A TALE OF TWO ARCHITECTURES

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us,……..."

Charles Dickens "A Tale of Two Cities"

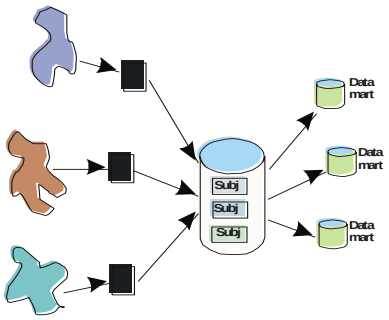
By W H Inmon

It was the best of times. It was the worst of times.

From an age of applications and the confusion over application based information in the corporation arose the concept of a data architecture and data warehousing. Into the miasma came Bill Inmon's best selling book - BUILDING THE DATA WAREHOUSE. And there was Kimball's software company – RedBrick Systems. And soon the world of data warehousing was born. It was the late 1980's and the world was about to witness the rise of analytical processing, business intelligence and a whole host of technologies never before seen that would change the world forever.

THE CORPORATE INFORMATION FACTORY DATA WAREHOUSE

The industry accepted definition of a data warehouse – "a subject oriented, integrated, non volatile, time variant collection of data for management's decision making" – appeared in BUILDING THE DATA WAREHOUSE. Later books by Inmon soon appeared which described the architecture into which the data warehouse fit. The architecture – sometimes called the "corporate information factory" (or simply "Inmon's architecture") is seen in a simple form in Fig 1.



SINGLE VERSION OF THE TRUTH

The nexus of the corporate information factory and its foundation – the data warehouse - is the notion of the single version of the truth. Centered in the data warehouse and described by the definition of the data warehouse is the granular, integrated historical data - the "single version of the truth" - which is the essence of the corporate information factory. With the corporate information factory, the data

warehouse has a place where there is a “final word” as to what data is right and what data is wrong. At the heart of the confusion over information that preceded the data warehouse is the inability of the organization to understand what data is correct and what data is not correct. It is hard to make proper decisions on data that is unreliable. Prior to the corporate information factory, organizations had a plethora of data, but they had no idea what data was correct and what data was incorrect. With the corporate information factory, there was a definitive source of data to which the corporation could turn – the “single version of the truth”. It is true that the corporate information factory solved many other problems. But the single most important aspect of the corporate information factory was that it contained the “single version of the truth.”

The corporate information factory includes an architecture that centers around the data warehouse. It is in the data warehouse where the “single version of the truth” resides. Other features of the corporate information factory architecture include legacy, operational systems, ETL and data marts. ETL is the technology that reads in raw data from applications and writes out corporate data (or data that constitutes the “single version of the truth”). Data marts are those data bases created for the analytical needs for different departments and different groups of people doing analytical processing. In the corporate information factory the only source of data for the data marts is the data warehouse.

The biggest issue in creating the corporate information factory is that of the integration of application data into corporate data. Data that comes from applications must be recast into a corporate form and structure. That is how the “single version of the truth” is created. The integration of old legacy, operational unintegrated data is a complex and time consuming job. In many cases, old legacy data is undocumented. In many cases old legacy data lies in technologies that have been unsupported for years. In many cases old legacy applications must be merged where a merger of application data was never an objective of the designer of the legacy application. In many cases the very definition of data sitting in an old application must be recast. All of the work required for integration is tedious and must be performed in a disciplined and in an exact manner. As such, building a data warehouse for the corporate information factory is not an easy or a fast thing to do. But the result is integrated data – a “single version of the truth” for the organization.

The focus of the corporate information factory is data across the enterprise. Data from many different places and applications – the “legacy systems environment” is all integrated and included into the data warehouse. One of the reasons why the corporate information factory is not built quickly is that data from lots of places needs to be integrated. For a small organization there may be very little integration that needs to occur. But for a large organization the process of integration can be laborious, tedious, and time consuming.

As a rule, the data in the corporate information factory data warehouse is stored in a normalized relational format. Generally speaking, the data in the corporate information factory relational data base is granular, historical and is “lightly” denormalized.

Stated differently, building the corporate information factory is a long term proposition and the result is a long term infrastructure that the corporation can rely upon.

The corporate information factory (and its evolved form - DW 2.0) is the architecture that is espoused and developed by Bill Inmon and expressed as the corporate information factory in his book in 1999 and later in the book DW 2.0 – ARCHITECTURE FOR THE NEXT GENERATION OF DATA WAREHOUSING.

THE DIMENSIONAL MODEL DATA WAREHOUSE – THE KIMBALL APPROACH

But there was another related architecture that arose in roughly the same time frame. That architecture is the one that can be called the “Kimball” architecture. It is the Kimball architecture that is associated with Red Brick Systems . The Kimball architecture has evolved over time, like all architectures evolve. The first stage of evolution of the Kimball architecture began with what is known as a “dimensional” (or star schema) architecture. In the context of this paper we will call the first stage of evolution of the Kimball architecture a “simple” dimensional model. Fig 2 shows the bare bone essence of a simple dimensional architecture.

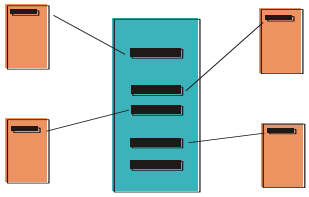
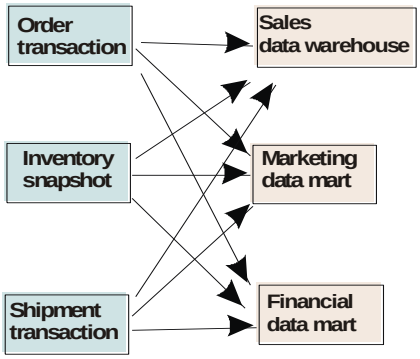


Fig 2 shows a fact table surrounded by several dimensions. In general, the facts are a cluster of attributes that are physically colocated and the dimensions are the separate tables that describe the facts. The fact table and its dimensions form what is termed a “star schema”. As a rule there are many facts in the fact tables and relatively few occurrences of data in the dimensions. The data that comes from applications is placed in a star schema and is used to create what is termed a “data mart”.

The data that populates the simple dimensional model comes directly from applications. In fact, Kimball draws a diagram that shows how application data enters the simple dimensional model. The diagram is taken from an article published by Kimball in 2004, along with Margy Ross[1]. Fig 3 depicts the diagram showing how the simple dimensional model is used to produce multiple data marts from multiple sources.



In Fig 3 it is seen that there are legacy applications and data marts in the simple dimensional model . The many different data marts are populated directly from the many different applications. Kimball goes on

to give his definition of a data warehouse. Kimball's definition relates to the first phase of the Kimball architecture – the simple dimensional model - "a data warehouse is nothing more than the union of the data marts."[3] Kimball refined the definition of the data warehouse at a later point in time, saying that the definition of a data warehouse was a "a copy of the data specifically structured for query and analysis."[2] It is easy and fast to merely copy data from one data base to the next.

Kimball's Stage 1 simple dimensional architecture was never designed for enterprise integration. The Kimball Stage 1 simple dimensional architecture was designed for immediate applications and immediate data marts, where the scope of the effort was limited. Because the scope of the Kimball Stage 1 simple dimension architecture was limited and because only the copying of data was involved, Kimball's Stage 1 simple dimensional data warehouse is fast and easy to construct.

The biggest selling point of the Kimball simple dimensional architecture is the speed with which the data marts can be constructed. Indeed, around the world, people like architectures that are easy to construct and quick to be used. The problem with the simple dimensional architecture (and the nexus of the difference between Inmon and Kimball) is that nowhere in the Kimball Stage 1 simple dimensional architecture is there the notion of the "single version of the truth". At best, Kimball says that application data should be **copied** from the application environment. Inmon, on the other hand, suggests that a fundamental and rigorous transformation of legacy data is necessary in order to create the "single version of the truth".

When comparing the Kimball Stage 1 simple dimensional architecture versus the Inmon corporate information factory, Inmon's data warehouse requires that there be a "single version of the truth" while Kimball's data warehouse is a collection of data marts consisting of data that has been copied from applications. And therein lies the difference between the Inmon approach to data warehousing and the Kimball approach to data warehousing.

DIFFERENCES BETWEEN THE MODELS

The fundamental differences between the Kimball Stage 1 simple dimension architecture and the Inmon corporate information factory architecture can be summed up as –

- The corporate information factory (Inmon) addresses the need for integration of data across the organization creating what can be called the "single version of the truth." The focus of the Inmon corporate information factory is the integration of data across the corporation.
- The Kimball Stage 1 dimensional architecture is quick to build and allows reports to be built quickly but does not require a "single version of the truth" be built, only that a "copy" of data from the legacy environment be made. The focus of the Kimball Stage 1 simple dimensional model is on a few immediate applications from which data marts can be built. Since the focus in the Kimball Stage 1 dimensional architecture is on the speed with which a data mart can be produced across a few applications, there is no time to build a "single version of the truth" across the enterprise.

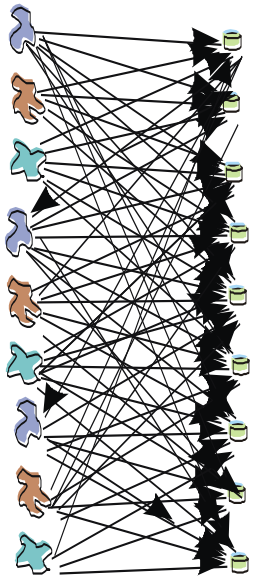
There is no denying that a corporate information factory requires much more time and many more resources to build than a simple dimensional architecture, primarily because the scope of the corporate

information factory is enterprise wide. The Kimball style simple dimensional architecture is unquestionably faster and easier to build. But the Kimball Stage 1 simple dimensional architecture does not contain the “single version of the truth” for the enterprise.

For small organizations with a small amount of data the Kimball Stage 1 simple dimensional architecture may be perfectly adequate. But for larger organizations with larger amounts of data and a need for integration of data cross the enterprise, the Kimball Stage 1 simple dimensional architecture soon becomes problematic. When the Kimball Stage 1 simple dimensional architecture is applied to large systems, the lack of the “single version of the truth” and the lack of the ability to integrate data across the organization becomes a large issue.

THE SIMPLE DIMENSIONAL MODEL IN THE LARGE ENTERPRISE

Consider what happens to the simple dimensional model in the face of a lot of data – there are lots of legacy sources and lots of data marts. The model **as illustrated by Kimball and Ross[1]** in Fig 3 merely expands. In the face of a large organization, the diagram drawn by Kimball and Ross that depicts the simple dimensional model simply grows larger. And with that expansion comes some major architectural problems. Fig 4 depicts a Kimball Stage 1 architecture for a large organization.



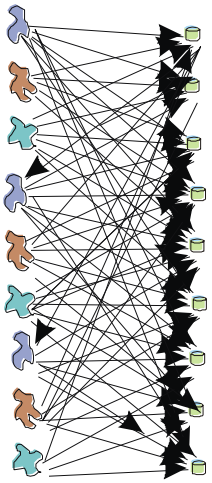
It is at this point that the Kimball architecture began to evolve into the next stage.

Evolution occurs because of the pain of problems. And there were adequate points of pain for large organizations that tried to implement the Kimball Stage 1 simple dimensional architecture for an evolution to occur.

One of the motivations for evolution is that there are many interface programs that are needed to support a Kimball Stage 1 simple dimensional architecture in a large organization. More pain arises when it comes time to maintain those interface programs. When the Kimball Stage 1 architecture is built for a large organization, there is enormous redundancy of data, from one data mart to the next. Another

motivation for evolution occurs when it is time to refresh data into the data marts. The window of opportunity for refreshment continues to shrink on a nightly basis. But perhaps the most pain with the Kimball Stage 1 simple dimensional architecture occurs because there is no corporately understood value of data, no "single version of the truth". In a large scale implementation of a Kimball Stage 1 simple dimensional architecture, when an end user wants to find a value of data, the end user literally has hundreds of places to turn to find that single value of data. In the Kimball Stage 1 simple dimensional architecture there is no one definitive place that states where a value of data is or is not. Consequently, a given value of data can reside anywhere (or nowhere) in a Kimball Stage 1 simple dimensional model. Since there is no definition of where there is a proper value of data, there can be many versions of the same value of data in a Kimball Stage 1 simple dimensional model in a large organization. Needless to say, large confusion results when large organizations turn the Kimball Stage 1 simple dimensional architecture into reality. If – as Kimball suggests (in his own words) – "a data warehouse is a union of all the data marts" – then there is a real problem with the data warehouse when it is based on the Kimball Stage 1 simple dimensional model.

Fig 5 suggests the major problems that arise with the Kimball Stage 1 data warehouse for a large organization. (Note – a small organization may not experience anywhere near the amount of grief that a large organization may experience. The size and the sophistication of the organization make a real difference in the amount of pain felt by an organization when it struggles with a Kimball Stage 1 dimensional architecture.)



- complex interface
- no reconciliation
- volumes of data

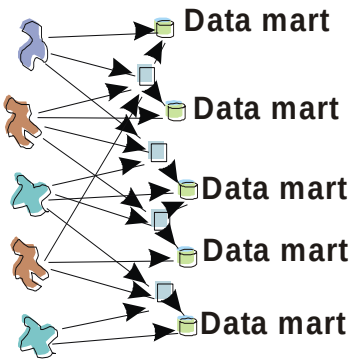
ENTER THE CONFORMED DIMENSION KIMBALL STAGE 2 ARCHITECTURE

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.

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.



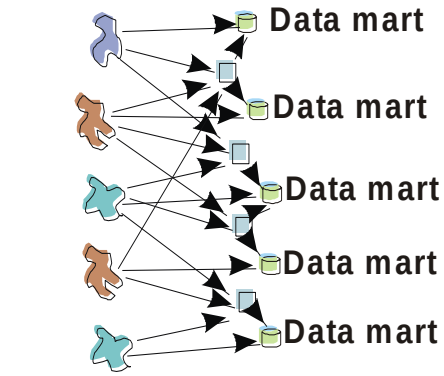
With a conformed dimension

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 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 problem 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.

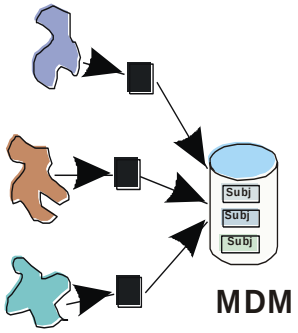
These issues with a Kimball Stage 2 conformed dimension architecture are seen in Fig 7



- redundancy within the data mart
- no reconciliation

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.

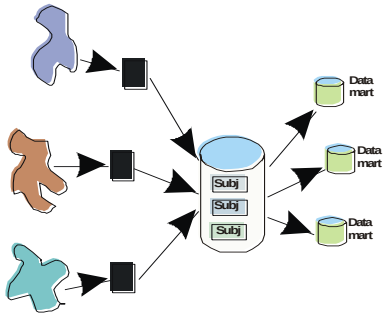


In the Kimball Stage 3 MDM architecture it is seen that there is at last corporate, enterprise wide integration of data. With MDM, now the “single version of the truth” exists. At this point, the focus on speed of building is completely lost because trying to integrate data across the enterprise is not 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.

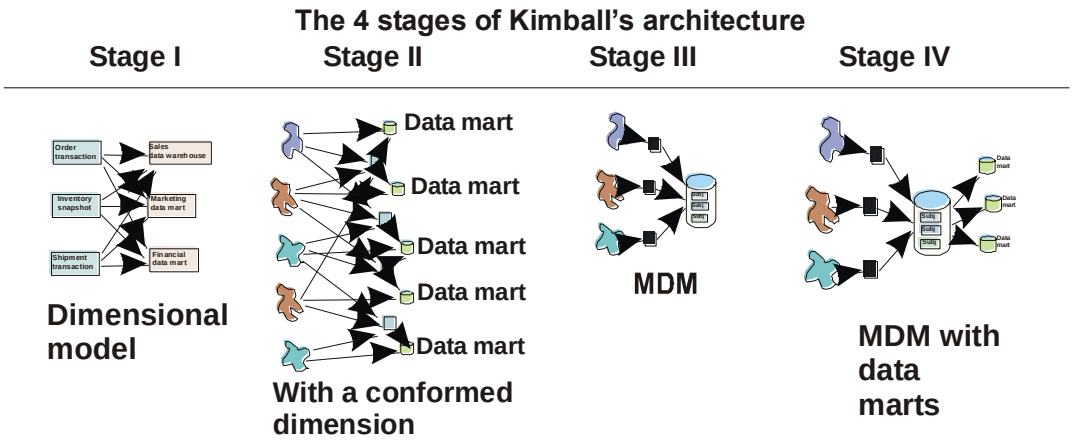
But there is yet another problem with the Kimball Stage 3 MDM architecture . This issue presages a next stage of evolution for the Kimball architecture.

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. or 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. Fig 9 depicts the predicted Kimball Stage 4 hub and spoke [4] 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

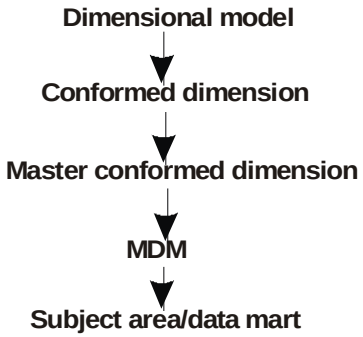
There is a certain irony here. Compare the predicted Kimball Stage 4 hub and spoke architecture with the corporate information factory architecture that was published by Inmon a decade earlier and it is seen that they in fact are the same. The emphasis for the predicted Kimball Stage 4 hub and spoke 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 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.

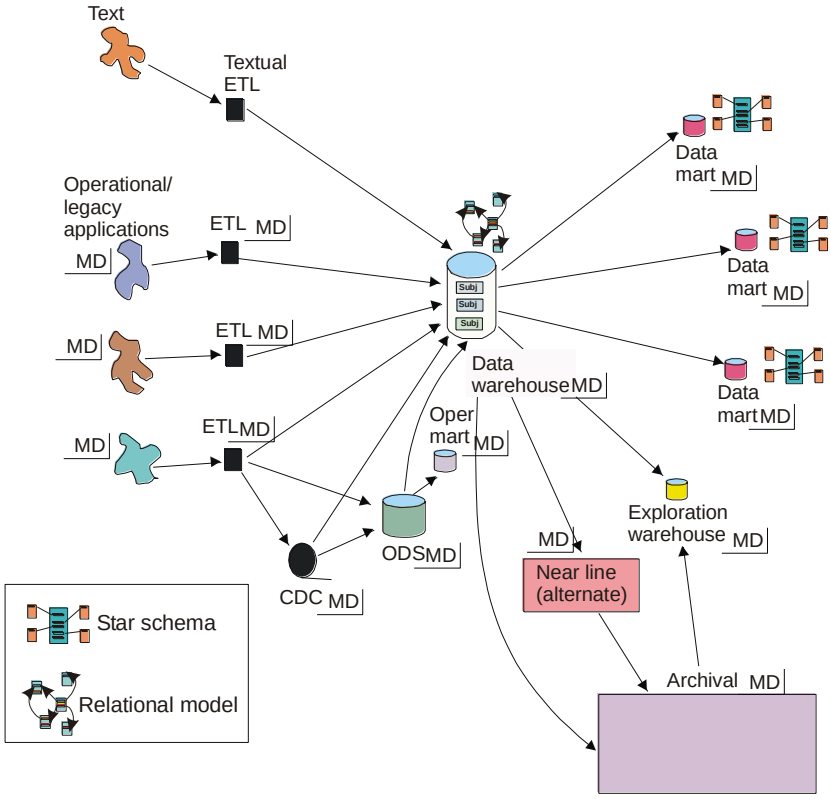


Fig 12 shows that in the center of the hub is the Inmon corporate information factory. In the Inmon corporate information factory is the ”single version of the truth”. The data here is granular, historical and integrated. The data here is cast in the form of the relational model.

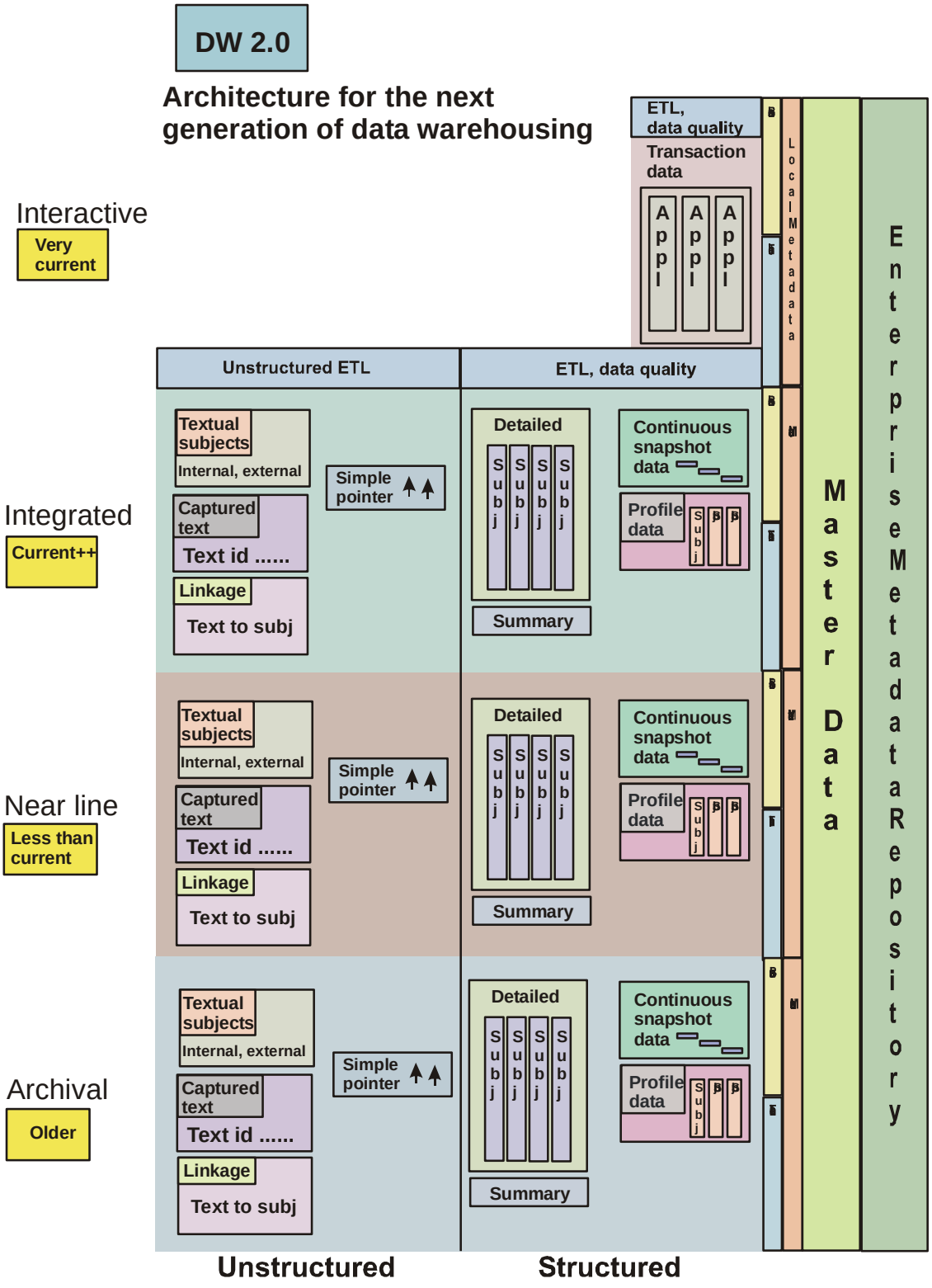
Surrounding the “single version of the truth” are the data marts. The data marts are cast in the form of the Kimball star schema architecture. In the star schema architecture, each data mart is optimized to meet the analytical needs of the end user. The source of data for each data mart is the data warehouse.

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
- 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.

Bibliography

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

[2][3] 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

[1] – Differences of Opinion: Comparing the Dominant Approaches to Enterprise Data Warehousing, Intelligent Enterprise magazine, 2004

[4] Internet – Planning MDM and EDW with Dr Kimball for 2010- Informatica