Data Virtualization and ETL

Revision 20200604

NOTE
This document is confidential and proprietary of **Denodo Technologies**.
No part of this document may be reproduced in any form by any means without prior written authorization of **Denodo Technologies**.

Copyright © 2021
Denodo Technologies Proprietary and Confidential

Denodo North America & APAC: 525 University Avenue, Suite 31, Palo Alto, CA 94301. USA
Denodo Iberia & Latino América: Montalbán 5, 28014 Madrid, Spain
Denodo EMEA: 21st Floor, Portland House, Bressenden Place, London SW1E 5RS. UK
Denodo DACH: Karlstraße 10, 80333 München. Germany
www.denodo.com

CONTENTS

1 GOAL

To help understand the difference between Data Virtualization and ETL technologies and answer some frequently asked questions on the same, i.e.:

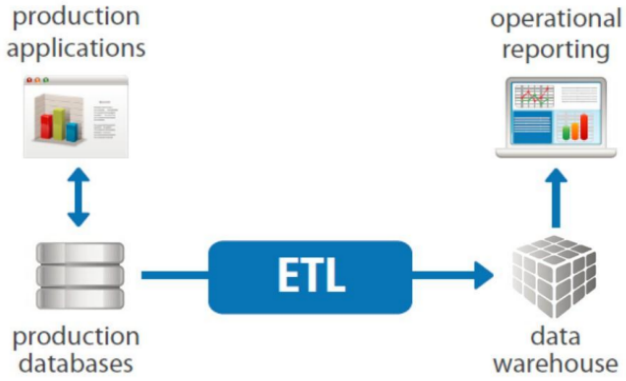
- When should I use Data Virtualization and when should I use ETL tools?
- “Does Data Virtualization replace ETL?
- I’ve already got an ETL in place, why do I need data virtualization?

2 CONTENT

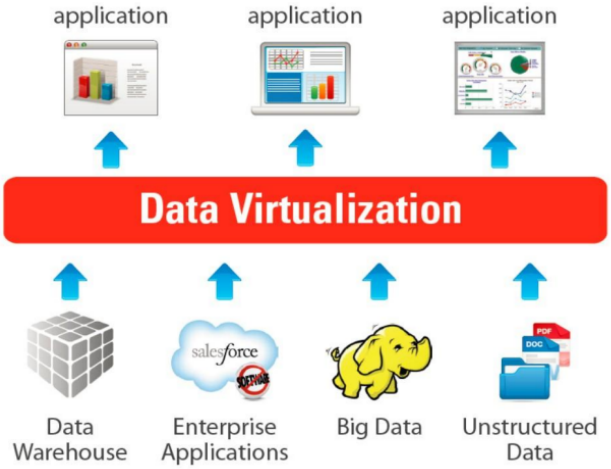
Data virtualization and ETL offer different capabilities to access, integrate, and deliver data. Although they can offer competing approaches for some specific scenarios, they are both useful additions to the data integration toolset of any organization. In general, Data Virtualization is more agile, flexible, versatile, and cost-efficient than ETL.

A simple takeaway is not to use ETL when data virtualization is a viable approach. Deciding whether data virtualization is a "viable approach" depends on the circumstances of the particular project and its unique constraints.

Extract, Transform, and Load (ETL) is a good solution for physical data consolidation projects which result in duplicating data from the original data sources into an enterprise data warehouse (EDW) or a new database. However, applications using the data from the resultant data store are working with data that is as old as the last ETL operation. In this case, the best scenario could be the data from yesterday's end-of-day ETL operation. This may be acceptable for applications performing data mining or historical analysis supporting strategic planning or long term performance management but is less acceptable for operational decision support applications, managing inventory levels or applications that need to use intraday updates of the data.



Data virtualization, on the other hand, abstracts, federates and publishes a wide variety of data sources to consuming applications in an array of different formats. The Denodo data virtualization platform consumes virtually any type of data, including SQL, MDX, XML, Web Services (REST and SOAP/XML), flat files, and unstructured data in Hadoop and NoSQL databases, and publishes the data as relational tables or Web Services. When users submit a query, the data virtualization platform calculates the optimal way to fetch and join the data from remote heterogeneous systems. It then queries the relevant data, performs the necessary joins and transformations, and delivers the results to users – all on the fly without the users knowing about the true location of the data or the mechanisms required to access and merge it.



Key differences:

- Data virtualization leaves the source data where it is and delegates the queries down to the source systems while ETL copies the data from the source system and stores it in a duplicate data store. In the Data Virtualization platform, when applications request data by querying a data service, the underlying data sources are queried in real-time by the platform and the results are aggregated before being returned to the application.
- Data virtualization is more agile in dealing with modifications in the underlying logical data model, including adding new data sources. It also supports rapid development iterations with the value being added to the solution in each iteration (e.g. every week or two weeks). A typical ETL/EDW project usually takes many months of upfront planning and data modeling before any data can be consolidated in a data warehouse and, after deployment, it is very hard to change.

Although the two solutions are different, data virtualization and ETL are often complementary technologies. Data virtualization can extend and enhance ETL/EDW deployments in many ways. How data virtualization can enhance existing ETL/EDW deployments is detailed later in this document.

2.1 WHAT IS DATA VIRTUALIZATION GOOD FOR?

Data virtualization is an excellent solution, and many times the only option, in scenarios such as:

- Structured, semi-structured, and unstructured data from disparate data sources need to be combined and queried. The consuming applications are isolated from the intricacies of accessing and formatting this disparate data as the data virtualization platform presents it via standard interfaces, such as SQL, Web Services (REST and SOAP/XML), and so on.
- Data needs to be accessed and delivered in real-time. This is very important for decision support applications, such as those managing inventory levels or providing intraday portfolio risk analysis.
- Highly regulated environments where data duplication can be perceived as a security and privacy risk like in GDPR.
- The requirements to expose business data entities demand multiple diverse consuming formats that decouples applications from data sources.

The Denodo Platform can also support data transformations and data cleansing:

- Data transformations can be complex and can also include hierarchical data structures, as in XML/JSON documents. Integration with external transformation tools is also supported via an API.
- The Denodo Platform can also perform data cleansing and quality operations, such as data enrichment, value normalization via mapping tables, data redaction in support of privacy requirements, and so on. As the Denodo Platform is extensible, custom data quality routines can be created and added to the platform or the platform APIs can be used to invoke external data cleansing and data quality tools.

2.2 WHAT APPLICATIONS CAN BENEFIT FROM DATA VIRTUALIZATION?

A wide variety of applications can benefit from the agility and flexibility provided by data virtualization. Obvious applications include those requiring real-time (or near real-time) access to the most up-to-date data available. Examples of these applications are operational decision support systems, such as inventory control, risk management, and so on.

Applications that are subject to changing requirements and the addition of new data sources – including unstructured data sources that are typically not handled by traditional data tools – are also good candidates for data virtualization. The flexibility and agility provided by the data virtualization platform to easily connect to new data sources of all types and to combine these new sources into the existing data views enable a rapid iteration process which allows the development team to react quickly to new data demands from the business.

BI and analytics applications that are traditionally the preserve of data warehouse deployments can also use data virtualization to extend the types of data being analyzed to include unstructured data sources typically not supported by data warehouses. Examples of this include pulling social media data from the web to analyze influencer behavior for consumer buying patterns. This combines the normal transaction data (purchases) with social media influencer or influenced data pulled from, for example, Twitter streams or Facebook posts.

Finally, new web and mobile applications that need to access corporate data sources are ideal candidates for data virtualization. Typically these applications need to be isolated from the underlying data schemas and traditional access methods such as SQL queries. Mobile applications especially are more likely to use REST Web Services to access any data from the corporate data stores and the ability of the data virtualization platform to expose the underlying data as REST Web Services makes them better suited for newer web and mobile applications.

Typical projects where data virtualization is a ‘must’ include:

- Building a logical data warehouse

It can be created in two ways: augmenting and enhancing an existing data warehouse deployment or it can be a ‘virtual’ data warehouse created from existing data sources (reducing the time and expenses of creating a physical data warehouse). In either scenario data virtualization is a critical piece of the solution for connecting, combining, and delivering the data in a consumable format for the dependent applications.

- Big Data initiatives

Data virtualization allows you to integrate the data with the information pulled from your CRM, from your DSR, from your data warehouse, and so on. Not only that, it can also pull in and combine the Big Data with unstructured data sources, such as social media (Twitter, Facebook), weblogs, etc. Without this ability to integrate the data coming from Big Data systems, you end up with data silos – one for the business and one for the data scientists – and you won't be able to realize the full potential of Big Data.

- Practical MDM

Data virtualization provides flexibility and time-to-value for any MDM project, whether you are using an MDM tool or not. Data virtualization platforms can quickly adapt to new 'master' sources of data. For example, if you find that your CRM data for customer contacts is getting stale, you can use the data virtualization platform to access new data sources (e.g. social media) to refresh the virtual master data from this new source. (Of course, you can feed this data back to your other data sources if you need to update their data). Alternatively, if you already have an MDM solution in place, you can extend and enrich the data from the MDM solution by using the data virtualization layer to access other data sources, such as unstructured data from social media and the web.

- Enterprise Information Governance initiatives

Information governance is increasingly important to organizations, especially as regulations control the access and usage of private and confidential data – with significant penalties for organizations and people who breach these regulations. A data virtualization platform acts as an abstraction layer between the data sources and the consumers. Data consumers connect to the data virtualization platform rather than the data sources directly and this provides a point of control for monitoring and enforcing data access policies. These policies allow you to control who accesses certain data, how and when they access it (e.g. office hours and from office locations), and how they can use the data. The data virtualization platform also allows you to configure different views of the same data for different users or user roles. This can be by simply not allowing access to certain data elements or by redacting parts of the data (e.g. all but the last four digits on a social security number).

- General Data Protection Regulation

Many of the Global companies struggle with complying with GDPR law as the data is present in many different countries across the world. Having a Data Virtualization technology, helps in connection to and utilizing only the data required, without having to make copies of data in multiple locations. It also creates a Unified access Layer(Single point of entry) for giving controlled access so that the organization can easily comply with regional data security laws and perform complete audits of data access when required.

With the data virtualization layer providing a control point for the underlying data sources, it is easier for organizations to implement information governance programs to bring them into compliance with industry regulations.
Using data virtualization can also provide additional benefits such as data lineage reporting, so that it is easy to determine the originating source of information in case of errors or other problems at the consuming end. The ability to trace the information back to its data source and see how it was modified or manipulated between the source and consumer is invaluable.
In the same way, it can also provide a reverse view of this lineage, showing where data is consumed i.e. going from the source to the consumers. This is important for impact

analysis when planning a change to a data source. It is easy to see what consuming applications are affected by changes to a data source and, therefore, to plan for these impacts.

2.3 WHAT IS ETL GOOD FOR?

ETL is an excellent tool for the bulk copying of complete data sets and transforming them into a predefined enterprise data model (EDM). It is designed and optimized to handle very large datasets with millions (or even billions) of rows. Typically ETL handles very structured data sets, such as relational databases, or XML. However, ETL does not handle semi-structured or unstructured data very well, if at all.
If the data needs to be transformed prior to loading in the target data warehouse, ETL processes support complex data transformations. These can be complex multi-pass transformations that even require a rules engine or workflow process to orchestrate the transformation. Data can also be staged by the ETL process to support the multiple passes over the data. Similarly, ETL supports complex cleansing operations, including value matching, de-duplication, conflict resolution with human intervention, etc. These transformations and data cleansing capabilities usually result in better quality data in the target data warehouse.

2.4 WHAT APPLICATIONS CAN BENEFIT FROM ETL?

ETL (and the resulting data warehouse) is best suited for applications that require access to the complete consolidated data set. For example, historical trend analysis, data mining operations, etc. These applications need to process (not just access) complete data sets to perform their analysis. Applications that can use 'old' data – for example, yesterday's 'end of day' data set – are also candidates for ETL/EDW. Examples are end of day sales reconciliation processes in retail, or close of business updates on portfolio positions in finance. These applications can use the 'last available' data provided by ETL operations. Finally, applications that are 'read-only' are candidates for ETL/EDW. ETL is typically in a single direction i.e. from the original source to the data warehouse. This provides read-only access to the consolidated data in the data warehouse, but any updates to the data (if even permitted) are typically not transferred back to the original source. Data updates need to be made at the source so that subsequent ETL processes (or any Change Data Capture processes) will pick up the changes and replicate them to the consolidated data warehouse.

2.5 WHAT TECHNOLOGY FITS MY NEEDS BETTER?

This table summarizes a comparison of the differences between typical ETL/EDW projects and data virtualization projects. It can be used to determine the technology which fits better to a project's needs.

Category	ETL/EDW	Data Virtualization
Time to value	Long term project, typically 12 months or longer.	Need a solution in production in days to weeks, and then rapidly iterate to increase benefits to the organization. Data models are dynamic and executable.

Project cost	$$$$$	$$
Data models stability	Requirements are very well defined and not predicted to change significantly after deployment or not to change at all.	Requirements are well understood but expected to evolve after deployment in ways that will require modifications, even adding new data sources.
Replication constraints	There are no policies or regulations governing replication. The cost of creating and managing data source replicas is not a limitation.	Privacy regulations(GDPR), internal policies, or security concerns don't allow physical replication and consolidation of data. Replication minimization is a business goal.
Source data availability	Source systems are often offline, and/or the network connection between them is not reliable, making direct access to the data sources difficult.	Source systems are generally available, and the network is reliable. Direct access to data sources is usually available. Note that Caching by the data virtualization platform can alleviate source unavailability.
Source system load	Source Systems do not have the capacity for additional load during 'business hours' resulting in the need for 'out of hours' processing	Source systems can handle additional controlled access. Caching (and sophisticated cache management policies) can reduce any additional load on the source systems.
Data cleansing	Complex multi-pass data cleansing (e.g., matching, de-duplicating, conflict resolution with human intervention).	Usually single pass data cleansing operations (e.g., enrichment, normalization, redaction, etc.). APIs support integration to external data cleansing and data quality tools.
Data transformation	Complex and multi-step, often requiring workflow engines and/or BPM pipelines.	Complex transformations, managing hierarchical data structures (e.g. XML/JSON). API for integration to external transformation systems.
Application uses	Heavy analytical BI – Historical analysis and/or data mining to facilitate strategic planning and long-term performance management.	Mixed informational and operational uses. Tactical decision-making visibility into operational data (e.g., current risk, inventory levels, device status). Also operational applications with moderate transactional requirements. Can also be used to Prepare Data for Machine Learning projects.
Data formats	Limited to structured data.	Wide range of data source connectivity: Structured (RDBMS,

		XLS), Semi-Structured (File systems, noSQL documents, web services.), and Unstructured (email, documents, etc.)
Data freshness	End of day and/or End of last load	Near real-time (“right time data”)
Data volume	Each query needs to read and process a very large amount of data (millions of rows)	Queries can range from simple SELECT statements on single views all the way to complex queries returning result sets containing millions of rows.
Data consumption	JDBC, ODBC, MDX	JDBC, ODBC, Web Services (SOAP/XML,REST, ODATA and GraphQL) , Portlets, SharePoint Web Parts, HTML, etc.

2.6 EXTENDING ETL/EDW WITH DATA VIRTUALIZATION

Frequently an ETL/EDW system already exists and there is a lot of pressure on new projects to make use of this existing infrastructure. The organization might have spent a lot of money and time to set up the ETL/EDW infrastructure and projects are pushed to make use of these tools to justify the investment. These results in projects that don’t fit into the classic ETL/EDW use case being force fitted into the infrastructure, which usually results in overly complex designs that fail to meet the user’s expectations. In these circumstances, a better solution would be to extend and enhance the existing system capabilities with data virtualization.
The following are examples of how data virtualization can coexist and enhance existing ETL/EDW systems. This is not intended to be an exhaustive list of the ways in which ETL/EDW can be enhanced with data virtualization – it simply provides some examples of these types of scenarios.

- Extending existing data warehouses with new data sources

In situations where a data warehouse already exists within the organization, but the business users need to add new data sources to enhance their reporting or analytics. In this situation, a data virtualization platform is layered over the data warehouse and also connects to the new data source(s). The reporting tools use the data virtualization platform as the data source for their reporting needs, effectively including the existing data (in the data warehouse) with the data from the new data sources. The same line of thought can be applied in creating a Virtual Data Lake project or to extend multiple Data Lake projects using Data Virtualization Platform.

- Federating multiple data warehouses

In a situation where there are multiple existing data warehouses within an organization – for example, regional data warehouses – and the business needs a single view of the data, a data virtualization platform can be used to provide this single view.

- Acting as a virtual data source to augment ETL processes

In the same way that data virtualization can provide a virtual view onto multiple data sources, it can also become a virtual data source for an ETL process. In this situation, the data virtualization platform accessing and federates the underlying data sources – including those containing unstructured data that is typically not handled by an ETL

process – and delivering the data as derived views that can be accessed via SQL by the ETL tools.

- Isolating applications from changes to underlying data sources

One of the core benefits of data virtualization is that it isolates the consuming application from the complexities and location of the underlying data sources(On-Prem or Cloud) The applications should not care where the data comes from and how it is retrieved. In addition to making application development quicker and easier, the data virtualization platform also isolates the applications from changes to the underlying data sources.

2.7 CONCLUSION

Data virtualization plays a key role in modern enterprise integration stacks to cover some strategic needs. Factors such as exponential data growth, new data source types that create new information silos (NoSQL, Cloud), and the extensive use of Big Data require a new infrastructure that is not covered with traditional ETL solutions alone.

The answer to the original question of “when should I use data virtualization and when should I use ETL tools?” really is “it depends on your circumstances”.

As a summary:

- In many cases, especially those combining structured data with unstructured data or requiring real-time access to up-to-date data, Data Virtualization is a better option.
- In some cases, where it is really necessary to copy massive amounts of data for complex analytics or historical data marts with no concerns about data freshness, ETL is still the best option.
- Very often, the line is not that clear. Cost of data storage, operational costs of the solution, and time to market can tip the balance to data virtualization, even for projects that have traditionally used ETL solutions.

Having said this, data virtualization can often be used to increase the value of existing ETL/EDW deployments by extending and enhancing the ETL/EDW tools and processes. Some typical scenarios of a hybrid solution – combining data virtualization and ETL/EDW – were described earlier in this document.