DIGITAL MOUNTAIN®

SPRING 2022 E-NEWSLETTER

At Digital Mountain, we assist our clients with their electronic discovery, digital forensics, cybersecurity, and data analytics needs. For this E-Newsletter, we discuss the business trends and cost explosion in eDiscovery.

The Anatomy of a Gigabyte – Preventing Cost Explosion

With data sizes growing 40% per year, it's imperative for eDiscovery practitioners, litigants, and investigators to understand data explosion and how vendor positioning of a Gigabyte as the price metric can greatly influence total costs. If spreadsheets aren't your forte, pivoting your skills to be able to perform apples-to-apples comparisons properly or engaging a skilled staff member to assist you is vital to understanding and controlling eDiscovery costs. When we talk about measuring data for storage purposes, we speak of bytes: kilo-, mega-, giga-, tera-, peta-, exa-, zeta-, all the way up to yottabytes. Current price sheets are primarily in gigabytes (GB). As a result, that is the primary unit of measurement discussed in this article.

Breaking Down Bytes and Compression

As a quick review, one gigabyte is comprised of 1,024 megabytes. Depending on the type of files being created, the quantity represented can vary dramatically. Here's a rough guesstimate of how 1GB can be viewed based on various data sources:

- 9,000 emails with 3,000 attachments
- 500 photographs
- 50 song downloads
- 1 hour of HD video streaming

The above summary is a simplification of how we measure stored data. For example, text files can yield a much higher volume. By compressing files using applications such as WinZip, 7-Zip, etc., we can create .zip, .rar, .tar, .7z, and other compressed file types. If compressed data is 1GB, the data set may yield two times more data on average when unzipped. As more advanced compression technologies and innovations materialize, the average compression rate may

increase. Compressing data is helpful for speeding up data transfers over the cloud such as FTP utilized in electronic discovery.

What Gigabyte Are You Paying On?

All vendors have unique pricing schemes. At the end of the day, it is important to consider what is the unit of measure on which you are paying and how that impacts the total costs. Here is an example workflow and data sizes that may occur:

- <u>1GB Compressed Data</u> – this is the data that was purposely put into a zip or compressed format for quicker transmission on the cloud (e.g., 1GB). This data may contain compressed data of previously compressed data, e.g., zip files of zip files. It may also be referred to as "incoming GBs."
- <u>2GB Uncompressed Data</u> – this is the same data that is now uncompressed from the compressed container after being downloaded from the cloud (e.g., 2GB if using a 50% compression rate). This is typically the size at which the data was maintained in the ordinary course of business.
- <u>4GB Expanded Data</u> – this is the data size when expanding data from .zip and other compressed files as held in the ordinary course of business. This may, depending on how the data is processed, also include extracting messages from .pst containers, so that they can be properly searched and culled (this may cause the 1GB of data to explode to the example 4GB, but the actual range can vary dramatically). For example, the message extraction process from .pst files may explode data by 5%-50%, whereas uncompacting compressed files may explode data by double, but the amount really depends on the specific data set.
- <u>10GB Processed Data for a Review Tool</u> – assuming no culling/filtering occurs, the following additional processing may need to happen and for which hosting companies may or may not charge for certain expansion steps:
  - Since review tools are in Web-based environments, an HTML conversion needs to happen to make the data viewable, causing further data explosion. This assumes PDF conversion does not occur on all of the data which can also be handled by web browsers.
  - In order to review attachments to emails as separate records or unique IDs, they need to be extracted from emails causing further data explosion.
  - If case teams require TIFF or PDF images, this will explode data dramatically. As a result, most of the industry went from image conversion on all data to only converting to images for redaction and production purposes (including Bates numbering and/or designation stamping). These images are stored in the web-based review environment, and monthly hosting fees are charged on images as well as the fully processed data. A black and white TIFF image will generally be less than 80KB, while a full-color .jpg will be as much as 10MB. If color is required versus monochrome images, this could cause 125x more data storage to be utilized.

The above example is just a hypothetical, but it demonstrates the need to understand how data explosion can affect pricing. For example, suppose you are paying $200 per GB on the compressed data size in an all-inclusive pricing model that bundles a year of hosting. In that case, the total cost may be much less than if the data explodes to 10GB in a review tool for which you are paying $50/GB for processing on uncompressed data (2GB), i.e., $100, plus a year of hosting

costs on 10GB at $20/GB per month for an aggregate cost of $2,500. Depending on the complexity of the matter, hosting may be required for multiple years, during which hosting costs accrue.

GB-to-GB or Apples-to-Apples – Preventing Unanticipated Costs

While the prices and GB units may vary, it's important to compare apples-to-apples and evaluate what your processes need to be based on your case. This will allow you to perform proper planning and determine what the best pricing model is for each case. A quote may look superficially lower until you examine the true potential costs by running the numbers. Most vendors do not base pricing on the compressed GB size as the uncompressed data size is so unpredictable, so they'd be guessing and gambling. In addition, pricing on uncompressed data incentivizes the customer to find the most aggressive compression technology available and leverage that to reduce the incoming GBs in the container for processing. While this may feel like a win for the customer, over-compressing data is not a desirable practice as it can result in loss of data fidelity.

When performing electronic discovery, it's important to take advantage of early case assessment capabilities, culling, and filtering prior to putting data into a hosted environment in order to minimize costs. If you have a large number of video and picture files, it may be better to negotiate hourly-based pricing or value-based pricing to cull the data as the GB-based pricing model may not be cost effective in certain circumstances based on your data set. Proper planning and understanding of the steps and phases of data explosion, as well as the different definitions of GBs, will help manage client expectations. We all want to keep that dashboard green for our clients and prevent unanticipated sticker shock.

Please direct questions and inquiries about electronic discovery, digital forensics, cybersecurity, and data analytics to info@digitalmountain.com.

UPCOMING INDUSTRY EVENTS

ADFSL CONFERENCE ON DIGITAL FORENSICS, SECURITY AND LAW
Virtual: May 23-25, 2022

NETDILIGENCE CYBER RISK SUMMIT
Philadelphia, PA: June 1-3, 2022

INTERNATIONAL CONFERENCE ON DIGITAL FORENSICS AND INVESTIGATION
San Francisco, CA: June 2-3, 2022

RSA CONFERENCE 2022
San Francisco, CA & Virtual: June 6-9, 2022

BLACK HAT USA 2022
Las Vegas, NV: August 6-11, 2022

Digital Mountain, Inc. Founder and CEO, Julie Lewis,
will be presenting at various upcoming industry events.
Please send requests for speaker or panel participation
for her to _marketing@digitalmountain.com_.

DIGITAL MOUNTAIN, INC.

4633 Old Ironsides Drive, Suite 401
Santa Clara, CA 95054
866.DIG.DOCS

www.digitalmountain.com

Contact us today!

FOLLOW US AT: