Data analysis: tools and methods

PROKOPOVA ZDENKA, SILHAVY PETR, SILHAVY RADEK
Department of Computer and Communication Systems
Faculty of Applied Informatics
Tomas Bata University in Zlin
nám. T. G. Masaryka 5555, 760 01 Zlín
CZECH REPUBLIC
prokopova@fai.utb.cz, psilhavy@fai.utb.cz, rsilhavy@fai.utb.cz http://web.fai.utb.cz

Abstract: - The paper outlines an overview about contemporary state of art and trends in the field of data analysis. 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. Database structures are a major instrument in prevailing applications. These structures have everyday thousands or millions entries. The objectives of analytical tools is obtaining necessary and useful information from collected data and consequently utilizing them for active control and decision making. The main aim of this contribution is to present some possibilities and tools of data analysis with regards to availability of final users.

Key-Words: - Data Analysis, Business Intelligence, Data Mining, OLAP, Datawarehouse

1 Introduction

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.

2 Tools of analytical systems

2.1 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).

2.1.1 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).

2.1.2 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.

2.2 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 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 business 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

2.2.1 Basic data stores

- *Data Warehouse (DWH)*
  Data warehouse is wide (extensive) central business database in which are saved transformed data coming from various service systems and external databases. Mentioned data are intended to following analyses.

- 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.).

2.2.2 Auxiliary data stores

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

2.2.3 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 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.

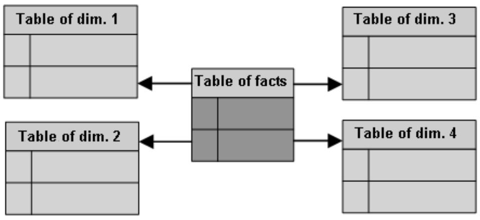


Fig. 1. Star schema

- Snowflake schema – in this schema 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.

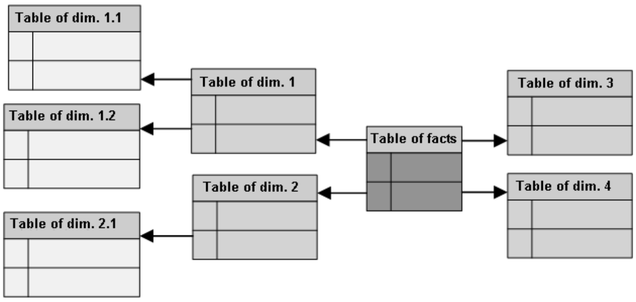


Fig. 2. Snowflake schema

2.3 Analytical components

2.3.1 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 business 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.

2.3.2 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.

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.

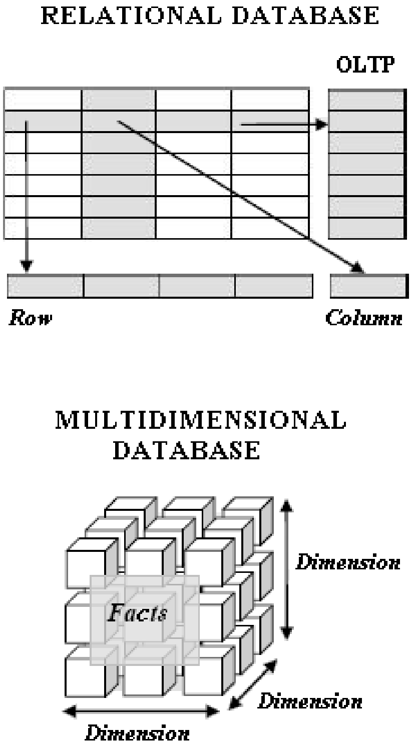


Fig. 3. Structures of relational and multidimensional databases

2.3.3 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.

2.3.4 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.

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 preprocessed and cleaned data.

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

2.4 Tools for end - users

2.4.1 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 business analyses.

2.4.2 Data analysis user tools - MS Excel

The simplest and most obtainable analysis proceeding of business 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 business 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

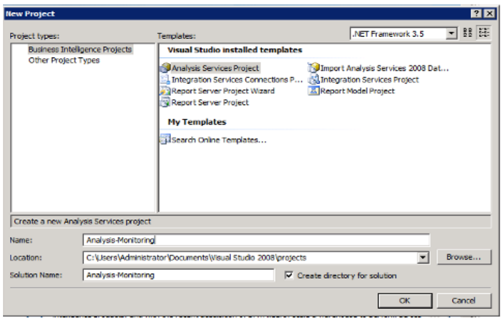
Pivot tables are one of the most powerful tools of MS Excel. Enable data summarization, filtration and ordering. There is possible to create a lot of different views, reports and graphs from one data source. Created pivot table is easily variable - we can add or delete data, columns, rows or change summaries 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.

3 Example

From the manufacturing processes point of view it is interesting utilization of data mining or OLAP 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 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).

If we want to create a new project we must choose Analyses Service project.


New Project
? ×
Project types:
Business Intelligence Projects
Other Project Types
Templates:
.NET Framework 3.5
Visual Studio installed templates
Analysis Services Project
Import Analysis Services 2008 Dat...
Integration Services Connections P...
Integration Services Project
Report Server Project Wizard
Report Model Project
Report Server Project
My Templates
Search Online Templates...
Browse...
OK
Cancel

Create a new Analysis Services project

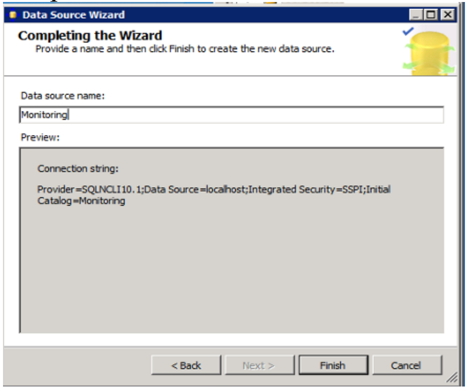
Name: Analysis-Monitoring

Location: C:\Users\Administrator\Documents\Visual Studio 2008\projects

Solution Name: Analysis-Monitoring ☑ Create directory for solution

Fig. 4. Project creation in BIDS

Second step is creation of Data Source connection.



Completing the Wizard

Provide a name and then click Finish to create the new data source.

Data source name: Monitoring

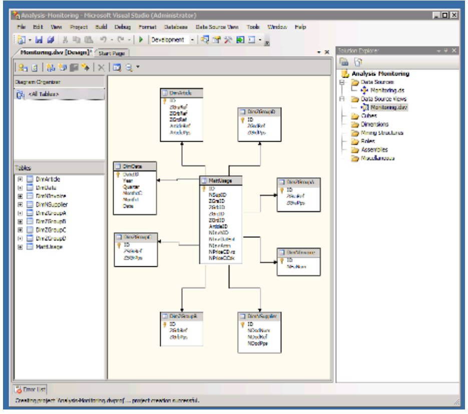
Preview:

Connection string:

Provider=SQLNCLI10.1;Data Source=localhost;Integrated Security=SSPI;Initial Catalog=Monitoring

Fig. 5. Definition of Data Source connection

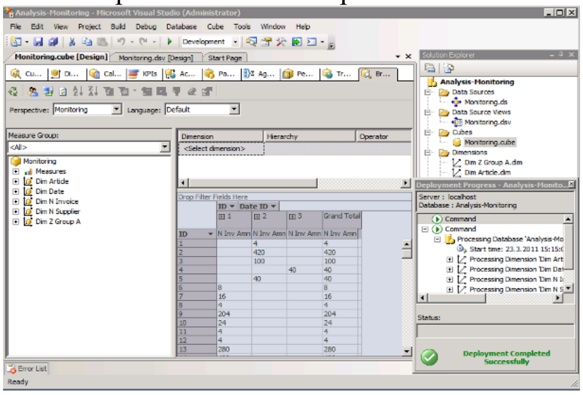
Third step is Data Source View setting.



Creating project 'Analysis-Monitoring.dwproj'... project creation successful.

Fig. 6. Setting of Data Source View

The last step is Data cube composition.



Perspective: Monitoring Language: Default

	ID ▾ Date ID ▾			
	⊞ 1	⊞ 2	⊞ 3	Grand Total
ID ▾	N Inv Amn	N Inv Amn	N Inv Amn	N Inv Amn
1		4		4
2		420		420
3		100		100
4			40	40
5		40		40
6	8			8
7	16			16
8	4			4
9	204			204
10	24			24
11	4			4
12	4			4
13	280			280

Server : localhost
Database : Analysis-Monitoring

Processing Database 'Analysis-Mo
Start time: 23.3.2011 15:15:(
Processing Dimension 'Dim Art
Processing Dimension 'Dim Dat
Processing Dimension 'Dim N I
Processing Dimension 'Dim N S

Deployment Completed Successfully

Fig. 7. Data Cube definition

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

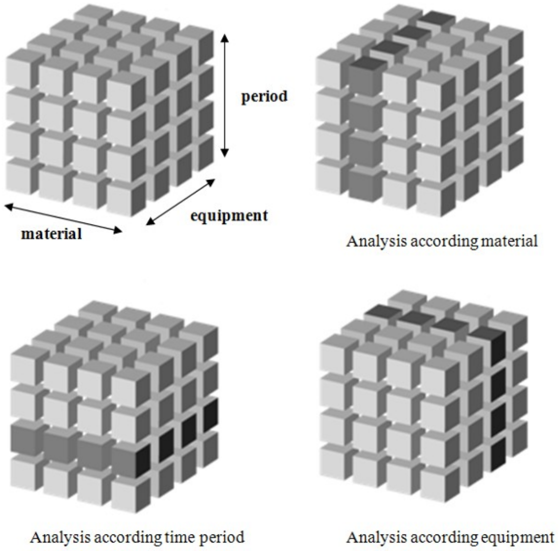

period
material
equipment
Analysis according material
Analysis according time period
Analysis according equipment

Fig. 8. Data analysis by the help of data cube

4 Conclusion

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.

Acknowledgments

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*. 2005. ISBN 0-321-32136-7

[4] G. Shmueli, N. R. Patel, P. C. Bruce, *Data Mining for Business Intelligence*. 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>.