

Data Warehouse Myths And Misconceptions

Howard Ong

Senior Consultant

Aurora Consulting Pty Ltd

ABSTRACT

For many years, the quest for competitive advantages has prompted many organisations to attempt the paradigm shift from data processing to the new exciting arena of information analysis. The advent of Data Warehousing promises to overcome many of the shortcomings of early Decision Support and Executive Information System. However, like any new IT discipline, Data Warehousing is flawed with myths and misconceptions. Whether it is concepts, techniques or technology, the list of myths and misconceptions reads like a fairy tale. In this paper the author examines some common myths and misconceptions; scrutinises their apparent validity; and presents a detailed list of counter arguments to dispel them.

About The Author

A principal of Aurora Consulting Pty Ltd and a frequent presenter at Oracle Conferences throughout the region, Howard has been working on Data Warehouses since 1991, before the term "Data Warehouse" was even coined. Howard possesses in-depth experience in the planning and development of Data Warehouses; and has been an expert user of Oracle database and tools for the past 6 years. Harboring a keen interest in the application of Oracle Technology in Data Warehouse development, Howard and his team of consultants have helped many organisations deploy Data Warehouses using technology such as Oracle Express, Discoverer, advance features of the Oracle8 Server and Aurora's proprietary DataWare and DataView products. Aurora's most recent success includes a major Data Warehouse initiatives at Education Department Western Australia. In other projects throughout his many years of consulting experience, Howard has worked with a wide variety of organisations such as government departments; and companies from mining, finance and transportation industries. In addition to Data Warehouse development, other services that Howard rendered to these organisations include business analysis, relational application development, multidimensional database design and database administration.

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

Introduction

Since the dawn of the data processing age, organisations have been building application systems aimed at addressing their operational needs. The IT industry has reached a stage where many organisations now possess years, if not decades, of detailed operational data, buried in a repertoire of legacy systems. The constant quest for a competitive edge has prompted many organisations to attempt the paradigm shift from sheer data processing to the new exciting arena of information analysis. However, like any new IT discipline, the Data Warehouse field is flawed with myths and misconceptions. The IT worker wading his first tentative step into this new arena will soon find the field booby-trapped with conflicting ideas and inconsistent concepts. In this paper, the author hopes to help smoothen the bumpy ride by dispelling some of the common myths and misconceptions.

Myths & Misconceptions

Myth 1: Every Organisation Needs A Data Warehouse.

Over-zealous software sales representatives may have you believe that building a Data Warehouse is the panacea to all your organisation's IT problems. To a varying extent and depending on your particular situation, they could be right.

With the wealth of information captured in operational systems that ranges from Payroll to Order Processing, many organisations today possess the information from which they could reap enormous benefits. Strategically deployed, a Data Warehouse could harness these information to help a Marketing Manager devise appropriate advertising strategies; a Mine Production Manager identify cost-effective mining techniques; a Government Policy Maker allocate public resources more effectively; ... and the list goes on.

So why *shouldn't* every organisation build a Data Warehouse ? To explain this paradoxical issue, the author would like to draw on the parallel concept of Maslow Hierarchy of personal needs. According to Maslow Hierarchy, an individual is said to progress through 5 levels of needs, from most basic to the most sophisticated. Like an individual, an organisation's needs for information also progress from basic to sophisticated. With few exceptions,

the need for sophisticated analytical information arises only upon the fulfilment for more basic information needs. Data Warehouses are a sophisticated breed of systems that would only benefit organisations with a certain degree of IT maturity. To embrace Data Warehousing prematurely, while an organisation is still striving to meet its day-to-day operational information needs is virtually an invitation to disasters. The likely pitfalls that such an organisation may encounter includes :

- Lukewarm support from Management, owing to inability to appreciate the need for Data Warehousing.
- Insufficient resource and funds allocated to the Data Warehousing Project owing to more pressing needs for operational systems.
- User expectations which could not be met due to immature operational systems and insufficient resources.
- Excessively long or total inability to develop satisfactory Extraction, Transportation and Transformation (ETT) processes.

Any of the above can cause the Data Warehouse Project to fail miserably, leaving the Management and IT Worker disillusioned with this new breed of systems.

Myth 2: A Data Warehouse Is A Data Archival System.

Data Warehouses and Data Archival Systems shared one common characteristic - they both store non-volatile historical data for query purposes. However, this is where the similarities stop ! Beyond the storage of historical data, Data Warehouses and Data Archival Systems have many qualitatively distinctive characteristics. Some of the key differences are :

- Functionally, a Data Archival System often serves as a fallback when an outdated system is decommissioned. A Data Warehouse, on the other hand, is a strategic initiative requiring the support and sponsorship of senior management.
- Typically, a Data Archival System stores data from a single Operational Source System. Even where multiple operational systems are involved, little or no effort is

made in integrating data from the heterogeneous source systems. Cross-system data integration, on the other hand, is one of the key characteristics of a Data Warehouse. As such, the data loading activities of a Data Warehouse is often a complex and meticulous process involving extensive data cleansing and transformation.

- The source systems of a Data Archival System are often Operational Source Systems that have been put out of use. A Data Warehouse, however, takes its data feeds from live Operational Source Systems. This attributes to a key qualitative difference between the data stored in these 2 types of systems : while the Data Archival System stores a single snapshot of its Operational Source Systems, the Data Warehouse is constantly updated with time-variant snapshots of its Operational Source Systems.
- As it is no more than a fallback measure for decommissioned operational systems, the number of queries against a Data Archival System tends to be few and far between. In fact, it is not uncommon to find a Data Archival System residing on slower, cheaper media such as tapes or even microfiches. On the other hand, a Data Warehouse, like its Operational Source Systems, is very much a live system. In fact, so intense are the query activities on the Data Warehouse that it is often deployed on optimally tuned databases. In addition, very specific and specialised technologies such as star schemata and multidimensional databases are usually employed.
- With the exception of archiving additional Operational Source Systems, the design and architecture of a Data Archival System rarely changes. The Data Warehouse, however, is a live system that constantly evolves as the organisation's business needs change and new requirements emerge.

Myth 3: Data Warehousing Is OLAP.

Data Warehouse applications exist well before the term *Data Warehouse* is even coined. Some of terminology used in the earlier years for essentially

similar concepts includes *Decision Support System* (DSS) and *Executive Information System* (EIS). The recent advent of Online Analytical Processing (OLAP) concepts and tools provide a superior implementation vehicle whereby some of the earlier promises of DSS and EIS could finally be realised. Superior as they are, OLAP tools and techniques are *not* the only way to implement Data Warehouses.

Inmon, widely regarded as the “father of Data Warehousing”, defined a Data Warehouse as “a subject-oriented, integrated, non-volatile, and time variant collection of data in support of management’s decisions”. Corey & Abbey defined a Data Warehouse as “a collection of corporate information, derived directly from operational systems and some external data sources”. Inmon’s definition examines the Data Warehouse from a qualitative perspective; while Corey & Abbey defines the Data Warehouse from a “content” point of view. It is not a coincidence that none of these definitions specifies the Data Warehouse by its underlying implementation technology. A Data Warehouse does not necessarily need an OLAP repository. In fact, depending on the circumstances, more traditional technologies such as the 3rd Normal Form (3NF) relational database can be an appropriate medium for a Data Warehouse. In a recent project undertaken for a government agency, the author implemented a Data Warehouse by integrating selected information from several operational systems into a single 3NF relational schema. From a functional point of view, and certainly by the 2 definitions quoted above, this relational schema is in every sense a Data Warehouse.

Myth 4: ROLAP Is Better Than MOLAP.

Relational OLAP builds upon established relational concepts and technologies to deliver OLAP capabilities. Concepts such as star schema implemented on Oracle8’s star query and bitmap index technologies make the Oracle8 Server a powerful ROLAP engine.

The advantages of ROLAP are :

- Facilitation of the OLTP-OLAP paradigm shift. Even with an appropriate level of IT maturity, it is sometimes difficult for an organisation to initiate the paradigm shift from OLTP and OLAP. Based on the same relational technology as OLTP, ROLAP technology facilitates the

paradigm shift by minimising licensing costs and technology transition.

- Support for lower granularity. ROLAP Data Warehouses tends to store data at a lower level of granularity, sometimes down to the level of the atomic records. This supports greater flexibility when performing ad hoc analyses. For example, by storing student enrolment numbers against specific schools, you have a choice to view enrolment numbers by school categories (i.e. types of schools), or by geographical districts, or in fact by any other means by which a school can be classified. On the other hand, if enrolment numbers are stored against school categories directly, you would then not be able to break enrolment numbers down by geographical districts.
- Rich dimensionality. With ROLAP, it is not uncommon to store a data measure against all its available dimensions. At query time, dimension that are not required can be omitted. For example, if enrolment numbers are stored against schools, age groups and ethnicity, it would not be hard to display enrolment numbers broken down only by schools and age groups. The extra ethnicity dimension can be omitted by aggregating over all ethnicity.

However, ROLAP solutions are not without their problems :

- Poor query performance. The added flexibility of lower granularity and rich dimensionality comes at a cost - that of poor query performance. Most ROLAP query tools perform on-the-fly aggregation to present data at the required granularity and dimensionality. In the real-world, where it is not uncommon to find 1,000,000 atomic records stored against 20 dimensions, even the most optimally configured Oracle8 database would take rather long to perform the on-the-fly aggregation.
- Complex summary management. A number of ROLAP tools such as Oracle Discoverer support summary management and query redirection to improve query performance. Summary management pre-calculates and stores measure value at frequently accessed combinations of dimension levels. At query time, query

redirection run the query against the pre-populated summary tables to shorten query time. However, summary management is a complex and tedious exercise. Consider a modest data measure with 6 dimensions, each consists of 2 levels. The number of potential summary combinations for this modest data measure is $2^6 = 64$! The problem is further accentuated by the typical user whose query patterns are often unpredictable.

- Less customisable query tools. Unlike their MOLAP counterparts, ROLAP query tools in the likes of Oracle Discoverer, Cognos Impromptu and Business Objects are less customisable. With more customisable ROLAP query tools yet to emerge, developers have to resort to more general-purpose tools such as Application Server PL/SQL Cartridge and Oracle Developer to develop customised ROLAP query environments.

Myth 5: MOLAP Is Better Than ROLAP.

MOLAP solutions based on natively multidimensional databases such as Oracle Express Server and Cognos Powerplay represent a greater technology shift from the OLTP-relational paradigm.

The advantages of MOLAP are :

- Superior query performance. MOLAP tools such as Oracle Express Server pre-aggregates all combinations of dimension levels to deliver superior query performance.
- Simple summary management. In Oracle Express Server, aggregation is achieved by a simple command - the ROLLUP command.
- More customisable query tools. Tools for custom-development are abound in the MOLAP arena. For example, Oracle offers an object-oriented development environment in Oracle Express Objects (OEO), Web publishing capabilities in Express Web Agent and Project Walden promises to deliver a Web-enabled version of OEO in the not-too-distant future.

Every road has its hazardous bends, and the MOLAP solution is no exception :

- Significant paradigm shift. Not only is OLAP a totally new breed of systems from OLTP, MOLAP tools is a whole new category of software tools from relational ones. Like any significant technological shift, it brings with it higher licensing costs, training overheads and stronger user resistance.
- Rigid data structures. Once built, the granularity and dimensionality are etched into the database structure. Changes such as altering the granularity level and the addition/removal of a dimension requires the database to be redesigned and rebuilt. Even a change of parenthood within a dimension can incur substantial overheads. For example, consider a data measure that tracks outlet sales of a fast food chain. Lets assume that with effect from October 1998, Outlet A was rezoned from Suburb X to Suburb Y. A simple requirement to associate Outlet A with its new suburb for all sales made from October 1998 and at the same time continue to aggregate sales data prior to the rezone into Suburb X is extremely tedious to fulfil and is associated with high maintenance overheads.
- Poor build performance. The superior query performance and simple summary management come at a cost of lengthy build time. Using the earlier example of a modest data measure with 6 dimensions, each consists of 2 levels, all 64 summary combinations are calculated during the build, without exception !
- Extremely large databases. Cross-dimensional sparsity, where data value does not exist in a high proportion of data cells, is a common symptom in MOLAP databases. For example, where there is no children of certain age groups attending certain schools, sparsity exists between the dimensions age group and school. In many circumstances, sparsity of 95% or more is rather common. Thus, for many ROLAP databases, the combination of exhaustive summary generation and cross-dimensional sparsity give rise to extremely large databases. In spite of features such as Oracle Express' Composite and Cognos Impromptu's Compression, which help reduce database sizes, it is common for even moderately sized MOLAP databases to run into tens of gigabytes.

Thus, between ROLAP and MOLAP, there is no clear answer on which is more superior. Often, the choice between ROLAP and MOLAP hinges upon the type of application on hand. Below are some guidelines :

- ROLAP is useful in certain Proof Of Concept projects where there is a need to introduce the concept of OLAP without a shift from familiar relational technology.
- The versatility of MOLAP query tools makes it ideal for Proof Of Concept projects where a quick win with a flashy front-end is the goal.
- ROLAP's ability to cater for details and structural changes makes it ideal for the Data Warehouse Global Repository, from which subject-specific Data Marts with specific granularity and dimensionality are derived. Global Repository and Data Marts are explained later in this paper.
- MOLAP's superior query performance and customisability presents as a excellent vehicle for certain types of subject-specific Data Marts.

Myth 6: Successful Data Warehouse Is All About Buying Great Software.

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 guarantee successful Data Warehouses. In fact, a Data Warehouse is not one, or even a combination of software products - it is an architecture of system components. Listed below are some common components of a Data Warehouse architecture. Figure 1 depicts 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. (For an explanation of these components and how they may be evolved, the reader is referred to another paper by the author entitled "The Evolution Of A Data Warehouse Architecture - One Size Fits All ?")

- Operational Source Systems.
- Operational Data Store (ODS).

- Extraction, Transportation and Transformation (ETT) processes.
- Global Repository.
- Data Dictionary.
- Administrator Console.
- Data Marts.
- Query Environment.

explosion of ideas and advances from both a technological and an academic perspective. This paper hopes to contribute to the advances of this fast exploding field by dispelling some of the commonly misunderstood concepts. It is worth noting myths and misconceptions are abound and this paper is far from exhaustive. Where possible, organisations contemplating the development of a Data Warehouse environment should seek the advice of experienced and qualified Data

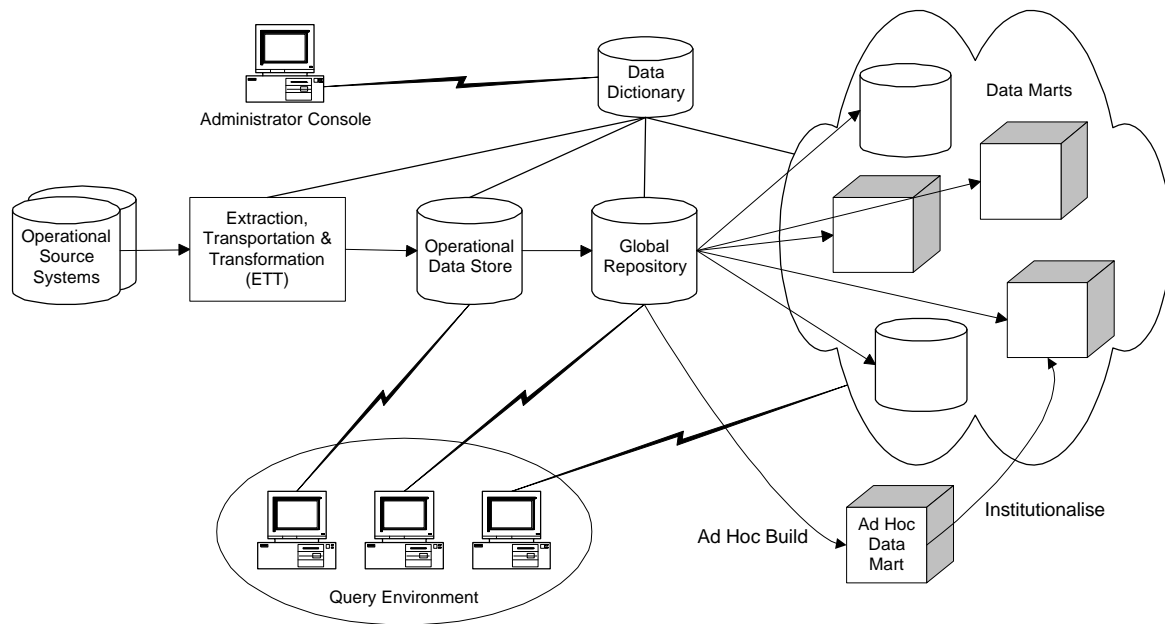


Figure 1 A Matured Data Warehouse Architecture

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.

Conclusion

Being a relatively new IT discipline, we can expect Data Warehouse to continue experiencing an

Warehouse specialists. Remember the old saying :
“A stitch in time saves nine” !

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.