Data Warehouse
Conceptual Design

Munawar, Ir. MMSI. M.Com, PhD

Conceptual Design

- Is intended to derive an implementation-independent and expressive conceptual schema for a data mart (DM) or DW
- It accommodates a high degree of abstraction in representing the process and architecture of a DW in all facets involved and is intended to realise independent implementation
- It allows having closer ideas about the ways that a user can perseive an application domain (Saxena and Agarval, 2014).
- It enables developers to describe the requirements for DW development from a user's perspective

Conceptual Design Process

DW Dev Phase	Input	Processes	Quality drivers	Tools	Deliverables
Conceptual Design → to abstract the users' request to some information structures, which act as the bridge connecting the real world and the machine world					
• **Multidimensional Modeling**	• Fact • Preliminary workload	• Identifying the fact of interest • Identifying the dimensions hierarchies • Identifying measures • Identifying aggregations	• Comprehensiveness • Currency • Speed	• ME/R • ER Model • DFM • UML class diagram	Dimensional Schema Dimensional schema is designed to store data in a way that: • Emphasizes understandibility • Enhances query performance • Accommodates change

Basic Concept of Conceptual Design

- A fact is a collection of data items related to business transactions or represent business items. A facts consist of measures and context data.
- A dimension is a collection of data related to one business dimension. Contextual background for the facts are defined by the dimensions; parameters to perform OLAP are also defined by dimensions.

Basic Concept of Conceptual Design

- A measure is a numerical attribute of a fact. Performance or behaviour of the business can be represented by a measure relative to the dimensions. An essential decision in a measure definition is the lowest level of detail (sometimes called the grain) in order to determine the type of analysis that can be performed.
- Aggregation is pre-calculated summaries of data came from the most granular fact table. Aggregation is applied in the case when the analysis needs computation through a number of dimensions and lots of rows of each dimension to calculate metrics of fact table. Query performance can be improved using aggregate fact tables without increasing overall storage space.

Tools for Conceptual Design

- Entity/relationship-based (E/R-based)
- Object-oriented
- Ad hoc models (Sen and Sinha, 2005).

Benefits of ER Extensions

- E/R has been tested for considerable time (years);
- E/R is a commonly used tool by many designers;
- a variety of application domains can be flexibly adapted by E/R;
- substantial research results have been derived for E/R (Sapia et al, 1999; Tryfona et al, 1999).

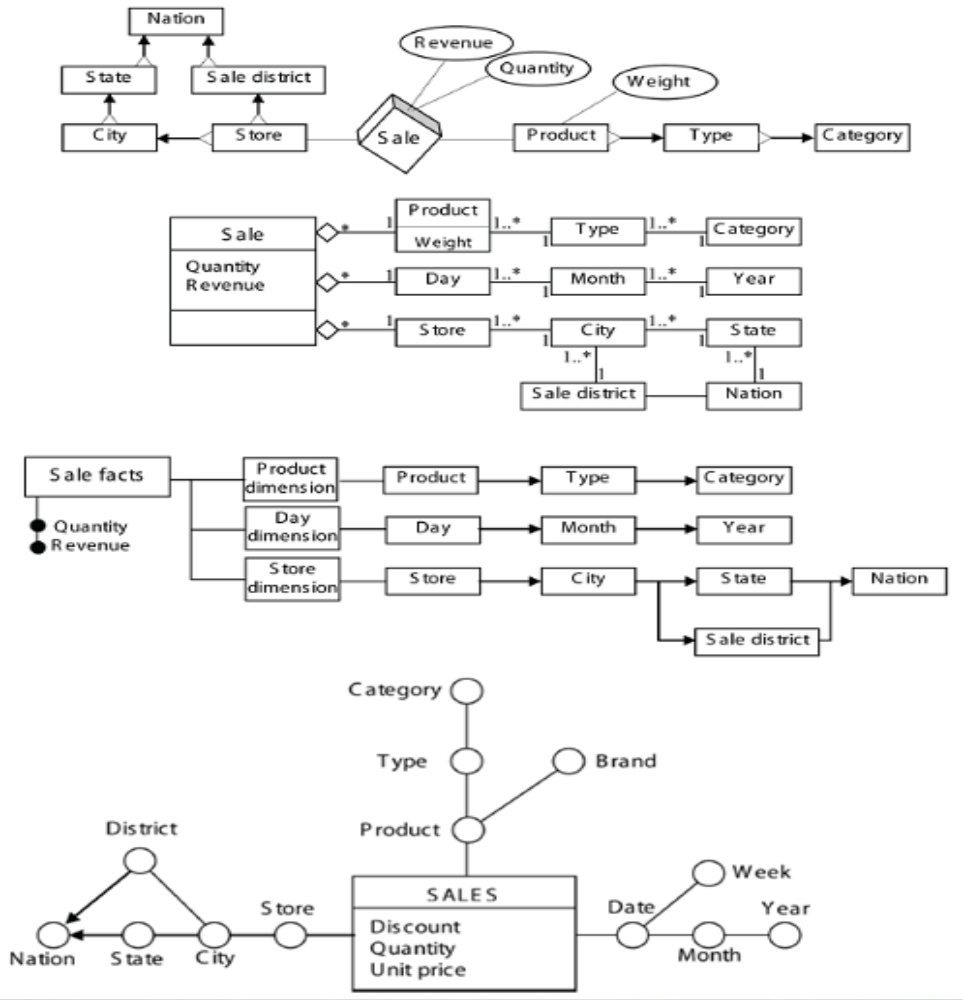
Benefits of OO Models

- the static and dynamic properties of information systems can be better represented with these models;
- requirements and constraints can be expressed in a powerful mechanism;
- data modelling is currently dominated by object-oriented approaches;
- UML, in particular, is a standard and is extendable (Lujan-Mora et al, 2002).

Benefits of Adhoc Models

- notational economy is more effectively achieved;
- specific multidimensional modelling can be appropriately emphasised, thereby making ad hoc models
- the most intuitive representations and the most readable by non-expert users (Rizzi, 2009).

Sales Fact Model



- M E/R (Sapia et al, 1999)
- UML class diagram (Lujan-Mora et al, 2002)
- fact schema (Husemann et al, 2000)
- dimensional fact model (DFM) (Golfarelli, 2010)

Facts Identification

Fact identification is the most difficult task in the DW design process and is commonly done manually. Some techniques that can be used to find facts are as follows:

- A fact table of a star schema can be derived from the many-to-many relationships in an E/R model that contains numeric and additive non-key facts (Kimball, 1997).
- Candidate measures can be found in business queries for data items that indicate business performance (Ballard et al, 1998).
- The most frequently updated entities can be identified as facts (Golfarelli et al, 1998).
- Fact properties can be summarised (or aggregated); thus, they are usually found in numerical data (Tryfona et al, 1999).

Dimensional Rule Mapping

UML diagram components	Snowflake schema
UML classes	Fact and dimension tables
Aggregation classes	Dimension hierarchies
Class attributes	Measures/dimension attributes
Generalisations	Aggregation levels

Identify Dimension Hierarchies

Data in dimensions should be organised into hierarchical levels (Agrawal et al, 1997).

Navigation path for drilling up and drilling down can be defined by a hierarchy.

- A *simple* hierarchy contains precisely one linear aggregation path within a dimension (e.g. path day → month → year in dimension time).
- A *multiple* dimension hierarchy consists of at least two different aggregation paths in a dimension (e.g. path account can be account → customer and account → organisation in a dimension account).

Identify Measures

- Measures are properties of fact collected about business operation which calculations (e.g. sum, count, average, minimum, maximum) can be made (Kimball, et. all, 2008).
- Measures are organized by dimensions.

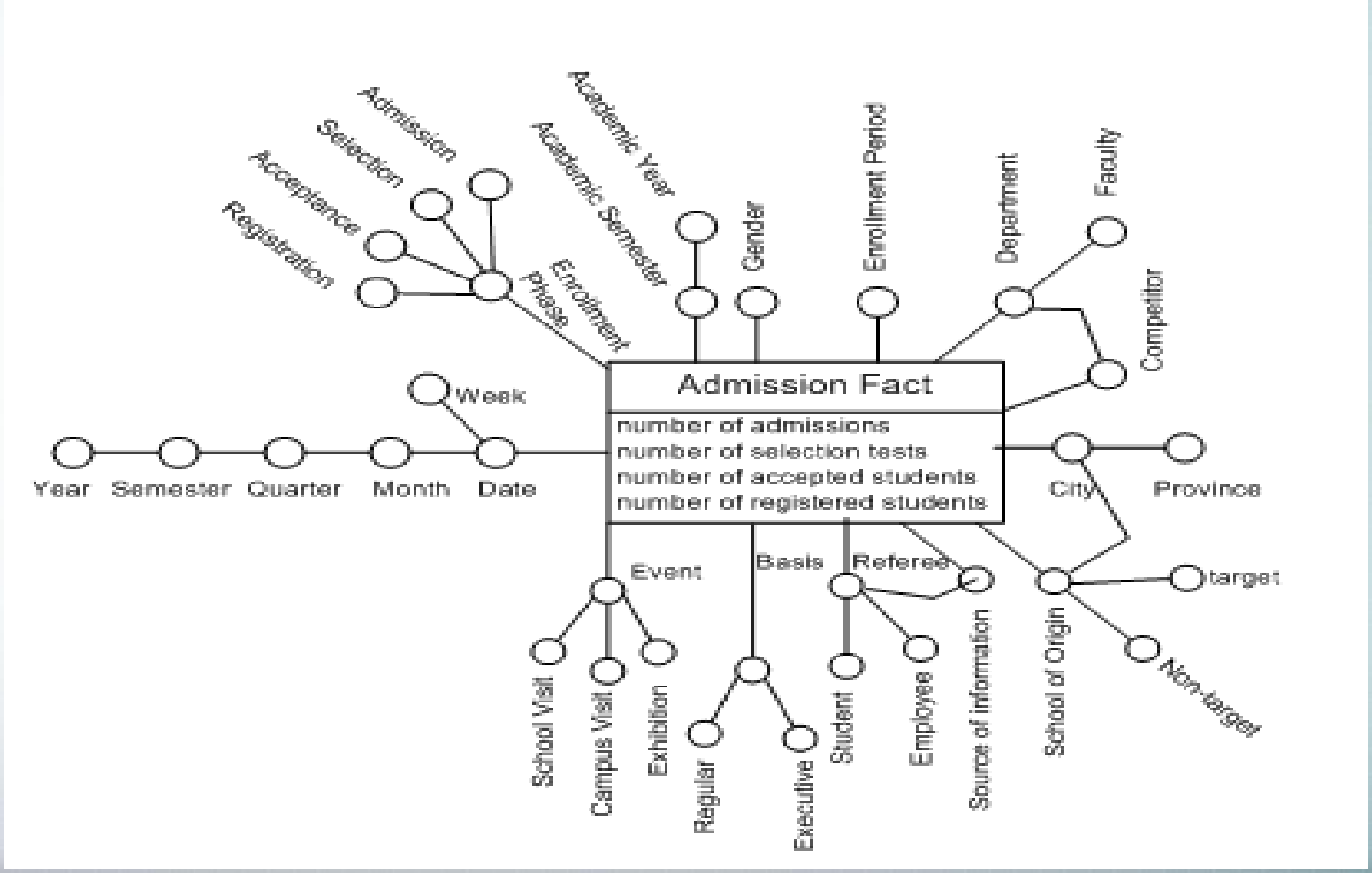
Identify Aggregation

- Aggregation is the central means to summarize and condense the information contained in the various sources (Cabot, et.all, 2010).
- *Summarisability* (the guaranteed correctness of aggregation results) is an essential requirement for OLAP queries. Consequently, any multidimensional schema should be arranged so that summarisability is available at the highest possible level. Furthermore, if summarisability is violated along certain aggregation paths, then a schema should explicitly explain this constraint.

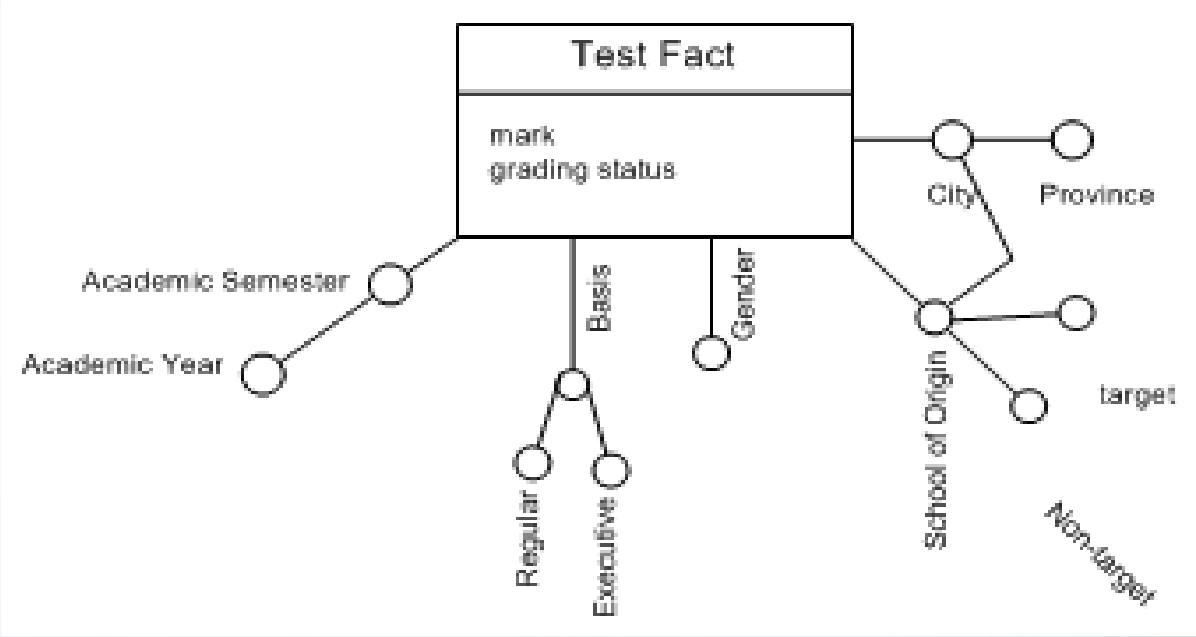
Practical Evidence

To illustrate the proposed model, a case study on the student admission process that is specifically related to marketing activities was conducted. A private university in Jakarta intends to build a monitoring system for student admissions. A series of related marketing activities have been carried out to increase student enrolment. Improvements in decision making related to the admission system is the expected benefit from the implementation of a DM.

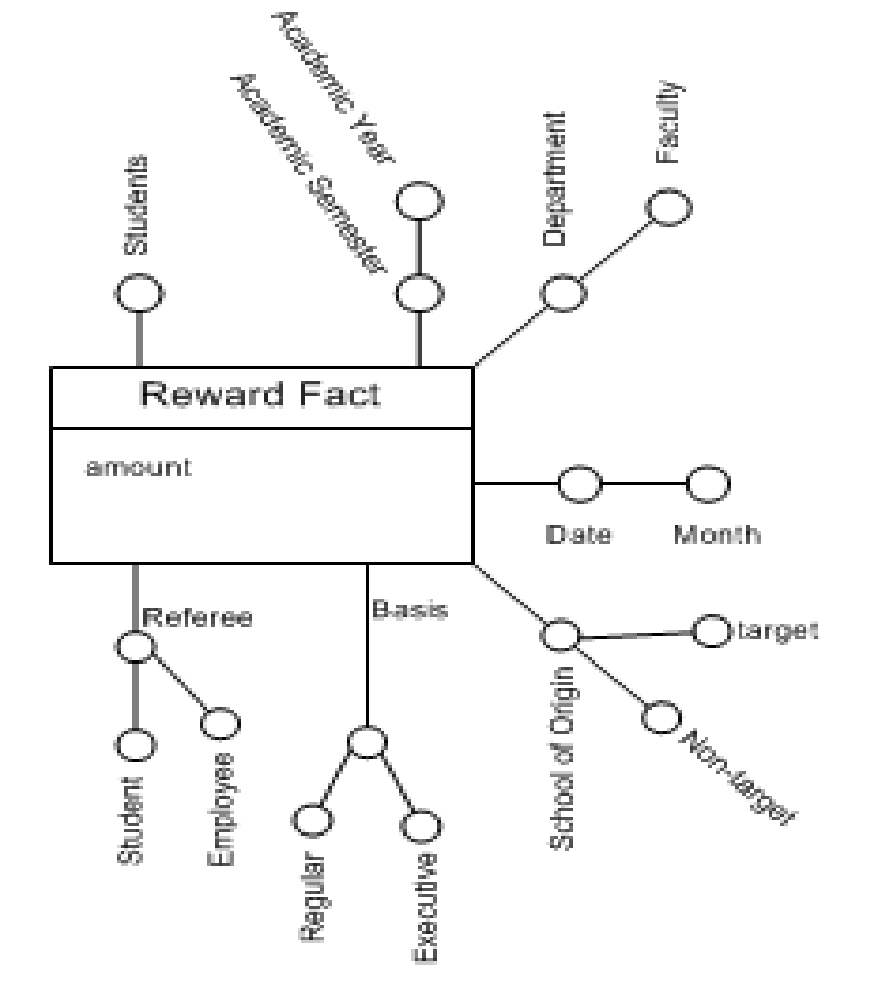
DFM for Student Admission



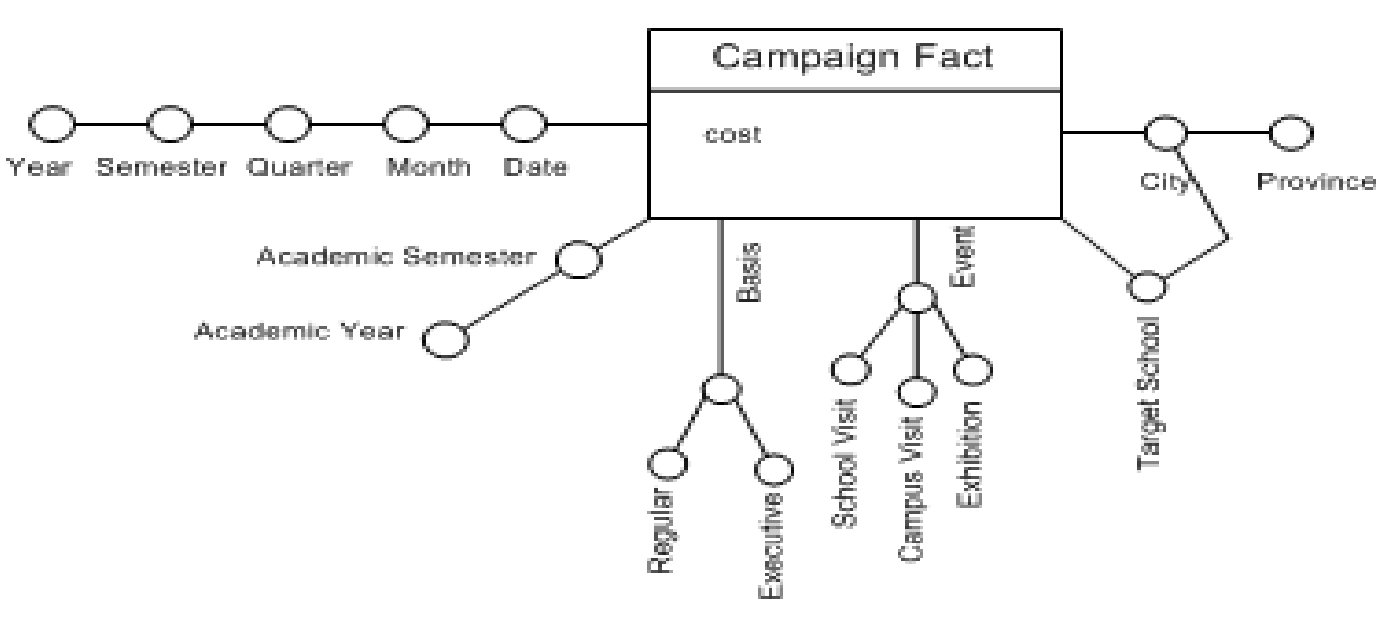
DFM for Student Admission ...



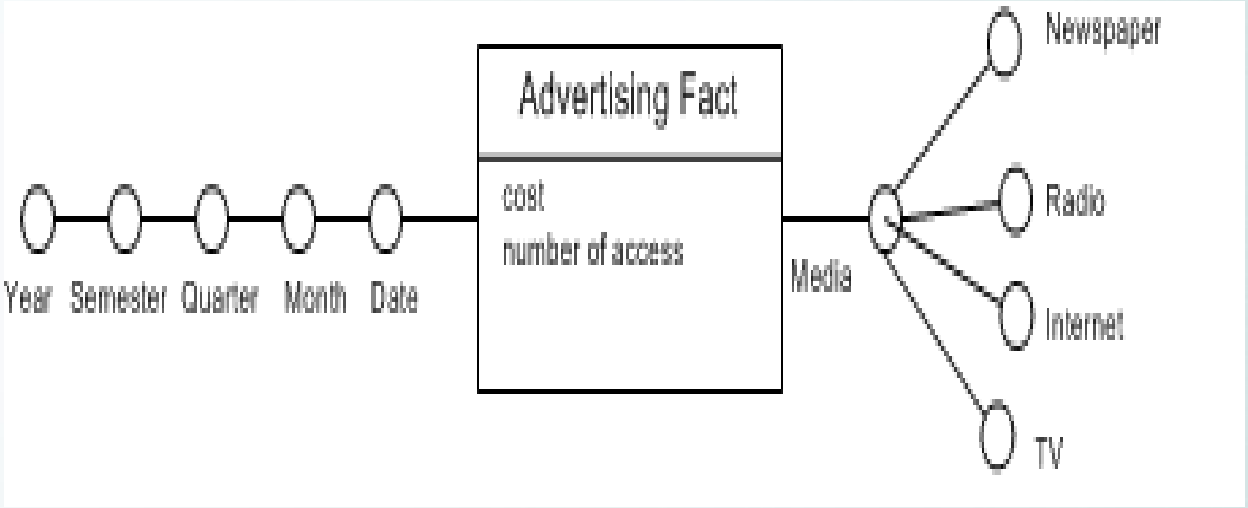
DFM for Student Admission …



DFM for Student Admission ...



DFM for Student Admission ...



Thank You !

Munawar, PhD