International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | Jun 2022 www.irjet.net p-ISSN: 2395-0072

# Comparing and analyzing various method of data integration in big data

**Rahul thakur[1], kewal krishan[2]**

*[1]Rahul thakur rahulthakurhm@gmail.com*
*[2]kewal krishan kewal.krishan@lpu.co.in*

---

**Abstract -** *Data integration is a challenge that involves combining different data from multiple sources and providing a viewer with a uniform or consistent picture of the data. In the modern world, combining multiple heterogeneous sources into a single truth is one of the biggest challenges. Everyone wants to look at and feel the data in one way. Different methods are there, but there are still some challenges. We have compared and analyzed so that the appropriate methods can be used for the system. This document presents an overview of different data integration techniques appropriate for the system and their challenges. This article pays special attention to comparative analysis of different techniques of data integration. A special spotlight on the following aspects: data integration techniques to deal with unreliable data.*

***Key Words***: **Big Data, Big Data Integration, Extract Transform Load.**

## 1. INTRODUCTION

Integration of data is a collection of techniques for retrieving and integrating data from multiple data sources to produce meaningful information. Nowadays, a large amount of data is gathered from a range of heterogeneous data sources in real time, resulting in data of varying quality. This is referred to as "Big Data." Big data integration is highly challenging, particularly when conventional data integration solutions have failed. Big data integration varies from traditional data integration in so many ways, including volume, velocity, variety, and veracity, which are the primary features of big data.

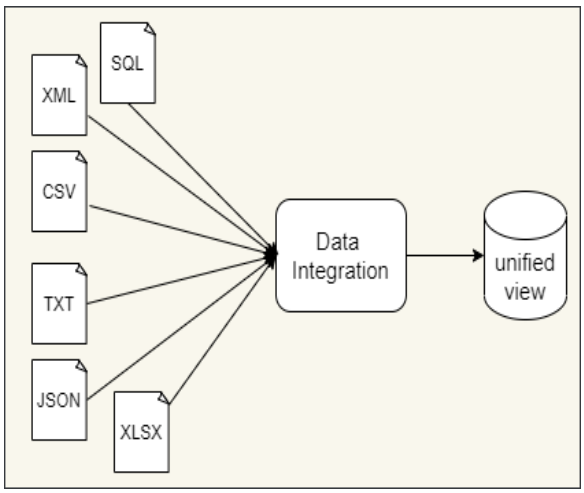


**Fig 1:Data Integration**

### 1.1 Different types of big data integration processes.

Big data integration involves combining data from various sources and displaying it in a single interface (BDI). Building an enterprise data warehouse, transferring data between databases, and keeping data synchronised across platforms all require BDI. To present an overall view, the BDI integrates data from both internal and external sources.

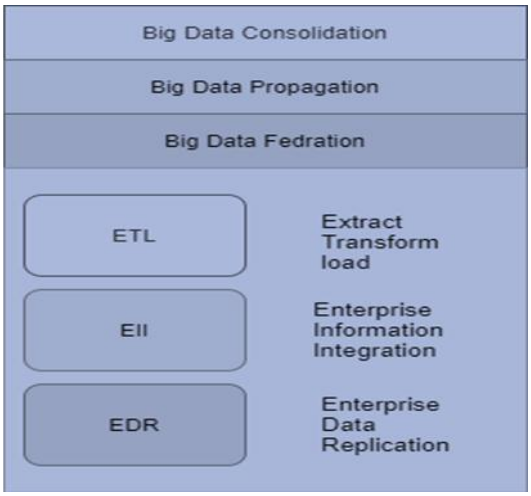


**Fig 2:BDI Technologies**

### 1.2 BDI technologies

**Extract, transform, and load (ETL)**

ETL is a method of integrating data. Extracting information from one system and loading it into another fter it has been transformed is known as ETL.

**EII**

This method of data integration is used to convey compiled data sets on demand.EII allows programmers and enterprise clients to combine data from multiple sources into one database.

**EDR stands for enterprise data replication (EDR).**

Data migration from one storage platform to another is a part of the EDR process. In its most basic form, while maintaining the data's structure, EDR moves it from one repository to another.

## 2. Types of Data Integration

**Data Consolidation**

Data consolidation is the procedure of combining data from several sources into a single storage area. Data consolidation uses networking servers as data sources to reduce the figure of locations for data storage required by an organisation. Data consolidation uses ETL (Extract, Transform, Load) to gather data from main databases. Data is collected from various sources, cleaned, normalised, and stored for processing.In most cases, ETL consolidations are processed in batches of 24 hours or less. A "holding area" is used for batch consolidation prior to delivery to an integrated data storage facility. Large datasets should avoid this processing due to its high latency.Data consolidation, unlike other data integration solutions, relies on delay. The amount of time it requires for data to move from one location to another is referred to as data latency.

**Data Propagation**

An integration technique known as "data propagation" involves copying data from one source to another. Local access databases are provided data from source data warehouses through propagation rules.

Data propagation rules three categories:

**Bulk extraction**

FTP is used to retrieve and transmit data from a source (file transfer protocol). The extracted data may need to be re-formatted to fit into the destination data storage. Bulk extraction is perfect for small source files or large changes. This method does not distinguish between modified and unaffected entries.

**Comparing files**

Unlike bulk extract, file comparison produces an incremental change record. Small files with few changes may benefit from this strategy to track changes over time.

**Change Data Capture (CDC)**

Change data and capture (CDC) is a real-time approach for identifying and replicating changes in the source store. CDC propagation keeps store databases updated quickly (seconds or minutes). Minor changes or new data can be discovered quickly without having to update the entire data warehouse. Trigger-based and log-based CDC are examples.

**Data federation**

Consider data federation as "linking up" or "becoming one unit." It's a term for "middleware" technology that connects data from various sources and formats into a single picture. With a relational database management system (RDBMS), analysts can create tables with rows and columns of data. At the endpoint, the Federation uses a data model to generate a single-view visualisation. Using the RDBMS's SQL (Structured Query Language) interface,

## 3. CONCLUSIONS

Integration is now the world's largest It challenge with 1 of 6 It dollars globally spent on integration.83% of executives have admitted to data silos being present in their organizations, and 97% say it harms overall decision making and 57% of marketers recognize integrating disparate technologies as the most significant barrier to success. Data integration provides the potential to produce more timely, more disaggregated statistics at higher frequencies than traditional approaches alone. Data integration activities will therefore only increase. With ever more data sources becoming available and increased capacities of IT and data infrastructure, the need for integrating different sources will grow.

Data integration is the process of combining data from many sources. Data integration must contend with issues such as duplicated data, inconsistent data, duplicate data, old systems, etc. Manual data integration can be accomplished through the use of middleware and applications. You can even use uniform access or data warehousing. There are several tools available on the market that may be used to do data integration.

In this paper, we provided a high-level summary of the constraints and problems that data integration must overcome with comparison . There is no one answer to any

of these issues. They're all connected in some way or another. Each data integration difficulty demands a distinctly different solution, which must be recognized in order to be successful in the long run. Attempts have been made to collect as many obstacles and concerns as feasible in this document so that additional work may be done in the future to solve these issues.

## REFERENCES

[1] Hasliza, N., Hassana, M., Ahmada, K. &Salehuddina, H. (2020). Diagnosing the Issues and Challenges in Data Integration Implementation in Public Sector, International Journal Advanced Science Engineering Information Technology, 10(2).

[2] Zhang, Y. (2020). The Integration of Professional Ethics of Modern Etiquette Students under the Background of Big Data, Journal of Physics: Conference Series 1574.

[3] Bansal, S. K. (2014). Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data Integration, IEEE International Congress on Big Data, 978-1-4799- 5057-7/14 © 2014 IEEE, DOI 10.1109/ BigData.Congress.2014.82

[4] Zheng, Y. (2015). Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions On Big Data, 1(1).

[5] Munné R. (2016). Big Data in the Public Sector. In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. https://doi. org/10.1007/978-3-319-21569-3_11.

[6] Camargo-Perez, J. A., PuentesVelasquez, A. M., & Sanchez-Perilla, A. L. (2019). Integration of big data in small and medium organizations: Business intelligence and cloud computing, J. Phys.: Conf. Ser. 1388 012029.

[7] Stonebraker, M., & Ilyas, I. F. (2018). Data Integration: The Current Status and the Way Forward, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

[8] Sazontev, V., &Stupnikov, S. (2019). An Extensible Approach for Materialized Big Data Integration in Distributed Computation Environments, Ivannikov Memorial Workshop (IVMEM), 978-1-7281- 4623-2/19/ ©2019 IEEE DOI 10.1109/ IVMEM.2019.00011

[9] Alsghaier, H., Akour, M., Shehabat, I., &Aldiabat, S. (2017). The Importance of Big Data Analytics in Business: A Case Study. American Journal of Software Engineering and Applications, 6(4), 111-115.

[10] Alam, J. R., Sajid, A., Talib, R., & Niaz, M. (2014). A Review on the Role of Big Data in Business. International Journal of Computer Science and Mobile Computing, 3(4), 446-453.

[11] Fikri, N., Rida, M., Abghour, N., Moussaid, K., &Omri, A. I. (2019). An adaptive and real-time based architecture for financial data integration. Journal of Big Data, 6(97).

[12] Bucea-Manea-Tonis, R. (2018). Deductive systems for Big data integration, Journal of Economic Development, Environment and People, 7(1).

[13] Chen, W., Wang, R., Wu, R., Tang, L., & Fan, J. (2016). Multi-source and Heterogeneous Data Integration Model for Big Data Analytics in Power DCS [Paper Presentation]. International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.

[14] Hussain K., Prieto E. (2016). Big Data in the Finance and Insurance Sectors. In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. https://doi. org/10.1007/978-3-319-21569-3_12

[15] Avi V., Kamaruddin S. (2017). Big Data Analytics Enabled Smart Financial Services: Opportunities and Challenges. In: Reddy P., Sureka A., Chakravarthy S., Bhalla S. (eds) Big Data Analytics. BDA 2017. Lecture Notes in Computer Science, vol 10721. Springer, Cham. https://doi. org/10.1007/978-3-319-72413-3_2

[16] Nabrzyski, J., Liu, C., Vardaman, C., Gesing, S., &Budhatoki, M. (2014). Agriculture Data for All - Integrated Tools for Agriculture Data Integration, Analytics and Sharing. IEEE International Congress on Big Data. 978-1-4799-5057-7/14 © 2014 IEEE DOI 10.1109/BigData.Congress.2014.117

[17] Kim, J. K., & Tam, S. (2020). Data integration by combining big data and survey sample data for finite population inference. arXiv:2003.12156v3.

[18] Saggi, M. K., & Jain, S. (2018). A survey towards the integration of big data analytics to big insights for valuecreation. Information Processing & Management, 54.

[19] Ribarics, P. (2016). Big Data and its impact on agriculture. Eco cycles, 2(1), 33-34.

[20] Sarker, M. N., Islam, M. S., Murmu, H., &Rozario, E. (2020). Role of Big Data on Digital Farming. International Journal of Scientific & Technology Research, 9(04).

[21] Kaur, H., & Kushwaha, A. S. (2018). A Review on Integration of Big Data and IoT. 4th International Conference on Computing Sciences. 978-1-5386-8025- 4/18/$31.00 ©2018 IEEE DOI 10.1109/ ICCS.2018.00040.

[22] Huang, E., Quiroz, A., &Ceriani, L. (2014). Automating Data Integration with HiperFuse [Paper Presentation] 2014 IEEE International Conference on Big Data

[23] Nuaimi, E. A., Neyadi, H. A., Mohamed, N., &Jaroodi, J. (2015). Applications of big data to smart cities. Journal of Internet Services and Applications, 6(25).

[24] Gomes, E., Dantas, M. A., Macedo, D. D., Rolt, C. D., Brocardo, M. L., & Foschini, L. (2016). Towards an Infrastructure to Support Big Data for a Smart City Project [Paper Presentation]. 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, 2016, pp. 107-112, DOI: 10.1109/ WETICE.2016.31

[25] Alshawish, r. A., Alfagih, S. M., &Musbah, M. S. (2016). Big data applications in smart cities. 2016 International Conference on Engineering & MIS (ICE), Agadir, 2016, pp. 1-7, DOI: 10.1109/ ICEMIS.2016.7745338

[26] Ahmed, F., Samorani, M., Bellinger, C., &Zaiane, O. R. (2016). Advantage of Integration in BigData: Feature Generation in Multi-Relational Databases for Imbalanced Learning, 2016 IEEE International Conference on Big Data (Big Data), 978-1-4673-9005- 7/16/$31.00 ©2016 IEEE

[27] Bennani, N., Ghedira-Guegan, C., Musicante, M. A., & Vargas-Solar, G. (2014). SLA-Guided Data Integration on Cloud Environments [Paper Presentation]. 2014 IEEE International Conference on Cloud Computing, Alaska, United States. 934-935.

[28] Qi, Q ., & Tao, F. (2018). Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison. IEEE Access, 6, 3585-3593.

[29] Hufnagel, J., & Vogel-Heuser, B. (2015). Data integration in manufacturing industry: Model-based integration of data distributed from ERP to PLC [Paper Presentation]. 2015 IEEE 13th International Conference on Industrial Informatics (INDIN), Cambridge, 2015, pp. 275-281, DOI: 10.1109/INDIN.2015.7281747.

[30] O'Donovan, P., Leahy, K., Bruton, K., & T. J. O'Sullivan. (2015). Journal of Big Data, 2(20). DOI 10.1186/ s40537-015-0028-x

[31] Hardiman, G. (2020). An Introduction to Systems Analytics and Integration of Big Omics Data, Genes, 11(245).

[32] Bhandari, S., Lewis, P., Craft, E., Marvel, s. W., Reif, D. M., & Chiu, W. A. (2020). HGBEnviroScreen: Enabling Community Action through Data Integration in the Houston–Galveston– Brazoria Region, Int J Environ Res Public Health, 17(4): 1130.

[33] Dhayne, H., Haque, R., Kilany, R., & Taher, Y. (2019). In Search of Big Medical Data Integration Solutions - A Comprehensive Survey. IEEE Access, 7.

[34] Eftekhari, A., Zulkernine, F., & Martin, P. (2016). BINARY: A Framework for Big Data Integration for Ad-hoc Querying, 2016 IEEE International Conference on Big Data (Big Data), 978-1-4673-9005- 7/16/©2016 IEEE

[35] Vidal, M., &Sakor, A. (2019). Semantic Data Integration Techniques for Transforming Big Biomedical Data into Actionable Knowledge, 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS).

[36] Husain, S., Kalinin, A., Truong, A., &Dinov, D. (2015). SOCR Data Dashboard: An integrated Big Data archive mashing Medicare, labour, census and econometric information. Journal of Big Data, 2(13).

[37] Cheng, Y., Zhou, K., Wang, J., & Yan, J. (2020). Big Earth Observation Data Integration in Remote Sensing Based on a Distributed Spatial Framework. Remote Sens. 12, 972.

[38] Wang, Z., Wei, G., Zhan, Y., & Sun, Y. (2017). Big data in telecommunication operators: data, platform and practices. Journal of Communications and Information Networks, 2(3). DOI: 10.1007/ s41650-017-0010-1

[39] Yayah, F. C., Ghauth, K. I., & Ting, C. (2017). Adopting Big Data Analytics Strategy in the Telecommunication Industry. Journal of Computer Science & Computational Mathematics. 7(3). DOI: 10.20967/jcscm.2017.03.002

[40] Nwanga, M. E., Onwuka, E. N., Aibinu, A. M., &Ubadike, O. C. (2015). Impact of Big Data Analytics to the Nigerian Mobile Phone Industry. Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management Dubai, United Arab Emirates (UAE), March 3-5, 2015.

[41] Antonio, A. C., Luis, M. S., Santos, M. Y., Guilherme, A. B., & Jose, A. O. (2020). Supply chain data integration: A literature review. Journal of Industrial Information Integration 19 100161.

[42] Ostrowski, D., & Kim, M. (2017). Semantic-Based Framework for Big Data Integration [Paper Presentation]. 2017 IEEE 11th International Conference on Semantic Computing

[43] Awwad, M., Kulkarni, P., Bapna, R., & Marathe, A. (2018). Big Data Analytics in Supply Chain: A Literature Review. Proceedings of the International Conference on Industrial Engineering and Operations Management, Washington DC, USA, September 27-29.

[44] Lia, Q ., Liu, A. (2019). Big DataDriven Supply Chain Management, Procedia CIRP 81 ScienceDirect 52nd CIRP Conference on Manufacturing Systems, 1089-1094.

[45] Benabdellah, A. C., Benghabrit, A., Bouhaddou, I., &Zemmouri, E. M. (2016). Big Data for Supply Chain Management: Opportunities and Challenges. International Journal of Scientific & Engineering Research, 7(11).

[46] Li, J. (2020). Research on the Integration of Chinese and Russian Original Ecological Dance Elements and Modern Elements Based on Computer Big Data Analysis. Journal of Physics: Conference Series 1578.

[47] Arputhamary, B. &Arockiam, L. (2015). Data Integration in Big Data Environment. Bonfring International Journal of Data Mining, 5(1), 1-5.

[48] Kadadi, A., Agrawal, R., Nyamful, C., &Atiq, R. (2014). Challenges of Data Integration and Interoperability in Big Data. 2014 IEEE International Conference on Big Data, 978-1-4799- 5666-1/14/$31.00 ©2014 IEEE

[49] Ostrowski, D., Rychtyckyj, N., MacNeille, P., Kim, M. (2016). Integration of Big Data Using Semantic Web Technologies. 2016 IEEE Tenth International Conference on Semantic Computing, 978-1-5090-0662-5/16 © 2016 IEEE DOI 10.1109/ICSC.2016.101

[50] Sottovia, P., Paganelli, M., Guerra, F., &Vincini, M. (2019). Big Data Integration of Heterogeneous Data Sources: the Research Alps CaseStudy. 2019 IEEE International Congress on Big Data (BigData Congress), 978-1-7281- 2772-9/19 ©2019 IEEE DOI 10.1109/ BigDataCongress.2019.00027

[51] Portugal, I., David, P. A., & Cowan, D. (2016). Towards a ProvenanceAware Spatial-Temporal Architectural Framework for Massive Data Integration and Analysis, 2016 IEEE International Conference on Big Data (Big Data).