MySQL V5 – Ready for Prime Time Business Intelligence

Seth Grimes

Alta Plana Corporation

A Pentaho^{TM} Sponsored White Paper

pentaho™ open source business intelligence™

Table of Contents

Executive Summary

Business intelligence (BI) creates competitive advantage for organizations of all sizes and in all industries. It is an important business tool, a means of exploiting data previously locked in operational databases.

With the fall 2005 release of MySQL Version 5, conditions for introducing and expanding open-source enterprise BI programs have never been better. MySQL capabilities have grown to meet the most demanding BI performance and scalability requirements. Open-source now similarly presents an attractive alternative to expensive, propriety, closed-source BI software.

Business intelligence is the key to fact-based decision making in the intelligent enterprise. Techniques and tools are designed to unlock knowledge from data. They will help you:

- Understand past performance
- Monitor current activities and respond quickly to changing conditions
- Optimize business processes and performance
- Manage risk
- Forecast future prospects and plan accordingly.

If your organization generates data, BI will help you use it for competitive advantage.

BI relies on strong database management as provided by DBMSs such as MySQL, which with Version 5 now competes in performance, capabilities, and reliability with other market-leading relational database systems. New and established MySQL features – a choice of engines for optimal performance, ACID transactions, views, stored procedures, triggers, and partitioning – match those previously found only in closed-source, commercial database systems. MySQL users and developers can now confidently extend their investments to provide even greater benefit to their organizations. MySQL is now not only an established DBMS standard for Web sites and operational applications, it is also an exceptional platform for high-value business intelligence functions.

Evaluators should look for software options offering a full range of BI capabilities. Many organizations will want to start simple with basic desktop reporting functions. Others will prefer to immediately implement a full analytics program with enterprise reporting, dashboards, OLAP, and data mining. Workflow facilities, offering the abilities both to manage BI components and to allow external orchestration of BI processes, are another important component of modern BI systems.

Pentaho Business Intelligence takes advantage of MySQL 5 scalability and performance enhancements and provides a comprehensive suite of best-of-breed BI tools – the only complete open-source suite – that matches the capabilities of expensive, closed BI systems. Pentaho is a MySQL Gold Partner; the software is designed to integrate seamlessly with MySQL databases and support a wide range of business initiatives including sales and profitability analysis, customer analysis, HR reporting, financial reporting, KPI dashboards, supply chain analytics, and operational reporting, all with unique workflow capabilities. Users may download individual programs or the complete suite, all free of charge and community backed but also enhanced by optional commercial support and an optional Professional version, at *www.pentaho.org*.

1 Evolution of the DBMS & BI markets

The Open Source Revolution and Database Management

The database management system (DBMS) and business intelligence (BI) markets are mature, yet much room remains for innovative technologies and business models. The two markets are really quite similar. Both feature a number of large, well-established vendors that offer roughly comparable technologies, with value-added services provided by the companies themselves and an assortment of consultants and developers.

MySQL emerged as a low-cost entry in a crowded DBMS market, but MySQL's selling point was never just or even primarily price: It was ease of use and administration and absolute suitability for important applications, in particular as the database back-end of choice for Web publishing. MySQL has matured in recent years. The technology now rivals that of established closed-source vendors for the broad set of traditional database applications, which we can classify as publishing, operational, and analytical. MySQL is able to compete in these areas because its modular architecture allows you to choose the storage engine that performs best for a spectrum of needs:

- InnoDB for transactional systems
- MyISAM for analytical systems including data warehouses and data marts
- Memory, formerly known as Heap, for high-performance applications
- NDB, the Cluster Storage engine, for high availability and scalability
- Archive for efficient storage of large data volumes
- Federated, providing for local access to remote data tables
- Merge, also known as MRG_MyISAM, which collects identical MyISAM tables for unified access.

The different engines share common administration and query interfaces, and MySQL even allows you to select engines on a table-by-table basis within a database. MySQL has a single version of SQL and has a smart optimizer that insulates developers and users from the technical details that distract from their focus on delivering the best possible applications.

MySQL Version 5, released in October 2005, addressed important enterprise concerns. Version 5 adds must-have features including ACID (atomic, consistent, isolated, and durable) transactions, distributed transaction processing conforming to the X/Open XA specification, triggers and stored procedures, and views. Version 5 also notably brings MySQL's internal metadata schema into line with relational standards.

Further enhancements are slated for forthcoming releases. Among these enhancements are support for table partitioning, an important scalability feature for “big-data” uses such as data warehouses.

Given its new capabilities, MySQL is now a DBMS of choice for data warehouses and data-analysis applications.

Next Up, BI

The business intelligence market is primed for an open-source make-over akin to what has occurred in the operating-system, database, application-server, and development-tool markets.

Business intelligence is a set of tools and techniques designed to unlock knowledge from data. BI lets organizations derive the maximum value from their information assets. It helps users:

- Understand past performance
- Monitor current activities and respond quickly to changing conditions
- Optimize business processes and performance
- Manage risk
- Forecast future prospects and plan accordingly.

BI exploits organizational data for competitive advantage. It is the key to fact-based decision making in the intelligent enterprise.

BI has long been dominated by a number of commercial-product vendors in areas including reporting, on-line analytical processing (OLAP), data mining, extract, transform, and load (ETL), and dashboards and visualization.

The best BI tools are process-centric, providing integrated workflow options. Best-of-breed workflow involves not only the ability to manage processes within the overall BI platform, it also allows the BI functions to be externally orchestrated by larger applications that require embedded analytics. Workflow capabilities provide for managed BI and for the flexibility to offer users BI within line-of-business applications.

Open-source BI alternatives were slow to emerge, but the market has flowered in recent years with the appearance of tools in each of the major categories.

Open-source reporting options are now plentiful. They include Eclipse BIRT (Business Intelligence and Reporting Toolkit), JasperReports, and JFreeReport. Several projects cover the spectrum of charting, dashboard, and visualization functions. Mondrian is a “relational OLAP” server that relies on a back-end RDBMS, accessed via JDBC, for data management. JPivot provides a Java Server Page tag library that can utilize Mondrian and other analytical services via MDX queries. Weka is the most prominent open-source data mining and machine learning software, and the Kettle ETL project has gained popularity for its ease of use and ability to integrate real-time and historical data.

Open-source business intelligence options actually offer technical advantages over the closed-source tools. They can incorporate features missing from mature tools, such as integrated workflow management (which treats BI as the central business process it is) and designed-in interoperability with commercial and open-source tools of the same category. Open-source BI tools share both cost and technical advantages with open-source stars including MySQL, JBoss, Apache, and Linux.

2 MySQL As a Business Intelligence Platform

Analytical and Transactional Databases

Business intelligence is a process that encompasses acquiring and managing information, designing analytical interfaces, and delivering actionable business insights. Because BI's value hinges on availability of timely, accurate, accessible information, strong data management is essential. The quality and performance of analytical databases – data warehouses, data marts, and operational data stores (ODSs) for real-time analytics – are critical.

Data warehouses are databases optimized for data analysis rather than for transaction processing. They typically house many gigabytes of data and are structured using dimensional modeling techniques – star schemas – to provide rapid response to complex queries. Many data warehouses store textual and geospatial data in addition to numerical data. By including harmonized, cleansed metadata – information that describes the tables, fields, and value sets – data warehouses are able to host diverse applications that range from structured reporting and performance dashboards to ad-hoc query and intensive statistical data mining.

Whether you are creating data warehouses or data marts – datasets specially designed to respond to the analytical needs of particular users or applications – or both, MySQL's MyISAM engine provides the fast bulk, incremental data loading and indexing, and the responsiveness, needed to support large data volumes and diverse, complex queries. MySQL is also, of course, fully capable of supporting real-time data analysis that works directly off operational data stores managed with InnoDB or another of MySQL's engines.

Because MySQL provides a standard SQL implementation and application programming interfaces (APIs) usable across the complete set of engines, the user has the flexibility to run analyses off an appropriately structured data warehouse, data mart, or operational data store.

Design Choices

Data-warehouse and ODS designers must make important decisions that start with selection of a database platform. With new Version 5 capabilities, MySQL is an option for data warehouses and ODSs extending into hundreds of gigabytes. Capacity, features, and ease of administration make MySQL an excellent choice for new users and the best bet for sites that already use MySQL for publishing and transactional systems.

MySQL database size is limited only by the operating system's support for large files and by hardware capacity. The software will support anything from small databases to ones containing terabytes of data. Performance options include standard techniques such as locating indexes and data files on different disk volumes and index caching, also possibilities not found in competing systems, notably the choice of static, dynamic, and compressed record structures. The record-structure possibilities allow the database designer flexibility in balancing performance and space-utilization considerations. Additionally, MySQL's Merge engine allows database partitioning by collecting distributed MyISAM tables for unified access. The forthcoming Version 5.1 will include explicit table-partitioning options.

MySQL supports embedding analytical functions in the database close to the data. This feature allows the user to create derived measures including key performance indicators (KPIs) that are then available for shared use across applications. Extending the database in this way is a convenient alternative to implementing functions in application code, one that may have significant performance advantages. There are three options:

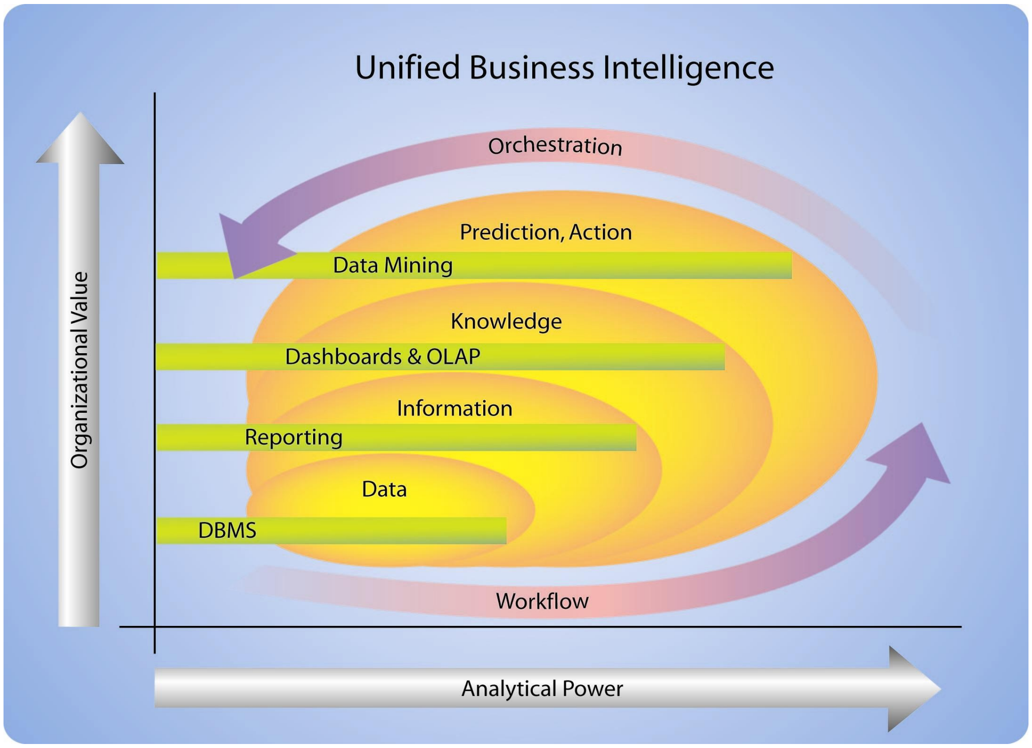
- User Defined Functions act on single rows or aggregate across rows, returning string or numerical values. They must be written in C or C++ and compiled into dynamically loaded, shared object code.
- Stored procedures are sets of SQL statements managed by the server. MySQL implements SQL-2003 syntax, which is also used by IBM's DB2. And stored procedures may be invoked via triggers that are activated in response to designated events.
- Native functions are built into MySQL's source code; MySQL is, after all, open source, designed for extension by a community of contributors.

The new functions and procedures are usable via SQL queries.

The system further offers integrated full-text indexing and search as well as r-tree indexes on spatial data with a number of functions that return geometric results. Any column of standard char, varchar, and text types can have a full-text index. Text-search functions using Boolean expressions and two-pass expanded searches are built into out-of-the-box SQL. MySQL's spatial extensions are a subset of the *SQL with Geometry Types* environment that conforms to specifications published by the Open Geospatial Consortium (OGC).

Lastly, Version 5 adds updatable views to the mix. Views, or virtual tables, can prove a plus for data-analysis applications. They are queries that function like tables, retrieving selected data columns and rows on request from joined database tables.

Given MySQL's capabilities and capacity, data warehouse, data mart, and ODS developers can rely on the platform to support the data-loading procedures, star schemas, and analytical interfaces needed to deliver a broad range of business intelligence functions.



3 Extending MySQL with Pentaho Business Intelligence

An Easy Add-On

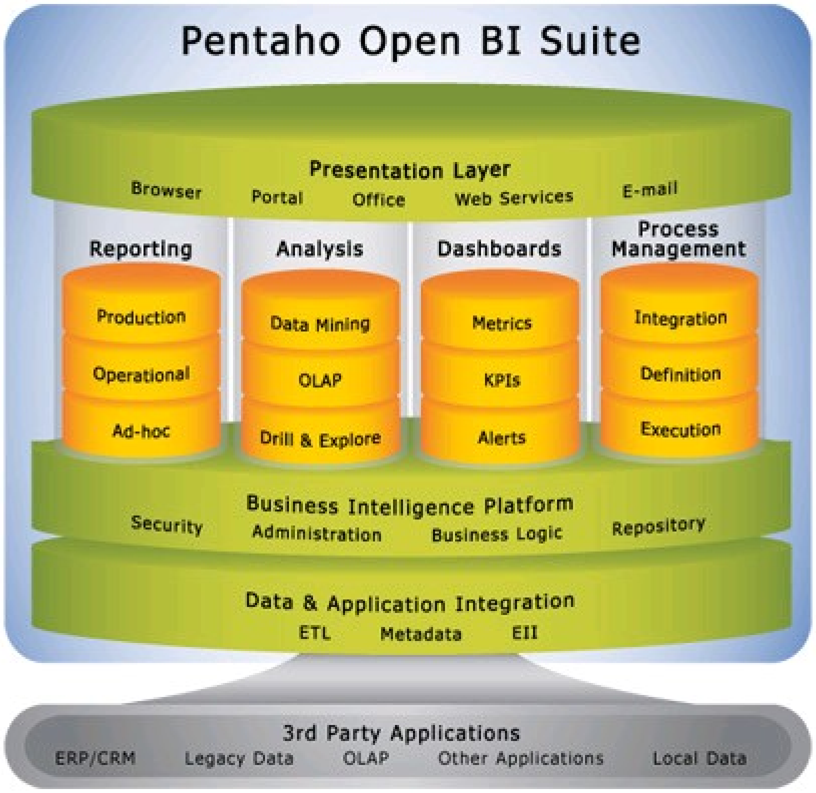
Pentaho Business Intelligence answers the question: "I'm generating and managing business-critical data. How do I unlock the greatest value from it?"

Pentaho supports the spectrum of BI functions (enterprise-class reporting, analysis, dashboard, and data mining) with workflow capabilities that help organizations operate more efficiently and effectively.

The software is flexible. You can embed components in other applications, create customized BI solutions, or deploy the suite as a complete, out-of-the-box, integrated BI platform.

Pentaho is an easy add-on for MySQL users. It is open-source and available as a free download. It is coded in Java and operates on all leading application servers including open-source options JBoss and Tomcat. Pentaho's development and production environments are designed for openness, scalability, and interoperability. The software relies on industry standards and open components such as JDBC, MDX, XML/A, Apache, and the JSR 168 portlet specification.

You can exploit the whole suite or you can choose to start simple and grow.

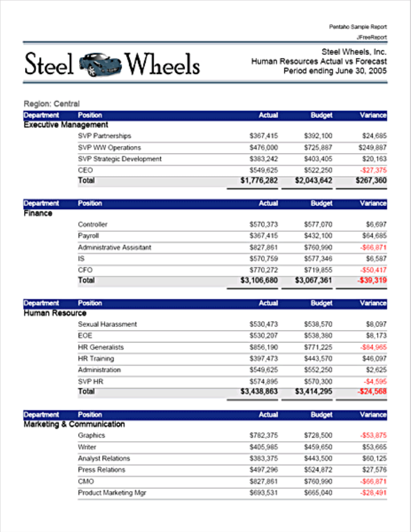


Quick-start: Reporting

Reporting operational and analytical data is *the* basic BI function. With Pentaho, you can start with desktop reporting and then optionally deploy reports within Web-based applications using the Pentaho Reporting Server. Some companies will simply want to download and run Pentaho in standalone mode to start building, scheduling, and distributing reports. Others will want to embed and integrate Pentaho into their own applications, using only the components they need, and customize it to fit their application.

Pentaho reports are defined in XML. Developers can work from samples directly or within the open-source Eclipse integrated development environment using Pentaho's Eclipse plug-in. Pentaho also provides a reporting wizard that guides business users through the basic steps of:

- Picking the data source
- Selecting the information for reporting
- Defining the layout template
- Choosing where the report should go, when it should run, and who gets it.



Pentaho Sample Report
JFreeReport

Steel Wheels

Steel Wheels, Inc.
Human Resources Actual vs Forecast
Period ending June 30, 2005

Region: Central

Department	Position	Actual	Budget	Variance
Executive Management				
	SVP Partnerships	$367,415	$392,100	$24,685
	SVP WW Operations	$476,000	$725,887	$249,887
	SVP Strategic Development	$383,242	$403,405	$20,163
	CEO	$549,625	$522,250	-$27,375
	Total	$1,776,282	$2,043,642	$267,360

Department	Position	Actual	Budget	Variance
Finance				
	Controller	$570,373	$577,070	$6,697
	Payroll	$367,415	$432,100	$64,685
	Administrative Assisitant	$827,861	$760,990	-$66,871
	IS	$570,759	$577,346	$6,587
	CFO	$770,272	$719,855	-$50,417
	Total	$3,106,680	$3,067,361	-$39,319

Department	Position	Actual	Budget	Variance
Human Resource				
	Sexual Harassment	$530,473	$538,570	$8,097
	EOE	$530,207	$538,380	$8,173
	HR Generalists	$856,190	$771,225	-$84,965
	HR Training	$397,473	$443,570	$46,097
	Administration	$549,625	$552,250	$2,625
	SVP HR	$574,895	$570,300	-$4,595
	Total	$3,438,863	$3,414,295	-$24,568

Department	Position	Actual	Budget	Variance
Marketing & Communication				
	Graphics	$782,375	$728,500	-$53,875
	Writer	$405,985	$459,650	$53,665
	Analyst Relations	$383,375	$443,500	$60,125
	Press Relations	$497,296	$524,872	$27,576
	CMO	$827,861	$760,990	-$66,871
	Product Marketing Mgr	$693,531	$665,040	-$28,491

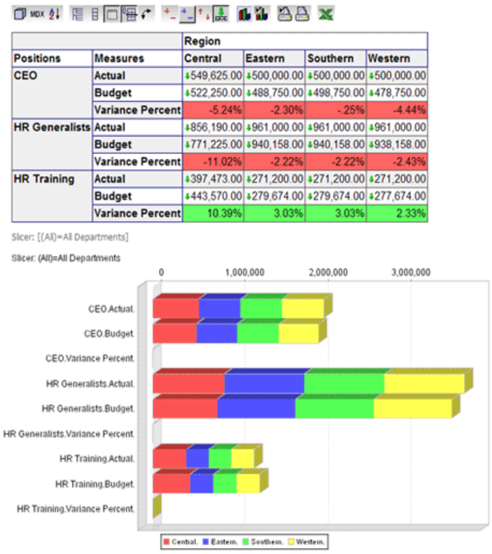
Report components – styles, data sources, queries, parameters and parameter groups – can be saved and loaded when designing a report.

The Pentaho Reporting Server was designed for high-volume, high-performance, secure enterprise reporting, both on-demand and via an integrated scheduler. It supports bursting – running and delivering components of a single report tailored for specific recipients – and other business-rules based functions. In addition to the unique workflow components, the server supports a variety of report formats: not only Pentaho's own JFreeReports but also JasperReports, and Eclipse BIRT. Pentaho is designed for scalability and interoperability.

Reporting is only a start. It is easy to provide further BI capabilities as users demand them.

Pentaho Analytics

Online analytical processing (OLAP) is a first step beyond reporting into true analytics. OLAP provides interactive, slice-and-dice analysis of data structured in multi-dimensional cubes.



		Region			
Positions	Measures	Central	Eastern	Southern	Western
CEO	Actual	549,625.00	500,000.00	500,000.00	500,000.00
	Budget	522,250.00	488,750.00	498,750.00	478,750.00
	Variance Percent	-5.24%	-2.30%	-.25%	-4.44%
HR Generalists	Actual	856,190.00	961,000.00	961,000.00	961,000.00
	Budget	771,225.00	940,158.00	940,158.00	938,158.00
	Variance Percent	-11.02%	-2.22%	-2.22%	-2.43%
HR Training	Actual	397,473.00	271,200.00	271,200.00	271,200.00
	Budget	443,570.00	279,674.00	279,674.00	277,674.00
	Variance Percent	10.39%	3.03%	3.03%	2.33%

Slicer: [(All)=All Departments]

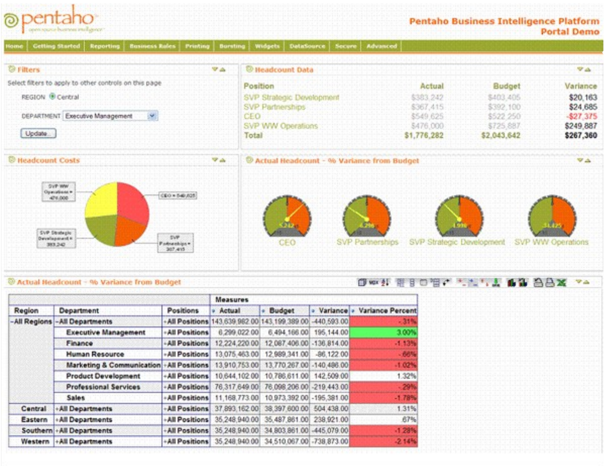
Slicer: (All)=All Departments

Pentaho's OLAP capabilities provide for visually driven exploratory data analysis with rich roll-up data aggregation, drill-down, and drill-through to underlying data records.

Mondrian, which provides the core of Pentaho Analysis Services, is a relational OLAP (ROLAP) engine, relying on a back-end relational database for data storage and query execution. MySQL is an ideal OLAP DBMS although Mondrian can use any JDBC data source. Developers can get started by downloading a sample schema and XML files that define the data cubes, the "business objects." Developers can define new objects by mapping their own database schemas into cubes with XML. The Pentaho interface will provide all the necessary end-user data selection and manipulation capabilities.

Dashboard Delivery

Dashboards are the leading method of packaging reports and analyses with charts and other visual elements for on-line delivery to staff throughout the enterprise. With Pentaho, dashboards can be customized based on business role (such as executive, manager, or analyst), department or subject matter, and even for particular individuals. And with Pentaho, dashboards can integrate with other BI components via Pentaho's unique workflow-management capabilities including task lists, escalation, routing, and tracking.


pentaho
open source business intelligence
Pentaho Business Intelligence Platform
Portal Demo
Home
Getting Started
Reporting
Business Rules
Printing
Bursting
Widgets
DataSource
Secure
Advanced
Filters
Headcount Data
Headcount Costs
SVP WW Operations = 476,000
CEO = 549,625
SVP Strategic Development = 383,242
SVP Partnerships = 367,415
Actual Headcount - % Variance from Budget
CEO
SVP Partnerships
SVP Strategic Development
SVP WW Operations
Actual Headcount - % Variance from Budget
Update...

Select filters to apply to other controls on this page

REGION ● Central

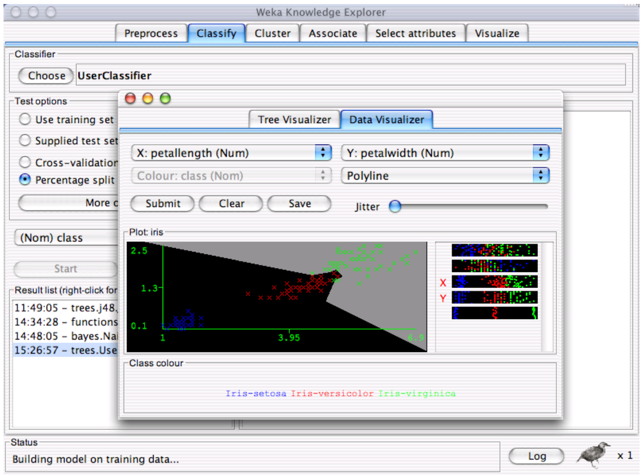
DEPARTMENT Executive Management

Position	Actual	Budget	Variance
SVP Strategic Development	$383,242	$403,405	$20,163
SVP Partnerships	$367,415	$392,100	$24,685
CEO	$549,625	$522,250	-$27,375
SVP WW Operations	$476,000	$725,887	$249,887
Total	$1,776,282	$2,043,642	$267,360

			Measures			
Region	Department	Positions	Actual	Budget	Variance	Variance Percent
-All Regions	-All Departments	+All Positions	143,639,982.00	143,199,389.00	-440,593.00	-.31%
	Executive Management	+All Positions	6,299,022.00	6,494,166.00	195,144.00	3.00%
	Finance	+All Positions	12,224,220.00	12,087,406.00	-136,814.00	-1.13%
	Human Resource	+All Positions	13,075,463.00	12,989,341.00	-86,122.00	-.66%
	Marketing & Communication	+All Positions	13,910,753.00	13,770,267.00	-140,486.00	-1.02%
	Product Development	+All Positions	10,644,102.00	10,786,611.00	142,509.00	1.32%
	Professional Services	+All Positions	76,317,649.00	76,098,206.00	-219,443.00	-.29%
	Sales	+All Positions	11,168,773.00	10,973,392.00	-195,381.00	-1.78%
Central	+All Departments	+All Positions	37,893,162.00	38,397,600.00	504,438.00	1.31%
Eastern	+All Departments	+All Positions	35,248,940.00	35,487,861.00	238,921.00	.67%
Southern	+All Departments	+All Positions	35,248,940.00	34,803,861.00	-445,079.00	-1.28%
Western	+All Departments	+All Positions	35,248,940.00	34,510,067.00	-738,873.00	-2.14%

Pentaho Dashboards can be deployed out-of-the-box for immediate use by casual business users, or Java developers can exploit components for custom solutions using Java Objects or Java Server Pages (JSPs). These solutions can be tightly integrated with other applications or with JSR 168-compliant portal solutions.

Data Mining



Classifier: UserClassifier

Test options

○ Use training set

○ Supplied test set

○ Cross-validation

● Percentage split

(Nom) class

Result list (right-click for

11:49:05 – trees.j48.

14:34:28 – functions

14:48:05 – bayes.Nai

15:26:57 – trees.Use

X: petallength (Num)

Y: petalwidth (Num)

Colour: class (Nom)

Polyline

Plot: iris

Class colour: Iris-setosa Iris-versicolor Iris-virginica

Status: Building model on training data...

Weka data mining completes the Pentaho platform, offering sophisticated pattern recognition and predictive analytics found in few competing BI offerings, whether closed or open source. Weka collects clustering, segmentation, decision tree, random forest, neural network, and principal component analysis algorithms, all integrated with Pentaho for delivery either free-standing or via Pentaho reports or dashboards linked with Pentaho workflow.

BI Differentiation

Pentaho Business Intelligence is unique in the BI market. It is the only comprehensive BI platform, offering reporting, OLAP, dashboards, and data mining, that is built from the start to support workflow at every level and across all functions. It is also highly scalable and interoperable, providing enterprise-class features such as an object repository, report bursting, and reporting services for JasperReports, BIRT, and Pentaho's own JFreeReports. It supports out-of-the-box BI and may also be deployed as a component of larger business solutions. And Pentaho is open source with the same cost and flexibility advantages offered by MySQL, JBoss, Apache, and Linux. It is the only open-source BI platform with such widely ranging capabilities, scalability, and flexible deployment, plus a unifying workflow engine.

Download and Go

Pentaho Business Intelligence is available as a free download. You have access to source code, pre-built installations, packaged demos, and a comprehensive set of development and deployment documentation. You can choose to download the whole suite, individual components, or the SDKs you need to create your own applications.

The easiest way to get started – to add business intelligence capabilities to your MySQL database platform – is to visit *pentaho.org* and download the Pre-Configured Installation. The download comes with a pre-installed and configured JBoss Application Server. In addition, it contains a ready-to-use solution repository including sets of reports demonstrating functionality, which will help you get started. In ten minutes, you'll have the software running and accessible via a Web browser. You can then modify the pre-configured installation to tap your own data sources and provide your own data objects as the basis for your own reports, cubes, dashboards, and other BI components.

About Pentaho

Pentaho is the leader in open-source business intelligence. The company manages, facilitates, supports, and leads development of the Pentaho BI Project. The core project team has been together for many years, through success after success. It includes highly experienced industry leaders with a strong record of creating successful BI products for top-tier commercial vendors, including Business Objects, Cognos, Hyperion, IBM, Oracle, and SAS.

Working with others in the Open Source community, Pentaho aims to achieve positive, disruptive change in the BI space by building a leading-class BI platform and making it available to everyone by releasing it into open source.

About the Author

Seth Grimes is a business intelligence, data warehousing, and decision systems expert. He has been working with relational databases for almost twenty years, with MySQL for six years, and with open-source data-analysis tools for about as long.

Seth founded Washington, D.C.-based Alta Plana Corporation in 1997. He consults, writes, and speaks on information-systems strategy, data management and analysis systems, industry trends, and emerging analytical technologies. Seth is a contributing editor and writes the Breakthrough Analysis column for *Intelligent Enterprise* magazine.