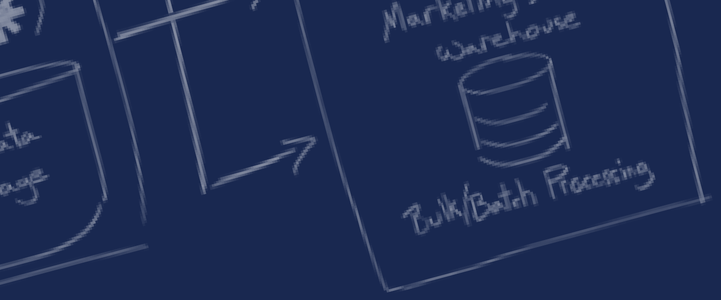
amazon web services SUMMIT SHENZHEN

企业转型分论坛

TRANSSNET

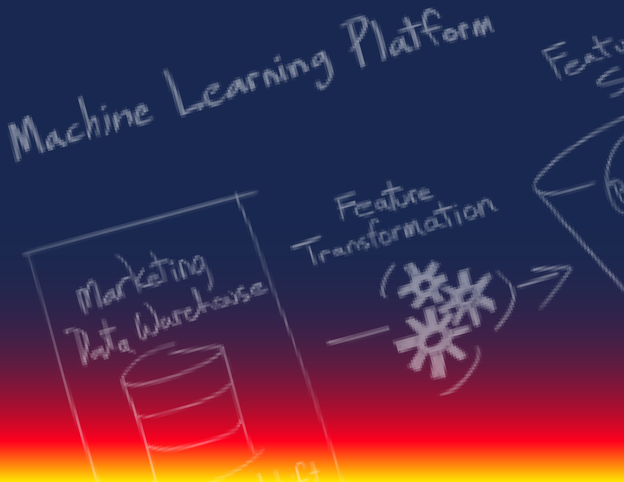
传易金融数据平台介绍

陈中杰
深圳传音控股有限公司 技术总监



amazon web services

SUMMIT
SHENZHEN



公司介绍

传易集团是由中国领先的互联网上市公司网易集团和全球智能终端产品移动增值服务提供商传音控股集团共同创办的合资公司，聚焦移动互联网领域，以互联网金融、音乐、短视频、应用市场、在线游戏等为主要业务方向。

旗下网聚了非洲第一大在线音乐平台Boomplay， 非洲排名第一的短视频平台 Vskit，非洲第二大移动应用分发平台 Palmstore，以及互联网金融信贷理财和支付产品Palmcredit、Palmsave、Palmpay等已在非洲崭露头角的优秀互联网产品，是中国互联网最先开始非洲本土化的企业之一。

TRANSSNET

Boomplay

Palmstore

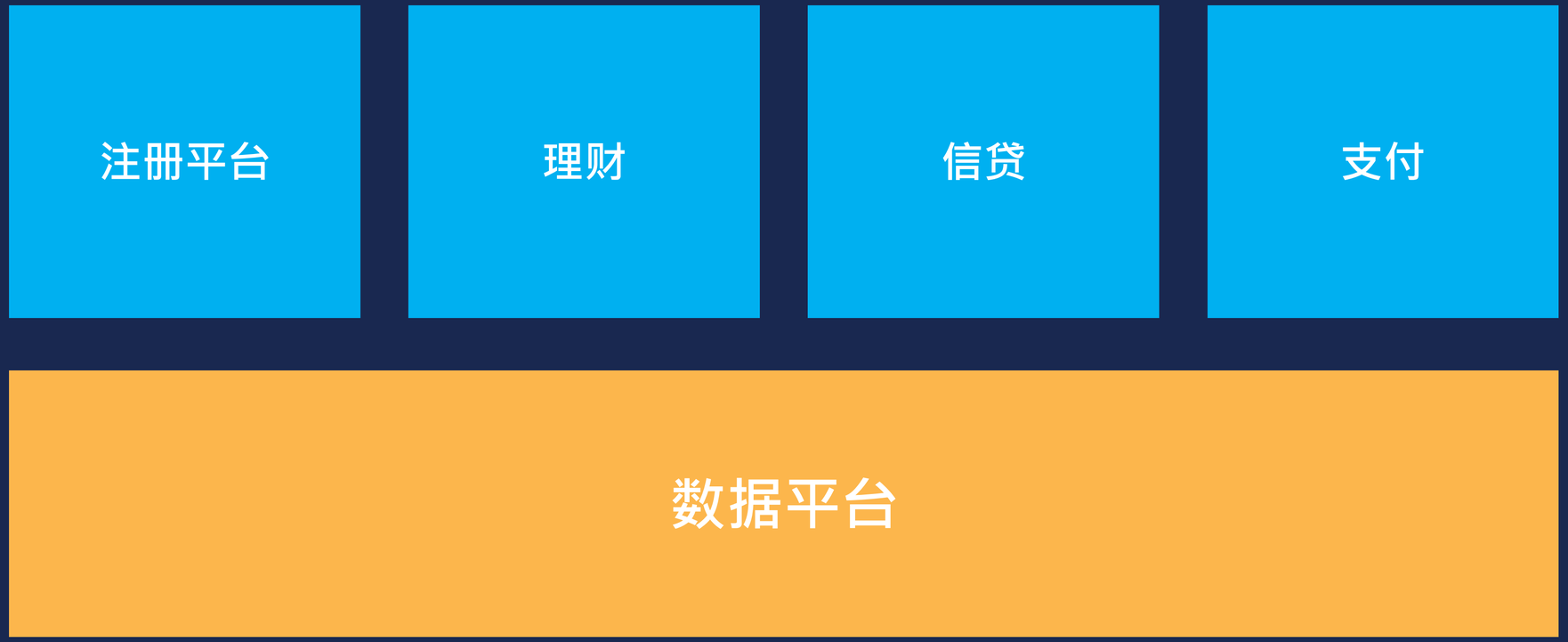
Vskit

Palmcredit

PalmSave

palmpay

业务整体框架



业务特点

- 海量的线上用户
- 丰富的应用场景
- 快速的授信、审批、交易和高度自动化
- 实时反欺诈和快速迭代

数据平台特点

数据种类多

各种采集数据、业务数据、三方数据

数据类型多

结构化、半结构化、非结构化

处理方式多

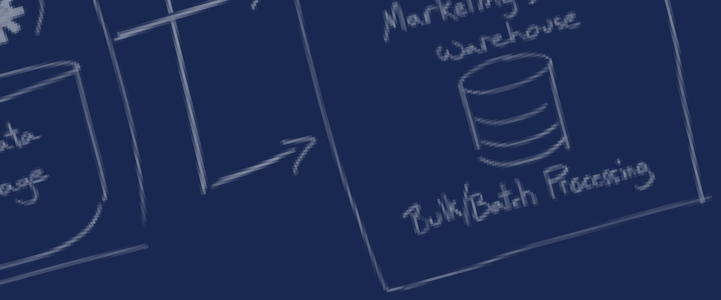
实时、准实时、离线

支持场景多

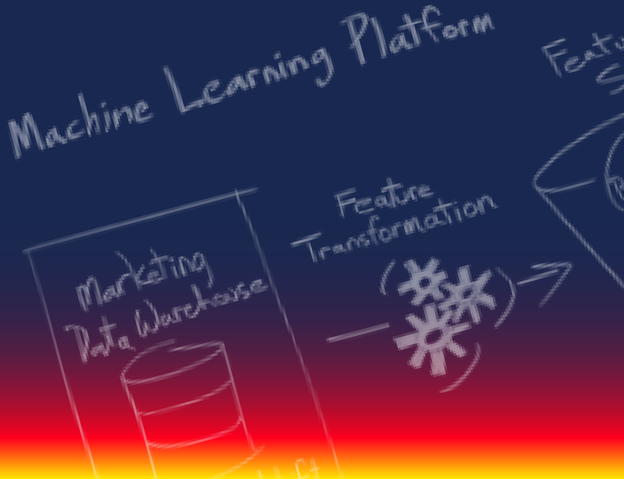
申请、审批、交易、分析

数据平台目标

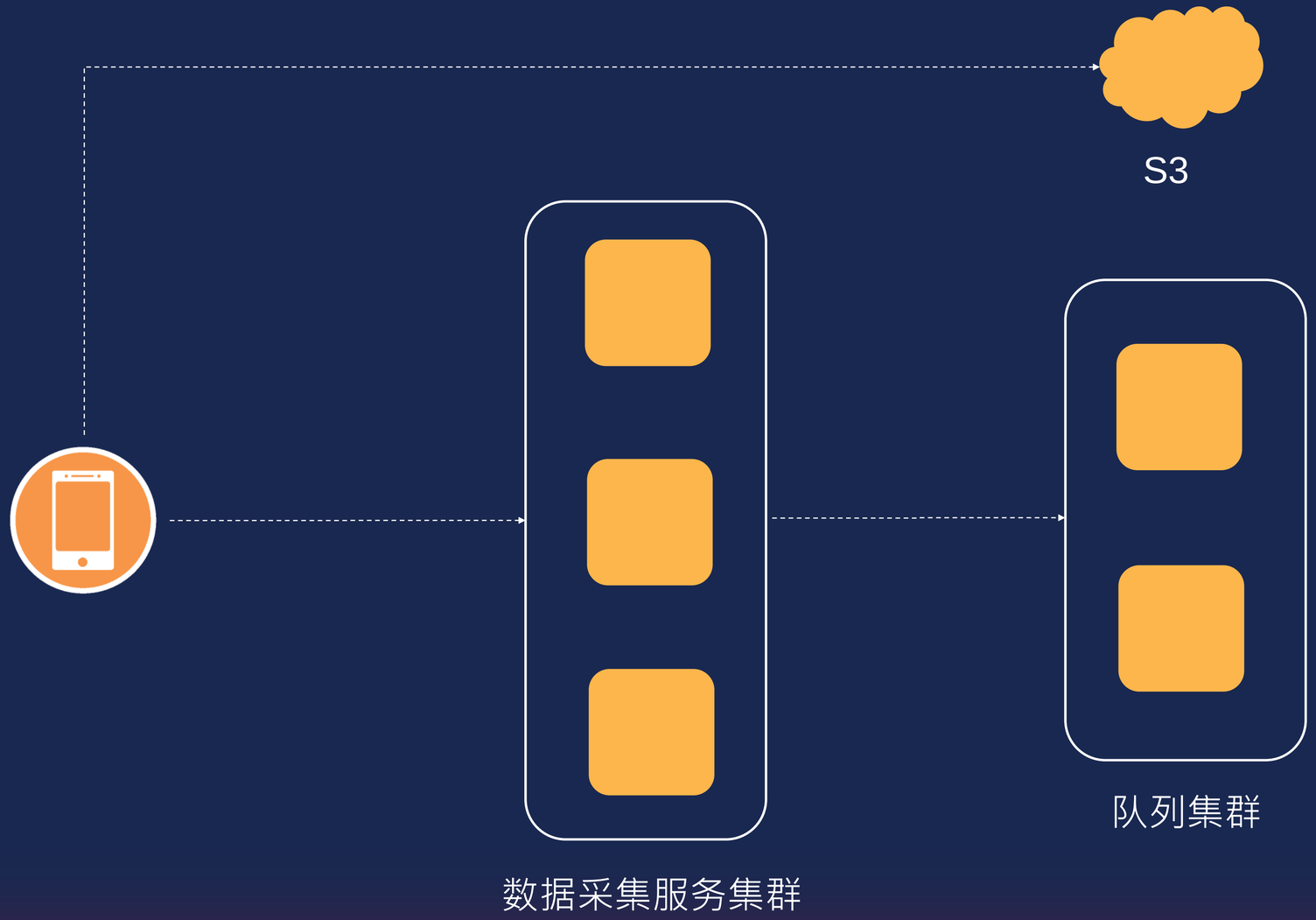
- 采集客户端10+种数据
- 实时和准实时解析200+风控变量
- 风控分析师能够灵活快速分析所有数据
- 高可用，支持横向伸缩


Marketing
Warehouse
Bulk/Batch Processing

我们的规划方案


Machine Learning Platform
Marketing
Data Warehouse
Feature
Transformation

数据采集



数据采集

数据上报

埋点数据、授权数据

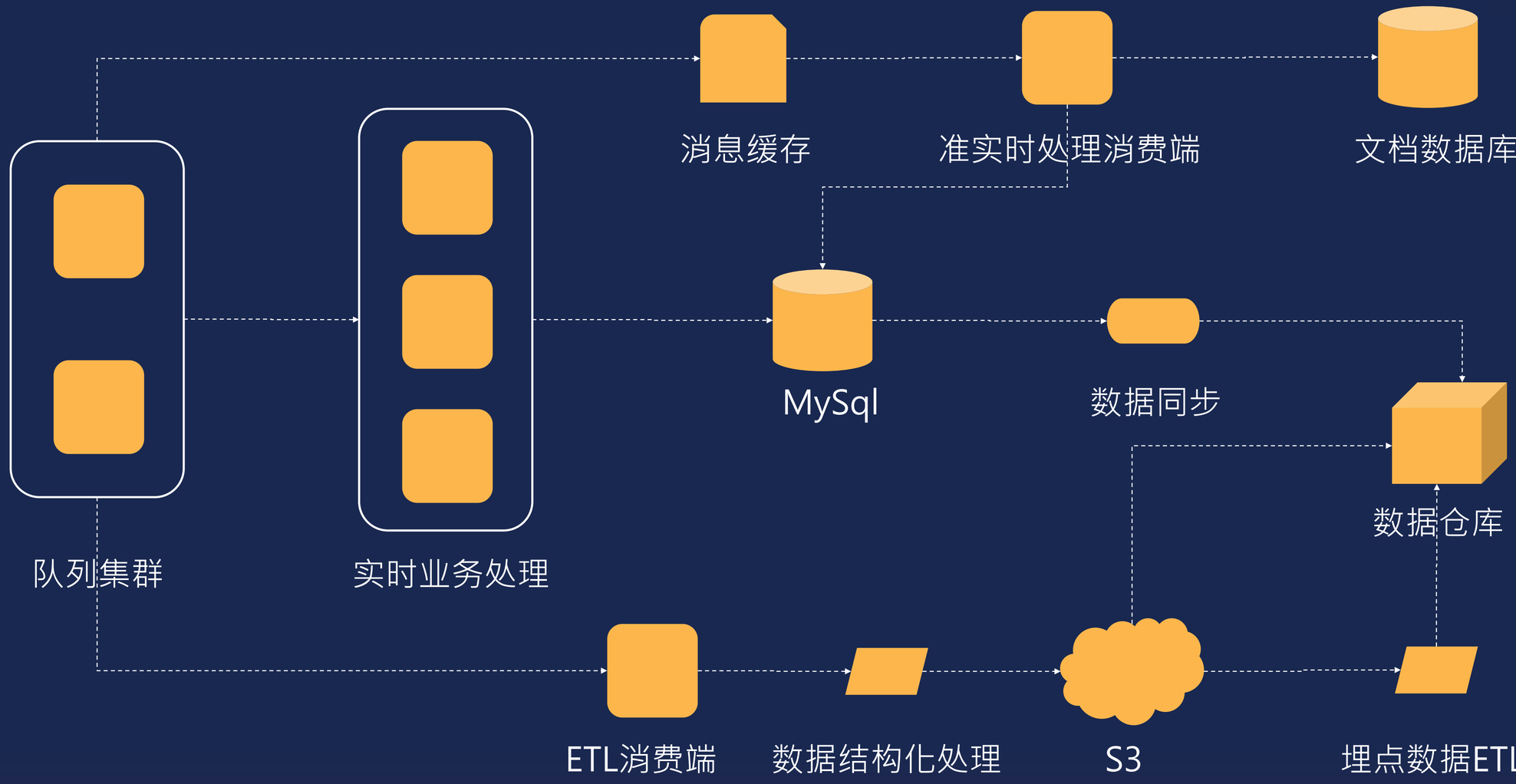
数据采集服务

接入节点，将数据转发到队列集群

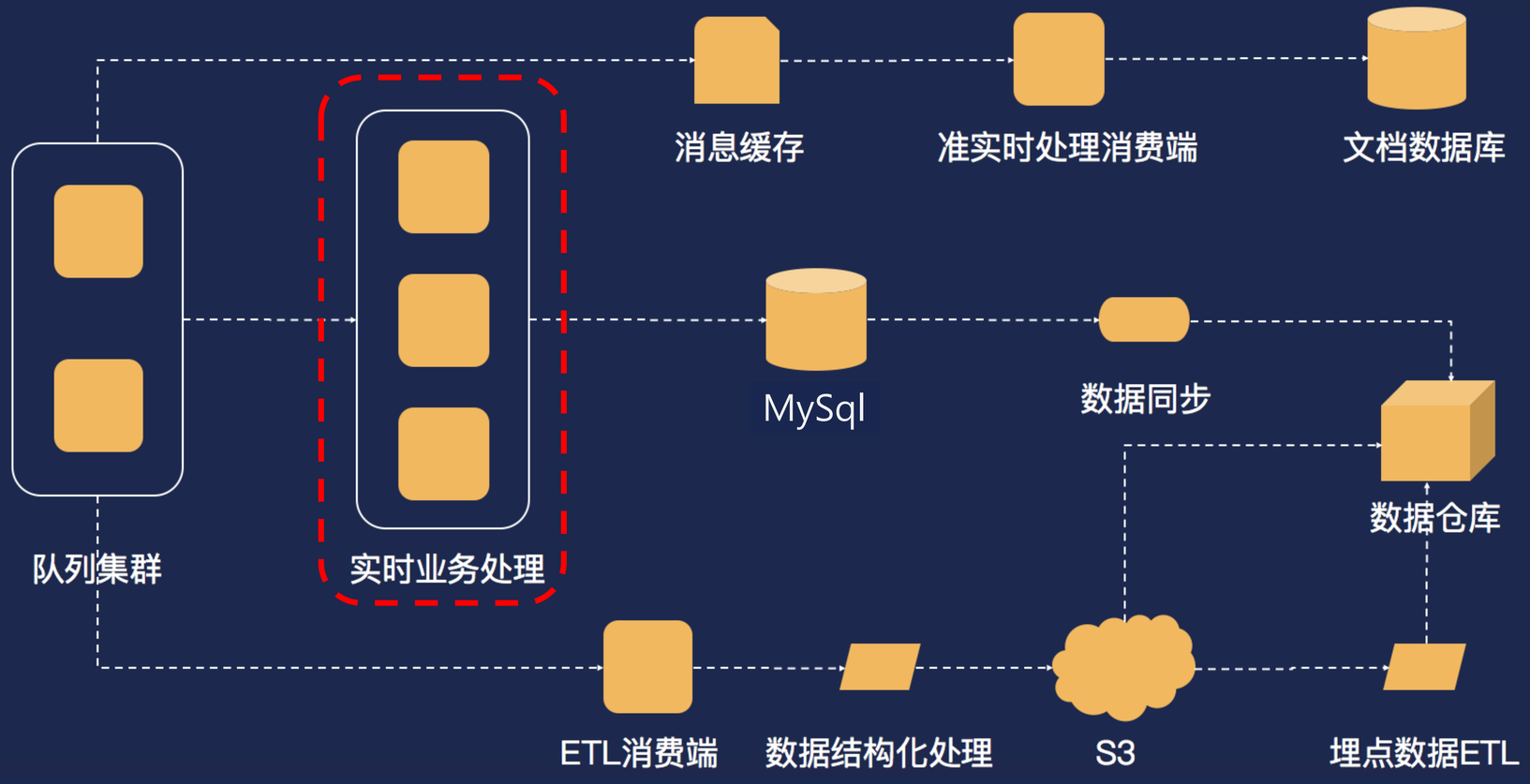
队列集群

NSQ消息队列

数据处理

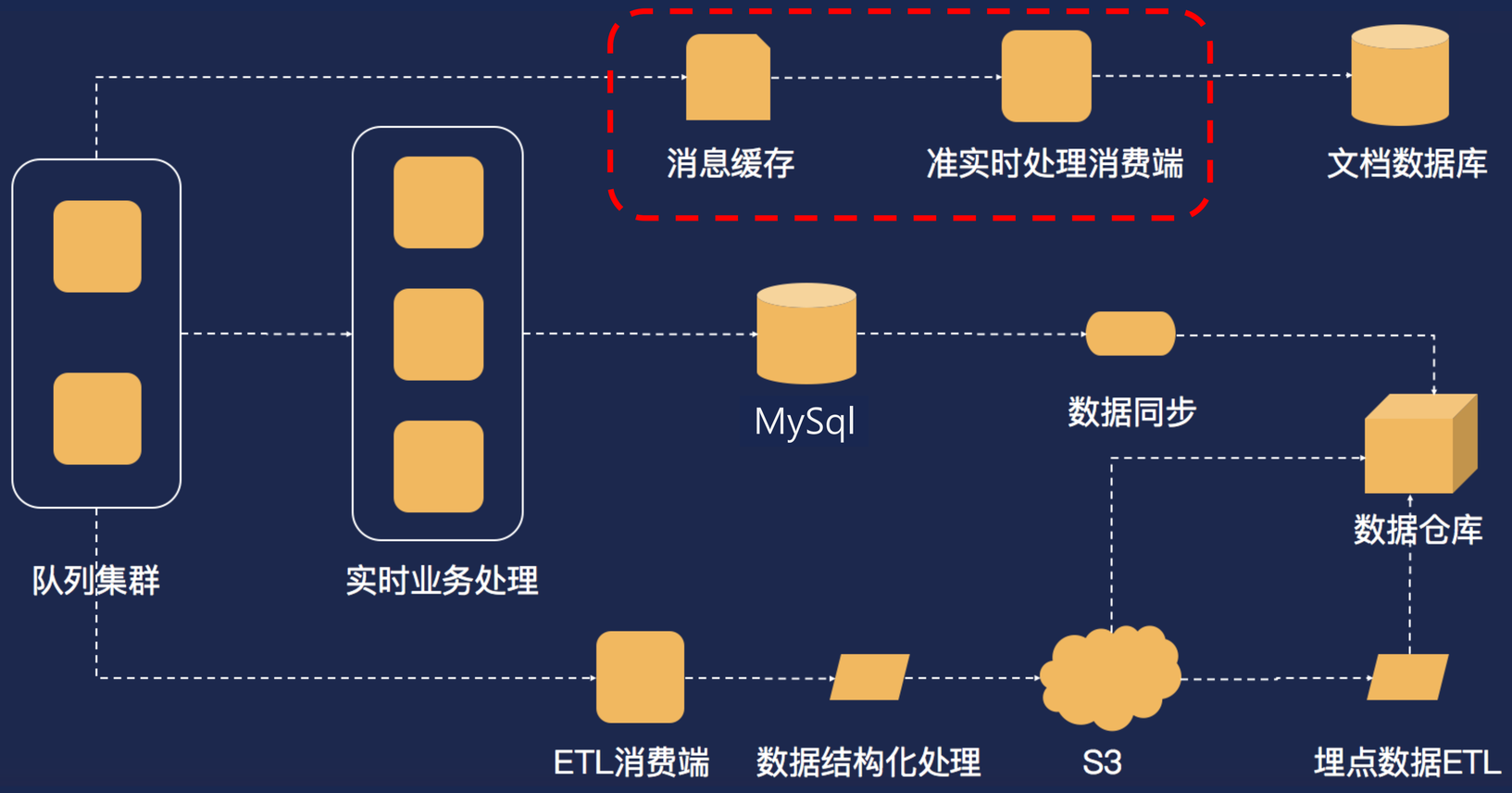


实时处理



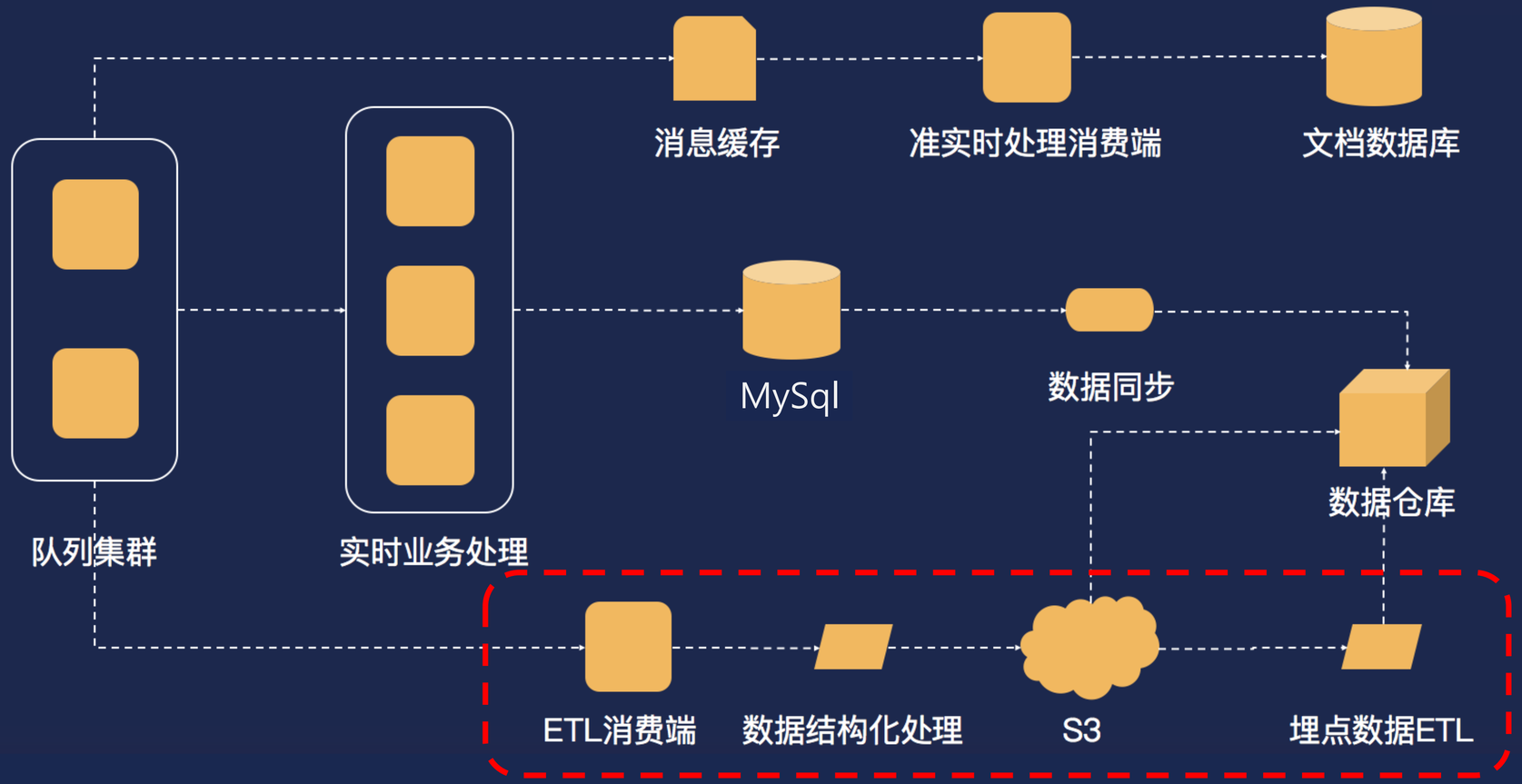
- 业务风控指标实时解析
- 用户授信、交易申请，根据申请时刻采集的数据出授信额度和交易审批
- 实时解析用户间的关系数据，识别欺诈交易

准实时处理



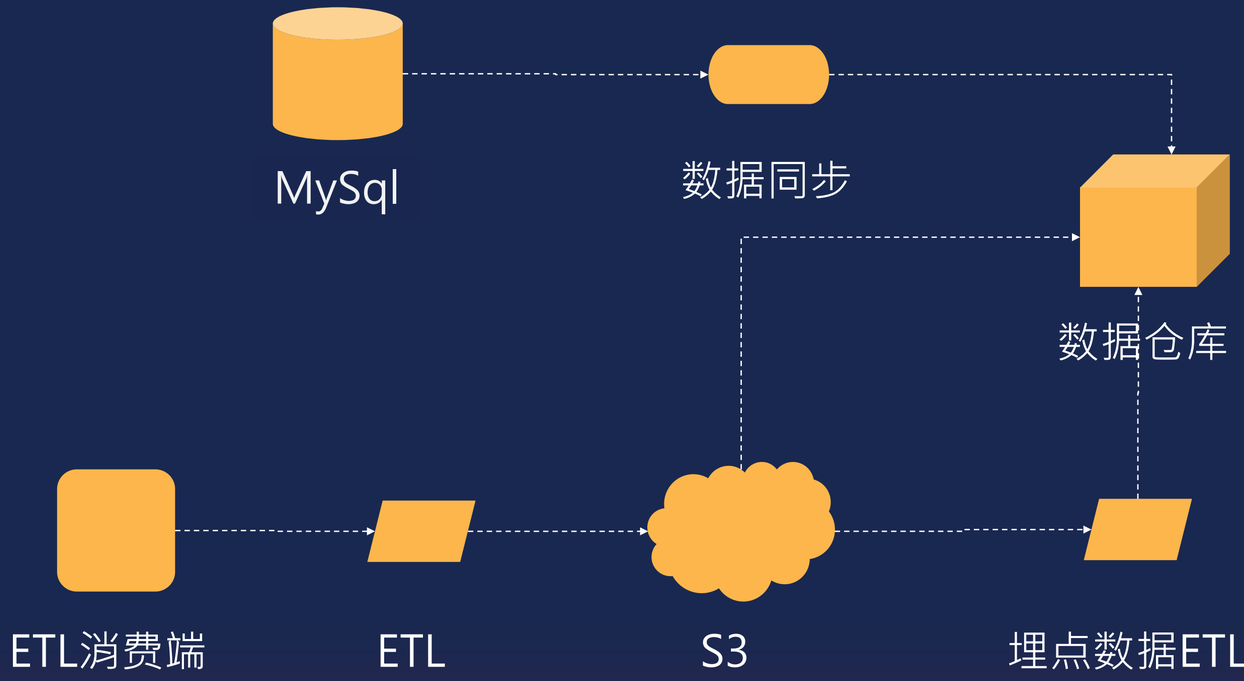
- 处理用户通讯录、通话记录、短信等信息
- 消费端故障，数据可以延迟处理，但不能丢失
- 支持慢消费和数据回溯
- Kafka可以满足消息缓存

ETL

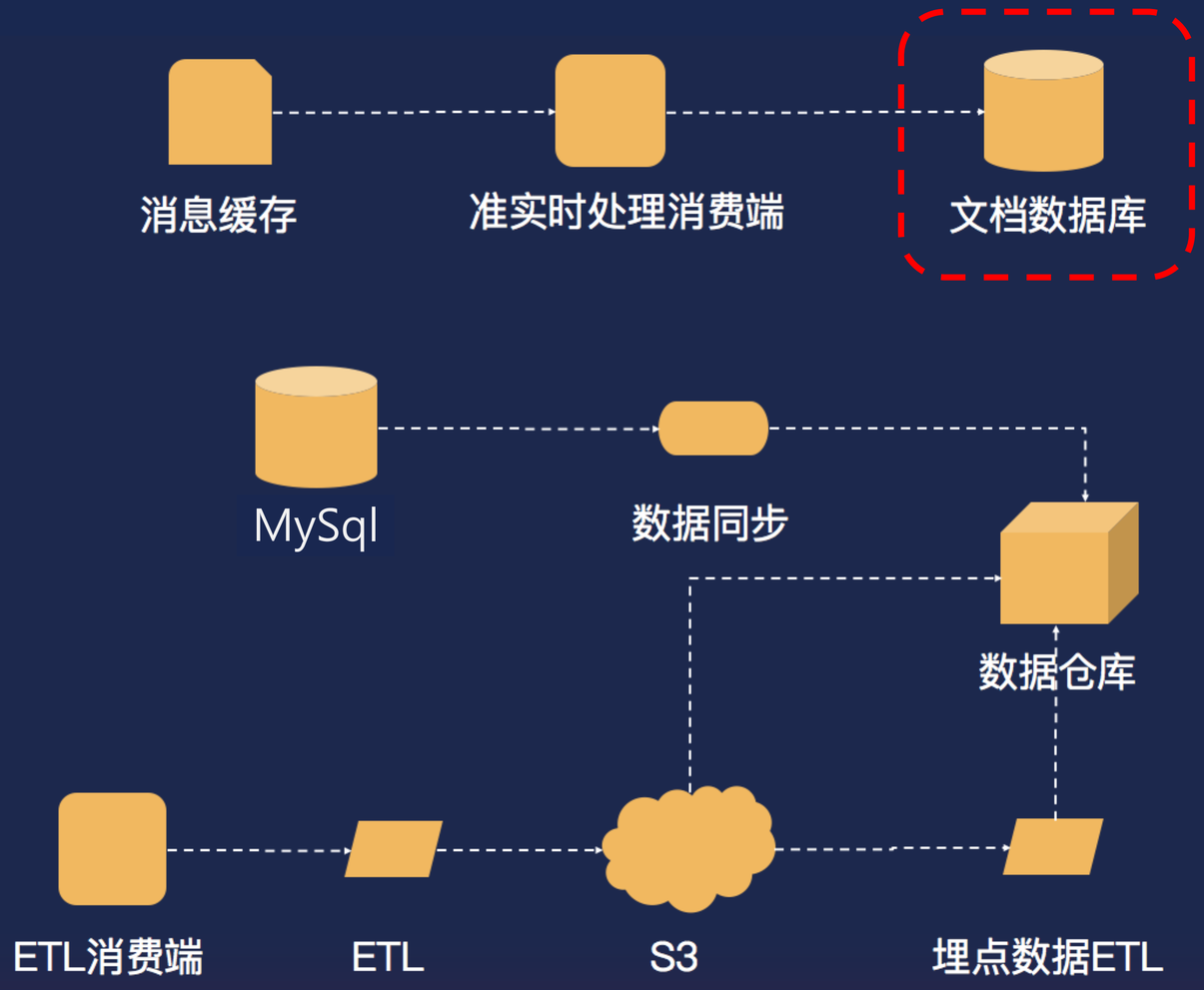


- 原始数据直接备份S3
- 将数据结构化处理，暂存S3
- S3结构化数据直接转入数据仓库
- 埋点数据ETL后，存入数据仓库

数据储存

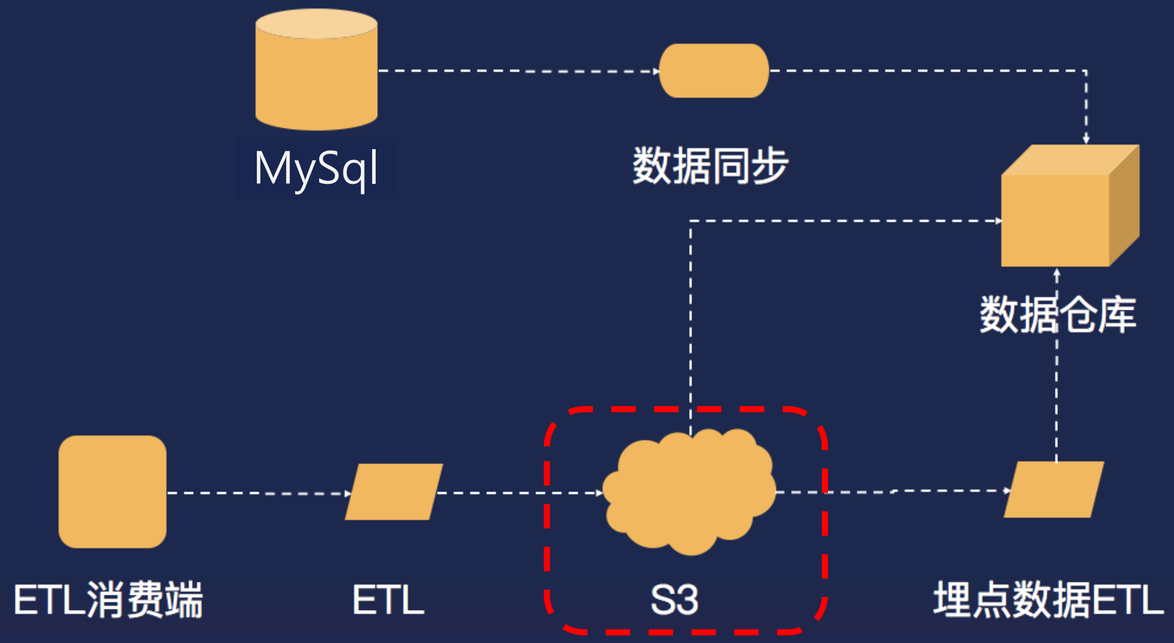


文档数据库



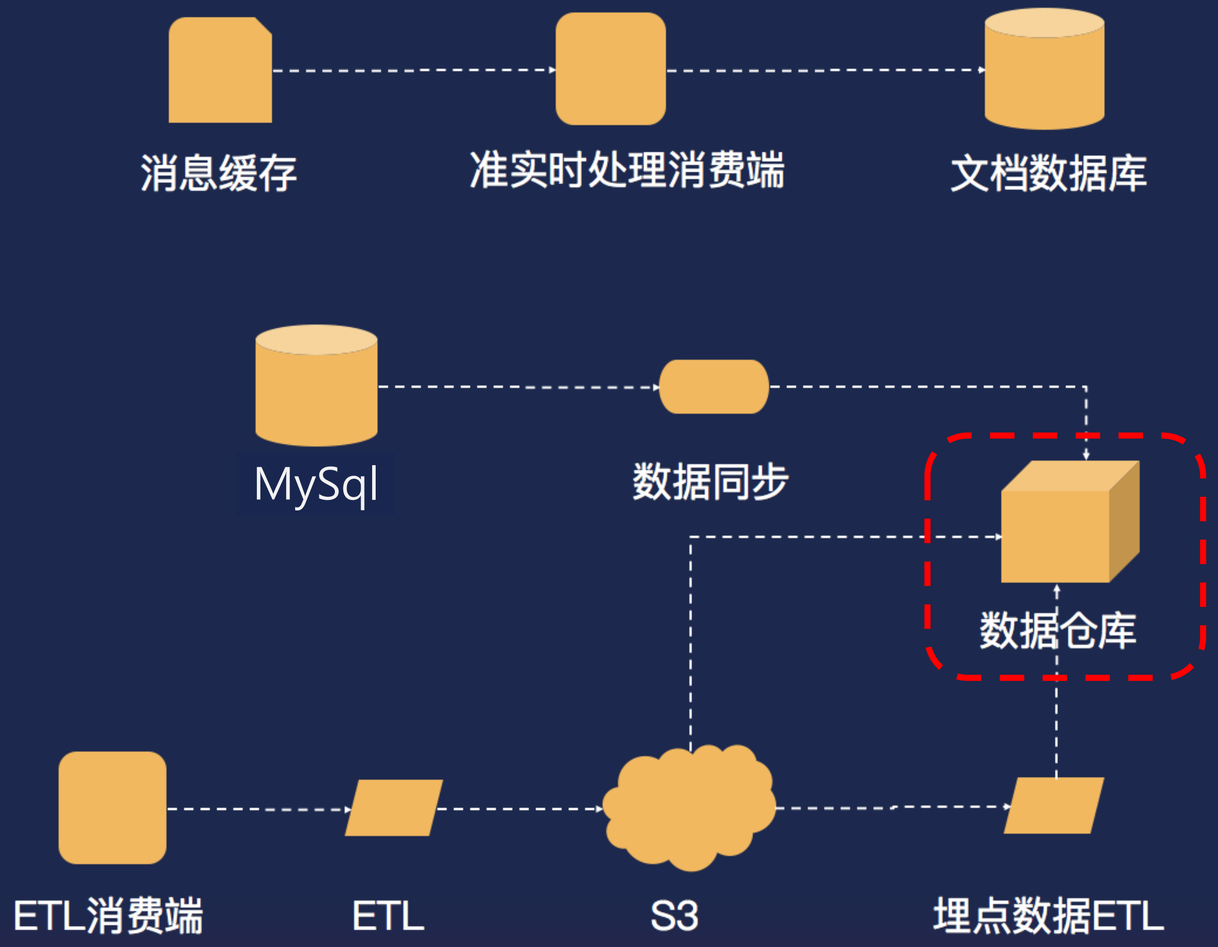
- 主要用于通讯录、短信等大块数据存储和解析
- 用户关系表的存储
- 要支持灵活扩展
- 可使用MongoDB来实现

S3



- 数据湖
- 所有原始数据
- 中间的处理数据

数据仓库



- 离线处理（OLAP）
- BI、风控分析
- 大规模并行处理
- 跟其他服务之间无缝对接
- 升级容易，易维护

数据平台方案

开源方案

Kafka + MongoDB + MySql + Sqoop + Hadoop + Hive

云服务方案

Kinesis + DynamoDB + Aurora + DMS + Redshift

数据平台方案

开发现状：

- 仅3名开发人员
- 基本无数据开发经验
- 开发人员的技术栈不同
- 两个月上线

数据平台方案

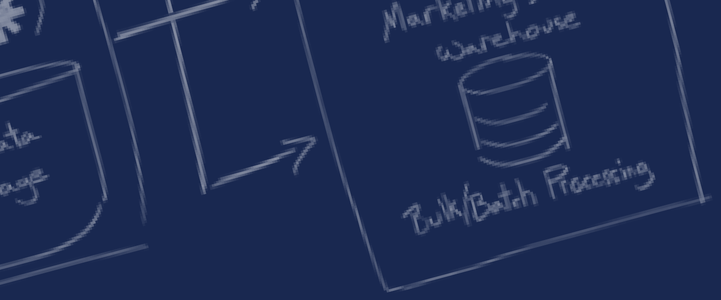
开源方案问题：

- 技术难度高，后期的维护和升级麻烦
- 稳定性低，开源问题较多，需要升级来保证
- 与其他服务的兼容性不好，需要额外开发保证
- 需要有经验的开发人力投入

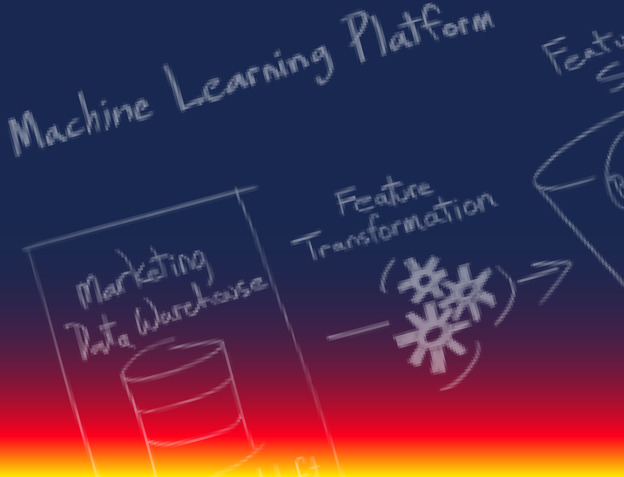
数据平台方案

原则：

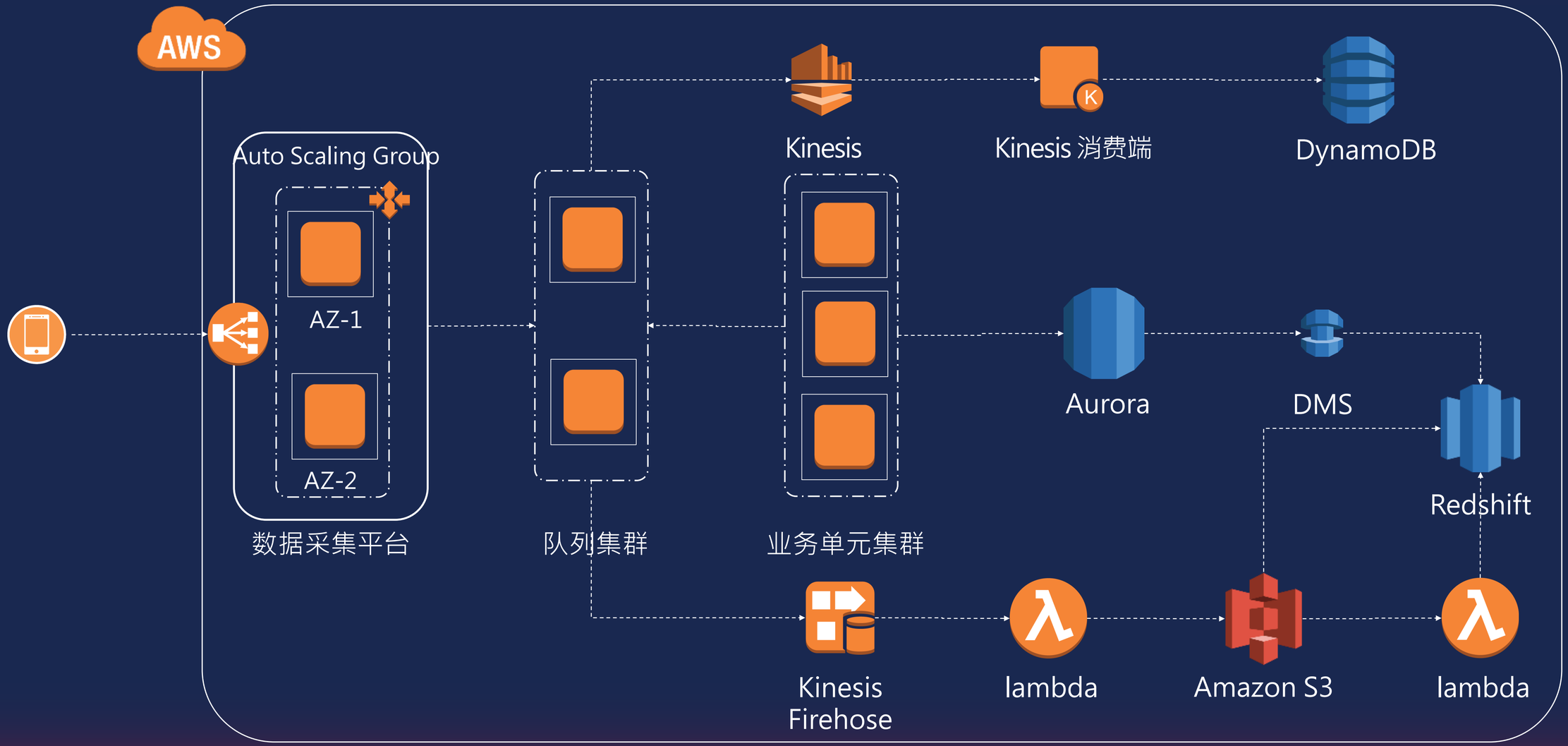
- 充分利用其他项目中已实现组件
- 优先选择开发人员熟悉的和上手快的
- 尽量整合AWS服务，减少基础服务人力投入
- 聚焦业务实现

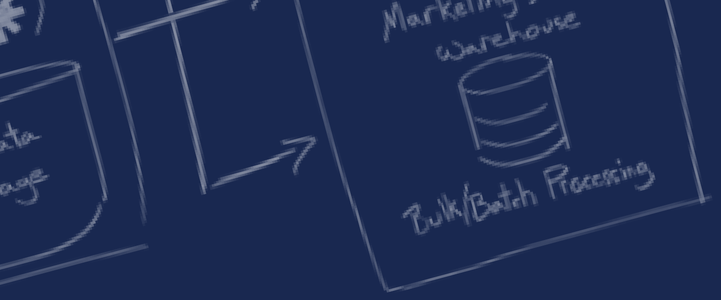

Marketing
Warehouse
Bulk/Batch Processing

最后方案的架构

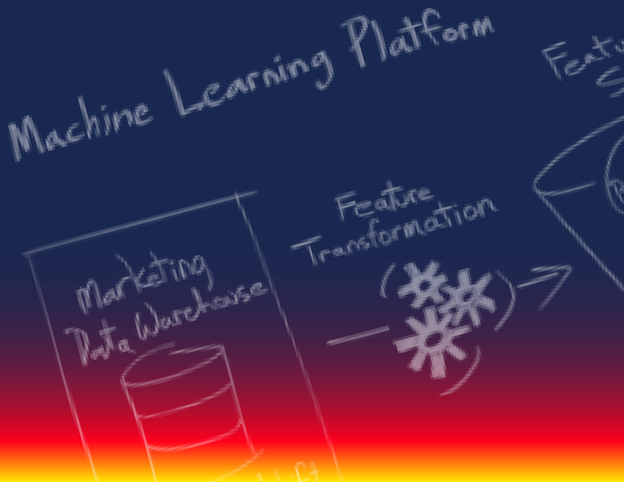

Machine Learning Platform
Marketing
Data Warehouse
Feature
Transformation

传易金融数据平台架构




Warehouse
Bulk/Batch Processing

一些使用注意事项


Machine Learning Platform
Marketing
Data Warehouse
Feature
Transformation

Kinesis

- 类似于Kafka
- Kinesis Data Streams API不能够处理多个应用间的负载均衡和记录已处理数据检查点等功能
- Kinesis Client Library (KCL)可以实现以上功能
- KCL支持的语言有限，不支持GO语言

Kinesis Firehose

- 将实时流传送到S3，Redshift
- 单次批量PUT限制500条
- 从S3 COPY数据到Redshift，遇到不兼容字符会导致任务阻塞，阻塞期间的数据会丢失
- 存储到S3的目录会自动加上年月日时
- 如果数据量不大，存储到S3的小文件较多，影响使用效率

DynamoDB

- 通过非主键scan，出现过读不到数据的情况
- Hash key和Range key的选择比较重要
- 支持通过指定Hash key批量查询，不支持批量查询Hash key

DMS

- 用于不同数据库之间数据同步和迁移
- DMS 对Redshift性能影响很大, 可以通过调整参数BatchApplyTimeoutMin来控制更新频率
- 如果Redshift性能持续偏高，会影响DMS同步，同步超时导致丢失数据而不报错
- 数据源改了数据类型会导致任务失败，整个表需要重新同步
- 只能从Aurora读写节点同步，不支持从只读节点同步

Redshift

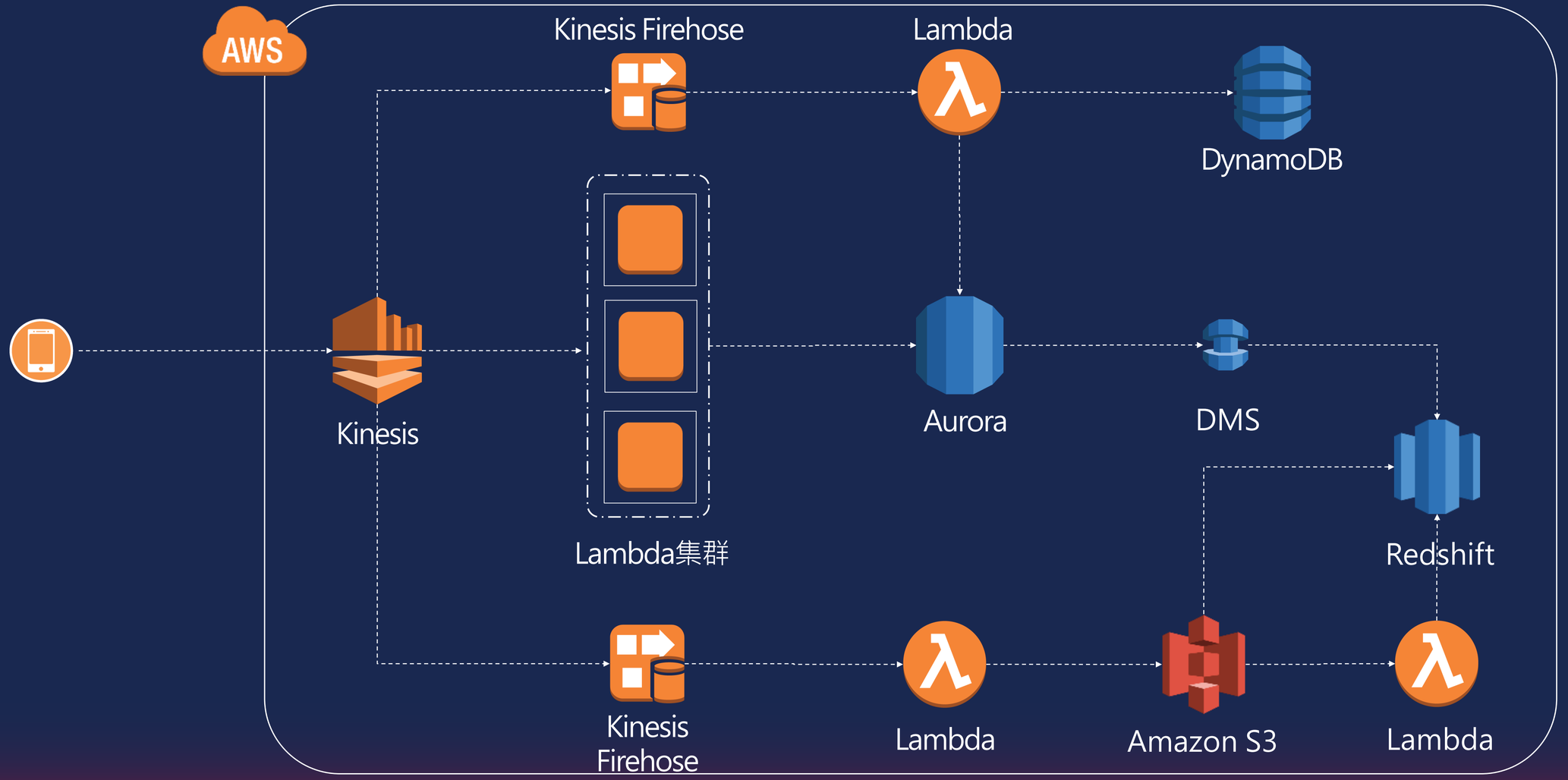
- 使用COPY命令，varchar类型长度必须是正常长度的3倍
- 设计表的时候 distkey 和 sortkey 要在一开始想好，创建表后无法更改
- 表创建后，字段的类型无法修改
- 尽量不要使用sql直接做数据清洗操作

架构迭代

存在问题和优化点：

- 数据采集平台需要额外维护，如果手机端直接上报数据到kinesis，可以直接干掉数据采集平台
- Kinesis直接替换掉队列集群
- 用户的复杂关系需要上图数据库处理
- 业务集群可以考虑使用lambda实现，实现无服务器
- 随着业务量的增长可以上EMR做ETL和流处理
- Redshift数据根据时间分表，冷数据转移到S3，挂载回Redshift

传易数据平台架构演进-无服务器架构



一些体会

AWS云服务方案：

- 技术难度低，重点学习产品应用方法
- 利用托管服务，团队专注于业务，项目快速上线
- 稳定性有保证
- AWS服务之间的集成性，缩短了开发周期，在构建应用时有很灵活的选择
- 提供专业的服务
- 在运营过程中， AWS的弹性机制，灵活控制成本

aws SUMMIT SHENZHEN

Thank you!

aws SUMMIT SHENZHEN

谢谢

扫码下载演讲资料