

IT Executive Exchange

Data Modeling, Quality and Security

Moderator: David Hutchins, Keithly Corporation

Executive Summary

Our wide-ranging discussion covered “a day in the life of data warehousing,” and was far more extensive than can easily be summarized. Topics and insights included:

- * **Data Modeling and Off-the-Shelf Packages:** The main problems with data modeling include how comprehensive the scope of the project should be, and getting various different constituencies to agree. There are pros and cons of the bottom-up approach using Data Marts and the top-down approach using a single Enterprise Data Warehouse. Using the data model of an off-the-shelf package can be useful, but should be seen as a starting point.
- * **Governance and Re-engineering:** Strong governance is required, both for the project itself and to deal with issues that may arise with more fundamental business processes. Cross functional teams must be empowered by business ownership of the project. Data warehouse projects may reveal business process weaknesses, and so may entail business process re-engineering.
- * **Quality:** Data warehousing projects are worth little unless the underlying data is accurate and comprehensive. Data quality should be considered well before the data are to be loaded. Early use of a profiling tool can save time and effort later.
- * **Globalization:** Globalization presents particular problems, especially with the need to have multiple translations in the enterprise-wide data warehouse.
- * **End Users:** Recognizing different types of end users and providing appropriate training is a must.
- * **Security:** Although sometimes difficult to administer, role-based security allows necessary distinctions deep into the data.

The IT Executive Exchange (ITEE) is a group of IT Executives and College of Business Administration professors at The University of Akron that meets about every six weeks to discuss pressing and leading edge IT issues faced by IT executives. The purpose of this forum is to have a healthy exchange of ideas that will be useful to all attendees. It is sponsored by the Center for Information Technologies and eBusiness (CITe) of The University of Akron's College of Business Administration. For previous topics and summaries, refer to <http://cite.uakron.edu>

This summary was prepared by Prof. William McHenry, CBA, The University of Akron.

COMPLETE SUMMARY OF THE SESSION

PARTICIPANTS.....	2
GETTING STARTED WITH DATA WAREHOUSING	3
INFLUENCE OF A PACKAGE	4
GOVERNANCE	5
DATA WAREHOUSES & BUSINESS PROCESS REENGINEERING	6
END USERS	7
DATA QUALITY.....	7
GLOBALIZATION.....	9
SECURITY	10
NEXT IT EXECUTIVE EXCHANGE.....	12

PARTICIPANTS

We had a relatively small but rather intense group for this session. One set of participants were from companies that are accumulating more and more data at present, and are looking to take the next steps in data management, whatever they may be. The amount of data in these firms is approaching “critical mass.” A big challenge is how to take all the available data and make sense of it using data warehousing, analytical tools (analytics), and business intelligence. Too much aggregation can render the data less useful, so how does one pinpoint where the value will come from within all this data?

Other firms have a longer history of data management, but are now facing problems brought on by globalization. One \$1B firm, for example, has a centralized data warehouse (running Oracle OnDemand products) that feeds global sales (70% of which are done outside the U.S.). There are also local in-house reporting requirements that have to interact with that environment. A persistent problem is localizing data representations into equivalent terms in local languages, ensuring enforceable data standards around the world, making it possible to share customer information across markets, providing a sufficient analytical model, guarding intellectual property, and ensuring the necessary levels of security for all of this. In this case the CRM module is in the forefront, with the desire to share customer information, contact information, transaction information, etc. What is the “information model” to provide this?

Another firm faces typical problems of growth with mergers and acquisitions. For example, there are now numerous formats for customer numbers throughout various divisions, and this firm needs a better data integration model. Integration problems may also be present in firms using ERP, CRM, and other enterprise packages via which data must be stored, analyzed, and shared. The data warehouse (DW) implementation has to be done so that customer expectations for performance are met, but at the same time the load on production systems should be minimized.

Two participants had extensive experience in master data management, customer data integration, and Data Warehouse/Business Intelligence consulting, bringing to the table a

lot of experience with various clients in and out of Ohio. After some introductions, we decided to structure the discussion as “a day in the life of data management,” which is roughly how the following sections are organized.

GETTING STARTED WITH DATA WAREHOUSING

A typical problem that firms face when doing enterprise data modeling is determining how comprehensive the effort should be. Should a firm take 18 to 24 to 36 months to agree on all the data representations throughout the firm? This might be the precursor to establishing the enterprise data warehouse (EDW), a single physical or logical representation of all data of interest. The Enterprise Data Warehouse approach was first conceived and publicized by William Inmon, and he still has many “followers.”

However, by the time 18-24-36 months have passed, the users may have moved on to other concerns. Therefore, a somewhat contrasting position is taken by followers of Ralph Kimball. Kimball advocates building separate data marts (DM) that cover specific subject areas of the business. He then ensures that data from these Data Marts can be integrated by “conforming” the dimensions (i.e. made consistent with the same definitions of common elements.) As one of the proponents of this view in attendance remarked, Business Intelligence (BI) is a journey, and data warehousing is the road you take to get there. You may never “arrive,” but you keep expanding and building as you go. This participant focuses on subject areas. About three hands went up when I asked how many people are followers of Kimball, and none went up when I asked about Inmon.

An example of problems at this level: finance builds a data mart that has a “bill-to” view of customer sales, but marketing wants to look at the “sell-to” version of the customer dimension. Rather than dealing with the conformity issues so that marketing can make use of the common fields from finance, marketing will build its own separate data mart. One participant said he has seen this happen many times. How then can an enterprise-wide customer dimension be defined? Then sales and finance can walk into the room with sales figures that does not agree. Or sales pulls data without even realizing that some of it comes from the finance data mart, and the finance people change the definitions, leading to confusion about the data.

Another firm has an integration problem that involves two different lines of business. Until now they have been separate. (Or they have arisen because of acquisitions.) A new product, however, will combine elements of both services, which will create new data flows that go across each division. It is a struggle to figure out how to provide unified data for this process. This firm is discussing new standards across divisions.

How is integration achieved in such an environment? Data modeling is key, whatever approach you use. One solution is to let the individual Data Marts use the local representations (e.g. customer numbers). But then define a certain consolidation model level, above which all consolidations will be done. The work goes into mapping the Data Mart representations into the consolidation model, and then creating higher level Data Marts to

drive these consolidations. One of the participant firms is just now engaged in trying to develop an Enterprise Data Warehouse model on top of the existing Data Marts. They plan to do it subject by subject, with the whole project lasting 3-5 years. First up will be sales and orders. At first the firm had adopted the viewpoint of Kimball, and was working on conforming all the dimensions in the existing data marts. (This participant called Kimball's ideas a "religion"; I likened them to an ideology.) Now, however, this firm is reconsidering in light of a presentation made by Teradata.

Teradata provides hardware and software for data warehousing. The hardware uses massive parallelization, and can support high speeds for 3rd normal form representations of the data. (In a nutshell, this means that entities are represented by one or more tables, where the tables follow certain design rules to minimize redundancy and thereby increase the integrity of the data storage.) In Kimball's worldview, entities are represented mainly by single dimensional tables, which include a lot of redundancy to increase speed of processing. This "star schema" minimizes the number of joins between tables at the expense of more storage, whereas the Teradata machine can supposedly allow you to have your cake and eat it too, i.e. fast performance with minimal redundancy. Teradata makes it possible to view the data as star schemas, but without having to store it that way. All the "rags" say that the hybrid approach is now the way to go. Is there a difference between performance of the physical and logical stars? No, Teradata claims not, because of the massive parallelization.

There is a lot of integration that has not happened in corporations despite the use of ERPs. Why don't the firms use an Enterprise Data Warehouse approach to solve these issues up front? They do try, but firms should approach creating an Enterprise Data Warehouse with caution. Creating one has been said to be harder than solving world hunger.

INFLUENCE OF A PACKAGE

We spent some time talking about the pros and cons of using a package to drive the Data Mart or Data Warehouse project. I gave an example of a firm in which the IT department found a package that seemed to do a lot of reporting within a particular industry vertical that current systems were not doing. This IT director won approval for a project to bring the package on board, and started by asking all of the user groups to map their information needs and current data fields to the data model inherent in the package. Since the package had been sold to—and presumably refined through the experience of—a number of similar organizations, the data model was felt to be comprehensive and worthwhile. However, after the initial efforts were made, the project appeared to be several months behind schedule, with difficulties in getting end user buy-in and not enough personnel to handle a project of this scope.

One participant thought the situation of firms using packages to help them define the data model may be pretty common. Participants reacted to this story by expressing concerns about using a package in this manner. Some felt that it is beneficial to have the data

model of a package like this, but others did not feel that the exercise of conforming to this model would necessarily be best for the firm. Vendors do have viable products. They can be very tempting, but the pre-package data model does not always work. And using a package like this does not necessarily build constituencies for the data model across the firm. The design has to be driven by end user needs. You can't have people just collecting information for the sake of collecting it. Uncertainty, discomfort with the data, and a lack of understanding are exacerbated if the tool is perceived as the solution. The ultimate design may be smaller than the package could support. But if the design is relevant, it is possible to build end-user confidence. Emphasis should be on accuracy, conformity of data, and relevance rather than on using a tool that no one understands. These projects get very big very fast, which is not in the best interests of the business drivers behind them. And then the result can be non-use by the users. The vendors may claim that they bring a pre-packaged data model that makes it possible to conform everything, but this is not necessarily true. There may be someone who is buying this who wants to believe this, but the hard work is non-technical: building the agreements across the constituencies of people who just have inherently different views of the world that have to be reconciled. If the Data Warehouse team takes the project on themselves, end users may look at the end reports, and reject them because the tech people "changed the data." The onus falls on the tech team, which is not where it belongs. This is where the problem was recognized, but not where it belongs.

One participant said that he has experienced numerous instances with Business Intelligence software where the customer thought it would just work right away, off-the-shelf. But it does not work that way. He creates a standard data model with an Enterprise Data Warehouse so that he can evolve it. When he goes to the solution model, he uses those standard packages as design guides. Not to say you have to fit in it. The model is adjusted according to the business drivers. He only uses the model as a starting point. When you get into discussions with people about customers, etc. – they don't know what you are talking about. But if you can use this as a prototype or design guide, this can be helpful. Understanding comes faster.

It is not useful to think of a Data Warehouse design exercise as coming exclusively from the top down or the bottom up. The top drives the WHAT, whereas the bottom drives the HOW. These two approaches have to interact with each other. A cookie cutter approach eventually leads you down the wrong path.

GOVERNANCE

A crucial aspect of establishing the data model is setting up the right governance model from day one. In one participant firm, the governance board owns the data model. This was agreed upon at the start of the project, before they even decided to move to an Enterprise Data Warehouse. This has to be a cross-functional team that is chartered at the beginning. Another participant said that the data model has to be owned by the business. It must include the a good chief architect and a good analyst. The team should have finance,

sales, marketing, and probably operations people at the least. A governance board should define how the data will be used.

DATA WAREHOUSES & BUSINESS PROCESS REENGINEERING

We also spoke about the relationship of data warehousing to business process reengineering. Data warehousing cannot rectify flawed underlying business processes. Indeed, it often uncovers them. (This can be a reason why Data Warehouse/Business Intelligence projects fail.) The differences in data definitions and data needs from one division or group to another shows differences in their processes and should cause the firm to ask why those differences exist. Without establishing these understandings, users of the data marts or data warehouse may have a flawed understanding of what data they are actually getting.

A related question is whether or not doing an ERP first can solve all the data definition problems, vastly simplifying the subsequent implementation of the Data Warehouse. One participant related a quote: you can either conform the data with an ERP or with an Enterprise Data Warehouse. Although correct in theory that you can do it with an ERP, firms do not necessarily use a single image of the ERP everywhere. Acquisitions may also have to be re-platformed. Even single images of ERPs can have differences in various locations. An item code can have thirty characters (everyone has agree to this), but there can be differences in sub-groupings and definitions within the codes. There can be different groupings of the data at various levels in the hierarchy, making it impossible to do customer, channel, or product rollups. So if there are still differences in data coming out of ERPs, the Data Warehouse level may be the only place to reconcile them.

The experience of a smaller organization was related (\$20M sales). Here the CIO brought the heads of each division (sales, marketing, finance, CEO) into a room to define how they were going to measure the business. They started with a sales profitability DM. The smallness of the organization made this feasible. They had trouble keeping all the definitions (conformity) within the ERP. They created audit and integrity checks using an Enterprise Data Warehouse and 3rd-normal form, but built virtual star schemas on top of this to deliver something to users that they would actually use. The Enterprise Data Warehouse serves as the mast model that keeps consistency. Example: the hard and soft integrity checks caught a case where a customer had been assigned by marketing to one channel which was in direct conflict with the policy set by the governance board. There was no check for this in the ERP so it was not caught until the ETL stage. They called in the relevant business units that were responsible for the definitions, and let them duke it out until they reached an agreement about what reality is going to be (the so-called single definition of truth). If you throw this case into “undefined,” then this category tends to grow over time and eventually will undermine the Data Warehouse. Also, when you stop the process because of conformity issues, and the CEO does not get the morning snapshot, then all of sudden conformity because a hot-button issue. Transactional discipline becomes better because they do not want to be caught as the people who blew up the data

warehouse load. This participant advised that if you, as a developer, find discrepancies in the data, stop the project and make the owners decide!

Another example given was a project where SAP-DW could not do a customer profitability application that was “tied to everything.” They had to build a separate DM for this. This was a \$3B company with a POS data warehouse. They did the same thing: took the director of sales, finance, and customer care and said OK, this is the definition of customer, these are the problems we are going to have, and then they duked it out for about three-four weeks until everyone finally agreed on the definition of customer. If you (IT) own the data definitions, the business people will always have you as a crutch or scape-goat for not doing this hard work. Make the business people own the definitions. It’s a business data warehouse.

END USERS

Several participants raised questions about training users to use Business Intelligence tools. A big problem here is making available explanations of the metadata to the users so that they really know what they are looking at. The results depend on the quality of the metadata and the skills / training of the end users. Two people can use the same Data Mart and try to do exactly the same analysis. However, because of different skill levels, they may come out with two different results. This is always an issue, no matter how well done are the front, middle, and tool levels.

There are five-six types of users for Data Warehouse/Business Intelligences. They range from people who just want to see dashboards all the way to people who will be deep data miners. It is useful to make a matrix with the user type on one axis and the subject areas for the Data Marts on the other. Then checks can be placed for each square in which there are users/customers for the Data Mart in that subject area. One tool does not fit all. Cognos or Business Objects may solve some, but then the data miners may need something else.

Ultimately firms like Wal-Mart, etc. are making use of fantastic volumes of data. Transactions that occur in a single store are visible within minutes to the products’ manufacturers. Closed-loop CRM takes the results of data mining and cycles them back into the CRM for additional marketing decisions. How can Data Warehouse/Business Intelligence best be set up to facilitate such information flows?

DATA QUALITY

The ultimate quality of data in a Data Warehouse depends on the quality of the data in the underlying production systems. You can have standard definitions for things, but end users can end up putting different data into different fields anyway. (“You used that field for *what*...?”)

A good practice is to invest in a pre-processing “profiling” tool that will assess the state of the data before doing the data warehouse design. This is a low cost entry into any Data Warehouse or ERP effort. A lot of companies do not understand what this approach can do as a quick hit to save lots of expenditures down the road.

A skilled SQL or Informatica user can examine some of these things because he or she knows what to look for. However, the value in profiling is finding the unexpected. It creates a wide spectrum of output. This participant was skeptical about this at first, and entered into it with “an army” of ETL programmers. They all walked out of it saying there was no way they could build all the needed queries in the available time. The benefit of this profiling shows up when you start marrying the ETL you have built with the data that you finally get from the source systems. All the ETL starts bombing out and you say, this is why it’s bombing out. You could have found that out months prior. At least you could then have gone on record as saying that these are the issues, and they are not all IT’s fault. And they will continue unless you do something about it. Or that data will not be loaded and the value will be diminished for the analytical capabilities. Then you can set reasonable expectations with the business community about what can be achieved. You can load the data with the errors it will produce, or mask it by writing ETL, but that does not solve the reason why it is bad in the first place. These can be weeks’ worth of effort that mitigate months of effort later.

In particular, you will eventually have to design extract, transform, and load (ETL) procedures that specifically connect the production systems to the data warehouse. It is often only when the Data Warehouse has been designed and the ETLs are being developed that it becomes apparent that there were unknown data quality problems. For example, if a field for U.S. State code has more than 50 records, then either there were errors or international (Virgin Islands) or XX or its definition morphed into something else. They tell you null values, min-max, etc. More complex rules can be defined in order to test the consistency of data across fields, e.g. if the birth date is greater than X, the pension plan cannot be “defined benefits.” They can be set up in batch as an initial data discovery exercise but you can also implement them as part of the ETL as data is trickle fed into the data warehouse or you can intercept data on an ERP screen before it is submitted and stored in the ODS. So “Mickey Mouse” can be caught as an invalid name.

An Enterprise Data Warehouse design can help improve data quality by permitting hard and software integrity checks.

This approach can help with conformance, governance and stewardship, and higher quality ETL. Fewer cases have to be treated as exceptions. The ETL design is sped up. The users know where and when to expect what levels of quality.

The problem is getting management to recognize its value and fund it. It is something that needs to happen before many IT activities that involve the data model, such as re-platforming or incorporation of mergers and acquisitions. Although there are a million reasons to do non-technical work before the project begins, the clients of one of the participants have a “helluva” time doing it. Getting agreement on the data model is a critical

risk factor for the project, but getting agreement on it up-front is still very difficult. Such activities do not involve big hardware purchases or “press-release-worthy” victory celebrations at the end. So people will put this off. Example: finance and marketing and service may decide they want to change the definition of something in the middle of the year, but then HR comes along and says, no, the sales people are paid based on this variable compensation plan, and if we change the definition in the middle of the year, and if we change everything underneath their feet, that’s a big problem. So the spider-web of concerns spirals out, and people want to ignore these things. Have firms been able to tackle this, putting it on the critical path as a sort of “internal victory”?

This participant continued: don’t wait until all the erroneous data has been loaded into the Data Warehouse and shows up on an end-user report. This may cause the end users to immediately lose faith in the IT department and the whole project, which could lead to irreparable harm for the project.

Specialized packages that really deliver in this case are available for \$25K-\$1M. They are not vaporware. They have dictionaries and bring a lot of value. Some specific vendors were named. Essential Software is a good firm, just bought by IBM. Evoke Software and Similarity Systems were bought by Informatica. FirstLogic was recently bought by Business Objects. SAS has a solution. You can strike deals to use these tools for free if you buy maintenance, for example. They do not require terribly technical capabilities to run these tools. They are migrating towards business analyst usage instead of ETL designer usage.

Some firms establish stewards for the data dimensions who have responsibility for their quality. The data steward role is separate from the gatekeeper person who guards the definitions of the data model. The metadata gatekeeper (like a librarian) tries to ensure that definitions do not migrate over time. One participant firm is discussing assigning stewards, but has not decided on it yet. Another firm has a person serving as a gatekeeper, but this role has come from the bottom up, and this person has not been given authority to control the process (just responsibility so far). This was a warning, because he does not have the corporate fiat/deed for this role, and we all know how difficult it is to have responsibility without authority.

GLOBALIZATION

As mentioned above, a number of companies are struggling with how to globalize their data warehouses. One participant firm has one global ERP (Oracle) and one global transactional database. They are multi-org, multi-language, multi-currency all at the transactional level, but providing the transparency at the transactional or the data warehouse level, where for example there is a guy in Malaysia trying to understand a lead developed in the German office. How can they leverage the need to understand the global customer needs, be it a firm like Nokia or other firms? To get that information into an understandable form at the higher corporate level is one thing, but to get it in forms that it can be

used across locations is another. (Oracle gives you pieces and parts and their view of best practices but you have to do it yourself.)

One participant faced a challenge when some Chinese firms refused to use the ERP of the parent firm. One executive said “No SAP. No SAP!” The Chinese firms were doing manufacturing and the US operation was doing sales, marketing, and distribution. They wanted to do three data consolidations: 1) financial (they just now went off a manual system last year). 2) data warehouse – this involves pulling sales from the various channels. Although there is not a lot of cross-channel or cross-location marketing opportunity, they still want to define it in a way that will be ready for it in the future. Europe and Latin America are on the radar screen. 3) integrate systems from an item, supply chain perspective. They created a multilingual portal for both sides. He has a team of translators in China. They created a bi-lingual metadata layer and took key pieces of information that said exchange lead, or expedite order, or whatever. This firm has found that there may be a common set of about 15-20% of the phrases that show up enough that it is worthwhile to record them in some sort of dictionary. Standard translations are built (since a direct one would possibly lead to a mis-translation), and drop down menus can be used to enforce their selection. So free formatting is replaced by using an abridged dictionary. The portal had toggles to select language. The portal is a translation area. They attack it subject area by subject area. They created their own translation competency in the firm. (You have to do this sooner or later. They had five translators previously handling the emails of 1000 people. That is just not transparent enough.) As a side note, this participant said that he has a team of six developers who do all the customer development on the portal in China. The team lead speaks English. He does the design and specifications here. So now, instead of translating email after email, these five translators and the developer lead are helping him to maintain the metadata level.

Prof. McHenry noted that he was familiar with a Russian firm that does large-scale localizations of Microsoft products for the Russian market. They do this kind of translation so often that they built tool suites to manage the workflow processes. It may be possible to seek out such firms and acquire such tools.

China was easy for another firm compared to Japan. In the CRM application of this firm, the use of Kanji makes it very difficult to find the right customer records when a customer calls the call center. First names are not used very much over there. Kanji characters can have numerous interpretations depending on context, but the context is absent when the call is just beginning. They decided that one way around this is to store the English translations with the names so that they can differentiate once a search returns numerous names based on the imprecise Kanji representation. Vijay noted that the same problem exists in India.

SECURITY

We got to the question of security without too much time left, and did not discuss all aspects. One participant illustrated a problem that arose when a developer removed security

constraints on some data cubes at the sub-division level, making it possible for all divisions to see data about other divisions' employees. The developer was doing some development and intended to restore the security, but forget. It persisted in this state for a year until a senior manager came across it and went ballistic. How do you have the controls to ensure this doesn't happen?

Where are the security controls? At the SQL level? At the level of the Enterprise Data Warehouse model? Then how does this impact the Business Intelligence level? There is not a one tool solves all approach.

Are there people in participant firms who are working full time on ensuring that the proper access controls are in place and updated as needed? Is it role-driven? Is there a challenge in assigning roles to people? Sure this is a challenge. But people (users) do not understand roles, and question why they can't do something today they could do yesterday.

The question of implementing security deep into the infrastructure is the tough part, one participant claimed. Plus external requirements are tough. Because of Sarbanes-Oxley and HIPPA, firms are looking at security from a somewhat different viewpoint. The levels of the audits are different. Global or multinational corporations can find that each year a different country has changed its auditing requirements, necessitating not only changes in that subdivision, but at headquarters and elsewhere.

The biggest challenge is that the team security in the Data Warehouse is under-budgeted and understaffed, said another participant. You are obviously going to have data level securities at the outset, but as the data model grows with cubes and multiple dimensions, etc. – whether you use the star schema or not – it becomes more complicated. For each way you are going to slice and dice the data, a security profile attaches to the resultant data. So yeah, you can see sales data, but only from your region. So the real challenge is not setting it up, but, as with every dynamic growing organization, the sales channel, the sales and marketing matrixes are changing on an annual basis. It's reorganized. And the security models that have been assigned can get pretty convoluted. Reorganizations can also make security very confusing and difficult to maintain. There needs to be a talented person who understands how the back and front-ends are related and how things need to change to go forward. This may get dumped to the DBA or the network administrator, rather than put where it belongs – on the Data Warehouse team. The DBA and the network administrator may do OK on the bottom side, but then they don't know the roles. And they get really confused. You have no choice but to staff and budget for it.

Are users shown a metric that lets them know what percentage of the data they are seeing? The group did not know of this explicit capability. But this danger does not seem to be too great, because the users who use the base reports can see informational bars showing what filters are on. This can be a corporate standard for footnoting templates.

NEXT IT EXECUTIVE EXCHANGE

The next IT Executive Exchange meeting will possibly be the most important yet, and we are hoping for quite a bit of participation. Although the curriculum committee has been making suggestions about the relationship between what we teach and the knowledge skills our graduates need to have, we would like the whole group of IT Executives to discuss this question. The meeting will be on June 13, from 1:00 p.m. to 3:00 p.m. in Bowers Conference room (4th floor, CBA).