大数据流式技术及实践

TalkingData研发副总裁 阎志涛

About Me

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

极速发展的
TalkingData

数据积累

30亿
20亿
10亿

2012
扬帆

进军移动互联网行业，
广告监测行业；完成对
20%主流App覆盖。

2013
精进
A轮 千万美金

拓展至
游戏、电商、金融 行
业；成最大数据服务提
供商

2014
锋芒
B轮 数千万美金

开拓地产行业
金融行业夯实
第一数据品牌

2015
蜕变
C轮完成

2016
谋势
D轮即将完成

积极推动
分行业数据应用
生态圈

◆ 布局App行为数据;

◆ 布局垂直深度数据;
交易数据

◆ 布局线下数据;
线下位置

地产、银行、证券、
保险、基金、零售、
投资、航空、汽车、
奢侈品等行业
快速突破

◆ 布局数据交换;
多维度数据
ID映射数据

◆ 布局生态融合;
智能数据

产品迭代

SAAS	SAAS	PAAS	DAAS	
● 统计分析平台	● 游戏分析平台	● 移动观象台	● DMP	● DMP Plus
● 广告监测平台	● 电商监测平台	● 数据中心	● 营销云	● Data Cloud

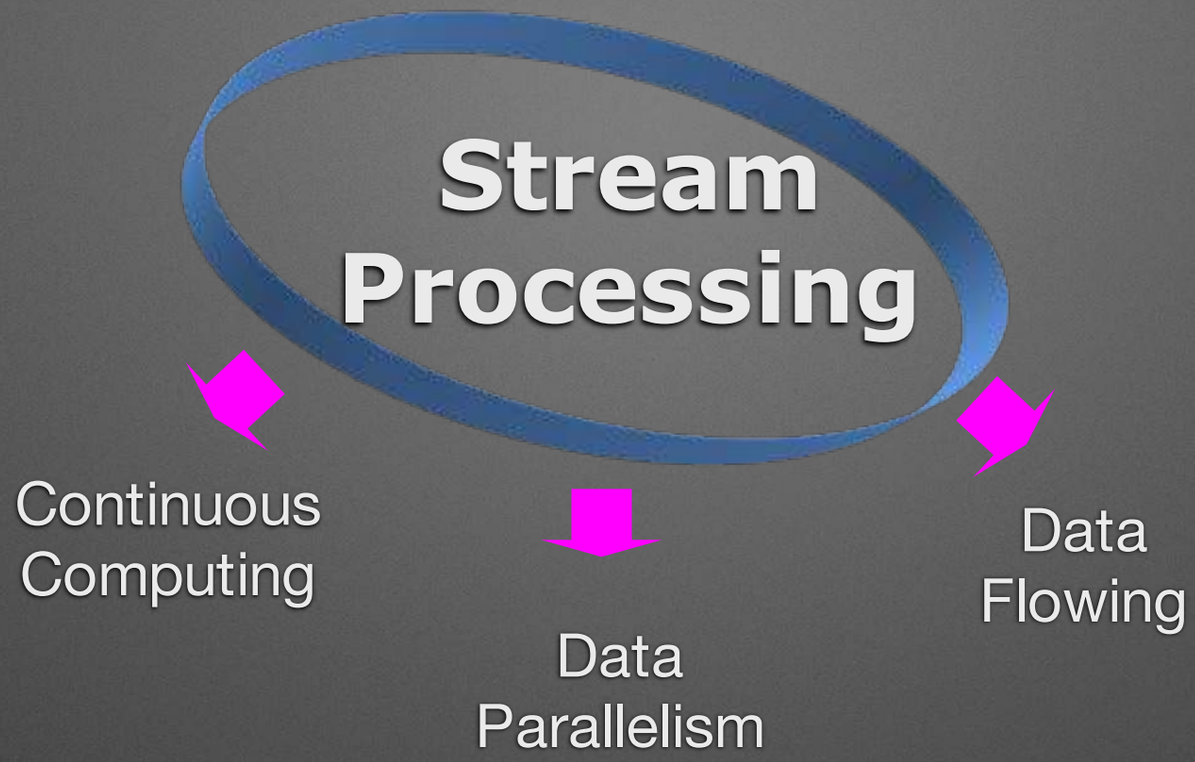
About TalkingData

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 Abstration for data system

Event Log

Event Sourcing

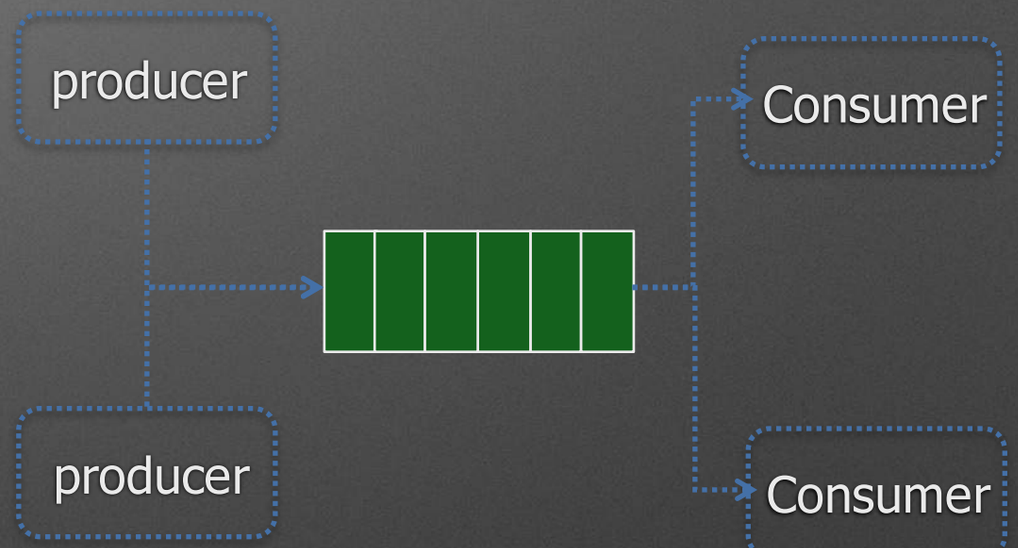
WAL

Command-Query Responsibility Segregation

Replication

Message Queue

- Message queues are the Core Integration tools
  - Decoupling
  - Parrallelism
  - Reliability

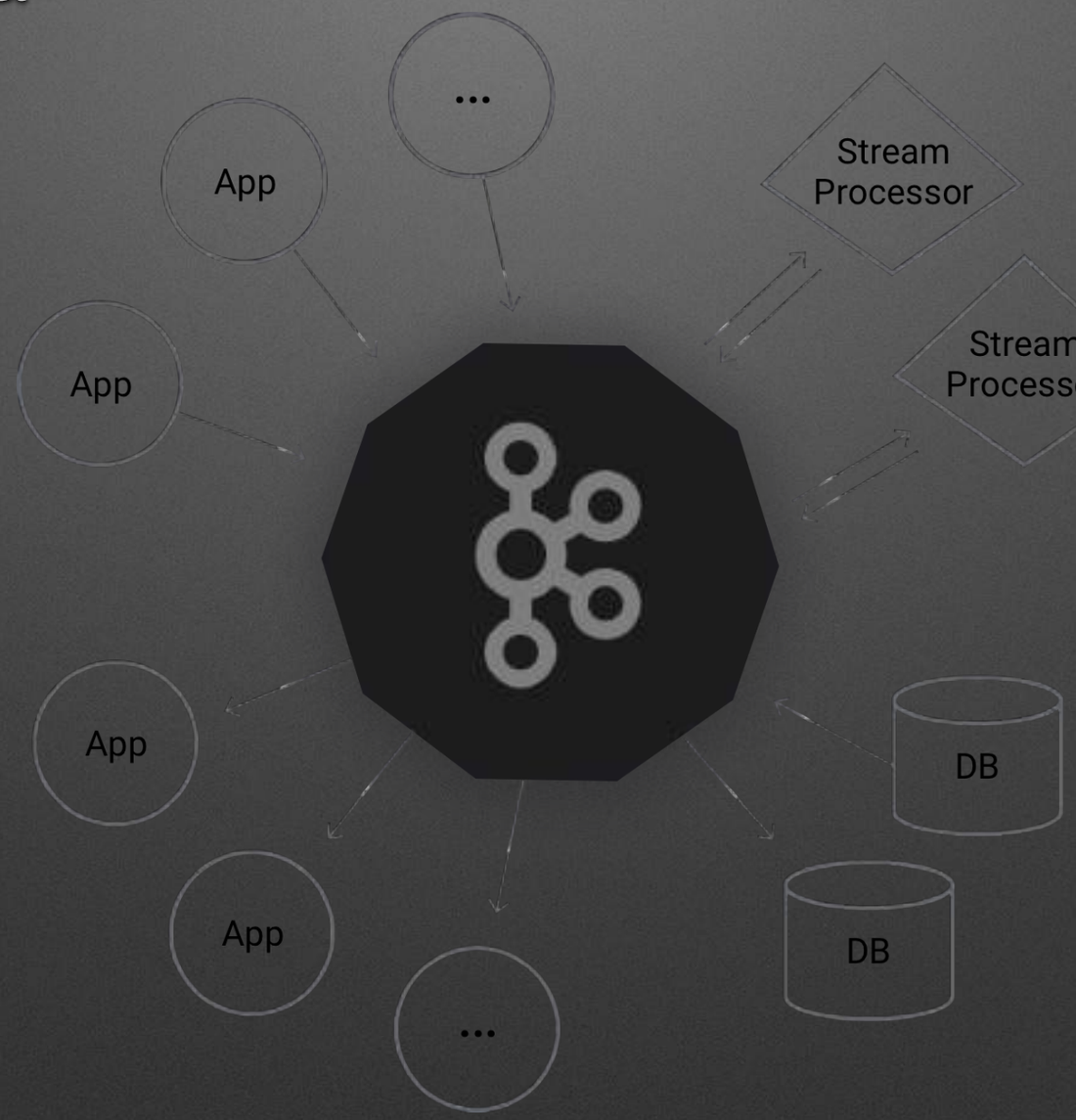


Message Queue

- At most once
- At least once
- Exactly once

Message Queue

- Apache Kafka



Stream Computing

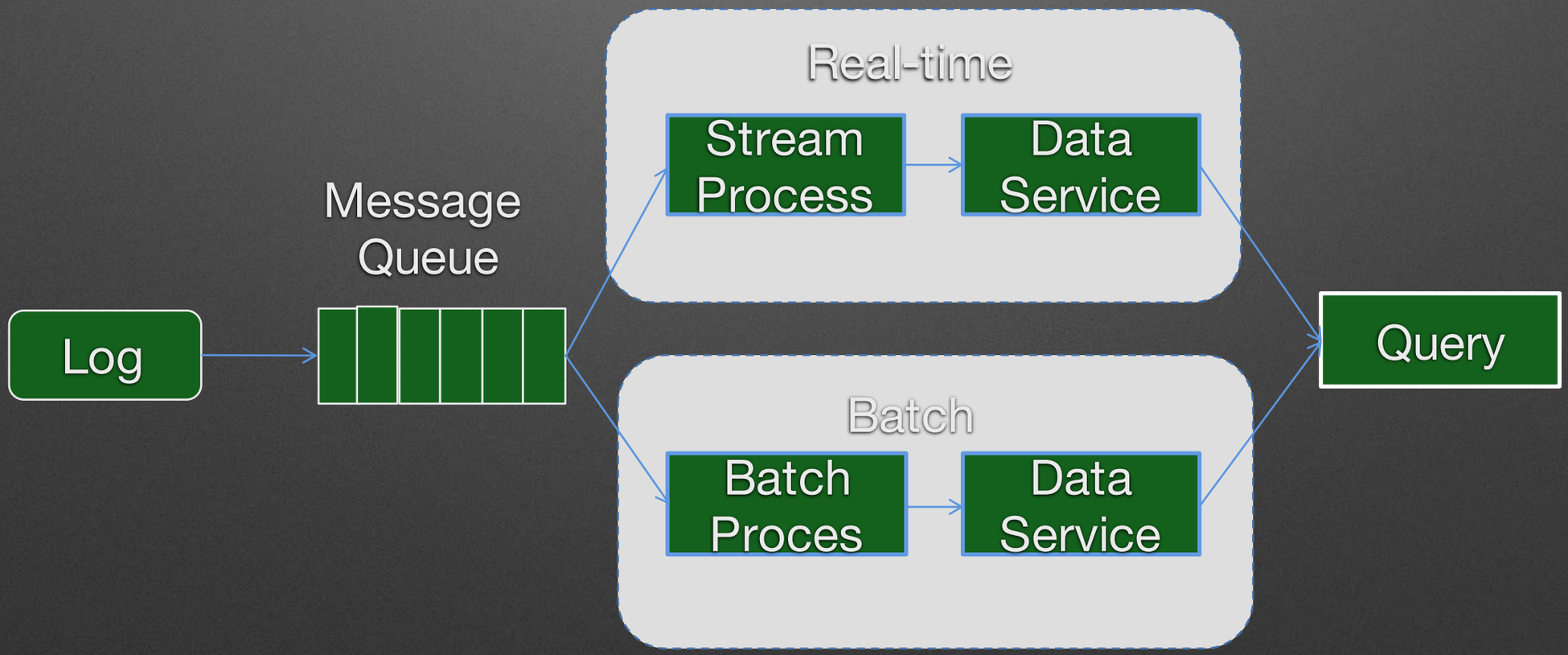
- Stream computing platforms are Core Data Processing tools
  - Apache Spark
  - Apache Flink
  - Apache Storm
  - Apache Apex
  - Apache Beam

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

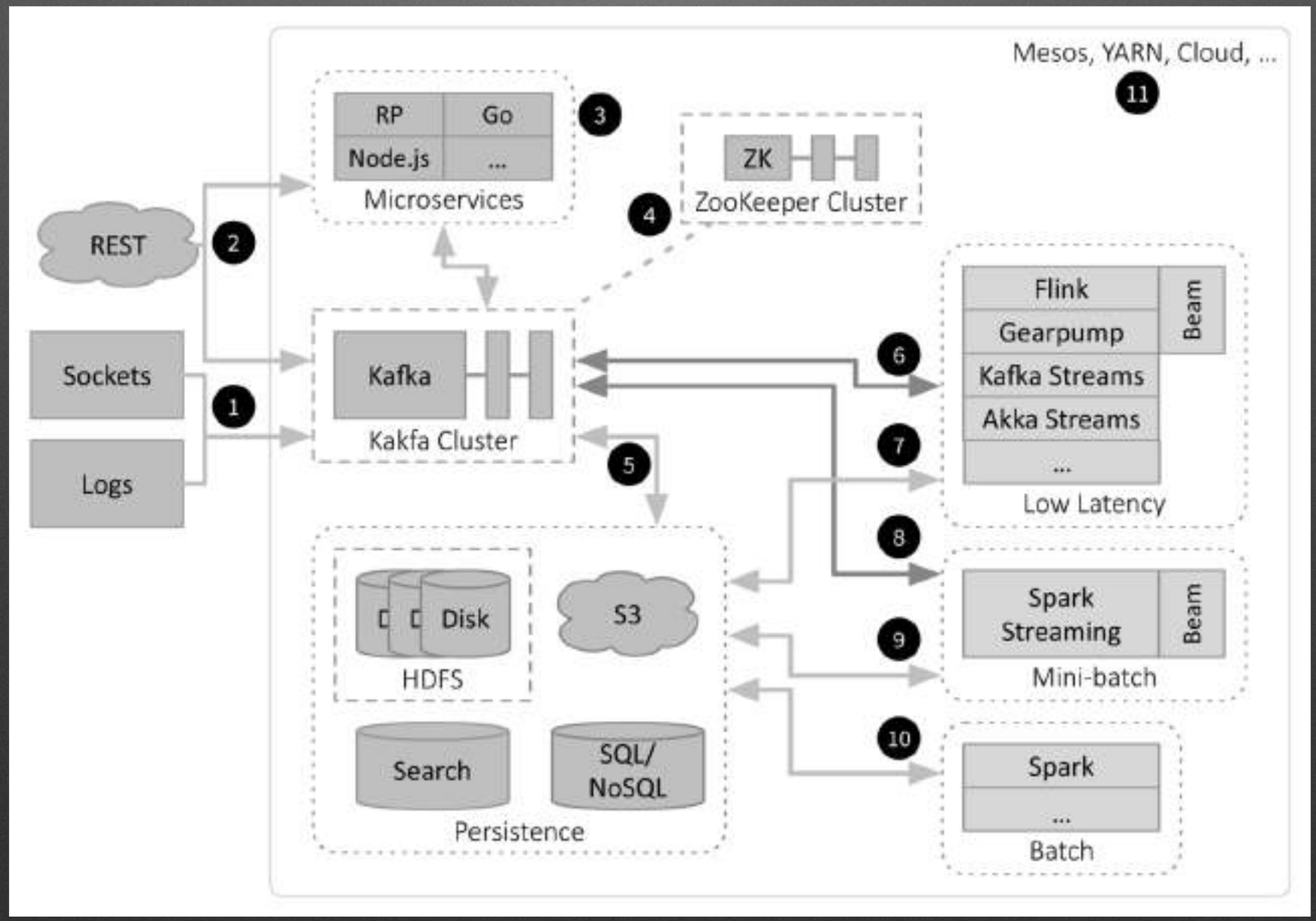


Reference Architecture

- Kappa Architecture

Reference Architecture

- Comprehensive Architecture

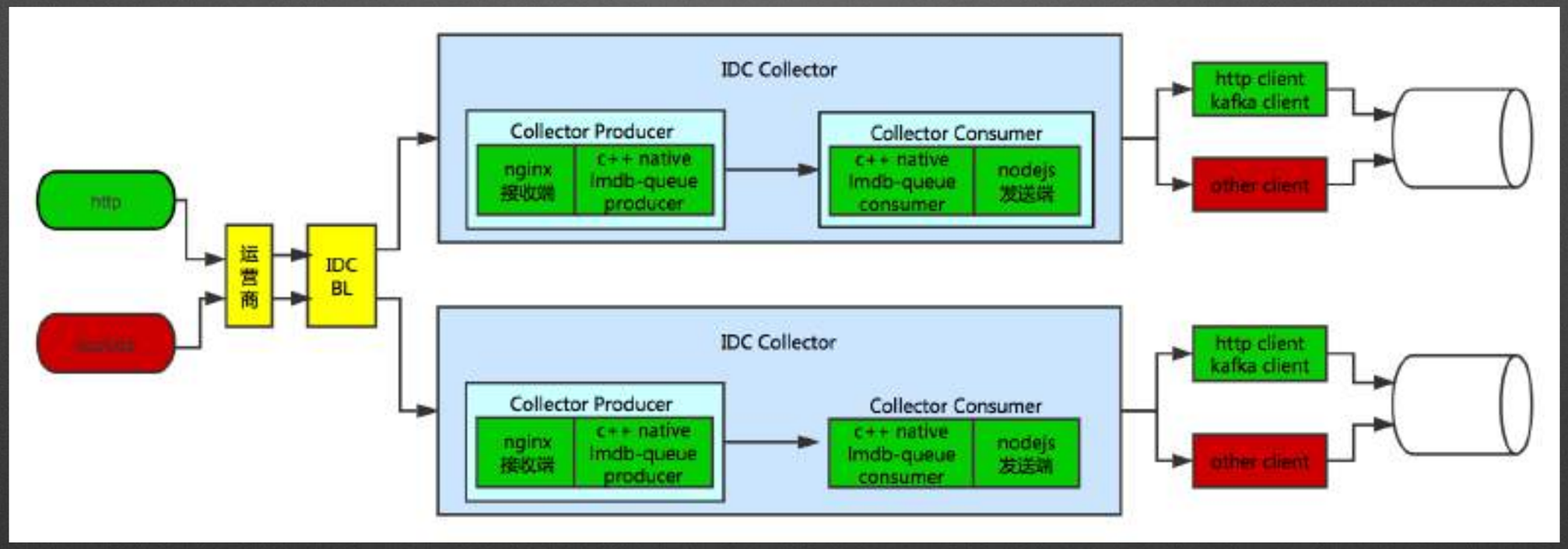


Stream Processing At TalkingData

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

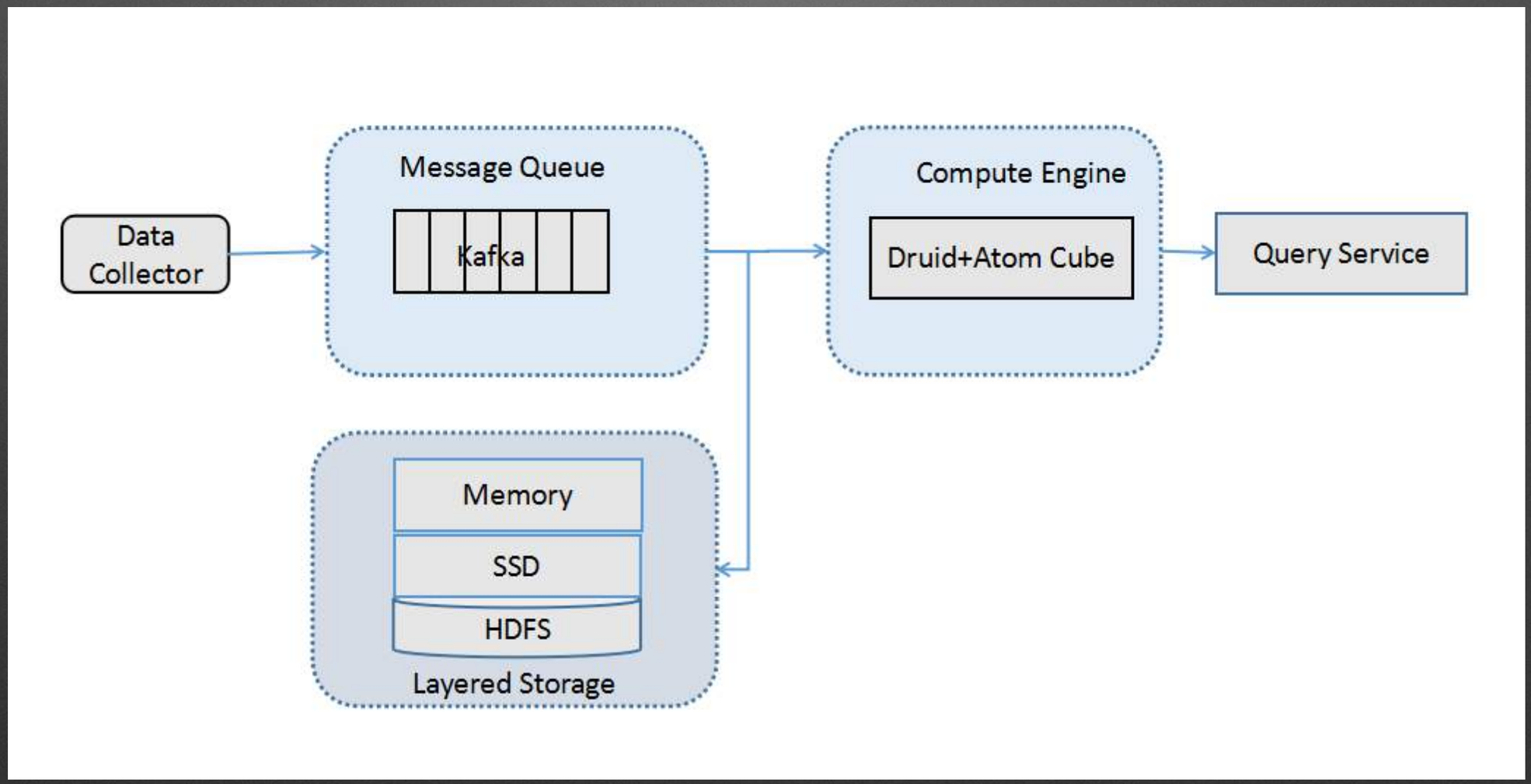
Stream Data Collection

- Data flows as stream



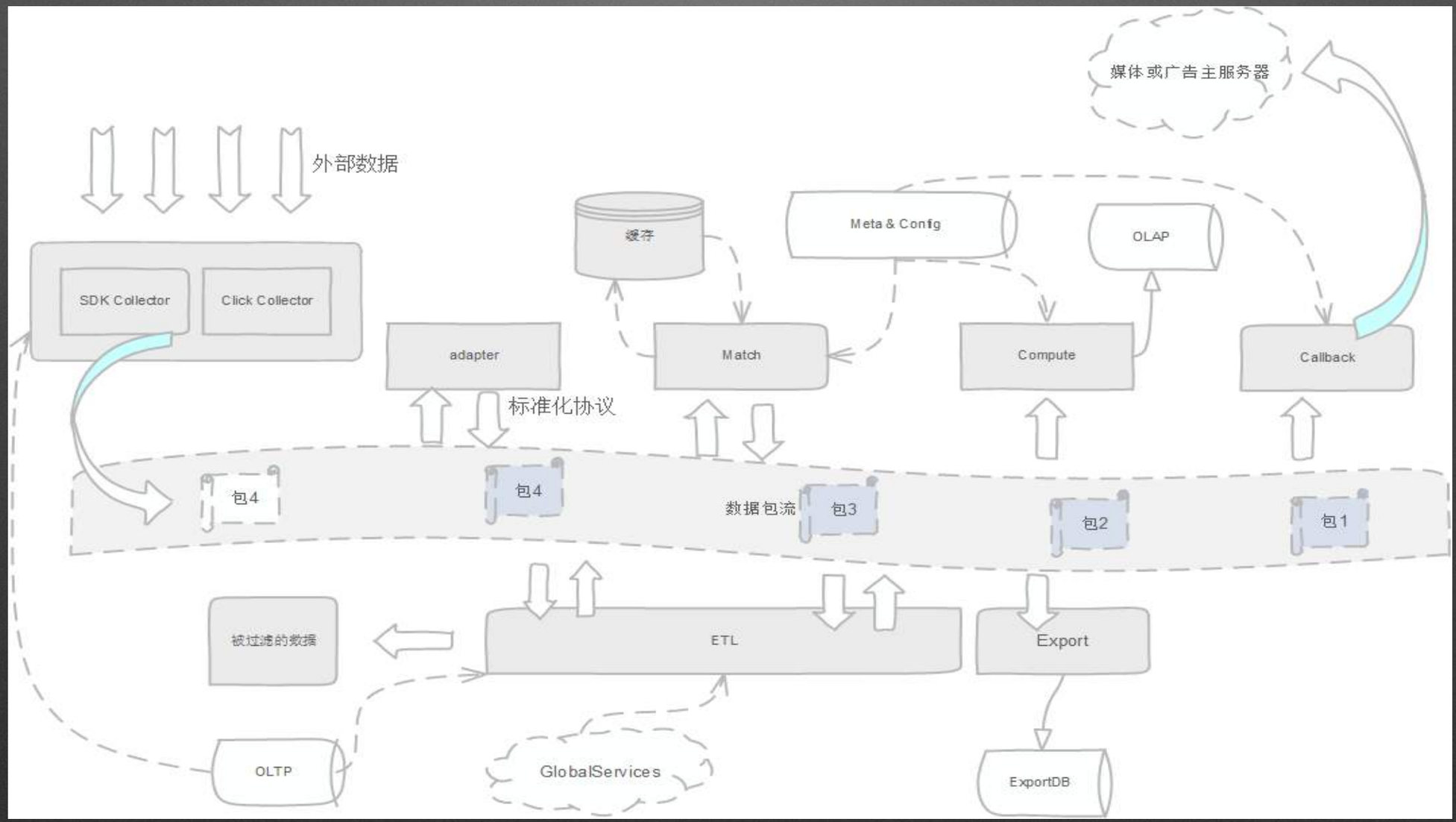
Stream Analytics

- TalkingData analytics architecture



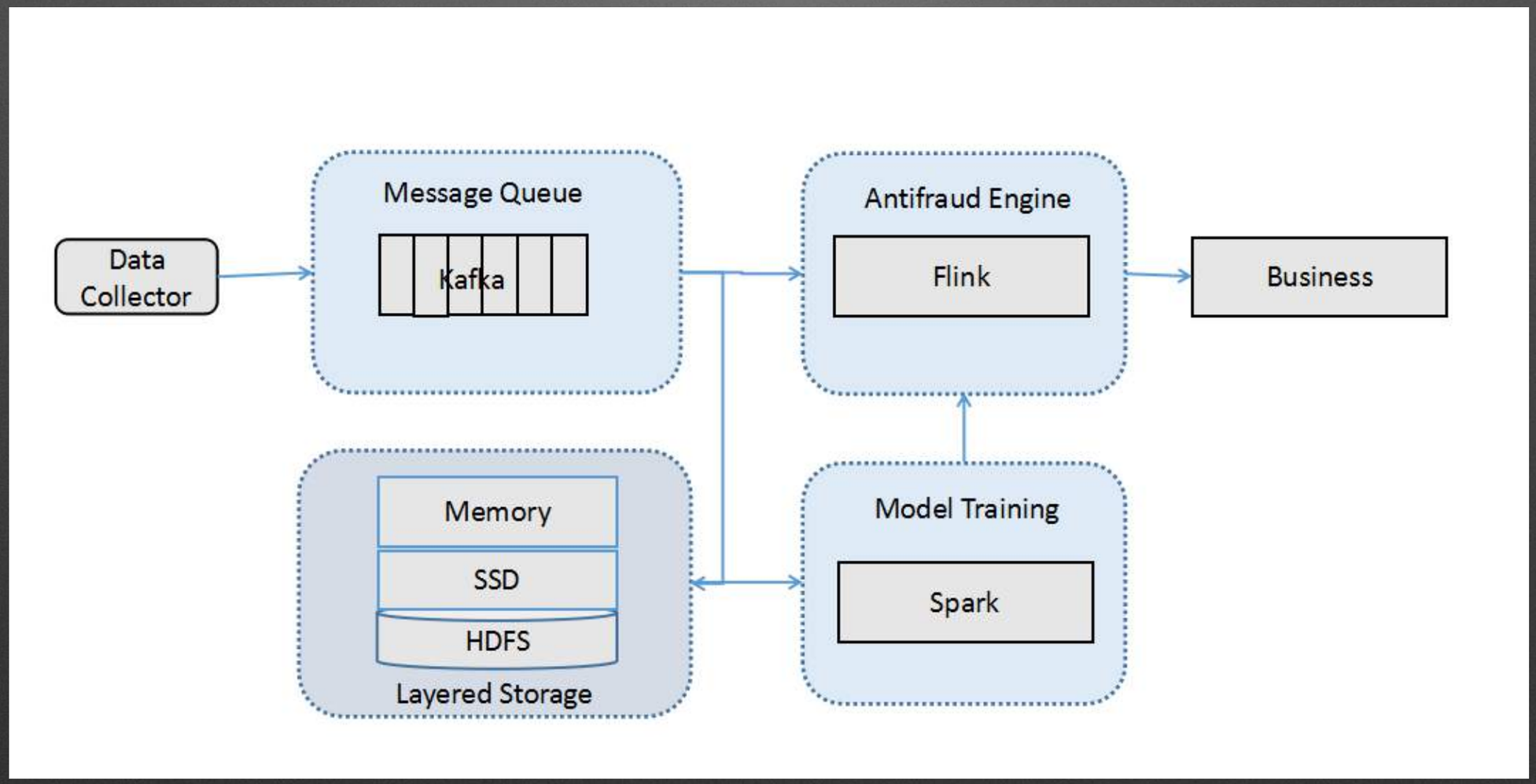
Stream Attribution

- Advertisment 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