A DataGravity White Paper:

Unstructured Data: Friend Or Foe?

THE CHALLENGE IS FINDING VALUE

CONTENTS

DataGravity, Inc

100 Innovative Way, Suite 3410
Nashua, NH 03062
P 603.943.8500

info@datagravity.com
datagravity.com

DATAGRAVITY

@DATAGRAVITYINC

DATAGRAVITY

WELCOME TO UNSTRUCTURED DATA

The era of Big Data is here and it's rewriting the rules for how corporate IT stores, manages, replicates, protects, mines, and analyzes the volumes of data that, even five or 10 years ago, were once unthinkable in size, complexity, and format. *Every day* the challenge grows, and every second the data changes. The biggest driver? *Unstructured data* – human-created text and a litany of file types.

In years past, IT largely revolved around database-centric or spreadsheet-centric data that fit neatly into rows, columns, and tables. That data had rules, definitions, and objective characteristics and formats. It was based on discrete transactions that could be aggregated, sliced, and analyzed: credit-card transactions, telecom call records, diagnosis codes, P&L statements, stock trades, commodity prices, and application records (e.g. ERP). For decades, IT has gradually evolved into teams of specialists – there are storage specialists, virtualization specialists, and database and application administrators – all ultimately aiming to manage structured data. Today? In many cases, that paradigm is being overwhelmed – massively and rapidly – as customers, suppliers, and employees create and consume unstructured data in many forms and formats.

- Written documents
- Spreadsheets
- PDFs
- Text files
- Presentations
- Email
- Tweets and Facebook posts
- Images
- Audio and video

The fact is, unstructured data may be the world's best-kept secret as companies have been slow to recognize that valuable information lies within the documents, emails, and social streams that are being constantly created. The value of structured data continues to grow, of course – but the new battleground for competitive advantage revolves around unstructured data.

As the Big Data gold rush progresses, the low-hanging fruit – the analysis of structured or semi-structured data – isn't yielding the competitive advantages it once did. Organizations are increasingly turning their focus to unstructured data – sources that were once considered too difficult to effectively use – to improve efficiency, reduce rework, and uncover insight from across the business.

THE SCOPE OF UNSTRUCTURED DATA

While we've all seen some of the eye-popping statistics, it's still valuable to pause and consider a few more. According to IDC, in 2011, there were 1.9 zettabytes of electronic data in the world – enough to fill more than 50 billion iPad 32GB tablets. By 2015, that's expected to more than quadruple to 7.6 zettabytes.

Office workers spend an average of 2.6 hours per day reading and answering emails, according to a survey by McKinsey Global Institute. That adds up to about 27 days per year.

Time spent reading and answering e-mails

2.6 HOURS
ON AVERAGE PER DAY

27 DAYS
ON AVERAGE PER YEAR

A recent survey by Microsoft found that 55 percent of office workers report they often experience information overload while 43 percent feel stressed as a result - and 34 percent describe themselves as overwhelmed. Knowledge workers – people who sit at desks for a living – spend an average of 20 percent of their time each week looking for unstructured data or re-creating it if they can't find it.

> *Within this massive and growing mountain of data lies exceptional, transformational value*

Companies of all sizes and industries understand that within this massive and growing mountain of data lies exceptional, transformational value. Unfortunately, unstructured data presents significant challenges to corporations seeking to tap that value.

- **Not Machine Readable –** Unstructured data is typically produced for human consumption and is not, generally, easily interpreted by a computer. It contains an ambiguity of language, subtleties, nuances – even hyperbole and sarcasm. And it can be presented in any of dozens of languages –all of which makes it difficult for a computer to parse, decode, and process.
- **Difficult to Automate –** Unstructured data doesn't lend itself to repeatable or automated processes. It's difficult to identify the attributes/metadata that's most interesting and valuable to the organization and link unstructured sources to structured ones. E.g. How do you relate social media posts against sales performance? How did your technical marketing affect the number and severity of support cases?

Dark Data: An Episode of Corporate 'Hoarders'

Like the reality series, many companies' IT systems resemble an episode of Hoarders, thanks to "dark data." Typically, most dark data is unstructured data, a series of documents, tables, databases, spreadsheets, and other files that aren't well-managed, monitored, or maintained (file shares are great hiding spots for dark data). Some examples include:

- Data in the shadow IT organization – Pockets that you're unaware of
- Dropbox accounts that move data in and out of firewall
- Huge presentation files created and used by a former employee
- Mailing databases that haven't been updated for years

Data usually "turns to the dark side" because its hidden value or risk are not generally known or the data is difficult to access or transform. That turns an asset into a cost – especially as the pile of dark data grows over time. Organizations should redouble their efforts to shine a light and analyze their dark data to determine whether it is worth developing and extending.

- **A Proliferation of Sources –** Simply navigating the access and security frameworks to bring the disparate formats together is a major undertaking.
- **A Proliferation of Formats –** From Google Docs to OpenOffice to Microsoft Office, and more, there are countless formats and standards that are conflicting – and changing frequently.
- **Questionable Accuracy –** As a result, there are endless concerns about the reliability and accuracy of unstructured data. For instance, data in an in-process contract is inherently less reliable than data in a completed/closed contract. Or Word files marked "draft" have lower reliability than those marked "final" in a different folder.
- **Unstructured Locations –** While there are many exceptions, structured data is largely housed in data centers that are managed by IT professionals. By contrast, unstructured data seems to live almost everywhere – in the data center, on user laptops, on tablets and smartphones, and in the cloud. Too often, "unstructured" means "unmanaged," representing significant risk exposures or lost hidden value to the organization.

Some IT veterans may harbor doubts about the value of unstructured data. After all, its initial value is elusive since it's less quantifiable than structured data. But it's important to remember that employees – knowledge workers, in particular – spend the vast majority of their time working with unstructured data – and that's a bidirectional dynamic that holds true for customers, of course, in both B2B and B2C domains, with data moving inside and outside the organization. Remember, some of the most important documents in the company are stored in unstructured formats, from corporate strategy documents and product roadmaps to sales plans, resumes, decision records, customer records and more.

Some of the most important documents in the company are stored in unstructured formats

What's more, not all of the interesting information is created inside your walls. Support cases, knowledge bases, RFPs, social interaction, and more – they can all start as external unstructured data.

It's no longer debatable: IT faces a new imperative: harnessing the untapped power and value of unstructured information – the information that the vast majority of knowledge workers interact with on a daily basis. Responding to those challenges will require new tools, new strategies, and new thinking. From simply managing the crushing volumes of information and devising automated processes to responding to the plethora of formats and formats – IT must adapt to this rapidly changing landscape.

OVERCOMING THE LIMITATIONS

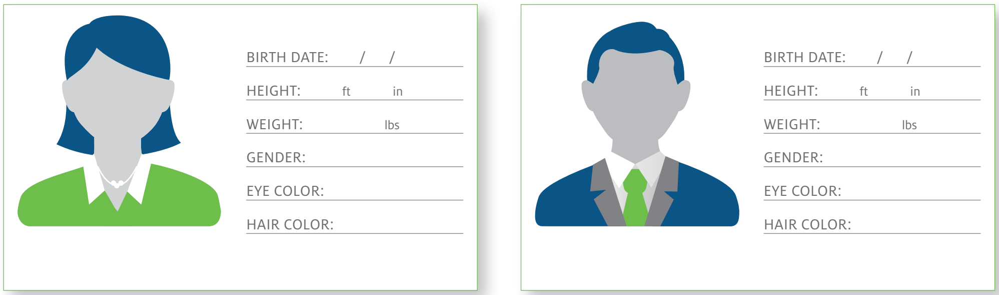
One of the challenges of unstructured data and its exponentially increasing volume is that it skews the "signal-to-noise" ratio. Simply put, it's harder for companies to find the needle in the haystack. In contrast to the typical/classic Internet search problem where Google attempts to bring you "*an* answer," business users need "*the* answer," which may exist in a single document or email chain. You want to make better decisions – and, invariably, those are fueled by more and better data. Today, there are more inputs for the same disciplines and decisions, which means companies must whittle through and eliminate much more "noise" to find newer and higher-value "signals" (whether it's the right document or the right aggregation of soft data) to make optimum business decisions.

One way we can do that is by addressing the unstructured data's *context*. That starts by using the metadata that accompanies the unstructured data to help identify how it fits into the many data silos:

- **Users, creators, owners, and editors**
- **File types**
- **File properties**
- **Content**
- **Extracted metadata**

What is MetaData?

Metadata can be seen as a set of attributes or data points that can be used help to describe or classify an object. For example, when looking at a person you could describe that person by age, occupation, educational background, birthdate, hair color, eye color, personality type (extrovert/introvert), etc. In the case of business content, you could assign attributes for including type of content, intended audience, the file format, etc.



For instance, a sales rep who's engaging with a client to renew a contract may want to know a broad range of interesting data points:

- **How many support cases were opened by the client?**
- **Who was involved in the last contract?**
- **What customer-service rep has been most involved?**
- **What discount schedules were applied previously?**
- **What is their level of satisfaction?**

A rich querying capability is essential for searching, navigating, exploring, and discovering within unstructured data.

"Discovery" is a particularly important concept, in that unstructured data often means that users learn things that are fundamentally new and unexpected – insights that weren't initially sought or predicted. You don't query about something you didn't know existed.

Text analytics is one key discipline that can help. It can encompass a variety of sophisticated techniques, including sentiment identification, language identification, topic, entity, and fact extraction, and many more. These routines identify the topics, languages, and other metadata points and offer important ways to give unstructured data its essential context. Regardless of the type of unstructured data, it's important to consider the keys:

- **Tracking –** Where did the data come from? How fresh is it? Do you have confidence in the source(s)?
- **Inspecting –** Parse the content to find useful metadata, such as sensitive fields (e.g. Social Security numbers or credit card numbers), customer interactions, and links to structured data.
- **Accessibility –** The utility of text-based unstructured data hinges on users finding and re-using it. The discovery process helps you identify experts and high-value content that matters to you – everything from budgets and payroll figures to strategy and goals. Ultimately, the goal is to increase the usability of the unstructured data by finding and relying on its curators *and the downstream users* who add and derive further value. Even better, a sophisticated system can learn from your activities to make your results more relevant.
- **Visualization** - Similarly, visualization tools help users literally look at their data from new perspectives to gain new insights and draw new conclusions. That can encompass two distinct abilities – the ability to quickly draw conclusions and identify patterns from large sets of data as well as the ability to quickly drill into specific data pathways to get more details.
- **Consolidation vs. Decentralization –** There's a tradeoff between unifying your unstructured data and abiding by a "data-in-place" strategy. Consolidation is the best practice to facilitate better data management. But with cloud platforms, shadow IT organizations, and the many places that data lives, strategies may be needed to support data-in-place paradigms that leave owners in control of their data. In any case, a meaningful file-sharing environment will help you get a handle on what you have. Although there will still be pockets of unstructured data from websites, social media, and other sources, it's still a smart strategy to build and maintain an on-premise, consolidated store of unstructured data that can be properly managed with appropriate retention/deletion policies – to prevent that data from turning into dark data.

Gartner’s Take:
File Analysis - Knowing What You Have

File analysis is Gartner’s term for one of the key disciplines in the world of unstructured data. With traditional structured data, file analysis involves simple reporting on a files’ attributes – such as size, file type and date of creation/update. With file analysis of unstructured data, you can also index, search, and report on each file’s detailed metadata and contextual information to improve our storage management, compliance/ governance actions, and decisions. This can even involve some analysis of the file content itself.

Properly analyzing and understanding unstructured files reduces corporate risk by identifying locations, the owners of files, and who has access to them. That enables companies to lessen risk exposures regarding personally identifiable information, tighten controls on intellectual property and sensitive data, and eliminate outdated or conflicting data (e.g. different versions of contracts). Analyzing this file-based data can also support litigation and due diligence processes, which rely heavily on unstructured documents that contain valuable information.

While most sources of structured data are inherently defined and understood, unstructured data must be carefully analyzed, monitored, and managed to deliver optimum value. File analysis can support numerous corporate processes:

- Classifications for information security
- Enforcing information retention and governance policies
- Archiving protocols
- E-discovery and litigation
- Business reporting
- Optimized storage management
- Data center consolidation
- Migrating to the cloud
- M&A data consolidation
- Data deletion, deduplication, and cleanup

Investments in analyzing this file-based data pays dividends in other ways –employees across the company can more easily find the data they need, when they need it. This improves efficiency, and reduces time-to-response, which can be the difference between winning a new account and losing to the competition.

THE NEW DIFFERENTIATOR: MASTERING UNSTRUCTURED DATA

In this changed and evolving environment for corporate data and decision-making, the advantage will accrue to the companies that can quickly and effectively take charge of and find the value within their unstructured data. That hinges on several key dimensions:

- **Faster Time-to-Data –** With its huge volumes and free-form nature, unstructured data can be a difficult discipline to harness. The challenge is to bring the data to decision makers with efficiency, agility, and accuracy. By enabling knowledge workers and executives alike to see new insights sooner, competitive advantages and opportunities emerge sooner and at higher value.

- **Data-Driven Decision-Making –** Leading companies continue to evolve from pure "gut instinct" to infuse structured data analysis into their decisions, and using traditional rows and columns of data to inform these analyses has brought significant rigor and benefit. But we mustn't leave behind the important value that the company derives by integrating unstructured data into its decisions as well. Soft metrics, subjective measures, and the vast majority of data that can't be compressed into tables and pie charts can have an outsized impact on company performance. In fact, MIT researchers have found that decision-making driven by unstructured data can be directly responsible for gains of corporate performance of 5-6 percent – a sizeable gain.
- **Data Quality –** The same principle holds true with unstructured data as with structured data: Your decisions and outcomes will only be as good as the data you base them on. It's essential to understand and account for the accuracy, reliability, and consistency of the unstructured data across your organization.
- **Measurement –** To understand the net gains from insights based on unstructured data, it's important to benchmark the state of your data and your business performance prior to pursuing initiatives. That benchmarking starts with focusing on a specific business problem and identifying the data you're using *and* the untapped (typically unstructured) data that's available to you – including so-called "dark data" (see sidebar).

CONCLUSION: UNCOVERING THE BEST-KEPT SECRET

Companies are sitting atop a vast reservoir of untapped value – the zettabytes of unstructured data that can drive key processes and strategic decisions to achieve competitive advantage. Unstructured data moves the company beyond traditional decisions based solely on rows and columns and tables of numbers, embracing the documents, spreadsheets, presentations, emails, and social media posts that can house vital information and insights. In short, companies are now being challenged to leverage the "best-kept secret" in corporate IT.

> *Companies are now being challenged to leverage the "best-kept secret" in Corporate IT – a vast reservoir of untapped, unstructured data]*

Unlocking that value will require IT to respond to challenges such as parsing, automation, formats, sources, integration, accuracy, and increasingly complex storage infrastructures. One key strategy is to address the data's context through file analysis to analyze its metadata and by using tools to search, navigate, explore, and discover unstructured data. Text analytics and visualization also play prominent roles in successfully leveraging unstructured data. The result is a faster time-to-data and an increase in data-driven decision-making that can significantly improve corporate performance.