WHITE PAPER

HIGH-SPEED EDISCOVERY PROCESSING & PRODUCTION

Conversion of the 53 GB, 5 Million-Page EDRM Enron Dataset into TIFF images in 5 hours, at a 24 Million pages/day rate, using the scalable Lexbe eDiscovery Processing System

August 2014

Introduction

In this white paper, we briefly describe a demonstration conducted to evaluate the speed and effectiveness of the Lexbe™ eDiscovery Processing System ("LEPS") to process and convert the EDRM Enron Dataset: a standardized real-life large collection of email and attachments to TIFF: a format commonly used for document review by law firms in popular litigation document review applications, and a notoriously slow and difficult conversion to make. We'll provide background for the eDiscovery industry, outline the goals of the study, describe the methodology used in the demonstration, and, finally, summarize the results.

LEPS converted the entire 53 GBs of the EDRM Enron Dataset, with 5 million page equivalents, into industry-standard TIFF images in only 5.3 hours. To accomplish this task in this short time, LEPS programmatically deployed and utilized over 60 parallel server instances, and maintained a sustained throughput rate of 240 GBs/day (23 Million pages/day) for Native to TIFF processing. This is a substantial increase in industry TIFFing capability. By comparison, an existing industry leader, Xerox Litigation Services, has a company-wide capacity of 5 Million pages per day, or less than 25% of Lexbe's capacity.

The substantial speed increase of LEPS does not compromise security or quality, as LEPS operates in a redundant, highly-secure environment, and the automated architecture of LEPS enhances quality and quality control as compared with traditional processing approaches.

Goal of the Demonstration

Our goal was to evaluate the capability of LEPS to perform high-speed, automated processing of native electronically stored information (ESI) to TIFF, on a large, real-life data set, under conditions that are testable and repeatable. This demonstration documented the steps that are typically undertaken in a large eDiscovery job, start to finish. In particular, we wanted to show eDiscovery processing under real-life stress conditions, utilizing the full range of eDiscovery processes.

Background of eDiscovery Processing & Production

Activities

eDiscovery processing represents a series of operations applied to ESI in connection with litigation to prepare data for document review or production. These CPU- and I/O-intensive processes include file decompression, file extraction and separation, file type identification and extension repair, metadata extraction and database fielding, system file identification and deNISTing, file deduplication, native text extraction, optical character recognition (OCR) of image files, database indexing, conversion of files to easily reviewable formats (TIFF, Native-extracted or PDF), embossing of files with Bates numbering and other designations, and the creation of load files for litigation document review applications.

Professionals

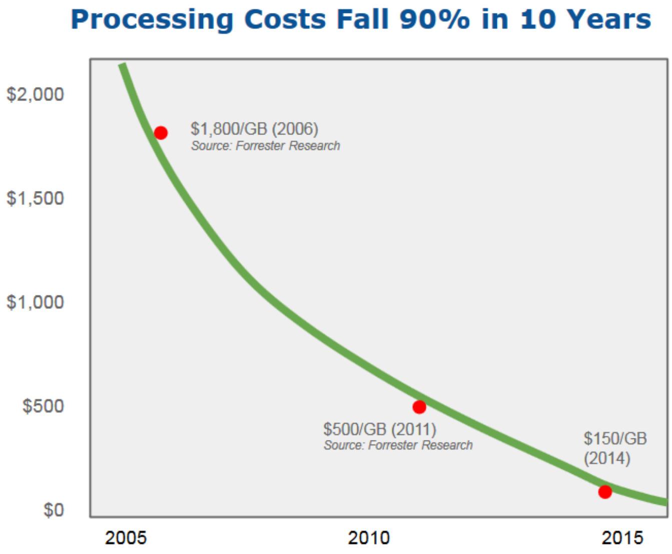
eDiscovery Processing and Production is typically performed by Litigation Support or IT personnel, often utilizing various specialty software applications, who work for the companies and organizations actually involved in litigation, their law firms, or with specialty eDiscovery Service providers, consultants or vendors. There are in excess of 10,000 Litigation Support Professionals in the U.S., working in law firms, large organizations, in private service companies, and in government agencies.[1] A principal function of litigation support professionals is managing the conversion of ESI to reviewable form -- including Native to TIFF conversion -- and loading the converted data into litigation management tools for attorneys and other legal professionals to conduct document review.

The eDiscovery Processing Market Continues to Grow Rapidly

Worldwide Processing & Production revenue involved with eDiscovery is estimated from various sources to be $0.77 Billion in 2013, and is projected to increase 19% per annum to $1.34 Billion in 2017. U.S. Processing and Production Service revenue represent 70% of the worldwide market. Processing and production services represent 19% of total eDiscovery Services revenue generally, which is $5.5 Billion in 2013, projected to increase 15.5% per annum to $9.8 Billion in 2017.[2] Two countervailing factors drive the revenue growth of this market; the amount of ESI needing to be processed and the cost to process it.

Processing Costs are Falling

The cost to process and produce a fixed volume of ESI has declined substantially over time, reflecting improvements in computational power and related efficiencies. The cost to process a GB of raw ESI (approximately 50,000 pages) in 2006 was $1,800, had declined to $500 by 2011, and today is $150 in volume for various types of processing, a 90% drop in 10 years.[3] The increasing aggregate processing and production costs associated with the rapidly expanding eDiscovery industry suggests that the volume of ESI to process is rising quicker than the cost to process and produce that data is falling. TIFFing is more processing intensive and is more expensive. Current market pricing for TIFF conversion is $300-$1,000/ per GB, or $.01-$.05 per page/image, depending on the vendor, volume and particular job specifics.



[1] Estimate by Litigation Support Today magazine (2014).

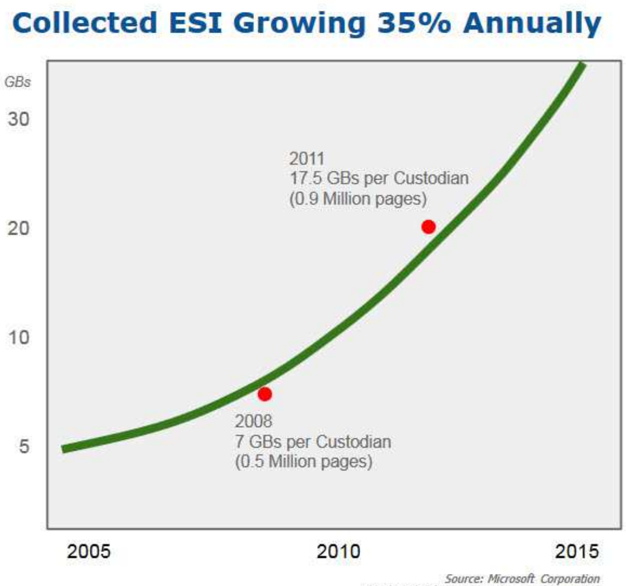
[2] Complex Discovery (www.complexdiscovery.com), aggregating estimates from Gartner, Inc. (2013), The Radicati Group (2012), Transparency Market Research (2012), Rand Institute For Civil Justice (2012) and IDC (2012).

[3] For 2006 and 2011 estimates: Murphy, Believe It - eDiscovery Technology Spending To Top $4.8 Billion By 2011 (Forrester Research, Inc. December 11, 2006).

ESI Collected, Processed and Produced is Increasing even Faster

The total amount of ESI GBs processed continues to grow quickly over time, as more ESI is created, stored, and potentially usable in litigation. As storage has become cheaper, more ESI is retained for longer periods and is then available for use in future litigation. Additionally, new types of potentially discoverable data are being created and stored, with social media sites and activities, text messaging, and cloud storage recently adding to the (relatively) older mix of email, computer files and databases.

A primary driver of increased ESI in litigation is the amount stored and collected from employees involved in commercial litigation. Organizations continue to try to manage the litigation risk and expense of retaining large amounts of potentially discoverable ESI, but the overwhelming tide of computerization and ubiquitous, cheap storage makes that a losing battle. Microsoft corporation, for example, recently found that its average collection per individual custodian involved in litigation increased from 7 GBs (~0.5 Million pages) in 2008 to 17.5 GBs in 2011 (~0.9 Million pages), an astonishing 150% increase in just three years (35% a year, compounded).[4]



Source: Microsoft Corporation
Legal Technology Leadership Summit (2011)

Rapidly increasing collection volumes of ESI and use of the data in litigation continues to outpace the improvements and volume-based cost decreases offered by eDiscovery Processing and Production Services, resulting in overall rising eDiscovery costs for law firms, private companies and governmental agencies and organizations. This means that even though per GB eDiscovery costs are falling to some degree, the volume of ESI data that must be handled as part of discovery is increasing at a much higher rate. The result is an overall increase in the total cost of eDiscovery.

Review Tools and ESI Processing Approaches

Different Processing Needed for Different Review Tools

Differing litigation document review tools require different types of processing to prepare documents for loading and review. Litigation review applications can generally be classified in four categories by how ESI must be prepared and processed before loading/ingesting: TIFF, PDF, Processed Natives and Raw Natives. Some review tools can load different types of ESI, but might work best or in high-volume with only certain types.

[4] Legal Technology Leadership Summit (2012).

‘TIFF-Based Review Applications’ were the first litigation document databases and review tools created and remain the most ubiquitous in law firms and organizations today. CT Summation™ (now AD Summation™) and LexisNexis Concordance™ were early specialty litigation review applications and updated versions remain popular today. These programs were initially developed in the 90s as specialty image, document management databases, and their software architecture was based on documents being scanned (and later converted from electronic sources) to single-page TIFF images, with text files representing OCR text and file metadata associated in ‘loadfiles’.[5] In addition to Concordance and Summation, other TIFF-based tools include CaseLogistix™, iConect™ and FTI Ringtail™. These tools generally require that ESI be processed first to single-page TIFF images, with associated text and load files, before loading to the litigation database and review tool.

‘PDF-Based Review Applications’ include many legal and non-legal industry specific document management systems, as Adobe’s portable document format is universally used in legal and other industries. PDF based systems work exclusively or best when Native documents are first converted to PDF, before loading and using. Examples of this class of application in the legal industry include WorldDox™ and LexisNexis CaseMap™, as well as Adobe Acrobat Pro™.

‘Processed Native Review Applications’ require that ESI be processed to a ‘Native Load file’ before being adding to the system. This requires less processing than TIFFing or PDFing, but still requires that raw ESI be expanded from containers, metadata extracted, email and attachments separated and associated, and load files generated. These applications then generate ‘near native’ versions of documents to review in HTML within the software. They may have an integrated native file viewer for a range of native file types, or require that native files be opened with other native applications installed on the user’s local computer. An advantage of this class of review tools is that pre-loading processing time and costs are reduced for processing systems that are not highly scalable. But the reviewable document in the program is in HTML format, or accessible only in a viewer, or must be downloaded in native format and opened, rather than a paginated and ‘Bates-stampable’ file version like TIFF or PDF. This can slow down document review when having a paginated or Bates-numbered copy is needed. Examples of programs of this class are Kcura Relativity™ and iPro Allegro™.

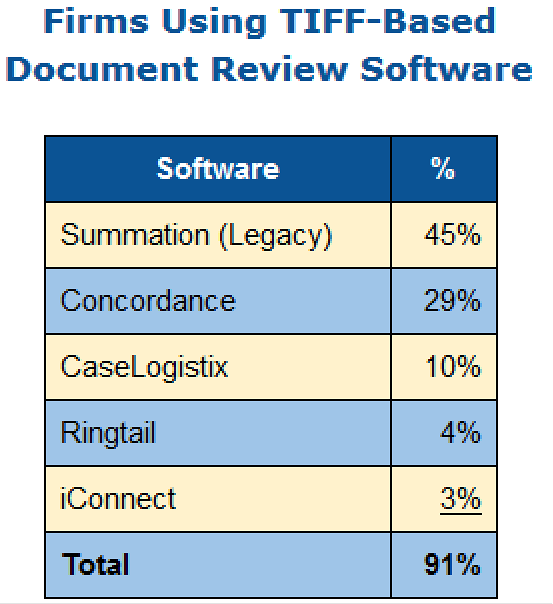
‘Native-based Litigation Review Applications’ are the most recent generation of review tools and allow loading of raw native files, which are processed within the program for review. Examples of these review tools are Digital Warroom™, Lexbe eDiscovery Platform, and Nextpoint™. These applications accept raw native files, and process to a viewable version in a paginated version, in PDF, TIFF or PNG within the application. These tools may also display text, HTML or native views as well within a document viewer.

[5] Loadfiles (or ‘load files’) legal-specific loosely typed databases, usually text-based, that associate the pages of a document (images and text), a Bates or other control number, and sometimes extracted file metadata and links to native file versions of documents.

Popularity of TIFF Review Applications

TIFF Review Application Usage

While in recent years the number of applications that allow effective review of documents without being TIFFed has increased, a majority of law firms still use TIFF-based application review tools in their practice. In a 2013 survey[6] of members, the International Legal Technology Association (ILTA) found that 91% of member law firms and corporations currently utilize in-house review tools that require pre-processing to TIFF images to load: Summation Legacy (45%), Concordance (29%), CaseLogistix (10%), FTI Ringtail (4%), iConnect (3%).



Firms Using TIFF-Based Document Review Software

Software	%
Summation (Legacy)	45%
Concordance	29%
CaseLogistix	10%
Ringtail	4%
iConnect	3%
Total	**91%**

Creating TIFF files has the advantage of providing a reviewable image file for attorneys to examine at fast review rates, without introducing the inefficiency of requiring attorneys to deal with multiple installed applications to review in various formats and avoid time-consuming complications if files do not open or view properly. As reviewing attorneys can charge from $100/hr to $700/hr or more in some cases, attorneys doing 'IT work' can be quite expensive and inefficient. A document reviewer may target completing review of 50 or more documents an hour, so any complication in getting documents to display immediately means that document review rates can plummet, with resultant increases in discovery costs and frustrations. RAND Corporation supports this conclusion in its finding that attorney review costs are 70% of eDiscovery costs, with processing costs representing only 19%.[7]

While a controversial subject among lawyers and litigation support professionals, TIFF-based review systems have remained popular among law firms, companies and users for a number of other reasons as well:

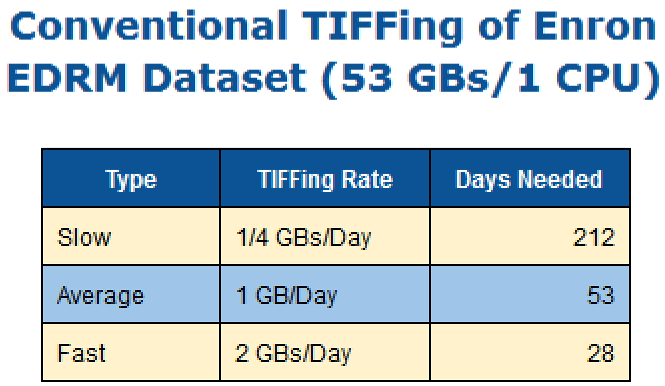
- Bates numbers can be applied at a page level to TIFFs (and PDFs as well), but native files or HTML near native versions can only be Bates named at the file level
- Concerns that opposition may inadvertently (or intentionally) alter native files
- Easy redaction of TIFF files
- Nothing but a TIFF viewer is needed to view TIFF images

Law firms may also prefer the ability to host most TIFF-based review tools in-house and support with their law firms' own in-house litigation support staff.

[6] 2013 Technology Survey of the International Technology Support Association
[7] RAND Corporation, Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery (2012).

The Challenge of Fast TIFF Processing

Processing and converting native files to TIFF images for litigation review tools that require TIFFs has traditionally been an expensive and time-consuming process, as TIFFing requires I/O and CPU-intensive imaging and rasterization[8] of documents.



Conventional TIFFing of Enron EDRM Dataset (53 GBs/1 CPU)

Type	TIFFing Rate	Days Needed
Slow	1/4 GBs/Day	212
Average	1 GB/Day	53
Fast	2 GBs/Day	28

As the amount of ESI has continued to explode, the requirement that TIFF processing be done quickly and inexpensively for law firms using TIFF-based review tools has increased. While many applications on the market can convert native files to TIFFs, they run generally as stand-alone applications, are slow, and require regular user (Litigation Support or IT) supervision and intervention (often referred to as 'babysitting' the application) to maintain a constant throughput rate. Industry estimates are that processing rates to convert native files to TIFF can be done at an average rate of about 1 GB per day per computer, but ranging from .25/GBs a day to 2 GBs a day, depending on the computer used, the software, the original ESI, and the availability and continued focus of the litigation support operator. This means that converting the 53 GBs of native files in the Enron EDRM Dataset to TIFF, by no means an extra-ordinarily sized data collection today, can take from 28 to over 200 machine days.

The time allowed for document review in modern commercial litigation cannot be extended easily. Even with drastically larger data volumes, law firms are under pressure to have ESI processed and ready for review in a few days, rather than in weeks or months. Of course, multiple machines can run simultaneously on large jobs, and this may decrease total processing time, but using too many machines adds costs and inefficiencies to the processing job. Running multiple machines concurrently also requires increased capital and operator cost including extensive user set-up, batching, monitoring, reassembly, and additional quality control to complete a job. A large job must be batched out to separate machines, and then accurately reassembled on completion. Individual machines must be separately monitored and quality control procedures become more complicated, time-consuming and error-prone. Each of the separate machines must be kept updated with processing, operating system and other software. All this makes running a job like the Enron EDRM Data Set in a few days very difficult using conventional hardware and software approaches.

[8] Rasterization is the process of converting an image stored as an outline into pixels that can be displayed on a screen or printed. When used in connection with TIFFing, it refers to converting letters stored as vector outlines (e.g., text in a Microsoft Word document - .DOC) into pixels as part of a an image file (e.g., image displaying text in a TIFF document).

Lexbe eDiscovery Processing System

Hardware and Software

Lexbe operates a proprietary eDiscovery processing and production software system, called the Lexbe eDiscovery Processing System (“LEPS”). LEPS runs as a scalable, automated, fault-tolerant cloud-based service, designed to handle large processing jobs quickly and with a minimum of operator supervision and interaction. The LEPS is run on Amazon Web Services (“AWS”) and utilizes the Amazon Elastic Compute Cloud ("EC2"), Amazon Simple Storage Service (“S3”), and Elastic Block Store (“EBS”) technology to facilitate automation, scalability and redundancy in a secured environment. LEPS consists of a database, queue controller, and multiple coordinated processing servers, all accessing shared storage. LEPS scales by dynamically starting and stopping server instances as needed, so resources are efficiently used, and server and IT staffing costs are not incurred when jobs are not being run.

LEPS is able to make use of other architecture advantages to speed litigation data processing in general, and TIFFing in particular. An example is the use of the latest generation of Solid State Drives on its servers. This innovative storage architecture substantially speeds data access, up to 4,000 IOPS[9]. A traditional workstation would generally have a single magnetic drive, with only 100-150 IOPS, and perform much slower.

Data Security

LEPS maintains security using the Amazon AWS, and incorporates EC2, Amazon S3, and EBS technologies to this end. Servers are maintained in secure, limited access, physically isolated data centers, and monitored from network operating centers 24x7. AWS data center personnel do not have logical access to LEPS or to the processed data. Data is transmitted and stored using strong 256-bit (AES-256) encryption. LEPS operates redundant firewalls to protect servers and data from potentially malicious network traffic.

LEPS servers and networking components are redundantly configured for persistent reliability. Amazon AWS data centers provide Service Organization Controls 1 & 2 reports published under SSAE 16 and ISAE 3402 professional standards (formerly SAS70 Type II audits). This enables HIPAA covered entities to leverage the secure environment to process, maintain, and store protected health information. Additionally, the data centers have achieved ISO 27001 certification, and have been successfully validated as a Level 1 service provider under the PCI Data Security Standard (DSS).

[9] IOPS is an acronym for 'Input/Output Operations Per Second' and is a common performance measurement used to benchmark computer storage devices, including hard disk drives, solid state drives, and storage area networks.

eDiscovery Processing Functions

LEPS takes raw ESI as input and prepares, processes, and converts ESI to be ready to load in industry standard eDiscovery review platforms. LEPS provides the following eDiscovery Processing functions: Archive/Container decompression, file repair, metadata extraction, MD5 hash code generation, system file identification and deNISTing, email attachment extraction and parent email association, native text extraction, optical character recognition (OCR) of image files, full-text indexing, Bates stamping, PDF & TIFF creation, placeholder creation, Native Extracted, PDF and TIFF loadfile generation in multiple formats: XLSX (Lexbe), DAT/OPT (Case Logistix, Concordance, iPro Allegro, Ringtail, Kura Relativity) and DAT (Summation), and quality control reports.

While the Performance Demonstration involved Native to TIFF, the LEPS is a flexible processing and production system and can also process to Native Extracted and PDF, for review systems that require or work best with those file formats.

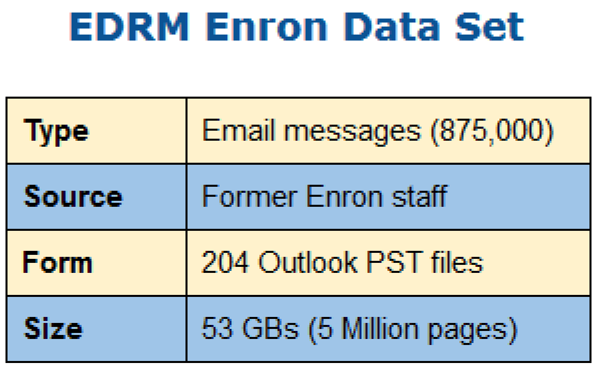
Quality Control

LEPS includes tools to check the quality of processing, OCR and productions. These include:

- Programmatic batching of processing to individual servers (reduces human error)
- Custom QC flag creation and filtering
- Integration with Excel for reporting and analysis
- Pivot table analysis and charting
- Ability to view all documents including parent containers (email and attachments) together
- Ability to verify image quality
- Filtering and reporting by any captured or calculated fields including failed to convert, words in document, placeholders, etc.
- Native files are extracted and provided for linked load and review
- Statistical sampling and reporting

Demonstration Methodology & Results

<u>The EDRM Enron Data Set</u>



EDRM Enron Data Set

Type	Email messages (875,000)
Source	Former Enron staff
Form	204 Outlook PST files
Size	53 GBs (5 Million pages)

The EDRM Enron Data Set is an industry-standard collection of email data that has been used for many years for electronic discovery performance testing and training. The EDRM Enron data set is maintained by the EDRM Data Set Project[10] and “provides industry-standard, reference data sets of electronically stored information (ESI) and software files that can be used to test various aspects of e-discovery software and services.” This data set was collected from the Federal Energy Regulatory Commission’s investigation into the collapsed energy firm, Enron. The data consists of email messages sent and received by Enron staff in the course of day-to-day business. The EDRM Enron Data Set consists of approximately 875,000 email messages with attachments, organized in 204 Microsoft Outlook PST[11] files by Custodian[12], aggregating 53 GBs of ESI.

<u>Performance Demonstration</u>

The Performance Demonstration started with the data from the EDRM Enron Data Set, in raw PST files, and the above eDiscovery Processing functions were applied. This consisted of ingesting the 204 PST files (53 GBs) from the EDRM Enron Data Set into LEPS, running the processing steps above, and outputting extracted and expanded Native files and single-paged TIFF files, suitable for use in TIFF-based litigation review applications.

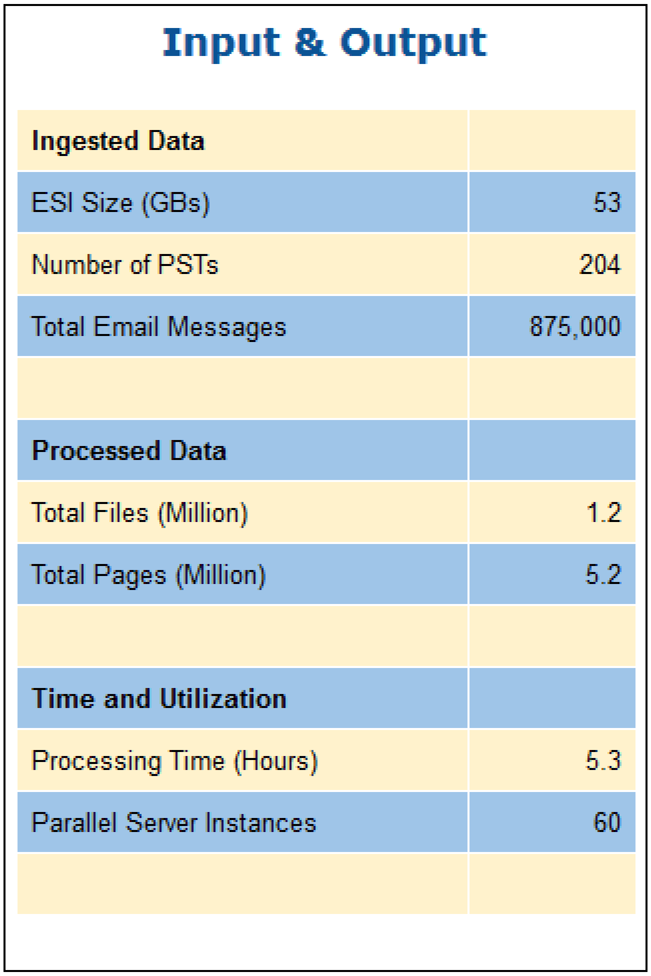
[10] http://www.edrm.net/projects/dataset

[11] PST is an acronym for 'Personal Storage Table' and is an Microsoft developed file format used to store copies of email messages (MSGs), calendar events, and other items within Microsoft Outlook.

[12] A custodian is a person who has responsibility or control over ESI data. Often the sender and recipients of email are custodians of email they have sent or received (e.g., it remains on their computers in PST email archive files).

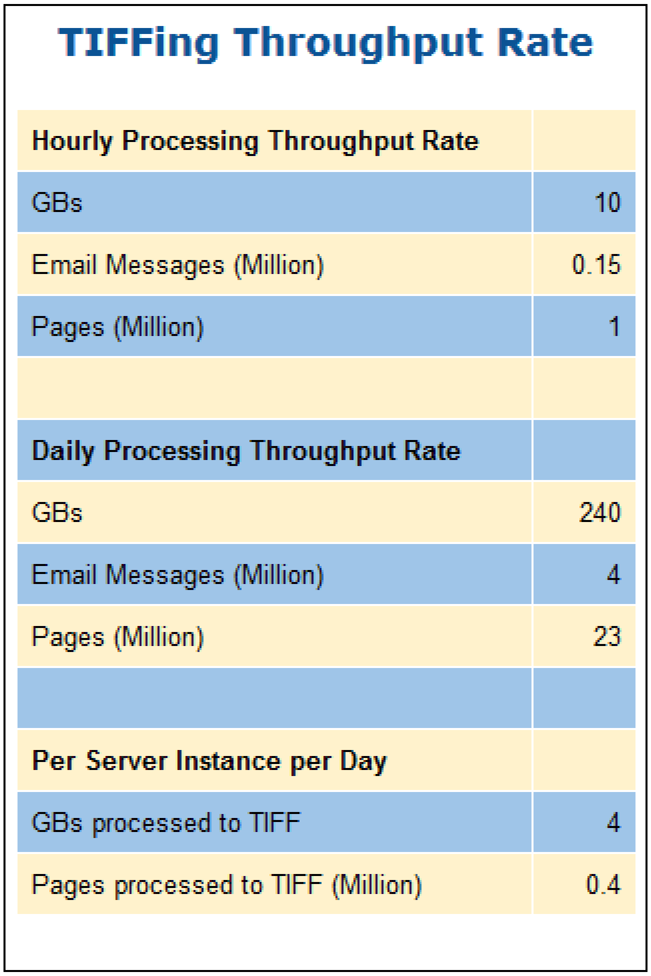
Performance Results

The following tables summarize the results from the performance demonstration.



Input & Output

Ingested Data	
ESI Size (GBs)	53
Number of PSTs	204
Total Email Messages	875,000
Processed Data	
Total Files (Million)	1.2
Total Pages (Million)	5.2
Time and Utilization	
Processing Time (Hours)	5.3
Parallel Server Instances	60



TIFFing Throughput Rate

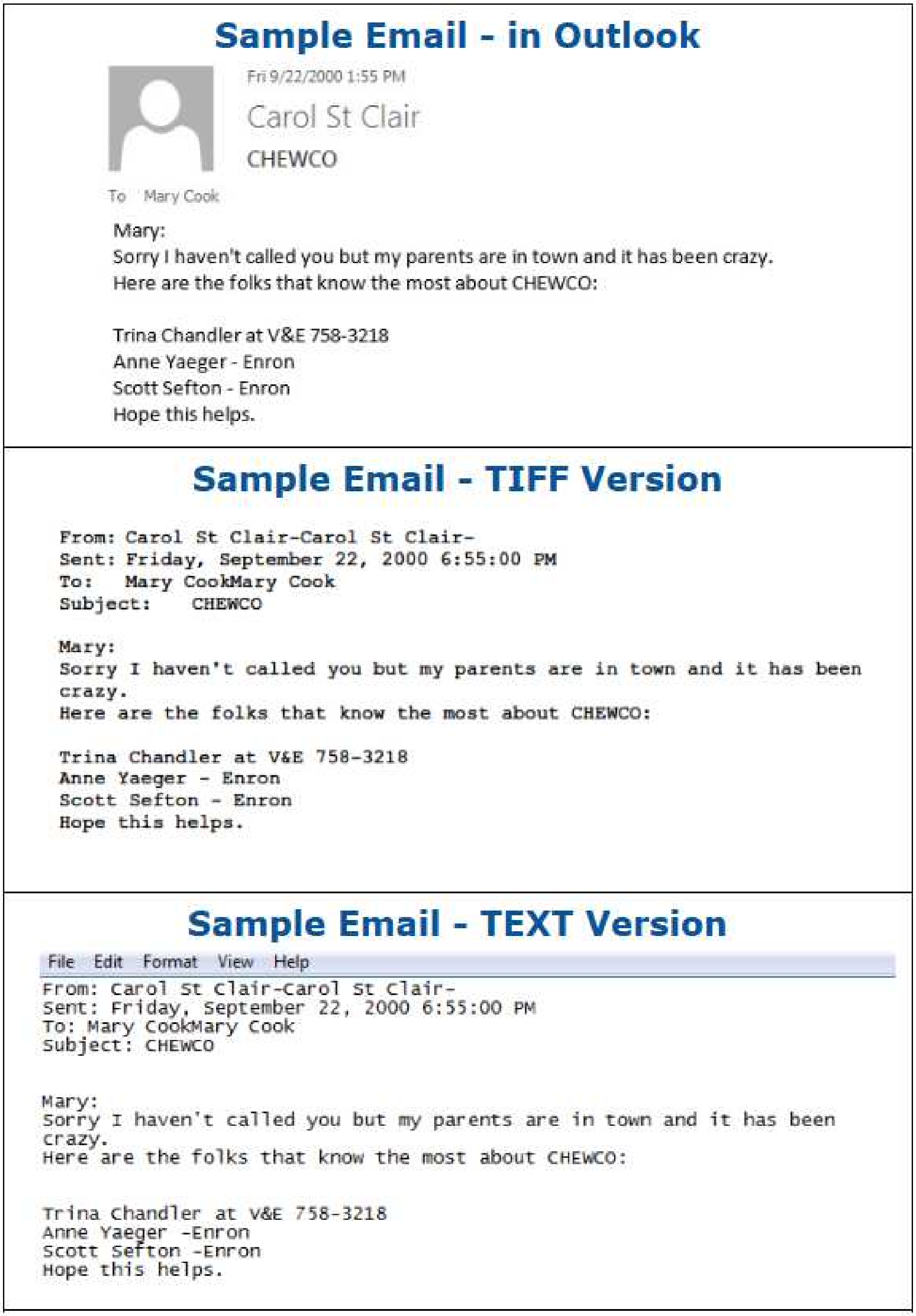
Hourly Processing Throughput Rate	
GBs	10
Email Messages (Million)	0.15
Pages (Million)	1
Daily Processing Throughput Rate	
GBs	240
Email Messages (Million)	4
Pages (Million)	23
Per Server Instance per Day	
GBs processed to TIFF	4
Pages processed to TIFF (Million)	0.4

60 servers were operated in parallel and programmatically started, run for the job, and terminated. The entire job was completed on these servers in 5.3 hours and consisted of Archive/Container decompression, file repair, metadata extraction, MD5 hash code generation, system file identification and deNISTing, email attachment extraction and parent email association, native text extraction, optical character recognition (OCR) of image files, full-text indexing, placeholder creation, and TIFF creation.

The hourly processing throughput rate was 10 GBs and 1 million pages. This equates to a daily throughput rate of 240 GBs, or 23 Million pages (with rounding). Each of the 60 servers averaged 4 GBs per day, or approximately 400,000 pages. This 4 GB per server instance per day rate is several times faster than the typical rate of .5 - 2 GBs a day with traditional eDiscovery processing software performing TIFFing functions (see above).

Sample Input and Output

The following shows a sample email from the EDRM dataset in three formats: 1) the original MSG format, as displayed in Outlook, 2) the TIFFed version after processing with LEPS, and 3) the extracted text version after processing with LEPS.



Sample Email - in Outlook

Fri 9/22/2000 1:55 PM

Carol St Clair

CHEWCO

To Mary Cook

Mary:
Sorry I haven't called you but my parents are in town and it has been crazy.
Here are the folks that know the most about CHEWCO:

Trina Chandler at V&E 758-3218
Anne Yaeger - Enron
Scott Sefton - Enron
Hope this helps.

Sample Email - TIFF Version

```
From: Carol St Clair-Carol St Clair-
Sent: Friday, September 22, 2000 6:55:00 PM
To:   Mary CookMary Cook
Subject:   CHEWCO

Mary:
Sorry I haven't called you but my parents are in town and it has been
crazy.
Here are the folks that know the most about CHEWCO:

Trina Chandler at V&E 758-3218
Anne Yaeger - Enron
Scott Sefton - Enron
Hope this helps.
```

Sample Email - TEXT Version

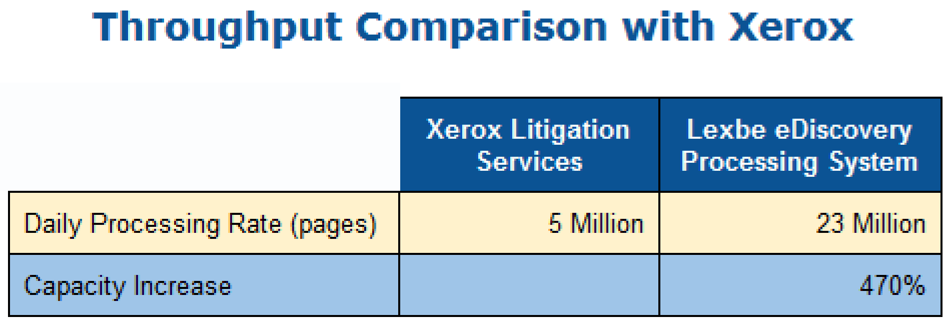
```
From: Carol St Clair-Carol St Clair-
Sent: Friday, September 22, 2000 6:55:00 PM
To: Mary CookMary Cook
Subject: CHEWCO


Mary:
Sorry I haven't called you but my parents are in town and it has been
crazy.
Here are the folks that know the most about CHEWCO:


Trina Chandler at V&E 758-3218
Anne Yaeger -Enron
Scott Sefton -Enron
Hope this helps.
```

Comparison with Xerox, a Traditional High-Capacity eDiscovery Service Provider

Xerox Litigation Services is one of the largest service providers in the industry, with worldwide distributed data centers, 80 IT/Litigation Support professionals, and 250 employees overall. Xerox Litigation Services is known for its high-volume processing and production capacity. Xerox states in its service literature that its production capacity is 5 million pages a day[13]. As the Enron 53 GB data set processes and produces in excess of 5 Million pages, this means Xerox would require using its entire production capacity for over a full day, assuming Xerox could quickly bring to bear this capacity across its data-centers and coordinate its personnel. Lexbe has over 4X this capacity with LEPS, at 23 Million pages a day without the need for a large litigation support staff.



Throughput Comparison with Xerox

	Xerox Litigation Services	Lexbe eDiscovery Processing System
Daily Processing Rate (pages)	5 Million	23 Million
Capacity Increase		470%

Conclusion

Lexbe has created an industry-leading eDiscovery processing and production system with LEPS, as illustrated by the demonstration described in this white paper. A large real-life set of files, the EDRM Enron Dataset, was processed and converted to TIFF in just a few hours. As TIFFing is one of the most difficult data conversion and processing challenges there are in eDiscovery, these results are impressive and unmatched by even large vendors using traditional methodologies.

[13] Source: Xerox Litigation Services Fact Sheet (available 5/23/2014, at http://www.xerox-xls.com/pdf/xls-fact-sheet.pdf).

The advantage of a scalable cloud-based eDiscovery processing solution like LEPS to law firms, companies, organizations and governmental agencies is compelling:

- Save time and accelerate case timelines by processing large and complex data in hours and days vs. weeks.
- Save time and costs for human resources by automating the setup and scaling of computing resources, eliminating the time and people cost to monitor eDiscovery processing jobs.
- Save on capital expenditures by substituting dedicated in-house resources (which can be idle much of the time) with low cost, cloud-based computing resources provisioned only when needed.
- Reduce litigation database hosting costs for law firms, companies, and governmental agencies by enabling TIFF-based document review on data collections larger than traditionally thought, with faster load and less costly conversion.
- Increase flexibility to meet rapidly changing demands and case priorities for processing.
- Increase ability to meet case deadlines.
- Significantly lower industry processing costs by moving to low cost cloud-based infrastructure.
- Take advantage of industry leading technology to to increase eDiscovery efficiency, lower costs, and increase processing accuracy and quality control.

Contact:

Gene Albert
Lexbe LC
8701 MoPac Expressway
Suite 320
Austin, TX 78759
512-686-3460 (direct)
gene@lexbe.com