data virtuality

# The Comprehensive Handbook of Data Integration

How to enable a flexible data supply chain
for faster time-to-insights in this digital era

Nick Golovin, PhD

The evolution of data integration nicely mirrors the different kinds of challenges that organizations are facing when turning data into insights. Starting from the question of how to move the data from a data source into a central repository, data integration is now trying to answer complex strategic topics that reflect many aspects of the entire data architecture such as data governance, the democratization of data, and the enablement of business users.

This e-Book examines the different data integration approaches that have evolved over the past 30+ years. You will learn about the various benefits and challenges that each of these approaches brings and how or if these approaches can help you to overcome your data management challenges incl. data governance, breaking data silos, and real-time connection to meet the current business objectives.

**Audience**:
Data Architects
Enterprise Architects
BI Managers
Data Warehourse Practitioners Project Managers

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

# The Comprehensive Handbook of Data Integration

## How to enable a flexible data supply chain for faster time-to-insights in this digital era.

### The Evolution of Data Integration

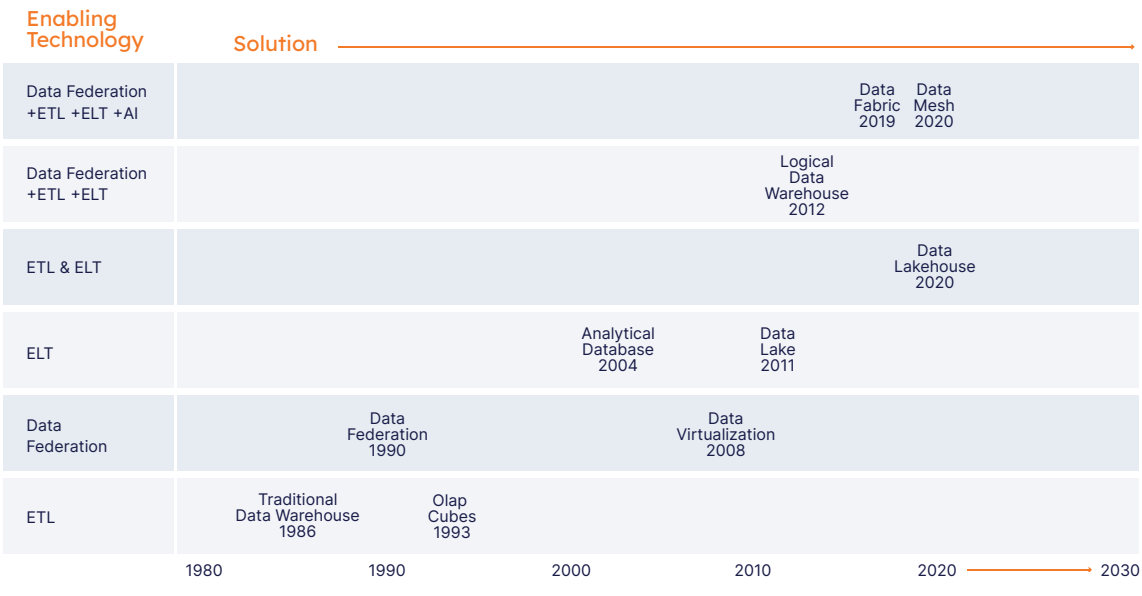
Even after several decades, the reason why many organizations still cannot transform the data into valuable insights, according to the two big research companies Gartner and Forrester, is because they fail to integrate the data.[1] The most common way to integrate data is still ETL. Considering the distributed nature of data in organizations and the fact that business users are increasingly involved, it isn't viable anymore. Further aspects that are underestimated:

• Data in organizations are and will stay distributed. It's the data architecture that needs to take this into account and support it.

• Ultimately, it's the business users who need to work with the data and turn them into insights. While some of the technologies empower business users to work independently from IT, some can create growing friction among the IT users and the business users, causing them to work with competing objectives in data governance and flexibility. Bridging this chasm by enabling a flexible "data supply chain" is essential for organizational efficiency and maturity.

• The way data is consumed by business users has changed as well. In the past, the requirements for data teams was to deliver data prepared in dashboards and reports. But for a few years now, business users are more interested in working directly with the data (sets) making web-based business data portals and SaaS deployments an essential part of the data architecture strategy. Data is becoming a first class citizen.

The pros and cons of ETL will follow in the next chapters. But what the recent experience and knowledge that we could gain teach us are that we need to consider several different types of data integration in our data architecture to keep up with the ever-changing demands.




* ELT: Extract Load Transform | ETL: Extract Transform Load

[1] Stephen Pritchard, "Data integration dogged by complexity and vital to business", Computerweekly, accessed April 7th, 2022. https://www.computerweekly.com/feature/Data-integration-dogged-by-complexity-and-vital-to-business.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

## Data Warehouse and ETL

In a traditional data warehouse environment, data is ingested, modeled, and stored through an Extract, Transform, and Load process (ETL). These ETL jobs are used to move large amounts of data in a batch-oriented manner and are most commonly scheduled to run daily. Running these jobs daily means that, at best, the warehoused data is a few hours old. But more typically, it is stale by a day or more. And because ETL jobs consume significant CPU, memory, disk space, and network bandwidth, it is difficult to justify running these jobs more than once per day. In a time when Application Programming Interfaces (APIs) and other myriad ways to access data were not as prevalent as they are now, ETL tools were the go-to solution for operational use cases. With APIs now in the picture and the sheer variety of data they represent, the ETL method is becoming less practical outside of non-time-critical large-scale data movement of APIs and big data, ETL tools posed significant challenges. Mainly because they require comprehensive knowledge of each operational database or application. Interconnectivity is complicated and requires thorough knowledge of each data source all the way down to the field level. The greater the number of interconnected systems that need to be included in the data warehouse, the more complicated the effort is. As such, ETL-based data warehousing projects became infamous for their appallingly high failure rates.

Further shortcomings are:

- Cost overruns and delayed implementations
- Perpertual incompleteness due to the lengthy onboarding cycle for new data sources, combined with ongoing changes to existing data pipelines
- Data teams playing catch up with the constantly changing business requirements
- Time consuming processes to conceptualize the database and thoroughly define requirements to avoid re-works

In order to overcome the shortcomings of data warehouses infrastrcutures, **data warehouse automation (DWA)** has been evolving for half a decade now. DWA solutions enable processes to accelerate and automate data warehouse development cycles while assuring quality and consistency of the data stored. Thereby, data storage and computing power can be saved leading to improved productivity and overall quality as well as cost reductions.

Gartner estimated that between 70 and 80 percent of corporate business intelligence projects failed to deliver the expected outcomes.

Due to the lack of technical possibilities, data warehouses were built in multi-purpose databases which were originally designed for operational use cases rather than analytical use cases. Complex database architecture work was needed in order to ensure the optimal performance for analytics. This situation improved with the emergence of new approaches, including multidimensional OLAP and analytical databases.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

**Online Analytical Processing (OLAP)** and cubes were other words for multi-dimensional sets of data and were a mechanism used to store and query data in an organized and multi-dimensional structure that is specifically optimized for analytic use cases. OLAP databases were designed to pre-calculate as many queries and combinations of data fields as possible to provide fast query responses. However, while these solutions performed better than classical relational databases, their multi-dimensional structure made them inflexible and unable to accommodate changes easily. In addition, storing large amounts of data in a cube caused performance bottlenecks.

Another way to organize data for multi-dimensional querying is **Relational Online Analytical Processing (ROLAP)** which is a form of OLAP that performs multi-dimensional analysis of data stored in a relational database. Although ROLAP technology performs better than OLAP databases when processing large amounts of data, it cannot beat the speed and efficiency of OLAP when it comes to smaller amounts of data. ROLAP databases require a great deal of manual maintenance and are difficult for business users to operate. So ROLAP is more inflexible than OLAP cubes. OLAP and ROLAP are largely out of popularity today, but can still be met in some legacy environments.

## Analytical Databases

Progressive software vendors sought to overcome the limitations of data warehouses and cubes by working towards products that were both flexible and able to process large analytical workloads. These analytical databases, or column stores, were the next step in the trend to provide business analysts the tools and flexibility they need. These analytical databases have evolved into Massively Parallel Processing (MPP) analytical databases that are more flexible and more performant than OLAP cubes. Even in cases where large amounts of data are being stored and queried.

However, these analytical databases require data to be copied into them using processes that are the same or very similar to ETL processes including similar drawbacks. The load processes for such databases are often bulk-oriented and require putting data into intermediate storage before loading them into the database.

With the analytical databases, the switch from ETL to ELT (Extract, Load, and Transform) was enabled. Unlike ETL, where data is transformed before it's loaded into the database, ELT significantly accelerates load time by ingesting data in its raw state. The rationale behind this approach is that the analytical databases (and other stores that allow ELT such as data lakes) are not picky about the structure of the data. Therefore, no development time is required to transform the data into the right structure before it is ingested. All further operations and transformations could occur within this database when it is needed.

MPP analytical databases were originally provided as on-premises solutions. But the costs of setting up and operating such databases on-premises were relatively high. With the rise of cloud-based analytical databases, the deployment has shifted to cloud solutions and is now the dominant type in the market.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

## Data Governance[1] and the Data Warehouse

Although a data warehouse can easily be governed in itself, overall, you can only govern what you can control. And there will always be at least some data not yet onboarded to the data warehouse. Therefore, the ability to govern data with a centralized data warehouse approach falls short. It is not a matter of whether some data is not governed or protected but how much data is exposed to the risk of breach, difficult to find, understand, etc. It is this ungoverned data, that when needed by the business, is problematic. When business reporting requirements necessitate sourcing data from outside the data warehouse, the business will be forced to do anything necessary to fulfill expectations for reporting. This may include unprotected storage of data (such as on local PC drives), an inappropriate joining of multiple data sets, or misrepresented metrics and KPIs due to lack of data quality, formal modeling, and so on.

## Self-Service BI Tools

Because the data warehouse approach falls short of expectations for speedy and comprehensive analytical data access for business users, a new approach surfaced. Self-Service Business Intelligence (SSBI) technologies, like Qlik and Tableau, introduced an approach to data analytics that enabled business users to access and work with corporate information without the IT department's involvement. These SSBI tools have the capability of 'blending' or locally integrating data from the data warehouse with any other data sources not stored in the data warehouse. This is accomplished by pulling copies of the data sources into the local data store of the SSBI tool where the analyst can 'blend' or integrate the data as needed.

These self-service tools are flexible and relatively easy to implement and they provide a good level of independence for data analysts. But there are clear disadvantages to these tools. The most prominent disadvantage is that data analyses performed in this manner do not scale, creating further silos. The outcome is redundant work, inconsistent results and, in short, chaotic reporting practices when used throughout the organization. Since everyone can define their own rules and calculations, it is both possible and likely for different groups and individuals to calculate the same KPIs and metrics in different ways leading to an array of conflicting results and to the publishing of both confusing and contradictory information.

Because these solutions have little to no permissions structures, there is no security layer to protect sensitive data which is a severe vulnerability since analysts frequently and casually exchange data files. Also, the ability to transform the data is relatively limited in most cases. Furthermore, because many machines are doing the same work for different users in parallel, powerful computer resources are being used inefficiently, contributing to higher costs and lower system performance. For these reasons, pure SSBI tools can fill a limited and short-term need, especially for prototyping, but fall short of being an end-to-end enterprise-level analytical solution.

[1] **Data governance**, briefly defined, is the formal organization and authority over the processes and methods used to manage and ensure the quality, integrity, discoverability, protection, and understanding of data. In order to achieve the objective of data governance, data must be under the control of a system responsible for governing access, usage, metric definitions, etc., which typically falls under the purview of a data warehouse, data mart, or some other managed data platform or process.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

### Applying Data Governance within a Self-Service Environment

The key to successful data governance within a self-service BI capability is to provide all the data required by your user community along with the freedom to explore and consume data to create reports and perform ad-hoc requests as needed.

By providing pre-defined and certified data tables, dimensional objects, data marts, and other data assets, data governance can be applied to a self-service environment. This ensures that the business has access to data they need which has been cleansed, modeled, and for which the metrics and KPIs are standardized and pre-created to avoid any misinterpretation of the data. Accomplishing this is challenging to say the least. Self-service BI is prone to the exact same problems as the traditional data warehouse model.

Due to the perpetually unfinished state of the data assets supporting self-service BI, at some point, your business audience will be left on their own to source data not yet onboarded and that leads us back to ungoverned data residing in an ungoverned location with ungoverned metrics being calculated in an ungoverned manner and dashboards getting shared and socialized without proper oversight.

## Data Lakes and ELT

Next came the data lake strategy. Data lakes are storage repositories that can hold a vast amount of raw data in its native format until it is needed. Traditionally, data lakes have been Hadoop-based systems and they represent the next stage in both power and flexibility. A compelling benefit of this approach is that there is no need to structure (transform) the data before querying (referred to as 'schema on write'). In fact, you can assign structure to the data at the time it is being queried (referred to as 'schema on read').

While it is a tantalizing approach, the data lake falls short of expectations for several reasons. A primary objective of the data lake is to simplify and accelerate. However, the approach often complicates matters with extra steps to prepare data for analytics, more complicated environments, and additional levels of staff and expertise to support. Although an analytical data lake provides significant reductions in labor for data loads, it still requires all data to be moved or copied to a single location prior to being accessed for analytical purposes. This drawback is also shared with the traditional data warehouse that uses an ETL approach since data load latency cannot be eliminated from the analytical data supply chain. But the load time latency is greatly reduced for the data lake as compared to a data warehouse. Another disadvantage of the data lake is a phenomenon that has come to be known as the 'data swamp' or 'data graveyard'. The data lake approach often leads to dumping and storing much more data as compared to ETL because of the lower cost of ingest and storage. The 'save everything' approach leads to loading and storing much more data than businesses are prepared to analyze. Since any data load takes time and consumes disk space and network bandwidth, unnecessary loads can be expensive and cause additional latency that delays other more valuable data from being analyzed in a timely manner. Another disadvantage to the 'save-everything' approach is the difficulty in cataloging data for ease of location and access.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

Apache Spark and Hadoop work together to offer impressive in-memory data processing for big data applications.

Although there has been hope that the in-memory capabilities of Spark would solve many of the latency issues related to Hadoop, the advent of cloud-based MPP (massive parallel processing) platforms providing quicker data movement capabilities and faster data processing, have all but rendered Hadoop irrelevant. Regardless, both Hadoop and Spark have limitations and fall short of a one-size fits all solution.

Although data lakes and ELT bring data together into one place quickly, they cannot provide fast query response as analytical databases do, nor can they provide access to data in real-time.

Recently, data lake and data warehouse/analytical database approaches have been combined into a hybrid concept called Data Lakehouse (see box for more information).

**Data Lakehouse**

For about two to three years, a new hybrid architecture called data lakehouse is gaining some traction. The idea of the data lakehouse is to bring the best of the worlds of data warehouse and the data lake by combining different elements of both. Data structure and management features of data warehouses allow for simplified schema and data governance. Data lakes are typically more cost-effective for data storage.[1]

In a data lakehouse, structured as well as unstructured data can be handled in a single system, allowing less complex and time-consuming administration. The efficient and secure working with the vast amount of data that is stored in the data lake makes this approach popular for Artificial Intelligence (AI) and Business Intelligence (BI).

As an early-stage concept with a limited number of real-world deployments, we still need some time to fully understand the downturns of the data lakehouse.

Some of the challenges that are already visible are:

- Limited integrations: The number of ready connectors is still limited and data load processes need to be developed manually.
- Monolithic structures: Data lakehouses are forming massive, monolithic structures. In the long run, these can become inflexible and difficult to work with.

[1] Thomas Hazel, "Data Lake vs Data Warehouse", Chaossearch, accessed April 7th, 2022. https://www.chaossearch.io/blog/data-lake-vs-data-warehouse

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

## Data Governance and Data Lakes

The very nature of the data lake paradigm is in stark opposition to the governed architecture of a data warehouse. Although this is part of the attraction due to quicker data ingestion, the data lake can be difficult or impossible to govern, especially regarding data protection interests such as GDPR, HIPAA, and CCPA. As such, data lakes should be considered a base layer and point of ingest, locked down to only the staff with both the skills and need for direct access to the masses of uncleansed, unmodeled, raw, and sensitive data. Typically, this staff may be limited to administrators, operations, and power users with the highest level of trust such as data scientists requiring access to the largest datasets available for the training and operating of machine learning models.

# From Data Federation to Data Virtualization

## Data Federation

While most data analysts were busy exploring the progression from relational databases to cubes, analytic databases, and data lakes, another camp was looking into using data federation to integrate data for analysis. Data federation allows to instantly run queries that join multiple disparate, mostly relational databases without the need to copy or move data at all. Significant time savings emerged without the need to copy data from the original operational sources to a central analytical repository. This approach is clearly a significant improvement on all predecessors in terms of the immediacy with which data can be analyzed and through reduction of effort to make data accessible.

While the idea is sound and the value is self-evident, data federation alone isn't scalable for large amounts of data, nor for large numbers of simultaneous queries. In addition, because it relies heavily on the speed and stability of the source systems and network, its performance is commonly diminished for both, data analysis and production operations. So while data federation is quick and flexible it is not scalable or particularly dependable. But it was an important step in the right direction.

## Data Virtualization

Taking data federation several steps further, data virtualization expands the virtual connectivity aspect to include a wider spectrum of sources from RDBMS to data appliance, NoSQL, Web Services, SaaS and enterprise applications. It also combines data federation with caching repositories to address the shortcomings of data federation. This approach is often used as part of a big data solution to complement data warehousing. The result is a combination of repositories, virtualization, and distributed processes for data management that delivers the best capabilities from several technologies. But it still falls short of the expectation for a robust, agile, and performant data warehouse. Caching can be problematic due to the need to schedule cache loads around performance concerns of source systems, as well as the fact that the cache is loaded into a single repository that may or may not be optimized for different data sets and/or data types.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

Still, in moving closer to modern data warehouses, virtual data technology is essential. First data federation and then data virtualization popularized the notion of virtual views, indices, and semantics. It also introduced the somewhat radical idea that data needs not be physically copied or relocated before it is accessed. In addition, virtual views can be altered without the need to transform and reload data, as in earlier data warehouse integration approaches, and this means that the changes can be presented immediately, without waiting for the data to populate through an overnight process. It is the virtualization of data integration that enables extreme agility in analytical development and significantly reduces build times and costs, all of which leads us to the next breakthrough in data warehousing.

## Modern Data Architecture

Looking back at the traditional data warehouses and the data lakes (including the latest data lakehouse), they share one commonality: they all follow a monolithic approach and rely on having all data in a physical and central repository. Before you can work with the data, you have to corral it into a single location. This assumption, however, has been a barrier to accelerating data accessibility and it is what is fundamentally wrong with each approach previously discussed.

A modern data integration strategy moves away from the monolithic approach and employs what's known as "best-fit engineering", whereby each part of the data management infrastructure utilizes the most appropriate technology solution to perform its role, including storing data determined by business requirements and Service Level Agreements (SLAs). Unlike the previous concepts, these new architectures follow a distributed approach, aligning information storage selection with information use, and leveraging multiple data technologies that are fit for specific purposes. A hybrid approach can also significantly reduce costs and time to delivery when changes or additions in the warehouse are required.

## Logical Data Warehouse

One concept in the direction of modern data architecture is the Logical Data Warehouse. Another is the Virtual Data Lake. In either case, the premise is that there is no single data repository. Instead, the logical data warehouse is an ecosystem of multiple, fit-for-purpose, repositories, technologies, and tools that interact synergistically to manage data storage and provide performant enterprise analytical capabilities.

The original, and so far unfulfilled, analytical requirements of the traditional data warehouse were to be able to retrieve data using a single query language, get speedy query response, and to quickly assemble different data models or views of the data to meet specific needs. By combining data federation, physical data integration, and a common query language (such as SQL), the logical data warehouse approach achieves all three of these goals without the need to copy or move all the data to a central location.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

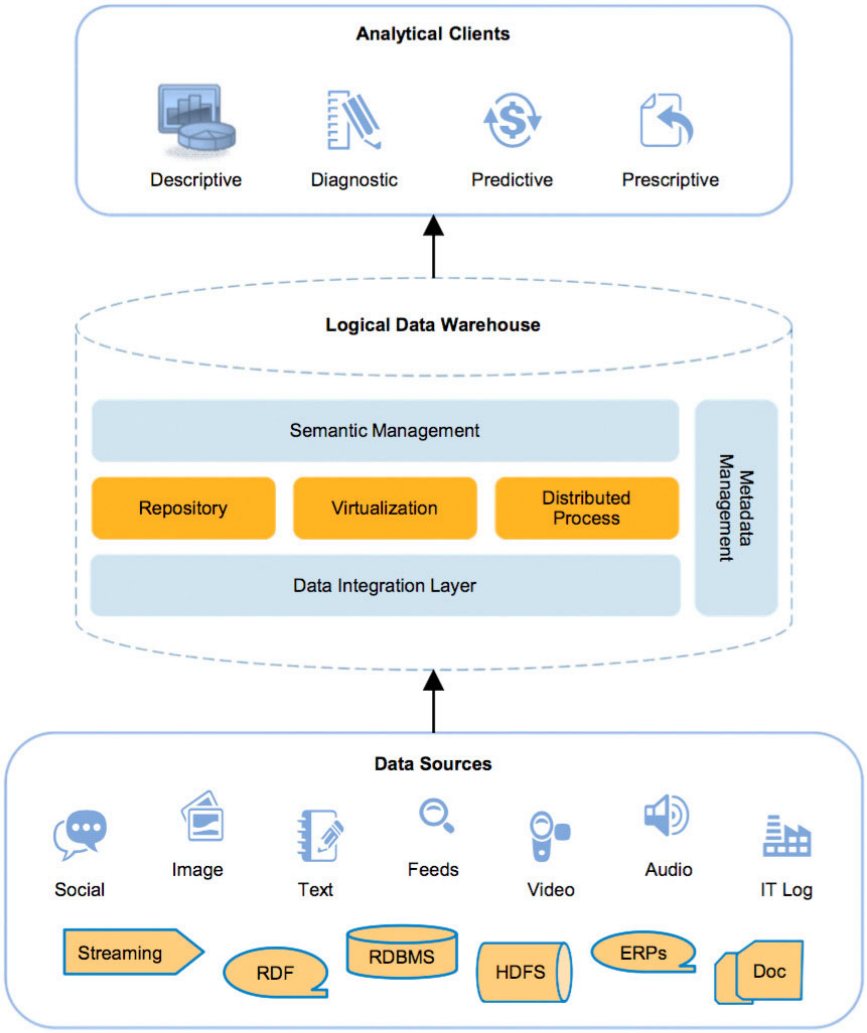
Conclusion

Data Virtuality

Use Cases

Industries

Physical data integration is a robust feature of the logical data warehouse that ensures fast query response while decoupling performance from the source data stores and moving it to the logical data warehouse repository. In this manner, the effort-intensive, physical transfer of the data is minimized and simplified, effectively removing lengthy data movement delays from the critical path of data integration projects.




Relational database management system (RDBMS)
Hadoop Distributed File System (HDFS)
Resource Description Framework (RDF)

Source: Gartner (September 2014)

While the LDW already delivers value to the users, Gartner sees shortcomings in the process that still needs to be implemented manually. Humans have to design the data structures, monitor and tune the system's performance, and document the contents of the system. Data Fabric is taking this into account and extends the foundations of the Logical Data Warehouse with automations.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

## Data Fabric

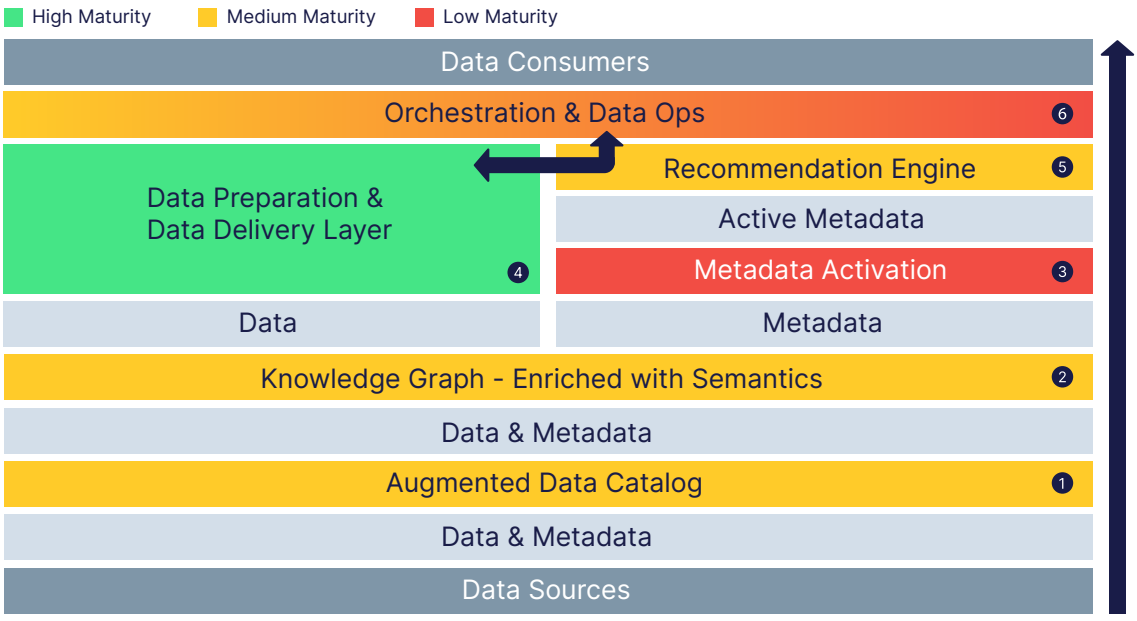
According to Gartner, organizations are looking into automating data integration (49%) and data preparation (37%) by the end of 2022. This has led to an increasing interest in the concept called Data Fabric. Evolving from the logical data warehouse, the data fabric is looking at automating the process integration, transformation, preparation, curation, security, governance, and orchestration to enable analytics and insights quickly for business success. Organizations can accelerate use cases such as customer 360, data science, fraud detection, internet-of-things (IoT) analytics, risk analytics, and he- althcare insights. Also, the pressure on data integration teams is alleviated by reducing repetitive and manual tasks of data optimization and provisioning processes.

Besides automation, the data fabric framework integrates the participating transactional systems and also incorporates them for better modeling capabilities through knowledge graphs. So while the logical data warehouse was still focused on the analytical use cases, the data fabric takes it one step further and enables operational use cases on top of the analytical ones.

While traditional data warehouses and big data platforms were mostly made for developers and architects, elements of the data fabric specifically aim to better enable business and non-technical users to efficiently leverage data, e.g. the ability to explore and manipulate data through a common semantic layer based on a knowledge graph is unique to data fabric.

The framework of the data fabric is still emerging and evolving. As shown in the image- from Gartner below, many of the elements are still at a low or medium level of maturity. But since this concept gained a lot of traction in the last couple of years many vendors claim that they provide data fabric (technology) creating a lot of confusion in the market. Ultimately, to fully assess this concept with all its weaknesses, more real-world implementations need to be realized and analyzed.

### Maturity of Data Fabric Components




Source: Gartner

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

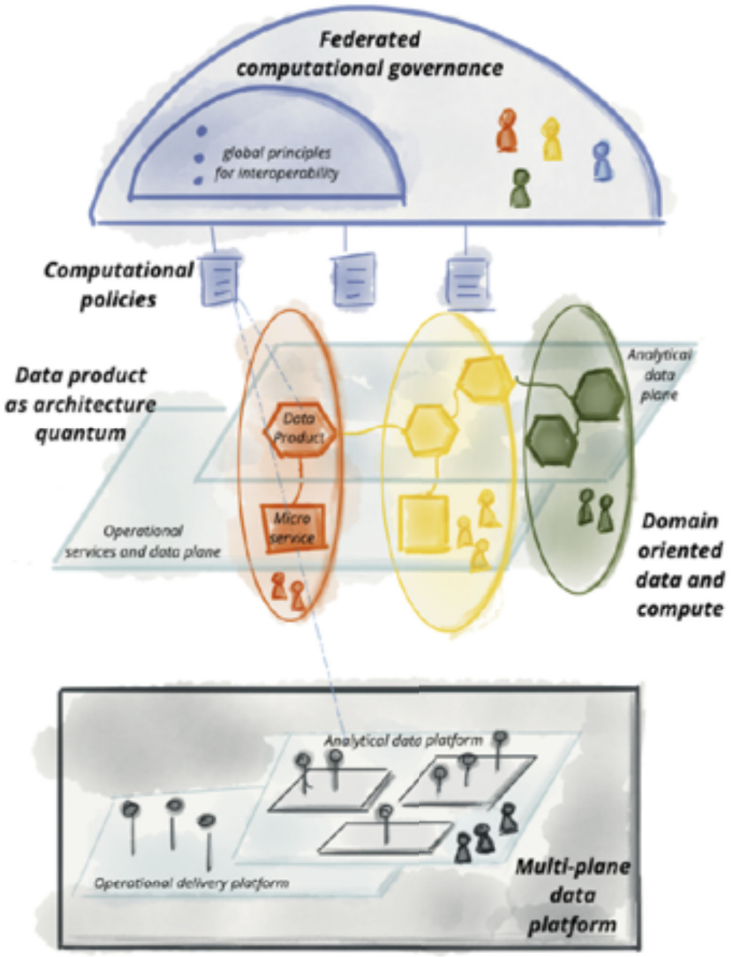
Data Virtuality

Use Cases

Industries

## Data Mesh

A socio-technical data management paradigm, called Data Mesh, proposed by Zhamak Dehghani is also challenging the traditional monolithic data architectures. It introduces a decentralized approach in which
1. data is owned by the teams that are closest to it as they are most familiar with it
2. data is a product that should be used and not an asset that needs to be treasured
3. a platform enables self-service data infrastructures with capabilities for data integration and transformation, implementation of security policies, data lineage, and identity management.
4. federated computational governance is supported for a healthy and interoperable ecosystem.




This concept differs from the other ones as it is technology-agnostic and focuses more on the human/socio part of the data management challenges. The data mesh framework claims that the previous concepts concentrated too much on the technology and thereby missed to fully understand and address the needs of the business that ultimately uses the data for insights. With the decentralized approach, the data mesh tries to bridge the gap between the business needs and the technology. On the technical side, it recognizes and respects the distributed nature and topology of the data and the different use cases that it can enable. On the human side, it looks at the individual personas of data consumers, their diverse access patterns, and their domain specific knowledge.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

Data Mesh is an interesting concept for companies that deal with a large number of disparate data sources. It is also relevant for companies that deal with different sets of consumers coming from operational and/or analytical systems of the business.

This framework surely has a lot of potential. However, it is still in an early emerging stage with very little real-life implementations. Many weaknesses and challenges are still unknown and unclear. It will be interesting to see how it evolves in the next few years!

## Conclusion

Data integration plays an essential role for the success of an organization's data strategy. To fully benefit from the value of the data, organizations need to break data silos and integrate multiple disparate data source in a flexible and agile way. At the same time, data governance and data quality aspects need to the ensured, self-service initiatives supported, and business users involved. The importance of the latter is well reflected in the latest trends such as Data Mesh. If data teams don't get the buy-in from the business side, where the actual users of the data reside, the data strategy is at risk. While our ability to integrate data still outstrips our ability to effectively involve the different teams, we are making great progress in balancing the two. Data integration vendors are starting to take this into account and provide features to support data teams in this regard, e.g. low code development, SaaS deployment, and specialized data portals for business users.

The ability of data integration to not only integrate data but also to integrate the different teams will be a major key for the future of enterprise data strategies.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

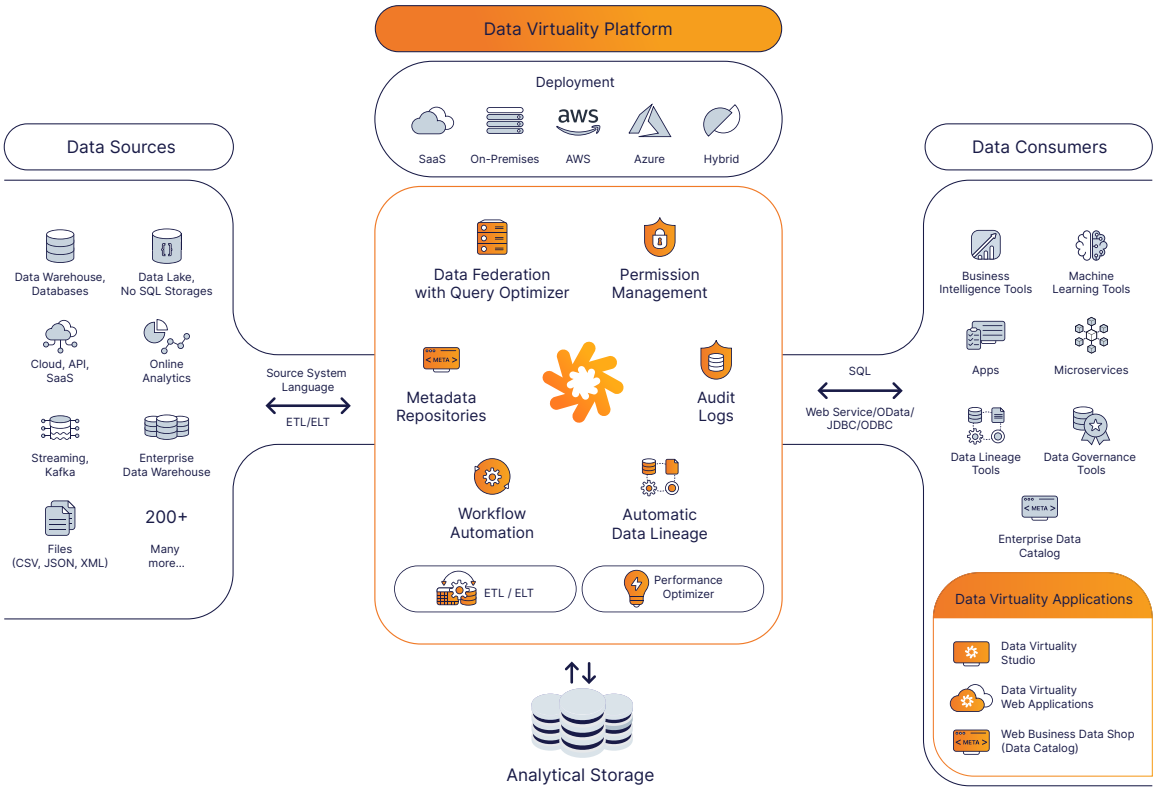
# Data Virtuality

## Smart Data Virtualization for Flexible Data Architectures

Decades of experience have shown that data in organizations is distributed. That's why the latest concepts incorporate these findings in their designs of data architectures. Data Virtuality supports distributed infrastructures with a centralized approach and thereby enables modern data architectures such as data fabric, data mesh, and the logical data warehouse. Data virtualization and ELT/ETL are combined in a unique way so that data teams have several data integration possibilities in a single tool and can choose the right method for the specific requirement.

By consolidating relational and non-relational data sources, including real-time data, Data Virtuality Platform enables immediate analysis using SQL. Integrated connectors to all data-producing and data-processing systems, including ERP and CRM systems, web shops, social media applications, and just about any SQL and NoSQL data sources allow data to be immediately processed using analysis, planning, or statistics tools, or written back to source systems as needed. With instant access to the data, users can begin experimenting with these connections and joins until they achieve the results they want.

By decoupling the semantic unified data access layer, in which the business users interact, from the actual data sources, changes that occur in the original data source can be isolated from interfering with analytical processes. In a profound departure from past data accessibility strategies, business users can interact with data comfortably and easily, focusing on their objectives rather than the technological underpinnings.



The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

## Data Governance Success through the Data Virtuality Platform

The Data Virtuality Platform simplifies and secures hybrid environments wherever data is stored, even when dispersed throughout various locations such as on-premises, cloud-based, third parties, or anywhere else. This greatly increases the likelihood of success for any data governance program. To promote data governance, governed data assets need to be provided that are so easy to use that there is no reason to search out alternative methods to source data for data consumption. By making the governed data complete, accessible, and easy to find and use, it is easier for business users to adopt the use of governed data, leaving them no reason to undertake efforts to source data themselves in an ungoverned manner. This willingness and voluntary adoption will be a key driver to success for data governance and greatly increase the likelihood that data consumption is done in a safe and accurate manner leading to accurate results.

By utilizing the Data Virtuality Platform as the preferred data access gateway throughout your organization, you can centralize data governance administration. This approach simplifies the fulfillment of data protection, regulatory, and compliance requirements for GDPR, HIPAA, CCPA, BCBS239, SFTR, FRTB, Solvency II such as data lineage, responsibility and accountability, timeliness, access monitoring, system auditing, and data classification, all while reducing the risk of a data breach.

Further, the Data Virtuality Platform adds flexibility regarding where and how you apply data governance functions to ensure aspects such as data quality and improve Master Data Management (MDM) in support of your data integrity objectives.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

# Use Cases

### Hybrid- and Multi-Cloud Architectures

The cloud data ecosystem plays an essential role in fully enabling and benefiting from multi- or hybrid-cloud architectures. The gaps in regard to data integration, metadata management, data governance, and data quality cannot be filled solely by one single cloud provider. The Data Virtuality Platform helps to plug in these gaps of the cloud platforms with the different data management capabilities such as data integration, data quality, data governance, master data management, and metadata management.

### Semantic Layer for Data Science and Self-Service BI

Built in the virtual layer of the Data Virtuality Platform, all data sources are connected and the data is accessible (also real-time) in a central data access layer. This secured and governed environment provides access to a large number of users including data scientists and BI. The metadata is searchable so users can easily find and work with the data, even in their web browser.

### Customer Data Integration and Customer 360 Degree View

Understanding the customers and prospects' behavior as individual entities is the key idea of a comprehensive data-driven marketing. Data Virtuality helps to simplify data management and data integration for a 360 degree view on customers and prospects during their whole journey. The marketing channels can be optimized and customized so that the marketing team can create memorable moments for the customers and achieve the following: Lower customer acquisition costs, better conversion rates, and increased retention rates.

### Digital Marketing

Digital marketing is extremely data-driven, relying on the volatile flow of real-time data. The Data Virtuality Platform offers a viable way to manage complexity of this kind, easily connecting to a host of digital marketing data providers for affiliate marketing, performance marketing, personalization, and other approaches.

The Evolution of Data Integration

Data Warehouse and ETL

Self-Service Business Intelligence Tools

Data Lakes and ELT

From Data Federation to Data Virtualization

Modern Data Integration Architectures

Logical Data Warehouse

Data Fabric

Data Mesh

Conclusion

Data Virtuality

Use Cases

Industries

# Industries

### Financial Services

Stricter regulatory requirements such as BCBS 239, SFTR, FRTB, Solvency II, and GDPR have created the need for an increasingly intricate data management environment for financial services institutions. Legacy best practices typically implemented for financial firms are outdated and can't meet the increasingly demanding regulatory requirements. The data warehouses and data marts populated via ETL are inflexible and lead to protracted time-to-market cycles and lack of transparency, accountability, and auditability. By combining automated ETL and data virtualization, the Data Virtuality Platform enables a flexible data supply chain for financial services institutions and helps to accelerate implementation by up to 75% and cut costs by up to 70%.

### Retail and E-Commerce

Modern data integration offers a compelling solution for e-commerce and retail organizations with a great number of different systems in their IT landscape. For example, a typical e-commerce business has an ERP system, CRM, web and mobile apps, email analytics programs, online marketing, social media marketing, and other tools. With the Data Virtuality Platform, all these data sources can be joined quickly and flexibly to provide comprehensive views of any data related to customers, products, etc.

### Healthcare

Healthcare providers are faced with operational challenges that come with severe economical impacts such as high costs due to long wait times, inefficient staff planning, and high staff turnover rates due to high stress and burnouts. The integrative Data Virtuality Platform solves the challenges by aggregating data from different systems for transparency and automation of workflows. This opens new possibilities such as dashboards that operational teams can use to track patients in real-time and better manage the patient-flow. The results are profound: reduced costs, lower staff turnover rates, and more efficient management/planning of the bed occupancy, to name a few.

### Data Marketplace

A data marketplace is an online transactional platform where data is shared and monetized. Integrating data from many different sources and providing self-service access while ensuring security, consistency and high quality of data is essential to these data marketplaces. A central data access and delivery layer in the Data Virtuality Platform allows a secure and governed transaction for both parties.

Nick Golovin, PhD, CEO of Data Virtuality, oversaw the BI at Koch Media during the period in which the importance of digital information began to grow exponentially. He quickly realized that the tools available to connect and manage data from multiple data sources were not capable of meeting the rapidly changing needs of the businesses.

Nick came to the conclusion that in-house development is too slow, and that the current data integration tools available in the market were too inflexible. Rising to the challenge, Nick decided to pair his work experience with 8 years of academic R&D to build an innovative solution, what we now call Data Virtuality Platform.

## About Data Virtuality

- Founded:
  2012 by Nick Golovin (PhD) in Leipzig, Germany after 8 years of research
- Offices:
  Munich, San Francisco, Leipzig
- Solutions:
  Data Virtuality Platform SaaS
  Data Virtuality Platform On-Premises
  Data Virtuality Pipes Professional
  Data Virtuality Pipes
- Acknowledgements:
  Honorable Mention in 2022 Gartner Magic Quadrant for Data Integration Tools
- Awards:
  Most Innovative Data Management Provider 2022, 2021 and 2019 (A-Team Insights)
  2020 and 2019 Deloitte Technology Fast 50

**Message:** info@datavirtuality.com
**Visit:** datavirtuality.com
**Request Demo:** demo@datavirtuality.com
**Data Virtuality Platform SaaS Free Trial**: https://eu.platform.datavirtuality.com/#/start-trial