INST733-IM01: Database Design-Spring 2023

Written Report

Team 9: Data Integration and Federation

1.0 Introduction

Data integration refers to the process of combining data from different sources and presenting it as a unified view. This can involve extracting data from different systems, transforming it into a consistent format, and loading it into a centralized repository such as a data warehouse or data lake. The goal is to provide a single, comprehensive view of data that can be used for reporting, analysis, and decision-making.

Data federation, on the other hand, is a technique for accessing data from multiple sources in real-time without the need for a centralized repository. With data federation, data is accessed "on-the-fly" from its original location and presented to the user as if it were a single, integrated dataset. This approach can be useful when dealing with large or distributed datasets that are difficult or expensive to centralize.

2.0 Importance

As individuals studying database technologies and design, understanding data integration and data federation is important for several reasons:

- Holistic view of data: By learning about data integration and data federation, one gains a deeper understanding of how to combine data from different sources to create a comprehensive view. This is a critical skill for anyone working with data because it enables one to see patterns and trends that may not be visible when looking at individual data sources in isolation.
- Improved decision-making: With a unified view of data, one can make better-informed decisions based on accurate and complete information. By understanding data integration and data federation, one can design and implement database solutions that support data-driven decision-making.
- Competitive advantage: As businesses increasingly rely on data to drive their operations and decision-making, those who can effectively manage and analyze data have a distinct competitive advantage. By developing expertise in data integration and data federation, one is well-positioned to provide valuable insights to organizations about database technologies and design.
- Cost savings: Data federation can help organizations save costs associated with maintaining and managing large, centralized databases. By accessing data in real-time from its original location, organizations can avoid the costs associated with building and maintaining a centralized data repository.
- Reduced complexity: By integrating data from multiple sources, one can reduce the complexity of data management processes. Rather than having to maintain and manage multiple databases or systems, one can centralize data into a single repository. This

simplifies the process of data retrieval and management, making it easier to ensure data quality and consistency.

3.0 History

Combining heterogeneous data sources, also known as data silos, has always been a difficult issue to resolve. In the early 1980s, computer scientists began designing systems to allow these databases to operate under a single query interface. The first data integration system was created by the University of Minnesota in 1991, using a data warehousing approach. This method extracts, transforms, and loads data into a single repository. This makes querying quick, but it is less efficient for frequently updated datasets due to the repetition of the ETL process. Another issue with this approach is that data warehouses are difficult to construct when there is no access to the full data, just the query interface.

Data integration experienced a shift in 2009, when using a unified query interface over a mediated schema was favored over coupling the data to retrieve information directly from original databases. This approach maps the mediated schema to the original sources and transforms queries to match the schema of the original sources.

Since 2011, data hub approaches have gained more traction and become more popular Enterprise Data Warehouses. While EDWs are typically fully structured and relational, data hubs serve as points of mediation and data sharing rather than focusing solely on data analysis.

In 2016, the U.S. Data Federation project was created to support coordination and collaboration across organizational boundaries in the government. More specifically, it encouraged federated data efforts across complex organizational boundaries such as federal, state, and local government entities. Data collected and shared in these exchanges can be used to support policy or budget decisions, increase operational efficiencies, or be aggregated for other users to derive meaning.

Since 2013, data lake approaches have increased in popularity over data hubs. These methods do not require a relational schema to structure and define the data, but still combine the unstructured data into one location. This creates a vast “lake” of data in different formats that can still be used for analytics.

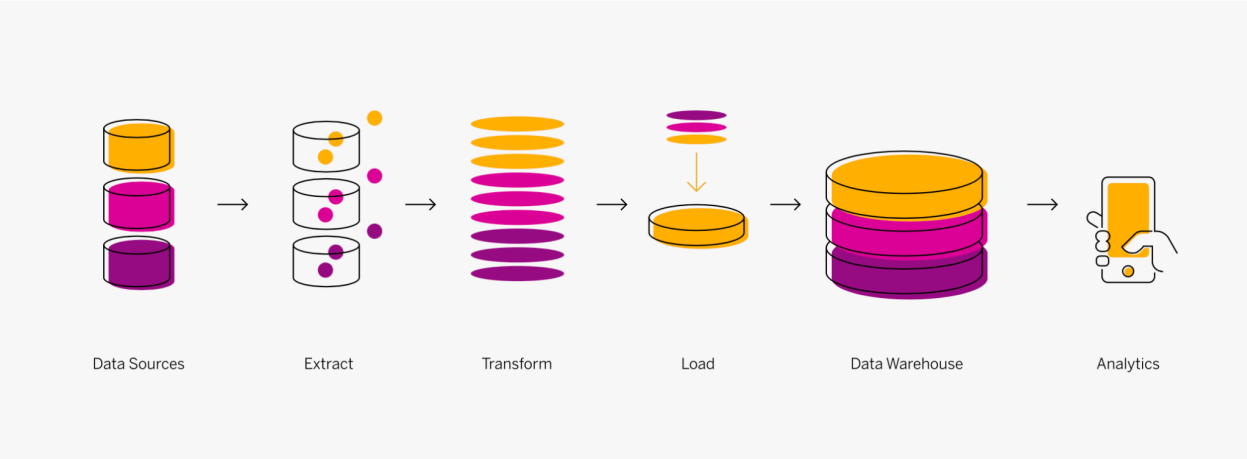

Figure 1. Data Integration Process: SAP Intelligence Cloud

4.0 Data Integration Steps

There are many different approaches to data integration, but there are a few common steps.

1. The first step is to identify the data sources that need to be integrated. This may include databases, applications, files, and other sources of data.
2. The next step is to extract the data using tools such as ETL (extract, transform, load) software. These tools can extract data from various sources.
3. Once the data has been extracted, it must be transformed through cleaning and standardizing the data, as well as combining data from multiple sources.
4. The final step involves loading the transformed data into a central repository, allowing users to access and analyze the data.

4.1 Data Federation Steps

Data federation overlaps greatly with data integration, but it is more specialized to organizations.

1. Similar to data integration, the first step is to identify the data sources. This may include databases, applications, files, and other sources of data.
2. The next step is to create a virtual layer that sits on top of the data sources, often done using data virtualization software or other techniques. This allows for a unified view of the data.
3. With metadata to define the virtual layer and describe the sources and relationships between them, users can understand and analyze the data in a meaningful way.
4. Users can then query the data using SQL or other query languages through the virtual layer. The virtual layer accesses the data from the underlying sources and provides a unified view of the data to the user.
5. The final step is optimization. Improving the performance of the data federation solution can include making sure the data sources are properly maintained, optimizing queries, and tuning the virtual layer.

4.2 Data Integration Methods and Strategies

Although different sources identify different types of methods and strategies, the major data integration methods include:

- Manual data integration
- Middleware data integration
- Application-based integration
- Uniform access integration
- Common storage integration

Ashraf (2020) in his article determines different data integration approaches, and identifies ETL as one of the approaches to be used. Some of the authors also mention other methods such as point-to-point, data virtualization and data propagation.

The major practice for data integration is to define clear data integration goals and objectives, which might involve all stakeholders in the planning process. It is best to choose the appropriate data integration strategy based on the organization's needs and resources.

For example, it's best to use the manual data integration, when there is a small amount of data sources and data needs to be merged for a basic analysis. In order to automate and translate communication between legacy and modernized systems the middleware data integration is the best option. However, according to the level of integration needed and complexity of the analysis the data integration technique to be used will be complex as well. To present the data uniformly, create and store a copy, and perform the most sophisticated data analysis tasks the common storage integration is appropriate.

5.0 Data Integration Challenges

Data integration challenges for the major part include the data quality, sources and its scalability. The integration process can be challenging as the data from different sources can come in various formats and structures. Ensuring the data quality is very important, since the data can be incomplete, inconsistent or outdated. As organizations grow and their data sources increase, maintaining an efficient and scalable data integration process is difficult and real-time processing will be long as well. The implementation and maintenance of data integration solutions can be expensive, both in terms of financial investment and human resources.

5.1 Alternative Technologies and Approaches

Data virtualization is the technology to abstract data from various sources, while creating a unified data layer for users to access. Although this process simplifies the data integration, it may not be suitable for handling high-volume and high-velocity data. Another approach is a data lake, which stores the raw data in a native format until needed, reducing the need for upfront data integration.

The major data integration tools and platforms include commercial solutions and open-source alternatives. The top ten ETL Framework alternatives include Microsoft SQL Server Integration Services, Sabermetrics, Fivetran, Workato, Adverity and others. However, as

mentioned above the best practice is to identify the needs and resources of the organization first. For example, Cleo Integration Cloud is an ecosystem integration platform that makes it easy to build, automate and manage B2B, application, cloud, and data integrations, while Fivetran is an ETL tool, designed to reinvent the simplicity by which data gets into data warehouses. Adverity is the fully integrated data platform for automating the connectivity, transformation, governance, and utilization of data at scale.

6.0 Conclusion

To conclude, data integration is the process of combining data from different sources and making it available for unified analysis and decision-making. This involves extracting, transforming, and loading data (ETL) from various systems into a common format, which enables businesses to gain insights from a consolidated view of their data. Data federation, on the other hand, is a method that allows querying and accessing data from multiple sources without physically moving or transforming it. Both approaches have their merits and are essential in today's data-driven business landscape. To learn more about data integration and data federation, individuals can explore books, online courses, blogs, and tools to deepen their understanding and develop the skills needed to implement these strategies effectively. Books such as “Data Integration Life Cycle Management with SSIS" by Andy Leonard and “Principles of Data Integration” by AnHai Doan, Alon Halevy and Zachary Ives are comprehensive textbooks of data integration, covering theoretical principles and implementation issues as well as current challenges raised by the semantic web and cloud computing. Apart from this, an online Data Integration course from the websites as DataCamp and Coursera will be very helpful to get the general idea about this topic.

7.0 References

(1) Bhattacharjee, S. (2023, April 14). The Ultimate Guide to Data Integration. DiGGrowth. Retrieved April 27, 2023, from https://diggrowth.com/blog/data-integration/

(2) Data aggregation: Definition, process, tools, and examples. KnowledgeHut. (n.d.). Retrieved April 29, 2023, from https://www.knowledgehut.com/blog/data-science/data-aggregation

(3) Data Integration. Data Integration - CIO Index. (n.d.). Retrieved April 27, 2023, from https://cio-wiki.org/wiki/Data_Integration

(4) Federal Enterprise Data Resources . (n.d.). Data Federation. Retrieved April 29, 2023, from https://resources.data.gov/data-federation/

(5) Ashraf, P. byS., Nadeem, P. N., & Ashraf, P. S. (2020, July 30). Data Integration Approaches – which one is right for business? Data Integration Blog. Retrieved May 3, 2023, from https://dataintegrationinfo.com/data-integration-approaches/

(6) 5 data integration methods and strategies. Talend. (n.d.). Retrieved May 3, 2023, from https://www.talend.com/resources/data-integration-methods/

(7) ETL framework alternatives for enterprise businesses in 2023 | G2. (n.d.). Retrieved May 3, 2023, from https://www.g2.com/products/etlframework/competitors/alternatives/enterprise