About One NoSQL Mechanism For Accessing Panel Data

Zaza Gamezardashvili*
Giorgi Ghlonti**
Hakan Ergun***

Abstract

In the presented paper, the problems connected with the support of the life cycle of information resource represented by panel data are considered. An example of a logical markup of a statistical document emerged at the stage of information resource planning is given, as well as a relational formalism describing the structure of the statistical table included in the document. Examples of NoSQL statements for data manipulation and processing, based on this relational formalism are presented. The problem of building a cyber infrastructure intended for the accumulation and distribution of an analytical information resource is discussed.

Keywords: Panel data; Information resource lifecycle; Multidimensional data model. Relational formalism; NoSQL mechanism

Introduction

In statistics and econometrics, the term panel data is used for referencing time-tracking multicomponent samples representing various aspects of organization and activeness in complex socio-economic systems and other objects of interest. Hence panel data may be used for the presentation of macroeconomic information and the information about economic activities, or records of medical content as well as data arising in forensic science, meteorology, etc.

They are noteworthy in making possible to examine objects of research, simultaneously, from different points of view, allowing the study of interaction and influence of factors, tracking the change of objects in time, analyzing the diversity and heterogeneity of phenomena that permeate economic or social activity.

Some experts consider the emergence of information systems for accumulation and analysis of panel data to be the main achievement of the twentieth century (Heckman, 2001).

When designing models for collecting, storing, processing and analyzing panel data, the heterogeneity of objects of research leading both to large volume and variety of information, as well as the need to perform not only deep analysis for knowledge discovery, but also calculations of economic and other indicators, necessity to ensure openness of models with the possibility of subsequent inclusion of additional variables, data structures, and aggregates, additional methods of data analysis should be taken in account.

Management Issues For Panel Data Life Cycle

As a consequence, in the life cycle of the information resource represented by panel data, the planning stage of resource gains

* Assist. Prof. Dr., Caucasus University, Caucasus School of Technologies, Tbilisi Georgia. E-mail:gam_zaza@yahoo.com
**Ph.D. Faculty Computer Technologies and Engineering, International Black Sea University, Associate Professor. Muskhelishvili Institute of Countable Mathematics of Georgian Technical University, Tbilisi, Georgia. E-mail: gg59ster@gmail.com
*** Ph.D. c Faculty Computer Technologies and Engineering, International Black Sea University, Tbilisi, Georgia. E-mail: hergun@ibsu.edu.ge

decisive importance. Planning allows to identify similarities and differences in objects of research, to split the information into more or less independent components that in turn allows to appropriately schedule collection and subsequent storage of data, to take measures ensuring compatibility and comparability of information coming from various sources, to create prerequisites for openness and replenishment of the information space.

At the planning stage, identification and classification of sources of information and objects of observation, the definition of the structure and semantics of the information resource, specification of criteria for integrity and consistency of data, specification of access rights to data for different categories of users, scheduling the information collection process takes place.

As a result, a multidimensional model emerges incorporating at least three dimensions: attributes - objects - time. Objects often constitute a hierarchy such as the hierarchy of subordination or inclusion, or some sort of classification hierarchy.

In properly managed subject areas results of information resource planning remains appropriately documented. They are presented in the form of a statistical document, which later serves as the main reference source for semantics, structure, and content of data. Further, it will be referred to as Foundation Statistical Document (FSD).

In FSD semantics of the information resource is presented as a set of statistical statements expressing a number of judgments about the objects of research or their components (Dujenko, 1975). These statements are grouped into statistical tables and questionnaires, in accordance with the structure of the objects under study and the organization of the information collection process. The statement consists of a statistical subject and a statistical predicate. The subject indicates the item in question. The predicate expresses some judgment about this item.

For example, (Glonti, 1977) contains a fragment of the table that represents a part of the annual report of a health care service provision facility. This table is reproduced in Tab. 1.

In this particular example, the subject specifies the staff positions in the institution, the predicate is a statement about the number of this posts in the staffing table and the number of employees for reporting year actually occupying them, and then, in the same line, that information in further split on different aspects.

Thus, in fact, it's a grouped frequency distribution table - each line of the working area of it describes different characteristics of a sample within the unit of observation – in this particular case, the provision of the healthcare facility. The grouping feature is the post in the staffing table, and the attributes that characterize the samples are the number of objects within them and the number of objects in some of their subsets.

It should be noted that the samples covered by the observation are not absolutely independent of each other. They constitute a hierarchy of inclusion - some of them are subsets of others - for example, therapists or surgeons are different subsets of physicians in all, and district therapists make up a subset of the whole multitude of therapists. Keeping information about the specified partial order is important for data integrity management. Therefore, in the

source document already, the code of the data aggregate bears the information on the specified inclusion relationship indicating the relative position of the given group in the general hierarchy. (In our example, this code is specified in the column titled "Code"). Visually this order is reflected in the stacking of the table rows in a certain sequence.

Relational Data Model For Statistical Table

At a certain stage in the design of information systems, it becomes necessary to map the original data structures to conceptual schemas supported by DBMSs. Due to the fact that the working area of the table is a matrix or a set of strings containing values of identical attributes, it is natural to choose in favor of relational formalism.

The basis of the conceptual model of a relational database is a relation defined as a (finite and unordered) set $R = \{t_i \mid 1 \le i \le n\}$ of functions that map the set $A = \{A_j \mid 1 \le j \le m\}$ of attribute names to $\bigcup_i Dom_i$ - the set union of domains (or sets of values) of the specified attributes (Maier, 1983). (Dom_i is the domain of the attribute A_i). The additional restriction imposed on the function t_i consists of condition $t_i(A_j) \in Dom_j$.

The set A of attribute names is called the schema of the relation. Functions t_i $(1 \le i \le n)$ are called tuples.

A subset of a relation's schema is called a primary key of the relationship if it uniquely defines a tuple within the given relation. Thus, $K \subseteq R$ will be the key to the relation R , if

$$\forall t_i \in R, \forall t_j \in R, \forall T \in A \quad \left(t_i(K)\right) = t_j\left(K\right)\Rightarrow (t_i(T) = t_j(T))$$

If an attempt to make abstraction from the semantics of information contained in the above mentioned statistical table, we'll try to treat it as a relation, the headings of the columns will make a set of attribute names, the rows of the table will be the tuples of the relation, and the primary key of the relation would naturally be an attribute specifying the name of the main grouping characteristic - in this particular case, staff position in the organization.

But unlike the relation, that is treated as an unordered set of tuples, in our case, the domain of the key attribute is ordered.

Naturally, it's possible to split this relation horizontally representing it as a union of several relations, each containing the tuples of the same level of the hierarchy. But since information about above mentioned partial order is crucial for the maintenance of data integrity, it would become necessary to introduce additional data elements for storing it. Therefore, fragmentation should be discarded at this point. And in this case, one additional advantage is awarded.

Namely, as the identifier of the group and its code are in a one-to-one functional relationship, the code, in turn, could be considered as the primary key of the relation and the tuples could be accessed through this code. That would give some advantages in the design of data manipulation language for DBMS.

But there is an opportunity to go further, to introduce a full order in the domain of the key attribute of the relation, referring to the tuple not through the code of it, but through the line number under which

this tuple is presented in the original statistical table. And this restriction should not be perceived too rigid, depriving the tuples of independence, since, in fact, it merely emphasizes the order originally implied in the domain of the key and that reflects the semantics of particular subject area.

If this restriction is extended to the set of attribute names that are by now partially ordered in accordance with the semantics - while putting full order in it and identifying the attributes by the numbers of the columns under which they are presented in the table, it would become possible to refer an attribute through the locator **Doc**xxx**Tab**yyy**Line**ll**Gr**dd,

where **Doc**xxx**Tab**yyy specifies the name of the relation, **Line**ll is equivalent to the value of the primary key, and **Gr**dd is the reference to the attribute itself.

NoSQL mechanism for data manipulation and metadata management.

Such a locator can be put in the basis of an excel-like user interface. For example, the conditions of data integrity for our table might be represented in the following form:

DocXXX**Tab**YYYСтр1 =
SUM (DocXXX**Tab**YYY**Line(**2:3)) +
+ **SUM (Doc** XXX**Tab**YYY**Line**(6:10)) +
+ **SUM (Doc**XXX**Tab**YYY**Line**(12:16));
(1)

DocXXX**Tab**YYY**Line**3 = **SUM**
(DocXXX**Tab**YYY**Line**4:5); (2)

DocXXX**Tab**YYY**Line**11 ≤
DocXXX**Tab**YYY**Line**10;
(3)

This approach has been realized in the system of analytical information processing described in (Glonti, 1977).

The formulas presented above confirm that the use of line numbers as identifiers gives an advantage, since it allows to use in formulas the ranges of values for rows and graphs while using the code of data aggregate, and even more attribute names as identifiers are deprived of this advantage.

But in this case, a natural need to refuse the SQL mechanism of data access emerges, since using SQL along with above-mentioned locators would be unnecessarily resourced intensive.

And this, in turn, will require the design of a powerful semantic metadata management system.

The initial content of the system of semantic metadata is given by the lexical component of the statistical document, presented, as mentioned above, as the set of statistical statements, consisting from statistical subjects and statistical predicates.

In the heading of the table, this collection is presented as the union of two hierarchies, called the subject and the predicate of the table, and it's necessary both to store information about the structures themselves and to provide ability of using their content as source of metainformation and to provide ability of processing the text itself.

In XML format, the description of the subject and predicate of the above table has the following appearance:

<The subject of statistical table>

<Text> Positions </Text>
<Decomposition of the subject of statistical table >
<Text> Doctors of all
<Text> From within specialists: /n CEOs of institutions and their deputies
</Text>
<Text> Therapists of all
<Text> from within: /n districts therapists </Text>
<Text> workshop therapists </Text>
</Text>
<Text> Surgeons </Text>
<Text> Neurologists </Text>
<Text> Endocrinologists </Text>
<Text> Neuropathologists </Text>
<Text> Pediatricians of all
<Text> from within district pediatricians</Text>
</Text>
<Text> Allergists </Text>
<Text> Infectious disease specialists </Text>
<Text> Physiotherapists </Text>
<Text> Urologist </Text>
<Text> Dentists </Text>
</Text>
</Decomposition of the subject of statistical table>
</The subject of statistical table>

<The predicate of statistical table>
<Text> The number of posts in the whole institution
<Text> allocated </Text>
<Text> occupied </Text>
</Text>
<Text> From within in the clinic
<Text> allocated </Text>
<Text> occupied </Text>
</Text>
<Text> The number of individuals /n in the institution, as a whole,
/n in the positions of core workers
<Text> Core workers </Text>
<Text> External part board </Text>
</Text>
</The predicate of statistical table>
And finally, the whole statistical table might be presented in following way:
<Statistical Table>
<Table Identifier> </Table Identifier>
<Table header></Table header >
<The subject of statistical table> </The subject of statistical table>
<The predicate of statistical table> </The predicate of statistical table>
</Statistical table>

With further development of information technologies, proper management of analytical information resource, as condition of knowledge accumulation in subject area and ability of motivating decision-making, acquires more and more urgent significance.

Conclusion

Analytical information is an interdisciplinary and intersectoral resource and it is necessary to ensure the development of this resource in an industrial mode when information is systematically collected and gradually accumulated, based on coordinated models and can be used by all interested parties in the future.

From a technical point of view, it is necessary to consider the possibility of building a cyberinfrastructure that supports an active information space capable of taking initial information from primary information providers, aggregating it according to specified multidimensional models, and

ensuring compliance with both routine and specialized information requests of various users (Ghlonti. 2012).

The construction of a cyber infrastructure requires solving problems related, on the one hand, with ensuring the quality criteria of an information resource, and on the other hand, building architectures that meet the requirements of openness, scalability, transparency of the information space, and minimization of information stability.

These problems should be solved within the framework of the system approach.

References

Heckman J.J. (2001). Micro Data, Heterogeneity and Evaluation of Public Policy. Nobel Lecture // Journal of Political Economy. Vol. 109. № 4.

Dujenko G. A. (1975). Documentnaia Lingvistika. M. “Statistika”. (in Russian) Available: http://genling.ru/books/item/f00/s00/z0000024/st000.shtml (May 20 2019).

Glonti I. G. (1977). Ob odnom opite razrabotki informacionnoi bazi otrasli na primere zdravooxranenia. Akademia nauk GSSR, Vichislitelnyi centr, Trudi XVII:2. (in Russian).

Maier D. (1983). The theory of relational databases, Computer Science Press Inc, 1983, ISBN 0-914894-42-0.

Ghlonti G. (2012). A Unified Analytical Information Space as the Condition for Sustainable Development of a Subject Area (Tbilisi-Batumi, SRBSU, MESDG, IBSU, 7th Silk Road International Conference: ”Challenges and Opportunities of Sustainable Economic Development in Eurasian Countries”, 24-26 May.

Positions	Line №	Code	The number of posts in the whole institution		From within in the clinic		*)The number of individuals in the institution, as a whole, in the positions of core workers	
			allocated	occupied	allocated	occupied	Core workers	External part board
Doctors of all	1	1.0						
From within specialists: CEOs of institutions and their deputies	2	1.1						
Therapists of all	3	1.2						
from within: districts therapists	4	1.2.1						
workshop therapists	5	1.2.2						
Surgeons	6	1.3						
Neurologists	7	1.4						
Endocrinologists	8	1.5						
Neuropathologists	9	1.6						
Pediatricians of all	10	1.7.0						
from within district pediatricians	11	1.7.1						
Allergists	12	1.8						
Infectious disease specialists	13	1.9						
Physiotherapists	14	1.10						
Urologists	15	1.11						
Dentists	16	1.12						

Tab. 1. Staff positions at the end of the reporting year

*) this column is filled every 5 years.