西安电子科技大学
XIDIAN UNIVERSITY

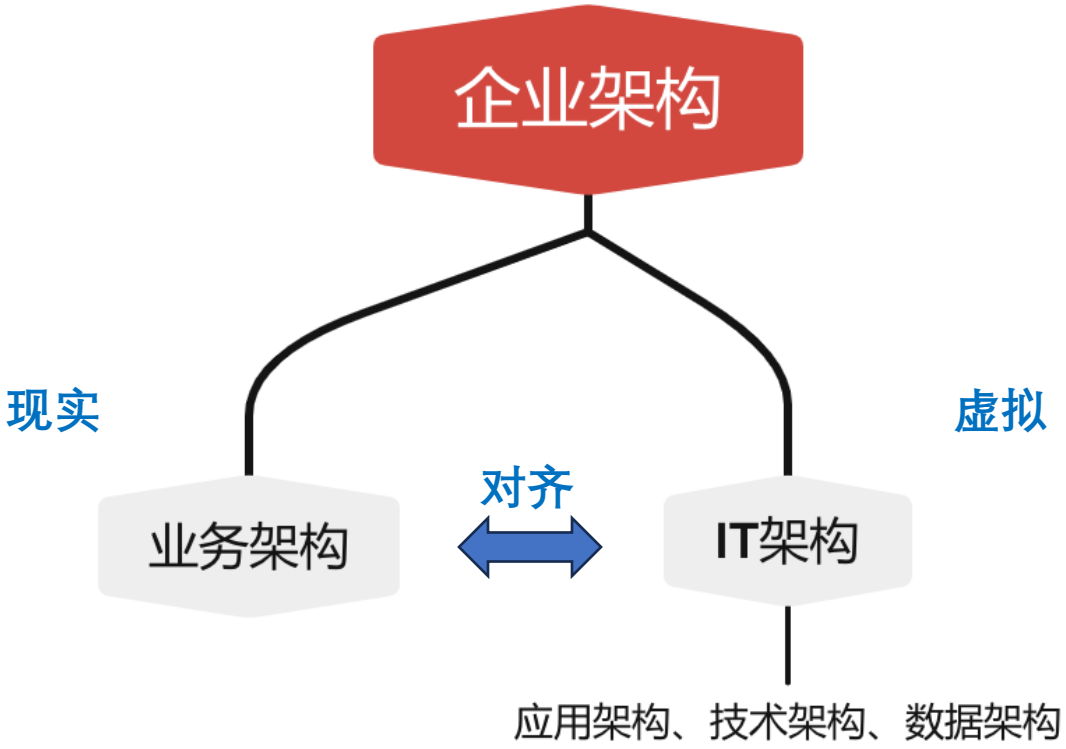
第二章 大数据架构管理

马 晶

经济与管理学院 信息管理系

Email: majing@xidian.edu.cn

数据架构概述



数据架构概述

定义：数据架构是一套指导组织内部数据需求、整合、控制和资产投资与业务战略相匹配的规范。

与企业架构的关系：数据架构是企业架构的重要组成部分，反映企业内部各种数据资产的组织、管理和利用。

核心作用：在支持整个企业的信息需求时，数据架构起到了至关重要的作用，助力企业更有效地管理、整合和利用其数据资产。

数据架构的价值主要体现在：

（1）从**企业运作**的角度，数据架构定义了企业运作过程中所涉及到的各类对象和其治理模式，成熟的数据架构，可以迅速地将企业的业务需求转换为数据和应用需求；

（2）从**数据资产**的角度，数据架构是管理数据资产的蓝图，帮助管理者梳理企业数据资产分布，将组织复杂的数据和信息传递至整个企业；

（3）从**数据管理**的角度，数据架构是企业各部门的共同语言，是数据管理的高层视角，是企业不同部门在数据层面保证业务和技术的一致性，有助于整个组织实现一致性数据的标准化和数据的整合，最终为企业改革、转型和提高适应性提供支撑。

数据架构的框架

架构框架提供了一种思考和理解架构的方法，以及需要架构的结构和系统。

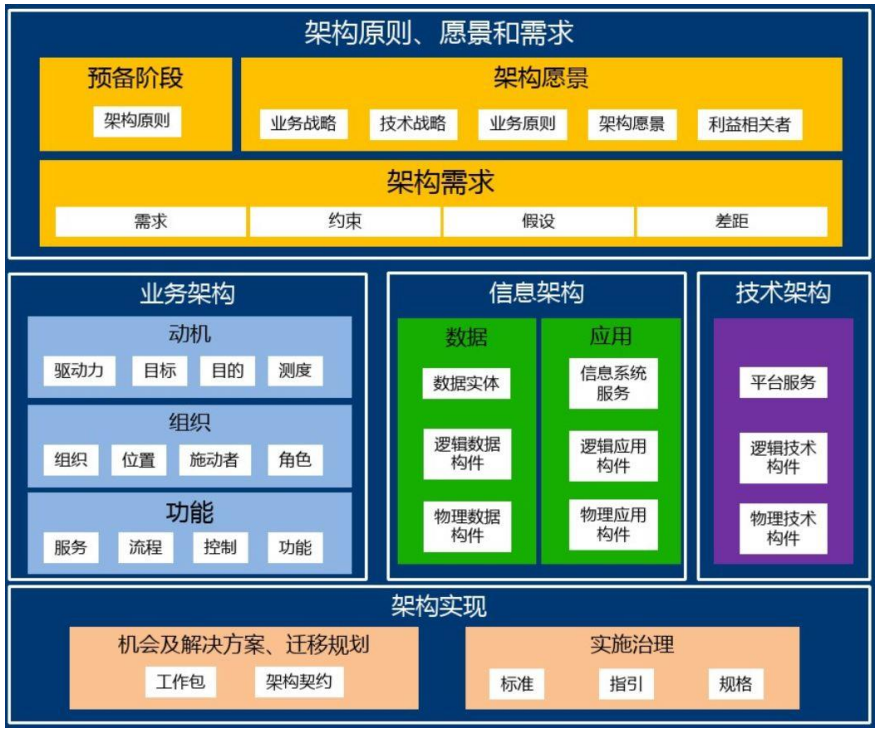
（1）**分类框架**—将指引企业架构的结构和视图组织起来。框架定义构件的标准语法来描述以上视图以及视图之间的关系。构件大多数是图形、表格和矩阵。（2）**流程框架**—规定业务和系统规划分析，以及流程的设计方法。有些IT规划和软件开发生命周期（SDLC）包括其自定义的复合分类。所有场景均通用的流程框架并不存在，要根据需求设计专有流程。

目前已经存在一些框架，例如：

（1）TOGAF。（2）ANSI/IEEE 1471-2000。（3）Zachman企业框架。

从1987年的Zachman Framework开始，多位专家与组织都试图对企业架构的内涵进行定义，影响力比较大的有Zachman架构框架、开放组体系结构框架（TOGAF）等。

- 无论企业还是组织，其存在必定有相应的业务目标，也就是创造价值的方向。同时围绕这一目标，会设置相应角色、环节，形成一定的流程。不管是固化的流程，还是动态调整的流程，“凡经过、必有痕迹”，借此，可对数据资产进行探寻。
- 参考开放组体系结构框架（TOGAF），可以建立从业务架构、应用架构到数据架构和技术架构的整个企业架构体系，这样就能够以业务为重心，建立其利益相关者可理解的、稳定且完整一致的方式，用于定义主要的数据类型和所需数据源。



数据架构的基本内容

随着数据治理理念的不断升级以及数据架构的不断完善，目前企业、数据治理组织的理论准备工作已经逐步收敛到四个方向：

- **数据资产目录**
- **数据标准**
- **数据模型**
- **数据分布**

即**数据架构体系的“四个基本内容”**。

其具体内容包括：**梳理企业的数据资产、制定数据标准并持续维护、建立数据模型，包括概念模型、逻辑模型和物理模型、管控数据分布，包括数据源头和流向。**

1.数据资产目录

关于数据治理与数据资产目录

- 数据治理为数据资产管理提供指导。
- 数据资产目录是实现这一指导的具体形式。

数据资产目录的层级结构

- **主题域分组**: 企业数据管理的顶级分类。
- **主题域**: 互不重叠的数据分类。
- **业务对象**: 数据架构中的重要实体。
- **实体**: 描述业务对象的特性集合。
- **属性**: 描述业务对象的性质和特征。

与数据架构的关系

- 数据架构的所有组件都基于数据资产目录。
- 数据资产目录定义了数据架构的边界和核心骨架。

2.数据标准

数据标准的挑战

- 实现统一标准对大型企业特别是数字化转型中的企业是一大挑战。
- 需要考虑未来的对象与属性命名规范与过去的数据整合。
- 既要适应业务部门，又要考虑技术部门的原则。

数据标准的层次

- **业务术语:** 业务部门提出的业务词汇，是数据标准的初级形式。
- **数据标准:** 通过编码、业务定义、分类分级和质量规范升华的业务术语。
- **数据字典:** 技术部门为管控数据模型而产生的表结构和字段定义规范。

数据标准的价值

- 纵向: 提供新的解读、应用、和价值给历史数据。
- 横向: 通讯不同部门和团队，消除数据的重复和歧义。
- 让数据真正转化为企业资产。

3.数据模型

定义与作用

- 数据模型：描述数据、数据语义、数据关系及约束的工具
- 用途：将实际问题的需求转化为描述性需求

数据模型分类

- 概念数据模型：需求分析产出
- 逻辑数据模型：需求分析产出
- 物理数据模型：设计活动产出

核心术语

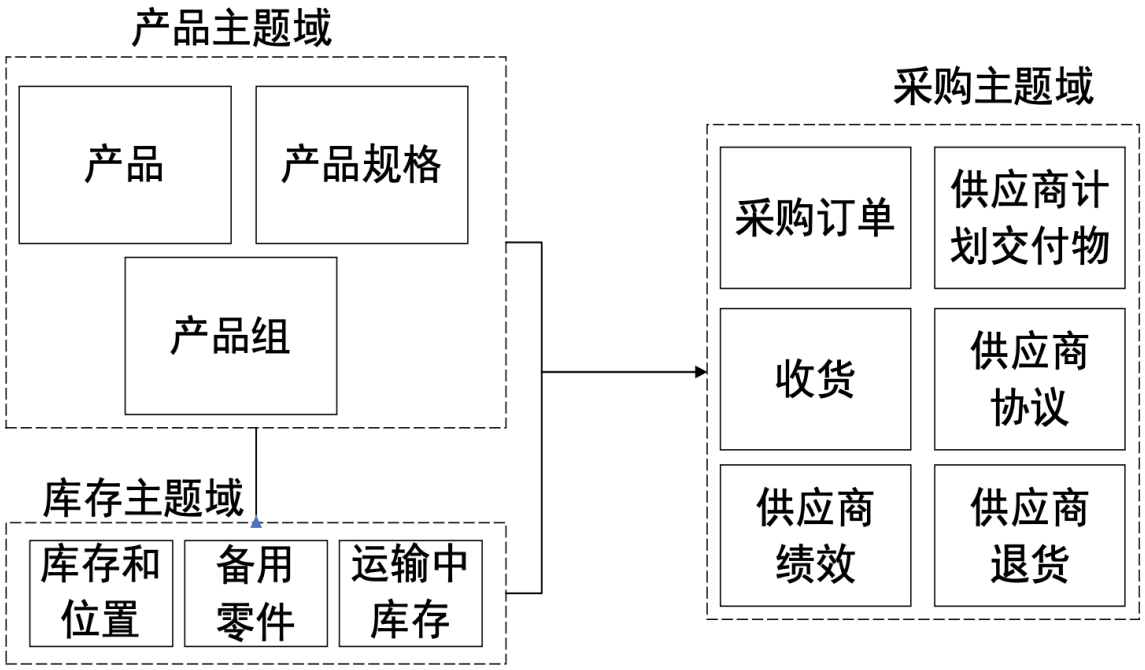
- **实体 (Entity)**：对象如采购订单、产品、客户等
- **属性 (Attribute)**：描述实体，如产品编号、客户电话
- **键属性 (Key Attribute)**：唯一识别数据实体，如客户编号
- **关系 (Relationship)**：实体间的关系，辨识主键和外键
- **范式 (Normal Form)**：规范属性间的依赖与分解关系

3.数据模型

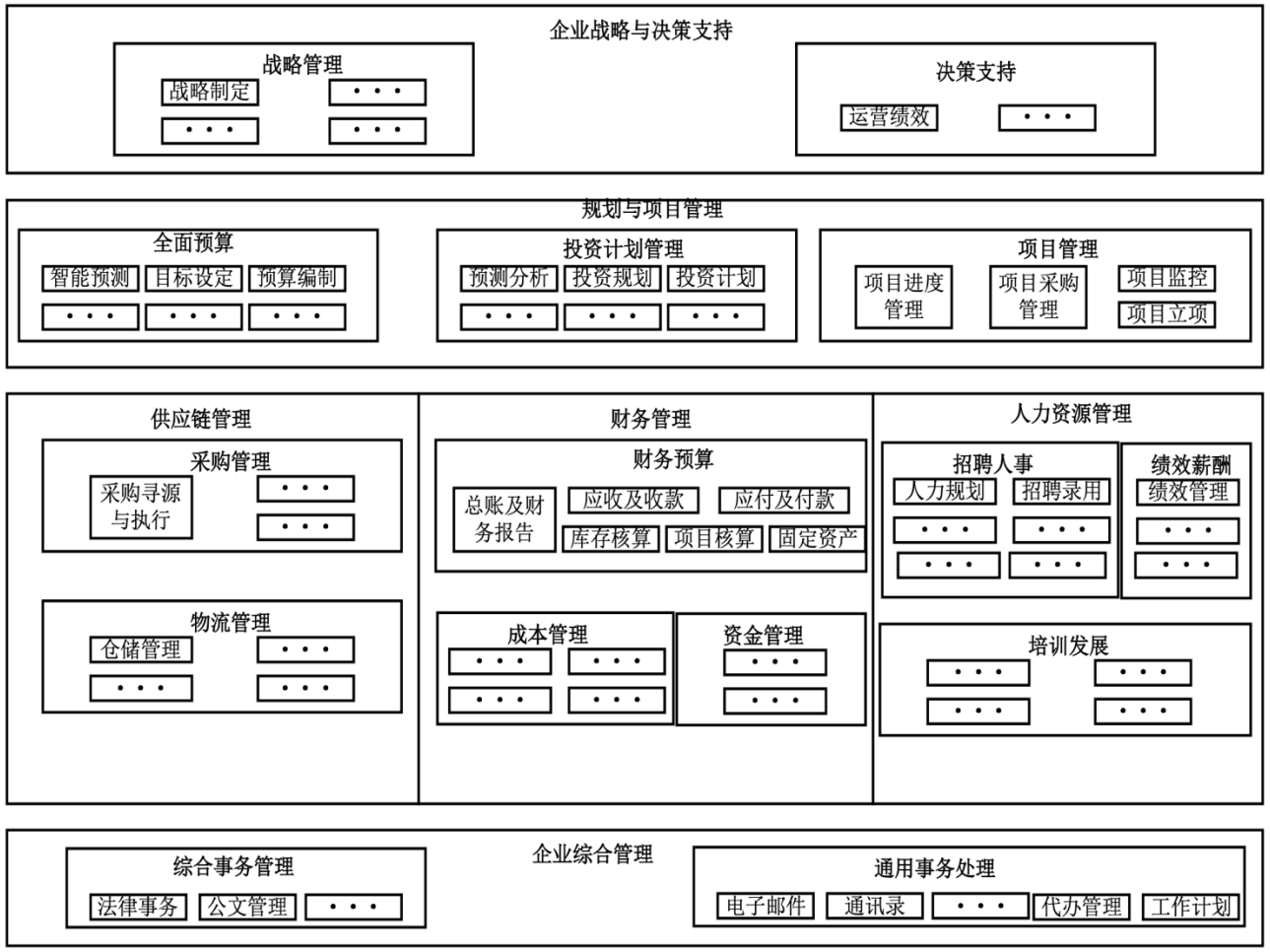
（1）概念数据模型

定义：面向数据库用户描述现实世界的模型。

目的：描述世界的概念化结构，摆脱具体技术问题，专注数据及其联系分析。高阶的概念数据模型可以是数据实体和主题领域的目录清单及组成关系。



构建概念数据模型首先需要对企业数据实体进行梳理



参考行业最佳实践，结合企业实际情况，划分数据主题域

采购与供应商	供应商、采购、合同
项目	投资计划、项目
物资	物料、实物、仓储、库存、配送、第三方物流、供应商管理库存
财务与资产	财务核算、成本、固定资产、预算、资金
人力资源	人事管理、招聘、培训与能力发展、薪酬福利、人力绩效
企业综合	企业战略、企业绩效、法律事务、安保信息、内审内控、公共关系、质量管理、知识管理、创新管理、工会管理、党群管理、后勤管理、公文管理、档案管理、会议管理、督察督办、纪检监察
通用业务	位置、企业协同、信息发布、门户

采购与供应商主题域的详细说明

主题域	主题域描述	实体类别	实体描述	实体名称
采购与供应商	采购与供应商主题领域涉及在采购管理中与供应商发生的相关信息。主要包括采购信息，供应商信息以及合同信息	采购	采购类别主要描述采购管理应用中涉及的有关需求预测，需求提报，采购计划，采购寻源，采购执行，采购评估，跟踪监控等方面信息	采购需求预测、物料使用量预测、物料需求提报信息、物料满足方案、采购申请、采购计划、采购进度、寻源方式、寻源进度、中标信息、投标信息、招标信息、评标信息、评标专家信息、采购订单、提前到货订单、无订单采购信息、采购目录、采购评估、暂收订单
		供应商	供应商类别描述供应商相关信息，主要包括供应商基本信息，供应商绩效，供应商分级等	供应商基本信息、供应商绩效、供应商认证信息、供应商评级、潜在供应商信息
		合同	合同类别主要描述与供应商发生的具有法律强制效力的合同信息，以及在合同履行过程中发生的违约，纠纷等信息	合同模板、合同基本信息、合同审批信息、合同立项审核信息、合同财务审核信息、合同法律审核信息、合同违约与争议、合同变更、合同解除、合同跟踪信息

（2）逻辑数据模型

定义：一种图形化的展现方式，一般采用面向对象的设计方法，有效组织来源多样的各种业务数据，使用统一的逻辑语言描述业务。

功能: 逻辑数据模型借助相对抽象、逻辑统一且稳健的结构，实现数据仓库系统所要求的数据存储目标，支持大量的分析应用，是实现业务智能的重要基础，也是数据管理分析的工具和交流的有效手段。

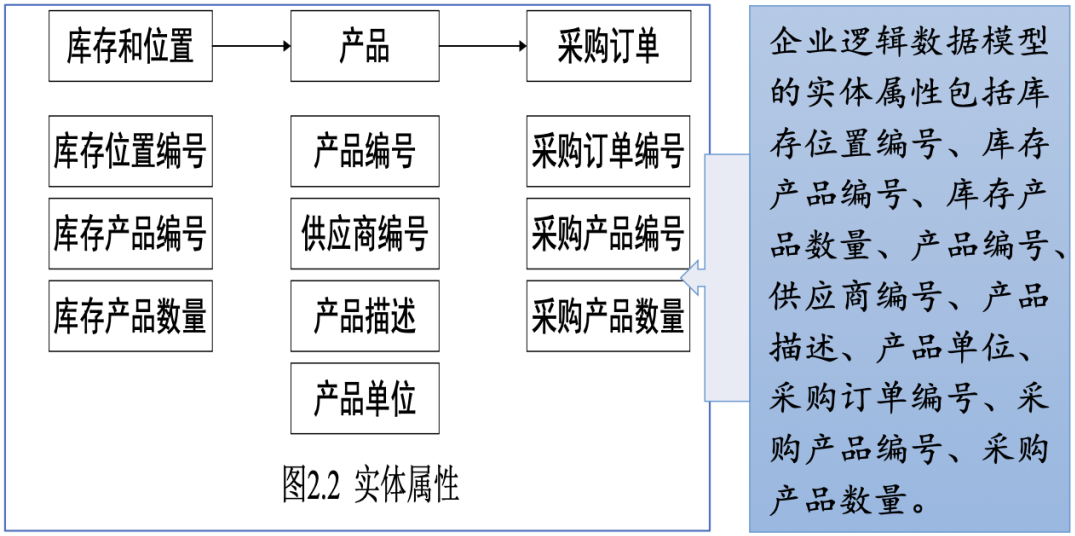


图2.2 实体属性

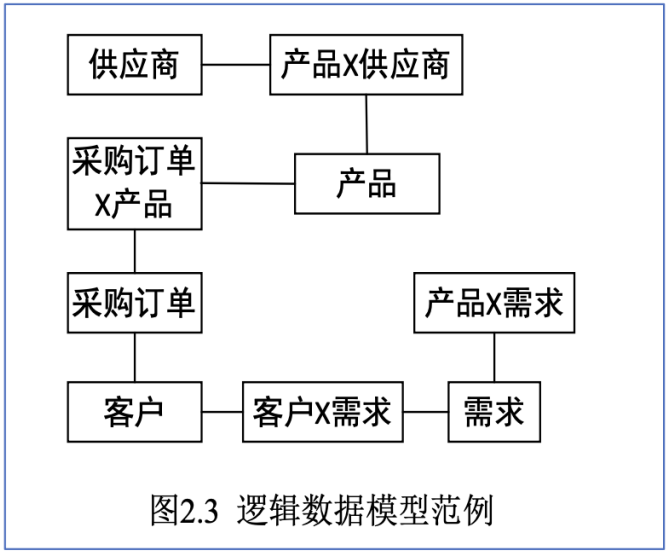
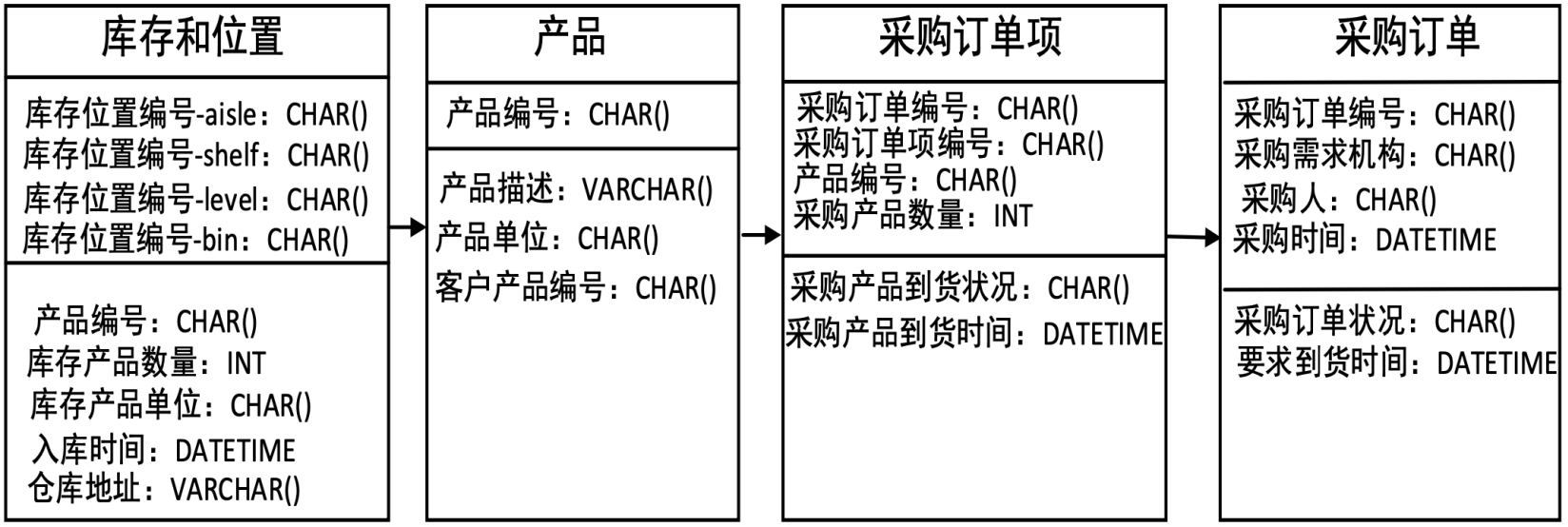


图2.3 逻辑数据模型范例

（3）物理数据模型

定义：物理数据模型是指提供系统初始设计所需要的基础元素，以及相关元素之间的关系。物理数据模型是在逻辑数据模型的基础上，考虑各种具体的技术实现因素，进行数据库体系结构设计，真正实现数据在数据库中的存放。

功能：描述记录结构、顺序和访问路径；可用于系统层实现数据库。



4.数据分布规划

数据分布规划描述企业数据模型在企业信息系统如何分布，通过了解数据分布可以清楚定义企业数据在信息系统中如何产生和使用。

第一步：明确数据主题域所归属的系统，系统及模块。

系统	企业门户	规划与项目管理系统		供应链管理系统		MIS系统		财务管理系统			人力系统	企业综合管理系统		
模块	门户	全面预算	投资计划	供应商管理	采购管理	HR MIS	财务MIS	作业成本管理	综合查询	资金管理	人力绩效	战略管理	内审内控	知识管理
采购与供应商				✓	✓									
财务与资产		✓					✓	✓	✓	✓				
人力资源						✓					✓			
项目			✓											
企业综合												✓	✓	✓
通用业务	✓													

第二步：明确CRUD。CRUD 是建立 （Create）、读取 （Read）、更新 （Update）及删除（Delete）这四项操作的缩写。

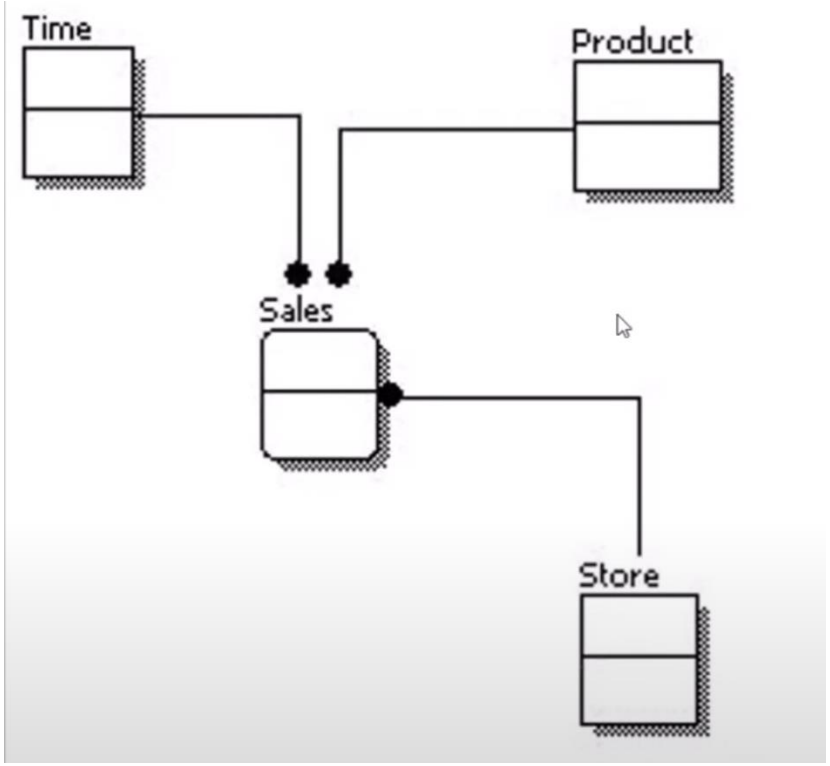
主题域	实体类别	实体名称	项目管理	供应商管理	采购管理	物理管理	合同管理	财务MIS	内审内控	综合办公	协同工作	决策支持
采购与供应商	供应商	供应商基本信息	R	R	R	R	R	CRUD			R	R
		潜在供应商信息			CRUD							R
		供应商绩效			CRUD			R				R
		供应商认证信息			CRUD							R
		供应商评级			CRUD			R				R
	合同	合同模板					CRUD					R
		合同基本信息	R		R	R	CRUD	R	R	R		R
		合同审批信息					CRUD					R
		合同立项审核信息	CRUD				R					R
		合同财务审核信息					R	CRUD				R
		合同法律审核信息					R			CRUD		R
		合同违约与争议					CRUD			R		R
		合同变更	R		R	R	CRUD	R	R	R		R
		合同解除	R		R	R	CRUD	R	R	R		R
		合同跟踪信息	R		R		CRUD					R

数据模型练习

关于产品的销售场景：

已知销售记录（Sales）与产品（Product）、店铺（Store）以及时间（Time）存在关联

其中以销售记录关联产品、店铺以及时间三个实体

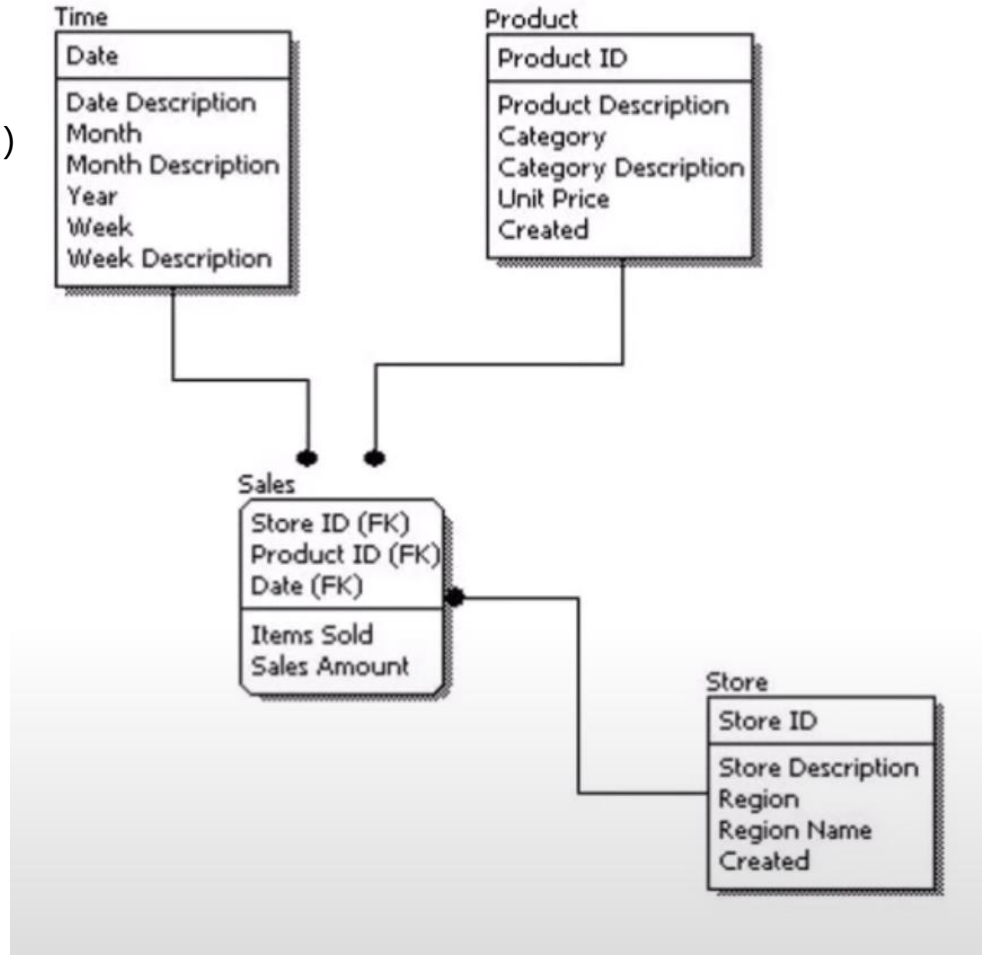


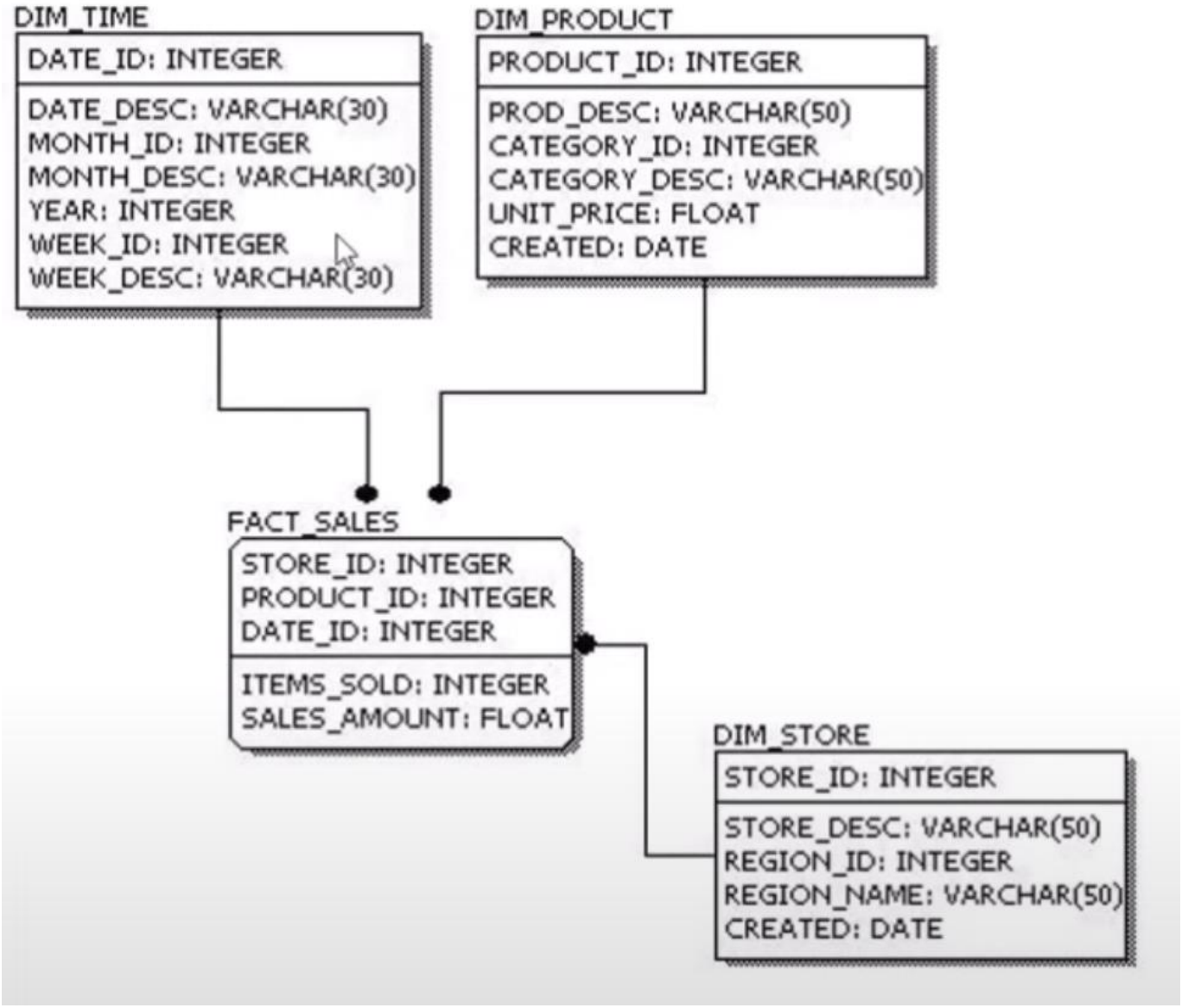
数据模型练习

结合逻辑模型梳理业务及数据规则：

1. Time实体对日期进行（Date为主键）
2. Product实体对产品进行描述（Product ID为主键）
3. Store实体对店铺进行描述（Store ID为主键）
4. Sales实体对销售记录进行描述（仅外键）

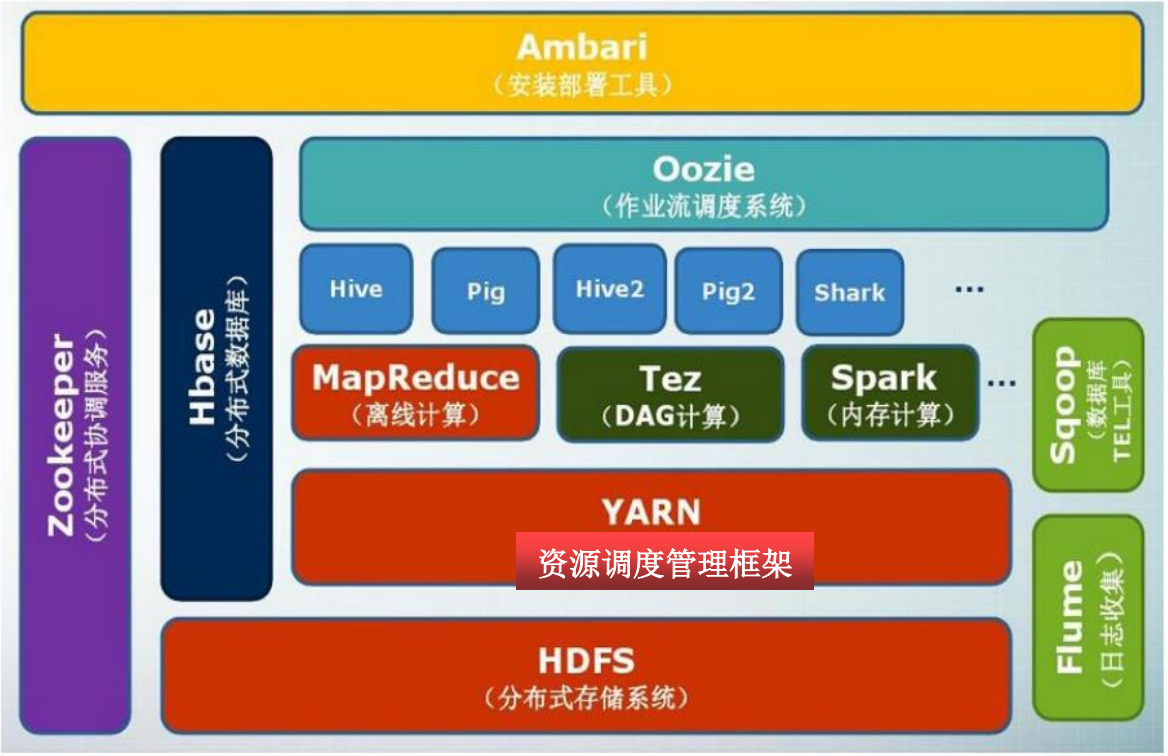
1. 每天可售卖来自多家店铺的多种商品
2. 每次可售卖同一商品的多件
3. 记录每种商品在不同店铺每天的销售总额




DIM_TIME
DATE_ID: INTEGER
DATE_DESC: VARCHAR(30)
MONTH_ID: INTEGER
MONTH_DESC: VARCHAR(30)
YEAR: INTEGER
WEEK_ID: INTEGER
WEEK_DESC: VARCHAR(30)
DIM_PRODUCT
PRODUCT_ID: INTEGER
PROD_DESC: VARCHAR(50)
CATEGORY_ID: INTEGER
CATEGORY_DESC: VARCHAR(50)
UNIT_PRICE: FLOAT
CREATED: DATE
FACT_SALES
STORE_ID: INTEGER
PRODUCT_ID: INTEGER
DATE_ID: INTEGER
ITEMS_SOLD: INTEGER
SALES_AMOUNT: FLOAT
DIM_STORE
STORE_ID: INTEGER
STORE_DESC: VARCHAR(50)
REGION_ID: INTEGER
REGION_NAME: VARCHAR(50)
CREATED: DATE

基于Hadoop的大数据架构实现

Hadoop通过大量高效的硬件集群和标准接口构建大规模分布式计算系统，为大数据提供存储和计算。Hadoop 的核心组件是 HFDS（Hadoop Distributed File System）和Hadoop MapReduce，其他组件为核心组件提供配套和补充性服务。



本章要点

1. 掌握数据架构包括哪些基本内容?
2. 掌握数据模型的内容与不同模型的区别?