数据仓库、 OLAP及数据立方体计算

什么是数据仓库

有多种但并不严格的定义

1. 与操作数据库相隔离并单独维护的一个用来支持决策过程的数据库。
2. 一个用来对整理过的历史数据进行分析以便支持信息处理的固定平台。
3. “数据仓库是面向主题的、集成的、时变的、非易失的数据集合，它用来支持管理部门的决策过程”—W. H. Inmon

数据仓库的特征

1. 面向主题的：围绕主题组织, 如消费者（customer）、产品（ product）， 销售量（sales）等。主要目的是对数据建模与分析，以便于决策者的决策过程，而不是日常操作与事物处理。

2. 集成的：集成多个、异构数据源

3. 时变的：数据仓库跨越的时间比操作数据库要长的多。
   操作数据库: 当前值数据。
   数据仓库: 从历史的视角提供信息（如过去5-10的数据）

4. 非易失的：与操作数据库分隔存储。操作数据库的数据更新不在数据仓库环境出现。仅仅需要以下2种操作: 数据的初始装载与数据访问。

数据仓库 vs. 数据库管理系统

联机事物处理 (OLTP， on-line transaction processing)

- 传统关系数据库的主要任务
- 日常操作 : 购买, 存货, 财务等.

联机分析处理 (OLAP，on-line analytical processing)

- 数据仓库的主要任务

- 数据分析与决策支持

	OLTP	OLAP
用户	员工, IT专业人员	知识工作者
功能	每天的日常操作	决策支持
DB设计	面向应用+ER	面向主题+Star
数据	当前的，详细的数据	历史的, 汇总的, 多维的集成的, 整理过的
使用	重复的	特定的
访问	读/写、索引	多次扫描
工作单元	短的, 简单的事务处理	复杂查询
记录数/查询	几十	百万
用户数	上千	百
DB规模	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

为什么要建立隔离的数据仓库

1. 使得操作数据库与数据仓库都获得高性能
   DBMS—OLTP: 访问方法, 索引, 并发控制, 数据恢复。
   Warehouse—OLAP: 复杂OLAP查询, 多维视图, 整理。

2. 对数据与功能的要求不同:
   丢失的数据: 决策支持需要历史数据，而传统数据库并不一定维护历史数据。
   数据整理: 决策支持需要对异构数据源进行数据整理 。
   数据质量: 不同的数据源常常具有不一致的数据表示，编码结构与格式。

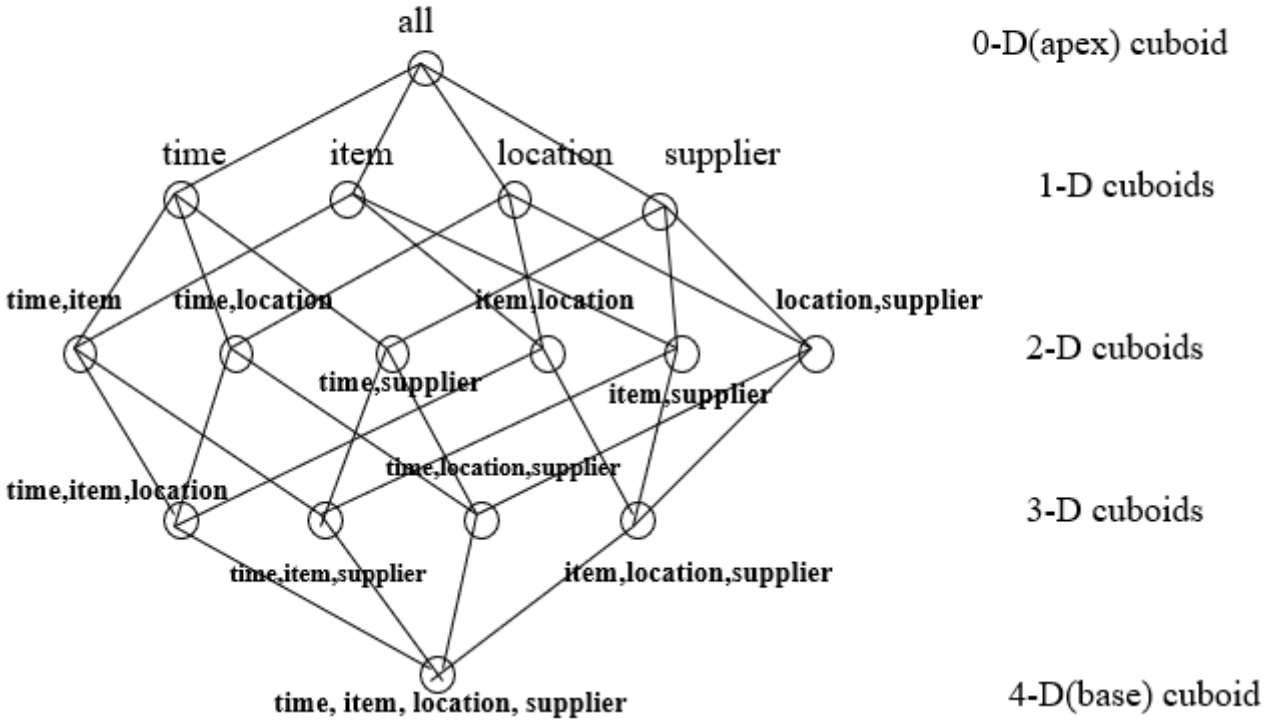
多维数据模型

数据仓库基于多维数据模型，以数据立方体的形式对数据进行观察。
数据立方由**维和度量组成**

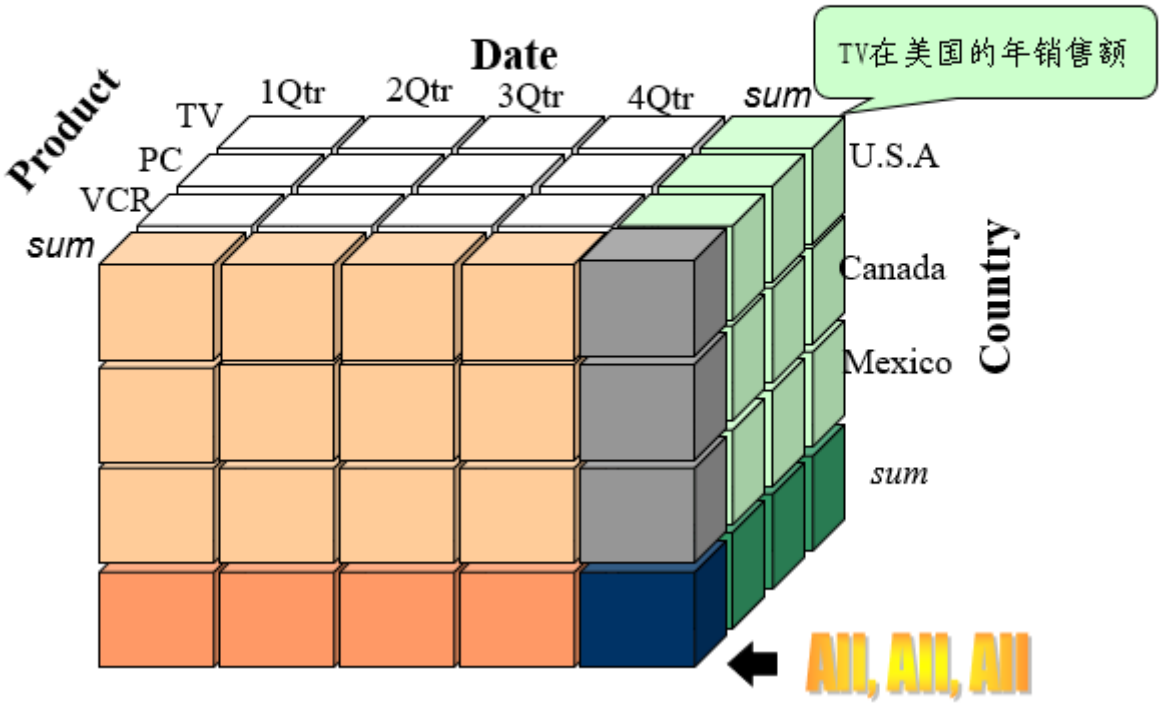
- 维表：如维item (item_name, brand, type)，或维time(day, week, month, quarter, year) 。
- 事实表包含度量（measures）：如销售额以及每个相关维表的关键字。

立方体：方体格

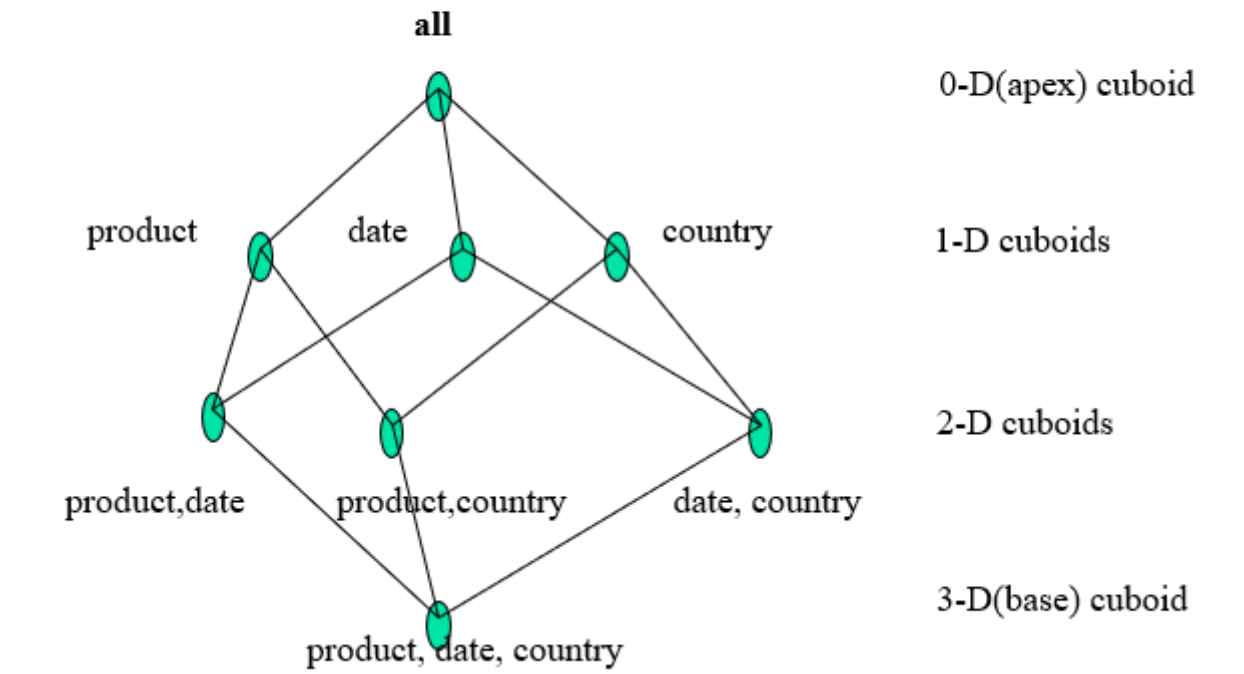
立方体：方体格



示例：数据立方体



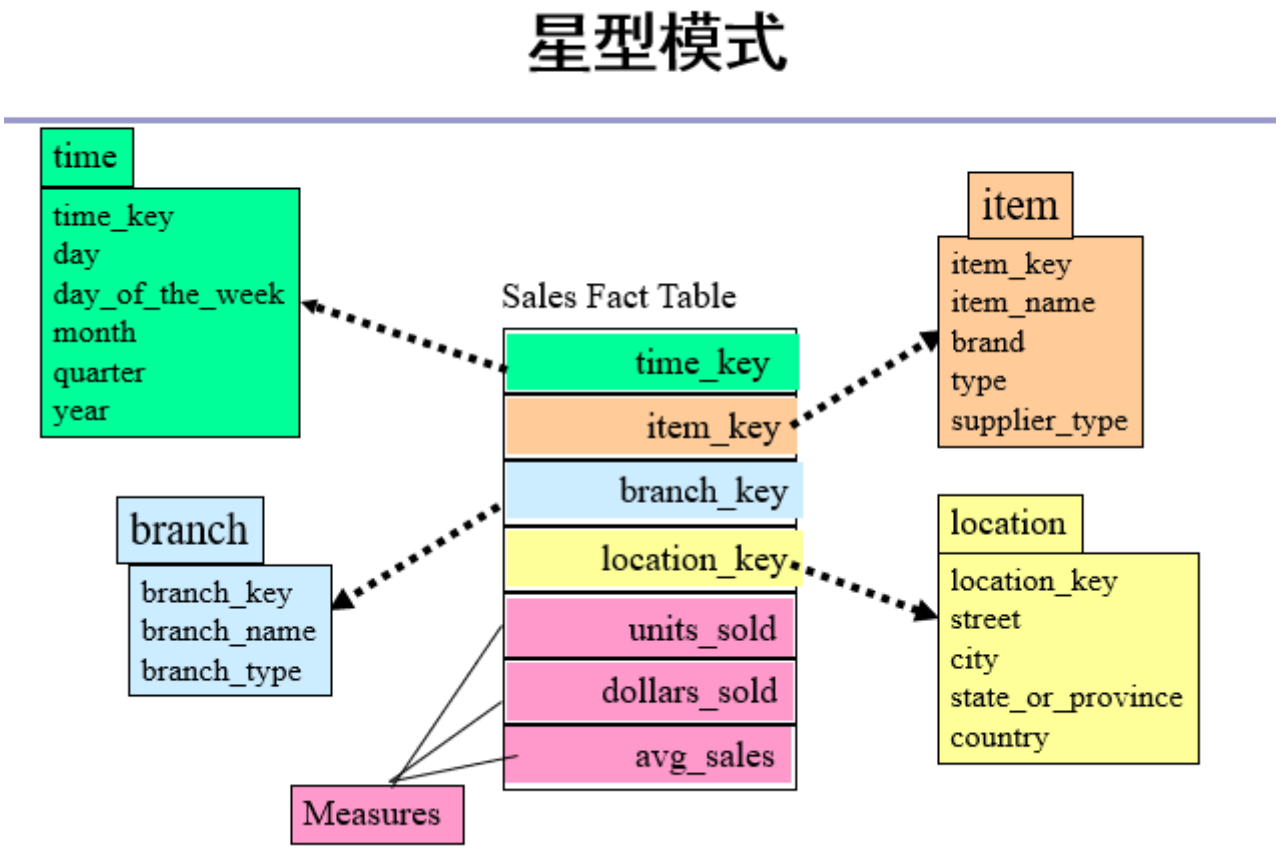
generated by haroopad



数据仓库概念模型

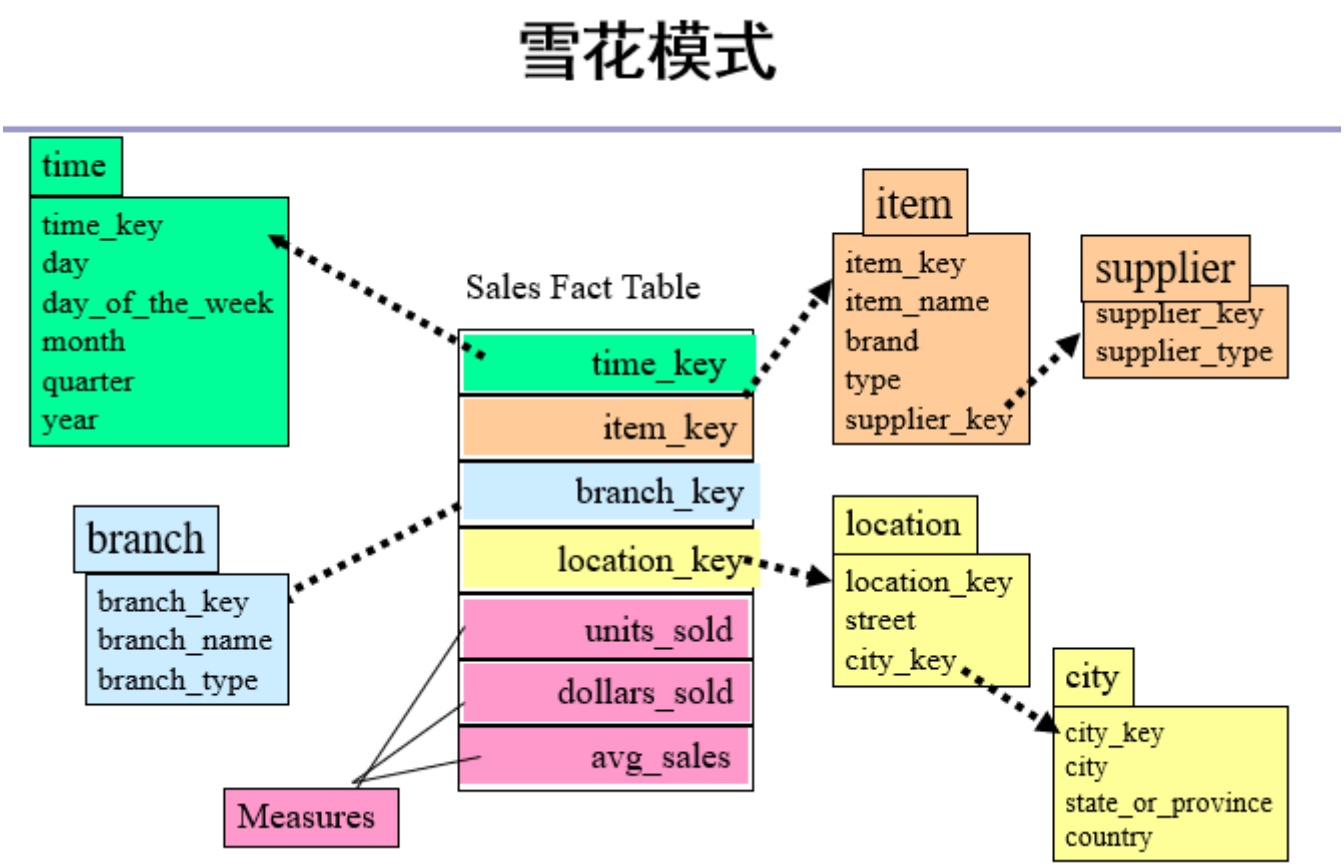
建模数据仓库: 维 & 度量

- 星型模式（Star schema）：一个事实表以及一组与事实表连结的维表。

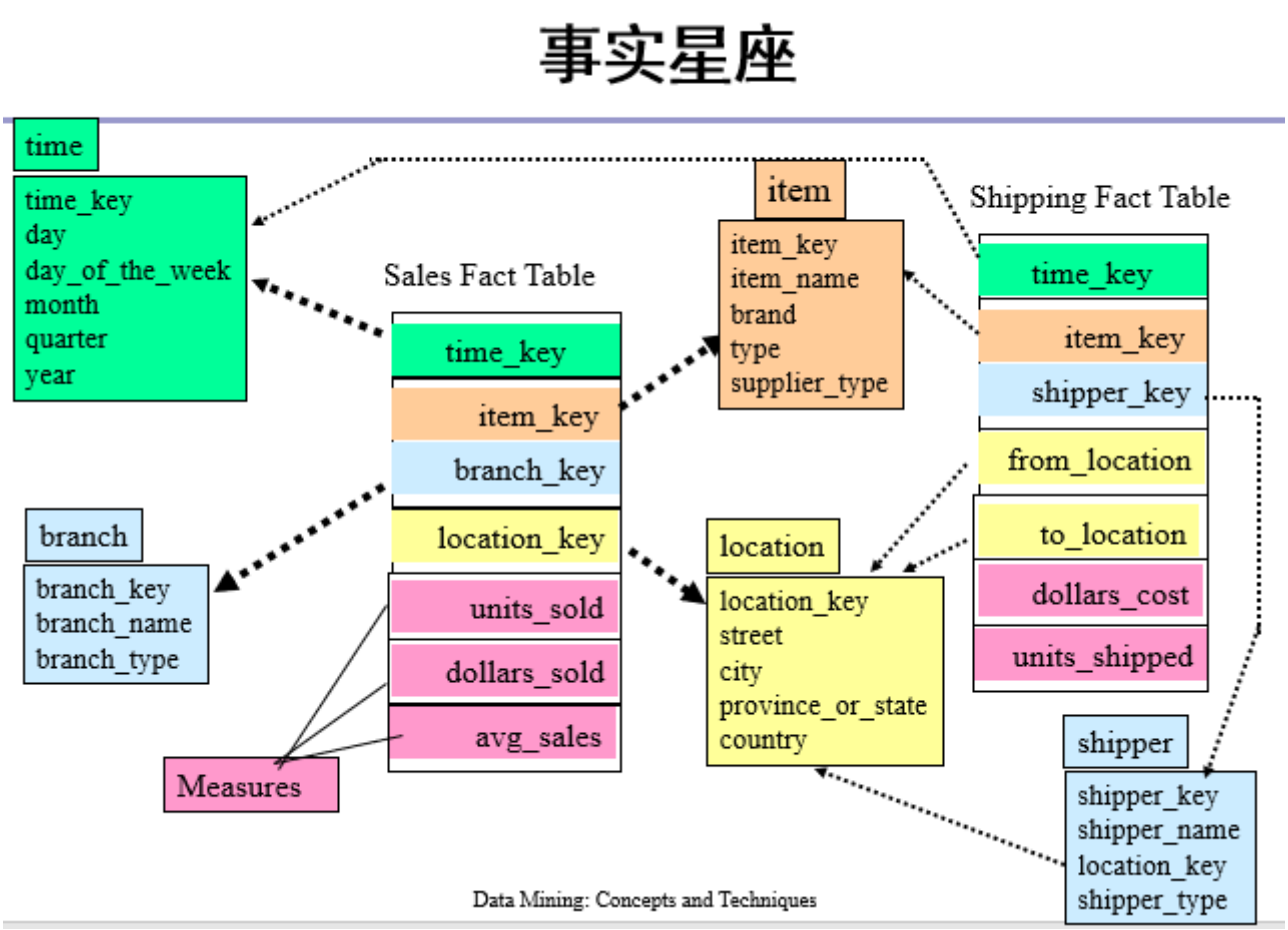


generated by haroopad

- 雪花模式（Snowflake schema）：雪花模式是星型模式的变种，其中某些维表是规范化的。（normalized），因而把数据进一步分解到附加的表中。



- 事实星座（Fact constellations）：多个事实表分享共同的维表，这种模式可以看作星型模式的集合，因此称为星系模式（galaxy schema）或事实星座。



度量的分类

- 分布式的（distributive）：一个聚集函数是分布的，如果它能以以下分布式进行计算:如果将函数用于n个聚集值得到的结果，与将函数用于所有数据得到的结果一样，则该函数可以用分布式计算。
  如, count(), sum(), min(), max().
- 代数的（algebraic）：一个函数是代数的，如果它能够由一个具有M个参数的代数函数计算（其中M是一个有界整数），而每个参数都可以用一个分布聚集函数得到。
  如, avg(), standard_deviation().
- 整体的（holistic）：如果描述它的子聚集所需的存储没有一个常数界，即不存在一个具有M个参数的代数函数进行这一计算（其中M是常数）。如, median()（中位数）, mode()（出现次数最多的数，众数）等。

常见的OLAP操作

- 上卷Roll up (上钻drill-up):
  通过一个维的概念分层向上攀升或通过维规约，在数据立方体上进行聚集。
- 下钻Drill down (roll down): 上卷的逆操作，它由不太详细的数据得到更详细的数据。可以通过沿维的概念分层向下或引入新的维实现。
- 切片Slice与切块dice：投影与选择。
- 转轴Pivot (rotate)：是一种目视操作，它转动数据的视角，提供数据的替代表示
- 其它操作：
  钻过drill across：执行涉及多个事实表的查询。

钻透drill through：使用SQL的机制，钻到数据立方的底层，到后端关系表。

数据仓库体系结构

数据仓库的多层结构