



COMPUTER
CORPORATION
OF AMERICA

THE VIRTUAL
DATA WAREHOUSE:
A PHASE ONE
BUSINESS
INTELLIGENCE
SOLUTION

For more information on using CCA Analytics to maximize your business intelligence efforts, please contact Computer Corporation of America at 508-270-6666, or visit us on the Web at www.cca-int.com.

© 2001 by Computer Corporation of America. All rights reserved.

Model 204 is a registered trademark of Computer Corporation of America. All other trademarks and tradenames are used to identify entities claiming the marks and names of their products.

Introduction

I hope no one disputes the business value of a well-built data warehouse. Just imagine all of that wealth of operational data - from any number of disparate systems throughout the enterprise - coming together to form an information source that quickly and easily allows any business manager to understand how the business is performing. Imagine too the non-intuitive relationships between data and systems that will just leap off the screen or page, and enable you to create new business initiatives for competitive advantage. It's a beautiful thing.

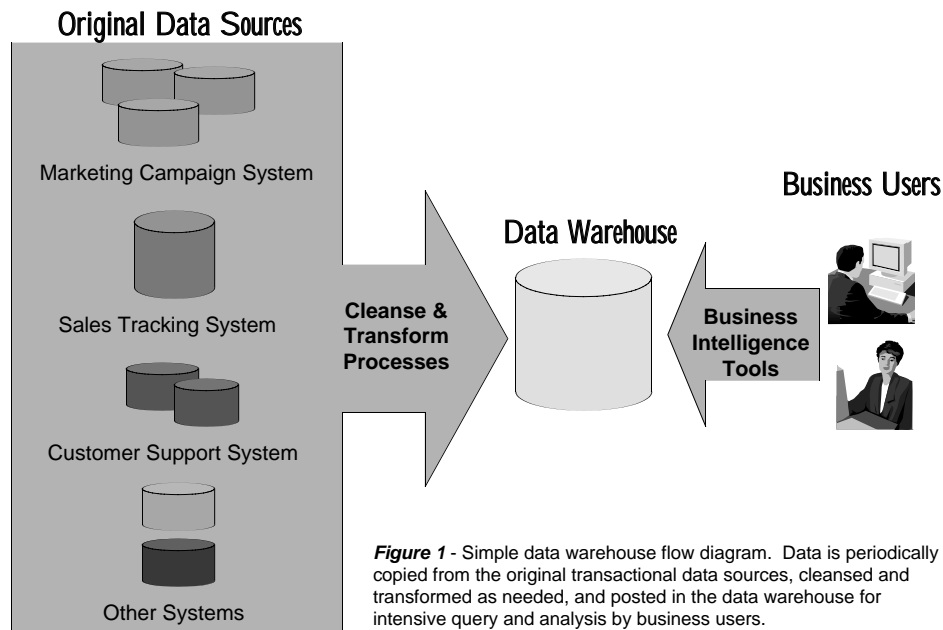


Figure 1 - Simple data warehouse flow diagram. Data is periodically copied from the original transactional data sources, cleansed and transformed as needed, and posted in the data warehouse for intensive query and analysis by business users.

It's simple isn't it? All of your data is cleansed and transformed. Customer addresses are consistent, accurate and de-duplicated. All of those Deutschmarks Francs and Schillings from those European companies you took over in the 90's are accurately expressed in US dollars. Data from that system that has been running your shipping operation in North Africa on something called a 'Prime' since 1972 is right there at your fingertips. Want to know the lifetime value of each customer by State, by Age and by Gender? It's there in a heartbeat.

And do you have to go cap-in-hand to the Information Technology department for this valuable business information? No. They've already done their job. They've built the data warehouse and now they're either soaking up the sun in Barbados or (more likely) they are continuing the sisyphian task of applying new technologies to your operational systems. They might be integrating systems from newly acquired companies. They might be modifying old systems or developing new ones in response to competitive pressure, consumer demand, or new government legislation. Perhaps they're implementing a brand new system as a result of a brilliant initiative resulting from your clever use of your data warehouse. Dead easy isn't it?

But we all know enough about data warehousing to realize that the implementation isn't nearly as simple and straightforward as the concept. First, we must analyze our business and understand the entities of which our enterprise consists. Of itself, this is a 'relatively straightforward' task. Whether you're a million-dollar or a billion-dollar organization, soup is soup and nuts are nuts. But however simple the logical model of our business may be, the physical representation of the model as expressed in our operational systems (computerized or otherwise!) can be much more complex. The analysis of existing operational systems can take many person-months, if not years. For a well-established enterprise, some operational systems might be 20 to 30 years old. In many cases the systems are undocumented or even worse, the documentation is inaccurate. Where is the data held? How current is it? How many systems source it? What forms does it come in? How clean is it? Do the operational systems validate it at source? Do they even care? All of these questions must be asked - and answered - anew. Hmmm! Perhaps it's not so easy after all.

If we find ourselves in the happy position of understanding the logical model underpinning our business and we have built an accurate map of our data sources, then we can start building our 'data warehouse'. We must design it (having, of course, determined our chosen flavor of 'industry standard' architecture), we must build systems to populate it with nice, clean data (yes, that means writing programs ... lots of them), and we must design, build and document mechanistic systems to keep our warehouse up-to-date - including all of those brand new aggregates that never existed in the operational systems because they didn't need them. (And yes, that means writing lots more programs that aren't necessarily the same as the ones we used to build the warehouse in the first place.) But it will be a beautiful thing, won't it?

Let us assume that we have gone through all of the pain, angst and expense of building our data warehouse. How are we going to use it? Any data warehouse/business intelligence vendor will help you answer that question ... for a few dollars more.

Imagine if you will that you have successfully achieved all of the above. You have modeled your business. You have analyzed your operational systems. You have built your warehouse. You have built your operational update mechanisms. You have chosen your end-user tools. You have educated your end-users in how to use them. (Phew! That took some time, didn't it?)

Congratulations! You are now the proud owner of the most complete and powerful competitive tool in the information age. Of course, during the many person-years (probably several elapsed-years) it has taken you to get to this point, your operational systems have remained stagnant and your end-user business units have been happily playing Scrabble and looking up the word 'sisyphian' (from the Greek legend of Sisyphus - not to be confused with that of Tantalus, from which we derive the word 'tantalize').

Alternatively, so tantalized at the prospect of the benefits they can glean from the data warehouse, your end-user departments might have independently paid consultants to build them some 'data marts'. Anyone remember the nightmare of 'distributed departmental systems'?

In the real world of course the chances are that you have out-sourced your data warehousing project to a respected consulting organization while your own in-house IT department has continued to maintain your operational systems. In the meantime you eagerly await the final, your ultimate Business Intelligence solution. By the way, what does Business Intelligence mean?

Many vendors and consultants adopt a Humpty Dumpty approach to data warehousing and business intelligence terminology: "When I use a word," Humpty Dumpty said, in a rather scornful tone, "it means just what I choose it to mean - neither more nor less." Or as the Gartner Group more conventionally puts it; "Data Warehouse, Data Mining, Business Intelligence (BI), Decision Support and Executive Information Systems are all terms that are misused to describe everything from the transformation and cleansing of data to complex analytical processing."

Usefully, the Gartner Group gives some independent definitions that I will use here.

"Data Mining refers to a process rather than a technology, with the goal of that process being to explore a large amount of data to discover new trends, relationships and categories in that data."

"Business Intelligence describes the enterprise's ability to access and explore information, analyzing that information and developing insight and understanding, which leads to improved and informed decision making."

"Data Warehousing is the means for creating and managing a data architecture for user access and analysis, besides data and information delivery. This is the critical foundation for both Business Intelligence and Data Mining."

The rewards to be gleaned from running data mining and business intelligence tools against a data warehouse can be enormous. One classic application is marketing. Successful users of data warehousing in this arena will tell you that better targeting can result in cost savings that alone can justify the investment in a few short years. That does not even take into account any incremental sales. Other high-return applications of data warehousing include enhanced fraud detection and predictive stock control.

Beyond the potential direct benefits that can be gleaned from data warehousing, don't forget that the extensive analysis that went into understanding the data before you put it into the warehouse will give your IT department a solid foundation for building and enhancing operational applications more rapidly than ever before. You will have created a sound platform not only for the development of new business system initiatives, but also for acting on those initiatives. You will have taken a significant step toward becoming a state-of-the-art, 'agile enterprise'.

Build it and the dollars will come ... if you are one of the lucky organizations who get it right that is. If you get it wrong, however, you will have wasted an awful lot of money. We've all heard the horror stories of multi-million dollar failures.

Wary of the risks involved, many organizations are not prepared to embark on a full-blown data warehousing project and will settle for the compromise of departmental data marts or of feeding their users small, transportable metacubes. Either of these approaches will doubtless deliver some benefit to the business, but they both have significant drawbacks. If you settle for departmental data marts, you will lose the opportunity of discovering non-intuitive significant relationships that cross departmental boundaries. (Perhaps shoe size is important in cat food marketing!) Deploying localized desktop solutions has the further disadvantage that a user can come to significant conclusions that might be valid ... if only the metacube they were working on wasn't significantly out of date.

So how do you go about minimizing the risks involved in building a data warehouse while at the same time delivering real benefit to the business in the shortest possible time? Let us now consider the concept we are calling the "virtual data warehouse" - a working prototype that will gradually evolve into the Real McCoy.

The Board of Cooperative Education Services (BOCES) in Syracuse, NY recently embarked on such a project. In two weeks, they had infrequent computer users - including school principals and administrators - accessing the virtual data warehouse and analyzing data on their own. As a bonus, the use of the virtual data warehouse is helping the IT department architect the "true" data warehouse because users are getting a better sense of their informational needs. These requirements are being passed on to the IT department, which is helping them design the permanent solution correctly the first time.

Defining the Virtual Data Warehouse

Dataquest defines data warehousing as "the establishment and maintenance of an IT architecture that provides data for end-user access, decision support, and business analysis. Typically, it uses operational data in raw or summarized form but stored and accessed separately."

Unlike a true data warehouse, the virtual data warehouse does not initially store data separate from the original data sources. It involves the application of appropriate business intelligence tools against operational data, providing the user with some amount of decision support capabilities quickly and easily. Clearly this is neither a profound nor original idea. It is the apparent limitations of this approach (performance, redundancies, unclean data, etc.) that spawned the data warehouse concept in the first place. Regardless, for some applications, and with the right tools, the virtual data warehouse can be an excellent interim solution. And because users are satisfying most of their own informational requests, the IT department suddenly has more bandwidth to devote to the design and implementation of the final data warehouse.

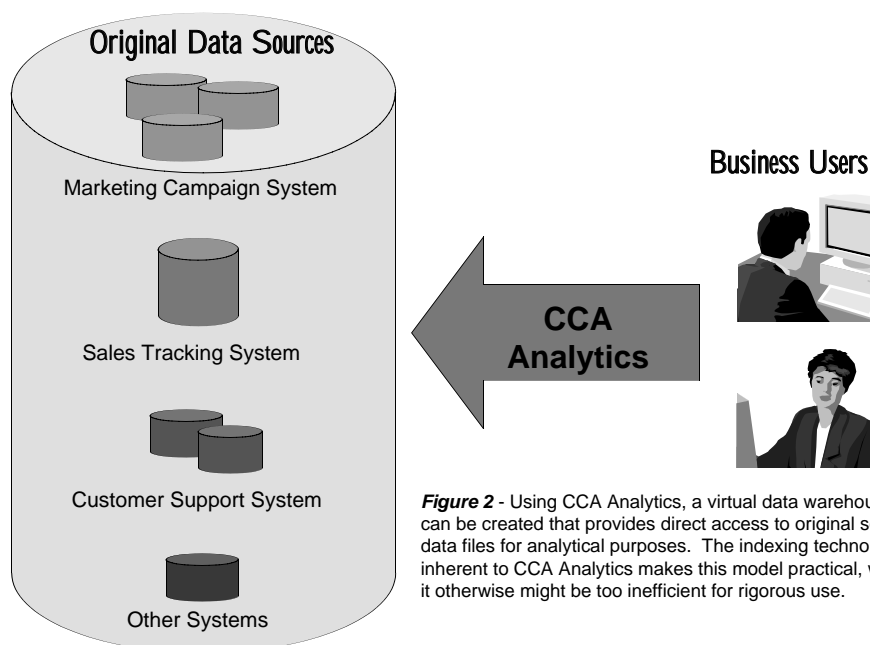


Figure 2 - Using CCA Analytics, a virtual data warehouse can be created that provides direct access to original source data files for analytical purposes. The indexing technology inherent to CCA Analytics makes this model practical, where it otherwise might be too inefficient for rigorous use.

The Challenges at BOCES

After making the decision to build a data warehouse for their internal clients, the IT department at BOCES had a real challenge before them. They knew building the data warehouse was absolutely the right solution for the future. After all, they were responsible for providing student-oriented research data to approximately 2,500 school principals, administrators and researchers in nearly every school district in the state. Every request was different, involving vast information about any number of different students, and requiring that results be displayed in diverse formats. Recently, despite their best efforts to keep up with the ever-growing demand for information, they found they no longer had the ability to completely satisfy their customers' thirst for knowledge. As a result, they were losing customers. Hence, they resolved to build a data warehouse.

A data warehouse would not only empower the users to obtain their own answers, but would significantly reduce the turnaround time and encourage them to explore the data at depths they never before considered or thought possible. And of course, the over-burdened IT department would have more time to devote to new critical initiatives such as e-business applications.

But building a well-architected data warehouse would not be an easy task. BOCES has 10 years of research data amounting to 30 million records of 3,000 fields, utilizing 161 different record types. All of this is stored in 58 Model 204 databases on the organization's IBM S390 MVS enterprise server. It would take some time just to figure out what data the users need to access in the data warehouse, let alone integrate the other components such as data cleansing, transformation, and analytical tools. The programming staff estimated it would take between 8 and 18 months to fully construct the data warehouse. But most troubling to them was this Catch-22: they still had to continue to serve their large user base, which left very little time for building the data warehouse that they hoped would eventually reduce their workload. Realizing that this created a no-win situation, BOCES began considering an interim, virtual data warehouse that would hopefully free up IT resources to devote to the true data warehouse project.

Choosing a Business Intelligence Tool

The first task was to choose a business intelligence tool that was up to the challenge of turning BOCES' operational data into a virtual data warehouse. After a brief review of popular tools, BOCES chose a freshly introduced product from Computer Corporation of America (CCA) called CCA Analytics. "CCA Analytics was attractive to us for several reasons", explains Larry Dismore, director of the Central New York Regional Information Center at BOCES. "First, like several other tools on the market, it provided all of the query, reporting, and analytical functions that we felt our users would need. Second, it was from our database vendor, so we knew there would be no integration problems and that the learning curve for producing custom functionality would be minimal. But third, and probably most important, the performance was pretty much unbeatable. We were not naive about the risks of throwing large numbers of researchers at operational data, and we were very concerned about the performance of both the analytical and the operational applications. The indexing technology inherent to CCA Analytics made this a non-issue."

CCA Analytics relies on the advanced bit-map indexing originally invented by CCA in the late 1960's. Today, after decades of development CCA's bit-map technology is capable of allowing rapid, in-core analysis of large databases with unparalleled performance. With CCA Analytics every data item in every database can be indexed. It also has the unique concept of 'Invisible Data'. Using 'Invisible Data', new discriminators can be added for analysis without storing data-proper in the physical records. This means two things of great significance for Analytics users:

- 1) Entities can be 'pre-joined' for enhanced performance. For example a 'CUSTOMER' file should only contain data-proper relating to attributes of the customer (date of birth, gender, etc.). Similarly a 'PURCHASE' entity should only contain data-proper relating to each purchase (date, department, value, etc.). By defining 'Invisible Data' fields for DATE, DEPARTMENT and VALUE as index-only attributes of the CUSTOMER file, questions such as 'how many female customers between the ages of 25 to 35 bought jewellery above the value of \$500 during the Christmas promotion period?' can be answered (unknown to the end-user) from a single physical B-tree-cum-bit-map structure without carrying out expensive joins. In this way, this type of query can be answered in a matter of seconds even against databases of tens of millions of customers with hundreds of millions of transactions.
- 2) CCA Analytics does not need to store data at all! In other words, to take advantage of the software's extreme performance profile for rapidly analyzing data, external sources of many gigabytes in size can be replicated solely as index structures using only a tiny fraction of the storage of the original sources.

Both of these factors are key to the secrets of how CCA Analytics can perform at great speed and so cost-effectively. From a technical perspective, just imagine that every logical I/O is accessing around 50,000 records at a time.

For complex queries and analyses, CCA Analytics combines index table information through the use of relational and logical operators, producing similarly fast results. In a sense, the index tables ARE the data warehouse, with the actual data records being used only when a full drill-down is required. This, of course, is completely transparent to the user, who doesn't need to worry about whether their operations are being completed at the index table or detail record level.

With the enabling technology chosen, the BOCES IT team was ready to begin implementing their virtual data warehouse.

Architecting the Virtual Data Warehouse

Like the building of a true data warehouse, constructing the virtual data warehouse required some understanding of the users' informational needs. But because the virtual data warehouse would consist mainly of index tables and logical views of the data, rather than physical copies of operational data, BOCES did not have to be too concerned with getting the views right the first time. If they built some views that users didn't really need, there was no significant consequence. And if they omitted a view initially which later surfaced as a real requirement, it could be built on the fly, without impacting the use of the rest of the warehouse.

BOCES' approach was to construct numerous views of all current databases, providing the users maximum flexibility for data analysis. This would help the users better understand what data was available to them, which would hopefully translate into more accurate user requirements for the true data warehouse. In just two weeks, the IT department completed the initial implementation of the virtual data warehouse, and delivered it to the users.

Accessing the Virtual Data Warehouse

The early results of using the data warehouse are very promising. Working in an intuitive, Windows-based environment, users have found it relatively easy to access and analyze the data, given the appropriate database view. Here are some tasks they can accomplish:

Query and Segment Data - Users can easily perform simple or highly complex queries in a point-and-click environment, without really knowing anything about the data. The values associated with each field are provided to the user upon request, helping the user find the exact data needed. Users can even search the data based on the result of a macro or formula, versus a specific value. Because users don't always know what they're looking for when they first begin to access the data, queries can be iteratively refined, allowing them to eventually narrow down the database in a step-by-step fashion. Once the ideal selection set of records and fields is found, the user can name and save the selection criteria for easy retrieval at a later date.

Summarize the Data - Users can quickly obtain summaries of the data such as counts, averages, and maximum and minimum values. Summaries can be grouped by any number of different fields, or even by computed values. As noted earlier, computations are extremely rapid, regardless of database size, because they are executed against the efficient index table structures, rather than against the actual database records.

Display the Data - The virtual data warehouse provides numerous tools for viewing selected data, including reports, charts, and graphs. Not only do these output formats display the data, but they are interactive, allowing the user to manipulate them to form new queries, drill down to underlying data, or view the data from different perspectives. If you require more specialized viewing of the data, the virtual data warehouse lets you easily export the data into other analytical packages such as spreadsheets or OLAP tools.

Success Factors

Not all environments can gain as much benefit so rapidly from a virtual data warehouse approach. BOCES' environment had several characteristics that contributed to their swift success, including:

Clean, Usable Data - BOCES' operational student data did not have serious integrity problems that prevented it from being effective for business intelligence purposes. Likewise, the data did not need to be substantially transformed. Aggregations that will likely be performed for the true data warehouse were not deemed essential due to the ease and speed of dynamic summaries through the index tables.

Appropriate Hardware Platform - Because BOCES' operational student databases are stored on their IBM S390 MVS enterprise server, it was reasonable to assume that the platform could handle the added user load. Had a smaller departmental server been in use, the number of users accessing the virtual data warehouse would most likely have to be limited, perhaps to the point of not meeting the original requirement.

Appropriate Database Platform - Likewise, the BOCES operational student databases are stored in Model 204, an enterprise database server known for its ability to support unusually high numbers of users. If a significantly less robust DBMS were in use, again, the number of users accessing the warehouse would most likely have to be limited.

Network Support - The infrastructure was already in place to connect the users' PCs to the mainframe, thus allowing them fast access to the new virtual data warehouse.

Homogeneous Data - Although the chosen business intelligence tool can easily support any data source, BOCES did not need to worry about database integration problems since all of their data was stored on the mainframe in a single DBMS. The more heterogeneous the database environment, the greater the likelihood that incompatibility issues will arise.

Pragmatic IT Staff - With a virtual data warehouse it is essential that the IT staff do not get bogged down in the paralysis of analysis. BOCES' IT staff had the benefit of being familiar with the technology underlying CCA Analytics and trusting it to be sufficiently flexible and resource efficient for the task in hand.

All of the above factors combined to allow BOCES to deliver end-user analytical capabilities against their full database in a remarkably short period of time. If you don't have all of these ducks lined up, implementing a virtual warehouse remains a possibility for your organization, but don't expect a two-week delivery time.

Moving Ahead

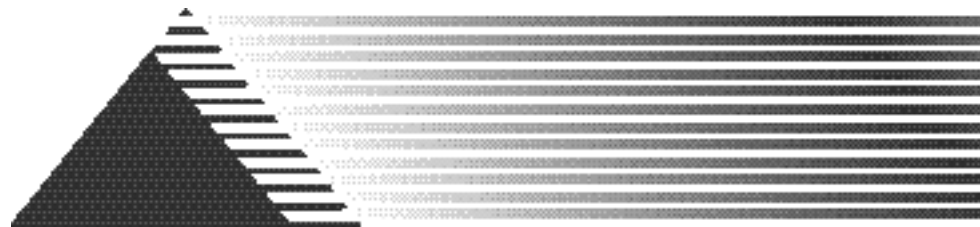
As noted earlier, the use of the virtual data warehouse is not only solving a short-term business intelligence need, but also contributing significantly toward the effective design of the true data warehouse. This side benefit will save both time and money for the IT department.

Furthermore, when the true data warehouse is up and running, the virtual data warehouse can continue to evolve into the second-generation virtual data warehouse and will remain a power weapon in your business intelligence armory. Even the most successful data warehouse implementations will reveal weaknesses and shortcomings over time. As an interim solution, these areas can continue to be addressed by the virtual data warehouse. Or taking a more optimistic viewpoint, the strengths and victories of the true data warehouse will undoubtedly generate more interest in its use by other lines of business that will require enhancements to the data warehouse to reflect their specific needs. As in the beginning, new requirements can be temporarily and immediately satisfied by the virtual data warehouse, which because of its association with operational data, continues to stay one step ahead.

This article has focused on a single product that can accelerate your data warehousing and business intelligence initiatives dramatically. But it is very important to remember that you can not buy a data warehouse 'out of the box'. No software vendor has 'the complete solution'. As the Gartner Group rightly observes; "A data warehouse is an architecture, and enterprises that focus on a single product for implementing a data warehouse increase the risk of failure."

All too often in my travels I come across organizations who tell me; "We already have our business solutions in place. We use Company Y, or Brand X." On probing a little further, I typically find that whatever benefits are accruing from the use of Company Y or Brand X, there are always some elements of disappointment. There is always some pain that an approach like virtual data warehousing might be able to salve. Remember, the object of the exercise is to discover something precious and create something valuable therefrom. Panning for gold, mining it, and the skills of the goldsmith all require separate, but ultimately related tools. Together they can make a beautiful thing.

The author, Chris Ramsdale, is the Director of Strategic Product Planning at Computer Corporation of America, the developers of Model 204 and CCA Analytics. He may be reached via e-mail at chris_ramsdale@cca-int.com.



**COMPUTER
CORPORATION
OF AMERICA**

500 Old Connecticut Path ♦ Framingham, MA 01701
508.270.6666 ♦ 508.270.6688 Fax ♦ www.cca-int.com