Process of transformation, storage and data analysis for data mart enlargement

Zdenka Prokopova, Petr Silhavy and Radek Silhavy

Tomas Bata University in Zlin,
nám. T. G. Masaryka 5555,
760 01 Zlín, Czech Republic
phone: +420 57 603 5015; email: rsilhavy@fai.utb.cz.

Abstract-**Creation of the information system is the complex and long process but it is only the first step in its existence. Most of information systems go through a certain development during its life cycle. Very often the users have defined the requests for the enlargement of the functionality and the volume of the displayed data. The process of the transformation, storage and analysis for data mart enlargement based on the users' requests; concretely the data mart of the personal transport is presented in the paper. The expressions from the field of the Business Intelligence and the systems used for gaining data, data analysis or creation of forms and reports are explained.**

I. INTRODUCTION

Modern technologies infiltrate into almost all fields of the human activity. Data warehouse and data mart is a consolidated repository of structured data. Serving Business Intelligence (BI hereinafter) analytical tools that can draw data for analysis from data warehouses or data marts [1]. Data marts are thematically oriented data warehouse designed to mediate information to a department of the organization or geographic location. The difference between data warehouse and data marketplace is determined on the basis of occupied disk space, data warehouse takes order of magnitude larger disk capacity than the data mart.

At the present time Business Intelligence is one of the branches under the most rapid development. The underlay provided by it serves to the support of the objective and qualified decision-making. That is the reason why it has been applied in many different establishments [2], [3]. The BI under data warehouse provides two ways of displaying data. The first way is to view reports using the Report Server, which is part of the MS SQL server. Reports are intended for ordinary users, without detailed knowledge of data and data structures. These reports are output reports prepared by the programmer in the authoring environment. The second way is Online Analytical Processing (OLAP hereinafter) data analysis. This method of displaying information requires the detailed knowledge of data and data structures. For both ways it is necessary to create a relational and analytical database before the presentation of information as in [4], [5].

In the article we will describe data warehouses of unnamed company, which includes data warehouse Economy, Personal transport data mart and Prints data mart. A data warehouse is built on Microsoft technologies, specifically Microsoft SQL Server 2008 Standard [6]. Personal transport data mart is provided for users to view information about their underlying business trips, but also for detailed analysis of data. Data mart provides information to employees based on their inclusion in hierarchical organizational structure, i.e. managers can also view information about the business trips of their subordinates. Information's from the data mart are used to support approval process for business trips and allocation of official vehicles.

II. BUSINESS INTELLIGENCE

Business Intelligence is a set of concepts and methodologies to improve decision making process using metrics or metrics based systems. The purpose of the process is to convert large volumes of data to knowledge that are needed for the end user. This knowledge can then be effectively used for example in the decision process and can form a very important competitive advantage.

A. The main functionalities of BI

- Data analysis
- Reports – list of tables and graphs
- Dashboards - interactive panels
- Data Mining - mining in data
- Corporate Performance Management (CPM)
- Predictive analysis

B. Components of BI

- Data Marts - DMA
- Data Staging Area - DSA
- Operational Data Store – ODS
- Enterprise Application Integration – EAI
- Data Mining

- Tools for metadata managing
- Tools for data quality insurance

III. DATA TRANSFORMATIONS

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 technologically heterogeneous data sources and then choose relevant data and centralize, integrate and aggregate them each other. This process is sometimes referred to as the data pump. Data pumps serves to collection and transmission of data from source systems to data stores and dumping ground.

They include:
- ETL process for extraction, transformation and transmission of data,
- EAI process for application integration (works in contrast to ETL tools in real-time).

A. ETL process

ETL process starts by data extraction from primary sources (extraction). During this phase there are seek out and remove various data inconsistency. After the extraction follows data transformation (transformation) which will convert data obtained from single data sources into unified data model. 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. By reason that ETL process 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 process

ETL is an abbreviation of the extraction, application, and 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.

IV. DATA STORES

The philosophy of data store (warehouse) has published for the first time by Bill Inmon in [8] (first edition was published in 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 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 types of basic data stores and two types of auxiliary stores as in [8], [9].

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.
– 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.

B. Auxiliary data stores

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

V. ONLINE ANALYTICAL PROCESSING

OLAP is method, which works with multidimensional databases that include data structures and analytical tools designed for analysis of large amounts of data. The result of analysis and aggregation of data is usually multidimensional data structure, called an OLAP cube. The basic building blocks for creating OLAP cubes are the facts and dimensional data as in [10], [11].

A. Fact data

Fact data are numerical values of quantities; these quantities can be primary or may be calculated from other fact data. The fact tables are stored and analyzed fact data. These are values that we are monitoring and using for analytical calculations, sorting and aggregations. In connection with the fact data is using the term granularity, which determines the depth of the data details stored in fact tables. Lower granularity allows more detailed analysis, but increasing amount of stored data. There are much more data in fact tables than in the dimensional tables.

B. Dimensional data

Dimensional data are information's about the values stored in fact tables. They contain information about the hierarchical, logical and organization data structure. Using the dimensional tables we usually ask questions when, where, why and etc. In comparison with the volume of fact data tables these are usually much smaller. There are often used hierarchical (tree structure) registration data as a dimensional data. A typical example of such data is date range or geographic dimension.

VI. APPROVAL PROCESSES

A. Request for a business trip

Analysis of the requirements for passenger transport system functionality required identifying and describing the approval process for a business trip request. The main requirement was to fill a minimum amount of data to application and the maximum amount of data take up from the other systems. The flowchart of such approval process is shown on Fig.1.
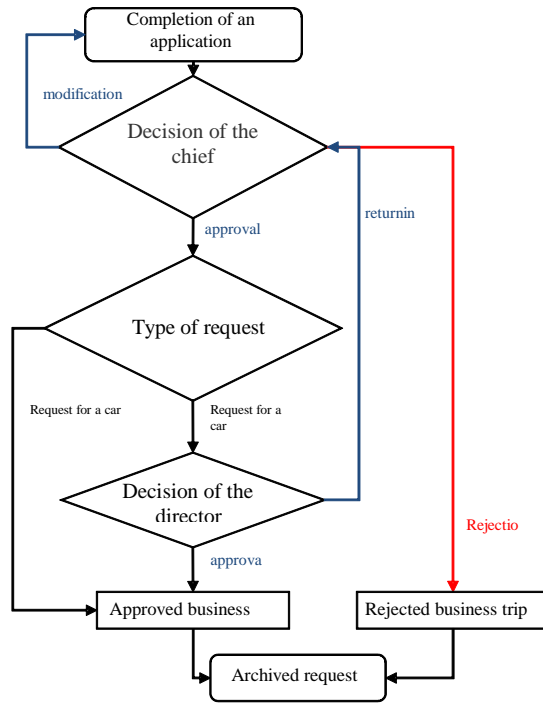

Fig. 1 The approval process of request for business trip

The approval process applications for a business trip can be done repeatedly.

B. Business trip statement

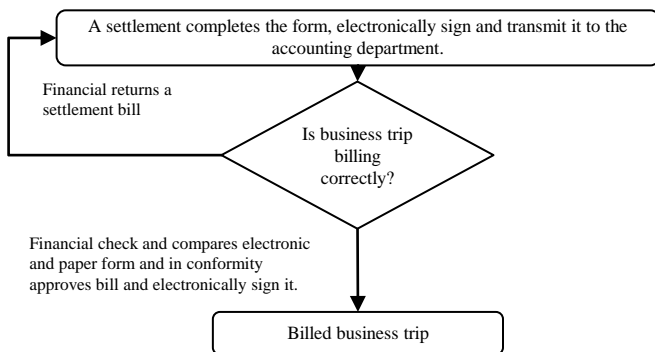The business trip billing is possible only on the basis of an approved request form.

Fig. 2 The approval process of business trip statement

It can be possible to charge several business trips in one statement. Compensations for travel expenses are calculated based on the entered data. Flowchart of business trips statement is shown on Fig.2.

VII. REQUIREMENTS FOR DATA MARKET FUNCTIONALITY

Data mart is intended for all staff within the organization. Information access is hierarchically structured in terms of organizational structure. There are available only limited information to ordinary workers, to managers are available extended information and data of their subordinates. To defined group of users are available all information [12]-[14]. Requirements of the original data mart:

- The data will be periodically transmitted from the data store to BI at least once a day.
- Will be created a logical and physical data model of data marketplace in accordance with established rules.
- Will be created ETL processes of data marketplace in accordance with established rules.
- Will be created multidimensional OLAP cubes in accordance with established rules.
- Will be created reports in accordance with established rules.
- It will set the security policy and security documentation will be updated.

A. Personal transport data mart ETL process

ETL process includes the extraction, transformation and storage. There are often problems with:

- Violation of referential and domain integrity
- Duplication of data with very poor identification of possible duplicates
- Absence of dials, or their misuse.
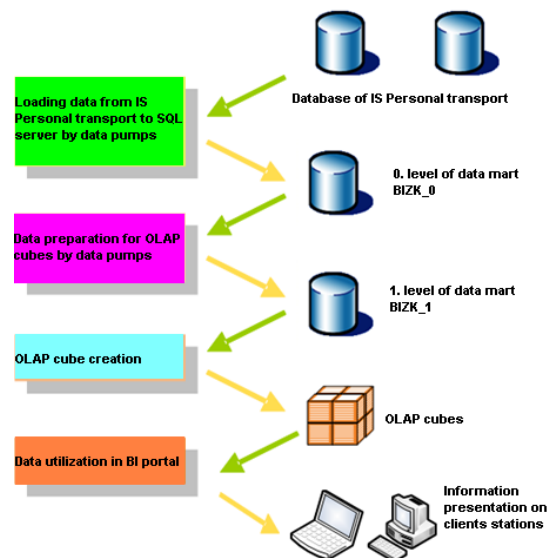- Random errors
- The problem of varying dimensions.

Fig. 3 Principle of personal transport data mart ETL process

The main task of the ETL process is to select the necessary data, clean them and ensure data integrity and localization in time. When creating a data warehouse relational database it is appropriate to use the basic principles:

- First load data from source information systems to 0th level data warehouse database.
- Then do the cleaning, transformation and loading to 1st level data warehouse database.

ETL process of personal transport is shown in figure Fig.3. The first level is used for multidimensional data model within there are designed OLAP data cubes for data warehouse and data marts. Data from this level can also be used for reports. Data from the Personal transport information system will be regularly pumped into the data warehouse daily.

1) Zero-th level data warehouse database

The main task of this level is to copy data from source systems into the data warehouse database. It means a few steps:

- Creation of copied tables in the data warehouse database
- Conversion of data types when copying (Conversion of data types is necessary especially in case of copying data from database systems of different manufacturers, or file types csv, xls, txt and others.)

These steps are created in the tool called Intelligence Business Development studio. So-called SSIS packages are used for ETL processes creation. When the ETL process is creating the whole process should be divided into several packages that are logically connected, due to clarity and starting times.
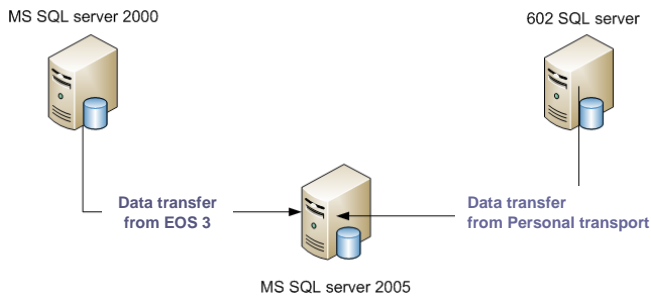

Fig. 4 Schema of database server's connections and copying the source data to 0-th level data warehouse database

2) 1-st level data warehouse database

The main task of the first-level data warehouse database is to create a database for OLAP cubes and report server. It creates suitable group of records from the previous level at this level in the desired granularity. Granularity is word for particularity of observed records. Granularity in presented data marked is at the level of person's trip, individual prices, etc. It is appropriate to create records in terms of dimensional and fact data at the same time.
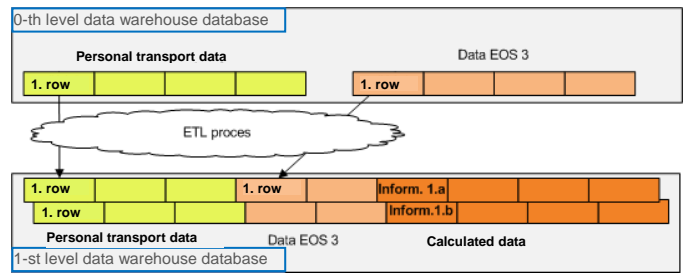

Fig. 5 Schema showing the extraction and transformation of data from level 0 to level 1

3) Extension of SQL query for data transfer

Another modification of ETL process is interlacing the columns of source tables to columns of target tables. It was needed to assign corresponding columns in the mapping setting. Assignment of source columns with the target columns is performed automatically in the case that they have the same name. This fact could help us in defining the data model. On the other hand, it could bring errors in automatic interconnection of the source and target columns with the same name, which contains other data.

4) The modified ETL process testing

Business Intelligence Development Studio (BIDS) allows testing SSIS package after creating or editing ETL process. Calling the context menu for the selected package can make it and using item Execute Package. BIDS is automatically switched to the Control Flow after confirmation of this option. On the Control Flow are displayed particular jobs. Jobs are color-coded according to the processing status:

- White background color - the job is not processed yet.
- Green color background - the job has been completed.
- Yellow background color - the job is processed now.

Automatic running of different packages (after creating and testing them) is ensured by means of two steps:

- Import package in Integration Services MS SQL.
- Create job in the Database Engine MS SQL server (determination of packages order and setting the time interval job's implementation).

Both steps are performed in SQL Server Management Studio, which allows connecting to several "Server type":

- Database Engine
- Analysis Services
- Report Services
- Integration Services.

B. Auto startup of ETL process on SQL Server

After creating and testing packages it needs to run them automatically. Two steps can achieve this:

- Import the package in Integration Services MS SQL.
- Create job in Database Engine MS SQL server, determinate the running order of packages and set the time interval of job implementation.
- Both steps are done in SQL Server Management Studio, which allows connection to several server types.

C. Import of packages to Integration Services MS SQL

Import of packages is done in Integration Services. In the Stored Packages folder are displayed already stored packages. With the context menu "Import package" new package is imported. It must be completed:
- Package location:
  - SQL Server
  - File System
  - SSIS Package Store
- Server – makes sense only with selection „SQL Server" or SSIS Package Store
- Authentication:
  - Windows Authentication
  - SQL Server Authentication
- Package Name

D. OLAP cubes

We dealt with ETL process of creation a relational database in zero and first levels in the previous chapter. We can design and create a multidimensional database which draws data from the first-level relational database after this step.

Two above-mentioned kinds of databases differ in a way of data storage. We are trying to create the highest normal form when creating a relational database, but in a multidimensional database we mainly use non-normalized tables. OLAP cube always contains facts and dimensional attributes.

Creating of OLAP cubes is done in MS Business Intelligence Development Studio, which has already been used to create and modify the ETP process. OLAP cube is created in the following order:
- Data Sources
- Data Source Views
- Cubes
- Dimension

OLAP cubes are attached as a data source of BI portal where are stored basic views.

VIII. REPORTING SERVICES

The role of reporting services is to provide in an appropriate form and in time data for decision support at all levels of organizational infrastructure. The main reason for the deployment of reporting services is generating outputs in electronic or paper form. It allows employees at all positions effective access to data and thus supports them in their activities.

Reports can be divided according to the criteria of "control - adaptation reports" on two variants:
- Static - report shows the preset information displayed content cannot control. These reports are different from hard copy.
- Interactive - reports show the information that you can adapt by using of special elements (filters) to meet current need.
- Another criterion of the distribution is "field of philosophy and reports deployment":
- Enterprise – present data in business informatics.

- Embedded – report generation is an integral part of the application.
- B2B (Business To Business) – report generation for business partners.
- Hierarchy of staff reporting services distribution [5]:
- Consumers of information - the largest group that receives mainly extracts data in the form of simple reports.
- Workers actively working with the information - are able to analyze and process data. In this group of reporting tools are used and to some extent, analytical tools.
- Analysts - the smallest group of users, but in terms of expertise at the highest level. Their main task is to define new areas of requirements analysis, to analyze existing data support the interpretation of analyzed data used in decision-making process. Based on current trends, they are able to predict trends for the future. They use analytical tool.

It is appropriate to define the life cycle of the report when creating it. Life cycle of report consists of three basic phases:
- Draft report - it can be divided into other sub-phases:
  - Collection and evaluation of requirements,
  - Creation of reports in BI Development Studio (result is an XML file in the Report Definition Language - RLD). At this stage it is necessary to define:
  - Data Source,
  - Design of query,
  - Design of graphic look.
- Report management.
- Repot delivery - the method of delivery and format of the report is solving in this phase. The user can deliver reports on demand, or according scheduled tasks sent by email.

A. New report creation

Based on the monitoring of costs arose request to create a new report. This should display information about the compensation paid for the organizational units in individual years. Report should allow subdivision users in organizational units and months in selected years. Reports were created in BIDS development environment using the wizard.

In the first step of the wizard, we must select a data source (Data Source). In the next step (the Query Design) is necessary to define the SQL query. It is possible to write directly Query string entries, or use the Query Builder.

On the next screen (Select the Report Type) we can select the type of report:
- Tabular - classic display
- Matrix - matrix display.

After selecting the type of report we need to define the other items in the next screen (the Design Matrix):
- Page
- Columns
- Rows
- Details.

It is necessary to define the order of multiple items.

The last step is selection of reports style visualization. Development Studio allows to view and test the created report before placing in the report server. Testing is done in the "Preview" tool.

Each report displays a menu that allows:
- Paging report
- Increase or decrease the displayed report
- Search
- Data export.

B. Reports for employees

The first sample report is the report Travel orders for employees showing of all information (years, places, accounts etc.) obtained by clicking on "Extract All" as you can see on the figure Fig. 6.

Fig. 6 Sample report – Travel orders for employees

C. Reports for compensation paid

Example of sample report aggregate amounts paid by organization from 2008 to 2010 for compensation paid is shown on the figure Fig. 10.

Fig. 10 Sample report aggregate amounts paid from 2008 to 2010

IX. CONCLUSION

The experience with the application of the Business Intelligence and launching the data warehouses and data marts confirm the big benefit for the processes of the decision-making. Data market through user-friendly interface, opening up important and necessary information to all employees. Most important task was to provide managers with a tool for managerial decision-making and show further possibilities of this technology.

Additional requirements to extend the functionality of the data market, as well as passenger information system appeared during the work. The most important of them is the request to create a new data market of the organization structure that would include the information on the users, functional locations and their inclusion into the organization structure. These data markets should serve not only to managers as the basis to make decisions about these areas, but also to all subjects providing the data.

REFERENCES

[1] M. Golfarelli, S. Rizzi, *Data Warehouse Design: Modern Principles and Methodologies*. McGrow-Hill, 2009. ISBN: 978-0-07-161039-1.

[2] D. Power, DSSResources.com [online]. 2005 [cit. 2012-06-07]. "What is business intelligence?" From WWW: <http://dssresources.com/faq/index.php?action=artikel &id=4>.

[3] H. P. Luhn, "A Business Intelligence Systems". *IBM Journal of Research and Development*, 1958, pp. 314-319.

[4] M. Berthold, D. Hand, *Intelligent Data Analysis*. Springer, Berlin, 2007. ISBN: 978-3-540-4306-5.

[5] J. Joe Celko, *Celko's Analytics and OLAP in SQL*. Elsevier Inc. 2006. ISBN: 978-0-12-369512-3.

[6] B. Larson, *Delivering Business Intelligence with Microsoft SQL Server 2008*, The McGraw-Hill Companies, 2009. ISBN: 978-0-07-154944-8.

[7] J. Singh, "Understanding ETL and Data Warehousing: Issues", *Challenges and Importance: Role of ETL routines in Quality Data Warehouse*. Lap Lambert, 2011. ISBN: 978-3843390934.

[8] W. H. Inmon, *Building the Data Warehouse*. Wiley, 2002, ISBN: 978-0471081302.

[9] 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.

[10] E. Thomsen, *OLAP Solutions: Building Multidimensional Information Systems*. John Wiley&Sons, Inc., 2002. ISBN: 0-471-40030-0.

[11] Z. Prokopová, P. Šilhavý and R. Šilhavý, "Preview of methods and tools for operating data analysis", *International Journal of Mathematical Models and Methods in Applied Science*, vol. 5, no. 6, 2011, pp. 1102-1109. ISSN 1998-0140.

[12] I. Lungu and A. Mihalache, "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.

[14] Z. Prokopova, P. Silhavy and R. Silhavy, "Data transfer, storage and analysis for data mart enlargement", *Advances in Mathematical and Computational Methods*, 2012, pp. 225–230. ISBN: 978-1-61804-117-3.

[15] P. Kopecky, *Data mart enlargement - data transfer, storage and analysis*. Diploma thesis. Tomas Bata University in Zlín, Faculty of Applied Informatics. 2011.