数据仓库

- 一、数据仓库定义及特性
  - 🔴数据仓库的提出，主要是为了解决多重数据复制带来的高成本，无论是性能上的，还是硬件上，数据仓库之父Bill Inmon
  - 1、定义：是一个面向主题的、集成的、反映历史变化、相对稳定的数据集合，公司处理数据都在数据仓库框架中进行处理
  - 2、特性
    - （1）面向主题：数据仓库侧重于数据分析工作，是按照主题存储的
    - （2）集成的：将多个分散的数据源统一成一致的、无歧义的数据格式，再放置到数据仓库中
    - （3）相对稳定的：数据仓库中的数据是不可修改的
    - （4）反映历史变化的：数据仓库中不可更改，索引保留了历史数据，通过历史数据，可对企业的发展历程和未来趋势做出定量分析和预测
  - 🔴数据仓库是企业的核心，是一种过程/架构，由很多数据库组成，数据仓库系统包含数据源、数据存储与管理、数据访问，数据仓库分布在企业内部各处的业务数据整合、加工和分析的过程
- 二、数据仓库架构
  - 1、典型的企业数据仓库系统，包含数据源、数据存储与管理、数据的访问三个部分
  - 2、数据仓库的架构
    - （1）数据源：整个数据仓库系统的数据源泉，来自企业内部数据和外部数据，包括了生产运营的数据、办公数据等内部数据、调查数据、市场信息等外来数据等
    - （2）数据的存储与管理
      - a.元数据的存储：关于数据的数据，主要包括数据仓库的数据字典、数据的定义、数据的抽取规则、数据的转换规则、数据的加载频率等信息
      - b.数据的存储：各操作数据库中的数据按照元数据库中的定义规则，经过抽取、清理、转换、集成，按照主题重新组织，依照相应的存储结构进行存储。
      - c.数据集市：数据仓库的一个子集（部门级数据仓库），含有较少主题且历史时间更短数据量更少，一般只能为一个局部范围的管理人员服务
    - （3）数据的访问：由OLAP(联机分析处理)、数据挖掘、统计报表、即席查询等几部分组成
    - 🔴OLAP（联机分析处理）
      - 1、定义：针对特定的分析主题，设计多种可能的观察形式、设计相应的分析主题结构（即进行事实表和维度表的设计），使管理决策人员在多维数据模型的基础上进行快速、稳定和交互性的访问，并进行各种复杂的分析和预测工作
      - 2、OLAP的分类（按存储方式来分）

- 1. MOLAP (Multi-Dimension OLAP):将OLAP分析所需的数据存放在多维数据库中。分析主题的数据可以形成一个或多个多维立方体
- 2. ROLAP（Relational OLAP) :将OLAP分析所需的数据存放在关系型数据库中。分析主题的数据以"事实表-维表"的星型模式组织。

- 3、数据的存储与管理架构
  - （1）DB（数据来源，各个系统的源数据），可以为mysql、SQLserver、文件日志等，为数据仓库提供数据来源的一般存在于现有的业务系统之中；
  - （2）ETL（Extract-Transform-Load ）：将数据从来源迁移到目标的过程
    - 1. Extract ，数据抽取，把数据从数据源读出来
    - 2. Transform，数据转换，把原始数据转换成期望的格式和维度，使之转变成适用于查询、分析的形式
    - 3. Load，数据加载，把转换后的数据加载到目标处，比如数据仓库
    - 抽取方式：全量抽取和增量抽取
    - ETL作用：将凌乱、分散、标准不一的数据整合到一起，把异构换成同构，没有ETL，将不能对异构数据进行程序化分析
  - （3）ODS（Operational Data Store）
    - 操作性数据，是数据库到数据仓库的缓存层，其数据结构与数据来源保持一致，便于减少ETL的工作复杂性，周期短，其数据最终流入DW
    - 特点：存储明细数据、操作的可变性、当期数据（一般3-6个月）
  - （4）DW（Data Warehouse）数据仓库，保持所有从ODS的数据，并长期保存，而且不被修改
  - （5）DM（Data Mart）数据集市，为了特定应用目的或应用范围，从数据仓库中独立出来的一部分数据，称为部门数据或主题数据，主要面向应用

- 三、数据仓库处理数据流程
  - 1、典型的企业数据仓库系统数据处理流程
    - 第一步：用ETL工具从源系统把数据抽取、转换、加载到ODS层，即数据过度层；
    - 第二步：从ODS层把数据按主题分类导入数据仓库的DW层
    - 第三步：从DW层把明细数据汇总到各DM，即数据集市
    - 第四步：报表工具、BI工具或其他数据分析应用从数据集市调用或查询数据
  - 2、ETL
    - （1）ETL目标：是将企业中分散的、零乱的、标准不统一的数据整合到一起，把异构数据转换成同构的,如果没有ETL，不可能对异构的数据进行程序化的分析。
    - （2）抽取、转换、装载的英文缩写
      - a. 抽取(Extract)：从操作行数据源获取数据，两种抽取方式：增量抽取、全量抽取
      - b. 转换(transform)：转换数据，使数据转变成适用与查询和分析的形式和结构
      - c. 装载 (load)： 将转换后的数据导入到最后的目标数据仓库

- （3）常见的ETL工具：Kettle、datastage、imformatic、ssis

- 3、ODS层（oparational data store操作数据存储）
  - （1）定义：(oparational data store)一个面向主题的、集成的、可变的、当前的细节数据集合，用于支持企业对于即时性的、操作性的、集成的全体信息需求，是数据缓存层，又称操作数据存储
  - （2）ODS的作用
    - a. 充当业务系统与数据仓库之间的过渡区：其数据结构、数据粒度、数据之间的逻辑关系与业务系统源数据一致，抽取逻辑简单，不需要数据转换，降低数据转换的复杂性，最小化对业务系统的侵入
    - b. 转移部分业务系统细节的查询功能：某些由业务系统产生的报表、细节数据的查询可以在ODS层进行，降低业务系统查询压力
    - c. 数据是可变的、当期数据，可以做更改，能确保数据的准确性
  - （3） ODS层数据存储时间一般为3-6个月，然后会把数据打包压缩起来
- 4、DW层（数据仓库DATE WARE HOUSE ）
  - （1）定义：按主题存储从ODS抽取过来的最细粒度事实表明细数据，面向主题的、集成的、反应历史变化的、稳定的数据集合
  - （2）特征
    - a. 面向主题的，数据仓库中的数据是按照一定主题进行存储，每一个主题对应一个宏观分析领域
    - b.稳定的，DW的数据只允许增加，不允许删除和修改，数据仓库主要是提供查询服务，删除和修改在分布式系统
    - c. 数据质量高，企业所有系统只能从DW取数据，所以要定期对DW里面的数据进行质量审查，保证数据的唯一性、权威性、准确性
    - d. 效率足够高，要对进入的数据快速处理。
- 5、DM（Data Market）数据集市
  - （1）定义：是为满足企业特定部门的分析需求而专门建立的数据集合，数据来源是DW层，数据在进入部门数据集市时可能进行聚合汇总
  - (2)数据集市使用多维模型设计，用于数据分析
  - （3）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
  - （4）特征
    - a. DM结构清晰，针对性抢，拓展性好，因为DM仅仅是单对一个领域而建立的，容易维护修改
    - b. DM 建设任务繁重，因为公司业务众多，每个业务单独建立表
    - c. DM的建设消耗更多的存储空间，因为DM的数量众多，数据量也会增加多倍
- 6、ODS（操作型数据存储）和DW（数据仓库）的区别

- （1）DW 是面向主题的、集成的、相对稳定的、反映历史变化的数据，用于支持管理决策，时效是 T + 1；
- （2）ODS 是面向主题的、集成的、可变的、当前的细节数据集合，用于支持企业对即时性、操作性、集成的全体信息的需求，时效是 实时的
- （3） ODS是数据库体系结构的一个可选部分，是DB与DW之间的中间层，ODS具备数据仓库的部分特征和OLTP系统的部分特征。
- 🔴抓住重点：DW是反映历史变化，ODS是反映当前变化

四、数据建模相关概念

1、粒度和维度

- （1）粒度：数据仓库中数据的细化程度级别，越细，粒度越低，越粗，粒度越高
- （2）维：人们观察数据的特定角度，是考虑问题时的一类属性，属性集合构成一个维

2、维度表、事实表

- （1）维度表： 存储维度信息的表
- （2）事实表：在维度建模的数据仓库中，事实表保存了大量业务度量数据，表中的度量值称为事实，事实表中最有用的事实是数据类型的事实和可加类型的事实，事实表的粒度决定了数据仓库中数据的详细程度
- 🔴总结
  - 1、事实表就是你要关注的内容；
  - 2、维度表就是你观察该事务的角度，是从哪个角度去观察这个内容的。
  - 3、上线时，都是先跑维度表，再跑事实表

3、关系型数据系统范式

- 🔴为了规范化关系型数据模型，关系型数据库系统在设计时必须遵循一定的规则，这种规则称为关系型数据库系统范式
- 🔴常用的范式：三范式
  - 🔴（1）第一范式（1NF）：字段必须具有单一属性特性，不可再拆分
    - 如果字段中的值已经是无法再分割的值，则符合第一范式。
    - 例如，在员工表 中，姓名字段一般仅包含员工的正式姓名，这是符合第一范式的，但是如果要在姓名字段 中包含中文名、英文名、昵称、别名等信息，就意味着姓名字段是可再拆分的，此时就不符合第一范式。
  - 🔴（2）第二范式（2NF）：表要具有唯一的主键列，且要求其他字段都依赖于主键
    - 第二范式(2NF)要求数据库表中的每个实例或行必须可以被唯一地区分，为实现区分通常需要为表加上一个列，以存储各个实例的唯一标识
    - 第二范式是在第一范式的基础上的进一步增强，在数据库设计时一般使用唯一性主键来唯一地标识行，且要求其他字段都依赖于主键。

- 比如在员工表中 定义了以工号作为主键，因为公司员工的工号通常用来识别某个员工个体，不能进行重复； 在部门表中通过部门编号作为主键，来唯一地区分一个部门。

- （3）第三范式（3NF）：表中的字段不能包含在其他表中已出现的非主键字段
  - 第三范式(3NF)是在前两个范式的基础上的进一步增强，主要用来降低数据的冗余
  - 比如，员工表中包含了部门编号，它引用到部门表中的部门编号这个主键，符合第三范式。 如果在员工表中又包含一个部门名称，那么表中的字段就包含了其他表中已出现的非主键 字段，造成了数据的冗余，不符合第三范式
- 注：一、二范式是必须要遵循，第三范式可选；
- （4）范式的作用
  - 范式主要用来规范数据库的设计，使得设计出来的数据库结构清晰，简洁易懂，避免 了数据冗余和操作的异常。在设计数据库模型时，灵活地应用范式是创建一个优秀的数据 库系统的基石

- 4、数据模型
  - （1）星型模型
    - 1） 定义：当所有维表都直接连接到"事实表"上时，该模型称为星型模型
    - 2） 特征：非正规化的结构，逻辑简单，多维数据集的每一个维度都直接与事实表相连接，不存在渐变维度，所以数据有一定的冗余，一般来说性能会更好
  - （2）雪花模型
    - 1） 定义：当一个或多个维表没有直接连接到事实表上，而通过其他维度表连接到事实表上时，称雪花模型
    - 2） 优点：逻辑清晰，最大限度的减少数据存储量、联合较小的维表来改善查询功能，减少数据冗余
    - 3） 因为跟维度表要关联多次，所以效率不一定有星型模型好
  - （4）星型模型和雪花模型的优缺点比较
    - 1） 效率方面：星型模型效率比雪花模型高
      - 星型模型由于数据冗余，所以很多数据查询不需要外部连接，一般情况下效率比雪花模型效率高
      - 雪花模型由于去除了冗余，有些统计需要通过表的连接才能产生，效率不一定比星型模型高
    - 2）设计和维护：星型模型比雪花模型更简单
      - 星型模型是非正规化结构，不需要考虑很多正规因素，设计和实现比较简单
      - 雪花模型要求正规化，相应的数据库设计、数据ETL，以及后期的维护都要更复杂
    - 3）在冗余可以接受的前提下，实际运用中星型模型使用更多，也更有效率（空间换易用与效率）。

- 5、缓慢变化维SCD（Slowly Changing Dimensions）
  - （1）缓慢变化维SCD定义：维度表里的数据并非始终不变，总会随着时间发生变化
  - （2）保存历史数据的处理方式,(主要用方式2)
    - Type1：不记录历史数据,直接用新的维度数据覆盖旧的维度数据；少用
      - 难以记录历史变化
    - Type2： 添加新的维度数据行，同时保留原有记录，并在原表新建一个主键列（新增代理主键列）
      - 添加代理主键列，，增加新的数据行，保留历史变化
    - Type3：添加历史列，用不同的字段保存变化痕迹.它只能保存两次变化记录.适用于变化不超过两次的维度；
  - （3）数据表
    - 1）全量表
      - ❶每天的所有的最新状态的数据
      - ❷特点
        - （1）全量表，有无变化，都要报
        - （2）每次上报的数据都是所有的数据（变化的 + 没有变化的）
    - 2）增量表
      - ❶每天的新增数据，增量数据是上次导出之后的新数据。
      - ❷特点
        - （1）记录每次增加的量，而不是总量；
        - （2）增量是指在一定时间内的增量；
        - （3）增量表，只报变化量，无变化不用报
    - 3）拉链表
      - ❶维护历史状态，以及最新状态数据的一种表，通过拉链表可以很方便的还原出拉链时点的客户记录。
      - ❷特点
        - （1）记录一个事物从开始，一直到当前状态的所有变化的信息
        - 2）拉链表每次上报的都是历史记录的最终状态，是记录在当前时刻的历史总量；
        - （3）当前记录存的是当前时间之前的所有历史记录的最后变化量（总量）；
        - （4）封链时间可以是2999，3000，9999等等比较大的年份；拉链表到期数据要报0；
    - 4）快照表
      - ❶按日分区，记录截止数据日期的全量数据

- 2️⃣特点
  - （1）快照表，有无变化，都要报
  - （2）每次上报的数据都是所有的数据（变化的 + 没有变化的）
  - （3）一天一个分区

- 6、业务主键和代理主键
  - （1）业务主键
    - 优点：
      - 1.具有更好的检索性能。
      - 2.直观，更好可读和便于理解。
      - 3.数据迁移更加容易。
    - 缺点：
      - 1.关联性能相对不好，占空间。
      - 2.某一业务属性发生变化，会牵连很多表，修改代价大
  - （2）代理主键
    - 优点：
      - 1.纯数字，占用空间少，关联性能好。
      - 2.在业务属性发生变化时，减少了对系统的影响范围。

        举例：产品编码规则发生变化。此时，产品编码不是主键，所以只需要按照新的编码规则更改产品实体表内的“业务编号”，而不会影响到其他实体。
    - 缺点：
      - 1.数据迁移比较麻烦，存在重复ID。
      - 2.展现时需要与对应的维表关联，多做一次映射转换的动作。
      - 3.代理主键不能被改变。
  - 🔴对业务主键和代理主键的取舍，更多的是需要从系统、应用环境、实体属性与关系、开发效率、系统性能和维护成本等多方面去思考。

- 课堂笔记1：数据仓库系统：数据源-（进行ETL处理）-数据仓库DW层--根据不同部门存在数据集市DM层--数据访问（数据挖掘、数据报表）
- 课堂笔记2：数据仓库处理流程：数据源--（通过ETL工具数据）--ODS层（数据缓存层，1、表结构和数据源一致，最细维度明细数据，保证数据的准确性，抽取逻辑简单，简化DW层数据集成Join的性能，转移部分业务系统的查询压力；2、当期数据；3、数据可变）--DW(ODS层数据抽取到DW层，抽逻辑主要是join合成集成表，主要通过tiger实现)--DM（根据需求搭建不同的数据集市，大部分是汇总数据）