WHITE PAPER

The Modern Data Imperative: Staying Ahead of Critical Data Challenges

How big data analytics can transform your ediscovery landscape.

Contents

Introduction

The current state of ediscovery is complex, inefficient, and cost-prohibitive as data formats and volumes continue to explode without bounds. Organizations of all sizes are bogged down in enormous amounts of electronically stored information (ESI) that must be managed during ediscovery. And even with widely used technology, a staggering amount of non-responsive and duplicative data continues to persist for review – the most expensive phase of ediscovery.

Right now, teams start from square one every time a new legal matter or investigation arises. Potentially relevant data is collected from different sources into a single database, where legal teams broadly cull irrelevant data before spending hundreds to thousands of hours reviewing documents one by one, manually categorizing and flagging for responsiveness, privilege, or data that must be redacted before being produced. Each time a new matter comes up, this inefficient process is repeated. Often the same data is collected, culled, and reviewed again from previous matters, duplicating the cost without adding value.

Organizations typically hold a wealth of previously reviewed data from past cases that sit unused after the close of the case or investigation. With this current approach, review insights remain trapped in silos. Without a portfolio-level view, there is loss of actionable insights, unnecessary costs, and increased risk. With tight deadlines, increasing data regulation, and the pressing need to efficiently handle a large volume of data, organizations must find a way to reap insights across their matters.

This overall ediscovery story is a big data analytics problem. The good news is that a significantly improved picture comes into focus with the implementation and use of cutting-edge analytics data. Effectively using a robust big data analytics product will enable organizations to revamp their inefficient and non-repeatable ediscovery workflows. A powerful big data analytics tool that utilizes artificial intelligence (AI) and machine learning can create customized data solutions by applying a number of models for a comprehensive analysis of the data, and by enabling knowledge transfer from prior cases onto current and future matters.

To understand how big data analytics can be used to create an adaptive workflow and permanently improve the ediscovery process, let's examine how the use of big data analytics methodologies optimize workflows across compliance, ediscovery, and information governance. We'll discuss the current data challenges facing the legal industry today, the importance of understanding the right solution, and the real-world application of big data analytics in ediscovery.

Challenges of Data in Legal

With the rapid increase and changes in data being created today, organizations across industries are seeing not just the benefits, but also the need, for moving to products and processes that directly address issues that arise with this influx of data. The legal industry is no different. Gone are the days that an extra swath of documents can be reviewed, just in case. Data in ediscovery presents numerous challenges as data formats and volumes grow, and increasing with them is the cost of review – already the most expensive phase of discovery.

As data generation continues to surge, organizations must use products built for big data to effectively manage the efficacy of review. Without the ability to transfer knowledge and learnings between matters, the same data is reviewed over and over, generating additional cost without generating additional value. What's more, this is exacerbated without implementation of a purpose-built product. There are four primary challenges when it comes to analyzing big data – also known as the 'four Vs' of data:

1. The volume, or scale of data
2. The variety, or different forms of data
3. The velocity, or analysis of data
4. The veracity, or uncertainty of data

Let's take a deeper look at each of these and how they factor into ediscovery.

THE VOLUME/SCALE OF DATA

175 ZETTABYTES (175 TRILLION GIGABYTES) OF DATA WILL BE CREATED BY 2025, GLOBALLY[1]

MOST COMPANIES IN THE U.S. HAVE AT LEAST **100 TERABYTES** (100,000 GIGABYTES) OF DATA STORED[2]

AN ESTIMATED **2.5 QUINTILLION BYTES** (2.3 TRILLION GIGABYTES) OF DATA ARE CREATED EACH DAY[3]

In the era of big data, the amount of data generated is a significant and growing issue for ediscovery cases. By 2025, worldwide data will grow 61 percent to 175 zettabytes[4], and 49 percent of the world's stored data will reside in public cloud environments[5]. As more and more data are generated every second of every day – including the large amounts of data being generated beyond the traditional realm of email, Word, Excel, PowerPoint, and PDFs – organizations are forced to consider better ways to retain and analyze their data. It's no longer practical to keep everything and rely on old-fashioned searches and analytics to find what you need.

Those who successfully manage this volume of big data will see measurable benefits in cost and time. This is particularly relevant to ediscovery, where the volume of legal holds, collections, processing, and hosting has grown exponentially, and would greatly benefit from a proactive solution that maximizes efficiency on current and future matters. Additionally, the sheer amount of data will make consistency difficult to achieve without an overarching view.

THE VARIETY/DIFFERENT FORMS OF DATA

6.95 BILLION PEOPLE HAVE CELL PHONES[6]

THE NEW YORK STOCK EXCHANGE CAPTURES **1 TERABYTE** OF TRADE INFORMATION DURING EACH TRADING SESSION[7]

THE NUMBER OF DEVICES CONNECTED TO IP NETWORKS WILL BE **MORE THAN THREE TIMES** THE GLOBAL POPULATION BY 2023[8]

In addition to the volume of ESI dramatically expanding, the diversity and variety of the data are significantly increasing. This is a big piece of the challenge involved with managing big data – the varying kinds of data the world is now generating. Gone are the days in ediscovery where the biggest challenge was processing and reviewing computer-based data such as email, spreadsheets, and documents.

Now, employees are regularly communicating across a variety of devices and channels, and ESI encompasses everything from email to mobile messaging and a new universe of collaboration tools like Microsoft Teams, Instant Bloomberg, and Slack. Any one of these would prove challenging – however, now consider that the average corporation maintains dozens of cloud-based collaboration tools. It's easy to see how these new data types represent the most complex part of the scope of discoverable data. This next phase of ESI is very different, but it can be just as relevant, and implementing an advanced analytics platform is the key to managing all new data types in an efficient and cost-effective manner.

THE VELOCITY/ANALYSIS OF DATA

AS OF 2011, THE GLOBAL SIZE OF DATA IN HEALTHCARE WAS ESTIMATED TO BE **150 EXTABYTES** (161 BILLION GIGABYTES)[9]

MICROSOFT TEAMS HAD **44 MILLION** DAILY ACTIVE USERS IN 2020[10]

SLACK HAD **12 MILLION** DAILY ACTIVE USERS IN 2019[11]

Contending with large amounts of data creates another significant issue: how to handle the velocity of the data – the speed in which the data is generated as well as the rate at which that data is processed for collection and analysis. The old way of managing data is to put everything into a database and analyze it later. But in the era of big data, the old way is expensive, time consuming, and can miss key information compared to big data analytics-enabled real-time analysis. This is especially true for compressed ediscovery timelines, as in Hart Scott Rodino (HSR) Second Requests for mergers and acquisitions and the often-accelerated court deadlines.

Implementing a big data analytics product that can rapidly process and analyze iterative volumes of collected data on an ongoing basis will allow for an immediate application of findings. The faster you can process this data into an effective big data analytics platform, the greater flexibility you will have to analyze it and quickly find the most relevant documents. Analyzing documents at the outset for relevance also means savings in hosting costs.

THE VERACITY/UNCERTAINTY OF DATA

POOR DATA QUALITY COST THE U.S. ECONOMY AROUND **$3.1 TRILLION** IN 2016[12]

1 IN 3 BUSINESS LEADERS DON'T TRUST THE INFORMATION THEY USE TO MAKE DECISIONS[13]

IN ONE SURVEY, **27%** OF RESPONDENTS DID NOT KNOW THE ACCURACY OF THEIR DATA[14]

IN 2017, **ONLY 3%** OF COMPANIES' DATA MET BASIC QUALITY STANDARDS[15]

When regularly collecting, processing, and generally amassing large amounts of data, it is most imperative that the data is reliable and trustworthy. In the case of ediscovery, veracity refers to the quality of data. Data must be accurate and untainted before it can be analyzed to deliver accurate insights for use in making decisions.

The veracity of data depends on knowing how to make the right decisions during collections and processing. To verify the veracity of data in ediscovery, correct collection must be ensured, as well as a check for missing information based on any errors made during collection and processing.

By truly understanding the veracity of your data, you can improve the accuracy of your data analysis and make correct decisions for the scope of ediscovery based on that analysis.

Individually, these challenges cannot be appropriately addressed using traditional methods and will only grow as worldwide data creation continues to surge. Combined, they pose new, even greater obstacles that must be addressed with cutting-edge technology, built specifically to address these issues.

Applying Big Data Analytics-Based Methodologies

There are a multitude of benefits that come from analyzing ediscovery data within and across matters. As detailed in the previous section, there are many challenges that come with data as we know it today. Adopting methodologies specifically for big data are necessary to get the most out of your data today and tomorrow.

Benefits of leveraging big data analytics include:

- Pre-classification to efficiently pinpoint important documents for review.
  - Build highly effective privilege screens based on previously coded privileged documents;
  - Quickly detect personally identifiable information (PII) and protected health information (PHI) for redaction; and
  - Protect sensitive company information, such as source code and trade secrets.
- Portfolio-level view to unlock value and impose efficiencies across matters.
  - Document accountability to improve consistency;
  - Reduce cost and timeline by cutting document re-review; and
  - Removing frequently occurring yet objectively non-responsive documents commonly pulled into reviews.
- Reporting to inform ediscovery workflow and information governance optimization decisions.

PRE-CLASSIFICATION

Documents that occur frequently across matters can be identified and pre-classified to decrease risk and save the cost of attorney review. This can be done using the algorithms created by data scientists either out-of-the-box or honed using a company's historical matter data. A key benefit of using big data analytics products is the ability to establish effective rules and processes for data that are statistically supported. An example of such a rule or process is correlation analysis: identification of key actionable rules between document features and coding decisions (e.g., documents with more than Y number of recipients are never privileged).

High-ROI use cases for pre-classification include:

- **Privilege documents** – Documents that are coded again and again as containing privileged information can most likely be directly coded as privileged. This will decrease privilege review costs, even if some documents are sampled to confirm accuracy of the analysis.
- **PII/PHI** – Information that must be redacted to protect individuals and to comply with data regulation rules. Quickly and consistently identifying and redacting this information is critical to maintain compliance.
- **Sensitive information** – Information such as source code and trade secrets that is critical to remove from production, where it could end up in the hands of competitors. At the same time, review is generally conducted by counsel who have more subject-matter expertise in the company's industry, which comes at a high cost.

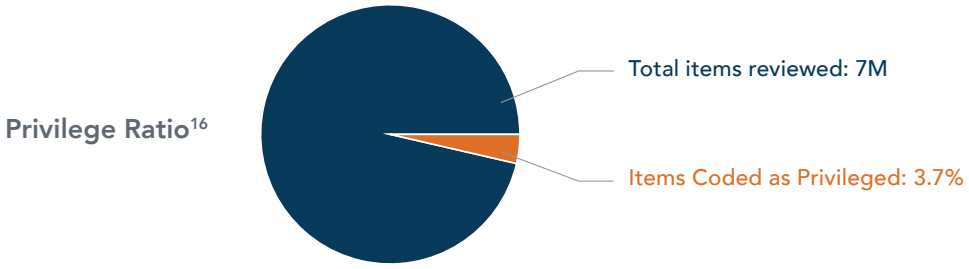
Below, we'll further explore the use of pre-classification for privilege documents as a specific example of how early detection can benefit companies.

PRIVILEGED DOCUMENTS

The current process of using search terms to find privileged information is highly imprecise and inefficient, leaving a significant opportunity for improvement through predictive modeling.

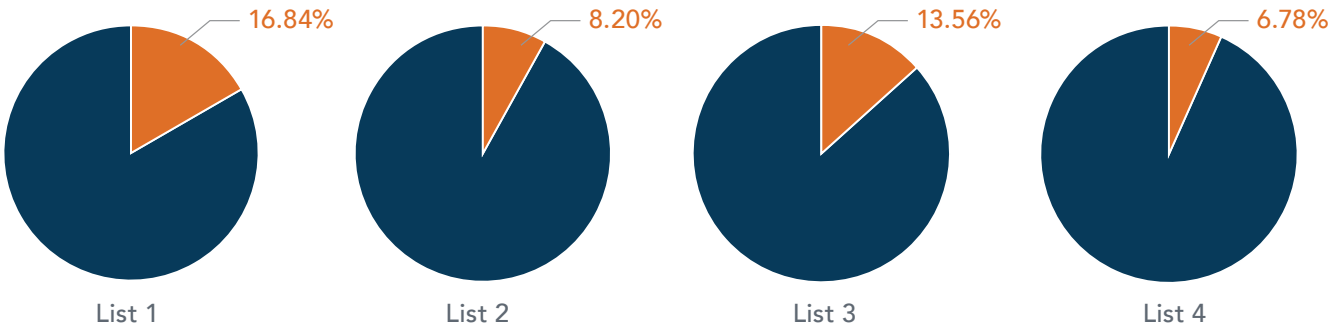
To illustrate, below is a general privilege analysis showing how an optimized data-driven privilege review process compares to the current baseline, using data from a pharmaceutical company's matter portfolio.

Using search terms to find privilege, more than 96% of the nearly 7.7 million documents reviewed across all of the company's matters were done so unnecessarily.



Percentage of documents coded as privileged

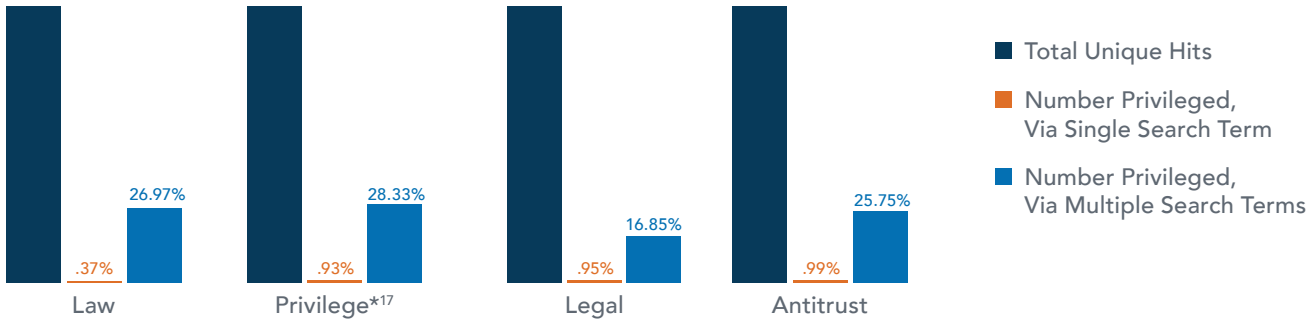
The performance for general keyword lists varies greatly and is still overly broad. Some of the more precise search term lists are shown below:



Percentage of documents coded as privileged per keyword list

Even looking within one of these lists, it is obvious that precision across terms varies, and all are subpar. Most of these terms on their own are less than 1% precise in finding actually privileged information.

Sending the lists through a big data analytics tool revealed that the precision of these commonly used search terms can be significantly increased by combining them with another term, thereby optimizing the search term process.



Percentage of documents coded as privileged per term and when combined with another term

Further, using data from prior matters, such as prior attorney work product, combined with these more precise search terms will pinpoint actually privileged documents more effectively, efficiently, and consistently.

PORTFOLIO-LEVEL VIEW

Using big data analytics technology allows the team to have one view over siloed data sets to bring consistency across the entire portfolio. Review teams must deliver consistent coding and tagging decisions even when reviewing thousands to millions of documents. Having a view into document treatment across a portfolio can help bring work product consistency and can even feed into process improvements based on trends.

Taking a cross-matter view enables the transfer of knowledge between cases and identification of junk documents, in addition to honing the AI algorithms to analyze data.

Tapping into prior work product brings efficiency to current review. A big data analytics program can identify opportunities to reuse work product from a duplicate document that was previously reviewed, and also provide details to help make a judgment call on when to apply prior work product. This is particularly compelling in cases of multidistrict litigation or other matters stemming from similar issues, where there could be few differences between document collections and responsiveness across matters.

In the example below, nearly a third of documents contained in a matter at hand have already been reviewed, and nearly 19% of documents were reviewed in multiple matters, representing opportunities to remove the documents from review.

Reuse of Work Product

Items with cross-matter duplicates	29.33%
Reviewed in multiple matters	18.91%

Percentage of documents that have been previously reviewed

Junk documents are often included due to a search term hit or family relationship, even though they have no relevance. These will be routed for review by the current search term-based screening method, but they can be removed when analysis reveals that the documents have never been found to be responsive over time. These documents are eligible to be diverted from review.

For example, if an email was included as having potential relevance due to a search term hit, any attachments, as well as embedded documents within those attachments, would also be included, regardless of actual relevance.

Early identification of these documents means cost savings. Below is an example of documents – identified by their hash values – that can be removed from review, as they have already been reviewed thousands of times without being found to be responsive.

Hash Value	Count	Workspaces	Responsive	Ratio
Hash 1	19,052	3	0	0%
Hash 2	7,910	5	0	0%
Hash 3	4,877	6	0	0%
Hash 4	2,744	4	0	0%
Hash 5	2,466	5	0	0%

Number of reviews per document with ratio of responsive coding

REPORTING

With a big data analytics product in place, you can gain a better understanding of your legal and data landscape at different levels. From a detailed view of a matter portfolio, to benchmarking in order to add measurement to process improvements, reporting can be customized to showcase what's top of mind.

Descriptive reporting – A detailed understanding of a historical hosted data universe, including number of matters, number of documents, and treatment (if reviewed, if coded responsive, if coded privileged, etc.) is important because it can help pinpoint opportunities to improve data management across discovery, compliance, and information governance.

Benchmark reporting – A view of key benchmarks of historical document reviews through which analytics can be used to measure efficacy. These include the number of documents that were coded privileged versus not privileged, responsive versus non-responsive, responsive versus produced, documents reviewed per gigabyte, etc.

The example benchmark report below shows an overview of historical document reviews detailing key aspects of review across this organization's entire matter portfolio. This snapshot can be used to identify opportunities for greater efficiencies and as a benchmark while improving those processes, such as privilege review. In the sample situation below, attorneys historically found only about four privileged documents for every 100 that they reviewed. This shows a big opportunity to save on review costs by improving the privilege screen workflow. And having this baseline information allows for measurement of improvement as processes are changed to optimize this part of review.

Category	Metrics	Ratio
Unique documents	7,733,387	61.3%
Items reviewed	7,675,111	60.8%
Responsive items	3,211,464	41.8%
Privileged items	280,522	3.7%
Produced items	3,077,583	40.1%
Redacted items	89,849	1.2%

MATTERS: 13
TOTAL DOCUMENTS: 12,617,016

Real Benefits of Big Data Analytics

To further explore the benefits of using big data analytics technology in legal and specifically in ediscovery review, let's take a look at a simulation performed by Lighthouse in partnership with a major pharmaceutical company. The Lighthouse analytics team analyzed the pharmaceutical company's real data with Prism, a proprietary big data analytics technology that uses AI to aggregate and analyze document data and prior attorney work product. Prism applies multiple algorithms to the matter data for a comprehensive analysis. When available, as it was in this case, prior attorney decisions are used to tailor the algorithms to an organization's unique dataset and review decisions.

While there are many long-term benefits of using big data analytics technologies, such as gaining trend data and insights, this example focuses on the short-term, immediate benefits that can be reaped.

APPLYING THE TECHNOLOGY

Key data from 22 of the pharmaceutical company's past legal matters were ingested into Prism, including duplicate hash values, metadata, document text, and production information, as well as attorney responsiveness, privilege, and redaction coding. Prism was used to holistically analyze the matter portfolio to unearth actionable insights for improving the review process. It also "learned" from the ingested data to tailor its recommendations. Prism was then applied to a separate large review matter for analysis, and the results were compared to the results from the legacy process.

REAL-WORLD RESULTS

Overall, the analysis revealed that many documents were being unnecessarily reviewed – in some cases hundreds or thousands of times – across different matters over time. Many of these documents contained keywords commonly used to identify potentially relevant documents using search term screening, but were themselves non-substantive and never identified as responsive in prior reviews. In the case of privilege review, the company's process for identifying potentially privileged information returned 90% false positives in addition to failing to identify many documents that were actually privileged.

The analysis of the single matter uncovered opportunities to repurpose attorney work product from earlier matters. In total, more than 350K non-substantive and non-relevant documents were flagged as not needing review, equating to an estimated $620K in potential document review savings. Using multiple inputs and algorithms, Prism identified opportunities to improve privilege review efficiency. Prism generated privilege predictive scoring – if used, it would have resulted in an additional $412K in privilege review savings, for a total of $1.03 million in savings across the entire review[18].

Additionally, opportunities to improve consistency within matters and across the pharmaceutical company's entire portfolio utilizing prior work product were uncovered by gaining insights into how documents are treated and coded across matters.

The Right Approach

The right combination of products, people, and processes is different for each organization and depends on need, data expertise, and available resources. It's also important to consider an approach from the view of how future-proof it is. In other words, will it be able to scale and evolve with the changing data landscape?

An effective big data analytics technology can integrate case information, documents, and attorney work product across all matters into a single analytics platform. This holistic view enables organizations and their representative counsel to make data-driven decisions and strategies resulting in more favorable case outcomes. In particular, a big data analytics product should provide:

- **Comprehensive analysis** – Using multiple algorithms to analyze various aspects of documents for a fast, multidimensional understanding of the data;
- **Unique learnings** – Leveraging AI and machine learning to hone the algorithms using real work product in real time to naturally tailor results to a company; accessing experts to efficiently use cutting-edge technology to your advantage;
- **Insightful view** – Integrating siloed data sources into a single, holistic view to map trends and provide actionable insights;
- **Risk mitigation** – Storing information as abstract pieces of data (instead of full documents) to retain work product value without keeping documents indefinitely.

COMPREHENSIVE ANALYSIS

Efficiency can be improved across all matters using meaningful insights and portfolio-level and industry-level trends uncovered by a big data analytics product. The best application considers the context of the document. It is not limited to analyzing one aspect of the data (such as text) or using a single algorithm for ranking and prioritization. Instead, this comprehensive analysis should be supported by true data science techniques designed to handle massive amounts of information, as well as a variety of data formats and aspects, including text, metadata, and prior decisions.

UNIQUE LEARNINGS

A benefit of using AI and machine-learning technologies in ediscovery is the capacity to add an additional layer of efficiency through the ability to hone algorithms and strategies based on current and prior attorney work product. This work product includes attorney coding and other document treatments, such as privilege calls, redactions, privilege log entries, PII and PHI, and other key document designations.

INSIGHTFUL VIEW

Applied to millions of data points, the right algorithms will provide truly tailored insights into an organization's unique data patterns, which can be leveraged to enhance discovery, information governance, and compliance efforts. Taking an end-to-end approach to data can ensure that privileged and sensitive company information is secured from the point of creation to (potential) production.

RISK MITIGATION

A comprehensive view of matter data should not come at the expense of data security. Key strategies to look for in an approach include:

- **Abstracting key data** – When data is abstracted from the documents, users are able to take advantage of the data insights without gaining access to entire documents contained within the matter portfolio.
- **Robust security protocols** – Products should be engineered to adapt to ever-evolving IT requirements, regulatory standards, and cybersecurity threats, with a security program comprising of policies, procedures, training, and controls based on industry best practices.
- **Optimize information governance** – A truly holistic view of ediscovery data will uncover trends and insights that can be used to update information governance processes to protect data at the source.

Apart from the technological aspect of the product, processes and people are also critical for success. Having data scientists and expert analytics professionals available can help you efficiently onboard new technology, make the most of the product via processes, and scale a team's ability to manage and optimize the review process. Further, the right team can take the insights gained from analyzing a matter portfolio and create a program to optimize compliance and information governance efforts.

Conclusion

The biggest mistake organizations make in ediscovery today is forgoing products that use AI and machine learning to drive efficiencies that result in short- and long-term cost savings. The industry is struggling to keep up with today's data landscape, by using yesterday's tools and processes. The ediscovery challenge is, in fact, a big data analytics challenge.

Big data analytics technologies are now more accessible than ever, expanding the ability to use data to make workplace decisions. Organizations that want to keep up as the availability and complexity of data continues to grow need to share data science tools across their entire organization, as well as spread data expertise and responsibilities across teams, including legal.

The application of big data analytics technology will show significant benefits, particularly for those with high ediscovery and document review volumes. A comprehensive analysis of data can pinpoint the most important information, enable knowledge transfer between matters, remove non-responsive documents from review, and generate insights to holistically create an optimized data management process for organizations.

Taking an end-to-end approach will deliver the highest ROI when using big data analytics for ediscovery. The right combination of technology, expertise, and processes can comprehensively analyze matter data, tailor the algorithms, deliver multilevel insights, and include risk mitigation tactics. Most importantly, big data analytics can mine value from otherwise trapped matter data that can extend beyond ediscovery into compliance and information governance, delivering consistency and efficiency across the full data lifecycle.

REFERENCES

1 David Reinsel, John Gantz, John Rydning. "The Digitization of the World: From Edge to Core." IDC. US44413318. 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

2 IBM. "The Four V's of Big Data." https://www.ibmbigdatahub.com/infographic/four-vs-big-data.

3 See above.

4 David Reinsel, John Gantz, John Rydning. "The Digitization of the World: From Edge to Core." IDC. US44413318. 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

5 See above.

6 S. O'Dea. "Forecast number of mobile users worldwide 2020-2024." 2020. https://www.statista.com/statistics/218984/number-of-global-mobile-users-since-2010/.

7 IBM. "The Four V's of Big Data." https://www.ibmbigdatahub.com/infographic/four-vs-big-data.

8 "Cisco Annual Internet Report (2018–2023)." White Paper. March 2020. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

9 IBM. "The Four V's of Big Data." https://www.ibmbigdatahub.com/infographic/four-vs-big-data.

10 Tom Warren. "Microsoft announces Teams features as usage skyrockets nearly 40 percent in a week." The Verge. 2020. https://www.theverge.com/2020/3/19/21186452/microsoft-teams-new-features-noise-supression-user-increase-coronavirus.

11 Brian Elliott. "Not all daily active users are created equal: Work is fueled by true engagement." Businesswire. 2019. https://www.businesswire.com/news/home/20191010005813/en/.

12 IBM. "The Four V's of Big Data." https://www.ibmbigdatahub.com/infographic/four-vs-big-data.

13 See above.

14 "The Data Advantage: How accuracy creates opportunity." Experian. 2013. https://www.experian.co.uk/assets/marketing-services/white-papers/wp-qas-the-data-advantage.pdf.

15 Tadhg Nagle, Thomas C. Redman, and David Sammon, "Only 3% of Companies' Data Meets Basic Quality Standards." Harvard Business Review. 2017. https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards.

16 Lighthouse proprietary analysis based on a pharmaceutical company's matter data.

17 Note: Using an asterisk denotes a "wildcard" search term, used to find documents that contain derivatives of the word. In this example, privilege, privileged, etc.

18 Lighthouse Global proprietary Prism proof of concept, 2020.

LIGHTHOUSE

About Lighthouse

For 25 years, Lighthouse has provided innovative software and services to manage the increasingly complex landscape of enterprise data for compliance and legal teams. Lighthouse leads by developing proprietary technology that integrates with industry-leading third-party software, automating workflows, and creating an easy-to-use, end-to-end platform. Lighthouse also delivers unique proprietary applications and advisory services that are highly valuable for large, complex matters, and a new SaaS platform designed for in-house teams. Whether reacting to incidents like litigation or governmental investigations, or designing programs to proactively minimize the potential for future incidents, Lighthouse partners with multinational industry leaders, top global law firms, and the world's leading software provider as a channel partner.

Contact us to find out what Lighthouse can do for your business.

(206) 223–9690 | lighthouseglobal.com | info@lighthouseglobal.com