

INFOBRIGHT



***Infobright for ISVs/SaaS:
The Ideal Embedded Analytic
Database***

Worldwide market for enterprise software hit \$220B in 2008*

Worldwide SaaS revenue projected to grow 22 percent in 2009*

Business Intelligence remains #1 CIO technology priority*

Enterprise applications ranked #2 CIO technology priority*

(* Source: Gartner)

How do these relate to each other? They each indicate a tremendous increase in demand for business intelligence and analytics within enterprise applications, whether the application is developed in-house, licensed as a packaged application or provided as a Software-as-a-Service (SaaS) offering.

Industry analysts report that enterprises are requiring the capability to enhance and customize licensed applications with advanced analytics. This includes the ability to add highly flexible reports, queries and dashboards, but increasingly includes the ability to add new sources of data into the application's database as a way to avoid the need to create an in-house application data mart. This requirement, combined with the explosive growth in the volume of data being collected and budget pressures across all organizations, presents a challenge to software companies looking to deliver differentiated capabilities.

Delivering these capabilities requires an embedded database foundation that can provide very fast response times across large volumes of data, without adding significant cost, administrative effort or hardware and software footprint.

Software providers must consider the following factors when choosing the embedded analytic database:

- *Performance of queries and reporting, and consistency as the amount of data grows*
- *Scalability (data volumes, number of users)*
- *DBA and administrative effort to implement, optimize, manage and tune the database*
- *Ease of integration with applications and BI tools*
- *Data load options and speed*
- *Deployment flexibility*
- *Hardware and software footprint*
- *Database licensing and maintenance costs*

Faster Performance, Less Work, Lower Cost

As the first commercial open source analytic database, Infobright is ideal for use as an embedded database. The combination of its column-oriented configuration and the unique Knowledge Grid architecture provides massive data compression, high performance, minimal administration and a very small footprint. In addition, Infobright's support for a broad range of BI tools, data integration tools and open and standard application interfaces enables easy application integration and delivery of flexible reporting, dashboards and query capability.

- Easily Embeddable***
- **Small download** – less than 15MB in size
 - **Zero-configuration** – simple install and setup
 - Simple, easy to use **open API**
 - Based on **open source**
 - **Self-contained** – no external dependencies, not reliant on any third party add-ons or products
 - **Complete** – the entire database is stored in a limited number of cross-platform files

- Easy to Manage***
- **No:** indexes, data partitioning, data duplication, or database tuning needed
 - **No:** aggregate tables prepared in advance
 - **Any schema** accepted
 - **Very small hardware footprint** - can support up to 50TB of data on a single, industry standard server with far less storage than alternatives

- Easy to Use***
- **MySQL interface**
 - **Infobright Loader** option – up to about 250 GB/hr for binary and 175GB/hr for CSV files
 - **MySQL Loader** option – expanded flexibility and features for varying file types

- Excellent Performance and Scalability***
- **Fast load speed** – remains constant as the size of the database grows
 - **Fast query speed** on large volumes of data
 - **Very high data compression** (from 10:1 to over 40:1), which drastically reduces I/O (improving query performance) and storage requirements – supports terabyte-sized databases that can be compressed by as much as 40X, saving disk space
 - Fast ETL load times with **parallel load**
 - Broad **SQL and feature set support** for industrial strength, information intensive applications
 - **ACID compliance** – transactions are atomic,

- consistent, isolated, and durable (ACID) in the event of system crashes and power failures
- Support for a dual peer-to-peer server deployment with **no single point of failure**, providing high-availability with automatic failover
- High Availability**
- **Cross-platform** – popular **Linux** distributions, **Solaris**, and **Windows (Win32 and Win64)** are supported out of the box
 - Works with major commercial and open source **Business Intelligence and ETL tools** from vendors such as **Actuate/BIRT, Jaspersoft, Pentaho, Talend, MicroStrategy, Cognos, Informatica, Business Objects**, and others
 - Supports many languages including **Java, C/C++, Ruby, Perl, 4th GL's, ODBC** and **JDBC** compliant solutions, etc
- Highly Flexible**
- **90% less work**
 - **50% or less the cost** than alternatives
 - Requires significantly **less hardware** than other products
 - **OEM program** and pricing allows you to start small and grow
- Low Cost**

The advantage of Infobright's technology has been of great value to Polystar, a leading telecommunications supplier based in Stockholm, Sweden, who provides world-class Customer Experience Management (CEM) and Service Assurance solutions to leading telecom operators and service providers world-wide:

*"In the highly competitive **telecommunications** industry, service performance is a key differentiator. As a result, our customers expect our applications to be able to process **large amounts of user data and call details**. They also demand **fast query response times**, so that they can stay on top of any issues that can potentially impact their business.*

*We were very impressed with the Infobright solution's load speed, its high-data compression rate and its ability to maximize query performance. With Infobright Enterprise Edition in place, we expect to be able to **deliver the critical network information that our customers require faster than ever.**"*
- Jan Öhman, CEO of Polystar

Infobright Architecture

Infobright's architecture combines a columnar database and a unique Knowledge Grid architecture optimized for analytics. The following is an overview of the architecture and a description of the data loading and compression technology.

Why Columnar?

Infobright at its core is a highly compressed column-oriented database, which means that instead of the data being stored row-by-row, it is stored column-by-column. Traditional row-based database systems have typically been used to store application data, but they were designed for transaction-oriented applications, not for analytics.

A row-oriented design used for analytics has the *disadvantages* of:

- Forcing the database to retrieve all column data for each row regardless of whether or not it is required to resolve the query.
- Requiring indexes in order to improve efficiency. These must be planned ahead of time and customized for pre-determined reporting. The use of indexes is unsuitable for analytical queries, which require flexibility in data retrieval patterns.
- Degradation in load speeds for growing amounts of data. As new data is added to existing tables, indexes must be recreated each time resulting in very large sort operations. This problem only gets worse as the database grows.
- Degradation in query performance for growing amounts of data and/or an increasing number of concurrent queries and users. Even when reporting on indexed columns, there is an increasingly large amount of data retrieval required as all columns are returned for each data set, regardless of relevance.

Alternatively, a column-based database is a much better fit for analytics because it retrieves only the data relevant to the query, rather than the entire row.

In addition to leveraging the column-oriented approach, Infobright employs a unique method of compressing, storing and retrieving data using our Knowledge Grid that provides a scalable, flexible analytical environment without the need for any indexes.

The *advantages* to Infobright's column-orientation include:

- **Faster load times.** As new data is added there is no degradation in load speed since no indexes need to be recreated. The load times are simply proportional to the load size, not the size of the existing tables.
- **Faster query response times.** Most analytic queries only involve a subset of the columns of the tables, and a column-oriented database focuses on retrieving only the data that is required. Infobright has further improved query response times by organizing information about the data in the Knowledge Grid ahead of time.
- **Industry-leading data compression.** Each column stores a single data type, as opposed to rows that typically contain multiple data types. This allows compression to be optimized for each particular data type, to which we apply our patent-pending compression algorithms. On average, Infobright users experience 10:1 compression (raw data versus size on disk) while some report compression as high as 40:1.

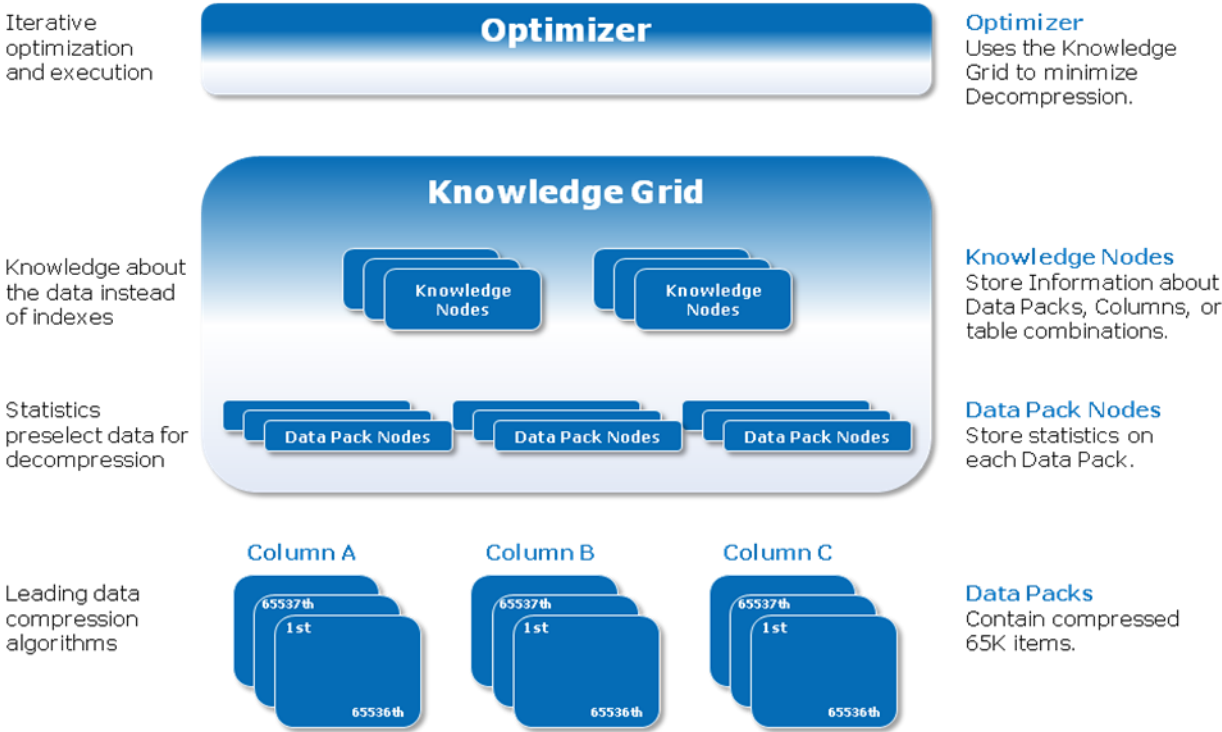
Data Organization and the Knowledge Grid

The Infobright database resolves complex analytic queries without the need for traditional indexes, data partitioning, manual tuning or specific schemas. Instead, the Knowledge Grid architecture automatically creates and stores the information needed to quickly resolve these queries.

Infobright organizes the data into 2 layers: the compressed data itself that is stored in segments called Data Packs, and information about the data which comprises the components of the Knowledge Grid.

For each query, the Infobright Optimizer uses the information in the Knowledge Grid to determine which Data Packs are relevant to the query before decompressing any data.

Figure 1 – Architecture Layers



Data Packs

The data within each column is stored in 65,536 item groupings called Data Packs. The use of Data Packs improves data compression, since there may be less variability within a subset of data in a column.

Data Pack Nodes (DPNs)

Data Pack Nodes contain a set of statistics and aggregate values of the data from each Data Pack: MIN, MAX, SUM, AVG, COUNT, and No. of NULLs. There is a 1:1 relationship between Data Packs and the DPNs that are created automatically during load. The Optimizer then has permanent summary information available about all of the data in the database that will later be used to flag relevant Data Packs when resolving queries. In the case of traditional databases, query resolution is aided by indexes that are created for a subset of columns only.

Knowledge Nodes (KNs)

This is a further set of metadata that is more introspective of the data within the Data Packs, describing ranges of numeric value occurrences and character positions, as well as column relationships between Data Packs. The

introspective KNs are created at load time, and the KNs relating the columns to each other are created in response to queries involving JOINS in order to optimize performance.

The DPNs and KNs together form the **Knowledge Grid**. Unlike the indexes required for traditional databases, DPNs and KNs are not manually created, and require no ongoing care and maintenance. Instead, they are created and managed automatically by the system. In essence, the Knowledge Grid provides a high level view of the entire content of the database with a minimal overhead of approximately 1% of the original data. By contrast, classic indexes may represent as much as 20% to 100% of the size of the original data.

The Infobright Optimizer

The Infobright Optimizer is the highest level of intelligence in the architecture. It uses the Knowledge Grid to determine the minimum set of Data Packs needed to be decompressed in order to satisfy a given query in the fastest possible time by identifying the relevant Data Packs. In some cases the summary information already contained in the Knowledge Grid is sufficient to resolve the query, and nothing is decompressed.

How do Data Packs, DPNs and KNs work together to achieve fast query performance?

For each query, the Optimizer uses the summary information in the DPNs and KNs to group the Data Packs into one of the three following categories:

- *Relevant* Packs – where each element (the record's value for the given column) is identified, based on DPNs and KNs, as applicable to the given query,
- *Irrelevant* Packs – where the Data Pack elements hold no relevant values based on the DPN and KN statistics, or
- *Suspect* Packs – where some relevant elements exist within a certain range, but the Data Pack needs to be decompressed in order to determine the detailed values specific to the query.

The Relevant and Suspect packs are then used to resolve the query. In some cases, for example if we're asking for aggregates, only the Suspect packs need to be decompressed because the Relevant packs will have the aggregate value(s) pre-determined. However, if the query is asking for record details, then all Suspect and all Relevant packs will need to be

decompressed.

An Example of Query Resolution Using the Knowledge Grid

Lets look at a table of employees with the following 4 columns: salary, age, job, and city. Now lets apply a query that asks for a count of the number of employees fitting a particular description.

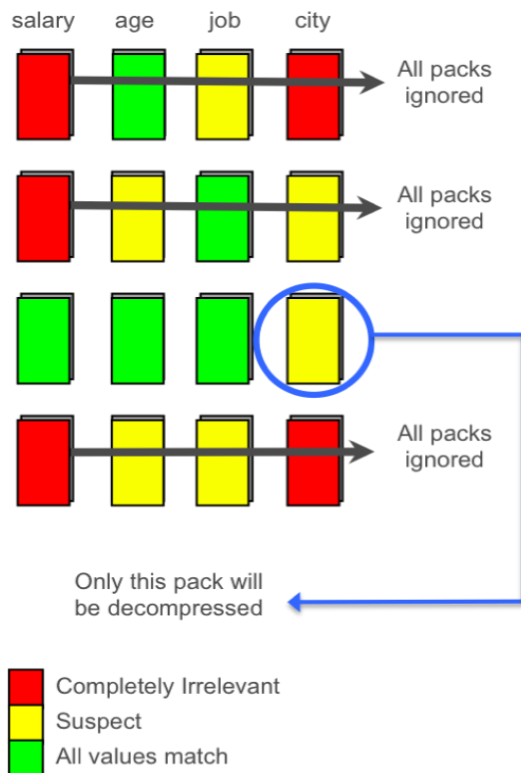
In this case we want to know the number of employees that have a salary over \$50k, are below the age of 35, have a job description of Accounting and are working out of the Toronto office.

Here is the sample query:

```
SELECT count(*) FROM employees
WHERE salary > 50,000 AND age < 35
AND job = 'Accounting' AND city = 'TORONTO';
```

The Optimizer uses the specific constraints of the query to identify the **Relevant**, **Suspect**, and **Irrelevant** Data Packs.

Figure 2 – Example of Query Resolution



In our sample table, the first constraint of `salary > 50000` eliminates 3 of the 4 Data Packs by using the MIN and MAX information stored in the Data Pack Nodes to tell us that all values in these Data Packs are less than or equal to 50000, making them *Irrelevant*.

Using similar logic for the column `age`, we can determine that 2 Data Packs contain employees with age under 35, making them *Relevant*, and 2 Data Packs contain some employees with age under 35 making them *Suspect* since we need more details about the date to determine how many.

We continue this logic to the `job` and `city` columns, but since we've already identified *Irrelevant* Data Packs in the first column based on `salary`, only the 3rd row of Data Packs for the entire table needs to be examined. We essentially use the Knowledge Grid to eliminate all Data Pack rows that have any columns flagged as *Irrelevant*.

In this example we found that only the Data Pack of the column `city` actually needs to be decompressed since the other 3 were found to be *Relevant*, so of the entire table we now only have 1 Data Pack to decompress.

This is the key to the fast query response Infobright delivers – eliminating the need to decompress and read extraneous data in order to resolve a query.

Data Loading and Compression

Infobright includes multiple options for loading data into the database:

- The Infobright high-speed Loader
- The MySQL Loader
- Loading data using third party ETL tools

The Infobright Loader and MySQL Loader are delivered as part of Infobright Enterprise Edition. Connectors to popular ETL tools such as JasperETL, Talend Open Studio and Pentaho Data Integration are available as free downloads from the Infobright connector core Java library (API) at <http://www.infobright.org/Downloads/Contributed-Software/>.

Since it was designed for fast data loading, the Infobright Loader has stricter requirements in terms of data formatting and less error checking than the MySQL Loader, as it assumes that the incoming data is aligned with the target database table and suitably formatted. The high speed Infobright Loader can be used for both text and binary files, achieving up to

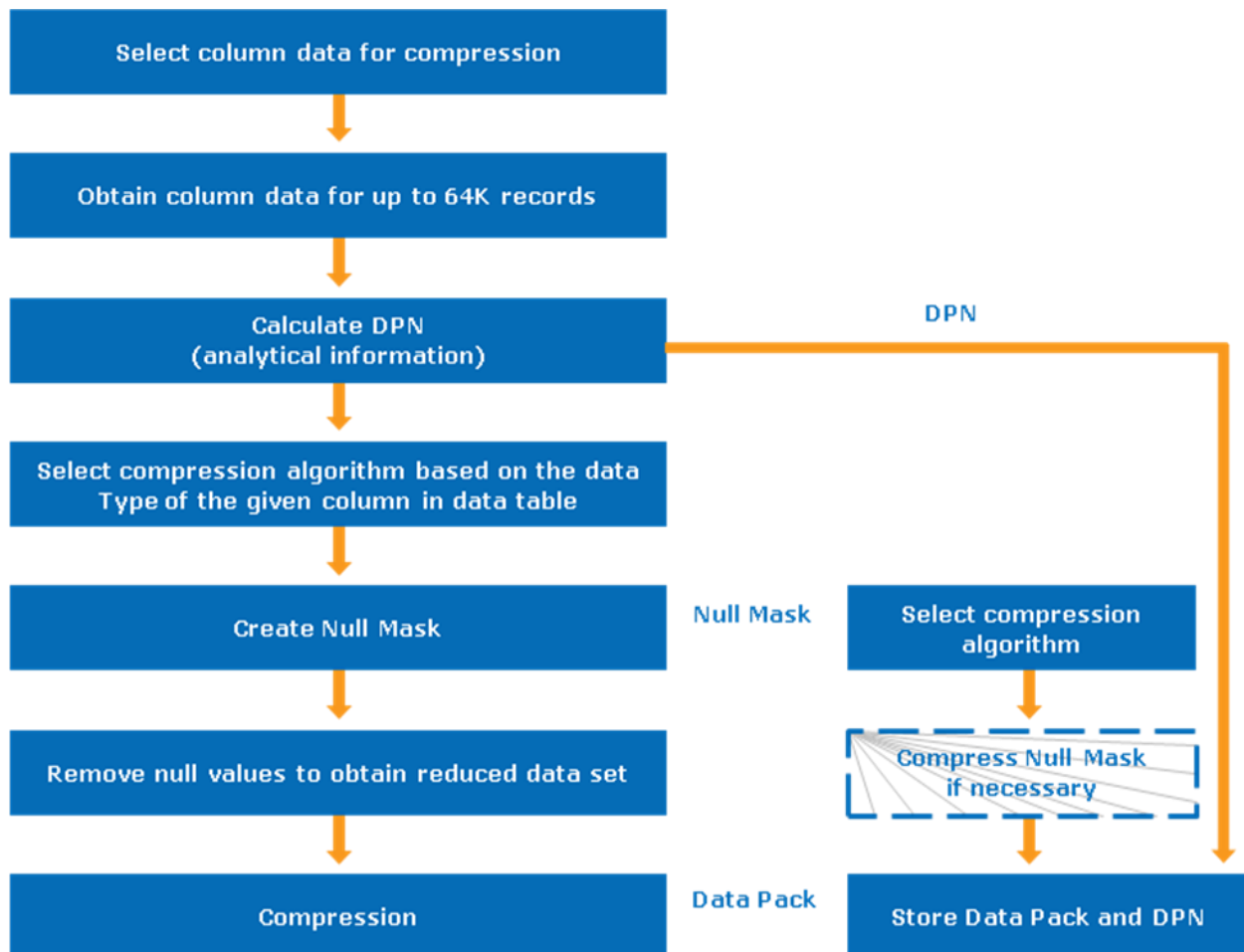
approximately 250GB/hour on a single server with parallel loads.

The MySQL Loader supports text-based loads and has additional features such as the ability to use variably formatted files and more extensive error handling. However, it is significantly slower than the Infobright Loader and is typically used during application development or for root cause analysis in the event of load failure.

For either loader, the mechanism of creating and storing Data Packs and their DPNs is the same, as illustrated below.

During the load, 65,536 values of a given column are treated as a sequence, with zero or more NULL values occurring anywhere in the sequence. Information about the NULL positions is stored separately and the remaining stream of non-NULL values is compressed, taking full advantage of any regularities within the data.

Figure 3 – Data Load and Compression Process



High Data Compression

An average compression ratio of 10:1 is achieved after loading data into Infobright. For example 10TB of raw data can be stored in about 1TB of space, including the 1% overhead associated with creating the Data Pack Nodes and the Knowledge Grid.

The compression algorithms for each column are selected based on the data type, and are applied iteratively to each Data Pack until maximum compression for that Data Pack has been achieved. Within a column the compression ratio may differ between Data Packs depending on how repetitive the data is. Some customers have reported an overall compression ratio as high as 40:1.

How Infobright Leverages MySQL

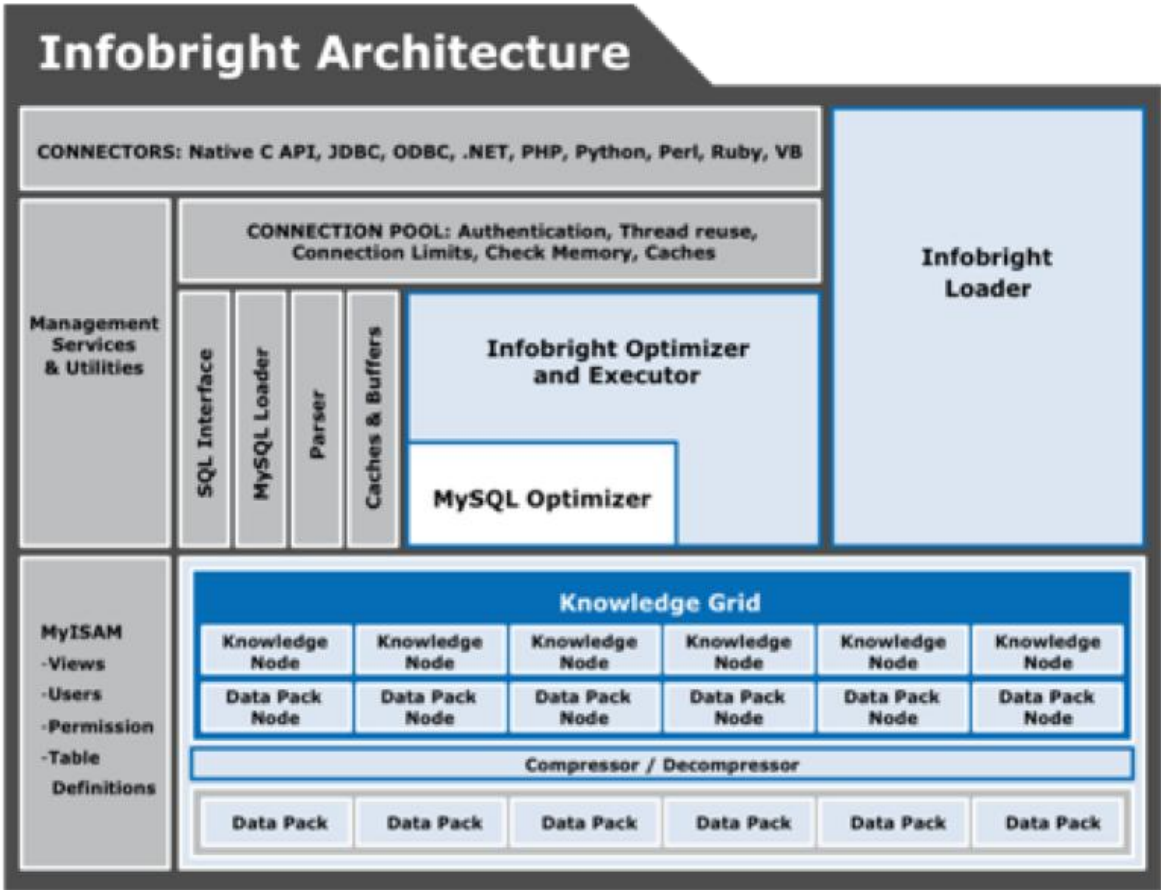
Infobright is built within the MySQL architecture and leverages the connectors and interoperability of MySQL standards. This integration with MySQL allows Infobright to tie in seamlessly with any ETL and BI tool that is compatible with MySQL and to leverage the extensive driver connectivity provided by MySQL connectors (including ODBC, JDBC, C/C++, .NET, Perl, Python, PHP, Ruby, Tcl and others).

MySQL also provides cataloging functions such as table definitions, views, users, permissions, etc. which Infobright stores in a MyISAM database.

Although MySQL provides a storage engine interface which is implemented in Infobright, the Infobright technology has its own advanced Optimizer that uses the column orientation and the Knowledge Grid to resolve queries. This integration provides a simple path to a high performance analytic database for MySQL users with no new interface to learn.

In addition, the MyISAM row-based transaction-oriented storage engine is available for use in conjunction with the Infobright storage engine. Customers have found this useful in providing an end-to-end solution requiring real-time data collection from a transaction-based source, and analytic capabilities in a data warehouse setting.

Figure 4 – Infobright and MySQL Architecture



Using ETL and BI Tools

Infobright has partnered with many of the leading open source and commercial ISVs in the data integration and business intelligence market. Examples include but are not limited to Talend, Jaspersoft, Pentaho, Actuate and MicroStrategy.

In order to ensure these solutions interface and interoperate well with our database solution, we have joined forces to certify and test our joint solutions. Furthermore, Infobright has adopted open and standard interfaces with popular solutions including ODBC and JDBC API’s to ensure a rich set of interoperability between our data integration and business intelligence partners that support these standards.

As previously mentioned, Infobright provides free connectors for Pentaho Data Integration, Talend Studio, JasperETL and their open source equivalents.

Hardware Considerations

Infobright’s industry-leading compression ratios mean an extremely low hardware footprint, and TCO. The database can support large volumes of data on a single server, with the following requirements and considerations:

Technical Considerations

Requirement	Description
Processor Architecture	* Intel 64 Bit or 32 Bit * AMD 64 Bit or 32 Bit
CPU Speed	* 2.0 GHz minimum * Dual or quad core recommended
Memory	• 4 GB minimum • 8 GB or more recommended
Operating Systems Support	• Red Hat Enterprise 5 • Debian "Lenny" • CentOS 5.2 • SUSE Linux Enterprise 10 • Windows Server 2003, 2008 • Solaris 10

CPU Considerations

Infobright supports both Intel and AMD architectures. Intel XEON Processors are recommended, and high performance systems can realize processing benefits from Intel "Nehalem" XEON processors.

The average number of concurrent queries is an important factor in determining the number of processors required to support Infobright. Typically, about 10% of connected users are running queries concurrently. For example 100 connected users will be running an average of 10 concurrent queries. To maximize performance, Infobright recommends 1 CPU core per concurrent query.

Database loads should be considered in addition to the CPU core concurrency value. Two cores per load process should suffice if loads are anticipated

when queries are running on the system. If loads are anticipated during a dedicated load window, then additional cores are not required.

Memory

A minimum of 2GB of RAM per concurrent query is required with a recommendation for 8GB or more depending on the type of data and queries running on Infobright. Additional memory should be made available if queries are anticipated to retrieve large amounts of data or are exceedingly complex in nature.

Infobright will utilize as much memory as possible and will retain the most recently used Data Packs in an uncompressed state. Additional queries using the same data will therefore gain performance enhancements, as this data is already available in an uncompressed state.

On-Disk Temporary Workspace Storage

Infobright utilizes an on-disk directory for temporary files necessary to complete query process. The amount of space available within this volume is very important for proper operation of the database, and depends on the query structure and the data size.

For example, a subquery run on a main fact table of 200 million rows returning 4 columns (2 numerical and 2 char) would require at least 16GB.

In the case of `COUNT(DISTINCT)` on large data sets the temporary space may be up to 20 bytes for each distinct value. For a billion random URLs, Infobright may need up to 20GB of temporary space to check whether or not all of them is distinct. For aggregations, the temporary files may be as big as the entire result of the aggregation. In the case of sorting, the temporary space may reach 32 bytes for each row for Infobright v3.2. This space requirement will increase with the upcoming support of UTF-8.

Best Practices for Embedding Infobright

Infobright technology is backed by a strong and responsive services and support organization dedicated to the successful adoption of our technology by partners and customers. Our expertise goes beyond the Infobright database to include other third party solutions including data integration and business intelligence solutions.

Infobright's best practices therefore cover not only our own product line but also those of our data integration and business intelligence partners. Our

methodology combined with a deep level of experience in data integration and business intelligence captures the best techniques and practices gathered by our service professionals through a wide variety of OEM experiences and implementations. These experiences have been very helpful to our OEM partners and our partners and customers frequently call us upon for our expertise in these product areas.

Points to consider about embedding Infobright:

- **Data Volumes.** Infobright is typically used for data volumes ranging from approximately 100 GB up to 50TB.
- **Analytic Environments**
 - Because of the column-orientation and the way the Knowledge Grid works to summarize and categorize relevant data for analytical reporting, Infobright best targets environments that are application-driven and analytical in nature rather than a traditional OLTP database environment.
- **Analytical Queries.** The highest level of performance is achieved for these type of queries:
 - Questions **about the data**
 - Aggregates: sums, counts, stats
 - Requests for some data that can be **optimized** by the Knowledge Grid

The following is a list of best practices for optimizing performance with Infobright:

- **Minimize JOINS.** As Infobright is a columnar database, it performs best when joins are minimized.
 - Reducing joins can be achieved by denormalizing, and because of Infobright's use of Data Packs and compression, this will have a minimal impact on the overall database size.
 - Infobright also improves JOIN performance by creating pack-to-pack Knowledge Nodes that describe a relationship between columns from different tables as defined by the query, and are created at query execution. With pack-to-pack KNs Infobright is able to maintain a high level of performance for queries that include multiple joins.
- **Increase efficiency of text-based searches.**
 - When using larger text based fields as constraints, performance can be improved by splitting the field up into 64 character column chunks.

- BLOBS can be broken up into small chunks that can be pieced together when needed.
- Always consider the 'LOOKUP' attribute for character fields where the number of unique values is less than 10,000.
- **Increase efficiency of date/time field searches.** When filtering on date/time fields, always use the complete format for that data type to avoid processing in MySQL execution path.
- **Increase efficiency at the ETL and load stages.**
 - Whenever possible, use the Infobright Loader to perform database population as opposed to direct database INSERTS.
 - Consider pre-sorting tables in the staging area by the most commonly filtered conditions.
 - Create/maintain surrogate keys, if needed, in the ETL/ELT/staging process for large text-based fields.
 - Consider occasional re-organization if a table has massive amount of DELETES or UPDATES.
- **Avoid, if possible:**
 - VIEWS with very little selectivity (filtering) on large tables.
 - Mixing INNER and OUTER JOINS.
 - Filtering with text strings that contain wildcard characters anywhere before the end of the string.

Migrating from MySQL/MyISAM to Infobright

Many users of MySQL turn to Infobright as their data volumes and analytic needs grow since Infobright offers exceptional query performance for analytic applications against large amounts of data. Migrating between MySQL's MyISAM storage engine, or other MySQL storage engines, to Infobright's column-oriented database is quite straightforward.

Infobright contains a bundled version of MySQL and installing Infobright installs a new instance of MySQL along with Infobright's Optimizer, Knowledge Grid, the Infobright Loader and the underlying columnar storage architecture. This installation also includes MySQL's MyISAM storage engine.

Unlike other storage engines that work with MySQL, it is not necessary to have an existing MySQL installation nor can Infobright be added to an existing MySQL Server installation. When installing Infobright, the assumption is that any previously existing MySQL or MyISAM database will exist in a separate installation of MySQL, installed in a different directory with a unique data path, configuration files, socket and port values.

Infobright has created a Unix-based shell script to assist with this scenario. Installation guides are available at <http://www.infobright.org/wiki>.

Specify the Database

As with all MySQL databases, Infobright has a unique storage type, or engine type that must be specified when creating tables within the Infobright schema by adding the following: `ENGINE=BRIGHTHOUSE`.

Schema

Infobright neither needs nor allows the manual creation of performance structures with duplicated data such as indexes or table partitioning based on expected usage patterns of the data. When preparing the MySQL schema definition for execution in Infobright, the first thing to do is simplify the schema. This means removing all references to indexes and other constraints expressed as indexes including `PRIMARY` and `FOREIGN KEYS`, and `UNIQUE` and `CHECK` constraints.

In addition, due to Infobright's extremely high query performance levels on large volumes of data, one should consider removing all aggregate, reporting and summary tables that may be in the data model as they are unnecessary.

Data Types

Infobright supports a large subset of MySQL data types and DDL. However in some cases, inconsistencies exist between Infobright and MyISAM tables in data type support, e.g. `UNSIGNED INTEGERS` and the `AUTO_INCREMENT` attribute. These and other cases are being addressed in development for upcoming Infobright releases. Detailed information about supported data types is available at

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

Data Loading

Finally, the last step is the physical transportation of the data. Infobright provides migration utilities for this purpose, and several other migration tools exist with which Infobright is compatible. ETL tools can also satisfy the task of one-time data migrations and Infobright has created high-speed connectors for these popular ETL tools: 1. Pentaho Data Integration (PDI), and 2. JasperETL from Jaspersoft (which is OEMed from Talend) or Talend Open Studio.

But for a one-time data migration, the simplest method might be the best, which is exporting the data from MyISAM tables and loading into Infobright

with the Infobright Loader. Both “data out” and “data in” methods use standard MySQL syntax.

Detailed information about data loading syntax is available at http://www.infobright.org/wiki/Data_Loading/

Database Maintenance

Once loaded an administrator can interact with the database just as they would any other in MySQL. Commands such as `SHOW DATABASES`, `USE <database>`, `SHOW TABLES`, `SHOW TABLE STATUS`, `SELECT COUNT(*) FROM TABLE`, etc., all work as expected.



Infobright's OEM Partner Program

The Infobright OEM Partner Program provides a full suite of product, pricing, service and support offerings specifically designed for ISVs and SaaS providers. The program includes a flexible licensing model that provides access to Infobright technology during product development, and scales to support the partner's growing business requirements.

Summary

Infobright is the ideal embedded analytic database for ISVs and SaaS providers. Using Infobright you can:

- Keep more of your customers' data on-line
- Enhance the functionality of your application or SaaS offering with greater flexibility for reporting and analytics
- Load data faster
- Reduce administrative effort by up to 90% by eliminating the need for indexes, data partitioning or database tuning
- Reduce storage requirements and hardware footprint
- Speed time-to-market for your application or service
- Reduce costs

The reduction in hardware infrastructure and administrative efforts combined with our special OEM pricing means that you can offer your customers

enhanced functionality at lower cost, increasing your product margins and providing an opportunity for greater revenue.

For more information or to download a free trial of Infobright Enterprise Edition, please go to www.infobright.com or contact us via email at oem@infobright.com.