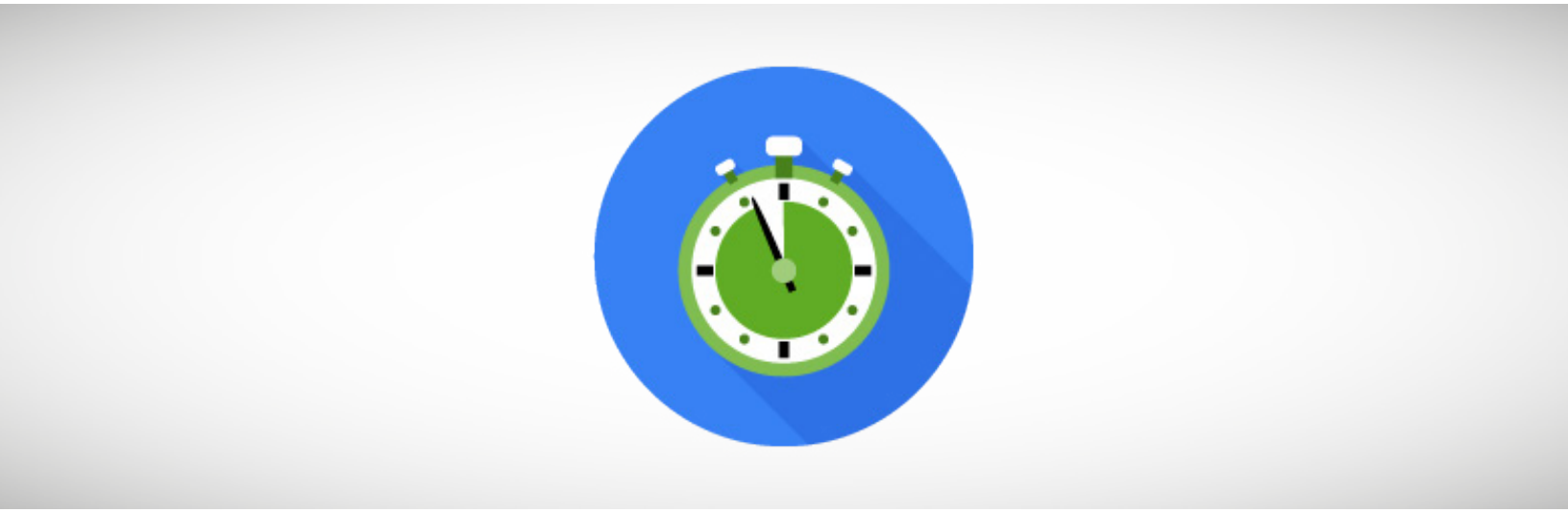
ZAPPROVED



Inside 1 Terabyte Per Hour

Building a High-Performance Pipeline for Heterogeneous Data Processing on AWS

By: **Lee Harding**, CTO, Zapproved, Inc.

DISCOVERY

Legal Hold **Pro** Data Collect **Pro** Digital Discovery **Pro**

About Lee Harding

CTO | ZAPPROVED INC.

Lee has worked in software technology for more than 20 years and brings experience from Autodesk, Intel and Sunstone Circuits. He provides high-level technical guidance in opportunities ranging from analysis and remediation strategies for regulatory mandates to fine-grained decisions such as team tool selection. His technical experience is varied and includes e-commerce, corporate web presence and ECAD application development. Lee holds a Bachelor's Degree in Mechanical Engineering from Oregon State University, a MS in Mechanical Engineering from Portland State University, and has been awarded four patents for his work. Lee is very involved in the entrepreneurship community including the Software Association of Oregon, the Austin Entrepreneurship Program at Oregon State University, and FIRST Robotics.

Disclaimer

This technical paper is provided for general information and educational purposes only. The contents should not be construed as legal advice or opinion. While every effort has been made to be accurate, the contents should not be relied upon in any specific factual situation. This technical paper is not intended to provide legal advice or to cover all laws or regulations that may be applicable to a specific factual situation. If you have matters to be resolved or for which legal advice may be indicated, you are encouraged to contact a lawyer authorized to practice law in your jurisdiction.

About Zapproved Inc.

Founded in 2008 in Portland, Ore., Zapproved Inc. is a pioneer in developing cloud-based software for corporate legal departments. The Z-Discovery Platform returns power to in-house corporate legal teams and helps them navigate electronic discovery with minimal risk and cost, and it sets new standards for scalability and intuitive design. The company's flagship product, Legal Hold Pro, is widely adopted by Fortune 500 and Global 2000 corporations and has earned recognition as the Best E-Discovery Legal Hold Product at the 2015 Legaltech News Innovation Awards, in 2014 - 2016 Best of the National Law Journal and the 2013 and 2014 Best of Legal Times. Zapproved was recognized in the 2014 Inc. 500 as one of the fastest growing private companies in the U.S. and was named as a "vendor to watch" in the 2015 Gartner Magic Quadrant for E-Discovery.

Inside 1 Terabyte per Hour– Building a High-Performance Pipeline for Heterogeneous Data Processing on AWS

In February 2016, we announced that our data processing software Digital Discovery Pro had exceeded the processing rate of one terabyte per hour — which exceeded known performance by more than 20x. The development team worked closely with AWS architects to achieve this rate by exploiting the AWS platform to bring to bear the compute power necessary to achieve these rates and solve a significant pain point in the legal industry.

Some key indicators of performance of the processing pipeline are:
- About 5.1 million documents and ~78.7 million events processed in less than one hour
- Multiple tasks (i.e. clients) simultaneously with no degradation in performance

This technical paper provides an overview of the following:
- Market need for improved processing
- Phases of data processing for legal
- Challenges for improving processing speed
- Methodology for measuring processing performance

Why Focus on Processing?

When we looked at the landscape around processing in the e-discovery marketplace, we saw processing time was a significant factor in the overall turnaround time to provide information from data collected from custodians.

We looked critically at the E-Discovery Reference Model (EDRM) for opportunities to improve processes, software and typical workflow employed by corporations. Through talking with and surveying our customers, prospects, and the industry at large, we identified that the legal team's expectations were not being met and that long delays and painful iteration cycles were adding risk and uncertainty early in the matter life-cycle. In fact, it may not be that their expectations "were not met" because without any alternative the users were conditioned to waiting days or weeks.

While other pain points included the expense and maintenance of on-premise software, complexity of software which typically required dedicated resources or a reliance on outside experts — topics we address with this software — we will focus on our approach to solving the issue of speed. In the e-discovery process it is **time to information** that is the critical measure. We adopted the goal for Digital Discovery Pro to deliver instant access to information to the corporate legal team. This required innovation in two areas: ease of use and processing speed. To achieve our goal of instant access to information it was clear that Digital Discovery Pro needed to eliminate the processing delay and put the software directly in the hands of the legal team, regardless of the volume or complexity of data.

The distinction here of **information** versus **data** is necessary:
- Data refers to the raw digital content as collected from computers, cloud services, mobile devices and portable media.
- Information, on the other hand, refers to the knowledge and contextual content needed in the decision making process around a matter.

We found that although software vendors marketed aspirational processing rates that were measured in the hundreds of gigabytes per day, much lower processing rates are realized in practice. It was not uncommon to find processing rates in the mere

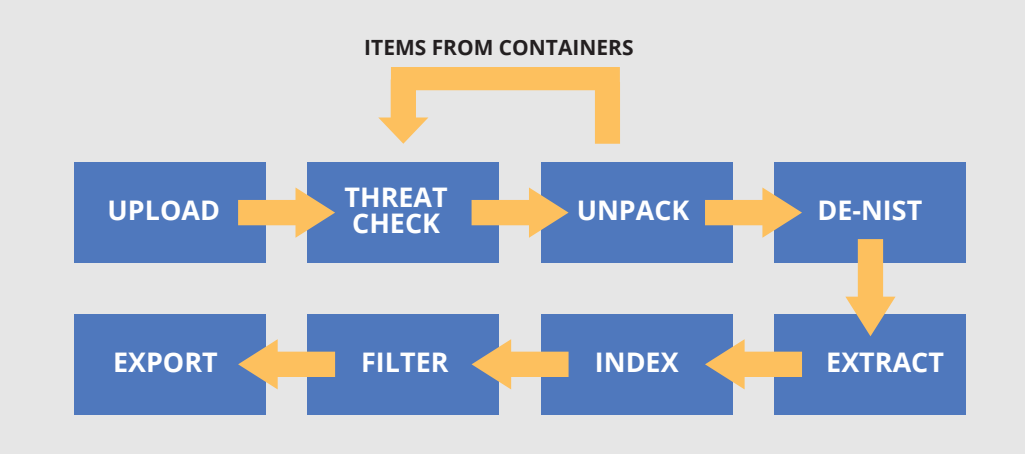


FIGURE 1.

dozens of gigabytes per day. Typical delays in processing multi-terabyte data sets were weeks or months.

The time to information is also constrained by the ability to move data into processing systems. While collecting data from custodians can add considerably to the time to getting information (which is addressed by other technology outside of the scope of this software), the ingestion time was the critical delay.

What Are the Phases of Data Processing for Legal?

When responding to litigation, corporate legal teams collect any data that could be potentially responsive to the case. Typically, the litigation support team collects broadly in order to minimize the risk of missing a critical document or communication. Once this data is collected, the loose files are identified and indexed so that legal professionals can begin investigating business records that are pertinent to the case.

Additionally, a typical corporation spans more than one office location, more than one data center, and stores information across a wide diversity of systems internally and externally — all potentially subject to discovery. This is where the job of processing begins.

Moving the Documents

If 1 TB of data (files, emails, text messages, designs, financials, contracts) is out there, the first step is to move that data set to where Digital Discovery Pro can access it. In the past, this information may have all been in one place, for example, Microsoft Exchange. Today, it's common for documents to be distributed between a combination of mobile, server and cloud locations. Collecting from these disparate sources to a single storage becomes a bottleneck, and an expensive proposition if the storage system is not efficient and scalable.

Amazon's S3 meets these criteria with the ability to retain trillions of objects arriving at very high rates from anywhere in the world. Modern browser software directly accesses S3 as the point of ingestion for Digital Discovery Pro process making uploading as easy as using "drag and drop" for files of any size. While performance is dependent on upload bandwidth, direct collection methods such as Data Collect Pro help automate and distribute the collection process efficiently.

Unpacking (All) the Documents

The process of unpacking documents involves the recursive expansion of containers such as compressed archives (ZIP, 7-ZIP, etc.), disk images (ISO, VHD, etc.) and mail containers. These files can contain documents as well as other containers, and the process of "unpacking" must produce all of the documents even at the deepest levels of nesting. The Enron v2 data is, for example, ZIP files that each contain one or more PSTs. Upon ingestion by Digital

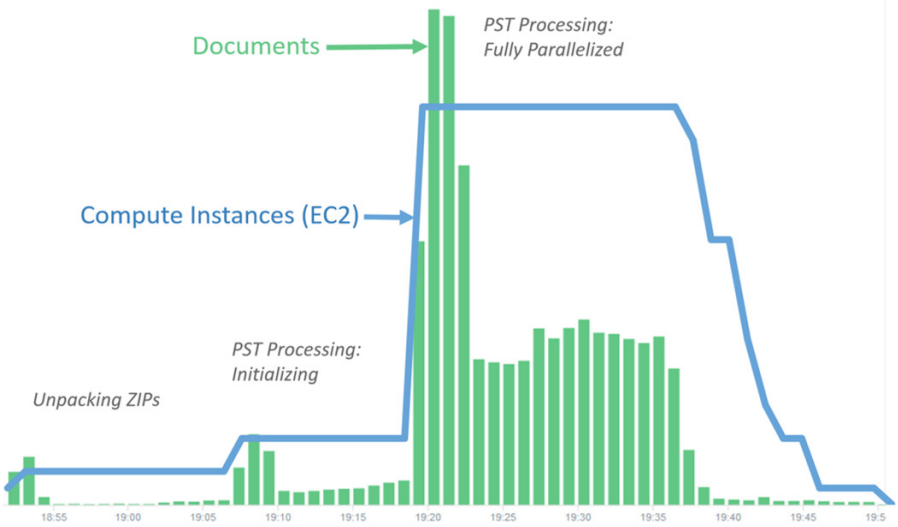


FIGURE 2. Graph of Documents vs. EC2 Compute Instances

	Enron v2	3 x Enron v2
Custodians	151	453
Zip Files	153	459
Compressed File Size	37.0 GB	111.0 GB
Unpacked PST Files	170	510
Unpacked PST Files	75.8 GB	227.4 GB
Unpacked Documents	1,712,079	5,136,237
Unpacked Documents	290.8 GB	872.4 GB

FIGURE 3. Overview of Test Data Set

Discovery Pro, these ZIP files are assigned multiple unpacking workers to share the work and ultimately produce millions of discrete documents (such as mail messages).

Threat Detection

Upon arrival in S3, all documents are inspected by up to three threat detection services to identify malware, viruses and other malicious content. These checks are performed on both the original uploaded objects, as well as any "unpacked" files from containers including email attachments.

Identifying NIST and Duplicate Files

In litigation, system files and other items that have no evidentiary value, i.e. the files that are known not to be valuable can be called out. We compare files against the NIST National Software Reference Library, or NSRL, to locate and exclude (if desired) system files. This task uses the same CloudFront technology as websites that serve millions of requests, producing results faster by acting in parallel rather than one at a time.

Extracting Relevant Information

The next step in the pipeline is to open each file with software that understands the content. For example, MS Word files need software that understands MS Word files. Similarly the same requirement exists to read email and compressed archives, presentations, images, spreadsheets and the hundreds of common file types used by corporations. Images may need to be sent through OCR (optical character recognition) if they contain text.

Making it Search-Friendly

Take the extracted information and normalize it — give it a consistent presentation to make it easier to find. Structure all the information in a way that keeps it organized and quick to access.

Exporting the Relevant Subset

Re-assemble what is found (and not excluded) into a form that can be used by downstream applications. The documents are assembled into groups, until all needed documents are included in downloadable images.

Our Challenge: How Fast Can We Make The Processing Pipeline?

We knew that to radically change processing speeds we needed to do things differently. In addition, we were facing a new reality of Software-as-a-Service and hosted data that meant we also needed a new way to **collect** data. The volume of data relevant to a matter is growing proportionally to the overall data managed by corporations. By any statistical measure that growth is fast — and accelerating.

Typical data collection volumes in legal and regulatory matters range from a few gigabytes all the way up to tens of terabytes. Large legacy collections archives of 50 to 100 terabytes or more are not uncommon. With these data volumes in mind, and also considering the rapid growth of information, we understood that to provide instant access meant being able to process typical collections in less than an hour.

What we know is this:

- To be fast means being **distributed**.
- To be cost effective means being **elastic**.

Being **distributed and elastic** required revisiting back at what engineers have wanted to build in the past but couldn't due to lack of appropriate infrastructure. Now with modern SaaS infrastructure, we can take advantage of AWS Elastic Compute Cloud (EC2) to deploy an event-driven stateless transformation engine. The inherent challenges of this model are:

- Efficiently assessing the data in queue to deploy precisely the amount of compute resources needed to perform the requested task in an appropriate time frame.

- Managing asynchronous events to have a correct order of execution to ensure complete and accurate processing.

Since the early 1970s, the concept of the Actor Model has been explored as an inherently concurrent model, but not widely nor affordably available to commercial software developers.

With the introduction of AWS Lambda compute service in December 2014, that changed. Lambda introduced a distributed stateless code execution service which allowed our developers to design and implement an event-driven stateless transformation engine. Based on Actor principals, this system automatically expands to meet workload demand with each discrete operation sharing nothing and free of dependencies.

In addition to Lambda, Digital Discovery Pro employs more than a dozen services from AWS including: CloudFront, CloudWatch, EBS, EC2, ELB, Kinesis Firehose, RDS, Redshift, SNS, SQS, VPC and S3 (see Glossary p. 6). Each service chosen to maximize the overall throughput of transformations and storage while processing ingested data.

A critical challenge is to instantaneously and accurately provision the correct amount of compute resources so that it can process the job in an acceptable time frame. Our software service calculates a Maximum Projected Time in Queue (MPTQ) value. This is derived by looking at the amount of data waiting to be processed and then automatically spinning up AWS workers. It continues doing so until the MPTQ drops below a predetermined threshold. As soon as a worker has no more work, it is automatically spun down. Figure 2 illustrates that there is a ramp-up phase as the system begins processing a job. While a large task, such as one terabyte will conclude in an hour, a smaller job – say 1/60th of a terabyte or 16.6GB – won't take just a minute. It will get processed extremely quickly but the ramp up time and allocation of resources can result in slightly lower speeds.

Methodology for Measuring 1 Terabyte per hour Performance

A common dataset used to test electronic discovery system performance is the public domain set called "Enron Email Data Set v2" which is available at EDRM.net. The data set is a widely used reference for e-discovery software and process testing. It is comprised of more than 1.7M messages from 158 employees that were collected by the Federal Energy Regulatory Commission after the company collapsed. See Figure 3 for a detailed view of the Enron v2 data set. In order to provide a robust test, we ingested three copies of the data set concurrently when conducting performance testing. Each set of files were processed in parallel as separate client matters to simulate three concurrent projects with a total ingestion size approaching 1TB. The data when ingested as ZIP files needs to be decompressed and then all files — and even files within files such as email attachments — must be extracted. See Figure 4 on the previous page for complete results of pipeline performance test.

	Results
Total Processed Data Set (GB)	983.4 GB
Time to Ingestion by S3 copy	10.6 seconds
Effective Ingestion Rate	296.3 TB/hr
Time to First Document	12.52 seconds
Time to 100,000 Documents	12 minutes 55 seconds
Time to Last Document	53 minutes 46 seconds
Overall Effective Processing Rate	1.097 TB/hr

FIGURE 4. Results from Test of Pipeline Performance

Conclusion

To achieve these performance levels of processing — nearly 80 million tasks on 5.1 million documents — in under an hour required a new architecture from existing approaches that are currently available in the e-discovery market. By building a system from the ground up to take full advantage of the latest developments in Amazon Web Services, our developers created a robust data pipeline that resets the norms for this space.

Another accomplishment to note is that we've confirmed that Digital Discovery Pro can readily handle multiple high-volume processing tasks with no degradation in performance on any single processing job.

In time, our team will continue to push performance standards even higher in terms of processing speed since it is a matter of tapping the resources available to us in AWS. These performance tests confirm that the architecture is capable of scaling in order to meet our current performance objectives and can meet future requirements as data sets increase exponentially.

GLOSSARY OF AWS SERVICES

CloudFront — Content delivery network (CDN) service to distribute content to end users with low latency and high data transfer speeds.

CloudWatch — Monitoring service to collect and track metrics, collect and monitor log files, set alarms, and automatically react to changes in AWS resources.

EBS (Elastic Block Store) — Provides persistent block level storage volumes offering high availability and durability.

EC2 (Elastic Compute Cloud) — A web service that provides resizable compute capacity in the cloud that allows services to quickly scale capacity, both up and down, as computing requirements change.

ELB (Elastic Load Balancing) — Automatically distributes incoming application traffic across multiple EC2 instances to enable greater levels of fault tolerance and provide the required amount of load balancing capacity needed to distribute application traffic.

Kinesis Firehose — A platform for streaming data that offers services to make it easy to load and analyze streaming data by continuously collecting, storing and processing large volumes of streaming data.

Lambda — An AWS service that runs code without provisioning or managing servers — also called a serverless microservice — all with zero administration.

RDS (Relational Database Service) — Provides cost-efficient and resizable capacity that makes it easy to set up, operate, and scale a relational database in the cloud.

Redshift — A fast, fully managed, petabyte-scale data warehouse that makes it simple and cost-effective to analyze data using your existing business intelligence tools.

SNS (Simple Notification Service) — A fast, flexible, fully managed "pub-sub" (or publish and subscribe) messaging service to work as enterprise-messaging infrastructure or to send push notifications, email, and SMS messages.

SQS (Simple Queue Service) — A scalable, fully managed message queuing service to transmit any volume of data, at any level of throughput, without losing messages or requiring other services to be always available.

VPC (Virtual Private Cloud) — Provisions a logically isolated section of the Amazon Web Services (AWS) cloud to launch AWS resources in a virtual network of which the developer has complete control over the networking environment.

S3 (Simple Storage Service) — Secure, durable, highly-scalable cloud storage to store and retrieve any amount of data from anywhere on the web.

Start your Smarter E-Discovery Strategy Today >>

visit: www.zapproved.com

call: (888) 806-6750

ZAPPROVED™
info@zapproved.com | (888) 806-6750