Challenges of Data Integration and Interoperability in Big Data

Anirudh Kadadi, Rajeev Agrawal, Christopher Nyamful, Rahman Atiq*
Department of Computer Systems Technology
North Carolina A & T State University, Greensboro, NC, USA
*University of Arkansas at Pine Bluff, Pine Bluff, AR, USA
akadadi@aggies.ncat.edu, ragrawal@ncat.edu, cnyamful@aggies.ncat.edu, rahmana@research-cs.org

***Abstract*— The enormous volumes of data created and maintained by industries, research institutions are on the verge of outgrowing its infrastructure. The advancements in the organization's work flow include data storage, data management, data maintenance, data integration, and data interoperability. Among these levels, data integration and data interoperability can be the two major focus areas for the organizations which tend to implement advancements in their workflow. Overall, data integration and data interoperability influence the organization's performance. The data integration and data interoperability are complex challenges for the organizations deploying big data architectures due to the heterogeneous nature of data used by them. Therefore, it requires a comprehensive approach to negotiate the challenges in integration and interoperability. This paper focuses on the challenges of data integration and data interoperability in big data.**

Keywords-Data Integration, Data Interoperability

I. Introduction

Evolution of large data sets from major industries is termed as big data in the field of data science. Big data can be classified as the large volumes of datasets with a higher complexity levels. Example:- A food product borne disease and its ties with the animal, weather circumstances, temperature, cattle food, etc. The data extracted from multiple sources can be used to analyze and gain meaningful insights; however, it requires data integration at various levels. Organizations like Facebook, Google, and Twitter tend to generate more than 500 terabytes of data each day. Given a scenario where two big organizations merge and tend to operate centrally, data integration and interoperability could be the major area of focus as the two organizations might have had different data management techniques before the merger and the data exchange involved is enormous. Data integration plays a key role in determining the efficiency of an organization, be it at the level of backend systems integration or integration of processes, administrative tasks, and databases [1]. The complexity of data integration and interoperability emphasizes on the levels of data storage, structure and the levels at which the data can be integrated and operated as a single entity. Collecting and maintaining the large data sets is costly, therefore organizations tend to adapt to cloud methodologies for storing the data and reuse [2]. KARMA, an open source tool, is a data integration tool which enables to integrate data from different sources like XML, text files, web Application Programming Interface (API's). Karma allows the users to integrate the information, select the GUI, combine the classes, allowing the user to transform the data in different formats, finally providing the user with desired integrated data [3].

The interoperability arises only after a successful integration. The data is to be integrated and optimized so as to ensure a smooth operation at each level of data integration. JNBridgePro tool serves as a good example for interoperability, where it bridges Java and .NET.

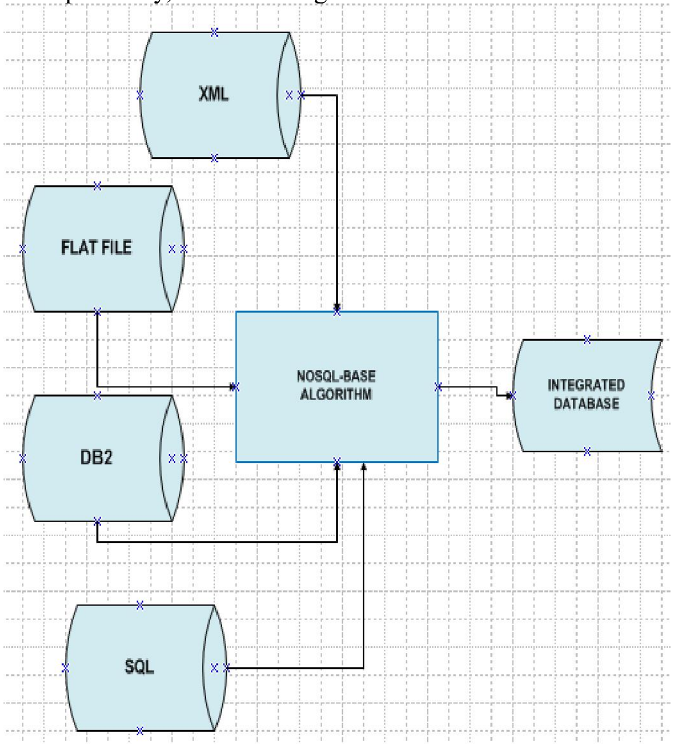

Figure1. Integration of different types of data using NOSQL algorithm.

II. RELATED WORK

The data integration challenges were proposed in the past which included query execution and algorithms to execute the query. Overall, data integration followed the query processing methodologies, which could involve the implementation of complex algorithms. This challenge has been addressed in the paper in [4], which proposed a new system for query execution. The lack of statistical data and analysis methodologies for evaluating huge data sets is a challenge in implementing date integration. This challenge has be addressed currently by analyzing the huge data sets

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

evolved from different sources like; flood statistics, weather data from numerous geographical locations, data from organizations like Facebook and Twitter. Evolution of the tools like tableau has eased the challenge in data analysis.

The low data transfer rate of communication systems may be a challenge in big data integration and interoperability due to the network constraints like bandwidth [4, 5]. This problem may be further compounded due to the packet loss and congestion in the network. This challenge is addressed by the evolution of IPV6 and use of high-end routers for connection establishment between the entities, implementation of 10 Gigabyte Ethernet and hybrid network systems.

III. Challenges of Big data integration and interoperability

The challenges of big data integration and interoperability are many. Some of the important challenges are listed below.

1) Accommodate scope of data:

The need to scale to accommodate the sheer scope of data and creation of new domains within the organization could be another challenge for interoperability, This challenge can be addressed by implementing the high performance computing environment and advanced data storage systems like hybrid storage device which has the features of hard disk drives (HDD) and solid state drives (SSD), featuring reduced latency, high reliability and rapid access to the data. This could accumulate large data sets from numerous sources. Finding the common operational methodologies between the two domains for integration and implementing the query operations and algorithms could help meet the challenges posed by the large data entities [1].

2) Data Inconsistency:

The data from heterogeneous sources could lead to inconsistency in data levels, therefore requires more resources to optimize the unstructured data. Structured data allows performing the query operations to analyze, filter and use this data for business decisions and organization capabilities [2]. In this scenario, where the large data sets are involved, unstructured data resides in higher volumes. This can be located using the tag and sort methods, allowing to search the data using keywords. Also the hadoop methodologies like MapReduce, Yarn which modulates the large data sets into sub divisions for ease in data conversions and schedule processes individually. Flume can be implemented for streaming of large data sets.

3) Query Optimization:

Query optimization at each level of data integration and mapping components to the existing or a new schema which could influence the existing, and new attributes. This challenge can be addressed by reducing the number of queries by using strings, joins, aggregation, grouping all the relational data. Parallel processing, where the asynchronous query operations are performed on individual threads, can influence the latency and response time positively. Implementing the distributed joins (hash, sort, merge) and determining the data which involves larger processing, consumes more storage resources, which depends on the type of data. Using the higher level query optimization techniques like data grouping, join algorithm selection, join ordering can be implemented to overcome this challenge [4].

4) Inadequate Resources:

Inadequate resources for implementing data integration, this refers to the lack of financial resources, lack of skilled professionals, implementation costs. Every organization must analyze their investment abilities in order to implement a new phase of work environment to their existing work system. Lack of financials is generally faced by the small segment organizations which are limited to only certain domain; Example: An organization limited to the consulting. In real time these organizations can deploy the changes in small intervals as this could fetch them additional time to gain back the invested resources. Lack of skilled professionals can not only slow down the projects but also demoralize the organization's abilities to handle projects. Skilled professionals in big data are hard to find as data integration requires high level of experienced professionals who in the past would have dealt with integration module. This can be curbed if organizations set up training modules for its employees. The implementation costs could be higher for implementation of data integration as this involves licensing new tools and technologies from the vendors. This cost could be shared when two different organizations are involved in the process of integration after the merger.

5) Scalability:

Scalability issues crop up when the new data from multiple resources is integrated with data from legacy systems. The organizational changes could influence the efficiency of legacy systems as it passes through many updations and modifications to meet the requirements of new technologies so as to be integrated with it. The mainframe, one of the oldest technologies from IBM has been integrated with big data tool hadoop turning it to a high performance computing experience. The mainframe's ability to accumulate large data sets and higher data streaming rate makes it adaptable to the new technology hadoop. Hadoop delivers the high performance computing experience by enabling the operations on huge data sets. The multidimensional feature of hadoop provides a better focus on data forms like unstructured, semi structured and structured data. Hive can be implemented as a query processing in hadoop which typically involves the steps: query language to web interface to JDBC ODBC to Metadata and finally the hadoop cluster**.** Overall, this approach provides a better insight for the organizations looking for large data storage systems integrated with big data tools [3, 6]. Figure 2 explains the integration between the legacy systems and the data from numerous sources.

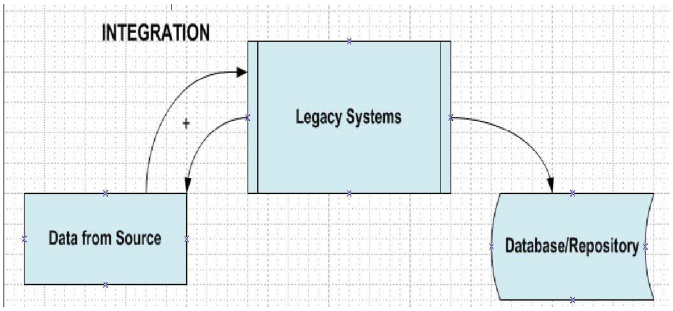

Figure 2. Data Integration between the new sources and legacy systems.

Data from sources is integrated with the data in legacy systems and configure the legacy systems to adapt with the new technology which runs the database after integration. Finally, the integrated data is stored into the database.

6) Implementing support system:

Organizations need to establish a support system for handling updates and error reporting in each step of data integration and this would need a training module for training the professionals to handle error reporting. This may require huge investment for an organization. To address this challenge, organizations should implement advancements in their work system to adapt themselves to the growing market trends. Implementation of Support system could help analyze the flaws in the existing architecture which could give them a scope for further updates or modifications. Although a high investment, this could still prove beneficial for organizations post the changes.

7) ETL Process in big data:

Each data item goes through the typical Extract Load Transform (ETL) process, therefore transforming into a huge data set on a whole after the integration, this could affect the data storage abilities of database. The ETL process is recommended for data integration as it ETL ensures a systematic and step by step approach for data integration. Extract performs the data retrieval from the sources which is its key feature. This operation does not affect the data sources. The data extraction is performed when there is an update or incremental notification from the data systems. In the cases, where it fails to determine the notification, the extraction is performed on complete data. This optimizes the data into a standardized form. The Transform step involves the implementation of a set of rules for transforming the data from source to the target. This also involves the data joining from all the sources, sorting, deriving values and applying the rules. Load applies the resources necessary before loading the data into the database and it ensures the minimal usage of the resources. The load process can be more efficient when the key constraints are disabled, and enabled after the process. Therefore ETL process is preferred for the data integration in large volumes [7].

The above mentioned challenges impact the traditional organizational practices for data integration and interoperability. If all these challenges are addressed, this could provide a good scope for organizations which integrate the domains within itself and also integrate at the levels of multiple organizations.

IV. Conclusion

We are analyzing the existing data integration and interoperability techniques and exploring their usage in big data scenario. We are working on to design a big data integration architecture that can handle the challenges as listed in section- III.

ACKNOWLEDGMENT

This research work was funded by the Office of National Consortium of Data Science (NCDS).

References

[1] M. Lenzerini, “Data Integration: A theoretical Perspective,” Proc. *twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,* 2002, pp. 233–246.

[2] Klischewski, R.; Scholl, H.J., "Information Quality as a Common Ground for Key Players in e-Government Integration and Interoperability," *System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on* , vol.4, no., pp.72,72, 04-07 Jan. 2006.

[3] Rattapoom Tuchinda, Pedro Szekely, and Craig A. Knoblock. 2007. Building data integration queries by demonstration. In *Proceedings of the 12th international conference on Intelligent user interfaces* (IUI '07). ACM, New York, NY, USA, 170-179.

[4] Zachary G. Ives, Daniela Florescu, Marc Friedman, Alon Levy, and Daniel S. Weld. 1999. An adaptive query execution system for data integration. *SIGMOD Rec.* 28, 2 (June 1999), 299-310.

[5] Knoblock, Craig A., and Pedro Szekely. "Semantics for Big Data Integration and Analysis." *2013 AAAI Fall Symposium Series*. 2013.

[6] Shvachko, K.; Hairong Kuang; Radia, S.; Chansler, R., "The Hadoop Distributed File System," *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on* , vol., no., pp.1,10, 3-7 May 2010.

[7] Panos Vassiliadis, Alkis Simitsis, and Spiros Skiadopoulos. 2002. Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*(DOLAP '02). ACM, New York, NY, USA, 14-21.