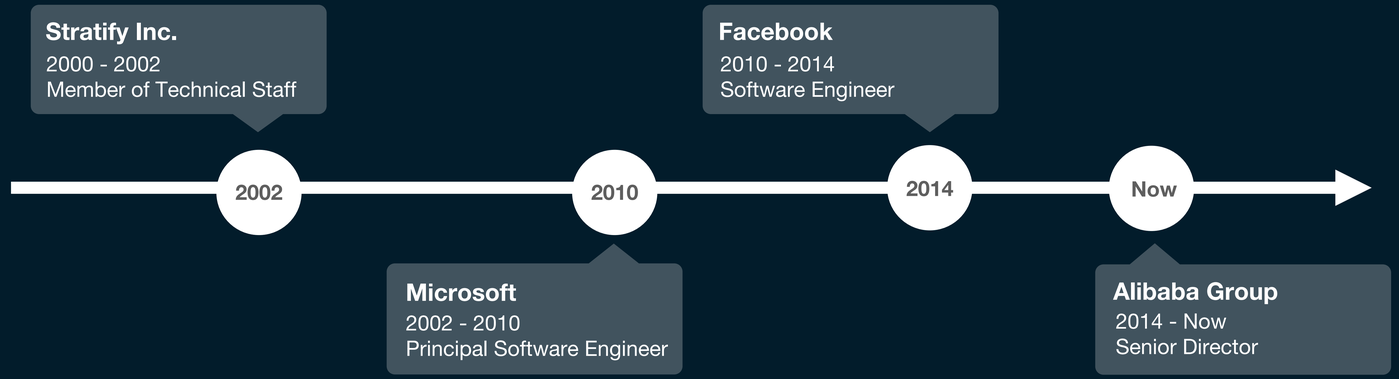
Apache Flink® - Redefining Computation

Xiaowei Jiang

Senior Director

Alibaba

FLINK FORWARD



FLINK FORWARD

聚划算 淘宝网 Taobao.com 天猫 TMALL.COM 阿里云 aliyun.com Alibaba Group 支付宝 ALIPAY CAINIAO 菜鸟 AliExpress

EB Total

PB Everyday

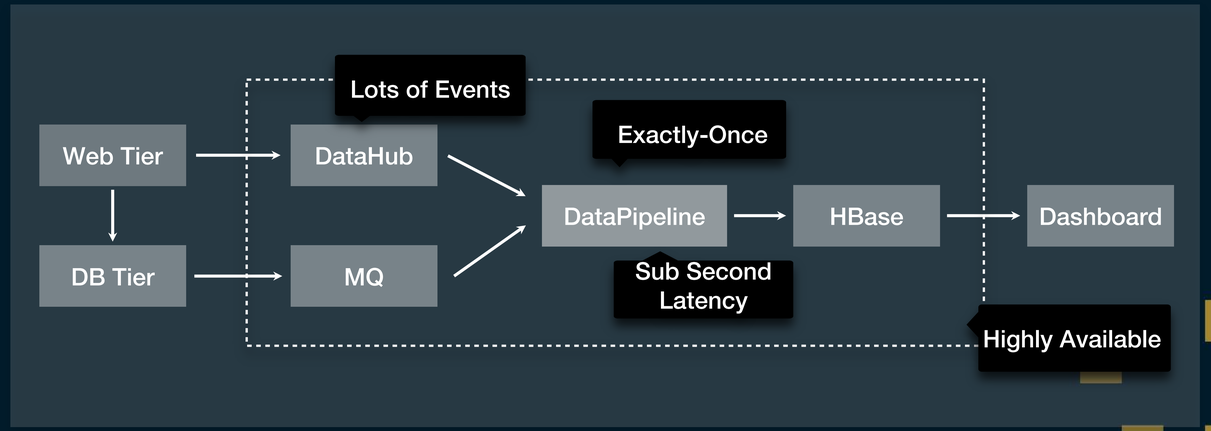
1T Event/Day

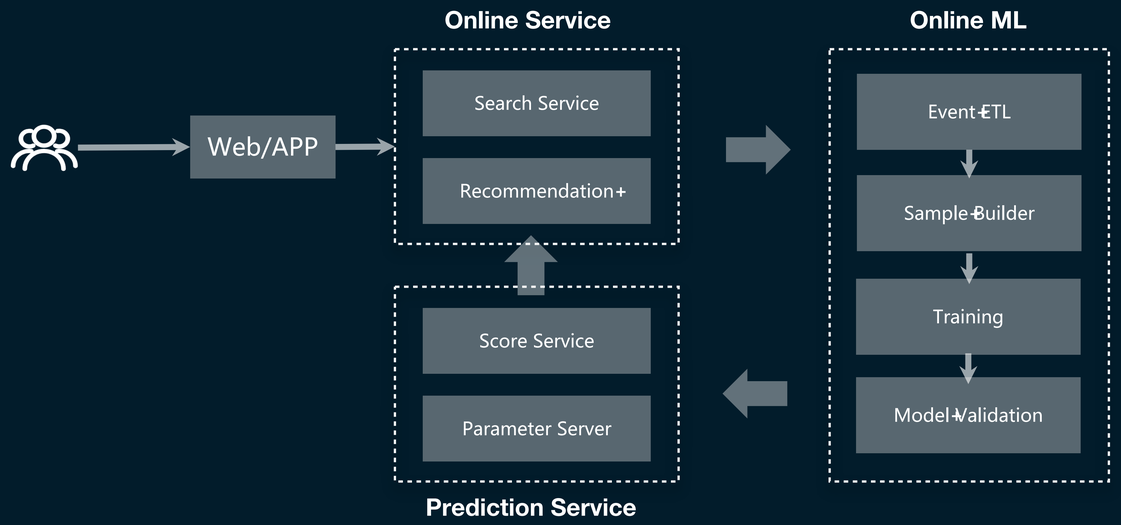
1.7B Events/sec

FLINK FORWARD

1. Stream Processing


2018天猫双11全球狂欢节
24:00:00
2018年天猫双11
成交额
2135


Lots of Events
Web Tier
DB Tier
DataHub
MQ
Exactly-Once
DataPipeline
HBase
Dashboard
Sub Second Latency
Highly Available



Real-Time Personalization

Large Scale: 100 Millions of Events、100 Billions of Features

Low Latency: Second Latency from End to End

Complex Logic: Real-Time Training, Feature/Model Update

FLINK FORWARD

What is Flink

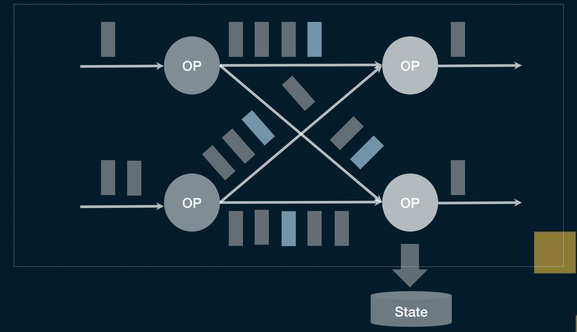
An open source stream processing framework that unifies real-time event-driven applications and real-time analytics.

Flink Program

```
>_

SELECT
  word, count(*)
FROM
  stream
GROUP BY
  word;
```

Flink DAG



FLINK FORWARD

Flink Stream Process

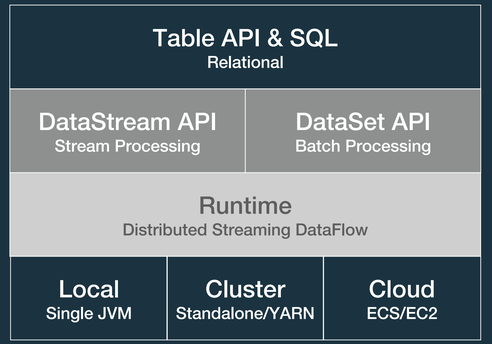
Exactly-Once

High Throughput

Low Latency

Fault Tolerant

FLINK FORWARD



Flink Architecture

Flink Runtime Improvements

Distributed Architecture
Rework Cluster Management [FLIP-6/FLINK-4319]

Fault Tolerance
JobManager Failover [FLINK-4911]
Region-based Task Failover [FLIP1/FLINK-4256]

Performance
Incremental Checkpoint [FLINK-5053]
Async I/O [FLIP12/FLINK-4391]
Credit-based Flow Control [FLINK-7282]

Flink SQL Improvements

Semantics

Functionality
Agg/w Retraction
Window
UDX Support
DDL Support
Connector Support

SQL

FLINK FORWARD

Status of Production

Subsecond Latency

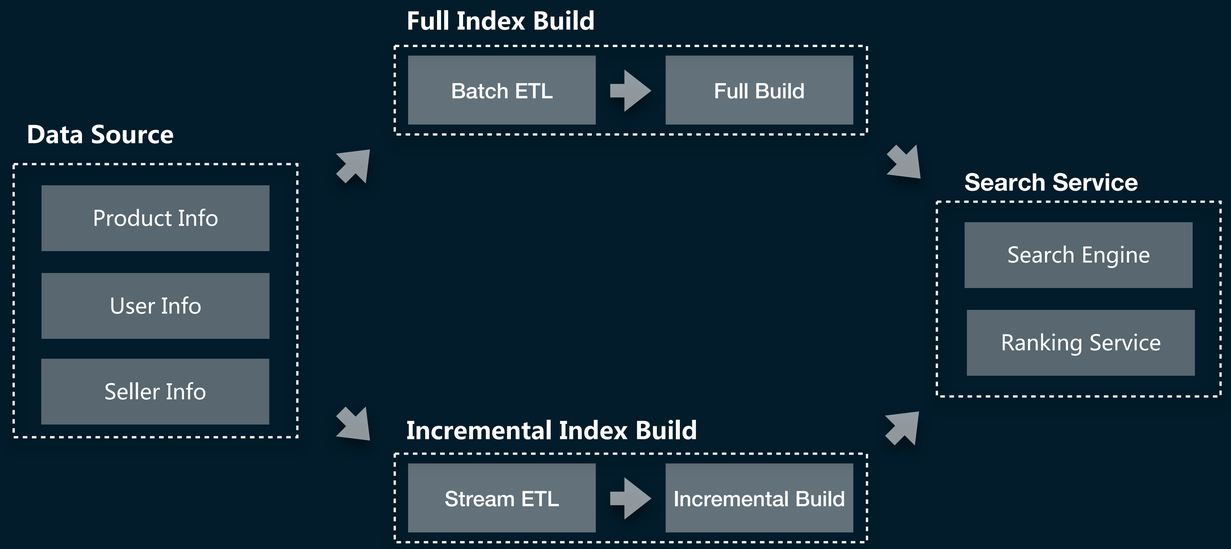
Billions Events/sec

Ten Thousand Nodes

Ten Thousand Jobs

FLINK FORWARD

2. Unified Engine



Index Pipelines for Search

Development Efficiency: Full/Incremental Index Build

Challenge: Consistency

FLINK FORWARD

Low Latency
Fixed Query

Stream Processing

Periodic/Continuous
Batch Jobs

Progressive Processing

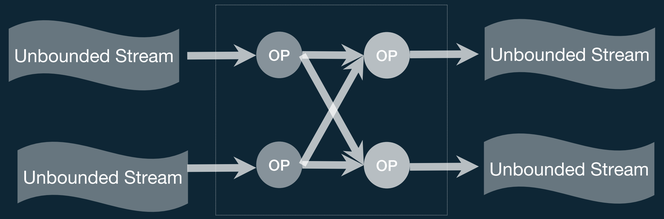
High Throughput
Flexible Query

Batch Analytics

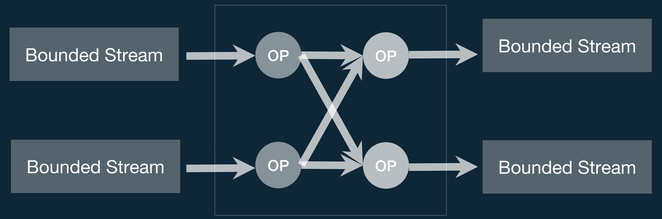
Streaming as the **core abstraction**, Batch as a **special case** of streaming

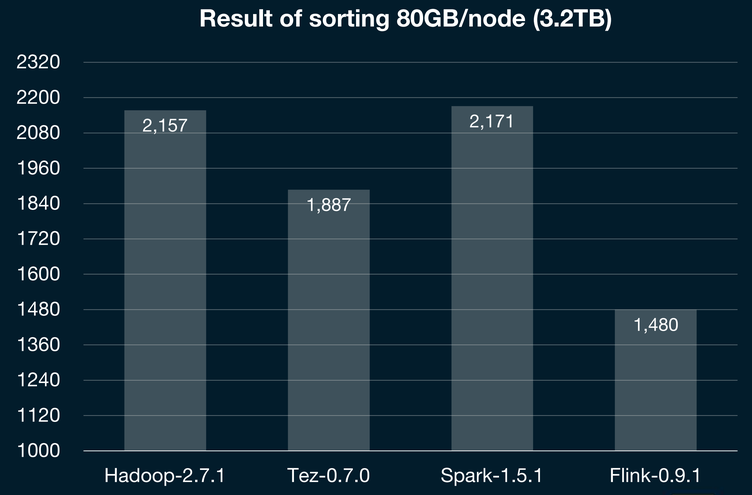
FLINK FORWARD

Stream Job



Batch Job





Flink is the fastest due to its pipelined execution

Tez and Spark do not overlap 1st and 2nd stages

MapReduce is slow despite overlapping stages

A Comparative Performance Evaluation of Flink, Dongwon Kim, POSTECH, Flink Forward 2015


Table API & SQL
Relational
DataStream API
Stream Processing
DataSet API
Batch Processing
Runtime
Distributed Streaming DataFlow
Local
Single JVM
Cluster
Standalone/YARN
Cloud
ECS/EC2
Old Design
DataStream
API
TableAPI
Relational
SQL
Relational
Query Processor
Query Optimization & Query Execution
Runtime
DAG API & Stream Operators
Local
Single JVM
Cluster
Standalone/YARN
Cloud
ECS/EC2
New Design

FLINK FORWARD

Improvements in New Design

Runtime

New Operator Framework
Customizable Scheduling
Flexible Chaining

Query Execution

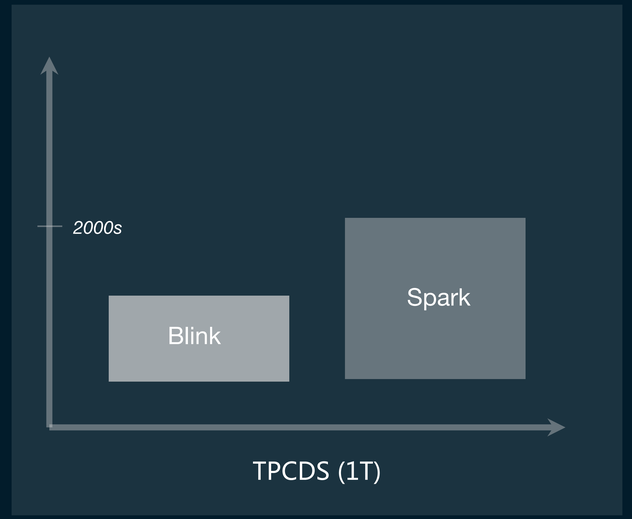
Expression Optimizations
Performant Operators
Resource Optimizations

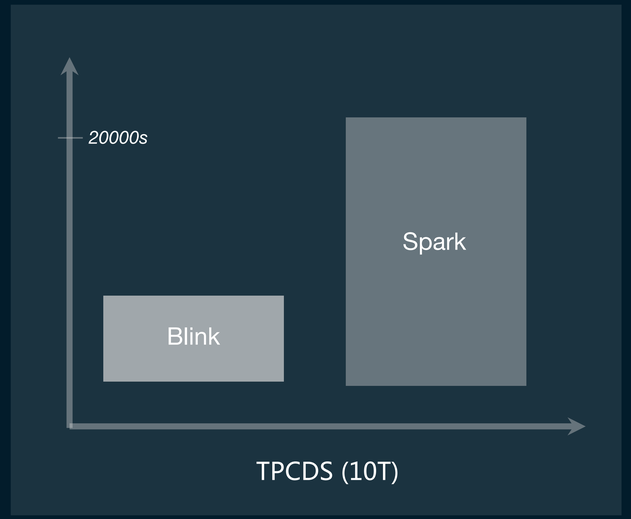
Query Optimizer

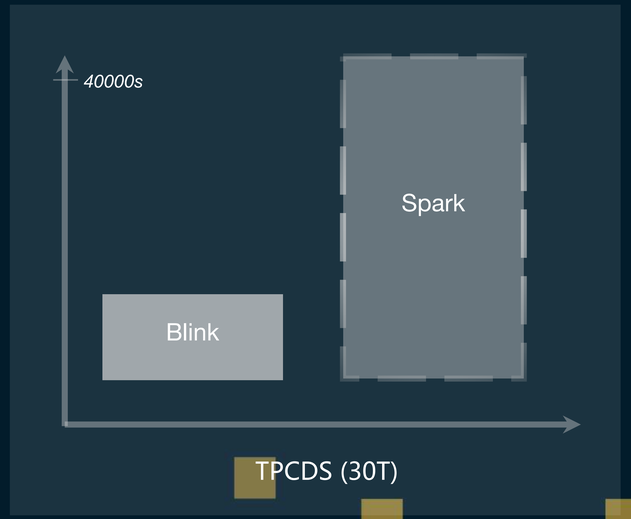
Cost Based
Advanced Rules
Rich Stats

FLINK FORWARD

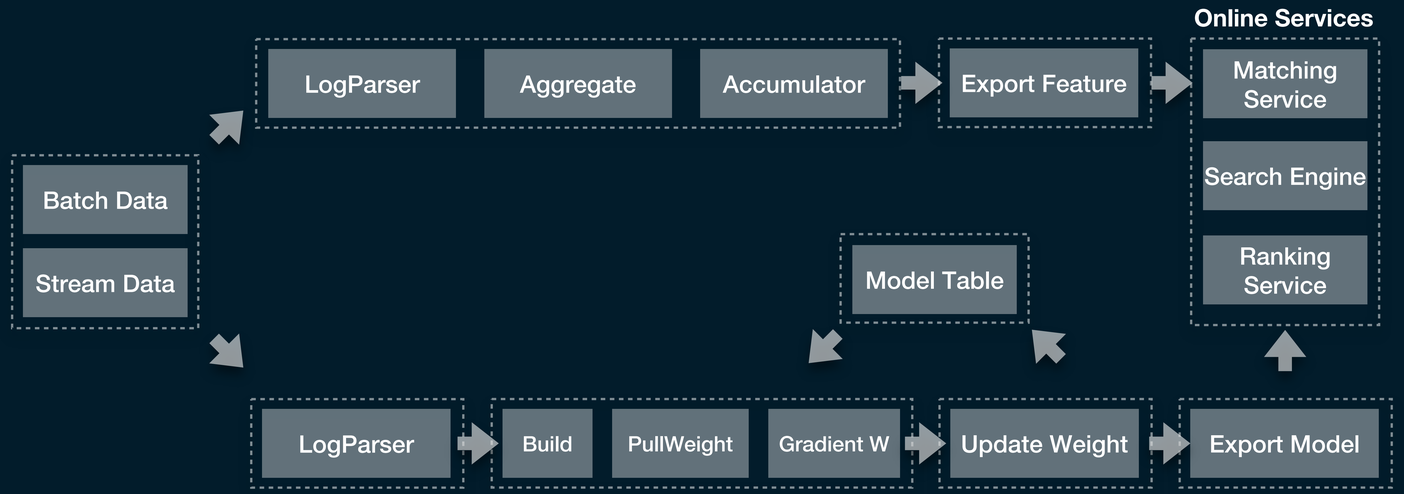
TPC-DS Performance *(the Lower, the Better)*







Spark选用社区2.3.1版本

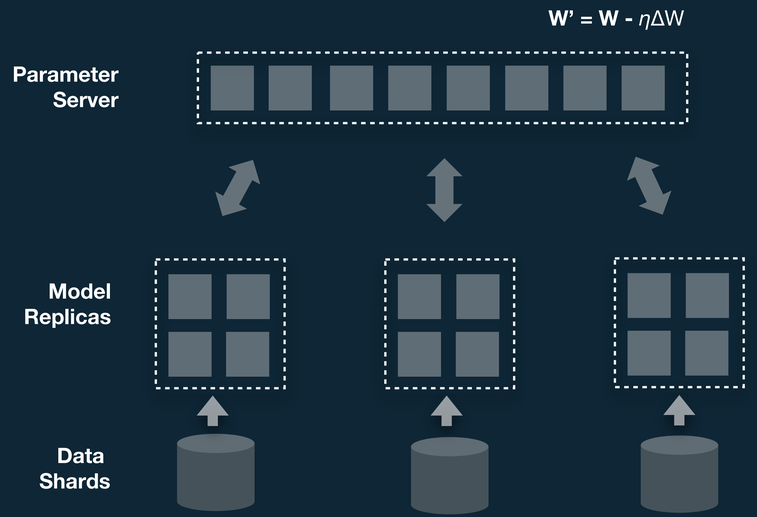


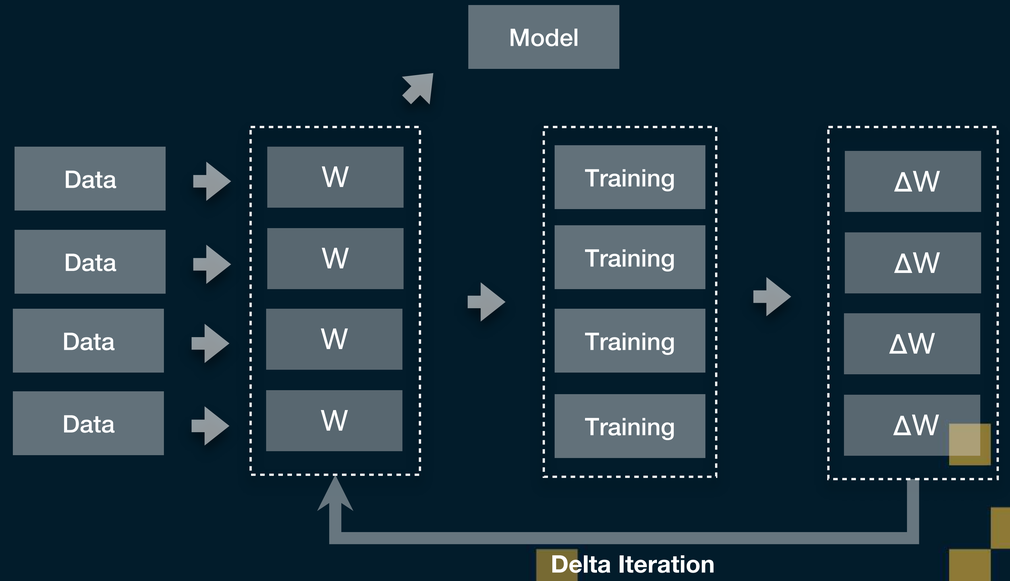
Search’ s Algorithm Platform

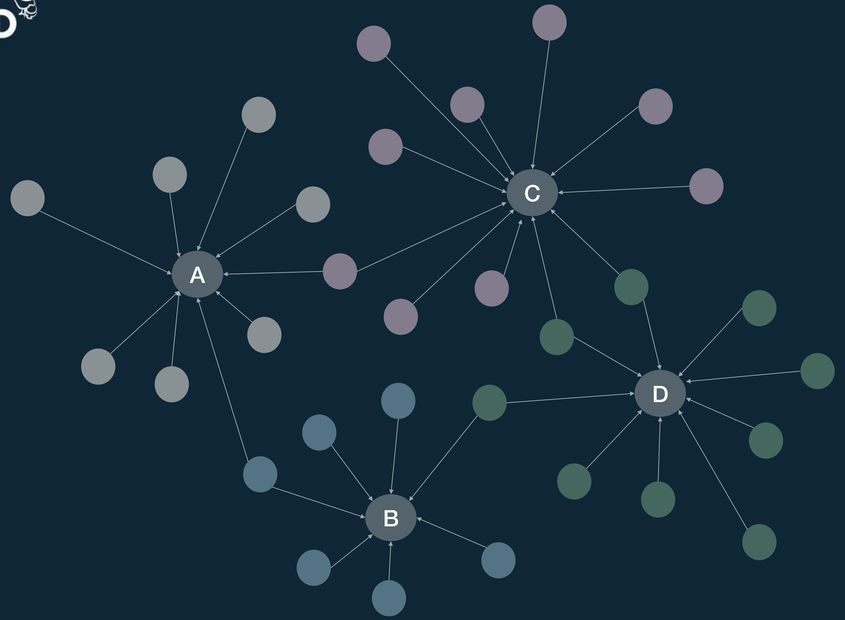
Unified Pipeline for Batch & Streaming
Streaming: 100M QPS, 100B features
Batch: Over 400TB in a single job

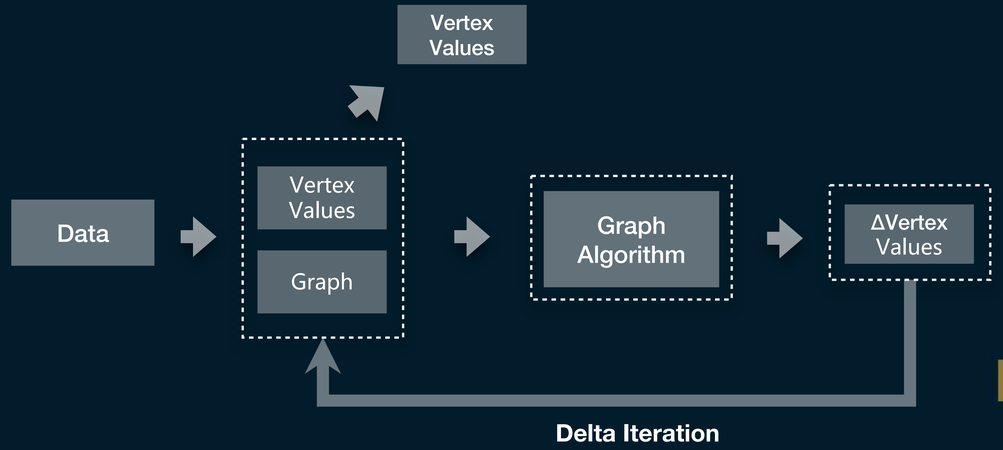
FLINK FORWARD

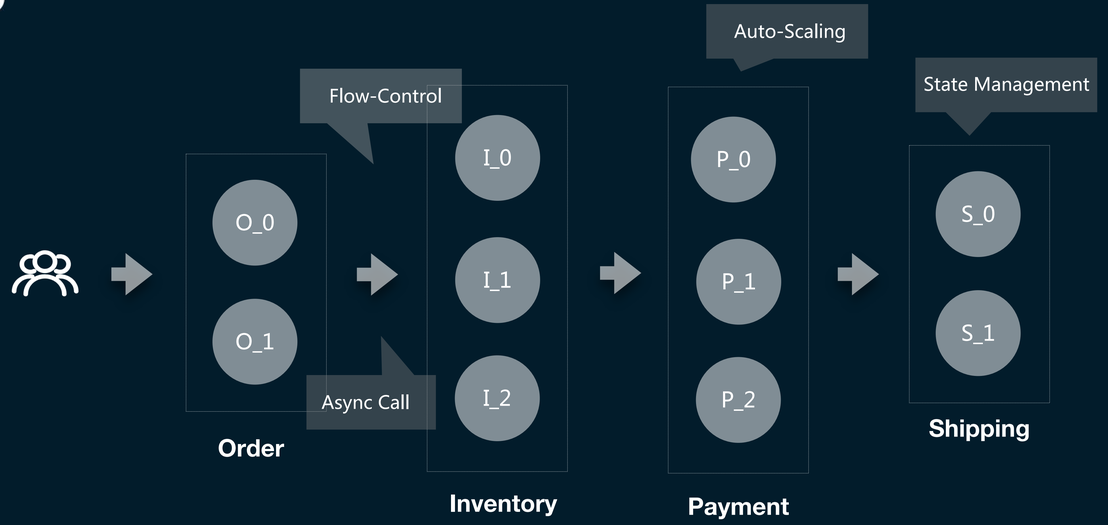
3. Future


W' = W - ηΔW
Parameter Server
Model Replicas
Data Shards


Model
Data
Data
Data
Data
W
W
W
W
Training
Training
Training
Training
ΔW
ΔW
ΔW
ΔW
Delta Iteration


A
B
C
D


Vertex Values
Data
Vertex Values
Graph
Graph Algorithm
ΔVertex Values
Delta Iteration

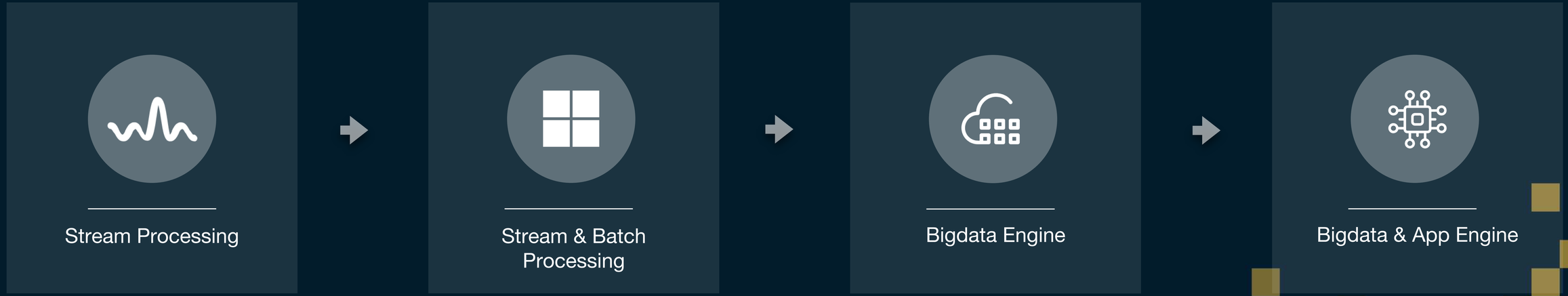


MicroService

Event Driven/Async Processing
Back Pressure & Flow Control
Auto-Scaling
State-Management & Atomicity

FLINK FORWARD

Apache Flink – Streaming Technology Redefining Computation



THANKS