Review on Intelligent Data Engineering for Migration to NoSQL Based Secure Environment

Samruddhi Pund[1] Prof. V. M. Lomte[2]
[1]Student [2]Head of Department
[1,2]Department of Computer Engineering
[1,2]R.M.D.Sinhgad School of Engineering, Warje, Pune, Maharashtra, India

Abstract— In the increasing technology of today's world, the data size is increasing exponentially, with variety of data being generated. Such an exponential growth of massive form of structured and unstructured data is termed as "Big Data". The efficiency of handling of big data is increased by preferring big databases like Oracle, MongoDB, Cassandra, Neo4j and NoSQL. NoSQL provides scalability, high performance, data modeling and distribution, with continuous availability and overcomes the limitations of the relational database. It appears as challenge to many organizations to convert their existing database to NoSQL databases by preserving its complexity, integrity and heterogeneity. Along with big data comes the data cleansing operation. Here we apply an approach, which consist of two modules. The first module uses data transformation to convert relational database to Oracle NoSQL database through model transformation. The second phase consist of data cleansing operation. This proposed approach successfully converts relational database to Oracle NoSQL and quality of data is being improved due to data cleansing.
Keywords: Migration, Oracle, NoSQL, Database, Data Cleansig, Model Transformation, Relational Database
Motivation: The main aspect is to handle large amount of variety and exponentially increasing data. This can be achieved by converting our traditional database systems to NoSQL database systems.
Which is basically available, soft state and eventually Consistent. It allows storage of structured, semi-structured And unstructured data. Authors provides a transformation Model method to convert the relational database into Oracle NoSQL database.

I. INTRODUCTION

Initially, relational databases prove to provide excellent performance in storage of data until the data became vast.
Even though any form of unstructured data came up, it was handled by ETL(Extract Transform and Load) tools. But now it has became nearly impossible for relational databases to handle the volume, velocity and variety of data and to store it efficiently and retrieve back.

The big data along with cloud computing proves to be great challenge for relational database. It is nearly impossible for relational databases to access petabytes of data.

The data growth was observed significantly after 1970, when there was increase of data from gigabytes to megabytes and later organizations were holding terabytes of data. Firms in today's world carries petabytes or even exabytes of data. To handle such a vast amount of data, NoSQL proves to be efficient.

NoSQL can handle structured, semi-structured unstructured forms of data. It provides horizontal scalability

to store the data. Data analysis proves to be a challenge to the big data organizations and firms. In case of relational databases, the data is first preprocessed and then stored. But, in NoSQL the data is stored first and then processed.

The main aim is to present an automated approach for conversion of existing relational database to NoSQL database and apply data cleansing technique to improve data quality. The approach uses model to model transformation to convert SQL server database to Oracle NoSQL database, for this transformation metamodels of both source and target databases is required.

The metamodel of relational database is available but the metamodel of NoSQL is not available, so we propose an initial version of the Oracle NoSQL metamodel. NoSQL can handle data more efficiently and do fast data processing as compared to relational database. As relational database consist of vertical scalability it is unable to manage huge growth of data.

The transformation of existing relational database to NoSQL database is a big challenge, automatic migration to NoSQL database can be achieved by exploiting the data and query characteristics of relational database.
Technology used in the system are:
1) Relational Database
2) Oracle NoSQL
3) Big Data
4) Data Cleansing

Brief Description about Basic concepts about relational databases, Oracle NoSQL,Big data and data cleansing:

A. Relational Databases

The basic building block of relational databases is rows (tuples), columns (attributes), and tables (relation) and each row is unique. There are two main keys in relational databases, they are primary key and foreign key.

One of the main concept of relational databases are ACID properties. They are Atomicity, Consistency, Isolation and Durability that ensures liable transaction has been taken place.

The data query language used in relational database is structured query language.

Data can be accessed by using relationships and stored by using relationships. These relationships are in the form of "one to one", "one to many" and "many to many". These are known as entity relationship.

B. Big Databases

As the data has been increasing exponentially since 1970, and variety of data is observed in this huge data.

The data storage, data processing and data analysis requires more proficiency. Such a massive growth of structured and unstructured data is termed as "Big Data".

C. Oracle NoSQL

In 1998, Carl Strozzi introduced "NoSQL". It consists of column store, document store, graph store and key-value store.

Oracle NoSQL consist of key-value store which provides horizontal scalability, monitoring, transactional semantics for improved data manipulation, and simple administration of data.
The Oracle NoSQL database is schema free.

D. Data Cleansing

Data cleansing is an operation which involves removal of inaccurate, incomplete and unreasonable data.
Data cleansing consist of following steps:
1) To define and determine the error types.
2) Then search and identify the error instance
3) Then to correct errors
4) To identify the error instances and types
5) Then modify the data entry procedures and reduce errors.

II. METHODOLOGY

In this approach, two major modules are to be performed one after another.
It consists of:
1) Model Transformation
2) Data Cleansing

A. Model Transformation

In this approach, a model is provided as input to the transformation engine and output model is obtained and transformation rules are used by the engine for that purpose. For this transformation, SiTra engine is used.

A Model Driven Architecture is a model based approach which allows model to model transformation by mapping the concepts of input model with the output model.

The Oracle NoSQL database consists of tables which are similar to that of tables of SQL Server database, the data stored in the Oracle NoSQL is in the form of rows and fields which are equivalent to the SQL Server database as it stores in the form of rows and columns.
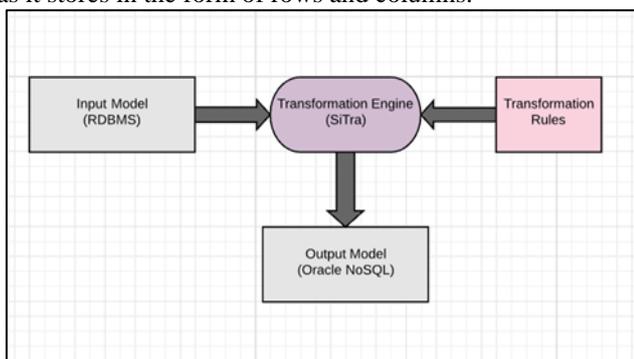

Fig. 1: Transformation Model

Transformation algorithm is used for implementation of transformer interface. The transformation algorithm is extended to keep the record of transformed objects and also to avoid duplicate creation of target object.

Then mapping of the relational database concepts to Oracle NoSQL takes place. SiTra engine uses transformation rules for transformation.

B. Data Cleansing

The data cleansing operation detects and removes the inconsistent errors and enhances the quality of the data transformation.

Data cleansing can be carried out by using the following steps:
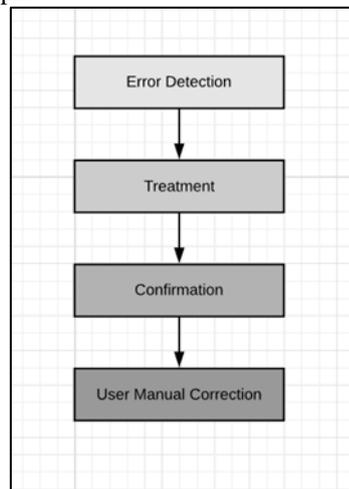

Fig. 2: Data Cleansing Steps

1) Error detection
The data which is incomplete, inaccurate and incorrect is detected using various techniques.
Steps involve in error detection are:
1) Parsing: To check the syntax error.
2) Clustering: The records that have abnormal distances are identified and clustering algorithm adjust the size of the cluster.
3) Duplicate identification: It identifies the duplicate records and makes comparison between the records.

2) Treatment
Whenever an incorrect, irrelevant and inaccurate data is detected then a decision is taken whether to leave the data unchanged, correct the data or delete the data. So the incorrect and incomplete data is remained as unchanged.
The duplicate data is deleted.

3) Confirmation
Correction confirmation is absolute necessary to ensure that the data is consistent, correct and relevant. After confirmation, we can add correction in the temporary file.

4) User Manual Correction
This step allows user to do manual correction in the data.
This step consist of:
1) Data Enhancement: Add more related information.
2) Data Harmonization: Place the actual code.
3) Data Standardization: Use standard codes.

III. LITERATURE SURVEY

Sr.No	Published By	Published Year	Paper Title	Research Gap
1.	T. Jia, X.Zhao Z.Wang, D.	2016	Model transformation and data	Long and hectic procedure

	Gong, G Ding		migration from relational database to MongoDB	which involves transformation and then migration[1]
2.	C.-H. Lee Y.-L. Zheng	2015	Automatic SQL-to-NoSQL schema transformation over the MySQL and HBase databases	Only 47% of access performance is improved[2]
3.	D. Serrano, D. Han, E. Stroulia	2015	From relations to multi-dimensional maps: Towards an SQL-to-HBase transformation methodology	Subset of migration problems are covered[3]
4.	H.Gonzalez, J. Han, and X. Shen	2007	Cost-conscious cleaning ofmassive RFID data sets	DBN-based cleaning method takes tag readings as noisy observations of a hidden state[6]
5.	Zhibin Guan Tongkai Ji Xu Qian Yan Ma Xuehai Hong	2017	A Survey on Big Data Pre-processing	Approach applies to social science research, it also has obvious drawback that its datasets must be given[5]
6.	Arati Koli Swati Shinde	2017	Approaches used in efficient migration from relational database to NoSQL database	Less efficient for small amount of data migration[7]
7.	Leonaro Rocha, Fernano Vale, Elder Cirilo, Darlintn Barbosa,and Fernano Mourao	2015	A Framework for Migrating Relational Datasets to NoSQL	Less efficient for small amount of data[8]
8.	Bogdan Walek, Cyril Klimes		Expert system for data migration between different database management systems	Some tools have problems with migration of foreign key. Tools cannot be modified or extended [9]
9.	M.Elamparithi , V.Anuratha	2015	A review on Database Migration Strategies,Techniques and Tools	Target database is systematically weaker and poorer quality than source[10]
10.	Virender Kumar, Cherry Kosla	2018	Data Cleaning – A thorough analysis and survey on Unstructured data	-

IV. COMPARISON

Comparison between RDBMS and Oracle NoSQL[11]:

Parameter	Relational DB	Oracle NoSQL
Query Language	It uses Structured Query Language	It uses Non-Structured Query Language
Schema	It has a pre-defined schema	It has dynamic schema
Scalability	Vertical Scalability	Horizontal Scalability
Big Data Support	Not efficient for Big Data	Very much efficient for Big Data Storage
Importance	Used when data validity is important	Used when data retrieval required is faster
Access Data	Through primary key and foreign key	Through key value pair
Storage property	Data is stored in tables	Document store, Graph store, key-value, Column store

The difference in the properties of both relational databases and Oracle NoSQL databases is shown in the below diagram[11]:
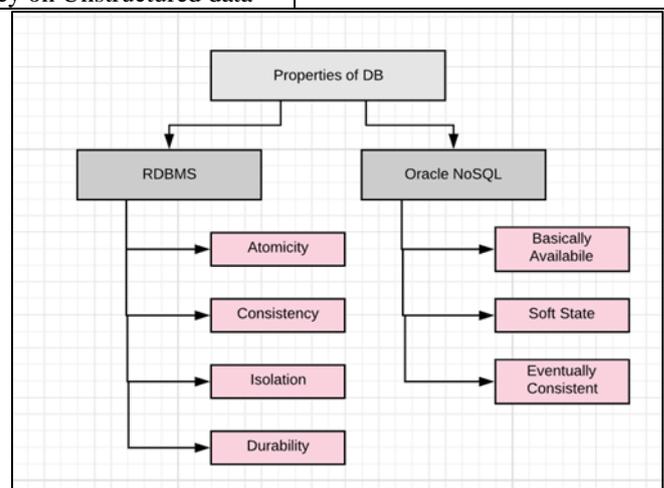

Fig. 3: Properties of RDBMS and Oracle NoSQL

V. CONCLUSION

This approach consist of two different modules that are Transformation and Data cleansing. The transformation module automatically converts SOL Server into Oracle NoSQL. For this transformation we need metamodel of both of them and transformation rules are applied and SiTra engine is used to execute the rules.RDBs have standard

metamodel but NoSQL databases have no metamodel therefore, we also proposed an initial version of the Oracle NoSQL metamodel.

Then data cleansing operation is carried out and corrections are stored in temporary file and manual corrections are also allowed and finally the file is uploaded in Oracle NoSQL database.

And all this carried out successfully by using the above modules.

REFERENCES

[1] T. Jia, X. Zhao, Z. Wang, D. Gong, and G. Ding, ''Model transformation and data migration from relational database to MongoDB,'' in Proc. IEEE Int. Congr. Big Data (BigData Congr.), Jun./Jul. 2016, pp. 60–67.

[2] C.-H. Lee and Y.-L. Zheng, ''Automatic SQL-to-NoSQL schema transformation over the MySQL and HBase databases,'' in Proc. IEEE Int. Conf. Consum. Electron.-Taiwan, Jun. 2015, pp. 426–427.

[3] D. Serrano, D. Han, and E. Stroulia, ''From relations to multi-dimensional maps: Towards an SQL-to-HBase transformation methodology,'' in Proc. IEEE 8th Int. Conf. Cloud Comput., Jun./Jul. 2015, pp. 81–89.

[4] Virender Kumar, Cherry Kosla, "Data Cleaning – A thorough analysis and survey on Unstructured data," in Proc. IEEE 8th International Conference on Cloud computing , Data Science & Engineering, 2018

[5] Zhibin Guan Tongkai Ji, Xu Qian, Yan Ma ,Xuehai Hong," A Survey on Big Data Pre-processing," in Proc. IEEE 2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD).

[6] H. Gonzalez, J. Han, and X. Shen, ''Cost-conscious cleaning of massive RFID data sets,'' in Proc. IEEE 23rd Int. Conf. Data Eng., Apr. 2007

[7] Arati Koli, Swati Shinde," Approaches used in efficient migration from relational database to NoSQL database," Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering

[8] Leonardo Rocha, Fernando Vale, Elder Cirilo, Darlinton Barbosa,and Fernando Mourao," A Framework for Migrating Relational Datasets to NoSQL," 2015 ICCS International Conference on Computational Science.

[9] Bogdan Walek, Cyril Klimes," Expert system for data migration between different database management systems," Advances in Data Network, Communication, Computers and Materials.

[10] M.Elamparithi, V.Anuratha," A review on Database Migration Strategies, Techniques and Tools," 2015 World Journal of Computer Application and Technonlogy.

[11] https://www.guru99.com/sql-vs-nosql.html