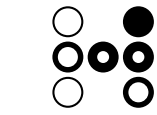
Large-Scale Data Warehousing Using Hyperion Essbase OLAP Technology

January 2000



Contents

Overview

Data warehousing has traditionally focused on relational technology. While well-suited to managing transactions and storing large amounts of data, relational databases are typically unable to handle ad hoc, speed-of-thought analytical querying for large user communities. Online analytical processing (OLAP) technology, however, provides the scalability, performance and analytic capabilities necessary to support sophisticated, calculation-intensive queries for large user populations. For these reasons, relational and OLAP technologies are often combined for maximum benefits.

This paper focuses on the terabyte-scale test performed at IBM's Teraplex Center in Rochester, MN, a facility to research and demonstrate large-scale data warehouse implementation. The test examined the role of OLAP technology in a large-scale data warehousing architecture. In summary, the Hyperion and IBM solution delivered the fastest results of any comparable demonstration, even with large numbers of simulated users accessing the system at the same time. When the number of simulated users increased from one user to 100 users, the average query time per user decreased significantly. Average query time per user was 0.0027 seconds for 100 simulated users. This ability to handle large numbers of simulated users in an ad hoc environment proves that the Hyperion-IBM solution is suitable for even the most sophisticated, large-scale, Web-enabled data warehouses.

"However, as enabling for end users as these new Relational Database Management System (RDBMS) products and associated tools and interfaces have been, there are still significant limitations to their efficacy …commercial Database Management System (DBMS) products do have boundaries with respect to providing functions to support user views of data. Most notably lacking has been the ability to consolidate, view and analyze data according to multiple dimensions, in ways that make sense to one or more specific enterprise analysts at any given point in time. This requirement is called 'multidimensional data analysis.'"

Dr. E.F. Codd
"Providing OLAP to End Users, an IT Mandate"

Data Warehousing and OLAP Technology

Two-Tiered Data Warehousing

The two-tiered approach to data warehousing consists of a data warehouse, where multiple sources of data have been extracted, transformed and cleansed, and one or more data marts, where subject-specific data is deployed to business users. To understand why companies have adopted the two-tiered model, it is useful to examine the historic development of data warehousing.

"The need which exists is NOT for yet another database technology, but rather for robust OLAP enterprise data analysis tools which complement the enterprise's existing data management system and which are rigorous enough to anticipate and facilitate the types of sophisticated business data analysis inherent in OLAP."

Dr. E.F. Codd
"Providing OLAP to End Users, an IT Mandate"

Many early efforts at data warehousing failed because companies focused more on the supply of data and all its sources than the demand for data and end-user requirements. Corporations built large warehouses integrating multiple sources of data, but did not provide the application-specific data models, interactive performance or rapid deployment that many end users required.

As end users became increasingly dissatisfied, many project managers and consultants shifted their focus to subject-specific data marts. Data marts typically offered faster performance and targeted data models/schemes, but data marts were often "point-to-point" solutions, meaning that different data marts were built from different systems using different business rules. For example, gross margin might be defined one way in a financial data mart and quite differently in a sales data mart. These marts also had "custom piping" linking data from operational sources to specific data marts, which led to data inconsistencies and increased maintenance costs as centralized IT departments struggled to maintain multiple systems.

To provide the best solution for business user and IT communities, many companies have turned to a two-tiered architecture, often called hub-and-spoke. In this model, IT technology departments construct a centralized data warehouse, which serves as a repository for extracted, transformed and cleansed data, integrated from multiple sources. Data marts scoped to specific business needs are deployed to leverage the information stored in the data warehouse. This solution enables centralized IT organizations to build, maintain and deploy data warehouses efficiently, and also to meet business users' requirements for analysis and rapid response times. Metadata integration is critical in ensuring that data definitions are consistent across data marts so that the data and metadata stay synchronized. As elements are added or reorganized in the data warehouse, data in the data marts is automatically updated due to tight metadata integration.

In this two-tiered model, the role of the data warehouse is to provide a central, cleansed repository for large amounts of heterogeneous data, and relational databases are deployed as the underlying database technology.

In contrast, the role of the data marts is to deliver focused business user applications such as balanced scorecards, e-Business analytics and performance management. To enable these analytic applications, data mart technology should include the following:

- The ability to scale to large volumes of data and large numbers of concurrent users
- Consistent, fast query response times that allow for iterative speed-of-thought analysis
- Integrated metadata that seamlessly links the data mart and the data warehouse relational database to ensure that data and metadata stay synchronized
- The ability to automatically drill from summary and calculated data to detail data stored in the data warehouse relational database, without exposing end users to the complexities of SQL querying. These queries should be dynamic and driven by end-user context, so that end users analyzing specific information can drill through to reports that automatically support only the relevant detail.
- A calculation engine that includes robust mathematical functions for computing derived data
- Seamless integration of historical, projected and derived data
- A multi-user read/write environment to support what-if analysis, modeling and planning requirements
- The ability to be deployed quickly and maintained cost-effectively with minimal user training
- Robust data-access security and user management
- Availability of a wide variety of viewing and analysis tools to support different user communities, application types and information delivery architectures (client-server, spreadsheet, Java, HTML, ActiveX)

Online Analytic Processing (OLAP)

Because OLAP technology provides user and data scalability, performance, read/write capabilities and calculation functionality, it meets all the requirements of a data mart. Two other options—personal productivity tools, and data query and reporting tools—cannot provide the same level of support. Personal productivity tools such as spreadsheets and statistical packages reside on individual PCs, and therefore support only small amounts of data to a single user. Data query and reporting tools are SQL-driven, and frequently used for list-oriented, basic drill-down analysis and report generation. These tools do not offer the predictable performance or robust calculations of OLAP. The OLAP technology option supports collaboration throughout the business management cycle of reporting, analysis, what-if modeling and planning.

Most important in OLAP technology are its sophisticated analytic capabilities, including:

Aggregations, which simply add numbers based upon levels defined by the application. For example, the application may call for adding up sales by week, month, quarter and year.

Matrix calculations, which are similar to calculations executed within a standard spreadsheet. For example, variances and ratios are matrix calculations.

Cross-dimensional calculations, which are similar to the calculations executed when spreadsheets are linked and formulas combine cells from different sheets. A percent product share calculation is a good example of this, as it requires the summation of a total and the calculation of percentage contribution to total sales of a given product.

Procedural calculations, in which specific calculation rules are defined and executed in a specific order. For example, allocating advertising expense as a percent of revenue contribution per product is a procedural calculation, requiring procedural logic to properly model and execute sophisticated business rules that accurately reflect the business.

OLAP-aware calculations, which provide the analytical intelligence necessary for multi-dimensional analysis, such as the understanding of hierarchy relationships within dimensions. These calculations include time intelligence and financial intelligence. For example, an OLAP-aware calculation would calculate inventory balances in which Q1 ending inventory is understood not to be the sum of January, February and March inventories.

OLAP technology may be either relational or multidimensional in nature. Relational OLAP technologies, while suitable for large, detail-level sets of data, have inherent weaknesses in a decision-support environment. Response time for decision-support queries in a relational framework can vary from minutes to hours. Calculations are limited to aggregations and simple matrix processing. Changes to metadata structures—for example, the organization of sales territories—usually require manual administrator intervention and re-creation of all summary tables. Typically, these relational solutions are read-only due to security and performance concerns, and therefore cannot support forward-looking modeling, planning or forecasting applications. In addition, resolving simple OLAP queries, such as: "Show me the top ten and bottom ten products based on sales growth by region, and show the sales of each as a percentage of the total for its brand," can require hundreds of SQL statements and huge amounts of system resources. For these reasons, many sites that initially deploy these technologies to support ad hoc reporting and analysis are forced to disable access and limit the number of concurrent queries.

For analytic and decision-support applications, implementation and maintenance are often more cumbersome in a relational environment. There are very few tools to define, build or manage relational schemes, forcing developers and consultants to manually design and continually optimize databases, leading to long implementation times. Furthermore, a large IT support staff is required to implement, maintain and update the environment, increasing the overall cost and limiting the IT organization's capacity to address other strategic information systems projects. Yet another concern is security, as a Relational Database Management Systems (RDBMS) provides table/column security only and cannot easily control access to individual facts in a star schema. The result is that it is often difficult or impossible to provide robust user data access security in an analytic relational database other than at the report level.

Multidimensional technology is free from the limitations that relational databases face in decision-support environments, as multidimensional OLAP delivers sub-second response times while supporting hundreds and thousands of concurrent users. In addition, it supports the full range of calculations, from aggregations to procedural calculations. Companies using Hyperion Essbase are able to rapidly deploy data marts and adapt to changing business environments. Since Hyperion Essbase is a server-centric technology, companies can share information readily and securely, with protection down to the most granular levels. Multiple users can update the database and see the impact of those updates, which is essential in planning and forecasting applications.

Data Warehouse Integration

Data warehouse integration with data marts speeds deployment, eases maintenance and enables users to navigate seamlessly between the mart and the warehouse. Therefore, tight integration is key to delivering a rapid and sustained return on investment (ROI). With data warehouse integration, infrequently used information can be stored and leveraged as needed in its appropriate context. Hyperion Integration Server™ provides a reusable metadata layer that aids in the rapid deployment of OLAP applications from relational sources, and provides the metadata linkage to support drill-through to warehouse data. In some cases, when individual power users have urgent but temporary analytical needs, temporary "disposable" data marts can be created. For example, users may need to quickly analyze a problem with a specific supplier, or understand the initial impact of a new product rollout. With Hyperion Integration Server, they can create temporary data marts on the fly that solve these problems, and then discard them when they are no longer needed. Because temporary data marts can be created from existing metadata mappings and business definitions, there is no need for specific project planning or IT intervention and almost no associated development cost.

Project Description

In tests at the IBM Teraplex Integration Center, Hyperion and IBM simulated a real-world environment with high concurrency and a generalized data model. By testing various user concurrency levels and response times for a variety of queries, they were able to project the response times and configurations of a very large database environment.

To create the environment used in the test, they generated one terabyte of data using an industry-standard generator utility and populated an IBM DB2 Universal Database which resided on a 32-node Netfinity cluster. The generator they used created random data in a defined schema based on the target size of the database.

Table	Row Count
Region	5
Nation	25
Customer	150,000,000
Orders	1,500,000,000
Lineitem	6,000,000,000
Partsupp	800,000,000
Part	200,000,000
Supplier	10,000,000

To best determine the most useful OLAP database structure, it is helpful to examine the ranges and max/min values for each field of interest based on the 1-terabyte size data warehouse.

Table	Field	Values
Customer	C_MKTSEGMENT	5 total, regardless of scaling
Orders	O_ORDERDATE	1/1/92 – 12/31/99
Orders	O_CLERK	1500 orders per clerk
Orders	O_ORDERSTATUS	F,O or P
Orders/Customers		Approx. 10 orders per customer
Lineitem	L_RETURNFLAG	A,N or R

Lineitem	L_LINESTATUS	O or F
Orders/Lineitem		Approx 4 per line item
Parts	P_BRAND	25-5 per P_MRGR
Parts	P_SIZE	50
Parts	P_TYPE	150
Parts	P_MGRG	5 total

Furthermore, the model needed to support four queries:

Query A: *Price Summary Report*
This query reports the amount of business that was billed, shipped and returned.

Query B: *Local Supplier Volume*
This query lists the revenue volume done through local suppliers, where the customer and supplier are from the same nation.

Query C: *Volume Shipping*
This query determines the value of goods shipped between certain nations to help in the renegotiation of shipping contracts.

Query D: *National Market Share*
This query determines how the market share of a given nation within a given region has changed over two years for a given type.

Building the OLAP Model

The Hyperion Essbase model was designed to answer the following questions:

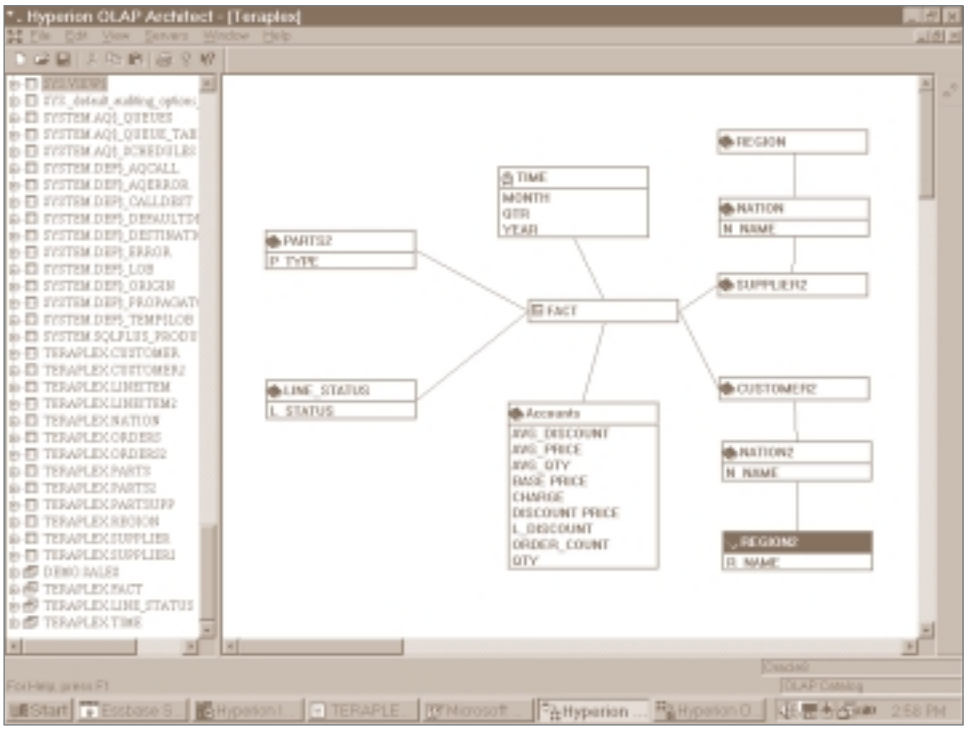
1. How much business was billed, shipped and returned?

2. How much revenue was done through local suppliers (where the supplier nation = the customer nation)?

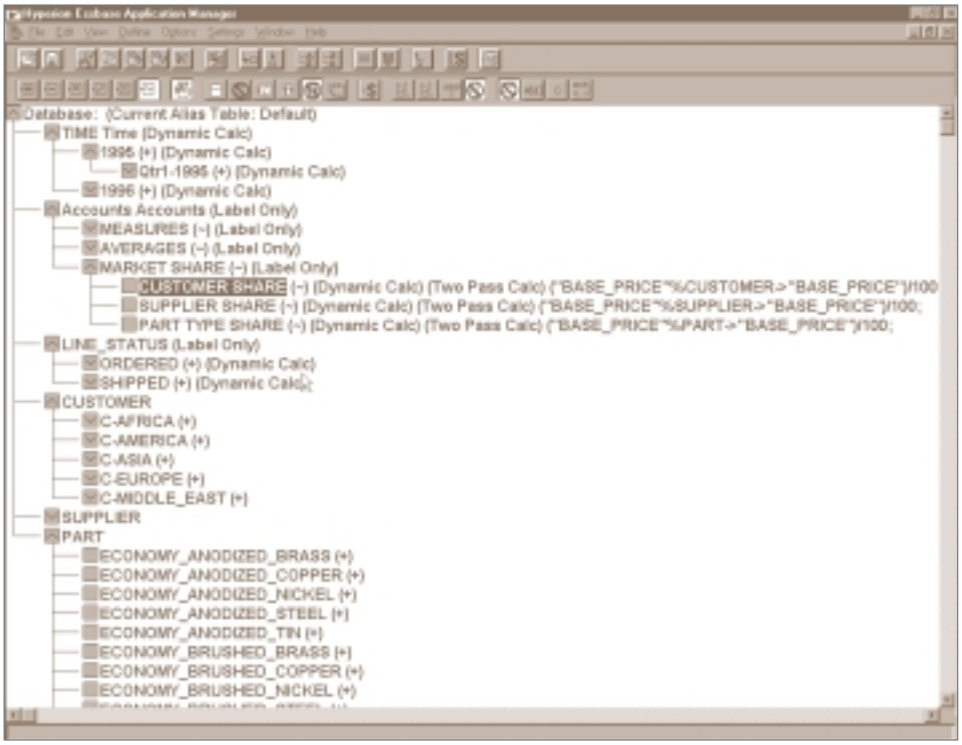
3. What were the gross discounted revenues for any two given nations (to assist in the renegotiation of contracts)?

4. How has the market share of a specific region changed over two years for a given part type?

Using Hyperion Integration Server, Hyperion experts mapped the relational data stored in the DB2 Universal Database to create six reusable dimension objects: time, accounts, line status, customer, supplier and parts. These objects were then assembled using the Hyperion Integration Server to create a multidimensional model for supplier analysis. In a real-world situation, business users or data mart designers would re-use and re-assemble these objects into additional applications such as billings, bookings and backlog analysis. Because the metadata is mapped back to the relational data warehouse, Hyperion was able to demonstrate context-specific drill-through capabilities.



The mapping used to create the Hyperion Essbase model.



The analytic application created in Hyperion Essbase using Hyperion Integration Server.

Server Configuration

Thirty-two clustered Netfinity 7000 M10 systems with four 450-megahertz (MHz) Pentium II Xeon processors each, 2 GB RAM and a 2-megabyte (MB) L2 cache were used to house the IBM DB2 Universal Database data warehouse and Hyperion Essbase OLAP server.

The Netfinity 7000 platform was selected because it is powerful, versatile and fast enough to handle the demands of large data warehouses. With four-way SMP support and sub-systems balanced to take advantage of Intel's fastest chips, the Netfinity 7000 M10 handles extreme demands, such as clustering, sever consolidation, e-Business intelligence and enterprise resource planning.

All database disk drives were controlled by HP NetRAID-3Si Disk Array Controller cards.

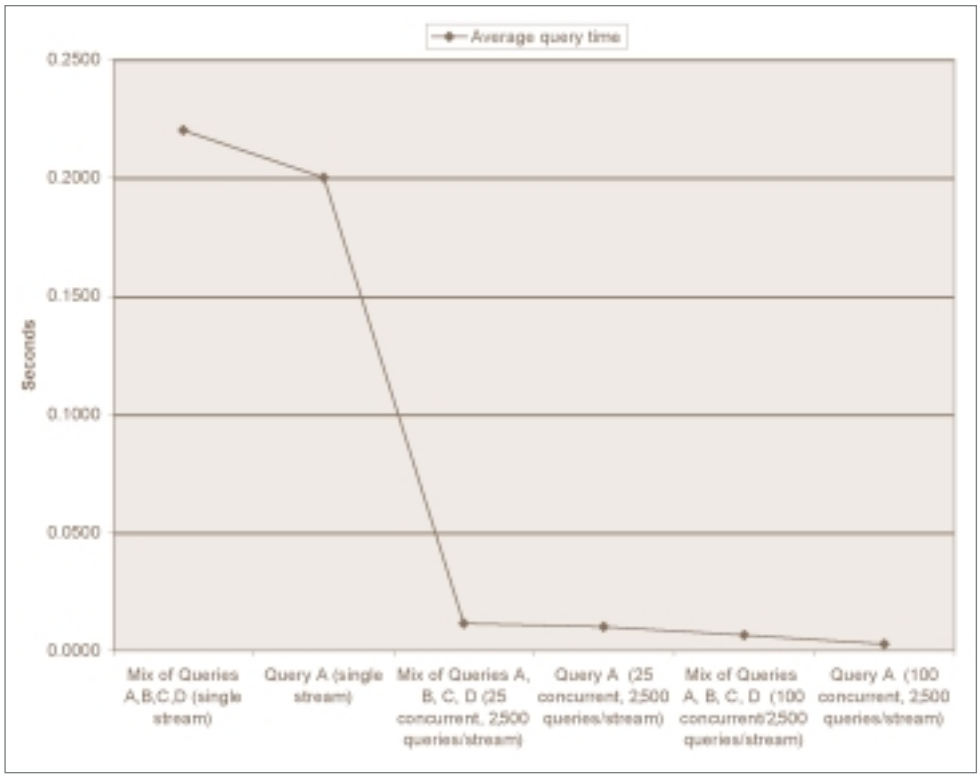
Software used for the test consisted of commercially available versions of the following:

- IBM DB2 Universal Database EEE V5.2
- Hyperion Integration Server 1.1
- Hyperion Essbase OLAP Server 5.0.2

Timing Test

Six groups of queries were run to demonstrate the effects of high user concurrency and to compare the results of a single query with a mix of four queries. For all query runs, Hyperion Essbase running on the IBM Netfinity Server delivered sub-second response times, even with 100 simulated concurrent users. Hyperion used a query execution program written in the C programming language, leveraging the Hyperion Essbase Application Programming Interface (API) to execute the queries. The timed tests were audited and declared valid by Jerry Lagorio of Lynx Consulting.

This lightning-fast response, coupled with the ability to support high levels of concurrent users, is crucial for large-scale Web deployments. Typically, in an RDBMS, as the number of concurrent users executing decision-support queries increases, the average query time per user increases as well. In contrast, the Hyperion Essbase server's average query time actually decreased dramatically as the number of concurrent users increased, as shown in the following chart. When query A with a single user is compared to query A with 100 simulated, concurrent users, the query time decreases from 0.20 seconds to 0.0027 seconds, a performance improvement of 9,865 percent. These results verify that Hyperion Essbase is ideally suited to support large enterprise deployments of analytic applications and Web-enabled data warehouses.



Hyperion Essbase OLAP Server's response times decreased as the number of concurrent users increased.

In addition to the query timing tests, Hyperion demonstrated context-specific, drill-through query capabilities. For 99 percent of end user analysis needs, the information required to answer the query is stored in the Hyperion Essbase server. In the remaining 1 percent of queries, end users can seamlessly access the relevant data stored in the data warehouse using built-in, drill-through functionality. In the Teraplex demonstration, Hyperion executed the following query, "How has the market share of a specific region changed over two years for a given part type?" From the result set, Hyperion then drilled-down automatically into the relevant data stored in the warehouse to see the details on the specific type of part, "economy anodized steel." This ability to drill-down to the context-specific detail is essential in delivering an integrated analytic solution.

Conclusion

This test demonstrated that Hyperion Essbase and DB2 on the Netfinity deliver a complete data warehousing solution, which supports sub-second response times for thousands of users.

In addition, the Teraplex proof of concept showed:

- Hyperion Essbase supports a 1-terabyte data warehousing environment.
- Hyperion Essbase OLAP Server delivers sub-second response times in high concurrency environments, making it ideal for large Web deployments.
- Large data warehouse systems, using a two-tiered model, can deliver rapid performance and high scalability with superior analytic capabilities, increasing end user adoption and dramatically enhancing ROI.
- Hyperion Essbase, when used with IBM DB2 Universal Database, provides this complementary solution.

Appendix I: Test Results

Query	Query Streams	Total Time Elapsed	Total Number of Queries	Queries per Second
Mix of Queries A,B,C,D (single stream)	1	22	100	4.55
Query A (single stream)	1	4	20	5.00
Mix of Queries A,B,C,D (25 concurrent, 2,500 queries/stream)	25	719	62,500	86.93
Query A (25 concurrent, 2,500 queries/stream)	25	634	62,500	98.58
Mix of Queries A,B,C,D (100 concurrent/ 2,500 queries/stream)	100	1,627.31	250,000	153.63
Query A (100 concurrent/2,500 queries/stream)	100	675	250,000	370.37

Appendix II: Auditor's Statement

Steps Taken by Independent Auditor Jerry Lagorio, Principal at Lynx Consulting:

a) Reviewed the hardware used in the test and verified that all components used in the test are commercially available from IBM.

b) Reviewed the database to ensure that a terabyte or more of storage was occupied by the base data.

c) Reviewed the steps taken to summarize this data from the terabyte to the Hyperion Essbase model.

d) Verified that the process used to replicate the data to a summary star schema and then to the Hyperion Essbase summary cube could be replicated in a standard production environment, guaranteeing consistent query results.

e) Verified that the timings for this exercise were accurate and run in accordance with standard business practice.

f) Verified that the dimensionality of the production cube and structures of the base table conformed to those outlined in the TPC-H document.

g) Verified that the software used to simulate standard query access to the data represented a fair implementation of a SQL standard, even though it was used to access a nonstandard relational data store.

h) Certified that this query generator actually provided somewhat slower access to the stored data, based on the overhead needed for standard query translation, than the standard Excel API.

i) Reviewed the batch processes used to simulate multiple user access to verify that this code did nothing to affect query performance, and provided a valid simulation of multiple users accessing the cube at the same time.

j) Timed each execution of the query sets, validating independently that the times recorded for their execution are true and correct.

k) Audited the results of these queries to ensure totally random execution of the query statements for the multiple query execution.

l) Verified that the query result did, in fact, tie back to the detail data stored in the terabyte data store.

m) Verified that in multiple query tests, the number and type of queries executed match the claims made in this report.

Conclusion

This test is a valid demonstration of the performance of Hyperion and IBM products handling sample queries, conforming to industry specifications. This audit confirms that the test was valid, and that the timings quoted in this document are accurate. In summary, Hyperion and IBM delivered sub-second response to 100 simulated no-think concurrent users in a 1-terabyte database environment with no response time degradation due to high levels of concurrency.

Hyperion Headquarters
Hyperion Solutions Corporation
1344 Crossman Avenue
Sunnyvale, CA 94089

tel: 408 744 9500
fax: 408 744 0400

info@hyperion.com
www.hyperion.com

European Headquarters
Hyperion Solutions Europe
Enterprise House
Greencourts Business Park
333 Styal Road
Manchester M22 5HY
United Kingdom

tel: 44 161 498 2200
fax: 44 161 498 2210

Asia-Pacific Headquarters
Hyperion Solutions Asia Pte. Ltd.
#24-01 IBM Towers
80 Anson Road
Singapore 079907

tel: 65 323 3485
fax: 65 323 3486

1470_01299TA_300KIN