Preview of methods and tools for operating data analysis

Zdenka Prokopova, Petr Silhavy, and Radek Silhavy

***Abstract*—** The main aim of the paper is to present a preview of methods and tools for operating or business data analysis with regards to availability of final users. The objective of analytical methods and tools is obtaining necessary and useful information from collected data and consequently utilizing them for active control and decision making. The paper outlines an overview about contemporary state of art and trends in the field of data analysis.

***Keywords*—**Data Analysis, OLAP, Business Intelligence, Data Mining, Datawarehouse.

I. Introduction

COLLECTING, storing, merging and sorting enormous amounts of data have been a major challenge for software and hardware facilities. Increasing number of companies and institutions has solved and developed tools for saving and storing tables, documents or multimedia data. During the years database structures became a major instrument in prevailing applications nowadays known as Business Intelligence applications.

The origin of Business Intelligence principles is connected with the name Hans Peter Luhn – experimental staff member of IBM. He published in the IBM Journal article entitled "A Business Intelligence System" in 1958 where main principles and cogitations were formulated [1]. The main philosophy idea consists in the principle that commercial aims of the company should have been defined on the bases of present facts evaluation. These presumptions led in various software program implementations intended to administration of manager information. In 1989, the term Business Intelligence defined by analyst Howard J. Dresner was introduced to the wider public awareness. He described them as a set of concepts and methods intended for improving the quality of analytical and decision-making processes in organizations. He focused on importance of data analysis, reporting and query tools, which offer to user amount of data and help him with synthesis of valuable and useful information [2], [3].

Early information systems in large companies and banks were operated since 60th years of last century. In spite of the title Management Information Systems there were only common routine agenda specialized to accounting data processing. Special systems intended not only for everyday operational control but especially for strategy management began to create a new discipline from seventies. These types of applications are known as Decision Support Systems. Their basic imposition was providing of information and tools for the modeling and evaluation of various business alternatives and strategies. The development of decision-making systems was supported also with expansion in the hardware and software area. Two points can be seen as a key factor in the development. The first one is the data access speed changes. The second one is revolutionary proposal of relational data model introduced by E. F. Codd. This model is based on mathematical set theory. With entrance of graphic-oriented user interface, the third wave of tools for helping in control processes appeared. There are so called Executive Information Systems (or Executive Support Systems) which offer on-line access to actual information about state of controlled organizations for top managers. First applications of this type worked right on the purchased data. However, it was a big primary system workload and therefore came to separation of service data and data for analyses [4], [5].

The depth data analysis of Business information systems and their subsequent utilization at company control can be labeled by the common mark – Business Intelligence. Analytical and planning characters of Business Intelligence applications differ from the ordinary operating systems in user's look on data. While operating systems work with detailed information then analytic exercises work with aggregate data. So that the analytical look into data imposed the necessity of changing the data access technology. Operating systems work with transaction entity-relational databases analytical systems work with data warehouses and multidimensional databases. The graphical interpretation of brief history of Business Intelligence we can see on Fig. 1. [6]

Manuscript received June 26, 2011. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the Research Plan No. MSM 7088352102.

Z. Prokopová is with the Tomas Bata University in Zlin, nám. T. G. Masaryka 5555, 760 01 Zlín, Czech Republic (corresponding author to provide phone: +420 57 603 5011. e-mail: prokopova@fai.utb.cz).

P. Silhavy is with the Tomas Bata University in Zlin, nám. T. G. Masaryka 5555, 760 01 Zlín, Czech Republic (phone: +420 57 603 5015; e-mail: psilhavy@fai.utb.cz).

R. Silhavy is with the Tomas Bata University in Zlin, nám. T. G. Masaryka 5555, 760 01 Zlín, Czech Republic (phone: +420 57 603 5015. e-mail: rsilhavy@fai.utb.cz).

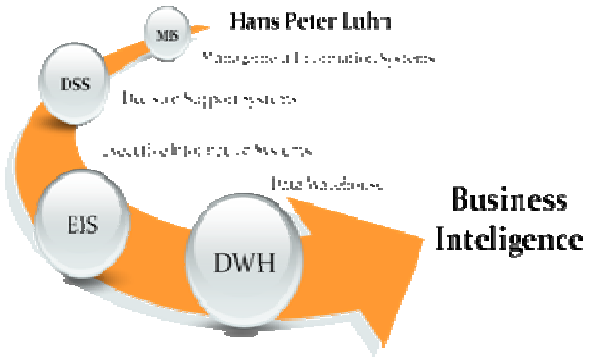

Fig. 1. Brief history of Business Intelligence

II. Data Transformations – Data Pumps

Data acceptable for next analysis have to be extracted from operational systems and put into data store. After that we can perform analyses by the help of OLAP technology, Data Mining technology or by the help of reporting services to create reports. This action is at the creation of data stores most important as well as more exacting. It is necessary to ensure analysis of contain and technologically heterogeneous data sources and then choose relevant data and centralize, integrate and aggregate them each other. Data pumps serves to collection and transmission of data from source systems to data stores and dumping ground. They include:

- ETL systems for extraction, transformation and transmission of data
- EAI systems for application integration (work in contrast to ETL tools in real-time).

A. *ETL – Extract, Transform and Load*

Data store filling (ETL process) starts by data extraction from primary sources (Extraction). During this phase there are seek out and remove various data inconsistency. Before their transformation into the data schema extracted data can be loaded in temporary dumping ground. Temporary dumping ground data component (Data Staging Area – DSA) used to be most frequently a part of those solutions of data stores which has a source in heavy transaction systems. By using of DSA will reduce requirement of transaction systems utilization in the ETL process and they can be used at business processes service. DSA is possible to use also in the case when is necessary to transfer data from for example text file into the required database format. After the extraction follows data transformation (Transformation) which will convert data obtained from single data sources into unified data model. This model makes it possible to create aggregations and clustering.

The final phase of ETL is data transmission from source data memories or temporary dumping ground to database tables of the data store. At the primary filling it can be a gigantic quantity of data. Because ETL works in batch mode next regular updating brings only such amount of data which corresponding with used time period (day, week, month, year).

B. *EAI – Enterprise Application Integration*

EAI tools are exploited in source system layer. Their aim is integration of primary business systems and reduction of a number of their reciprocal interface. These tools work on two levels:

- at the level of data integration where there are used for integration and data distribution
- at the level of application integration where there are used for sharing of selected functions of information systems.

III. Database Components – Data Warehouse

The philosophy of data warehouse (stores) has published for the first time by Bill Inmon in the book Building the Data Warehouse in the year 1991. Genuine reason of data warehouse occurrence had connection especially with massive setting of server business systems and their conception of separate and independent application at the end of eightieth years of last century. Data warehouses ware established as independent information systems set above business or operating data. While data warehouses are subject-oriented (data are separated according to types) data markets are problem-oriented. For the purpose of data storage served new multidimensional database model which enabled easily and quickly create various views on data by the help of special cuts of data cube. This technology is the bases of today analytical tools of Business Intelligence. By connection of BI with tools of business planning was created a new type of application called Corporate Performance Management (CPM).

Data warehouses are special types of databases which contain consolidated data from all accessible service systems. There are not optimized for quick transaction processing but quick administration of analytical information obtained from big amount of data. Data warehouses ensuring processes of storing, actualization and administration of data. There are exists two basic types of data stores and two types of auxiliary stores

A. *Basic data stores*

- Data Warehouse (DWH)
  Data warehouse is wide (extensive) central database in which are saved transformed data coming from various service systems and external databases. Mentioned data are intended to following analyses. [7]
- Data Marts (DMA)
  The principle of data marts is similar as the principle of data warehouses. Difference is only in one point of view - data marts are decentralized and thematic oriented. Provided analytical information are aimed to specific user group (marketing, selling etc.).

B. *Auxiliary data stores*

- Operational Data Store (ODS)
- Data Staging Areas (DSA)

C. *Schemes for data stores*

Data models of working systems used to be very coplicated because they contain a lot of tables and relations. It was appeared an effort to simplify ERD diagrams and their conformation to data stock requirements. There were created two types of dimensional models for data type structure. We can distinguish them according to connection between tables of dimension and table (tables) of facts:

- Star schema – in this schema (see Fig. 2) are data insert in one table so called "non-normal". Hierarchies of dimensions are created only by levels whose items are in one table. It causes complicated ETL process but on the contrary offers high query performance.

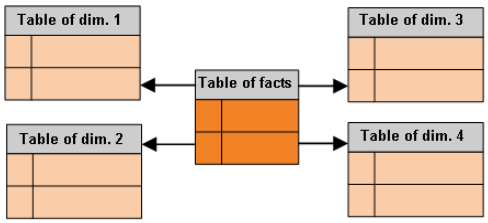

Fig. 2. Star schema for data stores

- Snowflake schema – in this schema (see Fig. 3) are data widespread in several related tables with cardinality 1:N. Obviously are tables in third normal form. It causes restriction of redundant data but by reason of more connections between tables is decreasing the query performance.

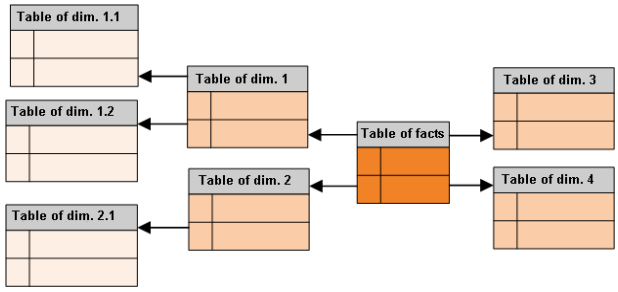

Fig. 3. Snowflake schema for data stores

IV. Analytical Components

A. *Analysis of multidimensional data - OLAP*

Data in data warehouse are cleaned out and integrated but often very voluminous. There are use special data structures and technology for their analysis known as OLAP (On-line Analytical Processing). OLAP tools are simple, readily available and very popular susceptible to create multidimensional analysis. There are for example pivot tables from MS Excel. There were defined 12 rules for OLAP by Dr. Codd in 1993:

- *Multidimensional conceptual view* - the system should offer multidimensional model corresponding to individual needs and enable intuitive manipulation and analysis of gained data.
- *Transparency* - the system should be connected to front-end systems.
- *Availability* - the system should offer only data needed to analysis. Users are not interested in the way how the system approaches to heterogeneous sources.
- *Consistent effort* - the system effort mustn't depend on the number of system dimensions.
- *Client-server architecture* - OLAP system has to be client-server type.
- *Generic dimensionality* - each dimension of data has to be equivalent in structure and operational abilities.
- *Dynamic treatment of sparse matrices* - the system should by able to adapt its physical scheme to analytical model optimizing treatment of sparse matrices.
- *Multi user support* - the system should by support team work of users and parallel data processing.
- *Unlimited crosswise dimensional operation* - the system has to distinguish dimensional hierarchy and automatically execute associated calculations.
- *Intuitive manipulation with data* - user interface should be intuitive.
- *Flexible declaration* - the system should be allows changes in rows and columns disposals (according the analysis needs).
- *Unlimited dimension number and aggregate levels* - OLAP system shouldn't implement any artificial restriction of dimensions or aggregation levels.

B. *Description of the OLAP technology*

The OLAP technology works with so called multidimensional data. In contrast to two dimensional data storage in relation databases (columns and rows) here is using n-dimensional Data Cube. The Data Cube can be considered as an n-dimensional hypercube known from analytic geometry. Comparison of structures of relational and multidimensional databases is shown on figures Fig. 4 and Fig. 5.

Multidimensional database is not normalized. It is formed from tables of dimensions and facts organized into schema. Every dimension represents other visual angle on data. Data could be organized not only logically but also hierarchically. Numerical data came from process are in table of facts. [8]

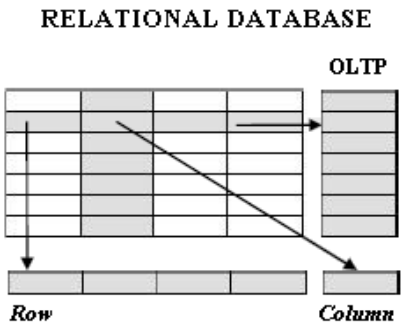

Fig. 4. Structure of relational databases

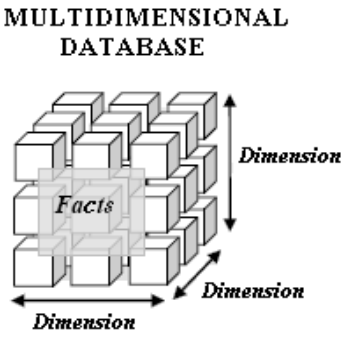

Fig. 5. Structure of multidimensional databases

C. Physical realization of multidimensional data model

MOLAP – Multidimensional OLAP

It needs for its work special multidimensional database which is periodically actualized by data from data warehouse. MOLAP is useful for small and middle sized data quantity.

ROLAP – Relational OLAP

It works above data warehouse or dada mart relational database. Multidimensional queries automatically translate to corresponding SQL queries (SELECT). ROLAP is useful for extensive data quantity.

HOLAP – Hybrid OLAP

It is specific combination of both approaches. Data analysis works with relational databases but aggregations are stored in multidimensional structure (in data warehouse).

DOLAP – Dynamical OLAP

This is special type of OLAP when the multidimensional Data Cube is constructed virtually in RAM memory. Basic advantage of this solution is unlimited flexibility and disadvantage id significant demands on RAM memory.

D. Knowledge mining from data

Data mining is process of looking for information and hidden or unknown relations in big mass of data. Development of this analytical method has connection with enormous data rising in companies databases. There are increasing not only data but also the number of errors (bugs) in data. Data mining work on the intuitive principle when on the basis of real data are created possible hypotheses. These hypotheses need to be verified and according solutions adopt or reject. [9]

Data mining arose by connection of database and statistical discipline. It utilizes various complicated algorithm whereby it is possible to predicate development or segment (or cluster) related data. From mathematical and statistical theory point of view there is based on correlations searching and hypotheses testing.

For the data mining is very important quality of input (incoming) data. If data do not contain some important statement the analysis solution couldn't be correct. For this reason it is very important preparation of data intended for analysis. Usually there is created one table from data warehouse which contains obviously preprocessed and cleaned data. [10], [11]

Objective setting

Ordinarily, there is the same real problem which is the impulse to start the data mining process. At the end of this process should be amount of information suitable for solving the defined problem. Perhaps marketing is area of largest use of the Data Mining.

Data selection

In this phase it is necessary to choose data for the Data mining not only according alignment point of view (demographical, behavioral, psychological etc.) but source databases too. Data are usually extracted from source systems to special server.

Data preprocessing

Data preparation is most exacting and most critical phase of the process. It is necessary to choose corresponding information from voluminous databases and save it to simple table. Data preprocessing consist of next steps:

- Data clearing – solving of missing or inconsistent data problem,
- Data integration – various sources cause problems with data redundancy, nomenclature,
- Data transformation – data have to be transformed to suitable format for data mining,
- Data reduction – erasing of unneeded data and attributes, data compression etc.

Data mining models

Previously prepared data can be processed by special algorithm to obtain mathematical models.

- Data exploration analysis – independent data searching without previous knowledge.
- Description – describe full data set. There are created groups according behavior demonstration.
- Prediction – it is trying to predict unknown value according to knowledge of the others.
- Retrieval according to template – the analyst aim is to find data corresponding to templates.

Data mining methods

- Regression methods – linear regression analysis, nonlinear regression analysis, neural networks,
- Classification – logistic regression analysis, decision trees,
- Segmentation (clustering) – clustering analysis, genetic algorithms, neural clustering,
- Time series prediction – Box-Jenkins method, neural networks,
- Deviation detection.

V. TOOLS FOR END - USERS

A. Analytical tools of MS SQL server 2008

From the beginning of OLAP Microsoft made effort to create the model of self-service analytical tools. In the version MS SQL Server 2005 were joined all analytical levels into Unified Dimension Model. In the version MS SQL Server 2008 is the focal point in Analysis Services which are containing OLAP, Data Mining, Reporting Services and Integration Services.

Integration Services

SQL Server Integration Services (SSIS) works as a data pump ETL. It allows creating applications for data administration, manipulation with files in directories, data import and data export.

Reporting Services

SQL Server Reporting Services (SSRS) provides flexible platform for reports creation and distribution. It cooperates with client tool MS SQL Server Report Builder which is complexly free for end-users.

Analysis Services

SQL Server Analysis Services (SSAS) is a key component of data analysis. It consists of two components:

- OLAP module for multidimensional data analysis enabling loading, questioning and administration of data cubes created by Business Intelligence Development Studio (BIDS)
- Data Mining module which extended possibilities of data analyses.

B. Data analysis user tools - MS Excel

The simplest and most obtainable analysis proceeding of business or operating data offers MS Excel. Certainly it is too the cheapest way because there is no manager or chief executive without this program installed on their notebooks or PC. That why there is not necessary to by license for specialized software. Users could create analytical reports and graphs immediately. Data analyses created by MS Excel are very dynamic and effective. They enable a lot of different views and graphical representations. Data into MS Excel we can obtain by several ways. Most common is the manual table filling form reports. The second way is easier and it is data import from business information system. The third way represents direct connection to database of business information system. This way is most operative.

Data analysis by pivot tables and graphs

Pivot tables are one of the most powerful tools of MS Excel. Enable data summarization, filtration and ordering. There is possible to create a lot of different views, reports and graphs from one data source. Created pivot table is easily variable - we can add or delete data, columns, rows or change summaries without influences of data source. Pivot tables are very often use as a user tool for work with data cube used by MS SQL Server.

VI. Example

From the manufacturing processes point of view it is interesting utilization of data mining, data warehouses, OLAP (Business Intelligence) at analysis of technological process stage, prediction and diagnostic of abnormal stages and looking for technological connections in historical data rising as a secondary product of monitoring. As an example is mentioned utilization of SQL Server Analysis Services as a key component for operating data analysis. For multidimensional data analysis enabling loading, questioning and administration of data cubes we used OLAP module created by Business Intelligence Development Studio (BIDS). [12], [13]

A. Data preparation

On the beginning we had data in various formats of Excel tables. It was necessary to sort data and design uniform structure. On the ground of utilization of analytical tools for data analyses we created data warehouse with several tables of facts and several tables of dimensions.

Structure of the designed data warehouse we can see on the figure Fig.6.

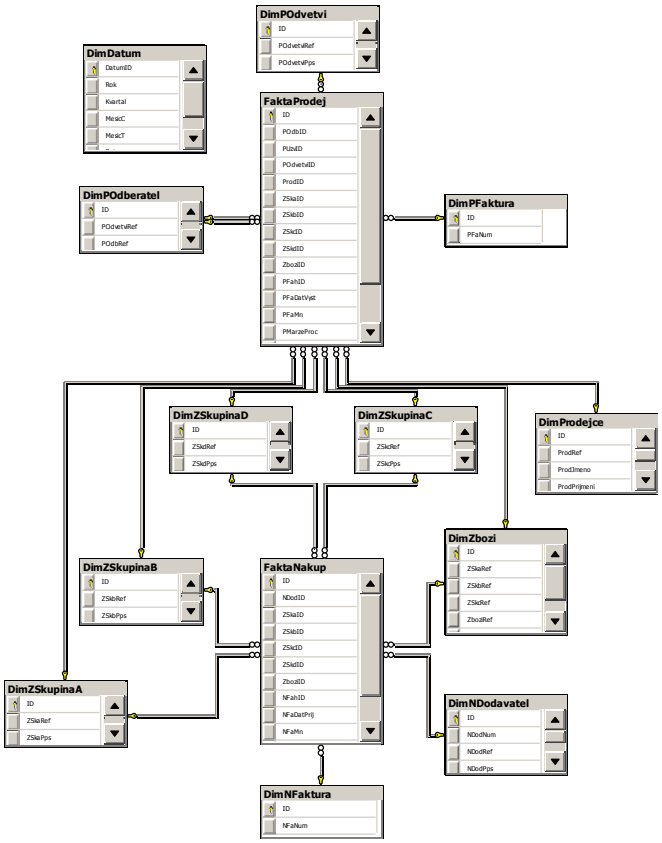

Fig. 6. Structure of the data warehouse

For interlocking of connections among tables of dimensions and tables of facts it was necessary to prepare and create all needed tables of keys.

B. Data cube creation

Data cube creation by Business Intelligence Development Studio consists of next steps:

- Definition of data sources
- Definition of the data source view
- Design of the data cube
- Dimensions configuration
- The data cube equalization

If we want to create a new project we must choose Analyses Service project and after definition of name and location of the project it will open the window of development environment for data cube modeling as we can see on the Fig. 7.

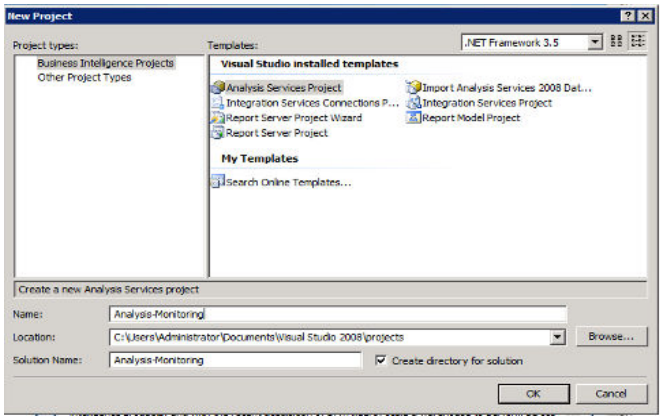

Fig. 7. Project creation in BIDS

1) Definition of data sources

The first step in data cube modeling is creation of Data sources (Fig. 8). Data source refer to database, from which project approaches to data. Data source can be connected with any database available by OLE DB or ODBC.

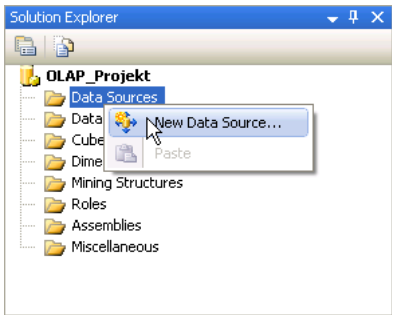

Fig. 8. Definition of data sources

Data source can be defined in Solution Explorer window where we can choose offered data source or we can create a new one by Data Source Wizard. In the case of creation of new data source we have to choose server, database, and type of desired security.

2) Definition of the data source view

As in the previous case we can use Data source View Wizard (Fig. 9).

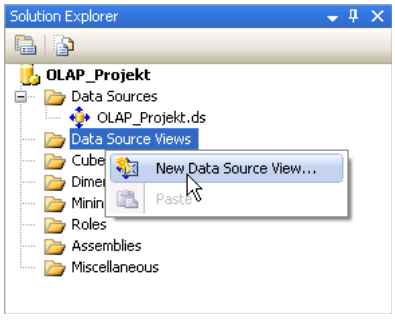

Fig. 9. Definition of the data source view

In the first step the appropriate data source is chosen. Then we choose needed tables from which data source consists of. The next step of completing the wizard is providing the new data source by the name – see Fig. 10.

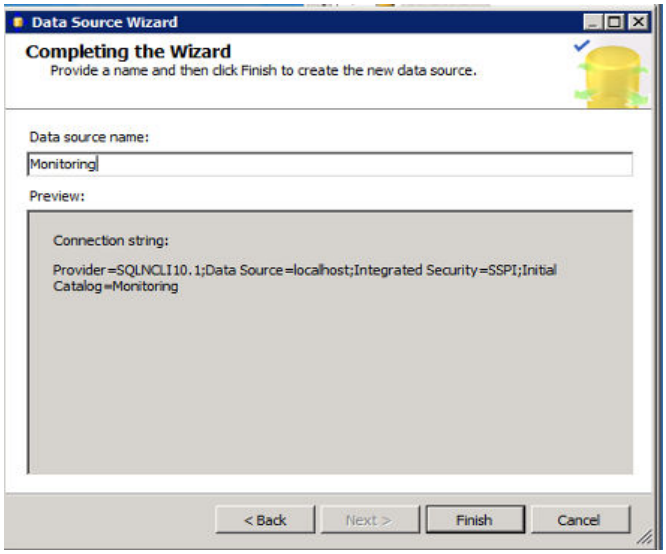

Fig. 10. Definition of Data Source connection

After wizard finishing we have the new data source view (as dsv file). In the same time is appearing structure of data diagram in the design window. There are tables of dimensions, tables of facts and their connections as we see on Fig. 11.

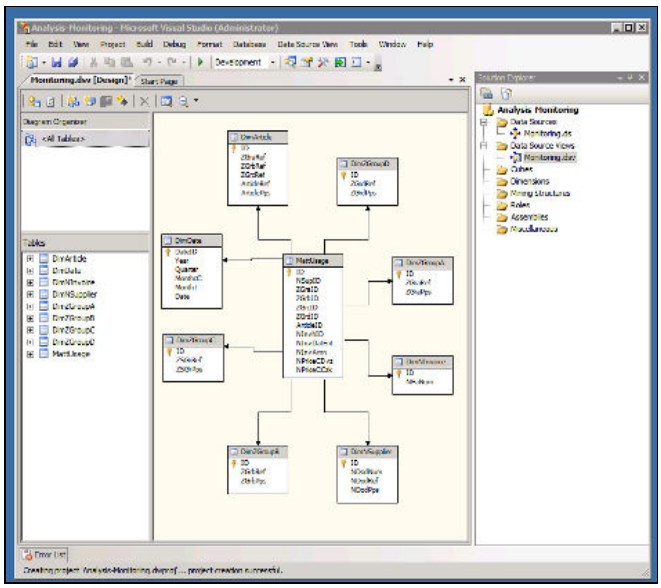

Fig. 11. Setting of Data Source View

3) Definition of the data cube

The last step is Data cube composition. In this step it is possibility to use wizard once again and create new data cube as we can see on Fig. 12.

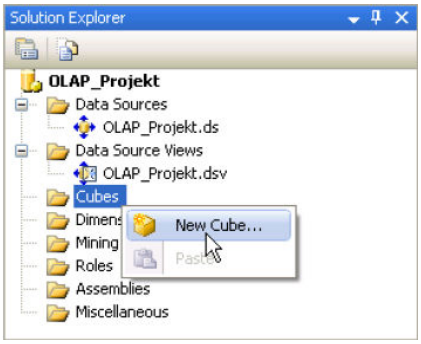

Fig. 12. Definition of the data cube

In the first step is necessary to choose way of data cube creation. We can create empty data cube by the help of data source. After that we ought to choose measures from tables of facts and dimensions from tables of dimensions. Demonstration of new data cube composition is shown on Fig. 13.

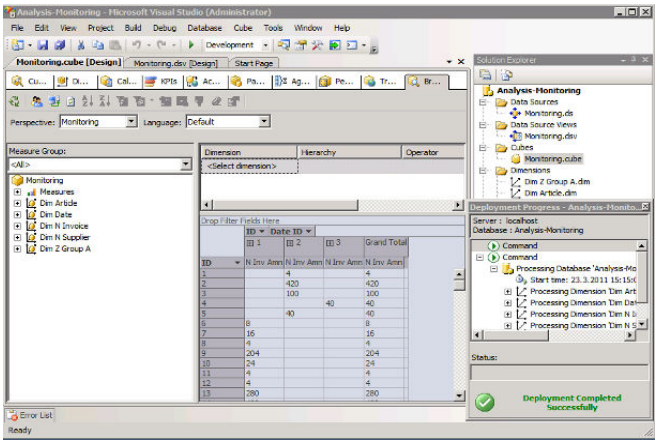

Fig. 13. Data Cube composition

4) Dimensions configuration

The last step before project publication is dimension configuration. In this phase are characteristics set up and hierarchies defined. Adjustment is providing in Dimension Designer which serves firstly to end users.

5) The data cube equalization

Final phase of project consists of two parts. Part Build covers development and preliminary tuning on local developer computer, part Deploy boots debugged program to analytical server and hand down him to end users. End user can choose between access of thick client, when client application runs on their local computer (MS Excel, Report Builder) and thin client, which employs web browser for approach to server application and data. In the event of thick client there is not so big server loading as in the case of thin client, because computation achievement divides between the local computer and the server. For thin client all application logic is on server and its computational possibilities are unused.

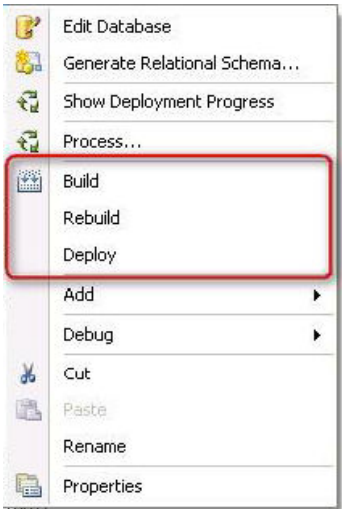

Fig. 13. Project publication

Completed Data Cube we can see in browser environment or we can draw it for better understanding as a three dimensional cube.

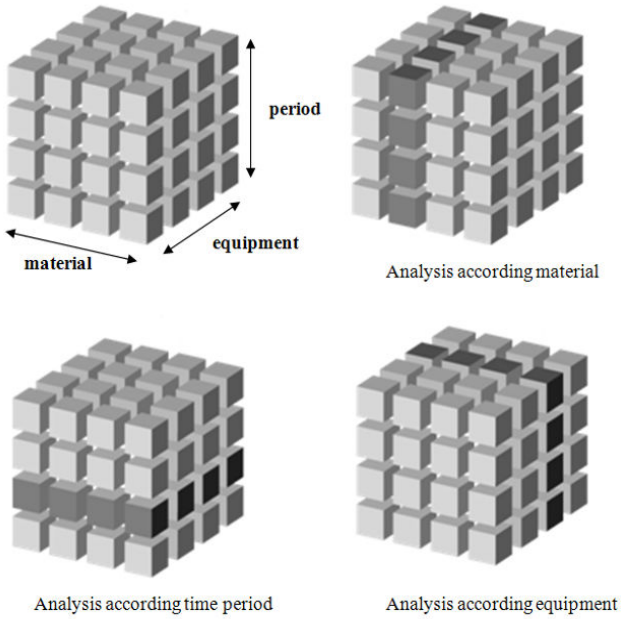

Fig. 14. Graphical interpretation of data analysis by the help of 3-dimensional data cube

VII. CONCLUSION

However, data analysis is not only for top managers and their strategy decision making, but find exercise on all control levels. Tools allowing depth business or operating data analysis are already integrated into database engines and are so direct part of business information systems. They stand above transactional databases and communicate with ordinary office programs. Thereby offer business or operating data access possibilities to all company users.

High - quality data analysis and level of gained information stands on background of all correct manager decisions. Good managers are able to use it for improvement of efficiency and

company competitive advantage by prediction of trend and future development tendencies. There are able to disclose the market anomaly and focus on suitable interest client groups.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic in the range of the project No. MSM 7088352102.

REFERENCES

[1] H. P. Luhn, "A Business Intelligence Systems". *IBM Journal of Research and Development*, 1958, pp. 314-319.
[2] M. Berthold, D. Hand, *Intelligent Data Analysis*. Springer, Berlin, 2009.
[3] P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005. ISBN 0-321-32136-7.
[4] G. Shmueli, N. R. Patel, P. C. Bruce, *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with Xlminer*. 2006. ISBN 0-470-08485-5.
[5] D. Pokorná, "Business Data Analyses Possibilities". *Diploma thesis*. Faculty of Applied Informatics, Tomas Bata University in Zlín. 2010.
[6] D. Power, Dssresources.com [online]. 2007 [cit. 2010-06-07]. "A Brief History of Decision Support Systems". From WWW: <http:// dssresources.com/history/dsshistory.html>.
[7] S.-K. Choi, T. Lee and J. Kim, "The genetic heuristics for the plant and warehouse location problem," *WSEAS Transactions on Circuits and Systems*, vol. 2, no. 4, 2003, pp. 704–709.
[8] P.Sajda, A.Gerson, K.R.Muller, B.Blankertz and L.Parra, "A Data Analysis Competition to Evaluate Machine Learning Algorithms for use in Brain-computer Interfaces", *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 11, no. 4, Dec. 2003, pp. 422-426.
[9] Tzung-Pei Hong, "Some Issues and Approaches in Data Mining", *Proceedings of the 10th WSEAS International Conference on Applied Informatics And Communications* (AIC '10) and the *3rd WSEAS International Conference on BIOMEDICAL ELECTRONICS and BIOMEDICAL INFORMATICS* (BEBI '10), Taipei, Taiwan, August 20-22, 2010.
[10] J. Skala and I. Kolingerova, "Faster Facility Location and Hierarchical Clustering", *International Journal Of Computers*, Issue 1, Volume 5, 2011, pp. 132-139.
[11] Y. Mizuno, H..Mabuchi, G. Chakraborty and M. Matsuhara, "Clustering of EEG data using maximum entropy method and LVQ", *International Journal Of Computers*, Issue 4, Volume 4, 2010, 193-200.
[12] I. Lungu and A. Mih*alache*, "An adaptive modeling approach in collaborative data and process-aware management systems", *International Journal Of Computers*, Issue 4, Volume 4, 2010, pp. 145-152.
[13] J. Savkovic-Stevanovic, L. Filipovic-Petrovic and R. Beric, "Network service systems for chemical engineers", *International Journal Of Mathematical models And Methods In Applied Sciences*, Issue 1, Volume 5, 2011, pp. 105-114.

Zdenka Prokopova was born in Rimavská Sobota, Slovak Republic in 1965. She graduated from Slovak Technical University in 1988, with a master's degree in automatic control theory. Doctor's degree she has received in technical cybernetics in 1993 from the same university.

She worked as assistant at Slovak Technical University from 1988 to 1993. During years 1993-1995 she worked as programmer of database systems in Datalock business firm. From 1995 to 2000 she worked on position lecturer at Brno University of Technology. Since 2001 she has been at Tomas Bata University in Zlin, Faculty of Applied Informatics. She presently holds the position of associating professor at the Department of Computer and Communication Systems. Her research activities include programming and application of database systems, mathematical modeling, computer simulation and control of technological systems.

Petr Silhavy was born in Vsetín in 1980. He received a B.Sc. (2004), M.Sc. (2006), and Ph.D. (2009) in engineering informatics from Faculty of Applied Informatics, Tomas Bata University in Zlin. He is a senior lecturer and researcher at the Computer and Communication Systems Department. His Ph.D. research was on the electronic communication and services in a medical information systems.

Major research interests are data mining, database systems and web-based services.

Radek Silhavy was born in Vsetín in 1980. He received a B.Sc. (2004), M.Sc. (2006), and Ph.D. (2009) in engineering informatics from Faculty of Applied Informatics, Tomas Bata University in Zlin. He is a senior lecturer and researcher at the Computer and Communication Systems Department. His Ph.D. research was on the verification of the distributed schema for the electronic voting system.

Major research interests are software engineering, empirical software engineering and system engineering.