RESEARCH PAPERS
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2016 Volume 24, Number 39

DATA STORING PROPOSAL FROM HETEROGENEOUS SYSTEMS INTO A SPECIALIZED REPOSITORY

Andrea VÁCLAVOVÁ, Pavol TANUŠKA, Ján JÁNOŠÍK

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA,
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA,
INSTITUTE OF APPLIED INFORMATICS, AUTOMATION AND MECHATRONICS,
ULICA JÁNA BOTTU 2781/25, 917 24 TRNAVA, SLOVAK REPUBLIC
e-mail: andrea.vaclavova@stuba.sk, pavol.tanuska@stuba.sk, jan.janosik@stuba.sk

Abstract

The aim of this paper is to analyze and to propose an appropriate system for processing and simultaneously storing a vast volume of structured and unstructured data. The paper consists of three parts. The first part addresses the issue of structured and unstructured data. The second part provides the detailed analysis of data repositories and subsequent evaluation indicating which system would be for the given type and volume of data optimal. The third part focuses on the use of gathered information to transfer data to the proposed repository.

Key words

Hadoop, Data Lake, Data Warehouse, Unstructured data

INTRODUCTION

The need for structured data storage and processing associated with typically used volumes nowadays can be addressed in many ways that are sufficient under certain circumstances. The problem arises with increasing volumes of data and when structured and unstructured sets of data get mixed.

Today, we face a huge increase in the amount of data in all sectors. This increase is mainly due to the introduction of new automated machinery and Internet of Things based technologies enabling direct communication with higher levels of management. Structured data is clearly defined and easy-to-manage, unlike the unstructured data that is more difficult to process. Analysis of proposed systems for the storage and processing of Big Data tries to determine which of them is most suitable for data collection at higher levels, subsequent processing and storage in data warehouses. Standard relational databases are no longer sufficient for such quantities and types of data. Therefore, we will analyze three data warehouse systems to find out which one would be most appropriate for storing large quantities of structured as well as

unstructured data gathered from various management levels being able to manage such quantities and types of processed data.

This paper describes the issues of storing structured and unstructured data and analyzes selected storages such as a Data Warehouse, a Data Lake and Hadoop. Based on the results we will define the most appropriate repository for storing data from heterogeneous systems at different levels of management.

First we introduce what is the structure of data we are working with. Big data is often characterized by an extreme volume of data, a wide variety of data types and the velocity at which the data must be processed. Although, big data doesn't mean any specific volume of data, the term is often used to describe terabytes, petabytes and even exabytes of data captured over time (1).

If the company did not address the processing of data that is collected, their volumes would have grown enormously. This condition will cause problems, if the company needs to obtain specific information in the short term. Software is primarily designed to deal with archiving and scanning, however, the main objective is always to ensure that the processing and management of data is a feasible task (2).

Before we can begin to find out which storage system is best suited for storage, we will describe the difference between structured and unstructured data. Structured data is usually presented as a text file where information is displayed in the named columns and rows. They have a relational key and can be easily mapped into per-designed fields. Such a construction of the data appears to be ideal and therefore, it is effectively used, even if the structured data constitutes only 5 to 10% of all the information data (3).

Unstructured data represents up to 80% of the data. It is raw, unorganized information that companies are saving. With text and multimedia content, we are talking about photos, videos, recordings or e-mails. It may seem that these types of files have an internal structure, yet we consider them unstructured because the data they contain, by their very nature does not correspond with data in databases (3).

By combining structured and unstructured data, companies can create more insight in their existing data. The best for the company is therefore to take advantage of both structured and unstructured data simultaneously.

SOLUTION PROPOSAL

After summarizing the differences between structured and unstructured data, the analysis of individual data stores follows. We analyzed three systems: the Data Warehouse, Data Lake and Hadoop.

The Data Warehouse is a core of the architectural environment and keystone of all DSSs. It is subject-oriented, integrated, non-volatile and has a time-variant collection of data in support of management's decision. The Data Warehouse contains granular corporate data. Data is entered into the data warehouse in such a way that many inconsistencies are undone at the application level. A Data Warehouse has a multi-level structure. It consists of old details, current details and partially and fully aggregated data (4).

If we want the Data Warehouse to process unstructured data, then the data has to be restructured. This process can be highly demanding. Nevertheless, there are methods to process unstructured data. For example, by setting up the medium for the storage volume as near-line storage. We can link structured and unstructured data by indexing. Another method is to create two storage sites while one of them contains all non-structured and the other having a capability of containing a certain set of data only (4).

The next analyzed repository is the Data Lake which unlike a Data Warehouse supports data storage in the original, unprocessed form. It can be seen as a unified repository that captures

data of any size, of any type and speed without the need for pre-processing data. The resulting data may not be used immediately but can be postponed indefinitely (5).

When using a Data Lake, you can only use what you really need not considering the rest of the information. So, each domain development can proceed independently. Users across the company can view data across the company. They are not limited only to data relevant to them or an immutable, fixed scheme. On the basis of the analysis we can assess which of the two storage systems would no longer comply with the specified requirements (5).

The Data Warehouse receives pre-treated data, having a form and structure unlike the Data Lake which is capable of storing the raw data, i.e., data that has not been pre-processed in such a form as required by the Data Warehouse. While processing the data, the Data Warehouse accepts only the written scheme - data in the correct form and structure while the Data Lake generates the correct data structure only when it comes to its use. This is the approach of the scheme for reading (6).

Another factor is the cost. A Data Warehouse is expensive when storing large amounts of data, whereas a Data Lake is designed for such storage and therefore, it is a low cost system. Regarding safety, for a Data Lake it is still under development therefore, the options currently associated with Data Warehousing are at a much higher level (7).

Based on this comparison we can state that a Data Lake is more suitable for processing of big data. Data based on the principle of Lake Hadoop provides a central data repository for raw data. The Data Lake collects and retrieves data from heterogeneous database systems and stores them in their original form. Therefore, there is no need to transform data to store them in a Data Lake. Figure 1 shows the inflow data from the ERM, MES and SCADA systems that can be processed in the Data Warehouse or in a Data Lake (Hadoop) and also the flow of unstructured data from autonomous systems and facilities which are processed exclusively in the Data Lake (Hadoop). And thanks to information obtained from processed data we can make various discoveries.

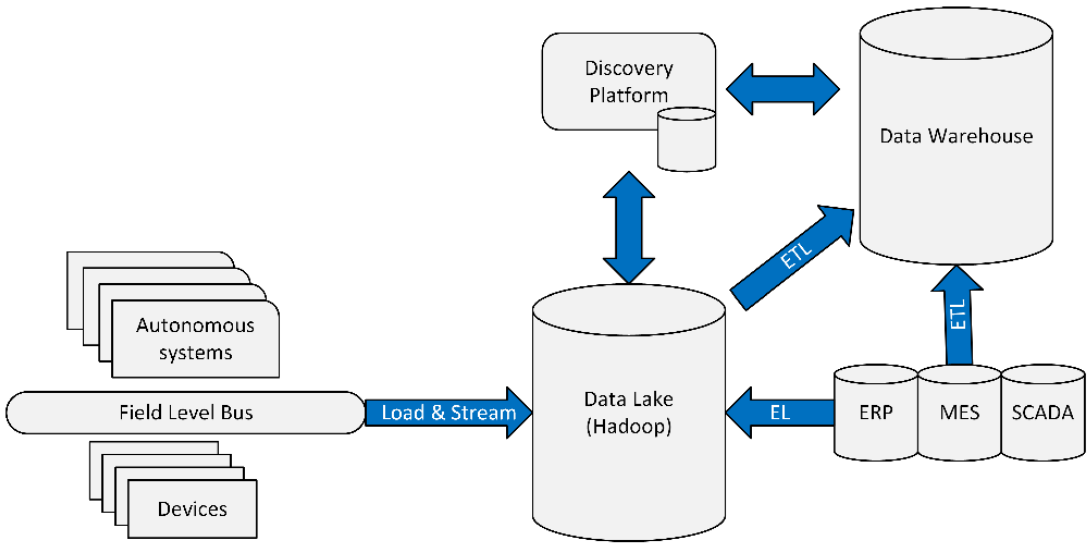

Fig. 1 *Coexistence of Data Lake and Hadoop*

The final analyzed repository was Hadoop. It is an open-source software framework for storing data and running applications on clusters of commodity hardware. With constant increases of volume and variety of data, Hadoop has an advantage of having the ability to quickly store and process vast amounts of data of any kind. The additional advantage would be

its computing power and fault resistance. Another feature of Hadoop is flexibility. Unlike traditional relational databases, you don't have to pre-process data before storing them. You can store as much data as you want and decide how to use it later. This includes unstructured data such as text, images, and videos. Since Hadoop is freely accessible the additional positive thing is that it is for free (8).

Hadoop works with four basic modules from the Apache Foundation. The first is the Hadoop Common which contains libraries and utilities used by other Hadoop modules. Second, Hadoop Distributed File System (HDFS) is the Java-based scalable system that stores data across multiple machines without prior organization. The third module is YARN (Yet Another Resource Negotiator) that provides resource management for processes running on Hadoop. Finally, a module called MapReduce. It is a framework for parallel processing software that works in two steps (Map Step a combination of responses when creating output) (8).

Entering data into Hadoop can be done in a few ways. Such as using third-party vendor connectors using Sqoop, importing structured data from relational database to HDFS, Hive and HBase, using Flume to continuously load data from logs into Hadoop, loading files to the system using simple Java commands or using HDFS as a file system and copy or write files there (8).

Based on analysis, the Hadoop seems to be the most suitable system. Therefore, we attempted to import data into the Hadoop environment. First, the data was exported from the Microsoft Dynamics database. Afterwards the data was stored in the correct format with correct delimiter characters. We subsequently checked the exported data in order to find out if the export was completed without damage. Figure 2 shows the BPMN model of data export.

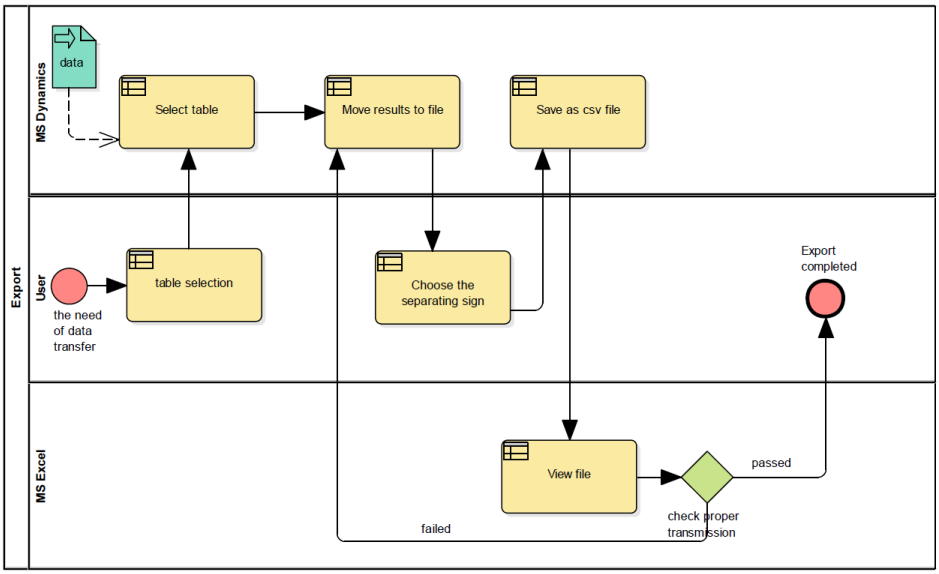

Fig. 2 *Model proposal of data export*

We imported data to the Hadoop environment and ensured the import worked properly. Since the imported data was intact, in correct types, the table was already available within the Hadoop environment and we were able to work with it. Figure 3 shows the process of importing data to the Hadoop environment.

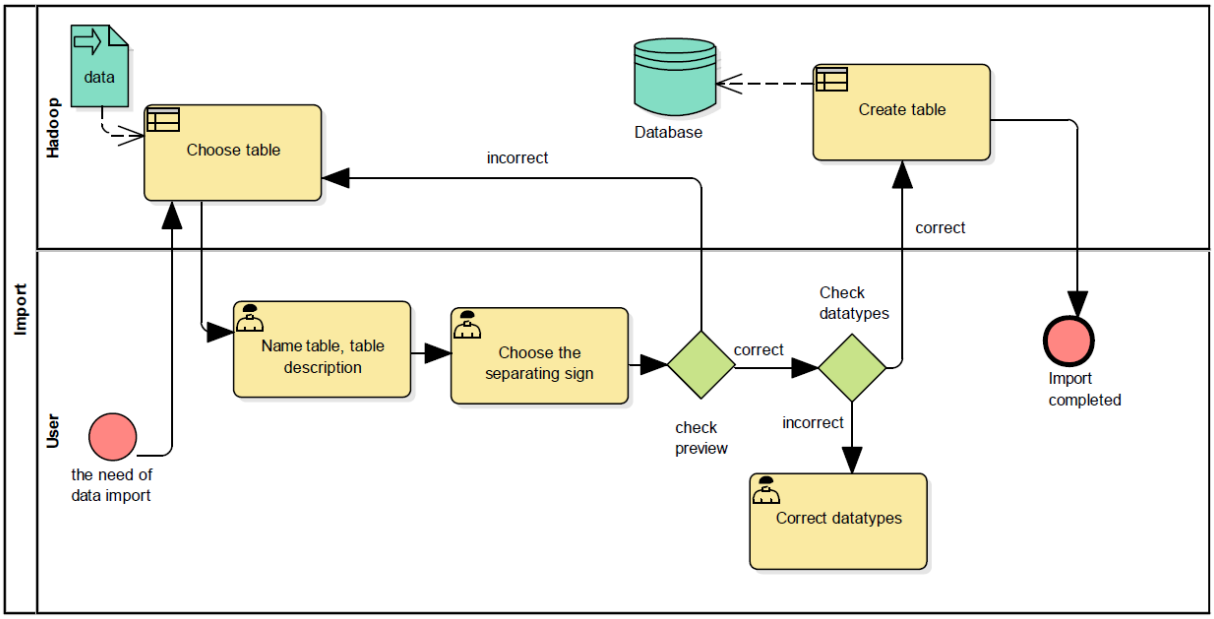

Fig. 3 *Model proposal of data import*

CONCLUSION

The aim of the work was to clarify the issues of Big Data which means structured and unstructured data and based on analysis to propose the data store that would be most suitable to store this type of data.

Traditional relational databases that do not work with unstructured data are not sufficient for the processing of structured and unstructured data. In the case that we would like to enter this type of data in such a database it would be necessary to transform that data into structured data before loading. Therefore, we chose to analyze data storage that can handle unstructured data.

First we analyzed the Data Warehouse. Although it can work with unstructured data it works slowly and intricately. The Data Lake is another analyzed system which handles processing unstructured data relatively quickly. However, Hadoop is more appropriate. The Data Lake is also used within Hadoop. Hadoop works swiftly and it is a flexible and reliable system. As far as processing and storing of structured and unstructured data of huge volumes is concerned Hadoop is the most appropriate of all these systems.

Following the work, we completed the import of data from the MS Dynamics environment into Hadoop. The import was performed without issues and data were transferred without any damage.

Acknowledgement

This publication is the result of implementation of the project VEGA 1/0673/15: “Knowledge discovery for hierarchical control of technological and production processes” supported by VEGA.

This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

This publication is the result of implementation of the project: “Research into monitoring and assessing the non – standard states in the vicinity of a nuclear power plant” (ITMS:

26220220159) supported by the Research & Development Operational Programme funded by the ERDF.

References:

1. ROUSE, M., 2014. big data, Available at: <http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data> (Accessed on 27th September 2016)
2. SHERPA SOFTWARE. 2015. *What's the Difference Between Structured & Unstructured Data?* Available at: <http://www.sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/> (Accessed on 27th September 2016)
3. RONIK, J., 2014. *Structured, semi structured and unstructured data*. Available at:< https://jeremyronk.wordpress.com/2014/09/01/structured-semi-structured-and-unstructureddata/> (Accessed on 27th September 2016)
4. INMON, W.H., 2002. *Building the Data Warehouse Third edition*. Canada: John Wiley & Sons, Inc. ISBN 0-471-08130-2
5. STEIN, B. and MORRISON, A., 2015. *Data Lakes and the promise of unsiloed data.* Available at:<http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/features/data-lakes.html> (Accessed on 27th September 2016)
6. DULL, T., 2015. *Data Lakes vs Data Warehouse: Key Differences*. Available at:< http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html> (Accessed on 27th September 2016)
7. CITO RESEARCH. 2014. *Putting the Data Lake to Work*. Available at:< https://hortonworks.com/wp-content/uploads/2014/05/TeradataHortonworks_Datalake_White-Paper_20140410.pdf> (Accessed on 27th September 2016)
8. SAS. 2015. *Hadoop. What it is and why does it matters*. Available at:< http://www.sas.com/en_us/insights/big-data/hadoop.html> (Accessed on 27th September 2016)