

数据仓库理论基础与企业应用场景

实体关系（ER）建模理论及
应用场景案例

Contents // 目录

01 认识数据仓库

02 数据仓库理论基础

03 实体关系（ER）建模理论
及应用场景案例

04 数据仓库与维度建模

05 实战案例-偏业务型行业数
据仓库设计

实体关系（ER）建模 理论及应用场景案例



PART 01

基础概念：模式、模型、
数据建模与数仓建模



PART 02

数仓分层

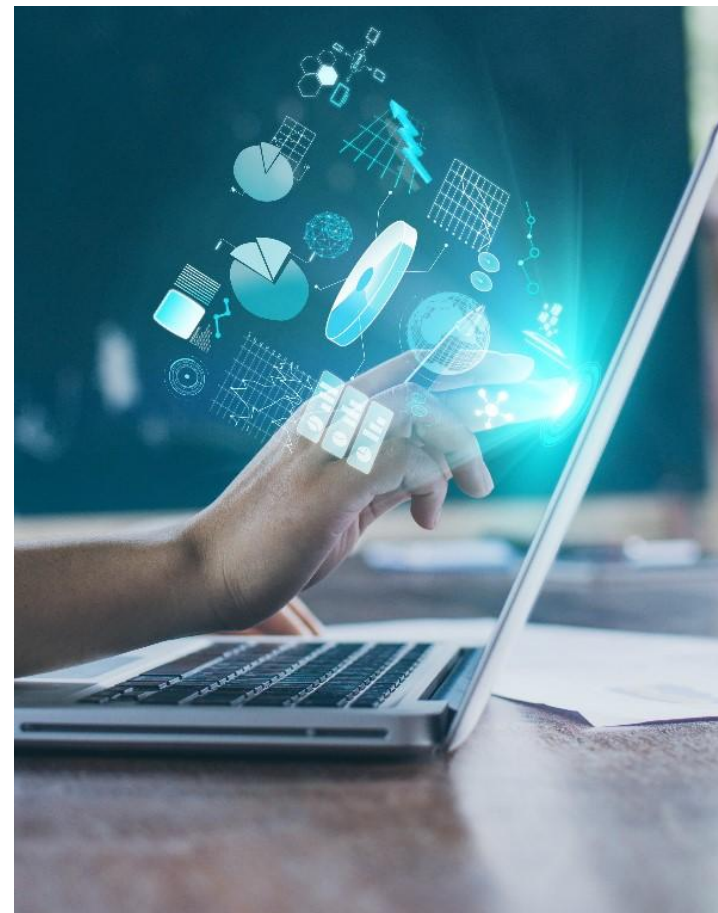


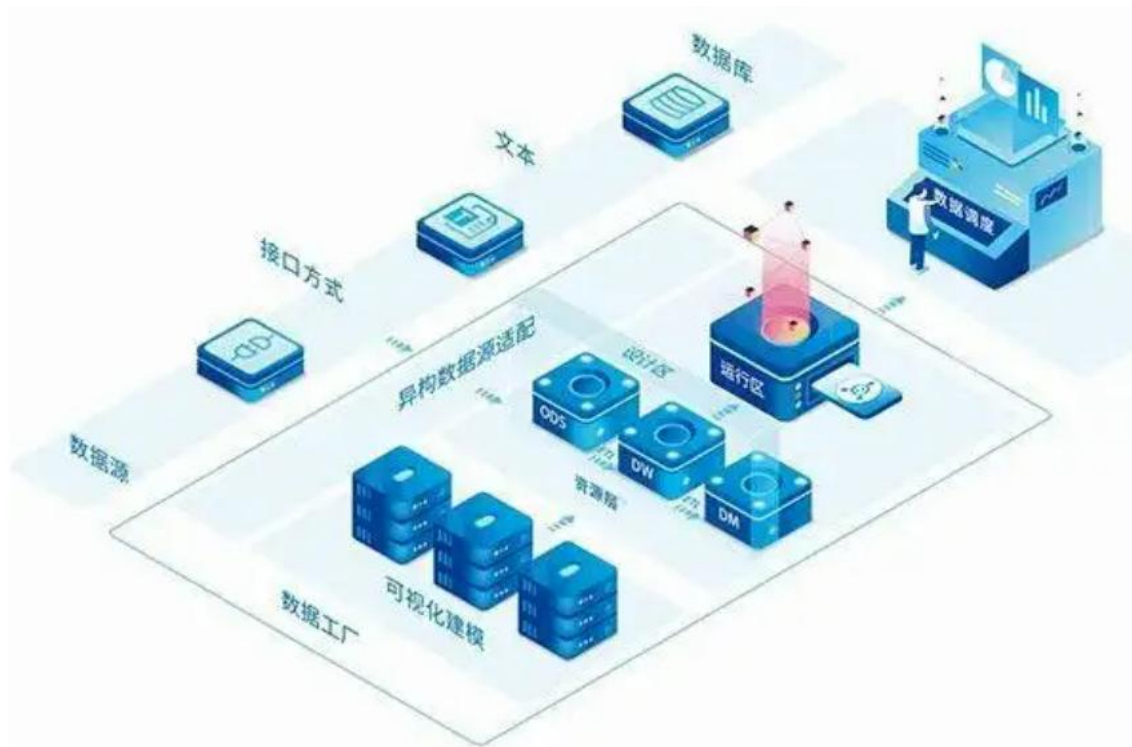
PART 03

数仓建模：范式建模



建模，就是建立模型，就是为了理解事物而对事物做出的一种抽象，是对事物的一种无歧义的书面描述。建立系统模型的过程，又称模型化。建模是研究系统的重要手段和前提。凡是用模型描述系统的因果关系或相互关系的过程都属于建模。





- 什么是模式？
- 什么是模型？
- 模式与模型的区别
- 数据建模
- 数仓建模

Part-01：基础概念：模式、模型、数据建模与数仓建模

基础概念：模式、模型、数据建模与数仓建模



- 什么是模型？

说到模型，还有另外一个比较容易搞混的概念：什么是模式？

从字面的意思理解，“模”一种标准，或者一种套路，“式”方式，方法，形式。两个字连接在一起就可以解释为，一种可以重复使用，具有参考性的方法、知识体系。

● 什么是模式？

- 互动百科：模式是指从生产经验和生活经验中经过抽象和升华提炼出来的核心知识体系。模式（Pattern）其实就是解决某一类问题的方法论。把解决某类问题的方法总结归纳到理论高度，那就是模式。模式是一种指导，在一个良好的指导下，有助于你完成任务，有助于你作出一个优良的设计方案，达到事半功倍的效果。而且会得到解决问题的最佳办法。



基础概念：模式、模型、数据建模与数仓建模



- 什么是模型？

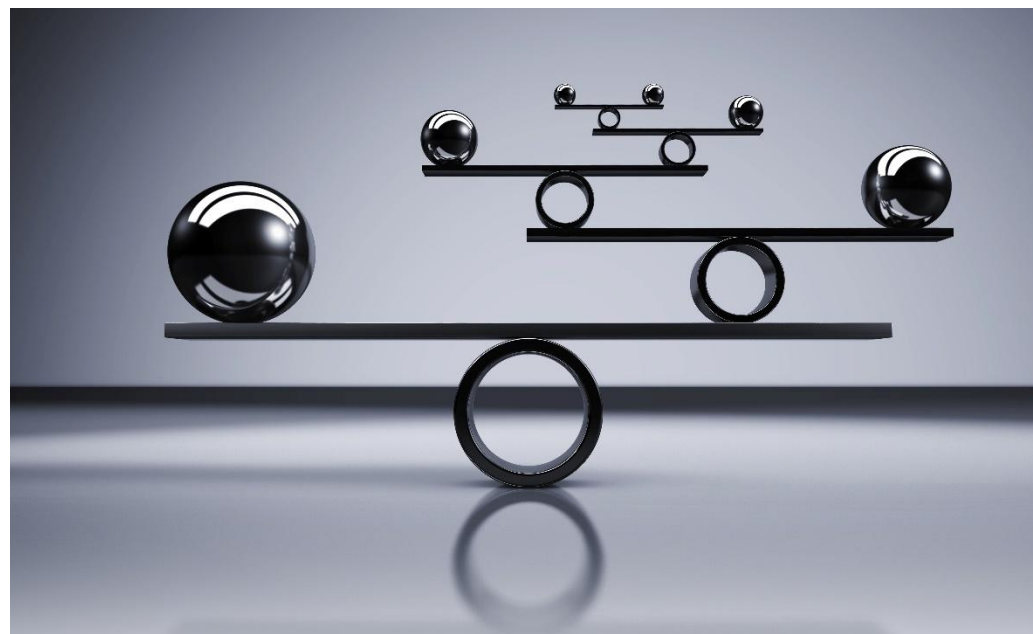
- 现实中我们经常听到各种各样的模型，比如数学模型、算法模型、数据模型、概念模型、内存模型、领域模型、编程模型、行业研究模型。。。。。



到底什么是模型？

● 什么是模型？

- 模型（model）是对现实客观事物的一种表示和抽象，它可以是文字、图表、公式，也可以是计算机程序或其他实体模型。可以说，模型是把对象实体通过适当的过滤，用适当的表现规则描绘出的简洁的模仿品，通过这个模仿品，人们可以了解到所研究实体的本质，而且在形式上便于人们对实体进行分析和处理。



模型不只可以描述实物，还可以描述虚拟物件。比如建筑模型、汽车模型、飞机模型、订单模型等。

基础概念：模式、模型、数据建模与数仓建模

● 模型与模式的区别？

- 模型、模式都是对现实事物（也就是“象”）的特征提取。模型更接近于实体（/“象”），模式更接近于人脑（/“术”）、更抽象。于是，模型更易上手，但适用范围狭窄；模式更难落地，但适用范围宽广。
- 因此，两者的结合更显重要。一个是拐杖（“模型”），一个是地图（“模式”），都是我们改造世界、正确前行的依靠。



基础概念：模式、模型、数据建模与数仓建模

● 什么是数据模型？

■ 百度百科的定义

数据模型

 语音

 编辑

 讨论

 上传视频

 本词条由“科普中国”科学百科词条编写与应用工作项目 审核。

数据模型（Data Model）是数据特征的**抽象**，它从抽象层次上描述了系统的静态特征、动态行为和约束条件，为数据库系统的信息表示与操作提供一个抽象的框架。数据模型所描述的内容有三部分，分别是数据结构、数据操作和数据约束 ^[1]。

■ wiki MBA智库的定义：

什么是数据模型

[编辑]

数据模型是现实世界数据特征的抽象，用于描述一组数据的概念和定义。数据模型是数据库中数据的存储方式，是数据库系统的基础。在数据库中，数据的物理结构又称数据的存储结构，就是数据元素在计算机存储器中的表示及其配置；数据的逻辑结构则是指数据元素之间的逻辑关系，它是数据在用户或程序员面前的表现形式，数据的存储结构不一定与逻辑结构一致。

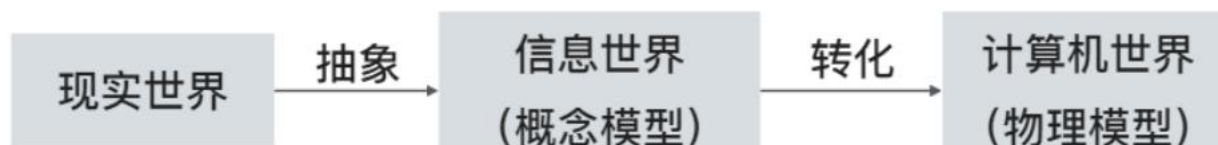
● 什么是数据模型？

■ 通俗版的解释

- ✓ 数据模型是现实世界或业务逻辑在数据层面的投影，是将数据元素以标准化的模式组织起来，用来模拟现实世界的信息框架和蓝图。

■ 目的和作用

- ✓ 方便人与人之间信息的传递和沟通。
- ✓ 方便人们通过数据模型去理解现实世界。
- ✓ 计算机通过算法模型、规则模型，可以预测客观虚拟事物的发展或轨迹。
- ✓ 现实世界的虚拟事物，抽象到信息世界逻辑模型，再转换成计算机世界的物理模型，而计算机能够存储和识别的是物理模型。



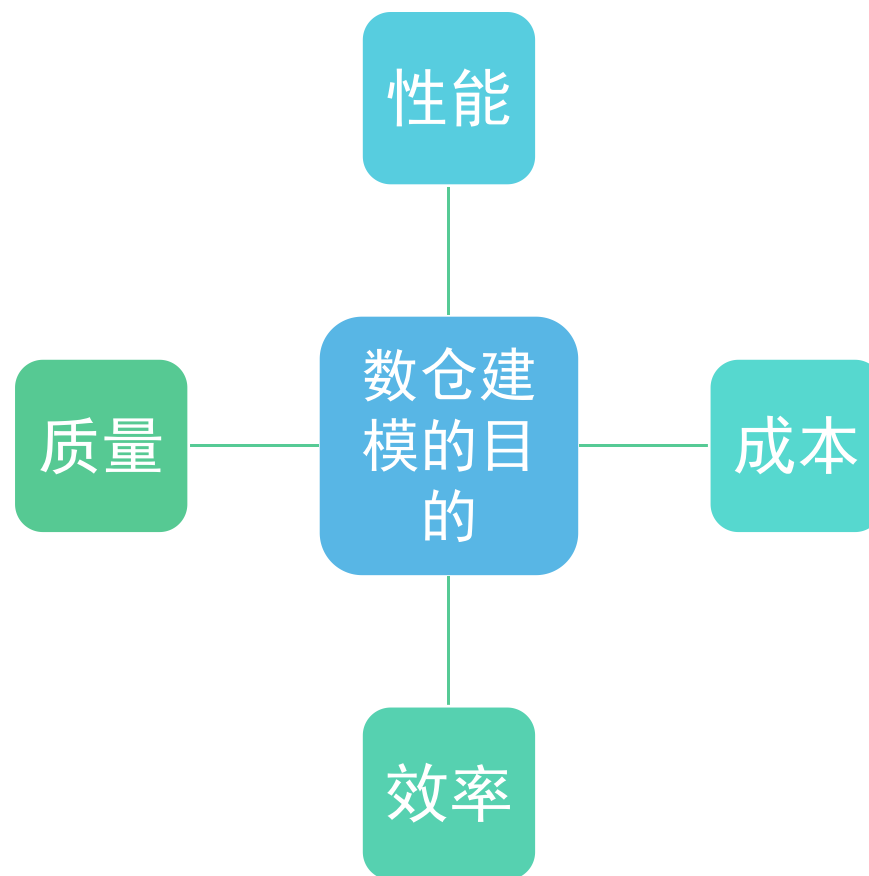
● 数仓建模

■ 什么是数仓建模

- ✓ 我们所说的数仓建模，实际上就是构建一种数据存储模型，用于结构化存储我们日常业务行为或信息化系统存储下来有价值的数据。

■ 数仓建模的目的

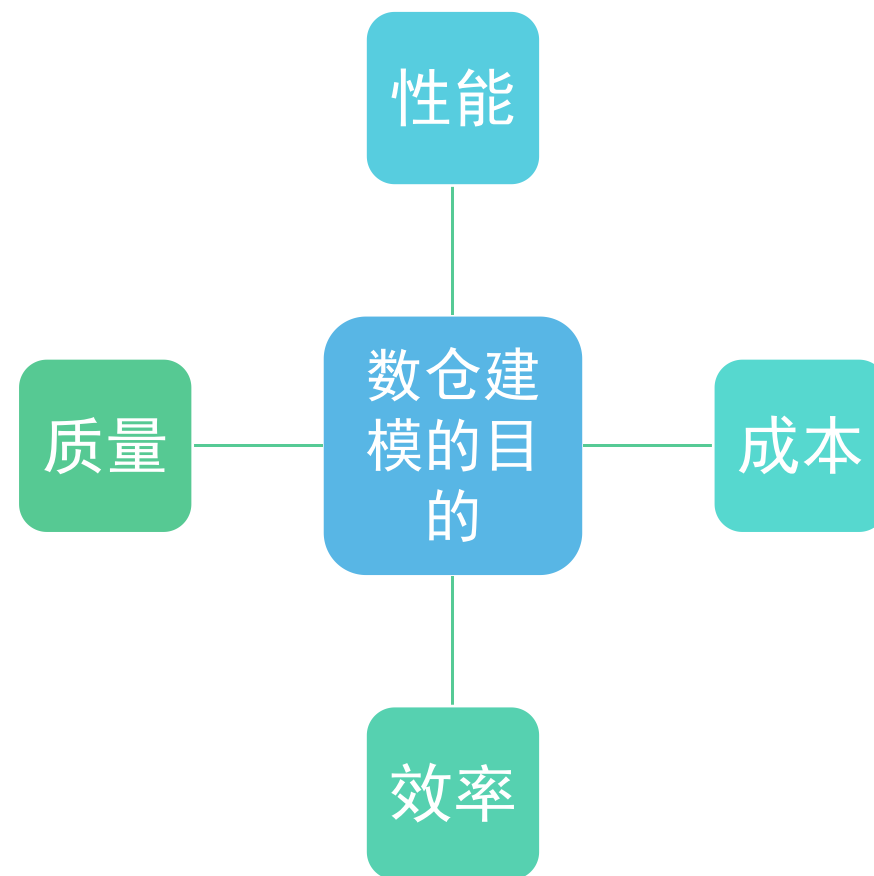
- ✓ 是使用数据建模方法来帮助更好地组织和存储数据，以便在性能、成本、效率和质量之间取得最佳平衡方便人们通过数据模型去理解现实世界。



● 数仓建模

■ 数仓建模的目的

- ✓ **性能**，良好的数据模型能帮助我们快速查询所需要的数据，减少数据的吞吐。
- ✓ **成本**，良好的数据模型能极大地减少不必要的数据冗余，也能实现计算结果复用，极大地降低大数据系统中的存储和计算成本。
- ✓ **效率**，良好的数据模型能极大地改善用户使用数据的体验，提高使用数据的效率。
- ✓ **质量**，良好的数据模型能改善数据统计口径的不一致性，减少数据计算错误的可能性。



■ 高质量数据建模的意义

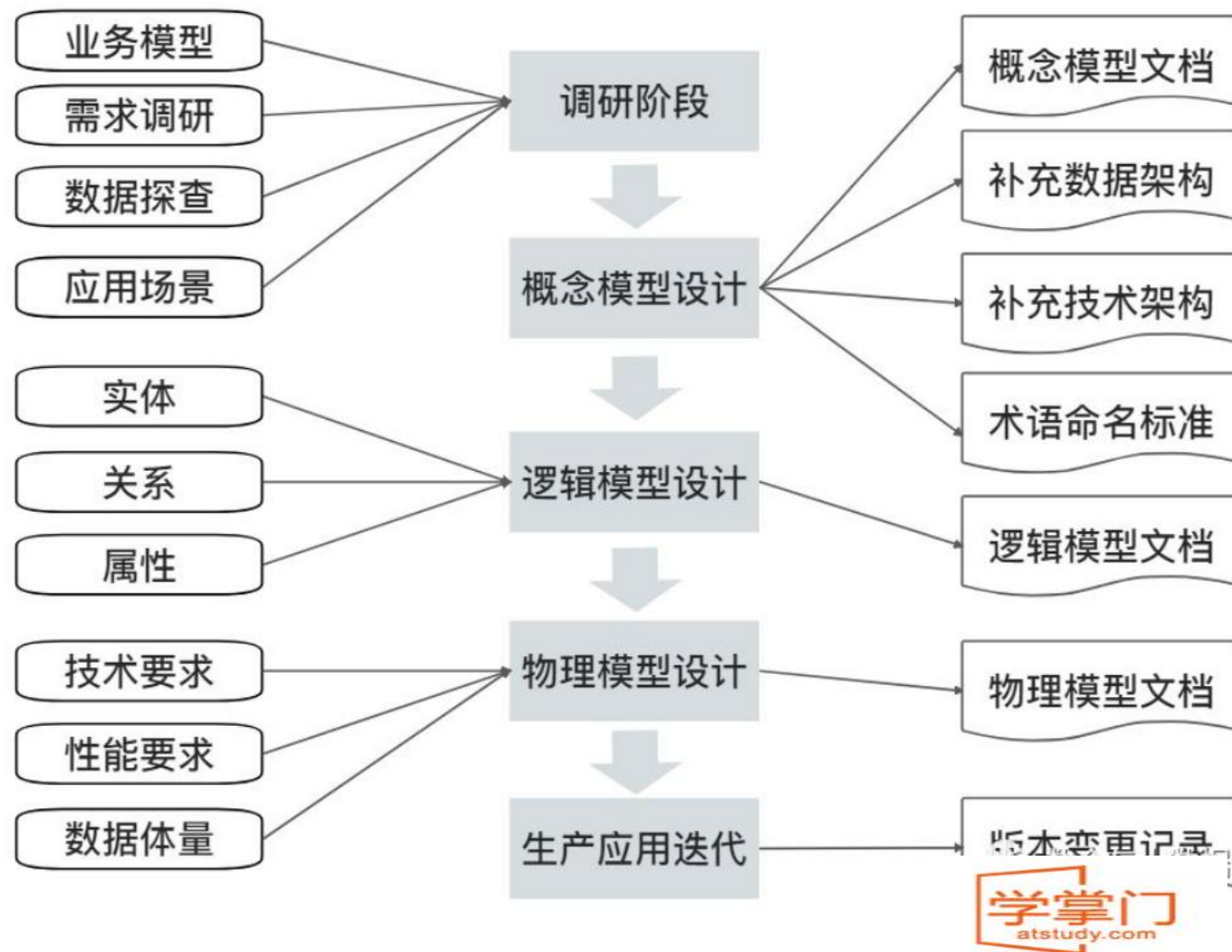
- ✓ 更低的存储成本
- ✓ 更高的查询效率
- ✓ 清晰明了的数据结构方便理解和使用
- ✓ 简化的ETL处理逻辑
- ✓ 更好的数据质量保障（一致性、准确性、完整性、时效性）
- ✓ 更灵活的应对变化
- ✓ 更好的满足客户需求



基础概念：模式、模型、数据建模与数仓建模

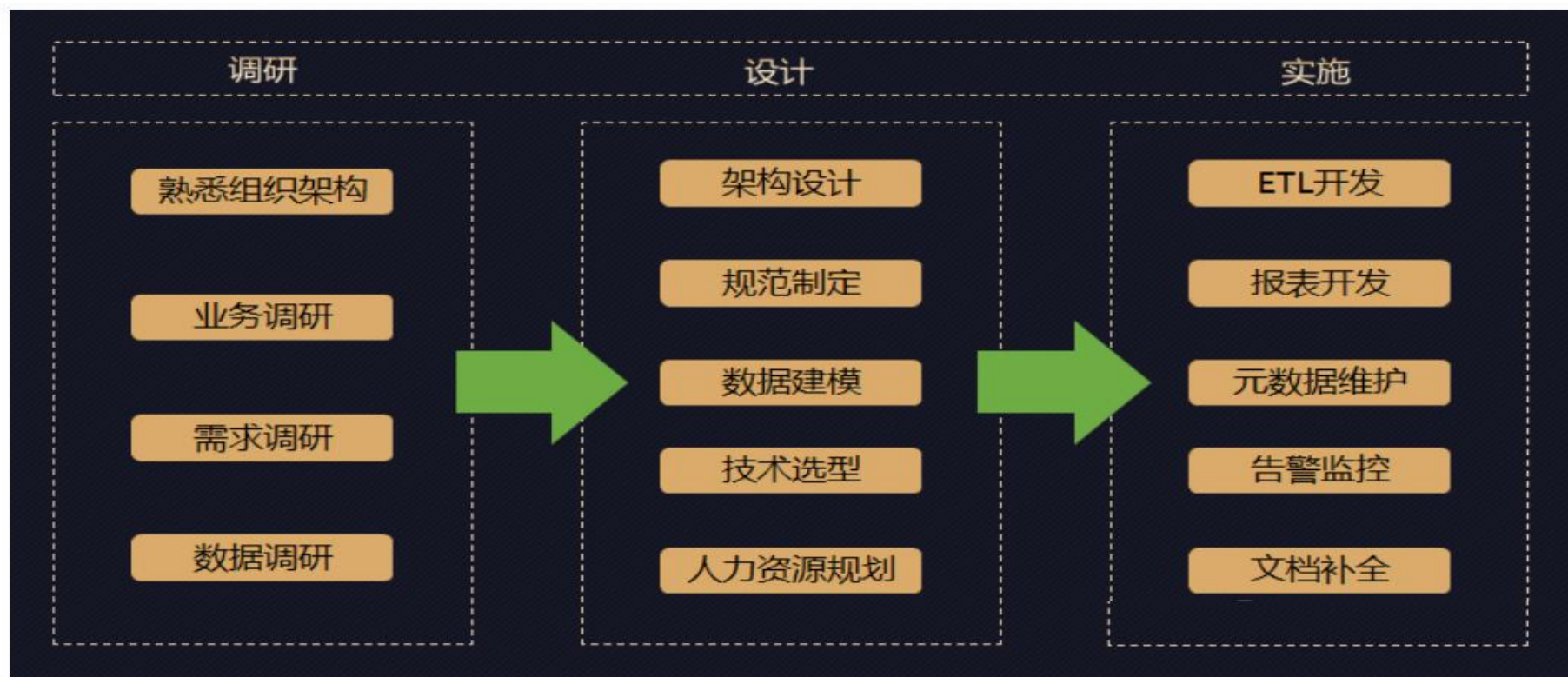
● 通用的数据建模流程

- 调研阶段 -> 概念模型 -> 逻辑模型 -> 物理模型 -> 生产应用迭代

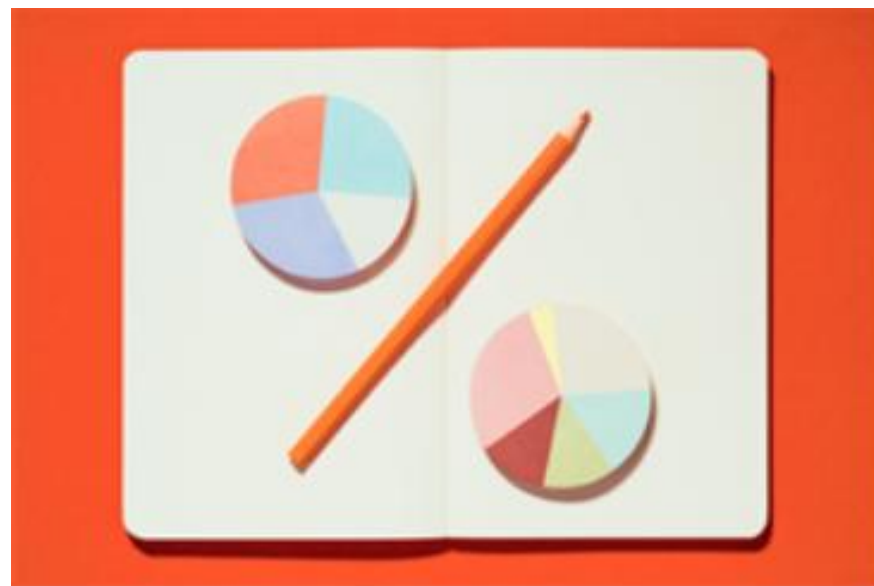


基础概念：模式、模型、数据建模与数仓建模

- 数仓在数据建模中的位置



- **模式**是一种可以重复使用，具有参考性的方法、知识体系。
- **模型**是指对于某个实际问题或客观事物、规律进行抽象后的一种形式化表达方式。
- **数据模型**是现实世界或业务逻辑在数据层面的投影，是将数据元素以标准化的模式组织起来，用来模拟现实世界的信息框架和蓝图。
- **数仓建模**，实际上就是构建一种数据存储模型，用于结构化存储我们日常业务行为或信息化系统存储下来有价值的数据。





- 数仓分层的原则
- 数仓分层的架构
- ODS层、DWD层
- DWM层、DIM层
- DWS层、APP层

Part-02: 数仓分层

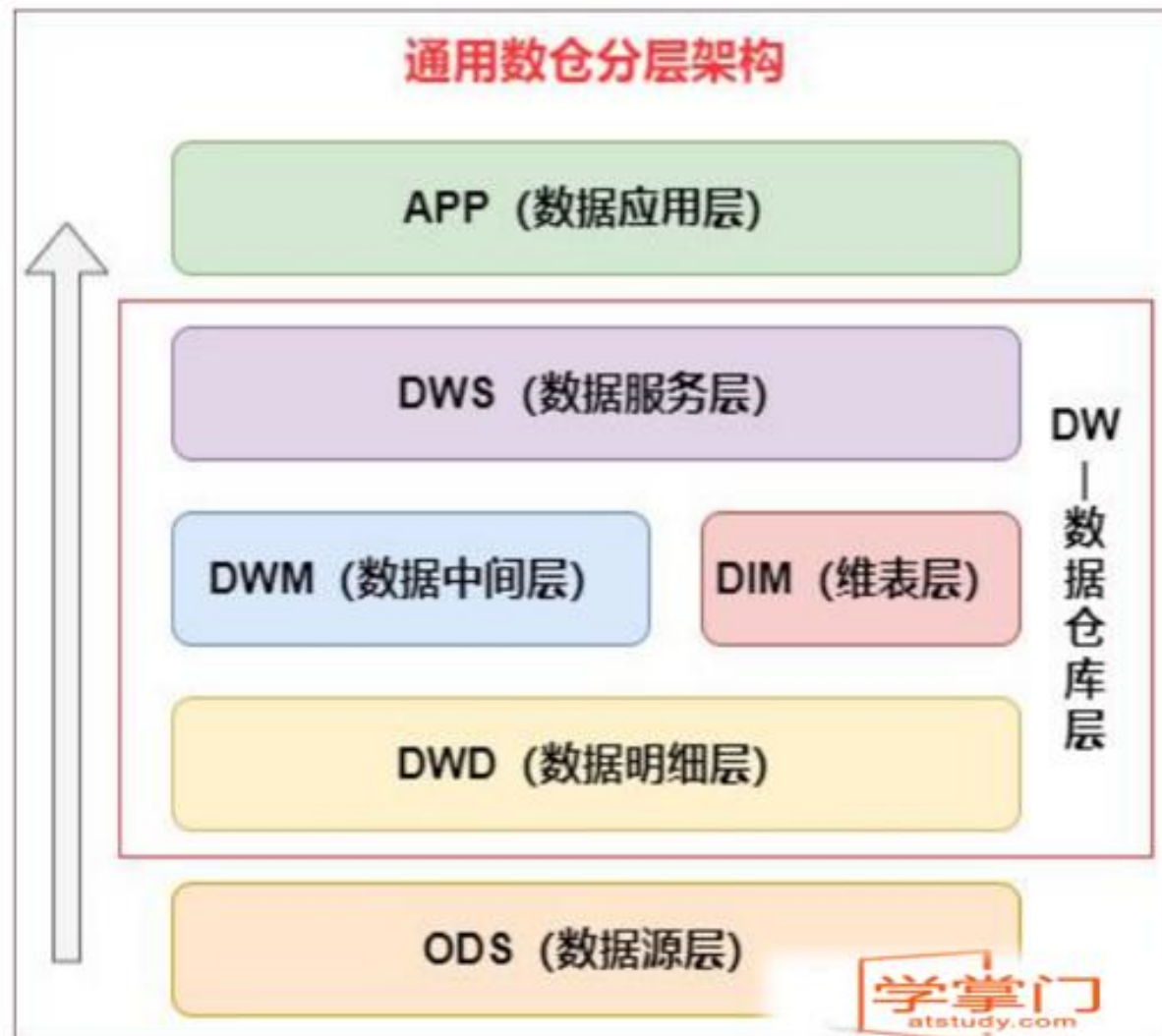
● 数仓分层的原则

- 为便于数据分析，要屏蔽底层复杂业务，简单、完整、集成的将数据暴露给分析层。
- 底层业务变动与上层需求变动对模型冲击最小化，业务系统变化影响削弱在基础数据层，结合自上而下的建设方法削弱需求变动对模型的影响。
- 高内聚松耦合，即主题之内或各个完整意义的系统内数据的高内聚，主题之间或各个完整意义的系统间数据的松耦合。
- 构建仓库基础数据层，使底层业务数据整合工作与上层应用开发工作相隔离，为仓库大规模开发奠定基础仓库层次更加清晰，对外暴露数据更加统一。



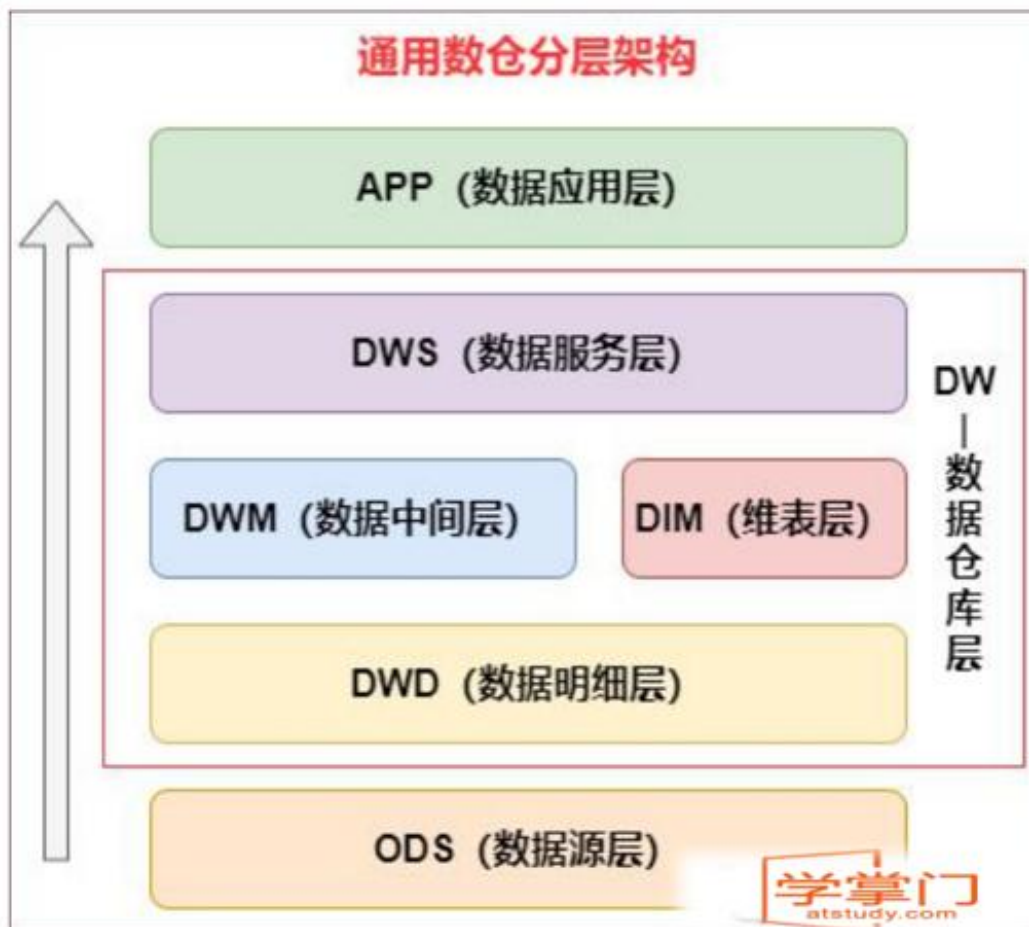
数据仓库架构

- 通用数仓分层架构
(CIF层次架构)



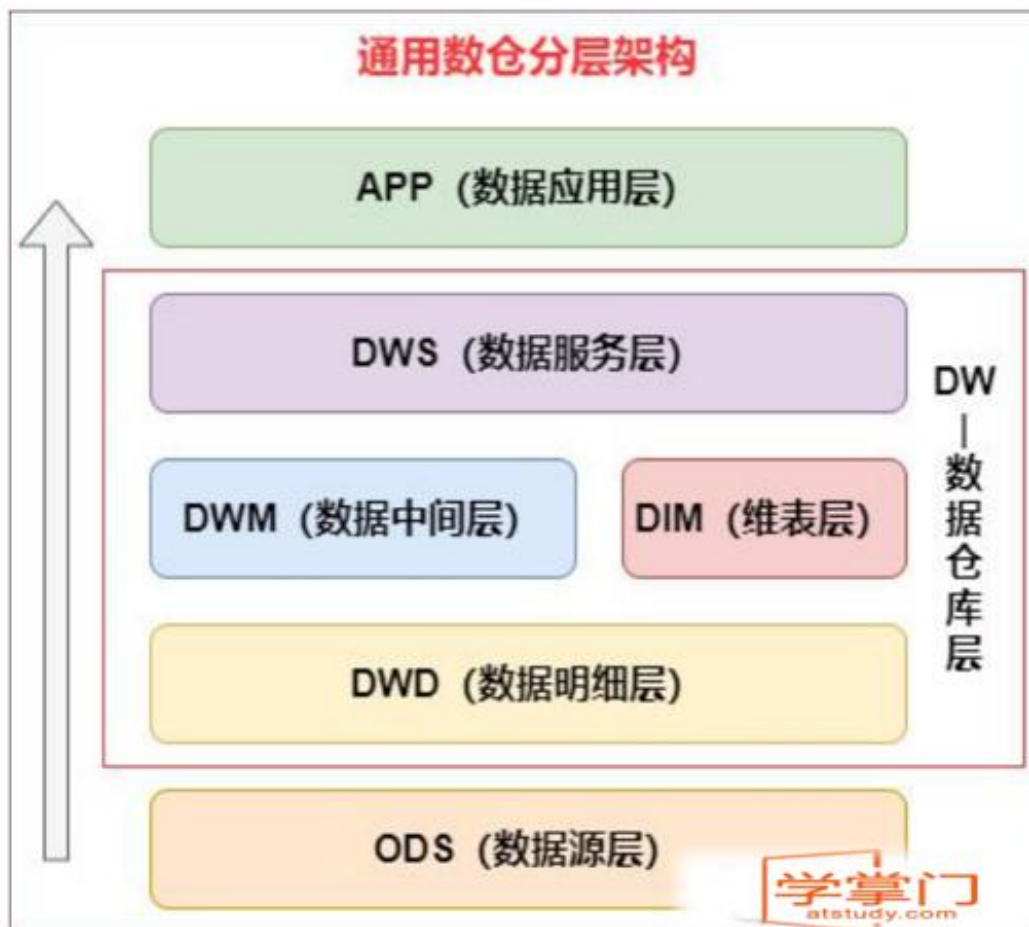
数据仓库架构

- 数据源层（ODS）
 - ODS 层，是最接近数据源中数据的一层，为了考虑后续可能需要追溯数据问题，因此对于这一层就不建议做过多的数据清洗工作，原封不动地接入原始数据即可，至于数据的去噪、去重、异常值处理等过程可以放在后面的 DWD 层来做。



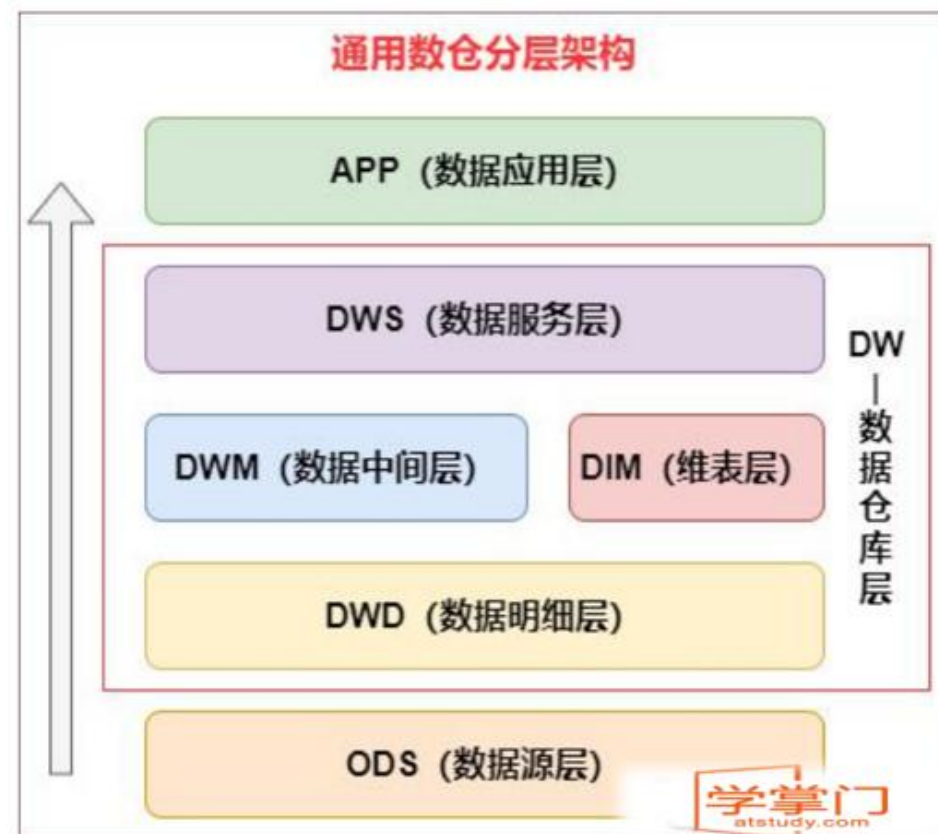
数据仓库架构

- 数据仓库层： DW （Data Warehouse ）
 - 数据仓库层是我们在做数据仓库时要核心设计的一层，在这里，从 ODS 层中获得的数据按照主题建立各种数据模型。
 - DW 层又细分为 DWD （Data Warehouse Detail）层、 DWM （Data WareHouse Middle）层和 DWS （Data WareHouse Servce） 层。



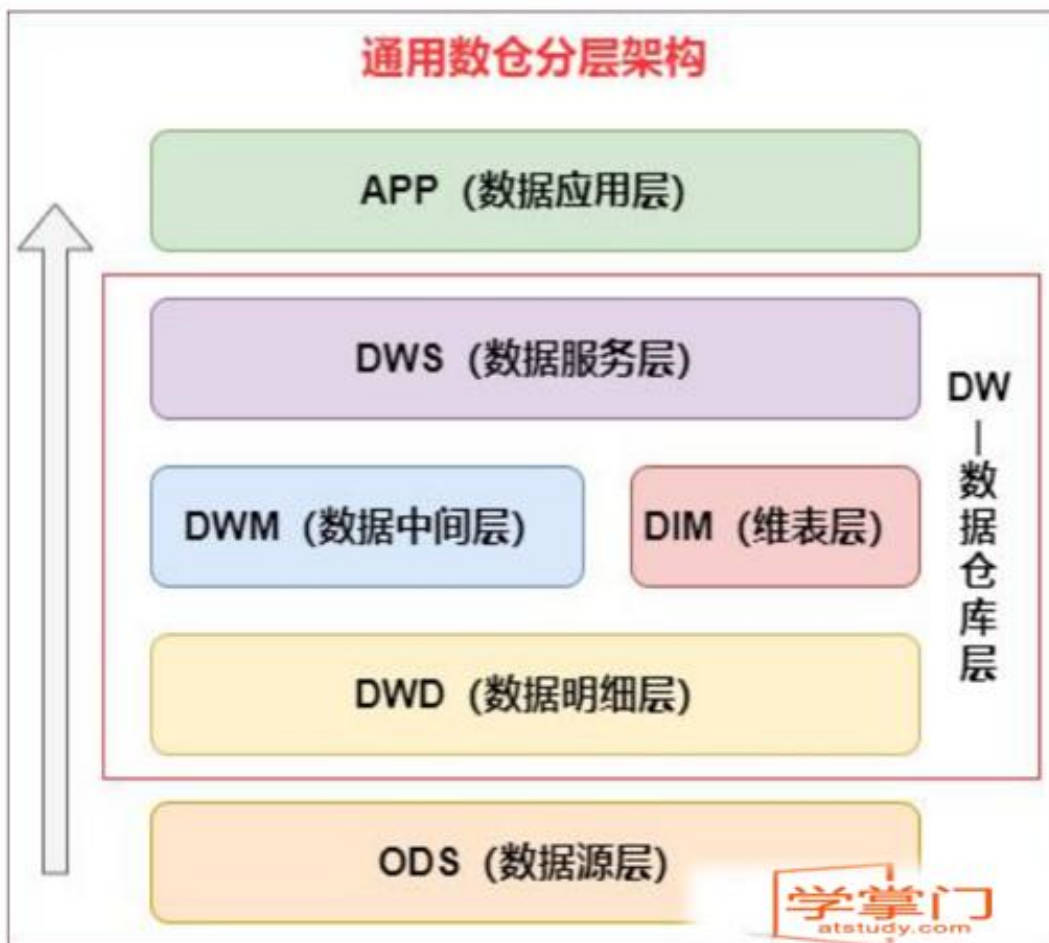
数据仓库架构

- 1) 数据明细层： DWD (Data Warehouse Detail)
 - 该层一般保持和 ODS 层一样的数据粒度，并且提供一定的数据质量保证。
 - DWD 层要做的就是将数据清理 、 整合 、 规范化 、 脏数据 、 垃圾数据 、 规范不一致的 、 状态定义不一致的、命名不规范的数据都会被处理 。
 - 同时，为了提高数据明细层的易用性， 该层会采用一些维度退化手法，将维度退化至事实表中，减少事实表和维表的关联 。
 - 另外，在该层也会做一部分的数据聚合，将相同主题的数据汇集到一张表中，提高数据的可用性。



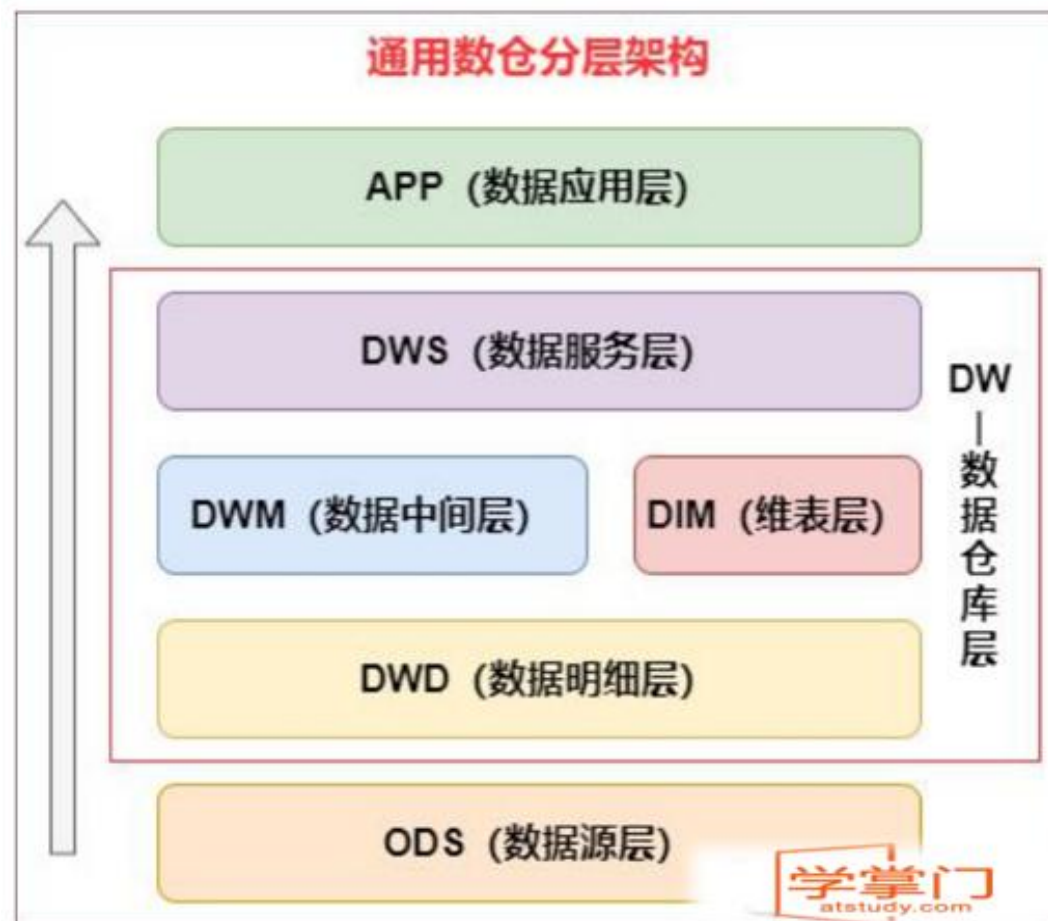
数据仓库架构

- 2) 数据中间层： DWM (Data WareHouse Middle)
 - 该层会在 DWD 层的数据基础上，数据做轻度的聚合操作，生成一系列的中间表，提升公共指标的复用性，减少重复加工。
 - 直观来讲，就是对通用的核心维度进行聚合操作，算出相应的统计指标。



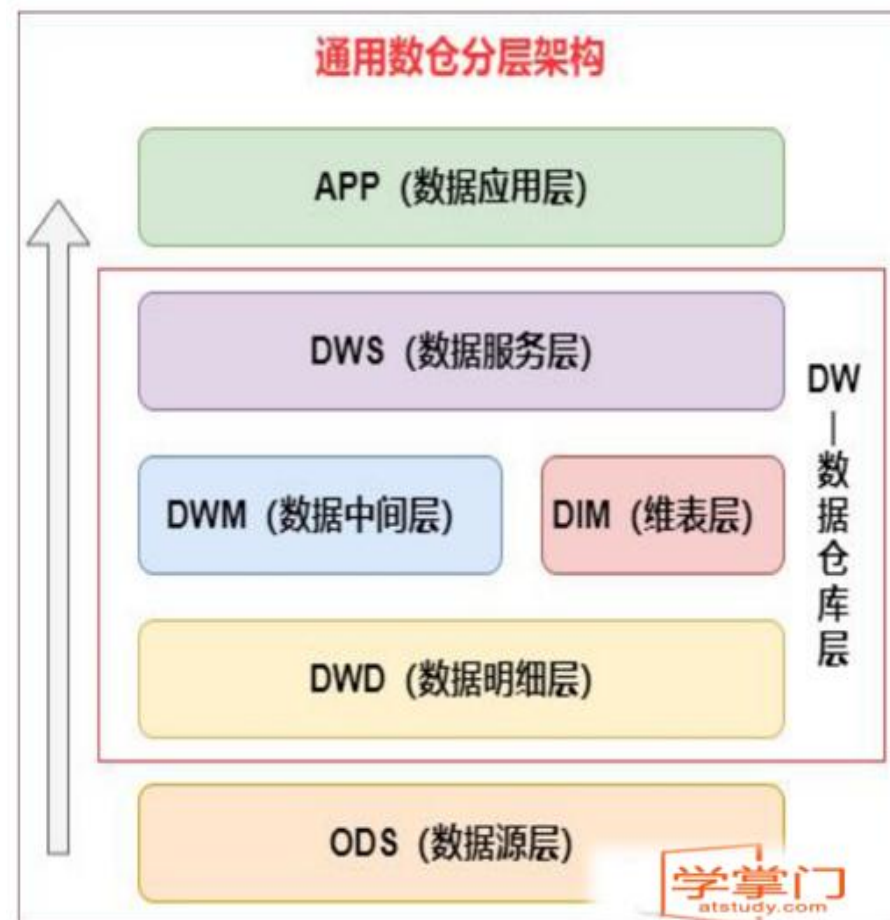
数据仓库架构

- 3) 维表层：DIM (Dimension)
 - 如果维表过多，也可针对维表设计单独一层，维表层主要包含两部分数据：
 - ✓ 高基数维度数据：一般是用户资料表、商品资料表类似的资料表。数据量可能是千万级或者上亿级别。
 - ✓ 低基数维度数据：一般是配置表，比如枚举值对应的中文含义，或者日期维表。数据量可能是个位数或者几千几万。



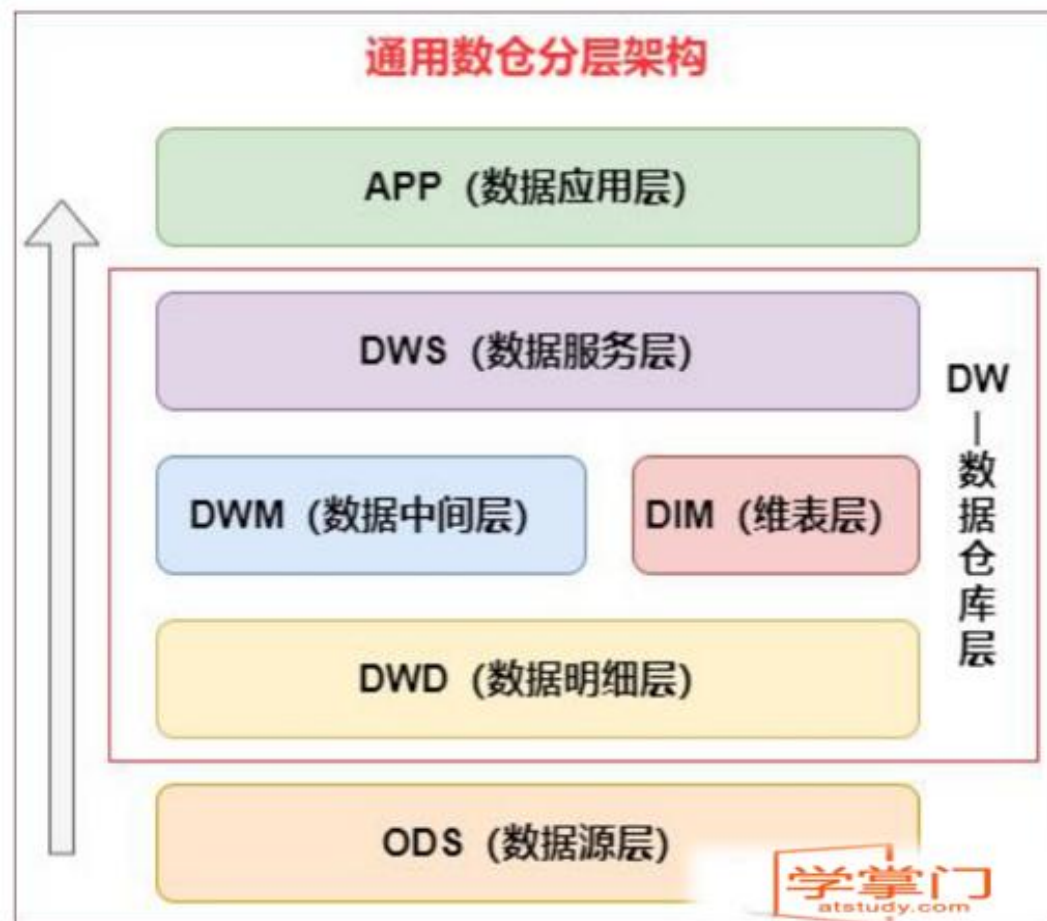
数据仓库架构

- 4) 数据服务层： DWS (Data Warehouse Service)
 - DWS 层为公共汇总层，会进行轻度汇总，粒度比明细数据稍粗，基于 DWD 层上的基础数据， 整合汇总成分析某一个主题域的服务数据 ， 一般是宽表 。
 - DWS 层应覆盖 80% 的应用场景。又称数据集市或宽表。
 - 按照业务划分，如主题域流量、订单、用户等，生成字段比较多的宽表，用于提供后续的业务查询，OLAP 分析，数据分发等。



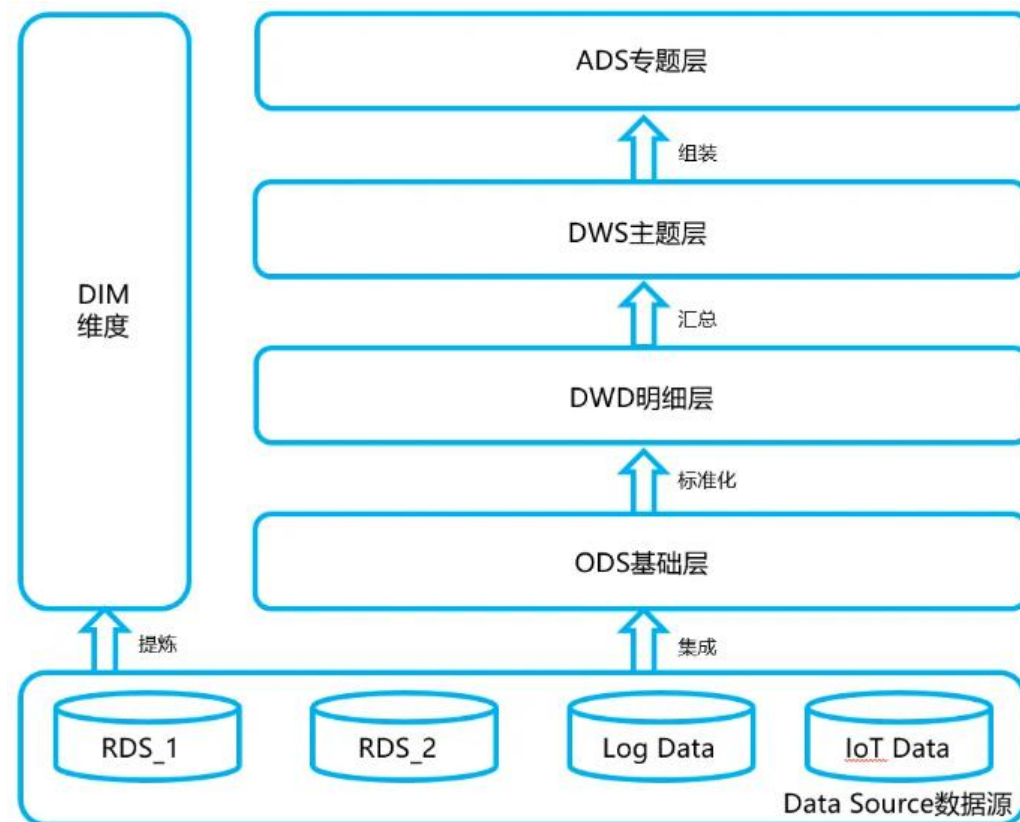
数据仓库架构

- 数据应用层： APP （ Application）
 - 数据应用层， 也有公司或书把这层成为ads层、dal层、dm层，叫法繁多。
 - App层主要是提供给数据产品和数据分析使用的数据，一般会存放在 ES、PostgreSql、Redis 等系统中供线上系统使用，也可能存在 Hive 或者 Druid中供数据分析和数据挖掘使用。比如我们经常说的报表数据，一般就放在这里。



小结

- 数仓分层的目的：把复杂问题简单化、数据结构清晰、提高数据的复用性、隔离原始数据。
- ODS 不用建模直接用源端数据存储结构。
- DWD 范式建模，保证 ODS 到 DW 信息不丢失，如果 DWD 也采用维度建模 ODS 数据一定要长期保留。
- DWS 维度建模面向需求设计，存储一些全域经常被使用到的数据。
- APP层面向实际的数据需求，以DWD或者DWS层的数据为基础，组成的各种统计报表。统计结果最终同步到RDS以供BI或应用系统查询使用。



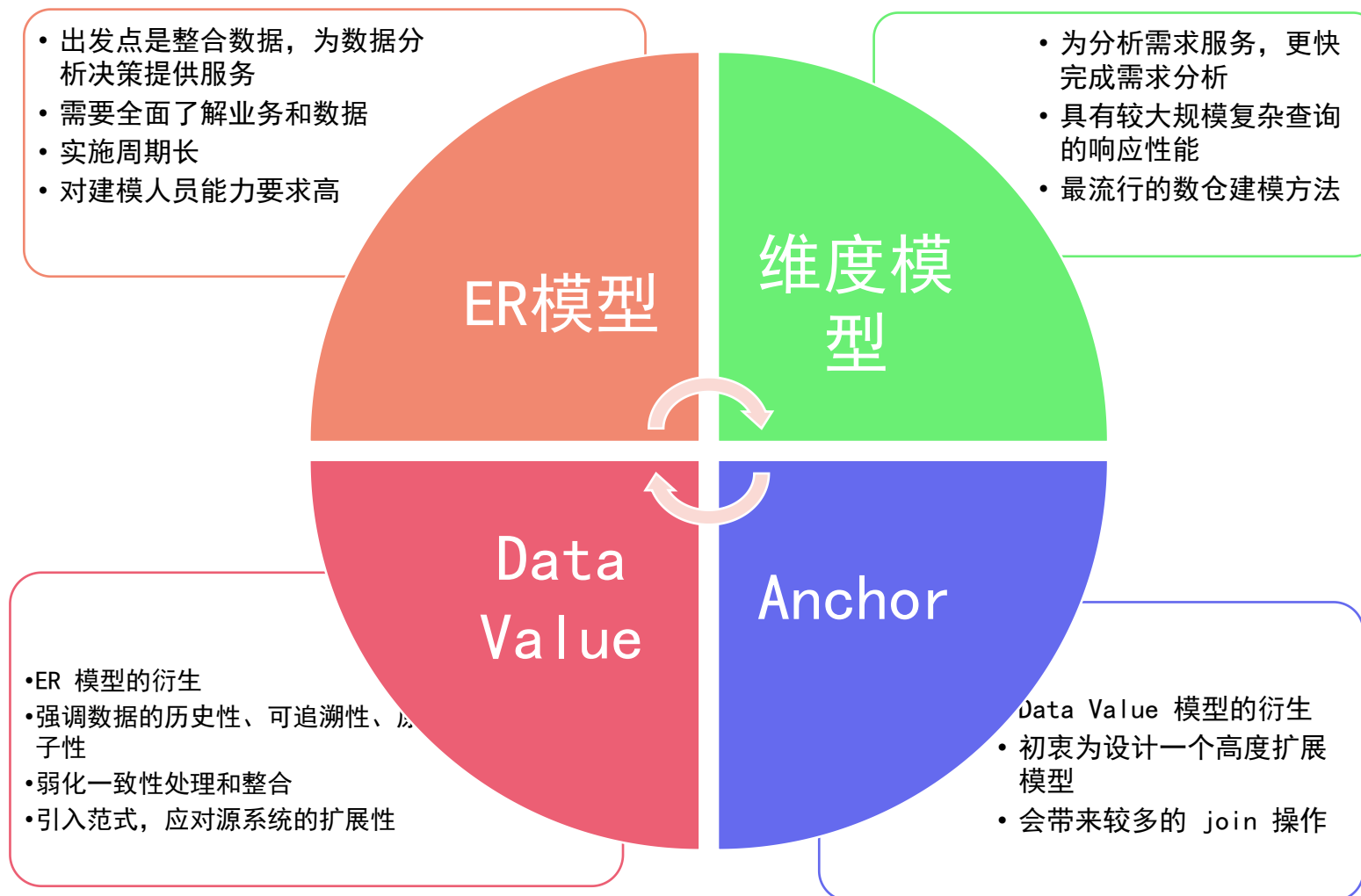


- 数仓建模的四种方法
- 关系模式范式
- ER实体模型
- 范式建模案例
- 范式建模应用场景

Part-03: 数仓建模 —— 范式建模

● 数仓建模方法：

- 数据仓库建模有如下四种：ER 模型、维度模型、Data Value、Anchor。
- 重点掌握的是 ER模型与维度模型。



- 关系模式范式

- 关系型数据库设计时，遵照一定的规范要求，目的在于降低数据的冗余性和数据的一致性，目前业界范式有：

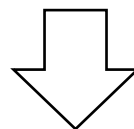
- 第一范式 (1NF)
- 第二范式 (2NF)
- 第三范式 (3NF)
- 巴斯-科德范式 (BCNF)
- 第四范式 (4NF)
- 第五范式 (5NF)



- 第一范式(1NF)

- 域都应该是原子性的，即数据库表的每一列都是不可分割的原子数据项：

部门	姓名	家庭信息
001-前端	张三	江苏-未婚
002-后端	李四	浙江-已婚
003-策划	王五	重庆-已婚

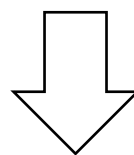


部门编号	部门名称	姓名	户籍	是否婚配
001	前端	张三	江苏	未婚
002	后端	李四	浙江	已婚
003	策划	王五	重庆	已婚

- 第二范式 (2NF)

- 在1NF的基础上，实体的属性完全依赖于主关键字，不能存在仅依赖于主关键字一部分的属性：

学号	姓名	课程号	年龄	学分
001	张三	101	18	3
002	李四	103	17	2
003	王五	102	19	3

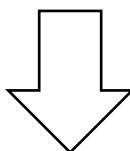


学号	姓名	年龄
001	张三	18
002	李四	17
003	王五	19

课程号	学分
101	3
103	2
102	3

- 第三范式 (3NF)
 - 在2NF的基础上，任何非主属性不依赖于其它非主属性：

学号	姓名	年龄	学院ID	学院名称	学院电话
001	张三	18	lg	理工	xxxxxx
002	李四	17	rw	人文	xxxxxxxx



学号	姓名	年龄	学院ID
001	张三	18	lg
002	李四	17	rw

学院ID	学院名称	学院电话
lg	理工	xxxxxx
rw	人文	xxxxxxxx

- 三大范式总结

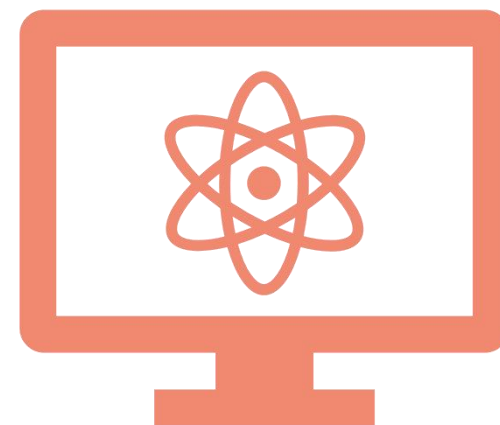
- 第一范式（1NF）：字段不可分，原子性。 字段不可再分, 否则就不是关系数据库;; ;
- 第二范式（2NF）：有主键，非主键字段依赖主键，唯一性 。一个表只说明一个事物;; ;
- 第三范式（3NF）：非主键字段不能相互依赖。每列都与主键有直接关系，不存在传递依赖。

- 不满足三大范式会带来问题：

- a、数据冗余
- b、删除异常;
- c、插入异常
- d、更新异常

● ER实体模型

- 在信息系统中，将事物抽象为“实体”、“属性”、“关系”来表示数据关联和事物描述；
- 实体：Entity，关系：Relationship，这种对数据的抽象建模通常被称为ER实体关系模型；
- 实体：通常为参与到过程中的主体，客观存在的，比如商品、仓库、货位、汽车，此实体非数据库的实体表。
- 属性：对主体的描述、修饰即为属性，比如商品的属性有商品名称、颜色、尺寸、重量、产地等。



● ER实体模型

- 关系：现实的物理事件是依附于实体的，比如商品入库事件，依附实体 商品、货位，就会有“库存”的属性产生；用户购买商品，依附实体用 户、商品，就会有“购买数量”、“金额”的属性产品。
- 实体之间建立关系时，存在对照关系：
 - ✓ 1:1、即1对1的关系，比如实体人、身份证，一个人有且仅有一个身份证号
 - ✓ 1:n、即1对多的关系，比如实体学生、班级，对于某1个学生，仅属于1个班级，而在1个班级中，可以有多个学生
 - ✓ n:m、即多对多的关系，比如实体学生、课程，每个学生可以选修多门课程，同样每个课程也可以被多门学生选修

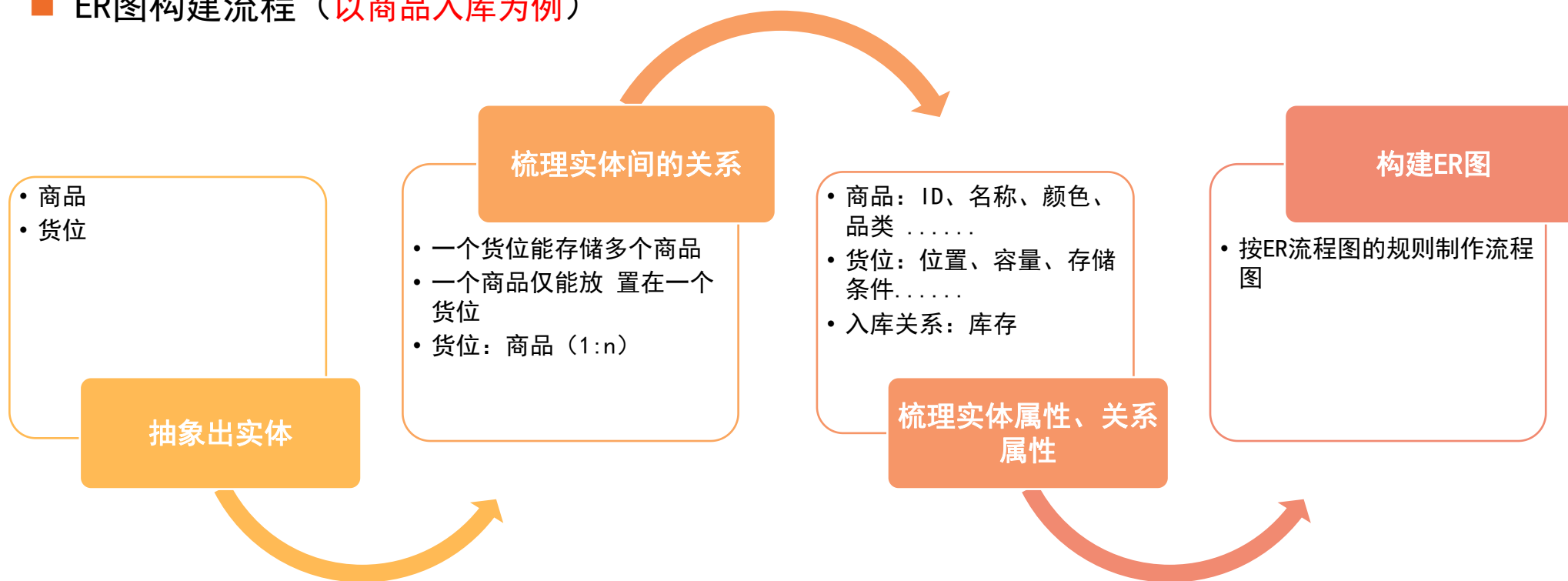
- ER实体模型

- 建模工具中，“实体”用矩形表示，“关系”用菱形表示，“属性”用椭圆形表示。
- 所以ER实体关系模型也称作E-R关系图



● ER实体模型

■ ER图构建流程（以商品入库为例）



● ER实体模型

■ 案例（课程管理系统建模）

✓ **案例背景：**该系统主要用来管理某校教师、学生、课程，其中包括课程选修、考试、教师授课、学生班级管理功能，现需要完成数据库逻辑模型设计。

✓ 建模流程

- 1，抽象出主体
- 2，梳理主体之间的关系
- 3，梳理主体的属性
- 4，画出E-R关系图



- ER实体模型

- 案例（课程管理系统建模）

- ✓ step-01、抽象出主体

- 学生
 - 课程



学生



课程

● ER实体模型

■ 案例（课程管理系统建模）

✓ step-02、梳理主体之间的关系

- 学生与课程之间的关系是“选修”关系
- 1个学生可以选多门课
- 1个课程也能被多个学生选
- 学生：课程 是“多(n)对多(n)”的关系



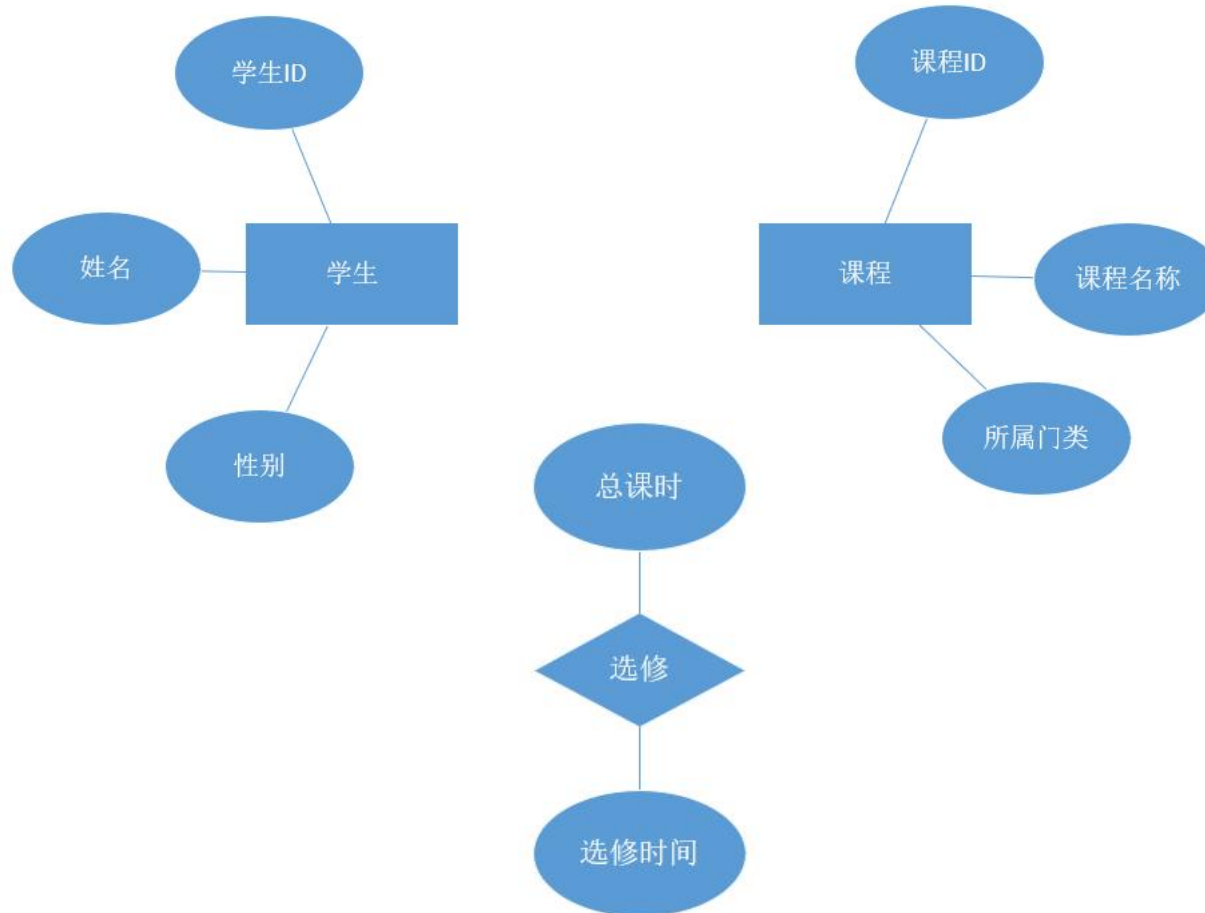
数仓建模 —— 范式建模

● ER实体模型

■ 案例（课程管理系统建模）

✓ step-03、梳理主体的属性

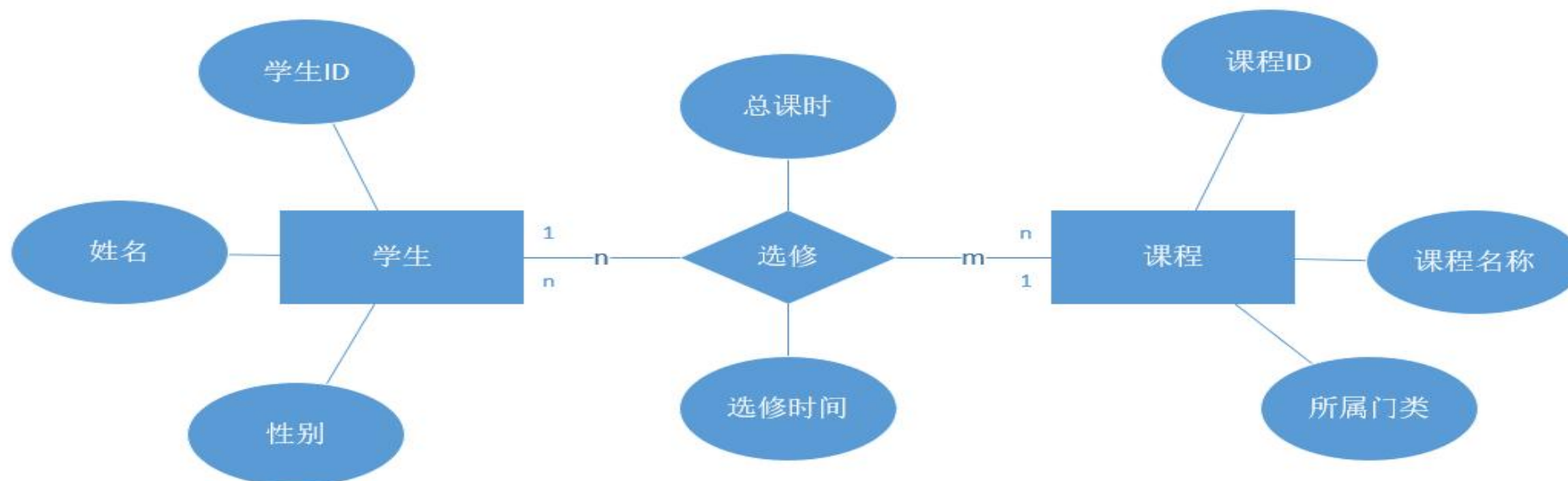
- 学生：学生ID、姓名、性别
- 课程：课程ID、课程名称、所属门类
- 选修：总课时、选修时间



● ER实体模型

■ 案例（课程管理系统建模）

✓ step-04、画出E-R关系图



● 案例（课程管理系统建模）

■ IDEF1X

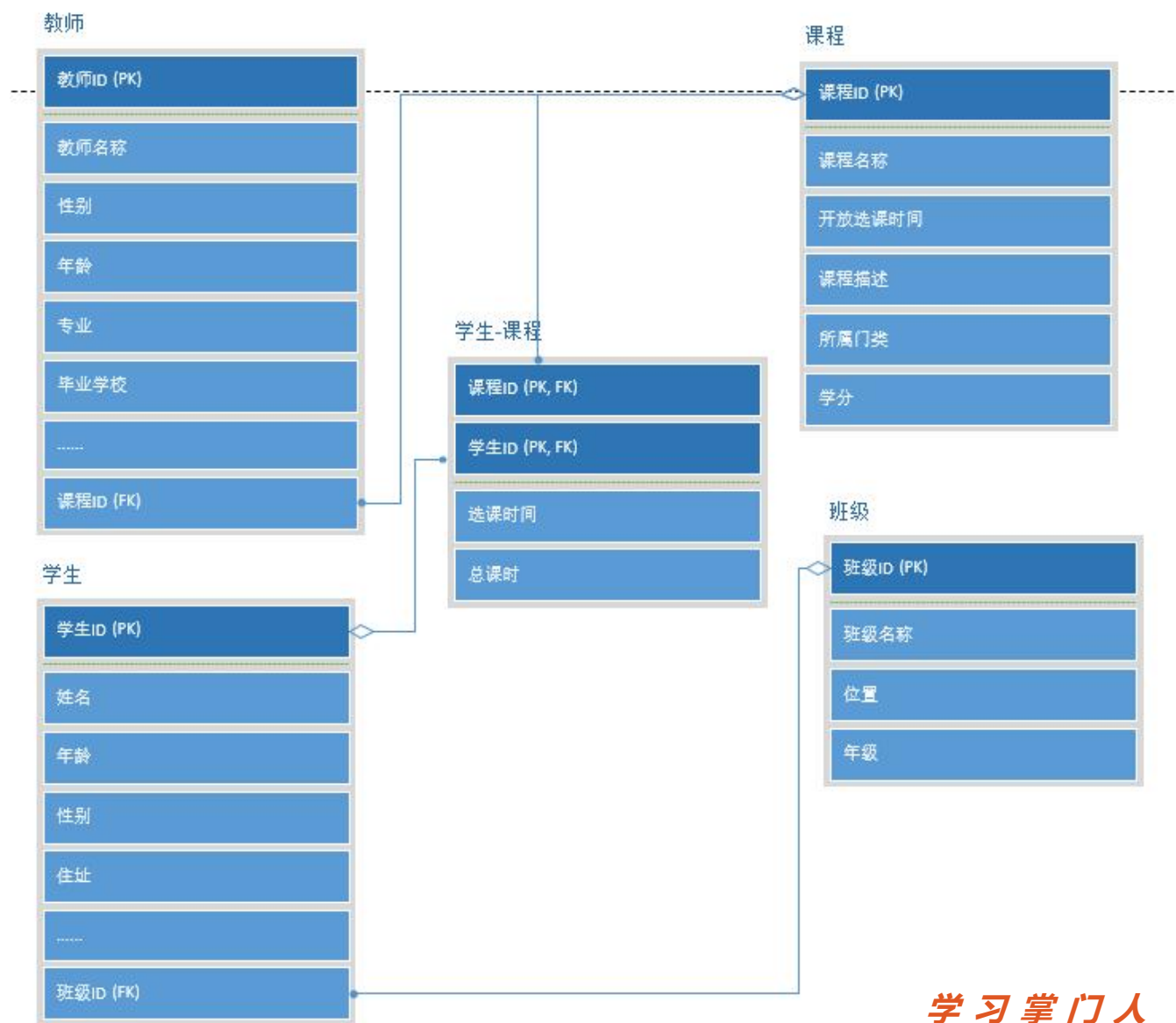
- ✓ IDEF1X 是语义数据模型化技术, 可用于概念模型设计, 有一致性、可扩展性、简洁的特点, 易于用户掌握。
- ✓ IDEF1X 图, 即 IDEF1X 是 IDEF 系列方法中 IDEF1 的扩展版本, 是在 E-R (实体联系) 法的原则基础上, 增加了一些规则, 使语义更为丰富的一种方法。用于建立系统信息模型。



数仓建模 —— 范式建模

● 案例（课程管理系统建模）

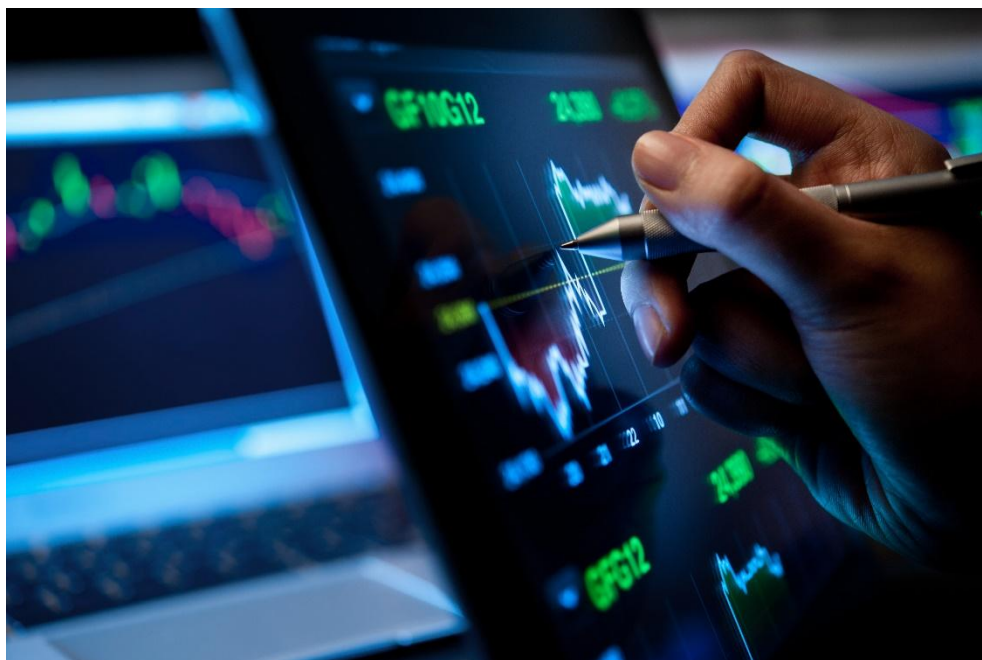
■ IDEF1X



数仓建模 —— 范式建模

● 应用场景

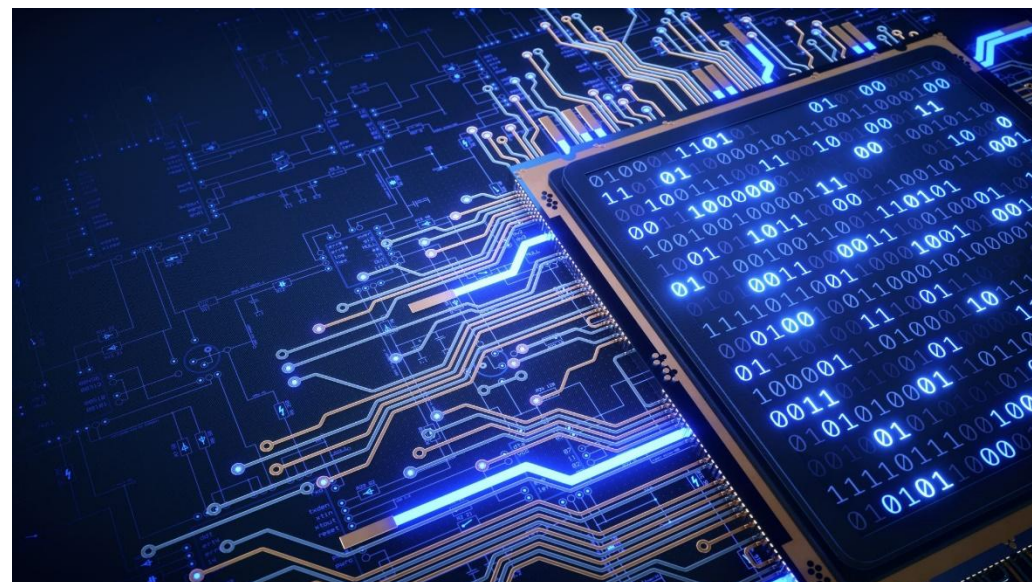
- ER模型是数据库设计的理论基础，当前几乎所有的OLTP系统 设计都采用ER模型建模的方式
- Bill Inom提出的数仓理论，推荐采用ER关系模型进行建模
- BI架构提出分层架构，数仓底层ods、dwd也多采用ER关系模型进行设计



● 动手练一练

■ 需求：某个工厂仓储管理系统，系统要求包含如下功能

- ✓ 1、仓库（仓库号、面积、联系电话）能存放多种零件（零件号、名称、规格、单价、描述）
- ✓ 2、一个仓库有职工（职工号、姓名、年龄、职称）当保管员
- ✓ 3、职工之间有领导与被领导关系
- ✓ 4、供应商（供应商号、姓名、地址、电话号码、账号）可以供应给若干个项目（项目号、预算、开工日期）多种零件



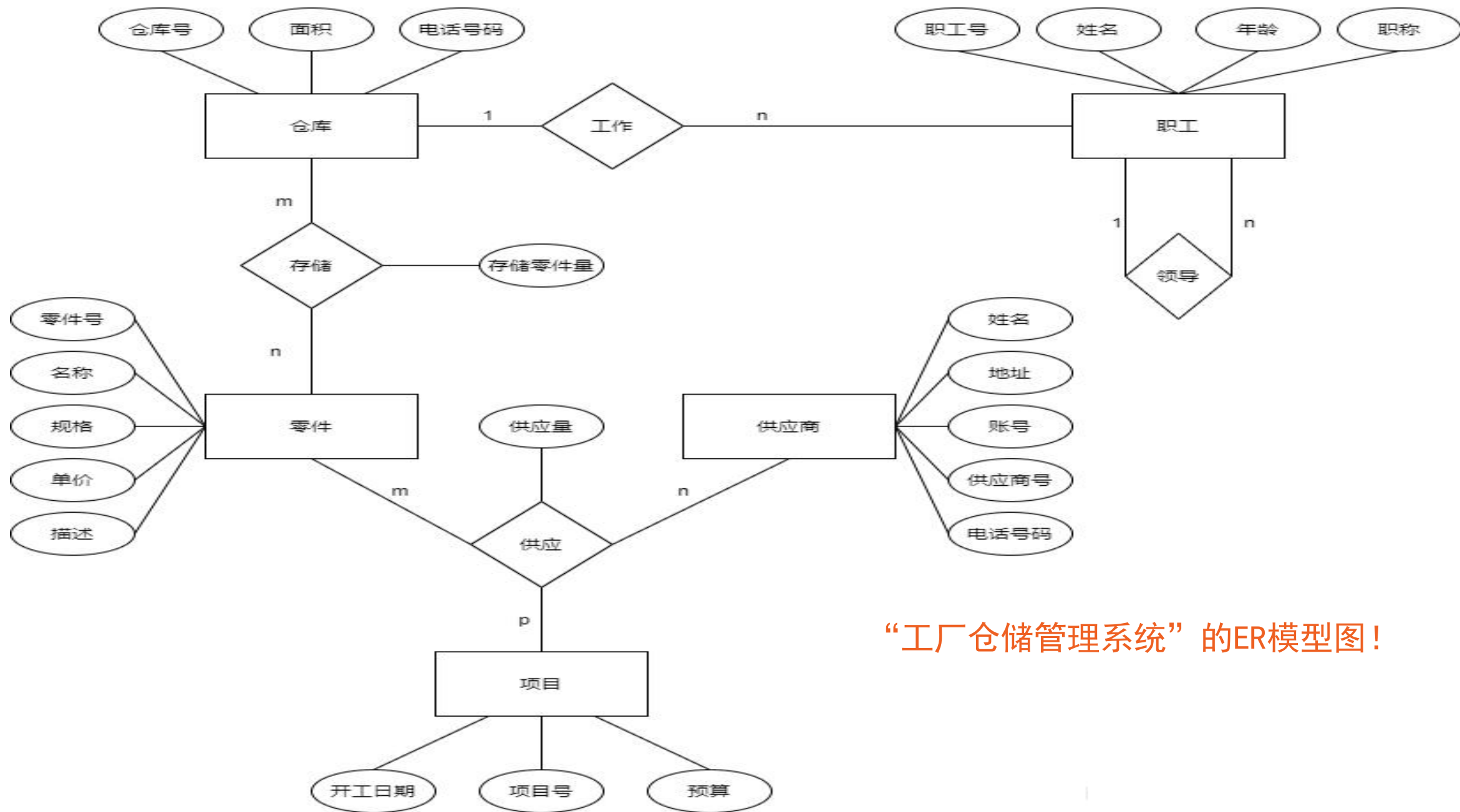
● 动手练一练

■ 它们之间的关系：

- ✓ 一个仓库可以存放多种零件，一种零件可以存放在多个仓库中，因此仓库和零件具有多对多的联系。用库存量来表示某种零件在某个仓库中的数量。
- ✓ 一个仓库有多个职工当仓库保管员，一个职工只能在一个仓库工作，因此仓库和职工之间是一对多的联系。
- ✓ 职工之间具有领导与被领导关系。即仓库主任领导若干保管员，因此职工实体型中具有一对多的联系。
- ✓ 供应商、项目和零件三者之间具有多对多的联系。即一个供应商可以供给若干项目多种零件，每个项目可以使用不同供应商供应的零件，每种零件可由不同供应商供给。

- 动手练一练

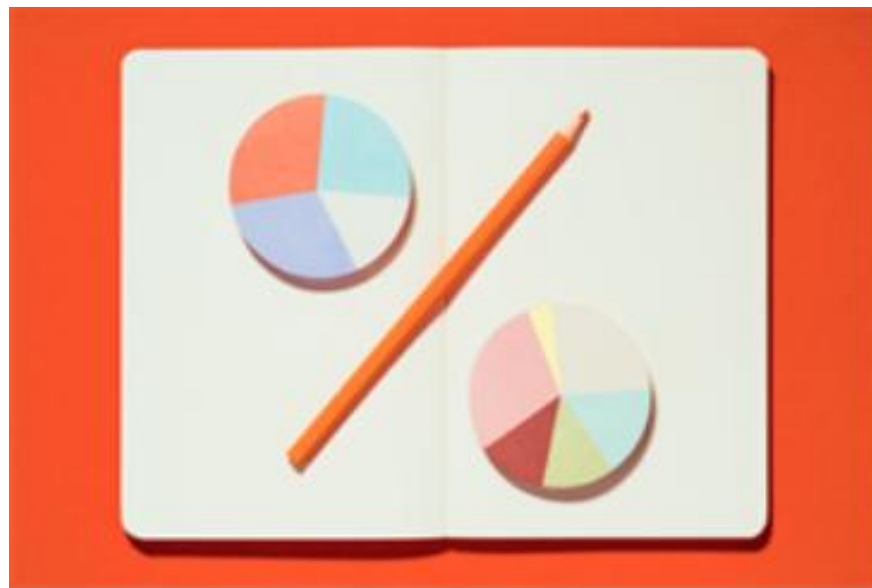
请根据如上需求设计出“工厂仓储管理系统”的ER模型图！



“工厂仓储管理系统”的ER模型图！


小结

- 数据仓库建模常见的有四种：ER 模型、维度模型、Data Value、Anchor。
- 关系型数据库设计时，遵照一定的规范要求，目的在于降低数据的冗余性和数据的一致性，目前业界范式有：五大范式，重点掌握前三范式。
- 数仓建模——范式建模的流程：
 - ✓ 1，抽象出主体
 - ✓ 2，梳理主体之间的关系
 - ✓ 3，梳理主体的属性
 - ✓ 4，画出E-R关系图
- ER模型是数据库设计的理论基础，当前几乎所有的 OLTP 系统 设计都采用ER模型建模的方式



扩展学习

- 数仓规范建设指南
 - 01-数仓公共开发规范
 - 02-数仓各层开发规范
 - 03-数仓命名规范



扩展学习
具体见附件

谢谢观看
Thanks