DW 2.0 represents a long term architectural blueprint

A tale of two architectures (2)

W. H. Inmon

Recognizing the problems that arise with realities of the implementations of the Kimball Stage 1 simple dimensional model in large organizations, Kimball next suggests that what is really needed is a "conformed dimension" in addition to the star schema. The conformed dimension sets the stage for the next stage of evolution of the Kimball architecture, the Kimball Stage conformed dimension architecture.

The conformed dimension "contains descriptive attributes and corresponding names". The purpose of the conformed dimension is to integrate the many data marts that are produced by the simple dimensional data model.

Enter the conformed dimension Kimball stage 2 architecture

With conformed dimensions Kimball starts to address the issue of integration. And with the issue of integration comes the issue of integration across the enterprise. And once the subject of integration across the enterprise is addressed, the speed with which the Kimball architecture can be implemented slows down exponentially. You simply cannot quickly and easily integrate data across the enterprise. So the attraction of speed of development of the Kimball architecture changes drastically in the face of a Kimball Stage 2 conformed dimension architecture. In the face of a small organization, the need for integration across the organization may not be a large issue. But in the face of a large organization, the issue of integration across the organization is a very real and pressing issue.

The result of introducing conformed dimensions to the Kimball Stage 1 dimensional architecture is the Kimball Stage 2 architecture. Fig 6 shows the Kimball Stage 2 conformed dimension architecture.

The Kimball Stage 2 conformed dimension architecture addresses the problem of integration of data across the organization by introducing conformed dimensions. With conformed dimensions it is possible to achieve a degree of integration. But there still are problems with a Kimball Stage 2 conformed dimension architecture. The problem with the Kimball Stage 2 conformed dimension architecture arises from the fact that conformed dimensions address only some attributes of the corporation, not all attributes of the corporation. There are many other attributes and data elements in the corporation that are not found in conformed dimensions and those attributes need attention when it comes to integration. But conformed dimensions do not address all data

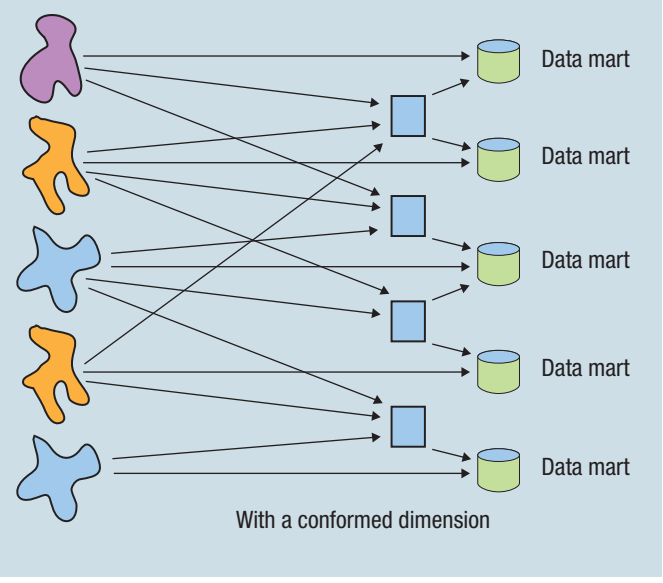

Figure 6.

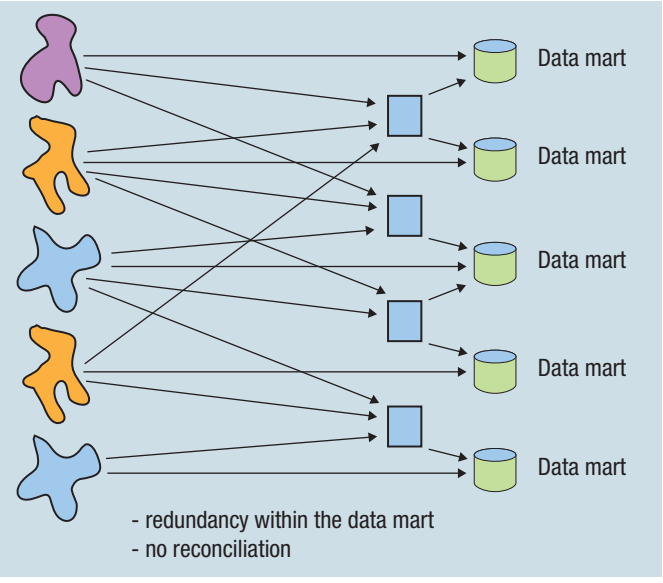

Figure 7.

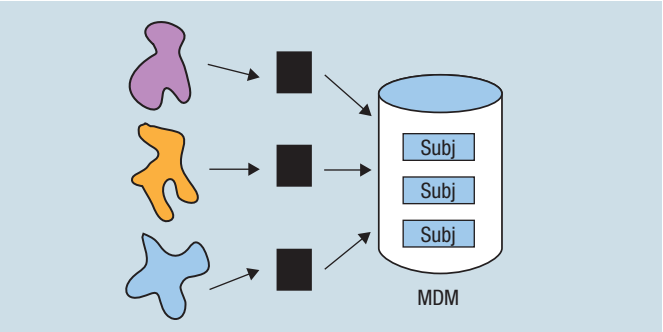

Figure 8.

elements, only some data elements. Fig 6 shows that in the portion of the Kimball Stage 2 conformed dimension architecture that is not contained in conformed dimensions that there is tremendous redundancy of data, that there is a tremendous amount of unintegrated data, and that addressing conformed dimensions only addresses a small part of the general problem of lack of integration of application data. In short, the data not found in a conformed dimension is not integrated in a Kimball Stage 2 conformed dimension architecture.

But there was another major issue with the Stage 2 conformed dimension model. The problem arises from the data marts that are connected by a conformed dimension. The data marts are process oriented collections of data – order processing, inventory, shipping – and so forth. As such, many data elements appear in more than one process oriented data mart. Even though the problems of integration of some of the data elements were resolved by the creation of conformed dimensions, the problem of integration of data elements that were not in the conformed dimension arose because of the process orientation of the data marts. These issues with a Kimball Stage 2 conformed dimension architecture are seen in Fig 7.

Enter MDM and the "golden record"

While conformed dimensions are a first step to integration of corporate data, they are just that – only a first step. What is needed is complete integration of ALL the corporate data needed for analytic processing. The key to creating a basis for all integration is MDM or master data management. With MDM there is the creation of what is sometimes referred to in MDM as the "golden record". (NOTE: the term "golden record" is not a term that widely appears in the Kimball architecture, but is a term that appears in many other conversations regarding MDM. The term nevertheless describes the most salient aspect of MDM – the need for a single, believable source of corporate data.) The golden record in an MDM architecture is the place where the single version of the truth lies. Fig 8 shows a Kimball Stage 3 MDM architecture.

In the Kimball Stage 3 MDM architecture it is seen that there is at last corporate, enterprise wide integration of data. With MDM, now the "single version of the truth" exists. At this point, the focus on speed of building is completely lost because trying to integrate data across the enterprise is not a speedy exercise under any scenario. Even though the "single version of the truth" has been established in the Kimball architecture by the introduction of MDM, the evolution of the Kimball architecture is not complete.

But there is yet another problem with the Kimball Stage 3 MDM architecture. This issue presages a next stage of evolution for the Kimball architecture.

The problem with the Kimball Stage 3 MDM architecture is that many departments across the organization need to use the data found in the non redundant MDM generated "golden records" for their analytic processing. In the world of MDM the orientation is to an organization around integrated subject areas. Data is organized according to the major subject areas of the corporation, such as CUSTOMER, PRODUCT, ORDER, SHIPMENT and so forth. Across all of the MDM subject areas there is little or no redundancy of data. When organizations go to use the subject area data, they find that they need to recast the subject area data into a form and structure for their own parochial processing needs. Stated differently, even though the MDM subject area does support the "single version of the truth", the MDM "golden records" do not support the many different ways that data needs to be viewed by the different departments of the organization. End users need to take the MDM subject areas and recast them into a form and structure for their own parochial processing needs. For this purpose, there is a simple architectural answer. In order to use the "golden record" across the organization for analytic processing in many different ways, departments may copy (but not update or otherwise alter) the data from the "golden record". These customized copies of data from the "golden record" can be called data marts. Those data marts receive data that comes from the MDM golden records (i.e., the "single version of the truth" records) that are found in the Kimball Stage 3 integrated MDM data). The data marts are then recast into a form and structure suitable for the individual departments that need to do analytical processing.

The result is the predictable next evolution of the Kimball architecture after the MDM has been established – the Kimball Stage 4 hub and spoke architecture. Note that it is only a prediction that the Kimball Stage 4 hub and spoke architecture will evolve.

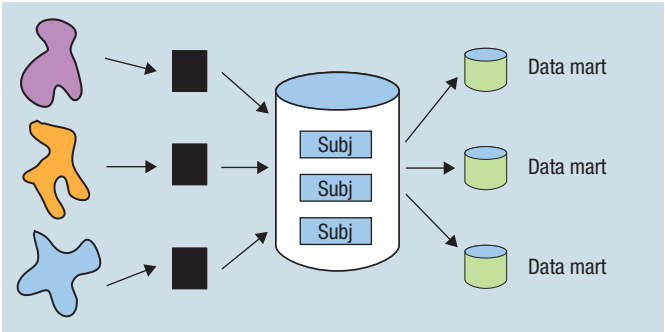

Figure 9.

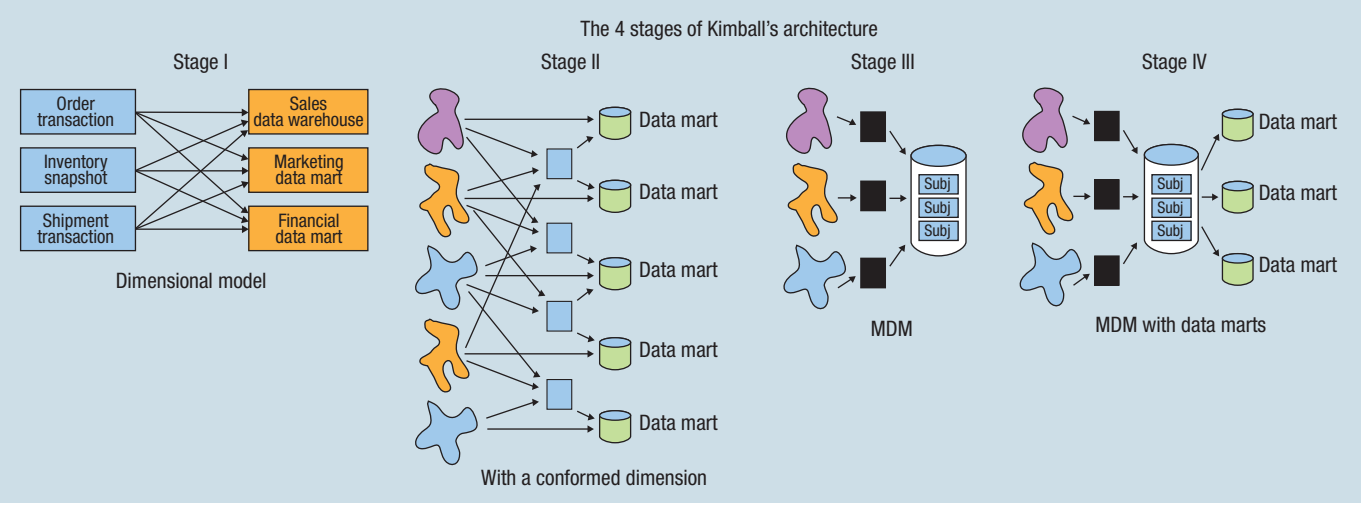

Figure 10.

Fig 9 depicts the predicted Kimball Stage 4 hub and spoke [1] architecture.
The different stages of evolution of the Kimball architecture can be seen in Fig 10.

Some of the notable events/papers/books/definitions of the different stages of evolution of the Kimball architectural approach are:
1992 – Kimball Stage 1 – simple dimensional model phase:
- Formation of Ralph Kimball Associates;
- "A data warehouse is a union of all its data marts";
- The data warehouse toolkit, 1998;

2002 – Kimball Stage 2 – conformed dimension/master conformed dimension phase:
- Data warehouse toolkit: the complete guide to dimensional modelling, 2002;
- Kimball Group/Kimball University: Kimball Design tip #48, De-Cluster with Junk (Dimension), Aug 7, 2003;

2007 – Kimball Stage 3 – MDM phase:
- Intelligent Enterprise: Kimball University, Pick The Right Approach To MDM – Feb 2007;
- The Need For Master Data;
- The Conformed Data Warehouse;
- The MDM Integration Hub;
- The Enterprise MDM System;
- Four Steps to MDM.

The evolving Kimball architecture

There is a certain irony here. Compare the predicted Kimball Stage 4 hub and spoke architecture with the corporate information factory architecture that was published by Inmon a decade earlier and it is seen that they in fact are the same. The emphasis for the predicted Kimball Stage 4 hub and spoke architecture is now on integrated data, not on speed of development.
The next irony is that the predicted Kimball Stage 4 hub and spoke architecture cannot be created quickly and easily. There has been a change in emphasis from Kimball Stage 1 architecture to the predicted Kimball Stage 4 architecture. In Kimball Stage 1 the emphasis was on speed of development. But in the predicted Kimball Stage 4 with the need for true enterprise development and the creation of the "golden record", building the Kimball Stage 4 environment is no longer speedy. The emphasis on the Stage 1 Kimball architecture is on a few legacy systems. The emphasis on the Kimball Stage 4 architecture is on the enterprise. The emphasis for the predicted Stage 4 Kimball model – the need for integration across the enterprise – was the one that Inmon recognized 10 years earlier.

Predicted Kimball stage 4 = corporate information factory

The predicted Kimball Stage 4 architecture has evolved (and is still evolving) to the Inmon corporate information factory. The Kimball Stage 3 architecture and the predicted Kimball Stage 4 hub and spoke architecture is being discussed in 2010. And the Inmon Corporate Information Factory was created in the 1990's, more than a decade earlier.
Over time, the basic Kimball dimensional architecture has undergone several major intellectual revolutions, all started by the realization that the basic dimensional architecture did not work in the face of large scale systems and that the simple dimensional model was not a true enterprise solution. That intellectual evolution is depicted by Fig 11.
First there was the dimensional architecture. Then there was the

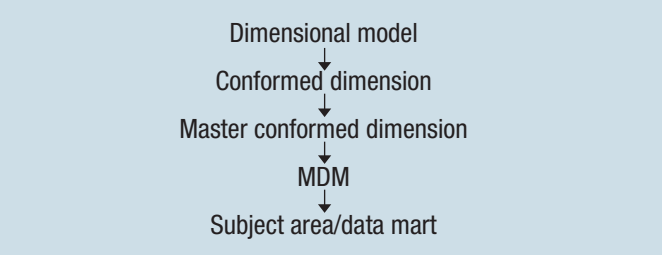

Figure 11.

conformed dimension. Then there was the master conformed dimension. Then there was MDM. Finally there is the predicted Kimball Stage 4 hub and spoke architecture.

Throughout the renditions of the Kimball Stage 1 – Stage 4 approach to data warehousing, the Kimball approach has been particularly popular with software vendors. In particular the Business Intelligence data mart software vendors have been drawn to the original Kimball Stage 1 simple dimensional architecture. There is a reason why data mart and Business Intelligence vendors are drawn to the Kimball Stage 1 simple dimensional architecture. That reason is the Business Intelligence and data mart vendors care most of all about making a sale. Consider the sales cycle for the data mart vendor in the face of an Inmon style corporate information factory architecture. In the Inmon architecture before the data mart can be built, a data warehouse has to be built. But building the Inmon style data warehouse is going to take a while. Therefore, building an Inmon style data warehouse gets in the way of the data mart vendor making a fast sale. On the other hand, with a Kimball dimensional model approach, the data mart is needed almost immediately. Is it any wonder then that the data mart, Business Intelligence vendors gave all their support to Kimball? It was in their own best interest to do so. Stated differently, the data mart, Business Intelligence vendors cared nothing for the long term architectural interests of their customers. All the data mart, Business Intelligence vendors cared for was their own immediate bottom line – making a quick sale, at the expense of their customers long term architecture. The Kimball dimensional Stage 1 simple dimensional architecture was a natural fit for the fast building of data marts.

Fitting the two architectures together

It is seen that there is a significant architectural difference between the Inmon corporate information factory "single version of the truth" architecture and the Kimball Stage 1 simple dimensional architecture. Despite the differences, there is a juxtaposition of the two architectures that makes sense. Fig 12 shows this arrangement.

Fig 12 shows that in the center of the hub is the Inmon corporate information factory. In the Inmon corporate information factory is the "single version of the truth". The data here is granular, historical and integrated. The data here is cast in the form of the relational model.

Surrounding the "single version of the truth" are the data marts. The data marts are cast in the form of the Kimball star schema architecture. In the star schema architecture, each data mart is optimized to meet the analytical needs of the end user. The source of data for each data mart is the data warehouse.

The basic architecture seen in Fig 12 meets the needs for a single version of the truth and for the different analytical needs of the different departments. And the architecture seen in Fig 12 blends the Inmon and Kimball architecture, taking the best features of each architecture.

However, the architecture seen in Fig 12 has been extended over the years into a much more robust, much more sophisticated architecture. The architecture seen in Fig 12 has been extended into what can be called DW 2.0.

DW 2.0

Over the decade between the creation of the corporate information factory and DW 2.0, the Inmon corporate information factory architecture has evolved as well. Today the Inmon architecture is best described by the body of work known as DW 2.0. Written in 2007, DW 2.0 is described in a book entitled "DW 2.0 – Architecture for the next generation of data warehousing".

The essence of the DW 2.0 architecture is depicted in Fig 13.

The DW 2.0 architecture contains many different architectural components that have been added on to the basic corporate information factory. Some of the more salient aspects of the DW 2.0 architecture include:

- Unstructured data as an essential and granular ingredient in the data warehouse;
- An exploration warehouse;
- Near line (or alternate) storage;
- An archival component;

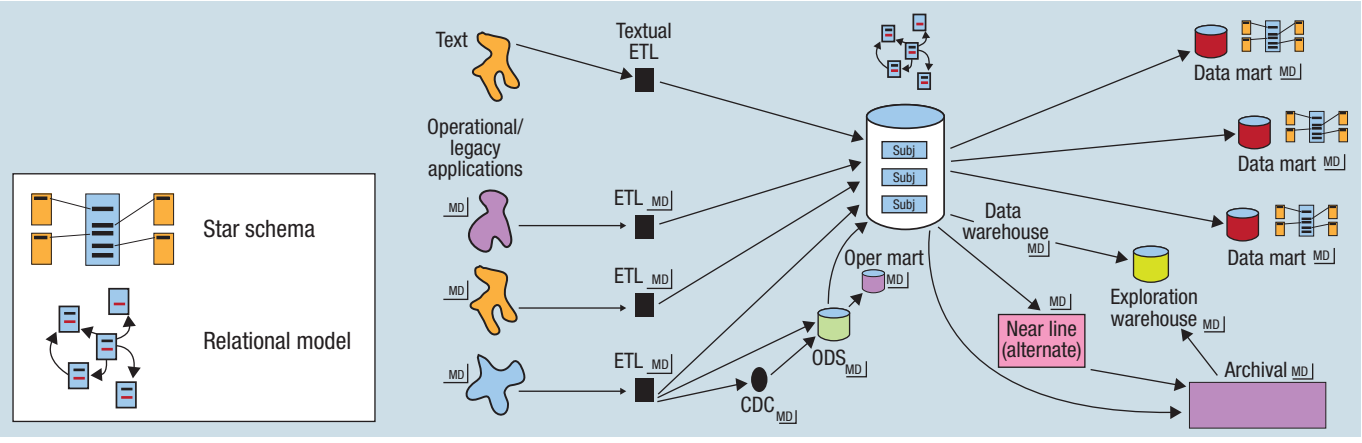

Figure 12.

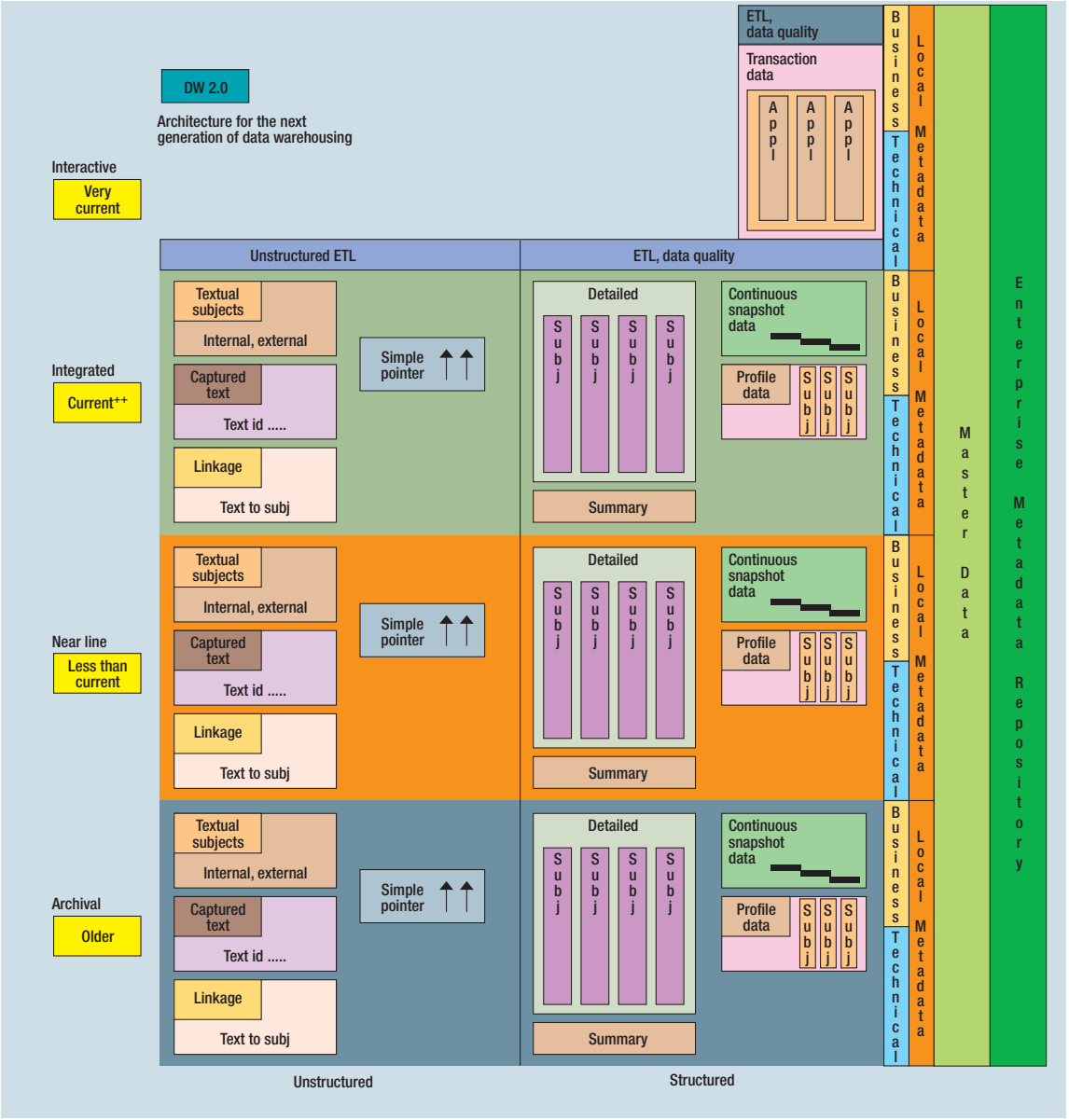

Figure 13.

- Oper marts;
- An ODS;
- Metadata as an essential component of the architecture;
- Taxonomies;
- Changed data capture;
- Recognition of the life cycle of data within the data warehouse.

The DW 2.0 architecture then represents the evolving architecture for data warehouse. It contains the best features of the Inmon architecture and the Kimball architecture can be combined very adroitly. DW 2.0 represents a long term architectural blueprint to meet the needs of modern corporations and modern organizations.

Bill Inmon

William H. Inmon (binmon@inmondatasystems.com) is oprichter en CEO van Inmon Data Systems, gevestigd in Castle Rock, Colorado.

Literatuur

Inmon;

- *Building the data warehouse, John Wiley, 1991.*
- *The corporate information factory, John Wiley, 1999.*
- *Operational data store, John Wiley, 1995.*
- *Business metadata: capturing enterprise knowledge, Morgan Kaufman, 2007.*
- *Tapping into unstructured data, Pearson, 2007.*
- *DW 2.0 – Architecture for the next generation of data warehouse, Morgan Kaufman, 2007.*
- *Building the unstructured data warehouse, Technics Publications, Nov 2010.*

Kimball;

- *Data warehouse toolkit, John Wiley, 1998.*
- *Data warehouse toolkit: complete guide to dimensional modeling, John Wiley, 2002.*
- *Data warehouse toolkit: building the web enabled data warehouse, John Wiley, 2000.*
- *Differences of Opinion: Comparing the Dominant Approaches to Enterprise Data Warehousing, Intelligent Enterprise magazine, 2004.*
- *[1] Internet – Planning MDM and EDW with Dr Kimball for 2010 – Informatica.*