CONTIAMO

White paper

Accelerate data experimentation with data virtualization

Agile data projects: the key to business agility

The pace of change in today's business landscape is faster than ever. To keep up employees must maintain their situational awareness and make informed data-driven decisions, while business processes need to be optimized and automated.

The advent of artificial intelligence and machine learning opens up opportunities to achieve these goals. Through applying these new technologies in successful data projects, organizations can more efficiently allocate resources, drive revenue and keep up with the ever-changing business environment.

However, with 60% of data projects failing[1], organizations must introduce speed and a culture of experimentation into their approach. The most successful organizations will be those that introduce true agility into their data projects.

The challenge with agile data projects

The agile approach to traditional project management has become standard practice from startup software development teams to corporate project management teams. However, implementing this approach for data projects is a significant challenge for both SMEs and large enterprises, largely due to data access issues. Data access is often a slow and frustrating process for both the data engineer or data steward and the data consumer. An average of 70% of time spent on data projects is focused on accessing and preparing the required data[2].

70% of time on data projects is spent on data discovery and preparation

Data engineers face several challenges that inhibit the timely provisioning of data. Firstly, the data required for a single data project may reside in multiple systems, some of which are locked behind access controls, with often unclear ownership. Secondly, where data is accessible, there may be multiple inconsistent copies, or even "shadow data" being shared as a CSV or JSON file. This method operates outside of enterprise data governance controls, often breaching internal data security regulations or external data privacy regulations, such as GDPR. Additionally, the role of the data engineer becomes not only providing the data but ascertaining the most reliable data set. Even once underlying data access is attained, data engineers must engage in time-consuming and CPU-heavy Extract, Transform, and Load (ETL) processes, where the prepared data is copied into a new data storage solution to be consumed by a data scientist, BI analyst or another data consumer.

For the data consumer, the process begins even earlier. When discussing potential new projects, there is often a lack of transparency into what data sets exist that could provide business benefit. This lack of transparency can blur project scope or even hinder the idea generation process. This results in some project ideas not even being conceived of, while others garner unrealistic expectations given the insufficient volume, challenging structure or existence of certain data. Once data is requested, data consumers can regularly expect a wait of days or weeks for the receipt of data. As a result, projects may have lost momentum, executive buy-in or the bandwidth of project team members.

This process is resource-intensive in terms of employee time, but also system cost. Standard ETL processes require the copying of data from source systems into new storage systems. This cost is amplified by the requirement for extremely large data volumes for projects involving big data and machine learning, compared to traditional BI and analytics projects. For many organizations, this makes experimentation in machine learning cost-prohibitive. In parallel, most IT

organizations are inundated with requests for data from business teams. Current processes and technologies prevent the IT organization from keeping up with these requests in a timely manner.

45% less time is spent on data integration by those who leverage data virtualization

Data virtualization provides a solution to a number of problems that inhibit organizations' shift to experimentation and agility in data projects. By 2020, organizations leveraging data virtualization for data delivery will spend 45% less on their data integration processes, than their counterparts[3].

What is data virtualization?

Data virtualization is a modern solution to data integration. Data is logically integrated from underlying data sources, such as a traditional on-premise databases. Logical integration means that data is generally not moved or replicated, but rather the data virtualization solution only stores the information on how to access and interpret the integrated data.

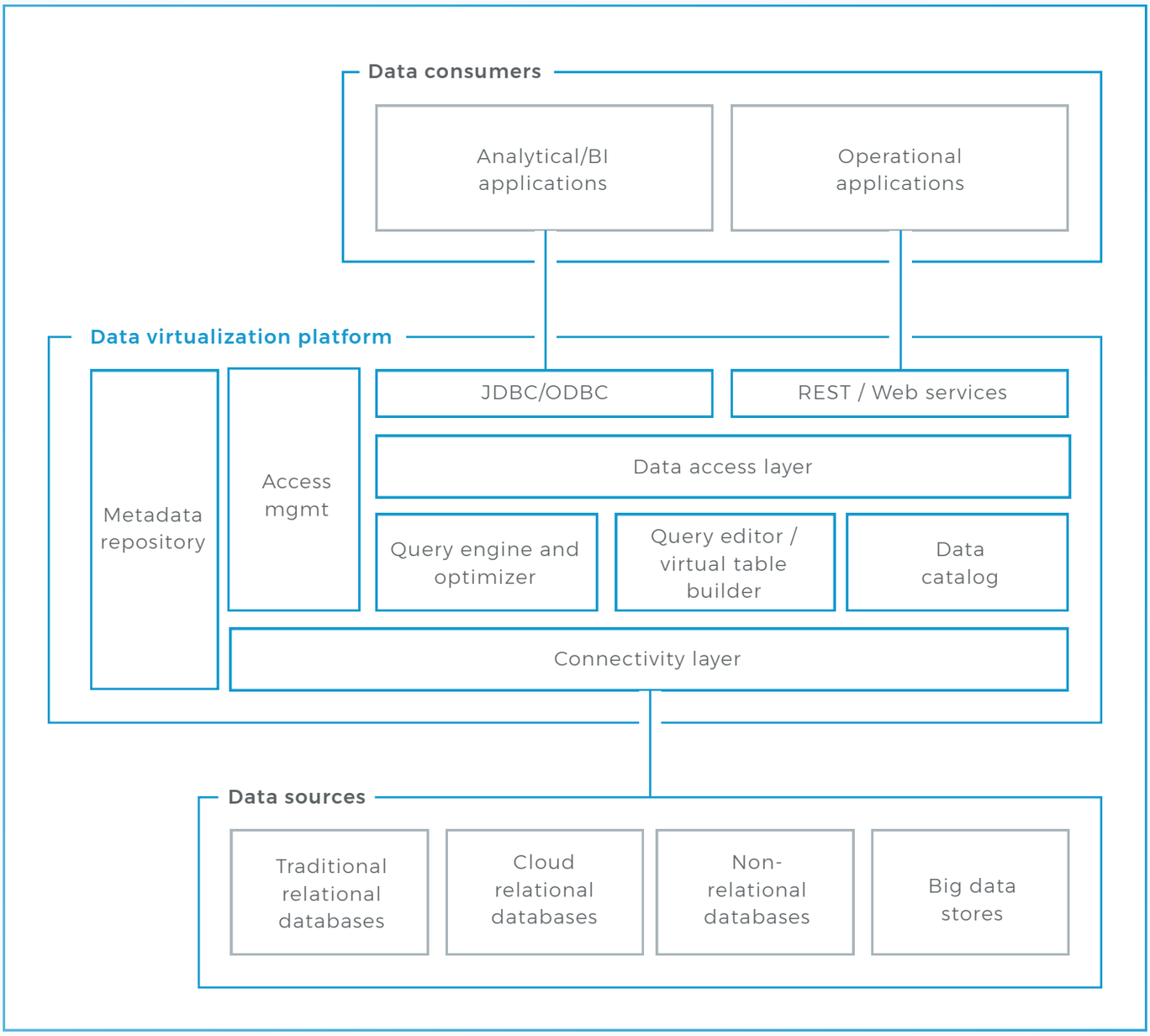
> *Data virtualization is a modern solution to data integration. Data is logically integrated from underlying data sources, such as a traditional on-premise databases.*

The creation of these new, distinct data sets results in what are known as virtual tables. The data, often from multiple systems, resides in those source systems, while the data consumer can access the data as if it were a single data source.

Many data virtualization platforms can catalog both the underlying data sources and the newly created virtual tables. Thus, data virtualization can become a single point of access to share and explore data.

Virtual tables are mostly shared outside of the data virtualization platforms and used to power analytical and operational use cases. The data delivered is always fresh, as data is queried and delivered from underlying systems on-demand.

Data virtualization in your architecture



Business benefits

Speed

Both the setup and deployment of data virtualization is significantly faster than traditional methods. The integration of data sources into data virtualization platforms is usually aided by a wizard so integration is done in a few clicks. Configuring virtual tables is fast and comfortable, with queries being written in SQL, conducted in a graphical user interface (GUI) or sometimes in a platform's proprietary language. The speed of data access is 10x faster with data virtualization versus traditional methods.

Resources

As data is not replicated or moved into new systems, organizations can make significant savings on data storage costs. Additionally, expensive ETL processes that require substantial computing power can be replaced by data virtualization, which runs only on-demand versus a regular refresh.

IT resources can also be significantly reduced as the process of responding to data requests becomes more efficient. They can then be applied to other important initiatives, increasing the value that IT resources provide to the organization as a whole.

Data quality

Implementing data virtualization for data projects, especially pilots and proof-of-concepts eliminates the sending of CSV and JSON files for data projects. This elimination of shadow data prevents data inconsistencies and ensures all projects are working with the latest data. The enhanced data quality ensures better outcomes for data projects.

Transparency

As data virtualization platforms can also act as powerful data catalogs, data consumers can search and discover new data sets. This ability to discover data sets enables consumers across the organization to generate new ideas and pursue new innovative opportunities that will drive business value.

Key features

Native data connectors

Data virtualization platforms are characterized by their ability to easily integrate data. The first capability backing this characterization is the native, out-of-the-box connectors to common data sources. The built-in connectors generally cover traditional relational databases (Oracle DB), cloud relational databases (Postgres), big data stores (Hadoop), and non-relational data stores (NoSQL).

The connection to these data sources is assisted by simple wizards that walk the user through the process of providing the necessary information to add the data source. Some data virtualization platforms enable custom data sources to be added to augment integrated data. Among others, these custom sources include CSV and JSON files. The structure of both custom and real-time data sources is generally detected by the data virtualization platform, with the intelligent discovery of data types, where the relevant metadata is not provided by the underlying system.

Data aggregation	Data virtualization acts as a layer to centrally manage the business meaning of data, enforce data governance and gain visibility into the details of data sources across the data architecture. This is achieved through a centralized metadata repository. In this repository, the physical schema of underlying data sources is explained in a logical schema. This logical schema explains the business meaning of a physical data set and can link data sets through common identifiers. For example, a user in the retail industry benefiting from data virtualization can take three disparate data sources containing customer loyalty data, store sales data and digital marketing campaign data and link the three data sources using a common customer ID. This data aggregation is achieved without any data movement or replication.
Query interface	The implementation of data virtualization requires a query interface whereby the user can query data from across data sets using SQL or a proprietary schema language. This interface enables simple data exploration as query results are displayed as simple tables and potentially visualizations. Advanced data virtualization platforms profile data sets and underlying data sources within this query interface. As a result, data practitioners can spot trends, data inconsistencies or errors within their data.
Virtual tables	In addition to data exploration, query interfaces are used to create virtual tables. Data practitioners can build complex queries, where data is joined, transformed, filtered and more to ensure that virtual tables contain only the data required for a given use case. To the consumer, these virtual tables appear as single data sources. The data virtualization platform stores only the metadata and queries, while the data consumer accesses fresh data from source systems, on-demand.
Data cataloging	Data virtualization platforms can also act as a catalog for both physical data sources and virtual tables. Augmenting the necessary capabilities of documenting the virtual tables, advanced data virtualization platforms can become internal marketplaces, whereby data consumers can discover relevant data sets. Both physical data sources and virtual tables can be made searchable, with

information about access rights, data structures, data profiles, comments or ratings, and even current implementations and use cases.

This enables data consumers from across the organization to easily discover data sources previously unknown to them. Access request processes become far more efficient with transparency to who owns or "stewards" the data. The ability to request and grant access to virtual data sets can be carried out directly in the data virtualization platform.

Materializations

Under the hood, data virtualization boasts several enhancements on traditional integration methods that ready it for a new class of use cases.

Data virtualization leverages query execution engines, such as Apache Spark or Apache Calcite, to ensure that queries are executed according to the most efficient plan. Some data virtualization platforms have intelligent routing of queries to different execution engines, depending on data volumes, use case requirements and query frequency.

When setting up virtual tables, data virtualization enables the materialization of data where needed. Materialization is the process of storing data from virtual tables in new systems. These systems may be far more adept at dealing with near-real-time data delivery.

Where near-real-time data is required and the data is stored in a traditional data source that can only return data at high latency, a data practitioner may choose to materialize data in a more modern system.

Materializations often have several parameters, such as data refresh scheduling, load-limiting, partial materialization targeting specific rows or columns, and intelligent materializations that learn which parts of the table require frequent refreshment.

Downstream integrations

Virtual tables can be consumed in several ways. Firstly, some data virtualization platforms enable simple data exploration with data tables and elementary visualizations.

The majority of data virtualization use cases require data to be consumed by upstream applications.

The consumption of data in analytics or BI software is achieved primarily through JDBC or ODBC drivers.

Other deployment options include making virtual tables available through web services, such as REST or XML.

Data virtualization vs. ETL

In most cases, data virtualization and ETL are not competing technologies but rather two solutions to distinct challenges.

The primary use of ETL is to bulk copy complete data sets, transforming them to be consistent with an enterprise data model (EDM). This process is optimized to handle very large, structured data sets. Running these ETL processes is resource-intensive and expensive so they are generally run on a regular, but infrequent, schedule. Due to the high load of source systems, ETL processes are often run outside of working hours to prevent downtime risk or strain on systems competing for the same resources.

To support the large-scale transformation of data supported by ETL processes, workflow engines are often part of the ETL process. Data cleansing and data transformation is integrated into the ETL process and occurs each time the data is refreshed.

> *Data virtualization is a modern solution to data integration. Data is logically integrated from underlying data sources, such as a traditional on-premise databases.*

While ETL processes are complex and time-consuming to set up, the results are extremely well-suited for data mining, historical analysis, and strategic planning. However, ETL is not set up for operational use cases and the ever-changing data requirements of most modern data projects.

Data virtualization is a much more agile option. The setup of virtual tables in data virtualization platforms can be often completed in minutes. This makes experimentation in data projects much cheaper, so data virtualization is the ideal solution for pilots and proof-of-concepts where the data model is still in flux and fast iteration is imperative.

Additionally, where running ETL processes continuously is prohibitively expensive, queries from data virtualization solutions run on-demand. This ensures that data is always as near-real-time as possible. In operational and analytical use cases that require fresh data, data virtualization is more well-suited than ETL.

Finally, many data sets have strict restrictions on how and how often data is replicated or moved. Unlike ETL, data virtualization does not replicate or move data from the source systems. Thus, data virtualization is a potential solution for any use cases that require data with restrictions on replication or movement.

Category	ETL	Data virtualization
Implementation time	Slow (week to months)	Fast (minutes to days)
Implementation cost	$$$$$	$$
Data structure stability	Not subject to change	Fast iteration
Data replication	Yes	No (unless desired)
Data cleansing	As part of ETL process	Precursor to data virtualization
Data transformation	Very complex, with workflow engines	Complex, manually-written
Data freshness	Only fresh from end of last, often infrequent, refresh cycle	Near real-time
Use cases	- Business intelligence - Data mining - Historical analysis - Strategic planning	- Data project pilots and proof-of-concepts - Near real-time analytical and operational use cases - Use cases involving data with replication/movement restrictions

Data virtualization in the past, present & future

The concept of data virtualization first arose in the early 2000s, primarily focused on data federation or Enterprise Information Integration (EII). Much like modern data virtualization, these early technologies aimed to logically integrate physical data sources into a central data store. However, the technology lacked the usability and optimizations that make data virtualization feasible for use in the modern data architecture.

Data virtualization has matured in recent years and, for many leading organizations, has now become an indispensable tool in their data integration architecture, due to its flexible and agile properties. Modern enhancements and optimizations of the technology have expanded the set of use cases for which data virtualization is advantageous.

For many production analytical and operational data use cases, it has many benefits over traditional integration methods, boasting near-real-time access to fresh, reliable data. Indeed, over 35% of enterprise organizations are already using data virtualization in production. Additionally, for pilots and proof-of-concepts, many organizations are choosing data virtualization for data delivery. The time to value and ability to iterate make it the tool of choice for agile data projects.

With innovative solution providers disrupting the data virtualization space with new features, distinct approaches, modern interfaces, and unprecedented simplicity, the prominence of data virtualization looks set to rise. By 2020, 60% of organizations will implement data virtualization as one of the key delivery methods in their data integration architecture[4].

Find out about data virtualization and speak to an expert at www.contiamo.com.

References

1. Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics; Laurence Goasduff; 15 Sept 2015; https://www.gartner.com/en/newsroom/press-releases/2015-09-15-gartner-says-business-intelligence-and-analytics-leaders-must-focus-on-mindsets-and-culture-to-kick-start-advanced-analytics

2. What's Your Data Strategy?; Leandro DalleMule & Thomas H. Davenport; June 2017; https://hbr.org/2017/05/whats-your-data-strategy

3 & 4. Market Guide for Data Virtualization; Ehtisham Zaidi, Mark Beyer, Ankush Jain, Sharat Menon; 16 Nov 2018; https://www.gartner.com/doc/3893219

CONTIAMO

White paper

About Contiamo

Contiamo is a data virtualization platform that radically accelerates data access to deliver agility and flexibility to data projects. Through its SQL-only setup, it is the fastest way for data engineers to deliver data. It is also offers comprehensive data catalog functionality, resulting in the simplest way for data consumers to search and access data.

For more information, go to www.contiamo.com