



Inmon Consulting

SOME STRAIGHT TALK ABOUT THE COSTS OF DATA WAREHOUSING

An Inmon Consulting White Paper

Inmon Consulting
PO Box 210
200 Wilcox Street
Castle Rock, Colorado
303-681-6772

By W H Inmon

SOME STRAIGHT TALK ABOUT THE COSTS OF DATA WAREHOUSING

By W H Inmon

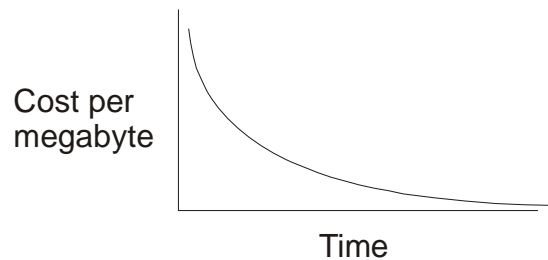
Everybody knows that data warehouses cost a lot of money. Everybody knows that when you have a data warehouse that the corporate budget will be affected. But what everybody doesn't know is that the costs of a data warehouse do not have to eat you out of house and home. There are some good ways to effectively mitigate the large amounts of dollars that a data warehouse is reputed to swallow up.

DOES A DATA WAREHOUSE HAVE TO COST A LOT OF MONEY?

The truth of the matter is that the long-term cost of the data warehouse depends more on the developers and designers and the decisions they make than on the actual cost of technology. Stated differently – a data warehouse can cost a lot of money, but a data warehouse does not have to cost a lot of money. In order to understand what some of the options in the building and operation of a data warehouse are and why the design options are so important to the cost of the data warehouse, consider some time honored conventional wisdoms that have been put forth into the community of technology and business. One of those conventional wisdoms is that storage costs are getting cheaper all the time. Another of those conventional wisdoms is that processors are getting cheaper all the time, as well as storage. But what happens when cheaper storage and processors are not the problem and the costs associated with the management and expertise and additional technologies are? That's when it's time to change the rules...when conventional wisdom must give way to innovation.

STORAGE IS GETTING CHEAPER ALL THE TIME

The diagram in Fig 1 illustrates this well-repeated conventional wisdom.



The classical curve of declining cost of storage

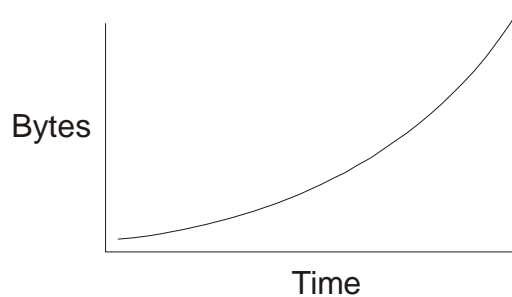
Fig 1

The diagram in Fig 1 shows that the unit cost of storage is getting less expensive every year. The purveyor of this diagram quotes Moore's law and then draws the conclusion that we should not worry about the costs of storage for our future systems design and development effort. The implications drawn from this diagram are that in the future the costs of storage will tend to be free or close to it. When this costing of storage occurs, storage becomes commoditized.

This conclusion – that we should not worry about the costs of storage in the future – is a gross distortion of the facts. Reality – as we shall see – is something quite different. It is no surprise that Moore's law is quoted by hardware and storage vendors – who have the most to gain by the gross misinterpretation of the lowered unit costs of storage on the spending habits in the marketplace.

THE CONSUMPTION CURVE

There are several other pieces of information that must be taken into account when considering the long-term costs of storage. One piece of information is that while the unit cost of storage is dropping, the demand for storage far outstrips the drop in the costs of megabytes of storage. Fig 2 shows the increase in the demands for storage.



The storage consumption curve

Fig 2

The storage consumption curve is not as widely known and quoted as the declining cost of storage curve. But industry sources such as IDEMA INSIGHT (James Parker, Vol. IX, no 5) support this consumption curve of storage. The truth is that storage is being consumed at a rate faster than the rate at which storage price is dropping.

And along with the consumption curve of data is a corresponding consumption curve for processors. It is not just that we have more data. We need more processors to do something with that data. And it is not just the volume of data and the cost of the infrastructure that presents a challenge. The pathways that allow the data to be accessed need to be increased as well. With a large amount of data and only a limited pathway into the data, it is inevitable that bottlenecks to performance will develop.

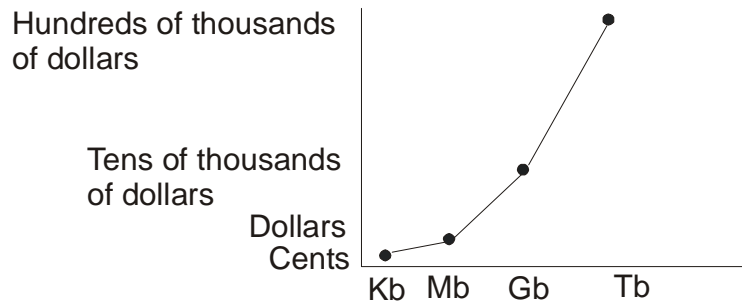
THE INFRASTRUCTURE COST OF STORAGE

But the real culprit in the storage/cost conundrum is the fact that as the volume of data increases, the infrastructure cost for the management of data increases far, far greater than the rate of the drop of the unit cost of storage.

In order to see this equation at work, you only have to go to Radio Shack and buy some storage for your computer. You can buy several gigabytes for your personal computer. The cost will be minimal. And the cost of your computer – which will ultimately house the storage – is several hundred dollars or maybe even a few thousand dollars.

So for smaller volumes of data, the cost of the management infrastructure is minimal. But for several terabytes of storage managed by IBM or Teradata in a data warehouse environment the cost of the infrastructure to manage those terabytes of data may be from \$500,000 to \$1,000,000 per terabyte. The actual cost of storage is a rounding error compared to the cost of the storage management infrastructure for the larger volumes of data.

Fig 3 shows this curve.



The storage infrastructure cost per unit of storage

Fig 3

It is noted that the curve shown in Fig 3 is logarithmic, not linear. The cost of the storage management infrastructure skyrockets as the volume of data increases. Stated differently, when you buy storage for a data warehouse, you are really paying for the storage management infrastructure, not the storage itself.

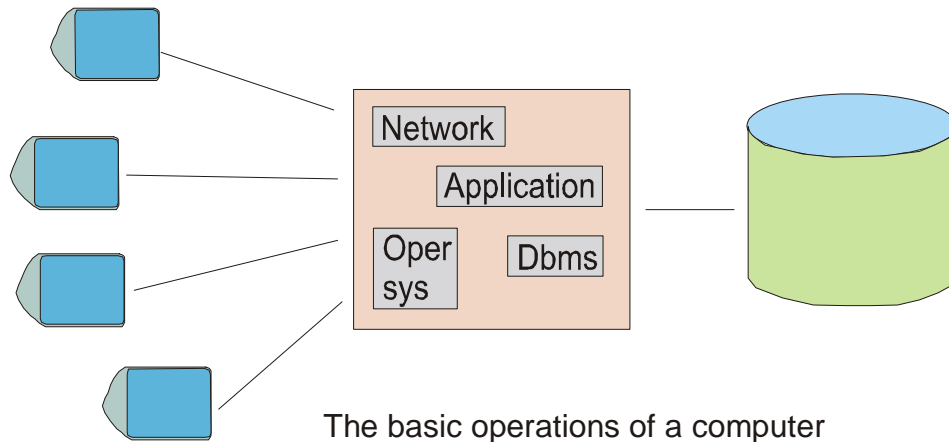
The same conditions hold true for processors. The most expensive processor cycles are those that are managed by the largest computers. The smaller the computer, the less expensive the processing cycle. Stated differently, the unit cost of processing cycles is more expensive, the larger the processor that you get. And when you have lots of data in a classical processing environment, the more processing cycles you need.

THE FUNCTIONALITY OF THE STORAGE MANAGEMENT INFRASTRUCTURE

So what kind of processing is going on that is so expensive in large processors and in large collections of data? In addition to the sheer volume of data driving the costs of storage management infrastructure higher, the actual functionality of the storage under management is a factor as well. It is not just the total volume of storage that needs to be managed that is a factor; it is what is being done to the storage by the infrastructure that costs as well.

Stated differently, it is a lot more expensive to manage an OLTP infrastructure of a terabyte of data than it is to manage an archival environment of ten terabytes of data. The degree of the functionality of processing supported by the infrastructure plays a big part in the cost of the infrastructure.

In order to understand the role played by the functionality of the processing being done to the data under storage management, consider the simple diagram of Fig 4 that shows the basic components of the modern computer.



The basic operations of a computer

Fig 4

Fig 4 shows the basic components of the computer. There is the network manager, which sends and receives messages from a network. There is the application function where specified processing is executed. There is the operating system that manages the functions and their priorities. And there is the dbms that manages the access and interchange of data to and from the computer.

These functions are pretty much generic to any computer. They are found in a \$1,000 personal computer and they are found in the largest and most sophisticated parallel processor that sells for millions of dollars.

It is a temptation to say that you can go to Radio Shack and buy a terabyte of storage. Indeed you can do just that. But when you come home and start to use the terabyte of data, you are going to find out that your personal computer operates quite differently from a large-scale parallel processor. You cannot run a world wide OLTP environment from your home computer.

THE FUNCTIONALITY OF A FULL-SCALE PARALLEL PROCESSOR

So what infrastructure services does a full-scale parallel processor offer? What is the functionality that drives up the price of the data management infrastructure? Fig 5 shows some of the functionality found in a full-scale parallel processor.

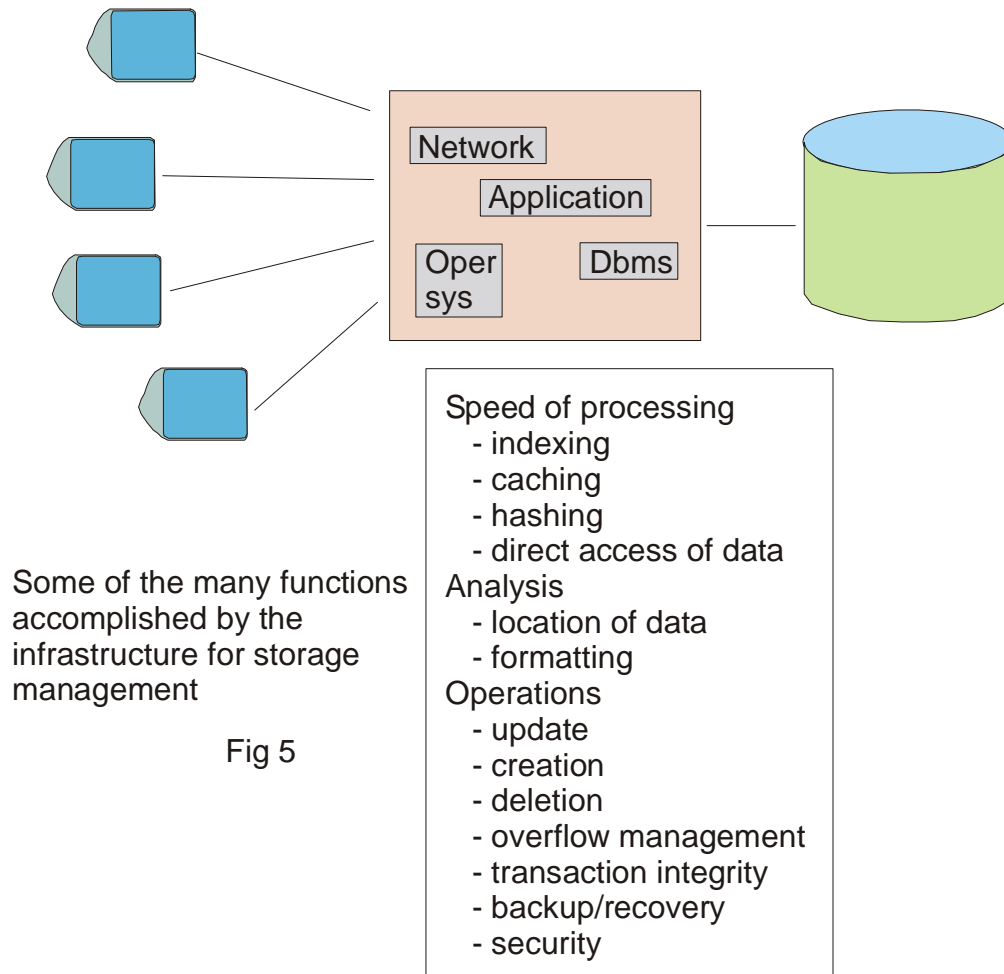


Fig 5 shows that a full-scale parallel processor offers quite a bit of functionality.

The first service offered is that of running in parallel. Running in parallel requires quite a bit of sophistication. Running in parallel requires the coordination and integration of many processors operating separately. But there are other features to the large scale management of storage.

Some of these features include –

- speed of processing
- indexing
- caching
- hashing
- direct access of data
- analysis
- location of data
- formatting of data
- operations
- update of data
- creation of data
- deletion of data
- overflow management
- transaction integrity
- backup and recovery
- security

and so forth.

It is seen that for large amounts of data that the data management infrastructure is quite sophisticated. It is no surprise then that the costs of storage rise as the volume of data and the sophisticated use of that data escalate.

Stated differently, it is no surprise that the costs of storage have much more to do with the management infrastructure than they do with the unit cost of storage.

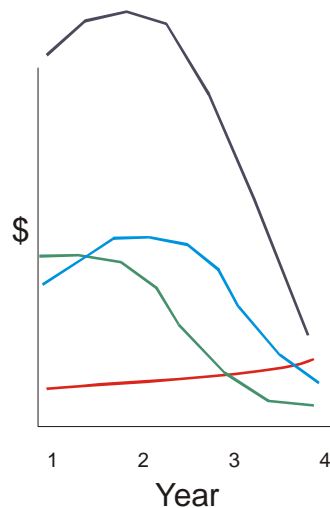
DATA WAREHOUSE COSTS

Experience has shown that there are four main components to the cost of a data warehouse. These components are –

- Consulting and development costs
- Storage and storage infrastructure costs
- ETL costs
- Dbms and other infrastructure costs.

Consulting and development costs are those costs that relate to the building of the data warehouse. These costs include data modeling, design, mapping, ETL and data population costs. These costs can be done in-house or by contracting to a consulting firm, or both. Storage costs include the cost of storage and the infrastructure. ETL costs are software costs that include the ETL software, metadata capture, transformation logic, source and target mapping, and so forth. Dbms and other infrastructure costs include software licensing, network, processor, and other infrastructure costs.

Fig 6 shows the short-term (up to 4 year) costs associated with the building and operation of a data warehouse.



The short-term costs of a data warehouse



It is seen that consulting and development costs are major ticket items. A lot of money will be spent on the design, development, and deployment of the data warehouse. The good news is that these costs diminish over time. While a certain amount of ongoing work will need to be done to the data warehouse, the major part of the work will be done in the first few years of the life of the data warehouse. After the first few iterations are built, the development work degenerates to maintenance work. And as maintenance is done, the high cost of development goes away.

The cost of storage in the first few years of the data warehouse is not high. In the first few years there is relatively little historical data, and there are only a few subject areas that are built into the data warehouse. It is over time that large amounts of history appear and that lots of subject areas are entered into the data warehouse.

ETL (extract/transform/load) processing is the processing that occurs as data is read from the legacy application environment and is transformed into the data warehouse environment. ETL is normally done by software. However, for a really small data warehouse ETL processing can be done manually. But when ETL processing is done manually, its long-term costs exceed that of doing ETL by software.

The costs of ETL are seen to diminish as the data warehouse becomes mature.

Dbms and other infrastructure costs include software and hardware. In the diagram shown in Fig 6 it is assumed that there is a new license of software that is needed and that there are more processors that are needed. If there already is a software license or where software licensing can be done on an incremental basis, then the initial costs will be less. If the initial CPU and/or CPU license cost is one-time then the maintenance costs are perpetual and accumulating unless there is a site license agreement in place. In addition, there is a cost to burn in the software. If the software is already being used, then there will be much less of a burn in cost.

In addition, training is needed for the installation and usage of both ETL and dbms software.

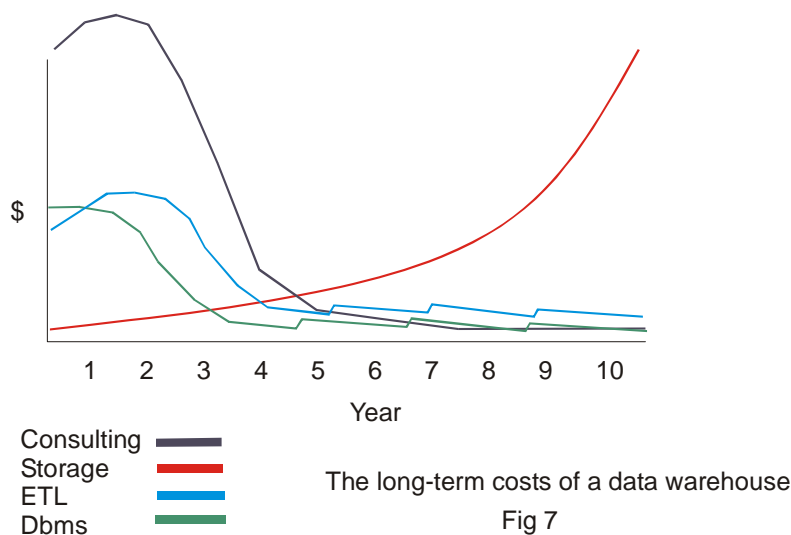
It is seen that as time passes the costs for dbms and processors diminish after the initial purchase and acquisition are made.

Looking at Fig 6, it is seen that the cost of storage is not particularly significant in the life of the building and usage of the data warehouse in the short term. So why should a person worry about storage costs?

In fact the costs of storage are so significant that in many organizations only summarized or aggregated data is archived. The details of data are lost because it is simply too expensive to store them over time. And unfortunately the details of data are the place where many of the most interesting and most useful data lie.

THE LONG-TERM COSTS OF A DATA WAREHOUSE

The long-term picture for the costs of data warehousing looks considerably different however. Fig 7 shows the long-term picture of the costs of the data warehouse.



The picture of costs of a data warehouse seen in Fig 7 is considerably different than the picture of costs seen in Fig 6. Fig 7 depicts the costs of data warehousing over a ten-year time frame. In Fig 7 it is seen that at the end of ten years – far and away – the costs of the data warehouse are dedicated to storage. In fact it is because of the large amounts of storage and the infrastructure needed to manage that volume of data that the cost of data warehousing is so large over a lengthy period of time.

But do the costs of data warehousing have to rise at the rate seen in Fig 7? The answer is not at all. It is possible to mitigate the long-term costs and the effective ability to use the data found in a data warehouse by the introduction of what is termed a “data warehouse appliance.”

A data warehouse appliance is a combination of hardware and software that is designed to manage very large amounts of storage but at a significantly lower rate than traditional storage.

The data warehouse appliance – such as that offered by Dataupia – allows the data warehouse environment to remain intact. If an organization starts with an Oracle license, they make no changes to the environment as the data warehouse appliance is implemented. If an organization starts with DB2, then they make no changes to the environment as the data warehouse appliance is implemented. The deployment of the data warehouse appliance is independent of and transparent to the dbms. In the case of Dataupia, the data warehouse appliance will act as an MPP “foundation” for your current rdbms vendor

The dbms accesses data as it has always done. Only some or all of the data being managed is being managed by the data warehouse appliance. And the cost of storage management under the data warehouse appliance is a fraction of that required by traditional storage management equipment.

Fig 8 shows what happens to the costs of storage in a data warehouse environment upon the introduction of the data warehouse appliance.

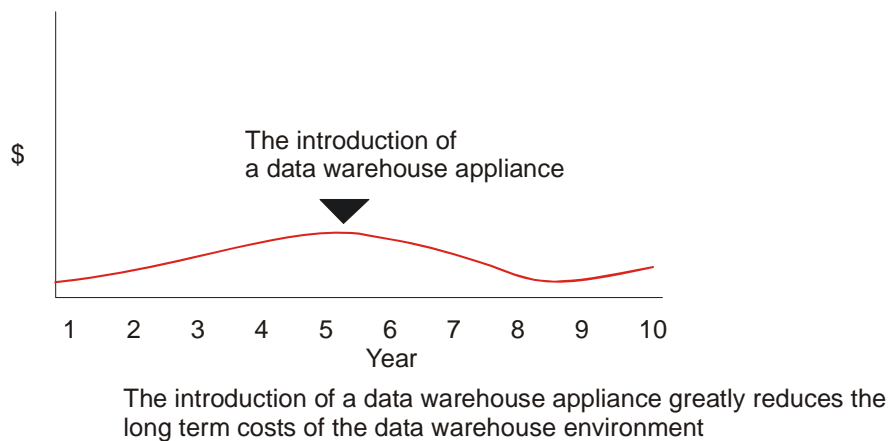


Fig 8

In Fig 8 it is seen that the escalating costs of storage are mitigated by the introduction of a data warehouse appliance. Once the costs of storage are mitigated, the long term costs of the data warehouse do not spiral out of control. An organization is free to build as large a data warehouse as they need with no real consideration to the costs of the data warehouse.

Another way to look at the phenomenon of the introduction a data warehouse appliance into the equation is seen in Fig 9.

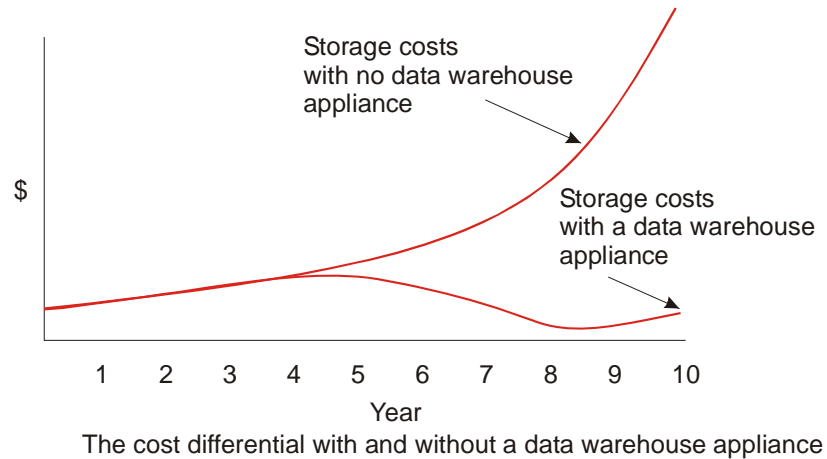


Fig 9

Fig 9 shows explicitly what the effect of the introduction of a data warehouse appliance into the data warehouse environment in the long term is. The amount of money needed to build and operate the data warehouse is greatly minimized by the usage of a data warehouse appliance.

It is because of this decision – to use or not to use a data warehouse appliance – that it is asserted that the costs of the data warehouse depend more on the design and development decision-makers than on the actual costs of equipment for the data warehouse. If a designer or developer chooses to make the costs of a data warehouse high, then a designer or developer has the ability to do just that. But if a designer or developer chooses to make the costs of a data warehouse low, then the designer/developer has the ability to do just that as well.

The money saved by designers allows them to buy other technologies and/or technologies which allow them to further increase ROI and the business value through its data warehouse user.