

ОСЛАДР-ТЕХНОЛОГИИ

Хранилища данных

Хранилище данных — это интегрированный накопитель информации, собранной из других систем, на основе которого строятся процессы принятия решений и анализа данных.

Ральф Кимбалл (*Ralph Kimball*), один из авторов концепции хранилищ данных, описывал хранилище данных как "место, где люди могут получить доступ к своим данным". Он же сформулировал и основные требования к хранилищам данных:

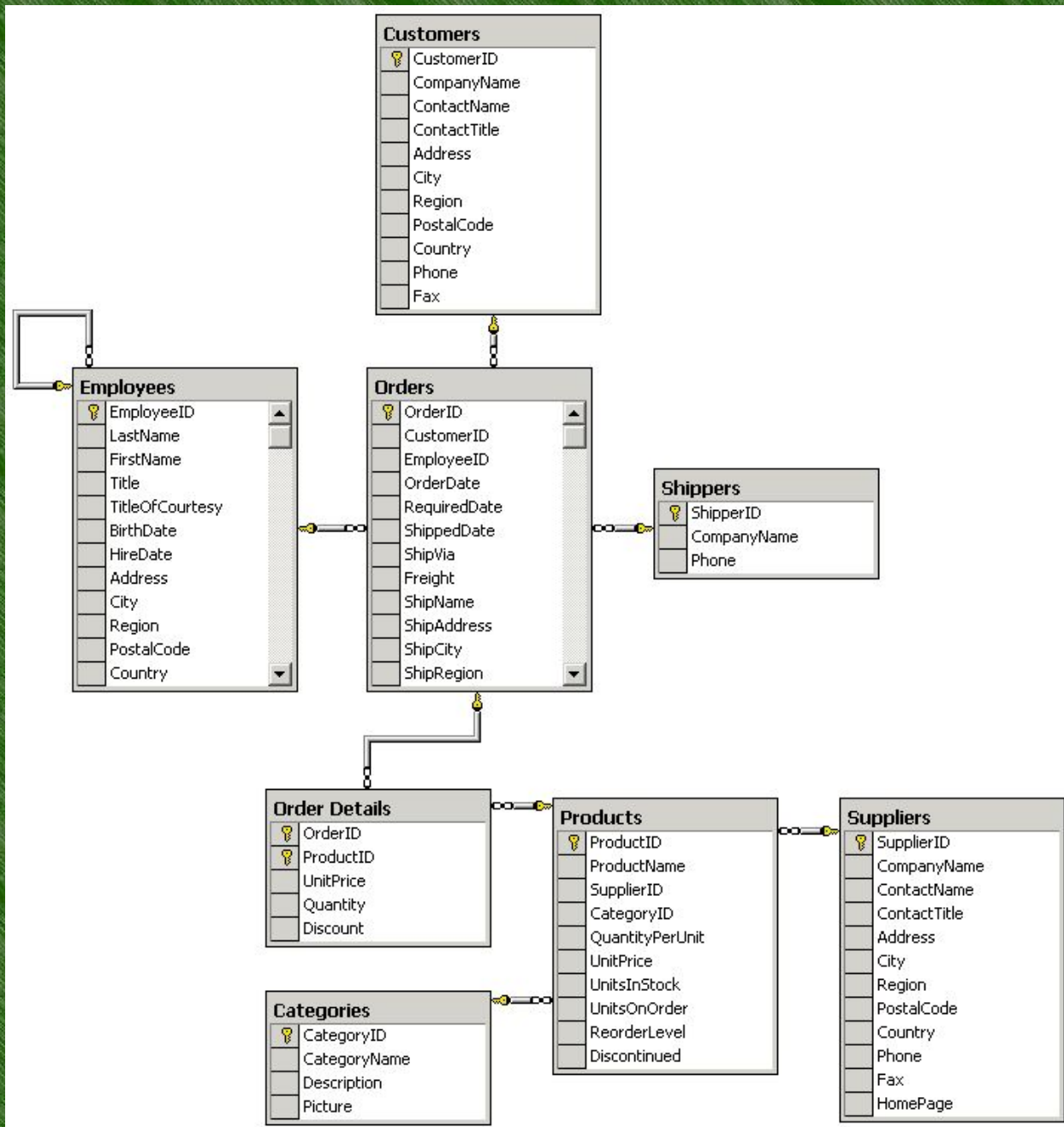
- поддержка высокой скорости получения данных из хранилища;
- поддержка внутренней непротиворечивости данных;
- возможность получения и сравнения так называемых срезов данных (*slice and dice*);
- наличие удобных утилит просмотра данных в хранилище;
- полнота и достоверность хранимых данных;



Оперативные данные собираются из различных источников, очищаются, интегрируются и складываются в реляционное хранилище. При этом они уже доступны для анализа при помощи различных средств построения отчетов. Затем данные (полностью или частично) подготавливаются для *OLAP*-анализа. Они могут быть загружены в специальную БД *OLAP* или оставлены в реляционном хранилище. Важнейшим его элементом являются метаданные, т. е. информация о структуре, размещении и трансформации данных. Благодаря

Типичная структура хранилища данных существенно отличается от структуры обычной реляционной СУБД. Как правило, эта структура денормализована (это позволяет повысить скорость выполнения запросов), поэтому может допускать избыточность данных.

Для дальнейших примеров мы снова воспользуемся базой данных *Northwind*, входящей в комплекты поставки *Microsoft SQL Server* и *Microsoft Access*. Ее структура данных приведена на далее.



Основными составляющими структуры хранилищ данных являются таблица фактов (*fact table*) и таблицы измерений (*dimension tables*).

Таблица фактов

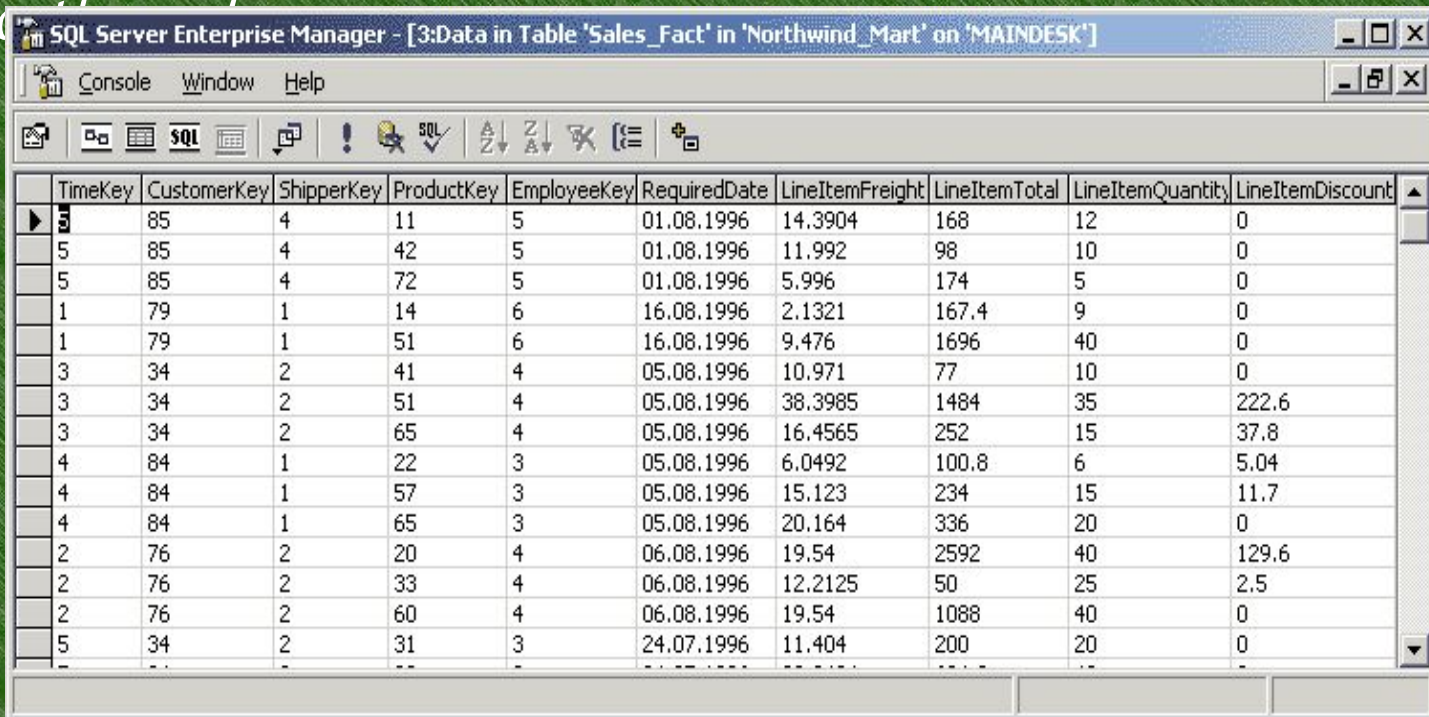
Таблица фактов является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться. Обычно говорят о четырех наиболее часто встречающихся типах фактов. К ним относятся:

- факты, связанные с транзакциями (*Transaction facts*). Они основаны на отдельных событиях (типичными примерами которых являются

- факты, связанные с «моментальными снимками» (*Snapshot facts*). Основаны на состоянии объекта (например, банковского счета) в определенные моменты времени, например на конец дня или месяца. Типичными примерами таких фактов являются объем продаж за день или дневная выручка;
- факты, связанные с элементами документа (*Line-item facts*). Основаны на том или ином документе (например, счете за товар или услуги) и содержат подробную информацию об элементах этого документа (например, количестве, цене, проценте скидки);
- факты, связанные с событиями или состоянием объекта (*Event or state facts*). Представляют

Пример таблицы фактов, которая может быть построена на основе базы данных

Northwind



The screenshot shows a window titled "SQL Server Enterprise Manager - [3:Data in Table 'Sales_Fact' in 'Northwind_Mart' on 'MAINDESK']". The window displays a table with the following columns: TimeKey, CustomerKey, ShipperKey, ProductKey, EmployeeKey, RequiredDate, LineItemFreight, LineItemTotal, LineItemQuantity, and LineItemDiscount. The table contains 15 rows of data.

| TimeKey | CustomerKey | ShipperKey | ProductKey | EmployeeKey | RequiredDate | LineItemFreight | LineItemTotal | LineItemQuantity | LineItemDiscount |
|---------|-------------|------------|------------|-------------|--------------|-----------------|---------------|------------------|------------------|
| 5 | 85 | 4 | 11 | 5 | 01.08.1996 | 14.3904 | 168 | 12 | 0 |
| 5 | 85 | 4 | 42 | 5 | 01.08.1996 | 11.992 | 98 | 10 | 0 |
| 5 | 85 | 4 | 72 | 5 | 01.08.1996 | 5.996 | 174 | 5 | 0 |
| 1 | 79 | 1 | 14 | 6 | 16.08.1996 | 2.1321 | 167.4 | 9 | 0 |
| 1 | 79 | 1 | 51 | 6 | 16.08.1996 | 9.476 | 1696 | 40 | 0 |
| 3 | 34 | 2 | 41 | 4 | 05.08.1996 | 10.971 | 77 | 10 | 0 |
| 3 | 34 | 2 | 51 | 4 | 05.08.1996 | 38.3985 | 1484 | 35 | 222.6 |
| 3 | 34 | 2 | 65 | 4 | 05.08.1996 | 16.4565 | 252 | 15 | 37.8 |
| 4 | 84 | 1 | 22 | 3 | 05.08.1996 | 6.0492 | 100.8 | 6 | 5.04 |
| 4 | 84 | 1 | 57 | 3 | 05.08.1996 | 15.123 | 234 | 15 | 11.7 |
| 4 | 84 | 1 | 65 | 3 | 05.08.1996 | 20.164 | 336 | 20 | 0 |
| 2 | 76 | 2 | 20 | 4 | 06.08.1996 | 19.54 | 2592 | 40 | 129.6 |
| 2 | 76 | 2 | 33 | 4 | 06.08.1996 | 12.2125 | 50 | 25 | 2.5 |
| 2 | 76 | 2 | 60 | 4 | 06.08.1996 | 19.54 | 1088 | 40 | 0 |
| 5 | 34 | 2 | 31 | 3 | 24.07.1996 | 11.404 | 200 | 20 | 0 |

Sales_Fact

- TimeKey
- CustomerKey
- ShipperKey
- ProductKey
- EmployeeKey
- RequiredDate
- LineItemFreight
- LineItemTotal
- LineItemQuantity
- LineItemDiscount

В данном примере измерениям будущего куба соответствуют первые шесть полей, а агрегатным данным — последние четыре.

Отметим, что для многомерного анализа пригодны таблицы фактов, содержащие как можно более подробные данные (то есть соответствующие членам нижних уровней иерархии соответствующих измерений). В данном случае предпочтительнее взять за основу факты продажи товаров отдельным заказчикам, а не суммы продаж для разных стран — последние все равно будут вычислены *OLAP*-средством.

Таблицы измерений

Таблицы измерений содержат неизменяемые либо редко изменяемые данные. В подавляющем большинстве случаев эти данные представляют собой по одной записи для каждого члена нижнего уровня иерархии в измерении. Таблицы измерений также содержат как минимум одно описательное поле (обычно с именем члена измерения) и, как правило, целочисленное ключевое поле (обычно это суррогатный ключ)

Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов.

Отметим, что скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов.

SQL Server Enterprise Manager - [4:Data in Table 'Product_Dim' in 'Northwind_Mart' on 'MAINDESK']

Console Window Help

SQL

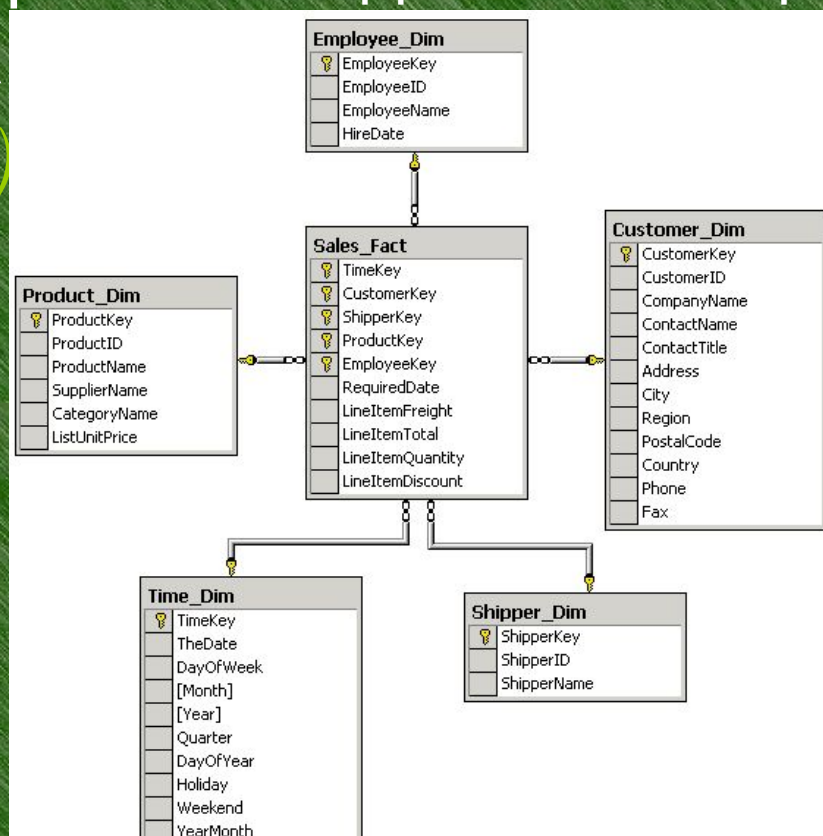
| | ProductKey | ProductID | ProductName | SupplierName | CategoryName | ListUnitPrice |
|---|------------|-----------|---------------------------------|------------------------------------|----------------|---------------|
| ▶ | 1 | 1 | Chai | Exotic Liquids | Beverages | 18 |
| | 2 | 2 | Chang | Exotic Liquids | Beverages | 19 |
| | 3 | 3 | Aniseed Syrup | Exotic Liquids | Condiments | 10 |
| | 4 | 4 | Chef Anton's Cajun Seasoning | New Orleans Cajun Delights | Condiments | 22 |
| | 5 | 5 | Chef Anton's Gumbo Mix | New Orleans Cajun Delights | Condiments | 21.35 |
| | 6 | 6 | Grandma's Boysenberry Spread | Grandma Kelly's Homestead | Condiments | 25 |
| | 7 | 7 | Uncle Bob's Organic Dried Pears | Grandma Kelly's Homestead | Produce | 30 |
| | 8 | 8 | Northwoods Cranberry Sauce | Grandma Kelly's Homestead | Condiments | 40 |
| | 9 | 9 | Mishi Kobe Niku | Tokyo Traders | Meat/Poultry | 97 |
| | 10 | 10 | Ikura | Tokyo Traders | Seafood | 31 |
| | 11 | 11 | Queso Cabrales | Cooperativa de Quesos 'Las Cabras' | Dairy Products | 21 |
| | 12 | 12 | Queso Manchego La Pastora | Cooperativa de Quesos 'Las Cabras' | Dairy Products | 38 |
| | 13 | 13 | Konbu | Mayumi's | Seafood | 6 |
| | 14 | 14 | Tofu | Mayumi's | Produce | 23.25 |
| | 15 | 15 | Genen Shouyu | Mayumi's | Condiments | 15.5 |
| | 16 | 16 | Pavlova | Pavlova, Ltd. | Confections | 17.45 |
| | 17 | 17 | Alice Mutton | Pavlova, Ltd. | Meat/Poultry | 39 |
| | 18 | 18 | Carnarvon Tigers | Pavlova, Ltd. | Seafood | 62.5 |
| | 19 | 19 | Teatime Chocolate Biscuits | Specialty Biscuits, Ltd. | Confections | 9.2 |

Product_Dim

| | |
|---|---------------|
| 🔑 | ProductKey |
| | ProductID |
| | ProductName |
| | SupplierName |
| | CategoryName |
| | ListUnitPrice |

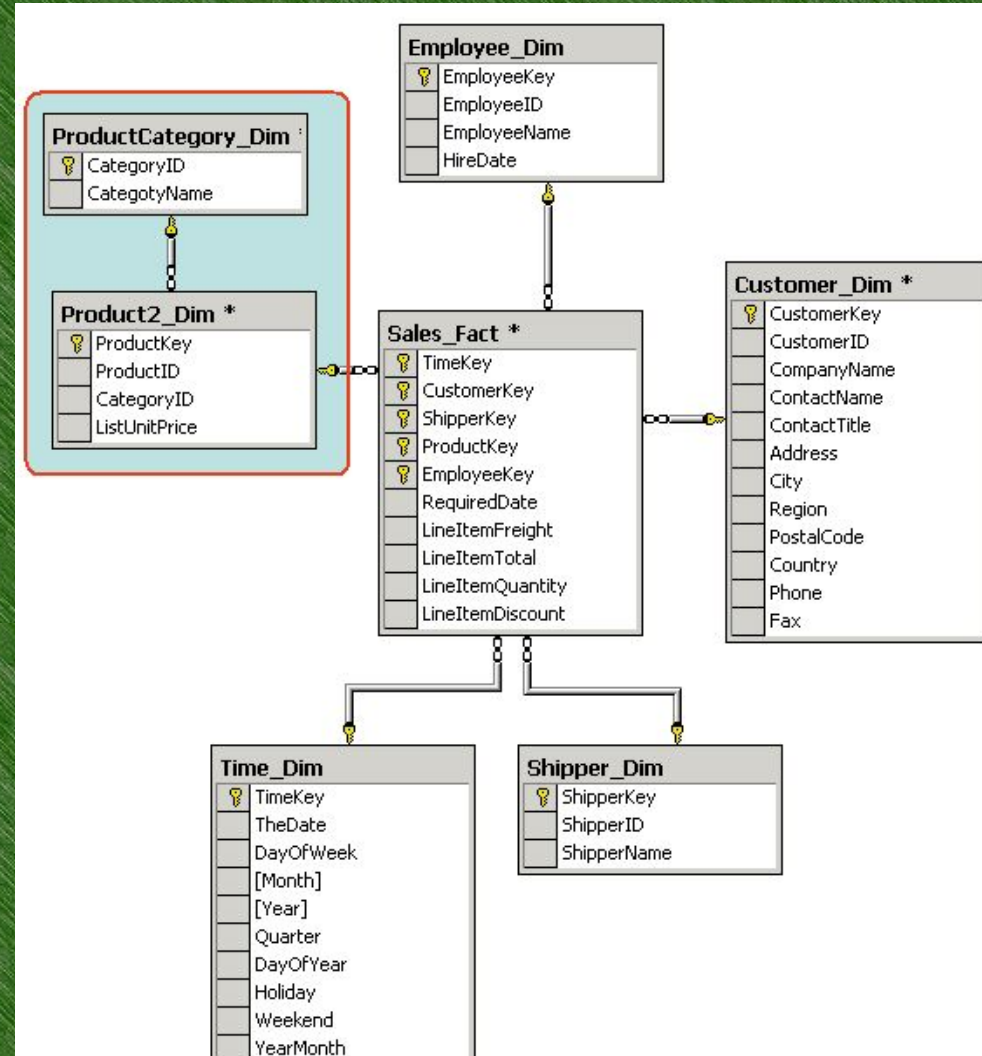
таблица измерений

Одно измерение куба может содержаться как в одной таблице (в том числе и при наличии нескольких уровней иерархии), так и в нескольких связанных таблицах, соответствующих различным уровням иерархии в измерении. Если каждое измерение содержится в одной таблице, такая схема хранилища «звезда» (*star schema*)



Если же хотя бы одно измерение содержится в нескольких связанных таблицах, такая схема хранилища данных носит название «СНЕЖИНКА»

(snowflake schema).
Дополнительные таблицы измерений в такой схеме, обычно соответствующие верхним уровням иерархии измерения и находящиеся в соотношении «один ко многим» в главной таблице измерений, соответствующей нижнему уровню



Кубы

ОЛАР предоставляет удобные быстродействующие средства доступа, просмотра и анализа деловой информации. Пользователь получает естественную, интуитивно понятную модель данных, организуя их в виде многомерных кубов (*Cubes*). Осями многомерной системы координат служат основные бизнес-показатели.



| | Март | Февраль | Январь |
|------------------|--------|---------|---------|
| | США | Канада | Мексика |
| Напитки | 10 000 | 2000 | 1 000 |
| Продукты питания | 5000 | 500 | 250 |
| Прочие товары | 5000 | 500 | 250 |

Спасибо за внимание!