FLINK FORWARD # ASIA 2022 ONLINE | 实时即未来

基于Flink ML搭建的智能运维算法服务及应用

张颖莹 | 阿里云算法专家

01 阿里云大数据平台的智能运维

02 智能运维算法服务应用场景

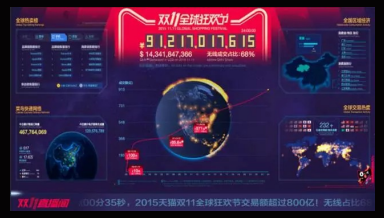
03 传统算法工程链路的局限性

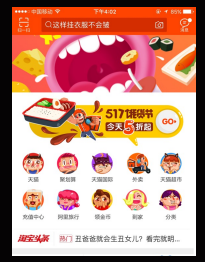
04 使用Flink ML搭建智能运维算法服务

01 阿里云大数据平台的智能运维

阿里云大数据平台

典型业务场景





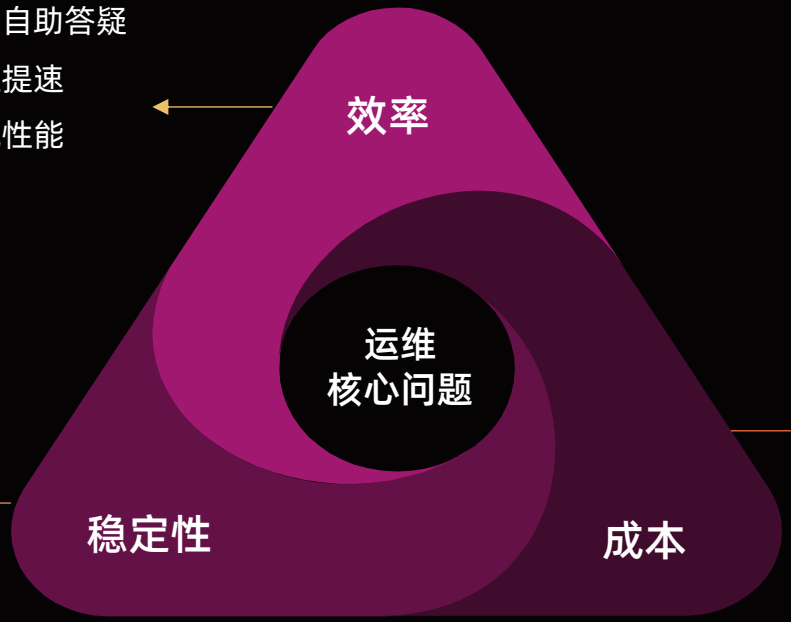
大数据平台

大数据计算服务 MaxCompute	实时计算 Flink版	实时数仓 Hologres	……
快速、完全托管的 TB/PB级数据仓库	企业级、高性能 实时大数据处理系统	交互式分析产品 一站式实时数据仓库引擎	

阿里云ABM 运维中台

大数据平台智能运维

- 用户自助答疑
- 作业提速
- 系统性能



- 异常发现
- 告警收敛
- 根因定位
- 自愈

- 资源配置优化
- 应用排布

算法需求

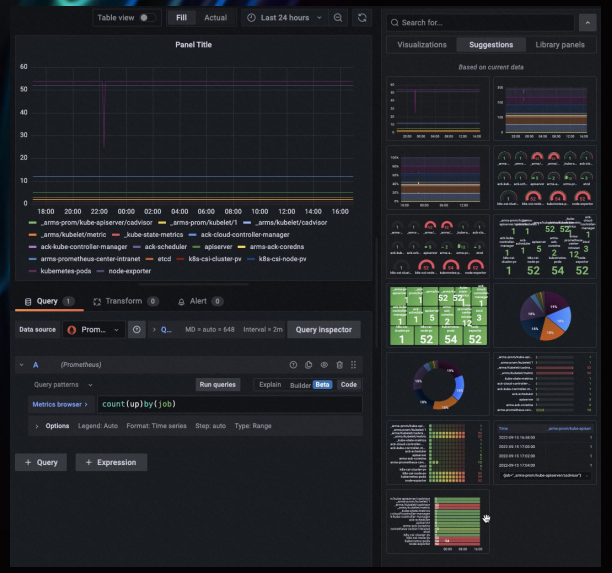
- 多源异构数据
- 海量数据
- 实时分析

可观测性支柱

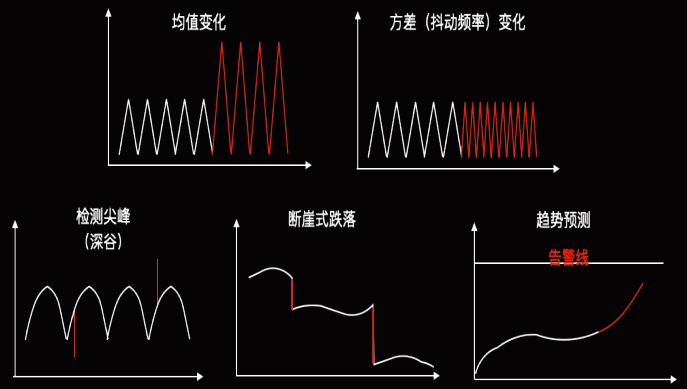
- 指标
- 日志
- Trace

02 智能运维算法服务应用场景

稳定性：集群核心指标监控



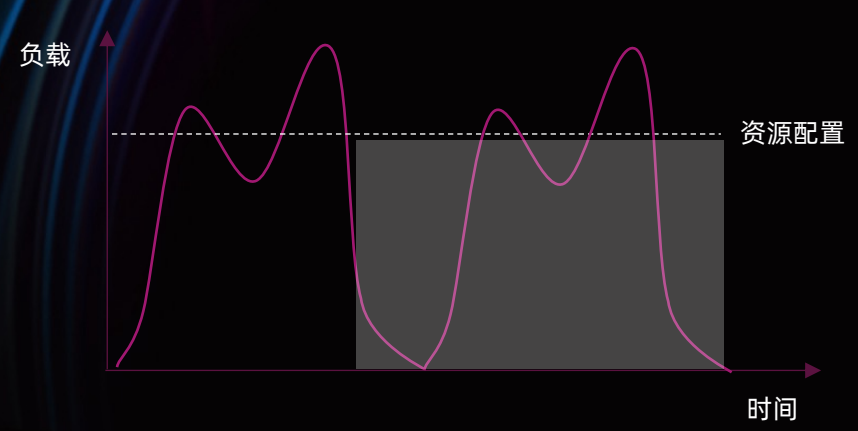
核心指标监控大盘



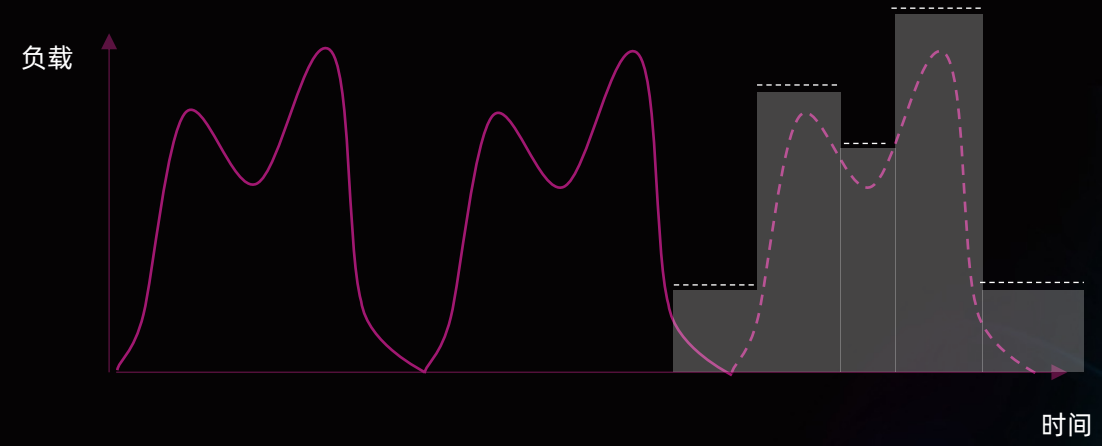
时序异常检测

缩短MTTD
保障SLA（Service Level Agreement）

成本:基于负载的资源自动扩缩容



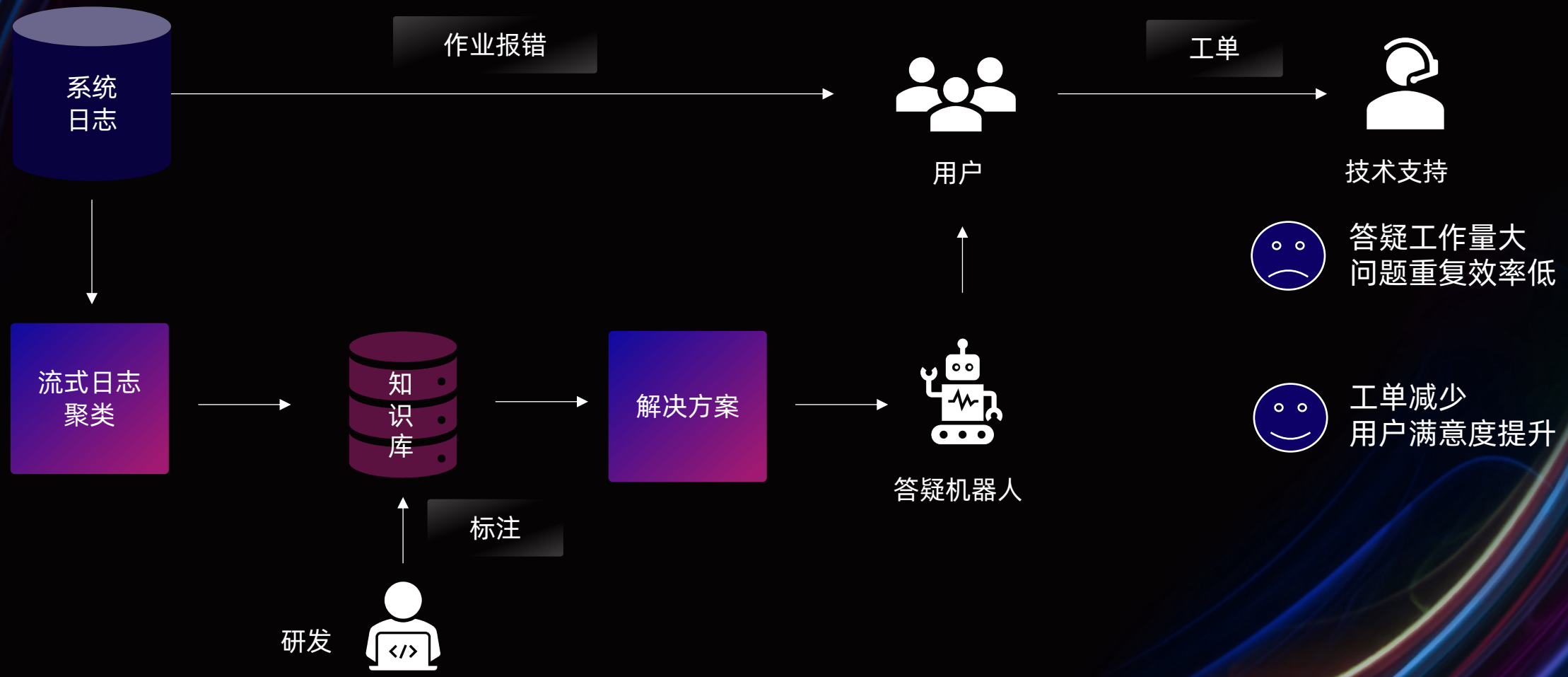
精准预测
自动扩缩容



低峰时段资源浪费
高峰时段需求无法满足

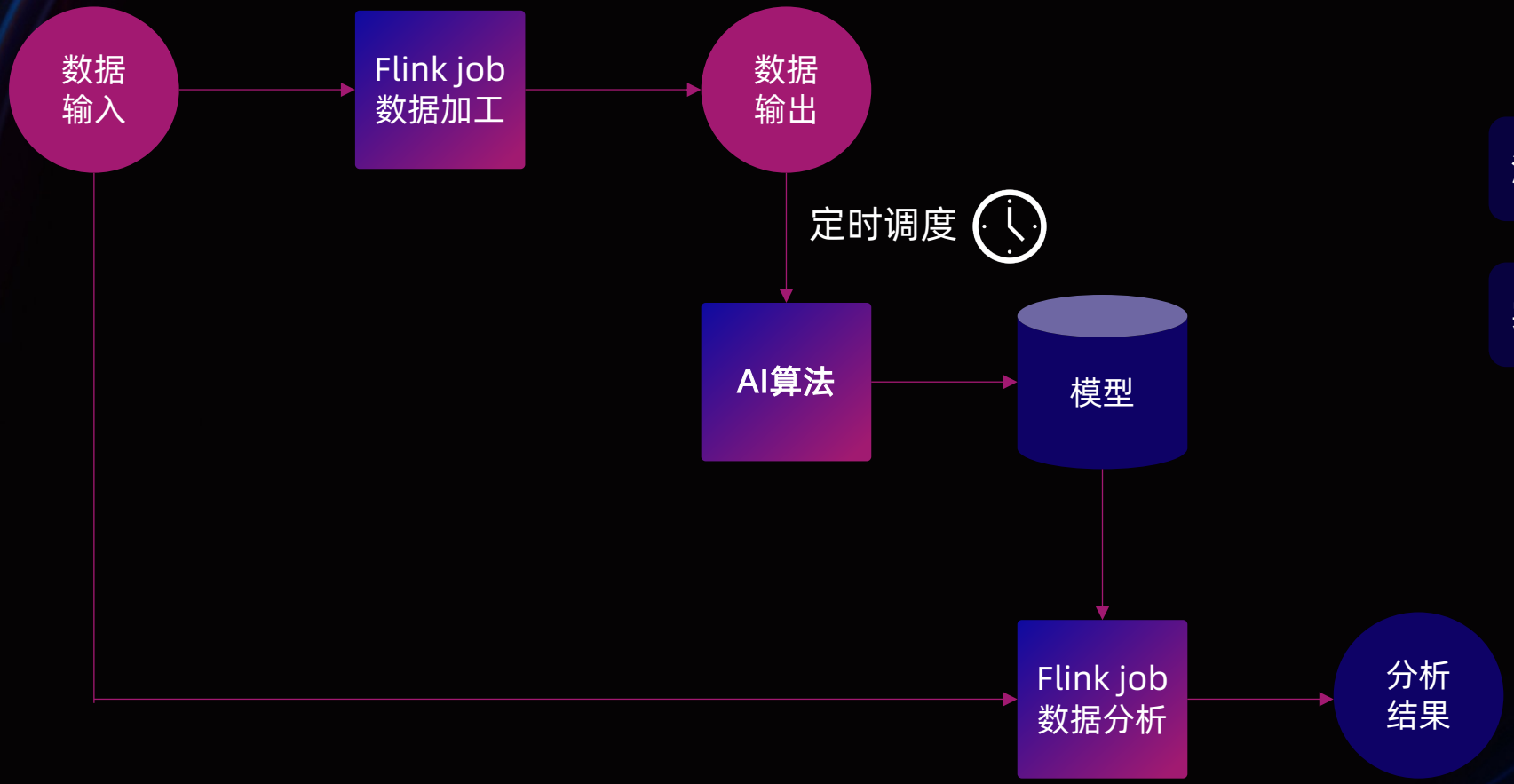
资源节省
需求得到充分保障

效率:基于日志的作业智能诊断



03 传统算法工程链路的局限性

传统算法工程链路



局限性

- 流程冗长
- 运维成本高
- 实时性低
- 性能难保证

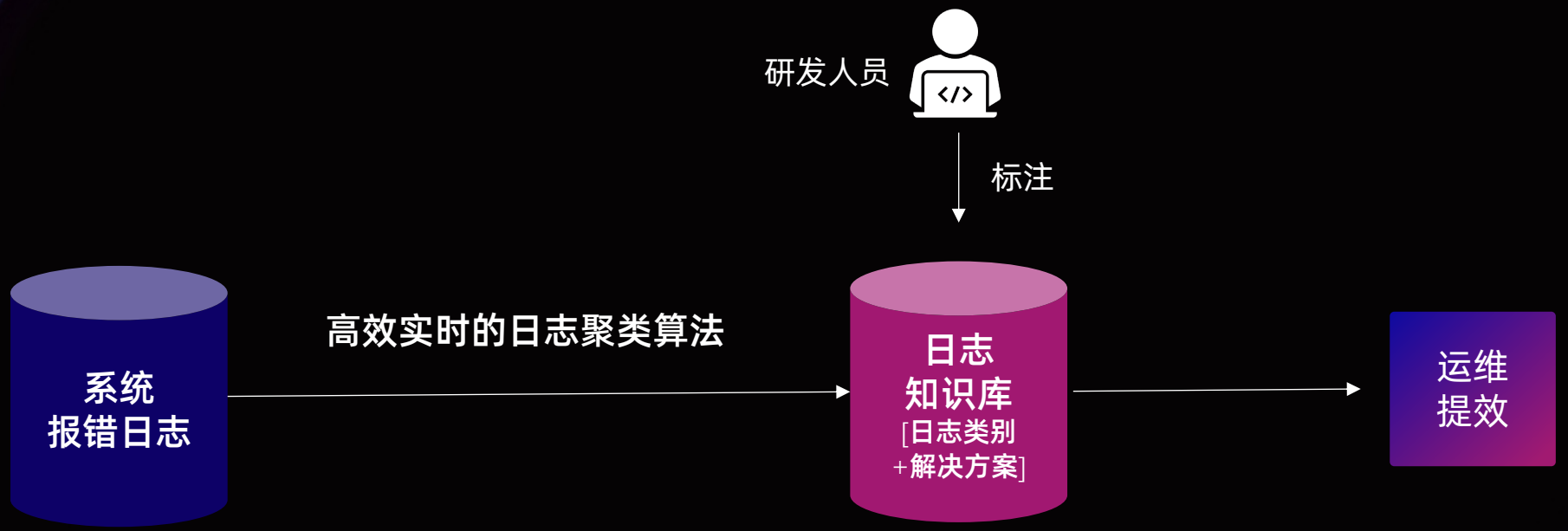
04 使用Flink ML搭建智能运维算法服务

面向运维场景的算法设计

Flink ML特性

- 实时性
- 流批一体
- CDC 增量读取
- Flink ML API
- Flink ML 基础设施

日志聚类业务背景和挑战



特点和挑战

- 海量且信息密度低
- 非结构化数据
- 包含的变量多
- 一定的格式规范+语义性
- 实时生成

聚类目标

- 自动实时地聚合为有限的日志类别
- 考虑日志的语义性
- 能应对新的日志类型的出现

日志聚类示例

日志原文

① Table meta.shop can not be found

② Table meta.merchant can not be found

③ Table meta.customer can not be found

④ Can't find table meta.cluster

⑤ Can't find table meta.project

⑥ Can't find table meta.machine

预处理和编码
（日志模板）

① Table * can not be found '895ca77eb163456b4bb4cf17fce8a81b'

② Can't find table * '1526796e4dc4afd42d93b8d759b05457'

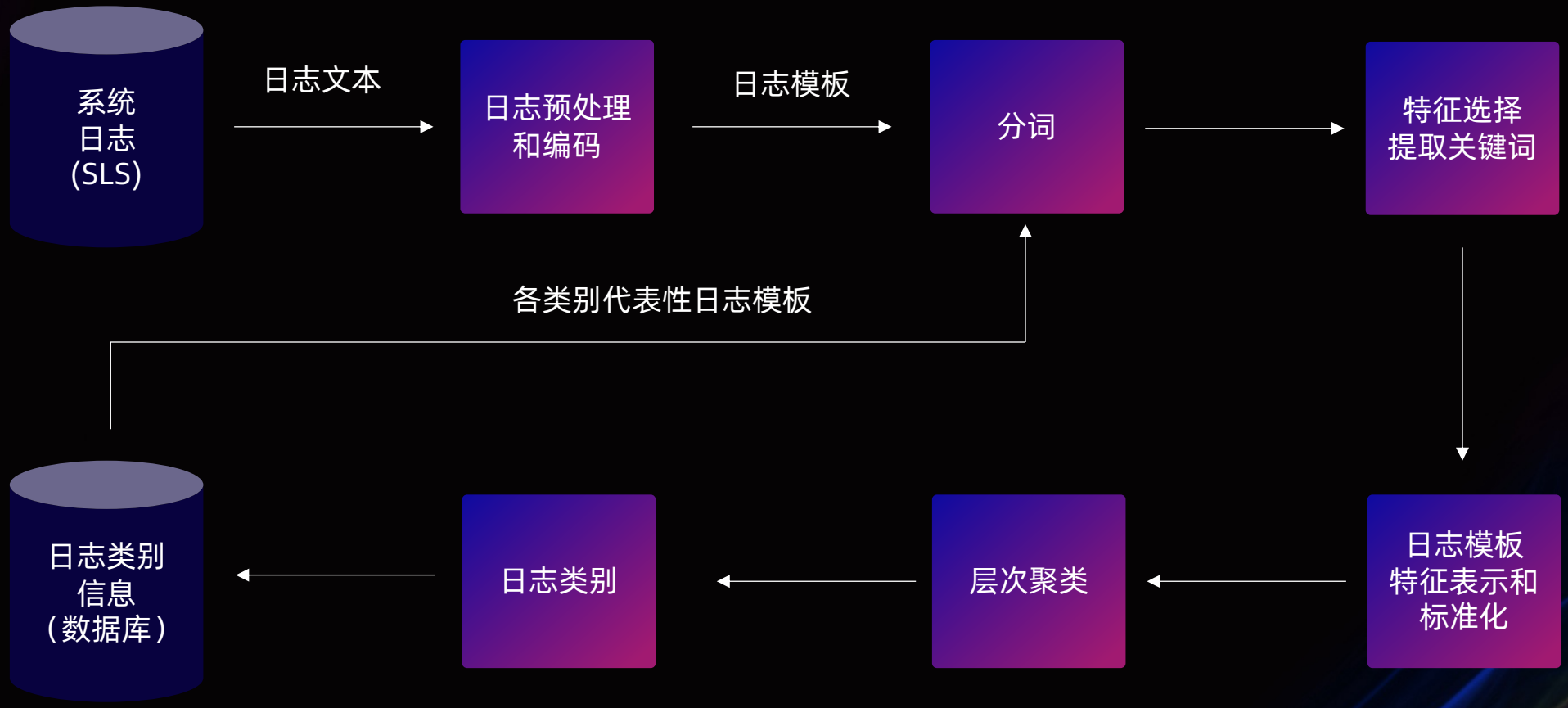
分词和特征表示

日志模板	table	can	not	be	find
'895ca77eb163456b4bb4cf17fce8a81b'	1	1	1	1	1
'1526796e4dc4afd42d93b8d759b05457'	1	1	1	0	1

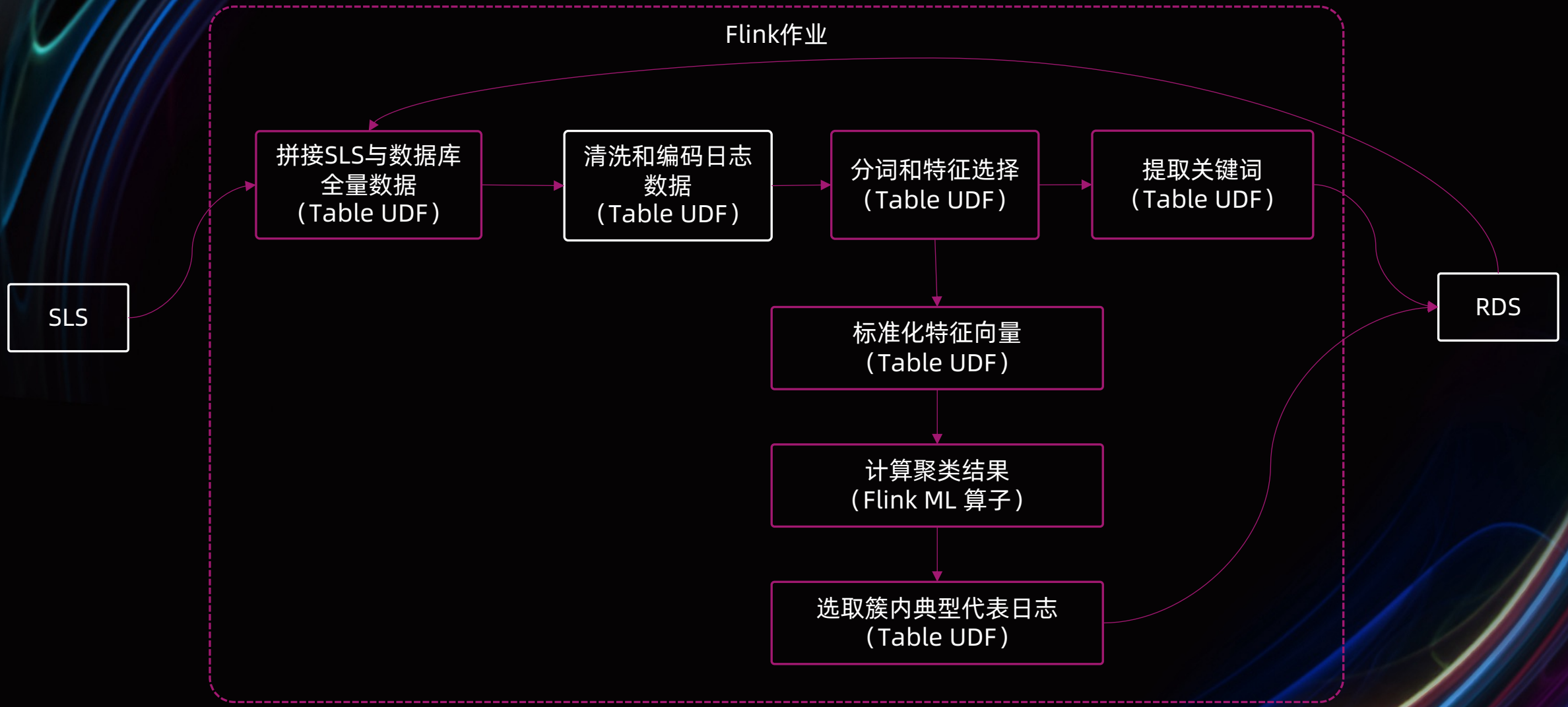
聚类和标注
（日志类别和解决方案）

日志模板	日志类别	关键词	解决方案
'895ca77eb163456b4bb4cf17fce8a81b'	类别A	table，can,not,find	请检查相关的表是否存在
'1526796e4dc4afd42d93b8d759b05457'			

日志聚类算法流程


系统日志(SLS)
日志文本
日志预处理和编码
日志模板
分词
特征选择提取关键词
各类别代表性日志模板
日志类别信息（数据库）
日志类别
层次聚类
日志模板特征表示和标准化

使用Flink ML构建流式日志聚类



算子的可复用性

```
# 数据读入
■ 数据库FlinkCDC增量读取
■ SLS流式读取
inputTable = ...

# 预处理相关UDF

# 特征工程 UDF
■ 分词、日志文本向量化（CountVectorizer）
■ 特征选择（TF-IDF）
■ 特征标准化（ StandardScaler）
```

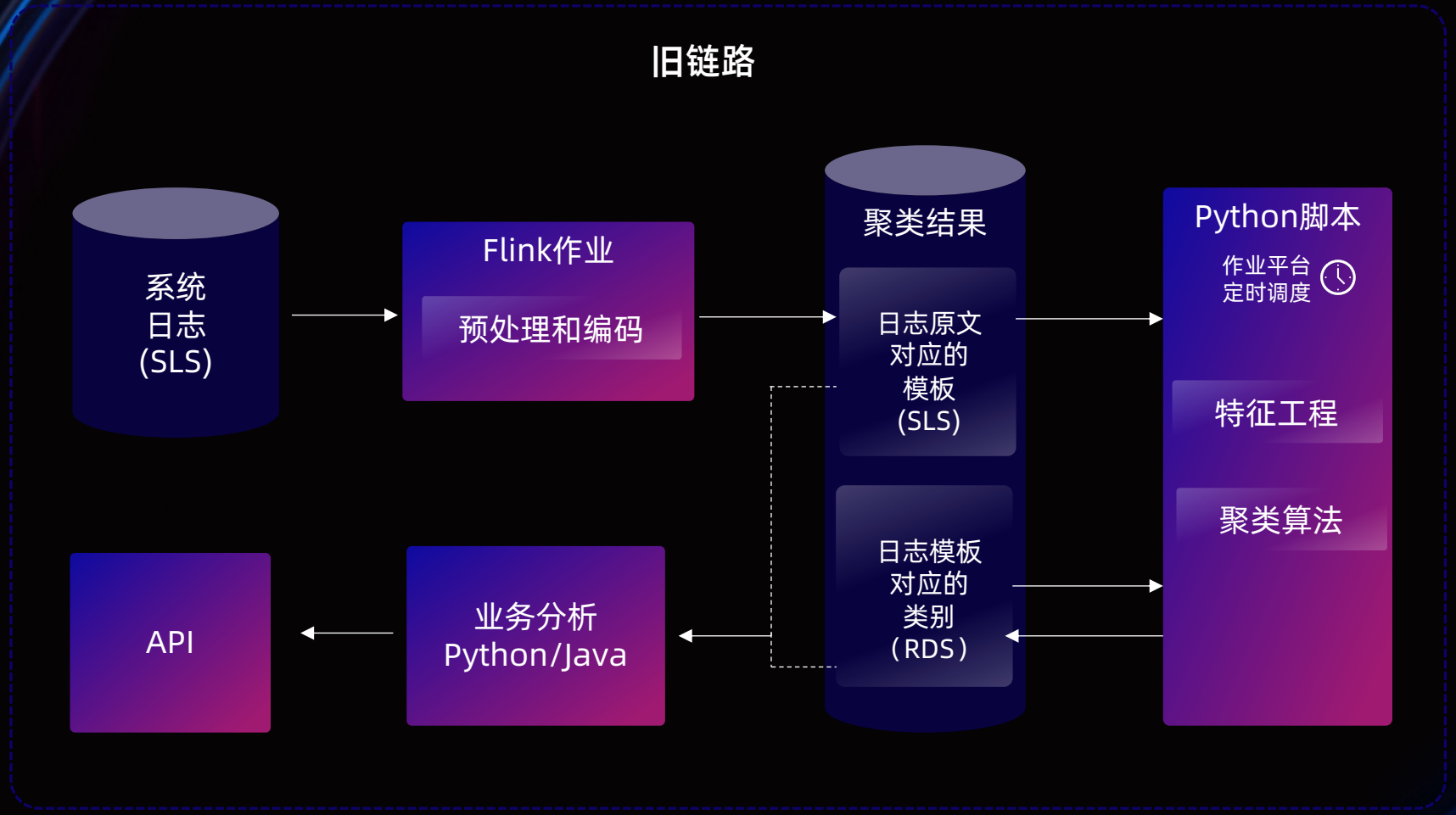
```
# 创建Flink ML层次聚类算子实例，并配置所需参数。
agglomerative_clustering =
AgglomerativeClustering()\
 .set_features_col('features')\
 .set_prediction_col('prediction')\
 .set_linkage('average')\
 .set_distance_measure('euclidean')\
 .set_windows(EventTimeTumblingWindows.of(Tim
e.minutes(1))) \
 .set_compute_full_tree(True)

# 输出每个向量的类别
outputTable =
agglomerative_clustering.transform(inputTable)[0]
```

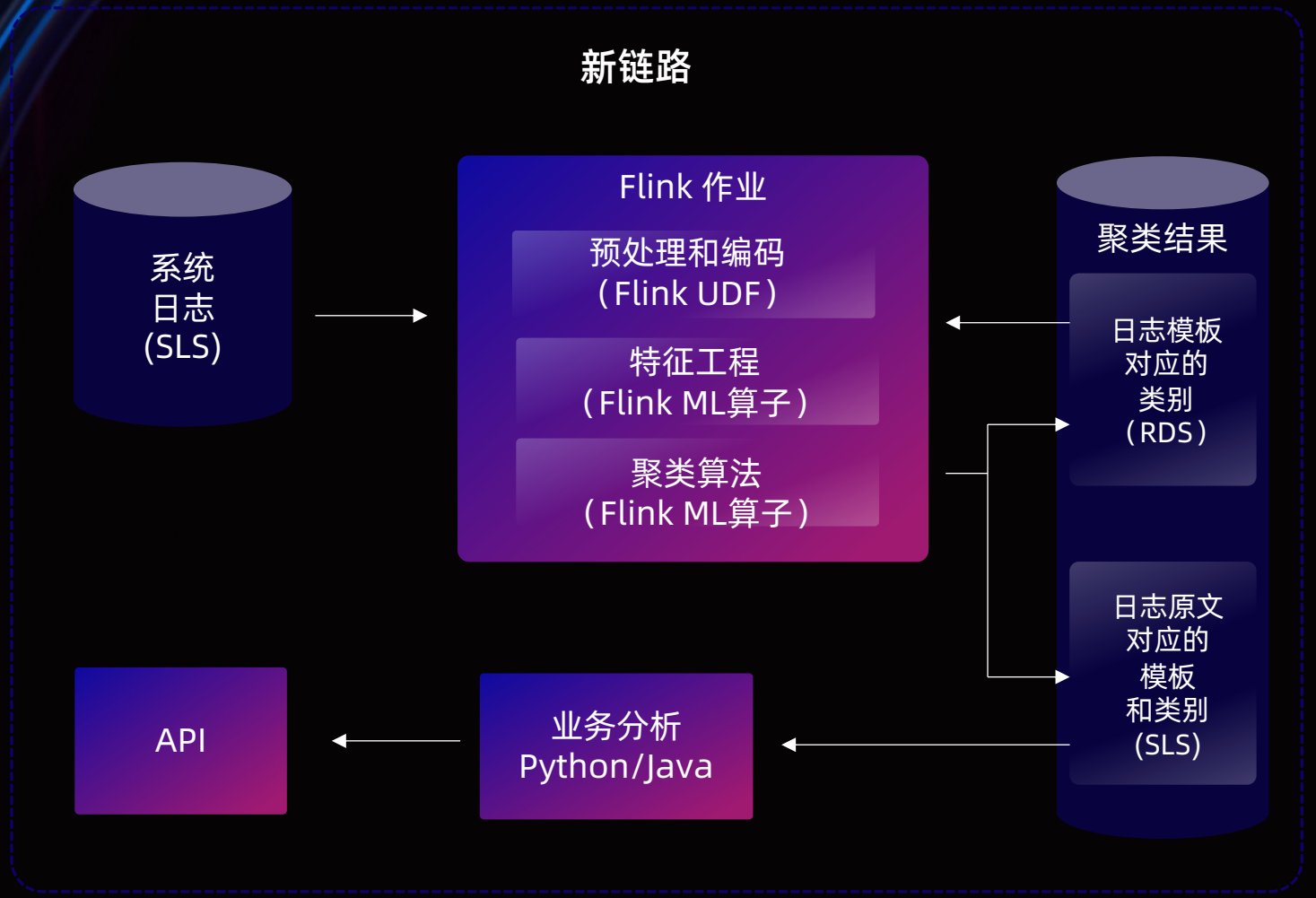
无限数据流上的实时聚类

- 多种窗口可供选择
  - GlobalWindows
  - CountTumblingWindows
  - EventTimeTumblingWindows
  - ProcessingTimeTumblingWindows
  - EventTimeSessionWindows
  - ProcessingTimeSessionWindows
- 根据实时性需求，灵活调节窗口大小

日志聚类算法链路升级的收益



日志聚类算法链路升级的收益

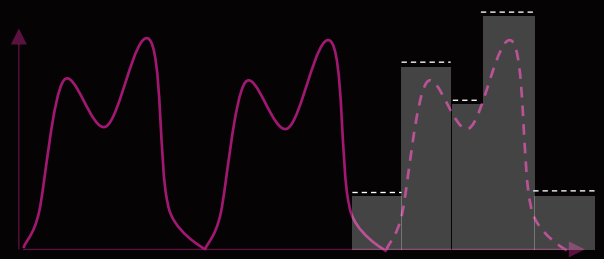


收益

- 链路延迟降低
  (5min ——> 30s)
- 运维成本降低
  (只需维护一个Flink作业)
- 分析成本降低
  (减少了RDS和SLS的联合分析)
- 算法性能提升
  (单机——>Flink 算子)

总结和开源计划

- 时序异常检测
- 时序预测
- 日志聚类



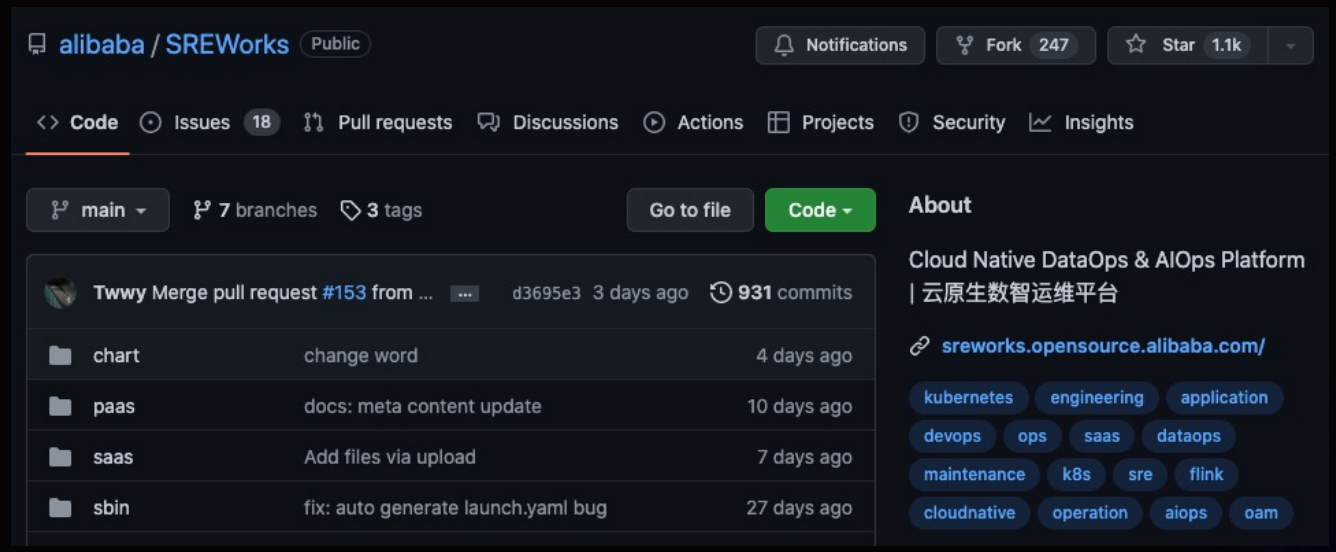
- 实时性
- 流批一体
- CDC 增量读取
- Flink ML API
- Flink ML 基础设施

收益

- 链路延迟降低
- 运维成本降低
- 分析成本降低
- 算法性能提升

展望：基于Flink ML的算法服务开源计划

SREWorks



https://github.com/alibaba/SREWorks

部分基于Flink ML构建的算法服务及框架在未来将通过SREWorks输出，欢迎关注！

FLINK FORWARD # ASIA 2022 ONLINE | 实时即未来

THANK YOU

谢 谢 观 看