Data Vault as a Decision Support Platform for an Electricity Supplier in the Open Electricity Market

Dragoljub Krneta, Sofija Krneta
Dwelt software
Banja Luka, Bosnia and Herzegovina
dragoljub@dwelt.net, sofija@dwelt.net

Abstract - **In this paper, an approach to designing data warehouse based on Data Vault concept is shown on the example of supply of the electricity market. The Data Vault concept made certain improvements to data warehousing, most of them regarding increased performance on data loads, scalability and ease of changes regarding load and new dimensions. The motivation for using this approach in this example is need for using those benefits of using Data Vault for decision support in the open electricity market.**

Keywords - **Data warehouse, Data Vault, decision support, supply of electricity**

I. Introduction

The Data warehouse (DW) contains data that is conveniently extracted from different data sources. Data sources can be relational or object databases, network databases, as well as data in different files [1]. As written in [2], data warehouse is a subject oriented, non-volatile, integrated, time variant collection of data in support of management's decisions. Data marts are small portions of data warehouses focused on a subset of data contained in the data warehouse [3]. The process of Extracting, Transforming, Loading (ETL) data comprises extracting data from transactional databases, transforming, unifying and cleaning the data, transforming data into suitable formats and loading the transformed data to the data warehouse [4].

Business intelligence is a decision support concept based on data, which is actually technical assembly used for providing decision makers in organizations whole information and statistical data for making decisions. The main goal of implementing a business intelligence solution is providing support and defining procedures of business decision making [3].

Business intelligence is focused on determining knowledge based on data used for decision making, often not suitable for human brain to conclude because of the amount of information summarized. Data is transformed into facts connected to strategic and tactical moves which can make advantage on the market. Core of all business intelligence systems are databases, usually data warehouses.

Data warehouses are notable by the type of their architecture. Bill Inmon proposed in [2] and [5] a data warehouse concept named CIF (Corporate Information Factory) defined as a unified data warehouse, with all source databases being in the third normal form (3NF). Other notable architecture is the bus architecture, described by Ralph Kimball et al in [7] and [8] as an assembly of data marts with different dimensions. From these notions, Data Warehousing 2.0 defined a standard regarding Data warehouse architecture. One of the competences emphasized in this standard is its capability of supporting variations in data structure and content over time.

II. Data Vault Approach

The Data Vault approach in creating data warehouse was introduced with the purpose of addressing flexibility and overall performance of analysis process challenges, providing better upkeep of database and records. The concept was introduced by D. Linstedt in [9] and [10]. The Data Vault (DV) model has specific structure, consisted of entities called Hub, Link and Satellite. A Hub entity represents all identities, typically business keys, Links are foreign key references. A Satellite represents significant attributes of a Link or a Hub.

Data Vault is highly normalized in terms of data, modelling parts of the environment data and low sensitivity to variations in the environment that is describing, without the need for restructuring [11] [12].

Databases which are decomposed to basic relational elements are called databases in Sixth Normal Form (6NF). Sixth Normal Form is especially important when database is holding time variant or seasonal, interval temporal variables [13]. C. Date et al presented Sixth Normal Form as "A relvar (table) R is in 6NF if and only if R satisfies no nontrivial join dependencies at all, in which case R is said to be irreducible". In the implementation where satellite tables are consisted from just one attribute, The Data Vault model, is designed in a 6NF.

A formal background of Data Vault model formalized in the scientific paper of M. Golfarelli et al [14], as a graph-based formalization and multidimensional schemata:

Definition 1 (Data Vault Schema) [14]:

A data vault schema *(briefly, dvschema) is a directed graph* $\nu = (T, F)$ *where* $T = T_H \cup T_L \cup T_S$ *and:*

1. T_H*,* T_L*, and* T_S *are, respectively, sets of hub, link, and satellite tables;*

2. each arc (t, t') in F represents a functional dependencies (FD) from a foreign key of table t to the primary key of table t', which we will denote with $t \rightarrow t'$ *to emphasize that one tuple of t determines one tuple of t';*

3. $F \subseteq (T_S \times (T_H \cup T_L)) \cup (T_L \times T_H)$*;*

4. exactly one arc exits from each satellite $s \in T_S$ *(entering a hub or a link);*

5. at least two arcs exit from each link [14].

Automation of data warehouse creation based on Data Vault approach, by using data model description and database metadata and user defined rules, is introduced as an algorithm in the scientific paper of Krneta, Jovanovic, and Marjanovic [12]. Algorithm for direct access to physical design of data warehouse formalizes, generalizes and largely automates the process of physical design for data warehouses based on data model description of database. This includes structured, semi-structured and non-structured databases. The paper introduced an agile solution for design of high in volume data warehouses based on a Data Vault approach. The algorithm which is introduced is intended for direct, but incremental design of data warehouse using data model description, metadata of the model and user defined rules.in the paper presented developing a prototype case tool for implementing Data Vault based data warehouse [12]. The solution presented fully addressed the challenged of creating a permanent solution which is flexible and immune to changes in volume and environment parameters.

III. Electricity Market, Supply and Distribution

Bosnia and Herzegovina has been an open electricity market since January 1st, 2015. Since then, all buyers have a free choice of choosing an electricity supplier. Supply of the market with electricity is done by energy stakeholders who are registered for trade and supply of the Bosnian market with electricity [15].

Supplier is a name for a utility provider with the permit of electricity supplying activity to tariff buyers and an entity with the permit of trading and supplying electricity issued by regulatory body. Public supplier is a utility provider supplying small buyers and households that had not chosen a supplier with electricity. Supply is done in a completely regulated way and with regulated prices. Distributor (distribution system operator) is a utility company focused on the activities related to distribution network. Distributors has a responsibility of enabling all buyers to connect to the distribution network and to transfer electricity from the network to the buyers [15].

With the opening of the electricity market, buyers of electricity have a right to choose a supplier and a price directed by the market. All countries in the region have opened the market of the electricity for legal trade, and have done the division of distribution and supply (or they are in the process of division).

In the future, utility companies are going to be facing increased market complexity, which will be harder to deal without bigger strategic and operational changes. Since currently market liberalization is ongoing. Utility companies need to develop a complete market and corporate strategy. In order for a utility entity to answer to all challenges on the market, it is necessary for them to have suitable tools and means as support in business processes of electricity supply in domain of strategic, tactical and operational decision making

This paper explains a method of designing a data warehouse based on data vault concept in supply of electricity.

IV. An Approach to Data Warehouse Design Based on Data Vault in Supply of Electricity

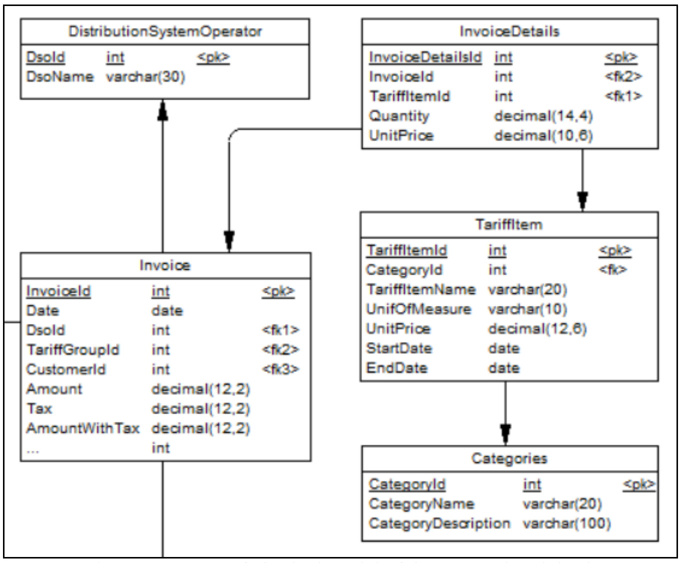
Data warehouse, as a part of data engineering needed for creating decision support system in a company is a crucial basis for implementation of a solution for supply of electricity.

Common methods for integration of data consist of pruning, deriving data from databases, which usually takes a lot of physical, manual work and cooperation of experts in the domain, due to semantics and business rules. Data Vault concept avoids traditional approach of pruning, by using satellites for keeping all data, without reconstruction, which is necessary for enterprise systems. Further research is focused on adding additional Links for possible merging of Data Vaults in the future, as stated in [12]. The before mentioned advantages of Data Vault are crucial in the case of the electricity supplier who receives the data from several different distribution system operators, emphasizing the ability to change the structure, at different time intervals. For the illustration purposes, a part of transactional database will be used. Focus will be on the distribution system operators, bills, categories and tariff elements. A part of database model of the enterprise operation system database is shown in Picture 1.

To transform the relational into a Data Vault model, Physical Data Vault (PDV) will be used, a design automation algorithm of a data warehouse introduced in [12]. Rules, which are defined form user and from the metadata are used by the algorithm for data warehouse design. A general algorithm for identification of Link, Hub and Satellite tables based on mapped original data is described by the following [12]:

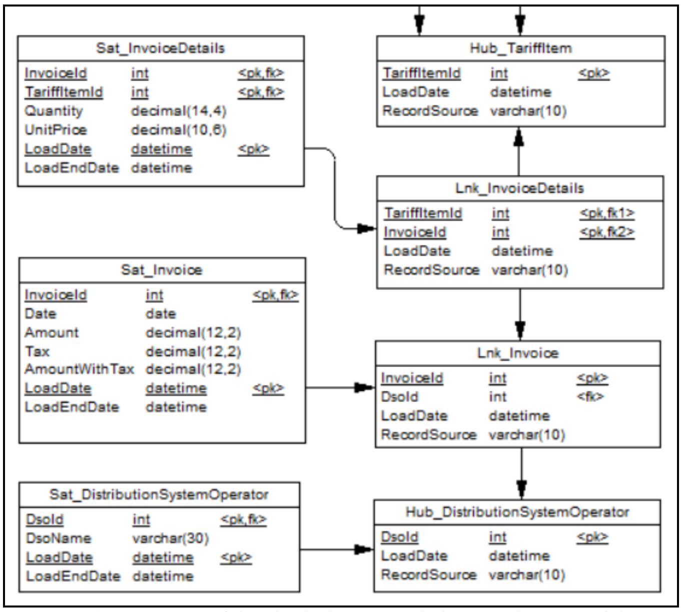
```
I- For each source:
  a) For each Table: the user confirms Hubs by selecting a Business Key
  b) For each Table, if not already selected as a Hub, create Link
  c) For each Table:
      i. for each FK create Link
  d) for each non-key attribute: create Satellite
```

II- For each Hub (PK)
a) For each source
i. For each Table (search for matching PK): if found, create Link (to Hub ad II; this only illustrate possible integration).



Picture 1. A part of physical model of the transactional database

In the model from the picture 1, the following business keys have been identified: Name of the distribution system operator (DsoName), name of the category (CategoryName) and tariff item name (TariffIItemName). After identification of business keys, with the use of previously mentioned algorithm, with appropriate tools, Hub tables are identified and relevant attributes. After that, Satellite and Link tables are identified as well as their relevant attributes. Part of the physical model of the data warehouse is shown in the picture 2.



Picture 2. Part of the physical Data Vault data warehouse model

After creating physical model and data warehouse, conditions for Extraction, Transformation and Load (ETL) process from transactional database into data warehouse are met. There is also an alternative to ETL processes, which is the ELT (Extract, Load, Transform) process. In this case, after extraction, data is first loaded into the data warehouse and then transformed. The ELT approach is most commonly used in the case of large data warehouses. In this way, we obtain a permanent and complete audit system of records in the data warehouse.

The Data Vault approach in creating a Data Warehouse represents a solution for complete manipulation and analysis of data, which includes integration, storage, and protection of data. On the other hand, this infrastructure is not suitable for querying, reporting or further analysis. That is the reason why a Data Vault is upgraded with data mart layer using star or snowflake schemas intended for reporting causes [16]. Data marts are often designed in a star or snowflake schema that consists of a single measure table and the set of dimension tables all of which are associated only with the fact table and, in some situations, the particular dimension of the relation resulting in the snowflake schema [2] [17] [18].

Measures or facts are key sources of statistical outputs, which are intended for decision-making process. Usually, records are result of events happened in an enterprise. A fact is usually represented in Entity-Relation diagram as entity F or *n*-ary relationship between entities E1,..,En. Later development of the model had taken n-ary relationships and transformed them into an entity F with a binary relationship between F and Ei, with cardinality of (1,1) - (m_i,M_i) where $m_i \in \{0,1\}$ and $M_i \in \{1,n\}$ are the minimum and the maximum multiplicity of branch Ei. The attributes of the relationship become attributes of F; the identifier of F is the combination of the identifiers of E_i, i=1,..,n [19].

The definitions in [19] state that a good nominee for measuring entities the ones that are often changed and that can represent a transaction of the system. Such entities, can be described as most similar to invoices, their items or order items. Dimensions determine the granularity adopted for the presentation of the facts in the decision-making process. Dimension tables are used in a denormalized form and they use facts of interest in the analysis. In most of the areas mentioned, the most commonly used is the time dimension, with different levels of detail (year, quarter, month, day).

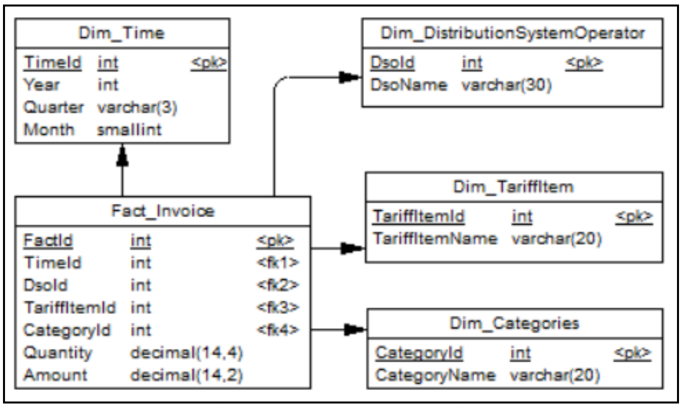
In the case of the Data Vault data warehouse model, data mart model is created in a way that dimensions of the star schema are created from Hubs and Satellite tables combined, and the fact tables are created from Satellite and connected Link tables [20] [21]. To transform the Data Vault into a data mart model, a Physical Data Mart (PDM) design automation algorithm will be used. A proposition for this algorithm is presented in [20]. The algorithm is consisted from data mart creation rules. A general algorithm for identifying Fact and Dimension tables and their relations based on the mapped Data

Vault model is described by the following. Also, by taking into account that the time dimension is the only dimension that is almost always there in every data mart, the algorithm consists of the following steps [20]:

```
(i) For each Data Vault table
    (a) For each Link table
        • For each Satellite of Link: user confirms fact column by selected business fact
    (b) For each Hub table
        • For each business key column of Hub: system confirms dimension by selected business key
        • For each Satellite of Hub: user confirms additional dimension by selected other Satellite attribute(optional)
(ii) Creating and filling time dimension table
(iii) For each confirmed fact
    (a) User determined aggregation of facts
```

The automation process is based on rules for creation of data mart and metadata model. Only condition for creating a data mart from the Data Vault model is existence of relationships between entities. An individual data mart can be created from conceptualization of model metadata.

By choosing appropriate business measures, such as in this case quantity and unit price of electric energy, according to the abovementioned algorithm, with the appropriate tool, dimensions, measures and attributes are identified. Physical model of generated data mart is shown on the picture 3.



Picture 3. Data mart in star schema

After creating physical model of data mart, conditions are met for extraction, transformation and load of data from data warehouse into the data warehouse based on measures and dimensions concept.

V. Conclusion

In this paper, a design of a data warehouse was shown, focusing on applying Data Vault concept on part of the relational model of the enterprise transactions from transactional database in the area of electricity supply activities with the focus on distribution system operators, categories and tariff elements. Also, a proposition for designing a data mart using automation algorithm based on previously created data warehouse is shown.

Data Vault approach in creating data warehouse allows separation of permanent data from time sensitive, transaction-based data into Hubs and Links, and separation of attributes into Satellites, which is very important for the electricity supplier who receives the data from several different distribution system operators at a different time interval. The implementation should provide better support for business decision making in energy companies in the process of market supply with electricity on the open energy market.

References

[1] B. Lazarević, Z. Marjanović, N. Aničić, S. Babarogić, Baze podataka, Fakultet organizacionih nauka, Beograd 2008.

[2] W. H. Inmon, Building the Data Warehouse. Wiley Computer Publishing, 1992

[3] D. Krneta, D. Radosav, B. Radulovic, "Realization business intelligence in commerce using Microsoft Business Intelligence", 6th International Symposium on Intelligent Systems and Informatics (SISY), pp. 1–6, 2008..

[4] B. Larson, Delivering Business Intelligence with MSSQL Server 2008, Mc Graw Hill, 2009.

[5] W. H. Inmon, Building the Data Warehouse: Gettiing Started, White Paper BillInmon.com, 2000.

[6] W. H. Inmon, D. Strauss D, G. Neushloss, DW 2.0 The Arcitecture for the Next Generation of Data Warehousing, Morgan Kaufman, 2008.

[7] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, Wiley, 2002.

[8] J. Mundy, W. Thornthwaite, R. Kimball, The Microsoft Data Warehouse Toolkit, Second edition, Wiley, 2008.

[9] D. Linstedt, Data Vault Modeling & Methodology, http://www.learndatavault.com, 2011.

[10] D. Linstedt, Super Charge your Data Warehouse, Kindle Edition, 2010.

[11] V. Jovanović, I. Bojicic, "Conceptual Data Vault Model", Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA, March 2012.

[12] D. Krneta, V. Jovanović, Z. Marjanović, "A Direct Approach to Physical Data Vault Design", Computer Science and Information Systems, Vol. 11, No. 2, pp. 569–599, 2014.

[13] C.J. Date, H. Darwen, N. Lorentzos T, Temporal data and the relational model: A detailed investigation into the application of interval and relation theory to the problem of temporal database management. Morgan Kaufmann Publishers, Amsterdam 2002.

[14] M. Golfarelli, S. Graziani, and S. Rizzi, "Starry vault: Automating multidimensional modeling from data vaults", ADBIS, p.137–151, 2016.

[15] https://ers.ba/wp-content/uploads/2019/11/Katalog-za-krajnje-kupce-elektricne-energije.pdf

[16] M. Casters, P. Bouman, J. van Dongen, Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration, Wiley Publishing, 2010.

[17] H. W. Inmon, Building the Data Warehouse, Wiley Computer Publishing, 1992.

[18] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, Wiley, 2002.

[19] M. Golfarelli, D. Mario, S. Rizzi, "Conceptual Design of Data Warehouses from E/R Schemes", System Sciences, 1998.

[20] D. Krneta, V. Jovanović, Z. Marjanović, "An Approach to Data Mart Design from a Data Vault", Infoteh, Jahorina, 2016.

[21] S. Govindarajan, Data Vault Modeling The Next Generation DW Approach, 2010.