FORESTRIM
TECHNOLOGY

LEARN MORE AT:
forestrimtech.com/case-studies

WHITE PAPER

Combining Structured & Unstructured Data

by W H Inmon, Forest Rim Technology

Text has always presented a challenge to the IT professional. There are many reasons why text is challenging. The first and biggest challenge is the fact that text does not fit comfortably with standard database technology. Trying to place unstructured text into a highly structured data base is like trying to put the proverbial square peg into a round hole. Text inside a database is an uncomfortable fit in the best of circumstances.

Another reason why text is so challenging is that language is terrifically complex. The syntax and rules of language are numerous and convoluted. In order to make any sense of language, the organization has to deal not just with text but the context of text as well. And there are MANY other complications that arise when dealing with text.

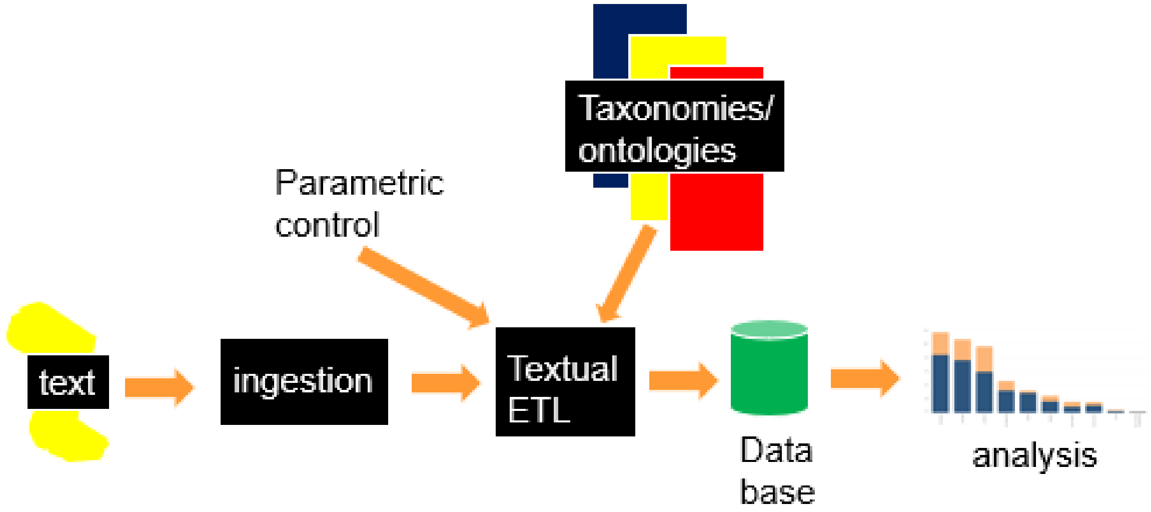
Because of these complicating factors, most organizations have ignored text as a source for information for many years. This is unfortunate because –

- Text represents the largest volume of data – by far - in the corporation.
- Some of the most important data in the corporation comes in the form of text.

Fortunately, there is new technology which mitigates the problems of reading and interpreting text in the corporation.

That technology is called TextualETL (or Textual Disambiguation)

With Textual ETL text can now be read and turned into a standard data base. This opens doors that have never before been opened –



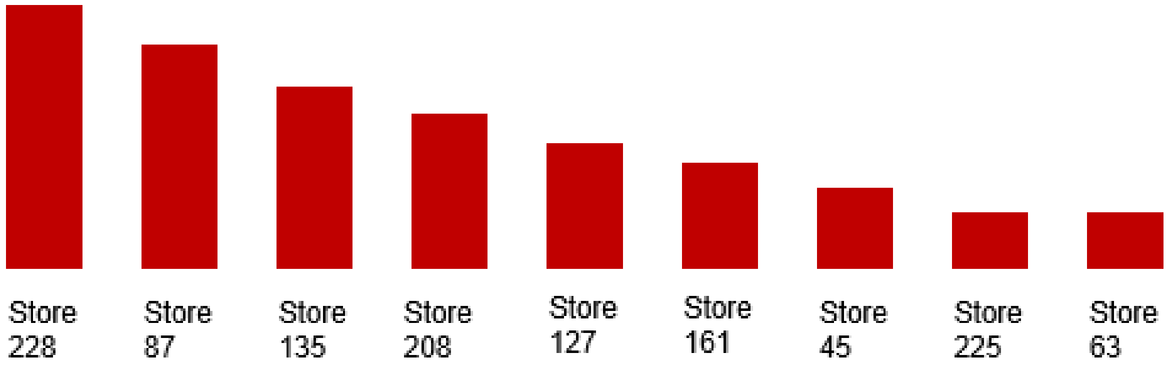
The database that text is restructured into is any standard database – SQL Server, Oracle, Teradata, DB2, Hadoop, and others. There is no bias for or against any standard database. Textual ETL works with them all.

By opening text up to standard database processing, the door is opened to many kinds of analysis. One of the most important types of analysis that is possible is the hearing of the voice of the customer. Organizations that listen to the voice of the customer prosper – simply stated.

And where is the voice of the customer found? For many organizations the voice of the customer is heard on the Internet. The customers of organizations leave comments on the Internet. Another place where the voice of the customer can be heard is the call center that many corporations have.

Wherever and however the organization can get their take on the voice of the customer, it is valuable.

As an example of hearing the voice of the customer, a restaurant chain took publicly available data from the Internet and fed the data into TextualETL. A database was created and a simple analysis was done –



The analysis was created by counting the complaints (negative sentiment) that came from the different stores in the restaurant chain. The diagram shows that store 228 had the most complaints, followed by store 87, then store 135, and so forth. Indeed, a raw count of text messages by store yields some interesting information. Looking at the data generated by examining text is quite useful. One conclusion that could be drawn is that the management of stores 228, 87, and 135 needs to be replaced or at least retrained. Something is going on at those stores.

The problem is that this information was gathered by solely looking at text. In order to get a more incisive perspective on the information, it is necessary to combine data from both the unstructured, text world with the classical structured world.

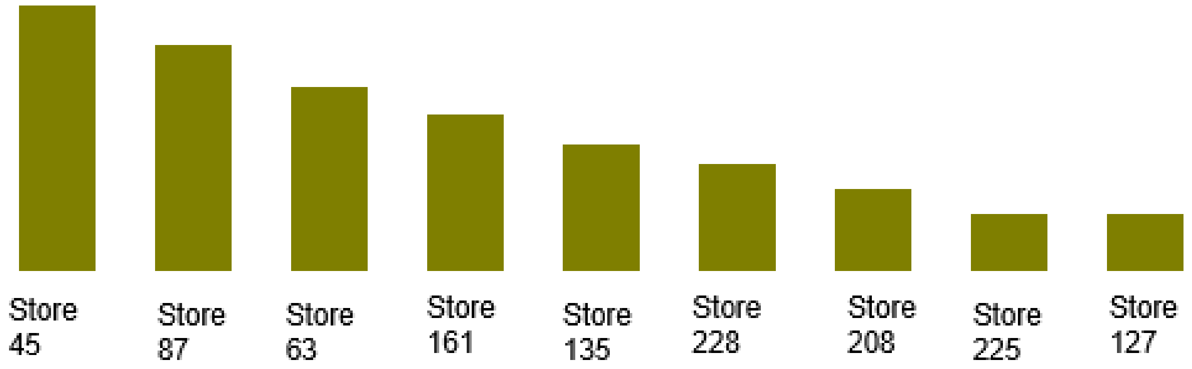
The reason why looking at the raw count of complaints on a store by store basis is misleading, is the fact that the volume of customers served by each store is very different. Store 228 is in downtown Manhattan, New York. Store 135 is in San Francisco. Store 45 is in Roswell, New Mexico. Store 127 is in Pueblo, Colorado. It stands to reason that the store in Manhattan is going to have more complaints than a store in Roswell, New Mexico, simply based on the difference in the number of customers that are served.

In order to account for the difference in store size, it is necessary to factor in the size of the store. One way to do this is to divide the number of complaints by the monthly store revenue. As an example of, Manhattan store 228 had a total number of 656 complaints and a monthly revenue of $1,260,473. In Roswell, store 45 had 125 complaints and revenue of $75,209 dollars.

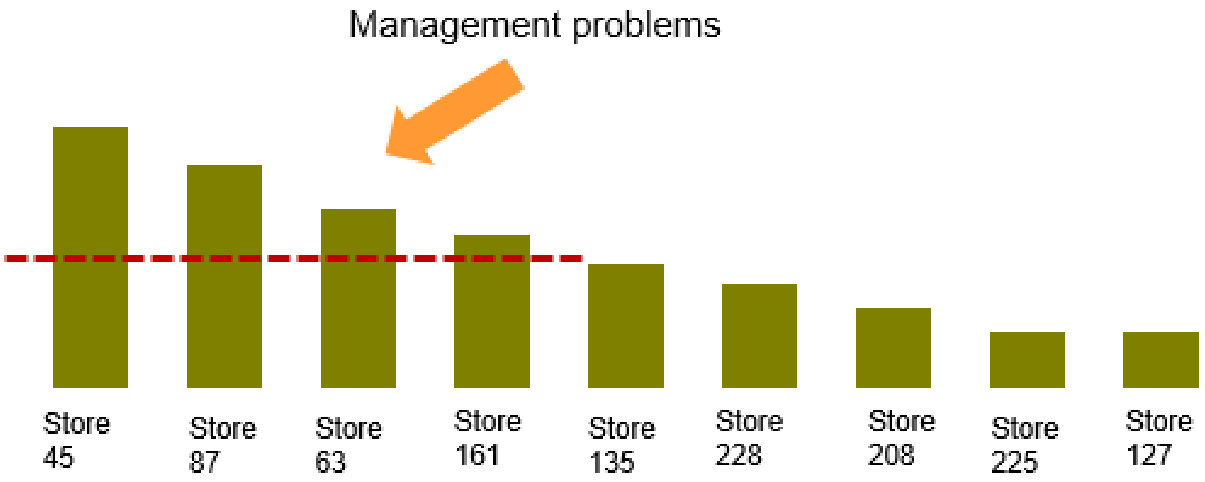
The ratio for Manhattan is .000520 and the ratio for Roswell New Mexico is .00166. When viewed this way, the management competency ratio of the stores in Manhattan and Roswell are seen completely differently. When the data is viewed this way, the management of the Roswell store--not the Manhattan store--is circumspect.

Once classical structured data is combined with textual data, the perspective is entirely different.

A very different measurement based on the ratios developed for each store looks like –



By combining both text and structured data, a much more enlightened comparison of stores and the competency of their management is achieved. Another way of looking at the data is –

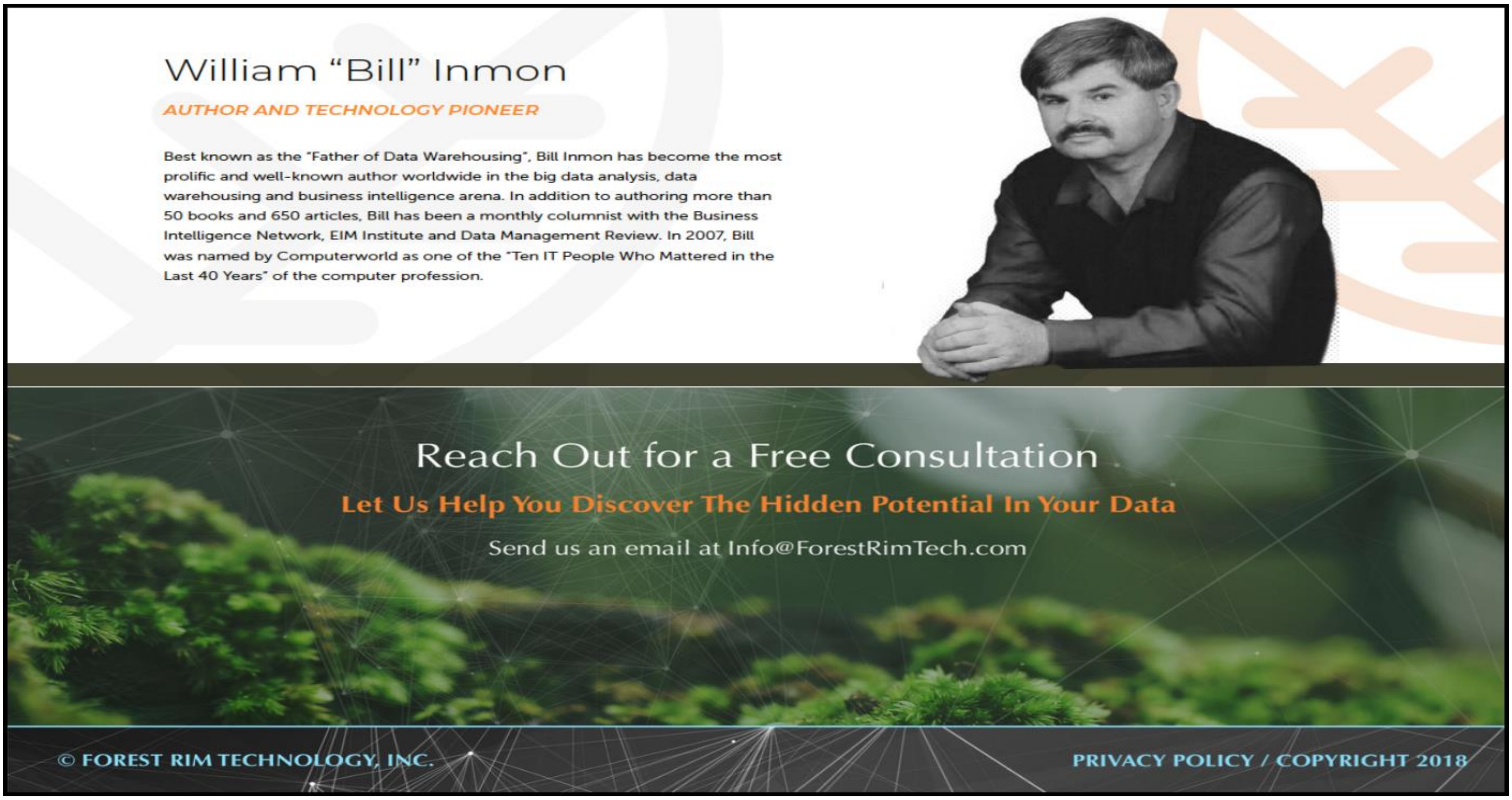


Using an analysis of the ratios, the management of the restaurant chain now knows which stores need top management attention.

As important as information derived from text can be, it can be even more insightful when combined with classical structured information.

It is noted that without the ability to read text and turn text into a database, the insights that have been discussed simply would not be possible.

Forest Rim Technology was formed by Bill Inmon in order to provide technology to bridge the gap between structured and unstructured data. Forest Rim Technology is located in Castle Rock, Colorado.



William “Bill” Inmon

AUTHOR AND TECHNOLOGY PIONEER

Best known as the "Father of Data Warehousing", Bill Inmon has become the most prolific and well-known author worldwide in the big data analysis, data warehousing and business intelligence arena. In addition to authoring more than 50 books and 650 articles, Bill has been a monthly columnist with the Business Intelligence Network, EIM Institute and Data Management Review. In 2007, Bill was named by Computerworld as one of the "Ten IT People Who Mattered in the Last 40 Years" of the computer profession.

Reach Out for a Free Consultation

Let Us Help You Discover The Hidden Potential In Your Data

Send us an email at Info@ForestRimTech.com