Data Warehouse for Professional Skills Required on the IT Labor Market

Cristian GEORGESCU
cristian.georgescu@ugal.ro
Corina SBUGHEA
csbughea@ugal.ro
Dunarea de Jos University of Galati, Romania

This paper represents a research regarding informatics graduates professional level adjustment to specific requirements of the IT labor market. It uses techniques and models for data warehouse technology to allow a comparative analysis between the supply competencies and the skills demand on the IT labor market.

Keywords: IT labor market, data warehouse, star schema, facts constellation schema
JEL Code: C82, M15

1. Introduction

In a less formal environment, an old colleague from faculty asked me if I agree anymore with the slogan "man - the most precious capital." Of course, I said, mostly because I had just told him that I am a teacher at the Faculty of Economics and Business Administration, in Galati. Then he replied to me, why does it not appear anywhere in accounting the human resource as an investment element? Everywhere we register labor costs; even when we are doing a training, which raises the value of human resources, we recorded it also as an expense. Moreover, a software product resulted from the overwhelming effort (as importance) of human intelligence, is recorded as an intangible asset, while the human resource which transferred his value is considered an expense.

Without clearly expressing my point of view, because I'm not an expert in accounting, I concluded that regardless of how companies reveal their relations with labor force, investment aspect is present in at least three levels: at the personal level by self motivation for improvement, at the family level by parent-child relationship and, in most cases, at the society level through the education system.

This event determined us to begin a research, with some of our students, regarding informatics graduates professional level adjustment to specific requirements of the IT labor market. For three weeks a group of students sought the skills required by employers for IT positions, in the most popular websites for jobs.

The work was materialized in a gross volume of data, minimal organized, with many inaccuracies and errors of collecting. Nothing surprising for such an activity, but the volume, diversity and dynamics recorded convinced us that the idea, with which we have started, deserves to be developed.

The only technology suitable for such circumstances is that of data warehouses, which through the volumes of data managed, by showing the time evolution of data, through diversity of sources and non-unified data formats, can lead to meaningful and documented analysis of the studied phenomenon.

This paper will start from the presentation of a current image of the phenomenon studied by investigating the literature in this area and will propose in section 3 a model for the amount of data already collected. On this basis, a case study will be done in Chapter 4, which will be the starting point for the development in Chapter 5, of a new model for data warehouse.

2. Data warehouse – concepts and literature review

It can be considered that the data warehouse concept dates from the late '80s, when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse", which was meant to ensure the flow of information from operational systems to decision support environments.

As the complexity of the organizations continued to grow in the 1990s, it became critically important for large enterprises to process data and use it in order to take strategic business decisions.

Nowadays, collecting large amounts of data, required for the day-to-day activity and for supporting proper decisions, is no longer a question from the technical point of view. Today the

question is how to extract only the relevant information for the current problem, with minimal amount of time and money. Thus was born the idea of structuring information in data warehouses, to facilitate data analysis for management, which operational databases could not perform without severely affecting the response time.

Data warehouse have been defined in many ways over time. In accordance with W.H. Inmon, a leader in building data warehouse systems, a data warehouse is "a source of data that is subject-oriented, integrated, nonvolatile, and time-variant for the purpose of management's decision processes". The feature subject-oriented refers to the fact that the data in the database is organized so that all the elements relating to the same event or object are connected and can be used to analyze a particular subject area. The data warehouse integrates data from multiple operational sub-systems of the organization, and time-variant means that all the changes in the data are kept over time and reports can be produced to show these changes. The non-volatility property refers to the fact that data in the warehouse is never over-written or deleted, once acquired should never be altered.

Another definition is offered by Ralph Kimball: a data warehouse is "a copy of transaction data specifically structured for query and analysis." Kimball proposes a bottom-up methodology for data warehouse, by creating individual data marts that provide thin views of the organizational data, which can be combined into a more comprehensive data warehouse. (Kimball, 1997)

Inmon has exactly the opposite view, namely that the data warehouse should be designed from the top-down to include all organizational data, so the data marts must be created only after the data warehouse is completed. (Inmon, 1998)

According to Larson "a data mart is a body of historical data in an electronic repository that does not participate in the daily operations of the organization. Instead, this data is used to create business intelligence. The data in the data mart usually applies to a specific area of the organization". As a source for business intelligence, the data mart is built considering more the speed of access, rather that the rules of normalization, although it is still a relational database, but with fewer table joins needed for data analysis.(Larson, 2009)

Ian Abramson (Oracle expert), concluded that data warehouse is the single organizational repository of enterprise wide data across many or all lines of business and subject areas. This means that contains massive and integrated data and represents the complete organizational view of information needed to run and understand the business.

In the scientific literature on data warehousing, according to Ariyachandra and Watson, architectures can be grouped into five categories: independent data marts, bus architecture, hub and spoke, centralized and federated. (Ariyachandra and Watson, 2005)

Independent data marts are orientated to particular applications, providing data to specific issues, and are considered that they do not offer a single version of the truth. (Alsqour, Matouk, Owoc, 2012)

A bus architecture (which Kimball promotes) is a collection of integrated data marts based on *conformed dimensions* (a set of data attributes that have been implemented in multiple database tables using the same structure, attributes, domain values, definitions and concepts) and *conformed facts*. Once defined and implemented, the conformed dimensions mean the same thing wherever they are utilized.

Hub and Spoke Architecture (which Inmon promotes) intends to iteratively develop, subject by subject, an enterprise view of data, where atomic (detail) level data are maintained in the warehouse in 3rd normal form. In this architecture, the dependent data marts need addresses, because they receive data from central source - the warehouse.

Centralized data warehouse is similar to the hub and spoke architecture but has no dependent data marts (it consolidates data marts into data warehouse). The warehouse contains detailed level data, summarized data, and logical dimensional views of the data.

A federated data warehouse is a combination of *existing structures (*heterogeneous business intelligence systems), in order to achieve the "single version of the truth" across the organization. It aims to integrate the key business metrics, measures and dimensions. The federated data warehouse architecture provides the basis for business analysis and decision making in this heterogeneous environment.

In 2005, Ariyachandra and Watson conducted a study which aimed to respond the question "Which Data Warehouse Architecture Is Most Successful?" The authors established as criteria for assessing the success of the architectures, four measures: (1) information quality, (2) system quality, (3) individual impacts, and (4) organizational impacts, using a seven-point scale for each of them. (Ariyachandra and Watson, 2005)

The survey included four hundred fifty-four respondents, from small to large companies, most of them located in the United States (60%), representing a variety of industries, with the financial services industry (15%) providing the most responses. We represented in the chart below, the average values determined by the authors.

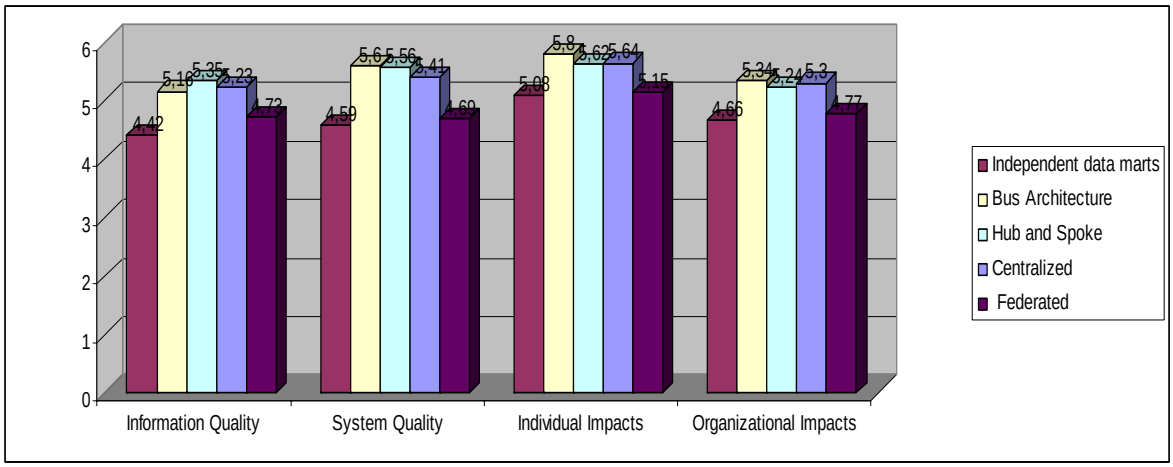


Fig.1.The average scores for the measures across the architectures
(based on data from H. J. Watson., T. Ariyachandra, 2005 - Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures)

As we can see, the lowest scores are on independent data marts architecture, which fact reinforces the belief that independent data marts are a poor architectural solution. Also, the authors of the study suggest that the federated architecture is not an optimal solution on the long term.

In the other three cases, the scores did not recorded major differences, so the study cannot proclaim the superiority of one of them.

But the research revealed that the most used architecture in practice was the hub-and-spoke (39%), followed by the bus architecture (26%), centralized (17%), independent data marts (12%), and federated (4%).

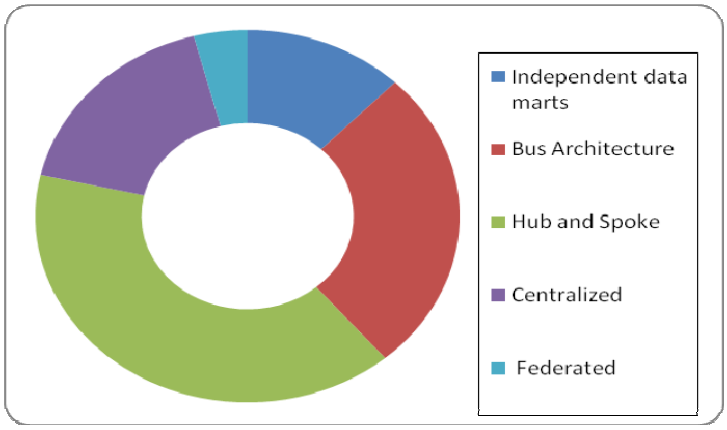


Fig.2. The predominant architectures
(based on data from H. J. Watson., T. Ariyachandra, 2005 - Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures)

In 2010, Ariyachandra and Watson made another research in which they investigated the factors that influence the choice of specific data warehouse architecture. Four hundred experts responded the survey, and provided various combinations of organizational factors that influence the selection of data warehouse architecture. The authors also calculated the distribution of four main architecture: independent data marts 11%, bus architecture 26%, enterprise data warehouse 58% and federated architecture 5%.

Sandeep Sing and Sona Malhotra (2011) concluded that, if resources are limited, more likely to be selected is independent data marts architecture, but if there is a high need to share information between departments, then the choice is for bus architecture. The hub and spoke or centralized architecture is chosen as a strategic solution. In another study from 2011, Mawilmada stated that in order to improve decision-making, there is a need of centralized data management.

Closer to the current article topic is a study of van Gils, Vermandere, and Booghmans, which aim to introduce the data warehouse as an important tool for analyzing mobility in the labor market. (Gils, Vermandere, and Booghmans, 2005)

Also, between 2005 and 2009, Berghman Jozef conducted a project in Belgium, which had as a main objective, the achieving of a consolidated, harmonized and high quality data warehouse Labor Market and Social Protection.

Regarding the choice for data warehouse architecture, there is no single generally valid solution. Most specialists agreed that, what can be the best solution for one organization, doesn't work for another one.

3. Identification of elements for the multidimensional model

By analyzing the necessities of processing, informing and cognition were identified five dimensions: time, location, organization, actor and competence. Time dimension allows the introduction of the dynamic aspect regarding the required skills, in the multidimensional analysis. It allows us, in a long-term study, to make a historical and perspective analysis on the level of demand for certain skills. The identified competencies are organized into competence dimension, that allows analysis operations at a certain moment, or historical, of one or more skills.

The refining of competences distribution, from a territorial point of view, will be done using location dimension, which will allow local or regional analysis on the evolution of the studied phenomenon.

Organization dimension focuses the analysis on applicant of skills, providing specific information about the profile of organization. Since recruitment companies cannot be excluded from the organizations that require the skills, we will find a greater uniformity of competences at the direct employers and a low one at the recruitment companies.

The actor dimension represents the man, the function or generally the actor's role in the organization. This represents the way to collect by functions the skills which employers require. The absence of this dimension would lead us to heterogeneous and arbitrary groups of competences, from the employee with one skill to the man who has it all.

Facts are the basic values in the dimensions space and are entered as number of appearances of each dimension realization. In consequence the measure is given by the instantaneous value, which indicates the number of requests at a time, in a specific location by a specific organization, for an arbitrary function with certain competences.

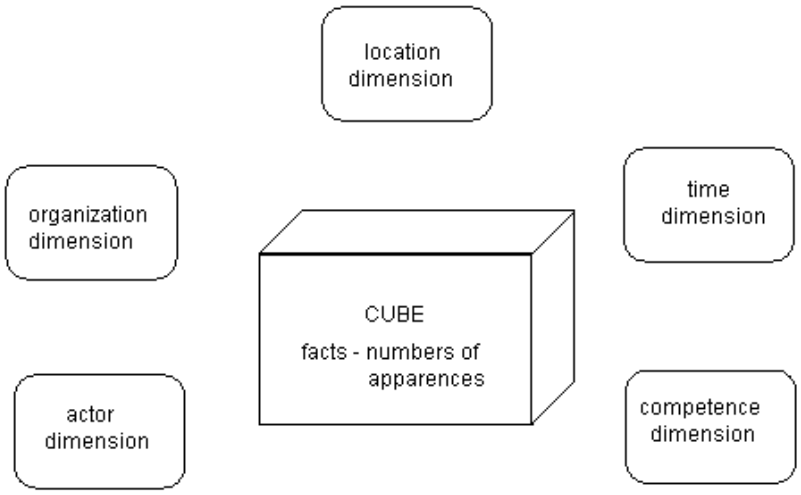


Fig.3. Cube dimensions

The metric of the model, the domain towards which the analysis focuses on, is the number of vacancies, required by organizations in the five dimensions space previously defined. Detailing the analysis of data warehouse design, leads us to an initial identification of each dimension attributes.

Thus, for each dimension was chosen a functional attribute id_ ***, which allows the implementation of the relationships in the multidimensional database. The other attributes are obvious and suggestive named, and they can be supplemented with new ones, based on iterative development principle in designing the data warehouse.

Attributes realizations are the values collected from various sources and made compatible in terms of representation with attributes formats declared in the multidimensional database. For the description of the five dimensions cube illustrated above (Fig.3) was chosen a star schema representation, figure 4, which includes the dimensions with their attributes and their relationships with fact table called "posturi".

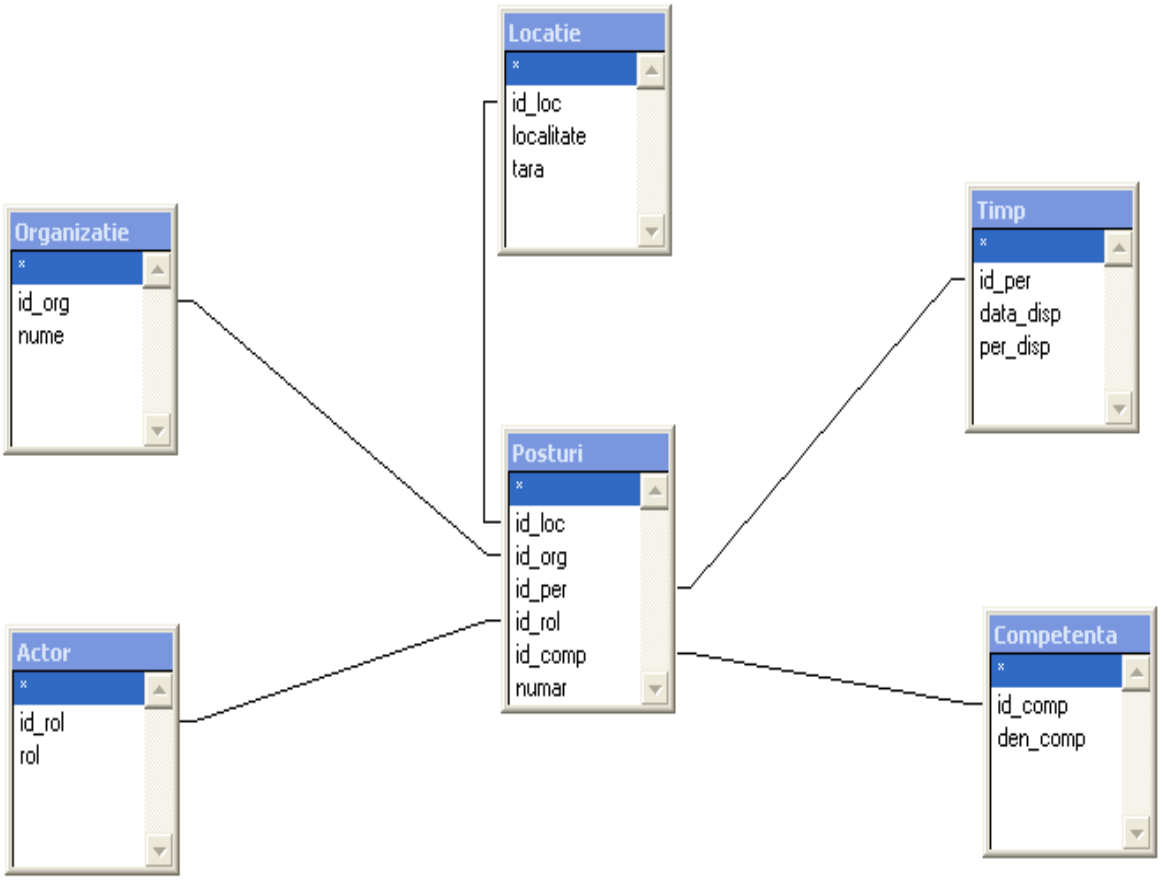


Fig.4. Star representation schema

4. Case study

For this study, were consulted 45 recruitment sites, identifying 1160 job offers in IT. The details followed were: the job title, location, company, date of this poster and the list of competencies required for each job separately.

In the chart below we can see the query results based on location of jobs offered. The share of jobs in Romania is higher because of all sites found, over half were Romanian sites.

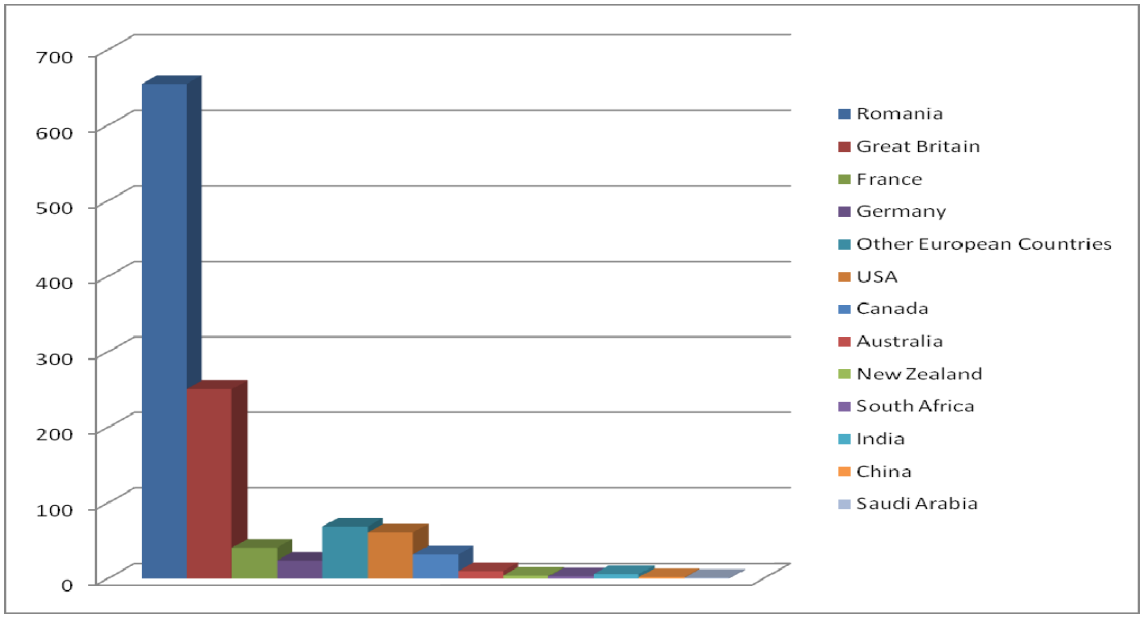
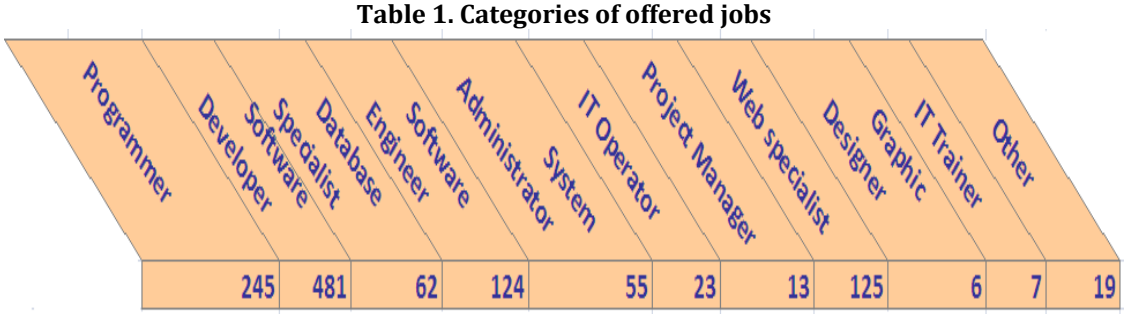


Fig. 5. Jobs distribution by country

Based on job titles, we made a grouping of requests by categories, and their structure is summarized in the table below:

Table 1. Categories of offered jobs



Queries made for the required skills have enabled a hierarchy of the most popular competencies.

Table 2. Number of requests for competencies

Current No.	Competence	Number of requests
1	SQL	566
2	Java Script	417
3	HTML	402
4	CSS	312
5	PHP	286
6	C#	240
7	C++	219
8	Windows	202
9	Java	184
10	OOP	181
11	Unix	172
12	Oracle	169
13	JQuery	136
14	C	130
15	AJAX	118
16	Microsoft Office	116

Current No.	Competence	Number of requests
17	ASP.NET	104
18	Linux	63
19	XML	57
20	.NET Framework	51
21	Python	49
22	MySQL	43
23	Network management	39
24	Adobe Photoshop	35
25	Microsoft Visual Studio	27
26	LAN/WAN	26
27	Android	22
28	iOS	22
29	Perl	16
30	WEB Services	14
31	Operating System	11
32	MVC	10
33	C++Builder	9
34	ITIL	7
35	J2EE	7
36	SOAP	7
37	Symphony	7
38	TCP/IP	7
39	X-Code	7
40	JDBC	6
41	Magento	6
42	Shell Scripting	6
43	GUI	5
44	Protocols	5
45	SEO	5
46	UML	5
47	API	4
48	Delphi	4
49	FTP	4
50	JSF	4
51	Ruby	4
52	SVN	4
53	3D Studio Max	3
54	ADO.NET	3
55	CISCO	3
56	CMS	3
57	Corel Draw	3
58	Eclipse	3
59	Hibernate	3
60	HTTP	3
61	Joomla	3
62	JSON	3
63	MS Visio	3
64	POP3	3
65	Share Point	3
66	XSLT	3
67	ABAP	2
68	Adobe Flash	2
69	Django	2
70	Drupal	2
71	Hardware	2
72	JS	2
73	JSE	2
74	MFC	2

Current No.	Competence	Number of requests
75	Silverlight	2
76	Swing	2
77	VBA	2
78	Visual FoxPro	2
79	VMWare	2
80	WCF	2
81	XSL	2
82	Active Directory	1
83	ActiveX	1
84	Adobe Flex	1
85	Adobe Ilustrator	1

Only for the jobs offered in Romania, the most requested skills are:

Table3. Top ten requested IT skills in Romania

Current No.	Competence	Number of requests
1	SQL	275
2	HTML	251
3	PHP	208
4	C++	106
5	Java	106
6	C#	87
7	Oracle	79
8	Java Script	78
9	Unix	77
10	C	68

5. Model development

This multidimensional model allows only a limited analysis of the phenomenon, namely, the existence and the request of skills in defined space. Reaching this level of formalism, developments are obvious in terms of representation techniques but also in terms of conceptual analysis enlargement. This can be safely switched to a fact constellation representation scheme by adding a table of facts related to the same dimensions. The technical and formal aspects of the action only involve ensuring the semantic consistency of attributes and dimensions already set. From a conceptual point of view we focused on the supply comparative analysis with the skills demand.

Additional fact table "Locuri" represents the number of actors in an organization, from a specific location, at a certain moment of time and with certain skills, that can occupy the previously analyzed offer.

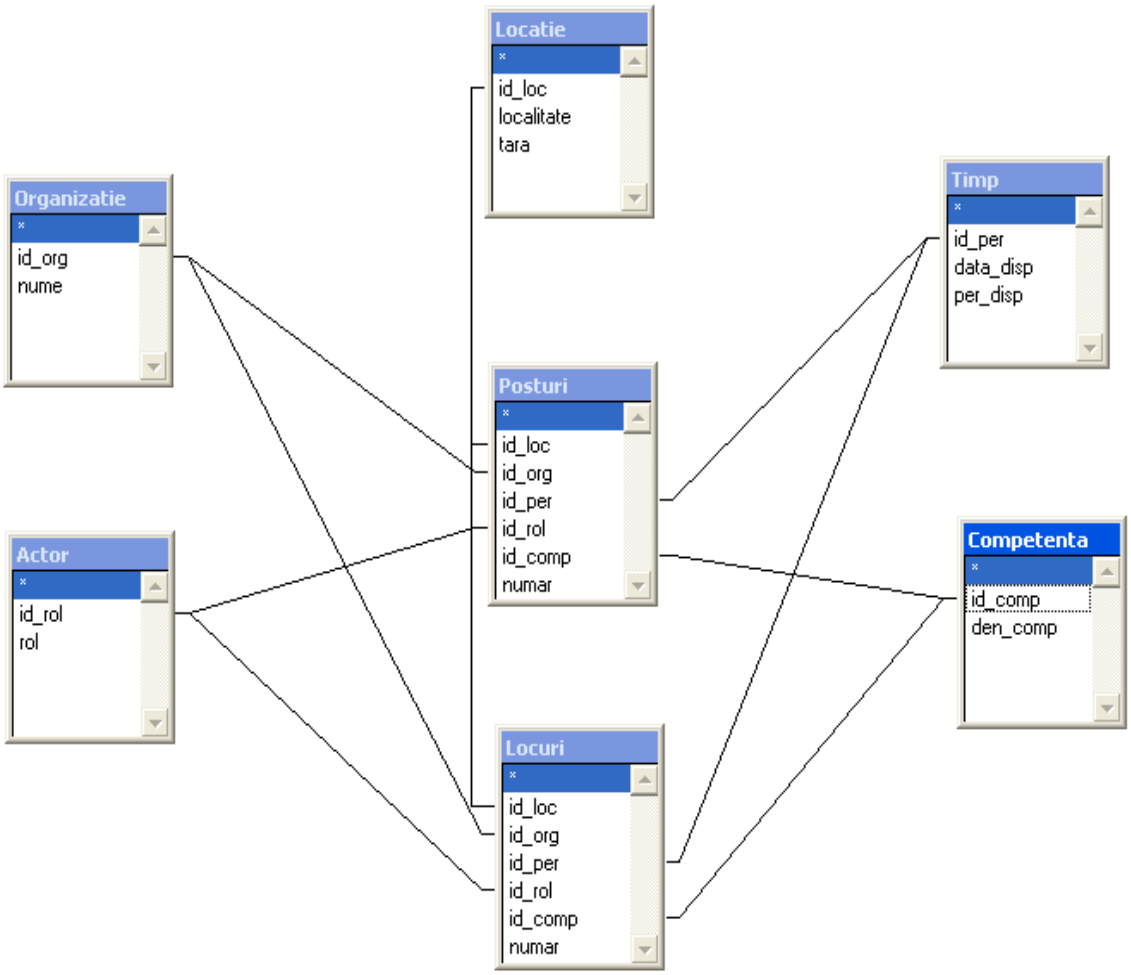


Fig.6. Facts constellation representation schema

In this case, organizations are also the educational institutions, actors are those who play the role of a student or postgraduate student, and the skills are declared and insured through the formal education curriculum. "Locatie" and "Timp" dimensions have the same meaning, except that the attribute "data_disp" represents the moment of competence certification: the diploma or a certificate evidencing the skills.

7. Conclusion

Although the case study was done for the skills demanded in the IT domain, the analysis can be extended to other expertise areas. This situation would require a transition to a more developed representation scheme for data cube: snowflake schema with multiple facts, a fact constellation with developments at the dimension tables' level.

Disposing of this new formalism we can expand the analysis also for the adjustment of educational system curriculum, with the skills required in the real labor market. This initiative may be materialized through the identification of skills acquired at each educational level from disciplines curricula.

Developing the model can lead to the completion of some multiple fact tables, one for each professional field, and to the organization of skills categories, building them in connection with the professional domains. Cross analysis between the skills required in the labor market and those offered by the educational system, as well as the correlations between supply and demand of skills from related fields, will allow an accurate and undoubted targeting of the educational system.

Analysis can be made on the geographical distribution of training level and labor use, and therefore can be carried out studies on the directions of labor migration. Finally the research will be

useful to those who invest in their own self-improvement, as well as to those who make the most dedicated investment in training and developing the human resource: the parents.

References

1. *B. Larson (2009) - Delivering Business Intelligence- Microsoft® SQL Server™ 2008, The McGraw-Hill, 2009*
2. *H. J. Watson., T. Ariyachandra (2005) - Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures, Terry College of Business, University of Georgia, (July 2005)*
3. *Moh'd Alsqour, Kamal Matouk, Mieczysław L. Owoc (2012) - A survey of data warehouse architectures - preliminary results, Proceedings of the Federated Conference on Computer Science and Information Systems, 2012*
4. *P. K. Mawilmada, (2011) - Impact of a data warehouse model for improved decision-making process in healthcare. Masters by Research thesis, Queensland University of Technology October.*
5. *R. Kimball, M. Ross (2002)- The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, Wiley 2002.*
6. *S. Gils, M. Booghmans and C. Vermandere (2005) - Making mobility in the labour market measurable. The possibilities of the Datawarehouse Labour market and the PMWP-database: Analysing career paths. TLM.NET Conference Paper, Budapest. 2005*
7. *S. Singh, S. Malhotra (2011) - Data warehouse and its methods, Journal of Global Research in Computer Science, Volume 2, No. 5, May 2011*
8. *T. Ariyachandra, H. J. Watson (2010)- Key organizational factors in data warehouse architecture selection, Decision Support Systems 49 (2010)*
9. *W.H. Inmon (2002) - Building the Data Warehouse, Third Edition, New York: John Wiley & Sons, 2002.*