Realtime publishers

The Shortcut Guide™ To

Large Scale Data Warehousing and Advanced Analytics

Mark Scott

Introduction to Realtime Publishers

by Don Jones, Series Editor

For several years now, Realtime has produced dozens and dozens of high-quality books that just happen to be delivered in electronic format—at no cost to you, the reader. We've made this unique publishing model work through the generous support and cooperation of our sponsors, who agree to bear each book's production expenses for the benefit of our readers.

Although we've always offered our publications to you for free, don't think for a moment that quality is anything less than our top priority. My job is to make sure that our books are as good as—and in most cases better than—any printed book that would cost you $40 or more. Our electronic publishing model offers several advantages over printed books: You receive chapters literally as fast as our authors produce them (hence the "realtime" aspect of our model), and we can update chapters to reflect the latest changes in technology.

I want to point out that our books are by no means paid advertisements or white papers. We're an independent publishing company, and an important aspect of my job is to make sure that our authors are free to voice their expertise and opinions without reservation or restriction. We maintain complete editorial control of our publications, and I'm proud that we've produced so many quality books over the past years.

I want to extend an invitation to visit us at http://nexus.realtimepublishers.com, especially if you've received this publication from a friend or colleague. We have a wide variety of additional books on a range of topics, and you're sure to find something that's of interest to you—and it won't cost you a thing. We hope you'll continue to come to Realtime for your educational needs far into the future.

Until then, enjoy.

Don Jones

Copyright Statement

Chapter 1: Exploiting the Power of Large Enterprise Data Warehouses

For professionals that have been working with databases, data warehouses, and reporting applications, the dramatic and inevitable growth of the data that you must manage comes as no surprise. The amount of information that businesses store grows explosively from year to year. In 2007, more than 280 exabytes of data were generated. That's 10% more than anyone predicted. And enterprises are responsible for storing and securing 85% of that data (Source: http://www.computerworld.com.au/article/208773/idc_report_data_creation_outstrips_storage_first_time/). To make strategic use of that information is the challenge faced by information architects and database engineers in every segment of the economy.

As the amount of data becomes increasingly more difficult to manage, the processes by which we store and retrieve that data changes. And, more significantly, the information that we extract from this data has increasingly more impact on the organizations that we serve. This guide is targeted specifically to address the challenge of large data sets. For our purposes, that means data sets with more than 10TB of data. If you have dealt with these large databases, you know how long it can take to load the data, perform backups, query the data, and provide accurate answers in a timely manner to your organization's key decision makers. You may have a system that works fine at the moment, but you can see the trends and you know that you are nearing the limits of your current architecture. This guide can help you lay a foundation that will allow your data to continue to provide the vital insights into your organization that will help the business thrive.

The Rationale for Large Scale Data Warehouses

To address the issues of large scale data warehouses, it is best to start with the understanding of how they become large. The growth patterns and reasons for their growth will help define the best approaches for dealing with them.

Proliferation of Data Sources

Working with data warehouses, it never ceases to amaze me the number and diversity of systems that are used to provide data to the warehouse. As the data warehouse evolves, it needs data from more systems within the organization. As the organization grows, it adds new departments and new systems to support these departments: new order entry, manufacturing control systems, inventory monitoring systems, ERP systems, CRM systems—the list seems endless.

Take, for example, placing an order with a company. Once, it would have been a single point of entry (for example, a green screen terminal) into a single order entry system. Orders coming from different sources were likely contained on paper until they reached an order entry clerk who actually typed the order into the system. By today's standards, this approach seems ridiculous. Now the order comes in through a salesperson using his or her laptop. Or a telesales person who is sitting at a workstation while placing calls. Or through an Internet order placed directly by a customer. Or through an EDI document exchange with a business partner. All of these systems do the same thing, but none of them do it in the same way. And although the end result, an order, is (hopefully) placed in a single system that handles all order entry for the enterprise, each system has details that are significant to the patterns contained in the sales process. Thus, they each become an independent source of data to the data warehouse.

The effect is that the warehouse must handle a large number of data sources. This reality places interesting requirements on the ETL systems and the ability of the warehouse to add new data. Many of these systems have small maintenance windows during which the data is static and can be extracted. The maintenance windows occur when the systems are at their lowest usage. Most of these systems will be at their maintenance windows at the same time, creating a bottleneck for grabbing the data and getting it into the data warehouse. Although each source system may be able to divulge its treasure of knowledge in a short period of time, the data warehouse must absorb the data from all these sources quickly and transform the data to organize it for daily reporting. The ETL process and data loading thus becomes a key factor in the design.

Data Volume

The volume of data entering the data warehouse is another key component. Data warehouses are tackling much larger problems and larger data sets than ever before, working with telecommunication companies, credit card and payment processing organizations, financial institutions that analyze national and international stock transactions, and others; the volume of data is simply staggering. The growth of computer capacity has allowed systems to be developed to process terabytes of data in the course of hours. This functionality allows these large sets of data to be seen in their entirety and opens doors to new types of analyses.

I did work for a telecom that was receiving so much data on a daily basis that the data warehouse was taking more than a day to add the data to the system. It was a losing battle and required a redesign of the data structure and the process for adding data to the fact and dimension tables in order to keep up with the loading process. As the business grew, the shear amount of data grew and made the initial design inadequate.

> **Note**
>
> I have found that data sets increase in size geometrically rather than linearly. It is much more common to underestimate the capacity requirements for a data warehouse than overestimate them.

Thus, the second requirement of a large data warehouse is its ability to handle the growing amount of data it must absorb. A well-designed and implemented system will balance the cost (in hardware and software) of handling the load at hand and have a clearly articulated plan for expanding as the demand grows. This process will optimize the efficacy of the system and keep the costs of the system in control.

Users and Queries

The next challenge is one of users and queries. It centers on several key factors: the number of simultaneous users, the nature and diversity of the queries themselves, and the total volume of the data being queried.

Early in the development of data warehouses, the information within the warehouse was accessed only by a small number of analysts who would carefully query the data and prepare reports for the users. The reports were distributed as static information to the various interested parties. Technology, computing power, and the need for specific tactical data have increased; users expect more interactive access to the data stored in the data warehouse. Self-service business intelligence (BI) and user-driven queries demand that the data become more accessible. A wider range of users access the data, and their queries are more diverse. Different uses of the data put different demands on the data. The data is more difficult to optimize because the queries are more diverse.

> **Note**
>
> The growing demand for self service BI and the need to provide data to much larger numbers of users puts greater demands on the query engines of the data warehouse. System scaling can be difficult if the demands placed on the query engine cannot be accurately projected. Systems that can scale to meet these needs are critical.

While this increase in demand grows, the size of the data itself grows each day. The tables and indexes become larger. The time needed to sort through terabyte-size fact tables grows with each passing week. Although there are many techniques for optimizing the queries, the changing needs and complexity of the user community makes optimization for everyone's needs increasingly challenging. Many of us have seen queries that took 5 seconds to return when the system was initially loaded grow to 5 minutes, 10 minutes, or more. There is only so much cache, so many indexes, and so many resources to distribute within the system to meet the demand.

The next requirement for a large scale data warehouse is ability to scale. The system needs to handle a growing number of users, a wide diversity of queries, and the ability to scale up to tables that contain an ever-growing number of rows. Well-tuned database systems are typically disk I/O bound. Thus, systems that can expand across multiple disks subsystems while maintaining the integrity of the entire data set are well suited to meet this challenge. The system needs to scale out the bottlenecks of disk I/O, memory, processing power, and network bandwidth to meet the needs of the business.

Timely Analytics

Many of us have received calls from a credit card company asking whether or not we made a purchase. These calls protect us from illegitimate charges on our credit accounts and help protect us against identity theft. To accomplish this, the credit card company must analyze millions of transactions and find the buying patterns of their customers. They use these patterns to detect outlier transactions, those that fall outside the pattern of regular usage. To accomplish this, the analytic system must continually process and update mining models against the current data. The models locate these outliers. It takes a lot of processing power, and it must be run quickly in order to be effective.

Financial institutions track individual transactions within the markets. Millions of transactions are processed and used to populate models that predict the trends within the market. If a financial firm can ride the crest of these trends, they can make a great deal of money. If they cannot keep up with the deluge of information, they will help fund the firms who can.

> **Note**
>
> Although most data warehouse design is based on the use of historical data, it never ceases to amaze me how quickly business owners want information available. A system that can process data and make it available rapidly will help satisfy this demand.

Analytics require the aggregation of data. The aggregate data can be used to show trends over time. The detailed source data can be mined using advanced algorithms to predict trends or reveal correlations in data that provide profound insights in the activities of the organization. A large data warehouse needs to support these aggregations and data mining processes. Although many can occur cyclically (weekly or monthly), there is an increasing need to perform the analysis in real time, or near real time. Enterprises that can obtain this data more quickly have a distinct advantage over their competitors. Thus, large scale data warehouses often need the ability to scale the aggregation and mining of data. They need to be able to add the processing power to running mining algorithms and perform on-the-fly exploration of the data to give the organization this type of key insight in an expedient manner.

Data warehouses grow because the businesses that build them grow. They gain new data sources, the data volume increases, the number of users and diversity of queries expands, and the need for more advanced and timely analytics increases. Thus, a series of data marts that merge into an enterprise data warehouse or an enterprise data warehouse that continues to grow with increasing detail will get larger. As the size increases, the initial architecture often reaches its limits. The wisdom of allowing a data warehouse to outgrow its platform may then be brought into question. Perhaps allowing the data to grow into the terabytes is not the right answer. However, the only means by which an enterprise can clearly see the entire picture of its activities is to view them as a whole and understand how each composite function affects the others. Although it might be unwieldy to handle all this data, the benefit of having it available in a single, queryable format more that makes up for the effort.

Why Do Businesses Need Large Scale Data Warehouses?

Data warehouses and analytical applications exist to serve the business needs of the organizations that build them. The question then becomes, Is there a real need to let a data warehouse grow into the 10+ terabyte range? They are large, costly, and provide their own special technological challenges, so does the business really need large scale data warehouses? After all, no one is going to use all 10+ TB of data at once. What are the business drivers that lead an enterprise to build these warehouses? By understanding the drivers, we can gain further insight into the requirements for these systems and help define the type of system that will best meet the needs of the organization.

Understanding the Entire Organization

Every department in the organization has its own software to help it perform its work. By design and convenience, each element of the organization has silos of information that they use to do their jobs. This arrangement is often necessary because the applications that collect the information and serve the various departments need to be close to those departments and responsive to the specific needs they serve.

The problem is that some of the detailed information in one department is often useful to another department. The data warehouse becomes a common hub where that information can be stored and easily shared. The data warehouse can help abstract the needs for shared information from all the detailed information that is relevant only to the source system. This can help reduce the overwhelming amount of detail to a higher level that can be more easily shared throughout the organization.

The gathering of information from throughout the entire organization can help provide cross-departmental insight. The information that flows between elements within the enterprise can provide a vantage point that cannot be seen from within the limited view of a single department. I worked on a system that had distributed manufacturing with captive partners and independent vendors located throughout the world. Information on the cost to ship, the availability of manufacturing capacity, the location of raw materials, and the current contractual arrangements with the potential sources all needed to be considered before a decision could be made where to manufacture an order. No single system captured enough information to feed the parameters of the decision algorithm, but all the information required was stored in the data warehouse. Although ERP systems often contain much of this information, they frequently do not have everything that is necessary. A data warehouse that can contain all these cross-departmental data sources and keep a longer view of the data will provide unique insight into the operation. As the data in these warehouses grows, there is a need for a larger scale warehouse and analytical system to manage it.

> **Note**
>
> Gathering information from systems throughout the organization is often what leads to data warehouses growing in size. It can be difficult to predict the size from all the data sources that will prove useful (assuming you can even determine them all in advance). Finding a system that can scale out to meet the demand allows the system to grow to meet this need.

Most really large data warehouses, however, grow from really large data sets. Large organizations that handle millions of transactions per day simply have large data sets. As data collection systems have become more automated and sophisticated, they collect more useful information with each transaction. That creates larger and larger data sets. Combined with increases in the number of transactions, it makes the data grow. If the organization wants to continue its growth, it needs to take full advantage of all the information at its disposal. That means handling larger data sets. Consider, as an example, the advances in inventory tracking. Using RFID, automated tracking of inventory and products allows each movement of those products throughout the distribution chain. That tracking can help keep tabs on inventory and optimize its movement—if an enterprise has the capability of capturing and analyzing all the additional data related to that movement.

Sometimes, data grows quickly overnight. When two organizations merge, they have a great deal of data that needs to be combined to report on the new organization. Each of the constituent organizations may have a data warehouse. But most data warehouses will struggle to double overnight. The effort to normalize data and charts of accounts from multiple organizations often puts demands for extra metadata and requires additional ETL processing. The work required of the combined data warehouse is often greater than the sum of the work required from the individual data warehouses from which it is derived.

I helped work on a data warehouse for an organization that grew through merger and acquisition. The warehouse collected data from more than 50 distinct ERP systems. The enterprise was really a federation of businesses working across the globe. The only realistic means of seeing the organization as a whole was to aggregate the data into a central repository that could collate all the data from the various organizations into a single cohesive source. Although each individual system could meet the reporting needs of their individual organizations, only a large scale data warehouse could handle the aggregation of this data into a single entity and show a picture of the performance of the enterprise as a whole.

To rationalize information from multiple systems, be they departmental or divisional, the data needs to be collated and organized. Even if it is later reduced into smaller data marts that are focused to direct specific questions, putting all the data in one place at one time allows the entire picture to be painted on a single canvas. For many organizations, that is a large amount of data. And that requires a system that can handle that volume of data and grow with the business as it grows.

> **Note**
>
> Creating unified semantics for the data that can be disseminated throughout the enterprise is one of the most challenging hurdles for making use of information. Fragmenting the data across disjointed data marts can make it even more difficult. Starting with a single coherent data source and then extracting smaller units of data makes this much easier and more consistent.

Enterprise Data Correlation

For the business analysts to get the most value from the data warehouse, they must be able to reliably and effectively query across data throughout the data warehouse. Conformed dimensions and clearly defined relationships between tables within the database allow the enterprise to relate the activities of its divisions and departments together and determine how the constituent entities interact as a whole.

When correlating data from multiple systems, there is always the problem of correlating the data in one system to the data in another system. The CRM system has a distinct method of identifying customers. The accounts payable system uses another identifier. The shipping system requires its data to be correlated with the order entry system before it can be matched to the account payable system.

This has two effects on the data warehouse. The data warehouse must contain all the data required to build the chain from one system to another. The CRM, Accounts Payable, and Shipping system tables need to be contained within the data warehouse. There are often mapping tables required to bridge the gaps and connect data from one system to another. The data warehouse will also require the processing bandwidth to correlate the data and process it into the requisite fact tables and dimension tables.

When working with source system data, I have often found that it is inaccurate or incomplete. When correlating data across multiple systems, the key fields that correlate data from one system to another must be present. This leads to the need for data cleansing. Data cleansing will require processing power to accomplish and metadata used to determine how to correct errors in the data and to fill in the gaps. Data mining may be used to choose likely substitutes for missing data. Additional lookups might use fuzzy logic to isolate good data. A variety of other techniques, all of which require additional space and computing power, can be used to make the data as usable as possible. Data de-duplication also becomes an issue. Data duplication can occur because of simple misalignments in data that result in erroneous joins in master data that allow a single transaction to be represented as multiple transactions within the framework of systems. Although removing the duplicate rows will ultimately reduce the total size of the data, it requires the data warehouse system to have the capacity to temporarily manage the duplicate rows and the processing power to find and delete the rows. The additional sorting and filtering can place strong demands on the warehouse system.

> **Note**
>
> Many source systems do not maintain clean internal data. The key objective in most source systems is handling individual transactions, so they may not work as hard at completeness—particularly when the information is not strictly required to complete the transaction. Having the data warehouse cleanse and purify the data will speed report development and help provide better information to the business users.

Many organizations obtain external information. Whether it is common market trends, lists of leads from advertisers or other data services, data dumps from other organizations, payment transaction records, Web server access logs, and so on, it needs to be stored. Some of these external feeds are large. They need be carefully archived because they not are available later. That often means storing them in their original form and making copies of the data as it is blended into the data warehouse. The extra storage and copies all add size and require bandwidth.

Once the data is clean, correlated, and presented, the business has a comprehensive source for data. They can compare activities in sales with activities and purchasing and manufacturing. This provides the insight that helps the organization view itself as a cohesive whole.

A Single Source for the Truth

One of the most interesting dilemmas encountered in a business is when two reporting systems provide different answers for the same data. Source systems often exchange data and then modify or transform the data. One of the most difficult tasks in any reporting system is to keep de-normalized data consistent and synchronized.

A central data warehouse is often used to identify the authoritative source of data for a given data element. If multiple source systems have a copy of the data, a single representation of the data stored in the warehouse can serve as the arbitrator of the truth. Another related issue is overloaded terms. For instance, there can be three systems that each receives an order, one for the order entry system, a second for a partner vendor, and the third for the accounts payable system. Each could have a date entered field, each field could be distinct, and each data point important. The user community needs to be able to locate the individual data point that they need so that the data is consistent across reports. Most organizations use multiple reporting systems that make different uses of the same data. Some seek detailed, transactional level data, while other systems aggregate the information. Drawing from a single source helps keep the data consistent across the systems and reduces confusion and reporting errors.

Having a centralized source for information also helps simplify the process of creating new reports. The data is located in one place and correlation of data is already completed, so it is much simpler to build the report. The efficiency of locating the reporting data in a designated repository makes the completion of new reporting projects much faster and more economical.

Of course, the more reporting done from this data source, the more demand is put on the service. The need to cache data and respond quickly to queries that vary makes tuning a different proposition than it is on smaller siloed systems that contain a less comprehensive breadth of data. The efficiencies of providing a single enterprise data source can also create demand for one that is tuned for a wider gamut of data access.

> **Note**
>
> Although having a single source for the truth is obviously beneficial, it can be difficult to achieve. People become territorial about their information. Providing a central warehouse for information that is consistent and performant can help people surrender "their" data—in exchange for a broader-based view of the organization as a whole and insight into interactions with other departments.

The demand for these reports can also drive how quickly the system needs to make reporting information available. The moving of data from the source systems to the data warehouse, and the ETL and analytical processing required to make the data available, will help define the capacity and performance requirements for the data warehouse.

Centralizing enterprise data can make it much faster and simpler to find data. It helps make the development of new reports simpler and more economical. It can improve communication and reduce errors and confusion.

The business receives many advantages from a single source of data that is consistent and readily available. As the business grows larger and more diverse, the system will inherently become larger and grow into the category of a large data warehouse. Although the business may benefit from a large, single source of enterprise information, should IT consider alternatives that might overcome some of the technical challenges to maintain a large, single source of business critical information? Is it practical to keep all this information in a single system?

Why Should IT Consider Large Data Warehouses?

We have established why having a single, large data warehouse with powerful processing capabilities can help the business deal with large data sets and build a single, comprehensive view of the entire enterprise. From an IT perspective, building such a large, high-power system can be complex and costly. But there are distinct advantages to building a large data warehouse.

Addressing the Real Needs of the Business

Businesses want accurate information about the activities of the organization. They want it immediately, and they often do not consider what it takes to produce this information. From an IT perspective, the details of providing such an environment are where the devil lives.

Most high-level analysis of data is done at an aggregate level. A large retailer may not care that Jane Doe purchased a box of paper clips in a store in Buzzard's Breath, Wyoming. They often care that the sale of paper clips in the Western district is increasing. More likely, they care that office supply sales are starting to increase. To build the tens of thousands of individual transactions into the aggregates that help purchasing realize that they need to increase purchases of office supplies before they run out and customers find somewhere else to buy them, the system needs a robust method of building analytical structures.

Note

Sometimes the retailer *does* care that Jane Doe bought a box of paper clips. Data mining can help use this information to classify Jane, and send her targeted ads to increase her interest in office products. The data warehouse often needs to support both scenarios.

I have been working for the past decade on building these types of analytical structures. Regardless of the technology or product used, it is a challenge to build systems that can take millions of rows of data and convert them into the analytical structures that provide this information. When OLAP cubes are being built, they place a strong demand on the database to provide information rapidly. They work best when the data is stored in an analytic structure, such as a star schema. This often leads to multiple copies of the data: a normalized operational data store and the reporting star schema. This, in turn, leads to more data storage requirements and more need for data warehouse processing power.

The business needs access from all its various data systems. The data should be conformed so that information from HR can be connected to information in accounting to information in shipping to information in quality control. The information needs to be detailed enough to point to specific incidents, but then aggregated to provide insight from a higher point of view. The data needs to be collected over time so that trends can be measured.

The data warehouse provides that common point, where detailed and aggregate data can be stored and provided for every part of the organization in a single, easily accessed location. This can help business analysts quickly and cost effectively develop new reports and visualizations of the information. In addition, it can help key decision makers develop the correct responses to changes within the organization, taking advantage of successes and minimizing missteps.

To support this type of vision, with pervasive reporting and analysis of the entire enterprise, the data warehouse must have the capacity to grow to meet this need. For successful organizations, this will soon mean dealing with terabytes of information. Planning a data warehouse architecture that can scale to meet this need will allow the information systems to keep pace with the expansion of the enterprise.

Developing from a Single Source

Many of the products used to build OLAP cubes can draw from multiple data sources. It is not uncommon for a system to draw from multiple data marts or smaller data warehouses to obtain the data they require. However, systems inevitably work best when the data is located in a single source. Data located on separate systems is seldom well synchronized, requiring additional ETL to affect the processing. Drawing data from multiple systems is typically slower than from a single source.

Note

This situation seems counter-intuitive because using multiple source systems would seem to parallelize the operations across multiple servers to improve performance. In practice, such is very rarely the case.

The nature of analytics is that someone will inevitably want to drill through the aggregates and into the base data. They will want to use the aggregates of interest to isolate a specific subset of transactions and then display (or print or extract to Excel) that subset. Without question, if the aggregate data and the detailed data are stored in the same system, this isolation is a simpler proposition.

A single source of data that stores both aggregate data and detailed transactional data will be the easiest to maintain and operate. That means that building large data warehouses that can store and process those large amounts of data, and that can easily scale with the growth of the data, provide IT with a very potent analytic resource.

I frequently need to scour an enterprise to locate the data that I need. When business requirements are cast, the business users think in terms of the answers that they require—not the constraints on where that data is located. A report based on a single source of data is much easier to define and produce.

Placing all the data one needs in one place clearly makes it easier to find and use. But it also helps align data in terms of grain. Different source systems have different methods of capturing data over time. Single transactions occur in a specific point in time, such as the moment a product leaves the production line or a shipment leaves the loading dock. But for many significant business events, the process that needs to be tracked takes place over a period of time. The entire sales cycles may take weeks, months, even years. The steps in the cycle are captured in distinct systems—CRM, order entry, shipping, manufacturing, payment processing, and so on. The tricky bit is to get these systems to track from beginning to end so that the activities in the cycle can be related to one another.

The question I am beginning to raise is one of grain. Grain is a broad topic that deals with how much of one quantifiable business event (for example, a sale) should be measured against the factors that drove it (for example, advertising, sales activity, manufacturing, shipping, inventory, and so on). If grain can be rationalized and captured within the structure of the database, it provides a very rich depth of relationships between the data sources within the organization.

Building these bridges is often trickier than it sounds at first. The accounting system provides its data once a month when the books close. The manufacturing systems provide information at the end of each day. The HR system provides data on a bi-weekly basis. It is challenging to integrate the information in these systems to allow their output to work together harmoniously.

> **Note**
>
> Although getting the grain in a data warehouse is difficult, it is mainly difficult in the design phase. Once implemented, it provides a great deal of benefit to those who maintain and enhance the reporting systems moving forward.

Diversity in data from all these source systems presents a significant challenge to the information architecture team. But it is dealing with this diversity of source data and fusing it into a cohesive whole that is the basis of the most dramatic business insights.

Building this broad landscape of data sources is greatly simplified when there is a single location where the data is combined and rationalized. Integrating new data sources into the single corporate warehouse mandates that the grain issues within the data be addressed so that the data can be added. And although it might be difficult to add the data initially, once it is done, the data can be used across the board to divine business insight.

The data warehouse system then must be able to scale to continually add new source data. It will need the processing power to transform and integrate the data into the conformed enterprise data schema. It will also need headroom to manage the metadata (data source, date extracted, transformations, and so on) that inevitably accompanies this wealth of data sources. Ideally, the system can be sized to meet immediate needs and grow incrementally to accommodate growth within the business.

Operational Considerations

Many of the organizations with which I have dealt start with a departmental data mart. They plan ahead and use conformed dimensions so that additional departmental data marts can be built. They plan a database infrastructure that seems ridiculously too large for their immediate need, foreseeing that the demand on the database will grow and the data mart will evolve into an enterprise data warehouse.

Two things occur during the course of this evolution that disrupt this simple approach. One is that someone in another department starts a similar project. They build their own data warehouse and it starts to grow. Now two (or three or four) systems begin to grow across the enterprise. The second is that most organizations do not accurately estimate how fast the system will grow. They start with an infrastructure that seems ridiculously too large (knowing how difficult it will be to re-scale later) and outgrow it much faster than they ever imagined that they could.

Eventually, these systems need to merge to form a single version of the truth. Political and territorial issues aside, the platform on which they merge must be able to handle the load. Many organizations start with a platform and then they are bound to it. The cost to move to another platform is prohibitive, even though the platform on which they are operating cannot handle the challenges of the workload. Re-platforming the data warehouse is a vast undertaking and should be done as infrequently as possible. Thus, understanding the scaling capabilities of the platform up front will help determine the right platform on which these data warehouses should be merged (or built in the first place).

> **Note**
>
> My experience has shown that data warehouses grow much faster than anyone predicts. It is a difficult decision to build for a point 3 years from now when the data warehouse will host 15TB of data when today it barely hosts 1TB. However, it is often less difficult than trying to re-platform once the data *is* 15TB and everyone is complaining about the poor performance.

The cost in terms of licensing, support staff, and effort to maintain multiple data warehouses can be high. It is also interesting to note that these data warehouses are often maintained on competing platforms, so some of the cost advantages of purchasing more software licenses and similar hardware from a single vendor are lost. If different platforms are used, then the IT staff must maintain a wider diversity of skills, which can lead to higher staffing costs. Also, the database systems that are economical for hosting data marts will often not scale to the large sizes of enterprise-wide data warehouses.

This is not a critique of using data marts to grow into an enterprise data warehouse. It is a quite effective technique that has a great deal of merit and can pay for itself as it grows. Rather, the point is that the selection of the platform on which the data marts are built should be considered carefully. Can the platform absorb multiple data marts? Can it grow to meet demand over time?

This situation creates somewhat of a distinctive requirement. The data warehouse platform needs to be able to scale out to meet growing demand. Finding a system that can provide a high level of performance at the onset of the project and scale out to maintain those levels of performance as the workloads continue to evolve and expand becomes the preferred system.

IT Governance

The need to maintain reporting systems in a consistent and controlled manner has caused many organizations, from governments on down, to implement regulations and strict procedures on how major IT systems are changed and maintained. The reaction is quite understandable (if sometimes poorly implemented). Thus, IT needs to keep a close watch on all of their systems, but an even closer eye on the reporting systems that supply key information to the key decision makers, regulators, and stockholders.

When reporting systems are scattered throughout the organization, the task of keeping them well audited and regulated can be stressful. The systems often keep de-normalized versions of the same data. The de-normalized data does not always agree across systems. An authoritative source of data needs to be defined. To get the reports to harmonize, they should all draw their information from the single source of truth. As the source systems are inevitably heterogeneous systems, a common platform for this authoritative data, the data warehouse, is often the most practical solution.

Having a single data warehouse makes the control and auditing of this reporting information a much simpler task. Monitoring the activities and changes in a single system is always preferable to monitoring multiple systems. Change in a single system, the data warehouse, can easily control the source of authoritative data. Changes in the source are promulgated to all the key reporting systems that use that data without a wide variety of changes scattered throughout the infrastructure. Tracing data lineage and auditing data collection becomes much simpler in a single data warehouse scenario.

The data warehouse then will need to house a number of additional structures. It needs to supply space for maintaining metadata for the data imports. It needs to provide auditing for changes made to the system. It may also need to audit data access. This will increase the demands on the system. But a single system that provides this information will be much easier to manage. It will be easier to keep in regulatory and corporate compliance than several independent systems. Just the reduction in effort in tracking changes within the information infrastructure on a single system versus several distinct, independent systems can represent a significant savings in time, cost, and stress.

Data Security and Availability

Most people have witnessed the negative effect that data breaches have on a corporation. There are two issues in particular that are important to address in securing data in a warehouse. The data warehouse must be able to control who can access what data within the data warehouse (data entitlement). The other issue is providing business continuity and disaster recovery (data availability).

In terms of data entitlement, the system needs to grant access to some users while denying access to others. With a large scale data warehouse, there is a wider range of data to protect and typically a larger number of distinct user groups. The system needs to work from end to end—from the staging tables through the operational data stores and analytical and reporting structures.

There is also the question of sensitive data, such as personally identifiable information (PII), HR data, and corporate-sensitive information. This data needs to be more rigorously defended and carefully audited than other types of information and will be mixed into the data structures of the warehouse. It may require encryption to adequately protect.

It may seem dangerous to place all this tempting data in one location where a malefactor may gather it all in one fell swoop. However, security is a volatile, moving target. People come and go. They change job responsibilities. Keeping security set is always a challenge. Setting it on multiple systems is much more challenging.

> **Note**
>
> If it is difficult to secure one source of information, it is much more difficult to secure multiple sources for information. For instance, a user may be able to see information from the labor reporting system but not the hourly wage of those employees from the HR system. Nonetheless, these are both staged in the data warehouse so that labor costs can be analyzed.

A single data warehouse system that can manage security throughout the chain of data custody provides the simplest, most easily managed form of data security. A single security system is more easily maintained and kept updated. A large data warehouse that houses the single version of the truth can be configured to be one of the most secure. PII can be encrypted or hashed so that the privacy of individuals can be protected. Data can also be aggregated so that information concerning any given individual is not visible, yet the information can still be used in analysis.

Data availability becomes critical because, once the data warehouse becomes the primary source of reporting data for the enterprise; the enterprise cannot afford to do without it. This will impact several areas. This first affects backup and restoration of data. Large data warehouse systems need backup processes that can operate while maintaining performance for the users. Although a large single system is more challenging to back up than a group of smaller systems, it is much easier to keep the backup data synchronized. It is simpler to manage the backup operation and ensure the data is protected.

> **Note**
>
> Performing backups and restorations of large data sets can be difficult in the narrow time windows typically available. When designing a large data warehouse system, the time required for these operations must be carefully considered and addressed.

The second is availability. A single, cohesive system can become a single point of failure. In most cases, however, a system can be designed that eliminates single points of failure. It is a consideration that must be given careful thought. First, the data warehouse system should protect itself from internal failure. Second, if remote location disaster recovery is appropriate (and for data warehouses of this type, it typically is), the system must allow for operation at a remote site. Synchronizing data with a remote site is a major issue that is exacerbated in large data warehouses by the sheer volume of data. To support a large data warehouse, these issues need to be properly engineered.

The advantage to a single system is once again the ease of management. Keeping multiple systems available 24 × 7 is much more burdensome than a single, well-designed system that has clear processes and procedures for maintaining its availability. From an IT standpoint, the burden of maintaining a single larger system is typically lighter that that of maintaining several smaller systems.

Analytic Advantages of Large Data Warehouses

Analyzing data, from aggregation to data mining, provides some of the most profound insights into the business. Analytics can be used to detect trends and help forecast upcoming events. They can be used to fill in the missing gaps in information. They can help identify causal relationships and provide the organization's key decision makers with the best data on which to make decisions.

Thus, the data warehouse should provide more that a place to store data. It should help organize and build data structures that extract and refine this information. Large data warehouses can provide a strong foundation for this type of analytical data.

Full Range of Business Data

Finding key trends and information within a data set requires a full range of data. The more thorough and detailed the data is, the more accurate the analytics can be. The implications for this are that keeping more data online and available to the system provides a richer source for the analytic system. It also means keeping more sources of conformed data can provide the analytics a broader reach across the organization.

I have worked with many systems where detail data is archived. With higher volume data, this is a common practice. Once the data is aggregated, there may not be a need to keep the full range of detailed data. Often, the details are removed and higher-level aggregations are kept because the system simply cannot handle the volume of data. However, if data mining is required, the detail must be retained and available to build and train the mining models. I have worked with systems that when the detailed data from the past year or 2 years ago is required, it needs to be laboriously restored from archives. Or the detailed data from 2 years ago is simply not available. The ability of the system to handle large volumes of data can dictate the type of analyses that can be performed.

> **Note**
>
> Consider a system designed for a retailer that needed to forecast trends for certain products that are seasonal and linked to specific promotional activities. Although the sales data for these products was available, it was aggregated at the departmental level. In order for the retailer to provide the specific reporting they desired to plan their sales, they needed to restore detailed information for the last three years. The system did not keep the information online because they were running out of room. Although this system archived the detailed data, many other systems do not, and that would have made the analysis impossible. The design did not fully meet the need of the organization because the limits of the physical architecture drove their decision-making processes.

Enterprise data can often be affected by external factors. External events like stock market trends, weather, and commodity market shifts can have a major affect on the data internal to the business. If the data warehouse has the capacity to add this data, both in terms of storage and processing power to incorporate and then utilize the data, then it can add real business value. Conversely, if the system is already running near capacity, adding such data becomes unrealistic.

The ability to store large amounts of data is only part of the issue. Large data warehouse systems that support analytics need to provide enough processing power to aggregate data and build mining models while still remaining responsive to users. Thus, the scale of these systems to perform concurrent operations such as loading data while querying the data securely is an important consideration. Yet the systems that can meet these challenges provide the greater value to the business.

Many organizations require a rich mix of aggregate data and detailed, transaction-level data. Aggregates often identify a subset of transactions that exemplify the desired behavior, or those that indicate trouble within an organization. A system that can drill down into the details and help the analysts find the successes or the missteps will help the enterprise make better informed decisions.

This again helps define the requirements of the large data warehouse. It needs not only to handle large data volumes but also must have the capacity to load it quickly and remain responsive to users. It must support detailed transactional data structure, aggregate structures, and data mining structures that help analysts gain key insights.

Common Data Dictionaries and Information Schemas

Having a large amount of data is not the key to a successful data warehouse. The data needs to be understood by the people who use it. Each source system will have its own designations for various data elements. An order entry date may be called the sale date on a system, the ED on another system, and just Date on a third system. The people who use the source systems will think of the data element in the terms of the system they use most. A common reporting system needs to build the bridge across all these departments and systems and help the users enterprise-wide to gain a common understanding.

Data warehouses in general serve this purpose quite well. They allow all the data within the organization to be generalized into a canonical schema. Different inputs from different systems can be used to build this schema. As long as the schema is well thought out and conformed across the data within the organization, it will serve as the central repository and help people report.

Although it needs to be consistent, the information also needs to be flexible. Businesses change and that means that the information they use to monitor and operate themselves changes. An adaptable schema that can be extended to include new business processes, expansions, mergers, and acquisitions will help the enterprise maintain its view of itself. That requires a system that is adaptable and can grow as the business grows. IT needs enough spare capacity to add new sources of data and enough processing power to update changes to the internal information schema without becoming overloaded.

> **Note**
>
> I have worked with analysts who have tried to hunt down data to create reports. Because the data was not well defined, and the analysts did not fully understand the source systems from which the data was drawn, the reports were fatally flawed. It led to bad decisions and cost tens of thousands of dollars. A single source of data that is well documented and understood by the analysts can prevent these types of errors.

Providing the data in a single place will help analysts make use of the information. With reliable information that is clearly defined, the analysts can reach across the organization to find important relationships within internal operations. The system can grow to include external data sources. This allows the analysts better opportunity to find the leading indicators and trends that will help make better decisions.

The canonical information schema for the entire enterprise will help rationalize and relate data from all the systems across the organization. Building on a platform that can scale to meet the growth of demand to house the schema will allow the business to grow and adapt to changes in its markets and internal operations.

Development of New Reports

I have found that when reports are required, the business states what they want. They seldom tell you where to find it. They do not care what you need to go through to gather the raw data or what it will take to process that information into a format that they can use. They just want the report.

An outgrowth of the previous point is that new reports can be developed much more quickly from a single source of data. Searching through the corporate source systems to locate the required data elements for a report is time consuming and tedious. And if a central reporting repository is not established, it will be repeated for the next report.

That is one of the basic reasons to build a data warehouse. This becomes more important with the growing trend for self-service BI. Users have become much more sophisticated in their use of reporting technology. Report-building tools and spreadsheets have added significant capacity to connect to databases, query for data sets, and visualize that data for the user. Letting the user who wants to use the information build his or her own reports can save time and money and enable users to publish information more quickly.

Whether the reports are developed by the development team or business users, the process is much faster from a central data source. Having the data defined, organized, and processed makes producing the reports much faster. The ETL processes that cleanse and de-duplicate the data make it more complete and reliable. The additional analytical structures provide richer insights into the operations of the enterprise.

Pulling all this data together often requires a large system. The system needs to multi-task, adding new data, building new analytic structures, and responding to user queries.

Cost Considerations for Large Scale Data Warehouses

Most people know that in IT terms, bigger means more expensive. As systems get larger, the costs tend to increase dramatically. This may cause many organizations to resist the establishment of a large data warehouse system. Although some of these generalizations are true, there are some offsetting factors that need to be considered when determining the total cost of the data warehousing system

Building on a Single Cohesive Platform

Many large organizations build several (often competing) data warehouses. They are frequently built on different platforms and use different software. This has a number of implications on the entire price of data warehousing.

I have worked with many organizations that keep smaller, distinct data marts rather than a large scale central data warehouse. These data marts often contain copies of the data found in another data mart. It is not uncommon for one data mart to use another data mart as a source for data. This leads to extra consumption of disk storage, increased use of network bandwidth to move duplicate data, and additional ETL processing power to get the same data into multiple data stores. Although the cost of any one data mart may seem small, the cost of all of them can amount to a significant investment.

When companies have multiple data marts of data warehouses, they often host them on diverse platforms. The distinct hardware platforms each need distinct maintenance. Caring for more servers of different makes and models can be more costly. If the servers are mission critical, the problem increases as the hardware is duplicated to provide redundancy. Different hardware often means different operating systems (Oss) and different database management systems. This causes lower quantities of licenses to be purchased and can mean higher unit costs.

Although a single, large scale data warehouse system may have a higher initial cost than a data mart-size operation, it often proves more economical in the long run. A single data warehouse system that can scale can be sized to meet the current needs and add capacity as the needs of the organization warrant it.

Operational Efficiency

If you are maintaining multiple servers hosting multiple database systems, there are several operational costs that need to be considered:

- Training staff on multiple systems
- Monitoring multiple systems
- Troubleshooting on multiple platforms
- Patching and updating multiple systems

It is simpler and less costly to operate a single system with a unified OS and hardware platform. Monitoring is much easier and troubleshooting is greatly simplified.

Summary

This chapter discussed the rationale for building large scale data warehouses. In some cases, the size is driven simply from the size of the data set. Large volumes of data need to be organized and processed to harvest the value. In other cases, there are options. Using multiple data marts or smaller, segmented data warehouses can be an option. Many organizations are better served if the intelligence for the entire enterprise is gathered into one consistent, secure whole.

If the engineering challenge of building, operating, and managing a large scale data warehouse can be met economically, it will often provide the best alternative for storing and analyzing the data collected from within and outside of the organization.