edt. | Discovery Simplified.

EDT 7

System Requirements & Deployment Options

Contents

Executive Summary

EDT is a modular, server-based application with central data repositories that operates on a Microsoft technology stack. Power users process source data using desktop applications that store document text and metadata within Microsoft SQL Server databases. Standard users access and interact with processed data in EDT via a web interface that runs on Microsoft IIS 8. Back-end tasks are performed by Windows Services running on one or more application servers. The components of an EDT deployment include:

- Data Processing Applications – data ingestion and exception management via desktop applications; document OCR and export via windows services;
- Near Native Rendering & Markup – near native rendering is conducted via Accusoft PrizmDoc Server (a third-party application included with EDT)
- Database – storage of document metadata and user work product in Microsoft SQL Server
- Website – delivery of the user interface via Microsoft IIS configured for HTTPS using TLS 1.2
- File Services – storage of original source data, EDT working files and exported data on a network file server, NAS or SAN via CIFS

Due to its modular nature, EDT may be deployed in a range of configurations. The manner of deployment will influence performance, availability and recoverability.

Deployment Options

The manner in which EDT is deployed within your organisation will depend on a variety of factors, such as the number of reviewers, volume of data to be processed, and the predicted document review rate. EDT supports the following deployment options:

- Portable – all components and modules deployed on a high-end laptop, workstation or portable server. This option is typically used when it is necessary for data to be processed and reviewed in situ.
- Single Server – all components and modules deployed on a single physical or virtual server installed in the organisation's datacentre or cloud provider. This option is typically selected for low data throughput and/or review scenarios.
- Distributed – components and modules deployed across two or more physical and/or virtual servers in the organisation's datacentre or cloud provider. This option is commonly used when the data throughput and/or review requirements cannot be supported by a single server, or functionality is to be segregated for security purposes.

There are operating requirements and security considerations that are factors that will determine which deployment option is the most appropriate for each environment.

Operating Requirements

The following factors are major influencers in the resources required to operate EDT successfully:

INGESTION THROUGHPUT	The rate at which the organisation needs to ingest data. This is a CPU and Disk IO intensive task that will impact any other process running on the same server. While ingestion rates will vary depending on the data being processed, throughput of between 150 and 250 GB per day per dedicated server can be expected.
CONCURRENT USERS	Each user accessing the system places CPU and RAM resource load on the Web and Database server as they navigate through screens and perform queries on the processed data.
REVIEW RATE	Viewing documents in the Near-Native and Markup views requires that the system retrieve and render them into HTML5 / SVG in real-time. To provide a consistent, responsive experience for end users it is imperative to understand peak document review rates across the environment and deploy one or more PrizmDoc servers with enough resources to meet peak demand.
OCR THROUGHPUT	Recognition of text from scanned documents and/or images is a very CPU intensive task. The OCR task will consume all available CPU resources and will directly impact the performance of any other process running on the server. The speed and quantity of CPUs available to the OCR task will directly impact the number of pages processed per minute.
EXPORT THROUGHPUT	During the operation of an Export task, EDT will extract original natives from their source, required metadata from the database, perform any required rendering operations (such as generating PDFs, TIFFs or JPEGs), redaction operations (using PrizmDoc) and output the results to disk.

Software Requirements

CLIENT COMPUTERS

End user computers accessing the Analyst or Reviewer require Microsoft Internet Explorer 11 (with JavaScript enabled), Microsoft Edge or Google Chrome browser to access and operate EDT.

EDT SERVERS

The table below lists the required software for the server and administrator applications.

Software	Web Application (IIS)	SQL Database Server EDT Server	PrizmDoc Server	Loader	Agent Service	Importer	QA Manager
Windows Server 2012 R2 or newer	■	■	■	■	■	■	■
IIS 7+ (Internet Information Services)	■						
SQL Server 2012, 2014 or 2017		■					
Ghostscript 9.05+					■		
PDF Printer. One of the following: • Adobe® PDF Printer • Bullzip PDF Printer • bioPDF PDF Writer					■		
IBM® Lotus Notes® Client 8.5 (Standalone, Messaging) [1]				□	□		□

[1] The IBM® Lotus Notes® Client is required to load .NSF files.

Mount Image Pro v6 [2]				□	□		□
Microsoft Access Database Engine 2010 or 2016 (64 bit)					■		
Microsoft .NET Framework 4.7.1	■	■	■	■	■	■	■

■ Required □ Optional

[2] Mount Image Pro is required to load Forensic Image files

Hardware Requirements

STANDALONE MACHINES

This information is provided by way of a guideline only as there are many ways the solution can be implemented to service different client environments and requirements. Detailed discussions should take place with an EDT technical consultant prior to infrastructure procurement and implementation to ensure capacity will meet client needs.

Description	Hardware Example
Laptop for small, portable document processing and review cases with a single reviewer	Intel Core i7 32 GB RAM 256 GB Solid State Drive
Server for small document processing, review and export cases with up to 5 reviewers	8 Cores, 128 GB RAM 128 GB SSD – Operating System 2 TB SSD – Source Data 1 TB SSD – SQL Database Files 256 GB SSD – SQL Transaction Log Files 256 GB SSD – EDT CFS & Export Path
Server for small document processing, review and export cases with up to 15 reviewers	16 Cores, 256 GB RAM 128 GB SSD – Operating System 2 TB SSD – Source Data 1 TB SSD – SQL Database Files 256 GB SSD – SQL Transaction Log Files 256 GB SSD – EDT CFS & Export Path

DISTRIBUTED ENVIRONMENTS

This information is provided by way of a guideline only as there are many ways the solution can be implemented to service different client environments and requirements. Detailed discussions should take place with an EDT technical consultant prior to infrastructure procurement and implementation to ensure capacity will meet client needs.

Assumptions	Moderately Sized Matters	Larger Matters	Very Large Matters
Concurrent Reviewers	1 – 20	20 – 50	50 +
Ingestion	100 GB per day	500 GB per day	Up to 1TB per day (Metadata only, filtering on load, and multiple case loading)
Expected Case Size	500 GB (~5,000,000 docs)	1 TB (~10,000,0000 docs)	4 TB (~40,000,000 docs)
Collective Size of Cases	10 TB	20 TB	100 TB
Native Documents Reviewed per Minute	10	30	60
Production	20,000 docs per day	100,000 docs per day	250,000 docs per day

Hardware	Moderately Sized Matters	Larger Matters	Very Large Matters
Agent, Loader & Importer server(s)	4 Cores 32 GB RAM	2 x Servers 8 Cores 32 GB RAM	5 x Servers 8 Cores 32 GB of RAM
Web server(s)	4 Cores 8 GB RAM	2 x servers – load balanced 4 Cores 16 GB RAM	4 x servers – load balanced 8 Cores 16 GB RAM
PCC Server	8 Cores 32 GB RAM	2 x Servers 16 Cores 64 GB RAM	4 x Servers 16 Cores 64 GB RAM

SQL Server(s)	4 Cores 64 GB RAM	8 Cores 256 GB RAM	16 Cores 1 TB RAM
Storage	16+ TB 10 TB Source 4 TB SQL DB 1 TB SQL TX 1 TB EDT CFS / Export	30+ TB 20 TB Source 7 TB SQL DB 2 TB SQL TX 1 TB EDT CFS / Export	150+ TB 100 TB Source 40 TB SQL DB 5 TB SQL TX 5 TB EDT CFS / Export

General hardware recommendations:

- Obtain the fastest CPUs available within your budget.
- Some case-level database processes are CPU intensive. Additional memory on the SQL Server will improve the execution time of case-level operations. More memory allows SQL Server to cache more database content, thereby increasing performance.
- Follow other SQL Server best practices. For example, distribute database-related files/logs/etc across dedicated LUNs/spindles/RAID arrays.
- General virtualisation recommendations:
- Dedicate sockets, memory and separate hard disks to each virtual machine i.e. do not share or over allocate physical resources

General storage recommendations:

- Use fast drives such as 15K RPM SAS and Solid State Disk (SSD) drives with, where appropriate, RAID configurations to maximum disk I/O performance.
- As a rule of thumb the storage space required for each case is double the original source data size. The storage space should initially be distributed among the Source data, the Common File Store and the database server. Additional storage is required by the web server for file caching and by the Agent for the export destination.
- Segregate Data Repositories by physical hard drive spindles. The data storage repositories include the Source data for each case, Common File Store data, Export destination, and the databases. Segregating the data reduces competition for storage resources.
- Additional storage may be required for external Optical Character Recognition (OCR) applications.

Optical Character Recognition

EDT supports both an internal OCR workflow (via the Tesseract OCR engine) and an external OCR workflow (via the use of third-party products). Users can select which workflow to use during data processing:

- Internal OCR – the internal workflow utilizes the Tesseract OCR engine founded by Hewlett Packard in 1985 and development sponsored by Google since 2006. There are no additional fees for the use of this workflow. Documents requiring OCR may be processed using a selected Agent. Extracted text is added to the database and is available for searching and/or export. Text searchable PDF documents are not produced for documents processed using this workflow.
- External OCR – users may utilize the OCR tool of their preference using the external workflow. Documents requiring OCR are exported from EDT, processed for OCR by the user, and then ingested back into EDT as text files or text searchable PDF documents. We recommend the purchase of ABBYY Recognition Server (running on a dedicated workstation on the same network as EDT) for the external workflow.

Optical Character Recognition is a CPU intensive task and should be conducted either outside of normal operation hours, or using application servers dedicated to the task.

Security Recommendations

The deployment type, user base and confidentiality of data processed into EDT will all impact the level of hardening and other security measures you may wish to implement. EDT recommends each customer perform their own risk assessment in line with internal IT Security policy. At a minimum, EDT recommends that customers:

- Use HTTPS with TLS 1.2 – it is good practice to encrypt all web traffic between the browser and the web server by permitting only HTTPS communications. Deploy an SSL certificate on your web server and enforce HTTPS connections to your website using TLS 1.2.
- Block all unnecessary traffic – to use EDT standard users only need to be able to connect to the EDT website. Power Users who perform data processing will need to be able to upload data and connect to the desktop of the application server(s).
- Segregate Components with a DMZ – if external access to EDT is required, we recommend implementing a distributed environment with the web server in a DMZ. Inbound port 443 is required from the internet to the web server. TCP ports 1433 (SQL default instance) and 18681 (PrizmDoc Server) will need to be open between the web server and the database and PrizmDoc servers respectively. It may also be desirable to segregate the entire EDT environment from the organisations internal network via a secondary DMZ.
- Patch Operating Systems and Applications – patching should be applied in a regular and timely fashion to address known vulnerabilities.
- Protect Against Anti-Virus & Malware – implement a respected AV & malware solution

- Enforce Password Complexity – either integrate with Active Directory or use EDT's inbuilt controls to enforce password complexity requirements in line with current industry recommendations
- Consider Two Factor Authentication – EDT includes the option to enforce two factor authentication for users accessing the web interface. Where external access is being provided, EDT recommends enforcing this option for all users.
- Rename or Disable Administrator Accounts – the inbuilt "Administrator" account on all servers should be renamed or disabled
- Monitor Service Accounts – the modules of EDT that operate as windows services require elevated privileges and will need to authenticate using an account that has local administrator access to the application servers and the various data repositories in the environment. Enable monitoring and alerting for this account.
- Penetration Test – before confidential or sensitive data is transferred into the environment, EDT recommends performing an external penetration test to validate that the external facing components and modules have been correctly configured and hardened.
- Vulnerability Assessment – it is recommended that a vulnerability assessment is performed within the EDT network(s), including a port, operating system and application scan to identify any known or potential vulnerabilities in your environment.

Business Continuity & Disaster Recovery

Each organisation has its own Recovery Point and Recovery Time Objectives ('RPO' and 'RTO') based on the importance of the system to the organisation and the impacts of loss of data or availability. An EDT environment contains several components and data repositories that are required for basic operation. Availability of some modules may not impact immediate access and therefore not require the same continuity mitigation measures. When conducting a BC&DR risk assessment exercise, customers should consider:

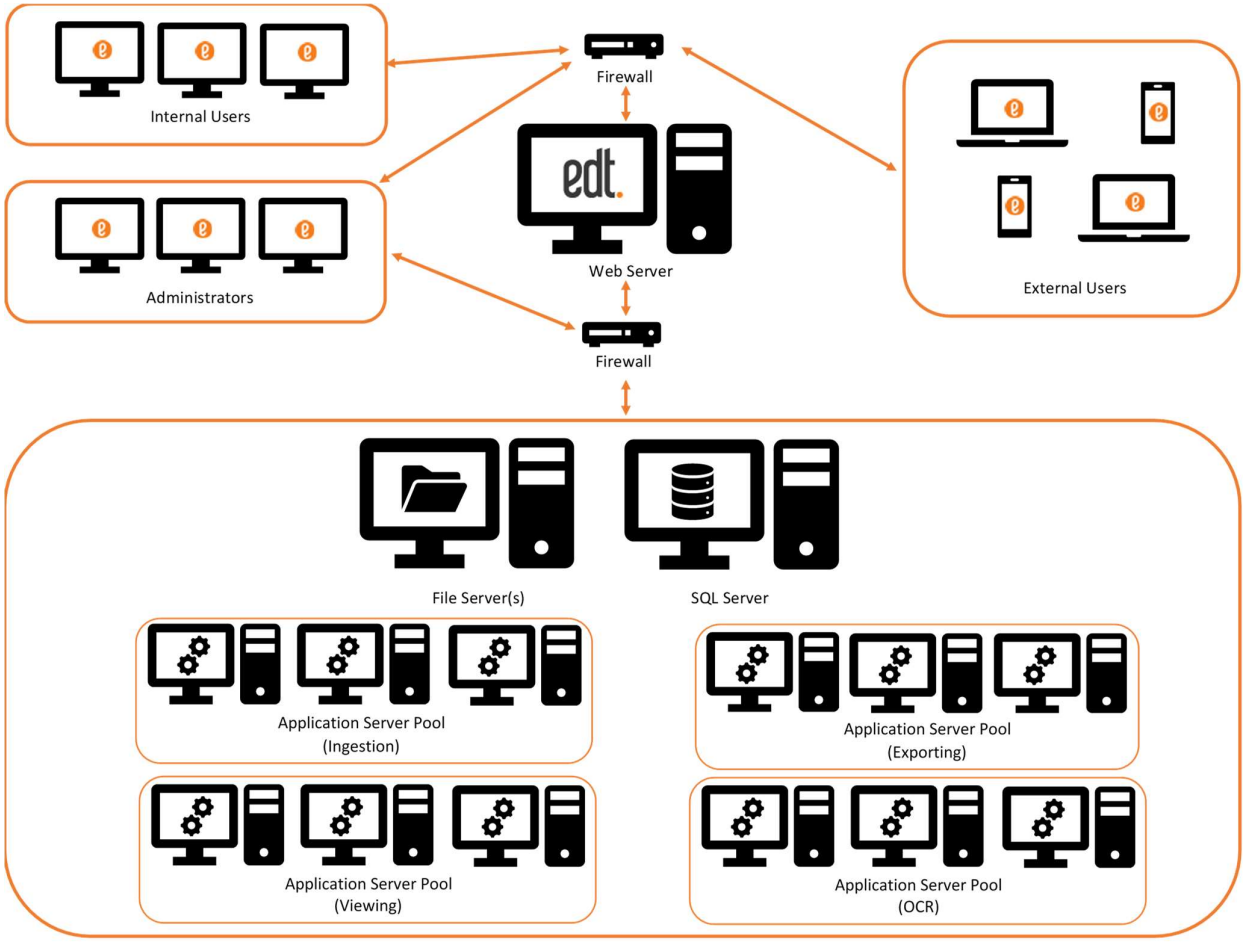
SQL DATA	The metadata from all processed data, along with user work product (document coding, tagging, redactions, etc.) is all stored in Microsoft SQL Databases. Two central databases are used that contain the information the operation of the entire environment, and a separate database is created for each Case. The availability of this information via Microsoft SQL Server is essential for ongoing operation of EDT.
SOURCE DATA	When data is processed into EDT only the text and metadata is stored in the SQL database. Operations such as viewing, OCR, export, etc. require access to the original document in the location that it was processed.
STORE DATA	Some processing functions, such as the OCR and QA workflows, cause EDT to store different versions and formats of documents in a central location known as the "Store". Correct function of EDT will be hindered if data that is expected to be located in the Store is not available.
APPLICATION MODULES	The Importer, Loader and QA Manager applications are used to ingest and process data in a batch method initiated by Power Users. The EDT Agent module performs background tasks initiated by users interacting with the web interface, and the EDT Server module performs database creation and modification tasks. The EDT Web module resides on the web server and provides the interface to EDT for standard users. Loss of availability of these modules will have different impacts on the application and each organisation's tolerance of these will vary.

Distributed Architecture

EDT is most commonly deployed in a distributed fashion, with one or more dedicated servers for each role within the environment.

WEB SERVER	Consisting of Microsoft Windows 2012 R2 or higher, with the Web Server role configured, and both the EDT Web module and Prizm Application Services installed. The EDT Web module contains the website content and applications that interact with Prizm Application Services and Microsoft SQL Server as required. Prizm Application Services interacts with the PrizmDoc Server as required.
DATABASE SERVER	Containing central configuration databases and a database for each Case, Microsoft SQL Server 2-012 or higher running on Microsoft Windows 2012 R2 or higher is at the heart of an EDT deployment. Each of the EDT modules will interact with the database server to authenticate users and read or write data. We recommend that the EDT Server module is installed on the database server.
APPLICATION SERVER	The EDT Importer, Loader, QA Manager desktop applications and the EDT Agent Windows service are installed on one or more Microsoft Windows 2012 R2 servers. These perform ingestion of structured loadfiles, ingestion of native data, processing of native data, and background processing tasks (document retrieval, rendering, OCR) respectively.
PRIZMDOC SERVER	This third-party server product running on Microsoft Windows 2012 R2 is used by EDT to provide near-native document viewing, annotation and rendering within the application. As requested by Prizm Application Services (on the web server) or the EDT Agent (on the application server), PrizmDoc will convert native documents to HTML5 compatible versions for display in the web interface and/or render redacted documents to PDF for export.
FILE SERVER	Whether a Microsoft Windows 2012 R2 or higher server, or a device such as a NAS or SAN that can provide a UNC addressable resource, the File Server contains the original source data, the EDT Store, and any data exported from the system.

Distributed Architecture Diagram



High Availability Architecture

The technologies that underpin EDT may be configured in a highly available fashion, providing for almost seamless failover and recovery from a variety of situations.

LOAD BALANCED WEB SERVERS	Utilising multiple web servers behind a session-state aware load balancer can provide seamless failover to cater for loss of a web server.
LOAD BALANCED PRIZMDOC SERVERS	Accusoft support having two or more PrizmDoc servers behind a round-robin style load balancer. EDT is simply configured with the address of the load balancer, which passes the request to the next in queue PrizmDoc server. Retrieval of cached documents from the appropriate PrizmDoc server is handled internally by PrizmDoc.
ALWAYS-ON HIGH AVAILABILITY CLUSTERED SQL SERVERS	When installed in a Windows Failover Cluster, Microsoft SQL Server can be configured to use SQL Listeners and Synchronous, Always-On High Availability Groups without sharing underlying storage. While this configuration may have a significant cost impact on the setup and operation of your EDT environment, it provides almost instant failover in the case of SQL server loss.
WINDOWS DISTRIBUTED FILE SYSTEM	Microsoft Windows Server includes a feature known as “Distributed File System”. This automatically synchronises data across two or more file servers and offers a high level of redundancy.
FILE SERVER	Whether a Microsoft Windows 2012 R2 or higher server, or a device such as a NAS or SAN that can provide a UNC addressable resource, the File Server contains the original source data, the EDT Store, and any data exported from the system.

An EDT environment that is configured across two data centres (or AWS Availability Zones) and is deployed with the above recommendations will continue to be available with minimal downtime even in the event of a total datacentre loss. Organisations should consider their RPO and RTO requirements before investigating these advanced configuration options.

High Availability Architecture Diagram