Data Warehouse Efforts and Metadata Foundations

Charles R. Thomas
Senior Consultant
NCHEMS

ABSTRACT

In 1997 a survey of 40 institutions with active data warehouse projects was summarized for a paper entitled "Information Architecture: The Data Warehouse Foundation". Seven years later many more institutions have data warehouse efforts in production, in development or planned. This paper summarizes the state of data warehouse projects in higher education institutions and the underlying metadata supporting those efforts.

INTRODUCTION

A very small number of institutions had data warehouse efforts underway in 1997, and most of those had little, if any, executive sponsorship. Seven years later many institutions have data warehouse projects, and many of the commercial ERP software products include data warehouse options. To assess the status of data warehouse projects in colleges and universities, the chief information officers of colleges and universities were sent a five-question survey concerning data warehouse efforts on their campus. They were asked if they had a data warehouse effort in production or planned, what data was included, and the executive sponsor. They were also asked about their institution-wide data dictionary. Appendix A contains a copy of the electronic mail survey questions sent in March 2004. Responses were received from almost 300 institutions, and this paper summarizes the results. Smmaries of the responses are detailed in Appendix C.

DATA WAREHOUSE DEFINITION

Some years ago Bill Inmon, a data warehouse industry expert, published the following definition:

> *"A Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process."*

Corporate data warehouses typically capture every detail of every transaction in every operational system into a massive tera-byte database. In fact, industry publications are beginning to discuss peta-byte sized databases. A peta-byte is a quadrillion bytes, shown as "1,000,000,000,000,000" bytes. (See Appendix B for a chart of really large and really small

numbers.) By comparison, most college and university data warehouses are much more modest in size. As Dennis Jones, NCHEMS president points out, "In higher education policy deliberations, ten percent of the data will usually answer ninety percent of the executive questions".

One major reason is the significant difference between the basic operational systems in the corporate environment and those in higher education.

	Corporate	*Academic*
Application Systems	Few	Many
Transactions	Many	Few
Mission	Clear	Hazy
Management Style	Autocratic	Collegial

Figure 1: Corporate versus Academic Information Systems

Most companies have a relatively small number of major application systems, while even a small college might have fifty to one hundred different application systems. In the corporate environment a small number of application systems will process billions of transactions daily, while in the academic environment the general accounting application might process a couple of million transactions in an entire year. As pointed out in an earlier paper, the juxtaposition of the number of systems and the number of transactions creates a very different set of information economics. These differences, when combined with the significant differences in both mission and management style between corporations and higher education institutions, have resulted in a less than integrated institutional information architecture at most colleges and universities. The same comparison illustrates the difference in complexity of data warehouse projects between the corporate and academic environments.

Corporate data systems typically operate on well-defined and common operational schedules. Since the data warehouse can be a critical component in client relationship management (CRM), it can be important to have it kept up-to-date on-line as the transactions occur. Most higher education data warehouses are populated on natural cycles that differ between application areas. Admissions and student data are usually transferred to the data warehouse at the beginning and end of terms, while financial data are transferred at the end of the month and annually.

It should also be pointed out that a data warehouse does NOT replace the need for electronic archives for all of the operational data within the many applications in higher education. Data archiving is a separate problem with its own set of requirements. As data enters a data warehouse it usually undergoes a "normalization" process that makes it more readable and makes it consistent from a longitudinal perspective. Because definitions and codes change over time, archived data will typically have the metadata for the same time period stored in the same place, and data from one time period may not match that from another. In fact, many times the formats may differ from one time period to another. Also, archival data is usually stored off line on a media such as compact disc, while recent (five years) data in the data warehouse is usually online and available for ready access by administrators.

The technical perspective introduces further complexity. The database structures that are efficient for transaction processing are very different than those that are efficient for data access and extraction. This situation adds complexity to the data normalization process as information is moved from the operational systems into the data warehouse, but provides justification for maintaining the data warehouse separate from the active data files.

One interesting aspect of data normalization we discovered involves the term "disambiguation". This is "the process of resolving the conflict that occurs when the names of two or more items have the same natural title." The example of disambiguation cited is the word "mercury;" is it the planet, the element, the model of an automobile, the NASA space project, or the Roman god? A more mundane example in higher education might be the disambiguation of a departmental abbreviation such as "ENG;" is it the English department or the Engineering department?

METADATA DEFINITION

MetaData is information about data. At a minimum it includes identification, definition, uses, sources, field values categories and descriptions, and linking information. It may also include information concerning when it was last updated, and by whom. Some metadata systems also include technical information about each data element, including such dimensions as field size and type. Masks for formatted printing may also be included, as well as short titles for report titles.

Some elements of metadata may adhere to standards set by external organizations such as government agencies or national standards bodies. Other elements will be determined by the individual organization. In any organization, it is important to establish an official forum for coming to agreement on the definitions, codes, categories, and descriptions maintained in the metadata files. MetaData is typically maintained by a central staff and made available to the rest of the organization on the Intranet.

DATA WAREHOUSE EFFORTS IN HIGHER EDUCATION

Seven years ago only a quarter of the EDUCAUSE members reported data warehouse efforts of any kind. In our 2004 survey, 75% of all responding institutions reported data warehouse efforts in production, in development, or planned. Significantly, only 1% of the responding universities and 4% of the responding large institutions report "no plans" for data warehouse efforts.

Public institutions reported more data warehouse efforts than private, and larger institutions report more efforts than small institutions. The responding universities also report more efforts that either four-year or two-year institutions.

DATA WAREHOUSE DATA

In response to the question about the specific data included in the data warehouse, responding institutions reported Student data most of the time. Financial data was the second most reported area, with Human Resources data a close third. In the Other category, Financial Aid, Facilities, and Grants were also reported frequently.

Only about 10% of the responding institutions reported data from four or more applications included in their data warehouse, and many of those listed five or six different areas. Another 15% of the institutions reported data from three applications included in their data warehouse, and all of those included both Student and Financial data. A quarter of the responding

institutions reported the inclusion of data from two applications, and over 80% of those included only Student and Financial data. Another quarter of the responding institutions reported that they included data from a single application, and most of those include either Student or Financial data only.

EXECUTIVE SPONSORS OF THE DATA WAREHOUSE

In all categories of the responding institutions, the Chief Information Officer is most reported as the executive sponsor of the data warehouse. Only the responding Universities report the sponsoring executive to be the Chief Academic Officer or the Chief Financial Officer in any significant numbers. Only a few institutions reported more than one executive sponsor for their data warehouse effort. This does create a perceptual problem in some institutions that the data warehouse is a "techie" effort that really doesn't have a lot to do with institutional policy decisions.

DATA DICTIONARY

It is surprising that even though a majority of the responding institutions have a Data Warehouse effort underway, less than half of them have a formal institution-wide data dictionary in place, under development or planned. Analysis of the survey responses indicates that the public, larger, and more complex institutions are more likely to have a formal Data Dictionary, with two-thirds of the smaller institutions reporting no plans.

The Information Technology department usually maintains the institutional Data Dictionary, with Institutional Research reported as a distant second. Planning and other departments were barely mentioned.

While it was not covered in the survey, the key to an effective institution-wide Data Dictionary is the forum in which agreement is achieved on definitions. This forum is usually an administrative systems advisory committee with permanent representatives of the major administrative offices, and rotating members representing academic departments and other levels of administration. The functioning of this forum is more important than the department responsible for the maintenance.

DATA DICTIONARY SOFTWARE

Oracle is clearly the dominant choice for software used to maintain the institutional Data Dictionary of MetaData. Over half of the responding universities use Oracle, with"Other" a close second, and Microsoft SQL representing half of that entry. Microsoft Access is the only other software mentioned by any significant number of institutions, particularly the 2-year institutions. Most of the institutions have developed their own Data Dictionary application software using the local database system.

The second edition of the CHESS CD, *Data Definitions for Colleges and Universities* has just been released through NCHEMS, and this edition includes "free-license" web-based software for building and maintaining an institution's Data Dictionary. It provides a means for institutions to integrate data from disparate operational systems into the Data Warehouse. The Access database is a starting point for an institutional Data Dictionary, with metadata for over 780 data elements, 120 Excel tables of field values, and a model Taxonomy of Activities. Details are available form NCHEMS at www.nchems.org.

CONCLUSIONS

While the 282 institutions who responded to the NCHEMS survey might be biased towards those who do have Data Warehouse efforts underway, clearly there are many more developments in progress now than there were seven years ago. From institutional descriptions of their Data Warehouse efforts, it is also clear that colleges and universities take a very different approach than companies. The corporate Data Warehouse is usually much more comprehensive, including all transactions in all operational systems. Colleges and universities tend to adopt an incremental approach and be selective about what data are included. It appears that the selection of what data are included is not directly related to the interests of the executive sponsor as one might expect, but is more related to size; larger institutions include more areas than others.

APPENDIX A

Data Warehouse & MetaData Email

Email to: Chief Information Officers
From: Chuck Thomas
Subject: Data Warehouse & MetaData

We are updating our 1997 NCHEMS paper on Data Warehouse efforts in Higher Education for a presentation at the 2004 EDUCAUSE-Midwest Conference and would appreciate you or someone on your campus answering five questions by return email. We will provide a summary of the results to all respondents.

1. Do you have a Data Warehouse effort?
   [ ]in production [ ]in development [ ]planned [ ]no plans.

2. What data is included in your Data Warehouse?
   [ ]Student [ ]Financial [ ]Alumni [ ]Other:_____

3. Who is, or will be, the executive sponsor of your Data Warehouse effort?
   [ ]Chief Executive Officer [ ]Chief Academic Officer
   [ ]Chief Financial Officer [ ]Chief Information Officer
   [ ]other:_____.

4. Do you have a formal institution-wide Data Dictionary of MetaData?
   [ ]yes, maintained by: [ ]IT [ ]Planning [ ]Instnl Research [ ] Other
   [ ]in development or planned
   [ ]no.

5. What software do you use, or plan to use, to maintain your Data Dictionary?
   [ ]Microsoft Access [ ]Oracle [ ]other database system:_____.

You may be interested in looking at the second edition of the new CHESS Data Definitions for Colleges and Universities with the free-license web-based MetaData Administrator Software at www.nchems.org.

Thanks,
Chuck Thomas, Senior Consultant, NCHEMS

APPENDIX B

Really Large and Really Small Numbers

REALLY LARGE NUMBERS

Prefix	Power of 10	Units	Number
kilo-	3	thousands	1,000
mega-	6	millions	1,000,000
giga-	9	billions	1,000,000,000
tera-	12	trillions	1,000,000,000,000
peta-	15	quadrillions	1,000,000,000,000,000
exa-	18	quintillions	1,000,000,000,000,000,000
zetta-	21	sextillions	1,000,000,000,000,000,000,000
yotta-	24	septillions	1,000,000,000,000,000,000,000,000

Downloading a one gigabyte file using a 10mb Internet connection would take about 17 minutes

Downloading a one-yottabyte file using a 10mb Internet connection would take over 30 billion years!

REALLY SMALL NUMBERS

Prefix	Power of 10	Units	Number
milli-	-3	one thousandth	.001
micro-	-6	one millionth	.000001
nano-	-9	one billionth	.000000001
pico-	-12	one trillionth	.000000000001
femto-	-15	one quadrillionth	.000000000000001
atto-	-18	one quintillionth	.000000000000000001
zepto-	-21	one sextillionth	.000000000000000000001
yocto-	-24	one septillionth	.000000000000000000000001

It takes 300 zeptoseconds for light to pass over a simple atom!

Source: National Institute of Standards and Technology,
quoted by Steven Beck in the New York Times.

APPENDIX C

Data Warehouse Survey Results Tables

	ALL INSTNS			BY CONTROL			
	All	*Percent*		*Public*	*Percent*	*Private*	*Percent*
Count	282			191		91	
Q1: Data Warehouse Effort							
1.1 In Production	102	36%		74	39%	28	31%
1.2 In Development	59	21%		41	21%	18	20%
1.3 Planned	54	19%		34	18%	20	22%
1.4 No Plans	70	25%		42	22%	28	31%
Q2: Data Warehouse Data							
2.1 Student	188	67%		137	72%	51	56%
2.2 Financial	141	50%		101	53%	40	44%
2.3 Alumni	50	18%		23	12%	27	30%
2.4 Other	65	23%		50	26%	15	16%
2.5 Human Resources	41	15%		34	18%	7	8%
Q3: Executive Sponsor							
3.1 Chief Executive Officer	17	6%		13	7%	4	4%
3.2 Chief Academic Officer	24	9%		20	10%	4	4%
3.3 Chef Financial Officer	23	8%		14	7%	9	10%
3.4 Chief Information Officer	130	46%		93	49%	37	41%
3.5 Other Executive Sponsor	31	11%		21	11%	10	11%
Q4: Data Dictionary							
4.1- Yes	55	20%		34	18%	21	23%
4.1a Maintained by IT	41	15%		25	13%	16	18%
4.1b Maintained by Planning	4	1%		1	1%	3	3%
4.1c Maintained by Instnl Research	10	4%		7	4%	3	3%
4.1d Maintained by Other Office	3	1%		2	1%	1	1%
4.2- In Development or Planned	76	27%		60	31%	16	18%
4.3- No Data Dictionary	149	53%		96	50%	53	58%
Q5: Data Dictionary Software							
5.1 Microsoft Access	22	8%		13	7%	9	10%
5.2 Oracle	93	33%		64	34%	29	32%
5.3 Other Data Dictionary Software	83	29%		60	31%	23	25%
5.4 Microsoft SQL	37	13%		30	16%	7	8%

APPENDIX C

Data Warehouse Survey Results Tables

	BY SIZE							
	1. Large 18,000 and over	*Pct*	*2. Med-Large 8,000 to18,000*	*Pct*	*3. Medium 2,000 to 8,000*	*Pct*	*4. Small less than 2,000*	*Pct*
Count	53		64		118		47	
Q1: Data Warehouse Effort								
1.1 In Production	33	62%	26	41%	30	25%	13	28%
1.2 In Development	12	23%	18	28%	21	18%	8	17%
1.3 Planned	6	11%	12	19%	28	24%	8	17%
1.4 No Plans	2	4%	9	14%	39	33%	20	43%
Q2: Data Warehouse Data								
2.1 Student	48	91%	51	80%	67	57%	22	47%
2.2 Financial	34	64%	38	59%	56	47%	13	28%
2.3 Alumni	7	13%	10	16%	24	20%	9	19%
2.4 Other	24	45%	20	31%	18	15%	3	6%
2.5 Human Resources	17	32%	16	25%	6	5%	2	4%
Q3: Executive Sponsor								
3.1 Chief Executive Officer	5	9%	4	6%	5	4%	3	6%
3.2 Chief Academic Officer	7	13%	7	11%	7	6%	3	6%
3.3 Chef Financial Officer	6	11%	7	11%	9	8%	1	2%
3.4 Chief Information Officer	34	64%	31	48%	50	42%	15	32%
3.5 Other Executive Sponsor	3	6%	8	13%	14	12%	6	13%
Q4: Data Dictionary								
4.1- Yes	13	25%	13	20%	18	15%	11	23%
4.1a Maintained by IT	10	19%	9	14%	13	11%	9	19%
4.1b Maintained by Planning	0	0%	0	0%	3	3%	1	2%
4.1c Maintained by Instnl Research	4	8%	3	5%	2	2%	1	2%
4.1d Maintained by Other Office	0	0%	1	2%	2	2%	0	0%
4.2- In Development or Planned	16	30%	27	42%	28	24%	5	11%
4.3- No Data Dictionary	24	45%	24	38%	70	59%	31	66%
Q5: Data Dictionary Software								
5.1 Microsoft Access	5	9%	3	5%	11	9%	3	6%
5.2 Oracle	22	42%	30	47%	31	26%	10	21%
5.3 Other Data Dictionary Software	21	40%	20	31%	29	25%	13	28%
5.4 Microsoft SQL	9	17%	10	16%	12	10%	6	13%

APPENDIX C

Data Warehouse Survey Results Tables

	BY TYPE					
	1. Univ	*Percent*	*4. 4-Yr*	*Percent*	*2. 2-Yr*	*Percent*
Count	67		118		97	
Q1: Data Warehouse Effort						
1.1 In Production	40	60%	30	25%	32	33%
1.2 In Development	19	28%	26	22%	14	14%
1.3 Planned	9	13%	27	23%	18	19%
1.4 No Plans	1	1%	36	31%	33	34%
Q2: Data Warehouse Data						
2.1 Student	58	87%	73	62%	57	59%
2.2 Financial	50	75%	56	47%	35	36%
2.3 Alumni	15	22%	30	25%	5	5%
2.4 Other	33	49%	20	17%	12	12%
2.5 Human Resources	22	33%	11	9%	8	8%
Q3: Executive Sponsor						
3.1 Chief Executive Officer	6	9%	7	6%	4	4%
3.2 Chief Academic Officer	12	18%	8	7%	4	4%
3.3 Chef Financial Officer	14	21%	5	4%	4	4%
3.4 Chief Information Officer	35	52%	51	43%	44	45%
3.5 Other Executive Sponsor	5	7%	14	12%	12	12%
Q4: Data Dictionary						
4.1- Yes	16	24%	28	24%	11	11%
4.1a Maintained by IT	13	19%	20	17%	8	8%
4.1b Maintained by Planning	0	0%	4	3%	0	0%
4.1c Maintained by Instnl Research	3	4%	4	3%	3	3%
4.1d Maintained by Other Office	1	1%	2	2%	0	0%
4.2- In Development or Planned	26	39%	26	22%	24	25%
4.3- No Data Dictionary	25	37%	64	54%	60	62%
Q5: Data Dictionary Software						
5.1 Microsoft Access	4	6%	8	7%	10	10%
5.2 Oracle	39	58%	42	36%	12	12%
5.3 Other Data Dictionary Software	18	27%	29	25%	36	37%
5.4 Microsoft SQL	6	9%	10	8%	21	22%