FLINK FORWARD #ASIA 2020 | 实时即未来

Flink 在实时标签系统中的实践

杨涵冰

上海数禾信息科技有限公司 大数据部

SHUHE GROUP 数禾科技

#1

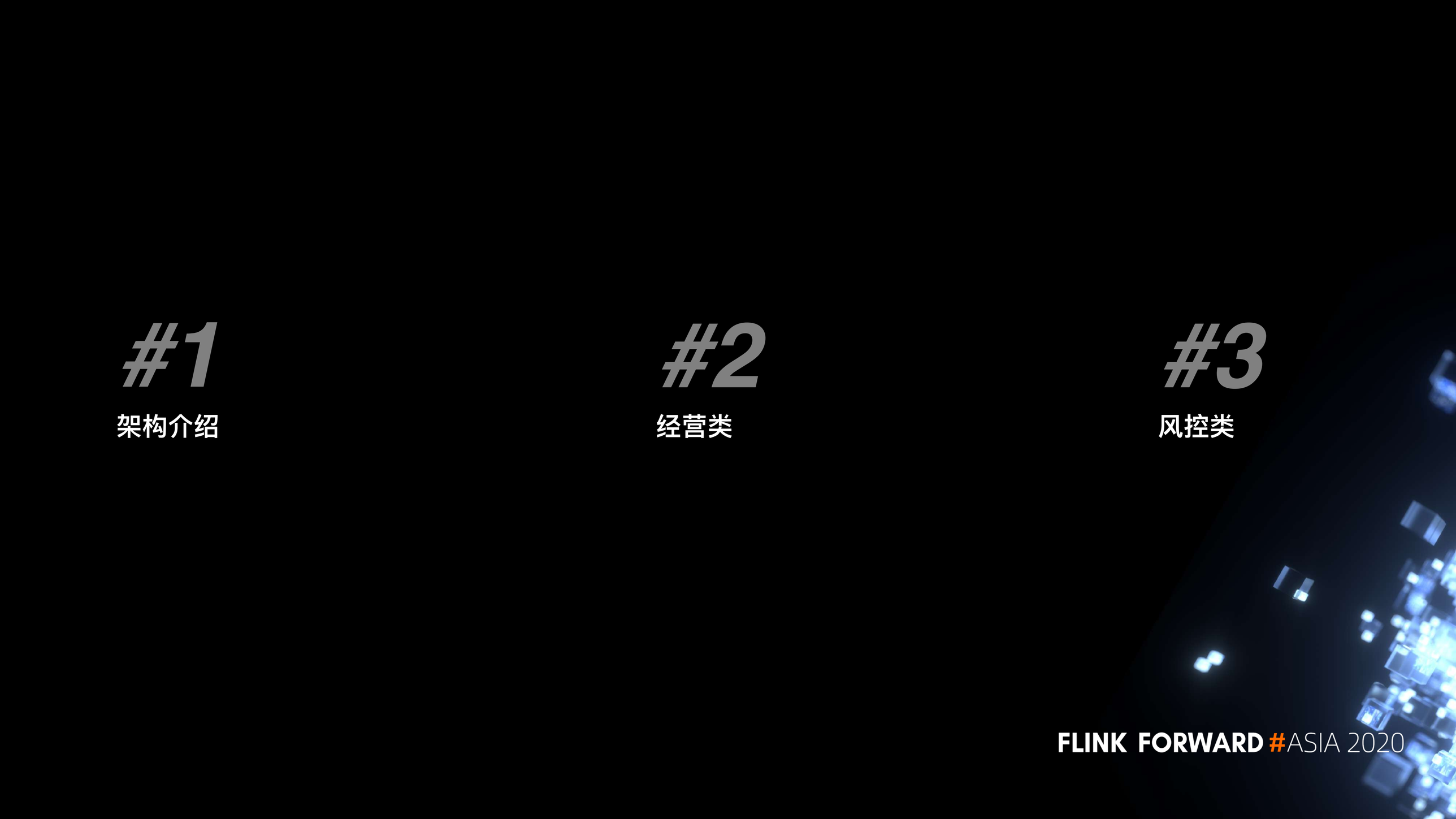
架构介绍

#2

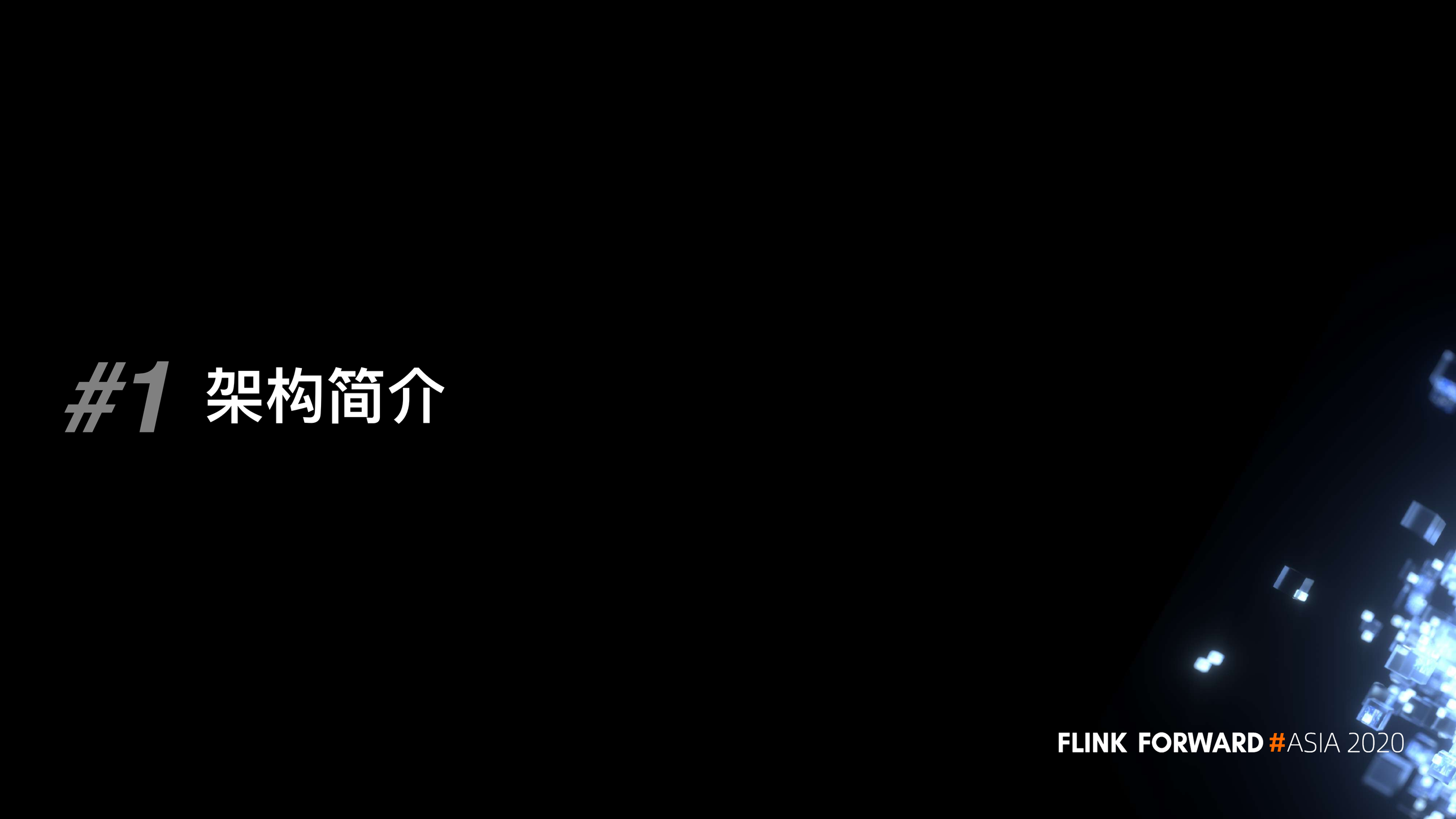
经营类

#3

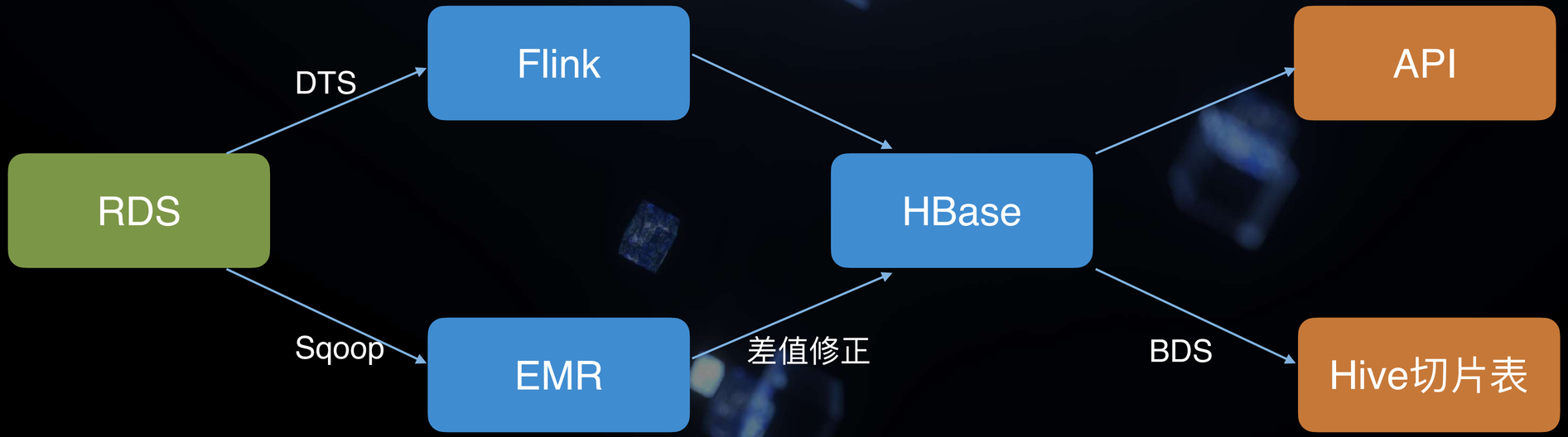
风控类



#1 架构简介



架构简图



四大标签类型

原生

同步线上数据

实时写入，离线修正

即时计算

API调用时运算

线下批量计算，逻辑一致

实时

传统实时链路

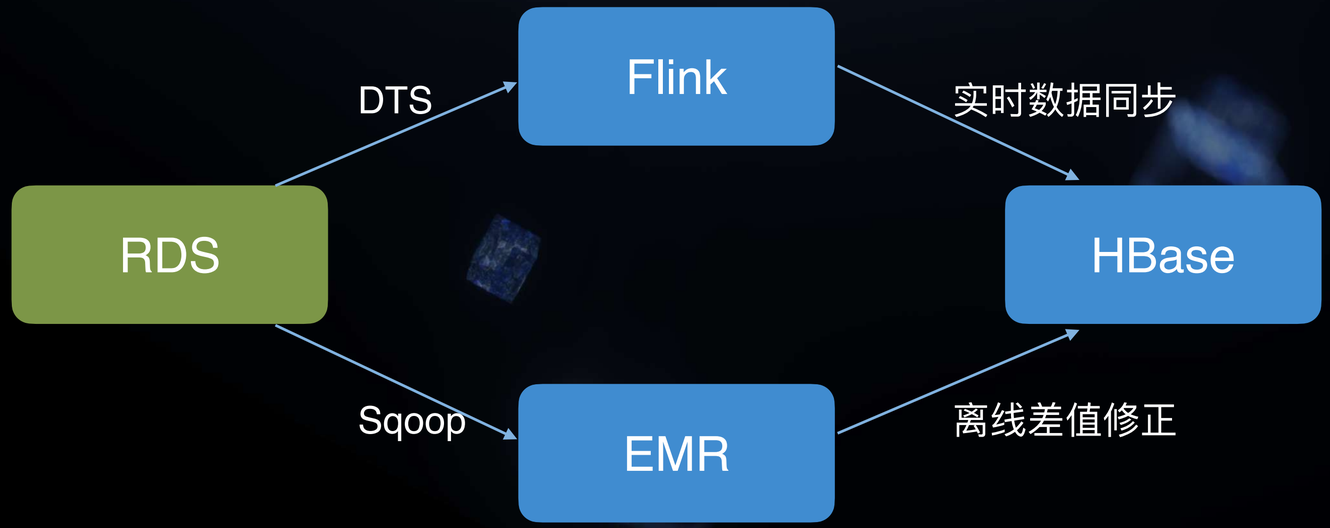
实现复杂逻辑

离线

传统离线链路

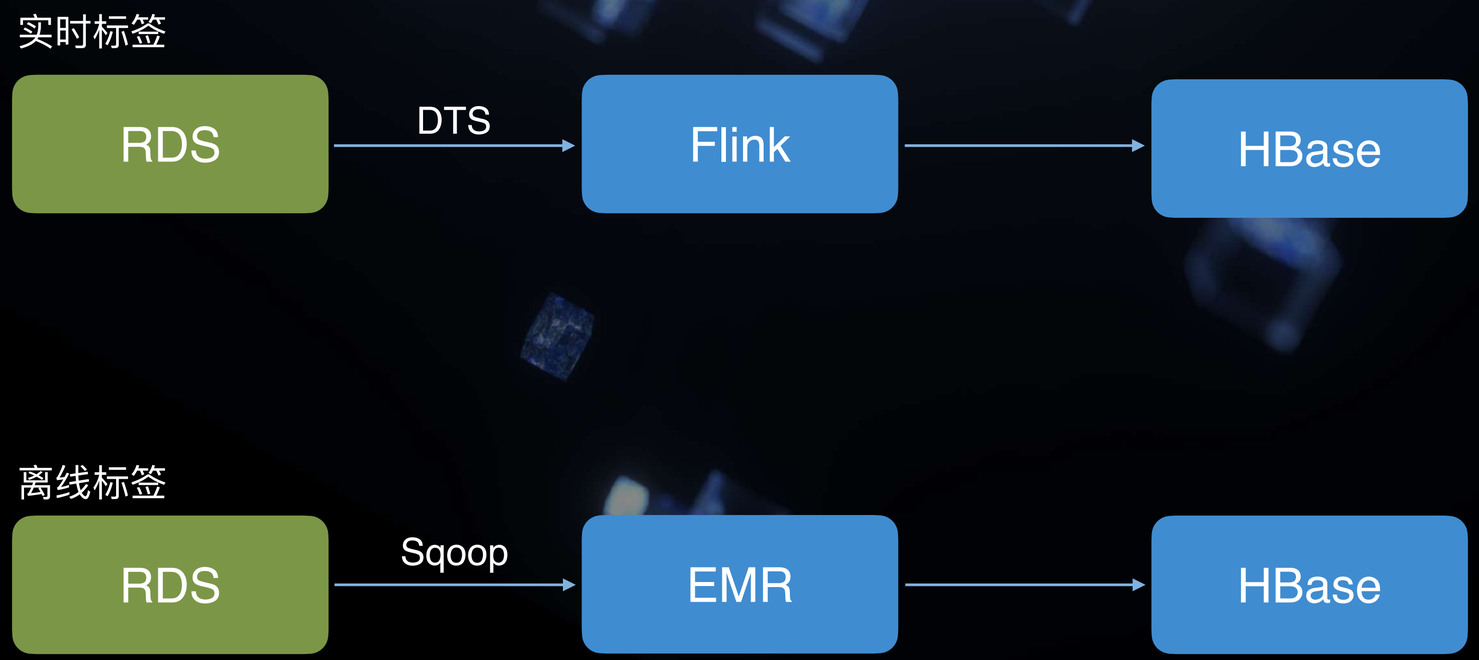
实现复杂逻辑

原生标签



即时计算标签

实时标签与离线标签



#2 经营类

经营类

#1

逻辑相对简单

#2

多数为单一维度的

简单加工

#3

迭代速度较快

大量探索类需求

经营类案例

用户经营案例

需求在决策流引擎中获取用户的各类信息，根据用户信息进行相应营销。

决策流所需信息中，大部分为分散在各个业务系统中的用户信息。比如分散在各系统的手机号：注册手机号、审核手机号、绑定手机号等。通过配置原生标签，可以简单的将各业务系统的用户信息同步到标签系统中。

还有一部分为需要进行简单运算的数据。比如，资方授信额度维护在各个资方对接系统中，此时将他们分别配置为原生标签，然后配置一个即时计算标签将这些数据进行即时合并运算，便可得到最大授信额度等衍生标签。

#3 风控类

风控类

#1

逻辑非常复杂

#2

长达数月甚至全量数据的聚合处理、排重聚合处理

#3

迭代比较稳定

需求一般经过验证

常见问题

排重统计

根据某个维度对一段时间事件进行排重统计运算。

1、单维度数据量较少。将明细数据存储在 HBase 中，直接进行统计。

2、单维度数据量虽然多，但单日新增数据量较少。每日运算离线统计值和明细值，实时存储当日明细，进行差值统计。

3、大数据量计数排重。需要使用有损统计。

图关系统计

根据数据的图关系进行计算。

1、一阶图关系可以将边数据存储在 HBase 中，直接进行统计。

2、二阶、三阶等低阶图关系通过多次 HBase 查询统计。需要注意的是随着阶数升高，查询量级会迅速增高。

更新时效

由于整条实时流链路较长，可能会有时效性发生波动的情况。如果下游系统需要根据标签时效性精确控制行为，需要通过一些额外属性来解决。

1、标签更新时间。判断当前标签值的更新时间。

2、标签整体更新时间。判断该标签整体更新情况。

风控类案例

三个月内登陆不同IP数

1、每日离线计算 T-1 数据，将统计值和明细值写入 HBase

2、实时流作业将实时明细值写入 HBase

3、标签实时流作业读取统计值信息，获取离线统计值及该值的更新时间。读取实时明细值，并与离线明细表进行差值统计。合并统计值得到最终结果。

被填为第一联系人数量

1、图数据实时流作业将关联数据更新至 HBase

2、每日图数据离线作业修正该实时数据。

3、标签实时流作业以当前用户去 HBase 中查询明细数据进行统计。

最近还款金额

1、即时计算标签获取标签值及其更新时间，若在 10 秒内，可以认为标签值为最新数据，直接返回。

2、若更新时间在 10 秒以上，获取标签整体更新时间，若在 10 秒内，则认为 10 秒前数据已更新，用户确实在 10 秒内无还款，直接返回标签值。否则说明标签发生时效性波动，返回标志值，通知调用方重试。


FLINK FORWARD
#ASIA 2020
实时即未来
谢谢