

湖仓数据模型设计与治理

北京数语科技有限公司

目录

01 公司介绍

02 湖仓模型设计与数据建模产品

03 Doris数据底座模型设计案例

01

公司介绍

公司介绍

数据治理驱动数据赋能之路

敏捷，生态，智能



datablau

数据资产管理平台



- 公司总部位于北京，全国设有4家分支机构，覆盖北、上、广、深，具有多地支持服务能力
- 公司人员200+，本科以上学历80%、研究生以上学历30%、研发和交付人员占总人数80%
- 人员多为名校毕业：北大、北邮、中科院、哈工大、西安交大、海外留学背景
- 名企工作经验：IBM、HP、Oracle、CA、VMware、AIA
- 技术实力雄厚，交付质量高，支持华为、交行等多个大型客户，客户评价高，续签率高，60%客户为复购客户
- 已成为阿里云、华为云、腾讯、百度、建信金科等多家顶级供应商的战略合作伙伴，提供云平台的整体解决方案

优秀企业的共同选择

1

华为本部

9

中国Top10
汽车行业有9家

4

中国Top5
家电企业有4家

3

中国6大
国有银行有3家

17

中国Top20
证券有17家

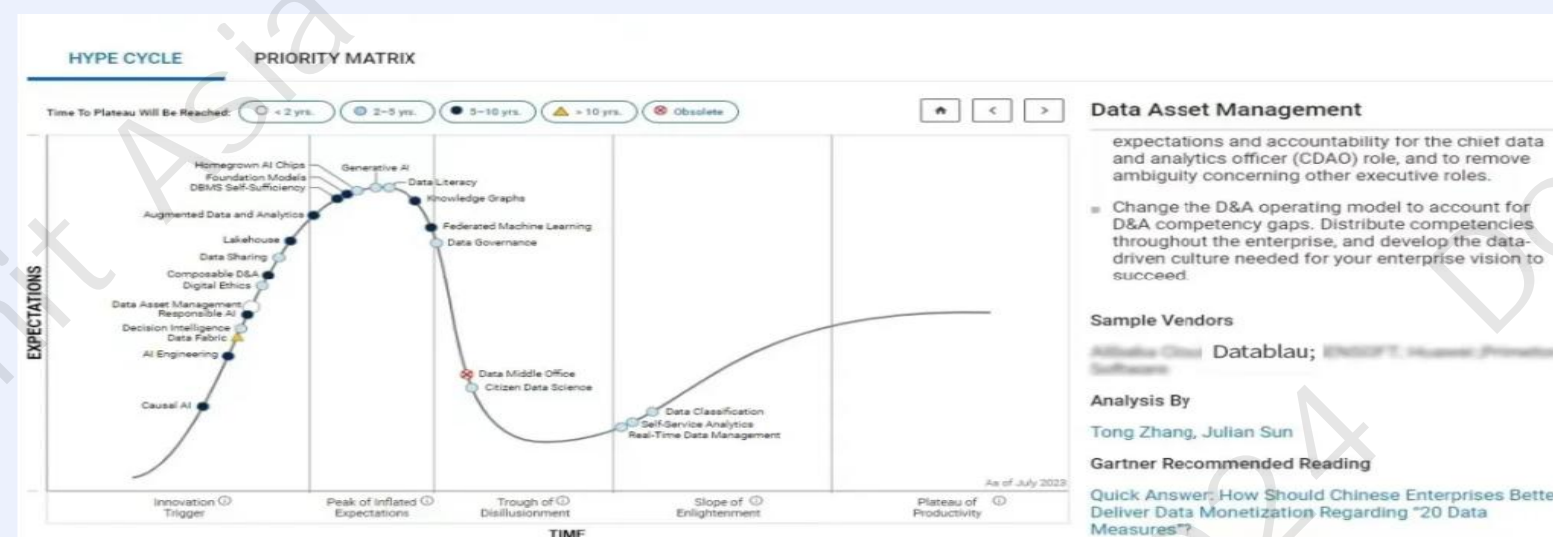
制造业

- 京东方、天马微电子、记忆存储、长安、吉利、长城、上汽通用、广汽丰田、广汽本田、金风、隆基、天合光能、宁德时代

技术领先，紧跟人工智能的时代浪潮

Gartner评选为数据资产管理代表厂商

◆ 被国际权威机构Gartner评为数据资产管理的代表厂商，充分体现了中国市场和国际权威机构对Datablau的认可，同时彰显了Datablau在数据资产管理领域领先技术实力和出色的产品能力，在数据模型、数据目录、数据质量、数据标准和元数据等各个领域具有丰富的实践经验，帮助企业解决数据治理工作中的耗时费力、效率低等问题。Datablau一直致力于通过数据模型管控、数据资产管理帮助企业实现数字化转型，在过7年多时间已经覆盖到众多行业。面对企业数据应用高速增长需求，Datablau将继续推进在更多领域的拓展落地，助力数智新时代。



01 权威发布

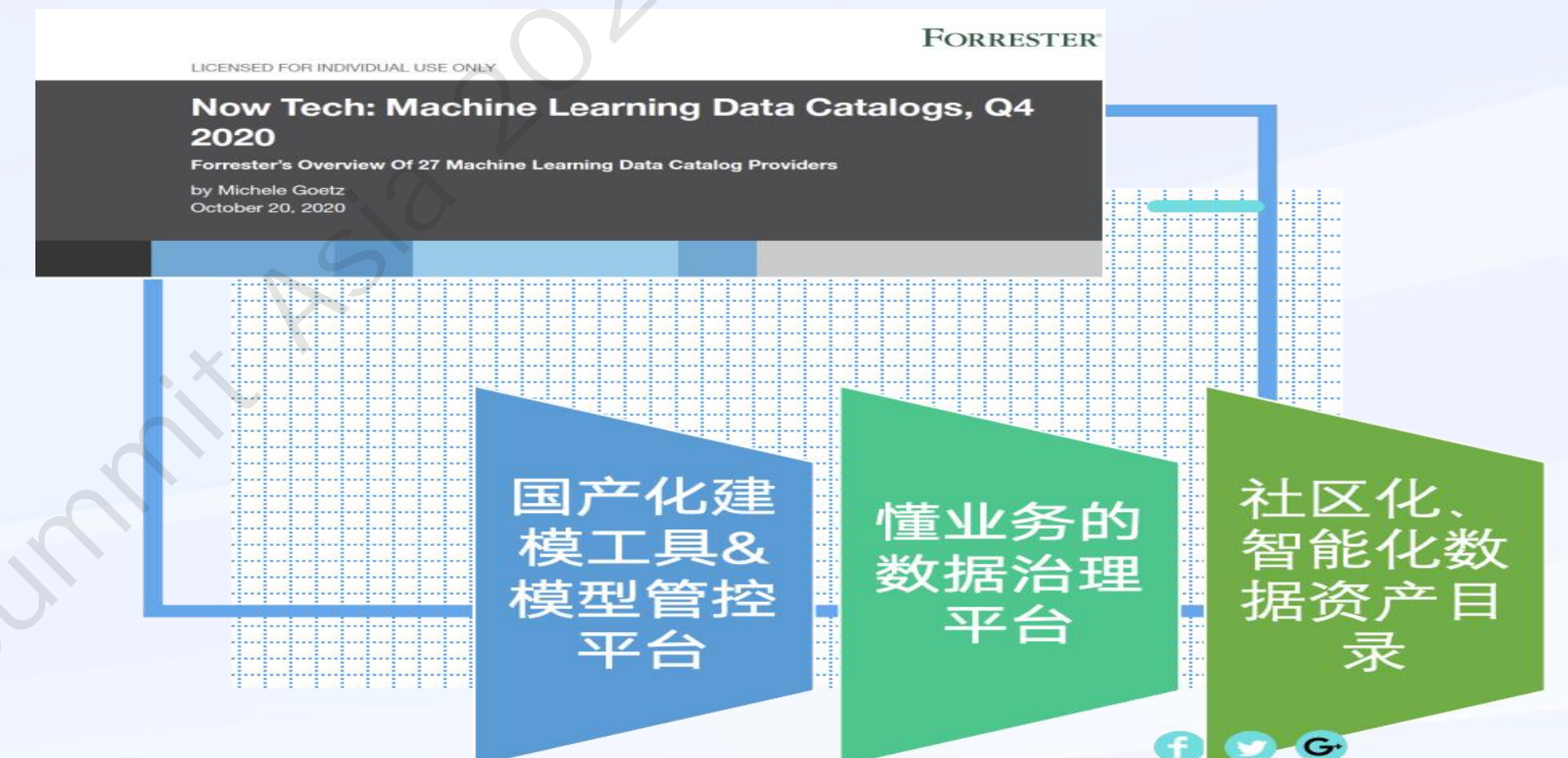
Gartner 发布

02 报告

《Hype Cycle for Data, Analytics and AI in China, 2023》

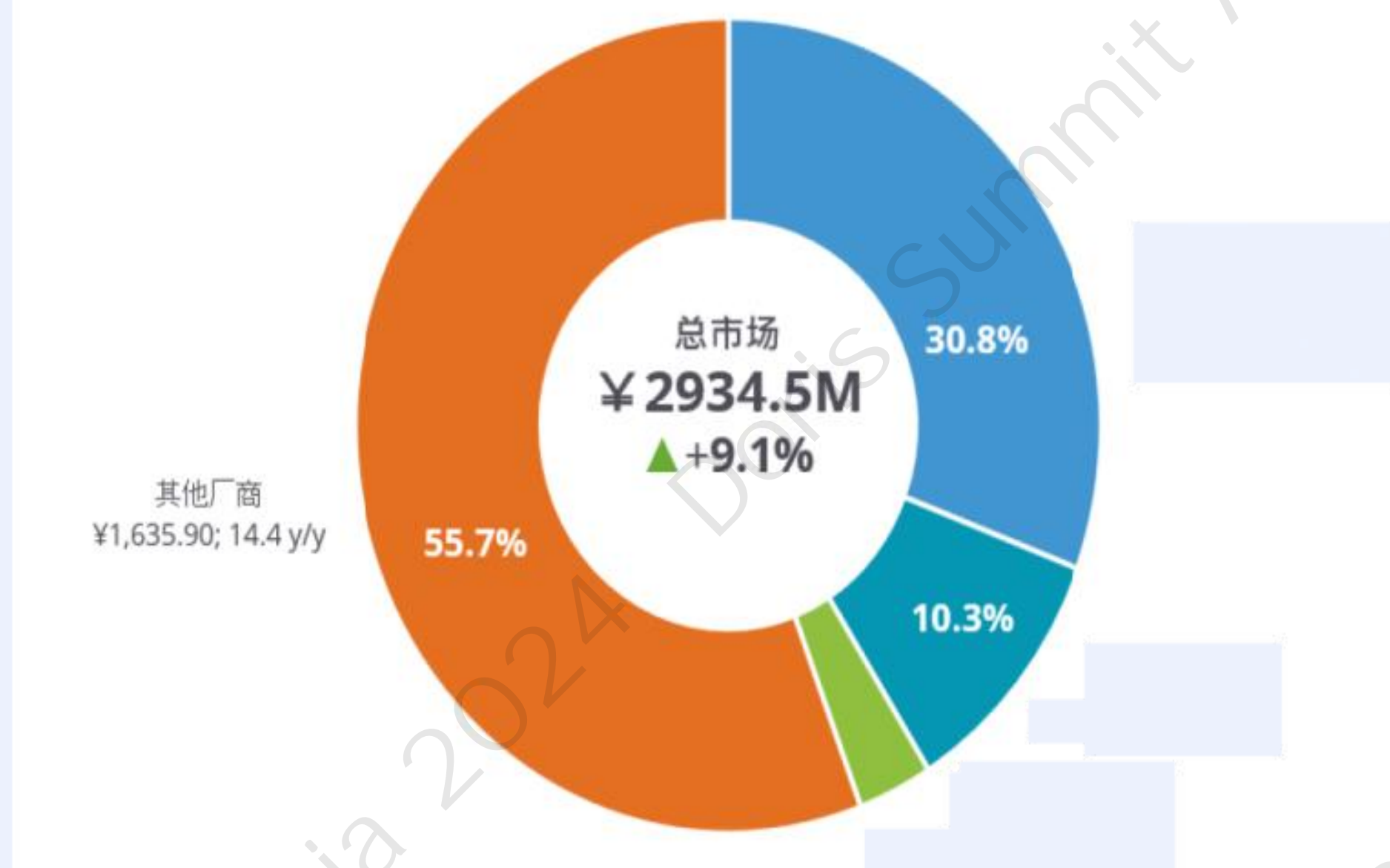
亚太地区唯一入选Forrester MLDC的厂商

◆ Forrester认可Datablau在数据管理领域的探索和实
践，在数据模型、数据目录、数据质量、数据标准和
元数据等环节自动化应用实现，帮助企业解决 AI 应用
门槛高、开发效率低等问题。Datablau一直致力于通
过数据模型管控、数据资产管理帮助企业实现数字化
转型，在过去4年多时间已经覆盖到众多行业。面对企
业数据应用高速增长期的到来，Datablau将继续推进
MLDC在更多领域的拓展和落地



市场排名第一的数据治理平台专业服务厂商

中国数据治理平台市场份额概况，2023



IDC：中国数据治理平台专业厂商市场份额第一

◆ 国际权威机构IDC对数语科技的评价：DDM融入了数据治理理念的数据模型设计与管理工具，把数据治理流程推进到数据开发流程中，进行源头治理，解决了标准落地的难题，从根本上管控企业数据质量问题。DAM企业级数据资产管理平台开创性的将数据标准、数据质量和元数据融合在一起实现闭环管理链路。DDC数据资产目录服务平台，从数据资产业务化视角出发，基于内置体系和数据自学习技术，形成企业统一的数据资产目录，依托自动数据分类分目、数据资产检索、数据资产地图等核心功能，极大提高了数据利用效率和提升业务数据应用水平。DDS数据安全平台定义数据资产安全级别，建立数据访问控制体系和动态脱敏引擎，确保数据质量、数据服务、数据查询等场景的安全性。数据链路监测平台：基于元数据采集和血缘解析，提升数据治理的透明度和效率。数据血缘解析成功率大于95%，响应速度达到毫秒级。数据资产开发平台：提供端到端的数据资产开发能力，规范数据开发流程，提升数据质量。核心功能包括数据仓库建模、项目管理、智能程序开发等。DDM Archy基于Datablau DDM推出的架构建模套件，统一贯穿业务到数据、高端架构到初级项目实施，提升数据治理成熟度和数据价值释放效率。Datablau AIC集成海量行业知识库，赋能元数据补全、数据质量规则构建等数据治理工作

01 权威发布
IDC 发布

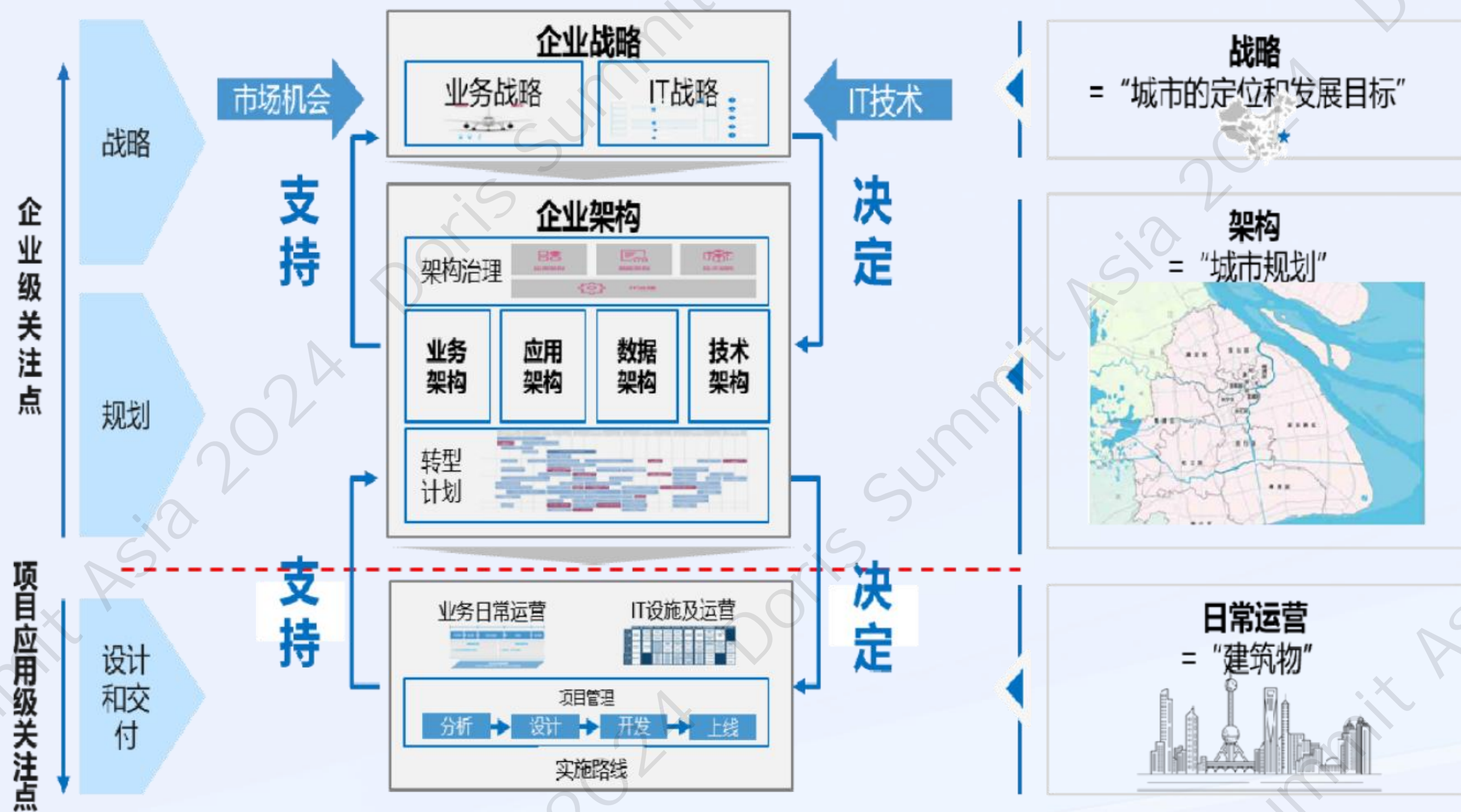
02 报告
《中国数据治理市场份额，
2023——为GenAI做好准备》

02

湖仓模型设计与数据建模产品

为什么需要数据模型？

企业架构就像企业的“城市总体规划蓝图”，在它的指导下，各个IT系统的建设得以有序开展



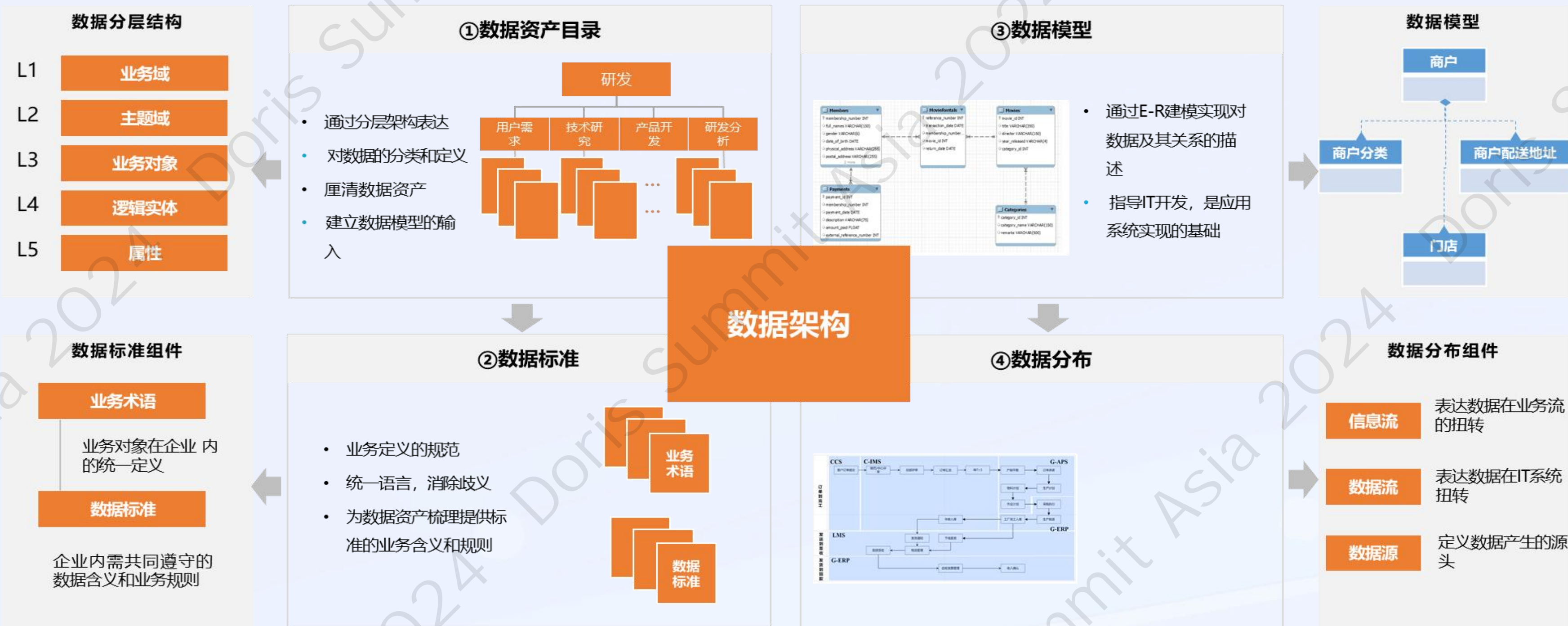
- 1 明确IT系统建设方向，确保IT功能对齐业务蓝图需求
- 2 打造可复用的业务或IT/数据能力，避免重复建设
- 3 明确系统职能边界，构建全局统一的IT能力地图，避免各自为战导致信息孤岛现象
- 4 监管IT实施交付过程，持续提升交付质量

代码千行不如好的架构图一张

为什么需要数据模型？

数据治理工作内容【数据架构建设】

- ◆ 数据架构是指以结构化的方式，描述在业务运作和管理决策过程中，所需的各类数据及其关系的一套整体组件规范
- ◆ 数据架构包含数据资产目录、数据标准、企业级数据模型和数据分布四个组件，如下图所示



数据架构灵魂五问

数据格式和定义是什么？数据从哪来到哪去？数据怎么被加工和消费的？数据被谁消费？数据何时获取和变更？

为什么需要数据模型？

VS

数据架构

- ❑ 组成元素：参与业务活动的数据实体、数据实体间关系、数据层级结构及提供报表加工或者外部数据共享的数据结构（视图）
- ❑ 价值作用：通过标准化数据结构设计保障系统的数据满足业务活动交互及数据分析的需要；通过提炼公共的数据结构设计解决数据分散、数据孤岛、数据口径不统一的问题；指导应用开发，确保系统实际运行产生的数据满足业务需要
- ❑ 产出物：数据模型、数据标准、数据分布关系/数据血缘图，数据目录
- ❑ 应用场景：系统需求分析、数据需求分析、系统详细设计

应用程序架构

- ❑ 组成元素：应用程序模块、程序交互关系，应用层级结构及对外服务接口
- ❑ 价值作用：解耦业务复杂性，通过提炼标准化程序模块功能实现系统低耦合；指导应用程序开发，确保程序功能满足业务场景的需求；通过服务接口方式实现系统对外的业务能力输出的标准化和自动化
- ❑ 产出物：应用架构设计图、设计文档
- ❑ 应用场景：系统需求分析，系统详细设计

数据实体的标准化、共性化以解耦业务复杂性 应用程序模块的标准化、共性化以解耦业务复杂性

数据模型：数据架构设计的产出物，用以结构化描述系统之间数据结构和数据流转关系，既可用作业务需求分析的载体，确保产品设计人员与业务分析人员之间就信息的格式标准、使用场景达成共识，也可作为架构设计规范指导应用程序开发，确保数据的正确性以能够保障系统按照预想的方式执行

数据建模-从蓝图规划走向设计试试

生活中的模型:

房屋平面图、地图等，用不同的符号向相关用户清晰的展示



数据模型:

数据的关系错综复杂，成千上万个表通过各种关系或约束互联形成复杂的结构。没有数据模型，利益相关者很难看到现有数据库的结构、理解关键的概念

概念模型

主要用来描述世界的概念化结构，是一个高层次的数据模型，由核心的数据实体或其集合，以及实体间的关系组成

员工

逻辑模型

对概念数据模型进一步的分解和细化，描述实体、属性以及实体关系。

| 名称 | 类型 | 注释 |
|--------|-----|--------|
| 员工编号 | 字符串 | 唯一标识员工 |
| 员工姓名 | 字符串 | |
| 员工身份证 | 字符串 | 身份证号 |
| 员工性别 | 数字 | 描述员工性别 |
| 员工生日 | 时间 | 描述员工生日 |
| 员工学历 | 字符串 | 描述员工学历 |
| 员工用工形式 | 字符串 | 描述用工形式 |

物理模型

面向特定的数据库，结合数据库特征，便于计算机实现的模型。

| name | type |
|-------------|--------------|
| Emp_id | Char(8) |
| Emp_name | Varchar2(50) |
| emp_idcard | Char(17) |
| Emp_gender | integer |
| Emp_bth | timestam |
| Emp_bachler | Char(2) |
| Emp_type | Char(1) |

数据模型是数据底座建设成功与否的关键因素

数据模型是企业的核心数据资产，直接决定数据从设计开发到汇聚加工全生命期的效率和质量



Datablau数据建模产品核心能力

1

可视化建模
数据架构设计

2

智能数据标准
落标核标

3

模型设计
影响变更管理

4

数据规范检核
自动构建DDL脚本

5

多人协作建模
能力

6












模型与元数据
一致性比较

【Part 1】完备的Doris数据模型特性支持

1

完备的数据模型语义支持【数据库-Schema-表-字段-索引】

| 门店维表 |
|---|
|  门店编码 : VARCHAR(255) |
|  门店名称 : VARCHAR(255) |
|  门店状态 : CHAR(1) |
|  门店地址 : VARCHAR(255) |
|  开店时间 : DATEV2 |
|  闭店时间 : DATEV2 |
|  门店类型 : CHAR(4) |

| 订单交易表 |
|---|
|  OrderID (订单编码) : BIGINT |
|  SKUID (商品编码) : CHAR(32) |
|  ProdCnt (商品数量) : FLOAT |
|  ProdName (商品名称) : VARCHAR(255) |
|  ProdUnit (商品单位) : CHAR(1) |
|  ProdPrice (商品单价) : DECIMALV3(12,4) |
|  TotalPrice (订单总价) : DECIMALV3(12,4) |
|  Currency (结算币种) : CHAR(3) |
|  ShopRegion (门店区域) : VARCHAR(255) |
|  OrderTime (下单时间) : DATE |
|  门店编码 : VARCHAR(255) |

物化

Doris数仓模型样例

4

聚合模型支持如AGGREGATE|UNIQUE|DUPLICATE

2

完备的数据类型支持

3

Doris内置及自定义索引支持

DDL 脚本预览

```
1  /* ===== */
2  /* Table: 门店维表 */
3  /* Definition: */
4  /* ===== */
5  CREATE TABLE IF NOT EXISTS 门店维表 (
6      门店名称    VARCHAR(255),
7      门店状态    CHAR(1),
8      门店地址    VARCHAR(255),
9      开店时间    DATEV2,
10     闭店时间    DATEV2,
11     门店类型    CHAR(4),
12     门店编码    VARCHAR(255) NOT NULL
13 )
14 ENGINE=OLAP;
15 /* ===== */
16 /* Table: 订单交易表 */
17 /* Definition: */
18 /* ===== */
19 CREATE TABLE IF NOT EXISTS 订单交易表 (
20     OrderID    BIGINT NOT NULL COMMENT '订单编码',
21     SKUID      CHAR(32) COMMENT '商品编码',
22     ProdCnt    FLOAT COMMENT '商品数量',
23     ProdName   VARCHAR(255) COMMENT '商品名称',
24     ProdUnit   CHAR(1) COMMENT '商品单位',
25     ProdPrice  DECIMALV3(12,4) COMMENT '商品单价',
26     TotalPrice DECIMALV3(12,4) COMMENT '订单总价',
27     Currency   CHAR(3) COMMENT '结算币种',
28     ShopRegion VARCHAR(255) COMMENT '门店区域',
29     OrderTime  DATE COMMENT '下单时间',
30     ShopID     VARCHAR(255) COMMENT '门店编码'
31 )
32 ENGINE=OLAP
33 PARTITION BY RANGE (OrderTime)
34 (
35     PARTITION Part0 VALUES LESS THAN ("2025-01-01"),
36     PARTITION Part1 VALUES LESS THAN ("2024-01-01"),
37     PARTITION Part2 VALUES LESS THAN ("2023-01-01"),
38     PARTITION Part3 VALUES LESS THAN ("2022-01-01"),
39     PARTITION Part4 VALUES LESS THAN ("2021-01-01")
40 )
41 DISTRIBUTED BY HASH (OrderID) BUCKETS 10
42 AGGREGATE KEY (SKUID, ShopID, ShopRegion)
43 PROPERTIES ('replication_num' = '3');
```

5

性能优化特性支持如物化视图、分区分桶

【Part 2】Git 模式的多人协作建模提升开发效率

设计态（DataBlau）



运行态（元数据）

开发

SIT

UAT

版本

【Part 3】智能数据关系构建助力企业级数据架构建设

结合LLM语义分析能力探查元数据结构，综合业务架构智能还原高阶数据架构与数据间逻辑关系

SQL探索

- 获取DDH内的SQL运行日志，建立业务关系，并通过汇总总结，建立完整的ER关系。

语义引擎

- 建立同义词库和语料库，运用向量模型算法，对主键的外键关系进一步推定。

数据剖析

- 分析每个表中的数据，了解其特征。查找不同表中列之间的模式和相关性。例如，如果两个表都包含日期和产品ID列，您可以根据这些值的模式推断它们之间的关系

数据血统

- 查看描述数据仓库中数据血统的文档或元数据。这可以提供有关不同表之间关系的见解，以及数据如何在系统中流动的信息。

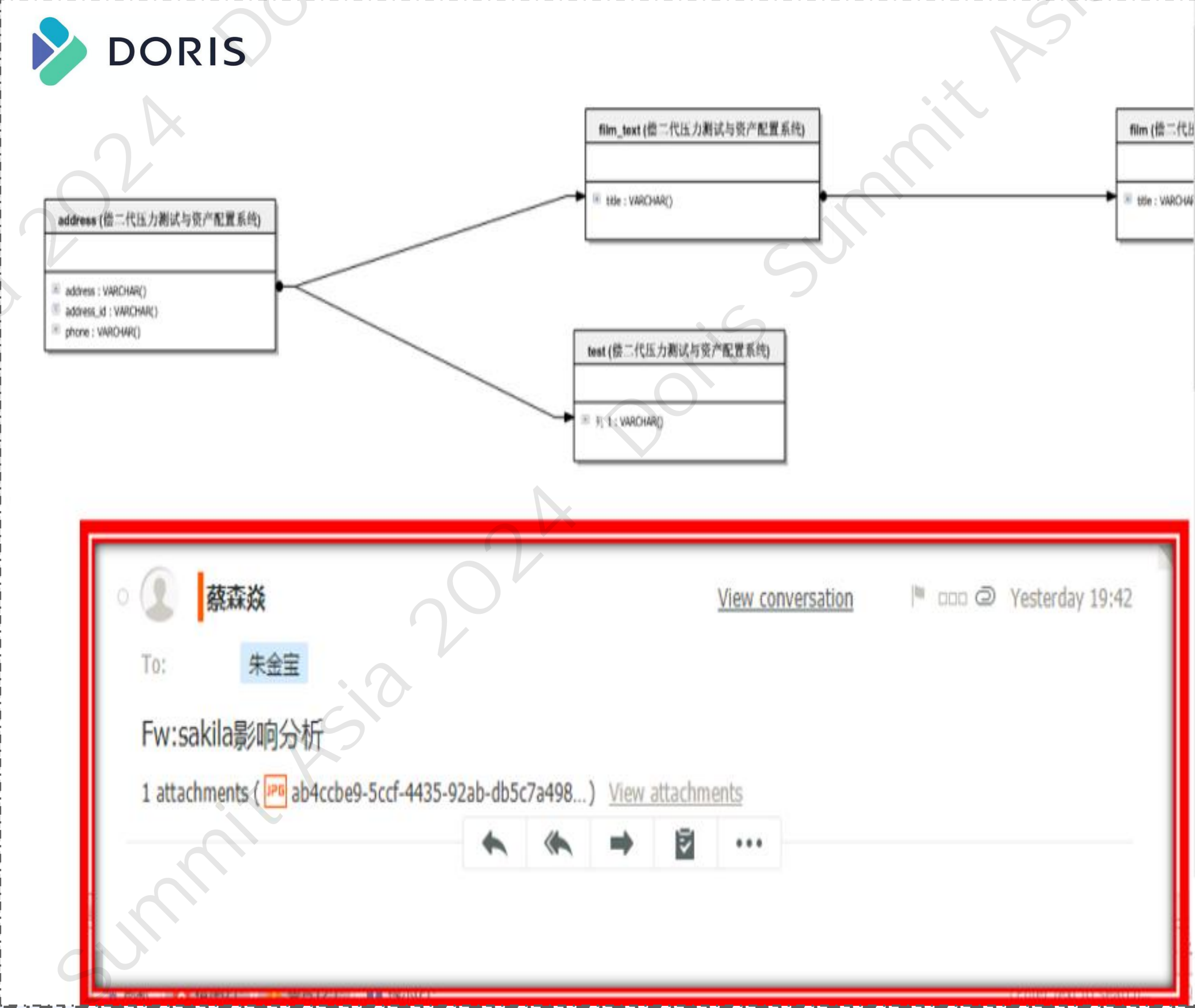
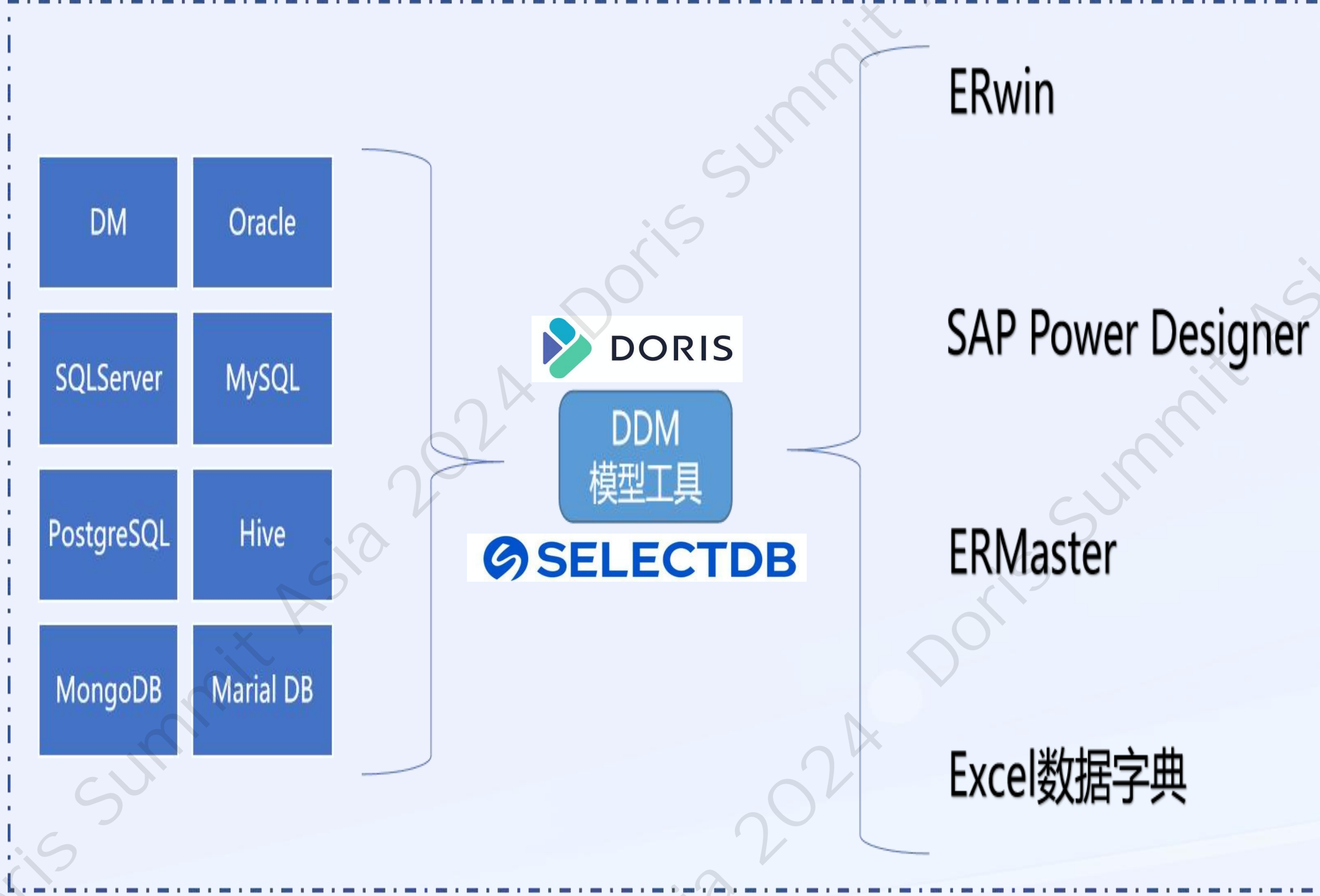
业务调研（人工修正）

- 利用您对业务领域的了解，推断表之间的关系。了解每个表中数据的含义以及它们与数据仓库中其他数据的关系。

【Part 4】异构数据源支持及涵盖数据全链路的智能设计变更管理

多数据源支持适应存量异构系统

基于Doris数仓血缘构建数据变更自动化管理机制



应用场景：模型驱动、设计治理一体化的Doris数据底座开发流程

设计阶段

数据开发项目

- 1.模型导入 - 通过导入工具，将PD、ERWin等工具的模型导入DDM中。
- 2.反向工程 - 通过直联数据库的方式，反向生成模型。
- 3.信息补全 - 补充模型中缺失的字段信息，例如字段中文名称

模型设计阶段

- 4.模型设计 - 使用客户端设计器进行模块设计与维护
- 5.影响分析 - 设计阶段能够显示模型的修改对下游系统的影响
- 6.字段引标 - 设计工具中能够引用数据标准

开发阶段

模型开发阶段

- 7.模型评审 - 模型的变更必须经过线上评审，信息补全率，落标率
- 8.数据开发 - 根据模型的映射和转换规范，书写代码进行开发和调试
- 9.任务编排 - 根据开发任务进行任务编排。

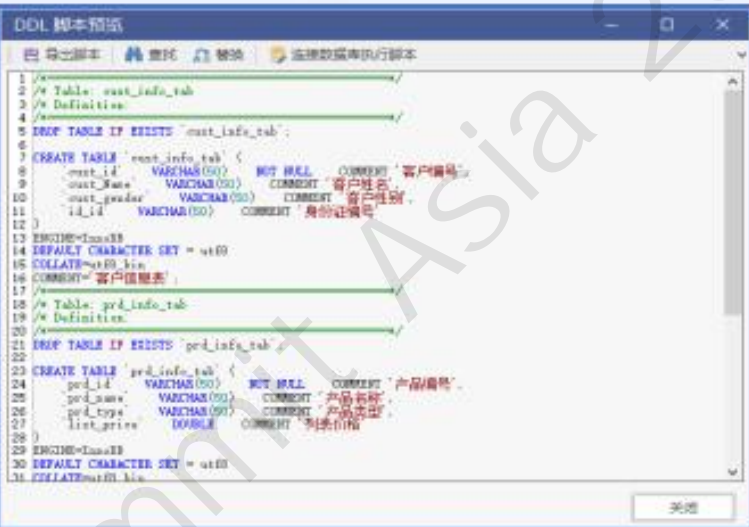
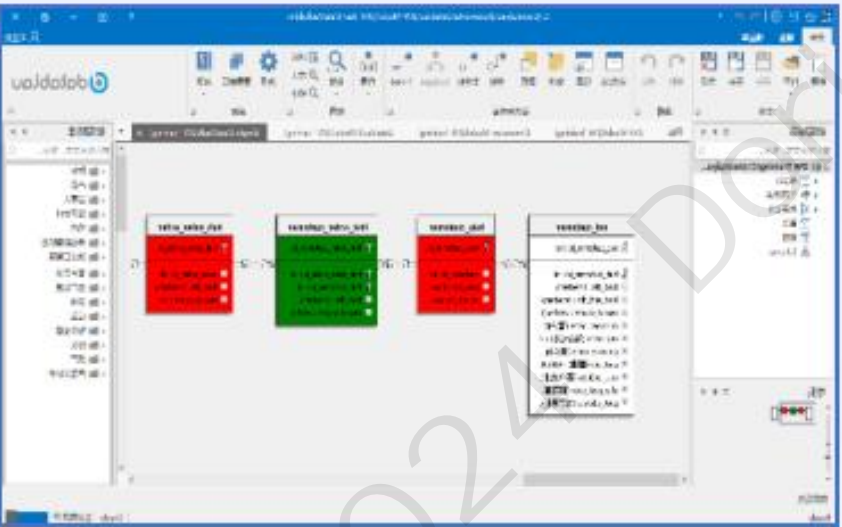
投产上线阶段

- 10.发布项目 - 检查和评审将投产DDL和DML代码，合格后，发布到调度工具
- 11.调度管理 - 对整体任务进行编排，切换生产库信息，进行上线投产。

投产后阶段

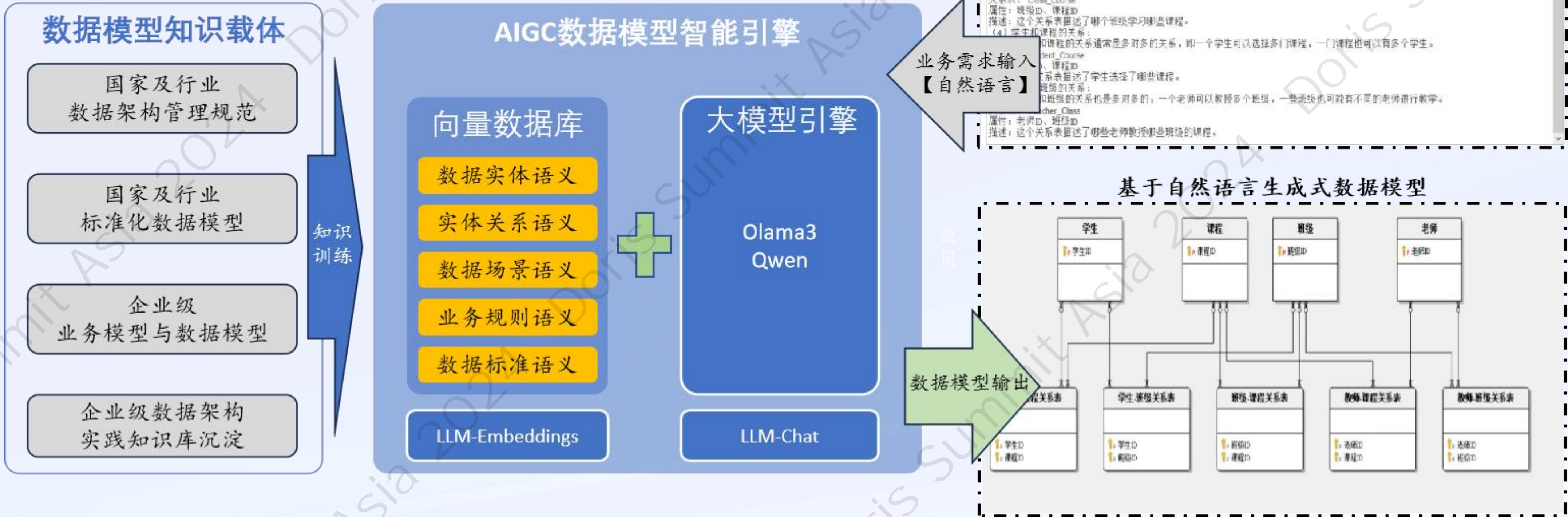
产后阶段

- 12.模型与元数据一致 - 将投产D模型与元数据比对，报告产生的不一致情况，进行后督改进。
- 13.发布数据资产 - 将模型信息更新到元数据中，进而发布到仓库数据资产中。



AI时代下的数据架构设计模式：业内首创的生成式数据模型

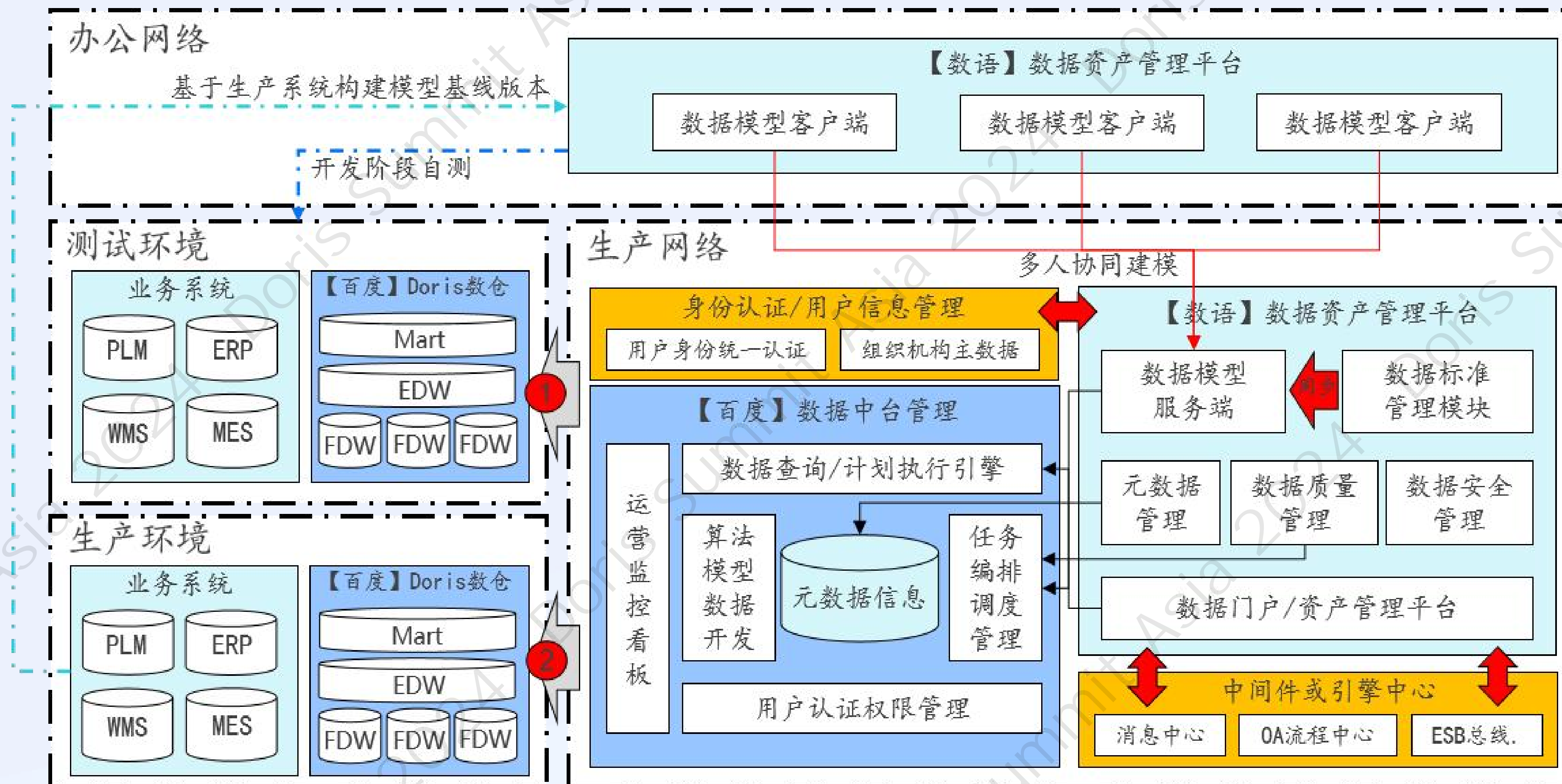
一分钟完成专业数据模型设计



03

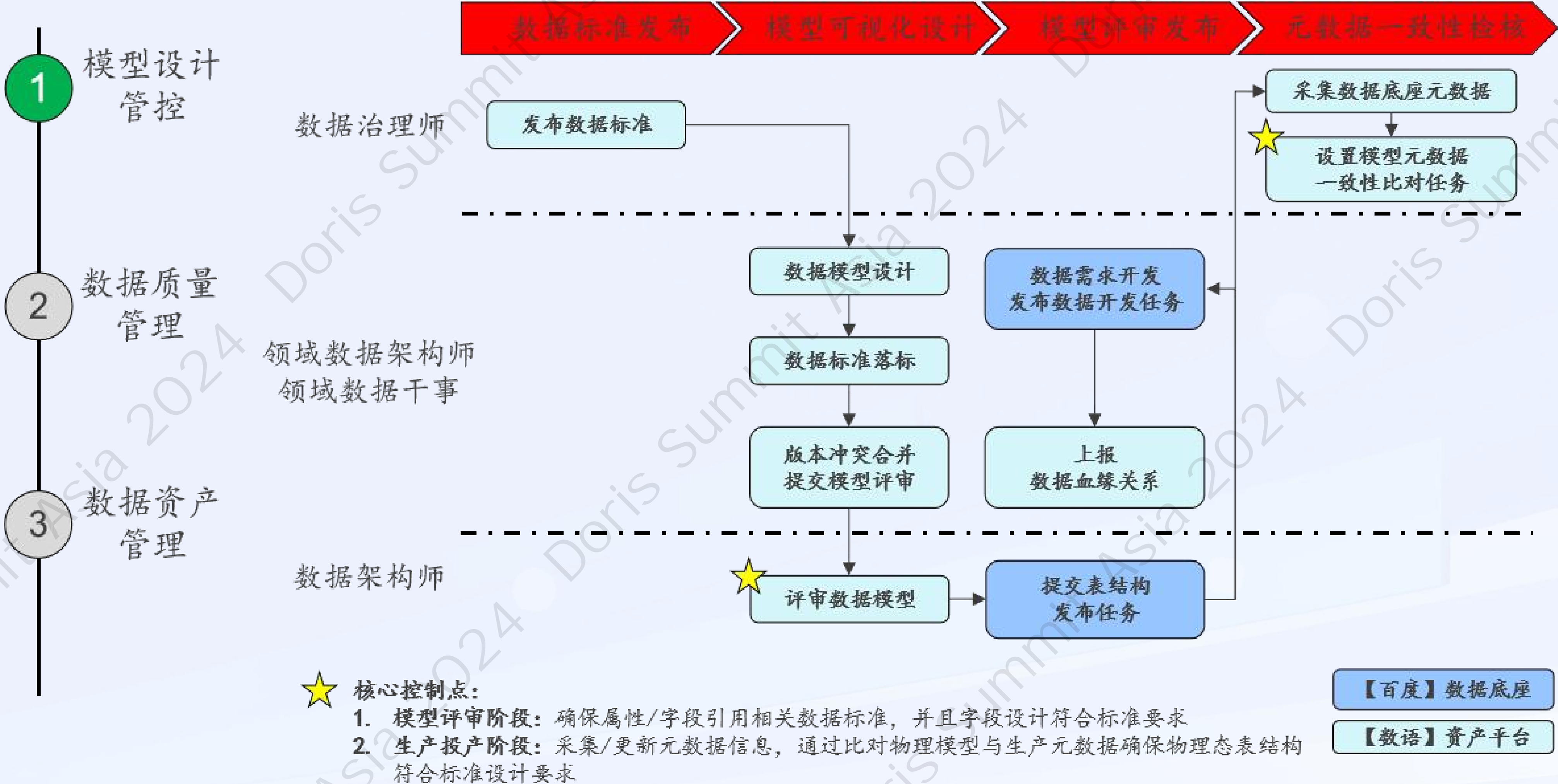
Doris数据底座模型设计案例

案例分享：XXX全球领先的大型PCB板制造集团案例

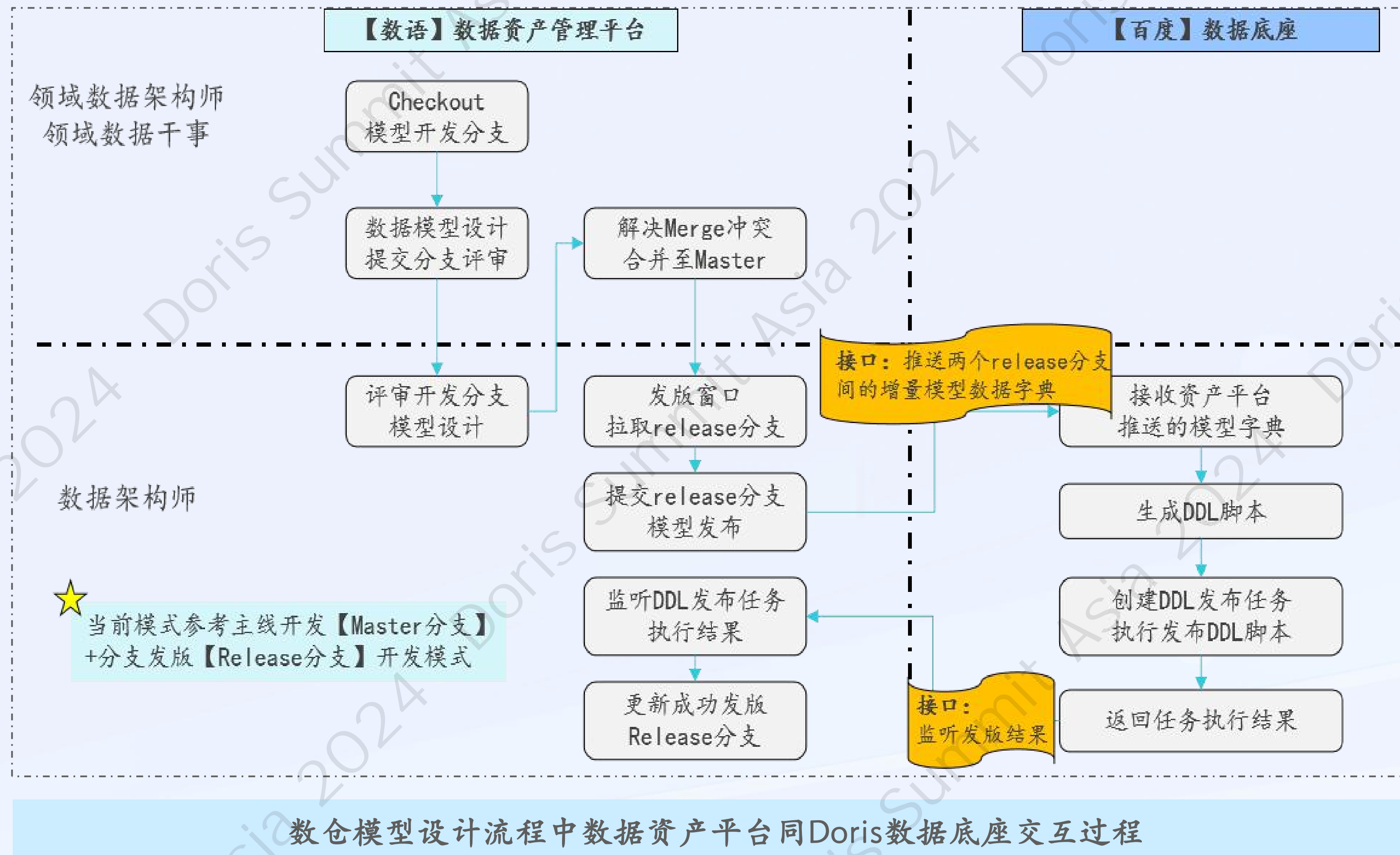


- 1、基于中台进行测试环境建表、数据开发及数据任务调度验证
- 2、基于中台进行生产环境建表、数据开发任务发布及调度配置

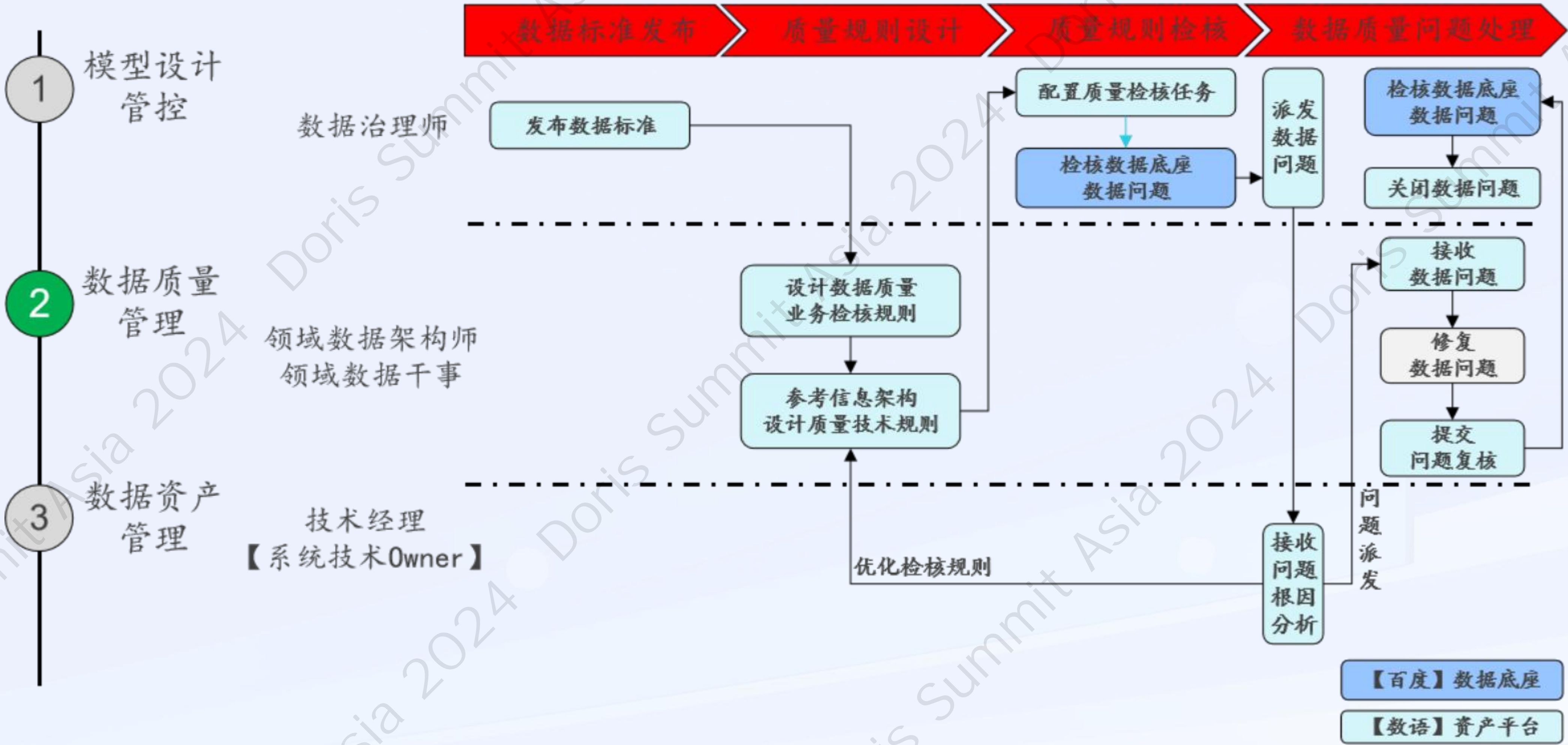
案例分享：XXX全球领先的大型PCB板制造集团案例



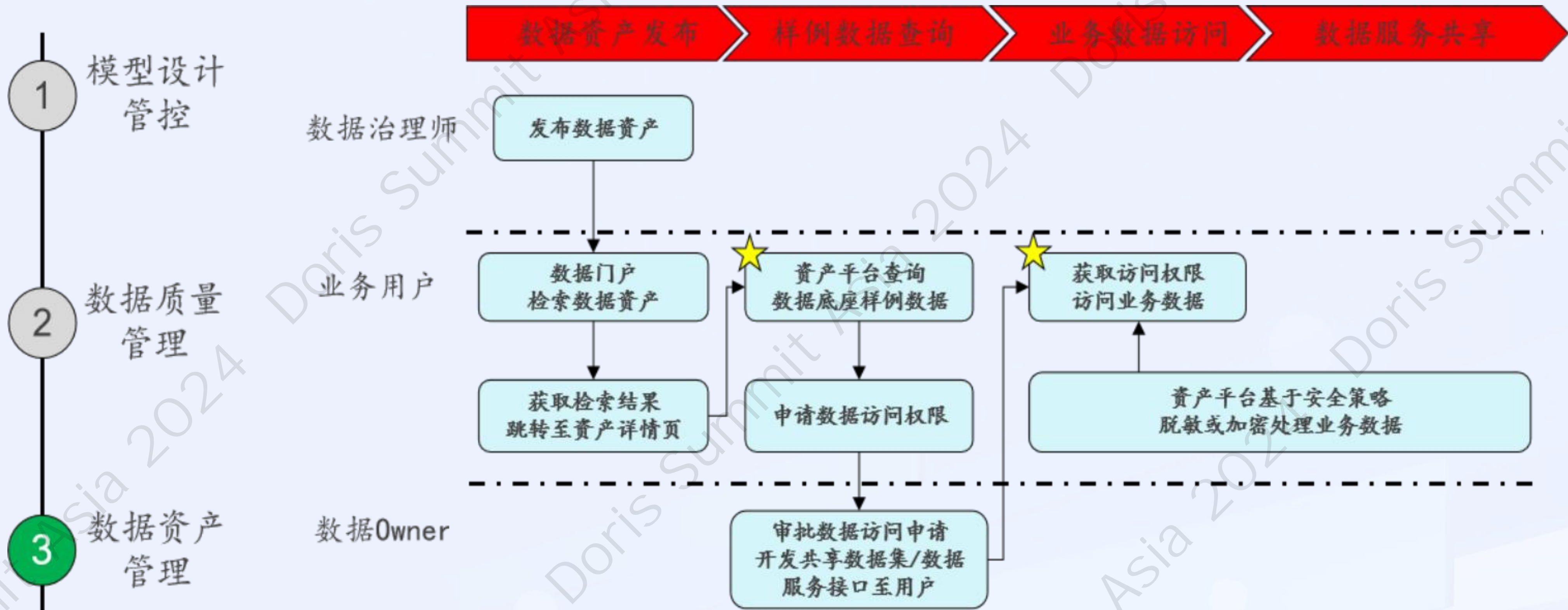
案例分享：XXX全球领先的大型PCB板制造集团案例



案例分享：XXX全球领先的大型PCB板制造集团案例



案例分享：XXX全球领先的大型PCB板制造集团案例



【百度】数据底座

【数语】资产平台

Thanks for Watching!