

Harmonizing Industry: Techniques, Challenges, and Case Studies in Industrial Data Integration

Lydia Lakhdari^a, Leila Zemmouchi-Ghomari^{a*}

^aIndustrial Engineering and Maintenance Department, Ecole Nationale Supérieure des Technologies Avancées, ENSTA, Algiers, Algeria

Abstract

Industrial data integration is the process of gathering, combining, and analyzing data from various sources to create a single, cohesive picture of the data for analysis and decision-making. Many industries, such as manufacturing, energy, transportation, and healthcare, depend on this integration because it helps businesses maximize productivity, cut expenses, and raise quality.

Industrial data can come from various sources, including sensors, machinery, devices, and databases. It can also be of different types, including unstructured, semi-structured, and structured data. Creating a coherent and significant whole out of all these disparate data sources is the challenge of industrial data integration.

Industrial data integration requires the application of expert methodologies and methods. The aim is to build a data infrastructure that can manage massive amounts of data, process it fast and precisely, and deliver real-time insights into industrial process performance.

This paper explains industrial data integration techniques and their challenges. Some case studies reported in the literature draw attention to the insights gained from successful data integration initiatives and use their benefits and best practices.

Keywords: Industrial data; data integration; data interoperability.

1. Introduction

Data-driven strategies are now a competitive imperative to succeed in every business and industry in the information economy, where data has become the most critical asset of every organization. Organizations increasingly rely on data for decision-making to support business goals like revenue growth, profitability, and customer satisfaction (Sbiti et al., 2021; Wang et al., 2019).

* Corresponding author.

E-mail address: leila.ghomari@ensta.edu.dz

Data integration creates a unified view of data from multiple sources, increasing its value and usability for stakeholders who access it (Reeve, 2013; Doan et al., 2012; Ziegler & Dittrich, 2007; Slater et al., 2008; Prasser et al., 2018). Organizations need successful data integration that combines speed and integrity at scale to meet the demands of their businesses today with dependable data and keep up with innovations and exponential data growth.

Integrating industrial data has become essential for businesses to streamline processes, cut expenses, and raise standards. Industrial data integration is the process of gathering, analyzing, and combining data from different sources in an industrial environment to produce a cohesive picture of the data for analysis and judgment. Thanks to this unified view of data, a more comprehensive understanding of their operations and the ability to make well-informed decisions that improve productivity and profitability are possible for manufacturing, energy, transportation, and healthcare sectors.

Data comes from various industrial sources, including sensors, equipment, databases, and devices. It can be organized into different categories, including unstructured, semi-structured, and structured data. Bringing all these disparate data sources together into a coherent and meaningful whole is the challenge of industrial data integration. This integration calls for specific tools and methodologies to build a data infrastructure that can manage massive volumes of data, process it fast and precisely, and offer real-time insights into the operation of industrial processes.

This paper will discuss the problems and difficulties associated with industrial data integration while outlining the methods utilized. Additionally, we will examine some prominent case studies that showcase the lessons learned from successful data integration projects and how companies can take advantage of this practice's benefits and best practices. This paper aims to provide a comprehensive overview of industrial data integration and equip readers with the knowledge to explore this critical field.

2. Benefits of data integration in the industrial domain

Through data integration, organizations can combine information from several sources to produce a single source of truth. This unified view offers a trustworthy and consistent picture of the business's operations, improving data-driven decision-making. Data integration enables organizations to utilize their data to its full potential. Enhanced data quality, enhanced decision-making, heightened efficiency, and innovation are just a few advantages it provides. Businesses integrating their data will be more competitive in the industrial landscape as data volumes increase. Companies can gain a competitive edge by discovering new avenues for innovation and product/service development by integrating data from various domains. Additionally, data integration promotes interdisciplinary cooperation, which makes it easier to solve complicated problems and make well-informed decisions.

By identifying and correcting errors and inconsistencies, data integration improves the security and quality of project data, producing more dependable data. Furthermore, the integration process makes data cleansing—a crucial component of data quality assurance—more accessible. Data cleansing comprises identifying and resolving missing, incorrect, or irrelevant information from a dataset.

A solid basis for well-informed decision-making is provided by data cleansing, which guarantees that the integrated data is accurate, complete, and consistent. Additionally, data integration ensures that the data is safe, private, and compliant with regulations by enhancing data quality and preventing errors that may result in project failures or security breaches.

Businesses can also use machine learning and advanced analytics with data integration. These technologies need a sizable and varied dataset to train models and produce precise predictions, which can be obtained through data integration.

Incorporating data from sensors and IoT devices can also offer real-time operational insights, facilitating proactive maintenance and cutting downtime.

Developing an organization-wide data-driven culture is another advantage of integrating data. Employees can better identify opportunities, solve problems more quickly, and make educated decisions with access to accurate and pertinent data. As a result, the company becomes more responsive and agile, able to change with the times and capitalize on new opportunities.

Finally, data integration can assist in pinpointing areas where cost savings can be achieved by eliminating inefficiencies and simplifying procedures. Identifying operational bottlenecks and inefficiencies facilitates process optimization and cost reduction for companies, enhancing manufacturing processes and operational efficiency. Additionally, businesses can detect emerging issues sooner by integrating data from various sources, facilitating early problem detection and resolution.

3. Data Integration Approaches

An approach to data integration is neither singular nor standard. Nonetheless, data integration solutions typically share a few components, such as a controller server, a network of data sources, and clients that access data through it (Sbiti et al., 2021). There are several dimensions through which we can classify data integration approaches (figure 1), including:

3.1. Layers

Given the business size, the requirements to be met, and the resources at hand, the integration issue can generally be handled on each system level mentioned above. The six data integration approaches listed below are the foundation upon which actual integration solutions are built in practice (Ziegler & Dittrich, 2007):

Manual Integration: Users manually integrate the chosen data while interacting directly with all pertinent information systems. That is, users have to deal with disparate query languages and user interfaces. Users must also understand the data's location, logical representation, and semantics in great detail.

Common User Interface: In this scenario, users can access a uniformly designed standard user interface, such as a web browser. Users (e.g., search engines) are still required to homogenize and integrate the data because the data from the relevant information systems is still presented separately.

Integration through Applications: This strategy uses integration apps connecting multiple data sources and providing users with integrated outcomes. Only a few component systems can effectively use this solution. However, as the number of system interfaces and data formats that need to be standardized and integrated rises, applications become more complicated.

Integration via Middleware: Middleware, such as SQL-middleware, offers reusable functionality generally utilized to address specific components of integration problems. While applications are spared from implementing standard integration functionality, they still need to make integration efforts. Furthermore, developing integrated systems typically requires combining various middleware tools.

Uniform data access: This is used to carry out logical data integration at the data access level. At this level, only virtual data is available, but global applications are nonetheless given a unified global view of physically distributed data. In addition to maintaining their independence, local information systems can facilitate additional data access layers for external applications. However, because data access, homogenization, and integration must be completed at runtime, the overall provision of physically integrated data may take longer than expected.

Common Data Storage: In this case, data integration is accomplished physically by moving the data to a new storage; local sources can be eliminated or re-operated. Physical data integration typically provides fast

access to data. If the programs that utilize the local data sources are removed, they must be migrated to the new data store. When the local data sources remain operational, periodic refreshment of the shared data storage should be considered (Wang et al., 2019).

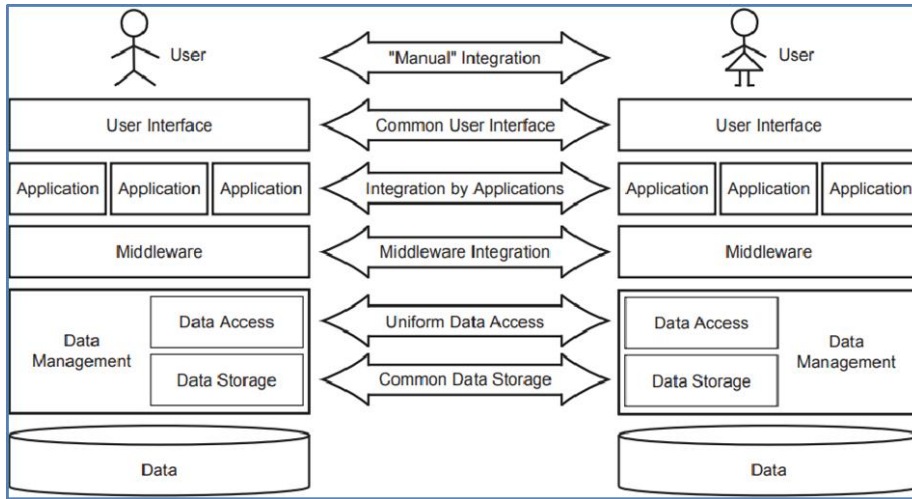


Fig. 1. General Integration Approaches on Different Architectural Levels (Wang et al., 2019)

3.2. Solutions

Data generation is happening at an exponential rate. Businesses need a practical data analysis and storage solution to handle this data effectively. The various forms of data storage solutions, such as data lakes, hubs, and warehouses, will be examined in this section.

3.2.1. Data Warehouse

As seen in Figure 2, the data warehouse is a conventional method of data storage that keeps organized data from multiple sources. Providing quick access to data for analytical purposes is a data warehouse's main goal (Nambia & Mundra, 2022).

Reports and trend analysis are frequently performed using the data housed in data warehouses. An analytical query-optimized star or snowflake schema is used to store data. Furthermore, the information kept in a data warehouse is frequently combined and aggregated to present the data comprehensively.

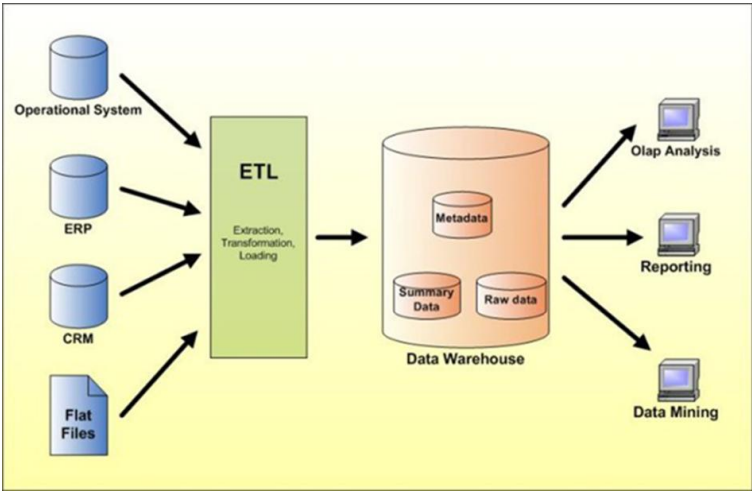


Fig. 2. Datawarehouse Architecture (Nambia & Mundra, 2022)

3.2.2. Data Lake

Unlike the data warehouse, the data lake (figure 3) is intended to hold any type of data, including unstructured, semi-structured, and structured data (Clay et al., 2021). A data lake's primary function is to hold unprocessed, raw data in its original format, which enables deeper analysis (Ravat & Zhao, 2019). Research and data analysis frequently use the data kept in data lakes. Without being combined or aggregated, data is stored in its unprocessed state. Instruments such as Hadoop, Spark, Pig, or Hive can be used for data analysis (Stein & Morrison, 2014).

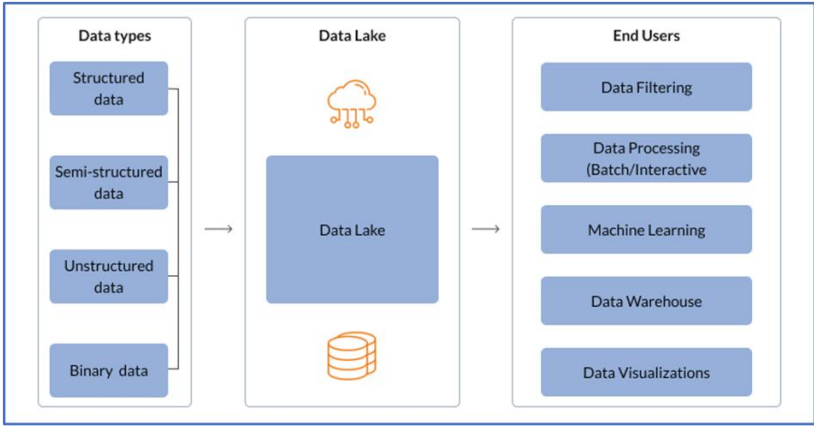


Fig 3. Data Lake Architecture[†]

[†] <https://www.n-ix.com/data-lake-vs-data-warehouse/>

3.2.3. Data Hub

The data hub (figure 4) is a hybrid data storage solution combining the advantages of a data warehouse and a data lake. Storing structured, semi-structured, and unstructured data while facilitating quick and simple access to the data for analysis is the primary function of a data hub (Bhardwaj et al., 2015).

Data can be quickly transformed into analytical data, so data hub data is frequently used for real-time analysis. Data is converted into analytical data and stored in its original format for easier-to-understand analysis. A snowflake or star schema is used to store data and is optimized for analytical queries.

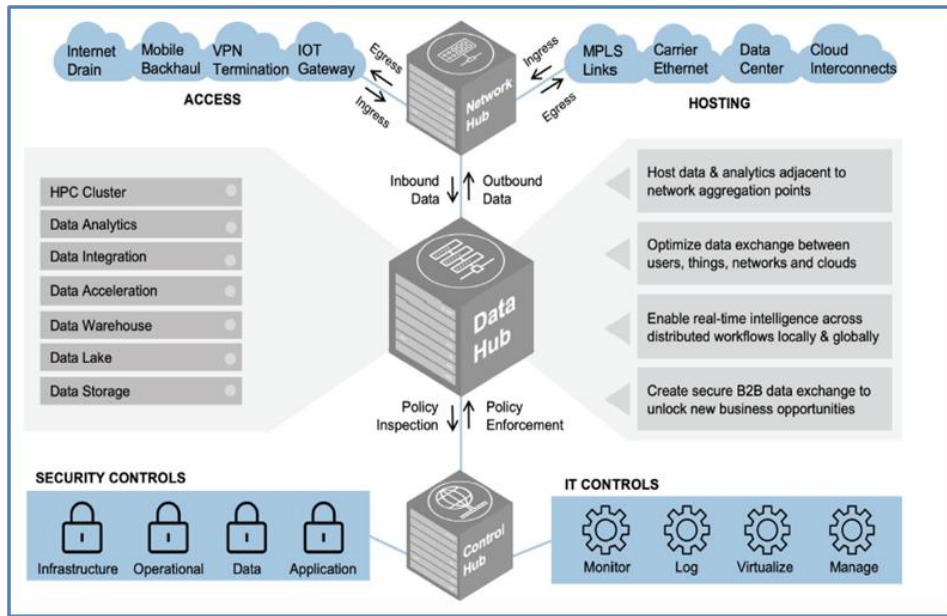


Fig. 4. Data Hub Architecture[‡]

Table 1 compares the tree data structures to several criteria, such as data type, organization, volume, and usage.

[‡] <https://www.prod.digitalrealty.fusionary.io/platform-digital/data-hub>

Table 1. A comparative table between the Data warehouse, Data Lake, and Data hub

Criterion	Data Lake	Data Hub	Data Warehouse
Data Type	Structured, unstructured.	Structured, semi-structured, unstructured.	Structured
Data Organisation	Not organized into a specific schema.	Organized flexible, extensible schema	Organized into a specific schema.
Data Volume	Suitable for handling significant data volume	Suitable for handling medium to large data volumes.	Suitable for handling smaller data volumes.
Querying and Analysis	It is not optimized for querying and analysis.	Optimized for data integration.	Optimized for querying and analysis
Updating	It may or may not be updated regularly	Regularly updating	Regularly updating.
Users	Data scientists, analysts.	Data analysts, data engineers.	Business analysis, decision-makers.
Primary Use Case	Data exploration, experimentation.	Data integration, ensuring data quality and consistency.	Business intelligence, reporting.

4. Case Studies

Data integration has received particular attention in the life sciences (Sbiti et al., 2021), where diverse data are produced at increasing rates. Nonetheless, data integration use cases are emerging in every industry and organization where technology solutions can now perform functions with vast amounts of information and in unavailable real-time situations. Here are some of them:

4.1. Healthcare

(Prasser et al., 2018) Describe a European research project called DIFUTURE data integration for Future Medicine (figure 5), which aims to improve personalized medicine by integrating genomic and clinical data from various sources, such as clinical trial data, electronic health records, and biobanks. This project is seen from the perspective of healthcare and medical research.

Common elements and ontologies are used to standardize and harmonize the data to facilitate consistent and precise data analysis.

A shared platform for storing, organizing, and integrating clinical and genomic data from diverse sources makes up the shared data infrastructure utilized in the DIFUTURE project. Researchers and clinicians can

easily find, access, and reuse the data for various purposes because this platform is based on the principles of FAIR data (Findable et al.).

The project also uses cutting-edge technologies like artificial intelligence, machine learning, and natural language processing to conclude the combined data. Machine learning algorithms, for instance, can be trained to find new biomarkers for specific diseases or to forecast disease risks based on clinical and genomic data. The data integration centers of the DIFUTURE project will integrate and exchange structured, unstructured, and omics data and images from clinical and research environments using a three-step process.

The first step is importing and harmonizing data using shared data and interface standards. The second step preprocesses, transforms, harmonizes, and enriches the data. By national interoperability standards, the third step entails importing the data into popular analytics models and platforms and making it available. Furthermore, the project will implement secure data access and sharing through cutting-edge privacy-enhancing technologies and distributed computing techniques.

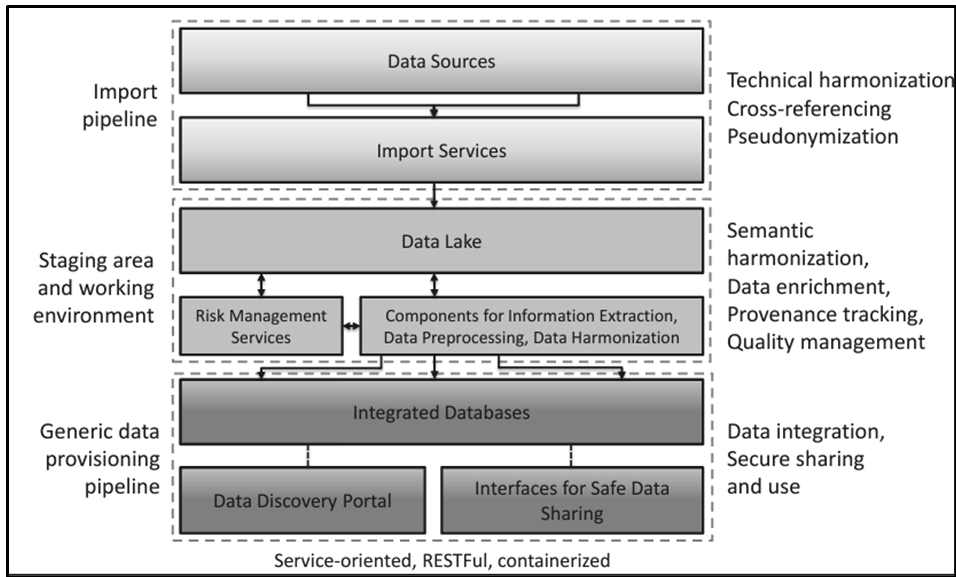


Fig. 5. Basic Architecture of DIFUTURE data integration center (Prasser et al., 2018)

By integrating various data types, this case study demonstrates how future medicine can be predictive, preventive, personalized, participatory, and digital, resulting in better patient outcomes and more informed decision-making.

4.2. Transport and Construction

In the context of transport infrastructure design, the authors (D'amico et al., 2020) presented a novel method for integrating BIM (Building information modeling) and GIS (Geographic information system) data for decision-making (figure 6). Combining data from many sources and domains, transforming it between formats, aligning it, and controlling the complexity of data integration are all part of integrating BIM and GIS data. To overcome these difficulties, the authors suggested a unified data model that consistently and uniformly integrates GIS and BIM data.

The unified data model provides a standard framework for organizing and managing data from different sources, enabling project stakeholders to access and analyze data more effectively. The model consists of interconnected data entities representing different aspects of the transport infrastructure project, such as the physical assets, the spatial context, and the environmental factors. A standardized schema that outlines their relationships and the attributes they contain is used to organize these data entities.

The authors suggest that middleware can enable data exchange between software applications supporting BIM and GIS technologies.

According to the authors, middleware can facilitate data interchange between software programs that support GIS and BIM technologies. Middleware provides a set of shared services and interfaces that simplify the integration of different applications and technologies. For example, in this case, it can be used to enable the transfer of data from a BIM software application to a GIS software application or to synchronize data between different BIM or GIS software applications.

The middleware selection is contingent upon the particular demands and specifications of the transportation infrastructure project and the resources and knowledge available to facilitate the middleware solution's implementation and upkeep. Large-scale transportation projects could benefit significantly from the performance, quality, and sustainability that BIM and GIS data integration can bring.

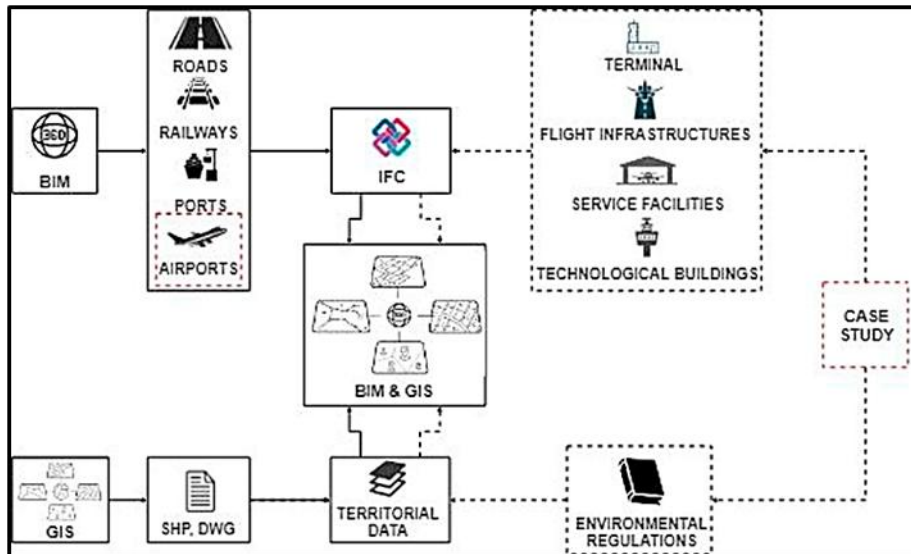


Fig. 6. Methodology of developing an airport infrastructure by integrating BIM and GIS data (D'amico et al., 2020)

4.3. Manufacturing

The study of (Mami et al., 2020) presents a method for integrating semantic data into the surface mount technology (SMT) manufacturing process using the SANSA stack (figure 7). The SANSA stack is an open-source semantic data processing infrastructure that enables high-level operations on semantic data. The authors applied this method to an electronic board manufacturing case study to integrate different data sources, such as production, quality, and real-time inspection data.

They used an ontology to represent a general schema of the various data sources. Ontology enables the representation of data semantically and formally; this allows the querying of the dataset within the light of the semantics and automatically. In addition, the inferencing mechanism allowed by ontology makes it possible to deduce additional knowledge.

The results showed that the proposed method allows for a better understanding of data and can help improve the quality and efficiency of the SMT manufacturing process.

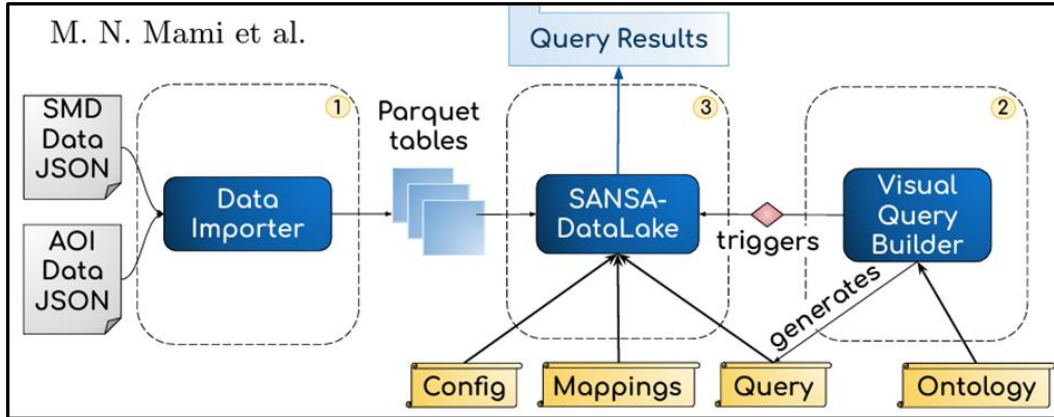


Fig. 7. General Architecture for the proposed solution (Mami et al., 2020)

5. Data Integration Challenges and Issues

Data integration poses several challenges and issues, primarily related to data heterogeneity, system compatibility, and high cost.

According to the user's needs, "fitness for use" defines data quality. As a result, the following seven factors should be taken into account:

The first factor that establishes relevance is how healthy data fulfills its intended purpose.

The second is accuracy, which denotes whether the features the data is intended to measure are accurately described by the data.

Thirdly, credibility measures users' faith in the data's objectivity and confidence in data products.

Fourth, timeliness is the amount of time that passes after the occurrence of the phenomenon that the data describes before it becomes available.

The ease of finding and accessing data is the fifth factor of accessibility.

Sixth, interpretability dictates how users appropriately comprehend, employ, and evaluate data. The last measure of coherence is how logically and consistently linked the data are.

Ensuring and enhancing data quality can be achieved by considering these dimensions to fulfill their intended function. Further complicating the integration process are potential incompatibilities between the IT systems utilized by various supply chain stakeholders. The high cost of data integration is another issue; a substantial investment is needed to create compatible systems and guarantee data security and quality.

Concerns about data security are also raised by integration, which raises the possibility of privacy violations and cyberattacks. Furthermore, it might be challenging to integrate the data if specific data is confidential due to confidentiality or competitive concerns. Conflicts over data ownership and use that emerge from data integration among supply chain participants may necessitate inappropriate conflict management.

6. Data Integration best practices

The researchers suggested several research directions to enhance data integration across different domains, including the application of cutting-edge technologies like blockchain, AI, and IOT that can transform data integration and offer fresh methods for data management (Slater et al., 2008; Maiti et al., 2022; Vieira et al., 2020).

Standardization and thorough documentation of data and analysis techniques are necessary to support integration and improve data integration even more (Clay et al., 2021). Furthermore, critical tactics for adapting to shifting client needs include data analytics, semantic modeling, and automated and scalable integration procedures. To define requirements and best practices, cross-industry collaboration is also advised (Maiti et al., 2022).

Scholars have proposed the use of lean production planning (LPS) and building information modeling (BIM) in construction project management to help with data integration (Sbiti et al., 2021). Stakeholder collaboration has improved with BIM's ability to integrate data from multiple sources and provide an all-encompassing view of the construction project. On the other hand, LPS offers a framework for process optimization and ongoing improvement, making it possible to manage construction projects effectively. These methods and tools can lower expenses, enhance project results, and simplify data integration. Combining BIM and LPS makes effective data integration possible, resulting in more productive and efficient project management.

Generally speaking, the six primary functions of collection, aggregation, and management can be used to distill the best practices of data integration into a manageable, dependable, and effective process:

1. Identification of data sources: Identifying all available data sources before beginning data integration is crucial. Deciding which data is most crucial and which is required is crucial.
2. Establish data quality guidelines: To guarantee that the integrated data is trustworthy and consistent, data quality guidelines must be established. As part of this, data formats may be defined: Selecting the integration tool that best suits the business's needs is crucial because many of them are available.
3. Form a data integration plan: A thorough plan will guarantee that each phase of the data integration process is handled correctly. For instance, the plan may comprise data collection, transformation, and validation steps.
4. Organizing metadata: Metadata details the data, including its source, format, and content. Ensuring sufficient documentation and comprehensibility of the integrated data is imperative for metadata management.
5. Monitor and audit data integration: Ongoing oversight and auditing of data integration can assist in identifying possible flaws and issues before they become serious.
6. Employee education: To guarantee that employees know the company's best practices and procedures, employee education is crucial for those working in the data integration process.

7. Conclusion

Companies must carefully plan and carry out the integration process by adhering to best practices involving data collection, aggregation, management, metadata management, data quality rules, data integration planning, monitoring, and auditing to achieve successful industrial data integration.

The management of metadata, staff training, dealing with various data sources, and guaranteeing data consistency and quality are just a few of the challenges that businesses must overcome in the data integration process.

To overcome these obstacles, businesses should consider investing in specialized equipment and software, setting up efficient lines of communication, using big data and cloud computing platforms, and routinely checking and auditing the data integration procedure.

References

- Sbiti, M., Beddiar, K., Beladjine, D., Perrault, R., & Mazari, B. (2021). Toward BIM and LPS data integration for lean site project management: A state-of-the-art review and recommendations. *Buildings*, 11(5), 196.
- Wang, J., Zheng, P., Lv, Y., Bao, J., & Zhang, J. (2019). Fog-IBDIS: Industrial extensive data integration and sharing with fog computing for manufacturing systems. *Engineering*, 5(4), 662-670.
- Reeve, A. (2013). Managing data in motion: data integration best practice techniques and technologies. Newnes.
- Doan, A., Halevy, A., & Ives, Z. (2012). Principles of data integration. Elsevier.
- Ziegler, P., & Dittrich, K. R. (2007). Data integration—problems, approaches, and perspectives. In *Conceptual modelling in information systems engineering* (pp. 39-58). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Slater, T., Bouton, C., & Huang, E. S. (2008). Beyond data integration. *Drug discovery today*, 13(13-14), 584-589.
- Nambiar, A., & Mundra, D. (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data and Cognitive Computing*, 6(4), 132.
- Clay, I., Angelopoulos, C., Bailey, A. L., Blocker, A., Carini, S., Carvajal, R., ... & Goldsack, J. C. (2021). Sensor data integration: a new cross-industry collaboration to articulate value, define needs, and advance a framework for best practices. *Journal of Medical Internet Research*, 23(11), e34493.
- Ravat, F., & Zhao, Y. (2019). Data lakes: Trends and perspectives. In *Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I* 30 (pp. 304-313). Springer International Publishing.
- Stein, B., & Morrison, A. (2014). The enterprise data lake: Better integration and deeper analytics. *PwC Technology Forecast: Rethinking integration*, 1(1-9), 18.
- Bhardwaj, A., Karger, D., Subramanyam, H., Deshpande, A., Madden, S., Wu, E., ... & Zhang, R. (2015, August). Collaborative data analytics with DataHub. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* (Vol. 8, No. 12, p. 1916). NIH Public Access.
- Altexsoft. Data science, machine learning, and AI: Altexsoft blog. <https://www.altexsoft.com/blog/data-hub/>, 2021. (Accessed: December, 2023).
- Prasser, F., Kohlbacher, O., Mansmann, U., Bauer, B., & Kuhn, K. A. (2018). Data integration for future medicine (DIFUTURE). *Methods of information in medicine*, 57(S 01), e57-e65.
- D'Amico, F., Calvi, A., Schiattarella, E., Di Prete, M., & Veraldi, V. (2020). BIM and GIS data integration: a novel approach of technical/environmental decision-making process in transport infrastructure design. *Transportation Research Procedia*, 45, 803-810.
- Mami, M. N., Grangel-González, I., Graux, D., Elezi, E., & Lösch, F. (2020, May). Semantic data integration for the SMT manufacturing process using SANSa stack. In *European Semantic Web Conference* (pp. 307-311). Cham: Springer International Publishing.
- Maiti, M., Vuković, D., Mukherjee, A., Paikarao, P. D., & Yadav, J. K. (2022). Advanced data integration in banking, financial, and insurance software in the age of COVID-19. *Software: Practice and Experience*, 52(4), 887-903.
- Vieira, A. A., Dias, L. M., Santos, M. Y., Pereira, G. A., & Oliveira, J. A. (2020). Supply chain data integration: A literature review. *Journal of Industrial Information Integration*, 19, 100161.
- Halevy, A., Rajaraman, A., & Ordille, J. (2006, September). Data integration: The teenage years. In *Proceedings of the 32nd international conference on Very large data bases* (pp. 9-16).
- Daraio, C., & Glänzel, W. (2016). Grand challenges in data integration—State of the art and future perspectives: An introduction. *Scientometrics*, 108(1), 391-400.