Data Vault based system catalog for NoSQL store integration in the Enterprise Data Warehouse

Extended Abstract†

Katerina Černjeka
University of Rijeka,
Department of informatics
P.O. Rijeka 51000
Croatia
kcernjeka@inf.uniri.hr

Danijela Jakšić
University of Rijeka,
Department of informatics
P.O. Rijeka 51000
Croatia
dsubotic@inf.uniri.hr

Patrizia Poščić
University of Rijeka,
Department of informatics
P.O. Rijeka 51000
Croatia
patrizia@inf.uniri.hr

ABSTRACT

Our research is in its initial phase and this poster presents a general idea of our current work. A new comprehensive data warehouse (DW) architecture integrating different data sources (relational databases and NoSQL stores) is presented. According to proposed DW architecture, a new conceptual meta-data vault model of DW system catalog will be introduced.

KEYWORDS

NoSQL; data warehouse; meta-data; Data Vault; system catalog; data warehouse architecture

1 INTRODUCTION

Nowadays, it is not enough to store and analyze only structured data stored in relational databases but also all the other data (like semi-structured and unstructured data) from which valuable corporate values could be extracted. Relational databases are not fully adapted for storing unstructured and semi-structured data. That is why NoSQL stores are used, with its different store types and ability to store big datasets of semi-structured and unstructured data. In this context of data heterogeneity, our DW system will provide not only the ability to query relational and NoSQL data but to integrate, store and preserve history of all the corporate data and their changes into a single system of records. The reason why we've chosen a DW, instead of a data lake is because a DW is a more sustainable solution in field of data governance and auditing.

2 RESERACH DETAILS

We aim to answer the following research question: *can our system catalog, built upon the proposed DW architecture, be able to track and store changes of data (and metadata) from relational source and NoSQL source through the whole DW architecture, and by doing so serve as a basis for data auditing and data governance?*

The current system catalog from [2] tracks the origin and history of changes of relational data sources through the whole DW architecture. It provides basis for fast and simple migration, integration and transformation of relational data without the loss of information [3], [4]. We aim to extend the current system catalog in order to integrate NoSQL data sources with relational ones in Data Vault (DV) [5] based Enterprise data warehouse (EDW) and by doing so track the origin and history of changes for both, data and metadata of NoSQL stores and relational databases, as well as their schemas. The proposed DW architecture is based on a three-layer type DW architecture [6] and consists of the following parts: *data sources*, *DV based central EDW and data marts*. D*ata sources* include relational databases and NoSQL stores. *DV based central EDW* consists of a single DV model, partially oriented towards the data sources side as *raw data vaults (RDV)*, and partially oriented towards reporting side as *business data vault (BDV)*. RDV contain unchanged copies of the originals and BDV is created by updating and consolidating this raw data with application of business rules and standardized master data. Since copies of the originals are permanently kept in RDV there is no loss of information and the basis for audit process is created. Because of the separation between RDV and BDV we distinguish reversible and irreversible transformations which allow us to track data back to its source and reconstruct them, if necessary. This is the key idea for getting an integrated central EDW system of records and basis for data governance. Finally, the meta-data vault (MDV) is the basis for our extended system catalog. It integrates meta-data from all layers of DW architecture with an emphasis on integration of RDV and BDV in the EDW. Finally, our research has 3 main contributions: a) new integrated DW architecture, b) meta-data vault model of EDW system catalog for relational and NoSQL data sources, and c) an EDW system catalog prototype that stores metadata of both, relational and NoSQL data sources, as well as their schemas.

REFERENCES

[1] R. Kimball, *The Data Warehouse Lifecycle Toolkit*. New York: Wiley, 2008.

[2] D. Jakšić, "Metadata repository model for data warehouse schema evolution and integration with master data management system," Sveučilište u Rijeci, 2016., doctoral thesis

[3] D. Subotic, V. Jovanović, and P. Poščić, "DATA WAREHOUSE AND MASTER DATA MANAGEMENT EVOLUTION – A META DATA VAULT APPROACH," *Issues Inf. Syst.*, vol. 15, no. Ii, pp. 14–23, 2014.

[4] D. Jakšić, V. Jovanović, and P. Poščić, "Integrating Evolving MDM and EDW Systems by Data Vault Based System Catalog," in *MIPRO 2017*, 2017.

[5] D. Linstedt and O. Michael, *Building a Scalable Data Warehouse with Data Vault 2.0*. Todd Green, 2016.

[6] M. Golfarelli and S. Rizzi, *Data Warehouse Design, Modern principles and methodologies*. The McGraw-Hill Companies, S.r.l.-Publishing Group Italia, 2009.