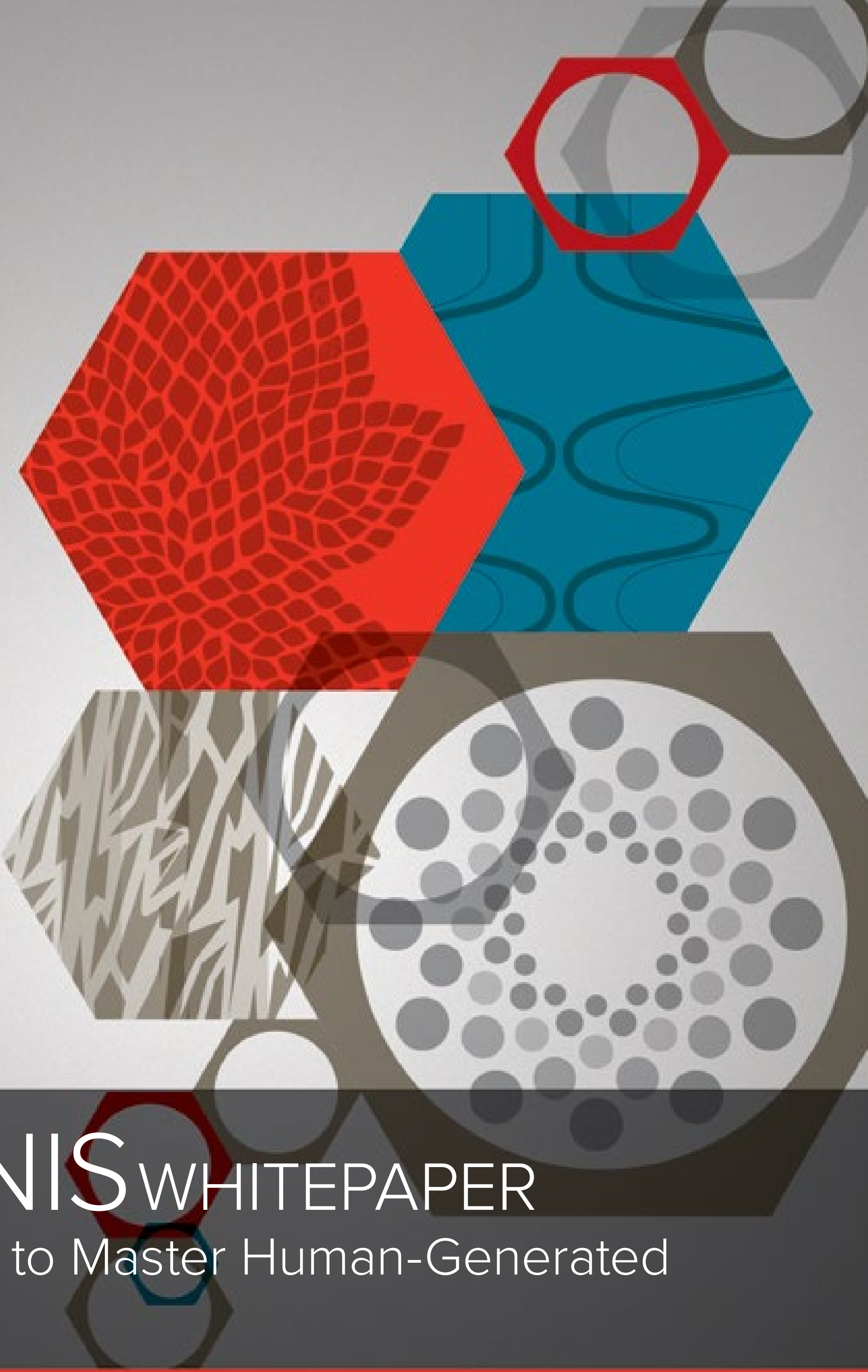

VARONIS WHITEPAPER

Why IT Needs to Master Human-Generated Big Data

CONTENTS

WHY IT NEEDS TO MASTER HUMAN-GENERATED BIG DATA

by Yaki Faitelson

This white paper discusses the challenges associated with the continued generation of human-generated content through digital collaboration, and why big data analytics techniques present both a remedy for current management and protection challenges, as well as opportunities for more efficient and informed collaboration.

HUMAN-GENERATED BIG DATA

The files and emails that most of us create every day are “human-generated data.” These are emails, word processing documents, spreadsheets, presentations, audio files. Not only do these files take up the lion’s share of digital storage capacity in most organizations, we usually keep them around for a long time, and there is an enormous amount of metadata associated with them.

Human-generated content is big; the metadata is bigger. Interesting metadata about a file might be who created it, what type of file it is (spreadsheet, presentation), what folder it is stored in, who has been reading it, who has access to it, or who sent it in an email to someone else. Over its lifespan, a file is usually accessed by many people, copied, sent or moved around to many places in many file systems. This metadata is so big that if you collect and store it all in its raw form, before long its size will dwarf the files themselves. The metadata associated with human-generated content is human-generated Big Data.

Just as analyzing machine generated data has practical applications for business, analyzing the “big metadata” associated with human-generated content has enormous potential. More than potential, harnessing the power of big metadata has become essential to manage and protect human-generated content. Those that fail to adopt these technologies report that they have little confidence that their data is protected[1], that they don’t know where critical information resides, do not know who it belongs to, and are no longer able to keep up with fundamental data protection activities.

WHY DO WE CALL IT BIG DATA?

Whereas a spreadsheet can be processed in memory, and a database can be processed by a single server, big data analytics operations generally require platforms that allow multiple systems to collect, store, and analyze the information. Wikipedia defines Big Data as follows: “In information technology, big data consists of data sets that grow so large and complex that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analytics, and visualizing.[2]”

Capturing, storing, searching, analyzing, and visualizing human-generated content at scale requires a big data analytics platform that is capable of handling the enormous amounts of metadata involved.

WHERE IS HUMAN-GENERATED BIG DATA STORED?

For organizations, human-generated big data is stored in file servers, email systems, and intranets. These platforms facilitate collaboration—they make it easy to store, find and share content, almost without thought. These stores don't seem to be going anywhere, either— over the next decade, IDC estimates that data centers will be responsible for over 50 times the data they currently manage; while there staff will only grow by a factor of 1.5[3].

As more employees work remotely outside of the traditional LAN infrastructure, on laptops, tablets, and smart phones where files shares and intranets are difficult to access, many have turned to cloud-based file sharing solutions to keep up their collaborative pace. This represents another growing data store—one that brings with it new management and protection challenges because data may now be stored in multiple locations inside and outside the organization's physical borders.

THE (HUMAN-GENERATED) BIG DATA IMPERATIVE

Unfortunately, file shares, emails, and intranets have made it so easy for end users to save and share files that organizations now have more human-generated content than they can sustainably manage and protect using small data thinking.

Many organizations face real problems because questions that could be answered 15 years ago on smaller, more static data sets can no longer be answered. These questions include: Where does critical data reside, who accesses it, and who should have access to it? IDC estimates that only half the data that should be protected is protected[4].

The problem is compounded with cloud based file sharing, as these services create yet another growing store of human-generated content requiring management and protection—one that lies outside corporate infrastructure with different controls and management processes.

Publicized data breaches, data protection and retention regulations have brought these problems into focus; as has the realization that data theft/loss can have severe consequences—not only for the organization, but also for its customers, business partners, and shareholders. As a result, organizations are scrambling to implement controls and processes for their human-generated Big Data.

Organizations survive through digital collaboration with human-generated content, but in order to manage and protect so much content, and even to collaborate efficiently, new thinking and technology must be adopted.

Before we can truly face the challenges and opportunities in Human-generated Big Data, we must accept that collaboration, by its nature, is chaotic. Organizations cannot constrict or control collaboration too much, or productivity will suffer. With too much chaos, on the other hand, the organization is put at risk, suffers an unsustainable management burden, and again the collaboration suffers—we can't collaborate with the right people, can't find the material we need, etc.

Big Data Analytics for human-generated content can control the chaos. By analyzing the content and its associated metadata across all the platforms where human-generated content is stored, organizations can understand better fundamental collaboration questions: who collaborates with whom, what data they share, who uses it, and where it is stored. From management and protection perspectives, Big Data analytics can determine who has and should have access to human-generated content, who owns it, where sensitive and regulated content is exposed to too many people, who might be abusing data, and which data is no longer needed.

Big Data Analytics provides the intelligence to automate many data management and protection activities that are manual today. Data can be made accessible to only the right people in the right places, moved to the appropriate data stores, protected from exposure, and archived safely. It is more efficient to let technology handle the technical idiosyncrasies of each platform than for IT to move data around manually, configure access controls, and look through access logs.

With analytics and automation, it's possible for humans to pay less attention to the mechanics and operational overhead associated with collaboration, and pay more attention to the content and the collaboration itself. If humans can easily create and share their content with whomever they need, and organizations can be confident that collaboration is controlled (data is in the right place, accessible to only the right people, use is monitored, etc.) then not only can we avoid being crushed under the weight of our human-generated Big Data, we can continue to realize the productivity gains that digital collaboration makes possible.

THE VARONIS METADATA FRAMEWORK: PURPOSE BUILT FOR HUMAN-GENERATED BIG DATA

There are many Big Data Analytics platforms available today; some are generic and some are purpose built. The Varonis Metadata Framework is a purpose-built Big data Analytics platform to manage and protect human-generated content—the files and email that resides on corporate infrastructure.

The Varonis Metadata Framework non-intrusively collects critical metadata about human-generated content, generates metadata where existing metadata is lacking (e.g. its file system filters and content inspection technologies), preprocesses it, normalizes it, analyzes it, stores it, and presents it to IT administrators in an interactive, dynamic interface. Once data owners are identified, they are empowered to make informed authorization and permissions maintenance decisions through a web-based interface—that are then executed—with no IT overhead or manual backend processes.

The intelligence and automation provide immediate answers to critical questions about human-generated content that IT departments struggle to answer every day, such as:

- Who is accessing data?
- Who has too much access?
- Who may be abusing their access?
- Where is sensitive data, and where is it exposed to too many people?
- How can access by reduced safely?
- Which business units align with each data set?
- Which data sets are no longer used?

The Varonis Metadata Framework will scale to present and future requirements adding standard computing infrastructure as Big Data Analytics nodes, even as the number of functional relationships between metadata entities grows exponentially. As new platforms and metadata streams emerge, they will be seamlessly assimilated into the Varonis framework, and the productive methodologies it enables for data management and protection.

[1] *See The State of Data Protection, http://hub.varonis.com/data-protection- survey-results/*

[2] *http://en.wikipedia.org/wiki/Big_data*

[3] *Extracting Value from Chaos IDC Digital Universe Study (Sponsored by EMC)*

[4] *Extracting Value from Chaos IDC Digital Universe Study (Sponsored by EMC)*

ABOUT VARONIS

Varonis is the leading provider of software solutions for unstructured, human-generated enterprise data. Varonis provides an innovative software platform that allows enterprises to map, analyze, manage and migrate their unstructured data. Varonis specializes in human-generated data, a type of unstructured data that includes an enterprise's spreadsheets, word processing documents, presentations, audio files, video files, emails, text messages and any other data created by employees. This data often contains an enterprise's financial information, product plans, strategic initiatives, intellectual property and numerous other forms of vital information. IT and business personnel deploy Varonis software for a variety of use cases, including data governance, data security, archiving, file synchronization, enhanced mobile data accessibility and information collaboration.

Free 30-day assessment:

WITHIN HOURS OF INSTALLATION

You can instantly conduct a permissions audit: File and folder access permissions and how those map to specific users and groups. You can even generate reports.

WITHIN A DAY OF INSTALLATION

Varonis DatAdvantage will begin to show you which users are accessing the data, and how.

WITHIN 3 WEEKS OF INSTALLATION

Varonis DatAdvantage will actually make highly reliable recommendations about how to limit access to files and folders to just those users who need it for their jobs.

WORLDWIDE HEADQUARTERS
1250 Broadway, 31st Floor, New York, NY 10001 **T** 877 292 8767 **E** sales@varonis.com **W** www.varonis.com
UNITED KINGDOM AND IRELAND
Varonis UK Ltd., Warnford Court, 29 Throgmorton Street, London, UK EC2N 2AT **T** +44 0207 947 4160 **E** sales-uk@varonis.com **W** www.varonis.com
WESTERN EUROPE
Varonis France SAS 4, rue Villaret de Joyeuse, 75017 Paris, France **T** +33 184 88 56 00 **E** sales-france@varonis.com **W** sites.varonis.com/fr
GERMANY, AUSTRIA AND SWITZERLAND
Varonis Deutschland GmbH, Welserstrasse 88, 90489 Nürnberg **T** +49 (0) 911 8937 1111 **E** sales-germany@varonis.com **W** sites.varonis.com/de