



The MIT 2008 Information Quality Industry Symposium



Improving your Data Warehouse's IQ



Derek Strauss
Gavroshe USA, Inc.



The MIT 2008 Information Quality Industry Symposium



Outline

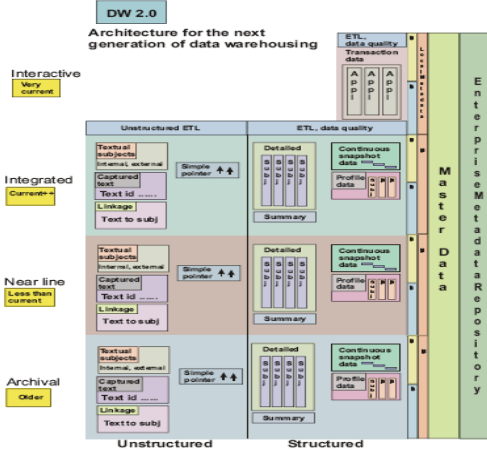
- **Data quality for second generation data warehouses**
- **DQ tool functionality categories and the data quality process**
- **Data model types across the DW2.0™ database landscape**
- **Challenging top-down from the bottom**
- **Deriving an interlocking set of models**



The MIT 2008 Information Quality Industry Symposium



DW2.0™ – Architecture for the next generation of data warehousing





DW 2.0 is a trademark of Bill Inmon. All rights reserved
 © "Architecture for the next generation of data warehousing" is copyrighted by Bill Inmon, 2006



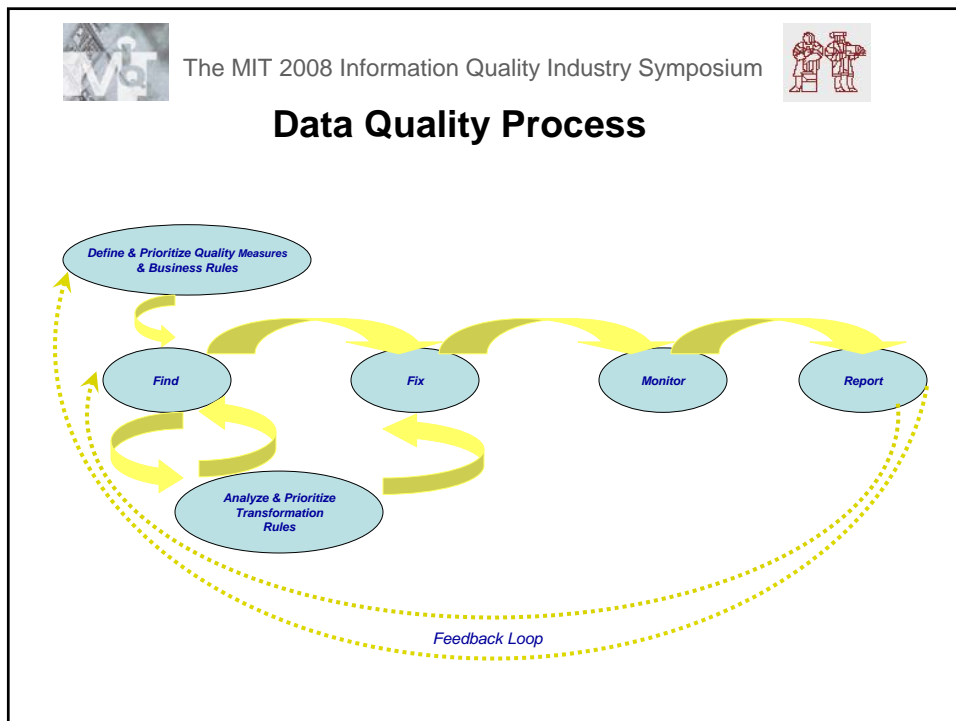
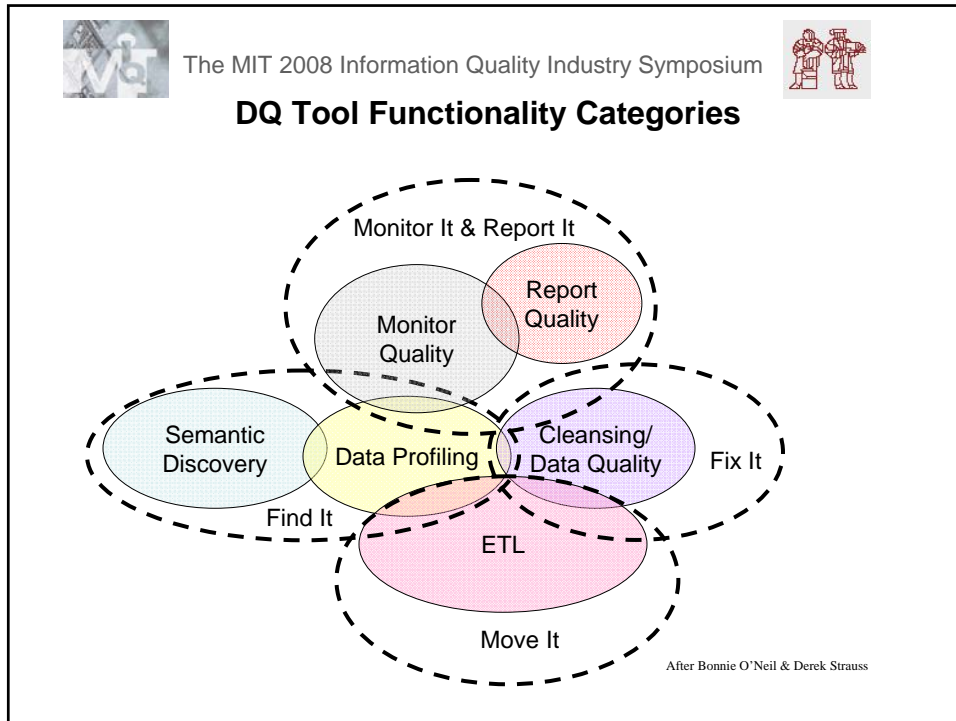
The MIT 2008 Information Quality Industry Symposium



Data Quality for second generation data warehouses

- Getting away from “code, load and explode” 
- The “data scouts” (team with business and IT representation) 
- On finding significant DQ problem, choose from strategies such as:
 - fix the data at the source (actually go into the data store and physically zap the data)
 - fix the program the source (apply the correct edits in order to validate the data)
 - fix the business process (a broken business process is very often the main cause of poor quality data)
 - recognize and resolve situations where data attributes are being used for a purpose other than their original intent (e.g. a gender code, which has more than two distinct values)
 - transform the data on the way into the data warehouse (this is the most common of strategies, but should not be the only strategy employed)*

* In the case of the latter strategy, it is important to note that there are two alternative implementations for transforming data on the way into the integrated sector. The first implementation is to simply change the data and load it into the warehouse. The second implementation scenario does that and more: it will actually load the unchanged data alongside of the changed data. There are many times when this may be a preferable route to go.





The MIT 2008 Information Quality Industry Symposium



DATA PROFILING TOOLS AND THE REVERSE ENGINEERED DATA MODEL.

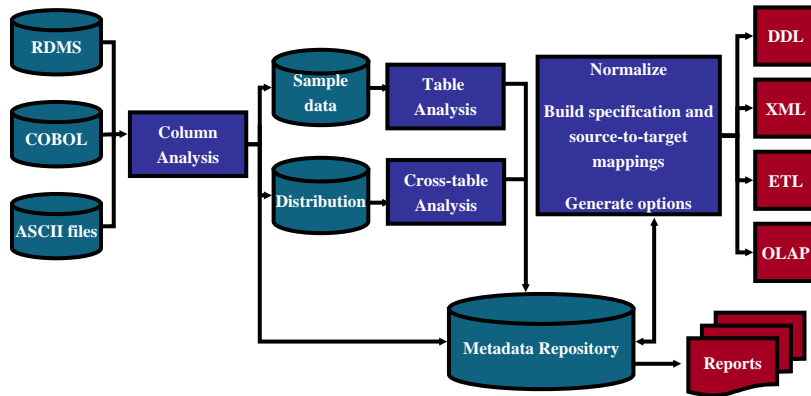
- Today there are many data profiling tools to assist the data quality team.
- The tools facilitate analysis of the data values held in a column; sometimes looking simultaneously at multiple columns in a table; sometimes even looking across tables, and even across systems to see if there are any patterns in the values held in the selected columns.
- These patterns can uncover hidden business rules, e.g. every time the value in column 1 is "a" we see that the value in column 5 can be "x" or "y".
- The best of these data profiling tools will go one step further: having analyzed the actual data values in the columns of a system, the tools can suggest a normalized schema.
- What in effect happens is the tool develops a third normal form data model, based on bottom up analysis, abstracted from the actual data values in the physical database.
- This abstracted data model is very useful input into the top down modeling process, which should be happening in parallel with the development of the warehouse.
- In fact, in the case of a DW 2.0 warehouse, we want to ensure that a high quality data model architecture is being developed, as this will greatly assist in improving the data quality of the enterprise data warehouse.

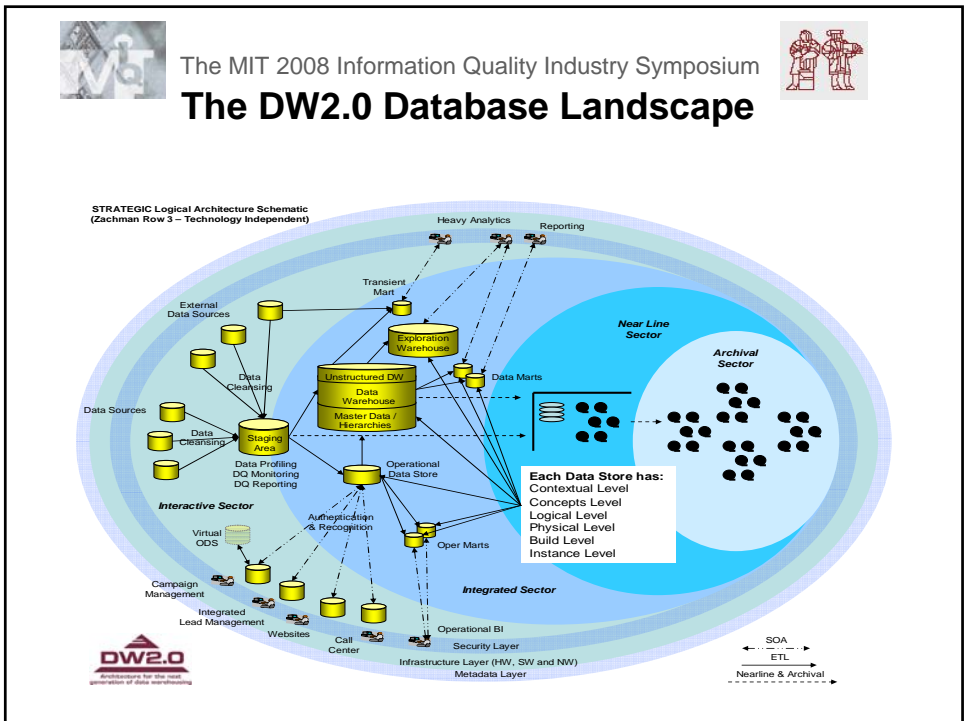
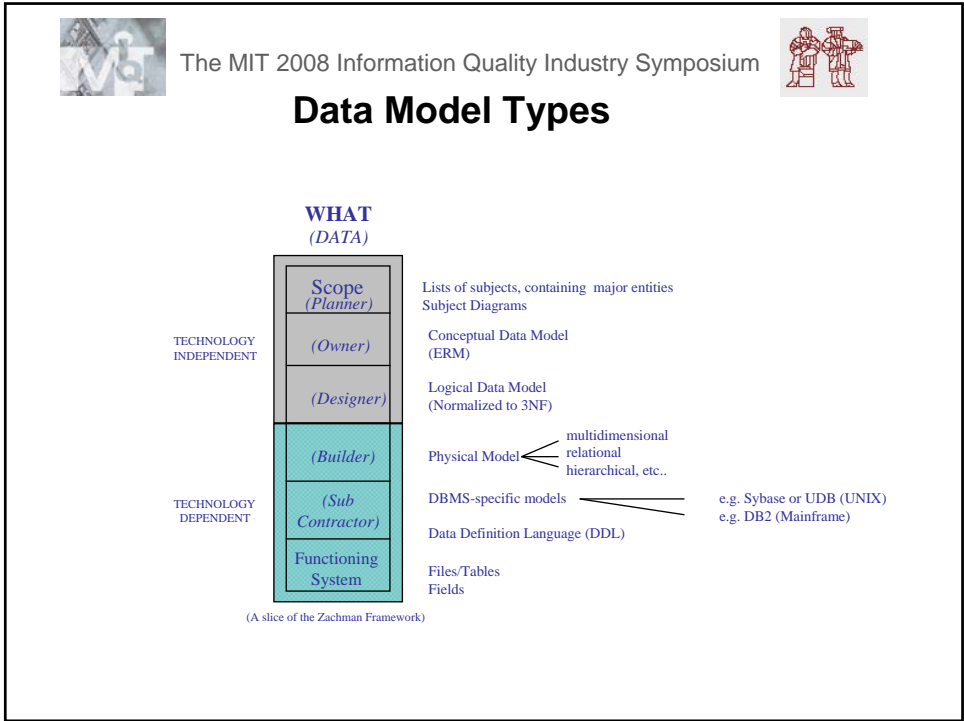


The MIT 2008 Information Quality Industry Symposium



Automated Methods







The MIT 2008 Information Quality Industry Symposium



DATA MODELS ACROSS THE FOUR DW2.0 SECTORS

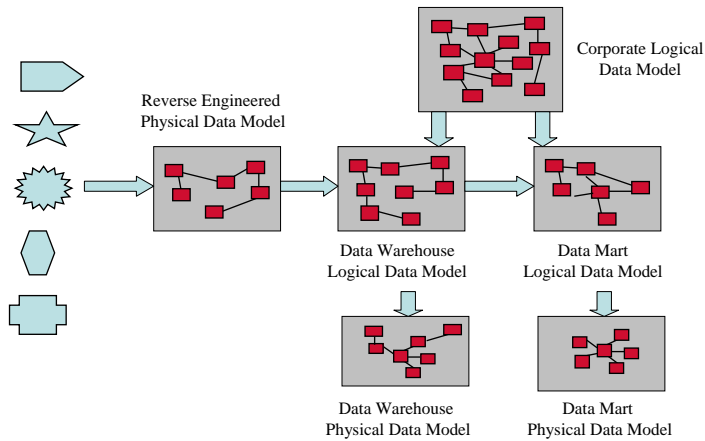
- A good concepts data model helps you to understand the major concepts in the business and how they interrelate.
- A good third normal form logical data model helps you understand all the attributes pertaining to the business entities and also the cardinality and optionality of the relationships between those entities. This model gives a great logical view of the business and its data and should be the starting point for the third model type – i.e. the physical data model.
- Physical data models for DW 2.0's integrated sector can differ widely in their structure. They will range from normalized and near normalized models for the data warehouse hub through to star schema and snowflake schema models for the data marts. Still other structures would be best suited for exploration warehouses, data mining warehouses, operational data stores, and opermarts.
- Data moving to the nearline sector should be kept as close to third normal form structure as possible; it is normal for data to be restructured as it enters the archival sector. It is important to the DW 2.0 world that there should be multi-directional traceability between these models: it should be possible to navigate from a physical model back up through the logical model and up still further to the concepts model; in like manner, we should be able to move from the top down from a concepts model to the logical data model and on to the physical data models.
- A rigorous set of interlocking models will go a long way towards improving the quality of the data in the enterprise, linking business meaning and structural business rules to physical instantiations of data entities and attributes.

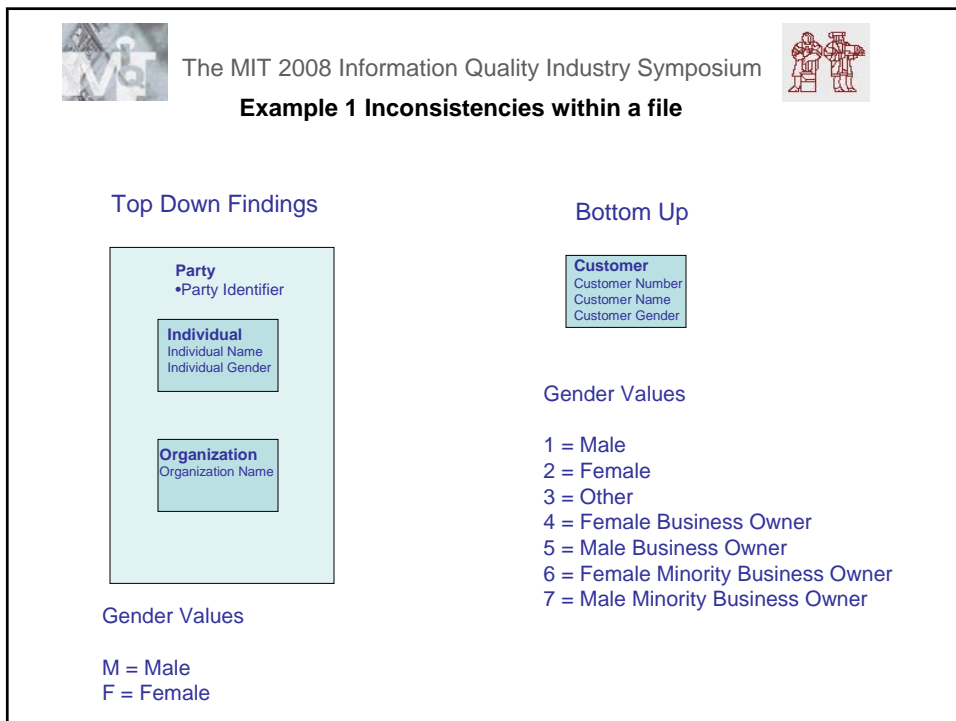
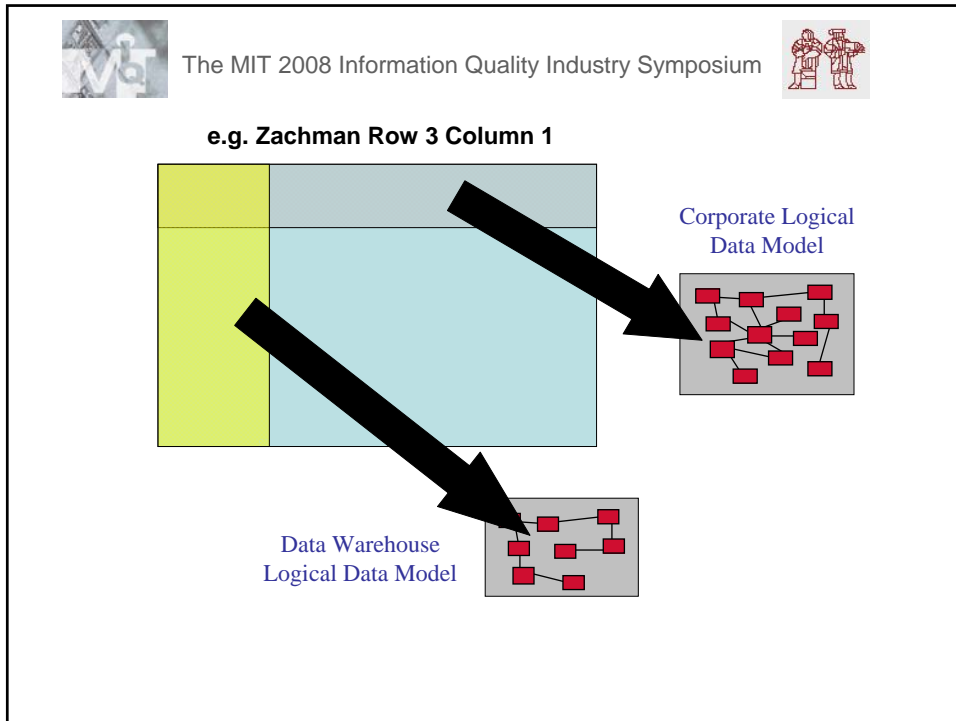


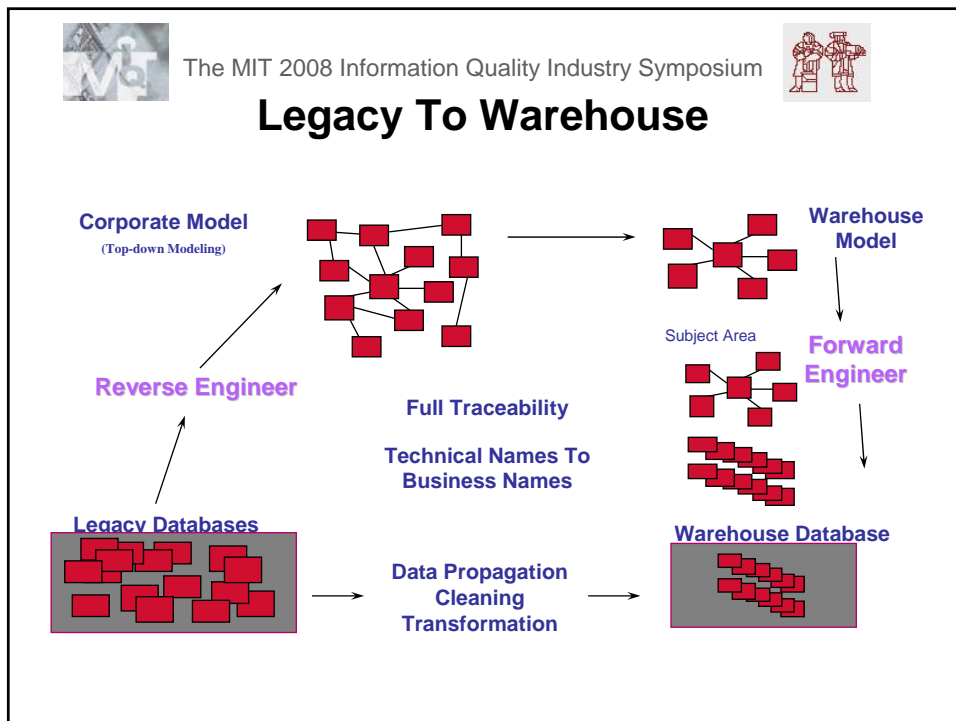
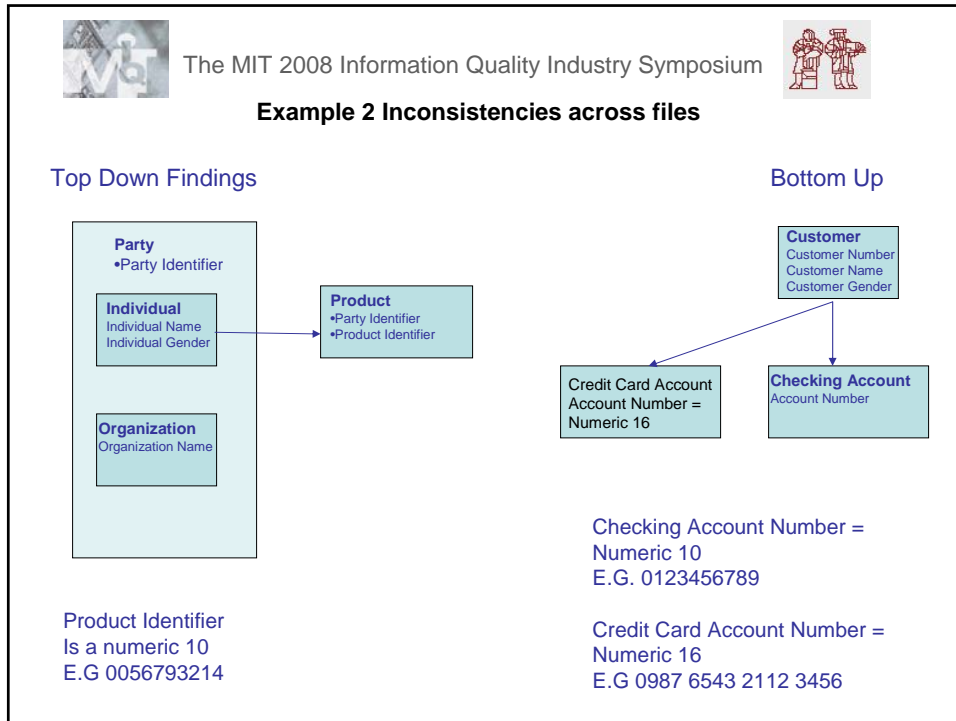
The MIT 2008 Information Quality Industry Symposium

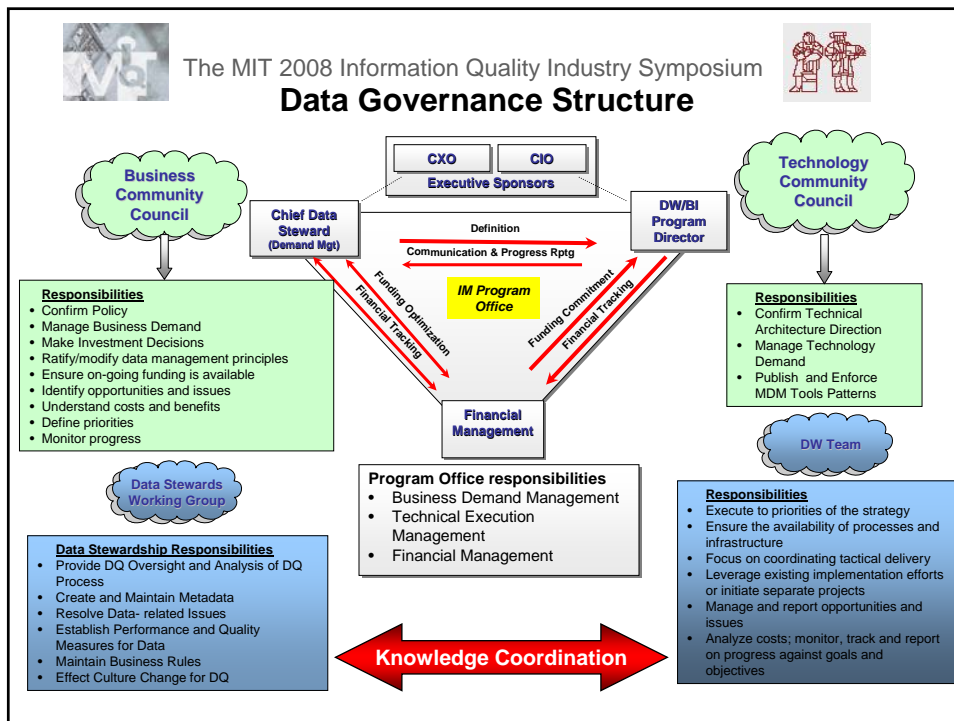
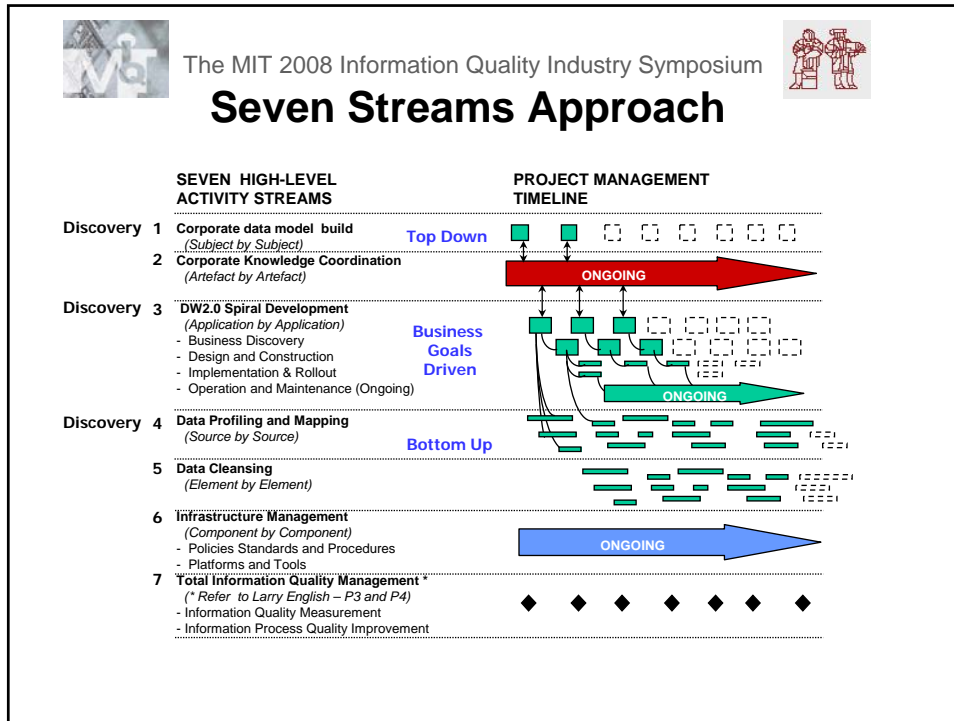


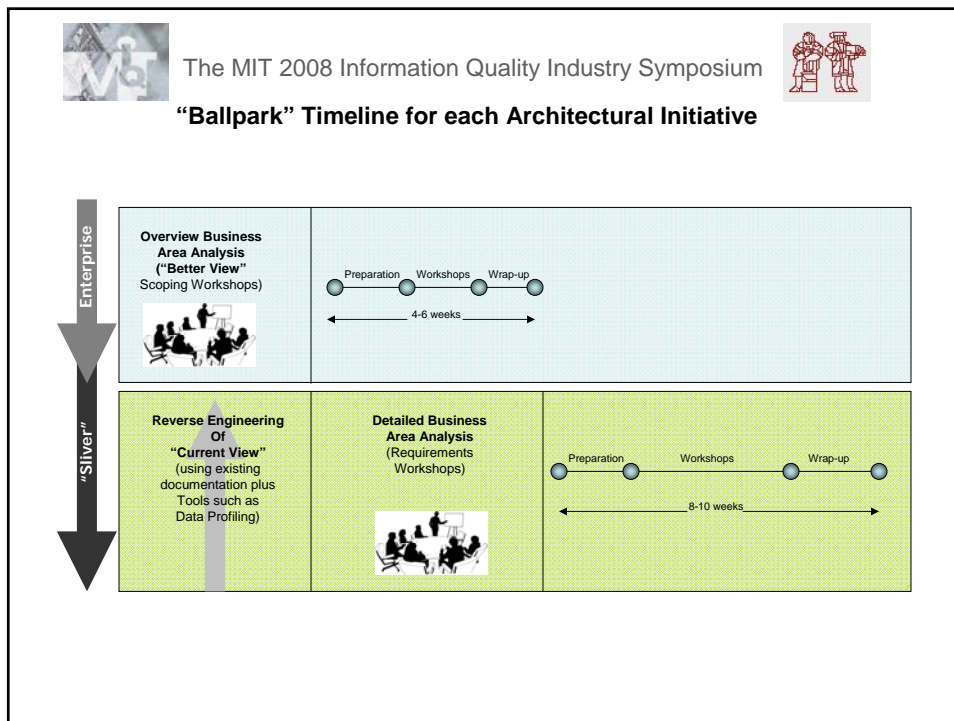
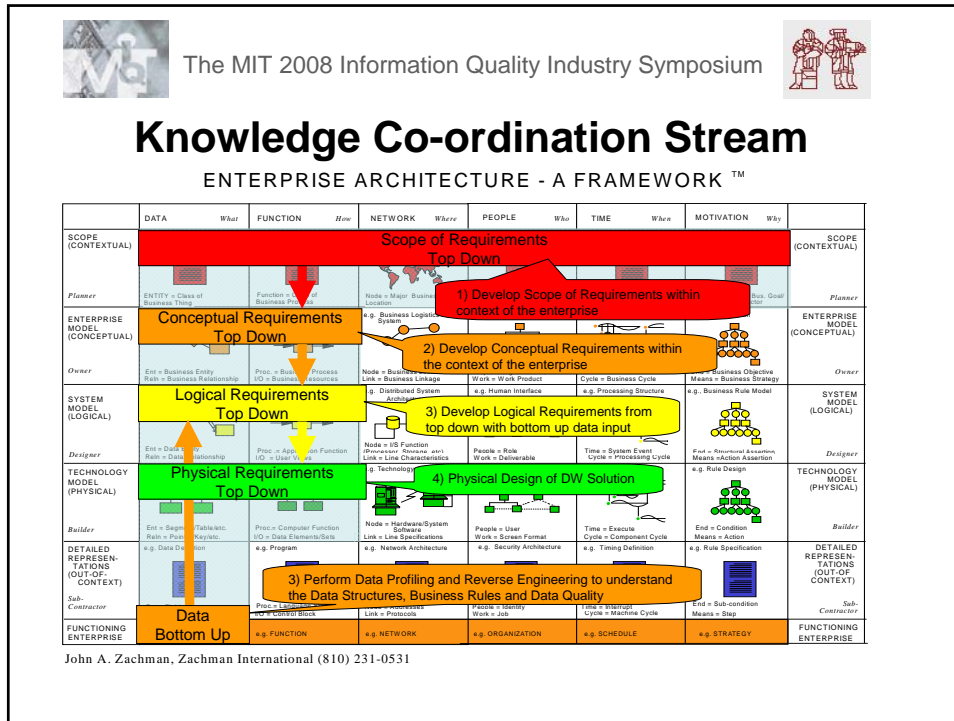
Data Models Needed

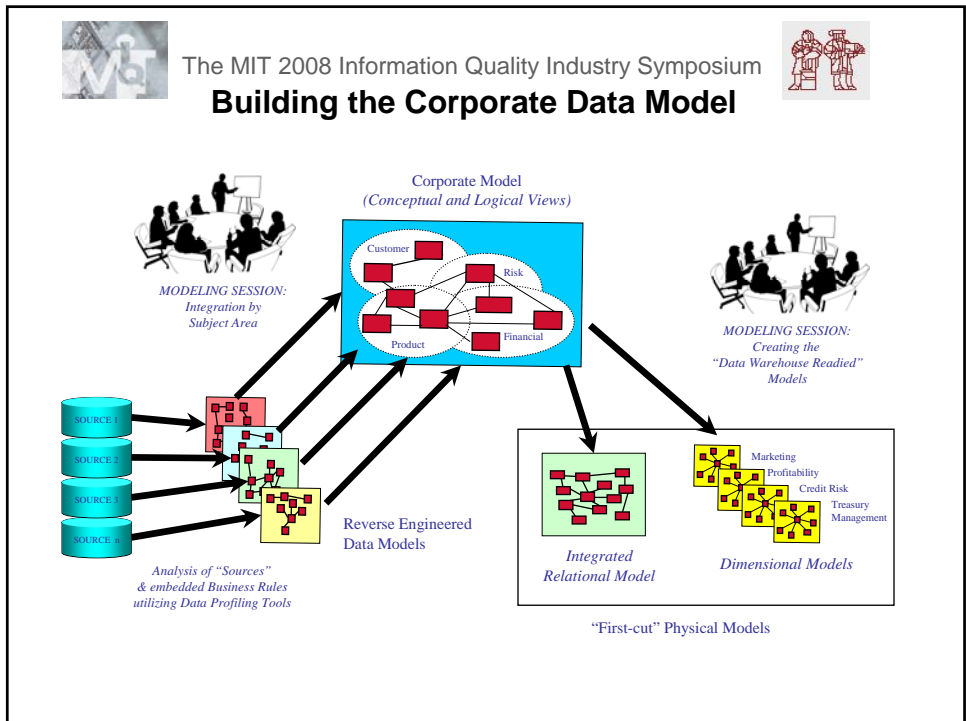
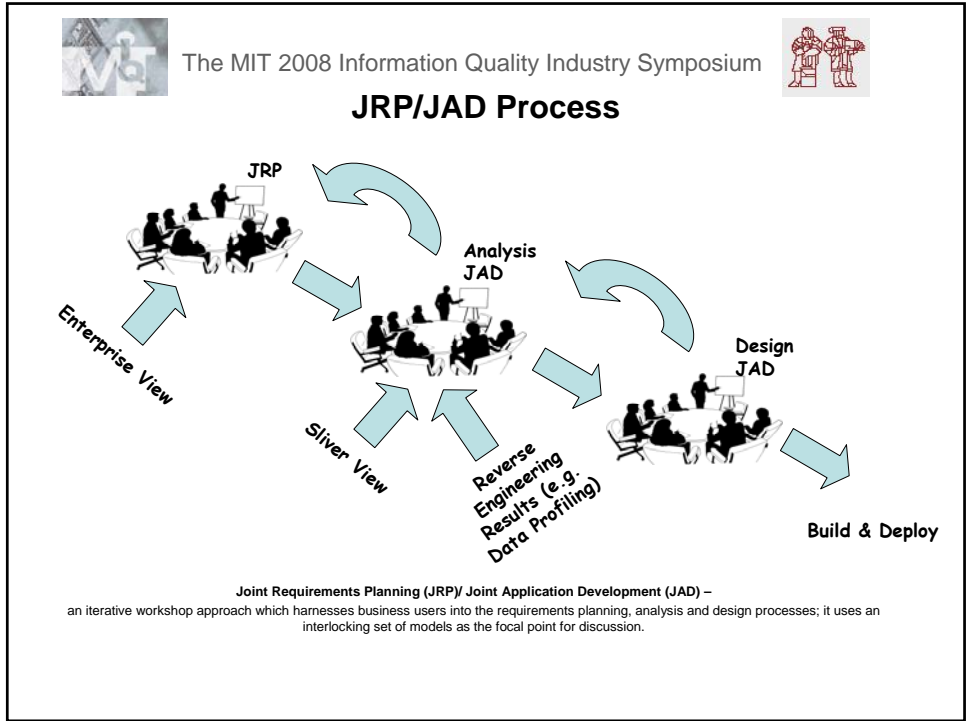


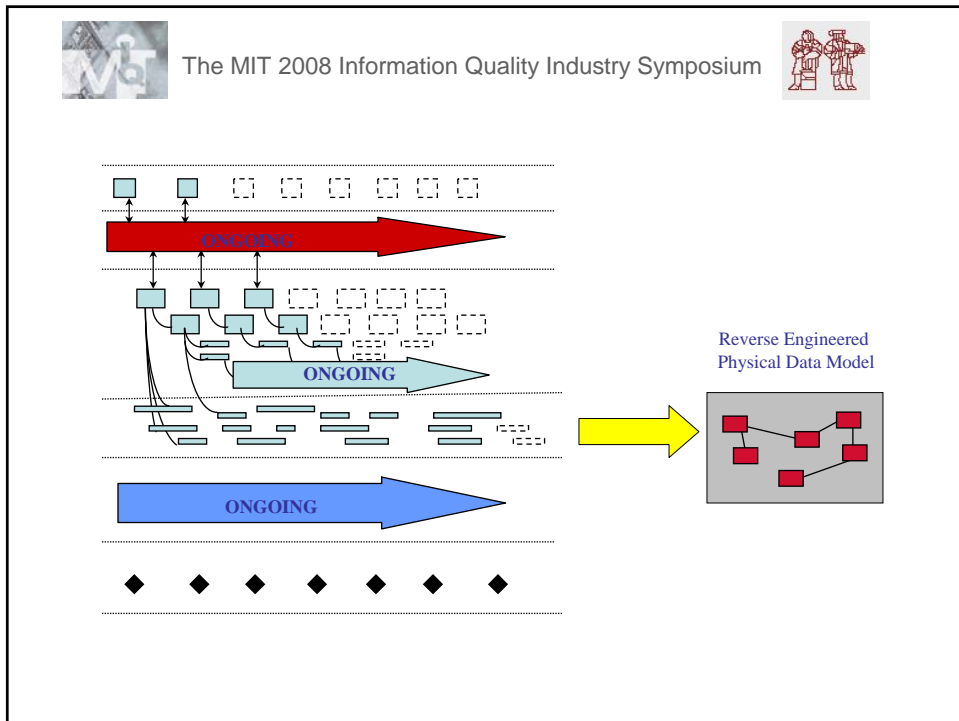
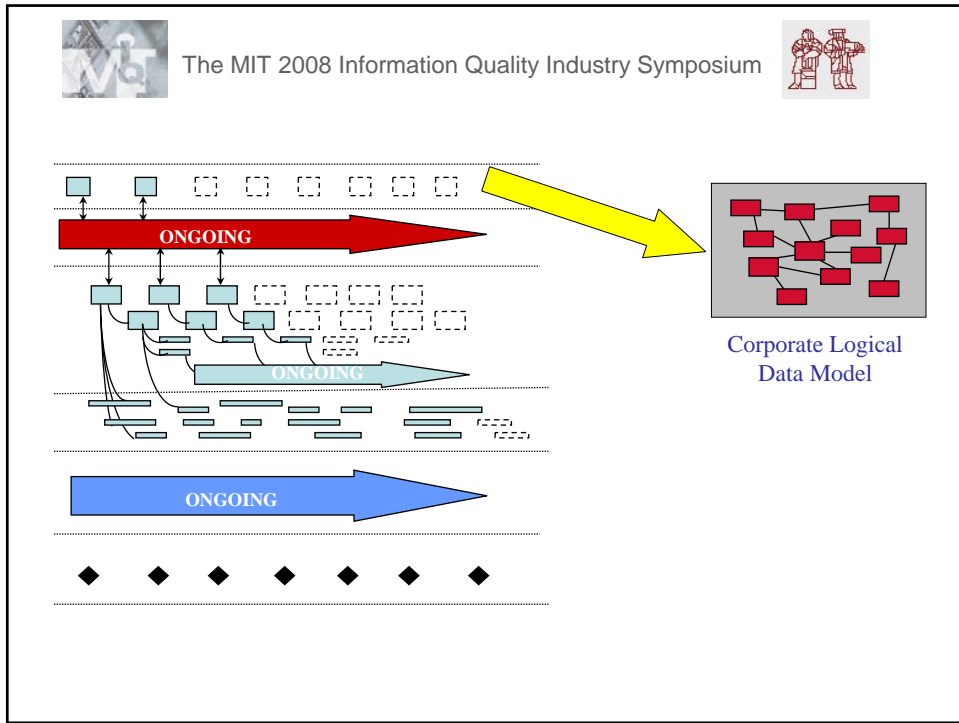


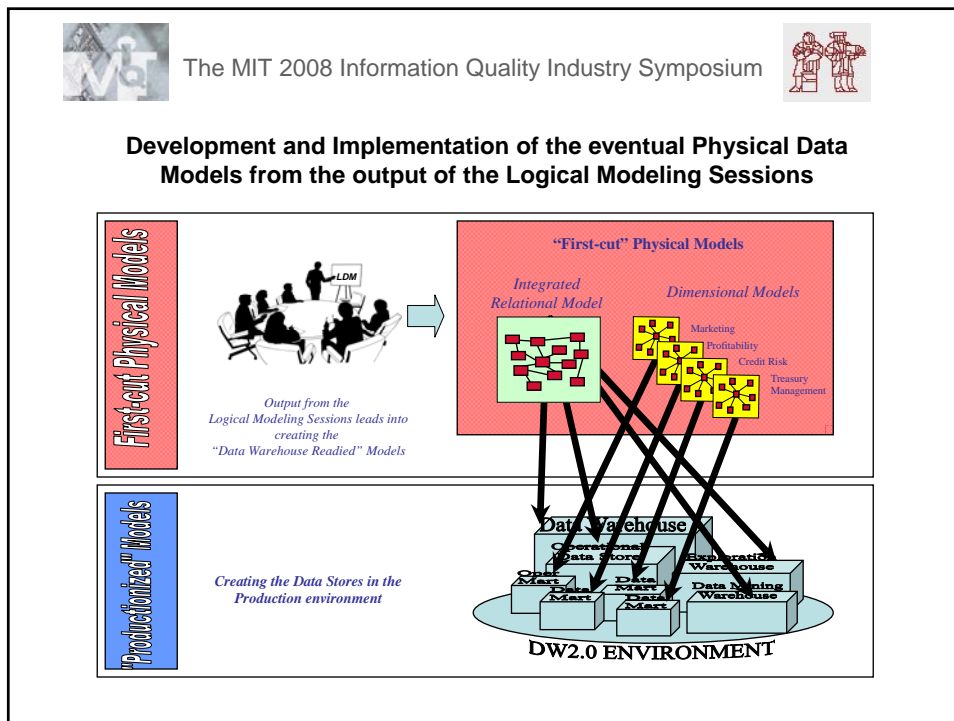
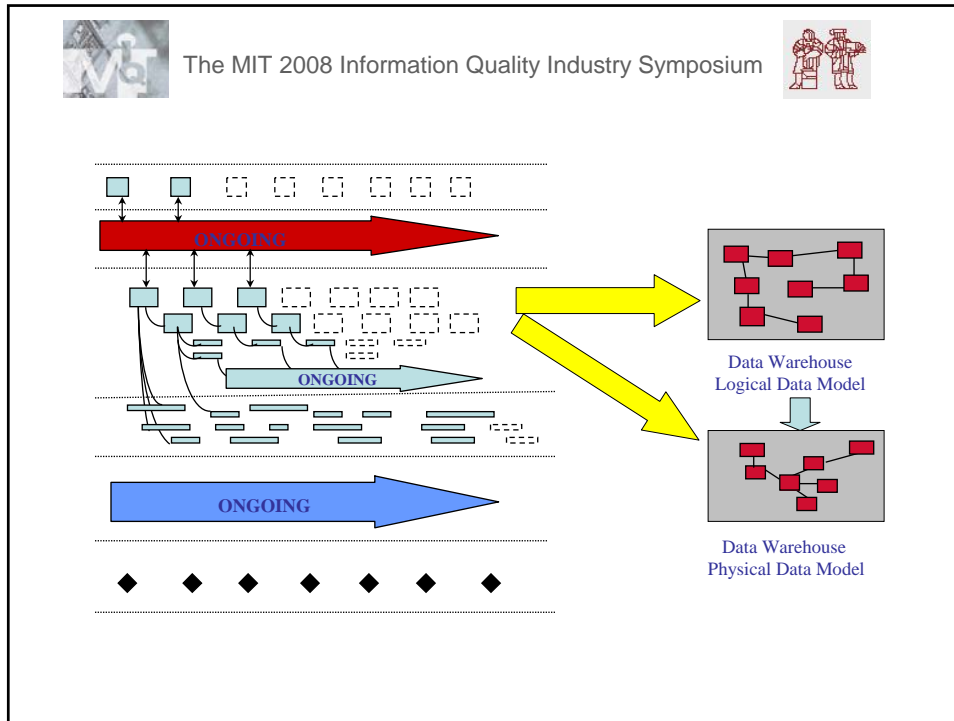














The MIT 2008 Information Quality Industry Symposium



Summary

- **Reusability is a critical success factor for second generation data warehousing and there is a much-needed focus on the quality of the Data Models which underpin the program.**
 - The models must accurately reflect the business and they must be reusable in all future releases of the program.
 - Sustained success of a DW/BI program requires a robust data architecture.
- **The foundational model is the Corporate Data Model.**
 - Traditionally, this model was derived using a top down approach and by utilizing Joint Requirements Planning and Joint Application Design techniques.
 - These techniques can deliver a good model relatively quickly.
 - The problem with models derived in this way is that they are based purely on business rules as perceived by management and senior analysts.
 - In reality, the systems that use the data may have a different set of rules. This is due to the fact that the systems are often 20 years old (and sometimes older). Undocumented changes have been made to the data and in the majority of cases the people that made the changes are no longer with the organization.
- **The only way to uncover what the data actually looks like is to reverse engineer the data into an abstracted logical data model.**
 - First generation data warehouse initiatives attempted this in the past but the tools available to help were limited.
 - Today a new set of tools has evolved – data profiling tools. These tools are an ideal aid to reverse engineer data and build a data model from the bottom up.
 - When a model is built in this way it is based on actual data content and the chance for errors and omissions in the data modeling process is reduced.
 - This “bottom-up” model is used as an input into the creation of the model that results from the “top-down” approach; in effect the former is used to challenge the latter model being drawn up by the business.



The MIT 2008 Information Quality Industry Symposium

