Veritas eDiscovery Platform™

Performance Guide

10.1

VERITAS

Veritas eDiscovery Platform™: Performance Guide

The software described in this book is furnished under a license agreement and may be used only in accordance with the terms of the agreement.

Last updated: 2022-01-25

Legal Notice

Documentation by the U.S. Government shall be solely in accordance with the terms of this Agreement.

Veritas Technologies LLC
2625 Augustine Drive
Santa Clara, CA 95054
http://www.Veritas.com

Contents

Chapter 1

About this document

Many factors influence the performance of eDiscovery Platform. For example, the size of data that is being processed, the file types, caching, tagging, and so on. Every customer has different environment that can affect these factors resulting in varied performance of eDiscovery Platform.

This document provides guidelines on performance expected while running Veritas eDiscovery Platform. Performance details of the individual eDiscovery features are provided in separate sections.

Note: This document provides the details only for deployments with distributed architecture. The details for standalone deployment would be added in near future.

Supporting documents

The table below lists the end-user documentation that is available for the Veritas eDiscovery Platform product.

Veritas eDiscovery Platform documentation set

Document	Comments
Installation and Configuration	
Installation Guide	Describes prerequisites, and how to perform a full install of the Veritas eDiscovery platform.
Upgrade Overview Guide	Provides critical upgrade information, by version, useful prior to upgrading an appliance to the current product release
Upgrade Guide	Describes prerequisites and upgrade information for the current customers with a previous version of the Veritas eDiscovery platform.

Document	Comments
Utility Node Guide	For customers using utility nodes, describes how to install and configure appliances as utility nodes for use with an existing eDiscovery Platform software setup.
Distributed Architecture Deployment Guide	Provides installation and configuration information for the Review and Processing Scalability feature in a distributed architecture deployment
Getting Started	
Navigation Reference Card	Provides a mapping of review changes from 10.x compared to 9.x, 8.x compared to 7.x and the user interface changes from 7.x compared to 6.x
Administrator's QuickStart Guide	Describes basic appliance and case configuration
Reviewer QuickStart Guide	A reviewer's reference to getting started using the *Analysis & Review* module
Tagging Reference Card	Describes how tag sets and filter type impact filter counts
User and Administration	
Legal Hold Guide	Describes how to set up and configure an appliance for Legal Holds, and use the Legal Hold module as an administrator.
Identification and Collection Guide	Describes how to prepare and collect data for processing, using the Identification and Collection module
Case Administration Guide	Describes case setup, processing, and management, plus pre-processing navigation, tips, and recommendations. Includes processing exceptions reference and associated reports, plus file handling information for multiple languages, and supported file types and file type mapping.
System Administration Guide	Includes system backup, restore, and support features, configuration, and anti-virus scanning guidelines for use with Veritas eDiscovery Platform
Load File Import Guide	Describes how to import load file sources into Veritas eDiscovery Platform

Document	Comments
User Guide	Describes how to perform searches, analysis, and review, including detailed information and syntax examples for performing advanced searches
Export and Production Guide	Describes how to use and produce exports, productions, and logs (privilege and redaction logs).
Imaging Tool Upgrade Guide	Provides details about the Imaging Tool Upgrade feature and how to perform Imaging Tool Upgrade after the eDiscovery Platform appliance is upgraded to version 10.1, workflows affected when the cases are upgraded or not upgraded, and frequently asked questions (FAQs).
Transparent Predictive Coding User Guide	Describes how to use the Predictive Coding feature to train the system to predict results from control set data and tag settings.
Audio Search Guide	Describes how to use the Audio Search feature to process, analyze, search and export search media content
Reference and Support	
Audio Processing	A quick reference card for processing multimedia sources
Audio Search	A quick reference card for performing multimedia search tasks
Legal Hold	A quick reference card of how to create and manage holds and notifications
Collection	A quick reference card of how to collect data in Veritas eDiscovery Platform
OnSite Collection	A quick reference for performing OnSite collection tasks
Review and Redaction	Reviewer's reference card of all redaction functions
Keyboard Shortcuts	A quick reference card listing all supported shortcuts
Production	Administrator's reference card for production exports
Online Help	
Includes all the above documentation (excluding Installation and Configuration) to enable search across all topics. To access this information from within the eDiscovery Platform user interface, click **Help**.	

Document	Comments
Release	
Release Notes	Provides latest updated information specific to the current product release

For the latest product information: https://www.veritas.com/product/a-to-z.html

Technical Support

Technical Support maintains support centers globally. All support services will be delivered in accordance with your support agreement and the then-current enterprise technical support policies.

For information about our support offerings and how to contact Technical Support, visit our website:

https://www.veritas.com/support

You can manage your Veritas account information at the following URL:

https://my.veritas.com

If you have questions regarding an existing support agreement, please email the support agreement administration team for your region as follows:

Worldwide (except Japan): CustomerCare@veritas.com

Japan: CustomerCare_Japan@veritas.com

Documentation

Make sure that you have the current version of the documentation. The latest documentation is available from:

- Documentation link at the bottom of any page in the eDiscovery Platform landing page.
- Veritas Products Web site: https://www.veritas.com/product/a-to-z

Documentation Feedback

Your feedback is important to us. Suggest improvements or report errors or omissions to the documentation. Include the document title, document

version, chapter title, and section title of the text on which you are reporting. Send feedback to: eDiscovery.InfoDev@veritas.com

You can also see documentation information or ask a question on the Veritas community site: https://vox.veritas.com/

Chapter 2

Setup Requirements

Following are the setup requirements for distributed architecture deployments:

- MySQL database server (shared remote database) must meet requirements stated in *Distributed Architecture Deployment Guide*.
- Nodes must be part of a cluster and communicating to a shared remote database.
- All data sources must be visible to all review nodes and/or processing nodes, including contained PSTs/NSFs and converted files.

Hardware Requirements

In a distributed architecture, eDiscovery Platform supports the following types of appliances, their corresponding functions, and minimum hardware requirements for physical or virtual machines.

Table 1: Appliance Compatibility Options

Appliance Type	Function	CPU and RAM**	Disc##	Cluster required	VM*
Utility	Caching, Retrieval, Export	8-Core CPU x 32-GB RAM	500-GB on D:	N	Y
Case Home and Processing	Appliance or node where case was first created and Processing	32-Core CPU x 128-GB RAM	1.5-TB (1500 IOPs)	Y	Y
Review and Processing	Review and Processing	32-Core CPU x 128-GB RAM	1.5-TB (1500 IOPs)	Y	Y
Case Home	Appliance or node where case was first created	16-Core CPU x 64-GB RAM	1.5-TB (1500 IOPs)	Y	Y

Review Only	Review	16-Core CPU x 64-GB RAM	1.5-TB (1500 IOPs)	Y	Y
Processing Only	Processing	16-Core CPU x 32-GB RAM	1.5-TB (1500 IOPs)	Y	Y
Standalone Primary	Cluster Primary node (minimum	32-Core CPU x 128-GB RAM	1.5-TB (1500 IOPs)	Y	Y
Shared Remote Database MySQL	Database Server	32-Core CPU x 128-GB RAM	1.5-TB (1500 IOPs)#	Y	N***

Note:

* VM performance is lesser than a physical machine.

** An appliance configuration of 16-Core CPU x 64-GB RAM when the appliance is not used for jobs such as Caching, Production/Export, and Bulk redaction. An appliance with this configuration should not be used for any imaging-related jobs. Also, the Imaging role should be disabled. The appliance configuration must be 32-Core CPU x 128-GB RAM for using the appliance for Imaging-related jobs with the Imaging role enabled.

*** A physical machine is recommended, but if local IT policies demand, you can use a VM of equivalent performance.

You might need 2-TB or more storage space if the number of cases and the number of items are too high and if multiple concurrent activities happen on the system.

We recommend using SAN storage, but NAS can also be used if it satisfies the IOPS requirements.

Important: Existing appliances can be repurposed for a distributed architecture deployment. For information about other appliance types, contact your Solutions Consultant or Technical Support.

Chapter 3

Considerations for Deployment on VMware

If the following considerations are supported, then it is likely that deploying eDiscovery Platform in a VMware environment is a good fit for your organization:

- eDiscovery Platform is heavily dependent on CPU and memory resources. In a typical server configuration, it is normal for the CPU to run at 90% or higher utilization while ingesting data, running an OCR job, or exporting data.

  Generally, the more powerful the processor, the better the ingestion rates and retrieval rates.
- For an eDiscovery Platform server running Collections, Legal Holds and Cluster Master (with no cases), Pre-processing, Processing, Analysis, and Review, the minimum recommended configuration is as follows:
  - 32-Core CPU
  - 128-GB RAM
- We recommend that CPU and memory resources are dedicated (reserved) and locked to the eDiscovery Platform server, and not shared with other virtual machines on the host.
- Other system components such as network and storage need to be sized accordingly to prevent them from becoming a bottleneck. Certain recommendations on configuring ESXi environment are mentioned in the following section.

Sizing eDiscovery Platform for VMware

While sizing eDiscovery Platform, a thorough understanding of the expected workload on each of the eDiscovery Platform servers is important. The primary consideration should be given to customer requirements for collecting, processing, reviewing, and exporting.

It is outside the scope of this document to provide a design and sizing introduction to eDiscovery Platform. But in general terms, once the customer requirements are understood, a close look at the function of each eDiscovery Platform server helps to determine what minimal server resources are required.

A common mistake while designing eDiscovery Platform is to size for capacity, as opposed to size for performance. The following sections provide details on how to design various components for optimal configuration.

Guidelines to set up VMware environment

Following guidelines to set up VMware environment serve to get better performance from eDiscovery Platform.

Configuring ESXi

The BIOS should be up to date. All virtualization features such as VT-x should be enabled in the BIOS. For best performance, power-saving features should be disabled. Unused hardware features, such as parallel ports, should be disabled. The settings that are chosen in the BIOS screens can have a substantial effect on the performance of the machine. If your hardware does not perform to your expectations, consult the vendor documentation or support services for information on the optimal configuration of your hardware.

- Do not enable Hyperthreading—in most cases, hyperthreading provides little or no benefit to multi-CPU virtual machines. Internal testing has shown that Hyperthreading provides no performance benefit.
- Ensure that the total number of vCPUs assigned to the virtual machines is equal or less than the total number of cores on the ESX host.
- For all other hardware and software recommendations, refer to *Veritas eDiscovery Platform Installation Guide*.

- Do not combine servers having reserved and locked resource with the servers having non-reserved and locked resource.

Storage

The storage technology that is used can be any of the usual choices—local, NAS, or SAN. But, care should be taken not to overburden storage with too many VMs. Veritas eDiscovery Platform is very IO-intensive and performance suffers if the storage system cannot meet the demand, especially during processing. Local storage on a fast RAID array can be as fast as an FC SAN. But, to use the advanced fault tolerance features of ESXi, you need to use SAN or NAS. We recommend using SAN storage, but NAS can also be used if it satisfies the IOPS requirements.

Memory

Sufficient memory should be dedicated to ESX hosts apart from all the VMs running on it. Requirements of CPU and Memory for eDiscovery Platform are mentioned in Table 1. For better sizing of ESX hosts for hosting such VMs, refer to ESX documentation.

Networking

Ideally, the machine should have at least two physical NICs. One should be dedicated to the service console and ESXi infrastructure. If possible, dedicate a NIC to each VM that needs to access different physical machines.

For best performance, all the eDiscovery Platform nodes in a cluster should use the same virtual switch.

Configuring virtual machines

Virtual CPUs

Important things to keep in mind when configuring vCPU allocations for a VM:

- If you allocate vCPUs to a virtual machine then that VM can receive a time slice if there are idle CPU cores available on the host machine at that time. For example, if you have a host machine with 24 cores, and you create a set of VMs with 8 vCPUs allocated, then at most 3 VMs can run simultaneously. The VM receives a time slice if the hypervisor can dedicate eight cores. This point is true even if the VMs are not using all those cores. And therefore, over-allocating vCPUs negatively affects performance of the VM by reducing the likelihood that the VM receives a time slice.
- In practice, the hypervisor requires significant processing time to maintain the virtualization environment, in particular, to perform IO

virtualization tasks. Thus, in the previous scenario, often only two VMs can run simultaneously.

- Avoid allocating more cores to VMs on a single host than physically available on the host if those VMs will be processing simultaneously. The best performance is realized when CPU resources are dedicated to the eDiscovery Platform VM, but this limits flexibility of the system.

  It does not matter if you configure a single CPU with multiple cores, or multiple CPUs with fewer cores. There may be some theoretical performance difference, but it is unlikely there will be any in practice. The configuration option is for running the software that is licensed per-CPU (so you can configure a single CPU, but with four or six cores or more).

Memory

When allocating memory to your virtual machine, it is possible to allocate more memory collectively than what physically exists in the VM server. It is called memory overcommitment. If the VMs do not use all of their allocated memory at the same time, it can work well. However, if multiple VMs simultaneously do need to use all their allocated RAM, performance is poor. Because, data in RAM is forced out into the Windows page file (if using the VMware balloon driver) or into the ESXi swap file. eDiscovery Platform is tuned to use as much memory as is available. Therefore, it is important that the eDiscovery Platform VMs have its full memory allotment available. For best performance, eDiscovery Platform VMs should have all of its memory reserved and locked. This is particularly important if multiple eDiscovery Platform VMs on a given host are processing concurrently, otherwise disk thrashing occurs.

If memory is over-committed, the VMware swap file and Windows page file should be located on a different storage device from the VM virtual disks to minimize the performance effect.

Additionally, if any Review related workflows (Caching, Production Folder Lock, Production Export, Native Review) are being performed, then the memory should be 4x the number of CPU cores.

Disk

The drives should be thick-provisioned to avoid fragmentation. Eager zeroing the drive does not have a large effect, but you might as well do that too.

To improve locality of reference, swap files should be located on separate devices – both the ESXi swap file and the windows page file.

Network

For best performance, all of the eDiscovery Platform nodes in a cluster should use the same virtual switch.

Disable the unnecessary hardware

Disable the unnecessary hardware like USB controllers, serial ports, parallel ports, CD-ROM drive, and sound cards. Disabling eliminates the need for the hypervisor to emulate unused devices when the operating system polls them for status.

Snapshots

Snapshots have a negative effect on performance. Avoid keeping them longer than necessary.

MySQL database

The database host must be within the same hypervisor and virtual switch. A physical server is recommended for database, but it can be a virtual machine of equivalent performance if required by the local IT policies. Refer to Table 1 for hardware requirements of database server.

Alternatively, you can choose the traditional configuration which uses standalone servers with a shared resource pool of utility nodes. Both options use a resource pool, and have a primary appliance designated as the Resource Manager.

Chapter 4

Deployment for Performance Measurements

The following distributed architecture deployment is used for measuring the performance of individual eDiscovery features:

Role	CPU	Memory	OS	SAN Storage
Master	32 cores	128 GB	Windows Server 2016 Standard	1 TB
Worker Node (Pre-Processing, Analysis, Review, and Export Server)	32 cores	128 GB	Windows Server 2016 Standard	2.5 TB
Utility Node (Imaging, retrieval, Caching, Export)	8 cores	32 GB	Windows Server 2016 Standard	1 TB
Shared DB	32 cores	128 GB	Windows Server 2016 Standard	1 TB

Note: For performance measurements, eDiscovery Platform was run on virtual machines with dedicated (reserved) CPU and memory resources. Better throughput and response time may be achieved on physical servers.

Chapter 5

Processing

Using Veritas eDiscovery Platform, corporations, government agencies, and law firms perform early case assessments and rapidly cull down data, thereby reducing overall electronic discovery costs. As an integrated part of the eDiscovery Platform, the Processing and Analysis Module supports the iterative workflows required during real-world electronic discovery.

Performance Numbers

The following table shows the 'Processing Rate' for scenarios tested:

Document Type	No. Of Documents	Avg. Document Size	Deduplication %	Hourly Processing Rate (Documents)	Hourly Processing Data Rate (GB)
EML/MSG	20,000	70.5 KB	18% (3%)	59,454	3.86
Word	20,000	159 KB	0% (0%)	54,752	8.21
PDF	20,000	327.7 KB	0% (0%)	45,169	14.22
Excel	20,000	534.9 KB	0% (0%)	42,105	21.47
PPT	20,000	634.8 KB	0% (0%)	22,346	13.51
Mix*	50,000	292.7 KB	7% (0%)	51,575	14.33
Mix*	100,000	268.9 KB	8% (0%)	57,297	14.66
Mix*	1,000,000	449.4 KB	1% (0%)	77,347	32.20
Mix*	5,000,000	211.6 KB	4% (2%)	96,847	19.49
Mix*	10,000,000	204 KB	8% (13%)	96,651	18.77

* Mix = 40% of EML & MSG, 15% each of Word, PDF, Excel, and PPT.

The throughput of the processing job depends on:

- The Document Type
- The Document Size
- Underlying Hardware
- Deduplication %
- Number of Attachments

Chapter 6

Caching

Caching enables a reviewer or case manager to convert a set of documents to their native format preemptively through (or by) a background process. The set of documents can be either a search result or the content of a folder. Use of caching has a dramatic effect on the review process by eliminating the real-time conversion that the reviewer sees. Instead, the response (or display) to the reviewer is immediate. Caching also helps increase the speed of locking a Production folder.

Performance Numbers

The following table shows the 'Caching' rate for scenarios tested:

Item Type	No. Of Items	Avg. Item Data Size	Hourly Caching Rate (Items)	Hourly Caching data rate (GB)	Hourly No. Of Pages Cached
Mix*	99,629	268.9 KB	11,543	2.95 GB	127,892

* Mix = 40% of EML & MSG, 15% each of Word, PDF, Excel, and PPT.

The throughput of the caching job depends on:

- Item Type
- Item Size

Chapter 7

Lock and Produce

The Veritas Review and Production Module accelerates the privilege and response review processes for legal professionals in corporations, government agencies, and law firms.

Performance Numbers

The following table shows the 'Lock & Produce' rate for scenarios tested:

Item Type	Hourly Lock & Produce Rate (Items)	Hourly Lock & Produce Data Rate (GB)	Hourly No. Of Pages Produced
Cached (Mix*)	97,346	18.84 GB	1,168,517
Non-Cached (Mix*)	8,589	1.66 GB	97,566

* Mix = 40% of EML & MSG, 15% each of Word, PDF, Excel, and PPT.

The throughput of 'Lock & Produce' job depends on whether the items are already cached or not.

Chapter 8

Export

Case administrators can produce documents in native, image (TIFF or PDF), and mixed mode formats (for example, Excel spreadsheets in native, but all others as images). Flexible export options allow for multiple metadata formats (EDRM XML, DAT, CSV, and so on), “reduplicating” of data by custodian, and preservation of original folder hierarchies.

Performance Numbers

The following table shows the ‘Export’ rate for scenarios tested:

Export Type	Export Format**	No. Of Items	Avg. Item Size (KB)	No. of Items Exported Per Hour	Source Data Exported Per Hour (GB)	No. Of Pages Exported Per Hour
Metadata (Mix*)	N/A	71,793	292.7	71,554	13.85	N/A
Production (Mix*)	PDF	71,793	292.7	28,549	5.52	338,514
Production (Mix*)	TIFF, DPI = 100	71,793	292.7	16,706	3.23	198,097
Production (Mix*)	TIFF, DPI = 300	71,793	292.7	11,540	2.23	136,835

* Mix = 40% of EML & MSG, 15% each of Word, PDF, Excel, and PPT. 10% of entire corpus is redacted.

** Export Options Selected for the test are as follows:

Export Type	Section/Field	Value
Metadata	Template	Extended CSV UTF
	Metadata Fields	DocID, ParentID, DocType, MimeType, Subject, From, To, CC
	Format	Default Values as it is: Field Delimiter: ,(44), Text Qualifier: '(34), Encoding (UTF-8), File Extension: CSV, Date & Time: By default
	Tags	Column Format: All tags in one column Include full hierarchies with tag names: Checked Include Tag Event Comments: Checked
	Folders	Column Format— All folders in one column Include full paths with folder names - Checked
	Options	Include batching information – Checked Include item notes – Checked Include complete history – Checked Include discussion thread – Checked Include file flags – Checked Include custodian & location info – Checked One row per custodian – Checked Include: Checked – (Messages in native format & native loose files) Include Journal Information – Checked (In Original Format) Include extracted text – Checked –Break out embeddings as separate items – Checked **If any of the options are not listed here, they were Unchecked.**
	Output	Access from appliance only – Selected (With local disk path)
Production	Template	Default
	Production Options	Production: Final, File Naming: Production Number,

		Include: Notes as HTML, all other pure native, Include Journal Information – Checked (In Original Journal Format), Production folder type - MIXED Include images - Image Format – PDF/TIFF (depending on test) - DPI – 100/300 (depending on test) - Color Style - Black & White - Color Depth - 24 - Each File – Single Page - Page Size – Letter (11 x 8.5 in) - Include Opticon image loadfile – Checked Include Markup Types - Redactions - Checked Include extracted text—Checked – If no extracted text includes an empty text file - Checked – Include file metadata in extracted text – Checked
	Metadata Fields	CC, DateModified, DateSent
	Format	Field Delimiter: (174), Text Qualifier: (254), Encoding (UTF-8), Date & Time: By default, File Extension: dat
	Tags	Column Format: One tag per column Include full hierarchies with tag names: Checked
	Folders	Column Format— One folder per column Include full paths with folder names-Checked
	Options	Include Item notes – Checked Include Complete History – Checked Include discussion thread – Checked Include custodian & location info - Checked **If any of the options are not listed here, they were Unchecked.**
	Output	Access from appliance only – Selected (With local disk path)

The item export rate that an eDiscovery server can deliver depends on several factors like:

- Item Type
- Item Size
- Export Disk Location Performance
- Export Format

The size of export data depends on:

- Export Format
  (Export in TIFF would have higher size than export to PDF)
- DPI Value
  (Higher the value, higher is the size of exported content)

Chapter 9

Native Review

Multiple reviewers accessing the same document can reduce the time it takes to review and make coding decisions. The new native viewer of eDiscovery Platform increases reviewer productivity by viewing documents in a near-native format without requiring each application to be loaded on a reviewer's workstation.

Performance Numbers

Average time to load an item to render an item in redaction view while you navigate from item to item:

Following matrix time includes the time that is required to save the markups that the users apply on an item. The matrix time also includes the time that is required to load the next item in the redaction view with caching done for the items being opened.

Response Time (in Seconds)		
Single User	**25 Concurrent Users**	**50 Concurrent Users**
1.11	2.53	3.68