贝壳一站式大数据开发平台实践

人工智能技术中心@贝壳找房

日期：2020年9月

个人简介

仰宗强

- **贝壳找房　人工智能技术中心-大数据平台**
- 趣店 数据应用
- 百度 流量平台
- 当当 数据平台
- 西安科技大学 学士

目录

背景介绍-贝壳数据业务

➢ **将贝壳数据 分为物的数据、人的数据、行为数据三大块来进行研究。**

物的数据

- 楼盘字典：2亿+套房屋信息

行为数据

- 线上行为
- 多样的线下行为

人的数据

- 买家 -> 上亿级别
- 业主
- 经纪人 -> 40万+
- 品牌

➢ **为了满足各类数据获取/计算等场景，业内涌现出了很多解决方案**

- 降本（降低数据处理成本）
- 增效（合理提高数据使用/计算效率）
- 规范化（提供统一的编程范式）

探索历程-历程回顾

最初阶段

- 搭建hadoop集群
- 采用HIVE构建数据仓库
- 定制化报表开发

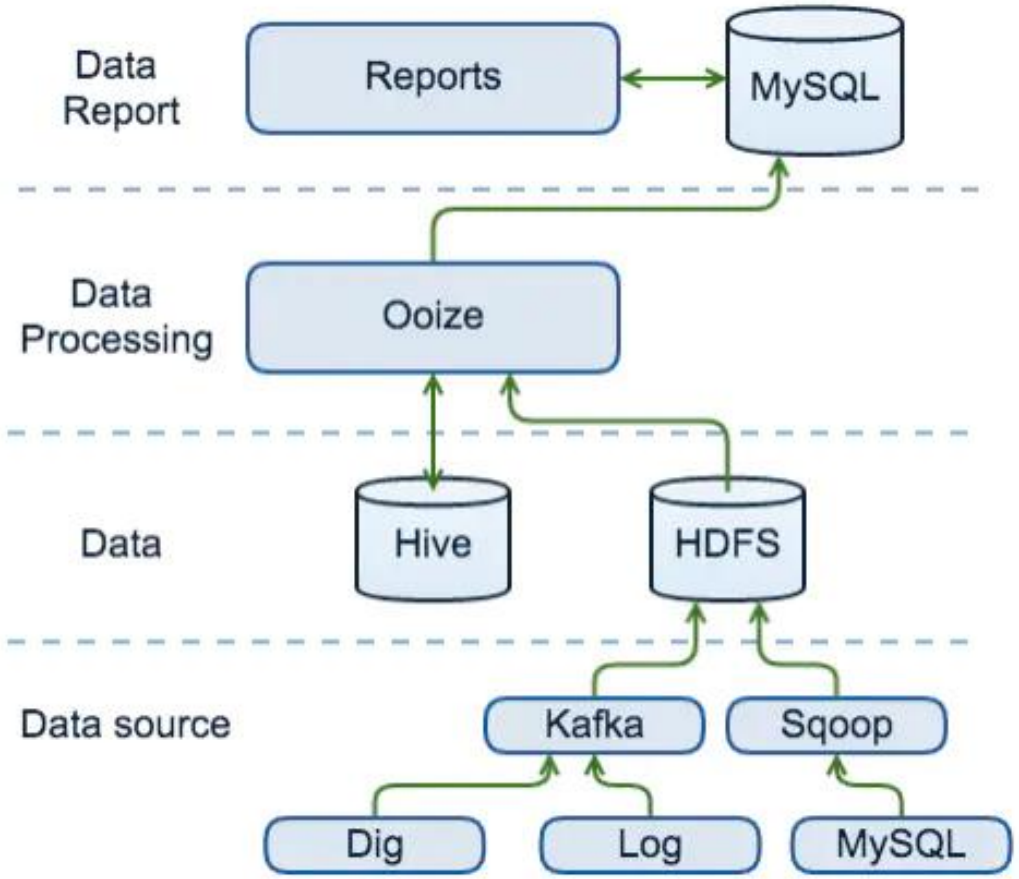
优化提升

- 元数据管理
- Adhoc自助查询
- 系统运维，简单监控

飞跃阶段

- 数据资产管理
- 数据开发套件
- 统一监控，智能运维

探索历程-初期的架构



优点

1. 开源组件，方便扩展和运维
2. 业界成熟的数据仓库方案，分层模型设计
3. 有利于技术人员培养

缺点

1. 需求迭代跟不上节奏
2. 数据仓库工程师累成狗。
3. 集群作业运维困难
4. 数据泄露

探索历程-历程回顾

最初阶段

- 搭建hadoop集群
- 采用HIVE构建数据仓库
- 定制化报表开发

优化提升

- 元数据管理
- Adhoc自助查询
- 系统运维，简单监控

飞跃阶段

- 数据资产管理
- 数据开发套件
- 统一监控，智能运维

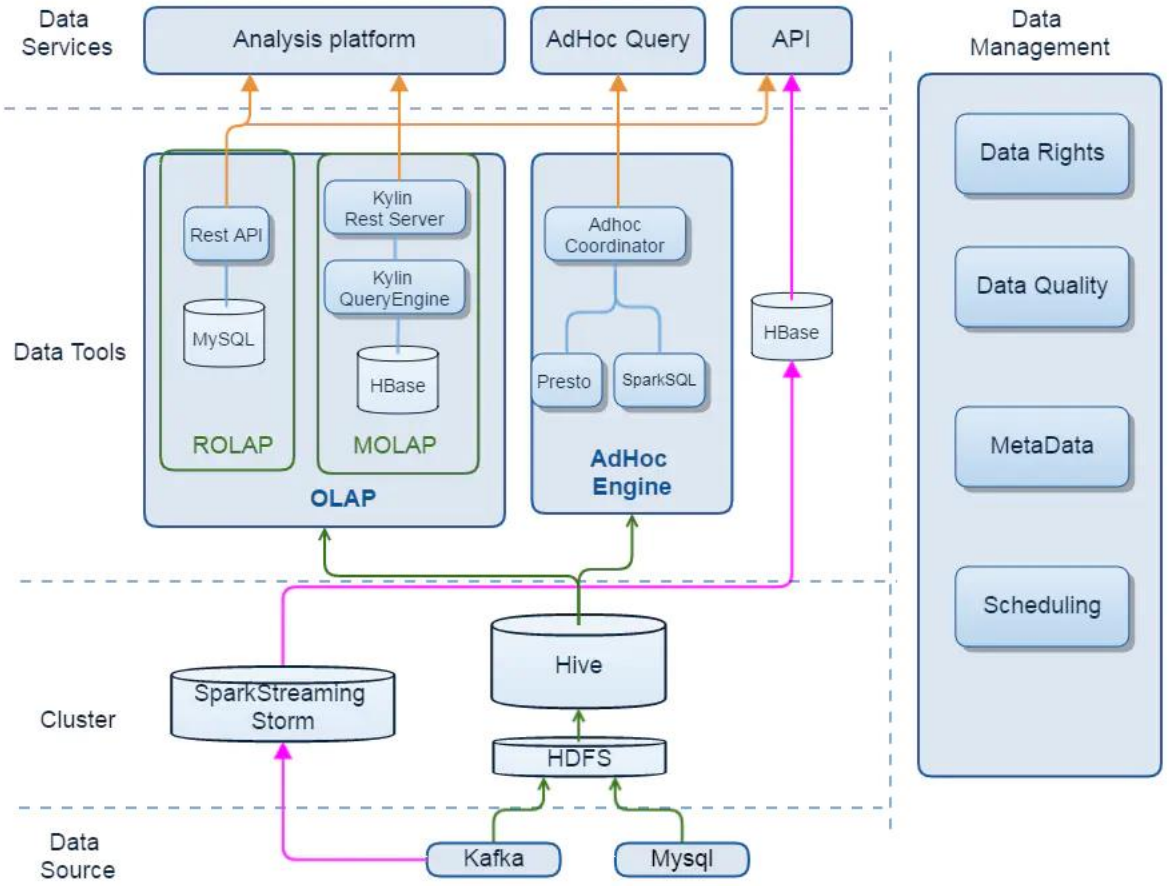
探索历程-平台化架构

优点

1. 解决了数据仓库的瓶颈问题。
2. 业务数据产品可以直接使用数据服务平台提高效率。
3. 自助运维，快速定位

缺点

1. 数仓表/任务量剧增，造成集群资源相当紧张。
2. 集中的几个人 -> 分散到业务，质量无法把控



探索历程-历程回顾

最初阶段

- 搭建hadoop集群
- 采用HIVE构建数据仓库
- 定制化报表开发

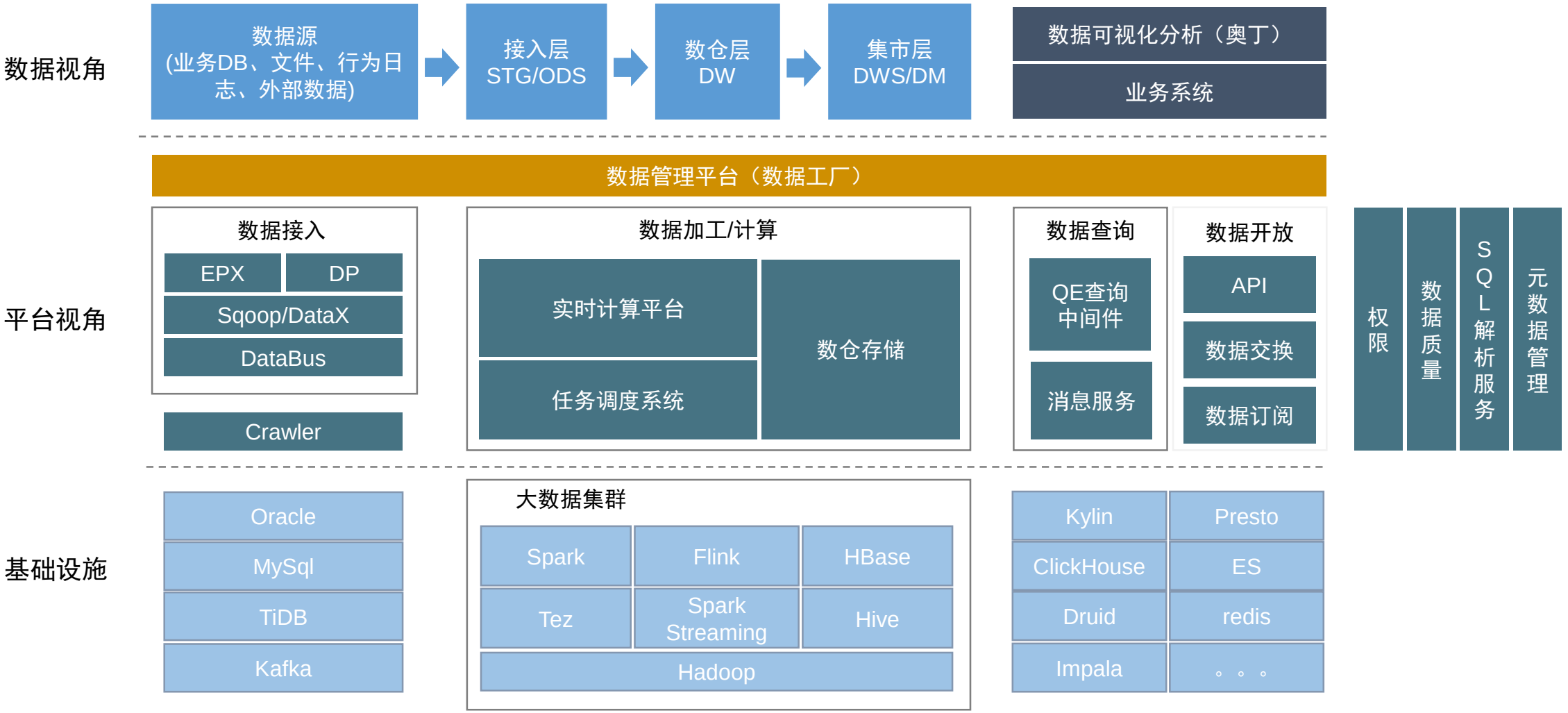
优化提升

- 元数据管理
- Adhoc自助查询
- 系统运维，简单监控

飞跃阶段

- 数据资产管理
- 数据开发套件
- 统一监控，智能运维

平台架构


数据视角
数据源
(业务DB、文件、行为日志、外部数据)
接入层
STG/ODS
数仓层
DW
集市层
DWS/DM
数据可视化分析（奥丁）
业务系统
平台视角
数据管理平台（数据工厂）
数据接入
EPX
DP
Sqoop/DataX
DataBus
Crawler
数据加工/计算
实时计算平台
任务调度系统
数仓存储
数据查询
QE查询中间件
消息服务
数据开放
API
数据交换
数据订阅
权限
数据质量
SQL解析服务
元数据管理
基础设施
Oracle
MySql
TiDB
Kafka
大数据集群
Spark
Flink
HBase
Tez
Spark Streaming
Hive
Hadoop
Kylin
Presto
ClickHouse
ES
Druid
redis
Impala
。。。

平台介绍-功能划分

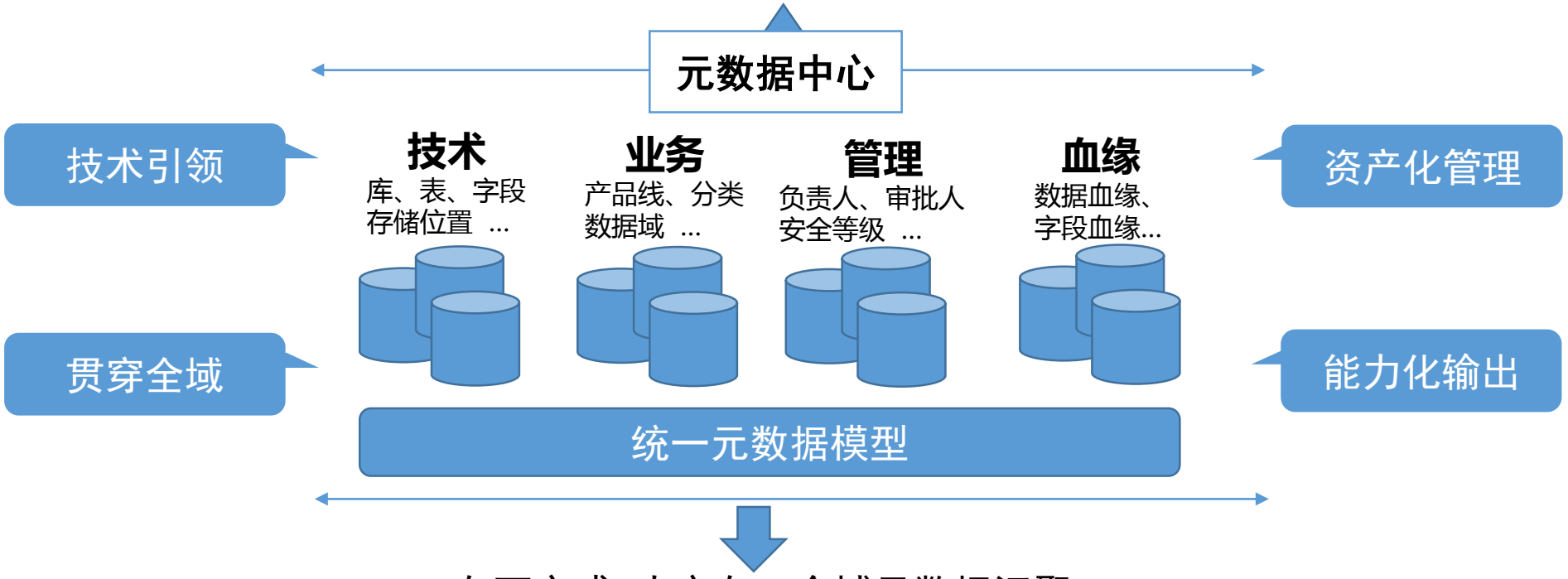
1. 数据管理
2. 数据集成
3. 作业调度
4. 数据质量
5. 数据开放

平台介绍-1. 数据管理

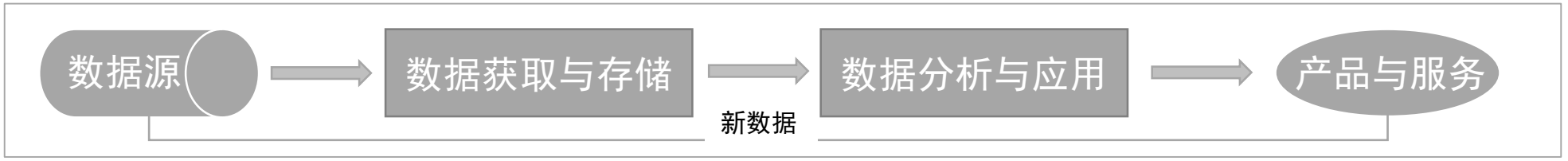
数据管理：以元数据为基础来构建贝壳数据资产的统一管控平台

元数据服务　能力开放　对内应用　对外价值释放

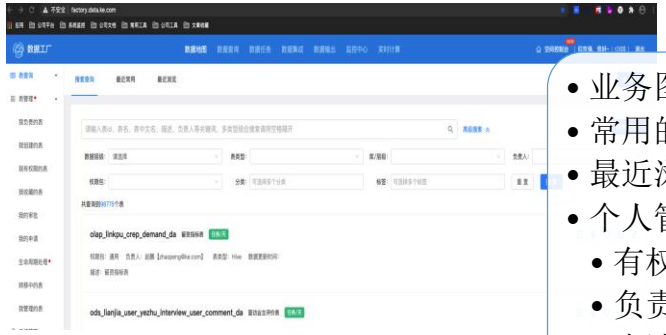
向上对内进行数据资产管理，对外提供价值开放



向下完成4大方向，全域元数据汇聚



平台介绍-1. 数据管理

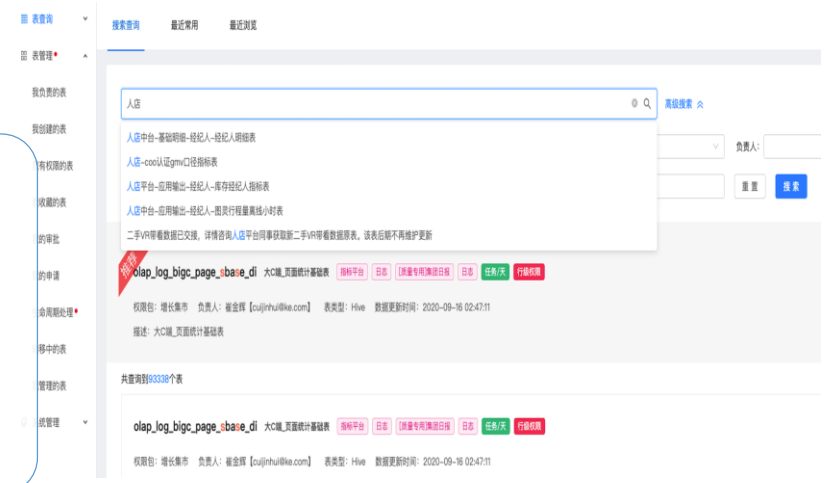


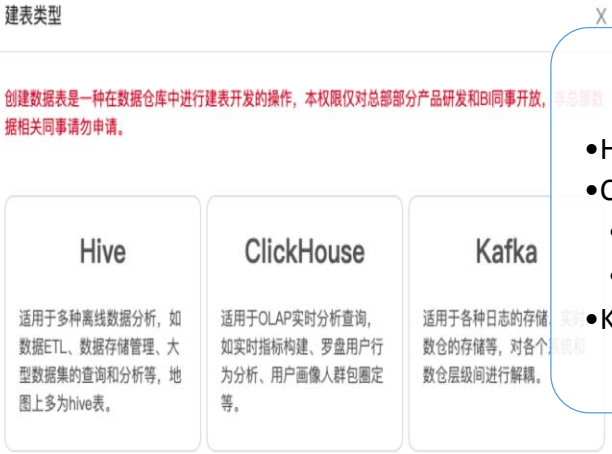
- 业务图谱
- 常用的表
- 最近浏览的表
- 个人管理模块
  - 有权限的表
  - 负责/创建的表
  - 申请/审批
  - 表交接转移

数据导航

智能搜索

- 根据表中文搜索
- 根据表英文名搜索
- 根据描述、标签、负责人搜索





数据录入

- Hive
- CK
  - 物理表
  - 视图
- Kafka

资产管理

- 数据生命周期管理
- 空间信息展示
- 空间资产管理分析
- 团队协作

平台介绍-2. 数据集成

数据集成最终目标是交付一套有效**提升集成效率**和**覆盖度的工具**

数据接入效率&覆盖度

多元集成

支持多种数据类型的集成，满足公司99%以上业务

能力提升

自动化：新建库/表自动接入（DB）

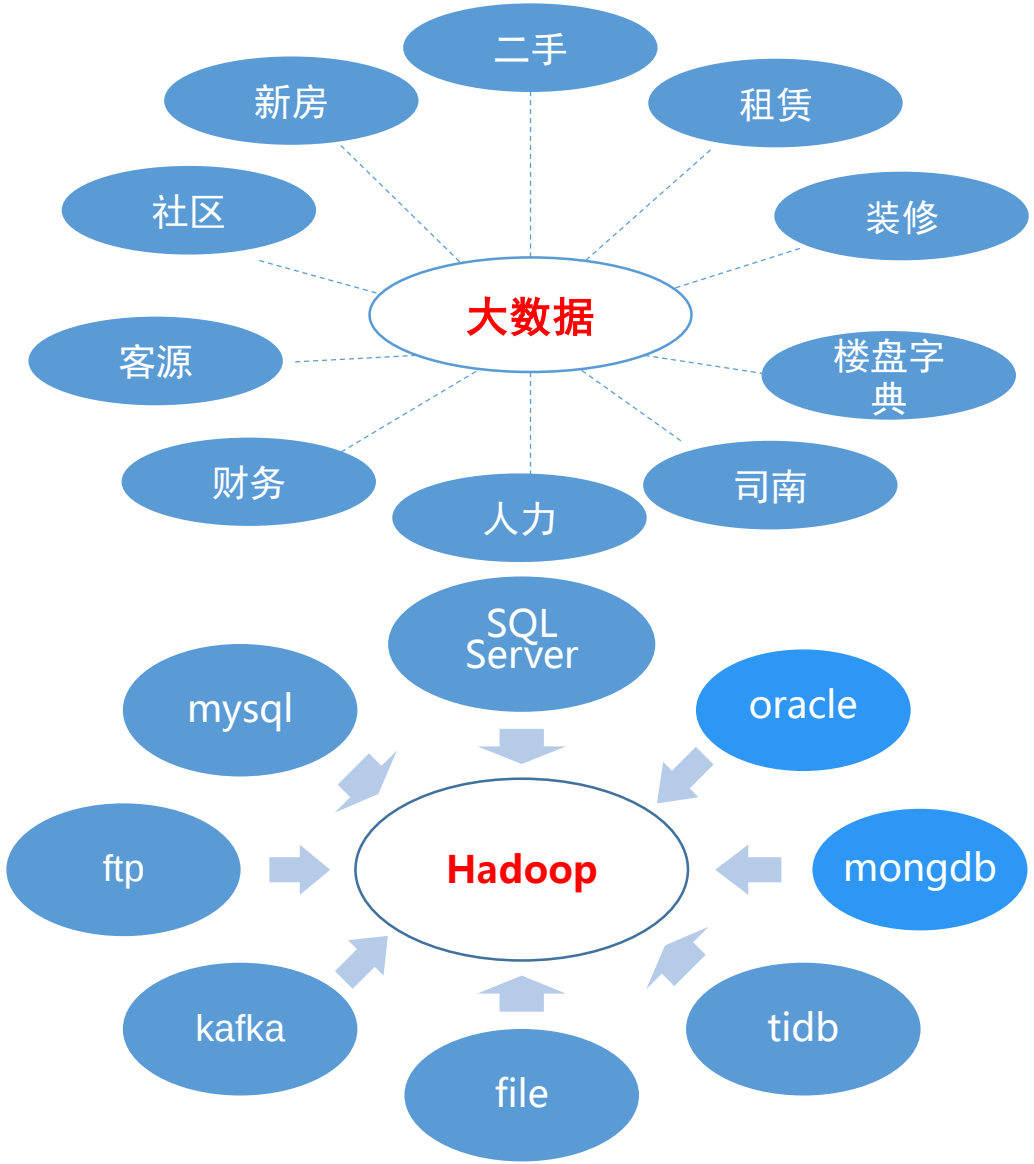
迁移提效：整库/表迁移自助

接入速率：T+1时效

长效保障

变更自动同步&提醒

接入任务监控机制保障



平台介绍-3. 作业调度

创建表 → 基本信息 → 字段信息 → 数据逻辑 → 任务调度 → 版本测试 → 数据上线

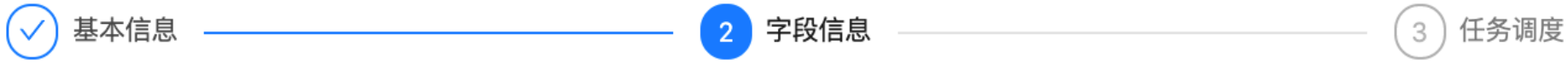
1 基本信息 — 2 字段信息 — 3 任务调度

基本信息

* 数据层级：	请选择	* 集群名：	hive集群
* 库/层级：		* 表名称：	表名称（小写英文字母）
* 表中文名称：	表中文名称	* 权限包：	
分类：	请选择	* 负责人：	请输入负责人
* 安全等级：	C级	* 审批人：	请输入审批人
* 所属空间：			
* 描述：	请输入描述		

平台介绍-3. 作业调度

创建表 → 基本信息 → 字段信息 → 数据逻辑 → 任务调度 → 版本测试 → 数据上线



添加字段方式： 可视化添加 | 语句添加

序号	*字段名	参照表	参照字段	*字段中文名	描述	枚举值		操作
1	pt	请输入参照表	请输入参照字段	分区字段	hive分区标识		stri	删除

添加新字段

上一步 仅创建表 创建并配置调度

平台介绍-3. 作业调度

数据源

* insert语句：

```
1
```

平台介绍-3. 作业调度

基本信息 — 字段信息 — 3 任务调度

* 任务周期: 请选择 crontab形式

* 任务有效期: 2020-10-13 19:00:52 2022-10-13 00:00:00

* 触发方式: 时间触发（按照crontab指定时间运行） 依赖触发（上游依赖就绪时，自动触发）

重试次数: 3

任务并发数: 1

* 启动超时: 设置启动超时报警 不设置启动超时报警

* 运行超时 任务运行时长超过 120 分钟后即可推送报警

* 任务队列:

* 需求方: 请选择需求方

依赖数据: 添加依赖 推荐配置

任务失败报警通知设置:

平台介绍-3. 作业调度

创建表 → 基本信息 → 字段信息 → 数据逻辑 → 任务调度 → 版本测试 → 数据上线

平台介绍-3. 作业调度

支撑批计算任务的管理、分发和执行，充当数据全链路中的心脏中枢

效果

- 任务配置简单，在页面上简单配置一个表单即可操作。
- 提供常用的ETL组件，零编码
- 依赖关系可视化，一键修复追溯，将排查问题修复数据的时间由一人天缩减到10分钟。
- 智能调度，错峰运行，保证高优先级任务优先运行。

平台介绍-4. 数据质量

保障数据资产健康，促进数据资产流通

数据准确性痛点

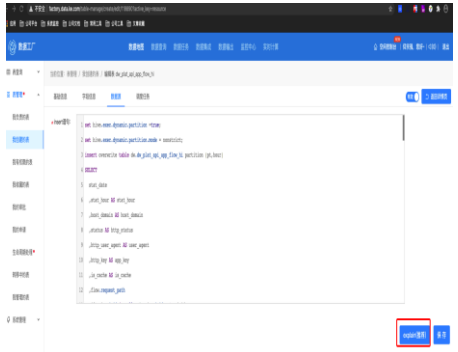
- 数据修改无法测试
- 数据例行无法监测

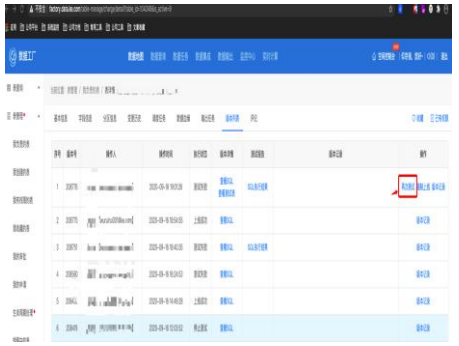
数据及时性痛点

- 例行任务失败、延迟无法感知
- 任务上游下线无法感知
- 怎么保障任务及时产出

解决方案

1 设计完善的开发流程，覆盖研发-测试-变更的事前全流程，识别并拦截潜在风险





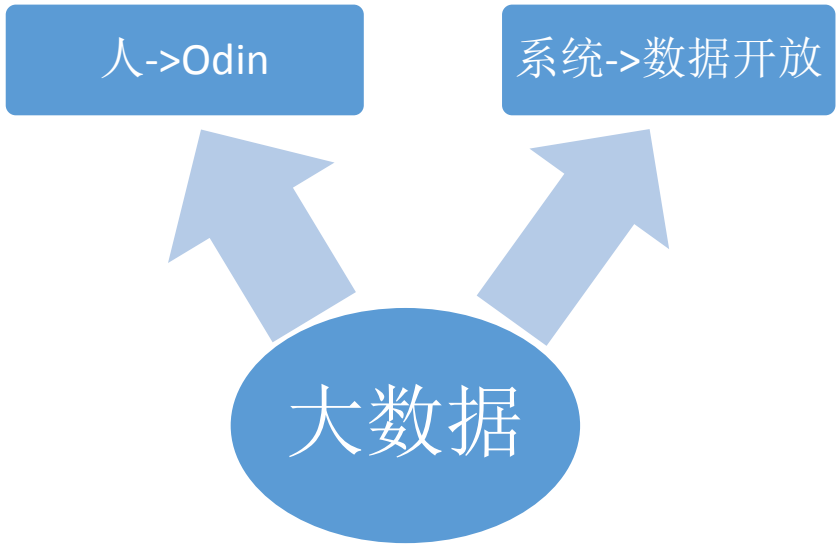
2 提供完善的任务监控服务，保证任务问题及时预警

- 通过任务失败报警，将失败的任务及时通知用户
- 通过超时报警，保证用户在规定时间未成功能够及时感知
- 通过表/任务链路报警，保障业务链路能够正常及时产出

3 提供多样性的数据质量监控，保障数据资产的准确性

平台介绍-5. 数据开放

提供 HIVE数据API、指标API、数据订阅、数据交换、QE代理 等各种能力和大数据输出的体系化解决方案。



应用场景

应用到APP场景

调用量随着业务变动

- HIVE表明细数据
  - 支持Mysql、Hbase、CK 引擎、QE代理
- 指标数据 → 指标API
- 实时数据
  - 支持redis、hbase、CK引擎

做业务融合

定时拉取HIVE数据，并发比较集中，有限定

- 数据交换
  - 支持端对端传输能力
- HIVE数据API（不推荐）
  - 支持Mysql、Hbase、CK 引擎、QE代理

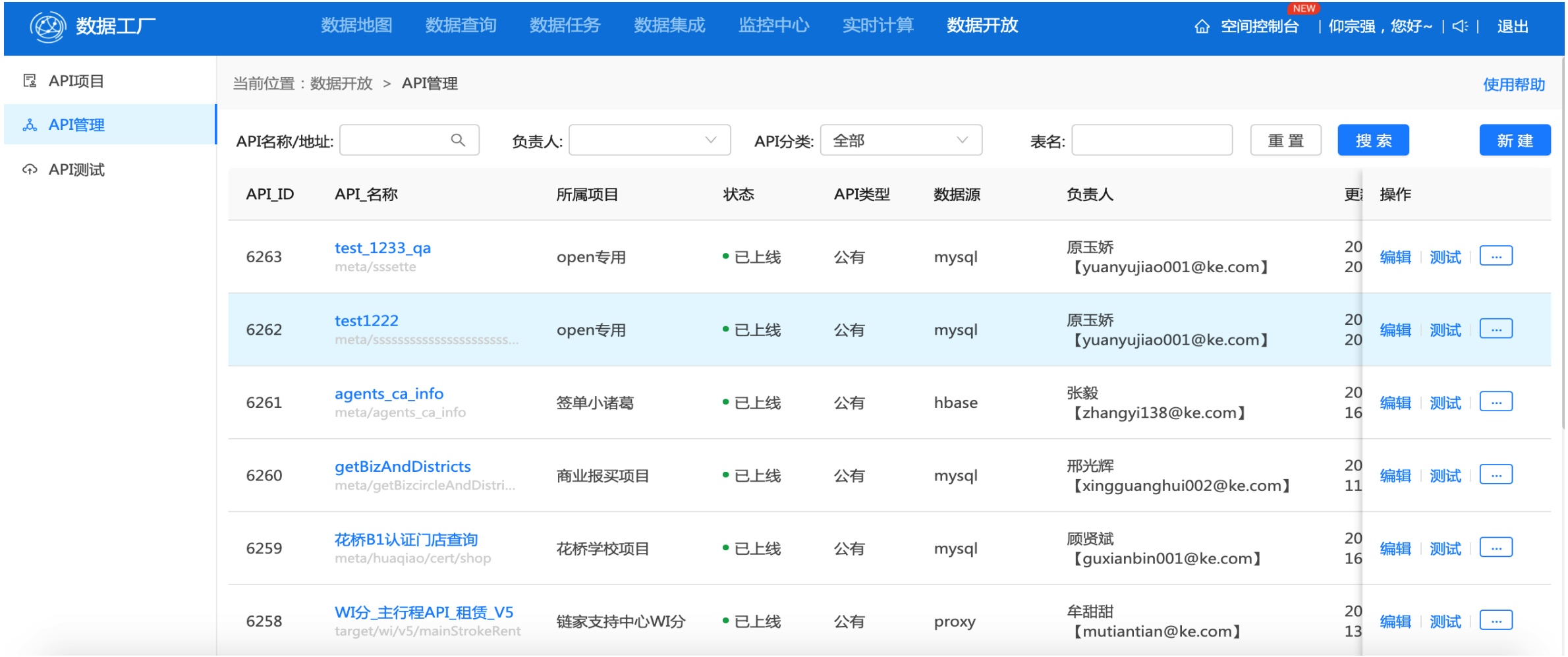
业务通知/push类场景

订阅数据类需求，实时消费能力

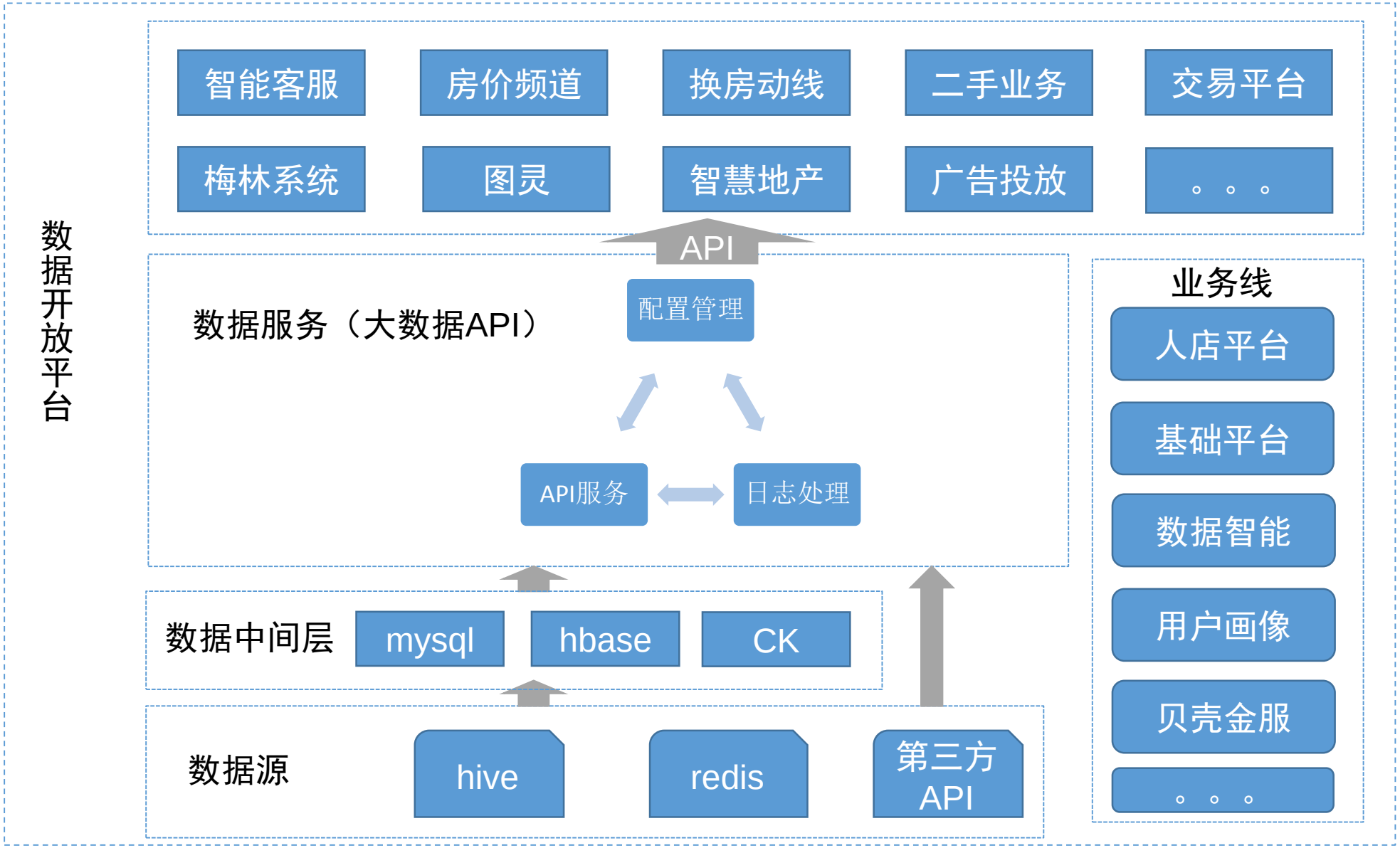
- EPX 数据订阅

平台介绍-5. 数据开放

创建API → 基本信息 → 请求信息 → 返回信息 → 排序分组 → Sql预览 → 生成上线



平台介绍-5. 数据开放


数据开放平台
智能客服
房价频道
换房动线
二手业务
交易平台
梅林系统
图灵
智慧地产
广告投放
。 。 。
API
数据服务（大数据API）
配置管理
API服务
日志处理
业务线
人店平台
基础平台
数据智能
用户画像
贝壳金服
。 。 。
数据中间层
mysql
hbase
CK
数据源
hive
redis
第三方API

总结与展望

- 数据资产化管理，全链路数据追踪和分析，提升数据价值
- 加密、脱敏、敏感监控等多种安全策略，全方位保障数据的存储、访问、传输过程
- 沉淀一系列的技术能力和组件集合，构建共性能力和通用服务，打造企业级大数据平台

THANK YOU