第十七章

数据仓库

数据分析

- 数据仓库
- 在线分析处理
- 数据挖掘

数据分析

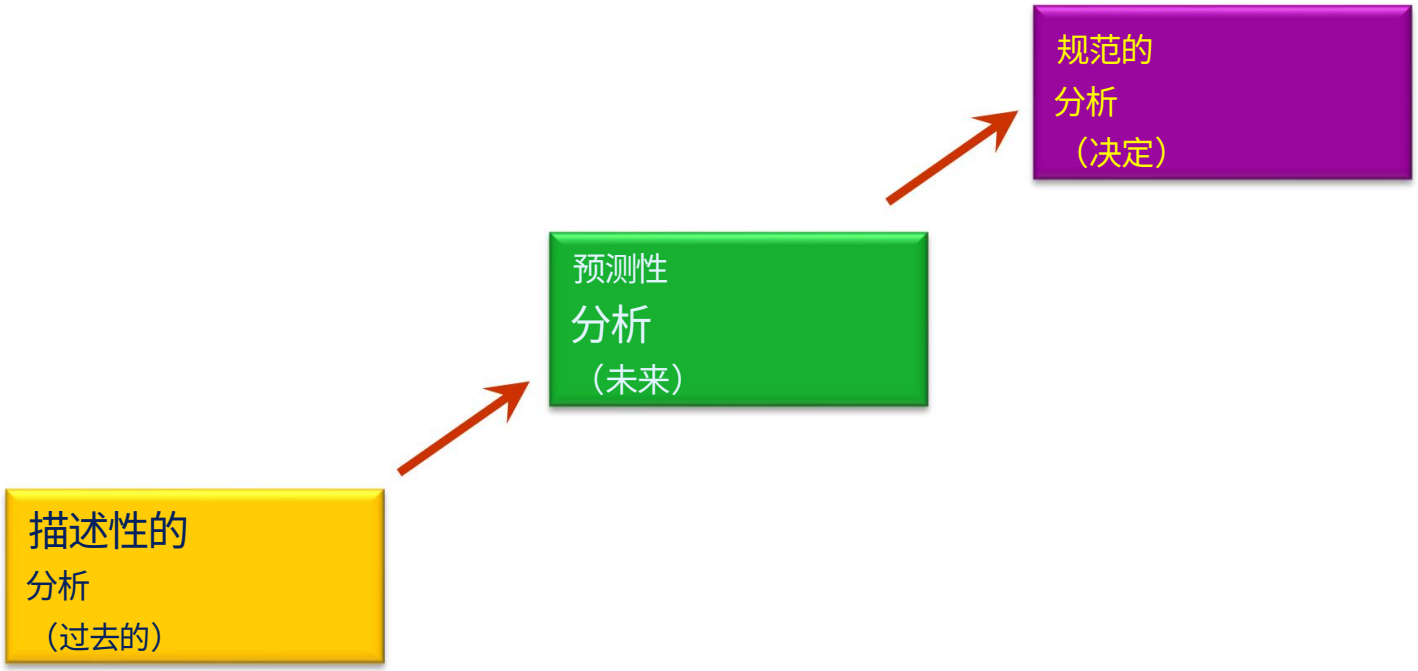
- 数据分析是指处理数据以推断模式，
  相关性或模型，其结果用于推动业务决策

- 预测模型被广泛使用
  - 例如，使用客户资料特征（例如，收入、年龄、性别、
    教育，就业）和客户的过去历史来预测贷款违约的可能性
    - 使用预测做出贷款决策
  - 例如，使用过去的销售历史（按季节）来预测未来的销售
    - 决定生产什么/多少/库存
    - 瞄准客户
- 业务决策示例：
  - 要库存哪些物品？
  - 要收取什么保险费？
  - 向谁发送广告？

数据分析 Analytics

- 机器学习技术是发现数据模式和做出预测的关键
- 数据挖掘扩展了机器学习社区开发的技术以在非常大的数据集上运行它们
- 商业智能 (BI)一词广泛用于类似于数据分析的意义
- 决策支持一词与BI 相关但狭义,侧重于报告和汇总;相关系统是DSS（决策支持系统）

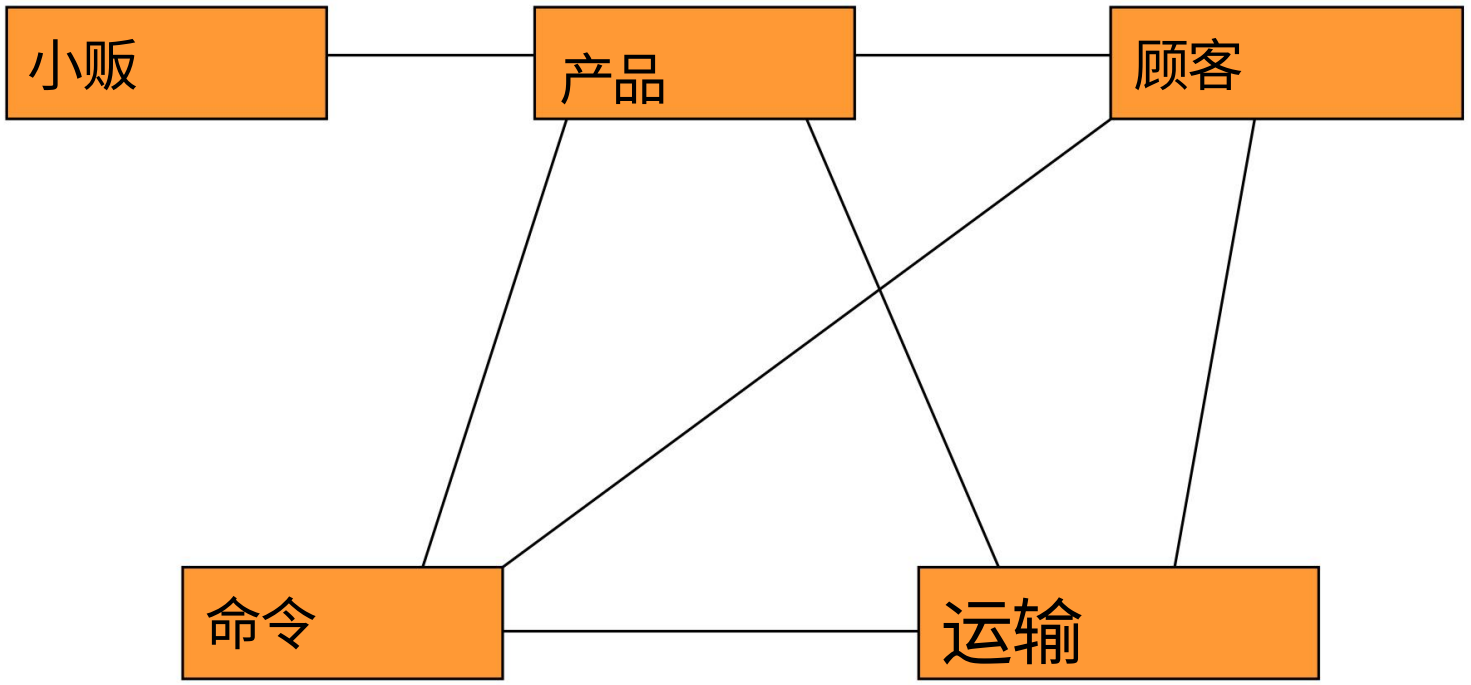
数据分析



多维数据和仓库模式

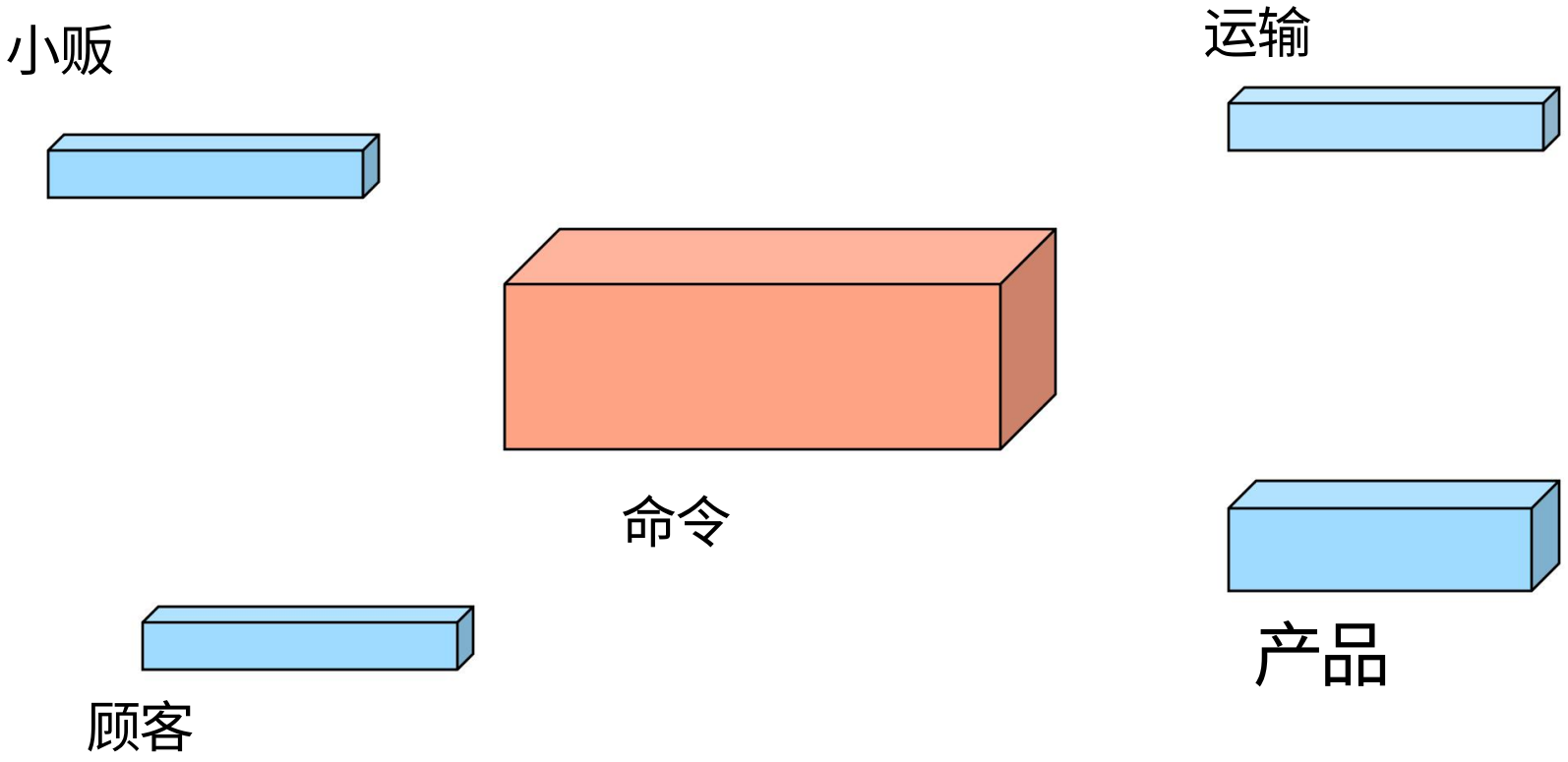
- 数据仓库通常使用星型模式，有时也使用雪花模式图式
  - 与维度表连接的事实表
  - 对维度表属性进行分组
  - 汇总事实表的度量属性
- 一些应用程序不认为将数据带到一个公共的图式
  - 数据湖是允许数据存储在多个格式，没有模式集成
  - 前期工作量较少，但查询期间的工作量较大

实体关系建模的局限性

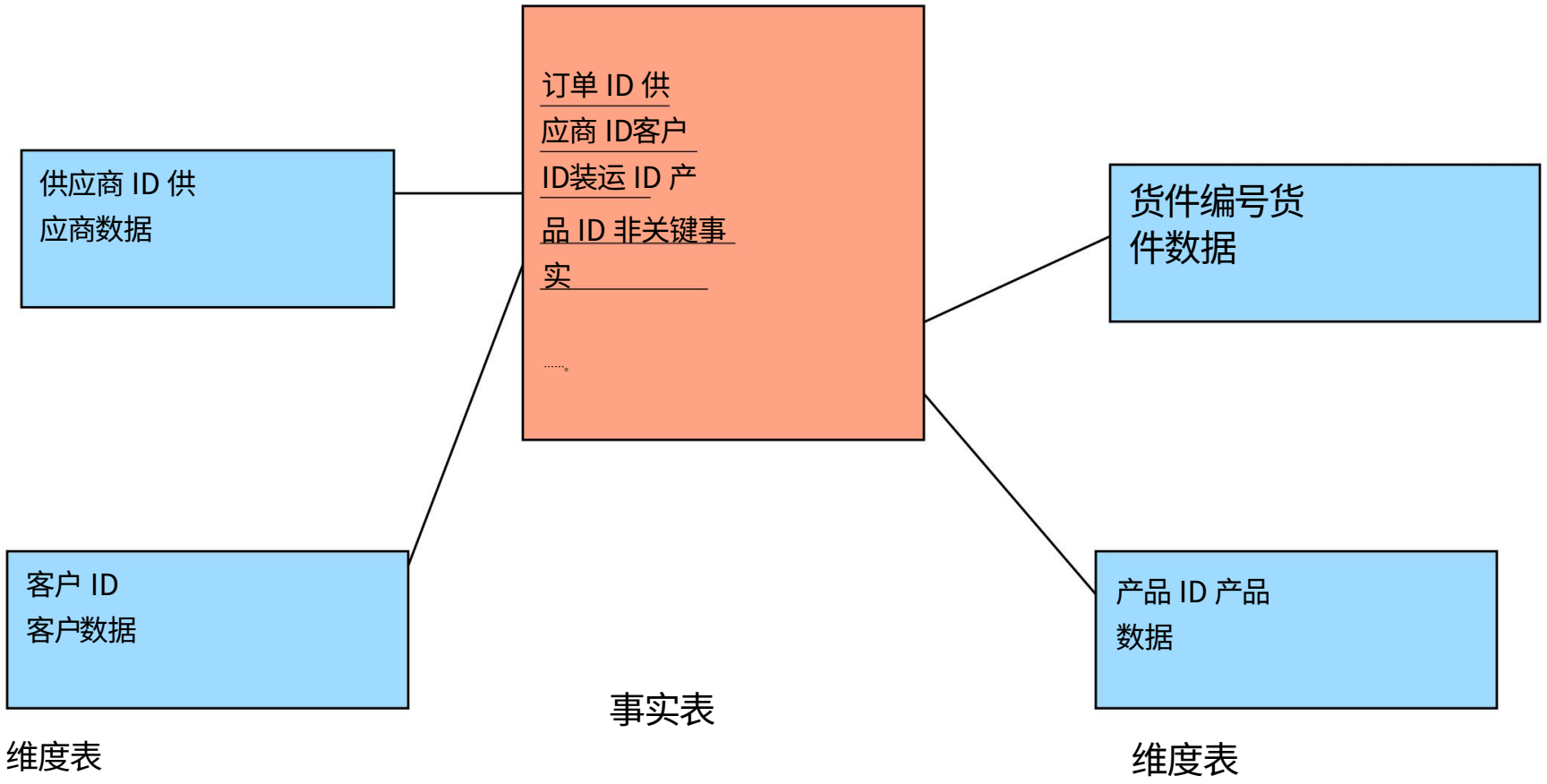


- Very symmetric
- Cannot tell which table is most important or largest
- Cannot tell which tables hold static or dynamic business information
- Joining of any tables is possible by user

3-D 透视图：
订单数量更多



星型模式:事实和维度表



明星时间表

- 星的中心由事实表组成,星的点是维度表
- 星型模式的特点是包含数据仓库中的主要信息的非常大的事实表和许多小得多的维度表（或查找表）,每个维度表都包含有关事实表中特定属性的条目的信息
- 星型查询是事实表和多个查找之间的连接表
- 每个查找表都使用主键到外键连接连接到事实表,但查找表不相互连接

明星时间表

- 非常不对称
- 事实表是唯一有多个连接将其连接到其他表的表
- 所有其他表只有一个连接将它们附加到
  中央桌子
- 常用于数据仓库

事实表 Fact Table

- 事实表往往包含附加事实
- 事实表具有复合键
- 所有其他表都是维度表
- 每个关键值的组合都会产生不同的记录在事实表中
- 事实表自然是高度标准化的

维度表 nsion Tables

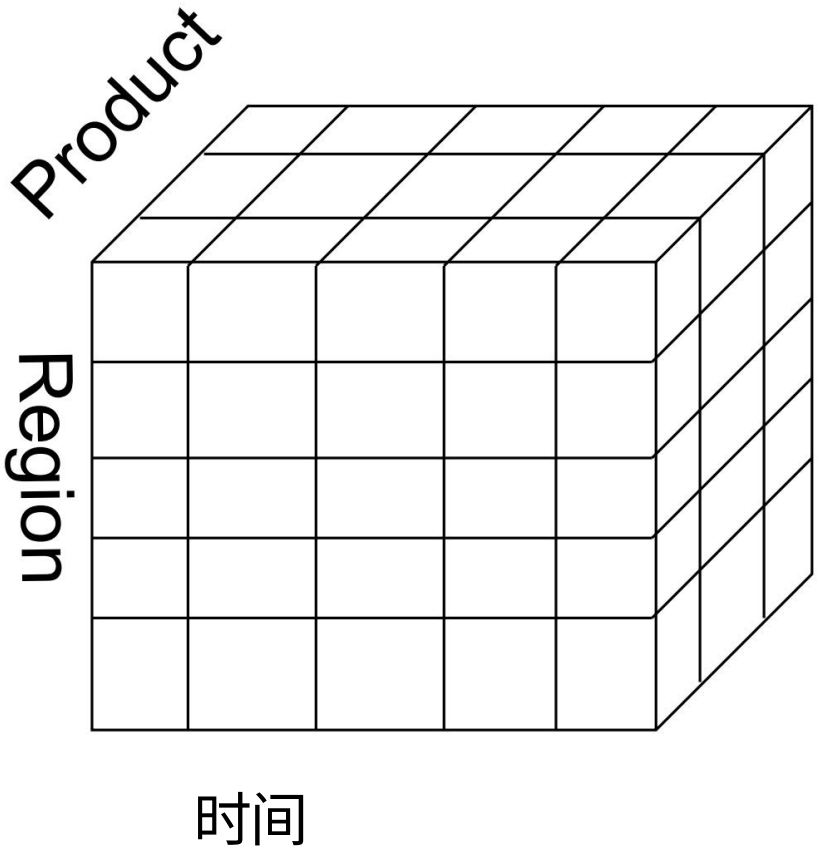
- 维度表倾向于包含文本或非附加事实
- 维度表不应该被规范化
- 规范化维度表破坏了
  浏览

加入

- 通常，一个维度数据库中只有几个连接（通常将事实表与一个或多个维度表连接）

- 每个连接都表达了它们之间的基本关系
  基础业务中的项目
- ER 数据库中原则上可以进行任何连接
  - 大多数意义不大

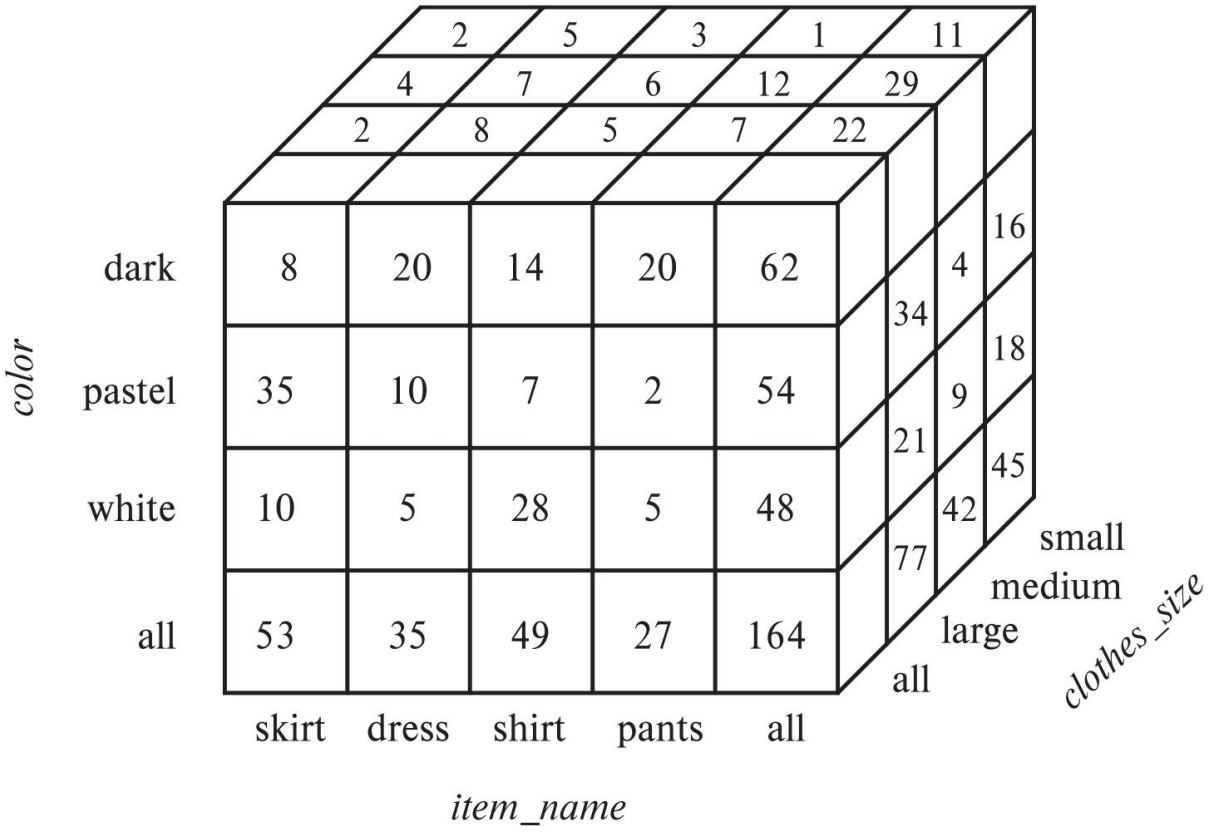
超立方体/数据立方体视图



多维信息的概念视图是超立方体（也称为数据立方体）。

事实表中的每个键组合对应一个小立方体（骰子）

超立方体/数据立方体



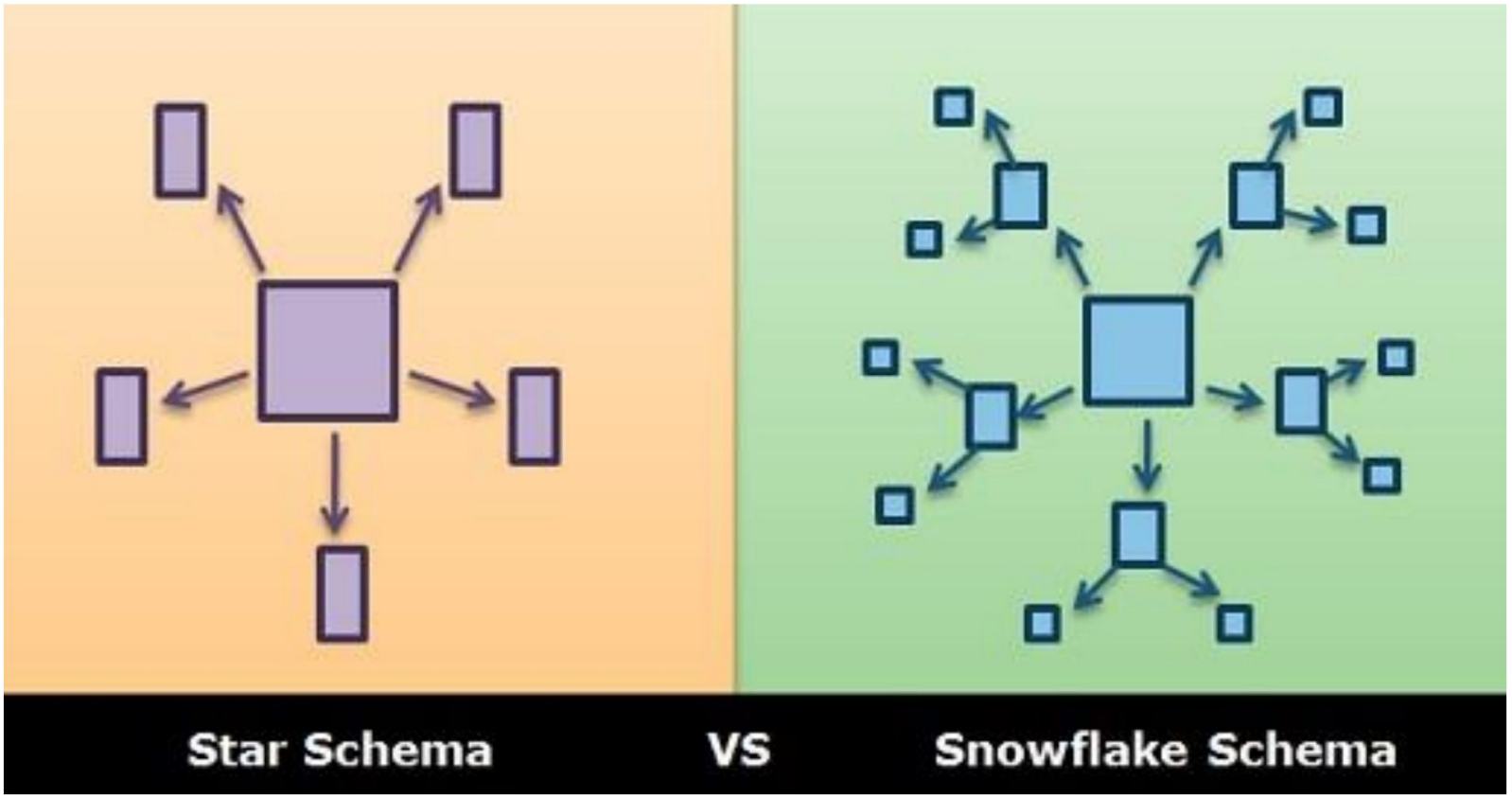
明星加盟

- 星型连接是主键到外键连接的
  维度表到事实表
- 事实表通常在键列上有一个连接索引,以促进这种类型的连接
- 星型模式的主要优点是:
  - 在最终用户分析的业务实体和模式设计之间提供直接和直观的映射
  - 为典型数据提供高度优化的性能
    仓库查询

雪花模式 Snowflake Schema

- 雪花模式是比星型模式更复杂的数据仓库模型,是星型模式的一种
- 它被称为雪花模式,因为
  架构类似于雪花
- 雪花模式标准化维度以消除冗余
- 维度数据被分组到多个表而不是一个大表中

雪花模式



雪花模式 Snowflake Schema

- 雪花模式不受欢迎
- 例如,星型模式中的产品维度表可能被规范化为 Product 表、Product_Category 表和

  雪花模式中的 Product_Manufacturer 表

  - 给定产品可能是小型家用产品,但编码为
    Product 表中的产品类别 5,以及单独的 Product_Category
    表还会提供其他信息,例如重量、颜色、儿童安全等。(并且 category_id 上的外键连接是获取此类信息所必需的)

  - 制造商信息可以存储在一个单独的表格中,给出
    制造商的详细信息(并且需要在manufacturer_id 上加入外键才能获得完整的制造商信息)

- 虽然这节省了空间,但它增加了维度表的数量并需要更多的外键连接

- 结果是更复杂的查询和降低的查询性能

数据和信息

- 信息不易从数据中获取
- “今年的账户活动与过去五年的每一年？”
- 存储的历史数据不足以满足 DSS 要求

运营数据与 DSS 数据

▪面向应用	▪以主题为导向
▪详细	▪总结
▪准确	▪表示一段时间内的值 ▪流程中使用的大量数据
▪使用少量数据过程	
▪服务于文职社区 ▪经常更新 ▪重复运行	▪服务于高管社区 ▪未更新 ▪启发式运行 ▪分析驱动 ▪宽松的可用性
▪交易驱动	
▪高可用性	
▪性能敏感	▪表现轻松

数据仓库

- 数据仓库是一个
  - 面向主题
  - 集成
  - 非易失性
  - 时间变量

收集数据以支持管理层的决策

学科方向 Orientation

操作

- 关注组织的应用领域
- 保险公司
  - 自动
  - 生活
  - 健康
  - 事故

数据仓库

- 关注组织的主要主题领域
- 保险公司
  - 客户
  - 保单
  - 保费
  - 索赔

一体化

- 将多个操作数据库中的数据转换为数据仓库
- 问题
  - 数据编码
  - 属性测量
  - 多个来源 · 密钥冲突

非波动性

操作

- 高度可变
- 大量更新
- 逐个记录操作记录

数据仓库

- 没有更新
- 大量加载和访问
- 刷新数据仓库

时变 Time Variancy

操作

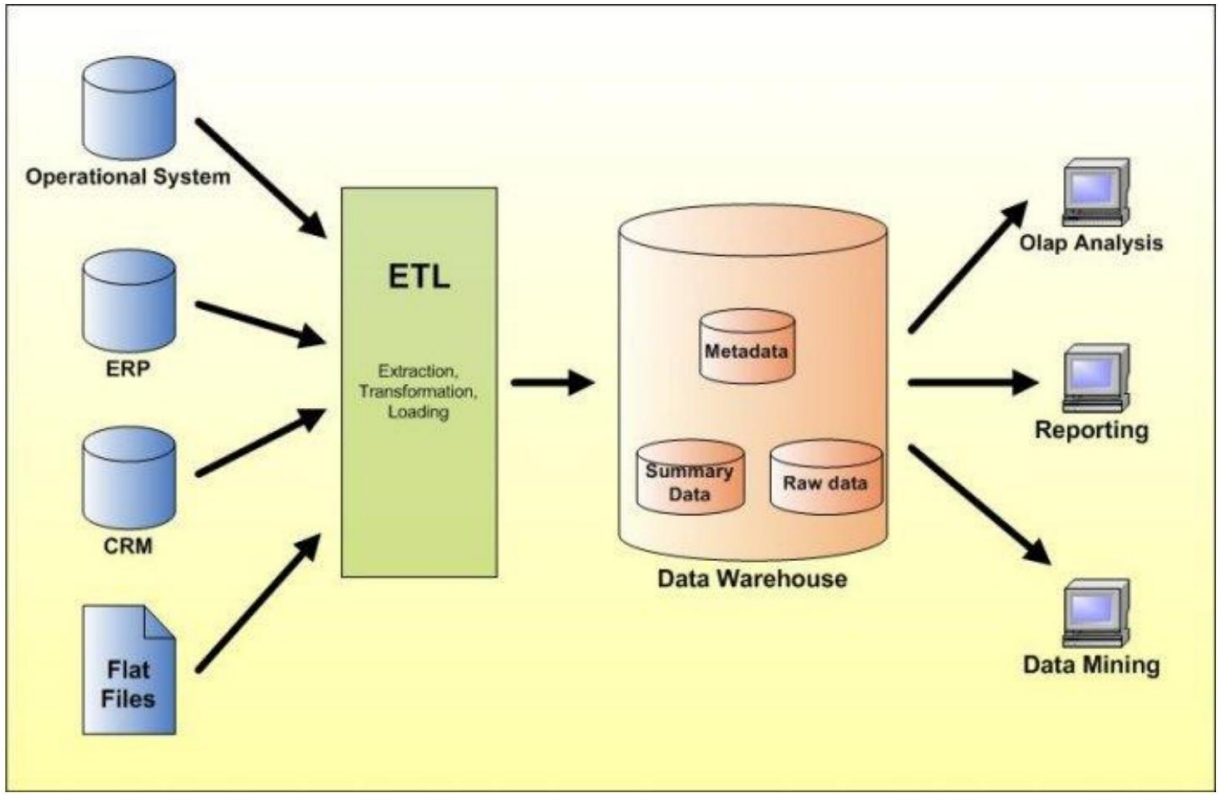
- 时间范围:当前至 60-90 天
- 准确的时间
  使用权
- 关键结构可能/可能不包含时间元素

数据仓库

- 时间跨度:5-10 年
- 复杂的数据快照
- 关键结构包含时间元素

数据仓库架构

Data Warehouse Architecture



- 结合多个数据源,例如,
  - 客户关系管理 (CRM)
  - 企业资源规划 (ERP)
  - 供应链管理 (SCM)
- 支持管理决策

数据库 Mart

- 数据集市是数据仓库的一种简单形式，专注于单一主题（或功能领域），例如销售或财务或营销
  - 数据集市通常由单个部门构建和控制在一个组织内
- 鉴于其单一主题，数据集市通常仅从一些来源
  - 来源可以是内部操作系统、中央数据仓库或外部数据
- 三种基本类型的数据集市
  - 相关数据集市从中央数据仓库中抽取数据，该数据仓库已经创建
  - 独立数据集市，是通过直接从运营或外部数据源或两者中提取数据而构建的独立系统
  - 混合数据集市可以从操作系统或数据中提取数据仓库