International Journal of Engineering & Technology, 12 (2) (2023) 26-30

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Data integration: "Seamless data harmony: The art and science of effective data integration"

Saloni Kumari *

Software Engineer II at EY (Ernst & Young), Hyderabad, India
**Corresponding author E-mail: salonisingh899@gmail.com*

Abstract

The idea of data integration has evolved as a key strategy in today's data-driven environment, as data is supplied from various and heterogeneous sources. This article explores the relevance, methodology, difficulties, and transformative possibilities of data integration, delving into its multidimensional world. Data integration serves as the cornerstone for well-informed decision-making by connecting heterogeneous datasets and fostering unified insights. This article gives readers a sneak preview of the in-depth investigation into data integration, illuminating its technical complexities and strategic ramifications for companies and organizations looking to maximize the value of their data as-sets.

Keywords: *Data Analytics; Data Processing; Data Storage; NoSQL Database; Distributed Computing; Scalability; Fault Tolerance; Data Warehousing; Data Ingestion; Workflow Scheduler; Coordination Service; Big Data Architecture; Hadoop Ecosystem.*

1. Introduction

It is obvious that our society is driven by data, and as we steadily move toward fully digitalized lives, data is becoming a valuable resource for the contemporary economy. We create important data whenever we use the internet to make purchases, view content, or share it on social media. Many of the biggest internet companies increasingly rely on data-driven business models to run their operations. However, without data integration, none of it is possible. Data integration is the glue that allows raw data to be transformed into an asset. Lack of data integration results in a host of business issues.

Due to fragmented data silos between organizations or departments within companies, it becomes necessary for users to rekey data or duplicate their efforts. Making decisions may be difficult in the absence of uniform data views. When individuals or departments only have partial access to data, they frequently make judgments that don't consider the overall process and are therefore suboptimal. Businesses waste a lot of money due to inefficiencies and bad decisions that result from poor data integration. Techniques for data integration aid in minimizing these issues. The process of obtaining data from many sources and transforming it into a data store or business application so that it can be used more efficiently is known as data integration. Three different kinds of data integration exist: Business-to-business integration entails cross-organizational linkages to improve the efficiency of business transactions between trading partners. The goal of application integration is to connect different corporate applications to create an integrated process. The data store itself is the level at which database integration takes place. Building pipelines to transfer raw data between data stores falls under this category. When developing data warehouses and business intelligence systems, this kind of integration is used. In the contemporary workplace, all three forms of integration are regularly used and are very beneficial to comprehend. In the current business environment, businesses that excel at data integration will have a significant competitive advantage.

Let's imagine a producer of agricultural equipment wants to use the information gathered by its tractors to improve crop yields. Perhaps sensors on the tractors can assess the soil moisture. The producer may gather this information from all of their tractors and integrate it with information from other sources, such as weather or commodity market prices. The ultimate result may be advice for farmers on how to improve the efficiency of their irrigation systems to maximize harvests, or the data could even be sold to outside parties like hedge funds to assist them in making smarter investment choices. This may develop into a new revenue stream for the corporation that would eventually complement or even outperform its current business strategy. That is an illustration of how digitization may be used to transform industries, and it all hinges on data fusion.

2. Research methodology

2.1. Business integration

Business to business, or B2B, integration permits the electronic interchange of commercial transactions between two or more trading partners, such as orders or payments. To accommodate certain scenarios, B2B messaging occasionally needs the extra flexibility provided

by XML or API standards. As businesses rely more and more on alliances or intricate supply chains to hasten entrance into new markets and boost competitive advantages without B2B messaging, B2B integration is becoming more and more crucial. The players in the supply chain ultimately communicate manually via email or by sharing Excel attachments.
B2B enables unconnected businesses to link their separate business systems into a cohesive workflow. A customer could send a B2B communication, including a purchase order, to the supplier's production system from their ERP system. A purchase order acknowledgement may be sent to the customer automatically by the supplier system after checking production schedules. Trading partners can handle large numbers of transactions with less labor-intensive manual labor and less error-prone automation.
Since it must handle the extra challenges of delivering data across corporate boundaries, B2B integration differs from application-to-application integration and intra-company database connections.
B2B integrations must check trading partner communications for compliance with CDI requirements, acknowledge trade partners, monitor the progress of messages, and transmit data securely.

2.2. Application integration

Any type of software used to carry out tasks is referred to as an application. This includes corporate applications like CRM or ERP, iPhone or Android mobile apps, and cloud services like MailChimp or Google Analytics. Due to how essential software has become to our every-day lives, it is usually required to use many applications to execute activities. These programmes are connected by application integration, which creates an efficient workflow.

2.3. Database integration

Simply merging data from numerous sources into one unified perspective to produce insights might be referred to as database integration. It gathers information from many sources and changes it into something more worthwhile and beneficial.
Most database integration strategies fit into one of these groups. Data from different storage locations is combined into one data repository through data consolidation. ETL, or extract, transform, and load, is a typical strategy for data consolidation. A specific type of data consolidation called data warehousing gathers data from numerous systems and merges it into a single storage engine that is intended to allow analytical queries.
Moving data from one place to another is known as data propagation, and replication is a frequent type of data propagation.
Replication tools that can automatically sync data from an origin source to a destination source are available in the majority of relational databases. This is frequently used to help with disaster recovery or boost data access performance. In contrast to data consolidation, which transfers data into a single data source, data virtualization offers a single view of data across multiple data sources. Users can operate with a façade that is created through data virtualization. Behind the scenes, it retrieves data from the many data sources.
In contrast to data virtualization, data federation enforces a single data model across all of the diverse data sources. Database workloads are frequently divided into one of two categories: LTP or OLAP.
A database that handles common corporate transactions, such as an ERP or CRM system, is referred to as old TPE, or online transaction processing. A mix of database reads and writes involving recent business transactions, such as orders or leads, is typical of old-style workloads.
Online analytical processing, or OLAP, is primarily concerned with reporting and analytics. OLAP tasks entail database reads across big data sets, such as queries on the daily average of orders over the last 12 months.
Data warehouses are made to serve these OLAP workloads. For instance, getting data from the billing system would require a very time-consuming query if we needed a report that compared the total revenue for this year to the previous year.
It might be necessary to add up millions of bills from the previous two years. Or perhaps there are different billing systems in use in various locations, and the revenue figures need to be added together to provide a whole picture. All of these data sources would be combined into a database with a data warehouse, allowing for rapid queries on millions of entries. These queries can be executed far more quickly than source systems using certain technologies. They sometimes combine the outcomes.
This implies that the billions of billing transaction rows are condensed in a batch process, and the annual revenue total is saved in the database as a single figure for easy retrieval.
Star schemas, which describe the arrangement of tables and columns in the database, are frequently used by data warehouses for their data models. In a star schema, data is kept in two types of tables: facts and dimensions. Fact tables keep track of important financial data like income. This design results in a large number of rows for the fact table and a relatively smaller number of rows for the dimension tables, which makes for simpler queries, query performance gains, and faster aggregations, operations where you sum up rows, like in our example, where we needed the total revenue for a year. This design results in many rows for the fact table and a relatively smaller number of rows for the dimension tables.
Data warehouses are frequently replaced by technologies like Hadoop or Spark. By splitting data over clusters of servers rather than transforming and loading data into a data warehouse, Hadoop enables rapid queries on extremely big datasets.

2.4. ETL extract, transform, load

ETL, which stands for Extract, Transform, and Load, is the procedure for loading a data warehouse and entails polling data from the source systems and putting it into a staging area.
To transform data is to prepare it ready for loading into the target system. And loading refers to putting data into the intended system.
The process of extracting involves taking data from multiple source systems and putting it into processing staging regions. On-site source systems are an example of a source system. ERP programmes such as SAP, cloud applications such as Salesforce, CSV files, or SQL databases. The business's operating data is contained in these source systems. For huge data sets, care must be taken not to negatively affect the performance of the source system. The extract process will read data from these systems using a variety of ways and write the data into a file system or database for the following stage in the processing pipeline. Typically, data is retrieved from these extracts in its original format and promptly stored in new staging storage in that format before any transformations are applied. This lowers the amount of computer power needed to extract the data from the source system. Although the full data set may occasionally be recovered in one batch from the source system, it is typically preferable to use a change data capture process in the source system. to handle newly added, updated, or removed records. Data must be transformed before being put into the target system. Data transformations can take a variety of forms. First,

data purification is required for this type of transformation to prevent the loading of faulty data into the target system. Eliminating faulty records, getting rid of duplicates, or resolving formatting issues are common cleansing chores.

Enrichment of data is frequently essential to enrich the source data before it is loaded. This entails adding information to the data that was not included in the source system but is required to be loaded into the target system. For instance, the customer's address might be provided by the source system, but the target system might need the GPS coordinates, latitude, and longitude of the address to geocode the address data and collect GPS coordinates prior to loading data.

Large data sets can be put into the target database after the data has been extracted and converted. Even while it's common to write data into a relational database using SQL statements like insert, update, and delete, loading 300,000 insert statements to load 300,000 records for an ETL task will be slow and use resources on the target system that could have an influence on performance. Many databases have specialized bulk load capabilities that assist in the effective loading of massive data sets.

Managing master data and foreign key relationships is one frequent problem. In a database, master data refers to reference information that is used across several tables.

A foreign key is a referential integrity database constraint that makes sure a reference value from one table is present in another related table. When loading new orders in an e-mail process, for instance, a customer mentioned on an order must be present in the customer master table. It is important to manage the key correctly since it links the orders customer field to the customer master table. In data warehouses, certain kinds of data links are handled in a specific fashion. In fact, tables linked to the dimensions by a foreign key, master data is frequently kept in dimension tables.

Master data changes over time, as well as their connections to business operations, are frequently captured using a technique termed "slowly evolving dimensions."

The most common way to enter data into data warehouses is through ETL. However, the procedure is frequently referred to as LTE, or extract, load, and transform, when working with data lakes.

This is because the data is directly fed into the data system after being extracted from the source system. The lake is transformed at query time. The ETL process is normally done during a period of low activity that will not affect business users of the business information commonly held in a data warehouse. Data warehouse ETL processes are typically batch-focused, possibly once a day or once a week. Given that a data warehouse is typically used for operational or financial analyses, this makes sense.

Technically speaking, real-time OLAP is far more difficult to construct than batch based typical OLAP systems. Real-time analysis can be done in many ways.

Apache Spark is a popular framework for implementing streaming analytics. Real-time streaming analytics uses Spark streaming to absorb small batches of data and transform them into a searchable data store. Most event-driven data stores, like Apache Kafka, have built-in handlers in tools like Sparke.

ETL procedures can be designed and carried out using a wide range of ETL tools, some of which are incorporated into database systems. An example of this is SQL Server Integration Services, which executes a workflow comprising data sources, data targets, and data flow activities. It may be used to connect to a wide variety of databases and data sources despite being strongly integrated with SQL Server. An open source ETL tool called Taloned supports many different types of databases. Open Studio for Windows or Mac can be downloaded for free. It also features a graphic designer and allows connectivity with SaaS providers, packaged software apps, and data sources like Dropbox.

Although Apache NiFi is not explicitly an ETL tool, given its versatility, it generally automates data transfers across systems. It could be used for both database and application integration. The vast array of data sources that NiFi provides processors for includes on-premises databases, big data platforms, and cloud services.

The two most well-known cloud providers, Amazon Web Services and Microsoft Azure, both include ETL tools. Its name is AWS Glue from Amazon. The fact that Glue is a cloud-native title tool means that it offers a visual designer that can be used in a Web browser. Python or Scala are two programming languages that can be used to create transformations.

Data formatting capabilities are one of AWS Glue's distinctive advantages. One of the most widely used methods for storing data such as files is the AWB cloud storage service, known as S03. These files can be crawled by AWS Glue and it can create a data catalogue that lists the data that is accessible in the data lake. AWS Glue makes it simple to transfer this data into different data warehousing services on our platform, such as Amazon Redshift or Amazon. Similar cloud ETL software on the Microsoft Azure cloud is called Azure Data Factory (ADF), which similarly offers a web based visual ETL builder. Building EDF ETL processes doesn't require programming, in contrast to AWS Glue. More than 90 data connectors are available through ADF, including sources from all the main cloud service providers, including Amazon and Google. ADF's ability to host SQL Server Integration Services packages makes it possible to run tasks created using SQL Server's standard ETL tool on Azure cloud infrastructure.

Data propagation, which is the process of moving data from one place to another, is frequently used to transfer a database's entirety or a specific subset from one location to another. Users at the target site can now access the data more quickly, and the source and target sites may benefit from redundancy as a result. Data propagation and data warehousing are sometimes used in tandem. Even though an organization may have an enterprise data warehouse where a large global data set is stored, this data is frequently propagated to regional data marts where a smaller portion is made available to local business units.

Better response times for regional users and a more pertinent data set that makes business intelligence jobs easier are two benefits of employing data marts. Edge computing is another possibility for data dissemination. Although moving data and computation to the cloud has been the general trend, businesses are discovering an increasing number of use cases that demand computing at local sites like retail storefronts or warehouses. Edge computing can improve the performance and dependability of services at these outlying locations. An application that uses facial recognition to identify workers entering a warehouse is a typical example of edge computing.

Data replication is a frequently used tool for carrying out data dissemination. Most database engines, including PostgreSQL, SQL Server, and Oracle, all have replication features built in.

This makes setting up replication and transferring data from a source database to a target database quite simple. Replication's functional implementations differ greatly among these databases. The replication solution may, in some situations, be centered on replicating a database to a backup location for disaster recovery. In other situations, the technology aims to transfer a portion of a master database to a read-only copy to facilitate reporting and analytics.

3. Results and discussion

This paper provides a thorough review of the many facets of data integration, including business integration, application integration, database integration, and the crucial ETL (Extract, Transform, Load) concept. Data integration is a crucial activity in the contemporary digital landscape. Based on the data in the paper, the following main findings and discussion points are listed:

- Importance of data integration:

In today's data-driven economy, the article emphasizes the importance of data integration. Organizations struggle with issues like data silos, redundant work, and ineffective decision-making without data integration.

- Types of data integration:

The paper discusses three distinct types of data integration: business integration, application integration, and database integration. These types are well-defined and essential in a variety of business contexts.

- Business integration:

B2B integration makes it easier for trading partners to communicate and exchange information about transactions. It emphasizes how crucial standards like XML and APIs are for facilitating these interactions.

- Application integration:

By connecting several software programs, application integration is said to build effective processes. The given example shows how this integration can increase productivity and streamline procedures.

- Database Integration:

Data from several sources are combined into one perspective through database integration, increasing its value and usability. The article addresses several database integration techniques, including data consolidation, propagation, virtualization, and federation.

- ETL process:

It involves removing data from source systems, converting it into a format that can be used, and putting it into a target system. Important steps in this process include data cleaning, enrichment, and bulk loading.

- Challenges in data integration:

In data integration procedures, issues with managing master data, foreign key relationships, and the dynamic nature of data are frequent. To ensure data consistency and correctness, these issues must be resolved.

- Technologies and tools:

Many ETL tools and cloud-based options, such as AWS Glue and Azure Data Factory tools, make data integration processes simpler and provide features like data formatting and broad data connectors.

- Real-time data integration:

It is known that real-time data integration is a trickier but more crucial component of data integration. In the study, real-time analytics solutions like Apache Spark are alluded to.

- Data dissemination:

Data replication and propagation are two ways for disseminating data that help move data to several sites for greater accessibility and redundancy.

Data integration is a vital procedure that enables businesses to fully utilize their data. The concept and its many elements are thoroughly discussed in this study, which also provides insightful information on the difficulties, methods, and tools involved in data integration. The capacity to convert and analyze data effectively can result in major competitive advantages in today's data-driven corporate environment, underscoring the crucial role data integration plays in this context.

4. Conclusion

This article concludes by delving deeply into the area of data integration and highlighting the crucial role it plays in our data-driven society. For organizations to succeed in the digital age, data integration acts as the keystone that turns raw data into an asset. Business integration, application integration, and database integration are the three main types of data integration that are thoroughly examined in the article. This information gives readers a thorough grasp of how these three types of data integration interact to maximize the value of data. The paper emphasizes how crucial data integration is in the modern economy, where data drives innovation, efficiency, and competitive advantage. Without efficient data integration, businesses must contend with fragmented data silos, duplication of effort, and inadequate decision-making procedures. Ineffective data integration can lead to big monetary losses and lost opportunities. In-depth analyses of each sort of data integration are provided, highlighting the various uses and advantages of each. B2B communication, which is the emphasis of business integration, automates procedures and lowers mistake rates by streamlining transactions between businesses. Application integration links many software programs to ensure efficient operation and smooth operations. Database integration gathers data from several sources, making it easily accessible for reporting and analysis. A crucial step in data integration is the ETL (Extract, Transform, Load) process, which is covered in detail in the article. Data warehousing and analytics require ETL because it ensures data quality and gets it ready for analysis. The debate over data lakes and warehouses also emphasizes how the field of data integration is constantly changing. It emphasizes the move toward real-time analytics and the effective use of tools like Hadoop and Spark for handling large information. The essay goes on to detail several ETL tools and cloud-based solutions, giving readers insights into how data integration methods are really put into practice. It emphasizes how crucial it is to pick the appropriate tools to make the integration process simpler and more efficient. Further demonstrating the adaptability of data integration techniques is the discussion on data replication, propagation, and edge computing. These methods enable businesses to move data where it is needed, speed up responses, and increase service dependability, especially in distant contexts.

This article essentially emphasizes the critical role that data integration plays in contemporary corporate processes. For businesses looking to make the most of their data, take wise decisions, and gain a competitive edge in a world that is becoming more and more data-centric, it is a priceless resource. By navigating the complexities of data integration and selecting the appropriate tactics and resources, one can succeed in the digital age and keep data from being an unmanaged resource but rather an asset.

Acknowledgement

I would like to express our sincere gratitude to my organization EY (Ernst & Young) for unwavering guidance, invaluable insights, and constant encouragement throughout this journey. Their valuable input and feedback significantly improved the quality of this research. I would like to acknowledge the contributions of my research colleagues and friends who provided valuable feedback, engaging discussions, and constructive criticism. Their diverse perspectives enriched this work significantly.

References

[1] "Data Integration Blueprint and Modeling: Techniques for a Scalable and Sustainable Architecture" by Anthony David Giordano.
[2] "Data Integration in the Life Sciences: 5th International Workshop, DILS 2008, Evry, France, June 25-27, 2008, Proceedings" edited by Alfonso Valencia and Paolo Romano.
[3] Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). The Data Warehouse Lifecycle Toolkit. Wiley.
[4] Inmon, W. H., & Linstedt, D. (2015). Data Architecture: A Primer for the Data Scientist. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-802044-9.00001-5.
[5] Inmon, W. H., & Hackathorn, R. D. (2007). Using the Data Warehouse. Wiley.
[6] Vassiliadis, P., Simitsis, A., & Georgantas, N. (2002). Conceptual modeling for ETL processes. International Journal of Data Warehousing and Mining, 8(4), 1-23. https://doi.org/10.1145/583890.583893.
[7] Duan, S., & Cercone, N. (2008). Data integration: A theoretical perspective. Journal of Computer Science and Technology, 23(4), 615-626.
[8] Piplai, T., & Piplai, S. (2014). Data integration: A theoretical perspective. International Journal of Advanced Research in Computer Science and Software Engineering, 4(12).