White Paper

Data Warehousing 2.0

Modeling and Metadata Strategies for Next Generation Architectures

By Bill H. Inmon
Forest Rim Technology, LLC

April 2010

Corporate Headquarters
100 California Street, 12th Floor
San Francisco, California 94111

EMEA Headquarters
York House
18 York Road
Maidenhead, Berkshire
SL6 1SF, United Kingdom

Asia-Pacific Headquarters
L7. 313 La Trobe Street
Melbourne VIC 3000
Australia

INTRODUCTION

Data warehousing has undergone a constant state of evolution since the beginning. Just when you think that everything has been discovered and developed, data warehousing evolves once again, mutating into a new form and structure.

EVOLUTION OF THE DATA WAREHOUSE

From the beginning the evolution of data warehousing has been shaped by powerful forces. First there was the need for access to data. Then there was the need for integration. Then came the need for a single version of the truth. Then different departmental needs for looking at the same foundation of data arose.

The general evolution of the first stages of the data warehouse environment is shown by Fig 1.

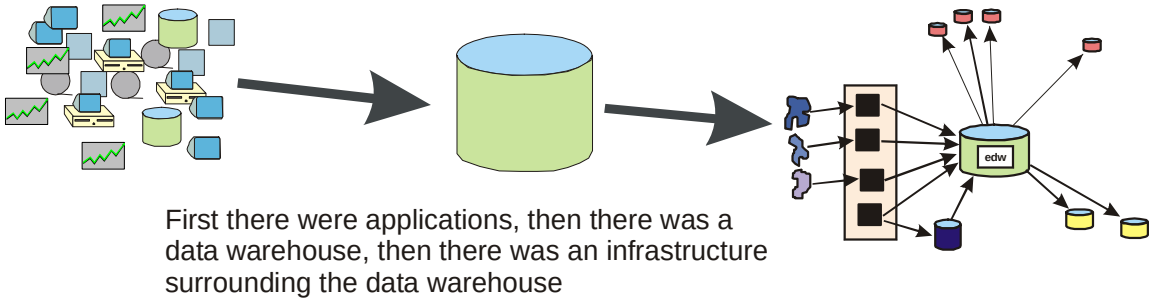


Figure 1

In the beginning there were applications. Applications grew so quickly that there developed what was termed the "spider web environment" where the same data was scattered all over the landscape. The frustration with the spider web environment led to the creation of the first data warehouse. The simple and singular data warehouse addressed many of the problems of the spider web environment.

But soon it was discovered that other components of the data warehouse infrastructure were needed. There was a need for ETL, the process that allows data to be read in an unintegrated application format and written out in a corporate, integrated format. Then there ware data marts, where different departments had their own version of the base data found in the data warehouse environment. Soon the ODS appeared where there was a need for high performance, transaction processing on integrated data.

An entire infrastructure grew up around the world of the simple data warehouse. An architecture called the "cif", or the corporate information factory, grew up around the data warehouse.

But the evolution of the data warehouse did not stop there. Soon the evolution of the data warehouse grew to include a broader set of requirements. Soon the notion of a data warehouse expanded to include a full set of data fulfilling a newly discovered set of requirements that reached well beyond the original notion of what a data warehouse should be.

DW 2.0: Architecture For The Next Generation of Data Warehousing

This new architecture was called "DW 2.0". Fig 2 shows the basic description of DW 2.0.

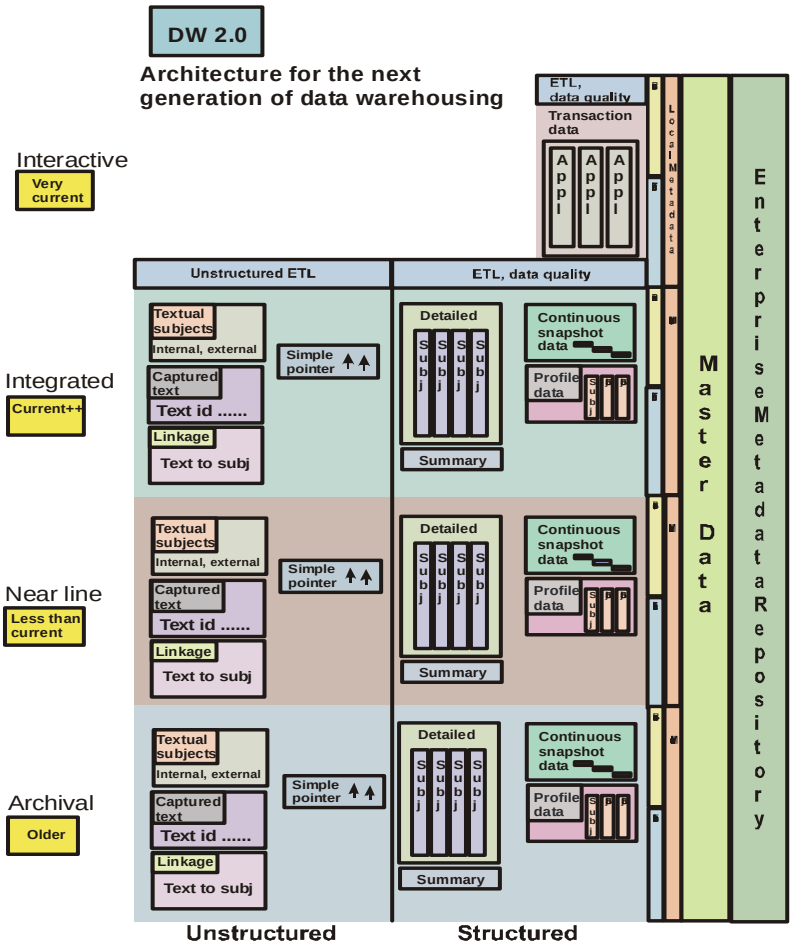


Figure 2

Like the earlier renditions of data warehouse architecture that preceded DW 2.0, DW 2.0 was shaped by powerful evolutionary forces.

The Life Cycle of Data

The first powerful evolutionary force that shaped DW 2.0 was the recognition that the data within the data warehouse contained its own life cycle. It was not enough to merely place data

on disk storage and call it a data warehouse. As time passed that data began to exhibit its own characteristics. The first manifestation of the life cycle of the data within the data warehouse was that over time, the probability of access to data in the data warehouse dropped. The older data became, the less that data was accessed. A second manifestation of the lifecycle of data with the data warehouse was that over time the volumes of data in the data warehouse grew rapidly. This led to a paradox: the larger a data warehouse became and the older the data in the data warehouse, the smaller percentage of data was being used.

UNSTRUCTURED DATA

The second manifestation of evolutionary forces in the data warehouse was the realization that unstructured, textual data belonged in the data warehouse. The original data that was placed in the data warehouse was transaction based, repetitive data. Many corporate decisions were based on this kind of data. But there is much important data that is not transaction based that belongs in the data warehouse.

METADATA

A third realization was that metadata belonged in the data warehouse as a formal and integral component. In first generation data warehouses, metadata was an afterthought. But there are powerful reasons why metadata needs to become an integral and formal part of the data warehouse environment.

Fig 3 shows the evolutionary forces and their effect on the DW 2.0 environment.

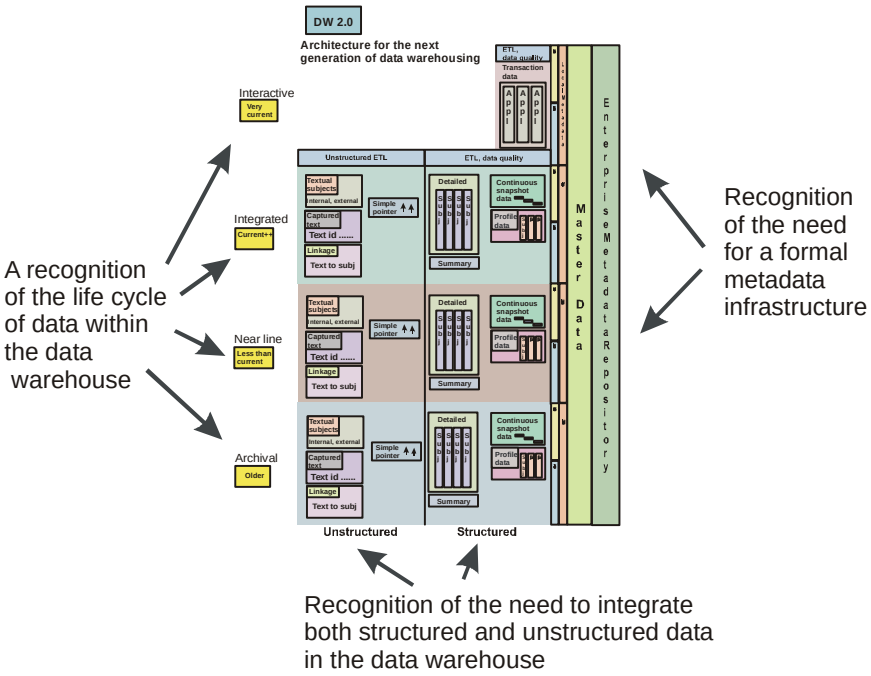


Figure 3

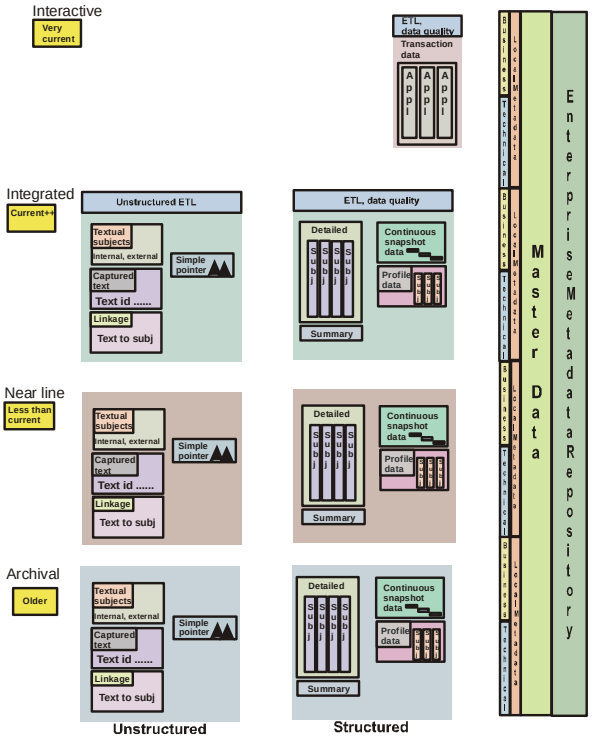
DW 2.0 has become the architectural paradigm for modern data warehouses. For a detailed description of DW 2.0 refer to the book DW 2.0 ARCHITECTURE FOR THE NEXT GENERATION OF DATA WAREHOUSING, Morgan Kaufman, 2008.

Like most large architectures, DW 2.0 is not build all at once. Instead an organization builds first one component of DW 2.0 then another component. Indeed some components of the data warehouse are optional, such as the near line sector.

COMPARTMENTS OF PROCESSING

In any case, the different components of DW 2.0 consist of sectors that perform one basic function. These sectors are in a sense their own small operating environments. Each sector has its own purpose and its own functionality.

Fig 4 shows that each sector is neatly compartmentalized.



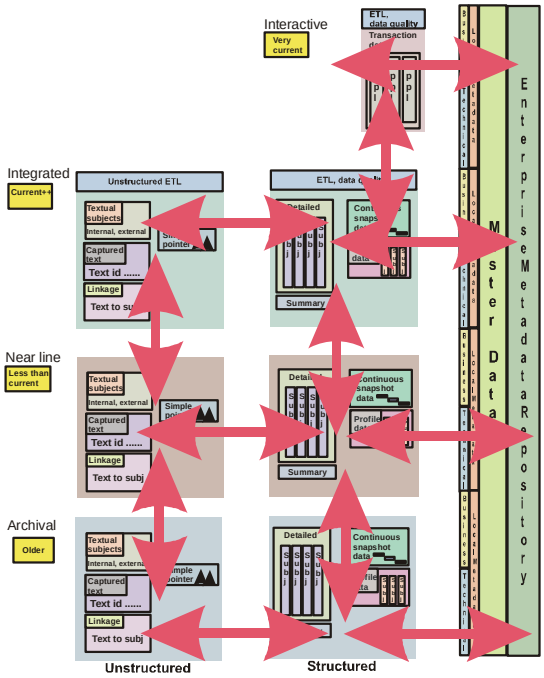
DW 2.0 is made up of different components

Figure 4

In light of the individual compartmentalization of the sectors of DW 2.0, the question then becomes – how is work distributed and coordinated across the different components? The answer is that work is distributed and coordinated by means of the passage of metadata from one component to the next.

METADATA AS THE GLUE

Fig 5 shows that metadata is the "glue" that binds the different operating components of DW 2.0 together.



There is a need for tight communications and tight coordination among the different components of DW 2.0

Figure 5

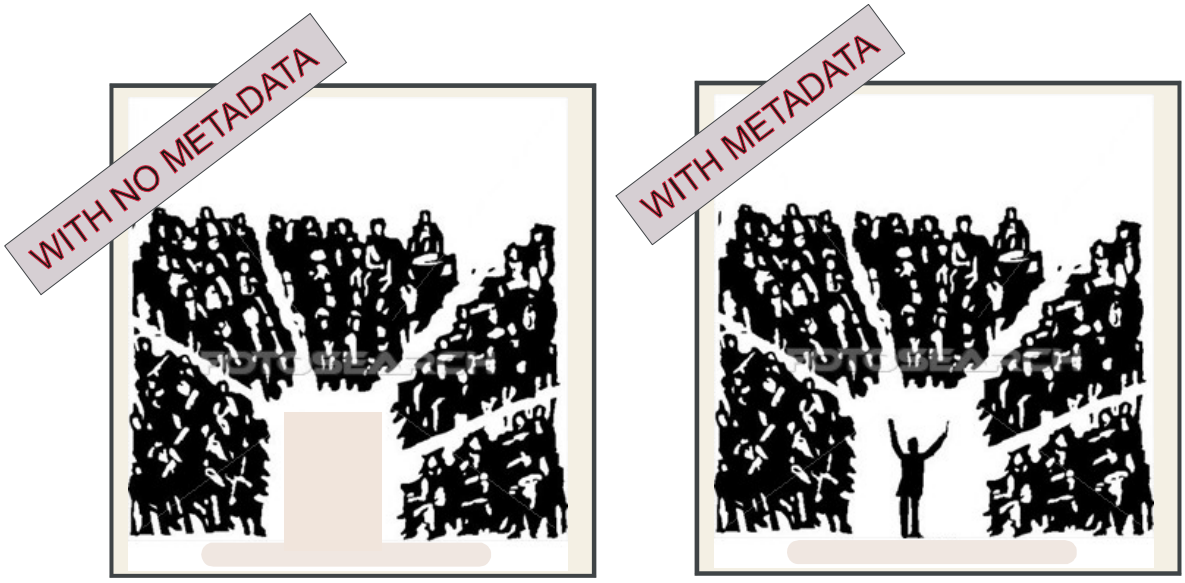
In a sense, metadata forms a lattice work that holds the components of DW 2.0 together. It is the passage of metadata that allows the different components of DW 2.0 to work in cooperation and in coordination with each other.

Stated differently, without metadata, the different components of DW 2.0 would not be able to achieve a coordinated and cohesive work flow.

Consider a symphony orchestra. What kind of music would be created if the violins were playing Beethoven's Fifth, the cellos were playing "Hotel California", the drums were playing Aretha Franklin's "Respect", the flutes were playing "Happy Birthday", and the trumpets were playing "Jingle Bells". The result would be a bunch of noise – nothing that anyone would want to listen to.

But now suppose a conductor steps in and gets everyone to play DeBussy's Le Mer. Now, suddenly the very same components that just a few seconds ago sounded awful are making very beautiful music. What is needed is a conductor, and it is metadata that plays the role of the conductor. Metadata of course does not direct a symphony orchestra. Metadata instead directs the DW 2.0 environment, with all of its different components and different technologies.

Fig 6 shows the role of metadata and the role of a conductor.



The orchestra with no conductor

The orchestra with a conductor

Figure 6

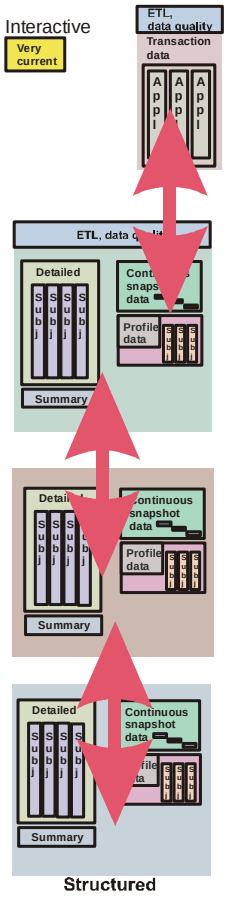
So what kind of metadata is used to be passed from one DW 2.0 component to another? In truth there are all sorts of metadata that are passed from one DW 2.0 component to another.

Different Types of Metadata

Typical types of metadata that are passed include:

- data descriptions
- data definitions
- formula and algorithms
- the timing of processing
- description of assorted volumes of data
- operating parameters
- encoded values
- processed/calculated/derived data
- ratios and statistics
- status codes, and the like

Fig 7 shows the different kinds of metadata that can be passed from one component to the next. (Note: this list is merely a representative sample of the different types of metadata that can be passed from one component to the next.)



The data that is passed from one component to the next includes -

- data descriptions
- parameters
- processed data
- ratios and statistics
- status codes
- and so forth

Figure 7

The data descriptions that can be passed from one component of DW 2.0 to another often reflect on the data model that exists for each component of DW 2.0. Indeed, there is a data model for each of the different components of the DW 2.0 environment.

DIFFERENT DATA MODELS

Fig 8 shows the different data models that exist for the different components of DW 2.0.

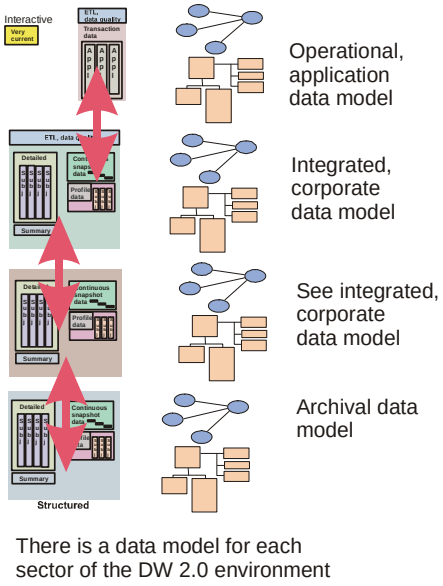


Figure 8

At the operational level (i.e., the interactive sector of DW 2.0) is an application oriented data model. This data model reflects the transactional nature of the work that is done here. While integration may occur here, it often doesn't. In addition, this level of data often includes the reception of data from applications that lie outside of it.

The second data model that is found is that of the data model for the integrated sector. The integrated sector is the place where corporate integration occurs. The integrated data model is a classical subject oriented, non volatile model.

The third data model that (optionally) appears is the data model for the near line sector. If the near line sector is part of the data warehouse environment, then the near line data model is IDENTICAL to the integrated data model. Stated differently, if the near line sector appears at all, then the data model for the near line sector is IDENTICAL, in every way, to the integrated data model.

The fourth data model found in the DW 2.0 environment is the data model for the archival environment. The archival data model is the data model that reflects several important aspects of design:

- The need to look at and measure data over lengthy periods of time
- the need to need to reflect a changing metadata structure over time
- The need to store metadata directly in the same physical volume of data as the actual data itself,
- The need to store calculation algorithms along with summarized or aggregated data
- The need to be able to free data from its software structuring

- The need to be able to further normalize data as it is placed in the archival environment, and so forth.

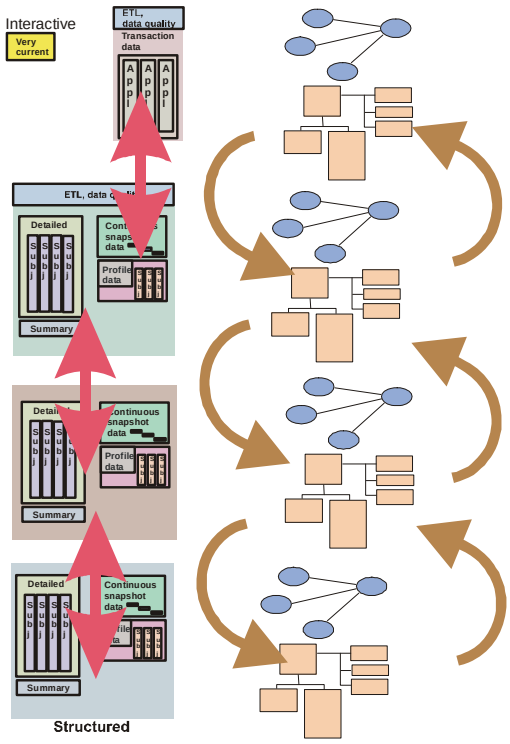
There are then some distinctive needs for a data model for each of the different layers processing that occur in the DW 2.0 environment.

A Common Theme

It is of interest that the different data models for the different sectors of DW 2.0 are in fact very different. Having stated that, there never the less is a common theme running through each of the different data models. For example, the integrated data model is undeniably akin to its interactive data model. And the data model found at the archival environment is undeniably related to the integrated data model.

The similarities of the different data models to each other are somewhat akin to looking at a grandfather, his daughter, and the child of the daughter. There will be certain facial and body similarities throughout the family. But no one will mistake a grandfather for his daughter, and no one will mistake a daughter for her child.

Fig 9 shows that there are definite familial similarities running one data model and the next.



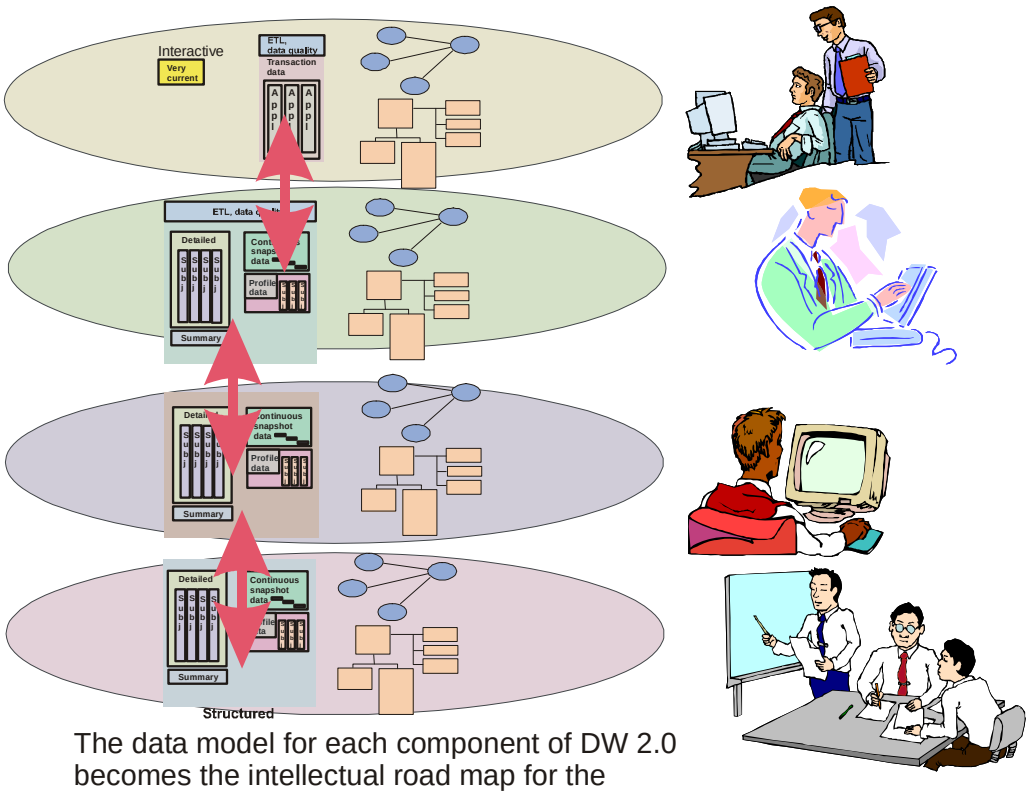
There is a great deal of commonality from one type of data model to the next

Figure 9

THE TECHNICAL COMMUNITY

So who do the metadata and the data model aspects of the DW 2.0 environment help? The first set of people that are helped by the data models and the metadata is the development and maintenance organizations. The technicians of the organization find that data models and metadata become the Bible of development and maintenance. The data models and the metadata become the true "intellectual roadmap" for the ongoing development and maintenance of the data warehouse environment. Stated differently, without metadata and a data model, the technician is like a mechanic working on a Yugo where the Yugo is no longer produced and where the manuals describing the Yugo are either all lost or all in Serbian, where the mechanic only reads and speaks English. Trying to do a repair job on a Yugo under those circumstances is questionable under the best of circumstances.

Fig 10 shows the use of the metadata and data models by the technical community.



The data model for each component of DW 2.0 becomes the intellectual road map for the building and the maintenance the data structures found in the component

Figure 10

THE END USER ANALYTICAL COMMUNITY

But there is another audience that is well served by metadata and data models, and that audience is the end user analyst.

Consider a newly hired end user analyst that has just received an office right across from you. The end user analyst has just received an MBA and wants to show his/her stuff. The end user

analyst starts to build an analysis and becomes perplexed. There are so many types of data, so much similar data, so much data that is of different ages that the analyst does not know where to start. Merely biting off a chunk of data may be very misleading if the analyst doesn't chose the proper data. And there is a lot to choose from.

So the analyst comes to your office and asks –"how do I find my way around this environment? There is a lot to choose from and I don't want to start with the wrong data."

It is at this point that the metadata and the data models become invaluable. Without the metadata and the data models the end user analyst may wander around the maze of data for a long time. But with the metadata and the data models, the end user analyst can determine what data is where and what data provides the best source for the analysis at hand.

Fig 11 shows that metadata and the data models are extremely useful to the analytical community as well as the technical community.

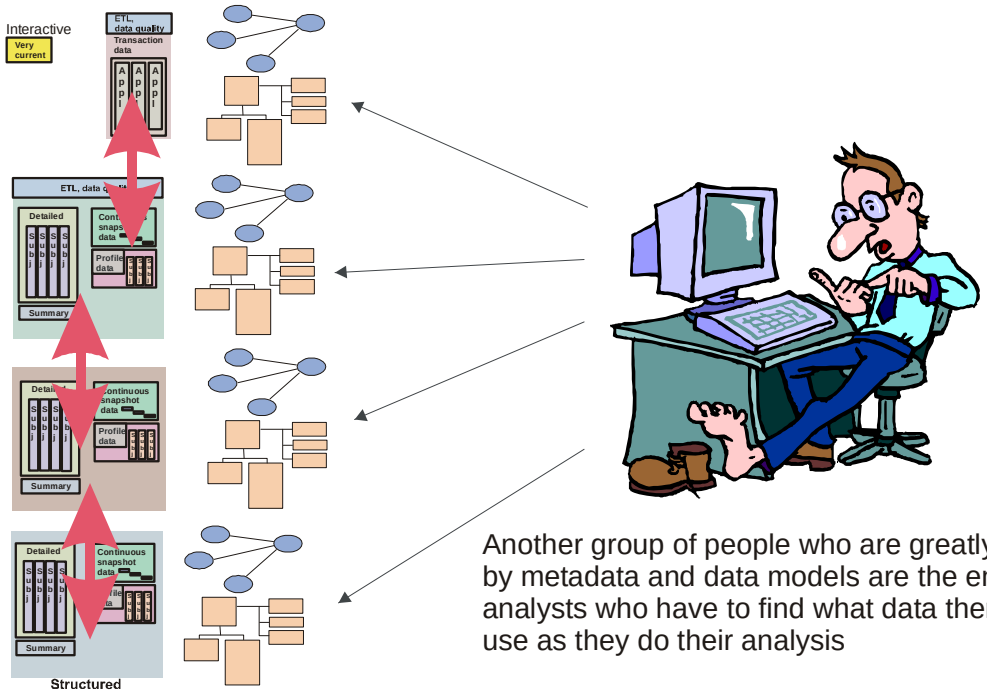


Figure 11

SUMMARY

In this paper we have discussed DW 2.0 and the evolution of the architecture surrounding data warehouse. DW 2.0 has different sectors. There is a need for metadata to control the activities in each of these sectors. In addition there is a data model for each of the sectors. Each data model is different and yet there is a common theme running through each data model.

The data models and the metadata serve two communities – the technical community and the end user analysis community. For the technical community the metadata and the data models act as an intellectual roadmap. For the end user analytical community, the data models and the

metadata serve as a guidepost for where the end user analyst should go to in order to find the proper data for analysis.

REFERENCES

Inmon, W H - DW 2.0 ARCHITECTURE FOR THE NEXT GENERATION OF DATA WAREHOUSING, Morgan – Kaufman, 2008.

ABOUT THE AUTHOR

Bill Inmon, "the father of data warehousing", has written 50 books translated into 9 languages. Bill founded and took public the world's first ETL software company. Bill has written over 1000 articles and published in most major trade journals.

Bill has conducted seminars and spoken at conferences on every continent except Antarctica. Bill holds three software patents. Bill's latest company is Forest Rim Technology, a company dedicated to the access and integration of unstructured data into the structured world. Bill's website – inmoncif.com - has attracted over 1,000,000 visitors a month. Bill's weekly newsletter in b-eye-network.com is one of the most widely read in the industry and goes out to 75,000 subscribers each week.

Embarcadero Technologies, Inc. is a leading provider of award-winning tools for application developers and database professionals so they can design systems right, build them faster and run them better, regardless of their platform or programming language. Ninety of the Fortune 100 and an active community of more than three million users worldwide rely on Embarcadero products to increase productivity, reduce costs, simplify change management and compliance and accelerate innovation. The company's flagship tools include: Embarcadero® RAD Studio, DBArtisan®, Delphi®, ER/Studio®, JBuilder® and Rapid SQL®. Founded in 1993, Embarcadero is headquartered in San Francisco, with offices located around the world. Embarcadero is online at www.embarcadero.com.