# Expert Trail: Query Performance Optimization

Revision 20240211



Denodo Inc.: 525 University Avenue, Suite 31, Palo Alto, CA 94301. USA www.denodo.com

CONTENTS



# 1 LOOKOUT

Expert trails guide Denodo users through all the relevant materials related to a specific topic, including official doc, KB articles, training, Professional Services offering, and more. The main goal is to give users a single place with references to all the information that they need to become a Denodo expert on any specific topic.

“What about performance?” is usually the first question that is raised when introducing someone to data virtualization.
This Expert Trail shows a curated selection of the different resources available to become a master in Denodo Query Performance Optimization.

# 2 THE HIKE

## 2.1 STAGE 1: MODELING

Let's start this path by checking the Best Practices and general guidelines on how to design views that are optimal from a performance point of view.

In this particular case, the focus is on a Big Data/Analytic use case.
This guidance is especially important in analytic use cases which require integrating large volumes of data from different sources:
Best Practices to Maximize Performance I: Modeling Big Data and Analytic Use Cases
**NOTE** that this type of scenario was selected as it is where Performance Optimization techniques can shine and make a big difference. Other scenarios i.e. operational do not require smart optimization techniques as the volumes of data involved are small and performance will be good.

The Expert Trail: Modelling and Metadata Organization explains how to do the modeling and how to organize the metadata.

## 2.2 STAGE 2: LEVERAGING THE OPTIMIZER

Let's focus now on the modeling best practices to handle Big Data Analytic use cases, dealing with horizontal and vertical partitioned federated data and other considerations.

Denodo Query Optimizer Engine analyzes the metadata and automatically determines the best execution plan to run a query in the most performant way. For this reason, it's essential to make sure that all the information that can be relevant for the optimizer analysis is available and it is accurate. This information includes PKs, Indexes, Referential constraints, statistics, etc.

Let's take a deep dive into the additional meta-information and configuration settings that are relevant for the Denodo Query Optimizer module.
Best Practices to Maximize Performance II: Configuring the Query Optimizer

## 2.3 STAGE 3: CACHING

With a better understanding of what Denodo can do to manage real-time queries even in analytical federated scenarios, it is time now to take a look at the Denodo Cache. Denodo allows configuring a Cache Engine to store local copies of the data retrieved from the data sources.

Caching in Denodo can be used for several purposes, such as enhancing performance. There might be situations that due to the client SLA´s, caching is required. This often raises a question that is developed in Comparing caching in Denodo with other forms of replication like ETL

Other uses of the cache would be to protect data sources from costly queries, and/or reusing complex data combinations and transformations.
Denodo also contains smart query acceleration techniques based on pre-stored data using an element called Summary.

Information about the Cache can be found in the Expert Trail: Cache

Let´s review the recommendations for different aspects of the Cache Module such as, how to choose the cache database, how to decide what views to cache, or what is the best cache mode and refresh strategy for each particular use case:
Best Practices to Maximize Performance III: Caching.

### 2.4 STAGE 4: TROUBLESHOOTING

Finally, the latest step is to learn a method to identify the bottlenecks of a Query in Denodo and the different options and actions to resolve or improve the performance of an existing query.
This is considering that the origin of this issue is the query itself regardless of server congestion situations.
Best Practices to Maximize Performance IV: Detecting Bottlenecks in a Query
The Expert Trail: Monitoring explains the details on how to monitor the performance of a query or server.

## 3 EXPLORATION

Fill up your backpack with additional gear:

### Performance overview

| | |
|---|---|
| Webinars | ● Myth Busters I: Can data virtualization uphold performance with complex queries? |
| Additional Resources | ● Achieving Lightning-Fast Performance in your Logical Data Warehouse - Post \| Data Virtualization blog<br>● Physical vs Logical Data Warehouse Performance - Post \| Data Virtualization Blog |

### Cache

| | |
|---|---|
| Official Documentation | ● Cache Module — Virtual DataPort Administration Guide<br>● Configuring the Cache — Virtual DataPort Administration Guide<br>● Monitoring - Cache — Diagnostic and Monitoring Tool Guide |
| KB Articles | ● Batch inserts in the cache and Scheduler<br>● Incremental Queries in Denodo<br>● Cache database size estimate<br>● Comparing caching in Denodo with other forms of replication like ETL<br>● What happens if the cache database goes down? |
| Additional Resources | ● Data Virtualization Basics Tutorial - Tutorial \| Denodo Community Site |

### Smart Query Acceleration (Summaries)

| | |
|---|---|
| Official Documentation | ● Smart Query Acceleration Using Summaries — Virtual DataPort Administration Guide |
| KB Articles | ● Best Practices to Maximize Performance III: Caching: Summaries vs Cache |
| Webinars | ● Accelerate your Queries with Data Virtualization |
| Additional Resources | ● Increase the Performance of Your Logical Data Fabric with Smart Query Acceleration - Post \| Data Virtualization Blog |

## Join Types

**NOTE:** The Cost Optimizer will automatically select the best JOIN Type to run a query when statistics are available. Nevertheless, it is useful to know how they work for query performance troubleshooting or understanding.

| | |
|---|---|
| Official Documentation | ● Optimizing Join Operations — Virtual DataPort Administration Guide |

## Denodo Optimizer Techniques

| | |
|---|---|
| Official Documentation | ● Automatic Simplification of Queries — Virtual DataPort Administration Guide (denodo.com) |
| KB Articles | ● **Example: Aggregation Push-down (full and partial)** Denodo Query Optimizations for the Logical Data Warehouse<br>● **Example: Partitioned Union** Denodo Query Optimizations for the Logical Data Warehouse (Part 2): Working With Partitioned Fact Tables |

## Statistics and Cost Optimizer

| | |
|---|---|
| Official Documentation | ● Cost-Based Optimization — Virtual DataPort Administration Guide |
| KB Articles | ● How to update the statistics of a view automatically |
| Additional Resources | ● Cost-based Optimization in Data Virtualization - Post \| Data Virtualization Blog |

## Massive Parallel Processing (MPP):

| | |
|---|---|
| Official Documentation | ● Denodo Embedded MPP - User Manual |
| KB Articles | ● How to configure MPP Query Acceleration in Denodo<br>● MPP Query Acceleration: Sizing guidelines<br>● Configuring an Autoscaling Denodo Embedded MPP Cluster in EKS |
| Webinars | ● Demo: Parallel In-Memory Processing and Accelerating Analytics Performance<br>● Unraveling the Data Lake: MPP integration within a Logical Data Fabric |

## Denodo Connects

| | |
|---|---|
| Official Documentation | ● Denodo Cloud Cache Load Bypass Stored Procedure<br>● Denodo Incremental Cache Load Stored Procedure<br>● Denodo Embedded MPP - User Manual |

## Denodo Test Drives:

Denodo Test Drive provides a private sandbox environment containing a preconfigured solution that demonstrates how data virtualization brings agility and flexibility to multiple use cases. In under an hour, and using a step-by-step guide you will experience how to quickly take advantage of multiple data sources independent of location and format with zero replication.

| | |
|---|---|
| Additional Resources | ● Test Drive: Agile BI and Analytics on AWS - Test Drive \| Denodo<br>● Test Drive: Agile BI and Analytics on Azure - Test Drive \| Denodo |

# 4 GUIDED ROUTES

## 4.1 DENODO TRAINING COURSES

Denodo training courses provide expert data virtualization training for data professionals, including administrators, architects, and developers.
If you are interested in Query Performance you should enroll in the following course:

- **Denodo Performance Best Practices**: Data Virtualization architectures need to be fully-performance. This course will talk about the internal details of the Denodo Optimizer to learn how to maximize the performance of the queries executed in the Denodo Platform.

## 4.2 TECHNICAL ADVISORY SESSIONS

Denodo Customers with active subscriptions have access to request Meet a Technical Advisory sessions.
These are the sessions available related to performance.

| | | |
|---|---|---|
| Platform Administration | **Performance Optimization: Administration**  | Assist in enabling optimizations capabilities:<br>- Enable Data Movements and Bulk Load.<br>- MPP (Massive Parallel Processing).<br>- Statistics: Gathering and refreshing policies. |
| Cache: Standards & Best Practices | **Cache Modes Overview** | Review and showcase how the Denodo Cache works:<br>- The different cache modes (Partial, Full, Incremental).<br>- Loading the cache and refreshing.<br>- Invalidating the cache contents.<br>- Indexes. |
| | **Cache Best Practices** | - Assist you in defining a policy to determine when to use the cache and what type to use, or review your current policy.<br>- Advice on selecting the best cache strategy for a specific scenario.<br>- Incremental caching strategies. |
| | **Summaries Overview and Best Practices** | Review and showcase how the Denodo Summaries works. |
| Performance Optimization | **Performance Optimization: Overview**  | Explanation of the different performance optimization techniques (Optimization features overview).<br>- static optimization |

| | | |
|---|---|---|
| | | - cost-based optimization<br>- cache and smart cache (summaries)<br>- remote tables<br>- MPP<br>- etc. |
| | **Performance Optimization: Best Practices** | In-depth best practices based on features applied to use optimizations effectively. Recommendations to decide what optimization is the most appropriate for each scenario and your use case. |
| | **Performance Optimization: Detecting Bottlenecks** | Guidance in the techniques to analyze and improve the performance of a query that is not reaching expected goals. |

### 4.3 PROFESSIONAL SERVICES

Denodo Professional Services can help you at the start or any part of your query performance trail. You can find information about the Denodo Professional Services offering in:
Professional Services for Data Virtualization | Denodo

If you are a Denodo customer, you can reach out to your Customer Success Manager for details about any Guided Route that you need.

# 5 BIG HIKE PREP CHECK

Let's see if you are ready to start your big trail. Take this 2-question questionnaire to check your readiness for an enjoyable hike.
Read the questions below, think about the solution and check if you got them right by looking at the solution. Have you become an expert?

1. If the same data is replicated physically in 2 places, is it possible for Denodo to use a unique view, so depending on the query Denodo uses the most convenient data source? How?
   Click here to check if you got it right

   Yes! This can be achieved through the Alternative Sources functionality. You can configure a base view to indicate that the same data is stored physically in different sources.
   At runtime, when this base view is involved in a query, the optimizer will select the source of the data of this base view that maximizes the number of operations that can be pushed down to the underlying source.

2. What views or types of views should have statistics? How often should those statistics be gathered?
   Click here to check if you got it right

   So the cost optimizer can be applied, it needs the statistics for the following views involved in a query:
   - All the base views
   - Derived views that have cache enabled
   - Flatten views
   - Summaries

   This is explained in the section What views require statistics? of the "Best Practices to Maximize Performance II: Configuring the Query Optimizer" KB article.

   Regarding how often to gather statistics, it is a good practice to keep the statistics as much updated as possible. They should be gathered from the underlying data source. Nevertheless, this is not always possible so it is convenient to gather the statistics during periods where the system is not expected to be under heavy load. Also, it is important to understand that statistics changes will affect when there is a significant change in the data they hold.
   For example, 50,000 new rows might represent a relevant change or not. If a table changes from 1000 rows size to 51000 rows. That is a significant change. If it changes from 5,000,000 rows to 50,050,000 rows the change is not relevant.