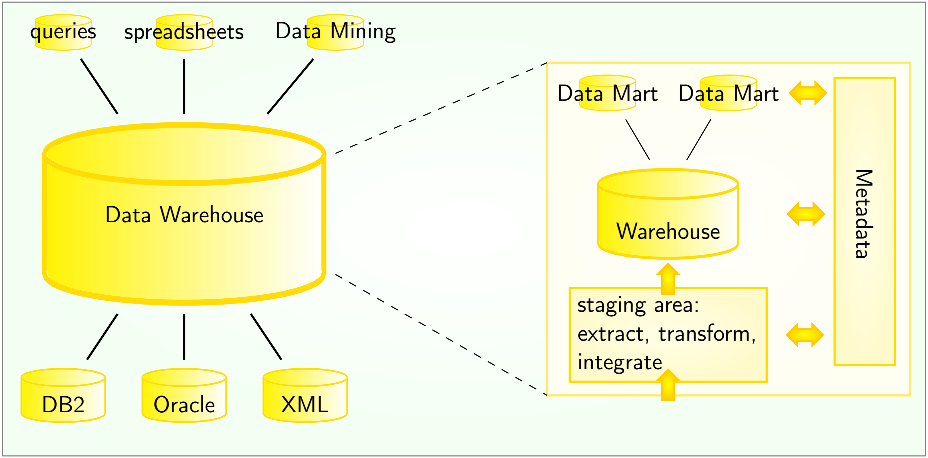
Conceptual design for DW: the multidimensional model

Chapter content:

- Components overview
- Multidimensional data model: facts,measures, dimensions,cube
- Multi-dimensional query concepts: OLAP operations

Table des matières

DW architecture



Birds' view on DW architecture

Table des matières

Why a new model?

- 3NF relational schema too complex for BI queries
- ... and suffer(ed?) from slow query performance

Multi-dimensional model

- is easy to understand for Business users
  (OLAP exploration, generalizes spreadsheets. . .)
- delivers fast query performance
- schemas will not need to be reorganized too much over time

From Spreadsheet to the Cube

Köln

	Beverages	Produce	Condiments	Seafood
Q1	24	18	28	14
Q2				
Q3				
Q4				

Berlin

	Beverages	Produce	Condiments	Seafood
Q1	33	25	23	25
Q2				
Q3				
Q4				

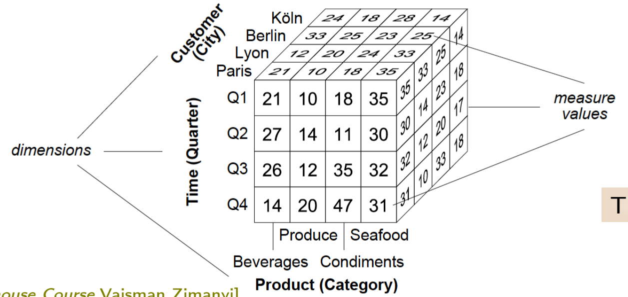
Lyon

	Beverages	Produce	Condiments	Seafood
Q1	12	20	24	33
Q2				
Q3				
Q4				

Paris

	Beverages	Produce	Condiments	Seafood
Q1	21	10	18	35
Q2	27	14	11	30
Q3	26	12	35	32
Q4	14	20	47	31

3^{rd} dimension = location



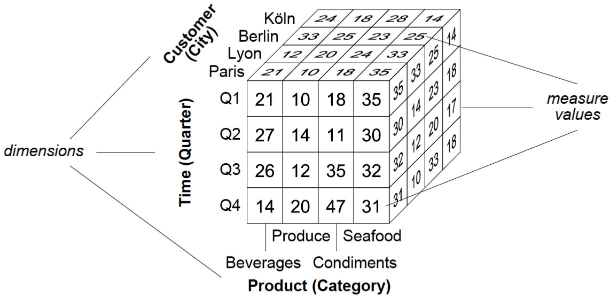
The *Data Cube*

The multidimensional model

Data viewed as n dimensional cube (here $n = 3$).

Associated to the cube are:

- Dimension: perspective used to analyze the data
- Cell: the intersection of dimension values
- Fact: non-empty cell
- Measure: numeric value of the cells



Facts

- concept relevant for the analysis: represents a measurement event (typically models a set of events taking place in the company)
- non-empty cell in the cube
- has a granularity = level of detail
- determined by the value of its dimension coordinates

Facts

Transaction fact

- Represent physical world events (at most detailed level)
- Exactly one fact per event
- Typically, events are independent and will occur at any time

Transaction fact example

- One fact for each sale of a wine bottle (detail)
- One fact for each day where at least one bottle sold (aggregation)
- sale fact tied to specific time/place

Snapshot fact

- Measures the current state of a process (possibly the result of a serie of events)
- Generally not independent
- Evaluated at specific interval/time

Snapshot fact example

- Inventory level per product,month,store
- a same product may appear in several inventories snapshots

Dimensions

- analysis perspective
- axis of the cube
- described by attributes

Dimensions can be used to define more than one cube.

Dimension attributes form a hierarchy of subsets (containment hierarchy): each level describes one degree of detail for the analysis on this dimension.

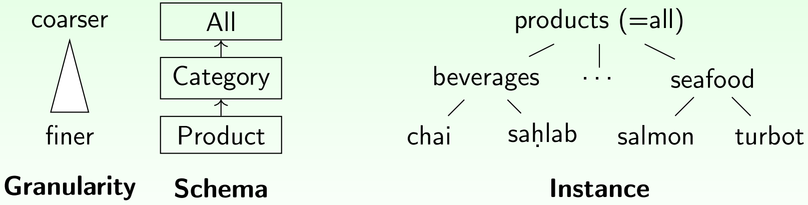
Dimensions

Dimension Schema

The classification schema of a dimension is a partially ordered set of category attributes: $(D.k_1, \ldots, D.k_n, \mathrm{Top}_D, \rightarrow)$.

- $\rightarrow$ is the functional dependancy
- Top_D is the maximal attribute regarding $\rightarrow$: $\forall i, D.k_i \rightarrow \mathrm{Top}_D$
- there exists a (unique) minimal attribute: $\exists i, \forall j \neq i, D.k_i \rightarrow D.k_j$. $D.k_i$ describes the finest granularity of the dimension.

Classification schema and instance of Product dimension



Dimensions

			Beverages	Produce	Condiments	Seafood
2010	Q1	Jan	4	5	0	6
		Feb	16	3	28	4
		Mar	4	10	0	8
	Q2	Apr	5	5	5	4
		May	12	5	5	10
		Jun	13	5	5	4
	Q3	Jul	20	10	15	7
		Aug	2	2	5	6
		Sept	3	3	10	4
	Q4	Oct	8	10	5	2
		Nov	3	10	20	8
		Dec	4	5	20	8
2011	Q1	Jan	10	25	40	10
		...				

Hierarchy (schema and instance) of temporal dimension?

Dimensions

The definition allows parallel paths in the schema (*parallel hierarchies*)

Classification schema of temporal dimension with (year,month,week,day)?

For this course we assume the classification schema has only one path.
We will also assume that items in distinct dimensions are independant (rules out $D.k_i \rightarrow D'.k_j$).

Measures

- describes a fact
- consists of two functions:
  - **numeric** property for each fact
  - function to compute measure at coarser aggregation levels

Several Measures can be associated to a fact.

A measure can depend on (other) measures on other facts.

Measure examples:

- number of units in stock
- value per unit
- quantity $\times$ price $\times$ turnover
- ...

Cube

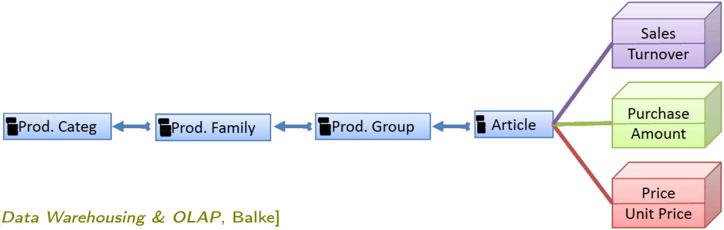
Cube

Cube Schema: set of n dimension schemas, and m measures $(\{D_1, \ldots, D_n\}, \{M_1, \ldots, M_m\})$.

Cube: instance of the cube schema= set of cells in $\mathrm{dom}(D_1) \times \cdots \times \mathrm{dom}(D_n) \times \mathrm{dom}(M_1) \times \cdots \times \mathrm{dom}(M_m)$
The cube contains *all* cells from its domain, not only the non-empty ones.

Can be observed at several granularities (determined by one level on each dimension)

Reminder: several cubes may share a dimension.



[*Data Warehousing & OLAP*, Balke]

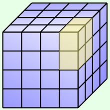
Table des matières

OLTP vs OLAP: queries

Queries

Operational DB	DW
transactions	analytical queries
INSERT, UPDATE, SELECT	SELECT, bulk insert
query deals with few tuples (often 1)	drill-down, slice. . .)

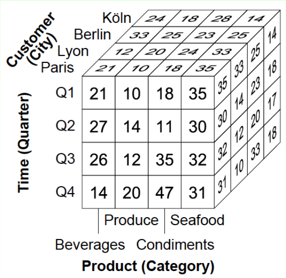
```
UPDATE Customers
SET City='Heidelberg'
WHERE CustomerName='H. Plattner';
```

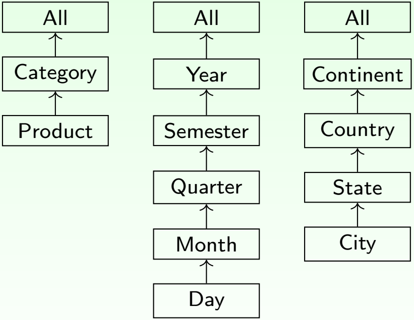


Multidimensional model

We next illustrate operations on:

Cube, and dimension schemas





Restrictions:

- we consider a single cube
- this cube has only 3 dimensions
- we consider a single measure (sales quantity)

OLAP operations

Typical OLAP operations (cube navigation):

- Roll-up
- Drill-down
- Slice and Dice

Typical OLAP operations (rearranging the cube):

- Pivot
- Sort

Advanced OLAP operations:

- Drill-across
- Drill-through

Other common operations: aggregate functions, ranking functions...

Roll-up

In short; *roll-up* summarizes data by navigating upward in the hierarchy.
Ex: (category,city,day)$\rightarrow$ (category,country,year).

Roll-up/Drill-up/Consolidate

Let C a cube with schema $(\{D_1,\ldots,D_n\},\{M_1,\ldots,M_m\})$ at granularity $G = (l_1,\ldots,l_n)$, where $1 \leq l_i < m+1$ is the level in dimension D_i : $(D_i.k_1,\ldots,D_i.k_m,\mathrm{Top}_{D_i},\rightarrow)$.
A Roll-up operation navigates toward a coarser granularity: some dimension $Dim^{up} \subset \{D_1,\ldots,D_n\}$ are rolled-up;
The new granularity is $G' = (l'_1,\ldots,l'_n)$.

- $\forall D_i \in Dim^{up}, l_i < l'_i \leq m+1$
- $\forall D_i \notin Dim^{up}, l_i = l'_i$

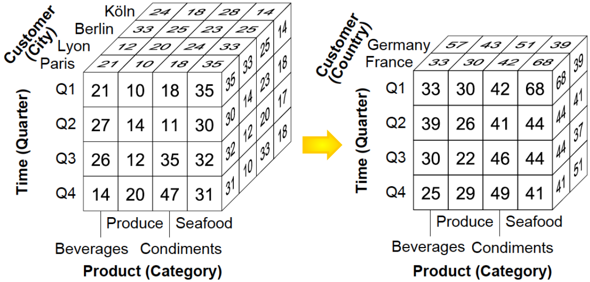
This definition allows

- to zoom-out on multiple dimensions at once
- to roll-up a dimension to the Top_D level*

*sometimes called *Drill-in* or *Merge*

Roll-up

ROLLUP(Cube,Dimension$\rightarrow$ Level, AggFunction(Measure))



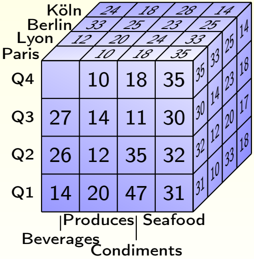
Roll-up to the Country level on Customer dimension:

ROLLUP(Sales, Customer$\rightarrow$ Country, SUM(Quantity))

Remark: beware semi-additive or non-additive measures.

Roll-up

Draw the result of the operation.



ROLLUP(Sales, Time$\rightarrow$ Semester, SUM(Quantity))

ROLLUP(Sales, Customer$\rightarrow$ Country,Time$\rightarrow$ Top$_{\text{Time}}$,
SUM(Quantity))

Hierarchical rollup, *Dimensional* rollup*, or both simultaneously.

*sometimes called *Drill-out* or *Split*

Drill-down

Inverse operation of Roll-up: zoom-in on the data.
Ex: (category,country,year)$\rightarrow$ (category,city,day).

Drill-Down

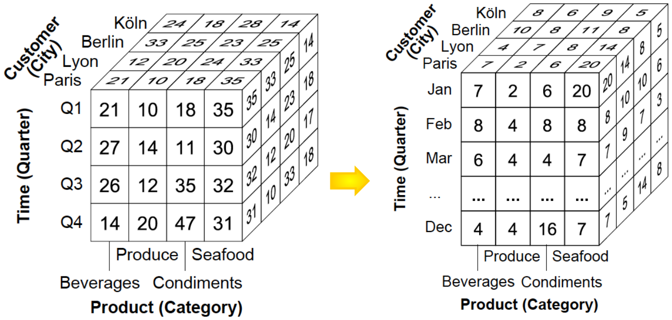
Let C a cube with schema $(\{D_1,\ldots,D_n\},\{M_1,\ldots,M_m\})$ at granularity $G = (l_1,\ldots,l_n)$, where $1 \leq l_i < m+1$ is the level in dimension D_i : $(D_i.k_1,\ldots,D_i.k_m,\mathrm{Top}_{D_i},\rightarrow)$.
A Drill-drown operation navigates toward a finerer granularity: some dimension $Dim^{down} \subset \{D_1,\ldots,D_n\}$ are drilled-down;
The new granularity is $G' = (l'_1,\ldots,l'_n)$.

- $\forall D_i \in Dim^{down}, 1 \leq l'_i < l_i$
- $\forall D_i \notin Dim^{down}, l_i = l'_i$

It's obvious but...you cannot drill down if you do not have the finer-grained data!

Drill-down

DRILLDOWN(Cube,Dimension→ Level)



Drill-down to the Month level on Time dimension:

DRILLDOWN(Sales, Time→ Month)

Drill-down/Roll-up Summary

Navigate between granularities.

- Roll-up: fewer details
  measure may be computed from input cube
- Drill-down: more details
  measure computed from the finest-detailed data

. . . assuming measures have not (yet) been materialized at other granularities.

The number of dimensions remains the same (except when at top level of dimension).

Slice and Dice

Slice and Dice definitions

Slice: returns a “slice” of the cube by selecting a *single* value on *one* of the dimensions.

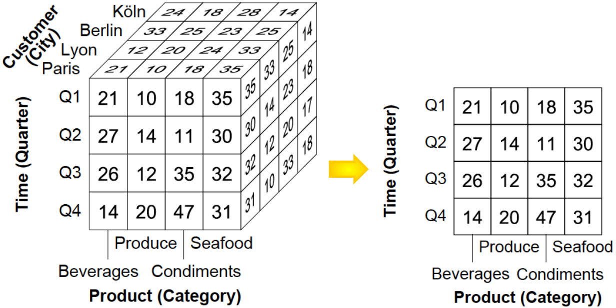
$\Longrightarrow$ corresponds to SQL's WHERE with equality selection.

Dice: returns a “dice" of the cube by selecting for each dimension a boolean combination of range or value conditions one one dimension *single* value.

⚠ definitions vary between authors.

Slice

SLICE(Cube,Dimension, Level=value)

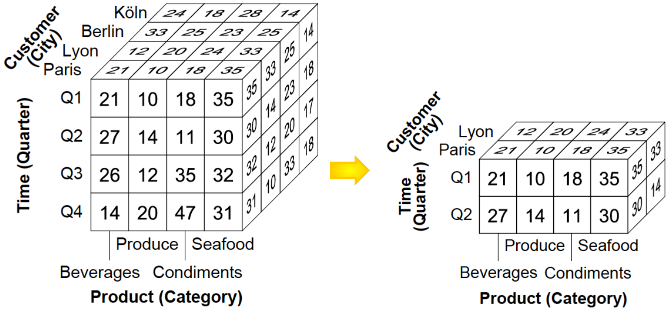


Slice on City='Paris':

SLICE(Sales, Customer,City='Paris')

Dice

DICE(Cube,Φ), with Φ: boolean combination



Dice on City=‘Paris’ or ‘Lyon’ and Quarter=‘Q1’ or ‘Q2’:
DICE(Sales,(Customer.City=‘Paris’ OR Customer.City=‘Lyon’)
AND (Time.Quarter=‘Q1’ OR Time.Quarter=‘Q2’))

Other operations

A few more typical operations on data:

- sorting the cube
- pivoting
- joining cubes

$\hookrightarrow$ but no official standard for OLAP queries.

Cross-tabulations

Also called crosstab, pivot table, contingency table.

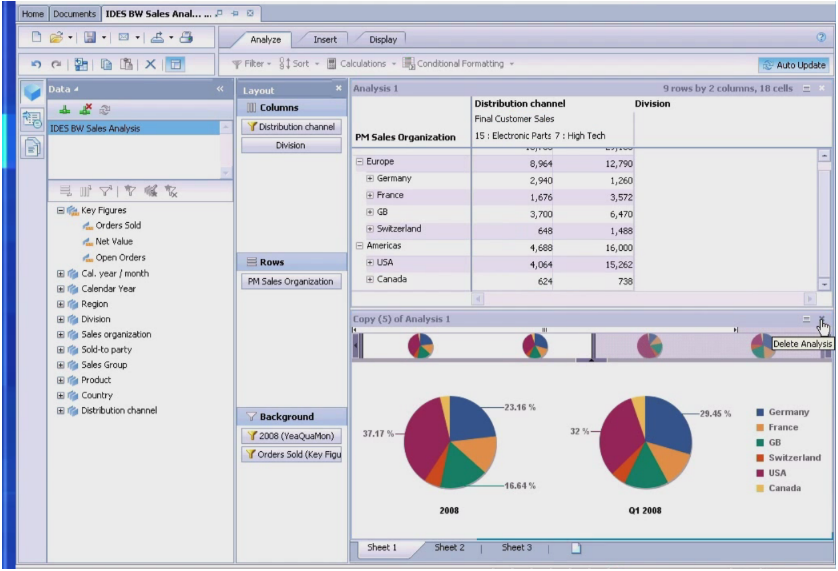
Simple tool to help find interactions between 2 variables.

Sales per city and day

	Mon	Tue	Wed	Thu	Fri	Sat	Row totals
Paris	10	20	30	40	30	10	140
Lyon	40	20	20	30	50	30	190
Lille	50	20	30	20	10	0	130
Col totals	100	60	80	90	90	40	460

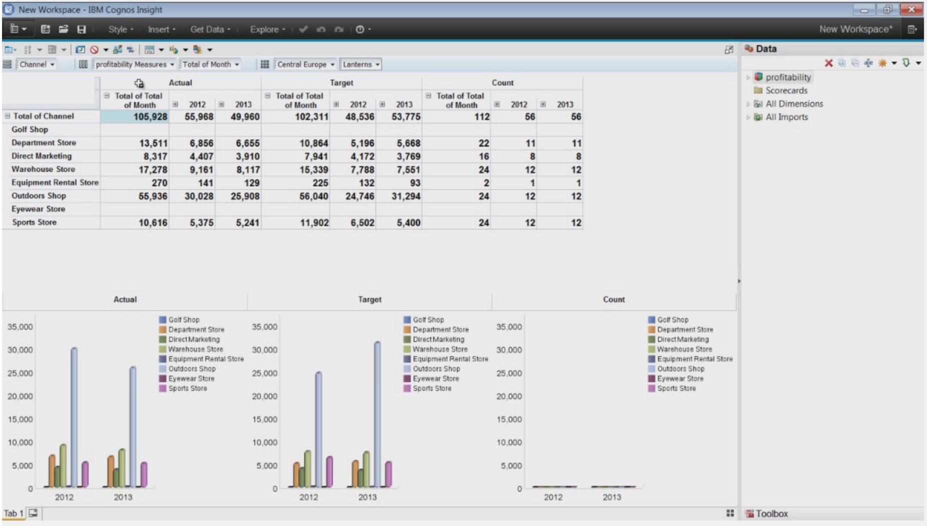
We can view pivoting as selecting 2 dimensions to aggregate some measure. Can be computed in Excel, and similar operation in many SQL implementations.

OLAP visualization in practice



SAP Business Object[https://www.youtube.com/watch?v=3rhiRjH6LiA]

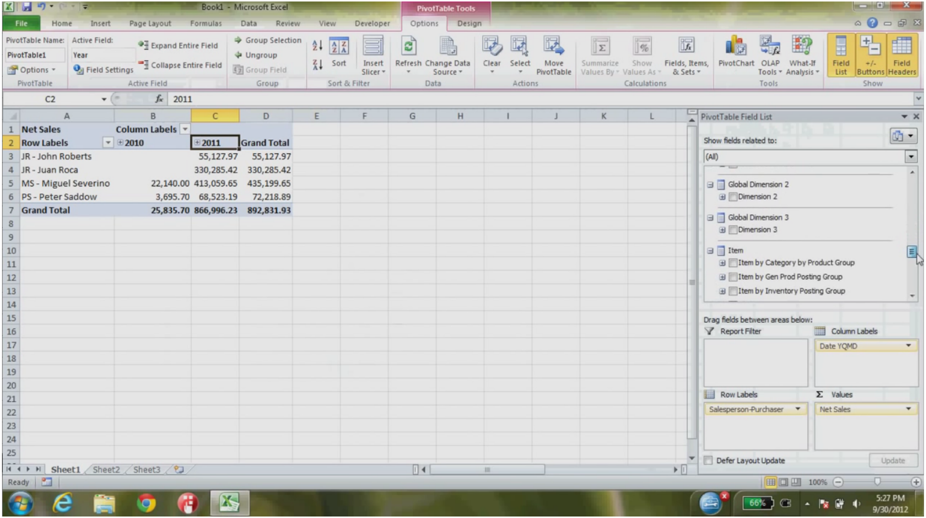
OLAP visualization in practice(2)



	Actual			Target			Count		
	Total of Total of Month	2012	2013	Total of Total of Month	2012	2013	Total of Total of Month	2012	2013
Total of Channel	105,928	55,968	49,960	102,311	48,536	53,775	112	56	56
Golf Shop									
Department Store	13,511	6,856	6,655	10,864	5,196	5,668	22	11	11
Direct Marketing	8,317	4,407	3,910	7,941	4,172	3,769	16	8	8
Warehouse Store	17,278	9,161	8,117	15,339	7,788	7,551	24	12	12
Equipment Rental Store	270	141	129	225	132	93	2	1	1
Outdoors Shop	55,936	30,028	25,908	56,040	24,746	31,294	24	12	12
Eyewear Store									
Sports Store	10,616	5,375	5,241	11,902	6,502	5,400	24	12	12

IBM Cognos[https://www.youtube.com/watch?v=FjKaRU5V1Rw]

OLAP visualization in practice(3)



Excel[https://www.youtube.com/watch?v=ygAs-6mEhBA]