18. 联机分析处理 OLAP

OLTP：联机事务处理，包括对数据的操作、频繁的询问修改（每次涉及少部分元组），有实时数据。

OLAP：联机分析处理：需要从数据中找到趋势，经常需要用到复杂聚合函数；可能会查询大量数据而且运行时间长；不一定要有一个完全实时的数据，只要相对较新即可。

数据仓库：传统的数据库系统为 OLTP 询问而创立，我们需要新的结构为 OLAP 询问创立。于是 OLAP 询问创造 OLTP 数据库的一个独立拷贝，称为数据仓库。

- 分析可能会用到各种来源的数据，所以这些数据必须被加入数据仓库中。
- 一个更加常见的场景：
    - 分支存储数据库解决 OLTP 询问；
    - 分支存储数据库定期拷贝至中央数据仓库；
    - OLAP 使用中央数据仓库进行分析。

看待数据的多维视角：主要分为两类变量：

- 依赖属性：分析中需要的数值属性，如销量、价格。
- 维度属性：分析中可使用不同维度属性从不同角度看待数值，如时间、空间。

星型模式 (star schema)：数据仓库的一种常见组成。包括：

- 一个**事实表 (fact table)** 包含非常大量的事实值，比如销量等等，经常只支持插入；
- 多个**维度表 (dimension tables)**：比较小，通常是事实中含有的一些静态的信息；
- 事实表通过外键 (FK) 与维度表进行连接。

例：

```
// A fact table
Sales(timeID, itemID, storeID, price, qty);
// Dimension tables
Times(timeID, day, week, month, year);
Items(itemID, item, size, color, manf);
Stores(storeID, store, city, province);
```

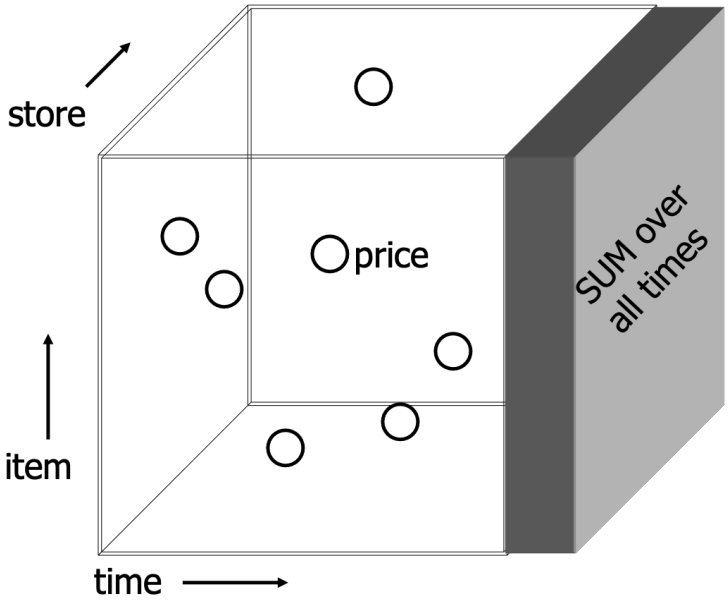
OLAP 实现

- 多维 OLAP (Multidimensional OLAP, MOLAP)：使用特殊的多维结构存储数据；
- 关系 OLAP (Relational OLAP, ROLAP)：使用关系以及星型模式存储数据；
- 混合 OLAP (Hybrid OLAP, HOLAP)：上述两种模式的结合。

数据立方体：OLAP 数据可以在多维空间以一下存储，每个维度属性可以看做超立方体的一个维度，而依赖属性存储在超立方体对应维度的区域中。

- **裸数据立方体** (raw-data cube)：只包含事实表中的原始数据。
- **正式数据立方体** (formal data cube)：同时包含代表裸数据聚合操作的点（聚合操作可以在任意维度子集上进行）。

- 需要提前计算聚合值；
- 对于分析询问的快速响应非常重要；
- 需要让聚合函数具有一定的实时性（及时更新，而不是实时更新）。



- 将每个维度“增加”一个额外的值 "*" 表示 "ALL"，即该维度上的聚合结果。
- 一个有若干坐标为 "*" 的点即相当于在这些坐标上进行聚合。

- **在 SQL 中建立数据立方体**：

```
SELECT time, item, store SUM(qty) FROM Sales
GROUP BY CUBE(time, item, store);
```

  - 相当于对三个维度共 2^3 子集进行聚合函数；
  - 使用空 `NULL` 代表 `*`。
  - 在 SQL 服务器中有时也写作：`GROUP BY ... WITH CUBE`。
- **存储数据立方体**

```
CREATE MATERIALIZED VIEW myCube AS
 ... cube generating statement ...;
```

- **数据立方体的变体**：`ROLLUP`。

```
SELECT time, item, store SUM(qty) FROM Sales
GROUP BY ROLLUP(time, item, store);
```

  - 相当于仅对于 `{}`, `{time}`, `{time, item}`, `{time, item, store}` 进行聚合。
  - 在 SQL 服务器中有时也写作：`GROUP BY ... WITH ROLLUP`。
- **数据立方体的操作**：
  - **切块** (dicing)：裸数据立方体在每个维度上被切分成一定粒度，将正方体切成“小块”。

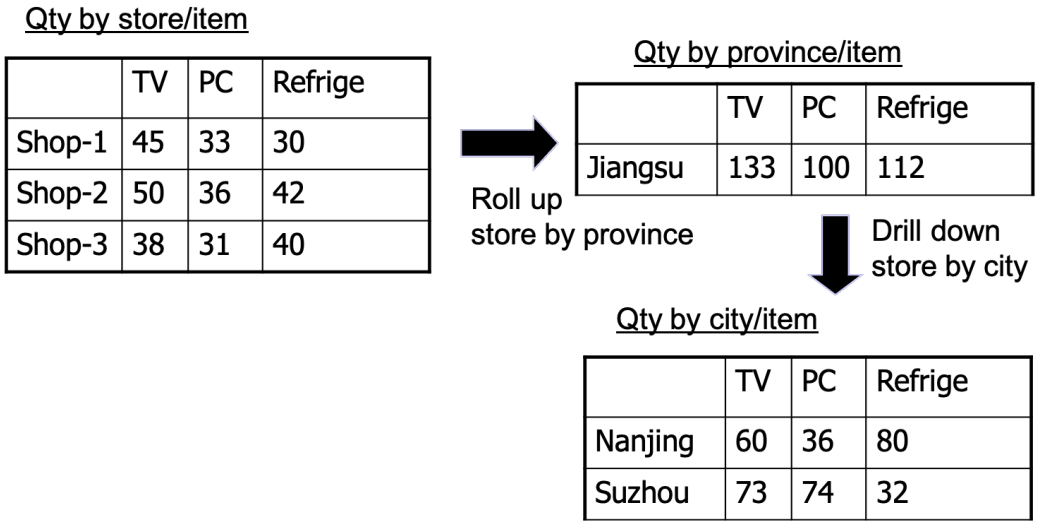
    切片 (slicing)：裸数据立方体在某个维度上被切分成一定粒度，将正方体切成“薄片”。

    例：

```
SELECT city, color, SUM(qty)
FROM (((Sales NATURAL JOIN Stores) NATURAL JOIN Items) NATURAL JOIN
Times)
WHERE year=2019
GROUP BY city, color;
```

    - 在 `city, color` 维度上进行切块；
    - 在时间维度进行切片。
- Roll-up：在一个或多个维度进行聚合，从细粒度到粗粒度，维度层级越来越高， `GROUP BY` 的维度越来越少（减维）。

  Drill-down：在一个或多个维度进行“逆聚合”，将聚合函数分解成其组成部分，从粗粒度到细粒度，维度层级越来越低，实际 `GROUP BY` 的维度越来越多（增维）。



	TV	PC	Refrige
Shop-1	45	33	30
Shop-2	50	36	42
Shop-3	38	31	40

	TV	PC	Refrige
Jiangsu	133	100	112

	TV	PC	Refrige
Nanjing	60	36	80
Suzhou	73	74	32