学掌门
atstudy.com

DESIGNED BY IBOTU

数据仓库理论基础与企业应用场景

数据仓库理论基础

学习掌门人 职场引路人

学掌门
atstudy.com

Contents
//目录

学习掌门人 职场引路人

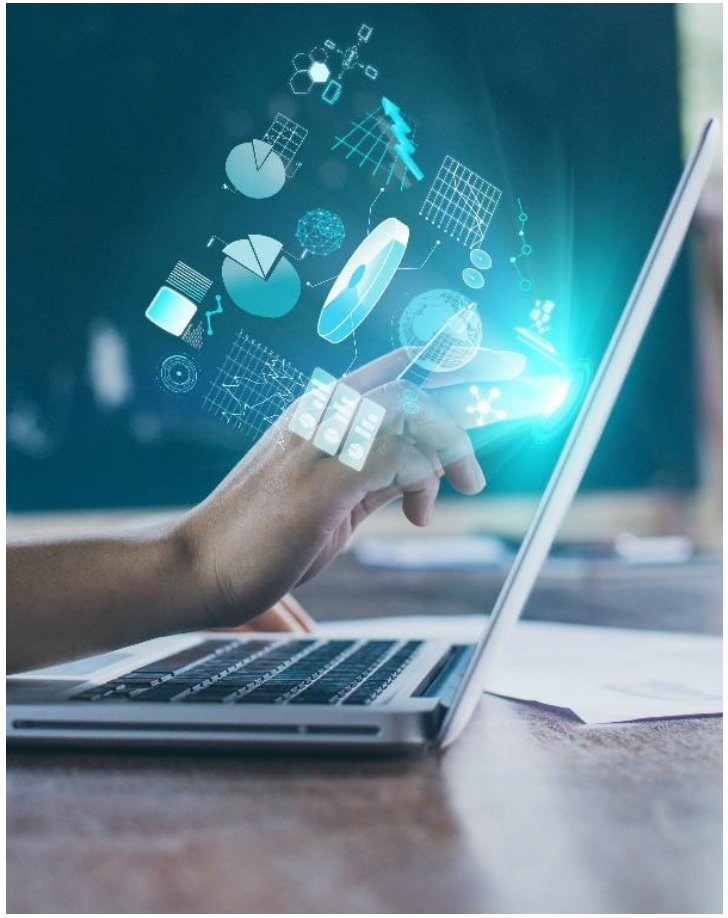
数据仓库理论基础

PART 01
数据仓库的发展历史

PART 02
数据仓库的架构

PART 03
数仓元数据及数仓术语解析

数据仓库，英文名称为Data Warehouse，可简写为DW或DWH。数据仓库，是为企业所有级别的决策制定过程，提供所有类型数据支持的战略集合。它是单个数据存储，出于分析性报告和决策支持目的而创建。 为需要业务智能的企业，提供指导业务流程改进、监视时间、成本、质量以及控制。



Part-01：数据仓库的发展历史

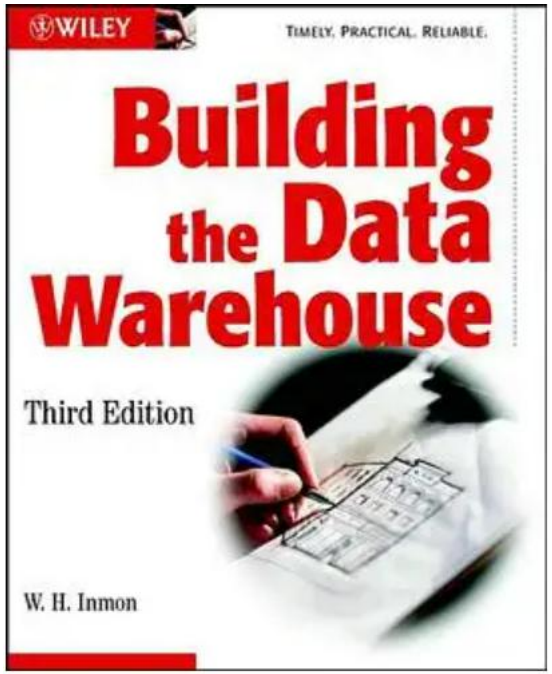
数据仓库的发展历史

- 数据仓库概念最早可追溯到20世纪70年代，希望提供一种架构将业务处理系统和分析处理分为不同的层次。
- 20世纪80年代，建立TA2(Technical Architecture2)规范，该明确定义了分析系统的四个组成部分：数据获取、数据访问、目录、用户服务。
- 1988年，IBM第一次提出信息仓库的概念：一个结构化的环境，能支持最终用户管理其全部的业务，并支持信息技术部门保证数据质量；抽象出基本组件：数据抽取、转换、有效性验证、加载、cube开发等，基本明确了数据仓库的基本原理、框架结构，以及分析系统的主要原则。

数据仓库的发展历史

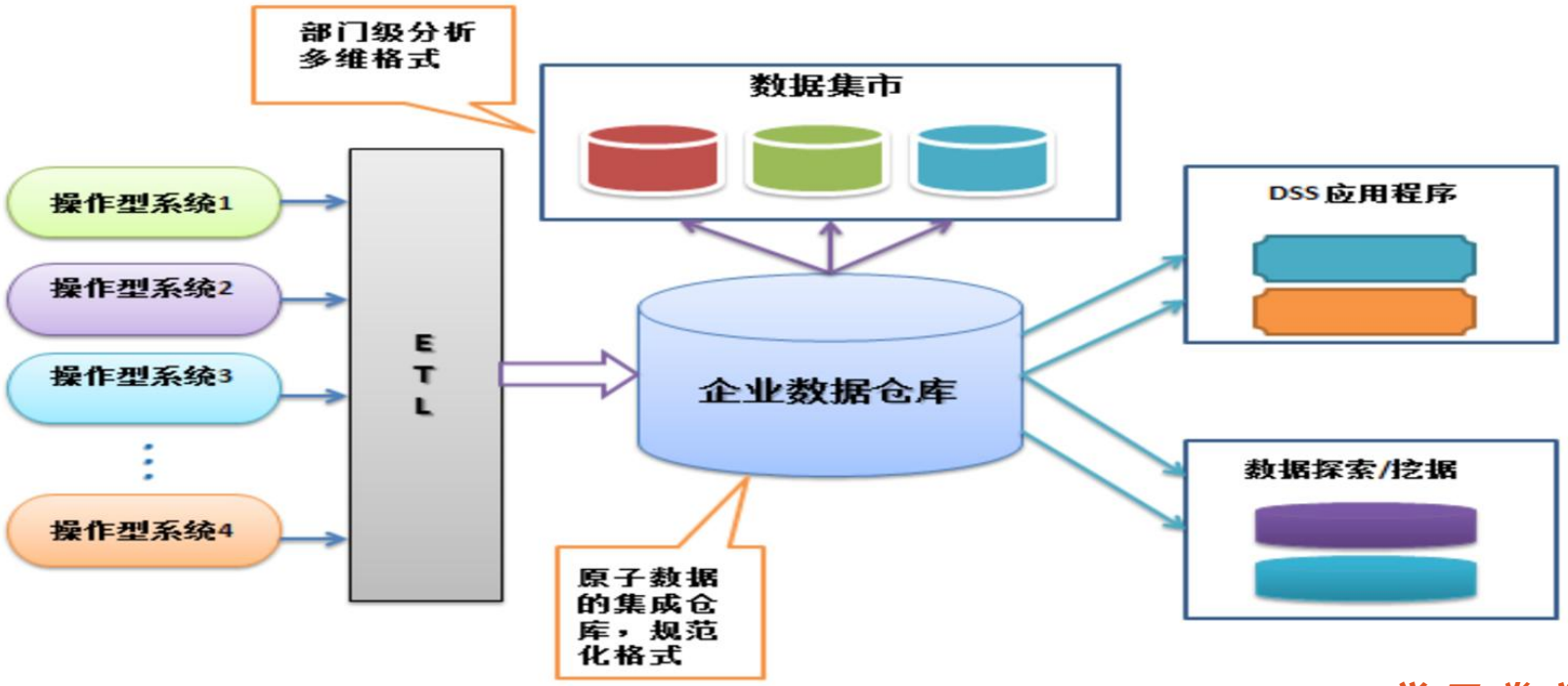
尽管有些理论目前仍有争议，但凭借此书获得“数据仓库之父”的殊荣

- 1991年，Bill Inmon出版《 Building the Data Warehouse 》提出了更具体的数据仓库原则：
  - 数据仓库是面向主题的
  - 集成的
  - 包含历史的
  - 不可更新的
  - 面向决策支持的
  - 面向全企业的
  - 最明细的数据存储
  - 数据快照式的数据获取



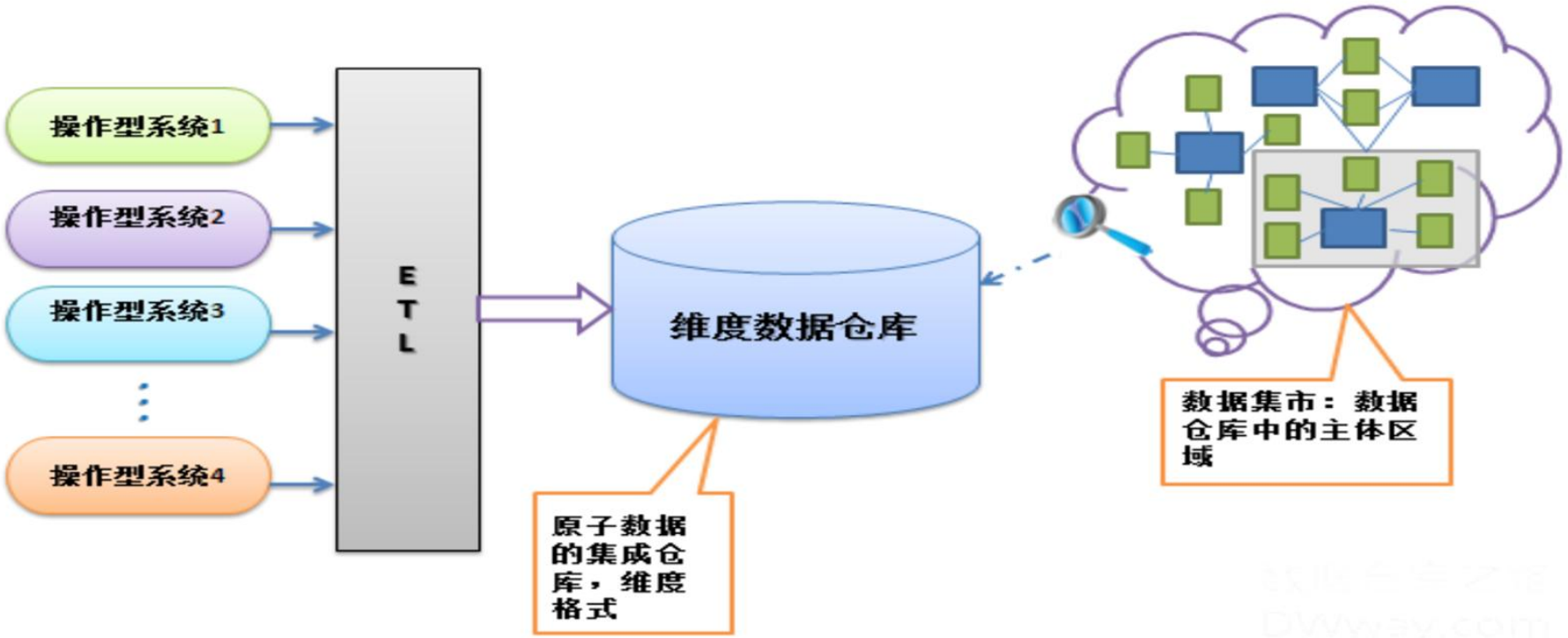
数据仓库的发展历史

- Bill Inmon主张 **自上而下** 的建设企业数据仓库，认为数据仓库是一个整体的商业智能系统的一部分。一家企业只有一个数据仓库，数据集市的信息来源出自数据仓库，在数据仓库中，信息存储符合第三范式，大致架构：



数据仓库的发展历史

- Ralph Kimball 出版《The Data Warehouse Toolkit》，其主张**自下而上**的建立数据仓库，极力推崇建立数据集市，认为数据仓库是企业内所有数据集市的集合，信息总是被存储在多维模型当中，其思路：



数据仓库的发展历史

- 比较自上而下、自下而上的优缺点

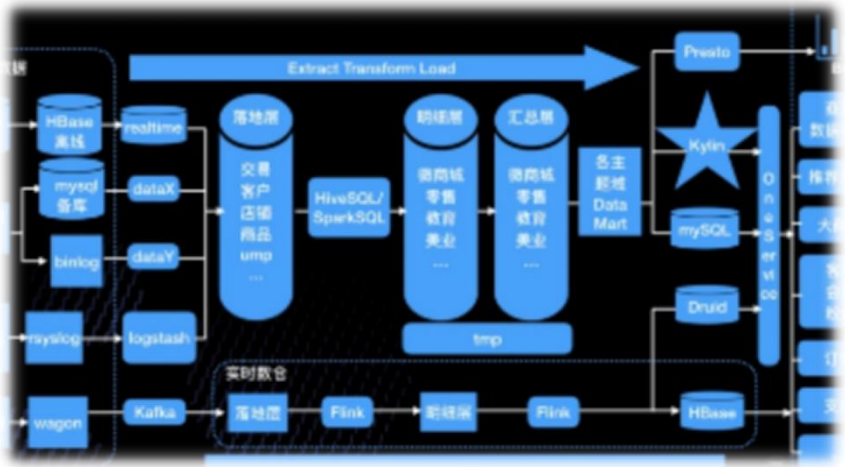
	优势	劣势
自上而下建立数据仓库（范式建模）	1. 数据的整合使得企业对数据有一个真正的企业范围级的观察； 2. 粒状的数据，可以满足不同分析角色的分析需求，支撑不同方式的分析过程； 3. 能够从整体上把握数据仓库的规模、粒度的级别和元数据管理，是一种系统的解决方法； 4. 易于维护，高度集成。	1. 部署周期较长，投入的资源比较大，导致数据化建设的效果不能立刻展现； 2. 结构死板，缺乏灵活性，因为整个组织的共同数据模型达到一致是很困难的。
自下而上建立数据仓库（维度建模）	1. 性能好，通过对各个维的预处理，能够极大地提升数据仓库的处理能力； 2. 比较直观，不需要经过特别的抽象处理即可完成维度建模，通过紧紧围绕业务模型，可直观地反映出业务模型中的业务问题； 3. 灵活性，花费低，能够得到快速的投资回报。	1. 数据孤岛现象。不能保证各个数据集市数据来源的一致性和准确性，可能会出现"孤岛"现象； 2. 数据抽取负担。当数据集市数量庞大时，抽取原始数据的负担会比较大； 3. 变更不可传递。当多个数据集市需要做一些相关或类似的变更时，需要在多个数据集市都要做一遍，会增加工作量和出错率； 4. 缺乏扩展性。当需要建立一个新的数据集市是，可能还要从头建起。

数据仓库的发展历史

- 两种方式在什么场景下会被分别选用
  - 什么情况下会选择范式建模？
    - ✓ 有公司层面的大量资源投入和部门间协调推进能力，属于公司自上而下发起的数据建设；
    - ✓ 对性能要求没那么高，可以接受查询性能的不足。
    - ✓ 有一些更高层次的要求，例如希望底层数据的一致性，无数据冗余，方便维护等。
  - 什么情况下选择维度建模？
    - ✓ 追求卓越的性能，希望能够有快速的查询和计算能力；
    - ✓ 希望数据建设项目快速上线，短期内不希望有大量的投入；
    - ✓ 能够接受后续比较高的维护成本，以及维度建模长期发展的一些劣势：数据孤岛、数据抽取负担、变更不可传递、缺乏扩展性的缺点。

小结

- 数据仓库概念最早可追溯到20世纪70年代，希望提供一种架构将业务处理系统和分析处理分为不同的层次。
- 1988年，IBM第一次提出信息仓库的概念。
- 1991年，数据仓库之父：Bill Inmon 出版《 Building the Data Warehouse 》提出了更具体的数据仓库原则。
- Ralph Kimball 出版《The Data Warehouse Toolkit》，其主张自下而上的建立数据仓库，极力推崇建立数据集市。



Part-02：数据仓库架构

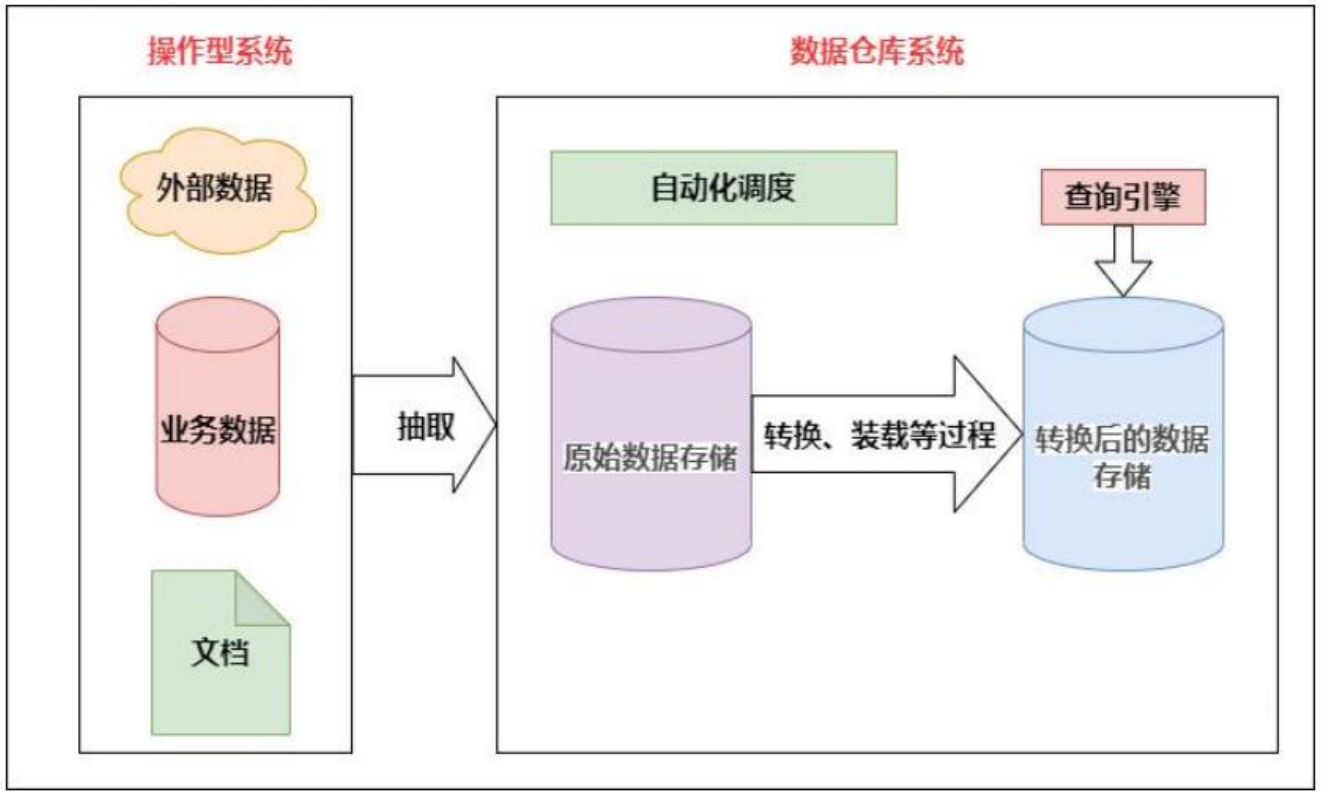
- 什么是架构？
- 数据集市架构
- Inmon 企业信息工厂架构
- Kimball 数据仓库架构
- 混合型数据仓库架构

数据仓库架构

- 什么是“架构”
  - “架构”是什 么？这个问题从来就没有一个准确的答案。这里我们引用一段话：在软件行业，一种被普遍接受的架构定义是指系统的一个或多个结构。结构中包括软件的构建 （构建是指软件的设计与实现），构建的外部可以看到属性以及它们之间的相互 关系。

数据仓库架构

- 数据仓库架构
  - 这里参考此定义，把数据仓库架构理解成构成数据仓库的组件及其之间的关系，画出下面的数仓架构图：



数据仓库架构

- 数据仓库架构
  - 在数据仓库技术演化过程中，产生了几种主要的架构方法，包括 数据集市架构、Inmon 企业信息工厂架构、Kimball 数据仓库架构、混合型数据仓库架构。

数据集市架构

Inmon 企业信息工厂架构

Kimball 数据仓库架构

混合型数据仓库

数据仓库架构

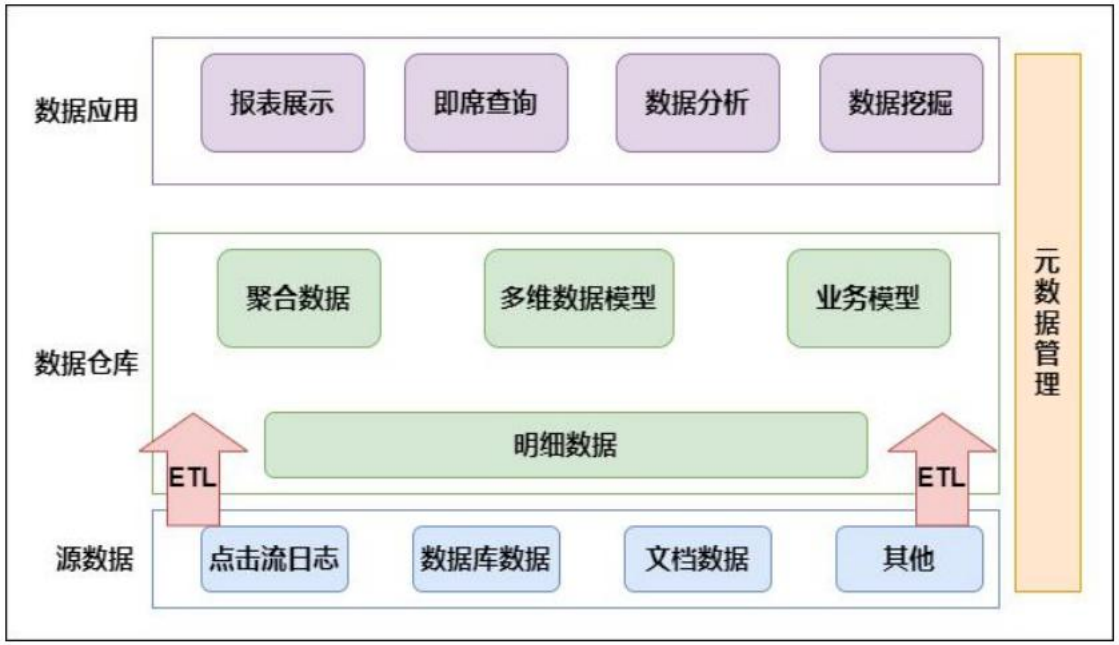
- 数据仓库分层架构
  - 再谈数据仓库与数据库的区别：
    - ✓ 数据库与数据仓库的区别实际讲的是 OLTP 与 OLAP 的区别：
      - 操作型处理，叫联机事务处理 OLTP（On-Line Transaction Processing，），也可以称面向交易的处理系统，它是针对具体业务在数据库联机的日常操作，通常对少数记录进行查询、修改。用户较为关心操作的响应时间、数据的安全性、完整性和并发支持的用户数等问题。传统的数据库系统作为数据管理的主要手段，主要用于操作型处理，像 Mysql，Oracle 等关系型数据库一般属于 OLTP。
      - 分析型处理，叫联机分析处理 OLAP（On-Line Analytical Processing）一般针对某些主题的历史数据进行分析，支持管理决策。

数据仓库架构

- 数据仓库分层架构
  - 再谈数据仓库与数据库的区别：
    - ✓ 数据库与数据仓库的区别实际讲的是 OLTP 与 OLAP 的区别：
      - 数据库设计是尽量避免冗余，一般针对某一业务应用进行设计，比如一张简单的 User 表，记录用户名、密码等简单数据即可，符合业务应用，但是不符合分析。
      - 数据仓库在设计是有意引入冗余，依照分析需求，分析维度、分析指标进行设计。 数据库是为捕获数据而设计，数据仓库是为分析数据而设计。
      - 数据仓库，是在数据库已经大量存在的情况下，为了进一步挖掘数据资源、为了决策需要而产生的，它决不是所谓的“大型数据库”。

数据仓库架构

- 数据仓库分层架构
  - 按照数据流入流出的过程，数据仓库架构可分为：源数据、数据仓库、数据应用：



- **源数据**：此层数据无任何更改，直接沿用外围系统数据结构和数据，不对外开放；
- **数据仓库**：也称为细节层，DW 层的数据应该是一致的、准确的、干净的数据。
- **数据应用**：前端应用直接读取的数据源；根据报表、专题分析需求而计算生成的 数据。

数据仓库架构

- 数据仓库分层架构
  - 为什么要进行数据仓库分层？

数据仓库架构

一、数据集市架构

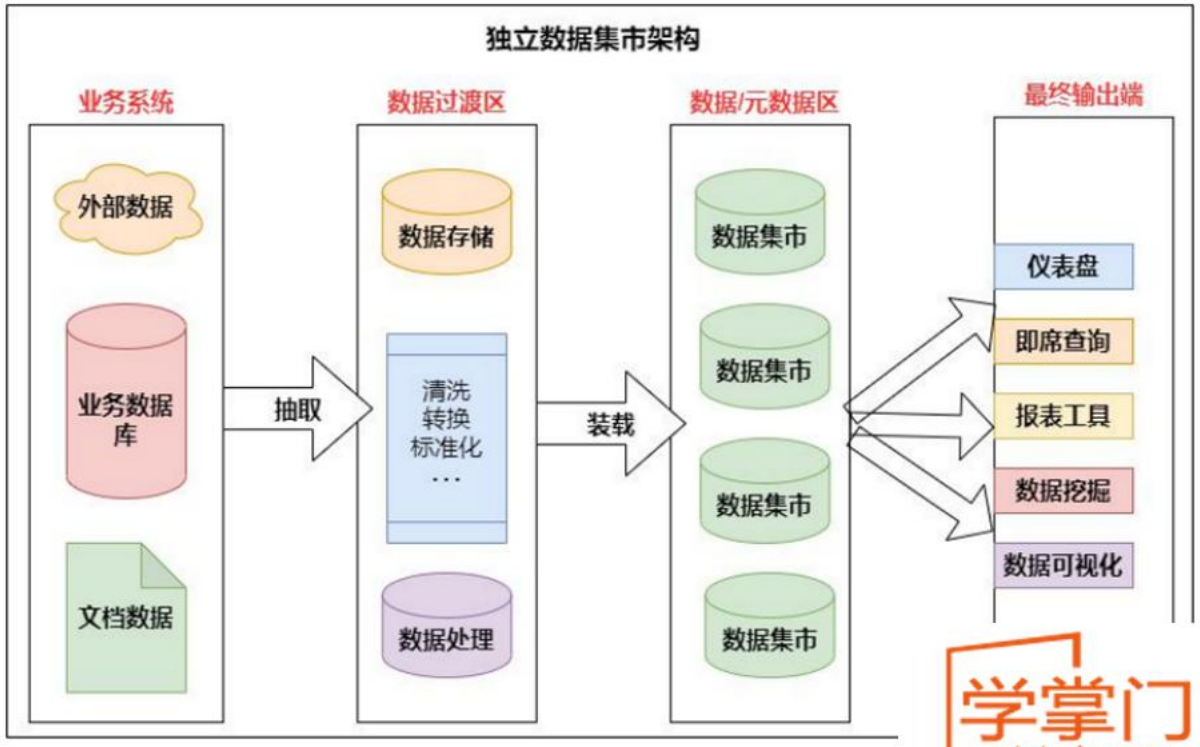
- 数据集市是按主题域组织的数据集合，用于支持部门级的决策。有两种类型的数据集市：独立数据集市和从属数据集市。

数据仓库架构

- 一、数据集市架构

  - 独立数据集市

    - ✓ 独立数据集市集中于部门所关心的单一主题域，数据以部门为基础部署，无须考虑企业级别的信息共享与集成。例如：制造部门、人力资源部门和其他部门都各自有他们自己的数据集市。



数据仓库架构

一、数据集市架构

独立数据集市

✓ 优点

- 因为一个部门的业务相对于整个企业要简单，数据量也小得多，所以部门 的独立数据集市具有周期短、见效快的特点。

✓ 缺点

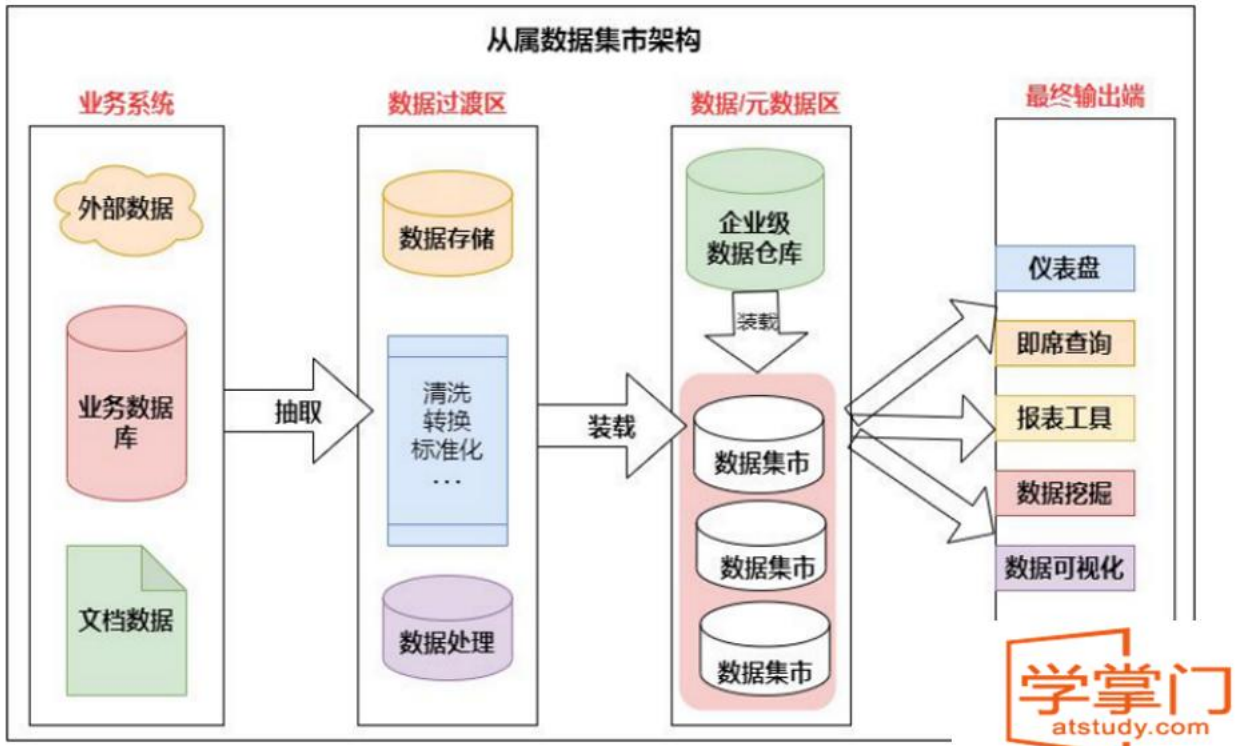
- 从业务角度看，当部门的分析需求扩展，或者需要分析跨部门或跨主题域 的数据时，独立数据市场会显得力不从心。
- 当数据存在歧义，比如同一个产品，在 A 部门和 B 部门的定义不同时，将 无法在部门间进行信息比较。
- 每个部门使用不同的技术，建立不同的 ETL 的过程，处理不同的事务系统， 而在多个独立的数据集市之间还会存在数据的交叉与重叠，甚至会有数据 不一致的情况。

数据仓库架构

- 一、数据集市架构

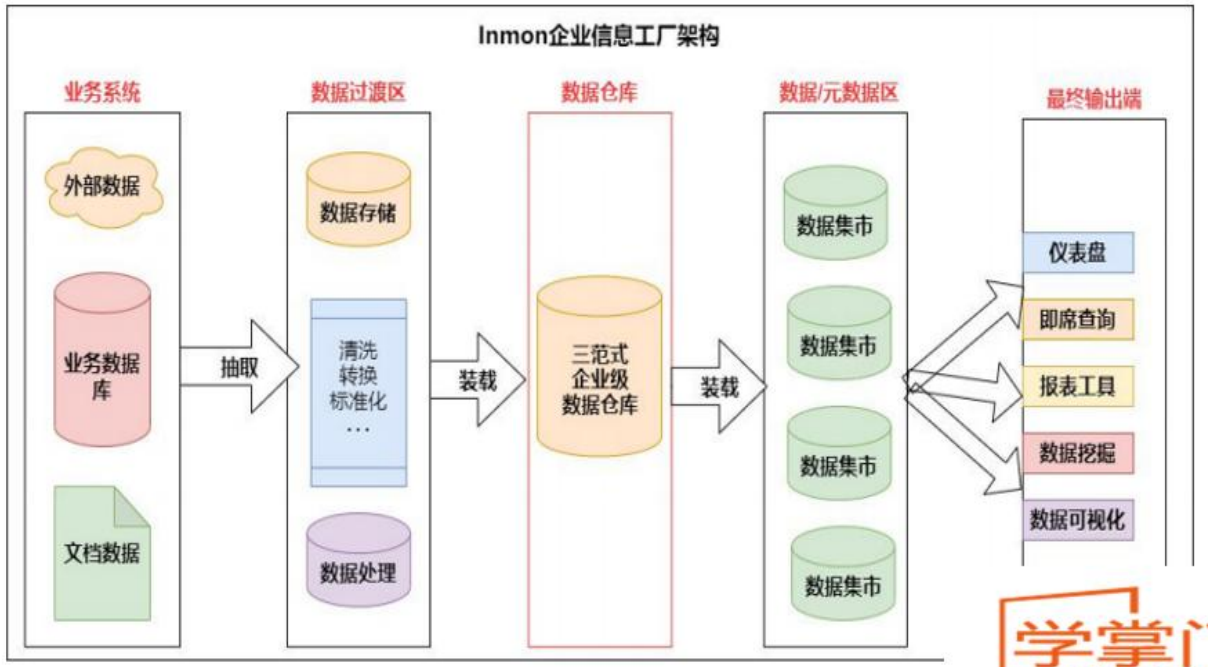
  - 从属数据集市

    - ✓ 从属数据集市的数据来源于数据仓库。数据仓库里的数据经过整合、重构、汇总 后传递给从属数据集市



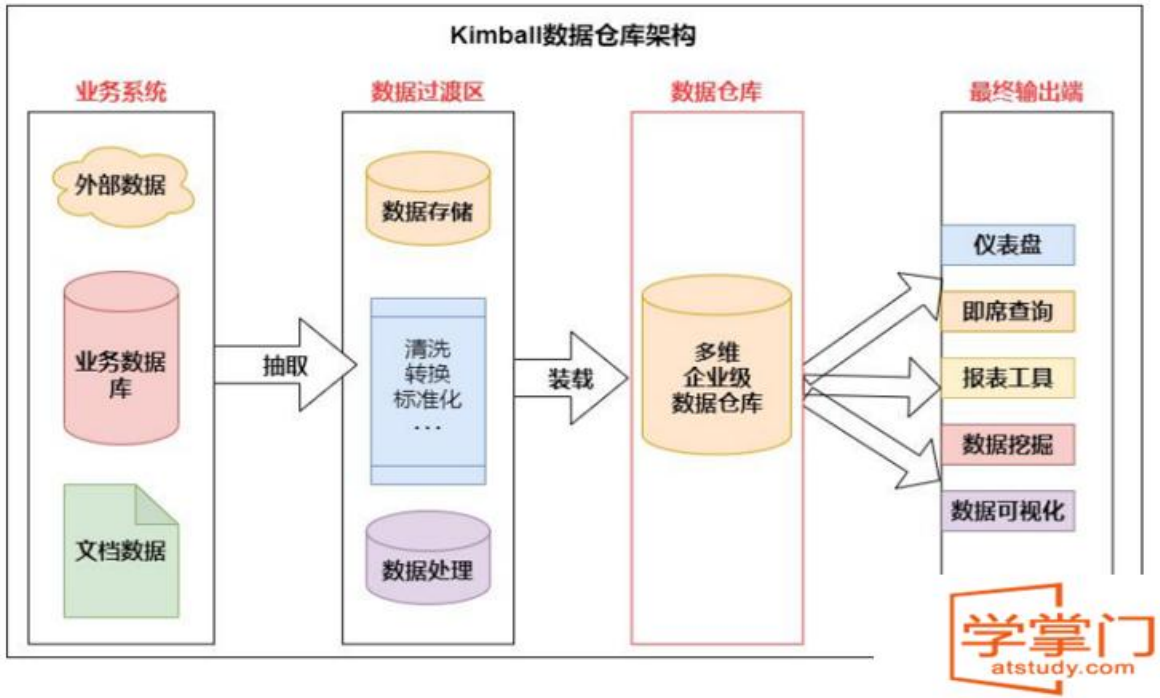
数据仓库架构

- 二、Inmon 企业工厂架构
  - 业务系统：这些应用是组织中的操作型系统，用来支撑业务。
  - 数据过渡区：数据存储、ETL清洗转换和处理
  - 数据仓库：企业级数据仓库，是该架构中的核心组件。
  - 数据/元数据区：部门级数据集市 ，是面向主题数据的部门级视图，数据从企业有数据仓库获取。
  - 最终输出端：所有的报表工具，BI工具或其他数据分析应用都从数据集市查询数据，而不是直接查询企业级数据仓库。



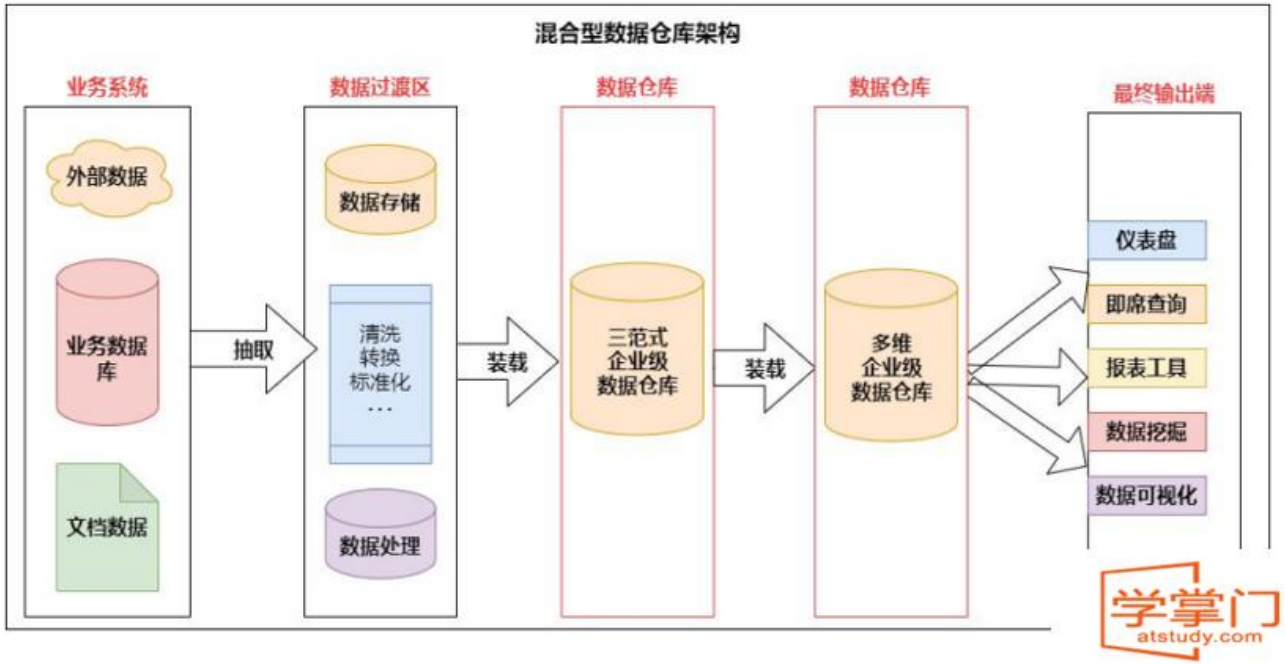
数据仓库架构

- 三、Kimball 数据仓库架构
  - Kimball 与 Inmon 两种架构的主要区别在于核心数据仓库的设计和建立。
    - ✓ Kimball 的数据仓库包含高粒度的企业数据，使用多维模型设计，这也意味着数据仓库由星型模式的维度表和事实表构成。分析系统或报表工具可以直接访问多 维数据仓库里的数据。
    - ✓ 在此架构中的数据集市也与 Inmon 中的不同。这里的数据集市是一个逻辑概念，只是多维数据仓库中的主题域划分，并没有自己的物理存储，也可以说是虚拟的数据集市



数据仓库架构

- 四、混合型数据仓库架构
  - 所谓的混合型结构，指的是在一个数据仓库环境中，联合使用 Inmon 和 Kimball 两种架构。
  - 使用这种架构的好处是：既可以利用规范化设计消除数据冗余，保证数据的粒度 足够细；又可以利用多维结构更灵活地在企业级实现报表和分析。



小结

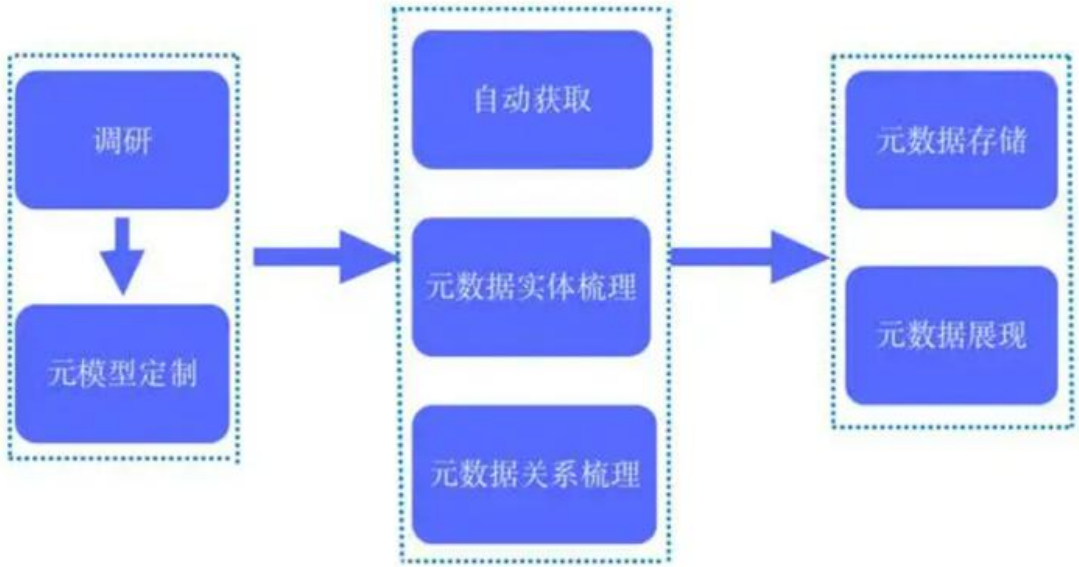
- 理解数据库与数据仓库的区别就先要理解 OLTP 与 OLAP 的区别。
- 数据仓库架构可以理解成构成数据仓库的组件及其之间的关系。
- 在数据仓库架构主要有四种：数据集市架构、Inmon 企业信息工厂架构、Kimball 数据仓库架构、混合型数据仓库架构，它们之间的关系是依次演变递进的。

- 数据仓库元数据
- 常见数仓术语解析
- 数仓名词之间的关系

Part-03：数仓元数据与常见术语解释

数仓元数据与常见术语解释

- 数据仓库元数据的管理
  - 什么是元数据？
    - ✓ 元数据（Meta Data），主要记录数据仓库中模型的定义、各层级间的映射关系、 监控数据仓库的数据状态及 ETL 的任务运行状态。
    - ✓ 一般会通过元数据资料库（Metadata Repository）来统一地存储和管理元数据，其主要目的是使数据仓库的设计、部署、操作和管理能达成协同和一致。



数仓元数据与常见术语解释

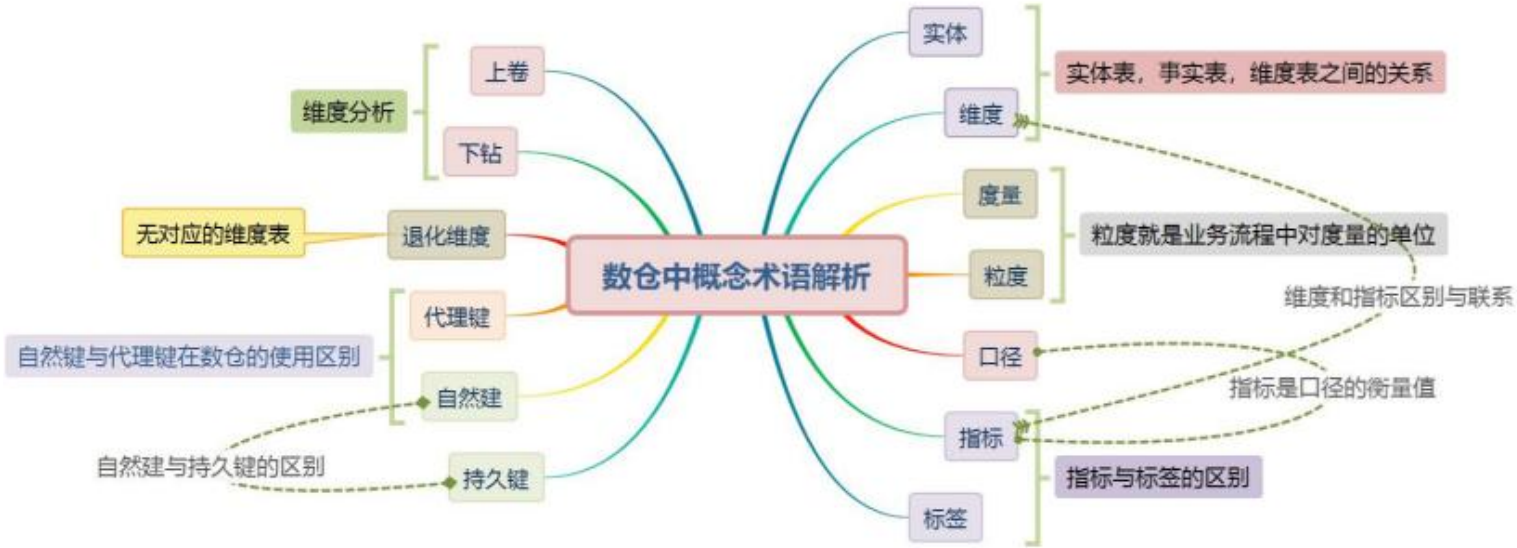
- 数据仓库元数据的管理
  - 数仓库元数据的重要性
    - ✓ 元数据是数据仓库管理系统的重要组成部分，元数据管理是企业级数据仓库中的关键组件，贯穿数据仓库构建的整个过程，直接影响着数据仓库的构建、使用和维护。
      - 构建数据仓库的主要步骤之一是 ETL。这时元数据将发挥重要的作用，它定义了源数据系统到数据仓库的映射、数据转换的规则、数据仓库的逻辑 结构、数据更新的规则、数据导入历史记录以及装载周期等相关内容。数据抽取和转换的专家以及数据仓库管理员正是通过元数据高效地构建数据仓库。
      - 用户在使用数据仓库时，通过元数据访问数据，明确数据项的含义以及定制报表
      - 数据仓库的规模及其复杂性离不开正确的元数据管理，包括增加或移除外 部数据源，改变数据清洗方法，控制出错的查询以及安排备份等。

数仓元数据与常见术语解释

- 数据仓库元数据的管理
  - 元数据的分类
    - ✓ 技术元数据
      - 技术元数据为开发和管理数据仓库的 IT 人员使用，它描述了与数据仓库开发、管理和维护相关的数据，包括数据源信息、 数据转换描述、数据仓库模型、数据清洗与更新规则、数据映射和访问权限等。
    - ✓ 业务元数据
      - 业务元数据为管理层和业务分析人员服务，从业务角度描述数据，包括商务术语、数据仓库中有什么数据、数据的位置和数据的可用性等，帮助业务人员更好 地理解数据仓库中哪些数据是可用的以及如何使用。

数仓元数据与常见术语解释

- 数仓常见术语解析



数仓元数据与常见术语解释

- 数仓常见术语解析
  - 1. 实体
    - 实体是指依附的主体，就是我们分析的一个对象。
    - 比如我们分析商品的销售情况：华为手机近半年的销售量是多少，那华为手机就是一个实体；我们分析用户的活跃度，用户就是一个实体。
    - 当然实体也可以现实中不存在的，比如虚拟的业务 对象，活动，会员等都可看做一个实体。
    - 实体的存在是为了业务分析，作为分析的一个筛选的维度，拥有描述自己的属性， 本身具有可分析的价值

数仓元数据与常见术语解释

- 数仓常见术语解析
  - 2. 维度
    - 维度就是看待问题的角度，分析业务数据，从什么角度分析，就建立什么样的维 度。
    - 所以维度就是要对数据进行分析时所用的一个量，比如你要分析产品销售情 况，你可以选择按商品类别来进行分析，这就构成一个维度，把所有商品类别集 合在一起，就构成了维度表。

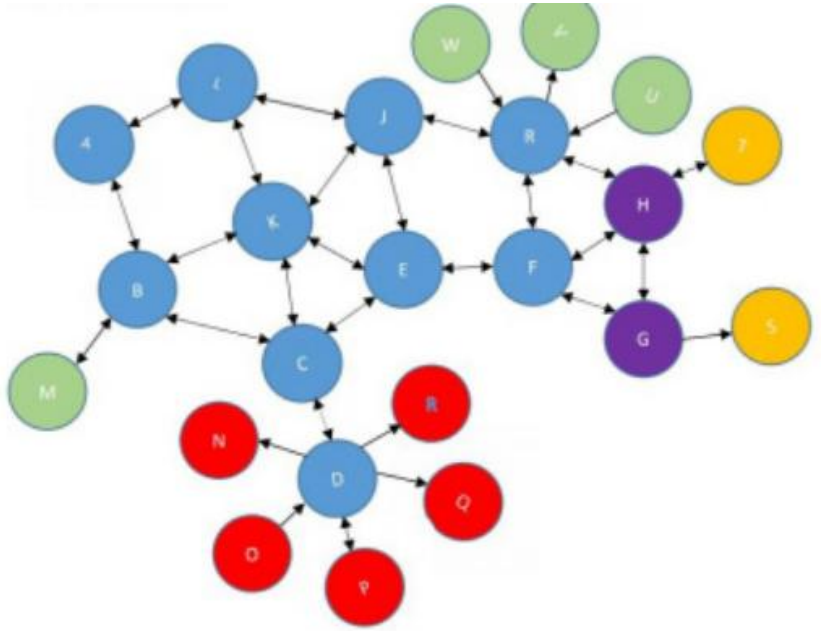
数仓元数据与常见术语解释

学掌门 atstudy.com

- 数仓常见术语解析
  - 3. 度量
    - 度量是业务流程节点上的一个数值。比如销量，价格，成本等等。事实表中的度量可分为三类：完全可加，半可加，不可加。
      - ✓ 1. 完全可加的度量是最灵活，最有用的。比如说销量，销售额等，可进行任 意维度汇总；
      - ✓ 2. 半可加的度量可以对某些维度汇总，但不能对所有维度汇总，差额是常见的半可加度量，它除了时间维度外，可以跨所有维度进行加法操作；
      - ✓ 3. 还有一种是完全不可加的。例如：比率。对于这类非可加度量，一种好的 方法是，尽可能存储非可加度量的完全可加分量，并在计算出最终的非可 加事实前，将这些分量汇总到最终的结果集中

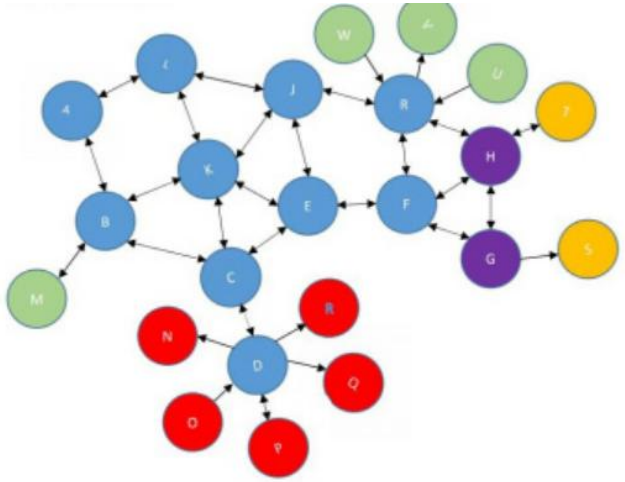
数仓元数据与常见术语解释

- 数仓常见术语解析
  - 4. 粒度
    - ✓ 粒度就是业务流程中对度量的单位，比如商品是按件记录度量，还是按批记录度量。
    - ✓ 在数仓建设中，我们说这是用户粒度的事实表，那么表中每行数据都是一个用户， 无重复用户；例如销售粒度表，那么表中每行都是一条销售记录。



数仓元数据与常见术语解释

- 数仓常见术语解析
  - 4. 粒度
    - 选择合适的粒度级别是数据仓库建设好坏的重要关键内容，在设计数据粒度时，通常需重点考虑以下因素：
      - ✓ 1. 要接受的分析类型、可接受的数据最低粒度和能存储的数据量；
      - ✓ 2. 粒度的层次定义越高，就越不能在该仓库中进行更细致的分析；
      - ✓ 3. 如果存储资源有一定的限制，就只能采用较高的数据粒度划分；
      - ✓ 4. 数据粒度划分策略一定要保证：数据的粒度确实能够满足用户的决策分析需 要，这是数据粒度划分策略中最重要的一个准则。



数仓元数据与常见术语解释

- 数仓常见术语解析
  - 5. 口径
    - 口径就是取数逻辑（如何取数的），比如要取的数是 10 岁以下儿童中男孩的平均 身高，这就是统计的口径。

数仓元数据与常见术语解释

- 数仓常见术语解析
  - 6. 指标
    - ✓ 指标是口径的衡量值，也就是最后的结果。比如最近七天的订单量，一个促销活动的购买转化率等：
      - ➢ 一个指标具体到计算实施，主要有以下几部分组成：
        - 指标加工逻辑，比如 count ，sum， avg
        - 维度，比如按部门、地域进行指标统计，对应 sql 中的 group by
        - 业务限定/修饰词，比如以不同的支付渠道来算对应的指标，微信支付的订单退款率，支付宝支付的订单退款率 。对应 sql 中的 where。

数仓元数据与常见术语解释

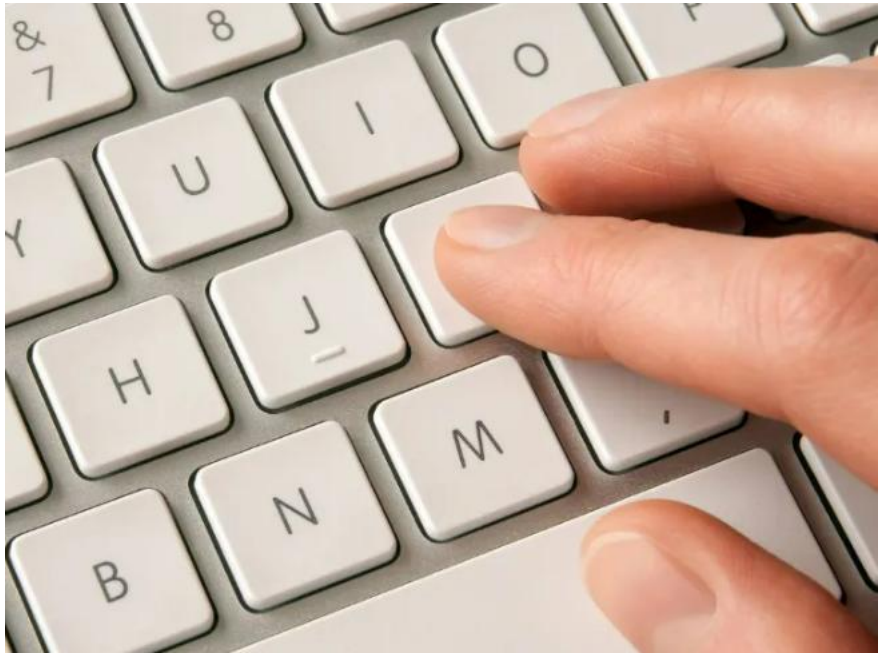
- 数仓常见术语解析
  - 6. 指标
    - ✓ 指标的分类：
      - 原子指标：基本业务事实，没有业务限定、没有维度。比如订单表中的订单量、订单总金额都算原子指标；
      - 派生指标：维度+修饰词+原子指标。 **店铺近 1 天订单支付金额** 中 **店铺** 是 维度，**近1天** 是一个时间类型的修饰词，**支付金额** 是一个原子指标；
      - 衍生指标：比如某一个**促销活动的转化率**就是衍生指标，因为需要 **促销投放人数指标** 和 **促销订单数指标** 进行计算得出。

数仓元数据与常见术语解释

- 数仓常见术语解析
  - 7. 标签
    - ✓ 标签是人为设定的、根据业务场景需求，对目标对象运用一定的算法得到的高度 精炼的特征标识。可见标签是经过人为再加工后的结果，如网红、白富美、萝莉。 对于有歧义的标签，我们内部可进行标签区分，比如：苹果，我们可以定义苹果指的是水果，苹果手机才指的是手机

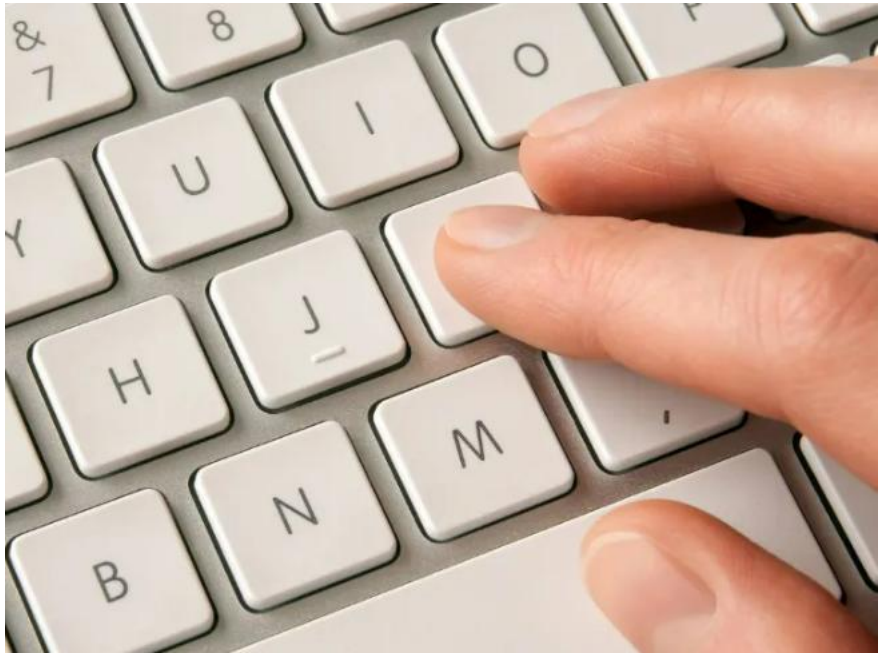
数仓元数据与常见术语解释

- 数仓常见术语解析
  - 8. 自然键
    - ✓ 由现实中已经存在的属性组成的键，它在业务概念中是唯一的，并具有一定的业 务含义，比如商品 ID，员工 ID。
    - ✓ 以数仓角度看，来自于业务系统的标识符就是自然键，比如业务库中员工的编号。



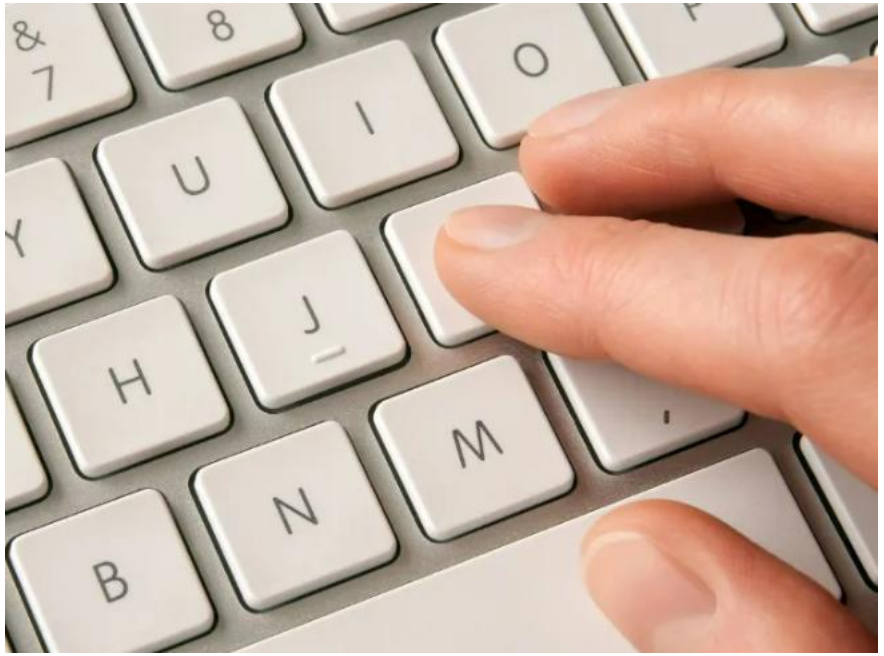
数仓元数据与常见术语解释

- 数仓常见术语解析
  - 9. 持久键
    - ✓ 保持永久性不会发生变化。有时也被叫做超自然持久键。比如身份证号属于持久键。
    - ✓ 自然键和持久键区别：举个例子就明白了，比如说公司员工离职之后又重新入职， 他的自然键也就是员工编号发生了变化，但是他的持久键身份证号是不变的。



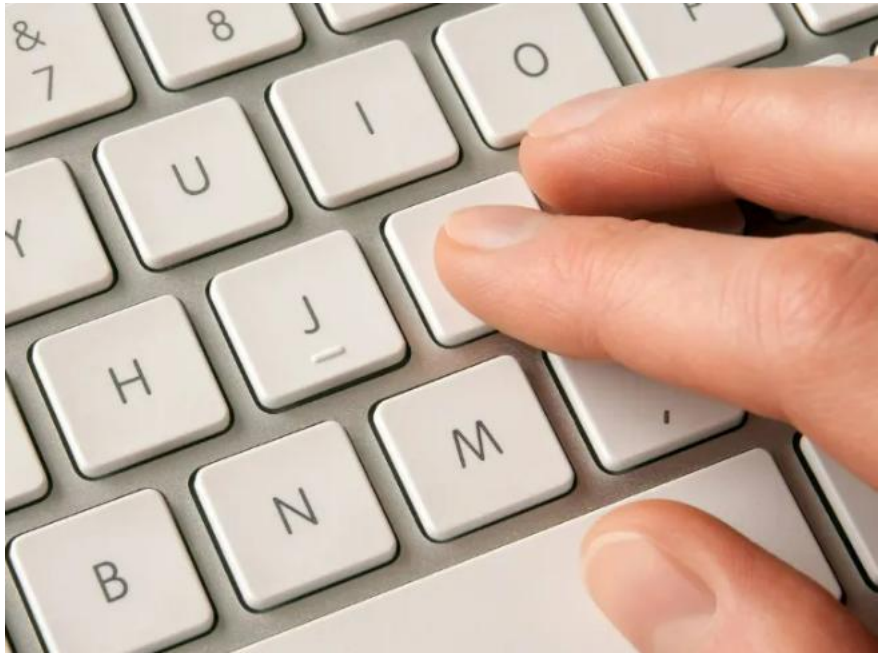
数仓元数据与常见术语解释

- 数仓常见术语解析
  - 10. 代理键
    - ✓ 就是不具有业务含义的键。代理键有许多其他的称呼：无意义键、整数键、非自然键、人工键、合成键等。
    - ✓ 代理键就是简单的以按照顺序序列生产的整数表示。产品行的第 1 行代理键为 1， 则下一行的代理键为 2，如此进行。
    - ✓ 代理键的作用仅仅是连接维度表和事实表。



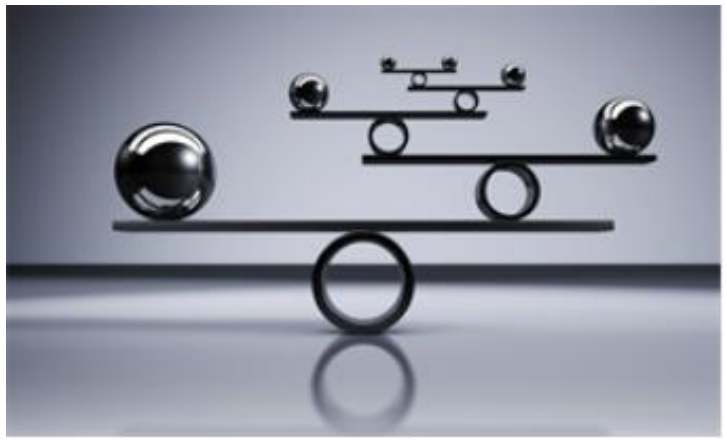
数仓元数据与常见术语解释

- 数仓常见术语解析
  - 11. 退化维度
    - ✓ 退化维度，就是那些看起来像是事实表的一个维度关键字，但实际上并没有对应的维度表，就是维度属性存储到事实表中，这种存储到事实表中的维度列被称为退化维度。
    - ✓ 与其他存储在维表中的维度一样，退化维度也可以用来进行事实表的过滤查询、实现聚合操作等



数仓元数据与常见术语解释

- 数仓常见术语解析
  - 12. 下钻
    - ✓ 这是在数据分析中常见的概念，下钻可以理解成增加维的层次，从而可以由粗粒度到细粒度来观察数据，比如对产品销售情况分析时，可以沿着时间维从年到月 到日更细粒度的观察数据。从年的维度可以下钻到月的维度、日的维度等。



数仓元数据与常见术语解释

- 数仓常见术语解析
  - 13. 上卷
    - ✓ 知道了下钻，上卷就容易理解了，它俩是相逆的操作，所以上卷可以理解为删掉维的某些层，由细粒度到粗粒度观察数据的操作或沿着维的层次向上聚合汇总数据。

数仓元数据与常见术语解释

学掌门 atstudy.com

- 数仓名词之间的关系
  - 1. 实体表，事实表，维度表之间的关系
    - ✓ 维度表：维度表可以看成是用户用来分析一个事实的窗口，它里面的数据 应该是对事实的各个方面描述，比如时间维度表，地域维度表，维度表是 事实表的一个分析角度。
    - ✓ 事实表：事实表其实就是通过各种维度和一些指标值的组合来确定一个事 实的，比如通过时间维度，地域组织维度，指标值可以去确定在某时某地 的一些指标值怎么样的事实。事实表的每一条数据都是几条维度表的数据 和指标值交汇而得到的。
    - ✓ 实体表：实体表就是一个实际对象的表，实体表放的数据一定是一条条客 观存在的事物数据，比如说各种商品，它就是客观存在的，所以可以将其 设计一个实体表。实体表只描述各个事物，并不存在具体的事实，所以也 有人称实体表是无事实的事实表。

数仓元数据与常见术语解释

- 数仓名词之间的关系
  - 2. 维度和指标区别与联系
    - ✓ 维度就是数据的观察角度，即从哪个角度去分析问题，看待问题。
    - ✓ 指标就是从维度的基础上去衡算这个结果的值。
    - ✓ 维度一般是一个离散的值，比如时间维度上每一个独立的日期或地域，因此统计 时，可以把维度相同记录的聚合在一起，应用聚合函数做累加、均值、最大值、 最小值等聚合计算。
    - ✓ 指标就是聚合运算的结果，一般是一个连续的值

数仓元数据与常见术语解释

- 数仓名词之间的关系
  - 3. 自然键与代理键在数仓的使用区别
    - ✓ 维度表的唯一主键应该是代理键而不应该是自然键。
    - ✓ 数据仓库中维度表与事实表的每个连接应该基于无实际含义的整数代理键。
    - ✓ 避免使用自然键作为维度表的主键。

数仓元数据与常见术语解释

- 数据集市和数据仓库的关系
  - 4. 数据集市和数据仓库的关系
    - ✓ 关系：数据集市是企业级数据仓库的一个子集，他主要面向部门级业务，并且只面向某 个特定的主题。数据集市可以在一定程度上缓解访问数据仓库的瓶颈。
    - ✓ 区别：数据仓库是企业级的，能为整个企业各个部门 的运行提供决策支持手段；数据集市是部门级的，一般只能为某 个局部范围内的管理人员服务，因此也称之为部门级数据仓库。

小结

- 元数据是数据仓库管理系统的重要组成部分，元数据管理是企业级数据仓库中的关键组件，贯穿数据仓库构建的整个过程，直接影响着数据仓库的构建、使用和维护。
- 数仓常见术语
  - 实体、维度
  - 度量、粒度
  - 口径、指标
  - 标签、自然键
  - 持久键、代理键
  - 退化维度、下钻、上卷
- 数仓常见术语解析
  - 实体表，事实表，维度表之间的关系
  - 维度和指标区别与联系
  - 自然键与代理键在数仓的使用区别
  - 数据集市和数据仓库的关系

学掌门
atstudy.com

谢谢观看
Thanks

学习掌门人 职场引路人