Universitas Esa Unggul

Smart, Creative and Entrepreneurial

Data Warehouse

Munawar, PhD

Session 10

Data Warehouse Physical Design

Agenda

- What is Physical Design?
- Popular DW Architecture
- Quality Based Physical Design
- Q/A ?

What is Physical Design ?

Physical Design

- Denotes all the problems particularly associated with the appropriateness of selected implementation tools, such as DW architecture.
- The DW design process ends in a physical design with respect to the architectural design of a DW and its association with DMs (Kimball et al, 2008)

Major Technique in DW Design

Top–down design: This design approach focuses on the construction of a centralised DW. A centralised DW is a single DW that supplies the needs of multiple departments by using a single model that satisfies the requirements of many divisions in a corporation

Bottom–up design: Constructing individual DMs is the main strategy adopted in a bottom–up design. Here, a DW can be regarded as the integration of different DMs.

Top Down Design

- In a centralised DW, feedback from departments or user groups can be used to customise requirements.
- Different requirements at different organisational levels are then combined to construct one schema for the entire DW.
- The failure rate of a centralised DW is much higher than that of a DM, particularly for organisations with limited budgets and resources.

Top Down…

- Powerful fundamental architecture that encourages centralised storage and control should be the foundation of the first approach.
- With this approach, a proof-of-concept system is difficult to develop because scope tends to be very broad and time consuming

Bottom Up Design

- DW is a collection of integrated DMs (Kimball and Ross, 2013), where each DM is dedicated to study of single subject in the organisation.
- Building a full-scale DW is very expensive and time consuming (Dimantini and Potena, 2012). The most important reason to choose this approach is faster and cheaper (Dimantini and Potena, 2012).

Bottom Up …

- A bottom–up technique is less risky than a top–down approach. Nevertheless, although the former can be more rapidly deployed and is more flexible, it may create redundancies and is difficult to integrate because DMs tend to provide a narrow perspective of corporate data.
- Organisations that opt for bottom–up approaches should be thoroughly aware of these problems and should ensure DM conformity to their business architecture; such conformity should be achieved following Kimball's recommendations (Kimball et al, 2008).

Bottom Up…

- With limited resources, a proof-of-concept system can be developed relatively fast, but it still satisfies the requirements of a specific user group or department.
- In this approach, only one DM can be constructed at a time because constructing all DMs in a single project is a highly demanding endeavour. Once the first DM is constructed, it can be used as a foundation on which other DMs are built in incremental development

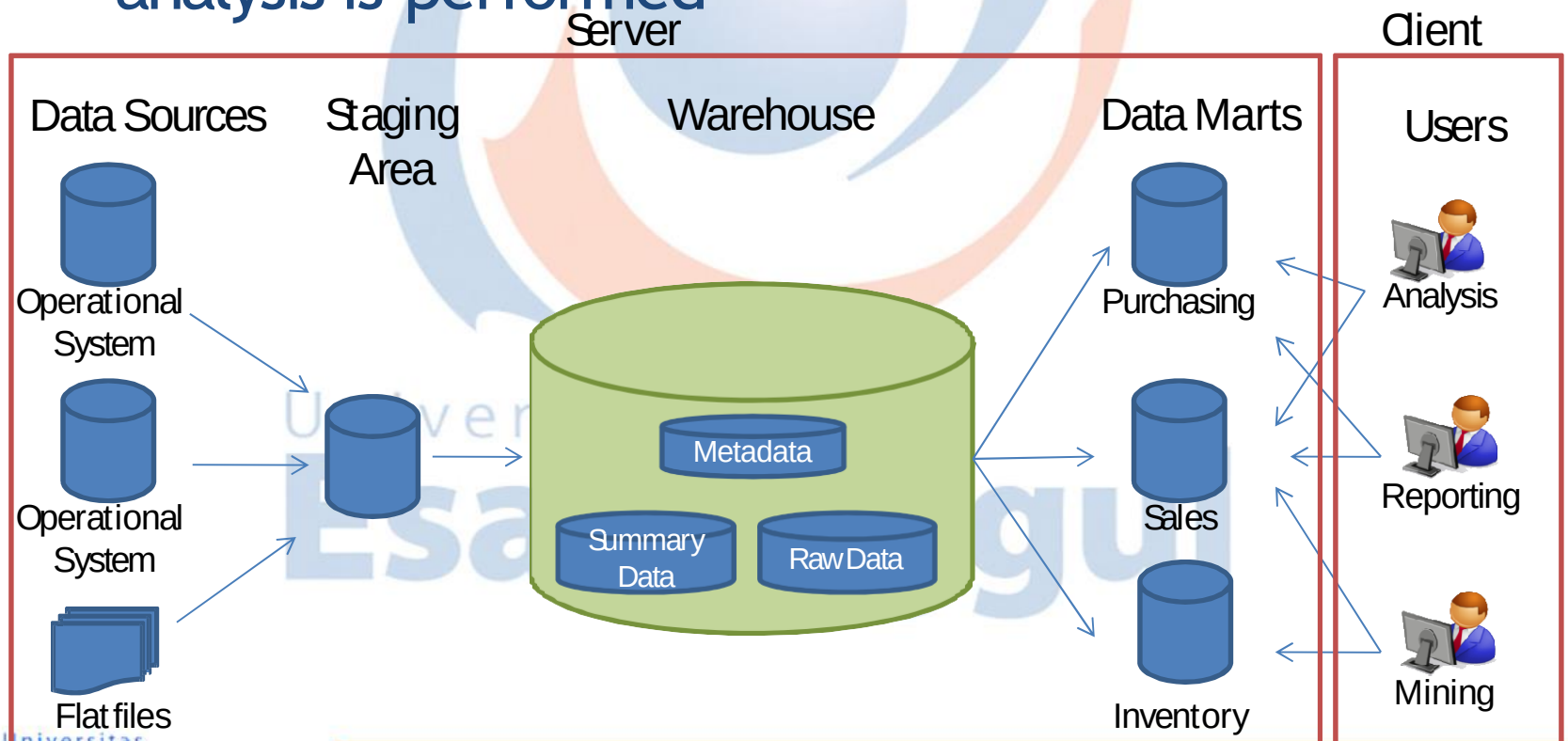
Popular DW Architecture

Architecture in Practice

- Popular DW architectures in practice
  - Vertical tiers
    - Generic Two-Tier Architecture
    - Three-Tier Architecture
  - Horizontal tiers
    - Independent Data Mart
    - Dependent Data Mart
    - Logical Data Mart

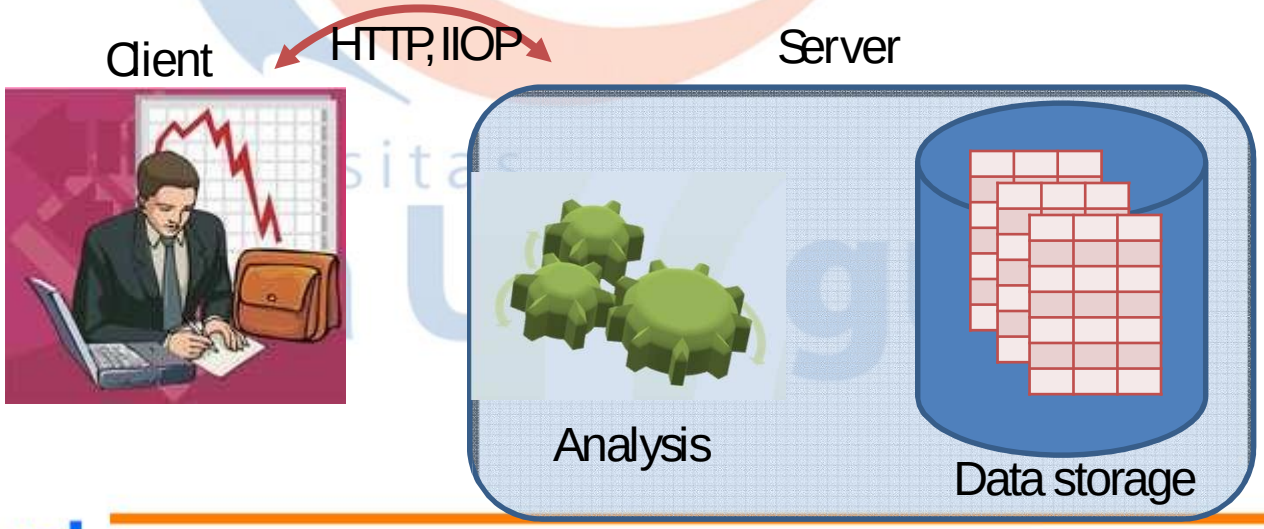
Two Tier Architecture

- Generic client-server architecture
  - Fat or thin client depending on where the data analysis is performed



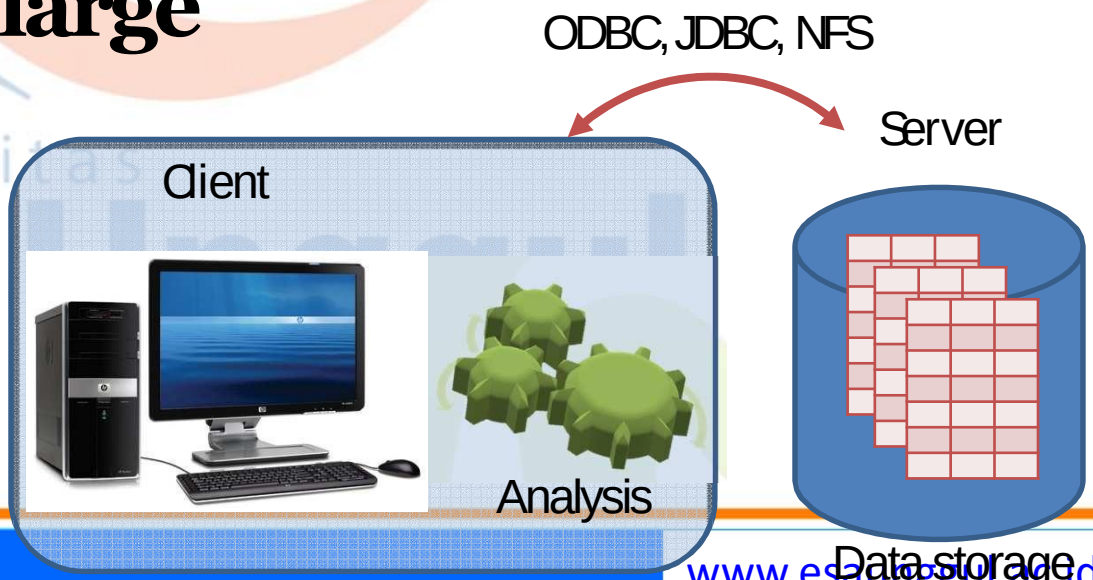
Thin Client

- Operations are executed on the server
- The client is just used to display the results
- This architecture fits well for Internet DW access



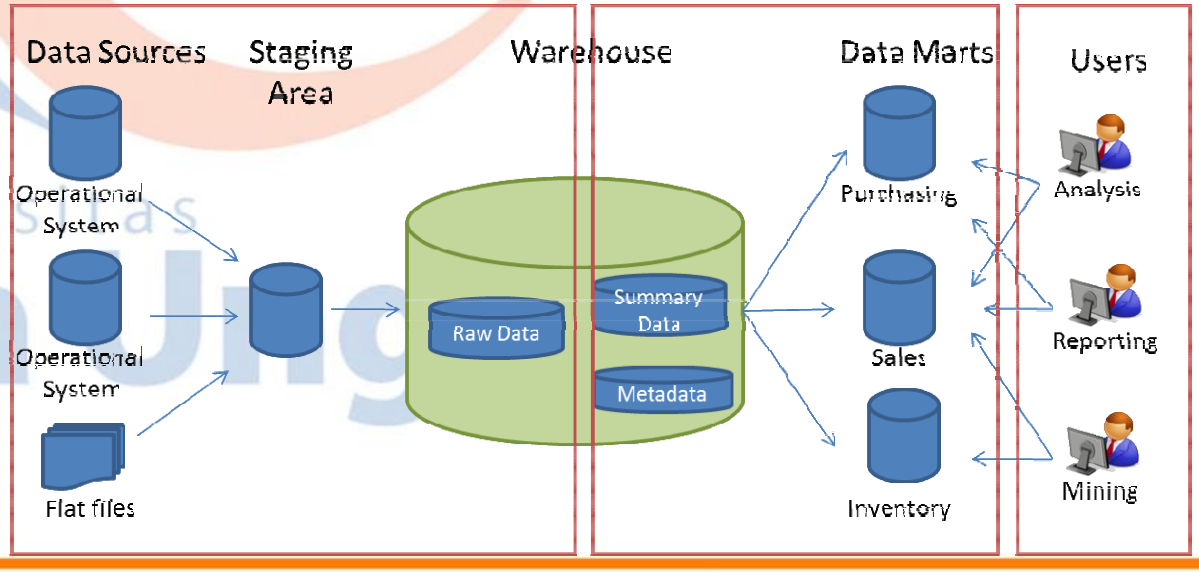
Fat Client

- The server just delivers the data e.g. the corresponding data mart
- Operations are executed on the client
- Communication between client and server must be able to sustain **large data transfers**



Three Tier Architecture

- Tier 1: raw and detailed data intended to be the single source for all decision support
- Tier 2: derived data that had been aggregated for DSS support
- Tier 3: reporting and analysis

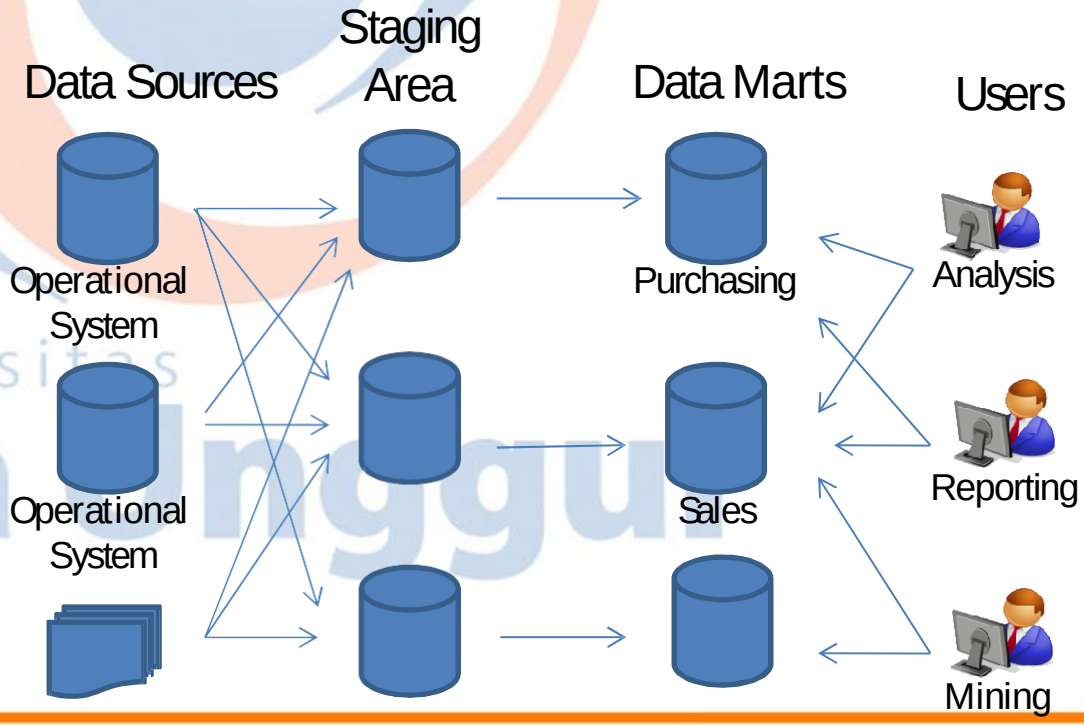


Other Architecture

- N-Tier Architecture
  - Higher tier architecture is also possible but the complexity grows with the number of tier-interfaces
- Web-based Architectures
  - Advantage: Usage of existing software, reduction of costs, platform independence
  - Disadvantage: Security overhead e.g. data encryption, user access and identification

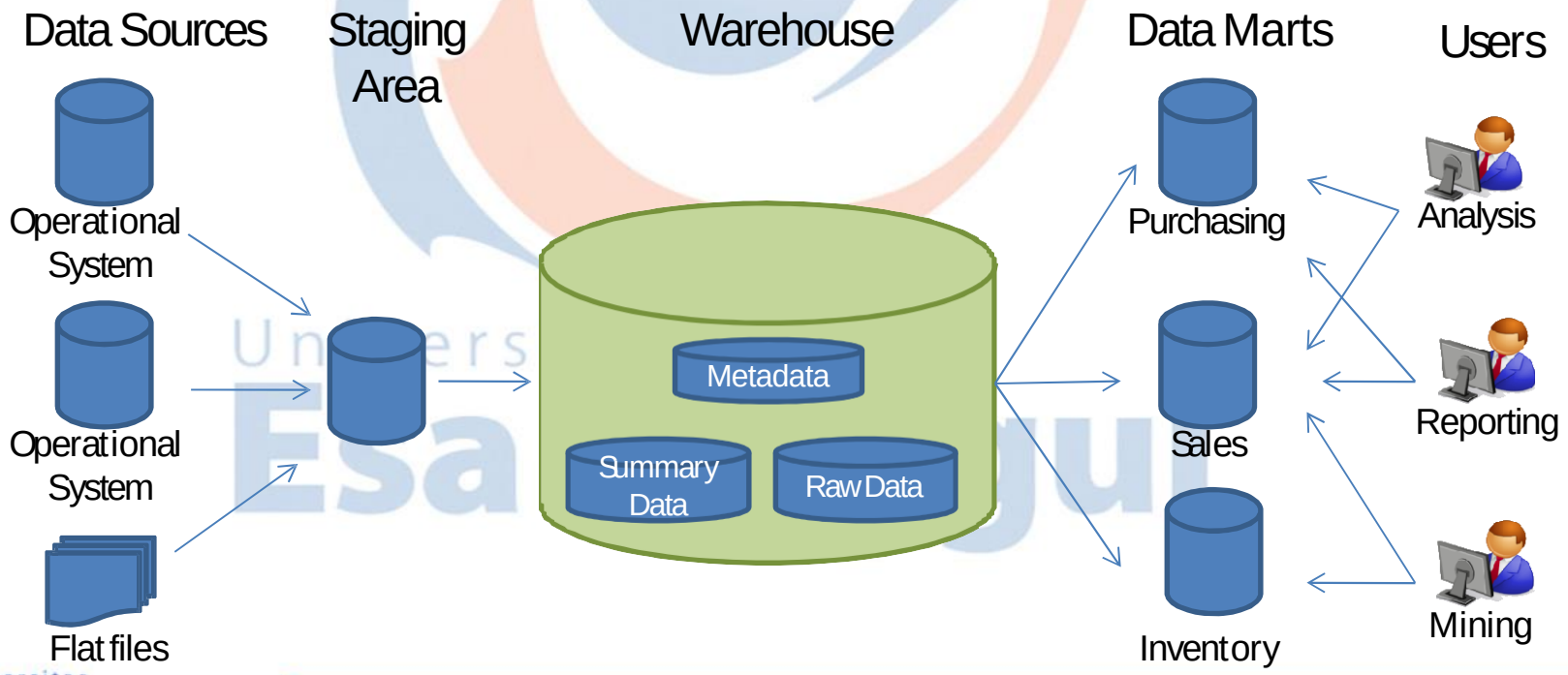
Independent Data Mart

- Mini warehouses – limited in scope
  - Faster and cheaper to build than DWs
- Separate ETL for each independent Data Mart
  - Redundant processing for each mart



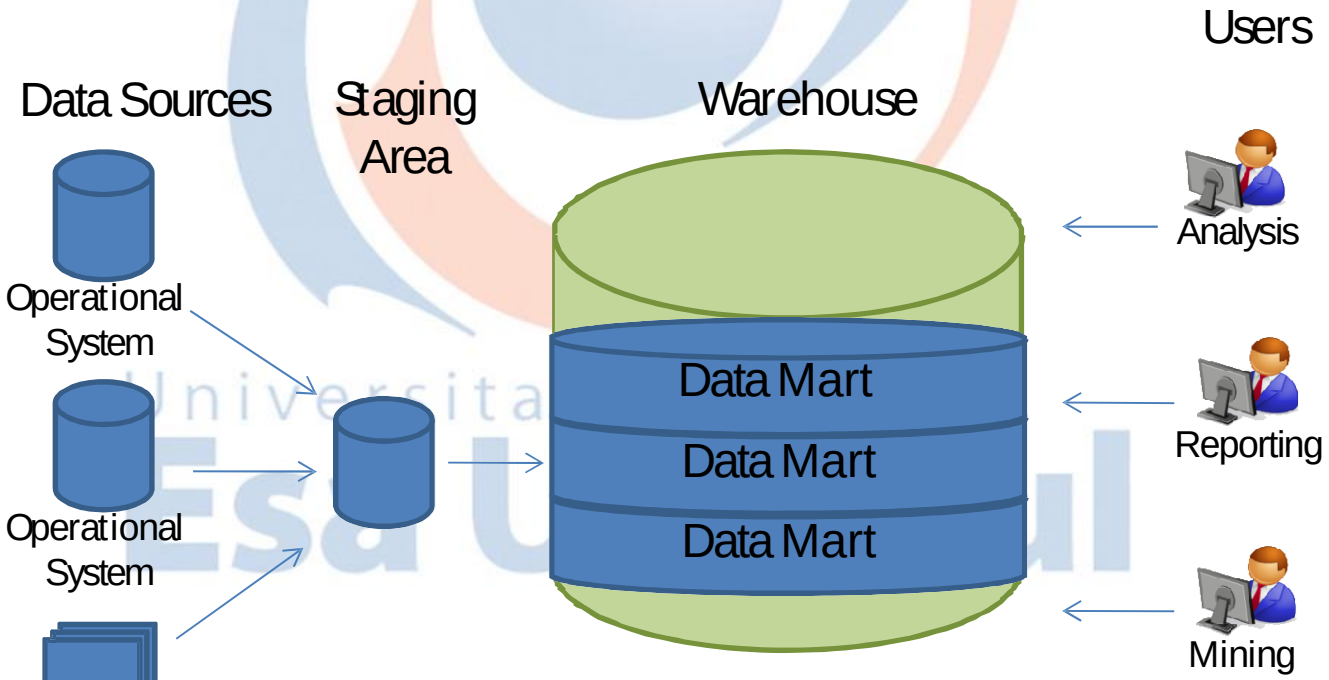
Dependent Data Mart

- Single ETL for the DW
  - No redundancy in the ETL process
- Data Marts are loaded from the DW



Logical Data Mart

- Data Marts are *not* separate databases, but logical views of the DW
  - Integrated view of the enterprise



DW vs Data Mart

Scope	
DW	**Data Marts**
Application independent	Specific DSS application
Centralized,	Decentralized by user area
Planned	Organic, possibly not planned

Data	
DW	**Data Marts**
Historical, detailed, summarized	Some history, detailed, summarized
Lightly denormalized	Highly denormalized

Subjects	
DW	**Data Marts**
Multiple subjects	One central subject

Sources	
DW	**Data Marts**
Many internal and external sources	Few internal and external sources

Other characteristics	
DW	**Data Marts**
Flexible	Restrictive
Data-oriented	Project oriented
Long life	Short life
Large	Start small, becomes large
Single complex structure	Multiple, semi-complex structure, together complex

Data Marts

- The first step to construct a DM is identifying all the facts to be placed in the DW (Golfarelli et al, 1998).
- The collection of facts required by the organisation will evolve over time, and can be in years to be completed.
- Kimball and Caserta (2004) recommends constructing the first DM that represents the minimum effort and risk. But, there is no detailed guidance to determine the priority of DM to be built

Data Marts…

- Many autonomous DMs will be developed over years in a DW project.
- The next significant area of concern in building enterprise DWs is integrating heterogeneous DMs.
- The problem of integrating heterogeneous DMs lies in identifying compatible dimensions

Data Integration in Data Mart

- Integration with dimensions sharing (Kimball, 2002) → shareable dimensions gives us the change to integrate the DM through the same dimensions in both of DM
- Integration with dimensions compatibility (Chhabra and Pahwa, 2014) → two dimensions in different DM can be said compatible when their common information is consistent
- Integration with generalization (Abello, Samos, and Saltor, 2002) via drill-across query even though the dimensions are not same

Centralized vs Distributed

- DW may be **centralized** or **distributed**
- Centralized DW (e.g. Volkswagen)
  - Analytical queries are run only at the main enterprise location - no need to transport data via network
  - High costs for large dedicated hardware
- Distributed DW (e.g. WalMart)
  - More natural form due to corporations being active all over the world and having different types of hardware and software
  - Higher overhead but lower cost

Distributed DW

- Types of **distributed DW**
  - **Geographically** distributed
    - Local DW/global DW
  - **Technologically** distributed DW
    - Logically one DW, physically more DW
  - **Independently evolving** distributed DW
    - Uncontrolled growth

Distributed DW…

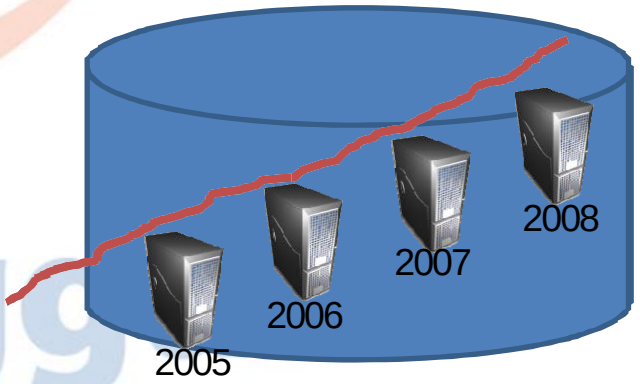
- **Geographically** distributed
  - In the case of corporations spread around the world
    - Information is needed both **locally** and **globally**
  - A distributed DW makes sense
    - When much processing occurs at the local level
    - Even though local branches report to the same balance sheet, the local organizations are somewhat autonomous

Distributed DW…

- **Technologically** distributed DW
  - Placing the DW on the distributed technology of some vendor
  - Advantages
    - Entry costs are cheap - large centralized hardware is expensive
    - No theoretical limit on how much data can be placed in the DW -new servers can be added to the network on demand

Distributed DW…

- As the DW starts to expand network **communication** starts playing an important role
  - Example: Let's simplify and consider we have 4 nodes each holding data regarding a specific year
  - Now let's consider a query which needs to access data from the last 4 years
  - Large amount of data has to be shipped to processing units

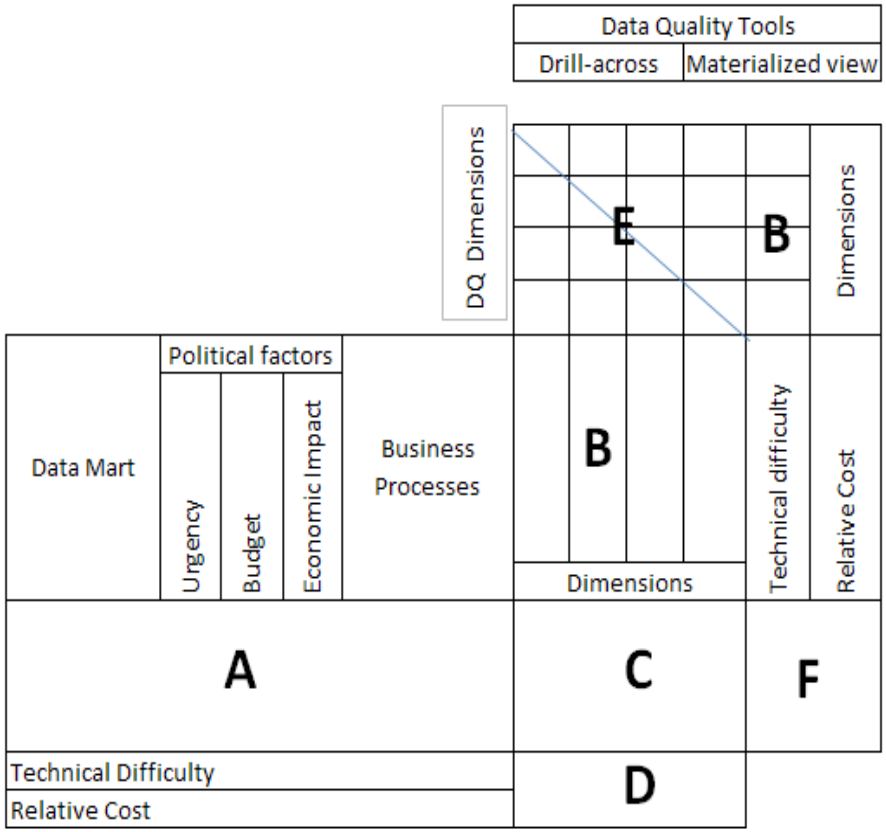


Distributed DW…

- **Independently evolving** distributed DW
  - In practice there are many cases in which independent DW are developed concurrently in the same organization
    - The first step in many corporations is to build a DW for financial or marketing
    - Once this is successfully set up, other parts of the organization follow independently

Quality Based Physical Design

Quality Based Physical Design



Area A →the DMs to be built and their political influence on the decision to develop a DW.
Area B → the dimensions required to support business processes
Area C → the mapping of the dimensions required in every DM
Area D → the technical difficulty and relative cost of constructing a dimension
Area E → the relationship between dimensions
Area F → the justification for ranking a DM construction strategy

Quality Based…

To illustrate the concept, it was applied in an academic affair of a private university

Data Quality Tool	
Drill-across Queries	Materialized View

Information Content / Information Access / Data Quality Dimensions	Dimensions
	Time
	Department
	Faculty
	Financial
	City
	Basis
	School of Origin
	Event
	Student
	Payment Scheme
	Curriculum
	Employee
	Subject
	Room
	Lecturer

Data Mart	Political Factors: Total	Urgency	Budget	Economic Impact	Business Processes	Time	Department	Faculty	Financial	City	Basis	School of Origin	Event	Student	Payment Scheme	Curriculum	Employee	Subject	Room	Lecturer	Technical difficulty	Relative Cost	Total
						Dimensions																	
Enrolment	28	10	8	10	Admision	X	X	X	X	X	X	X	X	X	X	X	X				5	15	
					Test	X	X	X													10	15	
					Registration	X			X					X	X						7	15	
					Reward Calculation	X	X	X	X		X	X		X	X		X				10	20	
																					32	**65**	**97**
Registration Semester	26	5	15	6	Course Offer		X	X			X					X		X			4	8	
					Scheduling		X	X			X					X		X	X		12	25	
					Registration Semester	X					X			X	X			X	X		4	7	
					Student Billing	X	X	X	X		X			X	X						5	10	
					Student Payment	X	X	X	X		X			X	X						5	10	
					Class Allocation	X	X	X			X			X	X	X		X	X	X	12	25	
					Lecturer Allocation	X	X	X			X					X		X	X	X	5	10	
																					47	**95**	**142**
Class Administration	29	4	20	5	Time Sheet	X	X	X			X			X				X	X	X	7	10	
					Teaching Invoicing	X	X	X	X		X							X		X	10	15	
					Teaching Payment	X	X	X	X		X			X				X		X	12	25	
					Exam Scheduling	X	X	X			X							X	X	X	8	17	
					Exam Grading		X	X			X			X				X		X	7	13	
																					44	80	**124**
Quality Assurance	9	1	5	3	Transcript Generation		X	X			X			X		X		X		X	5	8	
					Outstanding Billing Process		X	X	X		X			X	X						9	20	
																					14	**28**	**34**
Technical Difficulty						1	3	2	7	2	3	2	3	10			3	5	4	4			
Relative Cost						1	3	2	10	2	5	2	4	20			3	15	10	8			

Quality Based…

- From a technical aspect, time is the easiest dimension to address, whereas student may be the most difficult. The student dimension may be sourced from many tables in many databases that have evolved over several decades.
- The relative cost of building dimensions correspond to many elements, such as technical difficulty, number of related tables, joining or matching complexity and transformation complexity

Quality Based …

- A high total political factor (urgency, budget and economic impact) and low total cost can be used to rank the DM that should be constructed.
- Although the cost and technical difficulty of the Enrolment DM (total, 97) is higher than those of the Quality Assurance DM (total, 34), the political factor for the Enrolment DM (total, 28) is higher than that of the Quality Assurance DM (total, 9). Therefore, the combination of relative cost, technical difficulty and political factor suggests that the Enrolment DM (total, 125 from 28+97) can be ranked number one and be the first to be constructed.
- The second DM is likely the Class Administration DM. It represents lower technical challenge and cost (total, 124) than does the Registration Semester DM (total, 142). The political factor of the Class Administration DM (total, 29) is higher than that of the Registration Semester DM (total, 26). Therefore, the former can be the second DM to be constructed

Thank You…