THE BUCHAREST ACADEMY OF ECONOMIC STUDIES

ISSUE 3

Database Systems Journal

ISSN: 2069 – 3230 Volume II (March 2011)

Journal edited by Economic Informatics Department

DBJOURNAL BOARD

Contact
Calea Dorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro/
E-mail: editor@dbjournal.ro

Contents

Building a Spatial Database for Romanian Archaeological Sites

Aura-Mihaela MOCANU, Manole VELICANU
Economic Informatics Department, Academy of Economic Studies
Bucharest, ROMANIA
mocanuaura@yahoo.com, mvelicanu@yahoo.com

Spatial databases are a new technology in the database systems which allow storing, retrieving and maintaining geospatial data. This paper describes the steps which we have followed to model, design and develop a spatial database for Romanian archaeological sites and their assemblies. The system analysis was made using the well known Entity-Relationship model; the system design included the conceptual, the external and the internal schemas design, and the system development meant developing the needed database objects and programs. The designed database allows users to load vector geospatial data about the archaeological sites in two distinct spatial reference systems WGS84 and STEREO70, temporal data about the historical periods and cultures, other descriptive data and documents as references to the archaeological objects.
Keywords: *spatial databases, entity-relationship model, conceptual schema, external schema*

1 Introduction

Spatial databases have the ability of storing and manipulating geospatial data. They are usually extensions of relational databases which contain special geometry objects with several mandatory attributes and methods defined in the Open Geospatial Consortium's (OGC) standard "Application objects", spatial indexing mechanism, operators and functions to make queries, joins and other spatial analysis operations.

Geospatial data means raster data (Earth photography's made either from satellite, either from plane) or vector data (points, lines, polygons which describe the location of the objects on the surface of the Earth and their outline).

Some spatial available database systems are: commercial software Oracle Spatial or IBM DB2 Spatial Extender, the open source spatial databases such as: PostgreSQL/PostGIS, Spatial Box, Ingress Geospatial, H2 Spatial, Spatial Lite, MySQL Spatial. Spatial databases store the geospatial data by providing either a proprietary object such as Oracle's object SDO_GEOMETRY, either assuring the support for the standard storage format well-known binary (WKB). Paper [1] describes a detailed comparative study between Oracle Spatial and PostgreSQL/PostGIS.

Store the geospatial data is not enough; one must retrieve the data also, and fast. That is why the spatially-enabled databases also have defined a special type of indexes, named RTree indexes, which we have noticed to be available in most of the spatial databases which we have analyzed, both commercial and open source.

Spatial operations such as spatial queries, create, update, insert, and delete operations, conversions, and operations on the map or analysis on grid cells are very well documented in paper [2].

We will further describe the steps we have followed in order to build a spatial database for our collaborators from the Romanian National History Museum (MNIR) to manage the archaeological sites and their assemblies.

2 System analyses - Entity-Relationship Model

We have started the analysis of the new database by finding out what kind of data is it needed to be maintained by the specialists about the archaeological sites and their assemblies, having as example an

old application which was used by them, written in Microsoft Access.

Like in any other archaeology-related information system, the data model includes the following data categories [3]:

- temporal data - historical periods and cultures are assigned to any archaeological site,
- spatial data - location of the archaeological site (descriptive and vector geospatial data),
- archaeological objects – description of archaeological sites and assemblies,
- documents – attached to any site.

The most used technique to create the data model is the Entity-Relationship (ER). In order to create the ER diagram, we have identified first the entities (users, archaeological sites, assemblies, geographical coordinates, waters, counties, historical periods, historical cultures, assemblies' classes) and their attributes. The main entities and their attributes are described in below table.

Table 1. Main entities and attributes

Entity	Attribute	Description
site	id_sit	ID site
	cod_siruta	SIRUTA code
	cod_lmi	LMI code
	cod_ran	RAN code
	denumire	name of the site
	alternativ	another site's name
	ape	waters
	punct	a point mark
	repere	landmarks
	sursa	site's references
	observatii	observations
ensemble	id_ansamblu	ID ensemble
	denumire	name
	clasa	ensemble's class
	tip	ensemble's type
	cercetare	research
	reper_supl	landmark
	inventar	inventory
	perioada	historical period
	cultura	historical culture
	sursa	references
	observatii	observations
coordinate	id_coordonate	ID coordinate
	tip_date	data type (information about GPS receiver)
	elevat	elevation
	eroare	error
	geom	geographical coordinates
	observatii	observations

After describing all the entities and their attributes, we have drawn the ER diagram (figure 1) with WebRatio tool, commercial software which supports the Web Modeling Language (WebML), whose purpose is the design of data intensive web sites. We have chosen this CASE tool because after building the database, a geoportal will be also designed.

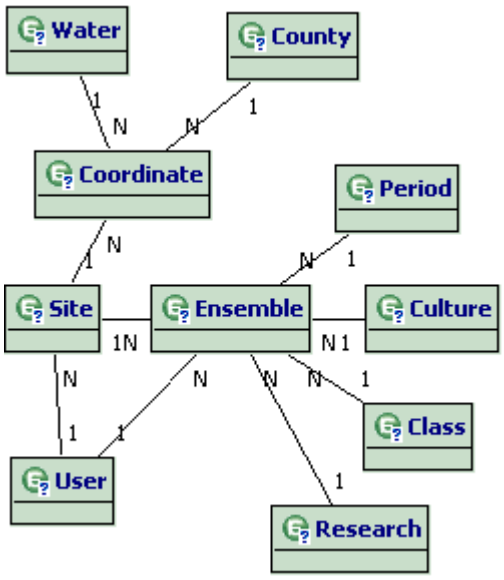


Fig. 1. ER diagram

The diagram has to be read as follows: one archaeological site is composed from one or more assemblies, which are loaded in the database by the system's users; every site has some geographical coordinates registered; the geographical coordinates refer not only to archaeological sites, but also to waters or Romania's counties boundaries; the

archaeological assemblies dates from a certain historical period and culture, they belong to a certain ensemble class and they are part of a certain research program.

The relationships between the entities are described in below table:

Table 2. Relationships between entities

Entity	Linkage Phrase	One/ Many	Entity
Site	is composed from	many	Ensemble
Site	has assigned	many	Coordinate
Site	is last maintained by	one	User
Coordinate	refer to	one	Site
Coordinate	refer to	one	Water
Coordinate	refer to	one	County
Ensemble	is part of	one	Site
Ensemble	is last maintained by	one	User
Ensemble	dates from	one	Period
Ensemble	belongs to	one	Culture
Ensemble	belongs to	one	Class
Ensemble	is part of	one	Research
Period	belongs to	many	Ensemble
Culture	belongs to	many	Ensemble
Research	studies	many	Ensemble
Class	describes	many	Ensemble
User	maintain	many	Site
User	maintain	many	Ensemble

The "geom" attribute of "coordinate" entity is the vector geospatial data: points, lines or polygons representing the archaeological sites in two different spatial reference systems (SRS): WGS84 (which is the world's most used SRS) or STEREO 70 (which is Romania's most used SRS).

The geospatial data describing the Romania's counties and waters is loaded from a shapefile (.shp) provided by the Romanian geospatial community (http://earth.unibuc.ro/). The shapefiles are the ESRI's proprietary storage format for geospatial data.

3 System design

PostgreSQL/PostGIS was chosen to be used as the spatially-enabled database by analyzing the following decision factors:

- technical capabilities: it includes support for all of the functions and objects defined in the OpenGIS 'Simple Features for SQL' specification,
- documentation: it is available online with a lot of coding examples,
- support: it has an online bug tracking mechanism (http://trac.osgeo.org),
- ease of use: it is easy to install, easy to develop the database,
- usage rate: many scholars have used with success this database for their projects,
- price: open source tools are very interesting for the archaeologists who usually deal with low budget projects.

3.1 Conceptual Schema

The conceptual schema of the database is designed starting from the previous modeled ER diagram. Each entity from the ER model is transformed into a database table, and for each relationship foreign keys (FK) are defined. The conceptual schema will be improved step by step, following the normalization technique, until a balance is reached between the maintenance requirements and system's exploitation performance [4].

The corresponding database tables for the main entities which we have described

in the previous section are detailed in below table.

Table 3. SITE table description

Column description	Column name	Data Type	PK	FK
ID site	id_sit	integer	X	
SIRUTA Code	cod_siruta	integer		X
LMI Code	cod_lmi	integer		X
RAN Code	cod_ran	integer		X
Name	denumire	text		
Alternative name	alternativ	text		
Point	punct	text		
Landmarks	repere	text		
References	sursa	text		
Comments	observatii	text		

Table 4. ASSEMBLIES table description

Column description	Column name	Data Type	PK	FK
ID site	id_sit	integer		X
ID ensemble	id_ansamblu	integer	X	
Name	denumire	text		
ID assemblies class	id_clasa	integer		X
ID assemblies types	id_tip	integer		X
ID research	id_cercetare	integer		X
Landmark	reper_supl	text		
Inventory	inventar	text		
ID historical period	id_perioada	integer		X
ID historical culture	id_cultura	integer		X
References	sursa	text		
Comments	observatii	text		
Last mutation	data_ultima_actualiz	timestamp		
Last mutation user	utiliz_ultima_actualiz	text		

Table 5. COORDINATES table description

Column description	Column name	Data Type	PK	FK
ID coordinate	id_coordonate	integer	X	
ID sit	id_sit	integer		X
Data type (GPS receiver)	tip_date	text		
Elevation	elevat	integer		
Error	eroare	numeric		
Geographical coordinates	geom	geometry		
Comments	observatii	text		

Because "geom" column of "coordinates" table is defined as a geometry data type column (being used to store vector geospatial data about the archaeological sites), it has to be included in special PostGIS table "geometry_columns". This table defines: all the tables containing geometry columns, the spatial dimension (2, 3 or 4 dimensional) of the geometry columns, the ID of the spatial reference system used for the coordinate geometry, and the type of the spatial object (POINT, LINESTRING, POLYGON, MULTIPOINT, MULTILINESTRING, MULTIPOLYGON, GEOMETRYCOLLECTION).

The spatial reference systems which could be assigned to the geospatial data stored in PostGIS are defined in PostGIS table named "spatial_ref_sys". This technical table lists over 3000 known spatial reference systems (SRS) and details needed to transform/reproject between them. The geospatial data which will be

filled by the users of currently described system might be in the following spatial reference systems: WGS84 whose corresponding PostGIS SRID is 4326 and STEREO70 whose SRID is 31700. One can find out the Well-Known Text (WKT) representation of a certain SRS by using the following select statement:

```
   select srtext from spatial_ref_sys
where srid = 4326;

   "GEOGCS["WGS
84",DATUM["WGS_1984",SPHEROID["WGS
84",6378137,298.257223563,AUTHORITY["E
PSG","7030"]],TOWGS84[0,0,0,0,0,0,0],A
UTHORITY["EPSG","6326"]],PRIMEM["Green
wich",0,AUTHORITY["EPSG","8901"]],UNIT
["degree",0.01745329251994328,AUTHORIT
Y["EPSG","9122"]],AUTHORITY["EPSG","43
26"]]"
```

The WKT representation of a spatial reference system offers a standard to describe, as a text, the information about the geospatial data projection system. The projection system must be specified in the geospatial data source (file or database), because it is very important especially when the data from different sources is used together. One can overlay on a map only layers in the same projection (or spatial reference system). The WGS84 (World Geodetic System of 1984) models the world as a spheroid. No ellipsoid/spheroid model perfectly the Earth, but corrections are made in order to get a better approximation for each territory. These special corrections of the initially spheroid are named datum [5]. In our country the Stereographic 70 (STEREO70) system was adopted.

3.2 External Schema

In order to draw maps in different spatial reference system, we have created several views (table 6) on the database table COORDINATES filtering the geospatial data by its type (point, line or polygon) and by the spatial reference system in which it was represented.

Table 6. Database views

View	Description
vw_point_wgs	The representation of the archaeological sites, as points, in 4326 projection (also known as WGS84).
vw_point_stereo	The representation of the archaeological sites, as points, in 31700 projection (also known as Stereo70).
vw_point_google	The representation of the archaeological sites, as points, in 900913 projection (also known as Web Mercator).
vw_polygon_wgs	The representation of the archaeological sites, as polygons, in 4326 projection (also known as WGS84).
vw_polygon_stereo	The representation of the archaeological sites, as polygons, in 31700 projection (also known as Stereo70).
vw_polygon_google	The representation of the archaeological sites, as polygons, in 900913 projection (also known as Web Mercator).

These views are important when the map layers will be defined in the Web Mapping Service (WMS) software, because one layer can be associated with a "geometry" database column only if this column contains data about one single type of geospatial data, in one single SRS. We will detail in section 4 of this paper how the geospatial data stored in the database is rendered on a map by using WMS software.

Regarding the security policy, there were designed two user group roles with different rights in the database. The select action on the tables is granted to user role „guest" and insert/update/delete actions are granted to the user role „admin". Each time a user is created, one of the two group roles („guest" or „admin") will be granted to the user, so that the access to the database is done in a rigorous way. Also for security reasons, each change on the 3

main tables (archaeological sites, assemblies and coordinates) will trigger an insert action into a history table, so that the „admin" role user will know when was it done an update on these tables, by which user and what data did he modified.

3.3 Internal Schema

In this step we have defined RTree indexes for each table which contains geometry columns. RTree approximates each geometry through a *Minimum Bounding Rectangle (MBR)*. The Postgres query optimizer will consider using an RTree index whenever an indexed attribute is involved in a comparison using one of the following geometric operators: <<, &<, &>, >>, @, ~=, && which are explained in below table:

Table 7. Geometric operators

Operator	Description
&&	Overlaps?
~=	Same as?
@	Center
<<	Is strictly left of?
>>	Is strictly right of?
&<	Does not extend to the right of?
&>	Does not extend to the left of?

The disk space can be estimated in three ways: using special SQL functions (table 8), using **VACUUM** information, and from the command line using the tools in **contrib/oid2name**. The SQL functions are the easiest to use and report information about tables, tables with indexes and long value storage (TOAST), databases, and tablespaces. Using psql on a recently vacuumed or analyzed database, queries could be written to see the disk usage of any table [6], such as:

```
SELECT relfilenode, relpages FROM
pg_class WHERE relname = 'coordonate';
```

Each page is typically 8 kilobytes and the **relfilenode** value is of interest to examine the table's disk file directly.

Table 8. Database Object Size Functions

Name	Return Type
pg_column_size(any)	int
pg_database_size(oid)	bigint
pg_database_size(name)	bigint
pg_relation_size(oid)	bigint
pg_relation_size(text)	bigint
pg_size_pretty(bigint)	text
pg_tablespace_size(oid)	bigint
pg_tablespace_size(name)	bigint
pg_total_relation_size(oid)	bigint
pg_total_relation_size(text)	bigint

4 System developments

We have first created the database tables, as shown in figure 2.

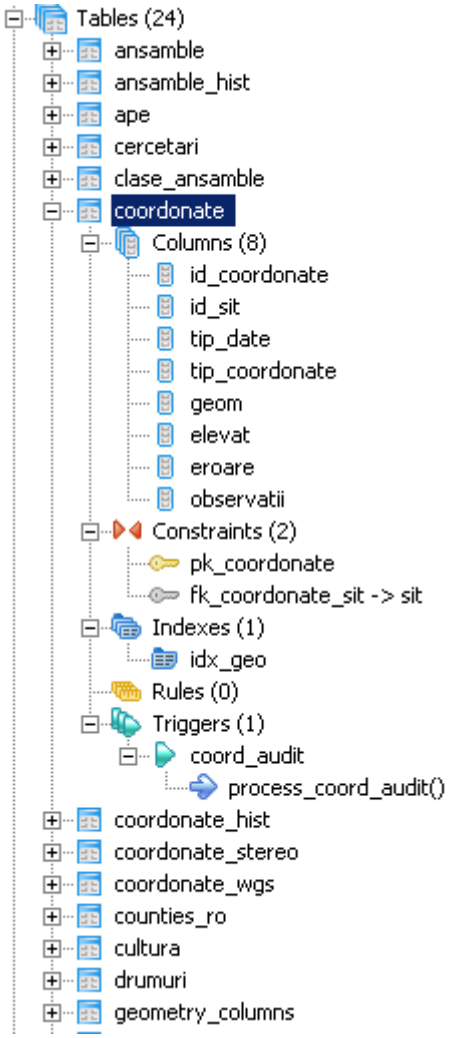


Fig. 2. Database tables

In PostGIS, there is used following syntax in order to add a geospatial column: AddGeometryColumn(<table_name>, <column_name>, <srid>, <data_type>, <dimension number>)

For example, we have used the following statement in order to add "geom" column from "coordonate" table:

```
AddGeometryColumn(coordonate, geom, 4326, point, 2);
```

The PostGIS storage format for the geometric object is Well Known Binary (WKB). The Well Known Binary Representation for Geometry (WKBGeometry) provides a portable representation of a geometric object as a contiguous stream of bytes. It permits geometric object to be exchanged between an SQL/CLI client and an SQL-implementation in binary form. The Well-known Binary Representation for Geometry is obtained by serializing a geometric object as a sequence of numeric types drawn from the set {Unsigned Integer, Double} and then serializing each numeric type as a sequence of bytes using one of two well defined, standard, binary representations for numeric types (NDR, XDR) [7].

Table 9. Integer codes for geometric types

Type	**Code**
Geometry	0
Point	1
LineString	2
Polygon	3
MultiPoint	4
MultiLineString	5
MultiPolygon	6
GeometryCollection	7
CircularString	8
CompoundCurve	9
CurvePolygon	10
MultiCurve	11
MultiSurface	12
Curve	13
Surface	14
PolyhedralSurface	15
TIN	16

The master data tables which store the geospatial data regarding the Romanian counties and waters were loaded from an ESRI shapefile, into the database, using the *shp2pgsql* tool according to the below schema.

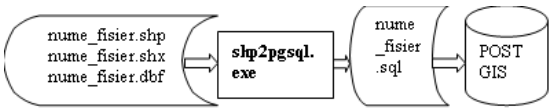


Fig. 3. Load geospatial data from a shapefile into the database

A GiST index was created on "coordonate" table:

```
CREATE INDEX idx_geo ON coordonate
 USING gist (geom);
```

GiST stands for Generalized Search Tree. It is a balanced, tree-structured access method that acts as a base template

in which to implement arbitrary indexing schemes. B-trees, R-trees and many other indexing schemes can be implemented in GiST [6].

The views on "coordonate" table, used to split geospatial data by its type and SRS, were created using some PostGIS spatial functions, such as ST_TRANSFORM. For example, to create a view with all the POINT geospatial data, either in WGS84, either in STEREO70, and reprojected in Web Mercator SRS, we have used below statement:

```
CREATE OR REPLACE VIEW vw_point_google
AS
SELECT coordonate.id_sit,
sit.denumire, sit.cod_siruta,
sit.repere, sit.ape,
coordonate.tip_date,
st_transform(coordonate.geom, 900913)
AS geom, coordonate.tip_coordonate,
coordonate.elevat, coordonate.eroare,
coordonate.observatii
FROM coordonate, sit
WHERE coordonate.id_sit = sit.id_sit
AND   geometrytype(coordonate.geom)   =
'POINT'::text;
```

PostGIS function ST_TRANSFORM uses the open source PROJ4 library in order to reproject the geospatial data from one spatial reference system to another. PROJ4 is a cartographic projections library used by many other GIS tools to reproject the geospatial data from one spatial reference system (SRS) to another. This function returns a new geometry with its coordinates transformed to spatial reference system referenced by the SRID integer parameter. The destination SRID must exist in the "spatial_ref_sys" table.

Some history tables of the main tables ("sit", "ansamble", and "coordonate") were created. These history tables will be filled with the old data from their corresponding tables when each maintenance action will happen. This automatically action is possible by developing some triggers.

The trigger will be associated with the specified table and will execute the specified function *funcname* when certain events occur. The trigger can be specified to fire either before the operation is attempted on a row (before constraints are checked and the **INSERT**, **UPDATE**, or **DELETE** is attempted) or after the operation has completed (after constraints are checked and the **INSERT**, **UPDATE**, or **DELETE** has completed). If the trigger fires before the event, the trigger may skip the operation for the current row, or change the row being inserted (for INSERT and UPDATE operations only). If the trigger fires after the event, all changes, including the last insertion, update, or deletion, are "visible" to the trigger [6].

For example, the trigger which was developed for "coordonate" table:

```
CREATE TRIGGER coord_audit
  AFTER INSERT OR UPDATE OR DELETE
  ON coordonate
  FOR EACH ROW
  EXECUTE PROCEDURE
process_coord_audit();

CREATE OR REPLACE FUNCTION
process_coord_audit()
  RETURNS trigger AS
$BODY$
  BEGIN
    IF (TG_OP = 'DELETE') THEN
       INSERT INTO coordonate_hist
SELECT OLD.*;
       RETURN OLD;
    ELSIF (TG_OP = 'UPDATE') THEN
       INSERT INTO coordonate_hist
SELECT OLD.*;
       RETURN NEW;
    ELSIF (TG_OP = 'INSERT') THEN
       INSERT INTO coordonate_hist
SELECT NEW.*;
       RETURN NEW;
    END IF;
  RETURN NULL;
  END;
$BODY$
  LANGUAGE 'plpgsql' VOLATILE
  COST 100;
ALTER FUNCTION process_coord_audit()
OWNER TO postgres;
```

PL/pgSQL can be used to define trigger procedures. A trigger procedure is created with the **CREATE FUNCTION** command, declaring it as a function with no arguments and a return type of trigger. The function must be declared with no arguments even if it expects to receive arguments specified in **CREATE**

`TRIGGER` - trigger arguments are passed via `TG_ARGV`.

In order to render the geospatial data about the archaeoloical sites on a map (figure 4), we have further configured the open source WMS **GeoServer**. The WMS has as output maps of spatially referenced data dynamically from geographic information. To make the interoperability possible between GeoServer and PostGIS, we have defined in GeoServer a connection to PostGIS database („data store") by providing the database name, location, port; the layers of the map („feature types") and the style in which the map will be drawn. In GeoServer, a feature type was created for each of the designed views containing the geospatial data. The style means how the layers will be drawn on the map: which colors and which symbols to use for our point or linestring or polygon data. To define a style, a XML based file named Style Layer Description (SLD) file must be developed. SLDs were developed for all the map layers, defining that the counties bounderies are drawn with black lines, the waters with blue lines and the archaeological sites with red dots (in case of POINT data) or red lines (in case of LINESTRING/POLYGON data).

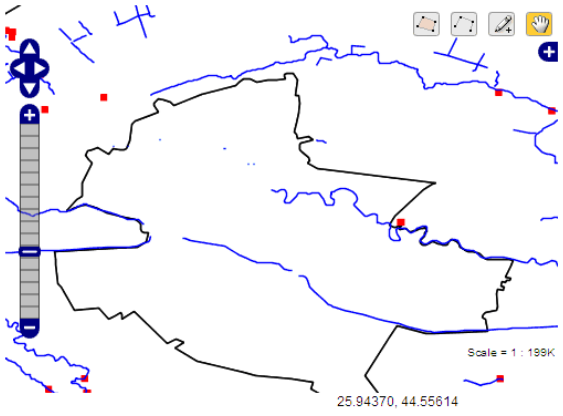


Fig. 4. Map rendered by GeoServer for the geospatial data stored in PostGIS

5 Conclusions

In this paper, we have shown that the spatial databases are modeled and designed like any other database, drawing the ER diagram, defining the conceptual, external and internal schemas, taking care though by one very important particularity: the "geometry" type column of some tables which store geospatial data as well. In order to use this kind of column further, when integrating the spatial database with a Web Mapping Server (WMS), the geospatial data has to be split by its type (point/linestring /polygon) and by its spatial reference system (SRS). Our proposal of creating different views for each type / SRS combination has proved to be an easy and efficient way for making possible the interoperability between PostGIS database and GeoServer WMS. Also, we have shown that the open source database PostGIS can successfully be used to develop a rigorous spatial database.

References

[1] D. Litan, A.M. Mocanu, S. Olaru, A. Apostu. "Modern Information Technologies Used In Market Research". *Proceedings of the 9th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics (CIMMACS'10)*, pp. 245-250, Merida, Venezuela, December 14-16, 2010

[2] A. Velicanu. "Spatial Operations". *Database Systems Journal*, vol. 1, pp. 5-8, 2010

[3] E. Meyer, P.Grussenmeyer, J.P. Perrin, A. Durand and P.Drap. „A web information system for the management and the dissemination of Cultural Heritage data". *Journal of Cultural Heritage*, vol. 8 (4), pp. 396-411, 2010

[4] M. Velicanu, M. Muntean, I. Lungu, S. Ionescu, *Sisteme de baze de date*, Ed. Petrion, Bucharest, 2003

[5] M. Bădut, *Sisteme geoinformatice pentru electroenergetică*, Ed. Polirom, Iasi, 2008

[6] PostgreSQL, *PostgreSQL 8.2.20 Documentation*, http://www.postgresql.org/docs/8.2/static/index.html

[7] Open Geospatial Consortium, *OpenGIS® Implementation Specification for Geographic information - Simple feature access - Part 1: Common architecture*, http://www.opengeospatial.org/standards/sfa

Aura-Mihaela MOCANU has graduated The Bucharest Academy of Economic Studies, Faculty of Cybernetics, Statistics and Economic Informatics in 2007. She holds a Master diploma in Databases - Support for business from 2009 and in present she is a PhD Candidate in Economic Informatics with the Doctor's Degree Thesis: Software technologies to build a Geographical Information System for a public institution. Her areas of interest are: Geographical Information Systems, Databases, Information Systems integration, Programming languages.

Manole VELICANU is a Professor at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. He has graduated the Faculty of Economic Cybernetics in 1976, holds a PhD diploma in Economics from 1994 and starting with 2002 he is a PhD coordinator in the field of Economic Informatics. He is the author of 18 books in the domain of economic informatics, 64 published articles (among which 2 articles ISI indexed), 55 scientific papers published in conferences proceedings (among which 5 papers ISI indexed and 7 included in international databases) and 36 scientific papers presented at conferences, but unpublished. He participated (as director or as team member) in more than 40 research projects that have been financed from national research programs. He is a member of INFOREC professional association, a CNCSIS expert evaluator and a MCT expert evaluator for the program Cercetare de Excelenta - CEEX (from 2006). From 2005 he is co-manager of the master program Databases for Business Support. His fields of interest include: Databases, Design of Economic Information Systems, Database Management Systems, Artificial Intelligence, Programming languages.

Conceptual and Statistical Issues Regarding the Probability of Default and Modeling Default Risk

Emilia ȚIȚAN
Department of Statistics and Econometrics
Romanian Academy of Economic Studies
Romana Sq, no 6, 1st district, Bucharest, Romania
Email: emilia_titan@yahoo.com

Adela Ioana TUDOR
PhD Candidate, Romanian Academy of Economic Studies
Romana Sq, no 6, 1st district, Bucharest, Romania
Email: adela_lungu@yahoo.com

In today's rapidly evolving financial markets, risk management offers different techniques in order to implement an efficient system against market risk. Probability of default (PD) is an essential part of business intelligence and customer relation management systems in the financial institutions. Recent studies indicates that underestimating this important component, and also the loss given default (LGD), might threaten the stability and smooth running of the financial markets. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default is more valuable than the standard binary classification: credible or non credible clients. The Basle II Accord recognizes the methods of reducing credit risk and also PD and LGD as important components of advanced Internal Rating Based (IRB) approach.
Keywords: *probability of default, stress test, PD buckets, pooled PDs, predictive analytics, data mining techniques, statistical methods, loss given default*

1 Introduction

Exposure to financial markets affects most of the financial organizations because they are involved in the risk business: there is either a possibility of loss, or an opportunity for gain. Risk can be defined as the volatility of unexpected outcomes. The risks associated with the banking sector differ by the type of service rendered and they are classified into five types: systematic risk, credit risk, liquidity risk, operational risk and legal risk.

Systematic risk (market risk) is the risk of asset value change associated with systematic factors. We mention accordingly interest rate risk and the foreign exchange risk.

Credit risk arises when counterparties are unwilling or unable to fulfil their contractual obligations. Of course, there is a risk of involuntary default, when the borrower may not have enough money to pay the loan or strategic default, when the borrower may simply refuse to pay up. The effect is measured by the cost of replacing cash flows. The real risk from credit is the deviation of portfolio performance from its expected value.

Liquidity risk has two meanings. There is market liquidity when a transaction can not be performed at current market prices due to insufficient market activity, and also we can refer to funding risk, as the inability to meet cash flow obligations. In both cases, the liquidity risk can be managed by setting limits on certain markets, products or cash flow gaps as well.

The liquidity risk is an important counterparty variable, beyond credit rating. If an obligor has high liquidity, then the one-year PD will be lower because liquidity is more important in the short term. The all-important *liquidity risk* arises from a variety of sources and, if left

unchecked, it has the potential to damage a firm's reputation.

Operational risk is associated with potential losses resulting from inadequate systems, management failure, faulty controls, fraud or human error.

Legal risks arise when new statutes, tax legislation, court opinions and regulations can put formerly transactions into contention. They include compliance and regulatory risks, which concern activities that might breach government regulations, such as market manipulation, insider trading and suitability restrictions.

The banking industry has long viewed the issue of risk management as the need of control the risks mentioned above, especially the credit risk. Understanding the various ways in which lenders manipulate and mitigate the default risk is the key to explaining some of the main features of credit markets.

Nonetheless, in a performant financial system, risk prediction is of great importance. This involves analytical processes and prediction models whose purpose is to use financial statements, customer transaction, repayment records and so on, in order to predict business performance or credit risk and to reduce the uncertainty and default. To forecast probability of default is a major challenge and it needs intense study.

In the next section, I have in view general considerations on default in the credit mechanism and several estimation methods. Section 3 refers to stress testing and stress probability of default. Section 4 is dedicated to data mining techniques which include statistical algorithms for PD evaluation and section 5 includes conclusions.

2. Default estimation and the role of PD as key factor

There is no standard definition of what 'default' means. Regulators and rating agencies define default as any of the following events: bankruptcy, write-down, 90 days past due loan or placement on internal non-accrual list. The obligor is considered defaulted as of the date of any of these accounting and financial failures.

Originally, the Basel Committee suggested that, to ensure consistent estimation of credit risk across the banking industry and provide for data sources concerning default statistics, a *default* be defined as involving one or more of four criteria:

- It is determined that the obligor is unlikely to pay its debt obligations (principal, interest, or fees) in full.
- There is a charge-off.
- The obligor is overdue more than 90 days on any credit obligation.
- The obligor has filed for bankruptcy or similar protection from creditors.

Subsequently, these four criteria have been reduced to only two: more than 90 days overdue, and unlikely to pay in full.

The IRB method, according to Basel II, allows the banks to set the capital requirements for different exposures, using their own estimations for the credit risk components. The best estimate of exposure to the counterparty will depend on:

- Probability of Default (PD), which is the likelihood that a loan will not be repayed and fall into default. PDs are largely based on credit ratings, whether internal to the bank or by independent agencies; but there are also other factors. Liquidity risk and credit risk (and therefore PD) correlate. The PD is both influenced by and impacts on liquidity.
- Loss Given Default (LGD), which is the loss recorded by the bank (as a percentage of the exposure value) when the debtor is in default.
- Exposure at Default (EAD), which is the amount of money involved in the default process.
- Effective Maturity (M) of the credit instrument.

Using their own methodology for estimating these components of credit risk is subject to approval by the supervising authority, and in some cases, banks will have to use values provided by the supervisor.
A bank may use its own values for PD and / or LGD, only if a strict set of regulations are accomplished. It settles minimum requirements to be fulfilled in order to implement a risk management system based on credit ratings internally generated.
The principle underlying these requirements is that the rating and risk estimated systems and processes should provide a relevant assessment of the counterparty and transaction characteristics, a significant differentiation of risk and a reasonably and consistent accuracy of the quantitative estimates of risk.
In addition, the systems and processes must be consistent with internal use of these estimates.
Basel Committee, recognizing the differences between markets, rating methodologies, products and banking practices in various countries, let at the discretion of national supervising authorities the development of the necessary procedures for implementation of the internal rating system.

2.1. The calculation of minimum capital requirements

In the banking system, the main tasks of the capital are:

- Protection of the deponents in the event of bank insolvency and liquidation;
- Absorbtion of the unanticipated losses to maintain trust, so that under the stress conditions, the bank can continue to work;
- Purchasing of the buildings and equipment for operation;
- Serving as a limit for the undue expansion of assets.

The regulatory capital is associated to minimum capital requirements that banks are obliged to held under the regulation of surveillance from the perspective of regulatory institution. The aim of the capital requirements is to ensure the stability and viability of the banking system.
The minimum capital requirements consist of three elements:

1. The capital definition (unchanged towards Basel I Accord).
2. The definition of the weighted assets towards risk (RWA).
3. The ratio between the capital and RWA.

The bank must maintain capital equal to at least 8% of its risk-weighted assets. For example, if a bank has risk-weighted assets of $100 million, it is required to maintain capital of at least $8 million. So, the minimum capital requirements are calculated by multiplying the amount of the weighted assets depending on risk and the percentage of 8:

$$\text{Capital} = \sum_{k}(\text{RWA}) * 8\%$$

RWA can be calculated based on two approaches: standard and internal rating .
I will focus on the second method because it makes the purpose of the article. In this case, RWA is based on the four components mentioned in the beginning: probability of default, loss given default, exposure at default and effective maturity.
For the foundation IRB approach, only PD is calculated by the bank, the remaining components of risk being provided either by the Basel Committee on Banking Supervision or by national supervising institution. In case of advanced IRB approach, all four components of risk are calculated by the bank.
Based on these four keys, for each product, RWA is calculated. For a given exposure, RWA is as follows:

$$RWA = 12{,}5 * EAD * K,$$

where K is the minimum capital for an exposure unit and it is calculated like:

$$K = LGD \cdot \left[N\left(\frac{N^{-1}(PD) + \sqrt{R}N^{-1}(0.999)}{\sqrt{1-R}} \right) - PD \right] \cdot MF(M, PD),$$

where:

- N() is the loss of the homogeneous portfolio with a probability of 99,9% and LGD of 100%. The loss is calculated based on a Merton method.
  In the Merton approach to modelling credit risk, it is assumed that a default happens if the value of an obligor's assets falls short of the value of debt. This provides financial analysts with the ability to forecast future, or *implied*, credit risk using information available at the current time. On a firm-by-firm basis, this is an important component of modern credit risk.
- LGD[N() – PD] is the unexpected loss of the same portfolio.
- R is the correlation coefficient between the assets (loans) for the same portfolio. R was estimated by the Basel Committee as:

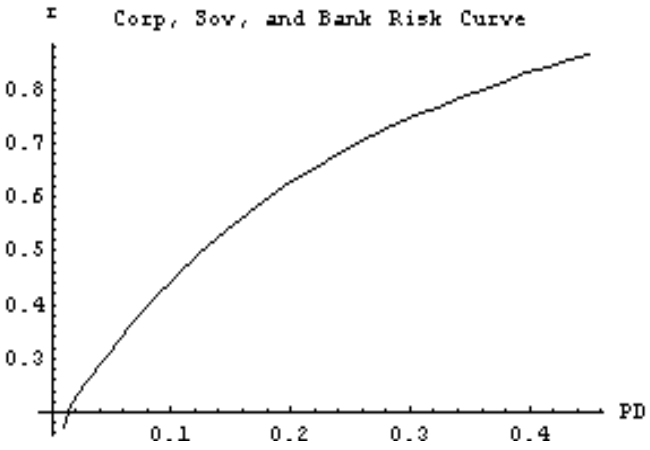
$$R = 0{,}12\left(\frac{1-e^{-50PD}}{1-e^{-50}} \right) + 0{,}24\left(1 - \frac{1-e^{-50PD}}{1-e^{-50}} \right).$$

R represents a decreasing function of PD and it has values between 12% and 24% . The debtors with a superior financial situation have a superior systemic risk towards the inferior quality debtors.

- MF is the maturity function and it is:

$$MF(M, PD) = \frac{1 + (M - 2{,}5) \cdot b(PD)}{1 - 1{,}5b(PD)}$$

where:
$b(PD) = [0{,}11852 - 0{,}05478 * \log(PD)]^2$
The MF function was obtained by the Basel Committee on Banking Supervision by calibration and it equals 1 for one year maturity.
As shown by Kiefer and Larson (2007), in case of bank exposures, the risk weight curve is generally a concave function in PD (see the graph below).



Source: Biases in Default Estimation and Capital Allocations under Basel II; Kiefer, Larson, 2007

Fig.1. The risk weight function

A calculation of the second derivative which is negative, proves the concavity and its meaning is explained in the next section.

The limits of PD are 0 and 1. The problem is often that ratings are not sufficiently responsive to changes in economic cycles, resulting in a certain overestimation or underestimation of likelihood of default over different periods. The value of 1 implies that the lender will recover all money in case of default by the counterparty, whereas the lender will recover nothing with 0. LGD is also bound between 0 and 1 and its measurement is not linear.

2.2. Estimating probability of default through probability of default buckets

The evaluation of the debtors is made through statistical models on an individual basis, by assigning individual PDs and/or individual scores. Obligors with similar

PDs/scores are then grouped into rating classes, or buckets ("PD-buckets"). Under Basel II, an IRB bank must assign obligors to risk buckets. Credit quality is the main principle that states at the basis of each bucket and it is defined by a mean value and a variance in credit standing.
The Basel Accord then requires that all obligors falling into the same bucket be assigned the same "pooled" PD (which can be thought as the mean of individual PDs). In this case, capital charges are calculated.
Related to the variance in credit standing, which is bound to exist with every pool, all banks have an interest in establishing a level of confidence at 99.9 per cent.
Banks with experience in the implementation of Basel II rules suggest that, to apply the method of PD buckets properly, user organizations should provide themselves with the means to continue drawing a distinction between the concepts of:
- A default probability linked to an individual obligor, and
- The pooled PD assigned to a credit risk bucket.
A PD associated with an individual obligor is a metric of the probability that this obligor will default during a one-year credit assessment. By contrast, the pooled PD assigned to a risk bucket is a measure of the average value of the PDs of obligors in that bucket.
Related questions have been addressed in the literature:
- How should pooled PDs be derived that reflect the PDs of obligors assigned to each risk bucket in an accurate manner?
- How should deviations from the pool's mean value be accounted for and presented?
- How should PD bucket mean values and variances in credit risk, among individual pool members, be stress tested?
There is no clear guideline on how this pooled PD could or should be stress tested. (Dimitris N. Chorafas, 2007)

PD is a continuous variable, taking values between 0 and 1, so there are infinitely many possible ways to partition the 0-1 interval into a set of discrete intervals (the PD-buckets). The choice of the "optimal" buckets (sometimes referred to as "PD bucketing") is rarely reached analytically by banks. Most of the times, banks offer a defining label of the rating buckets like "very good"or "AAA" and a set of rating criteria which help their analysts to sort obligors into different classes.
Therefore, the buckets should be chosen carefully, since all obligors falling into a given rating class will eventually be assigned the same PD. As individual PDs within a bucket are expected to be similar, but not exactly equal to one another, replacing them with a pooled PD obviously causes a loss of precision in the rating system.
On the other hand, the buckets must contain a high number of observations in order to have a precise assessment.

In this case, „the concavity of the risk weight curve (see fig. no. 1) means that, if two rating classes (having pooled PDs of p-k and p+k), containing an equal number of obligors, are pooled together into a single bucket with an average PD of p, the new capital charge $C(p)$ will be more than the sum of the capital charges on the two separate classes:
$C(p) > ½\ C(p-k) + ½\ C(p+k)$
Second, the use of pooled PDs instead of individual ones could cause opportunistic behaviour and adverse selection phenomena among a bank's customers. Indeed, if a single rating bucket were to include a very wide array of individual PDs, all replaced by the same pooled PD and treated as equally risky for pricing and risk management purposes, then the best customers in the bucket would feel they can get substantially lower lending rates elsewhere, while the worst customers in the bucket would stay as they feel their credit risk is significantly underpriced. "(T. Krink, 2008)

Basel Committee argues that the default probability assigned to each debtor depends strongly on the type of rating methodology and quantification techniques being used. There are several important approaches for quantifying pooled PDs. The significant ones include the historical default method, statistical model approach and external mapping.

The actuarial approach or the *historical default method* consists of recording the default events over several years assigned to the specific bucket. The algorithm is:

$DF_t = D_t/N_t$, where DF_t = default frequency;

D_t = number of defaults observed for a bucket over year t;

N_t = number of debtors assigned to that bucketat the beginning of year t.

In order to estimate a default probability for each obligor assigned to a bucket, there are used predictive statistical methods.

Therefore, the bucket's pooled PD is then calculated as the *median of obligor PD*. This approach to individually quantifying pooled PDs can produce accurate estimates of credit exposure, gaining an important advantage over the historical default alternative.

Within the *external mapping*, a bank simply establishes a connection between its internal rating system and an external scale such as that of big rating agencies, calculates a pooled PD for each external grade using an external reference dataset, and then assigns the pooled PD for the external grade to its internal grade by means of the mapping. Despite its apparent simplicity, this approach poses some difficult validation challenges for risk managers. To validate the accuracy of a bank's pooled PDs, supervisors and risk managers must first confirm the accuracy of the pooled PDs associated with the external rating scale. They must then validate the accuracy of the bank's mapping between internal and external grades. Quantifying pooled PDs for an external rating system poses the same estimation problems as quantifying pooled PDs for a bank's internal rating system. If a historical default experience approach is used, supervisors and risk managers must check to ensure that each bucket's pooled PD can be expected to approach its long-run default frequency over time. If a statistical models approach is used, supervisors and risk managers must validate the reliability of the underlying default prediction model. The main benefit of quantifying PDs using external ratings is that more data are likely to be available for calculating long-run default frequencies and/or estimating statistical default prediction models.

3. Stress testing

Stress tests are an important risk management tool that has been used for a number of years now, both by banks as part of their internal risk management practices and by supervisors to assess the resilience of banks and of financial systems in general to possible shocks (European Central Bank, 2010).

This method is also called *scenario analysis* and it consists of specific scenarios of interest in order to assess possible changes in the value of the portfolio.

In my opinion, the key role of the stress tests is to draw attention of how much capital might be needed to absorb losses in case of a financial crisis or other shocks and therefore increase the banks resistance in recession times.

The importance of these tests is bigger in a stable economy because, due to the fact that there are no special risks, the banks might not be aware of the major impact of a financial crisis upon their stability. Practically, stress testing forces management to consider events that they might otherwise ignore.

Fig. 2 shows a stressed and unstressed probability of default and it is based on the study by the Basel Committee on Banking Supervision in February 2005. In boom times, the stressed probability of default acts like a prevention of the unstressed probability of default in case of financial crisis. This way, it would be much easier to determine the value of the collateral that should be asked for.

It is quite obviously that SPDs tend to remain relatively stable over a business cycle compared with unstressed PDs. During economic expansion the unstressed PD declines and the obligor receives a higher rating; but during economic recession the unstressed PD increases, closely approximating stressed PD, and the obligor receives a lower rating.

Considering the stress scenarios associated with obligor-specific PDs, I believe that the economic environment is not sufficient in order to offer a pertinent result regarding the creditworthiness of the debtor. The tests should also contain both information relevant to assessing the obligor's ability and willingness to repay its debts, and macroeconomic variables (interest rate levels, market liquidity, inflation rates etc).

Basel Committee gives the following definitions for the classic PD and SPD:

- An *unstressed PD* is an unbiased estimate of the likelihood that an obligor will default over the next year, given all currently available information, including static and dynamic.
- A *stressed PD* (SPD) measures the likelihood that an obligor will default over the next year, using all available information, but assuming adverse economic and lender-specific conditions for the stress scenario.

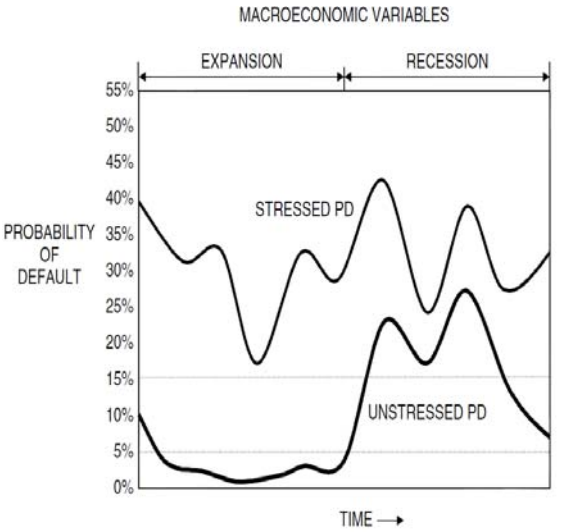


Fig 2. Unstressed and stressed probability of default, over time (time is expressed in years). Macroeconomic variables include a growth or downturn in gross domestic product, exchange rates and market psychology

Source: based on a study by the Basel Committee on Banking Supervision

But, the choice of scenarios may be affected by the portfolio position itself. For instance, one month the portfolio may be invested in a national fixed-income market; the scenario will then focus on interest rate shifts in this market. The following month, the portfolio may be invested mainly in currencies. If scenarios change over time, measures of risk will change just because of these changes. Also, stress testing does not specify the likelihood of worst case scenarios. Expected risk should be a function not only of the losses but also of the probability of such losses to occur.

The stress tests implemented by the banks have registered some deficiencies lately. The amplitude and the severe current financial crisis has determined many banking institutions and supervising authorities ask if the stress tests used before this crisis were quite efficient and helped the banking sector to face this real challenge.

The financial crisis showed several lacks in the stress tests systems of the banks

especially regarding the crisis scenarios and the methodologies used for crisis simulation.
In many banks, the stress tests were done only for specific activities or risks, without being considered an aggregation of results on the overall bank. Another issue is that most of the risk management methods, including stress simulations, use statistic data in order to assess the future exposures at risk. These data are based on long periods of economic stability and are not sufficient to identify a crisis. The banks underestimated the strong correlation between the lack of liquidities on the market and the financing pressure. Therefore, it is crucial to treat correctly the dependencies between different risks and integrate them on the overall financial group or bank.

4. Data mining techniques used for predictive default probability

In our days, data mining is an indispensable tool in decision supporting system and it is defined as "the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large databases" (Turban, Aronson, 2007). Practically, data mining is a technique for extracting knowledge from information. This analysis process has a significant role in probability of default estimation, credit scoring, customer services, fraud detection and market segmentation.
The most important techniques are: discriminant analysis, logistic regression, artificial neural networks and K-nearest neighbour model.

4.1. Discriminant analysis (DA)

DA or Fisher's rule is is a classification method that projects n-dimensional data onto a line, and performs classification in this one dimensional space. The projection is chosen so as to maximize the between-class mean, and minimize the within-class variance (R. Khemchandani, 2009).
In DA, a group of observations are used to measure parameter estimates of a discriminant function by minimizing the group misclassifications. This method is used in the decisional situations. For instance, DA provides data regarding the possibility of a loan application to default.

4.2. Logistic regression (LR)

A LR model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. (I-Chang Yeh, 2009)
This is a special case of linear regression models.
In logistic regression, there is no definition of the coefficient of determination (R^2) which is frequently used in the general linear model. R^2 has the desired interpretability as the proportion of variation of the dependent variable, which can be explained by the predictor variables of a given regression model.

4.3. Artificial neural networks

Artificial neural networks (ANN) are flexible computing frameworks and universal approximations that can be applied to a wide range of time series forecasting problems offering solutions in many fields, such as control and pattern recognition. ANN use non-linear mathematical equations in order to develop adequate correlations between input and output variables.
One of the major developments in neural networks over the last decade is the model combining or ensemble modelling. The basic idea of this multi-model approach is the use of each component model's unique capability to better capture different patterns in the data. Both theoretical and empirical findings have suggested that combining different models can be an

effective way to improve the predictive performance of each individual model, especially when the models in the ensemble are quite different (Baxt, 1992; Zhang, 2007).

4.4. K-nearest neighbor model (KNN)
Nearest-neighbour (NN) techniques are non-parametric classification systems based on learning by analogy. Given an unknown sample, a KNN classifier searches the pattern space for the KNN that are closest to the unknown sample. This means finding out the shortest distance. In learning systems, generalisation performance is affected by a trade-of between the number of training examples and the capacity (e.g. the number of parameters) of the learning machine. The major advantage is that it is not required to establish predictive model before classification.

Empirical studies in literature outline that in the predictive accuracy of probability of default, artificial neural networks show the best performance based on R^2, regression intercept and regression coefficient. Therefore, ANN should be employed to score clients instead of other data mining techniques, such as logistic regression. (I-Cheng Yeh, 2009).

5. Conclusions

The management of financial risks has many dimensions and involves many types of decisions. The importance of this article comes from the complex issue of credit risk management in order to assure financial stability. Credit risk is considered the most dangerous category of banking risk and in order to prevent it, banks must meet a series of regulations.
Recent studies show that default probabilities and average recovery rates are negatively correlated (see e.g. Altman et al. (2005); Acharya et al. (2007)). Both variables also seem to be driven by the same common factor that is persistent over time and clearly related to the business cycle: in recessions or industry downturns, default rates are high and recovery rates are low. Although the actual researches in this area are very elaborated, both in Romanian and foreign literature, through empirical and theoretical studies for loss predicting in case of default or sofisticated credit scoring models, at present the international financial crisis has revealed serious shortcomings and limitations in managing credit risk. Therefore, I believe a progressive research is required in terms of effects generated by the current international crisis.

Acknowledgements
This article is a result of the project POSDRU/88/1.5./S/55287 „Doctoral Programme in Economics at European Knowledge Standards (DOESEC)" . This project is co-funded by the European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies in partnership with West University of Timisoara.

References

[1] A. Feelders, H. Daniels, M. Holsheimer (2000). Methodological and practical aspects of data mining. Information and Management.
[2] Altman, E.I., Brady, B., Resti, A., Sironi, A., 2005. The link between default and recovery rates: theory, empirical evidence, and implications. Journal of Business 78.
[3] Balthazar, L., 2004. Pd estimates for Basel II.
[4] Basel Committee (2006) Basel II: International Convergence of capital measurement and capital standards: A revised framework.
[5] Basel Committee on Banking Supervision, Working Paper No. 14, *Studies on the Validation of Internal Rating Systems*, BIS, Basel, February 2005.

[6] Bielecki, T. & Rutkowski, M. (2002), Credit Risk: Modelling, Valuation, and Hedging, Springer, Berlin.
[7] Credit Risk Analytics: A Cornerstone for Effective Risk Management, *An Oracle White Paper, October 2008*
[8] Crosbie, P. & Bohn, J. (2002), 'Modeling default risk', KMV working paper. Available from http://www.kmv.com.
[9] Crouhy, M., Galai, D. & Mark, R. (2001), Risk Management, McGraw-Hill, New York.
[10] Delbaen, F. (2000), 'Coherent risk measures', lecture notes, Cattedra Galiliana, Scuola Normale Superiore, Pisa.
[11] Dimitris N. Chorafas (2007). Stress testing for risk control under Basel II.
[12] Du, Y. (2004), Credit rating, default probability and structural credit risk models. PhD. Queen's University at Kingston, Canada.
[13] George L. Head, Ph.D, CPCU, ARM, CSP, CLU. Risk Management- Why and How. An illustrative introduction to risk management for business executives. International Risk Management Institute, Inc.
[14] Hamerle, A., M. Knapp and N. Wildenauer. 2007. "Default and recovery correlations."
[15] Hull, John C. (2007) „Risk Management and Financial Institutions", Prentice Hall.
[16] I-Cheng Yeh, Che-hui Lien (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default. Expert Systems with Applications.
[17] Lando, D. (2004), Credit Risk Modelling: Theory and Applications, Princeton University Press, Princeton, New Jersey.
[18] Krink, T. Paterlini S. & Resti A. (2008) The optimal structure of PD buckets. Journal of Banking and Finance, 32.
[19] Moody's Investors Service, 2006, *Default and Recovery Rates of Corporate Bond Issuers, 1920-2005*, March.
[20] Nicholas M. Kiefer, C. Erik Larson. Biases in Default Estimation and Capital Allocations Under Basel II
[21] Tasche D., 2003, *A traffic lights approach to PD validation*, mimeo, Deutsche Bundesbank.
[22] Vasicek, O.A., 1997, *The Loan Loss Distribution*, Technical Report, KMV Corporation, San Francisco.
[23] Wilde, Tom, Jackson, Lee, 2006. Low-default portfolios without simulation.

Mrs. Emilia Ţiţan received the PhD title in 2005 and at present, she is professor at the Department of Statistics and Econometrics, in the Academy of Economic Studies, Bucharest. Since 2008, she has been a deputy dean of Cybernetics, Statistics and Economic Informatics faculty of the Academy of Economic Studies. Her research activity is found in over 50 articles and 20 books.

Adela Ioana Tudor has graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies in 2002. Two years later she graduated the MA in Financial Management and Capital Markets – DAFI, Faculty of Finance, Insurance, Banking and Stock Exchange Market. At present, she is a PhD student in the field of Economic Cybernetics and Statistics at the Academy of Economic Studies.

Modelling Financial-Accounting Decisions by Means of OLAP Tools

Diana Elena CODREANU
Universitatea "Constantin Brancoveanu", Piteşti
codreanudia@yahoo.com

At present, one can say that a company's good running largely depends on the information quantity and quality it relies on when making decisions. The information needed to underlie decisions and be obtained due to the existence of a high-performing information system which makes it possible for the data to be shown quickly, synthetically and truly, also providing the opportunity for complex analyses and predictions.
In such circumstances, computerized accounting systems, too, have grown their complexity by means of data analyzing information solutions such as OLAP and Data Mining which help perform a multidimensional analysis of financial-accounting data, potential frauds can be detected and data hidden information can be revealed, trends for certain indicators can be set up, therefore ensuring useful information to a company's decision making.
***Keywords*:** analytic proces, data warehousing, databases, decision support systems, OLAP cubes.

1 Introduction

In the current context of the world economy, the performance of an organization is both ensured and conditioned by the quality of the decisions made by its manager. Making the best decision relies upon a large amount of information as well as a complex process of its analysis and synthesis.

When analyzing the global efficiency of an enterprise, the quantity and quality of financial-accounting information in the system of the economic unit at a precise moment is specially important, preferably the information at more and more analytical levels [1].

To an enterprise, the use of financial-accounting information is the raw material when making the best managerial decisions.

A decision is the result of a conscious process of choosing an action direction and getting involved in it which usually implies the allocation of resources. A decision ensues as a consequence of processing certain information and knowledge, and belongs to a person or a group of people who has/have necessary authority and is/are responsible for the efficient use of resources in certain circumstances [2].

OLAP Tools

OLAP is the accronym of On-Line Analytical Processing and, as the words themselves show, it can be said that OLAP within an organization is used to provide easy and fast access to the analytical resources underlying decision-making and management processes.

Specialists think that an OLAP system is an information server which allows quick access to data (atomic and derived data) and complex calculation facilities [3].

OLAP has quickly become the basis of smart solutions and one can also mention the high-performing management in business environments, planning, resource allocation, budget allocation, predictions, financial report drafting, data discoveries and data storehouse reports.

The 60's -70's were hallmarked by the emergence of the concept of On-line Analitycal Processing used to model the financial activities in an organization by means of analytical functions. Thus, in the year 1962, Ken Iverson, in his paper called "*A Programming Language",* described multidimensional language for the first time, namely *APL* (*A programming Language*). The language has been

developed by IBM Company and used on mainframes ever since 1962. A large amount of the concepts in this language are still used nowadays and examples are the languages called Adaytum Planning and Lex 2000.

OLAP Council was founded in 1995, made up of companies concerned with developing OLAP products and having as main goal the removal of confusion and the desire for OLAP systems to become more interesting on the market by setting up open standards (OLAP API). OLAP Council has defined OLAP technology as a *"category of software tools which allows analysts, managers and directors to understand the essence of data by the rapid, substantial and interactive access to a wide range of potential visions upon information which has been obtained by the change of primary data so that to reflect an enterprise's real dimensions as perceived and understood by users"* [4].

The analytical functions and management facilities of data were integrated within a language in 1972 called Express Language. Express is still one of the main OLAP technologies in use, as its data model and concepts have remained unchanged.

OLAP provides users with the opportunity to analyze several dimensions on the spot, making sure all the necessary information is available to make the best decisions.

OLAP technology is characterized by a dynamic, multidimensional analysis which helps end-users by means of various activities [5]:

- in-depth analysis (drill-down);
- opportunity to make predictions during various periods of time;
- use of formulae and models for dimensions and rankings;
- extracting a sub-set of data to be viewed;
- rotations within dimensions.

OLAP systems have been included in data-oriented decision-making support systems and here are a few SSD notions since decision support systems (SSD) theory leaves its mark upon the OLAP systems theory:

- In early 70's, the first definition of SSD was stated by Little who asserted that an SSD is *a model relying on a series of procedures for data processing and for assisting a manager during decision making. An SSD should be simple, robust, easy to maintain, adjustable, available for communication etc.* The features identified by Little are still valid at present.
- In the paper entitled "Foundation of Decision Support Systems", authors Bonczek and Holsaple defined an SSD as: *"information system made up of three interacting constituents: buffer to users (Management Dialogue), Data Management and Model Management".*
- In 2002, Power defined an SSD as : *"interactive information system meant to help decision makers use data, documents and models in order to identify and solve problems, and make decisions".*

OLAP is a way to provide answers to complex inquiries of the data bases. OLAP is part of what is called Business Intelligence along with ETL (Extract, Transform and Load), relational reporting and data mining [6].

Since OLAP tools work with multidimensional data models, perform on-spot complex, analytical inquiries and have a high processing speed, certain specialists have suggested that they should rather be called FASMI (Fast Analysis of Shared Multidimensional Information).

Functional requirements of OLAP systems

OLAP systems must meet the following functional requirements [5]:

– *Dynamic analysis of data* – it implies the existence of various analysis tools as well as multiple dimensions, with an emphasis upon the control of an

enterprise's data models. The dynamic analysis of data helps better understand the causes of changes within an organization and it is also used to find solutions.
- *Quick access to data* – OLAP applications use a large amount of data and imply very quick access.
- *Multiple data sources* – most OLAP applications access data in several sources, including external ones and various applications performed in different programmes.
- *Data sources' synchronization* – if the data in an OLAP application come from several data bases, it is possible they are modified in various cycles.
- *Historic analysis* – almost all OLAP applications use time as a dimension and therefore the useful results come from time series analyses. In order for that to happen, it is necessary the data stored in data warehouses or data marts should last at least two, three years.
- *High generalization level* – the decision makers in an organization require the information to be grouped, aggregated and shown as synthetically as possible. As a rule, in order to ensure high efficiency and quick access to data, the latter are merged and aggregated at a high level and decision makers can also view detail levels if they request it.

Erik Thomson divided OLAP systems' functional requirements into two large categories: *logical and physical* [7].

Logical requirements ensure the possibility to process dimension data, structure data and render systems flexible. They are the following:
- Complete structuring of dimensions by ranking – an OLAP system should model the dimensions in an organizational environment according to certain hierarchies and at various levels, starting from the most detailed to the highest, generalized level.
- Efficient calculations and processing – OLAP systems should ensure the accomplishment of complex analyses, data comparison and prediction opportunities.
- Flexibility – the data obtained after processing should be shown according to users. OLAP systems should be flexible in all ways. Flexibility allows a model to be changed by a user without changing (redesigning) the entire system.
- Independence of presentations to a model's structure – an OLAP system should ensure the opportunity of not affecting data structure if a presentation changes.

Physical requirements refer to the access and response time of the system as well as to the multi-user support:
- *Quick and direct access* – OLAP systems should provide support for on-spot analytical requests for large data amounts. OLAP Council believes that the main goal of OLAP systems is to *"provide response duration of five seconds or less, irrespective of the request type or data base size, within a multi-user and shared environment"*.
- *Multi-user support* – due to the fact that the data stored in a data warehouse are accessed by several users, OLAP systems should ensure competitive, shared access to analytical processing.

E.F.Codd used the OLAP notion for the first time in 1993, in an article called "*Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate"*, and he was subsequently named the "father of OLAP concept". He identified twelve rules that OLAP systems should meet.

E. F. Codd outlined the processing difference between relational and multidimensional models, stating that *"irrespective of how powerful relational systems might be to users, the former are not designed to provide strong data synthesis, analysis and enhancement*

functions, collectively known as the multidimensional analysis of data".

Later on, in 1995, the twelve rules set up by E. F. Codd were improved and there were eighteen rules, shown below:

A. Basic features

Rule 1: A multidimensional conceptual image

Codd believed that a user's image of an organization is multidimensional and therefore the conceptual image of OLAP models should be multidimensional, too, and should rely upon the image or model existing within an organization.

Rule 2: Intuitive control of data

OLAP systems should allow data control operations (drill down, drill up, drill across operations), as many OLAP tools permit the intuitive control of data (for example, Microsoft OLAP, Express etc.).

Rule 3: Accessibility

OLAP systems should provide a single logical image of the data in an organization. Within an OLAP model, data sources should be transparent to users. E. F. Codd believed that even users can be a source of data.

Rule 4: Varied data sources

An OLAP system should have the ability to work with data stored in multidimensional data warehouses (MOLAP) or in relational data warehouses (HOLAP).

Rule 5: OLAP analysis models

There are four analysis models that OLAP tools should hold: explanatory, direct, contemplative and forming. Therefore, OLAP tools should at least ensure the drafting of standardized reports, analyses such as *"what happens if...?",* drill-down/roll-up and slice/dice operations.

Rule 6: Client/server structure

An OLAP system should provide the client/server structure, allowing the access of users by means of a client and aiming at performing multidimensional processing by means of a specialized server.

Rule 7: Transparency

OLAP systems should permit the access to heterogeneous data sources which must be transparent to users and include buffers to various client-tools, such as: table calculation sheet tools, text editors etc.

Rule 8: Multi-user support

OLAP systems should provide competitive, shared access to data, at the same time making sure the data are true and safe.

B. Special features

Rule 9: Data denormalization

When data are processed within an OLAP environment, the external data used as sources should not be affected. OLAP tools process large data collections upgraded periodically and in order to do that, they should be able to persistently connect to external data sources, in order to ensure the synchronization between external sources and data cubes. OLAP systems are generally separated from source systems and that is why the connections serve as transformation functions which indicate the way table data or table calculation sheet data are turned into multidimensional data.

Rule 10: Storage of results generated by OLAP systems

The data subjected to analysis have to be stored and processed separately from relational sources or the folders they come from. This condition should be met as a consequence of the differences between operational data and the ones meant for decision-making support.

Rule 11: Control of missing values

The *"spreading"* term has been used with the meaning of missing value, inapplicable value and zero value. The first two are invalid data (nule data). The third, where spreading means the existence of several zero values, is a special case of the way in which a large number of repeated values are stored, zero value here. Yet, zero value is as valid as any other number. The confusion emerged because there are many zero values in OLAP applications,

where there are also a lot of missing and non-valid data. The techniques for the physical improvement of many repeated values' storage are similar and sometimes identical with the ones used to physically improve the storage of numerous missing and non-valid data. However, missing and non-valid values are not valid data. They cannot be treated the same as any other value. Therefore, special techniques are necessary in such circumstances [8].

Rule 12: The way to treat missing values

To any analysis of any data series, either multidimensional or not, the accuracy of calculations is a major requirement. Treating spread data is very important and often debated upon in the field of data bases. The two types of data (missing and non-valid) should be however treated individually, as they affect calculations in different ways [8].

C. The way to show data

Rule 13: Flexibility of reports

The data which are subject to analysis should be accessibly shown to users, so that they can arrange their data according to various dimensions on available axes.

Rule 14: Performance of reporting

The performance of reporting should not be influenced by data's dimension or organization manner. However, there are two factors which affect the performance of reporting: the way calculations are made (they can be calculated before or during an inquiry) and the place where calculations are made (client/server). It can be stated that the importance of these factors is higher than the dimension of a data base or the number of dimensions.

Rule 15: Automatic adjustment of physical level

OLAP systems should automatically change the physical structure of data bases, according to the type of logical model and to the data amount.

D. Control of dimensions

Rule 16: Generic dimensionality

Dimensions should be structurally and operationally even. It implies that dimensions should allow multiple ranking as well as all types of multidimensional operations (adding/deleting a member, adding/deleting a hierarchy, changing a member or hierarchy etc.).

Rule 17: Unlimited aggregation dimensions and levels

Codd thinks that a maximum number of 15-20 dimensions should be used. In practice, there are many other requirements and limitations of OLAP tools, so that the issue of the maximum dimension number can become a minor, unsignificant requirement [5].

Rule 18: Operations between non-restrictive operations

OLAP systems, by means of the control language, should allow operations among various dimensions, without any restriction.

OLAP cube

OLAP technology uses data structures called cubes which are organized in multidimensional data bases. The process of defining the structure of cubes is called multidimensional modelling, the same as the design of a data warehouse structure. Since an OLAP cube is used to extract the information needed in the decision making within a data warehouse, a cube's constituents are similar to a data warehouses's.

Multidimensional data warehouses are optimized data structures used to exploit the data stored in data warehouses and for OLAP analyses.

One can say that multidimensional modelling is the core of OLAP technology. Multidimensional modelling helps one show the results of certain economic activities closely related to one or several factors that have been part in their make-up.

Specialized literature includes several definitions of the "cube" concept of which one is that of SQL Server Book Online, which states that a cube is a

”multidimensional data structure that contains dimensions and measures”.

A cube can be seen as a data sub-set within a data warehouse where the existing data have multidimensional structures.

Aggregated and hierarchical data are stored within a cube. For instance, if the sales of a specific product are inquired during a certain year term, an OLAP cube helps classify the sales of the respective product by months, weeks and days.

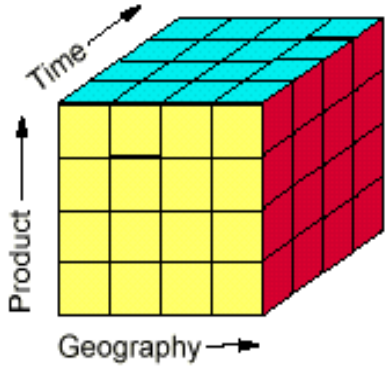


Fig. 1: OLAP Cub

A cube has the following constitutive elements:

- Dimensions are the features according to which the analysis of an activity’s measures is made [9]. Dimensions are the main elements used to define a cube. Dimensions are those which allow data to be viewed and they may contain a single data viewing criterion. Practically, dimensions are those which make an ”in-depth” analysis (drill-down) of data, starting from a lower detail level and gradually reaching more and more detailed levels. For example, time, or geographic dimension (such a dimension helps describe the organizational structures of a company from the perspective of its placement).

Dimensions are twofold: private (Private Dimensions) and public (Shared Dimensions). Shared dimensions are made independently from any cube whereas private dimensions are created along with a decisional cube and they are saved within the respective cube’s warehouse. One can choose private dimensions when a dimension implements the logical viewing of data which is available just for the respective cube. In all the other circumstances, it is recommended that shared dimensions should be defined by means of which the data can be viewed uniformly.

A dimension is characterized by the following elements:

- *Dimension name;*
- *Dimension structure* (also known as attributes): it is the set of elements that characterize a dimension;
- *Dimension members:* the values of dimension attributes;
- *Dimension levels:* a level emerges as a consequence of aggregating the members of a dimension that have a feature in common;
- *Dimension ranking:* the tree structure of father-child levels of the respective dimension.

- Measures are the elements that help describe the indicators meant for analysis (for example, the value of a company’s expenses, the value of a company’s loans etc.). When the measures of a cube are defined, it is necessary to take account of the following [10]:
  - Measures should reflect numeric values;
  - Measures should be stored in the cube’s fact values;
  - Measures should be associated to each level in each dimension.

It is compulsory that for each fact table within a data warehouse at least one measure should exist.

- Facts
- Fact tables
- Dimensional tables

From the architectural perspective, cubes can be organized in two ways:

a. *Star:* just like a data warehouse, a star-shaped cube has all its dimensional tables directly connected with the fact

table. In most cases, the starlike
b. is more frequently used as it has advantages such as: it is more easily maintained and provides higher performance at the same time.
c. *Snowflake:* within such a scheme, a dimension is represented by several dimensional tables. Unlike the star structure, the snowflake one has a more readable dimensional model when the amount of data is too large.

Architecture of OLAP systems

OLAP systems have a basic architecture structured in three main constituents:

a. Data base – is the data source used for an OLAP analysis. A relational data base may be used as a data base to ensure facilities for multidimensional storage,

structure
multidimensional data base, data warehouse etc.

b. OLAP server – is the one managing the multidimensional structure of data, simultaneously accomplishing the connection between a data base and an OLAP client.

c. OLAP client – is represented by the applications which ensure the exploitation of data and are also a support for generating results (graphs, reports etc.).

The following graph shows the architecture of OLAP systems which varies according to the way of data storage and to the type of their processing, yet seen in broad terms, three data levels can be identified: the level of data sources, the level of OLAP server and the level of data show or of the buffer to users.

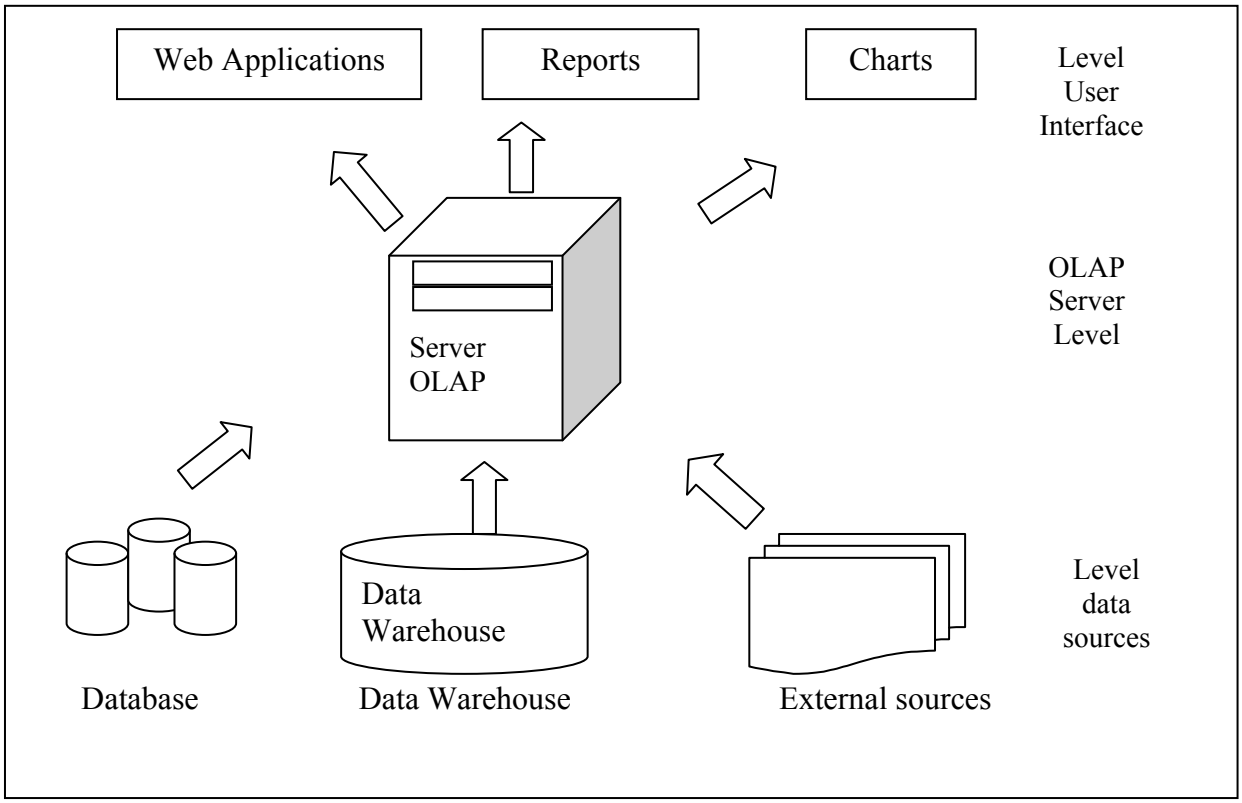


Fig 2: OLAP Architecture

According to the type of data base used for analyses, other versions of OLAP systems have emerged, namely:

– MOLAP systems
– HOLAO systems
– ROLAP systems

MOLAP systems (Multidimensional OLAP) – are seen as traditional solutions for multidimensional analyses. They store both basic and

aggregated data in a multidimensional data base which is called a cube and they are used as efficient tools for the operations during analyses as well as during complex calculations. Analyzing the space they cover on a disk along with the time they take in order to reply to complex inquiries, MOLAP cubes can be said to be the best performing.

Among the advantages of MOLAP systems are the ones below:

- Relational tables are not suited to multidimensional data;
- Multidimensional matrices allow the efficient storage of multidimensional data;
- SQL language is not adequate to multidimensional operations.

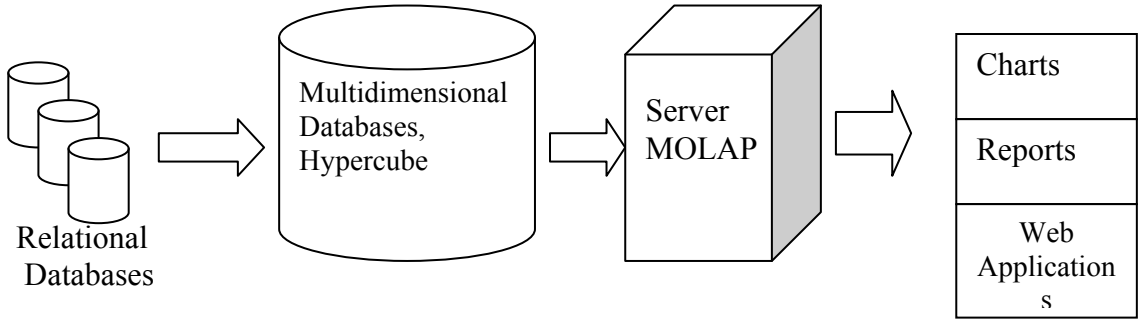


Fig 3: MOLAP Architecture

ROLAP (Relational OLAP) – here, the organization of the data meant for analysis is done in the form of a relational data base. Storing the basic and also the aggregated data is done within a relational data base and memorizing the dimensions' scheme is done within the ROLAP cube.

ROLAP helps exploit large amounts of data and sometimes data base management systems provide facilities typical of ROLAP. As a comparison between ROLAP and MOLAP, it can be said that ROLAP systems are much slower. Practically, a disadvantage of ROLAP systems is the time they take when responding to informational inquiries which can sometimes be quite long when it is about complex inquiries with multiple data sources.

HOLAP (Hybride OLAP) – can be called a compromise variant between the first two versions, which tries to combine the advantages provided by MOLAP and ROLAP, in order to ensure that users receive the best solution from the performance perspective. As to HOLAP, basic data are memorized in relational data bases whereas aggregated data bases are memorized in the HOLAP cube.

Relationship between OLAP and Data Warehouse

OLAP systems have been included in data-oriented decision support sytems, although they are hybrid decision support sytems, because they use simple analytical techniques (data's multidimensional analysis) in the analysis of large amounts of data. Most specialists in the field believe that data warehouses and OLAP tools provide the necessary support to turn companies' large amounts of data into information which is useful to decision makers.

A data warehouse mostly relies on the processes that help ensure the substantiality, truthfulness and validity of data, whereas OLAP systems rely on analytical requirements, modelling processes and necessary calculations. Bill Inmon, who was named "the father of data warehouse concept", emphasized the idea that such a data warehouse mainly aims at ensuring the substantiality and truthfulness of data used.

At present, due to the harsh competitions on the market, decision makers need information that should be provided constantly, rapidly and easily, information which can be given by the analytical techniques supplied by OLAP tools, data warehouses and Web facilities, too.

To ensure fast access to information necessary for managers, analysts in an interactive and flexible OLAP instrumetele use. We can say that completes OLAP and Data Warehouses are mutually OLAP making huge amounts of data stored in data warehouses, but the information necessary and useful to decision makers.

Several technologies are used to analyze the financial-accounting data stored in a data warehouse, among which the most frequent is OLAP technology (from the perspective of actual use and of the existing software support).

OLAP technology helps use the financial-accounting data stored in a data warehouse in an efficient way during on-line data analyses, supplying a fast reply to complex inquiries. OLAP multidimensional model along with specific aggregation techniques ensure the organization of large data series which allow an easy and prompt interpretation. OLAP provides data analysts with the necessary flexibility and work speed when underlying decisions in real time.

The connection between OLAP technology and data warehouses can be defined as follows: ”whereas data warehouses furnish data management, OLAP implements technologies that turn the data into strategic information” [9].

Using OLAP technology helps data warehouse users by several advantages, of which [11]:

- quick execution of inquiries – as it gives one the possibility to keep within a cube some values calculated before an inquiry, such as aggregated values;
- meta-data-based inquiries – an example is MDX Language which ensures the native generation of search criteria for inquiry results;
- calculation formulae similar to the ones used in table calculation applications – one of the advantages of a table processor is the fact that its users can generate formulae by the addresses of the cells where the necessary calculation values are stored. The use of such a formula is easy within the OLAP environment, too, since the address of any cube cell can be used in formulae.

The importance of a warehouse is not the amount of data, but the quality of stored information. Data warehouses were named *”data jailhouses”* by Aaron Zornes, a famous analyst.

Conclusions

Making the best decisions either helping to solve certain problems occurred in an organization’s economic-financial activities or to ensure an organization’s good running, depends on the quantity and quality of information provided by an economic information system.

Managers have always needed correct and rapid information in order to make the best decisions. Data have always been at hand, yet there has been an impediment when extracting and processing them which have been quite slow. Thus, in order to get rid of such an impediment, data have been organized into multidimensional cubes which have been easy to ”rotate” and ”slice”.

OLAP systems are helpful to managers who, after the process of data analyses, can make the best decisions related to the good running of their companies. OLAP systems help make predictions of revenues and expenses, analyses of sales, and identify the progress of financial indicators. OLAP technology helps managers have their own perspectives upon data in general and financial-accounting data in particular.

References

1. C. Crecană, *Profitability of Small and Medium Entreprises,* Editura Economică, Bucharest, 2000.
2. F.G. Filip, *Computer Aided Decision: Decision makers, basic methods and tools associeted*, Editura Tehnică, Bucharest, 2005.
3. M. Velicanu, M. Munteanu, *Characterization of Current OLAP Systems*, Economic Informatics Review, no.3(19)/2001.
4. The OLAP Council Definitions, January 1995, www.olapcouncil.org
5. I. Lungu, A. Bâra, *Executive Information Systems,* Editura ASE, Bucharest, 2007
6. www.marketwach.ro/artico/816/O_abordare_practică_a_bazelor_de_date_OLAP
7. E. Thomson, *OLAP Solitions: Building Multidimensional Information Systems*, John Wiley&Sons, New York, 2002, second edition
8. M. Muntean, *Introduction to OLAP Technology: Theory and Practice,* Editura ASE, Bucharest, 2004
9. Gh. Popa and others, *Microsoft SQL Sever,* Editura Economică, Bucharest*, 2006*
10. Appdev Products Company LLC, SQL Server 2000: OLAP Cubes and Queries Profesional Skills Development
11. R. Jacobson and others, *Microsoft SQL Server 2005, Analysis Services Step by Step*, Microsoft Press, 2006

Codreanu DianaElena is a graduate of ”Constantin Brâncoveanu” University of Pitești, Faculty of Management Marketing in Economic Affairs of Rm. Valcea, class 2004. At present, she is a Ph.D. candidate at the Academy of Economic Sciences in Bucharest, specialization in Economic Informatics.

Modeling Spatial Data within Object Relational-Databases

Iuliana BOTHA, Anda VELICANU, Adela BÂRA
Academy of Economic Studies, Bucharest, Romania
iuliana.botha@ie.ase.ro, anda.velicanu@ie.ase.ro, bara.adela@ie.ase.ro

Spatial data can refer to elements that help place a certain object in a certain area. These elements are latitude, longitude, points, geometric figures represented by points, etc. However, when translating these elements into data that can be stored in a computer, it all comes down to numbers. The interesting part that requires attention is how to memorize them in order to obtain fast and various spatial queries. This part is where the DBMS (Data Base Management System) that contains the database acts in. In this paper, we analyzed and compared two object-relational DBMS that work with spatial data: Oracle and PostgreSQL.
Keywords: *database, object-relational database, spatial data, GIS (Geographic Information System), spatial index*.

1 Introduction

Application that work with spatial data can be developed using various technologies (web services, desktop application, client-based application), architectures (SOA - Service Oriented Architecture, EDA - Event-Driven Architecture, cloud computing) or DBMS. Also the domains in which spatial database applications can have applicability are very vast: meteorology, secret services, army, agriculture (irrigations), geology, urbanism, utilities (energy, transport, telecommunications), health care, environment etc. An example of an application developed as a web service that uses spatial data as GPS coordinates is mentioned in the article [3].
Thus, in recent years are developed rapidly some new ways to store and manipulate spatial data. Since relational databases (RDB) have limitations in the case of spatial data, the most effective way proves to be the use of object-relational databases (ORDB) [4].

2. Some aspects regarding the object-relational databases

The existence of complex and comprehensive databases is an important requirement of the new type of economy, the digital one. Currently, the most widely used since the 80s are relational databases, characterized by simplicity, relatively easy implementation and easy data retrieval facilities through a powerful query language, SQL. In time, however, the complexity of the real world has led to failed attempts to represent it by simple models.
The relational databases limits led research in a new direction in programming that began to dominate - object-oriented technology, leading to a new generation of DBMS called object-oriented.
Developing object-oriented model was due to inability of the relational model to successfully deal with very large data volumes, of great complexity, encountered most often in new types of computer applications (multimedia, Internet, spatial applications etc.). However, although OODBMS (Object-Oriented DBMS) appear to meet the needs for better software required by the new economy, markets for their use remains relatively low, the reason most often cited being the difficult query with a large consumption of computational resources.
An intermediary level between relational and object-oriented databases comes through the new hybrid type, namely object-relational databases, which presents object-oriented features (in particular the

fundamental characteristics of objects: encapsulation, inheritance, polymorphism, etc.) as extensions of the relational model [8]. Thus, ORDB combines the benefits of both the relational and the object-oriented models such as scalability and support for complex data types (large objects, multimedia data, spatial data, user defined object types, etc.). Also called extended-relational, the object-relational data model is exemplified by the querying language version SQL-99.

In Fig. 1 is presented a very suggestive summary graph with a classification of the types of DBMS. This view was proposed in 1996 by Stonebraker [9], the ORDBMS being called "the next wave". However, the schema does not include the pre-relational models (hierarchical model and network model), considered obsolete at this stage of databases development.

The graph is simple and suggests how complex data and complex queries influence each other. It is a two-dimensional graphical representation: the abscissa refers to the ability to define complex data types and the ordinate presents the ability to query databases.

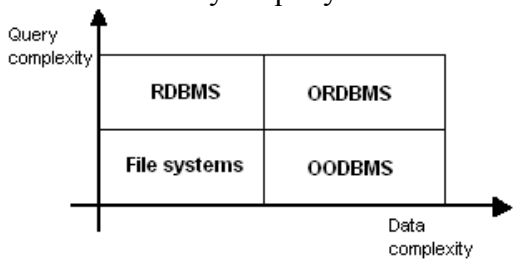


Fig.1. DBMS types classification - adapted from [9]

The bottom left quadrant contains the types of applications that process simple data types and requires no data query. These types of applications use directly the file system of the operating system for persistent data storage.

The top left quadrant refers to the relational database management systems (RDBMS), which processes simple data types, but allow complex queries.

Going forward, the lower right quadrant presents the object-oriented database management systems (OODBMS), which are handling with complex data types. However, resolving queries is quite difficult, since for each query must be provided the necessary relationships in the structure of objects.

In the upper right quadrant are included the object-relational database management systems (ORDBMS), which allow complex processing and resolving complex queries. Object-relational model is obviously the most complete, because allows both user-defined data types and complex queries. For this reason, the object-relational data model is also called extended-relational or universal and it can be seen as an attempt to capture as many as possible of the object-oriented concepts.

Using this hybrid type of database has its main reason for that:

- In many cases, the existing applications are already based on a relational data model. This calls for coexistence with the relational model as long as we do not want to redesign the applications based on a common object model to be included in a single OODB;
- Performance and scalability are important properties of an application, and in this respect, OODBMS have not yet shown advantages over RDBMS.

The current trends involve the transition from relational to the object-relational. Generally, this transition is achieved by gradually adding object model features to the relational one.

Regarding the ORDBMS, the offer is very generous and covers a wide scale of cost and performance, going from the DBMS that can be used free (unlicensed or with public license, such as PostgreSQL) to the commercial ones such as Oracle 10g, DB2 UDB 8, and SQL Server 2005. They all have adopted the object-relational data model to supports relational tables, abstract data types, and functions on abstract data types.

3. Some aspects regarding the spatial databases

Spatial objects consisting of lines, surfaces, volumes and objects of higher dimensions are commonly used in computer aided design applications, cartography or geographic information systems. These are described both by spatial attributes (length, pattern, perimeter, area, volume, etc.) and by non-spatial attributes (the time, the owner, membership of a superior structure, etc.). The values of objects' spatial attributes represent spatial data.

Spatial data can be divided into specific data and regional data. The specific data is a point, which is completely characterized by its location in a multidimensional space. It can be obtained directly from measurements or by converting for it to be more easily stored and retrieved.

The regional data is characterized by location (one fixed point in a region) and destination (a line in 2D or a surface in 3D).

Spatial data can also have a certain rank (e.g. "Cities in Europe"), approach (e.g. "Three nearby lakes in France"), junction (e.g. "Pairs of city, in Romania, located 50 km apart) [5].

The queries performed for spatial data can be: local, in the area or in the neighborhood (most difficult because it requires the evaluation of proximity - e.g. determining the road which passes closest to a specified region) [11].

Examples of spatial DBMS are:

- Boeing Spatial Query Server;
- PostgreSQL, which uses PostGIS spatial extension to implement standard "geometric" data and corresponding functions;
- MySQL, which implements "geometric" data type and some spatial functions. The functions that test spatial relations are limited to working with rectangles;
- Spatial Databox is a spatial front end to a relational system that offers spatial queries. It also offers real-time response to queries on interactive maps;
- Oracle Spatial (a feature of Oracle Database 11g Enterprise Edition) includes full support for Web and 3D services that allow managing all types of geospatial data, including raster and vector data, topologies and network models. It was designed to satisfy GIS requirements for applications such as territorial management, utilities, defense;
- SQL Server 2008 (spatial data stored in two types of data: GEOMETRY and GEOGRAPHY, for plane models and ellipsoid data).

The types of applications that work with spatial database systems are Geographic Information Systems (GIS), Computer Aided Design (CAD), Multimedia, etc.

A geographic information system is used to create, store, analyze and process information from the geometric space using a computer automated process. GIS technology can be used in various scientific fields such as resource management, studies on environmental impact, mapping, planning routes.

GIS has a unique way of organizing the database. There are two databases: one to store spatial distribution of geographic elements (through a system of x, y coordinates) and another to store the attributes of these elements (e.g. the length, the width, the number of bands, and the construction material of a road).

There are several ways to represent geographic data such as raster or vector. There is GIS software that works in either raster or vector mode, although most of them accept both types of formats.

4. Using spatial data for modeling object-relational databases

As shown above the object-relational databases are an extension of the relational ones. Thus, the logic of storing and retrieving data is the same like in the relational case. The main difference consists in new types of data, some of them user defined (like object classes), and in the ability of manipulate them. Spatial

data are important resources, which need to be manipulated, in order to use them in GIS applications.

However, one can observe that, in a table of such database, the main part of the columns have nothing in particular, being just standard columns. The exceptions (and the extensions) come from these columns that contain complex data: large objects, object types (with specific properties) or spatial data.

Therefore, regarding these features of the object-relational databases, several interesting metrics can be defined in order to use them for optimizations or just as statistics. Some examples of metrics are:

(1) – N_{OT}: the number of object types (known as object classes) involved in the definition of the columns from a user schema [6]:

$$N_{OT} = \sum_{i \in schema} ObjectType_i$$

(2) – P_{OT}: the percentage of inherited object types in the total number of object types (N_{OT}) that are involved in the definition of the columns from a user schema:

$$P_{OT} = \frac{\sum_{i \in Schema} InheritedObjectType_i}{N_{OT}} * 100$$

(3) – P_{CS}: the percentage of columns with spatial data type in the total number of columns from a user schema:

$$P_{CS} = \frac{\sum_{i \in Schema} SpatialTypeColumn_i}{\sum_{i \in Schema} Column_i} * 100$$

(4) – P_{CC}: the percentage of complex columns in a user schema [6]:

$$P_{CC} = \frac{\sum_{i \in Schema} ComplexColumn_i}{\sum_{i \in Schema} Column_i} * 100$$

(5) – P_{SC}: the percentage of spatial data in the total number of complex columns:

$$Psc = \frac{\sum_{i \in Schema} SpatialTypeColumn_i}{\sum_{i \in Schema} ComplexColumn_i} * 100$$

Using spatial data into object-relational databases is an important characteristic for the enterprises that develop or use GIS projects. Below, we provide an overview of spatial characteristics of Oracle and PostgreSQL, the ORDBMS mostly used.

Oracle and spatial data

Oracle DBMS has a feature that is called Oracle Spatial, which allows users to manage regional and geographic data in a native data type in the Oracle database. Thus, there can be developed a wide range of applications that may include: automated mapping, management facilities, geographic information systems or wireless location services.

Oracle Spatial provides facilities for storing, updating, querying spatial information from an Oracle database and it includes the following elements: MDSYS (Multi Dimensional SYStem) schema that governs the storage, syntax and semantics of data types supported by the geometric database; a mechanism for indexing spatial data by different ordering criteria; operators, functions and procedures for querying the areas of interest, the meeting space and other spatial analysis operations; functions and procedures specific for spatial data and adjustment operations; topographic data models for working with nodes, edges and faces in a topology; data networks models for representing the objects which are modeled as nodes and links in a network; GeoRaster, a feature that enables storage, indexing, querying, analysis and transferring the GeoRaster data.

Oracle Spatial uses a two-tier query model to resolve spatial queries and unions. The term is used to indicate that two distinct operations are used to perform queries. The result of combining the two operations is precisely the set of results. The two operations are: primary operation – meaning that the primary filter permits fast selection of records to the secondary filter; secondary operation – is based mostly on the operations performed by the secondary filter on the set of results from the primary filter and is usually more expensive.

Indexing spatial data is a mechanism to

decrease the number of searches, and a spatial index (considered logic) is used to locate objects in the same area of data (window query) or from different locations (spatial junction). Oracle Spatial uses two types of indexing: R-Tree (SDO_INDEX_TABLE table, maintaining the SDO_RTREE_SEQ_NAME sequence in the virtual table USER_SDO_INDEX_METADATA) and QuadTree (a tree structure, whose nodes have up to four children and is used to divide two-dimensional space, by recursively subdividing itself in four regions) [1].

To determine spatial relations, Oracle Spatial has several methods of secondary filtering:

- SDO_RELATE operator evaluates topological criteria;
- SDO_WITHIN_DISTANCE operator determines if two spatial objects are within a certain distance;
- SDO_INTERSECTION operator determines the topological intersection between two spatial elements;
- SDO_AREA operator calculates the area of a geometric figure;
- SDO_MAX_MBR_ORDINATE operator determines the maximum value for a coordinate (x or y);
- SDO_LENGTH operator calculates the length of a geometric figure;
- SDO_DIFFENECE operator determines the geometric element that results from the difference of two other spatial objects;
- SDO_CENTROID operator gets the centre of a polygon;
- SDO_NN operator identifies the nearest neighbor of a spatial object.

PostgreSQL and spatial data

PostgreSQL is an object-relational database developed by online community as an open-source alternative to the commercial databases like Oracle and Informix. PostgreSQL has native geometric types, built for academic research purposes, but they are too limited for using them in GIS projects and in spatial analysis.

The spatial extension to the PostgreSQL is called PostGIS and provides the ability to store relational attributes as well as spatial properties of the objects inside the database server. PostGIS supports the simple spatial features proposed by the Open GeoSpatial Consortium (OGC) - point, line, polygon, multipoint, multiline, multipolygon, and geometry collection [10]. In addition, PostGIS provides mechanisms for high speed spatial indexing using three types of access methods for indexes: B-Tree, R-Tree indexes and GiST (Generalized Search Tree).

In addition to normal B-Tree indexing, which is used only for data that can be sorted along one axis, R-Tree and GiST are used to speed up searches on all kinds of spatial data. R-Tree indexing algorithm breaks up data into smaller polygons, but according [7] the PostgreSQL R-Tree implementation is not as efficient and robust as the GiST implementation. GiST indexes decompose data according their spatial representation (e.g. "things to one side", "things which overlap", "things which are inside").

The SQL statements and functions provide the standard features for updating and retrieving spatial data from the database, for perform spatial operations (like area, length, union, intersection) and for determine spatial relations.

In terms of spatial analysis, one can make powerful queries such as location based queries (nearest neighbor) and sub-queries (searches based upon linear or spatial indexing of the subsets).

To determine spatial relations, PostGIS has a series of functions, of which we can specify:

- ST_INTERSECTION returns a geometry that represents the portion between two spatial elements;
- ST_AREA returns the area of a geometric figure;

• ST_LENGTH calculates the length of a geometric figure;

• ST_DIFFERENCE returns a geometry that represents a part of a spatial object that does not intersect with another object;

• ST_UNION returns a geometry that represents the common part of two spatial objects;

• ST_GEOMETRYTYPE returns the geometry type of an object.

Oracle vs. PostgreSQL

One can compare the two DBMS presented above considering their spatial characteristics and object-relational features.

Both Oracle and PostgreSQL with their spatial extensions, OracleSpatial and PostGIS, provides transactional spatial databases. Multiple users can access and edit the data stored simultaneously without file locking or data corruption issues. A main difference is that, unlike Oracle Spatial, PostGIS is an open-source project, which does not need licenses or restrictions for usage.

In Table 1 are presented some features that refer to Oracle and PostgreSQL historic, types of spatial data accepted in the database, spatial restrictions, spatial applicability, steps in developing a spatial application, object-relational functionalities.

Table 1. Oracle vs. PostgreSQL - synthetic comparison

	Oracle	**PostgreSQL**
Object-relational features	- type objects (similar with class objects) - each type has specific attributes and methods - there are implemented the following properties: type inheritance, polymorphism, encapsulation	- type objects (class objects) - each type has specific attributes and methods - there are implemented the following properties: type inheritance, polymorphism, encapsulation
Spatial features (**Oracle Spatial** vs. **PostGIS**)	*Oracle and PostgreSQL spatial history*	
	- Oracle 4 had spatial-data capability - Oracle 7 had "Spatial Data Option" or "SDO". - Since Oracle 8, "Oracle Spatial" extension exists. The primary spatial indexing system uses a standard R-tree index.	- PostgreSQL has its own native geometric data type, but this is unable for being used in GIS projects - spatial features since PostgreSQL 7.1.x database server, when it was released PostGIS 0.1
	Steps in building a spatial application in Oracle and PostgreSQL	
	1. Create the table in which the spatial data will be stored. 2. Adding the appropriate entries in the table. SDO_GEOMETRY is the spatial field. 3.Updating the USER_SDO_GEOM_METADATA view to reflect the dimensional information for spatial data. 4. Creating the spatial index (an R-	1. Create the table in which the spatial data will be stored. 2. Adding the appropriate entries in the table. GEOGRAPHY is the spatial field. 3. Updating the SDE_LAYERS, SDE_TABLE_REGISTRY, SDE_GEOMETRY_COLUMNS, SDE_COLUMN_REGISTRY tables to reflect the dimensional information

	Oracle	**PostgreSQL**
	tree index). 5. Executing spatial queries.	for spatial data. 4. Creating the spatial indexes (R-Tree or GiST). 5. Executing spatial queries.
	Operating systems on which Oracle and PostgreSQL can run	
	- Windows - Linux - Unix - Solaris - Mac OS	- Windows - Linux - Unix - Solaris - Mac OS
	Oracle and PostgreSQL spatial data type	
	SDO_GEOMETRY	GEOGRAPHY

In Oracle the spatial field has some parameters: SDO_GEOMETRY (polygon_dimension, latitude_longitude, SDO_POINT_TYPE, SDO_ELEM_INFO_ARRAY (SDO_STARTING_OFFSET, SDO_ETYPE, SDO_INTERPRETATION), SDO_ORDINATE_ARRAYV (variable number of parameters)). SDO_STARTING_OFFSET represents the offset from which to begin storing in the SDO_ORDINATE vector and starts at 1, not 0.

The step 3 in developing an Oracle Spatial application is required to be performed before the index is created. It is performed once for each level (table-column combination). It contains information about the name of the table that contains spatial data, the column from the table that is SDO_GEOMETRY type, the size of the geometry and a number (SID) which specifies the value of the coordinates system [12].

Spatial indexing is a mechanism that helps to increase the search performance on a table based on spatial criteria. An R-tree index approximates each geometry with the smallest rectangle that contains the geometry (called MBR - Minimum Bounding Rectangle). When there are more geometries, an R-tree index consists of a hierarchical indexing of MBR rectangles [2].

5. Conclusions

The integration of the spatial data into object-relational databases is an absolutely necessary characteristic for today's enterprises that uses a GIS. The paper provide a synthetic look of spatial features of two ORDBMS, namely PostgreSQL and Oracle. These two products cover the sector of open source and commercial object-relational technology today mostly used.

6. Acknowledgments

This paper is a result of the project POSDRU/6/1.5/S/11 „Doctoral Program and PhD Students in the education research and innovation triangle". This project is co funded by European Social Fund through The Sectorial Operational Program for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies.

This paper presents some results of the research project PN II, TE Program, Code 332: "Solutii informatice pentru asistarea procesului decizional in mediile incerte si cu evolutii putin predictibile in vederea integrarii in retele de tip Grid", financed within the framework of People research program.

References

[1] D. Abugov, N. Alexander, *Oracle Spatial User's Guide and Reference, 10g Release 1 (10.1)*, Oracle Corporation, 2003.

[2] T. Brinkhoff, H.P. Kriegel, B. Seeger, „Efficient Processing of Spatial Joins Using R-trees", International Conference on Management of Data, *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, Volume 22, Issue 2, pp. 237 - 246, 1993, Washington D.C., ISSN 0163-5808.

[3] A. Dioşteanu, L. Cotfas, "Agent Based Knowledge Based Management Solution using Ontology, Semantic Web Services and GIS", *Informatica Economica Journal*, Vol. 13, No. 4 / 2009, pp. 90-98, ISSN 1453-1305.

[4] W. Huibing, "Extending object-relational database to support spatio-temporal data", *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. XXXVII. Part B2. Beijing 2008, ISSN 1682-1750.

[5] J. LeSage, S. Banerjee, M. M. Fischer, P. Congdon, „Spatial statistics: Methods, models & computation", *Computational Statistics & Data Analysis*, Volume 53, Issue 8, 15 June 2009, pp. 2781-2785, ISSN: 0167-9473.

[6] M. Piattini, C. Calero, H. Sahraoui, "An empirical study with object-relational databases metrics", *Proceedings of the 7th International Conference on Software Quality*, London, 2002, pp. 298–309, ISBN 3-540-43749-5.

[7] PostGIS 1.5.1 Manual - http://postgis.refractions.net/docs/index.html

[8] Gh. Sabau, "Comparison of RDBMS, OODBMS and ORDBMS", *Proceedings of the 8th International Conference on Informatics in Economy,* Bucharest, 2007, pp. 792-796, ISBN 978-973-594-921-1.

[9] M. Stonebraker, D. Moore, *Object-Relational DBMS - The Next Great Wave*, Morgan-Kaufmann, 1996, ISBN:155-860-397-2.

[10] http://www.refractions.net/products/postgis

[11] http://www.spatial.cs.umn.edu/

[12] A. Yeung, B. Hall, *Spatial Database Systems: Design, Implementation and Project Management*, Springer, 2006, ISBN 1-4020-5393-2.

[13]

Iuliana BOTHA is an Assistant Lecturer at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2006 and the Databases for Business Support master program organized by the Academy of Economic Studies of Bucharest in 2008. Currently, she is a PhD student in the field of Economic Informatics at the Academy of Economic Studies. She is co-author of two books, 5 published articles (one article ISI indexed and another three included in international databases), 9 scientific papers published in conferences proceedings (among which 3 paper ISI indexed). She participated as team member in 3 research projects that have been financed from national research programs. From 2007, she is the scientific secretary of the master program *Databases for Business Support* and she is also a member of INFOREC professional association. Her scientific fields of interest include: Databases, Database Management Systems, Design of Economic Information Systems, Grid Computing, e-Learning Technologies.

Anda Velicanu is a Pre-Assistant Lecturer at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Economic Cybernetics, Statistics and Informatics of the Bucharest Academy of Economic Studies, in 2008. She is a PhD student in the field of Economic Informatics at the Academy of Economic Studies and since January 2009, she is a Pre-Assistant Lecturer. She teaches Database, Database Management Systems and Economic Informatics seminars at the following faculties: Economic Cybernetics, Statistics and Informatics, Commerce, Marketing and International Business and Economics. Her research activity can be observed in the following achievements: 5 diplomas, 2 scientific awards, 3 proceedings, 2 articles published in scientific reviews, 1 research contract, 1 book and 1 research grant. She is a member of INFOREC professional association. Her scientific fields of interest include: Databases, Database Management Systems, Programming, Information Systems.

Adela BÂRA is a Lecturer at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Economic Cybernetics in 2002, holds a PhD diploma in Economics from 2007. She is the author of 7 books in the domain of economic informatics, over 40 published scientific papers and articles (among which over 20 articles are indexed in international databases, ISI proceedings, SCOPUS and 2 of them are ISI indexed). She participated (as director or as team member) in 4 research projects that have been financed from national research programs. She is a member of INFOREC professional association. From May 2009, she is the director of the Oracle Excellence Centre in the university, responsible for the implementation of the Oracle Academy Initiative program. Domains of competence: Database systems, Data warehouses, OLAP and Business Intelligence, Executive Information Systems, Decision Support Systems, Data Mining.

Agile Development for Service Oriented Business Intelligence Solutions

Marinela MIRCEA, Anca Ioana ANDREESCU
Economic Informatics Department, Academy of Economic Studies, Bucharest, Romania
mmircea@ase.ro, anca.andreescu@ase.ro

Considering the evolution of information and communications technology, the necessity of alignment of public and private sectors to European Union requirements, the current economic crisis, and the global context, all organizations are trying to achieve major changes that would enable them to operate as intelligent organizations. For this purpose, agility and Business Intelligence are seen by most managers as a way to transform their organizations into intelligent organizations. The study highlights the importance of modern approaches (Service Oriented Architecture, Business Process Management, Business Rules, Cloud Computing, Master Data Management) in developing agile Business Intelligence solutions. The paper also presents the stages of developing an agile Business Intelligence solution in the case of public procurement.
Keywords: *Business Intelligence, agile development, service oriented architecture, business process management, business rules, public procurement*

1 Introduction

Given the characteristics of the knowledge society, the need to align legislation with the European Union (EU) requirements and the level of private sector development, the Romanian public sector made permanent change in its organizational structure, administrative practices and management systems. One current concern of public institutions is the creation of intelligent institutions through the efficient spending of public funds. This involves the necessity of monitoring the progress of institutions and the way they adapt to the changes in legislation as well as the necessity of creating an environment in which performance is properly evaluated.

In recent years, the Romanian public sector has moved from the traditional paradigm to models that meet the demands of knowledge-based economy, such as: flexibility, globalization, horizontal/vertical integration, innovative enterprises, organizational learning, customer-led strategy, electronic procurement. The new paradigms make the shift to the electronic procurement system, by creating a global collaborative and competitive network.

The present study highlights the importance of developing an agile solution for Business Intelligence (BI) and shows the stages of the solution in the case study of public procurement. The new trends in information technology and communications as well as those in the field of public procurement (legal principles, principles of quality management) have been considered in relation to achieving an agile BI solution.

The research methodology consisted in rigorous analysis of recent trends in the areas of interest, in practical documentation and in the authors' expertise in the areas of information technology (IT) and public procurement. This paper is continuing research in these fields, based on the (theoretical and practical) results, and is continuing with the steps of development of agile BI solutions.

2 Modern approaches in the development of agile Business Intelligence solutions

Developing a BI solution is an activity that involves many challenges, being constrained by the reality of information. Developers have to understand the business requirements, the format and the weaknesses of data sources, the existing systems, the various needs of the users etc. The development of the BI system has the purpose of ensuring comprehension of the factors affecting performance metrics and of providing managers with expressive representations of the information that shapes the business.

Creating a BI environment involves building an analytical data warehouse for managers. In many institutions, the most important decision metrics are calculated on the basis of information collected from various systems. For this reason, ❶ Business Process Modeling (BMP) is an important technique for gathering this information and, along with the data warehouse, is a method of integrating different sources of information.

Moreover, a data-centered approach on BI represents just a part of the picture of business. A process-oriented BI solution gives a complete picture of the business [1], providing information on data from the business process operations and the IT infrastructure, historical analyses and metrics on the history of business processes, the business plans, forecasts and budgets, data from external events in the form of key performance indicators, alerts, reports, and recommendations for corrective action.

The analysis of processes and ❷ Business Rules (BR) is necessary to further analyses for the creation of the BI solution, as the BR helps defining the dimensions and metrics. One of the key factors for the success of the development is the use of BPM and BPM/BR analysis for improving the database schema of the data warehouse. The combination between business rules and web services offers an adequate approach for applications integration and sharing of distributed information (details about main components of an e-Procurement system are presented in [2]).

❸ Service-oriented architecture (SOA) can provide numerous benefits, such as: promoting reuse, the ability to combine services to create new composite applications, use of decoupled services with a standard interface, while providing at the same time a technological method for the development of Business Intelligence solutions [3]. Implementation of ❹ Master Data Management (MDM) into SOA strategy [4] ensures data consistency, alignment of the organization's information resources, correct dissemination of information inside/outside the organization, and delivery of all the potential benefits of SOA initiatives.

Business rules adoption, together with a service-oriented architecture, allows the integration of strategic corporate applications between multiple business units. For example, the same business logic that has been explicitly defined in a Business Rules Management System (BRMS) may be shared in a Service-Oriented Architecture with other applications that need it. These applications communicate via XML with the Business Rules Services [5].

To accelerate the adoption of BI and BPM technologies, which involve relatively high costs for institutions, ❺ Cloud Computing may be used, considered to be a cheaper solution for providing intelligence and business process management ([6], [7]). Cloud computing is considered the next step in Internet evolution, providing for organizations a way to use IT services, against payment, from infrastructure and computing power, applications and business processes, to customized collaboration [8]. Given the complexity of these platforms, the agility of these solutions is difficult to test and validate.

The importance and utility of the audit of Business Intelligence solutions is measured in relation to effects on the

quality of economic activity and processes within organizations. The audit process helps us discover relatively quickly the weaknesses and the parameterisation problems of BI solutions in relation to the specific activity that is subject to implementation [9]. Thus, we can find answers to questions concerning the Business Solution solution's response times to data or information queries, the quality of data and information, how data are extracted from the system, the view mode, the user ring structure and content tree structure, the structure of data cubes and aggregation-disaggregation or synthesizing-desynthesizing queries within processes of ensuring information compatibility, etc.

Developing a BI system involves going through several steps [10], starting by identification of decision makers, of issues, entities and events that are necessary to make the decisions. Then follow the identification and analysis of business rules, the development of prototypes of BI dashboards and the underlining the cubes. In the final step of BI development process, the result will be certified. Metrics must be confirmed by management data and images. Finally, the BI system is released in order to be used by the managers.

3 Steps to develop an agile Business Intelligence solution

The further steps take into account the modern development solutions mentioned above. In addition, each step of development is exemplified in the case of procurement process within Romanian institutions.

3.1. Identifying decision-makers and defining performance metrics

In creating the department's performance evaluation system it should be taken into account the performance aspects of acquisitions, which are of interest to institutions' managers and key stakeholders, and the way in which they could be measured. The process of creating the system of indicators is built on three stages [11].

a) Identifying the overall objectives. The BI system should provide a representative picture of the objectives and of the indicators that justify them. For a purchasing function to be efficient and effective it has to ensure the three main objectives (savings, quality, convenience). One of the principles underlying the granting of a public procurement contract is the "efficient use of funds." Thus, the main function that a procurement service should provide is to ensure that funds are spent wisely, achieving savings in purchasing the goods and services required to meet the needs of the institution.

b*) Identifying key procurement processes necessary to achieve overall objectives*. To achieve the overall objectives of the department, it is necessary to ensure: compliance with best procurement practices; customer relationship management; supplier relationship management; procurement management.

c) *Identifying the necessary organizational resources*. The following elements must be present within the institution in order for the key processes to be developed: skilled human resources, resource allocation policy, and policies to stimulate employees.

The system of indicators has the purpose of encouraging progress and best practices in areas like ❶ efficiency, ❷ collaboration, ❸ compliance with the law, ❹ training, and ❺ electronic procurement. Taking into account the key performance issues identified in the three stages, the study [12] puts forward a system of indicators for the public procurement process.

3.2. Identifying the necessary information

BI for public procurement provides support to department decisions through the assessment indicators established at the

previous stage. Managers will use these metrics to evaluate the procurement performance level in the institution. Information gathered from different systems that contribute to the business model is needed to calculate these metrics. Each of the systems requires one or more processes. ETL (Extract, Transform and Load) processes will collect data for the business model.

The information necessary to calculate the performance metrics is collected mostly from of the operational source systems and, to a smaller extent, from additional source systems or manual processes. The institution has a medium level of informatization and there are several applications that administer data needed by the procurement process. The source systems that provide the most information and underlay the procurement process are: the relationship with customers, with suppliers, the accounting and the procurement system [13].

At each stage of the procurement process, users create data that can potentially be used for BI. Also, to complete the stage, users need BI. For example, to validate a request from a client, the demand must be within the limits of the available amount of money, and if it is not, the customer is offered alternative products or sources of funding. BI for procurement supports the procurement decisions through evaluation of the three metrics associated with the overall objectives of the department and through analyses that affect these metrics. Bellow is exemplified the diagram of BI concepts for calculating the savings indicator (figure 1).

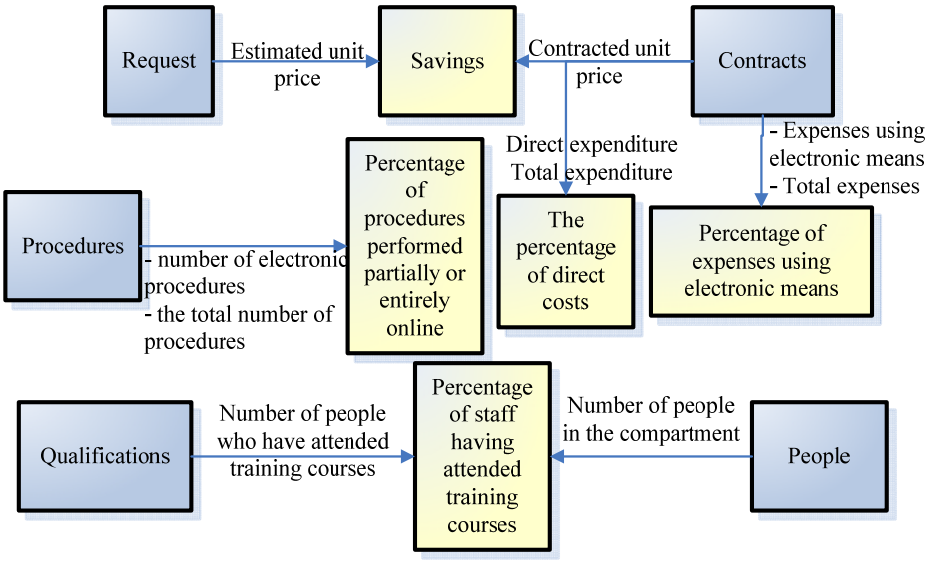


Fig. 1. Entities diagram for the calculations of savings indicator

3.3. Definition of business rules

BI experts use the term "business rule" in a variety of meanings and contexts. The definitions of this term can be focused exclusively on the business perspective or on the IT perspective. Ronald G Ross provides a description of the business rule that encompasses both sides. From a business perspective, the "business rules are literally the encoded knowledge of your business practices", while from the IT perspective, "the business rule is an atomic piece of reusable business logic."[1]

From the IT perspective, business rules are coded in certain cases in data warehouse ETL processes or in BI instruments, during the design of the specific reports. These are not the optimal choices for the integration of business rules. An optimal method for

[1] Ronald G. Ross. *Principles of the Business Rule Approach.* Boston: Addison-Wesley, 2003.

achieving business logic is independent description of rules in a separate module. This software component is dedicated only to the implementation of business logic and has four major advantages [14]:

- it is well designed and the business logic module can be transparent for business users;
- allows adapting business rules to the frequent changes;
- reduces duplication (if the IT department decides to replace an ETL or BI instrument, the implementation of rules does not change);
- allows interfunctionality, widely use of IT and business rules management.

Combining business rules and Web services provides an appropriate approach for integrating the application and sharing the distributed information. The adoption of business rules, along with implementing a service-oriented architecture allows the integration of institution's strategic applications between multiple business units. For example, the same business logic, which was explicitly defined by BRMS may be shared within SOA with other applications that need it. These applications communicate via XML with the Business Rules Service [5].

In this step: ❶ the rules affecting the modeled metrics should be identified, ❷ the rule for the limit values and their incorporation as dimensions within the warehouse data should be analyzed, ❸ numerical calculations should be extracted and added as facts in the data warehouse. Building a BI system according to these considerations will help managers to assess the effects of changes in the limit values and calculations. As all rules are centralized, it will be easy to find a rule and to use it in evaluating decisions.

Each defined business rule should classify, calculate, compare and control [10]. Thus, the business rule: • should classify the type, the division or the range (for example, suppliers may be classified as: approved, indifferent or unauthorized; the types of requests may be grouped into: approved, rejected, pending or completed); • should calculate formulas, should query data and statistics, transform and associate values (there are often numeric constraints - for example, applications must be within a client's maximum budget available, a conversion rule converts input values into useful data); • should compare the result with the limit values (the limit value is the key-value that must be met or must not be exceeded, or that is within a certain range), • should control what is true or valid, right or wrong and the associated messages (business rule example in [12]).

3.4. Defining business processes

Inclusion of time and performance into the data warehouse imposes the need to identify the processes associated with procurement, to find the time records for the steps and to incorporate changes in the time and date into the data warehouse. In order to define the business process we will use BPMN with SOA, which will provide many benefits. Separating the business in a number of central and discrete processes, an institution may provide a certain service to its customers or it may outsource the service. Control of business processes is done using business rules.

The BPM design software coordinates the details and activities, while execution environment invokes them at the right time. Because the BPM software coordinates the activity, the approach on business processes simplifies the creation of applications. The programming team writes the code which carries out the activities and the BPM software coordinates these activities. The flow controls connect the design details and become process activities. Then follow the identification of data or of business entities for the business processes. For important interfaces and business processes, business rules will be added to the diagram. Within a process, the rules may define policies, constraints or competitive business practices. The final step of the business

process is to define exceptions and points of failure. The result of the business process design is a prototype of the work process which, corroborated with interfaces, and confirms business entities.

Developing the main business processes

The central business processes, which are necessary to procurement activities, are the following: the issuing of the request, the approval of the request, the acquisition, and the completion of the request. These processes are independent from the systems that develop them; they execute an activity within the business architecture. Processes use the application as a service that moves transactions or adds a procedure in a system queue. Main business processes operate through data processing. BPM substantiate traditional business data with descriptive process information, mainly information related to activity time. Addressing the main business processes is an important modeling technique, essential in the construction of SOA. Monitoring of procurement procedures will be simplified. Thus, by using a process-oriented approach, the executive board will monitor high-level business processes, while managers will measure and control the subprocesses [13].

The process of issuing a request

For the purpose of customer relationship management, more specifically for the management of client requests, it may be used a combination between a web model and an internet shopping card (figure 2). Customers can view listed all the products proposed by the system and can add a product in their shopping cart (figure 3). Searching is done by locating a particular product or set of products based on several criteria: keyword, category (subcategory) of products, accounting record, storage identification number, etc. Customers are offered the facility to select multiple products and compare them based on a common set of criteria.

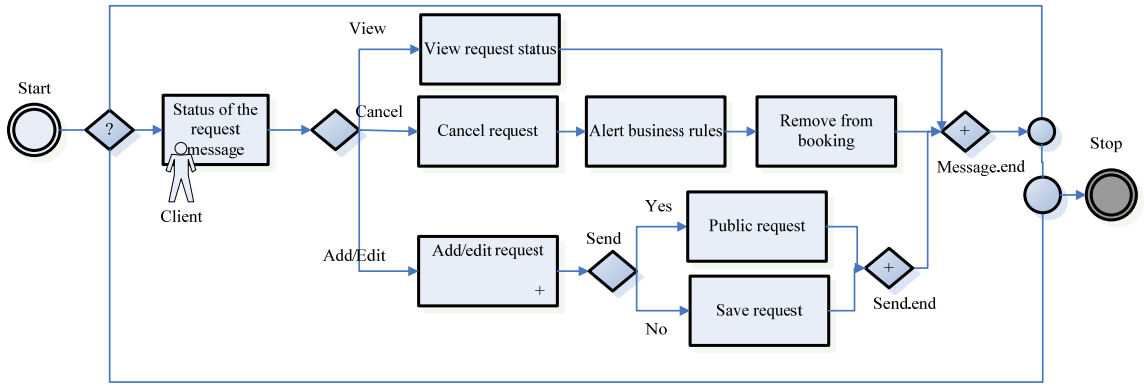


Fig. 2. Business process diagram for issuing the request

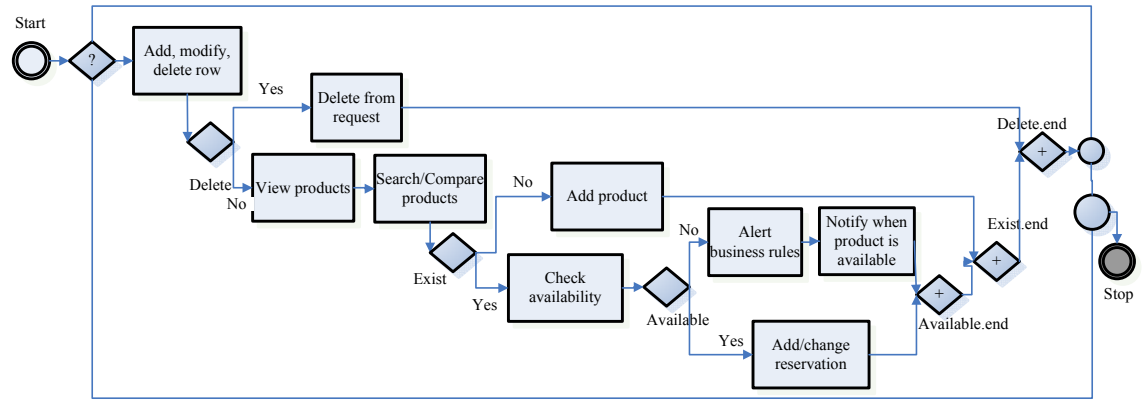


Fig. 3. Business process diagram describing the "Add/edit request" subprocess

Customers can add, within the request, a product that is not present in the list. In this case, the minimum information necessary in order to add the product is: the name of the product, the estimated unit price in lei, VAT excluded, and the minimum specifications required. For products that are not readily available to customers, the system will ask if they want to be notified via e-mail when the product becomes available. For each request shall be specified the necessary acquisition date and the place of delivery.

The system will allow customers to view the content of purchase requests and to add, change or delete a product from this content. The requests recorded and accepted by the system may be accessed and viewed for later use. At any time, the client can view the status of the request within the system. A customer who wishes to delete a request can do so only if the business rules acceptability conditions are met.

The process of approving a request

The system automatically determines the level of approval of a request, depending on the total value and the source of funding, according to a business rule, and notifies the approving manager on the existence of the request (figure 4). The manager is given the opportunity to view, select and examine the details of a request, in view of approval, modification or rejection. A request will be marked as a purchase request only after all the business rules are applied. The system automatically makes a log of requests and offers the option for issuing reports, according to selection criteria.

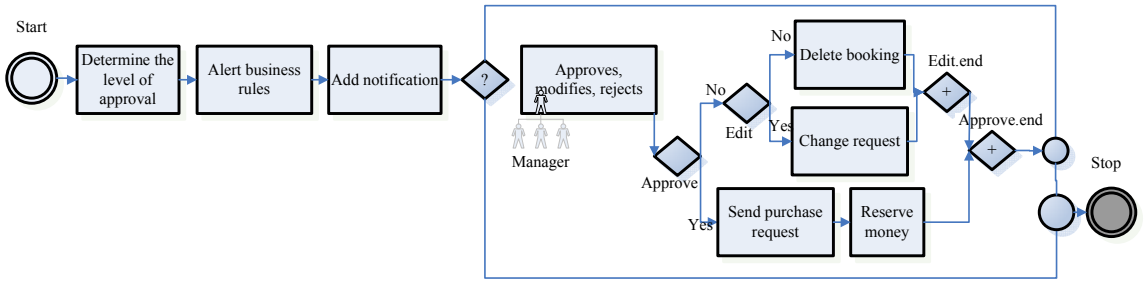


Fig. 4. The business process diagram describing the approval process

Processing the request

When a purchase request enters the system, it will be associated with a purchaser, depending on the type of contract, the category of products, its load factor, according to a business rule (figure 5). The purchaser may classify the products from the request (if they are available), may associate the request to an existing procurement procedure (for centralizing the information) or start a new procedure, according to business rules.

Starting a procurement procedure requires issuing all the necessary documents and their marking as approved only after the proposed sums are recorded and validated in the accounts associated to the requests (figure 6). Then follows the creation of tender documentation based on specifications and data from the associated requests, according to forms required by the law and by their transmission over the HTTP protocol. The acquisition system will submit for publication in SEAP (Electronic System for Public Procurement) the procurement documents needed for publication.

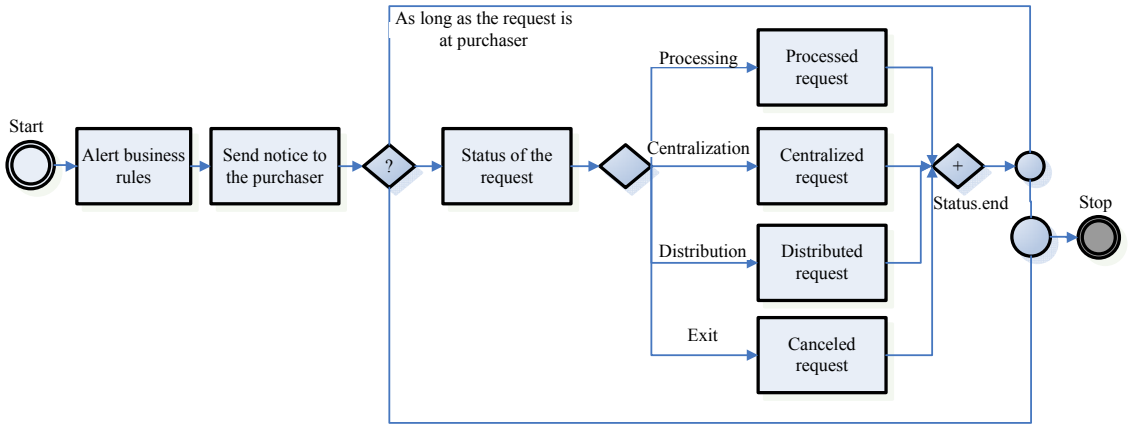


Fig. 5. The business process diagram describing the procurement process

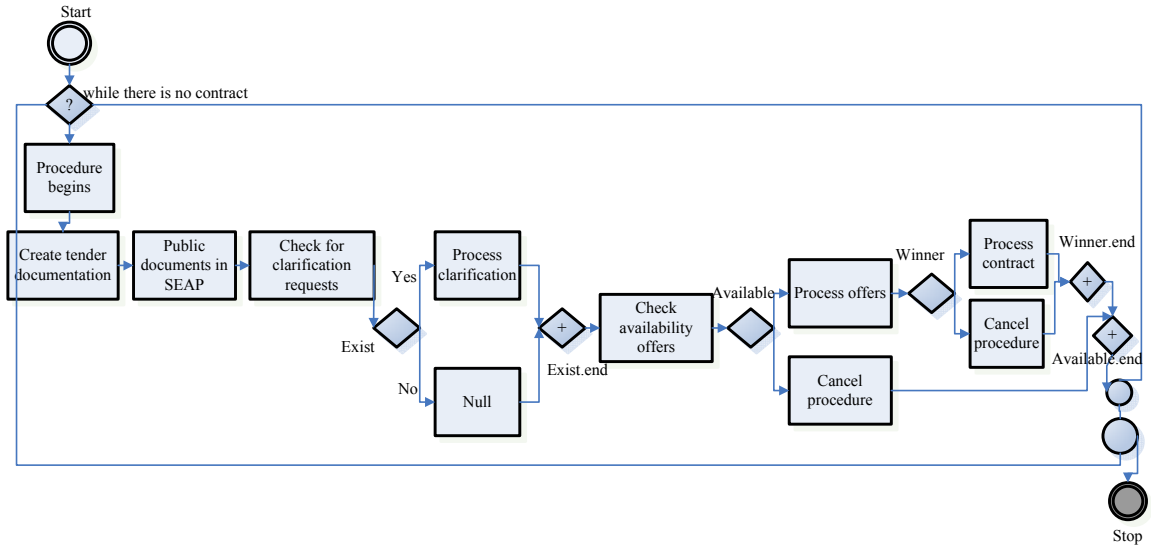


Fig. 6. The business process diagram, describes the "request under processing" subprocess

Information on the offers will be taken from SEAP, in the case of online procedures, or manually entered by the purchaser, in the case of an offline procedure or of a combination of offline and online procedures. The processing of offers follows next, involving the issuing of the evaluation documents according to forms required by the laws and to certain business rules, the announcing of the outcome of the procedure and the solution to possible disputes.

The procedure can be finalized by cancellation or the closing of one or more contracts and their registration into the system. On the basis of the contract and business rules, the contracting authority shall issue the firm orders to the supplier. The system automatically creates a log of procedures and allows the creation of reports listing all acquisition procedures, according to the selection criteria.

The process of completing the request

After inspecting the contracting authority, the accepted products are automatically added to the inventory management system through a unique identifier assigned to each product via a barcode reader (figure 7). If defective products are identified, they are returned to the supplier for correction or replacement. Products payment will be made by transferring the money from the clients' accounts to the account of the provider. The central deposit or the distribution centre delivers the actual products to the customer site and receives a delivery confirmation. The reception of the product by the customer, at the central deposit or at the distribution centre, represents the completion of the

acquisition process. The system creates the log of reception and enables the creation of reports listing all suppliers / products received from suppliers, according to certain criteria.

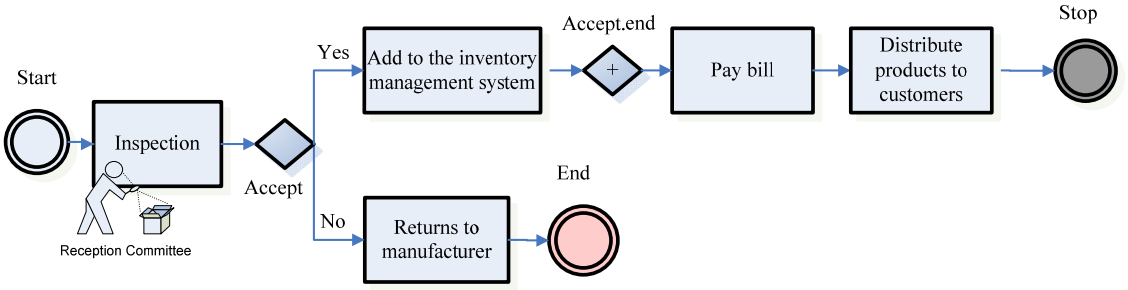


Fig. 7. The business process diagram for the process of completing the request

Flow control elements

BMNP instruments create a single document for each of the processes. BPMN coordinates services and applications through the link with the Web services, concentrating all logic on constructing a composite application that would be easily modifiable.

Data elements

In order to achieve a concise analysis, the data concept diagrams for the processes of issuing (figure 8), approval (figure 9), acquisition (figure 10) and completion of a request (figure 11) will not contain the attributes of entities. The concept describes the data design carried out during a BPM design process.

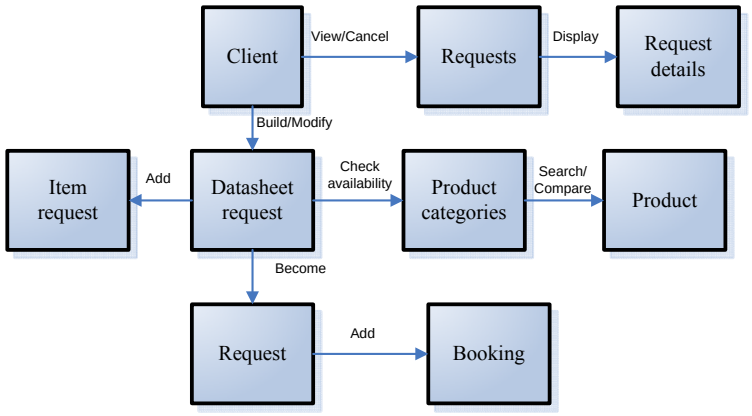


Fig. 8. The process of creating a request uses entities from different systems in order to manage customer requests

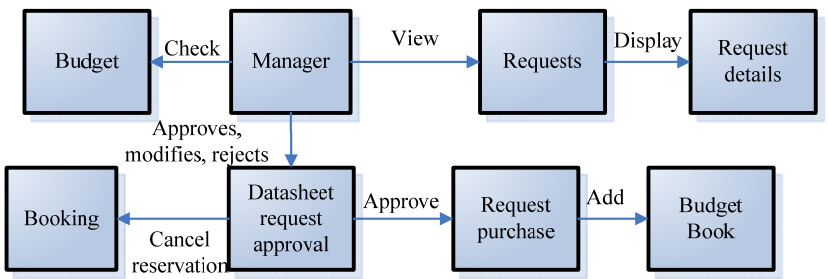


Fig. 9. The process of approving a request uses entities from different systems in order to manage the approved requests

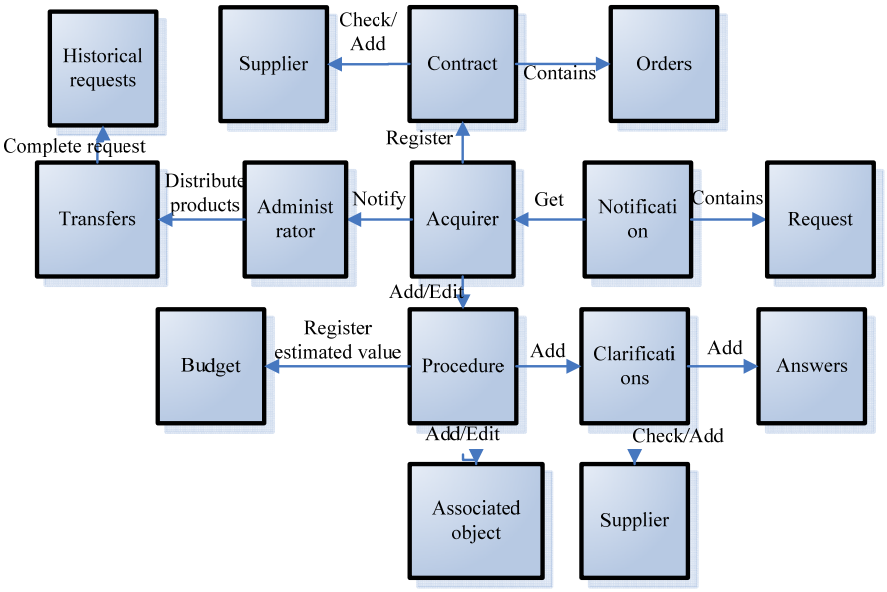


Fig. 10. The process of acquisition uses entities from different systems in order to manage the acquisition procedures

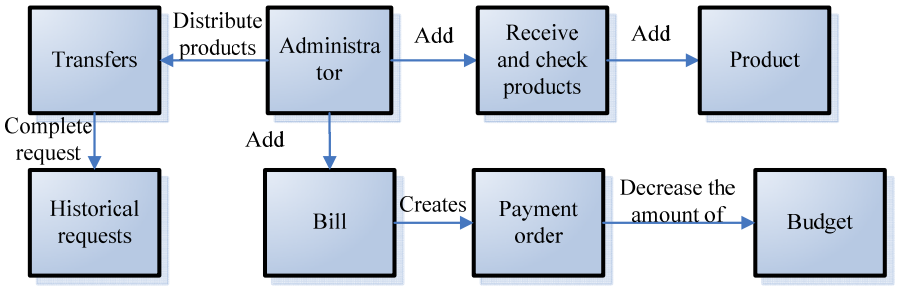


Fig. 11. The process of completing a request uses entities from different systems in order to manage the products

Establishing business rules of business processes

BPMN uses web services to invoke subprocesses, business rules and to interface with external systems. In the business processes only the data structure is needed to invoke a web service. An XML WSDL specifies the web service.

The following business rules are presented for the process of issuing a request:

- a business rule deciding whether the client will make an e-mail or phone notification when the requested object is close to the address of destination;
- a business rule deciding whether a request sent by the client may or may not be revoked / changed.

Procurement solution interfaces.

After detailing the central business processes, the project team has to design the interfaces. BPM / BR must ensure the link between current networks, business systems and existing applications into a single IT system – a composite application. Once the needs of the interface are established, the business forms or data entities have to be mapped according to the needs of the external systems. This task involves creating a series of process activities mapping the attributes of business forms according to systems interfaces.

Connecting the organization with external suppliers involves the use of open protocols, such as business protocols EDI (Electronic Data Interchanges) or B2B (Business to Business). At the time of execution the business process will register all the steps, from the first reception of a message until the final transmission of the message to/from the supplier.

Programming processes. BPM instruments provide integrated control and programming solutions for all business processes. In general, BPM solutions provide a control instrument that centralizes all businesses processes and the requirements of data processing within a simple application that is consistent and easy to manage.

Exceptions. To complete the business process the diagram must be detailed to include exceptions and the appropriate message structures. The system must be able to handle exceptions and allow transactions to return to original state.

3.5. Business Intelligence Dashboard

Design of the executive dashboard is made by controls based on measurements or dimensions. Using different combinations of visualisations and controls, the BI team must create a dashboard providing a view on the features of the business monitoring environment. Taking into account the need to underline indicators reflecting the different levels of aggregation, we will present different examples, using the dimension-based development. When the historical value is above or below the basic value, managers need to discover the performance characteristics. At this stage the “slice and dice” technique is used to break the cubes. Managers select the dimensions within the dashboard, and then choose the metrics to be calculated by the dashboard.

In designing the scheme of the data warehouse, the following types of schemes may be used: the "star" scheme, the "snowflake" scheme and the "constellation of facts." Given the fact that the "star" scheme provides a direct and intuitive link between the business entities and the performance of queries, we will use this type of scheme.

Based on the metrics identified in step one, we provide the example of the savings cube in the area of public procurement. All public institutions must measure the value of savings over a period of time in relation to the procurement procedure, the purchaser and the institution. This information is useful to managers in making timely decisions to reinvest or in the policy of providing incentives for purchasers. In order to calculate the *savings* indicator the following dimensions are taken into account: the time, organized by trimesters; the procedure, according to the similar products; the purchaser, the person responsible with the procedure (figure 12). Facts from the savings data warehouse include the estimated unit price, the contracted unit price and the quantity.

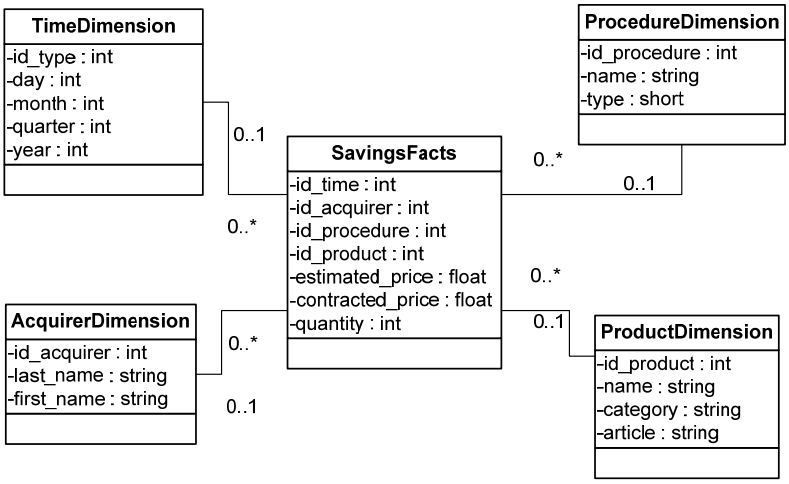


Fig. 12. Dimensional model for savings

3.6. Identifying ETL processes, testing and publishing

On the basis of the identified dimensions and measures, the ETL programs / software have to populate the data warehouse with information from the identified source entities. ETL is best done using business process techniques, because most ETL programs automatically decide when to insert or modify records. ETL processes for data collection may be: • iterative extractions of operational data, loading of a file, • data query by JDBC or *middleware*; • messages from a different business process.

4 Conclusions

Organizations are encouraged to become service-oriented and integrate agile and intelligent solutions that lead to increased efficiency and innovation. Integration of intelligent solutions depends on the outcome analysis of costs and benefits of Business Intelligence solutions and resources available for implementation. Also, the possibility of using a traditional Business Intelligence versus Cloud Computing solutions need to be considered.

5 Acknowledgment

This work was supported by CNCSIS-UEFISCSU, project PN II-RU (PD), "Modern Approaches in Business Intelligence Systems Development for Services Oriented Organizations Management", code 654/2010, contract no. 12/03.08.2010.

References

[1] M. Mircea, B. Ghilic-Micu and M. Stoica, "Combining Knowledge, Process and Business Intelligence to Delivering Agility in Collaborative Environment". L. Fischer, ed. 2010. *2010 BPM and Workflow Handbook, Spotlight on Business Intelligence*. Florida*:* Future Strategies Inc. & Workflow Management Coalition, pp.99-114.

[2] M. Stoica, B. Ghilic-Micu and M. Mircea, „Electronic biddings for public acquisitions of goods, services and work: Romanian approach". In: *The 9th WSEAS Int. Conf. On Mathematics and Computers in Business \and Economics (MCBE'08)*, WSEAS Press, Bucharest, Romania, ppg. 62-67, 2008;

[3] M. Mircea and A. I. Andreescu, "Agile Systems Development for the Management of Service Oriented Organizations". In: *11th International Conference on Computer Systems and Technologies, CompSysTech'10*, Sofia, Bulgaria, 17-18 June 2010, pp. 341-346;

[4] A. Andreescu and M. Mircea, "Actual Trends in Software Systems for Business Management", *CompSysTech'08, The Bulgarian*

Academic Society of Computer Systems and Information Technologies, 2008;

[5] G. Holden, "Reactive and Proactive Business Intelligence", 2007, http://www.b-eye-network.co.uk/view-articles/5899

[6] M. Mircea, B. Ghilic-Micu and M. Stoica, "Combining Business Intelligence with Cloud Computing to Delivery Agility in Actual Economy", *Journal of Economic Computation and Economic Cybernetics Studies*, 45 (1), 2011, pp. 39-54;

[7] M. Mircea and A. I. Andreescu, „Extending SOA to Cloud Computing in Higher Education". In: K. S. Soliman (ed.), *The 15th IBIMA conference on Knowledge Management and Innovation: A Business Competitive Edge Perspective*, Cairo, Egypt 6-7 November 2010, pp. 602-615,;

[8] M. Cunningham, "Cloud Computing Enables Self-serve BI", 2010, http://www.dashboardinsight.com/articles/business-performance-management/cloud-computing-enables-self-serve-bi.aspx

[9] B. Ghilic-Micu, M. Mircea and M. Stoica, "The Audit of Business Intelligence Solutions", *Informatica Economica*, 14 (1), 2010, pp. 66-77;

[10] T. Debevoise, *Business Process Management, witch a business rules approach, Implementing the Service Oriented Architecture*, United State of America, Tipping Point Solutions, 2007;

[11] B. Ghilic-Micu, M. Stoica and M. Mircea, Management of Acquisitions in Romanian Public Institutions. In: *Calitatea-acces la succes*, 93, Societatea Română pentru Asigurarea Calităţii, pp. 344-350, 2008;

[12] M. Mircea and A. Andreescu, "Using Business Rules in Business Intelligence", *Journal of Applied Quantitative Methods*, vol. 4, pp. 382-393, 2009;

[13] M. Mircea, "Development of Business Intelligence system for public acquisitions", ASE Printing House, pp. 427-432, 2009, In: *The Proceedings of the Ninth International Conference on Informatics in Economy*, may 2009;

[14] R. Blasum, "Business Rules and Business Intelligence", *DM Review Magazine*, April 2007, http://www.dmreview.com/issues/20070401/1079638-1.html

Marinela Mircea received her degree on Informatics in Economy from the Academy of Economic Studies, Bucharest in 2003 and his doctoral degree in economics in 2009. Since 2003 she is teaching in Academy of Economic Studies from Bucharest, at Informatics in Economy Department. Her work focuses on the programming, information system, business management and Business Intelligence. She published over 35 articles in journals and magazines in computer science, informatics and business management fields, over 15 papers presented at national and international conferences, symposiums and workshops and she was member over 15 research projects. She is the author of one book and she is coauthor of three books. In February 2009, she finished the doctoral stage, and her PhD thesis has the title *Business management in digital economy*.

Anca Ioana Andreescu is university lecturer in Economic Informatics Department, Academy of Economic Studies of Bucharest. She published over 20 articles in journals and magazines in computer science, informatics and business management fields, over 20 papers presented at national and international conferences, symposiums and workshops and she was member in over twelve research projects. In January 2009, she finished the doctoral stage, the title of her PhD thesis being: *The Development of Software Systems for eBusiness Management*. She is the author of one book and she is coauthor of four books. Her interest domains related to computer science are: business rules approaches, requirements engineering and software development methodologies.

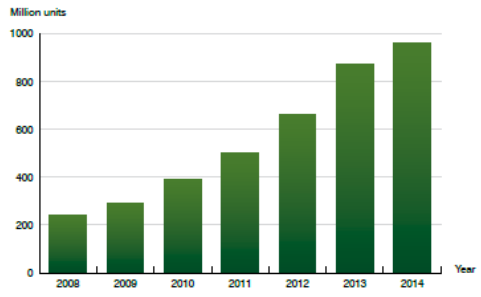
Proposing a Data Model for the Representation of Real Time Road Traffic Flow

Alex Alexandru SIROMASCENKO
Academy Of Economic Studies, Bucharest, Romania
Ingenios Soft Construct
alex_siro@yahoo.com, alex@ingenios.ro

Given recent developments in the fields of GIS data modelling, spatial data representation and storage in spatial databases, together with wireless Internet communications, it is becoming more obvious that the requirements for developing a real time road traffic information system are being met. This paper focuses on building a data model for traffic representation with support from the current free GIS resources, open source technologies and spatial databases. Community-created GIS maps can be used for easily populating an infrastructure model with accurate data; the spatial search features of relational databases can be used to map a given GPS position to the previously created network; open source ORM packages can be employed in mediating live traffic feeds into the model.
A testing mechanism will be devised in order to verify the feasibility of the solution, considering performance
Keywords: *data model, GIS, spatial database, open source technologies*

1 Introduction

Public interest in Web–enabled GIS applications has seen a rise in recent years, mainly due the decreasing cost of portable GPS devices, the increasingly sophisticated mobile devices and applications and the wider accessibility of wireless Internet services.
According to recent market research [1], GPS functions will be included in all but entry level handsets by 2014, after having grown from 8 percent to 15 percent of shipped handsets from 2008 to 2009. Shipments of GPS-enabled handsets with high-speed internet access capabilities are expected to reach 960 milion units, or 60 percent of all handset shipments, by 2014.



Annual shipments of GPS-enabled handsets (Worldwide 2008–2014)
Fig. 1. GPS-handset shipment prediction

As car ownership requires an average or above level of income, the prevalence of GPS-enabled handset ownership is expected to be higher among road traffic participants. Given these forecasts, it becomes apparent that real time collection of the positions of vechicles involved in urban traffic becomes feasible.
Traditionally, traffic data collection for traffic management systems was achieved through partially centralised networks of sensors, which were location-based, recording the traffic information for vechicles passing through a certain point/segment. The main disatvantages of such an approach compared to collecting real time data from each individual vehicle are the setup and maintenance costs of the sensors and the system`s unability to track individual journeys throughout the city.
This latter aspect is the most important, from a functional point of view, because traffic optimisation can benefit from the grouping of per vehicle data in the form of individual itineraries. With such information, a global image of traffic flows can be created, allowing planners to visualise traffic as a

collection of paths and allowing for optimisation through the allocation of itineraries.
Given these requirements, this paper will describe a potential data model for the quick manipulation and storage of traffic information.

2. Analysis of required entities

Given the requirements for real-time data collection and processing, the model will be split into 2 sub-models:

2.1 The static model

This sub-model will be conceived to deal with the representation of the road network and of the vehicles monitored by the system.

The main component of the road network is the **road segment**. A road segment is a traversable portion of a street between two points, open to vehicular traffic. Representing a segment in the data model requires considering two aspects: first, all the factors which can affect the segment`s ability to support vehicle traffic must be taken into account; second, the granularity of the segments will have a major impact on the performance of the system. All vehicle data will be represented in relation to a certain segment, and the speed of segment identification for a given vehicle is important for the process of data collection and mapping. As traffic load transfer can be carried out in any **intersection**, any given street must be split in relation to **crossroads**. The **number of lanes** also affects traffic flow, so a given segment must be further split in relation to any change in lanes number. Furthermore, if a street has two lanes, the flows on each of these lanes are independent and are part of different larger-scale flows, so a segment must be used to represent only one way of traffic.
Several road segments are grouped into a **street**. The street will be mainly used for communicating its name to the user, as it does not offer additional information over the one offered by a segment.
As a segment can be connected with more than two segments (previous and next segment, in the direction of traffic flow), navigating through segments requires a representation of intersections. A **node** is an entity which represents the meeting point of two or more segments. All the information contained in a node is also available in the segment entity, in order to minimise the number of joins required in querying. The redundance of information, in this case, is controlled – the quantity of redundant data can be determined during the generation of the road model, and does not change over time. Also, as the information is updated unfrequently, and its updates are not the subject of many use cases, the redundancy can be accepted, as it offers improved performance with minimal maintenance costs. The node organises the adjacent segments into two groups: the incoming segments and the ongoing segments. Upon arriving at a node through a given segment, a vehicle can be directed through any of the outgoing nodes. In order to optimise performance, the segments representing opposite ways on the same street can be joined together, in order to avoid “turning back” at an intersection.

In the first phase of the project, the road network model does not change over time. In a latter version, the information about the road system will be versionable, by adding a changelog system and history information about the segments, streets and nodes. As entries in these tables are marked as outdated, they will be moved to “history” partitions in order to keep a low number of records for current use, without affecting performance.
A **vehicle** is also part of the static submodel, because, although the positions of vehicles change very frequently, the number of individual vehicles taking part in traffic does not change considerably, sometimes not even on the long run.

2.2 The dynamic model

This sub-model is required to organise the data supplied during the real time usage of the monitoring system. When analysing the performance of the system, the granularity of data supplied to the dynamic model is of primary importance. As each individual vehicle is important in representing an accurate image of the real flow of traffic through the virtual model, the granularity will be controlled solely through the time interval between to consecutive reports of a vehicle`s position. In order to avoid the overcrowding of the system with update reports at certain moments, each client vehicle can be assigned its own reference moment from which data sending cycles will be calculated.

Thus, it is necessary for the dynamic model to contain a representation of the system **clock**. This entity`s purpose is to store information about a reference moment for calculating **reporting intervals**, as well as the size of these intervals. As a basic rule, given a reference moment and an interval, each newly logged in client will be assigned a moment in the last calculated reference reporting interval, in order to uniformly load the system, from a temporal point of view. Taking optimisation one step further, an algorithm can be employed in order to decrease the length of reporting intervals for more circulated road segments. In this manner, processing power is saved for the areas of the road network which can benefit from more frequent data updates, in order to maintain an accurate representation of the traffic flow. Clients will listen to changes in interval lengths while passing through different segments and adjust the communication with the server accordingly.

The dynamic model will combine the existing traffic representation paradigms. In traffic flow theory, modelling can take three approaches:

a.Microscopic model[2] – this category of models takes into account the motion of individual vechicles, with data communicated to the server on a timeframe basis. These models provide an accurate view of traffic conditions, but simulation errors can accumulate and processing costs are very high when applying them for large scale computations and predictions.

b. Macroscopic model – this type of models has been subject of intense research. *"Macroscopic traffic models consist of equations for a few aggregate quantities like the spatial density ρ, the average velocity V , and (in some cases) additional velocity moments. These equations are similar to fluid-dynamic equations, but some fundamental differences with respect to the dynamics of ordinary fluids have recently been recognized"*[3]. Macroscopic models allow for a greater efficiency in using computational resources.

c. Mesoscopic model – is a combination of the previous two models. *"Mesoscopic (meso) models tend to model the individual vehicles, but describe their behaviour in a simplified manner, on an aggregate level. A common way is to describe the vehicles' speed by using a speed/density relation that assigns an average speed based on the density of traffic ahead of the vehicle"* [4]

The mesoscopic approach brings together individual vehicle representation from the microscopic model, along with data aggregation from the macroscopic one.

Individual vehicle itineraries are important, because they can also bring additional traffic information at the macroscopic scale, other than single location segment statistics: they can be used to describe traffic flows from some regions of a city to the other ones, allowing for an accurate representation of infrastructure demands in certain timespans and for the prediction of short and long term traffic trends.

From a **microscopic** point of view, the **vehicle position** is the basis of the dynamic model. The number of vehicle positions, as mentioned before, is given by the **granularity** of updates, which can vary

depending on the congestion in the vehicle`s area. Furthermore, the efficiency of the system can be improved if the **time granularity** is replaced by the **distance granularity** in case the flow speed is very low. In this manner, frequent updates will not be received from vehicles stopped in a column at regular interval, but when these vehicles will advance a certain distance.On the other hand, a problem in maintaining an accurate view of the traffic flow might arise in case vehicles move very fast and pass through several segments during an update interval. In this case, information regarding the presence of vehicles in certain segments might be skipped. To overcome this issue, the **time referrence update system** can be complemented by **location referrence system**, which requires the client to know the start and end points of a segment and automatically report exiting/entering segments. In this case, time updating is important if it takes more time to pass a segment than the duration of the time frame.

From a **macroscopic** perspective, traffic data is represented through **segment loads**. These entities store information regarding the traffic conditions on a certain segment, during a certain interval. Segment load information is updated depending on a parameter called **load granularity**. This parameter, like the position granularity, can vary depending on the congestion of the segment, in order to better describe the conditions in more intensely circulated segments. Data describing the segment load is obtained by aggregating the microscopic information (vehicle position). The **average load** (number of vehicles / percentage of used length) and **average speed** are of primary interest when describing the flow through a segment.

When looking from a **mesoscopic** perspective, the model aggregates traffic data on a **per vehicle basis**.

The **vehicle behaviour** entity stores statistics regarding the manner in which a certain vehicle responds to different stimuli, such as the average speed of the segment and its congestion (this information is obtained from the macroscopic model). It mainly describes a relation between the flow throughout the segment and the vehicle`s speed. It allows for a behaviour-based itinerary allocation scheme, separating slower vehicles (with slower average speed or slower acceleration) from faster ones, which can use high speed segments more effectively. It can also be used to more accurately predict the travel time for each vehicle and it`s impact upon traffic, taking into account the differences between driving behaviours.

The **trip** entity groups the vehicle`s positions and creates a map of preferred departing points and destinations. Further aggregating this information into the macroscopic model, it becomes possible to correlate sources and destinations of traffic flow with respect to hourly patterns, in order to more accurately predict network loads.

3. Building the entity schema

In defining the tables, three aspects were taken into consideration:

3.1 Temporal navigation – given a certain level of aggregation (macroscopic, mesoscopic or microscopic), there must be a link between preceeding and following information and the current information. The vehicle positions must be linked as to easily reconstruct a trip and analyze the vehicle`s experience. Segment loads must be linked in order to easily emphasize the changes in load.

3.2 Inter-level navigation – navigation must be possible between the aggregation levels – a vehicle`s position (microscopic) must be linked to a traffic load (macroscopic) and to a trip (mesoscopic).

3.3 Data redundancy – because the system must deal with very large ammounts of data, processing must be done with a small number of joins between tables, in order not to affect performance. For example, the querying in the segment table in order to

navigate through the network`s state at a certain moment is optimised by joining the simultaneous segment loads through the **NodeLoad** entity.

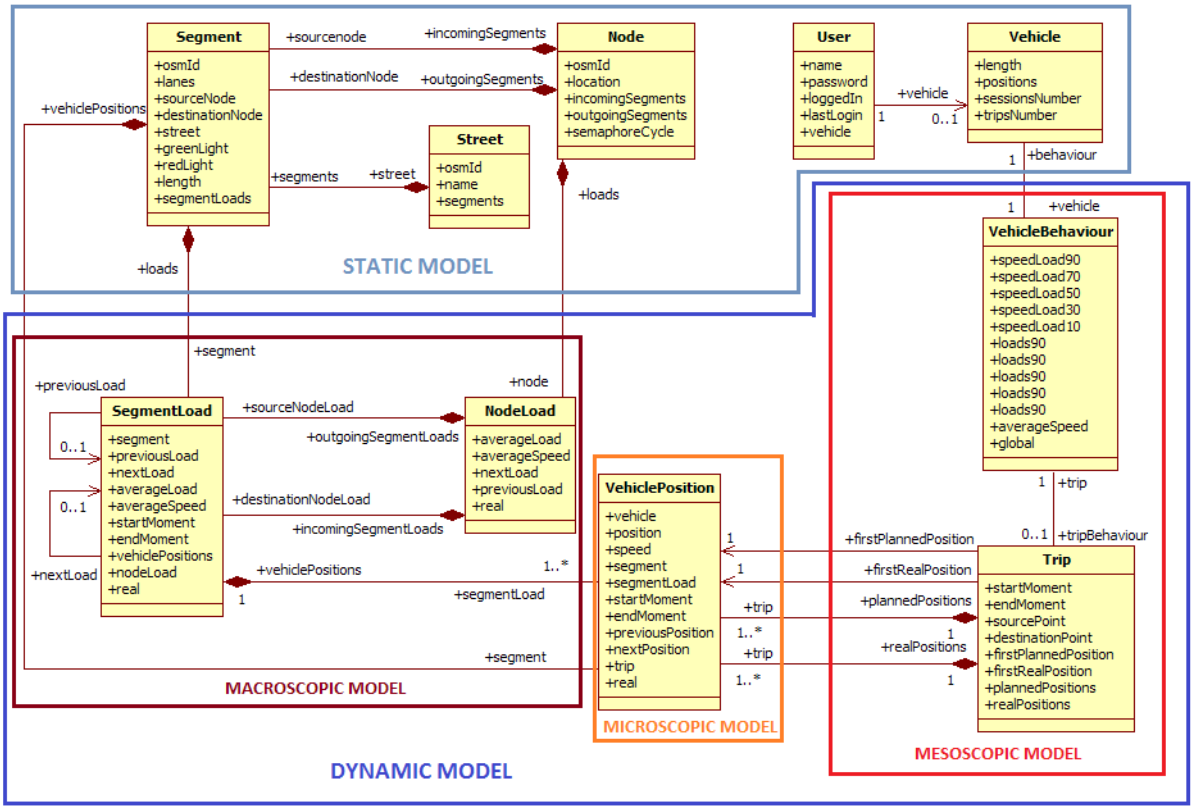


Fig. 2. Proposed data model

4. Spatial ORM

In order to allow object oriented manipulation of data during the model-view transformations and to facilitate the import of the street structure, the Hibernate Spatial ORM library was employed. This generic extension of Hibernate allows for Java representation of spatial structures defined in a spatial database (Oracle Spatial, PostgreSQL). Hibernate Spatial supports most of the functions from the OGC Simple Feature Specification.

Oracle Spatial features from the Oracle Database 11g Enterprise Edition were chosen for representing persistent data.

The Point and LineString implementations provided by Hibernate Spatial were used for mapping Java objects to SDO_GEOMETRY columns in the database[5]. Type matching was handled by the ORM, through the automatic generation of the table schema.

Example: a Java LineString attribute was mapped to an Oracle SDO_GEOMETRY column with the parameters:

SDO_GTYPE: 2002
SDO_SRID: 4326
SDO_ELEM_INFO[0]: 1
SDO_ELEM_INFO[0]: 2
SDO_ELEM_INFO[0]: 1

Additional spatial functionalities were employed through the spatial support offered by the Criteria API

The Hibernate Criteria API allows for the construction of queries from an objectual perspective. Support for spatial features is provided, through the SpatialRestrictions and OracleSpatialRestrictions objects. In this case SpatialRestrictions.within (for fetching all the segments intersecting a given area, represented by a rectangle). and OracleSpatialRestrictions.SDONN

(representation of the eponymous Oracle function, used for finding the nearest segment for a given point) were employed). These two functions make it possible to quickly map a driver on the road system, and to efficiently display real time information regarding the state of the traffic, querying only for the segments in concern.

5. Migrating data from a community project

In order to be functional, the static model must be populated with real-world data. In the context of growing interest in end-user GIS services, collaborative mapping projects have begun to appear. OpenStreetMap is an example of such a project, aiming to create a freely editable map of the world. Any type of source is accepted for map data, including GPS devices, aerial photography, local knowledge of the areas, other free sources or donations from former commercial applications.

The main component of the project is the OSM schema, running on a PostgreSQL database. The schema includes a representation of the submitted data in a proprietary form. Database information can be publicly downloaded or uploaded in the form of *.osm files, which have an XML structure.

Certain portions or the entire database can be downloaded using an online export tool

```
<way                          id="23125530"
user="me_my_self_and_I"     uid="18258"
visible="true"                version="1"
changeset="235251"  timestamp="2008-02-
29T13:36:13Z">
  <nd ref="248729665"/>
  <nd ref="129534986"/>
  <tag    k="created_by"    v="Potlatch
0.7b"/>
  <tag k="highway" v="residential"/>
  <tag  k="name"  v="Strada  Robescu  F.
Constantin"/>
 </way>
<node  id="248729665"  lat="44.4324625"
lon="26.111153"
user="me_my_self_and_I"     uid="18258"
visible="true"                version="2"
changeset="235251"  timestamp="2008-02-
29T13:35:37Z">
  <tag k="created_by" v="JOSM"/>
</node>
```

The main entities in the osm file structure are the way (representing a road, if the "highway" tag is present) and the node (representing GPS coordinates of points defining ways)[6]. In order to import the data, an XML parser was created. The goal was to transfer the way, node and way name to the entities in the proposed model: segment, node and street respectively. An efficient algorithm was devised, which prioritises "way" persistence, with all the available nodes pre-loaded into a transient map for quick fetching. Nodes referred by persisted ways are also persisted and marked as such. Thus, only required information, regarding road structure, is persisted. References to OSM ids are also persisted, in order to easily update the structure at a later point.

6. Testing the performance of algorithms employing the proposed data model

In order to test the performance of the import algorithm, and thus, the maintainability of the system, several tests were conducted, using differently sized *.osm files. The results are shown below:

Table 1. OSM file import performance

no	Segm	Nodes	Streets	Parse time	Persist time
1	5739	6127	712	1 sec.	4 sec.
2	2972	2477	284	<1 sec.	3 sec.
3	28951	52873	3523	1 sec.	16 sec.

The persistence of the entities also included spatial indexing of the nodes and segments.

Testing of the dynamic model was carried out through a random trip generation mechanism. The algorithm used a number of randomly generated pairs of departure and destination points. The points were generated within a user specified area and a trip was simulated for each pair of points. Given a source and a destination, a shortest path was found based on distance, using

the Dijkstra algorithm. After the shortest path was determined, the system simulated the actual trip. The trip involved position reports at each segment switch.

Simulation statistics:
trips: 1000
reports: 72902
total time: 2min 21sec
average report cost: 1.9 ms

The tests were conducted on a system featuring a dual-core Core i5-560M processor clocked at 2.67 GHZ and 4GB of RAM.

7. Conclusions

Live feed models for the representations of road traffic are becoming more feasible with the support of readily available resources.
In order for such a model to offer a balance between detailed data representation and quick processing of large volumes of information, the data must be split at three levels of abstraction: microscopic (vehicle positions) mesoscopic (vehicle behaviors and trips) and macroscopic (node and segment loads). The granularity of the data updates can be used to fine tune the model`s performance by increasing reports from more intensely used areas and by uniformly distributing reports, when time is considered.
Community projects such as OpenStreetMap offer an appropriate interface for getting access to an extensive database of spatial information. ORM tools with spatial features give the opportunity of bringing GIS operators to the end user.
When used in a simulation, the proposed model offers real-time performance in keeping track of the road network usage and can be quickly updated in order to respond to changes in the infrastructure.

8. Acknowledgment

This work was cofinanced from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards" (DOCCENT).

References

[1] Berg Insight AB Stockholm "GPS and Mobile Handsets – 4th edition, March 2010"
[2] Dirk Helbing "From microscopic to macroscopic traffic flow models", A Perspective Look at Nonlinear Media Lecture Notes in Physics, 1998, Volume 503/1998, 122-139, DOI: 10.1007/BFb0104959
[3] M. van den Berg, A. Hegyi, B. De Schutter, and J. Hellendoorn, "A macroscopic traffic flow model for integrated control of freeway and urban traffic networks," *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, Hawaii, pp. 2774–2779, Dec. 2003
[4] W. Burghout, "Hybrid Microscopic–Mesoscopic Traffic Simulation", *Doctoral Dissertation,* Royal Institute of Technology, Stockholm, Sweden, 2004.
[5] Hibernate Spatial API - *http://www.hibernatespatial.org/hibernate-spatial-oracle/apidocs/index.html*
[6] OpenStreetMap Wiki – Data Primitives *http://wiki.openstreetmap.org/wiki/Data_Primitives*

Alex Alexandru SIROMASCENKO graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2008 (Bachelor`s degree) and in 2010 (Master`s degree), specialising in Economic Informatics. He is currently a PhD candidate at the Academy of Economic Studies. His main domains of interest are road traffic optimisation, spatial databases and GIS technologies, data and business modelling, Web technologies and applications.