分布式快照：灵活可靠地处理 Apache Flink消息

万达网络科技集团大数据中心 大数据技术专家

李呈祥

概要

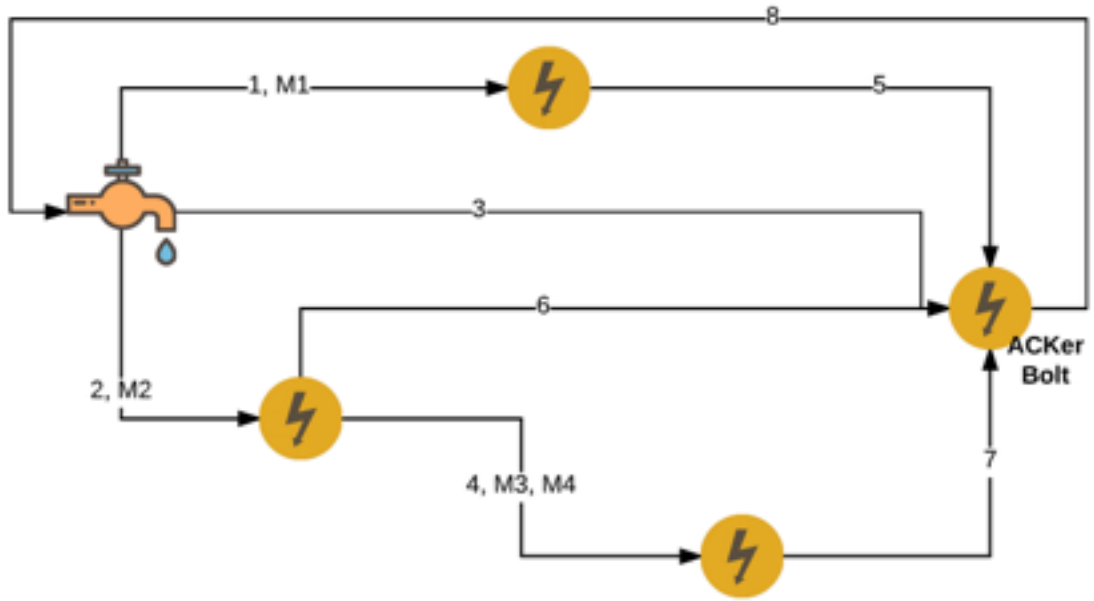
- 常见流处理框架的可靠性保障机制
- Flink分布式快照介绍
- 正确性与Exactly-Once消息处理

保证结果正确性是流计算框架在实际业务中应用的必要条件

ACK机制

Storm

- 保证at-least-once。
- 不保证中间状态结果。
- 对业务逻辑实现和性能都有影响。

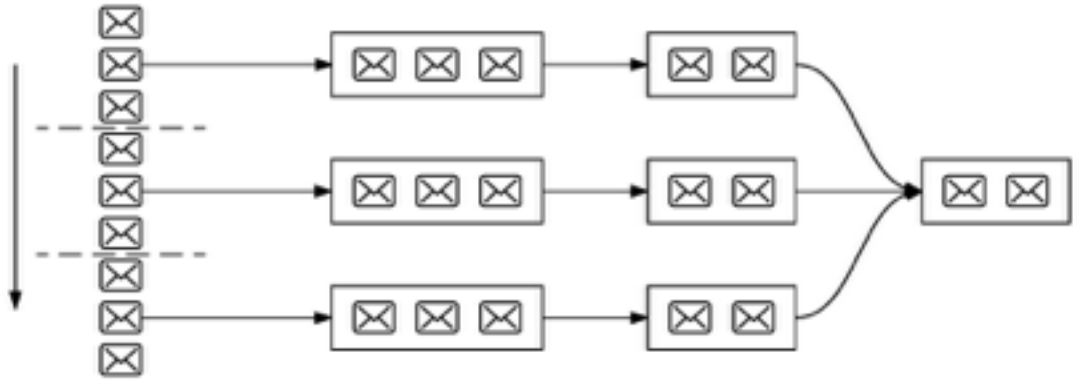


3: RootID -> M1 ^ M2
5: RootID -> (M1 ^ M2) ^ M1
6: RootID -> (M1 ^ M2) ^ M1 ^ (M2 ^ M3 ^ M4)
7: RootID -> (M1 ^ M2) ^ M1 ^ (M2 ^ M3 ^ M4) ^ (M3 ^ M4)
8: RootID -> 0

Micro Batch

Spark Streaming

Storm Trident

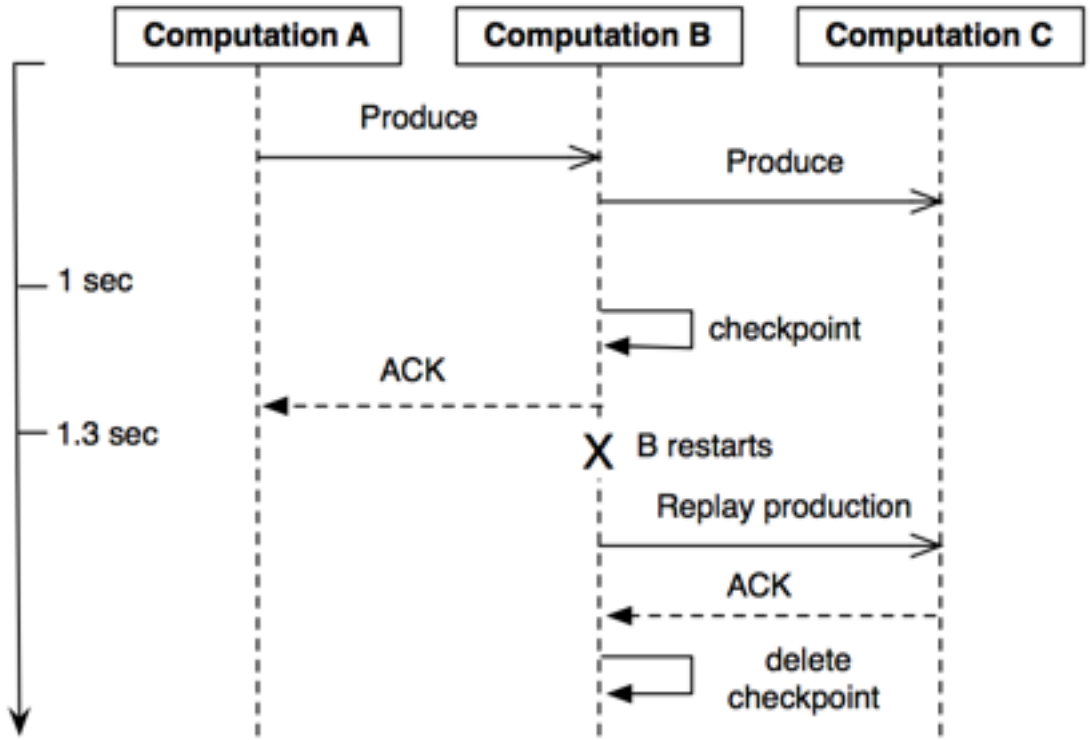


- 保证Exactly-Once消息处理
- 保证中间状态结果。
- 延迟，编程模型，流控

Transaction Update

Google Dataflow

Smaza

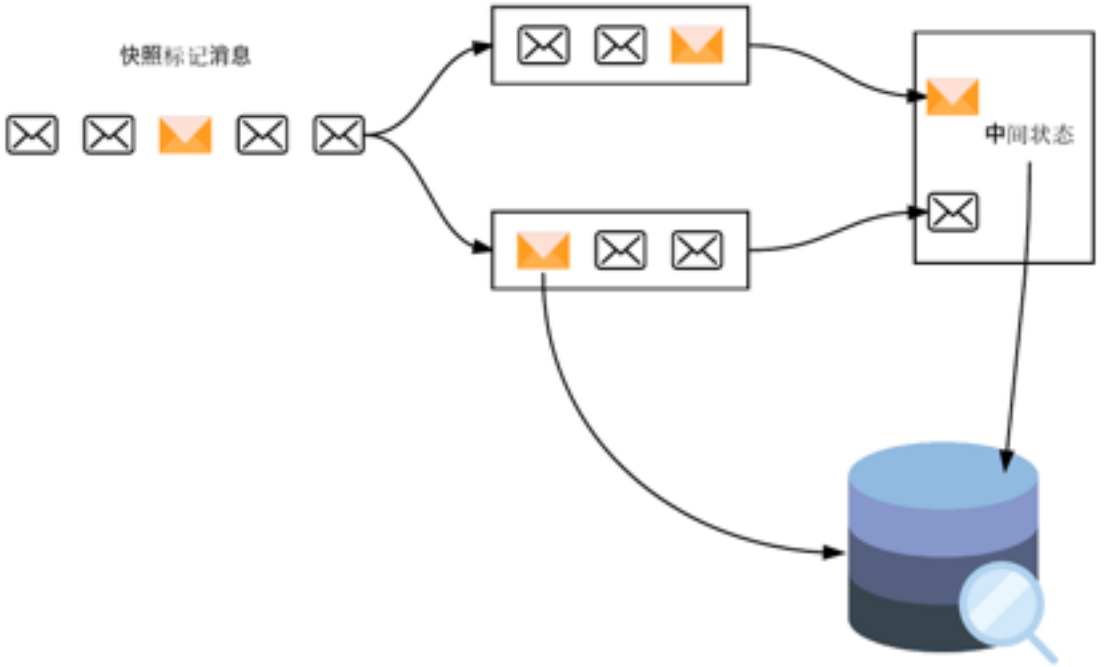


- 保证Exactly-Once消息处理
- 保证中间状态结果。
- 依赖写性能

分布式快照

Flink

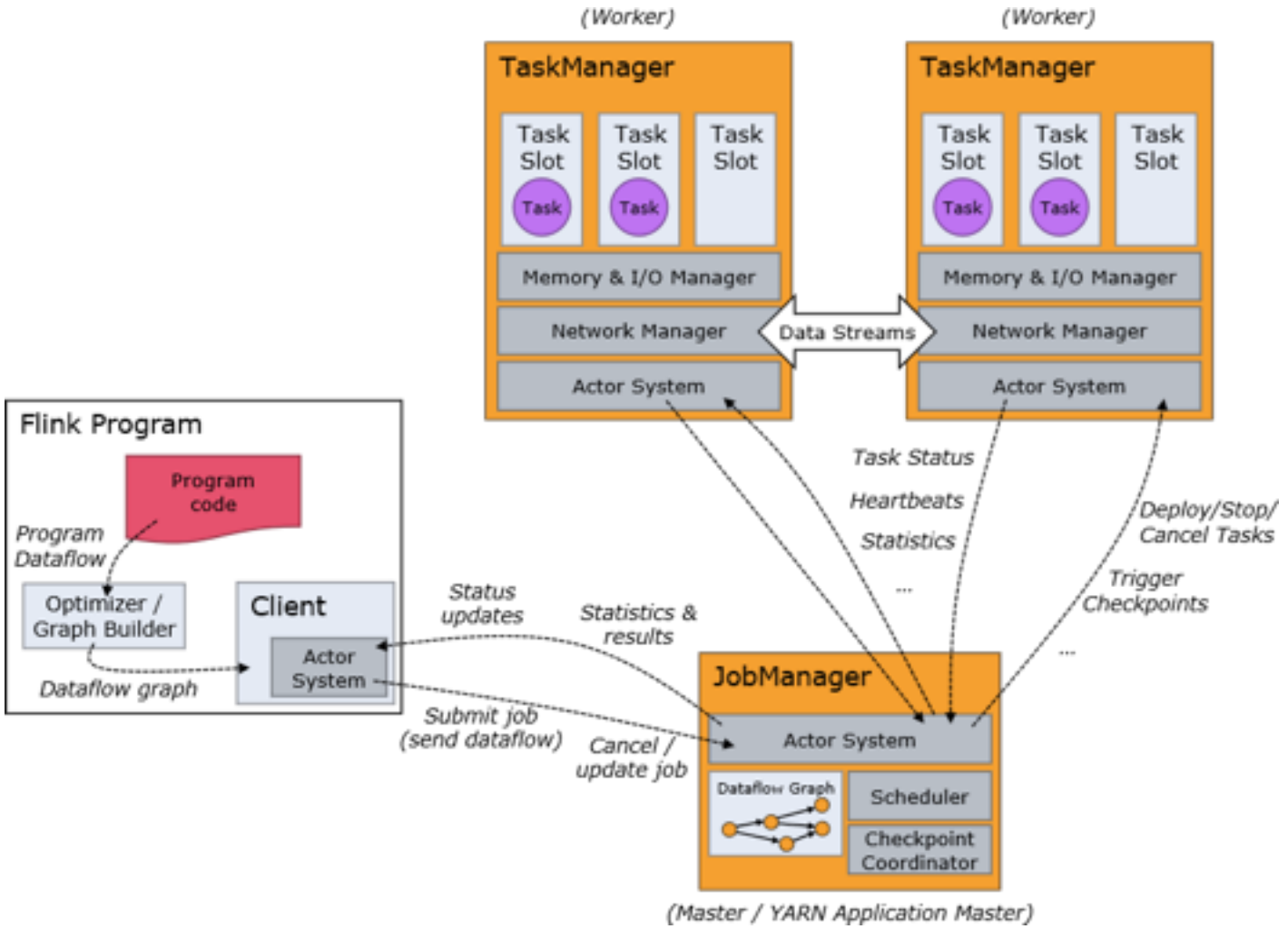
Gearpump



1. 定期保存分布式快照。
2. 发生节点宕机等异常。
3. 重启所有task,恢复到上一次分布式快照状态。
4. 从Source重发分布式快照之后的消息。

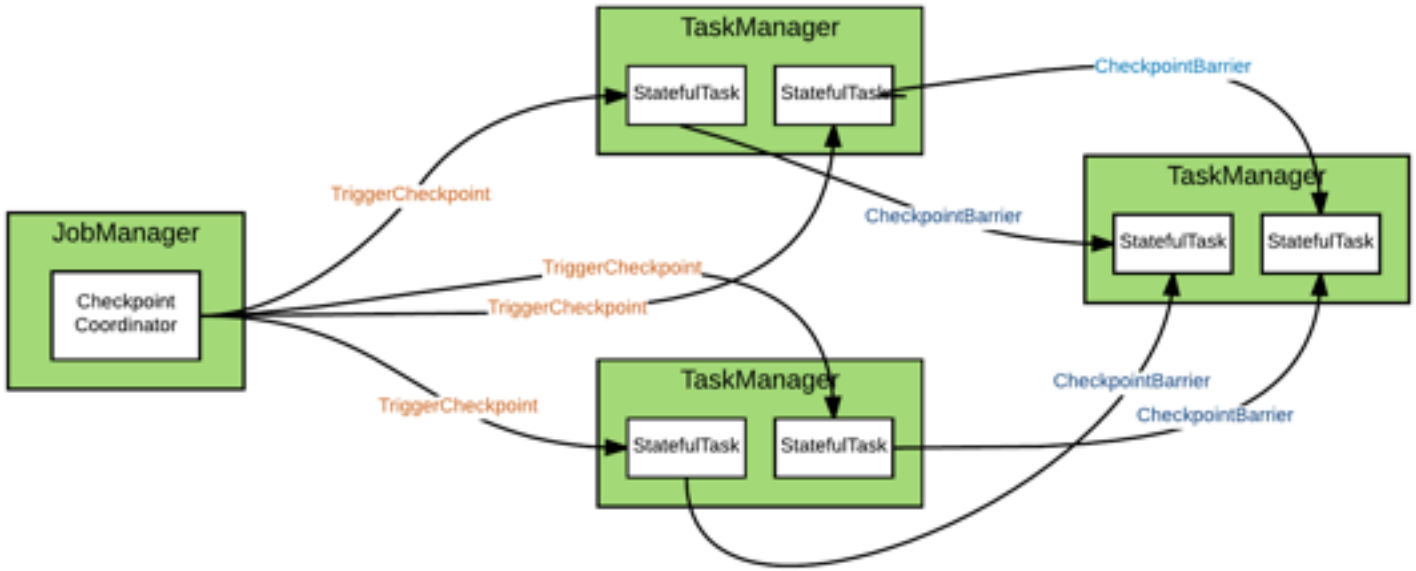
分布式快照 = 部分数据 * 全部计算

Flink 架构


(Worker)
TaskManager
Task Slot
Task
Memory & I/O Manager
Network Manager
Actor System
Data Streams
(Worker)
TaskManager
Flink Program
Program code
Program Dataflow
Optimizer / Graph Builder
Client
Actor System
Dataflow graph
Status updates
Statistics & results
Submit job (send dataflow)
Cancel / update job
Task Status
Heartbeats
Statistics
...
Deploy/Stop/ Cancel Tasks
Trigger Checkpoints
...
JobManager
Actor System
Dataflow Graph
Scheduler
Checkpoint Coordinator
(Master / YARN Application Master)

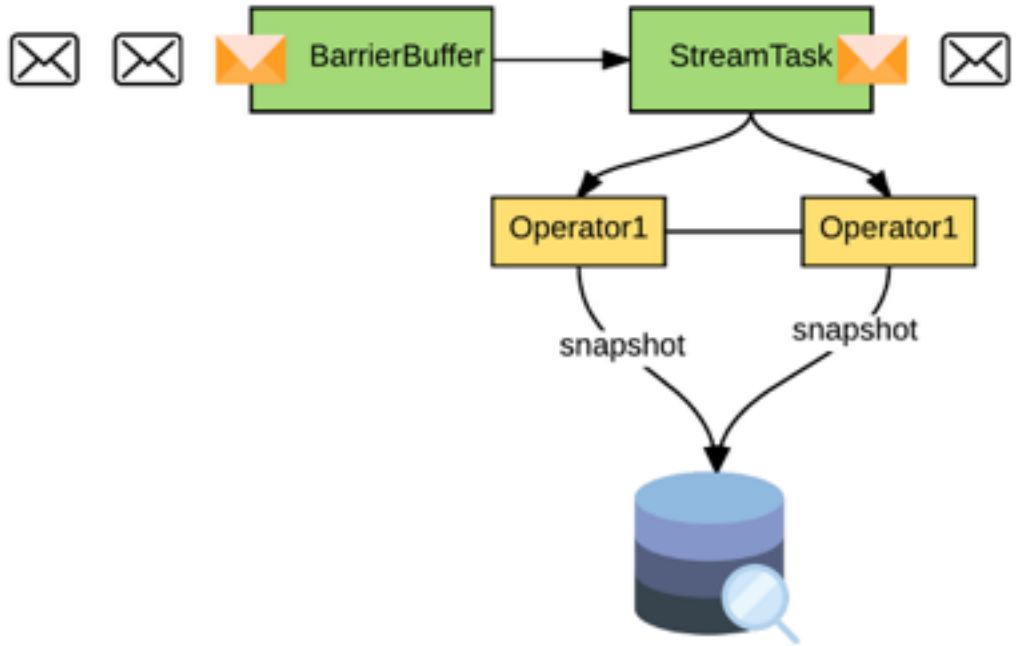
分布式快照：触发阶段

- Trigger Vertices: Input vertices
- ACK Vertices: all job vertices
- Commit Vertices: all job vertices



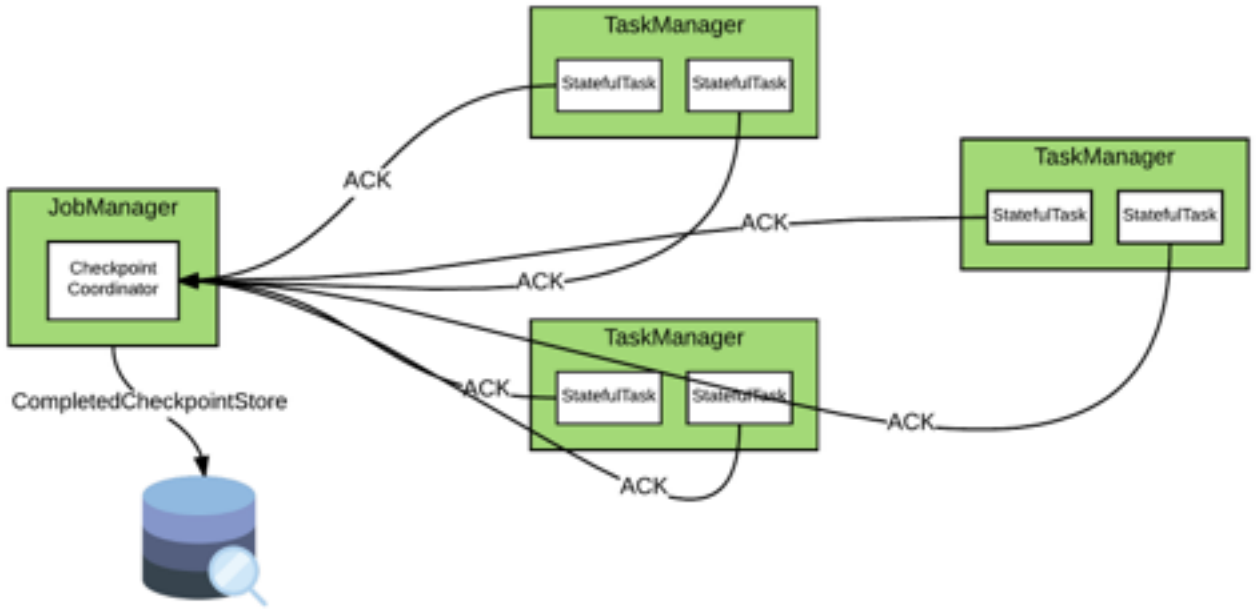
分布式快照：快照阶段

- BarrierBuffer: 处理所有Task的消息
- StreamTask：通知所有Operator快照
- Operator：执行UDF自定义snapshot方法



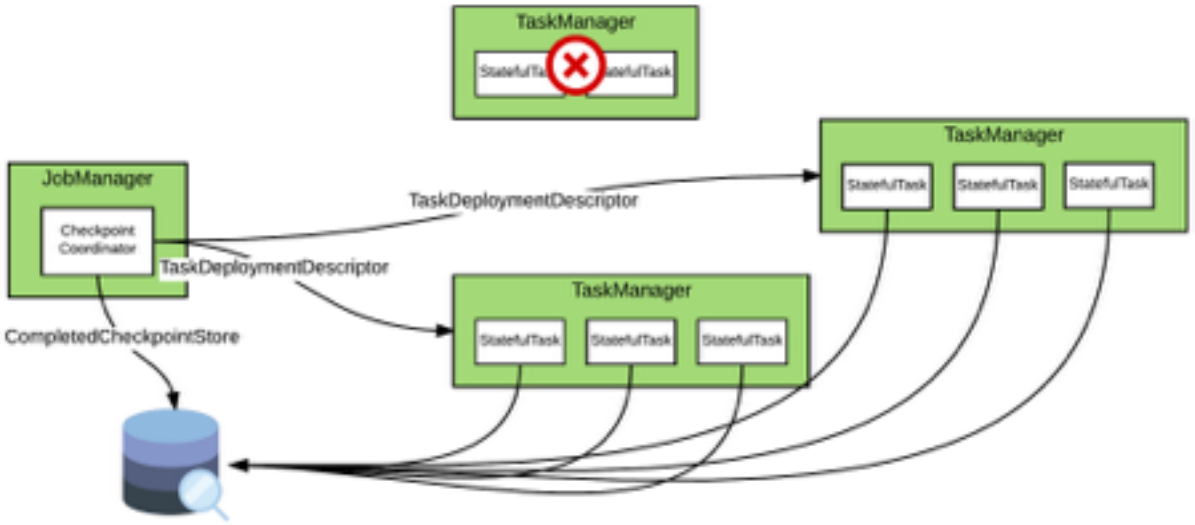
分布式快照：响应阶段

- Task: 完成快照后发送ACK
- CheckpointCoordinator：收到所有task快照后，将快照信息记录下来
- CheckpointCoordinator：发送commit信息给所有task



分布式快照：恢复阶段

- TaskManager崩溃
- CheckpointCoordinator：读取最近一次成功快照信息
- JobManager：将每个Task的快照信息保存在TaskDeploymentDescriptor中，重启所有Task
- TaskManager：Task启动时首先从存储中加载快照数据。

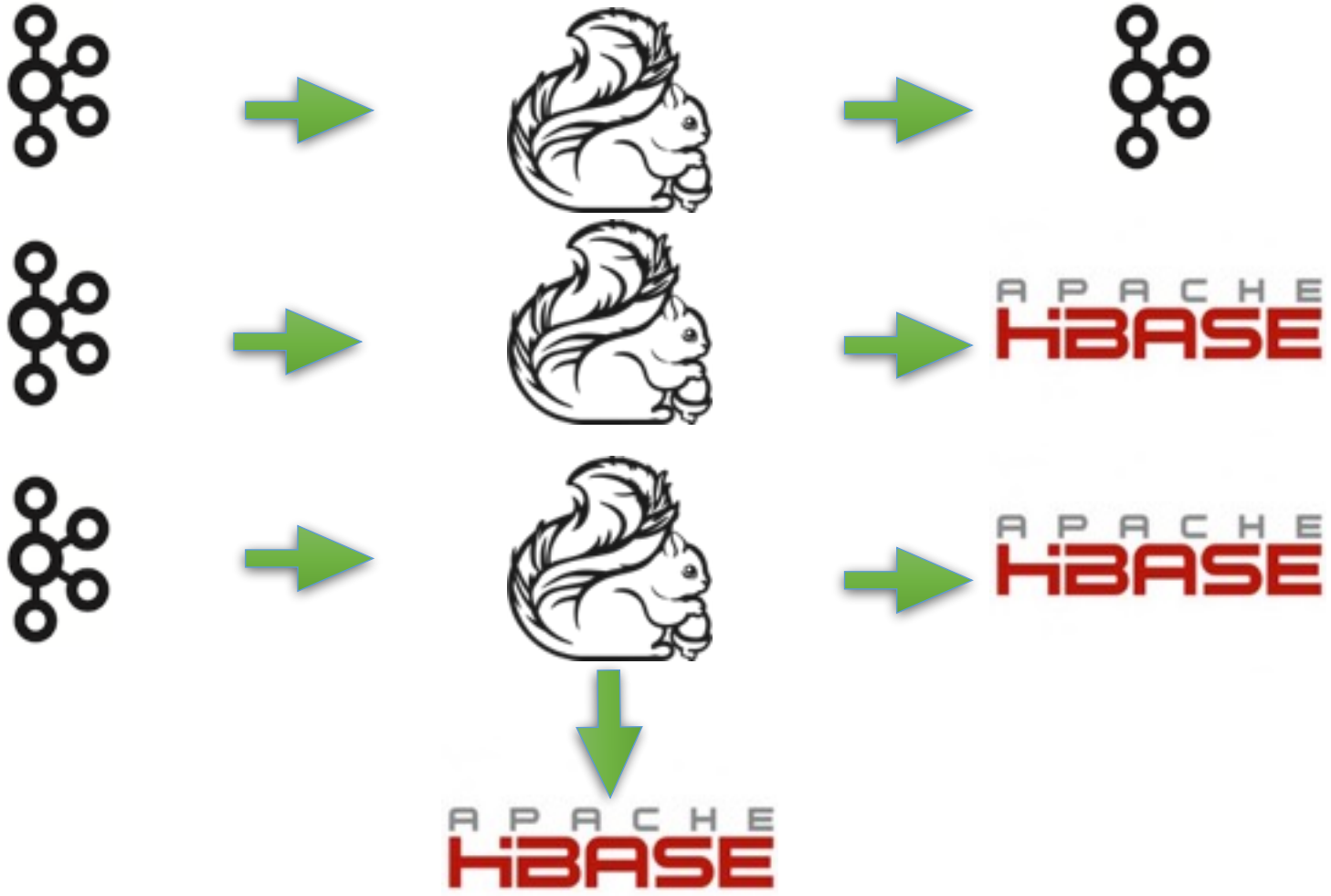


分布式快照：总结

- 正确性保证
- 低延迟
- 高吞吐量
- 强大的编程范式
  - 正确性保证的设计不影响流处理编程范式的实现
  - 对业务逻辑无侵入
- 较低的系统代价
- 流控

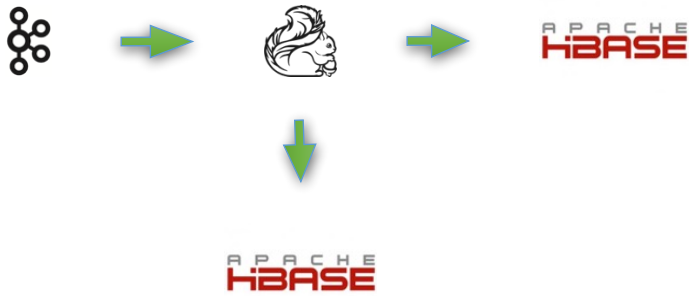
端到端的Exactly-Once业务实现

实际Flink应用



End to end exactly-once

Rolling update，唯一标识符作为Rowkey



Incremental update

- 消息包含唯一标识符,且有序。
- 以消息唯一标识符作为Hbase表的一个字段,
- 根据消息中标识符值与表中标识符值判断消息是否处理完成。

唯一标识符作为Key，Kafka Consumer后续处理

Q&A