



# 第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

## 架构革新 高效可控



北京国际会议中心 | 2020/12/21-12/23

# 元数据治理在企业中的实践

主讲人：何军

数据平台开发部负责人





# 目录

01

数据治理简介和解读

02

元数据管理实践分享

03

元数据质量管理的思考



# 数据治理简介和解读

---



# 数据治理的内涵与外延

**狭义：**数据治理是指对数据质量的管理、专注在数据本身。

**广义：**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，围绕数据生命周期而开展的业务。技术和管理活动都属于数据治理范畴。有的专家干脆把广义的数据治理称为数据资产管理。

## 数据治理的发展历史



## 利用比喻介绍数据治理各管理域



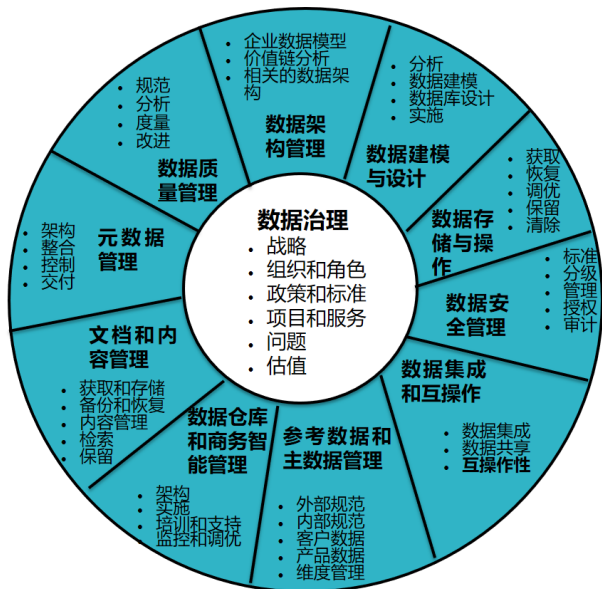
- ❖ **数据架构和模型管理**-----要把水果摆放的紧凑、稳定、便于取放
- ❖ **数据标准管理**-----水果规格一致
- ❖ **数据质量管理**-----保证没有烂水果
- ❖ **元数据管理**-----描述水果的品种、价钱和产地
- ❖ **主数据管理**-----往往是高价值、受欢迎而放在专区的水果
- ❖ **数据生命周期管理**-----水果保质期管理、下架管理
- ❖ **数据安全**管理-----防止被人偷吃、损坏
- ❖ **数据服务与应用**-----导购精准服务
- ❖ **数据资产管理**-----盘点有哪些水果，数量多少，价值多少



# 数据治理范围的权威参考

➤ DAMA是目前国际上权威的数据治理框架，2018年我国发布数据管理能力成熟度评估模型，提供数据管理能力的重要标准。

国际DAMA数据治理框架



国内数据管理能力成熟度评估模型



其他比较权威的参考：中国信通院2019年发布的《数据资产管理白皮书3.0》

# 企业中的数据治理工作范围

## 数据治理 流程

组织架构

职责分工

体系规划

管理制度

策略执行

考核评估

## 数据治理 专项工作

数据标准管理

元数据管理

数据架构管理

主数据管理

数据安全治理

数据质量管理

数据生命周期管理

## 数据治理 工具与技术

数据标准管理工具

元数据管理工具

数据架构工具

主数据管理工具

数据安全治理工具

数据质量管理工具

数据生命周期管理工具

流程管理工具



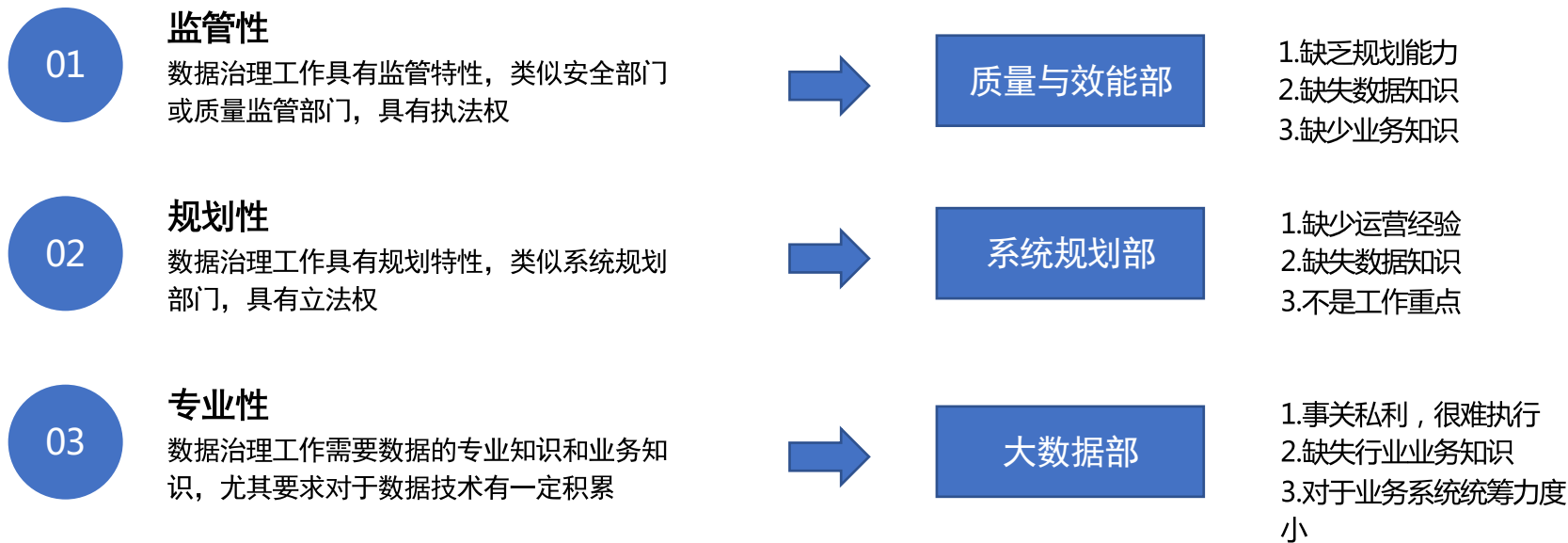
## 企业开展数据治理工作的切入点

很多企业在投资建设新的数据仓库、核心系统、大数据平台、数据中台时或者开展主数据管理时，会考虑开展数据治理工作，建立数据标准，进行数据质量检查，盘点各系统现状，保证数据迁移有序，保证新平台从建设之初数据就规范可控。



# 根据数据治理工作的特点，由哪个部门来做？

## 缺点



# 元数据管理实践分享

---



架构革新 11<sup>th</sup>  
高效可控



# 元数据的定义

□ **元数据** 是描述数据的数据，根据描述内容，可分为业务元数据、技术元数据、管理元数据。

## 业务属性

- 中文名称
- 英文名称
- 业务定义
- 业务规则
- .....

## 技术属性

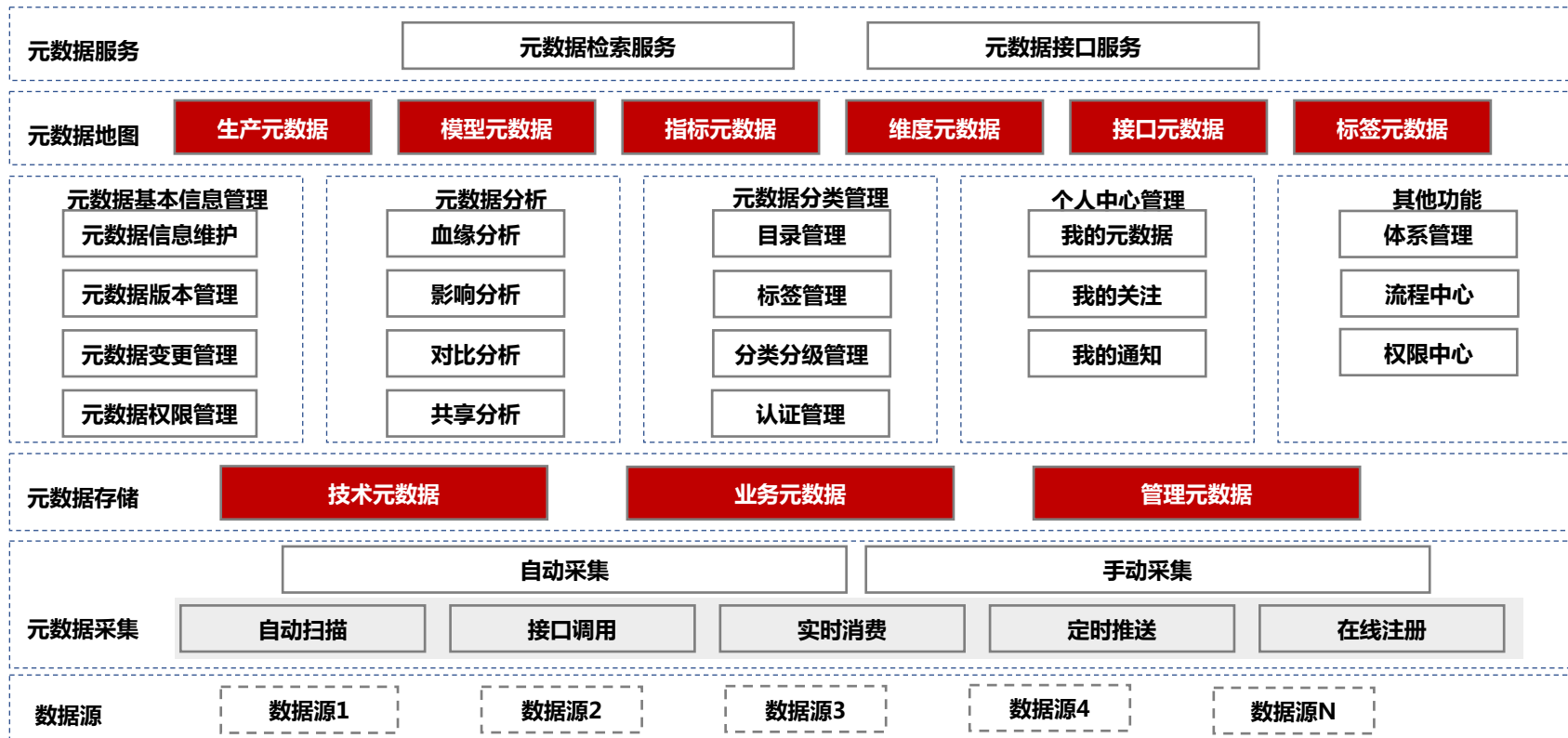
- 存储位置
- 存储周期
- 数据类型
- 数据长度
- .....

## 管理属性

- 数据开发负责人
- 数据业务负责人
- 数据使用部门
- 是否公司统一
- .....

元数据是帮助查找、存取、使用和管理信息资源的信息。元数据是业界公认的数据管理中的核心要素，做好元数据管理，更容易的对数据进行检索、定位、管理、评估。

## 元数据管理功能架构



通过筛选条件：业务体系、所在库、所属主题、所属业务，选择要关注的模型元数据

统一数据目录

数据模型目录

数据源目录

数据百科

个人中心

管理中心

模型

请输入模型英文/中文

跳至think

收起筛选

检索条件 >

业务体系: 京东物流 >

所在库: **cdm** >

业务体系	京东零售	<b>京东物流</b>																	
所在库	app	adm	<b>cdm</b>	gdm	fdm	dim													
物流主题	全部	客户	组织	流量_旧	商品	订单	销售	仓储	配送	客服	供应链	财务	社区_旧	时效	调度	门户			更多 >
售后	大件	合约	DIM公共维度																
所属业务	整体	仓储	快递																链路

清空选项

统一模型

为您找到相关结果 52 个

cdm\_cs\_afs\_ord\_survey\_det

已统一

创建时间: 2019-07-11 17:26:16

十张报表-cdm层评价明细

业务说明:

表描述: pop和自营的订单评价明细数据

表粒度: 该模型以订单号为粒度

指标: 明细数据表, 无指标

维度: 订单类型、评价客户端类型、用户下单级别、库房信息、站点信息、pop商家名称等等

时效: T+1

物流主题: 客服

cdm\_dis\_waybill\_process\_basic\_det

已统一

创建时间: 2019-07-11 17:26:16

配送中小件全程跟踪基础模型

业务说明:

1、表描述: 配送中小件全流程节点, 包含自营、POP、SOP、纯外单, 正向和逆向, 包含运单属性、预分拣、分拣、终端、3PL、自提柜等环节。京东物流承担多种业务类型还会承担非京东平台销售的运单的配送服务, 当然这些业务也可以入京东仓。SOP、LBP业务基本现在已经没有了。

此外, 受业务模式和国家相关法律法规限制, 京东的全球购业务、医药业务等, 也会相对京东商城业务有所区别, 走京东配送时, 也走外单业务模式配送。

在京东与一号店的融合过程中, 也是以运单形式交给京东物流承运, 体现出来的是纯外单业务, 一号店在青龙系统体现就是N个外单业务青龙B商家。

需要特别说明的是, 我们理解上认为的一个品牌商, 根据业务模式不同, 例如同时在淘宝天猫或者自己电商平台订单都归京东承运, 又加上商家的发货地不同, 或者品牌注册了多个法人机构, 一个品牌商可能衍生出来多个青龙B商家。例如“安路”这个商家, 有多达23个B商家ID。

2、表粒度: 该模型以运单号为粒度, 运单号唯一。

3、指标: 运单数、妥投运单数、出库运单数、指定妥投站点的一段时间运单数

4、指标维度: 销售模式、部门、区域、省份、城市、店铺、商家等。

5、时效: T+1。



## 模型元数据详情-模型详情

统一数据目录

数据源目录

数据源目录

数据源目录

个人中心

管理中心

数据模型目录 - 元数据详情

cdm\_dis\_waybill\_process\_basic\_det

配送中小件全流程跟踪基础模型

未审核申请

92

0

添加关注

模型详情

字段详情

集市详情

使用场景

表到任务

表到表

基础信息

数据信息

创建时间: 2019-07-11 17:26:16

更新时间: 2020-07-23 11:23:53

业务体系: 京东物流

所属集群: 10k

所属集市: mart\_coo

所在库: cdm

数据库类型: hive

所属主题: 配送

所属业务:

是否唯一: 是

使用部门:

所属发展中心: 技术与数据中心

所属区域分公司: 华南区域分公司

所属产品线: 快递产品线

技术支持部门: 西北区域分公司

体验保障中心:

数据信息

是否去重: 是

加工方式: 增量

开发负责人:

业务负责人:

业务说明:

1. 表描述: 配送中小件全流程节点, 包含自营、POP、SOP、纯外单、正向和逆向, 包含运单属性、预分拣、分拣、终端、3PL、自提相等环节, 京东物流承担多种业务类型的包裹运输服务, 还会承担京东平台销售的订单的配送服务, 当然这些业务也可以接入京东仓。SOP、LBP业务基本现在已经没有了。  
此外, 受业务模式和法规法规限制, 京东的生鲜购业务、医药业务等, 也会对京东商城业务有所区别, 走京东配送时, 也走外单业务模式配送。  
在京东与一号店的融合过程中, 也是以运单形式交给京东物流承运, 体现出来的是纯外单业务, 一号店在青龙系统体现就是N个外单业务商家。  
需要特别说明的是, 我们理解上认为的一个品牌商, 根据业务模式不同, 例如同时在淘宝天猫或者自己电商平台订单都交给京东承运, 又加上商家的发货地不同, 或者品牌注册了多个法人机构, 一个品牌商可能衍生出多个物流商家, 例如“安踏”这个商家, 有多达23个商家ID。  
2. 表粒度, 该模型以运单号为粒度, 运单号唯一。  
3. 指标: 运单数、签收运单数、出库运单数、指定签收站点的一段时间运单数  
4. 指标维度: 销售模式、部门、区域、省份、城市、店铺、商家等。  
5. 时效: T+1。

分区说明:

双层分区, dt, dp  
dp: 已签收运单由签收并自完成三天的运单进行归档  
dt: 每天增量加上近3天签收及逆向签收的运单和未完成签收的运单  
如果不确定运单是否归档了, 或者运单是未归档到那个分区了, 最准确的方法是卡dt=你想知道的下单时间的最小值。  
例: 想获取2016-08-01至2016-10-31号的下单情况。卡dt>= '2016-08-01' and create\_ord\_dt<= '2016-08-01' and create\_ord\_dt<= '2016-10-31', 如果担心运单不全, 可以把dt往前推

备注:

纯阿拉伯数字, 京东自营或POP类型运单, 例如: 50806946560  
T+阿拉伯数字, 京东自营或POP类型订单, 因用户取消或拒收后产生的“配送退货”运单, 例如: 750806946560  
“VA”+“V”开头的, 常见的运单号, 包含POP业务的POP运单, 以及其他类型运单。POP商家的京东订单号会单独生成这种类型运单号, 例如: VA34219392610  
“PVA”+“V”开头的, 一般的运单号, 因用户取消或拒收后产生的“配送退货”运单, 例如: PVE34070739286  
W或WA开头的, 售后取件单号, 例如: WA850238430742982656 (自营的是W或WA, 外单的是VY)

评论

请输入

## 模型元数据详情-字段详情

统一数据目录

数据模型目录

数据源目录

数据百科

个人中心

管理中心

数据模型目录 > 元数据详情

cdm\_dis\_waybill\_process\_basic\_det 配送中小件全程跟踪基础模型 [表权限申请 >](#)

编辑模型

数据预览

变更业务负责人

模型详情

字段详情

集市详情

使用场景

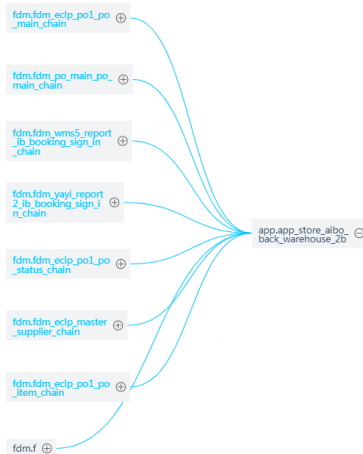
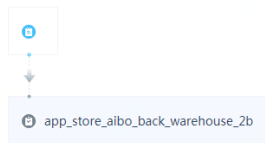
表to任务

表to表

字段名称	字段注释	数据类型	备注	操作
waybill_code	运单号	string		编辑
waybill_state	运单状态	string		编辑
subd_num	分公司编号	string		编辑
delv_center_num	配送中心编号	string		编辑
store_id	仓库ID	string		编辑
sale_ord_id	客户订单号	string		编辑
ship_bill_type_cd	运单种类代码	string		编辑
ship_bill_cate_cd	运单类别代码	string		编辑
products_flag	运单种类代码	string		编辑

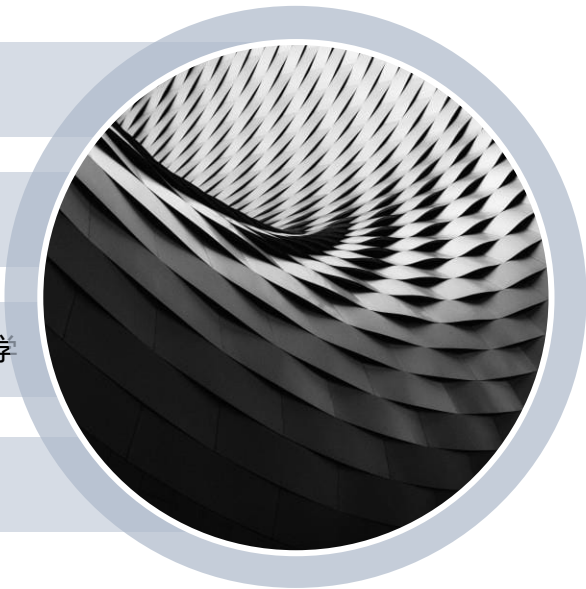


## 模型元数据详情-表To任务、表To表



# 元数据在企业中的价值

- 1 构建企业的数据知识库，可以跨产品线了解数据资源
- 2 是数据民主化的重要载体，非研发同学也可以了解企业的数据
- 3 血缘关系方便上下游数据问题定位，数据变更能精确通知下游同学
- 4 助力企业盘点数据资产，是数据资产化的基础

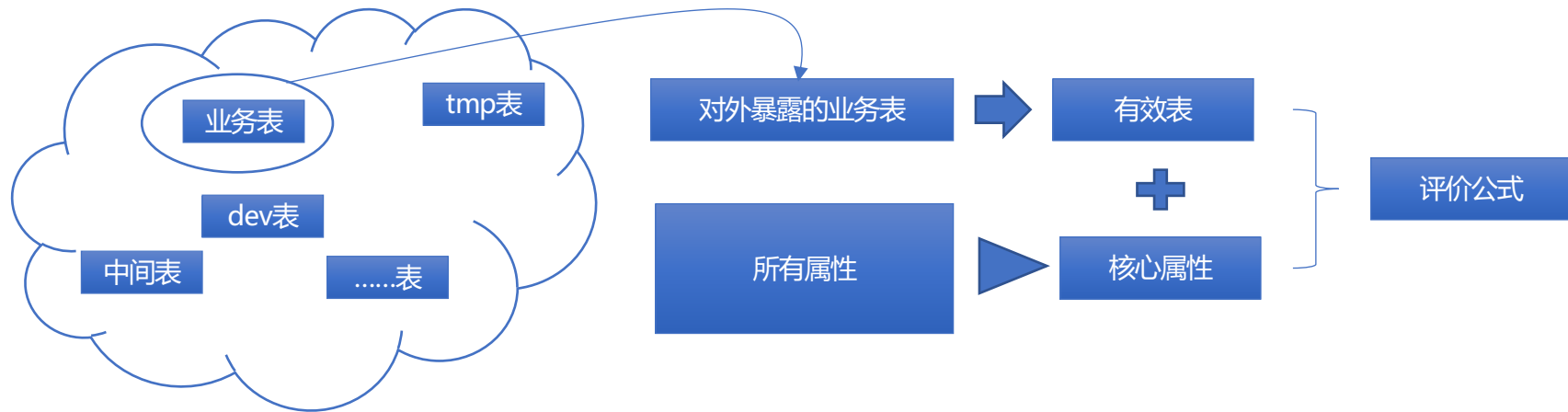


# 元数据质量管理的思考

---



## 衡量元数据质量的评价标准



评价公式

AVG ( 有效表的已经覆盖元数据核心属性总量/有效表元数据核心属性总量 )





# 针对核心属性，构建自动化检测工具

## 检查规则

缺少业务负责人

缺少所属主题

缺少所属业务

缺少是否统一标签

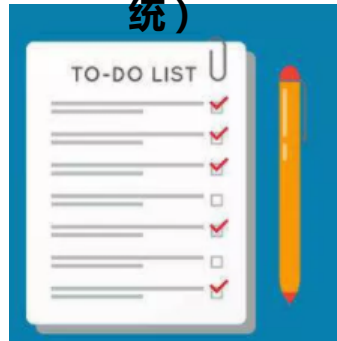
缺少字段注释

业务说明未按格式填写

.....



整改  
(会涉及到IT系统和DT系统)



## 保证元数据质量，引入长效机制

两个原则：



元数据采集尽量自动化



元数据维护尽量源头化

生产系统表设计阶段，明确表的中文名，表的业务说明，表字段的中文注释，表的开发负责人，表的业务负责人。



生产系统表建表阶段，要求DBA通过工具检查建表脚本，来审核上述各项是否具备。



数仓模型开发阶段：ODS表，继承了生产系统表相关的元数据属性，ODS以上各层表的开发，数据开发人员在大数据平台上，完成模型开发，建的表要包含表的中文名，字段注释等信息。之后，系统会自动把元数据信息，同步到元数据系统，开发人员在元数据系统中，维护业务说明等相关元数据信息。



数仓模型测试阶段：测试人员检查元数据系统中，相关元数据信息的完整性，如果发现有元数据信息缺失的情况，提交问题单，驱动开发人员补齐元数据。



# THANKS

