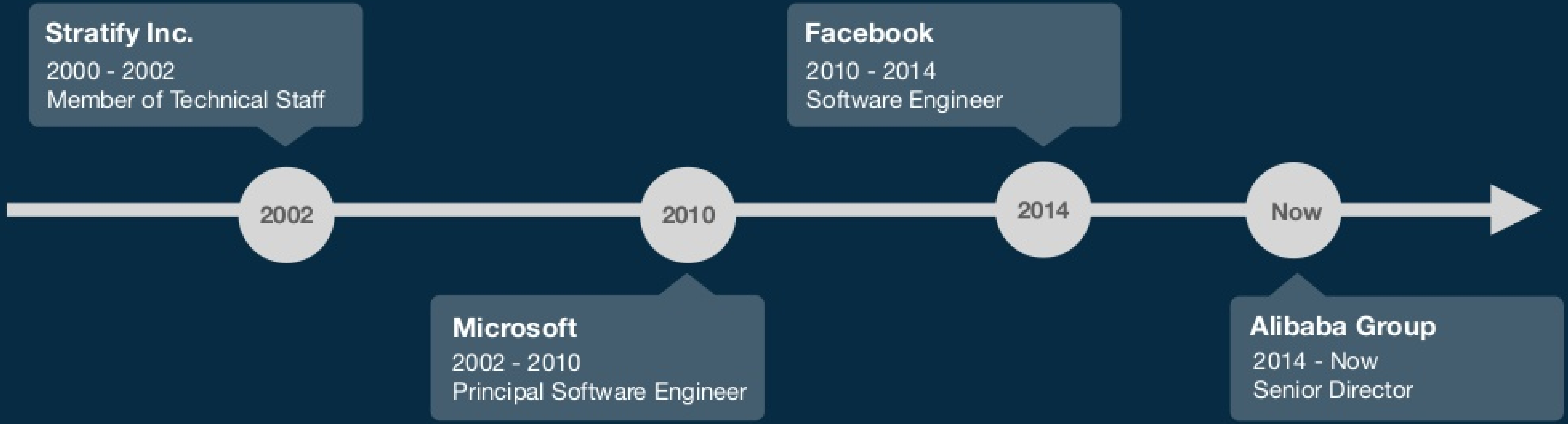
Unified Engine for Data Processing and AI

Xiaowei Jiang
Sept, 2018

About Me



Stratify Inc.
2000 - 2002
Member of Technical Staff

2002

Microsoft
2002 - 2010
Principal Software Engineer

2010

Facebook
2010 - 2014
Software Engineer

2014

Now

Alibaba Group
2014 - Now
Senior Director

About Alibaba

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

EB Total

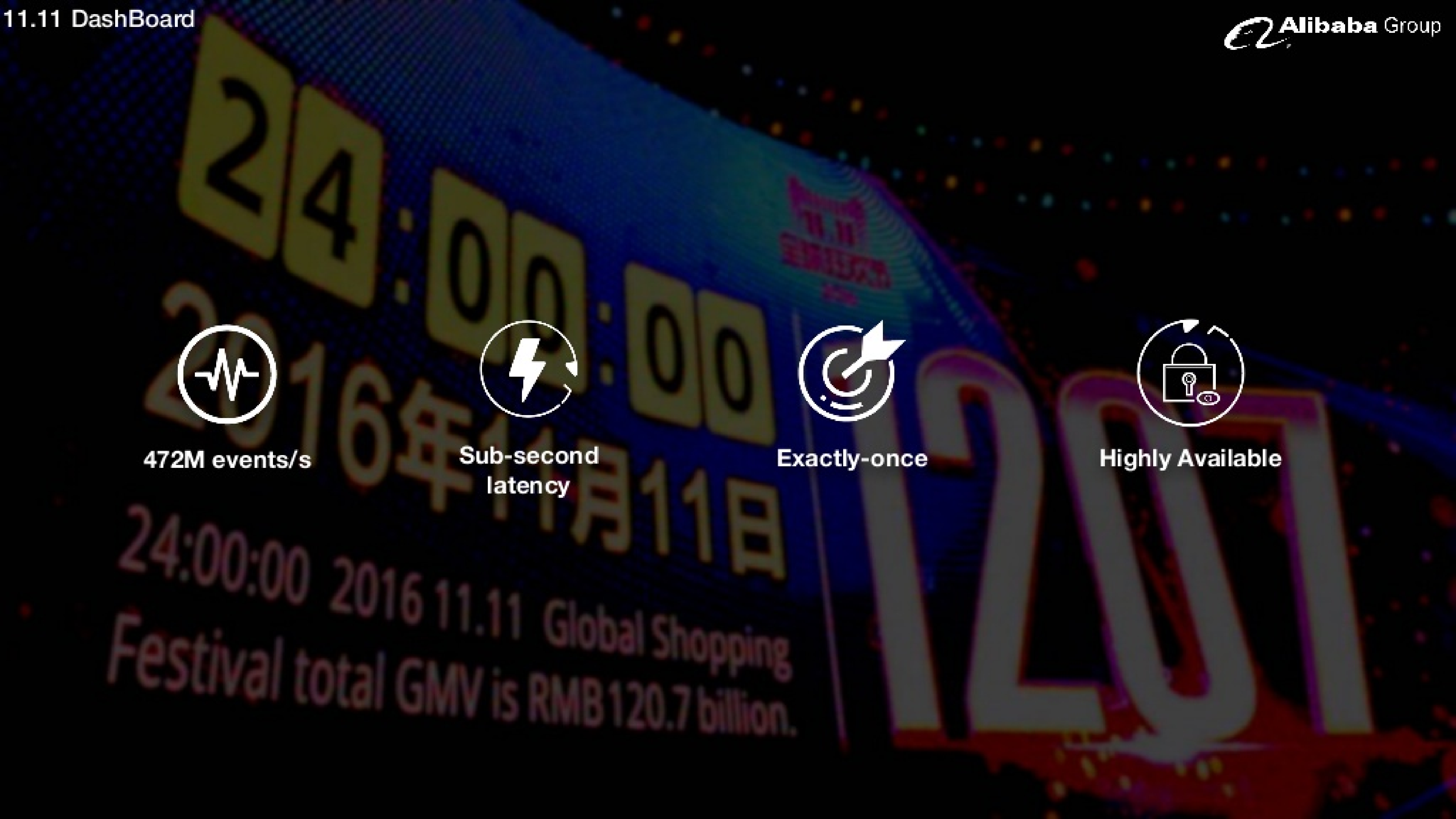
PB Everyday

1T Event/Day

472M Events/sec

11.11 DashBoard

Alibaba Group


24:00:00
2016年11月11日
24:00:00 2016 11.11 Global Shopping Festival total GMV is RMB 120.7 billion.
1207

- 472M events/s
- Sub-second latency
- Exactly-once
- Highly Available

Scenarios for Data Processing

Low Latency
Fixed Query

Stream Processing

Periodic/Continuous
Batch Jobs

Progressive Processing

High Latency
Flexible Query

Batch Analytics

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

Flink Architecture

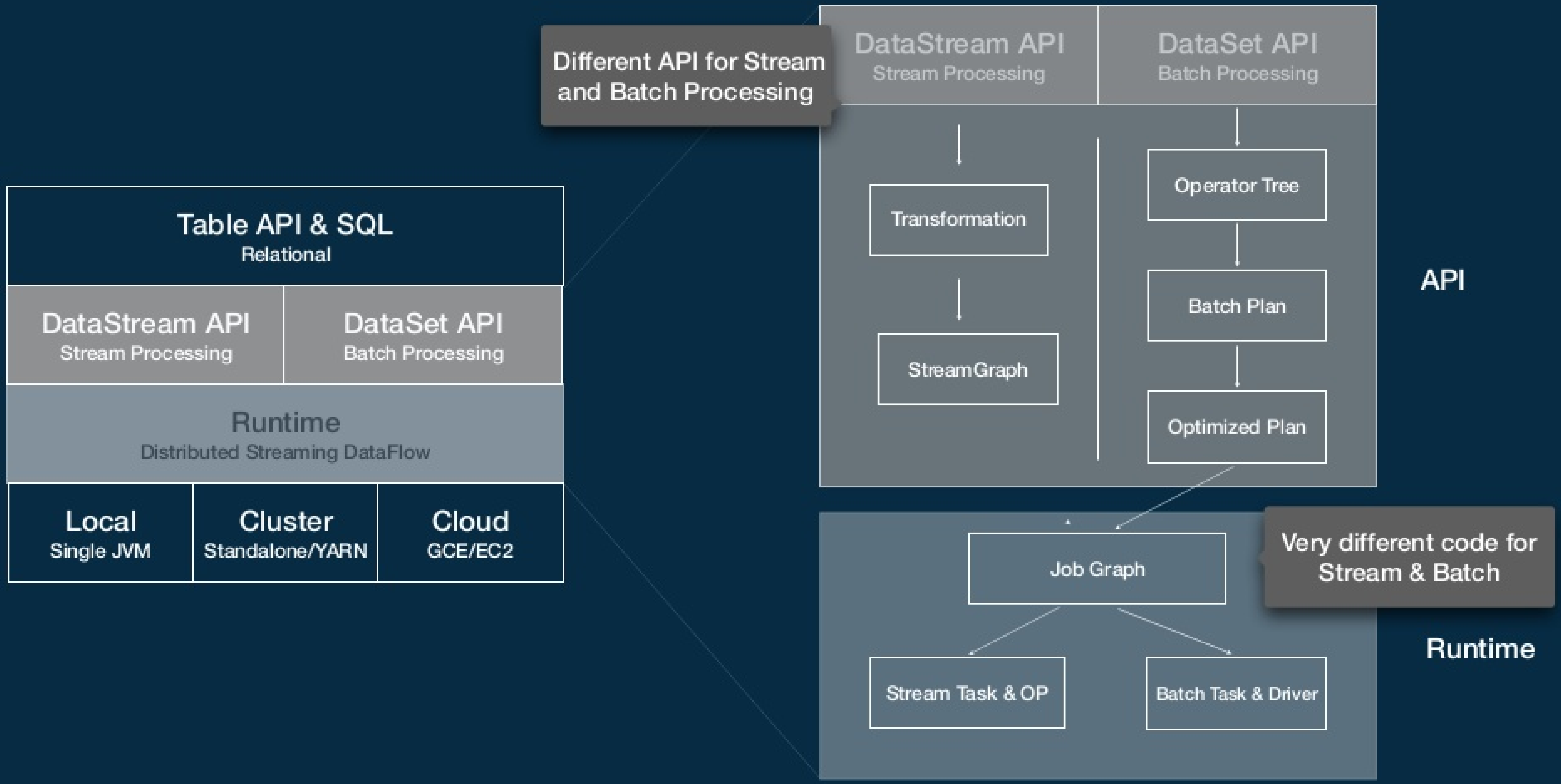


Table API & SQL
Relational

DataStream API
Stream Processing

DataSet API
Batch Processing

Runtime
Distributed Streaming DataFlow

Local
Single JVM

Cluster
Standalone/YARN

Cloud
GCE/EC2

Different API for Stream and Batch Processing

DataStream API
Stream Processing

DataSet API
Batch Processing

Transformation

StreamGraph

Operator Tree

Batch Plan

Optimized Plan

API

Job Graph

Stream Task & OP

Batch Task & Driver

Very different code for Stream & Batch

Runtime

Stream Processing Engines

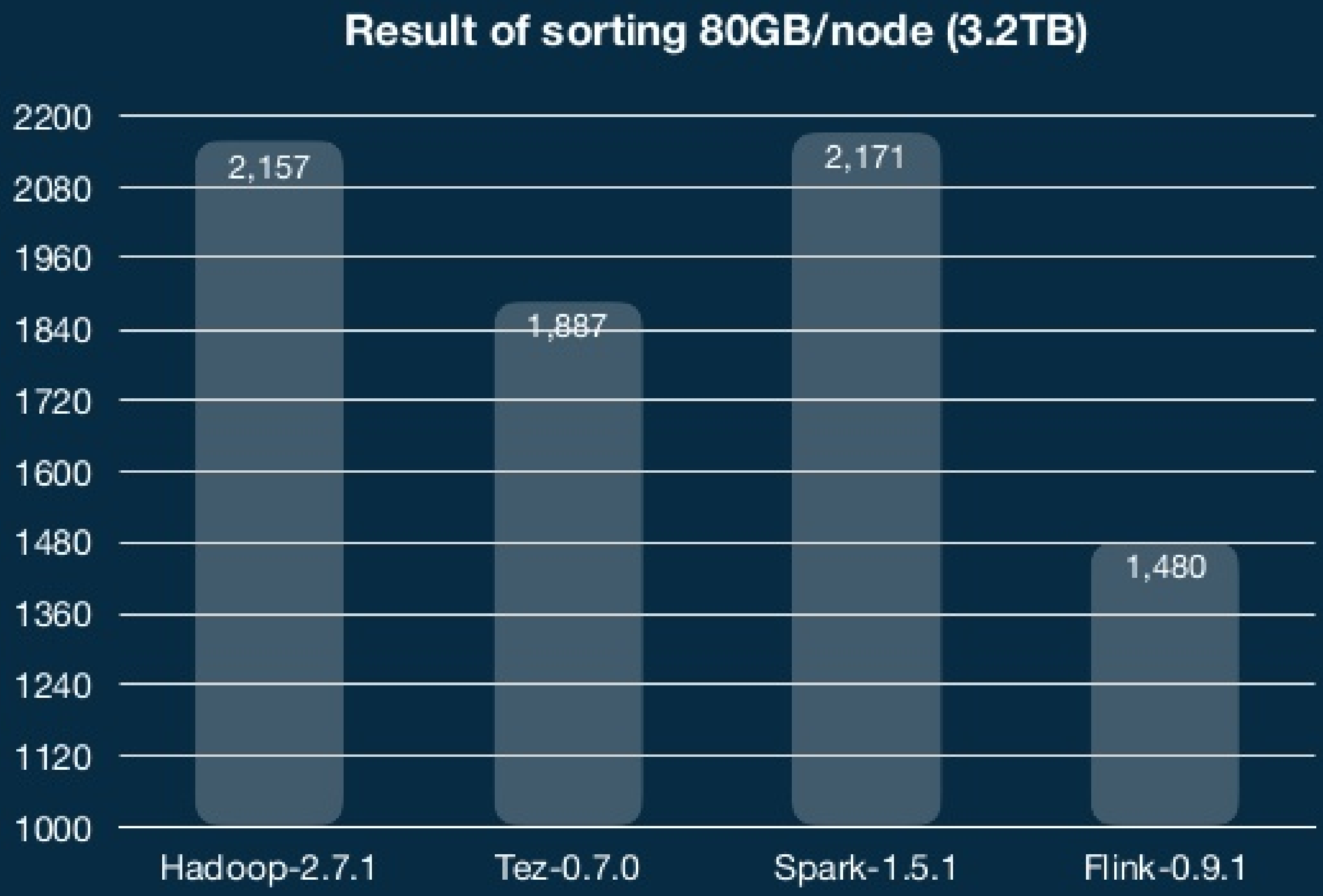
Apache Flink is the most sophisticated open-source Stream Processor

Batch Processing Engines

Can Apache Flink become the most sophisticated open-source batch processor?

Batch Micro-benchmark

Alibaba Group



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

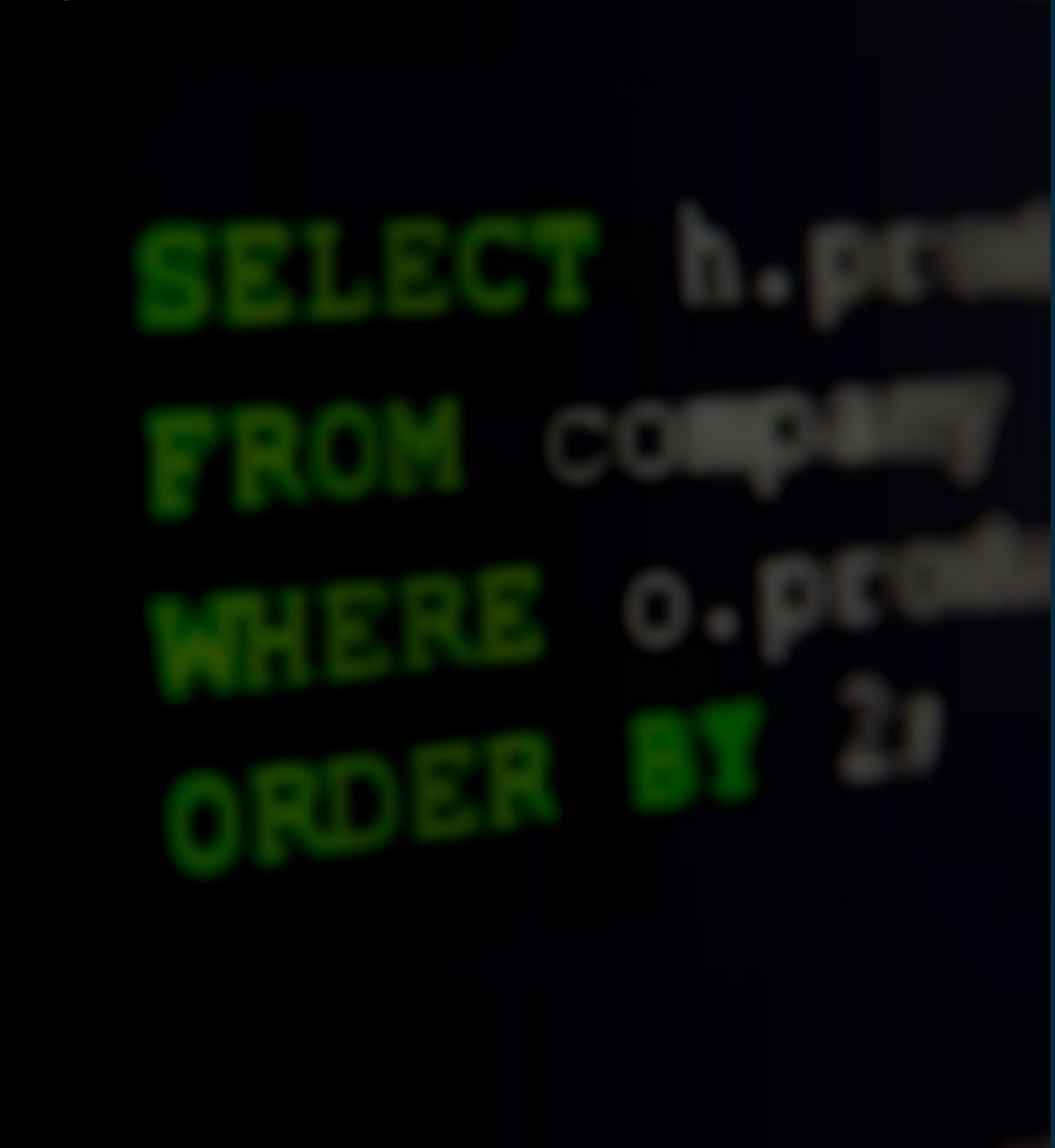
Unified Engine

Functionality

Performance

Reliability

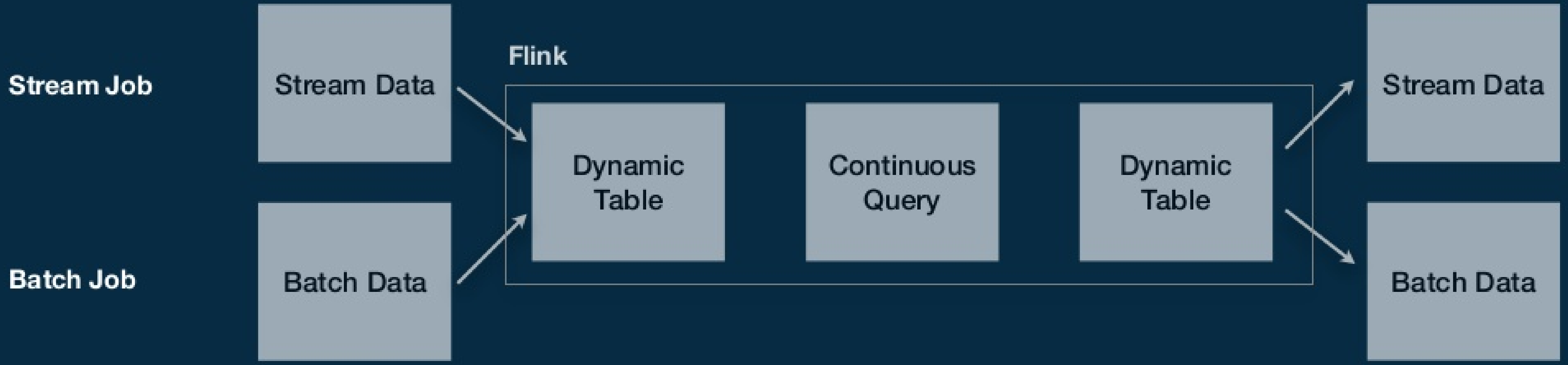
Why SQL



```
SELECT h.pr
FROM company
WHERE o.pr
ORDER BY 2;
```

- Declarative
- Optimizable
- Understandable
- Stable
- Unified

Dynamic Table



Streaming SQL Functionalities

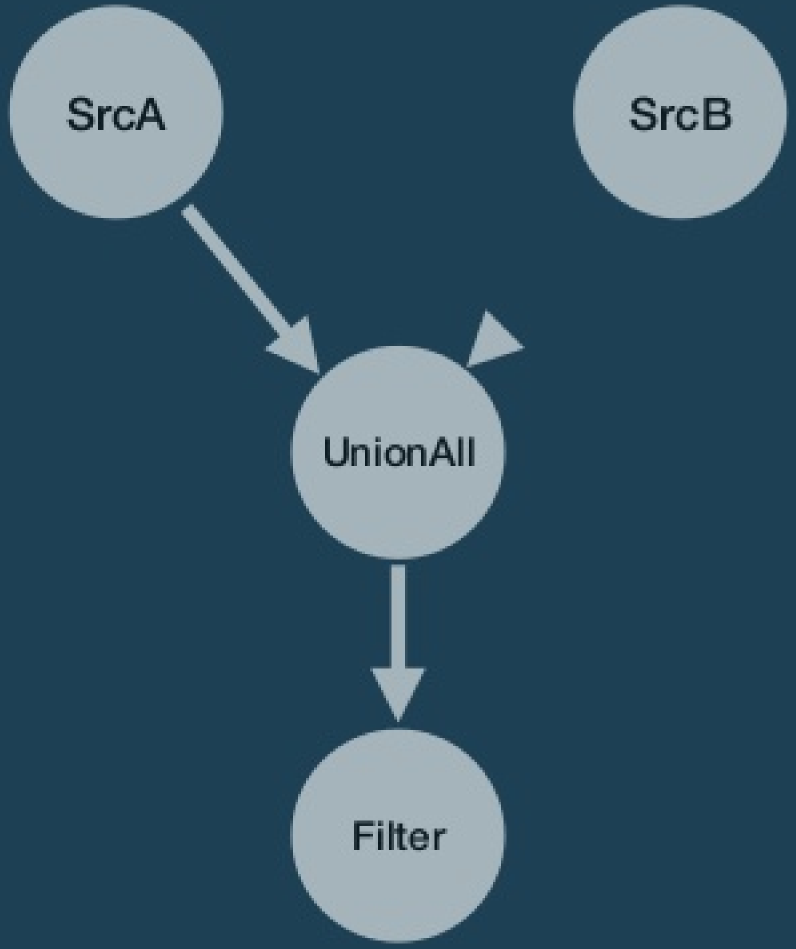
- Stream Join
- Agg w/ Retraction
- Window
- UDX
- DDL Support
- Connector

Batch SQL Completion

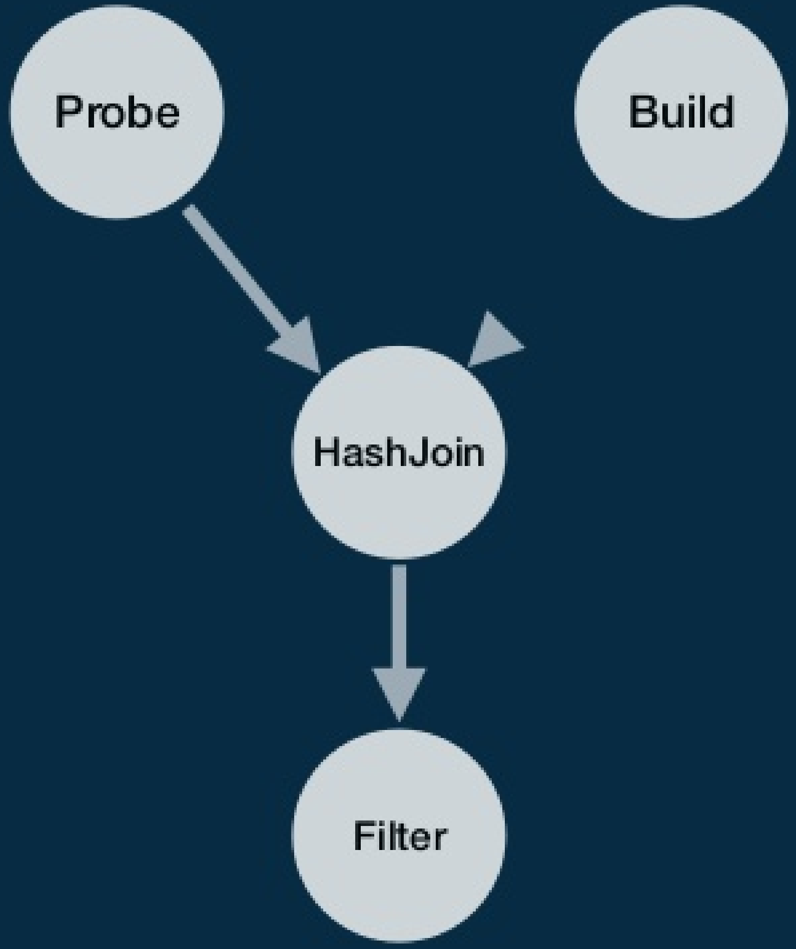
TPCH & TPCDS

2K+ CASE & SQL

Existing Execution Model

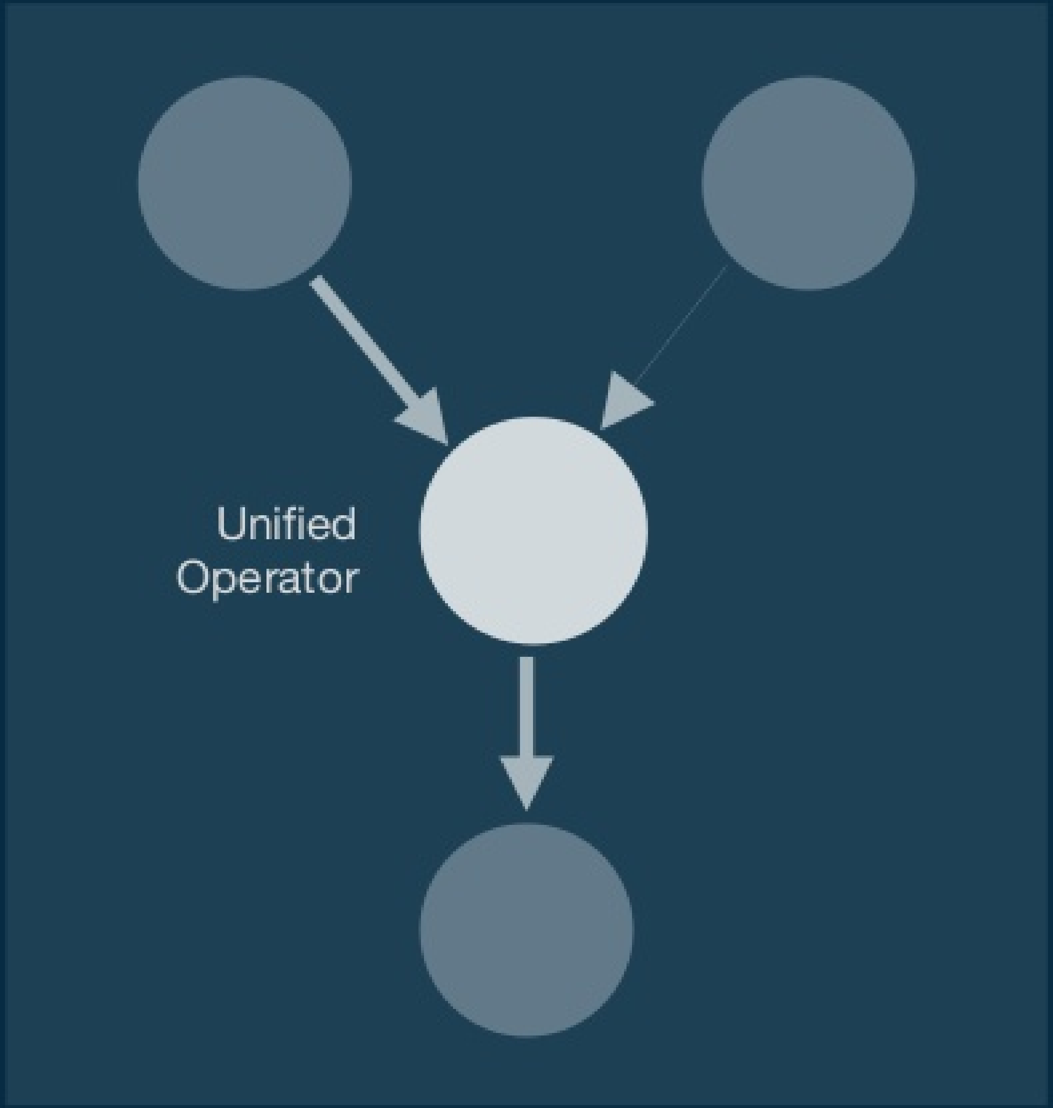


Stream Model



Batch Model

Unified Operator Framework



Unified Operator Abstraction

- *Operators can choose inputs*
- *Operators can be chained easily*
- *Helps batch as well as streaming*

Flink Architecture: New Design

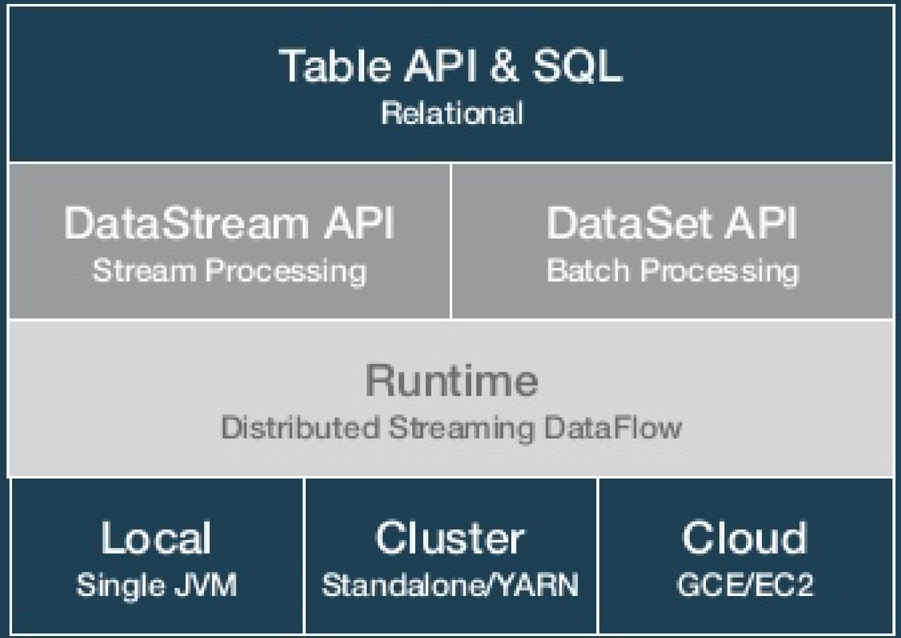


Table API & SQL
Relational

DataStream API
Stream Processing

DataSet API
Batch Processing

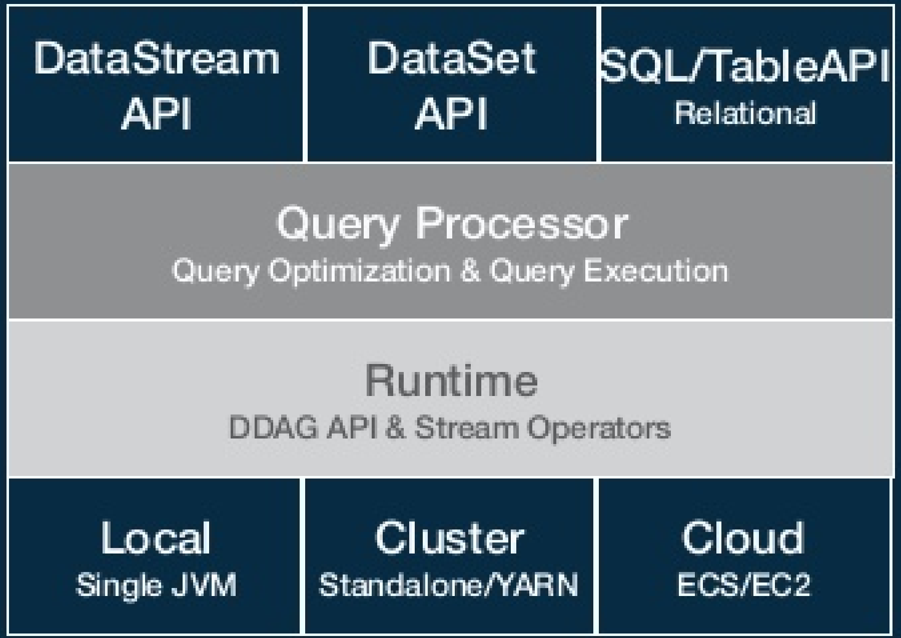
Runtime
Distributed Streaming DataFlow

Local
Single JVM

Cluster
Standalone/YARN

Cloud
GCE/EC2

Old Design



DataStream API

DataSet API

SQL/TableAPI
Relational

Query Processor
Query Optimization & Query Execution

Runtime
DDAG API & Stream Operators

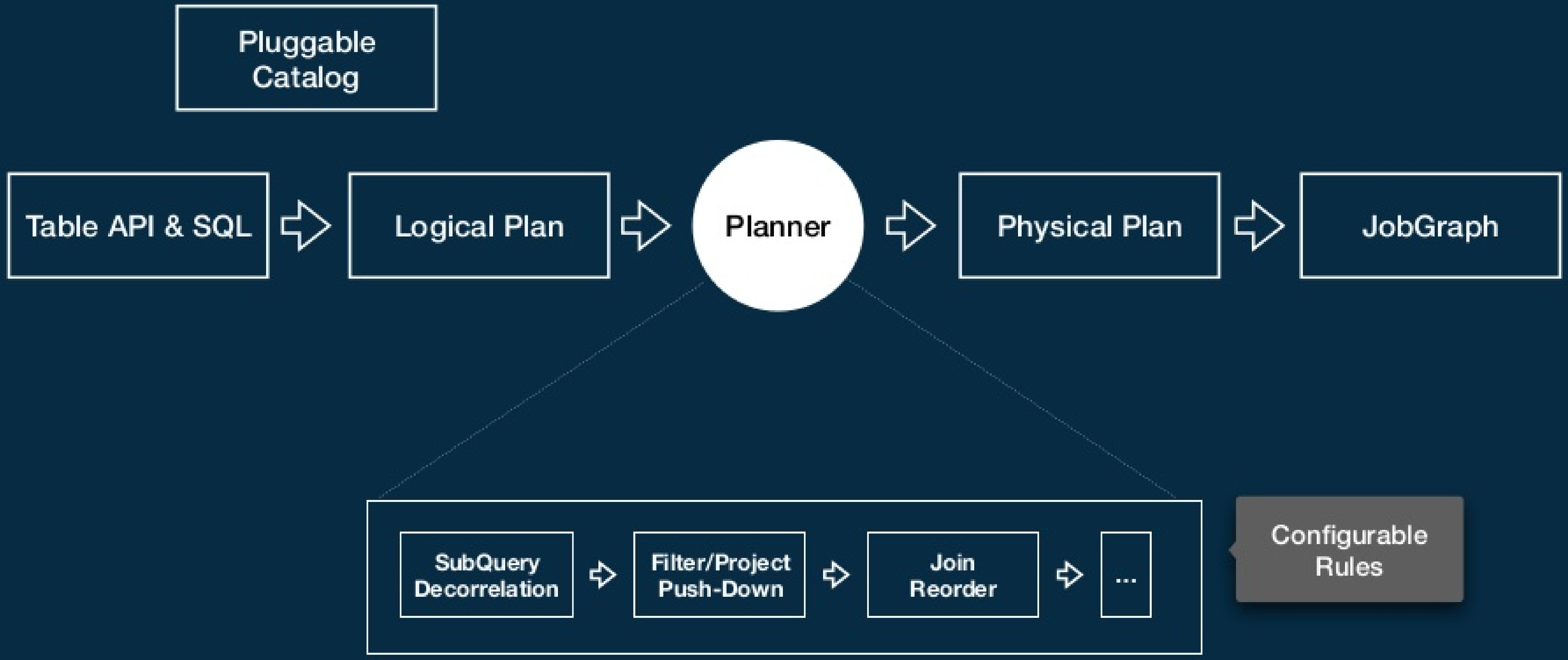
Local
Single JVM

Cluster
Standalone/YARN

Cloud
ECS/EC2

New Design

Query Processing



Configurable Rules

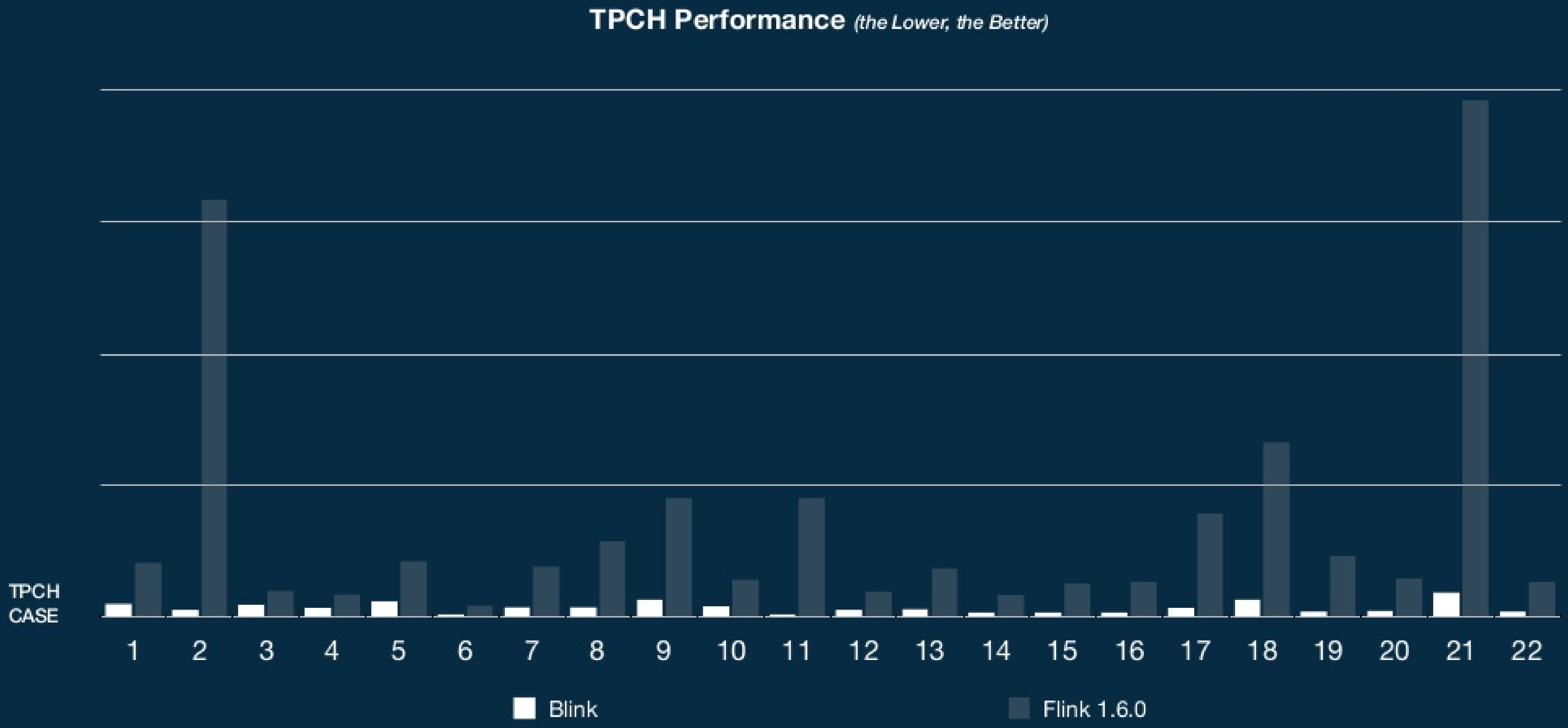
Unified Engine

Functionality

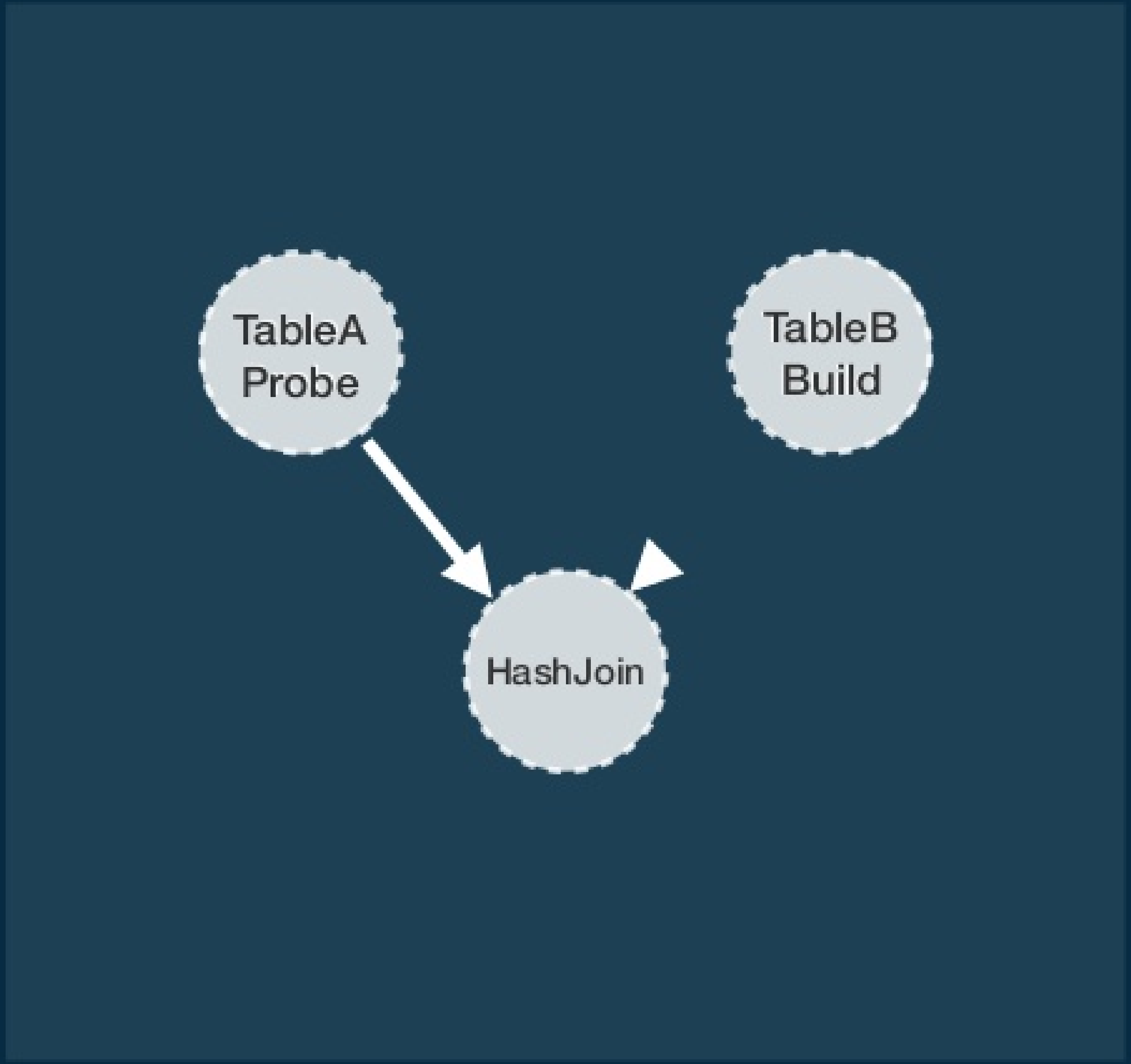
Performance

Reliability

Batch Performance



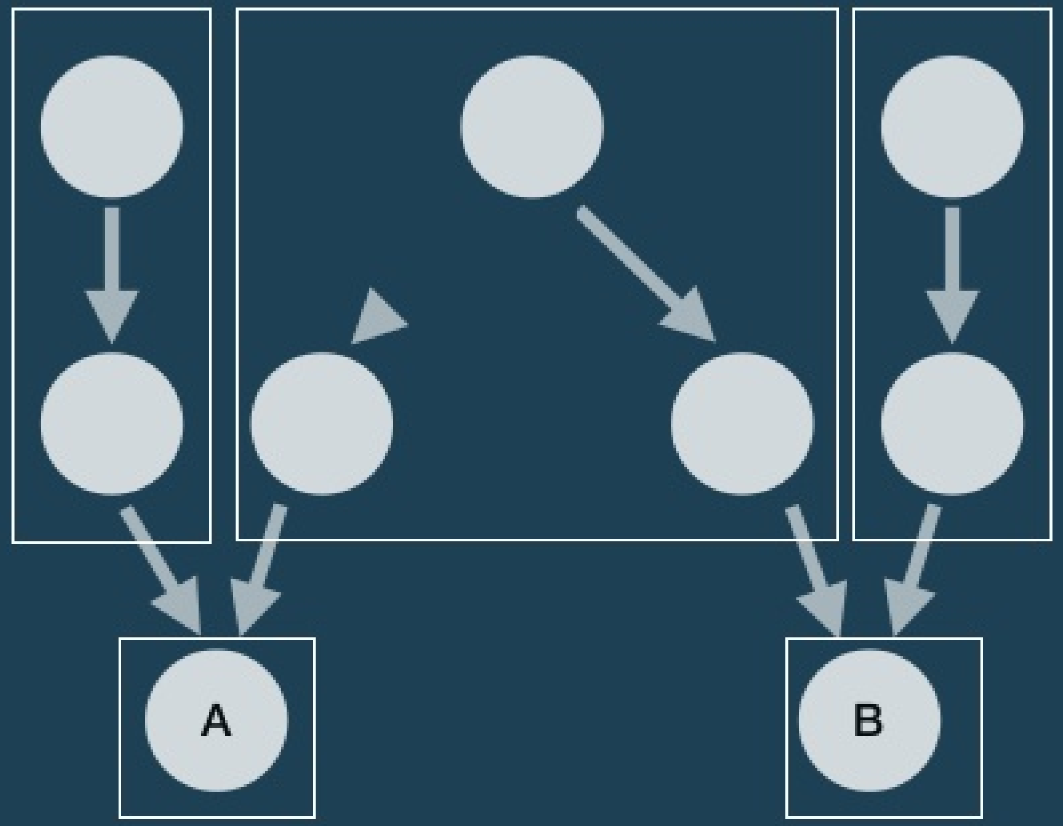
Scheduling



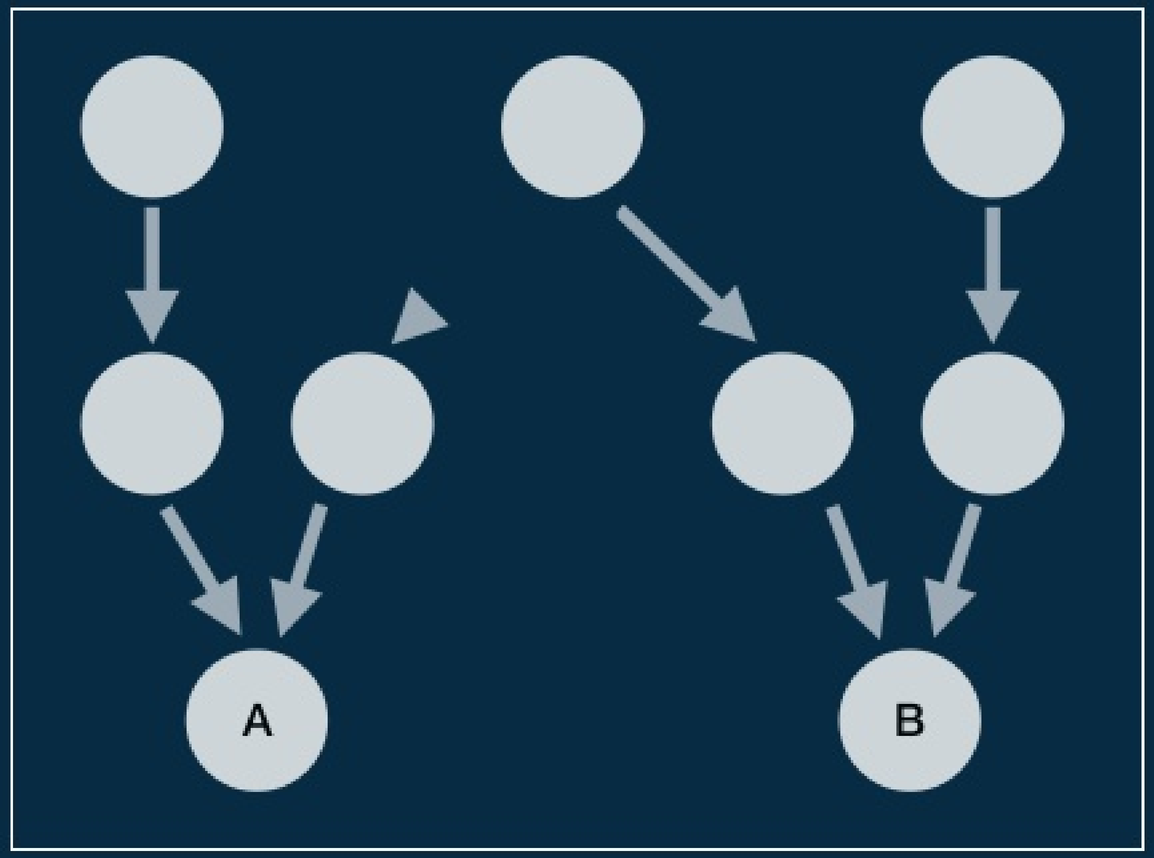
Customizable Scheduling

- *Flexible control over when tasks get scheduled*
- *Much better resource usage achieved*

Flexible Chaining



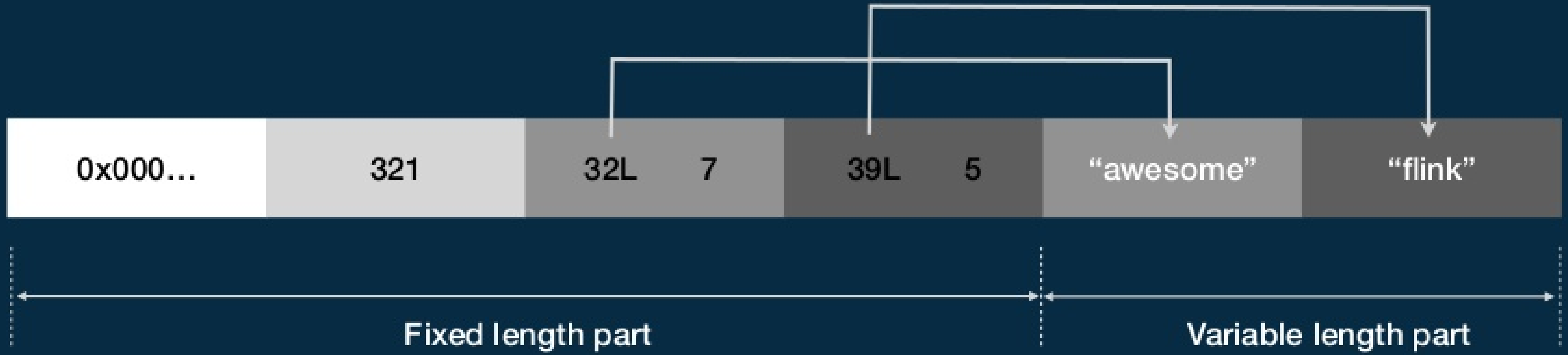
Only One-Input Operators can be chained



Multi-Input Operators can also be chained

Record Format

- Introduced new row format: BinaryRow
- Tight integration with memory management
- Avoid deserialization cost



Query Execution

Expression Optimizations

Operate directly on binary data
JVM intrinsics
Hot method codegen

Performant Operators

Operator codegen
HashAgg
Improved HashJoin
Semi/Anti join

Resource Optimizations

Stats based estimation
Dynamic memory allocation

Query Optimization

Alibaba Group

Cost Based

- Join order
- Join type
- Agg strategy
- ……

Advanced Rules

- Subplan reuse
- Join condition expansion
- Shuffle removal
- Distinct Agg rewrite
- ….

Rich Stats

- NDV
- NULL count
- Avg length
- Max length
- Min
- Max

Performance Improvements - Recap

10x *Overall* = **2** *Runtime* x **2+** *Query Execution* x **2+** *Query Optimizer*

Unified Engine

Functionality

Performance

Reliability

Reliability Improvements

Failover

- Region Based Failover
- JM Failover
- Blacklist
- ….

Shuffle

- Decoupled from TM
- Yarn Shuffle Service
- Async mode

More Details?

Sept 4th, 2018
5:10 PM - 5:50 PM
Maschinenhaus
Feng Wang, Alibaba

Runtime Improvements for Flink Batch Processing

Next Steps

Next Steps

Grand Unification of Data Processing

Switch between batch processing and streaming seamlessly

Flink Machine Learning/AI

PyFlink, TableAPI, DL Integration, Flink ML Improvements

Takeaway

Alibaba Group

Already the best stream processor

Becoming the best batch processor

Unified approach benefits both batch & stream processing

Working on seamless experience for AI

Flink Forward China

Flink Forward China

Dec 20th-21st @ Beijing National Conference Center

First Flink Forward Conference in Asia, 3000+ participants expected

Flink Community

Joint efforts by all major players in Flink community from China

Call For Talks & Sponsors

Submit your talks to flink-forward-china@list.alibaba-inc.com