Textual ETL – Opening Up New Worlds of Opportunity

ABSTRACT -----------------------------

For years computing has revolved around repetitive activities such as bank transactions, airlines reservations, and manufacturing processes. Recently it has been recognized that textual data is not being included in the decision making processes. There have been attempts at taking text and reshaping it into a form suitable for analytic processing. But text has so many forms that a fundamentally different approach is needed. This presentation is about textual ETL, the process that takes text, integrates text and produces the text in a form compatible with the analytical processes that already exist in the corporation.

BIOGRAPHY-----------------------------

William H. Inmon  
Inmon Consulting Services

Bill Inmon, is recognized as the "father of the data warehouse" and co-creator of the "Corporate Information Factory." He has 35 years of experience in database technology management and data warehouse design. He is known globally for his seminars on developing data warehouses and has been a keynote speaker for every major computing association and many industry conferences, seminars, and tradeshows.

As an author, Bill has written about a variety of topics on the building, usage, and maintenance of the data warehouse and the Corporate Information Factory. He has written more than 650 articles, many of them have been published in major computer journals such as Datamation, ComputerWorld, and Byte Magazine. Bill is currently a columnist with Data Management Review, and has been since its inception. He has published 45 books; one sold over half a million copies, 21 have been book club selections with publishers such as Prentice-Hall, John Wiley, and QED.

Translations of various books have been done in Chinese, Dutch, French, German, Japanese, Korean, Portuguese, Russian, and Spanish.
TEXTUAL ETL – OPENING UP NEW WORLDS OF OPPORTUNITY

A presentation by W H Inmon

Disclaimer

The technology about to be described is highly patented. If you are interested in licensing the technology, please contact Forest Rim Technology.
The informal systems of the corporation:
- unstructured data
  - .doc files
  - .txt files
  - .xls files
  - email
  - transcribed telephone

The formal systems of a corporation:
- structured systems
  - structured data
  - corporate transactions
  - corporate reports
  - corporate databases
  - customer files
  - audit reports

There is a gulf between the two worlds:
- technology
- business practice
- organizational
- historical
by moving textual data to the structured environment, you can take advantage of the infrastructure for analysis that has already been built –

- DB2
- Business Objects
- Cognos
- Hyperion
- Crystal Reports, etc

there is a very good reason for moving textual data to the structured environment
It seems I always have to keep buying things. Then I have to train people to use them. When does it end?

I can save a lot by reusing my existing infrastructure.

another good reason for textual ETL

Please do not confuse textual ETL with search. Search technology assumes that text is correct as written. Integration assumes that text must be integrated before it can be used for analysis.
Textual ETL is a necessary complement to ECM.

Some of the issues of textual ETL:
- Terminology of data
- Simple unstructured/semi-structured data
the kinds of documents that must be accounted for -

- **simple unstructured**
  - large documents with lots of text
    - books, reports, patents, contracts

- **semi structured**
  - smaller documents
    - resumes, recipe books, tables, inspection reports

perhaps the most important aspect of the preparation for textual analytics is that of the need to address terminology

- cardiologist
- orthopedics
- general practitioner

they are all talking about the same thing, but they are speaking different languages
when it comes time to do analysis, accessing words by categories is as important as accessing words by their actual value.

there are many ways that categorization can be done
“...he drove his Porsche and...”
“... the Ford dealership...”
“...ran by the Volkswagen...”
“...the manager of the Honda plant...”

Spanish
“...he drove his Porsche German product sports car and...”
“... the Ford car dealership...”
“...ran by the Volkswagen car German product...”
“...the manager of the Honda car plant...”

Unstructured ETL –
- stop word processing
- stemming
- alternate spelling
- synonym concatenation
- homograph resolution
- spell checking
- words and phrases
semi structured ETL –
- mapping the internal structure of text by textual ETL
- variable pattern recognition
- variable symbol recognition
- multiple types of indexes
- utilities
- raw data hidden character display
- multiple path processing
- final index trimming

what happens when you just send raw text over to the structured environment?
you get the Tower of Babel
electronic text
- .pdf
- .doc
- .txt
- .xls
- .ppt
- comments fields
- and many more

structured data integrated into a data warehouse –
- SAP
- DB2/UDB
- NT SQL Server
- Oracle
- Teradata

and you can use standard analytical tools –
- Business Objects
- Cognos
- MicroStrategy
- Crystal Reports
- SAS
- and many more

the integration of taxonomies into the data warehouse environment is an important component of integration

© Copyright Inmon Consulting Services, 2008
so who are some of the people using textual integration?

organizations that are concerned with safety –
- airlines, chemical manufacturers, oil and gas distributors, etc.

and what are they looking at?
- accident reports, inspection reports, repair reports, warranty data, etc.

a second important application is in terms of contracts.
what happens when a corporation has thousands of contracts?

handling a few contracts is one thing; handling thousands of contracts is something else
there are important business decisions that can be made once the textual data is integrated into the structured, data warehouse environment.

visualizations require ETL processing as well.
two kinds of questions are answered -

visualization – how can I discover what I need to know about?

unstructured data base – once I know what is of interest, how can I investigate in great depth the things that are of interest