Conceptual model of operational-analytical data marts for big data processing

Aleksey Raevich[1,*], *Boris* Dobronets[1] , *Olga* Popova[1] , *Ksenia* Raevich[1]

[1] Siberian Federal University, 660074, Krasnoyarsk, Borisova street 26, Russia

Abstract. Operational data marts that basically constitute slices of thematic narrowly-focused information are designed to provide operational access to big data sources due to consolidation and ranking of information resources based on their relevance. Unlike operational data marts dependent on the sources, analytical data marts are considered as independent data sources created by users to provide structuring of data for the tasks being solved. Therefore, the conceptual model of operational-analytical data marts allows combining the concepts of operational and analytical data marts to generate an analytical cluster that shall act as the basis for quick designing, development and implementation of data models.

1 Introduction

In modern and dynamically changing world, information, data and knowledge constitute the indisputable value and critical development factor both for a small enterprise and for a whole country.

Business analysis systems as a class of systems that implement the principles of decision support systems (DSS) focus on intelligent control of data and imply the use of a complex of technologies, software and practices aimed at achievement of the business objectives.

It should also be noted that detailed and quick analysis of input data for retrieval of tacit knowledge is possible only provided proper "understanding" of the data context [1]. Comparison of various data models demonstrates that specific units of information can be of small value as perceived by the end user. Information can be collected in an infinite number of sets and information elements that can co-exist in any number of various information kits. Such information should be contextualized in order to be brought to the level of the intellectual knowledge.

Therefore, the process of designing information structures to represent the data in order to fulfill certain tasks for business analysis systems shall consider the amount of data being processed, as well as the time required for a change of analytical model in case of business tasks being changed, including the changes caused by alteration of data structure in the source [2].

* Corresponding author: raevich.ap@yandex.ru

The suggested approach to generation of operational and analytical data marts shall provide availability of sources, promptness and structuring of data for tasks being solved.

2 Concept of operational data marts

Within the concept of operational data marts, data marts are considered as logically and physically decomposed subsets of data represented as sliced arrays of thematic narrowly-focused information oriented at the needs of the specific group of users.

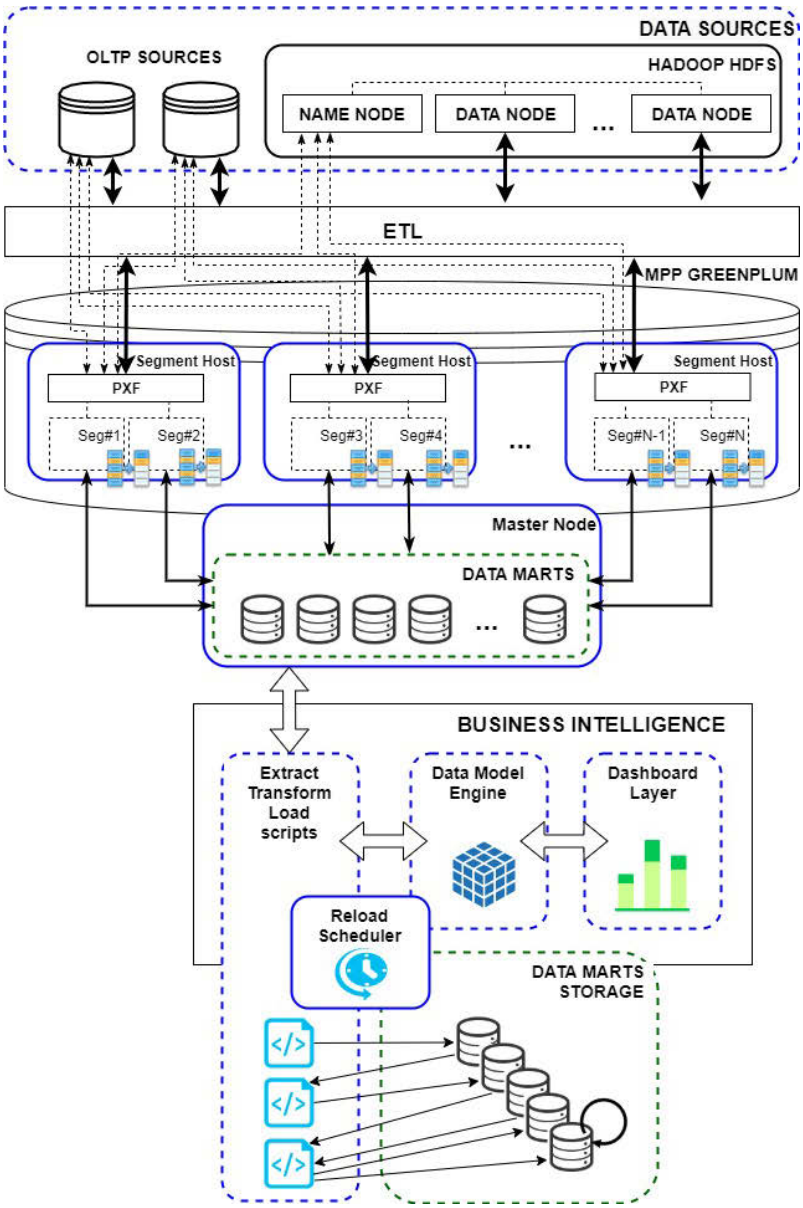

Fig. 1. Conceptual view of information structure of operational analytical data marts

In standard simulation each entity or object of the real world can be represented by a separate table with a variety of arguments. The set of jointly stored interrelated tables of minimum redundancy will be referred to as the data source.

The data source table can be designated as the information resource element a group of users shall require an operational access to. Then the information catalogue of resources shall comprise a tuple of information resources each element of which is referred to one of two classes: reference information class or detailed data class.

$$IK = \langle IR_i \rangle,\ \forall ir_{ij} \in IR_i,\ IR = \{ir_1, ir_2, ..., ir_N\} \tag{1}$$

It is obvious that relevance of different information resources elements based on data requests over a long time period T shall be different for different groups of users. Thus, the maximum frequency of requests for each information source element can be calculated for the preset time period:

$$f_{\max} = \max_{T}(\langle f_{t_1}(ir_{ij}), ..., f_{t_l}(ir_{ij}) \rangle) \tag{2}$$

The notion of a normalized index has been introduced to rank information resources in terms of their relevance. Such index is calculated as the maximum value of the aggregate function for the tuple of information resource relevance frequency values to the frequency of user references to the information resource element within a specified period:

$$w_c = \frac{f_c}{f_{\max}} \tag{3}$$

Therefore, the current frequency of each information resource element relevance for the preset time period can be defined based on the set of values [0; 1].

Transactional systems or other storages used for big data processing, such as Hadoop can be used as information sources for operational data mart storage (Fig.1).

The most requested tables shall be uploaded to the operational data storage via filtration of information catalogue elements using the normalized index. Thus, the upload of data from the sources shall be started from the class of most relevant resources including detailed data and reference information used for detailed data contextualizing.

3 Concept of analytical data marts

Compilation of logical data models within a centralized storage is often performed with account of two main requirements: elimination of redundancy and maximum improvement of data reliability, both of which derive from approaches to collective use of data marts within the storage by a group of users.

The concept of analytical data marts basically deals with selection of profile data for the specific area of operations. And as the selection of essences, attributes and fixation of interrelations between essences depends on the semantics of the subject field and is performed by a system analyst based on his or her own subjective understanding of the application task specifics, the sets of attributes describing information objects can overlap, partially overlap or never overlap in the course of generating analytical data marts.

Business application level contains parameters converted to the terms of the specific business logic that are also projected to end user reports grouped based on business tasks as part of objects prepared and calculated at the business logic level. The process also includes calculation of field-specific parameters, hierarchy building and compilation of fact tables and measurement tables in diagrams for each business area.

Hierarchy classes of essence attributes shall be defined based on analyst needs when generating data marts for the predefined set of business processes $\{M_1,M_2,\ldots,M_n\}$. The level of data detail shall be available according to the lowest level of data detail.

For instance, for cube H_n representing the solution of a user task for business process M_i, the tuple of analytical data marts shall be defined as follows:

$$(M_i, H_{n_i}) = \left\langle F_{d_i}, F_{a_i}, F_{m_i} \right\rangle \quad (4)$$

Where F_{di} is a mart of detailed (operational) data transferred directly from data sources or from the operational storage. They correspond to elementary transactional events; F_{ai} – aggregated marts that are basically generalized values of information objects attributes; F_{mi}– metadata mart acting as an index of detailed data contents for connection of data among the marts.

Data sources shall also include any objects that contain both the structured and non-structured data that can be useful to solve analytical tasks. Analytical platform shall have access to the data directly from the source or after it has been converted to another format.

ETL layer shall extract the data from various sources and convert the data to the agreed format. The data shall be converted from one type of information to another type using data processing methods. Data processing methods shall comprise operations of data collection, data formalization and filtration, data sorting, grouping and backup, data transportation and conversion.

The internal storage of analytical data marts shall be based on the concept of QlikSense QVD BI flat indexed tables that provides for data compaction and high speed of data reading up to 100 times as compared to other data sources.

The QlikSense system core generates an associative model for data interrelation providing “on-the-go” aggregation and indexation of data via interactive transformation of the model for user needs without the need to transfer the input data from the sources.

4 Research findings

A system based on MPP GreenPlum massive parallel architecture is used as a storage for operational data marts. The architecture incorporated in GreenPlum Database software and hardware complex is based on breaking down the complete array of data into the segments that can be worked with simultaneously.

The GreenPlum architecture has been initially developed for business analytics and for analytical processing of data using standard equipment. The data segments shall be automatically distributed among several segment servers, each of which holds and controls a specific part of the total array of data.

Configuration: «segment node»: [CPU] XEON 12 cores 2.66GHz [RAM] 64GB [HDD] Hitachi scsi 3x146GB 10000rpm RAID-5. 36 segments have been configured (12 per server).

Table 1 shows the results of test export of slices arrays of thematic information from Hadoop Apache Hive source to the GreenPlum mart.

Table 1. Duration of test operations at GreenPlum cluster

Operation	Average duration, ms
Creation of an external table for hive data warehouse for test collection of data, 585,000,000 lines, 50 columns (int, float, text, datetime)	160.262
Creation of a physical table distributed randomly from the external table to 1 Gbit networks	1733700.608

Operations of joining the test sets containing 505,000,000 and 7,757,000,000 entries physically distributed based on random key	884237.118
Sorting of the generated array containing 7,700,000,000 line by one field	761334.002

It should be noted that unlike conventional DBMSs that store date strictly by lines, Greenplum can store the data being processed both by lines and by columns. This radically decreases the load on disk subsystem during statistical analysis of data. Moreover, one of the main advantages of Greenplum is the use of PXF framework that enables each segment to exchange data with sources in parallel.

Qliksense BI platform is deployed at the server of the following configuration: [CPU] XEON 24 cores 2.33GHz [RAM] 1TB [HDD] Hitachi scsi 278GB 10000rpm.

Two sets of data have been generated as test sets: #1 = 9,000,000 lines and #2 = 11,000,000 lines containing 25 and 95 columns with corresponding data types: int, float, varchar (<= 100 char), datetime. “Source 1” - local disk storage of QlikSense server. “Source 2” - GreenPlum cluster.

Table 2 shows the results of test operations for specified sources and sets of data.

Table 2. Duration of QlikSense BI test operations

Operation	Source1, min		Source2, min	
	#1	#2	#1	#2
Data upload without field indexing	0.23	0.27	6.8	9.88
Data upload with field indexing	0.5	0.65	7.16	10.25
Data upload with synchronization of QlikSense visual data model without field indexing	1.77	1.83	8.1	11.13
Data upload with synchronization of QlikSense visual data model with field indexing	1.8	1.85	8.2	11.1

Relevant ODBC connectors shall be used for operation with various sources in Qliksense. When working with data sources via a visual editor, any changes in upload scripts shall require update of data and synchronization with the source. For analysis of data and compilation of models in visual environment it is reasonable to save the input data in QVD internal format in a local disk storage, which shall accelerate the data reading operation up to 100 times.

5 Conclusion

When used as an operational storage of data marts, Greenplum cluster enables efficient work with big data by parallel processing of data from the sources at the cluster segments. Greenplum contains the integrated libraries of analytical algorithms with open source code that perform computations with parallel processing of data by mathematical and statistical methods and machine learning methods for structured and non-structured data.

Impossibility of operational upload of big amounts of data from sources to be analyzed in BI is compensated by use of the internal storage of analytical data marts that can be initialized by complete export of sliced arrays of data or by partial (incremental) upload.

The suggested conceptual model of operational-analytical data marts tested at the base of MPP GreenPlum and BI QlikSense software and hardware complex enables various groups of specialists to transform the marts based on big data within a short period of time (close to realtime), gain operational access to the data and consequently transform the data model for the tasks being solved within the shortest possible time. Moreover, it allows not

only gain a better understanding and a deeper idea of data but also improve the data visualization for end users.

References

1. Raevich A.P., Dobronets B.S. Developing a conceptual model of operational-analytical data marts// Modelling, optimization and information technoogies. - 2019. - Vol.7 - No.4 - Pp. 1-13
2. Golov N., Rönnbäck L. Big Data normalization for massively parallel processing databases //Computer Standards & Interfaces. - 2017. - Vol. 54. - Pp. 86-93.