2

Description of the Data Market Enlargement Process

by

Zdenka **Prokopova**[1]

The aim of the presented paper is to familiarize the reader with the utilization and transformation of data and their storing into the relational database, the subsequent creation of analytical databases and description of used analytical tools. The main benefit of the work is description and demonstration of the data market creation and enlargement process.

Keywords:

Business Intelligence, Data Warehouse, Data Mart, ETL process, OLAP, information system.

[1] 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 E-mail: prokopova@fai.utb.cz

Introduction

Modern technologies infiltrate into almost all fields of the human activity. At the present time Business Intelligence (BI hereinafter) 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.

Business Intelligence involves a process of transformation of data from information systems to information. Access to such accurate and consolidated information allows users to change then their way of intuitive decision-making on an objective and qualified. BI includes issues of data warehousing, data marts and data mining. All this can be understood as not only the technology but also as a set of skills and methodologies for data processing and evaluation. At present the BI is one of the fastest growing industries of information technology [7].

Data warehouse and data mart is a consolidated repository of structured data. Serving Business Intelligence analytical tools that can draw data for analysis from data warehouses or data marts. 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. Another possible view, which specifies the difference between data warehouse and data marketplace is linked to the data source. Data mart usually has a single data source, while the data warehouse mainly draws data from multiple sources [3].

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 [5]. 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. 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 OLAP data analysis. This method of displaying information requires the user a 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 [11]. This is for the ETL process, sometimes also referred to as the Data Pump.

Through the information system Personal transport asking employees using electronic forms about a business travel and after their implementation perform billing business travel. Other functionality is the allocation of official vehicles to applicants who performs head of trucking. The approval process of applications for business travel and business travel bill is supported by an electronic signature that ensures the credibility of data.

Business intelligence

Business Intelligence is an integral part of modern strategic and operational management of companies, state and local governments. The main task of BI is to support decision making at all hierarchical levels. BI offers not only current information, but also look at historical data that allow modeling outlook for the future. BI is the process of transforming data into information [6].

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 [8].

BI is able to analyze data and help answer questions:

- What is the development of the observed problem, what happened?
- What is the main cause?
- What you need to do or change?
- What are better ways of solving?
- What are the prerequisites for the future?

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

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

ETL process

ETL is an abbreviation of the various stages of pumping data Extract, Transform and Load, also known as a data pump. The main task of the ETL process is obtained from the data source information systems thematically structured data suitable for analytical purposes. The majority of information systems are constructed in terms of the transaction data storage. It follows that the function of the ETL process is not just copying structures, but also the choice and transformation of data [10].

Data Extraction

From the information systems source data built on transactional databases is necessary to choose - extract data suitable for analytical purposes. When creating data marts there are mostly used multiple data sources such as relational databases, text files, XML files, CSV files and others.

Data transformation

Data transformation is mainly a process of verification, data cleaning and their time localization. In the phase of data transforming it applies a number of rules and functions to transfer data from source information systems to analytical data.

- During this process is done:
- Selection of appropriate columns
- Typecasting variables
- Calculation of new values
- Filtration and Assortment

- Combining multiple columns to one
- Division of one column into multiple columns
- Transposition - convert multiple columns on multiple lines
- Aggregation
- Verification of relevant data

Data loading

The final phase of the ETL process is usually store data in a data warehouse. Depending on the data warehouse, this phase may vary.

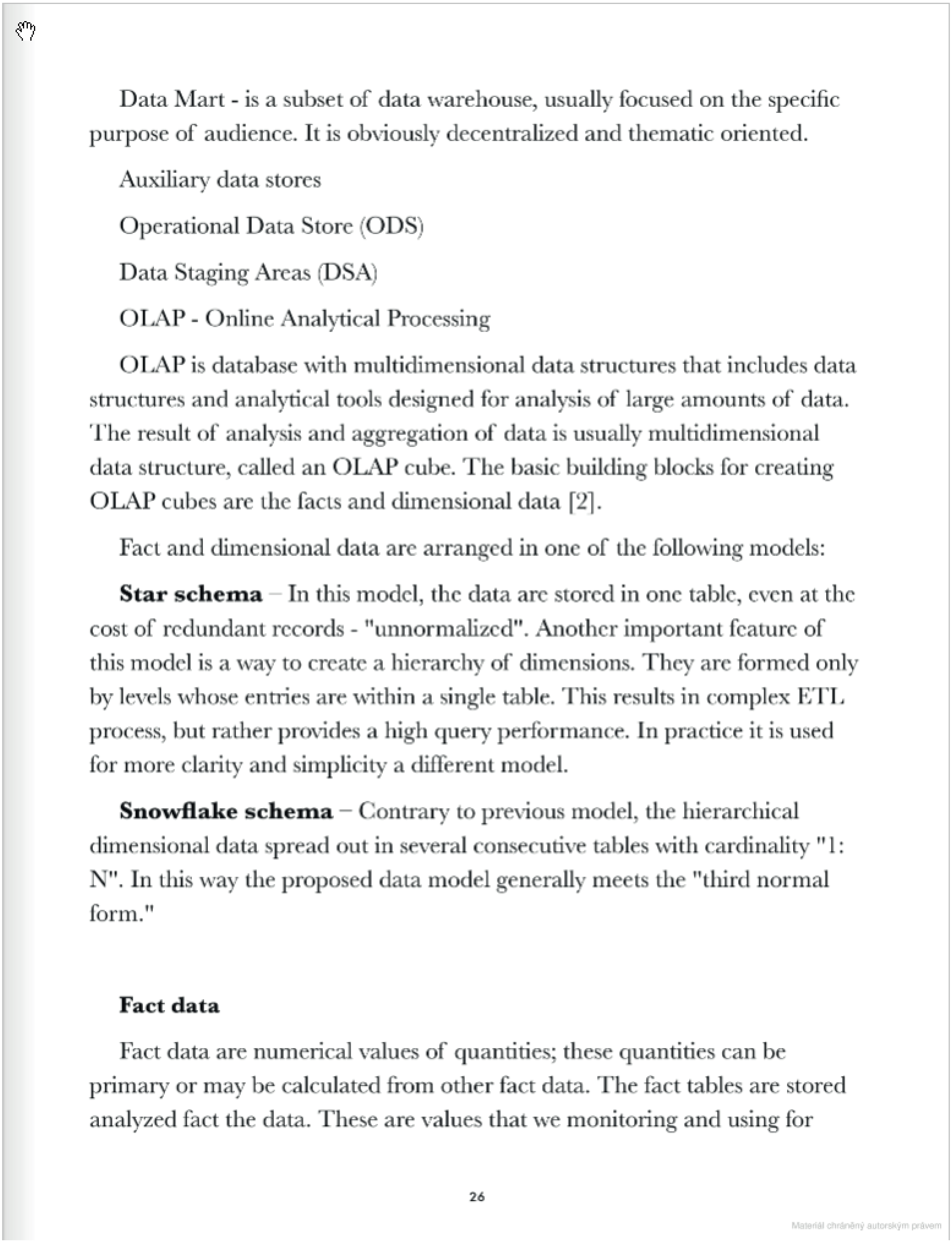
The dual approach to writing data into DWH:

- Incremental - data are added to the DWH in the form of historicized
- Overwrite all data - data are always overwritten, it can be done at certain intervals, such as day, month, year, etc.

Data warehouse

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 [3].

Data warehouse (DWH) is a special type of relational database that allows to solve problems focused mostly on analytical querying over large datasets. DWH is a central consolidated data repository that draws information from information systems. This information creates a regular pumping time series. The DWH database often does not in normal form, but adapts to the data structure for the best and fastest analytical inquiry. Such data storage often uses OLAP.



Data Mart - is a subset of data warehouse, usually focused on the specific purpose of audience. It is obviously decentralized and thematic oriented.

Auxiliary data stores

Operational Data Store (ODS)

Data Staging Areas (DSA)

OLAP - Online Analytical Processing

OLAP is database with multidimensional data structures that includes 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 [2].

Fact and dimensional data are arranged in one of the following models:

Star schema – In this model, the data are stored in one table, even at the cost of redundant records - "unnormalized". Another important feature of this model is a way to create a hierarchy of dimensions. They are formed only by levels whose entries are within a single table. This results in complex ETL process, but rather provides a high query performance. In practice it is used for more clarity and simplicity a different model.

Snowflake schema – Contrary to previous model, the hierarchical dimensional data spread out in several consecutive tables with cardinality "1: N". In this way the proposed data model generally meets the "third normal form."

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 analyzed fact the data. These are values that we monitoring and using for analytical calculations, sorting and aggregation. 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. In the fact tables is much more data than in the dimensional tables.

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. The dimensional data is often used hierarchical - tree structure of registration data. A typical example of such data is date range or geographic dimension.

Example of date range dimension:

- year
- quarter
- month
- week
- day

Example of geographic dimension:

- State
- Region
- District
- Village

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. To ordinary workers are available only their information, to managers are available their information and data of their subordinates. To defined group of users are available all information.

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.

ETL process of Personal transport data mart

ETL process includes the extraction, transformation and storage. At its creation 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.

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.

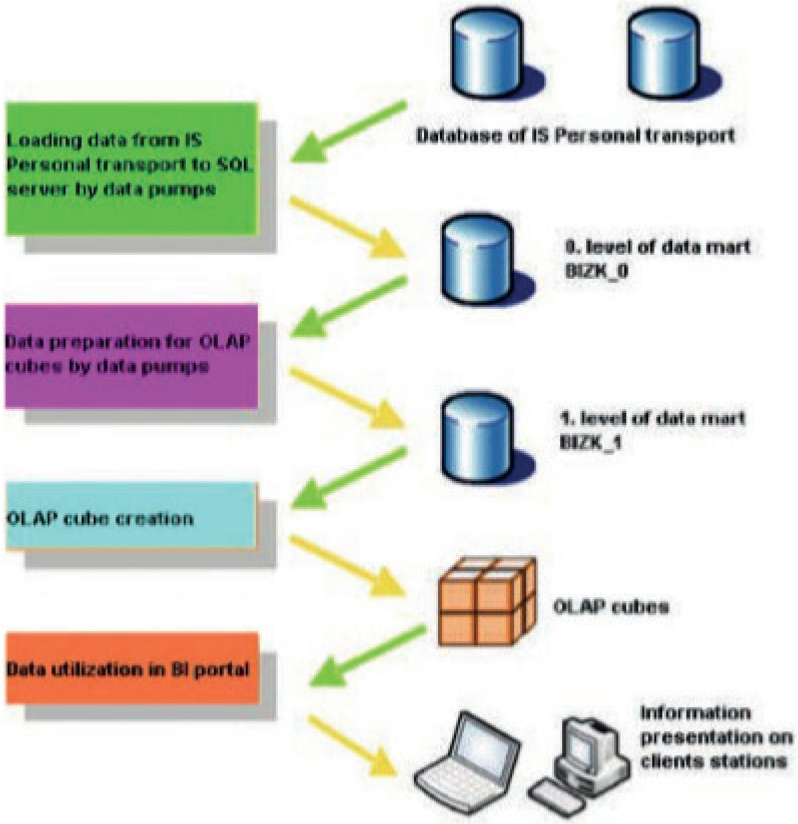

Loading data from IS Personal transport to SQL server by data pumps
Database of IS Personal transport
0. level of data mart BIZK_0
Data preparation for OLAP cubes by data pumps
1. level of data mart BIZK_1
OLAP cube creation
OLAP cubes
Data utilization in BI portal
Information presentation on clients stations

Fig. 1 Principle of personal transport data mart ETL process

ETL process of personal transport is shown in figure Fig.1. 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.

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.

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.

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 m odel. 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.

The modified ETL process testing

Business Intelligence Development Studio (BIDS) allows testing SSIS package after creating or editing ETL process. It can be made by calling the context menu for the selected package 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.

Auto startup of ETL process on SQL Server

After creating and testing the packages it needs to run them automatically. This can be achieved by two steps:

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: Database Engine, Analysis Services, Report Services, Integration Services etc.

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

- OLAP cubes

There are two OLAP cubes created in the analytical data warehouse:

- Cube_cares – used to data analyses in terms of utilization of cars, roads goals etc.

- Cube_travels – used to data analyses in terms of utilization of approved business travel, paid expenses and optimizing the of business trips approval.

OLAP cube always contains facts and dimensional attributes.

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 can not 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:

- 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 tools [5].

Reports for employees

Examples of reports created for employees you can see on the following figures. The first one is the sample report Travel orders for employees showing of all information obtained by clicking on "Extract All".

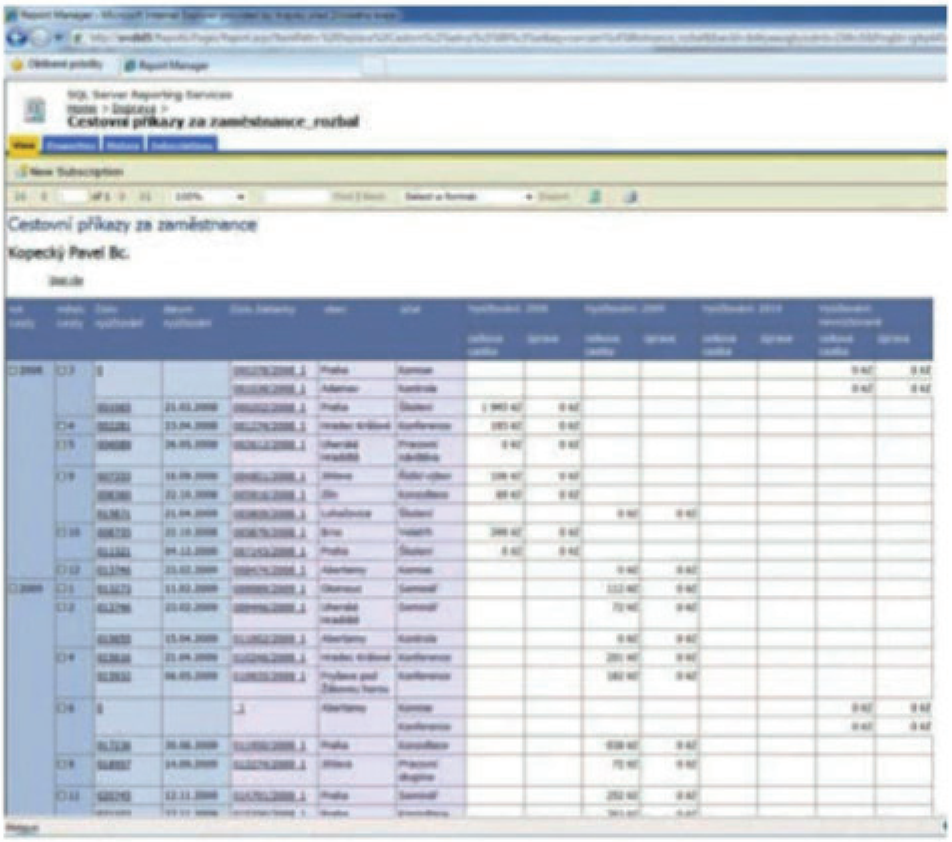
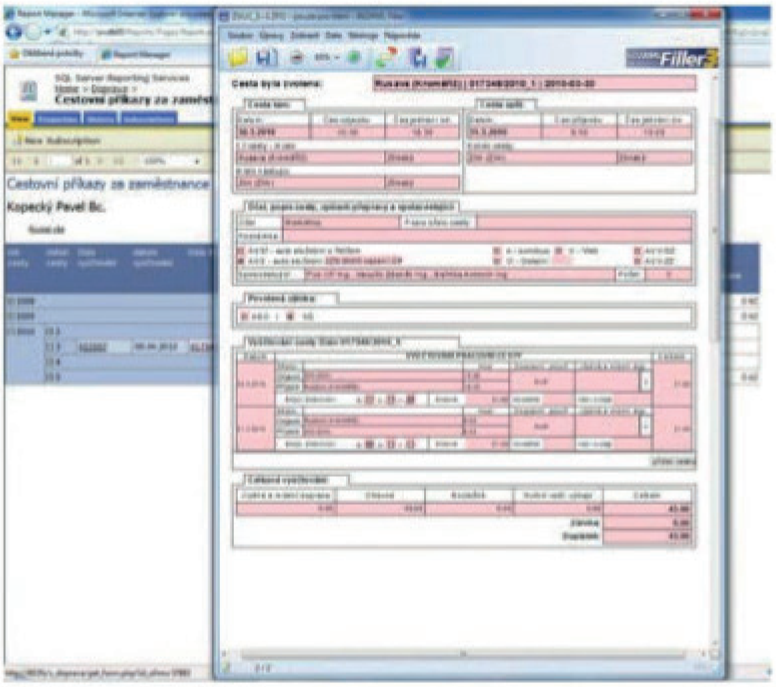

Report Manager
SQL Server Reporting Services
Cestovní příkazy za zaměstnance_rozbal
New Subscription
Select a format
Cestovní příkazy za zaměstnance
Kopecký Pavel Bc.

Fig. 2 Sample report – Travel orders for employees

The next one is the business trip form shown in the figure Fig. 3. It can be viewed by clicking on an active link with Applications for business trip or Billing business trip.



SQL Server Reporting Services

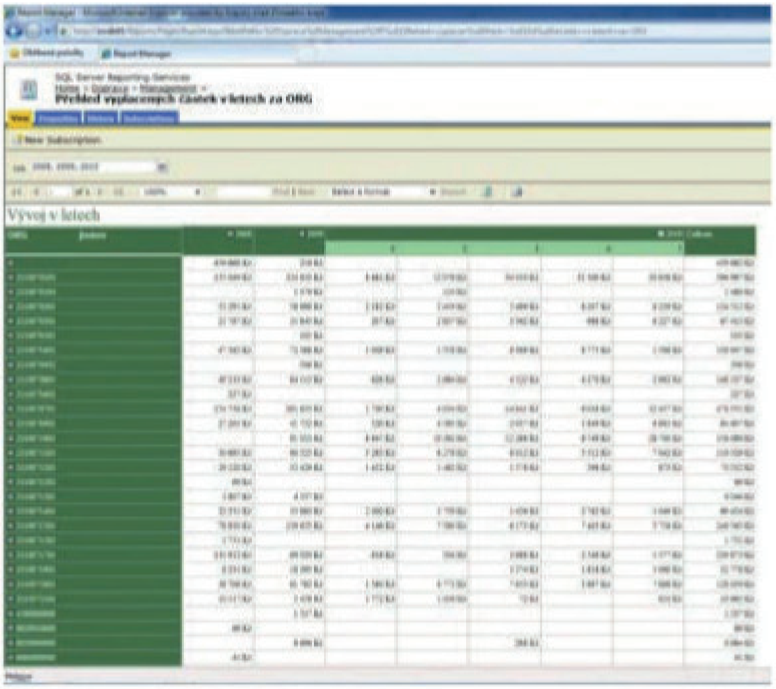
Cestovní příkazy za zaměstnance

Kopecký Pavel Bc.

New Subscription

Filler

Fig. 3 Business trip form



SQL Server Reporting Services

Přehled vyplacených částek v letech za ORG

New Subscription

Vývoj v letech

Fig. 4 Sample report aggregate amounts paid by organization from 2008 to 2010

Reports for compensation paid

Example of sample report aggregate amounts paid by organization from 2008 to 2010 for compensation paid is shown the figure Fig. 4.

Conclusion

Every information system goes through a certain development during its life cycle. Very often the requests for the enlargement of the functionality and the volume of the displayed data have been defined by the users. In the presented paper expressions from the field of the Business Intelligence are described. Here are explained systems used for gaining data, enlargement of data marts or creation of forms and reports.

During the work, additional requirements to extend the functionality of the data market, as well as passenger information system appeared. 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. Data market through user-friendly interface, opening up important and necessary information to all employees. It provides managers with a tool for managerial decision-making and showed further possibilities of this technology.

References

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

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

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

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

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

[6] Luhn, H. P. A Business Intelligence System. IBM Journal of Research and Development, pp. 314-319, 1958.

[7] Power, Dan. DSSResources.com [online]. 2005 [cit. 2010-06-07]. What is business intelligence? From WWW: <http://dssresources.com/faq/index.php?action=artikel &id=4>.

[8] Power, Dan. Dssresources.com [online]. 2007 [cit. 2010-06-07]. A Brief History of Decision Support Systems. From WWW: <http://dssresources.com/history/ dsshistory.html>.

[9] Shmueli, G., Patel, N. R., Bruce P. C. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with Xlminer. 2006. ISBN: 0-470-08485-5.

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

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