eDiscoveryJournal

eDiscovery Journal Report: Digital Reef & BlueArc eDiscovery Software Performance Benchmark Test

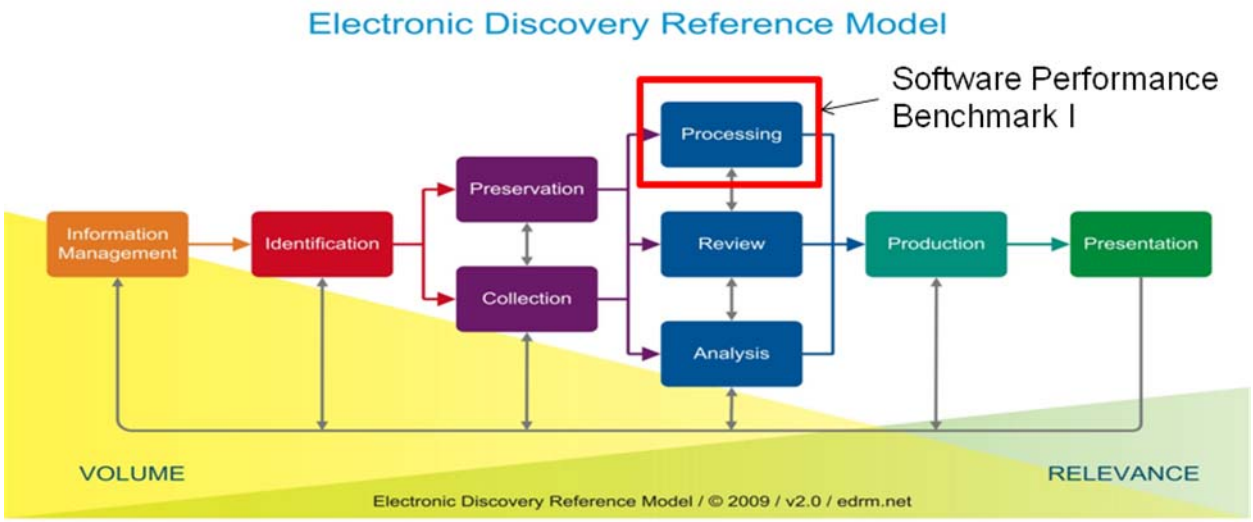
eDiscoveryJournal was engaged to review the testing methodology, execution, and results of the Digital Reef / BlueArc solution for certain components of eDiscovery.

Background

In the eDiscovery market, processing speed and scale are increasingly important given the volume and diversity of data sources. Digital Reef and BlueArc set out to create an open eDiscovery software performance benchmark test on a standardized set of eDiscovery data. The objective of the testing is to provide a software performance baseline for some of the most important aspects of along the Electronic Discovery Reference Model (EDRM).

The software performance benchmark (SPB) test was created to assess how eDiscovery activities might be performed in the real world. This particular testing scenario focuses specifically on data processing – creating a central index that allows organizations to quickly determine how much potentially responsive information exists as well get an idea as to the make-up of the information (e.g. types of files). Such an index can benefit both upstream eDiscovery activities (information management, identification, collection, and preservation) and downstream eDiscovery activities (analysis, review, and production).

The Open SPB I is a specific test suite focused on EDRM Processing. The test suite covers data processing for two levels: metadata level and full content indexing and classification.

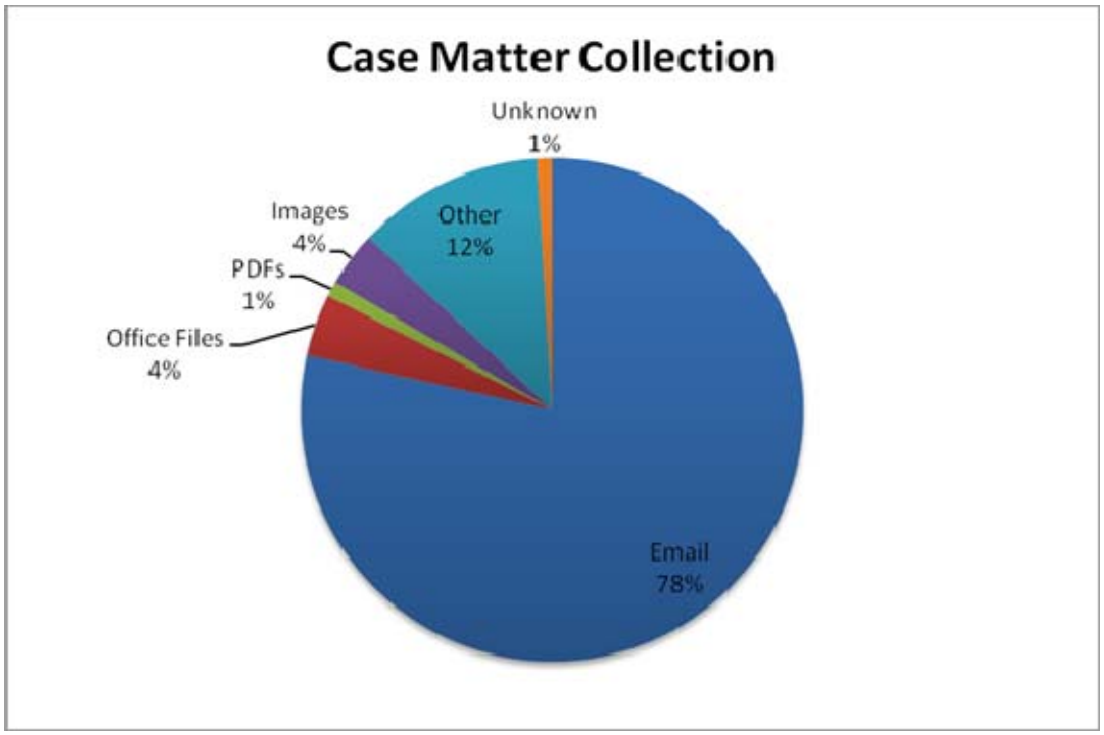


Testing Methodology

The goal for the software performance benchmark for eDiscovery processing was to show the processing speed and scale of the solution (as it might be used in the real world). To that, two data sets were created – the first to mirror the composition of the data set of an actual client matter, and the second to create a much larger data set typical of larger litigation matters.

Data Set Creation

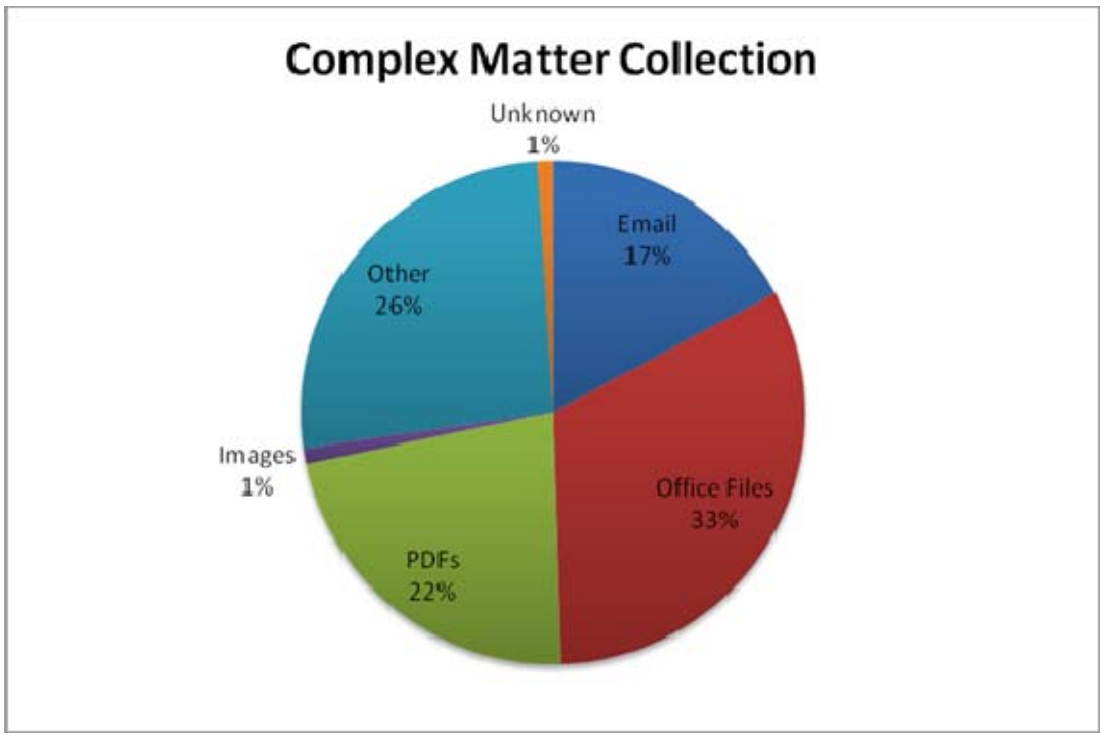
Two sets of data were created for the performance testing. The "case matter example" set of data was created to mirror the composition of a set of actual case data from a Digital Reef client. This data set had the following composition:



# of Files:	31 million
Data Collection Size:	1.2 TB; 2.3 TB when expanded (exploding out all archive files, e.g. PST ZIP, NSF, and pulling out attachments and embedded objects recursively to ensure complete capture)

Note: "other" includes files such as audio, e.g. mp3, and video, e.g. avi

The "complex matter sample" set of data was created for the purpose of having a large data set with the composition typical of many large litigation matters. It models what organizations typically find when crawling Electronically Stored Information (ESI) behind the firewall – many large files and fewer archive file formats like ZIP. To create this data set, Digital Reef used publicly available content, capturing large data sets from various organizations that publish material in standard format. Digital Reef used internet search engines to identify the remaining content necessary to make the data population fit the target composition profile. The crawls took place over a period of several months and resulted in a data set with the following composition:



# of Files:	9.8 million (smaller number of files than the case matter example set because the files themselves are larger and this data set had very few archive file formats such as ZIP and PST)
Data Collection Size:	4 TB; 4.2 TB when expanded

Note: "other" includes files such as audio, e.g. mp3, and video, e.g. avi

Testing Environment

The laboratory test environment consisted of both compute resources and storage, as represented visually below:

Access & Service Manager:

1x Dell 2950
OS: CentOS 5.4 (64-bit)
CPU: 4 Core 3Ghz
RAM: 16 GB RAM
Disk: 146 GB @ 10k (Raid 1)
Network: 1 Gb E

Content Analysis Servers:

17x Dell R710
OS: CentOS 5.4 (64-bit)
CPU: 4 Core 2.6Ghz
RAM: 16 GB RAM
Disk: 450 GB @ 7.2k (Raid 1)
Network: 1 Gb E

and

2x Dell 2950
OS: CentOS 5.4 (64-bit)
CPU: 4 Core 3Ghz
RAM: 16 GB RAM
Disk: 146 GB @ 10k (Raid 1)
Network: 1 Gb E

Network Switch:

Make: Dell
Model: 6224 PowerConnect
Add-on: 10Gb XFP Module

Software:

The software used in the testing was Digital Reef eDiscovery Solution version 3.1.

Storage:

Make: BlueArc
Model: Mercury 100
Disks: 96x 450G, 15K RPM drive
Storage: 17.3 TB

Note, the storage used in eDiscovery operations is critically important. If the storage system can't keep up with the demands imposed on it by a high-volume processing, both the speed and the quality of the final output will suffer. Data throughput performance will be stressed, for example, in enterprise-scale operations like eDiscovery that utilize tens or hundreds of clients hosted on modern multi-threaded, multi-core CPUs. Converting a number of large (~300GB or greater) source files to a much larger number of smaller target files requires a storage platform with capacity for high IOPS, as well as the capacity to perform many multiple concurrent read/write operations.

Some storage solutions don't provide the ability to accommodate the kinds of load balancing this shift in workload required to get optimal performance. Intelligent NAS solutions provide comprehensive virtualization tools that simplify the administration of file system functions, for example, allowing storage read blocks to be resized and fine-tuned for optimal performance. Quality of Service (QoS) is another area where storage performance plays a critical role across the entire content creation and delivery ecosystem. Because no software or hardware is 100 percent fail proof, it's essential that enterprise-class hardware and software offer failover support.

The software used in the testing was Digital Reef version 3.1.

The goal of the test environment was to create a laboratory composed of industry standard infrastructure indicative of the environment in many organizations' data centers.

Results of the Performance Testing

Digital Reef conducted the performance testing at the company's development center near Boston, MA. Testing began in mid-June 2010 and ended on July 31, 2010. The company used one development engineer and one lab engineer to configure, run, and report on the testing. For this software performance benchmark focusing on eDiscovery processing, the steps included: reading the files; exploding archives; and extracting email attachments and OLE embedded docs; NIST file detection; and duplicate file detection (dupes remain in the data set for full chain-of-custody reporting). While the search index is built, pattern detection is run to allow for, among other things, such activities as finding private/confidential information. When the processing steps are completed, a fully prepared, searchable case data set is available to work with.

The results of the software performance benchmark were as follows:

Case Matter Sample

Data set size: 2.3 TB
Processing time: 7 hours, 47 minutes
Processing Rate: 7 TB per day

Complex Matter Sample

Data set size: 4.2 TB
Processing time: 5 hours, 49 minutes for a full file content index
Processing Rate: 17.3 TB per day

Configuration Environment

Profile	Size (original / expanded)	# Files	Composition
Case Matter Collection	1.2 TB / 2.3 TB	31M	Email 78.5% Office Files 4% PDFs 1% Images 3.5% Other 12% Unknown 1%
Complex Matter Collection	4 TB / 4.2 TB	9.8M	Email 17% Office Files 32% PDFs 22% Images 1% Other 26% Unknown 1%

Results

Profile	Level	Size	Actual Hours	*Digital Reef v3.1 Results*
Case Matter Collection	Full file content	2.3TB	7:47 hours	***7 TB / Day***
Complex Matter Collection	Full file content	4.2TB	5:49 hours	***17.3 TB / Day***

Results Summary

All processing rates are stated in *volume* per day. Standardizing on this format simplifies performance comparisons across various testing configurations, e.g. different software versions or test suites. To arrive at the standard number, the SPB I participants calculate performance using the following formula:

(volume in Terabytes / minutes to process) * minutes in a day = volume in TB/ day

Conclusion

eDiscovery has become a massive legal expense and challenge for corporations, service providers, and law firms alike. Companies are spending billions of dollars on information management, identification, and the processing of electronically stored information, and in many cases they are doing so ineffectively and with software and storage infrastructure not well-suited for the task of processing massive quantities of diverse types of files. Being able to eliminate storage bottlenecks when processing these files for evaluation and review, and having the best software designed to collect, cull, process, analyze, and produce data for eDiscovery is critical. Any organizations seeking to deploy eDiscovery applications for activities like identification, collection, preservation, processing, or analysis should demand an apples-to-apples ERDM-focused comparison for evaluating their choices. This test is one such example and can be used as a template for others seeking a representative benchmark.

For Release August 19, 2010