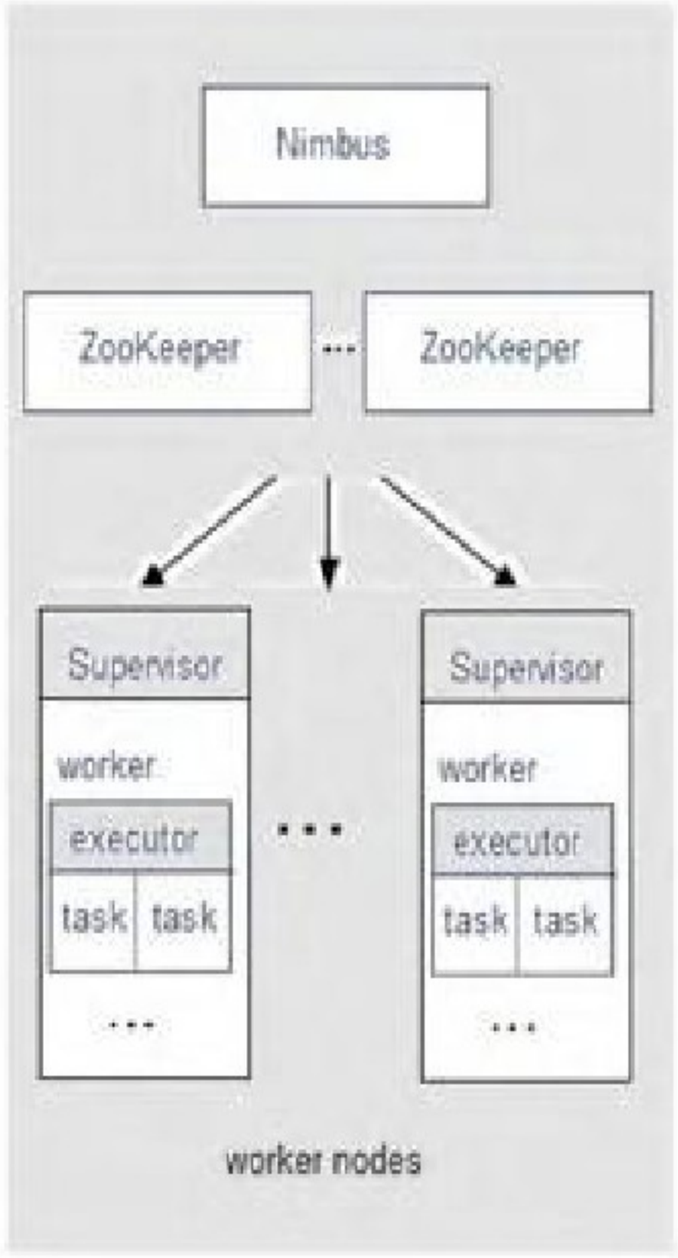
轻量级流式计算框架介绍

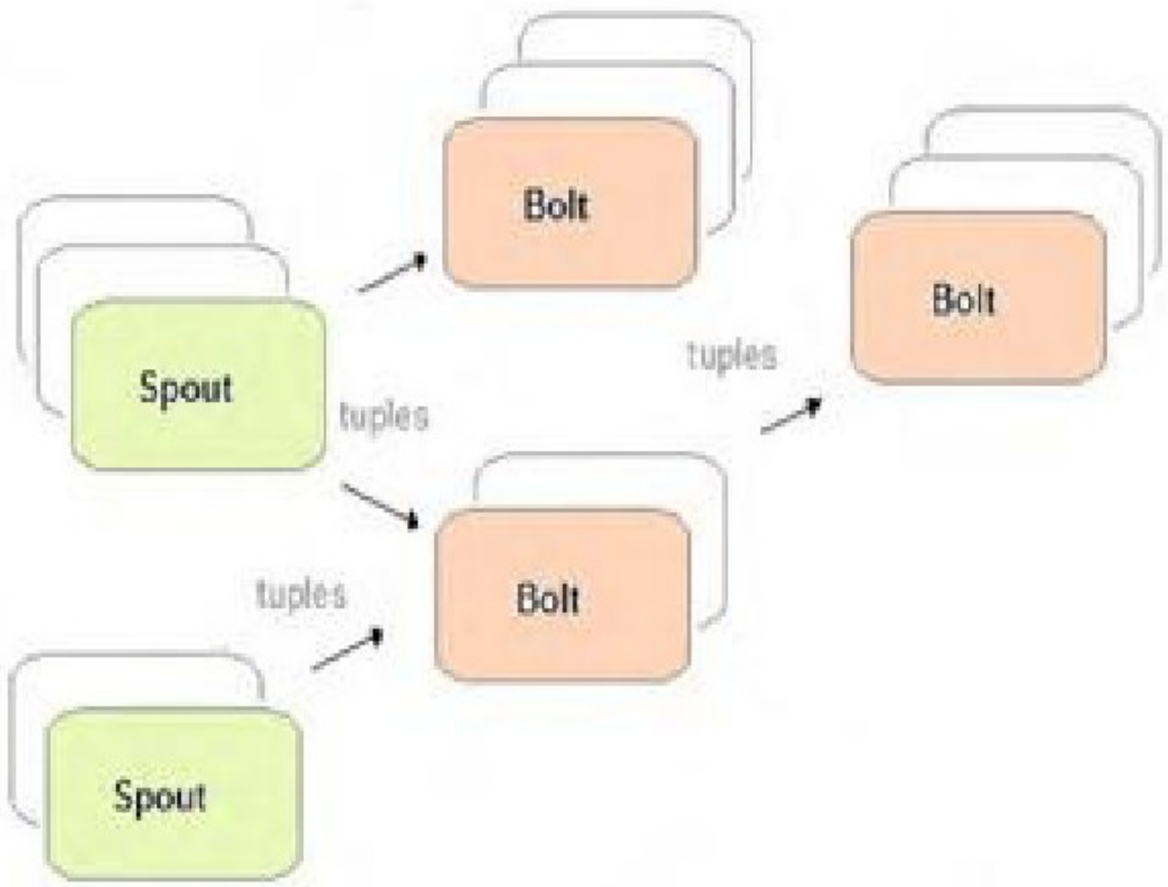
light_drtc

轻量级流式计算框架--light_drtc

- 当前主流流式计算框架
  - Storm
  - Spark Streaming
- 自研流式计算框架——light_drtc
  - 为什么自研
  - 逻辑架构
  - 物理架构
  - 实例：用户画像实时更新

流式计算框架Storm—原理





APACHE STORM™

流式计算框架Storm—实例

Storm UI

Topology summary

Name	Id	Status	Uptime	Num workers	Num executors	Num tasks
NewFeatureDump	NewFeatureDump-7-1466664749	ACTIVE	15d 20h 12m 22s	5	35	35

Topology actions

Activate | Deactivate | Rebalance | Kill

Topology stats

Window	Emitted	Transferred	Complete latency (ms)	Acked	Failed
10m 0s	34960	34960	1162.715	33760	80
3h 0m 0s	524280	524280	1164.344	523620	500
1d 0h 0m 0s	3087020	3087020	1205.315	3080280	2260
All time	54878020	54878020	1194.024	54831020	31880

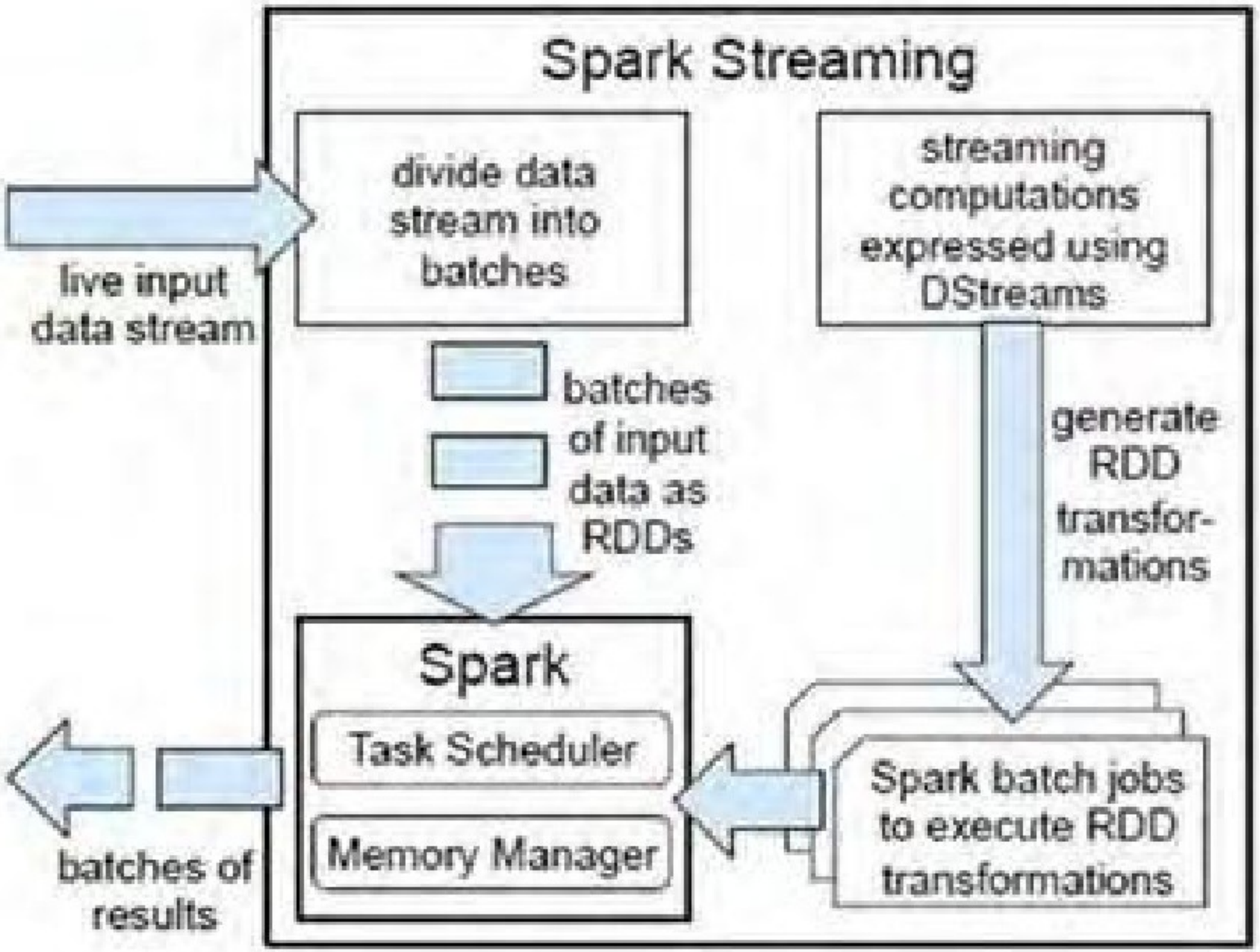
Spouts (All time)

Id	Executors	Tasks	Emitted	Transferred	Complete latency (ms)	Acked	Failed	Error Host	Error Port	Last error
feature-reader	5	5	54878020	54878020	1194.024	54831020	31880			

Bolts (All time)

Id	Executors	Tasks	Emitted	Transferred	Capacity (last 10m)	Execute latency (ms)	Executed	Process latency (ms)	Acked	Failed	Error Host	Error Port	Last error
feature-hbase	25	25	0	0	0.013	2.482	54862960	1194.496	54862860	0			

流式计算框架Spark Streaming—原理



流式计算框架Spark Streaming—实例

```
27    private static final Pattern SPACE = Pattern.compile(" ");
28
29    public static void test1() {
30        JavaStreamingContext jssc =
31                new JavaStreamingContext("local[2]", "JavaNetworkWordCount", new Duration(10000));
32        jssc.checkpoint(".");// 使用updateStateByKey()函数需要设置checkpoint
33        // 打开本地的端口9999
34        JavaReceiverInputDStream<String> lines = jssc.socketTextStream("localhost", 9999);
35        // 按行输入，以空格分隔
36        JavaDStream<String> words = lines.flatMap(line -> Arrays.asList(SPACE.split(line)));
37        // 每个单词形成pair，如 (word, 1)
38        JavaPairDStream<String, Integer> pairs = words.mapToPair(word -> new Tuple2<>(word, 1));
39        // 统计并更新每个单词的历史出现次数
40        JavaPairDStream<String, Integer> counts = pairs.updateStateByKey((values, state) -> {
```

Markers Properties Servers

```
<terminated> HelloSparkStreaming [Java Appli
17043 [task-result-getter-1] INFO
17043 [dag-scheduler-event-loop] IN
17043 [task-result-getter-1] INFO
17043 [streaming-job-executor-0] IN
-------------------------------------------
Time: 1466049610000 ms
-------------------------------------------
(hello,1)
(streaming,1)
(spark,2)
(hell,1)
```

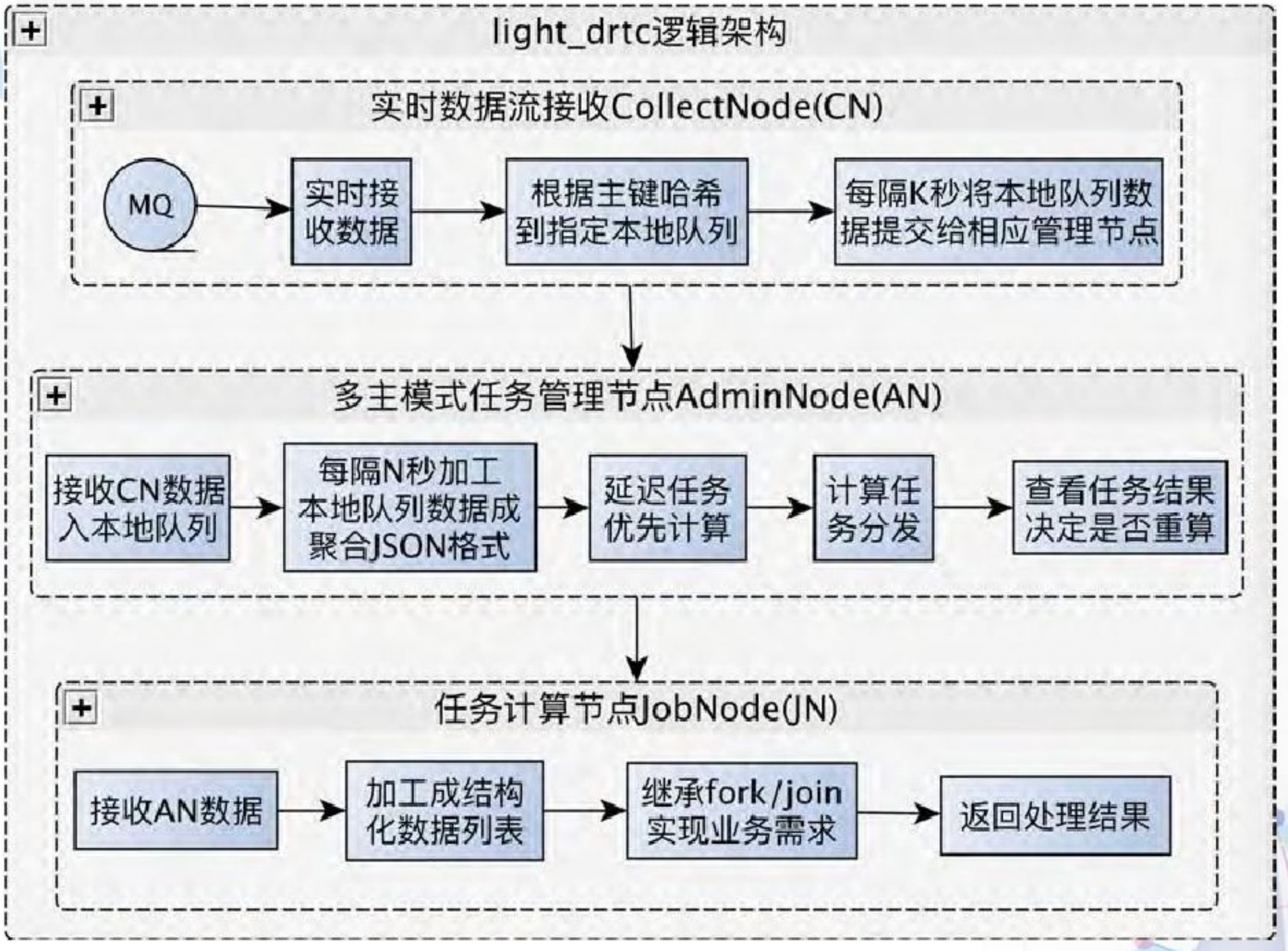
Info Close New

```
root@iZ25t4d589gZ:~ (ssh)        nc
adeMacBook-Pro-62:~ a$ nc -lk 9999
hello spark streaming
hell spark
```

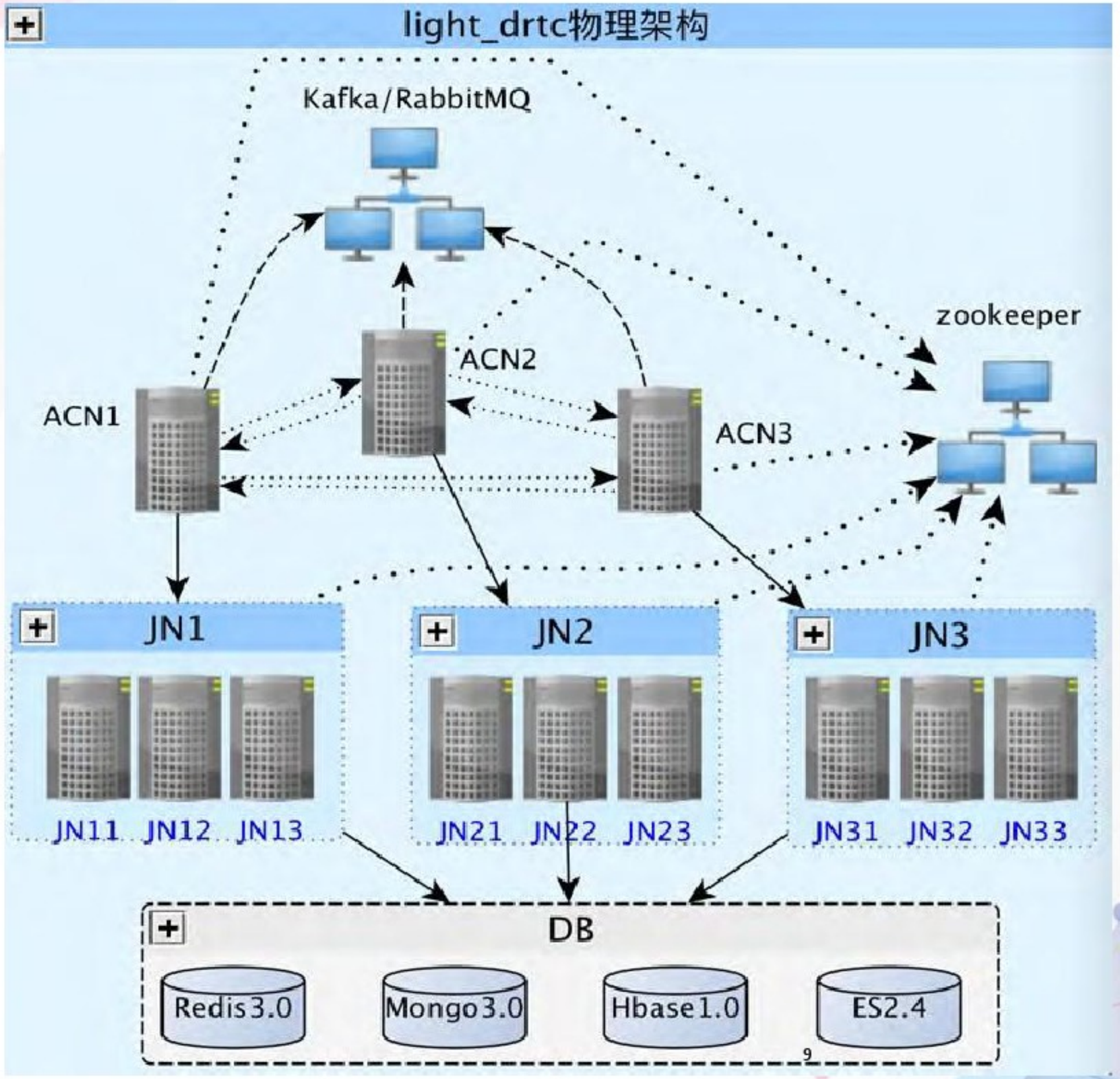
自研流式计算框架light_drtc—自研原因

- 主流流式计算框架一些缺点：
  - **Storm**
  - **Spark Streaming**
- 自研原因
  - 弥补国内对流式计算框架的空白
  - 更为轻量级，方便维护易用
  - 更好定位为题

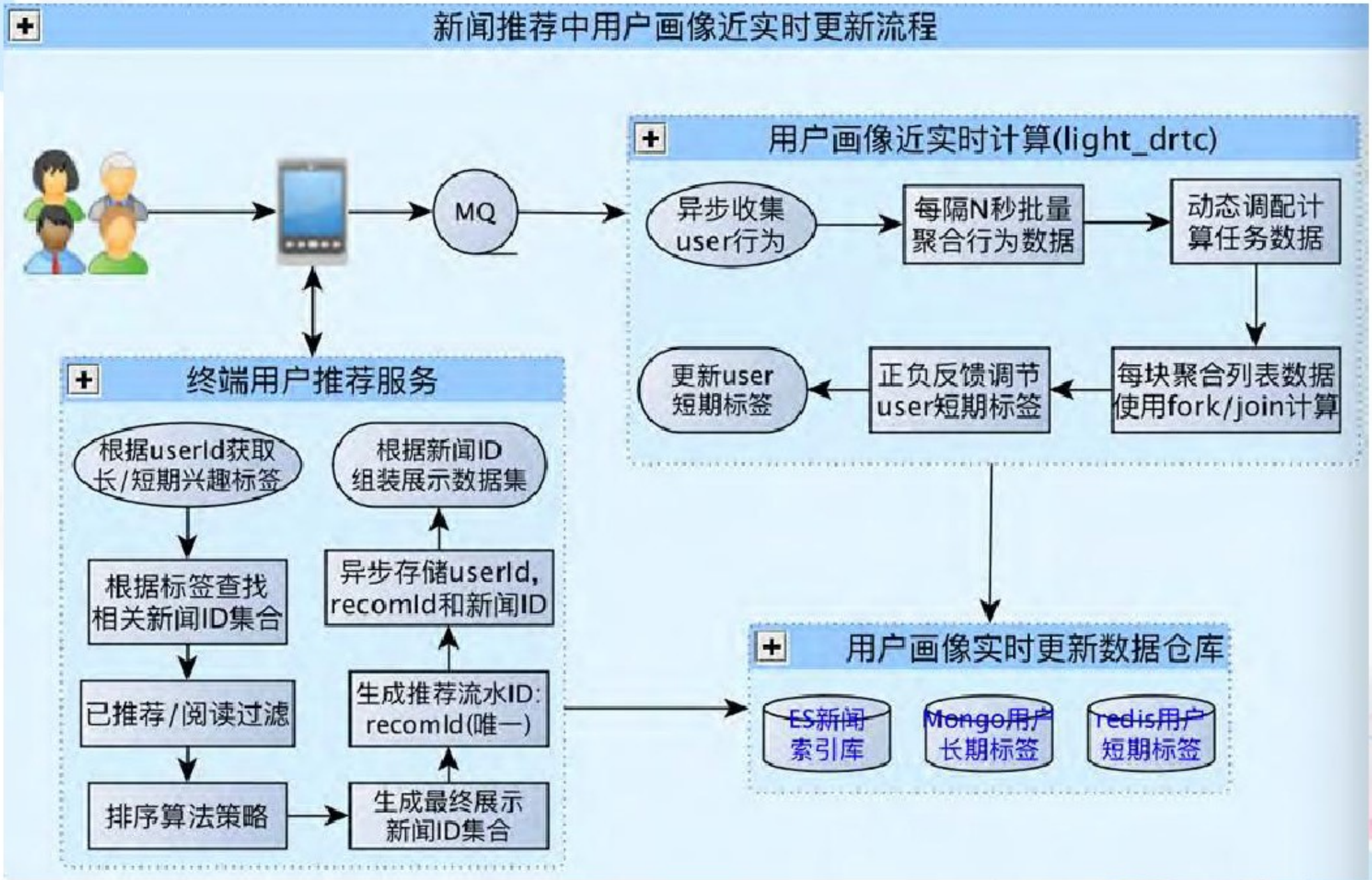
自研流式计算框架light_drtc



自研流式计算框架light_drtc



自研框架案例 - - 用户画像实时更新



自研流式计算框架light_drtc—运行实例