Alibaba Group 阿里巴巴集团

Blink:阿里新一代实时计算引擎

马国维

2017.4

Who am I?

✓ 2010 – 2017 Alibaba Search

- iStream
- Blink

✓ 2007 – 2010 Baidu Web Search

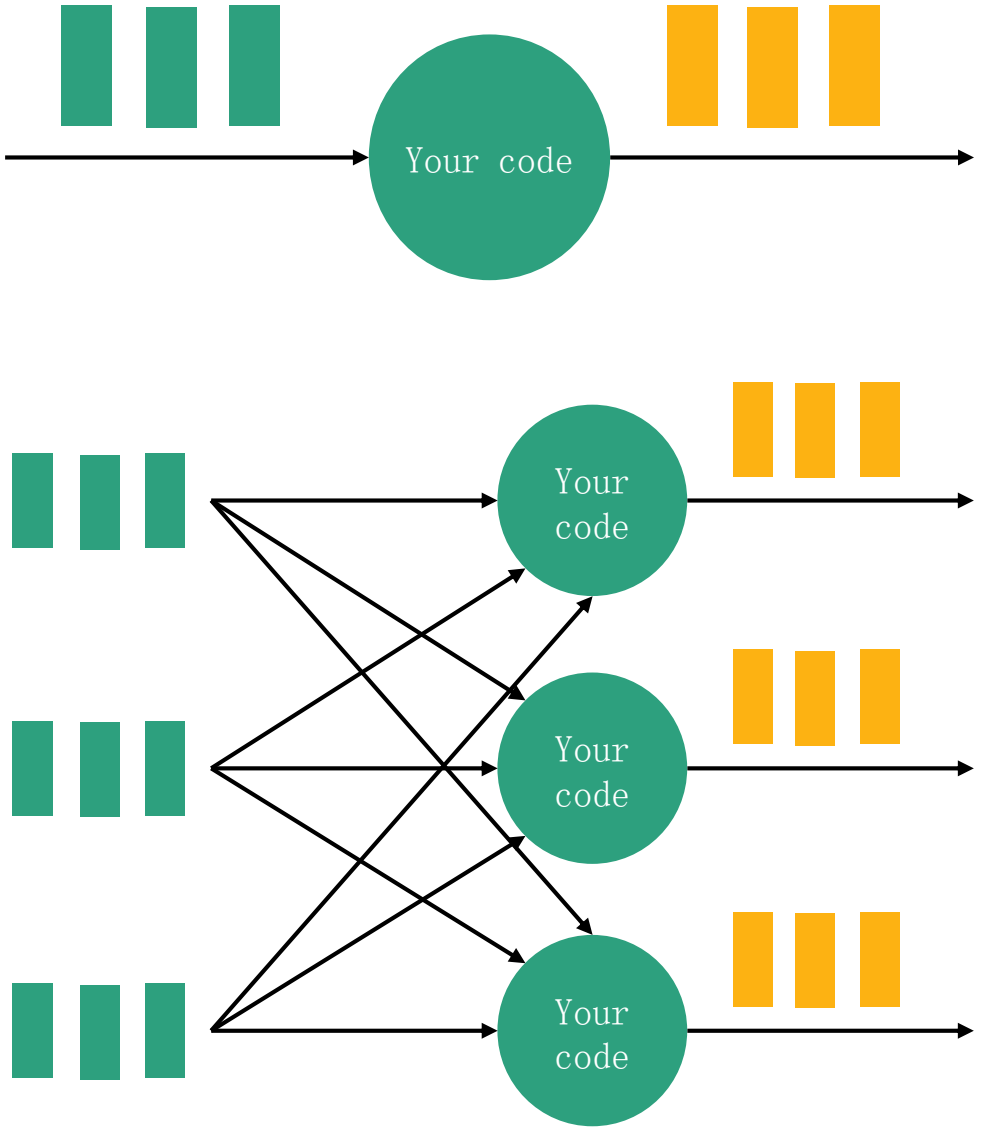
Outline

Alibaba Group
阿里巴巴集团

The streaming architecture

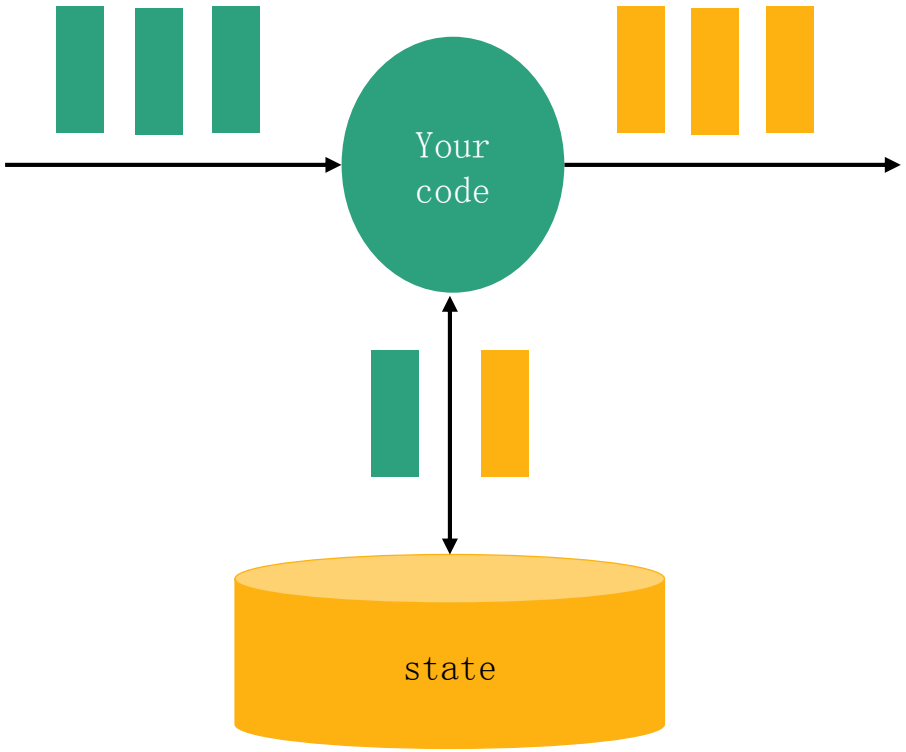
Part I

What is streaming?



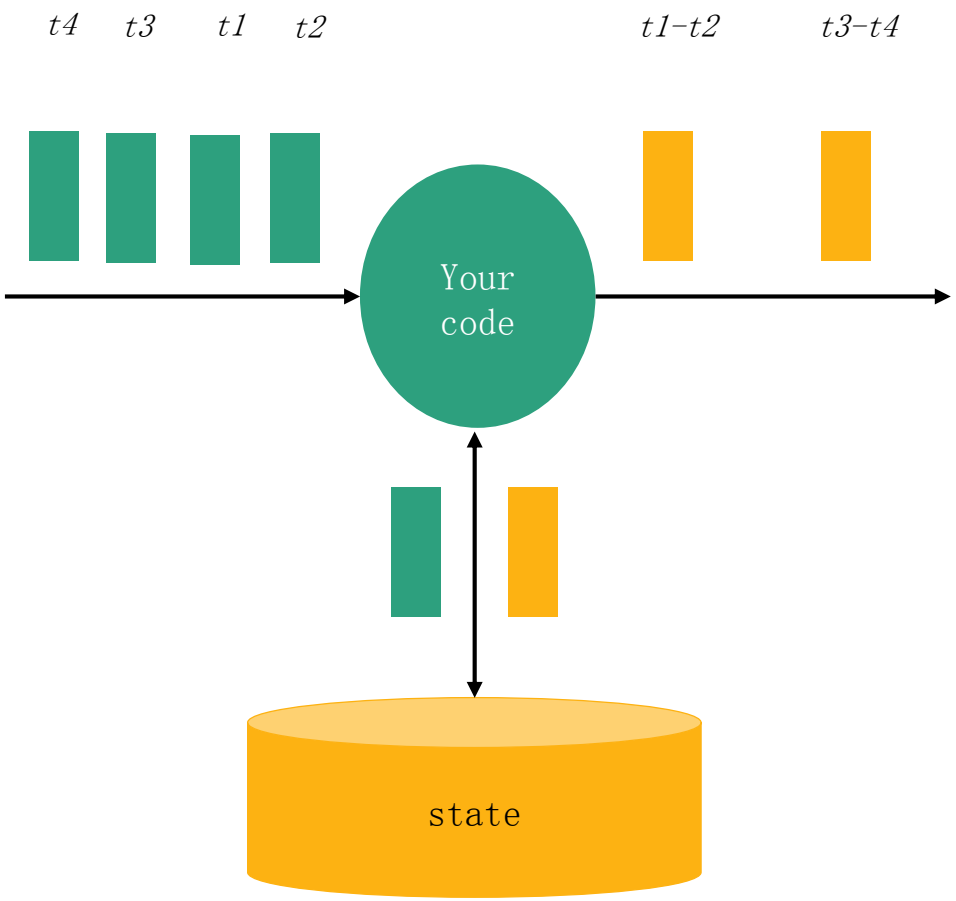
- ✓What is streaming?
  - Unbounded data
- ✓What is streaming process engine?
  - The data process engine that is designed with infinite data set in mind

What is stateful streaming



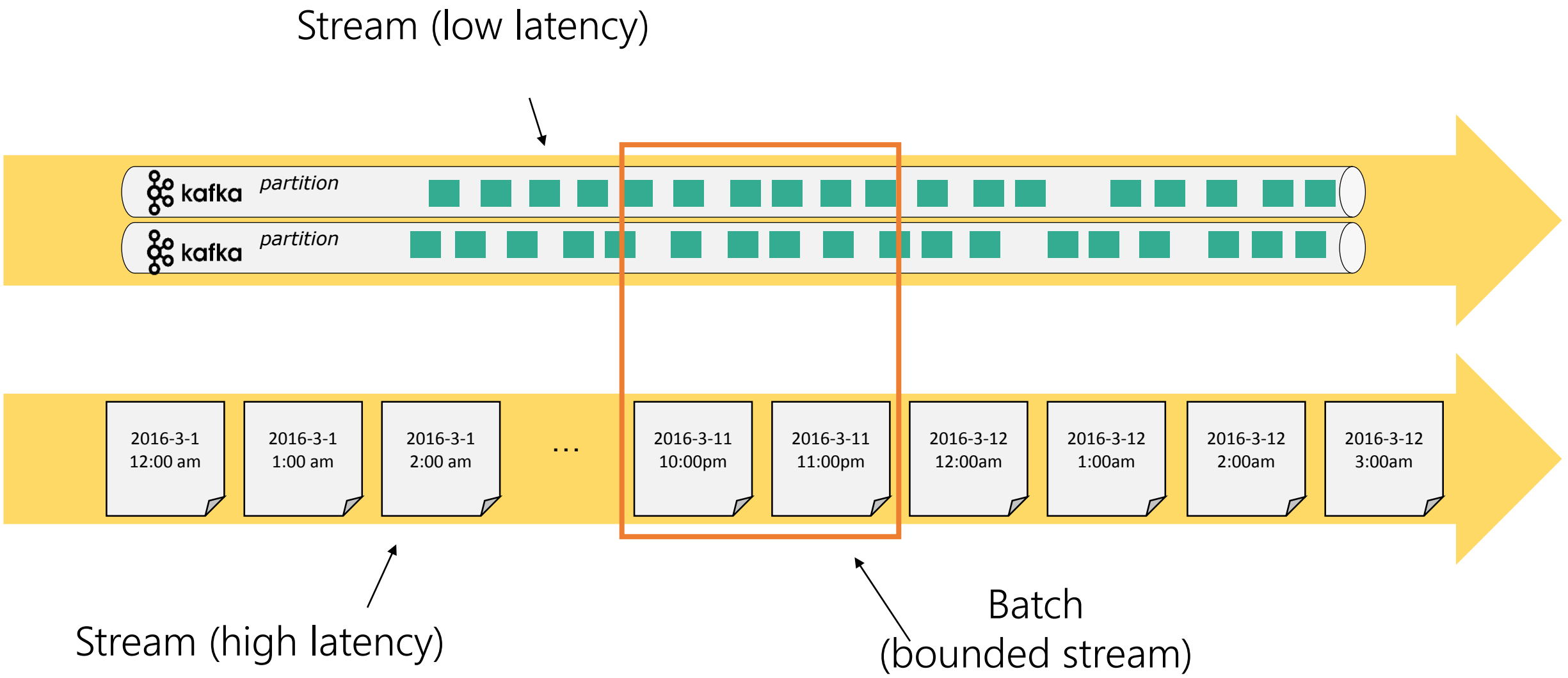
- ✓Computation *and* state
  - E.g., counters, windows of past events, state machines, trained ML models
- ✓Result depends on history of stream
- ✓Stateful stream processor gives the tools to manage state
  - Recover, roll back, version, upgrade, etc

What is event-time streaming



- ✓Data records associated with timestamps (time series data)
- ✓Processing depends on timestamps
- ✓Event-time stream processor gives you the tools to reason about time
  - E.g., handle streams that are out of order
  - Core feature is watermarks – a clock to measure event time

Streaming Subsumes Batch



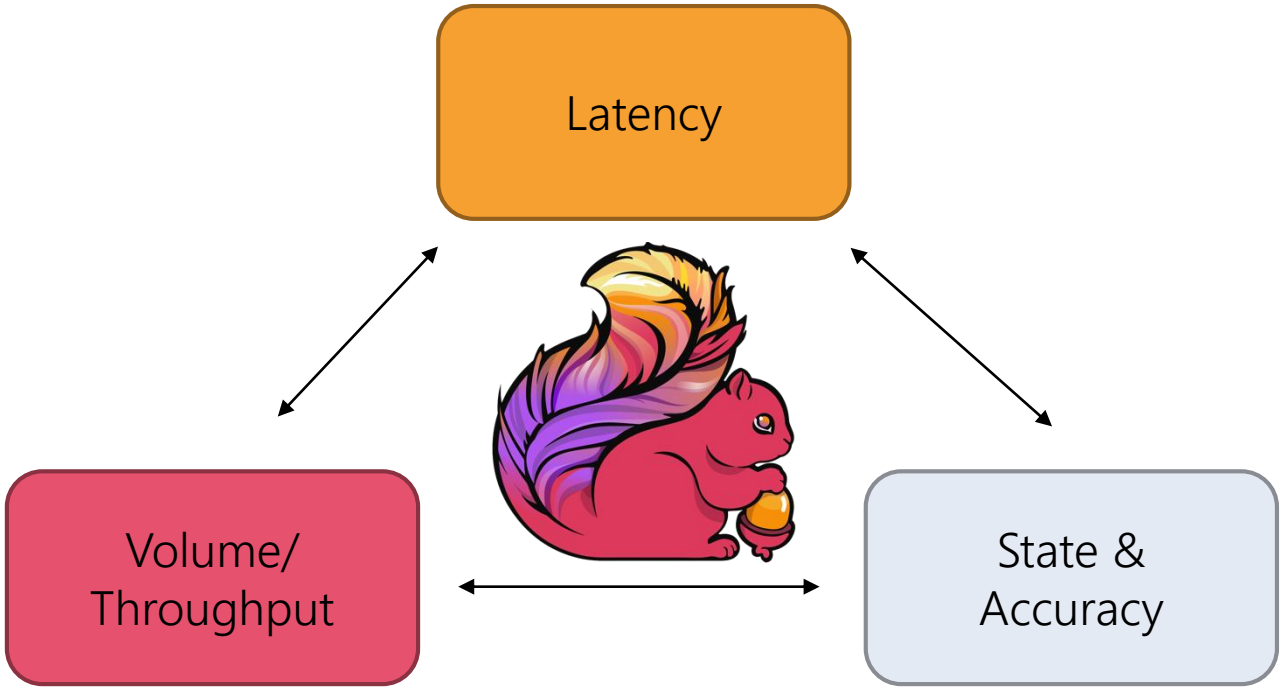
Alibaba Group
阿里巴巴集团

What is Flink?

Part 2

Flink - Streaming Compute Engine

Latency down to the milliseconds



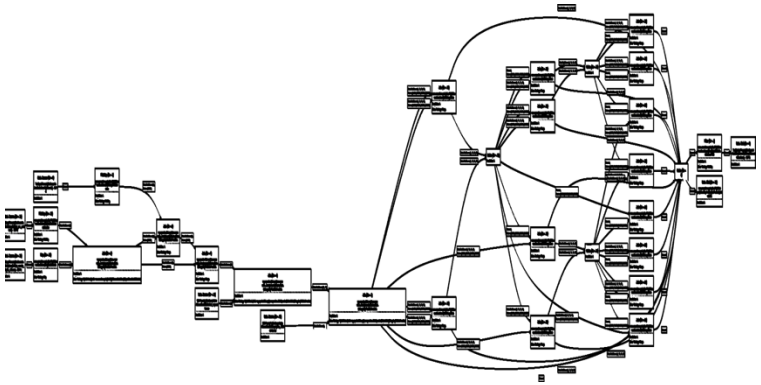
10s of millions evts/sec
for stateful applications

Exactly-once semantics
Event time processing

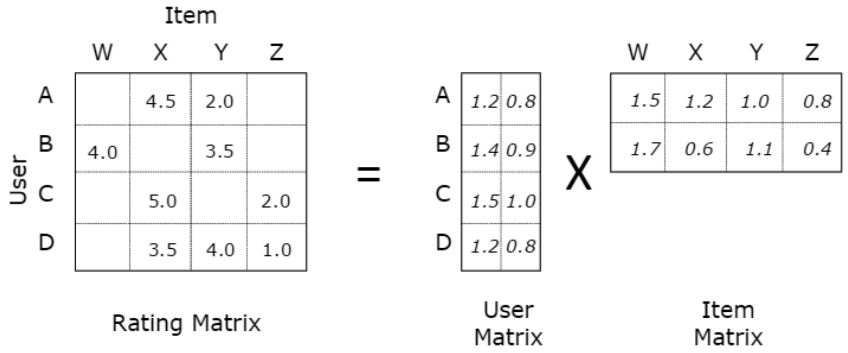
http://flink.apache.org

Flink – Unified Compute Engine

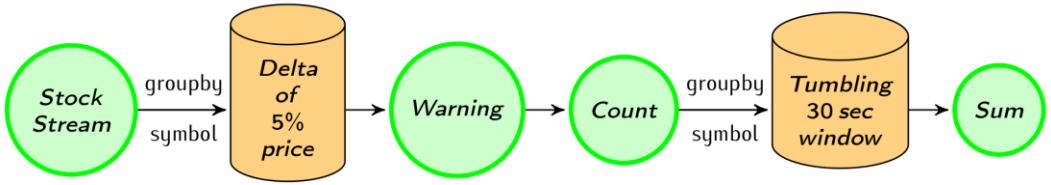
Long batch pipelines



Machine Learning at scale



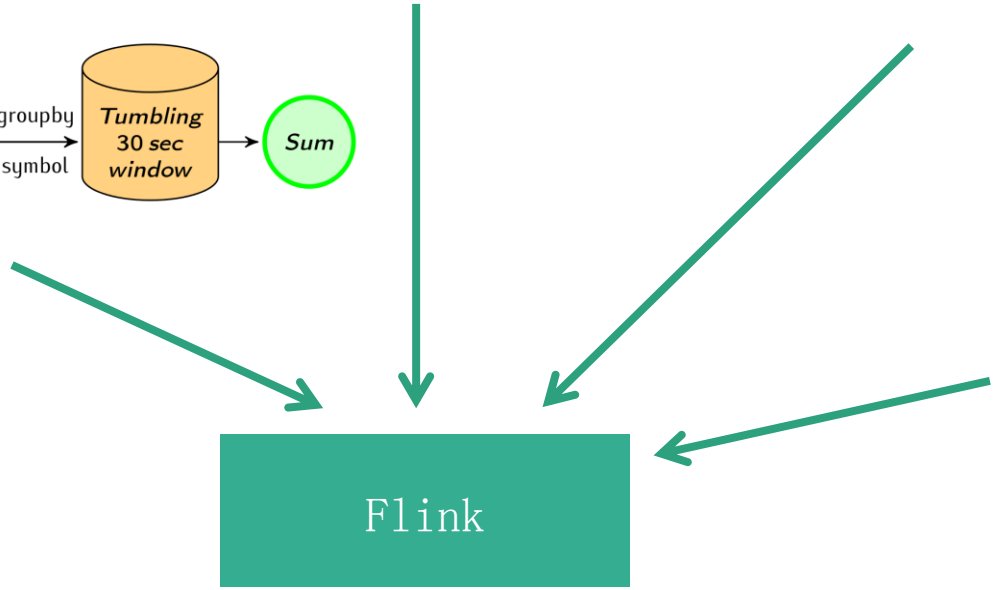
Streaming topologies



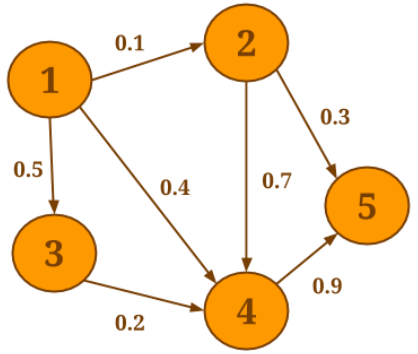
→ Low latency

→ resource utilization

→ iterative algorithms

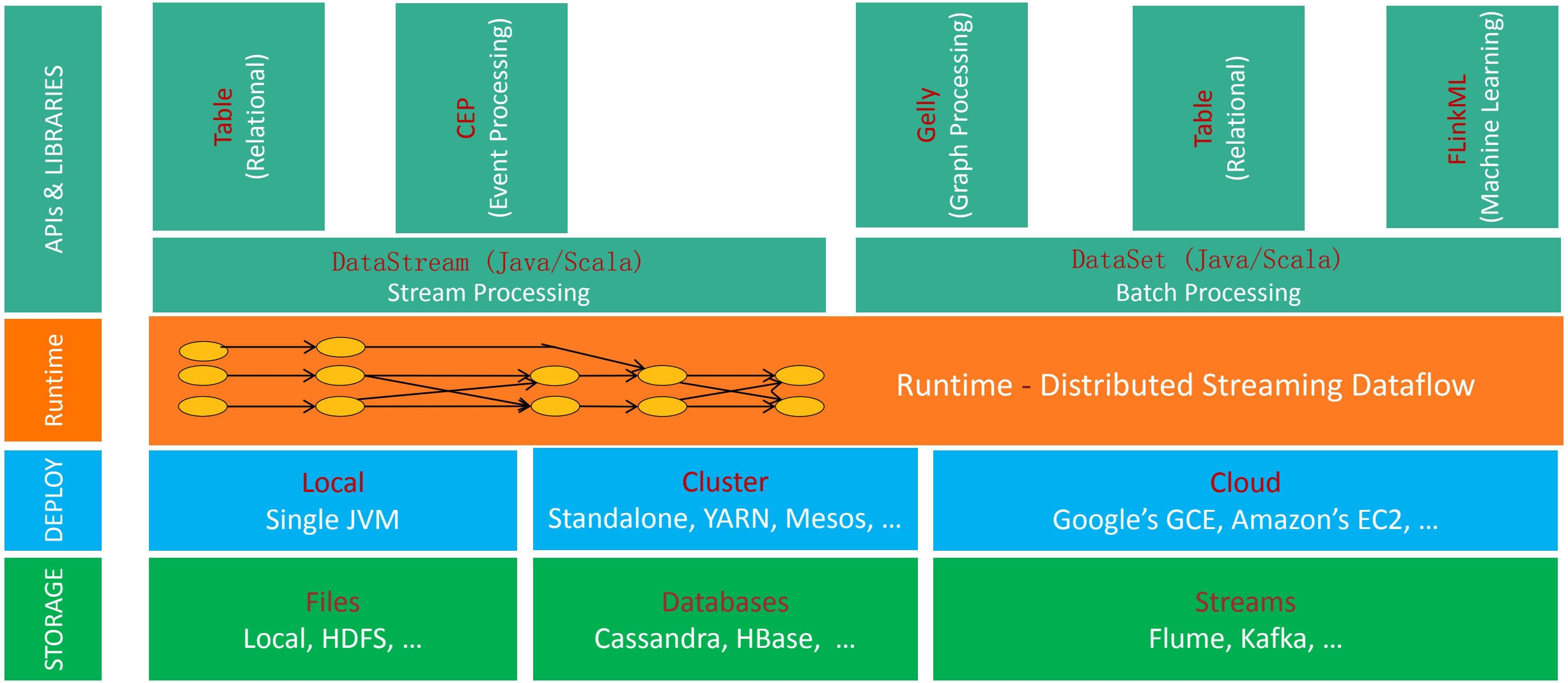


Graph Analysis



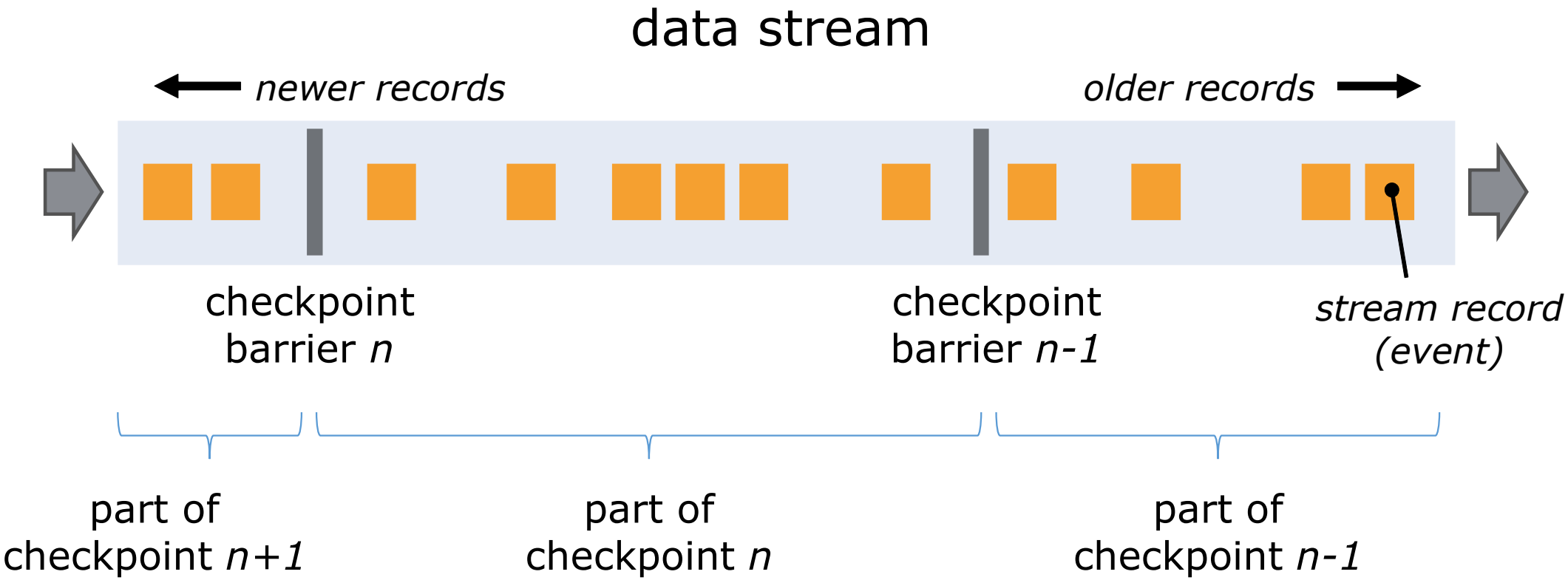
→ Mutable state

Flink Ecosystem

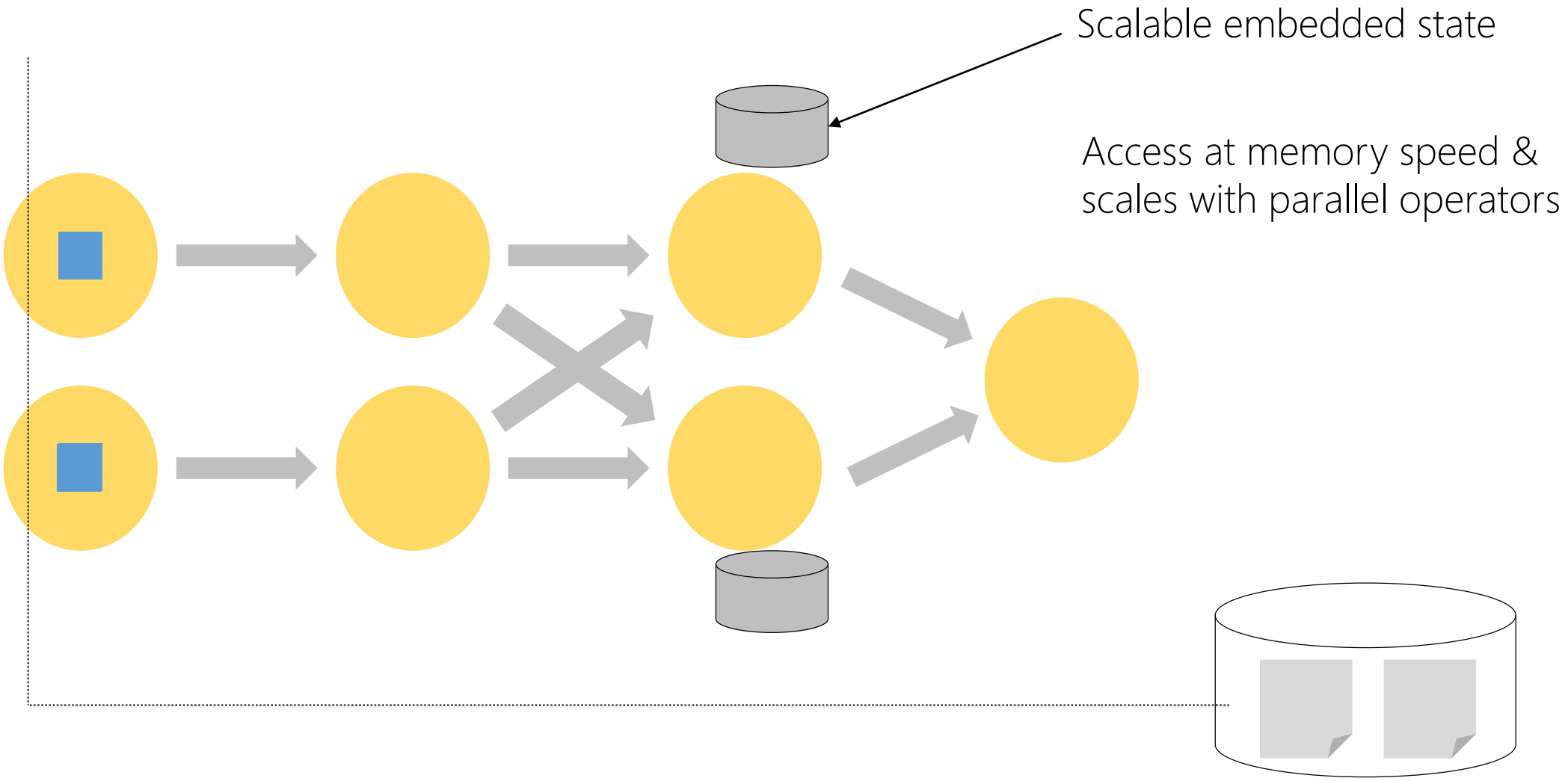


Checkpoint/Recovery

- Chandy-Lamport algorithm
- Periodic asynchronous consistent snapshots of application state
- Provide exactly-once `state` guarantees under failures



Stateful Steam Processing



Alibaba Group 阿里巴巴集团

What is Blink?

Part 3

Blink – Alibaba's version of Flink

✓Looked into Flink two years ago

- best choice of unified computing engine
- a few of issues in flink that can be problems for large scale applications

✓Started Blink project

- aimed to make Flink work reliably and efficiently at the very large scale at Alibaba

✓Made various improvements in Flink runtime

- native run on yarn cluster
- failover optimizations for fast recovery
- incremental checkpoint for super large state
- async operator for high throughputs

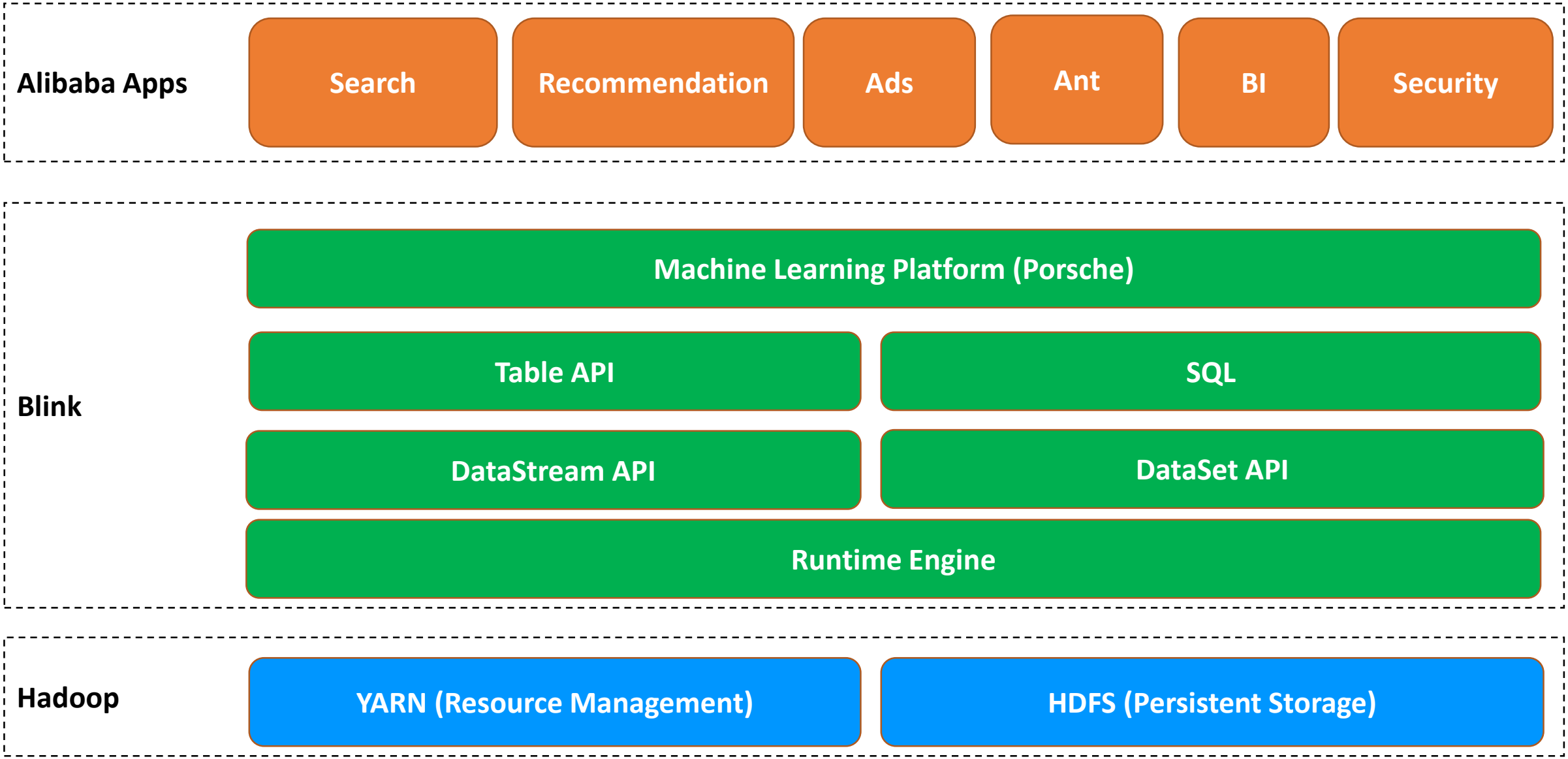
✓Working with Flink community to contribute back since last August

- several key designs
- hundreds of patches

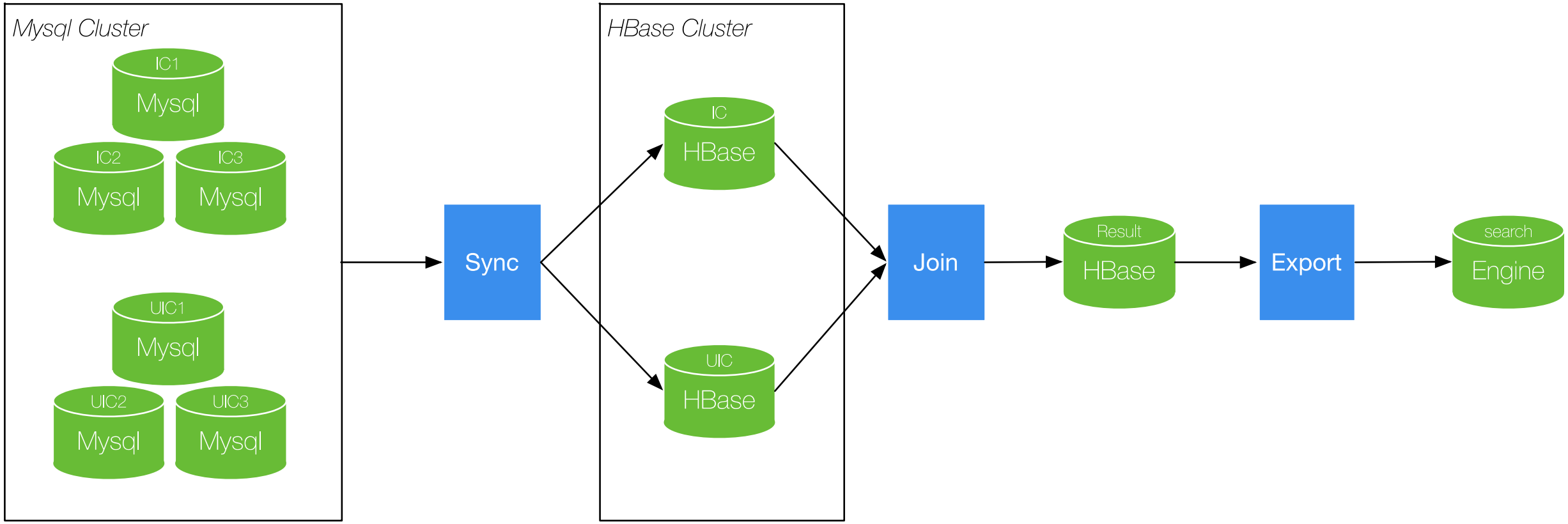
Blink in Alibaba Production

- ✓In production for almost one year
- ✓More than 3000 nodes are running Blink
- ✓The largest Blink cluster is more than 1000 nodes
- ✓There are hundreds of production jobs supported by Blink
- ✓Supported key online Service on last Nov 11th
  - The largest Blink job has 5000 concurrent, 10TB state, billions of QPS
  - Based on the Blink machine learning platform to significantly increase the transaction conversion

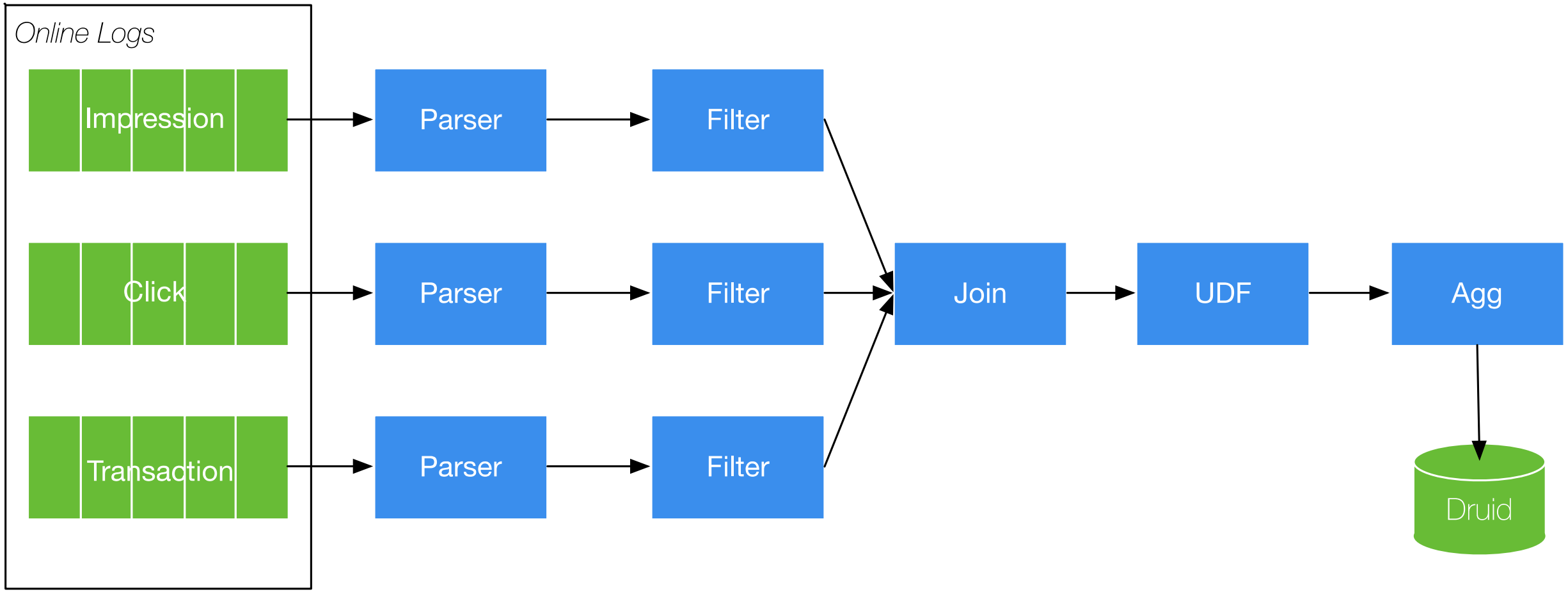
Blink Ecosystem in Alibaba



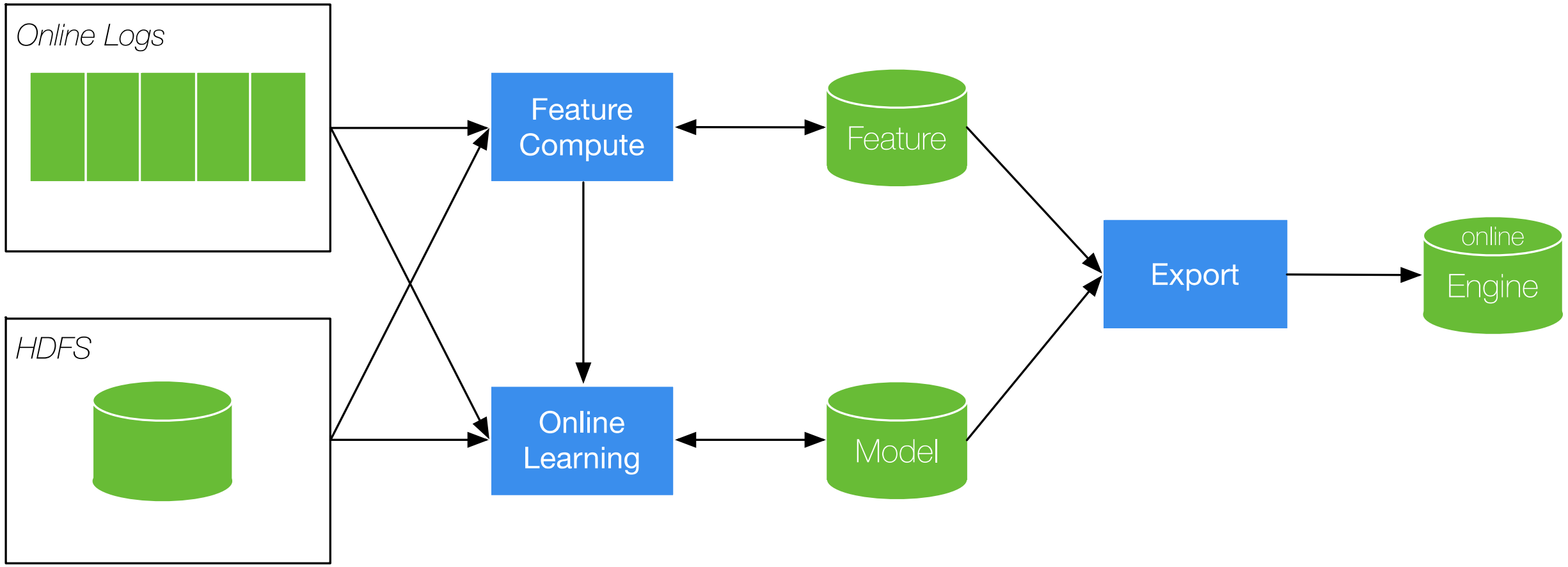
Use Case — Search Index Build & Update



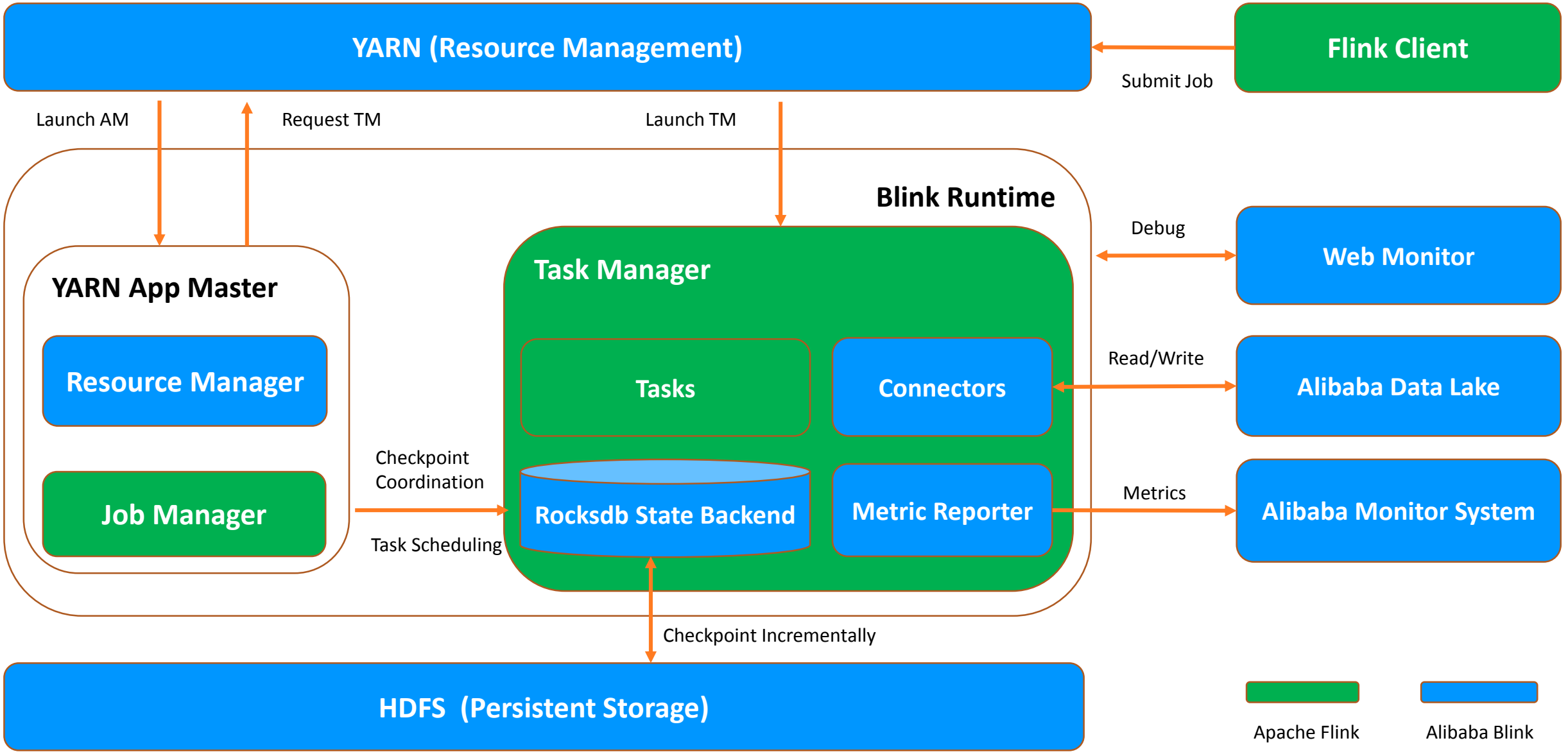
Use Case — Realtime A/B Test



Use Case — Online Machine Learning



Blink Architecture



Improvements to Flink Runtime

- ✓Native integration with Resource Managment for dynamic resource allacation and more larger scale
- ✓Performance Improvements
  - Incremental Checkpoint
  - Asynchronous Operator
- ✓Failover Optimization
  - Fine-grained Recovery for Task Failures
  - Allocation Reuse for Task Recovery
  - Non-disruptive JobManager Failures via Reconciliation

Alibaba Group 阿里巴巴集团

Future Plans

Section 4

Future Plans

- ✓Blink is already popular in the streaming scenarios
  - more and more streaming applications will run on blink
- ✓Make batch applications run on production
  - increase the resource utilization of the clusters
- ✓Blink as Service
  - Alibaba Group Wide
- ✓Cluster is growing very fast
  - cluster size will double
  - thousands of jobs run on production

Alibaba Group
阿里巴巴集团

THANKS

--------- Q&A Section --------