



TECHNOLOGY AND EVOLUTION: DATAUIA AND DW 2.0

An Inmon Consulting White Paper

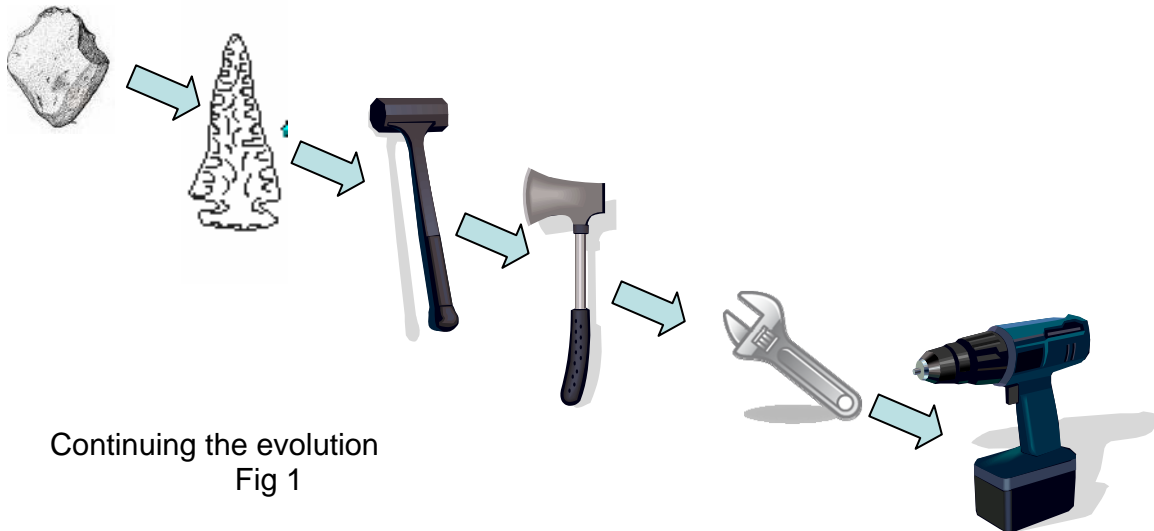
Inmon Consulting
PO Box 210
200 Wilcox Street
Castle Rock, Colorado
303-681-6772

TECHNOLOGY AND EVOLUTION: DATAUPIA AND DW 2.0

By W H Inmon

As you look around, the inevitable results of evolution are found everywhere. In 1920 we could drive the Model T. Today we have SUV's, sports cars, trucks and delivery vans. In 1950 we had the black and white TV. Today we have color TV, flat panel display, HDTV, in house television and a myriad of other forms of interactive display and transmission of images and sound. In the mountains of Chile are found caves that mankind inhabited 20,000 years ago. Today we have condos, trailers, apartments, and ranch style houses that we live in.

Everywhere you look you see evolution. It is either occurring now or has occurred in the past.



Continuing the evolution
Fig 1

In no small part evolution has found its way into technology. The world of technology today – the one that we all take for granted – is apparent and surrounds us all, and has been profoundly shaped by evolution.

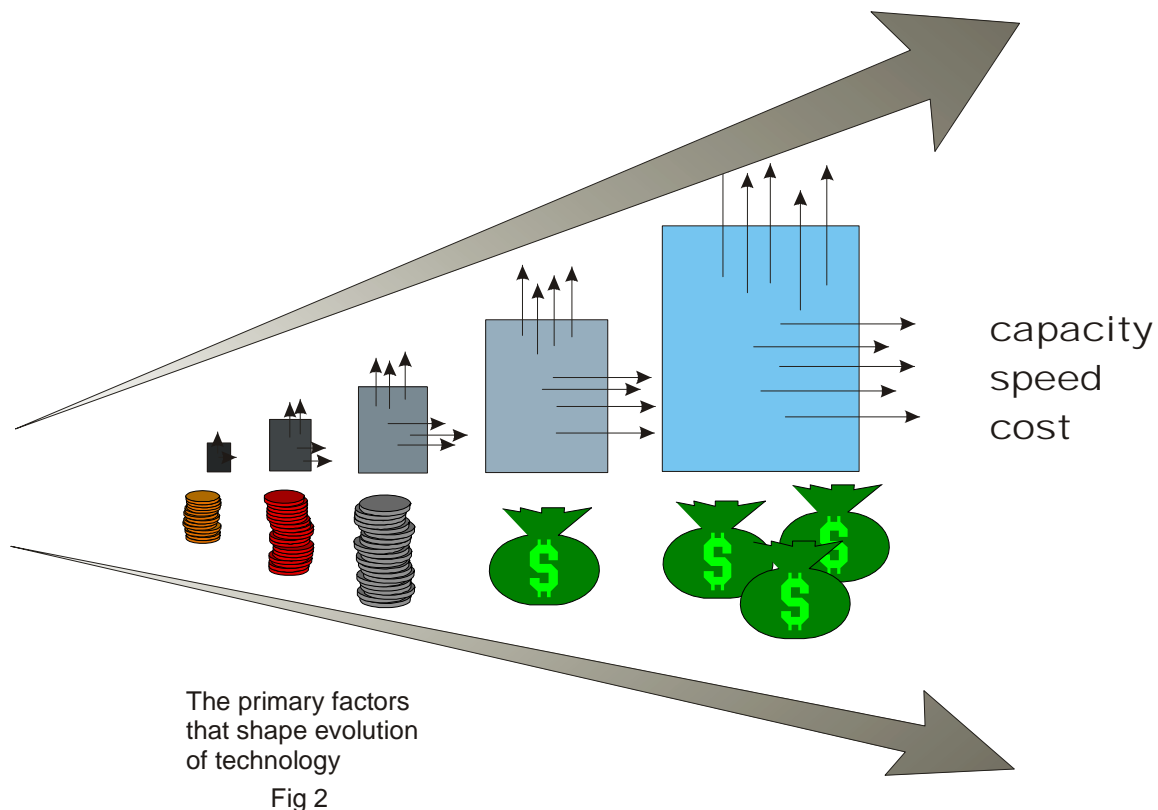
THE FORCES THAT SHAPE EVOLUTION

What are and have been the forces of evolution that have shaped technology? In truth there have been many evolutionary factors that have shaped technology. But there are three powerful forces of evolution that stand above all other evolutionary forces. Those three most powerful evolutionary forces that have shaped technology are –

- capacity – the ever increasing demand for more – more storage, more processors, more transmissions, and so forth.
- speed – it is not enough for technology just to have greater capacity. In addition technology must work at ever increasing speeds. There are more people using technology, more kinds of processing, more sophistication, more transactions, more

Internet transmissions every day. The number and type of interactions by people with the computer always goes up, never down.

- cost. For all of the benefits of technology and all of the functions that are performed by computers, cost is an issue. On the one hand the unit cost of computing has been going down for years (see Moore's law). But as fast as the unit cost of processing has been dropping, the rate of consumption is going up even faster. The cost of computing then is another shaping factor of evolution.



The cost of the continual dropping of the unit price of computing can be a misleading figure. While it is true that the unit cost of computing is dropping, the larger the processor, the greater the costs. The most expensive machine cycles occur on the largest machine. The least expensive machine cycles occur on the smallest machine. Therefore, before one blithely makes the assumption that the costs of computing are forever dropping, one must qualify that thought with the size of the machine and processor that is being considered.

THE EVOLUTION OF PROCESSORS

The evolution of technology is nowhere more obvious than in the processors and their associated storage technology that the computer industry has witnessed and has been a part of.

In the beginning there were very small, very crude processors. Early forms of storage included paper tape and Hollerith punch cards. Compared to what preceded these

technologies, these paper based technologies were powerful. But with these technologies came some severe drawbacks. Card decks were bulky and often got out of sequence. The punching of chads was less than a perfect science. Once a card was punched it was not reusable. And data could only be accessed sequentially. Processing was slow and expensive.

A vast improvement in information processing technology occurred when magnetic tape appeared. Data could be stored magnetically in a much smaller amount of space. The media on which data was stored was able to be reused. The reading of tapes was significantly faster than the reading of punched cards. But there were issues with magnetic tape. The oxide easily corroded off of the tapes making them unreadable. And magnetic tape could only be read sequentially.

At the same time as the advance to magnetic tape was being made, operating systems and speed and capacity were improving. There was starting to be some standardization of software as the early machines went from their early transistor state to faster and more powerful forms of processing. More data could be processed more reliably in a shorter amount of time.

Things really got to be interesting with the advent of disk storage. With disk storage data could be accessed directly. This meant no more horrendous latency times. In addition data could be accessed by means of software called a data base management system. The direct interface to the data was done through data management software, not directly by the programmer. Almost as soon as disk storage and data base management systems appeared, there appeared online transaction processing. With online transaction processing, transactions could be completed in a matter of seconds or less. And it is with online transaction processing that the computer worked its way into the business fabric of corporations. Prior to the advent of online transaction processing, the computer had been used in a “batch” mode. In the batch mode transactions were queued throughout the day. Then when enough transactions were queued they were run. This batching of transactions served the purpose of optimizing the computer time required for processing, but required the end user to wait – usually overnight – for the transactions to be run. In order to make business decisions, it was not realistic to ask the end user to wait 24 hours for the results of processing. The consequence was that prior to the online usage of the computer, there were only a handful of activities the computer could do – accounts payable, accounts receivable, human resources, and so forth.

THE DAY TO DAY INVOLVEMENT OF THE COMPUTER

But with the advent of online transaction processing, it was possible to directly involve the business with the usage of the computer. For the first time the computer became a part of day to day detailed business decisions. Bank teller systems used the computer to determine if a customer could cash a check. Airline reservations used the computer to manage ticket sales. Insurance companies used the computer to settle claims and to record premium payments. Manufacturers used the computer to manage inventory and manufacturing events.

In a word, online transaction processing married the computer to the business in a way that had never before been possible.

The close interaction with the business drove the demand for capacity and speed to places never before imagined. It is truly fortunate that the unit cost of processing was dropping because the demand for computers was nothing short of voracious.

Circuitry became smaller and faster. Memory was produced by machines. Storage became more reliable and faster. And the volume of transactions simply mushroomed.

NO INTEGRATION

With the explosion of transactions that occurred came an explosion in applications. And after enough applications appeared came the realization that the many applications were not integrated. The problem with unintegrated applications is that there is no definitive source of data. Applications turned into silos of data and the same data element was found in many silos. And in each silo was a different value of data for the same element. The result was a lack of integrity of data. No one in the organization ever really knew what the accurate amount of any given unit of data was at any point in time.

NO HISTORICAL DATA

Coupled with the lack of integrity of application based data came the realization that there was no historical data, to speak of. Transaction systems jettisoned historical data as soon as possible in the desire to achieve the highest performance possible. The result was that there was little or no historical data stored in the many applications the corporation had.

Not having historical data was of little concern to the clerical community who was primarily concerned with day to day detailed, up-to-the-second transactions and data. But to the management and the analytical community, the lack of historical data was a significant limitation to the kind of information the corporate analyst had access to. In years past the only significant amount of historical data that was available was externally generated data, such as statistical data relating to the economy or the prime lending rate. This historical data was usually gathered by government organizations.

ENTER THE DATA WAREHOUSE

The lack of integration and the lack of historical data led corporations to data warehousing. The data warehouse was a structure designed for management's information needs. Data in the data warehouse was integrated. In addition historical data was placed in the data warehouse. And finally data in the warehouse was at a low level of detail. The low level of detail allowed the same data to be used for many purposes and many functional areas in the corporation.

With the data warehouse, for the first time, the corporation had a basis for decision making. The data warehouse structure quickly spread across the landscape, to the point that data warehouses, in one form or another, are found in almost all major organizations and corporations.

DATA MART CONFUSION

There has been some confusion as to the difference between a data warehouse and a data mart. A data mart was a structure that served the colloquial interests of a group of people with similar needs for information. Typically there were data marts for accounting, sales, finance, marketing, management, engineering, human resources and so forth. Often times, data marts were silos of information, where there was no relationship between data in different silos. Each data mart had its own unique, uncoordinated set of data.

At first the differences between a data mart and a data warehouse were not apparent. But as the organization grew and the number of data marts grew, the differences between a data mart and a data warehouse became readily apparent. The data warehouse served the entire organization while the data mart was aimed at just one group of people. The data warehouse was not optimized for the information needs of any one organization, while the data mart was keenly optimized for the processing needs of a single audience.

VENDOR HYPE AND THE DEFINITION OF A DATA WAREHOUSE

Such was the attractiveness of the data warehouse that all sorts of vendors began to proclaim that their technology was suited to data warehousing. Even when there was no real affinity at all between the data warehouse and the vendors' products, the vendors saw that the business consumer was attracted to the data warehouse. In order to capitalize on the demand for data warehouses, vendors started to invent new forms of data warehousing (which may or may not actually be a data warehouse at all).

One form of the vendor invented data warehouse was the federated data warehouse. The federated data warehouse arose because people were reluctant to integrate their legacy data. Integration is a difficult topic for most organizations. Integration entails going backward in time and bringing together systems that were never meant to be integrated. Differences in definition, calculation, algorithms, presentation, storage, technology all have to be reconciled. As such integration is a large and complex task which organizations would rather avoid. The federated approach promised that there could be an "immediate" data warehouse where no complex and difficult integration was needed. A federated data warehouse was an amalgamation of data that was already in the form of a data base. Several data bases were "federated". In fact, a federated data warehouse is no data warehouse at all. There are many pitfalls to the idea of federating data.

Another form of a vendor-supported quasi data warehouse was the data mart data warehouse. Like the federated approach, the data mart approach was the promise that data marts could be created directly from legacy applications without bothering to go through the process of integration. The data mart vendors tried to convince people that there was no difference between a data warehouse and a data mart. In doing so, the data mart vendor could capitalize on the hunger for data warehouses by selling the uninformed consumer a data mart. In the early days of data warehousing there was a lot of confusion and consumers were taken in by this subterfuge.

There were indeed so many different forms of what vendors were trying to convince people was a data warehouse that the term data warehouse lost all or most of its meaning.

OTHER DATA WAREHOUSE ISSUES – THE NEED FOR MORE CAPACITY

At the same time that the term data warehouse was losing its meaning, other issues were arising with data warehousing. From an architectural standpoint, machines grew larger and larger until they reached a point after which it was difficult to grow any more. The uni processor reached a peak. The first advancement beyond the uni processor was the SMP architecture, where multiple uni processors were lashed together. Usually an SMP architecture consisted of four uni processors hooked together sharing a common memory. In doing so a certain amount of parallelism could be achieved.

The great advantage of parallel processing was that work could be done independently inside a processor. With four processors sharing the workload the elapsed time required to solve a problem was cut into fourths (in the best of cases). The workload was coordinated through shared memory.

SMP'S

There is no question that the evolution from uni processor to SMP architecture enhanced the processing power of the computer. SMP architecture represented a definite increase in capacity and speed over a uni processor architecture.

Unfortunately, because of the complexity of synchronizing processing over multiple computers with shared memory, the SMP approach to parallelism reached limitations.

MPP'S

However, parallelism was and is a powerful tool in the management of large volumes of data where response time is an issue. Even if the SMP architecture reached its limitation, the advantages of parallelism took another form, such as the allure and promise of a parallel approach to processing. The next form of parallelism was that of an MPP architecture. In an MPP architecture there are still separate and independent processors. In an MPP environment there is no shared memory. In fact nothing is shared. Each processor operates entirely independently from any other processor. The great advantage of the MPP shared nothing approach is that independent processors can be added indefinitely. And each time an independent processor is added, the amount of data that can be processed increases and performance improves.

With SMP architectures there is a very finite limit to the number of processors that can be added. But with MPP architecture, theoretically there is no limit to the number of processors that can be harnessed to work together. At least there is no technological limit as to the number of processors that can operate in an MPP configuration. The limitation of the number of processors that can work together becomes an economic limitation, not a technological one.

MPP AND COSTS

There was/is a real irony to the cost of the infrastructure associated with MPP architecture. While the unit cost of technology was dropping, the infrastructure required to hold together the different processors in the MPP environment was not dropping at all. In fact the cost to the consumer of an MPP approach was/is very expensive. In the

meantime, organizations were/are adding more data each day, and each day the addition of data takes the organization one more inexorable step toward an MPP architecture. The truth is that with a large enough volume of data an MPP architecture becomes mandatory. It is simply the evolutionary truth.

The world then arrived at a critical turning point in the evolution of both architecture and processing at approximately the same time –

- from an architectural standpoint the world was confused by vendors calling every product a “data warehouse”, and
- from a processor standpoint organizations were being driven to larger and larger configurations, toward an MPP solution. But the cost of the classical MPP infrastructure was prohibitive in the worst case and just plain outrageously expensive in the best.

True to form, the forces of evolution once again opened the door for innovation in architecture and in the processing infrastructure.

DW 2.0

From an architectural standpoint, evolution pushed the world of information management to DW 2.0. Other major aspects of technology experienced this maturation, as witnessed by Web 2.0. DW 2.0 is the next generation of architecture for information systems and architecture. There are several noteworthy features of DW 2.0 that set it apart from other architectural approaches. Some of the noteworthy features include –

- the recognition of the life cycle of data inside the data warehouse. As data in the data warehouse ages, the probability of access of data changes and the volumes of data increase. This has a dramatic effect on the way that analytical information systems are constructed. Entire types of functionality are rearranged depending on the life cycle of data that is being handled,
- unstructured, textual data is as important as classical transaction oriented data. In the early years of data warehousing data that appeared in the data warehouse came from the world of transaction processing. Transaction data is very structured. Standard data base management systems are geared for handling structured data where the same type of data appears repeatedly. But when it came to textual data there was no repeatability of data, and standard data base management systems were not geared for that kind of data. Nevertheless, there is a wealth of information that is buried in unstructured, textual data and that data certainly belongs in a corporate data warehouse,
- the recognition that there are many different parts to an enterprise data warehouse and that it is necessary to hold those parts together. The best mechanism for gluing together the many different parts of the data warehouse environment is metadata. It is metadata that allows older data to be cohesively managed along with newer data,

for example. For a variety of reasons metadata was not recognized as a part of the first generation data warehouse environment. Such an omission might have been OK for small, nascent data warehouses. But as data warehouses grew, their size and sophistication simply could not tolerate the lack of metadata as a binding element. Thus DW 2.0 emancipates metadata from the shadows and into the spotlight as an integral part of the information, analytical environment.

In truth there are many other aspects to the DW 2.0 definition. But these three features address the heart of the evolutionary advances represented by DW 2.0.

BEYOND MPP

At the same time that DW 2.0 was being formulated, the forces of evolution shaped changes in the processor marketplace. In particular the MPP marketplace was being reshaped. There was/is an evolutionary push toward MPP architecture. Most corporations are either at the point of needing MPP technology now or will be at that point in the very near future. MPP architecture is simply the technological answer for large volumes of data.

The problem with the classical MPP environment is its expense. The unit cost of storage may be falling but the cost of storage managed under an MPP umbrella is hardly inexpensive. So organizations are being compelled to move to an expensive set of technological choices.

DATA WAREHOUSE APPLIANCES

True to form, evolutionary forces have created an alternative to traditional MPP architecture. This alternative is called a “data warehouse appliance”. The data warehouse appliance takes advantage of the MPP architecture where independent, shared nothing processors can work in tandem with each other. As such there is linear scalability of processors, which is what is needed for the large volumes of data that organizations are facing. But the cost of data warehouse appliances is significantly less than classical MPP technology (which is exactly the result that is expected from the evolutionary forces that are at work). In fact the evolutionary forces may have already caused an evolution of MPP architecture into a “build – your – own” MPP approach and the appliance approach.

So MPP technology is unquestionably the wave of the future. But there is a problem associated with MPP technology, even when it is affordable and is in the form of an appliance. That problem that arises is that deploying most data warehouse appliances requires a conversion of data. Conversions require that data be moved from one environment to the next. The movement of data is actually a simple act. But the real challenge of conversions is that of the software that operates on the data. Queries, analyses, utilities all operate on data, and often times when data is moved these pieces of software must be reprogrammed, or at least modified. It is one thing to move data from one technology to another. It is another thing entirely to move all of the software that operates on the data from one environment to the next. To some extent the shock of the movement of the data can be mitigated by using BI tools that can operate in more than

one environment. But even in the best of cases, there is always unexpected unsettlement that goes along with a conversion of data.

THE “C” WORD

The problem with the conversion of data is the software that is affected, not just the movement of the data itself.

When organizations get ready to go to data warehouse appliances there are basically two scenarios under which the conversion is made. In some cases the conversion is made from a non MPP environment to the large scale MPP data warehouse appliance environment. The organization may be on an SMP that is overwhelmed. Or the organization may be simply planning ahead in anticipation of large volumes of data and may not even be on an SMP environment at all.

The other scenario is that an organization is already in a classical, expensive MPP environment and is tired of the costs of traditional MPP architecture. In the first case the organization has a “green field” set of circumstances. In the second case the organization does not have a green field set of circumstances.

In the case where the organization wants to move off of expensive classical MPP technology, there is yet another consideration. That consideration is this – is there a need to move all of the data off of the classical MPP technology or is there a need to move only some of the data off of classical MPP technology. In the case where there is a wholesale movement of data from the classical MPP environment, conversion of data and conversion of software is definitely an issue. In the case where there is the movement of only a fraction of the data off of classical MPP technology, conversion of software may not be such a large issue.

DBMS TRANSPARENCY

The ultimate on the evolutionary scale for MPP oriented, data warehouse appliances is the case where there is plug compatibility (or “transparency”) of the classical MPP technology with the data warehouse appliance. In this case the only conversion required is that of the physical movement of data (which is a very easy task to accomplish). Because the data warehouse appliance that runs the data warehouse is plug compatible, the software that the classical MPP runs on remains the same. To the end user the transfer of data from classical MPP technology to data warehouse appliance technology is truly seamless. Even the operating system and dbms do not know that such a conversion has taken place. The only person that really knows that a conversion from classical MPP technology to a transparent, plug compatible data warehouse appliance has taken place is the chief financial officer who suddenly is paying a lot less for technology and no one else knows there is any difference.

DATAUPIA

It is into this last class of data warehouse appliance that Dataupia fits. Dataupia is a data warehouse appliance operating under an MPP architecture and Dataupia has the potential to be a plug compatible, transparent storage infrastructure for MPP technology.

As such there is a very nice fit between DW 2.0 and Dataupia. Both are the most current ends of a long evolution. Both have surfaced in the world at approximately the same time. Both have the same goals – creating an environment for end user analytical freedom to operate and examine large amounts of information in an efficient and inexpensive manner. Both recognize the need for managing the life cycle of data in the data warehouse environment. Both recognize the need for being acutely aware of the costs of technology.

INDUSTRY LEADERSHIP

And finally, both Dataupia and DW 2.0 have taken a position of leadership in the industry. Often times positions of leadership are not popular and go against the then current conventional wisdom. But over time, when the vision is truly evolutionary the marketplace reinforces the wisdom of an evolutionary approach.

For more information –

DW 2.0 – ARCHITECTURE FOR THE NEXT GENERATION OF DATA WAREHOUSING, W H Inmon, Derek Strauss, and Genia Neushloss, ElSevier Press, spring of 2008

TAPPING INTO UNSTRUCTURED DATA, W H Inmon and Anthony Nesavich, Prentice Hall, Dec 2007

BUSINESS METADATA, W H Inmon, Bonnie O’neil and Lowell Fryman, ElSevier Press, Oct 2007

The web site – inmoncif.com – wherein a complete description of the DW 2.0 environment can be found.