

# Information - Analytic System for Problem Solving of Socioeconomic Monitoring

A. Chadyuk. G. Borodkin. G. Proskudina  
Institute of Software Systems of National Academy of Sciences of Ukraine  
40 Academician Glushkov Avenue, 03187, Kiev, Ukraine  
E-mail: [georgeb@miratech.com.ua](mailto:georgeb@miratech.com.ua)

**Abstract:** The experience of creation and research of information - analytic system is discussed. The system is oriented towards a solution of socioeconomic problems. The report describes “Monitor 2.1” system, which have been developed for The Ministry of Economy of Autonomous Republic of Crimea. The report contents the principles and approaches to the development of the next version of that system – “Monitor 3.0”. The concept of data warehouses is used for implementation of the last one. The needs for the solving of socioeconomic problems to integrate and accommodate large collection of socioeconomic data are discussed. The report contents a comparison between “Monitor 2.1” and “Monitor 3.0” systems. It shows the advantages of the “Monitor 3.0” system development approach. Papers summarize the last year’s author’s experience in the development of information - analytic systems for socioeconomic problems.

## 1 Introduction

Long-term reforming of economy caused sudden fall of a living standard of the considerable part of the population of Ukraine. Such situation led to the considerable growth of requirements for government administration bodies of different levels in scientifically reasonable decisions on principal directions of social and economic policy. Complexity of administrative solutions, demands to their quality, effectiveness and validity require analytical processing of large volumes of information, which are miscellaneous on their nature and arrive from many sources.

It is clear, that the more information is involved in process of decision-making, the more reasonable solution can be accepted. But here is a problem of use of large amount of accumulated information. For collection and storage of this information many efforts and resources are expended. Nevertheless, they, for whom it is most necessary - chiefs and decision-makers, can't use large part of this information. More often such information is accessible only those subdivisions, where it is collected and treated. Therefore, there is the need for special tools that will automate the decision support on reforming in a sphere of socioeconomic life of Ukrainian regions. Such tool can appear as the computer monitoring system of the living standard analysis of the population in Ukraine (system "Monitor"). This paper summarizes our study and development activities during last several years and shares our experiences.

## 2 Purpose and conditions of usage

The basic idea of the development of the “Monitor” system is creation of the high-performance high-speed computer system for the analysis, simulation and forecasting of society state, both in country as a whole and in its separate regions. It is based on a multidimensional model based on aggregated, historical and prediction data.

The activity of the executive and legislative bodies of different levels of an authority is directed at development, introduction into life and monitoring of the fulfillment of different legislative acts, including sphere of social protection. The living standard of the population of each separate region depends on its economic, climatic, demographic and diverse features. Therefore, government bodies should have initial information on a living standard of separate groups of the population in regions of Ukraine. Regional informational statistical centers should gather such information.

The feature of the initial information gathering is necessity of monitoring of a plenty of socioeconomic indicators in a profiles of regions, temporary periods, social groups etc. Besides the data arrive from miscellaneous information sources - statistical data, data of sociological inquiries, expert estimations and other. It causes necessity of an integration and organization of procedures of joint analytical processing of input information.

Thus, there is a problem of a storage and processing not only large volumes of miscellaneous factual information, but also information, which one has as multidimensional structure and different levels of aggregation. Really, for the analysis of a socioeconomic situation only in Autonomous Republic of Crimea, overseeing by 300 basic services, foodstuffs and manufactured goods, under inspection of 25 regions and 9 cities of republican subordination, at watching a price level 1 time per one week, the data volume of parameters of a price level on the goods and services will reach more than 1 million 100 entries per year.

The problems are solved by “Monitor” system ensure the account of estimation parameters of living standard of the population of regions, evaluate the tendencies of their dynamic modifications, realization of analytical matching of these parameters with itself and with normative. Using of the economic-mathematical models, the system allows executing forecasting of modification of the observable parameters. The final results are generated in form of the state and prognoses tables, maps of parameter’s state, estimation diagrams, parameter distribution frames, schedules of the trends of parameter modifications. All these analytical reports come to analytical control services.

After study of the obtained outcomes the governmental bodies work out solutions in a form of normative documents, which are directed on improving of a situation on places. The norms and specifications are installed by these normative statements, are entered in the normative-legal database through the system of guide shaping and are used for further problem solving by the system.

The main purpose of the “Monitor” system was: 1) The definition of Socioeconomic Data Model and the migration/evolution of existing data; 2) The design and implementation of databases covering the main sections of the Data Model (price levels; incomes of the population; physiological norms of consumption, administration etc.); 3) The design and implementation of data access services; 4) The evaluation of a living standard of different social population.

### 3 The basic system requirements

There are some basic requirements to functional capabilities of software packages for data analysis.

Data processing includes: data input from the documents, journals etc.; data import from different formats; data exchange with other systems; adding the new data during the analysis; input data adjustment; “empty” data filling.

Data analysis includes: construction of distribution tables and schedules; data tables editing according to user requests; data search and analysis according to user inquiry; aggregate of primary data and designing of the secondary one; storage optimization of primary data and secondary one; estimation of statistical errors; graphics methods of data representation with a color scale.

Result reporting includes: export of the tables, schedules, other materials and analysis outcomes to MS Windows applications; graphical data representation by the way of maps with color keying according to a selected ranking scale; possibility for creating of simple reports; availability of sets of report’s templates.

Information search consists of data searching through all databases and possibility of data searching through Internet.

Analysis tasks are devoted to Primary tasks (creating of distribution tables, schedules and diagrams; trends search under the selected options; analyses of existing features used in the rules; designing of new indications on the basis of available ones; data aggregation and new data analysis) and the tasks based on the primary one (search and analysis of the rules, which explain the investigate phenomenon; analyses of factors, which explain the investigate phenomenon; forecasting on the basis of the rules and regularities; the analysis of available indications, which are included in rules; analytical of a decision-making support).

The value of any analytical system is determined by entirety of the basic economy-mathematics model and quality of its soft-hardware realization. The economic part of system development keeps behind frameworks of this report. We shall stop on software realization.

### 4 “Monitor 2.1” system

The “Monitor 2.1” system was realized on the base of Excel 5.0, working under the Windows 3.1. The software was developed with Visual Basic for Applications (VBA) - object-oriented programming language build in Excel. It allows to support and process several computing models and to execute all necessary operations, to keep track of by data-refresh, to represent them graphically and to create reports. Possibility of a multidimensional data analysis was obtained due to PivotTable mean.

Any analytical scheme consists of three mandatory blocks - input and preservation of data, data processing and its presentation. The structure of “Monitor 2.1” system is represented in the Fig. 1. The system includes the following function boxes: Input and editing of income data; Management of the “Monitor” database; Solution of functional tasks; Creation and preservation of task results; Management of the guides and normative documents. All this components are joined in user's system interface.

The information component includes the following databases: Database “Monitor”; Legislative database; Task execution report database; Processing database of secondary data; Guide database.

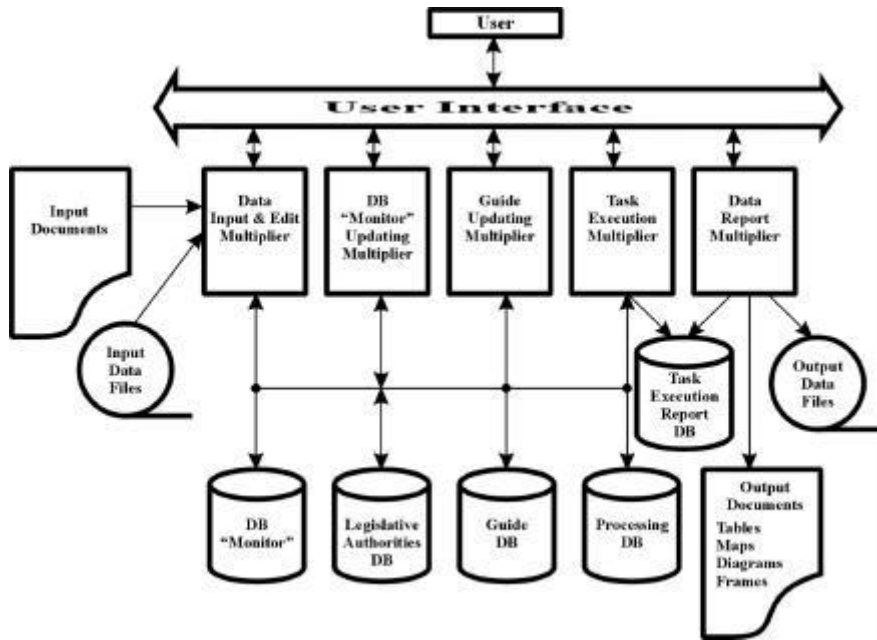


Fig. 1 - The skeleton diagram of the “Monitor 2.1” system

#### 4.1 Data input

Incoming data can be entered in the database “Monitor” in several ways. First way is through the initial forms or templates. Forms or templates represent electronic analogs of the paper documents, user familiar with. They are used for preserve a table built once in established format, with the form name, numbering etc.

Several kinds of the screen forms of input documents are stored in the system. Each respondent has its own group of such input forms. The forms are classified on groups of parameters (price, kind of trade, respondent etc.). The title of the form also is a full title of a parameter recorded in the system. All forms have the characteristics. They are: measurement unit of a recorded parameter, the table size, name of the relational database, in which these parameters will be saved.

The user selects the necessary form, inputs date and then parameters. The form with the parameters can be saved in the usual Excel format, and also through a command “Save in the database” in database format with an appropriate structure. The reference and normative data input is carried out in an approximately same way. These operations are executed through a component of Guide Input and Normative Document’s Input.

Other method of system data input is a usual copying of files with databases of parameters in the directory of system basic data. It is necessary to note, that the system

can work only with the files, the format which one it “remembers”. The formats of database’s files are saved in technological database and can be supplemented. The “Monitor” system can also work through a net using information resources of the remote computer. The DBMS core operates on remote computer, treating requests, executing incorporated operations, and returning outcome, in a form of data block, to client of “Monitor” system. The system acts as initiator of manipulations with the data, while DBMS core plays passive role of request service and data processing. Thus, the database “Monitor” can be formed as from databases, forming and filled in the system itself, and through other database files, which one arrive by data communication channels from miscellaneous sources, provided that the data format is known. We may say that the database “Monitor” is a virtual source, based on distributed databases.

## **4.2 Task execution**

The task execution subsystem gives the user a possibility to execute four types of tasks. The tasks have the universal nature and can be applied in a broad range of social-economy spheres. Tasks in a submenu arranged in such a manner that each subsequent task comprises determination parameters of the previous problems. The system allows carrying out the monitoring both observable parameters, as well as estimated ones. The task execution requires selecting some parameters. The system has user-friendly interface. A procedure of parameter selecting is identical to all tasks. The user determines such main specifications as Data domain; Respondent; Region; Observable and considered parameters; Accounting period. Automated selection of parameters required for tasks execution is carried out under tuning the system on one of the spheres of social-economy life (data domain). After a choice of all required parameters the task is executed. The procedure of account is executed stage by stage. First, an SQL-request is formed for primary databases. Through the ODBC interface the system provides access to databases of the definite format and their loading in a spreadsheet. Thus, the system reshapes primary parameter samplings. After this all necessary accounts are made. Finally, the system forms the account result in a shape of a table, which consists of two sheets. First sheet contains the list of task in a whole: name of data domain, task, titles of parameters, settlement dates, and regions. Other sheet contains all samplings and results of accounts. The system grants a possibility, behind an own reason of the user, to save outcomes of accounts in separate files, in the account’s directory.

## **4.3 Reporting**

The data samplings and account outcomes are structured in such a manner satisfying a PivotTable subsystem, providing further realization of the multidimensional analysis of the obtained data. Thus, we practically have received an OLAP-system [Go97][Lu98][Be98]. The system through the tools of table submission and EXCEL graphics possibilities allows presenting reports in the best way. The “Monitor” system offers a broad choice of the multidimensional tables, schedules, diagrams and maps (Fig. 2).

Through these means the target reports are formed. They can be saved in files or on paper carriers.

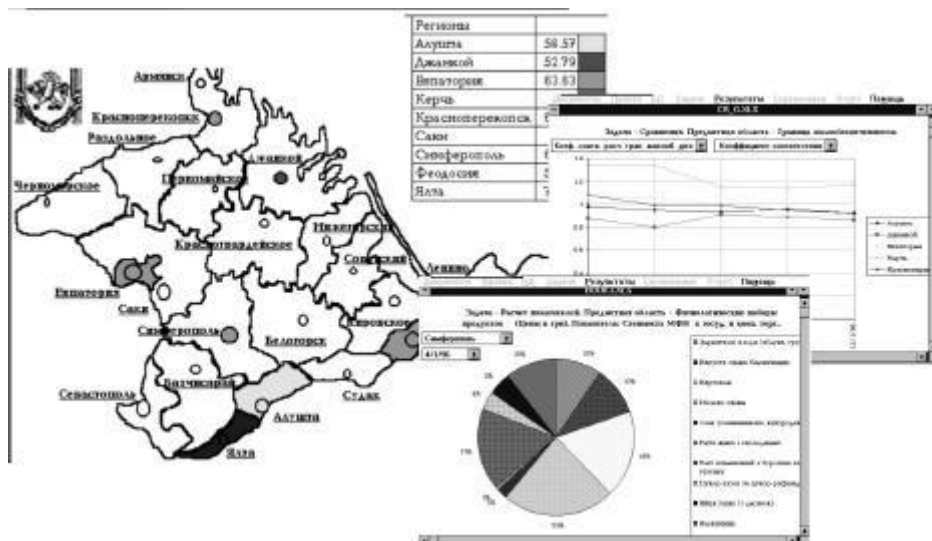


Fig. 2 Example of “Monitor 2.1” reports

#### 4.4 Imperfections

The system version (“Monitor 2.1”) was tested in the Ministry of Economy of Republic of Crimea. The developed system has showed a lot of restrictions. When a virtual source based on distributed databases is created, it induces difficulties [Sh98]. Each request at such source needs dynamically translation in requests at the initial databases. Obtained results require be agreeing, linking with one another, aggregating and returning to user. This approach showed a number of considerable shortfalls.

1. Time of request to distributed repository. Besides, the database structure, are significantly normalized, as a rule. Therefore analytical request to warehouse requires association of plenty of tables. Such situation also reduces the speed.
2. The integrated view on distributed database is possible only under a condition of continuous communication of all data sources in a net. Thus, the temporary inaccessibility even of one of sources can make an operation of the informational - analytical system impossible, or produce error outcomes.
3. The carrying-out of composite analytical request demands large resources of DB server and results in speed decrease.
4. The different databases can support miscellaneous data formats and coding. The data in them could be inconsistent. Very often some variants of the answers can be obtained on the same request. That will be the result of asynchronous moments of data-refresh, difference in treatment of separate events, concepts and data, modification of data semantics during development of data domain, input errors, by loss of pieces of

archives etc. In this case shaping of a uniform consistent view on object of management can be impossible.

5. Practical impossibility of consideration of long historical sequences can be considered as the main imperfection, because when the data are becoming obsolete, they will be unloaded from transaction DB to archive.

## **5 Data warehousing approach in “Monitor 3.0” system**

### **5.1 Overview**

As cited above, the beta test of an analytical “Monitor 2.1” system has revealed a number of essential defects, bound with shaping of a plenty of composite requests, execution of cumbersome accounts at processing large volumes of data.

In the “Monitor 3.0” system, data warehousing [In96] [Ki98] has been a natural evolution of the previous one. A considerable growth in the amount of data produced and managed by different organizations led in many cases to the introduction of multiple database systems within the same organization in order to deal with different aspects of the data. Nevertheless, the poor data analysis functionality, that the traditional databases offered, was the incentive for the advent and development of DW systems. These systems store consolidated, historical, and summarized data, and are designed to support complex, mostly ad hoc queries, which involve large portions of the stored data.

DW tends to be fairly big in size, usually orders of magnitude larger than operational databases. This happens because the DW gathers information from all the data sources, and in addition, this information is stored for long period of time. The data structures used in DW are specifically designed to support decision-making. Thus, they can be exploited in the knowledge extraction process. DW manages the aggregated knowledge, in the sense that they consolidate data from many sources and therefore store a richer data collection than any other database [Bi97].

The “Monitor 3.0” system is realized on the basis of MS Office 97, with the usage Visual Basic 5. In the new system version, the DBMS Access 97 is used as the database. The realization of an analytical part is assigned to the spreadsheet processor Excel 97.

### **5.2 Architecture**

It is known [Ki98], that whole DW (Fig. 3) should be separated in two parts: the data staging area or back room and the data presentation area or front room.

Fig. 5 shows a typical horizontal cut through an overall DW environment. At the extreme left you see the traditional transaction-oriented legacy systems. The DW responsibility starts when you extract data from the legacy systems and enter it into the data staging area.

The data staging area is the complete back-room operation for the DW in which we clean, prune, combine, sort, look up, add keys, remove duplicates, assemble households, archive, and export. The data arriving in the data staging area is frequently dirty, malformed, and in a flat-file format. The data arrives in pristine third normal form, but that's rare. When we are done cleaning and restructuring the data, we leave it in flat file

form or store it in third normal form. Flat-files, simple sorting, and sequential processing dominate the data staging area. Third normal form relational representations are wonderful, but they're mostly the final result of a lot of hard work done in non-relational formats.

The key architectural requirement for the data staging area is that it is off limits to all forms of end user inquiry. We must not distract ourselves by having to provide availability of data, indexes, aggregations, time series, and synchronous integration across subject areas, and especially user-level security.

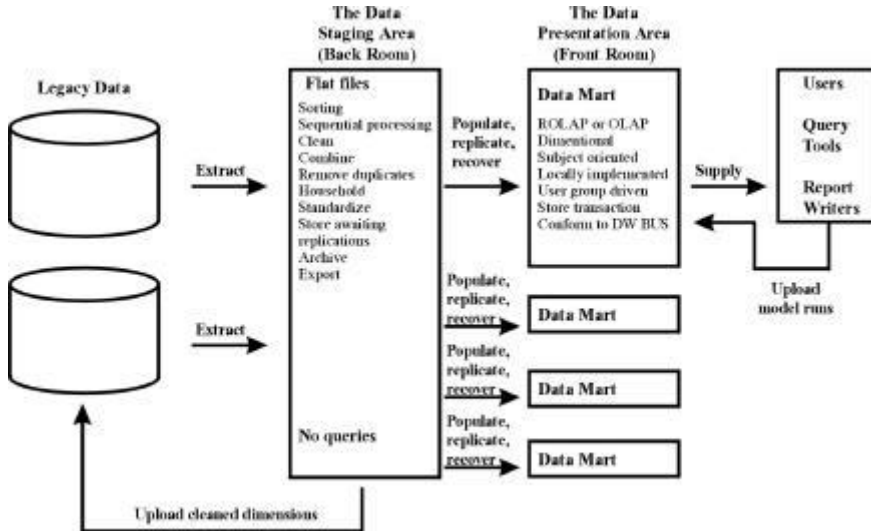


Fig. 3 DW showing the data staging area and the data presentation area

The tasks of data extraction, transformation, and loading take an enormous amount of time in a DW project. W.H. Inmon, in his books on data warehousing [In96], estimates that, on average, 80% of the efforts of building a DW go into these tasks. Today we are known more 200 tools aim to simplify those tasks.

The presentation area is the complete front-room operation for the DW. As its name implies, the presentation area is on stage and available at all times for end-user inquiry. The presentation area including ad hoc querying, drilling down, reporting, and data mining services all forms of end-user inquiry.

### 5.3 Data Mart and dimensional models

The presentation area is broken into subject areas, which are called data marts (DM). Each DM is organized entirely around effective presenting, which means dimensional models. Presenting encompasses all inquiry and analysis activities including ad hoc querying, report generation, high-end analysis tools, and data mining. All the dimensional models in all the DM look somewhat similar, and this suite of dimensional models must share the key dimensions. We call these the conformed dimensions.

Fig. 4 shows how to attach several independent DM together with conformed dimensions - a consistently defined set of dimensions that all DM that wish to refer to these common entities must use. Conformed dimensions typically include such things as date, product, region, food, and social group. We also have to consider conformed facts, which involve any measure that exists in more than one DM. Perhaps revenue, for example, is defined in several DM. To conform several instances of revenue, we must insist that the technical definitions of each instance be the same so that separate revenues can be compared and added. If two versions of revenue cannot be conformed, they must be labeled differently so users won't compare or add versions.

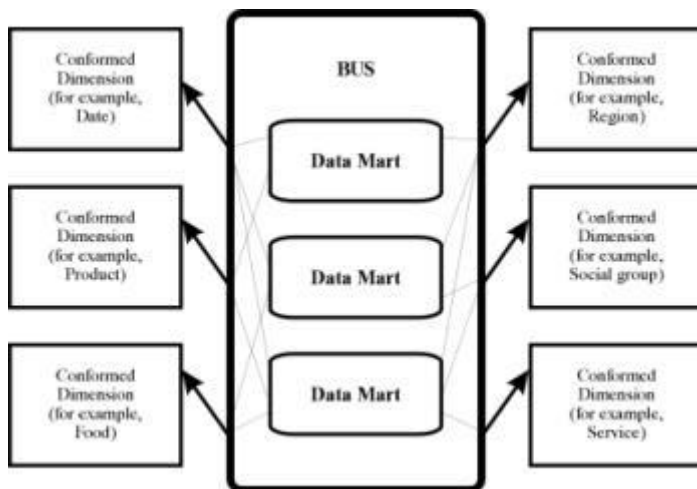


Fig. 4 The DW bus architecture, showing series of independent DM connecting to the conformed dimensions

The DW bus is a standard that allows separate data marts to be implemented by different groups in enterprise at different times. By adhering to the standard (conformed dimensions), the separate data marts can be plugged together.

The discipline of defining conformed dimensions and conformed facts before embarking on separate DM projects is the secret to incremental development. Anticipation of continuous change as needs and available data sources evolve. The similarity of all dimensional schemas lets us anticipate the effects of unexpected changes. Query tools and query strategies don't have to be reprogrammed when the user begins asking new kinds of questions. New data elements can be added to dimensional schemas in such a way that old applications continue to function without modification. It can add new data elements in this graceful way as new dimensional attributes, new additive facts, and entirely new dimensions.

Once the conformed dimensions and facts have been established, separate DM teams can proceed independently of each other. In many cases, it makes sense to build the first tables in each DM as dimensional images of single underlying sources. The DW bus architecture provides a formula combining these single source fact tables into higher-level combinations involving multiple sources. The charm of building single-source fact tables is that this is the fastest possible path to a partial deployment of DW data.

The separation of the back room from the front room is to implement the data staging area on a separate machine or on a separate process from the data presentation area. The final output of the data staging area is a set of load files for the data presentation area.

## **5.4 OLAP technology**

The standard logical model for representation the data in DW is the data cube. The data cube is a multidimensional view of all the data, which are represented as a set of numeric measures (or facts), and organized in several dimensions of interest. The numeric measures are the values of the data we are interested in, and the ones the analysis will be performed on. The dimensions are attributes of the data. They provide the context for the measures, and are organized in hierarchies of multiple levels. The measures are aggregated at each hierarchy level for each dimension to offer a more general view of the base data. The measures can quantities such as price, revenue, inventory, etc.

The physical design of the aforementioned data model is different for Multidimensional- and Relational- OLAP servers (MOLAP and ROLAP respectively). MOLAP servers directly support a multidimensional view of the data, achieved through a multidimensional storage engine. This approach results in more natural ways of data representation, but requires special care when storing data since data sets in high dimensions tend to be sparse. On the other hand, ROLAP servers take advantage of the existing relational database technology. It ensures high scalability when handling large databases with high dimensionality.

In difference from the traditional transaction-oriented systems, where the information is organized in the maximum normalized kind, data of DW is significant denormalized to ensure fast response time.

## **5.5 The star schema**

The database consists of a fact table, storing all the measures, and dimension tables around it. Having one table per dimension leads to the star schema, and by normalizing the dimension tables we get the snowflake schema.

In classical cases the table of the facts contains usually one or several columns giving numerical performance to any measures, and a few, it is usual integral, columns - keys for access to the dimensional tables. In the created system the developers have refused a possibility to task some columns of facts. At the central tables always is present only on one column with the facts. However, creation of an additional dimension always is possible, with the help of which can easily achieve demanded functionality (naturally by redundancy on a storage of keys of dimensions). In such a way some generality of work with the stored data was reached.

Usually such structures are developed so that then necessary formats of the reporting would be easily transformed to structures used for construction of DW [Pe98].

The Fig. 5 shows the schema “The food prices”. DW is organized as a set of such schemes. Each scheme has only one fact table, which can’t be included into any other scheme. But the dimension tables may be included into the various schemes (Fig. 4).

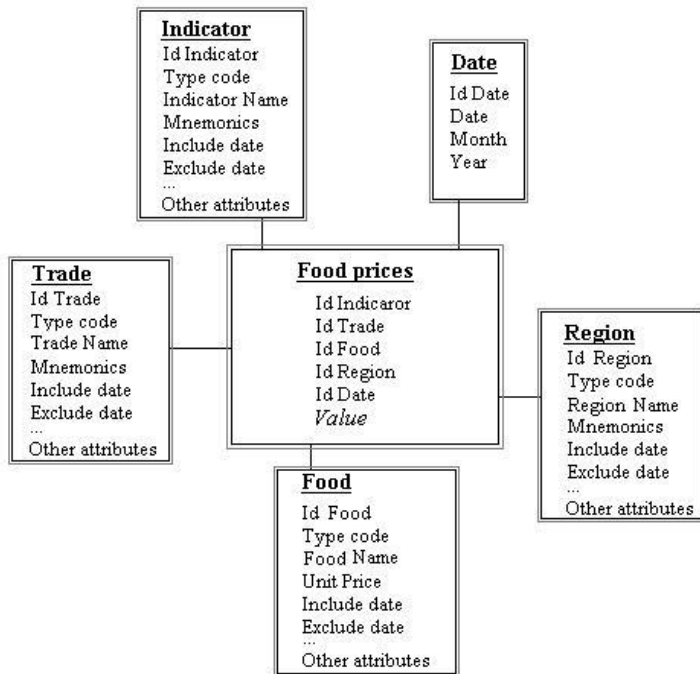


Fig. 5 The star-schema “The food prices”

The facts table “The food prices” forms schema with the same title. It is connected with Indicator, Trade, Food, Region and Date. In this case facts table has one column of currency (value) and five columns of keys for access to dimension tables.

## 5.6 Metadata

Metadata has been identified as a key success factor in DW projects [Sa96][MDC99]. It captures all kinds of information necessary to analyze, design, build, use, and interpret the DW contents.

In order to cope with the complexity of data structures and processes in data warehousing, a consistent management of metadata is required. Metadata (also called “data about data”) is defined as any information that may be used to minimize the efforts for administration and to improve the exploitation of a DW system. Typically, metadata is classified with regard to its use as business metadata, mainly needed by end-users, and technical metadata, produced and used by database administrators or by other software components of the DW system. The category of business metadata contains end-user-specific documentation, dictionaries, thesauri, and domain-specific ontological knowledge, business concepts and terminology, details about predefined queries, and user reports. Technical metadata includes schema definitions and configuration specifications, physical storage information, access rights, executable specifications like data transformation and plausibility rules, and runtime information like log files and performance results.

Consistent metadata management requires metadata to be captured and stored in a repository, which is shared both by various user groups and software components. Users need the repository as a consistent documentation in order to effectively and efficiently achieve their particular tasks.

## 6. Summaries

We summarize the experience gained in the developments of this project as follows:

1. It was constructed the socioeconomic data warehouse, which has the “Star” structure and integrates input data from 25 regions and 9 cities of Autonomous Republic of Crimea.
2. The incremental nature of the approach to both data integration and data warehousing deeply influences the design of methodology and software supporting tasks of Ministry of Economy of Autonomous Republic of Crimea (monitoring of price levels, incomes of the population, physiological norms of consumption).
3. It was developed the “Monitor 3.0” system. Last one was partially realized using data warehousing approach. It was built DW as some DM, (DM is a full logic subset DW). It was created simplest software for interfacing application, provided access to the data. It was adjusted connection with the data analysis system, which coincided with the one of the previous version “Monitor” system.

There are several important tasks on future:

1. Development a system of models of forecasting of socioeconomic processes;
2. Design and maintenance of software for interfacing applications and provide access to the socioeconomic data;

## References

- [Be98] Bednyak V.; Visual presentation of the fourth measurement. Computers+Programs, 1998; No.3; pp. 64-67; (in Russian).
- [Bi97] Birukov A., Decision Support System and Data Warehouses. DBMS; 1997; No.4; pp.37-45; (in Russian).
- [Go97] Golliet J. OLAP, Relational and Multidimensional Database Systems. Computers+Programs; 1997; No.3; pp. 36-40; (in Russian).
- [In96] Inmon, W.H. Building the Data Warehouse; QED Publishing Group; 1996
- [Ki98] Kimball, Ralph and Reeves, Laura and Ross, Margy and Thornwaite, Warren The Data Warehouse Lifecycle Toolkit: Tools and Techniques for Designing, Developing, and Deploying Data Marts and Data Warehouses; John Wiley & Sons; 1998.
- [MDC99] Meta Data Coalition. Open Information Model Version; 1.0 edition; August 1999; <http://www.MDCinfo.com>
- [Lu98] Lubinin M. The formulas for cube assembly. Computers+Programs; 1998; No.5; pp. 58-63; (in Russian).
- [Pe98] Petin P.A., Lovtsov I.V. and others. Information - Analytic System of Bank. Bank technologies; 1998; No.3; pp.15-22; (in Russian).
- [Sh98] Shevelyev L., Methods of analytical data processing for decision support. DBMS; 1998; No.4-5; pp.30-40; (in Russian).