The Evolution Of A Data Warehouse Architecture - One Size Fits All ?

Howard Ong

Principal Consultant

Aurora Consulting Pty Ltd

Abstract

A promising new star on the IT horizon, Data Warehousing overcomes many of the shortcomings of early Decision Support and Executive Information System. A key to successful Data Warehousing though is to understand that a Data Warehouse is not just a collection of technologies but an architecture. This paper explains the various components of a matured Data Warehouse architecture. It examines the different evolutionary routes that an organisation can take to developing a Corporate Data Warehouse solution. Where appropriate, the paper draws illustrations from real-life Data Warehouse applications the author has built over the past years. In conclusion, the paper suggests that there is no one right route to the evolution of a Corporate Data Warehouse Architecture. Organisations need to examine its short- and long-term analytical information needs, corporate IT maturity, and other political and organisational factors before deciding on an optimal evolutionary path.

About The Author

A frequent presenter at local and regional conferences, Howard has been a proficient user of database technology since 1991. In particular, Howard possesses in-depth experience in the planning and development of Business Intelligence and E-Business Applications. Harbouring a keen interest in the application of advanced database in these areas, Howard and his team have helped many organisations deploy BI and E-Business solutions using Oracle, Microsoft and IBM technology. Since 1997, Aurora Consulting has been delivering Data Warehouses and E-Business solutions to a wide variety of organisations including large government departments and ASX-listed companies.

For more information on Howard or Aurora Consulting, visit http://www.aurora-consult.com.au or email info@aurora-consult.com.au.

Introduction

Walk into any computer book store today and you will likely stumble upon a barrage of books written on Data Warehousing. Talk to a sales representative from any of the major software vendor today and you will likely be bombarded with a barrage of software and tools that are supposed to help you build the Data Warehouse of your dreams. It is not uncommon for organisations wading their first tentative steps into this new arena to be confronted with an overwhelming number of mind-boggling questions. Many of these questions would undoubtedly relate to the very definition of a Data Warehouse itself and how can one go about developing one. In this paper, the author hopes to shed some light on some of these questions by providing a practical guide to building a Corporate Data Warehouse solution by suggesting 4 evolutionary routes that one may consider.

The Data Warehouse As An Architecture

How often do you hear software sales persons equating great software products with successful Data Warehouses ? The truth is, choice of appropriate software, though important, does not

Data Warehouse is not one, or even a combination of software products - it is an architecture of system components. Described below are some common components of a Data Warehouse architecture and their interrelationship. Note that not all of these are essential components to a Data Warehouse architecture. In fact, even the infrastructure by which they are brought together may vary from warehouses to warehouses.

- Operational Source Systems. These are operational systems that supply the Data Warehouse with data. These systems can range from decade-old legacy systems to small standalone Access databases.

- Operational Data Store (ODS). The ODS is a repository of analysis data of a fine granularity and typically of a low retention period. It is commonly used in scenarios such as that of a retail outlet chain, where the Sales Director would like to track sales figures on a daily basis. However there is usually little likelihood of such fine granularity data to be required over a long period of time. For example, the Sales Director may wish to compare the current day's sales figure with that of last month or even last quarter, but probably not beyond. ODS' can play a key role in keeping the total size of the Data
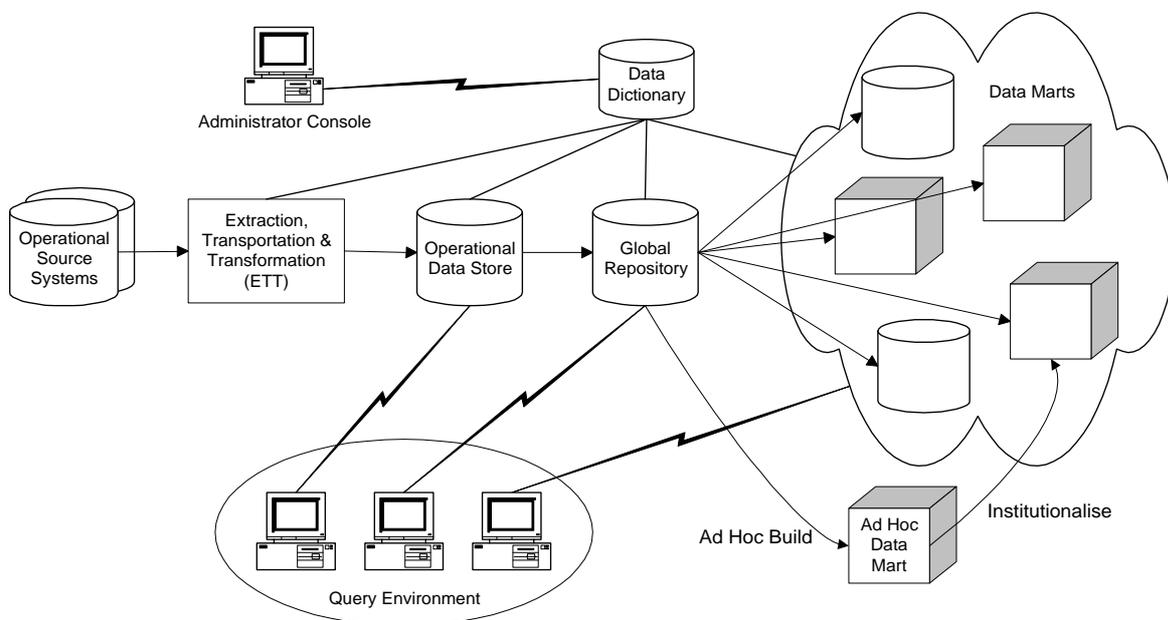

Figure 1 A Matured Data Warehouse Architecture

guarantee successful Data Warehouses. In fact, a

Warehouse manageable.

- Extraction, Transportation and Transformation (ETT) processes. The ETT processes transport and transform data from the Operational Source Systems into analysis data. It is equivalent to what Microsoft called *Data Transformation Services*. Common functions of the ETT processes include, automated data load, manual data entry, derivation, verification and data cleansing.

- Global Repository. The Global Repository is a repository of all Data Warehouse data. It is typically a ROLAP or a 3NF relational database.

- Data Dictionary. The Data Dictionary houses the Data Warehouse's meta data. two types of meta data may reside in the Data Dictionary : structural information such as measure dimensionality and dimension structures; and ETT process information such as derivation logic, validation procedures, status of data loading processes, etc.

- Administrator Console. The Administrator Console performs Data Warehouse administration tasks such as maintenance of dimensions and measures; definition of ETT processes, monitor ETT runs, etc.

- Data Marts. Data Marts house subject-specific subsets of information from the Global Repository. Depending on the requirements, these Data Marts can either be implemented using ROLAP or MOLAP technology. Some advantages of Data Marts over the Global Repository are simpler database structure, versatility of user interface, improved performance, richer functionality, and in some cases, portability. From time to time, additional Data Marts can be built on an ad hoc basis to analyse particular subject areas of interest. If desired, ad hoc Data Marts can be institutionalised and become part of the pool of permanent Data Marts.

- Query Environment. Depending on the needs, a wide variety of query tools can be deployed in the query environment. These can be traditional client-server tools (Oracle Developer, PowerBuilder); web-query tools (Oracle PL/SQL Cartridge, Cold Fusion); custom-built applications, ROLAP tools (Oracle Discoverer, Cognos Impromptu, Business Objects); or MOLAP tools (Oracle Express Objects, Cognos Powerplay).

Note that not all of the above-mentioned components are present in all Data Warehouses. In fact, of the 8 components, only the Operational Source Systems, ETT Processes and Query Environment are essential. However, some components, like the Global Data Warehouse Repository and the Data Dictionary, are highly desirable.

Evolutionary Routes

Rome was not built in a day, neither is the Data Warehouse. Like any other major project, building a Data Warehouse requires significant time and resource commitment. With these commitments comes the importance of planning and risk management. Drawing from past experience, the author would like to suggest 4 evolutionary routes for an organisation to develop its Data Warehouse environment. Note that *these evolutionary routes are meant to serve as suggestions only*. No two Data Warehouse environment are alike, and neither are the routes taken to evolve them. Depending on your particular situation, your organisation may choose to develop the Data Warehouse components in a different order; omit certain components; or mix and match facets from the different evolutionary routes. Whatever the case, remember this Golden Rule : A well-planned journey is more likely to give you a smoother ride.

MOLAP Centric Route

Probably the most popular route, the MOLAP centric route emphasises on utilising MOLAP technology to deliver flashy prototypes within a short period of time.
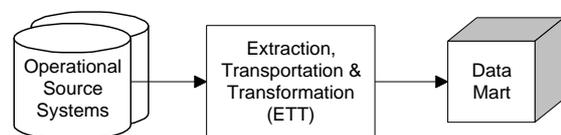

Figure 2 MOLAP Centric Route - Stage 1

Stage 1. A subject area is identified and a data mart is built by extracting data directly from one or more operational source systems and loading it into a MOLAP database. The prototype serves a very important function of getting the users acquainted with the new paradigm of OLAP analysis. As

users' confidence grows, what Corey and Abbey called "power users" start to emerge. These are users who have "a desire to learn technology and a willingness to teach others". Power users could be anyone from an IT manager to a technically oriented accountant. Power users help the Data Warehouse developers promote the Data Warehouse to the wider user-community, a crucial role in the success of the Project.
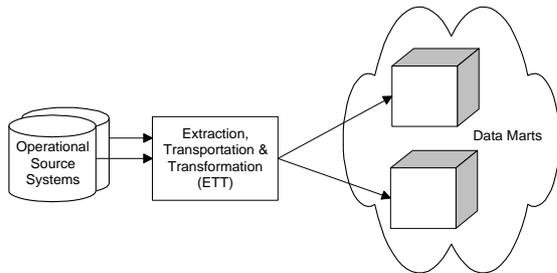

Figure 3 MOLAP Centric Route - Stage 2

Stage 2. Encouraged by the success of the initial data mart, more subject areas are identified, and more data marts are built in the same manner.
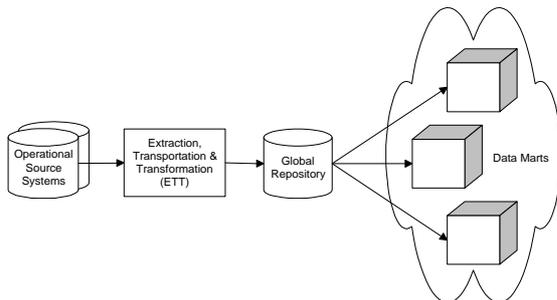

Figure 4 MOLAP Centric Route - Stage 3

Stage 3. As more data marts are developed, the need for a more integrated approach to loading these data marts arises. Consequently, a Global Repository, acting as an intermediary between the Operational Source Systems and the Data Marts, is developed. Through the Global Repository, a uniform set of data cleansing rules, standardised derived measure calculation, and centralised manual measure data entry are implemented. The development of the Global Repository represents a major milestone in the evolution of the Data Warehouse architecture. Often, it involves significant effort in the development of an integrated data model; redesign of the individual data marts; and in some cases, migration of cleansing and derivation routines from the data marts to the Global Repository.

The quick delivery of this approach has the advantage of gaining quick converts by delivering results within a short span of time. However, the disadvantages of the quick-and-dirty delivery of analysis solutions without first deploying an underlying supporting architecture should not be overlooked. Management, unaware of the resource intensive nature of Data Warehouse development, might be led to believe that the collection of MOLAP prototypes *is* the Data Warehouse, and hence withhold support for the substantial development effort required in evolving an integrated Data Warehouse architecture. Without an integrated architecture, the organisation would have difficulty in satisfying its future needs for other analytical information. In the long run, the organisation, having found itself moving from a situation of non-integrated operational systems to one of non-integrated data marts, could easily become disillusioned with Data Warehousing altogether.

ROLAP Centric Route

The motivation behind this approach is very similar to the MOLAP Centric approach, with the major difference being the employment of ROLAP technology instead of MOLAP.
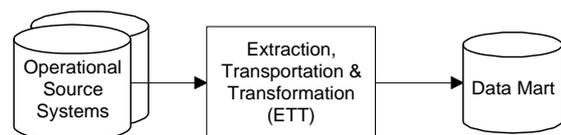

Figure 5 ROLAP Centric Route - Stage 1

Stage 1. As with the MOLAP Centric approach, an initial subject area is identified and a data mart is built.

Stage 2. As the initial data mart successfully helps users acquaint with the new paradigm of OLAP analysis, more subject areas are identified. This leads to the development of additional data marts.
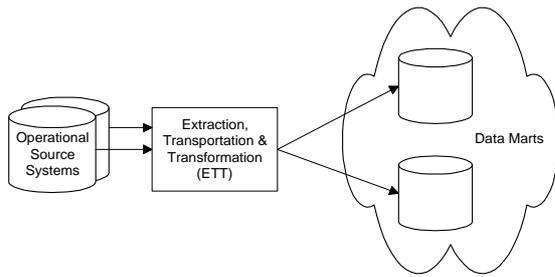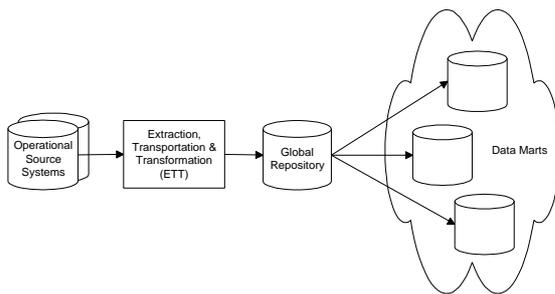
Figure 6 ROLAP Centric Route - Stage 2

Figure 7 ROLAP Centric Route - Stage 3

Stage 3. As the number of standalone data marts grows, data consistency and data sharing becomes an issue. This leads to a watershed realisation that while the data marts serve well to sell the initial concept of OLAP analysis to the users, a more integrated architecture is required to prevent the data marts from degenerating into islands of information. A Global Repository is introduced to address this concern.

The ROLAP Centric approach seeks to deliver quick-win OLAP prototypes by leveraging off an organisation's investment in relational technology. Compared to the MOLAP Centric approach, it enjoys cost-savings in terms of licenses and human resources. However, owing to the limitation of current breed of ROLAP tools, the data marts' user interface and system performance are likely to be less impressive compared to their MOLAP counterparts. Like the MOLAP Centric approach, it is important for the Data Warehouse developer to stress to the Management that the initial data marts are a mere first step in a long evolution journey. Significant development effort is needed before the long-term need for a information analysis environment can be addressed.

Life Cycle Route

Like the traditional Structured Development Life Cycle (SDLC) development methodology, the Life Cycle Approach seeks to develop the various Data Warehouse components in a logical order.
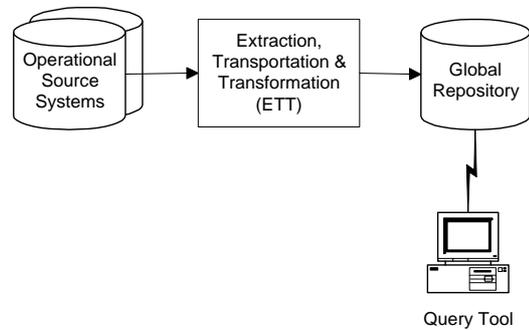

Figure 8 Life Cycle Route - Stage 1

Stage 1. One or more subject areas are identified. Using these as a starting point, a corporate data model is developed, forming the basis of the Global Repository design. Significant effort is expended in infrastructure development such as uniform data cleansing rules, standardised derived measure calculation, and centralised manual measure data entry. A query application is deployed to present one or more pilot subject areas.
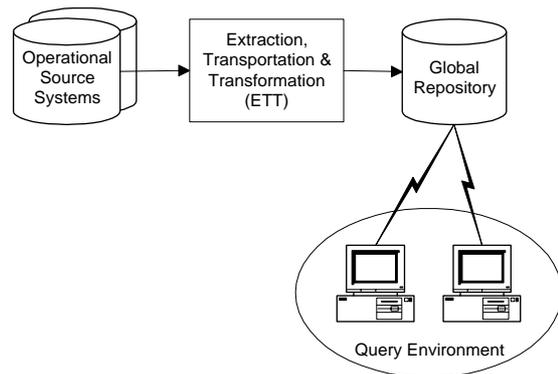

Figure 9 Life Cycle Route - Stage 2

Stage 2. As more subject areas are identified, they are integrated into the corporate data model of the Global Repository. More query applications are deployed to access information from the newly added subject area. A chief qualitative difference between the Life Cycle approach and other

approaches is that the newly added subject areas are *integrated* into the existing subject-areas. This avoids the problems of data consistency and data sharing, often the very reasons Data Warehouses are developed.
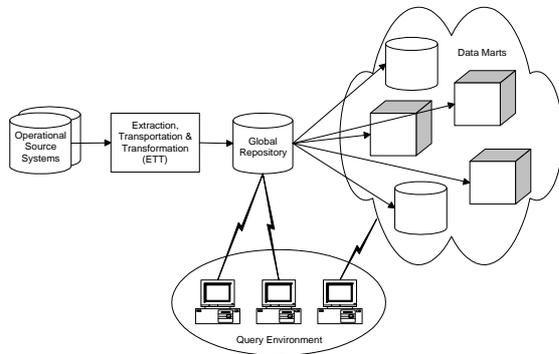

Figure 10 Life Cycle Route - Stage 3

Stage 3. As mentioned earlier in the paper, the low granularity and rich dimensionality of the Global Repository can be as much a handicap as a virtue. The desire for simpler database structure, more impressive user interfaces, better performance, richer functionality and portability provide the common impetus to the development of Data Marts.

In order to achieve quick-win solutions, both the MOLAP and ROLAP Centric approaches violates the logical order of Data Warehouse development by deferring the development of the Global Repository till late in the project. From a development cost point of view, the "back-filling" of the Global Repository at a later stage is usually more expensive than developing it early. In some cases, it is difficult to convince Management to commit to a development effort without a visually discernible end-product. Whether developed early or late, the development of the Global Repository is a time-consuming exercise. Developing it early, though cost-effective in the long run, is only suitable for organisations whose Management has a clear understanding of Data Warehouse development life cycle. Such Management would more likely support more lengthy development of the initial application for the long term benefit of the project.

Operational Centric Route

The Operational Centric Route is normally taken by organisations with analysis need for low granularity operational data. An example of such an organisation is a retail outlet chain whose prime information analysis needs are those of daily sales data.
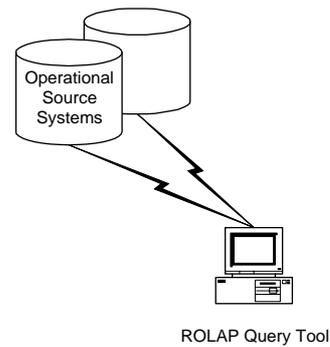

Figure 11 Operational Centric Route - Stage 1

Stage 1. A ROLAP tool is deployed over existing operational systems to provide low granularity OLAP analyses. This is a relatively inexpensive exercise, sometimes costing little more than the license cost of the tool itself. Many ROLAP tools today are shipped with pre-defined end-user catalogs for popular Enterprise Resource Planning applications such as Oracle, SAP and PeopleSoft.
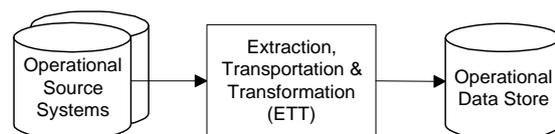

Figure 12 Operational Centric Route - Stage 2

Stage 2. The approach of deploying a ROLAP tool directly over the operational systems is fine as long as cross-system referencing of data is not required. Once the need for data sharing arises, the importance of data consistency becomes evident as well. This leads to the need for an integrated Operational Data Store. Like the Operational Source Systems, the content of the Operational Data Store is likely to be of low granularity and with a short retention period. Daily data with a 3-6 months' retention period is not uncommon.
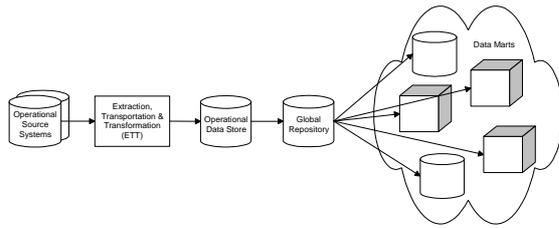
Figure 13 Operational Centric Route - Stage 3

Stage 3. As expected of data of low granularity, a short retention period is required in order to keep the size of the Operational Data Store manageable. However, while it is a common phenomenon that organisations do not normally perform low granularity analysis over long periods, they do require higher granularity analyses of older data. For example, it would not be unreasonable for a retail store outlet manager to request daily sales summary of up to 3 months old; and monthly sales summary of up to 5 years old. The natural solution to this requirement is the development of a Global Repository to house data of a higher granularity to the Operational Data Store. With the Global Repository set up, it opens up other options, such as the deployment of versatile Data Marts to better cater for the requirements of specific user communities.

Perhaps more so than the ROLAP Centric approach, the Operational Centric approach allows organisation to venture into the new paradigm of OLAP analysis without wading too deep into new concepts and technology. A chief reason for that is the architectural simplicity of the initial deployment of a ROLAP tool directly over the Operational Systems. The main risk of this approach is the tendency to treat Operational Data Store as the Data Warehouse, without considerations to the necessity of retaining low granularity data over a long period of time. By performing medium- to long-term higher granularity analyses on the low granularity data, system performance suffers. Over time, as the Operational Data Store becomes unduly large and cumbersome, the system performance of short-term low granularity queries suffers as well.

Beyond Stage 4

To bring the Data Warehouse to a matured state as depicted in Figure 1, other components such as ROLAP query tools, Data Dictionary and the Administrator Console may be built and deployed in the order they are required. It is important to stress again that no two Data Warehouses are alike. Similarly, not all organisations aspire to the level of Data Warehouse maturity depicted in Figure 1.

Conclusion

Data Warehouse being the new IT discipline that it is, will continue to enjoy the explosion of ideas and advances it is currently experiencing. Academic concepts will continue to be refined and new technology will continue to emerge. This paper aims to contribute the advances of this fast exploding field by providing practical guidelines to help organisations harness the potential of this new and exciting breed of application systems. It is clear that there is no one silver bullet to the successful evolution of a Data Warehouse architecture. It is imperative that organisations assess its specific needs and where possible seek the advice of Data Warehouse specialists to help it determine its optimal evolutionary route. Remember the Golden Rule : A well-planned journey is more likely to give you a smoother ride !

References

Corey, Michael J, and Michael Abbey. *Oracle Data Warehousing.* Osborne McGraw-Hill 1997.

Inmon, W H. *Building The Data Warehouse.* Second Edition. John Wiley & Sons 1996.

Kimball, Ralph, and W H Inmon. *The Data Warehouse Toolkit.* John Wiley & Sons 1996.