The Business Case for Using Unstructured Text Analytics on IBM Power Systems for Critical Decision Making

Stephen Markham, Ph.D.
Michael Kowolenko, Ph.D.
Poole College of Management
North Carolina State University

It's the question, not the technology.
--Michael Kowolenko

Although big data promises to dramatically alter the business environment, technology is only a decision enabler. Many firms apply structured approaches to big data for routine operational decisions. However, making un-programmed, critical and strategic decisions usually involves unstructured data. This paper describes the business value of using big data techniques to gather and analyze unstructured data. This approach uses critical thinking embedded in a process along with advanced servers and software to convert big data into business value. The process and results open new opportunities that require adaptive structures, culture and expertise to fulfill the promise of big data.

This article has three purposes. The first is to explain the difference in how structured and unstructured data are used in decision making. The second is to give examples of how companies realized business value using unstructured data. And the third is to show how companies can realize similar business value using unstructured data and why picking the right servers and software can make a difference.

1. The differences between how structured and unstructured data are used in decision-making

Unstructured text comprises about 80% of the data available today. Firms that use only structured data miss benefitting from the majority of available information. Unstructured data includes all the text contained in government reports from the SEC, NIH, NSF, DOE, as well as all academic research, business and financial analyst's reports, consultant research results and many other sources. Unstructured text is also found in a myriad of social media outlets such as Facebook, blogs, customer complaint logs, and Twitter, as well as news transcripts, the popular press, specialty magazines and many other outlets.

Unstructured data reveals customer needs, competitor actions, emerging trends, and other individual pieces of information necessary to make critical business decisions. Structured approaches to big data gather and aggregate rows and columns of numbers to inform decision makers. Large sets of numbers are analyzed using advanced statistical techniques to reveal valuable patterns in the data. These techniques allow decision makers to see what has happened or what is happening in real time. Analyzing data that can be added, subtracted, multiplied and divided relies on a structured approach to aggregate the data and are essential for making operational decisions such as pricing, distribution, and inventory.

Unstructured approaches on the other hand seek to isolate critical pieces of information. For example, unstructured big data finds announcements that a competitor is building a new facility, or that a customer is expanding operations. This will give decisions-makers time to react – before the structured data ultimately reveals a decrease in their sales revenue. To reliably find a "needle-in-the-haystack," one must use big data to gather vast amounts of unstructured text and use specialized programs to look for specific pieces of information across tens of millions of documents with the unflinching eye of a computer.

1.1. Critical Thinking Drives the Use of Big Data

The ability to make data-driven decisions assumes that people know what questions to ask, but our experience has shown that this is not always the case. Many organizations lack processes to apply critical thinking. In each of the industry-sponsored projects conducted by the Center for Innovation Management Studies (CIMS)[1], companies ranging from startups to members of the Fortune 500 struggle with generating strategic inquiries.

[1] CIMS is a graduated IUCRC. Formed in 1984, it is the only NSF-sponsored research center dedicated to investigating the organizational and managerial effects on innovation.

Thinking must drive the use of big data – big data cannot drive thinking. Critical thinking is the basis for finding the data sources and tools to recognize the underlying meaning in unstructured text. For example, the statement, “Our company needs to examine social media for sentiment analysis” can be broken down into a series of sub questions such as:

- Sentiment as it relates to our product
- Sentiment as it relates to the competition
- Customer likes and dislikes of the product class in question
- How our company’s new products address customer dislikes
- What the competition is developing that may address customer dislikes

Drawing on the principles of critical thinking, we helped develop a process to gain business value from unstructured text. In practice, the application of critical thinking to the development of the business decision-making process has been successful in numerous companies.

1.2. The Process for Using Critical Thinking to Analyze Unstructured Data

The process (Figure 1) requires the participation of a cross-disciplinary team of affected stakeholders from the start and has proven successful in a number of industries. This process uses critical thinking techniques to: 1) define the problem statement and form specific questions to inquire about; 2) identify sources of information; 3) identify the search terms and define the relationships between the terms (called rules); 4) apply big data technology to the terms and rules to gather, store and analyze massive amounts of data collected from external and internal sources; 5) assess the data for sufficiency, applicability and veracity; and finally; 6) evaluate the evidence that either supports or refutes the assumptions or conditions required to make the decision, once the filters have been applied to the data. It is important to emphasize that this is an iterative process. The act of collecting and filtering the data often results in new insight that requires further investigation. This requires the team to overcome individual and group bias when faced with new evidence.

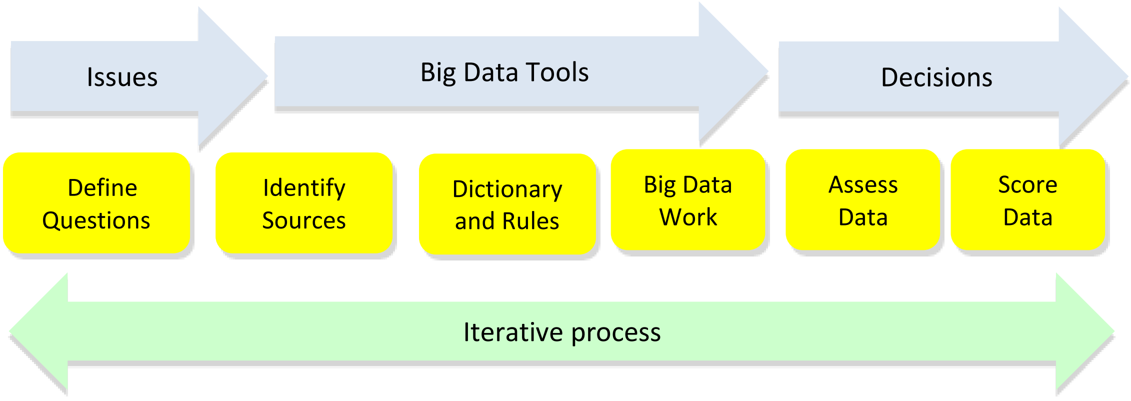

Figure 1. Process for unstructured data-enabled decision making

2. Examples of how companies realized business value using unstructured data

Just five of the dozens of examples of business questions successfully addressed with this process are outlined in the table below. Each example represents a different type of the most commonly asked questions. Another takeaway from these examples is that the decisions are not based on quantitative analysis of structured data but rather on the interpretation of the facts derived from analyzing unstructured data. The selection of the questions, use of terms and the creation of dictionaries and rules allows one to configure the software to answer a wide variety of critical decisions.

Industry: Company	Question	Sources	Outcome
Temporary workforce: Kelly Services	Develop new service offerings in healthcare staffing	SEC, URLs, trade journals, professional journals, insurance providers	Decision to move forward in an unexpected healthcare domain
Industrial Gases: Air Products	Find new customers and market opportunities	SEC, news feeds, industry publications, building permits	Identification of a new customer planning to build new facilities
University: NC State	Identify commercial partners for new technologies	SEC, URLs, industry publications	Potential partners identified for collaborations
Clinical Research Organization: PRA International	Provide business intelligence for new clinical trials	Clintrials, PubMed	Identify new physicians/hospitals with expertise in areas of clinical trials
Non-Governmental Organization: Clinton Health Care Access Initiative	Find the fit between new technologies and market opportunities for disease diagnostics	Clintrials, PubMed VC firms	Identification of research labs active in cutting edge diagnostic research

Table 2. Industry Examples

2.1. Kelly Services: Develop new staffing service offerings in healthcare

Kelly Services is a leader in temporary staffing and has expanded its offerings into a full-service staffing solutions company. They wanted to explore whether offering temporary staffing services to the healthcare industry would be viable. Working with internal and external subject matter experts, unstructured data about the need for this new service and competition and regulations not only revealed numerous opportunities but also provide guidance about how to proceed.

The research showed a great need for nurses to work remotely providing medical assessments and advice to a variety of clients including hospitals, nursing care facilities, insurance companies and self-insuring organizations. It was also found that there is a nursing shortage, and the need was not being met on a local or regional basis. Combining Kelly's national staffing capabilities with regional needs allows Kelly to uniquely meet market demands not yet discovered by their competitors.

It was also found that the ability for health care providers to get paid for these types of telemedicine services varies greatly by state. Unstructured text analytics was used to find the exact regulations in every state. This provided Kelly with the knowledge to enter the right states with favorable regulations with the right services. This new service is now one of two major innovative initiatives in Kelly's service portfolio.

2.2. Air Products and Chemicals: Find new customers and market opportunities

Air Products offers industrial gases and specialty chemicals in a dozen business verticals. They wanted to identify potential new customers and provide their sales force with specific information about these customers. In metal processing, it is important to identify customers early, even before production starts. Metal processing plants are designed with specific gas feeds, so it is important for gas suppliers to know as soon as possible in order to meet new customer's needs.

The data gathering process started by "reading" all of the SEC data using Natural Language Processing rules to identify companies that were planning capital expenditures on new plants and equipment. The crawls also read a variety of public press outlets in an attempt to find companies mentioning new business operations. Additionally, the search included reading foreign newspapers (representing 22 languages) in countries that produce metal processing equipment for announcements about new sales of equipment. Finally, the engines read all building permits in the U.S. to identify metal processing companies that were building new plants.

The research uncovered a metal processing company building a new plant in a region served by Air Products that ordered a specific type of equipment and announced a specific number of new jobs. Not only did Air Products identify a new prospect, they also learned what type of equipment they were going to use (thus defining the type of gas needed) and how much metal the company planned on processing (indicating the volume of gas needed) -- all eighteen months before construction started. This gave Air Products a distinct competitive lead in approaching this new customer. Additional analysis also revealed a number of existing customers expanding their operations, information that was not previously known.

2.3 Pentair: Find customers for new product

Pentair supplies equipment for applying chemical treatment and aeration for impaired water bodies including lakes, ponds, and water collection areas. It recently developed a new product that can automatically disperse chemicals uniformly over a lake or pond. State and local governments and their agencies determine whether a body of water is impaired. Owners responsible for remediation vary from counties and municipalities to homeowner associations and golf courses, etc. These exist in all geographic regions, but there is a concentration in certain states.

The business challenge: Find potential sales opportunities for a new water treatment system. Specific Questions:

- In which states should the effort be concentrated?
- Who are the people in the local agencies that are involved in remediation determination?
- What water bodies have been deemed unsuitable for water quality?
- Who is the responsible owner for that water body?

Using several key word lists and Natural Language Processing (NLP) rules, a model was developed to extract information of interest from a dataset. A dictionary of state names was used in combination with impaired water body terms to determine where to focus by region. By searching for regulations in these states, agency names and contact information could be found. Names of specific targets could be found by using the agency lists of impaired water bodies. Additionally, one of the rules created for the model included the “company finder.” This NLP rule was also generated to extract owner names based on patterns of naming conventions. By using both lists and rules, a model was developed and applied to query the indexed dataset.

This project was completed in approximately three months, which included web crawling and indexing thousands of documents, setting up filtering models to select relevant information to answer specific questions, and analyzing results. The project team worked closely with client subject matter experts to select information sources, review findings and improve the information filtering models through several iterations. Names of impaired water bodies, their responsible owners, and contact names were found and are being pursued by the client.

2.4 PRA: Provide business intelligence for new clinical trials

PRA Internal is a clinical research organization that conducts drug trials for pharmaceutical companies. It wanted to gain better insight into what other clinical research organizations were doing, and to better predict what their customers might want in the near future so they could better serve them.

Unstructured text analytics was used to read four million data files containing over 75 million web pages. These unstructured sources of information were then linked to structured information contained in www.clinicaltrails.gov and company specific data files to provide a much more in-depth picture of what was happening in their industry.

In one project, PRA wanted to know what was happening in myeloma research. Company leaders wanted to know a variety of things not commonly available, such as the reason for previous myeloma trial failures and what companies might be working on myeloma activities that might not show up in government and industry reports. This research uncovered several relevant findings. The majority of trial failures were due to lack of enrollment, but other important reasons emerged. Interestingly, seven companies were found using the same gene targets for a variety of other indications such as asthma, breast and lung cancer, Parkinson’s disease, Alzheimer’s, sickle cell anemia, and blindness. Also found were the names of the managers of these projects, which enabled PRA to form relationships to better advance their own research.

2.5. Clinton Healthcare Access Initiative (CHAI): Find a fit between new technologies and market opportunities for disease diagnostics

CHAI wanted to assess if its healthcare investments in diagnostics were effective. Managers were supporting a large number of initiatives and wanted to maximize the effect they were having.

Big data was used to assess not just the effectiveness of the new diagnostics but also what impact the new diagnostic tools might have. By assessing unstructured data from thousands of articles and reports it was found that the impact of developing new diagnostics for new diseases would have much less impact than making existing diagnostics for more common diseases. This prompted a reassessment of CHAI’s strategy and realignment of its resources to favorably impact more people.

Conclusion

These examples demonstrate the business value of using unstructured text analytics to find and assess new business opportunities, find and qualify new customers, provide vastly improved and more detailed business intelligence and to make strategic resource allocation decisions. Unstructured data can be used for many other applications where decisions require knowing specific information that cannot be added, subtracted, multiplied or divided.

3. How companies can realize business value using unstructured data

The advantage of using unstructured big data is that the tools and techniques have now been refined to allow people with critical thinking skills to answer important business questions without being software programmers or statisticians. The process can seem complicated at first and the software quite specialized, but it is no longer in the exclusive purview of data scientists.

An unstructured big data project requires teaming across the IT department, to support the software and help gather and store the data, and among the statistical analysts, to run the big data process. But by far the most important ingredient for successful unstructured data analysis is the critical thinking ability of the company’s subject matter experts (SMEs). This is an interdisciplinary decision-making process capability -- it is not a one-time event. Therefore, the company must commit the SME resources necessary to run the process and make the decisions and empower senior management with the authority to act on them.

This democratization of big data for use by business people is not just a benefit or even an aim of these tools; rather, it is a necessity. Business content must drive the questions, terms, sources and rules necessary to realize business value.

It is critical to choose the right software to conduct unstructured text analytics. There are many commercial and open-source programs that can be used. If you do not choose a program that has a graphical user interface and the ability to seamlessly interact with large data sets, you will need data scientists just to run the technical part of the software.

IBM Content Analytics Studio (ICA) software uniquely meets all the criteria to allow business content people to engage in big data analytics of unstructured data for themselves. With a few days of training, most business people learn to use ICA well enough to apply it on a continuing basis in their area. Thus, big data can be used routinely by a wide variety of people rather than on a special-project basis.

It is also important to use the right server platform for big data. Although we seek to isolate highly specific information to make decisions rather than aggregate large data sets, we can only identify that critical piece of information if we can gather it and process it in a timely manner. Consequently, choosing the right server platform can be a crucial aspect of deploying a successful unstructured big data project.

Some companies just getting started decide to use x86 because they already have the servers installed and the big data software, including ICA, runs on x86. There are profound limitations, however, to this approach. While x86 servers may suffice for small demonstrations, the low reliability of the platform hinders adoption of big data. You may not be able to actually demonstrate the fully value of big data on x86 servers. We use both x86 and IBM Power Systems servers for processing big data. There is no doubt that the Power servers are far superior. The x86 servers crash so often that we chose not to use them with our clients.

Power System server, with the POWER8 processor-based technology, was designed for big data to run more concurrent queries in parallel faster, across multiple cores with more threads per core. It also has increased memory bandwidth and faster IO to ingest, move and access data faster. This allows companies to run these data-hungry analytics queries faster.

3.1. Gaining a competitive advantage with big data

Unstructured data analysis can be the source of great business value if the appropriate processes, tools, skills and structure are implemented together. Frustration and disappointment await you if you do not implement these elements together. The cost of failure is to fall behind competitors that are successful at implementing big data.

The key to a successful implementation is to establish a simple process for business content experts and decision makers to use on a regular basis. This means that both the inputs and outputs of the decision-making process must be appropriately resourced and roles and responsibilities established. You must establish a reporting relationship with decision makers to ensure accountability. Performance and outcome expectations must be created for what your company wants from big data. The right infrastructure must be used to ensure the required levels of performance and reliability and met. A proper implementation promises business value and competitive advantage by arming decision-makers with more targeted information.