

实时流计算应用开发框架-天罡

□ 孔令西

- 阿里巴巴数据平台部
- 游泳，海鲜，金花，儿子
- 专注数据平台基础平台产品化及流计算
- lingxi.konglx@alibaba-inc.com
- weibo.com : <http://weibo.com/kennyccp>

- 1. 背景
- 2. 业界
- 3. 产品介绍
- 4. 架构设计
- 5. squirrel QL
- 6. 实践经验

- Big data数据量膨胀
- 业务快速变化，商业模式的创新
- SNS,移动互联网
- 用户体验个性化，实时化
- ...

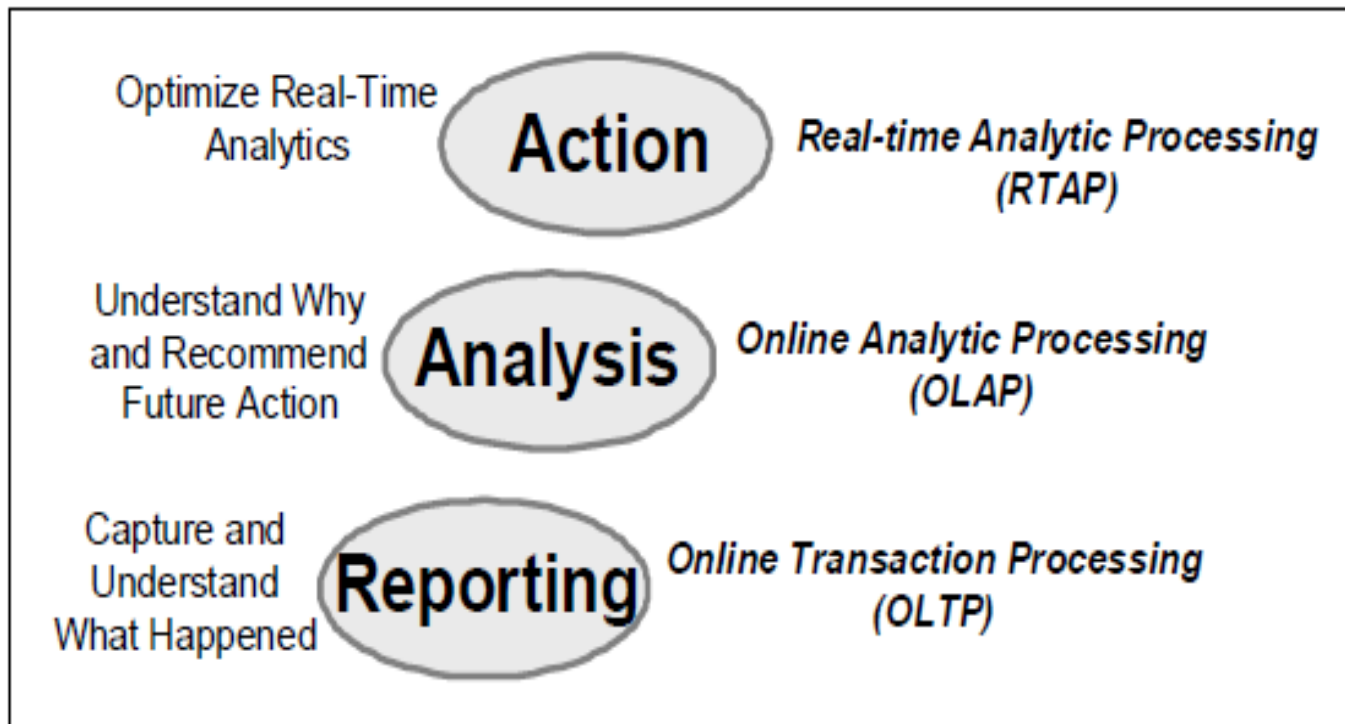


1.2 离线计算 vs. 流计算



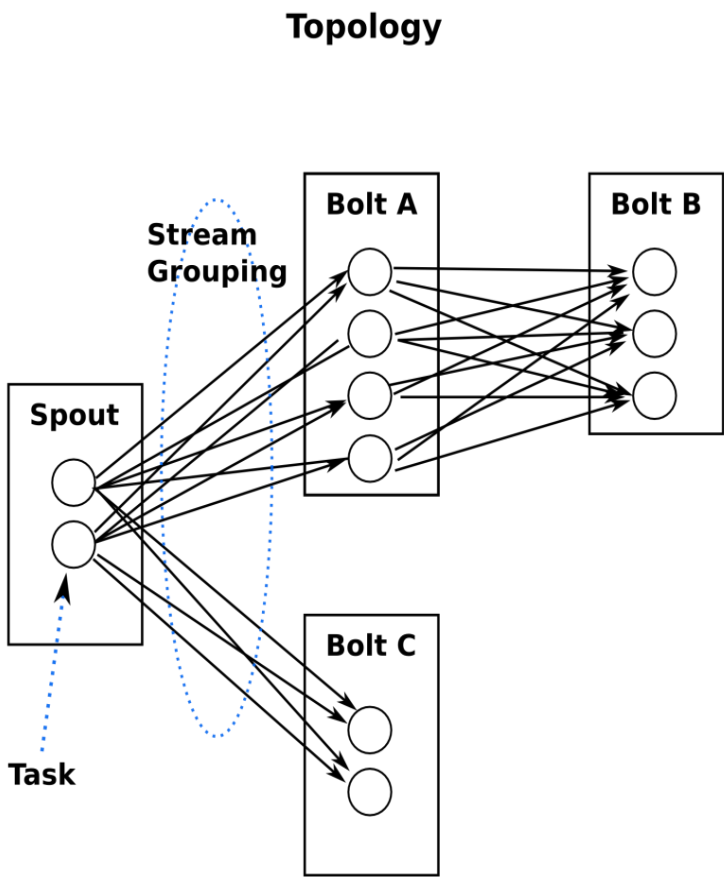
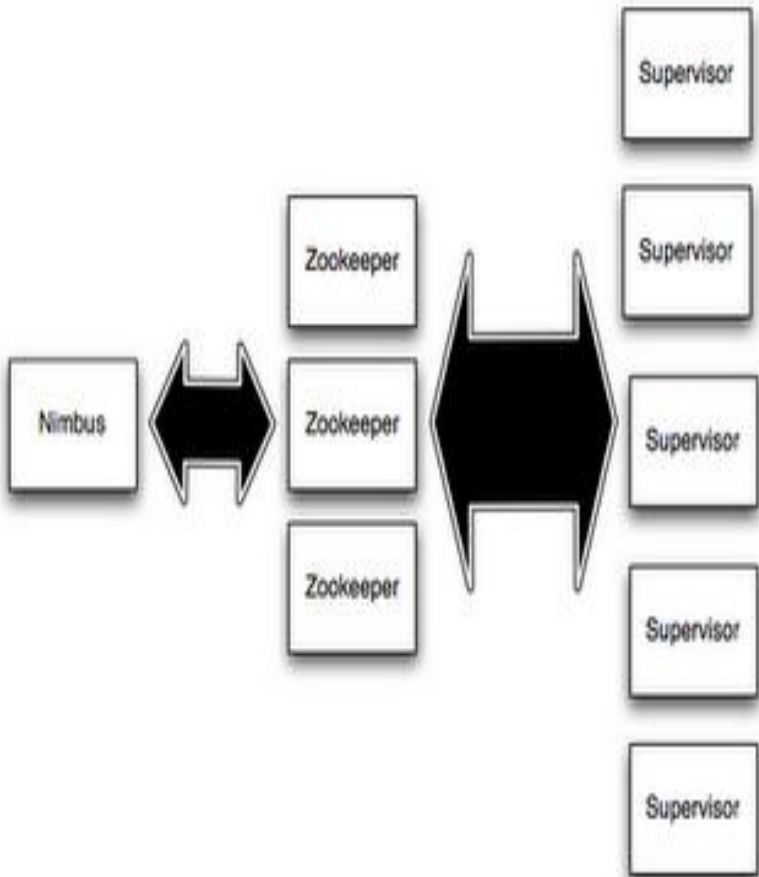
From <IBM InfoSphere Streams: Harnessing Data in Motion>

1.3 数据分析演进趋势



Feature	Puma	Storm	S4	Comments
Development	Java	Clojure	Java	Different systems towards different application requirement
HA	Primary Standby	Upstream Backup	Primary Standby	
Precise Recovery	No	No	No	
Architecture	Symmetric	Master-Slave	Symmetric	
Resource Utilization	Low	High	Low	
Recovery Time	Short	Long	Long	
State Persistence	Yes	No	Yes	
De-dup	Yes	No	No	
★Universal Data Stream Processing System				

- 最新版本storm 0.74
- 主要特性：
 - 适用场景广泛
 - 可伸缩性高
 - 保证无数据丢失
 - 异常健壮
 - 容错性好
- 一些问题：
 - 编程门槛对普通用户较高
 - 框架无持久化存储
 - 框架不提供消息接入模块
 - storm ui功能简单
 - 跨topology的bolt复用
 - nimbus单点
 - topology不支持动态部署



目前需求归类

需求特征

1

统计业务关键指标，客观反映当前的业绩现状，比如网站活动监控

分钟级延迟；不能漏算；不能错算；统计时长为当天；

2

跟踪业务指标的变化趋势，出现异常波动，能智能报警

分钟级延迟；不能漏算；不能错算；统计时长为当天；

3

业务闭环运营中的实时数据应用，比如事件营销，触发式服务

秒级延迟；允许漏算；不能错算；计算过程复杂（规则多）；

4

实时推荐

秒级延迟；允许漏算；不能错算；与推荐系统交互频繁；

5

实时数据信息服务

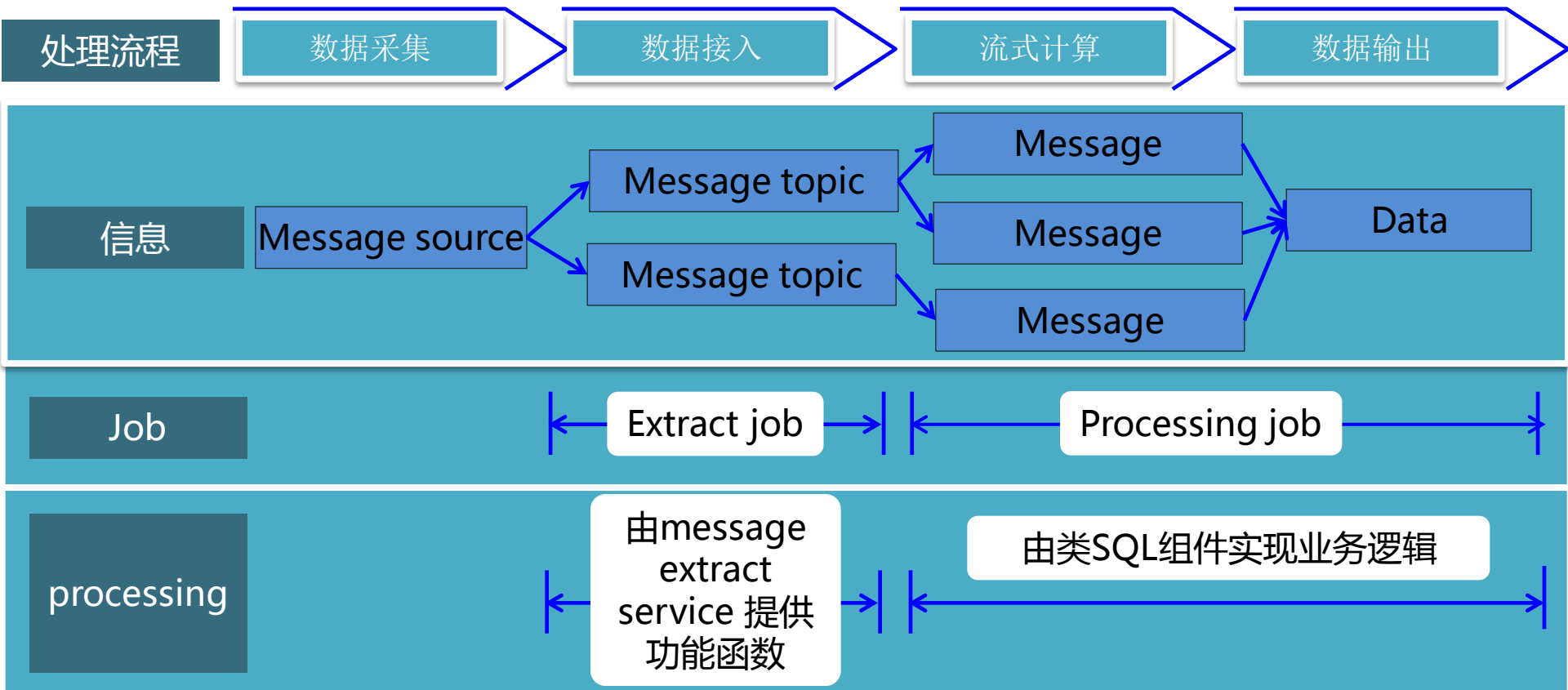
秒级延迟；不允许漏算；不能错算；计算过程复杂（指标定义复杂，指标个数多）；部分指标统计时长跨天；

天罡：实时流计算应用开发框架

深刻理解实时业务需求，提供实时计算的完整**应用开发框架**

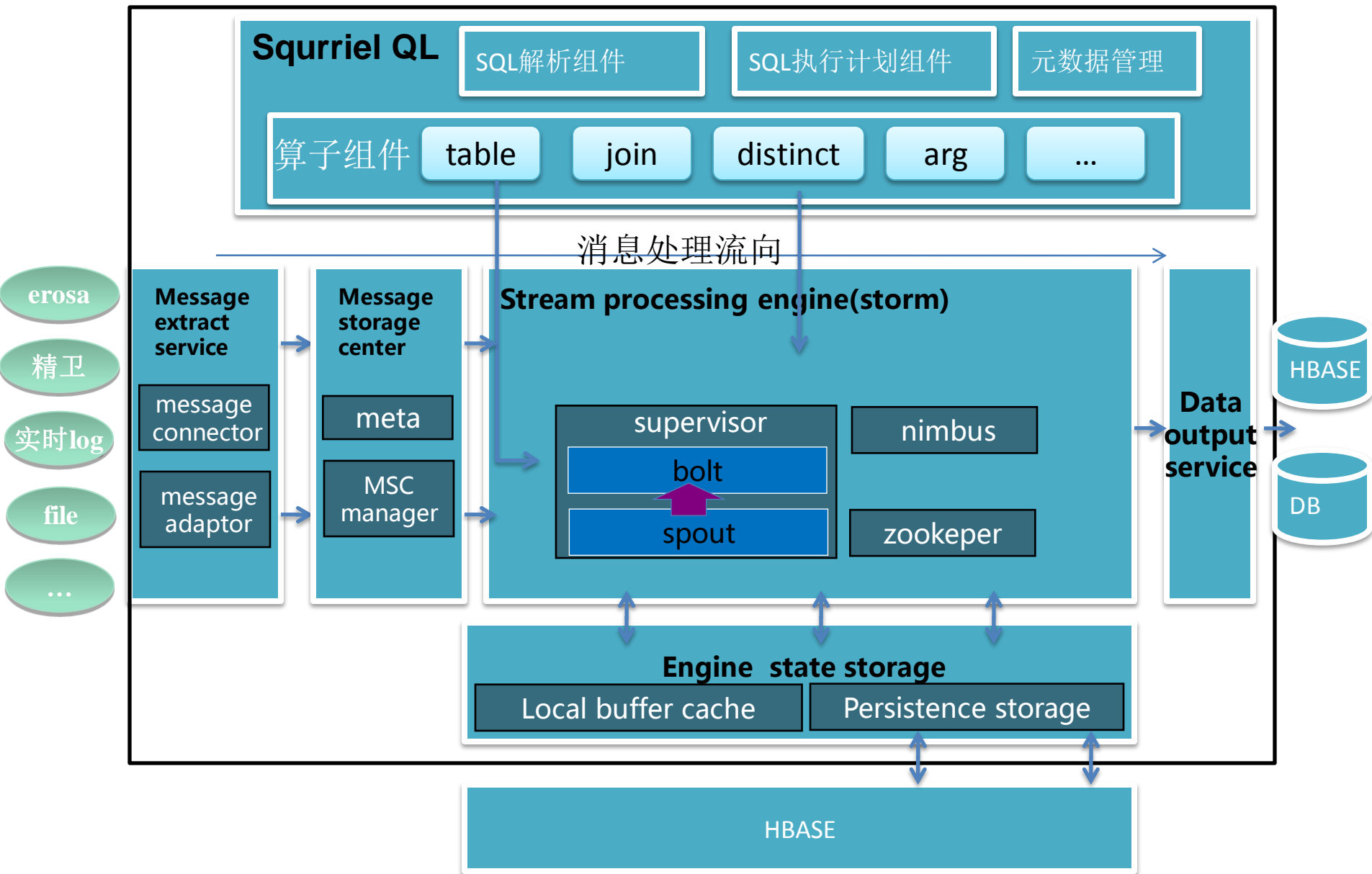
1. 屏蔽系统复杂性，**可配置方式**即可完成消息源接入
2. **类SQL工具**，封装引擎系统，降低实时计算任务的开发难度
3. 完整的任务管理系统，提供任务配置、发布、管控一条龙服务
4. 强大的运维管理系统，监控系统、任务、数据的状态，适时报警

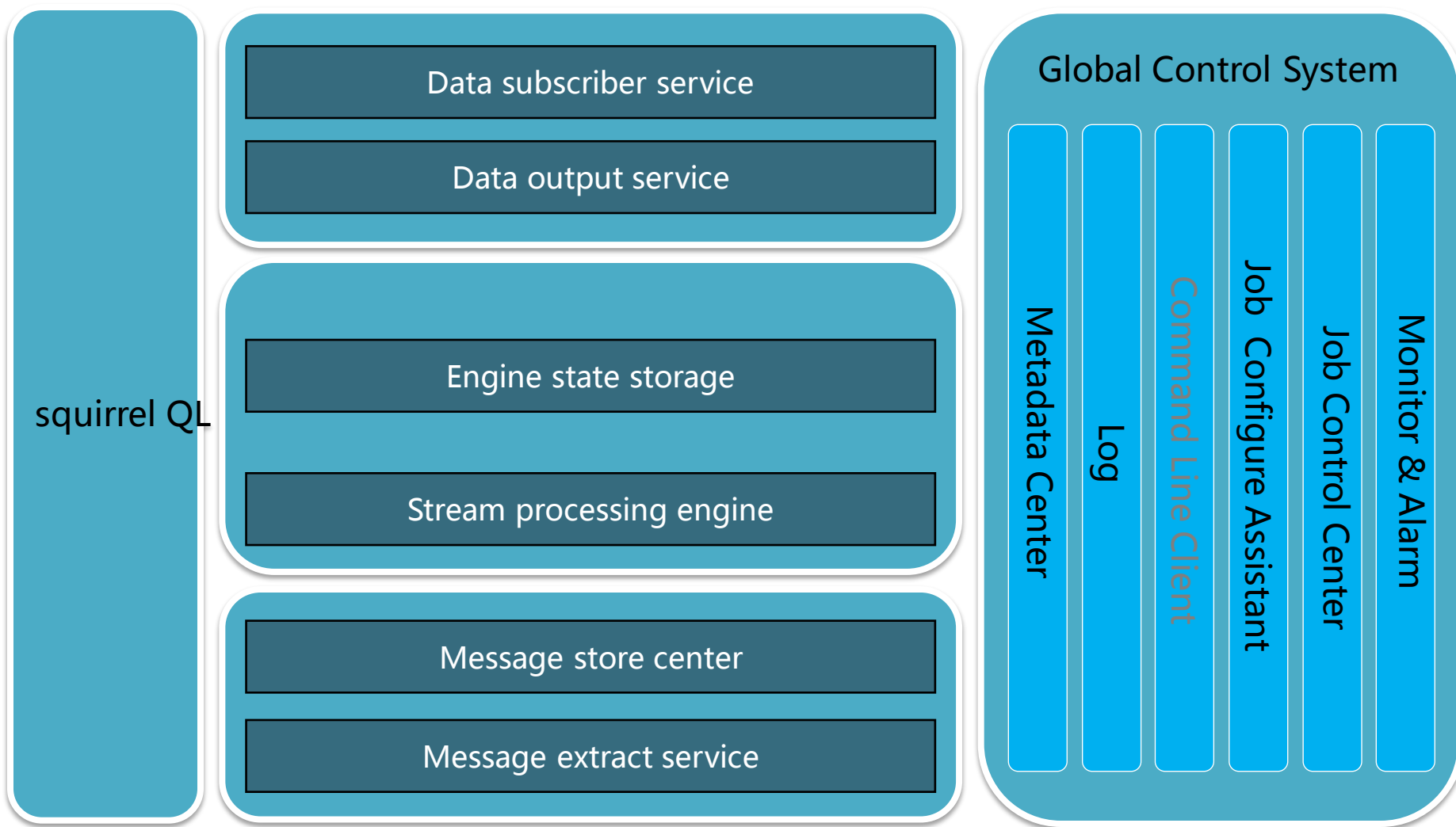
- 简单统计计算(包括时间窗口)
- 多流join计算(动态)
- 容错，事务
- 中间状态持久化
- 统一消息接入
- 支持类SQL
- 支持数据类型：int/long/string/double...
- 支持schema
- 支持join,distinct,group by,count, top N
- 支持常规函数
 - to_char ,substr...



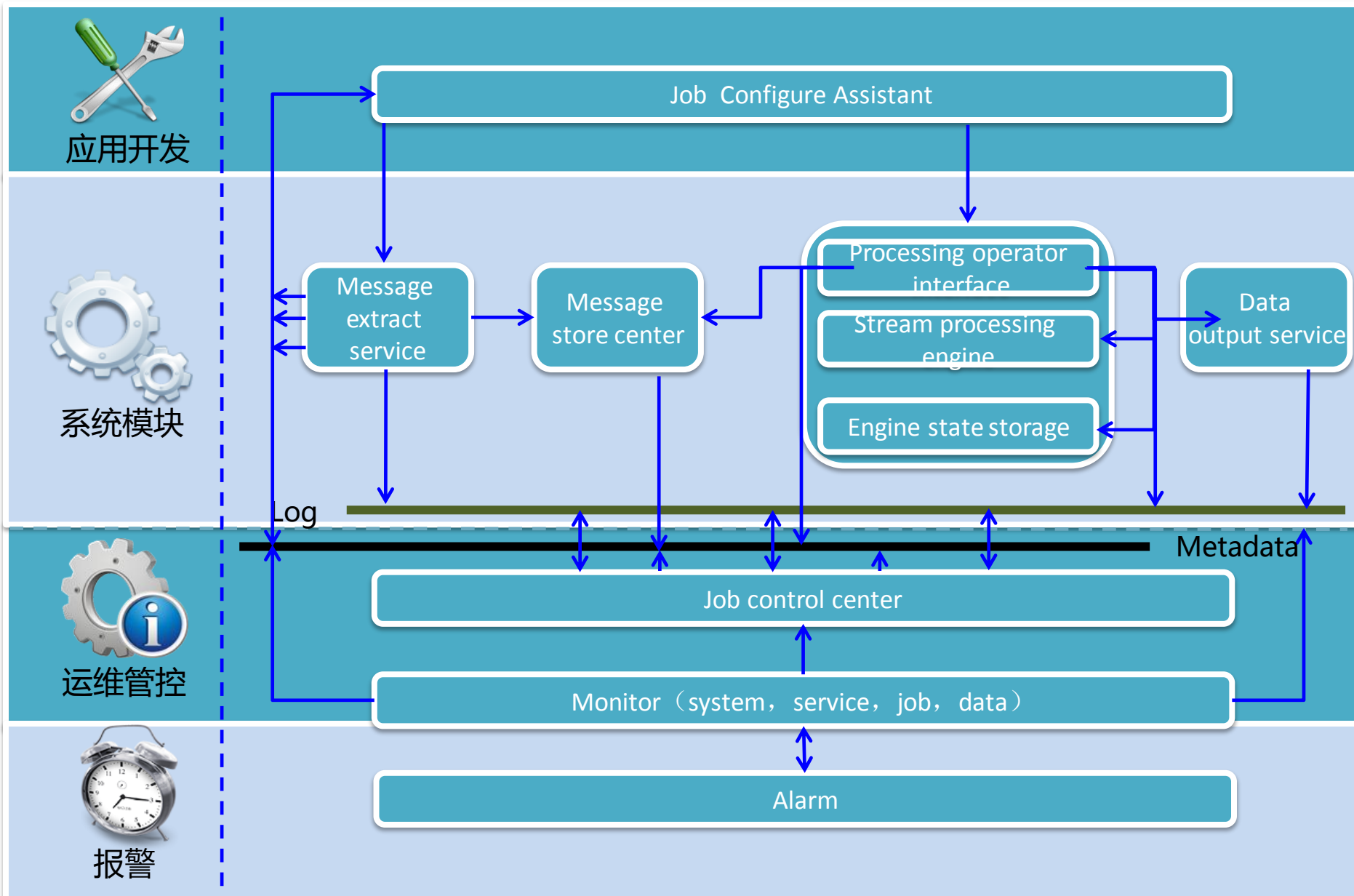
天罡系统里的相关概念说明：

1. 一个message source (相当于database) 包括1个以上的message topic(相当于table), 一个message topic由1条以上的message (相当于record) 组成。
2. 天罡系统有二类Job, 分别称为extract job和processing job。一个extract job负责一个message topic的消息接入, 一个processing job由1个以上的message topic参与计算。一个message topic可以被多个processing job引用。Processing job的输入内容称为message, 计算的结果输出, 称为data。



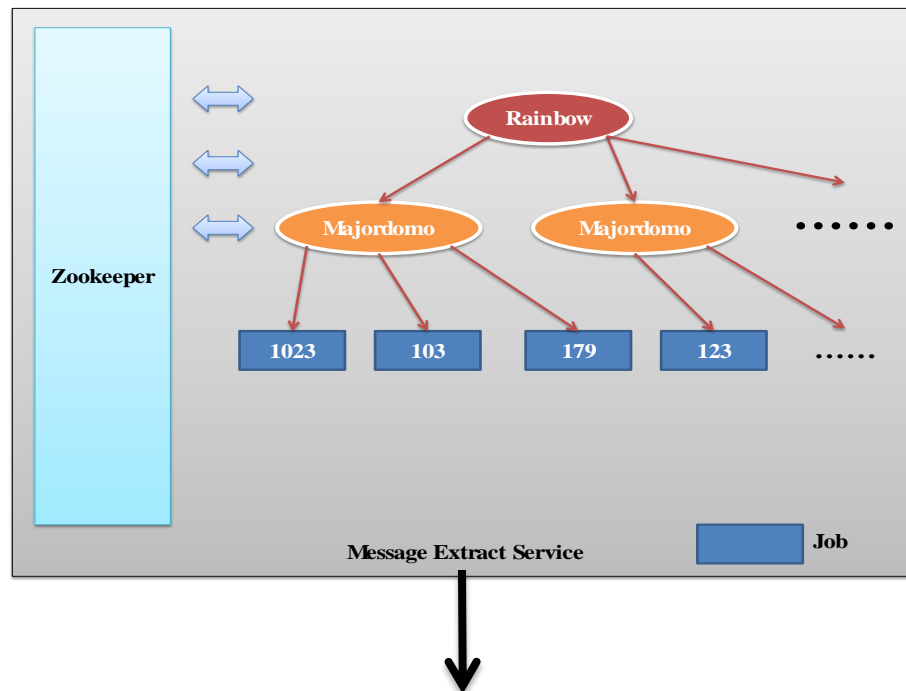


4.3天罡-功能模块关系



4.4 Message extract service

- Extract job
 - message连接器, 能够接入db、log、file、mq
 - message适配器, 提供字段选择、格式转换功能
 - 实现Job的配置和管理接口
- 服务
 - 分布式调度控制管理
 - 负载管理
 - Job启动, 停止
- 规则配置
 - Keyword search



```

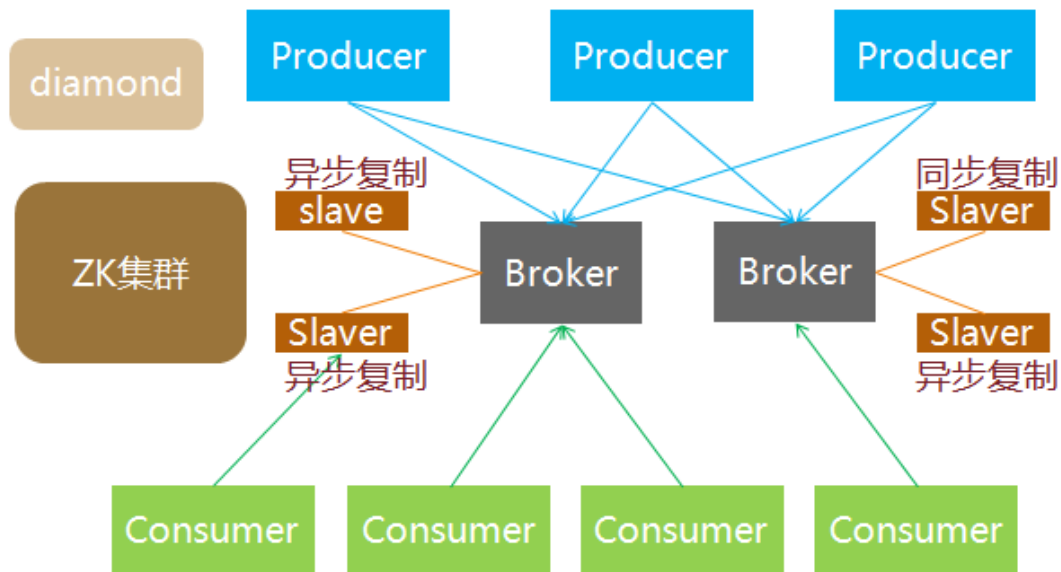
[admin@inc-dw-151-45 plough-mes]$ bin/plough-mes
Usage: plough-mes command [command_params].....
- submit job_type job_id @submit a job, you should provide the job_type and job_id that you want to submit.
- kill job_id @kill a job, you should provide the job_id that you want to kill.
- list [majordomo_ip] @list all active jobs without majordomo_ip or list the jobs of the specific majordomo.
- stats type [id] @get stats of something.
- stats cluster @get stats of the cluster.
- stats rainbow @get stats of the rainbow.
- stats majordomo ip @get stats of the specific majordomo.
- stats job job_id @getstats of the specific job.
- logs job_id @get all log file paths of specific job.
- rainbow @start the rainbow service on current machine.
- majordomo @start the majordomo service on current machine.
- help @how to use it.

Good luck!!!
[admin@inc-dw-151-45 plough-mes]$
    
```

- 统一消息存储方式

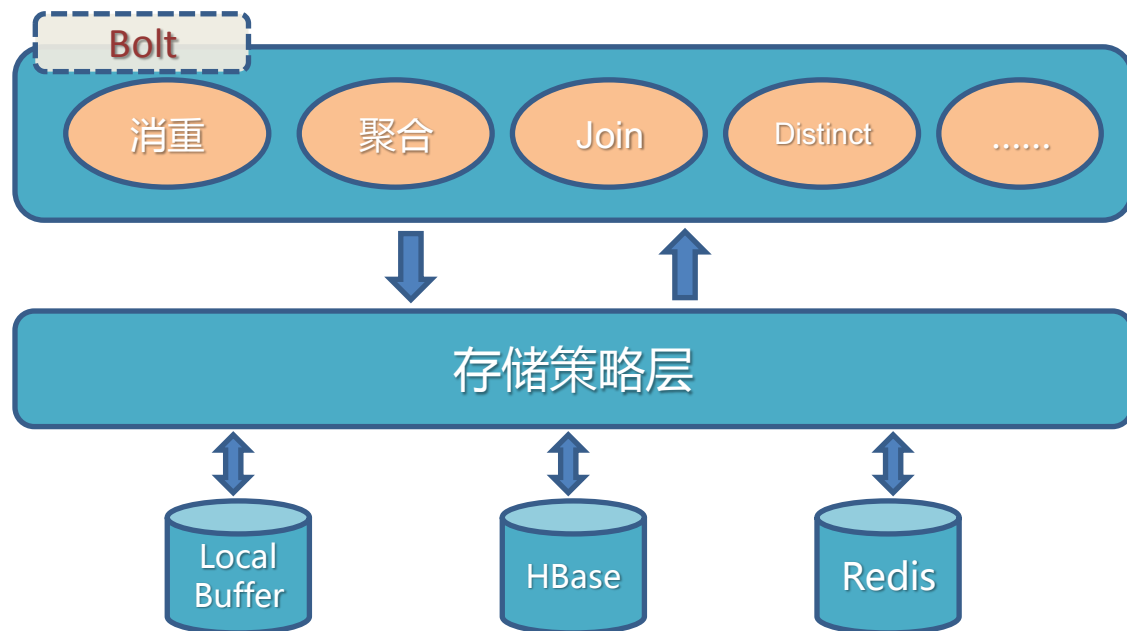
- 基于**Metamorphosis**: 淘宝一款类似**Kafka** 强大的通用消息中间件
- Pull模式的MQ
- 高吞吐量
- Meta spout
- 支持消息顺序

- Message read interface
- Message write interface
- MSC manager



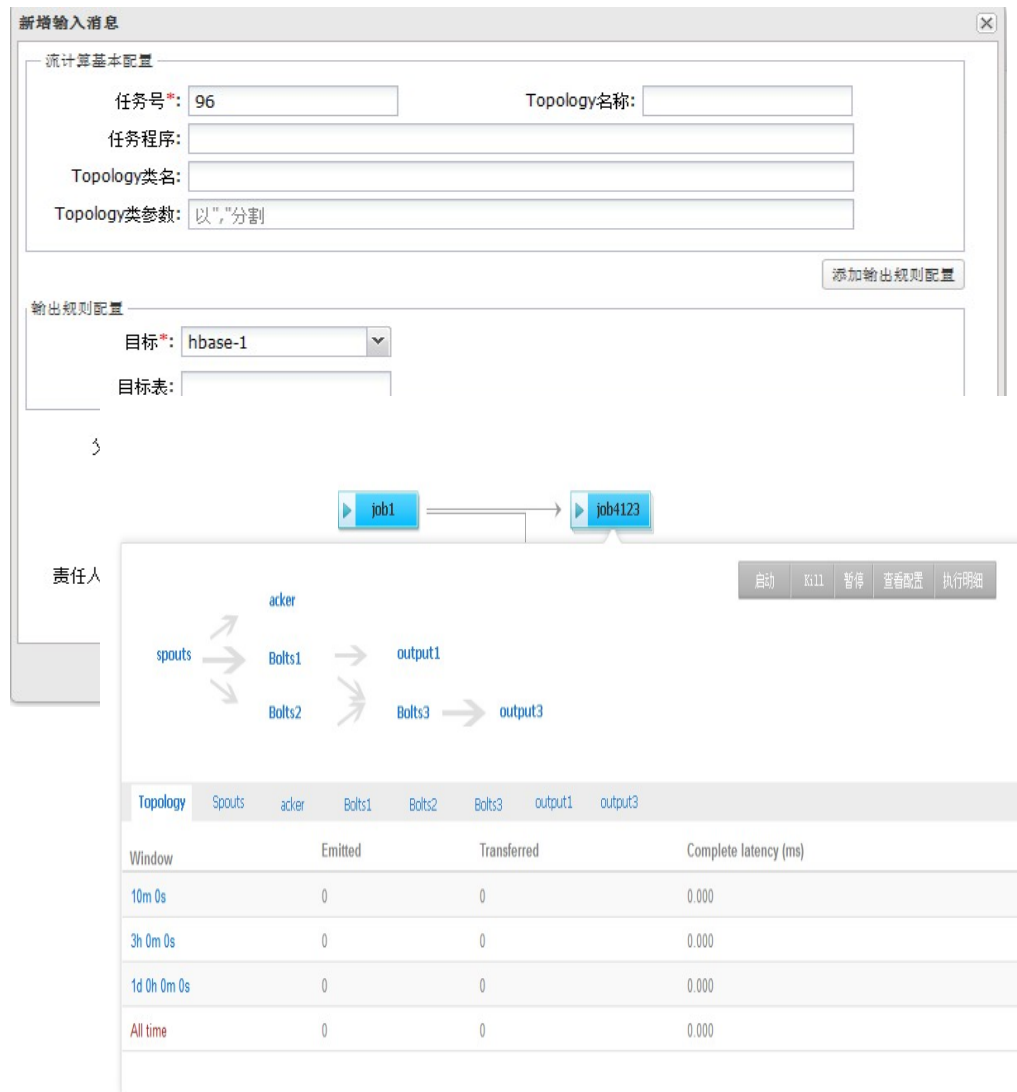
- 中间状态持久化需求场景

- join
- 聚合
- 原始消息存储
- 容错，事务



4.7 Global Control System

- 系统、任务等配置信息的统一存储管理中心
- 提供用户体验良好的任务研发和配置服务
- 任务管理中心，直观反应当前任务执行状态，能够中止、暂时、重设任务
- 监控：系统（内存，cpu，网络，io等），功能（redis，meta，storm等），任务，数据质量
- storm集群管理



The screenshot displays the '新增输入消息' (Add Input Message) window, which is divided into two main sections: '流计算基本配置' (Stream Computing Basic Configuration) and '输出规则配置' (Output Rule Configuration).

流计算基本配置 (Stream Computing Basic Configuration):

- 任务号* (Task ID): 96
- Topology名称 (Topology Name):
- 任务程序 (Task Program):
- Topology类名 (Topology Class Name):
- Topology类参数 (Topology Class Parameters): 以","分割 (Separated by commas)
- 按钮: 添加输出规则配置 (Add Output Rule Configuration)

输出规则配置 (Output Rule Configuration):

- 目标* (Target): hbase-1
- 目标表 (Target Table):

任务执行状态 (Task Execution Status):

The interface shows a task named 'job1' with a status of 'job4123'. Below this, there is a diagram of the topology structure:

```

graph LR
    spouts --> Bolts1
    spouts --> Bolts2
    Bolts1 --> output1
    Bolts2 --> Bolts3
    Bolts3 --> output3
  
```

责任人 (Responsible Person):

Buttons: 启动 (Start), Kill, 暂停 (Pause), 查看配置 (View Configuration), 执行明细 (Execution Details)

Table: Topology Performance Metrics

Topology	Spouts	ackerr	Bolts1	Bolts2	Bolts3	output1	output3
Window		Emitted		Transferred			Complete latency (ms)
10m 0s		0		0			0.000
3h 0m 0s		0		0			0.000
1d 0h 0m 0s		0		0			0.000
All time		0		0			0.000

简单统计场景（中文站UV,PV统计）：

```
select a,b,count(c),count(distinct d) from x where a=1 group by  
to_char(a,'yyyymmdd'),b
```

过滤

聚合

中间状态存储

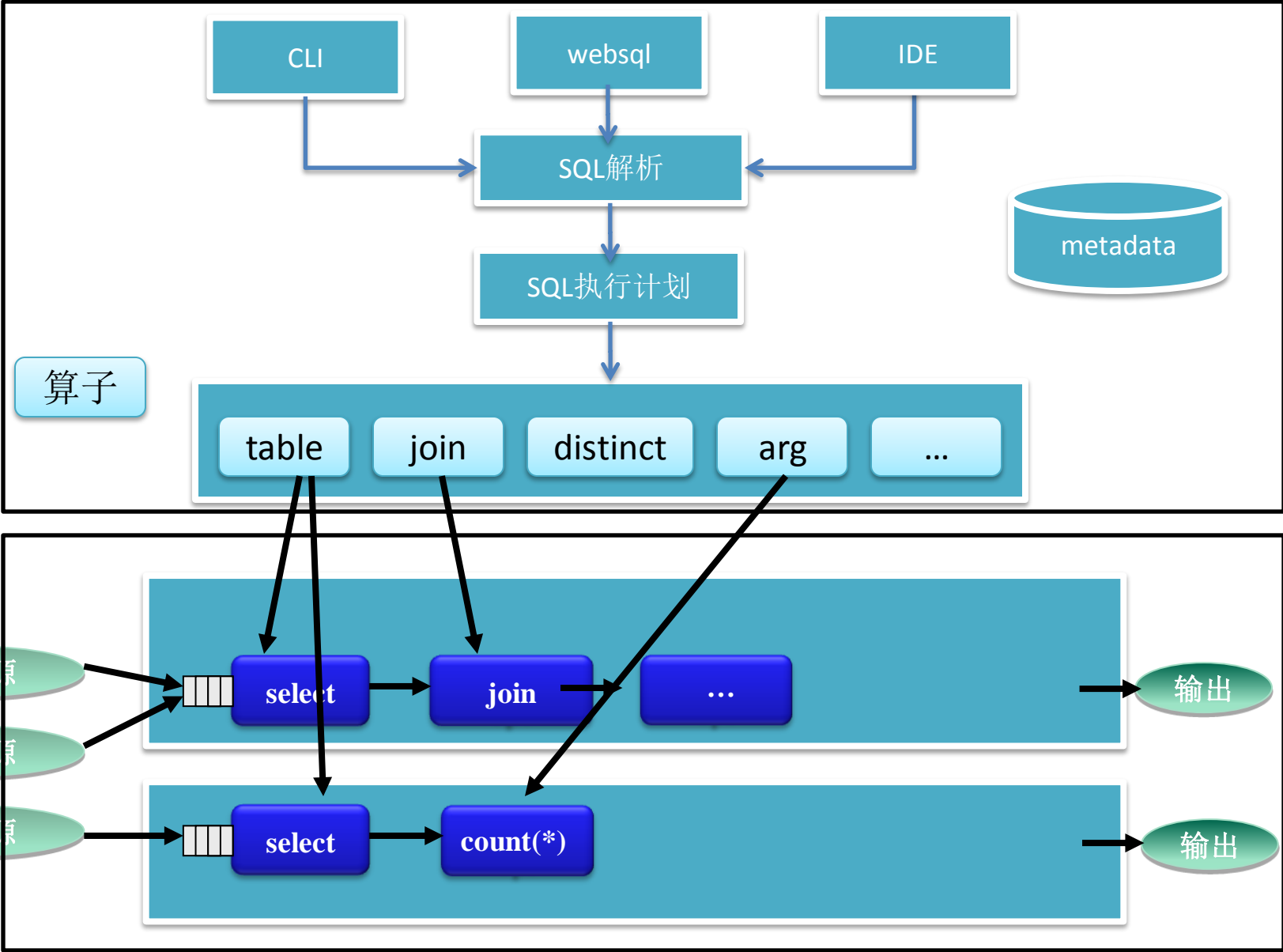
输出

容错

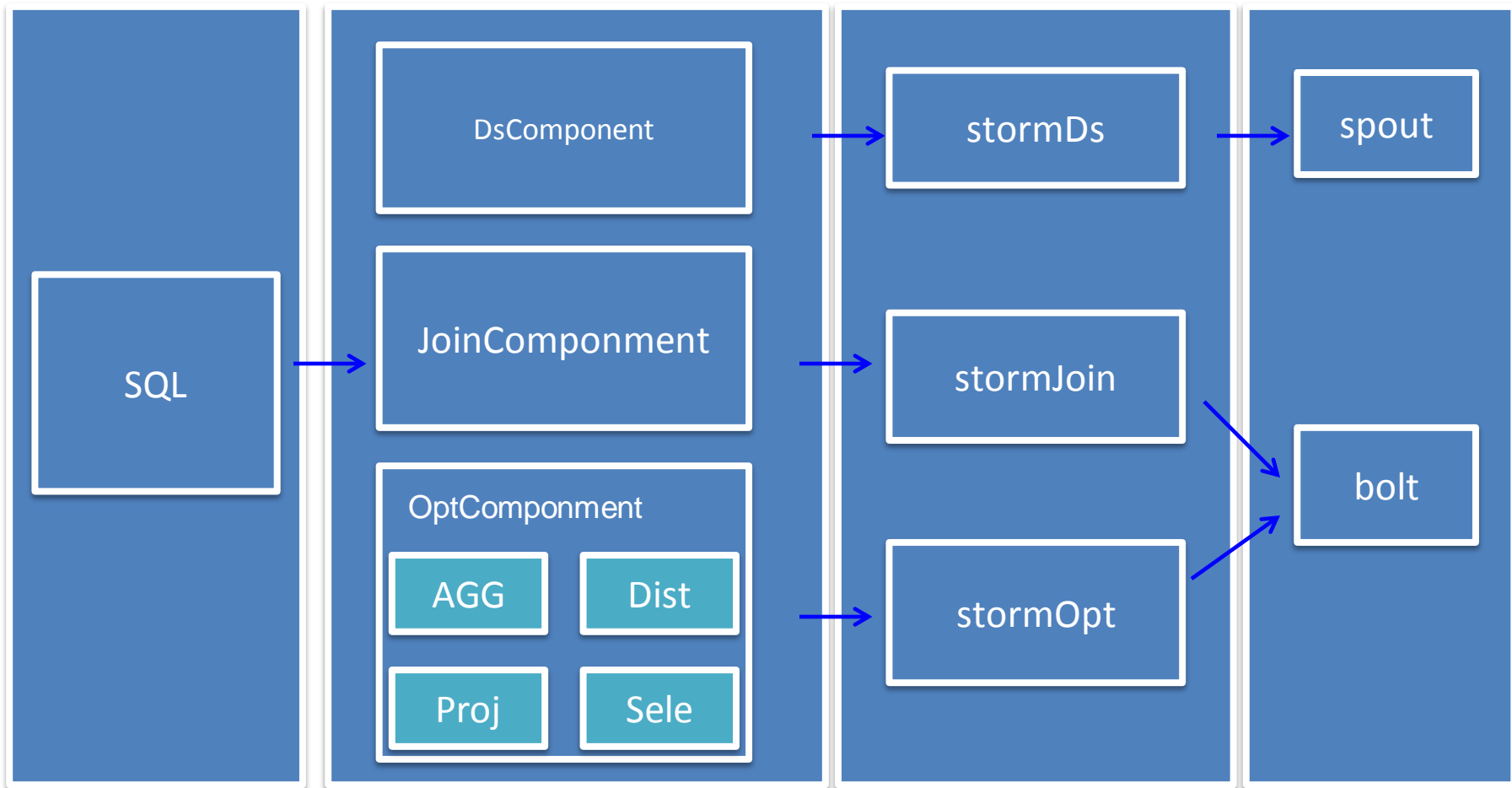
...



```
TopologyBuilder builder = newTopologyBuilder();  
builder.setSpout(1, cbuuvpvSpout(), 10);  
builder.setBolt(2, cbupvuvBolt(),  
3).shuffleGrouping(1);  
.....
```



5.3 squirrel QL 解析流程



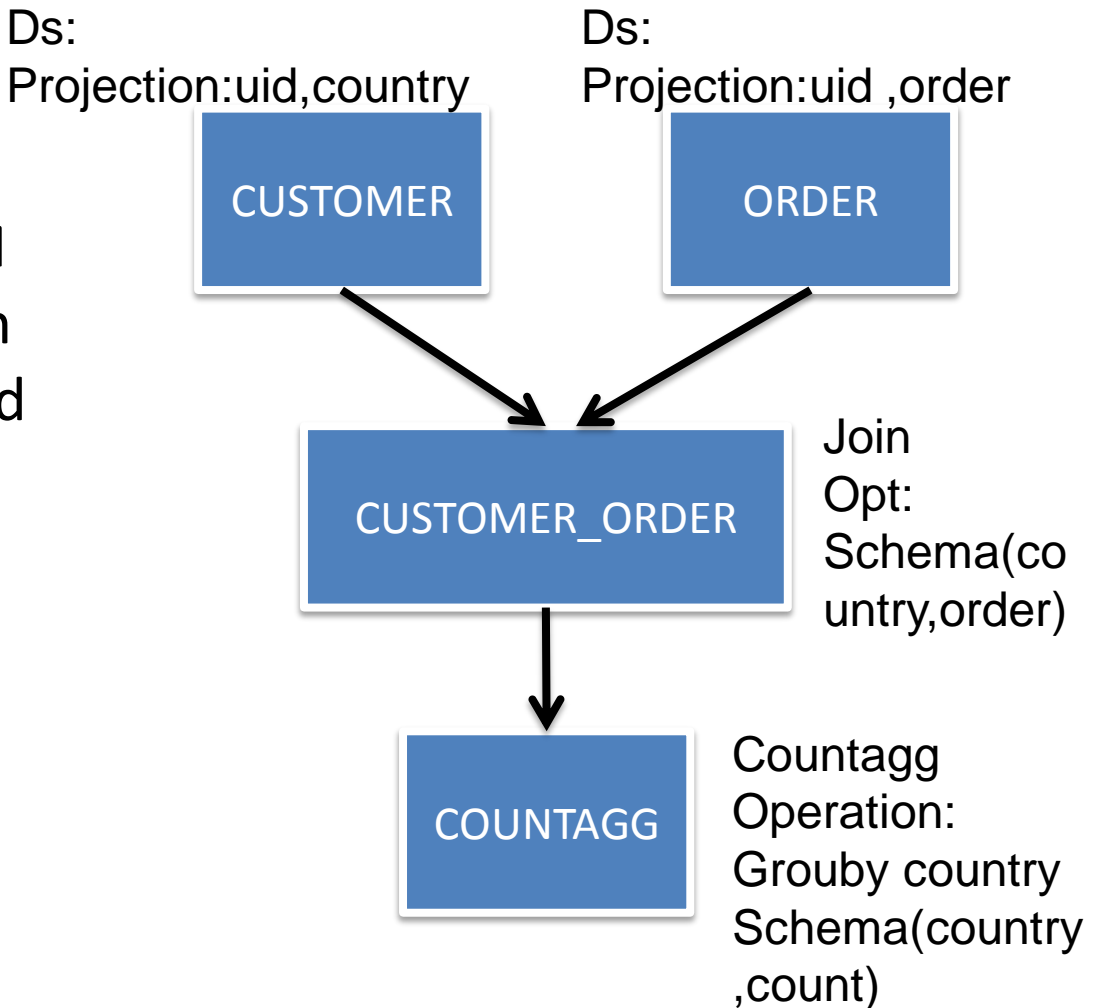
- SELECT uid,COUNT(order) FROM table
GROUP by uid



Ds:
Projection (uid, order)

Opt:
Group by uid
Schema(uid,count)

- SELECT
 c.country,
 COUNT(o.order) FROM
 customer c join
 order o on c.uid= o.uid
 GROUP BY c.country

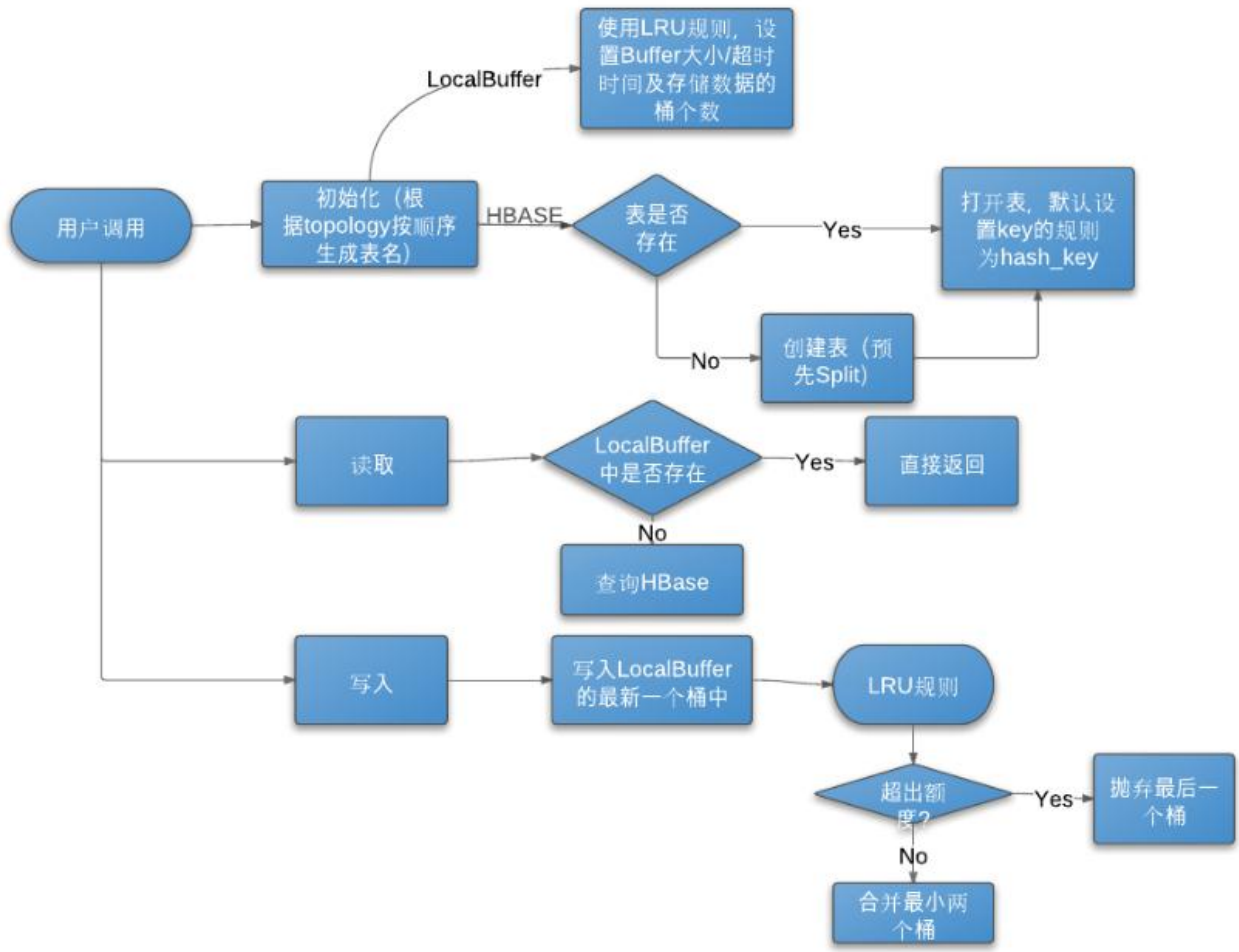


5.6 squirrel QL&Hive QL

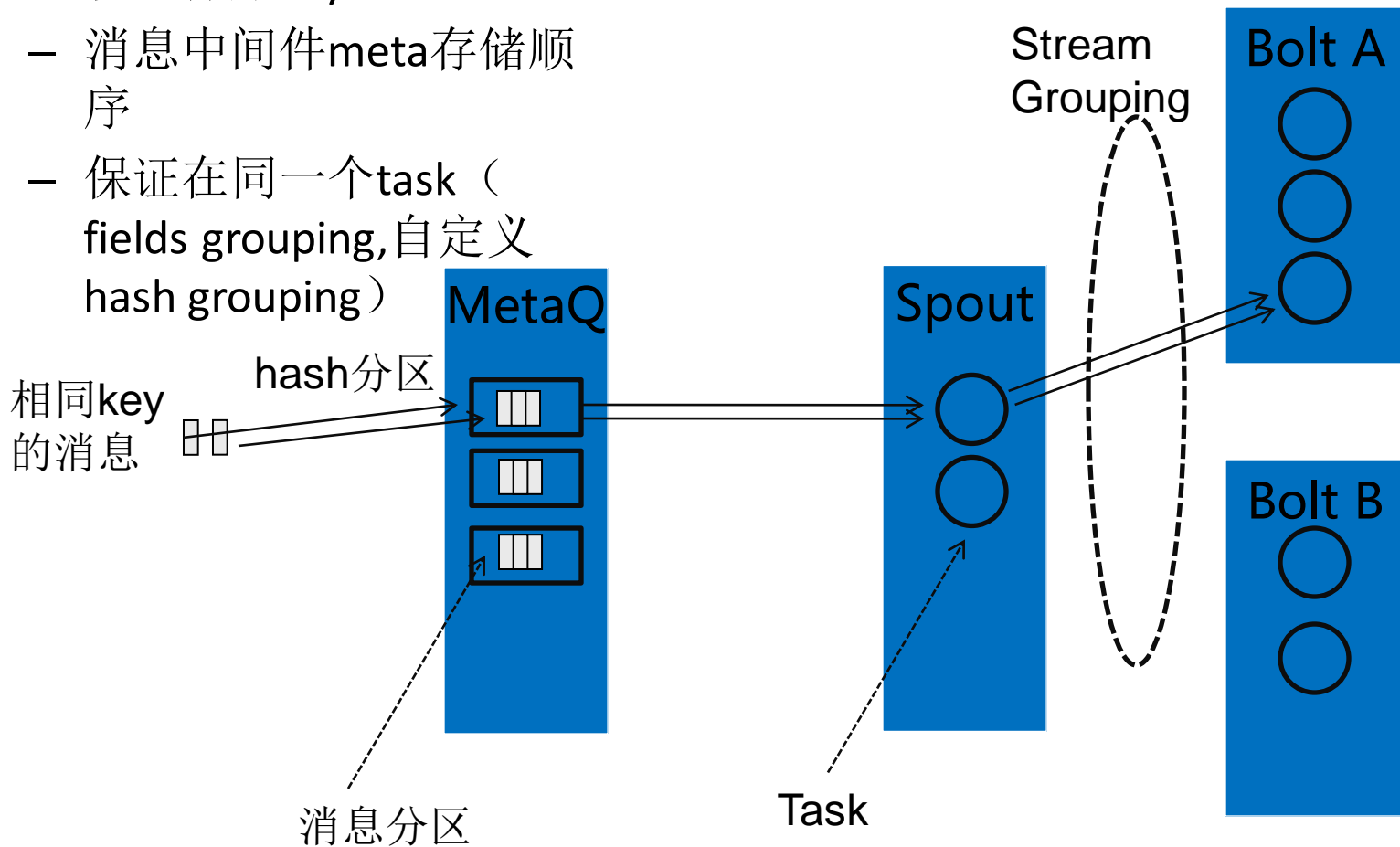
	Hive QL	Squirrel QL
数据存储	HDFS	HBASE
数据格式	用户自定义	用户自定义
数据更新	不支持	不支持
索引	有（0.8版之后增加）	无
执行	MapReduce	Topology
可扩展性	高（UDF，UDAF,UDTF）	高

中间状态读写策略

- 缓存,超时6秒
- LRU
- 命中率



- 消息顺序如何保证？
 - 全序，偏序
 - 设置保序key
 - 消息中间件meta存储顺序
 - 保证在同一个task（
fields grouping, 自定义
hash grouping）



- 时效与准确性间平衡
 - 纯流计算(时序不严格, 注重时效性, 准确性可通过批量计算补偿)
 - 精确计算(牺牲一定时效, 接近准确)
 - 消息保序
 - transactional topology (transaction id)
 - 消息重发
 - 应用去重
 - 幂等

业务监控

效果监控中心

实时GMV
效果监控

实时流量
效果监控

实时活动
效果监控

实时营销
效果监控

异常监控中心

日志异常实时报警

恶意点击
实时报警

攻击实时监控报警

推荐

网站实时推荐

首页实时推荐

404页面实时推荐

店铺实时推荐

Top likes/trends

贸易通实时推荐

热卖产品实时推
荐

偏好产品实时推
荐

广告实时推荐

卖家360

买家实时活动

实时访问动态

实时反馈动态

实时搜索动态

实时订阅动态

实时询盘动态

实时点击动态

实时意愿判断

Q&A
Thanks

会场用餐自理，以下用餐场所供大家参考

海外海皇冠假日酒店.主楼

名人名家餐厅：二楼（包厢）、三楼（大厅）

西餐厅：四楼

海外海酒店马路对面有各色小吃店

出酒店左拐直走10分钟左右，胜利河美食街

祝您用餐愉快！