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

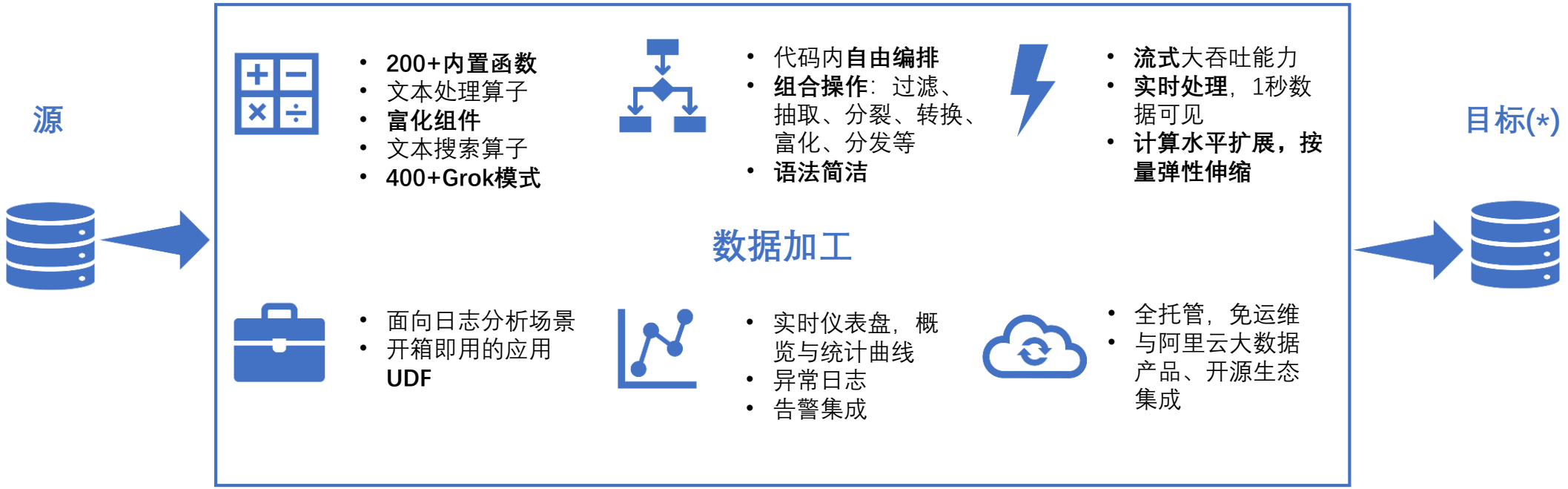
数据加工:数据映射富化实践

系列培训七

唐恺

日志服务-数据加工简介

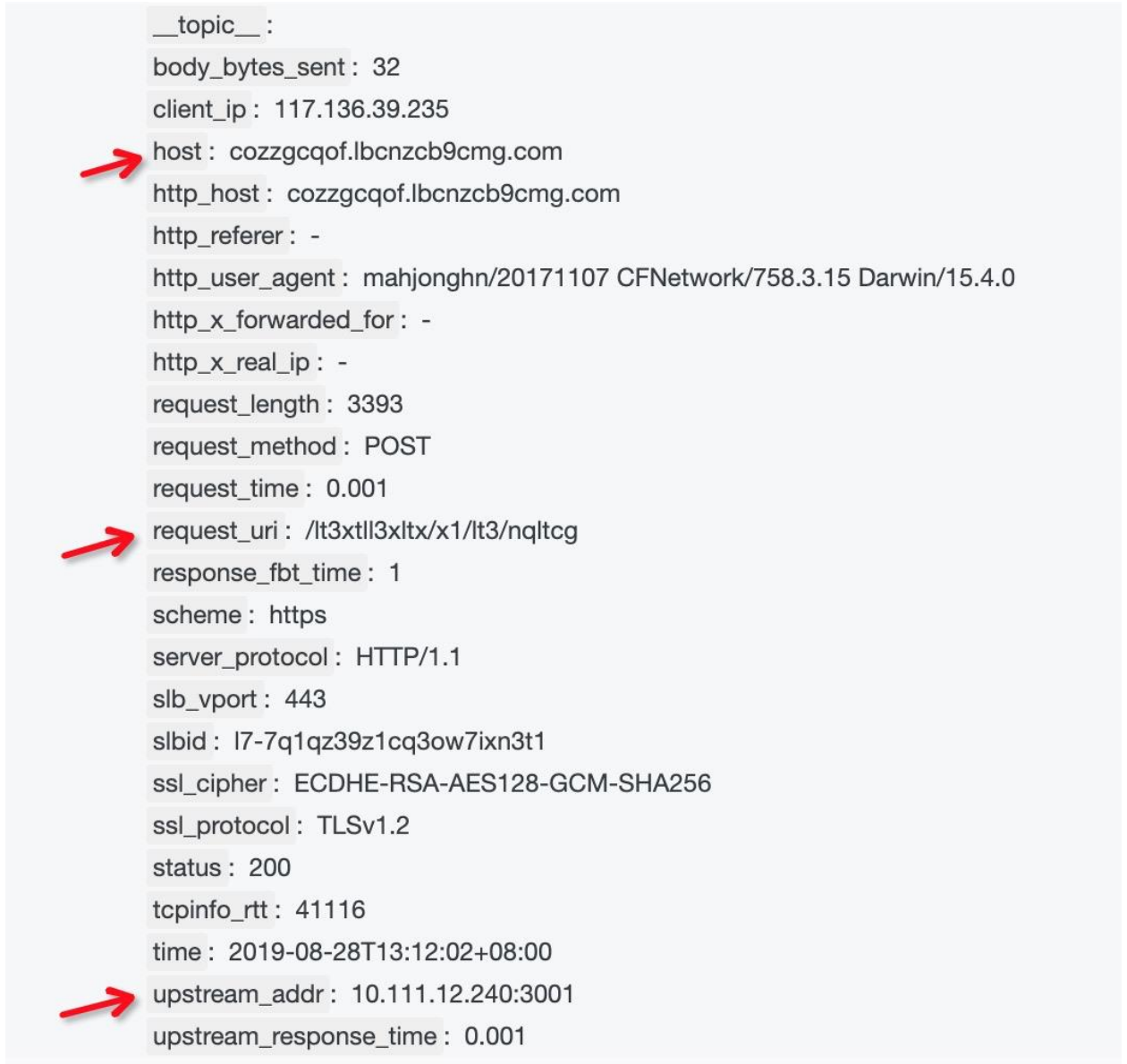
- 功能概述
  - 将各类日志处理为结构化数据，具备全托管、实时、高吞吐的特点
  - 面向日志分析领域，提供丰富算子、开箱即用的场景化UDF（Syslog、非标准json、AccessLog UA/URI/IP解析等）
  - 丰富的阿里云大数据产品（OSS、MC、EMR、ADB等）、开源生态（Flink、Spark等）集成能力，降低数据分析门槛
- 典型场景
  - 数据规整：对混乱格式的日志进行字段提取、格式转换，获取结构化数据以支持后续的流处理、数仓计算
  - 数据富化：日志（例如业务订单）与维表（例如用户信息MySQL表）进行字段join，为日志添加更多维度信息供分析
  - 数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



实践场景介绍

数据与需求

- 数据
  - SLB访问日志（实时、增量）
  - 两份维表（servers、pages）
- 运营需求
  - 统计top页面pv、uv
  - 页面标签分析
- 开发需求
  - 查看real server上的请求qps分布
  - 查看分别用于生产、测试的real server水位



维表(1)： MySQL servers表

重点字段

- detail.tag：标明机器用于生产、测试
- intranet_ip：VPC内子网ip

auto_id	id	intranet_ip	vpc	detail	ctime	uptime	status
1	i-2ze4h5q4jjgg9440t5g1	10.111.1.144	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	0
2	i-2ze4h5q4jjgg9440txgd	10.111.1.152	vpc-az018cntjnzvqm8xd9g7t	{"cpu":"32", "mem":"128G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	1
3	i-t4e4h5q4jjgg9440t5g1	10.111.1.142	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"dev"}	1566877884	1566964284	1
4	i-2ze4xge4jjgg9440t5g2	10.111.11.13	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	1
5	i-2ze4h6gegeg9440t5g1	10.111.12.144	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	1
6	i-2ze4h5q4jjgg9443513	10.111.1.11	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"32", "mem":"128G", "storage":"512GB", "network":"10Mbps","tag":"dev"}	1566877884	1566964284	1
7	i-2ze4h5q4j13251325g1	10.111.1.151	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"dev"}	1566877884	1566964284	1
8	i-3153513cdge9440t5g1	10.111.1.141	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"dev"}	1566877884	1566964284	0
9	i-2345334jjgg9440t5g1g	10.111.1.10	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"dev"}	1566877884	1566964284	1
10	i-2ze4h5q43513g9440t5	10.111.1.12	vpc-wx118cntjnzvqm8xd9atx	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"dev"}	1566877884	1566964284	1
11	i-2ze1534jjgg9440t5g1z	10.111.1.145	vpc-az018cntjnzvqm8xd9g7t	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	1
12	i-2ze4h5q4jjgg944015dt	10.111.12.24	vpc-az018cntjnzvqm8xd9g7t	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	1
13	i-1534h5q4jjgg9440t5g1	10.111.12.23	vpc-az018cntjnzvqm8xd9g7t	{"cpu":"32", "mem":"128G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	1
14	i-2ze4h5q4jjgg9440t153	10.111.143.132	vpc-az018cntjnzvqm8xd9g7t	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"product	1566877884	1566964284	0
15	i-2ze415q4jjgg9440t5gd	10.111.143.144	vpc-az018cntjnzvqm8xd9g7t	{"cpu":"64", "mem":"256G", "storage":"512GB", "network":"10Mbps","tag":"product"}	1566877884	1566964284	1

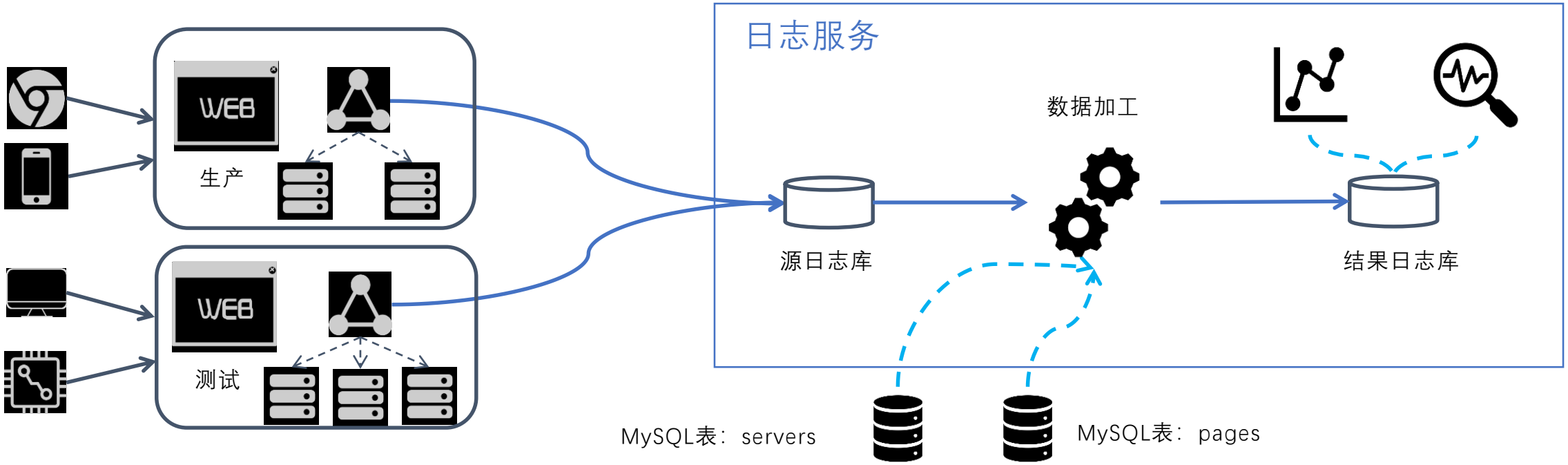
维表(2)： MySQL pages表

重点字段

- author：标注人
- page_tag：标签
- host、uri、author级别唯一，一个页面可能有多个维度的标签

auto_id	search	host	uri	author	page_tag	ctime	uptime	status
1	host=="n7q2.3gzzo7big.com" and request_uri=="/gnztf_lt33xf3"	n7q2.3gzzo7bi...	/gnztf_lt33xf3	LiLei	release	1566360389	1566878789	1
2	host=="qclbclb.3gzzo7big.com" and request_uri=="/cqx"	qclbclb.3gzzo7...	/cqx	LiLei	release	1566360389	1566878789	1
3	host=="n7q.3gzzo7big.com" and request_uri=="/gnztf_lt33xf3"	n7q.3gzzo7big....	/gnztf_lt33xf3	LiLei	release	1566360389	1566878789	1
4	host=="rrr.nmm56.com" and request_uri=="/goo-xcx3t-r3cxxb-cqq/xcx3t3r/onllxS3cxxb"	rrr.nmm56.com	/goo-xcx3t-r3cxxb-...	LiLei	release	1566360389	1566878789	1
5	host=="rrr.nmm56.com" and request_uri=="/cgx3xl-w37/nr3x/33lAgx3xlBgPtrxlxtf"	rrr.nmm56.com	/cgx3xl-w37/nr3x/3...	LiLei	release	1566360389	1566878789	1
6	host=="rrr.nmm56.com" and request_uri=="/xcx3t-ftlx9g-cqq/rn7rxxx73Gxtnq/33lSn7rxxx73"	rrr.nmm56.com	/xcx3t-ftlx9g-cqq/r...	LiLei	release	1566360389	1566878789	1
7	host=="rrr.nmm56.com" and request_uri=="/goo-xo-cqq/f3wloM3rrc33/lxcg3/x3x3xx3M3rrc33"	rrr.nmm56.com	/goo-xo-cqq/f3wlo...	LiLei	dev	1566360389	1566878789	1
8	host=="zo9.nmm56.com" and request_uri=="/lt3r/nqltcg"	zo9.nmm56.com	/lt3r/nqltcg	LiLei	dev	1566360389	1566878789	1
9	host=="rrr.nmm56.com" and request_uri=="/goo-nx-cqq/r3xxxx3/33lUr3xC3fl3xGxtnqIl3orWxlbPcxc"	rrr.nmm56.com	/goo-nx-cqq/r3xxx...	LiLei	dev	1566360389	1566878789	1
10	host=="qclbclb.3gzzo7big.cn" and request_uri=="/cqx"	qclbclb.3gzzo7...	/cqx	LiLei	dev	1566360389	1566878789	1
11	host=="rrr.nmm56.com" and request_uri=="/goo-xcx3t-r3cxxb-cqq/tqlxtfrCtf9x3/33lCcx3tCbcff3lEfl3xcfx3Ctf9x3"	rrr.nmm56.com	/goo-xcx3t-r3cxxb-...	LiLei	staging	1566360389	1566878789	1
12	host=="rrr.nmm56.com" and request_uri=="/goo-xcx3t-cqq/xcx3t3r/g3lcxl"	rrr.nmm56.com	/goo-xcx3t-cqq/xcx...	LiLei	staging	1566360389	1566878789	1

基于日志服务的解决方案



数据加工：映射富化函数

e_dict_map

功能：使用多个字段的值根据字典映射出一个新的字段.

语法:

e_dict_map(data, field, output_field, case_insensitive=True, missing=None, mode="overwrite")

参数名称	字段属性	是否必填	说明
data	Dict	是	映射字典, 标准的关键字-值的配对, 要求关键字必须是字符串; 进一步参考使用字典进行映射
field	String/ String List	是	一个字段名或者多个字段名的列表. 特殊字段名的设置, 可以参考事件类型; 多个字段时, 对每个字段依次做映射. 进一步参考下文多字段操作
output_field	String	是	输出字段的名称; 特殊字段名的设置, 可以参考事件类型
case_insensitive	Bool	否	匹配时是否是否大小写不敏感. 默认True(不敏感)
missing	String	否	不匹配时的目标字段的值, 进一步参考下文字典中的默认匹配与missing
mode	String	否	默认overwrite, 请参考覆盖模式

e_dict_map示例

```
1 e_dict_map({"400":"参数错误", "200": "正常", "500":"内部错误", "*": "其它错误"}, "status", "error_code")
2 e_keep_fields("status", "error_code")
```

预览任务开始时间： 2019-08-28 16:29:00

原始日志 | 数据加工 new

>	输出目标	时间	内容
1	target0	08-29 15:30:32	__source__ : __topic__ : error_code : 正常 status : 200

e_table_map

功能：根据输入字段的值, 在表格中查找对应的行, 返回对应字段的值.

语法:

e_table_map(data, field, output_fields, missing=None, mode="fill-auto")

参数名称	字段属性	是否必填	说明
data	表格	是	值映射表. 多列的表格. 进一步参考使用表格进行映射
field	String/ String List/ Tuple List	是	事件中映射到表格的源字段, 可以是单个字符串, 多个字符串列表或者其名称映射的元组列表（一个字段时: "user_id"； 多个字段时: ["province", "city", ...]； 一个字段与表格列名不一样时: ("userid", "user_id"); 如果事件中不存在对应字段, 则不进行任何操作.
output_fields	String/String List/Tuple List	是	表格中映射出的字段, 可以是字符串, 列表或者其名称映射元组的列表。
missing	任意	否	不匹配时的默认值, 如果目标字段是多列, 可以是一个默认值列表(长度必须与目标字段数一致)
mode	String	否	默认fill-auto， 请参考覆盖模式

e_table_map示例

```
e_table_map(tab_parse_csv("status,code,message\n400,参数错误,修改参数\n200,正常,无错误\n500,内部错误,请提工单"),
            "status",
            [("code", "error_code"), "message"],
            missing = ["其它错误", "请提工单"])
e_keep_fields("status", "error_code", "message")
```

预览任务开始时间： 2019-08-28 15:28:00

原始日志 | 数据加工 new

>	输出目标	时间	内容
61	target0	08-29 15:33:00	__source__ : __topic__ : error_code : 正常 message : 无错误 status : 200

e_search_dict_map

功能：关键字是查询字符串, 值是匹配的值的字典进行映射.

语法:

e_search_dict_map(data, output_field, multi_match=False, multi_join=" ", missing=None, mode="overwrite")

参数名称	字段属性	是否必填	说明
data	Dict	是	值映射表. 标准的关键字-值的配对, 要求关键字必须是字符串且是搜索字符串； 参考搜索字典
output_field	String	是	输出字段的名称； 特殊字段名的设置, 可以参考事件类型
multi_match	Bool	否	是否允许匹配多个, 允许时, 使用multi_join来拼接多个匹配的值. 默认否, 会返回匹配的第一个.
multi_join	String	否	匹配多个时, 多值的连接字符串, 默认空格
missing	String	否	不匹配时的目标字段的值. 默认None, 不做任何操作.
mode	String	否	默认overwrite， 请参考覆盖模式

e_search_dict_map示例

```
e_search_dict_map({'status ~= "2\d+" and request_method ~= "G.*"' : 'A', 'status == "500"' : 'B', 'host : ".*com"' : 'C'},
                  "tags",
                  multi_match = True,
                  multi_join = "; ",
                  missing = ["no tag"])
e_keep_fields("status", "request_method", "tags", "host")
```

预览任务开始时间： 2019-08-28 15:37:47

原始日志 | 数据加工 new

	输出目标	时间	内容
1	target0	08-29 15:37:58	__source__ : __topic__ : host : tgntoft-xcqc.obn9ton9c3on.com request_method : GET status : 200 tags : A; C

运行结果信息汇总

源日志总数	分发条数
58139	58139

e_search_table_map

功能：关键字是查询字符串, 值是匹配的值的字典进行映射.

语法:

e_search_table_map(data, inpt, output_fields, multi_match=False, multi_join=" ", missing=None, mode="fill-auto")

参数名称	字段属性	是否必填	说明
data	表格	是	搜索表格. 要求某一列必须是搜索字符串； 参考搜索表格
field	String	是	表格中的用于匹配搜索的字段名, 内容包含的都是查询字符串
output_fields	String/ String List/ Tuple List	是	表格中映射出的字段, 可以是字符串, 列表或者其名称映射元组的列表； 多个输出字段希望在表格列名上重命名时: [("population", "pop"), ("gdp", "GDP")])
multi_match	Bool	否	是否允许匹配多个, 允许时, 对每个输出字段的多个匹配的值, 使用multi_join来拼接, 默认否, 会使用匹配的第一个.
multi_join	String	否	匹配多个时, 多值的连接字符串, 默认空格
missing	String	否	不匹配时的目标字段的值. 默认None, 不做任何操作.
mode	String	否	默认fill-auto， 请参考覆盖模式

e_search_table_map示例

```
e_search_table_map(tab_parse_csv('search,tag,message\nstatus ~= "2\d+" and request_method ~= "G.*",A,aaa\nstatus == "500",B,bbb\nhost : ".*com",C,ccc'),
                   "search",
                   [("tag", "tags"), "message"],
                   multi_match = True,
                   multi_join = "; ",
                   missing = ["no tag", "no message"])
e_keep_fields("status", "request_method", "tags", "host", "message")
```

预览任务开始时间： 2019-08-28 15:38:03

原始日志 | 数据加工 new

修改预览配置

	输出目标	时间	内容
1	target0	08-29 15:39:03	__source__ : __topic__ : host : tgntoft-xcqc.obn9ton9c3on.com message : aaa; ccc request_method : GET status : 200 tags : A; C

数据加工：资源函数

资源函数概览

类别	函数	描述
本地	res_local	从当前数据加工的任务参数中获取信息
RDS数据库	res_rds_mysql	从RDS-MySQL中获取特定数据库表格的数据, 支持定期刷新
日志服务	res_log_logstore_pull	从另外一个logstore中拉取数据, 支持持续拉取与表格维护
OSS	res_oss_file	从OSS中获取特定Bucket下的文件内容, 并支持定期刷新

res_local

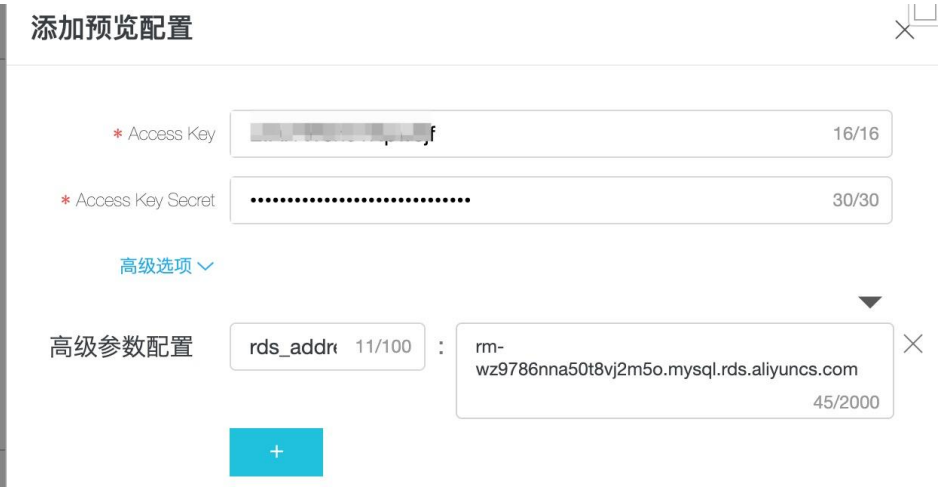
功能：从当前数据加工的高级参数中获取信息, 返回特定数据结构.

语法:

res_local(param, default=None, type="auto")

参数名称	字段属性	是否必填	说明
param	String	是	任务的参数名, 具体参考高级参数配置
default	任意	否	不存在时的默认值, 默认None
type	String	否	auto (默认): 首先转化为json格式, 当失败时, 直接返回其字符串形式; json: 将原始值转化为json格式, 参考json转换. 失败了, 会返回default值; raw: 原始形式(字符串);

res_local示例




```
1 e_set("address", res_local("rds_address"))
2 e_keep_fields("address")
```

预览任务开始时间： **2019-08-28 12:00:00**

原始日志 | 数据加工 new

	输出目标	时间	内容
1	target0	08-28 14:04:28	__source__ : __topic__ : address : rm-wz9786nna50t8vj2m5o.mysql.rds.aliyuncs.com

res_rds_mysql

功能：从MySQL中抓取表格或者SQL的结果返回数据表格. 支持按照一定间隔刷新全量.

语法:

res_rds_mysql(address, username, password, database, table=None, sql=None, fields=None, fetch_include_data=None, fetch_exclude_data=None, refresh_interval_max=60,refresh_interval=0)

参数名称	字段属性	是否必填	说明
address	String	是	域名或地址, 当端口号不是3306时,以地址:port来配置， 目前只支持公网地址
username	String	是	连接的用户名, 避免明文配置, 参考安全配置
password	String	是	账户密码, 避免明文配置, 参考安全配置
database	String	是	将要连接的数据库名
table	String	是	将拉取表格名, 如果传入了sql, 可以不传
sql	String	是	将拉取表格名, 如果传入了table, 可以不传
fields	字符串列表	否	数据列列表, 不填默认使用table或sql中返回的所有列
fetch_include_data	搜索字符串	否	搜索字符串,参考搜索字符串，黑白名单机制
fetch_exclude_data	搜索字符串	否	搜索字符串,参考搜索字符串，黑白名单机制
refresh_interval	数字字符串/数字	否	抓取间隔, 单位秒. 0表示只抓取一次, 默认只抓取一次.
refresh_interval_max	int	否	默认值为60s， 退火重试最大间隔。

基于数据加工的方案实施

富化servers表

```
e_keep(e_search('upstream_addr ~= "(.+)\:(\d+).*"'))
e_regex("upstream_addr", r"(?P<intranet_ip>[^:]+)\:.*")
e_table_map(res_rds_mysql("rm-w[illegible].mysql.rds.aliyuncs.com",
                          "tangkai", "[illegible]", "sls-test", table="servers",
                          fetch_include_data="status:1", refresh_interval_max = 60, refresh_interval = 1),
            "intranet_ip", [("id", "rs_id"), "detail"], missing = ["unknown", "{\"tag\":\"unknown\"}"])
e_json("detail", prefix= "rs_")
e_drop_fields("rs_cpu", "rs_mem", "rs_storage", "rs_network")
```

预览任务开始时间： **2019-08-28 15:47:29**

response_fbt_time : 1
rs_id : i-2ze4xge4jjgg9440t5g2
rs_tag : product
scheme : https
server_protocol : HTTP/1.1
slb_vport : 443
slbid : l7-7q1xx06z786tc3gowtb7x
ssl_cipher : ECDHE-RSA-AES128-GCM-SHA256
ssl_protocol : TLSv1.2
status : 200
tcpinfo_rtt : 55729
time : 2019-08-28T12:58:50+08:00
upstream_addr : 10.111.11.13:4040
upstream_response_time : 0.001

富化pages表

```
1  e_search_table_map(res_rds_mysql("rm-wz[illegible].mysql.rds.aliyuncs.com",
2                                   "tangkai", "[illegible]", "sls-test", table="pages",
3                                   fetch_include_data="status:1", refresh_interval_max = 60, refresh_interval = 1),
4                     "search", "page_tag", multi_match = True, multi_join = " ", missing = "null")
```

预览任务开始时间： 2019-08-28 15:49:39

输出目标	时间	内容
target0	08-28 12:59:19	__source__ : log_service __tag__:__receive_time__ : 1566979023 __topic__ : body_bytes_sent : 30 client_ip : 183.147.166.252 host : n7q.3gzzo7big.com http_host : n7q.3gzzo7big.com http_referer : - http_user_agent : okhttp/3.12.2 http_x_forwarded_for : - http_x_real_ip : - page_tag : release 广告 重要 request_length : 1214 request_method : POST request_time : 0.000

分析结果数据


数据加工-mysql lookuk （属于 etl-test-shenzhen ）
请选择
编辑
订阅
new
告警列表
刷新
分享
全屏
标题设置
重置时间
rs_tag qps 1小时（相对）
200
160
120
80
40
0
2019-08-28 18:00:40.000
2019-08-28 18:01:00.000
2019-08-28 18:01:20.000
2019-08-28 18:01:40.000
2019-08-28 18:02:00.000
2019-08-28 18:02:20.000
2019-08-28 18:02:40.000
2019-08-28 18:03:00.000
product
null
dev
page_tag uv 1小时（相对）
1.50%
9.62%
6.57%
42.41%
20.22%
1.58%
1.72%
6.45%
1.11%
8.81%
null
staging
release
dev
release 交易...
release 建议...
release 主页...
page_tag pv 1小时（相对）
release 广告 重要
release 主页 重要
release 交易 重要
release 建议 一般
release 文档
release
dev
staging
release 推荐 一般

总结

数据映射富化实践总结

维表选择

	本地配置	MySQL	OSS	LOG
DSL函数	res_local	res_rds_mysql	res_oss_file	res_log_logstore_pull
存储位置	数据加工，高级配置	阿里云RDS 自建MySQL	OSS	日志服务logstore
大小限制	单配置项：8KB 支持最多20个配置项	受限于数据加工内存 默认2GB，超出工单联系	受限于数据加工内存 默认2GB，超出工单联系	受限于数据加工内存 默认2GB，超出工单联系
数据类型	string	string二维表	text：csv文件 binary：ip词典等	string（key-value pair）
数据刷新	手动修改加工配置，重启作业	按时间触发刷新 覆盖，增量（计划中）	按时间触发刷新 全量覆盖	按时间触发刷新（流式订阅） 增量、覆盖
网络连通性	无要求	公网（开启ssl，ip白名单） 阿里云VPC（ETA：2019/9）	阿里云经典网络、VPC 公网（建议https endpoint）	阿里云经典网络、VPC 公网（建议https endpoint）

映射函数选择

- e_table_map/e_dict_map：简洁，最常用
- e_search_dict_map/e_search_table_map：支持multi_join，支持复杂join条件

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