



Upgrading Your Row-based Database to a Columnar Database in Infobright

Benefits and Migration Techniques

25-Feb-10

AGENDA

1 Infobright Overview

- What it is
- Column vs Row
- Technology Overview and Benefits

2 Migrating from Row-based databases to Infobright

- Considerations
- Data Loading
- Tools and Utilities
- Getting Started

Infobright

Innovation

- First commercial open source analytic database
- Knowledge Grid provides significant advantage over other columnar databases
- Fastest time-to-value, simplest administration

Strong Momentum & Adoption

- Release 3.3.1 generally available
- > 100 customers in 10 Countries
- > 40 Partners on 6 continents
- A vibrant open source community
 - > 1 million visitors
 - 35,000 downloads
 - 4,500 active community participants

Gartner.
Cool Vendor in Data Management
and Integration 2009

MySQL. 
Partner of the Year 2009

intelligent
enterprise
2010 Company to Watch


Infobright: Economic
Data Warehouse
Choice



Infobright Technology

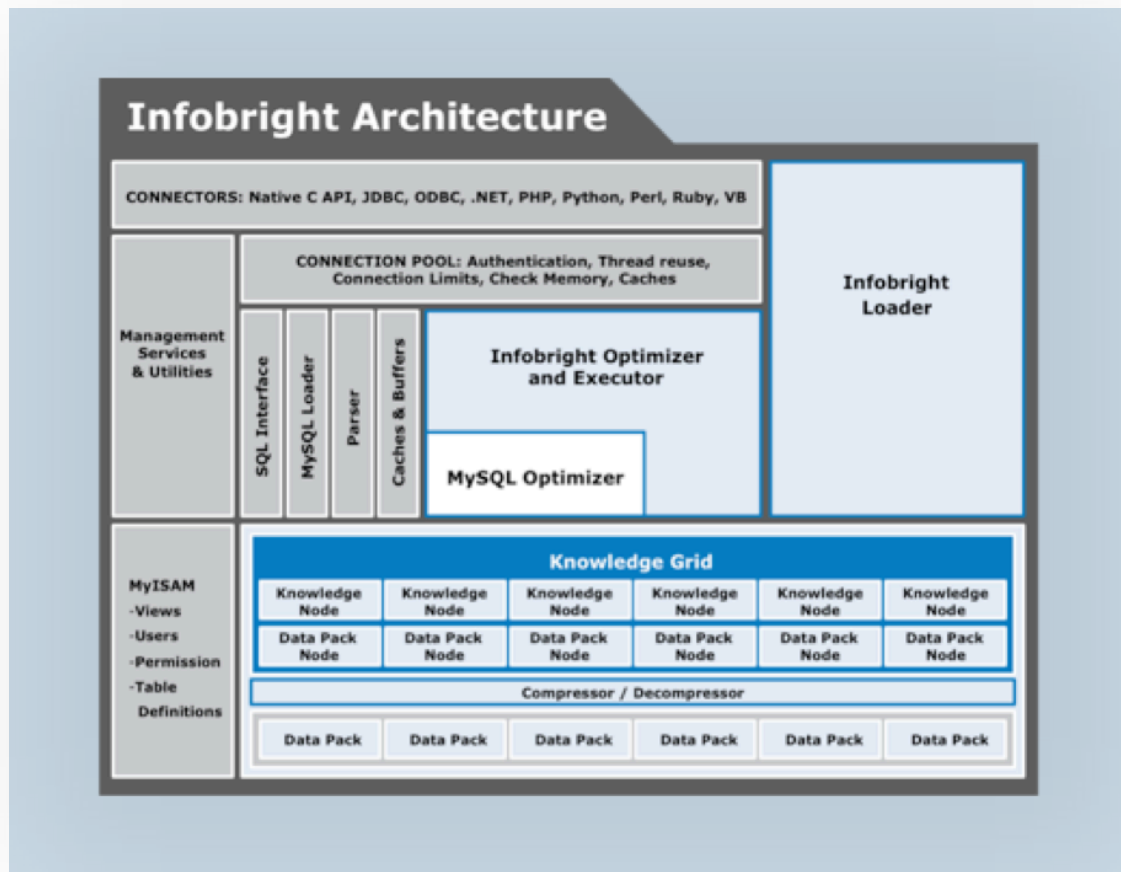
Infobright is a high performance
analytic database
that delivers fast query performance
against large volumes of data
with minimal IT effort

The Infobright Difference

Low cost	50% of alternative solutions
Reduction in storage costs	10x and better compression
Self-managing	90% less administrative effort
Significant reduction in maintenance	Delivers high performance query response time without indexes. Supports ad hoc and standard queries without manual tuning
Scalable, high performance	Up to 50TB using single server

Infobright and MySQL

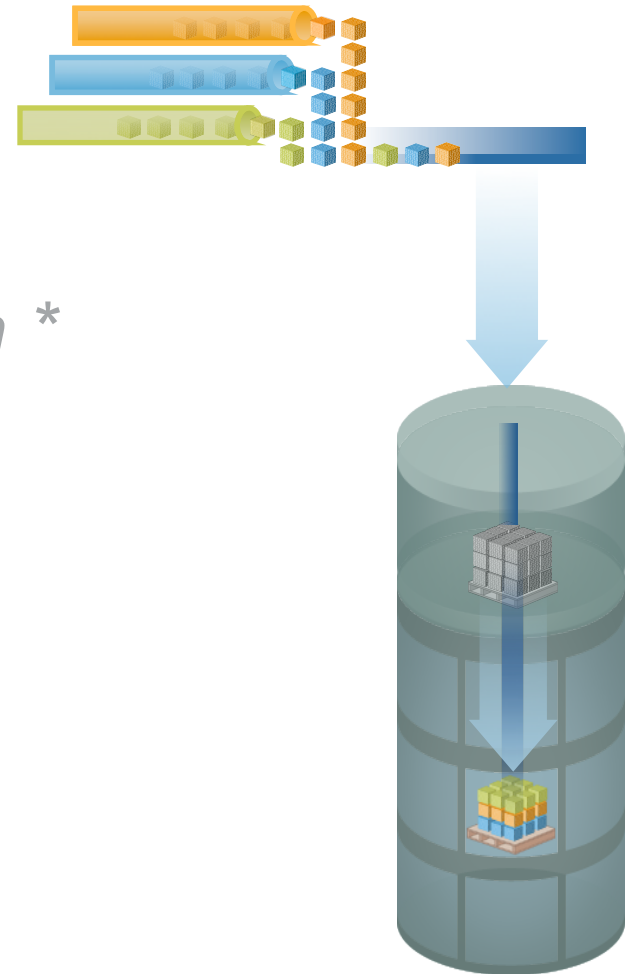
- Infobright is architected on MySQL, “*the world’s most popular open source database*”



- Provides a simple scalability path for MySQL users and OEMs
- No new management interface to learn
- MySQL integration enables seamless connectivity to BI tools and MySQL drivers for ODBC, JDBC, C/C++, .NET, Perl, Python, PHP, Ruby, Tcl, etc.

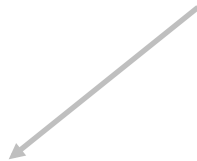
Infobright Technology: Key Concepts

1. Column orientation
2. *Data packs and Compression **
3. *Knowledge Grid **
4. *Optimizer **



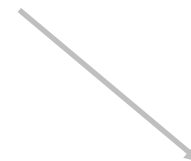
1. Column vs. Row Orientation

<i>Employee_ID</i>	<i>Job</i>	<i>Dept</i>	<i>City</i>
1	Shipping	Operations	Toronto
2	Receiving	Operations	Toronto
3	Accounting	Finance	Boston



Data stored in rows

1	Shipping	Operations	Toronto
2	Receiving	Operations	Toronto
3	Accounting	Finance	Boston



Data stored in columns

1	Shipping	Operations	Toronto
2	Receiving	Operations	Toronto
3	Accounting	Finance	Boston

Column vs. Row Orientation - Use Cases

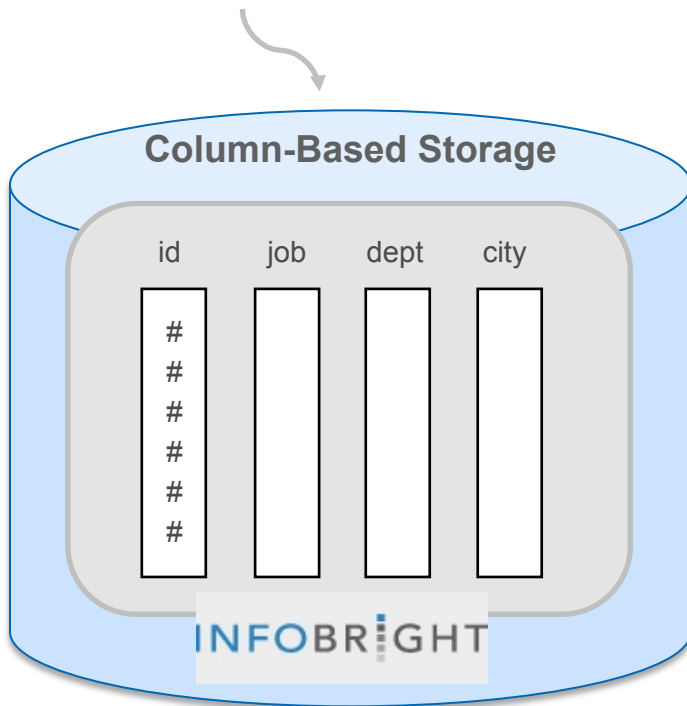
ID	job	dept	city
#			
#			
#			
#			
#			

Row-Based Storage

ID	job	dept	city
#			
#			
#			
#			
#			
#			

Row Oriented works if...

- All the columns are needed
- Transactional processing is required



Column Oriented works if...

- Only relevant columns are needed
- Reports are aggregates (sum, count, average, etc.)

Benefits

- Very efficient compression
- Faster results for analytical queries

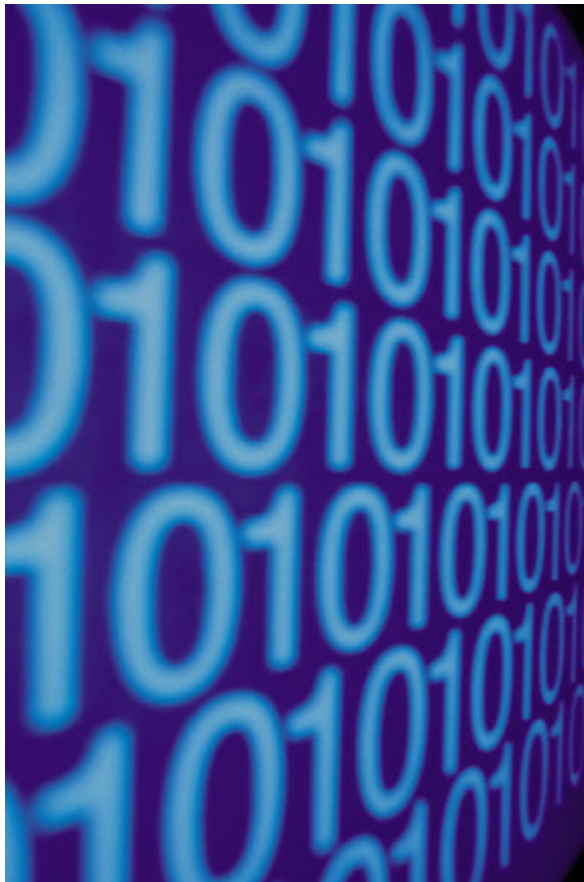
Best Use Cases

- **Analytic applications with large data volumes**
 - Examples: Web/online analytics, mobile analytics, customer behavior analysis, marketing/advertising analysis
- **Log/event management**
 - Examples: Telecom CDR analysis and reporting, systems/network/security analysis
- **Data Marts**
 - Application or business unit specific
 - Data warehouse for SMB
- **Embedded analytic database for ISVs/SaaS providers**

What Your Data Looks Like Now

Original data

500GB



Compressed data

50 GB

Avg compression ratio of 10:1

=



+



Knowledge Grid

< 0.5 GB

< 1% of compressed data

Why Some Customers Have Bought Infobright – Some Performance Statistics

- **Fast query response with no tuning**

Customer's Test	Row-based RDBMS	Infobright
Analytic queries	2+ hours	< 10 seconds
Query (AND – Left Join)	26.4 seconds	.02 seconds
Query set	10 sec – 15 min	0.43 – 22 seconds
BI report	7 hours	17 seconds
Data load	11 hours	11 minutes

- **Fast and consistent data load speed** as as database grows. Up to 300GB/hour on a single server

“Infobright is 10 times faster than [Product X] when the SQL statement is more complex than a simple `”SELECT * FROM some_table”`. With some more complex SQL statements, Infobright proved to be more than 50 times faster than [Product X].”
(from benchmark done by leading BI vendor)

Mobile Analytics Example: bango.com

Results – *Quantitative Measures*

- Compression
 - 1 month of data (actual customer data)
 - Row-based RDBMS 450GB (includes indices)
 - Infobright 10GB
- Queries

Query	Row-based	Infobright
1 Month Report (5M events)	11 min	10 secs
1 Month Report (15M events)	43 min	23 secs
Complex Filter (10M events)	29 min	8 secs

Business Intelligence Example: Austin Energy

Changes	Before	After
<i>Data Volumes Available</i>	1 month of data	2 years of data (2.5 million rows, 73 columns)
<i>Compression</i>	1 : 1 (none)	98 : 1
<i>Data Footprint</i>	5+ GB	5 GB now occupies ~100 MB
<i>Age of Data Available</i>	6 weeks old	1 day old
<i>Reporting Time / KPI</i>	Once per month	< 10 seconds on demand
<i>End Users</i>	Limited, isolated business units	Organizationally secure access to data

Bear in Mind

The unique attributes of column orientation in Infobright are transparent to developers.

The benefits are obvious and immediate to users.

- Infobright *is* a relational database
- Infobright observes and obeys SQL standards
- Infobright observes and obeys standards-based connectivity
 - Design tools
 - Development tools
 - Administrative tools
 - Query and reporting tools

Upgrading Your Row-based Database

“Upgrade”, in this case, is a business concept, not an actual technical upgrade.

- You will experience design benefits from a simplified schema
- You will experience administrative benefits
- You will experience cost benefits
- Your users will experience performance benefits
- *But ...* you will have to migrate your data

Migration of Existing Databases



So, how do I move my data in a traditional row-oriented RDBMS to Infobright?

The Basic Concept

All database migration projects, regardless of what tools are used to manage the mechanics, do two (2) essential tasks:

- ***Export the data from the original source database***
- ***Import the data into the target database, Infobright***

All the rest is left to you as a choice of convenience, expediency and/or what you have experience with or may already own.



Migration Approaches

1. Manual execution of individual tasks
2. Utilities provided by the vendor
 - usually scripts or lightweight programs
3. One-time migration tools
4. Data movement tools
 - most notably ETL tools
 - open source software (OSS) options
 - commercial off-the-shelf (COTS) options

Infobright (and MySQL) Considerations

As a purpose-built analytical database engine, as opposed to a general purpose database, there are unique differences between schemas in Infobright and other technologies.

- Declaration of storage engine type
- Lack of need for indices or partition schemes
- Lack of referential integrity checks
- Removal of constraints
- Minor data type differences
- Supported character sets and collations

Storage Engine Types

As with all MySQL databases, Infobright has a unique storage type that must be specified when creating tables by adding the following:

ENGINE=BRIGHHOUSE

Example:

```
CREATE TABLE imbright (  
    id    INTEGER,  
    value CHAR(10)  
)  
ENGINE=BRIGHHOUSE  
DEFAULT CHARSET=ascii COLLATE=ascii_bin;
```

Indices and Partition Schemes

Infobright neither needs nor allows manual creation of performance structures with duplicated data or constraints that are typically expressed as indices.

In your Infobright DDL -

- remove **PRIMARY KEYS**
- remove **FOREIGN KEYS**
- remove **UNIQUE** constraints
- remove **CHECK** constraints
- remove **PARTITION BY** clauses

An “aside” On Performance Structures

Because of Infobright’s high query performance levels against atomic detail data, ***start out simple*** and add performance structures only as needed.

Drop from your schema –

- all aggregate tables
- all reporting tables
- all summary tables

Data Type Support

Infobright supports a large subset of MySQL data types and DDL statements. However, there are some inconsistencies. *All* are on the Roadmap for resolution.

- UNSIGNED INTEGERS
- DECIMAL(65,30)
- ENUM
- AUTO_INCREMENT
- BLOBs

Detailed information on supported data types can be found here

http://www.infobright.org/wiki/Supported_Data_Types_and_Values/

Supported Character Sets and Collations

Infobright currently* supports the ASCII character set and ASCII BINARY collation. These are the default values and need not be specified.

Example:

```
CREATE TABLE imbright (  
    id    INTEGER,  
    value CHAR(10)  
) ENGINE=BRIGHHOUSE  
DEFAULT CHARSET=ascii COLLATE=ascii_bin;
```

* Full Unicode/UTF8 support will be delivered by mid-2010

Manual Execution – Data Sourcing (aka export)

Every database has an export function. With manual exporting, consult your DBA for the appropriate method. But, every table that needs to be migrated must be written to a separate file for loading.

- Microsoft SQL Server and Sybase provide `bcp`
- MySQL provides `SELECT INTO OUTFILE` (all engines)
- Oracle provides spooling – `SPOOL 'path/file'` then `SELECT`
- Informix provides `dbexport`
- IBM DB2 provides `export`
- PostgreSQL-based systems provide `dbdump`

Manual Execution - Data Loading (aka *import*)

From <http://www.infobright.org/wiki/Data>Loading/>

Import your data into an Infobright table by using the following load syntax (all other MySQL Loader syntax is not supported):

```
LOAD DATA INFILE '/full_path/file_name' INTO TABLE table_name
  [FIELDS
    [TERMINATED BY 'char']
    [ENCLOSED BY 'char']
    [ESCAPED 'char']
  ];
```

The data is committed when the load completes if AUTOCOMMIT is set to on. *This is the default setting*, but you can make it explicit by setting:

```
mysql> SET AUTOCOMMIT=1;
```

Utilities Provided By Infobright

All utilities described here can be found on the Contributed Software page of the Downloads section on infobright.org

<http://www.infobright.org/Downloads/Contributed-Software/>



HOME

DOWNLOAD

COMMUNITY

BLOGS

RESOURCES

SUPPORT

Home > download contributed software

Download Contributed Software

Utilities Provided By Infobright

These include:

- The ICE Breakers (work with Infobright Enterprise Edition)
 - Microsoft SQL Server
 - Oracle
 - MySQL
- MyISAM Export
 - Table level only from single instance
- InnoDB Export – Import to Infobright
 - Table level only for two instances of MySQL
 - InnoDB instance and Infobright instance
- Tutorial: Data Transfer from MySQL to Infobright
 - Submitted by an Infobright customer

The ICE Breaker

The screenshot shows a Windows-style application window titled "ICE Breaker for SQL Server - V 0.4 - Alpha". The window features the "INFOBRIGHT" logo at the top. Below the logo, there are two main sections: "SQL Server" and "File Locations".

SQL Server

Server Name:	<input type="text" value="192.168.1.75"/>
Database:	<input type="text" value="EDW01"/>
Login:	<input type="text" value="sa"/>
Password:	<input type="password" value="*****"/>

File Locations

Export File Directory:	<input type="text" value="C:\temp"/>	<input type="button" value="..."/>
Import File Directory:	<input type="text" value="/home/mysql/"/>	
Script File:	<input type="text" value="C:\Users\carl\Desktop\load.txt"/>	<input type="button" value="..."/>

At the bottom of the window, there are three buttons: "Go!", "Forum", and "Cancel".

Utilities Provided By Infobright

MyISAM Export August 6, 2009

For Linux/Unix use only

This is a Unix (bash) script that works on a table-by-table basis and exports data from a MyISAM table into a data file and imports (via LOAD) into an Infobright table of the same name and schema. It could be adapted for multiple tables can be purged using the same filter criteria.

InnoDB Export - Import to Infobright (VERSION 0.9) Added January 29, 2010

Script to export/import data from InnoDB table in one MySQL instance to an Infobright table in another.

Assumption(s): table to be exported and imported have same name the whole table is exported - edits will need to be made for higher selectivity

Caveats: there is no validation of the "type" of either schema there are very few sanity checks in this script such as file system space and record counts out and in.

Tutorial: Data Transfer from MySQL to Infobright September 22, 2009

Tutorial on how to get data out of mysql into a CSV file, work on the data using awk and finally loading it into Infobright.

The emphasis is on the checking and discovery of the data in the CSV file.

Utilities Provided By Infobright

See also:

Migration Guide: MySQL/MyISAM to Infobright

<http://support.infobright.com/Support/Resource-Library/Whitepapers/>

One-time Migration Tools

This class of tools provides multiple functions.

- Introspects the source database's metadata and re-create it in the target database
- Attempts to re-create all performance structures in Infobright, which it interprets as MyISAM
 - indices
 - aggregate tables
 - partition schemes
 - as well as, referential integrity and other constraints
- Converts stored procedures (if possible)
- Transfers the data

One-time Migration Tools

Because these tools see Infobright as MyISAM, most are more work than convenience because the schema must be modified before data transfer begins to comply with Infobright's schema requirements. When this is not an issue, they are a good fit.

Most commonly seen tool

- **SQL**Ways from Ispirer
- <http://www.ispirer.com/>

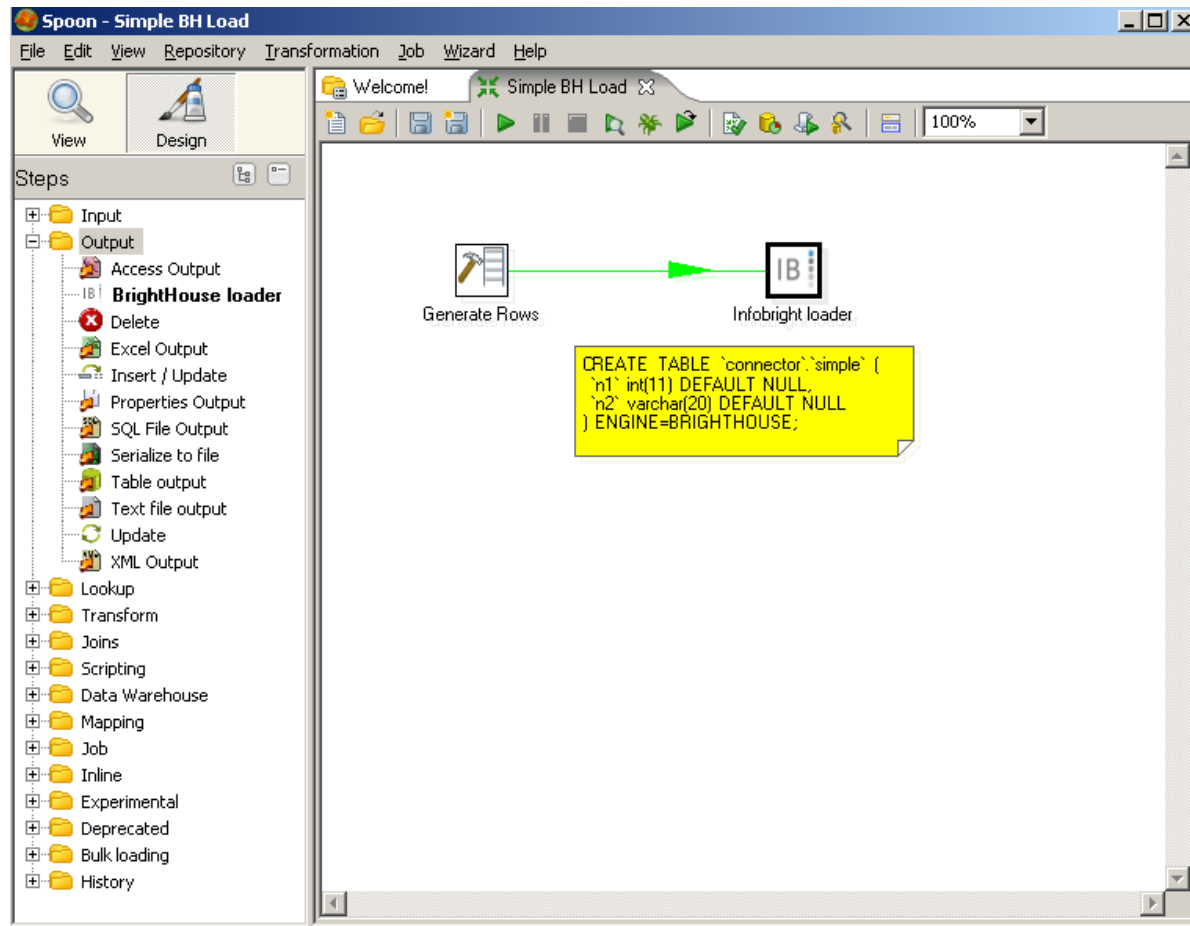
Data Loading with Custom ETL Connectors

Custom Connectors for Open Source ETL tools

- Pentaho Data Integration, or PDI (aka Kettle) from Pentaho
- Jaspersoft ETL from Jaspersoft
- Talend Open Studio

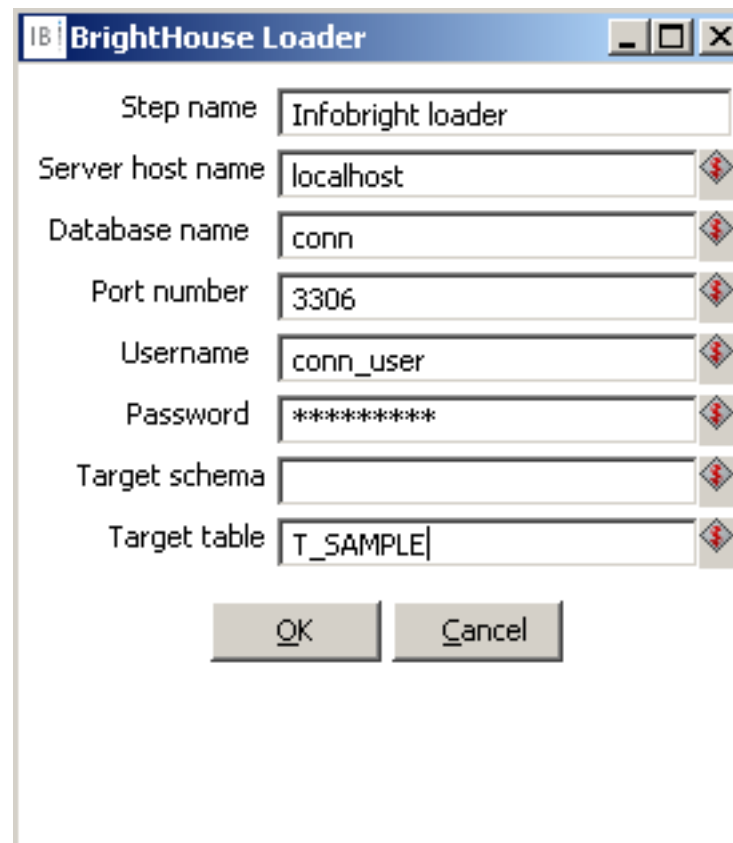
Data Loading with Custom Pentaho Connector

Infobright provides a custom connector for Pentaho Data Integration, PDI



Configuring the PDI Connector

- Setting parameters for the PDI Connector is easy
- Just configure the options in the dialog box



The screenshot shows a dialog box titled "BrightHouse Loader" with the following fields and values:

Field	Value
Step name	Infobright loader
Server host name	localhost
Database name	conn
Port number	3306
Username	conn_user
Password	*****
Target schema	
Target table	T_SAMPLE

Buttons: OK, Cancel

Data Loading with Custom Jaspersoft Connector

- Under the covers, the Jaspersoft ETL tool uses the ETL engine from Talend
- The Talend connector can also be invoked from JasperETL

The screenshot displays the Talend Open Studio interface. The main workspace shows a job design for 'fact_sales_load 0.1'. The job flow consists of the following components: tFileInputDelimited_1, tInfobrightOutput_1, LogSuccess, and LogFail. The tInfobrightOutput_1 component is selected, and its properties are shown in the bottom panel. The properties include Host: localhost, Port: 3306, Database: connector, Username: gtF, Password: gtF, and Table: fact_sales. The right palette shows various components, including tInfobrightOutput.

Property	Value
Host	localhost
Port	3306
Database	connector
Username	gtF
Password	gtF
Table	fact_sales

Get Started



HOME

DOWNLOAD

COMMUNITY

BLOGS

RESOURCES

SUPPORT

- Join the forums, learn from the experts
- Sign up for a webinar
- Download a white paper
- Download ICE (Infobright Community Edition)
- Download integrated VMs with Pentaho, Jaspersoft or Talend at www.infobright.org
- Download a free trial of Infobright Enterprise Edition

info@infobright.com
www.infobright.com
www.infobright.org