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

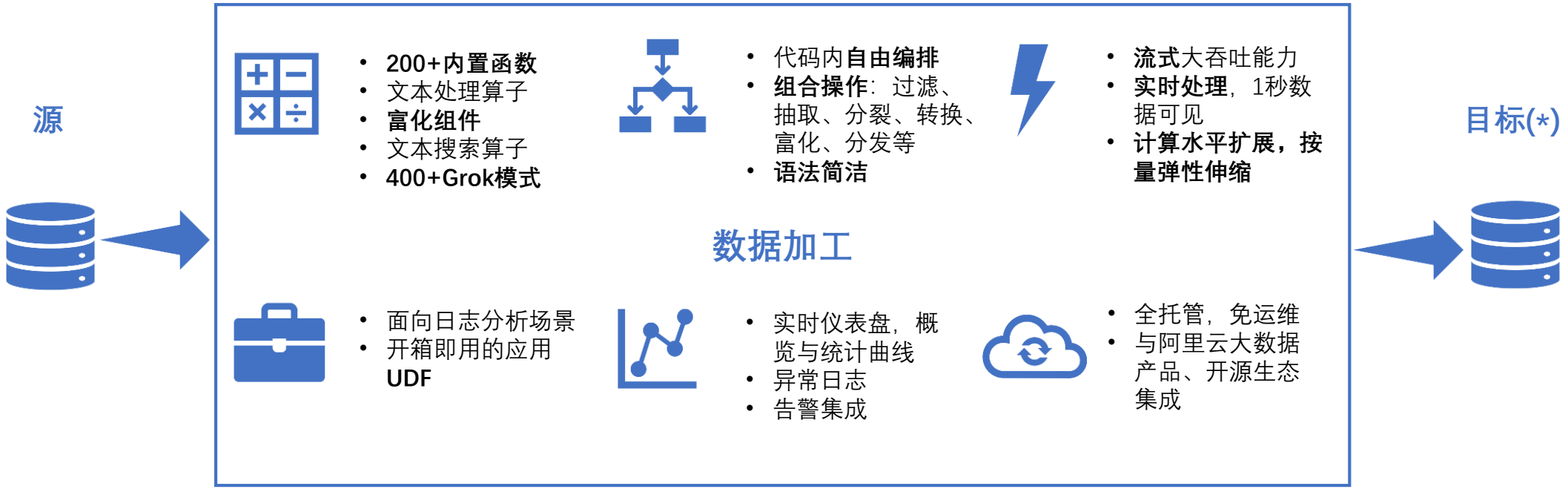
数据加工: 数据分发汇集实践

系列培训四

唐恺

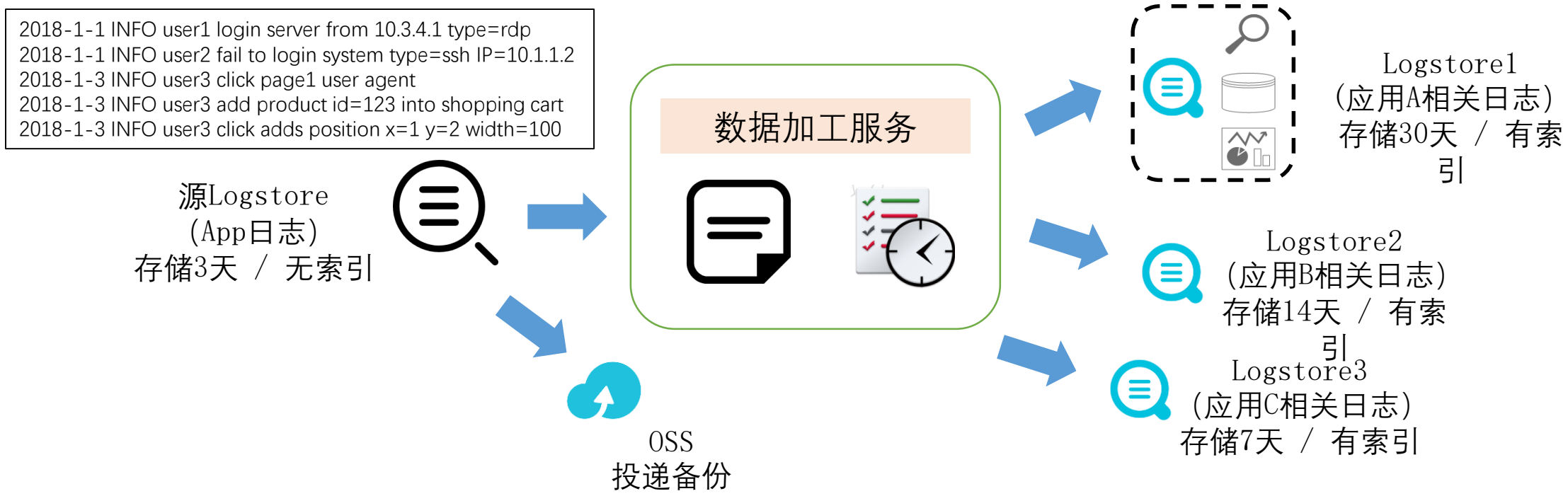
日志服务-数据加工简介

- 功能概述
  - 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



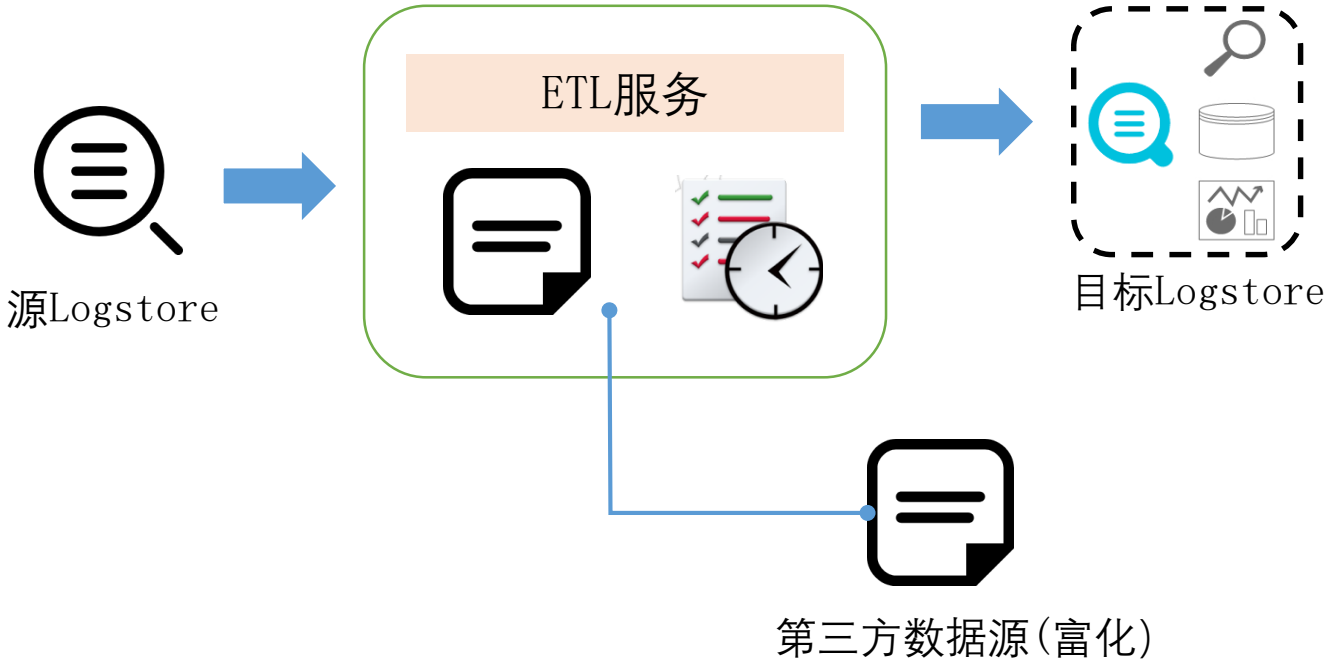
相关语法与原理

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

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



源与目标的细粒度权限

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"
}
```

流程控制函数

类型	函数	说明
组合	e_compose	组合一系列操作
流程控制	e_if	条件与操作的配对组合, 根据条件判断, 依次进行操作, 不满足任何条件时不进行对应操作
流程控制	e_if_else	根据条件判断, 进行真假对应的操作
流程控制	e_switch	条件与操作的配对组合, 根据条件判断, 根据条件判断, 进行操作, 只要满足条件了就执行对应的操作(或列表), 然后直接返回. 不再进行参数中后续的条件检查与操作.

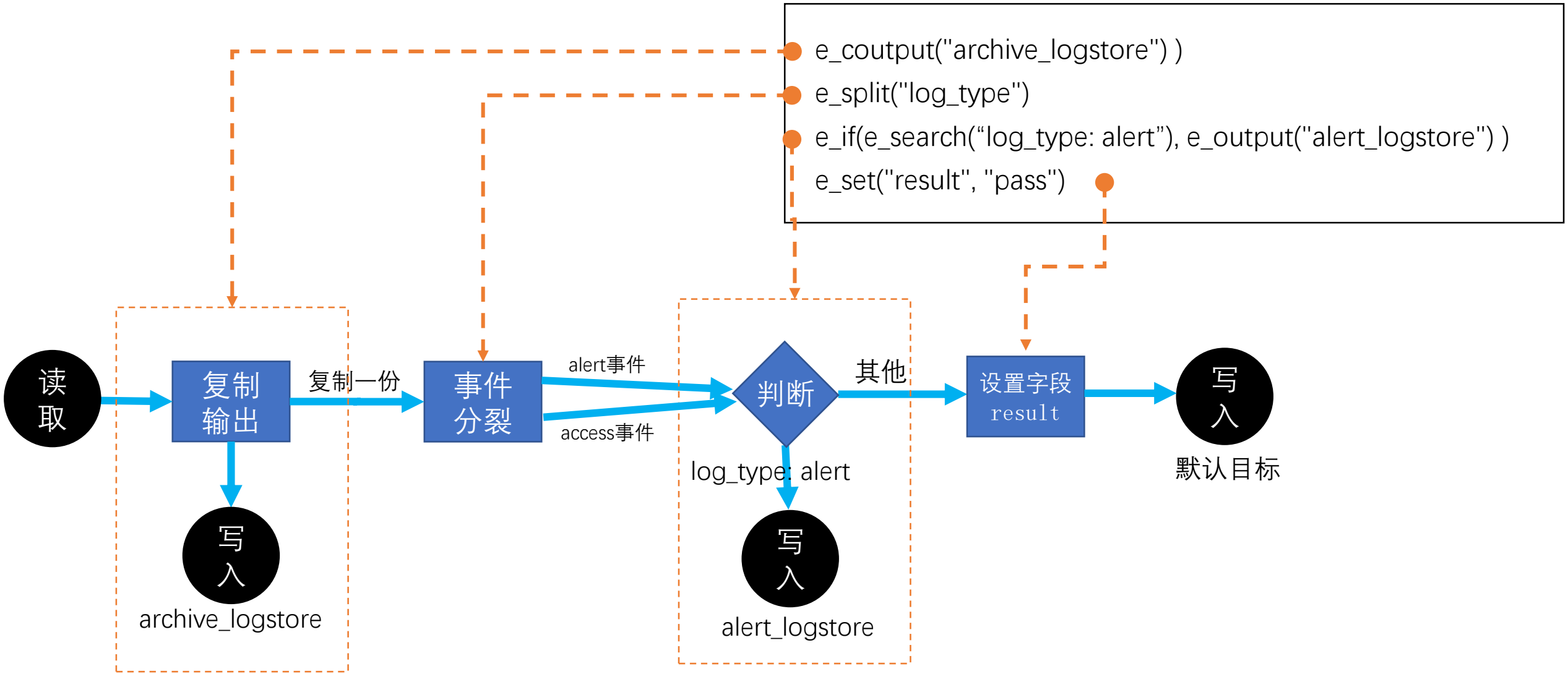
事件操作类函数

类型	函数	说明
事件操作	e_drop, DROP	根据条件, 丢弃事件
事件操作	e_keep, KEEP	根据条件, 保留事件
事件分裂	e_split	基于字段的值进行分裂出多个事件. 事件所有值都一样, 除了基于的字段的值是具体某一项。也支持基于JMES提取字段后再进行分裂更多参考复杂JSON处理和JMES的使用
输出事件	e_output	输出事件到配置的特定目标， 并配置输出时的topic、source、标签等信息。输出事件后会删除事件
输出事件	e_coutput	输出事件到配置的特定目标， 并配置输出时的topic、source、标签等信息。删除事件后, 会保留事件, 继续进行后续操作

检查与比较函数

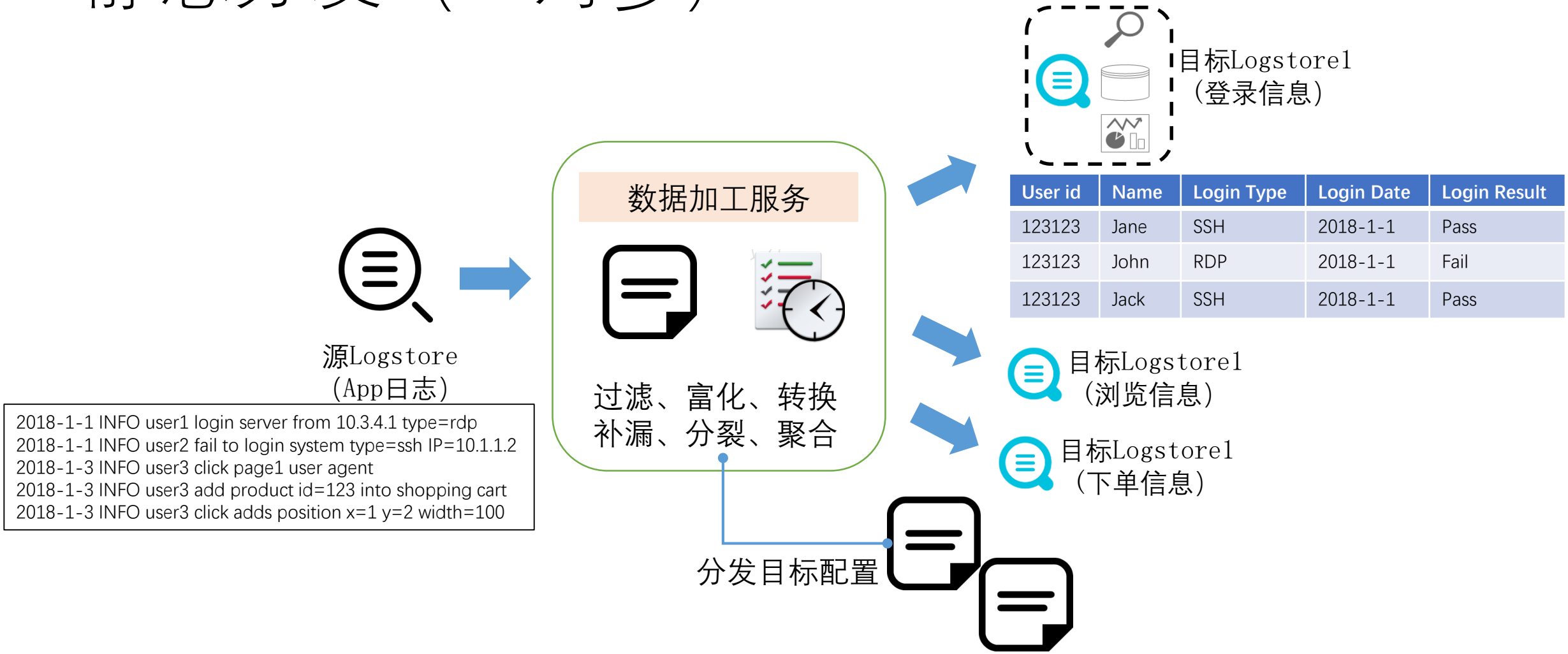
类型	函数	说明
事件检查函数	e_has e_not_has	获取事件字段值, 或判断字段或字段值是否符合特定内容
	e_match e_match_all e_match_any	使用正则匹配值
	e_search	接受搜索字符串, 支持正则
基础操作函数	op_and op_or op_not 等op_*系列函数	比较, 条件判断, 容器类计算, 一般性多值操作

规则引擎原理：输出，复制与分裂



实践1：基于静态配置的分发

静态分发（一对多）


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
分发目标配置
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
目标Logstore1
(浏览信息)
目标Logstore1
(下单信息)

DSL：e_switch/e_output结合运用

```
e_switch(
    # test log
    op_eq(v("__topic__"), "test"),
    e_output(name="foreign-target",
             project="etl-test-shenzhen-devops",
             logstore="test-log",
             topic="数据加工",
             tags={"upstream_project":"etl-test-shenzhen",
                   "upstream_logstore":"slb-layer7-accesslog"}),
    # slb log
    e_has("slbid"),
    e_output(name="default",
             tags={"upstream_project":"etl-test-shenzhen",
                   "upstream_logstore":"slb-layer7-accesslog"}),
    # metric log
    e_search("__topic__==metric and Level ~= 'error|info|warning|debug'"),
    e_compose(e_set("__topic__", v("Level")),
              e_drop_fields("Level"),
              e_output(name="foreign-target",
                       tags={"upstream_project":"etl-test-shenzhen",
                             "upstream_logstore":"slb-layer7-accesslog"}))
)

```

DSL：e_if/e_output/e_coutput结合运用

```
28# test log
29e_if(op_eq(v("__topic__"), "test"),
30     e_output(name="foreign-target",
31              project="etl-test-shenzhen-devops",
32              logstore="test-log",
33              topic="数据加工",
34              tags={"upstream_project":"etl-test-shenzhen",
35                    "upstream_logstore":"slb-layer7-accesslog"}))
36# slb log
37e_if(e_has("slbid"),
38     e_output(name="default",
39              tags={"upstream_project":"etl-test-shenzhen",
40                    "upstream_logstore":"slb-layer7-accesslog"}))
41# metric log
42e_if(e_search("__topic__==metric and Level ~= 'error|info|warning|debug'"),
43     e_compose(e_set("__topic__", v("Level")),
44               e_drop_fields("Level"),
45               e_coutput(name="foreign-target",
46                         tags={"upstream_project":"etl-test-shenzhen",
47                               "upstream_logstore":"slb-layer7-accesslog"}))
48)
49# after e_coutput, send to default
50e_output(tags={"upstream_project":"etl-test-shenzhen",
51              "upstream_logstore":"slb-layer7-accesslog", "type":"test_e_coutput"})
```

实践2：基于数据的动态分发

动态分发（一对多）

源Logstore
(App日志)

2018-1-1 INFO user1 login server from project=a logstore=b
2018-1-1 INFO user2 fail to login system type=ssh project=c logstore=d
2018-1-3 INFO user3 click page1 user agent project=e logstore=f
2018-1-3 INFO user3 add product id=123 into shopping cart project=g logstore=h
2018-1-3 INFO user3 click adds position x=1 y=2 width=100 project=a logstore=b

数据加工服务

过滤、富化、转换
补漏、分裂、聚合

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

目标Logstore1
(浏览信息)

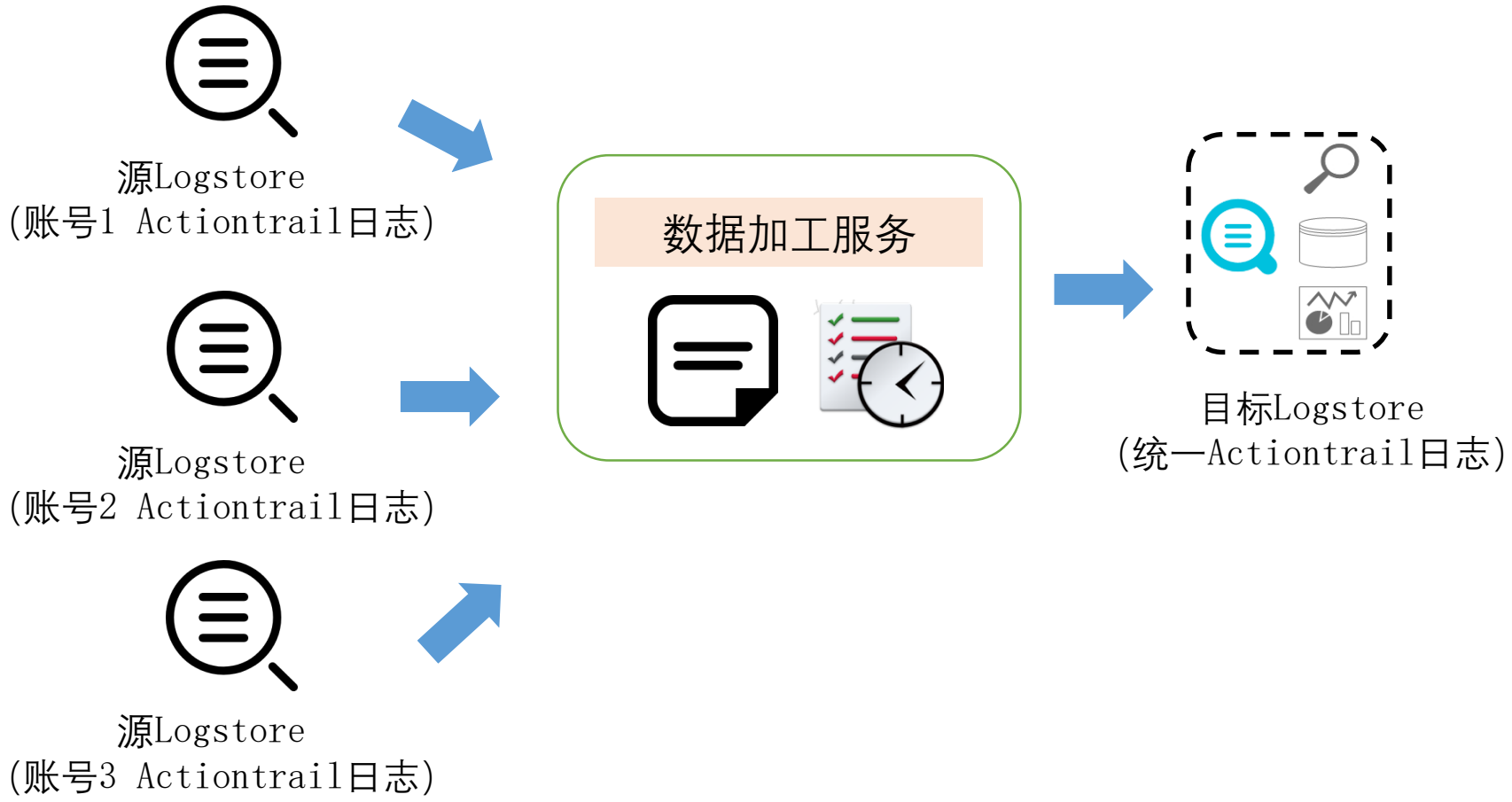
目标Logstore1
(下单信息)

DSL：根据日志内容动态计算目标位置

```
55e_keep(e_match("__topic__", "metric"))
56e_output(name="foreign-target",
57          #project="etl-test-shenzhen-devops",
58          logstore=str_format("metric_{}", v("Level")),
59          tags={"upstream_project":"etl-test-shenzhen",
60                "upstream_logstore":"slb-layer7-accesslog",
61                "type": "dynamic_dispatch"})
62
```

实践3：多源的数据汇集

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)

DSL：三个源logstore汇聚到一个目标

```
65#merge level:debug metric
66#no DSL code for only copy
67
68#merge level:warning metric
69#overwrite upstream tag
70e_set("__tag__:upstream_project", "etl-test-shenzhen-devops",
71      "__tag__:upstream_logstore", "metric_warning",
72      "__tag__:type", "merge_dispatch",
73      "__topic__", "")
74
75#merge level:info metric
76#overwrite upstream tag and split array
77e_set("__tag__:upstream_project", "etl-test-shenzhen-devops",
78      "__tag__:upstream_logstore", "metric_info",
79      "__tag__:type", "merge_dispatch",
80      "__topic__", "")
81e_if(e_has("Processes"),
82     e_split("Processes", output="Process"))
```

总结

分发实践总结

- 准备工作
  - 干活前前充分了解数据（基于前提的加工操作）
  - 合理规划源、目标logstore的shard数目，保证读取、加工、写入能力
- 特性
  - 支持跨账号、跨project
  - 注意e_output与e_coutput的区别：终止与非终止性操作
  - 使用分支：多目标分发一般会与流程控制语句一起使用
  - 善用e_compose减少重复判断，合并步骤
  - e_output是非必需的，当只有一个下游时并不需要
- 限制
  - 目前是Region化（跨Region未来会推出）
  - 支持最多20个静态目标
  - 无限个动态目标（考虑写出放大影响，需要合理规划shard数目）

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