Big Data Mining and Data Integration: Techniques, Challenges, and Solutions

John A. Smith[1], Emily J. Wilson[2]
[1]Department of Computer Science, Massachusetts Institute of Technology, USA
[2]Department of Information Technology, Stanford University, USA

ABSTRACT

The paper discusses few of data mining techniques for data integration. Also Discuss challenges in data Integration of data in big data mining to improve their businesses and found excellent result. Data integration can be performed by several organisational level . Manual data , middleware data, Data warehouse etc. Benefit is that they have ability to easily manage history of data and also ability to combine data from very different sources. Some challenges visaged during its integration include uncertainty of data management ,big data mining talent gap, getting data into big data mining structure, reset access data sources

Keywords: *Big Data, data mining,techniques, challenges and Data integration.*

I. INTRODUCTION

Big data mining is data mining with big data.
Data Mining is an analytic process designed to explore data (usually large amounts of data - typically commertial, business or market related - also known as "big data"). Data mining is process of finding out pattern in large data set. Big data is often described in terms of the variety of data, the velocity with which it changes and the volume of data. It's been a big year for Big Data. As more businesses are accepting that data is intrical for decision-making. We continue to see the systems that support relational, non-relational or unstructured,,structured forms of data also massive data volumes are maturing to operate well inside of Bussiness Enterprise IT systems. Example Apache Spark

Data integration is the combination of technical and business processes used to combine data from different sources into meaningful and extremely useful information.

There is a need of Data integration when a business decides to implement a new application and migrate its data from the award systems into the new application. It becomes even critically important in cases of company mergers where two companies join together and they need to compact their applications

One of the most commonly known use of data integration is to make a data warehouse for a company which enables a business to have a unified view of their data for analysis and business intelligence (BI) needs.

A whole data integration solution tranforms trusted data from a variety of sources.
Benefit access to the unified governance and integration platform offerings by way of curve-point licensing. It supports your rapidly evolving business requirement by providing flexible access to the offerings included in the platform. Buying entitlement depend on your expected needs. When your business requirements change, gain the pliability to add or stop using a product and apply the same curv points to another offering within the platform, as long as you stay within total curve points buy.

Data Integration process is about taking data from many different sources (such as files, various databases, mainframes etc.,) and combining that data to provide a unified view of the data for business intelligence.

A broad term for large and intrical datasets are Big Data mining where traditional data processing applications are inadequate. The integration of this large data sets is quite intrical. There are several challenges one can face during this integration such as analysis, data curation, capture, sharing, search, visualization, information privacy and storage. The core elements of the big data platform is to handle the data in new ways as compared to the traditional relational database. Accuracy in managing big data will lead to more confident decision making. In this, we discuss the integration of big data and six challenges that can be faced during the process.

Characteristics of Data Integration

Data integration involves a framework of applications, techniques, technologies, and products for providing a unified and consistent view of enterprise business data (see Figure 1).

- *Applications* are custom-built and vendor-developed solutions that utilize one or more data integration products.
- *Products* are off-the-shelf bussines purpose solutions that support one or more data integration technologies.
- *Technologies* implement one or more data integration techniques.
- *Techniques* are technology independent approaches for doing data integration.

Data integration framework

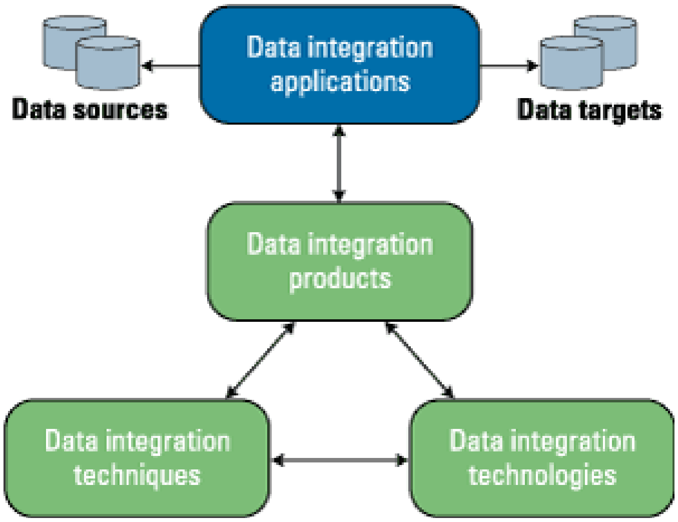
Data sources
Data integration applications
Data targets
Data integration products
Data integration techniques
Data integration technologies

Figure 1. Components of a data integration

II. DATA INTREGRATION TECHNIQUES

There are many intellectual ways the unitied view of data can be created today. No more ETL is the only way to reach the goal and that is a new level of complexity in the field of Data Integration.

There are several organizational levels on which the Data Integration can be performed and let’s discuss them briefly.

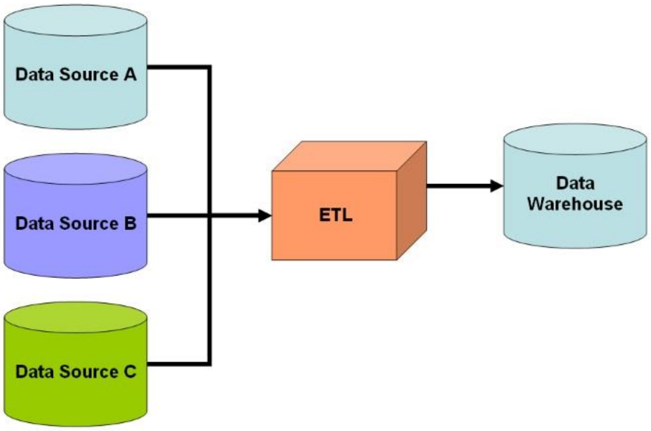

Figure2. Data Integration Technique

1. **Data Integration in Manually**

With regard to technique, this is really not a Data Integration. In this approach, a web based user interface or particular purpose is created for users of the system to show them all the relevant information by right to entering all the source systems directly. There is separation of data in reality.

2. **Data Integration in Middleware**

A data integration solution in middleware is essentially a layer between two different systems allowing them to communicate. Integration in Middleware can act like a glue that holds together multiple award applications, making coherent connectivity possible without requiring the two applications to communicate directly.

3. **Data Virtualization Integration Approach**

Data Virtualization allows us to leave data in the source systems while allowing to create a new set of unified views. This provides a way for users to access the unified view of disparate source system's data across whole enterprise.

Now a days many organizations today prefer this approach because of the benefits and technologies that exist today to support this approach. The main benefit of the virtual integration approach is near real time view of data from the source systems. It eliminates a need for separate data store for the compact unified data.

However, that doesn't mean it's the best way to do Data Integration although it certainly has a short term benefit. The drawbacks of this approach include limited possibility of data's history availability or data version management and unwanted load on the source systems involved which may have an adverse effect on the performance of the source systems.

4. **Data Integration with Data Warehouse Approach**

According to Ralf Kimball and/or Bill Inmon, this is the most commonly known approach to Data Integration.

This approach requires creation of a new Data Warehouse of Data Marts which stores a unified version of data extracted from all the source systems involved and manage as well handle it independent of the original source systems.

The benefits of this approach include ability to easily manage history of data (or data versioning), ability to combine data from very different sources (mainframes, databases, flat files, etc.) and to store them in a central repository of data. i.e. Big data mining.

III. CHALLENGES IN BIG DATA INTEGRATION

The handling of data in big data mining is very intrical. While handling some challenges faced during its integration include uncertainty of data Management, big data mining talent gap, getting data into a big data mining structure, syncing across data sources, getting useful information out of the big data mining, volume, skill availability, solution cost etc.

1. **The Uncertainty of Data Management:**

One unquit aspect facet of big data mining management is the use of a wide range of progressive data management tools and frameworks whose designs are devoted to supporting operational and analytical processing. The NoSQL (not only SQL) frameworks are used that differentiate it from conventional relational database management systems and are also largely designed to fulfil performance urgent request of big data mining applications such as managing a large amount of data and quick response times. There are a variety of NoSQL approaches such as hierarchical object representation (such as JSON, XML and BSON) and the concept of a key-value storage. The wide range of NoSQL tools, developers and the status of the market are creating uncertainty with the data management.

2. **Talent Gap in Big Data mining:**

It is difficult to win the respect from media and analysts in tech without being bombarded with content touting the value of the analysis of big data mining and corresponding reliance on a wide range of unquiet technologies. The new tools evolved in this sector can range from traditional relational database tools with some alternative data layouts designed to maximize access speed while reducing the storage footprints, NoSQL data management frameworks, in-memory analytics, and as well as the broad Hadoop ecosystem. The reality is that there is lack of skills available in the market for big data mining technologies. The typical expert has also gained experience through tool implementation and its use as a programming model, apart from the big data mining management aspects.

3. Getting Data into Big Data mining Structure:

It might be apperent that the intent of a big data mining management involves diagnosing and processing a large amount of data. There are many people who have raised expectations considering analyzing huge data sets for a big data minig platform. They also may not be aware of the complexity behind the transmission, access, and delivery of data and information from a wide range of resources and then loading these data in a big data mining platform. The intricate aspects of data transmission, access and loading are only part of the challenge. The requirement to navigate transformation and extraction is not limited to conventional relational data sets.

4. Synchronise Across Data Sources:

Another challenge is that , Once you bring data into big data mining platforms you may also realize that data duplication migrated from a wide range of sources on different rates and schedules can rapidly get out of the synchronization with the originating system. This implies that the data coming from one source is not out of date as compared to the data coming from another source. It also means the commonality of data definitions, concepts, metadata and the like. The conventional data management and data warehouses, the sequentially arrangement of data transformation, extraction and migrations all arise the situation in which there are some risks for data to become unsynchronized.

5. Extracting Information from the Data in Big Data Integration:

Practically ,the most use cases for big data mining involve the availability of data, augmenting existing storage of data as well as allowing access to end-user employing business intelligence tools for the purpose of the discovery of data. This business intelligence must be able to connect different big data mining platforms and also provide transparency of the data consumers to eliminate the requirement of custom coding. At the same time, if the number of data consumers grow, then one can provide a need to support an increasing collection of many simultaneous user accesses. This increment of demand may also spike at any time in reaction to different aspects of business process cycles. It also becomes a challenge in big data integration to ensure the right-time data availability to the data consumers.

6. Many Other Mixed Challenges:

Many mixed challenges may occur while integrating big data mining. Some of the challenges include integration of data, availability of skill, cost of solution, the volume of data, the transformation rate of data, veracity and validity of data. The ability to merge data that is not similar in source or structure and to do so at a reasonable cost and in time. It is also a challenge to process a large amount of data at a reasonable speed so that information is available for data consumers when they need it. The validation of data set is also fulfilled while transferring data from source to destination as well as consumers .

This is about the big data integration and some challenges arises during the implementation. These points must be considered and should be taken care of if you are going to manage any big data mining platform.

IV. SPARK 'S MAKING SENSE

Apache spark is a distributed computing engine for data processing. A lot of advantages are derived by its memory centric processing model. Initially, it was developed to make machine learning easier, but it was not designed to be a nice multi-tenant system. Owing to the needs of machine learning for normalized data sets, a lot of ancillary efforts and use cases have followed.

"Machine learning has a need for normalized data sets," said Matei Zaharia, the developer of Spark and CTO at Databricks, a vendor of a commercial version of Spark. This means that when developers build an application that takes data from different databases, it has to be transformed or normalized so that the algorithms can efficiently perform analytics. As it turns out, this is one of the core functions of ETL systems required for data warehousing.

Heudecker said the number one use case for Spark today is data integration and log processing, not machine learning. He cited one example of an enterprise that improved ETL processes where Spark reduced the time to 90 seconds from four hours. This kind of capability is significant because it means an enterprise can ask a question constantly throughout the day rather than once or twice.

Uber had an issue where the original system was built to ingest trip data, normalize it and put it into a data warehouse. This did not scale across multiple cities. Spark was one of the key elements that allowed them to bypass the legacy systems for putting data into the warehouse. This made it possible to format the data to drive new types of analytics and machine learning applications.

Consider Spark for different types of computations

One of the challenges enterprise architects face lies in optimizing data for different types of processing. As a result, many companies implement one architecture for transactions, another for operational analytics and a third for business intelligence. The wide interest in Spark has led to the development for these different types of computation on top of unified underlying data architecture.

The Spark platform addresses batch, interactive and real-time use cases. Before Spark, an enterprise would need to implement three separate platforms for these different use cases. In addition to that, machine learning options are included with Spark and are available in a wide variety of implementations on standalone clusters, in conjunction with Hadoop infrastructure, or as a cloud service.

The Spark stream processing infrastructure allows enterprises to perform analytics across streams of data in batch, which allows applications to apply batch-oriented analytics methods to data in motion. Spark SQL allows developers to do interactive analytics on SQL data sources. Heudecker said this enables more than what organizations are used to thinking about with SQL and siloed data access.

Spark also supports graph computation, which is good at identifying the links between entities described in large data sets. It's commonly used in social network analysis, recommendation engines, and fraud detection. Heudecker said he does not hear much about graph computation being adopted by enterprises today, but it is rising in importance.

Consider an in-memory integration tier

A number of tools are emerging to enable new architectures to leverage Spark integration. For example, Alluxio has developed an open source memory-centric cache that works in conjunction with Spark. This allows the enterprise to create an index of metadata describing data stored throughout the enterprise, which can be queried more quickly.

Traditionally, enterprises would put data into a big cluster to derive value, but would lose the historical context of the data because it had been moved. Spark uses metadata to tag these different sources of data to provide the concept of a just-in-time data warehouse. Heudecker said, "This is more than a data warehouse, this is a data warehouse with analytics." Many companies are built from acquisition and will not get rid of separate data warehouses. They can use Alluxio as a repeater station.

The Chinese search giant Baidu has leveraged this kind of approach to speed its query processing for data stored across servers throughout China. It would take four to eight hours to process specific complex queries. By implementing a virtual integration tier on top of Alluxio, Baidu reduced query time to 10 seconds.

V. CONCLUSIONS

We have studied big data mining from definitions to various cases and derived challenges as well as data integration for data mining with big data. In order to explore Big Data mining, we have analyzed several challenges at the data.

We regard Big Data mining as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate data to social media.There are possibilities for designing big data integration for identifying social media releted trend . We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived

REFERENCES

[1] . Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (references)

[2] Rajaraman and Ullman, 2011, A. Rajaraman and J. Ullman,

Mining of Massive Datasets, Cambridge University Press, 2011.

[3] https://iaonline.theiia.org/data-mining-101-tools-and-techniques

[4] http://sixminutes.dlugan.com/six-simple-techniques-for presenting-data-hans-rosling-ted-2006/

[5] A., Statistical Pattern Recognition, Wiley, 2002.

[6] Data Mining Techniques Paperback – 2010

[7] By Arun K. Pujari (Author)

[8] Data Mining: Concepts and Techniques Paperback –2007 by Han (Author)

[9] Data Mining: Practical Machine Learning Tools and Techniques Paperback – 2010 by Ian H. Witten (Author), Eibe Frank (Author), Mark A. Hall (Author)

[10] http://www.slideshare.net/marin_dimitrov/semantictechnologies for-big-data

[11] [Online] Srinath Srinivasa, Big Data and the Semantic Web: Challenges and Opportunities

[12] http://www.slideshare.net/srinaths/big-data-and-thesemantic-web - challenges-and-opportunities

[13] [Online] http://www.tutorialspoint.com/neo4j/

[14] [Online]http://franz.com/agraph/allegrograph/

[15] [Online]Marin Dimitrov, Semantic Technologies for Big data

[16] [online]http://searchcloudapplications.techtarget.com

[17] http://www.ibm.com

[18] http:// www.wikipedia.com

[19] Big data minig and Semantic Technology :Challenges and Opportunity 2015 by Ms.Yesha Mehta ,Dr. Sanjay Buch.