William McKnight

Data Driven Design for Data Warehousing

Custom Research Report
Prepared for Wherescape
By William McKnight

www.mcknightcg.com

Success Through Speed and Minimizing Cost

provided by:
William McKnight
www.mcknightcg.com

INTRODUCTION

Data warehousing has long been a staple of organizations of all sizes. Early systems ran the business, but as data became moderately important to access for company gain, the architecture limitations of the systems were soon exposed and copies of the data were made for reporting. As it turned out, copying that data was not enough, nor was data warehousing as simple as it sounded. Some best practices emerged, books were written and some science and homogenous approaches were adopted to accommodate the growing data volumes, user communities and data access requirements.

Where once the software vendors were completely focused on operational transactional systems, like a herd, they refocused on the new opportunities in reporting and the advanced reporting that became analytics. From that filling of the marketplace, it is clear that data warehousing is mainstream, or even beyond, into early maturity. It is spoken of as a de facto foundation of most businesses. However, though the value is clearly there, organizations continue to struggle with some of the basic concepts.

Maximum success in data warehousing remains elusive. Value creation can come about through development clearly focused on what the business can readily utilize. That value can be enhanced through cost-effective measures to achieve that goal.

Even 'best practice' shops routinely leave value on the table. The primary challenges that remain present in a data warehouse project can be found in the statement:

> The tendency of technologists to be lured to elegance and disregard the political need to deliver quickly on an important business goal and within budget - the ultimate primary determinants of success.

Table of Contents

Book elegance meets real-world demands and real-world demands win every time. Wouldn't it be better (even more elegant) to incorporate real-world demands into the business of data warehousing from the start? If we were to do this, the profile of data warehousing would improve:

- The development cycles would correspond to business cycles and each would stand on its own for justification
- The overall cost of data warehousing would come down
- The needed cross-departmental business involvement, sponsorship and agreement would be attained up front
- The data model will not take an excessively long time to build
- The data model will not be built relying on either business conversations, ignoring the reality of available data, or simply relying on source environments without a focus on limiting scope to a reasonable quick deliverable

Wouldn't it be better (even more elegant) to incorporate real-world demands into the business of data warehousing from the start?

The skilled data warehouse practitioner is not there to be led by the nose in a detailed step-by-step from business interests who are undoubtedly less familiar with data warehousing. However, considering the history of miscommunication and even failure that accommodate these projects, it should be unsurprising that the business has taken a keen interest in the development process. Many business areas outsource or staff the development themselves![1]

While not absolving the data warehouse team from the need to quickly deliver timely functionality, the value-add of the team comes in the form of doing their work in a manner that builds the foundation for future work and minimizes rework. This we call building scalably ("with scale"). This side of the responsibility means:

- Disk space will not run low in the first few months, when the team is still busy with productionizing and tuning the solution
- The additional concurrent workload of real production does not cause a degregation of performance
- Additional reports and means of data access do not require IT involvement; users should be trained to build
- Additional users can be added anytime; they do not need to wait for a "window" to get their ID and get active
- Data acquisition cycles do not need constant attention and intervention
- The production support workload that was estimated prior to production is about what the real production support workload turns out to be
- High quality data model constructs in the areas that will be used immediately upon going to production
- The data model implemented is fully populated

The data warehouse team mission: Delight the customer, lead the way and do it scalably

The lack of this full life-cycle approach to data warehousing is evident in so many data warehouse projects.

[1] Then, it can start to become confusing as to when this "IT advisement" applies to that group!

However, experienced leaders operate as if they are in a "free market" economy, delivering frequent value and communicating the value of the deliverable as an iteration of working, scalable software with the expectation of future iterations.

Tightening scope and increasing deliverables is required. The utilization of a conscious methodology is necessary to probe failure and success points. The journey should see many points from both camps, especially early on, but learn quickly from the failures and limit repeat failures. These are points of iterative progress towards the value proposition of data warehousing.

The shorter timeframes for delivery than most projects are accustomed to mean standards must be built and maintained across the spectrum of data warehouse components. One of the most important components is the data model.

Data Modeling is the Data Warehouse Foundation

In the data modeling area, rather than starting out with the grandiose goal of building the "enterprise data model", as if it were a respectable end in itself, to be successful, data warehouse teams must leave the spotlight firmly on the business deliverables. The data model, being a means to an end, is grounded in reality and constructed through a series of iterative progressions, staying in synch and not ahead of the partner components which include:

- Data Integration – the data warehouse iteration must be able to populate all data built in the data model
- Data Access – there must be reports or other user-based methods of access to the data

With a series of low-risk steps, the team can test the validity of its approach. Each success brings feedback into the process that leads to lessons learned and ways to improve. This iterative, agile, development approach for the data model is the optimal model.

The data model is the most leveragable component of the data warehouse; whatever aspects of the model that need to be worked on for an iteration need to be done with high care and quality

The trick is to balance the methodology so that you are afforded the best of both worlds. With such a balanced methodology, you maintain the structure that is required to prevent hitting a wall while, at the same time, being agile and able to keep up with the pace of business need. You adhere to both the short-term customer service aspect of modeling while also adhering to the higher calling and responsibility of information management.

A balanced methodology creates modeling environments that are characterized by:

- High Quality, Low Maintenance Models with Minimal Breakage
- Limited Model Rework
- Fewer Working Hours with more Productivity
- Low Stress more Predictive Environment
- Resource Leveling (just enough resources in the right places)

- On-time and on-budget results, resulting in a more satisfied business community
- Easier Justification Processes and Business Approval of New Solutions
- More Opportunity for Career Path, Employee Development and Training
- Faster Problem Resolution with Calculated Predictive Research Actions
- Realistic Expectations Set at all Levels with Open Communications
- IT Viewed as a Strategic Business Unit at the Corporate Level

While the data model is the most important component of the data warehouse due to its high leverage, the agile, iterative approach has become the optimal model for many organizations for not just the data model, but also entire projects.

The waterfall approach to modeling is anathema to the speed value that I am emphasizing here. In this approach, full requirements are gathered up front, next comes design with all of the associated gatekeeping reviews and possibly even physical sign offs. This model and associated development is then moved through unit testing, functional testing, performance testing, regression testing, end-to-end testing and possibly other types of testing that the program feels is necessary. Often, the end goal being thought of is an enterprise data warehouse.

Speed is king with data warehousing; deliver early and often; eliminate waste in the methodology.

The plan with this approach is that the model will still hang together for the business need in the multi-months that have passed since requirements. It never does. Modeling is not linear. Modelers preferring this approach may get upset when reality intervenes and changes are required. It can be overheard that "management doesn't know what they're doing, they keep changing their mind." That's the nature of business.

Recent advances in data integration and business intelligence methodologies and technologies have broken through some critical, long-standing roadblocks and made it possible for data warehouse projects to achieve rapid business gain. The use of "agile" techniques accelerates business intelligence development, drives collaboration between information technology and end users and ultimately conserves and maximizes budget.

Agile literally means "quick and well-coordinated in movement; lithe"[2]. Some methodologies, such as Scrum, Extreme Programming and Agile Unified Process, have emerged that put some structure around the agile concept. Using one of these approaches certainly helps a company be agile and their standard training and materials, available from the market, catapults the need to do full methodology definition internally. However, these methodologies are susceptible to semantic gaps[3], misinterpretations and misunderstandings. And they can become the dogma the company tried to leave behind in waterfall without proper leadership in place. Any data warehouse methodology should delight the customer, lead the way and do it scalably.

While I am not extensively covering any one of the methodologies here, with modeling, it's important to keep in mind that it's the final product – the application deliverable - that is important. Likewise, with the project, it's the business gain that is important.

[2] From Dictionary.com

[3] From Wikipedia.com: the difference between ambiguous formulation of contextual knowledge in a powerful language (e.g. natural language) and its sound, reproducible and computational representation in a formal language (e.g. programming language)

Data Driven Design

Removing excess cycles from the modeling effort means focusing the modeling effort on the current iteration, not on modeling constructs of the future for which data is not currently available. The modeler should keep one eye on the business and the other eye on the source data. Since the data warehouse originates little to no information[4], it must rely on other data stores for its data. Those stores have a finite capacity of information; almost certainly none approach an enterprise level for the business or even a single subject area.

Time has a tendency to change requirements. Moderate pursuit of modeling beyond the current iteration.

Therefore, the body of data that one has to work with in a data warehousing effort is limited to data that exists in appointed sources. Data profiling is a must in order to understand the quality of the source data. Data existence does not imply quality. The data warehouse will illuminate the data for perhaps the first time. The data warehouse will also cause the data to be viewed in a broader sense. While data quality issues with single records can be viewed as immaterial and unrepresentative in an operational system, seeing high percentages of invalid data through the lens of a data warehouse is another story.

Though the data defects are not caused by the data warehouse, it is at the time of sourcing the data for a data warehouse that data quality problems come to light. It's often therefore the data warehouse teams that undertake the data quality initiative and drive the data defect remediation efforts, whether the correction is done in the operational system or the data warehouse itself.

Therefore, it's after source system analysis for the interesting data for the iteration, data profiling and data quality efforts when the data warehouse understands the data it has to work with. Though there may be ongoing – and usually lengthy – remediation efforts in the source systems spawned from the analysis, data warehouse iterations move forward with the quality data they have control of.

Analyze the source systems, profile the data and apply quality efforts to determine the data from which to drive design

It is time to refocus modeling energies for the speed of business today and apply focus where it needs to be and where it has been missing - and that is on the data. This is what is known as Data Driven Design. It aligns with the speed of business and the speed that modeling must adhere to. We can know, from experience, what causes the problem of misguided modeling energies, and we can therefore know the solution.

Data Driven Design Principles

1. The data warehouse team mission is to delight the customer, lead the way and do it scalably
2. Get quickly to a shared understanding with the business of what is possible
3. Build the warehouse in a set of quick-turn iterations
4. Model in lockstep with the data warehouse iteration plan
5. Try out packaged or inherited models against the source environment

[4] Except for information derived from other sourced data

6. Build logical and physical models only for what the data integration and data access layers are able to process in the current iteration
7. Perform readiness (analysis, profiling) only on source data systems for those areas that will be processed in the current iteration
8. Come up with a data quality strategy that addresses quality defects at the earliest possible point
9. If changing data en route the data warehouse, bring the pre-cleansed data to the warehouse as well
10. Act according to timeline reality – source data cleanup and process change to keep data clean - will take longer than the data warehouse iteration

Get quickly to a shared understanding with the business of what is possible

With Data Driven Design, all modeling energies are geared to productive activity in support of the data warehouse. Data is given the equal seat at the modeling table that it deserves and modeling doesn't become the long pole in the data warehouse tent with only the promise of long-term benefits. It is the solid, practical foundation of a ROI-producing data warehouse throughout the data warehouse iterations.

About the Author

William functions as Strategist, Lead Enterprise Information Architect, and Program Manager for complex, high-volume full life-cycle implementations worldwide utilizing the disciplines of data warehousing, master data management, business intelligence, data quality and operational business intelligence. Many of his clients have gone public with their success stories. William is a Southwest Entrepreneur of the Year Finalist, a frequent best practices judge, has authored more than 150 articles and white papers and given over 150 international keynotes and public seminars. His team's implementations from both IT and consultant positions have won Best Practices awards. William is a former IT VP of a Fortune company, a former engineer of DB2 at IBM and holds an MBA.

William can be reached at 214-514-1444 or wmcknight@mcknightcg.com.

5960 W. Parker Rd., Suite 278-133
Plano, TX 75093
Tel (214) 514-1444