HITACHI
Inspire the Next

Pentaho Data Modeling and Storage

HITACHI
Inspire the Next

Change log (if you want to use it):

Date	Version	Author	Changes

Contents

This page intentionally left blank.

Overview

This document covers some best practices on how to design and build your Pentaho solution for maximum speed, portability, and knowledge transfer, as well as ease of reuse and maintenance.

Our intended audience is Pentaho or database administrators, or anyone with a background in data storage who is interested in optimizing speed and performance for storage, retrieval, and reporting.

The intention of this document is to speak about topics generally; however, these are the specific versions covered here:

Software	Version(s)
Pentaho	6.x, 7.x, 8.0

The Components Reference in Pentaho Documentation has a complete list of supported software and hardware.

This document assumes that you have knowledge about Pentaho and database administration.

© Hitachi Vantara Corporation 2018. All Rights Reserved

Best Practices for Data Modeling

To maximize the speed and usefulness of your Pentaho solution, there are several steps you can take to model your data. This section provides that information and covers operating and improving databases, tables, and values.

Table 1: Data Modeling Recommendations

Recommendation	Details
Use **dimensional models** whenever possible	Dimensional models are optimized for online queries and data warehousing, and allow Mondrian and Pentaho to perform at their best at high volumes. One common example of a dimensional model is a star schema.
Optimize your database server memory and processing power and adjust Database Management System (DBMS) kernel to use that capacity	Because Mondrian can only perform as fast as the database can return data, we recommend that you optimize your database server and instance for analytic workloads. Databases have specific parameters for analysis that do not apply to transaction workloads. Typically, you will want to provide the maximum amount of memory and processing power to the database server, and adjust the DBMS kernel parameters to efficiently use that additional capacity.
Apply **standard indexing** to your database	Apply standard indexing to your database, to allow Mondrian-generated queries to perform optimally. A common approach to indexing appears in Recommendations for database tuning. Create indexes on all primary keys of a dimension, and all foreign keys in a fact table. Create indexes for each level of each hierarchy in all dimensions of all cubes of all schemas. Indexes on keys are particularly important on high cardinality dimensions and levels. Primary and foreign keys should be single-column integers or BIGINT data type. Keys should not be string, GUID, or a combination of several fields.
Define all tables in the schema as database tables	Wherever possible, avoid using database views or structured query language (SQL) queries as tables in a schema. Instead, define all tables in the schema as database tables. Normally, when an SQL query or view is desired, this is an indication that more extracting, transforming, and loading (ETL) needs to be done to the data before analysis. If you do use a database view, the entire database view must be evaluated before the filters are applied to avoid poor performance. Normally, if the dimensional model is set up properly, these techniques are not needed. If you are unable or unwilling to create a dimensional model and use ETL, these may be your only options.
Prepopulate with default values	Prepopulate a record for all levels of all hierarchies with a default value. Use a value of N/A, unknown, or -1 to represent a not found value in a lookup. This will allow the data to flow into the analytic database without being lost. N/A records can later be found and updated as appropriate.

© Hitachi Vantara Corporation 2018. All Rights Reserved

Best Practices for Data Storage

To speed up data retrieval and reporting, there are several steps you can take, These include:

Table 2: Data Storage Recommendations

Recommendation	Details
Use **high-speed input/output storage**	Poor database performance can degrade the performance of your entire analytic project, so you will want your data to reside on high-speed input/output storage. Use a physically mounted drive or storage area network (SAN) with fiber channel in the same data center. Avoid using a virtual machine for a database unless you can address data storage concerns.
Choose a different database platform from the one provided with Pentaho	The database provided with Pentaho is for metadata around a Pentaho object, not for reporting data. Do not use this database for your reporting data. Instead, choose a database platform better suited for this workload.
Modify **memory and kernel parameters** before loading data	Most out-of-the-box DBMS configurations will only perform well on very small or demonstrative datasets. Therefore, avoid using an out-of-the-box configuration for your reporting data DBMS. Instead, modify your memory and kernel parameters soon after installation before you load in your data.

Related Information

Here are some links to information that you may find helpful while using this best practices document:

- Pentaho Components Reference
- Recommendations for database tuning
- Star schemas

© Hitachi Vantara Corporation 2018. All Rights Reserved

Finalization Checklist

This checklist is designed to be added to any implemented project that uses this collection of best practices, to verify that all items have been considered and reviews have been performed.

Name of the Project:____

Date of the Review:____

Name of the Reviewer:____

Item	Response	Comments
Did you set up a dimensional model design?	YES_______ NO_______	
Did you optimize the database server and instance for analytic workloads?	YES_______ NO_______	
Did you apply standard indexing to your database?	YES_______ NO_______	
Did you create indexes on all primary keys of a dimension and all foreign keys in a fact table?	YES_______ NO_______	
Did you create indexes for each level of each hierarchy in all dimensions of all cubes of all schemas?	YES_______ NO_______	
Did you define all tables in the schema as database tables?	YES_______ NO_______	
Did you prepopulate a record for all levels of all hierarchies with a default value?	YES_______ NO_______	
Did you set up high-speed input/output storage?	YES_______ NO_______	
Did you store your reporting data in a database platform other than the Pentaho default?	YES_______ NO_______	
Did you modify your memory and kernel parameters before loading in data?	YES_______ NO_______	

© Hitachi Vantara Corporation 2018. All Rights Reserved