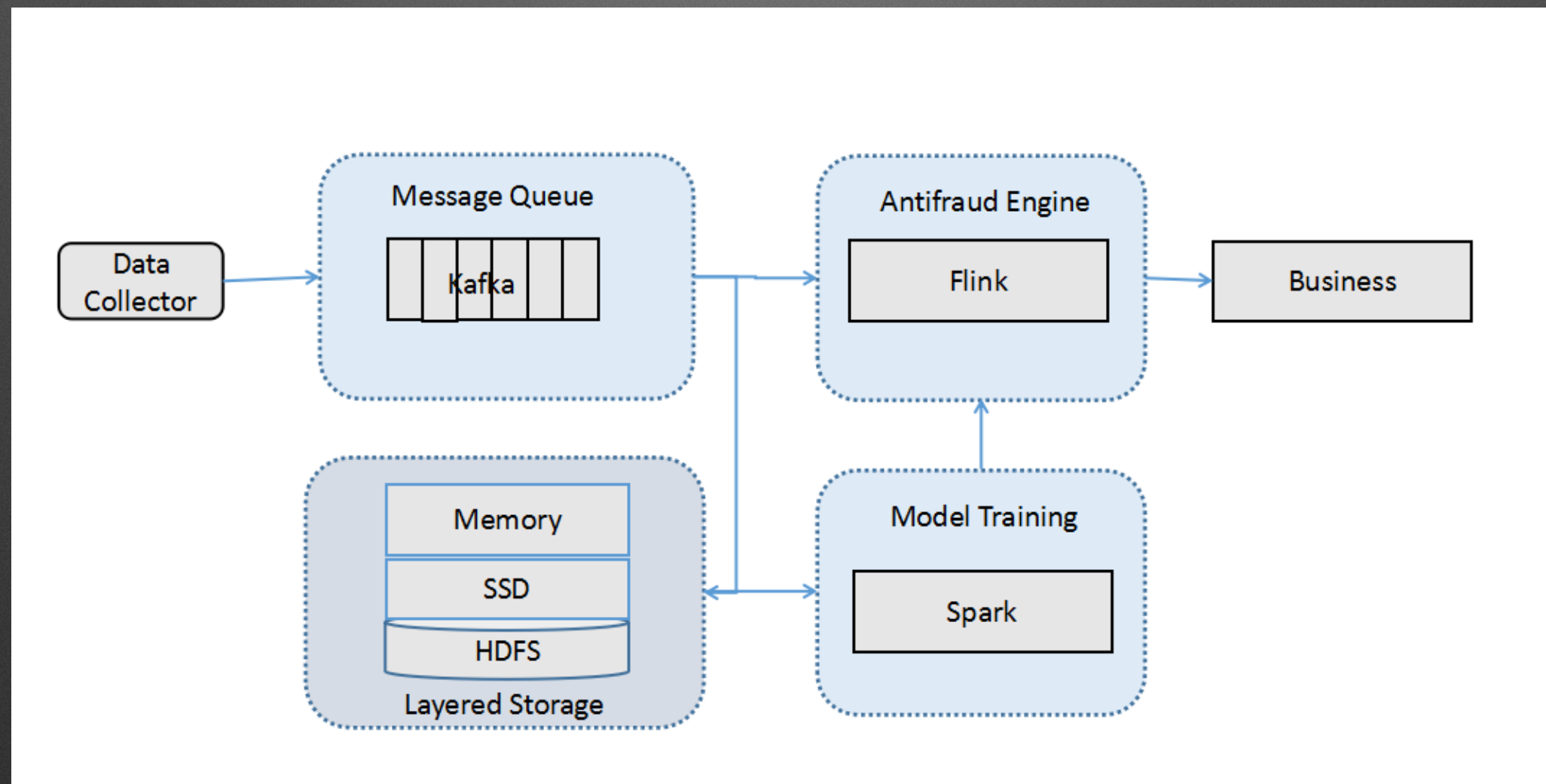
- Advertisement attribution need to be stream processing





# Stream Antifraud

- Real-time antifraud means save money



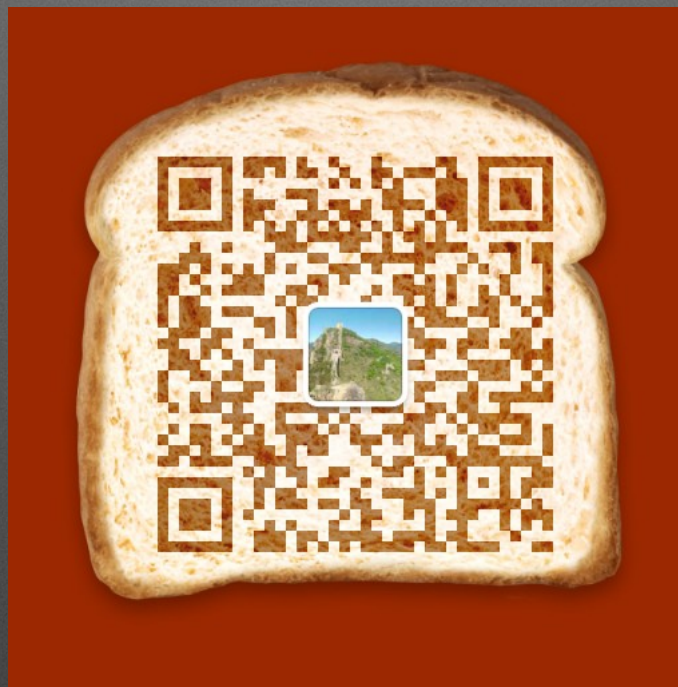


Stream Processing is new epoch of big data!





We are hiring!



hr@tendcloud.com



**SFDC**

SegmentFault  
Developer Conference