



# Sizing Study for Pentaho Open Source Business Intelligence Suite on Sun™ Platforms

*Krishnan Shankar*

*June 2009*

*Sun Microsystems, Inc.*

**BigAdmin\***

*This article was published by BigAdmin at:  
[http://www.sun.com/bigadmin/features/articles/pentaho\\_sizing.jsp](http://www.sun.com/bigadmin/features/articles/pentaho_sizing.jsp)  
To keep track of the latest content published by BigAdmin, subscribe  
to the BigAdmin newsletter: <http://www.sun.com/bigadmin/newsletter/>.*

Copyright © 2009 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, U.S.A. All rights reserved.

*U.S. Government Rights - Commercial software. Government users are subject to the Sun Microsystems, Inc. standard license agreement and applicable provisions of the FAR and its supplements. Use is subject to license terms. This distribution may include materials developed by third parties.*

*Parts of the product may be derived from Berkeley BSD systems, licensed from the University of California. UNIX is a registered trademark in the U.S. and in other countries, exclusively licensed through X/Open Company, Ltd. X/Open is a registered trademark of X/Open Company, Ltd.*

*AMD, Opteron, the AMD logo, the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices.*

*Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation in the United States and other countries.*

*Sun, Sun Microsystems, the Sun logo, Java, JDBC, MySQL, Solaris, Sun BluePrints, Sun Fire, and SunSolve are trademarks or registered trademarks of Sun Microsystems, Inc. or its subsidiaries in the United States and other countries.*

*This product is covered and controlled by U.S. Export Control laws and may be subject to the export or import laws in other countries. Nuclear, missile, chemical biological weapons or nuclear maritime end uses or end users, whether direct or indirect, are strictly prohibited. Export or reexport to countries subject to U.S. embargo or to entities identified on U.S. export exclusion lists, including, but not limited to, the denied persons and specially designated nationals lists is strictly prohibited.*

*DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.*

# Table of Contents

1. Introduction .....	4
1.1 Overview .....	4
1.2 How the Report Is Organized .....	4
1.3 Terminology .....	4
1.4 Key Performance Metrics.....	5
1.5 Project Evaluation and Requirements .....	5
1.6 Reference Links.....	5
2. Testing Activity .....	6
2.1 High-Level Description of Tests.....	6
2.1.1 Pentaho ETL.....	6
2.1.2 Pentaho Reporting .....	6
2.1.3 Pentaho Mondrian OLAP Server.....	6
2.2 Deployment Topology .....	7
2.3 Hardware and Software Specifications.....	7
3. Performance Results.....	8
3.1 Pentaho ETL.....	8
3.2 Pentaho Report Designer .....	9
3.2.1 Query 1: Simple Join of Two Tables on a 1-GB Data Warehouse .....	9
3.2.2 Query 2: Single Table With WHERE Clause on a 30-GB Data Warehouse.....	10
3.3 Pentaho Mondrian OLAP Server.....	11
3.4 Interpreting the Test Results.....	11
4. Summary of Results.....	12
4.1 Summary of Performance Data.....	12
4.2 Analysis and Conclusions .....	12
4.3 Recommended Hardware .....	13
5. Additional Information .....	13
5.1 Standard Tuning Parameters.....	13
5.2 Workload Details.....	14
6. Appendix .....	14
6.1 Detailed Deployment of Application Software.....	14
6.1.1 Configure VNC Software (Optional).....	14
6.1.2 Configure Apache Tomcat Server and the MySQL Connector Driver.....	15
6.1.3 Configure the Data Warehouse Workload Kit and the DBGEN Data Generator Tool.....	15
6.1.4 Configure the MySQL Database.....	15
6.1.5 Configure Pentaho ETL.....	15
6.1.6 Configure Pentaho BI-CE 2.0.0 Server .....	16
6.1.7 Configure Pentaho Report Designer.....	16
6.1.8 Configure Mondrian OLAP Server .....	16
7. Acknowledgements.....	16
8. Product Information.....	16
8.1 Sun Products.....	16
8.2 Pentaho Offerings.....	17
9. For More Information.....	17
10. Licensing Information.....	19

# 1. Introduction

## 1.1 Overview

This paper describes a sizing study of Pentaho Open Source Business Intelligence (OSBI) Suite on a Sun platform, including tests conducted with application tools from Pentaho. The purpose is to provide this guide as a reference point when users are choosing a suitable Sun platform for their Business Intelligence (BI) applications.

This test used the Solaris 10™ 5/08 OS, MySQL™ database, and a Sun Fire™ X4540 server. For more information on the Sun and Pentaho technologies in this study, please see Section 8 of this article, “Product Information,” and Section 9, “For More Information.”

## 1.2 How the Report Is Organized

The requirements of the project are listed, followed by the hardware and software requirements. Next, the deployment of Pentaho application tools and the setup of the tests are described. The performance results are shown in tabular form.

The report ends with an analysis of the results and a recommendation of hardware choice for the tested workloads.

## 1.3 Terminology

*Table 1: Terminology*

<i>Term</i>	<i>Definition</i>
Virtual CPU	A processing granularity (for example, lightweight process [LWP]) that shares a pipeline of a core in a physical CPU
Transformation	A process that describes to the Pentaho Kettle extraction, transformation, and loading (ETL) application how to modify the data from source to sink
Cube, Measures, Dimension, Levels	Pentaho Mondrian Server's Multidimensional Expressions (MDX) query language constructs

## 1.4 Key Performance Metrics

Table 2: Key Performance Metrics

<i>Performance Metric</i>	<i>Definition</i>
Throughput	GB/hour of data loaded for Pentaho ETL; GB/hour of data fetched for Pentaho Reporting
Response Time	Query time taken to process and retrieve output
CPU Utilization	CPU utilization information obtained from the Solaris 10 Operating System <code>vmstat</code> tool
Memory Utilization	Total memory less the computed “free” value shown in <code>vmstat</code> output
Average I/O Utilization	Information derived from the <code>kr/s</code> column, the <code>kw/s</code> column, or both from the Solaris 10 OS <code>iostat</code> tool

## 1.5 Project Evaluation and Requirements

This section describes the evaluation of the project.

- The scalability focus is vertical and for two to four processors on a non-cluster system. No high availability (HA) or failover is required. The configuration is sized for a processor utilization of 20% to 80%, with 10% headroom for unexpected load.
- There are no baselines for expected database growth and the peak load. These vary according to the customer's environment.
- The user mix is of an online analytical processing (OLAP) nature rather than an online transaction processing (OLTP) nature. The expectations are that a few hundred users would operate frequently on medium-to-large sized data (1–10 Gbyte to 30–100 Gbyte), and only rare cases of concurrent users querying on large data would occur. Metrics are defined for ETL, report queries, and Mondrian OLAP queries (Gbyte/hour and response time).
- System configuration requirements are required for medium-to-large ETL loads or for acceptable response times for medium report and OLAP queries, at both low and high processor utilizations.

## 1.6 Reference Links

Refer to Section 9, “For More Information,” at the end of this article for references.

## 2. Testing Activity

### 2.1 High-Level Description of Tests

#### 2.1.1 Pentaho ETL

1. Download a data warehouse workload kit (see Section 9, “For More Information”), and create a schema.
2. Build the DBGEN data generator tool from TPC.
3. Download and install Pentaho Data Integration Tool 3.1.0.
4. Create Pentaho Kettle transformations through the Pentaho Spoon transformation tool, and run the pan tool to load data into the database (see Section 9, “For More Information”).
5. Repeat with flat-file input loads of 1 to 100 Gbyte, varying system resources each time.
6. Measure the throughput, and monitor system resource utilization.

#### 2.1.2 Pentaho Reporting

1. Download and install Pentaho Report Designer (RD) 1.7.1.
2. Download and install Pentaho BI Server Community Edition (BI-CE) v2.0.0.
3. Create reports using the Pentaho RD GUI, and publish reports onto the BI-CE Server.
4. Perform repeated runs of typical user queries, varying system resources each time.
5. Measure the throughput, and monitor system resource utilization.

#### 2.1.3 Pentaho Mondrian OLAP Server

1. Download and install Apache Tomcat 5.5.27.
2. Download and install Pentaho Mondrian Server 3.0.4.
3. Configure Mondrian Server with Apache Tomcat.
4. Install Mondrian Server's demo database schema and tables.
5. Run the demo queries, and vary system resources each time.
6. Measure the response time, and monitor system resource utilization.

## 2.2 Deployment Topology

This section describes the hardware and software used.

The Pentaho applications and the MySQL™ database run on the same physical server. A Microsoft Windows client is used to create transformations through the Pentaho ETL Spoon GUI. Network connectivity is through a 1 Gbps, 10BaseT Cisco switch. There are no private interconnects and load balancers. Storage comprises internal, SATA-II disks.

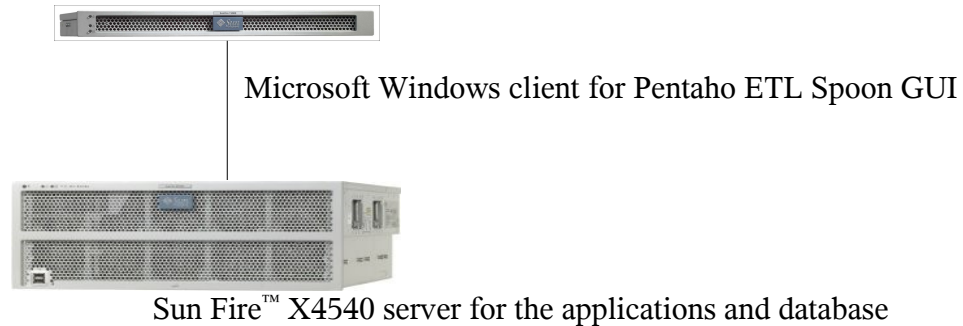


Figure 1: Deployment Topology

## 2.3 Hardware and Software Specifications

Table 3: Hardware and Software Specifications

Application Component	Software (With Version Number)	Hardware Configuration				
		System Model	Processor	Memory	Network	Storage
Applications and database are hosted on the Sun Fire X4540 server	<ul style="list-style-type: none"> <li>Pentaho Data Integration 3.1.0,</li> <li>Pentaho Report Designer 1.7.1,</li> <li>Pentaho Mondrian Server 3.0.4;</li> <li>Pentaho BI Server Community Edition 2.0.0,</li> <li>MySQL Community Edition v5.1.30,</li> <li>data warehouse workload kit,</li> <li>Apache Tomcat 5.5.27,</li> <li>Solaris 10 5/08 OS</li> </ul>	Sun Fire X4540 Server	2 x quad-core AMD Opteron processor 2356 @ 2300 MHz each;  Cache 2 GB (512 MB per core)	64 GB	1 GB NIC	48 pluggable, SATA-II, internal disks of 930 GB each

## 3. Performance Results

### 3.1 Pentaho ETL

Table 4 shows the throughput achieved when using Pentaho ETL Kettle transformations to load a flat file into the data warehouse schema of a MySQL 5.1.30 database.

- Disk: The database is on a different disk from the application and input files. No RAID is configured.
- Network: A 1-Gbyte NIC and no private interconnects were used.
- Database: The default MySQL 5.1.30 64-bit binary was used.
- Transform: A simple join and a WHERE clause were used.
- Process heap size: A heap size of 2,048 m was used.

*Table 4: Throughput When Using Pentaho ETL Kettle Transformations to Load a Flat File*

<i>Virtual Processors</i>	<i>Data Loaded Into the DB (GB)</i>	<i>Load Time From CSV Files (Minutes)</i>	<b><i>Load Throughput (GB/hour)</i></b>	<i>Processor Utilization (%)</i>	<i>Memory Utilization (%)</i>	<i>Max. I/O Utilization (kw/s)</i>
1	1	23	<b>2.6</b>	100	18	2323
2	1	19	<b>3.2</b>	63	17	2435
	10	188	<b>3.2</b>	66	20	2531
4	1	18	<b>3.3</b>	37	20	2442
	10	163	<b>3.7</b>	42	20	2573
	30	485	<b>3.7</b>	38	20	2558
	100	1637	<b>3.7</b>	33	23	2596
8	1	18.5	<b>3.2</b>	25	23	2108
	10	174	<b>3.5</b>	30	20	2515

## 3.2 Pentaho Report Designer

Table 5 and Table 6 show the response times achieved when Pentaho BI-CE Server runs a large query report published from the Pentaho Report Designer tool.

- Disk: The database is on a different disk from the application and input files. No RAID is configured.
- Network: A 1-Gbyte NIC and no private interconnects were used.
- Database: The default MySQL 5.1.30 64-bit binary was used.

### 3.2.1 Query 1: Simple Join of Two Tables on a 1-GB Data Warehouse

a.) Performance of the report (External Action Sequence [xaction] with SQLComponent and report component):

- Output: Two million rows at 256 bytes per row
- Amount of data retrieved: 488.28 Mbyte

*Table 5: Performance of the Report (Xaction With SQLComponent and Report Component)*

<i>Virtual Processors</i>	<i>Fetch Time From DB (Minutes)</i>	<i>Fetch Throughput (GB/hour)</i>	<i>Processor Utilization (%)</i>	<i>Memory Utilization (%)</i>
1	18	<b>1.59</b>	100	16.4
2	14.5	<b>1.97</b>	67	16.3
4	10	<b>2.86</b>	45	16.4
8	12.2	<b>2.35</b>	27	16.5

b.) Performance of the platform running the query (xaction with SQLComponent):

- Output: 200,000 rows.
- For 1, 2, and 4 CPUs, the fetch time was 3.5 minutes, and data retrieved was 0.82 GB/hour.
- For 8 CPUs, the fetch time was 4.3 minutes, and data retrieved was 0.67 GB/hour.
- Processor utilization was 100%, 52%, 26%, and 13% for 1, 2, 4, and 8 CPUs, respectively.
- Memory utilization was negligible.
- Querying for beyond 300,000 rows yielded long waits.

c.) Performance of the raw query (select query on shell command line):

- Output: 200,000 rows

*cpu[s]*      *Fetch time (mm:ss)*      *cpu utilization*

1	6:51	100%
2	6:42	52%
4	6:58	26%
8	7:14	13%

### 3.2.2 Query 2: Single Table With WHERE Clause on a 30-GB Data Warehouse

a.) Performance of the report (xaction with SQLComponent and report component):

- Output: 2 million rows at 146.85 bytes per row, from a 24-million row table
- Amount of data retrieved: 280.09 Mbyte

*Table 6: Performance of the Report (Xaction With SQLComponent and Report Component)*

<i>Virtual Processors</i>	<i>Fetch Time From DB (Minutes)</i>	<i>Fetch Throughput (GB/hour)</i>	<i>Processor Utilization (%)</i>	<i>Memory Utilization (%)</i>
1	19.5	<b>0.84</b>	100	18.5
2	19.3	<b>0.85</b>	60	18.7
4	18.3	<b>0.9</b>	34	17.8
8	18.6	<b>0.88</b>	29	17.5

b.) Performance of the platform running the query (xaction with SQLComponent):

- Output: 200,000 rows.
- The fetch time was less than 2 seconds for up to 8 CPUs.

c.) Performance of the raw query (select query on shell command line):

- Output: 200,000 rows

*cpu[s] Fetch time (mm:ss) cpu utilization*

1	9:15	8%
2	8:54	4%
4	9:00	2%
8	8:58	1%

### 3.3 Pentaho Mondrian OLAP Server

Table 7 shows the response times achieved when the Pentaho Mondrian Server runs a set of queries against a small database.

- Disk: The database is on a different disk from the application and input files. No RAID is configured.
- Network: A 1-Gbyte NIC and no private interconnects were used.
- Database: The default MySQL 5.1.30 64-bit binary was used.
- Query: FoodMart Demo queries and a database size of 135 Mbyte were used.
- Caching: Caching was disabled. The database and application server were restarted on each run.

*Table 7: Response Times When Pentaho Mondrian Server Runs Queries Against a Small Database*

<i>Virtual Processors</i>	<i>Fetch Time From Database (Seconds)</i>	<i>Processor Utilization (Max. %)</i>	<i>Memory Utilization (%)</i>
1	<b>588</b>	100	19
2	<b>522</b>	54	19
4	<b>502</b>	27	19
8	<b>484</b>	13	19

### 3.4 Interpreting the Test Results

For the Pentaho ETL and Pentaho Report queries, the throughput measure is the data loaded per hour in Gbyte. The load time is measured from the point where the first data warehouse table starts loading until the point at which the last data warehouse table completes loading. The ETL data load is from flat files.

For Pentaho Reporting, the response time is the metric to be measured. It is calculated as the sum of the sizes of the rows (in the join query) times the number of rows retrieved, and extrapolated to an hour.

For the Pentaho Mondrian Server MDX queries, the response time is the number of Gbyte fetched per hour. It includes the time from when the first FoodMart demo query is clicked in the browser until the point at which the results for the last query are retrieved and displayed on the screen.

## 4. Summary of Results

### 4.1 Summary of Performance Data

For Pentaho ETL, different sizes of data loads on a number of virtual processors constitute the scenarios. These are some typical workloads for medium-to-large scale software vendors. The processor, memory, and I/O utilizations are shown.

For Pentaho Reporting, one query scenario is a join of two tables, while the other is a single table query. These might be workloads for medium-to-large vendors using huge databases. The tables are chosen from large databases. The throughput is shown in Gbyte/hour. The tables are not indexed. The processor and memory utilizations are shown.

For Pentaho Mondrian, running the FoodMart demo is the only scenario. This might be a realistic workload for small-to-medium vendors. The response time is shown for various number of processors, and should be interpreted in the context of the workload. The processor utilization is monitored. Memory and I/O are not critical. The table shows virtual processors and response times. The processor and memory utilizations are shown.

### 4.2 Analysis and Conclusions

This section discusses the behavior of the system under typical workloads on the Sun hardware platform. It does not encompass all the types of workloads that a BI user generates.

For Pentaho ETL:

- Four processors yielded the maximum throughput regardless of the input load size.
- A 10-Gbyte load yielded the highest throughput.
- An 8-CPU setup yielded the lowest “per cpu” utilization at 3%.
  
- For a 1-Gbyte database, the average throughput was 3.1 Gbyte/hour.
- For a 10-Gbyte database, the average throughput was 3.5 Gbyte/hour.
- For a 30-Gbyte database, the average throughput was 3.7 Gbyte/hour.
- For a 100-Gbyte database, the average throughput was 3.5 Gbyte/hour.
  
- Average throughput for one processor was 2.6 Gbyte/hour.
- Average throughput for two processors was 3.2 Gbyte/hour.
- Average throughput for four processors was 3.6 Gbyte/hour.
- Average throughput for eight processors was 3.3 Gbyte/hour.
  
- The memory utilization was 17–23 % across all load types and processors.
- The I/O utilization was 16–19 % across all load types and processors.
  
- Launching startup copies of the input, or choosing “Running in Parallel” did not yield added performance.
- The application and database were kept on different disks in order to prevent an I/O bound constraint.

For Pentaho Reporting, performance of the report (xaction with SQLComponent and report component) yielded a maximum throughput of 2.86 Gbyte/hour at four processors for the join query on a 1- Gbyte database. Throughput, CPU, and memory utilization all scaled to four processors on both the queries.

For Pentaho Mondrian OLAP Server, eight processors yielded the lowest fetch time of 484 seconds. The processor utilization scaled with the number of processors. The memory utilization was constant.

### 4.3 Recommended Hardware

For a small ETL load, queries for reports of small complexity, and a situation where large response times are acceptable for Report and OLAP queries beyond 70 % CPU utilization, the recommended hardware is **a server with 1 quad-core Intel or AMD Opteron processor at 2.5 GHz and 16 Gbyte of memory.**

For a medium-to-large ETL load, queries for reports of medium-to-large complexity, and a situation where low response times are required for Report and OLAP queries below 70 % CPU utilization, the recommended hardware is **a server with 2 quad-core Intel or AMD processors at 2.5 GHz and 16 Gbyte of memory.**

## 5. Additional Information

For Pentaho ETL, the Spoon GUI needs `gtk` libraries, which were not present in the Solaris 10 OS for x86 platforms. A Microsoft Windows client was used to launch the Spoon GUI, in order to create the basic transformations.

For Pentaho Reporting, the Report Burst test requires the Pentaho BI-EE Server (Enterprise Edition v3.0.0 RC2). A partner evaluation license was obtained. After installation of BI-EE, however, the Enterprise Console tool did not start up successfully on the Solaris OS, and hence, the license could not be installed. As a result, the burst test was bypassed.

Use of the Tsung simulator tool was attempted for load simulation (see Section 9, “For More Information”). However, Pentaho BI-CE 2.0.0 outputs special characters in the simulated login sequence that Tsung did not recognize. Also, a large number of concurrent users is not the typical scenario for Pentaho BI. Instead, a small number of users with large OLTP queries is the likely mix. Hence, this test was not pursued.

For Pentaho Mondrian OLAP Server, the demo file `FoodMart.xml` was published to Pentaho BI-CE server from Mondrian Schema Workbench, but the entry did not show up on the Analysis screen.

Apache Tomcat Server and the Pentaho BI-CE server might not start up simultaneously, even if they are configured on different port numbers. To resolve this, the Pentaho team suggested that `CATALINA_HOME` in the BI-CE server's installation script be set explicitly.

### 5.1 Standard Tuning Parameters

For Pentaho ETL, 10 startup copies of the Filter Rows transform was applied. The maximum heap space was set to 2048m. Manage Thread Priority was enabled in the transformation.

For Pentaho Reporting, the maximum heap size was set to 1800m.

For Pentaho Mondrian Server, the heap space was set to 1800m. The cache was effectively disabled by restarting the Mondrian Server and the database prior to each run.

## 5.2 Workload Details

No indexes or constraints were enforced in the data warehouse schema for Pentaho ETL.

For Pentaho Reporting, a 1-Gbyte database and a 30-Gbyte database were considered. Two million rows were retrieved from a simple join of two data warehouse tables and also from a single table fetch.

Query 1:

```
SELECT LINEITEM.QUANTITY, SUPPLIER.NAME
FROM LINEITEM , SUPPLIER
WHERE LINEITEM.SUPPKEY=SUPPLIER.SUPPKEY
LIMIT 2000000
```

Query 2:

```
SELECT PARTKEY
FROM PARTSUPP
WHERE PARTKEY > 0
LIMIT 2000000
```

For Pentaho Mondrian OLAP Server, a large test data set is not available currently. Hence, data in the FoodMart database was increased. This was achieved by first removing the UNIQUE keys, and then adding data by three times in these 12 demo tables: customer, employee, product, salary, store, warehouse, expense\_fact, inventory\_fact\_1997, inventory\_fact\_1998, sales\_fact\_1997, sales\_fact\_1998, and sales\_fact\_dec\_1998.

Increasing data by four times caused a timeout on one virtual processor. Mondrian Server MDX queries consist of a cube, dimensions, and measures. Most queries involve cross-joins and are of simple-to-medium complexity.

## 6. Appendix

### 6.1 Detailed Deployment of Application Software

This section describes the steps to configure the software required to carry out the tests with Pentaho Open Source Business Intelligence Suite.

#### 6.1.1 Configure VNC Software (Optional)

1. As root user, obtain and install the Virtual Network Computing (VNC) software on the Solaris 10 OS (see Section 9, “For More Information”).
2. Point a web browser to the client on the default port 5801.
3. Open a terminal window on the client.

## 6.1.2 Configure Apache Tomcat Server and the MySQL Connector Driver

1. As `root` user, download and uncompress `apache-tomcat-5.5.27.tar.gz`.
2. Optionally, use a terminal window opened through the VNC software.
3. Edit `conf/tomcat-users.xml` and add an admin user:

```
<tomcat-users>
<role rolename="admin"/>
<user username="admin" password="admin" roles="admin,manager"/>
</tomcat-users>
```

4. Download and uncompress the MySQL Connector file, `mysql-connector-java-5.1.7.tar.gz`.
5. Copy `mysql-connector-java-5.1.7-bin.jar` to `apache-tomcat-5.5.27/common/lib`.

## 6.1.3 Configure the Data Warehouse Workload Kit and the DBGEN Data Generator Tool

1. Install a standard data warehouse workload kit, and build the binaries.
2. Create the load transformation (`.ktr`) files. (In this case, they were provided by Pentaho.)
3. Generate 1 Gbyte of data using the DBGEN tool (`dbgen -vFF -s 1`). Avoid using indexes during the load.
4. Install and configure MySQL as described in the next section, and create a data warehouse database.
5. Change the input file path names and the server name in the `.ktr` files. Load data through the transformations.

## 6.1.4 Configure the MySQL Database

1. Log in as user `root`, and create a `mysql` user and group.
2. Download and install MySQL Community Server 5.1.30.
3. Have `/usr/local/mysql/bin`, `/usr/local/mysql/libexec`, and a MySQL Connector jar file in the `PATH` environment variable.
4. Create a soft link from `/usr/local/mysql` to the MySQL installation directory.
5. Create the configuration file `/etc/my.cnf`, and set relevant parameters.
6. Execute `scripts/mysql_install_db`, and supply grants and privileges.

## 6.1.5 Configure Pentaho ETL

1. The ETL Kettle Spoon GUI requires `gtk` libraries to be able to launch. If they are not present on the server, use an external Microsoft Windows client instead to create ETL Spoon transformations.
2. Download and uncompress Pentaho Data Integration (`pdi-open-3.1.0-826.zip`) on the client, and launch `spoon.bat`. Follow the documentation to create a repository and a database connection.
3. Create or use an existing transformation for each data warehouse table with Table Input, Filter Rows, and Text Output icons. On Filter Rows, set filter conditions.
4. Save the transformations and copy them to the Sun Fire X4540 server.
5. Run the `pan` tool on the Sun Fire X4540 server using the following command, which provides the path name of the transformation XML file. The `pan` tool is a command line program that lets you launch the transforms designed in ETL Kettle Spoon.

```
# ./pan.sh -file=<ktr file> -level=Basic
```

### 6.1.6 Configure Pentaho BI-CE 2.0.0 Server

1. Install, configure, and start the BI-CE 2.0.0 server, per the documentation.
2. Ensure that reports can be opened on the BI-CE server.

### 6.1.7 Configure Pentaho Report Designer

1. Uncompress Pentaho Report Designer (`prd-open-1.7.1.tar.gz`), and start the tool (`./startDesigner_osx.sh`).
2. Configure a Java Naming and Directory Interface™ (JNDI) connection for the MySQL database. Refer to the documentation.
3. Click QueryDesigner to build the report, and supply a suitable query.
4. Save and publish the report onto the BI-CE server that was installed previously.

### 6.1.8 Configure Mondrian OLAP Server

1. Uncompress Mondrian OLAP server (`mondrian-3.0.4.11371.zip`). Copy `lib/mondrian.war` to: `<tomcat>/webapps`.
2. Configure `datasources.xml`, `mondrian.properties`, and `web.xml` by supplying the required data source name and parameters. Reference the documentation.
3. Install the demo database creation script, `foodmart_mysql.sql`. Discard `UNIQUE KEY` constraints.
4. Configure the demo files to establish data source connectivity, and access the Mondrian server: `http://<server:port>/mondrian/`.

## 7. Acknowledgements

I wish to thank these engineers from Pentaho Corporation for their timely help and support:

- **James Dixon**, Founder and CTO
- **Matt Casters**, Chief Architect, Data Integration, Kettle Project Founder
- **Brian Hagan**, Senior Support Engineer

## 8. Product Information

### 8.1 Sun Products

Organizations need to deliver large amounts of data quickly, reliably, and economically. Sun's servers and storage components, and its innovative technologies, can offer significant cost savings while providing enterprise-class data services, as well as strong scalability and performance.

The Solaris 10 OS, Sun's flagship OS, is multi-platform, scalable, and can yield performance advantages for databases, Web, and Java™ technology-based services. Its advanced features include security, system observability (DTrace), system resource utilization, an optimized network stack, data management, system availability, and interoperability tools.

Sun Fire x64 servers, which include AMD Opteron™ processors or Intel® processors, are also known as x86/x64-based systems. They can run the Solaris OS, Microsoft Windows, Linux, and VMware. Server types include rackmount servers, blade servers, and workstations. They can deliver extreme efficiency, intelligent scaling, reliability, and performance.

MySQL database, acquired by Sun, is the most popular open source database. It is multi-threaded and consists of an SQL server, client programs and libraries, administrative tools, and APIs. Java client programs that use Java DataBase Connectivity (JDBC™) connections can access a MySQL server through the MySQL Connector/J interface.

## 8.2 Pentaho Offerings

Pentaho OSBI Suite is an open source business intelligence product that provides a full range of business intelligence solutions to customers. It has reporting capabilities, data analysis capabilities, dashboards, and data mining capabilities, and it comprises tools such as data integration, which extracts, transforms, and loads data.

These features enable businesses to access a variety of information, including sales analysis, customer and product profitability, HR reporting, and finance analysis and reporting, and they enable delivery of complex information to top management.

Reporting can be on-demand or scheduled, and reports can be published in popular formats. Analysis provides pivot table views, graphical displays, and workflow integration. Dashboards provide re-usable display widgets to embed into applications. Data mining combines algorithms with OLAP technologies to provide intelligent data analysis to end users. Data integration provides a design GUI and high scalability and flexibility for data processing.

## 9. For More Information

Here are some additional resources:

- Pentaho Open Source Business Intelligence web site: <http://www.pentaho.com>
- Pentaho Spoon and Kettle: <http://etl-tools.info/en/pentaho/kettle-spoon.htm>
- TPC-H data warehouse workload kit and DBGEN data generator tool:  
<http://www.tpc.org/tpch>
- Mondrian FoodMart demo database:

<http://business-intelligence.phi-integration.com/2008/04/mondrian-mysql-setup.html>

- Tsung open source multiprotocol distributed load testing tool:  
<http://tsung.erlang-projects.org/>
- VNC software on the Solaris 10 OS:  
[http://www.salixtraining.co.uk/index\\_files/vncsol10.htm](http://www.salixtraining.co.uk/index_files/vncsol10.htm)
- Solaris OS web site: <http://www.sun.com/software/solaris/10/index.jsp>
- Sun Fire X4540 web site: <http://www.sun.com/servers/x64/x4540/>

- MySQL information:
  - On sun.com: <http://www.sun.com/software/products/mysql/index.jsp>
  - On mysql.com: <http://www.mysql.com/>
- Sun product documentation:
  - Documentation at <http://docs.sun.com>, such as Sun Fire X4540 Server documents, Solaris 10 System Administrator Collection, and MySQL documents
  - MySQL documents at mysql.com: <http://dev.mysql.com/doc/>
  - Sun Documentation Center: <http://www.sun.com/documentation/>
- Resources on BigAdmin, such as:
  - MySQL Resources for System Administrators: <http://www.sun.com/bigadmin/topics/mysql/>
  - Database resource collection (includes community submissions): <http://www.sun.com/bigadmin/collections/database.jsp>
  - Storage resource collection (includes community submissions): <http://www.sun.com/bigadmin/collections/storage.jsp>
  - Feature Article: Pentaho Business Intelligence and MySQL on Sun Storage 7000 Unified Storage System: [http://www.sun.com/bigadmin/features/articles/pentaho\\_7000.jsp](http://www.sun.com/bigadmin/features/articles/pentaho_7000.jsp)
- BigAdmin wiki at <http://wikis.sun.com/display/BigAdmin/Home>
  - Storage Tech Tips page: <http://wikis.sun.com/display/BigAdmin/Storage+Tech+Tips>
  - Databases page: <http://wikis.sun.com/display/BigAdmin/Databases>
  - Servers page: <http://wikis.sun.com/display/BigAdmin/Servers>
- Sun BluePrints™ papers (registration required) at the Sun BluePrints wiki (<http://wikis.sun.com/display/BluePrints/Main>), for example:
  - Running MySQL Database in Solaris Containers: <http://wikis.sun.com/display/BluePrints/Running+MySQL+Database+in+Solaris+Containers>
  - Improving MySQL Database Scalability: <http://wikis.sun.com/display/BluePrints/Improving+MySQL+Database+Scalability>
- Sun download site: <http://www.sun.com/download/>
- Sun training courses at <http://www.sun.com/training/>, for example:
  - Intermediate System Administration for the Solaris 10 Operating System (WSB-200-S10)
  - Sun Fire X4540 Server Administration (WET-6179)
- MySQL training courses on sun.com: <http://www.sun.com/software/products/mysql/training.jsp>
- MySQL training courses on mysql.com: <http://www.mysql.com/training/>

- Discussions:
  - Sun forums: <http://forums.sun.com/index.jspa>
  - BigAdmin Discussions collection: <http://www.sun.com/bigadmin/discussions/>
- Support:
  - Sun resources:
    - Register your Sun gear: <https://inventory.sun.com/inventory/>
    - Services: <http://www.sun.com/service>
    - SunSolve<sup>SM</sup>: <http://sunsolve.sun.com>
  - Community system administration experts:  
<http://www.sun.com/bigadmin/content/communityexperts>
- Events of interest to users of Sun products:
  - Worldwide Developer Events and Sun Tech Days:  
<http://developers.sun.com/events/>
  - Current Events: <http://www.sun.com/events/index.jsp>

## 10. Licensing Information

Unless otherwise specified, the use of this software is authorized pursuant to the terms of the license found at [http://www.sun.com/bigadmin/common/berkeley\\_license.html](http://www.sun.com/bigadmin/common/berkeley_license.html).