EXECUTIVE SUMMARY

Big data analytics offer organizations an unprecedented opportunity to derive new business insights and drive smarter decisions. It’s no wonder, then, that big data initiatives are a top investment area today and a strategic priority for forward-thinking organizations in every industry.

The outcome of any big data analytics project, however, is only as good as the quality of the data being used. As big data analytics solutions have matured — and as organizations have developed greater expertise with big data technologies — the quality and trustworthiness of the data sources themselves are emerging as key concerns.

Although organizations may have their structured data under fairly good control, this is often not the case with the unstructured content that accounts for the vast majority of enterprise information. IDC believes that good information governance is essential to the success of big data analytics projects. Good information governance also pays big dividends by reducing the costs and risks associated with the management of unstructured information.

This paper explores the link between good information governance and the outcomes of big data analytics projects and takes a look at IBM’s StoredIQ solution.

THE PROMISE OF BIG DATA ANALYTICS

The amount of information that enterprises create and manage today continues to grow at an astonishing pace. The sheer volume, variety, and velocity of that information are staggering — whether the information is generated by digital customer interactions; captured from mobile devices and embedded sensors; harvested from social conversations and emails; contained in documents, videos, audio clips, and other unstructured data types; or streamed in real time.

Innovative big data solutions are enabling organizations to leverage their wealth of structured and unstructured information to uncover trends, predict the "next best action," and improve business outcomes. Big data analytics give organizations the insights they need to grow revenue and
market share, reduce cycle times and costs, manage business and compliance risk, and create sustainable operational and competitive advantage.

It's no wonder, then, that big data initiatives are a top investment priority for many executives today. Spend on big data infrastructure, software, and services amounted to $16.6 billion in 2014, and IDC expects this number to grow to $41.5 billion in 2018 – a compound annual growth rate (CAGR) of 26.4%. This is about seven times the rate of growth of the worldwide information and communication technology (ICT) market. Organizations from every industry are leveraging big data analytics to sense and respond in real time. For example:

- Retailers are leveraging big data analytics to gain a deeper understanding of customer preferences, segment customers in new ways, and target buyers with tailored and personalized offers that increase conversion rates and order size.
- Manufacturers are using big data analytics to optimize their supply chains, anticipate product problems and warranty issues, and improve the performance of enterprise assets and equipment.
- Energy companies and utilities are leveraging big data to improve their demand forecasts, build smarter grids, reduce outages, and optimize production.
- Healthcare organizations are turning to big data to optimize care and improve patient outcomes.
- Research organizations are using big data analytics to accelerate the pace of medical and scientific research.
- Government agencies are exploiting big data for intelligence, national security, and mission support and planning.
- Financial services organizations are using big data to detect and prevent fraud.

As the amount of data continues to grow – and as organizations begin to leverage more of the unstructured information they collect – the use cases for big data analytics will continue to expand. Organizations are already using natural language processing to mine information in contracts, customer correspondence, call center transcripts, patient records, social conversations, industry journals and research publications, disclosure documents, emails, and many other sources.

Cognitive computing, which leverages artificial intelligence and machine learning to infer and predict, offers tremendous potential to augment human expertise and accelerate knowledge transfer around the globe. Indeed, big data analytics will be transformative for many industries, and unstructured information will play an increasingly important role.

**Governed Data Is Good Data**

We previously highlighted three of the four hallmarks of big data – volume, variety, and velocity. As organizations become more experienced with big data initiatives, they are beginning to pay greater attention to what IDC views as the fourth key attribute of big data: veracity. After all, the quality of the input data determines the trustworthiness of the analysis.
As IDC research shows, good information governance is a key component of the strategic use of big data analytics — especially for organizations that hope to progress beyond ad hoc and opportunistic use to repeatable, managed, and optimized use (see Figure 1).

**FIGURE 1**

**Good Information Governance Is a Key Component of Big Data and Analytics Maturity**

---

**Information Access**
- **Experimental**
  - Ad hoc siloed pilot projects; undefined processes; individual effort
- **Opportunistic**
  - Easily available information is utilized but is incomplete, and preparation requires substantial manual effort

**Data Analysis**
- **Intentional**
  - Defined requirements and processes; budgeted funding; efficient project management and resource allocation
- **Managed**
  - Information collection, monitoring, and integration processes are in place, but consistent governance and security practices haven’t been established

**Comprehensive Hindsight**
- **Repeatable**
  - Recurring projects; budgeted and funded program management; documented strategy and processes; stakeholder buy-in
  - Business Outcome: New product and service opportunities transition to business plans

**Actionable Insight**
- **Optimized**
  - Continuous and coordinated big data and analytics process improvement and value realization
  - Business Outcome: Previously unattainable business value is continuously produced

**Enterprise-wide Foresight**

---

**Governance**
- Metrics are in place to manage information quality, timeliness, and veracity and to govern collection, monitoring, and management processes

---

Source: IDC’s Big Data and Analytics MaturityScape, September 2014
Governance of Unstructured Information

Most organizations manage their *structured* data effectively. This is the data that fits neatly into the rows and columns of a relational database and is managed by the organization's enterprise applications, including enterprise resource planning (ERP), customer relationship management (CRM), human capital management (HCM), supply chain management (SCM), and other systems.

The same cannot be said, however, for the enterprise's *unstructured* information — the documents, images, rich media, and other content assets that reside in the organization's enterprise content management, collaboration, and email systems; on network drives and users' computers; and in enterprise application document stores whether on-premise or in the cloud.

Unstructured information accounts for about 90% of enterprise information, and many organizations lack the processes and systems required to effectively manage this information throughout its life cycle. This is a sizable problem.

The consequences of storing this ungoverned information can be severe. As IDC research shows, a quarter of companies suffer some sort of information leak each year. Leaked strategic plans, merger and acquisition information, product plans and other intellectual property, or customer information can damage the organization's brand, adversely impact customer loyalty, put the company at a competitive disadvantage, and expose the company to regulatory penalties.

Similarly, using information that is wrong, out of date, or incomplete — "bad data" — for analytics and decision making exposes the organization to risk. Ensuring the veracity of the unstructured information that is fed into big data analytics applications is thus becoming a top-of-mind concern.

Need for an Information Governance Solution

Governance of unstructured information is a more challenging problem than it might appear. Governance entails finding and cataloging all of the files and folders that are stored in disparate systems; identifying duplicative, confidential, and sensitive information; and assessing the value of all of that information to the organization so that toxic, out-of-date, and low-value information can be defensibly deleted. An information governance solution makes this manageable by bringing discovery, categorization, process, and best practices to bear — and ensuring visibility and auditability.

The need to better manage and govern unstructured information is becoming more apparent. As IDC research shows, about half of senior IT leaders in the U.S. and EMEA regions recognize the need for improvement (see Figure 2). Further education is warranted, however.
FIGURE 2

Growing Awareness of Information Governance Issues

<table>
<thead>
<tr>
<th>Percentage of respondents who agree/strongly agree on a scale from 1 to 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proliferation of content on team sites is a compliance/governance challenge</td>
</tr>
<tr>
<td>User adoption of cloud file sync and share solutions creates fresh governance challenges</td>
</tr>
<tr>
<td>We need to improve our retention and disposition processes</td>
</tr>
</tbody>
</table>

Source: IDC's ECM Strategy Survey 2013: Highlights, April 2014

Information Governance in the Era of Big Data Analytics

As we engage with big data analytics on unstructured information, the need for an information governance solution becomes even more acute when new needs are factored in. There is an inherent conflict between the priorities of compliance and governance professionals and the desires of big data analytics teams.

From a records and retention management perspective, it's desirable to delete (defensibly dispose of) out-of-date, inaccurate, duplicative, toxic, and low-value information. Deleting all of the clutter reduces storage costs and makes it easier to discover the high-value, relevant content that can generate new business value. Defensible disposition has become the antidote to escalating storage costs, information discovery challenges, and risk.

Big data analytics projects, however, benefit from scale, and big volumes of unstructured information are required to train cognitive systems. Because it's difficult to predict just what might be relevant for big data analytics down the road, big data proponents are inclined to keep everything – just in case it turns out to be useful in the future.

Finding the happy medium between these two extremes requires creating a healthy dialogue between the two camps. Initiating that dialogue, however, requires deep insights into the information the organization currently possesses. Some of that information will be easily discarded as irrelevant,
redundant, out of date, or of poor quality. Some of that information will be deemed valuable for ongoing business operations and analytics. And some of that information will be considered potentially valuable but problematic because it contains personally identifiable or confidential company data.

A sensible information governance approach enables diverse stakeholders to collaboratively decide the optimum course. A balanced approach typically includes sanitizing potentially useful information by redacting personal or confidential information that – if disclosed – would create risk. That way, the organization is protected while it benefits from the use of that information in big data analytics. Achieving this happy medium requires a common solution.

IBM STOREDIIQ

IBM StoredIQ helps organizations address the myriad challenges they face around the effective governance of unstructured information – challenges that have proven daunting and extremely costly to address using other approaches. StoredIQ gives organizations the comprehensive solution and methodology they need to establish sound and defensible information governance practices that not only address retention, risk, and eDiscovery needs but also position their big data analytics projects for success.

How StoredIQ Works

Enterprise content is highly fragmented today, often across multiple content repositories, team sites, shared drives, cloud services, enterprise applications, users’ hard drives, and other locations. StoredIQ uses a combination of rules and machine learning to identify and categorize content – regardless of type or location – including the "dark data" that organizations don't even know exists. That "dark data" can put the organization at significant risk of non-compliance with retention requirements, non-compliance with information privacy regulations, leaks of sensitive or confidential information or intellectual property, and even litigation due to over-retention.

Once information assets are identified and classified, StoredIQ enables stakeholders to understand those assets. StoredIQ visualizes what can be a daunting amount of data about the organization's information assets in highly intuitive heat maps that give users from legal, compliance, records management, IT, and other groups an at-a-glance understanding of the organization's content (see Figure 3). This helps diverse groups get on the same page when it comes to finding the balance between conflicting information governance needs – even as those needs evolve and change.
StoredIQ then helps stakeholders prioritize and take action – whether that means securing confidential information, optimizing tiered storage, applying retention policies to regulated content, or disposing of redundant, out-of-date, and toxic information. One of the strengths of StoredIQ is that it supports an iterative approach to improved information governance. That is, organizations can start with a limited scope or specific area and then expand in successive iterations. StoredIQ is much more than a remediation solution: Customers rely on StoredIQ to monitor and manage their unstructured information on an ongoing basis, increasing business agility and peace of mind.

Given its unobtrusive footprint, StoredIQ has little or no impact on running systems or IT service-level agreements (SLAs). StoredIQ rapidly indexes information in place and at scale – providing rapid time to value and eliminating the need to copy or move information.
Benefits of IBM StoredIQ

StoredIQ enables organizations to discern "good data" from "bad data" and improve the outcomes of their "big data" projects. Benefits of implementing StoredIQ include:

- **Improved insights, better decisions.** Using StoredIQ, organizations can maximize the value of their unstructured information by putting it to work in big data analytics systems. StoredIQ helps ensure that the information consumed by big data analytics applications is of high quality, and it enables organizations to maximize the potential of their unstructured information while minimizing the risks associated with over-retention or the use of information that contains confidential or personally identifiable data.

- **Improved compliance with retention requirements.** StoredIQ's automated classification enables organizations to quickly and accurately identify information that is subject to regulatory and board-mandated retention requirements — for both remediation and ongoing compliance assurance.

- **Defensible disposal.** StoredIQ's proven methodology, rich content intelligence and classification capabilities, and auditability give organizations the automated policy management they need for defensible disposal.

- **Better targeting of relevant information for litigation or audit.** StoredIQ helps organizations accelerate their collection efforts, ensure the completeness of the information collected, and reduce their external review costs.

- **Merger, acquisition, and divestiture support.** StoredIQ gives organizations the insight they need to effectively onboard content from acquired entities and accelerate consolidation — or offload content from divested entities and accelerate time to close.

- **Improved operational efficiency.** By enabling organizations to identify and confidently delete low-value, out-of-date, and redundant information, StoredIQ helps reduce storage costs, reduce backup time/costs, streamline data migration tasks, and reduce bandwidth costs for cloud migrations. StoredIQ also gives IT organizations valuable information for planning infrastructure investments.

- **Better SharePoint team site governance.** As noted previously, many organizations continue to struggle to define disposition strategies for content in SharePoint team sites. Large organizations have thousands (and sometimes tens of thousands) of user-provisioned team sites — many of which contain information of uncertain value and relevance. StoredIQ gives organizations the insight they need to define appropriate disposition strategies for SharePoint team site content and reduce costs.

- **Increased information worker productivity.** StoredIQ enables organizations to identify and eliminate the clutter that makes it so difficult for information workers to find the high-value, relevant information they need.
IBM Watson Curator

IBM recently announced a new SaaS packaged offering called IBM Watson Curator, which includes StoredIQ technologies. Companies using Watson Engagement Advisor for their big data analytics projects should consider this companion product.

With Watson Curator, IBM is extending the concept of the data refinery to unstructured information. Using Watson Curator, business users can quickly identify and collect relevant, trustworthy content to form the information collections that they need to make their Watson projects successful. Users can work collaboratively and iteratively to refine their collections; and Watson Curator serves as a system of record with governance, documenting precisely what information was used in each analysis.

Using the capabilities of StoredIQ, Watson Curator ensures that users have a complete view of all of the content that is available for their analyses – along with an assessment of the quality of that information and its sensitivity. This gives users greater confidence in their content collections, and they can readily determine whether the content they are using for their analyses requires redaction or scrubbing for personally identifiable or confidential information.

As we have noted previously, good information governance is a key requirement for managed and optimized use of big data analytics. IBM Watson Curator helps make good information governance for Watson projects a core competency.

CHALLENGES/OPPORTUNITIES

To be sure, successful big data analytics projects require more than high-quality information: They require the effective synthesis of information of different kinds from many different sources. They also require new skill sets, technologies, and processes. Operationalizing insights gleaned from big data analytics also entails cultural change – in addition to changes to existing systems. Nonetheless, we expect investment in big data technologies and services to continue at a rapid pace over the next several years given the strategic advantages that big data analytics can confer on organizations that adopt them.

As organizations expand their use cases for big data analytics – and mature as data-driven organizations – they need tools to help them identify, classify, and determine the value of the information they possess. This will be critical to finding the balance between keeping everything (in case it could prove valuable for big data analytics) and disposing of everything that isn't business critical or subject to legal hold and retention policies. Information governance tools will also be critical to appraising the value of information used in big data analytics projects and devising strategies to refactor – or sanitize – that information when it contains data that needs to be expunged.

This bodes well for IBM, already a leader in information management and analytics. The immense challenges that organizations face managing, retaining, and defensibly disposing of their unstructured information should ensure that IBM's StoredIQ solution continues to find a ready market. Growing investment in big data analytics projects – and the rise of cognitive computing – should further enhance StoredIQ's appeal.
To reap the full benefits of solutions such as StoredIQ, organizations should seek to make good information governance a core competency by establishing centers of excellence. Finally, StoredIQ is designed specifically for unstructured information: Organizations with structured data quality problems will need to seek out complementary solutions such as IBM InfoSphere Optim.

**CONCLUSION**

Good information governance practices and solutions help ensure the success of big data initiatives by enabling enterprises to discover, classify, and manage information according to its business value and triage "good data" (relevant, current, and trustworthy business information) from "bad data" (out-of-date, obsolete, or low-value information).

Bringing the organization's unstructured information under good governance requires a solution that can automatically discover and classify content regardless of type or location. In addition to making "good data" easier to find and ensuring information quality for big data analytics projects, information governance pays dividends by reducing storage and eDiscovery costs and reducing risk. In particular, an information governance solution enables the organization to identify sensitive, confidential, private, and toxic information that is inappropriately managed or should be deleted. This is an important aspect of the ability of an organization to demonstrate that it has robust processes in place to protect and preserve information that is subject to regulatory control.

Only a small percentage of an organization's unstructured information is subject to legal hold or regulatory retention. By some estimates, just a quarter of an organization's unstructured information has current business utility. Good information governance not only helps safeguard valuable business information and ensure that legal and regulatory requirements are met but also helps the organization determine what to do with the roughly 60-70% of unstructured information that may or may not be useful.

In the past, common wisdom suggested that disposal was the best policy. Today, in the era of big data analytics, that's not so clear.

Striking the balance between disposal and preservation — between minimizing clutter and risk and optimizing the potential for business insight through big data analytics — requires a deep understanding of enterprise information assets and an effective management strategy that factors in the needs of a very diverse set of stakeholders. Without the right information governance solution, this is an impossible task.

Good information governance needs to be a core competency in any organization contemplating investments in big data analytics projects. The outcome of a big data analytics project will be only as good as the quality of the data being used. Organizations must be able to ensure that the unstructured information they leverage in their big data analytics projects is relevant, complete, and trustworthy. This will become increasingly important as big data analytics are operationalized to drive decision making and optimize core business processes.
Bringing unstructured information under good governance pays huge dividends above and beyond helping ensure the success of big data analytics projects. It enables organizations to identify and dispose of out-of-date and non-business information, saving costs. It also helps organizations discover and remediate toxic and sensitive information – information that puts them at risk. These benefits should resonate strongly with the legal, compliance, privacy, records management, and IT professionals who are chartered with safeguarding enterprise information.

Organizations should assess their current information governance practices in light of the following questions:

- Is there visibility into all enterprise information, regardless of where it is stored – including enterprise repositories and team sites, shared drives, email systems, users’ desktop and laptop computers, and cloud applications?
- Is it difficult to find or identify high-value information because of information clutter?
- Are storage, backup, and eDiscovery costs escalating as the volume of information grows?
- What is the level of confidence that confidential, sensitive, and private information is adequately protected? Is the organization at risk of inadvertent disclosure of confidential information or of non-compliance with privacy regulations?

If one or more of these are pain points for an organization, then it's time for the company to get its unstructured information under control – especially if it is considering investing in one or more big data analytics projects.

Deciding where to start can be daunting. Most organizations benefit from an incremental, iterative approach. IDC recommends that organizations evaluate IBM's StoredIQ solution.
About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-insights-community.com
www.idc.com

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2015 IDC. Reproduction without written permission is completely forbidden.