

## 数据是企业的重要资产

- 公司内各条线的业务人员在业务开展的过程中往往会面临大量的、来自不同数据源的、异构的数据，如何有效管理和使用这些企业未来最重要的资产经常成为数据管理者和使用者的一大难题
- 典型的数据源包括：公司内部数据、政府平台数据、征信数据、银联数据、第三方数据供应商数据、社交网络数据、埋点数据等等



如何管理和使用数据



- 德勤全球数据中心（GDC）是一个专注于为客户提供（风险）数据的专业团队，目前设立在重庆
- 依托德勤丰富的风险管理项目经验以及对客户一线需求的深入理解，该团队搭建了德勤风险数据平台提供风险数据服务

## 数据来源

### 外部公开数据

- 针对外部数据的公开性，通过开发网络爬虫获取各监管单位、行业协会、类权威财经网等众多网站发布的数据
- 通过外包形式，人工收集部分逻辑复杂的定性/非结构化数据

### 第三方数据

- 通过API接口等方式接入第三方，针对性的获取第三方数据
- 通过付费方式购买第三方数据

### 有权机关数据

- 通过对接工商局、法院、教育部等政府平台，获取权威数据

### 德勤内部数据

- 德勤拥有丰富的风险管理项目经验，通过项目实施，在德勤内部数据库中积累了大量脱敏后的数据

## 数据维度

通过对原始数据进行清洗整合，根据不同的维度体系建立不同的数据仓库。  
以企业数据为例，主要数据维度如下：

### 25+行业

- 11+类金融行业
- 14类非金融行业

### 120+定性指标

- 包含企业基本信息与历史沿革(企业年限、股权结构与变更等)、经营模式(技术优势、主营业务等)、上下游企业(稳定性、集中度等)等定性指标

### 270+定量指标

- 涵盖规模类、杠杆比率、流动性、盈利能力、运营能力、成长性等6大类270+定量指标

## 数据仓库

### 宏观数据

- 全国；32个省、市、自治区、直辖市；400+地级市；2000+县
- 1000+指标

### 行业数据

- 集成自企业数据的业务逻辑，进行行业整合与分析，为各类企业定位提供支持

### 企业数据

- 涵盖所有银行、证券等金融企业
- 所有上市公司、发债企业
- 其他企业

### 外部数据

- 涵盖基本背景信息、舆情风险、监管诉讼等各类外部公开数据

## 数据服务

### 项目+数据

- 在项目实施过程中和后续服务期内，以定期推送的方式为客户提供相关数据

### 数据个性定制

- 根据客户需求，为其量身定制成套数据模板，进行定期更新推送

### 数据接口

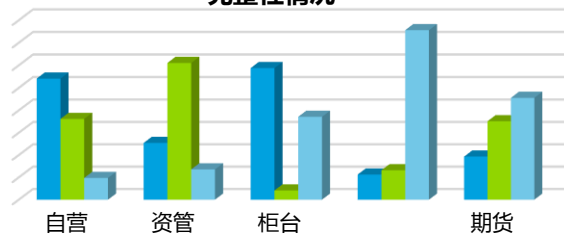
- 客户可通过实时/批量数据接口接入德勤风险数据平台进行数据查询与获取

### 数据订阅

- 客户可根据自身业务需求，订阅评级、预警、指数等资讯服务，德勤将第一时间为客户推送相关讯息

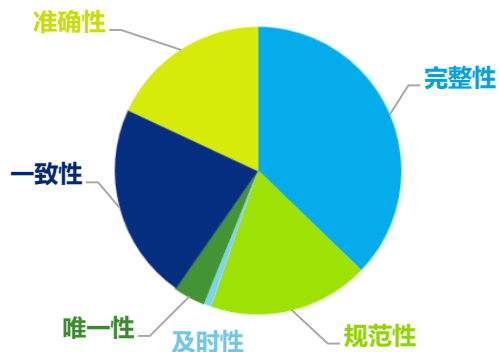
## 数据质量问题严重制约数据价值发挥

完整性情况

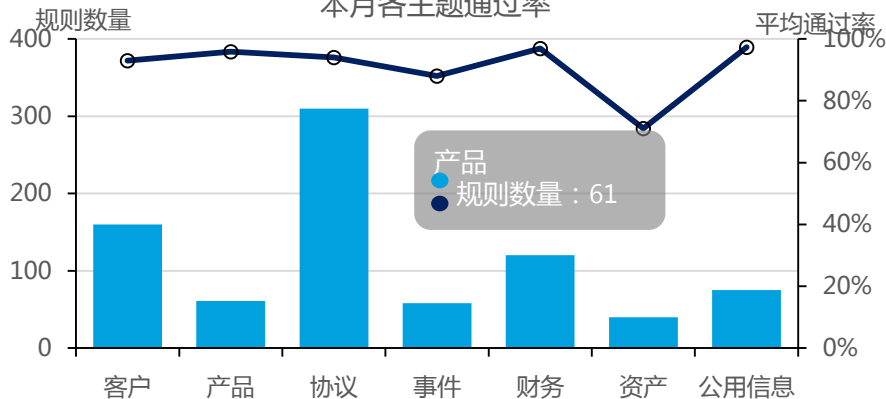


柜台	十万条数据中存在941条空值，其余全为数字0。分段结束阶段为空时，分段开始阶段同样为空，但是分段结束日期不为空
柜台	十万条数据中存在941条空值，其余全为数字0。分段结束阶段为空时，分段开始阶段同样为空，但是分段结束日期不为空
CRM	十万条数据中存在61779条数据为空，且在代理人标识不为空的情况下，仍有61540条数据为空
CRM	十万条数据中存在236条数据为空，该列为空时代理人证件类型、代理人姓名存在不为空的情况
柜台	51918条数据中存在7102条数据为空，且在登记人不为空的情况下存在6881条数据为空的情况
柜台	十万条数据中存在2382条数据为空，由于存在基准利率非空但是基准利率类型为空的情况，可认为基准利率类型填写不完善
柜台	十万条数据中存在99990条数据为空，保证金金额不为空的情况下，仍然有账号为空
柜台	十万条数据中存在17962条数据为空，涉及第三方1不为空时存在本字段为空的情况
自营	十万条数据中存在18291条数据为空，涉及第三方2不为空时存在本字段为空的情况
自营	十万条数据中存在99996条数据为空，涉及第三方3不为空时存在本字段为空的情况
自营	十万条数据中存在93998条数据为空，存在还款账号非空但是还款账户名为空的现象
自营	十万条数据中存在94026条数据为空，存在还款账户名非空但是还款账号为空的现象
资管	十万条数据中存在61779条数据为空，存在代理人证件类型标识不为空，但是代理人姓名为空的现象
资管	十万条数据中存在61896条数据为空，存在代理人姓名不为空，但是证件号为空的现象
期货	346数据中存在337条数据为空，存在联系人不为空，但是证件号码为空的情况
期货	十万条数据中存在97934条数据为空，存在姓名为空但是电话号码不为空的现象

数据质量问题各维度



本月各主题通过率



## GDC建设过程中曾遇到的各类数据问题

### 数据问题

#### 数据对接人员缺乏

各项目团队各自为战，重复工作

#### 时效性差

数据未能在第一时间及时更新

#### 样本缺失

使用过程中发现缺乏一些关键样本

#### 数据值缺失

一条记录里可能含有缺失值

#### 数据文件损坏

保存或处理方式不当，导致数据文件损坏

#### 数据文件遗失

电脑系统崩溃、遗失、操作不当等，导致文件遗失

#### 数据重复

相同或者部分相同的记录出现多条

#### 数据异常

#### 数据错误

数据没有严格按照规范输入，导致错误

#### 数据差异

定性数据的录入存在主观性差异

#### 数据无效

数据完整，但因格式等问题不可用

#### 数据口径不统一

数据统计口径存在差异，如财务数据，万元/元等单位不一

#### 数据处理技术落后

传统工具无法处理

.....

.....

### 原因

- 没有专业的数据工作**人员**及**团队**
- 没有健全的数据**样本**和数据**监测**机制
- 缺乏**异常值处理**机制
- 没有完善的处理机制、**备份**机制
- 数据**清洗**机制不健全
- 没有严谨的数据**校验**机制
- 人工数据**录入错误**
- 数据录入**不规范**
- 没有建立标准的**指标体系**
- 缺乏**专业技术**应对大数据时代的海量数据

### 解决方式

#### 组建数据团队

- 创建了重庆GDC数据中心，组建了截至目前数十人的专的数据团队

#### 建立健全的数据机制

- 以爬虫+API的建立监测机制实时监测数据动态
- 建立完善的数据流引擎进行数据清洗、数据校验、异常值处理
- 建立标准的数据库以及数据备份机制

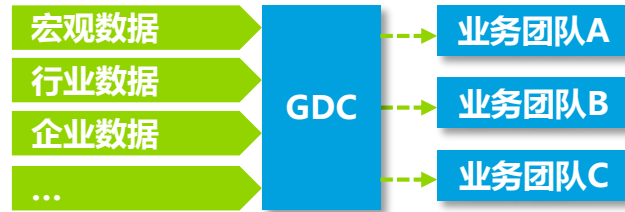
#### 规范数据标准

- 根据业务需求和规范，建立标准的指标体系
- 规范数据命名、类型、质量标准
- 搭建数据补录平台，规范数据录入流程及标准

#### 技术革新

- 结合R、SAS、Spark等专业数据处理软件
- 引进数据挖掘与机器学习算法
- 云服务器、分布式、并行等大数据解决方案
- .....

## 数据治理案例



### 人员、组织与架构

- 过去各业务团队独立收取所需数据，易出现重复收集以及资源不足的情况

- 通过建立GDC大数据中心对数据进行统一收集，再供数给各业务团队，实现共享化与专业化

### 标准、制度与规范

- 过去各业务团队自行收取数据导致数据标准不统一，数据处理整合难度大

- 通过建立GDC大数据中心，建立统一的各类数据标准与规范，提升数据管理效率

## 数据治理案例

序号	单位名称	负责人	年份	报告名称	产品多样性	服务区域多样性	应收账款多样性	应收账款账龄	企业信用评级
1	中国工商银行股份有限公司	王冠	2014	报告	2000	2000	2000	2000	2000
2	中国农业银行股份有限公司	王冠	2014	报告	2000	2000	2000	2000	2000
3	中国银行股份有限公司	王冠	2014	报告	2000	2000	2000	2000	2000
4	交通银行股份有限公司	王冠	2014	报告	2000	2000	2000	2000	2000
5	中国建设银行股份有限公司	王冠	2014	报告	2000	2000	2000	2000	2000



房地产补录模板



基础设施补录模板



建筑补录模板



交通运输补录模板



贸易补录模板



能源补录模板



轻工业补录模板



寿险补录模板



消费补录模板



银行补录模板

## 流程、活动与机制

- 根据项目需求清单整理相关报告，进而进行数据补录

数据补录平台 v2.3

我的首页 > 统计管理 > 补录统计

序号	机构名称	参与统计机构数	总记录数	已提交记录数	完成率
1	***分行	28	2011	1701	84.58%
2	***分行	29	1989	1569	78.88%
3	***分行	69	1830	1406	76.83%
4	***分行	51	1329	985	74.12%
5	***分行	22	1320	911	70.53%
6	***分行	41	1960	1308	66.73%
7	***分行	96	2641	1733	65.62%
8	***分行	37	2331	1399	60.02%
9	***分行	51	1721	954	55.43%
合计		424	17132	11808	68.96%

- GDC大数据中心集中进行数据自动化补录，通过数据补录平台，建立了完整的样本及数据监测机制、数据补录触发机制、数据清洗机制、数据校检机制，流程简洁高效

## 技术、平台与工具

- 采用人工的方式对数据情况进行搜索、下载和分析，耗时耗力
- 通过数据补录平台，实现数据的自动化监测、获取、存储以及初步分析

- 数据治理是成功的企业数据管理中不可或缺的重要组成

企业数据管理的4个组成部分



## 数据应用

- 基于可靠的信息行动决策
- 决策流程优化
- 预测与前瞻性分析
- 如：以客户为中心的产品研发，营销战略策略与执行，服新务开发等。

## 数据模型与分析模型

- 描述性分析
- 360度企业全景视图
- 商业智能应用程序
- 管理仪表盘，报告
- 如：客户洞察，客户统一视图

具体做法

## 数据治理

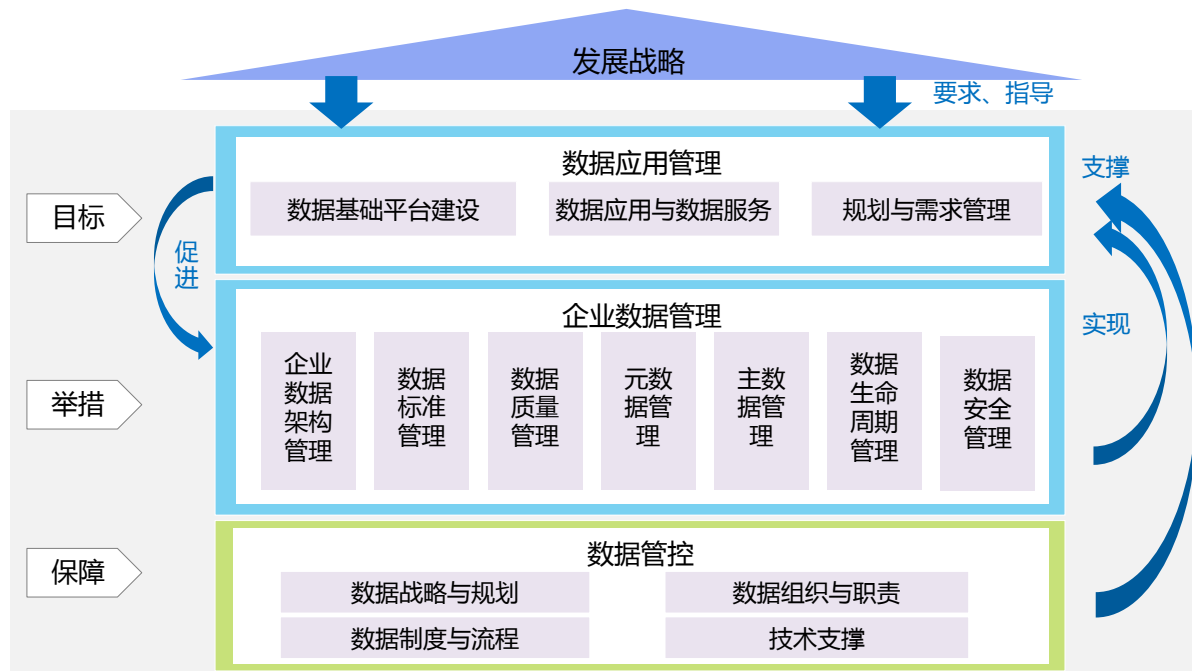
- 基于生命周期的数据管理
- 数据人员、组织与架构
- 数据流程、活动与机制
- 数据标准、制度与规范
- 数据技术、平台与工具

## 大数据基础架构

- 数据概要
- 数据清洗
- 数据整合

- 明确数据治理职责分工，明确不同部门在数据生命周期各个阶段的具体职责
- 建立数据治理的规则制度及流程，详细指导数据治理工作的开展
- 建立数据标准，建立统一的数据规范，统一的指标计算规则与逻辑
- 建立数据质量管理端到端的闭环管理机制，做到事前防范，事中控制，事后治理相结合，提升数据质量，提升数据应用的可靠性
- 结合内外部数据，力求发挥最大数据价值

### 证券行业数据治理工作框架



- 以证券公司发展战略为导向
- 围绕证券公司数据的生命周期
- 从数据管理和服务的整体角度出发
- 描述券商数据各项功能和活动



各领域工作内容分解



### 德勤在广发证券数据治理项目中的工作内容



工作方法

与流程

1 事实依据收集

重点部门访谈

问卷调查

现有资料文件整理



2 成熟度评估

工作现状

能力差距

关键问题

数据需求



3 实施路线图设计

目标设定

任务识别

项目优先级排序

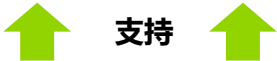

实施路线图



4 组织架构、制度流程设计

数据管理组织架构

数据管理流程、制度



支持



德勤数据治理体系模型



成熟度五级模型



同业优秀实践

5 数据治理体系实施评估(回访阶段)

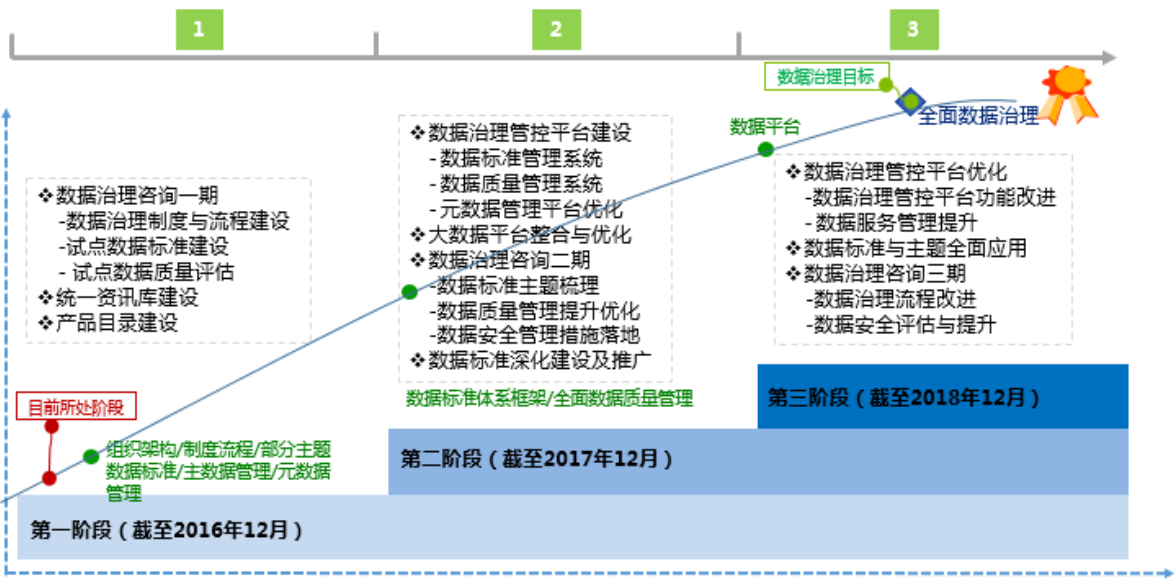
实施成效评估

评估分类	评估内容	评估标准
数据治理组织保障	数据治理组织架构是否健全	数据治理组织架构是否健全
数据治理制度保障	数据治理制度是否健全	数据治理制度是否健全
数据治理流程保障	数据治理流程是否健全	数据治理流程是否健全
数据治理技术保障	数据治理技术是否健全	数据治理技术是否健全
数据治理人才保障	数据治理人才是否健全	数据治理人才是否健全
数据治理文化保障	数据治理文化是否健全	数据治理文化是否健全
数据治理资金保障	数据治理资金是否健全	数据治理资金是否健全
数据治理风险保障	数据治理风险是否健全	数据治理风险是否健全
数据治理安全保障	数据治理安全是否健全	数据治理安全是否健全
数据治理隐私保障	数据治理隐私是否健全	数据治理隐私是否健全
数据治理合规保障	数据治理合规是否健全	数据治理合规是否健全
数据治理伦理保障	数据治理伦理是否健全	数据治理伦理是否健全
数据治理社会责任保障	数据治理社会责任是否健全	数据治理社会责任是否健全

数据治理实施路线图设计

通过数据治理工作，促进数据质量的标准化，实现数据的全面管控：

数据治理实施路线图设计：



阶段目标

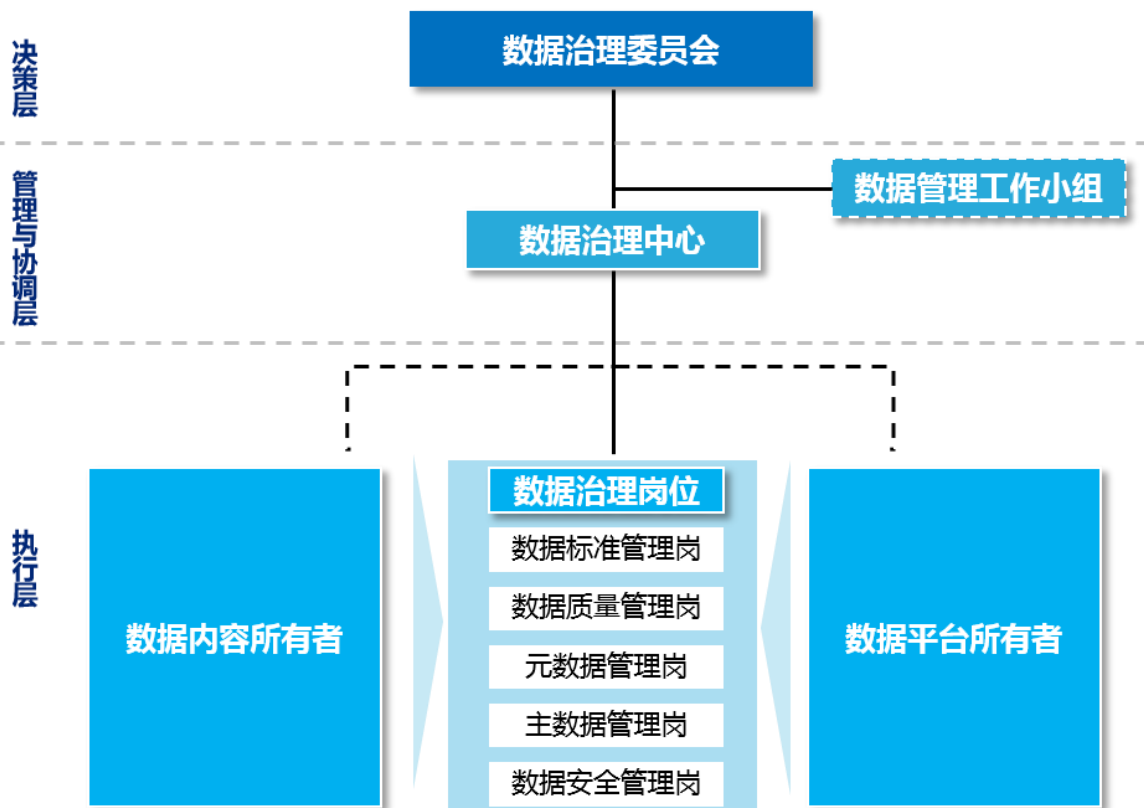
短期

大力开展数据治理、主数据、数据标准、数据质量、数据应用等领域相关工作，建立数据管理长效机制，夯实数据基础工作，支撑数据模型的运行。

中长期

完善重点领域管理能力，数据质量明显改善，加强数据模型应用，推进数据管理各领域工作全面开展和数据管理能力全面提升，全面提升企业的数据成熟度。

### 数据治理组织架构设计



#### 数据治理委员会

数据管理的最高决策机构。

#### 数据管理工作小组

数据治理工作的统筹协调与议事的组织。

#### 数据治理中心

数据治理工作的管理组织和推动的部门。

#### 数据内容所有者 (Content Owner)

业务部门与职能部门内部设置全职或兼职的数据治理岗位。

#### 数据平台所有者 (Platform Owner)

信息技术部各系统管理岗位或数据库管理岗位。

### 数据治理制度流程设计与编制



《数据治理制度》



《数据需求管理办法》

《数据标准管理办法》

《数据安全管理办法》

.....



《数据标准管理流程》

《数据质量管理流程》

《元数据管理流程》

.....



《数据治理操作手册》

1

### 模型健康性检查

- ✓ 架构层面
- ✓ 设计规范
- ✓ 管理流程
- ✓ 业务层面

2

### 模型优化

1. 结合行业通用数据模型的成果，扩充基础模型的覆盖范围
2. 充分考虑数据标准定义及大数据平台的特点进行优化设计
3. 主题及实体的定义更贴合业务实际，同时考虑到可扩展性的要求
4. 考虑目标应用是否能够方便、快捷支持

01

克服数据黑暗现象

通过清晰的数据模型管理让企业可以真正理解和运用自身的数据，并不断扩大应用和分析数据的范围和规模。

02

明确数据与流程的关系

了解数据访问与业务流程之间的关系，帮助企业业务使用者应用更好完成工作，推动全面数据化运营。

03

挖掘数据意义

连接和映射更多数据，充分发掘现有的数据之间的关系，扩大数据规模效应，让数据可以充分发挥其作用和价值。

04

各项数据活动的基础

其他的数据资产管理活动，包括数据质量、数据生命周期管理、数据操作、数据安全、主数据管理等提供一个高质量的基础。

客户主题

机构客户控股性质

➤

业务定义：公司控股主体的性质

➤

信息项类型：代码类信息项

国标

国家统计局  
关于统计上对公有和非公有控股经济的分类办法

第三条 控股经济分类与代码

- 100 公有控股经济
  - 110 国有控股
    - 111 国有绝对控股
    - 112 国有相对控股
  - 120 集体控股
    - 121 集体绝对控股
    - 122 集体相对控股
- 200 非公有控股经济
  - 210 私人控股
    - 211 私人绝对控股
    - 212 私人相对控股
  - 220 港澳台商控股
    - 221 港澳台商绝对控股
    - 222 港澳台商相对控股
  - 230 外商控股
    - 231 外商绝对控股
    - 232 外商相对控股

✓ 证监会标准

证监会  
公司控股情况分类标准

代码	控股情况
S	国有控股
11	中央国有控股
12	地方国有控股
C	社团集体控股
P	自然人控股
F	外商控股
O	其 他
11	控股主体性质不明确
12	无控股主体

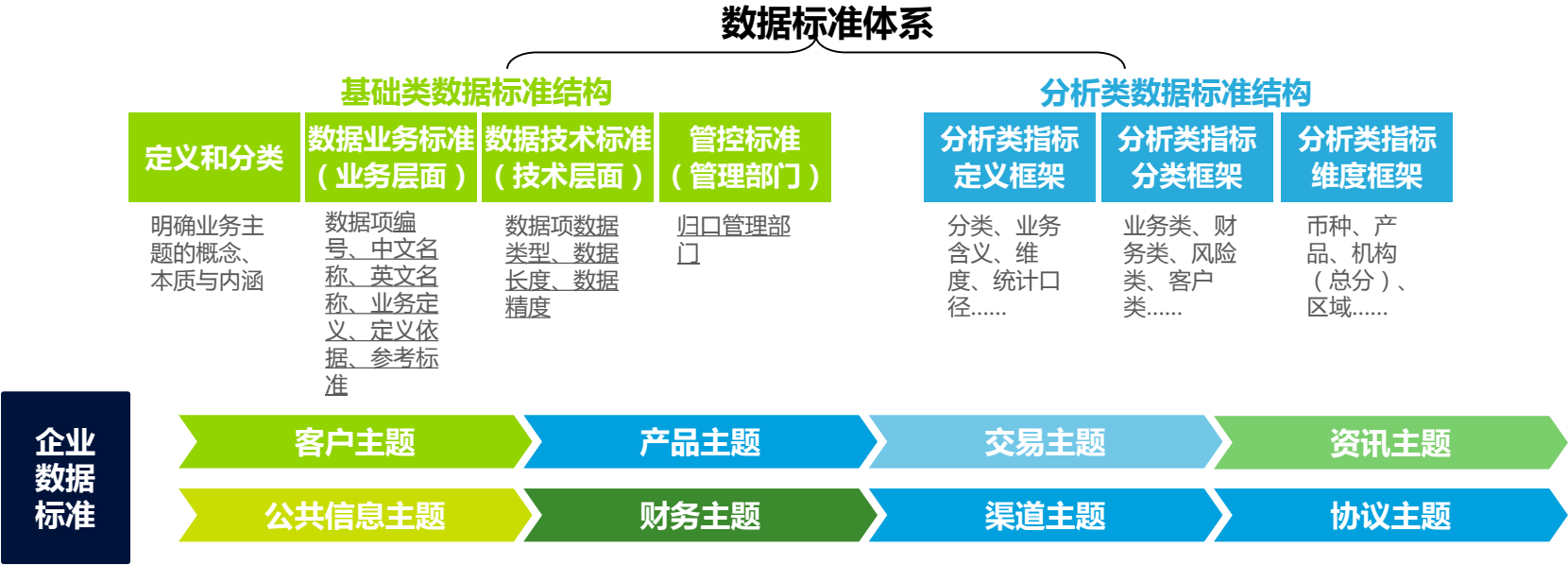
选用依据  
券商生产、报送、发布统计数据涉及的控股情况分类需遵循本标准。

系统现状码值

国家单位	10
国有独资	11
国有控股	12
集体企业	20
民营企业	30
港澳台投资	40
港澳台独资	41
中外合资	50
外资独资	51
其他	99



数据标准是企业或组织的数据项的分类、语义定义、值域和计算机应用的规范化集合，数据标准管理是建立、维护、应用数据标准的过程。

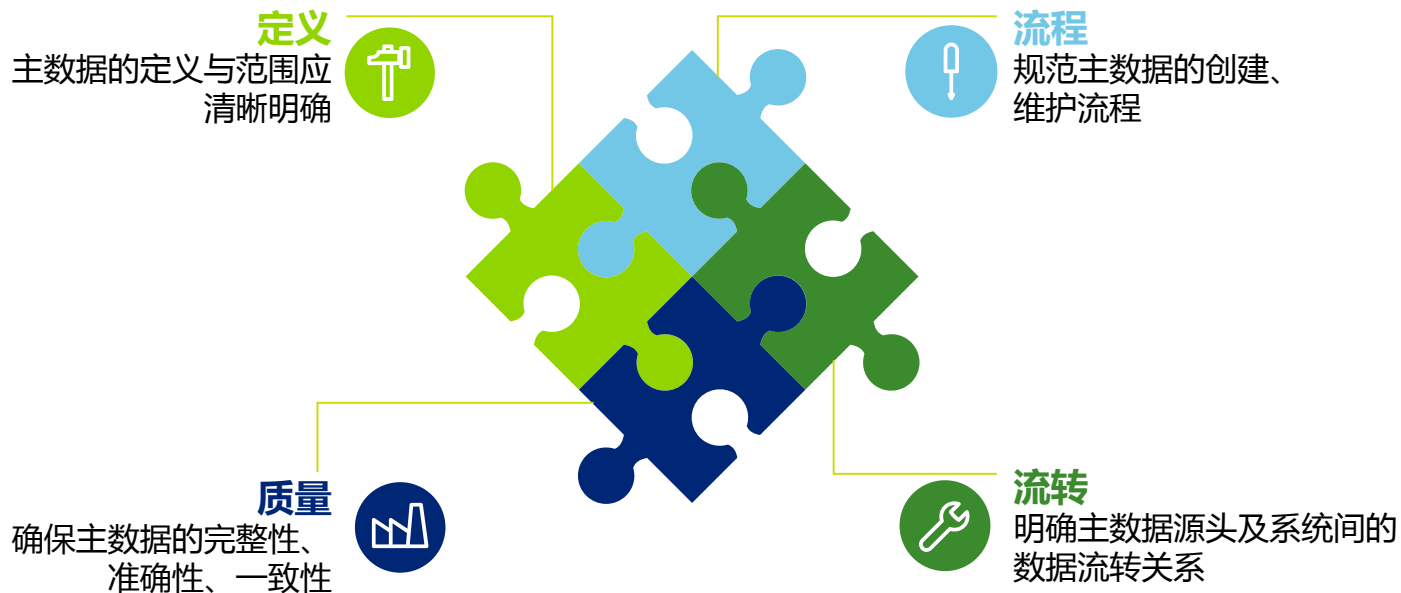


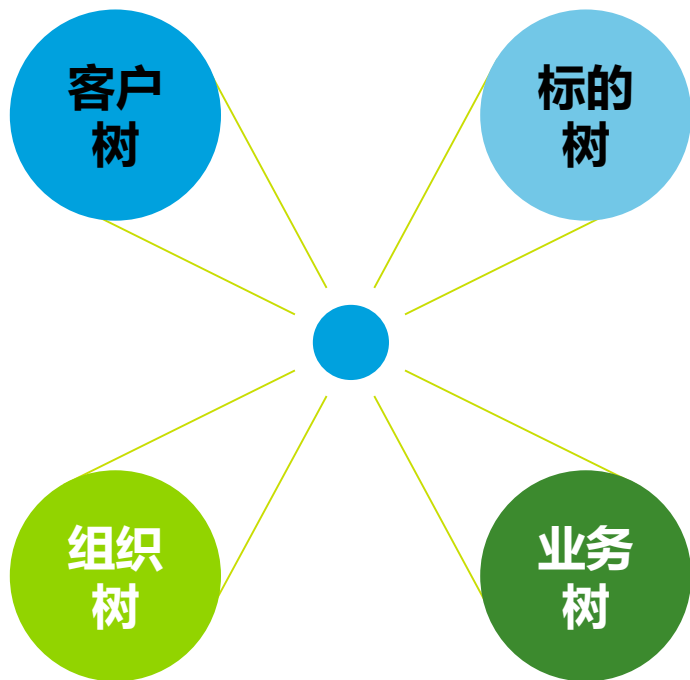
### 数据质量管理

- 数据作为全公司一项重要资产，质量管理是核心目标；
- 保证公司数据质量，数据认责和数据质量考核是抓手，数据标准、数据模型、元数据等是基础的必备条件；
- 通过建立数据质量管理端到端的闭环管理机制，做到事前防范，事中控制，事后治理相结合，全面主动的进行数据质量持续提升；



- ✓ **准确识别**企业的主数据，确保主数据在企业内部的**完整性、准确性和一致性**。
- ✓ 建立**主数据管理机制和平台**，为企业的主数据建立**统一的视图**。





### 产品目录设计目标

- 能覆盖证券公司母、子公司各业务条线业务、产品、客户；
- 能为各产品 / 部门提供统一产品衡量标准，便于部门之间沟通管理；

### 产品目录实现方式

- 按照四个维度进行理论上笛卡尔积的生成
- 对四个维度一起生成的笛卡尔积的数据进行合理性检查，删除不合理组合。

## 元数据管理

元数据应用(分析应用, 元数据查询展现等)

元数据管理(增删改查、统计管理、版本管理等)

公共接口  
(WebService、通用接口等)

元数据存储与计算

元模型管理

元数据采集

## 元数据分析

实现数据分布地图, 数据血缘分析和影响分析。

## 元数据查询与展现

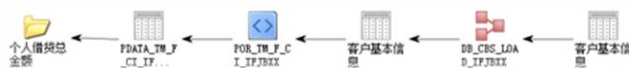
提供技术元数据, 业务元数据等信息查询的展现, 支持元数据的统计等。

## 元数据管理

对元数据版本进行匹配, 及时通知和提醒业务元数据的变更。通过元数据的登记修改流程对元数据进行管理。

## 元数据采集

实现对管理范围内的技术元数据(数据结构、ETL加工, 数据映射等)、业务元数据(指标报表、标准)的自动或手工采集, 完成自动匹配, 实现对无法自动采集的内容作补录



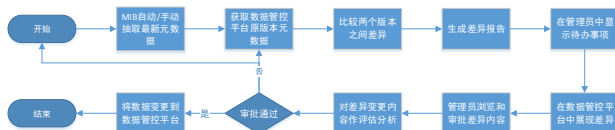
物理表: T8005820 / 托创明报表 使用状态: 在用 已有 8 个分片

英文名称: BTMMN  
描述:  
业务定义: 用途: 记录托创明创建信息; 范围:

更多信息 点选情况 字段 相关数据源方案列表

共 15 个字段

编号	编号	英文名称	中文名称	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型	数据类型
CL00000953252	1	TSCFBH	托创编号	CHAR	16	false	false	是	在用								
CL00000953253	2	YNYFJG	营业机构号	CHAR	4	false	true	是	在用								
CL00000953254	3	GUYYDH	柜员代码	CHAR	6	false	true	是	在用								
CL00000953255	4	JQYVRQ	交易日期	CHAR	8	false	true	否	在用								



数据库类型: sqlserver

数据库驱动: net.sourceforge.jtds.jdbc.Driver

连接地址: jdbc:jtds:sqlserver://10.253.1.47:22937;DatabaseName=basedb

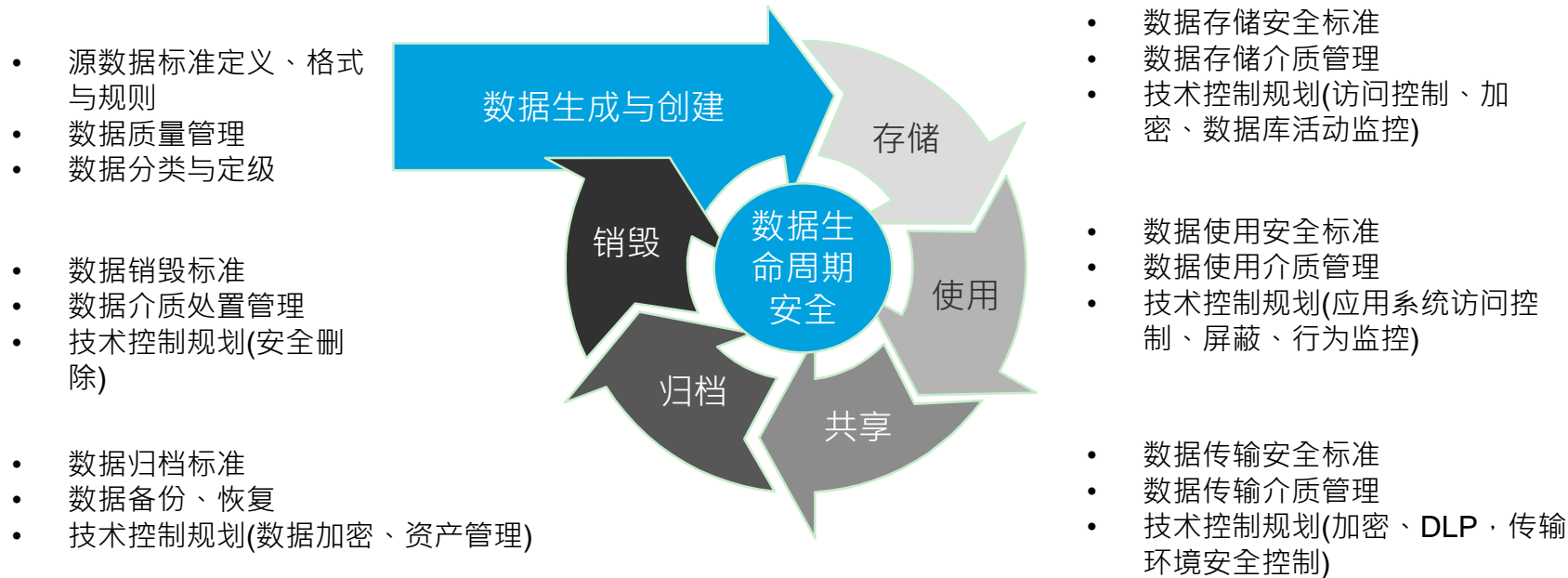
URL 格式: jdbc:jtds:sqlserver://{server}<port1433>;DatabaseName=<database>

用户名: sa

密码: \*\*\*\*\*

数据库版本:

- ✓ 企业应通过建立对数据及相关信息系统进行保护的一系列措施，确保数据免遭未经授权的访问、使用、修改或删除，保证数据完整性、保密性和可用性



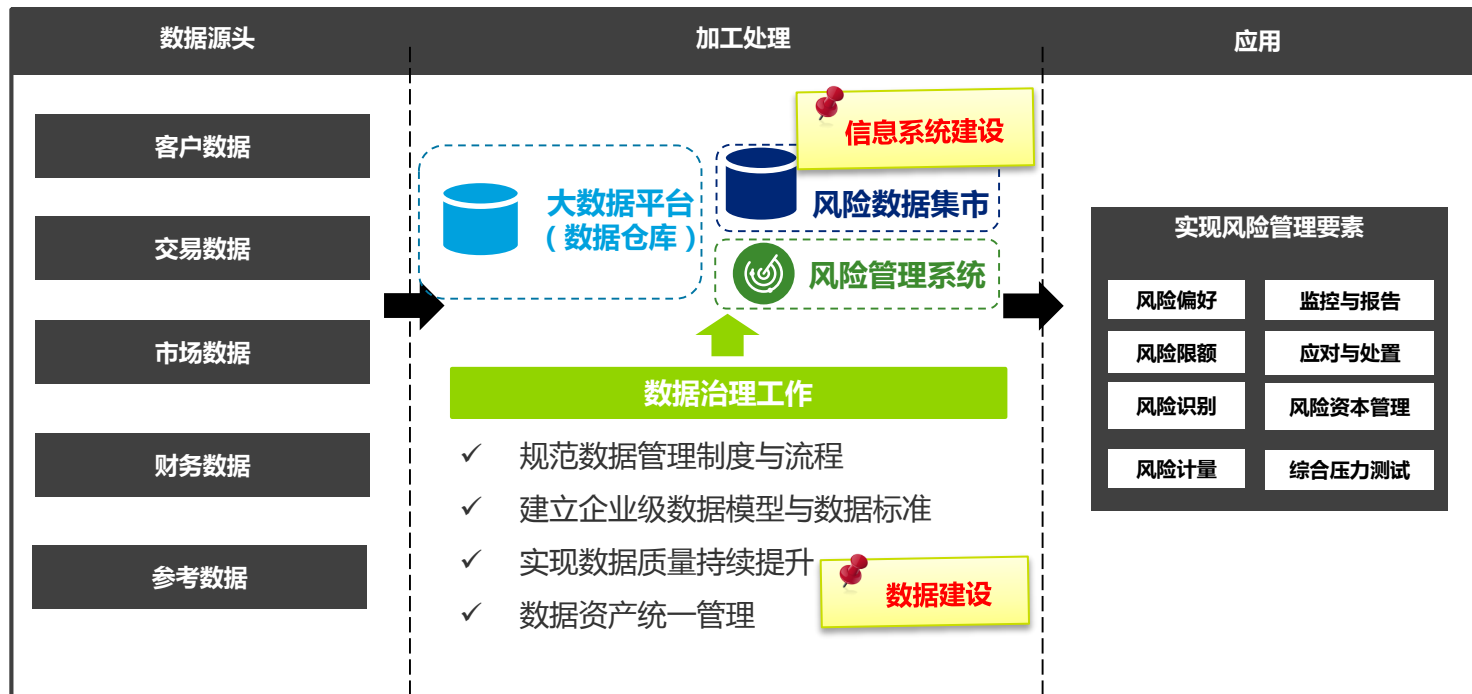






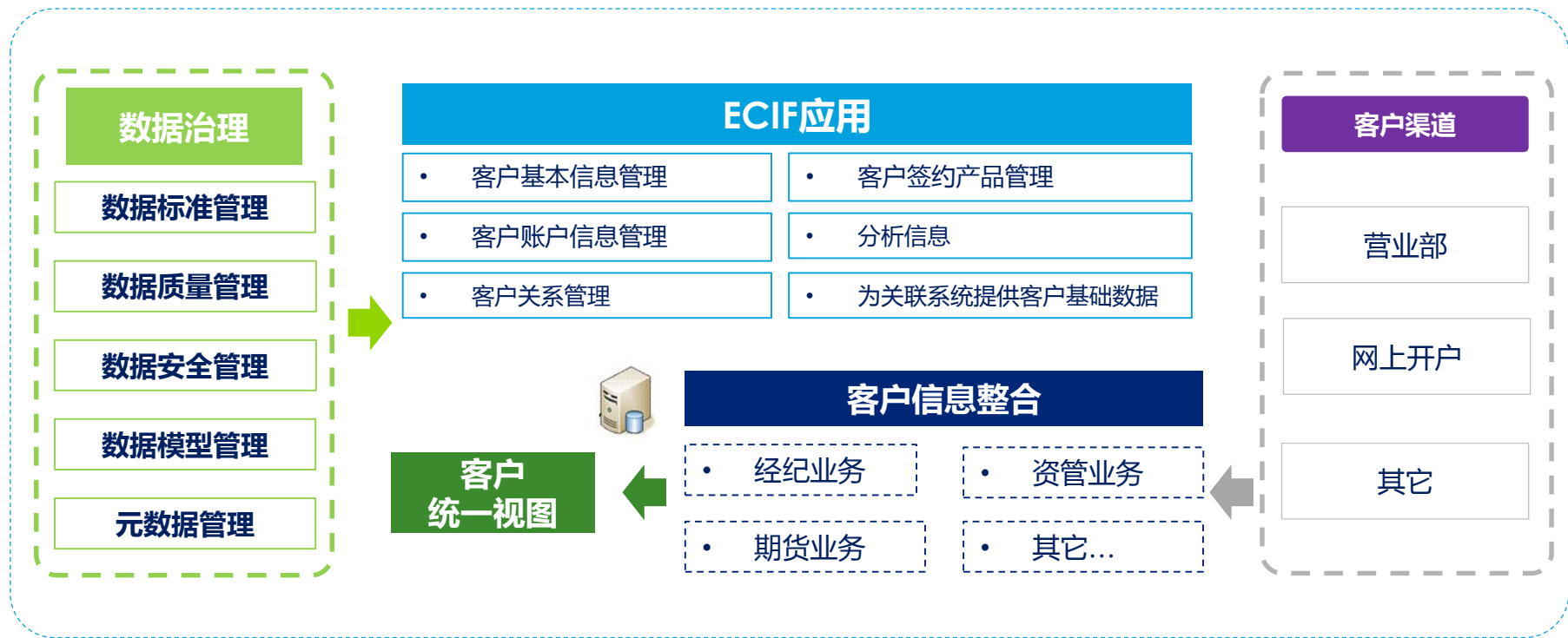
- 数据和系统建设是风险管理的基础设施

#### CASE II 建设风险管理基础设施



- 使用技术手段，实现“以客户为中心”的服务理念

#### CASE III 建设客户信息平台 (ECIF)



#### CASE IV 建立第一方DMP平台

#### 建立整合模型，实现营业部客户画像数据和CRM数据的贯通

