基础数据建设实践

美丽说 高玉石

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

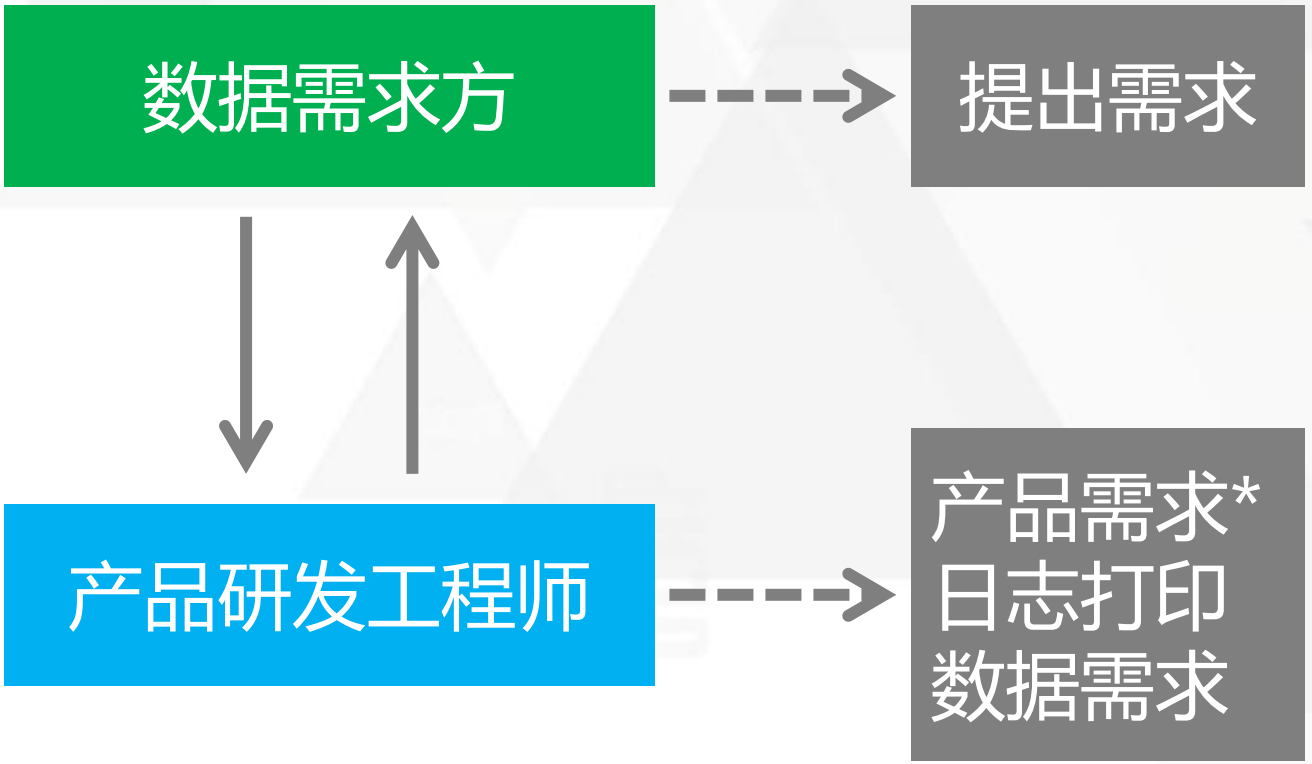
大数据技术探索和价值发现

互联网+

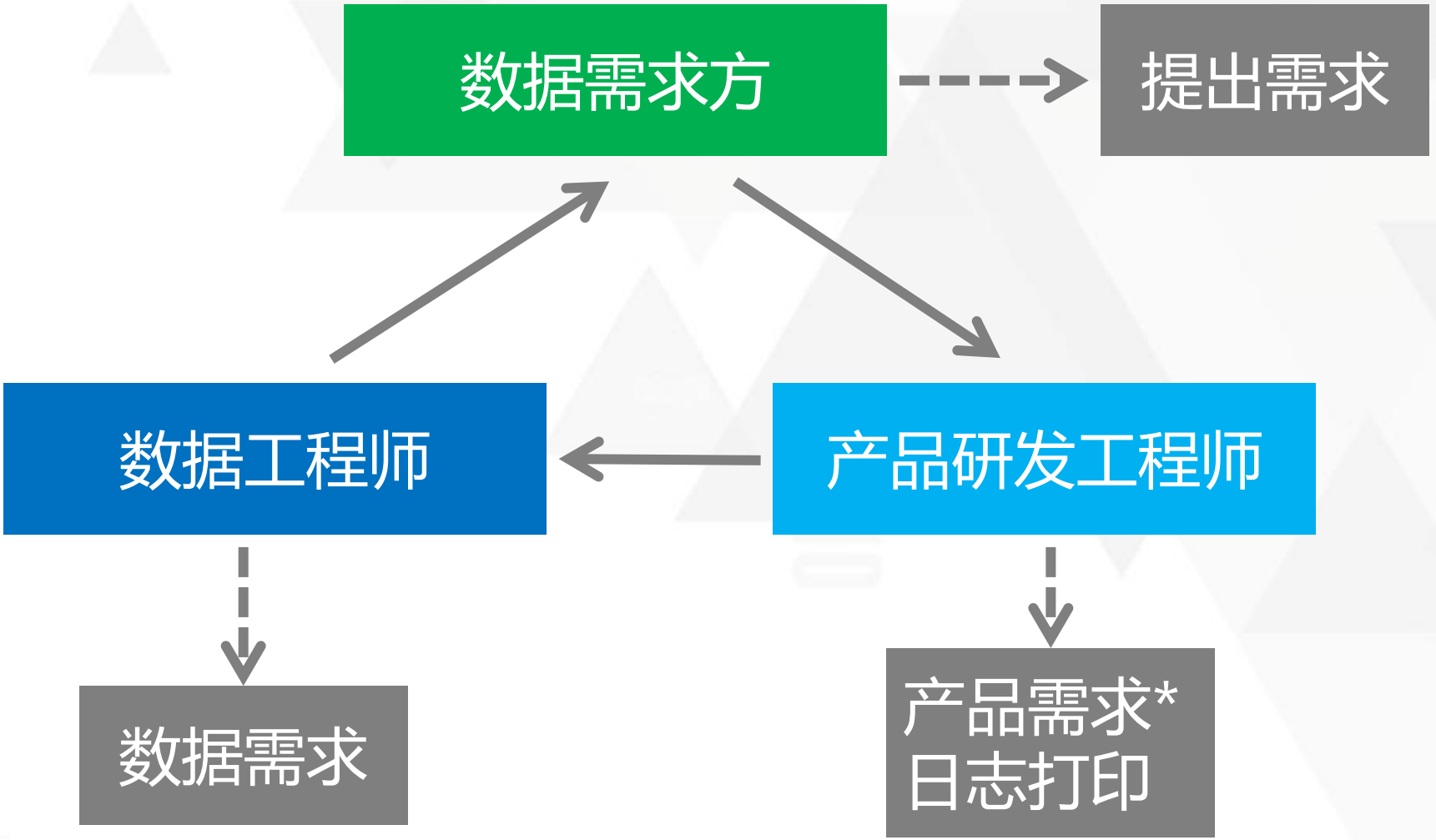
大数据

基础数据建设演进

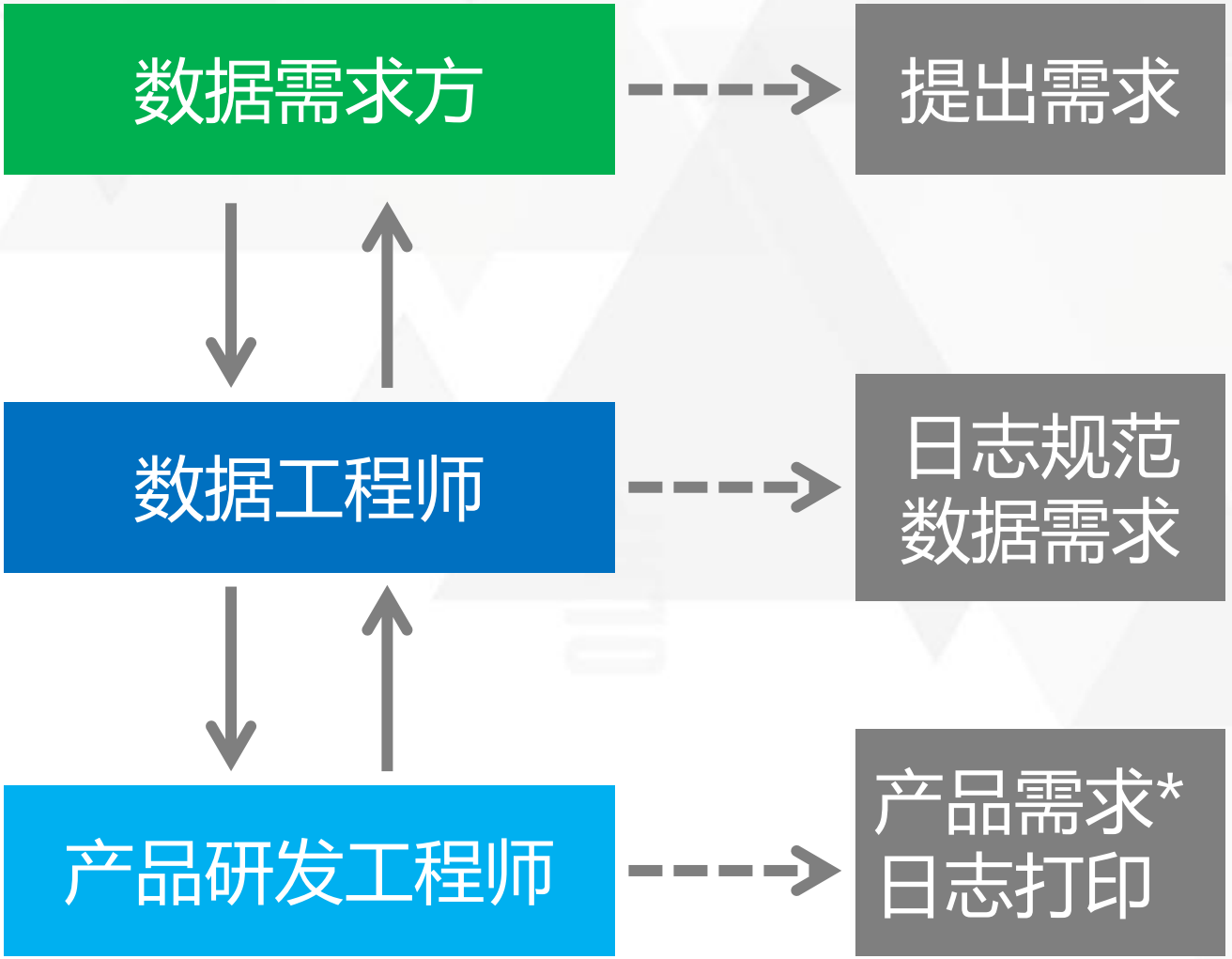
原始社会



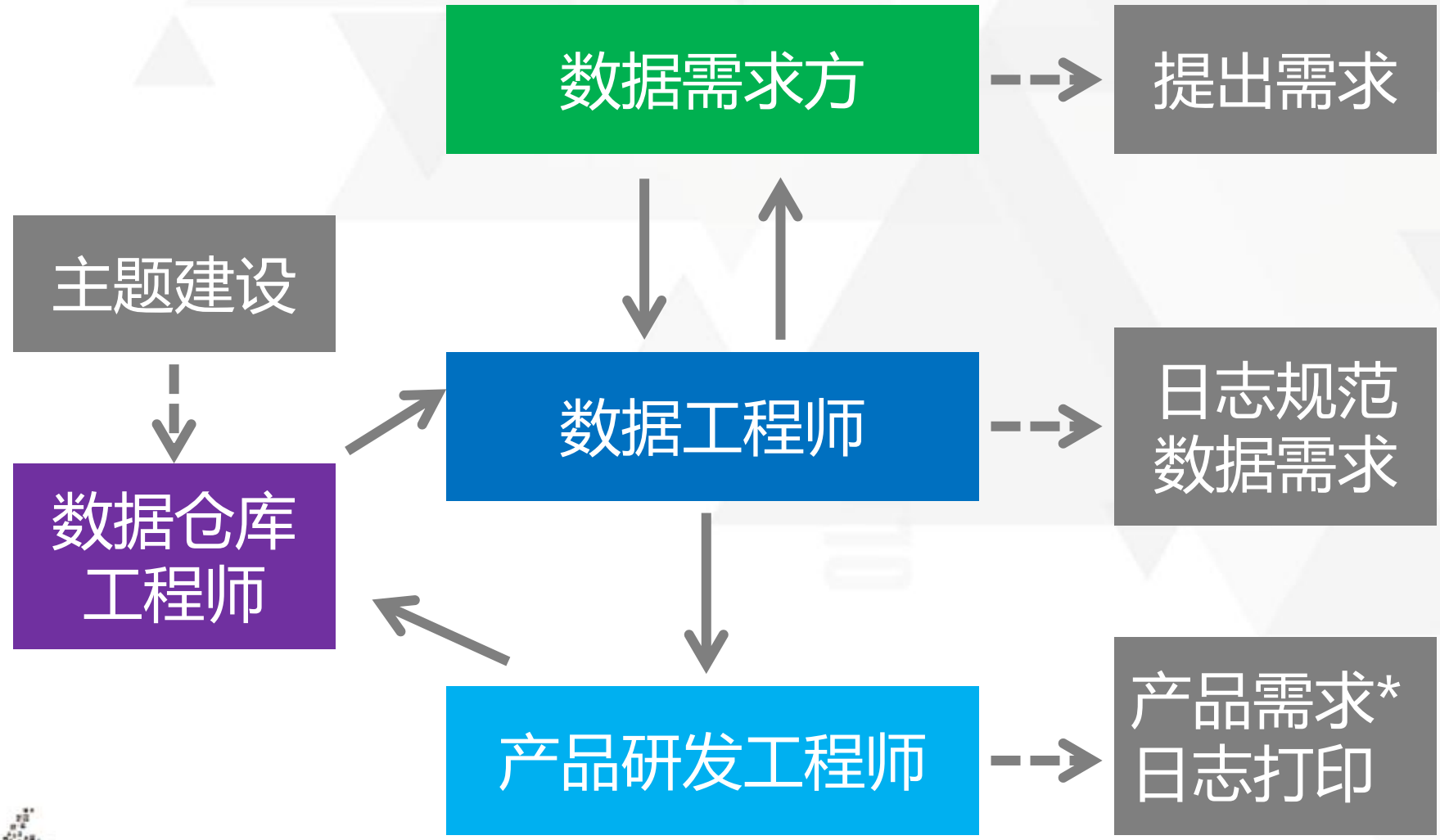
奴隶社会



封建社会

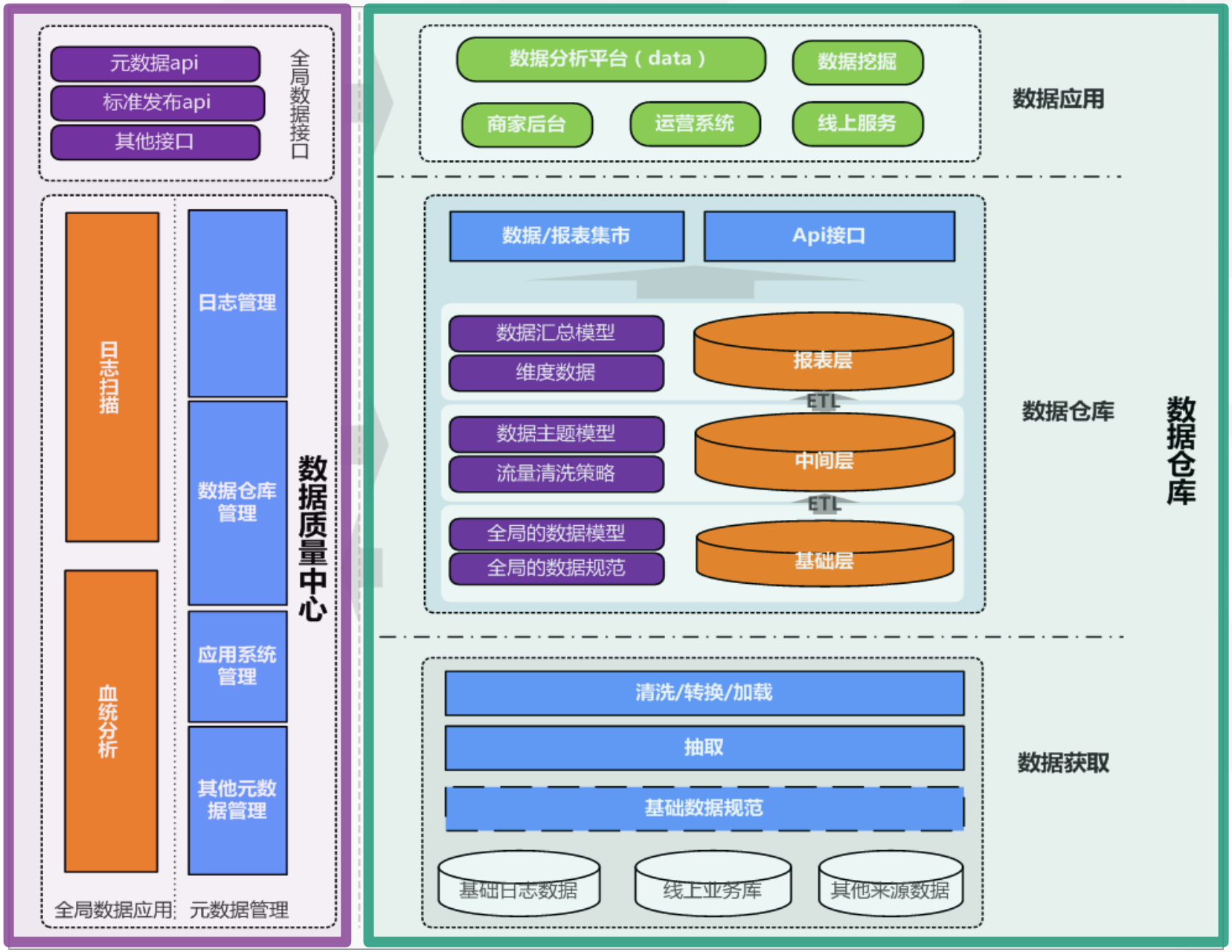


资本主义社会



共产主义社会

数据仓库+


元数据api
标准发布api
其他接口
全局数据接口
数据分析平台（data）
数据挖掘
商家后台
运营系统
线上服务
数据应用
数据/报表集市
Api接口
数据汇总模型
维度数据
报表层
ETL
数据主题模型
流量清洗策略
中间层
ETL
全局的数据模型
全局的数据规范
基础层
数据仓库
数据仓库
清洗/转换/加载
抽取
基础数据规范
基础日志数据
线上业务库
其他来源数据
数据获取
日志扫描
血统分析
日志管理
数据仓库管理
应用系统管理
其他元数据管理
数据质量中心
全局数据应用
元数据管理

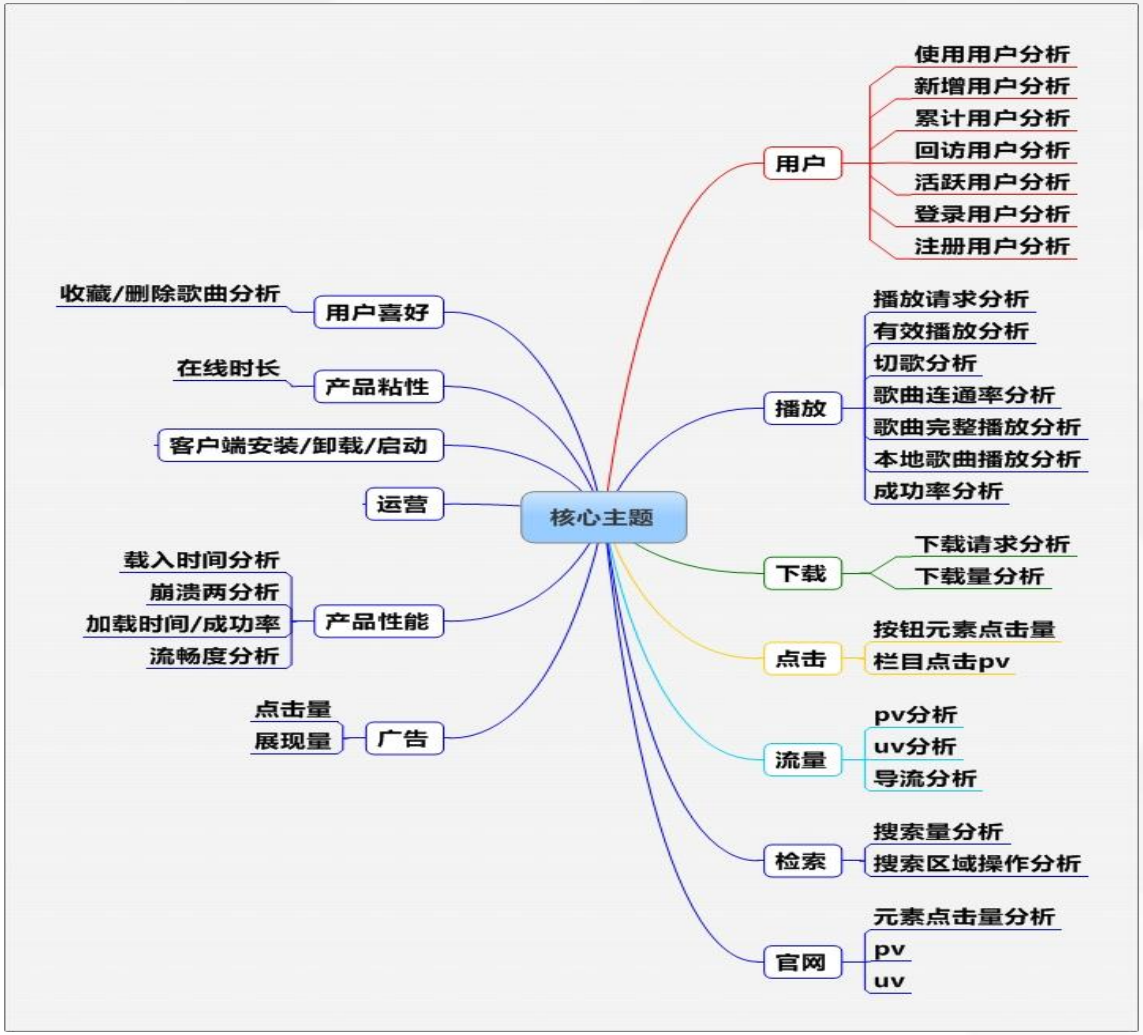
数据仓库建设


日志数据
数据库数据
文档数据
…
业务范围
日志管理
业务划分
…
结构化
扩展规范
数据标注
…
基础层

要点

- 主题划分
- 一致性
- 粒度VS灵活
- 维度模型

数据仓库建模



主题划分

高度汇总

指标唯一

数据一致

需求驱动

数据分析平台

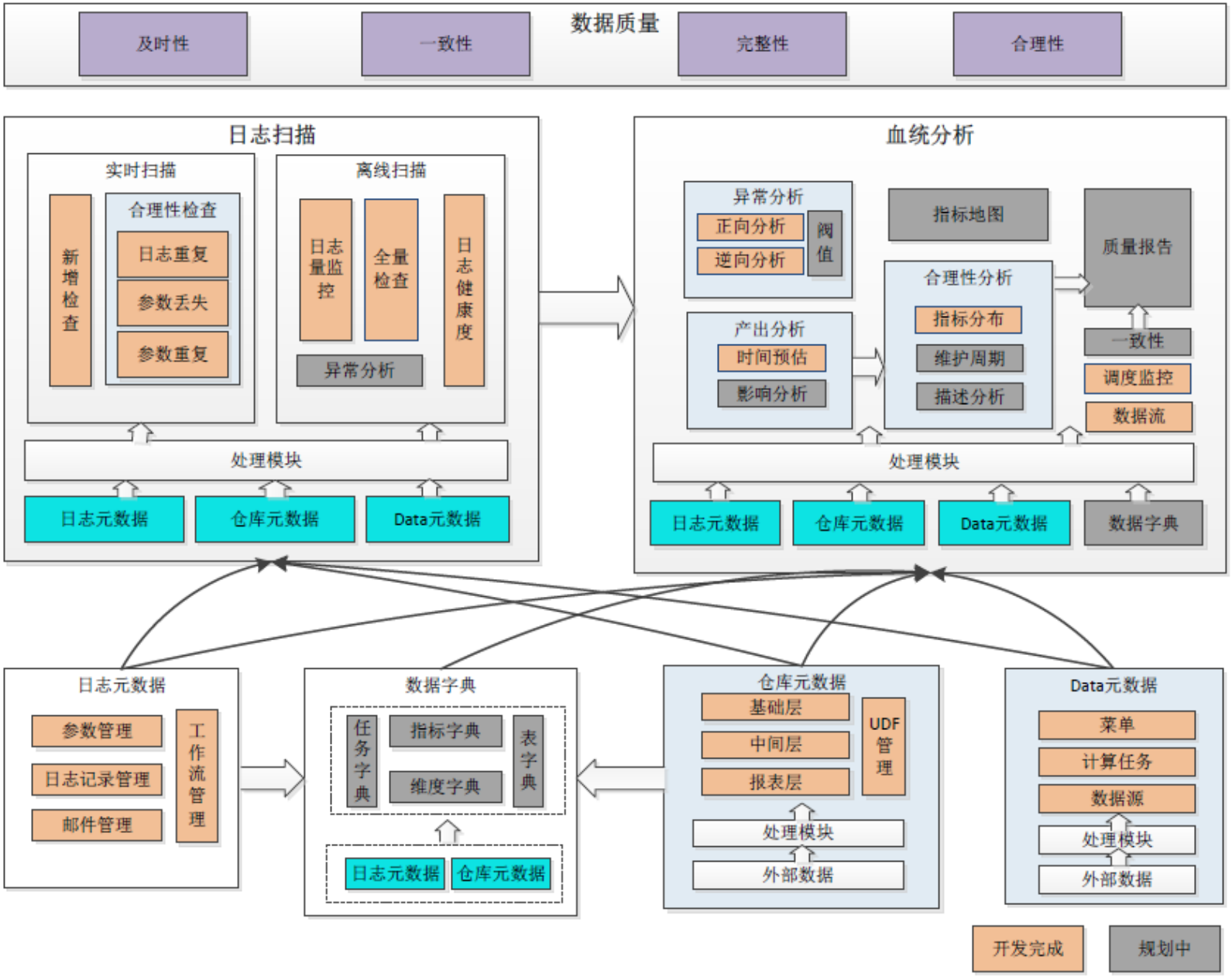
数据挖掘

线上服务

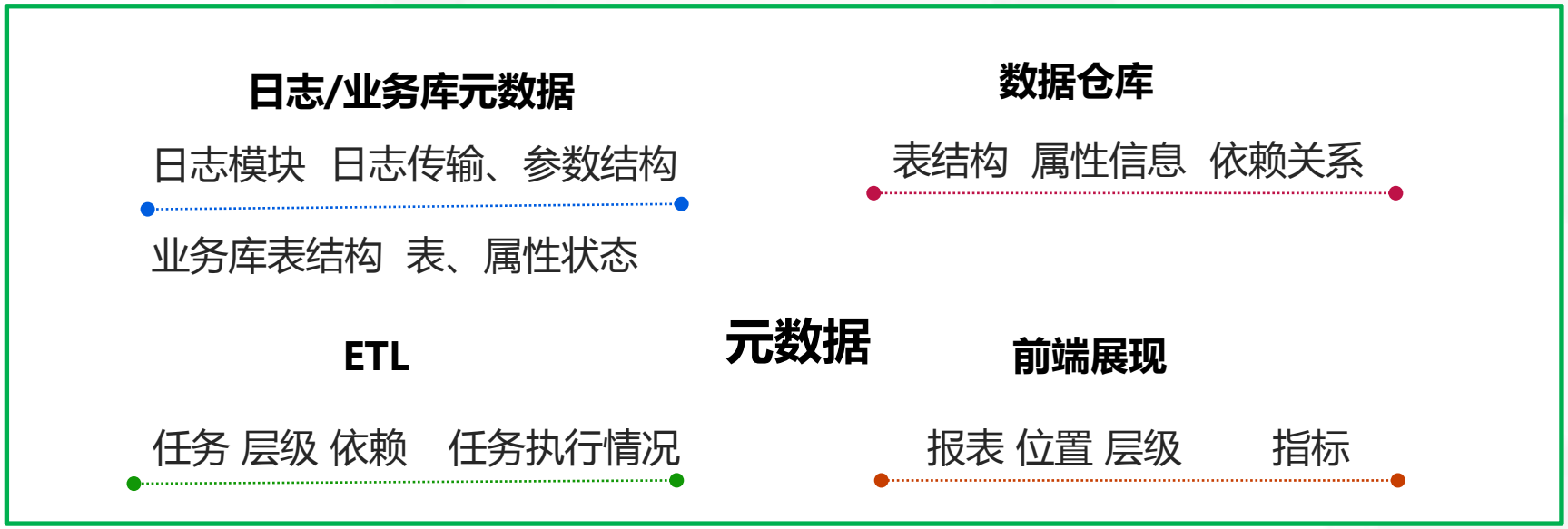
...

报表层

数据质量建设


数据质量
及时性
一致性
完整性
合理性
日志扫描
实时扫描
新增检查
合理性检查
日志重复
参数丢失
参数重复
离线扫描
日志量监控
全量检查
异常分析
日志健康度
处理模块
日志元数据
仓库元数据
Data元数据
血统分析
异常分析
正向分析
逆向分析
阀值
指标地图
质量报告
合理性分析
指标分布
维护周期
描述分析
产出分析
时间预估
影响分析
一致性
调度监控
数据流
处理模块
日志元数据
仓库元数据
Data元数据
数据字典
日志元数据
参数管理
日志记录管理
邮件管理
工作流管理
数据字典
任务字典
指标字典
维度字典
表字典
日志元数据
仓库元数据
仓库元数据
基础层
中间层
报表层
UDF管理
处理模块
外部数据
Data元数据
菜单
计算任务
数据源
处理模块
外部数据
开发完成
规划中

元数据管理


日志/业务库元数据
日志模块 日志传输、参数结构
业务库表结构 表、属性状态
数据仓库
表结构 属性信息 依赖关系
元数据
ETL
任务 层级 依赖 任务执行情况
前端展现
报表 位置 层级 指标

元数据管理


表基本信息 | 表字段信息 | 表维度信息
名称
music_mds_songid_bhv_day
所属层级
中间层
主题
行为主题
终端类型
不分端
抽取周期
天
来源表
music_mds_bhv_play,music_dim_quku_song_si,music_dim_quku_resource_sf,music_dim_quku_song_sl
对应程序
处理逻辑
从music_mds_bhv_play统计出每天各端的全量歌曲播放量，以songid作为主键和歌曲全量信息表全关联。歌曲全量信息表是指以songid作为主键将music_dim_quku_song_si,music_dim_quku_resource_sf,
music_dim_quku_song_sl关联到一起取出每首歌曲歌名，音频，歌词等信息。
功能描述
songid四端全线播放表
修改

日志扫描

类型

- 离线日志扫描
- 实时日志扫描

技术

- 实时计算
- 规则引擎

分析内容

- 日志错、漏、重复、异常等
- 日志使用情况
- 日志冗余

日志扫描

扫描结果报告

一. 扫描结果 (下载日志)

扫描完成时间： 2014-05-05 10:19:01

扫描成功,结果正常： 成功扫描到该字段，详情请关注《三》《四》《五》！

二. 扫描基本情况

主扫描字段: < albu

扫描规则: 无明确

扫描开始时间: 2014-0

三. 初级扫描结果详情

成功扫描到主字段(

详情日志如下：

第1条：

61.172.247.198 - - 05/May/2

hello&singer_id=1&singer_

_version=&linkCode=2200

ga_201404251715&userna

xcode=26cb151a204b141f8

ertype=lowflow&prestatus=

__m=mboxCtrl.playSong&_

NBID=6D0DEC6BB793873

四. 高级扫描结果详情

1.暂未发现有重复字段的现象发生！

2.暂未发现有重复记录的现象发生！

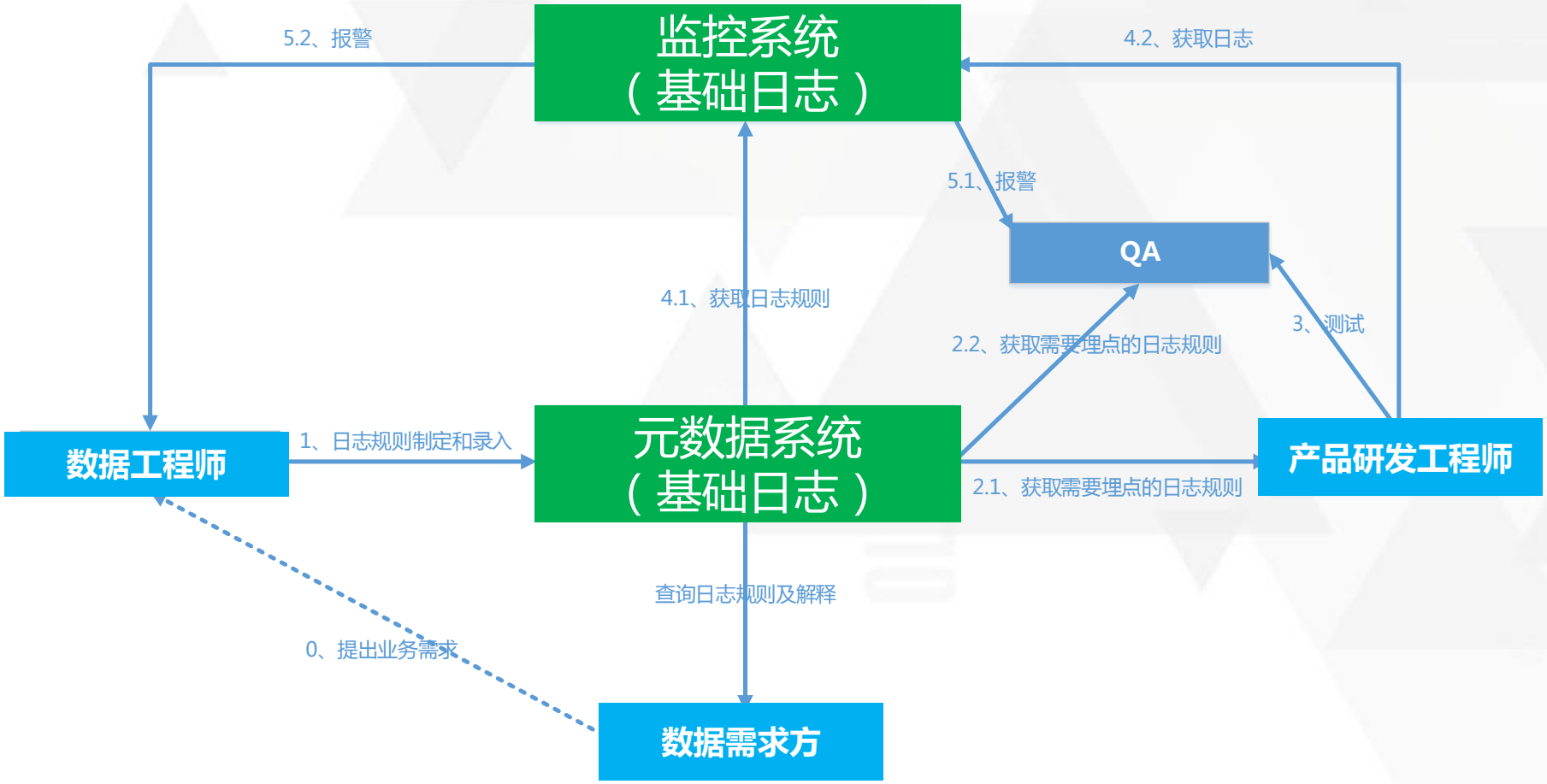
五. 基础扫描字段结果分析

总条数为：501条,基础字段的详细扫描结果分析如下表：

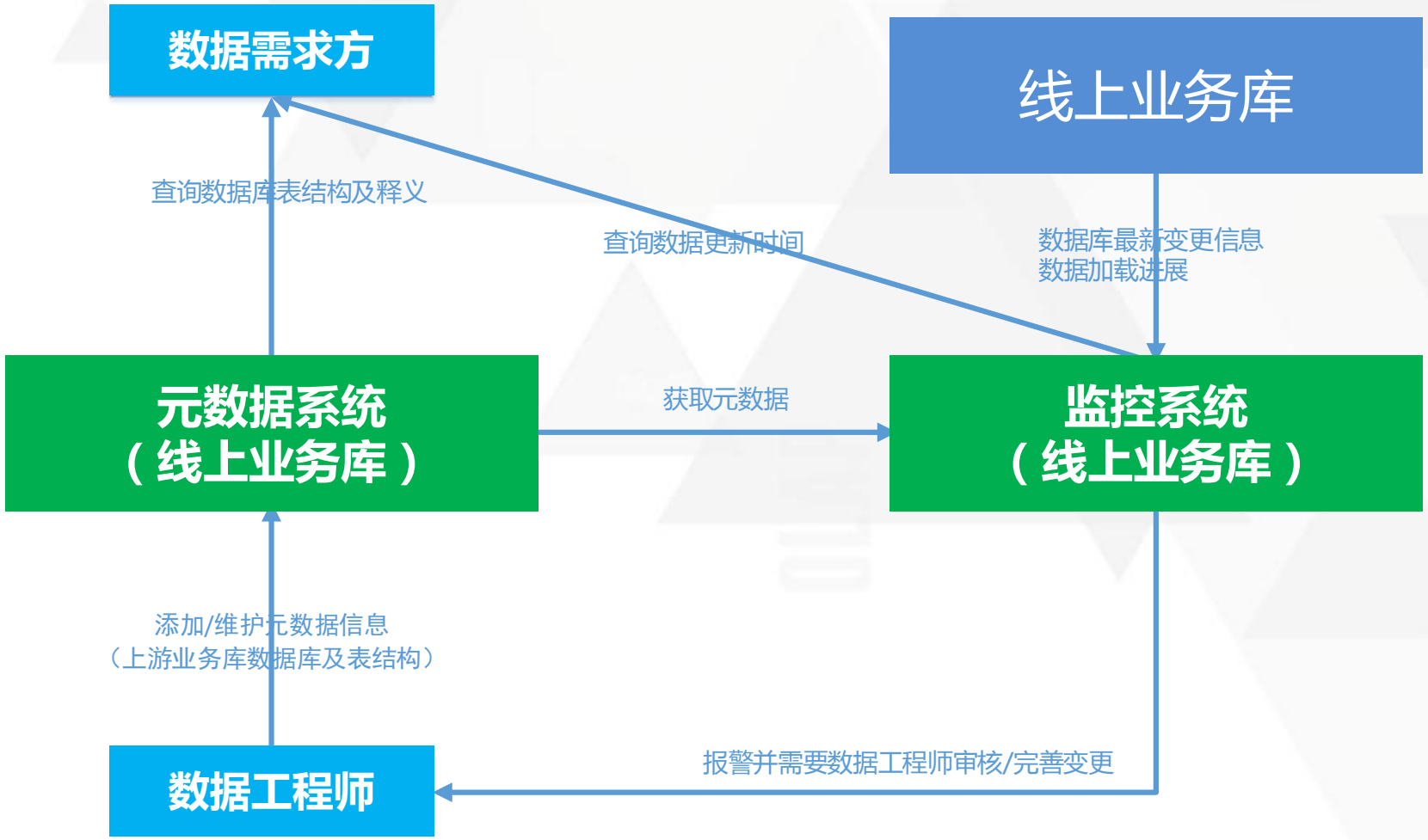
参数	规则	空值率	扫描结果分析（占比=存在数/总数）	操作
pid	=	0	正常(value=304)的条数：501条————占比：100%	查看详细日志
flow	in	0	未扫描到该字段的条数：98条————占比：19.56%	查看详细日志
			正常(value=ad_voice_1)的条数：403条————占比：80.44%	查看详细日志
ref	in	0	正常(value=tingbox)的条数：403条————占比：80.44%	查看详细日志
			正常(value=radio)的条数：85条————占比：16.97%	查看详细日志
			正常(value=aldsinglebox)的条数：12条————占比：2.4%	查看详细日志
			正常(value=pc_widgethao123)的条数：1条————占比：0.2%	查看详细日志

返回上层

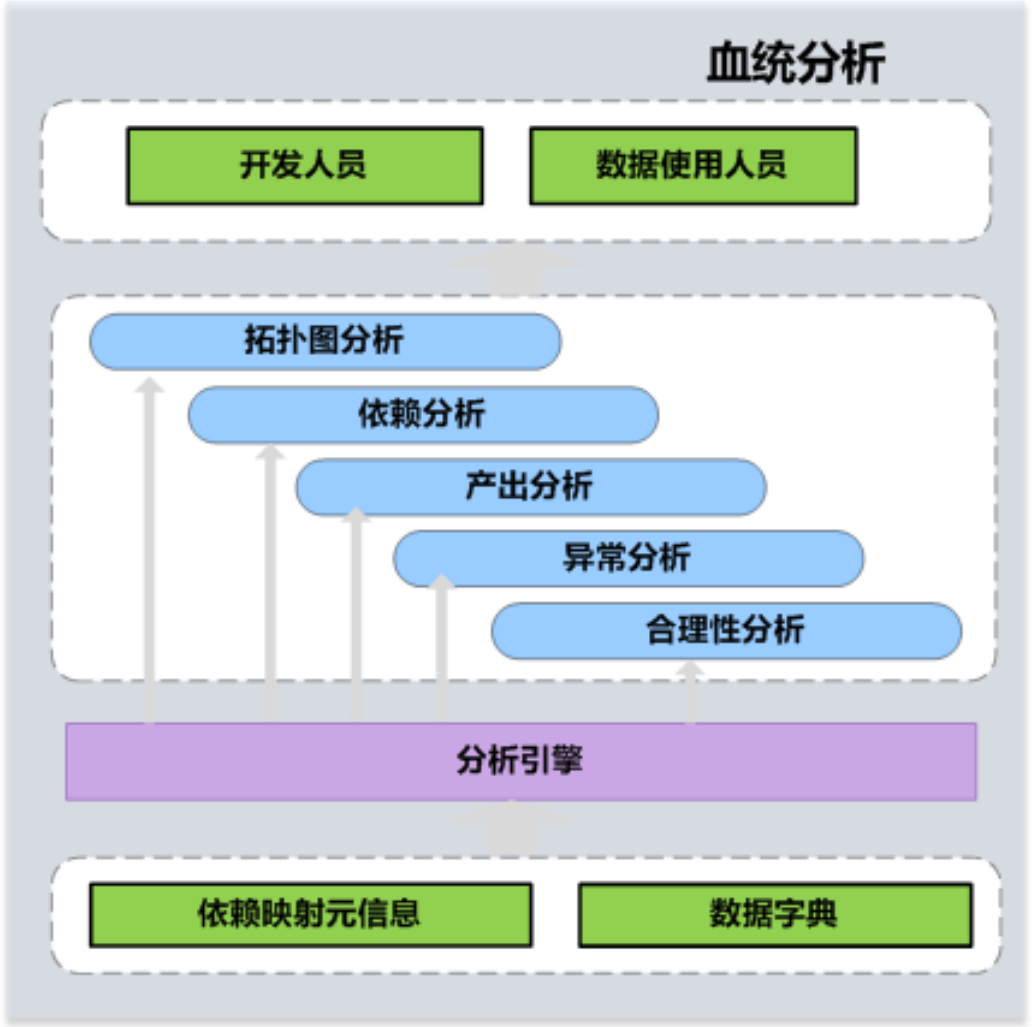
日志规范流程


监控系统
（基础日志）
5.2、报警
4.2、获取日志
5.1、报警
QA
4.1、获取日志规则
3、测试
2.2、获取需要埋点的日志规则
数据工程师
1、日志规则制定和录入
元数据系统
（基础日志）
产品研发工程师
2.1、获取需要埋点的日志规则
查询日志规则及解释
0、提出业务需求
数据需求方

业务库规范流程

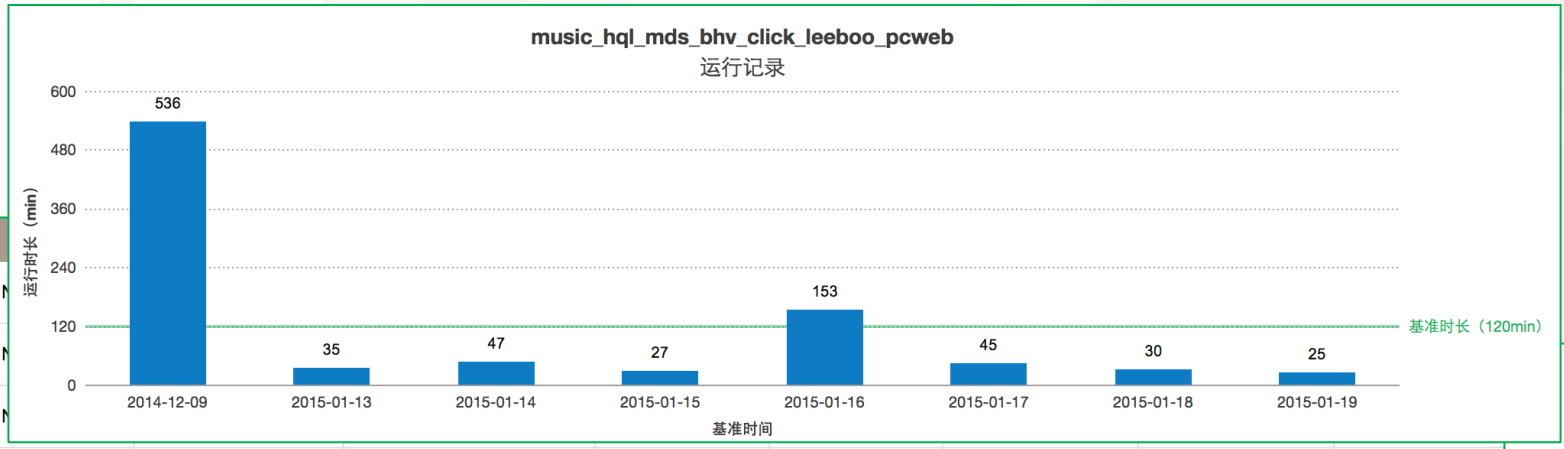

数据需求方
线上业务库
查询数据库表结构及释义
查询数据更新时间
数据库最新变更信息
数据加载进展
元数据系统
（线上业务库）
获取元数据
监控系统
（线上业务库）
添加/维护元数据信息
（上游业务库数据库及表结构）
数据工程师
报警并需要数据工程师审核/完善变更

血统分析

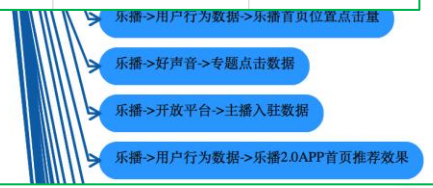


- **拓扑图分析**
  了解单点对于上下游甚至全局的影响
- **依赖分析**
  任务调度/运行异常、依赖异常、重写异常
- **产出分析**
  数据产出时间预估等
- **异常分析**
- **合理性分析**
- …

血统分析



NativeApp	Android核心数据	Android日使用用户数	core	10:00	2015-01-21 14:00:01	已产出	修改 | 手动更新
NativeApp	NativeApp月数据	nativeapp月使用用户	common	14:00		已下线	修改 | 手动更新
NativeApp	性能监控数据	性能监控数据	common	14:00	2015-01-21 14:00:01	2015-01-21 21:00	修改 | 手动更新



欢迎加入我们！

微信：kioskgao

邮箱：yushigao@meilishuo.com


ChinaUnix
ITPUB
THANKS