FOREST RIM TECHNOLOGY

ATSCALE

TEXT ANALYTICS SIMPLIFIED

Bill Inmon
Forest Rim Technology

Dave Mariani
CTO & Founder of AtScale

David Rapien
Partner at Forest Rim Technology

Ranjeet Srivastava
Chief Architect & Vice President of Coforge

...SO SIMPLE EVEN BILL INMON CAN DO IT...

Text Analytics Simplified

Bill Inmon
Dave Mariani
David Rapien
Ranjeet Srivastava

Published by self

Edited by external editor

Cover design by external designer

First Printing 2022

ISBN, print ed. 978-93-5680-061-8
ISBN, ePub ed. 978-93-5680-295-7
ISBN, PDF ed. 978-93-5680-504-0
ISBN, Kindle ed. 978-93-5680-643-6

Preface

AN INDUSTRY CRITIQUE

The path to text analytics that is presented in this book works and is practical. It shortens the cycle of text analytics enormously. In the most extreme example we know of, what took an organization ten years to do in old style text analytics was done in a day's time using the techniques that have been described.

So how is it even possible to condense ten years of work into a day? It sounds preposterous, but it really happened.

How is it even possible that a project that took ten years to be able to be done in a day? There are actually a lot of answers. Some of the factors for the shortening of the process of text analytics are:

1) The analysts that took ten years' time spent an inordinate amount of time to build a taxonomy/ontology. Instead, the analysts could have bought (for a very reasonable price) the same taxonomy commercially and greatly compacted the time that was consumed. An enormous saving here.
2) The focus was on text and parsing text. The focus should have been on finding context.

When you process text you MUST HAVE context, or else your understanding of text is compromised.

3) Once the taxonomy was created, the analysts didn't know what to do with it. If they had had Textual ETL they could have saved enormous amounts of time.

4) The time frame for old style text analysis was such that doing iterative processing was a real pain. But when the time frame is compacted enormously, iterative analysis is easy and natural to do. And iterative analysis is an essential part of text analytics.

5) Once the data had been created from the text, the analysts did not know what to do with it. It is a fact that once the text has been converted to data, that a whole other level of analytics ensues.

6) Ingesting text can take a lot of time and trial and error. This inevitably occurs the first time you do it. In many cases ingesting text is the most difficult part of text analytics.

7) There was no clear business value-based goal for the project. Many projects fail because there is no clear direction as to why the project is even being done.

8) The analysts got caught up in the analysis of text rather than the exploitation of text.

9) The focus was on exploring text rather than finding business value. There is a lot to explore, but much of the exploration of text has limited or no business value.
10) The analysts were sidetracked by the need to do inline contextualization.
11) The analysts wanted to do a perfect job with text. But when you deal with text you have to be prepared for some amount of imperfection because text itself is not perfect. In text analytics, perfection is the enemy of progress.

And there were other reasons why the first generation of text analytics was a failure.

The second generation of text analytics is here today. It is focused, fast, efficient, inexpensive and simple to do. It does not require a small army of highly paid consultants. And it can be done quickly and inexpensively.

Text is the backbone of language, yet organizations do little or nothing with their text. The problem is that there is a huge amount of important information buried in text. And that information never finds itself in the hands of the decision maker.

Some of the important information wrapped up in text include:

- **The voice of the customer** - What does the customer think about the products and services of a company?
- **Medical records** - Medical records are designed for a single doctor and a single patient. But when it comes to needing to look at 10,000 patients at a time, the medical record needs to be transformed.
- **Corporate contracts** - Corporate contracts are loaded with important information. But executives have no idea what is in their own corporate contracts.
- **Warranty claims** - Warranty claims hold a wealth of information for the manufacturer. But that information is seldom fed back to the corporation.
- And many more places.

The IT industry has a long record of either ignoring text or attempting to use text by having expensive and academic consultants conduct a long, expensive, laborious, academic exercise which usually does not result in anything terribly useful (but costs a lot of money and time).

It doesn't have to be this way.

Today – in the second generation of text analytics - there is a new and improved way to handle text –

- Simply
- Quickly
- Inexpensively

Today text can be reduced to a data base where it can then be analyzed. The process is fast, simple, and inexpensive.

This book describes in a step by step fashion how to do text analytics without going through a tedious, complex, expensive, academic exercise that requires expensive consultants.

This book is designed to be a description of the step by step process that you need to go through in order to start to do the second generation of text analytics.

This book is designed for both the technician and the business user. Both should be equal partners in the quest for turning text into actionable decisions. In addition, management may find the book useful if things start to go astray.

The authors have pared down the process of text analytics to the bare minimums that are needed. There is an introduction to all of the necessary steps and topics that you need in order to do textual analytics. In addition, the authors have covered the basic topics so that they are understandable to the average user.

We would also like to address who the intended users of this book and this technology are. The goal is for Business Professionals to be able to make better decisions. The purpose of Data Analysts, Data Engineers, Data Architects, and all of IT are to aide business in productivity and

decision making. It is NOT the other way around. The direct end users of this technology include all of these positions, but it should be able to be performed directly by the Business Professional to gain initial quick insights, even without the need of the other technical people. Technicians, you will use the output to delve deeper into insights to also aide business. Remember, that this is easy enough that even Bill Inmon can use it.

And that is saying something.

It is hoped that the mystery and the cost and overhead of doing text analytics has been removed. We should all be doing text analytics.

- **Bill Inmon**, Denver, Colorado, Forest Rim Technology 6/25/2022
- **Dave Mariani**, San Francisco, Atscale 6/25/2022
- **David Rapien**, Cincinnati, Ohio, Forest Rim Technology 6/25/2022
- **Ranjeet Srivastava**, Bangalore, India, Forest Rim Technology 6/25/2022

Acknowledgments

There have been quite a few people that have contributed to the work behind the scenes. The contributions have been both direct and indirect. But all contributions have been valued.

We wish to thank every one of them.

Carol Renne for being the chief cat herder and keeping us all in line.

Dr. Valerie Bartelt for keeping the train running when it runs off the tracks.

Tony Drake for his patience and counsel.

Patty Haines for friendship and wisdom and a lifetime of experience.

John Salazar for herding the Latin America cats.

Georgia Burleson (Ms. Taxonomy) for being the taxonomy goddess.

Sharon Mnich, for her talent and forebearance.

Ross Leher, (Mr. Taxonomy) for friendship and wisdom and advice on taxonomies.

Mary O'hara, for herding all the cats that Carol couldn't get into her cage.

Matt Zeringo, for making connections at the right time and right place.

We owe a debt of gratitude to all who contributed, directly and/or indirectly.

Thanks ever so much.

Contents

Preface _____ **i**

Acknowledgments _____ **vii**

Contents _____ **viii**

Some Preliminary Thoughts _____ **1**

Getting Started _____ **23**

Ingesting Text _____ **37**

Deidentification _____ **51**

Taxonomy Management _____ **59**

Other Mapping _____ **73**

Textual ETL _____ **81**

Phase I Database _____ **92**

Correlative Analytics _____ **100**

Milestones _____ **114**

The Semantic Layer _____ **119**

Data Future-proofing™ in line with the semantic layer ___ **131**

FDM™ _____ **154**

CHAPTER 1

Some Preliminary Thoughts

Text is the great enigma of high tech. Everyone knows that text is there. Everyone knows that important information is wrapped up in text. Everyone knows that they ought to be doing something with text. Text is the elephant in the room that everyone ignores.

But no one does anything about text. And intuitively they know they ought to be doing something about text. They just don't know what can and should be done.

Text just does not fit well with modern data base management systems. And text is inherently complex and difficult to deal with in any case, even if it fit well with data base management systems.

RULES

When we learn to speak and communicate, we learn a lot of rules. We learn sounds. We learn the alphabet. We learn words. We learn how to spell. We learn the meaning of words. We learn how to put sentences together. We learn

punctuation. In order to communicate we have to learn a lot of rules. And those rules are tucked away in our heads.

When we try to go to use the computer for the purposes of teaching the computer how to read and interpret language, we find that we have to teach the computer at least some of these rules. And the computer just is not designed to handle text well.

And – to make matters even more complicated - there is the issue of multiple meaning of words. The same word can mean a lot of very different things. What does the word "hot" mean? Is it referring to the way a jalapeno tastes? Is it referring to a summer's day? Is it referring to an attractive lady? Is it referring to a room's temperature? It is the same word, and it can have very different meanings and interpretations.

Simply stated, you cannot do text analytics just by looking at text. You HAVE TO account for context as well as text.

For these reasons and more, text presents some formidable obstacles to the person who wants to put text on the computer and to do analysis on the text.

Yet there is a world of opportunity awaiting the person coming to grips with these challenges.

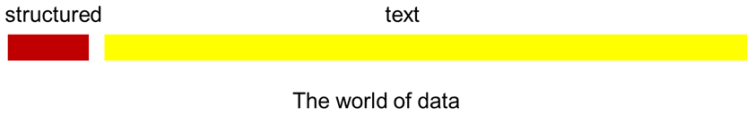
TRANSFORMATION

In order to use text on a computer, it is necessary to transform the text. The computer demands that the data that it processes be put into a very rigid, very structured format. And is it possible to actually put text into a rigid, highly structured format? Yes, it is.

This booklet is a guide to exactly what needs to be done to start to do textual analytics. The booklet is meant to be a paint-by-the-numbers document to be used as a means for doing textual analytics. The purpose of this book is to show how you can turn an enigma into an asset. An important and highly useful asset.

ANALYTICS TODAY

In order to use text, you need to be able to do analytics on it. So how is analytics done in the world of technology today? The starting point for our journey is – what data exists in the corporation today?



STRUCTURED DATA/UNSTRUCTURED DATA

The figure shows that there are two basic types of data – structured data and unstructured data. Structured data is data whose structure is repeated, over and over again. The content of structured data is different but the structure of data is identical. Structured data might be for the registration of a sale at Wal Mart. Structured data might be for the record of a phone call. Structured data might be for a deposit in the bank. Structured data might be for an airline reservation. There are indeed many repetitive activities that are captured by structured data.

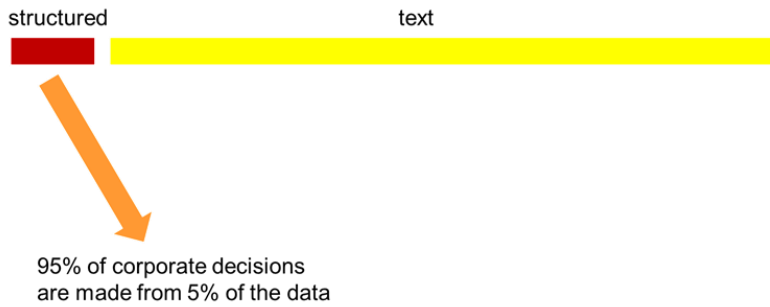
Unstructured data is data where the underlying structure of the data is not apparent. There are many different forms of unstructured data. One form of unstructured data is text, but there are many other forms of unstructured data. But – far and away – text holds the greatest amount of business value of unstructured data.

CORPORATE DECISIONS

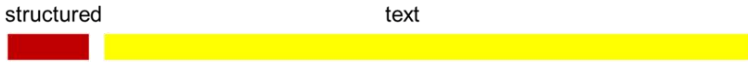
There is a fundamental problem with the way decisions are made in organizations.

In most corporations today, the vast majority of business decisions are made on a distinct minority of the data. Most business decisions are made on the basis of structured data.

But structured data is only a small part of the data that passes through the corporation. Does that make any sense? It is like standing beside a creek and declaring that you know all about water around the world because you understand the creek. You may well know a lot about the water in your creek. But there is a lot of water that you know nothing about. Rivers like the Mississippi. Oceans like the Pacific. Lakes like Lake Superior or the Great Salt Lake. Knowing all about a creek is merely scratching the surface of knowing all about the water of the world. And basing 95% of your corporate decisions on 5% of your data is just as foolish and short sighted.



There is something supremely short sighted in not making use of all the data you have in your corporation. All of your data.



95% of corporate decisions
are made from 5% of the data

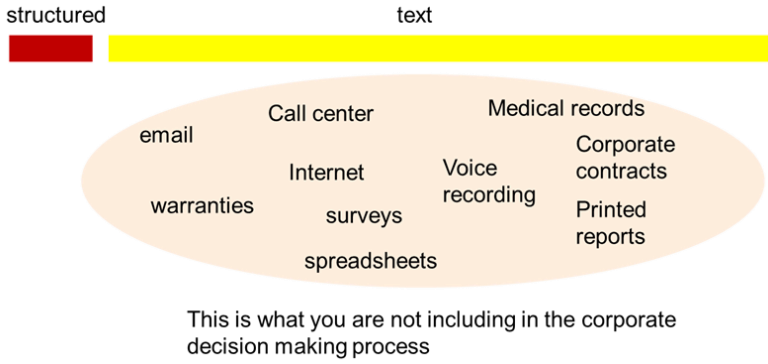
Something is really wrong
with this picture

WHAT'S MISSING

So, when you don't include text in the decision-making process, what are you missing? You are missing a LOT of important things. You are missing:

- Emails
- Internet conversations
- Corporate contracts
- Call center conversations
- Medical records
- Warranty claims
- Insurance claims
- And a whole lot more.

In short, there is a wealth of information that is tied up in the form of text. And that data is very relevant to the decisions that are made.



DATA MODELS AND TAXONOMIES

So, how do you go about getting a handle on all of this text? The classical starting point for trying to get a handle on a lot of disparate data is to build a data model. The problem is that classical data models do not work well (or at all) on textual data. There are actually a lot of reasons why data models are inappropriate for text.

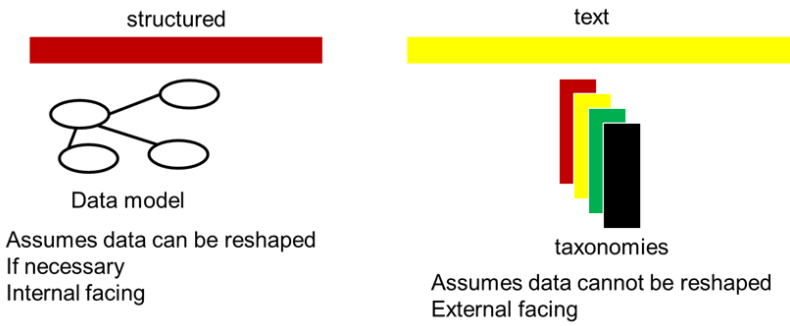
One of the basic reasons that data modelling does not fit well with text is that data modelling assumes that if something is wrong with the data, that the designer/analyst can simply go back and change the data. And with structured data, this assumption is usually true. But with

text you cannot make that assumption. It may even be illegal to go back and change text. Changing text is not an acceptable practice even if it may not be illegal.

Once text is written or spoken, it cannot be changed. It is what it is, right, wrong or indifferent.

Text can be parsed. Text can be classified. Text can be edited. But the raw source text cannot be changed.

A second reason why data models are not appropriate for text analytics is that data models are inward facing and text requires an external focus. Data models are good for looking at the data and systems internal to a company. But text analytics requires that the abstraction of text being outward facing, into the external world.



TAXONOMIES

Having stated that the data model is not appropriate for text, there still is a need for modelling text. In the place of a data model there is a taxonomy and/or an ontology. Taxonomies are similar to a data model in concept. But taxonomies and ontologies still have some essential differences from a data model.

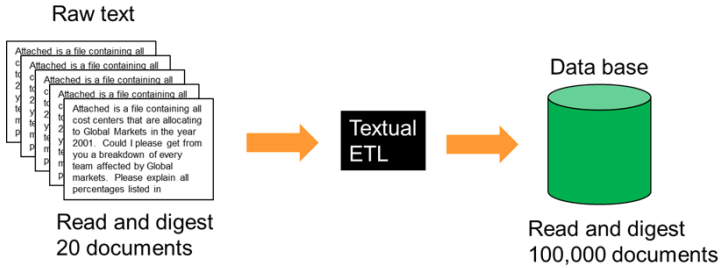
AUTOMATED ANALYSIS

So how is text analyzed? One approach is to analyze the text manually. This method has been practiced around the world for years. But there is a fundamental problem with trying to do the analysis of text manually. The problem is that a human must be trained to properly read a document if you want any consistency. Even then, there will not be consistency because of interpretation. Also, humans can only read so many documents. Depending on the document, a human may be able to read and digest 10 or 20 documents. In any case, it takes a long time for a human to read, much less interpret and analyze.

, a document. But what if you have thousands of documents to be read and analyzed? The human simply cannot read and digest thousands of documents.

In order to be able to read and digest lots of documents, automation – computation – is required. The computer can read and digest 10,000 documents (or more).

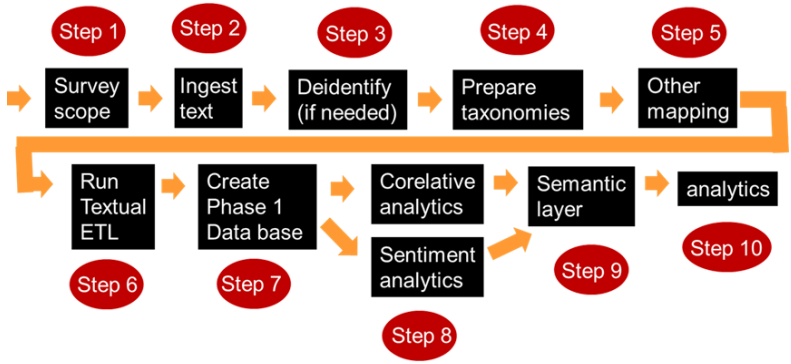
The problem with a computer reading and digesting thousands of documents is that that text must be transformed into a highly structured format. The computer requires this rigidity of format. And it is difficult to make that transformation. But that transformation must be done if the computer is to be used to manage text. And with today's technology that transformation can be made.



The major value of being able to process text automatically

DOING TEXT ANALYTICS – A PATH TO SUCCESS

So, what are the steps that are required to transform text into a format suitable for analytics? There is a well-travelled path to being able to read and analyze text. That path - the path to modern text analytics - looks like –



The first step is to survey and define the scope of analysis. The analysis may be exploratory or it may be very focused. But – for a variety of reasons – it is necessary to have a good idea of why you are going to do textual analytics before you ever begin.

After the scope has been established, the next step is to ingest the text. Text comes in many forms and flavors and it is necessary to transform the text from its originated form into electronic text. Once ingestion has occurred and the text is in an electronic format, the process of text analytics can then proceed to the next step.

Some text needs to be deidentified. Because of draconian government regulations there may be a severe penalty for not protecting text. In this case the text needs to be deidentified before further processing can occur. Note that not all text needs to be deidentified.

The next step in the path to text analytics is to prepare taxonomies that will be used for processing. There are several places to acquire taxonomies:

- From Forest Rim Technology – a limited amount of taxonomies
- From Wand Inc – an abundance of taxonomies
- Other vendors
- You can make your own taxonomies

There are several sources for taxonomies.

On occasion there are requirements for other mappings (other than taxonomies). These mappings are call "inline contextualization". These mappings look at the format of text, the placement of text, and so forth in order to understand and interpret the meaning of text. There are occasions where inline contextualization is not required. The need for other mappings than taxonomical resolution depends entirely on the data being analyzed.

After the text is readied for processing and after the mappings are done, it is time to run Textual ETL. As a rule, Textual ETL is the easiest, fastest part of the process of text analytics to execute. The data is identified to Textual ETL, the operating parameters are specified, and the processing commences.

The result of running Textual ETL is the creation of a Phase I data base. The Phase I data base represents that basic

transformation from text to data base. At this point the unstructured data has been transformed into structured data. The Phase I data base can be used for analytics by itself. Or the Phase I data can be transformed for other kinds of analytical processing.

The Phase I data base sets the stage for either correlative analytics or sentiment analysis.

The Phase I data can undergo more transformations. For example, sentiment analysis, sentence structure, punctuation, negation, and other features of text need to be extracted from the raw text. These might not have been originally extracted in the building of the Phase I data base.

After the Phase I data is extracted, the data can be sent to semantic layer analysis. Semantic layer analysis prepares the data for processing for further analytical processing.

Finally, the data can be visualized. Management demands that data be visualized before it becomes meaningful to them.

Each of these steps is straight forward. Each step will be fully described in this booklet.

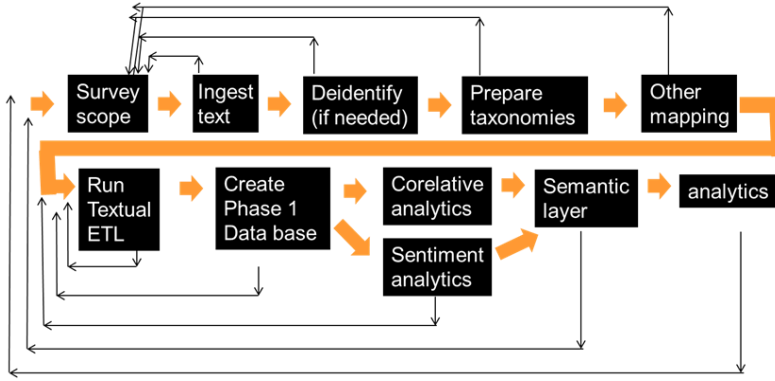
ITERATIVE PROCESSING

One of the features of the text analytics path that has been described is that it is iterative. It is almost guaranteed that the first time that a person travels the path of text analytics that has been described that some one or two details of data preparation will not be done properly. There is a misinterpretation. The taxonomy needs to be extended. Text needs to be edited differently. There are a hundred reasons why there needs to be an iterative return back to the beginning and another iteration of analysis performed.

Iterative processing is simply a way of life in doing textual analytics.

The good news is that there is no need to discard the analytical work that has been done when iterating the analysis. Instead, the analysis can be built upon the existing analysis to include the new features.

Another interesting feature of the iterative nature of analysis is that the process of reiteration can begin from anywhere in the path.



HOW LONG DOES IT TAKE?

One of the questions that everyone has is how long does it take to execute the path to text analytics that has been described? There are several factors that determine how long it takes to do textual analytics.

The first factor is how many times has the analyst executed text analytics. The first time through execution the analyst will need to become familiar with what needs to be done. There is a learning curve and it will take some time – some trial and error – to get through the learning curve. But after a few passes through the path, the learning curve subsides.

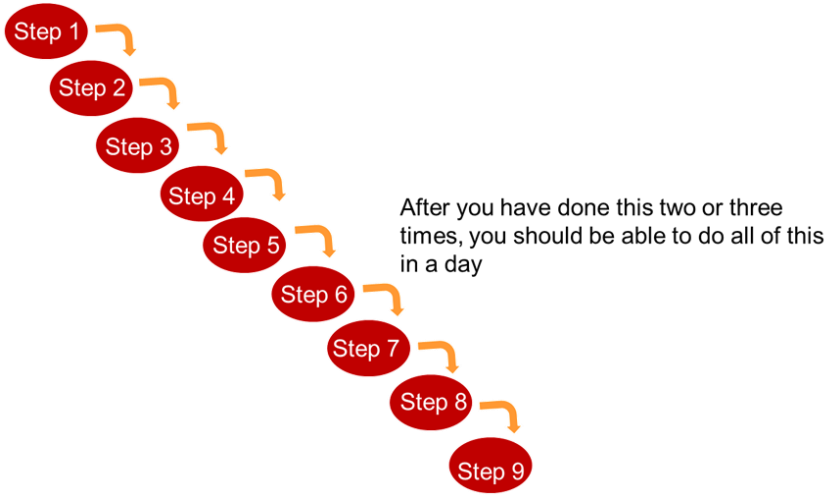
The second factor determining how long the execution of the path of text analytics takes is how much text must be processed. It is simply true that processing very large

amounts of text takes longer than processing moderate or small amounts of text. That is an inescapable fact.

The third factor determining how long it takes to execute the path is the simplicity or the complexity of the text. Some text is simple to process and some text requires a great deal of effort. It just depends on the text itself.

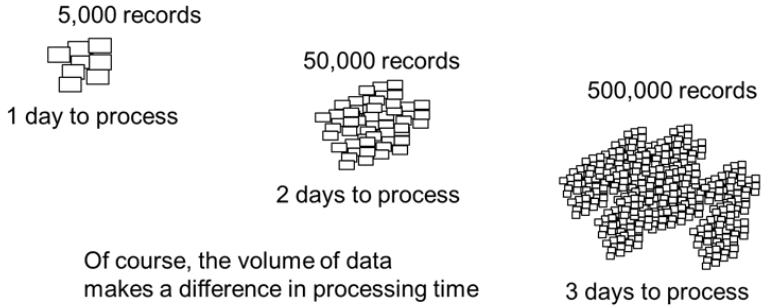
Having stated that, for an experienced analyst that has done text analytics before, for a small to moderate amount of data, the entire text analytics path should be able to be processed in a day's time. Or less than a day.

We do it all the time.

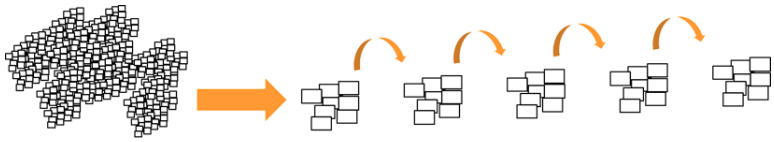


VOLUMES OF TEXT

Any amount of text can be processed in the path for text analytics that has been shown. There is no practical limit to the amount of data that can be processed.



But there is a very practical limitation that should not be ignored. That limitation is – the nature of textual analytics is iterative processing. It simply makes sense to process only a small amount of text at a time because the odds of that text having to be reprocessed in an iterative manner are very good. Stated differently, if you try to process a lot of text, you will find it easier and much more efficient to have to reprocess a smaller amount of text than a larger amount of text. So, a strategic decision on processing text is to break large amounts of text into smaller lots in order to be processed separately.



Because of the ability to do iterative processing it is highly recommended that large batches of text be broken into a series of smaller batches of text for processing

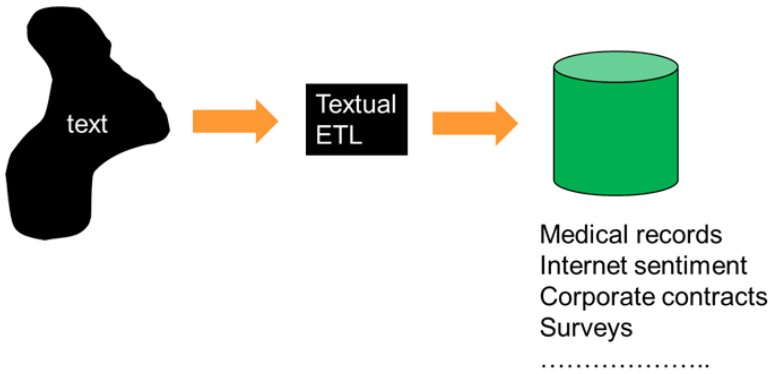
DIFFERENT KINDS OF TEXT

So, what kinds of text can be processed? The answer is – any kind of text can be processed and placed into a data base. Some of the many types of text that lend themselves to textual analytics include:

- **Medical records** - Medical records are designed for a single doctor and a single patient. But when it comes to research it is necessary to read and analyze 1,000 patient records at a time.
- **Internet sentiment** - The Internet is filled with sites where people voice their experiences and concerns about the products and the services of a corporation. Corporations find these comments invaluable in hearing the voice of their customer.
- **Corporate contracts** - Executives know that corporate contracts are important. But there are many contracts and many variations of contracts.
- **Surveys** - The comments section of surveys is where much of the valuable information lies, but

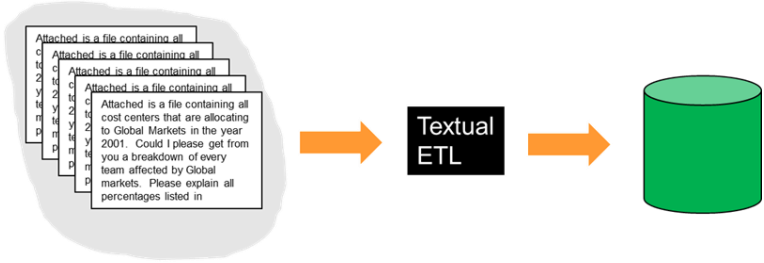
because the comments are in the form of text, they are often ignored.

- **Warranty claims** - Warranty claims can provide a wealth of information about the efficiency and the efficacy of the manufacturing process. But because warranty claims are in the form of text they are ignored.
- And there are many other types of text that contain really useful information.



TEXTUAL INTEGRITY

One of the issues about the processing of text is the integrity of the source text. In order to maintain integrity, the source text of textual analytics is NEVER altered. UNDER NO CIRCUMSTANCES IS THE SOURCE TEXT CHANGED. EVER.



The source text is never altered

TYPES OF ANALYSIS

There are two basic types of analytics that can be performed on the data bases of text that are produced. Those two forms of analytics are:

- Correlative analysis
- Sentiment analysis

In correlative analytics, the analysis centers around the occurrence of words and phrases in conjunction with each other. For example, in looking at 10,000 patients who have COVID, how many of those patients have also had:

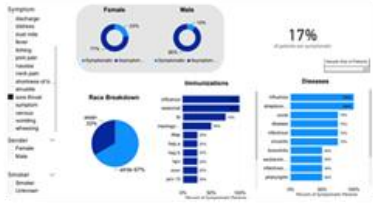
- Smoking
- Cancer
- Heart issues
- Diabetes

In addition, the 10,000 patient records can be correlated with:

- Medications
- Age
- Sex
- Race
- Weight
- And other relevant factors.

By finding related factors, researchers can find medical patterns that are otherwise hidden.

corelative



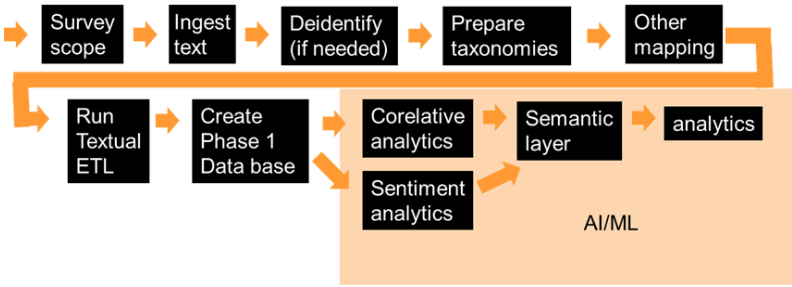
sentiment

What type of analytics do you want to do?

TEXT ANALYTICS AND AI

A final topic of interest is that of the relationship of textual analytics and AI and ML.

The relationship is shown as -



In many ways Textual ETL sets the stage for AI/ML.

CHAPTER 2

Getting Started

The starting point for text analytics is to define the scope of the analysis. The scope definition has many purposes:

- It directs the analyst to the taxonomies that will be used.
- It sets the stage for understanding and defining the context of the language to be processed.
- It limits the amount of data that needs to be processed.
- It defines what data should be examined.

Defining the scope of analysis is the first step in the execution of text analytics.

EXPLORATORY ANALYSIS AND BUSINESS FOCUSED ANALYSIS

There are two basic approaches to doing text analytics. One of those approaches is the exploratory approach. In the exploratory approach the approach attempts to find out what is there in the text that is worthy of further analysis. In the exploratory approach the idea is to survey all the things that might be of interest. The analyst does triage on what is in the document. An inventory is taken of what in the document needs to be to be analyzed and is worthy of further exploration.

In the exploratory mode of analysis, the categories of interest are wide. The exploratory approach is often referred to as the "heuristic" approach.

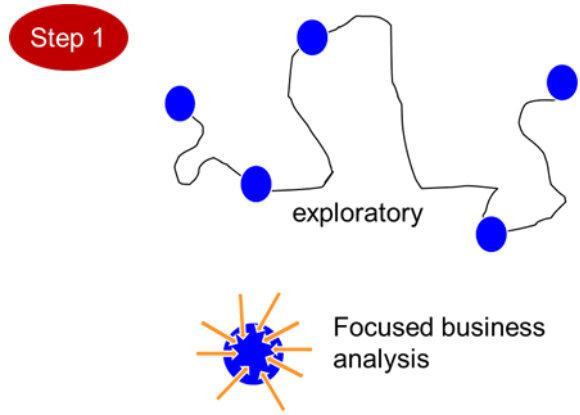
The other approach to analytics is to have one or two specific business-related questions that can be answered by textual analysis. In this approach, the analysis is focused on text that is relevant to the questions that need to be answered.

The categories of interest are usually much narrower than for the exploratory mode of analysis.

Both of these approaches are valid. And both of these approaches have great synergy with each other. For example, in the business focused approach it is normal to

discover business questions outside of the focus that needs to be explored. And in the exploratory approach it is normal to discover business questions that need to be explored.

So, neither of the two approaches are mutually exclusive. One approach can lead to the other and vice versa.



There are two basic types of analysis

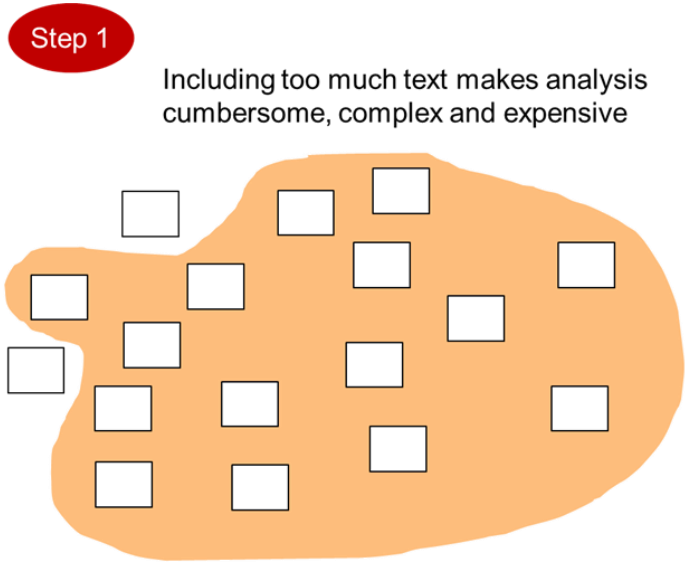
LIMITING THE TEXT TO BE SEARCHED

One of the many purposes of determining the scope of analysis is that of limiting the text that is to be studied. In other words, defining the scope of analysis sets the

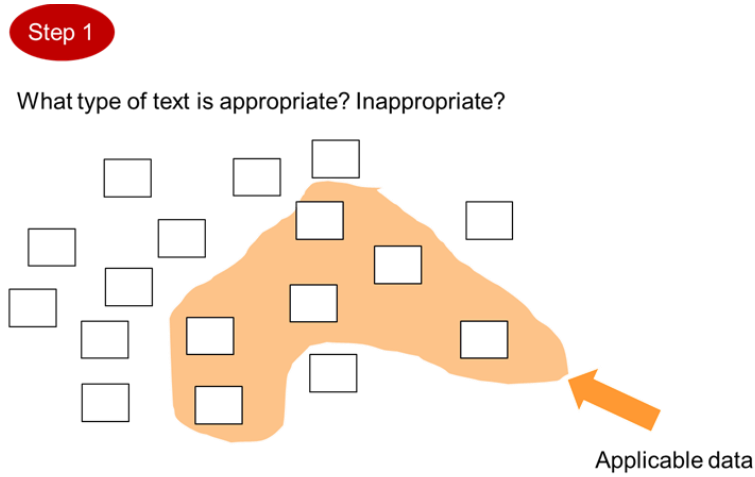
boundaries for determining what text needs further scrutiny and what text does not need scrutiny.

It takes time and money to do analysis on text. That means that the more focused you are, the faster you can expect to achieve results.

For example, suppose someone is interested in looking at customers attitudes towards a new product introduction – a new bicycle. Finding text on aeronautical engineering, religion, or the management of forests is not likely to have any impact on the analysis of the newly introduced product. Determining the scope of analysis then tells the analyst where to look productively for text that relates to the question at hand.



And by the same token, determining the scope of analysis tells the analyst what text is appropriate. If the analyst is looking for information about the introduction of a new product, it is probably productive to look at emails, the Internet, and other feedback from the customers or prospects. And when you look at these sources of text, you will have a clue as to what needs focus.



HAVING A SHARP FOCUS

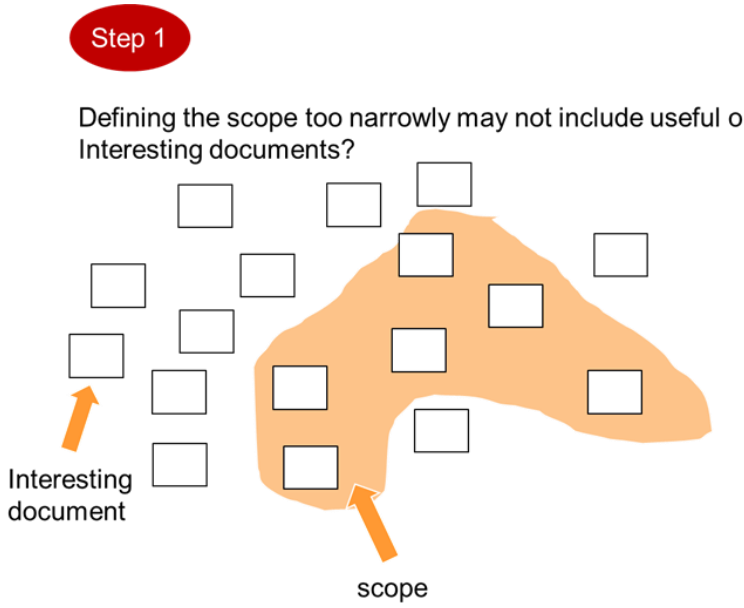
Having a sharply defined scope of analysis is usually not likely for exploratory analysis. But in the case of analysis to answer specific business questions, having a sharp focus can be a very good thing. Having a sharp focus on one or

two important business questions allows the organization to be very precise about which documents are of interest and need to be examined and which documents are not of interest. And limiting the documents that are of interest can greatly accelerate the progress of doing analysis.

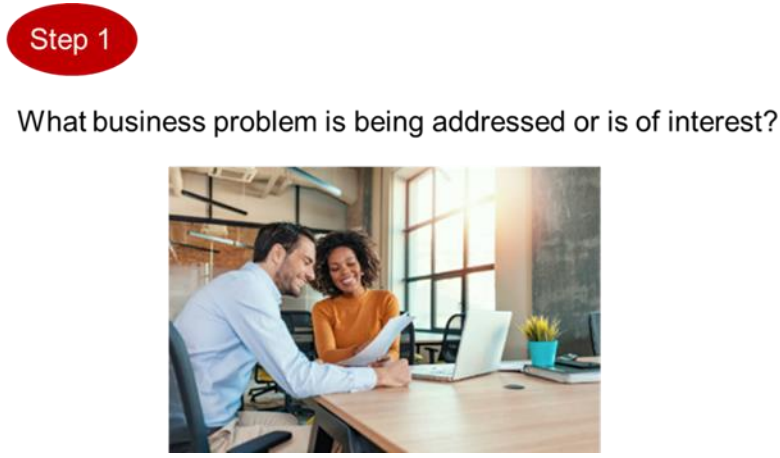
Having a well defined scope allows there to be focus

But whenever the scope of analysis is limited, there is always the danger that documents that are tangentially interesting will not be included. And often times the documents that are tangentially interesting contain very valuable pieces of information.

There is a tradeoff to be made in selecting documents and excluding documents in doing an analysis.



There is then real value in having focus when preparing to do an analysis. And the closer the focus is to business, the more likely that the end result will be of value.

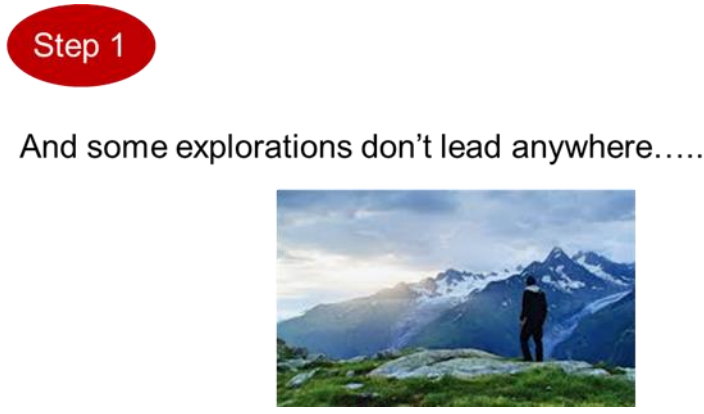


But there is an issue with focus. The issue with focus is that some analysis is of the exploratory variety. And having too narrow of a focus may cause the explorer to miss valuable information.

Step 1

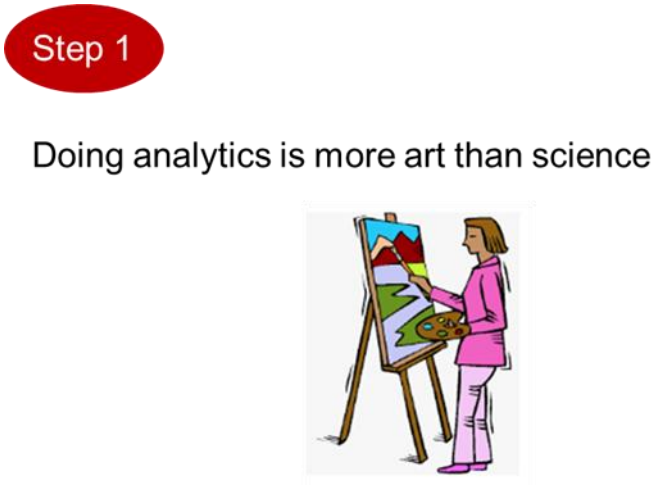
The problem with focus is that some analysis is just exploration…

One of the issues with exploration is that on occasion it leads nowhere. And on other occasions exploration produces truly profound results. The analyst must always be prepared to accept the fact that there is no answer to the questions that are being raised.



AN ART, NOT A SCIENCE

For all of the variables that are intertwined, doing analytical processing is as much an art as it is a science. The experienced analyst learns to trust instinct and intuition. And the only way to develop instinct and intuition is to practice the art of analysis. The more analyses that the analyst has done, the sharper the instincts and intuitions become.



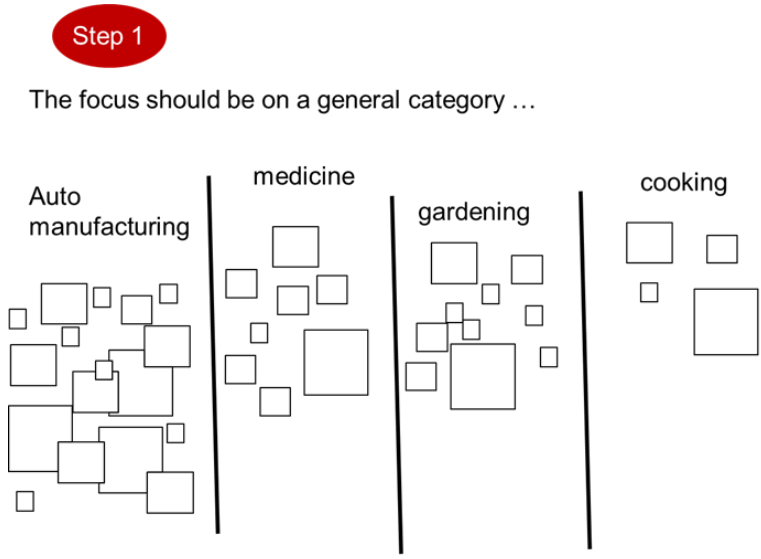
WHAT IS MEANT BY FOCUS?

When the subject of focus is discussed, it is instructive to examine what is meant by focus. As a rule, focus refers to large general categories of information.

Some large categories on which to center an analysis might include:

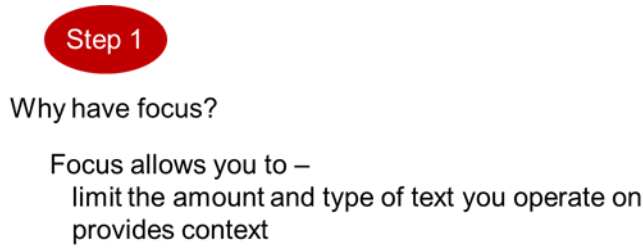
- Automobile manufacturing
- Medicine
- Gardening
- Cooking
- And so forth.

As a rule, there is very little if any overlap between the general categories of information.



CONTEXT

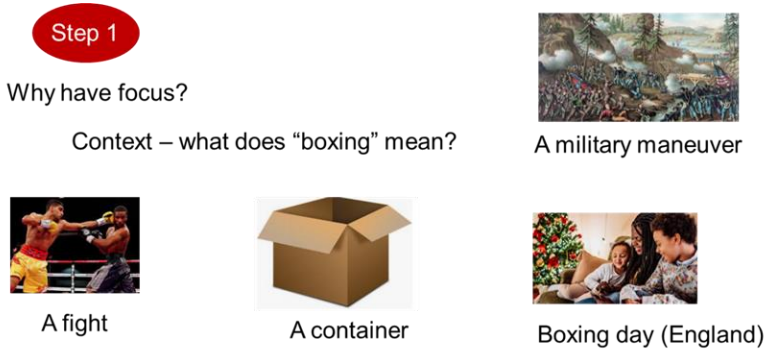
There are many benefits to having focus. The most obvious benefit is that focus allows you to limit the types and amount of text you want to process. But there is another significant benefit, and that benefit is that focus allows you to assign context to your data.



As an example of the value of context, suppose someone comes to you and says the word “seven”. Now what are they talking about? Seven days of the week? Snow White and the seven dwarfs? The seven seas? The mere word “seven” actually tells you nothing until the person tells you seven what. Only after context is provided does the number “seven” mean anything.

As another example of the value of context, consider the word “boxing”. What does boxing refer to? It may refer to a fistfight with gloves on. Or it may refer to a way of packaging some retail goods? Or it may refer to a military maneuver. Or it may refer to a holiday that is held in England the day after Christmas.

These interpretations of “boxing” are very different and refer to very different things. In order to understand what is meant, the analyst has to have context. Once the context is understood the meaning of “boxing” can be ascertained.



SURPRISE ENDINGS

One of the mysteries of doing analytical processing is that you never know what you are going to find. You can turn up lots of different unexpected things.

To highlight what is meant by surprise endings, contrast a banking transaction with an analysis of the search of likely neighborhoods in which to place a new retail store. When a bank teller goes to authorize the cashing of a check, the bank teller has a reasonable assurance that he/she is going to find out whether a customer has the funds to cash the check. But when a retailer looks for a new neighborhood in which to locate a new store, the analyst may or may not find such a neighborhood.

So, there may be no outcome for doing an analysis.

PROGRESSING TO THE NEXT STEP

So, how do you know when you are ready to progress from step 1 to step 2?

The types of questions that you need to answer in order to progress to the next step are shown -

Have sources of raw text been identified?
How much text is there?
Has business value been identified?
Has the general category of text been identified?
Have the boundaries of exploration been identified?
Have the success milestones been identified?
How accessible is the text?
Have security requirements been identified?

CHAPTER 3

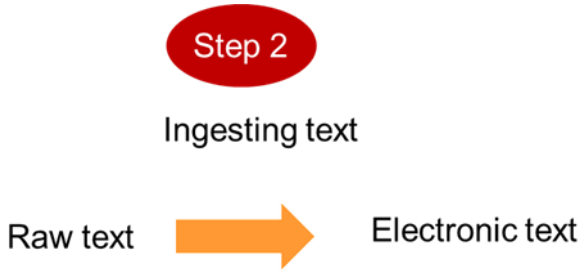
Ingesting Text

Text comes in all forms and all flavors. Text is found on the Internet. Text is in the newspapers. Text is in emails. Text is in voice records. Text is found in social media. There are many ways that text is expressed and created.

The next step in doing text analytics is to prepare text for processing by Textual ETL.

ELECTRONIC FORMAT

However, text is expressed, text needs to be in an electronic format before Textual ETL can process it. The next step in textual analytics then is in preparing the text for entry into an electronic format.

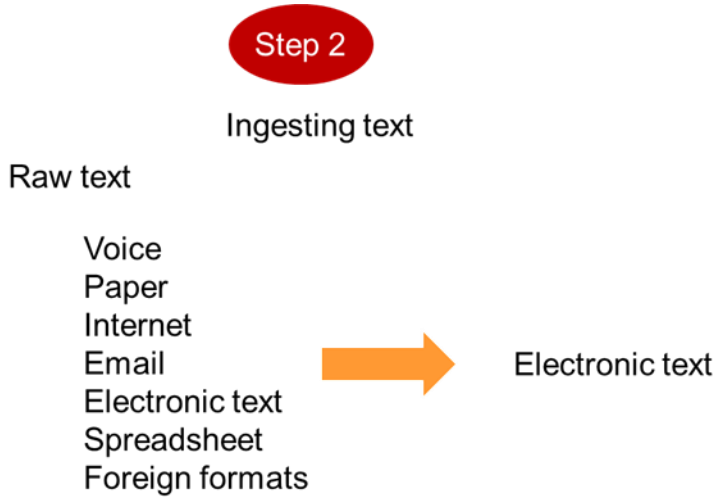


WHERE IS TEXT FOUND?

So, what are the standard ways that text is created and disseminated? The common ways that text is created and disseminated are:

- By voice recording
- On printed paper
- Over the Internet, in one of the many sites where people discuss their interaction with companies and their products
- Email
- Electronic text
- Spreadsheets

There are undoubtedly other places where text exists. But this list covers most of the common ways that text is created and disseminated.



Each of these formats has its own considerations.

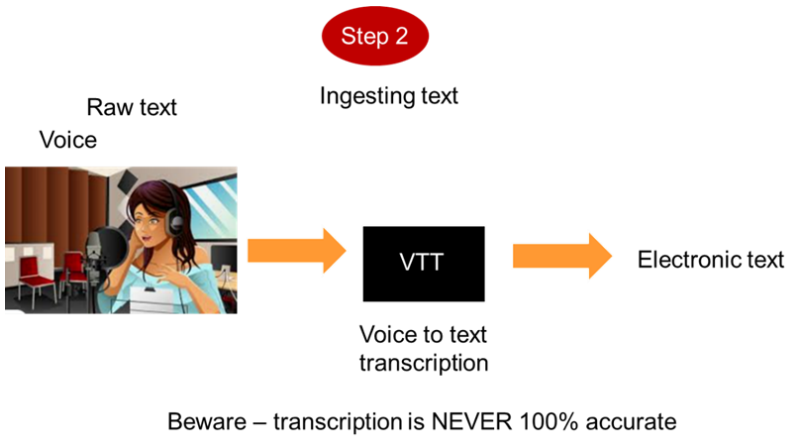
VOICE RECORDINGS

Voice recordings contain text. Typical of voice recording are hospitals and video recordings. It is normal to use VTT technology – voice-to-text transcription technology to convert recordings to electronic text. It is possible to transcribe voice manually. But manual transcription of voice is normally prohibitively expensive and time consuming. In addition, it is still possible for voice transcription done manually to contain errors. It is much more efficient and much less expensive to simply play the voice recording into VTT technology, and VTT technology converts the voice to electronic text.

One drawback of any voice transcription method is that the transcription to electronic text is never done perfectly. There are always misinterpretations that creep into the electronic text. As long as the error rate is not too high, VTT technology produces an acceptable and inexpensive result.

Some of the factors that affect the quality of the VTT transcription process are:

- The line quality the recording is made over
- The force with which people speak. Some people speak forcefully. Other people speak very softly.
- Accents - VTT can be confused by a heavy accent



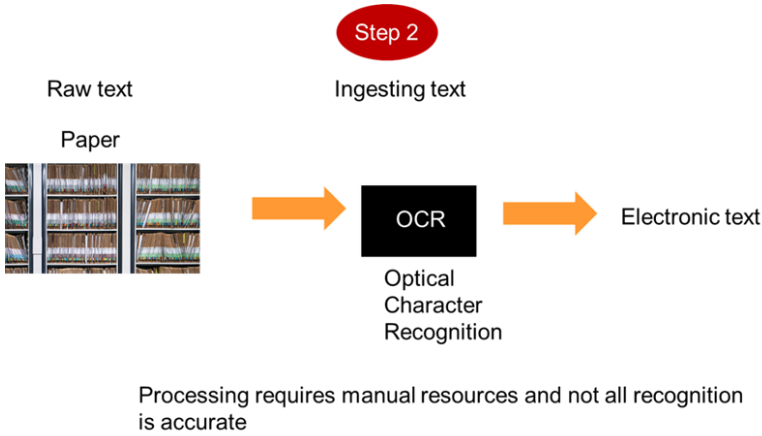
Beware – transcription is NEVER 100% accurate

PRINTED TEXT

Printed text can be ingested and placed in an electronic format as well. Much text ends up in a printed format. OCR – Optical Character Recognition – technology is used to transfer the text from the paper into an electronic format. The quality of OCR transcription is impacted by:

- The OCR manufacturer. Some OCR manufacturers are simply better than others.
- The age and quality of the paper being processed. Some paper disintegrates upon being processed
- The font the text is written in. OCR works better on standard fonts and not at all on other fonts
- The strength of the ink strike as the paper is being printed. At the end of a printing ribbon the ink strike often becomes very faint. When it becomes faint it is hard to read.

However, it is done, OCR conversion to electronic text involves a manual process. There is always a manual effort involved.



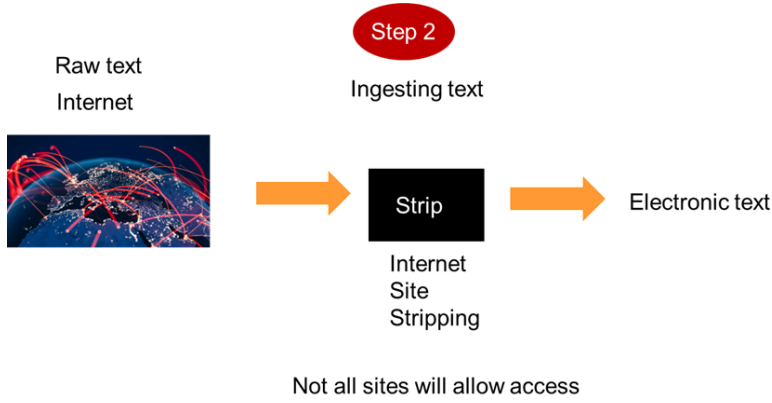
INTERNET AND SOCIAL MEDIA

A third viable source for the collection of text is the Internet and social media. There are many sites on the Internet where people talk about their experiences with a company, its products and its services. As such, the Internet is a wealth of information for a company that wishes to hear about what their customers are thinking.

Getting text off of the Internet is fairly straight forward. The text is already in the form of electronic text. But gaining access to the sites on the Internet is another matter altogether. Some sites are easy to access. Other sites go out of their way to prevent mass access.

In addition, the access of each site is usually different. So, accessing sites on the Internet is not necessarily an easy

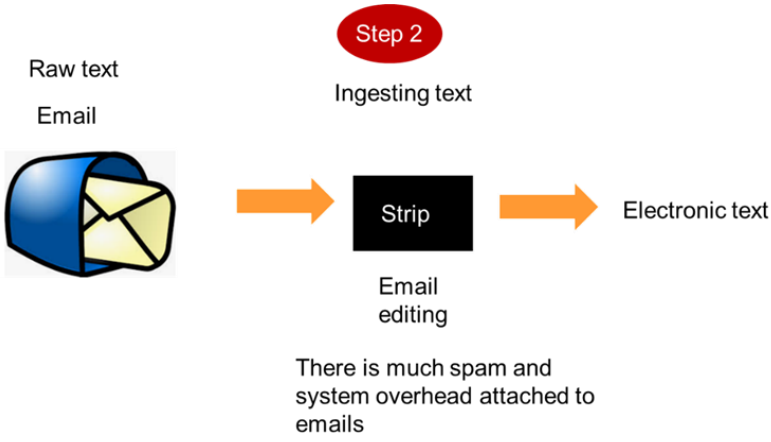
thing to do. Constant update to the program that does the site access is an ongoing need.



EMAIL

Another excellent source of text is email. Text in email is already in an electronic format. But email has its own limitations. The first limitation is that emails often contain spam. And spam is not what anybody needs to be analyzing. Spam needs to be removed from the email stream.

Another issue with emails is that email contain a very large amount of system overhead that needs to be removed. The excessive amount of system overhead burdens the system, clouds the results and wastes resources doing the processing.



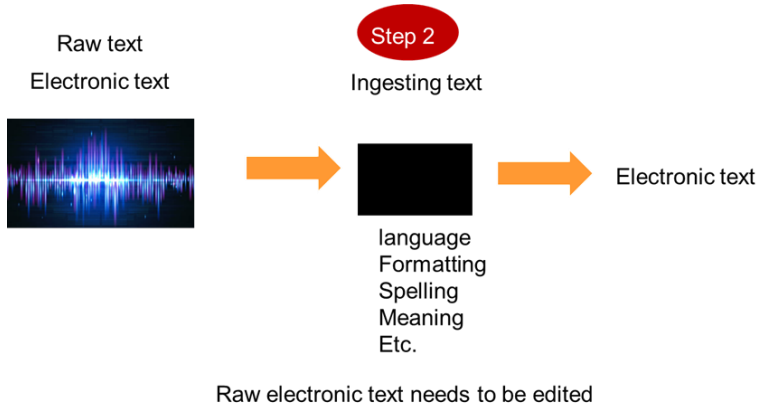
ELECTRONIC TEXT AS A SOURCE

In some cases, the text that is desired is already in the form of electronic text. In these cases, there is no need for the physical conversion of the data into an electronic format. But there are other considerations to the electronic text that is found. Some of those considerations are:

- The language the text is in – English, Spanish, French, etc.
- The formatting of the text – tables, acronyms, homographs, etc.
- Spelling – the English spell the same word differently than Americans – colour/color.
- Meaning what does "seven" mean? "Seven" by itself means nothing.

- Proximity analysis – the phrase – "Dallas Cowboys" written together is different in meaning from Dallas written on page 1 and cowboys written on page 13.

And these are just a few of the considerations of being able to make sense of the words that are found electronically.



SPREADSHEETS

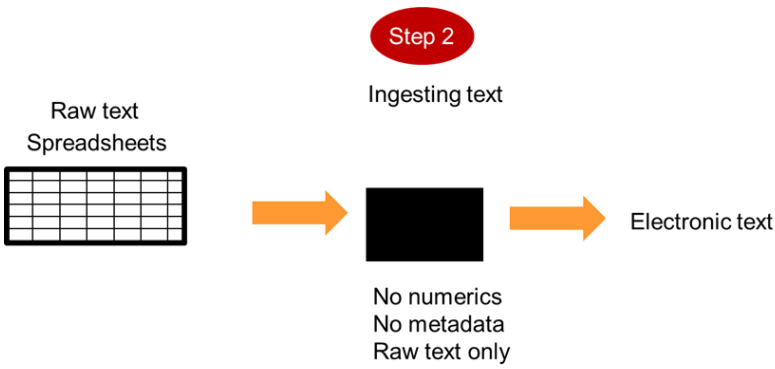
Spreadsheets are another common source of text. Spreadsheets are readily available to almost everyone. Spreadsheets are easy to use. And it is relatively easy to add or remove data – text and numeric data – from a spreadsheet.

But there are some rather severe downsides to using text off of spreadsheets. The first drawback is that no numeric data

should be lifted off of a spreadsheet. The problem with numeric data from a spreadsheet is that there is no good way of determining the context of the numeric data. And with no context, numeric data is useless. The second problem is that for text found on a spreadsheet, the analyst has got to be sure about the meaning and context of the text.

One of the problems with a spreadsheet is that anyone can put data onto a spreadsheet and anyone can change the data found on a spreadsheet. For this reason, data lifted from a spreadsheet is often questionable.

Despite all of the drawbacks, sometimes textual information resides on a spreadsheet that can and should be brought into Textual ETL.



There are great limitations to using spreadsheets

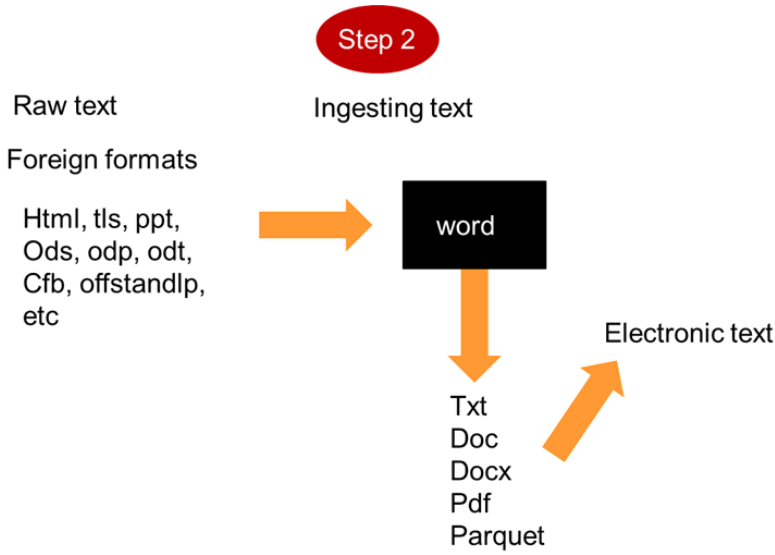
ELECTRONIC FORMATS

Another obstacle faced by the analyst wishing to bring text into Textual ETL is the obstacle of the actual format of the electronic data itself. Electronic data can exist in many formats. But in order for textual data to be read by Textual ETL, the textual data needs to be placed into one of the following basic formats for text:

- Txt
- Pdf
- Doc
- Docx

In fact, the txt format is the preferred format, but the other standard formats work as well.

Usually, it is very easy to convert text into a usable format. The normal way to do this is to read the text into Word in the foreign format and tell Word to write it out in the acceptable format.



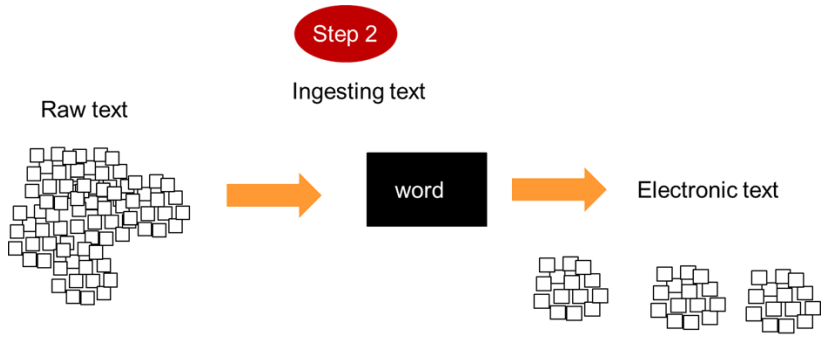
SIZING TEXT INTO SMALLER BATCHES

A final task of preparing text for Textual ETL is the task of sizing the data properly. If there is a large batch of text to be processed the text should be broken into smaller batches. There are a lot of good reasons for this separation of text:

- Small batches can be run quickly and iteratively. If a batch needs to be rerun (which is very likely) doing a rerun on a small batch is much more efficient than having to rerun a large batch.
- Small batches of text pass through the system very efficiently. Upon processing, combining the results

of processing from multiple small batches is an extremely easy thing to do with Textual ETL.

- Small batches of data are easier to work with than large batches of data.



Large collections of data are broken into a series of smaller collections of data

There is then a significant amount of work that needs to be done to text before it is ready for processing.

GOING TO STEP 3

How do you know you are ready to pass to step 3? The criteria for going to step 3 look like –

All of the sources of data have been converted to electronic text
All file types are readable
Large collections of data have been broken up into a series of smaller collections

CHAPTER 4

Deidentification

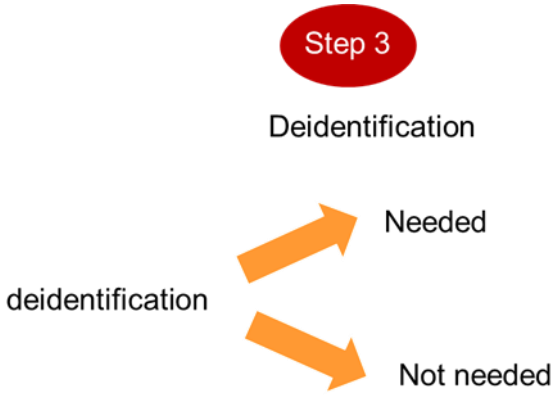
After raw text has been reduced to electronic text, the next step is to deidentify the text.

Deidentification

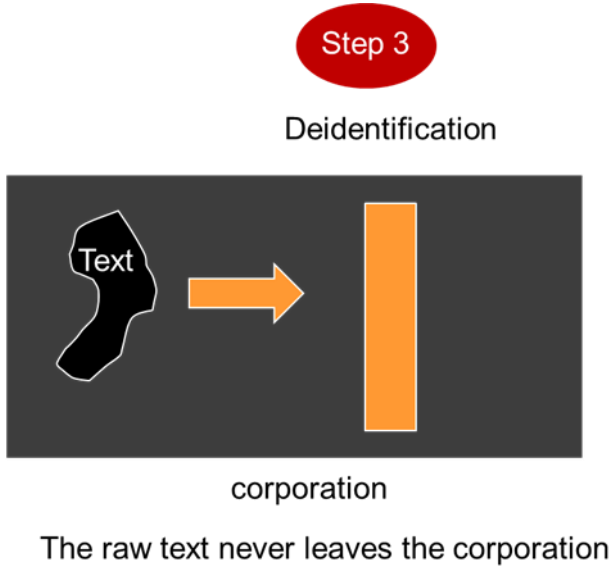
DEIDENTIFICATION IS OPTIONAL

The deidentification step only needs to be done if the text must be made secure. Typical organizations that need to deidentify their data include military, medical, and government. However, anyone can do deidentification if needed. But a lot of organizations simply do not need to deidentify their text.

If deidentification of data is not needed then this step can be omitted.



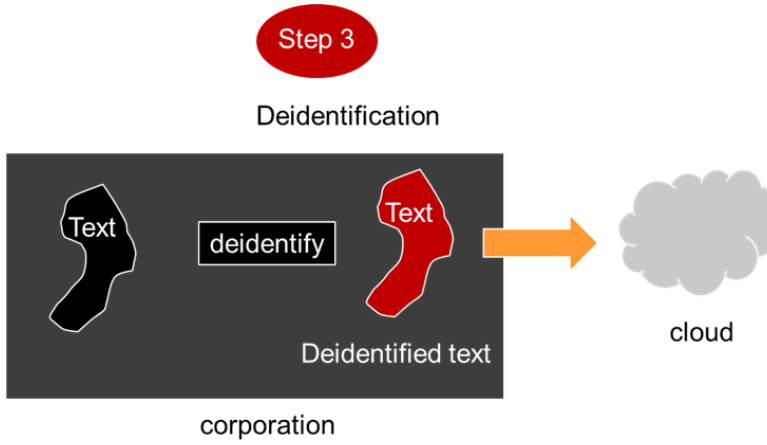
The point of deidentification of text is that no raw text ever leaves the corporate walls in a fashion where the subject of the text or the author of the text is apparent.



TWO APPROACHES TO DEIDENTIFICATION

There are two basic approaches to deidentification.

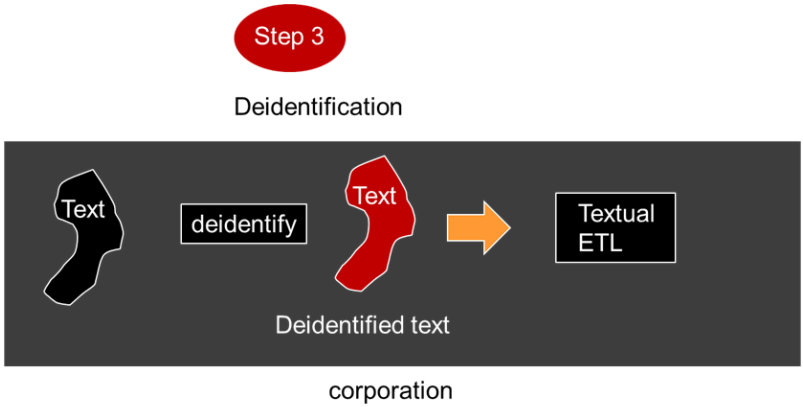
One approach to doing the deidentification is to do the deidentification inside the walls of the corporation, or within the corporation's intranet, and send the deidentified data to an external source. In many cases this external source is the cloud that deidentified data is sent to. Note that no raw data ever leaves the premises in this approach.



Another approach to deidentification is the approach where deidentified data never leaves the corporation. This approach is the ultimate in the securing of data. If raw data is leaked, it is leaked through other means than textual analytics.

There is one drawback to this approach and that drawback is – all of textual analytics must be done internally. The

corporation is completely responsible for the execution of textual analytics when all work is done internally



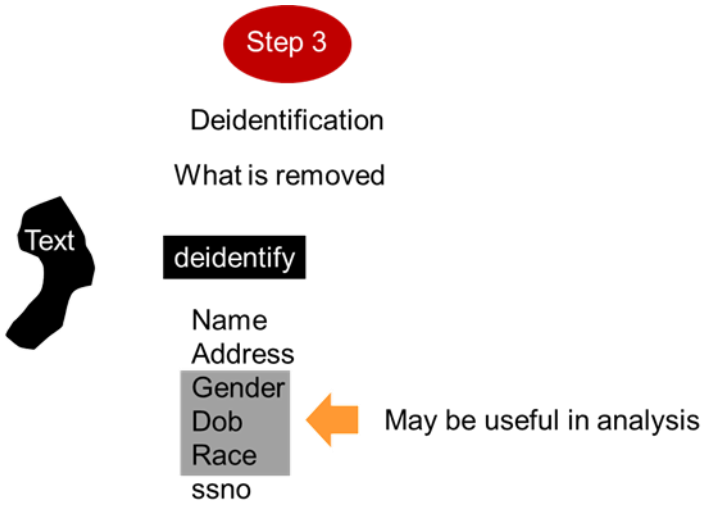
WHAT DATA IS REMOVED?

During deidentification, what data is removed? The answer is that all data that can be used for the identification of a person who wrote a document or is mentioned in the document is ether removed or blanked out. Typically, the types of data that are removed include:

- Name
- Address
- Social security number
- Age
- Gender

- Race

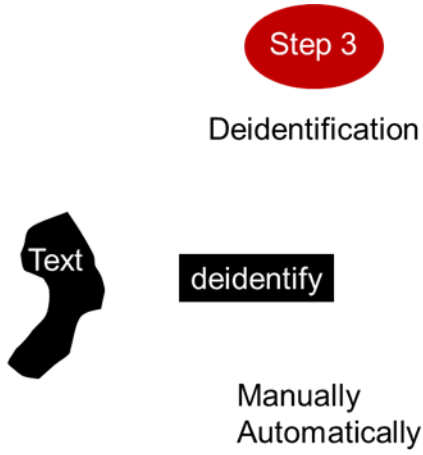
The text that is either removed entirely or replaced with xxx's is any data that could be used for the purpose of identification. However, care must be taken when selecting the data that is to be removed. Occasionally some data may be needed for analytic processing. Data such as age, date of birth and race may be needed for analytical processing.



There is then a tradeoff to be made. If data is needed for analytical purposes it should not be deidentified. But if data clearly identifies an individual, it needs to be removed.

MANUAL OR AUTOMATIC DEIDENTIFICATION

Deidentification can be done either manually or automatically. If you only have a small amount of data, you may be better off doing the deidentification manually. But if there is a large amount of data, it is best to do the deidentification using an automated tool.



GOING TO THE NEXT STEP

If data needs to be deidentified, this is the point that it needs to be done.

CHAPTER 5

Taxonomy Management

One of the essential ingredients for textual analytics is that of taxonomies.

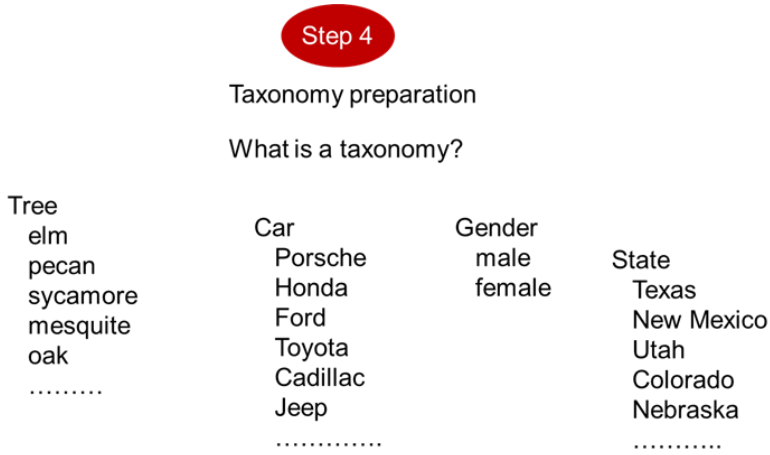
Taxonomies are like the light house for ships in the ocean. Taxonomies provide direction when direction is needed.

There is an old saying – "when a ship has no destination, any setting of the rudder will do." Taxonomies allow the rudder of the ship to be set properly.

Taxonomy preparation

WHAT IS A TAXONOMY?

In its simplest form, a taxonomy is just a classification of objects. The objects can be anything – trees, cars, houses, grasses, animals, birds, etc. Some sample taxonomies might look like –

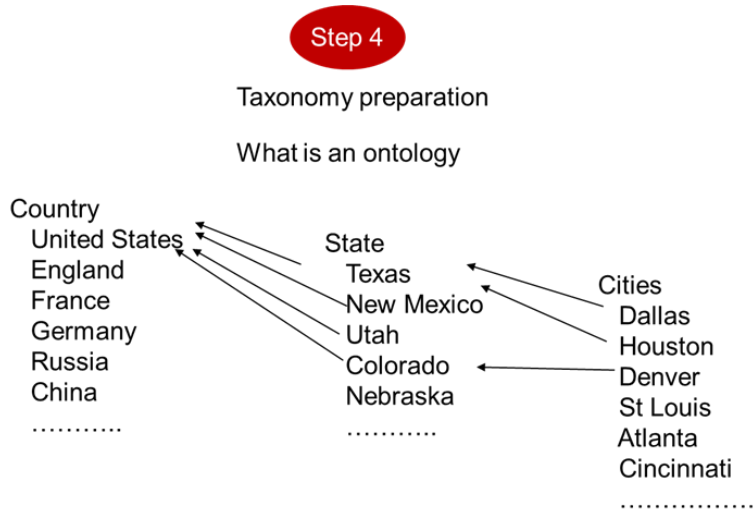


ONTOLOGIES

A related form of a taxonomy is an ontology. An ontology is nothing more than a related collection of taxonomies. For example, there might be a taxonomy of countries. Then there might be another taxonomy of states, where a state is shown being related to a country. And there might be another taxonomy of cities, where the cities are shown to be related to a state. Together the related taxonomies form an ontology.

As another form of an ontology, consider medicine. In medicine there might be medications, treatments, procedures, symptoms and the like. There might be medications, such as furosemide, januvia, and allopurinol.

The individual medications would be related to the higher classification of medications. There might be procedures like appendectomy, amniocentesis, and thoracic surgery.



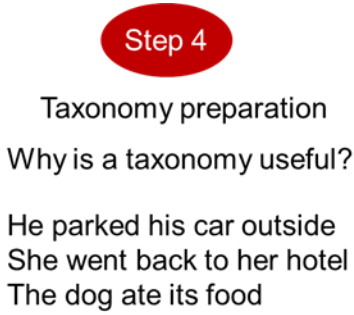
LANGUAGE AND TAXONOMIES

So why are taxonomies and classifications useful in doing textual analytics? The answer is that in using language, people do classifications all the time, and they don't even realize that they are doing classification. When you speak or write, classifications are done subconsciously, without even knowing you are doing it. Doing classifications is just a normal part of communication.

For example, take the simple sentence – "he parked his car outside." This sentence is filled with classifications. The word "he" more specifically could have been Jim Smith. The word car could have been "Porsche". The world outside could have been "on Eudora Street". So, the sentence – "he parked his car outside" could have been, in an unabstracted, unclassified format "Jim Smith parked his Porsche on Eudora street."

People – when communicating – abstract words without even knowing that they are doing abstractions. It is just a standard part of language.

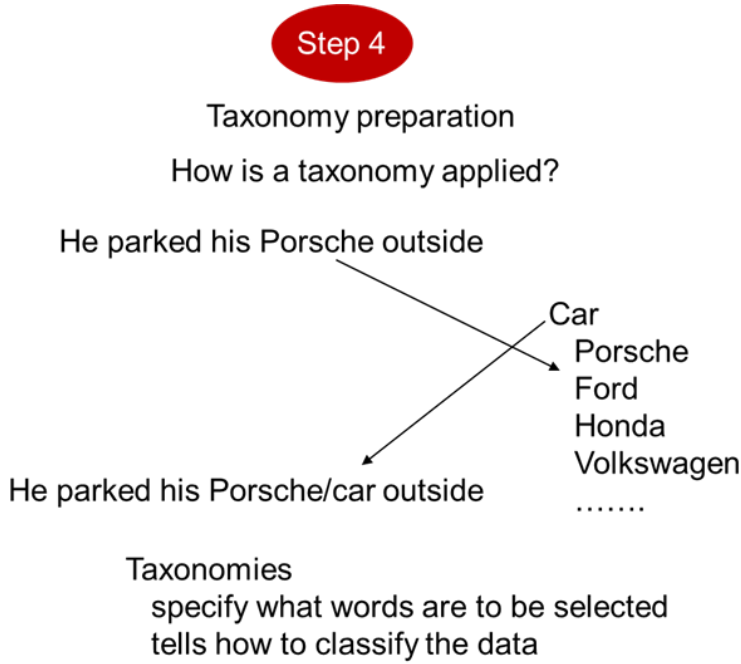
So, it makes sense that taxonomies become an important part of understanding communications.



TEXTUAL ETL AND TAXONOMIES

So how does Textual ETL use taxonomies in converting unstructured text to a structured format? Textual ETL does what is called taxonomical resolution. Textual ETL reads the raw text, finds the words that have taxonomies related to them, and then selects the word to go into the data base and supplies the hierarchical resolution to the data base.

The steps that Textual ETL goes through as shown by the following diagram -

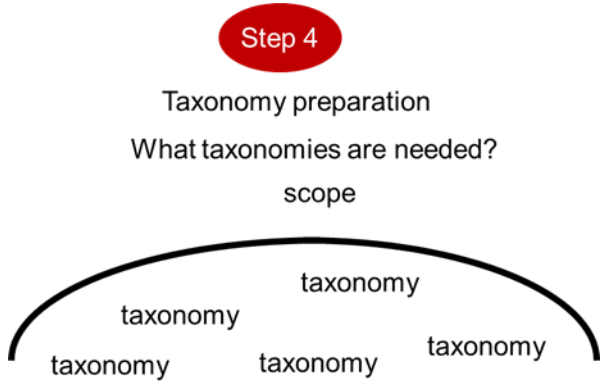


In the diagram the sentence is read. Textual ETL finds that the word Porsche is part of a taxonomy. Textual ETL then

notifies the system that a word of interest has been found in the taxonomy and what the classification of the word is. The word, its classification and other information are then written to the data base for further textual analytical processing.

TAXONOMIES AND THE SCOPE OF ANALYSIS

So, what taxonomies does the analyst need to supply to Textual ETL in order to process a document? The answer is that the analyst needs to supply the taxonomies that explain – encompass - the scope of analytical processing that has been described in chapter 1.



As a simple example of taxonomies encompassing the scope of the text, suppose the text that had been selected was automotive engineering. The taxonomies that might be found useful for automotive engineering might include:

- Bill of material
- Process control
- Strength of materials
- Suppliers of parts
- And so forth.

ACQUIRING THE TAXONOMY

Taxonomies are required for text analytics. Yet most organizations do not have people that are familiar with building and managing a taxonomy. So how does an analyst go about acquiring the taxonomies that are needed?

There are basically two sources for taxonomies: there are commercial taxonomies and there are manually built taxonomies. In almost every case, it is desirable to get a commercially built taxonomy. There is little value in creating something that has already been created. Acquiring a commercially built taxonomy is:

- Inexpensive
- Fast to acquire
- Readily available

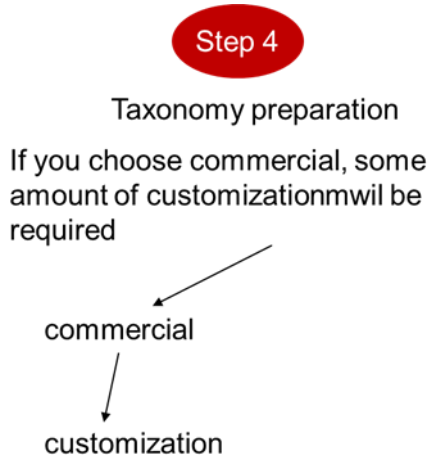
The only circumstance where a commercially available taxonomy is not preferred is the case where there simply are no taxonomies anywhere that have ever been built.

CUSTOMIZING THE TAXONOMY

One consideration of acquiring a commercially built taxonomy is that in almost every case, once acquired, the commercial taxonomy needs to be customized.

For example, suppose you acquired a commercial taxonomy for restaurant chain, if you were McDonald's you would have to add Big Mac to the taxonomy. But much of the other things that McDonald's has would already be found in the taxonomy – kitchens, stoves, tables, bathrooms, parking lots, etc.

However, it is always easier to customize a taxonomy than it is to build it from scratch.



TWO TYPES OF TAXONOMIES

There are two types of taxonomies – generic taxonomies and specific taxonomies. A generic taxonomy is one that contains words that can be used generically. For example, positive statements are generic. A positive taxonomy can contain words such as:

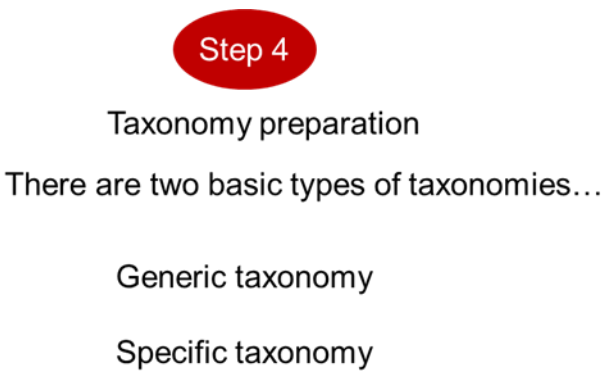
- Like
- Love
- Adore
- Cherish
- Want
- and so forth.

The words are generic because they can refer to anything.

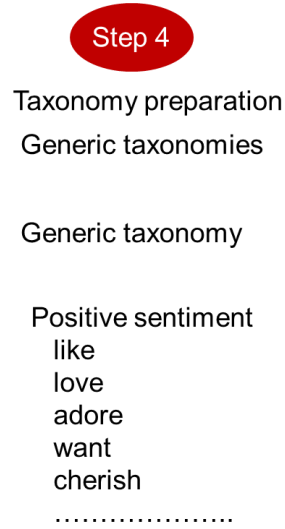
- I like enchiladas.
- I like golf
- I like freedom
- I like Mexico

The other type of taxonomy is specific. The specific taxonomy contains words and phrases that are specific to one discipline. For example, a specific taxonomy for automobile manufacturing may have the words:

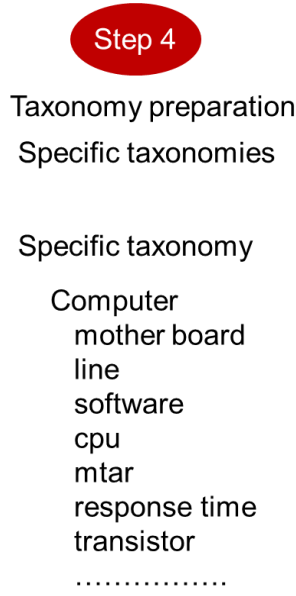
- Cam shaft
- Rotor
- Carburetor
- Brake
- Steering wheel
- Transmission
- and so forth.



As an example of a generic taxonomy, there is -



As an example of a specific taxonomy there is -



MERGING TAXONOMIES

One of the interesting questions is – can taxonomies be merged? The answer is absolutely yes – taxonomies – specific and generic can be merged. Furthermore, it almost always makes sense to merge taxonomies to form a complete picture of everything that is in the scope of analysis.

PURPOSING TAXONOMIES

Taxonomies have purpose as well as classification. Generic taxonomies with positive sentiment words and classifications need to be purposed for Sentiment Analysis, and further purposed for positive or negative sentiment analysis. Some taxonomies are purposed as important words needed as objects of that sentiment analysis. Some words trigger an opposite meaning. These words negate the original meaning of the sentiment. If we say:

- I don't like ice cream
- The pizza was not hot
- The was not happy with how server spoke to my children

We need to consider a different meaning or sentiment than the words "like", "hot", and "happy" mean. These negating words need to be classified as Reversal or Negation words.

This merging of taxonomies would normally make an ontology, but with Textual ETL's ability to further classify purpose and contextualization along with the merging of taxonomies, a new method is required. In Textual ETL, we call this contextualized merging of ontologies a Nexus. In truth, a Nexus also includes contracts, Inline contextualization, dentification protocols and many other things. But to begin, you must join and purpose your taxonomies to properly do text analytics.

UPDATING TAXONOMIES

One of the aspects of taxonomies is that they must be updated over time. Over time, language and expressions change. And as these changes to language occur, it is important that the taxonomy be made current.

The good news is that change to language does not occur often. So, updating taxonomies is not an everyday event.

So, what happens to text that has already been processed when a taxonomy has to be updated? The answer is – if you want previously processed text to be made current with the most up to date version of the taxonomy, you have to go back and rerun the older text with the current taxonomy. However, in most cases it is sufficient to merely run all future text with the current taxonomy. It is only on rare

occasions that you need to go back rerun older text against a new version of a taxonomy.

GETTING TO THE NEXT STEP

Have all the taxonomies been identified and made available that are needed to encompass the scope of analysis?
Has customization been done on the commercially acquired taxonomies?
Are the taxonomies up to date?
Are the taxonomies formatted to fit into textual ETL?

CHAPTER 6

Other Mapping

The standard mapping for text analytics is taxonomical resolution. In truth taxonomical resolution encompasses nearly everything that the text analyst needs in order to do text analytics. However, there are occasions where a different kind of mapping is necessary in order for the system to be able to process text.

Other mapping

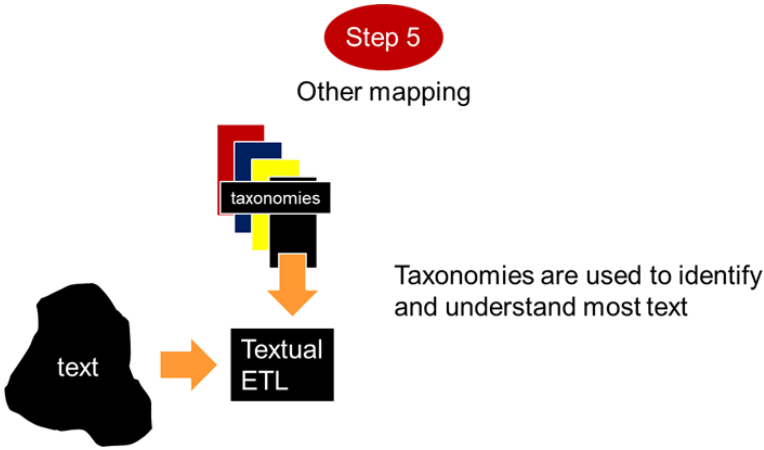
FINDING NAMES

As a simple example of alternate mapping, consider finding names in a document. It is very difficult for a taxonomy to encompass all of the first names that people have. (It can be done but it is a cumbersome and complex task.) And it is flatly impossible for a taxonomy to have an inclusion of all

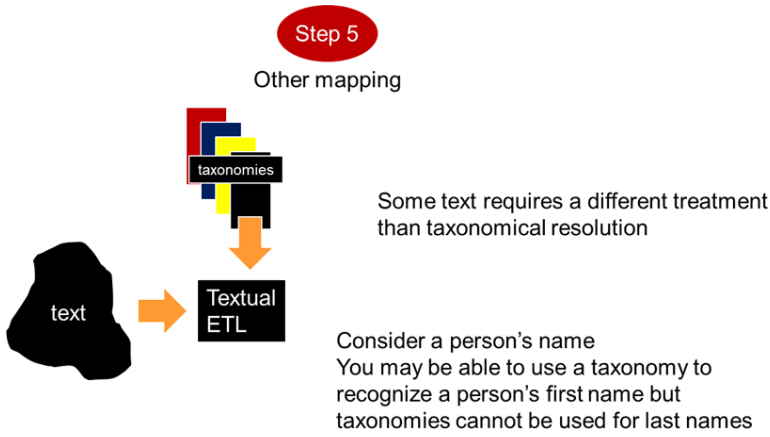
of the last names that there are. While first names are common, last names are not.

So, trying to find proper names in a document requires an approach other than taxonomical resolution.

For finding and identifying certain types of data (such as a name), an alternate mapping approach is required.



There needs to be a way to include a mapping of text other than through taxonomical resolution. It is through other forms of mapping that Textual ETL knows how to interpret the text found in a document.



MAPPING PREFACING INFORMATION

As a simple example of how other mapping works in Textual ETL, consider the prefacing information that is found on many Internet sites. Prefacing information typically includes:

- Date
- Location
- Rating
- From:
- Comments…..

Other mapping

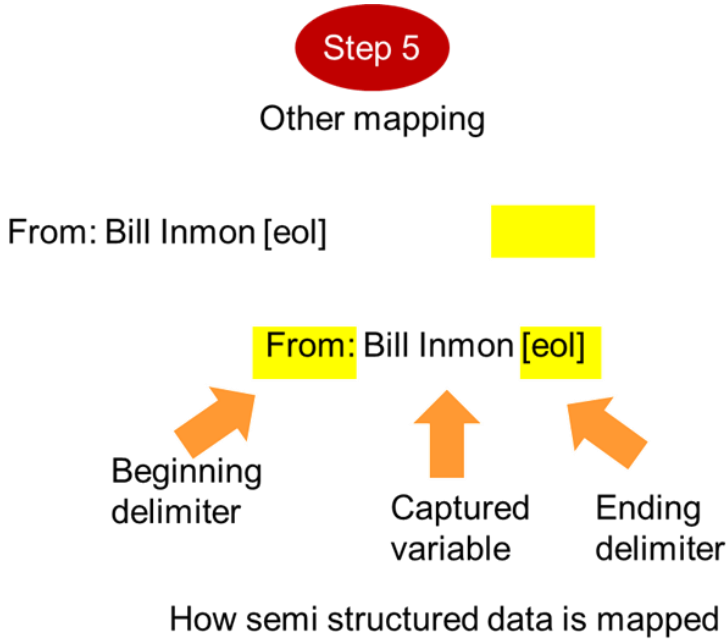
Date: xxxxxx
Location: yyyyyyyyy
Rating: zzzzzzzzz
From: nnnnnnnnnn
Comments: ……

How do you read and capture semistructured data?

So, how exactly do you go about capturing the content of data that is mentioned in the preface?

You define to the system three different things: A beginning delimiter, an ending delimiter, and a name of the data found. In the case of date, the beginning delimiter would be the string "Date:" In the case of ending delimiter the analyst would choose the special character of end-of-line. In this case the special character is marked as "[eol]". Then the analyst would assign a name to the variable – dateofcomment.

Now the system knows to pick up information found in the date preface.



How semi structured data is mapped

Of course, all other fields that need to be picked up would be defined in a similar manner. When the system reads the raw text and finds the beginning delimiter, the system reads until it finds the ending delimiter. Then the system picks up and stores all the text that was found between the beginning delimiter and the ending delimiter.

This example is the simplest form of Inline Contextualization. In practice, the concept of a delimiter can get quite complex. This means that in practice all types and concepts of delimiters must be available to perform Inline Contextualization properly. You may ask, "What do

you mean? Isn't a [eol] just an End of Line?" Well, couldn't the end of line be the end of a sentence, or even, within the sentence, the end of a thought? It can get complex quickly. Fortunately, these concepts have been considered and added to Textual ETL within the Nexus.

IN THE DATA BASE

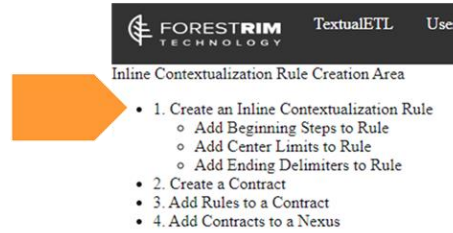
The data that has been found is placed into the data base just like all other data is placed into a data base. And the analyst can organize the data by document and by sorting on the document's id, which is attached to every entry in the data base.

In doing so – by sorting on bytes address of the document and the source of the document, the analyst can reconstruct the comment the way it was originally placed in the data base itself.

The other mappings – called inline contextualization – are found in Textual ETL where shown –

Other mapping

Here is where you define the parameters for other mappings



GETTING TO THE NEXT STEP

Other mappings (if needed) are defined

CHAPTER 7

Textual ETL

After the taxonomies are prepared and are in place, after the data has been reduced to an electronic text format, after deidentification has been done (if necessary) and after other mappings have been done (if necessary), it is time to execute Textual ETL.

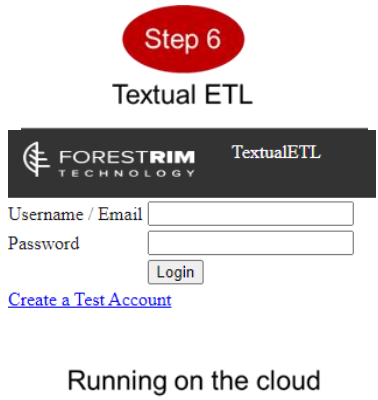
The execution of Textual ETL takes the parameters that have been specified and uses those parameters to transform text into a data base format.

Textual ETL

There are two ways to execute Textual ETL. The preferred mode is on the cloud. But it is possible to execute Textual ETL on premise if desired. But it is far easier and less expensive to execute Textual ETL on the cloud.

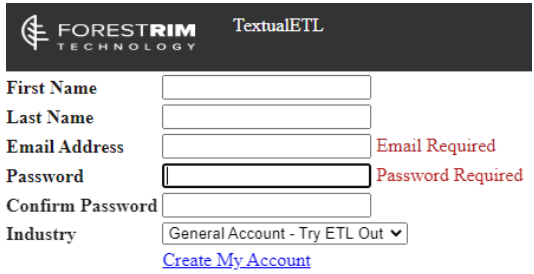
The entry point for Textual ETL on the cloud is to sign on into the cloud. Before you run on the cloud you need to have had Forest Rim set up an account for you.

You enter "RimETL.com" into your machine and the following screen appears. You enter your account number and your password.



If you have a Username and password already set up by ForestRim, enter them. If you do not, click on the Create a Test Account link.

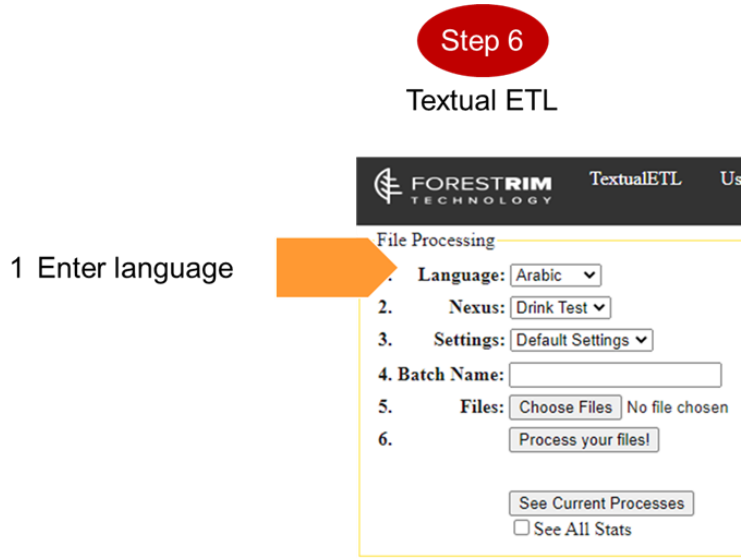
In the next screen,



Enter your correct information. Your email address, which will be verified, will become your Username. We suggest

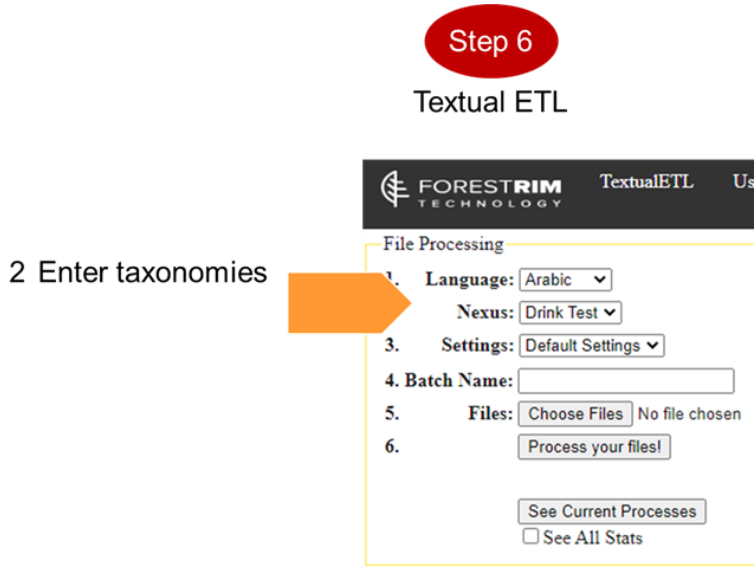
that for the Industry, use the "General Account - Try ETL Out" option. This will give you the ability to use Textual ETL with a limited number of files in each batch each day. You will have many options for to choose for a Nexus (a related group of taxonomies). If you need a Nexus for another industry, contact Forest Rim. Once you set up your account, you can log in.

When your account number has been entered correctly, you see the following screen –

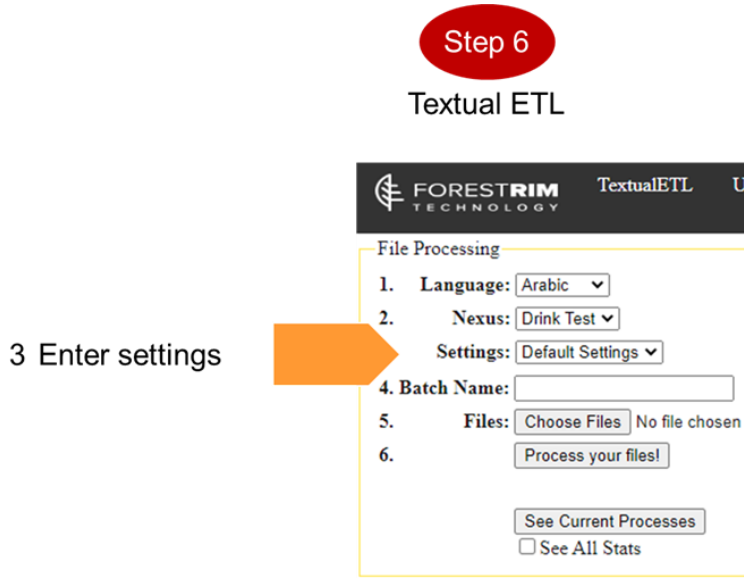


The first decision you have to make is to tell the system what language you will be operating in. You enter your language in this box.

Next you tell the system which Nexus you want to be using. Remember that a Nexus is a set of Taxonomies and rules that have been set up to provide focus and context to process your documents. A new Nexus can be created and customized for your particular purpose. Textual ETL's Nexus is, at its lowest root, one or more specific taxonomies.



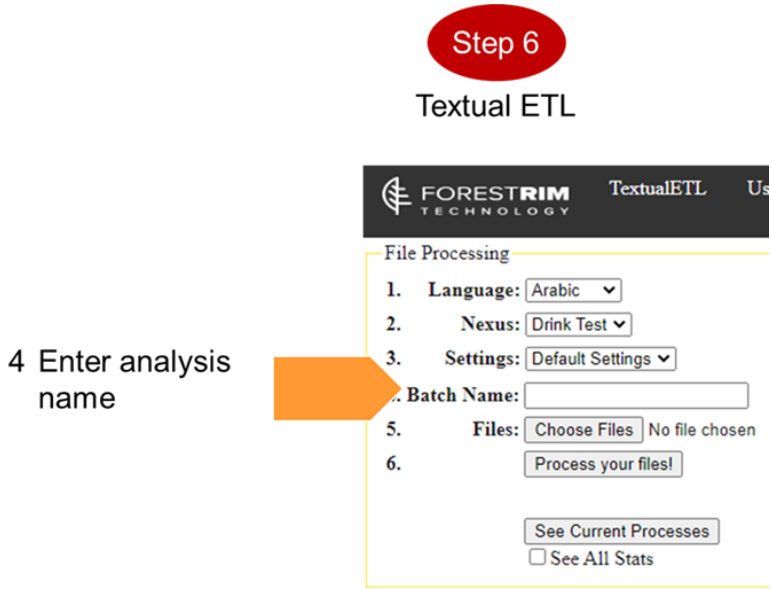
After you have specified the taxonomies you need, the next choice is to tell the system what system settings you want to use. You open the panel and select the settings.



After you have selected the settings you want to run under, the next step is to give your analysis a name. It is advised that you give it a name that is:

- Unique
- Meaningful
- Easy to recognize

Your analytical output will be registered and identified under this name.

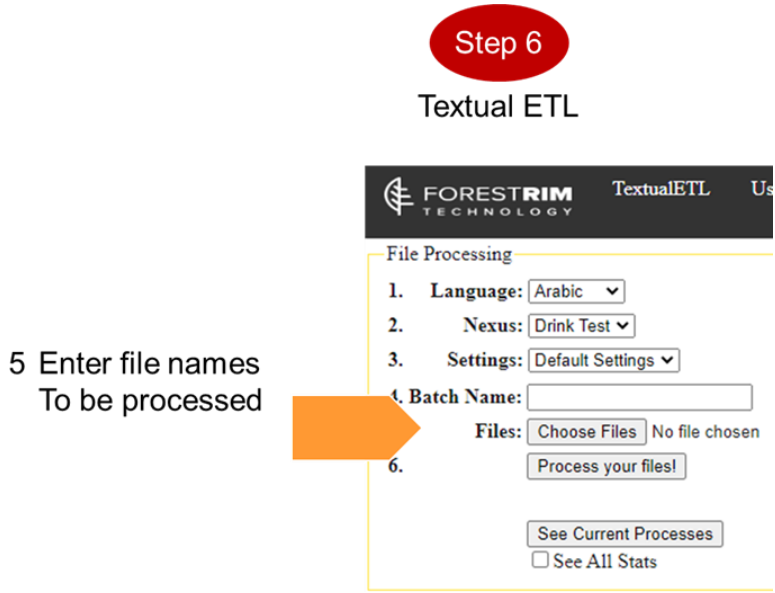


After you have assigned the name to your analysis, the next step is to tell the system what files you want to have processed.

You need to enter an amount of data that the system can process. At this point it really helps to have the system broken into smaller sets of text. Stated differently, the system – the cloud – has a limit on how much data it can swallow. You don't want to exceed this limit. The basic reason for this limit is not within Textual ETL, but rather the security nature of the World Wide Web. Microsoft, Google, and Apple limit the time a web page can run to around 200+ seconds. So, if you have more files than 200 seconds up uploading time can handle, then Textual ETL will work

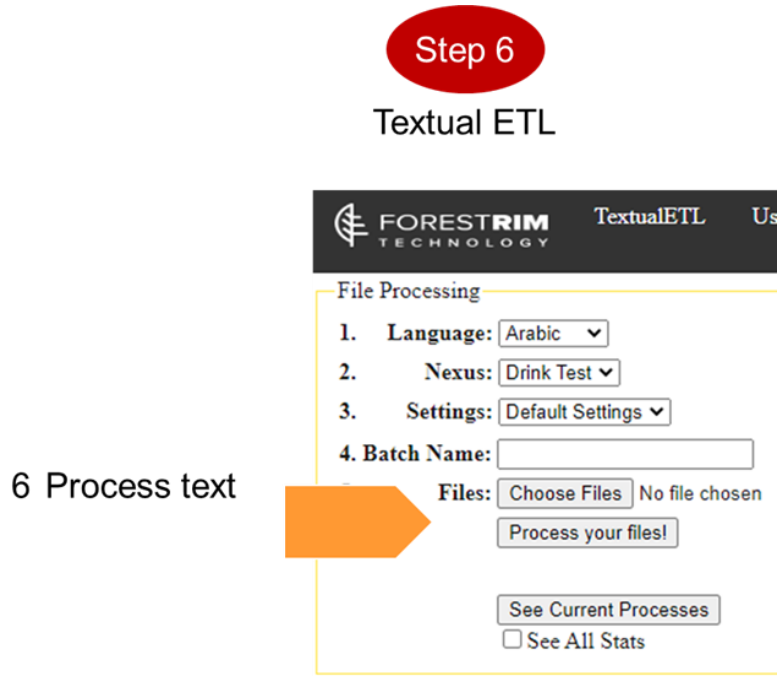
with what can be uploaded. Textual ETL has no limit on the actual behind the scenes processing time as that is done in a separate engine on the cloud.

Just so you know, if you need to process larger batches at one time, there is a desktop version of Textual ETL that will allow you to do this. Otherwise, if you have a lot of text to be processed, you simply can take your output from each iteration and merge it together after you have processed your individual small runs of analysis.



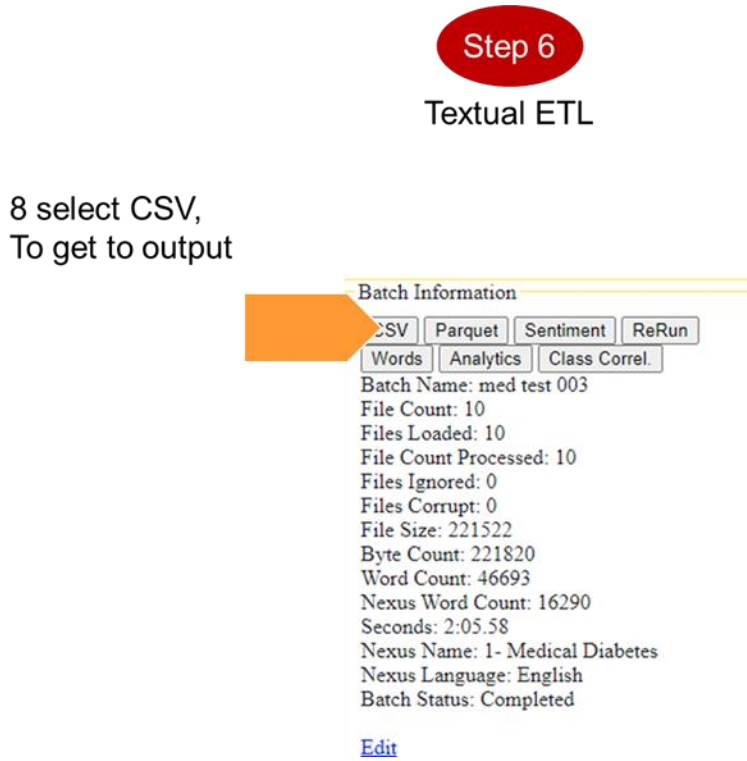
After you have identified the files that you want to have processing executed on, the next step is to execute the text processing. Depending on the amount of data you have specified, the system may take a while to do its execution.

The system will tell you when it has finished executing.

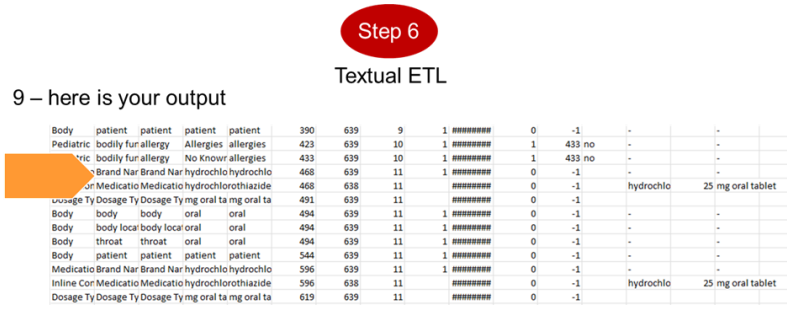


The system advises you of the status of an execution. When your input has been processed, the system will tell you it has completed the analysis.

Once the system has finished its processing, you find the button that allows you to look at your output. You press the CSV button to look at your output. You may choose to have your output in a parquet file if you wish.



The system generates a command to download your output. When you look at the output, this is what you see.

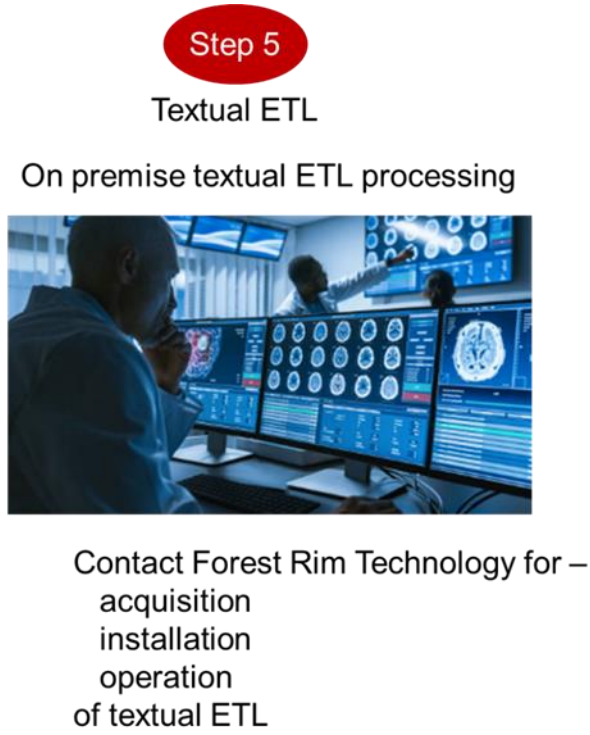


Body	patient	patient	patient	patient	390	639	9	1	########	0	-1	-	-	
Pediatric	bodily fur	allergy	Allergies	allergies	423	639	10	1	########	1	433	no	-	-
[illegible]ric	bodily fur	allergy	No Knowr	allergies	433	639	10	1	########	1	433	no	-	-
[illegible]	Brand Nar	Brand Nar	hydrochlo	hydrochlo	468	639	11	1	########	0	-1	-	-	
[illegible]on	Medicatio	Medicatio	hydrochlo	rothiazide	468	638	11		########	0	-1	hydrochlo	25	mg oral tablet
[illegible]osage Ty	Dosage Ty	Dosage Ty	mg oral ta	mg oral ta	491	639	11		########	0	-1			
Body	body	body	oral	oral	494	639	11	1	########	0	-1	-	-	
Body	body loca	body loca	oral	oral	494	639	11	1	########	0	-1	-	-	
Body	throat	throat	oral	oral	494	639	11	1	########	0	-1	-	-	
Body	patient	patient	patient	patient	544	639	11	1	########	0	-1	-	-	
Medicatio	Brand Nar	Brand Nar	hydrochlo	hydrochlo	596	639	11	1	########	0	-1	-	-	
Inline Cor	Medicatio	Medicatio	hydrochlo	rothiazide	596	638	11		########	0	-1	hydrochlo	25	mg oral tablet
Dosage Ty	Dosage Ty	Dosage Ty	mg oral ta	mg oral ta	619	639	11		########	0	-1			

Once you have finished the processing for text you can return and process another batch. Of course, at the end if you have multiple batches to be processed, you can easily combine the different iterations of output into a single file.

The other option than processing on the cloud is processing on premise.

Some organizations have to have the security of on-premise processing. This is not the preferable way for running Textual ETL. It requires operational training and acquisition costs. But it is possible to run Textual ETL entirely on premise if necessary.



In order to make arrangements for the running of Textual ETL on premise, you need to contact Forest Rim technology and make necessary arrangements.

Now that we have processed our text through the Textual ETL engine, we can get ready for the next step, working with the Standard data base output generated by Textual ETL.

CHAPTER 8

Phase I Database

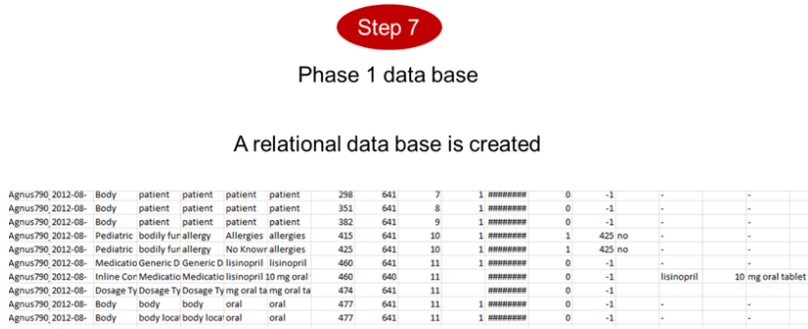
The result of running Textual ETL is the creation of a data base. This data base that has been produced can be termed a Phase I data base. It is called a Phase I data base because it sits at the root of many other kinds of analytics and is the direct output of Textual ETL.

THE PHASE I DATA BASE

The data base that is created is a relational like data base. As such it can be sent to many other types of storage of data – Teradata, Hadoop, Extreme, et al. And it can certainly be sent to a standard relational data base.

The relational data base is important because that is the format supported by popular analytical tools such as Tableau, Qlik, PowerBI, et al.

The data base that has been created represents the transformation from text to a data base. Among other things, this means that instead of looking at and analyzing a handful of records, once in the form of a data base, the analyst can look at and analyze an unlimited number of records. Putting text into the form of a data base unleashes the tools of automation.



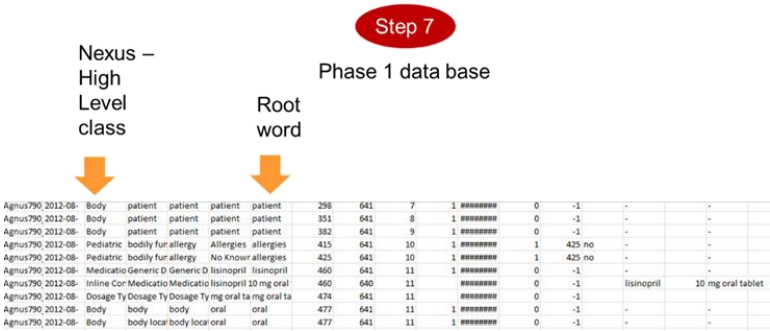
Agnus790_2012-08-	Body	patient	patient	patient	patient	298	641	7	1	########	0	-1	-	-	
Agnus790_2012-08-	Body	patient	patient	patient	patient	351	641	8	1	########	0	-1	-	-	
Agnus790_2012-08-	Body	patient	patient	patient	patient	382	641	9	1	########	0	-1	-	-	
Agnus790_2012-08-	Pediatric	bodily fur	allergy	Allergies	allergies	415	641	10	1	########	1	425	no	-	-
Agnus790_2012-08-	Pediatric	bodily fur	allergy	No Knowr	allergies	425	641	10	1	########	1	425	no	-	-
Agnus790_2012-08-	Medicatio	Generic D	Generic D	lisinopril	lisinopril	460	641	11	1	########	0	-1		-	-
Agnus790_2012-08-	Inline Cor	Medicatio	Medicatio	lisinopril	10 mg oral	460	640	11		########	0	-1		lisinopril	10 mg oral tablet
Agnus790_2012-08-	Dosage Ty	Dosage Ty	Dosage Ty	mg oral ta	mg oral ta	474	641	11		########	0	-1			
Agnus790_2012-08-	Body	body	body	oral	oral	477	641	11	1	########	0	-1		-	-
Agnus790_2012-08-	Body	body loca	body loca	oral	oral	477	641	11	1	########	0	-1		-	-

ROOT WORD AND NEXUS CONTEXTULIZATION

The data that has been placed in the data base is of great importance. Easily the most important data elements are the root word and the Nexus Contextualization. The root word is the word that has been lifted directly from the text. The multiple layers of context defined in the Nexus is the

interpretation of the context of the word as understood by Textual ETL used for analysis and aggregation.

Both of these types of data are very useful to the analyst.

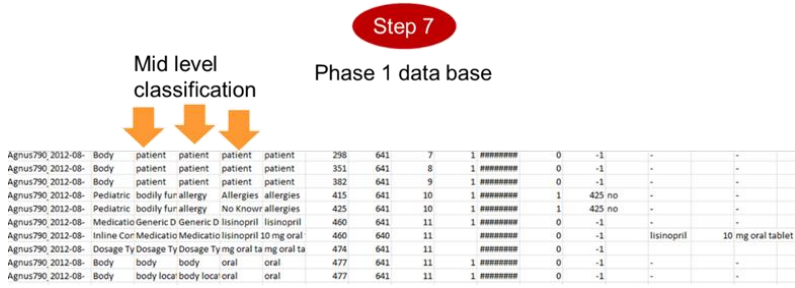


Agnus790_2012-08-	Body	patient	patient	patient	patient	298	641	7	1	########	0	-1	-	-
Agnus790_2012-08-	Body	patient	patient	patient	patient	351	641	8	1	########	0	-1	-	-
Agnus790_2012-08-	Body	patient	patient	patient	patient	382	641	9	1	########	0	-1	-	-
Agnus790_2012-08-	Pediatric	bodily fur	allergy	Allergies	allergies	415	641	10	1	########	1	425 no	-	-
Agnus790_2012-08-	Pediatric	bodily fur	allergy	No Known	allergies	425	641	10	1	########	1	425 no	-	-
Agnus790_2012-08-	Medicatio	Generic D	Generic D	lisinopril	lisinopril	460	641	11	1	########	0	-1	-	-
Agnus790_2012-08-	Inline Cor	Medicatio	Medicatio	lisinopril	10 mg oral	460	640	11		########	0	-1	lisinopril	10 mg oral tablet
Agnus790_2012-08-	Dosage Ty	Dosage Ty	Dosage Ty	mg oral ta	mg oral ta	474	641	11		########	0	-1		
Agnus790_2012-08-	Body	body	body	oral	oral	477	641	11	1	########	0	-1	-	-
Agnus790_2012-08-	Body	body local	body local	oral	oral	477	641	11	1	########	0	-1	-	-

Another type of data that is found in the data base are the mid-level classifications. On occasion these mid-levels of classification are useful in doing analytical processing.

As an example of a mid-level classification, consider the word "Tylenol". Tylenol is a medication taken for pain relief. Tylenol is sold over the counter. The medical name for Tylenol is acetaminophen.

In the broadest sense Tylenol is a form of medication. So, the different levels of abstraction might look like acetaminophen -> Tylenol -> pain reliever -> analgesic -> medication. In this case the words – "Tylenol", "pain reliever" and "analgesic" are all forms of mid-level classification. These mid-level words show the progression from "acetaminophen" to the high-level Nexus classification "medication".



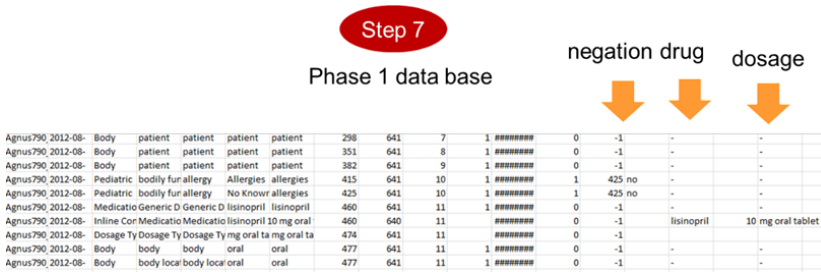
Other useful data that appears in the data base are location, file size and sentence sequence. Location identifies the byte in the source where the root word appears. File size indicates the size of the file, and sentence sequence indicates the sentence number in which that word appears.

Yet other useful data found in the data base includes negation, drug, and dosage.

Negation occurs when a word or phrase is negated. When a doctor says – "You don't have cancer", the word cancer is negated. Textual ETL will pick up the word cancer, but the

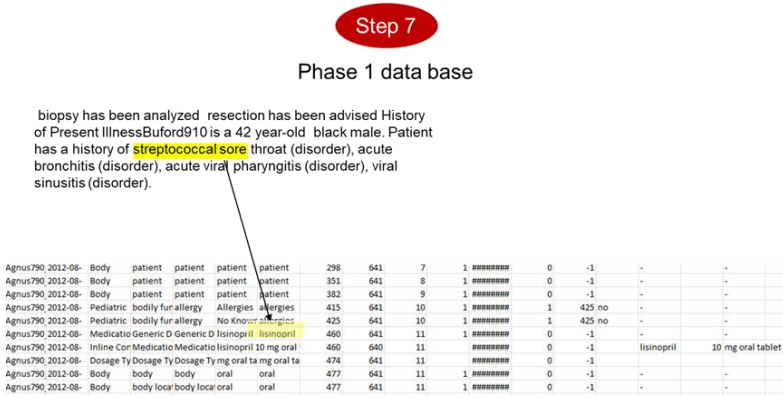
existence of the word cancer is negated by the statement made by the doctor.

Drug is shown for medical records, along with the dosage of the drug which is on the doctor's orders.

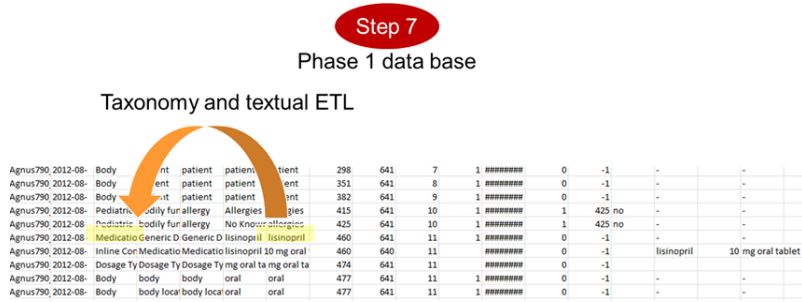


FROM RAW TEXT TO A DATA BASE

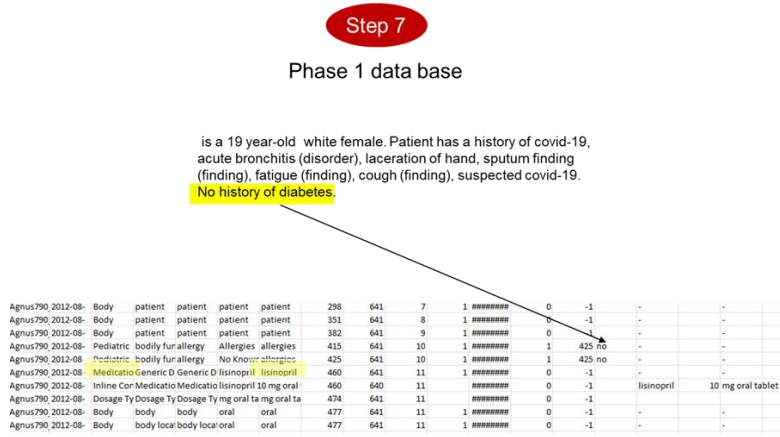
In case it isn't clear, the raw text shows up in the data base, as shown -



The Nexus and the different levels of abstraction are created by analyzing Textual ETL's taxonomies and other mappings.

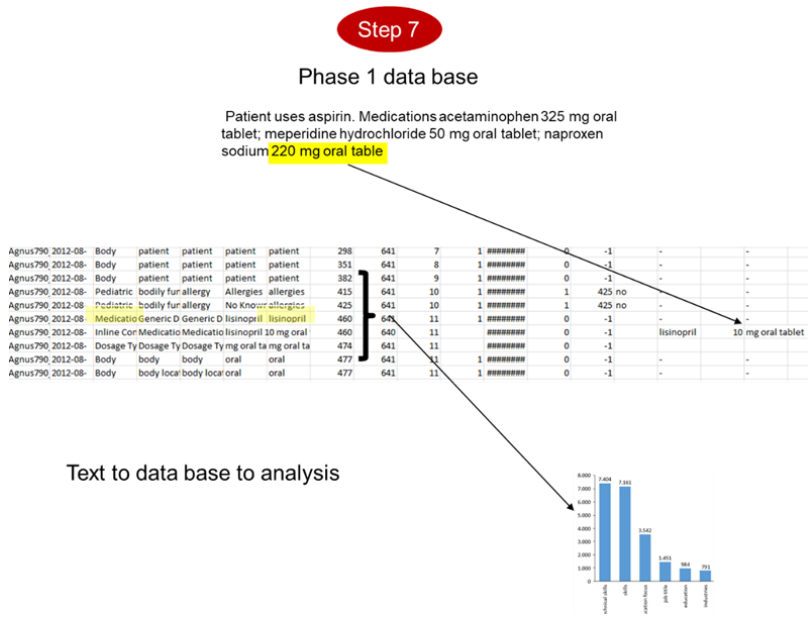


The incidence of negation is shown as well.



FROM RAW TEXT TO A DATA BASE TO ANALYTICS

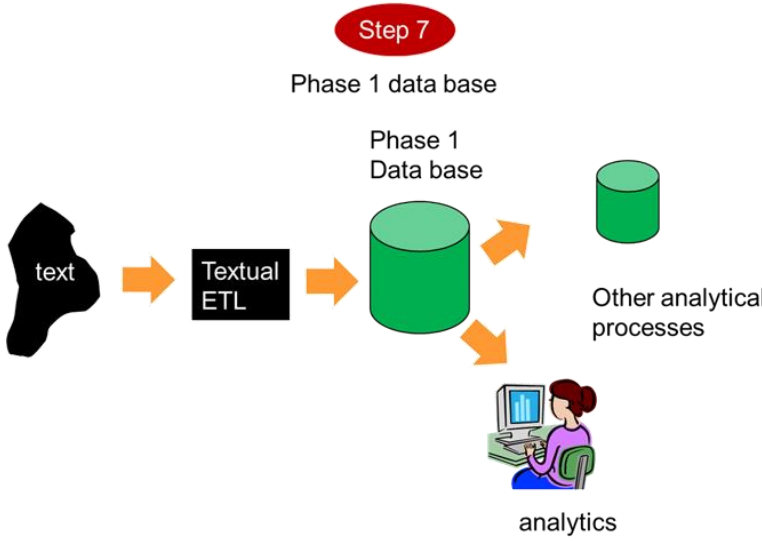
The complete cycle of data is shown in the figure. Raw text is read and is placed in a data base. Analytics are done from the data. This diagram shows the simple transformation by Textual ETL that enables analytic processing to be done on text.



Once the Phase I data base is built, it may either be used directly for analytics or used indirectly for other types of analytics.

As an example of where the Phase I data base is used as a stepping stone, consider sentiment analysis.

In sentiment analysis it is necessary to collect other parts of raw text that haven't been collected, such as punctuation.



GETTING TO THE NEXT STEP

Correlative Analytics

SENTIMENT ANALYSIS

After the Phase I data base is created, it is now time to do analytical processing with the text that has been transformed from text into a data base.

There are two types of analytical processing that will be discussed. But there are MANY other types of analytical processing that can be done. The ones that are discussed are really the most basic and used forms of analytical processing.

The two types of analytical processing discussed here are probably the most common type of analysis.

CORRELATIVE ANALYSIS – AN INTRODUCTION

Some types of text lend themselves to different kinds of processing. For example, in doctor's notes there is almost never any sentiment. Doctors express things such as medications, procedures, and symptoms. Doctors treat patients whether they like or dislike the patient. So, sentiment is not something that is done on doctor's notes. Instead correlative analytics is done on doctor's notes.

SENTIMENT ANALYSIS – AN INTRODUCTION

On the other hand, retail business and the hospitality industry are very interested in the sentiment of customers and prospects. Retailers and hospitality industries are most interested in both the good things that people say and the bad things that people say. Listening to the customer is the basis for increasing business and increasing revenue flow.

CORRELATIVE ANALYSIS

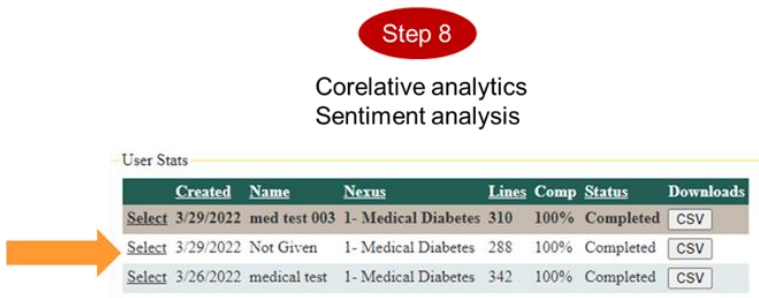
The first and simplest analysis is a correlative analysis. In a correlative analysis, the occurrences of different occurrences of text within a record are analyzed. For example, suppose the medical records of 10,000 patients are accessed. The analyst wants to find out:

- Out of the 10,000 people how many have had COVID?
- Of the people that have had COVID, how many are smokers?
- Of the people that have had COVID, how many are cancer patients?

- Of the people who have had COVID, how many are overweight?
- and so forth.

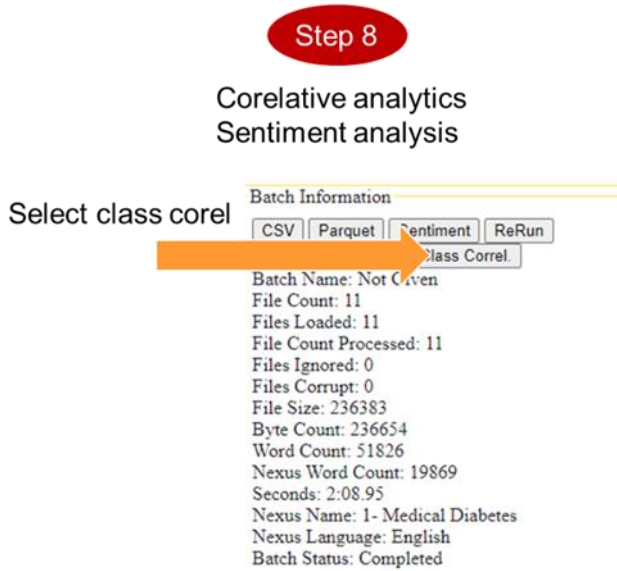
Once the records are found and analyzed, the occurrences are correlated together.

The correlative analysis begins with the selection of a study that has passed through Textual ETL. This is found by going to the cloud. Once on the cloud, the study that has been done is selected -



Select the analysis that you want to study

Once the study has been accessed, the next step is to press the Class Correl button.



Once the Class Correl button is pressed, the next step is to wait for the analytics process to finish. Once the process is finished you will see a screen.

The screen contains the Pearson Correlation Coefficient matrix for the study you have selected.

The Pearson coefficient shows the correlation of every variable encountered in the study matched against all other variables. Not only are the variables shown, but there is a description of the relationship itself. A green indication shows a positive relationship. The red indicates a negative relationship. A light green indicates a lightly positive relationship. A light red indicates a light negative

relationship. By putting your mouse over a number in the matrix, you can see the exact correlation, from -1 to 1, for those two classifications, categories, or words.

Step 8

Corelative analytics
Sentiment analysis

Pearson coefficient matrix

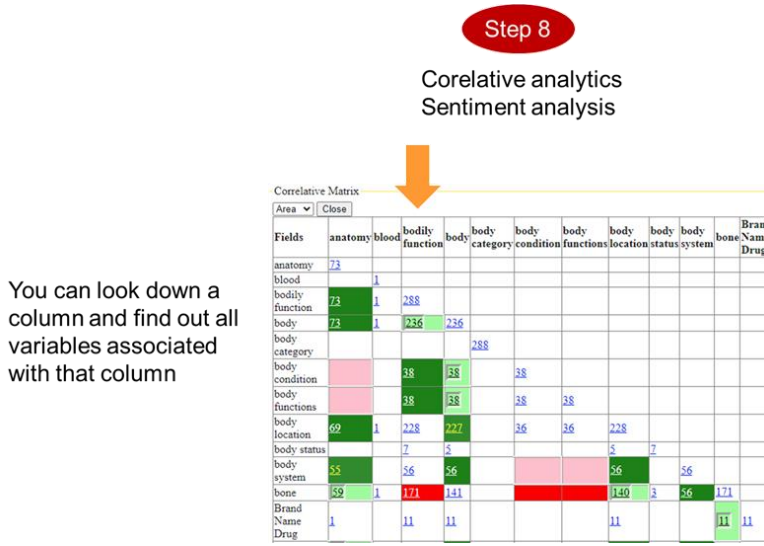
Correlative Matrix

Area Close

Fields	anatomy	blood	bodily function	body	body category	body condition	body functions	body location	body status	body system	bone	Brand Name Drug
anatomy	73											
blood		1										
bodily function	73	1	288									
body	73	1	236	236								
body category					288							
body condition			38	38		38						
body functions			38	38		38	38					
body location	69	1	228	227		36	36	228				
body status			7	5				5	7			
body system	55		56	56				56		56		
bone	59	1	171	141				140	3	56	171	
Brand Name Drug	1		11	11				11			11	11

The Pearson Correlation Coefficient matrix can be looked at two ways. One way is to look at the column and move down the column. In doing so you will access the data for all of

the correlations with the data element that appears in the column.



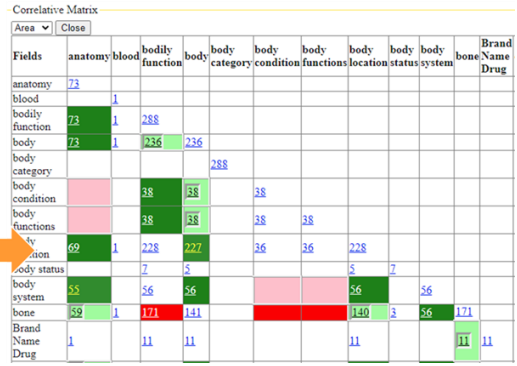
You can look down a column and find out all variables associated with that column

And when you look across the spreadsheet to see all of the columns that correlate against the row, you can see the strength of the relationships correlated against the row highlighted in green or red.

Step 8

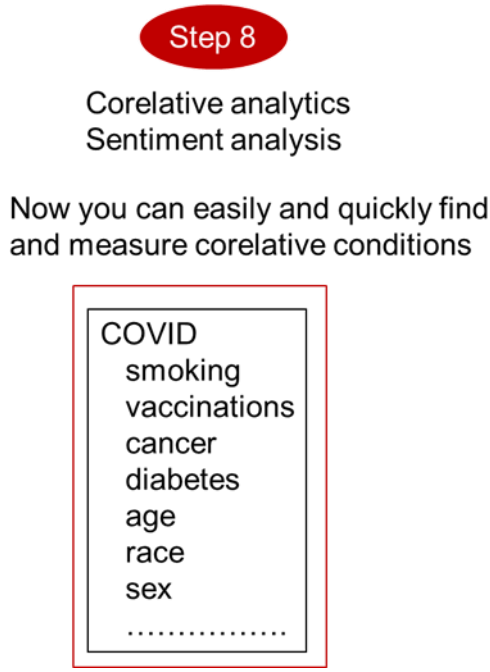
Corelative analytics
Sentiment analysis

You can look across a row and find out all variables associated with that row



Once you have the Pearson coefficient matrix created, it is a simple matter to study the relationships.

For example, suppose you were studying COVID. You could easily see how COVID related to the other factors found in the patient records that have been entered.

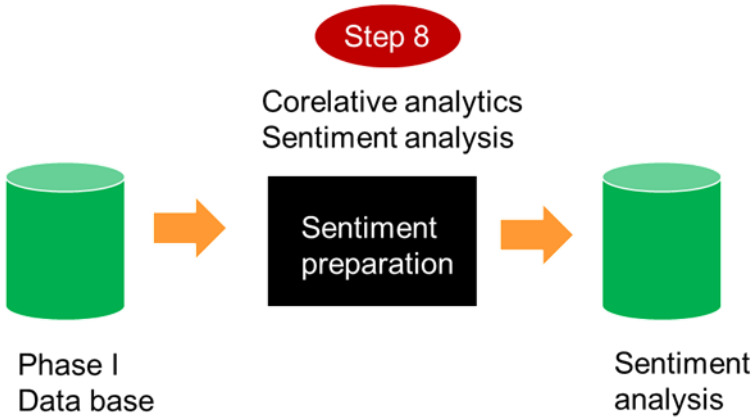


SENTIMENT ANALYSIS

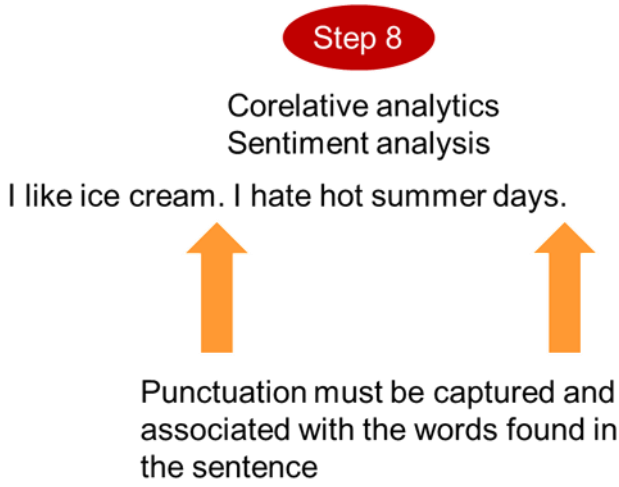
The second kind of common analysis that can be done is sentiment analysis. In sentiment analysis, you look for the expression of sentiment. In addition, you look for the object that is the source of the expression of sentiment.

Sentiment analysis is typically done where people are talking about their experiences with a company, its products, and its services.

The starting point for sentiment analysis is the Phase I data base. The Phase I data base has the records that have been selected for basic analysis. But the Phase I data base needs other data as well in order to do sentiment analysis.

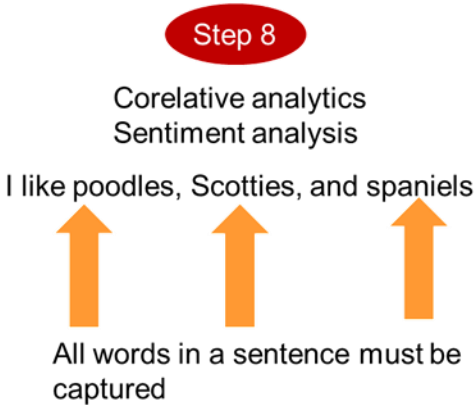


One piece of data that is needed for sentiment analysis is sentence punctuation, which includes periods, question marks and exclamation points. Punctuation marks are necessary for sentiment analysis because sentiment analysis must be done on a sentence by sentence basis, and punctuation marks are necessary to delineate the beginning and the ending of sentences.

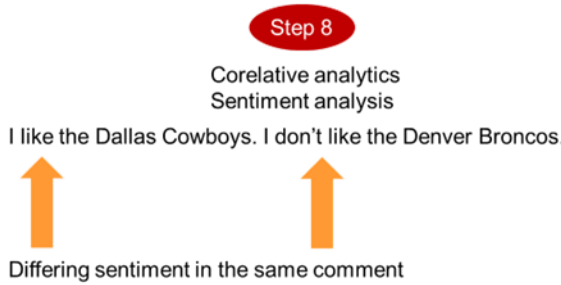


Another piece of information that is necessary for doing sentiment analysis is negation. If all you capture is positive expression, then you will interpret the negation of a positive expression as a positive statement, and this is incorrect. Therefore, there needs to be the capture and analysis of negation. And of course, the negation of a negative term is a positive expression. (E.g., "I do not hate ice cream." is a positive expression.)

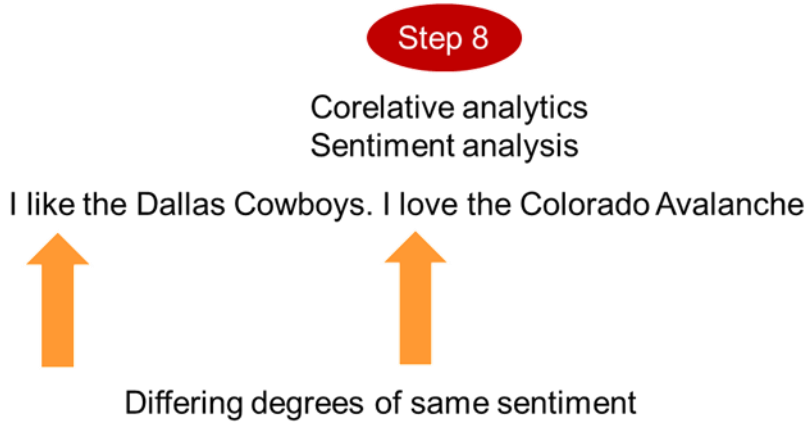
Another thing that needs to be captured is the inclusion of all the items that appear in an expression. It is not sufficient to merely look at the first item in a sentence that appears.



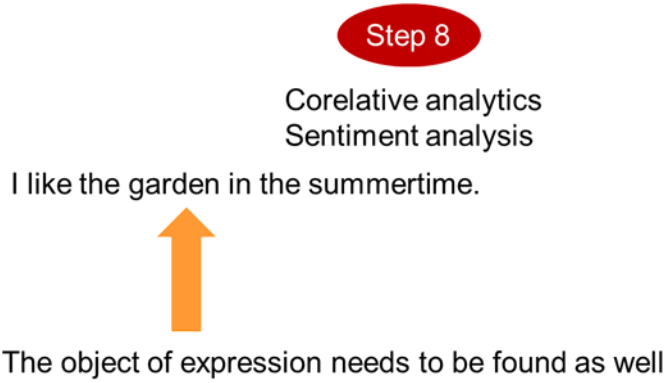
Another factor is the occurrence of different sentiment in the same comment. Differing sentiments in the same comment occur all the time.



Another analytical factor that must be taken into consideration in doing sentiment analysis is that of different degrees of like or dislike. I "like" something is a lower degree of positivity than I "love" something.

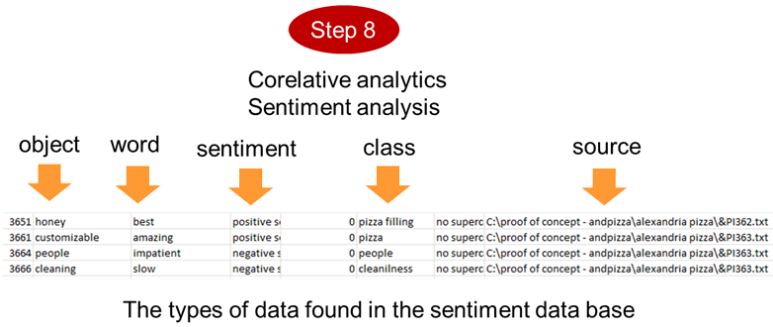


In addition to the recognition and interpretation of sentiment is the necessity of finding what has been expressed – what is the object of the expression.



THE SENTIMENT DATA BASE

The sentiment data base that is produced looks like -



The types of data found in the sentiment data base

The data that is found in the sentiment data base includes –

1) The word that is the root of the analysis

2) The object of the expression

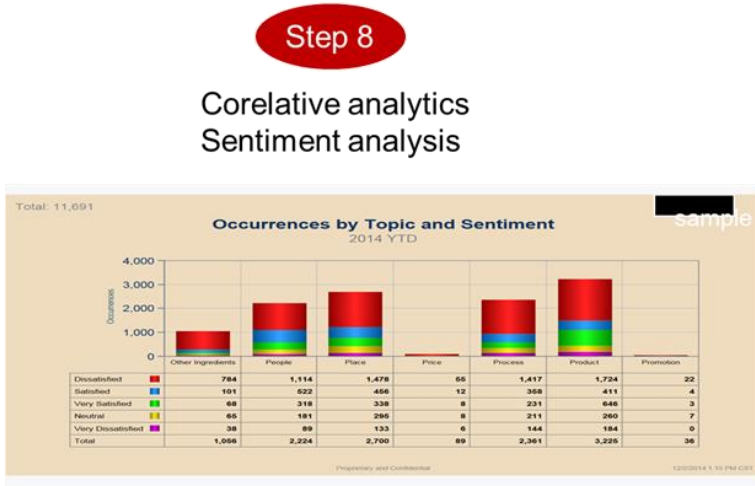
3) The interpretation of the expression

4) The higher-level classification of the object of the expression

5) The identification of the source of the word being analyzed

These fields are gathered by looking at the Phase I data base and the raw text.

As an example of the analytics that can be created by looking at the sentiment analysis data base, consider the following visualization -



The following visualization looks at customer comments gathered from customers.

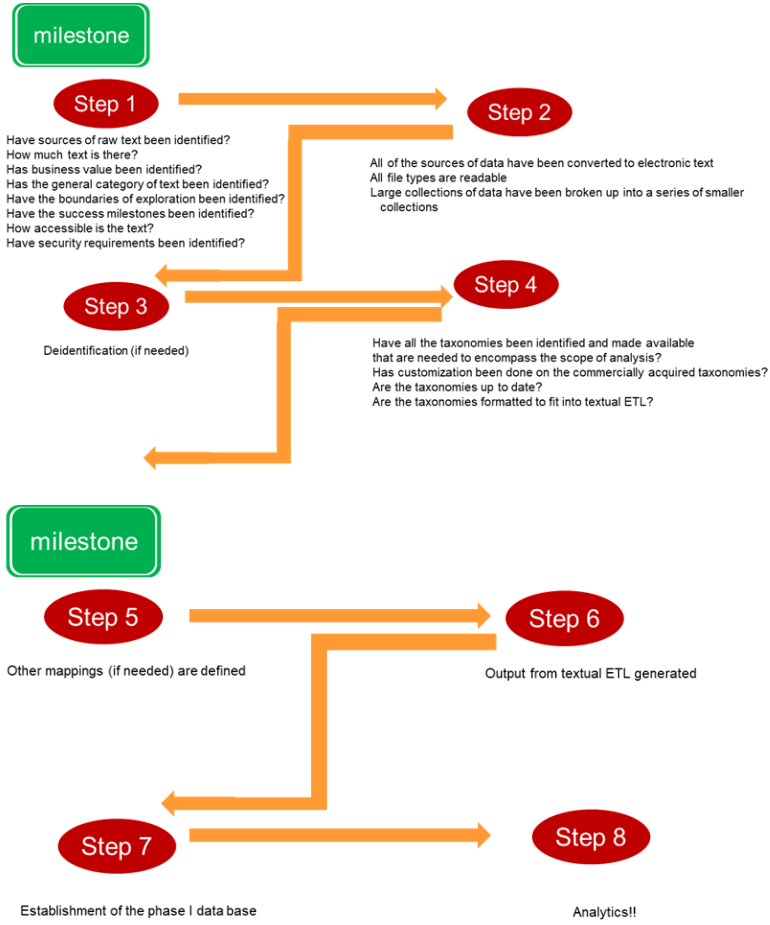
CHAPTER 10

Milestones

The path to doing text analytics is a very straightforward path. It does not have to be made the tortuous path that vendors and consultants make it to be. It is true that you can take many turns from the path and go down deep corridors of analysis. But such detours are neither necessary nor productive.

Instead, having a straightforward plan and a single-minded approach is the key to quick and meaningful results when doing text analytics.

The path to textual analytics and its milestones that has been described is shown -



At first glance the path seems to be complicated. Indeed, there are a number of steps. But, not all organizations have to pass through the steps. For example, many organizations do not need to do deidentification. And even if they need to do deidentification, there are automated processes for doing that task.

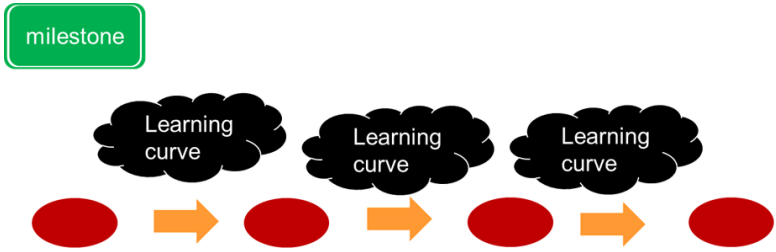
Or another optional task is other mapping. In many cases there is no need for other mapping. Or even when there is a need, there is only a small amount of data that needs to be mapped in a manner other than taxonomical resolution.

The first time an organization passes through the path to textual analytics there is a learning curve. And like all learning curves, it takes some time and effort to get through the learning curve. As with all learning curves, there is a certain amount of trial and error. The analyst tries one approach and sees how it works. After seeing the results, the analyst tries another approach and improves the way textual analytics is done. Such a trial and error approach is inevitable with text because text is inherently imprecise.

And it takes time to go through the trial and error approach.

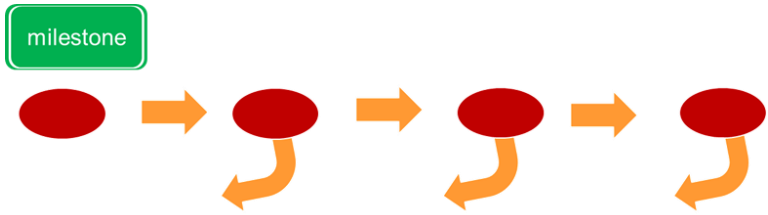
The consequence is that the first time through the path to text analytics, the journey will not be fast and efficient. But the journey is both inevitable and valuable. It is valuable because the lessons learned create a solid final product.

The path to text analytics is well documented and clearly laid out.



The first time through the steps, there is a learning curve.
It takes time and mistakes to go through the steps successfully

One of the really advantageous things about the path to text analytics is that path is iterative. At any point in time, the analyst can go back and start over. However, when the analyst goes back and starts over, he/she does not have to discard all the work that has been done. Instead, the analyst can build on the analytical work that has been done.

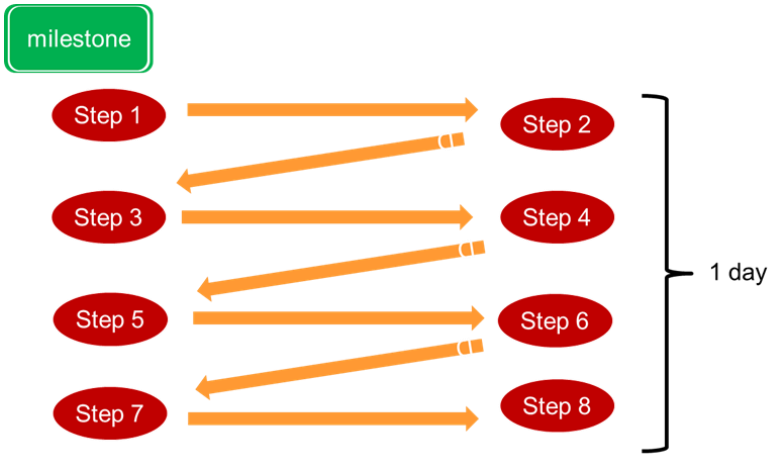


At any point you can always go back, rework analysis and repeat

In doing so, one level of iteration builds on the next level of iteration. There is no need to go back to the beginning and destroy all the work that has been done.

So, once the analyst has gone through the learning curves, how long does a text analytics study take? The answer is that – regularly – a full text analytical cycle can be done in a day, depending on the volume of data. If there are very large amount is data to be processed, it may take more than a day. But even then, when data can be processed in a parallel manner, the processing can still be done in a day's time. Even with reiterations, the full cycle, even with large amounts of data, can take only a few days.

There is no need to spend weeks and months on doing text analytics.



How long does it take to execute the full cycle after you have been through the learning curve?

CHAPTER 11

The Semantic Layer

Abstractions done correctly will factor complexity down to a level where the analyst doesn't have to spend any brain cycles thinking about it. Abstraction allows work with a well thought out interface accomplish more without having to always consider the system at a molecular level.

It turns out business people also like abstraction. This shouldn't be surprising as businesses model complex real-world concepts where the details matter. From calculation to contextual meaning, abstraction helps with correctness and understanding. Abstraction is inherent to the very process of speaking. So, abstraction in the management of data is natural and normal.

Take for example a simple definition such as an "Employee". Does employee include contractors? Remote people? What is "Net Sales"? Is net sales net of invoice line-item costs and/or net of rebates?

A small use case may contain tens of these calculations while a departmental model may contain hundreds. Without some level of abstraction, business is beholden to IT to generate and run reports or risk making big, costly and

worst of all hidden mistakes. Can you afford to have each of your employees independently trying to replicate this logic correctly in their spreadsheets and reports? Will you be able to catch the subtle yet impactful errors?

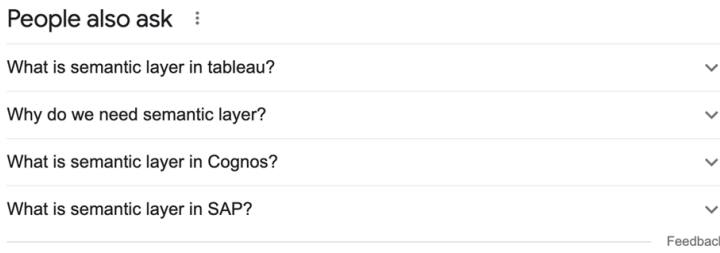
In business intelligence (BI), also sometimes referred to as analytics, the key abstraction used in the majority of implementations is called the "semantic layer".

A semantic layer is a business representation of corporate data that helps end users autonomously use common business terms. A semantic layer maps complex data into familiar business terms such as product, customer, or revenue. With a semantic layer there is a unified, consolidated view of data across the organization.

Semantic, in the context of data, means 'from the user's perspective'; which sounds like a nice clean solution to a nasty unbounded complexity problem. The semantic layer is not entirely new.

So, a Semantic Layer sounds great, what's the catch?

Check out this simple search that was done for 'semantic layer':



How many semantic layers do you need? (Hint: The correct answer is 1). Remember, the semantic layer is an abstraction and while having multiple abstractions for the same concept is sometimes right and useful, most of the time it goes against the engineering mantra of DRY (Don't Repeat Yourself).

Why Now?

To have great longevity, great ideas must evolve. The concepts of a BI semantic model introduced almost 25 years ago have been largely static. AtScale believes the value of a modern BI platform is in addressing known shortcomings and innovating on what originally made the semantic layer great: it is a force multiplier for data consumers. In-depth knowledge of calculations and underlying data structure took a backseat to generations of useful operational reports, dashboards, and other analysis built by folks that understand the business domains and not math.

What's Changed?

- Volume

More data is being generated than ever and its pace of growth is accelerating. Data is now at a size where pre-calculating analytical views of data in the form of cubes or data extracts just doesn't scale anymore. Solutions that bind the semantic layer to a specific tool or analytics platform will force compromises in data fidelity or breath to "make it fit" - unacceptable.

- Velocity

More data is coming in faster and faster. Older semantic layer approaches that require a static build (or pre-compute) phase are too slow to keep up with the onslaught of data now and in the future.

- Variety

More types of data are introduced every day. A modern semantic layer can make semi-structured data look structured; as though it is relational behind its abstractions. This means business users can keep using your standard visualization tools that have no concept of new or machine-generated underlying formats, like JSON or key-value pairs, without the need for complex, hard to maintain, no-value-adding repetitive data movement or ETL. If done right, a modern semantic layer avoids the need to retrain end-users

on a new visualization UX. They can continue to use their Tableau, Excel, Power BI, Looker, Jupyter Notebooks and the other tools that they'll only give up when you pry them from their cold, dead hands.

- Veracity

This speaks right to the heart of the semantic layer. Uncertainty comes in many forms, and business users have lost faith in the data they use to make decisions. The abstractions provide proven, tested structures and calculations that consumers can trust. The allusive "single source of truth" can only come with a universal semantic layer that encapsulates business logic and is used by all data consumers, not just one tool.

The Upside

The semantic layer isn't a single abstraction, it is a grouping of abstractions used to address different problems. Semantic layers allow for improvements in the following:

1. Usability/Understandability/Acceptance

One of the biggest complaints from the business is that it takes way too long for IT to build or alter reports for them. They want to take the helm and control their own destiny. A well-designed semantic layer with agile tooling allows

users to understand how modifying their query will result in different results, while at the same time giving them independence from IT – freedom from IT – while still having confidence their results will be correct.

2. Security & Governance

In this day and age, the enterprise has strong, and sometimes regulatory requirements that they track to know 'who' saw 'which' data and 'when'. A modern semantic layer allows users to appear as themselves to the underlying data platforms from any consumer tool, providing the ability to track the lineage of data from row level to every aggregate managed by the software.

3. Agility

Analytics agility, also thought of as "time to insight" is how long after the data lands somewhere that it can be used to make decisions. In legacy BI solutions, there was often a build process that could take anywhere from minutes for small data, to days/weeks for large data. A modern semantic layer leverages data virtualization to enable new data landing in your data warehouse to be queried by your BI tool within minutes, regardless of size. There are no full-build or full-rebuild times to wait for, and no manual ETL processes.

4. Performance & Scale

Cubes and data extracts were introduced to overcome the performance issues of data platforms. This approach introduces data copies, adds complexity, destroys agility and introduces latency. A modern semantic layer improves performance regardless of underlying data model, whether it's a snowflake, a star, or purely OLTP schema. By automatically creating and managing aggregates or materialized views inside the underlying data platform, a semantic layer learns from user query patterns and optimizes the data platform's performance and cost without data movement.

The Downside

The biggest downside of a semantic layer is that you have to build, maintain and manage it. The layer must be kept in sync with any database changes that occur. However, at the end of the day, it's a lot easier to maintain a semantic layer of definitions than it is to maintain 1,000's of reports, cubes and data extracts.

The Trend

In response to previous generations of semantic layer platforms' lack of usability, complexity, and IT's inability to

service requests in a reasonable timeframe, the market has moved to embrace a set of data discovery tools: The Tableaus, Power BIs, etc., which have done away with centralized semantic layers altogether. These tools make connecting into relational data sources a breeze. The fallout is that it means business users now need to learn how to be a data engineer, which arguably wasn't necessary with a good semantic layer.

Worse, logic is now distributed and duplicated, complex personal models have replaced a common semantic model and data includes mashups of data from the warehouse and other systems. As a result, data management has become a major problem. End users were given enough rope to shoot themselves in the foot.

Don't get me wrong. I love data discovery and machine learning tools and believe that business users and data scientists should query data using the tools they already know and love. A modern semantic layer is a solution that can deliver on both fronts - self-service data access with guard rails.

What You Need a Semantic Layer to Do

For a semantic layer to deliver value for the business, it must satisfy a number of requirements. You can use the following

as a checklist when evaluating semantic layer platforms or when building your own:

1. A semantic layer must support multiple consumer personas, including business analysts, data scientists and application developers, to deliver the full spectrum of data access and analysis.

2. A semantic layer must support multiple inbound languages to support a wide range of data consumers using their preferred protocols. Semantic layer solutions that only support SQL or JavaScript are unsuitable to serve as endpoints for a variety of popular consumption tools.

3. A semantic layer should not require the IT team to install additional client software on query consumers' machines.

4. A semantic data model must support a hub and spoke style for creating data products using an object-oriented data modeling language that allows subject matter experts to own and share data model components across teams.

5. A semantic data model must break down data silos by blending data sources on the service side to create rich, composite views across multiple business domains.

6. A semantic layer platform should support both code-based and graphical data modeling to allow engineers

and non-engineers alike to build and collaborate on data models and data modeling components.

7. A semantic data model must support data transformation expressions in the semantic data model using the native platform's SQL dialect.

8. A semantic data model must support the ability to perform pre-query & post-query calculations for handling calculations that summarize data at different levels of granularity.

9. A semantic layer platform should support inline data transformations using direct queries without data movement or creating copies of data.

10. A semantic layer must be backed by a multidimensional, cell-based engine to express complex business logic. Semantic layer solutions that use SQL-based calculation engines cannot express business constructs in a variety of contexts. A semantic layer must also support hierarchies to allow for intuitive drill paths and level relationships. Semantic layer solutions that only support dimensions and metrics do not provide an intuitive data navigation experience for end users.

11. A semantic layer must autonomously tune query performance to support interactive, live connections to data platforms. Semantic layer solutions that do not

automatically manage query performance are unsuitable for supporting direct (live) queries.

12. A semantic layer must deliver query performance at "speed of thought" with a live connection to data platforms without the need to create tool-specific extracts or imports or moving data to separate caching subsystems.

13. A semantic layer must support deep integration with IT identity management services and respect underlying data platform security policies by running queries with the user's account.

14. A semantic layer must enforce data governance in real-time for every query in order to provide comprehensive coverage and respond to frequently changing policies.

15. A semantic layer must apply query governance with dynamic filtering, column-level security and object-level security based on the query user's identity. Semantic layer solutions that lack row, column, and modeling object controls are not suitable for use cases where data access restrictions are required.

16. A semantic layer must work with a variety of data platforms equally well by supporting native platform dialects and optimizations.

17. A semantic layer must support modern data platform features and constructs to support analytics on unstructured and semi-structured data.

Do You Need a Semantic Layer?

Current BI and AI tools have focused on giving users flexibility and that often means the governance model is "no management required". Many folks we talk to are coming from a world where they write and maintain complex ETL pipelines to generate numerous extracts that are fed to BI and AI tools. Not only is this a maintenance headache, this gives the governance people a massive headache.

Businesses will have to define a semantic layer, no matter what. If you don't have experts do it, all your end users will do it for themselves in Tableau, Power BI, Excel, Jupyter Notebooks or whichever front-end they are using. A universal semantic layer will promote rational self-service, increase data adoption, reduce time to analysis and improve correctness.

CHAPTER 12

Data Future-proofing™ in line with the semantic layer

In terms of data, 'future proofing' can be defined as the process of anticipating the future and developing methods of capturing and arranging the data in a way that can minimize the gap due to missing data or relevant data for the purpose of relevance, data driven research, correlations, trends, patterns, data supported evidence, past incidents and experiences and many more. It can give you confidence to proof and support your past data findings and help reduce the unwanted shocks and surprises that can give business stresses due to missing future proof data.

"Data Future Proofing" is a new phrase. We are going to talk about this subject in detail here.

All data of an enterprise might not be relevant for that enterprise 10-20 or 50 years down the line. It is the responsibility of organizational stake holders within an

enterprise to coordinate with data architect to decide and earmark core entities and attributes that will be relevant and useful for the business benefits even in far future. Don't forget to consider the topic of privacy and confidentiality, we discussed just above while deciding the future proof data.

Future proofing of data is a vast subject. It deserves a separate book. We may even decide to write one if still needed. But we will try to cover its broader aspects here as much as we can considering the semantic layer.

We should not forget the importance of a data semantic layer for an enterprise. It is meant for long term, a long term beyond you can think of, as of now. Remember one very important thing here, technology will keep evolving, employees, consultants, vendors etc. may come and go within an enterprise, but accumulated data relevance will always be there for an enterprise. Next generation business is all about Data. All cognitive activities (including cognitive science) revolve around the data you accumulate. The more the data the better the outcome from your cognitive system. Outcome becomes more reliable and foolproof with quality and quantity of the learning data.

And – whether it is healthcare related data or insurance data. Aviation data, environmental or weather data. Social, socio economic data or behavioral data. Geographical, political or geopolitical data. Space data or the research data

of any field of study. More past data or the proven historical data can help in better future findings. Because, “Data is the New Gold”.

But biggest and most important question here is "Which Data" that need to be preserved/accumulated/stored/saved for future usage purpose. Till now while creating a data warehouse or data lake, we were thinking of a data storage for next few years. It was never considered for decades or centuries. Here we are going to consider the importance of data till eternity that can be passed from generation to generation. But we should be sensitive about its significance and relevance. We can't consider all gathered enterprise data a future proof data. All data might not retain its relevance in future or far future. It will be a silly consideration and may lead to various problems including but not limited to size/volume, policies breaches and misinterpretation or even misuse at latter stages.

While considering future proofing of data, we need to be sensitive about various aspects of data including its relevance, relevant grain level, context, format, dimensions for different perspectives, future views and viewpoints.

In this data context, a data view is what data you see, and a data viewpoint is where you are looking from. Data viewpoints are a means to focus on the business data for particular aspects of the business or the purpose. These aspects are determined by the concerns or business

purposes of a stakeholder with whom communication takes place to future proof the data. Remember one thing, the data viewpoints may sometimes depend upon stakeholder's perspective and may be subjective. Hence suggested to be little generic while deciding the prospective candidates (subject, entities and attributes to be captured) for Data Future Proofing. We understand that 'Little generic' is a subjective word, and every individual might have a different take on this word. We are calling it 'little generic' because a 'generic' or 'more generic' may lead you to capture data that may not be useful in future, and that will spoil the core purpose behind "Data Future Proofing

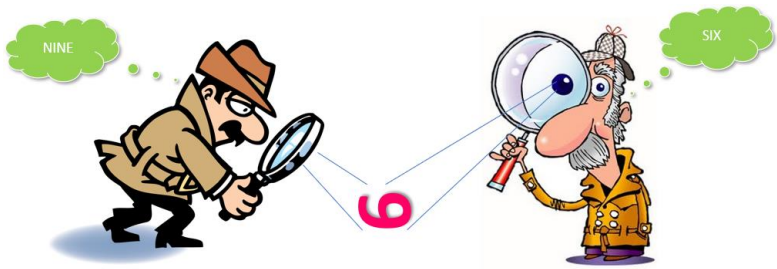


Figure – represents the two different Views of the same subject from two different Viewpoints.

Think about medical records. Capturing medical records is important. Do you know, what is the value and benefits of historical medical records? What all benefits it can bring to healthcare if utilized efficiently? Yes, you heard it right, efficiently and effectively too. It can help save human life. It can help in early diagnosis. It can help in medical research

in many ways - By helping in early medicine discovery. In advanced or faster vaccine creation for deadly health threats. It has potential to reduce the healthcare stress many folds.

Take even the Covid pandemic scenario. If the future proofing would have been done for all the past medical records of similar epidemics or pandemics like Ebola, Avian Influenza, Mers, H1N1 and if we would have efficiently and effectively, today's scenarios would have been far better. The world would have been recovered more quickly. Death could have been eliminated or minimized. Treatment of the Covid related Corona virus would have been found in advance or at very early stages. Vaccine could have been researched faster that it is done today. Lack of the historical data or more over lack of future proof data availability of the past, they would still have no clue about this corona virus. Still medical science has no direct treatment to the infection.

But do you know how much privacy and confidentiality clauses a medical record carries? Handling the medical records is a sensitive subject. Storing medical records at application level has different purpose than storing it into a data warehouse or a data lake or data lakehouse by leveraging the semantic layer. While you bring a medical record to your data lake or so, purpose is to retain it for different business needs than simply treating a patient. You certainly have the EMR application in place for the

treatment of the patient that has all applicable privacy and confidentiality policies applied on top of it. Once you capture those data into your data lake or warehouse, you might have captured it for various other purpose than only clinical or patient treatment. The organization might utilize the medical records in the data lake or data warehouse for various medical research purpose, early diagnosis purpose, accurate treatment purpose, preventive healthcare purpose, so on and so forth. And quite obviously by now we understand that how a semantic layer is going to play an important role in provisioning all desired key abstractions in the data for the purpose.

There is one thing, for the purpose of medical research or early diagnosis or proven treatment findings or preventive healthcare methodology findings, that you really need to know – 'what was the name and social security number of the patient who was diagnosed for Covid in 2021 and got the treatment in a New York hospital or the Apollo hospital, Bangalore?'. Do you really need – 'the door number or the phone number of a patient who was diagnosed with Influenza Virus in 1968 and was treated into a city hospital?'. What you might need is the Age (or age group), Sex, City/State and certainly the medical conditions including the symptoms, diagnosis, treatment and the result of that treatment. And of-course good if those records are supported with the time stamp (may be date or month and year or only year; depending upon the grain level of

data). In various cases, even the month and year of the treatment should be sufficient enough.

Another very common and scenario where the "Data Future proofing" can help an enterprise from data vulnerability - let us consider an enterprise, who captures both personal and sensitive data. Personal data like name, address, medical details, bank details and sensitive data like racial information, political opinion, religion, trade union association information, health, sex life, criminal activity etc is not relevant. The enterprise has been doing its business for last more than two decades. All of a sudden, the company is acquired by another business house from a different domain or sector – e.g. a retail company acquired a software development company or a Manufacturing company acquired a marketing research company. What will be the future of those personal and sensitive data captured by the 1st enterprise, if the Data Future Proofing was not done? Obviously, there are chances of accidental exposure to those tons of personal and sensitive data kept by the previous organization. You never know if the new owning organization has the support to handle such data.

The same thing could have been handled in a hassle-free way once you have already implemented "Data Future Proofing" into your organization. That keeps all such merger and acquisition hassles free from any such data vulnerability to the old or very old/historical personal and/or sensitive data/information.

So, consider the above medical records use case? Clarify the purpose, identify the entity and attributes to be captured and eliminate sensitive data or the data that might not be relevant for the purpose in future. Once all these things are successfully done, then we organize all data earmarked for future proofing and store it in a way we are going to discuss in details below.

Let us discuss how the future proofing can be done for the data.

We have divided "Data Future Proofing" processes in 5 different phases – Identification, Elimination, Future proofing, Organization and Storage.

FIVE PHASES OF "DATA FUTURE PROOFING":

FIGURE – FIVE PHASES OF "DATA FUTURE PROOFING"

1. Identification phase – Keeping the ultimate purpose of the future data in your mind, identify all entities and attributes that will have relevance and that is meaningful for various business needs and business benefits. Mostly such future proof data purposes include but not limited to analytics and research. You need to identify only those future proof entities and attributes of the subject. The subject can be your business domain like - healthcare,

insurance, aviation, manufacturing, education, transportation, hospitality, retail and what not.

For example – in a healthcare domain, the EMR might have hundreds of attributes, but you might need very few of them. The grain level of an EMR is up to every encounter of a patient, but for the future you might need the extraction level from the disease, number of patients diagnosed with that disease (maybe primary diagnosis), (maybe) their gender, age-group, cure status and ultimate treatment.

PID	NAME	DOB	AGE	ADDRESS	PHONE	SEX	ENCOUNT ERNO	ENCOUNTER TYPE	ENCOUNTER DATE	DOCTOR	DIAGNOSIS	SECONDARY DIAGNOSIS	SYMPTOMS	HOSPITAL	MEDICINE	PATIENT TYPE	DISCHARGE	MEDICAL CONDITION	RACE	EMAIL	REFERRED BY	STATUS
1	Sanjay	01-10-1976	45	MG Rd, New D	11111	123	1	Consultation	05-05-2021	Dr Sikka	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	OP	NA	Critical	White	s@aa.com	Fortis	Admitted
1	Sanjay	01-10-1976	45	MG Rd, New D	11111	123	2	Admission	06-05-2021	Dr Sikka	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	NA	Critical	White	s@aa.com	Fortis	Admitted
1	Sanjay	01-10-1976	45	MG Rd, New D	11111	123	3	Pharmacy	06-05-2021		Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	Yes	Critical	White	s@aa.com	Fortis	Discharged
2	Peter	01-01-1975	46	BG Rd, Bangal	22222	124	1	Admission	07-05-2021	Dr Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	Black	p@pp.com	Apollo	Died
3	Ramesh	15-05-2010	11	BG Rd, Bangal	33333	125	1	Admission	08-05-2021	Dr Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	White	p@pp.com	Apollo	Discharged
3	Ramesh	15-05-2010	11	BG Rd, Bangal	33333	125	2	Radiology	08-05-2021	Dr Rupendra	Covid	HA	Cough, Fever	Fortis, Banglore	CT Scan	IP	Yes	Admitted	White	p@pp.com	Apollo	Discharged
4	Jack	01-10-1976	45	MG Rd, New D	44444	126	1	Consultation	05-05-2021	Dr Sikka	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	OP	NA	Critical	White	j@ii.com	Fortis	Admitted
4	Jack	01-10-1976	45	MG Rd, New D	44444	126	2	Admission	06-05-2021	Dr Sikka	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	NA	Critical	White	j@ii.com	Fortis	Admitted
4	Jack	01-10-1976	45	MG Rd, New D	44444	126	3	Pharmacy	06-05-2021		Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	Yes	Critical	White	j@ii.com	Fortis	Discharged
5	John	01-01-1975	46	BG Rd, Bangal	55555	127	1	Admission	07-05-2021	Dr Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	Black	j@ii.com	Apollo	Died
6	Mahesh	15-05-2010	11	BG Rd, Bangal	66666	128	1	Admission	08-05-2021	Dr Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	White	m@rm.com	Apollo	Discharged
6	Mahesh	15-05-2010	11	BG Rd, Bangal	66666	128	2	Radiology	08-05-2021	Dr Rupendra	Covid	HA	Cough, Fever	Fortis, Banglore	CT Scan	IP	Yes	Admitted	White	m@rm.com	Apollo	Died
7	Sanju	01-10-1976	45	MG Rd, New D	77777	129	1	Consultation	05-05-2021	Dr Sikka	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	OP	NA	Critical	White	s@ss.com	Fortis	Admitted
7	Sanju	01-10-1976	45	MG Rd, New D	77777	129	2	Admission	06-05-2021	Dr Sikka	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	NA	Critical	White	s@ss.com	Fortis	Admitted
7	Sanju	01-10-1976	45	MG Rd, New D	77777	129	3	Pharmacy	06-05-2021		Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	Yes	Critical	White	s@ss.com	Fortis	Discharged
8	Pet	01-01-1975	46	BG Rd, Bangal	88888	130	1	Admission	07-05-2021	Dr Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	Black	p@pp.com	Apollo	Died
9	Isu	15-05-2010	11	BG Rd, Bangal	99999	131	1	Admission	08-05-2021	Dr Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	White	i@i.com	Apollo	Discharged
9	Isu	15-05-2010	11	BG Rd, Bangal	99999	131	2	Radiology	08-05-2021	Dr Rupendra	Covid	HA	Cough, Fever	Fortis, Banglore	CT Scan	IP	Yes	Admitted	White	i@i.com	Apollo	Discharged

FIGURE – SHOWS HOW TO IDENTIFY FUTURE PROOF DATA AND ATTRIBUTES

2. Elimination phase – Eliminate all those entities and attributes those has no significance for any future analysis or research. Eliminate sensitive, future sensitive and controversial data. Make a strategy to capture only meaningful and data to be useful in future, i.e. future proof data and not sensitive data. You can follow an elimination method to be more specific, it helps narrow down easily and quickly. Eliminate all other data that has no relevance in future. For example, in retail domain, how important or how much relevant it will be to capture that "how many pair of shoes a customer bought per order?" Rather we should try to focus that 'which type or color or design of shoes were liked most or in the trend, during early 80s?' or 'how the customer choice is changing every decade and what kind of shoes are in trend now a days?'. This may give the perspective of fashion trend changes over time. Such information can be captured and retained only by capturing the information like shoes type, brand (maybe), period, number of pairs sold etc.

Take an example of EMR data from healthcare domain. We had shown the Future proof data sets those were identified for future proofing. There are other types of data as well like Personal Data, Sensitive Data and Controversial Data.

In EMR example and the figure shown below, Personal, Future proof and Sensitive Data are shown in blue and red color respectively:

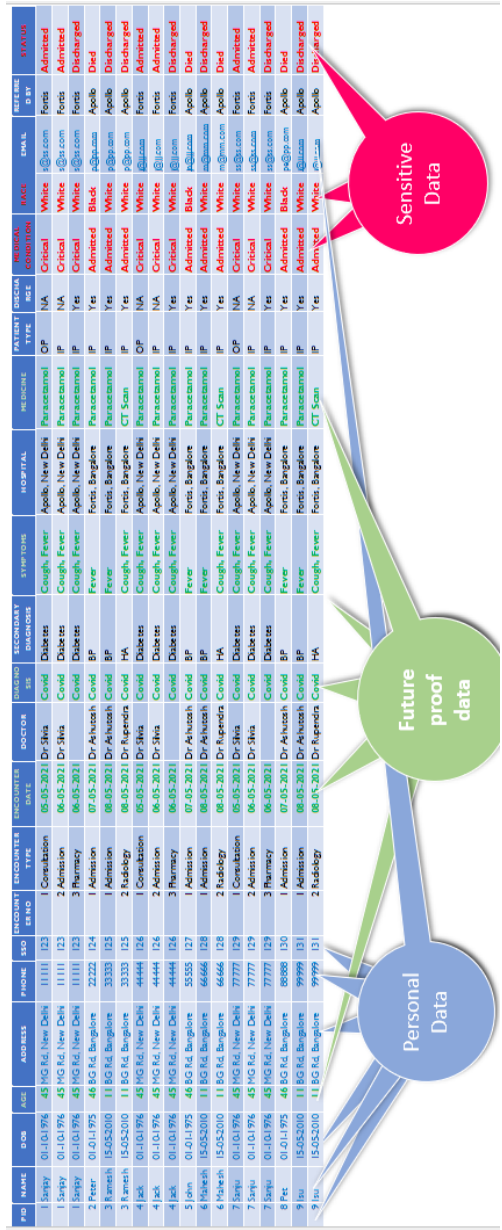


Figure – Shows Personal, Future proof and Sensitive data

From the above example you can understand that few attributes are supposed to be sensitive only if it is supported and served with the personal data. For example, a patient's medical condition and diagnosis can be sensitive information if it is supported and linked to the personal information like PID or SSN. Similarly, the patient's race can come under controversial category of data only if it is linked to the actual patient PID or Name or SSN and/or other personal data, else it has not any controversial exposure. Hence, you should make a note that a data will consist its sensitivity or controversy only if it is linked to the original or relatable personal information like Name, SSN, PID (Patient ID) etc. As far as the attributes that comes under sensitive or controversial category are not relatable to any individual person, place or thing; it will not retain its status of being sensitive or controversial.

3. Future Proofing – By following phase 1 and 2, half of future proofing job is done. You identified only the data that will have relevance in future. You eliminated all sensitive/ controversial and personal data. But still if you have some data that you think, might not be fair to be exposed for any future use, or in case it might be misleading, or might bias any data analysis once exposed for any future business usage or purposes, we should always anonymize it. Information that you need to be careful about can be classified in 2 different categories – (a).

Personal Data – like name, address, medical details, banking details etc and (b). Sensitive Data – like racial or ethnic origin, political opinions, religion, membership of a trade union, health, sex life, criminal activity

These above categories of data might create biased decisions or may create controversies in future. Specially category '(b)' is more sensitive for the political subjects. Hence data anonymization will be a handy tool for the purpose. With data anonymization, you retain the purpose and you don't expose the actual data.

In most of the business scenarios (except few), detailed level or the lowest grain level of data is not required for future use. Lowest cardinality of data is generally required in transactional environment. Hence, we should judiciously decide the data extraction level or the grain level of data while future proofing your data. Decision of the required grain level of data will be next very important aspect while future proofing the data. This will decide what will be the volume of your future proof data every year or every decade. BTW, more the aggregation, less is the chance of personal and sensitive data vulnerability. Data in its lowest grain level are the raw form of records that may incorporate all personal and sensitive data that are more vulnerable in nature if not handled responsibly.

Data Lake or Data warehouse or a Data lakehouse in general have lowest grain level of data, hence we need to be more

sensible and preferably give priority to "Data Future Proofing" features.

PID	NAME	DOB	AGE	ADDRESS	PHONE	ISO	ENCOUNT ER NO	ENCOUNTER TYPE	ENCOUNTER DATE	DOCTOR	DIAGNOSIS	SECONDARY DIAGNOSIS	SYMPTOMS	HOSPITAL	MEDICINE	PATIENT TYPE	DISCHA RGE	MEDICAL CONDITIO N	RACE	EMAIL	REFERRE D BY	STATUS
1	Sanjay	01-10-1976	45	MG Rd, New [	11111	123	1	Consultation	05-05-2021	Dr.Sikta	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	OP	NA	Critical	White	s@xcom	Fortis	Admitted
1	Sanjay	01-10-1976	45	MG Rd, New [	11111	123	2	Admission	06-05-2021	Dr.Sikta	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	NA	Critical	White	s@xcom	Fortis	Admitted
1	Sanjay	01-10-1976	45	MG Rd, New [	11111	123	3	Pharmacy	06-05-2021		Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	Yes	Critical	White	s@xcom	Fortis	Discharged
2	Peter	01-01-1975	46	BG R.d, Bangal	22222	124	1	Admission	07-05-2021	Dr.Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	Black	p@xcom	Apollo	Died
3	Ramesh	15-05-2010	11	BG R.d, Bangal	33333	125	1	Admission	08-05-2021	Dr.Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	White	r@xcom	Apollo	Discharged
3	Ramesh	15-05-2010	11	BG R.d, Bangal	33333	125	2	Radiology	08-05-2021	Dr.Rupendra	Covid	HA	Cough, Fever	Fortis, Banglore	CT Scan	IP	Yes	Admitted	White	r@xcom	Apollo	Discharged
4	Jack	01-10-1976	45	MG Rd, New [	44444	126	1	Consultation	05-05-2021	Dr.Sikta	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	OP	NA	Critical	White	j@xcom	Fortis	Admitted
4	Jack	01-10-1976	45	MG Rd, New [	44444	126	2	Admission	06-05-2021	Dr.Sikta	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	NA	Critical	White	j@xcom	Fortis	Admitted
4	Jack	01-10-1976	45	MG Rd, New [	44444	126	3	Pharmacy	06-05-2021		Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	Yes	Critical	White	j@xcom	Fortis	Discharged
5	John	01-01-1975	46	BG R.d, Bangal	55555	127	1	Admission	07-05-2021	Dr.Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	Black	j@xcom	Apollo	Died
6	Mahesh	15-05-2010	11	BG R.d, Bangal	66666	128	1	Admission	08-05-2021	Dr.Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	White	m@xm.com	Apollo	Discharged
6	Mahesh	15-05-2010	11	BG R.d, Bangal	66666	128	2	Radiology	08-05-2021	Dr.Rupendra	Covid	HA	Cough, Fever	Fortis, Banglore	CT Scan	IP	Yes	Admitted	White	m@xm.com	Apollo	Died
7	Sanju	01-10-1976	45	MG Rd, New [	77777	129	1	Consultation	05-05-2021	Dr.Sikta	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	OP	NA	Critical	White	s@xcom	Fortis	Admitted
7	Sanju	01-10-1976	45	MG Rd, New [	77777	129	2	Admission	06-05-2021	Dr.Sikta	Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	NA	Critical	White	s@xcom	Fortis	Admitted
7	Sanju	01-10-1976	45	MG Rd, New [	77777	129	3	Pharmacy	06-05-2021		Covid	Diabetes	Cough, Fever	Apollo, New Delhi	Paracetamol	IP	Yes	Critical	White	s@xcom	Fortis	Discharged
8	Pet	01-01-1975	46	BG R.d, Bangal	88888	130	1	Admission	07-05-2021	Dr.Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	Black	pe@xcom	Apollo	Died
9	Isu	15-05-2010	11	BG R.d, Bangal	99999	131	1	Admission	08-05-2021	Dr.Ashutosh	Covid	BP	Fever	Fortis, Banglore	Paracetamol	IP	Yes	Admitted	White	i@xcom	Apollo	Discharged
9	Isu	15-05-2010	11	BG R.d, Bangal	99999	131	2	Radiology	08-05-2021	Dr.Rupendra	Covid	HA	Cough, Fever	Fortis, Banglore	CT Scan	IP	Yes	Admitted	White	i@xcom	Apollo	Discharged

FIGURE – SHOWS FUTURE PROOF ATTRIBUTES IN AN EMR DATA SET

In a healthcare data use case. 50 years down the line, one might not be interested in knowing that whether 45 years 'Mr. Sanjay' (name) from 'New Delhi' (location/address) who was a 'Hindu' (religion) had Covid, and if he had, what was his CRT value?... It might not solve any industrial problem and it might not address any business issues at that time. Rather it might trigger some sensitive political issues or might fulfil biased interest. But by any chance if Covid or Covid kind of any such virus re-surfaces again after 50 years, the whole healthcare fraternity might be interested in knowing the figures like - What was the percentage of Covid positive patients for human between 40-50 years of age and what was their average CRT value? What was the overall mortality rate? What was the recovery rates for different age group of people? Which medicine or drug shown best result during its treatment? What was the infection spread ratio between male, female and kids? What was the recovery rates in people with comorbidities?... so on and so forth. And such information requirements and fulfilment will address lots of business problems at that time. It may help the healthcare professionals solve lots healthcare issues. It may save many humans lives. So, it is very much clear and evident that we may not have any reason to store data with lowest grain level. Why should we store such data to be used in future if there is no significance of it in future? Hence the Data Future Proofing. Data Future Proofing is to eliminate unnecessary data, reduce storage

expenses, reduce data vulnerability, reduce data complexity.

Next important thing in this phase of future proofing of the data is, future proof data capture cycle. Once you have decided the grain level and extraction level, you write extraction rules and then based on the capture cycle, you capture your future proof data. Your capture cycle can be either daily, monthly, quarterly or yearly. Mode of capture should be any supported storage mode, depending upon your destination system of storage, like your data warehouse or lake or your data lakehouse supported mode will be your mode of storage.

4. Organization phase – Unlike your conventional organization of data into the destination system and like any special data management organization e.g. MDM or CDM, we propose to have a separate FDM (Future-proof Data Management) layer into your destination data management, in this case into your data warehouse, lake or a lakehouse, or if required a separate data organization can also be proposed for FDM. Please remember that FDM has nothing to do with MDM and CDM design or architecture. Conceptually as MDM and/or CDM helps managing enterprise data in efficient ways and help the business, similarly FDM will helps managing future proof data to be treated specially for the future business benefits in an efficient and strategic ways.

So, this FDM layer should be treated special and should be designed by following all these 5 phases of Data Future Proofing discussed in this section.

FDM is an implementation of enterprise-wide system where business gets their historical information for any future use, from single managed place. A central repository is created and all requests for future proof data are satisfied from that one point.

Creation of FDM system is not very complex. It is a repository of past business facts and figures. Keep the design simple and follow the process of "Data Future Proofing", as discussed in this section. Adhere to all 5 phases of "Data Future Proofing" to shape the FDM.

FIGURE – SHOWS FDM (FUTURE-PROOF DATA MANAGEMENT) LAYER WITH A DATA LAKE

5. Storage phase – Once it comes to storage of FDM (Future-proof Data Management), an open format, generic platform based system, vendor independent infrastructure and a no proprietary format storage is recommended. It is so because, "Data Future Proofing" is for long term that might be passed from generation to generation. Hence a

proprietary or vendor locked format or platform will not meet the overall purpose of Data Future Proofing. Most of the Data lake or lakehouse platforms support open format of storage that is in line with the basic requirements of the FDM.

Next very important point in this phase is to ascertain accessibility of the future proof data sets. Who should have the access to the data, who can have permissions to insert, update and delete any data in FDM should be decided and ascertained in this phase only.

Then next important point in this phase that need to be discussed with emphasis is the availability of FDM. It is the nature of the FDM that it does not fall under 99.99999 availability requirement category. But yes, it should be available for any future use. Future can be even next year of your business. Because this year's data is a past data for you, very next year.

Once you have ascertained accessibility and availability, storage upgradation is the last but not the least important strategy under this storage phase. Make sure the storage upgradation is done for the FDM where you have kept all your future proof data so that it can be accessed seamlessly using the latest storage platform year over year and decades over decades. Remember, "Data Future Proofing" is for eternity or till your business or the subject (like healthcare, biomedical science etc.) exists.

Data future-proofing can be considered as the enterprise's most visionary steps in making sure that the future proofed data is going to help the business in past.

CHAPTER 13

FDM™

FDM™ BEFORE AND/OR AFTER SEMANTIC LAYER

FDM is going to be a new, innovative and futuristic discipline in which the business and technology think together to preserve the past for the future. It is for your business data that is required till eternity. And the cherry on the cake is the significance of an enterprise level semantic layer that will help in every stage(s) before and after phases of the FDM design and development.

As we discussed earlier in the previous chapter under different phases of data future-proofing where we discussed about FDM in Org phase as well as in Storage phase.

Unlike your conventional organization of data into the destination system and like any special data management organization e.g. MDM or CDM, we propose to have a separate FDM layer into your destination data management for the data lake/lakehouse, or if required a separate data organization can also be proposed for FDM. Please remember that FDM has nothing to do with MDM and CDM design or architecture. Conceptually as MDM and/or CDM helps managing enterprise data in efficient ways and

help the business, similarly FDM will helps managing future proof data to be treated specially for the future business benefits in an efficient and strategic ways. MDM is helpful when the enterprise holds more than one copy of the data about a business entity. Similarly, FDM will help business manage their data that is future proof. That doesn't change its significance even in future. That helps the business refer the FDM for their past needs for many types of business analysis including but not limited to the trend analysis, AI and ML based analytics for various decision supports.

So, this FDM layer should be treated special and should be designed by following all those 5 phases of Data Future Proofing discussed in previous chapter.

FDM is an implementation of enterprise-wide system where business gets their historical information for any future use, from single managed place. A central repository is created and all requests for future proof data are satisfied from that one managed source.

Creation of FDM system is not very complex. It is a repository of past business facts and figures. Keep the design simple and follow the process of "Data Future Proofing", as discussed in the chapter named 'Data Future Proofing'. Adhere to all those 5 phases of "Data Future Proofing" to shape the FDM.

FIGURE – SHOWS FDM (FUTURE-PROOF DATA MANAGEMENT) LAYER WITH A DATA LAKE

Once it comes to storage of FDM (Future-proof Data Management), a open format, generic platform based system, vendor independent infrastructure and a no proprietary format storage is recommended. It is so because, "Data Future Proofing" is for long term that might be passed from generation to generation. Hence a proprietary or vendor locked format or platform will not meet the overall purpose of Data Future Proofing. Most of the Data Lake or a lakehouse platform(s) support open format of storage that is in line with the basic requirements of the FDM.

Next very important point in this phase is to ascertain accessibility of the future proof data sets. Who should have the access to the data, who can have permissions to insert, update and delete any data in FDM should be decided and ascertained in this phase only.

Then next important point in this phase that need to be discussed with emphasis is the availability of FDM. It is the nature of the FDM that it does not fall under 99.99999

availability requirement category. But yes, it should be available for any future use. Future can be even next year of your business. Because this year's data will become past data for your business, very next year.

Once you have ascertained accessibility and availability, storage upgradation is the last but not the least important strategy under this storage phase. Make sure the storage upgradation is done for the FDM where you have kept all your future proof data so that it can be accessed seamlessly using the latest storage platform year over year and decades over decades. Remember, "Data Future Proofing" is for eternity or till your business or the subject (like healthcare, biomedical science etc.) exists.

See as we have discussed earlier too that a data Lake is built to house both structured and unstructured data.

These data lakes or the lakehouse might make use of intelligent metadata layers – that act as an intermediary between the unstructured data and the data user in order to classify the data in different categories. By identifying and extracting features from the data, it can effectively be structured, allowing it to be catalogued and indexed just as if it was in the form of data that could be analysed and in a tidy structured.

FDM IN CLOUD ERA

In this cloud era where enterprise prefers to be on private, public or a virtual cloud, due to various valid reasons, we must admit that at the end of the day nothing is free on cloud.

Even if you opt for Glacier, archive or any kind of cold storage or warm storage, it will cost and will cost till you occupy any space on the CSP-Cloud service provider's infra. Even if it is in cents, it can be cent per use or per hour or per quantity (kb/mb/gb/tb or so).

So even if a few 100 TB of enterprise data get accumulated year over year, and it is stored in the cloud and even using a cold storage and even if that cold storage is costing you damn cheap like $0.004 per GB/month and retrieval rate is at $0.04/GB then think of a total effective cost for cheapest mode of the cloud storage. The table below showcases the approximate estimated cost of cloud storage before FDM implementation and after FDM implementation:

Table: Representation of approximate estimated cost of cloud storage before FDM implementation and after FDM implementation

So you can note the figures provided above in the table, once the FDM is implemented in your enterprise, the

storage cost comes down significantly. Almost 90% reduction can be noticed if FDM is implemented in right way and in line with the right guidelines and phases proposed in this book.

THE IMPLEMENTATION STRATEGY

We highly recommend to implement the FDM centrally and enterprise wide for a better result and to avoid any redundancy of data and to avoid any conflicting data to be referred into the FDM layer.

Create a strategy and roadmap to answer these questions to help yourself while designing the FDM layer for your enterprise:

1. What purpose you want to achieve?
2. What all tasks can help you achieve it?
3. What should be the order of those tasks?
4. What benefits will you get out of it and when?
5. How much is it going to cost?

The purpose can be –

- Facilitating the business future decision support

- Research – like market research, healthcare and/or medical research
- Analytics - statistical analysis, trend analysis, future projection, roadmap creation
- Learning – various types of business learnings from past data

Tasks can be like –

- Relevant and future proof Entity and Attribute identification
- Anonymization (if required)
- Ingestion
- Transformation (if needed)
- Data Quality
- Storage and preservation in future proof formats

Order of the above tasks can be like –

Order is situational and completely based on business needs. Transformation might come during ingestion or after ingestion. Data Quality task positioning is also based on the situations, in some cases it can be taken care at source, sometimes at destination.

While we do future proofing of healthcare data, we might not need various attributes there or we might have already filtered out using our earlier discussed attributes by classifying attributes in 3 different categories like – Personal Data, Future-proof data and Sensitive data.

In most of the business cases, while we design the FDM, we eliminate personally identifiable data and sensitive data. And in case we need to maintain those attributes for future use, we prefer and recommend to apply anonymization for data privacy purposes if business allows based on its nature of business and the need. We recommend it due to various reasons. One of them is the risk of exposure of some sensitive data in future and/or the cost over time due to preserving or storing such data and applying all sort of data security over those data. It costs. And it costs damn high over time, every year and year over year. We have discussed various such use cases around the reasons behind such business use cases in our chapter on Data future-proofing.

So it is the architect's call what will come first.

Benefits can be –

- Reduced business liability over historical data
- Reduced storage due to elimination of personal and other sensitive data

- Reduced cost of storage in cloud or on premise
- Reduced cost on data security
- Faster retrieval by keeping and maintaining readily accessible data format

Cost can be –

- One-time design cost
- Storage cost
- Maintenance cost

READILY ACCESSIBLE DATA STORAGE FORMAT

We recommend designing the FDM in a readily accessible data storage format. That is ready to actively participate in the relevant business analytics at any time in future by avoiding vendor lock-in. A format that most of the analytics or visualization tools accept directly without further transformation in the format of the data. Like .csv, parquet and many such file formats are universally accepted by almost all sort of analytics and visualization tools. They are convenient for AI/ML based advanced analytics too.

NO VENDOR LOCK-IN

We discourage using any vendor lock in based product use to store the FDM and FDM related data.

Vendor lock-in is a big NO for FDM and FDM data storage. Avoid getting into the vendor lock-in by using any COTS (commercially of the self) product or other means. The FDM being the reference data for future usage, any such kind of lock-in can be fatal and costly. You never know which product will go out of order or get obsolete. We don't want to put the FDM in trouble to be maintained in future. FDM should be maintenance free.

AUTO PILOT DESIGN

FDM need to be designed in a way that is self-contained and maintained. Maintenance overhead might be acceptable for current business data, due to its transactional nature. FDM and maintenance of data in FDM will bring in more cost. It can be a show stopper for the business and can have further challenge of budget and funding. We don't want any such hurdle that hampers the accumulation and ingestion of FDM. Any such break or unwanted interval in FDM data will spoil the purpose of FDM.

Auto pilot means, the sources and targets are pre-defined. Extraction level and transformation logics are pre-written.

Jobs and trigger points are pre-set. No or minimal manual interventions.

AVOID MULTIPLE VERSIONS

Avoid multiple version of the same type of data or information in the FDM till it is obvious and mandatory to have more than one versions of similar attribute, or set of attributes. In case there is obvious business reasons for any attribute or some set of attribute(s) to be repeated or having more than one version, we recommend that all the versions are handled together while it comes to the associated data operations.

FDM AS BUSINESS MANDATE

FDM should be issued as the business mandate for its successful implementations and executions. It should be implemented as religion and culture in the organization rather than an adhoc IT initiative that goes off after few exercises or iterations.

We envision and wish that it becomes statutory instructions and mandates for every organizations/enterprise/corporations with futuristic data,

for the betterment and ease of the future of the business and associated research and analysis.

And to reach to that level, we must convey and discuss the benefits of FDM to the decision makers. We should openly discuss FDM's financial as well as business benefits and its futuristic nature.

We should clearly tell them, that if we don't implement FDM, how the unnecessary and avoidable cost overheads can impact the enterprise bottom line and how we have been or we will end up by wasting funds on large volume data storage for any future use.

One of my friends questioned that why FDM should become statutory? So let us answer – Think of a situation – one key stake holder starts the FDM initiative throughout the enterprise today, and after a few years, another/new executive stake holder joins who doesn't know more about it or didn't show his interest due to her/his own reasons or preferences, and s/he prefers not to pursue with FDM. You never know how far sighted or sort sighted the new leader is.

Hence, we can't afford or the business can't afford to let the FDM purpose get diluted due to the management change or other administrative or policy reasons. It should not fall at the mercy of the executive, may be due to the limited capability or so of the executive or the decision maker. So if

any such thing of this sort happens, FDM will not cater to the enterprise future data needs. And as we have discussed so far, that can cost fortune to the business. Because we ultimately end up by storing far more volume of data for the future than in is actually and practically required. And the difference in size in not small it is many times or many fold as we just discussed in this chapter only.

WHY FDM

FDM becomes necessary in this age where organizational growth is not limited to organic growth. Now a days organizations/ enterprise/ corporations are not only growing organically. Inorganic growth is quite rampant. In fact, such model of growth is working wonderfully for many enterprises who want faster growth and easy market coverage. The moment it comes to inorganic growth, following factors becomes the root cause for data future-proofing –

- Business unit bifurcation or separation
- Product line segmentation
- Merger
- Acquisitions
- Change in management

- Change in policies
- Change in business priorities so on and forth.

All these above factors can impact your data accumulation for future and it becomes trick to manage such data considering these factors that has potential to impact the process directly or indirectly.

Hence a FDM layer across enterprise and its mandate can address and mitigate such challenges due to above mentioned factor.

FORESTRIM TECHNOLOGY

ATSCALE

ABOUT THE AUTHORS

Bill Inmon

Bill is the father of data warehousing. Bill has written many books, published in 9 languages and has sold over 1,000,000 books worldwide. Bill developed textual disambiguation technology – textual ETL - at Forest Rim Technology. Textual ETL reads raw text and turns it into a data base. Bill lives in Denver with his wife and his Scotty dog, Jeb. And Bill knows that it is Jeb who runs the house.

Dave Mariani

Dave is the founder of AtScale and is the Chief Technology Officer. Prior to AtScale, he ran engineering and data at Klout and Yahoo! where he built the world's largest multi-dimensional cube.

David Rapien

David Rapien is a system architect and information flow analyst. As a founder of D&P Software, Schedule DR LLC, and Panther Productions, David was the mind behind Sports Scheduling Software, a global information systems solution that has been providing a scheduling engine for organizations worldwide since 1991.

Ranjeet Srivastava

Ranjeet is known for his contribution to enterprise architecture, product engineering and data engineering over the last 22 years. He has created a few IP and Trademark in Data management and Microservices design Patterns. Published author of 3 books in 3 languages. As Chief Architect and Vice President of Coforge, he helps his customers to exceed their business expectations in many critical, large scale, complex and multi-dimensional, wide prospect business problems.

ISBN 978-9-35-680295-7

US$ 49.95

9 789356 802957 >