Paper# : 3011

Decision Support and Data Warehousing in an HP 3000 Environment.

Christophe Jacquet
Hewlett-Packard Co.
19447 Pruneridge Ave
Cupertino, CA, 94087
Email: cjacquet@cup.hp.com

As corporate downsizing continues, flatter organizational structures give end users more decision making responsibilities. HP 3000 users can take advantage of powerful tools such as Image/SQL and ODBC drivers to gain direct access to the data in the databases. While these tools provide real benefits, the limitations of directly accessing operational data, in terms of performance, concurrency, ease of use and data quality, are rapidly becoming apparent.

A current approach to Decision Support is the consolidation of company-wide data into a separately designed environment, commonly called a Data Warehouse.

This paper covers the benefits and history of Data Warehouses and examines a variety of approaches available within the HP 3000 marketplace to implement such a Decision Support environment. It describes the elements needed to create a Data Warehouse environment from the Decision Support data store to the end user desktop, including extraction and cleansing tools.

Data Warehousing is just one example of how enhancements to the MPE/iX operating environment permit HP 3000 customers to take full advantage of the many new technologies and approaches emerging in the Information Technology marketplace.

Why is decision support important.

In today’s world, companies have to be able to make decisions much faster than before. Making the right decisions at the right time can provide significant competitive advantage for a company. Whether the marketing organization is struggling with pricing decisions, the finance department is contemplating a major change in investment strategy, or an Asian subsidiary needs to decide on opening a new store in a new area, it is imperative to have access to all relevant information.

What is a Data Warehouse?

A Data Warehouse can be viewed as a Decision Support database that is maintained separately from an organization's operational databases. A more rigorous definition is that a Data Warehouse is a subject-oriented, integrated, time variant, non volatile collection of data that is used primarily to aid organizational decision making. Data warehouses sometimes are referred to as "informational databases" as opposed to operational databases.

As we just saw, a Data Warehouse has four fundamental characteristics:

Subject Oriented

Operational data, such as order processing and manufacturing databases, are organized around business activities or functional areas. They are typically optimized to serve a single, static, application. The functional separation of applications causes companies to store identical information in multiple locations. The duplicated information’s format and currency are usually inconsistent. For example, in a delivery database, the customer list will have very detailed information on customer addresses and typically indexed by customer number concatenated with a zip code. The same customer list in the invoicing system will contain a potentially different billing address and be indexed by an accounting “Customer Account Number". In both instances the customer name is the same, but is identified and stored differently. Deriving any correlation between data extracted from those two data bases presents a challenge. In contrast, a Data Warehouse is organized around subjects. Subject orientation presents the data in a format that is consistent and much clearer for end users to understand. For example subjects could be “Product”, “Customers”, “Orders” as opposed to “Purchasing”, “Payroll” etc.

Integrated
Integration of data within a warehouse is accomplished by dictating consistency in format, naming, etc. Operational databases, for historic reasons, often have major inconsistencies in data representation. For example, a set of operational databases may represent "male" and "female" by "m" and "f," by "1" and "2," by "x" and "y". Frequently the inconsistencies are more complex and subtle. By definition, in a Data Warehouse data is always maintained in a consistent fashion.

Time variant
Data warehouses are time variant in the sense that they maintain both historical and (nearly) current data. Operational databases, in contrast, contain only the most current, up-to-date data values. Furthermore, they generally maintain this information for no more than a year (and often much less). By comparison, Data Warehouses contain data that is generally loaded from the operational databases daily, weekly, or monthly and then typically maintained for a period of 3 to 5 years. This aspect marks a major difference between the two types of environments. Historical information is of high importance to decision makers, who often want to understand trends and relationships between data. For example, the product manager for a soft drink maker may want to see the relationship between coupon promotions and sales. This type of information is typically impossible to determine with an operational database.

Nonvolatile
Nonvolatility, the final primary aspect of Data Warehouses, means that after the informational data is loaded into the warehouse changes, inserts, or deletes are rarely performed. The loaded data is transformed data that originated in the operational databases. The Data Warehouse is subsequently reloaded or, more likely, appended on a periodic basis (usually nightly, weekly, or monthly) with new, transformed or summarized data from the operational databases. Apart from this loading process, the information contained in the Data Warehouse generally remains static. The property of nonvolatility permits a Data Warehouse to be heavily optimized for query processing.

The benefits of implementing a data warehouse

Turning Data into Information

Data Warehouses are important to nearly all large businesses since they resolve several problems experienced by both IT and end users. Most organizations today have no shortage of data. The challenge, however, lies in transforming the data into useable INFORMATION. Most of the data that exists today is in a form that is difficult for human access or interpretation. The Data Warehouse process when properly designed and constructed, extracts information from the operational databases and puts it in a form that can be understood and interpreted by the people who need to make decisions. In the past, an organization could maintain its competitive advantage by using information technology to automate repetitive tasks and control complex processes. Now, however, these traditional uses are no longer sufficient to keep a company competitive. It is essential to leverage the information value of data to provide true Decision Support capabilities. Data Warehouses exist to do just that: to maximize the decision-making power that is hidden in every organization’s data.

Bringing Solutions to Users.

Historically, data was exclusively managed by a centralized IS department that reacted to user requests by providing processing and reporting services. Today those users are demanding faster access to the information. With the pervasive deployment of desktop systems, the role of the IS is changing. End users are now active participants in the use and manipulation of data. In many organizations today, end users create and manage their own databases. This is happening because of a need for faster data access, and the driving need to analyze data, not just read reports.

Making Strategic company-wide decisions.

Data Warehouses collect and maintain all of the required information and offer it in a format that can be easily used and manipulated. This greatly improves efficiency and improves the overall value of information extracted offering a way for IT to add value and leverage the use of information across different parts of the company.

What are the solutions on the HP 3000?

There are actually three different approaches to Decision Support and Data Warehousing on the HP 3000.

Decision support using an existing operational database

As discussed earlier, users who require data located in one or two HP 3000 resident databases and are seeking an inexpensive solution can continue to use their operational databases. Image/SQL, Allbase/SQL and Oracle on the HP 3000 allow the use of appropriate Client/Server tools to give information access to their users. This approach has the great advantage of being inexpensive (no new database, no duplication of the data into a separate data store). However, there are some issues with the performance impact of the Decision Support queries on OLTP activity.
A hybrid solution consists of enhancing existing databases by adding further datasets or tables dedicated to Decision Support in the operational database. Operational data can be separated from historical data and organized by subject in those tables to greatly reduce locking conflicts with OLTP applications. This provides the opportunity for the extract program to cleanse and summarize it. This type of approach is an appropriate low-end Decision Support solution, but should not be considered for an enterprise wide Decision Support system since it is often restricted to one database being use for one activity of the company.

Virtual Data Warehouse

The Virtual Data Warehouse is an alternative for companies wanting to expand the previous concept over several distributed databases in the company. The Virtual Data Warehouse depends on a piece of software, known as middleware, that links the end user tools to the physical database. The user transparently and simultaneously accesses multiple databases on multiple systems, functioning as a single coherent Data Warehouse. Often the middleware includes a data dictionary that allows access to intelligible filed name (i.e. “Employee Number” instead of “Emp_NUM”or “E#”or “Field_2562”). Middleware can also simplify the data presented to the user. For instance, if a field ‘customer name’ appears in several tables, the middleware will only show one unique field. Middleware will group the field by subjects. For instance, the user will see one table that contains customer information, whereas the table may be the vertical (or horizontal) aggregation of fields pulled from the marketing database, the invoicing database, the credit database, etc. The information is accessed in the original location without duplication, and the user has the perception of accessing a single central database.
As we have discussed earlier, a potential drawback is that the operational databases accessed are not typically optimized for Decision Support. The lack of standardization in the databases is also an obstacle for Virtual Data Warehouses. For example the social security number in the invoicing database can be an ASCII field with the format XXX-XX-XX, and the same social security number in the shipping database can be a numeric Field with the format 99999999. An improperly planned join of those two fields might cause a problem.

The HP 3000 can take advantage of Image/SQL, Allbase/SQL, or Oracle databases. Middleware products such as EDA/SQL from IBI can unify access to many additional databases potentially resident on other platforms, for example, Informix running on a remote HP 9000.

Discrete Data Warehouse

Given the definitive description of a Data Warehouse, the Discrete Data Warehouse is probably the only approach that truly warrants the title of a Data Warehouse. It is composed of a discrete database, dedicated to Decision Support activity and populated with data consistent with true Data Warehouse definition criteria (subject oriented, integrated, time variant, nonvolatile). Typically, building such a warehouse requires involving the users in identifying all required data, designing the data model, creating the extraction and cleansing routines, periodically populating the database and subsequently maintaining it.

The following tasks are critical to the success of a Data Warehouse project:

· Spending time with the end user to define data requirements is important to the success of the project. Users should be cautioned about selecting too many data items to populate the warehouse as the database may become too large and unmanageable. Additionally, loading time, backup and recovery time, as well as overall solution cost, will increase proportionally to the number of elements selected. The choice of data loaded into the database must be thoroughly discussed and accepted by the users as well as the architect of the solution.

· The design of a warehouse database demands a radical departure from the traditional database design approach. Data warehouses are often de-normalized and redundancy is added to improve response time. The overall design will also depend on the most frequently accessed data elements and the most commonly intended dimensional query type (time, geography, customer, etc.). Information to be aggregated or summarized will have to be identified. For instance, if the operational database contains all the invoices for each customer, the Data Warehouse might contain a summarization of invoices per customer per month or quarter. Another example would be to aggregate in a single record all the transactions of a particular company by region, state or county. All these choices must be made early in the implementation process. At this stage the data elements relating to the data contained in the warehouse must be defined; these elements, that will also assist the user in understanding the warehouse's structure and content, are known as Metadata.

· The implementation of the warehouse itself is usually done in a relational database. Allbase/SQL and Oracle have the appropriate characteristics to be HP 3000 based databases. These include scalability, performance, a degree of parallelization, high availability features and compliance to open standards. The HP 9000 provides additional RDBMS options (Informix, Software AG, Sybase, Red Brick etc.) In both cases appropriate extraction tools are required to populate the database. Examples of commercially available extraction tools appear later in this paper under "Choosing tools". Frequency of database population will have to be defined, based upon data volatility and usage.

· Finally, it is important to determine appropriate warehouse maintenance procedures. Given the dynamic nature today's business environment, change is inevitable. This change will place new and unforeseen demands on the content of the Decision Support System. Flexibility and adaptability are critical to the continued use and success of the database and therefore must be embedded into the design at the very earliest stages.

How Hewlett-Packard Co can help you?

Hewlett-Packard's OpenWarehouse Program has proven to be a very strong player in the Data Warehouse market. OpenWarehouse is HP's framework for delivering Data Warehouse solutions to customers based on best-in-class HP and third-party components. OpenWarehouse is made up of a basic technical framework, the user's choice of best-in-class components, and the consulting/integration services necessary to help an organization design and deploy the Data Warehouse solution.

HP's OpenWarehouse differentiates itself from other superficially similar offerings in several key areas: expertise, partnerships and Intelligent Warehouse:

· Breadth and Depth of Expertise - HP helps ensure solution success through its breadth and depth of expertise. This expertise is demonstrated by the following:

· According to industry analyst MetaGroup, HP’s PSO has the largest group of trained and experienced Data Warehouse consultants in the industry.

· HP has over nine years of internal experience in Data Warehousing and has been recognized at national conferences as a leader in enterprisewide Data Warehousing. This experience has been one of the primary sources of HP’s consulting methodology.

· Managed Partnerships - HP can help ensure solution success through its managed partnerships with nearly all major Data Warehouse components providers. These partnerships are built on both technical and executive level relationships that we take advantage of to help ensure that the full solution is successful.

· Intelligent Warehouse - HP can help companies in making Data Warehouses easy-to-manage, easy-to-use, and enterprise scaleable through its Intelligent Warehouse product, described later in this document. Intelligent Warehouse is OpenWarehouse’s key differentiator.

Choosing tools

RDBMS for Data Warehouses

A wide choice of RDBMS products are available on the market. The HP 3000 has many options with Allbase/SQL and Oracle while the HP 9000 enjoys a more varied selection. Despite the high performance and reliability of Image/SQL for OLTP applications, the lack of flexibility and of a Data Definition Language may not make it a strong candidate for a large scale Data Warehouse project. However, for limited projects where the structure of the database is fairly stable Image/SQL is a definite possibility especially in the context of a Virtual Data Warehouse.

On the HP 9000, the choice is much larger. HP works closely with Informix, Oracle, Red Brick, and SYBASE to ensure that each of their RDBMS’s is tuned for high performance load/indexing, query processing and parallelization. Another important consideration when choosing an RDBMS for the HP 9000 in an HP 3000 environment is the availability of gateways between the operational databases on the HP 3000 and the Data Warehouse. Allbase/SQL, with Allbase/Net, for example provides an excellent linkage between the two platforms. Last year Oracle released the Transparent Gateway to the Image/SQL product that enables easy interoperability and data communication between the two databases on either the HP 3000 or HP 9000. SYBASE is working with Proactive System on the port of Open Client/Open Server, and OmniSQL to the HP 3000, this should be available first half of 96. With other databases, communication with Image/SQL can be done through EDA/SQL.

Middleware

Intelligent Warehouse (IW) is the Data Warehouse management component of OpenWarehouse. IW is a software product that is made up of open middleware together with administrative and end-user tools that make Data Warehouses easy to manage and enterprise scaleable. As middleware, Intelligent Warehouse works in a three-tier architecture with most relational DBMS (i.e. Allbase/SQL, Informix, Ingres, Oracle, Red Brick, SYBASE or MVS/DB2) and with most ODBC compliant data access tools (this includes a large number of tools, including: Business Objects, Cognos Impromptu, HP Information Access etc...) IW helps database administrators to manage the performance and the security for diverse users and groups. With IW the system administrator has all the information necessary to make changes that will improve query performance in some cases from hours to minutes or even seconds.

Finally, IW optionally allows Data Warehouses to be enterprise scaleable. Intelligent Warehouse can support single logical Data Warehouses spanning multiple systems that can easily handle multiple terabytes of data. It can also tie together multiple Data Warehouses into a single enterprisewide federation of Data Warehouses.

Integration of the Information Builder Inc.'s (IBI) product into a Data Warehouse solution was referenced earlier in the document. With EDA/SQL's support of more than 50 databases and files, and more than 30 hardware platforms, IBI appears in the market as very strong player. Integration with HP 3000 databases and files makes it a definite winner in an HP 3000 environment. The HP 3000 can actually run the EDA/SQL Hub Server code that provides a single business directory of all data elements available to the user community. This allows a total data location transparency that simply gives the user access to information without requiring user knowledge of where the physical data resides or the network protocols required for access. The Hub server manages the cross platform joins and returns information to the user as if they were in a single database. This is a perfect tool for multi-platform or geographically distributed environments.

Data Extraction tools

The process of Data Warehouse management involves extracting data from operational databases, transforming (or cleaning) the data, moving the data to the server on which the warehouse is located, and loading the data into the warehouse. The Data Warehouse management process also typically should include the production of relevant warehouse Metadata. This activity can be handled by COBOL or 4GL programs written specially by a customer or integrator to handle a specific custom warehouse. In many cases, it is more appropriate to use special Data Warehouse management software for the following two reasons:

· Developing custom code to perform the extract, transformation, and load processes is typically a long, tedious process that greatly lengthens the time necessary to develop and deploy the Data Warehouse. Initially this process may seem easy, but the realities of most corporations' operational databases normally make the activity long and difficult. Adopting this approach can place disproportionately large demands on scarce IT resources and cause frustrations in management, who are typically looking for fast results.

· Custom coding does not produce organized Metadata ("data about data") that describes the Data Warehouse. Data Warehouse Metadata serves a number of functions, such as

· showing the mapping between the original operational data and the warehouse data,
· showing the relationship of fields to the underlying data model,
· showing the frequency of data element usage,
· business term description of the data,
· guidelines for drilling down through the data.

The function typically provided by the transformation tools are filtering, mapping, conversion, summarization, delta data selection, sequence restructuring, other algorithmic changes.

After the developer has specified extraction criteria, transformations, and target DBMS, the Data Warehouse extraction tool generates COBOL or C code, associated JOB, and relevant scripts. The warehouse manager uploads the extraction program (and associated JOB) to the operational system(s) and uploads the transformation program, compile/link script, program run script, and CREATE/LOAD script to the server on which the Data Warehouse will be located. On the operational system, the extraction programs are run periodically, extracting the relevant data into a large sequential file (or set of files) that is ready to be transported via tape or network connection to the server on which the Data Warehouse resides.

Network transport is generally done through higher speed file transfer mechanisms like FTP rather than the typically slower database gateways. This is an important point since extraction, transformation, transport, loading and indexing normally must all take place in a relatively narrow time window (generally a night or a weekend). For large volumes, database gateways often do not have the necessary bandwidth to transport data. For smaller volumes, database gateways can be an inexpensive and effective solution to the problem. Beside the network transfer users need to be made aware that there are other steps that need to take place within the available time window.

Following data transfer, the transformation program is then run, necessary transformations performed, and the CREATE/LOAD scripts run for the target database management system (i.e., the Data Warehouse).
The choice of extraction/transformation tools that support Image/SQL or ALLBASE/SQL on the HP 3000 as a data source are: Extract from Evolutionary Technologies Inc. (ETI), Data Movement and Warehouse products from Taurus Software Inc, and ReTarGet from Rankin Technology Group. Taurus software has the great advantage of having the ability to look in both source and destination databases. This feature is very useful for conditional moves.

Desktop Query and reporting tools

Data access and reporting tools are desktop products for graphically building SQL queries that operate against the Data Warehouse. In addition to allowing queries to be made, most of these tools have full report writing capabilities that allow the returned data to be easily formatted.
A multitude of data access and reporting tools exist on the market that are applicable to an HP 3000 environment. These tools support a wide variety of desktop systems including Microsoft Windows, Windows NT. The most common are Cognos Impromptu, HP Information Access, Esperant from Software AG and Focus for Windows from IBI. For HP 3000 data environments some data extraction tools like Data Express from M.B. Foster & Associates can also be considered.

Desktop executive information systems
An executive information system (EIS) is simply a very easy-to-use data access tool, and obviously not restricted to executives. Common characteristics in today's EIS products include drill-down, threshold and exception reporting. The term EIS is now losing favor in many parts of the world.
In contrast, much market excitement is now being directed towards On-Line Analytical Processing (OLAP) server and client tools. In some cases these are re-dubbed EIS tools and in other cases totally new. OLAP servers and clients provide an easy-to-use multi-dimensional interface for easily drilling down from high-level to low-level summary data and for data interrogation.
OLAP tools can provide extremely fast response time on small databases, handle complex calculations, and enable what-if analysis for budgeting and forecasting. Examples of popular EIS/OLAP systems include Cognos PowerPlay, Platinum Forest & Trees, SAS, and Speedware Media.
A further important consideration when choosing end-user tools is to verify that the tools can be used against the existing operational data. For example, verify that they can directly access Image/SQL databases and are ODBC compliant. This avoids multiplication of tools necessary to manage the desktop.

Other tools
As you will start using the data warehousing environment you will have to tune the database to make it more performant. A lot of tools are available on the market to improve query performances. Intelligent Warehouse from HP has some of those functionalities. Omnidex from DISC can greatly improve the performance of long queries by adding advance indexing technics to your databases.

Summary

We have discussed several aspects of Decision Support and Data Warehousing in an HP 3000 environment. Many of the tools exist, and the role of the IT manager consists of putting all these tools together. The HP 3000 can continue to be your mission critical platform. At the same time you can integrate new technologies like Data Warehousing to enhance your current investment. Whether you choose to do it on MPE or on UNIX becomes a secondary question. As we have discussed in this paper the right bridges are available to make it happen on either platform. As the technology moves forward, the HP 3000 continues to play an important role in your computing environment.