Executive Information Systems, Inc.

Data Warehouses and Data Marts: A Dynamic View

By

Joseph M. Firestone, Ph.D.

White Paper No. Three

March 27, 1997

Patterns of Data Mart Development

In the beginning, there were only the *islands of information*: the operational data stores and legacy systems that needed enterprise-wide integration; and *the data warehouse*: the solution to the problem of integration of diverse and often redundant corporate information assets. Data marts were not a part of the vision. Soon though, it was clear that the vision was too sweeping. It is too difficult, too costly, too impolitic, and requires too long a development period, for many organizations to directly implement a data warehouse.

A data mart, on the other hand, is a decision support system incorporating a subset of the enterprise's data focused on specific functions or actvities of the enterprise. Data marts have specific business-related purposes such as measuring the impact of marketing promotions, or measuring and forecasting sales performance, or measuring the impact of new product introductions on company profits, or measuring and forecasting the performance of a new company division. Data Marts are specific business-related software applications.

Data marts may incorporate substantial data, even hundreds of gigabytes, but they contain much less data than would a data warehouse developed for the same company. Also since data marts are focused on relatively specific business purposes, system planning and requirements analysis are much more manageable processes, and consequently design, implementation, testing and installation are all much less costly than for data warehouses.

In brief, data marts can be delivered in a matter of months, and for hundreds of thousands, rather than millions of dollars. That defines them as within the range of divisional or departmental budgets, rather than as projects needing enterprise level funding. And that brings up politics or project justification. Data marts are easier to get through politically for at least three reasons. First, because they cost less, and often don't require digging into organization-level budgets, they are less likely to lead to interdepartmental conflicts. Second, because they are completed quickly, they can quickly produce models of success and corporate constituencies that will look favorably on data mart applications in general. Third, because they perform specific functions for a division or department that are part of that unit's generally recognized corporate or organizational responsibility, political justification of a data mart is relatively clean. After all, it is self-evident that managers should have the best decision support they can get provided costs are affordable for their business unit, and the technology appears up to the job. Perhaps for the first time in computing history those conditions may exist for DSS applications.

So, data marts have become a popular alternative to data warehouses. As this alternative has gained in popularity, however, at least three different patterns or informal models of data mart development have appeared. The first response to the call for data mart development was the view that data marts are best characterized as subsets (often somewhat or highly aggregated) of the data warehouse, sited on relatively inexpensive computing platforms that are closer to the user, and are periodically updated from the central data warehouse. In this view, the data warehouse is the parent of the data mart.

The second pattern of development denies the data warehouse its place of primacy and sees the data mart as independently derived from the islands of information that predate both data warehouses and data marts. The data mart uses data warehousing techniques of organization and tools. The data mart is structurally a data warehouse. It is just a smaller data warehouse with a specific business function. Moreover, its relation to the data warehouse turns the first pattern of development on its head. Here multiple data marts are parents to the data warehouse, which evolves from them organically.

The third pattern of development attempts to synthesize and remove the conflict inherent in the first two. Here data marts are seen as developing in parallel with the data warehouse. Both develop from islands of information, but data marts don't have to wait for the data warehouse to be implemented. It is enough that each data mart is guided by the enterprise data model developed for the data warehouse, and is developed in a manner consistent with this data model. Then the data marts can be finished quickly, and can be modified later when the enterprise data warehouse is finished.

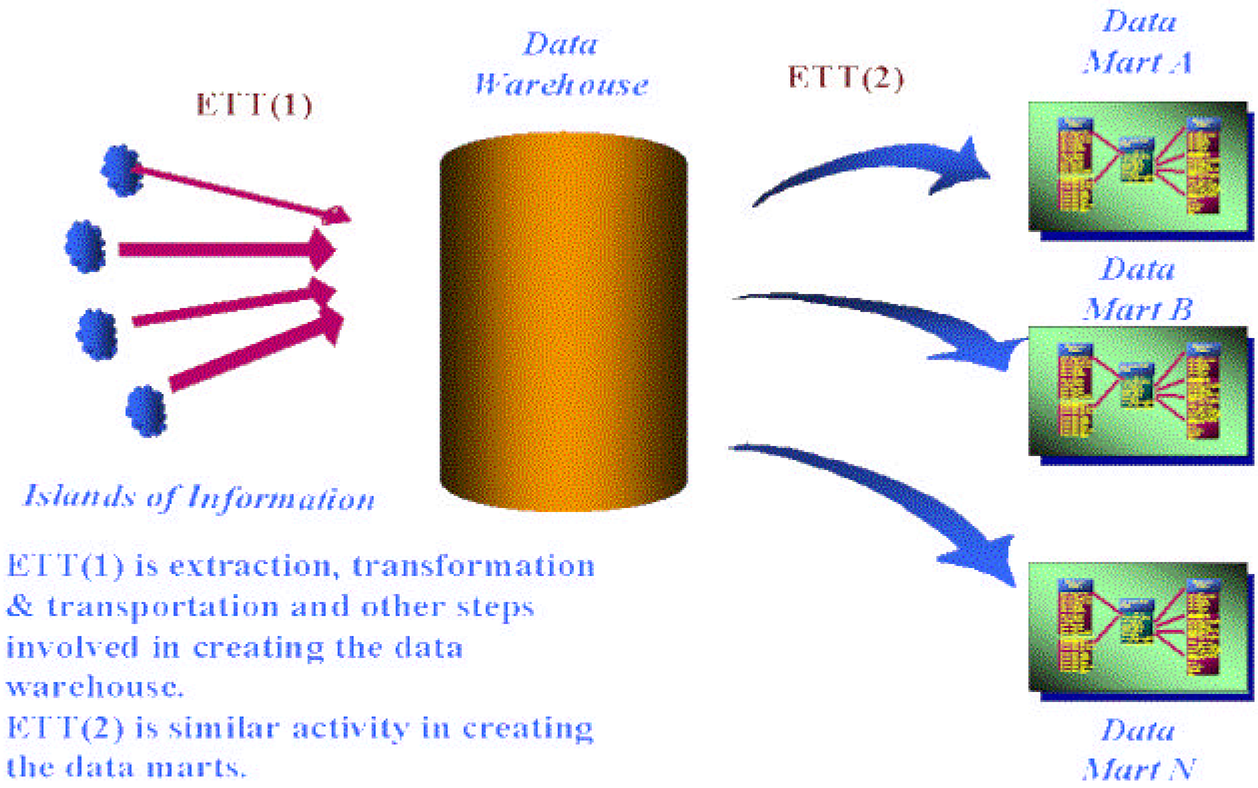
These three patterns of data mart development have in common a viewpoint that does not explicitly consider the role of user feedback in the development process. Each view assumes that the relationship between data warehouses and data marts is relatively static. The data mart is a subset of the data warehouse, or the data warehouse is an outgrowth of the data marts, or there is parallel development, with the data marts guided by the data warehouse data model, and ultimately superseded by the data warehouse, which provides a final answer to the islands of information problem. Whatever view is taken, the role of users in the dynamics of data warehouse/data mart relationship is not considered. These dynamics are the main subject of this white paper.

To develop this subject the original three models are first developed in a little more detail. This development is followed with a presentation of three alternative models that consider the role of feedback from users in the development of data warehouses and data marts. Lastly, an analysis of the usefulness of the six patterns of development is given in light of a particular viewpoint on organizational reality.

Development Models Without Explicit User Feedback

The Top Down Model

The top down model is given graphically in Figure One. The data warehouse is developed from the islands of information through application of the extraction, transformation and transportation (ETT) process. The data warehouse integrates all data in a common format and a common software environment. In theory all of an organization’s data resources are consolidated in the data warehouse construct. All data necessary for decision support are resident in the data warehouse. After the data warehouse is implemented, there is no further need for consolidation. It only remains to distribute the data to information consumers and to present it so that it does constitute information for them.



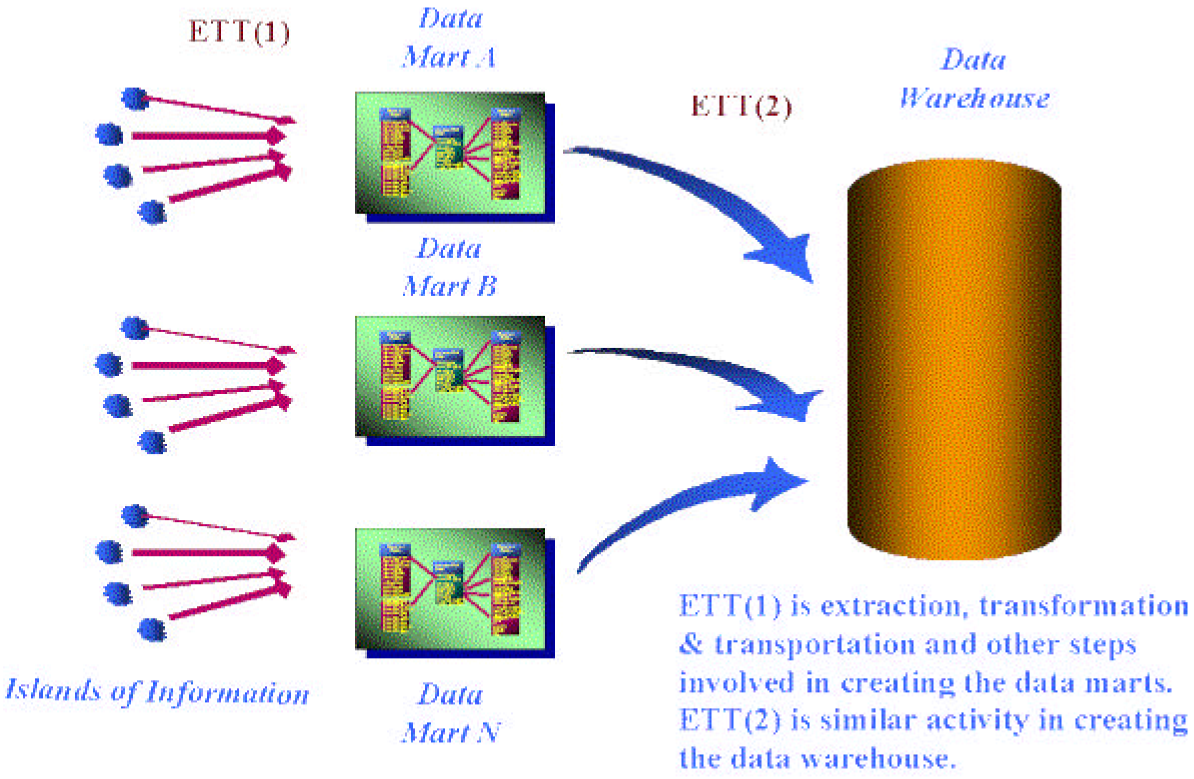
ETT(1) is extraction, transformation & transportation and other steps involved in creating the data warehouse.
ETT(2) is similar activity in creating the data marts.

Figure One -- The Top Down Flow from Data Warehouses to Data Marts

The role of the data marts is to present convenient subsets of the data warehouse to consumers having specific functional needs, to help with structuring of the data so that it becomes information, and to provide an interface to front-end reporting and analysis tools that, in turn, can provide the business intelligence that is the precursor to information. The relation of the data marts to the data warehouse is strictly one-way. The data marts are derived from the data warehouse. What they contain is limited to what the data warehouse contains. The need for information they fulfill is limited to what the data warehouse can fulfill. The data warehouse therefore is required to contain all the data that the enterprise or any part of it might need to perform decision support. And if users discover any need the data warehouse does not meet, the only way to fix the situation is for the users to get the enterprise level managers of the data warehouse to change the warehouse structure and to add or modify the data warehouse as necessary to meet user needs. The model contains no description or explanation of this process of recognition and fulfillment of changing user needs or requirements. But it is inconsistent with the model to assume that data marts would serve as a means of fulfilling changing user needs without changes to the data warehouse occurring first.

The Bottom Up Model

Figure Two depicts the The bottom-up pattern of development. In the left-hand portion of Figure Two, data marts are constructed from pre-existing islands of information, and the data warehouse from the data marts. In this model the data marts are independently designed and implemented, and therefore unrelated to one another, at least by design. Growth of this kind is likely to contain both redundancy and important information gaps from an enterprise point of view.



ETT(1) is extraction, transformation & transportation and other steps involved in creating the data marts. ETT(2) is similar activity in creating the data warehouse.

Figure Two -- The Bottom-up Flow from Data Marts to the Data Warehouse

While each data mart achieves an integration of islands of integration in the service of the data mart’s function, the integration exists only from the narrow point of view of the business function sponsoring the data mart. From the enterprise point of view, new legacy systems are created by such a process, and these constitute new islands of information. The only progress made is that the new islands employ updated technology. But they are no more integrated and coherent than the old islands were; and they are no more capable of supporting enterprisewide functions.

The right-hand side of Figure Two shows the data mart islands of information being used as the foundation of an integrated data warehouse. A second ETT process supports this integration. It will be needed to remove the redundancy in the data marts, to identify the gaps the process of isolative data mart creation will leave, and to integrate the old islands of information into the new data warehouse in order to fulfill these gaps. The possibility of using older islands of information in this way is not envisioned in this model, which tacitly and incorrectly assumes that the flow from data marts to data warehouse will be adequate to produce a data warehouse with comprehensive coverage of enterprise data needs.

The second model is vague on what happens after the data warehouse is built. Will the data warehouse suddenly become the parent of the data marts, and development proceed according to the top-down pattern? Or will the data warehouse continue to be the "child" of the data marts, which will continue to evolve and lead periodically to an adjustment in the enterprise data warehouse to make it consistent with the changed data marts? The second model doesn’t answer such questions, but instead ends its story with the creation of the data warehouse.

Parallel Development

The most popular pattern of development of the first three is the parallel development model. The parallel model sees the independence of the data marts as limited in two ways. First, the data marts must be guided during their development by a data warehouse data model expressing the enterprise point of view. This same data model will be used as the foundation for continuing development of the data warehouse, ensuring that the data marts and the data warehouse will be commensurable, and that information gaps and redundancies will be planned and cataloged as data mart construction goes forward. Data marts will have a good bit of independence during this process. In fact, as data marts evolve, lessons may be learned that will lead to changes in the enterprise data warehouse model. Changes that may benefit other data marts being created, as well as the data warehouse itself.

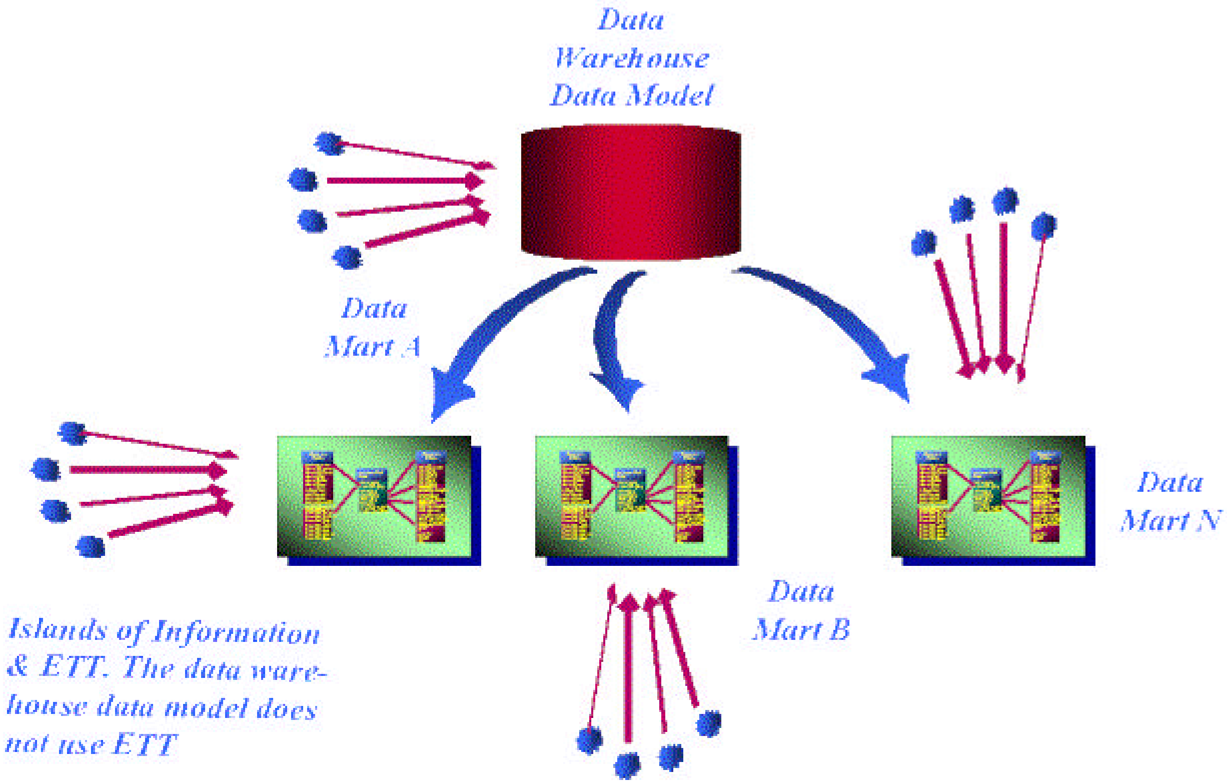

Data Warehouse Data Model
Data Mart A
Data Mart N
Data Mart B
Islands of Information & ETT. The data ware-house data model does not use ETT

Figure Three -- Data Mart Creation Guided By a Data Model of the Data Warehouse

Second, the independence of data marts is treated as a necessary and temporary expedient on the road to construction of a data warehouse. Once the goal is achieved, the warehouse will supersede the data marts, which will become true subsets of the fully integrated warehouse. From that point on, the data warehouse will feed established data marts, create subsets for new data marts, and, in general determine the course of data mart creation and evolution.

The third pattern begins to treat some of the complexities of the relationship between the data warehouse and data marts. Unlike the first pattern, it recognizes that organizational departments and divisions need decision support in the short-term and will not wait for data warehouse development projects to bear fruit. Thus data marts are necessary and desirable applications for organizations to pursue. Also unlike the first pattern, it sees the data marts as contributing to the data warehouse through evolution in the enterprise data model stimulated by the data marts.

Unlike the second pattern, the parallel view does not provide for uncontrolled growth in data marts. The contents of data marts are to be determined by the enterprise wide data model. Redundancies and information gaps are to be carefully tracked. The enterprise data model will track the activities and accomplishments of data mart projects and be adjusted accordingly. In the parallel development view, data mart activities will contribute to integration of islands of information within the data warehouse, by constituting islands of integrated information within the overall plan provided by the enterprise data model. These islands, in turn, will eventually enter the comprehensive integration of the data warehouse when it is completed.

The third view still retains difficulties. First, it hinges on the rapid development of the enterprise data warehouse model. Decision support consumers will not wait. Not when they have budgets and can support creation of data marts. Though waiting for a data model is a considerably shorter wait than waiting for a full-blown data warehouse, in large organizations the JAD sessions and requirements analyses preceding data model development can take many months. And the job must be done carefully. If the enterprise data model is to guide data mart development, it must be comprehensive in its coverage of data needs. Each time the enterprise model fails to identify a table or attribute necessary for a data mart, a little legitimacy is lost, and the feeling grows that it was not worth waiting for the enterprise data warehouse model, and that it will not be worth waiting for it to be adjusted.

Second, the parallel view, like the other two, also assumes that once the data warehouse is constructed, the data marts will become subsets of the data warehouse, rather than autonomous entities. Parallel development will end, and the data warehouse will fulfill everyone’s needs. This assumption is flawed, and envisions a degree of centralization of large enterprises that no longer exists.

The first three patterns of development all fail to explicitly consider continuous user feedback in response to data mart and data warehouse activities. In each view, user requirements are taken into account in constructing data marts or data warehouses, but user requirements are not static and tend to evolve on exposure to new applications and new technologies. Changes in requirements, further, are not limited only to faster hardware, or better techniques for data mining, or improved database software, or GUI interfaces. They may also encompass changes in information and in data requirements that necessitate adding new attributes and tables to data warehouses and data marts, and reorganizing old ones. New requirements may therefore impact data models at both the data mart and data warehouse levels. How will this be handled? Will all new requirements be processed through the data warehouse management? Or will local management implement most new requirements in local data marts first?

Whatever happens will be largely dependent on the nature and amount of feedback from users. The implications of user feedback for the three patterns of development produce three alternative patterns of data warehouse/data mart development. We now turn to these.

Development Models With Feedback

Top Down with Feedback

Suppose your organization is one of the pioneers that implemented a data warehouse before developing any data marts. Suppose the requirements analysis process was done carefully, and the enterprise data warehouse now contains all of the data and conceptual domains suggested or implied by that process. You are now given the assignment to develop an application to carefully measure the performance of your department during the last three years, and to forecast it three months into the future. What does the enterprise data warehouse have to contain to allow you to complete that assignment?

Certainly it has to contain indicators that will track the outcome of performance. Changes in sales, profits, and costs are obvious facts that need to be tracked. But what about *causal* variables, will they be among the attributes of the data warehouse?

The answer is some will be. But unless the effort of creating the data warehouse identified all of the domains within the database, and all of the attributes within those domains by referring to a comprehensive conceptual framework broad enough to encompass concepts and attributes contained in all of the causal models possibly relevant to your department's performance, it is a good bet that the data warehouse will not provide all of the attributes you need. The data warehouse, further, will not be constructed from the causal modeling point of view, unless you or some other representative of your department were considering a data mart at the time the requirements analysis was done for the data warehouse. There would have been no reason for the department's representative to either think in those terms, or to undergo the preparation necessary to think in those terms, in time for the data warehousing JAD sessions, or other requirements gathering tasks.

Your representative would almost certainly have specified essential facts to the data warehouse team, and obvious analytical hierarchies such as: company organization, geography, time, product hierarchies, and so on. But the full makeup of causal dimensions essential for measuring performance, distinguishing it from accident, separating it from either good or bad breaks is likely to be absent.

But you now have the assignment requiring at least some causal modeling, so what do you do? I think you get a subset of the data warehouse for a data mart. But data gathering does not end there. Either you gather data yourself if your department will support that, or you go to external services that sell data relevant for your problem. If you can do either of these two things, you will then supplement the data warehouse subset with the new data, go through a new, if limited, ETT process, and constitute a data mart that will work for your analysis problem.

The assignment your boss gave you has made your requirements change, and that has made the data mart change, and more specifically go beyond the bounds of the corporate data warehouse. You don't want to exceed these bounds, but if you don't, your departmental function suffers, and your job, and your boss's job depend on performing that function, not on maintaining the integrity of the enterprise data warehouse.

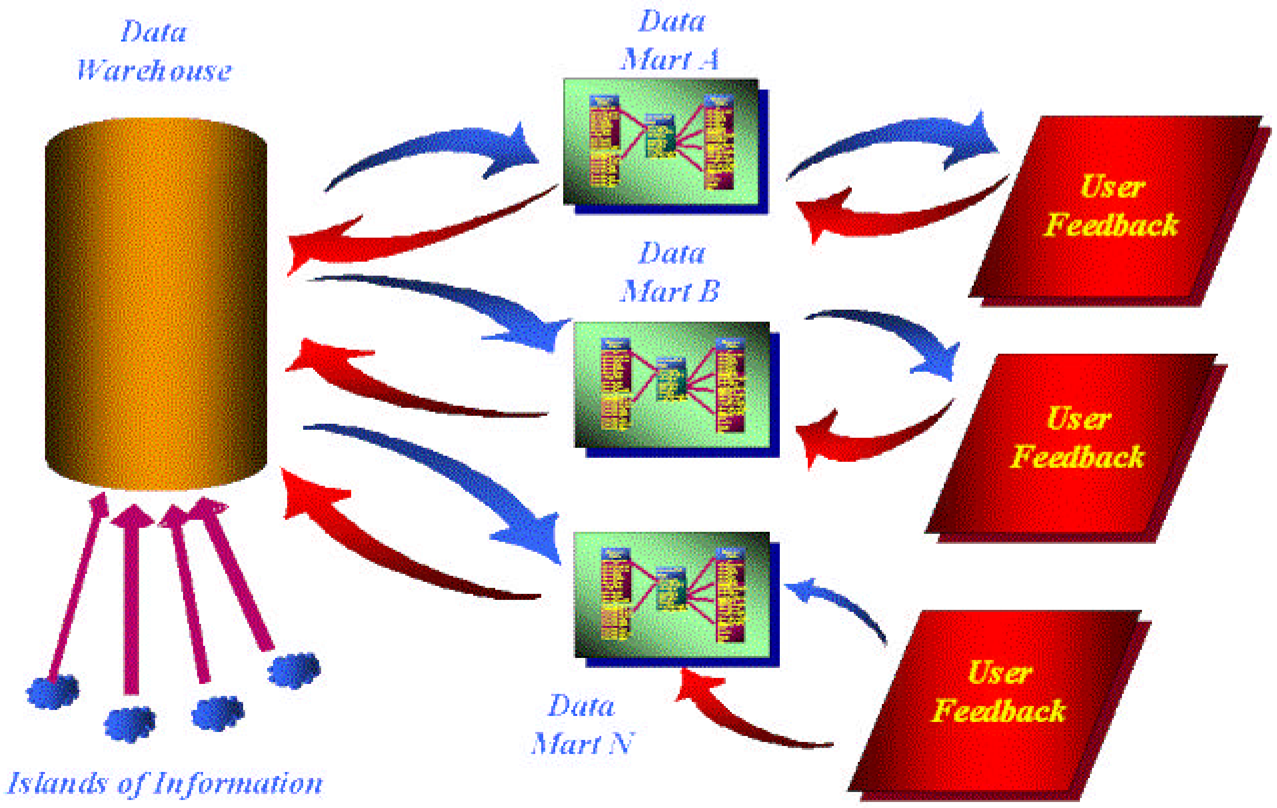


Figure Four -- The Top Down Model with End User Feedback

So this is the first stage of user feedback (See Figure Four). The second stage occurs when the changes made in your data mart are integrated with the enterprise data warehouse. This process can come sooner, or later. It will be sooner, if your company is wise enough to allow continuous feedback from departmental data marts to the data warehouse, and continuous integration of changes that seem necessary at the departmental level. Or alternatively, your company can refuse to recognize the changes being introduced into the data marts at the departmental level. If that’s the pattern, the changes could accumulate for years; until there is a new islands of information problem in the company. Then the changes will all come at once with both sides pointing fingers at the other for allowing the data warehouse to get so out of balance with reality.

Whichever pattern applies, the top down model will be subject to departmental user feedback, or adaptation to the top-down data warehouse by departmental data marts. If the continuous pattern of adjustment to departmental changes is adopted, a pattern of gradual evolution of the data warehouses and data marts will occur. The pattern will involve continuous feedback from the periphery to the center, and continual adjustment of both the periphery and the center to each other. The enterprise data warehouse will not bring a once and for-all decision support nirvana, but a much healthier process of continuous conflict and growth in business intelligence.

The Bottom-Up Model With Feedback

The three user feedback models are similar in the possibilities of adjustment to user feedback in the long run. Once the data warehouse is implemented, in each pattern there is the choice of building in a continuous adjustment process between the data warehouse and the data marts, or centralizing further DSS development in the data warehouse (migrating to the top down model). In the short run though, there is considerable difference between the three patterns.

In the top down pattern, user feedback before implementation of the data warehouse is through involvement in the system planning, requirements analysis, system design, prototyping, and system acceptance activities of the software development process. For reasons stated earlier, this involvement is likely to leave gaps in the coverage of domains and attributes that are causal in character, or for that matter that involve unanticipated side effects of departmental performance activities.

In the bottom-up pattern, in contrast, the effect of beginning development with data marts is to ensure much more complete coverage of causal and side effect dimensions. This also means that once the data warehouse is implemented, the bottom-up model with feedback will have little initial gap between user data mart requirements, and what is in the data warehouse. Paradoxically, this small gap could result in an enterprise level decision to migrate to the top down model for long-term development, once the data warehouse is in place. But if this danger is avoided, and the continuous adjustment path to development is followed, then the initial small gap between the data warehouse and data mart requirements will result in a much less painful adjustment process than will be experienced by organizations starting from the top down model. The future should be one of smooth continuous adjustments in the relationships between local data marts and the enterprisewide data warehouse.

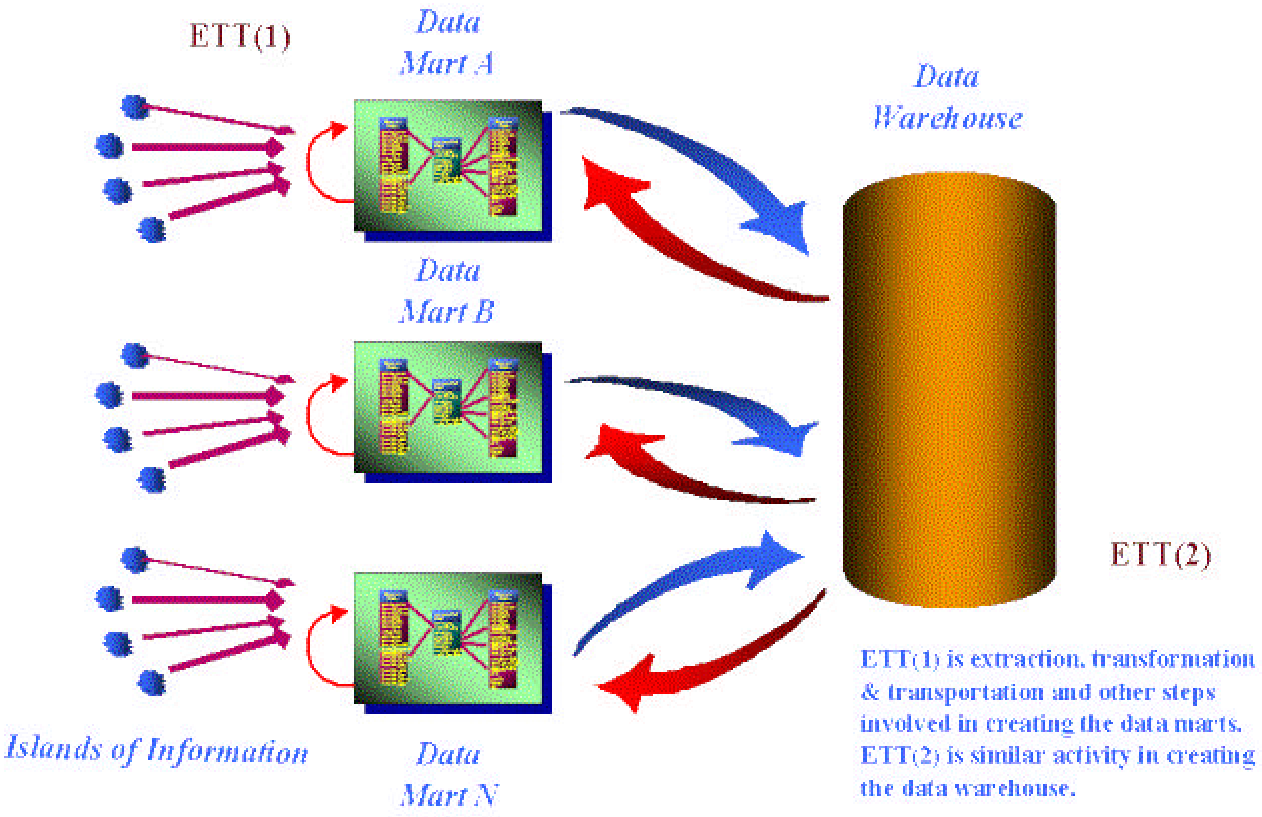

ETT(1)
Data Mart A
Data Warehouse
Data Mart B
ETT(2)
ETT(1) is extraction, transformation & transportation and other steps involved in creating the data marts. ETT(2) is similar activity in creating the data warehouse.
Islands of Information
Data Mart N

Figure Five -- The Bottom-up Flow from Data Marts to the Data Warehouse with Feedback

Don’t conclude though, that the bottom-up model with feedback (See Figure Five) is idyllic. It may not imply much pain once the data warehouse is in place, but if too many data marts are developed for too long in following the bottom-up model, the result is a set of new islands of information, and a painful process of handling redundancies and information gaps in data warehouse construction. So, if the top down model with user feedback means excessive pain in adjusting to data marts following construction of the data warehouse, the bottom-up model can mean excessive pain in integrating information and data from data marts during data warehouse construction.

The Parallel Model With Feedback

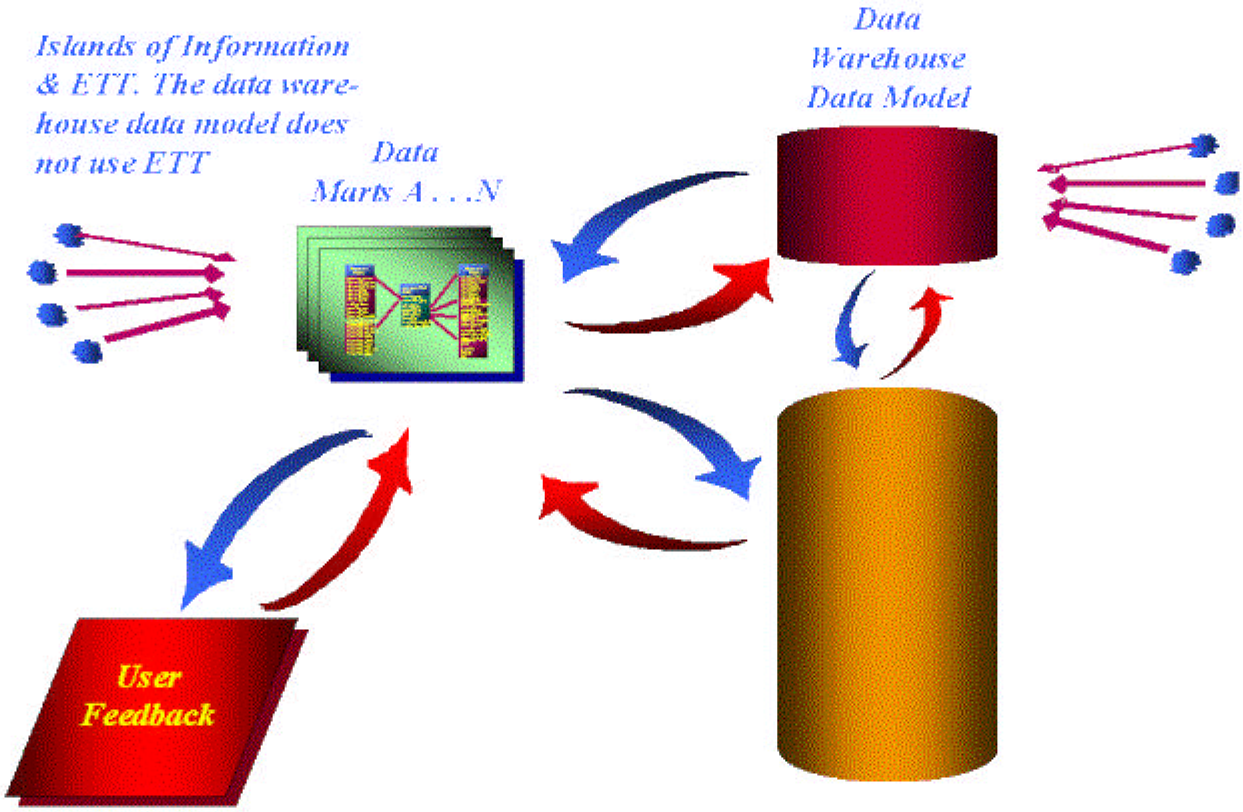

Islands of Information & ETT. The data warehouse data model does not use ETT
Data Marts A . . . N
Data Warehouse Data Model
User Feedback

Figure Six -- Data Mart Creation Guided By a Data Model of the Data Warehouse with Feedback and An Eventual Data Warehouse

Again, of the three alternative patterns, the parallel model (See Figure Six) offers the most promise. Development begins with a period of mutual adjustment between the enterprise data model and the data marts. As long as the center is open to data mart feedback and adjusts itself to the departmental perspectives on causal and side effect dimensions and attributes, the period of data warehouse development can be relatively smooth. While the data marts should be guided by the enterprise data warehouse model, in a very real sense, the enterprise level model should be guided by the individual and collective input from the data marts. Though the enterprise data warehouse data model is more than the aggregate of collected data mart models, it must certainly encompass those if it is to perform its long-term coordinative/integrative functions.

The danger implementing the parallel model is at the beginning of development. The model assumes completion of the data warehouse data model before data mart development begins, and therefore requires rapid development of the enterprise level model, and also requires the data marts to wait until this development is complete.

This assumption is not necessary for the parallel model. It is probably enough for the data warehouse data model to be in development at the same time as the first data marts, and for the data warehouse to adopt a coordinative and gentle guidance role in common efforts with data mart development staffs. A complete enterprise level data warehouse model is not necessary to monitor and evaluate interdepartmental redundancies, and to track information gaps. Nor is it necessary to coordinate data mart back-ends to ensure eventual compatibility. On the other hand, if data marts are coordinated by a central modeling team and encouraged to proceed with completion of their data marts with all deliberate speed, their results will inform the enterprise level team of what data warehouse requirements are much more effectively than the most carefully conducted JAD or requirements gathering sessions are likely to do.

The Dynamics of Data Mart Development

The three initial patterns of data mart development are unrealistic in their failure to take account of user feedback to data marts and data warehouses. By introducing explicit consideration of user feedback, one can see that the issue of centralized versus decentralized DSS development is one of long-term as well as short-term significance. All three patterns of development face the key decision of what to do once the data warehouse is developed. Will data marts then be handed down from on high, or will departments and divisions of enterprises have autonomy in evolving their data marts? It is clear that autonomy with central coordination is the most practical course for enterprises in the long run. But the three patterns of development are still distinct choices even if the same long-term policy of mutual adjustment of data marts and data warehouses is followed after data warehouse development.

The top down pattern will require a period of substantial adjustment to data mart needs after the data warehouse is constructed, to moderate centripetal forces and to adjust to the inevitable development of partly autonomous data marts. The bottom-up model will require an extra stage of significant ETT processing to accomodate development of the data warehouse from the data marts. The parallel development model will require rapid development of an enterprise level data warehouse data model unless it is moderated to require only simultaneous development of data marts and the data warehouse, along with coordination from the enterprise team. The parallel development model with feedback and less or no emphasis on a completed data warehouse data model prior to development, seems the indicated "rational" choice for a normative developmental pattern.

But the "rational" choice for development is frequently not a choice that organizations can make. So, an important question is, what will be the distribution of the different patterns of data mart/data warehouse development in organizations? First, none of the first three models will be represented. In neglecting user feedback, they ignore an essential empirical factor in the deverlopment process.

Second, of the alternative patterns, the top down pattern will apply to only a small percentage of enterprises, since it runs counter to the decentralizing forces pervading organizations today. The bottom-up pattern will be popular. Especially if it is supplemented with some coordination from an enterprise-level CIO sponsored data modeling group. Then the worst effects of uncoordinated bottom-up development would be avoided, and the eventual data warehouse would faithfully incorporate the requirements of the data marts.

Finally, the parallel model will also be popular, because it provides for both coordination and autonomy. It will be still more popular, if it is moderated to require coordination of developing data models, rather than guidance from a completed enterprise level data model. If the bottom-up development pattern is supplemented with coordination from an enterprise level data modeling group, and the parallel model is moderated to abandon the requirement that the enterprise-level data model be completed before beginning data mart development, then the distinction between these two models will blur, and real world cases will only have minor differences in the degree of central coordination of data marts they require. In the end we will see bottom-up and parallel models of data mart development merging, and the final pattern of development will be one of gradual evolution of data marts and data warehouses in a parallel process of mutual adjustment, change, and adaptation to the new problems facing organizations.

Biography

Joseph M. Firestone is an independent Information Technology consultant working in the areas of Decision Support (especially Data Marts and Data Mining), Business Process Reengineering and Database Marketing. He formulated and is developing the idea of Market Systems Reengineering (MSR). In addition, he is developing an integrated data mining approach incorporating a fair comparison methodology for evaluating data mining results. You can e-mail Joe at eisai@home.com.

[Up] [Data Warehouses and Data Marts: New Definitions and New Conceptions]
[Is Data Staging Relational: A Comment]
[DKMA and The Data Warehouse Bus Architecture]
[The Corporate Information Factory or the Corporate Knowledge Factory]
[Architectural Evolution in Data Warehousing]
[Dimensional Modeling and E-R Modeling in the Data Warehouse]
[Dimensional Object Modeling] [Evaluating OLAP Alternatives]
[Data Mining and KDD: A Shifting Mosaic]
[Data Warehouses and Data Marts: A Dynamic View]
[A Systems Approach to Dimensional Modeling in Data Marts]
[Object-Oriented Data Warehousing]