An Approach to Data Storing and Analysis in the Case of Using a Non-standard Enterprise Calendar*

Dmitry V. Zhuchkov
Institute of Computational Modeling of the Siberian Branch
of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia
dimkas@icm.krasn.ru

Abstract. The paper presents an approach to data storing and analysis, which can be applied in organizations that use a non-standard enterprise calendar with the original division of dates into categories and periods. For such cases it is proposed to use a specialized data structure – *local calendar*, which has a rigid connection with the commonly used calendar, and also takes into consideration the local specific of processing with calendar dates. The structure of the local calendar includes the *main calendar table*, which presents an ordered list of days with the most frequently used parameters, and additional *sub-calendar tables*, which contain a selected part of calendar rows used to solve individual business and analytical tasks. The proposed approach is quite universal for various cases of the non-standard enterprise calendar. The proposed structure of the local calendar functions well with normalized OLTP systems, denormalized data warehouses and OLAP systems.

Keywords: Calendar, Local Calendar, Database, OLTP, OLAP, Data Warehouse, Temporary Dimension.

1 Introduction

1.1 Problem of Using a Non-standard Enterprise Calendar

In this paper, *standard calendar* is the generally applied Gregorian calendar, which is used by almost all countries of the world [1]. Accordingly, the *standard division of dates* into categories and periods in this study is the division into years, half-years, quarters, months, etc. This division has its own difficulties for data analysis, but it is commonly used and supported by most analytical systems.

However, a large number of organizations use a non-standard division of dates into categories and periods. As a result, their analyzed dates have properties which cannot be obtained from the standard calendar. For example, in non-standard calendars, alternative chronologies can be used, days can be separated into “workdays” and "weekends", weeks can be divided into "even" and "odd" ones, data can be combined into 10-day periods, school year can begin in August and end July, etc. Non-calendar features of dates may change from year to year, and their set may significantly differ within individual data analysis tasks.

The main goal of this research is to develop a common universal approach to data storage and analysis, which will take into consideration the features of various enterprise calendars, and combine various non-standard calendars into a unified system.

1.2 Variants of Non-standard Calendar and Basic Idea of the Proposed Approach

Based on the analysis of non-standard calendars, there are four main features of such calendars which can be distinguished:

- Alternative chronology – ordinal number or duration of the year (month) is not equal to the Gregorian calendar year (month).
- Non-repeating analytical periods – the periods limited by randomly selected specific calendar dates.
- Repeated analytical periods – the periods that are not equal to the generally used calendar periods.
- Extended date attribution – assigning to calendar dates additional attributes and properties to be used later in data analysis.

After careful consideration of these cases, it becomes clear that the first three variants of the non-standard calendars are special cases of the fourth variant. If it is possible to add a sufficient number of additional attributes to each calendar date, it becomes possible to easily implement all three noted types of the non-standard calendars. As a result, the main idea of the proposed approach is to develop a special data structure which, being tightly linked to the generally applied standard calendar, will store an arbitrary number of additional attributes and properties for calendar dates.

2 Structure of the Local Calendar

Local calendar is a new entity, which is added to the database of the enterprise information and analytical system. The local calendar is a set of interconnected tables, namely the main calendar table, sub-calendars and reference tables. *Main calendar table* (hereinafter – the calendar table) is the central element of the local calendar. It contains an ordered list of days (calendar dates) from the beginning of an *era* of the local calendar. As the start date of the local calendar era any significant date can be taken which is back far enough in the past. For modern information systems this may be the date "01.01.2001", for systems containing the XX century data – "01.01.1901", for middle-historical systems – "01.01.1700", and so on.

The unique identifier (primary key) of the calendar table rows is the ordinal number of the day from the beginning of the era. If it is necessary to deal with events which occurred before the beginning of the era, negative key values can be used.

The calendar table also includes the following data:

- the actual date, represented by a indexed column with the “date” or “date-time” data type;
- canonical textual representation of the actual date – for example, in the format “day.month.year” or “year/month/day”;
- alternative textual representation of the date, if it is used in the system;
- other unchanged and most frequently used date parameters for the analysis – for example, “year”, “month”, “day of the week” in the numeric form, etc.

The direct inclusion of the mainly used date properties in the calendar table allows one to avoid additional calculations on the data analysis stage. In the case of systematically using an alternative chronology, the parameters of this chronology are also included into the calendar table. But if the system is supposed to use several equivalent chronologies, then they should be considered as *additional attributes* of the calendar dates.

Additional *sub-calendar tables* are created to include extended attributes of the dates, such as alternative chronology or business days flag. Sub-calendar tables have a structure which is similar to the satellite tables in the Data Vault 2.0 technology [2]. The unique primary key of the sub-calendar table is a direct link to the main calendar table. Thus, the sub-calendar is literally a part of the calendar table key values that are included in a separate table containing additional attributive columns.

The columns of the main calendar table and sub-calendars that have the property of enumerability must refer to the corresponding reference tables (master-data tables or classifiers) [3]. Also a personal reference table is created for each type of non-repeating analytical periods. Such reference tables contain a list of analytical periods, including the dates of the start and end of the period. After creating the reference list of analytical periods, a sub-calendar can be automatically generated, to include all the days related to the specified periods. To deal with the repeated analytical periods, similar reference tables of the periods should be generated. For the control purposes in this type of reference tables special parameters are introduced to limit the duration of the analytical period. The calendar table, as well as sub-calendars, can include links to the reference tables of periods and date categories, as well as simple attributive columns.

Thus, when information system contains the local calendar – a complex structure, including the calendar table and sub-calendars, having an equal structure which provides an extended attribution of dates, making it possible to combine the approach with the data analysis taking into consideration non-standard properties of the dates. The data tables in the information system can either directly refer to the calendar table, or access it via an external connection using any of the “date-time” columns.

3 Practical Aspects of the Local Calendar Realization

The proposed local calendar scheme has a snowflake-type form, which is typical for OLTP systems [4]. In Figure 1, the central element shown in red is a calendar table, the surrounding orange elements are sub-calendars, and the blue tables on the outer contour are reference tables. The geometric size of the elements in figure 1 corresponds to the dimensions of the tables in the database. The red table has a large number of rows, and the yellow tables have a large number of columns. The lines in figure 1 are foreign key links located in the database tables.

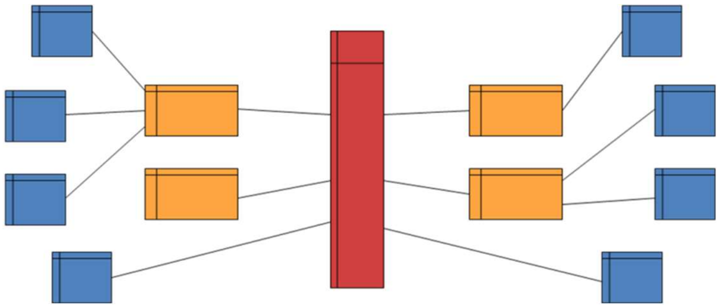

Fig. 1. General scheme of the local calendar.

Figure 2 illustrates in more detail a composition of the data in the local calendar tables. The main calendar table has its own key field, attribute fields, and links to reference books. The sub-calendars contain key values from the main calendar table, as well as attribute fields and links to reference lists. The reference tables are usually separated tables, but in some cases they can also be linked to others.

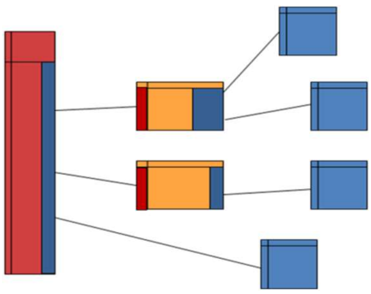

Fig. 2. Data intersection in the local calendar tables.

The actual dimensions of the local calendar tables arequite predictable. One century-size calendar table will contain 36525 rows, one decade-size subcalendar will contain 3653 rows and average one year-size sub-calendar will contain at most 366 rows. The size of reference tables can be different, but in general it is not a very big amount of data.

To create a local calendar in the existing information system, it is necessary to take the following steps:

1. to find out the needs of the organization in analyzing data related to calendar dates, and highlight the main characteristics of the calendar dates, as well as the main periods and categories of the dates used in the analysis.
2. based on the selected characteristics, to create tables of reference lists and classifiers to be used in the local calendar structure.
3. to determine the start and end dates of the local calendar. When choosing a starting date, one should make a reserve for the past, but not too large. Similarly, one should decide on the end date of the calendar, which is to lie far enough in the future. The optimum date for the end of the calendar is a 10-year gap from today. If necessary, the end date of the calendar can be moved further into the future at any time.
4. to create the main calendar table and fill it with rows having an ordinal number, starting from the start date of the calendar till the current end date. A basic set of columns in the main table is described above. One can expand the basic set of columns, but only the most frequently used calendar date parameters should be included in the main calendar table. The number of columns in the main calendar table should be limited, and the procedure for adding new columns to the calendar table should be regulated.
5. Sub-calendars are created in accordance with the defined objectives for data analysis. For each sub-calendar, its owner must be defined – a person or role in the organization which is responsible for the content of information in the sub-calendar. Changes to the sub-calendar must be approved by its owner.
6. For all the elements of the local calendar, it is recommended to develop a unified naming system for individual elements and data structures. The main rule of this system is that elements with a similar purpose should have similar names.
7. Two cases are possible using the local calendar storage data in the information system:

- data tables keep a direct link to the calendar row, i.e. a table with data directly includes the day number, and in order to determine a specific date one should to relate the table data with the link calendar.
- data tables store information on the dates in the "date" or "date-time" format and for the purposes of analysis, the connection with the calendar is established by linking two fields of the type "date".

The practical implementation of the local calendar concept should not cause significant difficulties. The main difficulties lie in the regulation of individual procedures and establishment of general corporate rules for handling the local calendar.

4 Conclusions

The proposed approach is quite universal for application in the case of various non-standard enterprise calendars. The proposed structure of the local calendar fits the normalized OLTP systems, as well as the denormalized data storages and OLAP systems [5]. In the case of using the presented approach with denormalized data warehouses, the data structure of the local calendar can easily be transformed from the original snowflake-type scheme to a pure star-type scheme, or even into a flat-table form.

It should be noted that the presented approach has a limited efficiency in solving problems requiring finer granularity of the calendar than one day. Dividing down to an hour can be implemented by analogy with the presented approach, but a smaller division of the calendar will require changes in the applied approach.

References

1. Gregorian calendar. Wikipedia page, https://en.wikipedia.org/wiki/Gregorian_calendar, last accessed 2020/05/20
2. Linstedt, D.: Data Vault 2.0 Being Announced. DanLinstedt.com. Dan Linstedt. Retrieved 2014-01-03, http://danlinstedt.com/datavaultcat/data-vault-2-0-being-announced
3. Loshin, D.: Master data management. 1st edn. Morgan Kaufmann (2008), https://doi.org/10.1016/B978-0-12-374225-4.X0001-X
4. Online transaction processing, Wikipedia, https://en.wikipedia.org/wiki/Online_transaction_processing, last accessed 2020/05/20
5. Online analytical processing, Wikipedia, https://en.wikipedia.org/wiki/Online_analytical_processing, last accessed 2020/05/20