Jurnal Kendali Teknik dan Sains
Vol. 2, No. 3 Juli 2024
e-ISSN: 2986-3228; p-ISSN: 2986-4429, Hal 08-19
DOI: https://doi.org/10.59581/jkts-widyakarya.v2i3.3223

Cloud Data Integration Tools As Services

Nisreen Nizar Raouf
College of Computer Science and Mathematics, University of Mosul
Mosul, Iraq

Mohammad A. Taha Aldabbagh
College of Computer Science and Mathematics, University of Mosul
Mosul, Iraq

Corresponding author: nisreen.21csp9@student.uomosul.edu.iq , m.a.taha@uomosul.edu.iq

Abstract: *In this paper, a variety of data integration, cloud computing, and web services-related topics were discussed. The conversation covered the advantages of using cloud-based methods for cloud-based data integration, such as enhanced data accuracy and completeness, as well as the challenges and considerations that must be addressed. In addition, the significance of document integrity and the use of auto-enhance document tools to ensure data accuracy were emphasized. The paper also provided a broad overview of the subject, touching on a variety of aspects and providing insights into the potential of cloud-based data integration methods in the cloud industry.*

Keyword's: *Data integration tools, cloud-based data, Auto-enhance document, Common Data Format, and web-based data integration.*

I. INTRODUCTION

A data integration tool is a piece of software that facilitates the integration of data from various sources into a unified view [1]. The objective of data integration is to provide a unified view of data that can be accessed, analyzed, and utilized by multiple applications and systems [2, 3]. Data integration tools are essential to the development and maintenance of software applications that require access to data from multiple sources. These tools can assist developers and engineers in overcoming the challenges of integrating data from multiple sources, such as distinctions in data formats, data schemas, and data semantics [4].

In order to integrate data from various sources into a cloud-based system, cloud data integration applications are used. Data connectors move data from one database to another and handle transformations [3]. Data catalogues help businesses locate and inventory data assets across multiple silos. One of the primary advantages of data integration tools is that they provide a unified view of data that is readily accessible, analyzable, and usable by multiple applications and systems [2]. This enhances the efficacy, precision, and efficiency of software development and maintenance processes. Also, data integration tools reduce the time and effort required to integrate data from various sources, which can be a complex and time-consuming task [5].

Received April 30, 2024; Accepted May 17,2024; Published July 31, 2024
* *Nisreen Nizar Raouf nisreen.21csp9@student.uomosul.edu.iq*

Without data integration technologies, current software systems cannot manage data [3]. They offer a comprehensive collection of features and functionalities that can significantly improve the efficacy, precision, and efficiency of software development and maintenance processes. Data integration tools help software programmers overcome the issues associated with integrating data from numerous sources, such as differences in data formats, data schemas, and data semantics, by giving a unified view of data from various sources [1, 4].

Cloud data integration solutions are software applications that help you combine diverse cloud-based data sources into a cohesive picture. Data mapping and transformation, data synchronization, data quality and validation, workflow automation, and security and compliance are some of the features and capabilities provided by these technologies to help organizations manage their cloud data integration requirements [3]. There are various possibilities for choosing the finest framework for Python-based web-based data integration services. Django, Flask, Pyramid, and Bottle are among prominent frameworks. Each of these frameworks has advantages and disadvantages, therefore it is critical to select the one that best meets the organization's goals. Consider the project's complexity, the level of customization necessary, the framework's scalability, and the level of community support available for the framework [2, 4, 5].

II.FRAMEWORK IN CLOUD METHODS FOR DATA INTEGRATION

Data integration is the process of incorporating data from multiple sources to aid data administrators and executives in analyzing the data and making more informed business decisions. This process involves a person or computer retrieving, cleaning, and presenting the data [1]. There are several methods and strategies for data integration, such as ETL (Extract, Transform, Load), EAI (Enterprise Application Integration), EII (Enterprise Information Integration), and others. [1].

Cloud-based frameworks facilitate the integration, analysis, and visualization of data. These services can be either end-user applications or platforms for the development of new services [2]. Integration of cloud-based data is comprised of tools and technologies that link disparate applications, systems, data repositories, and IT environments. It enables the real-time exchange of data and processes. Once unified, multiple devices can access data and cloud services via a network or the internet [3]. Cloud computing is the transmission of computing services such as servers, storage, databases, networking, software, analytics, and intelligence

over the Internet ("the cloud") to provide quicker innovation, flexible resources, and economies of scale [4, 5].

Data integration is the process of incorporating data from multiple sources to aid data administrators and executives in analyzing the data and making more informed business decisions. This procedure involves locating, retrieving, cleansing, and displaying the data. Methods and strategies for data integration consist of ETL (Extract, Transform, Load), EAI (Enterprise Application Integration), EII (Enterprise Information Integration), and others. [5].

Cloud-based frameworks facilitate the integration, analysis, and visualization of data. These services can be either end-user applications or platforms for the development of new services. Integration of cloud-based data is comprised of tools and technologies that link disparate applications, systems, data repositories, and IT environments. It enables the real-time exchange of data and processes. Once unified, multiple devices can access the data and cloud services via a network or the internet [6].

Integration of data is important because it enables organizations to combine data from diverse sources in order to gain insights that would not be possible otherwise [7]. It enables businesses to make more informed decisions by providing a more comprehensive view of their data [7, 8]. Data integration is challenging for these reasons:

Data quality issues: varied sources have varied formats, structures, and quality levels. This can cause integrated data discrepancies and mistakes [9].

Data security concerns: Data integration necessitates sharing data across systems and organizations, which increases the risk of data breaches and cyberattacks [10].

Technical complexity: Data integration requires data mapping, transformation, and synchronization. Specialized expertise and tools are needed [11].

Cost: Data integration requires expensive technologies and experience [11].

Time-consuming: Data integration requires data purification, transformation, and synchronization [9], [11].

Cloud-based data integration frameworks employ ETL techniques to combine data from diverse sources. ETL collects data from diverse sources, converts it into a format the destination system can utilize, and loads it [12]. Cloud-based data integration frameworks employ ETL techniques to combine data from diverse sources. ETL collects data from diverse sources, converts it into a format the destination system can utilize, and loads it [12]. Data extraction,

transformation, loading, validation, and verification are common framework phases. ETL (extract, transform, load) and API (application programming interface) cloud-based technologies may automate these procedures and integrate data easily. Frameworks in cloud data integration provide several advantages [13].

Improved data quality: The framework guarantees data is accurate, full, and consistent across all systems, enhancing data quality.

Reduced mistakes and redundancy: The framework eliminate data entry errors and redundancy.

Increased efficiency: Cloud-based solutions like ETL and API automate data integration procedures, enhancing efficiency.

Scalability: The framework can handle massive amounts of data and complicated data integration [12].

Overall, a cloud data integration framework is a crucial tool for organizations wishing to integrate data from many sources in the cloud, allowing them to make better decisions based on accurate, consistent, and trustworthy data. There are several examples of cloud computing and data integration. Here are some cloud computing examples [13]:

1. Amazon Web Services (AWS)
2. Microsoft Azure
3. Google Cloud Platform (GCP)
4. IBM Cloud

Here are some examples of data integration [9], [13]:Google Cloud

1. Informatica
2. NetApp
3. Informatica Cloud
4. Dell Boomi
5. Microsoft Azure Data Factory
6. SnapLogic
7. Talend Cloud
8. MuleSoft Anypoint Platform

There are various advantages of using cloud-based data integration technologies. Here are a few examples [13]: Lowering expenses and complexity; increasing agility and flexibility; improving performance and quality; scalability; and cost-effectiveness

The initial goal of data replication in data could was to keep two data stores in sync with one another while passing as little information as possible between them and having as little influence on the source application as possible.

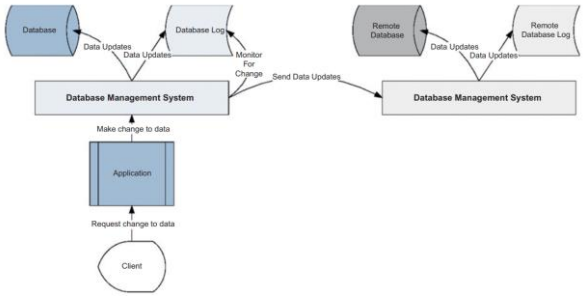

Fig 1. Change Data Capture [28].

Change data collection, as illustrated in Fig 1, is a very effective method of allowing data to be utilised across far sites; it was especially important at a period when wide area network access was quite sluggish. Emerging technologies are discovering that this is an effective solution to solve many additional latency and impact concerns. Because the source data structure is not accessible, the impact on the source system is limited, and it is helpful when the source system's reaction time cannot be altered [27]. Several benefits are associated with cloud-based data integration methods. According to LinkedIn, cloud-based data integration platforms offer several benefits over traditional on-premise or hybrid solutions, including reduced costs and complexity, increased agility and flexibility, and enhanced performance and quality [13]. In addition, cloud-based data integration can offer several advantages over traditional on-premise data integration, including scalability, flexibility, agility, and cost-efficiency [10].

III. AUTO-ENHANCE DOCUMENT AS A THIRD-PARTY DATA INTEGRATION TOOL

Auto-enhance document is not a tool for third-party data integration that requires special attention. To integrate data from multiple sources into a single system, however, third-party data integration tools may be used. These tools can assist organizations in enhancing their data quality and reducing the time and expense associated with manual data integration processes [14]. Salesforce Platform [14, 15], Snowflake [15], and Clearbit [16] are a few examples of third-party data integration tools. These tools can assist businesses in extracting data from various sources, transforming it as necessary, and loading it into a target system. Data cleansing and enrichment capabilities offered by third-party data integration tools can also aid organizations in enhancing the quality of their data.

Fig 2 depicts a hypothetical environment arrangement during application and conversion development. Application developers will require environments for unit testing as well as integrated system testing. In most cases, a single test environment may be used for both unit and integrated system testing. Depending on how application testing is organized, different environments for QA testing and user acceptability testing may be requested. However, this can generally be coordinated to build a single environment that can be utilized for various testing phases of the project. After the programmed has been turned on for production operation, at least two test environments normally exist independently from the production environment: the unit/system testing environment (also known as development) and the QA/user acceptance testing environment [15, 17].

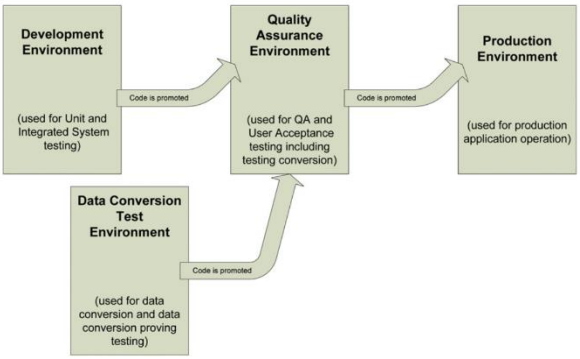

Fig 2. Environments Data Flow [27].

Utilizing a third-party data integration tool that can help organizations extract data from various sources, transform it as necessary, and load it into a target system is one method to enhance data integration. These tools can also assist organizations in enhancing the quality of their data by offering features such as data purification and enrichment [17]. Using auto-enhance document tools has the added benefit of ensuring that data entered into the system is accurate and consistent. The "gold standard" for a document is one that can be parsed to extract discrete data elements that can be incorporated into an electronic health record using standard mapping and conversion methods. This serves to ensure the accuracy and consistency of the data, reducing the likelihood of errors and inconsistencies that can occur when data is manually entered [14].

In addition to enhancing the precision and completeness of data integration, auto-enhance document tools can also enhance the integrity of documentation. This includes patient identification, validation of authorship, record amendments and corrections, and auditing. By outsourcing these processes, organizations can ensure that the data in the complete health record

is accurate and comprehensive, thereby enhancing the quality of patient care and operational efficiency [17]. Using a master data management (MDM) system is a second method for improving data integration. By providing features such as data governance, data quality management, and metadata management [18], an MDM system can help organizations manage their master data more effectively. Auto-enhance document tools can improve the precision and thoroughness of data integration by providing a standardized format for data entry and reducing the likelihood of data entry errors. A well-designed document may, for instance, include prompts or instructions for what data to capture and group logically related information near together [19]. In addition, the "gold standard" for a document is one that can be parsed to extract discrete data elements that can be incorporated into an electronic health record using standard mapping and conversion techniques, ensuring the accuracy and consistency of the data [20]. Documentation integrity, which includes patient identification, authorship validation, amendments and record corrections, and auditing, is also essential for ensuring the completeness and accuracy of the health record [19, 20]. By outsourcing the enhancement of documents, organizations can expedite data collection and ensure that the data entered into their systems is accurate and comprehensive, thereby enhancing the quality of patient care and operational efficiency [20].

In general, auto-enhance document tools are a valuable solution for increasing the precision and completeness of data integration in a variety of industries, including healthcare. By providing standardized formats for data input, reducing the likelihood of data entry errors, and automating the process of data collection, organizations can ensure that the data in their systems is accurate and consistent, thereby enhancing patient care and overall operational efficiency.

IV. IMPLEMENTATION OF COMMON DATA FORMAT TYPE

The Common Data Format (CDF) is a self-describing data format used for storing and exchanging data between applications. Many organizations have adopted CDF as a standard format for storing and exchanging scientific data [19]. CDF is extensively utilized in scientific research. It is a conceptual data abstraction used to store, manipulate, and access multidimensional data sets. CDF's fundamental component is a software programming interface that provides a device-independent perspective of the CDF data model. Besides the actual data being stored, CDF also retains user-supplied descriptions of the data, known as metadata [21]. The process of integrating data from various cloud-based sources. Implementing CDF for cloud-

based data integration can facilitate the exchange of scientific data between cloud-hosted applications [22].

The CDF is a platform- and domain-independent self-describing data format for storing scalar and multidimensional data [21]. The CDF metadata system enables data and its meaning to be shared readily across applications and business processes [22]. Using a common data format such as CDF can provide several benefits for data integration, including improved consistency and accuracy, reduced duplication and redundancy, increased completeness and coverage, facilitated integration and interoperability, streamlined management and maintenance, and enhanced usability and value [21]. The Common Data Model (CDM) metadata system enables data and its meaning to be readily shared between applications and business processes [22]. In Azure Data Factory and Synapse pipelines, users can transform CDM entity data stored in model.json or manifest format in Azure [21, 23]. The translation of several data sources in diverse formats into a single target data set is depicted in Fig 3 Many data transformations may be accomplished simply by changing the technical format of the data, but as illustrated in the image, additional information is usually necessary to determine how the source data should be converted from one set of values to another [28].

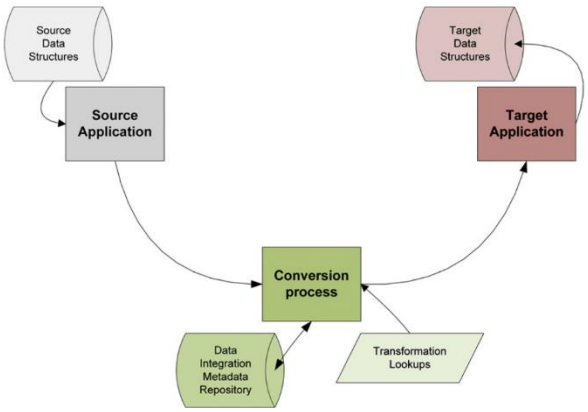

Fig 3. Migrating Data from One Application to Another [28].

While the data conversion process interacts with the source and target application systems to transport and transform data from the source system's technical format to the target system's format and structure. This is great practice, especially if developer want to execute a data update Transforming Data into a Common Format. What Exactly Is Data Integration? rather to directly modifying the target data structures, there are situations when the data migration procedure directly interacts with the source or target data structures rather than the application interfaces.

The Common tools and frameworks include [22, 23]: Apache Avro, Apache Parquet, Apache Thrift, Protocol Buffers Apache Arrow, and JSON (JavaScript Object Notation). While the popular CDF frameworks include: Hadoop, Spark, Kafka, and Flink. These technologies and frameworks help organizations create a standardized, interoperable CDF, improving data quality, efficiency, and interoperability [22].

V. PYTHON-BASED WEB-BASED DATA INTEGRATION SERVICES

Python-based web-based data integration services can be employed in real-time systems. You may obtain, analyses, update, and manipulate any online service's data using Python and REST APIs [23]. Python-based web-based data integration services allow customers to combine data from different sources utilizing Python programming language and online technologies. Data from databases, spreadsheets, and online services may be extracted, transformed, and loaded using these technologies. [23, 24, 25] are Python-based web-based data integration services.

Apache Nifi: This open-source data integration platform lets users collect, process, and disseminate data from several sources. It supports JSON, XML, and CSV data formats and has a simple interface for data integration.

Talend Open Studio: This free, open-source data integration application lets users develop, launch, and manage data integration procedures. Data integration processes may be designed using a drag-and-drop interface and support databases, online services, and cloud storage.

Apache Airflow: Users may schedule, monitor, and manage data operations using Apache Airflow, an open-source data integration platform. Data pipelines may be created using Python and support databases, cloud storage, and APIs.

The finest Python-based web-based data integration services framework relies on the project's needs. Pyramid, Flask, and Django are prominent frameworks. For small to medium-sized applications, Flask is a lightweight framework. For bigger applications, Django is a more complete framework with many built-in features [24]. Pyramid can be customized to meet project needs. Based on standard type hints, FastAPI is a contemporary, high-performance Python web framework for API development. Its main features are: Fast running: Thanks to Starlette and pydantic, it performs like NodeJS and Go. It's fast to code [24]. Python-based web-based data integration services may automate the onerous, repetitive procedures of ingesting, duplicating, and synchronizing data throughout the company, speeding up data integration in real-time systems. These services enable organizations to combine data from diverse sources

and formats [25, 26]. For instance, FastAPI is a contemporary, high-performance Python web framework for developing APIs using type hints. Thanks to Starlette and pydantic, it performs like NodeJS and Go. Development pace is greatly increased [27]. Web development and data integration are done with Django, another Python-based web platform.

VI. CONCLUSION

The Cloud computing and real-time systems require effective data integration solutions to unify data from many sources and formats. This can enhance decision-making, customer service, and cost-cutting. Data integration in cloud computing lets organizations instantly access data from many sources and formats. This can enhance decision-making and speed up corporate responses. Data integration helps real-time systems digest enormous volumes of data fast. This can boost operating efficiency and cut expenses.

Cloud data integration frameworks help ensure data quality and completeness. Auto-enhance document has made data entry more standardized and error-free, enhancing patient care and operational efficiency. The Common Data Format type has also improved data uniformity and accuracy. Real-time systems also use Python-based web-based data integration services. Cloud-based data integration technologies have improved data integration efficiency and accuracy, helping organizations make better decisions. Data integration will grow further with sustained study and development.

Advanced data integration tool and technique development has various research opportunities. These include developing more efficient data integration algorithms, using machine learning and artificial intelligence to automate data integration processes, and creating new data integration architectures that can handle large amounts of data in real time.

VII. REFERENCES

[1] 5 data integration methods and strategies (no date) Talend. Available at: https://www.talend.com/resources/data-integration-methods/(Accessed: April 29, 2023).

[2] Goldstein, L. Fink and G. Ravid, "A Cloud-Based Framework for Agricultural Data Integration: A Top-Down-Bottom-Up Approach," in IEEE Access, vol. 10, pp. 88527-88537, 2022, doi: 10.1109/ACCESS.2022.3198099.

[3] Ashraf, P. byS., Nadeem, P. N., & Ashraf, P. S. (2020, August 18). Cloud Data Integration: How It Works & Why is it needed? Data Integration Blog. Retrieved May 1, 2023, from https://dataintegrationinfo.com/cloud-data-integration/

[4] Pereira, J., Batista, T., Cavalcante, E., Souza, A., Lopes, F., & Cacho, N. (2022). A platform for integrating heterogeneous data and developing smart city applications. Future Generation Computer Systems, 128, 552-566.

[5] Data reference architecture. IBM. (n.d.). Retrieved May 2, 2023, from https://www.ibm.com/cloud/architecture/architectures/dataArchitecture/reference-architecture

[6] Data reference architecture. IBM. (n.d.). Retrieved May 2, 2023, from https://www.ibm.com/cloud/architecture/architectures/dataArchitecture/reference-architecture

[7] Michaels, G. (2022, March 21). Data integration: What it is and why it matters. Unito. Retrieved May 2, 2023, from https://unito.io/blog/data-integration/

[8] Murphy, G. (n.d.). Data Integration: Benefits, challenges, and considerations. Confluent. Retrieved May 2, 2023, from https://www.confluent.io/blog/data-integration-benefits-challenges-considerations/

[9] Farmer, D. (2022, December 15). 8 data integration challenges and how to overcome them. Data Management. Retrieved May 2, 2023, from https://www.techtarget.com/searchdatamanagement/feature/5-data-integration-challenges-and-how-to-overcome-them

[10] Bello, S. A., Oyedele, L. O., Akinade, O. O., Bilal, M., Delgado, J. M. D., Akanbi, L. A., ... & Owolabi, H. A. (2021). Cloud computing in construction industry: Use cases, benefits and challenges. Automation in Construction, 122, 103441.

[11] Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. Frontiers in medicine, 8, 3082.

[12] Seenivasan, D. (2023). ETL (Extract, Transform, Load) Best Practices. International Journal of Computer Trends and Technology, 71(1), 40-44.

[13] Making, D.-driven D. (n.d.). What are the benefits and challenges of cloud-based data integration platforms? Cloud-Based Data Integration: Benefits and Challenges. Retrieved May 2, 2023, from https://www.linkedin.com/advice/3/what-benefits-challenges-cloud-based-1e

[14] Tan, B., Anderson Jr, E. G., & Parker, G. G. (2020). Platform pricing and investment to drive third-party value creation in two-sided networks. Information Systems Research, 31(1), 217-239.

[15] Guides. Data Integration | Snowflake Documentation. (n.d.). Retrieved May 2, 2023, from https://docs.snowflake.com/en/user-guide/ecosystem-etl

[16] Ozsahan, H. (2023, February 21). 14 Best Data Enrichment Tools of 2023 & why you need them. Popupsmart. Retrieved May 2, 2023, from https://popupsmart.com/blog/data-enrichment-tools

[17] Goldfedder, J., & Goldfedder, J. (2020). Choosing an ETL Tool. Building a Data Integration Team: Skills, Requirements, and Solutions for Designing Integrations, 75-101.

[18] Ng, S. T., Xu, F. J., Yang, Y., & Lu, M. (2017). A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning. Procedia engineering, 196, 939-947.

[19] Anvari, Z., & Athitsos, V. (2021). A survey on deep learning based document image enhancement. arXiv preprint arXiv:2112.02719.

[20] Gangeh, M. J., Tiyyagura, S. R., Dasaratha, S. V., Motahari, H., & Duffy, N. P. (2019, November). Document enhancement system using auto-encoders. In Workshop on Document Intelligence at NeurIPS 2019.

[21] Trautwein, D., Raman, A., Tyson, G., Castro, I., Scott, W., Schubotz, M., ... & Psaras, Y. (2022, August). Design and evaluation of IPFS: a storage layer for the decentralized web. In Proceedings of the ACM SIGCOMM 2022 Conference (pp. 739-752).

[22] Liu, Y. K., Ong, S. K., & Nee, A. Y. C. (2022). State-of-the-art survey on digital twin implementations. Advances in Manufacturing, 10(1), 1-23.

[23] Kromerm. (n.d.). Common data model format - azure data factory & azure synapse. Common Data Model format - Azure Data Factory & Azure Synapse | Microsoft Learn. Retrieved May 2, 2023, from https://learn.microsoft.com/en-us/azure/data-factory/format-common-data-model

[24] 24- Real Python. (2022, May 7). Python and REST apis: Interacting with web services. Real Python. Retrieved May 2, 2023, from https://realpython.com/api-integration-in-python/

[25] Real Python. (2023, April 28). Using fastapi to build python web apis. Real Python. Retrieved May 2, 2023, from https://realpython.com/fastapi-python-web-apis/

[26] Orme, T. (2021, February 9). How data integration can improve data strategies long-term. TechRadar. Retrieved May 2, 2023, from https://www.techradar.com/news/how-data-integration-can-improve-data-strategies-long-term

[27] Middleware. 2023, April 28). (n.d.). How do web services improve data integration and interoperability for business? www.linkedin.com. https://www.linkedin.com/advice/1/howdowebservicesimprove-data-integration-interoperability

[28] Reeve, A. (2013). Managing data in motion: data integration best practice techniques and technologies. Newnes.