

# Performance Benchmark

---

*White Paper*  
*Equivio v2.3.7*



**equivio**  
zoom in. find out



# MANAGEMENT SUMMARY

## INTRODUCTION

This document describes the results of performance benchmark tests performed on the Equivio Version 2.3.7 software in May 2009. The tests were carried out in Equivio's R&D laboratories. The tests were conducted using standard hardware and software equipment to ensure replicability in customer production environments.

## TEST OVERVIEW

The benchmark tests were conducted to demonstrate the performance of Equivio v2.3.7 in detecting and grouping near-duplicate documents and email threads in a variety of common operational scenarios. Each test was run on one or more cases, each of which comprised a collection of documents from an actual enterprise e-discovery scenario. The scenarios were designed to examine the individual effect on performance of each of the following parameters:

- Database Types: Microsoft SQL Server, MySQL and Oracle
- Data configuration: local versus remote storage of input data
- Threads: number of processing threads on single machine
- EquiLevel: threshold of similarity for determining a near-duplicate

## SUMMARY OF FINDINGS

The benchmark tests clearly demonstrate the ability of Equivio v2.3.7 to process large sets of real-world data within a matter of hours. The test results indicate that Equivio can handle even the largest cases effectively. For example, **one million text documents can be processed by Equivio on one machine in less than 3 hours. A case of seven million documents can be processed in 30 hours.**

Following are the main conclusions drawn from the test results:

- Recommended configuration is based on one machine for database and one machine for processing
- Equivio scales efficiently to handle large cases with millions of documents
- The database type (SQL Server, MySQL, Oracle) has a minimal effect on performance
- Storing the input data locally improves performance by approximately 30%
- Due to the relatively minor impact of the EquiLevel on performance, it is recommended to set the EquiLevel based on business needs and the type of data being processed, rather than performance considerations
- Speed of data transfer from the disk/network is a key prerequisite for optimal performance





# TEST SCENARIOS

## EQUIVIO APPLICATION OVERVIEW

Equivio offers patent-pending software to detect and group near-duplicate documents and emails. The Equivio product is used to expedite the management and review of unstructured document repositories. The grouping of near-duplicates and email threads allows similar documents to be handled and treated together. The result: Equivio users significantly reduce the time and cost of document review, while ensuring the consistent treatment of similar documents.

Equivio supports a broad range of business applications. Equivio is offered as a specialized component for integration within third-party applications for e-discovery, internal investigations, data retention, email archiving and intelligence.

Equivio>NearDuplicates organizes near-duplicate documents into sets, allowing each set to be assigned to a reviewer for efficient and coherent handling. For each set, Equivio>NearDuplicates determines the pivot document, which is the most representative document of the set. In the review process, the user reads the pivot document, then reviews its near-duplicates by invoking a compare tool which highlights the differences in each document vis-à-vis the pivot document. This dramatically reduces the time required to review documents.

Equivio>EmailThreads captures and structures email threads, cutting the number of emails that need to be read in a review process. Beginning with an unstructured collection of emails, Equivio>EmailThreads groups emails belonging to a thread, and builds the thread hierarchy based on the original email and subsequent “events”, such as reply and forward. Equivio>EmailThreads analyzes the email content and verifies that the last email in the thread contains all preceding emails, allowing users to focus review efforts on this “inclusive” email. The use of the email content eliminates any dependence on metadata, which is often corrupt and cannot guarantee that the last email in a thread contains the previous layers of the thread.

## BENCHMARK CASES

In order to provide a test environment that reflects a real world implementation of Equivio, the benchmark tests used data from two different cases, both of which are available in the public domain. These cases, which are described below, vary in terms of volume and file types, covering a wide spectrum of operational scenarios. Each of the tests was run on one or more cases.



## TREC

The Text Retrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, supports research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. More information about TREC can be found [here](#).

The case used for the Equivio benchmark consists of documents that were made public during various legal cases involving US tobacco companies and contain a wide variety of document genres typical of large enterprise environments.

This case comprises 6.9 million text files, including both documents and emails.

## ENRON

This case consists of information made public as part of the Enron trial in 2004. It contains data from about 150 users, mostly senior management of Enron, organized into folders. More information about the Enron case can be found [here](#).

The dataset contains a total of about 517,000 email messages in [RFC2822](#) format.

# TEST SCENARIOS AND RESULTS

In line with standard e-discovery practice, the test scenarios assume that the input to Equivio is extracted text files, provided by the client.

## SCENARIO 1 - DATABASE TYPE

### GOAL

*To measure the effect of different databases on performance.*

The databases tested were:

- MSSQL Microsoft SQL Server 2005 - 9.00.3042.00 (X64) on Windows NT 5.2 (Build 3790: Service Pack 2)
- MySQL 5.1.33-community(InnoDB)
- ORACLE NLSRTL (10.2.0.1.0,Production) Oracle Database 10g Enterprise Edition (10.2.0.1.0,64bi) PL/SQL (10.2.0.1.0,Production) TNS for 64-bit Windows: (10.2.0.1.0,Production)



Once Equivio has processed and calculated the near-duplicates and email threads, an extract utility creates a load file for the target review system database. Database performance was measured for near-duplicate processing, as well as for the extract process.

Scalability was also examined in this test scenario by measuring the effect of increasing document volumes on performance over time.

### TEST DATA AND SETUP

The setup for this test scenario is presented in the following table:

Configuration	# of Threads	Application	EquiLevel	Data Location
1 machine for processing and 1 machine for database	8	Near-Duplicates	60	Remote*

\*Data stored remotely and accessed over the network.

Details regarding the hardware configuration can be found in Appendix A.

This test scenario was run on the TREC case.

### TEST RESULTS

The test results are presented below. The time units relate to days, hours, minutes and seconds.

*Table 1 Case: TREC – 6,910,192 documents*

Database	Processing Time	Extract Time	% of NDs
SQL Server	1:06:50:05	00:08:15:00	33
MySQL	1:02:49:28	00:09:30:00	33
Oracle	1:09:09:17	00:08:30:00	33

System scalability is illustrated in Figure 1, which shows the total number of files processed over time for each database type.

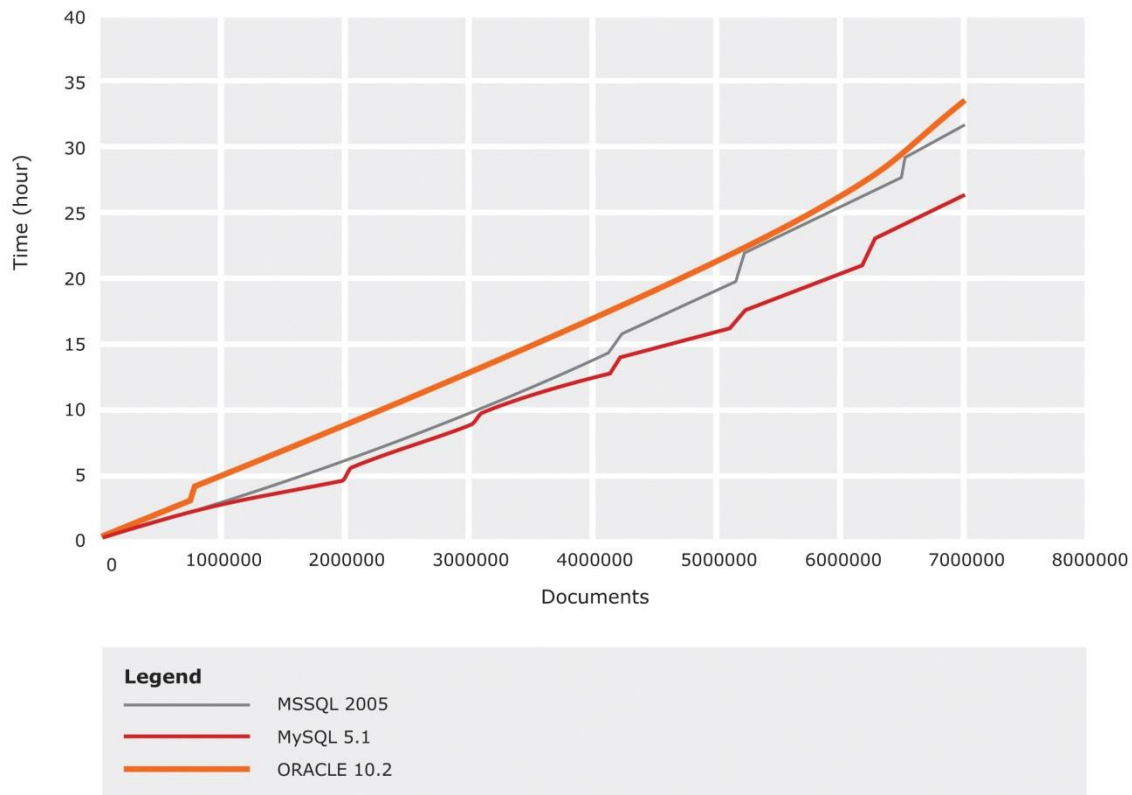


Figure :1 Total files processed over time per database

**Conclusion:** Equivio can efficiently process very large cases, such as TREC, on any of the three databases tested. The type of database has a minimal impact on the performance and scalability of the Equivio application. It should be noted that regardless of the database being used, **Equivio processed one million text documents – which is representative of the vast majority of e-discovery cases - in three hours or less.**

This test clearly shows that Equivio is an efficient and scalable solution for processing large-scale document sets.

## SCENARIO 2: LOCAL VS. REMOTE DATA

### GOAL

*To measure the effect of data access on performance.*

This test examines the impact of the proximity of the data collection to the processing module. In the first run, the test data was stored locally, on the processing machine, while in the second run the data was stored remotely.

## TEST DATA AND SETUP

The setup for this test scenario is presented in the following table:

Configuration	Database	# of Threads	Application	EquiLevel
1 machine for processing and 1 machine for database	SQL Server	8	Near- Duplicates	60

Details regarding the hardware configuration can be found in Appendix A.

This test scenario was run on the TREC case.

## TEST RESULTS

The test results are presented below. The time units relate to days, hours, minutes and seconds. The speed is measured in thousands of files per hour (K f/h).

*Table 2Case: TREC – 6,910,192 documents*

Local/Remote	Time	Speed (K f/h)	% of NDs
Local	0:21:52:14	314	33
Remote	1:06:50:05	222	33

**Conclusion:** I/O speed has a significant impact on performance. Storing the input data locally improves performance by approximately 30%. For best performance, it is important to ensure optimal access to the data.

## SCENARIO 3: PROCESSING THREADS

### GOAL

*To measure the effect of the number of processing threads on performance.*

The tests were run using 1, 4, 8 and 16 processing threads.

## TEST DATA AND SETUP

This test scenario was run on the TREC and Enron cases.

The setup for this test scenario is presented in the following table:



Case	Database	Application	EquiLevel	Data Location
TREC	SQL Server	Near Duplicates	60	Remote
Enron	SQL Server	Near Duplicates and Email threads	75	Remote

Details regarding the hardware configuration can be found in Appendix A.

### TEST RESULTS

The test results are presented below. The time units relate to days, hours, minutes and seconds.

*Table 3Case: TREC - 6,910,192 documents*

# of threads	Time	Speed (K f/h)
4	1:15:41:20	174
8	1:06:50:05	222

*Table 4Case: Enron - 517,431 documents*

# of threads	Time	Speed (K f/h)
1	0:03:39:41	141
4	0:01:48:35	287
8	0:01:22:24	378
16	0:01:15:13	413

**Conclusion:** Increasing the number of threads has a significant effect on performance, at no additional cost. The benchmark shows that using four processing threads rather than one (ENRON) halves the processing time. Increasing the number of threads from four to eight (TREC) reduced processing time by approximately 23%.

The recommended number of threads depends on the processing power of your machine. Older machines/laptops should use single thread processing. Generally speaking, four threads is a good starting point. Feel free to consult with Equivio regarding multi-threading for servers or other technical/configuration issues.

## SCENARIO 4: EQUILEVELS

### GOAL

*To measure the effect of EquiLevel settings on performance.*

The EquiLevel is the minimum percentage resemblance between two documents for them to be considered near-duplicates. Any two documents exceeding this threshold are considered near-duplicates.

### TEST DATA AND SETUP

The setup for this test scenario is presented in the following table:

Configuration	Database	# of Threads	Application	Data Location
1 machine for processing and 1 machine for database	SQL Server	8	Near-Duplicates	Remote

Details regarding the hardware configuration can be found in Appendix A.

This test scenario was run on the TREC case.


### TEST RESULTS

The test results are presented below. The time units relate to days, hours, minutes and seconds.

*Table 5Case: TREC – 6,910,192 documents*

EquiLevel	Time	Speed (K f/h)	% of NDs
50	1:09:37:25	206	40
60	1:06:50:05	222	33
75	1:00:40:39	280	21

**Conclusion:** Reducing the EquiLevel yields a greater number of near-duplicates with a relatively minor impact on performance. Since the performance differences between EquiLevels are not dramatic, it is recommended to determine the EquiLevel based on business needs and the type of data being processed, rather than performance considerations. This will enable your organization to derive maximum value from the Equivio application. For example, a lower EquiLevel might be used for OCR data, which



typically contains a significant number of OCR errors which reduce similarity percentages.

## CONCLUSION

The benchmark tests conducted on the Equivio v2.3.7 software examined how variances in several key parameters affect the performance of the application. The tests were run using two different sets of real-world data and were designed to reflect common operational scenarios.

**The results of these tests clearly demonstrate the ability of the application to process cases containing up to one million documents in a 3-hour window. The tests also demonstrated Equivio's ability to scale efficiently to process large-scale multi-million document cases.** In terms of performance, the optimal configuration is to use separate machines (1+1) for database and for processing. For very large cases, in the range of 10 or more million documents, and under the assumption that the file system is not a bottleneck, the use of more than one processing machine can increase throughput.

Increasing the number of threads significantly enhances system performance, at no cost. Storing the input data locally on the processing machine accelerates I/O speed, resulting in better performance.

Standard PCs with 2 GB of memory can be used for efficient Equivio processing. In order to eliminate potential processing "bottlenecks," it is recommended to use a robust machine with 8 GB memory (RAM) to host the database.



## APPENDIX A - PLATFORM DESCRIPTION

This appendix provides details regarding the hardware configurations used in the performance benchmark tests.

The Equivio v2.3.7 application runs on standard hardware and software platforms commonly deployed in enterprise environments.

### CORE PROCESSING SERVER – LOCAL DATA CONFIGURATION

CPU	2 X Intel Xeon Quad Core E5430 (2.6GHz)
Front side bus	1333MHz
Board	Intel Chipset EM64T (64-bit Extension)
Memory (RAM)	PC-2 5300 ECC DDR2 RAM 8GB
Controller	ServeRAID-8k , 256MB , Raid 0/1+0/5
Hard Disk	2 X 146.4GB 15k SAS 3.5" 4 X 300GB 15k SAS 3.5"
Operating System	Windows 2008 (x64)
Network (1 Gbps)	Realtek RTL8169/8110 Family Gigabit Ethernet NIC.

### CORE PROCESSING SERVER – REMOTE DATA CONFIGURATION

CPU	AMD Athlon 64 X2 Dual Core Processor 6000+ 3.0 GHz
Memory (RAM)	2 GB
Hard Disk	WD5000AAKS-22YGA0 465 GB
Operating System	Microsoft Windows XP SP2
Network (1 Gbps)	NVIDIA nForce Networking

### DATABASE SERVER

CPU	AMD Athlon 64 X2 Dual Core Processor 5600+ 2.81 GHz
Memory (RAM)	8 GB
Hard Disk	Western Digital WD1500AHFD—00RAR5 Size : 149 GB
Operating System	Microsoft Windows XP Professional x64 Edition
Network (1 Gbps)	NVIDIA nForce Networking





## ABOUT EQUIVIO

Equivio develops text analysis software for e-discovery. Users include the DoJ, the FTC, KPMG, Deloitte, plus hundreds of law firms and corporations. Equivio offers Zoom, an integrated web platform for analytics and predictive coding. Zoom organizes collections of documents in meaningful ways. So you can zoom right in and find out what's interesting, notable and unique. Request a demo at [info@equivio.com](mailto:info@equivio.com) or visit us at [www.equivio.com](http://www.equivio.com).

**Zoom in. Find out.**