数据仓库原理与设计

张冬松

信阳学院
大数据与人工智能学院

dszhang@nudt.edu.cn

讲课前

微信答疑群

DWDM2020答疑群

该二维码7天内(9月15日前)有效，重新进入将更新

SmartBI——商业智能最新进展

目录 content

第一节　数据仓库原理

- 数据仓库系统结构
- 数据仓库的数据模型
- 数据抽取、转换和转载
- 元数据

1.1 数据仓库系统结构

A.数据仓库结构

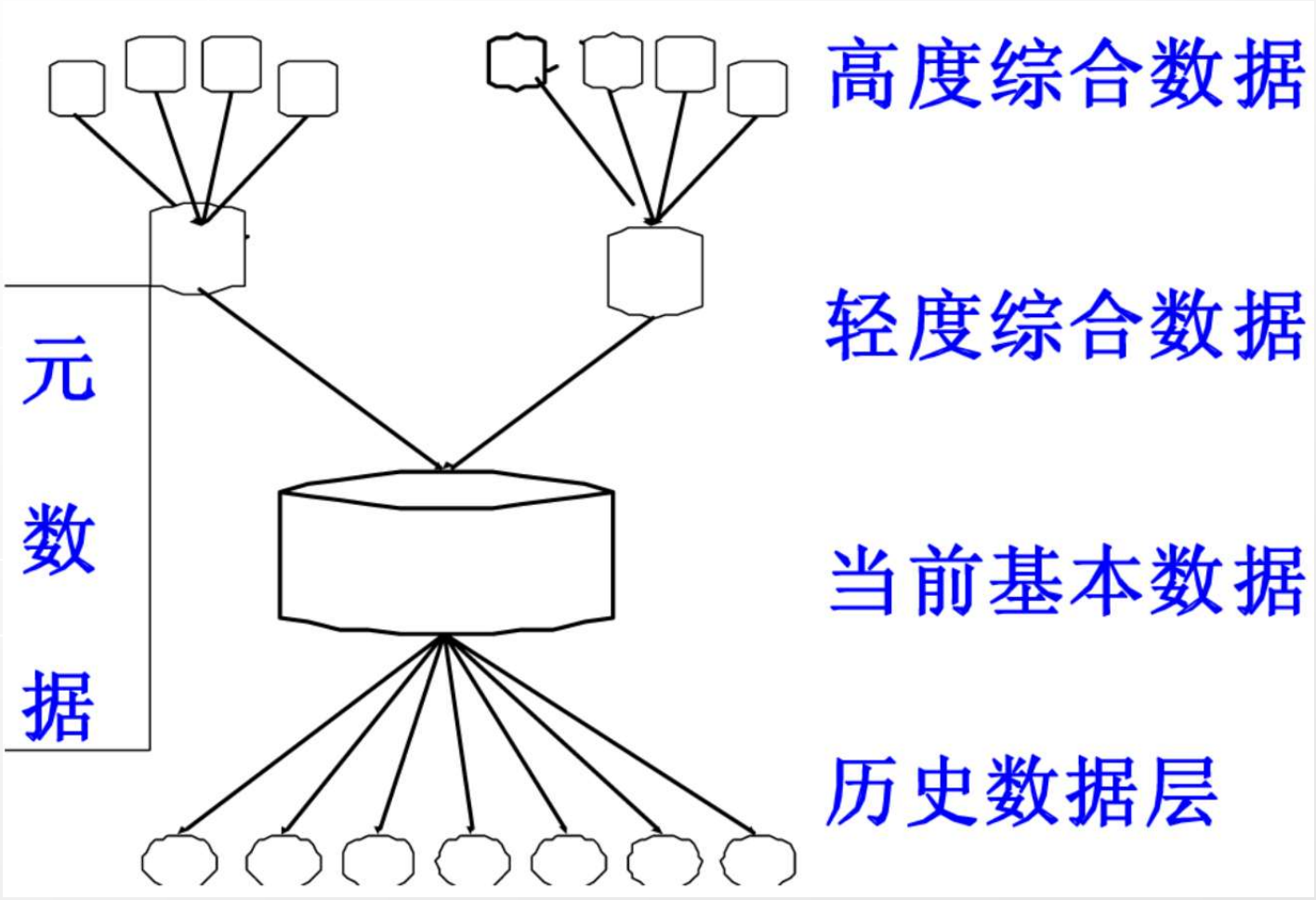
1.近期基本数据：是最近时期的业务数据，是数据仓库用户最感兴趣的部分，数量量大。

2.历史基本数据：近期基本数据随时间的推移，由数据仓库的时间控制机制转为历史基本数据。

3.轻度综合数据：是从近期基本数据中提取出来，或者按时间段选取，或者按数据属性和内容进行综合。

4.高度综合数据：是在轻度综合数据基础上的再一次综合，是一种准决策数据。

1.1 数据仓库系统结构



1.1 数据仓库系统结构

B.数据集市及其结构

1.通常开发数据仓库的代价很高，时间较长。

2.因此，提供更紧密集成的数据集市就应运而生。

3.数据集市（Data Marts）：是一种更小、更集中的数据仓库，为公司提供分析商业数据的一条廉价途径。

4.数据集市：还是指具有特定应用的数据仓库，主要针对某个应用或者具体部门级的应用，支持用户获得竞争优势。

1.1 数据仓库系统结构

C.数据集市与数据仓库区别

1.数据仓库是基于整个企业的数据模型而建立，拥有面向企业范围的主题数据库。

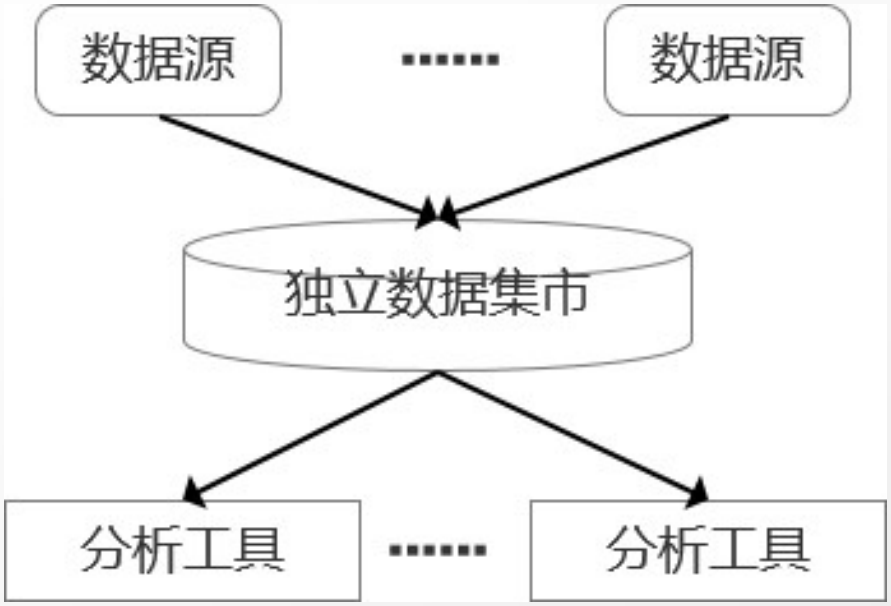
2.数据集市是按照某一特定部门的数据模型建立的，只拥有面向某个部门的主题数据库。

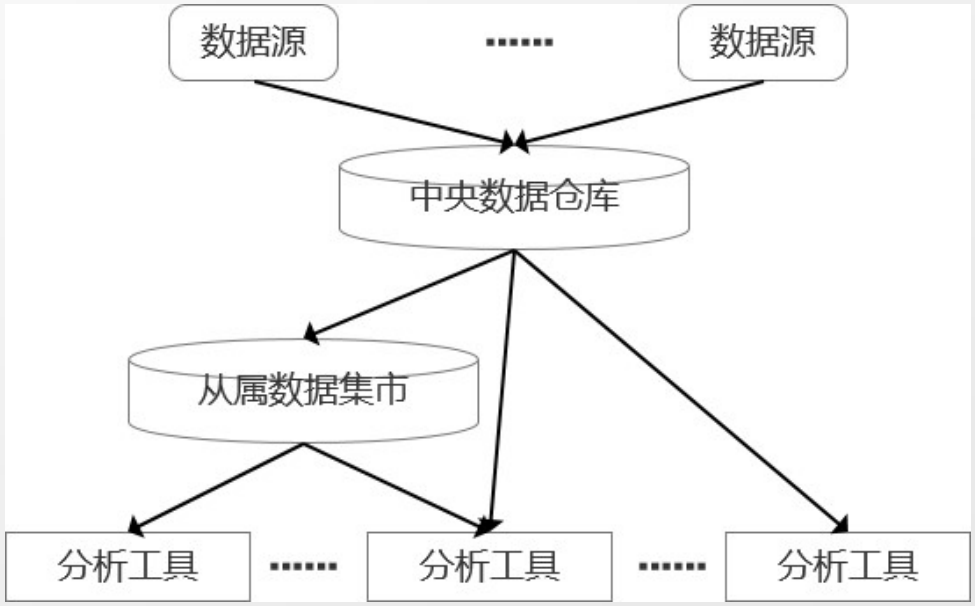
3.部门主题与企业主题之间可能存在关联，也可能不存在关联。

4.数据集市的数据组织通常采用星型模型。

1.1 数据仓库系统结构

C.数据集市与数据仓库区别

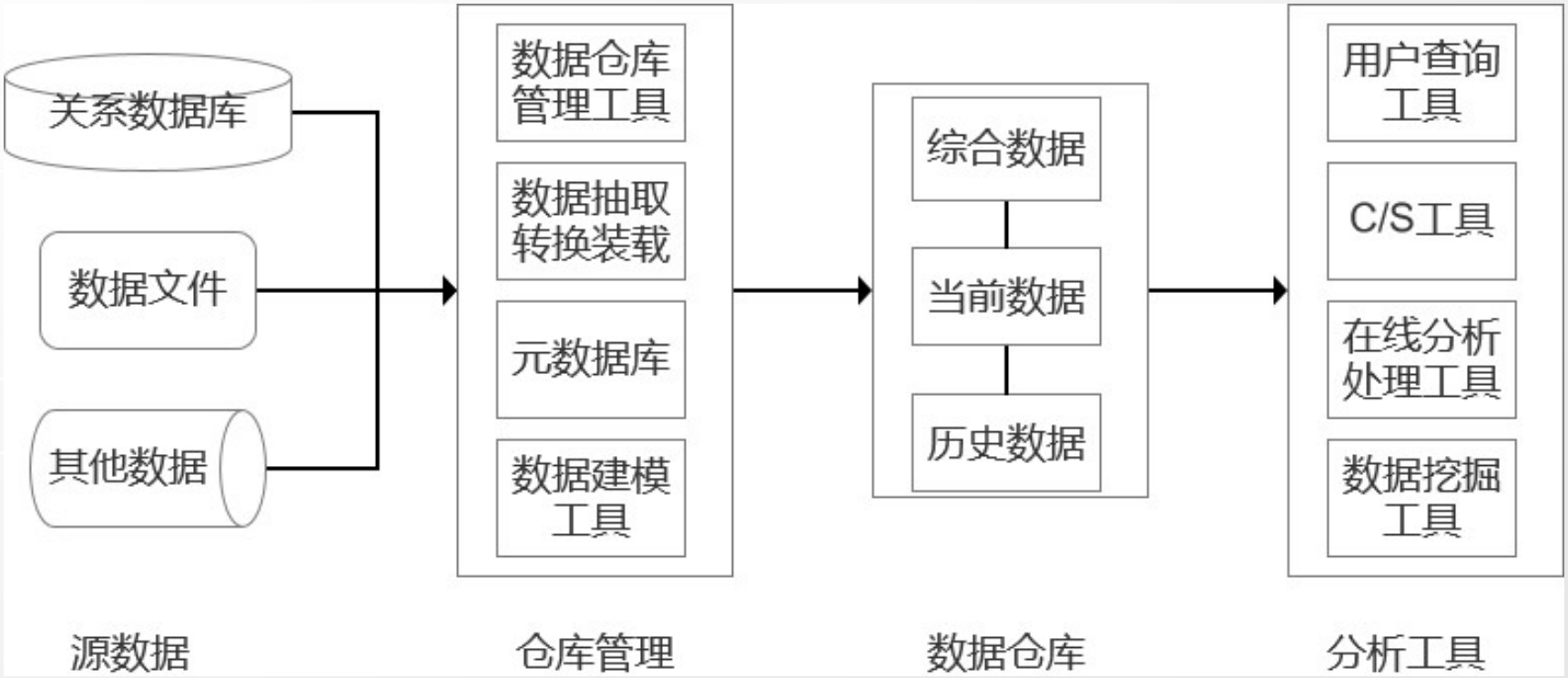




1.1 数据仓库系统结构

D.数据仓库系统结构

数据仓库系统是由数据仓库、仓库管理和分析工具组成。



第一节　数据仓库原理

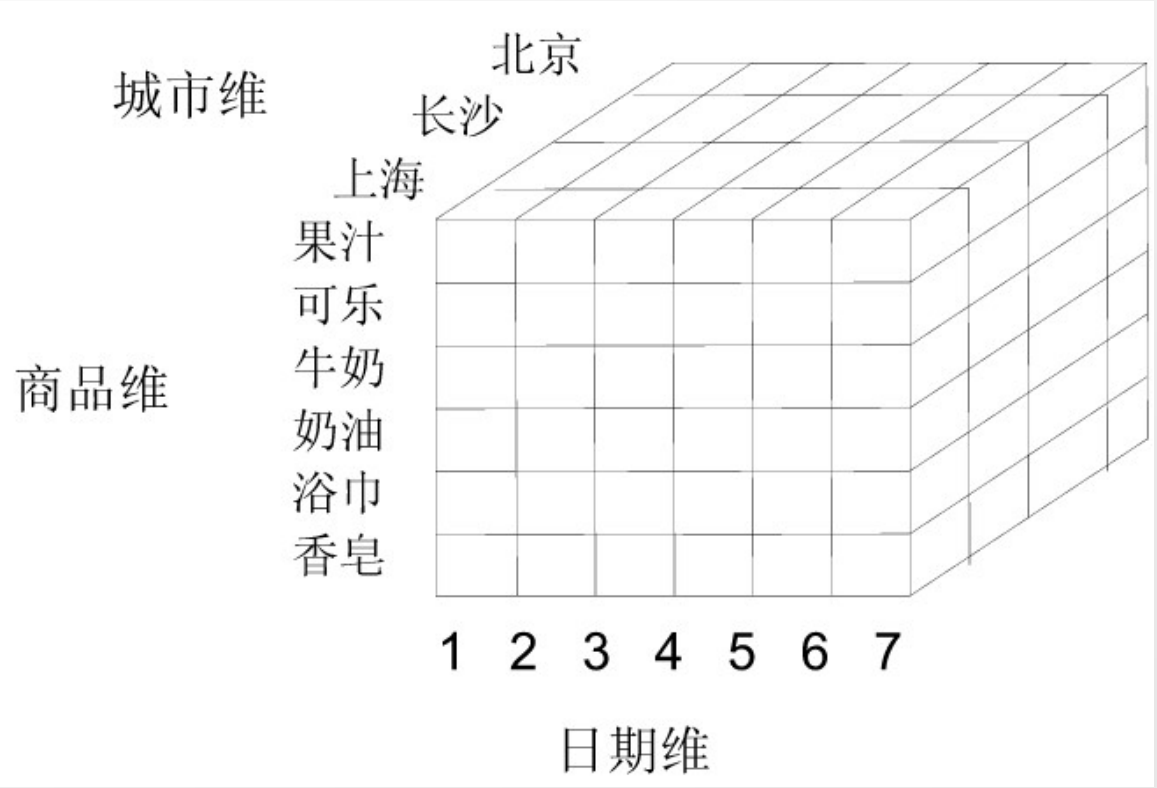
- 数据仓库系统结构
- 数据仓库的数据模型
- 数据抽取、转换和转载
- 元数据

1.2 数据仓库的数据模型

A.数据仓库采用多维数据模型

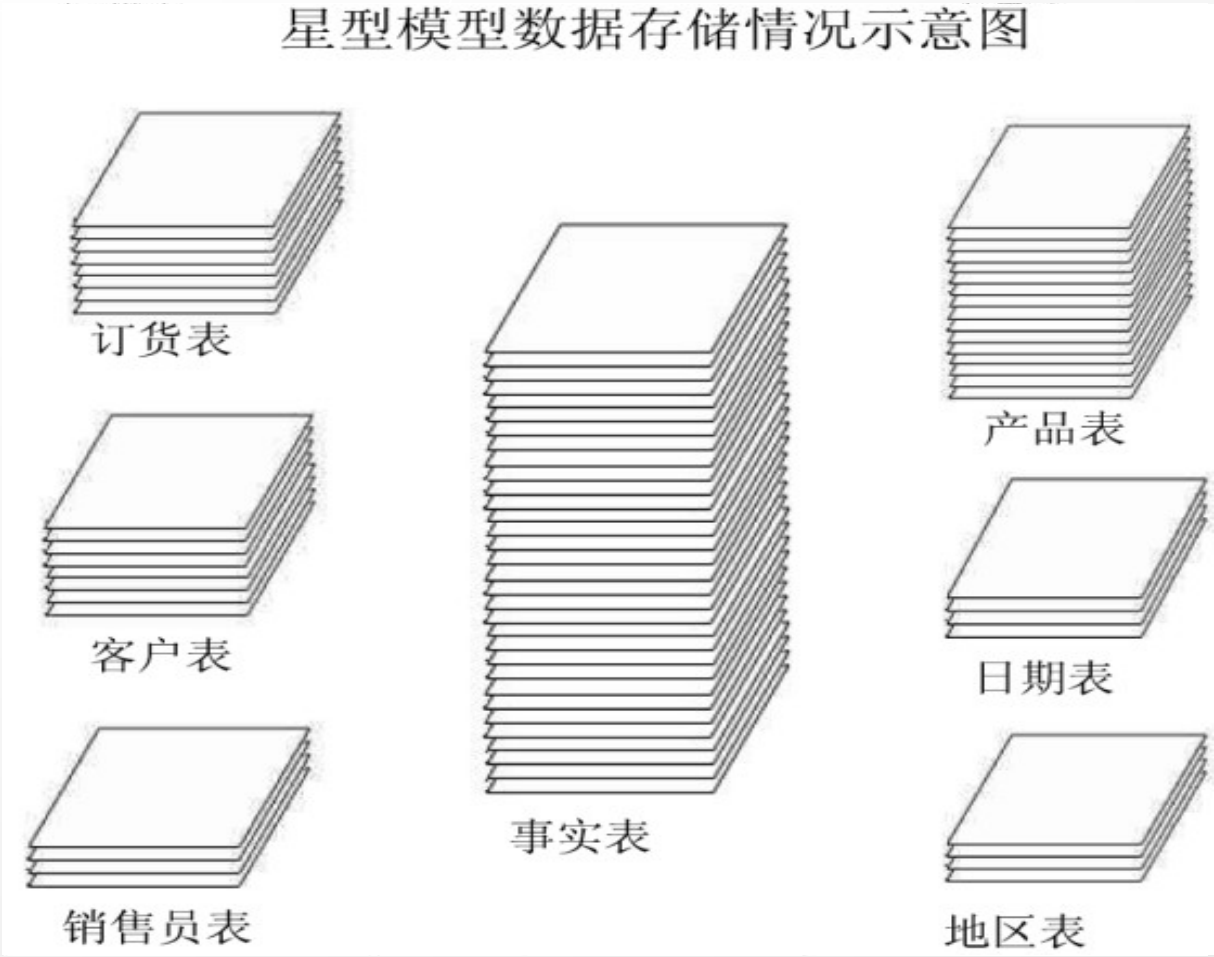
1.维就是同类数据的集合，城市、商品、日期都是维。

2.二维是电子表格，三维是立方体，四维及以上很难用图形显示。



1.2 数据仓库的数据模型

B.星型数据模型



1.星型模型是由大表和若干小表组成。

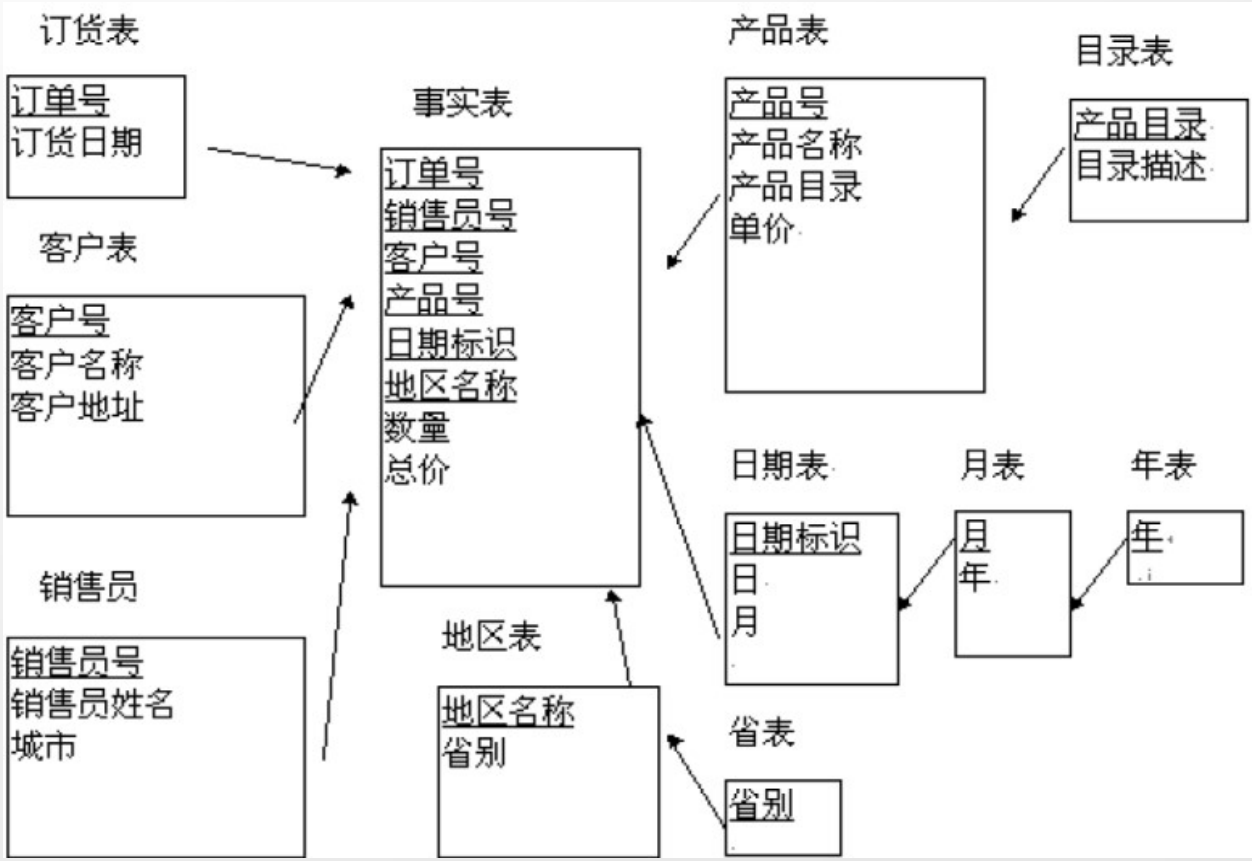
2.大表是事实表，保存大量关于企业的事实数据。

3.小表是维度表，保存描述性数据，通常是围绕事实表建立的较小的表。

1.2 数据仓库的数据模型

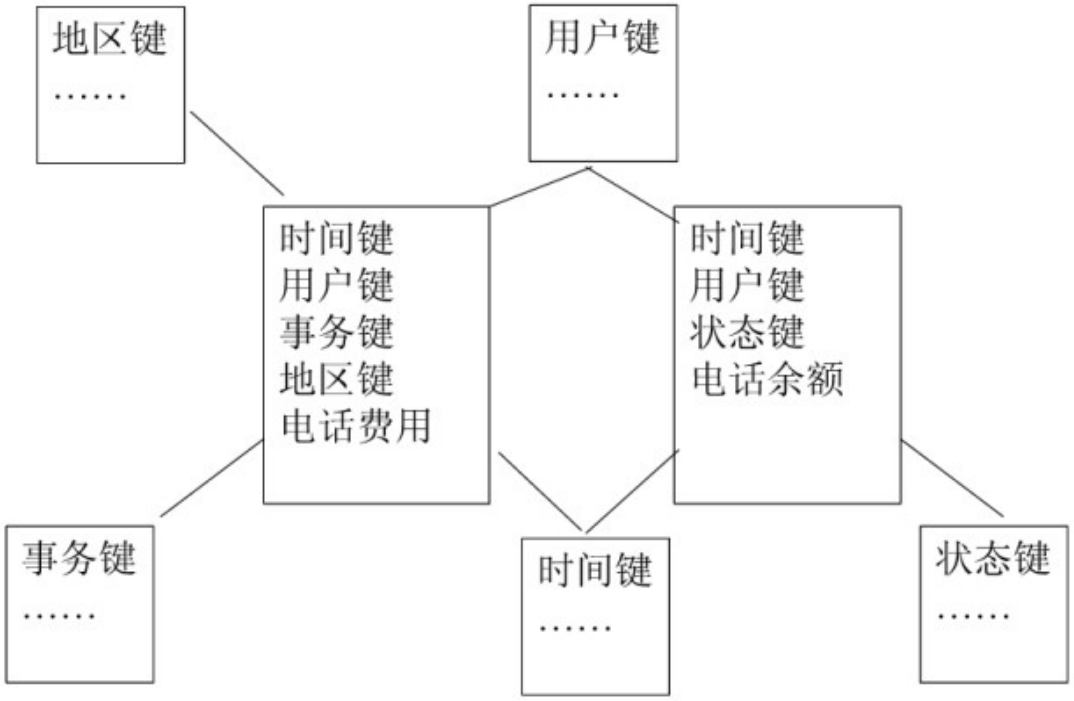
C.雪花数据模型

1.雪花模型就是对星型模型的小表进一步层次化，将原来的小表扩展为更小的事实表，形成一些局部的“层次”区域。



1.2 数据仓库的数据模型

D.星网数据模型



电话公司星网模型实例

1.星网模型是将若干个星型模型连接起来，形成网状结构。

2.若干个星型模型通过相同的维度，如时间维，连接多个事实表。

第一节　数据仓库原理

- 数据仓库系统结构
- 数据仓库的数据模型
- 数据抽取、转换和转载
- 元数据

1.3 数据抽取、转换和装载

A.ETL产生背景

1.数据仓库的数据来源于多个数据源，主要是企业内部数据、存档的历史数据，企业的外部数据等。

2.数据仓库的数据源可能是在不同的硬件平台上，使用不同的操作系统。源数据也可能是以不同的格式保存在不同的数据库中。

3.因此，数据仓库需要将这些源数据经过抽取（Extraction）、转换（Transform）、装载（Load）的过程，存储到数据仓库的数据模型中。这三个过程简称为ETL过程。

1.3 数据抽取、转换和装载

B.数据抽取

1.确认数据源

1.1 列出对事实表的每一个数据项和事实

1.2 列出每一个维度属性

1.3 对于每个目标数据项，找到源数据项

1.4 确认目标字段的多个数据源，建立合并规则

1.5 确认目标字段的多个数据源，建立分离规则

1.6 确定默认值

1.7 检查缺失值的源数据

2.数据抽取

2.1 当前时刻的值

2.2 周期性状态值

1.3 数据抽取、转换和装载

C.数据转换

1.基本功能

1.1选择：从数据源中选择整个或者部分数据。

1.2分离或合并：对源数据进行分离操作或者合并操作。

1.3转化：对数据源进行标准化和可理解化。

1.4汇总：对最低粒度数据进行汇总。

1.5清晰：对单个字段数据进行重新分配和简化。

2.如何实施转换

2.1自己编写程序

2.2使用转换工具

1.3 数据抽取、转换和装载

D.数据装载

1.数据装载方式

1.1基本装载：将转换后的数据输入到数据仓库的目标表中

1.2追加：追加过程在保留原有数据的基础上增加输入数据

1.3破坏性合并：用新输入数据更新目标数据

1.4建设性合并：保留已有数据，增加输入的新数据，标记为旧数据的替代

2.数据装载类型

2.1最初装载：第一次对整个数据仓库进行装载

2.2增量装载：数据仓库根据需要装载变化的数据

2.3周期装载：周期性重写数据仓库

1.3 数据抽取、转换和装载

E.ETL过程总结

第一节　数据仓库原理

- 数据仓库系统结构
- 数据仓库的数据模型
- 数据抽取、转换和转载
- 元数据

1.4 元数据

A.元数据定义

1.最基本的元数据就是数据库中的数据字典。

2.元数据定义了数据仓库有什么，说明了数据仓库中数据的内容和位置，刻画了数据的抽取和转换规则，存储了与数据仓库主题有关的各种信息，可以说，整个数据仓库的运行都是基于元数据的。

3.元数据不仅是数据仓库的字典，还是数据仓库本身信息的数据。

4.元数据（Meta data）：定义为关于数据的数据。换句话说，元数据描述了数据仓库的数据和环境。

1.4 元数据

B.元数据类型

1.元数据描述了数据的结构、内容、键、索引等项内容。

2.在数据仓库中，元数据定义了数据仓库中的许多对象——表、列、查询、商业规则或是数据仓库内部的数据转移。

3.元数据是数据仓库的重要构件，是数据仓库的指示图。

<table>
<tr><td colspan="8">元数据</td></tr>
<tr><td colspan="8">动态元数据</td></tr>
<tr><td>入库
时间</td><td>更新
周期</td><td>数据
质量</td><td>统计
信息</td><td>状态</td><td>处理</td><td>存储
位置</td><td>存储
大小</td><td>引用处</td></tr>
</table>

1.4 元数据

C.元数据组成

1.数据仓库中的数据字典

2.数据源的元数据

3.数据模型的元数据

4.数据源与数据仓库映射的元数据

5.数据仓库应用的元数据

1.4 元数据

D.两类重要元数据

1.数据仓库的用户最关心的是两类元数据：

（1）元数据告诉数据仓库中有什么数据，都从哪里来，即如何按主题查看数据仓库中的数据。

（2）元数据提供已有的可重复利用的查询语言信息。如果某个查询能够满足他们的需求，或者与他们的愿望相似，他们就可以再次使用那些查询而不必从头开始编程。

2.数据仓库使用的元数据能帮助用户到数据仓库中查询所需要的数据，用于解决用户遇到的问题。

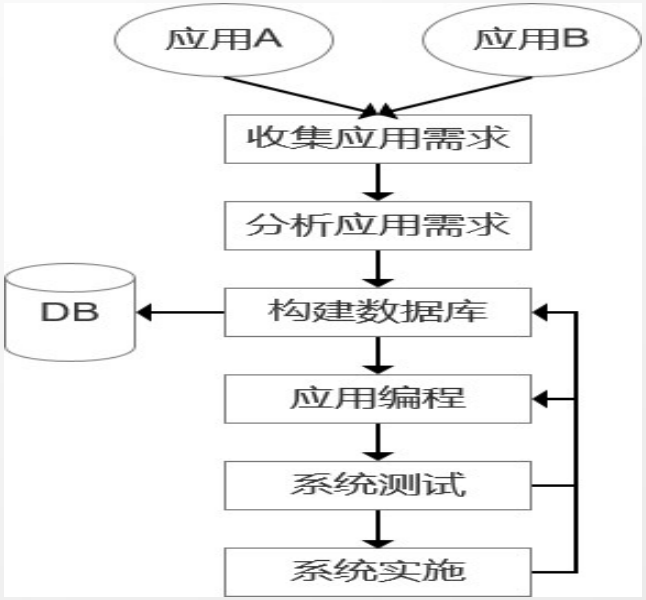
第二节　数据仓库设计

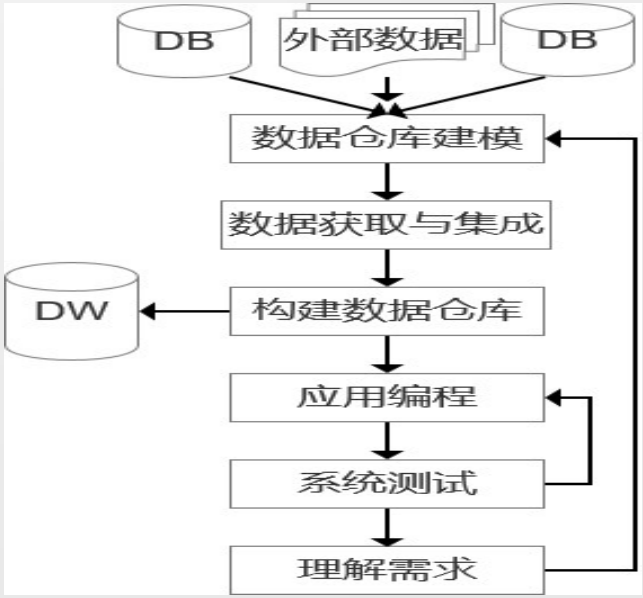
- 数据仓库的设计原则
- 数据仓库的设计步骤

2.1 数据仓库的设计原则

1.数据仓库的设计类似数据库设计，但又有不同。

2.不同在于：数据库是面向事务型处理，所以事务型处理性能是数据库设计的主要目标；而数据仓库是面向决策分析，所以更关心是建立起一个全局一致的分析型处理环境。





2.1 数据仓库的设计原则

A.面向主题原则

- 建立数据仓库的目的
- 数据仓库中数据的组织方法

B.数据驱动原则

- 数据来源
- 数据驱动方法

C.原型法设计原则

- 原型法即从构建系统的基本框架入手，不断丰富完善
- 一个逐步求精的过程，不断循环、反馈而要求决策者与开发者共同参与和密切合作

第二节　数据仓库设计

- 数据仓库的设计原则
- 数据仓库的设计步骤

2.2 数据仓库的设计步骤

1.明确主题

2.概念设计（高层设计）

3.技术准备

4.逻辑设计（中间层设计）

5.物理设计（底层设计）

6.数据仓库创建

7.数据仓库的运行与维护

第三节　报告题目列表

- 报告题目1
- 报告题目2
- 报告题目3

3.1 报告题目1

1.信阳学院学生学籍管理系统中的数据仓库建立方案设计研究

结合信阳学院学生学籍管理系统进行数据仓库的建立方案设计，列举有哪些主题数据库划分？包含有哪些数据表？并划分维度层次。

第三节　报告题目列表

- 报告题目1
- 报告题目2
- 报告题目3

3.2 报告题目2

2.信阳学院教学管理数据仓库系统的分析与设计研究

结合信阳学院教学管理工作，进行教学管理数据仓库系统的分析与设计，列举完整的教学管理数据仓库系统的体系结构，功能组成，使用方式等。

第三节　报告题目列表

- 报告题目1
- 报告题目2
- 报告题目3

3.3 报告题目3

3.基于Hapdoop的分布式数据仓库的关键技术与应用研究

结合当前主流的开源云平台Hadoop，分析并研究分布式数据仓库的关键技术与应用方法，以了解掌握工业界当前大数据应用发展为己任。

张冬松
dszhang@nudt.edu.cn

谢谢！ Q&A

THANKS FOR YOUR ATTENTION