# k2view

Dynamic Data Virtualization
A Practical Guide

This article defines data virtualization, goes on to discuss its drivers, challenges, benefits, use cases, and limitations, and concludes with a novel data product approach that promises to make the technology more dynamic and robust than ever before.

Intro

Real-time access to data is crucial for making business decisions and delivering value to customers fast. But today, enterprise data is dispersed across numerous systems, formats, and technologies. On top of that, new sources are continuously being added in pursuit of big data analytics and modernization of legacy apps.

To tear down data silos, and improve access to business-critical information, many organizations are turning to an advanced type of data integration called data virtualization – offering new levels of access, speed, and efficiency in moving data across disparate systems.
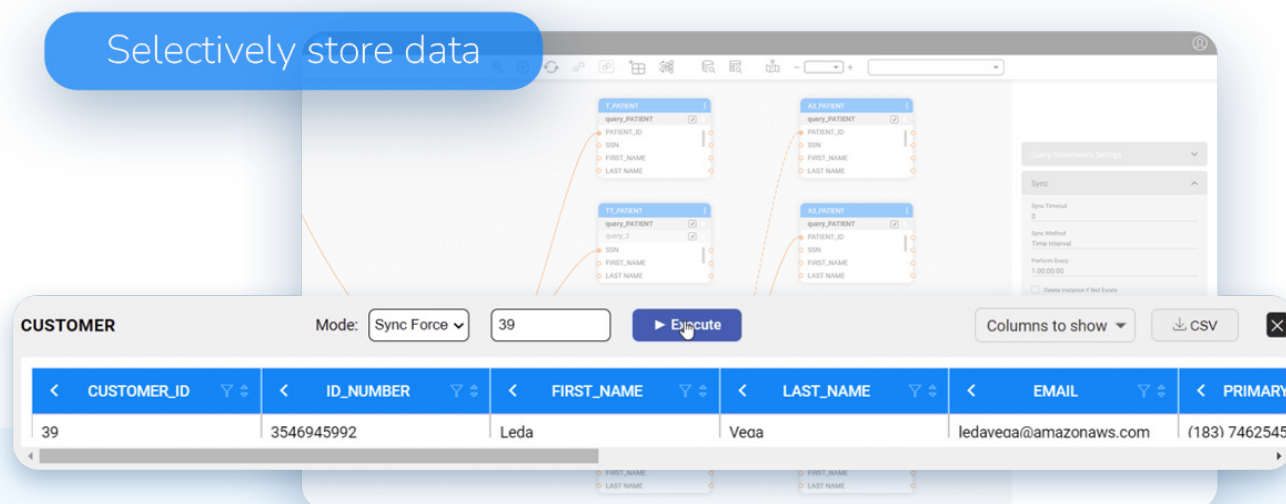
What is Data Virtualization?

According to analyst Gartner, data virtualization technology is based on a distributed data management architecture (such as **data mesh**). It is used primarily for queries against many different data sources, and the federation of query results into virtual data views – consumed by applications, query/reporting tools, message-oriented middleware, or other data management infrastructure components. The virtual views, which can be stored in memory, provide a layer of abstraction beyond the physical implementation of data, to simplify querying logic.

In short, data virtualization is a unified, logical data layer that integrates enterprise data across multiple systems, while allowing for centralized governance and security.

With data virtualization, applications and users can access enterprise data, regardless of its location, format, or protocol. Unlike traditional ETL tools, which physically copy data from separate sources into a target system, like a data warehouse or data lake, data virtualization enables business users to access data where it resides, without creating a physical replica along the way.

Data virtualization can support multiple use cases and business functions. However, it is not a silver bullet. Along with its many benefits, data virtualization software comes with several significant limitations. In this article, we'll examine the benefits of data virtualization, its drawbacks, and why a data product approach to data integration mitigates the limitations of data virtualization.

Data Virtualization Drivers

Data virtualization allows for federated and real-time data access, integration, and sharing – without the need for replication, or additional data silos.
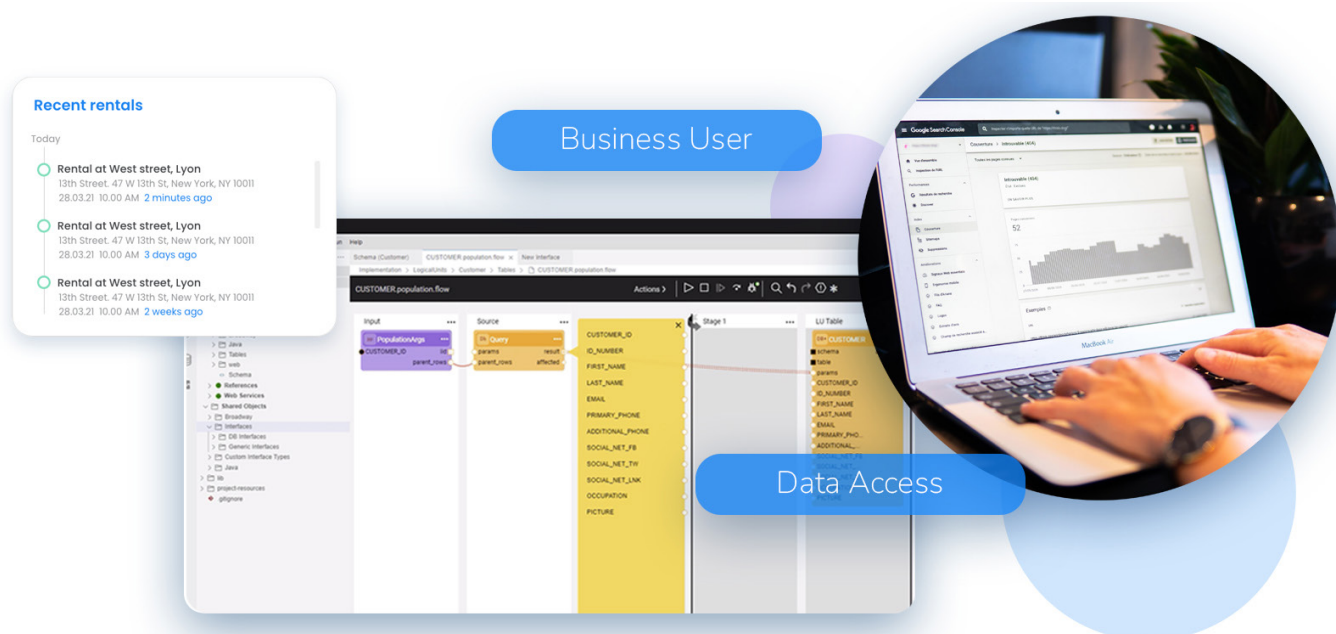
Data virtualization gives less technical business users access to data.

Here are the key business drivers:

- **Cost**
  Data virtualization is a cost-effective approach to data integration, which can even replace ETL when Transformation (the "T" in "ETL") requirements are light (or non-existent).

- **Data sharing**
  Data virtualization is useful for integrated data access and sharing, especially in light of regulatory constraints on the physical movement of data. It also allows for cross-platform data analysis, such as combining historical data from a data warehouse with data from an external third-party data source to respond to a particular query.

- **Experimentation**
  Less technical roles, like domain-based citizen integrators, can participate in federated data access, to obtain better business outcomes. An integrated dataset, quickly provisioned as a "sandbox", can provide citizen integrators with the ability to do experimental integration, and to get access to data for ad-hoc queries that arise.

- **Flexibility**
  Data virtualization lends flexibility to enterprise data integration architectures, by allowing users to virtualize and integrate different data models and technologies – without physically moving data into a repository – including unstructured and diverse types of data (e.g., IoT and time series data).

- **Reusability**
  With data virtualization, enterprises benefit from reusability of integration artifacts across a wide range of business needs, including logical data warehousing, virtualized operational data stores and master data management (MDM) use cases.

Data Virtualization Challenges

For maximum agility – in making faster decisions, serving customers more personally, and outdoing competitive offerings – enterprises recognize the need to use their data more effectively.

Exploiting the power of data analytics, business intelligence, and workflow automation is one way for companies to accelerate new revenue streams while reducing costs and improving the performance of data services.

But here lies the challenge – enterprise data is stored in disparate locations with rapidly evolving formats such as:

- **Relational and non-relational databases,** like Amazon Redshift, MongoDB, and MySQL

- **Cloud/Software-as-a-Service applications,** such as Mailchimp, NetSuite, and Salesforce

- **Semi-structured and unstructured data,** including social media data, call recordings, email interactions, images, and moreLegacy systems, such as mainframe and midrange applications

- **CRM/ERP data,** like Microsoft Dynamics, Oracle, and SAP

- **Flat files,** such as CSV, JSON, and XML

- **Big data repositories,** including **data lakes and data warehouses**

The demand for speed and higher volumes of increasingly complex data leads to further challenges such as:

- **Self-service capabilities** for data users
- **Time efficiency** in data management
- **Trusted data quality** in terms of cleanliness and freshness

In a virtualized environment, different domains can access the data they need, on their own, in real time.

To address these challenges, organizations recognize the need to move, from silos of disparate data and isolated technologies, to a business-focused strategy where data and analytics are simply a part of everyday life for business users.
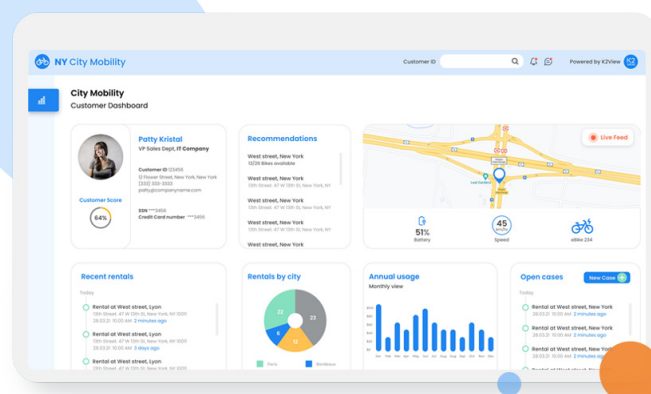
Cultural obstacles

Data virtualization needs to overcome certain obstacles before it is widely accepted. For example, some may think the semantic virtual tiers built into some of their applications can substitute for data virtualization, while others might assume that data virtualization is a replacement for ETL.

Both assumptions are faulty. Data virtualization is a cross-platform technology that must be used in conjunction with other data delivery styles (e.g., bulk/batch with ETL) and data transformation and orchestration – or it might limit performance, and inhibit its adoption.

Finally, data virtualization is less appealing to traditional data management teams because it challenges existing practices. For example, those who oppose more agile approaches, like "experiment-and-fail-fast", would most likely also resist data virtualization.

Data Virtualization Benefits

Here are the top 6 benefits of data virtualization:

Speed
Data virtualization enables data consumers to access data, wherever it resides (including traditional databases, the cloud, or IoT systems) in seconds.

Security and governance
Data virtualization enables a centralized approach to data security and governance, ensuring that all data is consistent, protected, and high-quality.

Efficiency
Data virtualization doesn't replicate data, so enterprises can save on governance and hardware, while enhancing the utilization of server and storage resources.

Access
Data virtualization allows for a self-service approach to accessing data, enabling quick access to data by any authorized data consumer

Cost savings
Data virtualization software requires fewer resources, and costs a lot less than building a separate repository for consolidating and storing data.

Analytics
Data virtualization lets business users apply visualized, predictive, and streaming analytics across many different data sources.

Data Virtualization Use Cases

Data virtualization supports 4 key use cases:

1. Data integration
Data consumers need access to data that is spread out across disparate data sources. With data virtualization, data consumers gain a holistic view of this data, regardless of its format or location.

2. Data analytics
Business domains require data analytics and business intelligence to support decision making. Data virtualization gives business users on-demand access to data through one centralized, virtual layer.

3. DataOps
Although many elements of application development are automated, data is not. Using data virtualization, DataOps teams can eliminate bottlenecks in data provisioning by giving users direct access to high-quality data, and the ability to collaborate cross-functionally.

4. Backup and production support
If a production issue occurs, development teams can create complete virtual data environments in which they can identify the cause. They can also validate that any changes made will not lead to unanticipated regressions.

Team work

Data virtualization helps data teams communicate and collaborate better.

Data Virtualization Limitations

Although the benefits of enhanced speed, access, and efficiency across business-critical use cases may sound enticing, data virtualization has several limitations you should be aware of.

1. **Inability to handle complex data transformations**
   Data virtualization is limited to simple data processing using database joins and rudimentary data processing. To support complex data transformations, it requires additional, complementary data pipeline tools such as ETL. It's also limited to offline/batch processing, making it less appropriate for operational workloads.

2. **Unreliable performance**
   Data virtualization performance is only as good as its data source performance and availability. In addition, when multiple joins are required to access data, performance is significantly impaired and often inadequate for real-time use cases.

3. **Distributed/federated deployments**
   Data virtualization across multiple data centers is very limited and complex.

4. **Not -designed for operational workloads**
   Data virtualization is great for returning large datasets, where response times aren't a critical factor. As a rule, it's more appropriate for analytical workloads, such as creating data lakes/warehouses.

5. **Lack of governance consistency**
   Data virtualization lacks capabilities to support consistent data governance and data cataloging.

Dynamic Data Virtualization: All the Benefits, Without Compromising

Dynamic data virtualization provides all of the benefits and capabilities of regular data virtualization, but enables a much higher level of flexibility and customization.

Whereas standard data virtualization "locks" customers into a 100% virtualization scenario, dynamic data virtualization is a hybrid solution, providing data teams the flexibility to determine which data should be virtualized, aand which to be persisted in a staging datastore.

Storing some of the data, rather than completely virtualizing it, is an advantage when you'd like to:

Minimize the burden on your source systems by storing data that doesn't change on an ongoing basis, instead of accessing it from the source every time.

Enrich the data with new fields that do not exist in your underlying source applications.

Maintain a log of the changes made to the source data by storing data that doesn't change on an ongoing basis, instead of accessing it from the source every time.

Implementing Dynamic Data Virtualization with Data Products

A data product approach to implementing dynamic data virtualization delivers complete, governed data-on-demand from a wide range of underlying data systems, formats, and structures. It provides access to live data through a logical abstraction layer, whose schema is the collection of all data tables and fields for a particular business entity (e.g., customer, device, supplier, or employee).

A data product creates and manages a "ready-to-use," complete set of data for a specific business entity, accessible for both operational and analytical workloads. It provides authorized users with everything they need to know about aspecific business entity in less than a second.

A data product approach gives you the flexibility to decide which data will be virtualized, and which will be stored physically.

And finally, a data product can be defined, created, and managed by a centralized data team in a data fabric architecture, or federated to business domains in a data mesh.
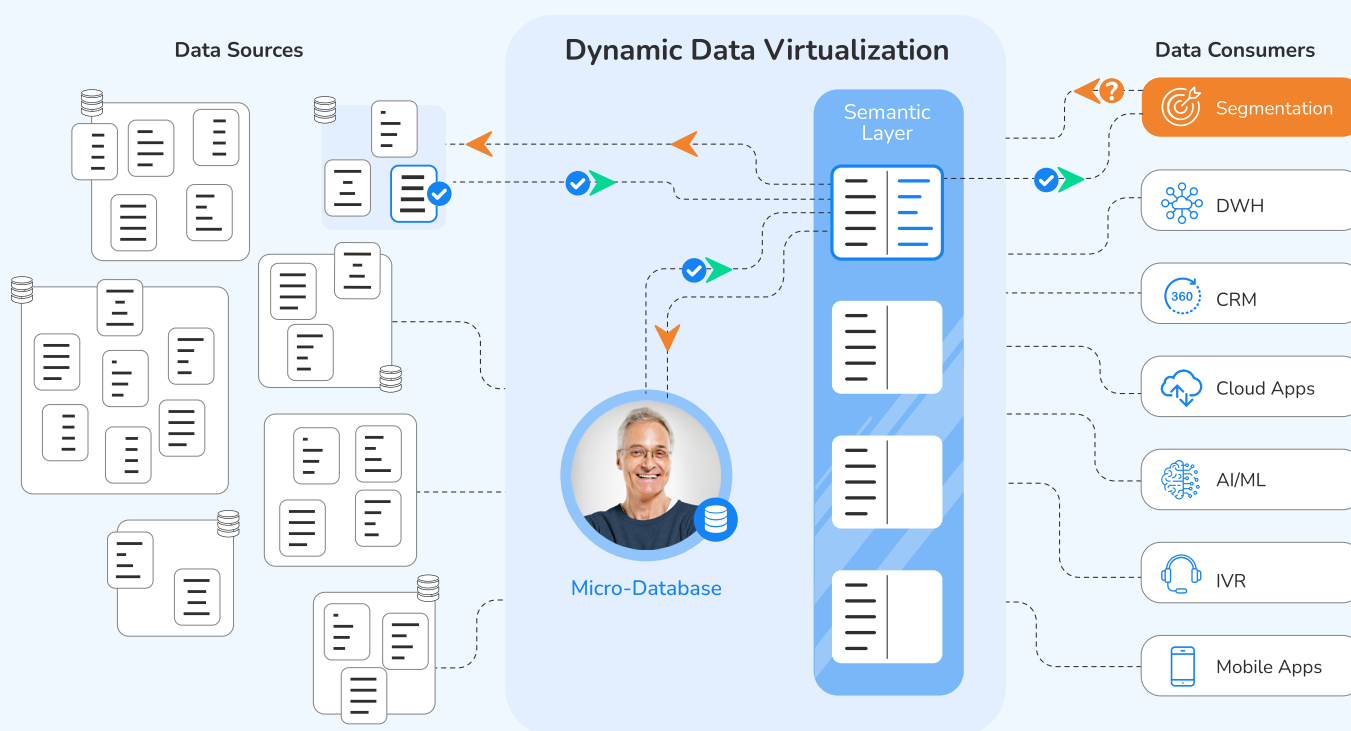
K2view Dynamic Data Virtualization ingests, unifies, transforms, and enriches data – via data products.

About K2view

At K2view, we believe that every enterprise should be able to use its data to be as disruptive and agile as Google, Amazon, and Netflix.

We make this possible by transforming all your data – wherever it is – into business-driven data products, which are defined and managed by business domains.

Data products could be customers, products, suppliers, orders – or anything else that's important to your business. We manage every individual data product in its own secure Micro-Database™, continuously in sync with all source systems, and instantly accessible to everyone.

This is all made possible by our operational data fabric, which delivers a trusted, real-time view of any data product. K2view Fabric deploys in weeks, scales linearly, and adapts to change on the fly. It supports modern data architectures, such as data mesh, data hub, and multi-domain MDM – in on-premise, cloud, or hybrid environments.

This one platform drives many use cases, including application modernization, cloud migration, customer 360, data privacy, data testing, and more – to deliver business outcomes in less than half the time, and at half the cost, of any other alternative.