Data Warehousing and Data Mining
– Case Studies –

- Inventory
- Order Management
- OncoNet
- MEDAN

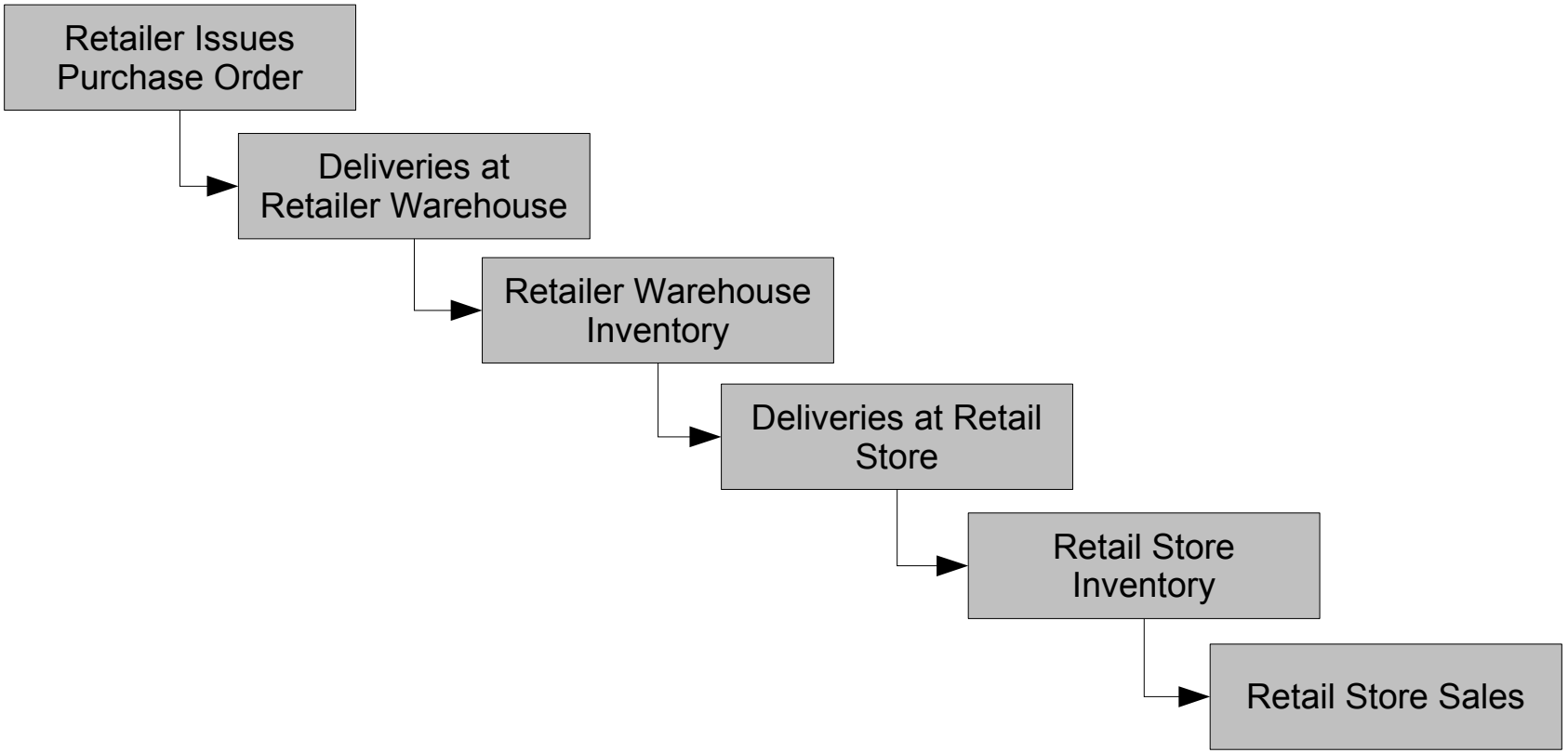
Acknowledgements: I am indebted to Michael Böhlen for providing me his slides, upon which these lecture notes are based.

Inventory Example/1

- Consider a large grocery chain with a central warehouse and several retail stores
- Advanced retail business requires inventory information.
  - Making sure the right product is in the right store at the right time minimizes out-of-stocks and reduces overall inventory carrying costs.
  - The retailer needs the ability to analyze daily quantity-on-hand inventory levels by product and store.
- Design dimensional models that support the analysis of inventories for retail businesses (grocery stores).

Inventory Example/2

- The **value chain** identifies the natural, logical flow of an organization's primary activities.
- Operational systems provide snapshots at each step with interesting data and performance metrics



Inventory Example/3

- 3 different inventory models
  - Inventory periodic snapshot
  - Inventory transactions
  - Inventory accumulating snapshot

Inventory Example/4

- **Model 1: Inventory Periodic Snapshot**
  - Every day (or at some other regular time interval) the inventory levels of each product is measured and stored as a new row in the fact table
- Business process: Analysis of retail store inventory
- Granularity: Daily inventory by product at each individual store
- Dimensions: Date, product, and store
- Facts/measures: Quantity on hand

Inventory Example/5

- Inventory generates **dense snapshot** tables
  - In contrast, POS Retail Sales table was sparse
- Consequently, inventory fact table is growing fast
  - 60.000 products x 100 stores = 6 Mio. rows each time
  - With a row width of 14 bytes, this is 84 MB each time
  - 1 year of daily snapshots would be 30 GB
- Reduce snapshot frequencies over time
  - Last 60 days of inventory at daily level
  - Weekly snapshots for older data
  - Instead of 1.095 snapshots in 3 years, only 208 snapshots would be required

Inventory Example/6

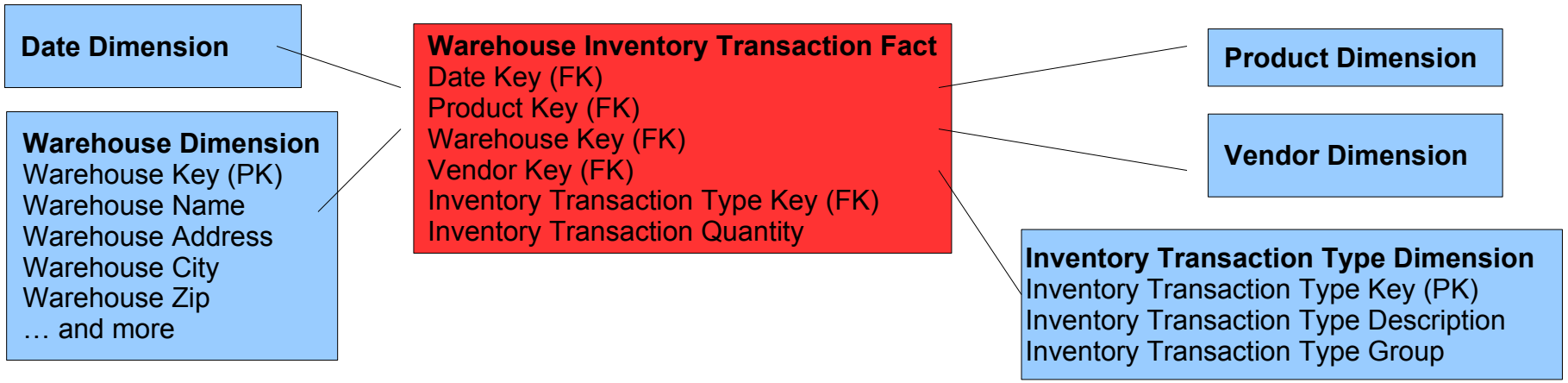
- The quantity on hand is a **semi-additive** measure
  - Can be summarized across products and stores, but not across time
  - Different in POS Retail Sales table: Sold entities are counted only once
- All measures that record a **static level** (inventory, financial account balance, measures of intensity e.g., temperature) are **inherently non-additive** across time and possibly other dimensions.
  - Can be aggregated along time dimension by averaging
- A note about SQL AVG function
  - Cannot be used to compute the average over time, since it averages over the number of rows
  - Avg inventory over a cluster of 3 products in 4 stores across 7 days would divide the summed value by 84

Inventory Example/7

- **Model 2: Inventory Transactions**
  - Record every transaction that affects inventory
- Inventory transactions in the store chain
  - Receive product
  - Place product in to inspection hold
  - Release product from inspection hold
  - Return product to vendor due to inspection failure
  - Place product in bin
  - Authorize product for sale
  - Pick product for shipment
  - Ship product to customer
  - Receive product form customer
  - Return product to inventory from customer return
  - Remove product from inventory

Inventory Example/8

- Model



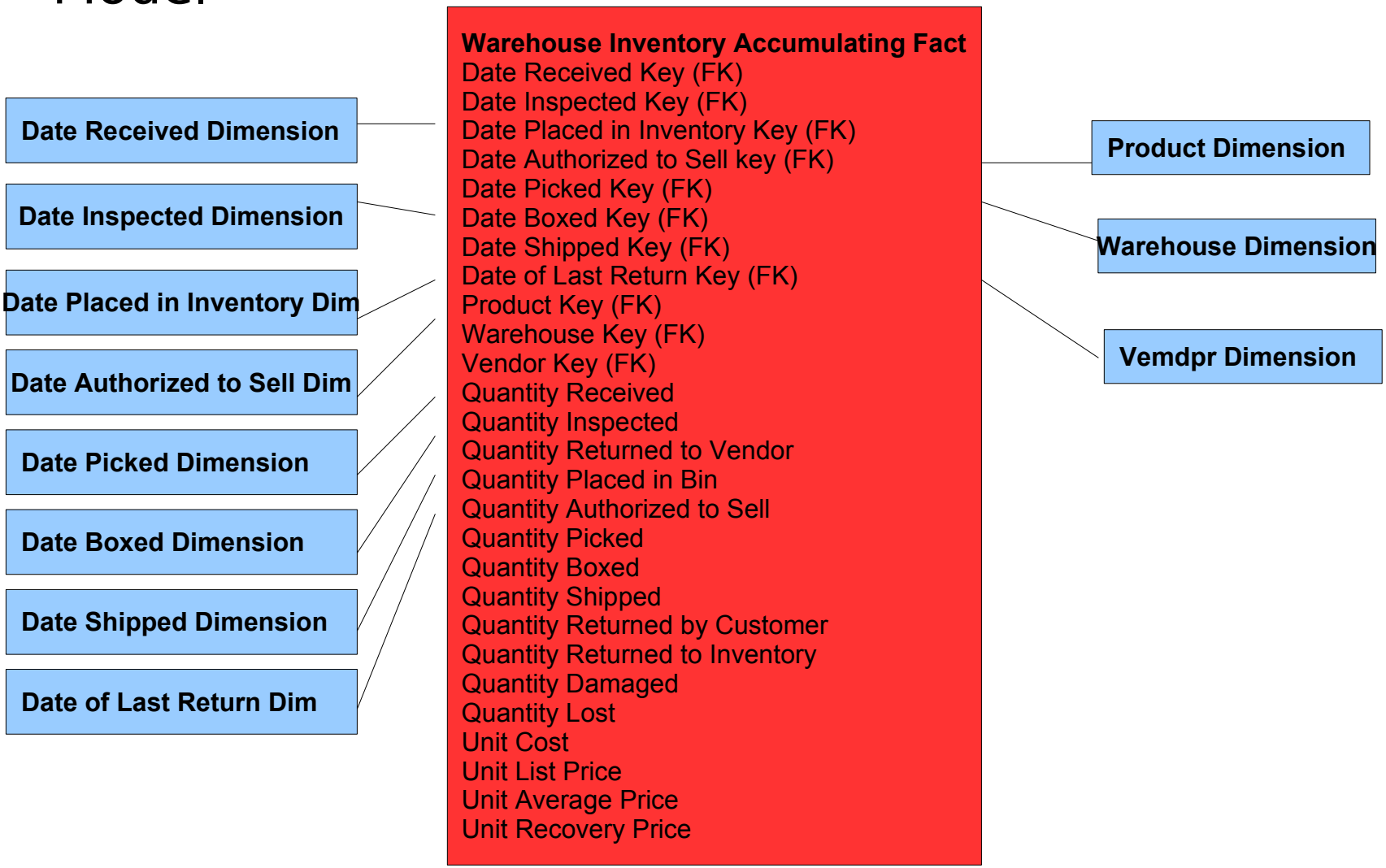
- Contains most detailed information, e.g.,
  - How many shipments from a given vendor?
  - On which products more than one round of inspection?
- Reconstruction of exact inventory numbers is possible, but not practical!
  - Used in combination with other fact table

Inventory Example/9

- **Model 3: Inventory Accumulating Snapshot**
  - One row in the fact table for each shipment of a particular product to the warehouse
  - This row tracks the disposition of the shipment through the warehouse
    - i.e., provides an updated status of the shipment as it moves through the warehouse
- Assume that the inventory goes through a series of events, e.g., receiving, inspection, bin placement, authorization to sell, picking, boxing, and shipping.

Inventory Example/10

- Model

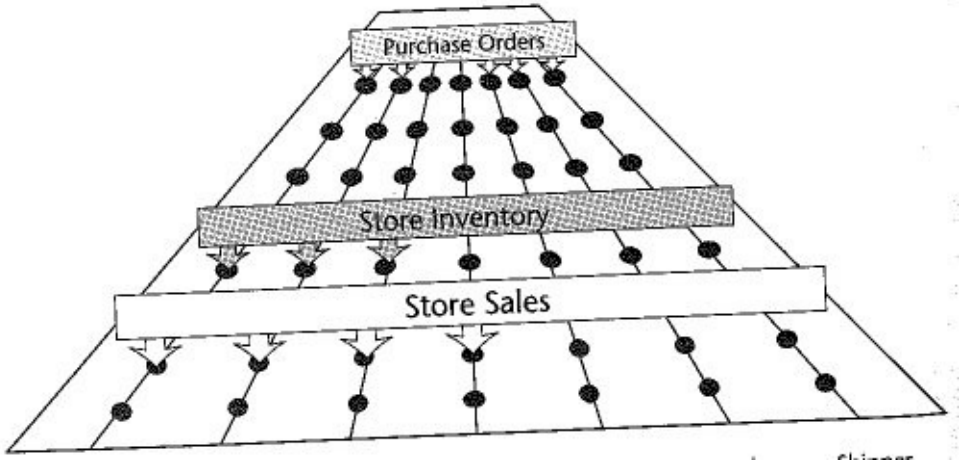


Inventory Example/11

- **Value Chain Integration**
  - Provide analysis across the business to better evaluate the performance (not just at the individual department level)
    - End-to-end perspective high-level management to customer
  - Requires integration and consistent handling/use of data
    - Individual fact tables for processes + shared dimensions
- **Shared dimensions** are crucial to design data marts that can be integrated
  - Allow to combine performance measurement for different processes (also known as **drill across**)

Inventory Example/12

- **Data Warehouse Bus Architecture**
  - A standard bus interface for a data warehouse environment that allows to implement separate data marts that can be successfully integrated
  - Based on conformed (similar) dimensions that are shared by the data marts



- Guides the overall design and breaks down the development process into small chunks (DMs)

Inventory Example/13

- **Data Warehouse Bus Matrix**
  - Tool to create and document the bus architecture
  - Rows represent business processes (translate into DMs)
  - Columns represent a suite of standardized dimensions

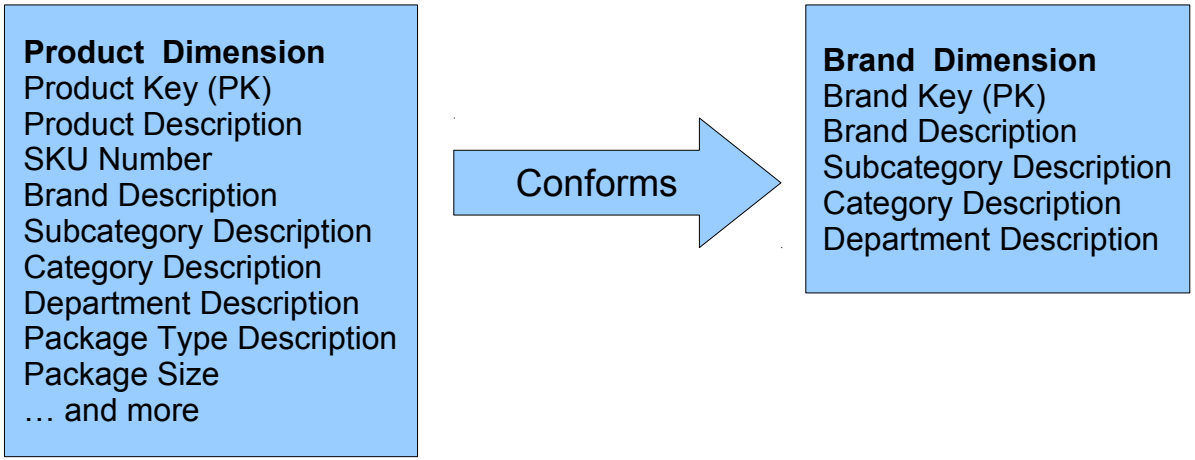
	COMMON DIMENSIONS							
BUSINESS PROCESSES	Date	Product	Store	Promotion	Warehouse	Vendor	Contract	Shipper
Retail Sales	X	X	X	X				
Retail Inventory	X	X	X					
Retail Deliveries	X	X	X					
Warehouse Inventory	X	X			X	X		
Warehouse Deliveries	X	X			X	X		
Purchase Orders	X	X			X	X	X	X

Inventory Example/14

- Creating the DW bus matrix is one of the most important up-front deliverables of a DW implementation
  - Create a comprehensive list of dimensions before filling in the matrix
- The rows provide a concise overview about the dimensionality of the individual DMs
- The columns show the interaction between the DMs and the common/shared dimensions

Inventory Example/15

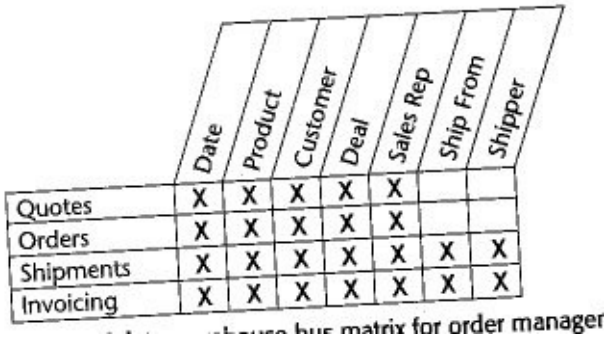
- **Conformed Dimension**
  - Either identical or strict mathematical subsets of the most granular, detailed dimension
  - Have consistent dimension keys, and consistent column names and values
- Roll-up dimensions conform to the base-level atomic dimension
  - Brand table is a strict subset product table



Order Management Example/1

- **Order management** consists of several critical business processes (order, shipment, invoice processing, etc.) and measures (sales volume, invoice revenue, etc.)

- Warehouse bus matrix

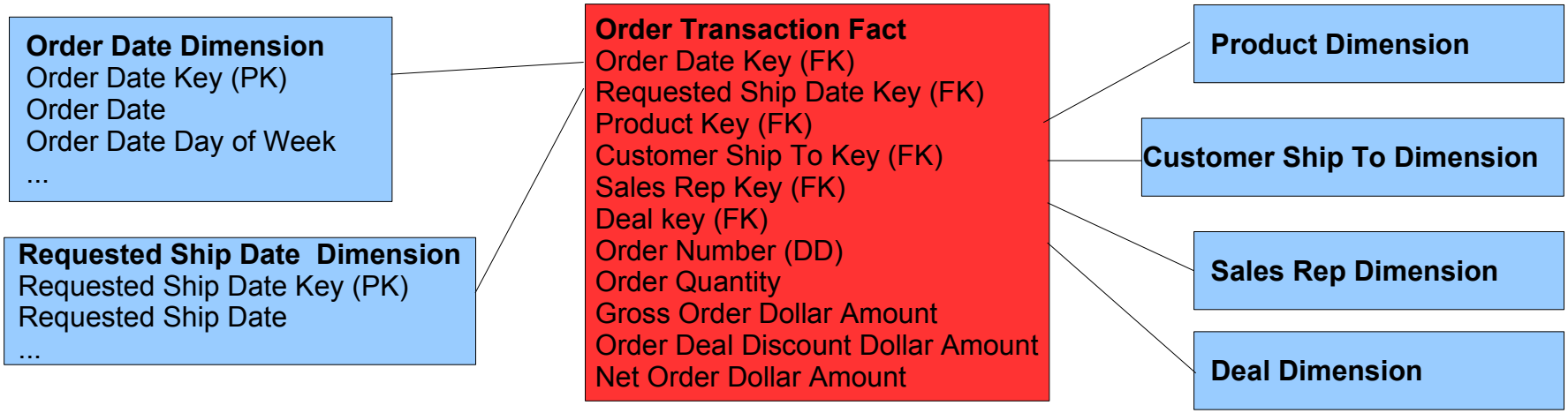


	Date	Product	Customer	Deal	Sales Rep	Ship From	Shipper
Quotes	X	X	X	X	X		
Orders	X	X	X	X	X		
Shipments	X	X	X	X	X	X	X
Invoicing	X	X	X	X	X	X	X

bus matrix for order manager

Order Management Example/2

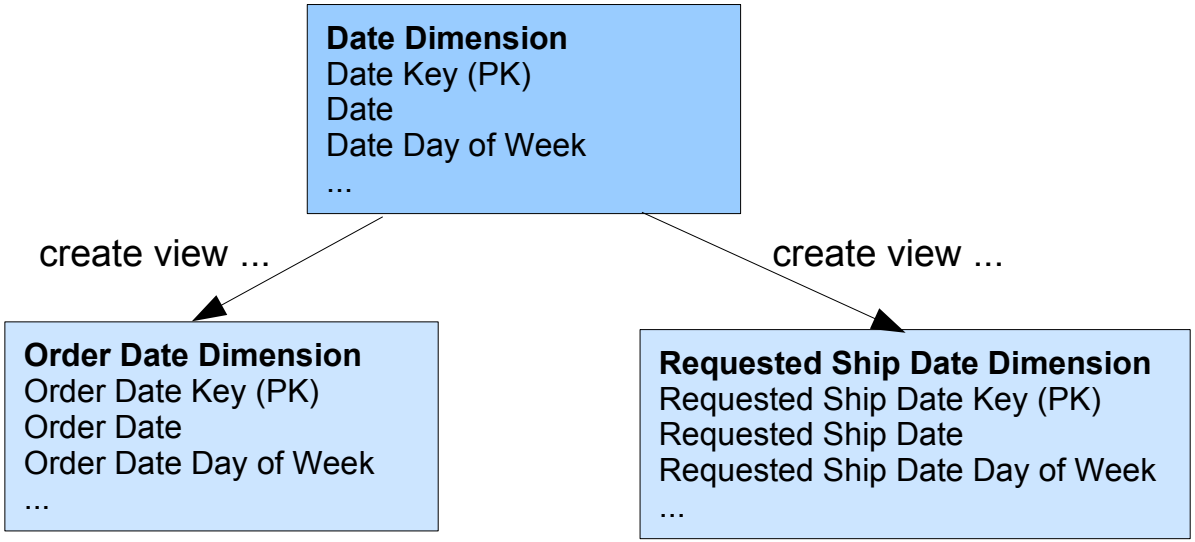
- Model



- Issues
  - Should there be a “Ship Date Key” in the fact table?
  - Can/should Order Date Key and Requested Ship Date Key be foreign keys to the same dimension table?

Order Management Example/3

- **Role-playing:** When a single dimension appears several times in the same fact table,
  - e.g., order date and requested ship date
- Should **not** be FK to the same dimension table
  - SQL would require the two dates to be the same
  - We might want to constrain the two dimensions differently
- Might be one physical table, but each of the roles should be defined as a view with different labels



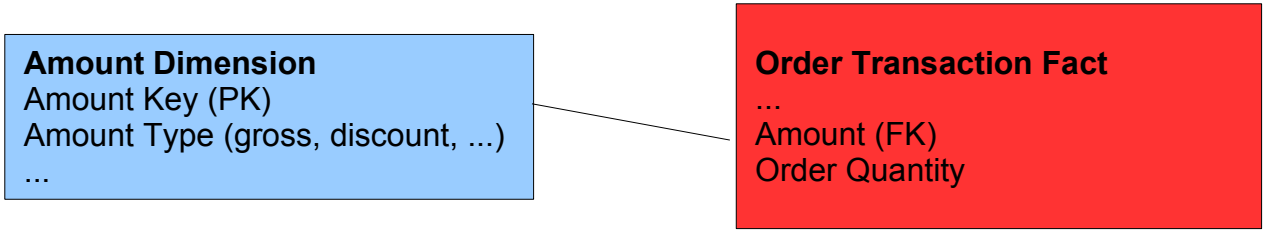
Order Management Example/4

- **Multiple hierarchies** often coexist in a dimension table, especially for customer-oriented dimensions
  - Natural geographic hierarchy
  - Zip Code defines a second hiearchy
- Can have different number of levels
- Should all be supported in a DW

Customer Ship To Dimension
Customer Ship to Key (PK)
Customer Ship To ID (NK)
Customer Ship To Name
Customer Ship To Address
Customer Ship To City
Customer Ship To State
Customer Ship To Zip + 4
Customer Ship To Zip Region
Customer Ship To Zip Sectional Center
Customer Bill To Name
Customer Bill To Address Attributes
Customer Credit Rating
Customer URL
...

Order Management Example/5

- **Fact normalization:** Further normalize the fact table and collapse all measures into a single measure along with a dimension that identifies the type of the measure.



- Makes only sense if
  - fact table is sparsely populated and no computations are made between measures of different type
  - e.g., medical tests where different things are measured and data is sparse

OncoNet DW/1

- Collaboration with Hospital Meran
  - BSc thesis of A. Heinisch
- OncoNet is an application for the management of patients undergoing a cancer therapy
  - Cancer therapy follows a treatment plan / protocol

Event type	Event description	Date
Data collection	Ct Thorax	Day 60, 100, 360, 720
Data collection	Ct Abdomen	Day 60, 100, 360, 720
Data collection	Hemogram	Day 110
Medication	Zofran	Day 1
Medication	Adriblastina	Day 1

OncoNet DW/2

- **Business process:** Analysis of cancer therapies
  - How many patients with normal blood pressure after …
  - Which dosages of drug A were successful to reduce …
- **Granularity**: Individual events of the chemotherapy
  - Includes measurements, examinations, questionnaires, etc.

OncoNet DW/3

- Patient Dimension and Drug Dimension

Date Dimension
Patient Key (SK)
Patient ID
Patient First Name
Patient Last Name
Patient Gender
Patient Address
Patient ZIP Code
Patient Phone Number
Patient Profession
Patient First Language
Patient Height
Patient Weight
Patient Body Surface Area
Patient Place of Birth
Patient Birthday Date key
Patient Death Date Key
Patient First Admission Doctor
Patient First Admission Area
Patient Smoking Indicator
Patient Cigarettes per Day
Patient Alcohol Indicator
Patient Alcohol Amount per Day
...

Drug Dimension
Drug Key (SK)
Drug ID
Drug Name
Drug Category
Drug Active Substance
Drug manufacturer
Drug Quantity Unit
Drug Quantity Unit Description
Drug Administration Type
Drug Administration Location
Drug Packaging
Drug Packaging AIC Code
...

OncoNet DW/4

- Normalized fact table
  - One measure only
  - Type of measure is described in Event and Investigation dimension

OncoNet DW/5

Patient Key	Patient ID	First Name	Last Name	Gender	Language	Weight	Height
1	522	Andreas	Heinisch	Male	German	88	185

Date Key	Date Type	Full Date	Day of Week	Weekday Indicator	Month	Year
1	Normal	30.05.06	Tuesday	Weekday	May	2006

Date Key	Patient Key	Therapy Key	Event Key	Investigation Key	Numverical Value
1	1	1	1	1	32,7
1	1	1	1	2	11,4
1	1	1	1	3	97,8

Therapy Key	Therapy ID	Therapy Name	Therapy Type
1	29854	NHL Chop 14	Profile

Investigation Key	Group Level	Label	Unit
1	Hemogram	MCH	1
1	Hemogram	HB	g/dl
3	Hemogram	MCV	pg

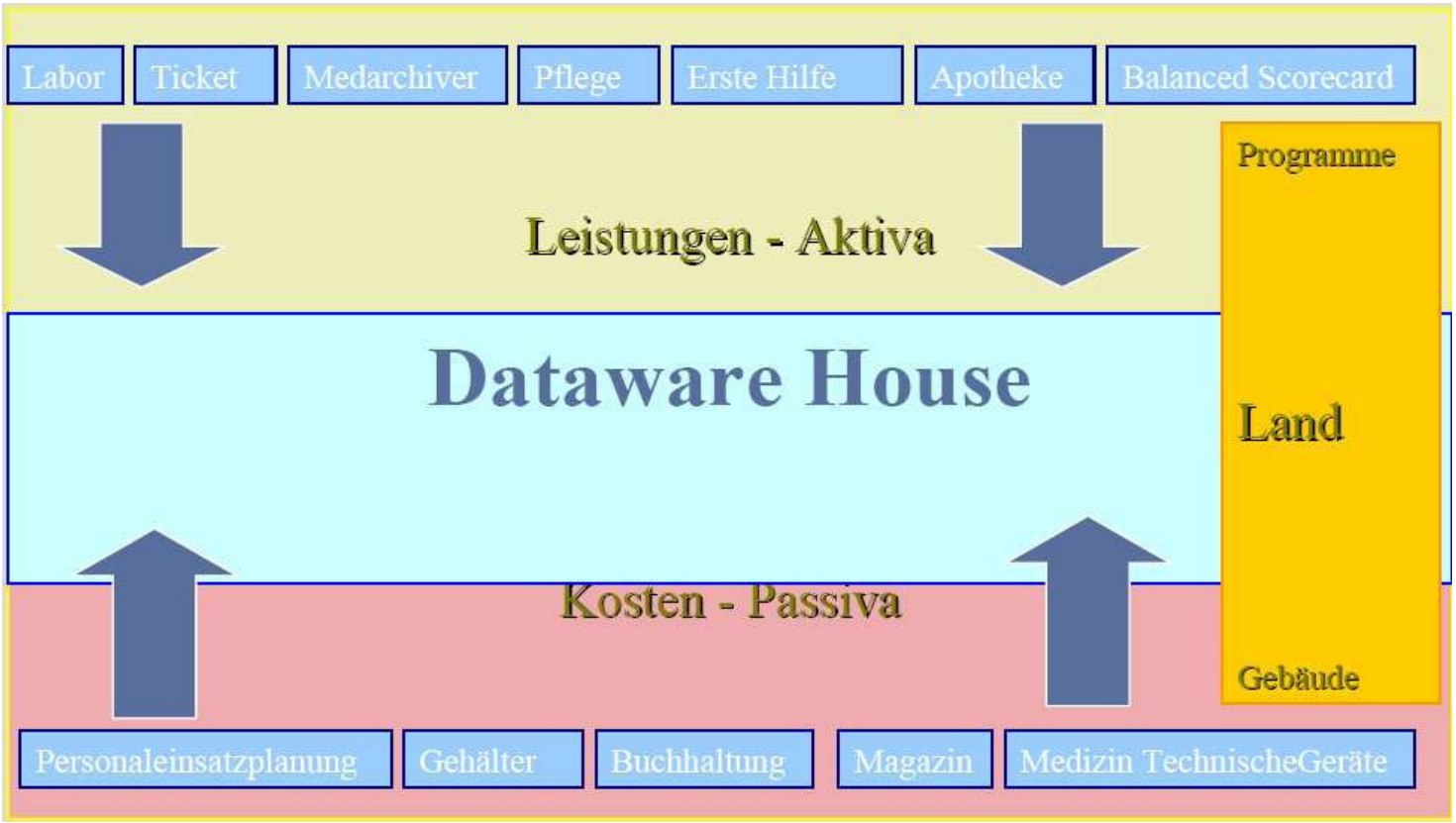
Event Key	Event ID	Name	Type	Responsible
1	9267	Urgent Laboratory	Laboratory Test	Nurse

MEDAN DW/1

- MEDAN
  - **ME**dical **D**ata Warehousing and **AN**alysis
  - Collaboration between the Hospital Meran and the FUB
  - http://www.inf.unibz.it/dis/projects/medan
- Objectives
  - Conduct research and create competences in the field of medical data warehousing and analysis
  - Build a BI/DW solution
    - Administrative DW
    - Medical DW
  - Develop and apply data analysis/mining techniques

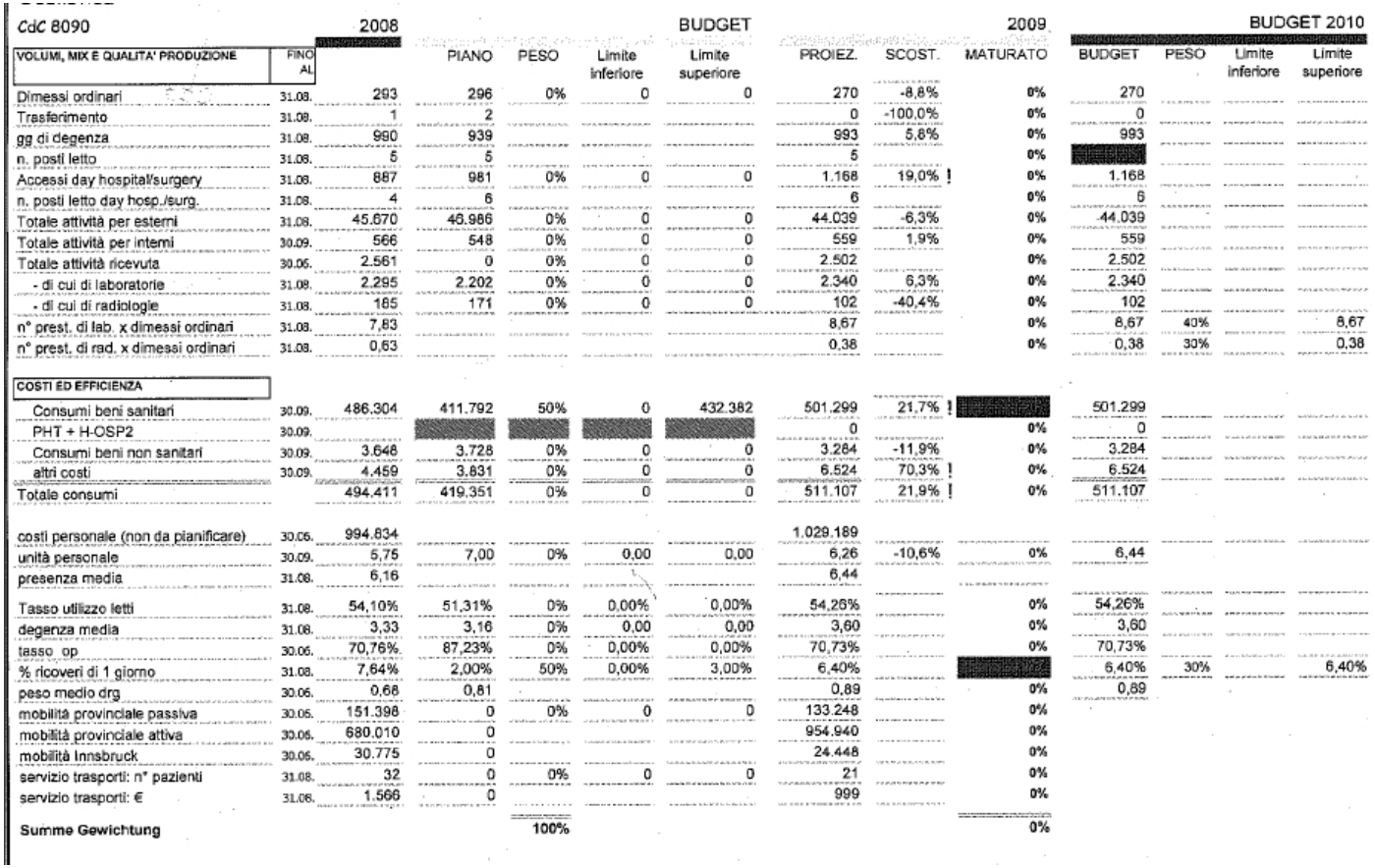
MEDAN DW/2

- Data sources in a health care environment
  - Internal production systems (SQL, Excel, Text files, ...)
  - External information systems



MEDAN DW/3

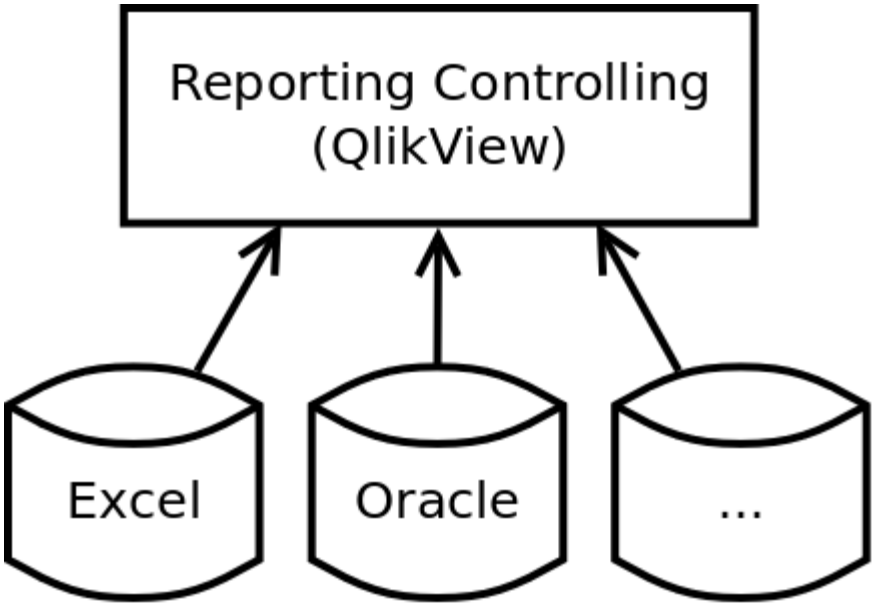
- **Example:** Budgeting in the controlling department
  - Complex Excel spreadsheets have been used in the past



CdC 8090		2008			BUDGET				2009	BUDGET 2010			
VOLUMI, MIX E QUALITA' PRODUZIONE	PINO AL		PIANO	PESO	Limite inferiore	Limite superiore	PROIEZ.	SCOST.	MATURATO	BUDGET	PESO	Limite inferiore	Limite superiore
Dimessi ordinari	31.08.	293	296	0%	0	0	270	-8,8%	0%	270			
Trasferimento	31.08.	1	2				0	-100,0%	0%	0			
gg di degenza	31.08.	980	939				993	5,8%	0%	993			
n. posti letto	31.08.	5	5				5		0%				
Accessi day hospital/surgery	31.08.	887	981	0%	0	0	1.168	19,0% !	0%	1.168			
n. posti letto day hosp./surg.	31.08.	4	6				6		0%	6			
Totale attività per esterni	31.08.	45.670	46.986	0%	0	0	44.039	-6,3%	0%	44.039			
Totale attività per interni	30.09.	566	548	0%	0	0	559	1,9%	0%	559			
Totale attività ricevuta	30.06.	2.561	0	0%	0	0	2.502		0%	2.502			
- di cui di laboratorie	31.08.	2.295	2.202	0%	0	0	2.340	6,3%	0%	2.340			
- di cui di radiologie	31.08.	165	171	0%	0	0	102	-40,4%	0%	102			
n° prest. di lab. x dimessi ordinari	31.08.	7,83					8,67		0%	8,67	40%		8,67
n° prest. di rad. x dimessi ordinari	31.08.	0,63					0,38		0%	0,38	30%		0,38
COSTI ED EFFICIENZA													
Consumi beni sanitari	30.09.	486.304	411.792	50%	0	432.382	501.299	21,7% !		501.299			
PHT + H-OSP2	30.09.						0		0%	0			
Consumi beni non sanitari	30.09.	3.648	3.728	0%	0	0	3.284	-11,9%	0%	3.284			
altri costi	30.09.	4.459	3.831	0%	0	0	6.524	70,3% !	0%	6.524			
Totale consumi		494.411	419.351	0%	0	0	511.107	21,9% !	0%	511.107			
costi personale (non da pianificare)	30.06.	994.634					1.029.189						
unità personale	30.09.	5,75	7,00	0%	0,00	0,00	6,26	-10,6%	0%	6,44			
presenza media	31.08.	6,16					6,44						
Tasso utilizzo letti	31.08.	54,10%	51,31%	0%	0,00%	0,00%	54,26%		0%	54,26%			
degenza media	31.08.	3,33	3,16	0%	0,00	0,00	3,60		0%	3,60			
tasso op	30.06.	70,76%	87,23%	0%	0,00%	0,00%	70,73%		0%	70,73%			
% ricoveri di 1 giorno	31.08.	7,64%	2,00%	50%	0,00%	3,00%	6,40%			6,40%	30%		6,40%
peso medio drg	30.06.	0,68	0,81				0,89		0%	0,89			
mobilità provinciale passiva	30.06.	151.398	0	0%	0	0	133.248		0%				
mobilità provinciale attiva	30.06.	680.010	0				954.940		0%				
mobilità Innsbruck	30.06.	30.775	0				24.448		0%				
servizio trasporti: n° pazienti	31.08.	32	0	0%	0	0	21		0%				
servizio trasporti: €	31.08.	1.566	0				999		0%				
Summe Gewichtung				100%					0%				

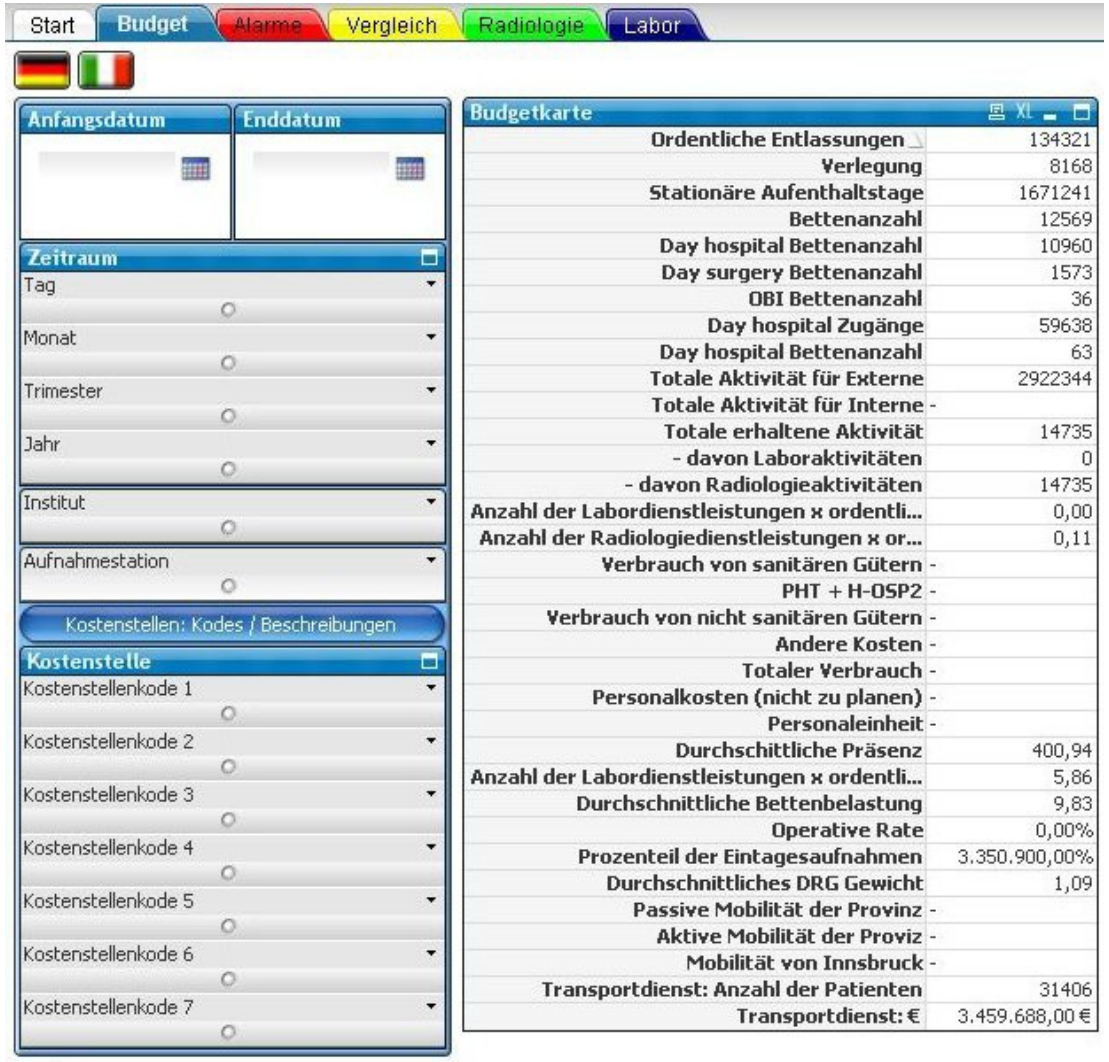
MEDAN DW/4

- **Example (contd.)**
  - A QlickView application to replace Excel
  - Direct access to the data sources
  - Data integration in QlikView
  - No DW in place



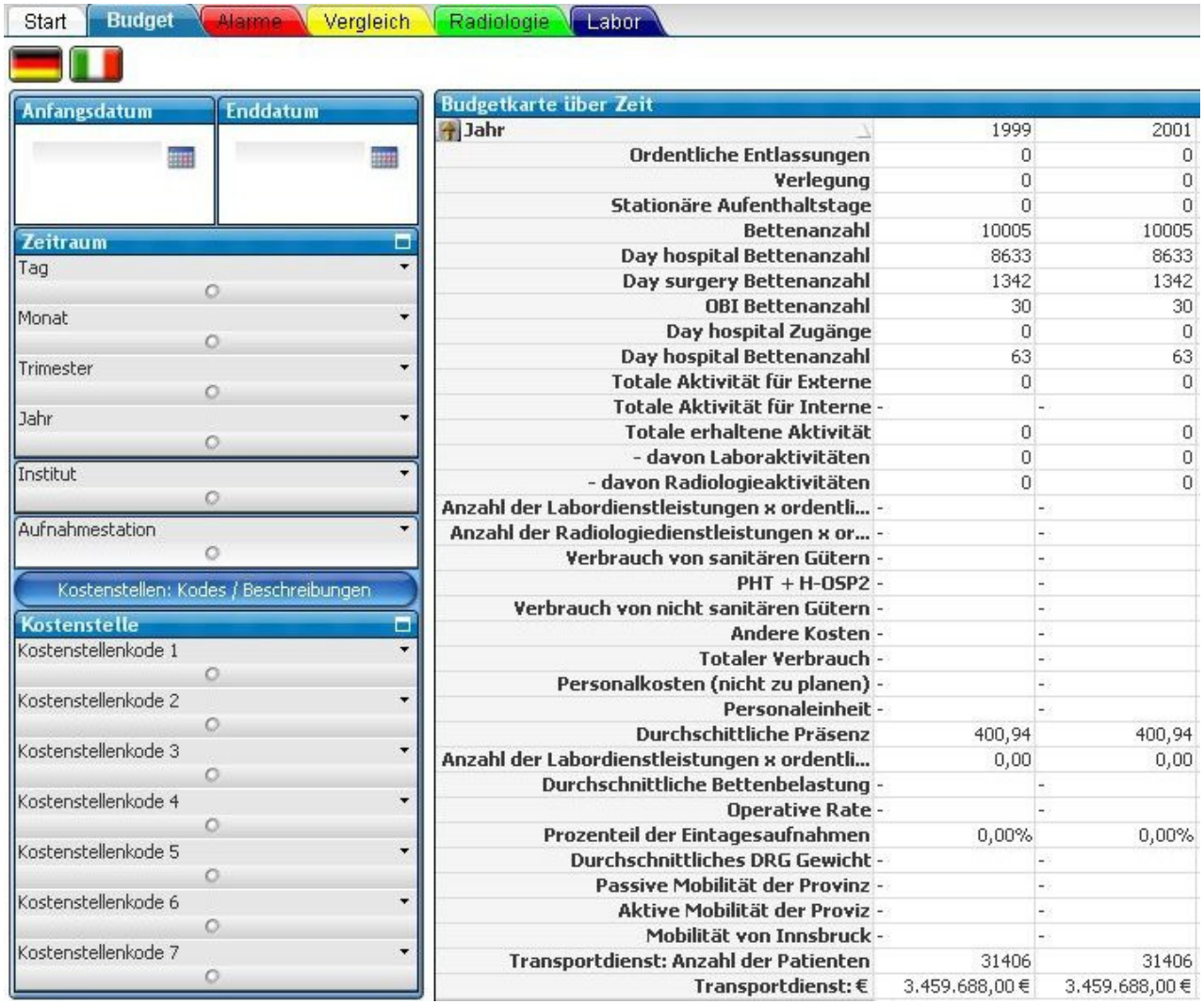
MEDAN DW/5

- **Example (contd.):** QlikView application



MEDAN DW/6

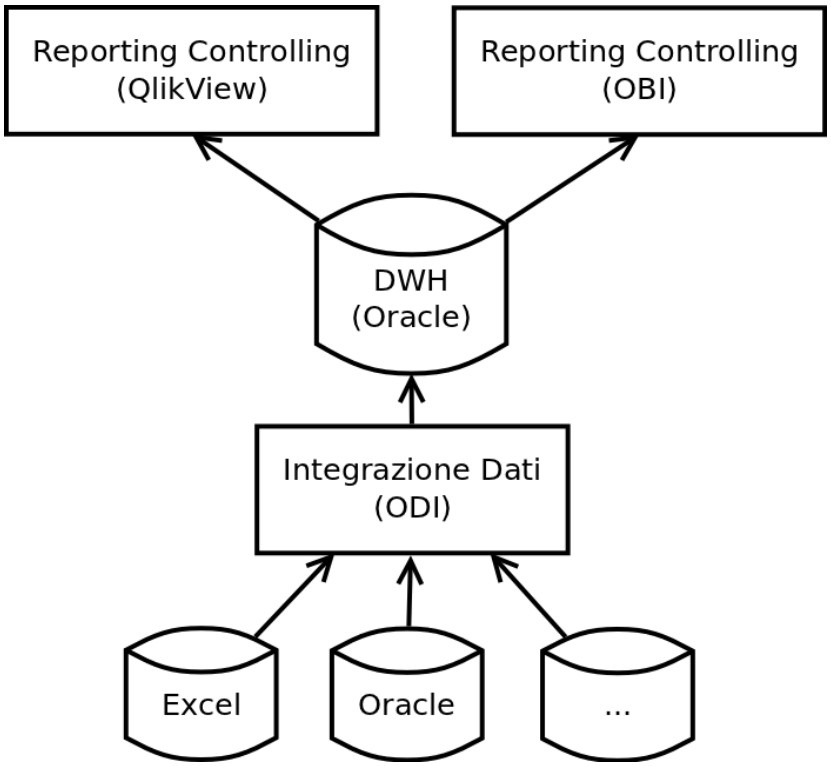
- **Example (contd.):** QlikView application



Jahr	1999	2001
Ordentliche Entlassungen	0	0
Verlegung	0	0
Stationäre Aufenthaltstage	0	0
Bettenanzahl	10005	10005
Day hospital Bettenanzahl	8633	8633
Day surgery Bettenanzahl	1342	1342
OBI Bettenanzahl	30	30
Day hospital Zugänge	0	0
Day hospital Bettenanzahl	63	63
Totale Aktivität für Externe	0	0
Totale Aktivität für Interne	-	-
Totale erhaltene Aktivität	0	0
- davon Laboraktivitäten	0	0
- davon Radiologieaktivitäten	0	0
Anzahl der Labordienstleistungen x ordentli...	-	-
Anzahl der Radiologiedienstleistungen x or...	-	-
Verbrauch von sanitären Gütern	-	-
PHT + H-OSP2	-	-
Verbrauch von nicht sanitären Gütern	-	-
Andere Kosten	-	-
Totaler Verbrauch	-	-
Personalkosten (nicht zu planen)	-	-
Personaleinheit	-	-
Durchschittliche Präsenz	400,94	400,94
Anzahl der Labordienstleistungen x ordentli...	0,00	0,00
Durchschnittliche Bettenbelastung	-	-
Operative Rate	-	-
Prozenteil der Eintagesaufnahmen	0,00%	0,00%
Durchschnittliches DRG Gewicht	-	-
Passive Mobilität der Provinz	-	-
Aktive Mobilität der Proviz	-	-
Mobilität von Innsbruck	-	-
Transportdienst: Anzahl der Patienten	31406	31406
Transportdienst: €	3.459.688,00 €	3.459.688,00 €

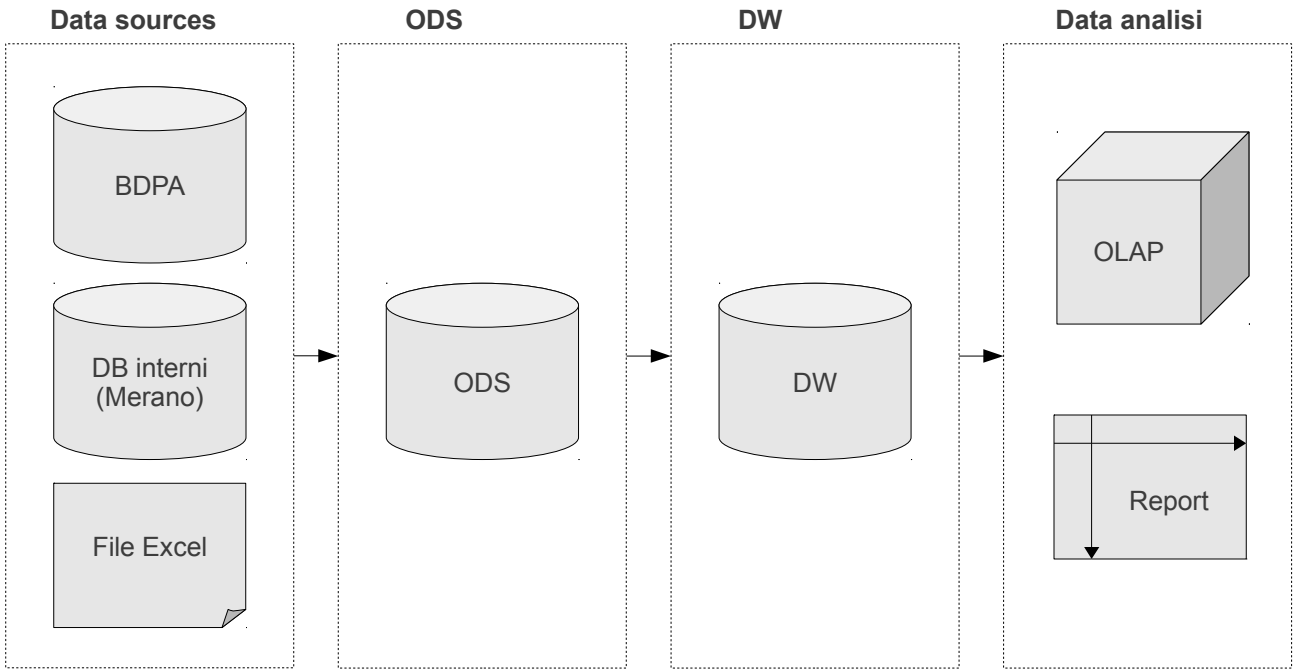
MEDAN DW/7

- **Difficult** to convince decision makers to build a **DW as the core of a BI** solution
- QlikView is mainly an analysis tool and cannot replace a DW
  - Rather ad hoc
  - Difficult to control data quality
  - Not scalable for many applications, changing sources, etc.



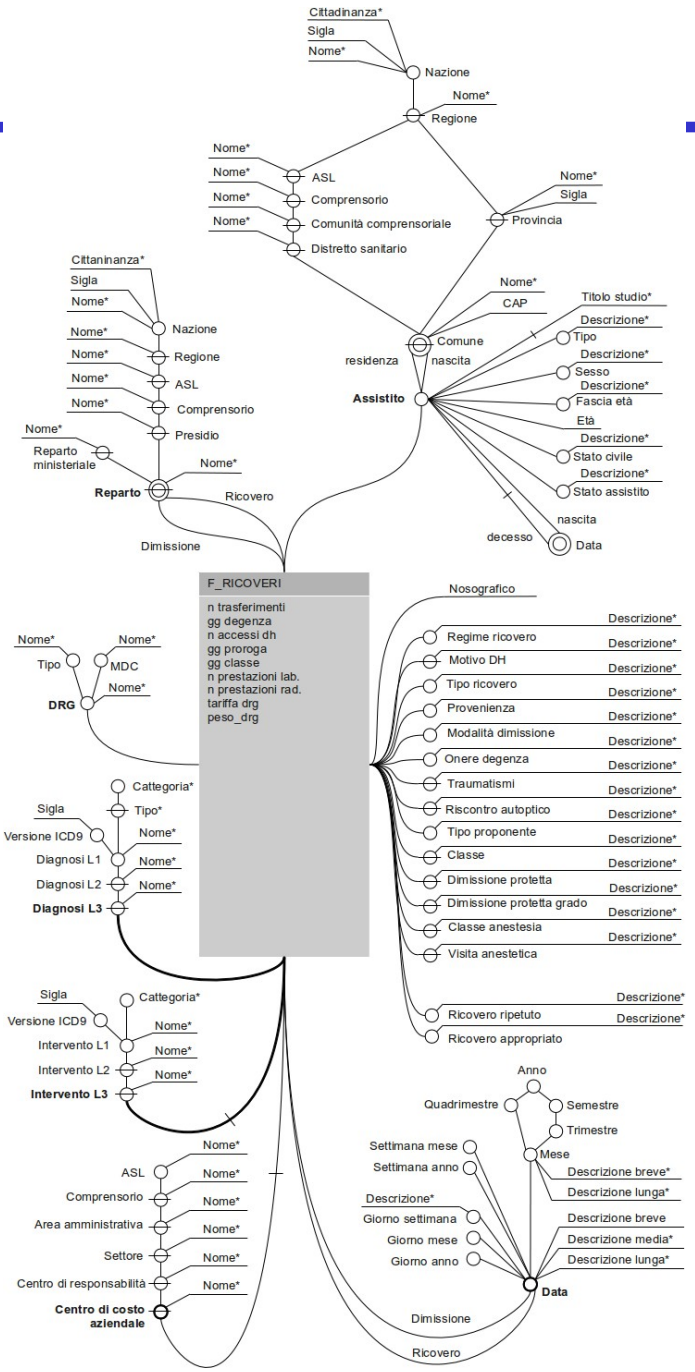
MEDAN DW/8

- Architecture
  - Oracle ODI for ETL part and ODS
  - Oracle DB for the DW
  - QlikView and OBI for data analysis



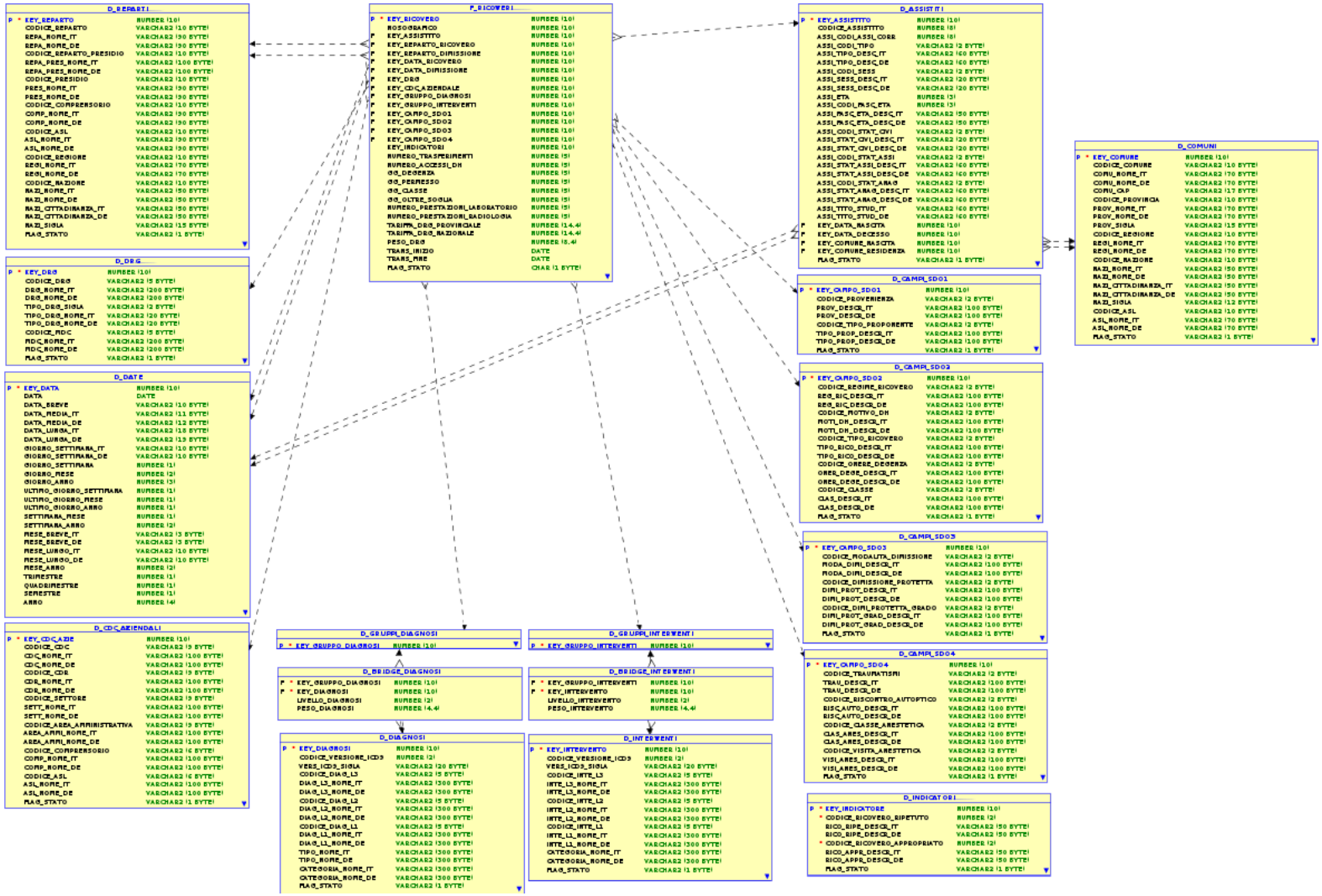
MEDAN DW/9

- Conceptual model of the data cube Hospital_Stay
  - Each event stores a hospital stay of a patient
- Similar model for data cube Services and data cube Transfer



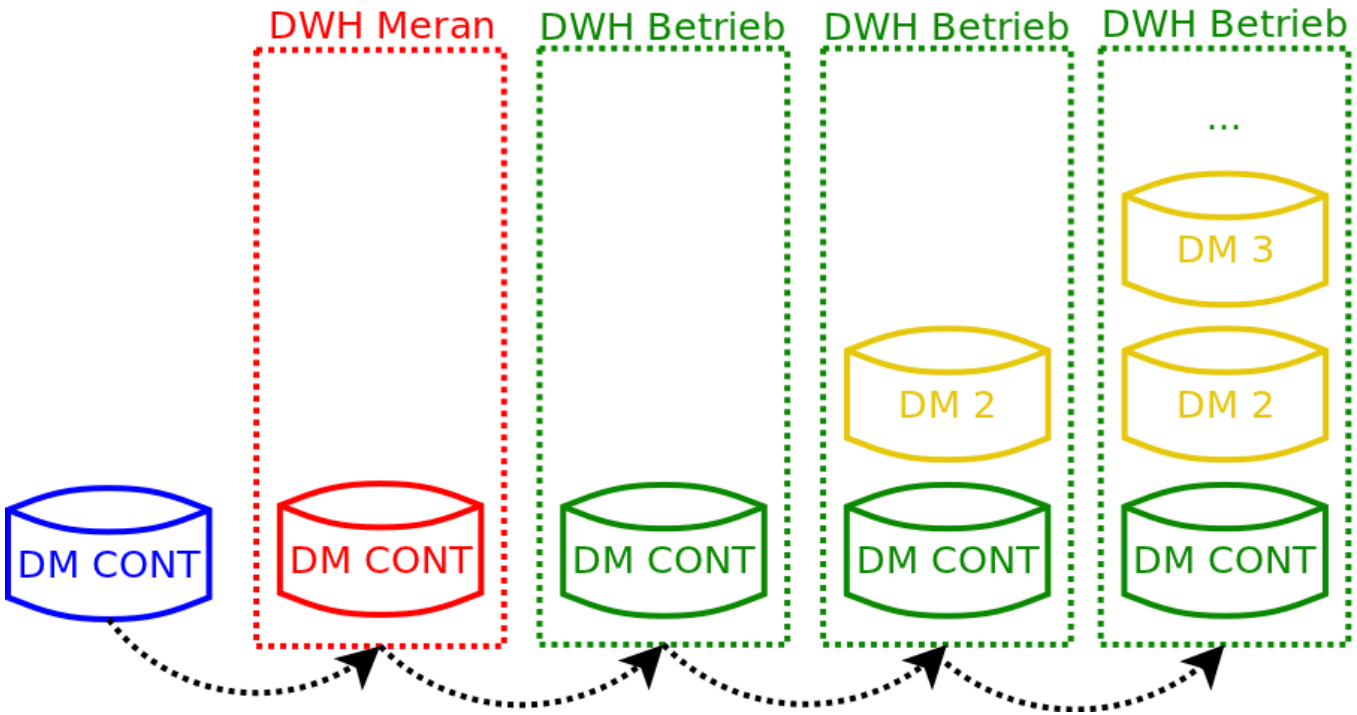
MEDAN DW/10

- Logical model of the data cube Hospital_Stay



MEDAN DW/11

- Bottom-up approach
  - Prototype of DM CONT with 3 cubes (hospital stay, services, transfer) for hospital Meran
  - Deploy DM CONT in hospital Meran, then in other hospitals
  - Repeat the same cycle for other Dms (DM Personnel, DM Pharmacy, DM Laboratory, etc.)



MEDAN DW/12

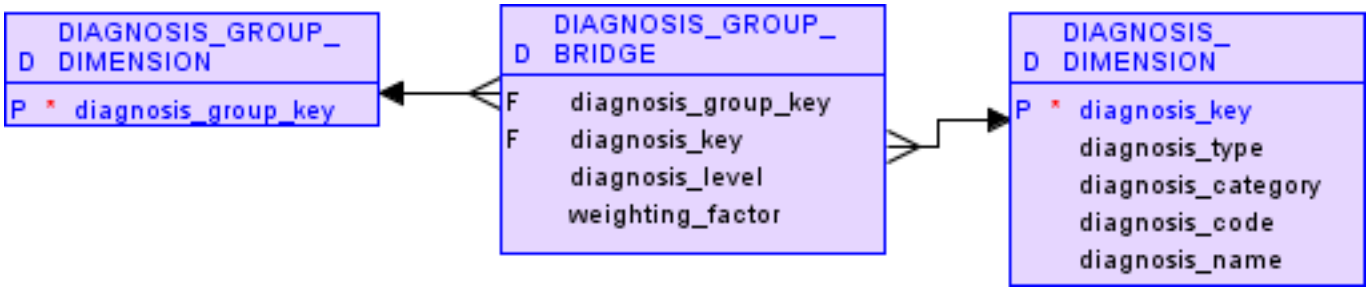
- Multi-valued dimensions
  - Diagnoses and procedures are examples of multi-valued attributes/dimensions
  - i.e., a patient typically has multiple diagnoses (up to > 10)
- Solutions
  - Reserve multiple columns, one for each diagnosis
    - Many empty cells, i.e., sparse fact table
  - Use several facts for a single recovery
    - Increases the number of tuples in fact table
  - Bridge tables

Budgeting_Recovery_Fact
Health_record_numbers (FK)
Health_record_year (FK)
Patient_key (FK)
…
Diagnosis_Key_1 (FK)
Diagnosis_Key_2 (FK)
Diagnosis_Key_3 (FK)
...

Budgeting_Recovery_Fact
Health_record_numbers (FK)
Health_record_year (FK)
Patient_key (FK)
…
Recovery_Number
Diagnosis_Key (FK)
...

MEDAN DW/13

- **Bridge table** for multi-valued dimensions



MEDAN DW/14

- Lessons learned
  - Developing a BI platform is a **process** that takes years
  - A well-designed and consistent DW is the foundation for BI
    - QlikView is a tool for quick analyses; it cannot replace a DW
  - Do not put anything in a **single data mart**
    - Use one DM for one business query (set of closely related business queries)
  - Different opinions on bottom-up vs. top-down, but bottom-up seems to have more acceptance
  - Data modeling is difficult but very important
    - Helps to get a conformed view on the business
    - e.g., What is an admission?
    - Different granularity by different users, e.g., Province, Hospital
  - Establishing a **Business Intelligence Competence Center** (BICC) is crucial
    - Coordinates the BI project
    - Combines business, IT and analytical skills