Pentaho Aggregation Designer User Guide

This document supports Pentaho Business Analytics Suite 4.8 GA and Pentaho Data Integration 4.4 GA, documentation revision October 31, 2012.

Help and Support Resources

If you have questions that are not covered in this guide, or if you would like to report errors in the documentation, please contact your Pentaho technical support representative.

Support-related questions should be submitted through the Pentaho Customer Support Portal at http://support.pentaho.com.

For information about how to purchase support or enable an additional named support contact, please contact your sales representative, or send an email to sales@pentaho.com.

For information about instructor-led training on the topics covered in this guide, visit http://www.pentaho.com/training.

Limits of Liability and Disclaimer of Warranty

The author(s) of this document have used their best efforts in preparing the content and the programs contained in it. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, express or implied, with regard to these programs or the documentation contained in this book.

The author(s) and Pentaho shall not be liable in the event of incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of the programs, associated instructions, and/or claims.

Trademarks

Pentaho (TM) and the Pentaho logo are registered trademarks of Pentaho Corporation. All other trademarks are the property of their respective owners. Trademarked names may appear throughout this document. Rather than list the names and entities that own the trademarks or insert a trademark symbol with each mention of the trademarked name, Pentaho states that it is using the names for editorial purposes only and to the benefit of the trademark owner, with no intention of infringing upon that trademark.

Company Information

Pentaho Corporation
Citadel International, Suite 340
5950 Hazeltine National Drive
Orlando, FL 32822
Phone: +1 407 812-OPEN (6736)
Fax: +1 407 517-4575
http://www.pentaho.com

E-mail: communityconnection@pentaho.com

Sales Inquiries: sales@pentaho.com

Documentation Suggestions: documentation@pentaho.com

Sign-up for our newsletter: http://community.pentaho.com/newsletter/

Contents

Introduction

This guide provides you with instructions and recommendations for designing aggregate tables for Mondrian ROLAP models. The use of aggregate tables can dramatically improve the query performance of analysis solutions.

Assumptions

This document is written for Database Administrators and consultants who design specific aggregate tables or get recommendations for aggregate tables based on an intelligent adviser algorithm. It is assumed that you, the reader, have a strong understanding of database design and concepts (such as database modeling, SQL security, and performance), and are familiar with aggregate table concepts.

Pentaho Aggregation Designer Overview

The Pentaho Aggregation Designer simplifies the creation and deployment of aggregate tables that improve the performance of your Pentaho Analysis (Mondrian) OLAP cubes. Pentaho Analysis is a pure, relational OLAP engine that works solely with the data stored in your relational database rather than providing its own multidimensional data storage model. This simplifies deployment and data management, but places limitations on performance when working with very large data sets (fact tables with more than 10 million records and/or cubes with a high cardinality of levels and members). To improve performance in these scenarios, Pentaho Analysis supports aggregate tables. Aggregate tables coexist with the base fact table and contain pre-aggregated measures built from the fact table. This improves performance by enabling the Mondrian engine to fulfill certain summary level queries from the smaller aggregate table versus aggregating a large number of individual facts from the base fact table.
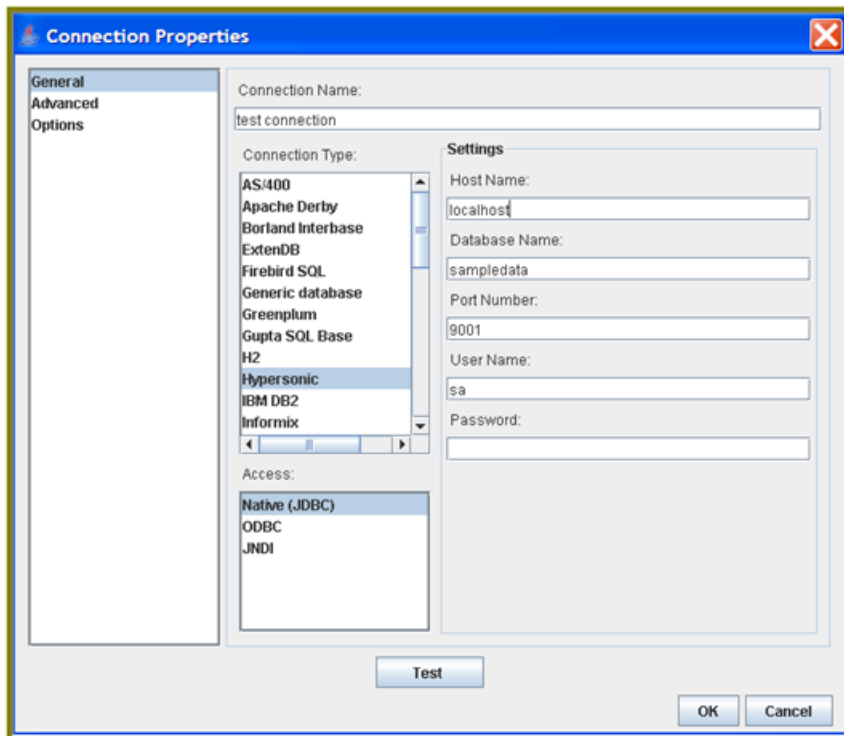
The Pentaho Aggregation Designer provides you with a simple interface that allows you to create aggregate tables from levels within the dimensions you specify. Based on these selections, the Aggregation Designer generates the Data Definition Language (DDL) for creating the aggregate tables, the Data Manipulation Language (DML) for populating them, and an updated Mondrian schema which references the new aggregate tables. If you are unfamiliar with aggregate table design concepts, the Aggregation Designer also includes an intelligent adviser that evaluates the structure and cardinality of your OLAP cube and recommends some initial aggregate tables to create for improving performance.

The components of the Pentaho Aggregation Designer workspace are shown below:
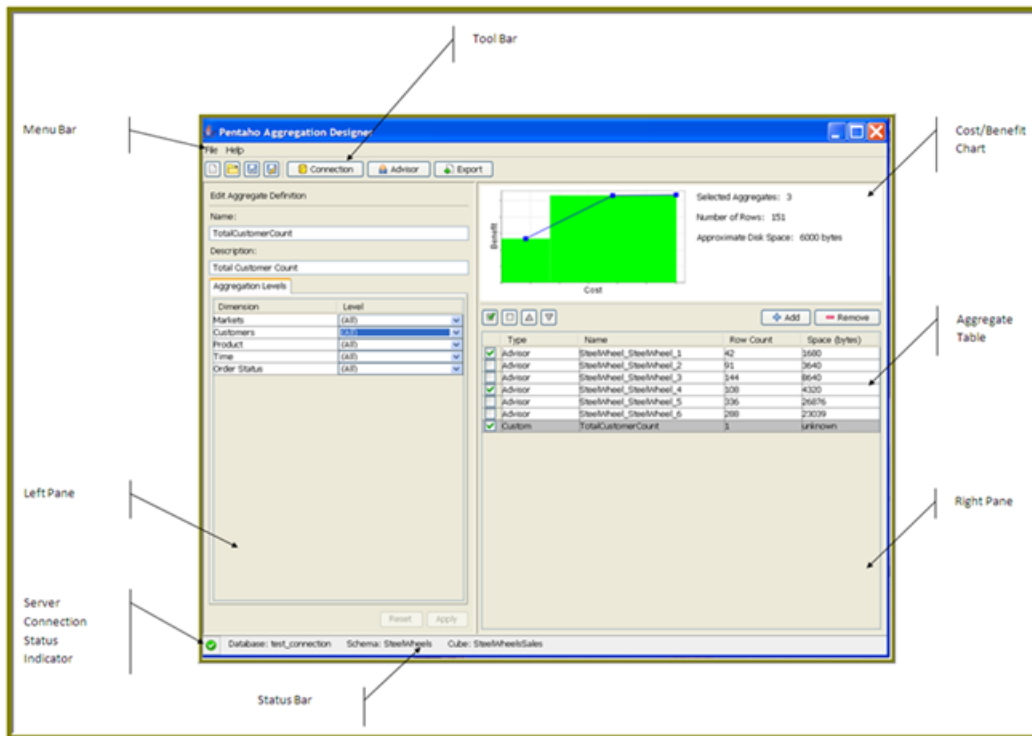
Defining the Data Source

To design an aggregate table, you must first establish a connection with your target relational database, then select the OLAP model to optimize. You can connect to any relational database that is supported by Mondrian. In some instances, you may need to define additional parameter-related values for your JDBC driver.

To define a data source connection...

1. In the Pentaho Aggregation Designer tool bar, click **Connection** to open the **Connect to Data Source** dialog box.
2. Click **Configure**. The **Connection Properties** dialog box appears.
3. In the **Connection Name** field, enter a name for your connection; this is a free-text field. A connection name uniquely defines a connection.
4. In the **Connection Type** list, select a database.
5. In the **Access** list, keep the default choice, which should be **Native (JDBC)**.
6. In the **Settings** section, type the host name of the database server into the **Host Name** field. In the **Database Name** field, type the name of the database you're connecting to. In the **Port Number** field, enter the TCP port number. Optionally, in the **User Name** and **Password** fields, type the user name and password used to connect to the database.
7. Click **Test**.
   If the settings you typed in are correct, a success message appears.
8. Click **OK**.

Adding a JDBC Driver

Before you can connect to a data source in any Pentaho server or client tool, you must first install the appropriate database driver. Your database administrator, Chief Intelligence Officer, or IT manager should be able to provide you with the proper driver JAR. If not, you can download a JDBC driver JAR file from your database vendor or driver developer's Web site. Once you have the JAR, follow the instructions below to copy it to the driver directories for all of the Business Analytics components that need to connect to this data source. See the *Compatibility Matrix: Supported Components* in any of the Installation guide for current version numbers.

> **Note:** Microsoft SQL Server users frequently use an alternative, non-vendor-supported driver called JTDS. If you are adding an MSSQL data source, ensure that you are installing the correct driver.

Backing up old drivers

You must also ensure that there are no other versions of the same vendor's JDBC driver installed in these directories. If there are, you may have to back them up and remove them to avoid confusion and potential class loading problems. This is of particular concern when you are installing a driver JAR for a data source that is the same database type as your Pentaho solution repository. If you have any doubts as to how to proceed, contact your Pentaho support representative for guidance.

Installing JDBC drivers

Copy the driver JAR file to the following directories, depending on which servers and client tools you are using (Dashboard Designer, ad hoc reporting, and Analyzer are all part of the BA Server):

👉 **Note: For the DI Server:** before copying a new JDBC driver, ensure that there is not a different version of the same JAR in the destination directory. If there is, you must remove the old JAR to avoid version conflicts.

- **BA Server:** `/pentaho/server/biserver-ee/tomcat/lib/`
- **Enterprise Console:** `/pentaho/server/enterprise-console/jdbc/`
- **Data Integration Server:** `/pentaho/server/data-integration-server/tomcat/webapps/pentaho-di/WEB-INF/lib/`
- **Data Integration client:** `/pentaho/design-tools/data-integration/libext/JDBC/`
- **Report Designer:** `/pentaho/design-tools/report-designer/lib/jdbc/`
- **Schema Workbench:** `/pentaho/design-tools/schema-workbench/drivers/`
- **Aggregation Designer:** `/pentaho/design-tools/agg-designer/drivers/`
- **Metadata Editor:** `/pentaho/design-tools/metadata-editor/libext/JDBC/`

👉 **Note:** To establish a data source in the Pentaho Enterprise Console, you must install the driver in both the Enterprise Console and the BA Server or Data Integration Server. If you are just adding a data source through the Pentaho User Console, you do not need to install the driver to Enterprise Console.

Restarting

Once the driver JAR is in place, you must restart the server or client tool that you added it to.

Connecting to a Microsoft SQL Server using Integrated or Windows Authentication

The JDBC driver supports Type 2 integrated authentication on Windows operating systems through the **integratedSecurity** connection string property. To use integrated authentication, copy the **sqljdbc_auth.dll** file to all the directories to which you copied the JDBC files.

The **sqljdbc_auth.dll** files are installed in the following location:

```
<installation directory>\sqljdbc_<version>\<language>\auth\
```

👉 **Note:** Use the **sqljdbc_auth.dll** file, in the x86 folder, if you are running a 32-bit Java Virtual Machine (JVM) even if the operating system is version x64. Use the **sqljdbc_auth.dll** file in the x64 folder, if you are running a 64-bit JVM on a x64 processor. Use the **sqljdbc_auth.dll** file in the IA64 folder, you are running a 64-bit JVM on an Itanium processor.

Adding a Simple JNDI Data Source For Design Tools

Pentaho provides a method for defining a JNDI connection that exists only for locally-installed client tools. This is useful in scenarios where you will be publishing to a BA Server that has a JNDI data source; if you establish the same JNDI connection on your client tool workstation, you will not need to change any data source details after publishing to the BA Server. Follow the directions below to establish a simple JNDI connection for Pentaho client tools.

1. Navigate to the **.pentaho** directory in your home or user directory.

   For a user name of **fbeuller**, typically in Linux and Solaris this would be `/home/fbeuller/.pentaho/`, and in Windows it would be `C:\Users\fbeuller\.pentaho\`

2. Switch to the `~/.pentaho/simple-jndi/` subdirectory. If it does not exist, create it.
3. Edit the **default.properties** file found there. If it does not exist, create it now.

4. Add a data source by declaring a JNDI name followed by a forward slash, then a JNDI parameter and its proper value.

Refer to *Simple JNDI Options* on page 8 for more information on parameter options.

```
SampleData/type=javax.sql.DataSource
SampleData/driver=org.hsqldb.jdbcDriver
SampleData/user=pentaho_user
SampleData/password=password
SampleData/url=jdbc:hsqldb:mem:SampleData
```

5. Save and close the file.

You now have a global Pentaho data source that can be used across all of the client tools installed on this machine. You must restart any running Pentaho program in order for this change to take effect.

Simple JNDI Options

Each line in the data source definition must begin with the JNDI name and a forward slash (/), followed by the required parameters listed below.

Parameter	Values
type	**javax.sql.DataSource** defines a JNDI data source type.
driver	This is the driver class name provided by your database vendor.
user	A user account that can connect to this database.
password	The password for the previously declared user.
url	The database connection string provided by your database vendor.

```
SampleData/type=javax.sql.DataSource
SampleData/driver=org.hsqldb.jdbcDriver
SampleData/user=pentaho_user
SampleData/password=password
SampleData/url=jdbc:hsqldb:mem:SampleData
```

Defining Additional Parameters

If you must define additional parameters for your JDBC driver, or if you want to enter your server settings manually, follow the instructions below:

1. Click **Options** in the left panel..

2. Enter the parameter name and value for the settings you need to specify. For example, **PORT** (parameter name), **1025** (parameter value).

3. Click **Test** when your settings are entered.
A success message appears if everything was typed in correctly.

4. Click **OK**.

Selecting a Model

After defining your data source, you must select the cube you want to use for defining and building aggregate tables.

To select a model...

1. In the **Connect to Data Source** dialog box, under **OLAP Model**, select **Mondrian Schema File**.
2. Click the ellipsis **(...)** to display a file dialog box.
3. Browse to locate and select your Mondrian schema file (SteelWheels.mondrian.xml if using sample data), then click **OK**.
4. Click **Apply**. The Cube list is populated with a list of cubes defined in your schema.
5. Select the Mondrian cube you want to optimize, then click **Connect**.
    When the Pentaho Aggregation Designer establishes a connection, it runs several validation tests to ensure that your database structure is ready to support aggregate tables. A validation summary dialog box appears with a list of test results. If you see an error message, contact your database administrator.

Getting Recommendations Using Aggregate Advisor

If you are unfamiliar with aggregate table design and need help creating aggregates to optimize a cube, you can rely on the Aggregate Advisor to provide you with a list of recommendations. The Pentaho Aggregation Designer uses your schema file and the data in your database to create aggregate definitions.
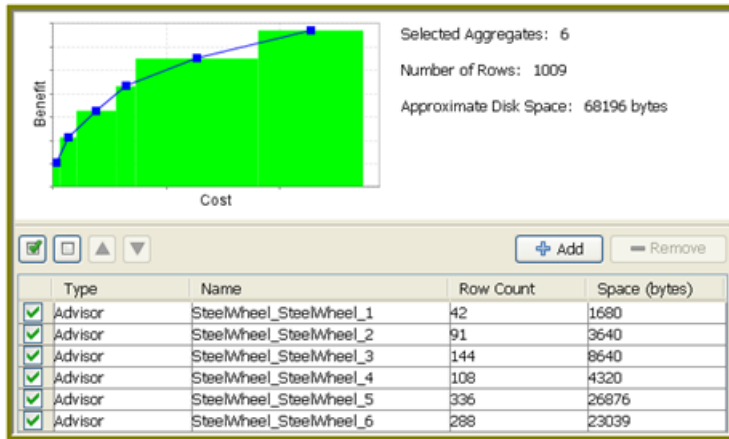
To display recommended aggregates...

1. In the Pentaho Aggregation Designer toolbar, click **Advisor**.
2. Specify your **Advisor Input Parameters**.

Setting	Description
Max Aggregates	Allows you to specify the maximum number of aggregates you want the Advisor to recommend.
Max Time to Run	Allows you to specify the maximum amount of time (in seconds) you want the Advisor to run before making recommendations. *Note*: Allowing the Advisor to run for longer periods of time allows for more potential recommendations to be evaluated and results in more accurate recommendations.

3. Click **Recommend**.
   The Advisor runs for a few seconds before it displays an initial list of recommended aggregates. The Advisor is designed to keep running until it finds an optimal solution. If you stop the Advisor prematurely, the Advisor returns the best set of recommendations it has found up to the point when it was stopped.

Customizing Aggregates

When you select an aggregate, the Pentaho Aggregation Designer pulls information from the schema file to display its dimensions and levels. You can modify any aggregate the Pentaho Aggregation Designer recommends, customizing it for your needs. You can also create an aggregate from scratch and delete aggregates you do not want.

Impact summary

The impact summary in the lower right pane provides you with information on the estimated impact for creating all of the currently selected aggregates. This summary includes the number of aggregate tables that will be created, the estimated number of rows contained in those tables, and the estimated amount of space it will occupy on the hard drive. The impact summary is automatically updated as you select and deselect aggregates from the list of proposed aggregates.

Cost/benefit chart

The Cost/Benefit chart provides a high-level comparison of the benefit of all currently selected aggregates relative to their estimated cost. The benefit scale represents the relative number of queries that can be fulfilled by an aggregate table versus having to be retrieved from the base fact table. The cost scale is an indicator of the impact in terms of number of tables and disk space needed to create the selected aggregate recommendations.

Saving your design

The Pentaho Aggregation Designer allows you to save all aggregate-related data (custom- or advisor-created) in your workspace at any time. Saving ensures that all of the data (your designs) in the workspace is retained; you are saving the state of your workspace as an XML file in a location you specify. To save, go to the **File** menu and click **Save As**. To open a saved file, go to the **File** menu and click **Open**, then navigate to the design you previously saved.

Customizing an aggregate

To customize an aggregate...

1. In the **Pentaho Aggregation Designer**, click on an aggregate in the proposed aggregate list to select it.

   Note: When you modify an aggregate created using the Advisor, the aggregate becomes a Custom aggregate as indicated by the Type column in the proposed aggregate list.

2. In the left pane, you can (optionally) modify the **Name** and **Description** for your custom aggregate.

3. In the **Aggregation Levels** tab, click the down arrows to make changes to the hierarchy and levels associated with the aggregate definition you are customizing.

4. Click **Apply** .
   The Pentaho Aggregation Designer updates the proposed aggregate list, cost/benefit chart, and impact summary.

Adding Aggregates

To add an aggregate...

1. In the right pane of the Pentaho Aggregation Designer, click **Add**.

2. In the left pane, type a **Name** and **Description** for your new aggregate.

3. Under **Level**, click the down arrows to define the hierarchy and levels associated with the aggregate you are creating.

4. Click **Apply**.
   Your aggregate is added to the aggregate list.

Deleting Aggregates

To delete an aggregate, select it from the proposed aggregate list and click **Remove**.

Exporting Aggregates

The Pentaho Aggregation Designer allows you to preview the DDL, DML, and schema (for relational databases) outputs before you build aggregate tables. You can also save the outputs and edit them later. If you are using OLAP, DML is the only available output.

To preview the DDL and DML outputs...

1. Select the aggregates that have DML/DDL output you want to preview.
2. In the Pentaho Aggregation Designer toolbar, click **Export**.
3. In the Execute and Publish dialog box, click **Preview**.
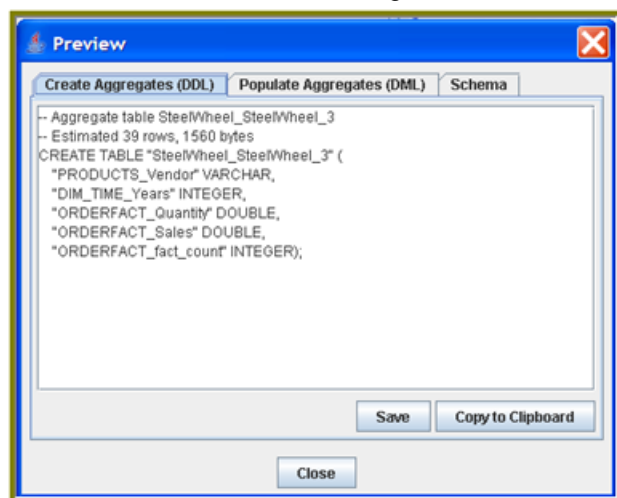
4. Click **Copy to Clipboard** or **Save** to retain the output.
5. If you examine the DDL/DML outputs and are satisfied with the results, you can allow the Pentaho Aggregation Designer to build (Execute/Publish) the aggregate tables. Follow the instructions for publishing and exporting included in the **Execute and Publish** dialog box.
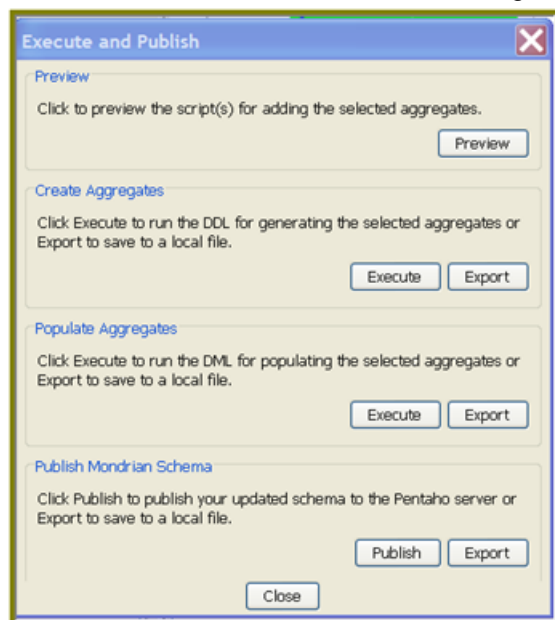
Glossary of Terms

Below is a list of terms used in this document.

Aggregate

Definitions for aggregate tables that help optimize a cube; also, summarized data.

Aggregate Tables

Coexists with the base fact table, and contains pre-aggregated measures built from the fact table. It is registered in Mondrian's schema, so that Mondrian can choose whether to use the aggregate table rather than the fact table, if applicable for a particular query.

Aggregation

The process of merging multiple data values into one value. For example, sales data collected daily can then be aggregated to the week level, the week data could be aggregated to the month level, and so on. The data can then be referred to as aggregate data. Aggregation and summarization are synonyms, as are aggregate data and summary data.

Data Definition Language (DDL)

Originally a subset of SQL, this language defines data structures, including rows, columns, tables, indexes, and database specifics such as file locations. DDL SQL statements are more a part of the database management system, and have large differences between SQL implementations.

Mondrian Schema

Defines a multi-dimensional database. A Mondrian schema contains a logical model, consisting of cubes, hierarchies, and members, and a mapping of this model onto a physical model. The logical model consists of the constructs used to write queries in the MDX language: cubes, dimensions, hierarchies, levels, and members. The physical model is the source of the data presented through the logical model. It is typically a star schema, which is a set of tables in a relational databases.

Relational Online Analytic Processing (ROLAP)

An alternative to MOLAP (Multidimensional OLAP) technology. While both ROLAP and MOLAP analytic tools are designed to allow analysis of data through the use of a multidimensional data model, ROLAP differs significantly in that it does not require the pre-computation and storage of information. Instead, ROLAP tools access the data in a relational database and generate SQL queries to calculate information at the appropriate level when an end user requests it. With ROLAP, it is possible to create additional database tables (summary tables or aggregations) which summarize the data at any desired combination of dimensions.

Snowflake Schema

A way of arranging tables in a relational database such that the entity relationship diagram resembles a snowflake in shape. At the center of the schema are fact tables which are connected to multiple dimension tables. Thus a snowflake simplifies to a star schema when relatively few dimensions are used. The star and snowflake schemas are most commonly found in data warehouses where the speed of data retrieval is more important than the speed of insertion. As such, these schemas are not normalized much, and are frequently left in third normal form or second normal form.