The Database for Analytic Applications

April 13, 2010

David Lutz
Director, Technical Sales Consulting
Agenda

- Infobright Technology Overview
- Use Cases and Case Studies
- Migration to Infobright
- 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 Generally Available
- > 120 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
Challenging Times

More data

- More online activity → more web data
- Growth of mobile → more call data, web data
- Servers/networks → lots of log/event data

With increasing value in the details

- Target individual customers
- Identify micro-segments
- Find security threats
- Identify fraud

“Enterprise data growth over the next 5 years is estimated to be 650%.” Gartner
Challenging Times

More requirements

- More users
- Diverse demands
- More data sources

With less

- Time
- Resources
- Money

“The universe of applications for which analytics is now an important component continues to expand.” Wells Fargo Equity Research
Analytic Infrastructure Requirements

- Handles large data volumes with less cost and complexity
- Meets business users needs
  - Fast query response – static and *ad hoc* queries
  - Fast access to new data
  - Access to detailed data, not just aggregates
- Takes less IT time
  - Easy to implement
  - No complex hardware configuration
  - No index creation, data partitioning or manual tuning
- Lower cost
Infobright is a high performance analytic database that delivers fast query performance against large volumes of data with minimal IT effort.
What is Unique about Infobright?

- Uses intelligence, not hardware, to drive query performance:
  - Creates information about the data (metadata) upon load, automatically
  - Uses metadata to eliminate or reduce the need to access data to respond to a query
  - The less data that needs to be accessed, the faster the response

- What this means to you:
  - No need to partition data, create/maintain indexes or tune for performance
  - ad hoc queries are as fast as static queries, so users have total flexibility
  - ad hoc queries that may take hours with other databases run in minutes; queries that take minutes with other databases run in seconds
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

<table>
<thead>
<tr>
<th>Employee_ID</th>
<th>Job</th>
<th>Dept</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shipping</td>
<td>Operations</td>
<td>Toronto</td>
</tr>
<tr>
<td>2</td>
<td>Receiving</td>
<td>Operations</td>
<td>Toronto</td>
</tr>
<tr>
<td>3</td>
<td>Accounting</td>
<td>Finance</td>
<td>Boston</td>
</tr>
</tbody>
</table>

Data stored in rows

Data stored in columns
1. Column vs. Row Orientation - Use Cases

**Row-Based Storage**

<table>
<thead>
<tr>
<th>ID</th>
<th>job</th>
<th>dept</th>
<th>city</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Row Oriented works if...**

- All the columns are needed
- Transactional processing is required

**Column-Based Storage**

<table>
<thead>
<tr>
<th>id</th>
<th>job</th>
<th>dept</th>
<th>city</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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
- Reading column takes similar CPU resources as reading a row
2. Data Packs and Compression

Data Packs
- Each data pack contains 65,536 data values
- Compression is applied to each individual data pack
- The compression algorithm varies depending on data type and distribution

Compression
- Results vary depending on the distribution of data among data packs
- A typical overall compression ratio seen in the field is 10:1
- Some customers have seen results of 40:1 and higher
- For example, 1TB of raw data compressed 10 to 1 would only require 100GB of disk capacity
2. 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
3. The Knowledge Grid

Knowledge Grid
applies to the whole table

Knowledge Nodes
built for each Data Pack

Information about the data

- Knowledge Nodes answer the query directly, or
- Identify only relevant Data Packs, minimizing decompression
Data Pack Nodes ...

This KN contains \textit{statistical} and \textit{aggregate} values for the Data Pack:

- MINIMUM value
- MAXIMUM value
- COUNT of all elements
- SUM of all values
- No. of NULLs

<table>
<thead>
<tr>
<th>MIN</th>
<th>MAX</th>
<th>COUNT</th>
<th>SUM</th>
<th>No. NULLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25000</td>
<td>65536</td>
<td>58003500</td>
<td>1000</td>
</tr>
</tbody>
</table>

- DPNs help \textbf{optimize} the search by \textbf{minimizing} the need to decompress data.
- DPNs alone often contain enough information to \textbf{resolve} a query.
3. Knowledge Grid Nodes - Histograms

**Numerical Histograms ...**

The MIN-MAX range from the DPN is divided into **1024 intervals**.

This KN is a *binary representation* of whether a numerical value exists within each interval.

If the MIN-MAX range is < 1024, then each ‘interval’ is a distinct value.

<table>
<thead>
<tr>
<th>1 - 24</th>
<th>25 - 48</th>
<th>49 - 72</th>
<th>...</th>
<th>24577 - 25000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

➢ Numerical Histograms are very efficient at minimizing the Data Packs required to resolve a query with numerical constraints.
3. Knowledge Grid Nodes - CMAPs

Character Maps ...

The first 64 positions of text fields are read.

This is a binary representation of the occurrence of every possible character within the first 64 positions.

CMAPs are very efficient at resolving text-based search queries that involve the beginnings of strings.
Pack-to-Pack Nodes (P-2-P) ...

A fourth type of Knowledge Node is created by a JOIN query.

P-2-P nodes describe *relationships* between the Data Packs of columns of joined tables.

- P-2-P Nodes are stored in memory and persisted during a session.
- Query performance improves as joins are created and re-used.
- Best practice is to “warm up queries” to pre-establish P2P??
4. Optimizer

1. Query received
2. Optimizer iterates on Knowledge Grid
3. Each pass eliminates Data Packs
4. If any Data Packs are needed to resolve query, only those are decompressed

Query:
Q: How are my sales doing this year?
A Simple Query using the Knowledge Grid

```
SELECT COUNT(*) FROM employees
WHERE salary > 100000
    AND age < 35
    AND job = 'IT'
    AND city = 'San Mateo';
```

1. Find the Data Packs with salary > $100,000
2. Find the Data Packs that contain age < 35
3. Find the Data Packs that have job = ‘IT’
4. Find the Data Packs that have City = ‘San Mateo’
5. Now we eliminate all rows that have been flagged as irrelevant.
6. Finally we have identified the data pack that needs to be decompressed
Infobright in the Microsoft Application Stack

Visual Studio (.NET)  SSRS  SSAS  SSIS  etc …

.NET, ADO, ODBC, JDBC, etc. …

MySQL-provided Drivers

Infobright
Fast query response with no tuning or indexes

<table>
<thead>
<tr>
<th>Customer’s Test</th>
<th>Alternative</th>
<th>Infobright</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytic queries</td>
<td>2+ hours with MySQL</td>
<td>&lt;10 seconds</td>
</tr>
<tr>
<td>1 Month Report (15MM events)</td>
<td>43 min with SQL Server</td>
<td>23 seconds</td>
</tr>
<tr>
<td>Oracle query set</td>
<td>10 seconds – 15 minutes</td>
<td>0.43 – 22 seconds</td>
</tr>
<tr>
<td>BI report</td>
<td>7 hours in Informix</td>
<td>17 seconds</td>
</tr>
<tr>
<td>Data load</td>
<td>11 hours in MySQL ISAM</td>
<td>11 minutes</td>
</tr>
</tbody>
</table>
## Designed For Analytics

<table>
<thead>
<tr>
<th>Primary Use Case</th>
<th>Best Fit Dynamic Analytics</th>
<th>Good Fit Static Analytics</th>
<th>Not A Fit Heavy OLTP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query Types</strong></td>
<td>• Many <em>ad hoc</em> queries</td>
<td>• Limited mixed workloads</td>
<td>• Heavy transactions</td>
</tr>
<tr>
<td></td>
<td>• Near real-time response</td>
<td>• End of day reports</td>
<td>• Many stored</td>
</tr>
<tr>
<td></td>
<td>• Fast data load speeds</td>
<td>• Some changing data</td>
<td>procedures</td>
</tr>
<tr>
<td></td>
<td>• Big Data / Fast Queries</td>
<td>• Simple JOINs</td>
<td>• Heavy referential</td>
</tr>
<tr>
<td></td>
<td>• Wide Tables</td>
<td>• Wide range of tool options</td>
<td>integrity</td>
</tr>
<tr>
<td></td>
<td>• Aggregates: COUNT, SUM, etc.</td>
<td>• Batch load feeds</td>
<td>• Zero downtime</td>
</tr>
<tr>
<td></td>
<td>• Deep Compression</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Rapid Deployment / Ease of Use</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Example Query</strong></td>
<td>• Analytic-intensive queries</td>
<td>• Mixed workload queries</td>
<td>• Lots of insert/deletes</td>
</tr>
<tr>
<td></td>
<td>• Standard data types</td>
<td>• Data mart-type BI queries</td>
<td>• Frequent changes</td>
</tr>
<tr>
<td></td>
<td>• Limited JOINs</td>
<td>• SQL standard queries</td>
<td>• Updates across</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tables</td>
</tr>
<tr>
<td><strong>Example Query</strong></td>
<td>• Average clicks per visit</td>
<td>• Sales numbers by region</td>
<td>• UPDATE balance in account(s)</td>
</tr>
<tr>
<td></td>
<td>• Total number of visits</td>
<td>• Average selling price by rep</td>
<td>• Abandon current cart</td>
</tr>
<tr>
<td></td>
<td>• Total visit time</td>
<td>• Num items sold by product</td>
<td>• DELETE all accounts</td>
</tr>
<tr>
<td></td>
<td>• Total bounce rate</td>
<td>• New customers in month</td>
<td>• over 1 year old</td>
</tr>
<tr>
<td></td>
<td>• UNIQUE values</td>
<td>• Products not sold</td>
<td>• INSERT new</td>
</tr>
<tr>
<td></td>
<td>• Largest sales made</td>
<td>• Top selling product</td>
<td>accounts</td>
</tr>
<tr>
<td></td>
<td>• Number of customers in region</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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**
A leader in mobile billing and analytics services utilizing a SaaS model

Received a contract with a large media provider
- 150 million rows per month
- 450GB per month on SQL Server

SQL Server could not support required query performance

Needed a database that could
- scale for much larger data sets
- with fast query response
- with fast implementation
- and low maintenance
- in a cost-effective solution

### Infobright’s Solution

- Reduced queries from minutes to seconds

<table>
<thead>
<tr>
<th>Query</th>
<th>SQL Server</th>
<th>Infobright</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Month Report (5MM events)</td>
<td>11 min</td>
<td>10 secs</td>
</tr>
<tr>
<td>1 Month Report (15MM events)</td>
<td>43 min</td>
<td>23 secs</td>
</tr>
<tr>
<td>Complex Filter (10MM events)</td>
<td>29 min</td>
<td>8 secs</td>
</tr>
</tbody>
</table>

- Reduced size of one customer’s database from 450 GB to 10 GB for one month of data
Project requirements

- Executive dashboard / reporting tool with flexible reporting options for business users with multiple levels of detail
- Required ability to consolidate large volumes of data from multiple sources
- Request had been outstanding for over a year – needed solution that could be implemented quickly, at low cost, without central IT effort
## Austin Energy Results

### Changes

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Volumes Available</strong></td>
<td>1 month of data</td>
<td>2 years of data (2.5 million rows, 73 columns)</td>
</tr>
<tr>
<td><strong>Compression</strong></td>
<td>1 : 1</td>
<td>98 : 1</td>
</tr>
<tr>
<td><strong>Data Footprint</strong></td>
<td>5+ GB</td>
<td>5 GB now occupies &gt;100 MB</td>
</tr>
<tr>
<td><strong>Age of Data Available</strong></td>
<td>6 weeks old</td>
<td>1 day old</td>
</tr>
<tr>
<td><strong>Reporting Time / KPI</strong></td>
<td>Once per month</td>
<td>&lt; 10 sec on demand</td>
</tr>
<tr>
<td><strong>End Users</strong></td>
<td>Limited, isolated business units</td>
<td>Organizationally secure access to data</td>
</tr>
</tbody>
</table>

### Notes

| **Future Plans**                      | Continued expansion into other departments |
| **Lessons Learned**                   | This new technology eliminates or reduces the need for aggregate tables! Software enables agile project approach |
| **Time to Deploy**                    | 2 Project Managers, part-time, over 6 weeks met all the requirements |
Migrating to Infobright

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.
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
Migration Approaches

1. **Manual execution of individual tasks**
   - Microsoft SQL Server and Sybase provide bcp
   - LOAD DATA INFILE '/full_path' INTO TABLE table_name [FIELDS [TERMINATED BY 'char'] [ENCLOSED BY 'char'] [ESCAPED 'char']] ;

2. **Utilities provided by the vendor**
   - Includes ICE Breakers for SQL Server, Oracle, MyISAM and InnoDB
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/
The ICE Breaker for SQL Server

SQL Server
Server Name: 192.168.1.75
Database: EDW01
Login: sa
Password: ****

File Locations
Export File Directory: C:\temp
Import File Directory: /home/mysql/
Script File: C:\Users\carl\Desktop\load.txt

Go!  Forum  Cancel
3. One-time migration tools
   • Most commonly seen tool is SQLWays from Ispirer
   • http://www.ispirer.com

4. Data movement tools
   ▪ open source software (OSS) options
     ▪ Jaspersoft ETL from Jaspersoft
     ▪ Talend Open Studio
     ▪ Pentaho Data Integration, or PDI (aka Kettle) from Pentaho
     ▪ CloverETL
     ▪ SwisSQL
   ▪ commercial off-the-shelf (COTS) options
     ▪ Informatica PowerCenter
     ▪ IBM Ascential DataStage
     ▪ etc.
When is Infobright a “Must Have”?

- Users Need it Now
- Load Window cannot be met
- Easy install and operations required
- Cost Pressures Point toward Open Source
- Data Growth Impacting Performance
- Mixed Workload of AdHoc and Standard Queries

Database for Analytic Applications
Get Started

- 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, Talend or Actuate/BIRT at www.infobright.org
- Download a free trial of Infobright Enterprise Edition

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