阿里巴巴在线技术峰会
Alibaba Online Technology Summit

Blink计算引擎

蒋晓伟
xiaowei.jxw@alibaba-inc.com

Alibaba Group 阿里巴巴集团 阿里云 aliyun.com

个人简介

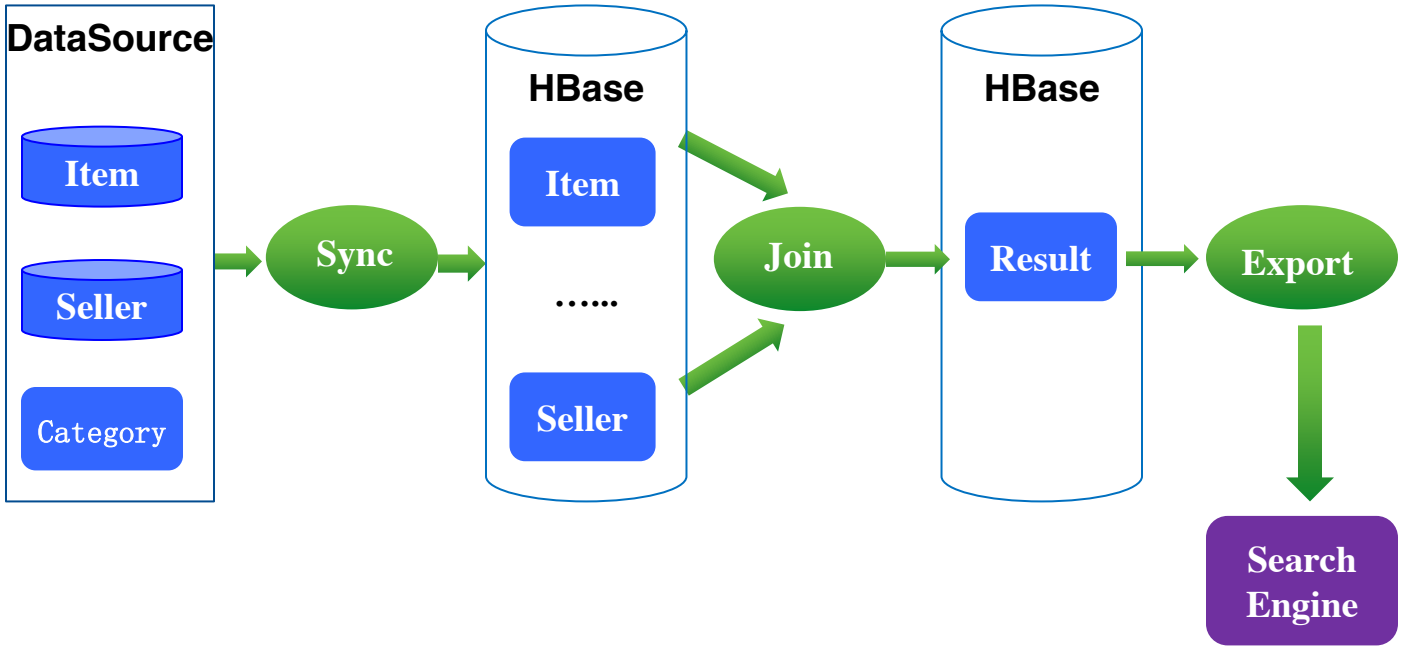
- 蒋晓伟
  - 2014 –– 现在 阿里巴巴
  - 2010 –– 2014 脸书
  - 2002 –– 2010 微软
  - 2000 –– 2002 Stratify

提纲

- 背景和用例
- 什么是Blink?
- Blink的改进
- 现状和计划

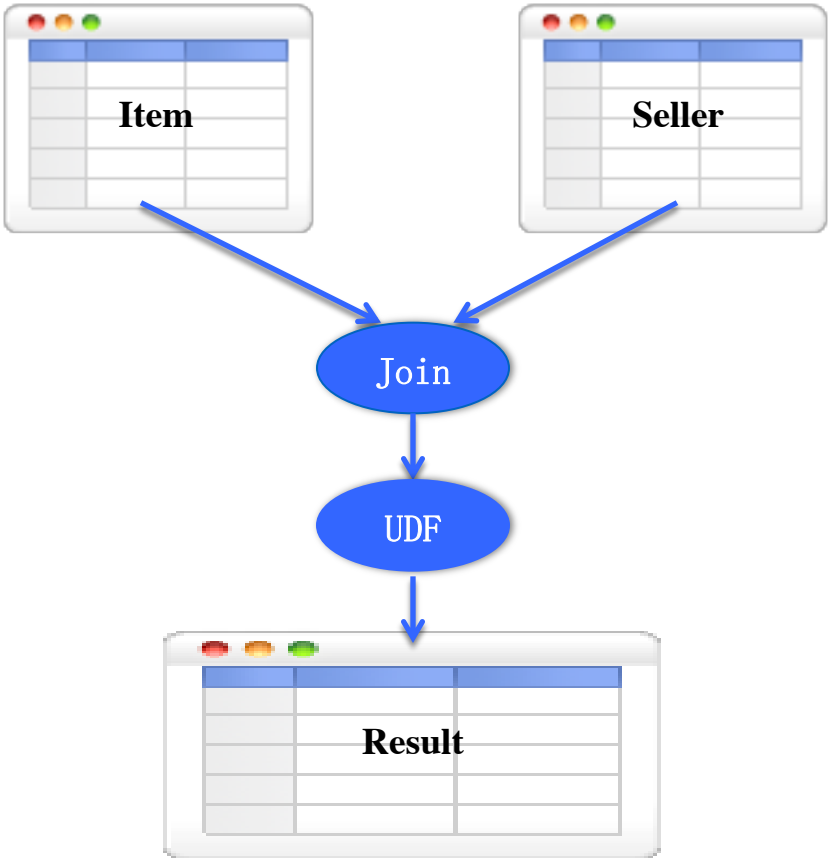
用例 - 搜索文档的创建和更新

- 开发效率
  - 全量增量一套代码
  - 高层次API
- 一致性
  - 至少一次
  - 恰好一次
- 低延迟
  - 亚秒级
- 成本
  - 高吞吐



全量增量一体化的抽象

```
CREATE MATERIALIZED VIEW Result AS
SELECT *, UDF(a, b, c)
    FROM Item JOIN Seller
    ON Item.uid=Seller.id
```

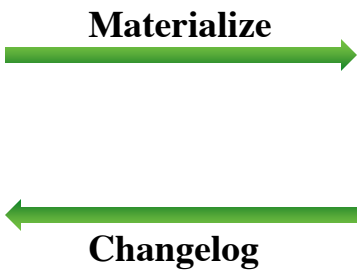


- 结果表 = 物化视图
  - 全量 – 索引的创建和重建
  - 增量 – 索引的维护

流和表的对偶性

流

word	count
Hello	1
World	1
Hello	2
Bark	1
Hello	3



历史表

word	count
Hello	3
World	1
Bark	1

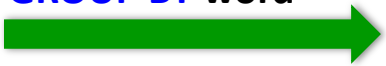
流的等价性: 两个流等价 <==> 它们产生相同的正则历史表

Word Count 例子(批处理)

Words

time	word
1	Hello
2	World
3	Hello
4	Hello
5	Bark

```
SELECT word, count(*)
FROM Words
GROUP BY word
```



WordCount

word	count
Hello	3
World	1
Bark	1

```
SELECT sum(count)
FROM WordCount
```

Total

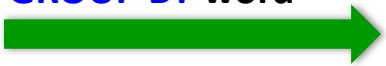
total
5

Word Count 例子(流处理)

Words

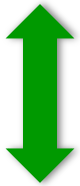
time	word
1	Hello
2	World
3	Hello
4	Hello
5	Bark

```
SELECT word, count(*)
FROM Words
GROUP BY word
```



WordCount

word	count
Hello	1
World	1
Hello	2
Hello	3
Bark	1



word	count
Hello	3
World	1
Bark	1

```
SELECT sum(count)
FROM WordCount
```

Total

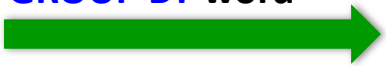
total
1
2
4
7
8

Word Count 例子(Retraction)

Words

time	word
1	Hello
2	World
3	Hello
4	Hello
5	Bark

```
SELECT word, count(*)
FROM Words
GROUP BY word
```



WordCount

word	count
Hello	1
World	1
Hello	1
Hello	2
Hello	2
Hello	3
Bark	1

word	count
Hello	3
World	1
Bark	1

```
SELECT sum(count)
FROM WordCount
```

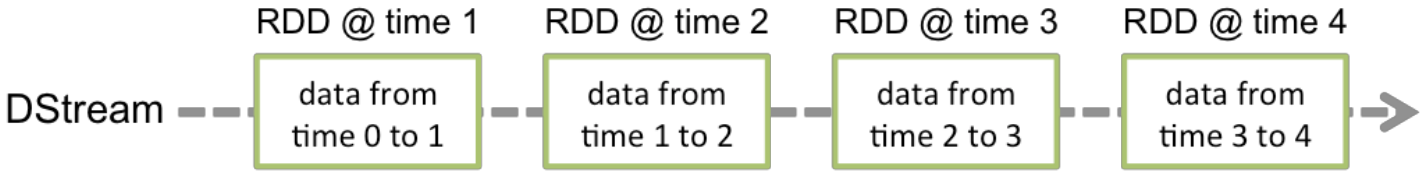
Total

total
1
2
1
3
1
4
5

total
5

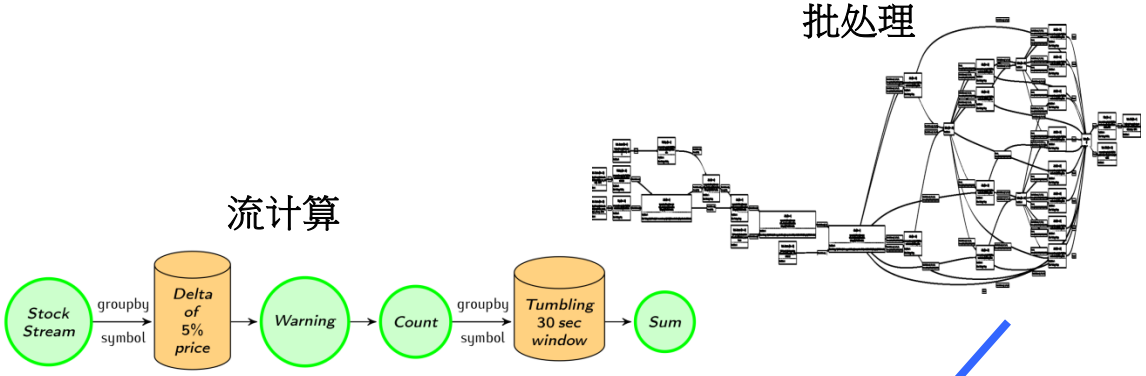
Apache Spark

- Spark
  - 高层次的API - 开发效率
  - Tungsten - 很多性能优化
  - 较活跃的生态
- Spark Streaming
  - Micro-batching
  - 很难做到亚秒级延迟



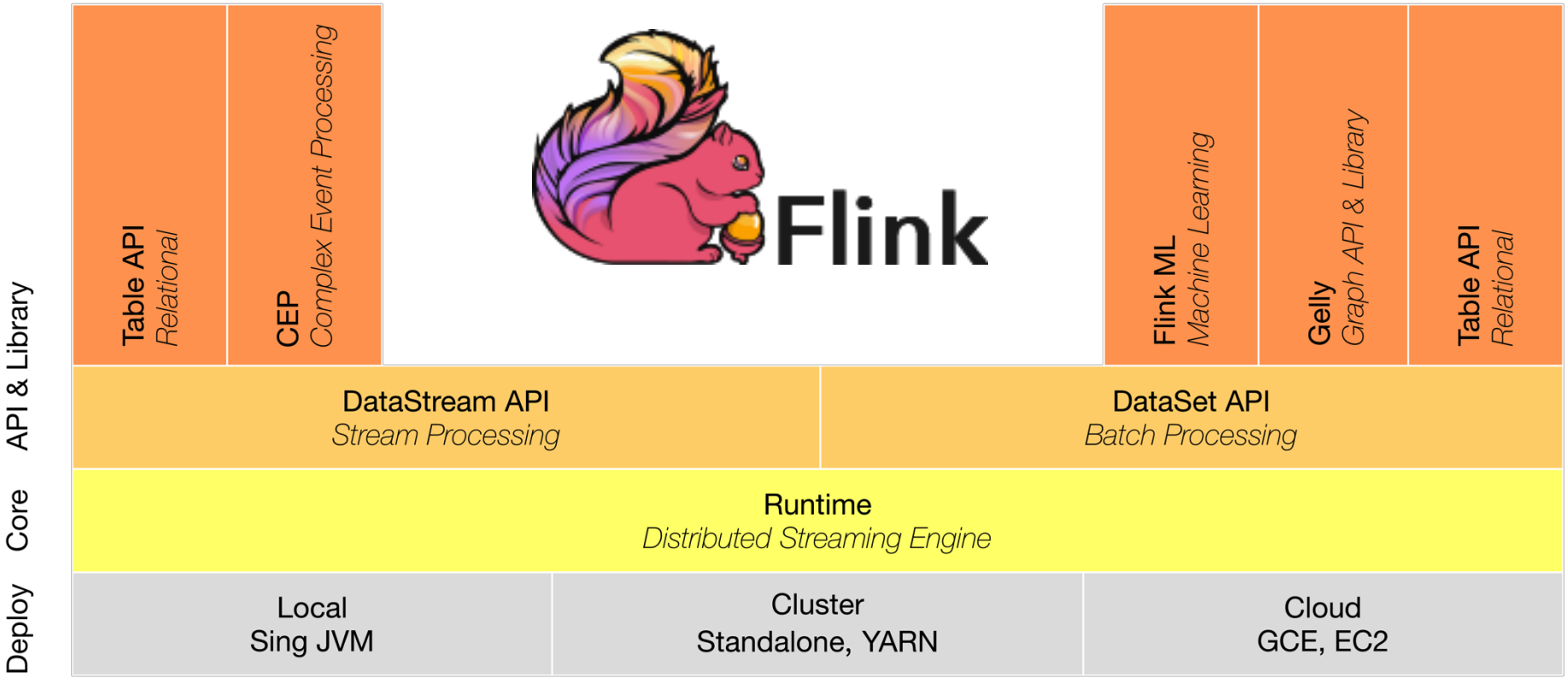
Apache Flink

- 流和批的一体化计算引擎
- 开发效率：高层次API
- 一致性：有状态的计算
- 低延迟
- 高吞吐




```
DataStream<String> text = [...]
DataStream<Tuple2<String, Integer>> wordCounts = text
    .flatMap(new LineSplitter())
    .keyBy(0)
    .sum(1);
```

Flink生态


API & Library
Core
Deploy
Table API
Relational
CEP
Complex Event Processing
Flink
Flink ML
Machine Learning
Gelly
Graph API & Library
Table API
Relational
DataStream API
Stream Processing
DataSet API
Batch Processing
Runtime
Distributed Streaming Engine
Local
Sing JVM
Cluster
Standalone, YARN
Cloud
GCE, EC2

什么是Blink?

- Blink – 阿里巴巴基于Flink开发的计算引擎
  - 批和流一体化的完备的Table API
  - 和Flink API以及生态兼容的重新开发的Runtime

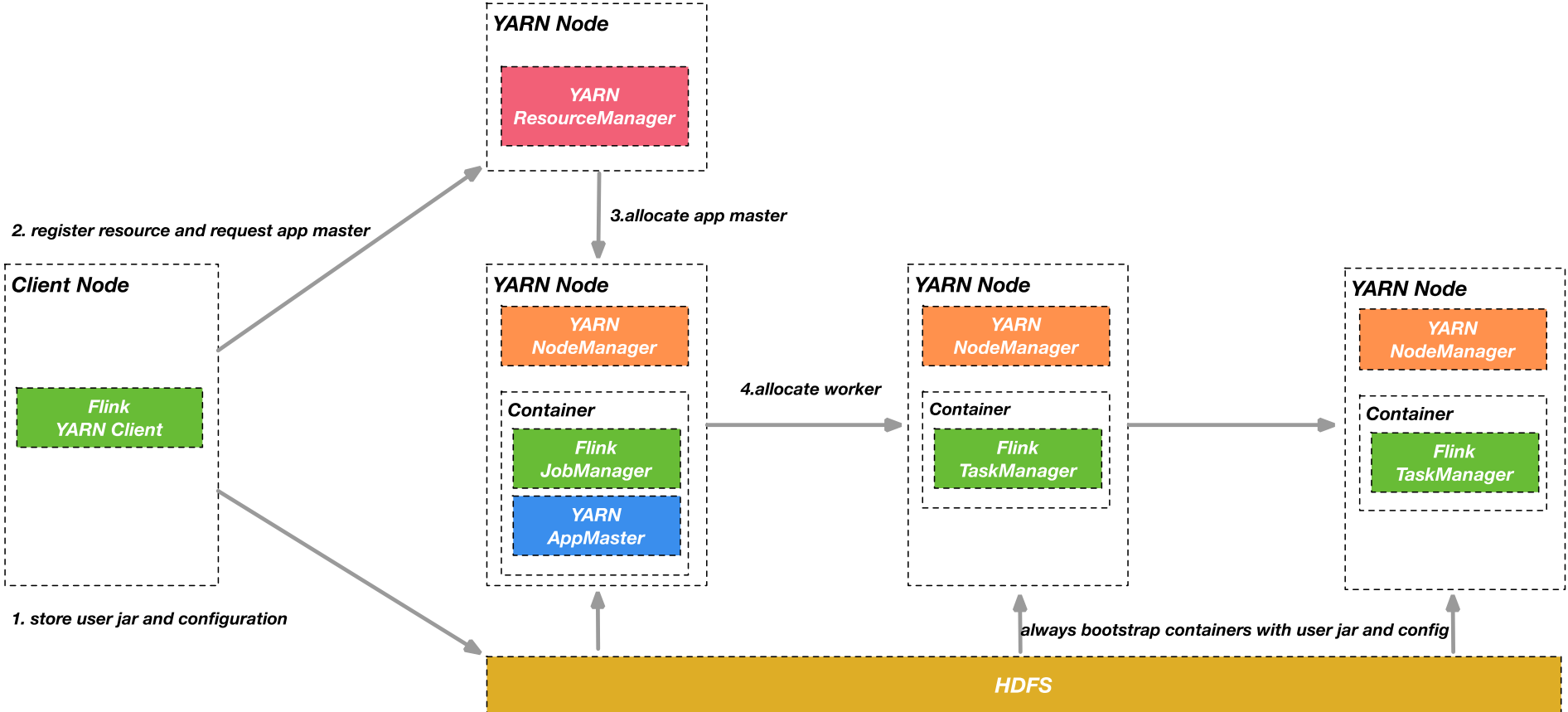
Blink的Table API

- 原则 – 流和批的一体化处理
- 功能
  - UDF/UDTF/UDAGG
  - 双流Join
  - Aggregation(min, max, avg, sum, count, distinct_count)
  - Windowing (time_window, count_window)
  - 撤回（Retraction）

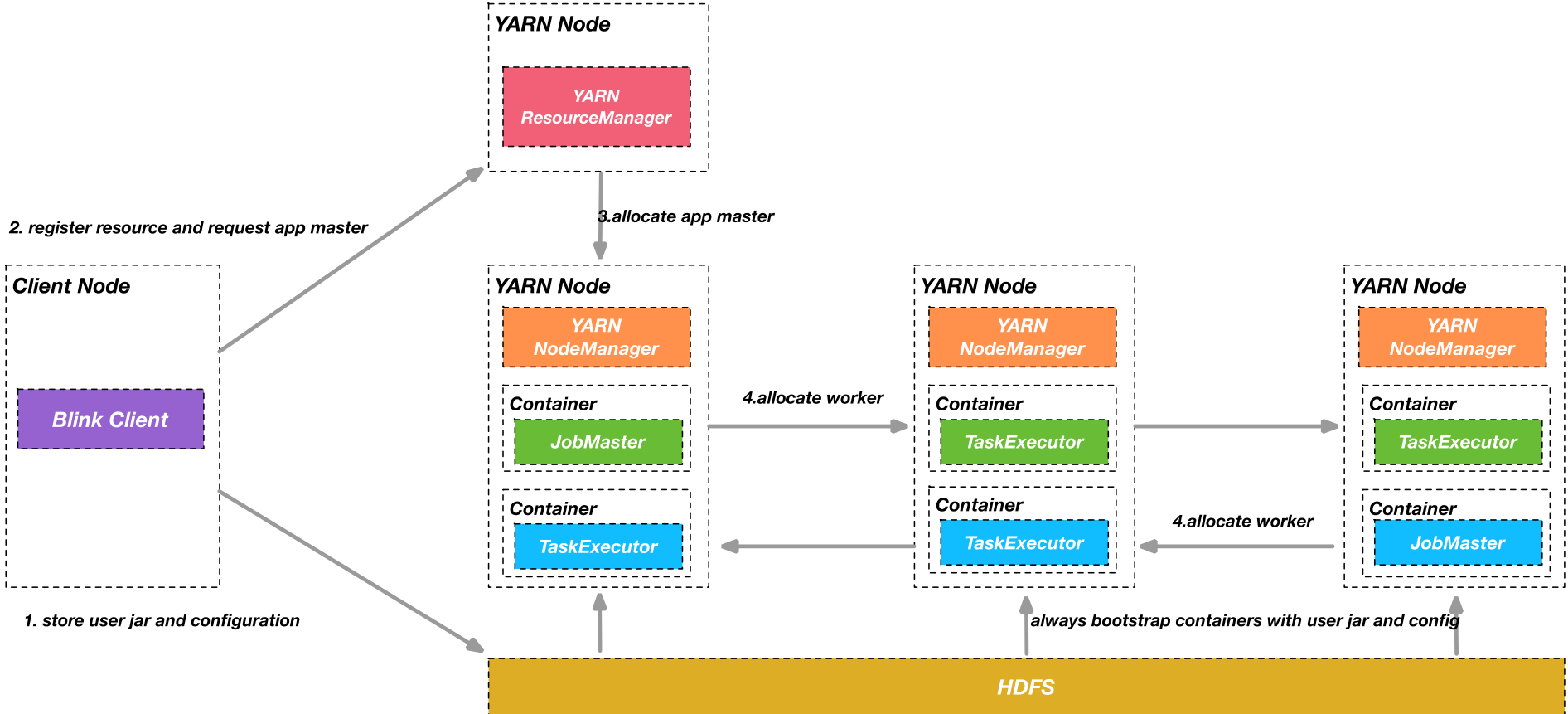
Blink的Runtime

- 和YARN的原生态整合
- Checkpoint和状态管理的优化
- 容错性和高可用性
- 动态伸缩
- 稳定性和可运维性的改进

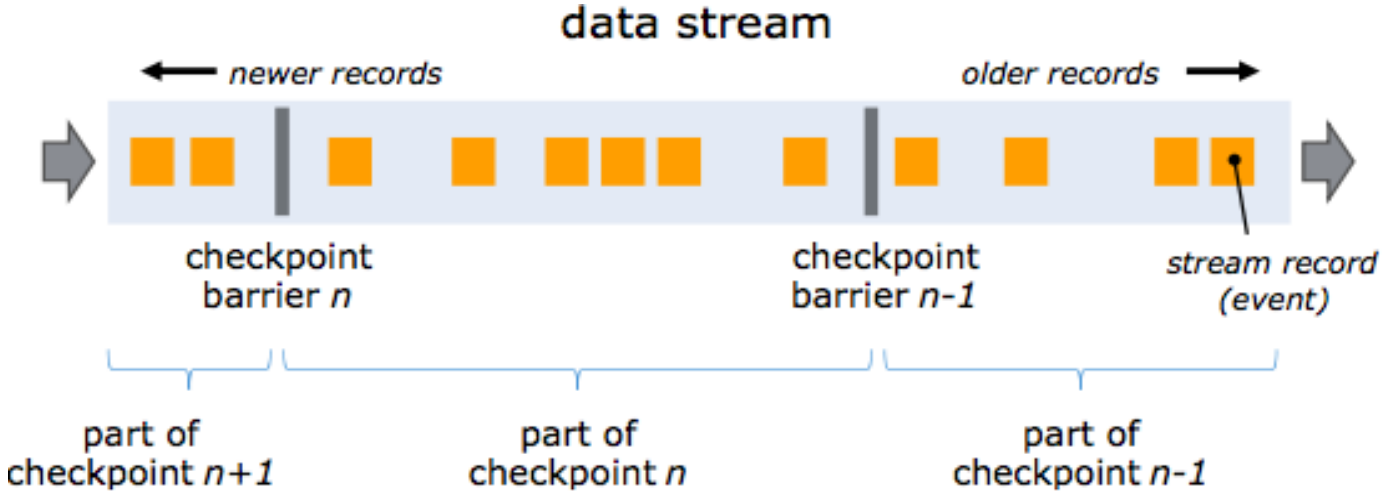
Flink on YARN

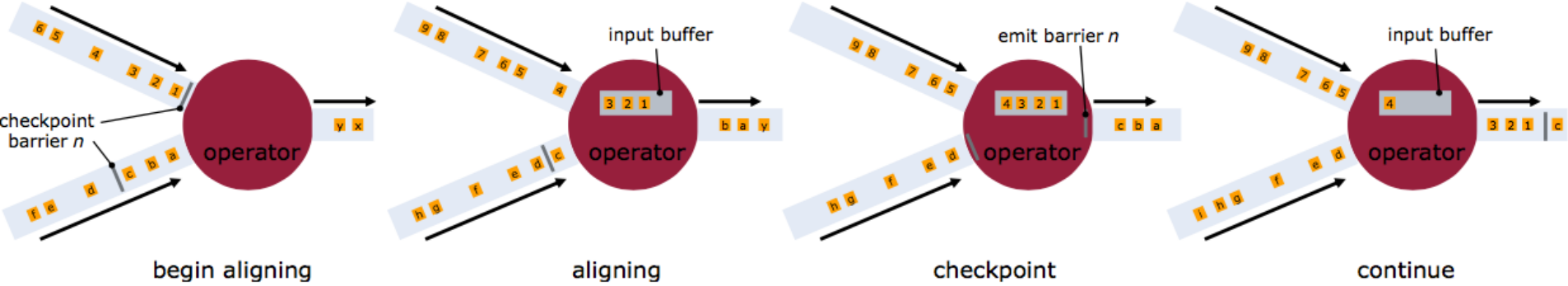


Blink YARN的原生态整合

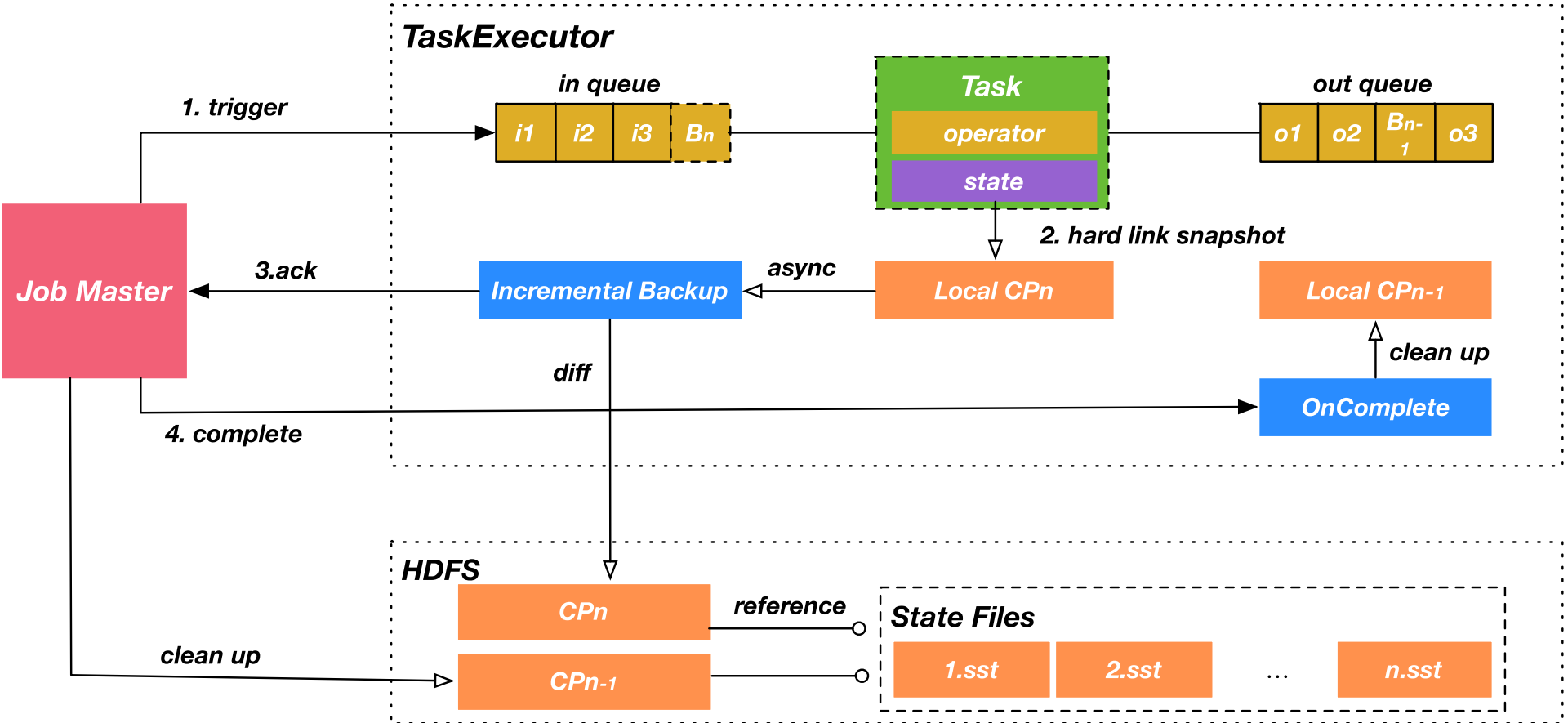


Chandy-Lamport Checkpoint

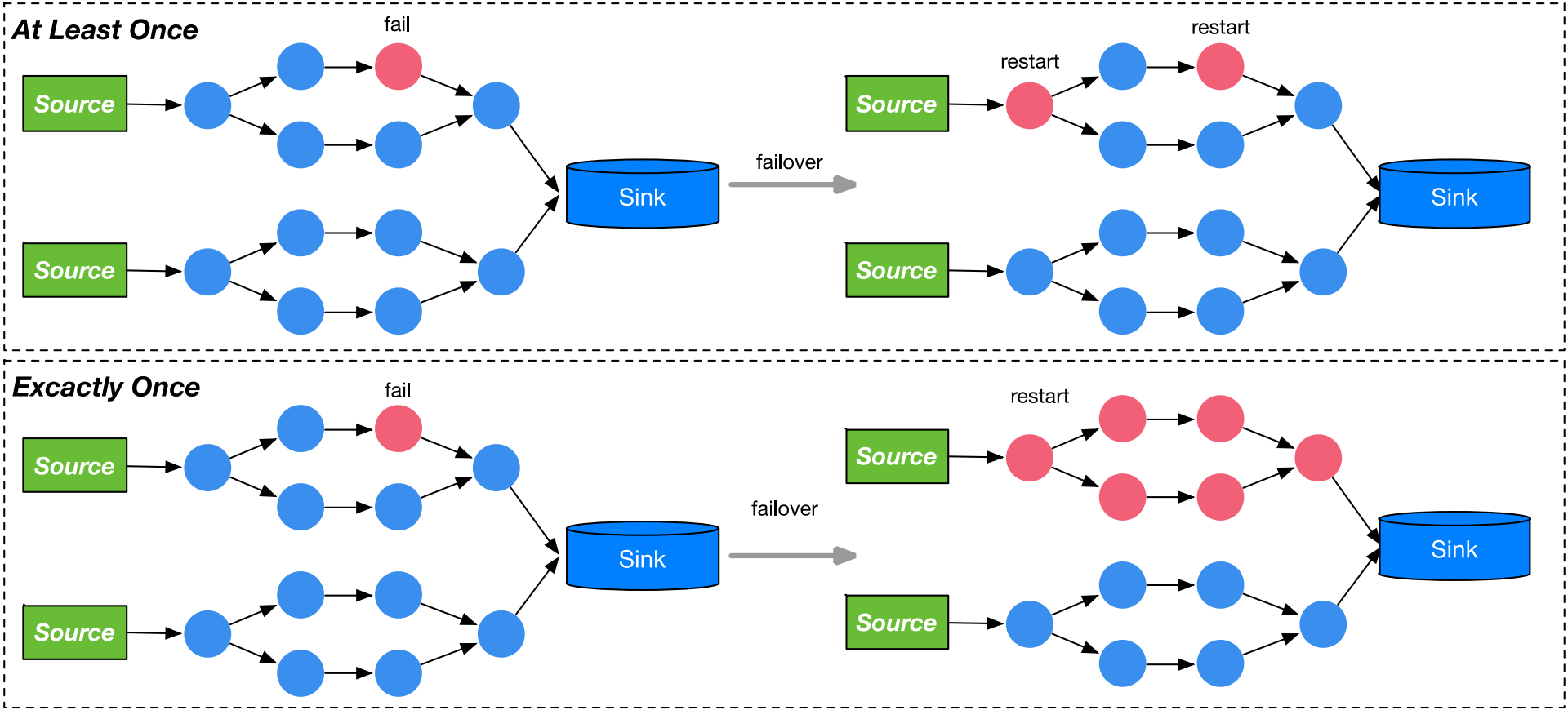




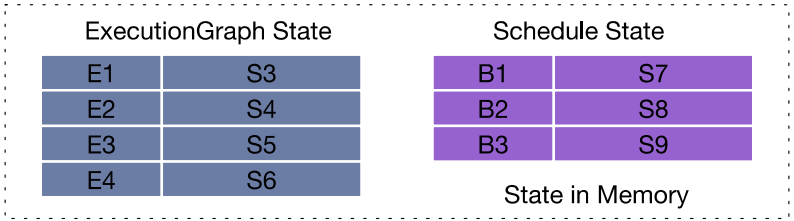
Blink的Checkpoint和状态管理

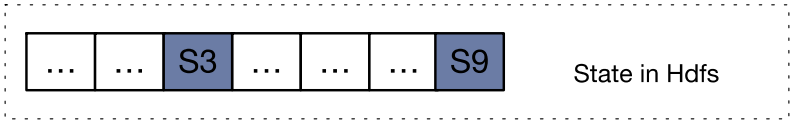


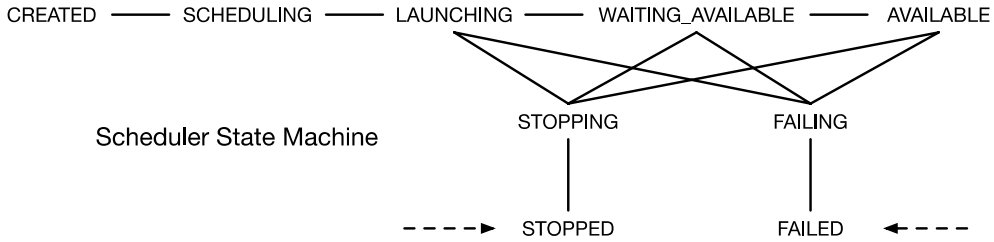
Blink Worker的错误恢复

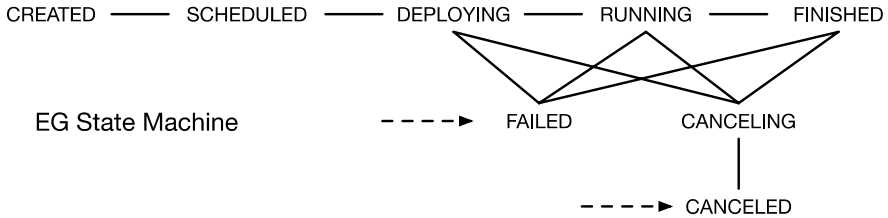


Blink Master的高可用性

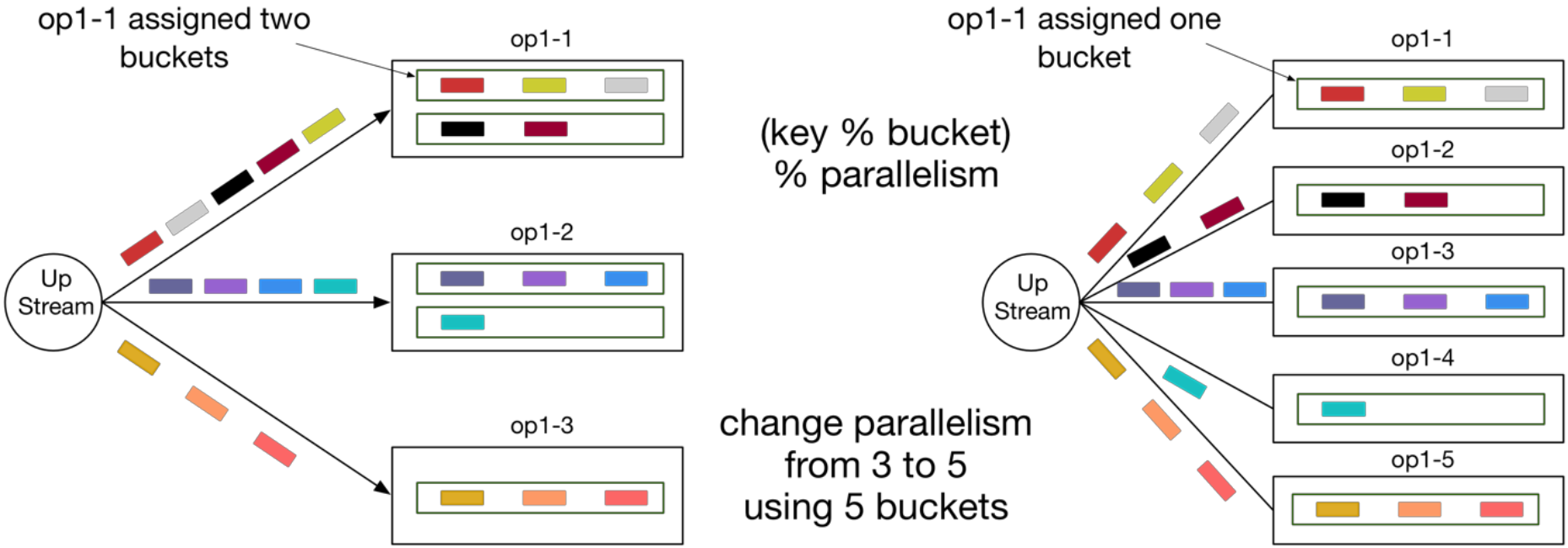




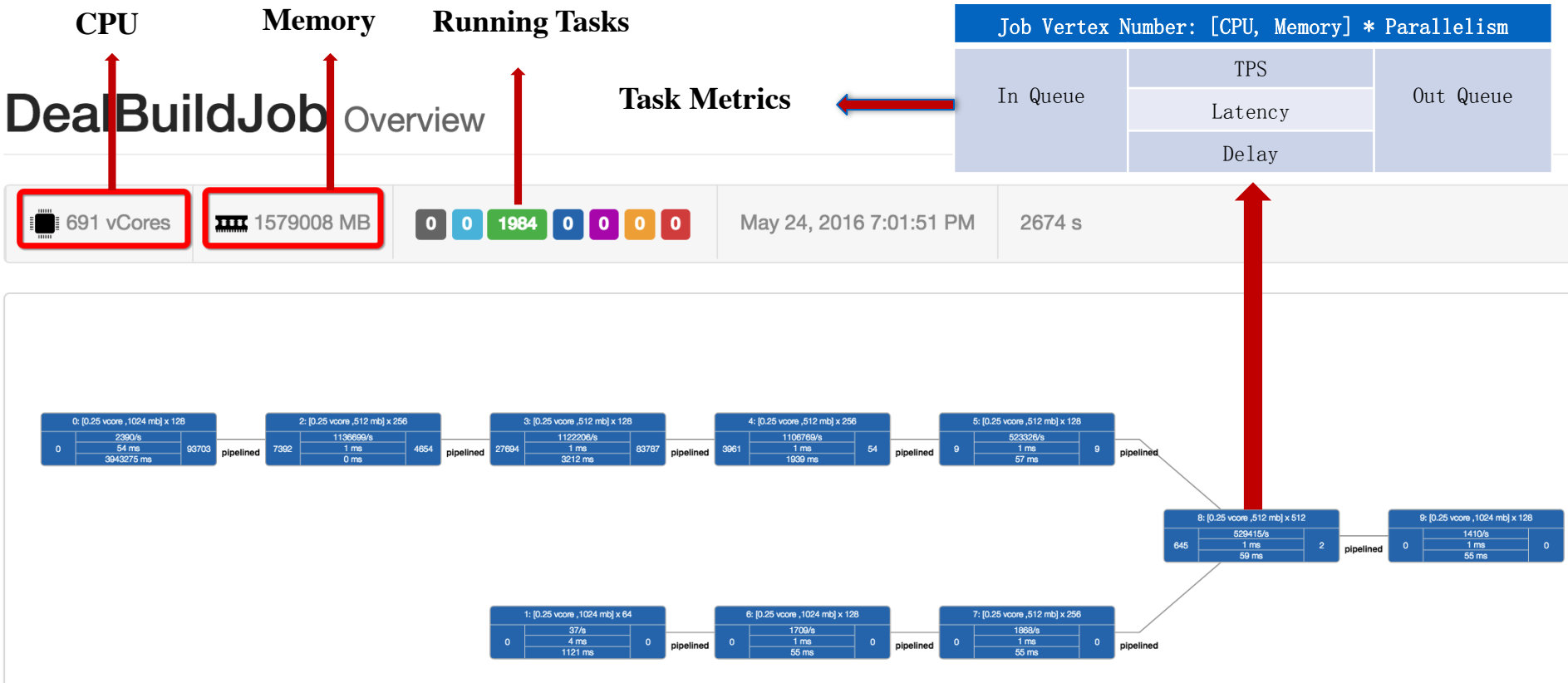




Blink的动态伸缩


op1-1 assigned two buckets
op1-1
op1-2
op1-3
Up Stream
(key % bucket) % parallelism
op1-1 assigned one bucket
op1-1
op1-2
op1-3
op1-4
op1-5
Up Stream
change parallelism from 3 to 5 using 5 buckets

Blink监控


CPU
Memory
Running Tasks
Task Metrics
Job Vertex Number: [CPU, Memory] * Parallelism
In Queue
TPS
Latency
Delay
Out Queue
DealBuildJob Overview
691 vCores
1579008 MB
0 0 1984 0 0 0 0
May 24, 2016 7:01:51 PM
2674 s
0: [0.25 vcore ,1024 mb] x 128
2390/s
54 ms
3943275 ms
93703
pipelined
2: [0.25 vcore ,512 mb] x 256
7392
1136699/s
1 ms
0 ms
4654
pipelined
3: [0.25 vcore ,512 mb] x 128
27694
1122206/s
1 ms
3212 ms
83787
pipelined
4: [0.25 vcore ,512 mb] x 256
3961
1106769/s
1 ms
1939 ms
54
pipelined
5: [0.25 vcore ,512 mb] x 128
9
523326/s
1 ms
57 ms
9
pipelined
8: [0.25 vcore ,512 mb] x 512
645
529415/s
1 ms
59 ms
2
pipelined
9: [0.25 vcore ,1024 mb] x 128
0
1410/s
1 ms
55 ms
0
1: [0.25 vcore ,1024 mb] x 64
0
37/s
4 ms
1121 ms
0
pipelined
6: [0.25 vcore ,1024 mb] x 128
0
1709/s
1 ms
55 ms
0
pipelined
7: [0.25 vcore ,512 mb] x 256
0
1868/s
1 ms
55 ms
0
pipelined

现状和计划

- 已在上千台机器的集群上线
- 支持搜索和推荐等核心业务
- 集团内外的浓厚兴趣
  - NetFlix, Airbnb, Uber, Facebook
- 反馈给社区


Thank you !