







Supply-Demand Matrix: A Process-Oriented Approach for Data Warehouses with Constellation Schemas

Luís Cavique¹  , Mariana Cavique² , and Jorge M. A. Santos³ 

¹ Universidade Aberta, BioISI-MAS, Lisbon, Portugal
luis.cavique@uab.pt

² Universidade Europeia, Lisbon, Portugal
mariana.cavique@universidadeeuropeia.pt

³ Universidade Évora, CIMA, Évora, Portugal
jmas@uevora.pt

Abstract. Star schema in data warehouses is a very well established model. However, the increasing number of star schemas creating large constellations schemas add new challenges in the organizations. In this document, we intend to make a contribution in the technical architecture of data warehouses with constellation schemas using an extension of the bus matrix. The proposed supply-demand matrix details the raw data from the original databases, describes the constellation schemas with different dimensions and establishes the information demand requirements.

Keywords: Denormalization forms · Data warehouses · Constellation schemas · Process oriented

1 Introduction

To build a data warehouse, two types of architecture can be found: the Inmon architecture (Inmon 2005) and the Kimball architecture (Kimball and Ross 2013). In a historic perspective, Inmon coined the term ‘data warehouse’ in 1990 and in 1996 Kimball published the first edition of the Data Warehouse Toolkit (Breslin 2004).

On one hand, Inmon strategy advocates a top-down approach which begins with the corporate data model. On the other hand, Kimball’s architecture uses a bottom-up approach based on the dimensional modeling, where the fundamental concept is the star schema. Most of the companies adopt Kimball’s strategies, given the reduce costs of creating a star schema, but aspire a corporate model with Inmon’s design. A detailed document reporting similarities and differences of the two methods can be found in Breslin (2004).

Since Kimball’s strategy is supported by the development of different data marts, by distinct teams, it risks losing the integrated vision of the organization. In this work we choose the bottom-up data warehouse approach and extend the study to the constellation schema, since most of the bibliography focuses only on the star schema (Shin and Sanders 2006; Caldeira 2012; Santos and Ramos 2017).

The goal of this paper is to develop a systematic procedure to transform more than one database into a constellation schema, given a set of requirements. This work defines data suppliers and information consumers and balances the supply and the demand of information flow.

The paper is organized in four sections. In Sect. 2, related work is presented. Section 3 presents the supply-demand matrix with a running example. Finally, in Sect. 4, some conclusions are drawn.

2 Related Work

In this document, we develop a procedure to support database denormalization and integration in a fact constellation schema of a data warehousing. In this section, first, we present a way to differentiate types of tables in a database. Then, we introduce a database denormalization process (Cavique et al. 2019). The bus matrix (Kimball and Ross 2013) and its extensions are reviewed. Finally, some aspects of technical architecture are reported.

2.1 Types of Tables

In the database denormalization process it is important to differentiate the types of tables in a database based on their relationships.

We reuse the work of Cavique et al. (2019) which identify three types of tables, using the following nomenclature, as shown in Fig. 1:

- lookup tables for tables only with cardinality equal to 1,
- intermediate tables for tables with cardinality 1 and N, and
- fact tables for tables only with cardinality equal to N.

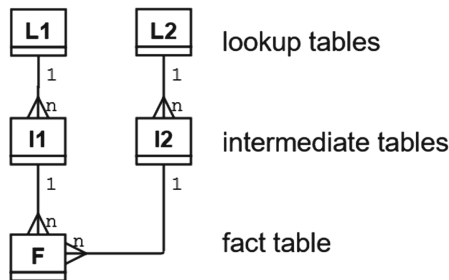


Fig. 1. Lookup, intermediate and fact tables

In this work we also draw all database tables with relation 1:N with the following rule - the table with a single line is drawn on top, while the table with multiple lines is drawn underneath.

In similar approaches, like modeling agile data warehouse with Data Vault (Linstedt and Graziano 2011) the authors also found three different types of structures: hubs, links and satellites.

A fact table in a database corresponds also to a fact table in a data warehouse, for that reason we use the same name.

2.2 Top-Down Database Denormalization Process

Cavique et al. (2019) present a top-down database denormalization process with two denormalization forms.

In the First Denormalization Form (1DF) given a database schema, in order to avoid multiple paths for the same query, a split strategy is applied aiming to find a poly-tree structure. In Fig. 2.a, in order to avoid multiple paths (L1-I1-F1-L2 and L1-I2-F2-L2) table L1 is duplicated and a poly-tree is found.

In the Second Denormalization Form (2DF) given a poly-tree the goal is to find for each fact table its own tree. In Fig. 2.b the poly-tree is divided into two trees which roots are F1 and F2.

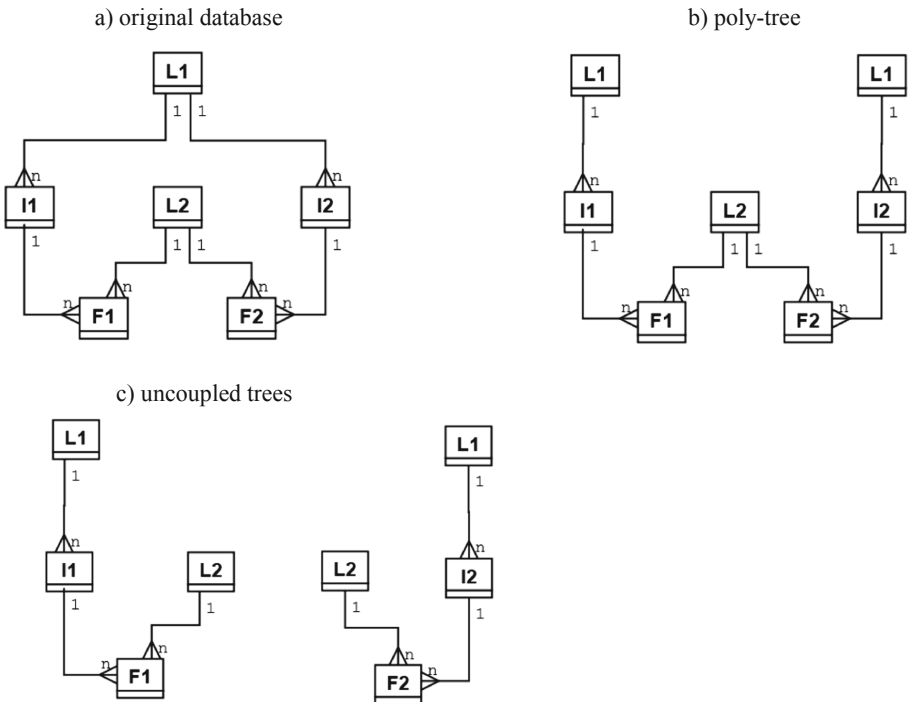


Fig. 2. Denormalization process from the original database to uncoupled trees

2.3 Bus Matrix Evolution

In process-based management, where processes are transversal to departments, the organizations can be represented by a matrix with processes versus departments.

The business process matrix, also called bus matrix (Kimball and Ross 2002), combines the business process with dimensions of the dimensional data model. The bus matrix is usually represented as processes versus dimensions.

The bus matrix is oriented to a single star schema. The complexity of the organizations leads them to store many star schemas or constellations. In order to show a constellation, the bus matrix can evolve in the dimensional data model and includes the fact tables instead of process (Shahzad and Sohail 2009). The constellation matrix is represented as facts versus dimensions, as represented in Fig. 3.

business process	fact table	dimension	dimension	dimension	dimension
		1	2	3	4
Process X	A	X			X
	B		X	X	
	C				X

Fig. 3. Constellation matrix

2.4 Technical Architecture

The technical architecture of data warehouse is proposed by Kimball and Caserta (2004) where the concepts of back-room and front-room are proposed.

The back-room corresponds to the data management, in particular the sources sub-systems and the staging area. The staging area is divided in two groups: (i) the ETL process of extracting, cleaning, conforming, and delivering data, and (ii) the storage of the dimensional tables ready to delivery atomic or aggregate data.

The front-room corresponds to the presentation area, where the user’s community is able to browse and analyze data, using standard reports or ad-hoc queries.

The back-room and front-room work out like two separated data silos. In our work we propose a process from data source to data presentation in order to avoid redundancy or lack of information.

3 Proposed Model

In this section we develop a procedure to find the supply-demand information matrix. First, based on the database denormalization process of Cavique et al. (2019) a new denormalization process is presented. Then, we show how to extract all fact tables from a database. Finally, we present the constellation matrix, followed by the supply-demand matrix with a running example.

3.1 Bottom-Up Denormalization Process

As already mentioned, Cavique et al. (2019) present a top-down database denormalization process with two denormalization forms.

A similar decomposition process, with two phases, can be described using the inverse strategy, i.e. the bottom-up method.

In the First Denormalization Form using bottom-up method (1DF_bu) given a database schema, all the fact tables are identified, i.e., tables only with cardinality equal to N. In Fig. 2.a tables F1 and F2 should be recognized.

To obtain the Second Denormalization Form (2DF), for each fact table, add the tables from the upper level and repeat the process until no more tables can be added. Figure 4 exemplifies for table F1, in the first iteration I1 and L2 are added, and in the second iteration L1 is also added, obtaining Fig. 2.c on then left. The procedure is repeated for table F2 obtaining the two uncoupled trees of Fig. 2.c.

Summarizing the two denormalization strategies, for the First Denormalization Form we have two ways: 1DF via top-down, or 1DF_td, and 1DF via bottom up, or 1DF_bu. The Second Denormalization Form, 2DF, is equal for both pathways.

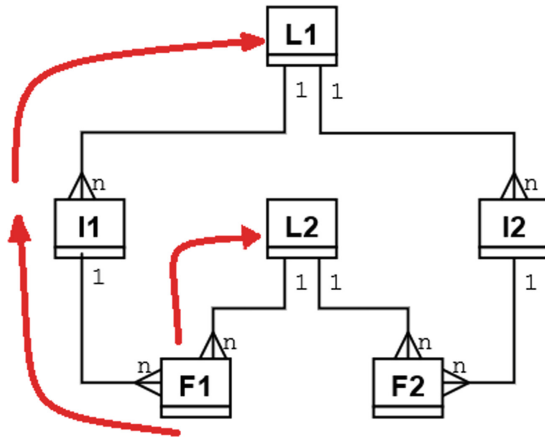


Fig. 4. Bottom-up denormalization process

3.2 Database Reduced Representation

In order to show how to extract all fact tables and dimensions from a database, we are going to exemplify the database reduction with the well-known Sakila database (2019), from the MySQL examples, which supports a DVD rental business.

First all the tables are categorized in lookup, intermediate or fact tables, using the definitions mentioned in Subject. 2.1. Table 1 shows columns with the name of the table, the type of table and the type of facts. The information about the fact tables is extracted without going into the details of the attributes of each table.

The numeric measures in a fact table fall into three categories (Kimball and Ross 2013): additive facts, semi-addictive facts and non-addictive facts. An extra category should also be mentioned, the fact tables without fact.

The reduced information from Sakila database retrieves a single fact table with additive facts, the Payment table.

Table 1 Sakila reduced database^u

table name	type of table	type of facts
country	lookup	
city	intermediate	
address	intermediate	
customer	intermediate	
store	intermediate	
staff	intermediate	
rental	intermediate	
payment	fact	additive
actor	lookup	
language	lookup	
category	lookup	
film	intermediate	
inventory	intermediate	
film_category	fact	without facts
film_actor	fact	without facts
film_text	fact	without facts

3.3 Constellation Matrix

Given the fact tables of Table 1 and applying the denormalization process described in Subsect. 3.1 it is possible to associate fact tables and dimensions.

In the running example we add a Human Resource database with the previous one. Figure 5 shows the constellation matrix (Shahzad and Sohail 2009) for Sakila and Human Resource databases, where facts and dimensions come together. The type of facts also reports: 'a' means additive, 'na' non-additive and 'wf' without facts.

Business database	Fact table	type	customer	staff	rental	actor	category	film	inventory
Sakila	payment	a	X	X	X				
	film_category	wf				X	X		
	film_actor	wf				X	X		
	film_text	wf							X
Human Resource	payroll	a		X					
	job_history	na		X					

Fig. 5. Constellation matrix for Sakila and HR databases

3.4 Supply-Demand Matrix

The constant arrival of new legislation and new business opportunities generates new requirements to the system that should adapt to change. By data warehouse requirement we mean a report or a data view to analyze or mine. Each requirement should be supported by source data, i.e. one or more fact tables of the constellation schema as shown in Fig. 6. Since a requirement can use more than one fact table, a correlation sub-matrix is shown on the right. This type of correlation is inspired by the House of Quality (Tapke et al. 2003).

Given the fact table Payment with additive facts, it is possible to answer to the requirements of a rental weekly report. The other requirement is a monthly payroll report which is possible to obtain given the fact table Payroll. Annually it is required a job analysis with needs additive facts and non-additive facts from table Payroll and Job_history.

#	fact table	type								requirements	1	2	3	4	5	6
			customer	staff	rental	actor	category	film	inventory							
1	payment	a	X	X	X					rental weekly report	X					
2	film_category	wf					X	X								
3	film_actor	wf				X		X								
4	film_text	wf							X							
5	payroll	a		X						monthly payroll report					X	
6	job_history	na		X						annual job analysis					X	X

(supply) (demand)

Fig. 6. Supply-demand matrix

Requirement oriented data warehouse is a challenge for the Kimball architecture which uses a bottom-up approach. In Jovanovic et al. (2014) the authors present a method to iteratively design the multi-dimensional schema of a data warehouse from requirements. Our systematic Procedure 1 follows a similar approach, by iterating the finding of new fact tables, followed by the matching with new dimensions and integrating with requirements, until the balance between supply and demand is established.

Procedure 1. Generation of the Supply-demand Information Matrix:

Input: files, databases

Output: supply-demand matrix

1. Iterate

1.1. Find new Fact Tables

1.2. Match with Dimensions

1.3. Integrate with Requirements

2. Until balance between supply and demand is established

To find all the fact tables in a database the procedure Subject. 3.2 is applied, which classifies each table into lookup, intermediate or fact. To match fact table with dimensions, creating a constellation matrix, the procedure in Subject. 3.3 can be used. Finally,

in order to integrate requirements, the information is mapped in the supply-demand matrix.

The emerging discipline of Organizational Engineering (Magalhães et al. 2007; Aveiro et al. 2011) advocates new principles. Organizational Engineering argues that each organization has its own identity and concerns for its interrelated subsystems. By developing meaningful meetings and business process KPI, organizations tend to be process-dependent rather than people-dependent. As a result, they can easily adapt personnel and they achieve high teams' performance.

Our process-oriented method of finding fact tables, matching dimensions and answer to requirement is iterative and incremental. On each iteration, new fact tables and/or dimensions should be added, to support new requirements. This approach goes beyond Technical Architecture design, with a back-room and a front-room working separately. The supply-demand matrix allows bird's-eye view of the data warehouse by representing the process from data source to data presentation, in order to avoid redundancy or lack of information.

The proposed systematic procedure follows also the Organizational Engineering by avoiding the human dependency, by establishing a set of rules to follow, strengthening aspects of systems engineering rather than constantly recreating new ways to solve the same problems for the purpose of personal appreciation.

In our data warehouse design a process is created from the data supply to the information demand. The process should iterate while the organization is learning and evolving.

4 Conclusions

Although star schema is a very well-established model, the increasing number of star schemas in large constellations adds new challenges in the organizations. Also, the constant arrival of new legislation and new business generates new requirements. Incremental demands, internal and external, cause the loss of the overall vision of the organization.

The goal of our paper is to develop a process-oriented procedure in the technical architecture of a data warehouses with constellation schemas using an extension of the bus matrix, in order to obtain a bird's-eye view of the system by representing the process from data source to data presentation. The integrated vision of supply and demand goes beyond technical architecture using a back-room and a front-room. This process view extracts information about the fact tables, without going into the details of the attributes of each table.

This work is also an attempt to bring together the visions of Kimball and Inmon, using a bottom-up approach to find fact tables and a top-down view to meet the requirements. The effort to match supply and demand of information avoids commitment on reports that does not correspond to actual data, causing the disappointment of the end users, and allows the deletion fact tables that are not used in the requirements.

An additional contribution regarding denormalization forms is reported. Given the top-down denormalization process by Cavique et al. (2019) we propose a bottom-up denormalization strategy, also with two denormalization forms, 1DF_bu and 2DF.

In future work, following the advices of Organizational Engineering we plan to establish KPI in our process-oriented data warehouse matrix.

References

- Aveiro, D., Silva, A.R., Tribolet, J.: Control organization: a DEMO based specification and extension. In: First Enterprise Engineering Working Conference, EEWK 2011, Antwerp, Belgium (2011)
- Breslin, M.: Data warehousing battle of the giants: comparing the basics of the Kimball and Inmon models. *Bus. Intell. J.* **7**, 6–20 (2004)
- Caldeira, C.P.: Data Warehousing: conceitos e modelos com exemplos práticos, 2ª edição, Edições Sílabo (2012)
- Cavique, L., Cavique, M., Gonçalves, A.: Extraction of fact tables from a relational database: an effort to establish rules in denormalization. In: Rocha, Á., Adeli, H., Reis, L., Costanzo, S. (eds.) *New Knowledge in Information Systems and Technologies, WorldCIST 2019. Advances in Intelligent Systems and Computing*, vol. 930, pp. 936–945. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16181-1_88
- Inmon, W.H.: *Building the Data Warehouse*, 4th edn. Wiley, New York (2005)
- Jovanovic, P., Romero, O., Simitsis, A., Abelló, A., Mayorova, D.: A requirement driven approach to the design and evolution of data warehouses. *Inf. Syst.* **44**, 94–119 (2014). <https://doi.org/10.1016/j.is.2014.01.004>
- Kimball, R., Caserta, J.: *The ETL Data warehouse Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*. Wiley, New York (2004)
- Kimball, R., Ross, M.: *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2nd edn. John Wiley & Sons, New York (2002). ISBN 0471200247
- Kimball, R., Ross, M.: *The Data Warehouse Toolkit: the Definitive Guide to Dimensional Modeling*, 3rd edn. Wiley, New York (2013). ISBN 9781118530801
- Linstedt, D., Graziano, K.: *Super Charge Your Data Warehouse: Invaluable Data Modeling Rules to Implement Your Data Vault*. Create Space Publishing Platform, Scotts Valley (2011)
- Magalhães, R., Zacarias, M., Tribolet, J.: Making sense of enterprise architectures as tools of Organizational Self-Awareness (OSA). *J. Enterp. Archit.* **3**(4), 64–72 (2007)
- Sakila. https://database.guide/what-is-a-database-schema/sakila_full_database_schema_diagram/. Accessed Nov (2019)
- Santos, M.Y., Ramos, I.: *Business Intelligence: da informação ao conhecimento*, 3ª edição, FCA - Editora de Informática (2017)
- Shahzad, K., Sohail, A.: A systematic approach for transformation of ER schema to dimensional schema. In: *Proceedings of the 6th International Conference on Frontiers of Information Technology, FIT 2009* (2009)
- Shin, S.K., Sanders, G.L.: Denormalization strategies for data retrieval from data warehouses. *Decis. Support Syst.* **42**, 267–282 (2006)
- Tapke, J., Muller, A., Johnson, G., Siec, J.: *House of Quality: Steps in Understanding the House of Quality*, IE 361. Iowa State University (2003)