Spark Streaming in Log Processing System

Shen Zhun
shenzhunallen@gmail.com

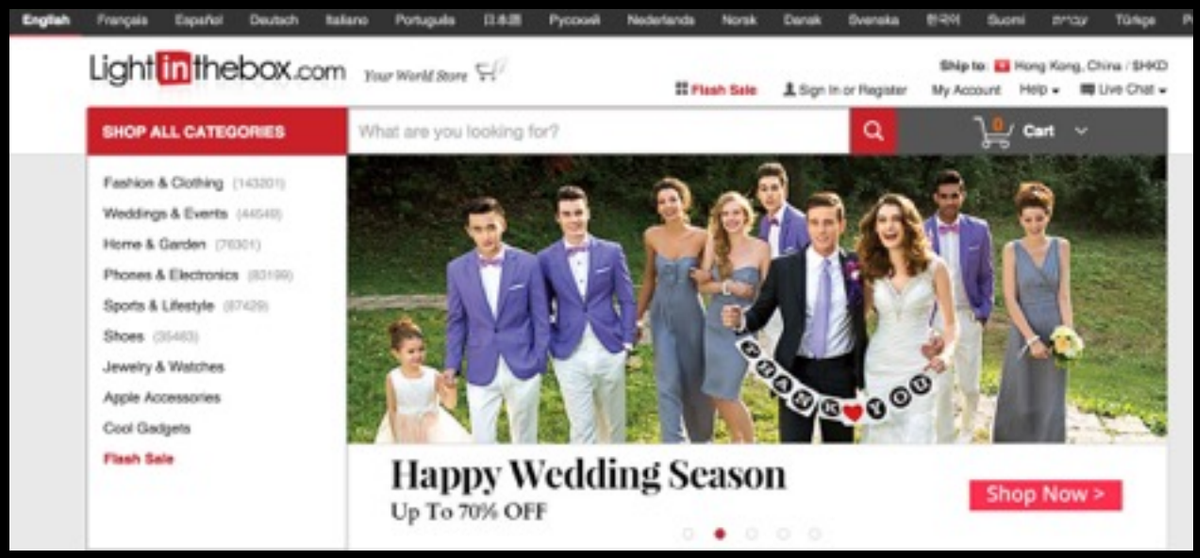
God made Heaven and Earth, and the rest was made in China

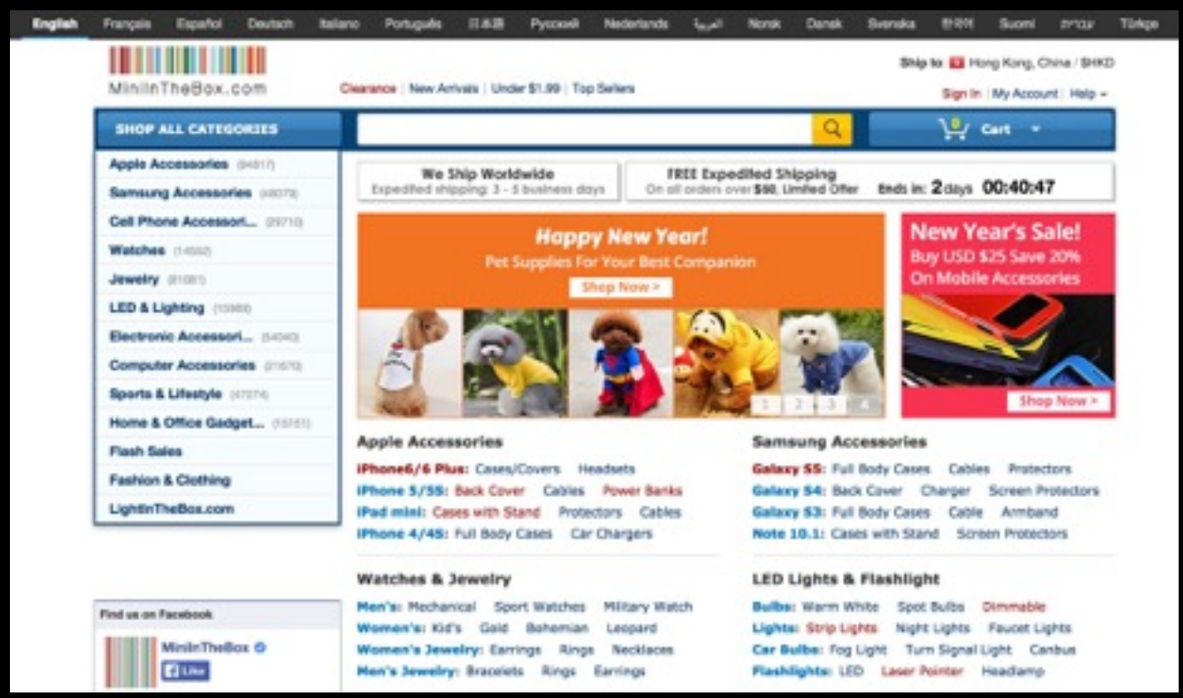
-Unknown

What's LightInTheBox

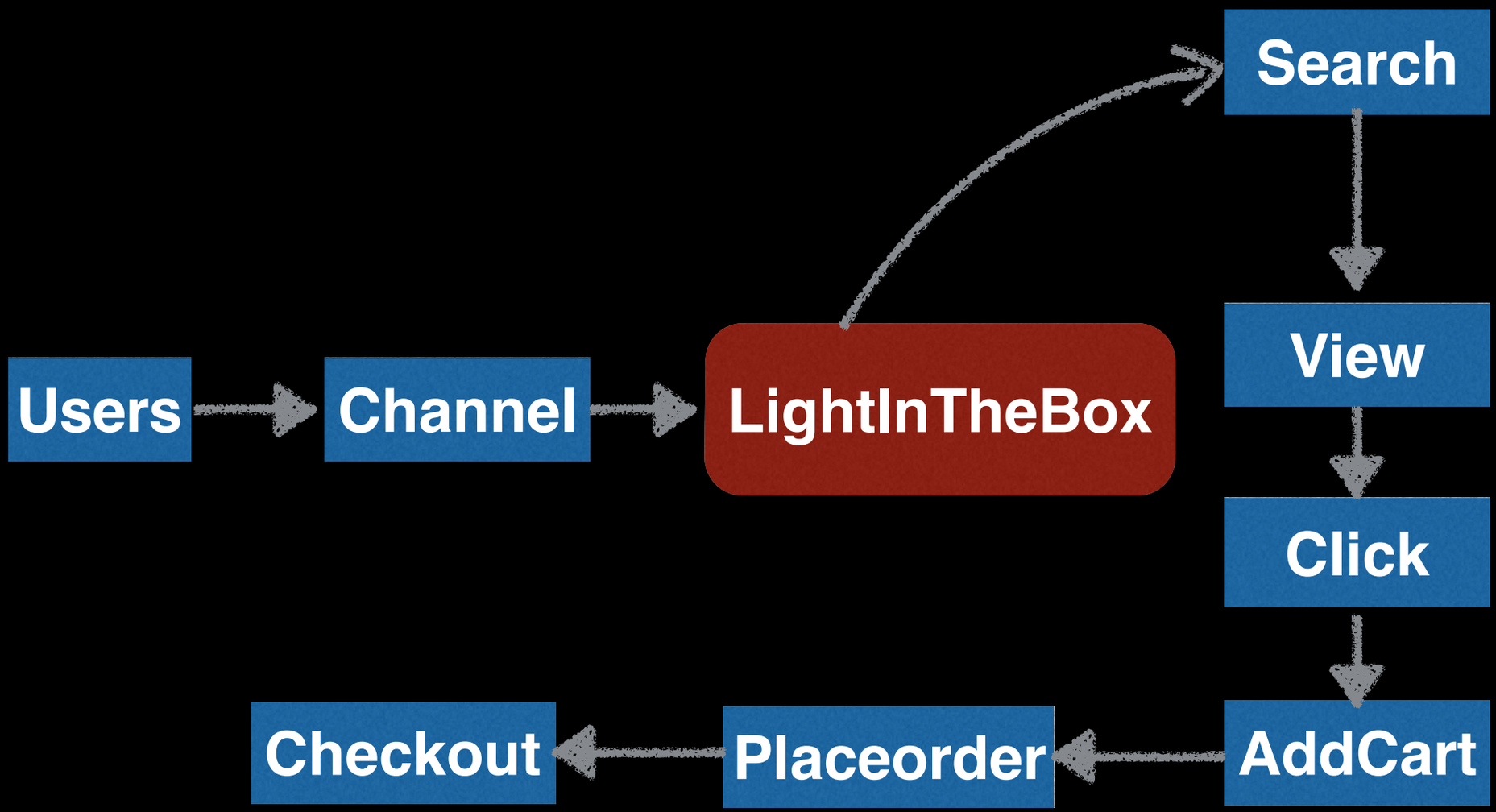
Founded in 2007

IPO in 2013

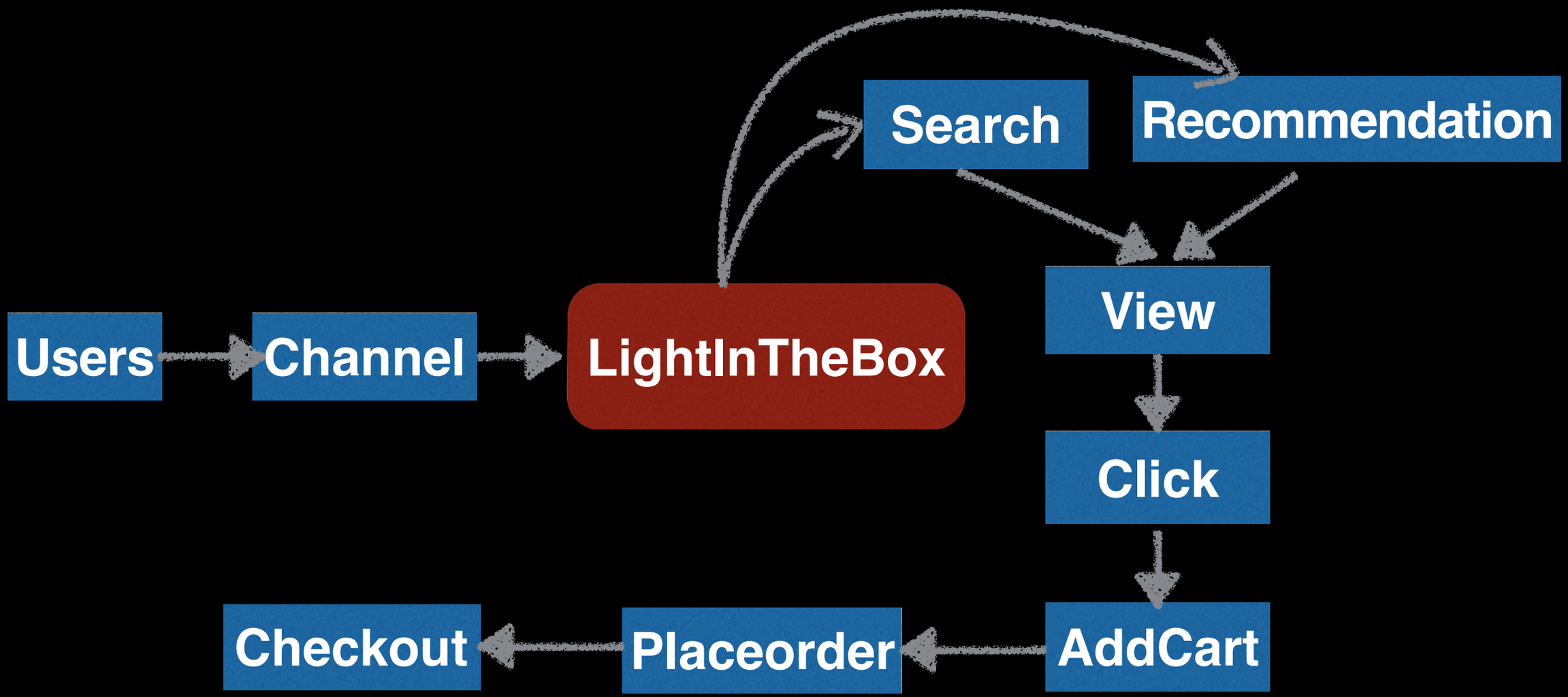




Users' Behavior(1)



Users' Behavior(2)



ClickStream Log

- Large volume log
- Users and bot mixed
- Analysis based on UV and PV
- 24x7 monitering

Technology Stack

- AWS
- Flume
- Kafka
- Zookeeper
- Apache Spark Stack (Streaming, SQL , MLlib)
- ElasticSearch
- Kibana

Processing Models

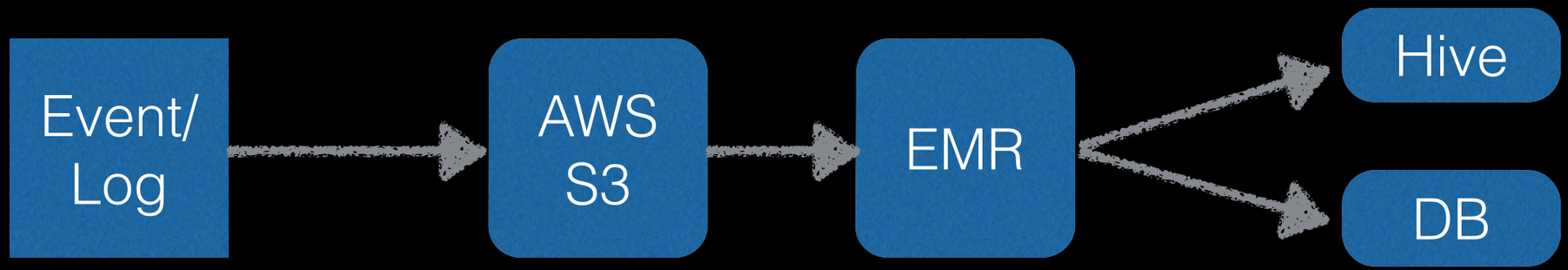
- Batch Processing
- Streaming Processing
- Micro-Batching

Batch vs Streaming

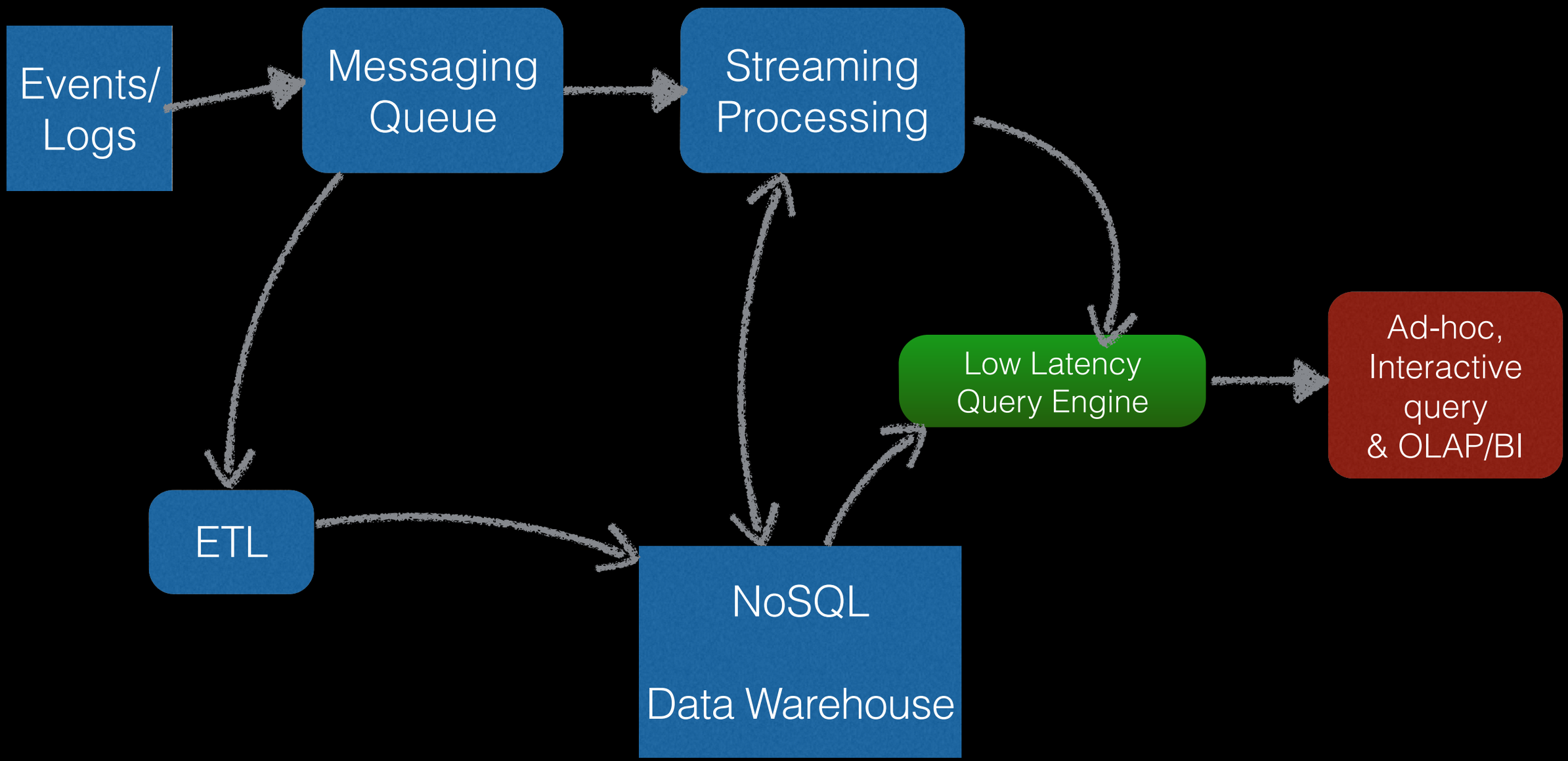
Daily reporting
Orders/Sales
Anything with strict SLA
Correctness > low latency

“Real-Time” reporting
Low latency to use data
Only reliable as source
Low latency > correctness

Batch Processing



New Data Pipeline



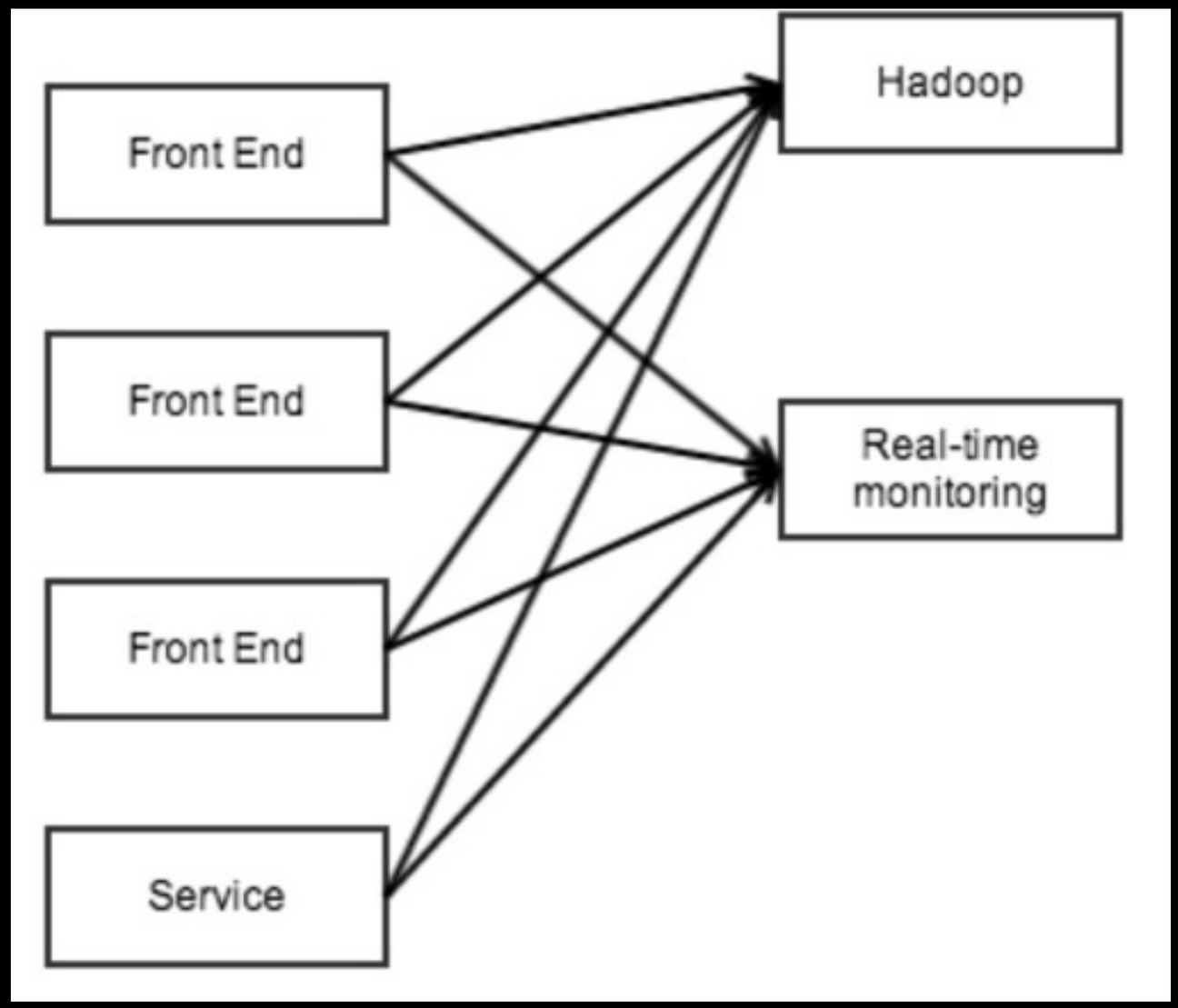
Why Spark Streaming(1)

- Liked theoretical foundation of mini-batch
- Scala codebase + functional API
- Batch model for iterative ML algorithms
- Scala + Java
- Young project with opportunities to contribute

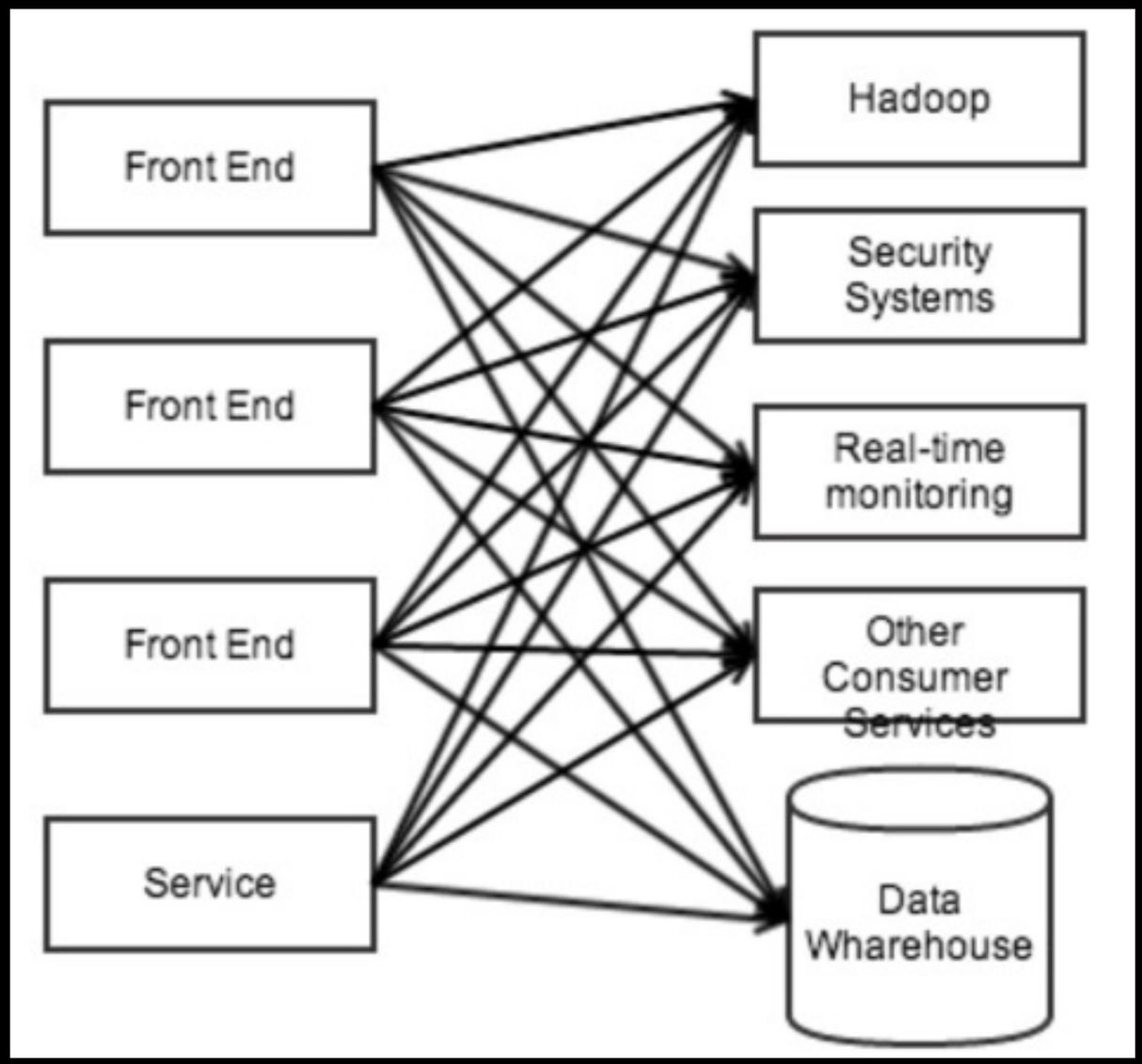
Why Spark Streaming(2)

- HA
- Zero Data Loss
- Fault Tolerance

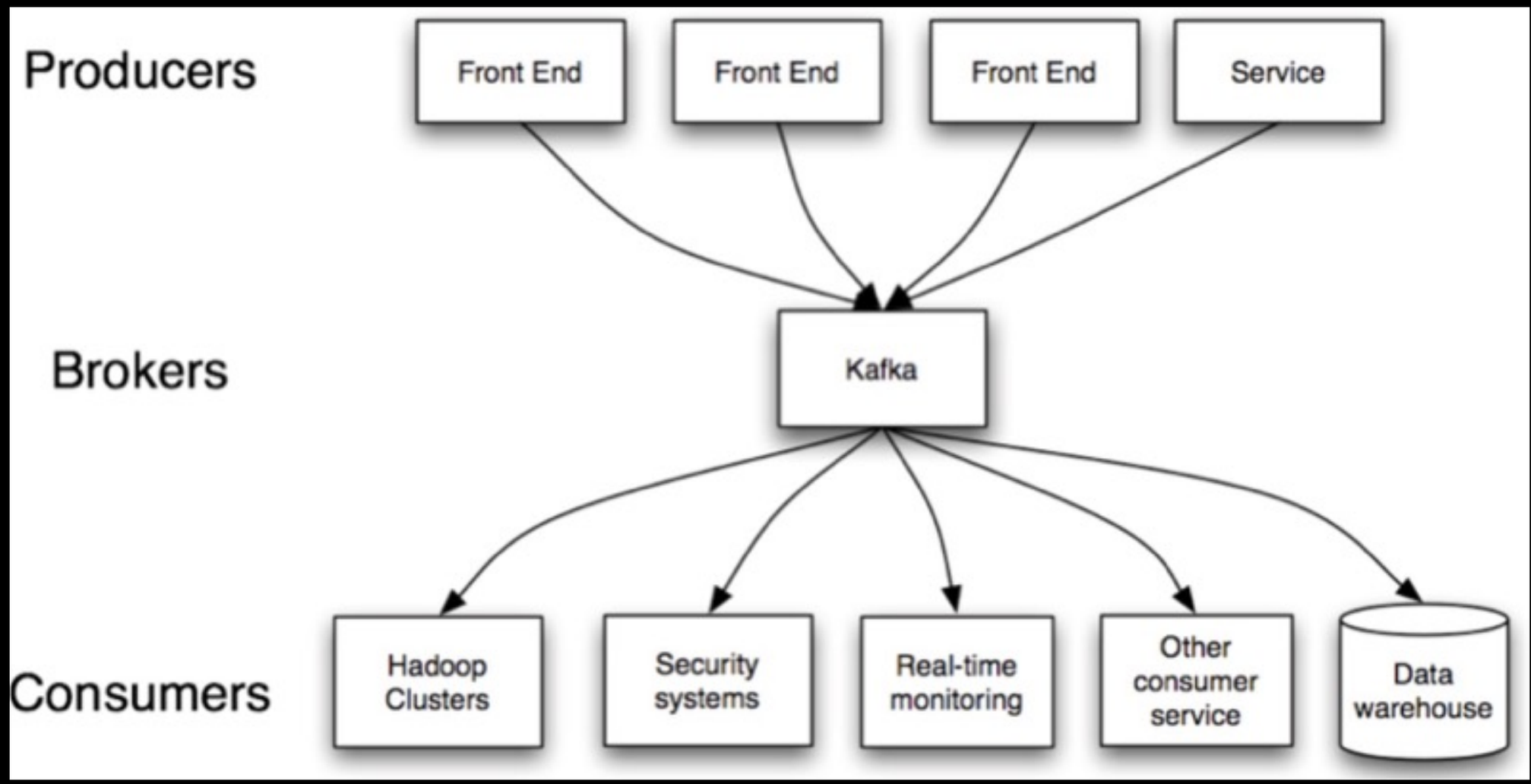
Why Kafka(1)


Front End
Front End
Front End
Service
Hadoop
Real-time
monitoring

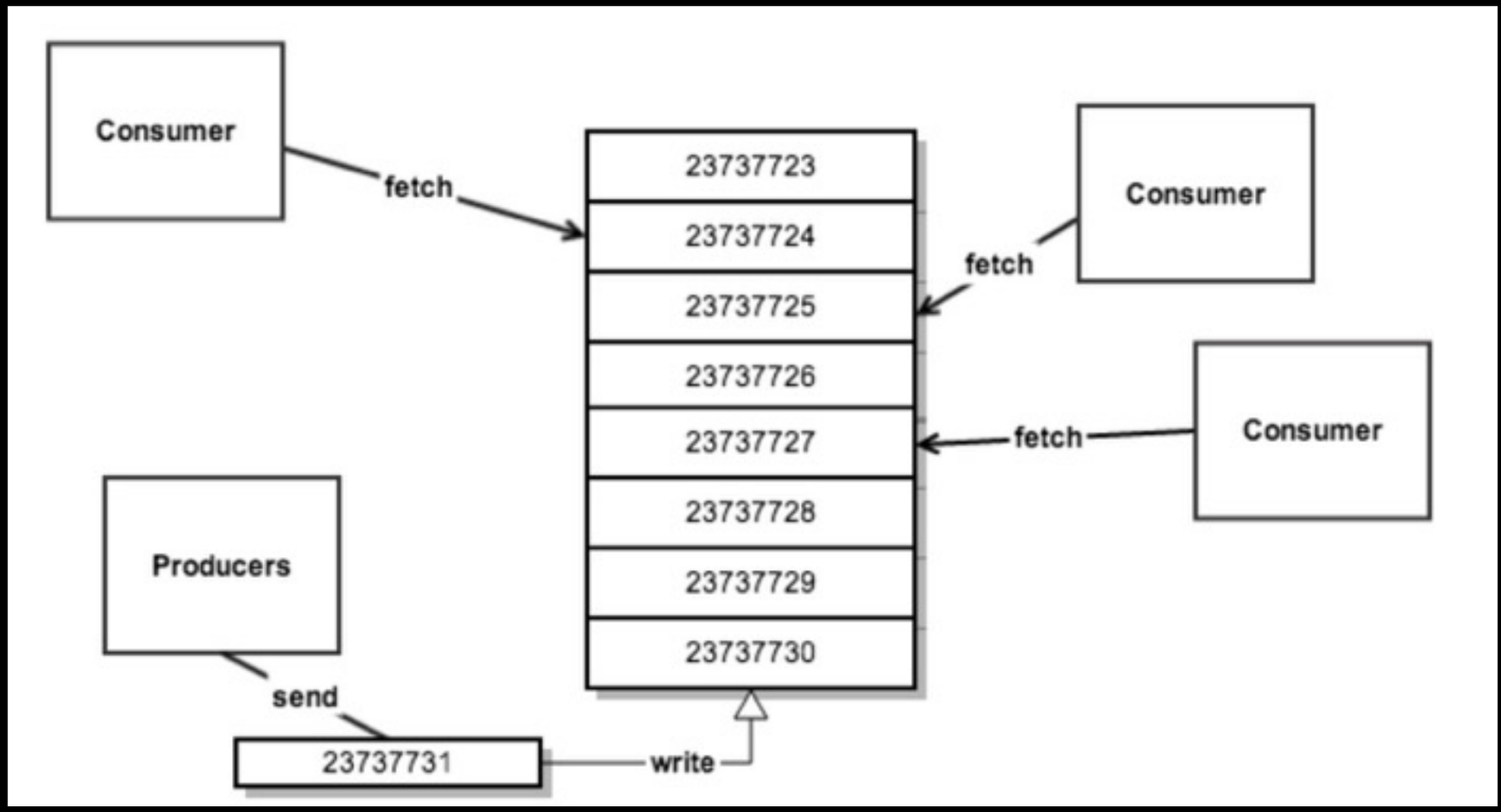
Why Kafka(2)


Front End
Front End
Front End
Service
Hadoop
Security Systems
Real-time monitoring
Other Consumer Services
Data Wharehouse

Why Kafka(3)

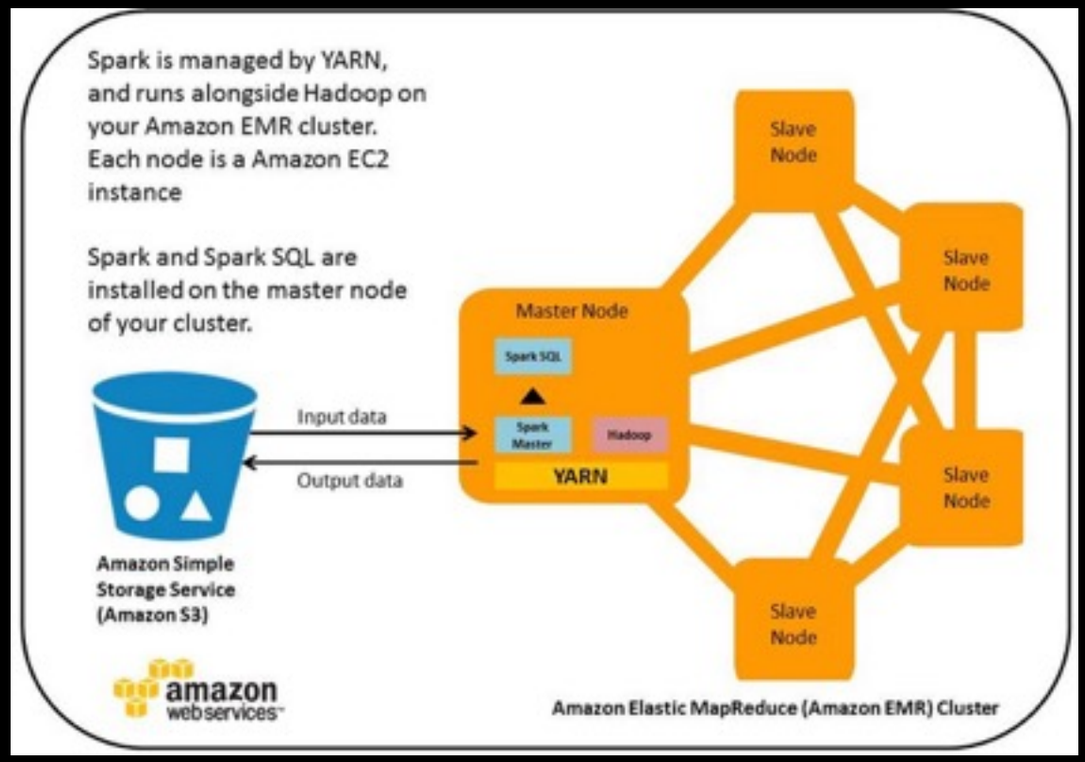

Producers
Front End
Front End
Front End
Service
Brokers
Kafka
Consumers
Hadoop Clusters
Security systems
Real-time monitoring
Other consumer service
Data warehouse

Why Kafka(4)


Consumer
fetch
23737723
23737724
23737725
23737726
23737727
23737728
23737729
23737730
Consumer
fetch
Consumer
fetch
Producers
send
23737731
write

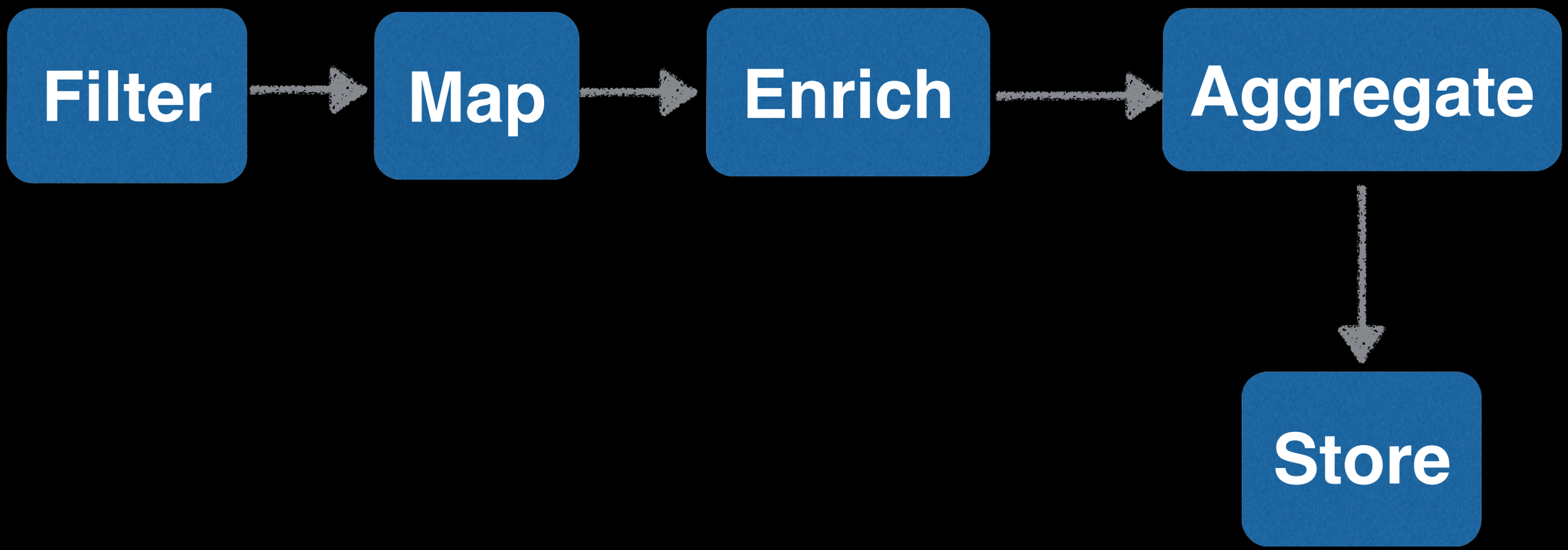
Amazon Web Service

EC2
S3
EMR
Spark with YARN

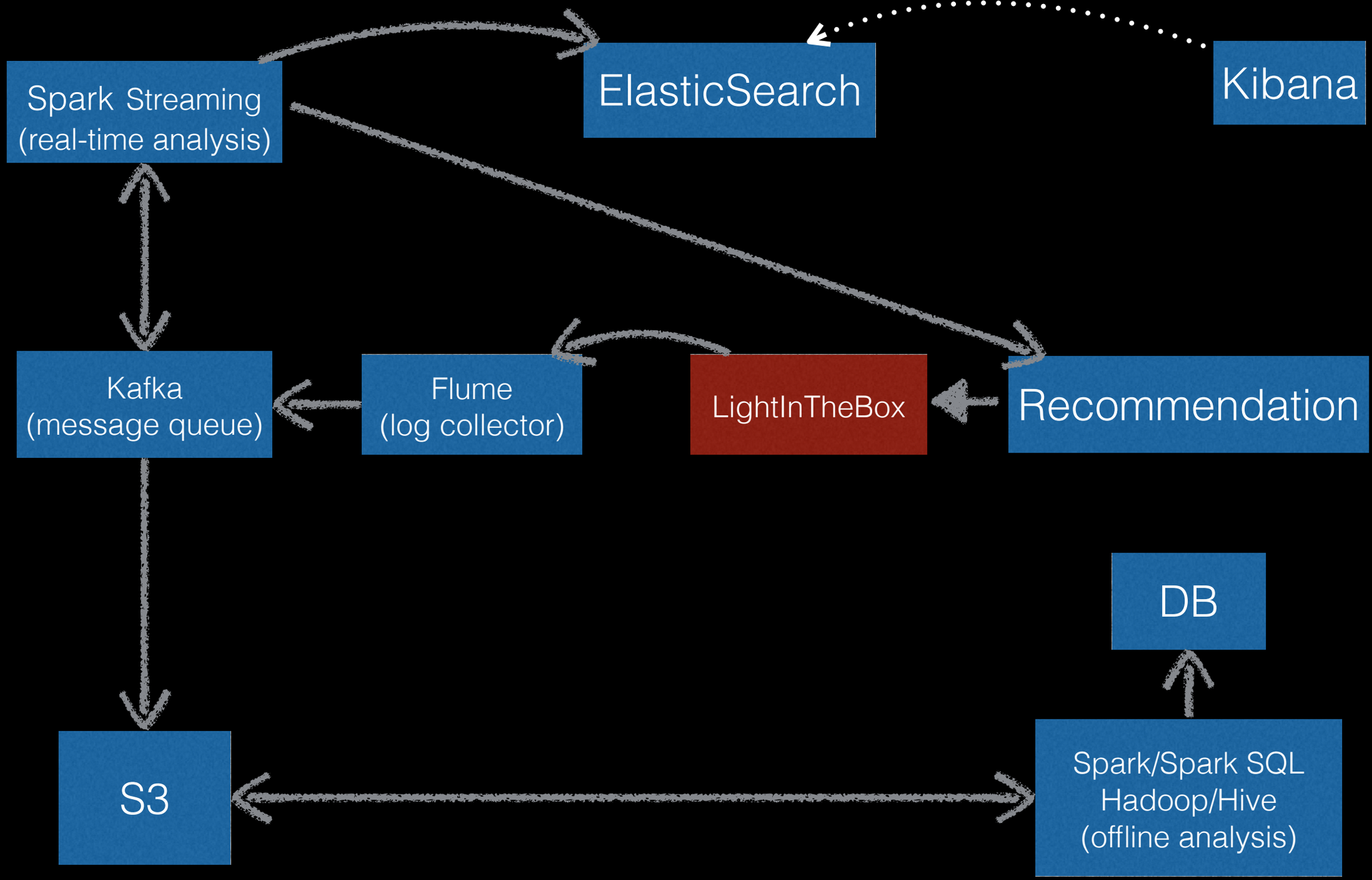


https://aws.amazon.com/articles/4926593393724923

Common Job Pattern



Final Design



Q & A