DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

数据治理-从无序走向秩序为企业数据资产增值提效

海保人寿 架构师 胡赵勇

PART 1

数据治理简介

数据治理定义,现状,痛点,常见误区以及推进模式,数据治理体系标准

1.1 数据治理的定义

- **狭义上**: 数据治理是指对数据质量的管理、专注在数据本身。

- **广义上**: 数据治理是对数据的全生命周期进行管理，包含数据采集、清洗、转换等传统数据集成和存储环节的工作、同时还包含数据资产目录、数据标准、质量、安全、数据开发、数据价值、数据服务与应用等，整个数据生命期而开展开的业务、技术和管理活动都属于数据治理范畴。广义的数据治理称为数据资产管理。

1.2 数据痛点

- 用数不知道找谁要
- 要的数据不知道准不准
  - 各部门的指标标准不一致
- 数据流向混乱
- 监管报送数据逻辑错乱
  - 保单登记平台,统信报送,east监管数据标准化报送 (建设时间不一,面向对象各异,保送标准逻辑不一样)
- 元数据和上下文的缺乏使得难以信任数据
  - 信任的缺乏使员工使用资源小心翼翼,害怕使用过时或不正确的信息

1.3 数据治理的的常见误区

- 误区一:缺乏统一的数据治理标准
  - 在企业中，数据是一种特殊的“资产”，它可以创造价值，但也可能导致风险。 这就需要我们用一套标准化的方式去管理、去治理
- 误区二:只做数据质量提升，忽略数据治理的过程管理
  - 很多企业数据治理过程管理不到位，对数据治理过程中的重要节点缺乏关注
- 误区三:数据治理不只是 IT 部门的宠儿
  - 也是公司领导层以及整个业务部重要事
  - 从源头抓起，找到源头部门,定责定权
- 误区四:不只是大数据平台单个系统的事情,而是整体信息架构的事情

1.4 数据治理的推进模式

- 数据来源于核心，从核心开始
- 由下至上,从细处着手
- 从业务中来,到业务中去

PART 2

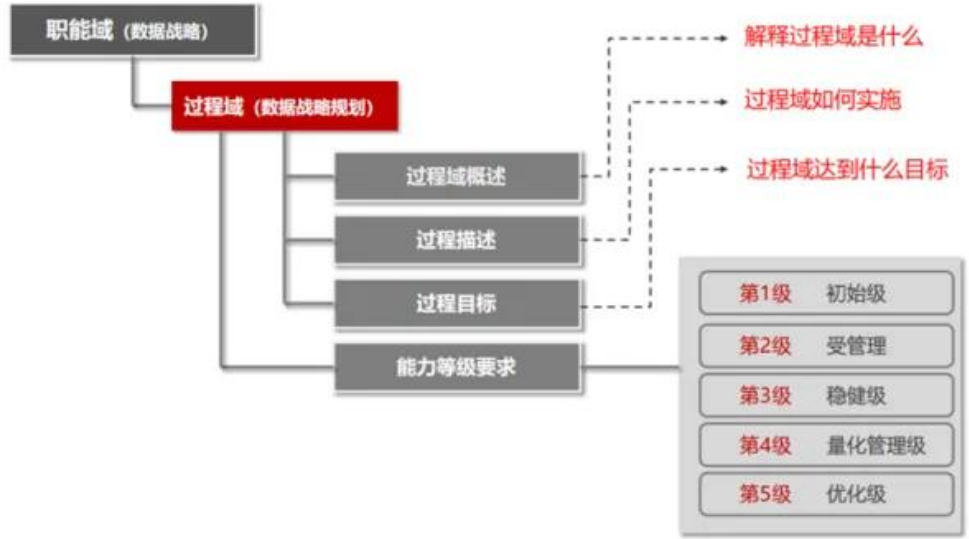
以DCMM为标准构建数据管理体系

DCMM内容,数据战略规划,数据战略指导数据治理,数据治理组织和制度建设

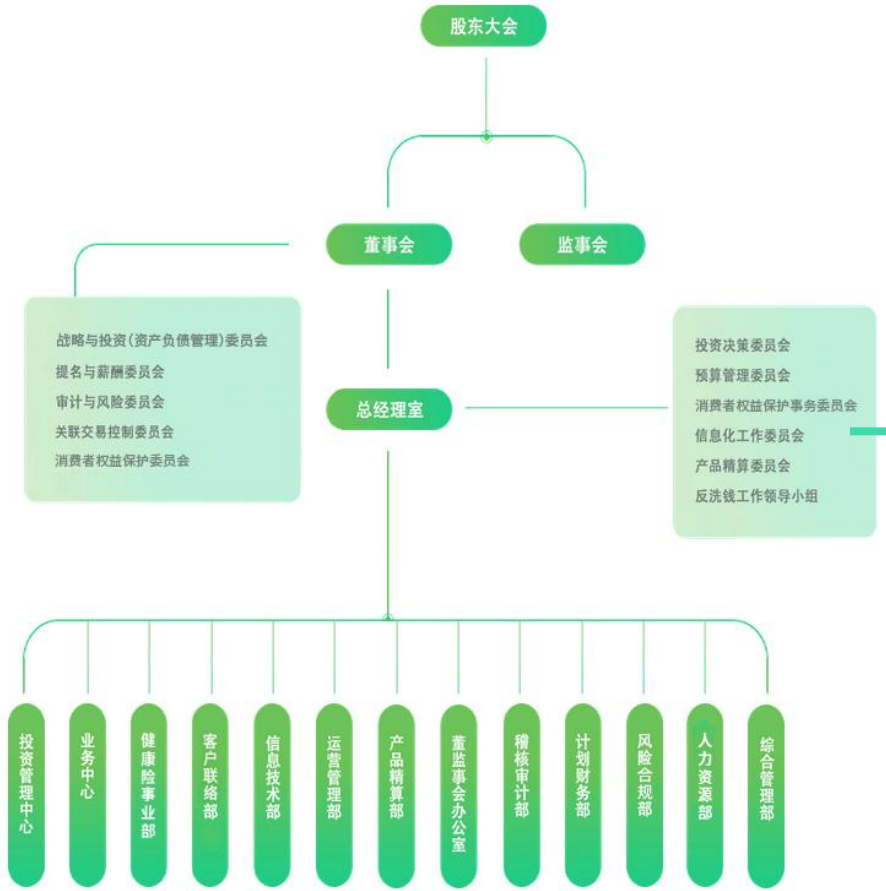
2.1 DCMM内容

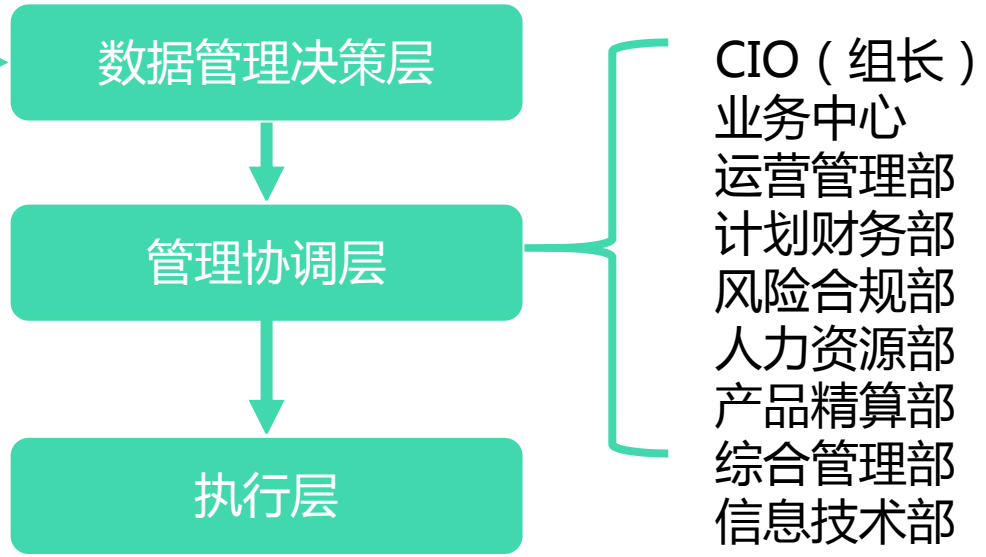
- 数据战略
- 数据治理
- 数据架构
- 数据应用
- 数据安全
- 数据治理
- 数据标准
- 数据生存周期

八个核心能力域和二十八个能力项



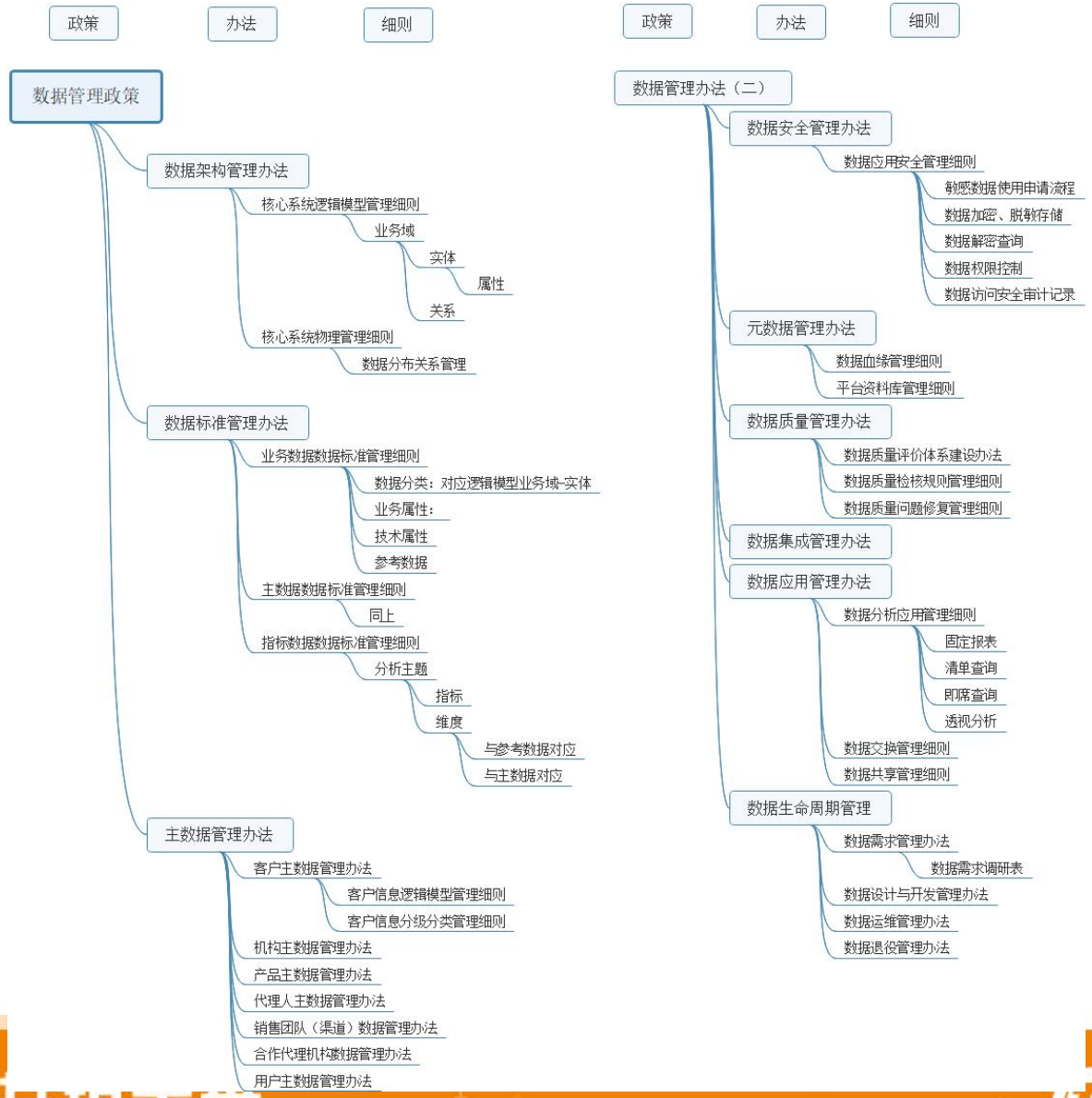
2.2 数据治理组织建设





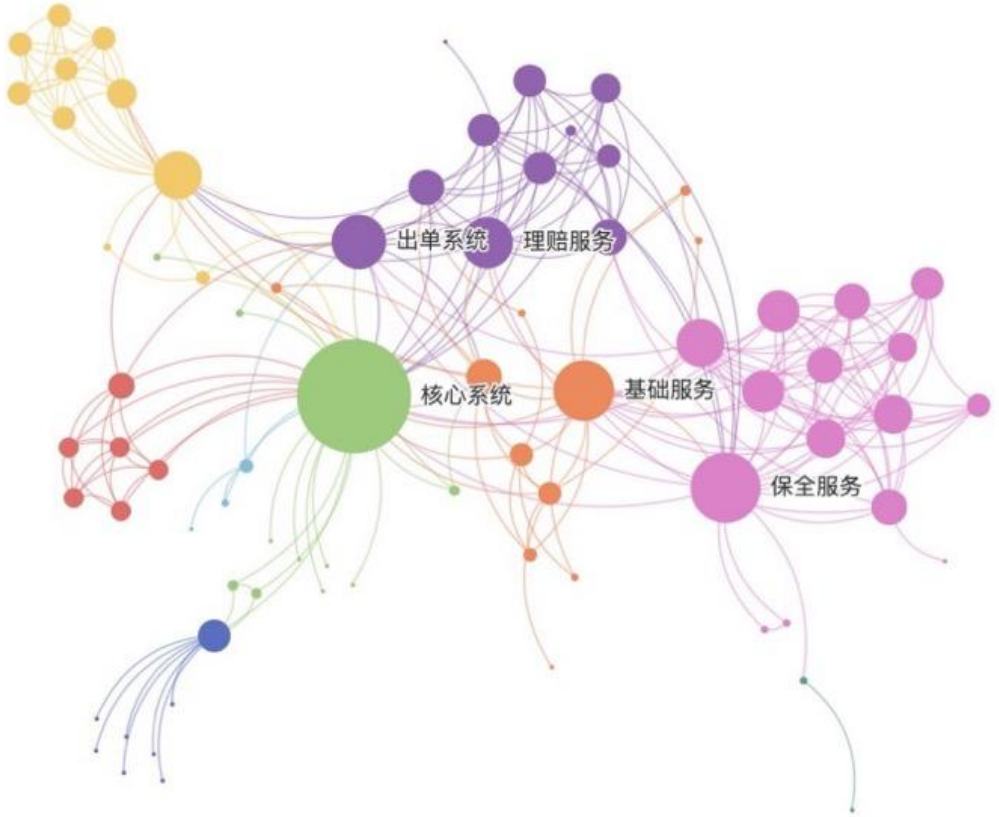
执行层由信息技术部负责人任组长、总公司各部门、各分支机构指派1人及信息技术部各二级部经理

2.3 数据治理的制度体系建设



2.4 数据管理之应用架构管理规范

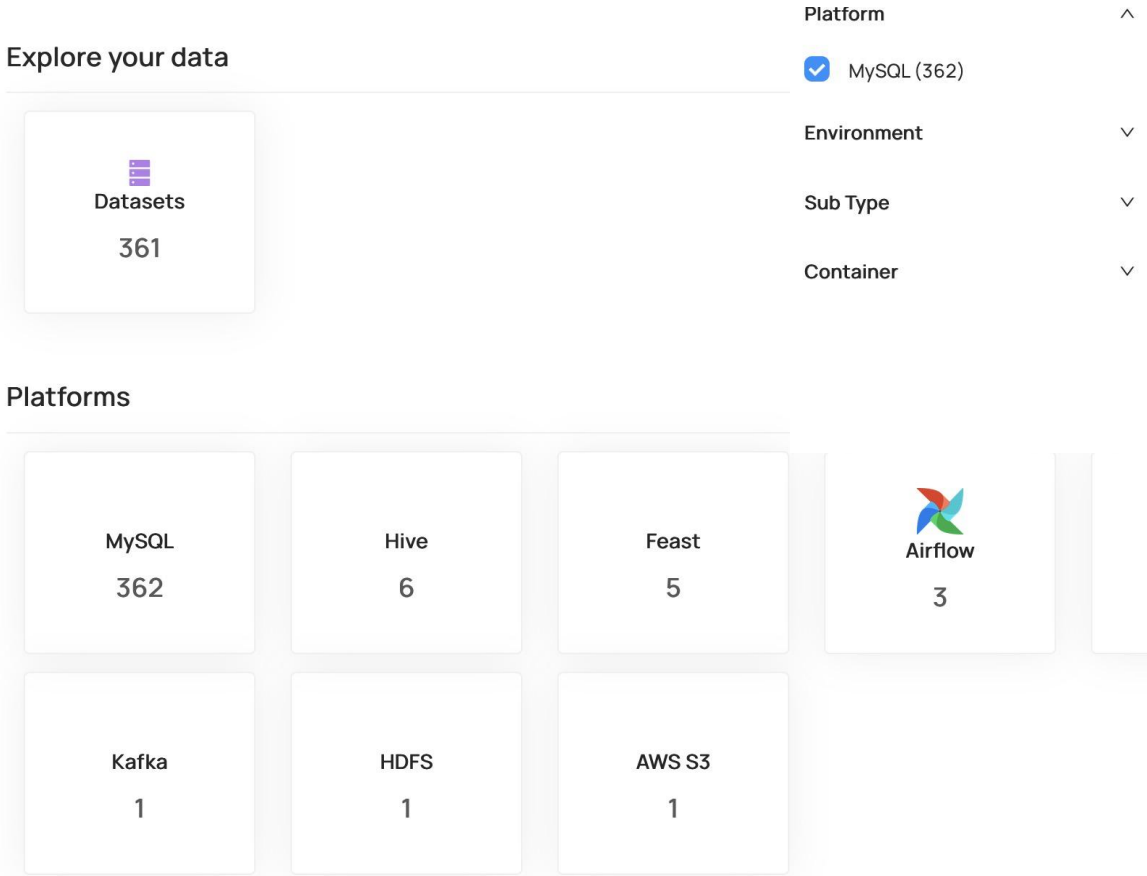
- 摸清家底
- 制定应用架构评审规范:应用间关系变更,API接口变化需走OA流程评审
- 建立应用架构管理平台, 记录系统间关系, 使用图数据库绘制系统间调用关系
- 按颜色分类标记系统间依赖关系
- 每季度出一次应用架构图



海保人寿2023年第二季度系统关系图

2.4 数据管理之模型管理

以核心主数据作为切入点，主数据的管理业务部门



1 系统建设形式多样

自建、采购成熟产品；

2 按照标准数据模型建设困难重重

3 大数据平台层进行数据收集整理

4 借助模型工具进行标准转换

PART 3

元数据管理工具

元数据管理的理解,元数据管理架构,元数据的来源,分类,元数据标签实践,数据生存周期

3.1 元数据管理的理解

- **技术角度**：元数据管理着企业的数据源系统、数据平台、数据仓库、数据模型、数据库、表、字段以及字段间的数据关系等技术元数据
- **业务角度**：元数据管理着企业的业务术语表、业务规则、质量规则、安全策略以及表的加工策略、表的生命周期信息等业务元数据。
- **应用角度**：元数据管理为数据提供了完整的加工处理全链路跟踪，方便数据的溯源和审计，这对于数据的合规使用越来越重要。通过数据血缘分析，追溯发生数据质量问题和其他错误的根本原因，并对更改后的元数据进行影响分析。

3.2 元数据分类

元数据类型	元数据	元数据描述	元数据实例
业务元数据	业务定义	数据的含义	客户的完整名称，具有法律效力
	业务规则	数据录入规则	企业的营业执照、组织机构代码证书、统一社会信用代码证书等具有法律效力的证明文件中的中文名称全称
	识别规则	识别规则	企业的组织机构代码、统一社会信用代码或者统一纳税号必须完全匹配，才能认为是同一客户
	质量规则	质量规则	客户名称为非空，并且与营业执照上的中文名称一致
技术元数据	存储位置	数据存储在什么地方	CRM 系统
	数据库表	存储数据的库表名称和路径	CRM/Customers
	字段类型	数据的技术类型	字符型
	字段长度	数据存储的最大长度	[200]
操作元数据	更新频率	数据的更新频率	每年更新一次
	管理部门	数据责任部门	客户管理部
	管理责任人	数据责任部门	客户管理部业务员

3.3 元数据-标签

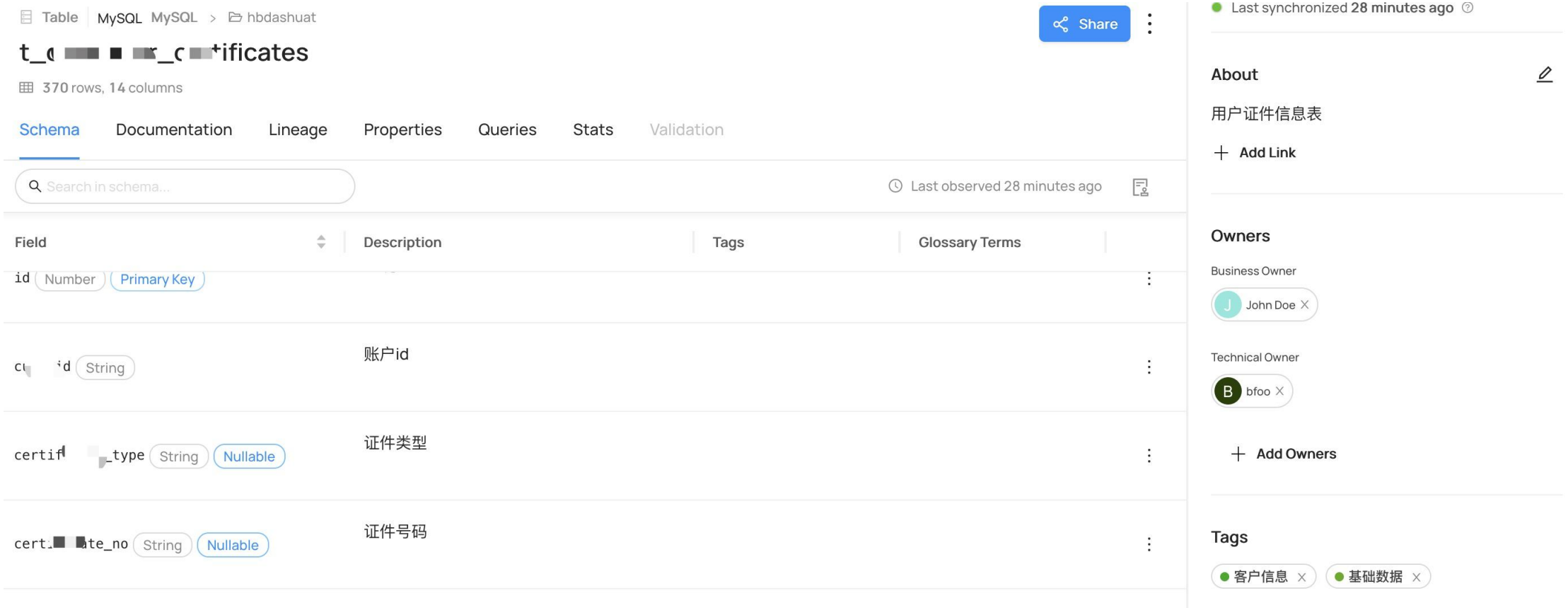


Table MySQL MySQL > hbdashuat
Share
370 rows, 14 columns
Schema
Documentation
Lineage
Properties
Queries
Stats
Validation
Search in schema...
Last observed 28 minutes ago
Field
Description
Tags
Glossary Terms
id Number Primary Key
String
账户id
_type String Nullable
证件类型
te_no String Nullable
证件号码
Last synchronized 28 minutes ago
About
用户证件信息表
Add Link
Owners
Business Owner
John Doe
Technical Owner
bfoo
Add Owners
Tags
客户信息
基础数据

3.4 数据生存周期

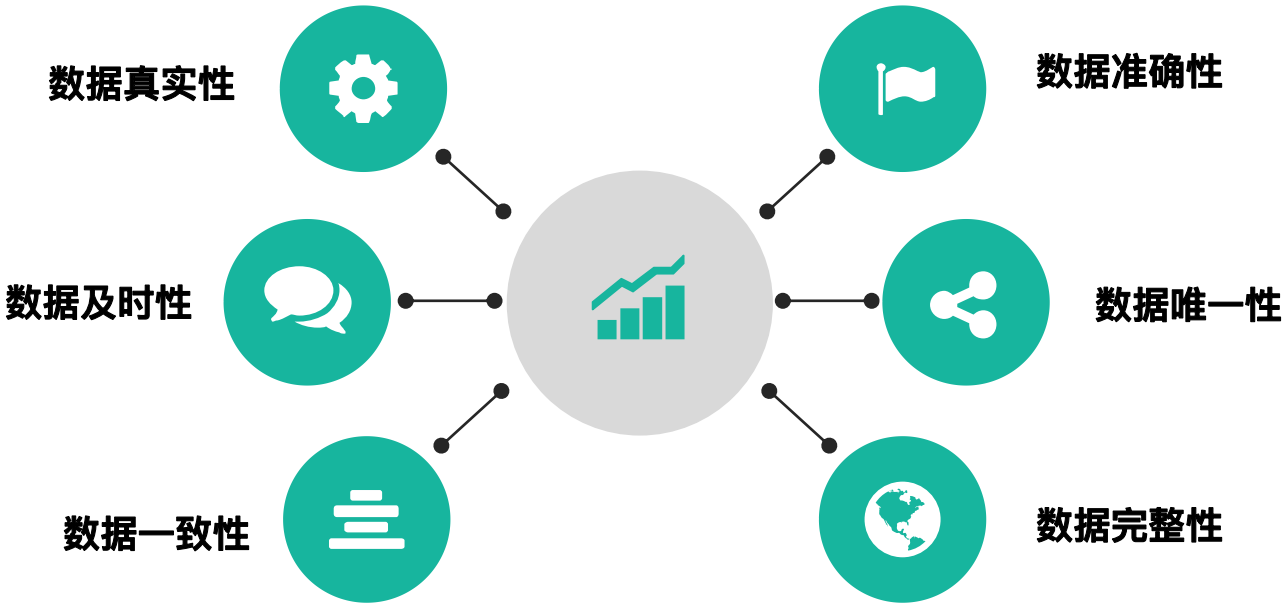
- 提升大数据平台的架构地位
  - 地位同核心业务系统
- 冷热数据存储机制
- 冷热数据迁移方案
  - 对保单的冷热数据迁移到大数据平台
  - 减轻核心系统的流量压力
  - 及短期的、失效的迁移，减轻核心业务系统改造压力

PART 4

数据质量检查

数据质量分类,数据质量应用架构,数据质量成果

4.1 数据质量分类



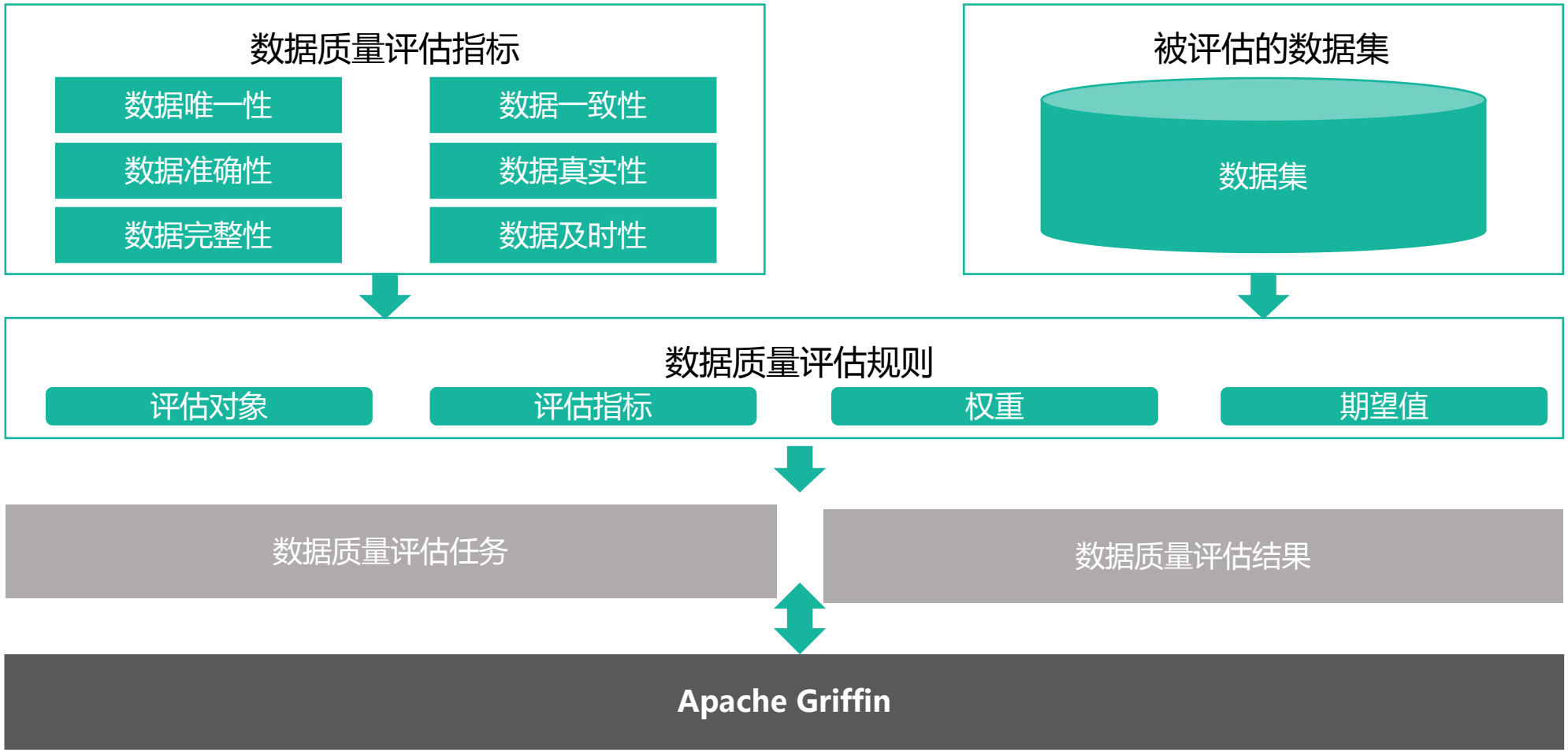
4.2 数据质量指标细则

从元数据管理平台

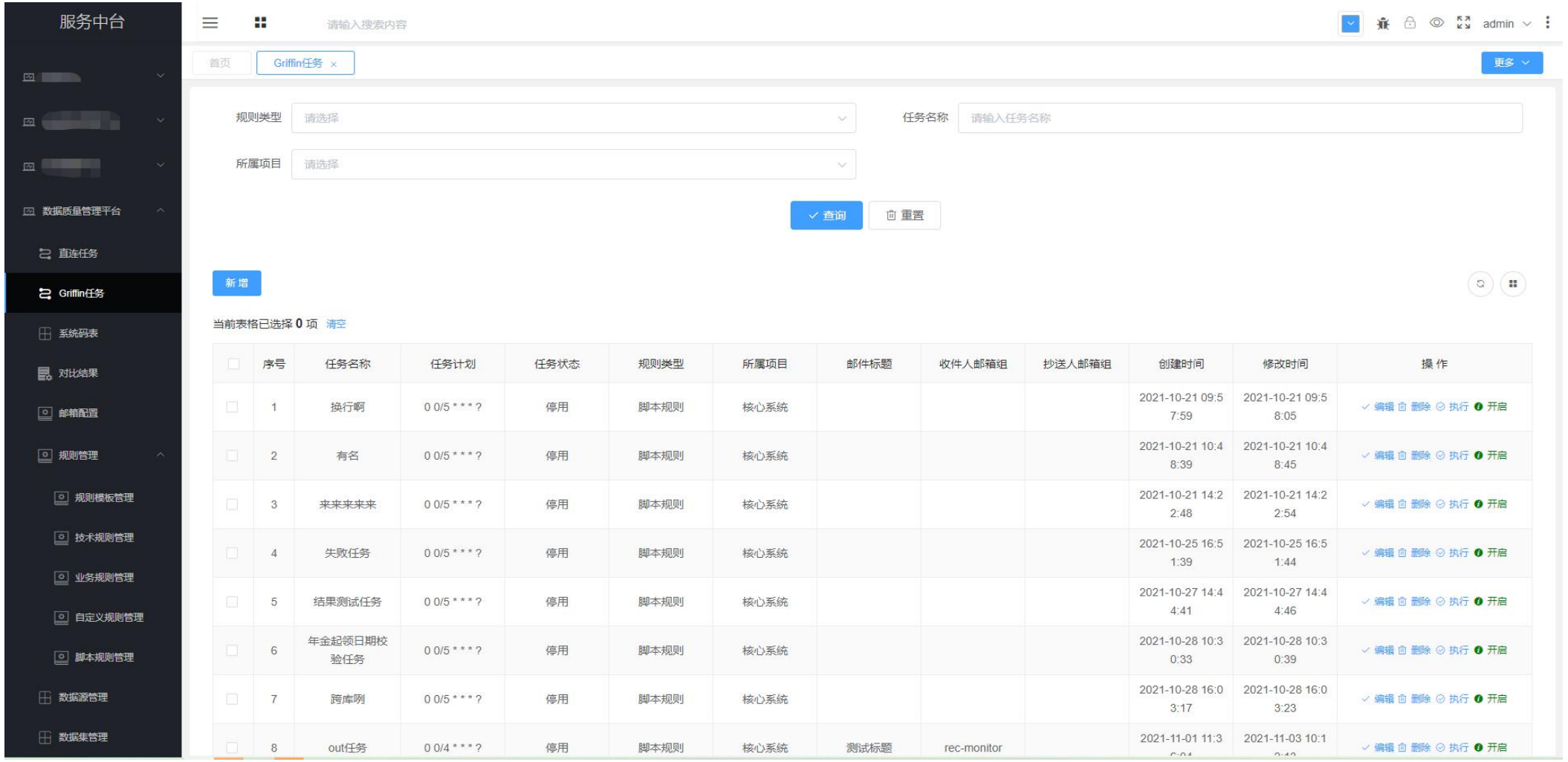
提取数据标准规则

序号	大类	指标类别	指标值	
1	准确性	主键唯一性	主键重复的记录数	主键重复的记录占比
2		非空检查	空值字段的记录数	空值字段的记录占比
3		类型检查	类型不符的记录数	类型不符的记录占比
4	规范性	长度检查	长度不符的记录数	长度不符的记录占比
5		格式检查	格式不符的记录数	格式不符的记录占比
6		值域检查	不满足要求的记录数	不满足要求的记录占比
7		精度检查	精度不符的记录数	精度不符的记录占比
8		字典码表检查	未通过检查的记录数	未通过检查的记录占比
9		自定义检查（例如：邮箱）	不符合规则的记录数	不符合规则的记录占比
10	完整性	外键约束	主键缺失的记录数	主键缺失的记录占比
11		自定义检查	不符合规则的记录数	不符合规则的记录占比
12	一致性	系统内信息的一致性	不符合系统内一致性的记录数	不符合系统内一致性的记录占比
13		系统间信息的一致性	不符合系统间一致性 的记录数	不符合系统间一致性 的记录占比
14		自定义检查	不符合规则的记录数	不符合规则的记录占比
15	及时性	数据传送的及时性	没有及时传送的表个数	没有及时传送的表占比
16		报表产生的及时性	没有及时产生的报表个数	没有及时产生的报表占比
17		自定义检查	不符合规则的记录数	不符合规则的记录占比

4.3 数据质量应用架构


数据质量评估指标
数据唯一性
数据一致性
数据准确性
数据真实性
数据完整性
数据及时性
被评估的数据集
数据集
数据质量评估规则
评估对象
评估指标
权重
期望值
数据质量评估任务
数据质量评估结果
Apache Griffin

4.4 数据质量成果展示



序号	任务名称	任务计划	任务状态	规则类型	所属项目	邮件标题	收件人邮箱组	抄送人邮箱组	创建时间	修改时间	操作
1	换行啊	0 0/5 * * * ?	停用	脚本规则	核心系统				2021-10-21 09:57:59	2021-10-21 09:58:05	编辑 删除 执行 开启
2	有名	0 0/5 * * * ?	停用	脚本规则	核心系统				2021-10-21 10:48:39	2021-10-21 10:48:45	编辑 删除 执行 开启
3	来来来来来	0 0/5 * * * ?	停用	脚本规则	核心系统				2021-10-21 14:22:48	2021-10-21 14:22:54	编辑 删除 执行 开启
4	失败任务	0 0/5 * * * ?	停用	脚本规则	核心系统				2021-10-25 16:51:39	2021-10-25 16:51:44	编辑 删除 执行 开启
5	结果测试任务	0 0/5 * * * ?	停用	脚本规则	核心系统				2021-10-27 14:44:41	2021-10-27 14:44:46	编辑 删除 执行 开启
6	年金起领日期校验任务	0 0/5 * * * ?	停用	脚本规则	核心系统				2021-10-28 10:30:33	2021-10-28 10:30:39	编辑 删除 执行 开启
7	跨库咧	0 0/5 * * * ?	停用	脚本规则	核心系统				2021-10-28 16:03:17	2021-10-28 16:03:23	编辑 删除 执行 开启
8	out任务	0 0/4 * * * ?	停用	脚本规则	核心系统	测试标题	rec-monitor		2021-11-01 11:3	2021-11-03 10:1	编辑 删除 执行 开启

PART 5

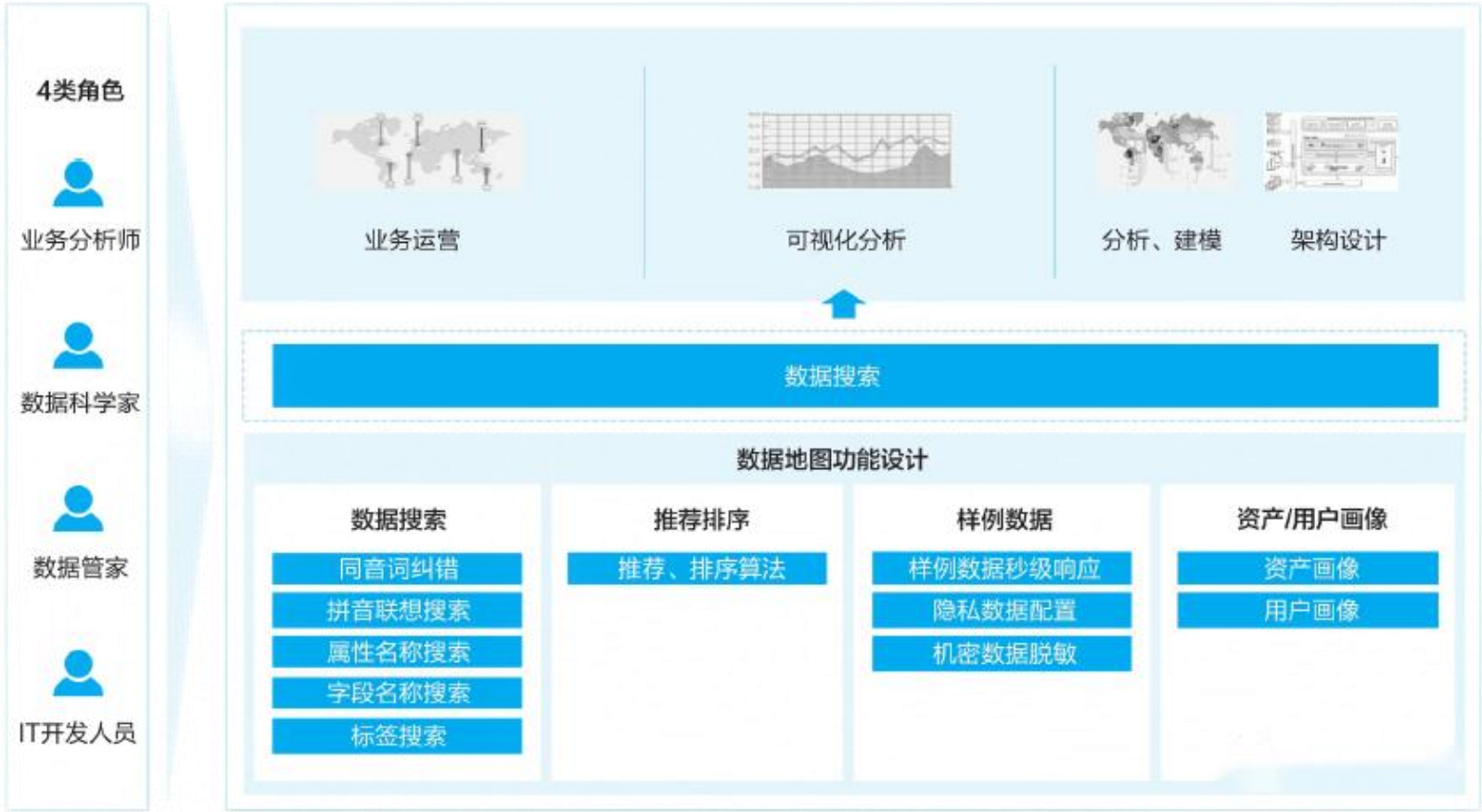
数据资产

数据资产之数据地图,数据血缘

5.1 数据地图的核心价值

- 数据供应者与消费者之间往往存在一种矛盾：供应者做了大量的数据治理工作、提供了大量的数据，但数据消费者却仍然不满意，他们始终认为在使用数据之前存在两个重大困难。
- 核心问题1 找数难
  - 企业的数据分散存储在上百个数据库、上万张物理表中，已纳入架构、经过质量、安全有效管理的数据资产也会超过上万个，并且还在持续增长中
- 核心问题2 读不懂
  - 企业往往会面对数据库物理层和业务层脱离的现状，数据的最终消费用户无法直接读懂物理层数据，无法确认数据是否能满足需求，只能寻求IT人员支持，经过大量转换和人工校验，才最终确认可消费的数据，而熟悉物理层结构的IT人员，并不是数据的最终消费者

5.2 数据资产之数据地图

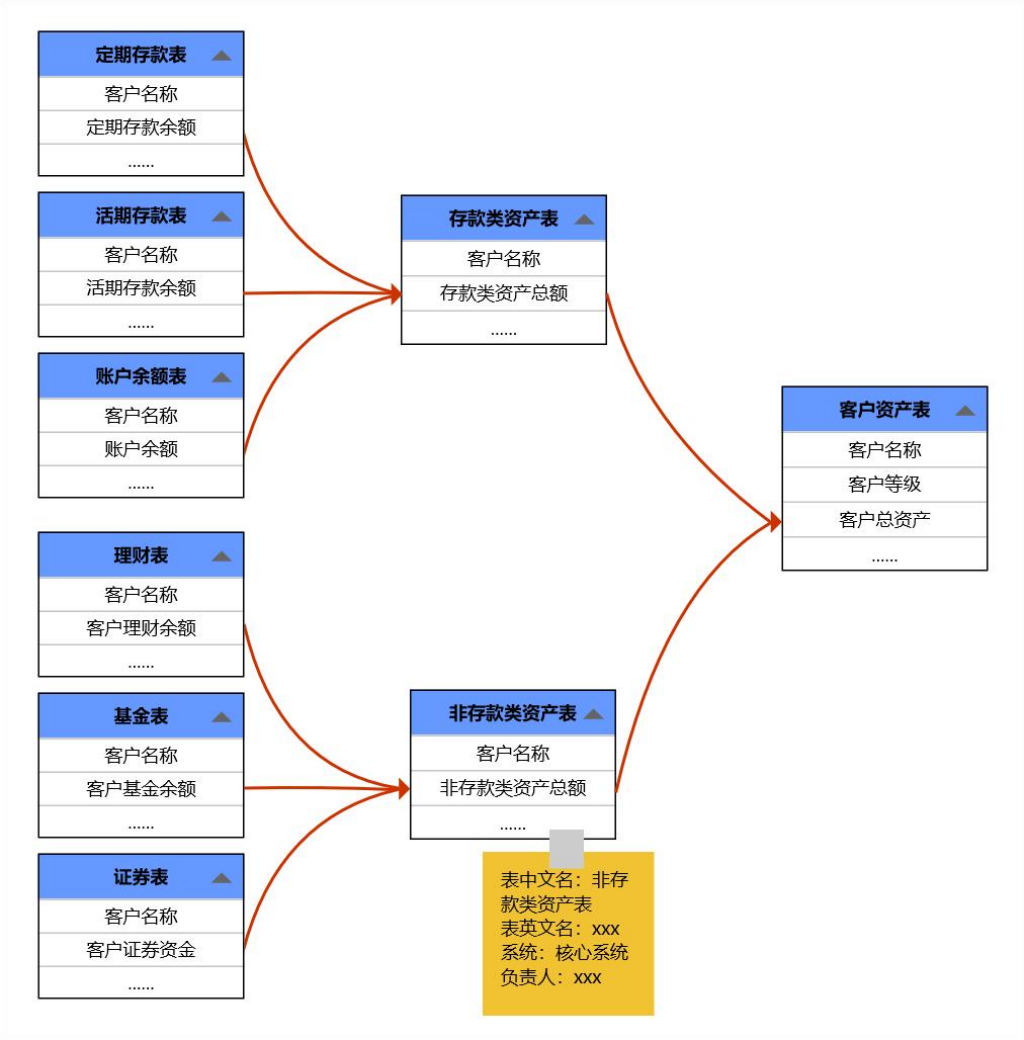


5.2 数据资产之数据血缘的价值

- 数据价值评估
  - 通过血缘分析和元数据可以从数据的集中度、分布、冗余度
    评估分析，从而初步判断数据的价值

- 数据质量评估
  - 判定数据的健康情况，是否存在大量的冗余数据、无效数据
    浪费等问题

5.3 数据血缘的可视化


定期存款表
客户名称
定期存款余额
.....
活期存款表
客户名称
活期存款余额
.....
账户余额表
客户名称
账户余额
.....
理财表
客户名称
客户理财余额
.....
基金表
客户名称
客户基金余额
.....
证券表
客户名称
客户证券资金
.....
存款类资产表
客户名称
存款类资产总额
.....
非存款类资产表
客户名称
非存款类资产总额
.....
客户资产表
客户名称
客户等级
客户总资产
.....
表中文名：非存款类资产表
表英文名：xxx
系统：核心系统
负责人：xxx

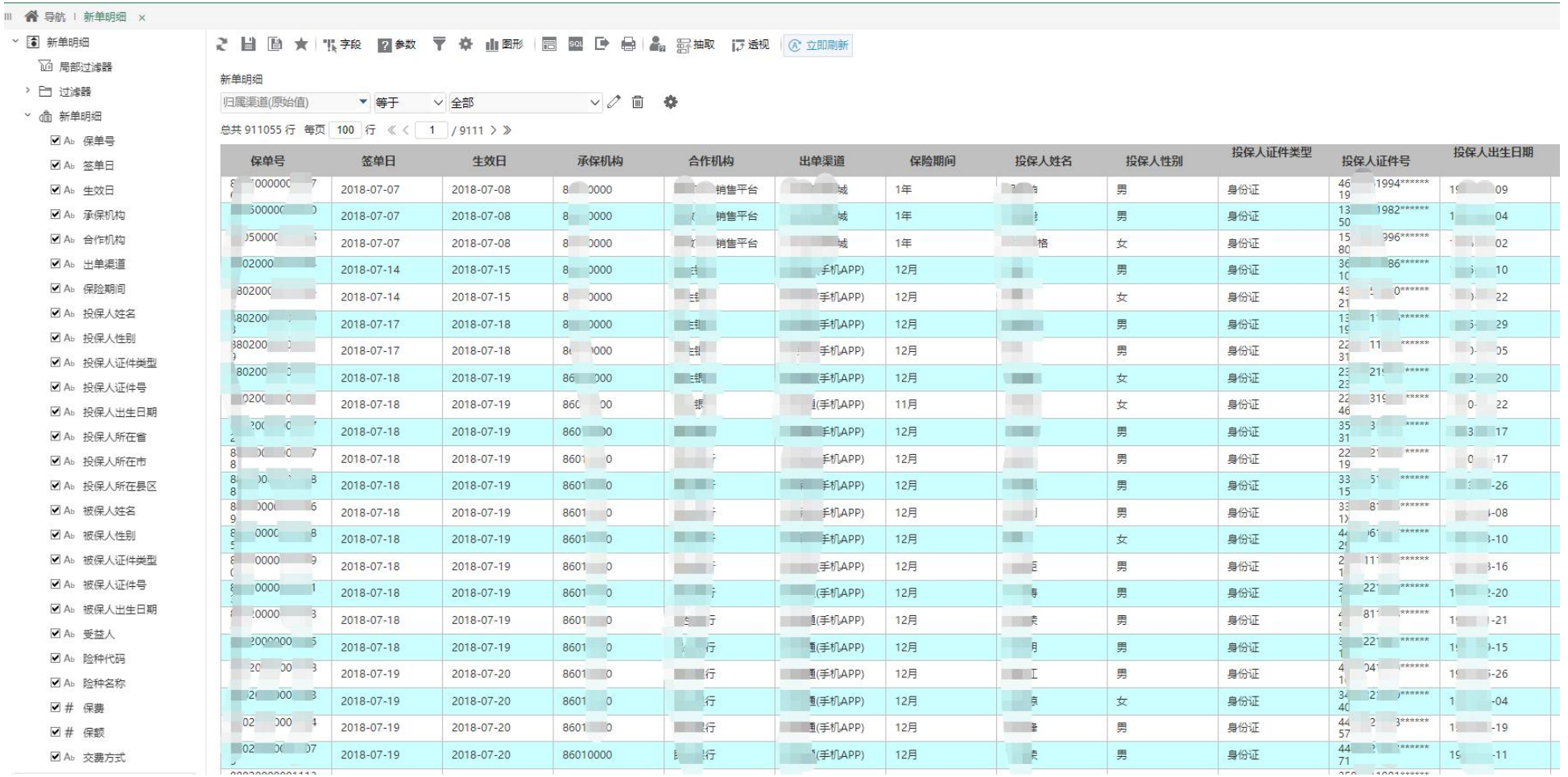
PART 6

数据应用

数据实验室,数据共享,数据交换,数据智能化探索

6.1 数据实验室之即席查询

- 例: 保单数据业务人员用即席查询功能自主选择字段展示
- 需求应变能力得到提升
- 安全得到提升:敏感字段脱敏、加密处理

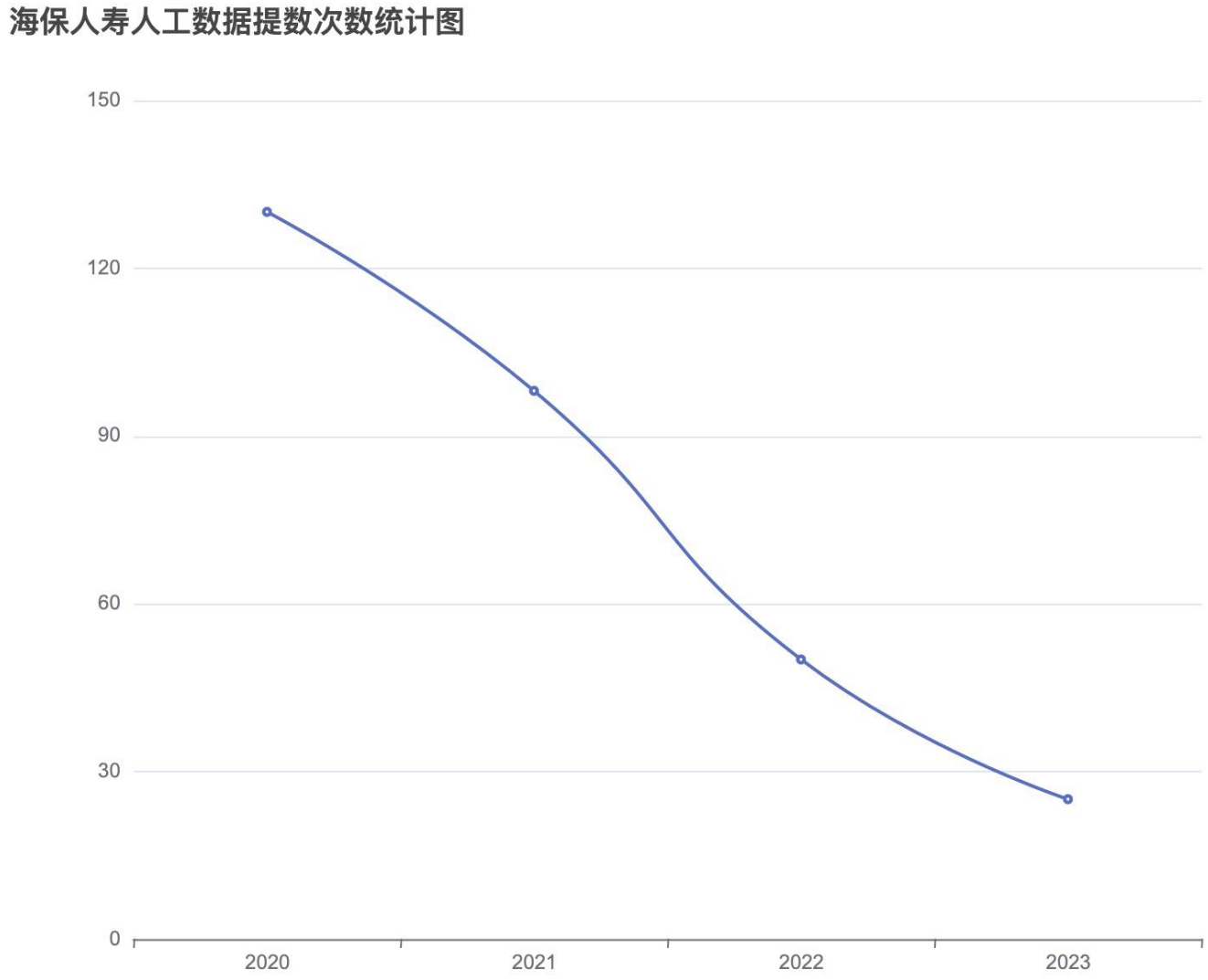


6.1 数据实验室之自助分析

- 例:对保险出险率的统计分析
- 增加水印,防止图片泄漏

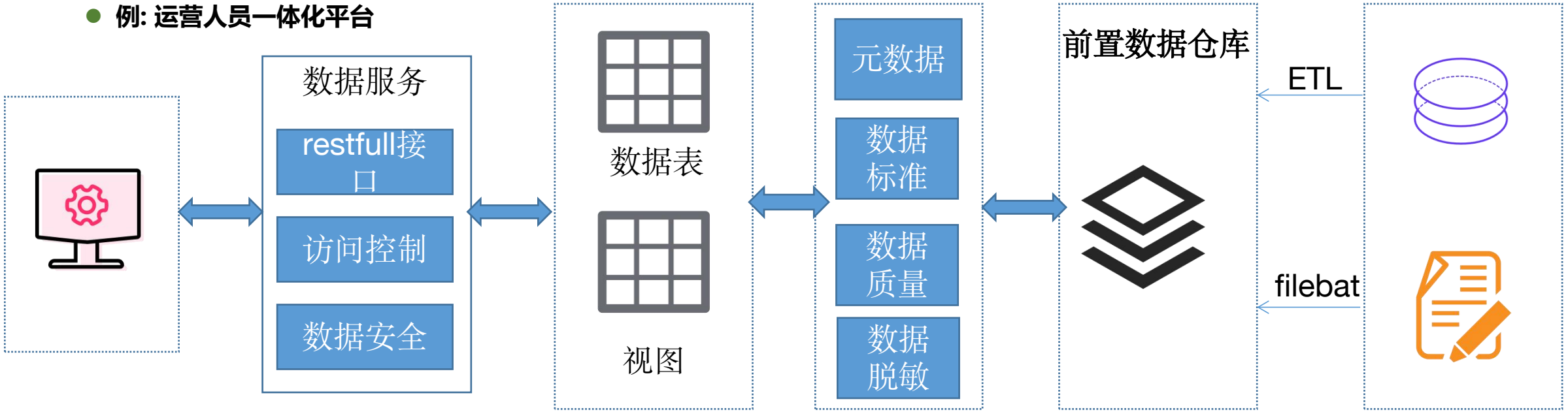
6.2 数据实验室之成果

- 提数效率显著提升,人工提数次数显著降低
- 数据提取集中在 审计、监管、业务手续费 数据核对业务场景



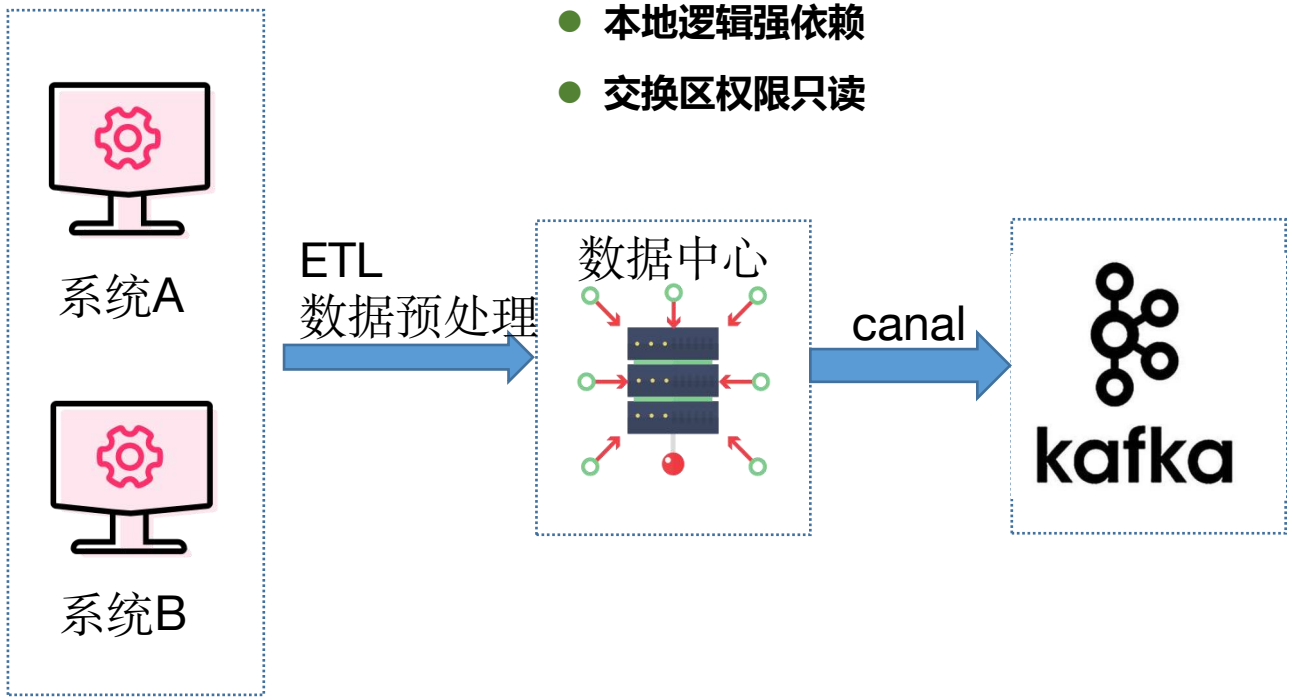
6.3 数据应用之数据共享

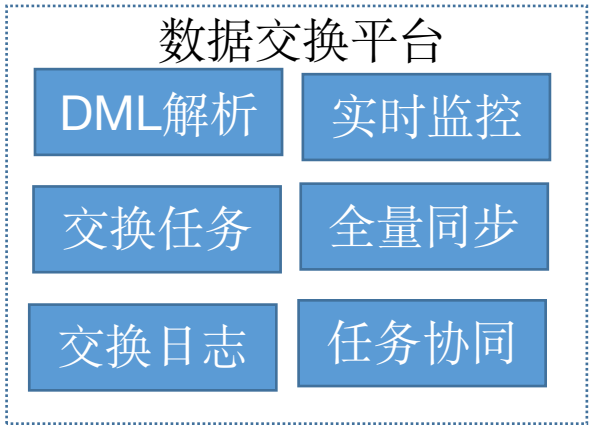
- 实时数据应用场景
- 支持数据分页
- 支持数据加密,数据脱敏
- 例: 运营人员一体化平台



6.3 数据应用之数据交换

- 适用于 非实时数据
- 本地逻辑强依赖
- 交换区权限只读





6.4 数据智能化探索-智能运维

- 借助人工智能技术对规则类数据语料训练出私有问答模型
- 为业务人员自动解答各种运维问题

6.4 数据智能化探索-非结构化数据的阅读理解

例如: 借助大模型的embending能力
理解保险产品文档
保险须知
保险条款等

```
Using embedded DuckDB with persistence: data will be stored in: db
Enter query (or 'exit' to quit): 保险责任有哪些
Query:保险责任有哪些
Answer:意外身故保险金，被保险人在认可的医疗机构因首次接受保险单上载明的择期手术时遭受手术意外、麻醉意外伤害而导致的手术过程中身故。

Enter query (or 'exit' to quit): 投保范围
Query:投保范围
Answer:被保险人范围为出生满30日（含）至70周岁（含），符合承保条件者均可作为被保险人参加本保险。投保人范围为被保险人本人或对被保险人有保险利益的其他人，符合承保条件者可作为投保人向保险公司投保本保险。

Enter query (or 'exit' to quit): 保险期间
Query:期间
Answer:保险期间为自被保险人实际接受保险单上载明的择期手术起至诊毕离开医院或者办妥住院手续当日二十四时止，保险期间最长为 10 日（门诊就医者）或自被保险人实际接受保险单上载明的择期手术起至办妥出院手续当日二十四时止，保险期间最长为 45日（住院治疗者）。

Enter query (or 'exit' to quit): exit
exit
(langchain) [root@VM-100-73-centos simple-chatpdf]#
```

追本溯源

```
(langchain) [root@VM-100-73-centos simple-chatpdf]# python qa.py
Enter query (or 'exit' to quit): 保险期间
[Document(page_content='3 保险金额 本合同的保险金额由您与我们在投保时约定，并在保险单或保险凭证上载明。 1.4 保险期间 门诊8就医者，除另有约定外，本合同的保险期间为自被保险人实际接受保险单上载明的择期手术起至诊毕离开医院或者办妥住院手续当日二十四时止，保险期间最长为 [illegible]日。 住院9治疗者，除另有约定外，本合同的保险期间为自被保险人实际接受保险单上载明的择期手术起至办妥出院手续当日二十四时止，保险期间最长为 45日。 \uf020 \uf08d\uf020我们不保障什么 这部分讲的是我们不承担保险责任的情况，请您仔细阅读。', metadata={'source': '119102.pdf'}), Document(page_content='我们在收到保险金给付申请书及有关证明和资料之日起 60日内， 对给付保险金的数额不能确定的，根据已有证明和资料可以确定的数额先予支付；我们最终确定给付保险金的数额后，将支付相应的差额。 3.7 诉讼时效 受益人及其他有权领取保险金的人向我们请求给付保险金的诉讼时效期间为 2年，自其知道或者应当知道保险事故发生之日起计算。 \uf078 如何退保 这部分讲的是您可以申请退保，但会有一部分损失。 4.1 您解除合同的手续及风险 在本保险合同成立后，除手术医院通知撤销手术的情况以外，您不得要求解除本合同。', metadata={'source': '119102.pdf'}), Document(page_content='3.6 保险金给付 我们在收到保险金给付申请书及合同约定的证明和资料后，将在 5 日内作出核定；情形复杂的，在 30 日内作出核定。 对属于保险责任的，我们在与受益人达成给付保险金的协议后 10 日内，履行给付保险金义务。 如果我们在收到保险金给付申请书及有关证明和资料后第 30 日仍未作出核定，对属于保险责任的，除支付保险金外，我们将赔偿受益人因此受到的损失。如果我们要求投保人、被保险人或者受益人补充提供有关证明和资料的，则上述30日不包括补充提供有关证明和资料的期间。 对不属于保险责任的，我们自作出核定之日起 3 日内向受益人发出拒绝给付保险金通知书并说明理由。', metadata={'source': '119102.pdf'}), Document(page_content='1 保险责任 在本合同有效期间内，本公司将承担如下保险责任： 意外身故保险金 被保险人在我们 认可的医疗机构1因首次接受保险单上载明的 择期手术2时遭受手术意外3、麻醉意外4伤害而导致的手术过程中身故， 我们按本合同约定的意外身故保险金额给付意外身故保险金，本合同终止。', metadata={'source': '119102.pdf'}), Document(page_content='6 未成年人身故保险金限制 为未成年子女投保的人身保险，因被保险人身故给付的保险金总和不得超过国务院保险监督管理机构规定的限额，身故给付的保险金额总和约定也不得超过海保人寿好安心手术意外伤害保险条款 第 7 页[共 7 页] 前述限额。 5.7 我们合同解除权的限制 前述规定的我们解除合同的权利，自我们知道有解除事由之日起，超过 30日不行使而消灭。 5.8 联系方式变更 为了保障您的合法权益，您的住所、通讯地址或电话等联系方式变更时，请及时以书面形式或双方认可的其他形式通知我们。', metadata={'source': '119102.pdf'})]
Query:保险期间
Answer:门诊就医者，保险期间为自被保险人实际接受保险单上载明的择期手术起至诊毕离开医院或者办妥住院手续当日二十四时止，保险期间最长为10日。住院治疗者，保险期间为自被保险人实际接受保险单上载明的择期手术起至办妥出院手续当日二十四时止，保险期间最长为45日。
```

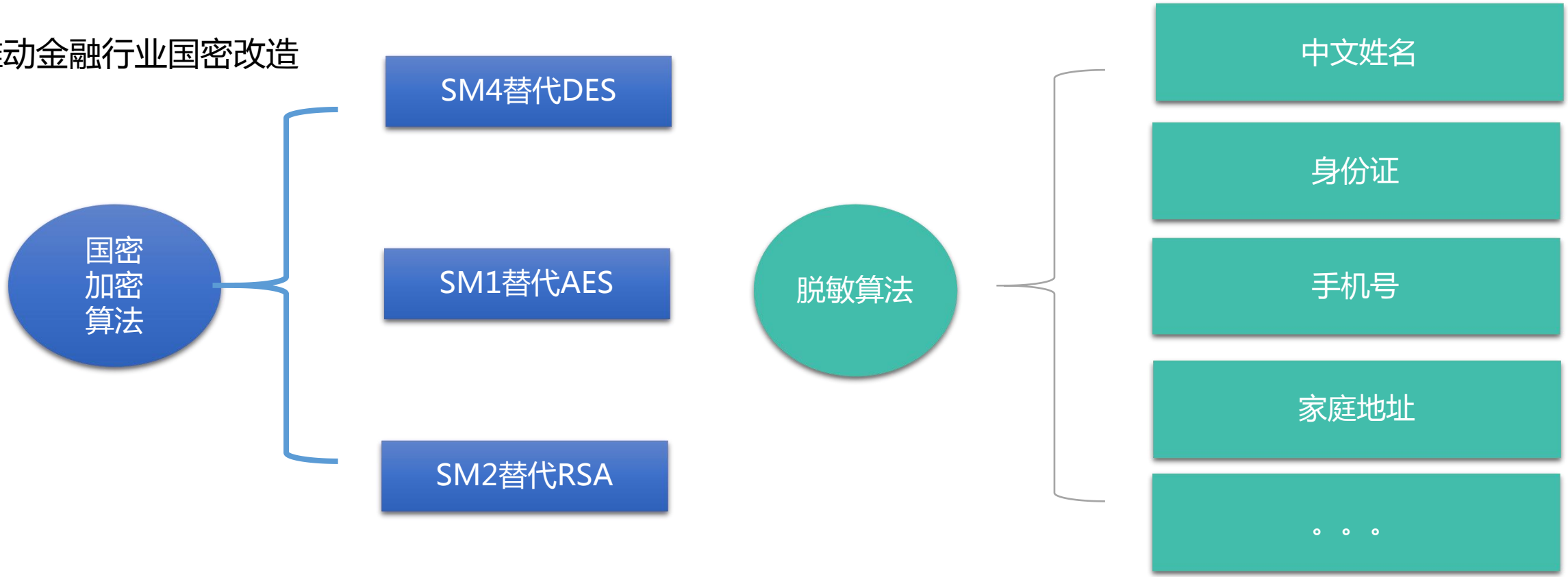
PART 7

数据安全

数据安全算法,数据安全问题,传输安全方案,数据存储安全

7.1 数据安全算法

推动金融行业国密改造



- ✓ 系统实现加密算法工厂设计模式，
- ✓ 通过工厂类加载不同的加密算法实现类来完成加密、解密处理；
- ✓ 后续可以通过扩展加密算法实现类来扩展新的加密、解密能力

- ✓ 系统实现脱敏算法工厂设计模式，
- ✓ 通过工厂类加载不同的脱敏算法实现类来完成脱敏处理；
- ✓ 后续可以通过扩展脱敏算法实现类来扩展新的脱敏能力

7.2 常见安全问题（一）

- API资产家底不清
  - 问题: 由于管理和技术原因，API接口作为非常重要的互联网暴露面，却长期处于安全监管盲区。在HVV行动期间，API应用便是一个重要突破口
  - 方法: 建立应用架构管理平台, 通过人工填报、主动探测、流量解析等方式发现API接口资产，并定期梳理和更新资产，实现API接口资产的全生命周期管理，快速发现未知资产，及时下线影子资产、僵尸资产，消除API接口安全监管盲区
- 接口防御能力不足
  - 问题: 由于设计缺陷或安全防御能力不足，对暴力破解、撞库攻击、注入攻击、遍历攻击、越权访问、未授权访问等网络攻击无能为力
  - 方法: API接口上线前或者接口业务调整后，开展安全风险评估，重点关注逻辑缺陷，找出安全漏洞和隐患，验证安全防御能力，减少API接口的数据互联网暴露面；

7.2 常见安全问题（二）

- 鉴权机制不够健全
  - 问题: 由于开发人员安全意识的缺失，API接口没有身份验证，或者前端硬编码方式实现，一旦突破身份认证这道防线后续取数便畅通无阻
  - 方法: 对API接口数据资产开展分类分级工作，明确重点保护对象和数据访问权限，为后续落实分级保护要求打下基础

- 接口返回数据过全
  - 问题: 由于开发人员安全意识薄弱，取数范围往往超过业务范围，接口返回数据字段过全/过多，增加敏感数据的互联网暴露面和数据泄露的风险
  - 方法: 遵循最小必要原则，根据用户访问权限，通过访问源、访问方式、访问总量、访问频次、访问时段、访问周期、访问时长等纬度强化访问控制，健全访问授权机制，约束，接口访问行为，防范API接口滥用和数据爬取

7.2 常见安全问题（三）

- API接口滥用盗用：
  - 问题: 用数部门为了贪图日后用数便利性，超出业务范围违规批量拉取数据，甚至开发网络爬虫持续读取并本地存储；不法分子窃取或第三方私存API接口认证账号，非法读取API接口数据，造成数据的泄露；
  - 方法: 记录和存储API接口访问日志，开展API接口访问常态化监测，从访问源、访问方式、访问总量、访问频次、访问时段、访问周期、访问时长等纬度借助威胁建模技术建立安全基线

- API接口二次流转
  - 数据使用部门对上一级共享数据交换平台开放的API接口进行二次封装，再对外开放以满足下级单位/第三方的用数需求，数据流转和使用范围肆意扩大和蔓延；
  - 方法1: 借助脱敏技术对敏感数据实施动态脱敏，避免敏感信息泄露,
  - 方法2: 在接口返回数据字段中植入数据水印，当发生数据泄露事件后，可通过数据水印进行版权宣示和溯源追责。

7.4 数据存储安全

背景：等保三级要求在存储层对敏感数据进行加密。

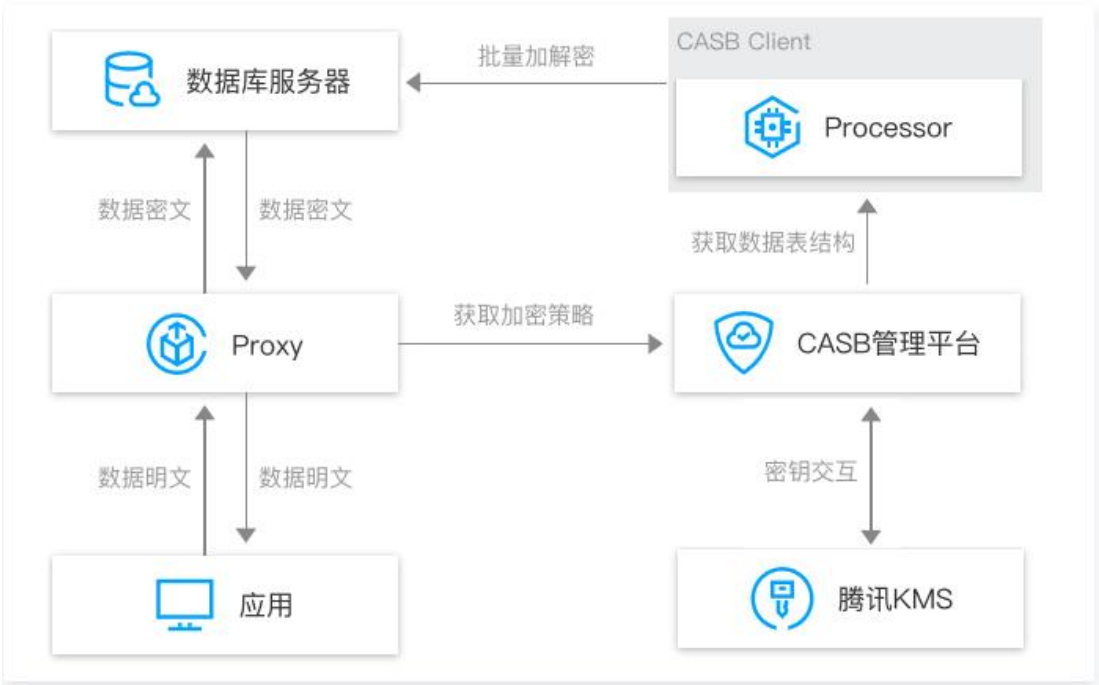
存储层加密的主要作用：防止备份或磁盘丢失引起敏感数据泄露。

存储层加密有两种：物理加密和逻辑加密：

- 物理加密：对物理磁盘存储空间进行加密，现有TDSQL实例存储层物理加密腾讯云不支持；
- 逻辑加密：对敏感字段进行加密。

CASB 简介

- CASB是腾讯云开发，提供面向服务侧的字段级数据存储加密防护，有效防护内外部数据安全威胁；
- 加密字段和加密策略在CASB管理平台进行管理，proxy获取加密码策略后，负责加解密工作；



数据治理-路漫漫而修远

- 不能 一蹴而就 而要 循序渐进
- 不只 三分激情 更需 持之以恒
- 不只有技术也有管理

THANKS