denodo

Whitepaper

What´s New in Denodo Platform 8.0



Contents

Introduction 3

Enhanced, Unified, Web-Based User Experience 7

Performance Acceleration for Complex Analytical Queries 9

PaaS Support for Cloud and Hybrid Environments 12

Enhanced Data Services APIs with Graph-Like Access to Denodo Views 13

Going beyond the Logical Data Warehouse: New Support for Data Science and ML 14

Enhanced User Experience in the Data Catalog 16

Conclusions 17

What´s New in Denodo Platform 8.0

Denodo Platform 8.0 is a significant step forward in realizing the vision of a logical data fabric as a unified data delivery platform that bridges the gap between IT infrastructure and consuming business applications.

Denodo Platform 8.0 is based on the same data virtualization technology that Denodo has developed over the years to unify data integration, metadata management, security, and data governance. Denodo Platform 8.0 greatly accelerates the delivery of governed data to all data consumers, in the most appropriate format for each, across multiple distributed heterogeneous systems.

At the same time, it serves as an abstraction layer that isolates business applications from the technology infrastructure underneath, facilitating changes, such as adopting new technologies or migrating to the cloud.



© 2020 Denodo Technologies

A logical data fabric based on data virtualization is the answer to today's data management challenges:



FASTER, MORE COMPLEX DEMANDS ON DECISION MAKING

Organizations are facing a significant increase in business speed and requirement complexity, and IT is under high pressure to deliver data to the business, while also receiving more requests that are increasingly complex, from more advanced users such as citizen analysts and data scientists.

Companies have tried to solve those challenges with self-service initiatives, but these initiatives have their own challenges. Analytical information is typically distributed across different systems, such as data warehouses, data marts, or data lakes, and they can be both on-premises or in the cloud. Data usually requires cleansing and transformations (e.g. to deal with specific coding at the data source), and it needs to be combined from different sources. This means that giving users direct access to analytical systems will force them to perform very complex data integration tasks, which is not feasible for them in most cases.

The other alternative is the classical approach of having BI teams first create pre-integrated, curated, physical data marts for every type of user and every type of information need. But this approach is too slow and costly when accomplished using traditional methods, and it does not scale.

Logical data fabric makes data delivery much faster and cheaper, as it enables the creation of virtual data models that expose data to consumers using semantic models and the naming and formatting conventions that are best for every type of data consumer. It does this while minimizing data replication (since it is based on data virtualization), so the process is much faster and cheaper than conventional approaches, such as building physical data marts.



ENTERPRISE-WIDE DATA GOVERNANCE AND SECURITY

Organizations aim at being more data-driven, exploiting their data assets as any other assets in the company, and they have new corporate functions such as the chief data officer. In many industries this comes as a result of pressing external restrictions from regulatory bodies (such as banking- and tax-related agencies).

To become data-driven, however, organizations need to guarantee that data delivered to users conforms to previously agreed-upon semantics and that security and data governance policies are enforced across the whole organization.

Companies try to achieve this by deploying data catalogs and governance tools, however those tools help with part of the problem only. They help to define business glossaries, data quality rules, security rules, etc., but they are disconnected from the actual data delivery

© 2020 Denodo Technologies

process. This means they cannot enforce those policies when they really matter, which is when delivering data to consumers. To enforce these policies effectively, the rules have to be applied in the actual production systems, and again this is very costly and slow when accomplished using traditional methods.

Logical data fabric helps organizations to solve data governance challenges by providing a single entry point for enforcing governance and security policies in the data delivery stage, across all systems.



THE COMPLEXITY OF DATA MANAGEMENT INFRASTRUCTURE: IT COST REDUCTION

As an additional challenge, data management technology is evolving very quickly. A perfect example is the adoption of big data technologies (Spark, Presto, etc.) or the current cloud revolution that has created hybrid data management architectures, with data systems at multiple locations. Cloud deployments usually bring many benefits, but managing the transition, while isolating end users from those changes, can add significant complexity to the cloud strategy.

Logical data fabric abstracts consumers from the underlying technology and the location of the data sources. Consumers can see all types of data according to a consistent model, as if all the data were in a single place. This reduces complexity for consumers and isolates them from the changes happening behind the scenes, where the technology is used in the underlying systems, for example when data repositories are moved to the cloud or when organizations adopt a new technology.

We have seen that a *logical data fabric* based on data virtualization can help solve the main issues that today's companies face in data management. Denodo Platform 8.0 is a significant step forward in accomplishing this vision.

On one hand, Denodo Platform 8.0 increases support for core data virtualization use cases: logical analytical architectures, logical data warehouse, and data services APIs.

On the other, Denodo Platform 8.0 goes beyond traditional data virtualization scenarios to better support new types of users and new types of use cases, such as data science and machine learning (ML) initiatives. It is also a big step forward in platform as a service (PaaS) cloud strategy with enhanced support for automating the management of infrastructure in the cloud, facilitating the management of hybrid environments.

© 2020 Denodo Technologies

THE NEW CAPABILITIES OF DENODO PLATFORM 8.0 CAN BE CLASSIFIED IN THE FOLLOWING WAY



Enhanced, Unified, Web-Based User Experience

- A full web-based interface for all Denodo tools with SSO support
- Full web-based Design Studio tool for developers



Performance Acceleration for Complex Analytical Queries in Logical Data Warehouse / Data Fabric Scenarios

- Smart query acceleration for analytics



PaaS support for Cloud and Hybrid Environments

- Automated infrastructure management for the cloud
- New adapters for the cloud



Enhanced Data Services APIs with Graph-Like Access to Denodo Views

- GraphQL support



Going Beyond the Logical Data Warehouse: New Support for Data Science and ML

- New and expanded tools for data scientists and citizen analysts: “Apache Zeppelin for Denodo” Notebook
- Use ML to automate steps in the data management process



Enhanced User Experience in the Data Catalog

- A redesigned UI and revised user experience
- Leverage ML for automatic recommendations
- Enhanced collaboration, profiling, and search features

Let´s review the different areas in more detail (for an exhaustive list of new features please visit **Denodo Platform 8.0 documentation**.

© 2020 Denodo Technologies

Enhanced, Unified, Web-Based User Experience

A full web-based interface for all Denodo tools with SSO support

The new Denodo Central Web Console, integrated in the Solution Manager, provides a single entry point for all Denodo tools, enabling all users to access all Denodo environments, both on-premises and in the cloud. It supports SSO using Kerberos, SAML, OpenID, and OAuth, for seamless connectivity across all Denodo tools.

Denodo Platform 8.0 also offers a tighter integration between Denodo tools. The Diagnostic and Monitoring Tool is now integrated with the Solution Manager, sharing the same catalog of information about the Denodo deployment, so users can navigate seamlessly between them.

This portal provides a more unified experience for the user across the different Denodo tools and environments.

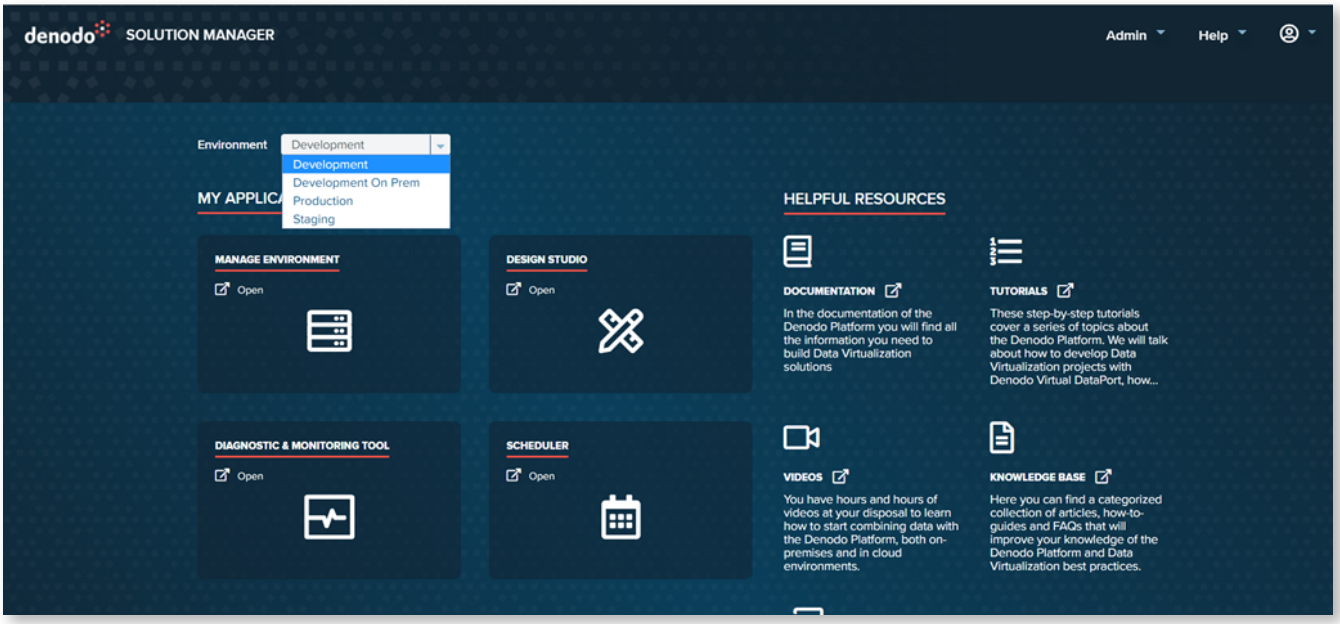


Fig. 1. The Denodo Central Web Console in the Solution Manager

New Web Design Studio: Advanced Web-Based Development Studio for Data Developers

Developers now benefit from a new web-based Studio tool with which to develop views and data services (which can be used in addition to the desktop version). This new tool has been designed to ensure that “ease of use,” one of the main benefits listed by Denodo users, is maintained and enhanced by this new interface. With the new Design Studio, users will experience how easy it is to connect to a variety of data sources, combine and transform them to create virtual views, and then publish them for access in multiple formats.

© 2020 Denodo Technologies

It includes:

- Connection to any data source at any location
- Graphical modeling wizards to easily define and publish business-friendly virtual data sets
- An SQL shell to run queries from the browser
- One-click publishing of secure data services using technologies like REST, OData, and GraphQL
- Integration with version control
- And many more features

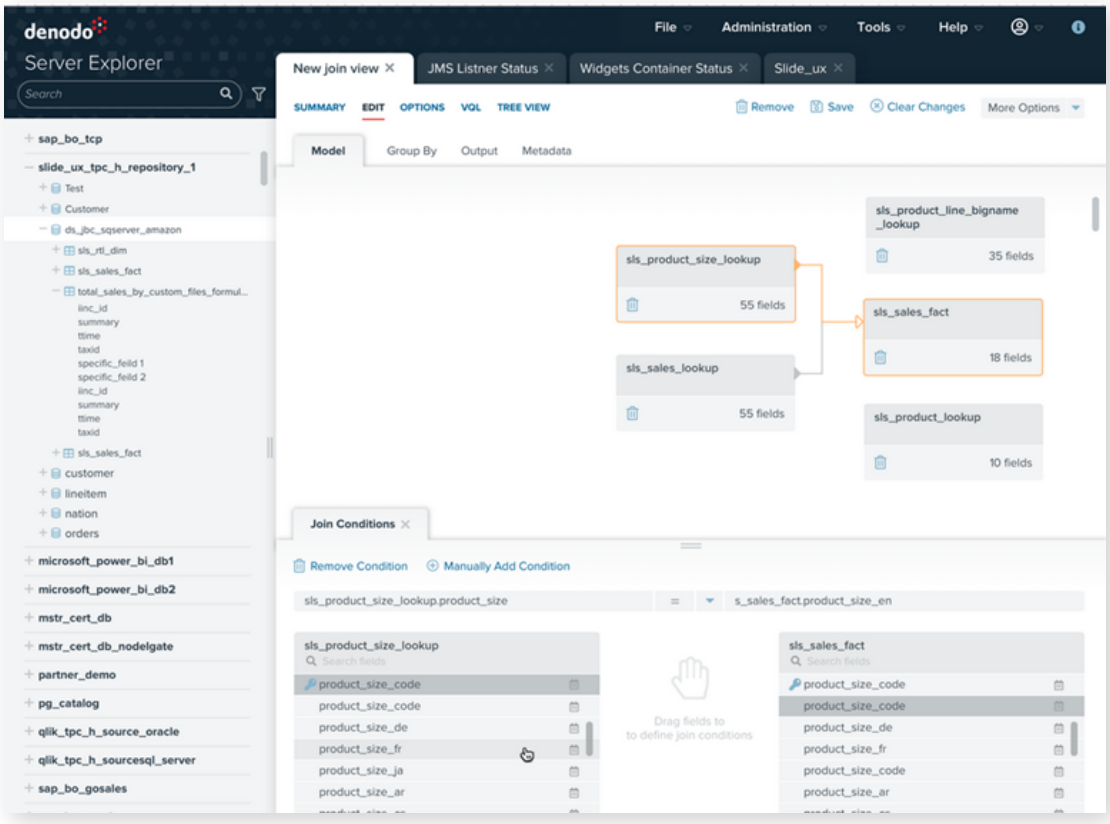


Fig. 2. The Denodo Studio tool

Other Administration Enhancements

In addition to the new web-based tool, Denodo Platform 8.0 incorporates many other enhancements and new features that simplify platform administration, such as:

- Simplified TLS/SSL configuration
- Enhanced support for managing cost statistics
- Management of JDBC drivers and other external libraries
- Finer-grained security in Solution Manager
- Encrypted metadata export /import
- And many more

© 2020 Denodo Technologies

Performance Acceleration for Complex Analytical Queries in Logical Data Warehouse / Data Fabric Scenarios

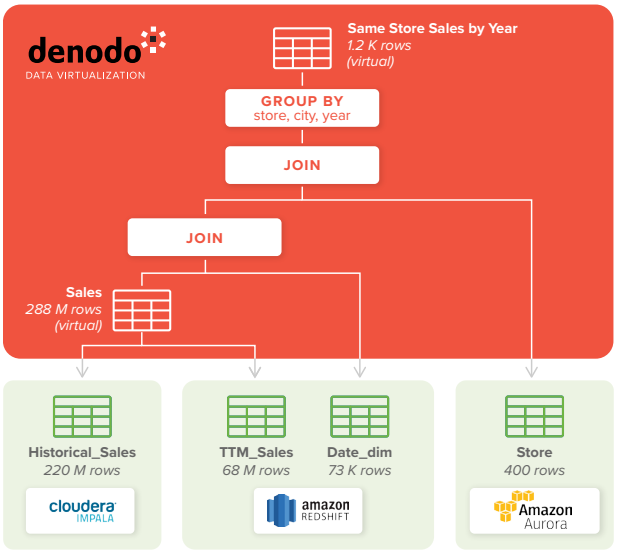
Smart Query Acceleration for Analytics

Denodo Platform 8.0 introduces a new concept, smart query acceleration, which accelerates the execution of queries in a logical data warehouse / data fabric architecture. Partial aggregates of fact and dimension tables, which are commonly joined together in many queries of a certain type, are precomputed, and used to accelerate future queries. This technique achieves significant performance gains and facilitates the building of ad-hoc queries in a self-service scenario, as Denodo Platform 8.0 handles all of these performance optimizations behind the scenes.

These partial aggregates are called “summaries” because they summarize the original datasets, and they are much smaller than original ones, yet they still contain enough information to answer relevant query subsets over them.

Summaries are created by the administrator and are also automatically suggested by Denodo Platform 8.0. By analyzing the different query patterns in a given scenario and the performance of past queries, the system applies an ML model to identify suitable “summaries” that will help to accelerate future queries. Summaries are persisted either in the cache or in a data source defined by the user. The user can make use of any supported store as a cache or as a data source in Denodo Platform 8.0, including in-memory stores, for further query acceleration.

This concept is similar to “aggregation awareness” or “materialized views,” which are used by some OLAP engines and some reporting tools, but with a key difference: Denodo can provide this acceleration technology for all data sources and all consumers, so all users will be able to benefit from it. We think that this type of technology belongs to the data virtualization layer, so as to be decoupled from specific data sources or reporting tools. In addition, smart query acceleration is fully Integrated with Denodo’s engine query rewriting rules and CBO, to provide features that are not available in any other product designed for enhancing logical data warehouse or data fabric scenarios.



Example Pattern:

- Combine all sales data (Redshift and Impala) with stores data in Aurora
- Compute metrics aggregated by different time periods (day, month, year...)

Proposed Summary: total sales by store and day, include only *store_id*

- Aprox. 300k rows (100 times smaller)
- Precomputed and stored in Redshift

Example Query: total sales by store and year, including store name and city

- Optimizer Joins summary with store and date dims. to obtain missing fields
- Then re-aggregates by year instead of day

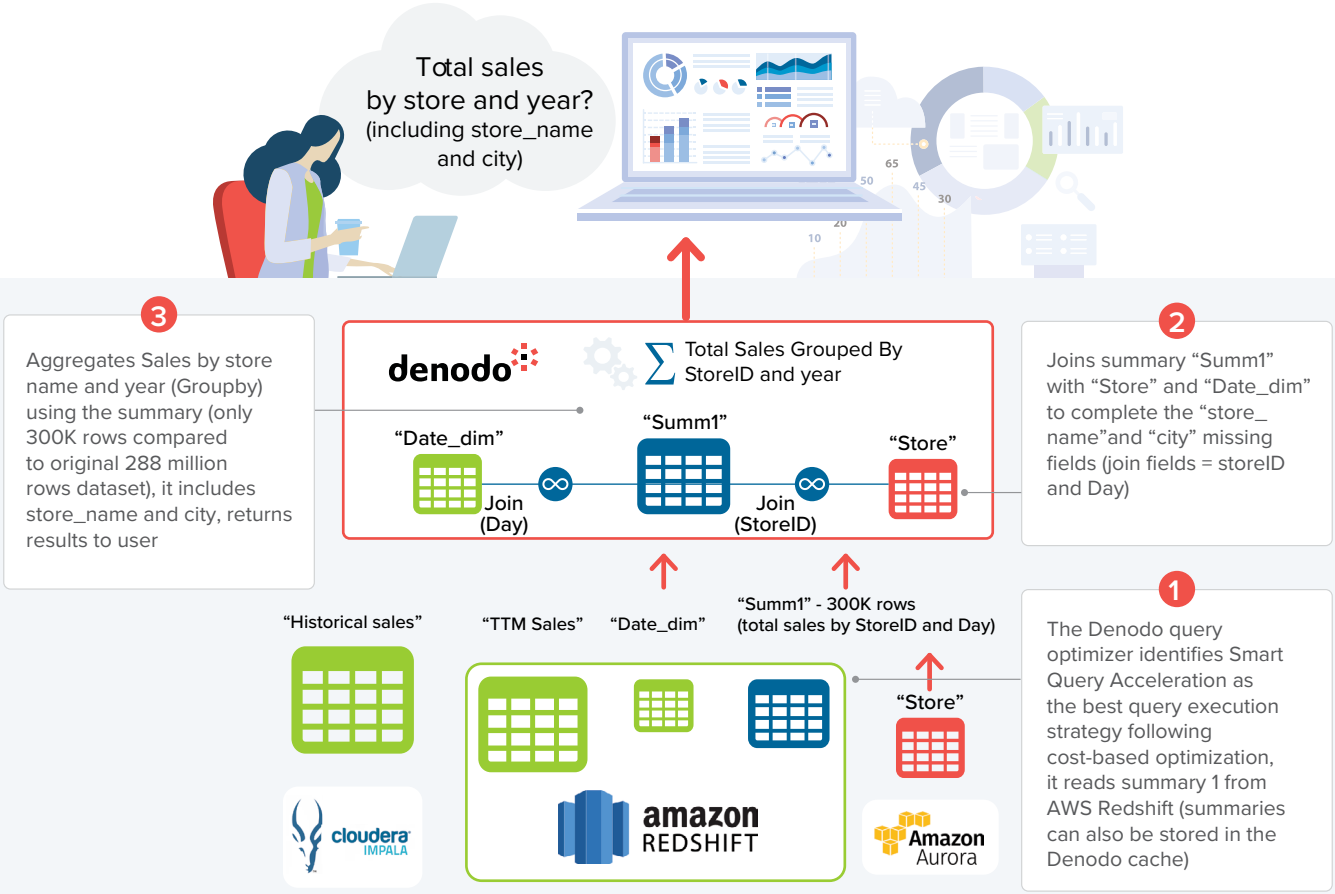
Fig. 3. Analytical Query Scenario with Summaries

© 2020 Denodo Technologies

Example: Let´s suppose we have an analytical scenario with sales information spread across Redshift (TTM sales) and Cloudera (historical sales), with an additional dimension in an Aurora database (store).

Let´s say we need to compute "total sales by store and year," including the store_name and city, and that we have the following summary already created in Redshift: "total sales by store id and day." Notice that this summary does not have any columns from the store and date dimensions. For instance, it does not have the store_name, city, or the year that the query needs, and the results are aggregated by day, while the query needs them aggregated by year.

The Denodo optimizer can still detect that this summary can be used as a starting point to compute the result. Rather than using the raw sales data, which is much bigger (nearly 300M rows), it uses the summary, which is only 300K rows. Next, it joins the summary with the store and date dimensions to get the additional needed columns,and then the optimizer can re-aggregate the results by year, rather than by day, to obtain the final results.



- "Historical Sales" - Fact table, 220M rows
- "TTM_Sales" -Trailing Twelve Months Sales - Fact table 68M rows
- "Date_dim" – dimension table – 73K rows
- "Store" - dimensión table - it contains StoreID, store_name and City
- "Summ 1" – summary, 300K rows, obtained with "Total Sales by StoreID and Day" query, stored in Redshift

Fig. 4. Smart Query Acceleration (Summaries)

© 2020 Denodo Technologies

With large data volumes, such an approach can be much faster than recomputing the entire query from scratch. In this example the query executes almost 10 times faster.

SYSTEM	EXECUTION TIME
Other systems	>500 secs
Denodo (no summaries)	~ 13 secs
Denodo (with summary)	~ 1.4 secs

Table 1. Performance improvement of Summaries

The key advantage of summaries is that the same summary can be useful for many different queries that follow a similar pattern. For example, Table 2 shows the results of executing five similar queries in the same scenario. In all cases, the query execution times are between 5 and 20 times faster.

QUERY	EXECUTION TIME (NO ACCELERATION)	EXECUTION TIME (ACCELERATION)	PERFORMANCE GAIN	SUMMARY USED
Total sales by year	15.45 secs	2.38 secs	6.5 x	summary_total_by_store_day
Total sales by quarter, store name and city	22.49 secs	2.62 secs	8.57 x	summary_total_by_store_day
Total sales by store and city for last quarter	14.71 secs	0.47 secs	31.1 x	summary_total_by_store_day
Total sales in a specific store	14.36 secs	2.66 secs	5.39 x	summary_total_by_store_day
Total sales in a specific store and year	14.32 secs	3.18 secs	4.0 x	summary_total_by_store_day

Table 2. Performance improvements over other analytical queries using the same summary

An additional benefit is that queries that use summaries (notice that in this case the summary has been created in Redshift) do not need to get access to the on-premises system (Cloudera Impala in this example), because the data needed from that on-premises system has been summarized in Redshift. This is a promising feature for hybrid scenarios, as users can generate summaries from data sources in remote locations, to minimize network traffic.

Also, whenever users need to accelerate query execution in a given environment (such as a Hadoop store), they can create “summaries” and store them either in the same store or in another one (e.g. an in-memory database) and use them to accelerate queries. In this case, they would be using Denodo Platform 8.0 as a real query acceleration engine for another system.

A last scenario is cost reduction in cloud systems that charge per use or per bytes scanned, as queries accelerated with summaries are significantly cheaper, since they need less data and fewer CPU cycles. This Denodo Platform 8.0 feature can therefore reduce your bill in a cloud system.

© 2020 Denodo Technologies

PaaS Support for Cloud and Hybrid Environments

Automated Infrastructure Management in the Cloud

One of the most significant features of Denodo Platform 8.0 is automated infrastructure management in the cloud, which automates all of the tasks related to installing, configuring, deploying, and upgrading Denodo Platform clusters.

Initially available for AWS (Azure support coming soon), this functionality is delivered through the new Solution Manager, which now provides a web-based UI with which users can define and configure clusters, indicating user preferences in aspects such as TLS configuration, load-balancing, autoscaling, etc. Once the clusters have been defined using this UI, users can simply press “Start,” and the clusters will be automatically installed and created. The system also automates the installation of updates in Denodo clusters and offers integrated monitoring of all cluster activity.

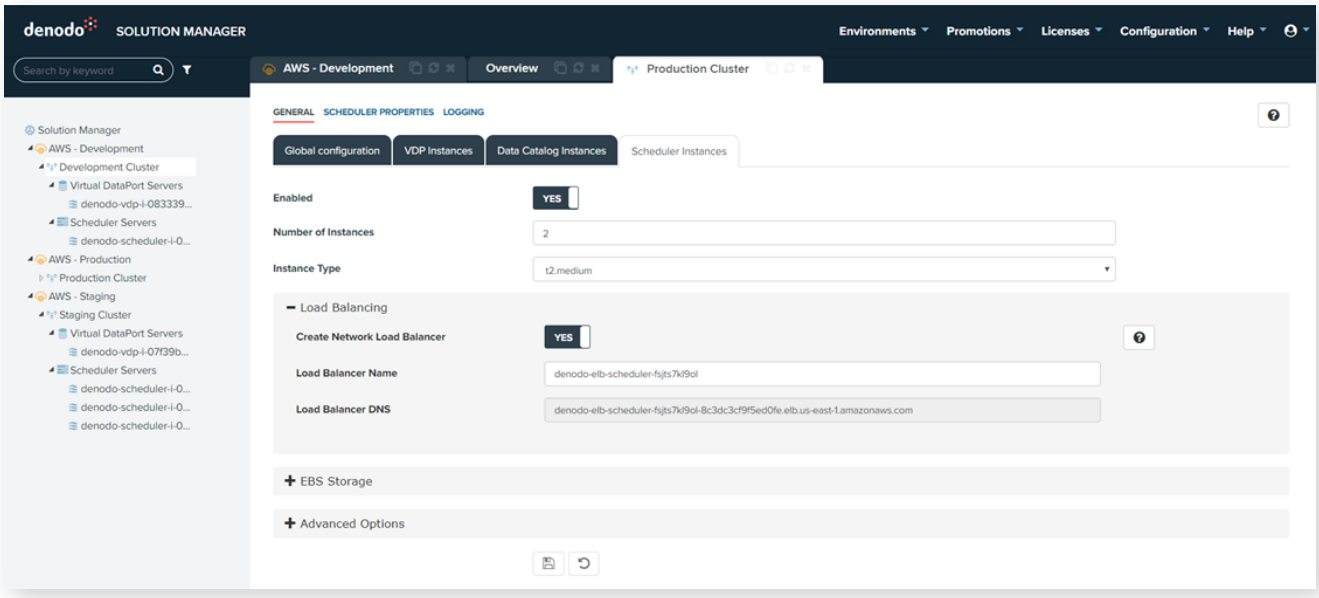


Fig. 5. Automated Infrastructure Management for the Cloud

In addition, the Diagnostic and Monitoring Tool is fully integrated in the Solution Manager, and it shares the same metadata. There is also an extended privileges system that controls what users can do in each cluster, who can start or stop a given cluster, who can monitor it, who can modify its configuration, etc. There are also important enhancements in import/export functionalities for backups as well as in other features such as managing the execution of Denodo Monitor or deploying promotions.

© 2020 Denodo Technologies

New Adapters for the Cloud

Denodo Platform 8.0 includes new adapters for Databricks Delta, Azure Synapsis, and Google BigQuery. In all cases, Denodo Platform 8.0 makes use of bulk data load APIs to efficiently load data for the data movement optimization mechanism or cache.

It also includes a new connector, the "Distributed File System Connector" for cloud data storage systems such as AWS S3, Azure ADLS/Blob storage, and Google Cloud Storage. This connector is able to read Parquet files in parallel, achieving very high performance (files, row groups, and columns are read in parallel), and both predicate and projection pushdown are supported by native Parquet APIs.

Some other enhancements include support for using IAM roles when accessing Redshift and Athena, as well as templates for accessing Marketo.

Enhanced Data Services APIs with Graph-Like Access to Denodo Views

GraphQL Support

Denodo Platform 8.0 includes support for **GraphQL**. GraphQL provides a query language for APIs, and it is one of the fastest growing data services standards.

GraphQL provides an abstraction layer on top of REST APIs in such a way that applications can perform a single call to the API to obtain all the data they need for a certain action, rather than performing multiple API calls, as it is usually required by conventional REST APIs.

This improves performance, and it also simplifies consuming applications by removing the orchestration and combination logic needed to integrate the output of the different API requests. For this reason, GraphQL is mostly used for web application development, as it simplifies the logic normally needed in the UI.

GraphQL offers many benefits for API consumers, as we have seen, but they have the following main drawbacks:

- GraphQL APIs are very costly to develop.
- The performance of hand-coded implementations suffers because combining data from multiple endpoints in real time is complex.

Denodo solves both issues for GraphQL data APIs: With Denodo Platform 8.0, development cost is virtually zero, because all virtual views in Denodo can now be invoked with GraphQL out-of-the-box. In addition, the Denodo optimizer takes care of the data combination from multiple end-points, so the performance is also improved.

© 2020 Denodo Technologies

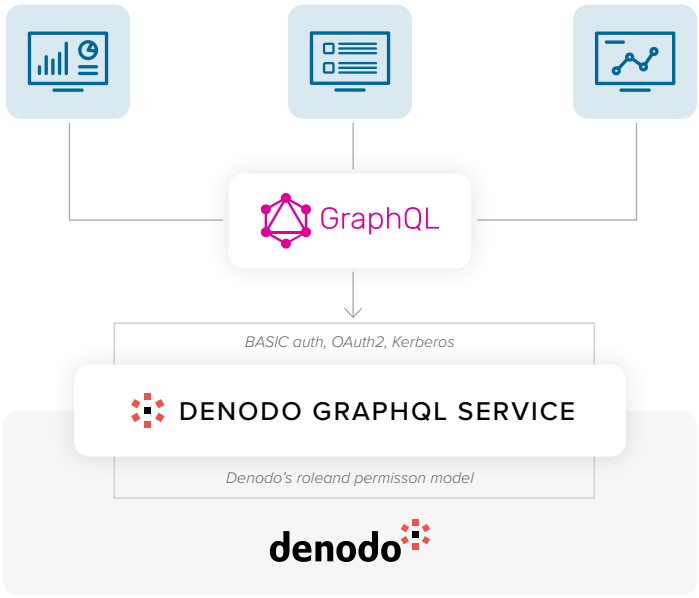


Fig. 6. Denodo GraphQL Service

GraphQL is normally used as an abstraction layer between UI and REST services.

- Decreases the number of API requests
- Removes orchestration from the UI when obtaining data

Denodo Platform 8.0 can provide declarative execution of GraphQL queries on top of Denodo's virtual data model, with zero code:

- Zero development time, with better performance than manual coding
- Resource management: for example assign an associated quota per user or role (10 queries per hour)
- Security, Optimization, Lineage...

Going beyond the Logical Data Warehouse: New Support for Data Science and ML

Denodo Notebook for Data Science

Denodo Platform 8.0 plays a significant role in data science projects. According to many studies, data scientists typically spend most of their time doing data preparation and data integration tasks (up to 80% of their time), rather than creating and refining their data science models.

Denodo minimizes this problem by providing a very agile way to expose data views and data services in formats that are friendly to data scientists, and it also abstracts them from complexities such as where the data is located, the native technologies that are used at each data source, and other data integration complexities.

© 2020 Denodo Technologies

To further facilitate data scientists' access to data, Denodo Platform 8.0 introduces a new tool for data scientists, called the Denodo Notebook.

Using the Denodo Notebook, data scientists can construct narratives that combine queries, code, text, and graphics, to aid in data analysis and to help data scientists explain their work and share it with colleagues.

The Denodo Notebook is based on Apache Zeppelin, one of the most popular notebooks, which has been fully integrated with Denodo Platform 8.0, so data scientists can easily access any data view from the Denodo Platform. It has also been integrated with the Denodo security system including SSO support, so all security and governance policies defined at the Denodo Platform layer are enforced when data scientists use the Denodo Notebook.

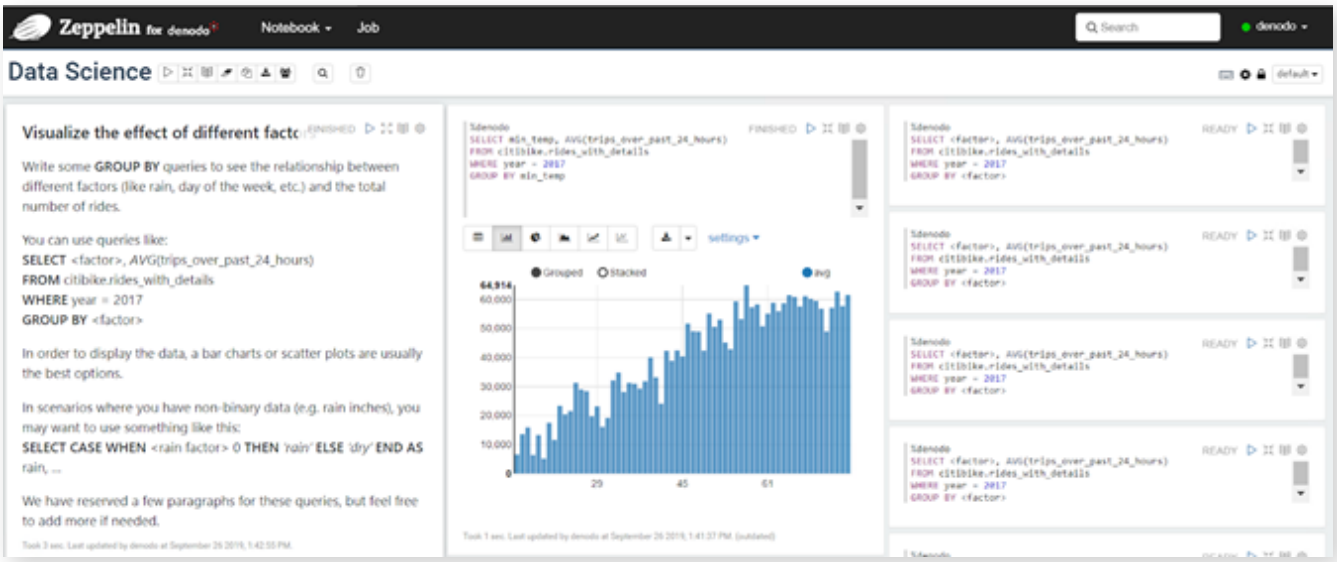


Fig. 7. "Zeppelin for Denodo" Notebook

Summary of "Zeppelin for Denodo" Notebook features:

- Combine queries, scripts, graphics, and text to create narratives
- Based on Apache Zeppelin
- Denodo users can create, save, and share their own notebooks
- Fully integrated with Denodo's security system and SSO capabilities
- Download it from the Denodo Connects section of the Denodo Support Site

© 2020 Denodo Technologies

Enhanced User Experience in the Data Catalog

Denodo Data Catalog

The Denodo Data Catalog provides a data marketplace for business users (citizen analysts, data scientists, etc.), with three main features:

- Business users can quickly discover interesting data sets (virtual or physical), by browsing over datasets classified in business categories and also by searching the metadata as well as the contents of the datasets.
- When business users find potentially interesting datasets, they can put them in context, to understand exactly what they provide and how they could be used: Users can see column descriptions, tags, business categories of the datasets, data lineage information, and also information about how other users and applications are using the dataset.
- Finally, the catalog also offers data preparation wizards for querying and customizing the data, which enables users to connect to the data with their favorite data visualization tools.

Denodo Platform 8.0 provides the following enhancements to the Denodo Data Catalog:

- A Redesigned UI offering a revised user experience
- Machine-learning-powered features that analyze user activity to provide personalized recommendations and shortcuts to select datasets.
- Enhanced collaboration features: Users will be able to endorse datasets or register comments or warnings about them. This will help users to further contextualize dataset usage and better understand how other users experience them.
- Improved information about past dataset activities, including which users and applications are querying which datasets, and which queries are used more frequently.
- Other enhancements include extended profiling information about datasets and columns and improvements in smart search (smart ranking of results).

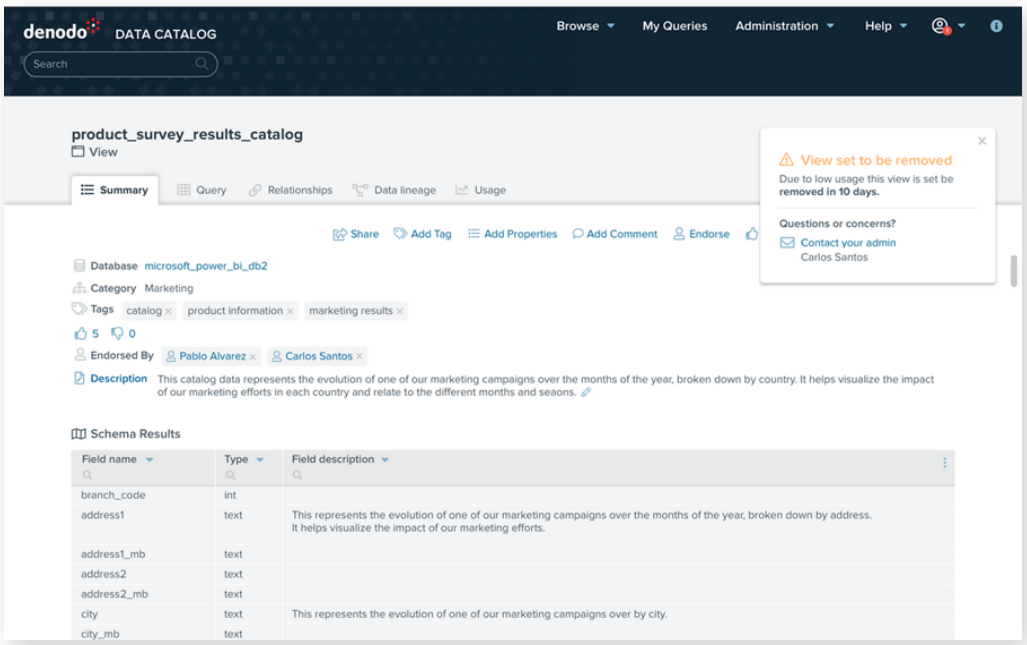
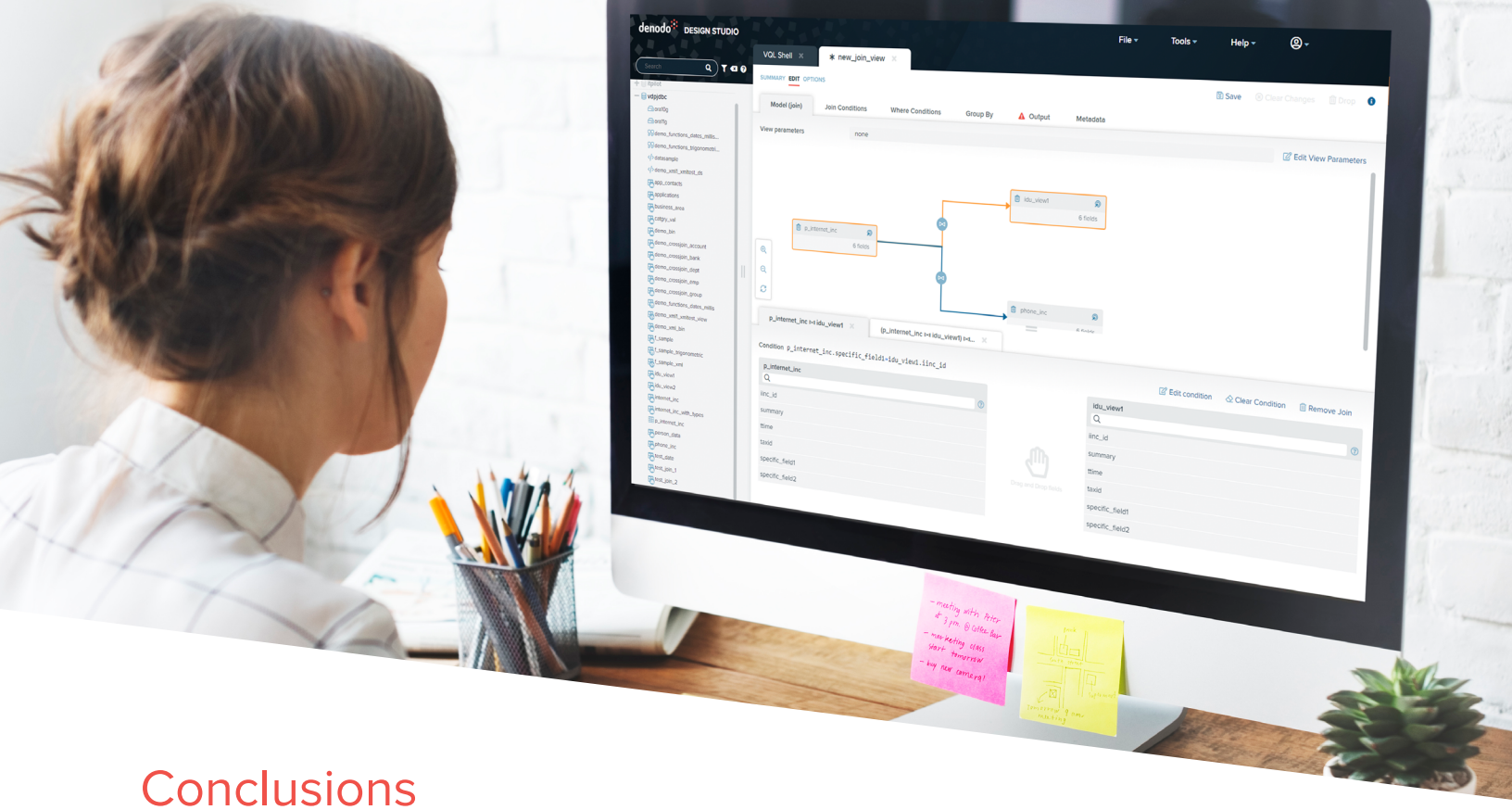


Fig. 8. Redesigned Denodo Data Catalog

© 2020 Denodo Technologies



Conclusions

Denodo Platform 8.0 and data virtualization play a crucial role in the most pressing data management challenges today, enabling agile, governed data delivery.

The new release of Denodo Platform 8.0 reinforces the Denodo Platform's strengths, such as ease-of-use and performance, extending the platform's reach into new areas like data science, and offering first-class support for cloud and hybrid scenarios. The main new features in Denodo Platform 8.0 include:

- A full web-based interface for all Denodo tools with SSO support: An integrated, web-based experience across all tools.
- Web-based Design Studio tool for developers, providing ease-of-use across all of the steps in the data-service development process.
- Smart query acceleration for analytics: Partial aggregates ("summaries") are pre-computed to accelerate future queries. Denodo Platform 8.0 provides this acceleration mechanism for all data sources and consumers.
- Automated infrastructure management for the cloud: PaaS support including cluster configuration (TLS, load-balancing, autoscaling, etc.), start/stop controls, automatic installation of updates, and integrated monitoring.
- GraphQL support: Zero-code creation of GraphQL data APIs with first-class performance by leveraging the Denodo query optimizer.
- "Apache Zeppelin for Denodo" Notebook: Data scientists can construct narratives that combine queries, code, text, and graphics, to aid in data analysis and to help them to explain their work and share it with colleagues.
- Redesigned data catalog with automatic recommendations and enhanced collaboration, profiling, and search features.