Data Warehousing and Dimensional Modelling

Stéphane Bressan

Online transaction processing applications are characterised by many short transactions involving updates and mostly point queries.

- Update account balance
- Enroll in course
- Add book to shopping cart

Queries touch small amounts of data (one record or a few records). Updates are frequent. Data must be up-to-date and consistent at all times. Concurrency is the biggest performance concern.

"The users of an operational system turn the wheels of the organization. They take orders, sign up new customers, and log complaints. Users of an operational system almost always deal with one record at a time. They repeatedly perform the same operational tasks over and over."
The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling
by Ralph Kimball and Margy Ross

Online analytical processing applications are characterised by long transactions involving complex queries.

- Report total sales for each department in each month
- Identify top-selling books
- Count classes with fewer than 10 students

Queries touch large amounts of data. Updates are infrequent (only at the beginning). Individual queries can require lots of resources. Operating on static snapshots of data my be acceptable. Approximate answers may also be acceptable.

"The users of a data warehouse, on the other hand, watch the wheels of the organization turn. They count the new orders and compare them with last week's orders and ask why the new customers signed up and what the customers complained about. Users of a data warehouse almost never deal with one row at a time. Rather, their questions often require that hundreds or thousands of rows be searched and compressed into an answer set. To further complicate matters, users of a data warehouse continuously change the kinds of questions they ask."
The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling
by Ralph Kimball and Margy Ross

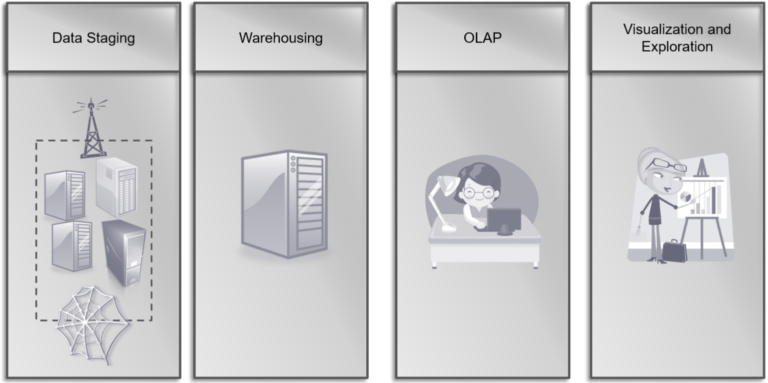
Doing OLTP and OLAP in the same database system is often impractical.

For example, an analyst asks a query that calculates the sum of all sales, then the query acquires locks on the sales table for consistency and new sales transactions are blocked.

The solution is to build a dedicated data warehouse.

- Copy data from various OLTP systems and streamline data cleaning and refreshing by using Extract, Transform, Load (ETL) tools;
- Simplify the design for OLAP by using dimensional modelling;
- Optimise the data organisation and tune the database management or use a dedicated system for OLAP;
- Leverage the data organisation to create user friendly visualisations by using reporting, visualisation and interactive exploration tools.

Data Warehouse



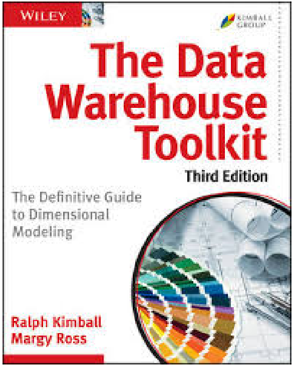
- Extract-Transform-Load
  - IBM InfoSphere DataStage
  - Oracle Warehouse Builder
  - SQL Server Integration Services
  - Pentaho Kettle
- Data Warehouse Management System
  - Oracle, IBM DB2, Microsoft SQL Server, PostgreSQL
  - Teradata
  - SAP HANNA
- Online analytical processing (building summaries) and Exploratory data analysis (looking for patterns)
  - Microsoft Excel andPower BI
  - Oracle Reports
  - Cognos
  - Tableau
  - Business Intelligence and Reporting Tools (BIRT)
  - Data mining tools
  - Machine learning algorithms

"A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. The data warehouse contains granular corporate data."
Building the Data Warehouse, 1992
by Bill Inmon

A data warehouse is one part of the overall business intelligence system. An enterprise has one data warehouse, and data marts source their information from the data warehouse. In the data warehouse, information is stored in third normal form.

"A data warehouse is a copy of transaction data specifically structured for query and analysis."
"In its most simplistic form, a data mart presents the data from a single business process. These business processes cross the boundaries of organizational functions."
The Data Warehouse Toolkit, 1996
by Ralph Kimball

A data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model.



The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling
by Ralph Kimball and Margy Ross

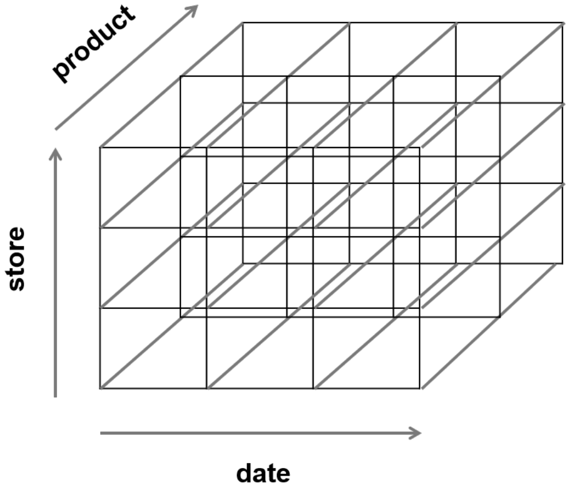
Acording to Kimball, the data warehouse must

- make an organisation's information easily accessible.
- present the organisation's information consistently.
- be adaptive and resilient to change.
- be a secure bastion that protects our information assets
- serve as the foundation for improved decision making.
- be accepted by the business community.

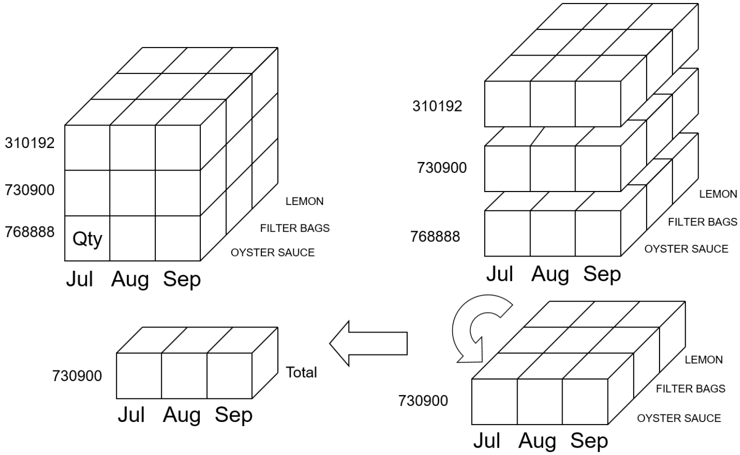
Simple Retail Sales Case

product	date	store	quantity
FAIRPRICE PREMIUM OYSTER SAUCE	12/03/13	Toa Payoh Lorong 4 Blk 192	12
FAIRPRICE PREMIUM OYSTER SAUCE	12/03/13	900 South Woodlands Drive	11
BONCAFE FILTERS BAGS - NATURAL	13/03/13	Toa Payoh Lorong 4 Blk 192	34
BONCAFE FILTERS BAGS - NATURAL	13/03/13	Toa Payoh Lorong 4 Blk 192	3
BONCAFE FILTERS BAGS - NATURAL	13/03/13	Toa Payoh Lorong 4 Blk 192	3
CLOROX BLEACH - LEMON	13/03/13	Yishun Ave 9, Blk 10	4
FAIRPRICE PREMIUM OYSTER SAUCE	13/03/13	Yishun Ave 9, Blk 10	12
...			

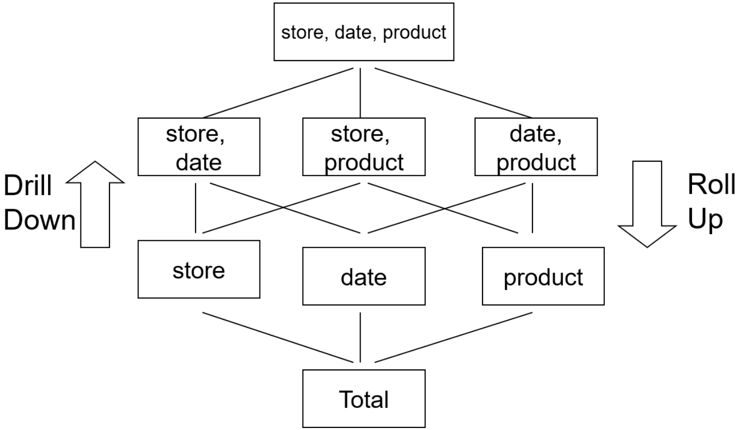
Cube



Slice and Dice



Roll Up and Drill Down

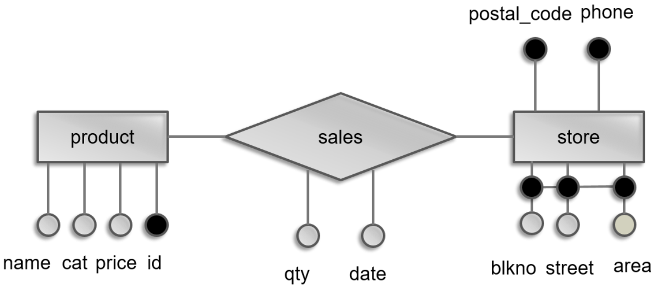


We now think the case of the retail business process. The transactions recorded in the data warehouse are the individual or aggregated sales (e.g. the line items on a point-of-sale receipt).

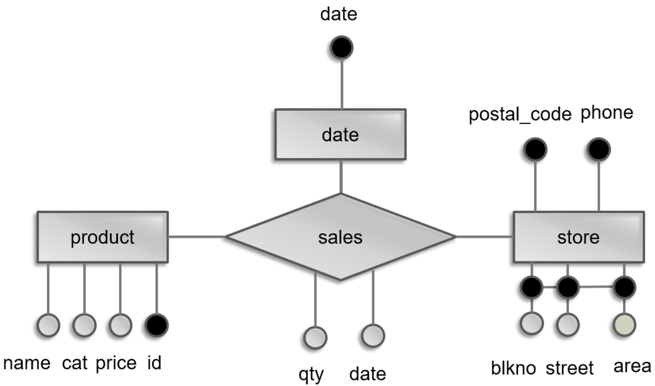
The Data

product	date	store	quantity
FAIRPRICE PREMIUM OYSTER SAUCE	12/03/13	Toa Payoh Lorong 4 Blk 192	12
FAIRPRICE PREMIUM OYSTER SAUCE	12/03/13	900 South Woodlands Drive	11
BONCAFE FILTERS BAGS - NATURAL	13/03/13	Toa Payoh Lorong 4 Blk 192	34
BONCAFE FILTERS BAGS - NATURAL	13/03/13	Toa Payoh Lorong 4 Blk 192	3
BONCAFE FILTERS BAGS - NATURAL	13/03/13	Toa Payoh Lorong 4 Blk 192	3
CLOROX BLEACH - LEMON	13/03/13	Yishun Ave 9, Blk 10	4
FAIRPRICE PREMIUM OYSTER SAUCE	13/03/13	Yishun Ave 9, Blk 10	12
...			

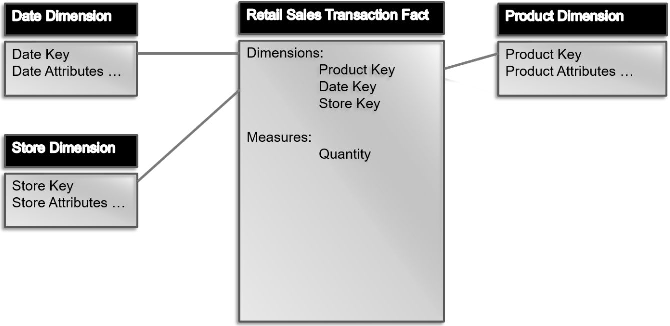
Entity-relationship Diagram



Entity-relationship Diagram



We design the database as a star schema.



A star schema is composed of one fact table and several dimension tables.

The fact table records the transactions of the business process at the finest available granularity (e.g. one line of the point-of-sale receipt).

product	date	store	quantity
1	1	1	12
1	1	2	11
2	2	1	34
2	2	1	3
2	2	1	3
3	2	3	4
2	2	3	12
...			

The fact table records, for each transaction of the business process, its measures (also sometimes called facts) (e.g quantity, total price, etc.) and the surrogate keys of the dimension rows in the different dimension tables that describe the transaction (e.g. the product, the date, the store etc.).

The dimension tables provide as comprehensive as possible a description of the dimensions for the sake of analysis. Each entry in a dimension table has a surrogate key used in the fact table to refer to it (a kind of foreign key).

product	sku	name	category	price
1	261721	FAIRPRICE PREMIUM OYSTER SAUCE	Groceries	2.6
2	263789	BONCAFE FILTERS BAGS - WHITE	Beverages	2.3
3	265147	CLOROX BLEACH - LEMON	Household Items	4.25
...				

For instance, the product dimension records for each product, its name, its stock keeping unit number, its category, its unit price etc. The product dimension table may also have rows for special (e.g. products without SKU) and unidentified products (better than null values).

store	postal code	Toa Payoh Lorong 4	blkno	area	phone
1	310192	Toa Payoh Lorong 4	Blk 192	TOA PAYOH	62508019
2	730900	South Woodlands Drive	NO.900	WOODLANDS	64582558
3	768888	Yishun Ave 9	Blk 10	YISHUN	67665009
. . .					

For instance, the store dimension records for each store, its name, the details of its location and address, its telephone number etc. The store dimension table may also have rows for unidentified stores.

date	date_actual	day_name	month_actual	...
1	12/03/13	Tuesday	3	...
2	12/03/13	Wednesday	3	...
...				

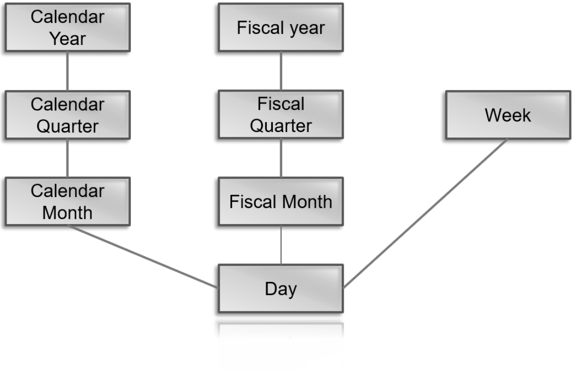
For instance, the date dimension records for each date, all the fields that can be used for analysis: the date in different formats, the day of the week, the position in the calendar year, the fiscal year and the academic year,a as useful for analysis, whether is is a public holiday in a given country etc. The day dimension table may also have rows for special and unidentified dates.

The Date Dimension

```
CREATE TABLE date(
  date                      INT NOT NULL,
  date_actual               DATE NOT NULL,
  epoch                     BIGINT NOT NULL,
  day_suffix                VARCHAR(4) NOT NULL,
  day_name                  VARCHAR(9) NOT NULL,
  day_of_week               INT NOT NULL,
  day_of_month              INT NOT NULL,
  day_of_quarter            INT NOT NULL,
  day_of_year               INT NOT NULL,
  week_of_month             INT NOT NULL,
  week_of_year              INT NOT NULL,
  week_of_year_iso          CHAR(10) NOT NULL,
  month_actual              INT NOT NULL,
  month_name                VARCHAR(9) NOT NULL,
  month_name_abbreviated    CHAR(3) NOT NULL,

  quarter_actual            INT NOT NULL,
  quarter_name              VARCHAR(9) NOT NULL,
  year_actual               INT NOT NULL,
  first_day_of_week         DATE NOT NULL,
  last_day_of_week          DATE NOT NULL,
  first_day_of_month        DATE NOT NULL,
  last_day_of_month         DATE NOT NULL,
  first_day_of_quarter      DATE NOT NULL,
  last_day_of_quarter       DATE NOT NULL,
  first_day_of_year         DATE NOT NULL,
  last_day_of_year          DATE NOT NULL,
  mmyyyy                    CHAR(6) NOT NULL,
  mmddyyyy                  CHAR(10) NOT NULL,
  weekend_indr              BOOLEAN NOT NULL);
```

Different fields of a dimension define some hierarchies.



For instance the year, quarter, month, week, day hierarchy in calendar, fiscal and academic years.

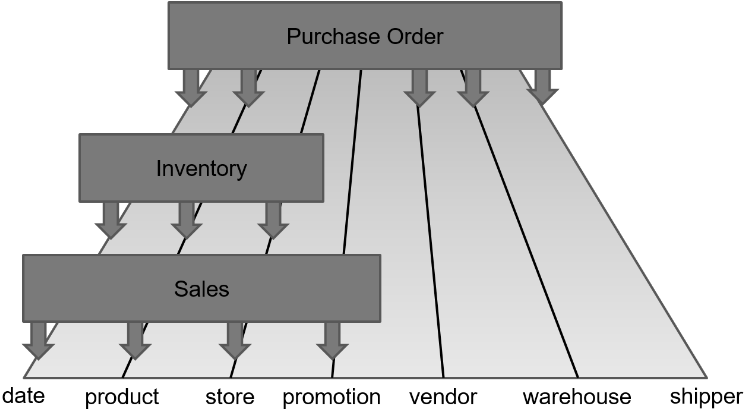
A dimension table may play several roles in a star schema.

For instance the date dimension can be used to indicate the order, delivery and payment dates of an order in a inventory star schema.

It is generally not preferred to further expand a star schema into a snowflake schema.

Information can be repeated in the dimensions for the sake of simplicity and efficiency. Storage is rarely an issue (views can be used if necessary). Normalisation is not a concern because the data should have been cleaned and there is no update.

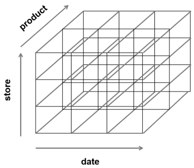
A data warehouse may contain several fact table that share common dimensions.



Data Warehouse Queries

Most data warehouse queries calculate aggregate functions on the natural join of the fact table with the dimension tables to which we add conditions on certain dimensions and that are grouped according to certain dimensions.

```
SELECT <aggregation of measures>
FROM sales
    NATURAL JOIN date
    NATURAL JOIN product
    NATURAL JOIN store
WHERE
    <conditions on the dimensions>
GROUP BY
    <dimensions>
```

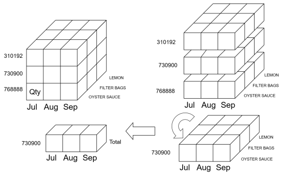


Queries may involve

- Complex Boolean conditions
- Grouping, partitioning and sorting
- Aggregate and window functions
- Statistical functions
- Time series functions
- Spatial data functions

CUBE is a subclause of the GROUP BY clause that defines multiple simultaneous grouping sets. CUBE generates all possible grouping sets of the input columns.

For instance, a clause `CUBE (A, B, C)` is equivalent to the (ordered) union of the query with no GROUP BY clause and the seven queries with the the clauses `GROUP BY A`, `GROUP BY B`, `GROUP BY C`, `GROUP BY A, B`, `GROUP BY A, C`, `GROUP BY B, C` and `GROUP BY A, B, C`, in that order.



Cube

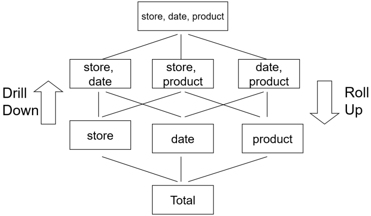
```sql
SELECT  p.category, st.area, ROUND(AVG(s.quantity*p.price),2) AS avgvolume
FROM product p, sales s, store st
WHERE p.product=s.product AND s.store=st.store
GROUP BY CUBE(p.category, st.area)
```

category	area	avgqty
		10223.97
	"YISHUN"	8249.97
"Toiletries"		6203.28
Groceries	"JURONG EAST"	3787.16
"Household Items"	"BISHAN"	5178.07
...		

In the example above, CUBE calculates the average sales overall, by area, by category, and by category and area.

`ROLLUP` is a subclause of the `GROUP BY` clause that defines multiple simultaneous grouping sets. `ROLLUP` generates all grouping sets considering the order of the input columns.

For instance, a clause `GROUP BY ROLLUP (A, B, C)` is equivalent to the (ordered) union of the query with no GROUP BY clause and the three queries with the the clauses `GROUP BY A`, `GROUP BY A, B` and `GROUP BY A, B, C`, in that order.



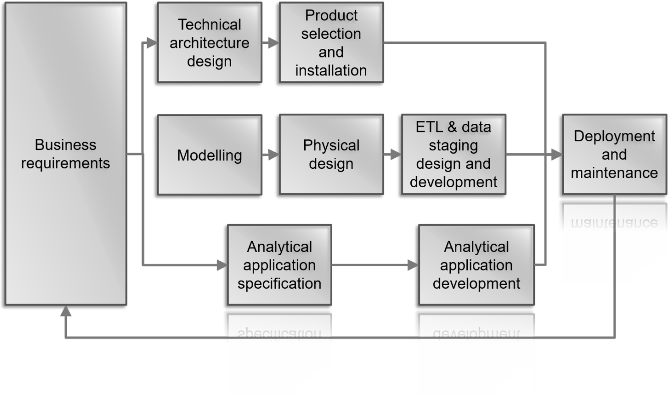
Rollup

```sql
SELECT  p.category, st.area, ROUND(AVG(s.quantity*p.price),2) AS avgvolume
FROM product p, sales s, store st
WHERE p.product=s.product AND s.store=st.store
GROUP BY ROLLUP(p.category, st.area);
```

category	area	avgqty
		10223.97
"Toiletries"		6203.28
"Household Items"	"YISHUN"	4110.39
Groceries	"JURONG EAST"	3787.16
"Household Items"	"BISHAN"	5178.07
...		

In the example above, `ROLLUP` calculates the average sales overall, by category and by category and area.

In conclusion, let us review the development and life cycle of a data warehouse.



Copyright 2022 Stéphane Bressan. All rights reserved.