阿里云开发者社区

日志服务数据加工系列培训

主题：扫平日志分析路上障碍，实时海量日志加工实践培训

讲师：丁来强（成喆）- 阿里高级技术专家 | 唐恺(风毅) - 阿里技术专家

分享介绍

8月7日 19:30-20:30	8月8日 19:30-20:30	8月13日 19:30-20:30	8月14日 19:30-20:30	8月20日 19:30-20:30	8月21日 19:30-20:30	8月28日 19:30-20:30	8月29日 19:30-20:30
数据加工 介绍与实战	数据加工DSL 核心语法介绍	数据加工DSL 语法实践	数据加工动态 数据分发汇集实践	非结构化数据 解析实践	结构化数据 解析实践	数据映射 富化实践	数据加工 可靠性与排错实践

数据加工：介绍与实战

系列培训一

丁来强（成喆）

议题

- 日志服务扫盲
- 数据加工：解决痛点与场景
- 功能配置与原理
- Demo：
  - 控制台操作
  - SLB数据加工实战

日志服务介绍

日志使用场景：运维、研发、运营、安全、BI、审计....

日志服务：阿里集团针对日志分析、处理自研产品

让用户专注在“分析”上，远离琐碎的工作

1. 数据实时采集 LogHub

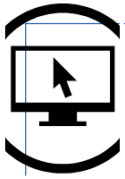
- 30+渠道
- 公网加速
- 断点续传
- 免维护、低成本

2. 智能查询分析 Search/Analytics

- SQL92语法
- 秒级分析亿级
- 所见即所得
- AI、预测与告警

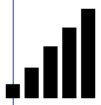
3. 数据分发 LogShipper/LogHub

- 免维护/免费
- 生态丰富
- 使用简单



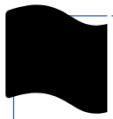
可靠

双十一、金融级考验



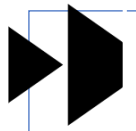
稳定

百万机器运行保证



无运维负担

0负担，0预留成本



迭代快

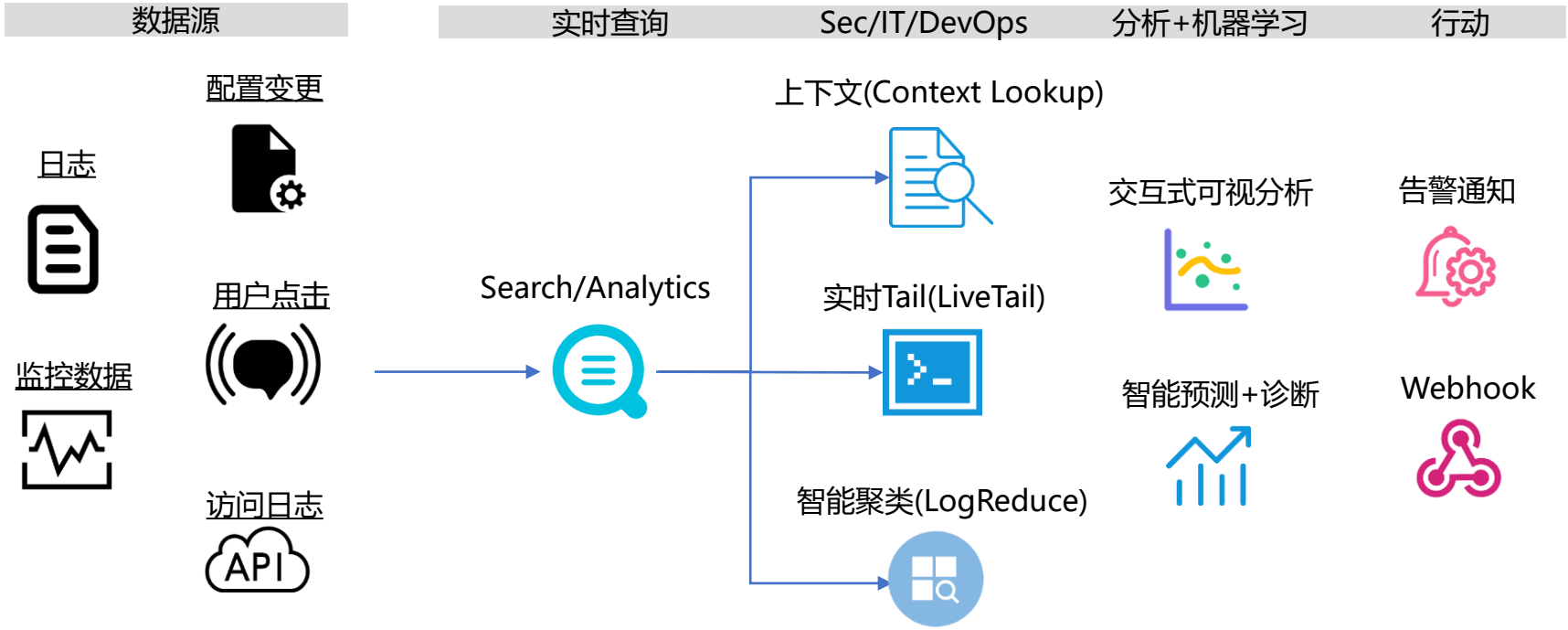
与阿里使用同一套产品

全面覆盖近30个阿里云产品的审计、网络、安全、数据等日志

资产分类	类型	用户类日志	系统类日志
基础	操作审计 (ActionTrail)	IAM/RAM登录日志 阿里云产品的资源操作日志(除存储类产品)	阿里云通过API操作资源的行为
计算	ECS	主机内登录日志, 进程打开日志, 网络连接日志, 账户/端口Snapshot等)	/
	K8S	资源操作日志	/
	函数计算	应用内日志	实例启停等
网络	SLB	网络日志	/
	VPC Flow	网络日志	/
	CEN	网络日志	/
	API Gateway	访问日志	/
安全	安骑士	7类日志 (主机登录, 进程, 网络连接, 账户/端口Snapshot等),基线, 报警, 漏洞日志	主机内登录/账户/网络的检查日志
	WAF	访问日志	攻击日志
	DDOS高防	访问日志(7层)	攻击日志
	态势感知	14类日志 (主机7类, 网络- DNS/ Web/Session,基线/告警/漏洞日志	对ECS资源的主机内或网络层的操作日志
	反爬管理	访问日志	攻击日志
	数据库审计	SQL审计日志	/
	风险识别	安全识别与结果日志	/
	云防火墙	互联网日志	/
存储/数据库	OSS	资源操作, 数据操作, 数据访问日志, 计量日志	过期文件删除日志, CDN回流日志
	CDN	访问日志	/
	RDS	SQL审计日志	/
	DRDS	SQL审计日志	/
	NAS	访问日志	/
	SLS	资源操作, 数据操作, 数据访问日志	数据采集, 报警等日志
	Table Store	资源操作	/

* 仅列出部分产品

面向Sec/Dev/IT Ops 日志处理，分析引擎

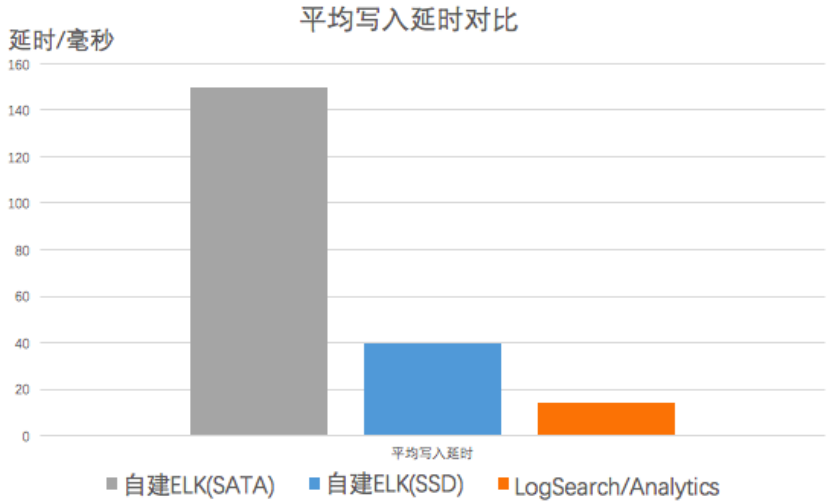


丰富可视化（所见即所得）　　效率（十亿秒级响应）　　智能（AIOps算法加持）

低成本（ELK方案15%费用）　　海量（PB/Day）　　功能丰富（阿里集团同一个版本，迭代快）

为日志处理设计存储

- 性能强劲：高可靠性SLA>99.9%；低延时（比ELK写入低70%）
- 存储功能：生命周期TTL调节（1 Day → 永久）；高可靠性（99.999999999%）；免运维 （MB → PB 弹性扩展）
- 压缩能力：高压缩率，存储空间比ES 低 55%
- 导出：与6+种流计算，5+大数据系统兼容；最大5GB/S导出速度



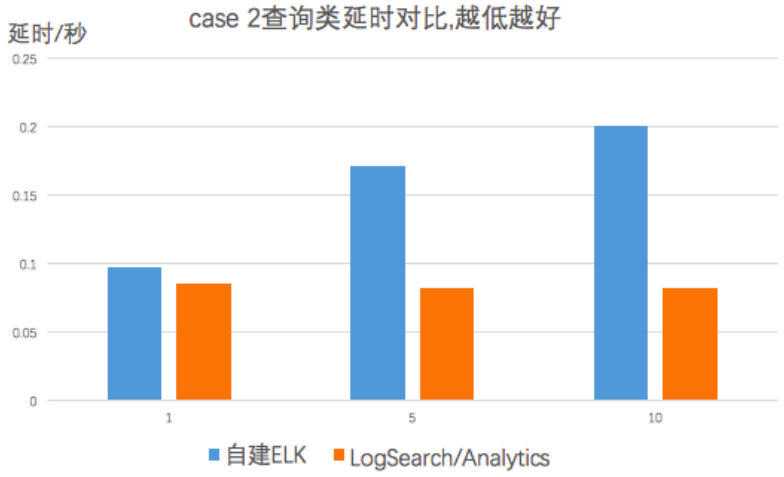
写入对比：
- 使用50KB测试数据，根据默认配置写入
- 延时比SSD自建ELK低 70%，稳定可靠

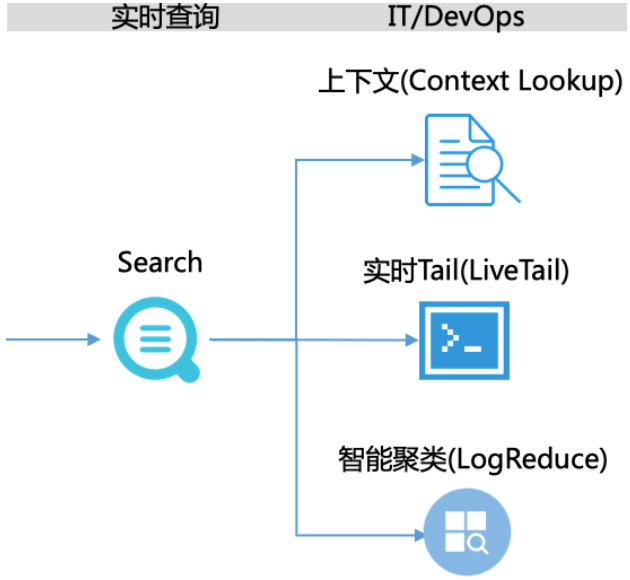
	ES	SLS（LogSearch/Analytics）
配置	单服务节点云盘	3 服务节点（对用户透明）
50GB存储	110GB（推荐2.2倍原始大小）	50G
存储单价	1元/GB*Month	0.35 元/GB*Month
月价	110 元	17.5 元
价格对比		15.9%
	需要预留磁盘大小	无需预留

存储对比：
- 在完全随机数据下，压缩率是ES 6.X版本53%
- 无数据膨胀（例如原始数据+索引Merge+最终数据等）
- 单位存储成本为ES 15.9%，并无预留空间（例如85%水位线）

日志原生查询

- 丰富类型：文本/JSON/Double/Long；原生支持中文；智能识别JSON类型
- 速度：100亿级数据 5+条件，5秒内返回
- 查询能力增强：
  - 上下文查询：原始上下文翻页，免登服务器
  - LiveTail功能：原始上下文tail-f，更新实时情况
  - 智能聚类：根据日志Pattern动态归类，合并重复模式，洞察异常



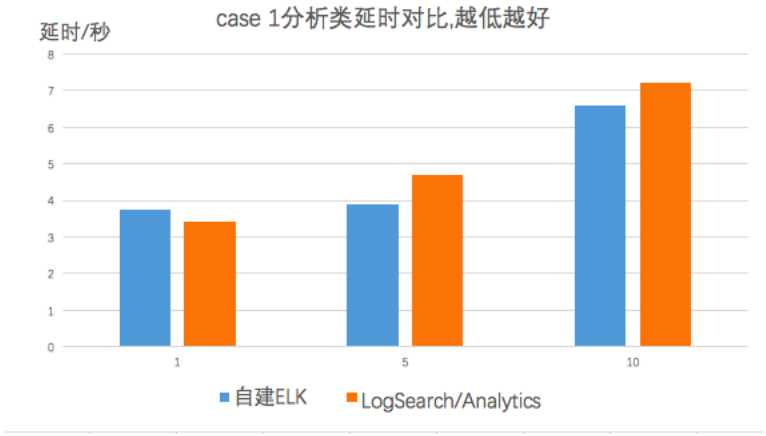


写入测试：
- 1.5亿条完全随机数据，并发1/5/10查询
- 延时优于ES
- 随着并发增加，延时保持稳定

日志聚类：
- 自动支持任意格式日志：Log4J、Json、单行（syslog）
- 亿级数据，秒级出结果，动态调整Reduce精度
- 支持筛选后聚类，原始日志查看

日志原生分析函数（SQL功能，ES速度）

- 完整SQL92标准，JDBC完整协议，支持Join
- 时序分析、机器学习与建模
- 支持外部数据源联合查询：OSS/MySQL
- 分析函数：同环比/IP经纬度/恶意IP访问/SQL注入分析等函数
- 性能：10亿级数据秒级返回



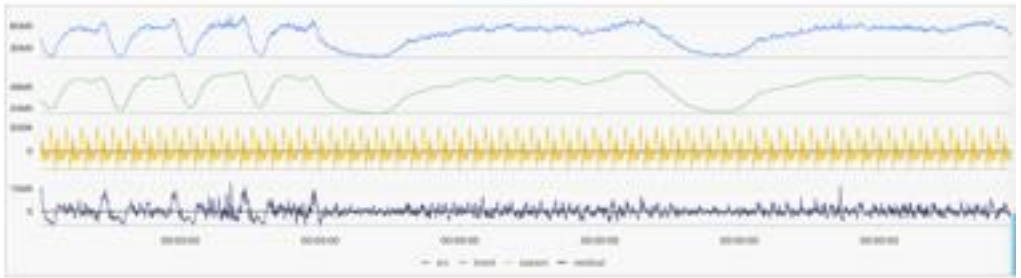
测试：

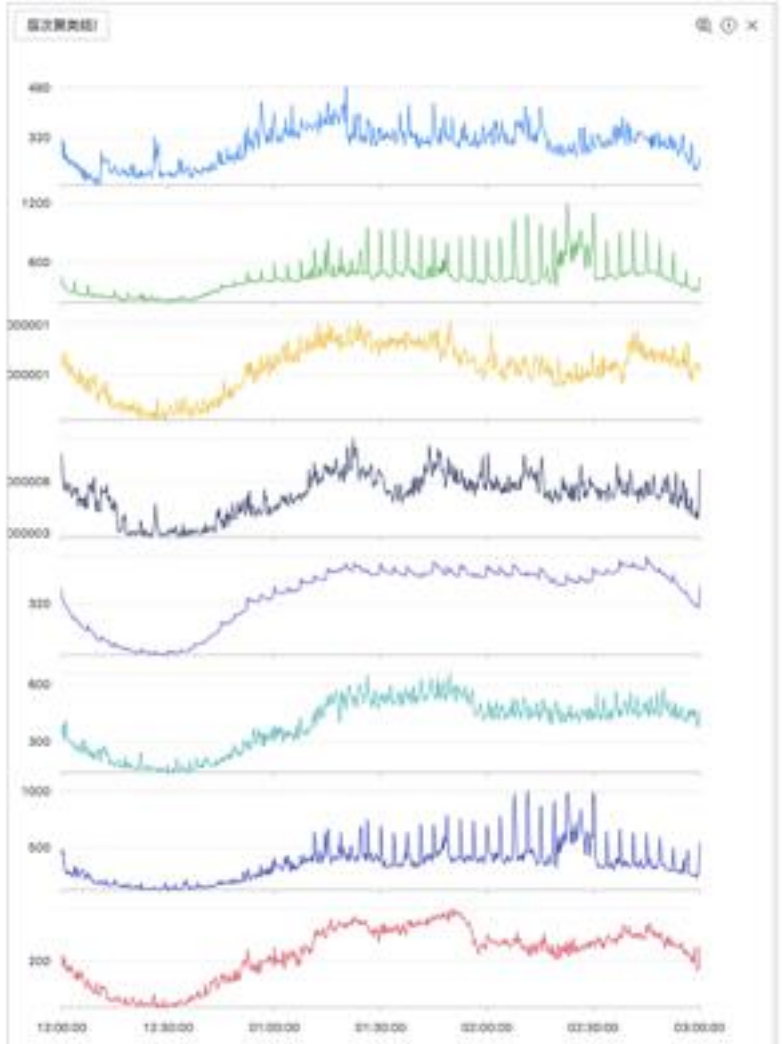
- 1.5亿条完全随机数据，并发测试5个维度聚合
- ES和日志服务延时处同一量级
- ES使用SSD盘，在读取大量数据时IO优势比较高

日志服务		ES
SELECT聚合计算函数： • 通用聚合函数 • 安全检测函数 • Map映射函数 • 估算函数 • 数学统计函数 • 数学计算函数 • 字符串函数 • 日期和时间函数 • URL函数 • 正则式函数 • JSON函数 • 类型转换函数 • IP地理函数 • 数组 • 二进制字符串函数 • 位运算 • 同比和环比函数 • 比较函数和运算符 • lambda函数 • 逻辑函数 • 空间几何函数 • 地理函数 • 机器学习函数 • 电话号码函数 GROUP BY 语法 窗口函数 HAVING语法 ORDER BY语法 LIMIT语法 CASE WHEN和IF分支语法 unnest语法 列的别名 嵌套子查询	VS	有限聚合计算 Group BY语法

机器学习与异常检测函数

- 预测：根据历史数据拟合基线
- 异常检测、变点检测、折点检测：找到异常点
- 多周期检测：发现数据访问中的周期规律
- 时序聚类：找到形态不一样的时序



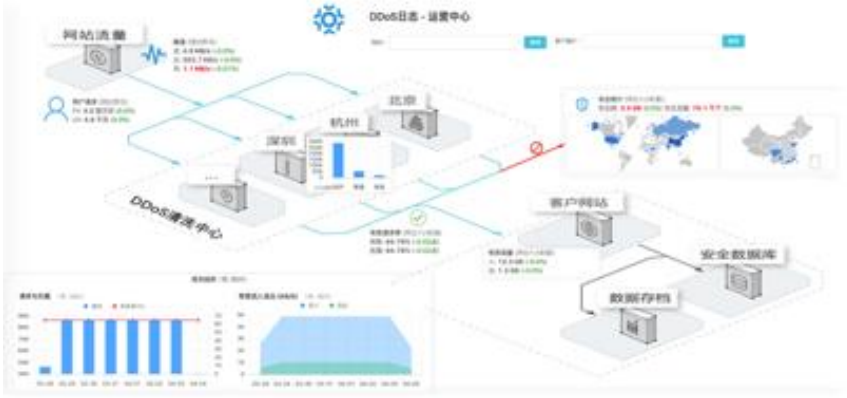


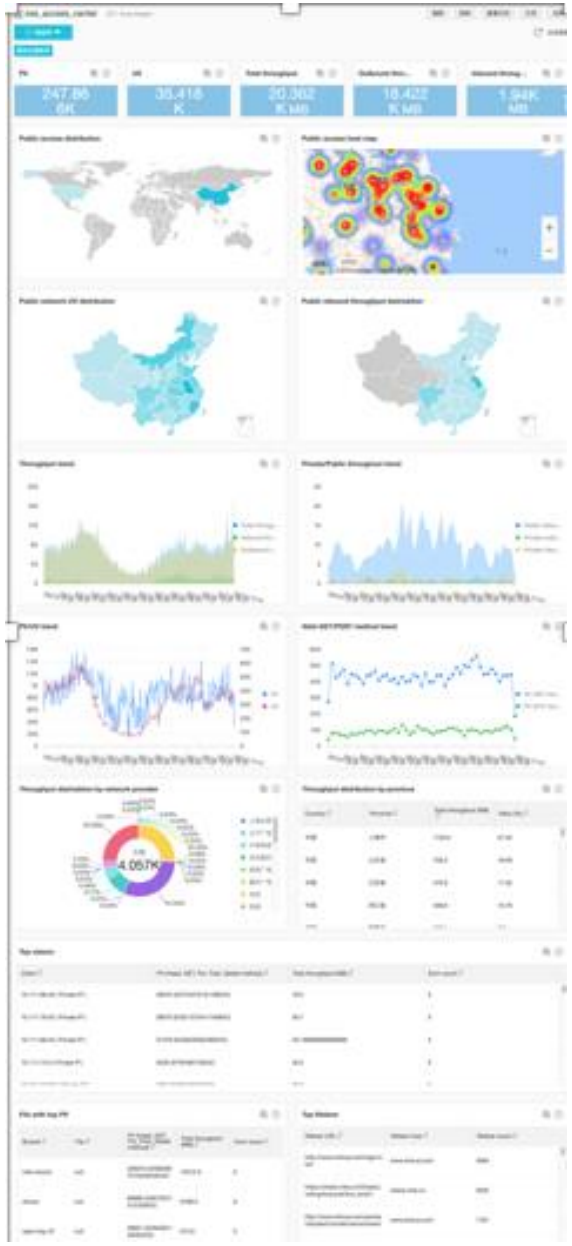
文档： https://help.aliyun.com/document_detail/93024.html

交互式可视分析仪表盘

原生：

- 所见即所得，支持Drill Down/Roll Up
- 20+图形，自定义画布（Canvas）
- 支持告警、报告自动发送





生态：

- API、JDBC两种接口
- 与生态打通（DataV、QuickBI、Grafana、Zipkin、Jeager）

生态大图（黄色部分为日志服务）

30+ 日志采集手段

采集

Mobile & Web
Mobile
Web

Text & Logs
logtail Logs
fluentd LOG4J

Services & Languages
Java JS C++ GO php

IoT & Devices
Camera
IoT

存储

Massive Structured, Unstructured & Semi-structured Data
OSS CSV JSON Parquet
Hybrid Storage Array
SQL、NoSQL

Log Service / LogShipper

Real-time Data Stream
Log Service / LogHub

分析

Big Data Analytics
MaxCompute
Tensorflow caffe PAI

Batch Processing
Pig Spark HIVE
EMR Hadoop

Interactive Analytics
DLA presto Impala

Log Service / Analytics

Stream Processing
Spark SparkStreaming
Function Compute

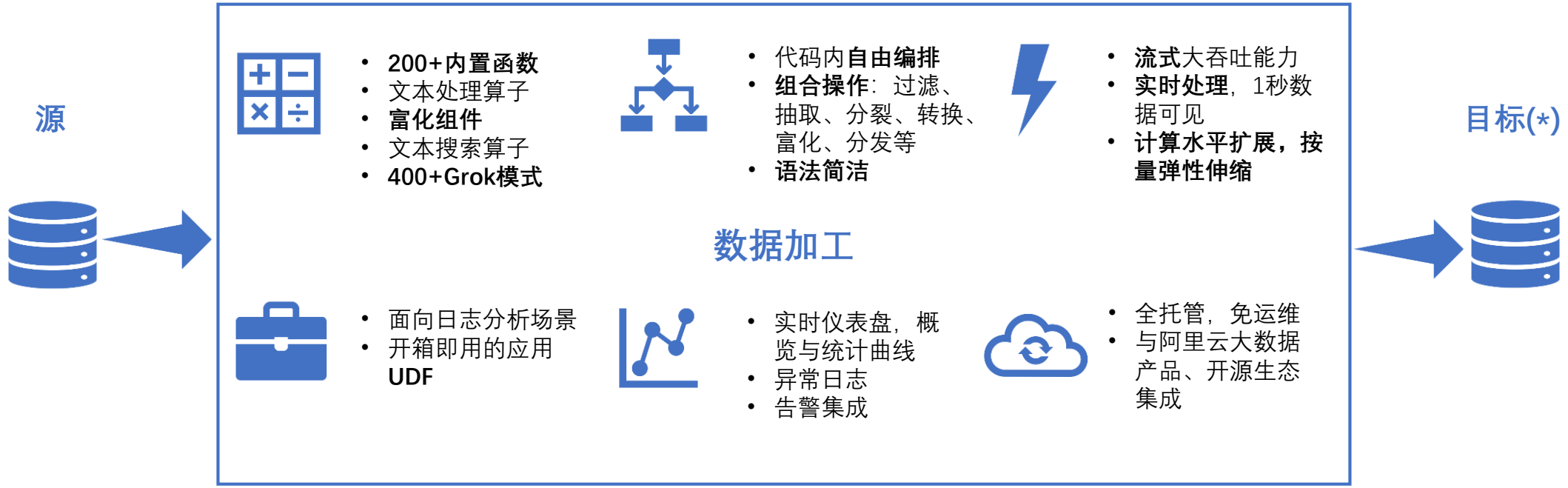
可视化

Visualization
QuickBI
DataV
JDBC +ableau

Log Service / Dashboard

数据加工：面向日志分析的实时、灵活加工

- 功能概述（参考链接）
  - 将各类日志处理为结构化数据，具备全托管、实时、高吞吐的特点
  - 面向日志分析领域，提供丰富算子、开箱即用的场景化UDF（Syslog、非标准json、AccessLog UA/URI/IP解析等）
  - 丰富的阿里云大数据产品（OSS、MC、EMR、ADB等）、开源生态（Flink、Spark等）集成能力，降低数据分析门槛
- 典型场景
  - 数据规整：对混乱格式的日志进行字段提取、格式转换，获取结构化数据以支持后续的流处理、数仓计算
  - 数据富化：日志（例如业务订单）与维表（例如用户信息MySQL表）进行字段join，为日志添加更多维度信息供分析
  - 数据分发：将全量日志按转发规则分别提取到多个下游存储供不同业务使用



数据加工：解决痛点与场景

解决数据加工的痛点

- 行业上80%的数据分析花费在数据规整上
  - 数据在接入、分析、投递、对接时存在各种ETL需求与痛点
  - 痛点参考：
    1. 数据源混合各种格式，难以简单提取，如交换机、服务器、容器、程序Logging模块等，通过文件、stdout、syslog、网络等途径收集的数据，混合了程序的各种格式的日志；如日志中的message字段，需要根据各种情况的正则匹配后提取；
    2. 单一场景，字段动态且不确定，例如NGNIX，的QueryString、HttpCookie、HttpBody信息中的字段，需要用自动KV或者多种正则提取；
    3. 源数据包含动态JSON格式的数据（如CVE数据、O365 Audit日志等），需要动态计算，提取合并字段，甚至分裂为多条日志进行处理；
    4. 某些常规日志包含了敏感信息（例如秘钥的密码，用户手机号、内部数据库连接字符串等），很难在提取时过滤掉或者做脱敏。
    5. 使用简单表格、CSV、OSS文件、RDS、外部API等进行数据富化。

场景1 – 数据规整（一对一）

源Logstore
（终端syslog日志）

```
2018-1-1 INFO user1 login server from 10.3.4.1 type=rdp
2018-1-1 INFO user2 fail to login system type=ssh IP=10.1.1.2
2018-1-3 INFO user3 use OAuth to login via O365
2018-1-3 INFO user4 login with 2 MFA {f1: mobile, f2: face}
2018-1-3 INFO user5 create project, delete object
2018-1-3 INFO user6 start process file://abc.exe
```

目标Logstore
（登录信息、
Audit信息、进程
信息）

User id	Name	Login Type	Login Date	Login Result
123123	Jane	SSH	2018-1-1	Pass

User id	Name	Op Type	Target	Result
123123	Jane	Create	project	Success

User id	Name	Process	Process id	Result
123123	Jane	Abc.exe	100	Success
123123	John	Xyz.exe	200	Fail
123123	Jack	Lass.exe	101	Success

注册用户信息

User id	Name	Device Sig	Register date	Mobile	…
123123	Jane	Nokia 100	2018-1-1	13912919191	…
123123	John	Huawei P20	2018-2-2	13912919192	…
123123	Jack	iPhone MX	2018-3-3	13912919193	…

场景2 – 数据分派（一对多）

目标Logstore1
（登录信息）

User id	Name	Login Type	Login Date	Login Result
123123	Jane	SSH	2018-1-1	Pass
123123	John	RDP	2018-1-1	Fail
123123	Jack	SSH	2018-1-1	Pass

源Logstore
（App日志）

```
2018-1-1 INFO user1 login server from 10.3.4.1 type=rdp
2018-1-1 INFO user2 fail to login system type=ssh IP=10.1.1.2
2018-1-3 INFO user3 click page1 user agent
2018-1-3 INFO user3 add product id=123 into shopping cart
2018-1-3 INFO user3 click adds position x=1 y=2 width=100
```

数据加工服务

过滤、富化、转换
补漏、分裂、聚合

目标Logstore1
（浏览信息）

目标Logstore1
（下单信息）

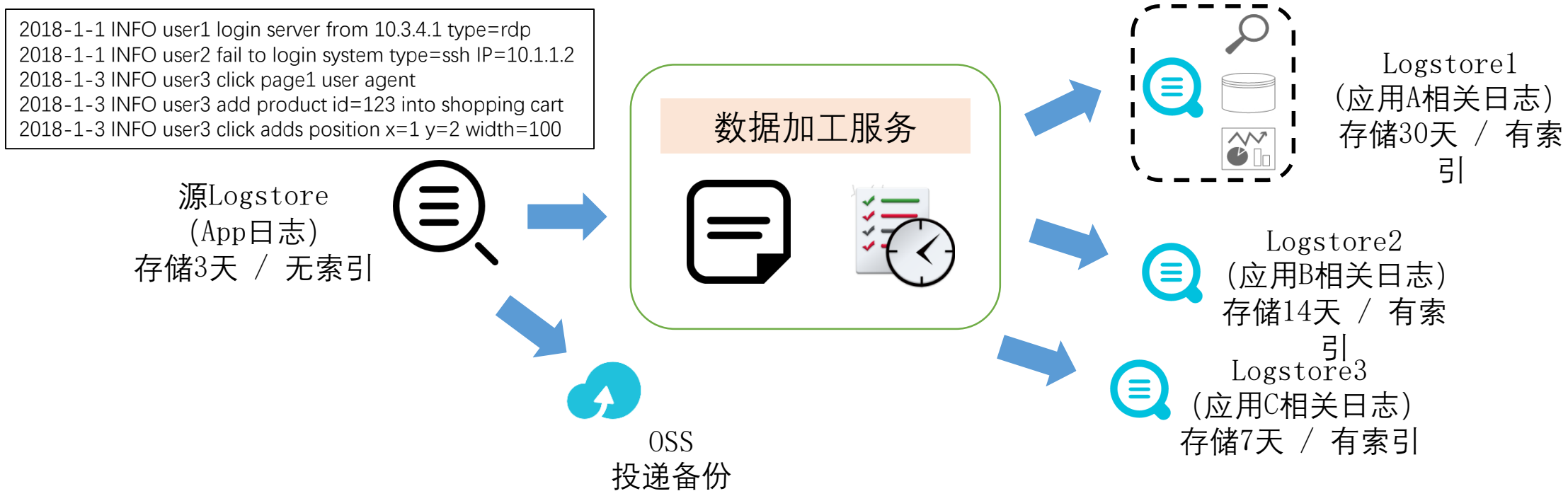
商品信息

Product ID	Name	Spec	Price	Stock	…
111	P20	Huawei P20	200	100	…
221	iPhone	iPhone MX	300	20	…

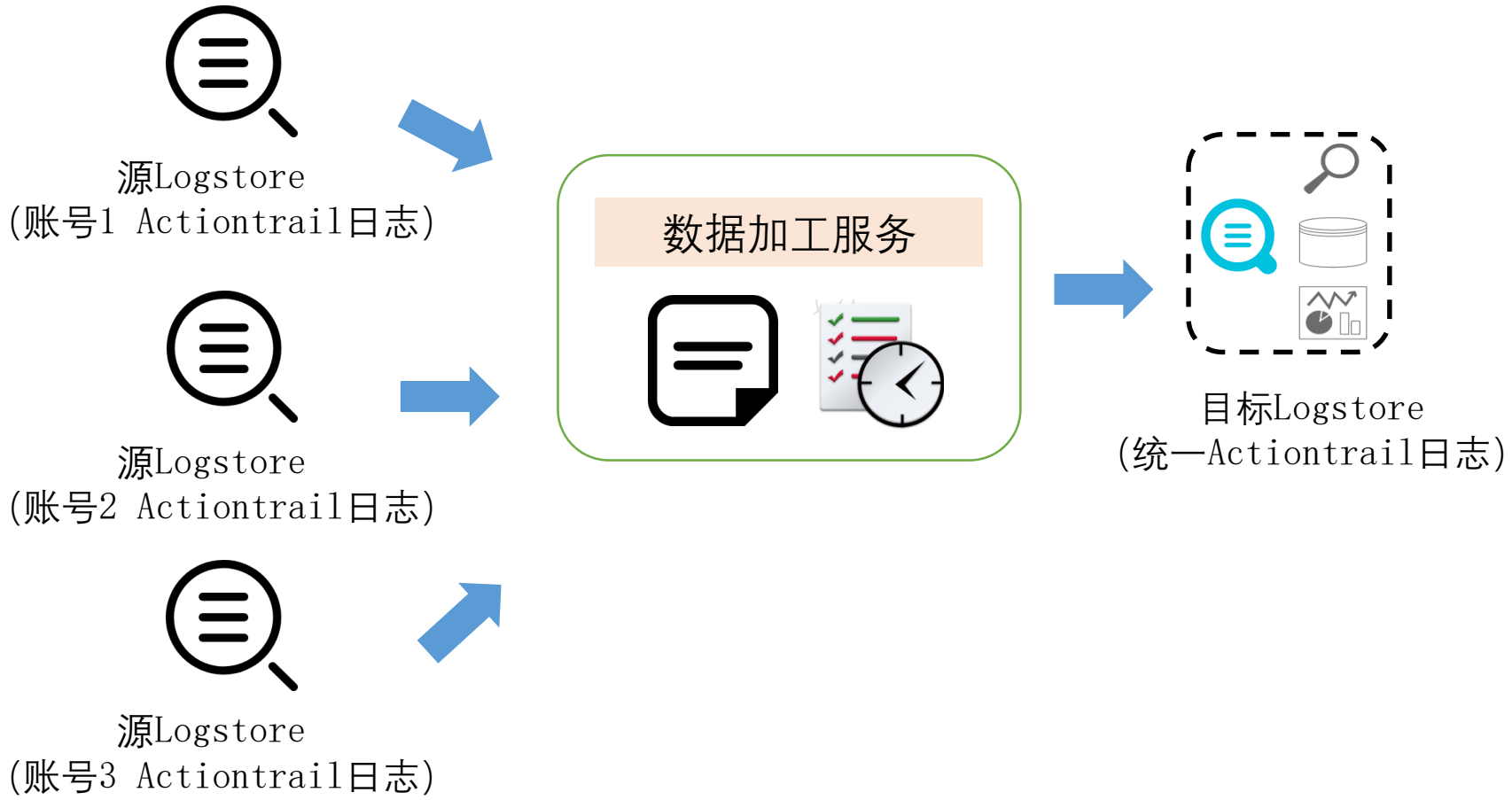
注册用户信息

User id	Name	Device Sig	Register date	Mobile	…
123123	Jane	Nokia 100	2018-1-1	13912919191	…
123123	John	Huawei P20	2018-2-2	13912919192	…
123123	Jack	iPhone MX	2018-3-3	13912919193	…

典型分发规划


2018-1-1 INFO user1 login server from 10.3.4.1 type=rdp
2018-1-1 INFO user2 fail to login system type=ssh IP=10.1.1.2
2018-1-3 INFO user3 click page1 user agent
2018-1-3 INFO user3 add product id=123 into shopping cart
2018-1-3 INFO user3 click adds position x=1 y=2 width=100
源Logstore
(App日志)
存储3天 / 无索引
数据加工服务
OSS
投递备份
Logstore1
(应用A相关日志)
存储30天 / 有索引
Logstore2
(应用B相关日志)
存储14天 / 有索引
Logstore3
(应用C相关日志)
存储7天 / 有索引

场景3 – 多源汇集（多对一）


源Logstore
(账号1 Actiontrail日志)
源Logstore
(账号2 Actiontrail日志)
源Logstore
(账号3 Actiontrail日志)
数据加工服务
目标Logstore
(统一Actiontrail日志)

场景4 – 数据加工场景

- 过滤（filter）：将特定的日志去掉
- 分裂（split）：将一条日志变成多条
- 转换（transform）：字段操作、内容转换等
- 富化（enrich）：关联外部资源，丰富字段信息等
- 聚合（Rollup）：特定维度做聚集，减少日志量（**计划推出**）
- UDF：完全自定义操作，如SQL模式解析、NLP解析等（内部开放）

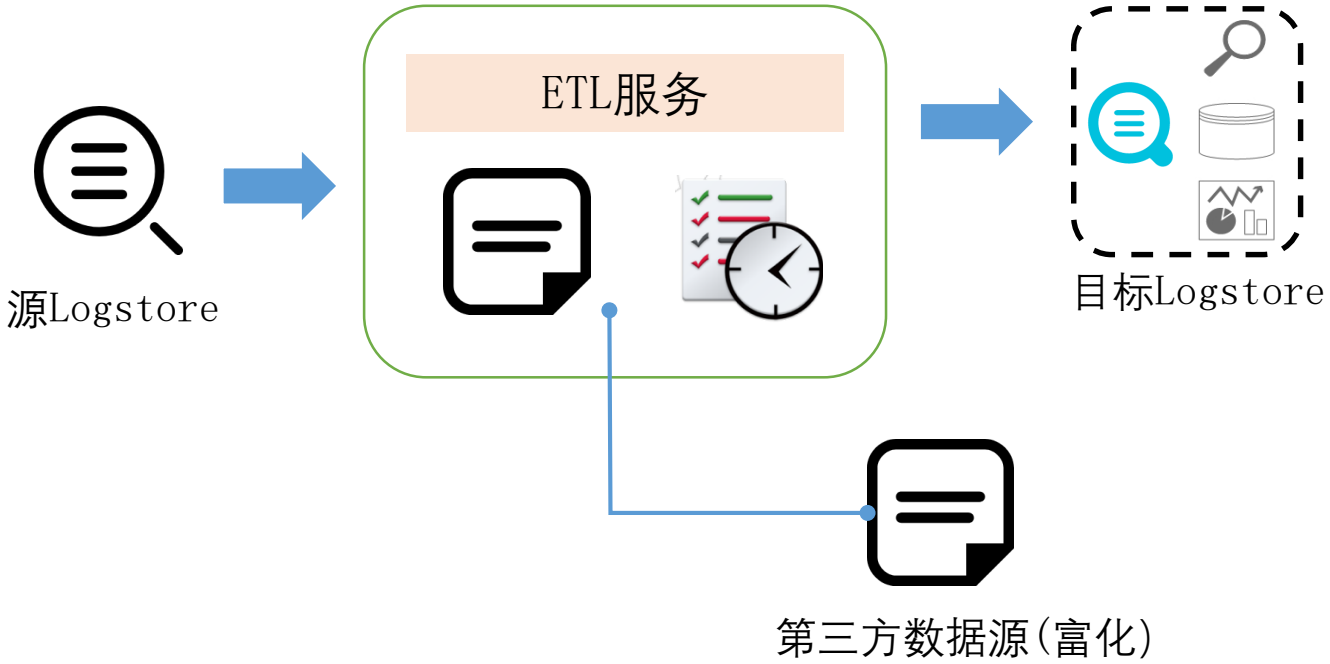
功能配置、原理

源与目标

- 目前是Region化（跨Region即将推出）
- 支持跨project
- 支持跨账号
- 支持最多20个静态目标或无限动态目标

配置-授权

- 当前操作需要授权以便读取源logstore或写入目标logstore
  - 通过AK授权
  - 通过角色授权（计划推出）
- 连接第三方数据用于富化的授权通过配置中的秘钥项目完成



源/目标logstore最小权限

源Logstore

```
{
  "Version": "1",
  "Statement": [
    {
      "Action": [
        "log:ListShards",
        "log:GetCursorOrData",
        "log:GetConsumerGroupCheckPoint",
        "log:UpdateConsumerGroup",
        "log:ConsumerGroupHeartBeat",
        "log:ConsumerGroupUpdateCheckPoint",
        "log:ListConsumerGroup",
        "log:CreateConsumerGroup"
      ],
      "Resource": [
        "acs:log:*:*:project/源project/logstore/源logstore",
    "acs:log:*:*:project/源project/logstore/源logstore/*"
      ],
      "Effect": "Allow"
    }
  ]
}
```

目标Logstore

```
{
  "Statement": [
    {
      "Action": [
        "log:Post*"
      ],
      "Effect": "Allow",
      "Resource": "acs:log:*:*:project/目标Project/logstore/目标Logstore"
    }
  ],
  "Version": "1"
}
```

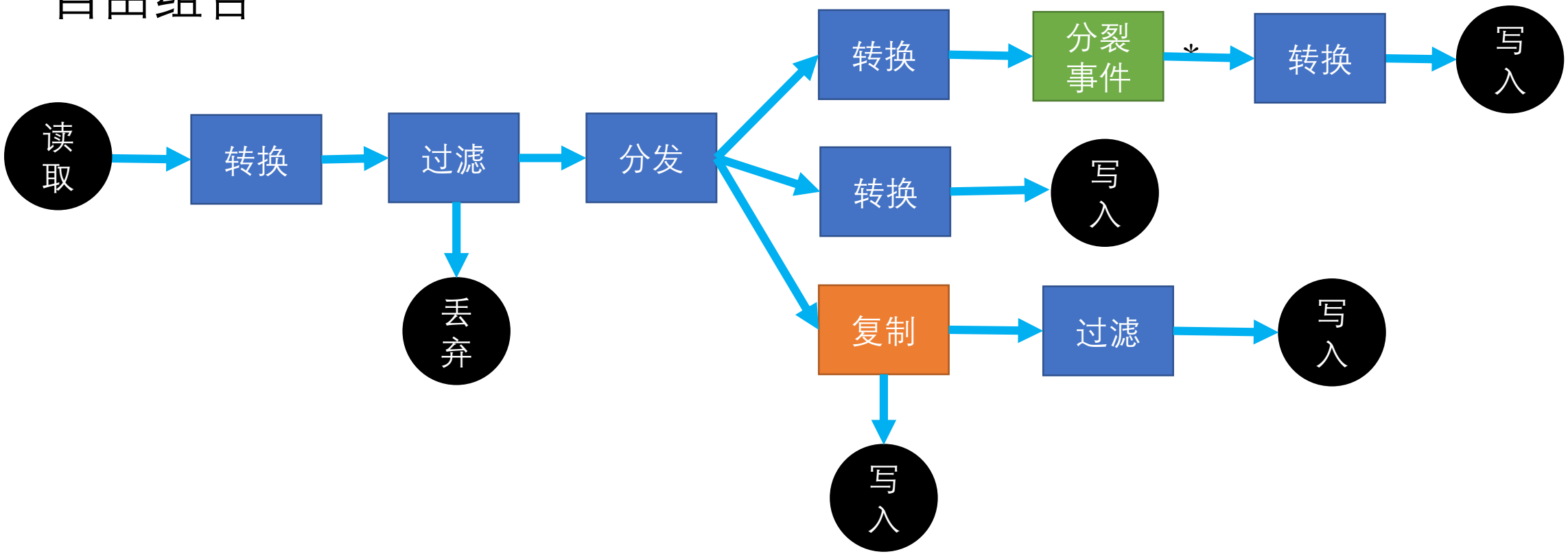
调度规则

方案与特点	持续加工	一次性加工
消费范围（配置）	消费时间起点（**日志接收时间**或begin）- 仅在初次消费时有用	时间起止点（**日志接收时间范围**）
启动与消费行为	启动后从指定点开始消费，持续下去，不会停止（除非人介入）	启动后消费指定范围（或无数据）后停止 – **任务结束**
实时性	实时	非实时
适用场景	常规实时加工	历史数据加工或特定场景（如校验、测试、演示等）

自由编排DSL

200+函数，400+GROK模式

自由组合



数据加工DSL能力

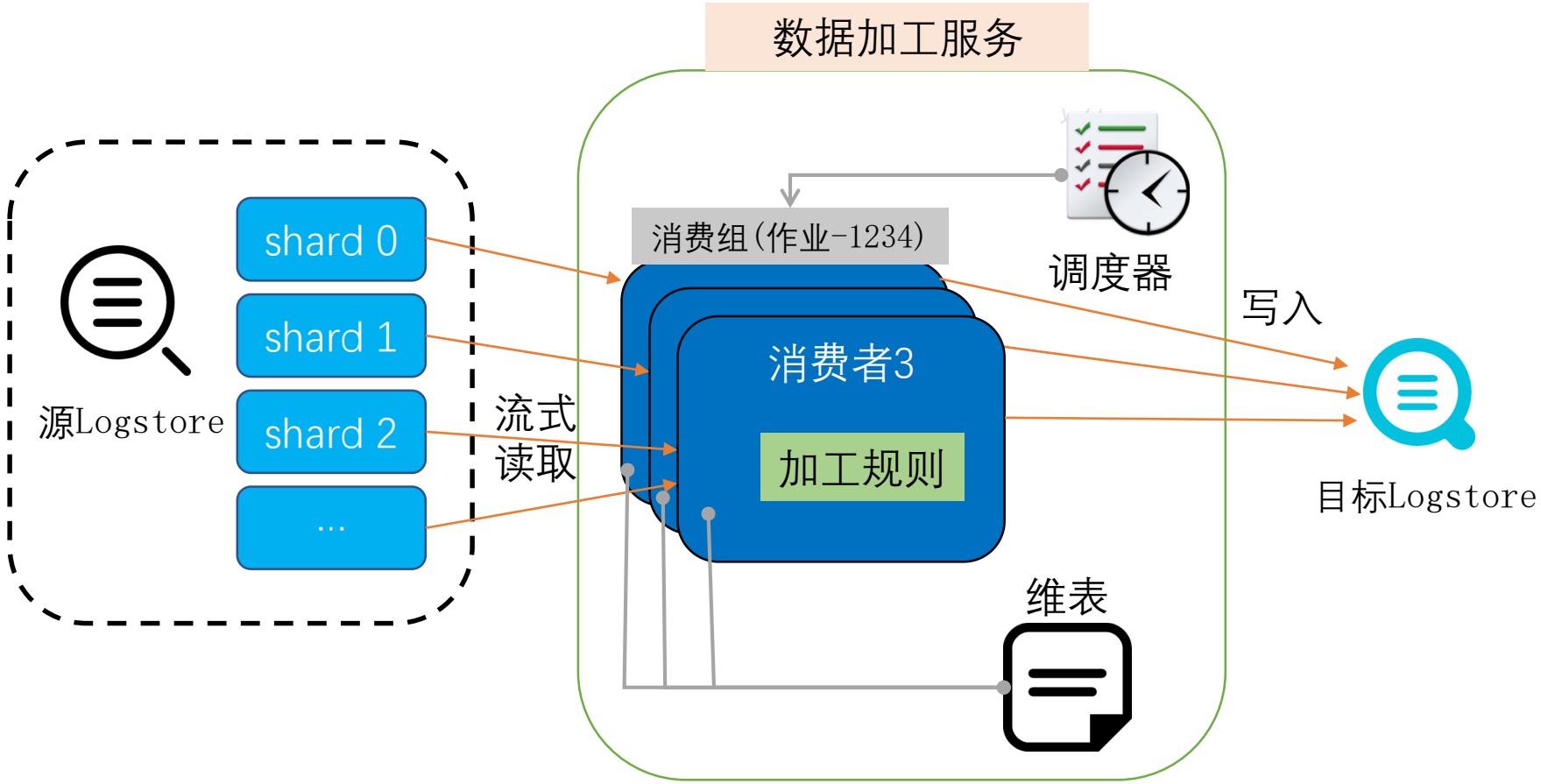
- 语法简洁且支持UDF(内部)
- 30+全局函数
- 200+内置函数
- 数学/文本/时间/IP/处理算子
- OSS/Logstore/RDS-MySQL/IP库/配置资源连接器
- 400+Grok模式
- 类Lucene事件搜索算子
- JMES语法支持
- 代码内自由编排
- 组合操作：过滤、抽取、分裂、转换、富化、分发等

类型	函数	说明
流程控制	e_if	多个条件操作的配对操作
流程控制	e_if_else	if-else操作
流程控制	e_switch	满足一个条件操作后跳出
流程控制	e_compose	组合操作
事件操作	e_drop	丢弃
事件操作	e_keep	保留
事件操作	e_split	分裂
事件操作	e_output	输出
事件操作	e_coutput	复制输出
字段操作	e_drop_fields	删除
字段操作	e_keep_fields	保留
字段操作	e_rename	重命名
字段值赋值	e_set	赋值
字段值提取	e_regex	正则提取
字段值提取	e_json	json展开或提取
字段值提取	e_kv	自动提取键值对
字段值提取	e_kv_delimit	基于分隔符提取键值对
字段值提取	e_csv	逗号或其他分隔符提取
字段值提取	e_tsv	tab分隔符提取
字段值提取	e_psv	pipe分隔符提取
字段值提取	e_syslogrfc	根据syslog协议提取头
字段富化	e_dict_map	字典映射
字段富化	e_table_map	表格映射
字段富化	e_search_map	搜索映射
任务配置	res_local_update	设置任务参数上下文

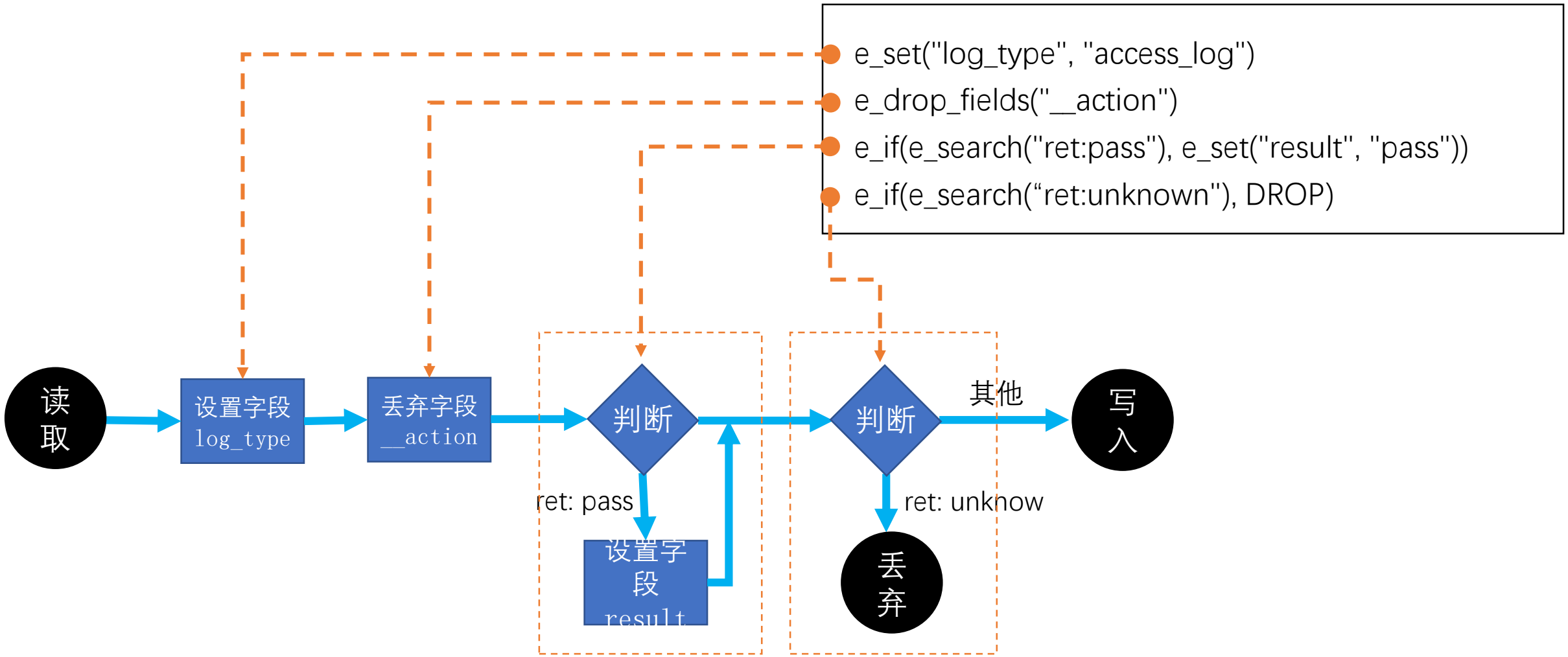
类型	函数	说明
事件检查函数	v, e_has, e_not_has, e_search, e_match, e_match_any, e_match_all等	获取事件字段值, 或判断字段或字段值是否符合特定内容
操作符函数	部分`op_*` 函数	比较, 条件判断, 容器类计算, 一般性多值操作
转换函数	`ct_*`函数	数字,字符串,布尔之间的转换, 数字进制转换
算术函数	部分`op_*`函数, `math_*`函数等	数字的`+-*/`幂等计算, 数学计算, 多值计算等
字符串函数	`str_*`函数	字符串的所有相关操作与判断搜索等
日期时间函数	`dt_*`函数	Unix时间戳, 日期时间对象, 日期时间字符串转化, 时区调整, 圆整等
正则表达式函数	`reg_*`函数	正则提取, 检索, 替换, 分裂多值等
GROK函数	`grok`函数	提取grok模式返回对应正则表达式
JSON, XML, Protobuf函数	`json_*`, `xml_*`, `pb_*` 函数	对应提取或解析
编码解码类函数	`url_*`, `html_*`, `md5_*`, `sha1_*`, `base64_*`函数	相关单向或双向函数
列表函数	`lst_*`函数	列表相关构建,获取, 修改, 操作等
字典函数	`dct_*`函数	字典相关构建,获取, 修改, 操作等
表格函数	`tab_*`函数	从文本解析出表格, 表格转字典等
资源函数	`res_*`函数	本地配置, RDS, Logstore等资源获取

原理

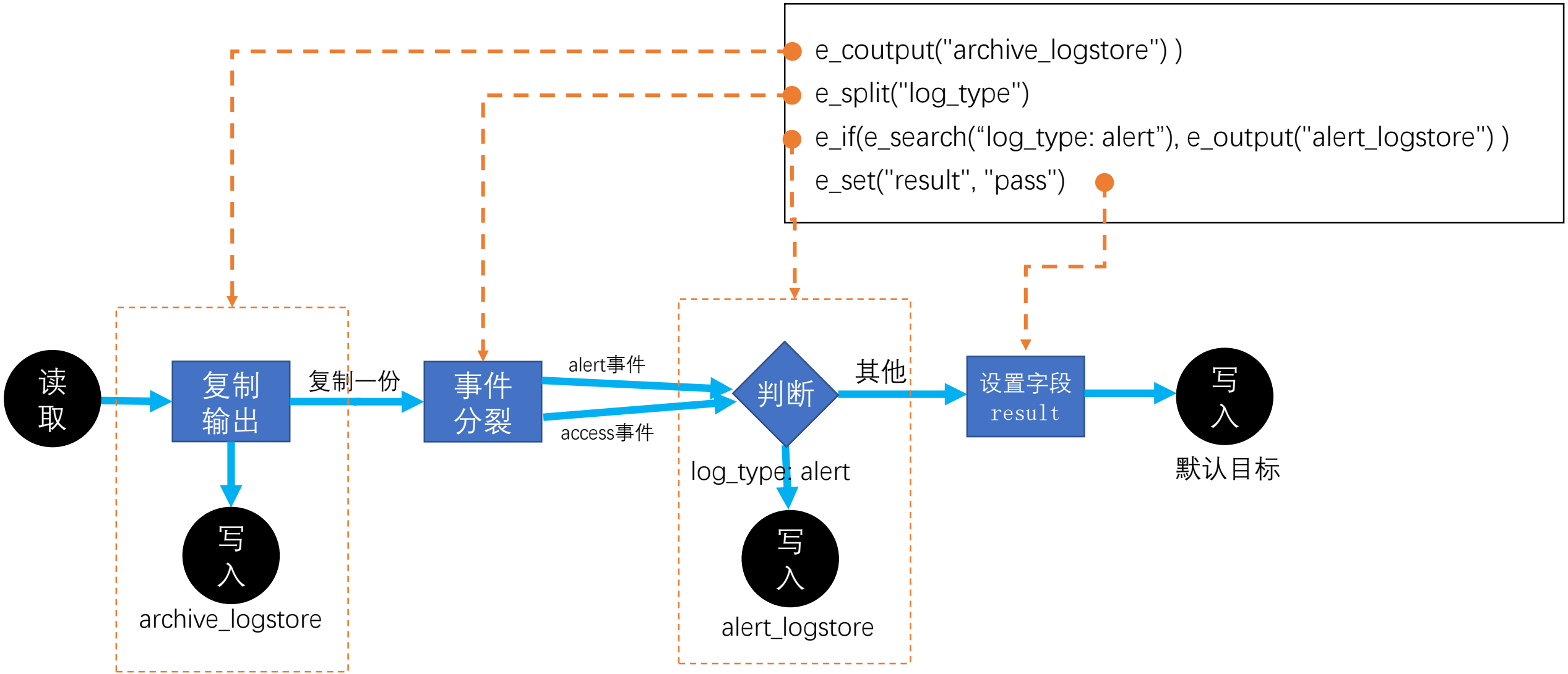
调度原理


数据加工服务
shard 0
shard 1
shard 2
...
源Logstore
流式
读取
消费组(作业-1234)
消费者3
加工规则
调度器
写入
目标Logstore
维表

规则引擎原理：基本操作


e_set("log_type", "access_log")
e_drop_fields("__action")
e_if(e_search("ret:pass"), e_set("result", "pass"))
e_if(e_search(“ret:unknown"), DROP)
读取
设置字段
log_type
丢弃字段
__action
判断
ret: pass
设置字段
result
判断
其他
写入
ret: unknow
丢弃

规则引擎原理：输出，复制与分裂

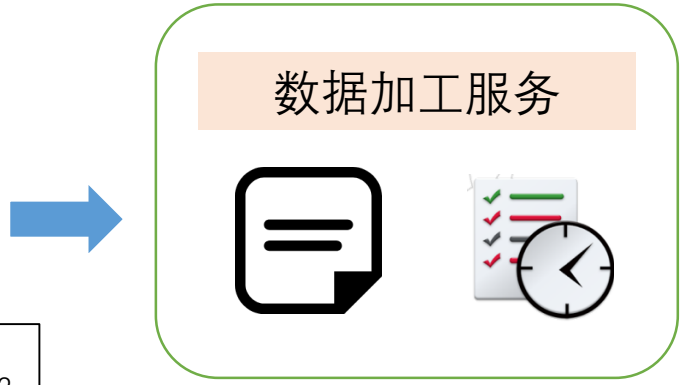


Demo

典型分发规划

SLB日志

2018-1-1 INFO user1 login server from 10.3.4.1 type=rdp
2018-1-1 INFO user2 fail to login system type=ssh IP=10.1.1.2
2018-1-3 INFO user3 click page1 user agent
2018-1-3 INFO user3 add product id=123 into shopping cart
2018-1-3 INFO user3 click adds position x=1 y=2 width=100



Logstore1
(应用A相关日志)
存储30天 / 有索引

Logstore2
(应用B相关日志)
存储14天 / 有索引

Logstore3
(应用C相关日志)
存储7天 / 有索引

阿里云开发者社区

日志服务数据加工系列培训

欢迎加入
（上海、杭州）

主题：扫平日志分析路上障碍，实时海量日志加工实践培训

讲师：丁来强（成喆）- 阿里高级技术专家 | 唐恺(风毅) - 阿里技术专家

分享介绍

8月7日	8月8日	8月13日	8月14日	8月20日	8月21日	8月28日	8月29日
19:30-20:30	19:30-20:30	19:30-20:30	19:30-20:30	19:30-20:30	19:30-20:30	19:30-20:30	19:30-20:30
数据加工 介绍与实战	数据加工DSL 核心语法介绍	数据加工DSL 语法实践	数据加工动态 数据分发汇集实践	非结构化数据 解析实践	结构化数据 解析实践	数据映射 富化实践	数据加工 可靠性与排错实践