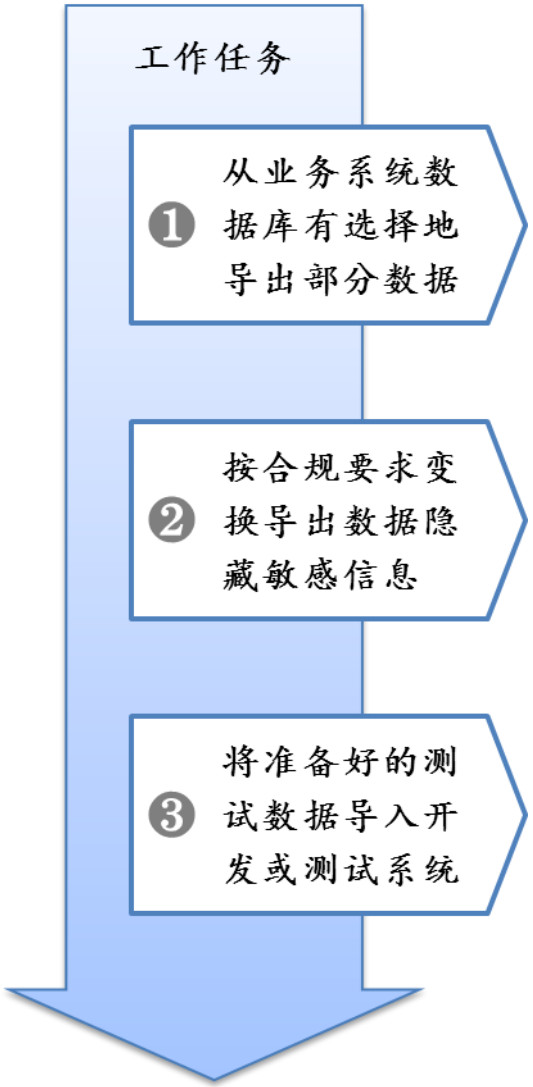
数 据 抽 取 和 脱 敏

银 行 测 试 数 据 管 理 和 隐 私 保 护 解 决 方 案

银行测试数据管理需求和难点



传统工作方式遇到的难以解决的问题

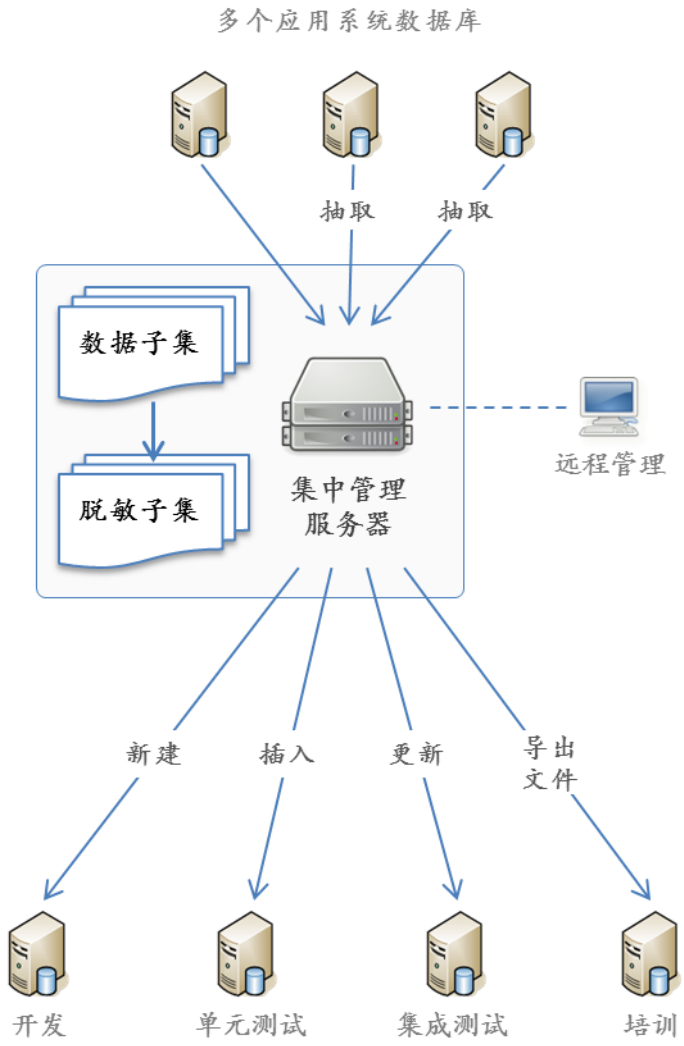
- 跨系统跨平台的测试数据难以准备，管理难度高
- 数据变形和脱敏手段缺乏或过于单一，且脱敏过程无法监控，测试数据安全性难以保证
- 手工处理方式效率低下，数据质量难以保证
- 测试数据版本管理难度大，复杂度高
- 难以保证如期提交高质量测试数据
- 复用测试数据的计划难以实现，经常需要反复准备数据，造成人力物力资源浪费
- 需要长期保持一支人员较多的数据准备团队，成本居高不下
- 难以统一记录数据准备过程中各项工作的日志，审计难度高

准备测试数据的最佳实践
需要建立集中数据管理平台

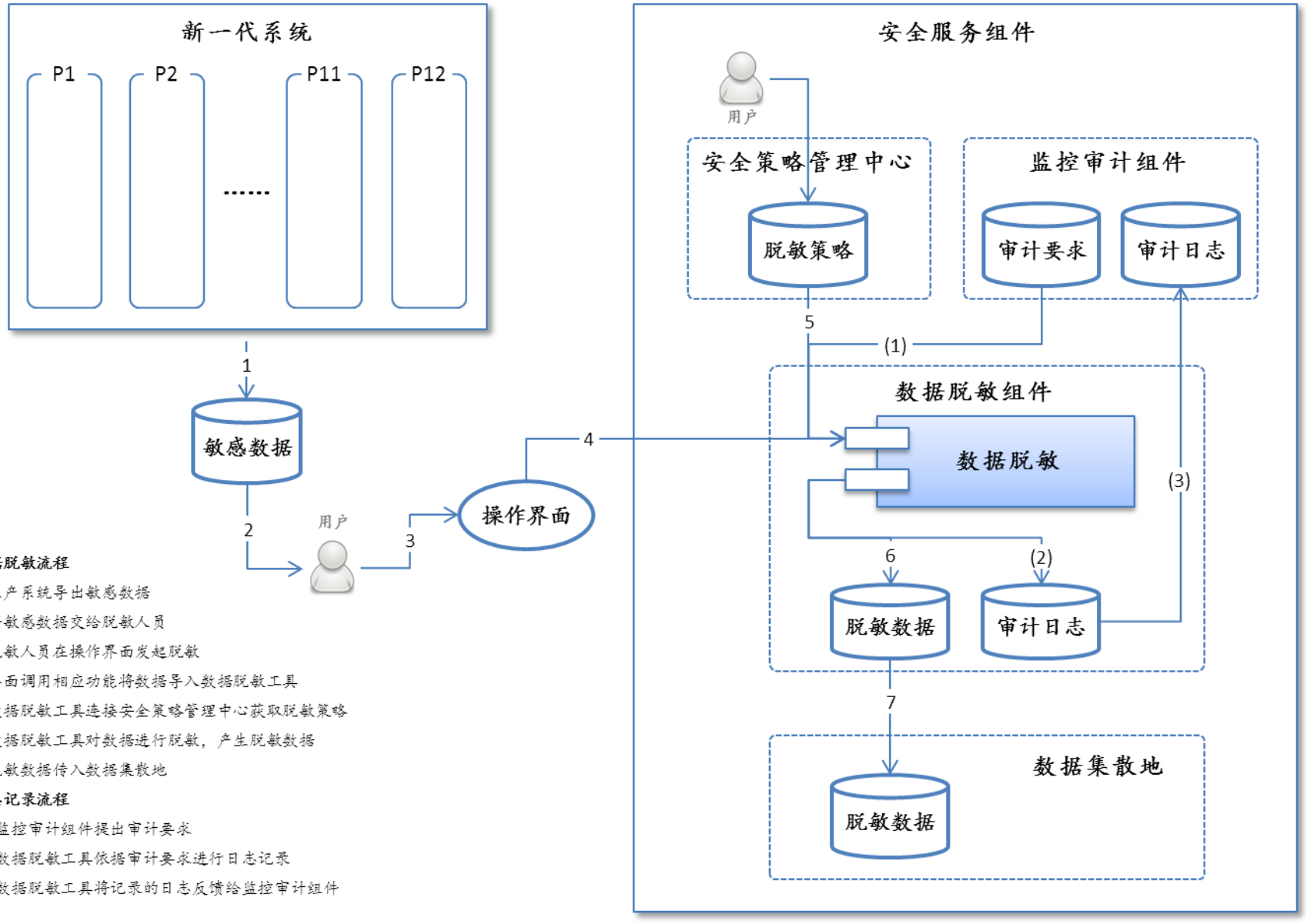
测试数据集中管理平台架构

集中管理数据抽取、脱敏、和加载的优势

- 基于预定义条件或者抽样数据抽取，为目标测试环境创建大小合适、规模适中、保留业务特性的测试数据子集
- 对数据子集中的敏感数据完成脱敏操作，以保护数据隐私，同时保持业务相关上下文语义，不影响数据的完整性
- 智能比较脱敏前后数据的差异变化，预防兼容问题
- 缩减测试数据准备时间，加快测试进程，提高效率
- 脱敏后的测试数据子集压缩存储，可反复使用，能随时加载到指定的数据库中，降低业务系统的压力
- 与生产系统隔离，支持用户权限管理和日志审计，保证敏感数据安全，遵从合规要求
- 本地支持服务协助用户配置子集抽取，有效降低用户繁琐的初始化和持续设置工作量



案例：某国有银行研发中心数据脱敏方案



数据脱敏流程

1. 生产系统导出敏感数据
2. 将敏感数据交给脱敏人员
3. 脱敏人员在操作界面发起脱敏
4. 界面调用相应功能将数据导入数据脱敏工具
5. 数据脱敏工具连接安全策略管理中心获取脱敏策略
6. 数据脱敏工具对数据进行脱敏，产生脱敏数据
7. 脱敏数据传入数据集散地

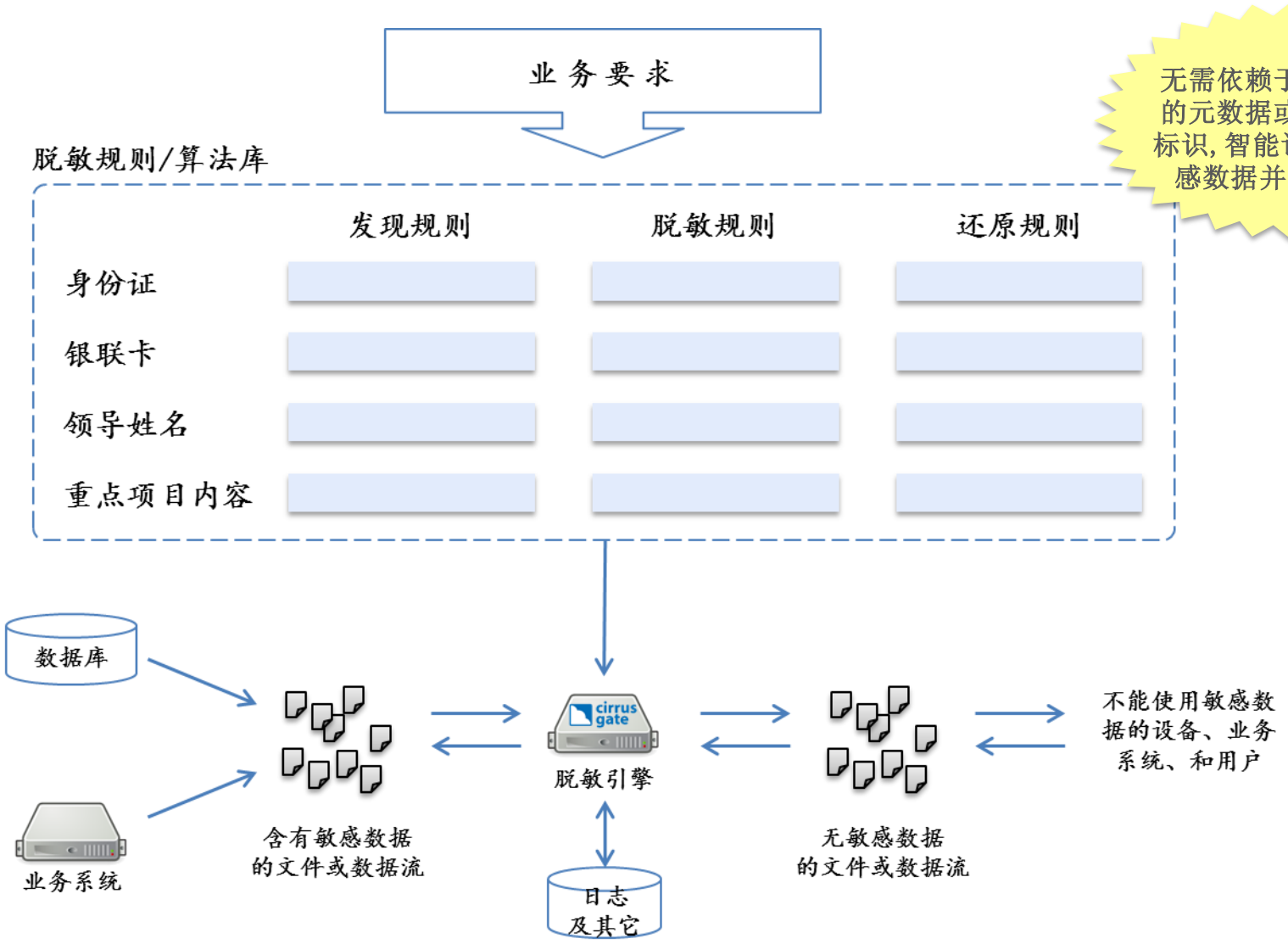
日志记录流程

(1) 监控审计组件提出审计要求

(2) 数据脱敏工具依据审计要求进行日志记录

(3) 数据脱敏工具将记录的日志反馈给监控审计组件

智慧数据脱敏与还原


业务要求
无需依赖于预先定义的元数据或其它数据标识，智能识别发现敏感数据并自动处理
脱敏规则/算法库
发现规则
脱敏规则
还原规则
身份证
银联卡
领导姓名
重点项目内容
数据库
业务系统
含有敏感数据的文件或数据流
cirrus gate
脱敏引擎
日志及其它
无敏感数据的文件或数据流
不能使用敏感数据的设备、业务系统、和用户

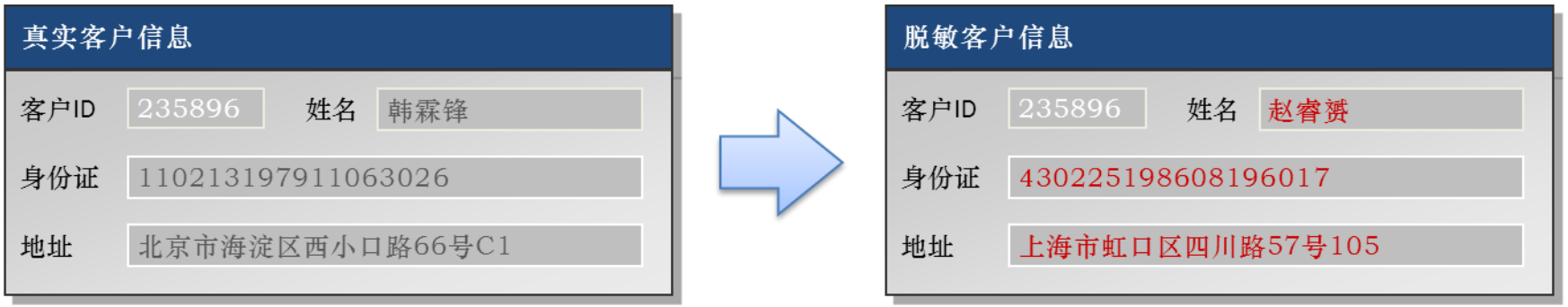
银行测试数据准备中常见的脱敏需求

智慧发现

- 不依赖元数据或其它数据标识，智能扫描数据内容发现敏感信息，支持结构特征明显的数据、以及自然语言文本
- 正确识别客户信息和交易信息等：身份证、地址、电话号码、邮件地址、银行账号、信用卡号
- 能正确识别多个敏感数据组合在同一字段，如“身份证+姓名+信用卡号”
- 隐藏在注释或者文本列中的敏感信息

语义保持

- 数据脱敏后仍能正确通过有效性验证，如身份证的校验码和生日区间
- 取值范围合理，如信用卡号变换后仍是本行卡号区间等
- 脱敏策略可以保持业务需求的特定信息，如按年龄段进行业务分析
- 保持数据长度、可读性、完整性、上下文数据关联性等
- 可根据安全策略设置进行全局变换或局部变换



附录：智慧数据安全产品简介

采用新技术解决困扰银行的海量数据管理难题

专注于
识别数据内容的
技术和应用

银行面临全新
数据管理挑战

- 数据治理 Governance
- 数据安全 Security
- 数据分析 Analytics

新技术方向

自然语言处理 | 数据挖掘 | 机器学习 | 数据可视化

我们的核心技术

- 采集
- 转换
- 加载

- 内容识别
- 分类
- 聚类

- 特征提取
- 内容摘要
- 图像匹配

- 可视化
- 展现
- 分析

我们的产品

① 数据治理 Data Governance

- 聚类分类 Clustering/Classification
- 内容管理 Content Management
- 数据梳理 Profiling

② 数据安全 Data Security

- 自动分类分级
- 数据脱敏 Masking/Tokenization
- 数据防泄漏 DLP
- 云网关 Cloud Gateway
- 加密 归档 销毁 Encrypt/Archive/Wipe
- 大数据保护 Big Data Protection

③ 数据分析 Data Analytics

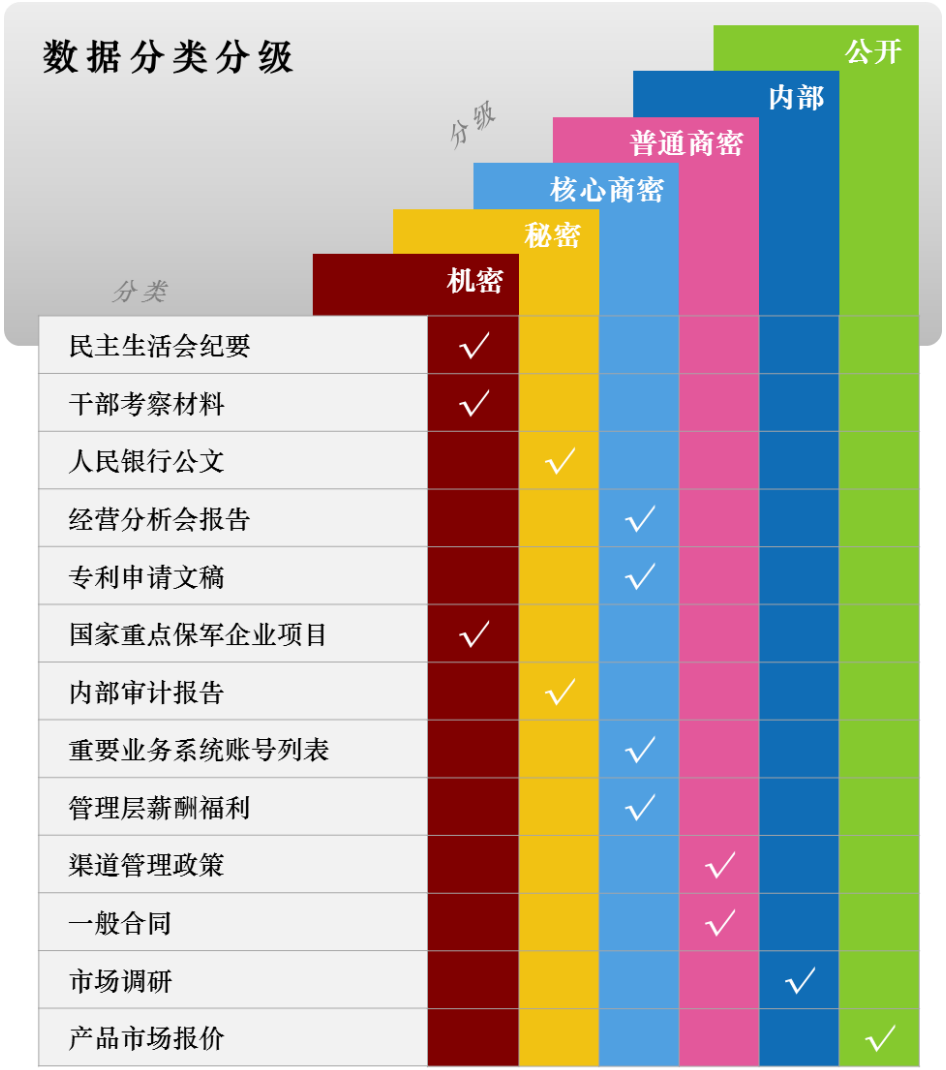
- 采集 转换 加载 ETL
- 聚类 分类 关联 Analytics
- 可视化展现 Visualization

统一数据管理平台

自动数据分类分级

- 信息安全国际标准认为，不同数据的重要性各不相同，高价值的数据需要更严格的保护机制
- 数据分类分级是信息安全风险评估流程和数据安全治理中的一个重要组成
- 国资委《中央企业商业秘密保护暂行规定》中要求对商密数据分类，实施分级管理，并进行标识
- 银监会《十二五信息科技发展规划监管指导意见》明确要求推进信息资产分类分级管理
- 工信部《公共及商用服务信息系统个人信息保护指南》2013年2月1日实施

基于内容识别的、机器智能的、可处理海量数据的、实时自动的数据分类分级



领先的内容识别技术

完全自主知识产权

简单关键字

- 入门级产品
- 误报漏报率极高，实际无法使用
- 没有任何改进空间

正则表达式组合

- 产品简单易理解，上手快
- 更适合结构化数据匹配
- 需要经验丰富的顾问进行人工归纳关键字和正则表达式，大量试错，后续维护困难，改进空间小
- 复杂正则表达式的匹配性能非常糟糕

自然语言处理

- 使用自然语言处理、数据挖掘、和机器学习技术的聚类/分类器，对以中文撰写的例如公文、会议纪要、经营计划等非结构化文档的分类效果十分出色
- 可根据客户行业特点和自身业务要求，划分至更加细分类别，处理非结构化数据的实际效果远远超过关键字和正则表达式产品
- 机器学习自动生成规则库，准确率和可靠性比人工总结正则表达式高得多

支持特征数

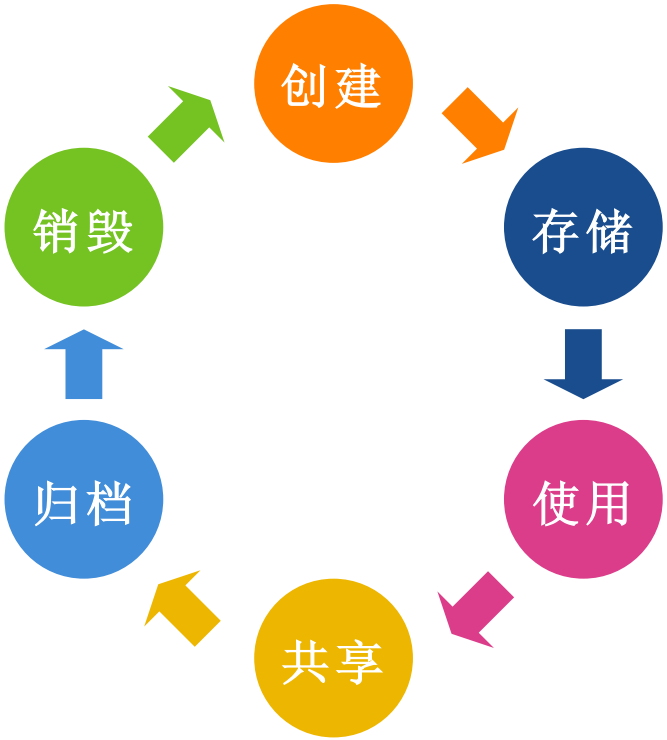
- < 5
- 或与关系

- < 20
- 布尔运算

- > 300
- 语义相似度

覆盖全数据生命周期的解决方案

数据生命周期



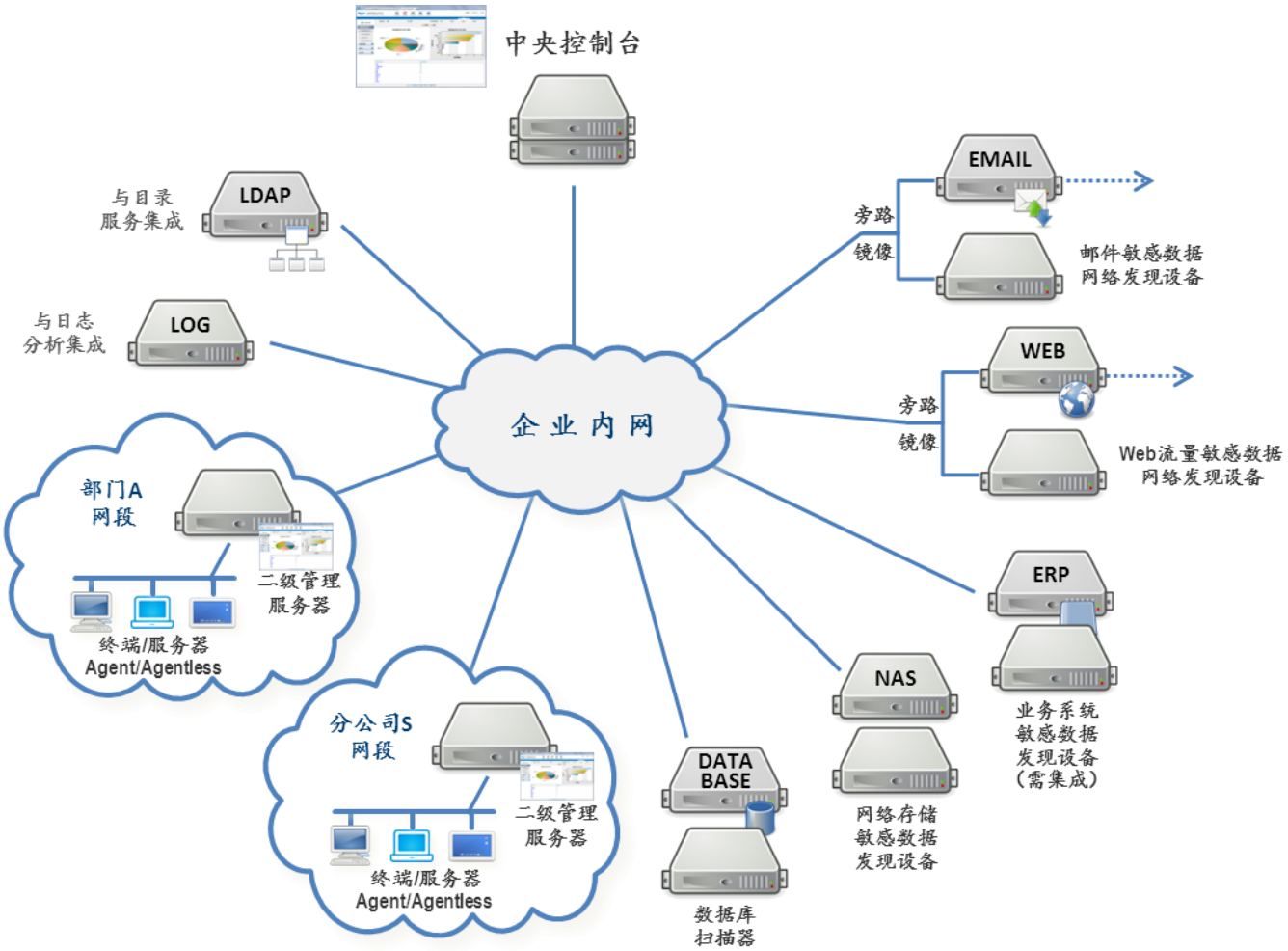
智慧数据安全平台架构

基于先进内容识别引擎 (ACCE)

可视化集中管理平台：洞察关键数据分布

市场上首个产品！
可视化展现关键数据的分布、
风险、和安全事件

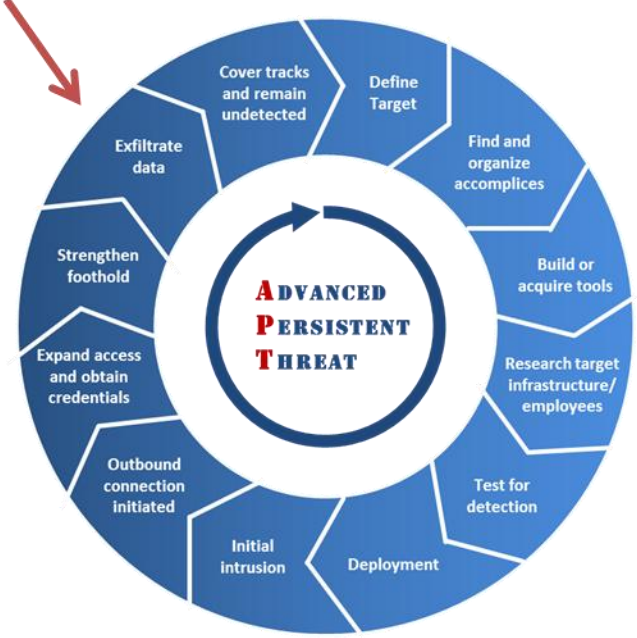
覆盖企业中所有关键数据出现位置



- 终端、数据库、网络存储、邮件、互联网流量、业务系统、数据仓库、数据集市、大数据系统等
- 可根据敏感数据管控需求任意组合，分期实施，逐步覆盖
- 树形分级部署管理服务器，集中管理，权限按角色分配，适合国内大型企业管理制度
- 完备的扩展接口，可集成定制化

智慧监控敏感数据以侦测和防御APT攻击

APT攻击的目标是窃取关键数据！



APT攻击的生命周期

针对APT攻击的根本，实时监控敏感数据的分布和流动，与安全基线比对，智能分析异常动向：

- 某台PC过去一周内保有的敏感数据数量突然剧增
- 凌晨时分大量敏感数据从某网段流向未经授权的另一网段
- 一周内某台设备持续向公司外通过各种途径(HTTP/FTP/SMTP)发送敏感数据
- 业务系统的大量客户信息被转移到数据集市，随后被下载到某台PC
- 更多风险因素和场景智慧分析！

云计算和大数据基础架构中的数据保护场景

进入和离开云的数据

- 使用URL过滤器和基于内容识别的数据防泄露DLP等技术，监控数据向云中迁移的过程
- 使用基于内容识别的“网络数据/文件活动监测”，侦测和预防敏感数据离开云
- 加密：链路/网络加密模式、客户端/应用程序加密、基于代理的加密
- 数据脱敏：使用特定规则对业务数据进行变换，以隐藏客户真实证件号码、账户、交易记录等关键信息，必要时仍可还原
- 权限控制

在云内存储的数据

- 采用内容发现产品，识别云存储的敏感信息，并加以管理和监控
- 加密：IaaS（卷存储加密、对象存储加密），PaaS（客户端/应用加密、数据库加密、代理加密、其它），SaaS（服务提供方管理加密、代理加密）
- 数据防泄露DLP（专用设备/服务器、虚拟设备、终端代理、Hypervisor代理、DLP SaaS）
- 数据库和文件活动监测
- 数据脱敏
- 权限控制

在云内迁移的数据

- 使用基于内容识别的“网络数据/文件活动监测”，管控大量数据的内部迁移
- 数据库活动监控器可以监测到大数据集移出或数据库复制，表明有迁移发生
- 加密：链路/网络加密模式、基于代理的加密
- 数据脱敏
- 权限控制