AIIM Market Intelligence

Delivering the priorities and opinions of AIIM's 65,000 community

aiim Industry Watch

Big Data

- extracting value from your digital landfills

Send to a friend

aiim®
The Global Community of Information Professionals
aiim.org | 301.587.8202

About the Research

As the non-profit association dedicated to nurturing, growing and supporting the Information Management community, AIIM is proud to provide this research at no charge. In this way, the entire community can leverage the education, thought leadership and direction provided by our work. We would like this research to be as widely distributed as possible. Feel free to use this research in presentations and publications with the attribution – "© AIIM 2012, www.aiim.org"

Rather than redistribute a copy of this report to your colleagues, we would prefer that you direct them to www.aiim.org/research for a free download of their own.

Our ability to deliver such high-quality research is partially made possible by our underwriting companies, without whom we would have to return to a paid subscription model. For that, we hope you will join us in thanking our underwriters, who are:

Actuate - Xenos Group
95 Mural Street, Ste. 201,
Richmond Hill, Ontario L4B 3G2
Phone: 905.763.5096
Fax: 905.709.0123
Email: lgamboa@actuate.com
Twitter: @xenos_groupWeb:
www.xenos.com

Attivio
275 Grove St.,
Newton, MA 02466
Phone: 857.226.5040
Sales: 857.226.5235
Email: karnowitz@attivio.com
Web: www.attivio.com

EMC Corporation
176 South Street
Hopkinton MA 01748
Phone: 800.222.3622
or 508.435.1000
Fax: 508.497.6904
Email: softwaresales@emc.com
Web: www.emc.com

IBM
3565 Harbor Blvd.,
Costa Mesa, CA 92626
Phone: +1 800.345.3638
Web: www.ibm.com/software/data/ecm

Process Used and Survey Demographics

While we appreciate the support of these sponsors, we also greatly value our objectivity and independence as a non-profit industry association. The results of the survey and the market commentary made in this report are independent of any bias from the vendor community.

The survey was taken using a web-based tool by 403 individual members of the AIIM community between March 30th, and April 25th, 2012. Invitations to take the survey were sent via e-mail to a selection of the 65,000 AIIM community members.

Survey demographics can be found in Appendix A. Graphs throughout the report exclude responses from organizations with less than 10 employees, and suppliers of ECM products or services, taking the number of respondents to 345.

About AIIM

AIIM has been an advocate and supporter of information professionals for nearly 70 years. The association mission is to ensure that information professionals understand the current and future challenges of managing information assets in an era of social, mobile, cloud and big data. AIIM builds on a strong heritage of research and member service. Today, AIIM is a global, non-profit organization that provides independent research, education and certification programs to information professionals. AIIM represents the entire information management community: practitioners, technology suppliers, integrators and consultants.

About the Author

Doug Miles is head of the AIIM Market Intelligence Division. He has over 25 years' experience of working with users and vendors across a broad spectrum of IT applications. He was an early pioneer of document management systems for business and engineering applications, and has produced many AIIM survey reports on issues and drivers for Capture, ECM, Records Management, SharePoint, Mobile, Cloud and Social Business. Doug has also worked closely with other enterprise-level IT systems such as ERP, BI and CRM. Doug has an MSc in Communications Engineering and is a member of the IET in the UK.

Table of Contents

Introduction

Use cases for the current applications around big data were portrayed as early as 2002 in the Tom Cruise movie, Minority Report. The film hinges on the ability of 3 psychics called "pre-cogs" to digest data of previous crimes and behavior across the city in order to predict potential murders just before they happen. Filmgoers may recall how cameras on the electronic billboards around the city were also able to recognize passers-by, and display personalized posters linked to the observer's preferences and buying habits. Today's reality is that event prediction and customer segmentation are two of the many possible applications for big data analytics; facial recognition and electronic posters are relatively commonplace.

Perhaps it is the almost sci-fi appeal of many big data applications that has made it such a current favorite with industry-watchers and thought-leaders. Much has been said about its potential applications, but there is much confusion as to what it actually means – even among the analysts and the product vendors.

Much of the recent interest has been triggered by some interesting technical developments in how to store and query very large datasets (greater than 1 petabyte), but we are more concerned here with the business insights that can be gained from a number of new (and some old) analysis techniques that can be run against unstructured content repositories, generally containing text or rich media files. We are particularly interested in applications where these un-structured datasets can be linked to structured datasets holding transactional or numerical data. We are thinking more about what happens when "big content meets big data".

In this report, perhaps as something of a reality-check, we record how users see the potential benefits of big data analytics. We explore some of the more practical and popular applications, and we also look at the issues that are holding users back, including skills shortages, product immaturity and implementation uncertainties. We have also compiled a glossary of terms – see Appendix 2.

Key Findings

Overall Drivers and Benefits

- **70% of respondents can envisage a killer application for big data that would be "very useful" or "spectacular" (18%) for their business.** The majority chose not to disclose what that application would be.
- **61% would find it "very useful" to link structured and unstructured datasets.** Currently, unified data access across content repositories is a struggle for most respondents.
- **56% would find it "very valuable" or "hugely valuable" (18%) to be able to carry out sophisticated analytics on unstructured content.** Particularly pattern analysis, keyword correlation, incident prediction and fraud prevention.
- **For 70% it's "harder" or "much harder" to research information held on their own internal systems compared to the Web.** A lack of standardized analysis tools is given as a significant issue for improving business intelligence.

Data Types

- **Detecting trends and patterns, and content categorization/migration** show the strongest level of demand for unstructured analysis techniques.
- **Analyzing comment fields from forms is the most popular potential application (68%).** Analyzing help desk or CRM logs is currently the most popular. Correlating key phrases from incident reports and case notes would also be widely useful.
- **9% of organizations are already making use of publically available or open data sets to extract longer-term business intelligence or solve problems, and 42% are keen to do so.** There is a similar willingness to link to external subscription data sets.
- **9% of organizations have seen value in analyzing their output print-streams, and 24% would like to do so.** A further 33% had not previously thought about the possibility of mining printouts and statements for financial trends.

- **Monitoring system logs, help desk conversations and web clicks are the most popular real-time or time-critical applications now,** with a future wish to monitor and analyse *all* incoming customer communications, as well as news channels and external social streams.

Issues

- **Many organizations (26%) are still struggling to organize their content. Many (30%) have poor reporting and BI capabilities.** Both of these factors will affect priorities and the ability to roll out big data projects.
- **There is some cross-over of priorities between search and analytics, but most users see equal value in both.** 55% of organizations currently have neither, only 8% have both.
- **Security within search and analytics is a major concern for 64%, including 19% who say it is a potential show-stopper.** This is also given as a strong requirement for any potential product purchase.
- **Lack of in-house expertise is the most prominent issue facing users, followed by cost, and then the difficulty of connecting their datasets.** There is an understandable enthusiasm amongst respondents to be better trained on big data and thereby enhance their value.
- **The terms "content analytics", "unified data access", "semantic analysis" and "sentiment analysis" are generally understood by around half of the survey respondents.** Specific big data technologies like Hadoop, NoSQL and Map Reduce are unfamiliar for three quarters of those responding.

Deployment

- **88% would be inclined to create big datasets for analysis on-premise or in a private cloud** rather than as a SaaS application (6%) or hosted on a public cloud infrastructure (3%).
- **There is no consensus as to how to source a big data capability. In-house development, open source, custom development, best-of-breed, and analytics suites are all contenders.** In fact, the majority are likely to use a mix of all of these.
- **7% of the respondents have already invested in big data or big content analytics tools, and this is set to double in the next 12 months.** Most respondents (48%) are looking at a 2-3 year timeframe.

Big data analytics versus search and conventional BI

One of the confusions amongst the user base is how "big data" analytics differs from normal reporting or BI (Business Intelligence), particularly in the context of structured data or transactional databases. For these extensions of conventional reporting, complexity and scope play a part, as well as the size of the data to be analyzed. The ability to scale the data yet produce query results within reasonable times may lead on from technical differences between conventional data warehouses, and some of the more recent database technologies.

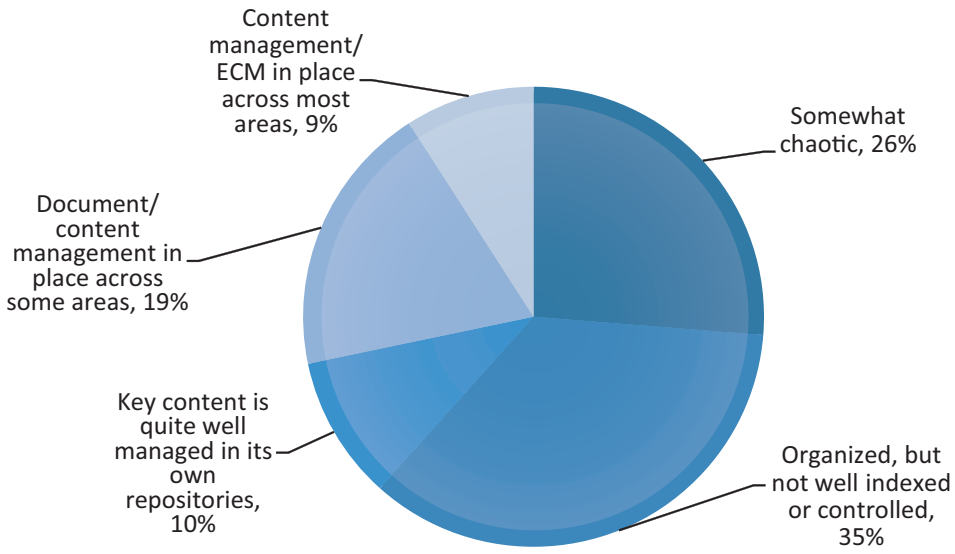
In the unstructured world, the confusion is more between search, particularly enterprise or federated search, and content analytics. Even when extended across multiple content sources, the search process can only bring back matches (or near matches) to the search criteria – perhaps a word or word-string, perhaps with a simple count. Analytics can count, sum, group, segment, trend and infer the results by context, author, document type, date, etc. Both can produce useful business benefits, but whilst search will produce a distillation of knowledge and past history, content analytics is more likely to provide deeper insight for business management. Although we may use the term "big content", the actual size of the dataset is probably less of a consideration in un-structured applications. We will return to the comparison of search products and analytics later in the report.

Content Management

Although not necessarily a pre-requisite, a degree of content organization will certainly make big data analysis somewhat more straightforward, and in terms of priorities, many organizations are looking to address this before embarking on big data projects. Even though one would expect survey respondents

from the AIIM community to be well aware of the need for ECM and information management, many of their organizations are well behind the curve for adoption. Only around a third feel their content is reasonably well managed for any degree of universal access, and 26% are still somewhat chaotic.

Figure 1: How would you best describe management of the unstructured content in your organization? (N=339)



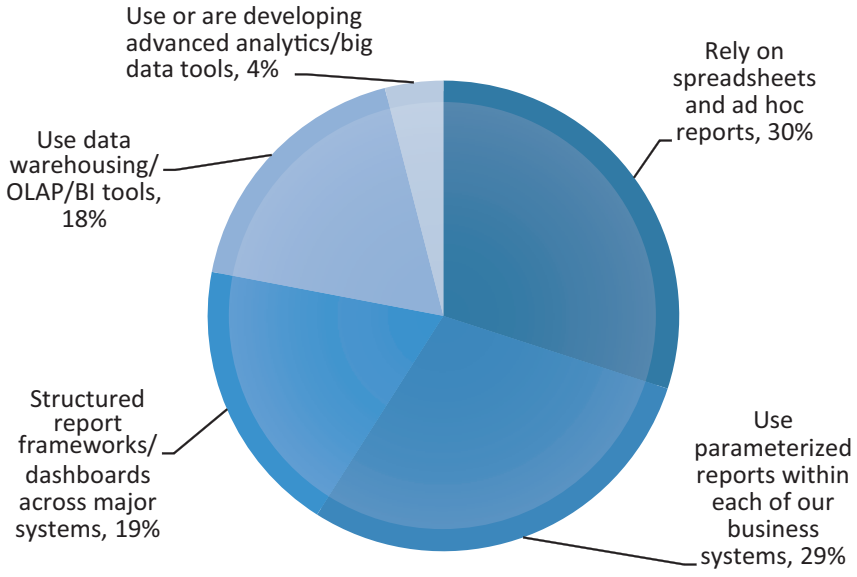
Search

As we might expect, this ongoing lack of content management is also reflected in the levels of search available across different content repositories. Only 20% have enterprise search or unified search capability across departmental content. Of these, 7% have extended search across the whole enterprise.

BI

When it comes to existing BI and analytics capability, we see a similar situation, with 30% relying on basic spreadsheets and reports. However, in this case 41% currently have a reasonable reporting and BI framework, with 18% using data warehouse techniques and 4% already embarking on advanced analytics or big data tools.

Figure 2: How would you describe your data analysis, reporting and BI capability across structured datasets - ERP, Financial, Line of Business systems, etc.? (Pick highest capability) (N=331)



So as we start our investigation, we can already see that most organizations are relatively immature in their existing content management and data analysis capabilities, and this will inevitably have a bearing on their ability and priority to adopt the more advanced big data and big content techniques.

Many organizations will struggle to deploy big data applications until they improve their current levels of information management and reduce content chaos.

Drivers for Big Data

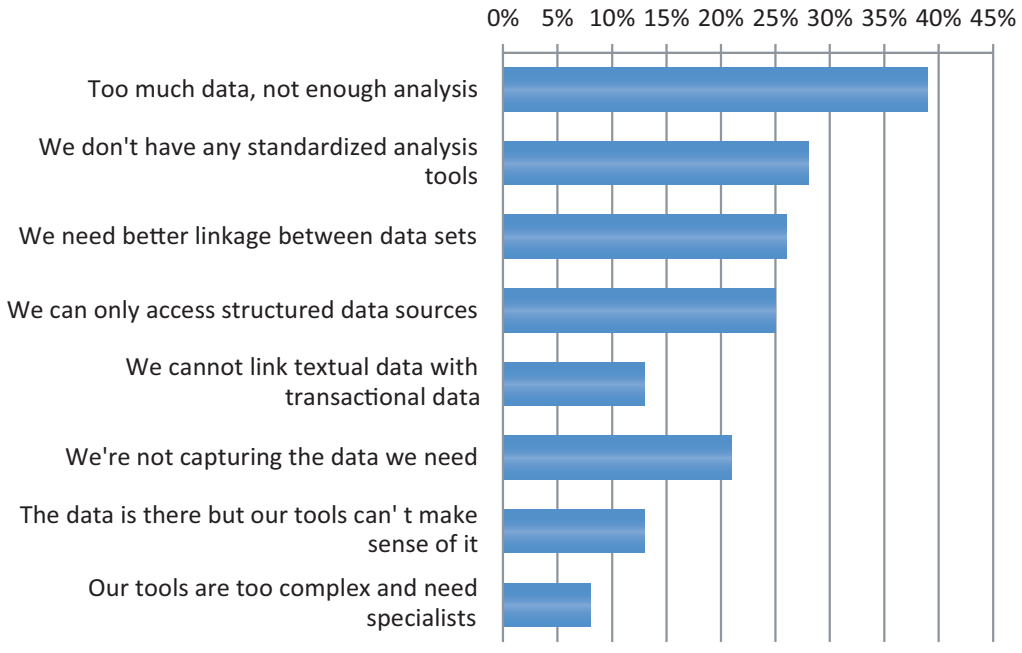
The key drivers for big data applications are, of course, to better exploit internal and external data in order to improve the running of the business, to avoid disruption, and to gain a competitive edge over the competition. These are very high-level targets, so we need to move down a step to consider what specific business insights might be possible. Listed below are some generic applications. At the next level down, we would find insights that are quite specific to a given industry sector, or an individual business within that sector.

- Modeling risk (or reward)
- Failure or incident prediction
- Threat analysis
- Non-compliance detection
- Diagnostic and forensic
- Analyzing customer churn
- Web targeting
- Sales transaction and customer profiling
- Content categorization, migration and deletion
- Public perception and sentiment monitoring
- Geographic, demographic and location tracking
- Product and service improvement
- Case-file mining and aggregation
- Expertise/candidate profiling
- Network monitoring and continuity
- Exploitation of public datasets

Data Exploitation

As we can see from Figure 3, many organizations have plenty of data – even if it is not always the right kind of data - but they struggle to put it to any constructive use, either because they don't have any analysis tools, or the tools they have aren't sufficiently capable or accessible. In the middle of the table, we see another root cause in that organizations are struggling to link together the datasets they do have, or are unable to make any connection to their unstructured data sources. This raises the technical issue of unified data access that we will keep coming back to in this report.

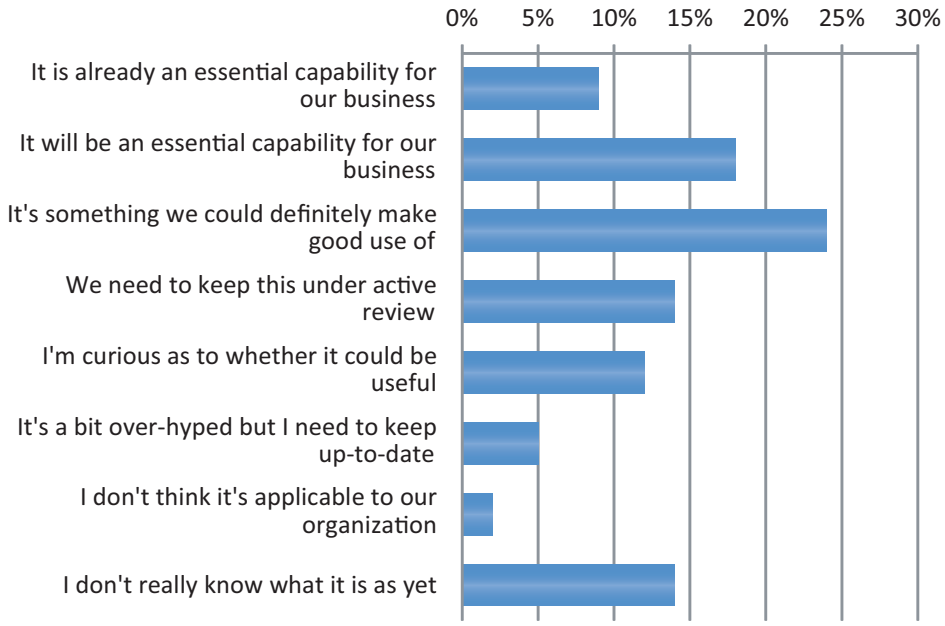
Figure 3: Which two of the following would you choose to describe your biggest data analysis and BI issues? (N=330)



The result of these problems is reflected in the "It's easier to do things on the web than in the office" effect. For this survey, we particularly asked about *researching* information, rather than just *searching* for it, but even so, for 70% it's "harder" or "much harder" (23%) to analyze content on their own servers than it is on the web. This is largely unchanged from a similar AIIM survey 2 years ago, and perhaps reflects the fact that the web has moved on apace in that time, with many users making the comparison with products such as Google Analytics and other web-presence analysis tools.

Only 14% of our respondents are prepared to admit that they don't really know what big data is as yet, despite electing to take the survey. We would expect that number to be higher in the IT population at large, although there can be few who have yet to come across the term. Having said that, 51% of the respondents are keen to make some use of big data with 23% feeling it will become an "essential capability" for their business – a strong endorsement of the perceived benefits.

Figure 4: How would you best describe your current interest in "Big Data"? *(N=339)*



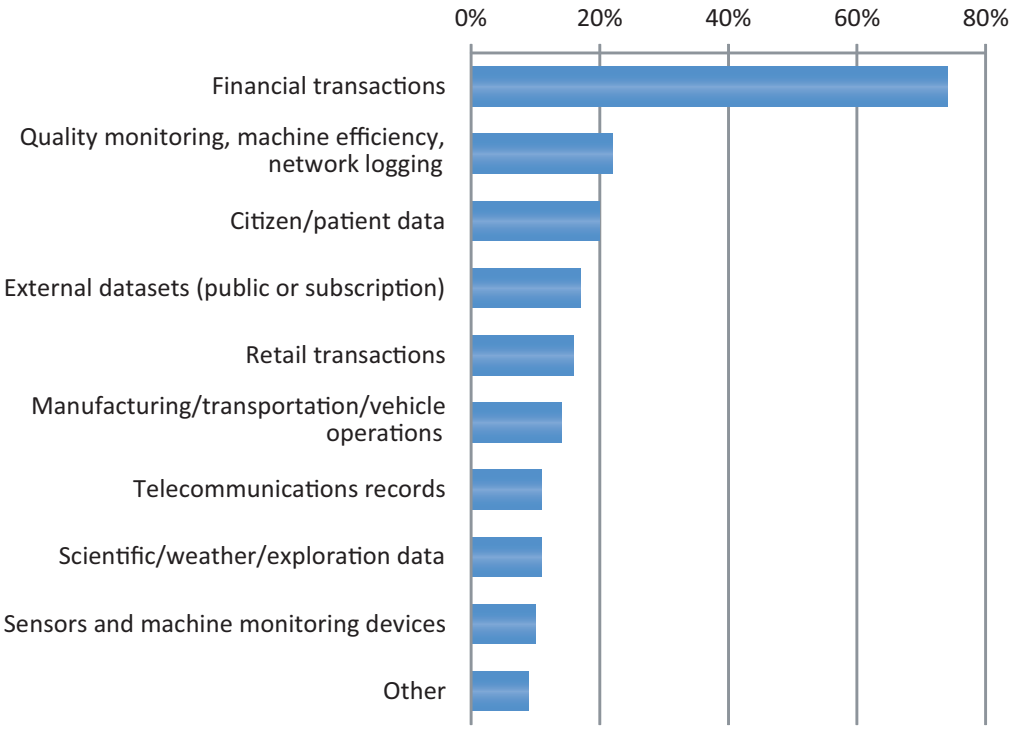
Over half of respondents are keen to make good use of big data techniques. 23% feel it will become an essential capability for their business, and for 9% it already is.

Structured, Unstructured and Linked Datasets

Structured

In order to set the scene, we asked respondents if their business generated or used large amounts of "structured" data. Financial transactions create the most ubiquitous set of data in most companies – although one might question if it actually counts as "big" or not for most sizes of organization. Beyond that, most responses were heavily related to the appropriate vertical industry. External datasets, either public or subscription are used by 17% of organizations. In the "Other" category, some respondents mentioned HR and staff data, which may be considerable in very large companies, and some were a little unsure of the distinction between structured and unstructured data.

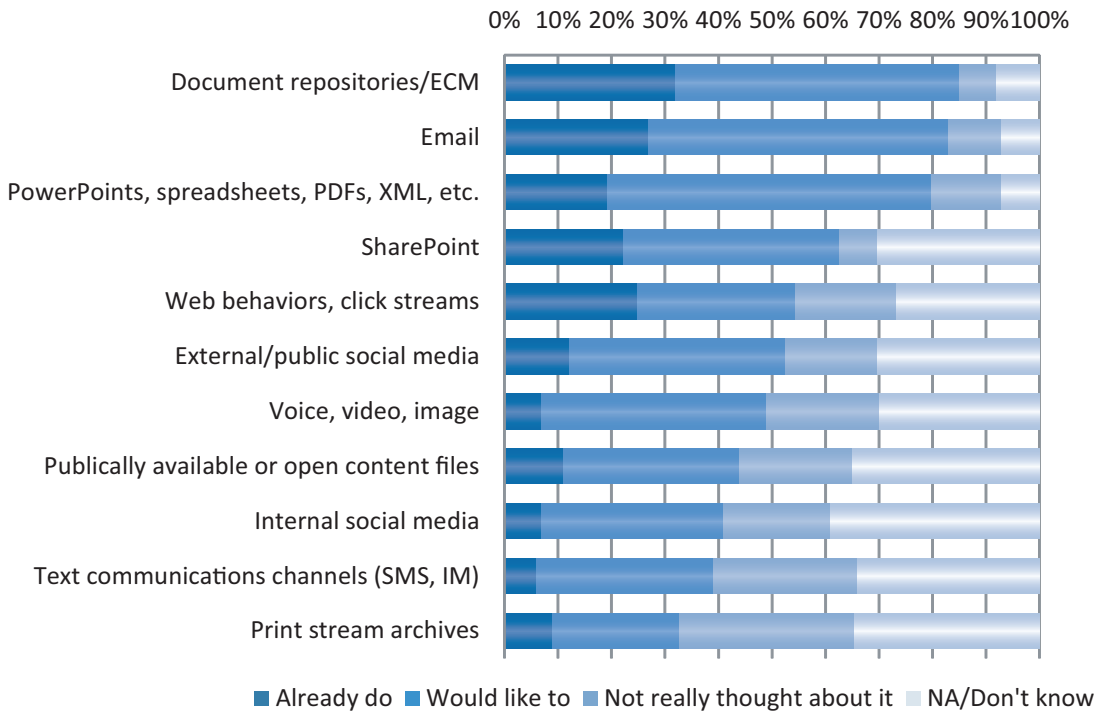
Figure 5: Are there areas of your business operations that generate or utilize large amounts of structured data (ie, transactional, numerical, etc.)? (N=300)



Unstructured

The picture on the unstructured side is more difficult to evaluate. Although we asked users to differentiate between analysis and search, it seems likely that there is some crossover, and to an extent, data types like video and audio tend to be inherently big data, but are also a struggle to search, let alone research. Equally, log files and click streams could best be described as semi-structured data, particularly if presented as XML files.

Figure 6: Are there large unstructured or semi-structured data repositories (ie, text, rich media, etc.) in your business that you would like to analyze, monitor or query - as opposed to search/retrieve? (N=300 users)



The most popular candidates for analysis are the standard document types, and, of course, email – although this is perhaps more likely to be search than analysis. Next come web applications such as click-stream tracking and social media.

Analyzing print-stream archives as in high volume transaction output or HVTO is an interesting application here. Although it is currently only being used by 9% of organizations, it is potentially applicable to 65% - albeit that half of those have not, as yet, given it any thought as a potential source of trends and indicators. 24% say they would like to do so. Utilities and financial organizations have considerable history locked up in the output archives of printed statements, although as we will see later, real-time or near real-time analysis of the customer communications output stream can provide live trend prediction.

Linked

Having introduced respondents to the structured and unstructured datasets they already have, we sought to find out how useful it would be to link the two types together for analysis – for example, linking case reports to geographical demographics, or customer web behavior to their history of product sales. Nearly 60% could see that linking would be very useful, although only 2% are doing so as yet.

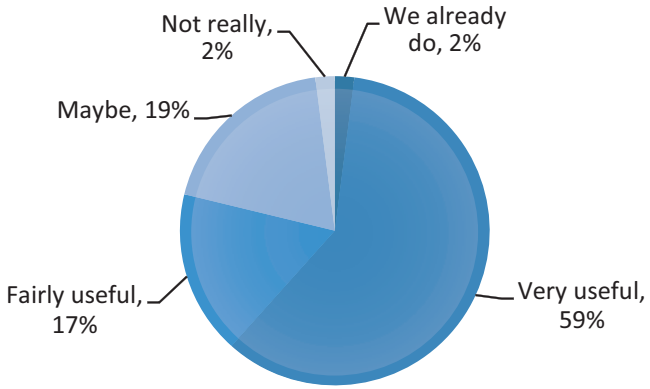


Figure 7: Would it be useful if you could link structured and unstructured data sets for analysis? *(N=285)*

Over 60% would find it very useful if they could correlate text-based data with transactional data, but only 2% are able to do so at present.

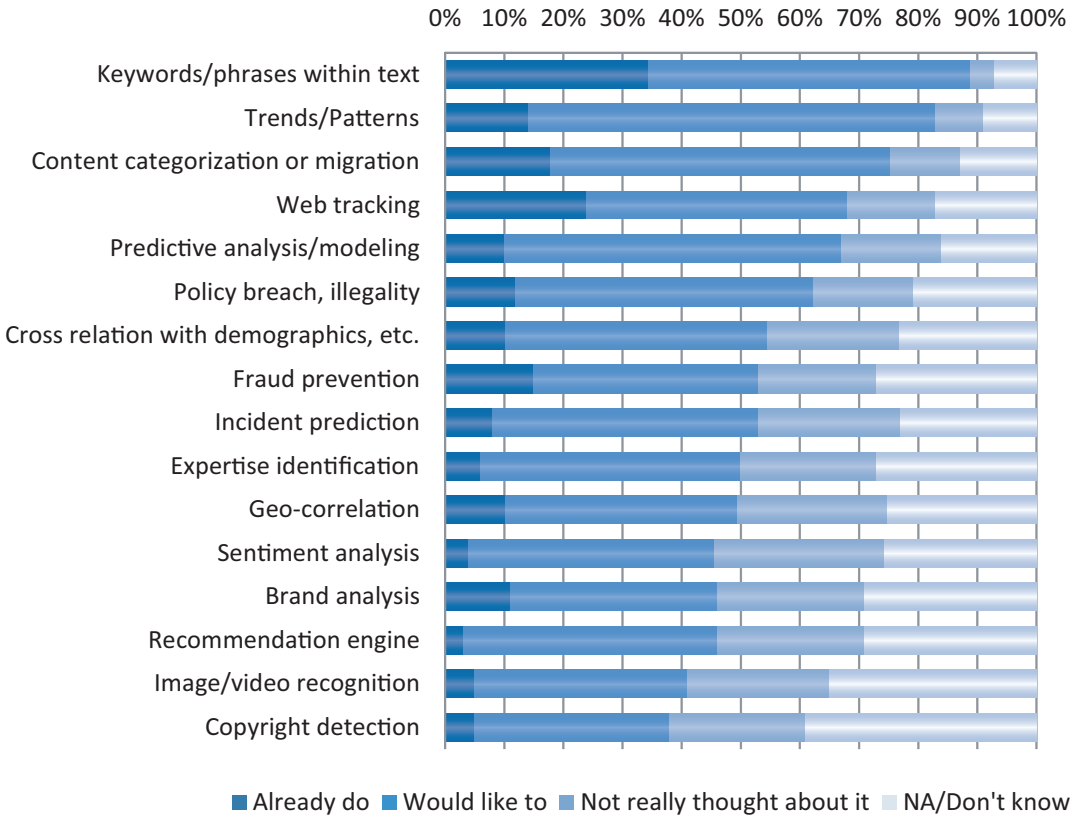
Unstructured Analytics

Techniques

There are many different analytic techniques that can be brought to bear on one or more repositories of unstructured data. Textual analytics and semantics have been around for many years. However, these techniques have been brought into the limelight by the ubiquity of born-digital documents and the fact that modern processors are able to execute very sophisticated algorithms in a timely manner. Added to this is the potential additional insight made available from cross-linking to conventional BI analysis.

Keyword search within context is one of the most common uses, and is on the border between advanced search and true analytics. Auto-classification and content categorization pick up on techniques developed in spam filtering to filter and sort high volume content such as emails and message streams into long-term archive, with automatic metadata creation and tagging. Sentiment analysis, copyright and IP theft, social media monitoring, and fraud detection are all based on semantic analysis. Going beyond that are cross-over applications such as web-tracking, geo-correlation, expertise identification and demographic segmentation. A full glossary of these terms is given in Appendix 2.

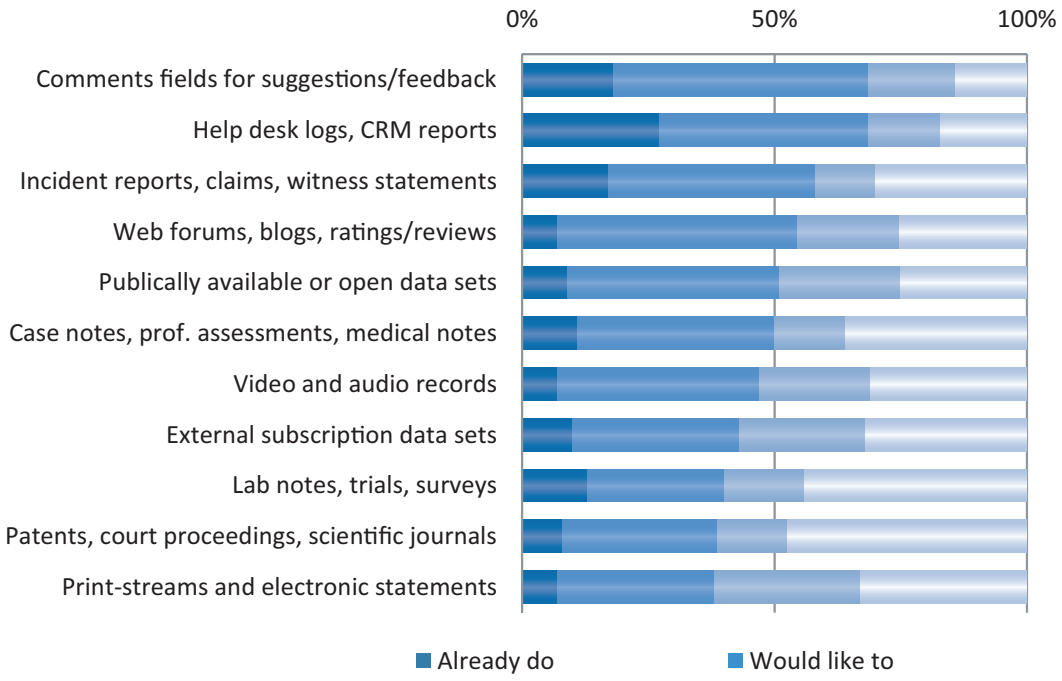
Figure 8: What type of analysis would you like to do/already do on unstructured data? (N=270)



Content Types

Picking up on the mainly text-based techniques above, we asked our respondents what document or content types they would like to analyze. Interestingly, the open-ended comment fields provided on most forms for comments and suggestions proved to be the most popular, with 68% looking to analyze them in a more formal way. Of course, if these comments are coming as hand script on paper forms, businesses might also need to invest in an ICR capture capability. Incident reports, claims forms and witness statements are seen as useful candidates for forensic search and cross correlation, as well as fraud detection. Case notes and professional assessments are quite widely seen as useful, and of course the medical notes application is one of the "poster boys" for big data.

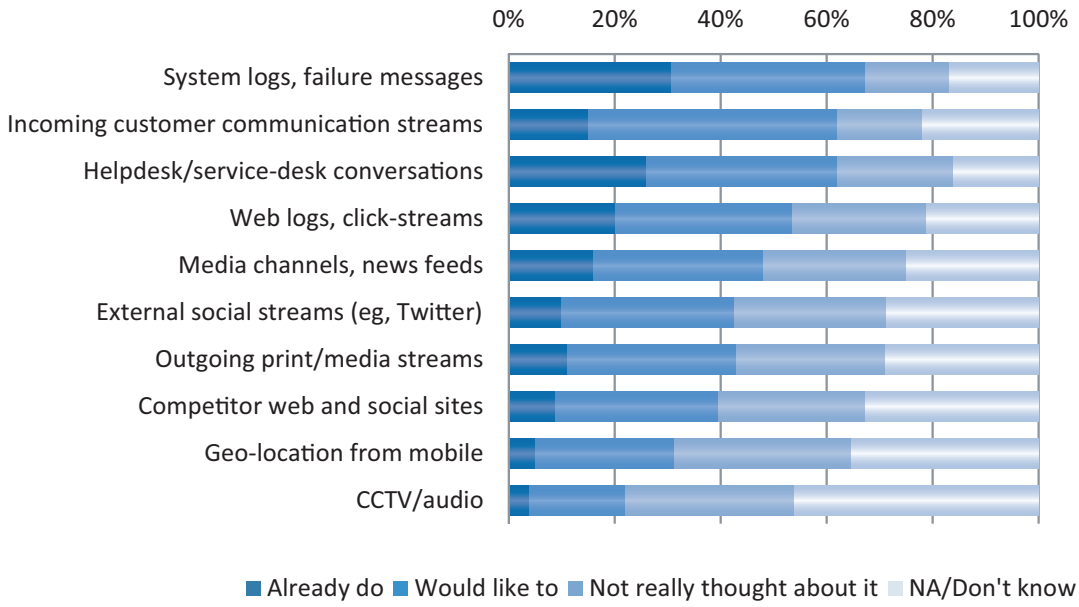
Figure 9: Have you considered analyzing any of the following document or content types to extract longer-term business intelligence or solve problems? (N=259)



Real-Time Analytics

One thing that characterizes the modern age is the constant stream of data that has become available inside or outside the business, often as a side effect of other activities. System logs, web logs, geo-updates, social streams, print streams and news feeds are to an extent spinning along, waiting to be exploited. On the other hand, the potential value in helpdesk transactions, incoming customer communications, and competitor website updates is obvious to all. In both cases, capturing and monitoring these data streams in a robust and repeatable way can prove somewhat overwhelming. Here the new technologies come into their own as data volume and data velocity combine to overwhelm conventional computing methods.

Figure 10: Have you considered analyzing any of the following to extract live or near-time business intelligence? (N=260)



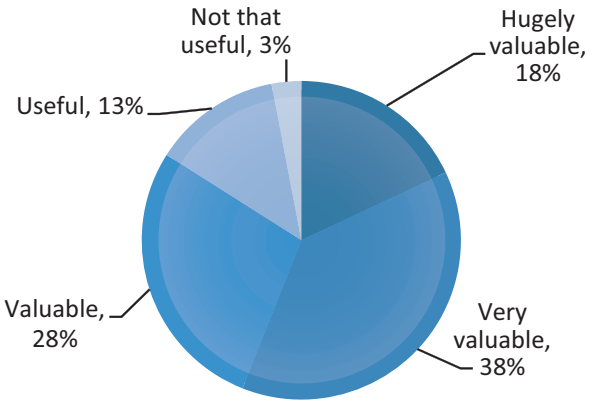
Monitoring of social streams is quite low on this list, perhaps because it is very much a marketing and PR issue. In some very customer-sensitive companies, staff are being specifically employed to monitor social sites for positive and negative comments about companies, brands and products, in order to quickly react and de-fuse negative comments or enhance positive comments. Automated sentiment analysis would be likely to provide a cheaper and more consistent alert mechanism, and one that can be better analyzed over time for the effects of brand promotions and PR activities. One UK bank provides a live sentiment analysis on their website based on customer comments, with a continuously updated positive/negative rating - currently 70% positive!

Monitoring customer communications, spotting trends and categorizing content are the most popular applications of big data techniques.

Value

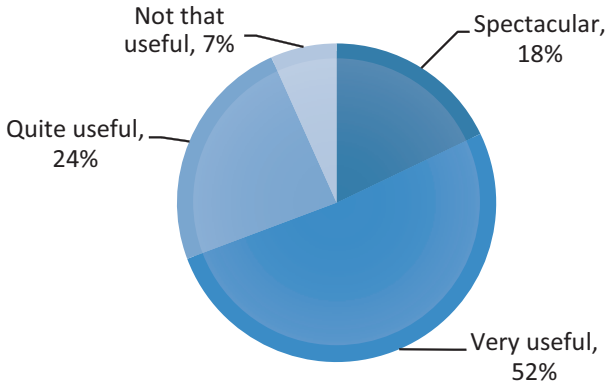
Our respondents see huge value in these kinds of sophisticated analytics on unstructured content and data streams, with 56% saying they would find it very valuable, including 18% who would consider it hugely valuable.

Figure 11: How valuable would it be/is it to your business if you could do the kinds of analysis suggested above? *(N=274)*



We then asked users to describe a "Killer Application" for their type of business - if they had suitable analytics tools. As might be expected, the answers were quite specific to various vertical industries, but we also gave the option to keep such an application to themselves. 85% took this "Not disclosed" option. We then asked how they would rate the usefulness of such an application to their business. 70% said it would be very useful, including 18% for whom it would be "spectacular".

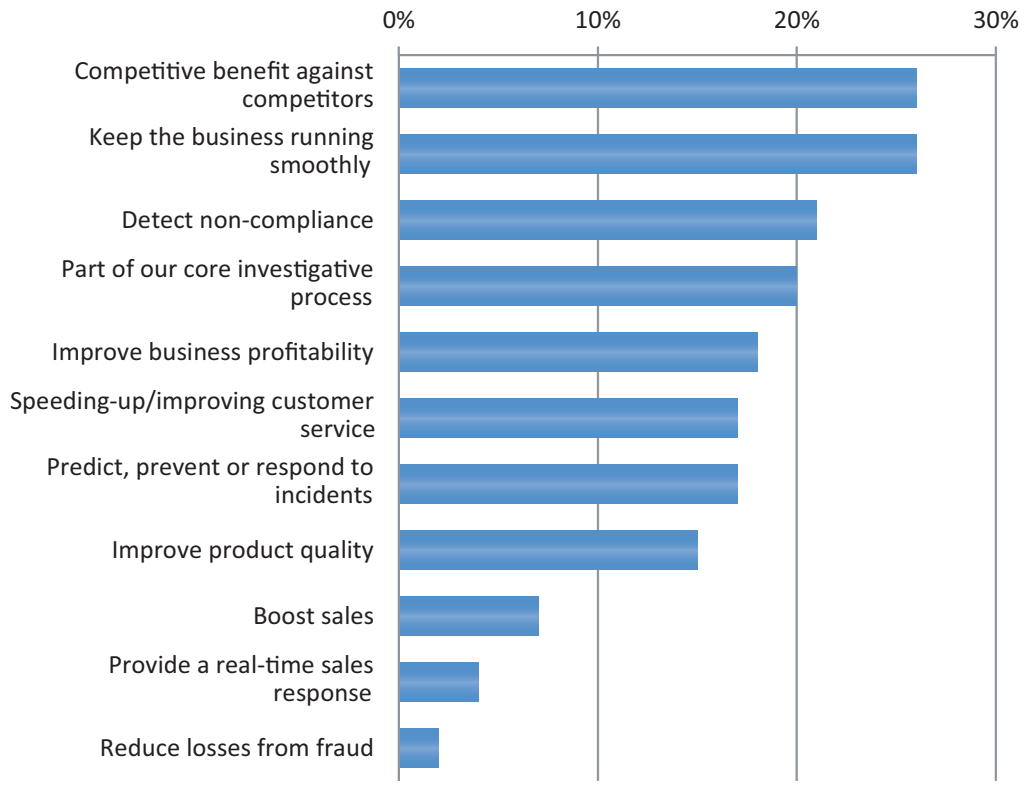
Figure 12: Based on your answer to the previous question (Killer Application), how would you rate the usefulness of such an application to your business? *(N=179 with such an application)*



To explore the value proposition more closely, we asked how such a Killer Application would help their

business. “Competitive benefit against competitors” is the strongest answer, and this explains why so many preferred not to disclose their ideas. Next highest, “Keeping the business running”, is likely to apply to those real-time or near real-time situations where incident prediction and prevention might come into play. Non-compliance is also quite high up, suggesting that big data can play a part in policy enforcement by detecting inappropriate use and beyond that, potential illegal or criminal behavior. For 20% of organizations, these kinds of techniques provide a boost to their core investigative processes, whether they are scientific or forensic.

Figure 13: How or where would such an application prove most useful for you? (Max Two) *(N=179 with such an application)*



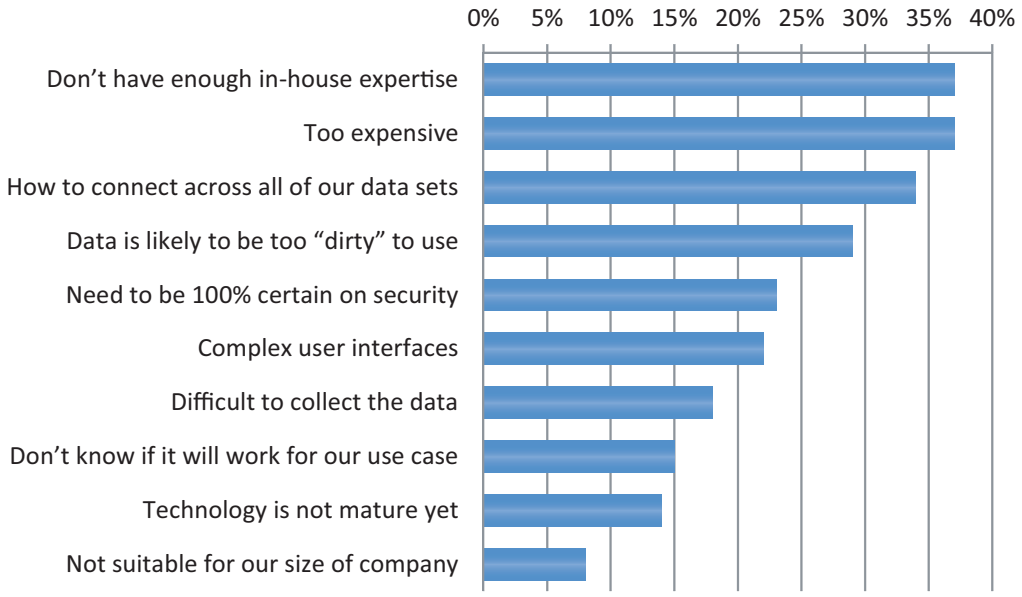
Most respondents can identify a “killer application” for big data that would enhance their competitive position, avoid business discontinuity, or detect non-compliance.

Issues

Expertise

As would be expected for a relatively new area of technology, users have concerns about how it works, how much it costs and how to get started. Advanced BI has always been one of the more challenging areas of IT. When we add the vagaries of unstructured content, and the need to connect across multiple repositories, we can see that users may be struggling to find the appropriate expertise within their business.

Figure 14: Which three of the following issues concern you most about big data/big content analytics tools? (N=250)



Skills shortages will always create a rush of interest, and 45% of our respondents are keen to get better trained, recognizing that this would increase their usefulness to their own and other companies.

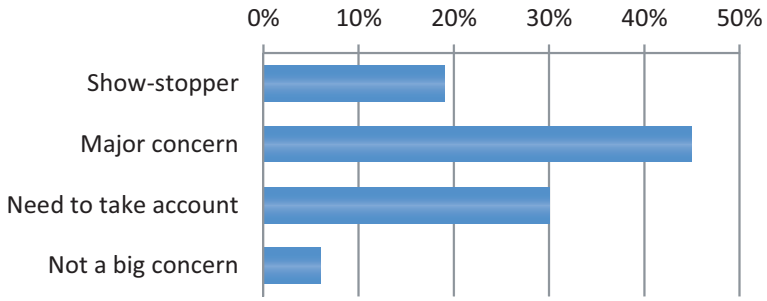
Connectivity

After cost concerns, how to connect to all these disparate datasets is the next consideration. ERP systems, CRM systems, document and records management systems, email systems, web content systems, and specialist line-of-business systems might all contain useful data for analysis and linking before we consider specific big data files such as log streams, social content, geo-tracking and so on. Some of the pioneering work associated with enterprise search products can be re-purposed, and the concept of "unified data access" is important to the working of many toolsets in this area. Datasets can be linked-in-place, or content may be migrated across to a single big data store, much as data warehousing operated in the past. Uniform connectivity is required in either case, but the issues of query creation will be different.

Security

Often missed as an impediment to both enterprise search and big data systems, security can be a show-stopper for many organizations (19%) and is a major concern for most (45%). Personally identifiable data must always be protected, and many potential applications for segmentation and personalization cannot be anonymized. Even if the raw data is not sensitive, given the competitive benefits we have described above, the results of analysis could be very company confidential. It is interesting to note that an increasing number of public or open datasets are providing some reassurance that otherwise sensitive data can be made safe.

Figure 15: Is security or permissions a concern for you in search and analytics? (N=274)

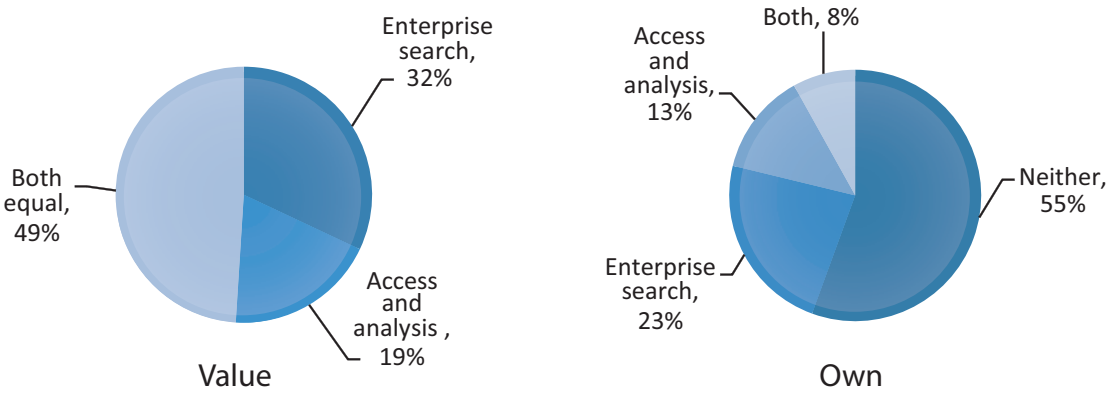


Lack of in-house expertise, connectivity and security are the biggest impediments to adoption of big data technologies – as well as a perception that costs are high.

Priorities

We outlined earlier a degree of crossover between search and analytics, and there is no doubt that in some organizations these two may compete for priority attention - albeit that they actually do different jobs, much as database searching has always been different from database reporting.

Figure 16: Which is/would be the more valuable to the overall competitivity of your business unit? Which capabilities do you currently have? (N=274)

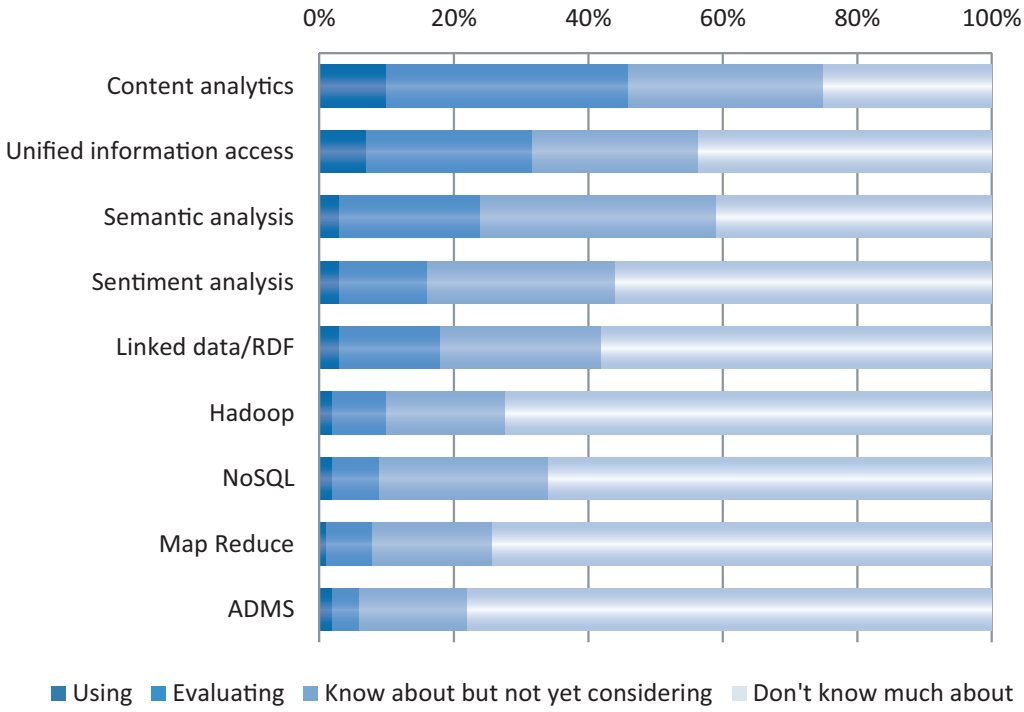


Most users seem to appreciate this difference, but would still see value in having access to both (49%). However, only 8% actually have both, and 55% have neither. Users are more likely to have invested in search than unified access and analytics.

Big Data Technologies

There are many new terminologies and technologies arriving in the mainstream as a result of the new interest in big data. Many are defined in the Glossary in Appendix 2. On the whole, our respondents are quite well informed as regards their understanding of analytics and unified data access (50%), but are less aware of the new database and query technologies such as NoSQL, Hadoop and Map Reduce (25%). Many of these are open source, but usage levels are quite low as yet.

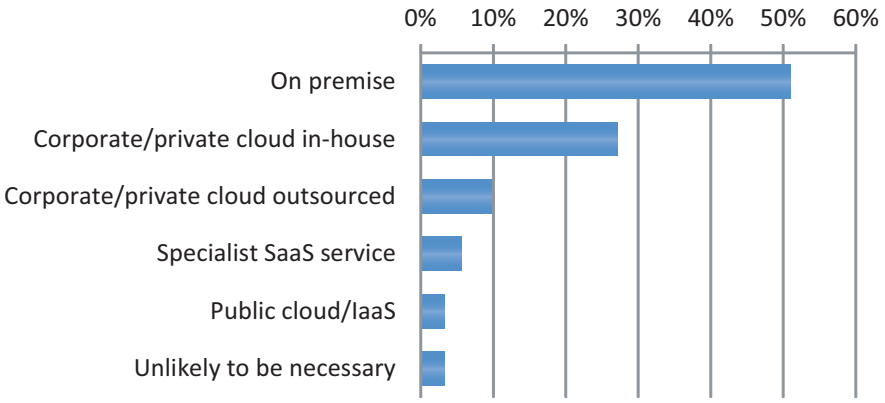
Figure 17: Which of the following technologies have you looked at? (N=269)



SaaS and Cloud

In theory, large and rapidly growing datasets and complex analytic products should lend themselves very well to Cloud and SaaS deployment. However, as we saw earlier, security is a big issue, and this plays against the current perceptions, particularly of private clouds. There may also be a concern that truly big data could quickly run up huge data hosting bills. However, given the shortage of trained expertise and the perceived expense of big data tools, Cloud and SaaS deployment would offer a quick entry-point for many organizations, reducing the need to pre-dimension servers, and providing access to very specific analytics techniques.

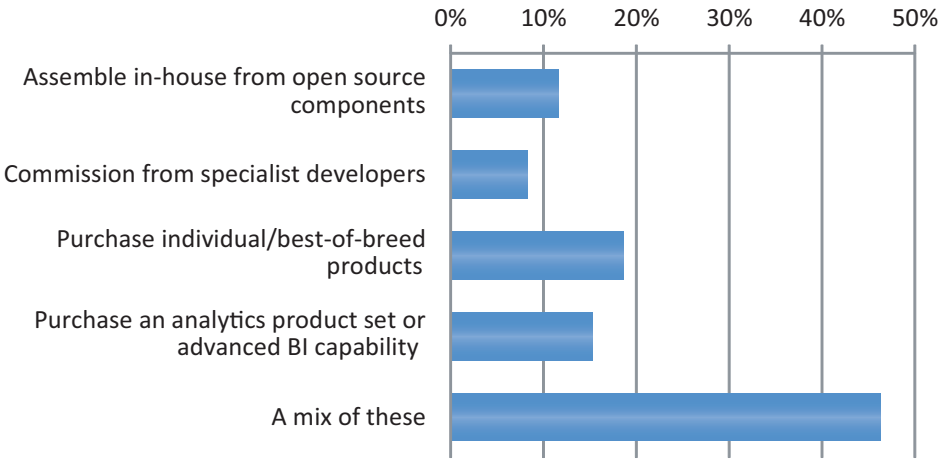
Figure 18: If you needed to create a "big data" dataset for analysis, where would you most likely create it? (N=214, excl. 55 Don't Know)



Big Data Product Types

Is big data a set of tools, a product or a technology? It would seem that all of these are applicable, and this reflects potential users' views of how they might acquire a big data capability. Vendors are moving quickly to provide packaged product sets, and this is driving a need for standardized connectors to provide unified data access to as many different databases as possible. The temptation to press ahead with in-house developments using open-source components may be driven by early-mover competitive advantage. However, as we have seen before in BI, usability outside of the technical department is important, and for big data, assurance of robust security is essential.

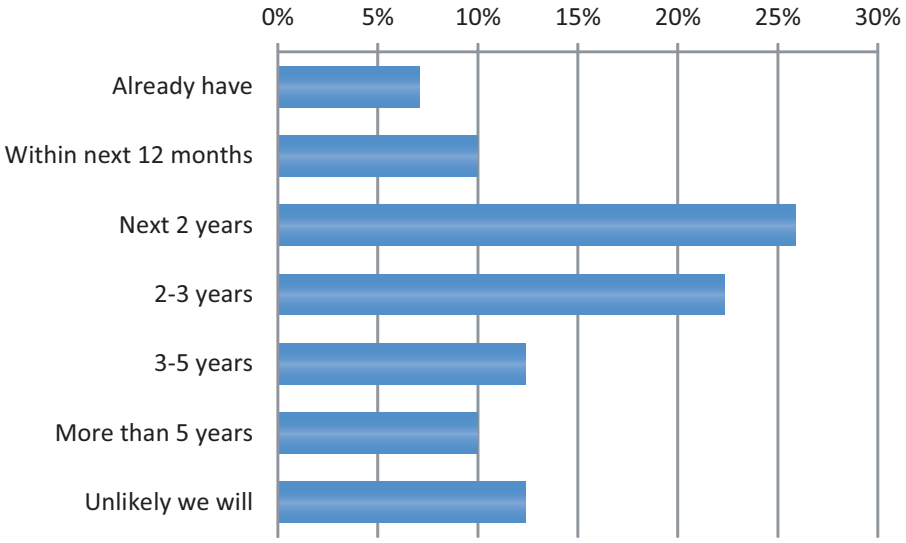
Figure 19: If you needed to create a "big data" dataset for analysis, where would you most likely create it?
(N=242, excl. 70 Don't Know/Unlikely)



Spend

Despite this relative immaturity, early adopters are keen to get ahead of the competition, with 10% of our survey respondents planning to make an investment in the next 12 months, adding to the existing user-base of 7%. Perhaps more interesting is that 48% of the organizations surveyed are looking to make a move into big data in the next two to three years. If this were to be reflected even partially in the population at large it would represent a massive growth.

Figure 20: When might you expect to make an investment in big data/big content analytics tools?
(N=170, excl. 92 Don't Know)



Overall big data spend is likely to be split between hardware, software and consultancy services. Estimates elsewhere have suggested that given its complexity, and the general lack of in-house expertise, consultancy services will take a higher share, with hardware and software fairly evenly split. The hardware component is, of course, highly dependent on the need for specialized or extended storage.

Big data is very much in the early adopter stage. Many organizations are convinced of the benefits and are poised to jump in as soon as more packaged products are available with simpler user interfaces and greater applicability across different data repositories.

Conclusion and Recommendations

Although surrounded by hype, the ability to analyze and correlate big data and big content repositories is deemed to be very useful by our respondents. In particular, the linking of unstructured text or rich media data with structured transactional data would be very attractive to many.

The range of available analytic techniques across unstructured and semi-structured data is considerable, and this, combined with the need to present a unified data connection across diverse datasets, presents a challenge for most organizations. Expertise is scarce, and generic analytic product platforms are in their infancy, with most organizations resorting to a mix of open source products, in-house development, and best-of-breed applications. For many, there is also a security issue of both a competitive and a compliance nature.

However, most of our respondents consider that there are one or more "killer apps" for big data in their organization that would provide a dramatic competitive benefit, or would provide much better prediction and prevention for business continuity.

For a third of organizations, the current levels of content chaos and the lack of conventional BI would suggest that they have higher priorities at the moment, but according to the spend intentions given by our respondents, the existing big data user base is set to double to around 17% in the next 12 months, with 48% looking to make investments in a 2-3 year timeframe.

Recommendations

- Ask blue-sky questions of your business such as "if only we knew…" or "if we could predict…" or "if we could measure…" Consider how useful that might be to the business *before* thinking about how it can be done or at what cost.
- Play those questions off against the data you already have, data you could collect, or data that you could source elsewhere.
- Include in your thinking structured transactional data, semi-structured logs and files, and text-based or rich media content.
- Incoming communications from your customers, outbound communications to your customers, and what customers (or employees) are saying about you on social sites can all be useful for monitoring sentiment, heading off issues and analyzing trends.
- Consider high volume streams such as telemetry, geo-location, voice, video, news feeds, till transactions, web clicks, or any combination of these.
- If your content is currently digital landfill spread across disparate file shares and content systems, consider how this could be rationalized prior to any big data projects.
- Content access for both search and analytics, and if necessary, content migration to dedicated big data storage, can be facilitated by unified data access products.
- Don't be tempted to rush into in-house developments or specific point-solutions without considering wider, more universal analytics platforms.
- Consider SaaS and cloud deployment as a faster way to acquire experience and focus activity.
- Big data is much more about the breadth of the analysis and the insights that can be achieved than it is about the size of the data and the underlying database technology.

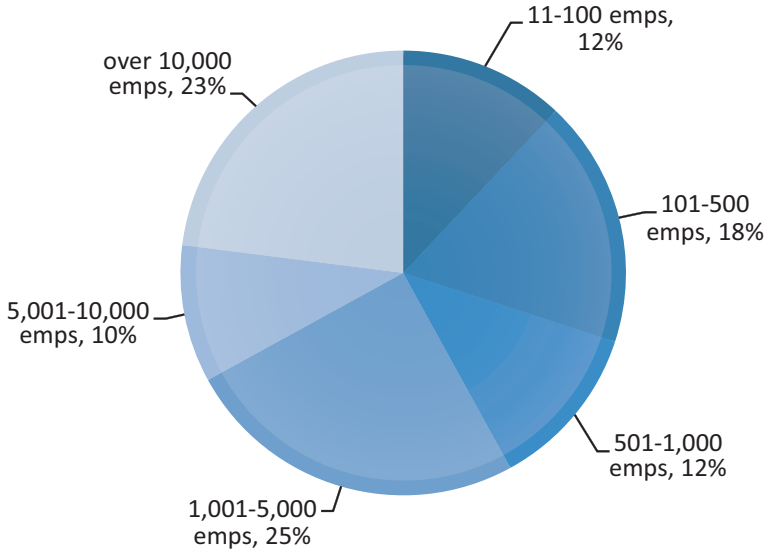
Appendix 1: Survey Demographics

Survey Background

402 individual members of the AIIM community took the survey between Mar 30 and Apr 25, 2012, using a Web-based tool. Invitations to take the survey were sent via email to a selection of the 65,000 AIIM community members.

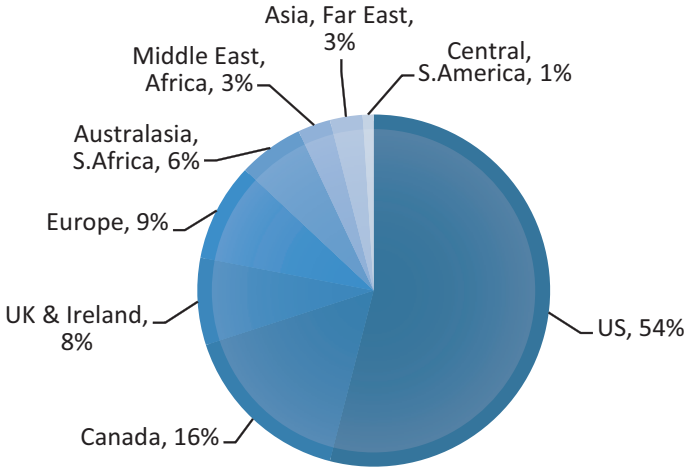
Organizational Size

Survey respondents represent organizations of all sizes. Larger organizations over 5,000 employees represent 33%, with mid-sized organizations of 500 to 5,000 employees at 37%. Small-to-mid sized organizations with 10 to 500 employees constitute 30%. Respondents (58) from organizations with less than 10 employees or from suppliers of ECM products and services have been eliminated from the results.



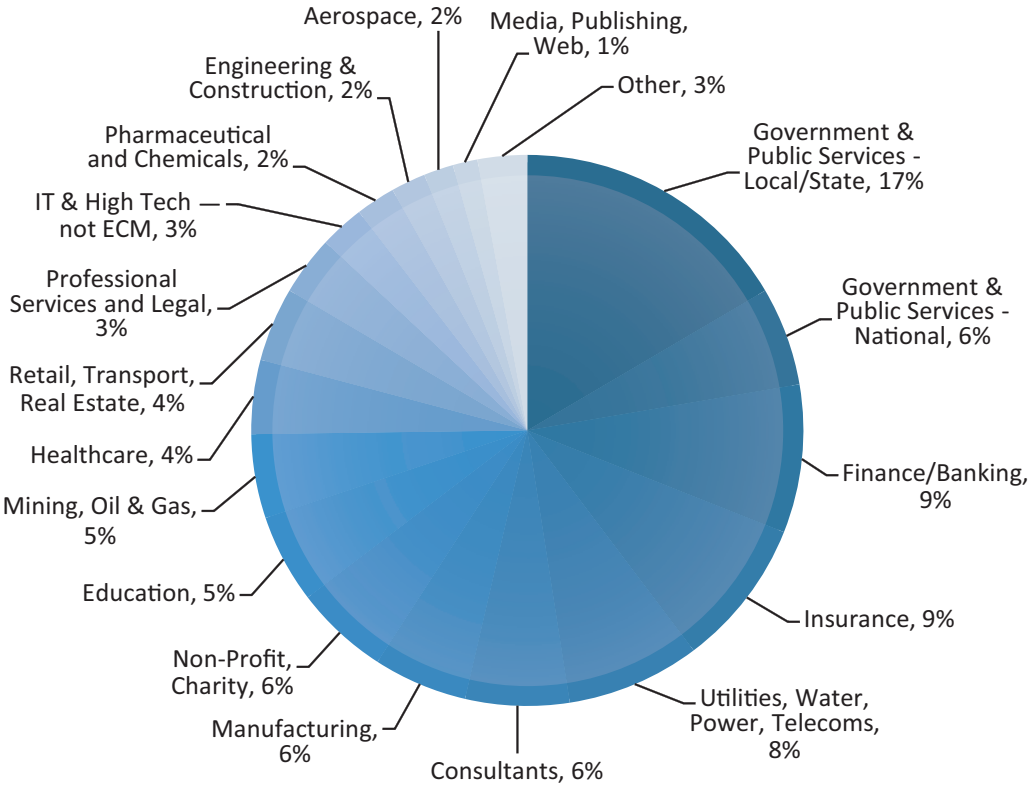
Geography

70% of the participants are based in North America, with most of the remainder (17%) from Europe.



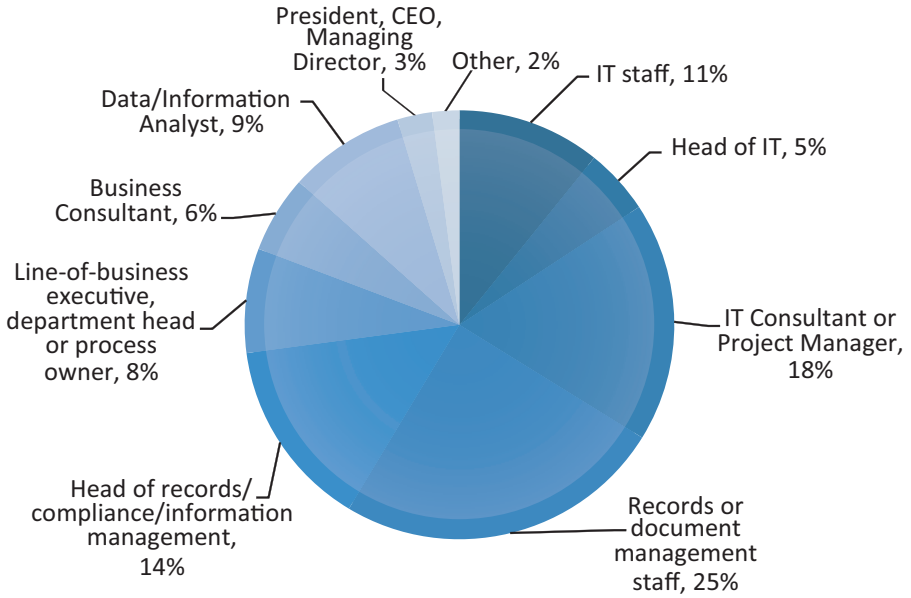
Industry Sector

Local and National Government together make up 23% and Finance, Banking and Insurance represent 18%, and Utilities 8%. The remaining sectors are fairly evenly split. To avoid bias, suppliers of ECM products and services have been eliminated from all of the results.



Job Roles

34% of respondents are from IT, 39% have a records management, information management or compliance role, 19% are line-of-business managers, and 9% are BI analysts.



Appendix 2 - Glossary of Terms

ADMS:

- **Automated redaction:** Search, matching and blanking across scanned images or electronic documents of personal details such as social security numbers, phone numbers, names, monetary amounts, etc.
- **Click stream:** Web logfiles that can be used to indicate the behavior and track the paths of website visitors. When aggregated provide ability to optimize content, or individually to optimize user experiences.
- **Content Analytics:** A range of search and reporting technologies which can provide similar levels of business intelligence and strategic value across unstructured data to that conventionally associated with structured data reporting.
- **Content assessment:** Trawling of stored documents and content to measure value, relevancy, currency, frequency or recency of access as an indication of the need to keep, or more particularly, migrate content to another system.
- **Content de-duplication:** Identification of exact or near-exact match of content stored within the same or different systems, albeit with different metadata as to who stored it and when. Scoring system allows automatic deletion of duplicates, saving space and reducing potential errors.
- **Copyright detection:** Web search to match unauthorized use of copyrighted images, sound files, etc., and to serve notice of infringement. Similar tools may also be used to detect out-of-date logos or product imagery across marketing collateral and company websites.
- **Demographic Segmentation:** Analysis of customer profiles, customer behavior, customer location or customer communications in order to better define and target market communications including web-page presentations, offers, etc.
- **Digital Asset Management (DAM):** Content management systems particularly geared up for rich media files such as images and sound which are characterized by large file sizes, proxy representations (low resolution thumbnails or clips) and complex coders, decoders or format transformers.
- **Digital forensics:** Investigative analysis to detect fraud or misbehavior through word usage, trends, links, photo or sound patterns, particularly with a view to providing evidence for potential legal action.
- **Digital Rights Management (DRM):** Technology that inhibits use (legitimate or otherwise) of digital content that was not desired or foreseen by the content provider or copyright holder, particularly to prevent unauthorized duplication.
- **Social media monitoring:** see also *Sentiment Analysis*. Monitoring of positive or negative comments being made on social media (eg Facebook, Twitter, etc,) in order to head off potential PR storms or negative brand impact. For internal social sites, can be used to monitor staff morale.
- **E-Discovery tools:** Search tools which analyze content for its likely relevancy to litigation by, for example, linking names, time periods and terms used. May also extend to legal hold and partitioning of content for further scrutiny.
- **E-mail trending:** Reporting of email activity to indicate unusual patterns or topics of particular current interest. May also reflect potential faults or incidents, poor service, media coverage or staff sentiment.
- **Expertise identification:** Sourcing of experts based on analysis of profiles, resumés, qualifications, etc.
- **Faceted search tools:** Ability to sub-divide within search results by a standard set of metadata tags, eg, as used by shopping websites to sub-set a product search by manufacturer, size, color, price range, etc.
- **Federated search:** Ability to interrogate more than one repository or index from a single search

screen. Might include linking internal ECM systems with subscription access to government databases or those of professional bodies.

- **Fraud Prevention:** Analysis of claims, statements, refunds, etc., for usage patterns, inconsistencies, correlations with past incidents and demographic analysis.
- **Hadoop:** An open source, Java-based programming framework on Apache that supports the processing of large data sets in a distributed computing environment.
- **Image and sound tagging:** Pattern recognition search to match and apply additional metadata, eg, faces, trees, voices, birdsong, to rich media files. May also be part of digital forensics.
- **Map Reduce:** A query or look-up technique optimized for large datasets on highly distributed or parallel servers.
- **NoSQL:** Database technology that does not resemble conventional relational SQL databases and where the ability to store and retrieve great quantities of data is more important than the relationships between elements.
- **Recommendation Engine:** An associative inference engine taking a specific set of inputs- person, age, sex – and returns a % likelihood that they will do something or like something, eg., books or music based on previous purchases.
- **RDF/Linked data:** Resource Description Framework – a unified metadata description that assists with connections to different data repositories.
- **Rich media:** Generally used term for non- text formats such as photo-images, graphics, video, sound and animation. They cannot be searched by their textual content so must be tagged and/or represented by thumbnails or audio samples.
- **Sentiment analysis:** Analysis of words used in comment or feedback to indicate satisfaction or dissatisfaction with products or services, aggregated to an overall score of satisfaction. Often extended to overall brand response, and monitored for trends or incidents.
- **Text analytics:** Lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining, etc., frequently as part of other processes described here.
- **Three Vs:** Often used in the description of big data. **Volume** is the volume of data relative to the ability to store and manage it. **Velocity** is the speed of calculation needed to query the data relative to its rate of change. **Variety** is the number of different formats the data is presented in.
- **Web analytics:** Advanced reporting over time of web behavior including content assessed, paths followed, time on-site, referring sites, access points, search terms, geographic origin and pre-purchasing trends.

Appendix 3

Do you have any general comments to make about the potential for big data analytics in your business? (Selective)

- Outstanding potential and the wave of tomorrow's management skills and expertise.
- Big data is classic "hype cycle" stuff. Useful? Yes. Ground breaking? Hell no!
- I think smaller companies like ours will benefit from generally available analysis of big data from public or semi-private research and analysis efforts.
- Would like to see and evaluate a few products.
- Big data for healthcare is critical on many fronts.
- The real benefit for us is being able to analyze data and provide business insight before the business "knows the questions to ask".
- Government moves too slow so it'll likely be another decade before the state catches up!
- We have been saying for over twenty years; "we're drowning in a sea of data." Every year we still create more databases, more applications that create more data and have not improved our ability to analyze the data we asked for.

UNDERWRITTEN IN PART BY

Actuate - Xenos Group

Actuate Xenos Group products enable organizations to develop, implement and manage enterprise-class content management solutions. Our technology is used to design, store and deliver high volume content such as statements, policies, and bills for top-tier organizations in financial services, insurance and telecommunications.

We differentiate ourselves by offering our customers a fully integrated application framework consisting of **Actuate BIRT for Statement Design** and interactive presentation, **Xenos Enterprise Server** for high performance processing and systems management, **Xenos Document and Data Transformation** for content agility and repurposing, and **Xenos Repository** for storage and multi-channel delivery.

Customers engage Xenos based on the performance, scalability and reliability of our products.

www.xenos.com

Attivio

Attivio's unified information access platform, the Active Intelligence Engine® (AIE®), redefines the business impact of our customers' information assets, so they can quickly seize opportunities, solve critical challenges and fulfill their strategic vision.

AIE integrates and connects diverse information sources so our customers not only understand "what" is happening, but also have the critical context to understand "why" it is happening. Attivio intensifies the value of big data with advanced text analytics such as entity extraction and sentiment analysis, text extraction from file formats such as Word and PDFs, multi-language support and simple intuitive access and analysis directly by business users.

Attivio correlates disparate silos of structured data and unstructured content in ways never before possible. Offering both intuitive search capabilities and the power of SQL, AIE seamlessly integrates with existing BI and big data tools to reveal insight that matters, through the access method that best suits each user's technical skills and priorities.

www.attivio.com

UNDERWRITTEN IN PART BY

EMC Corporation

EMC leads customers and partners on their journey to big data, helping them capitalize on the big data opportunity to accelerate business transformation. EMC offers a comprehensive big data solution that enables organizations to achieve unprecedented value from all their data sources, both inside and outside the organization, gaining new levels of efficiency, agility and business breakthroughs. EMC's Big Data solution is architected on the EMC Isilon elastic and scale-out storage foundation, runs the Greenplum Unified Analytics Platform (UAP) designed to process both structured and unstructured data and enable collaboration among data science teams, and provides a business process platform called Documentum xCP containing application development tools that drive actionable insight.

Businesses that can exploit Big Data to improve their strategy and execution will distance themselves from competitors. Big Data is different in scale and significance and demands a new approach.

• Scale: Big Data is measured in Exabytes and billions of files. According to a 2011 IDC report, more than 90% of Big Data is unstructured data. • Significance: Big Data transforms business through strategic insight. • New approach: Traditional architectures and tools cannot deliver on the big data opportunity.

Organizations that embed analytics into applications and use data in context, as a part of all decision-making processes, can gain foresight into future outcomes and transition into a predictive enterprise. EMC Documentum xCP rapidly composes and delivers Big Data-enabled processes in a fraction of the time than traditional methods. The xCP platform takes data and content-rich, collaborative interactions and puts them into a structured process with clear roles, next steps and outcomes. It uses data and analytics to trigger workflows and business processes within applications, allowing you to:

• Act Immediately on Insight • Transform Business in Real Time • Optimize for Continuous Improvement

www.emc.com

IBM

In this age of big data, organizations must be able to fully exploit all sources of data and content for insight. Executives need to make decisions based not only on operational data and customer demographics, but also on customer feedback, details in contracts and agreements, and other types of unstructured data or content. How can you manage all of this data, and give executives access to the visually compelling information they need to make timely, informed decisions?

When companies can align, anticipate, and act on all available information, rather than a subset, they gain a powerful advantage over their competition. IBM can help you expand from enterprise data to big data with enterprise content management solutions that enable organizations to put enterprise content in motion—capturing, activating, socializing, analyzing and governing it throughout the entire lifecycle.

While there is a lot of buzz about big data in the market, it isn't hype. Plenty of customers are seeing tangible ROI using IBM solutions to address their big data challenges:

- **Pharmaceutical:** Increased the ability to dispose of unnecessary information by ten times
- **Healthcare:** Identified patients likely for re-admission and introduced early interventions to reduce cost, mortality rates, and improve patient quality of life
- **Research:** Reduced the time to discover and process a research asset from three to six months to as little as seven days

Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of information, to make your business more agile, and to answer questions that were previously considered beyond your reach. Until now, there was no practical way to harvest this opportunity. Today, IBM's solutions for big data open the door to a world of possibilities.

www.ibm.com

- Demonstrate your ability to bridge IT and business
- Enhance your value to employers and clients
- Become part of the next wave of information management professionals

aiim
The Global Community of Information Professionals

Become a Certified Information Professional

There is a market need for information professionals. Independent market research by AIIM confirms that senior executives find value in the certification of the information professionals.

- 61% of surveyed business executives would prefer consultants that hold the Certified Information Professional (CIP) designation
- 64% of business executives would prefer to hire a CIP versus a non-certified candidate
- 76% of business executives would pay a CIP a salary premium.

Email certification@aiim.org
www.aiim.org/certification

aiim®
The Global Community of Information Professionals

AIIM (www.aiim.org) has been an advocate and supporter of information professionals for nearly 70 years. The association mission is to ensure that information professionals understand the current and future challenges of managing information assets in an era of social, mobile, cloud and big data. Founded in 1943, AIIM builds on a strong heritage of research and member service. Today, AIIM is a global, non-profit organization that provides independent research, education and certification programs to information professionals. AIIM represents the entire information management community, with programs and content for practitioners, technology suppliers, integrators and consultants.

AIIM
1100 Wayne Avenue, Suite 1100
Silver Spring, MD 20910
+1 301.587.8202
www.aiim.org

AIIM Europe
The IT Centre, Lowesmoor Wharf
Worcester, WR1 2RR, UK
+44 (0)1905 727600
www.aiim.eu