

恩
門

BILLINMON.COM

THE DATA WAREHOUSE ENVIRONMENT:
QUANTIFYING COST JUSTIFICATION
AND RETURN ON INVESTMENT

by

W. H. Inmon

November, 2000

INTRODUCTION

Intuitively the data warehouse environment makes a lot of sense when viewed from the larger perspective of gathering and accessing information. But a common question arises when the issue of the cost effectiveness of the data warehouse is raised. It is obvious that a data warehouse is not built for free. But are the benefits of the warehouse commensurate with the costs? Is, in fact, a data warehouse worth it?

Factored into this is the longer-term cost of ownership and operation of the data warehouse environment. Along these lines, Microsoft's SQL Server 2000 product offers very high value of ownership at a cost that makes it feasible to deploy data warehouses, data marts and OLAP cubes throughout an enterprise. It delivers a variety of data warehousing capabilities including ETL, meta-data repository, a powerful relational database, English Query and Analysis Services. Analysis Services offers both OLAP and data mining capabilities, making it the most comprehensive analytical offering in the market. Because of its lower price and ease of use, the initial and longer-term costs associated with SQL Server make it a very attractive platform upon which to build a comprehensive data warehousing and data mart environment.

VARIABLE COSTS

To begin with, the costs of a data warehouse are very variable. One organization builds a large data warehouse for a large amount of money. Another organization builds a large data warehouse for a significantly smaller amount of money. And ironically the organization that spends significantly less money for their data warehouse is usually much happier with their data warehouse than the organization that spent much more money.

Some of the factors that determine the cost and the satisfaction of a data warehouse are:

- does the organization understand that with the large volumes of data found in a data warehouse that the mass of the volumes of data should not be placed on high performance disk storage,
- does the organization understand that building a central architected data warehouse surrounded by data marts is the best long term approach,
- does the organization understand that the data warehouse must be built in small fast iterations of development rather than in a single large "big bang" approach,
- does the organization understand that for serious amounts of exploration that a separate structure called an "exploration warehouse" is required,
- does the organization understand that for true oltp response time that a separate structure called an ODS is required, and so forth.

If an organization does not understand these basic architectural issues then the costs of the warehouse will rise at an alarming rate along with the dissatisfaction with the warehouse. The first and most important key to managing the costs of the data warehouse is to understand the architecture of the warehouse and the environment that surrounds it. However, even if these architectural issues are understood, there is still a cost to the data warehouse.

COST JUSTIFICATION

There are many approaches to cost justification of the data warehouse. Several of those approaches will be addressed in this white paper.

The easiest and most straight forward way to address the cost justification of the data warehouse is to simply say that the data warehouse greatly reduces the cost of getting information into the hands of the user - across the board - for all of the organization. With a data warehouse, the cost of getting information shrinks dramatically. For every one. Every time there is a need to access information. Of any kind.

In other words, data warehousing reduces the cost of accessing information dramatically.

INFORMATION WHERE THERE IS NO DATA WAREHOUSE

In order to illustrate this dramatic and profound effect, consider the classical environment shown in Figure 1.

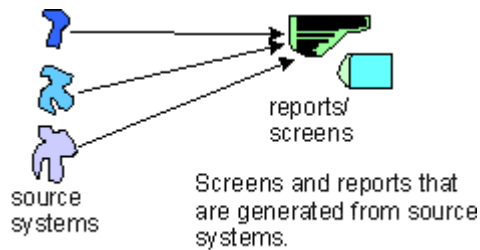


Figure 1

In Figure 1 there are some source systems, usually called the legacy environment. These source systems provide the data needed by the end user. Usually these legacy systems are transaction processing systems and reflect business requirements. There is a need for information from these legacy systems. That need is represented as the desire for a report or a screen. In order to fulfill that need, the legacy systems are accessed, information is gathered and that information is fed into the report or screen. This simple scenario is where most organizations are prior to the building of a data warehouse.

INFORMATION WHERE THERE IS A DATA WAREHOUSE

Now consider Figure 2.

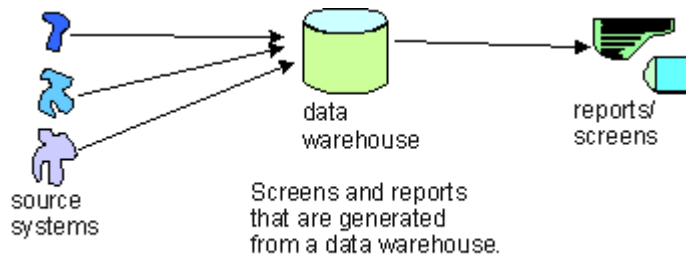


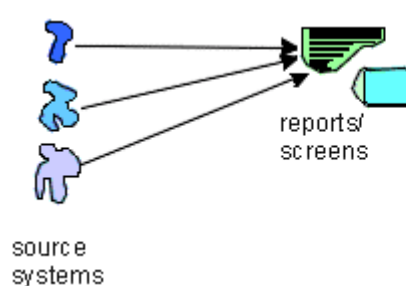
Figure 2

Figure 2 is the same as Figure 1 except that there is a data warehouse. The data warehouse sits between the legacy applications and the report. The legacy applications feed the data warehouse and the data warehouse in turn feeds the report or screen.

The only difference between Figure 1 and Figure 2 is the existence of the data warehouse. There is no data warehouse in Figure 1 and there is a data warehouse in Figure 2.

THE COST DIFFERENTIAL

In order to understand the cost differential between Figure 1 and Figure 2, consider the work that has to be done in order to produce a report. Figure 3 outlines the work that must be done to produce a report in the case of no data warehouse.



In order to produce the reports and screens it is necessary to:

- locate the data needed for the report,
- gather the data,
- convert/integrate the data,
- merge data from different sources,
- build the report.

Figure 3

Figure 3 shows that in order to feed the report from the legacy systems environment, the following activities must take place:

- the data needed for the report or screen is located in the legacy environment,
- once located, the data is gathered. This means accessing the data across a wide variety of technologies, such as IMS, IDMS, VSAM, ADABAS, Oracle, DB2, et. al.,
- once the data is gathered it must be converted or integrated. This means reconciling keys, internal encoding values, reference tables, data structures, operating systems, and so forth,
- once converted, if there are multiple sources of data, then data must be merged, and
- the report or screen is produced.

Depending on the state of the legacy environment, how many legacy applications must be produced, and the size and complexity of the report, the cost of producing the report from the legacy environment may be from:

- 2 or 3 months to 2 or 3 years, and
- \$100,000 to \$2 or \$3 million.

In fact the cost of getting information from the legacy environment is so high that many organizations simply give up in frustration.

GETTING DATA WHERE THERE IS A DATA WAREHOUSE

Now consider the costs and activities required to get the same report from the data warehouse environment. This scenario is shown in Figure 4.

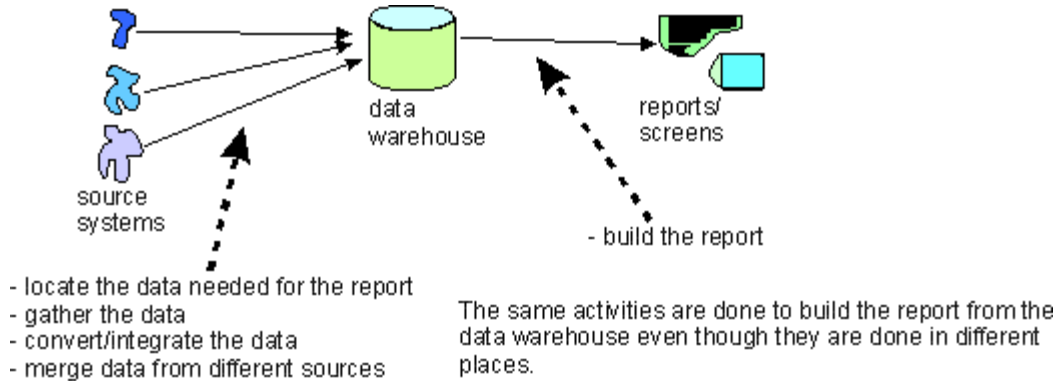


Figure 4

In Figure 4, there is a data warehouse. The report or screen is created from the data warehouse. But in order to create the data warehouse, it is necessary to:

- locate the data in the legacy environment that is needed for reporting,
- once located, gather the data in preparation for movement into the data warehouse. This means accessing the data across a wide variety of technologies, such as IMS, IDMS, VSAM, ADABAS, Oracle, DB2, et. al.,
- once the data is gathered it must be converted or integrated as it is moved into the data warehouse. This means reconciling keys, internal encoding values, reference tables, data structures, operating systems, and so forth,
- merging the data, once integrated, that comes from multiple sources.

In other words, the first four steps that are done in Figure 3 are also done in Figure 4. In Figure 3 the preparation steps are done to produce the report. In Figure 4 those same steps are done to produce the data warehouse. In fact the only real difference between Figure 3 and Figure 4 is that the report is issued from the data warehouse in Figure 4 instead of being directly issued from legacy systems as in Figure 3.

The costs of location, gathering, conversion and integration, and merging are the same in Figure 3 and Figure 4. So why is there any savings at all when the same activities must be done in Figure 3 and Figure 4?

MULTIPLE REPORTS

The answer is that if there were only one report to be created, then there is no savings to be made by creating a data warehouse. But the reality is that no corporation operates on a single report. Instead corporations operate on many reports and many screens.

Given the reality of the corporate need for many reports and screens, Figure 5 shows what corporations need.

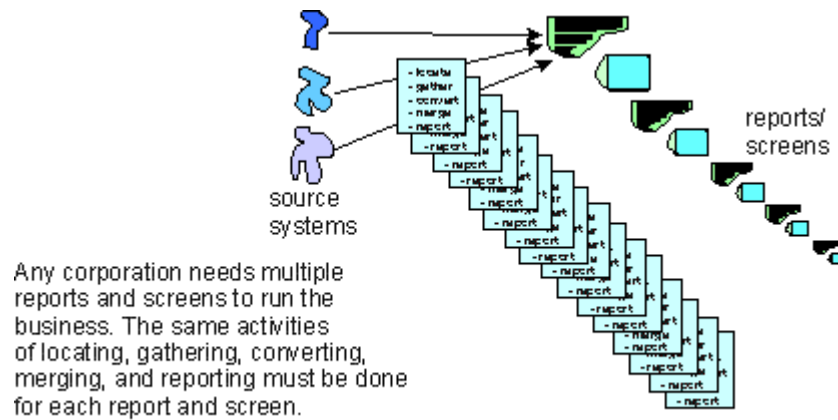


Figure 5

Figure 5 shows that corporations operate on many forms of information. The problem is that EACH report and EACH screen requires its own unique set of transformations. In other words, each report requires its own set of activities that include:

- locating the data
- gathering the data
- converting/integrating the data
- merging the data.

The cost of EACH report and EACH screen then is very high when the source of the data is the legacy environment.

Now consider the costs of reporting when there is a data warehouse. Figure 6 illustrates this set of circumstances.

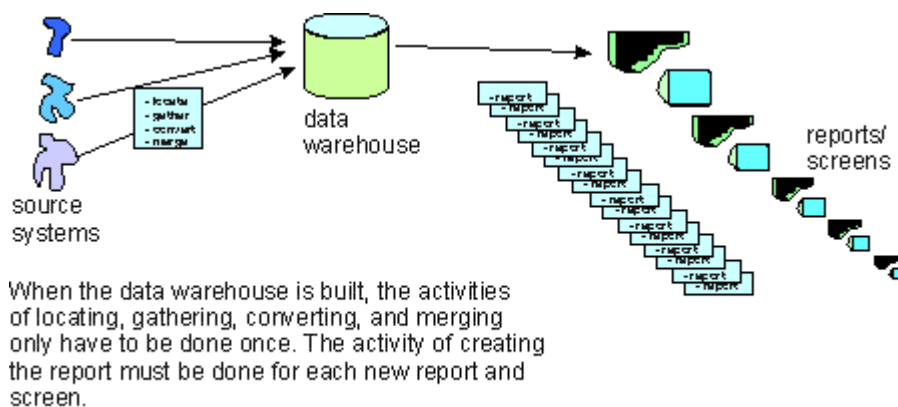


Figure 6

In Figure 6 data is gathered into a data warehouse. Once gathered into a data warehouse, the data is then available for reports. Many reports are written from the data warehouse.

There is a one-time cost of gathering and organizing the data into a data warehouse. But in order to produce the reports and screens the corporation needs, there is only the cost of reporting for each report. In other words, the gathering of data into the data warehouse is a one time only cost. When there is a data warehouse, data has to be located, gathered, integrated and merged only once. But when there is no data warehouse, the location, gathering, integration, and merging must be done for every report. And that is where the tremendous cost savings is made for a data warehouse.

Appendix A provides a simple spreadsheet for calculating your savings in the building of a data warehouse.

FARMERS AND EXPLORERS

There is another aspect of the cost justification of a data warehouse that needs to be explored. That aspect is: for whom in the corporation is the data warehouse justified in the first place?

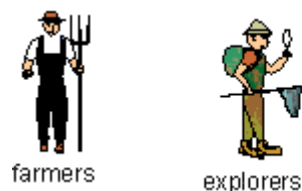
In order to answer that question, consider that the corporation has two very different kinds of users – farmers and explorers.

A farmer is someone who knows what they want before they set out to find it. A farmer is predictable. A farmer usually finds what he/she wants. A farmer submits small queries. A farmer seldom finds huge nuggets of information, but a farmer often finds small flakes of gold. If a farmer were a baseball player the farmer would hit a lot of singles, seldom strike out, but only infrequently hit home runs.

An explorer is very different from a farmer. An explorer is the original "out of the box" corporate thinker. An explorer is very unpredictable. An explorer may go six months submitting no queries at all, then one week submit ten queries. An explorer submits very large queries. Often times the explorer finds nothing for his/her efforts. But occasionally an explorer finds a priceless nugget of corporate wisdom that has been overlooked. If an explorer were a baseball player, the explorer would hit a lot of home runs. And correspondingly, the explorer would strike out often.

There are then very real differences between farmers and explorers even though farmers and explorers both use the data warehouse.

Figure 7 illustrates farmers and explorers.



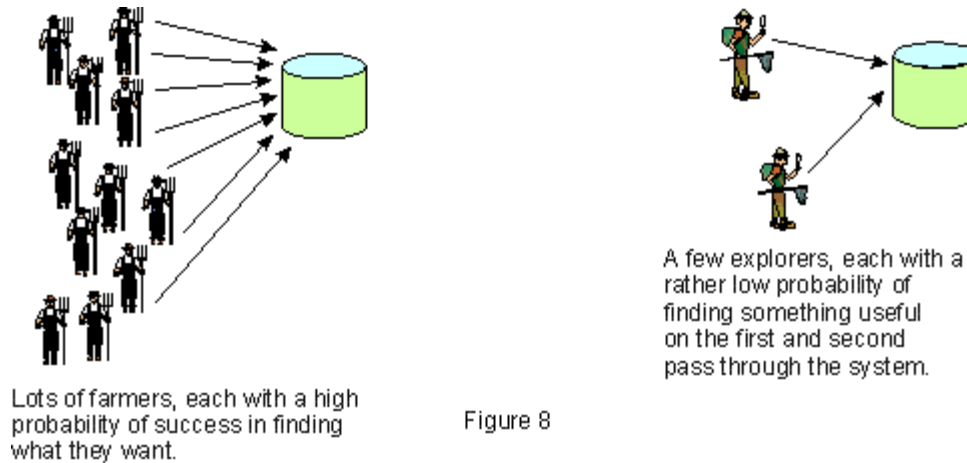
Two members of the DSS data warehouse community are farmers and explorers.

Figure 7

The issue of data warehouse cost justification is very germane to farmers and explorers. Doing cost justification for a data warehouse is almost always done on the basis of the results obtained by farmers, not explorers. Stated differently, doing a cost justification for a data warehouse based on the work of explorers is a very risky and unadvised thing to do.

FARMERS, EXPLORERS AND COST JUSTIFICATION

Figure 8 shows that cost justification for a data warehouse is best done for farmers not explorers.



Why is it that cost justification for a data warehouse is not done for explorers? The answer is that with explorers you never know what you are going to come up with. If by some chance you set out doing exploration and you achieve a spectacular result, then management feels very good about the sponsorship of the data warehouse. But if you set out to do an exploration activity and come up with nothing, then management will feel very bad about paying for the data warehouse. And truthfully, at the outset of an exploration activity, you never know what you are going to come up with. The chances are excellent that as an explorer, you will not hit a home run your first time at bat.

Management does not like to back perceived losers. Therefore, when it comes to cost justification for the data warehouse, it needs to be done on the basis of the expectations for farmers, not explorers. Of course, some organizations are very sophisticated and understand the risks. In this case, a data warehouse may be justified on the basis of the expectations of explorers. But the more normal case is such that it is very risky to try to base the results of explorers as a basis for the data warehouse.

DATA MARTS AND THE DATA WAREHOUSE

Another issue relevant to the cost justification of the data warehouse is that of whether the organization should build a data warehouse or a data mart. In order to understand the issues it is first necessary to understand what a data warehouse is, what a data mart is, and how they differ.

A data mart is a departmental structure. Typical departments having a data mart are the finance department, the sales department, the marketing department, the accounting department, and so forth. The data in the data mart is designed to be optimal for access for the different users of the data mart. The data warehouse is structurally different from a data mart in that the data warehouse must serve the needs of the entire corporation. The data warehouse is truly a corporate structure, serving many different needs.

Figure 9 shows the architectural positioning of the data marts and the data warehouse.

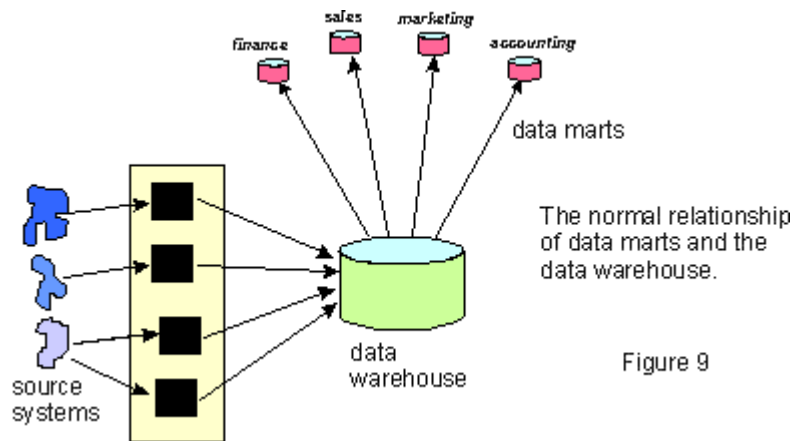


Figure 9

Figure 9 shows that the data marts are fed data from the data warehouse. In turn the data warehouse is fed data from the legacy applications.

There are many fundamental differences between a data mart and a data warehouse. Figure 10 depicts some of the major differences.

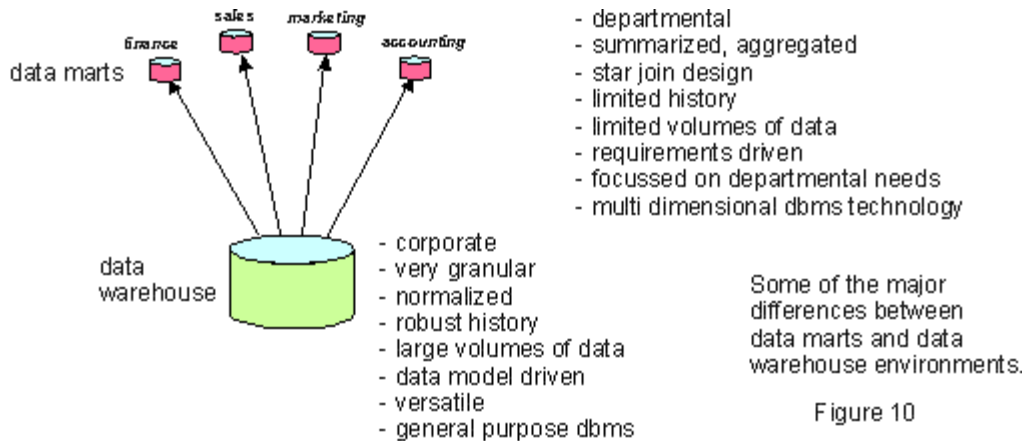


Figure 10

In Figure 10, it is seen that a data warehouse contains corporate data, granular data, and normalized data. The normalization of the data in the data warehouse allows the data in the warehouse to serve all purposes. The data in the data mart, on the other hand, is summarized and is very denormalized. The data mart data serves the needs of a department, not the corporation.

DATA MARTS AND DATA WAREHOUSES

Data in the data warehouse contains a deep, robust amount of history, typically from five to ten years worth of history. The data mart contains only limited amounts of history, from one month to a years worth of history. If there is any great length of history preserved in the data mart it is preserved at the summary level.

The data warehouse contains large volumes of data - up to several terabytes typically. The data mart contains substantially smaller amounts of data, say from five gbs to fifty gbs

The data warehouse is designed primarily from the data model. The corporate data model reflects the abstraction of the needs of the corporation for information. The data mart reflects specific requirement needs by a department. The different applications of the data mart user and the ways the data mart user needs to access data shape the design of the data mart. The data mart is requirement driven while the data warehouse is data driven.

The data warehouse is very versatile in its usage. The data warehouse serves first one user then another. The data warehouse allows data to be looked at one way, then in the next minute allows the data to be looked at another way. The data mart allows data to be viewed optimally in only one manner. While a data mart is optimal for access by one group of people - say finance, the data is non-optimal for access by anyone else - say accounting, sales, marketing, and so forth.

The data warehouse sits on top of general-purpose dbms technology. The primary purpose of the general-purpose dbms technology is to manage a lot of data. The data mart sits on top of technology that is optimal for end user access and analysis. Typically the data mart sits on top of multi dimensional technology or other technology useful for end user reporting and analysis. There are then a whole host of architectural differences between a data warehouse and a data mart

WHICH TO BUILD FIRST?

The question in many shops becomes one of which do I build first - a data warehouse or a data mart? In fact if I build a data mart first, do I even need to build a data warehouse?

In truth it is possible to build a data mart without building a data warehouse. And the data mart vendors greatly encourage that stance. Figure 11 shows that building a data mart without building a data warehouse is a real possibility.

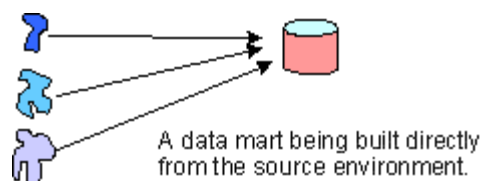


Figure 11

In Figure 11 there are some legacy systems feeding a data mart. Such a configuration is relatively easy to accomplish. And as long as there are only one or two data marts, such a proposition is acceptable. But the problem is that an organization NEVER builds one or two data marts. There are many parts of the organization that will need their own data mart:

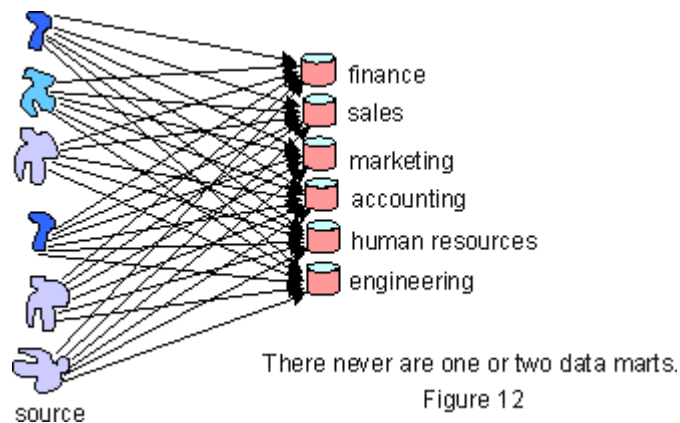
- Marketing
- Sales
- Finance
- Accounting
- Actuary
- Engineering

- Production Control
- Human Resources, and so forth.

It is never reasonable to assume that there will be a single data mart in an organization. Therefore, the simple diagram seen in Figure 11 is valid only for a short amount of time, while the corporation is building and operating the first data mart. Very quickly the diagram shown in Figure 11 changes.

MULTIPLE DATA MARTS

As the corporation grows, the diagram shown in Figure 11 turns into the diagram shown in Figure 12.



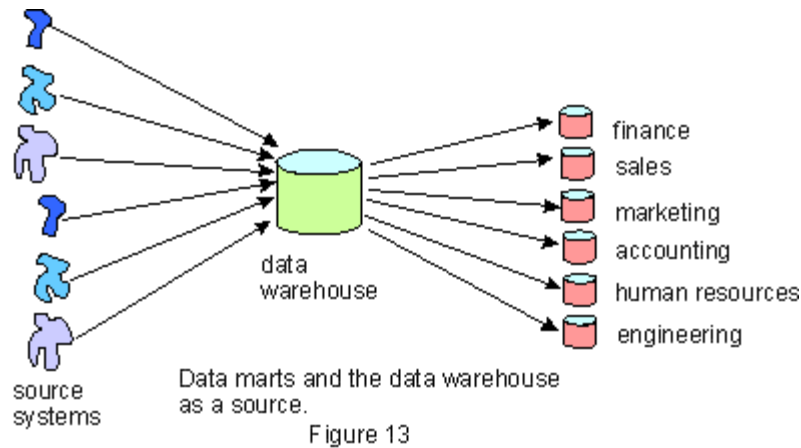
In Figure 12 it is seen that there are many sources of data and many data marts. Suddenly, the interface between the data marts and the source systems turns complex and ugly. If there are m sources of data and if there are n data marts, then the number of interface programs needed are $m \times n$.

But there are other problems with the architecture shown in Figure 12 as well. The amount of detailed redundant data among different data marts is truly significant. Each data mart collects and stores much of the same detailed data as each other data mart. The result is massive duplication of detailed data across all of the data marts.

But the largest inadequacy of the architecture in Figure 12 is the inability to reconcile data. When management asks what revenues were made last quarter, finance says \$10 million, sales says \$12 million, and marketing says \$15 million. It simply is a management nightmare to try to make decisions where there is conflicting information that cannot be resolved.

THE DATA WAREHOUSE CENTRIC ARCHITECTURE

An alternative to the data mart architecture shown in Figure 12 is the data mart/data warehouse architecture seen in Figure 13.



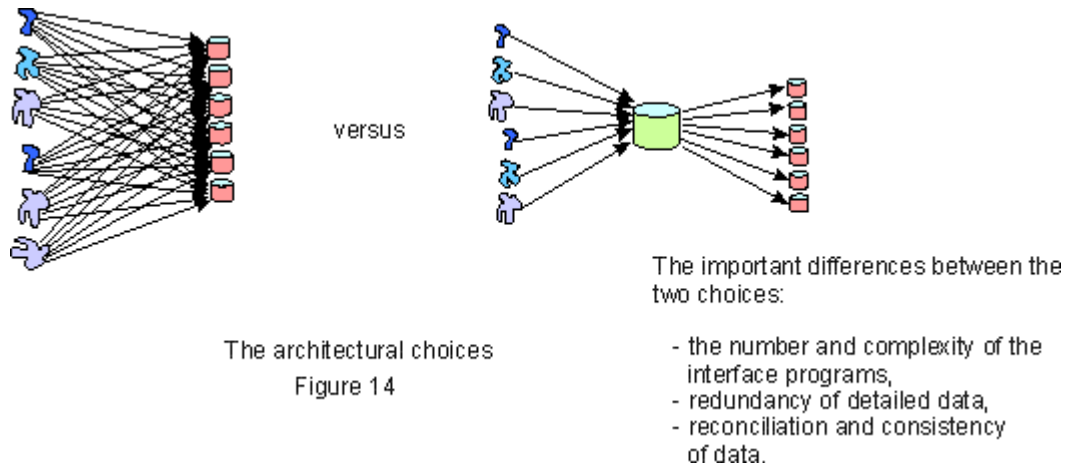
In Figure 13 there are source systems and data marts. But there is also a data warehouse. The different source systems feed data into the data warehouse. From the data warehouse data marts are created. But there are some surprising architectural differences between the architecture shown in Figure 12 and Figure 13. The data warehouse based architecture needs only a minimum number of interface programs. If there are m source systems and there are n applications, then the number of interface programs needed are $m + n$ (as opposed to $m \times n$ needed in the data mart architecture of Figure 12.)

There is no need for each data mart to store detailed data redundantly. The data warehouse contains the detailed data. In Figure 13 each data mart draws on the data warehouse to build its aggregated and summarized data. There is much, much less data found in Figure 13 than found in Figure 14.

And reconciliation of data is done easily when there is a data warehouse. If two or more departments have inconsistent results, there is the data warehouse to go to in order to reconcile the differences.

Architecturally there are then some very significant differences between the data mart architecture and the data warehouse based architecture.

Figure 14 illustrates the architectural differences.



THE COST DIFFERENCES

What then are the cost differences between the data mart architecture and the data warehouse based architecture?

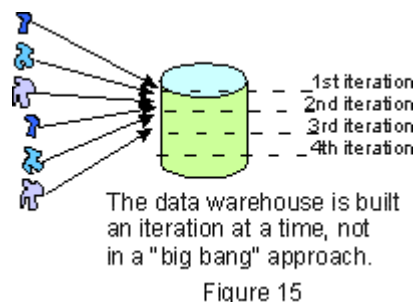
Appendix B quantifies those differences.

BUILDING THE DATA WAREHOUSE

One of the incorrect claims made by the data mart vendors is that the data warehouse must be built in an "all at once" proposition. In doing so the data warehouse development becomes an unworkable proposition. But a large development effort is not the way that data warehouses are built, when they are built properly. From the beginning the knowledgeable data warehouse practitioners have warned against the "big bang" approach to the building of the data warehouse. Instead the practitioners have always advocated the iterative approach to the construction of the data warehouse.

In the iterative approach, first one part of the data warehouse is quickly built and populated. Then another part of the warehouse is built and populated, and so forth. Throughout the development process the end user can use the warehouse and provide feedback to the developer.

Figure 15 depicts the iterative approach to the building of the data warehouse.



In Figure 15 it is seen that first one portion of the data warehouse is built. Then another portion of the data warehouse is designed and populated. Then another portion is built, and so forth. It is absolutely against the principles of proper data warehouse development to try to build the data warehouse in an “all at once” approach.

BUSINESS BASED COST JUSTIFICATION

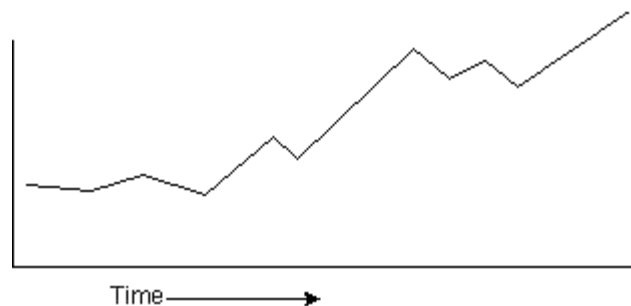
The approaches taken to quantify the development and usage of the data warehouse have previously focused on the "back office" costs, those seen by the IT developer. But there is an entirely different approach to the cost measurement and justification of the data warehouse. That approach is the business based approach.

If a data warehouse is effective, it allows an organization to:

- hold on to and increase market share
- maximize profitability
- minimize expenses.

Since a data warehouse has the potential to accomplish these very worthwhile objectives, then the worth of the data warehouse should be able to be measured in terms of the incremental fall and rise of these measurements.

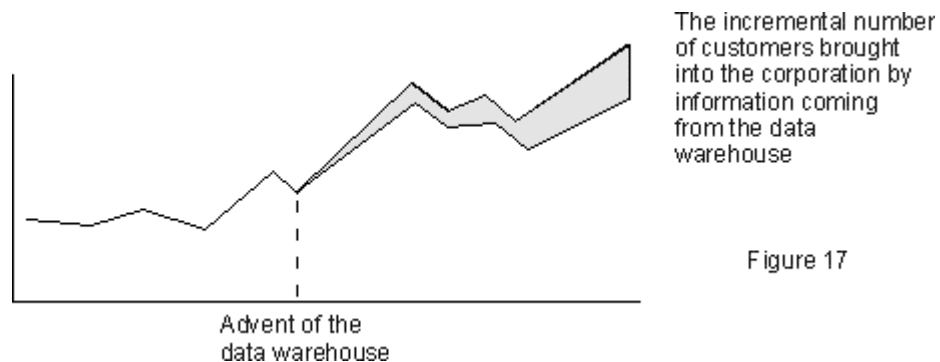
Consider a variable measured by every corporation – revenue over time. Figure 16 shows a sample measurement of revenue over time.



Total number
of customers.

Figure 16

Now suppose that a data warehouse is introduced to the corporation at point in time n . Figure 17 shows that the point n is marked.



The data warehouse administrator does an extrapolation to management that shows two things - the actual revenue made and the incremental revenue attributable to the information coming from the data warehouse. In such a manner the quantifiable argument for the cost justification of the data warehouse can be made.

But there are some severe problems with this argument, however true the argument might be. Some of the problems with this argument are:

- the incremental revenue after point in time n was due to the introduction of a new product
- the incremental revenue after point in time n was due to a change in the economy
- the incremental revenue after point in time n was due to a change in interest rates
- the incremental revenue after point in time n was due to the withdrawal of a product line by a competitor, and so forth.

The problem is that ALL of the above factors and more may account for the incremental revenue after point in time n .

It may well be that the data warehouse has accounted for significant revenue, but there are so many other relevant factors that trying to quantify the effects of the data warehouse this way is almost impossible.

SPECIFIC APPLICATIONS

Another approach to quantifying the impact of the data warehouse on the business of the corporation is to create specific applications that use the data warehouse at their core. Once built, the results of these application can be easily quantified. For example, a common data warehouse centric application is CRM (customer relationship management.) CRM uses the data warehouse to create an integrated, historical foundation of customer information.

Some of the graphics found in a typical CRM application are seen in Figure 18.

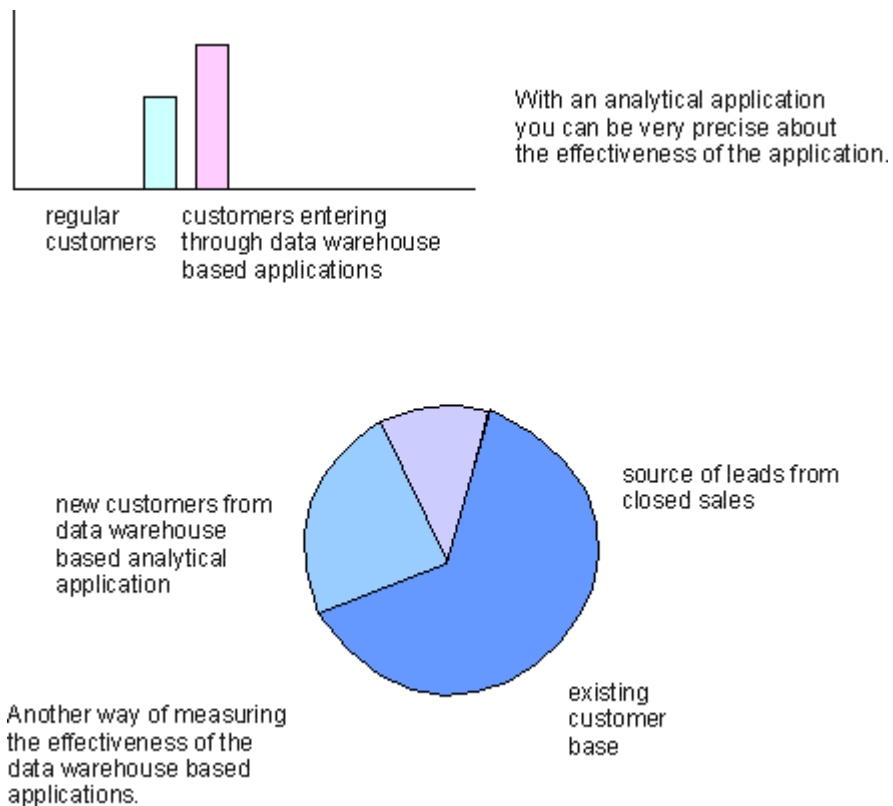


Figure 18

Once the CRM is created it is easy to measure such things as:

- how many new customers are added to the system
- how much revenue these new customers account for
- what kinds of products these new customers are attracted to
- how many times a customer has a touch point with the corporation, and so forth.

In short, with specific warehouse based applications, it is possible to quantify the business impact of the data warehouse.

Once the impact to the data warehouse can be quantified through an application, then the impact on the business can be quantified. Figure 18 shows some simple and typical qualifications that can be done based on an application.

A DUAL EFFECT

While applications based on the data warehouse are a good way to quantify the impact of the data warehouse, it is worthwhile noting that applications can only measure part of the worth of the data warehouse. The data warehouse always has the worth of lowering the cost of accessing information. Therefore the cost benefit of the data warehouse is BOTH the savings noted at the beginning of this paper AND the business benefit achieved by data warehouse centric applications.

In any case, none of the scenarios address the usefulness of the data warehouse to the explorer community. The ability to do exploration comes as a by-product of the building of the data warehouse. And ironically, when exploration is successful, it has the promise of a payback exceeding any of the measures discussed.

IN SUMMARY

This white paper has addressed the quantification of the benefits of data warehouse and data marts for the corporation.

The first approach to the cost justification of the data warehouse is that of quantifying the general costs of accessing information. When there is no data warehouse, each request for information requires the gathering of data, integration of data, and the merging and summarization of data before the request can be satisfied. But with a data warehouse, data must be gathered, integrated, and merged only once. Once the data warehouse is in place, reports can be run quickly and efficiently.

The general cost for access of information for the corporation applies to farmers, not explorers. Explorers have a random path for the access of information, which cannot be enhanced by any a priori arrangement of data. Therefore the economies of data access of the data warehouse apply to farmers where there is a predictable pattern of access.

The second discussion as to the efficiencies and economies of data warehousing relates to the differences between data warehouses and data marts. Where there is a data warehouse there is the opportunity to quickly and efficiently build the data mart. But where there is no data warehouse, each new data mart requires the same construction from the legacy application environment as the previous data mart. Taken from a long-term perspective, building a data warehouse greatly emancipates the development required to build data marts.

The third approach to the measurement of cost benefits for a data warehouse, are from a business perspective. A data warehouse contains integrated and historical data. It is a fairly simple matter for the business to turn integrated, historical data into business advantage through enhanced revenues, increased market share, and reduced expenditures.

Appendix A

Calculating the impact of a data warehouse on the access of information.

1. How many sources (applications) do you have that will feed the need for reports and screens?
 - 1.a ____ How many data bases/files/data sets are there on the average in the source systems (applications)?
 - 1.b ____ DS - different sources = 1.a x 1.b ____
2. What is the cost of locating source information for reporting and screens?
 - 2.a ____ (default = \$1,000)
What is the cost of writing code to gather source information?
 - 2.b ____ (default = \$10,000)
What is the cost of converting/integrating source information
 - 2.c ____ (default = \$10,000)
What is the cost of merging data after conversion?
 - 2.d ____ (default = \$1,000)
GC - gathering cost per source = 2.a x 2.b x 2.c x 2.d
3. What is the cost of creating a report or a screen?
 - 3.a ____ (default = \$2,500)
CR - 3.a
4. How many screens/reports will be needed across the organization over the next five years?
 - 4.a ____
RS - 4.a
5. Calculate the infrastructure costs of a data warehouse. Determine the target size of the data warehouse in gigabytes. Now multiply the gigabytes by \$7,500.
 - 5.a ____
DWC - 5.a
6. Now calculate the cost of reporting from legacy/operational applications.
REPORT COSTS = DS x (GC + CR) x RS
7. Now calculate the cost of reporting from a data warehouse.
DW REPORT COSTS = (DS x GC) + (CR x RS) + DWC

Usually the cost of reporting from legacy applications comes out so high that no reasonable organization would ever entertain spending that much money. And indeed they do not spend that much money. Instead the end users are frustrated in that they never get the information they need. This is why organizations without a data warehouse are said to be data rich and information poor.

Appendix B

Calculating whether a data warehouse or a data mart should be built first (or whether a data warehouse should be built at all).

1. How many sources (applications) do you have that will feed the data warehouse and/or data mart environment?
 - 1.a ____ How many data bases/files/data sets are there on the average in the source systems (applications)?
 - 1.b ____ DS - different sources = 1.a x 1.b ____
2. How many data marts will be needed in the long run?
 - 2.a ____ (default = 10) DM - 2.a
3. How many interface programs will be needed?
IP - DM x DS
4. What is the cost of creating an average interface program?
 - 4.a ____ (default = \$20,000)AIP - 4.a
5. Calculate the infrastructure costs of a data warehouse. Determine the target size of the data warehouse in gigabytes.
 - 5.a ____ Now multiply the gigabytes by \$7,500.
DWC - 5.a x 7,500
6. Calculate the cost of detailed redundant data that each data mart will have. Each data mart will have 1/2 of the detailed data found in the data warehouse.
 - 6.a ____ = 5.a / 2.
7. Estimate the number of gigabytes of redundant detailed data in an average data mart. Now multiply that number by \$7,500.
 - 6.a = 5.a x 7,500
DMRD - 6.a
8. Now calculate the cost of the data mart centric environment.
(IP x AIP) + (DM x DMRD)
9. Now calculate the cost of the data warehouse centric environment.
((DS + DM) x AIP) + DWC