Comparative Assessment of Data Virtualization Solutions

Published 17 November 2023 - ID G00799330 - 58 min read

By Analyst(s): Mayank Talwar, Ramke Ramakrishnan, Shubhankar Nandi, Yukti Mehta

Initiatives: Data Management Solutions for Technical Professionals

Data virtualization solutions have evolved to encompass broader themes such as data fabric, lakehouse technology and traditional data integration solutions. This research evaluates segments and the representative vendor solutions focused on this market.

Overview

Key Findings

- While the market for data virtualization (DV) has existed for over 20 years, these tools have naturally gravitated toward the logical data warehouse (LDW) and data warehousing market and the emerging themes such as data fabric and lakehouse. This expansion reflects the tools' adaptability and recognition of the multifaceted data challenges organizations face.

- Tools diverge into various areas and their capabilities overlap, which confuses customers during solution comparison. Each tool caters to a distinct set of capabilities, making direct comparison challenging.

- DV enables data access without data movement in modern, distributed, hybrid and multicloud architectures. DV technologies will transform how you manage, govern, integrate and consume distributed data assets.

Recommendations

As a data and analytics technical professional concerned with assessing DV solutions, you should:

- Assess the current data ecosystem before adopting a DV tool. Identify gaps that DV can bridge without redundancy.

- Align DV implementation with your data and analytics needs. Prioritize vendors that align with your existing architecture and long-term goals. Do not reverse engineer your systems based on the functionality of vendor tools.
- Evaluate the tools based on technical and nontechnical factors like query performance optimization (such as pushdown processing and massively parallel processing [MPP] support), data preparation capabilities, metadata and governance capabilities, and partnership with other vendors and licensing models.

Comparison

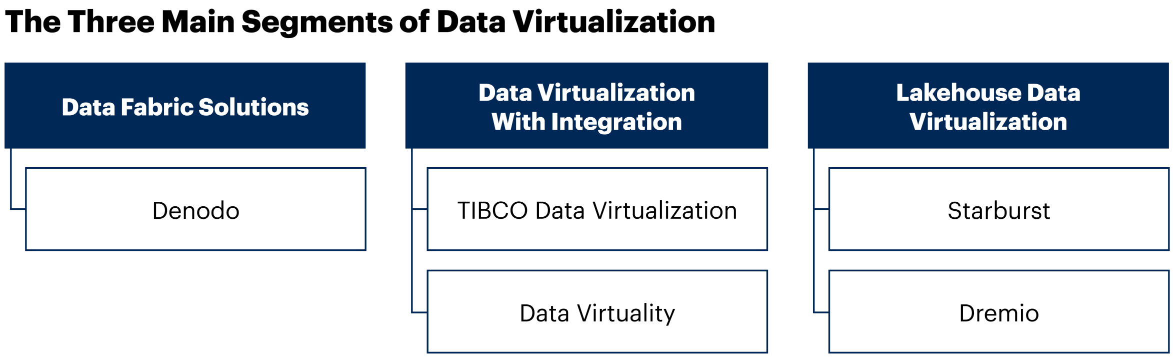
The DV market faced challenges in lack of differentiation until data virtualization solutions began expanding their focus beyond traditional realms. Solely offering data virtualization solutions limited their reach and impact. However, by venturing into areas such as data fabric, different data integration styles, and lakehouse architectures, vendors seized opportunities to address broader data management needs. This strategic shift revitalized the market and empowered vendors to offer comprehensive solutions that resonate with diverse customer requirements, fostering innovation and growth within the DV landscape.

The research focuses on three major segments in DV and the representative vendor solutions that are concentrated heavily on that market segment.

Although cloud service providers and other data management vendors offer similar capabilities through additional services, we limited this research to vendors focused exclusively on the DV market.

The vendors for this comparative assessment are a relative subset of the overall data virtualization market, and this list is selected based on the comprehensive set of criteria identified for the assessment.

Figure 1: The Three Main Segments of Data Virtualization


The Three Main Segments of Data Virtualization
Data Fabric Solutions
Denodo
Data Virtualization With Integration
TIBCO Data Virtualization
Data Virtuality
Lakehouse Data Virtualization
Starburst
Dremio

Source: Gartner
799330_C

DV Solutions Operating in the Data Fabric Space

Data fabric is an evolving data management design; there is no single data fabric solution in the market. However, certain tools can assist in building a data fabric design by providing some technology pillars that focus heavily on metadata utilization.

> **Gartner defines data fabric as:**
>
> *A data fabric is a data management design concept for attaining flexible, reusable and augmented data pipelines and services in support of various operational and analytics use cases. Data fabrics support a combination of different data integration styles, and utilize active metadata, knowledge graphs, semantics and machine learning to augment data integration design and delivery (see Figure 1).*

Denodo offers Denodo Platform, which focuses heavily on the data fabric space. The platform reflects Denodo's longtime focus and strength in DV and fabric.

Data Virtualization Solutions With Other Integration Styles

> **Gartner defines data integration as:**
>
> *The discipline comprising the architectural patterns, methodologies and tools that allow organizations to achieve consistent access and delivery of data across a broad spectrum of data sources and data types to meet the data consumption requirements of business applications and end users. Data integration tools enable organizations to access, integrate, transform, process and move data spanning various endpoints and across any infrastructure to support their data integration use cases.*

This category involves vendors such as TIBCO Software and Data Virtuality. TIBCO offers DV and other integration methods. But those are different products under the same portfolio. The Data Virtuality Platform, however, is a single product that offers different data integration methods such as extraction, transformation and loading (ETL) and change data capture (CDC) along with DV.

This category is recommended for those seeking a DV product that can also act as a data integration solution.

DV Tools With Data Lakehouse Capabilities

These are primarily lakehouse vendors that also offer DV capabilities, such as Dremio and Starburst. These tools are primarily adopted by customers who have a **clear strategy of building a lakehouse for their data and analytics needs.** The data sources with such tools can span beyond the regular lakehouse, but the major performance boost comes if you have a lakehouse-building strategy. This is because these tools are performance optimizers *and* virtualizers. These tools have gained a lot of popularity in the last couple of years. If implemented with proper strategy and lakehouse optimizations, they can provide good query performance.

Lakehouse engines are the technological solutions or platforms that establish the foundational infrastructure for enacting lakehouse architecture.

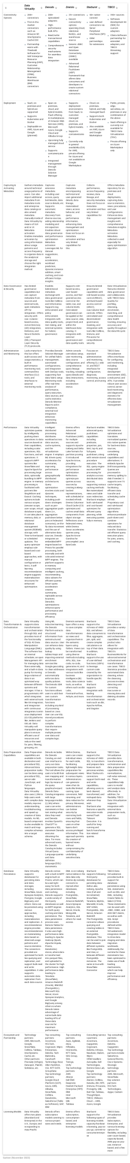
> **Gartner defines lakehouse as:**
>
> *A lakehouse is a converged infrastructure environment that combines the semantic flexibility of a data lake with the production optimization and delivery of a data warehouse. It supports the full progression of data from its raw, unrefined state, through the steps of refining it, to ultimately deliver optimized data for consumption.*

Key Differentiations in the Criteria Across the Vendors

Table 1 provides the key differentiations between the cloud providers that customers can use while assessing a DV tool. For a more in-depth assessment with drilled-down criteria, please download the attached Excel file.

Table 1: Key Differentiations Between DV Tool Cloud Providers

(Enlarged table in Appendix)



Analysis

Emerging in the early 2000s, data virtualization arose as a solution to the mounting complexities of data management. It drew inspiration from the concepts of federated databases and distributed query processing, both of which aimed to provide a unified method for accessing varied data sources without the need for data duplication or intricate integrations. The term “data virtualization” was coined to encapsulate the idea of creating a virtual layer over an organization’s disparate data sources, allowing for seamless and unified data access.

> **Gartner defines data virtualization technology as:**
>
> *Data virtualization technology is based on the execution of distributed data management (managed across multiple, geographically dispersed systems or locations). It is used primarily for querying against multiple heterogeneous data sources and federating the query results into virtual views. These virtual views are then consumed by applications, query/reporting tools, message-oriented middleware or other data management infrastructure components.*

Data Virtuality

Introduction

Data Virtuality, headquartered in Leipzig, Germany, is one of the emerging players in the DV market. It offers the Data Virtuality Platform and Pipes as its data integration tools. Data Virtuality has a practical approach toward data fabric as it realizes that DV cannot be used in every situation. Its platform creates a good balance between virtualization and persistence.

A Brief History and Recent Developments

Founded in 2012, Data Virtuality received its last round of funding in 2016.

Data Virtuality in 2023 launched SaaS-based DV software: Data Virtuality Platform SaaS. This provides a fully managed hardware and software stack in the cloud. This allows data architects, data developers and analysts to concentrate on harnessing the full potential of their data environment without worrying about infrastructure, maintenance, patching, or updates.

Data Virtuality has made the following recent improvements:

- Launched Data Virtuality Platform SaaS on January 31, 2023

- New business-friendly web UI for on-premises and SaaS customers
- Integrated data catalog (Business Data Shop)

Standout Features

- **Data Virtuality Platform Unites Data Virtualization and ELT/ETL Processes:** Data Virtuality Platform is a comprehensive data management solution that combines the benefits of DV with extraction, loading, transformation/extraction, transformation, loading (ELT/ETL) processes. DV is used for initial proof of concept (POC) for data and ad hoc querying, and ETL/ELT is used for scaling the queries created in DV in a production environment. Users can retrieve data in near real time from different data sources and assemble different data models and views using a single, unified query language.
- **Robust SQL Capabilities and Integration Features:** Data Virtuality supports data transformation and orchestration through SQL and provides nearly all of the functionality of SQL-92 DML with support for various features such as filtering, aggregating, joining, splitting, and performing calculations on data. The software also has a powerful internal scheduler, an open Management API for managing data flows externally, and a built-in data pump for moving large volumes of data in an optimized way between different systems and storages. It has a programmatic API for every aspect of the software, allowing integration with external tools, Git integration, and integration with CI/CD platforms like Jenkins and GitLab. Data Virtuality will manipulate query plans to ensure efficient use of resources and reduce the need for joins, filtering, grouping, etc.
- **Focusing on Data Storage and Replication:** Data Virtuality supports databases for persisting data and data storages, including Snowflake, Azure Synapse Analytics, Postgres, Google BigQuery, and others. Data can be persisted using various approaches, including materialization and replication. Data Virtuality also features a recommendation engine that provides recommendations on materializing data based on observed usage patterns and source statistics. The connectors provided by Data Virtuality are optimized for both live querying and replication, and support bulk read and write capabilities. It also supports automatic data replication from each data source.

- **Flexible Deployment Choices:** Different deployment options are available for the Data Virtuality Platform. It can be deployed as a SaaS solution in the cloud, on-premises within a local infrastructure, or as a hybrid model that combines both cloud and on-premises environments. This flexibility enables organizations to choose the deployment approach that best aligns with their infrastructure management, compliance, and data utilization needs.
- **Intuitive UI and Catalog:** The platform offers a new business-friendly web UI for on-premises customers and an integrated metadata repository called Data Shop. It is recommended to use a specialized data catalog along with the Business Data Shop.
- **Diverse Data Delivery Support:** With more than 200 connectors (Enterprise version), Data Virtuality's primary focus has been on supporting different data delivery styles. It supports DV, ETL, ELT, streaming and CDC, but not for all data sources.

Personas and Interfaces

Data Virtuality provides various interfaces for different personas like Business Data Shop for business users, Data Virtuality Platform SaaS for data teams and citizen data experts, and DV Studio for data engineers/architects.

- **Data Virtuality Platform SaaS:** Provides fully managed hardware and software stack in the cloud; users can connect and use it right away without having to worry about infrastructure, maintenance, patching, or updates. This is basically for data professionals and citizen integrators.
- **Business Data Shop:** This is a self-service web interface with features like metadata catalog and search, self-service data access for business users, write-and-run queries, download data, etc. This is basically designed for business users.

- **Data Virtuality Studio:** Available both as a web interface and thick client, this is for data engineers/architects. Here, clients can connect their data source and manage data in SQL. It offers features like:
    - SQL editor code completion on column level
    - Metadata dependency viewer (data lineage)
    - Metadata catalog and search
    - Graphical view builder
    - Wizards for easily connecting generic data using the formats XML, JSON, CSV, xSV

Data Virtuality provides focused training courses for all the types of personas supported in its product. These trainings will help the user understand and get acquainted with the Data Virtuality UI faster.

Those interfaces are used by different personas as follows:

- **Business Users:** Business Data Shop
- **Data Teams and Citizen Data Experts:** Data Virtuality Platform SaaS
- **Data Engineers and Architects:** Data Virtuality Studio

Pricing Models

- It comes with two plans — Standard and Enterprise.
- Both models are currently available in the U.S. and Europe market but are slowly expanding to Asia/Pacific.
- Data Virtuality offers monthly and yearly Standard and Enterprise subscriptions.
    - Standard subscription: data architects and engineers who want to centrally access, model, and govern organizational data in a single interface
    - Enterprise subscription with custom pricing plan
- Licensing: Depending on the pricing model, the license is available as software on-premises or SaaS.

- If installed on-premises, the package includes Data Virtuality Server, Data Virtuality Studio and/or web interface, Data Virtuality JDBC Driver, Data Virtuality ODBC Driver, access to various connectors, and documentation.

- If SaaS, the license includes a web interface to Data Virtuality Platform, access to connectors, and documentation.

After-Sale Support

Data Virtuality provides Standard and Enterprise support to its customers on different levels. The company also offers various types of training, both remote and on-premises, for customer teams based on different roles such as data users, data developers, administration, infrastructure, and security.

For SaaS customers, in-product support is available over web chat. For new customers, a dedicated solution engineer provides onboarding support for the first 90 days. This onboarding support is available for both Standard and Enterprise subscription types. Technical support is available from Monday to Friday between 4 a.m. to 10 p.m. CET (Berlin) and between 10 p.m. and 4 p.m. ET (New York) for both subscription types.

Strengths

- **Portfolio of connectors:** Data Virtuality stands out on data source connectivity regarding the number of ready-to-use first-to-market connectors. It also creates custom connectors upon request (after evaluating the business impact).

- **Agile data materialization:** Data Virtuality's automated ETL capabilities allow for seamless materialization, offering scalable solutions coupled with semantic, business-friendly data element naming and modeling. The platform ensures optimal performance while enabling the convenient facilitation of data historization.

Weaknesses

- **Upgrade Process Challenges:** Although Data Virtuality has overall positive feedback from customers, based on the Peer Insights data, customers continue to have challenges and difficulties with the upgrade process, which is deemed not that structured.

- **Limited Cataloging Capabilities:** Data Virtuality doesn't have a mature catalog. Unlike some competitors, it has a metadata repository but works best if used with a stand-alone data catalog.

Denodo

Introduction

Denodo, a company headquartered in Palo Alto, California, U.S. is widely recognized as one of the pioneers in the field of DV. Denodo's DV integrates all enterprise data from multiple systems, controls the unified data for centralized security and governance, and makes it available in real-time to business users, which helps them solve critical problems.

A Brief History and Recent Developments

Denodo was founded in 1999 and launched its first product, Denodo Platform v1, in 2002. It debuted in Gartner Magic Quadrants in the 2015 Data Integration Tools Magic Quadrant. Denodo announced in 2020 Denodo Platform 8.0, which is the latest major version. This version has a focus on hybrid/multicloud integration, a complete transition to web-based user interface, and advanced query acceleration techniques. It also includes support for Open Data Protocol (OData) 4 and GraphQL, thereby strengthening its capabilities for API creation.

In 2021, Denodo added a recommendation engine that learns from usage and automatically provides recommendations for performance acceleration and data exploration. In the same year they also added a new Data Catalog designed for self-service and data exploration across the entire data ecosystem, targeting business users, analysts and data scientists.

Last year, Denodo introduced a new security policy engine for tag-based security policies and attribute-based access control across data sources.

Denodo has its own MPP engine based on Presto, a high-performance, distributed SQL query engine for big data.

Standout Features

- **Strategic Data Integration:** Denodo adeptly realizes a logical approach to data integration and management that supports data fabric, data mesh and data hub architectures. Its chief strength is its comprehensive variety of features.

- **Decent Data Source Connectivity and Optimization Capabilities:** Denodo's platform can connect to a large range of data sources across traditional DB, cloud data stores, data lakes, cubes and files. It has a rich set of native connectors along with specific mapping for functions and operators (syntax and performance optimizations for each data source). It also offers data-source-native APIs that can be used for bulk loading, column descriptions, and statistics and indexes for the optimizers.
- **Empowering Business Data Access and Governance Through Denodo's Unified Semantic Layer:** Denodo offers a common semantic layer to easily expose data to users. Features include data discovery, search, classification and security and governance, using advanced features such as tags, endorsements, comments, activity usage, and more. The semantic layer security policies enable global access control management (masking, encryption, data restrictions, etc.) using security classifications, glossary terms and tags in security policies.
- **Intelligent Query Enhancement:** This focuses on AI/machine learning (ML) with features like a query optimizer, dynamic caching and an AI-powered query acceleration engine.
- **Cloud Automation:** The automation capabilities also extend toward automated cloud management for rapid deployment of Denodo Platform on Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP) by simplifying DevOps including launching cloud instances and clusters based on images and configuring security, load-balancing and auto-scaling parameters.
- **Business-Friendly Data Catalog:** A data catalog with a business-friendly way to find, access, curate and put trusted data to work with full lineage, metadata and governance. The data catalog also offers automatic recommendations based on usage and integration with an LLM (ChatGPT) for facilitating building queries for business users.
- **Virtual Query Language (VQL):** Denodo supports transformations in the data integration layer via functions within the Denodo Virtual Query Language (VQL). VQL is used to query and construct derived views against the underlying data. It is a superset of SQL-99 with a set of extensions that lets users handle data sources, hierarchical elements, and heterogeneous sources' query capabilities.

- **Cloud and Hybrid Deployments:** Denodo is designed to work in various deployment scenarios, including on-premises, cloud and hybrid environments, providing flexibility to adapt to changing infrastructure needs.

Personas and Interfaces

Denodo includes user interfaces for different personas. The Denodo portal integrates access to all different components of the Denodo Platform (web-based tools are integrated with single sign-on [SSO], so there's no need for extra authentication in any module). The modules are:

- **Design Studio:** for data integration, data transformation, data quality, data pipelining (design), data security, etc. The web-based Design Studio is also available as a graphical Java client.
- **Denodo Catalog:** a web-based interface for business users for data discovery and data preparation.
- **Denodo Scheduler:** a web-based tool for batch operations (periodical data exports, cache refresh, content indexing, etc.).
- **Denodo Solution Manager:** a web-based tool for managing complex deployments.
- **Denodo Diagnostic & Monitoring Tool:** a web-based tool for monitoring multiple environments.

Those interfaces are used by different personas as follows:

- Business Users: Data Catalog
- Data Architects, Data Engineers, Citizen Integrators, Developers: Data Catalog, Design Studio
- Data Scientists: Apache Zeppelin for Denodo notebook
- Data Stewards, DBAs: Design Studio, Data Catalog
- System Admins: Solution Manager, Diagnostic and Monitoring Tool
- Developers: Eclipse plugin (Denodo4E)

Licensing Models

Denodo offers multiple subscription options, including:

- Denodo Professional for small departmental projects.
- Denodo Standard, which is best suited for multiple use cases at the departmental level.
- Denodo Enterprise for enterprisewide deployment for multiple use cases and groups with large data volumes. It has features like Smart Query Acceleration and Data Catalog, which are not present in Professional and Standard.
- Denodo Enterprise Plus, which offers further automation, collaboration, global security policies, and parallel processing for enterprisewide deployments. This is the most expensive offering; it exclusively offers AI/ML capabilities, advanced semantics and embedded MPP.

There are two types of licensing options for Denodo Cloud:

- Infrastructure as a Service (IaaS)/Bring Your Own License (BYOL): The customer has to do the installation by leveraging cloud infrastructure services.
- Cloud Marketplace: This a lower entry point, reduced risk, and utility-based hourly pricing options (pay as you go).

After-Sale Support

Denodo offers comprehensive postsales support through Education, Support, and Professional Services. The customer success team provides tailored plans for successful implementation. The Denodo Community provides resources for learning and implementation, including a free Denodo Express version.

Strengths

- A broad set of features: Denodo offers versatile capabilities beyond DV, making it suitable for logical data warehousing and data fabric scenarios, including building user-friendly semantic models.
- Progressive and refined optimizer: Utilizes statistics and ML-driven DataOps to optimize query patterns, resulting in improved performance, quicker response times, and efficient resource utilization.

- Try-before-you-buy option: Denodo's presales approach and free product, Denodo Express, engage customers effectively, with around 80% of paying customers using the free product first, enabling tailored pricing based on core requirements.

Weaknesses

- DV tools are evolving: Although Denodo has been perceived as the leader in the DV space for a long time, the DV technology market has changed. Other DV vendors that operate primarily in the lakehouse realm also offer great performance and competing capabilities in the DV space.

- Pricing options and complexity: Denodo's rich feature set increased pricing. While flexible pricing models exist, using advanced capabilities often leads to high costs. Organizations considering pricier tiers should assess the required capabilities to costs based on their specific needs.

Dremio

Introduction

Headquartered in Santa Clara, California, U.S., Dremio offers an open data lakehouse, enabling self-service SQL analytics and combining data warehouse capabilities with data lake flexibility. It leverages Apache Arrow to eliminate data copies and persistent storage, granting access to diverse data sources, open and proprietary, with optimizations like storage indexes, materialized views, predictive pipelining, and bytecode generation.

Dremio's platform encompasses two primary offerings: Dremio Sonar and Dremio Arctic. Sonar provides a query engine for data lake along with unified analytics and self-service capabilities. Arctic is a data lakehouse management service that features data-as-code functionality, automatic maintenance, and open standards and technology. Dremio's query engine is also proficient in analyzing and seamlessly merging data from diverse nonlake origins, encompassing relational databases and NoSQL repositories.

Brief History and Recent Developments

Dremio was founded in 2015. In the same year, it raised its first round of funding valuing of $15M. Its latest funding round was a Series E round on January 25, 2022, which raised $160M. Some of the key investors in Dremio are Insight Partners, Lightspeed and Cisco.

In 2017, Dremio introduced its Dremio Self-Service Data Platform which aims to reduce or eliminate the need for traditional ETL, data warehouses, cubes, and aggregation tables, along with the associated infrastructure, data copies, and effort these systems require.

Developments over the past three years:

- In 2022, Dremio unveiled Dremio Cloud, a fully managed cloud version currently available on AWS and Azure (in preview). DML support for Apache Iceberg was also announced, allowing users to update data directly in their data lakes via standard SQL DML commands.
- In 2023, Dremio launched Dremio Arctic, a metadata and data management service for Apache Iceberg. Dremio Arctic provides a Git-like experience, allowing users to manage data as code for data versioning.
- In 2023, Dremio integrated Generative AI, thereby enabling users to directly convert text to SQL commands.
- In 2023, Dremio launched significant enhancements to its query acceleration technology, Dremio Reflections, designed to make it easier to use and manage. The update includes recommendations and optimized refreshes.

Standout Features

- **Enhanced Data Query Speed:** Dremio utilizes a query acceleration technology called Reflections, which provides a finely tuned materialization of a table or view that can be used to partially or wholly accelerate data processing. This leads to considerably faster data queries compared to traditional data warehouses.
- **No Vendor Lock-In:** One of Dremio's features is its nonrestrictive approach, ensuring no vendor lock-in. Dremio does not store any data, analyzing an organization's data where it resides, whether in the cloud or on-premises. Data remains open, by default, persisted in Apache Iceberg tables backed by Apache Parquet files in the company-owned Amazon Simple Storage Service (Amazon S3) bucket.
- **Unified View of Data:** This enables users to connect to all different data sources and expose the data through a unified business-friendly semantic layer that improves data discovery, ensures consistent reporting, and enables governed self-service data access.

- **Tailored for SQL Proficiency:** Dremio empowers users to seamlessly query data using any SQL client, including popular BI tools and data science notebooks, and construct data integration and consumption workflows exclusively through SQL commands. This approach simplifies the data analysis process, allowing users with SQL expertise to efficiently harness the platform's capabilities for seamless data exploration and utilization within a familiar SQL environment.
- **Contemporary and User-Friendly Interface:** Dremio offers a modern, intuitive user interface that empowers users to effortlessly generate novel data views, compute new columns, perform dataset comparisons, and enhance data organization by adding descriptions and tags to datasets.
- **Fast Response Time:** Dremio's query engine delivers subsecond performance through an innovative optimization technology that merges Apache Arrow and Reflections. This patent-pending approach ensures that queries are processed rapidly, providing users with near-instantaneous results and enabling swift decision making based on real-time insights.
- **Robust Concurrency Support:** Dremio achieves high concurrency through its MPP architecture, Workload Management (WLM) and its autoscaling multiengine architecture. This combination empowers the platform to efficiently handle multiple query patterns concurrently, allowing users to seamlessly and simultaneously execute diverse analytical tasks.

Personas and Interfaces

Dremio's user interface provides rich functionality for a wide range of personas and enables API integrations for those who prefer CLI or event-driven integrations. Their native user interface includes a variety of options from code to low-code and no-code. Users can interact with the data in several ways, with a variety of features within each:

- Via SQL Runner using point-and-click capabilities
- With SQL syntax generative AI in the semantic layer to autogenerate SQL from text
- Through one-click integrations to BI tools of choice with all query accelerations and telemetry provided by Dremio without the need for extracts

Strengths

- **Deliver Performance at Scale:** Dremio uses various techniques to provide speedy performance, including Reflections query acceleration, Apache Arrow, query optimization, and parallel processing.
- **Integrated Built-In Semantic Layer:** Dremio provides an integrated semantic layer with shared business logic, governed data access, and security, providing a single, consistent view of the data for all teams and tools.
- **Built on Open Standards:** Dremio uses open standards and is designed to be portable so that organizations can easily move their data and workloads to a different platform if they choose to do so.

Weaknesses

- **Challenges for Nontechnical Users:** Dremio is designed to be a self-service analytics platform, but some nontechnical users may need help with its features and interface. Dremio is a powerful platform with many features, some of which may be necessary only for technical users, so it may not be intuitive for nontechnical users.
- **Limited Data Source Support:** Although Dremio supports a wide variety of data sources, some less common or niche data sources are not included or fully compatible.

Licensing Model

- Dremio Cloud standard license comes with a fully managed lakehouse platform and a lot of features like built-in governance and lineage and global control pane.
- Dremio Cloud enterprise comes with a pay-as-you-go model. It comes with all standard features and also offers enterprise identity provider integration, data masking, SCIM synchronization, and has support options available.
- Dremio Software has a free community edition and is a self-managed software that users can deploy and run on their cloud or on-premises infrastructure.
- Dremio Software Enterprise edition comes with extra features.

After-Sale Support

Dremio provides multiple support options, starting with customer implementations, upgrades, overall platform and use case health and performance optimization, and overall customer success. Expertise is offered in-house thus eliminating the need for any third-party support contracts. Dremio's staff is prepared to assist with customer implementations, upgrades, platform health, and performance optimization. If customers opt for third-party deployment, Dremio can collaborate to tailor a suitable support approach.

Starburst

Introduction

Starburst, headquartered in Boston, is one of the prominent vendors in the DV and lakehouse space. Founded in 2017, Starburst takes a modern stance to solve the issue of data silos and slow access speed by harnessing Trino, a distributed analytics engine. Trino, formerly recognized as PrestoSQL, has its roots in the original Presto project initiated at Meta (then called Facebook). In 2020, this project was rebranded to Trino.

Brief History and Recent Developments

- Starburst got its Series A funding in 2019 and raised $250 million in a Series D funding round in 2022.
- In February 2021, Starburst introduced Galaxy, a distributed analytics platform. The fully automated, multicloud platform enables the fastest path to analytics across data silos to inform better business decisions.
- In June 2021, Starburst announced Stargate, a gateway for global cross-cloud analytics. The new technology is purpose-built to enable analytics across borders while ensuring data sovereignty and privacy regulations are met.

Starburst Galaxy is a fully managed data lake analytics platform that enables clients to access data from various sources. Starburst Stargate is a gateway that enables Starburst Enterprise customers to link catalogs and data sources supported by one Starburst cluster to other catalogs and data sources in remote Starburst clusters.

- In April 2023, Starburst announced a new integration with dbt Cloud to unlock cross-platform data transformations.

- In June 2023, Starburst unveiled its data lake analytics platform vision by introducing a series of capabilities in Starburst Galaxy. Starburst is empowering teams to modernize their data lake strategy with Starburst Gravity, which provides a single point of access for teams to discover, govern and analyze the data in and around their data lakes.
- In June 2022, Starburst acquired Varada to deliver the new standard of data lake analytics. Varada's technology uses patented techniques to accelerate data queries in data lakes via autonomous indexing and caching.

Starburst Galaxy and Enterprise

Starburst comes with two options

- **Starburst Galaxy:** Starburst Galaxy is the cloud-native and fully managed service, highly performant Trino SQL query engine. It includes a user interface that allows users to query a variety of data sources or join data across multiple data sources through a single query. Galaxy handles complex procedures, including deployment and upgrade, so the user doesn't have to. With Starburst Galaxy, users can use SQL and the business intelligence or reporting tool of their choice to analyze the data and gain insights quickly. It is recommended for organizations that don't yet have the resources to manage complex data clusters.
- **Starburst Enterprise platform:** Starburst Enterprise platform (SEP) offers third-party integrations, improves performance, and provides a platform to deploy, configure, and manage clusters. It is available in Amazon, Google, and Azure ecosystems as marketplace offerings or as direct Helm chart deployments in their respective Kubernetes services. Users can also run SEP on bare-metal or virtual machines. SEP is recommended for the following scenarios:
  - Workloads that require connectors not currently supported by Starburst Galaxy
  - Self-managed deployment on a customer's own infrastructure
  - Specific security requirements regarding authentication and authorization to SEP and the connected data sources

Standout Features

- **Starburst Boosts Trino's Performance:** Starburst has undertaken a series of improvements to the open-source Trino, resulting in significant advancements in query performance. Primarily, Starburst introduces enhanced connectors that expose, generate, and leverage statistics, enabling the query engine to devise swifter and more efficient query plans. These connectors also facilitate advanced query processing pushdown (e.g., predicate pushdown, projection pushdown, join pushdown, limit pushdown, top-N pushdown, full query pushdown, partial query pushdown, and remote stored procedure execution) to connected data sources, enhancing query performance. Furthermore, dynamic filtering optimizations substantially curtail network traffic and data source load. The introduction of Starburst Cached Views brings a suite of performance enhancements, offering various avenues to access frequently utilized data. Warp Speed typically accelerates data lake queries by 4-7x.

- **Enhanced Security Capabilities:** Starburst Enterprise integrates an array of security capabilities, including end-to-end encryption, various authentication options, granular access control, detailed security auditing, and more. This empowers organizations to seamlessly connect with a centralized security framework, enabling them to manage fine-grained access control across enterprise data lakes and other data sources.

Additionally, in conjunction with the platform's preconfigured security features, users can leverage the platform's robust integrations with premier security partners

- **Universal Search:** Starburst Galaxy's universal search lets users locate their data entities in Galaxy. It can be used to locate any data object by name or metadata name anywhere in the user's Starburst Galaxy account.

- **Data Products:** Starburst Galaxy and SEP data products provide a collection of curated, high-quality related datasets and relevant metadata. Data products increase the discoverability of users' data and are assigned to the data domains that define their business.

- **Starburst Stargate:** Starburst Stargate is a connector that lets users link a local catalog on their Starburst Enterprise cluster to a catalog on a remote Starburst Enterprise cluster with the same version. Starburst Stargate uses a Java Database Connectivity (Sun) (JDBC) connection to the remote cluster.

Some key features include:

- **Java CLI:** Connect to Starburst, issue queries and receive results from the console.
- **Trino REST Client:** Connect to Starburst, issue queries, and receive results programmatically.
- **PyStarburst:** Python client that provides a similar syntax to PySpark and Snowflake Snowpark for writing and running production-grade ETL pipelines and data transformations. PyStarburst makes it easy to not only build new pipelines with PyStarburst but also to migrate existing PySpark and Snowpark pipelines to Starburst without rewriting code.
- **Universal Search:** Helps locate any data object by name or metadata name anywhere. Clients can search for the names of catalogs, schemas, tables, views, columns, data products, tags, owners, or contacts.
- **Query Editor:** Offers a full environment to write and execute SQL statements in Starburst. Clients can browse and access all their configured data sources from here.
- **Data Product Explorer:** Allows users to view and use curated datasets from their organization on the data products pane in Starburst.
- **Catalog Explorer:** Information about the currently selected table or view:
    - Metadata
    - Columns
    - Metrics
    - Definition
    - Data preview
    - Query history
    - Audit log
    - Privileges
- **Administration:** Access the audit log, manage cluster availability, see usage and billing information, specify notification settings.

- **REST-based Administration:** Programmatically manages a Starburst cluster:
  - Create and manage data products.
  - Start, stop, create, modify, and destroy clusters.
  - Manage users, roles, policies, masks & filters, metadata, tags.

- **Access Control:** Allows you to manage the following data governance aspects of Starburst:
  - Users
  - Roles and privileges
  - SSO groups
  - Single sign-on
  - API authentication tokens
  - Service accounts
  - Tags
  - Row filters
  - Column masks

- **Partner Connect:** A compilation of BI and data integration tools from Starburst partners that connect directly to Starburst Galaxy clusters.

Personas and Interfaces

The Starburst Enterprise platform (SEP) web UI is the default web interface when the user accesses an SEP cluster that has a valid license configured.

The following features can be accessed with this SEP web user interface:

- **Query Editor and Saved Queries:** The query editor offers a full environment to write and execute SQL statements in Starburst. Clients can browse and access all your configured data sources from here.

- **Data Product:** Create, view and use curated datasets from your organization on the data products pane in Starburst.
- **Insights:** Shows metrics about queries, CPU usage, top users, etc.
- **Built-In Access Controls:** The built-in role-based access control makes it easy to configure any user's correct access rights to catalogs, individual schemas, and tables thereby makes security configuration better.

Starburst supports a range of personas, as detailed below. These personas represent a set of focused workflows and assume a nominal skill set.

- **Data Consumer:** CLI, Query Editor, Universal Search, Catalog & Data Product Explorer
- **Data Engineer:** CLI, PyStarburst, Universal Search, Catalog & Data Product Explorer, plus Partner Connect
- **Platform Administrator:** All of the above

Strengths

- **Impressive Performance**: Starburst is a data lakehouse platform with remarkable performance, offering subsecond query response times across a diverse spectrum of analytical inquiries.
- **Open-Source Advantage:** Starburst embraces an open-source model with extensive community support.
- **Key Player in Data Mesh:** In 2019, Starburst has built a community committed to advancing data mesh adoption. With DV, data governance and data sharing capabilities, Starburst offers extensive data mesh capabilities and has been a leading advocate for data mesh.

Weaknesses

- **Navigating Through Complexity:** Starburst's powerful capabilities can make it a challenge for users unfamiliar with distributed SQL query engines or data lakehouse concepts.
- **Missing Capabilities:** Starburst lacks embedded capabilities for advanced text mining, fuzzy matching, and advanced AI integration. Starburst has a prototype for ChatGPT integration and is actively working on fulfilling the missing capabilities.

- **Limited Data Source Support:** In our research, we came across more than 50 data sources for Starburst, which should cover most modern systems but is lower than other DV tools.

Licensing Model

- **Starburst Enterprise License:** Includes support and enables numerous features
- **Starburst Enterprise Elite License:** Unlocks premium features, such as Starburst Stargate and Starburst Warp Speed, in addition to the support and features included in the Starburst Enterprise license.

Pricing options include pay-as-you-go and commit to consume.

Pricing for Galaxy

- **Starter:** Perfect for basic data exploration and learning Starburst
- **Standard:** For launching to production with the option to scale
- **Premium:** For teams with more support and performance needs; comes with custom pricing

After-Sale Support

Starburst provides following services as a part of after-sale support:

Customer Success

- Success planning
- Onboarding project plan
- Scheduled progress tracking
- Adoption planning
- Quarterly business reviews
- Advisory services
- Value validation and optimization
- Continuous usage monitoring and optimization

- Roadmap sessions

Services

- Best-practice-driven deployment
- Use-case-driven methodology
- Accelerated onboarding experience
- Optimized architecture for scale
- Maximized query performance and efficiency
- ETL & Batch workload best practices
- Design for new use cases, data products
- Design and development of connectors
- Design and development of platform extensions

Enablement

- Self-paced and live training
- Certifications

Support

- 24/7 Expert Support
- Expert break/fix assistance
- World-class ticket response and resolution
- Timely product fixes

TIBCO

Introduction

- TIBCO Software established in 1997, is a prominent vendor that specializes in integration, analytics, and event-processing software. TIBCO Data Virtualization stands as a robust element within the broader TIBCO platform, centering on the facilitation of data integration and access. It empowers organizations to establish a cohesive, virtual representation of data from diverse sources, disregarding their location or format.

- TIBCO caters to a diverse array of industries, encompassing finance, healthcare, manufacturing, telecommunications, and beyond. Its client base spans across the globe, comprising prominent organizations of all sizes, from small-scale enterprises to large corporations.

A Brief History and Recent Developments

- The TIBCO Data Virtualization platform has roots in the early 2000s, with several mergers and acquisitions shaping its development. In 2002, TIBCO acquired DataDirect Technologies, which brought valuable DV technologies to the platform. This acquisition laid the foundation for the TIBCO Data Virtualization platform.

- In 2007, TIBCO expanded its capabilities by acquiring Spotfire, integrating its analytics platform with the TIBCO Data Virtualization platform. This integration enabled the use of DV for powering analytics applications.

- In 2017, TIBCO acquired the Cisco Data Virtualization business, incorporating additional data virtualization technologies, including the Cisco Integration Server. This acquisition expanded the platform's capabilities for accessing and managing data from various sources.

- In 2021, TIBCO acquired Information Builders, integrating their DV and analytics technologies, such as the Information Builders iWay product. This acquisition further enhanced the TIBCO data and analytics capabilities, enabling comprehensive data access, management, and the development of advanced analytics applications.

- In the last three years, TIBCO Data Virtualization has added:

    - MPP engine based on Apache Drill in 2019

    - Support for Snowflake, TIBCO Streaming, Microsoft SQL Server Integration Services (SSIS) adapter and OData V4 in 2020

    - New web UI for business users for self-service in 2021

    - Cloud-native containerized support

Standout Features

- The TIBCO platform, inclusive of Data Virtualization and other integral components, presents a comprehensive solution tailored for organizations keen on leveraging their data's potential. Seamlessly amalgamating data management, analytics, and event processing capabilities, TIBCO empowers businesses to make informed decisions driven by real-time data insights. The platform streamlines the harmonious integration of data across varied systems and applications, thereby enabling organizations to attain a comprehensive panorama of their operations and clientele.

- By delivering a unified layer for data using a Java middleware, TIBCO Data Virtualization facilitates prompt data accessibility, observability, and empowers users to delve into analysis and glean insights without the need for time-consuming data movement or replication. This approach significantly enhances agility, efficiency, and flexibility in data management, ultimately fostering improved business outcomes for organizations capitalizing on the robust TIBCO platform.

- From a product perspective, TIBCO has a rich set of components, including:
    - **Web UI**: Self-service data provisioning, catalog, and publishing for tailored views. Enables data manipulation, SQL queries, and secure workflows. User-friendly interface with drag-and-drop functionality. It is designed for citizen data engineers/business users.
    - **Developer Studio**: Agile data modeling and development tool for data-oriented developers, enabling modeling, transformations and resource management
    - **Business Directory:** Self-service data directory for searching and consuming IT-curated datasets, promoting sharing and reuse
    - **Discovery:** Advanced data exploration, entity identification, and relationship analysis across multiple sources. Accelerates meeting new business requirements
    - **Metadata Repository:** Centralized storage and management of metadata and data services life cycle
    - **Deployment Manager:** Streamlined migration of projects, including resources, cache settings, security profiles, across multiple instances. Simplifies and automates development life cycle
    - **Active Cluster:** Enhances availability and scalability with load balancers. Simplifies operations management through resource sharing and dynamic capacity adjustments
    - **Adapters**: Provide a wide range of data source connectivity for databases, files, big data, cloud sources, packaged applications. Integrate with data source optimizers to ensure more accurate queries and higher performance
    - **Cost- and Rule-Based Optimizers:** Work directly with data source optimizers to maximize query performance

Personas and Interfaces

TIBCO Data Virtualization provides a local integrated development environment (IDE) for power users like data scientists and data engineers to connect to various data sources, transform data, apply security and access controls, and build and publish virtualized views.

A lightweight, browser-based interface geared toward business users and citizen integrators is also available. This largely helps business users discover data using the data catalog and citizen integrators to create required flows for self-servicing of the data.

Additionally, a Deployment Manager interface is provided for a DevOps/DataOps persona to manage and seamlessly promote the artifacts between various TIBCO Data Virtualization environments.

Strengths

- **Complete Ecosystem:** TIBCO offers a wide ecosystem of data management solutions, integrating DV with other tools like integration, messaging, analytics, and event processing. This unified platform enables interoperability, empowering organizations with a cohesive data management approach.

- **Full Integration and Real-Time Access:** TIBCO Data Virtualization seamlessly integrates with diverse data sources, including databases, files, and cloud platforms. It provides virtualized access to data, facilitating agile decision making by quickly adapting to changing business requirements.

- **Data Governance and Scalability:** TIBCO Data Virtualization emphasizes data governance, ensuring data privacy, security, and compliance. It also scales effectively to handle large-scale deployments and high-performance demands, enabling efficient processing of complex data queries.

Weaknesses

- **Lack of a Built-In Recommendation Engine:** Unlike other DV products, TIBCO Data Virtualization does not provide a specific recommendation engine feature. Organizations requiring advanced recommendation capabilities may need to explore additional solutions or integrate third-party recommendation engines in addition to TIBCO Data Virtualization to meet their specific requirements.

- **Potential Integration Challenges and Changes:** TIBCO's history of multiple acquisitions and mergers while expanding its offerings may introduce challenges in terms of integration and transition. Organizations using TIBCO Data Virtualization may experience product direction, support, and compatibility changes due to these corporate activities that could impact their DV initiatives.

- **User Interface:** TIBCO's user interface's general look and feel needs refresh and update. The UI is not customizable, and the informational links can occasionally be slow to load.

Licensing Model

- Flexible Consumption Pricing Model
    - TIBCO BusinessWorks Container Edition and plug-ins for AWS Marketplace use a consumption-based pricing model.
    - Customers pay hourly charges based on the number of containers running per hour; usage is calculated and billed monthly.
    - Users incur associated AWS infrastructure charges depending on the services and infrastructure used; rates and fees are defined and controlled by AWS and can vary between regions.
    - Annual Premium Subscription: Users can contact TIBCO Sales to purchase Premium Subscription for TIBCO BusinessWorks Container Edition and plug-ins for AWS Marketplace.
- Subscription Model
    - Customers can subscribe to the numerous subscriptions available on TIBCO Cloud. This subscription gives subscribers role-based privileges.
- TIBCO Cloud Integration Subscriptions
    - Several subscription packages are available for TIBCO Cloud Integration, allowing access based on the subscription tier.
- TIBCO Cloud Passport Pricing
    - Subscription allows the customers to utilize TIBCO Cloud capabilities per their needs.
    - A single consumption model permitting scaling of the consumption: customers pay for peak demand capacity only when they use it, rather than continuously throughout the year.

After-Sale Support

TIBCO Data Virtualization provides an after-sale support plan that includes community pages, onboarding programs, product updates, social engagement, user groups, industry engagement, an idea portal, change advisory board (CAB) and user conferences, and an implementation partner program. The TIBCO product specialist team composed of proficient TIBCO technologists focuses on increasing customer adoption by offering comprehensive knowledge to aid the customer journey and influence product strategy. They collaborate with cross-technical teams to assist customers with removing technical blockers to accelerate product value and increase retention.

The after-sale support plan highlights include:

- Access to TIBCO Academy for product-learning resources
- Add-on professional services and customer success plans
- Access to the TIBCO Community forum for additional resources

Other DV Solutions

The following are some of the other DV solutions revealed during Gartner client inquiries and analysis that support both existing and new use cases.

Azure PolyBase

PolyBase was first introduced in the Microsoft Parallel Data Warehouse appliance in 2012 to query over Hadoop Distributed File System (HDFS), and then added to SQL Server 2016 to process Transact-SQL queries across SQL Server and external sources such as Hadoop and Azure Blob Storage.

PolyBase capability can be invoked by any BI tool, such as Microsoft Power BI, on top of SQL Server or Azure Synapse Analytics. PolyBase DV allows the data to stay in its original location and format. PolyBase allows virtualizing the external data through the SQL Server instance, so that it can be queried in place like any other table in SQL Server. This in turn minimizes the need for ETL processes for data movement. This DV scenario is possible with the use of PolyBase connectors.

PolyBase connects to external data sources like:

- Oracle
- MongoDB

- Teradata
- Generic ODBC
- Azure Storage
- Hadoop
- SQL Server,
- Amazon S3-compatible object storage

Amazon Redshift Spectrum

Amazon Redshift Spectrum operates on dedicated Amazon Redshift servers that are independent of the cluster. Amazon Redshift offloads many compute-intensive tasks, such as predicate filtering and aggregation, down to the Redshift Spectrum layer. As a result, Redshift Spectrum queries consume much less of a cluster's processing capacity than other queries. Additionally, Redshift Spectrum can scale efficiently by utilizing thousands of instances to leverage parallel processing, depending on your query workload.

Amazon Redshift Spectrum also allows users to efficiently query and retrieve structured and semistructured data from files in Amazon S3 without having to load the data into Amazon Redshift tables.

K2view

K2view offers the K2view Data Virtualization Tools as its DV capabilities. K2view ingests data from underlying sources, unifies it according to a semantic layer, transforms and enriches it, employs data masking, and ultimately makes it available to consuming applications and data analysts. The platform offers a capability called dynamic data virtualization, in which some data can be accessed from the source via virtualization and other data can be accessed from a Micro-Database, which is asynchronously synced with the sources. This ability to decide which data will be stored physically and which will be virtualized provides advantages in terms of performance optimization and load management on source systems.

IBM Data Virtualization

IBM Data Virtualization is a universal query engine that executes distributed and virtualized queries across databases, data warehouses, data lakes, and streaming data without additional manual changes, data movement, or replication. It is available through the data fabric capabilities of the IBM Cloud Pak for Data platform. IBM Data Virtualization is vendor agnostic and a key piece in modernizing data architecture to accelerate digital transformation.

It offers improved performance through intelligent caching and queries data wherever it resides thus minimizing movement and copies. IBM Data Virtualization can independently scale compute and storage with zero downtime and offers high flexibility along with the option of virtualization as a service or on-premises deployment. It also offers end-to-end data governance leveraging its cataloging technology and thus ensuring that queries are run on high-quality and valid data.

Alteryx Analytics Cloud Platform

Alteryx Analytics Cloud supports an insight-driven organization with unified solutions from automated data preparation and analytics, approachable machine learning, and AI-generated insights. It delivers the right analytic capability to each user and allows collaboration across teams to support end-to-end enterprise analytics solutions. Alteryx Analytics Cloud automates analytics at scale to enable intelligent decisions across the enterprise. It provides self-service for all by empowering teams with an easy, intuitive user experience allowing everyone to create analytic solutions that improve productivity, efficiency, and the bottom line. Alteryx Analytics Cloud is a unified cloud platform that helps to transform the data into insights with self-service data prep, ML, and AI-generated insights. It has the latest enterprise grade security standards and certifications, thus reducing risk and helps users to connect their data and applications with open API standards.

AtScale

AtScale enables smarter decision-making by accelerating the flow of data-driven insights. Its semantic layer platform simplifies, accelerates, and extends business intelligence and data science capabilities for enterprise customers across all industries. With AtScale, customers are empowered to democratize data, implement self-service BI, and build a more agile analytics infrastructure for better, more impactful decision making.

AtScale forms a semantic layer that stands independent from BI tools and cloud data platforms — centralizing governance and control while enabling decentralized analytics consumption and data product creation. The platform provides a modern, cloud-centric approach to deliver an online analytical processing (OLAP)-like experience. AtScale platform maintains live connection to data warehouse/lakehouse and helps in building interactive dashboards on different BI tools. It embraces data mesh/hub-and-spoke analytics frameworks.

Guidance

Technical professionals assessing a DV tool for data architectures should follow these recommendations:

- **Clearly Outline SLAs and Service-Level Objectives (SLOs) Before Choosing the Tool:** When evaluating a DV tool for data architectures, it is essential to begin by clearly defining your SLAs and SLOs before making a tool selection.
- **Evaluate Existing DBMS DV Services First:** If any of the organizational DBMS vendors provide a DV capability, evaluate them first. Extending a vendor relationship, rather than bringing in a new vendor, can be useful for maximizing existing investments. But if you're using multiple databases and need a single access/semantic layer on top of all of them, then you have to invest in a stand-alone DV product.
- **Familiarize Yourself With the Different Segments of the DV Market:** This research document covers three segments of the DV market; understand each of these and assess which makes sense for you.
- **Assess Criteria:** After understanding the segments, assess the criteria mentioned in the document and figure out the ones that are most important to you.
- **Leverage DV for Agile Development:** DV allows rapid prototyping. It also enables multiple modern BI tools, connected to the consumption layer of the DV tool, to access a common semantic layer. Without DV, IT teams that must satisfy new data requests face more repetitive and time-consuming processes. These processes involve either creating another data mart or building the model into the enterprise data warehouse — two approaches that are significantly less agile.

- **Become Familiar With DV Weaknesses and Prepare to Deploy Alternative Data Integration Tools:** Successful organizations are using a variety of data integration styles, not just data virtualization. DV can never scale to support the edge use cases requiring extreme query performance. It can't compete with a database appliance or purpose-built data warehouse.
- **Deploy Flexible, Open, Popular Platforms Ready for Your Future-State Architecture:** The DV tool or platform you select should be open enough to accommodate future architecture decisions. For example, it should be adaptable to data fabric architectures or composable enterprise/architectures that may exist on multiple clouds.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Assessing the Relevance of Data Virtualization in Modern Data Architectures

Extending LDW Principles Toward Data Fabric

Does My Organization Need a Data Lakehouse?

Data Engineering Essentials, Patterns and Best Practices

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's research organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner research may address legal and financial issues, Gartner does not provide legal or investment advice and its research should not be construed or used as such. Your access and use of this publication are governed by Gartner's Usage Policy. Gartner prides itself on its reputation for independence and objectivity. Its research is produced independently by its research organization without input or influence from any third party. For further information, see "Guiding Principles on Independence and Objectivity." Gartner research may not be used as input into or for the training or development of generative artificial intelligence, machine learning, algorithms, software, or related technologies.

Table 1: Key Differentiations Between DV Tool Cloud Providers

↓	**Data Virtuality** ↓	**Denodo** ↓	**Dremio** ↓	**Starburst** ↓	**TIBCO** ↓
Connectivity Options	▪ 200+ connectors ▪ First-in-the-market connectors such as Amazon SP, InterSystems Caché CDC ▪ Partnership exists with Theobald Software for SAP Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Business Warehouse (BW) connectors	▪ 200+ specialized adapters ▪ High-performance bulk APIs ▪ Read/write transactional capabilities ▪ Comprehensive security mechanisms for seamless data integration	▪ 25+ connectors ▪ Decent connectivity options spanning cloud, relational, NoSQL, and warehouse source ▪ Generic JDBC and Open Database Connectivity (ODBC) available ▪ Advanced Relational Pushdown (ARP) framework that allows data consumers and developers to create custom relational connectors	▪ 50+ native connect ▪ user-defined connectors via Serial Peripheral Interfaces (SPI) ▪ OEM partnerships for extensions	▪ 350+ sources ▪ Software development kit (SDK) for custom adapters ▪ CData Software partnership for extended adapters, and TIBCO Streaming support

↓	*Data Virtuality* ↓	*Denodo* ↓	*Dremio* ↓	*Starburst* ↓	*TIBCO* ↓
Deployment	■ SaaS, on-premises and hybrid are supported ■ Supports Kubernetes and Docker ■ Deployable on AWS, Azure and Google (planned)	■ Span on-premises, cloud, and hybrid models. PaaS offering in marketplaces of Amazon Web Services (AWS), Microsoft Azure, Google Cloud and Alibaba Group ■ Upcoming fully managed cloud service ■ Supports containerization ■ Integrated management through Denodo Solution Manager	■ Supports deployment environments from SaaS, to customer-managed cloud, on-premises, and hybrid models. ■ Supports containerization ■ Currently only in general availability (GA) for AWS, private offering for Azure and not available on Google Marketplace	■ Cloud, on-premises, hybrid, and fully managed SaaS ■ Supports both Kubernetes and Docker ■ Also available on AWS, Azure and Google Marketplace	■ Public, private, edge, serverless, on-premises or any combination of the above ■ SaaS offering is currently not available for TIBCO Data Virtualization (TDV). ■ Private offering on Azure Marketplace only

↓	**Data Virtuality** ↓	**Denodo** ↓	**Dremio** ↓	**Starburst** ↓	**TIBCO** ↓
Capturing & Activating Metadata	Gathers metadata around technical usage patterns of users. Can also import additional metadata from metadata tools and enterprise data catalogs or also amend metadata manually inside of Data Virtuality by means of GUI and/or UI. Metadata activation: Enables metadata activation by using information about usage patterns and access frequency to organize data in the analytical storage and choose the right integration method.	Captures extensive metadata including user access, query bottlenecks, data source details, etc. Supports automatic metadata discovery from diverse sources, aiding relationship modeling and introspection. Enables robust data governance and performance optimization Metadata activation: Leverages ML for dataset suggestions, query acceleration, workload prediction, dynamic instance addition, smart caching, and cost-efficient FinOps insights.	Captures metadata encompassing technical and user-generated details, data lineage, materialized views, access behavior, and query performance information, enhancing data management and optimization. Does not focus on metadata activation and has very limited capabilities for this.	Collects performance, access frequency, reviews, data lineage, and security metadata. Does not focus on metadata activation and has very limited capabilities for this.	Offers metadata repository for on-premises and cloud deployments, capturing data source statistics, views, procedures, and user configurations. Enhances data management and insights with comprehensive metadata insights. Metadata activation: Metadata collected is used especially for query optimization algorithm

↓	***Data Virtuality*** ↓	***Denodo*** ↓	***Dremio*** ↓	***Starburst*** ↓	***TIBCO*** ↓
Data Governance & Security	Has limited governance capabilities but can capture metadata from sources and external tools, open metadata API for integration. Offers security with row-/column-based access, Active Directory (AD) integration, Kerberos, and Secure Sockets Layer (SSL)/Transport Layer Security (TLS) protocols.	Enables bidirectional metadata exchange with data governance tools like Collibra and Informatica. Supports tag enforcement and policy enforcement. Integrates with data quality tools, text mining, and external systems for holistic governance.	Supports role-based access, user-based access, and fine-grained access control at the table, row, and column levels. Centralized governance can be applied at the data source, view, project level, and within the lakehouse data catalog. It integrates with external governance and data quality tools.	Ensures limited governance capabilities with role-based access control (RBAC), attribute-based access control (ABAC), row/column-level controls. Offers centralized and decentralized governance, masking, and integration with tools like Collibra, Immuta, and Privacera for comprehensive security	Data Virtualization features inherent data governance and collaborates with TIBCO Data Quality for profiling, cleansing, enrichment, matching, and comprehensive data health tracking, ensuring accuracy and quality throughout the data life cycle.

↓	***Data Virtuality*** ↓	***Denodo*** ↓	***Dremio*** ↓	***Starburst*** ↓	***TIBCO*** ↓
Administration and Monitoring	Lacks maturity in this but offers audit access and usage statistics, a built-in performance monitoring tool, command line interface (CLI) tools, and interfaces to external monitoring tools.	Provides Denodo Solution Manager for unified fabric management, automated configurations, and integration with DevOps tools, supporting fast data fabric setup. Monitoring tools offer real-time analysis of performance, query execution, data sources, and cache statistics. Denodo Monitor Reports aid in compliance; external tool integration enhances monitoring capabilities.	Admin console centralizes setup, user monitoring, data source configuration, and comprehensive query lineage tracking, revealing query details and data interactions, including acceleration status.	Focuses on standard administration and monitoring, offering CLI, web UI, and REST interfaces for configuration, diagnosis, monitoring, access control, and tuning	TIBCO Data Virtualization offers interfaces for configuring and monitoring, including integrated development environment (IDE), a web portal, and APIs. It provides robust user access control, server-level monitoring, and diagnostic statistics for effective data virtualization management.
Performance	Data Virtuality optimizes queries by intelligently pushing down operations to data sources based on their capabilities, including	Intelligent distributed processing distributes workload across data sources, cache systems, and the virtualization engine. It employs ML-	Dremio offers Advanced Relational Pushdown (ARP) for multiple sources and supports Apache Iceberg and Delta Lake formats for	Starburst enables advanced query pushdown capabilities including predicate, projection, join, limit, top-N, and full query	TIBCO Data Virtualization optimizes and translates normalized queries into native queries for back-end data sources,

	Data Virtuality	Denodo	Dremio	Starburst	TIBCO
	scalar operations, SQL functions, and set operations. It supports MPP systems like Snowflake and Apache Spark for processing large data, optionally bundling its MPP engine. In-memory processing is facilitated with partners like SingleStore and Exasol. Caching options include various scopes (session scope, user scope, virtual database scope). It can take place in memory, disk, or relational database management system (RDBMS) storage, with Time-to-live-based or scheduled updates. The query optimizer employs rule-based and cost-based approaches, with manual configuration or hints. A self-learning	driven query acceleration for summaries, improving performance across geographically distributed architectures. Query optimization involves static and dynamic techniques, including push-down strategies such as join-aggregation push-down (i.e., aggregation pushdown below joins in federated scenarios), on-the-fly data movement and the use of alternative data sources. Denodo integrates an embedded MPP engine based on Presto for parallel processing, both internally and with colocated external MPP	partition pruning. It introduced Apache Arrow, providing integrations for performance optimization. Reflections optimize join queries across sources by creating Iceberg-backed representations, with scheduled or invoked refresh. Dremio Columnar Cloud Cache (C3) optimizes and caches small data components from object storage. Reflections store federated views, while Dremio Sonar accelerates queries using Apache Arrow Gandiva for precompiled Java ByteCode.	pushdown, along with remote stored procedure execution. The SQL query engine excels in MPP processing for analytics across distributed data sources, supporting various caching methods like materialized views and table scan redirection. Starburst employs both rule-based and cost-based optimizers for efficient query planning, while executing queries fully in memory.	considering underlying capabilities and alternatives. It embeds Apache Drill transparently. Queries are processed in memory, spilling to disk if needed. It supports over 15 cache targets, facilitating efficient data transfer and scheduling cache refresh. Proprietary cost and rule-based optimization algorithms enhance predicate push-down and combine operations for reduced data transfer. Statistics aid efficient query execution plans for joins, unions, and caching.

↓	*Data Virtuality* ↓	*Denodo* ↓	*Dremio* ↓	*Starburst* ↓	*TIBCO* ↓
	recommendation engine suggests materialization structures for enhanced optimization.	engines. The platform supports in-memory computing and intelligent caching, recommending data subsets for efficient queries. Smart query acceleration creates summaries, replicable across locations. Denodo's optimizations enhance query performance and processing efficiency.			
Data Transformation & Orchestration	Data Virtuality supports data transformation and orchestration through SQL and provides most of the functionality of SQL-92 Data Manipulation Language (DML). The software has an internal scheduler, an open Management API for	Using ML inductive learning, the Denodo Platform extracts structured information from semistructured data sources. User can enhance data quality by using lookup tables/maps and importing external enrichment services like	Dremio's semantic layer offers a way to save complex data transformations as views and organize them in the semantic layers using folders. Views can be transformed using diverse methods including APIs, SQL, low-	Starburst supports the richness of ANSI SQL and allows data consumers to filter, aggregate, join, split, and create complex calculations on top of their data. In addition, Starburst supports the use of user-defined functions (UDFs) to	TIBCO Data Virtualization provides a variety of data transformation and orchestration capabilities. The product also allows for integration with TIBCO's BusinessWorks for advanced external orchestration and

	Data Virtuality	*Denodo*	*Dremio*	*Starburst*	*TIBCO*
	managing data flows externally, and a built-in data pump for moving large volumes of data in an optimized way between different systems and storages. It has a programmatic API which integrates with external tools like Git integration, and integration with continuous integration/continuous deployment (CI/CD) platforms like Jenkins and GitLab. Data Virtuality will manipulate query plans to ensure efficient use of resources and reduce the need for joins, filtering, grouping, etc.	Google geocoding to clean addresses or Denodo automatically enforces data quality policies. The Denodo SDK for custom functions allows users to add their own custom transformation functions including any kind of processing based on Java. User can use VQL stored procedures to perform complex transformations that require multiple passes over data and complex logic.	code, no-code, generative AI, and integrations with various tools like business intelligence (BI) tools, data science tools, as well as others such as dbt Labs, FivetranFiveTran, Airbyte, and more.	leverage custom code created by the organization to handle specific tasks. Starburst supports integration with leading transformation and orchestration tools such as dbt, Apache Airflow, and Dagster.	transformation use cases. TIBCO Statistica provides automated data cleaning, where the cleansing process involves detecting and correcting inaccurate or missing data and deleting obsolete records.
Data Preparation Capabilities	Data Preparation capabilities are available	Denodo includes the Denodo Catalog, an end-	Within Dremio, users can create individual views	Starburst supports the richness of ANSI SQL to	TIBCO Data Virtualization provides

	Data Virtuality	***Denodo***	***Dremio***	***Starburst***	***TIBCO***
	in both declarative and procedural SQL. Advanced data preparation tasks can be done using procedural SQL, server-side JavaScript, and Python languages. Data Virtuality also uses LLMs to help users work with complex data models, improve understanding, troubleshooting, and speed up creation of data models. An ML-based component is also available to match and map complex schemas into a target schema.	user interface to carry out data preparation tasks. The tool allows the users to build their own mapping and transformation logic by means of the query graphical wizard. Denodo Data Catalog integrates with large language models (LLMs) where users can write statements in free text stating what data they need and Denodo invokes OpenAI's ChatGPT to generate a SQL statement Denodo executes obtaining the data. For more advanced users, the Denodo Catalog includes Virtual Query Language queries and data definition language (DDL) statements.	for each table, facilitating lightweight data preparation for subsequent views in various layers. Administrators can generate views that do tasks like limited casting, type conversion, field renaming, and data redaction for privacy. Moreover, administrators can bolster security by restricting both rows and fields, ensuring that users in other layers don't access privileged information. This approach permits the creation of layered views without compromising the confidentiality of	transform and prepare data however each personas wishes. Most Starburst's connectors support both read and write capabilities, allowing users to land cleaned, transformed, and reshaped data wherever they wish. Starburst also features a specific execution mode named Fault Tolerant Execution. It is useful for long-running and batch/transformation related queries.	data transformation, orchestration and data preparation capabilities like normalization, null-value removal, etc. These capabilities help users to model and analyze data effectively. In addition, for flexibility TIBCO Data Virtualization supports integration with external data preparation tools, such as DataRobot.

↓	***Data Virtuality*** ↓	***Denodo*** ↓	***Dremio*** ↓	***Starburst*** ↓	***TIBCO*** ↓
			underlying sensitive data.		
Database Persistence	Data Virtuality supports databases for persisting data and data storages, including Snowflake, Azure Synapse Analytics, Postgres, Google BigQuery, and others. Data can be persisted using various approaches, including materialization and replication. A recommendation engine provides recommendations on materializing data based on observed usage patterns and source statistics. The connectors provided are optimized for both live querying and replication, and support bulk read	Denodo comes with a built-in MPP engine built on Presto, which can be used right away for both data persistence and caching. Denodo supports a wide variety of third-party databases as well as MPP engines for both query caching and data persistence. The data persistence process uses Remote Tables for data persistence. It also uses high-performance data loading to Apache Hadoop (Spark, iApache Impala, Presto, Databricks) stores, where data is transformed into parquet files, and	DML is on Iceberg native tables in the data lake only. Dremio dp natively connects to myriad other databases, including PostgreSQL, Snowflake, Amazon Redshift, Azure Synapse Analytics, SQL Server, Oracle, MongoDB, Teradata, and others for read-access.	Starburst supports following third-party databases for persisting data: Delta Lake, Iceberg, Apache Hive (Full and Incremental) Snowflake, Teradata, MySQL, MariaDB, Oracle, SAP HANA, PostgreSQL, SQL Server, Amazon Redshift The PostgreSQL connector allows querying and creating tables in an external PostgreSQL database. This can be used to join data between different systems like PostgreSQL and Hive, or between different PostgreSQL instances. The same is	TIBCO Data Virtualization provides built-in support for database persistence using SQL statements such as DDL and DML, including Create Table As Select (CTAS). These statements can be used with JDBC, ODBC, and ADO.NET clients, as well as with SQL Script procedures defined within TIBCO Data Virtualization's integration model to drive data staging and data migration workloads. Additionally, the runtime for these statements has been optimized to make use of bulk loading

	Data Virtuality	**Denodo**	**Dremio**	**Starburst**	**TIBCO**
	and write capabilities. It also supports automatic data replication from each data source.	uploaded to the cluster for high performance. High performance data loading to Teradata (Snowflake), Amazon Redshift (Oracle), IBM Db2 (PostgreSQL), Microsoft SQL Server, Azure Synapse Analytics, and Google BigQuery among others is where Denodo takes advantage of native bulk data loading APIs of these data sources for maximum performance.		true with Snowflake Connector.	mechanisms, which can help improve performance and efficiency.

	Data Virtuality	***Denodo***	***Dremio***	***Starburst***	***TIBCO***
Ecosystem and Partnership	Technology partners: AWS, Microsoft, Google, Snowflake, Collibra, Tableau, InterSystems Caché, etc. Other partners: Precisely (Infogix), Dataspot, PlainID, Solidatus, etc.	Top consulting partners: Accenture, HCLTech, Cognizant, Wipro Enterprises, Deloitte, Tech Mahindra, Carahsoft Technology, Booz Allen Hamilton, CGI, NTT DATA GroupData Top technology partners: Google Cloud Platform (GCP), Azure, AWS, Alibaba, Snowflake, Collibra, Databricks, Salesforce/(Tableau Software), SAP, IBM	Top consulting partners: Aays, Agilelab, drjve, Hiflylabs, Interworks, NTT Data, SDG Group, SVA, Twingo, Unify Top technology partners: AWS Alteryx Dell EMC Deepnote Hewlett Packard Enterprise (HPE) IBM Intel Microsoft Qlik Tableau	Consulting/services partners: Accenture, Capgemini, Cognizant, Deloitte, Infosys, Neudesic, Slalom, Wipro OEM/resellers: Aerospike, Dell Technologies, Domino Data Lab, Google Top technology partners: Alation, Collibra, Dataiku, dbt, HPE, Immuta, Privacera, Protegrity, Qlik, Red Hat, Tableau, ThoughtSpot, TIBCO, VMware Cloud — AWS, Azure, Google	Top consulting partners: Accenture, Deloitte, Capgemini. Top technology partners include: Salesforce, General Electric, Cisco Systems. Strategic technology partners: AWS, Microsoft, Snowflake. Key solution partners: HCLTech Technologies, Apgar, Cadeon

↓	***Data Virtuality*** ↓	***Denodo*** ↓	***Dremio*** ↓	***Starburst*** ↓	***TIBCO*** ↓
Licensing Models	Data Virtuality offers two plans (Standard and Enterprise) in the U.S., Europe, and is expanding to Asia/Pacific	Denodo offers subscription models catering to various project sizes and needs.	Dremio offers subscription, consumption-based, and enterprise licensing models.	Starburst supports the following types of licensing models: capacity/hardware licensing, pay-as-you-go, commit-to-consume/enterprise	TIBCO Data Virtualization provides Subscription, Enterprise, and various licensing options for flexibility, including user-count-based, capacity-based, AWS pay-as-you-use, instance-based, and a few more

Gartner (November 2023)