EXPLORING THE PRINCIPLES AND PRACTICES OF DATA WAREHOUSING

Rick D. Allemandi*

Executive Summary

I selected Data Warehousing, in Contra Costa County as the topic for my BASSC Internship and my associated case study. In this case study, I explore the technical concepts of existing data retrieval systems, data warehousing, and the history of the Contra Costa Data Warehouse Project. While this research will be helpful in Santa Cruz, the greatest insight I received as a result of this project was not what I had expected.

The greatest insights learned on this project were not actually derived from the technical issues as much as the "people" issues involved with Data Warehousing.

While reviewing Contra Costa County's data warehousing project, I was surprised to learn that at one point, they were on the verge of a crisis. They had spent large sums of money purchasing new hardware and software, and even on bringing "experts" into the project. When the "experts" left the project, they had an unusable system and no knowledge about how to complete the project.

Realizing that they lacked the skills internally, Contra Costa County again looked to an outside consultant. Moving forward would require establishing someone to serve as a Project Manager for the Data Warehouse project. A Project Manager was found and he immediately began to plan how he would build his team.

The first step was to decide the roles and responsibilities of the future members. Once this was defined, it was time to build the team. In order to proceed in a timely manner, and to ensure the transference of knowledge, it was decided that the new team would consist of four Contra Costa County employees, and four outside consultants.

Once the team had been assembled, the Project Manager created a "Functional Assignment Chart." This chart showed the different Data Warehouse tasks and how these tasks were to be spread between team members. This chart was used to ensure that all "players" had a chance to build upon their Oracle and Data Warehousing expertise. It was also utilized in documenting "technology transfer."

RECOMMENDATIONS

I have two recommendations to the director of Administrative Services and to the manager of Information Technologies. My first is aimed at what must be defined as a major concern with the Oracle / Data Warehouse Project in Santa Cruz County.

Santa Cruz County has one person assigned full time to their Oracle / Data Warehouse project. While that one person is proficient at what he does, there is no backup staff for this person. My recommendation is to immediately train existing personnel in Oracle and Data Ware-housing technologies. It is extremely important to have the technology understood by more than one

* Rick D. Allemandi is a Departmental Systems Analyst for the Human Resources Agency of Santa Cruz County

person, to sustain continued development and maintenance of the system in case of personnel changes.

My second recommendation is an outgrowth from the "Functional Assignment" chart created by the Contra Costa County Data Warehouse Project Manager. This chart is an excellent way to document the transfer and sharing of technology. I recommend that Santa Cruz County review their existing Information Technologies Department, and create something similar to this chart.

Recently, the HRA Information Technologies Department in Santa Cruz County lost the ability to generate maps, and charts that utilized the Are View GIS software. The only person who was trained and understood this complex system went on maternity leave. For the last 8 months or so, the county has been unable to use this software. A program that developed cross training or that encouraged technology sharing may have prevented this loss.

EPILOGUE

Several major changes have already been implemented in Santa Cruz County as a result of this internship project.

Santa Cruz County also utilizes outside consultants to aid in their Oracle software applications. To prevent the type of mistake experienced by Contra Costa County, special care is given to ensure that all consultants are highly monitored. Consultants are only brought in for well-defined tasks, and are paid only when they have completed them.

The Data Management Group has taken several steps to ensure that staff is sharing technologies. Specifically, in order to expose other staff members to the Oracle Software and to data warehousing technologies, the Data Management Group has promoted a Departmental Data Processing Coordinator in the Client Applications Department, to an Information Systems Analyst in the Data Management Group. In this new position, the analyst will be exposed to the Oracle software and data warehousing technologies.

The Data Management Group has also begun to share Internet technologies. Previously, one staff member was completely responsible for all aspects of the Agency's web sites and Intranet. A second staff member now shares the responsibility for web site creation and maintenance, with plans for another staff member to participate in these tasks.

The Data Management Group is also on the road to re-establishing the ability to work with the Are View GIS software. A staff member has been assigned to work on this project part time.

While a formal Functional Assignment Chart has not yet been created, the seeds have been planted. I expect a continuance in the sharing of technologies within the Santa Cruz County Human Resources Agency.

EXPLORING THE PRINCIPLES AND PRACTICES OF DATA WAREHOUSING

Rick D. Allemandi

INTRODUCTION

Governmental agencies are constantly under increased stress and public scrutiny to provide better services to our citizens, to reduce fraud, and to better manage our very limited resources. It is also becoming increasingly important for government to respond quickly to environmental and economic changes, to predict trends, and to provide status reports to other government entities and to the public at large.

The private sector has always faced the type of challenges that government is beginning to experience. Their well-being has long dictated the need to predict emerging trends, and respond quickly to market changes. Businesses must maximize profits by responding to their predicted trends, and by watching and tracking expenses. They must report to shareholders and regulating organizations, to justify their actions. To accomplish these tasks, the private sector has relied upon fast, accurate, and consistent data. More importantly, they have learned that they must understand the data they collect.

One thing that government agencies have always been "rich" with is data. All levels of government have always collected massive amounts of data, but understanding and using of this data has been a challenge. One of the biggest challenges comes with knowing "where to find what you need." What if all data was stored in a single repository? Is it possible to combine data from different sources? Once data has been stored, can you retrieve it into a usable form? What does it mean to warehouse data? What is a Data Warehouse?

In this case study I will examine these questions, and explore the theoretical principles and practical applications associated with Data Warehousing. I will research this information, based upon the successful creation and implementation of a Data Warehouse in Contra Costa County.

TYPICAL EXISTING SYSTEMS

To understand how a Data Warehouse functions, it is important to understand how data is currently processed, and what its limitations are.

Most current data collection and reporting systems are based upon either the On Line Transaction Processing (OLTP) model, or the On Line Analytical Processing (OLAP) model.

THE ON LINE TRANSACTION PROCESSING (OLTP) MODEL

The OLTP model is considered "the tool of choice" for the automation of high-volume, repetitive business processes. This model is often comprised of a system where transactions might include many short records, with minimal amounts of data. Good examples of an OLTP system include airline reservation systems, grocery store check out systems, credit card authorization systems, and even ATM withdrawals.

The OLTP system model utilizes a traditional centralized host computer, and a private network, with the private network being accessible by clients.

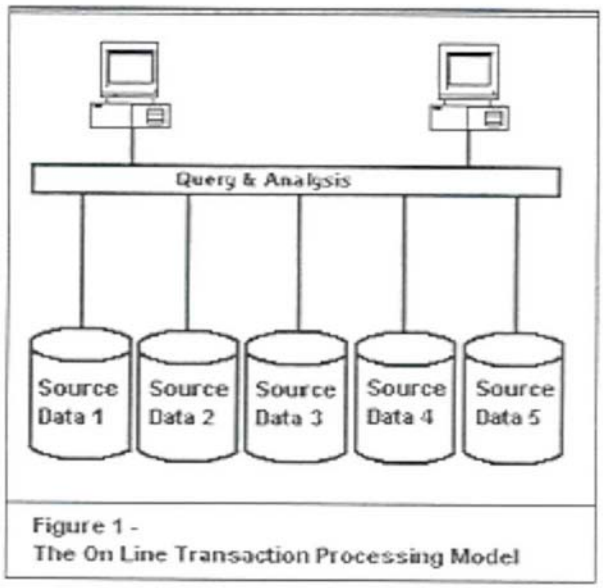

Figure 1 -
The On Line Transaction Processing Model

Brian Black, in his article "OLTP on the Internet"[1] defines the following four specific attributes required in any OLTP system:

- Continuous availability: round-the-clock access with very little downtime
- Predictability: major transaction response times do not vary significantly with time of day, with season of the year, or during any heavy, peak processing period
- Transaction integrity: the transaction processed by the system is immediately reflected in reality; when you make airline reservations and request a certain seat, you will get that seat
- Transaction security: the whole system is secure from unauthorized entry

The OLTP model is very different from the concept of a Data Warehouse; however, both the OLTP model and the Data Warehouse model are highly concerned about Data Accuracy. The difference is at the detail level and the timeliness of the data.

THE ON LINE ANALYTICAL PROCESSING (OLAP) DATA MODEL

The On Line Analytical Processing (OLAP) model is based on data reporting. This model also utilizes a traditional centralized host computer and a private network, with the private network being accessible by clients.

[1] Black, Brian, October 1996, "OLTP on the Internet." Internet Systems Magazine

Unlike the OLTP model, the OLAP model brings data together, from different sources, and allows users to combine data and create reports. The OLAP Council, on their web site[2], defines the OLAP functionality as a multi-dimensional system that allows users to work with data across many systems and to provide information to enhance navigational and analytical activities, including:

- Calculations and modeling applied across dimensions, through hierarchies and/or across members
- Trend analysis over sequential time periods slicing subsets for on-screen viewing
- Drill-down to deeper levels of consolidation in order to reach-through to underlying detail data
- Rotation to new dimensional comparisons in the viewing area

There are 2 major drawbacks to an OLAP system. The first is the size limitation. OLAP systems typically cannot exceed I OGB of input data, and 500,000 members in any single dimension.[3] This size limitation greatly limits the ability of an OLAP system to be used in arenas with large amounts of source data.

The second major drawback is related to basic design functions. OLAP products do not work with SQL. SQL is the industry standard language used to query data-bases. This effectively limits the tools that users can use, and forces users to use a single supplier for all aspects of their data system.

DEFINING THE DATA WAREHOUSE

The Data Warehouse is a hybrid, combining the best of the OLTP and OLAP systems. A Data Warehouse acts pretty much like its name would imply. The Data Warehouse is used to house, or store, "used" data. The Data Warehouse also anticipates where and how the stored data will be used. The Data Warehouse also contains a set of "tools" that allows users easy access to the data.

In his book "The Data Warehouse Toolkit -Practical Techniques for Building Dimensional Data Warehouses,"[4] Ralph Kimball identifies six goals of a Data Warehouse.

1. The Data Warehouse MUST provide easy access to the data.

---

[2] http://www.olMcouncil.ora/research/izlossary.htm

[3] Kimbal, Ralph (2000), The Data Webhouse Toolkit - Building the Web-Enabled Data Warehouse. N.Y.: John Wiley & Sons Inc.

[4] Kimbal, Ralph (1996), The Data Warehouse Toolkit - Practical Techniques for Building Dimensional Data Warehouses. N.Y.: John Wiley & Sons Inc.

2. Data in a Data Warehouse must be consistent.

3. The data in a Data Warehouse must be able to be separated and/or combined in any possible manner.

4. The Data Warehouse is not just data, but is also a full compliment of tools allowing users to query, analyze, and present information.

5. The Data Warehouse is where used data is published.

6. The quality of the data in the Data Warehouse is a driver of business reengineering.

PLANNING THE DATA WAREHOUSE

A successful data warehouse begins with planning. Initially, the most important planning decisions concern the following three user requirements:

1. What information does the end user need?

2. How often is data retrieval desired?

3. Where is that information currently stored?

Once user requirements have been defined, the remaining planning tasks are similar to any other data system, with special emphasis on establishing a "fault tolerant," redundant data system. Key elements in data warehousing are consistency, and the availability for quick and accurate data retrieval.

THE MECHANICS OF HOW A DATA WAREHOUSE OPERATES

All Data Warehouses operate along the same basic process:

1. Data from multiple sources is brought to a centralized repository.

2. Data is then manipulated, either manually or through custom software, to develop a consistent format.

3. The data is then moved to the Data Warehouse, either manually or through custom software.

4. Some of that data will be moved, either manually or through custom software to populate special summary tables. These summary tables are used to quickly build common client reports.

5. All of the data in the Data Warehouse represents a "snap-shot" of the data at a specified point in time.

6. The Data Warehouse is now "Open for Business" and is ready for users to query and access data.

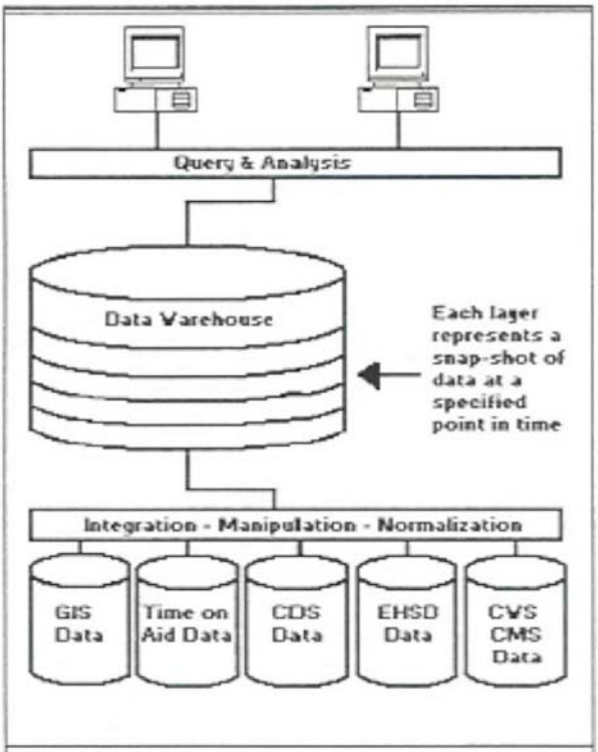

Figure 2 -
The Contra Costa County Data Warehouse Model

THE CONTRA COSTA COUNTY EXPERIENCE

Contra Cost County consists of 13 cities, with a population of almost 985,000. Employment services and social services in this county are handled by the Employment & Human Services Agency. This agency maintains a decentralized Information Services department. John Zimmerman (BASSC Graduate, Class of 2000) manages this department and was the prime reason I chose this county for my case study.

Not only is John one of the most highly respected "computer guys" involved with county government, he is also the founder and director of the County Oracle Groups (COG).

COG is a group of county representatives that meet monthly to share information on data warehousing and other Oracle applications.

Contra Costa County faced increasing requirements for more data reporting, and analysis. A review of their existing systems showed that they were using Microsoft Access as the database tool of choice. They were using this tool to assemble over 100 monthly reports from data out of various State systems. While Microsoft Access is a good desktop type database product, it was not designed to handle such large data sets. As the number of requests for more data grew, it became apparent that a much more sophisticated system was needed, and that a Data Warehouse was the next logical step.

The county formed a group to investigate and evaluate large database systems. It was decided that the ideal product would, not only power the proposed Data Warehouse, but also serve as the database engine in other applications. It was also important that the selected software have, not only the necessary tools to develop databases and reporting systems, but a supportive third party "following." This third party support might prove necessary for assisting the county in development areas. The selected software would also have to integrate with the upcoming Cal-WIN project.

The group evaluated several alternatives and selected the Oracle product from Oracle Corporation. The Oracle product was determined as the only product that would meet all of their requirements. A sole source letter was generated, and the Oracle license acquired. Additionally, Contra Costa County hired an Oracle Consulting Team from the Oracle Corporation.

The Consulting Team hired by Contra Costa County was comprised of four Oracle employees. Their job was threefold. First, they were to recommend the hardware and operating system platform on which to run the Oracle database. Second, they had to get the database installed, define summary tables, and create usable reports. Thirdly, the Consulting Team was supposed to transfer the technology to Contra Costa County employees, so that they could continue to maintain the system, once the Consulting Team was gone. The Consulting Team experienced difficulty almost immediately. They had problems installing and maximizing the main Oracle application, and it was not as easy to install the Oracle product on an IBM based server as everyone had expected. The initial installation was plagued with problems and required several "re-installs."

The Consulting Team also had trouble in identifying and understanding the different data types, and the data streams that they received. This lack of understanding caused the Consulting Team to prematurely recommend expansion of the data storage hardware associated with their Storage Area Network (SAN).

The most infuriating failure of the Consulting Team, however, was their lack of ability to complete the project. Their lack of understanding of "Government Type" data and systems caused them to proceed slowly. Lack of understanding associated with this specific knowledge forced the Consulting Team to spend too many resources in setting up and initializing the project.

The contract with the Consulting Team was funded upon a "Time + Materials" basis. Most Counties find that this saves money, as the County only pays for services they use. In an ideal situation, work is completed before spending all of the associated funds. In this case, funds for the contract were used up before the Consulting Team could complete this project. The Consulting Team also failed in successfully transferring the technology and the knowledge to the Contra Costa County employees. The knowledge and understanding of Oracle software and Data Warehousing technology, needed to be taught to County employees. These would be the personnel responsible for the maintenance and continued development of the system.

The Consulting Team was successful in several areas. They were able to guide the County in their selection of hardware and operating system platform. While they did experience some difficulties, they were eventually able to convert those difficulties into a successful installation. They were also very successful in helping to define user requirements, and in guiding the county to envision the completed project.

When funding for the Data Warehouse Project ran out, prior to completion, it became apparent that the Oracle Consulting Team had left Contra Costa County in a very bad position. The County was now heavily invested in this project, yet they were still a long way from finishing the data warehouse. Abandoning the project was not an option.

Contra Costa County knew that they needed to make their Data Warehouse work. They also knew that they needed to educate their own employees on the technology needed to facilitate ongoing system development and maintenance.

Realizing that they lacked the skills internally, Contra Costa County again looked to an outside consultant. To move forward, they needed someone to serve as a Project Manager for the Data Warehouse project. A Project Manager was found and he immediately began to plan how he would build his team.

The first step was to decide the roles and responsibilities of the future team members. In order to proceed in a timely manner, and to ensure the transference of knowledge, it was decided that the new team would consist of four Contra Costa County employees, and four outside consultants.

Figure 3. Functional Assignments & Percentage Break Outs								
	Larr	Ken	Wan	Sandi	Alana*	Mike	Sal	Teri
Technical Architect Database Administrator	40					100		
Warehouse Developer	20	25	10					
Reports Developer	15	40			25			
Lead Reports Developer		10						
Training Specialist								85
Project Manager							85	
Project Administrator				100				
Operations Manager			15					
Data Modeler	10		15					
Requirements Analyst	15	25			50		15	15
QA/QC Manager			60					
Totals	100	100	100	100	76	100	100	100
" Part Time Employee								

Once the team had been assembled, the Project Manager created a "Functional Assignment Chart (see figure 3)." This chart spells out the different Data Warehouse tasks and how these tasks were to be spread between team members. The chart was used to ensure that all "players" had a chance to build upon their Oracle and Data Warehousing expertise. It was also utilized in documenting "technology transfer."

The Data Warehousing project is nearing completion, and it is believed that the project will be turned over to Contra Costa County employees by June 2002. At this point, enough understanding of the associated technologies will have been transferred.

Once completed, the Data Warehouse will allow Contra Costa to automatically generate standardized reports, or create "ad-hoc" reports, in an unprecedented manner. They will be able to combine data from GIS, CDS, CWS/CMS, EHSD, and Time on Aid data. It will then be easier to analyze all aspects of social services within Contra Costa County.

LESSONS LEARNED

Contra Costa County was one of the first local counties to consider building a Data Warehouse. They have graciously shared their experiences and "growing pains" with me on this project. I learned more than I had expected to, however my greatest insights into this project were not what I had initially expected.

The greatest insights were not derived from the technical issues so much as the "people" issues involved with Data Warehousing.

The initial Consulting Team hired by Contra Costa County, was expected to perform at an extremely high level of competency, yet they did not perform at anywhere near that level. Their lack of experience dramatically affected the project. The four-man team, all employed as "experts" from the Oracle Corporation, experienced great difficulty in even installing the basic components associated with their company's product.

This showed me that there is no such thing as a standard, easily installed Data Warehouse application. The education and knowledge of the people involved with installing the software is extremely important to the building of a strong foundation for successful implementation.

The other great "people related" insight involves the transfer of technology. Contra Costa County was at the verge of a crisis. They had spent large sums of money on the purchase of new hardware, new software, and even on bringing the "experts" into the project. When the "experts" left the project, Contra Costa County had an unusable system and no internal knowledge about how to complete the project.

Contra Costa realized that, for this project to be successful, there would have to be some kind of technology transfer. Their own employees must be able to continue with the development and maintenance of this very complex system. The only way to guarantee the viability of such a large project was through this transference of knowledge.

RECOMMENDATIONS FOR SANTA CRUZ COUNTY

The Human Resources Agency (HRA) in Santa Cruz County has planted the seeds for their own Data Warehouse. A newer version of the Oracle database engine has been successfully installed and is running on a Linux based system. Several non-data warehousing database applications are in the process of being tested, and are being prepared for full implementation.

My first recommendation to the Director of Administrative Services, and to the Manager of Information Technologies, is aimed at what must be defined as a major concern with the Oracle / Data Warehouse Project in Santa Cruz County.

Santa Cruz County has only one person assigned full time to their Oracle /Data Warehouse project. While that one person is proficient at what he does, there is no backup staff for this person. My recommendation is to immediately begin training existing personnel in Oracle and Data Warehousing technologies. It is extremely important to have more than one person

understand the technology, in order to sustain the development and maintenance of the system in case of personnel changes.

My second recommendation is an out growth from the "Functional Assignment" chart created by the Contra Costa County Data Warehouse Project Manager. This chart is an excellent way to document the transfer and sharing of technology. I recommend that Santa Cruz County review their existing Information Technologies Department, and create a similar chart.

Recently, the HRA Information Technologies Department in Santa Cruz County lost the ability to generate the maps and charts that utilized the Arc View GIS software. The only person who understood this complex system recently went on maternity leave. For the last 8 months or so, the county has not had the ability to use this software. A program that developed cross training, or that encouraged technology sharing, may have prevented this loss.

EPILOGUE

Several major changes have already been implemented in Santa Cruz County as a result of this Internship Project.

Santa Cruz County also utilizes outside consultants to aid in their Oracle software applications. To prevent the type of mistake experienced by Contra Costa County, special care is given to ensure that all consultants are highly monitored. Consultants are only brought in for well-defined tasks, and are paid only when they have completed these tasks.

The Data Management Group has taken steps to ensure that staff are sharing technologies. Specifically, in order to expose other staff members to the Oracle Software and to data warehousing technologies, the Data Management Group has promoted a depart-mental Data Processing Coordinator in the Client Applications department, to an Information Systems Analyst in the Data Management Group. In his new position, this analyst will be exposed to the Oracle software, and to data warehousing technologies.

The Data Management Group has also begun to share intemet technologies. Previously, one staff member had been responsible for all aspects of the Agency's web sites and Intranet. A second staff member is now sharing the responsibility for web site creation and maintenance, and there are plans for another staff member to help with these tasks.

The Data Management Group is also on the road to re-establishing their ability to work with the Arc View GIS software. A staff member has been assigned to work on this project part time.

While a formal Functional Assignment Chart has not yet been created, the seeds have been planted. I expect a continuance in the sharing of technologies within the Santa Cruz County Human Resources Agency.

ACKNOWLEDGMENTS

I would like to extend my appreciation to the Contra Costa County C3BIS Team for their help and especially their patience in my education on Data Warehousing I would also like to extend a special "Thank You" to both Sal Bruno, Project Manager of the Contra Costa County Data Warehousing Project, and to John Zimmerman, the Information Systems Manager for Contra Costa County. John graciously facilitated my internship, and allowed me the freedom to work with his staff and to interrupt their already hectic schedules.

I would also like to thank those in Santa Cruz County who believed in me enough to send me through the BASSC Program, and who allowed me to juggle my assignments so that I could participate. I may have originally gone "kicking and screaming" but this program has been one of the best learning experiences I have ever had.

Last, but not least, I wish to thank the BASSC instructors and staff. There is no way that I can begin to describe how the program has changed my life. I have learned many things that were quite unexpected. I learned how to prioritize the things I need to do. I learned how to effectively evaluate new systems and ideas. I learned how my peer groups perceive me, and most importantly, I changed the way that I perceive myself. THANK YOU!