Data Integration

Chapter · January 2017

DOI: 10.1007/978-3-319-32001-4_54-1

2 authors:

Anirudh Kadadi

4 PUBLICATIONS **175** CITATIONS

SEE PROFILE

Rajeev Agrawal

Engineer Research and Development Center - U.S. Army

95 PUBLICATIONS **640** CITATIONS

SEE PROFILE

D

Data Integration

Anirudh Kadadi[1] and Rajeev Agrawal[2]
[1]Department of Computer Systems Technology, North Carolina A&T State University, Greensboro, NC, USA
[2]Information Technology Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, USA

Synonyms

Big data; Big data integration tools; Semi-structured data; Structured data; Unstructured data

Introduction

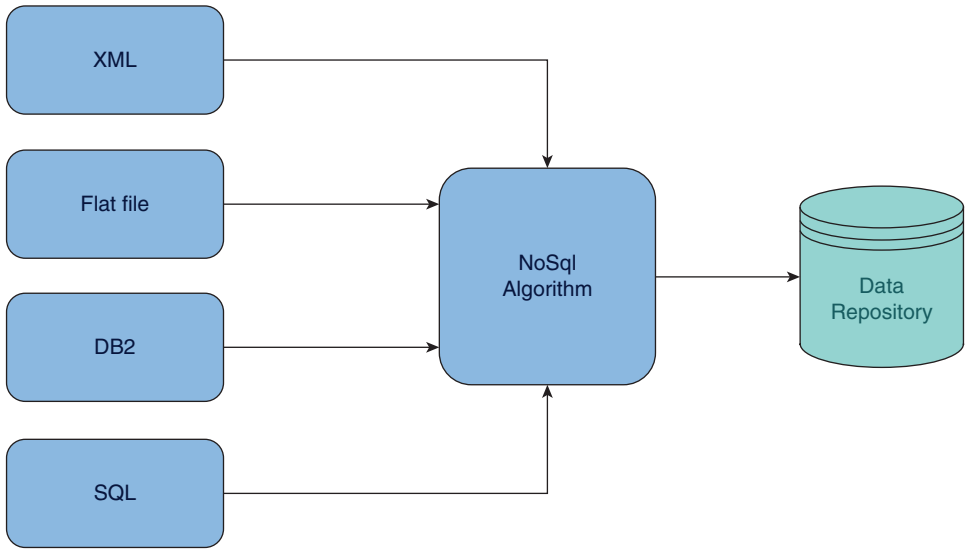
Big data integration can be classified as a crucial part of integrating enormous datasets in multiple values. The big data integration is a combination of data management and business intelligence operations which covers multiple sources of data within the business and other sources. This data can be integrated into a single subsystem and utilized by organizations for business growth. Big data integration also involves the development and governance of data from different sources which could impact organization's abilities to handle this data in real time.

The data integration in big data projects can be critical as it involves:

1. Discovering the sources of data, analyzing the sources to gain bigger insights of data, and profiling the data.
2. Understanding the value of data and analyzing the organizational gains through this data. This can be achieved by improving the quality of data.
3. Finally transforming the data as per the big data environment (Fig. 1).

The five Vs of big data can influence the data integration in many ways. The five Vs can be classified as volume, velocity, variety, veracity, and value: The enormous volume of data is generated every second in huge organizations like Facebook and Google. In the earlier times, the same amount of data was generated every minute. This variation in data-generating capacity of the organizations has been increasing rapidly, and this could motivate the organizations to find alternatives for integrating the data generated in larger volumes for every second. The speed at which the data is transmitted from the source to destination can be termed as velocity. Data generated by different jobs at each time is transmitted at timely basis and stored for further processing. In this case, the data integration can be performed only after a successful data transmission to the database. The data comes from numerous sources which categorizes them into structured and unstructured. The data from social media can be the best example for unstructured data which

L.A. Schintler, C.L. McNeely (eds.), *Encyclopedia of Big Data*,
DOI 10.1007/978-3-319-32001-4_54-1



Data Integration, Fig. 1 Data integration – different sources

includes logs, texts, html tags, videos, photographs, etc. The data integration in this scenario can be performed only on the relational data which is already structured and the unstructured data has to be optimized to structured data before the data integration is performed (Vassiliadis et al. 2002).

The trustworthiness and accuracy of the data from the sources can be termed as the veracity. The data from different sources comes in the form of tags and codes where organizations were lagging the technologies to understand and interpret this data. But technology today provides us the flexibility to work with these forms of data and use it for business decisions. The data integration jobs can be created on this data depending on the flexibility and trust of this data and its source. The value can be termed as the business advantage and profits the data can bring to the organization. The value depends solely on the data and its source. Organizations target their profits using this data, and this data remains at a higher stake for different business decisions across the organization. Data integration jobs can be easily implemented on this data, but most of the organizations tend to keep this data as a backup for their future business decisions. Overall, the five V's of big data play a major role in determining the efficiency of organizations to perform the data integration jobs at each level (Lenzerini 2002).

Traditional ETL Methods with Hadoop as a Solution

Organizations tend to implement the big data methodologies into their work system creating information management barriers which include access, transform, extract, and load the information using traditional methodologies for big data. Big data creates potential opportunities for organizations. To gain the advantage over the opportunities, organizations tend to develop an effective way of processing and transforming the information which involves data integration at each level of data management. Traditionally, data integration involves integration of flat files, in-memory computing, relational databases, and moving data from relational to non-relational environments.

Hadoop is the new big data framework which enables the processing of huge datasets from different sources. Some of the market leaders are working on integrating Hadoop with the legacy systems to process their data for business use in current market trend. One of the oldest contributors to the IT industry "the mainframe" has been

into existence since a long time, and currently IBM is working on development of new techniques to integrate the large datasets through Hadoop and mainframe.

The Challenges of Data Integration

In a big data environment, the data integration can lead to many challenges in real-time implementation which has the direct impact on projects. Organizations tend to implement new ways to integrate this data to derive meaningful insights at a bigger picture. Some of the challenges posed in data integration are discussed as:

(i) *Accommodate scope of data:*

Accommodating the sheer scope of data and creating newer domains in the organization are a challenge, and this can be addressed by implementing a high-performance computing environment and advanced data storage devices like hybrid storage device which features hard disk drives (HDD) and solid-state drives (SSD); possesses better performance levels with reduced latency, high reliability, and quick access to the data; and therefore helps accumulate large datasets from all the sources. Another way of addressing this challenge can be through discovery of common operational methodologies between the domains for integrating the query operations which stands as a better environment to address the challenges for large data entities.

(ii) *Data inconsistency:*

Data inconsistency refers to the imbalances in data types, structures, and levels. Although the structured data provides the scope for query operations through relational approach so that the data can be analyzed and used by the organization, unstructured data takes a lead always in larger data entities, and this comes as a challenge for organizations. Addressing the data inconsistency can be achieved using the tag and sort methods which allow searching the data using keywords. The new big data tool Hadoop provides the solution for modulating and converting the data through MapReduce and Yarn. Although Hive in Hadoop doesn't support the online transactions, they can be implemented for file conversions and batch processing.

(iii) *Query optimization:*

In real-time data integration, the large data entities require the query optimization at microlevels which could involve mapping components to the existing or a new schema which impacts the existing structures. To address this challenge, the number of queries can be reduced by implementing the joins, strings, and grouping functions. Also the query operations are performed on individual data threads which can reduce the latency and responsiveness. Using the distributed joins like merge, hash, and sort can be an alternative in this scenario but requires more resources. Implementing the grouping, aggregation, and joins can be the best approach to address this challenge.

(iv) *Inadequate resources and implementing support system:*

Lack of resources haunts every organization at certain point, and this has the direct impact on the project. Limited or inadequate resources for creating data integration jobs, lack of skilled labor that don't specialize in data integration, and costs incurred during the implementation of data integration tools can be some of the challenges faced by organizations in real time. This challenge can be addressed by constant resource monitoring within the organization, and limiting the standards to an extent can save the organizations from bankruptcy. Human resources play a major role in every organization, and this could pick the right professionals for the right task in a timely manner for the projects and tasks at hand.

There is a need to establish a support system for updating requirements and error handling, and reporting is required when organizations perform various data integration jobs within the domains and externally. This can be an additional cost for the

organizations as setting up a training module to train the professionals and direct them toward understanding the business expectations and deploy them in a fully equipped environment. This can be termed as a good investment as every organization would implement advancements in a timely manner to stick with the growing market trends. Support system for handling errors could fetch them the reviews to analyze the negative feedback and modify the architecture as per the reviews and update the newer versions with better functionalities.

(v) *Scalability:*

Organizations could face big time challenge in maintaining the data accumulated from number of years of their service. This data is stored and maintained using the traditional file systems or other methodologies as per their environment. In this scenario, often the scalability issues arise when the new data from multiple resources is integrated with data from legacy systems. Changes made by the data scientists and architects could impact the functioning of legacy systems as it has to go through many updates to match the standards and requirements of new technologies to perform a successful data integration. In recent times, mainframe stands as one of the best example for legacy system. For a better data operation environment and rapid access to the data, Hadoop has been implemented by organizations to handle the batch processing unit. This follows a typical ETL (Extract, Transform, and Load) approach to extract the data from number of resources and load them into Hadoop environment for the batch processing.

Some of the common data integration tools which have been in use are Talend, CloverETL, and KARMA. Both the data integration tools have their own significance individually for providing the best data integration solutions for the business.

Real-Time Scenarios for Data Integration

In the recent times, Talend was used as the main base for data integration by Groupon, one of the leading deal-of-the day website which offers discounted gift certificates, to be used at local shopping stores. For integrating the data from sources, Groupon relied on "Talend." Talend is an open-source data integration tool which is used to integrate data from numerous resources. When Groupon was a startup, they relied on an open source for more gains rather than using a licensed tool which involves more cost for licensing. Since Groupon is a public traded company now, they would have to process 1 TB of data per day, which come from various sources.

There is another case study where a telephone company was facing issues with phone invoices in different formats which were not suitable for electronic processing and therefore involved the manual evaluation of phone bills. This consumed a lot of time and resources for the company. The CloverETL data integration tool was the solution for the issue, and the inputs given were itemized phone bills, company's employee database, and customer contact database. The data integration process involved consolidated call data records, report phone expenses in hierarchy, and analysis of phone calls and its patterns. This helped organization cut down the costs incurred by 37% yearly.

Conclusion

On a whole, the data integration in current IT world is on demand with the increasing number of data and covers complete aspects of data solutions with the usage of data integration tools. Data scientists are still finding solutions for a simplified data integration with an efficient automated storage systems and visualization methods which could turn out complex in terms of big data. Development of newer data integration solutions in the near future could help address the big data integration challenges. An efficient data integration tool is yet to conquer the market, and evolution of these tools can help organizations handle

the data integration in a much more simplified way.

Further Readings

Big Data Integration. http://www.ibmbigdatahub.com/video/big-data-integration-what-it-and-why-you-need-it

Clover ETL. http://www.cloveretl.com/resources/case-studies/data-integration

Data Integration tool. http://blog.pentaho.com/2011/07/15/facebook-and-pentaho-data-integration/

IBM. How does data integration help your organization? http://www-01.ibm.com/software/data/integration/

Lenzerini, M. (2002). Data integration: A theoretical perspective, In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 233–246), New York, NY, USA.

Talend. https://www.talend.com/customers/customer-reference/groupon-builds-on-talend-enterprise-data-integration

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP(DOLAP '02)* (pp. 14–21). New York: ACM.