

Data Warehouse: Primary Concepts

By Thibault Dambrine

According to market research firm IDC, the total data warehouse market is expected to grow to \$13.5 billion in 2009 at a nine percent compound annual growth rate. Data warehousing technology is currently at a point of acceptance such that for most medium to large companies, it satisfies its own discrete part of strategic corporate data requirements. Effectively, it has proven so useful that no C-suite (Chief Executive Officer, Chief Financial Officer etc.) team would want to run a business without it. How does such a technology, with more of a back-room than front-office connotation, become so hot? In a word, it boils down to *business* (as opposed to IT).



Business Drivers

The raw material for good business decision-making is simply good data. “Good”, for the purpose of this topic, can be defined as accurate, reliable, organized and timely. If one of these elements is missing, the decision maker will either have to do more work to verify, organize, or update the data before making a decision—or risk taking a bad (read expensive) decision.

In what circumstances could data be so bad, when computerized systems are so prevalent? Here are some of the leading causes for having late, inaccurate, disorganized, or unverifiable data:

Data Organization:

- Mergers and acquisitions or major organizational shifts
- Data silos resulting in conflicting information
- Competitive pressure to maximize effectiveness of marketing efforts and production resources

Timeliness:

- Poor data access resulting in delayed decision making
- Inability to perform real-time business analysis

Accuracy and Reliability:

- Rising management costs for disparate systems and questionable data integrity verification (reliability)

Qualifying Questions

How would you know if a data warehouse could help in YOUR business? Here

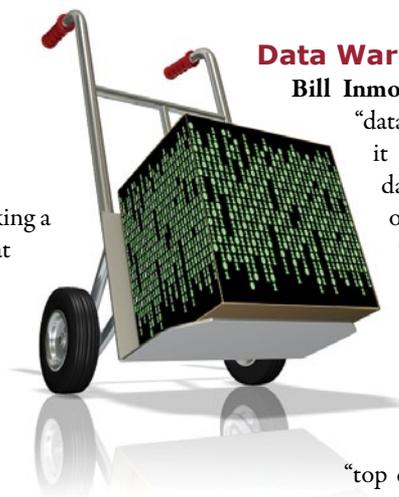
are some litmus test questions:

- Are you managing multiple, disparate databases?
- Is your company lacking a common data set that facilitates decision making?
- Does your IT staff struggle to satisfy business requests for access to data?
- Does your company have the capabilities to perform real-time business analysis?
- Do your competitors use real-time business analysis?

Over and above these questions, three business trends have come to dominate IT requirements:

1. Access to data 24/7, world-wide, for internal and external, web-based information customers.
2. The demand for business decision making data in real-time, at all levels of aggregation.
3. Sarbanes-Oxley and other similar legislation has led to an increased emphasis on financial controls.

The ease of satisfying the type of requirements described above with the help of DW technology has effectively spurred the growth in this discipline. In this article, I will lay out some theory on data warehousing and expose the three phases of a data warehouse project: the planning, the implementation and the operation.



Data Warehouse Theory

Bill Inmon first defined the term “data warehouse”. He defined it in the following way: “A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision making process.” Bill Inmon’s view of the data warehouse is also known as the “top down” design method, as it involves a lot of up-front end-in-mind planning before any results can be extracted.

Ralph Kimball, another well-known data warehouse author, defines a data warehouse as “a copy of transaction data specifically structured for query and analysis.” Kimball is a proponent of the bottom-up approach to data warehouse design. In this approach, data marts, or “mini data warehouse” data storage facilities are first created to provide reporting and analytical capabilities for specific business processes. Like bricks forming a wall, the data contained within these data marts can eventually be combined to create a more comprehensive data warehouse.

A large number of vendors have initially appeared on the market to bring to life Inmon and Kimball theories. As DW software vendors are maturing, a wave of consolidations and buy-outs similar to what has been seen in the ERP vendor space is taking shape. (See **Table 1.**)

DW Consolidations and Buy-outs

Date	Target	Acquired By	Valuation
2003/12	Crystal Reports	Business Objects	\$1.2 billion
2006/02	FirstLogic	Business Objects	\$96 million
2006/12	Knightsbridge Solutions	HP	700-person DW consultancy, undisclosed financial terms
2007/03	Hyperion	Oracle	\$3.3 billion
2007/05	OutlookSoft	SAP	Estimate: \$4-500 million
2007/09	Applix	Cognos	\$339 million
2007/10	Business Objects	SAP	\$6.8 billion
2007/12	Cognos	IBM	\$5 billion
2008/01	BEA	Oracle	\$8.5 billion

Table 1.

In the table presented, note the following points:

- Several acquirers, like Business Objects, who bought Crystal Reports and FirstLogic, and Cognos who bought Applix, were subsequently bought out themselves.
- The takeover events above are in order of date. One can clearly see a trend on the right side, to bigger and bigger valuations, pointing to a maturing and increasing valuation of corporations in this sector.

Data Warehouse Foundations: the Star Schema

The foundation of any data Warehouse starts with giving some thought as to how the data will be organized within. The classic data organization method for DW data storage is called the “Star Schema”. Here is how it works:

The first aim of the data warehouse system is to help decision makers find, earlier than their competitors, hitherto unforeseen trends that may affect their business. To support these decision making requirements, the data in a data warehouse is divided into “facts” and “dimensions”. Facts are tangible events which also carry inherent characteristics. Dimensions are any data elements that may affect the behavior of these facts.

To illustrate this point, the diagram in **Figure 1** shows a retail star schema.

- At the center of the star are the facts. Facts are tangible events. In this case, the facts are individual sales transactions.
- Around the facts are five dimensions:
 - 1) Customer Loyalty Dimension
 - 2) Geographic Dimension
 - 3) Product Dimension
 - 4) HR Dimension
 - 5) Time Dimension
- Further defining the Geographic dimension are two sub-dimensions, also known as “snowflake” dimensions because of the shape they give to the star:
 - 1) Tax Snowflake Dimension, which depends on the geographic location and the time when the fact occurred

2) Weather Snowflake Dimension, which also depends on the geographic location and the time when the fact occurred

Deciding when to snowflake a dimension or to include it should be above all a practical decision. For example, if you get weather data for 1,000 retail outlet postal codes every day from an outside supplier, it may be just easier to keep it in a separate table. Speed of retrieval is also a factor. If the extra time it takes to access snowflake data is too expensive, it may be advisable to add the information to the main dimension or even in the fact table if it is critical enough—effectively de-normalizing, trading storage for speed.

Having divided the data into facts and dimensions, one can mine the data for trends. In a retail environment, one could look for questions such as:

- What distance will the average loyalty card holding customer travel from their home to one of the company retail stores?
- Is there a correlation between the distance and the frequency of the visits?
- If a promotion flyer was distributed by mail to a given postal code, what was the loyalty card holder response?
- In the spring season, at what average temperature do customers purchase more cold drinks, like fruit juices than hot drinks, like coffee?
- If a customer bought a product in the “salty snack” category, what is the probability that they would also buy one or more cold drinks?
- Is there a typical “basket of goods” purchased on certain weekdays?
- What is the profile of the employees with the best sales?
- If a sales education course was provided for employees of a given territory, can the results be measured?

These are just examples of questions one could ask and conveniently get answers to using data stored in a fact/dimension-based star schema. Better yet, beyond having answers to questions that the marketers may be curious about, the secondary aim of the data warehouse star schema is to enable “data mining”. Effectively, data mining is

the use of software to uncover hitherto unknown trends, or trends not easily visible otherwise.

To make the point, here is an example: A large (big surface/many stores) retail grocer has its retail transactions organized in a star schema-based data warehouse. They use a software package to mine the data for cyclic sales trends. What if they discovered that eggplants sales go through the roof every time a full-moon is on a Thursday? The result is that they could prepare and stock up on that item in advance of every full-moon Thursday, enabling management to take advantage of a cyclical, predictable event.

If the eggplant story is not convincing, here is one that is real. In the book he wrote about “Leadership”, **Rudolph Giuliani**, former mayor of New York, describes how data mining actually helped reduce crime in the New York City prison system. The situation was as follows: The Prison system had point-of-sale systems to manage the prison concession stores, which sell items such as cigarettes, chocolate bars and such. They also had a completely separate system to handle criminal records, including the criminal history of each individual registered in the system, with a track record of their crimes both outside and inside the prison system.

One could describe these two systems as “silos” of information, as they were independent of each other. Using data warehouse technology and data mining tools, the NYPD systems analysts uncovered a hitherto un-noticed trend: prior to most prison riots or group crime events, there was a run up in concession sales. When they reported their findings to the wardens, effectively the “business” users of these systems, it made total sense to them. What happened was that a few prison kingpins and gang leaders would first stock up on concession goods, and then trigger a riot. As a predictable punishment for such events, the prison authorities would clamp down on the inmates and shut down the concession stores for a period of time, thereby creating a shortage of goods. The kingpins would then have a captive black market to re-sell what they had bought at regular price, at a

Sample: Retail Star Schema

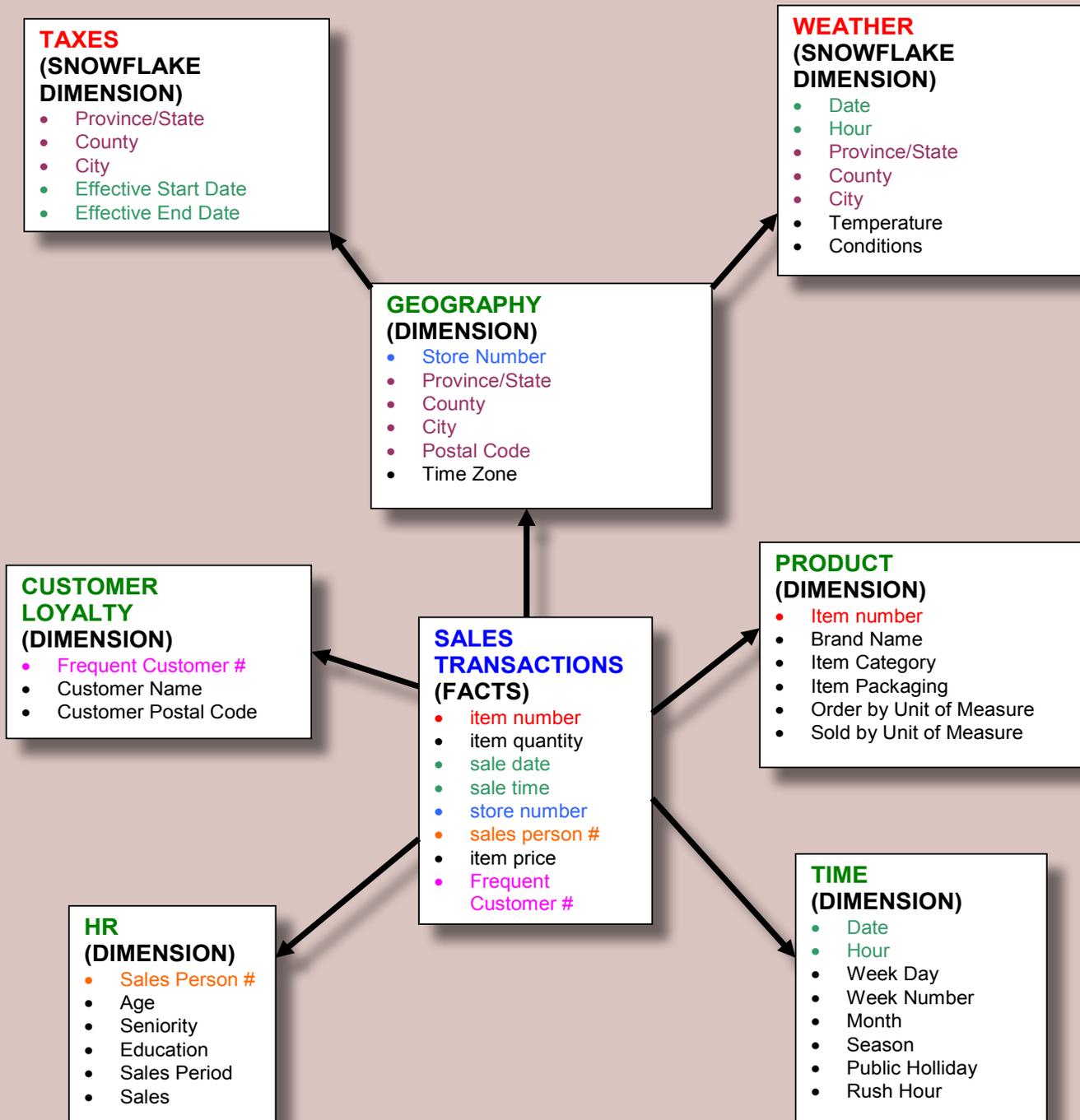


Figure 1.

premium to other inmates. By monitoring sudden spikes in concession sales, the NY prison wardens managed to reduce in-prison crime. When such trends were noticeable, they would move the known trouble-makers to un-familiar cells, beef up the security, do more in-depth searches and effectively destroy the ability of these

gang leaders to take advantage of their positions. The net effect was a significant reduction in prison rioting events.

Making It All Happen: Planning

Prior to starting a data warehouse in an IT department that did not previously use one;

the first step is to ensure the staff who will work in this area have a strong understanding of data warehouse technologies. Plan to get formal training for the team that will follow through with the implementation, development and maintenance of the DW. Training the Business Analysts is equally important, as they will convey the value

of the new techniques to the business data consumers—the decision makers.

Once you have been through the steps of planning your first star schema, your next step will be to plan for the physical data repository for your data. Most of the Fortune 500 software vendors offer some sort of DW hosting solution, sometimes as a stand-alone offering, sometimes in conjunction with operating system vendors, such as Teradata, HP or IBM for example. The same corporations typically offer consulting services, in addition to hardware and software. Getting consultants that understand well the technology that you currently run will help reduce the knowledge gap.

Equally, choosing the right hosting platform for your data warehouse is a decision not to be taken lightly, as it will be there for a long time. Factors to consider are scalability, cost and your own staff’s ability to support this new platform in your organization. Picking a platform your existing IT staff is not familiar with will likely translate in higher consulting, training and startup costs for the project. While the choice of platform should not be entirely driven by money, experience shows that the Business tends to look at costs until they see benefits. Higher up-front costs tend to make the project harder to sell.

When selecting any data warehouse software package, expect significant up-front costs. If anything is right, you will work with the chosen tool for many years, so you need to

be sure that you are buying the software that best fits your requirements. In real estate, the golden rule is “location, location, location”. For software selection, “research, research, and more research” is the number one rule before committing to a software vendor.

One more critical point in the planning phase is budgeting for the long term. Starting a new data warehouse operation means not only new hardware and software. There should be budget items for training, consulting, recurring license fees, disk space, and new staff. All this should not be a surprise to the sponsors of the project.

Designing the first star schema, installing the data warehouse host system and the disk resources are first two steps in the journey. Once the theoretical data model is created, the next step is to implement it, create the tables and create the indexes that will physically contain the data in the data warehouse as well as the indexes, which will allow programs to efficiently retrieve and join to the data within. At that point, the star schema will physically exist, but it will be empty. The next step will be to ensure data will flow into the data warehouse.

As described earlier on, the table relationships in a star schema are foundation information of the data warehouse. These relationships don’t just magically happen. They have to be recorded, tracked and managed. This type of tracking information is known as “metadata”, or data about data. Metadata management tools may span everything from star schema relationship, to file usage

frequency, data origin & business descriptions, and change management. Like data warehousing in general, the discipline of “metadata management” is a growth area.

The Big Picture

The picture in **Figure 2** summarizes the flow of data from external systems to the data warehouse and out again to the business users. Note the span of the governance process. It guides the data and guards integrity from the beginning to end.

Extract / Transform / Load – ETL

Data warehouse systems are typically hosted on systems that are separate and distinct from the source systems they pull data from. To move the data from one system to another, the data warehouse planners will require interface skills. These interfaces will bring the data from the source systems and verify that what has been sent to the DW matches the source system. While ETL work can be done with virtually any common programming language, the trend is to use specialized ETL packages that facilitate the three steps described above.

When designing your star schema, you have already made a decision about what data is important to you, what data you want to see correlations for, and what data will provide new value. While you may need to refine some of these concepts, by the time you reach the point where your data warehouse is designed, you should know what systems will feed your data warehouse.

Data Warehouse Data Flow Diagram

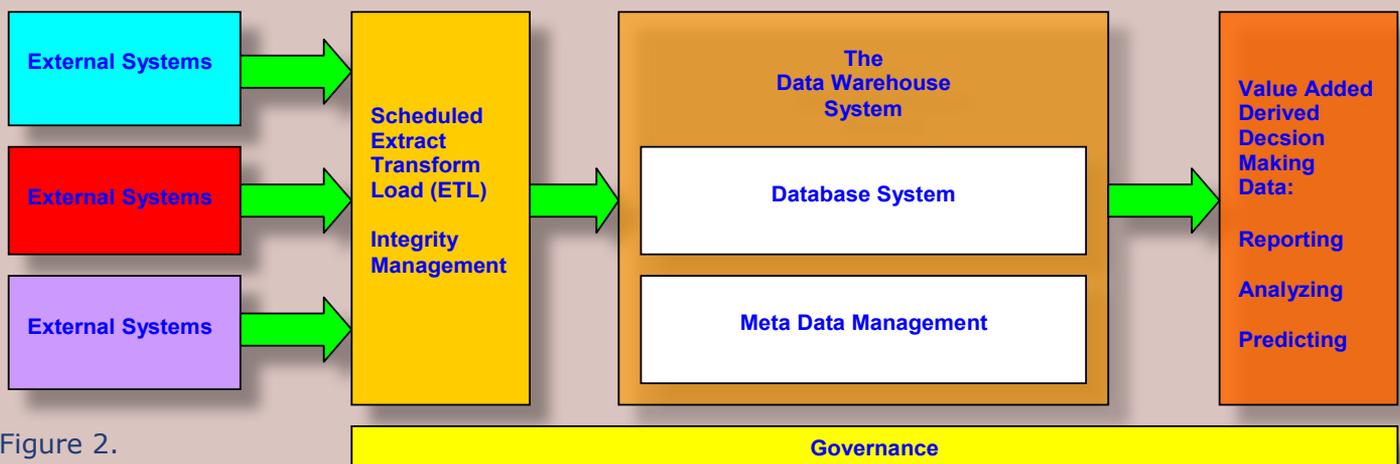


Figure 2.

The process of bringing data from your operational systems to the data warehouse is commonly known as ETL or Extract/Transform/Load. These three verbs suggest exactly what you will need to do to get that data into the data warehouse.

1. Data extraction involves:

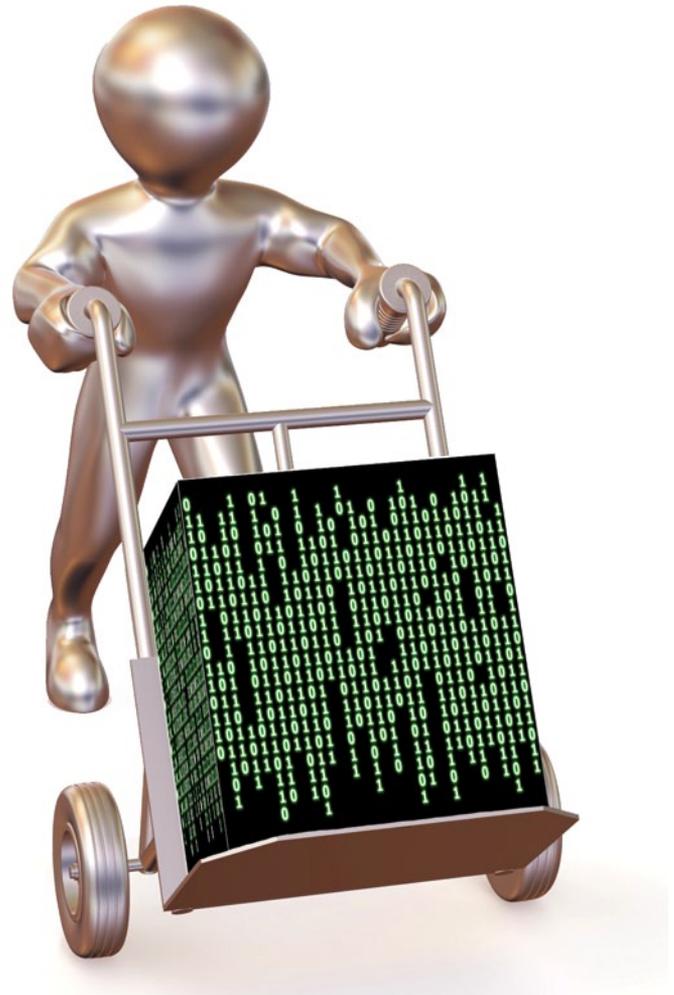
- a. Identifying extraction criteria and frequency of extraction
- b. Cleaning the data (ensuring there are no duplicates, ensuring there is a default value for missing elements, and ensuring that all the data follows the same rules.
- c. Extracting also means, most times, to physically pull data from a number of possible operational systems to the data warehouse, thereby involving software interfaces.

2. Data transformation involves:

- a. Bringing the data to its most granular shape: For example, in your financial system, your monthly data is stored in 12 columns per table. Making the data more granular means to transform the shape of this data and produce 12 individual rows (one for each month), This will enable joining at the GL fact data at year/month/account level to data from other parts of the business stored at a similar level of granularity and make comparisons/trending at that level.
- b. Making the data more meaningful, e.g. if “1” meant “Male” and “2” meant “Female” in the source system, having “Male” and “Female” in the data warehouse makes these values more readable and more usable.
- c. Deriving values from existing data e.g. “sales = qty * price” the result of which will be stored for quicker retrieval (no need to calculate it at retrieval time).
- d. Data cleansing is also typically part of the transformation. This activity includes omission of useless data as well as validation and possibly rejection (error reporting) for data that does not conform to established rules. Activities such as replacing nulls with default values or flagging and removing duplicate values are part of this discipline. On the topic of duplicates, one should keep in mind that some duplicates are trickier to identify than others. For example, having “GM” and “General Motors” as two different clients with the same address could be considered a duplicate, even if the names are completely different. This is the type of data cleansing that has to be addressed for the data warehouse to be effective. Even better, if possible, the cleansing routine should report such anomalies and help the source systems clean their data if possible. Some companies actually specialize in data cleanup.

3. Data Loading:

- a. Ensuring the new data is not a duplicate of the existing one already loaded in the DW.
- b. If the data is loaded multiple times per day, the load may involve re-calculating the running totals used in subsequent extract every time a new load is performed.
- c. Referential integrity is critical for the functioning of a data warehouse. This aspect must be verified at load time.



- d. Source to target integrity checking— verifying that what has been loaded in the DW does reflect what is in the source system.

At the end of the interfaces and ETL processes, one final step should always follow: The integrity check, which will ensure that the data loaded in the data warehouse truly reflects the contents of the business data. If that integrity is lost, then business decisions made with the help of the data warehouse will likely be unreliable— the credibility of the DW will be questioned, rightfully so and this will impact the support from sponsors.

Scheduling

One of the seldom-talked-about topics of data warehousing is scheduling. On the load side, gathering data on regular basis, as seldom as monthly and as often as hourly or better, does require automation. Most IT departments make use of some form of job scheduler. Your data warehouse ETL jobs will have to not only be scheduled but also be subject to business run dependencies. Attention to detail, when it comes to setting job scheduler dependencies is critical.

On the data retrieval side, while a simple data pull from the Business system to the data warehouse sounds like routine, many things can go wrong. Make sure you check your integrity reports daily or better yet, schedule integrity checking jobs that will notify a pager if any integrity is out. Keep a keen eye on any disruption or any job that runs abnormally slowly or quickly. If there is no programming or scheduling modifications in the DW systems, this is often a sign of trouble with source data. Many times, the data warehouse integrities will be the first to signal anomalies in the data and make Business Systems aware of a problem even before the Business Users know about it.

24/7 accessibility is also a requirement to consider, especially for operations that work across many time zones. This should be taken into account when planning scheduled jobs and architecting solutions. Conversely, in limited time-zone DW implementations, having the night hours with relatively low client demands opens the opportunity to do data-preparation in off hours. Examples of these could be

“canned reports” or cubes that take a long time to process. Preparing these at night or during off hours will ensure the data is ready for the users. When they come in the morning, the users will not have to wait for the reports, they will be pre-calculated.

Retrieving the Data – Providing the Value

This is the visible part of total effort that goes into the data warehouse, indeed, the one that will get the most attention. The information sourced from the data warehouse is broadly known as “Business Intelligence” (BI) - strategic decision-making data. Business decision makers will consume the reporting and data mining results from the data warehouse. Being visible to the high end-users who often sponsor data warehouse operations, reporting and data mining get a lot of research and development dollars from most BI vendors. Reporting from the data warehouse is done in two broad methods:

1. Reporting:

– which itself is divided into conventional reports, OLAP and dashboards

2. Data Mining:

– which comprises analysis of past events and predicting future events, on the base of the past trends

Reporting – Describing “what happened”

Reliable and timely Reporting of “what has happened”, in terms of business transactions will no doubt help management make educated decisions on what to do next. In this section, I will describe three levels of reporting:

- Conventional, two-dimensional reports (like one that could be printed on paper)
- OLAP cubes, which have the ability to report on more than two dimensions as well as pre-calculating aggregate values
- Dashboard reporting, which is typically geared to report at a high level (Chief Executive Officers, Chief Financial officers etc.)

Conventional Reports:

Two-dimensional query tools—“Simply Reporting”—many corporations use the data warehouse facility simply to enable users to draw their own data. The data, being typically stored in a star schema format, is in the best shape to facilitate fact/dimension reporting with sub-totals and categorization by dimension. A variety of specialized vendors offer typically point-and-click web-based tools which enable even the most junior users to produce very usable reports. These reporting tools typically offer scheduling options and the ability to provide sophisticated results at low user effort cost.

Spreadsheets are generally looked-down upon as a reporting tool. The truth however is that they are still widely used; primarily because of their low cost. The ease of use and flexibility offered by this tool is also its downfall. Spreadsheets enable users to conveniently manipulate data and potentially alter the original “version of the truth”. Use of spreadsheets, because of their flexibility and ease of change, also means that there will likely be inconsistencies between methods of extraction, source of data, calculations, and results. With so many potential inconsistencies, silos of information tend to appear. While



spreadsheets are easy to use, most experts agree on limiting their use for data warehouse analysis and reporting.

OLAP:

is short for “On-Line Analytical Processing.” The website www.olapreport.com has a more descriptive acronym: FASMI for Fast Analysis of Shared Multidimensional Information. OLAP Cubes are akin to reports, except that they can have more than two dimensions. It could enable for example, to see a representation of sales per cities over a time period (3 dimensions rather than 2) which is the best a “paper” report or even a conventional spreadsheet can do. Cubes can only be visualized using specialized software, much in the same way that “Microsoft Excel” can open an “.xls” document. OLAP Cube terminology refers to measures (equivalent of facts), categorized by dimensions.

Dashboards:

In an automobile, the dashboard has a simple function: to inform the driver at a glance of the status of all critical systems in the car. This includes the speed, RPM, fuel, and engine temperature. In all cases, there is a red line on each gauge to clearly indicate the unsafe zone of operation. Corporate dashboards, in a similar way, are geared to provide vital corporate data at a glance to the top decision makers. The information is provided in a format that shows how close or how far the corporation is operating from a pre-defined red line. When for example, inventory levels are critically low or how far from the pre-defined norm the key performance indicators such as sales or cash balance are. The idea is that the executive level management would be informed early and accurately – real time if possible – about the state of the company to enable action early on.

One of the best-known digital dashboard measurements is the “Norton Kaplan Balanced Scorecard”, which measures four critical aspects of a company:

- Financial perspective
- Customer perspective
- Internal process perspective
- Learning and growth perspective

By setting up ranges within which the company should operate on digital dashboards, the C-Suite staff can monitor

the corporate health and trends with relative ease. If one particular area would need attention, the Chief can drill down (another word for “get down to the details”) to see what is happening. For example, if the sales were unexpectedly down for a given month, the action to take would be to drill down in the sales by territory, to see if any specific territory is a problem, or by product category to see if a given product line was in trouble. Corrective action can then be taken early, before the trouble spreads.

Data Mining: Analyzing and Predicting

“Data mining”:

suggests a rigorous, rule-based statistical approach to examining the data with purpose to identify repeatable trends within. To effectively mine data, companies use purpose-built software tools. The trends they are looking for are often not visible to the naked eye. Think of it as “seismic prospecting” in your data.

Data mining has a somewhat serendipitous connotation. The general idea would be that data mining software would be able to auto-find trends hitherto un-noticed. While this does happen occasionally, most times, the findings are not completely unexpected.

Data mining is all about getting a better understanding of the data. In effect, the end-result of data mining is literally a form of “information about data”. To put this in simpler words, if a user says “I have a hypothesis about this data”, the data mining software tool can be programmed to verify the hypothesis, enabling stronger decision-making and possibly pro-active or pre-emptive action.

Analyzing – describing “why it happened”:

The first purpose of data mining is to describe “why it happened”. Effectively, to figure out, out of the mass of organized data within the data warehouse, how a change in a dimension has impacted the facts. If such a trend can be reliably recognized and if the change in dimension is cyclical in nature, there is a good bet that the event is predictable. A simple example could be

to determine at what point, between the winter and summer, sales of cold drinks become more prevalent than warm drinks. The driver here would be the temperature change, which is cyclical in nature and thus predictable.

Predicting – describing “what will happen” or rather “what will likely happen”:

This is a stage where the DW has enough past data to mine, and has attained a stage of maturity where it can reliably find predictable trends within the data. The rationale is to enable the decision makers to take pre-emptive action to capitalize on predictable events before the competition would.

Data mining can be applied to look for trends outside of the company’s control, weather or demographic changes for example. It can also be used to measure the impact of actions within the company’s own control – like pricing changes, store renovations

Customer retention is one of these areas that are often challenging. Do customers about to switch to the competition have identifiable pattern behaviors? Could customers about to leave be identified in advance? Could a phone call from a sales associate or customer care representative increase their chances of staying? Data mining can be used to look at past client purchasing patterns common to switchers and help recognize in advance those about to defect, enabling action prior to losing customers.

The Critical Value of a DW Governance Model

We have now visited the design, the interface and the data retrieval processes. Successful DW operations are typically busy. Managing the new requests as well as large data volumes that flow in every day does require strong coordination and leadership that can distinguish the difference between “urgent” and “important”.

Setting up a DW Governance team, especially at the inception of the project, is counter-intuitive. Many will say “why now?”. The fact is that the Governance Team effectively aligns the user requirements, the

A Data Warehouse Governance Model

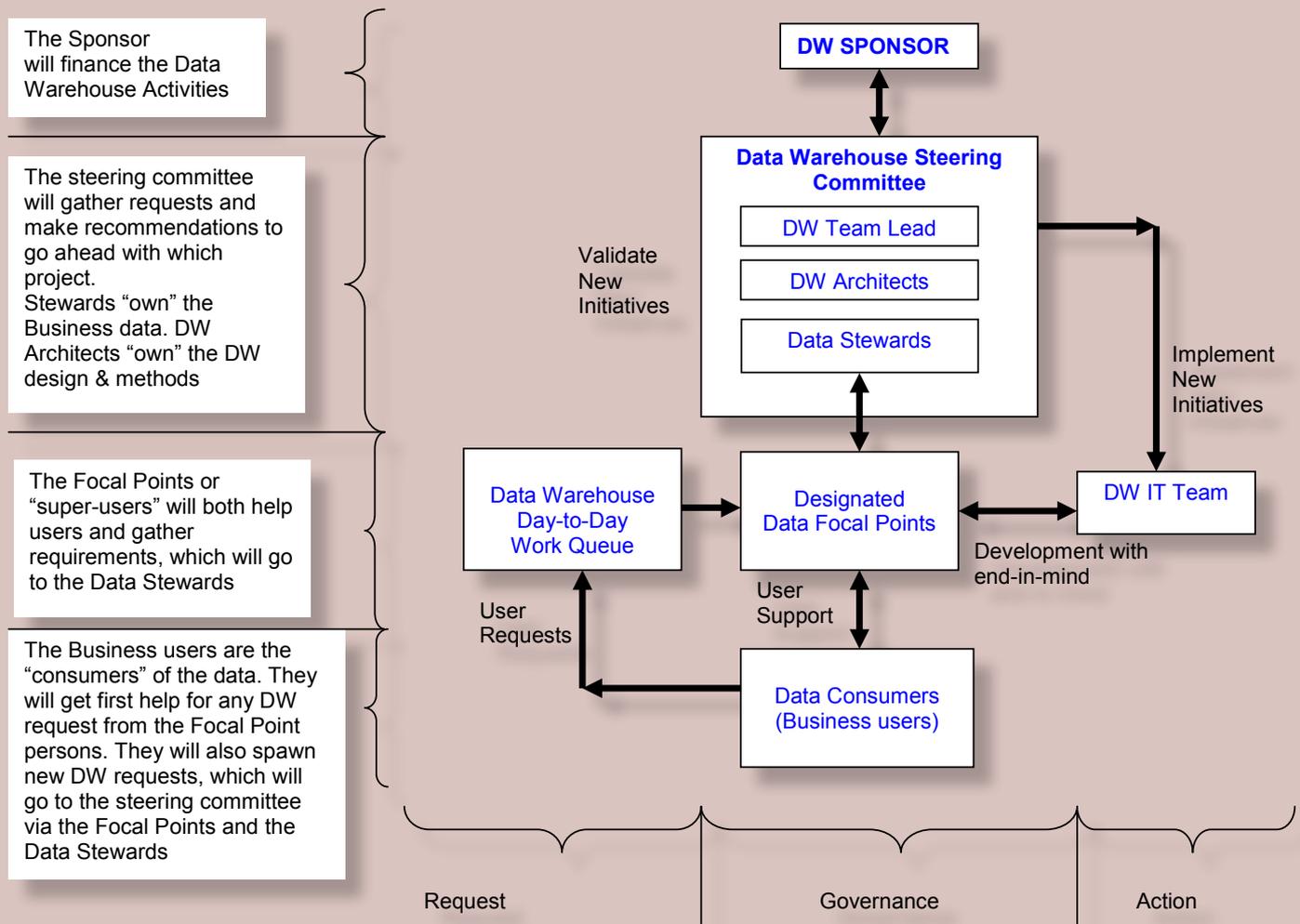


Figure 3.

corporate requirements and the best DW solutions for the aforementioned.

By definition, a DW operation gathers data from many sources and typically also has clients from all parts of the company. With such varieties of requirements and potentially diverging driving interests, the potential for going astray is strong. Typical pitfalls of data warehousing are the creation of “information silos”, or subject areas also known as “data marts” designed in such a way that they cannot be correlated with other data marts within the same data warehouse. Information silos often lead to the “many versions of the truth” syndrome. Constant, unrelenting governance is critical to data warehousing success.

To ensure these potentially diverging interests are properly prioritized and funded, it is wise to establish early on a

permanent governance team, composed of both technical and business leaders.

The mission of the DW governance team is to ensure each new DW project or initiative follows strict standards, is designed in harmony with the existing structure (avoiding isolated data silos) and with the duty to make recommendations for senior management when new investments are necessary. Too many times, when many conflicting interests require help from the data warehouse, “the squeaky wheel gets the attention first”.

At a high level, the DW Governance Team must establish priorities for DW services, enabling work on what is “important” as opposed to what is deemed “urgent” by some. At the ground level, to ensure the data quality is monitored and maintained. Beyond this, establishing strong processes

and a DW project life cycle will add value. While it does require some extra resources, consistent methods of bringing new data in will ensure predictable results. As in all things in IT, “no surprises” is desirable. Each new DW proposed initiative should also be subjected to a governance review to ensure prioritization, alignment with other data marts and methods of execution are in accordance with the direction agreed upon by the Governance team.

While all this sounds an awful lot like “red tape”, the mission of the governance team is to—above all—ensure the DW team best satisfies the Business requirements. The data warehouse must be built with Business requirements in mind. It should be—before anything else (including IT) a Business-sponsored activity. If this is to be a successful, long-term project, it will have to be done with the interest of the business

as a top-of-mind, overriding requirement. Everything else will be secondary in the governance model.

The diagram in **Figure 3** proposes a data governance structure that distributes the day-to-day DW requests to super-users, known as Data Warehouse Focal Points. These are typically business users with a good understanding of the data. When necessary, they will bring concerns or new DW requirements to the Data Stewards, also business users, but a higher level, who effectively have responsibility for the data. The Data Stewards are part of the steering committee. Together with the DW Data Architects, they will make the recommendations on what new DW projects will be implemented and how. These projects will be submitted to the sponsor for approval. Note that the governance process involves all the parties that have any involvement with the data warehouse, from financing (sponsor) all the way to data consumers (the decision makers who use the data provided by the DW).

Size and Scalability

One of the more salient properties of data warehouses at large is that they typically are BIG systems. They require lots of disk storage, processing power and they are geared to grow every day as a matter of existence. If one follows Bill Inmon's principles, data should never be purged from a DW database. On this point, while theory has to stand on principle, there are times when common sense may enable exceptions. If for example, data germane to an obsolete business unit is eating storage but not yielding value, it should be archived and removed from the disk. Part of the governance is to guide the DW staff in constantly re-evaluating an often overlooked question – how much data should be stored? How valuable is the oldest data? Is forever data really worth it? Maybe it is, maybe it is not, but the question should be asked.

Scalability is also an issue that should be considered every time new code is written. One often-found problem is code that works well with small volumes of data but slows down exponentially with the increased volume load. Testing DW applications, whether for ETL or reporting purposes, is critical. The value of having proper indexes, especially on large volume tables, cannot be overestimated.

Conclusion

One can say that data warehousing is often perceived as a simple concept. At a high level, operations consist in bringing data to a central repository at the back-end and retrieving it to do reporting and analysis at the front-end. The reality is that to build and maintain a successful data warehouse – read: useful to its users and sponsors – there are a lot of factors to consider.

Planning:

Number one recommendation – don't just jump in. This is a good spot to take a deep breath and ensure a strong plan is put in place.

Building:

Inmon vs. Kimball.... You don't have to make a choice. Most data warehouse models are hybrids of these two models. The success factor is reflected by the usage and the benefits the DW brings to the company.

Operating:

Ensure strong processes are in place, especially a governance model. The need for governance is proportional to the diversity and number of requests coming from all parts of the company. Without governance, DW operations will likely struggle to please too many customers and not achieve full value potential.

Data warehouses are expensive, large-scale operations. Since designing and extracting value from a data warehouse is not an exact science, one could wager that at least a percentage of DW implementations are mediocre to the point of not bringing any return to their sponsoring business. Data warehousing is not risk-free.

In this article, I have described examples of customer data, sales data, data related to the customer/business relationship. In actual fact, DW technology can be pointed to just about any type of data, including operational or security. Imagine what you could find if you mined the data produced by your firewall or Accounts Payable for example. What trends could you find? If it is well done, the DW may bring huge returns in terms of understanding data trends—both internal [operational] and external, ahead of the competitors.

For the Big Picture perspective, it is Human nature to not want to be the last to know. Because of this simple fact, the trend for more sophisticated Business Intelligence and data warehouse implementations will not slow down anytime soon. The Internet, 24/7 access to data and Sarbanes-Oxley will only accelerate this tendency for more information, delivered faster, covering more ground more accurately. IBM, HP, SAP and Oracle now own most of the more established DW outfits. No doubt, they will put their weight behind this technology and soon bring new, more powerful, more sophisticated tools to the market. Expect these vendors to take competition in this area to a whole new level. No doubt, they themselves use data warehousing services. This can only be good for current and future DW consumers.



Thibault Dambrine works for Shell Canada Limited as a senior systems analyst. He holds the ITIL Foundations as well as the Release and Control Practitioner's Certificates. His past articles can be found at www.tylogix.com.