DTCC 2010-2022

第十三届中国数据库技术大会

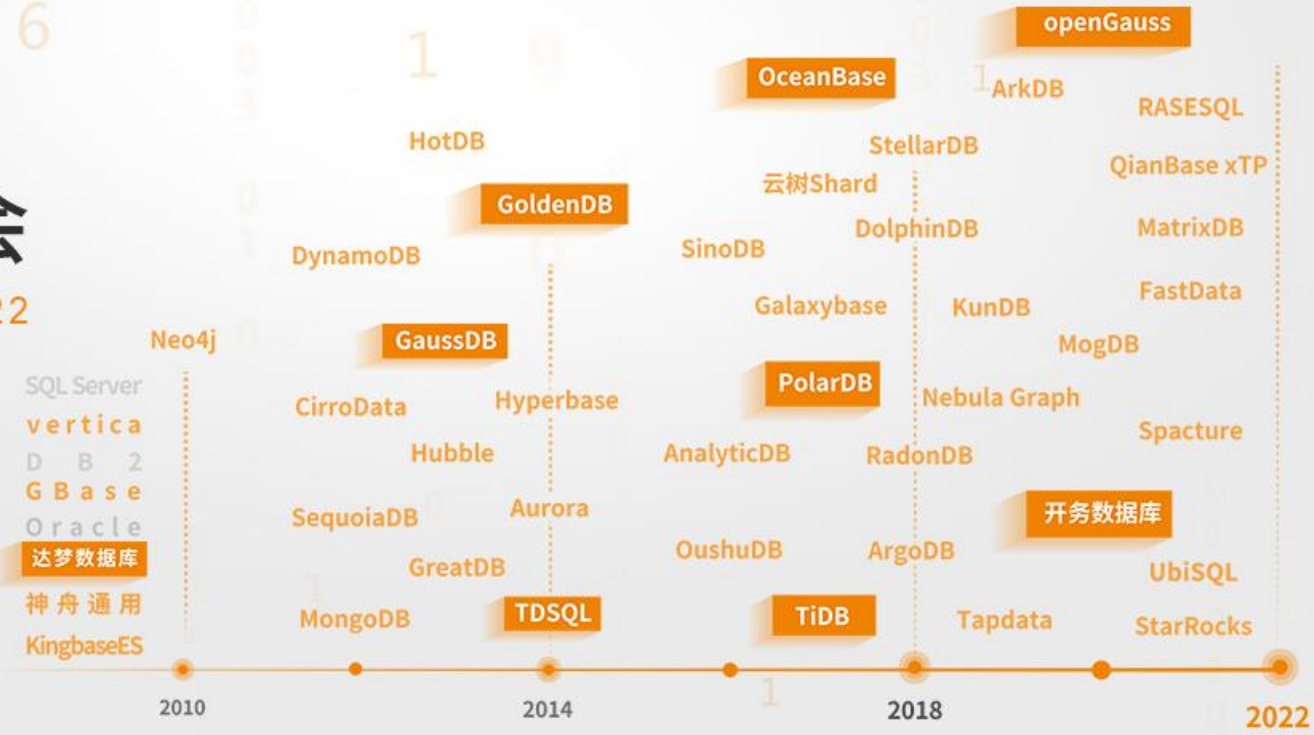
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



快手大数据安全平台建设与实践

马玲玲 快手 大数据安全负责人

自我介绍

马玲玲　　快手大数据安全负责人

- 主要负责大数据安全平台的体系化建设工作
- 主要关注大数据平台架构和大数据安全技术领域

关于快手

愿景

◆致力于成为全球最**痴迷于为客户创造价值的公司**

◆我们的使命是帮助人们**发现所需、发挥所长，持续提升每个人独特的幸福感**

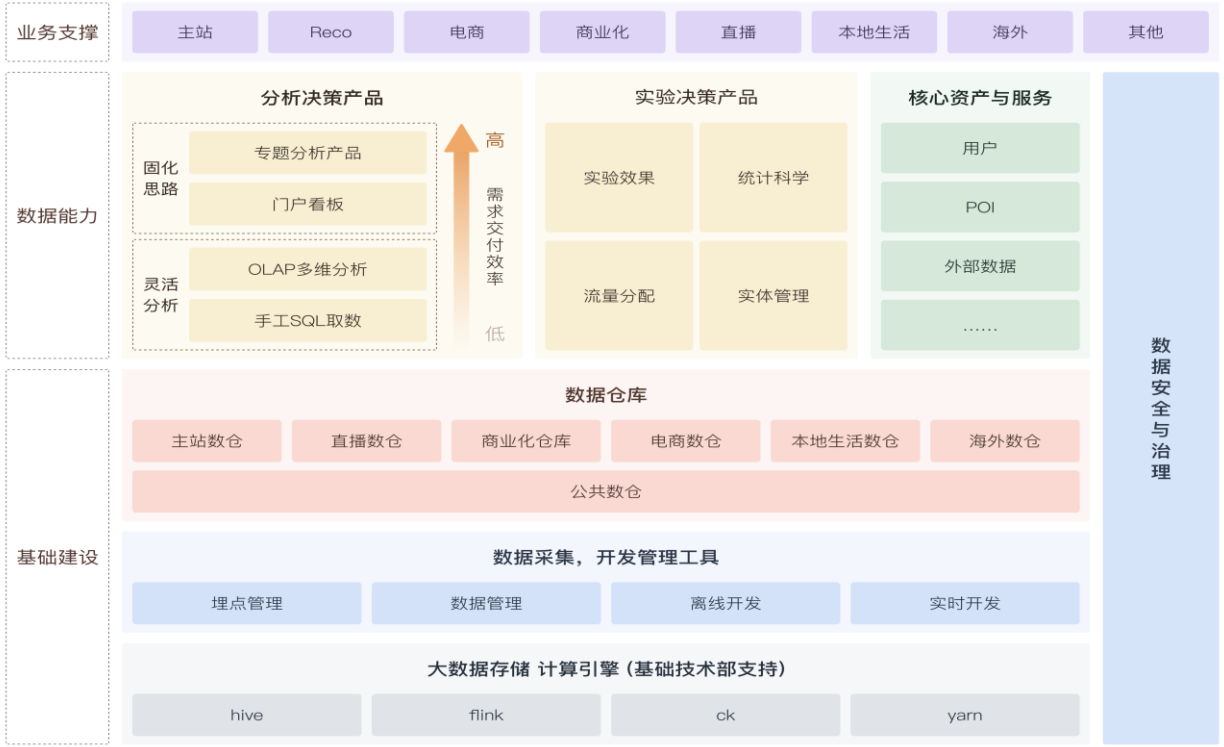
全民短视频社区

日活用户 3.63亿

月活用户 6.26亿

日均使用时长 129.3分钟

关于快手大数据平台



使命

提升数据决策效率，利用数据助力业绩提升

职责

通过大数据技术，对公司数据统一采集、存储、加工和挖掘形成高质量全域数据资产，以分析决策产品和服务的方式对外提供数据解决方案

万级

集群规模

EB级

总数据量

PB级

日净增数据量

十万级

日作业量

目 录

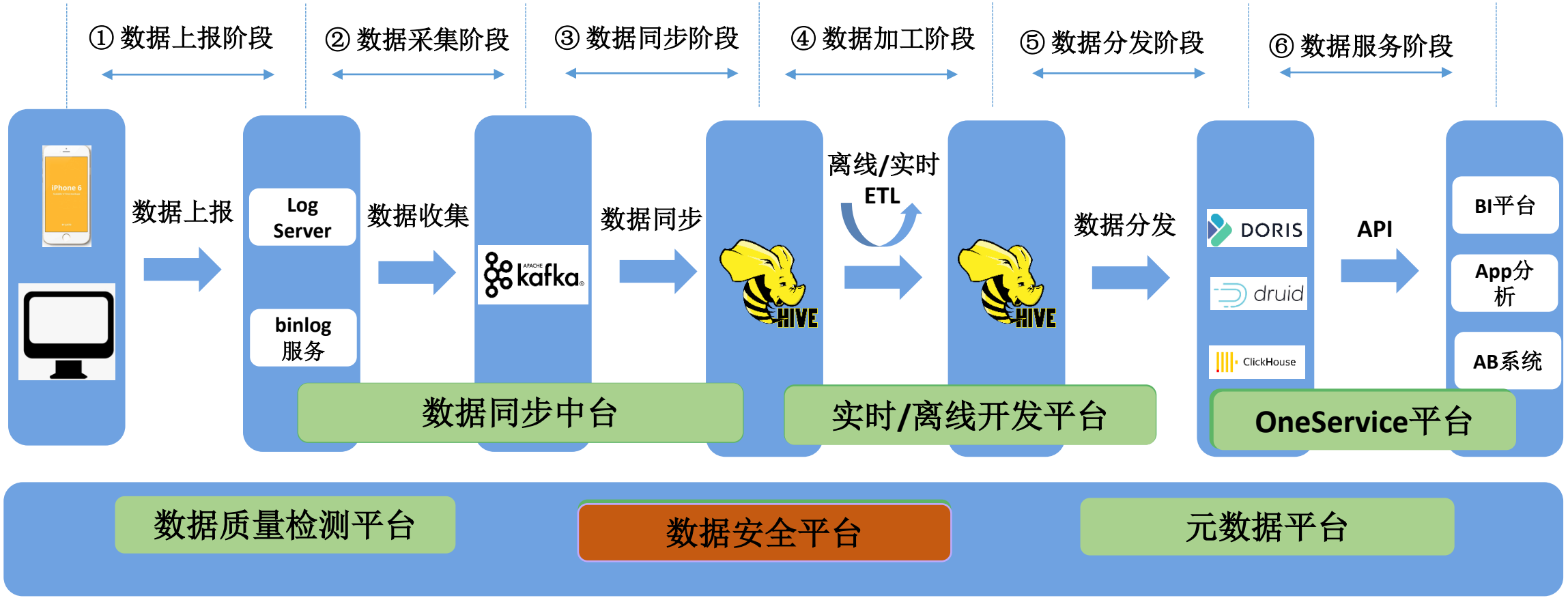
01

背景介绍

- 快手大数据安全平台定位
- 快手大数据安全面临的挑战
- 快手大数据安全建设思路

快手大数据安全平台定位

职责：为大数据全链路、全生命周期保驾护航，保障数据安全



快手大数据安全面临的挑战

系统覆盖度广

- 大数据计算和存储引擎
- 数据生产类平台
- 数据分析类平台

数据精细化管控

- 支持报表、数据集、指标、维度、库、表、行、列、文件等多种资源的权限控制
- 对数据的读、写等操作进行细粒度权限控制
- 满足多租户体系的数据隔离和权限管控

性能要求高

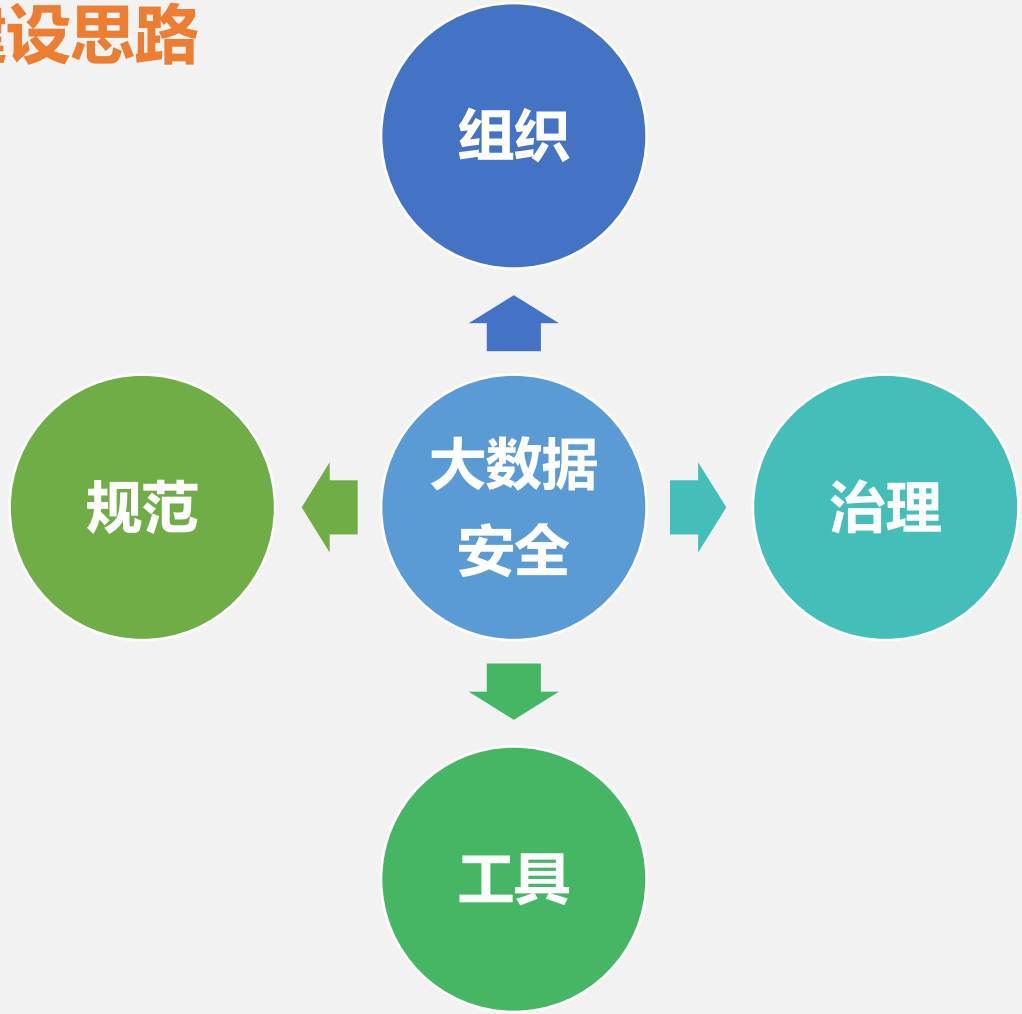
- 需要支撑千级用户、百万级资源的亿级权限关系，满足几十毫秒级鉴权延时
- 支持OLAP每天亿级查询，HDFS百万级QPS

业务灵活多变

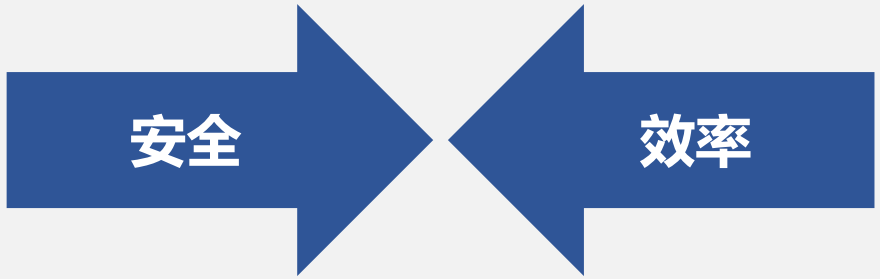
- 满足多种业务线的权限管控需求
- 满足数据分析类平台灵活多变的业务需求

快手大数据安全建设思路

建设思路



建设原则



保证安全兼顾效率

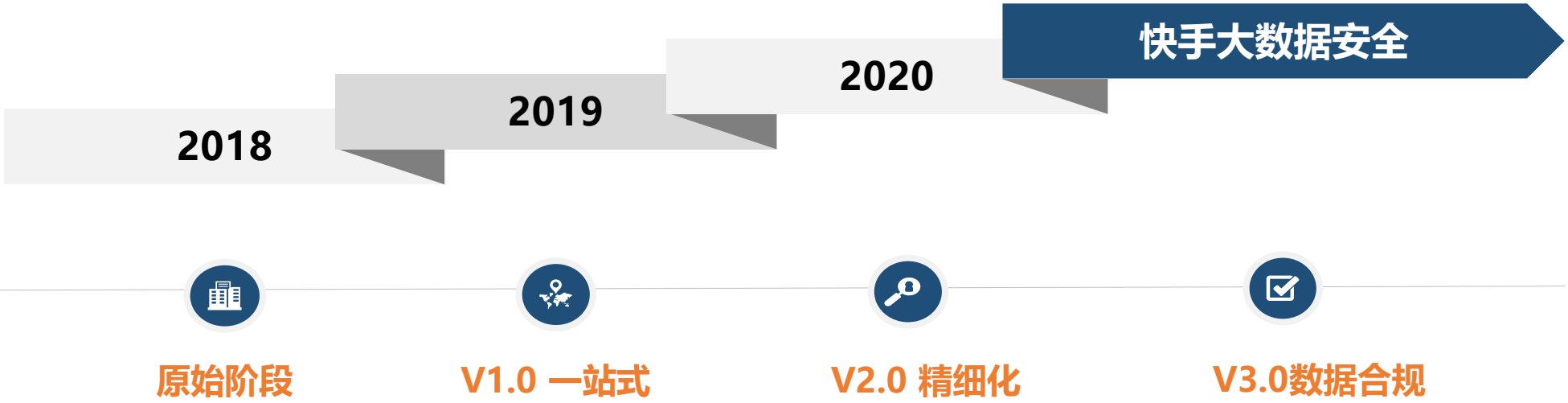
安全原则

02

平台建设

- 发展历程
- 解决方案
- 系统架构
- 关键技术

发展历程



权限模型	RBAC：资源和角色	PRBAC：资源、资源包、动作、用户组	PRBAC(行级权限)：行级权限、租户数据隔离	PRBAC(行列级权限+多模式)：列级权限、精细的管控模式
安全能力	2A能力：提供鉴权、申请、主动授权的能力（2A安全能力）	2A能力：申请、审批、授权、清查等一站式权限管理能力	4A能力：大数据统一认证、全链路审计	5A能力：加解密、脱敏、安全隔离仓
系统覆盖	分析类：报表平台	分析类：报表/分析工具/实验等 引擎类：HIVE	分析类：报表/分析工具/实验等 开发类：ETL开发/同步/API服务 引擎类：HIVE	分析类：报表/分析工具/实验等 开发类：ETL开发/同步/API服务 引擎类：HIVE/DRUID/CK/KAFKA/HDFS

解决方案

全域

数据生产平台

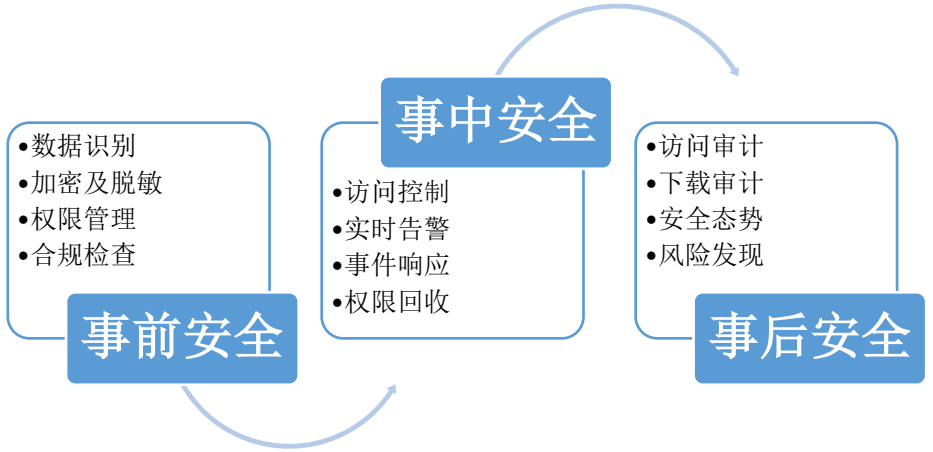
数据分析平台

大数据计算及存储引擎

大数据安全

全能力

全周期



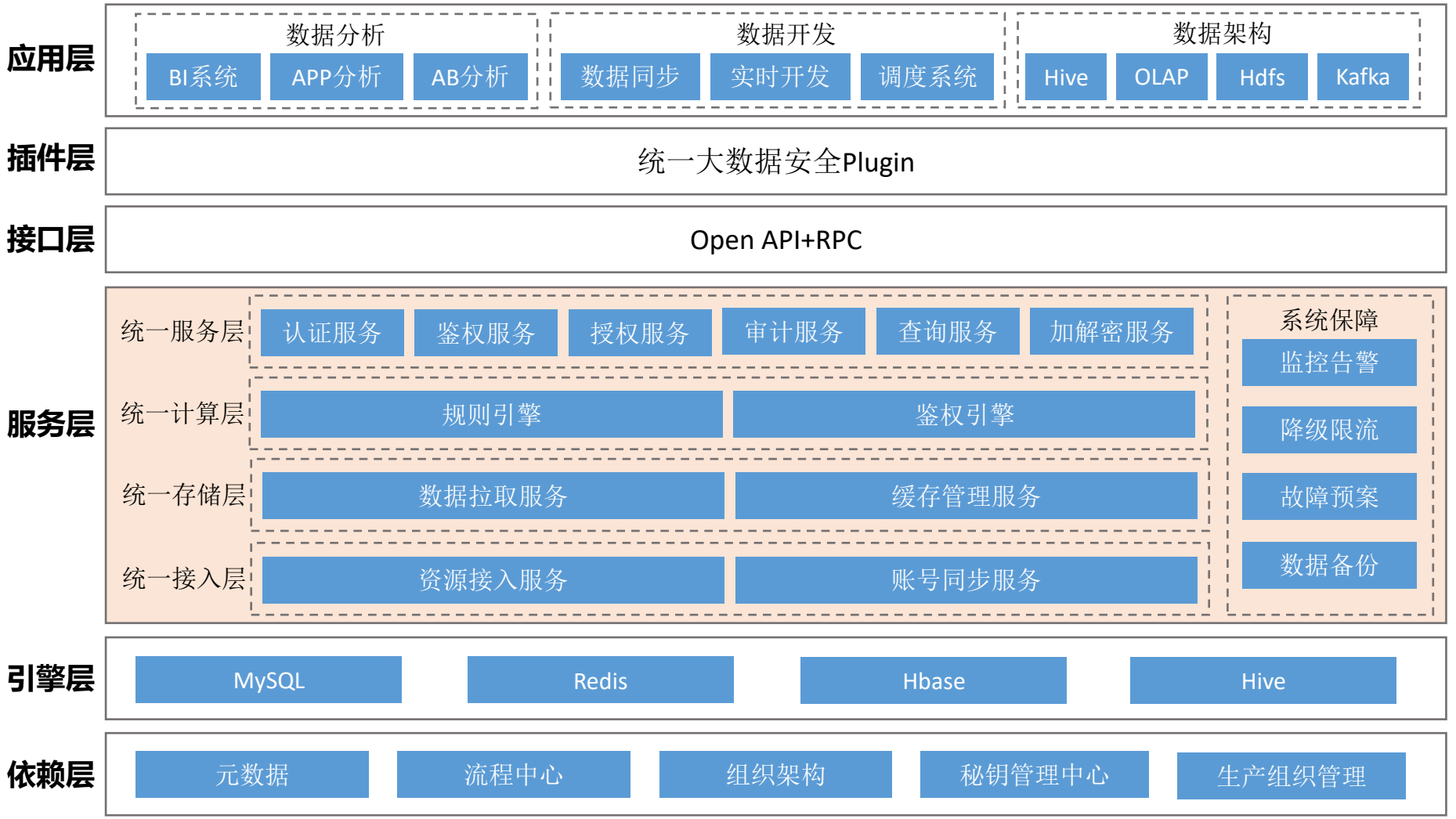
系统架构

统一化和插件化

- **统一服务：**提供通用的认证、鉴权、查询、审计等服务
- **统一计算：**鉴权和规则计算
- **统一接入：**资源通过元数据总线统一接入
- **统一存储：**提供缓存管理、缓存数据加载及版本管理等
- **引擎组件插件：**满足各个引擎自身特点，比如高QPS、低延时等

系统保障

- **高可用保障：**提供监控告警、降级容错、预案演练、限流等一些列措施，保障系统的高可用
- **高性能保障：**多级缓存等



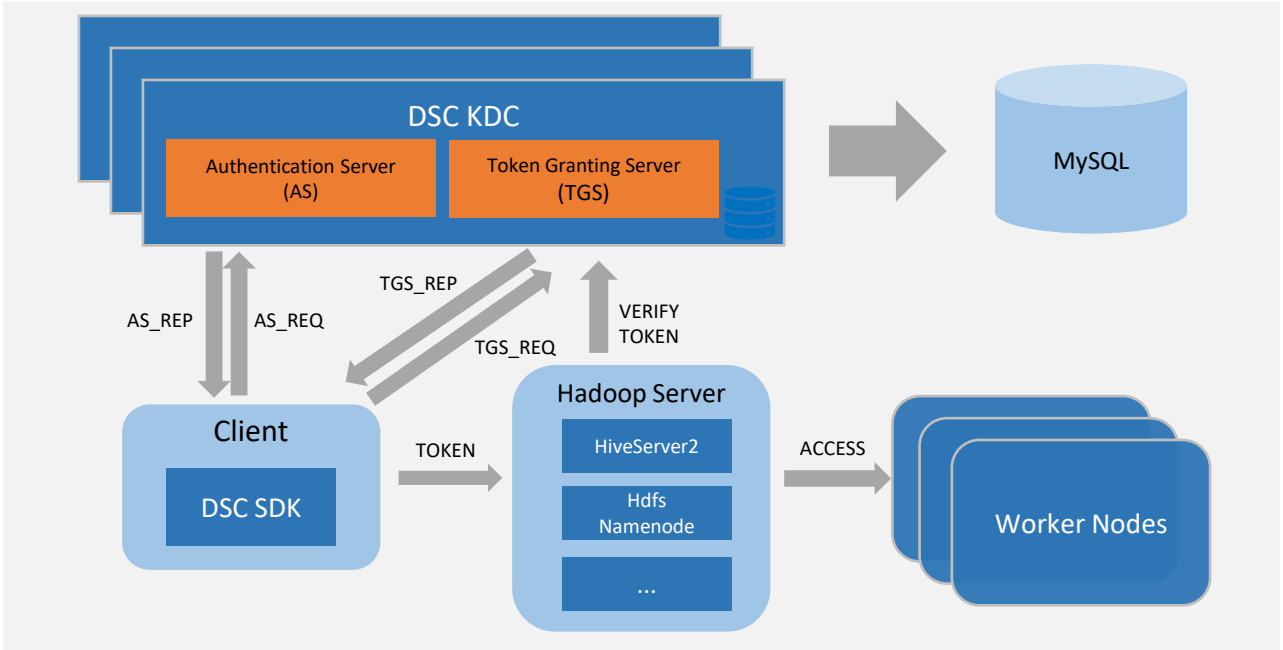
关键技术 – 认证体系

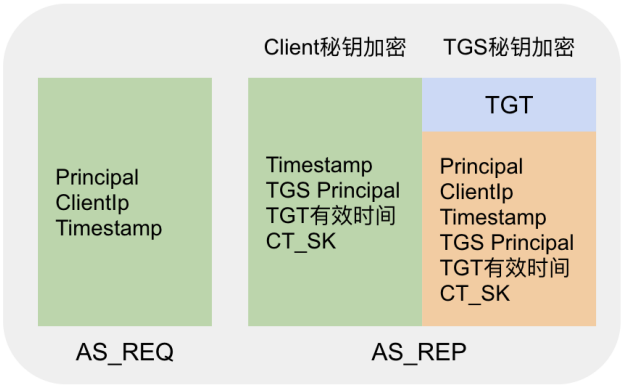
挑战

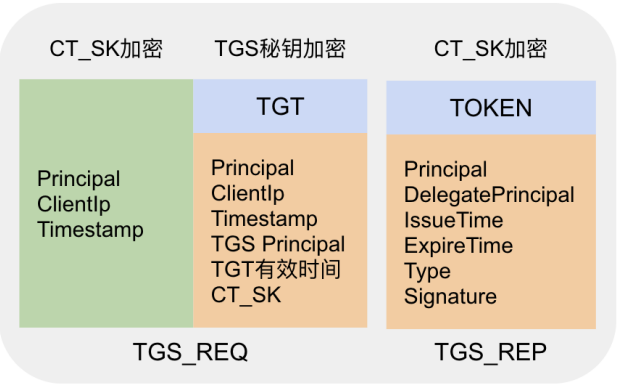
- **轻量级**：对现有接入系统入侵最小，对性能和稳定性影响小，原理简单具有良好的可解释性
- **本地化**：能够很好的与快手特有的生产组织管理体系相结合，相辅相成
- **易衍化**：能够很好的满足快手发展需求，尤其是大集群、国际化等

方案

- **账号类型**
  - **类型**：个人、项目组、代理账号
  - **表示**：使用principal表示
  - **格式**：principal_name/type@realm
- **令牌类型**：支持AccessToken、DelegateToken、DegradeToken



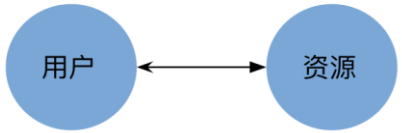




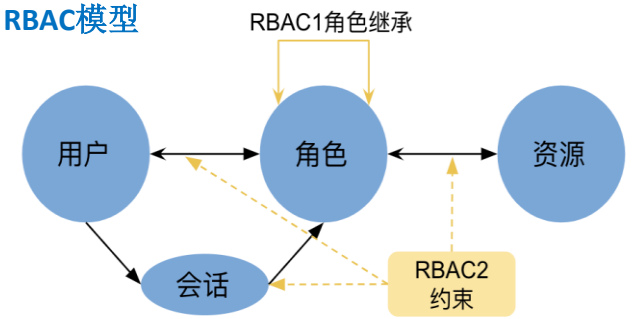
关键技术 – 权限模型

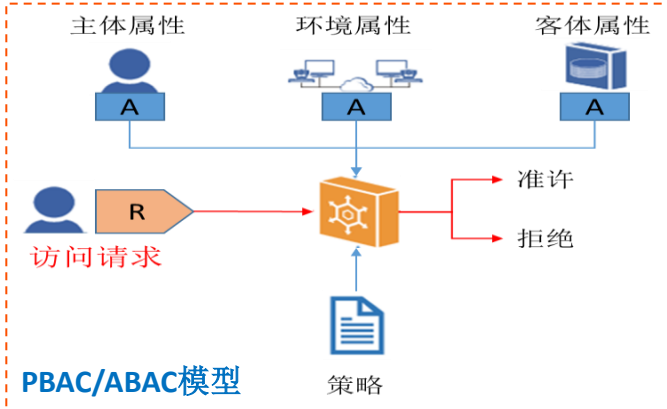
常见的权限模型

ACL模型



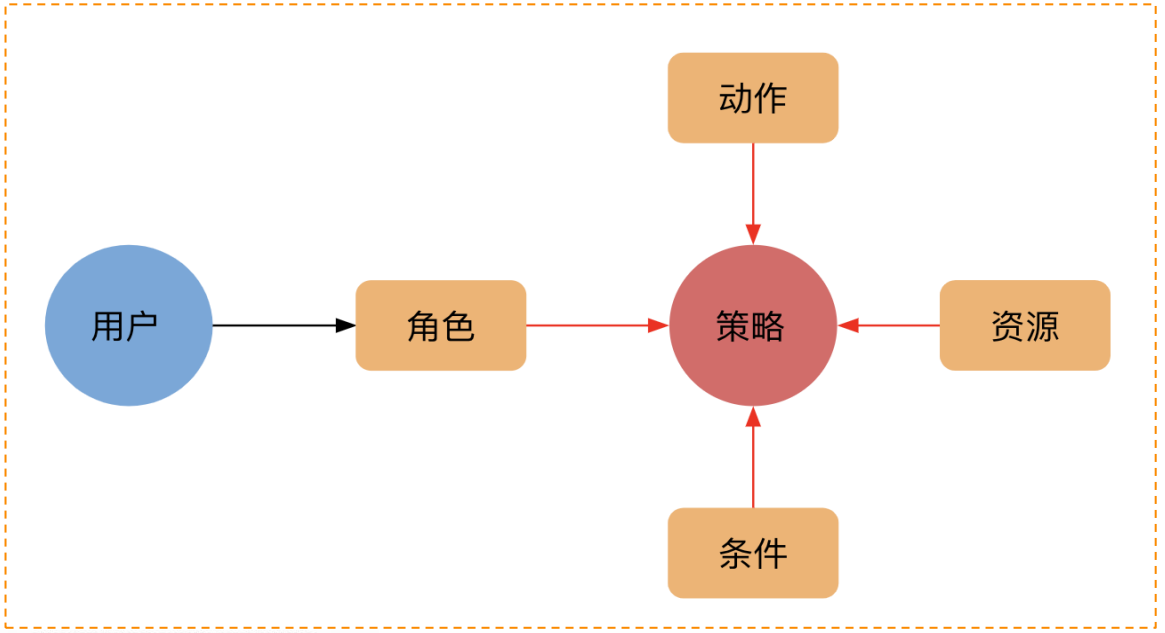
RBAC模型





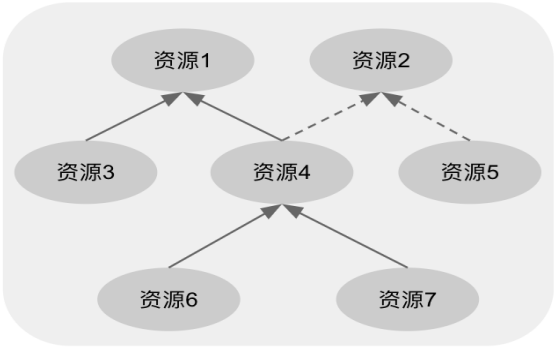
PBAC/ABAC模型

快手权限模型：组合RBAC和PBAC的自研的PRBAC权限模型



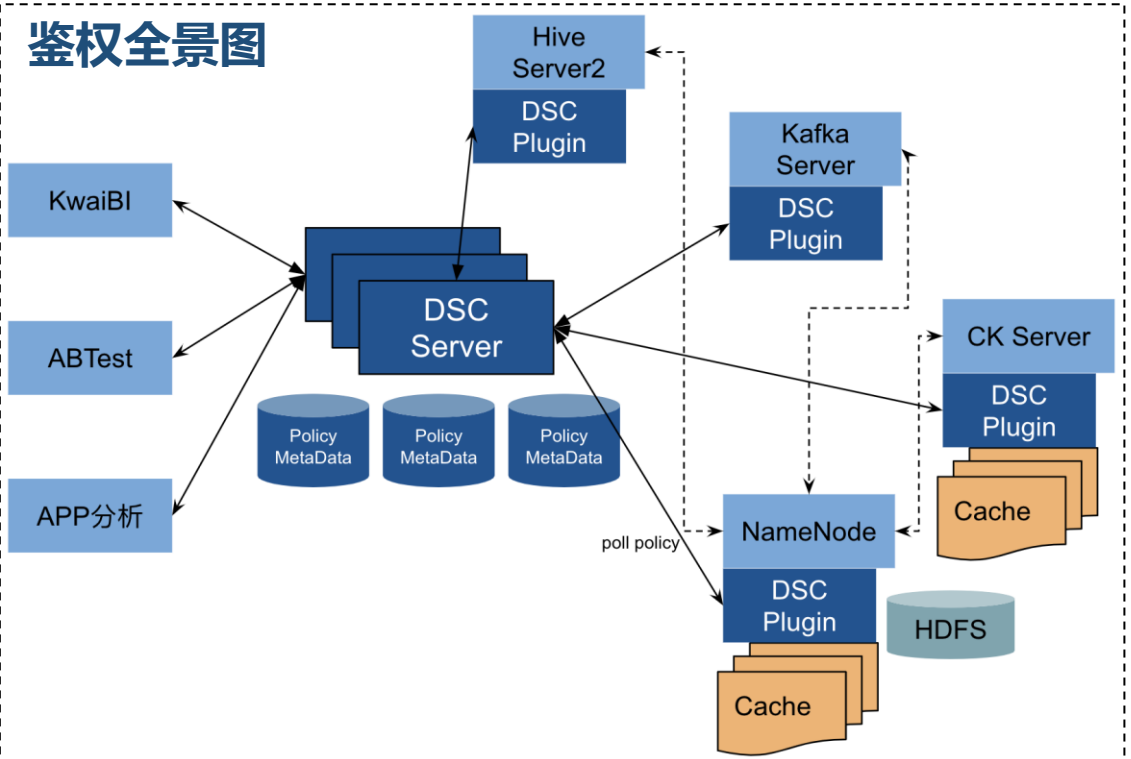
资源表示

- 全局唯一标识URN
- 三段式，由公司域、资源域和唯一ID构成

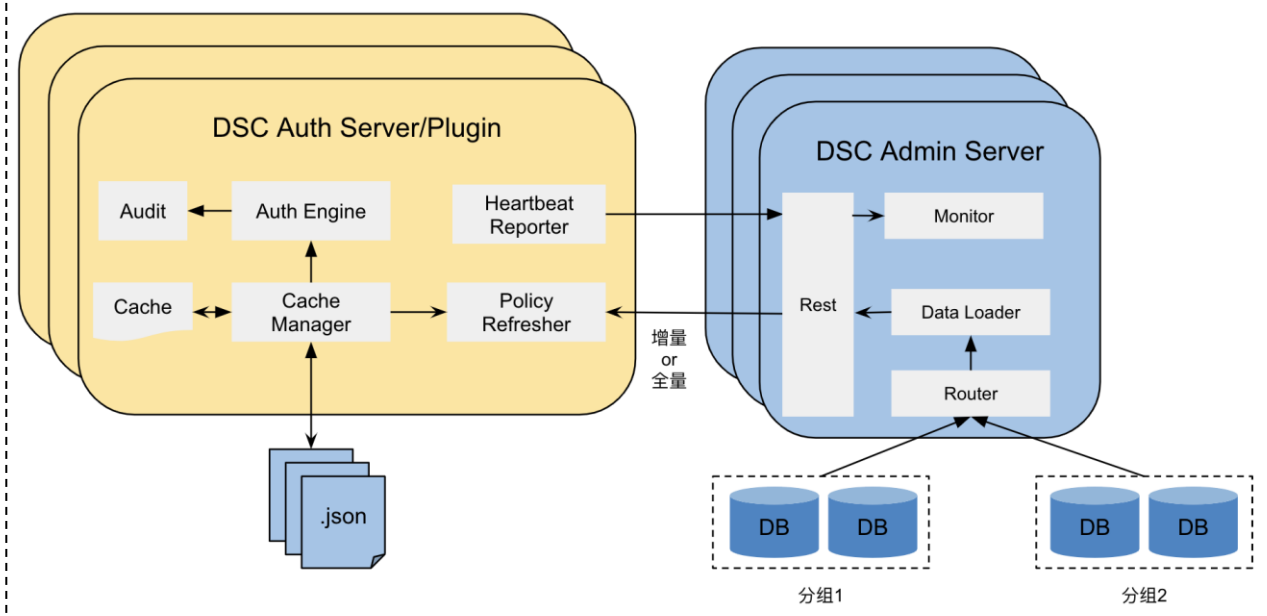


关键技术 – 统一鉴权

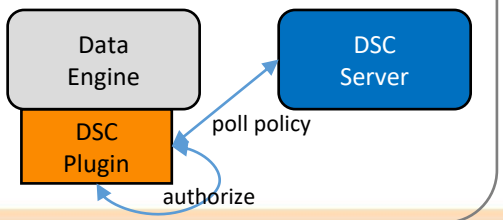
鉴权全景图



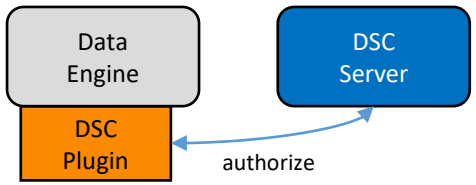
鉴权核心架构



鉴权模式1：本地鉴权



鉴权模式2：远程鉴权

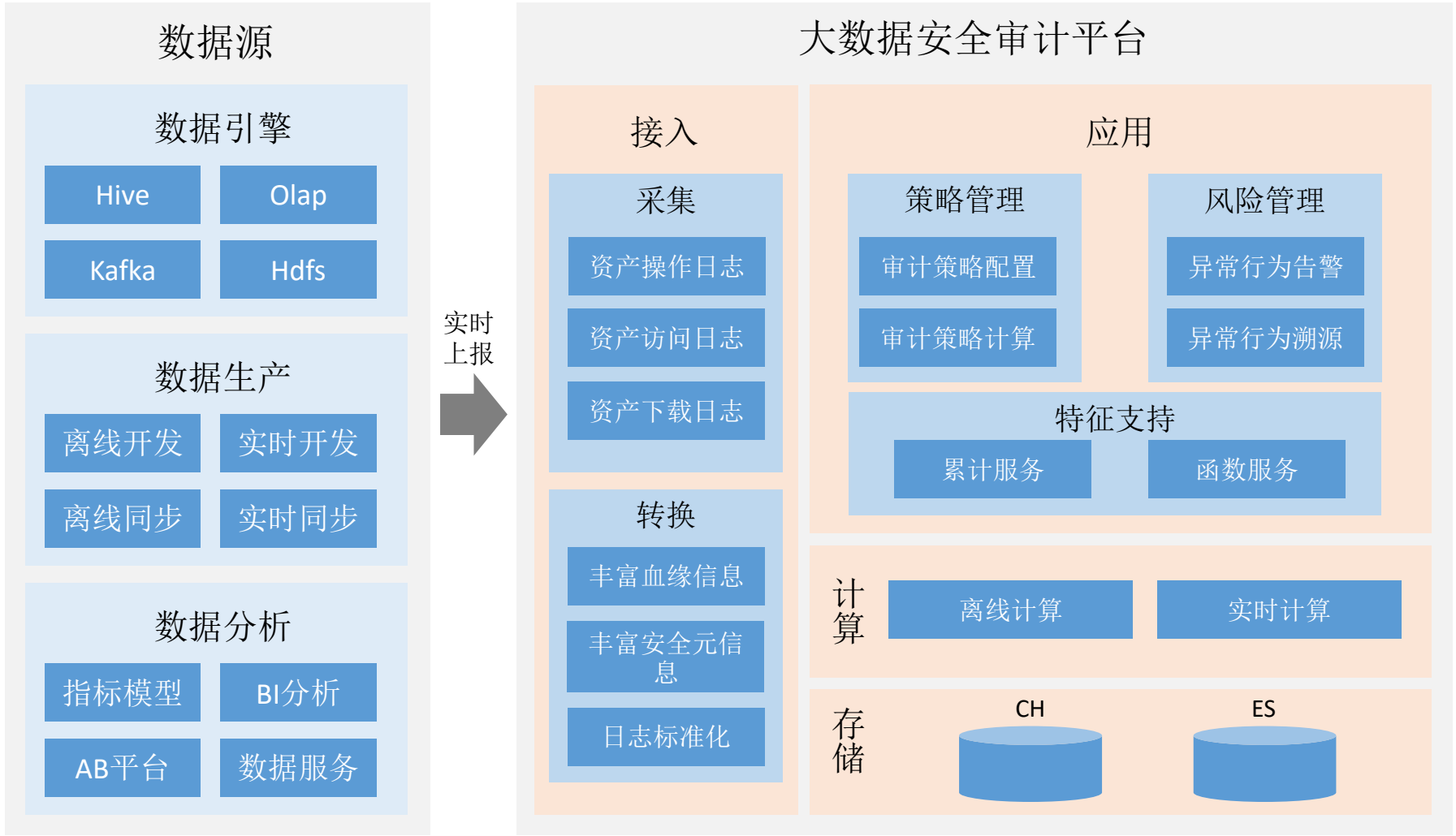


- **Auth Engine：** 鉴权引擎，负责鉴权模型的计算和策略规则的计算
- **Policy Refresher：** 负责策略的增量和全量的拉取
- **Cache Manager：** 负责鉴权服务本地缓存的管理，包括缓存的读写以及定时持久化到本地磁盘
- **Data Loader：** 负责从数据库加载策略相关的数据，并且根据路由策略查询不同的一组从库，做到存储的隔离

关键技术 – 全链路审计

特点

- 全链路覆盖
- 融合血缘信息
- 统一审计标准
- 风险识别告警



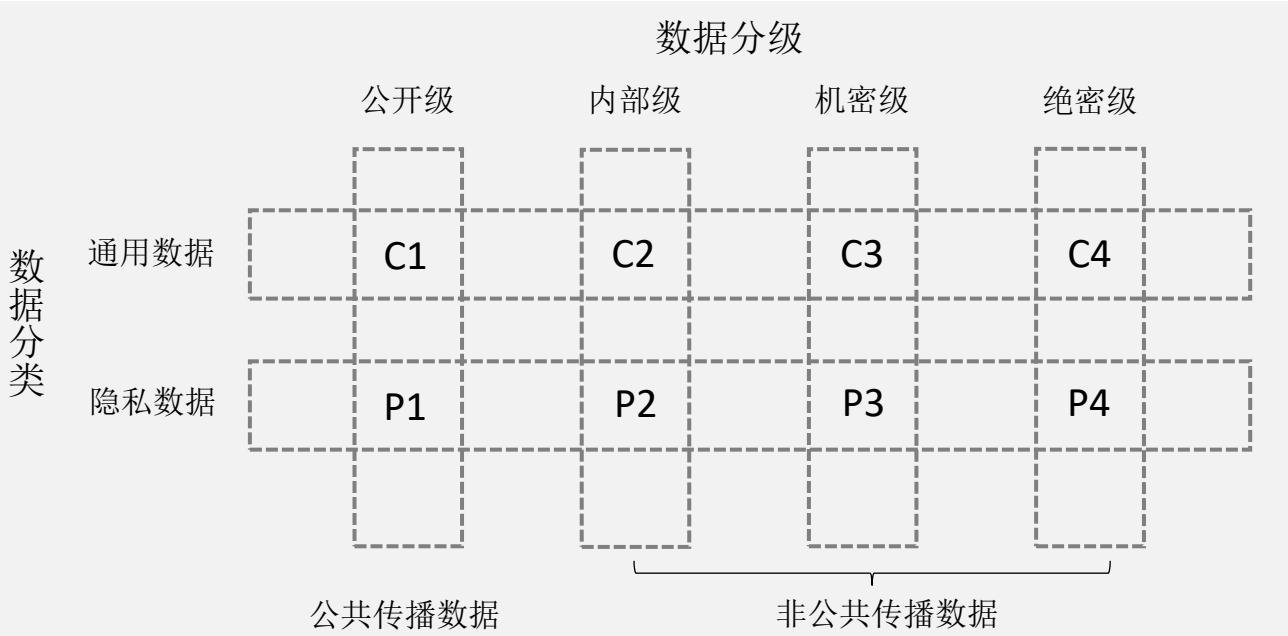
03

最佳实践

- 数据分类分级
- 数据引擎安全
- 敏感数据保护

数据分类分级 – 背景介绍

快手数据分类分级标准



快手数据分类分级原则

- 数据升级原则
- 数据降级原则
- 数据衍生原则

数据分类分级 – 解决方案

元数据采集

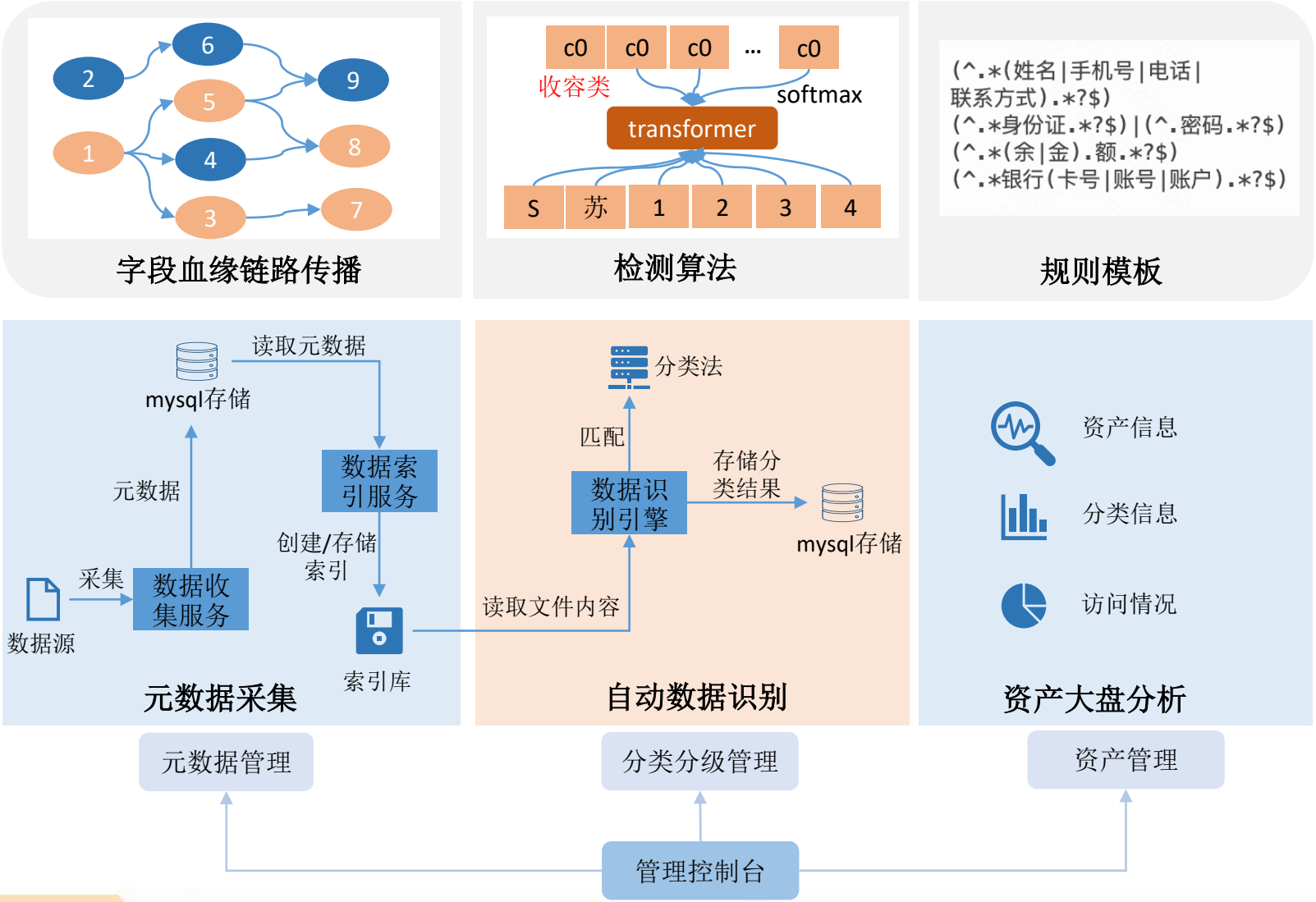
- **统一采集：**全链路资产的元信息统一上报到元数据
- **统一存储：**元数据及血缘信息统一存储至图数据库中

自动数据识别

- **血缘链路传播：**表/字段血缘继承
- **检测算法：**改进BERT模型、机器学习算法k-means、校验算法Luhn等
- **规则模板：**正则/关键字，内置50+个人敏感信息的识别规则模板

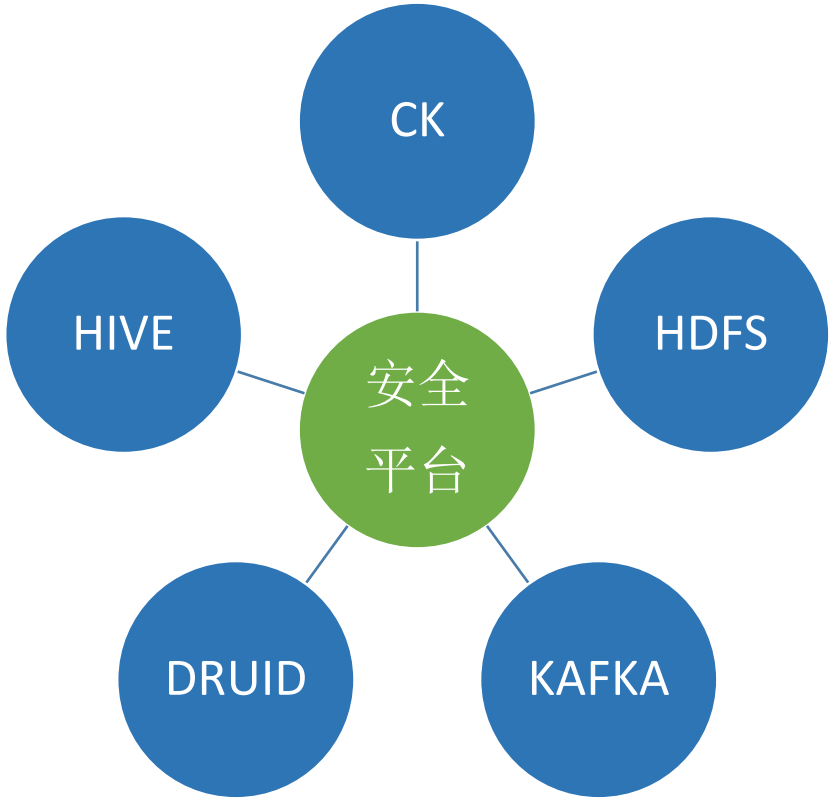
资产大盘分析

- **资产信息：**可以从个人、组织、部门三个视角查看不同级别资产的分布
- **分类信息：**展示资产各个分类的数量
- **访问情况：**高频访问资产的分类分级分布



数据引擎安全 – 问题及挑战

接入五大引擎



挑战

管理规范

1. 组织管理体系不清晰，账号体系未建设
2. 资产归属不清晰，无法定义资产的安全管理角色
3. 没有多租户的权限管理规范

安全能力

1. 身份认证能力缺失，没有安全审计及溯源能力
2. 没有权限控制，用户可查询任意数据，安全风险巨大
3. 数据引擎鉴权对平台的性能和稳定性要求高

运营治理

1. 引擎的查询无法定位到真实访问用户，导致推动用户改造困难
2. 各使用方平台领域知识复杂，导致沟通协作困难
3. 用户需求多样，需要支持灵活多样的灰度和降级策略

数据引擎安全 – 解决方案

规范

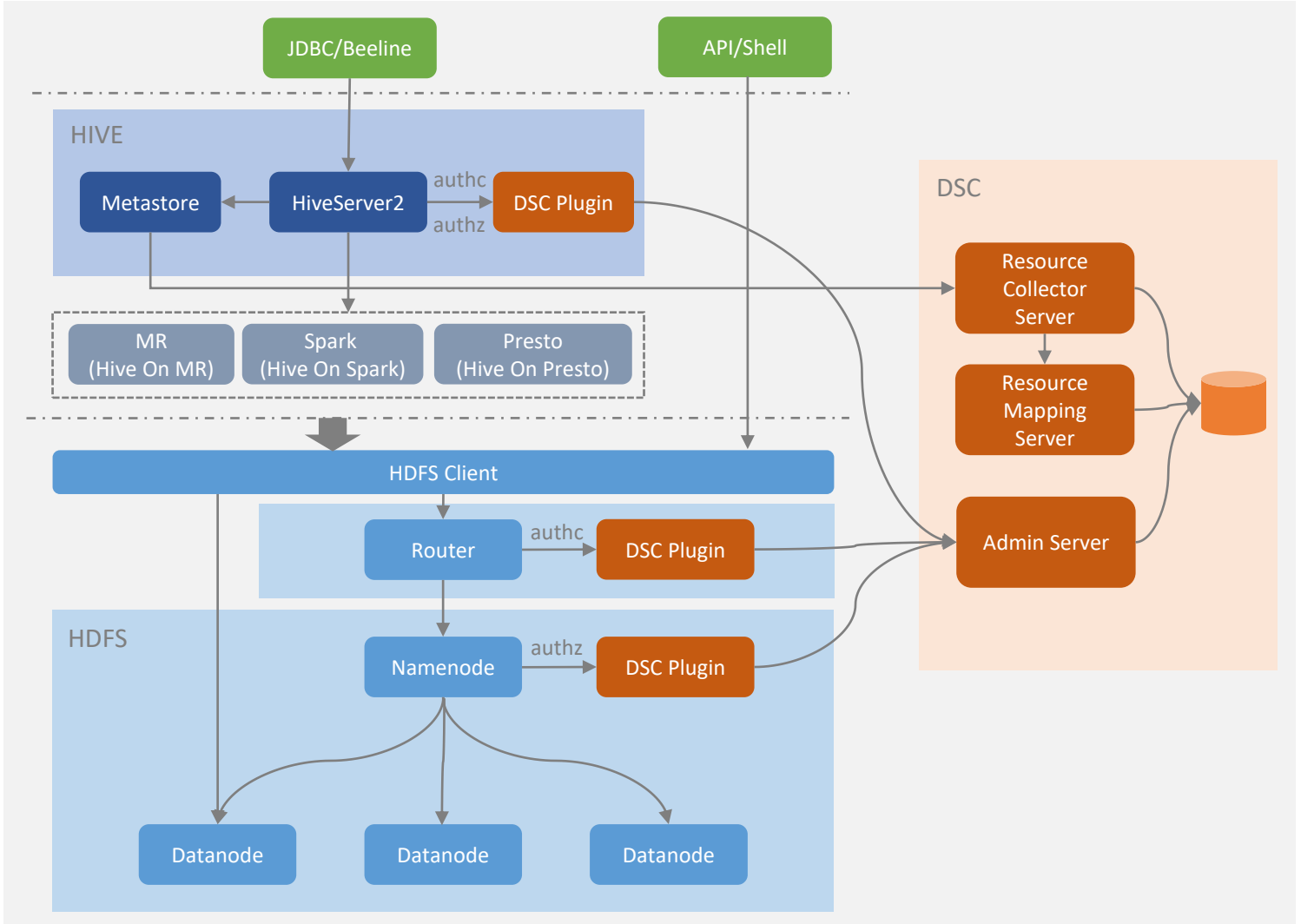
- **账号体系：**提供个人账号、生产账号和代理账号三种类型
- **管理角色：**安全接口人、租户管理员、项目组管理员和权限负责人四种角色
- **权限隔离：**租户之间权限隔离；租户的权限由归属和申请两种获权方式

工具

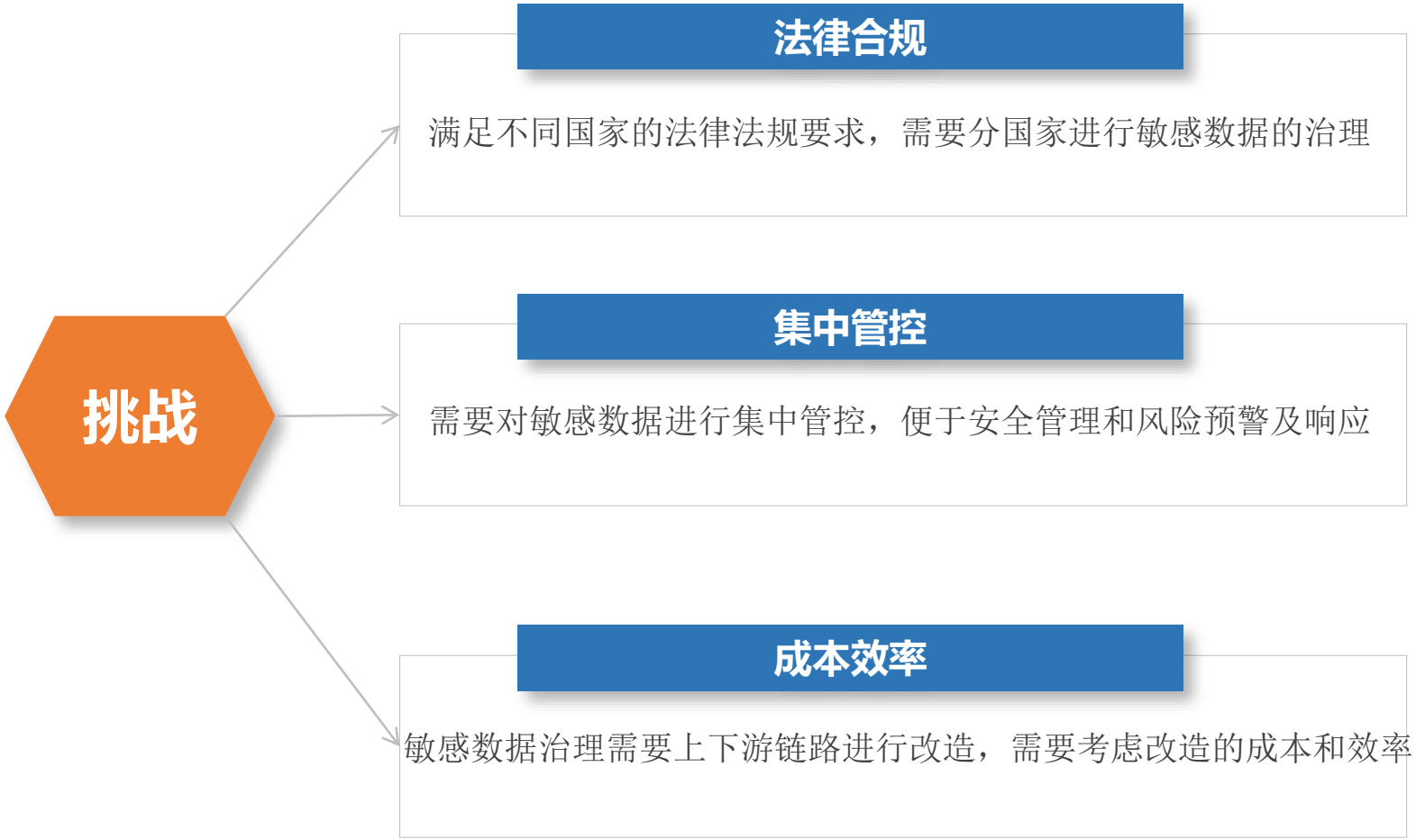
- **产品能力：**SQL类引擎行列级权限；租户体系的多种管控模式
- **鉴权模型：**HDFS及之上的其它引擎分层独立进行访问控制；大账号机制
- **安全元信息：**具有血缘关系的资产，安全元信息联动
- **鉴权plugin：**通用的鉴权插件，提升鉴权计算性能

治理

- **头部平台：**头部平台用户进行重点沟通
- **长尾运营：**采用多种运营渠道触达用户
- **灰度封禁：**丰富灵活的封禁策略



敏感数据保护 – 问题及挑战



敏感数据保护 – 解决方案

规范

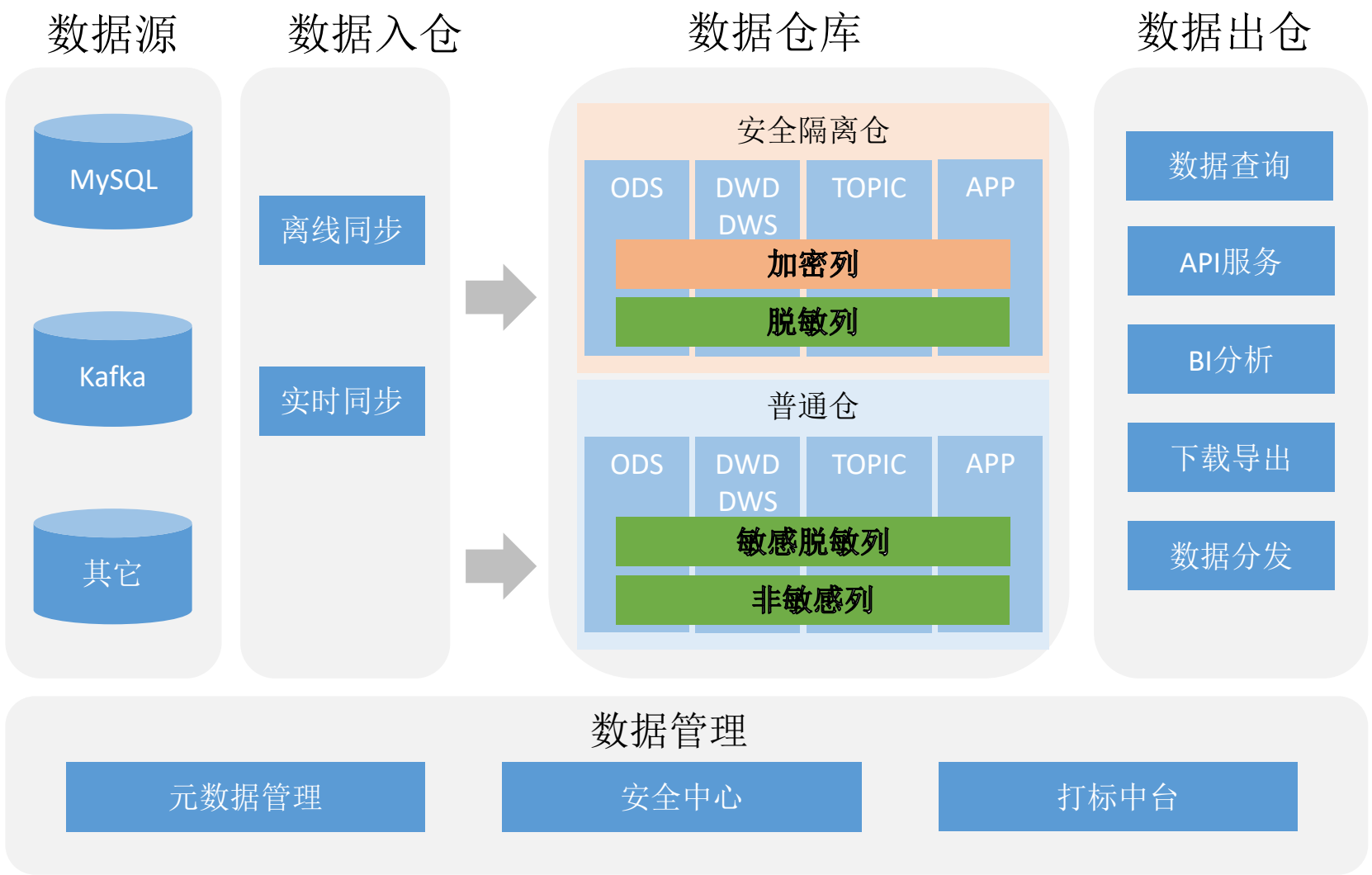
- **国内外高敏感信息：**根据国内外法律法规梳理高敏感个人信息
- **国内外脱敏规范：**定义各类敏感数据的脱敏方式和要求

工具

- **数据识别：**高敏感数据识别、文件/字段级加密、脱敏
- **数据保护：**字段级权限控制、严格的审批流程、安全隔离仓、精细管控模式
- **数据检测：**代码检测、数据内容扫描、下载监控
- **数据响应：**数据泄露应急预案、全链路异常监控告警及溯源

治理

- **存量治理：**上下游链路生产任务改造优化、存量及增量数据重刷
- **增量治理：**日常敏感数据识别、治理跟进、工具沉淀



04

总结与规划

- 成果总结
- 未来规划

成果总结

30+	千万级	百万级	千级
接入系统数	资源数	授权数	日均申请量

数据应用	BI系统	APP分析	AB系统	报表系统	… (20+系统)
数据开发	数据同步	服务化平台	数据开发平台	数据地图	指标管理
数据架构	HIVE	DRUID	CK	KAFKA	HDFS

未来规划

- ✓ 推动底层引擎的使用方100%接入认证和鉴权

覆盖度

态势感知

- ✓ 对数据资产分布、敏感数据访问行为进行多维度全方位分析，对异常行为进行检测

- ✓ 探索前沿的隐私保护技术，比如联邦学习、安全多方计算等，增强隐私数据保护，真正做到“数据的可用不可见”

新技术

智能化

- ✓ 通过机器学习算法实现数据的智能分类分级
- ✓ 持续提升数据分类分级的准确性

欢迎交流

快手数据中台公众号

快手大数据公众号

快手技术团队公众号


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
开务数据库
UbiSQL
神舟通用
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS