

## Finance Midmarket Client Unstructured Data Analysis Report

**Author : Alastair Williams**

**Version: 1.0**

**Date: 11 January 2010**

## Contents

<b>1.</b>	<b>EXECUTIVE SUMMARY</b>	<b>3</b>
<b>2.</b>	<b>INTRODUCTION</b>	<b>4</b>
2.1.	Purpose of the Unstructured Analysis	4
2.2.	Overview of Concerns	4
2.3.	Information Reviewed	4
2.4.	Exceptions	5
<b>3.</b>	<b>RESULTS – SYSTEM A</b>	<b>6</b>
3.1.	Exceptions	6
3.2.	Overview	6
3.3.	Capacity and File Prevalence	7
3.4.	Capacity and File Growth	9
3.5.	Data Persistence	10
3.6.	Modification Profile	11
3.7.	Data Dormancy	12
3.8.	Duplication Impact	13
<b>4.</b>	<b>RESULTS – SYSTEM B</b>	<b>14</b>
4.1.	Exceptions	14
4.2.	Overview	15
4.3.	Capacity and File Prevalence	15
4.4.	Capacity and File Growth	17
4.5.	Data Persistence	18
4.6.	Modification Profile	19
4.7.	Data Dormancy	20
4.8.	Duplication Impact	21
<b>5.</b>	<b>RECOMMENDATIONS</b>	<b>22</b>
5.1.	Implement HSM (Hierarchical Storage Management)	22
5.2.	Introduce Version Aware Backups	22
5.3.	Introduce Continuous Data Protection (CDP)	22
5.4.	Review Business Reporting Techniques	22

## 1. EXECUTIVE SUMMARY

Centiq undertook a high level Unstructured Data Analysis (UDA) to enable a finance client to understand the state of the two corporate file server systems and the value of data archiving within this environment. The conviction within the business units is that all data is required at all times and therefore archiving and removal of data from these systems is not possible. Storage growth was estimated to be around 10% per annum. The resulting design is now causing systems management and maintenance window pressures.

The UDA report showed that:

- Storage capacity growth is closer to 20% per annum,
- 12,000 new files are created a month and creation rate is increasing
- Over 250,000 spreadsheets exist across the 2 systems, (average 1000 spreadsheets per user)
- 60,000 duplicates files.
- The majority of the unstructured information is static (unchanging)
- A significant percentage is inactive and unused.

Our recommendations, (shown in section 5), include implementing hierarchical storage management and changing the current backup design to include version aware or continuous protection methods. Each of these alternatives would reduce significant portions of the capacity management and maintenance window pressures currently seen. A budget of around £20,000 - £25,000 should be considered to fully implement a single recommendation. Implementing recommendations 5.1 – 5.3 in parallel would give integration and management advantages and typically costs are between £35,000 and £40,000 for environments of this size.

To understand the full business costs of the current storage pressures and what future budget can be deferred or removed a detailed Infrastructure Assessment workshop is advised.

## 2. INTRODUCTION

### 2.1. Purpose of the Unstructured Analysis

Two options are available for the Unstructured Data Analysis. The first is a broad brush point in time analysis of the files held on an organisations files systems. The second is a more detailed engagement where the flow of information through an organisation is analysed by department and an information map is created in lines with the business process.

The point in time report has the benefit of quick deployment and results, and gives general recommendations typically based on IT's pain points of capacity growth, file type management, data duplication and non-business files.

The detailed consulting, because it engages directly with the business units, details how information is created, shared and managed through an organisation and as such provides insight to the file systems purpose for different departments. This more detailed engagement often highlights business risks IT are unaware of and process inefficiencies that could be removed with simple IT functions that also reduce the cost overheads to IT. The resulting report includes an information process map for each analysis business function and documents the files' current and long term value to the organisation. It can form the basis of a corporate information management policy document that addresses archive, retention and business continuity objectives.

Centiq ran the first level analysis for the finance client and the following sections are the results and recommendations based on those results.

### 2.2. Overview of Concerns

The main area of concern defined by the finance client was with regards to IT operations efficiency. This manifested in a number of sub-areas namely:

- Storage growth and planning - How capacity can be managed and costed effectively
- Archive definition – What data estate looks like prior to commencing any archive project
- Backup management – How to meet server availability constraints on growing backup demands
- Business recovery – How to prioritise data access and recovery in the event of system failure

A de-duplication device was considered to assist with the back end overheads of tape management. We recognize that this may have significant benefits with regards the backup management however, as the proposed unit post-processes, it is our expectation that the backup window pressure would not be relieved by this device.

### 2.3. Information Reviewed

Two systems are the subject of the UDA namely System A and. These servers run the file services for THE FINANCE CLIENT and store a range of files from traditional office documents, PDF output and user data. These servers were analyzed in March 2009 and all dates and figures relate to this date.

#### 2.4. Exceptions

It is our experience that the majority of objects in a typical file system are less than 2GB in size. We limit the import of data to files reporting as 2GB or less as incorporating these exceptionally large files would skew the results and reduce the effectiveness of the report. Where exceptions occur these outsized files are analyzed separately and highlighted in the overview statement however they are not included in any graphical presentation.

### 3. RESULTS – SYSTEM A

#### 3.1. Exceptions

There were six exceptions based on files exceeding the 2GB import limit. These are listed below and include one instance of file duplication. 32GB of capacity is associated with these files.

Name	Type	Length MB	Creation	Last Access	Last Write	Directory
ABCDE180107.bkf	.bkf	3786.33728	30-Jan-07	05-Aug-08	18-Jan-07	L:\User Data\Departmental Data\Information Systems\Support
ABCDE180107.bkf	.bkf	3786.33728	29-Jan-07	02-Dec-07	18-Jan-07	L:\User Data\Departmental Data\Information Systems\Personal\
USER_Z.pst	.pst	2307.933184	25-Jan-08	28-Jan-08	28-Jan-08	L:\User Data\Departmental Data\Information Systems\Personal\n\user .pst files
SCAN.tib	.tib	9220.273664	20-Jul-07	16-Feb-09	20-Jul-07	L:\User Data\Departmental Data\Public Data
Backup.bkf	.bkf	2654.394368	11-Jan-05	02-Jan-09	11-Jan-05	L:\User Data\Personal Data\User_A
VirtualBackup.bkf	.bkf	10507.50976	06-Oct-04	18-Jul-07	06-Oct-04	L:\User Data\Personal Data\User_B\Virtual Backup

#### 3.2. Overview

This section covers the highlights of each system, prominent file types and capacity utilisation of the systems.

The System A system comprises of 782,000 files accounting for 280GB (excluding the above exceptions) of stored capacity. Nearly  $\frac{3}{4}$  of this capacity is consumed by the top 10 file types with 27% of the capacity used in spreadsheets and 8% of capacity stored through Access databases.

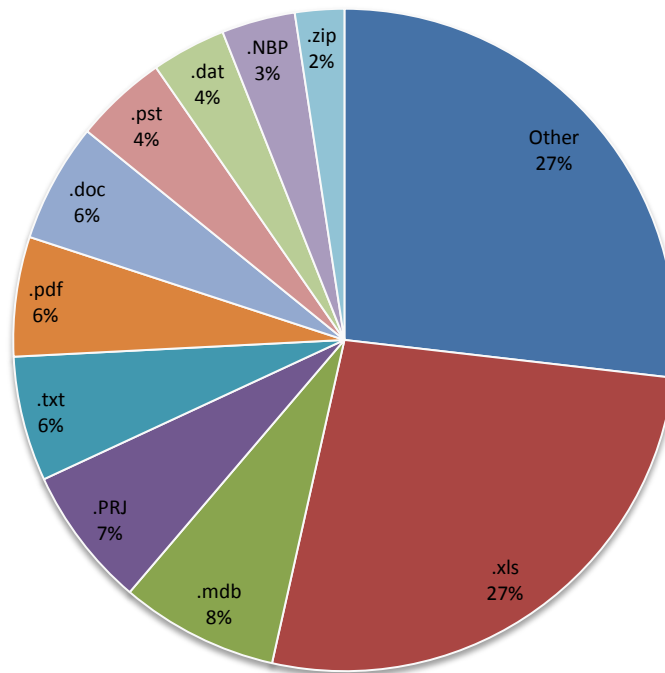
On average 300 new objects are added per day to this system and there are 1700 files which would be classed as backups. These include Zip and BKP files for example. It is likely especially in the case of zip files that there is additional duplication of data.

There are 6.5GB of media files, 12 GB of email “archives” including one lotus notes NSF archive

There are 3500 executable files (.exe) which may not be appropriately stored.

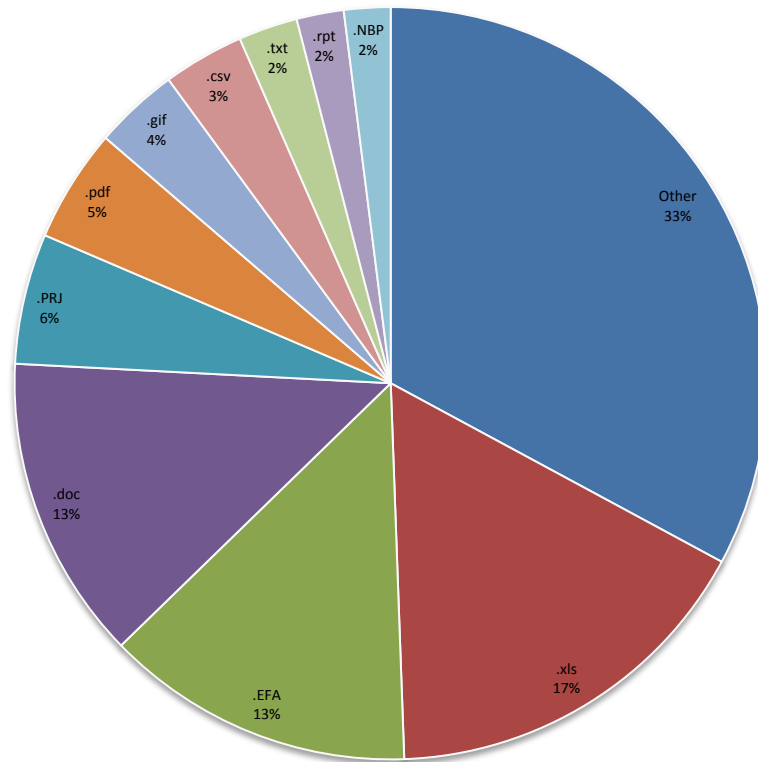
The analysis also scans for file names with inappropriate language that may bring the company into disrepute should they be forwarded outside the organisation. None of the searched words were found in this scan

### 3.3. Capacity and File Prevalence



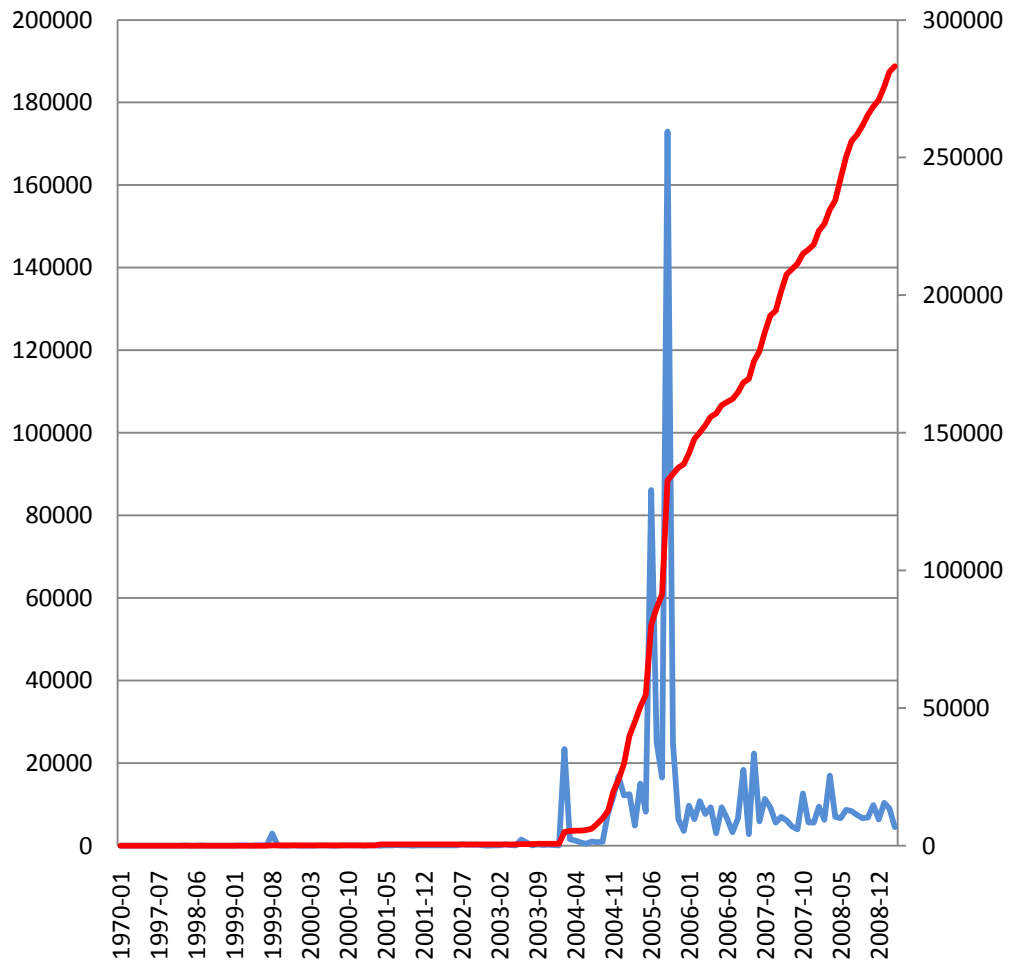
The above graph shows the capacity impact of the top 10 storage consuming file types.

This shows that over 1/3<sup>rd</sup> of the capacity is associated with tracking and report generation tools of Excel and Microsoft Access. How these spreadsheets are used cannot be determined however there is indication shown in later graphs that these are static documents suggesting one-off usage. Further understanding of how the business uses these tools is needed to understand whether this is the most efficient use of storage capacity and users time.



This graph shows the top ten files by count, showing where the majority of user effort and file creation occurs. This shows that around 230,000 files on the system are MS word or Excel.

### 3.4. Capacity and File Growth

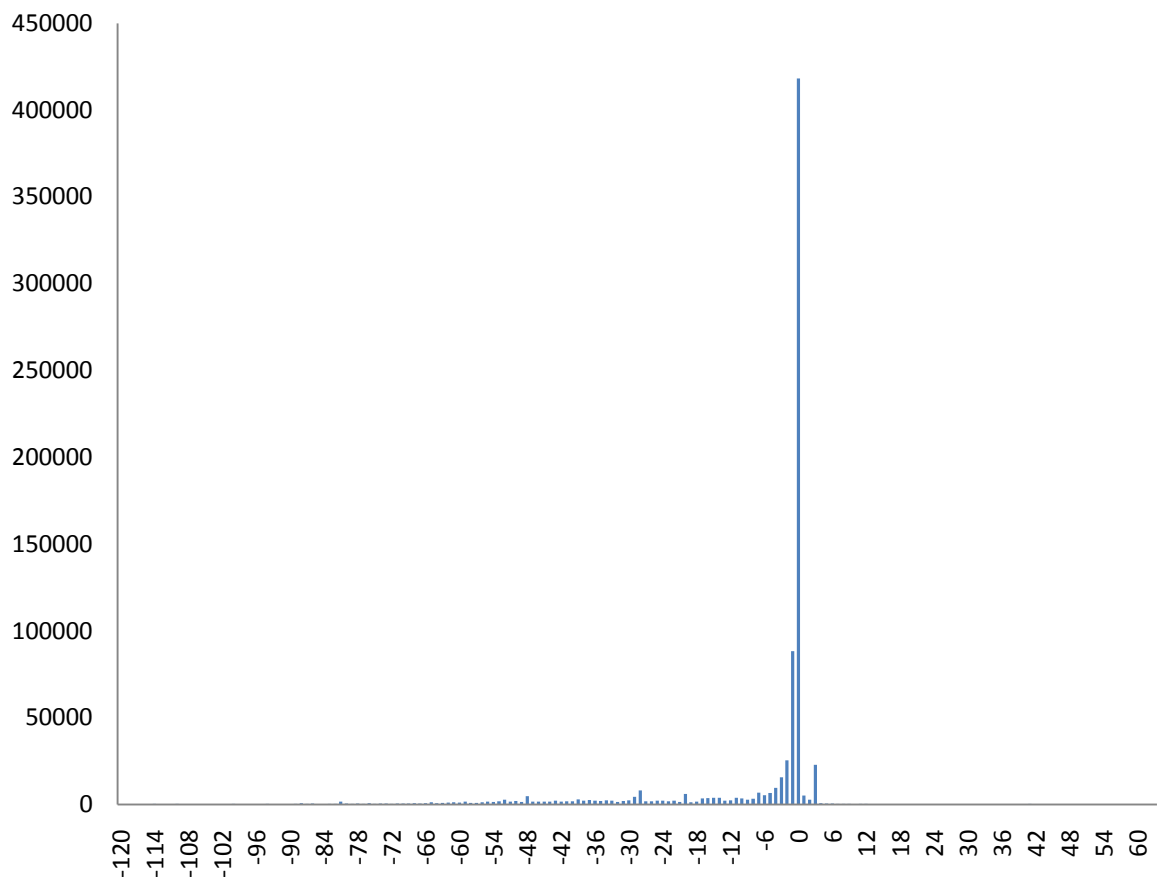


This graph shows the capacity growth and the file creation profile associated with this system (y2). The red line indicates the capacity increase which is currently running at around 50GB increase in the past 12 months and equates to around 20% growth per Annum.

The blue line indicates the number of files created or added to the system (y1). High peaks typically denote file migrations or system consolidation. This information is based on the file creation date. Significantly old files may be the result of operation system packages; however these typically have little impact on the capacity utilisation. In the three months running up to the data collection 24,000 files were created or added to this system

### 3.5. Data Persistence

This section covers how data is stored and modified on the system.



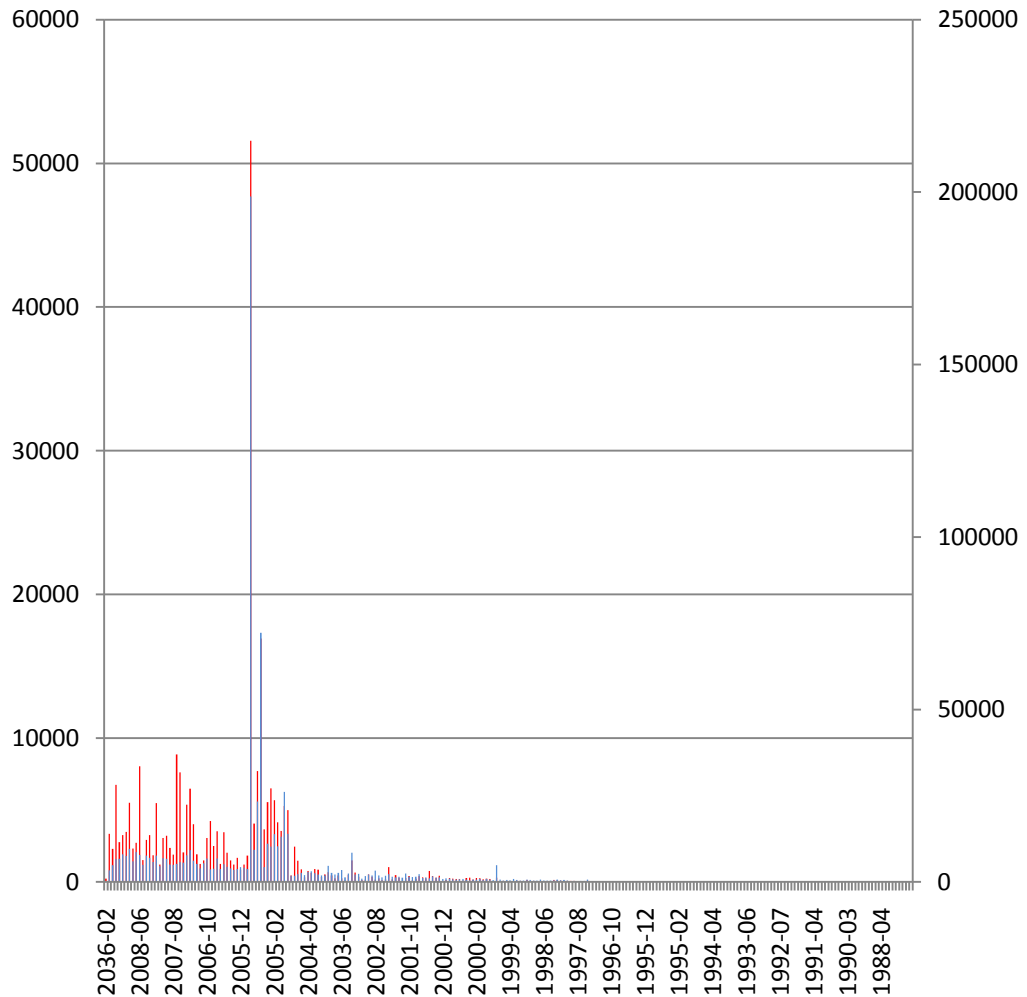
This graph shows the number of files (y) and whether they have a modified date older or younger than the creation date (the date that the file first appears on the system) (x) scale, in months. Negative numbers indicate that the file was modified prior to creation and indicates that the file was migrated, 0 indicates that the file was created and modified within the same month whilst positive numbers indicate that the file was modified at some point after the creation. This shows that whilst some files are modified the majority of files remain passive (unchanged) after the month they were migrated/created.

53% of files have the same creation and change dates and 41% of files indicate migration, 17% are less than a month old when loaded. This leaves only 6% of the system files modified post creation

When compared with capacity the unchanged or migrated data accounts for 86% of the storage usage.

### 3.6. Modification Profile

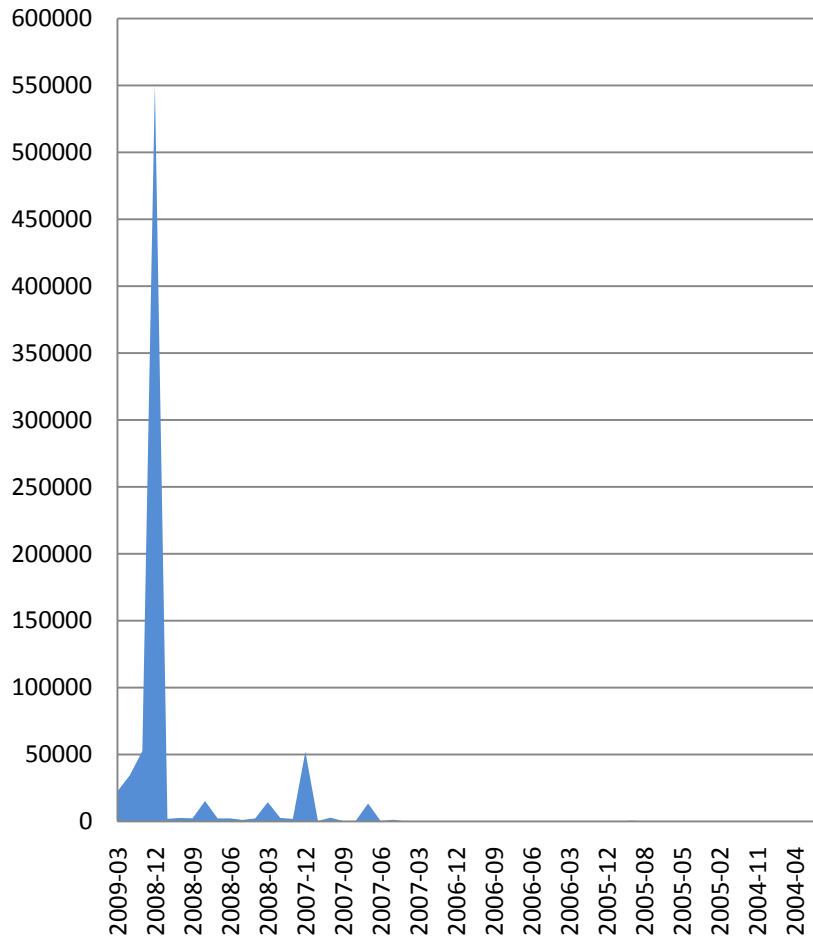
This section shows when files were last modified against the current date.



Number of Files (y1) shown in blue and capacity (y2) in red and the year month last modified.

This shows that 97.6% of the files have not been modified in the past 3 months. When reviewed the average capacity change per month excluding growth is 6.5GB per month or 0.8% change. New files account for 4GB per month. This implies an optimised backup of 10-15GB per month is required.

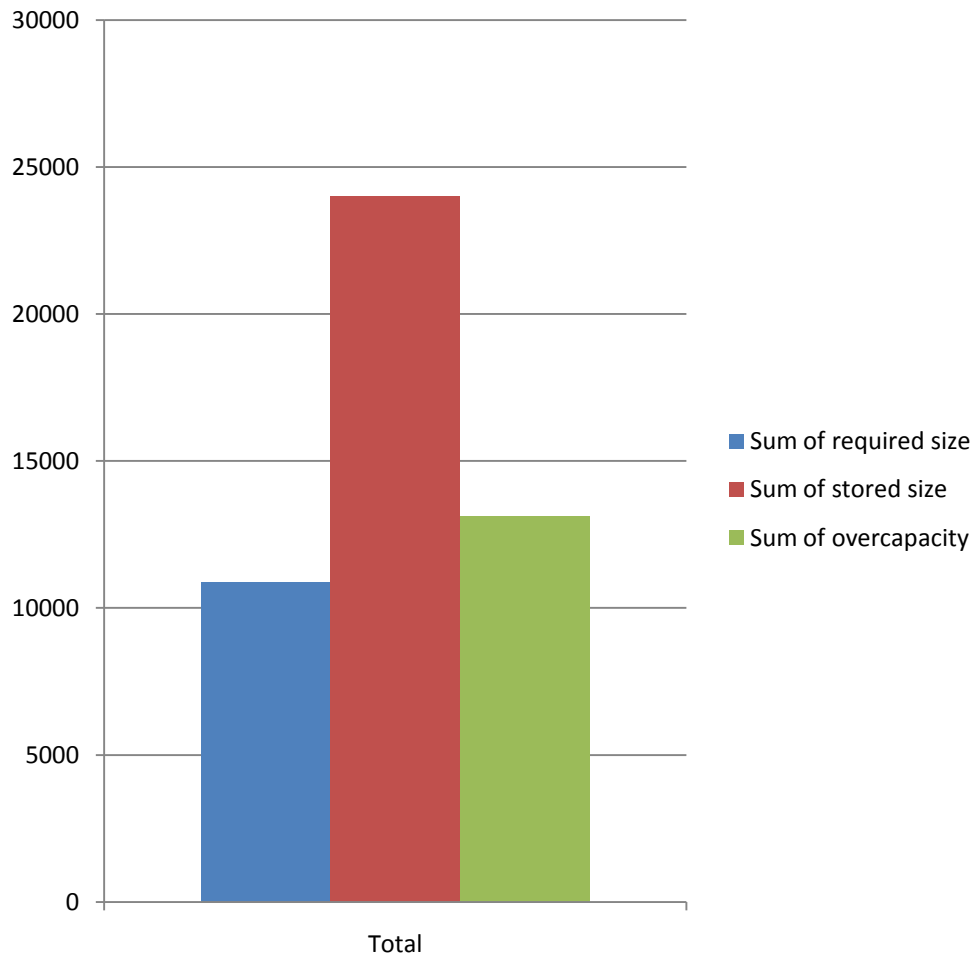
### 3.7. Data Dormancy



This graph relates to the date last access and indicates the ongoing value of the files for instant access, (as opposed to business value for compliance or data retention obligations such as Mifid). A significant opening of 65% of files in mid December over a period of three days has skewed this graph so it is not possible to make definitive conclusions. However this still shows that 86% of files had not been accesses in the past 3 months and that equates to 81% or 230GB of the system capacity.

Further analysis showed that 3% of files were not accessed after last modified or created. (\*skewed by file opening)

### 3.8. Duplication Impact



This graph shows the impact of files that have the same name and size. Whilst not definitive it is a good indicator of file duplication across the system. This shows that 8% of the capacity is wasted due to file duplication with these files needing around 11GB of capacity yet with the multiple copies stored 23GB or double the needed capacity is in use. This equates to around 50,000 files with 2700 of these being spreadsheet duplicates and account for 15GB of the stored capacity.

Further investigation showed that around 40,000 files had the same name but different sizes, (10GB capacity of which 3GB is spreadsheets). Whilst these may be different files it can lead to poor decision making within a business unit if the wrong copy or version of the file is used as the source.

## 4. RESULTS – SYSTEM B

### 4.1. Exceptions

There were 13 exceptions based on files exceeding the 2GB import limit. These are listed below and account for 65GB of capacity. The majority are .BAK files

Name	Extension	Length MB	Creation	Last Access	Last Write	Directory
SecurityArchive.rar	.rar	4732.625259	08-Nov-06	08-Nov-06	08-Nov-06	E:\Data\User Data\Departmental Data\IS
Backup 061106.bkf	.bkf	5941.045248	06-Nov-06	06-Nov-06	06-Nov-06	E:\Data\User Data\Departmental Data\Information Systems\pc backups\USER_T
Backup060904.bkf	.bkf	4397.883392	04-Feb-06	10-Jun-06	06-Sep-04	E:\Data\User Data\Personal Data\USER\PC Backup pre upgrade
outlook.ost	.ost	3077.915648	20-Feb-09	27-Feb-09	20-Feb-09	E:\Data\User Data\Personal Data\USER_U
VirtualBackup.bkf	.bkf	10507.50976	04-Feb-06	10-Jun-06	06-Oct-04	E:\Data\User Data\Personal Data\USER_Z\Virtual Backup
Tracking_db_200902062200.BAK	.BAK	4513.991168	06-Feb-09	06-Feb-09	06-Feb-09	E:\SQL Backups\ Tracking
Tracking_db_200902070200.BAK	.BAK	4503.505408	07-Feb-09	07-Feb-09	07-Feb-09	E:\SQL Backups\ Tracking
Tracking_db_200902071333.BAK	.BAK	4506.323456	07-Feb-09	07-Feb-09	07-Feb-09	E:\SQL Backups\ Tracking
Tracking_db_200902080201.BAK	.BAK	4497.213952	08-Feb-09	08-Feb-09	08-Feb-09	E:\SQL Backups\ Tracking
Tracking_db_200902090200.BAK	.BAK	4494.854656	09-Feb-09	09-Feb-09	09-Feb-09	E:\SQL Backups\ Tracking
Tracking_db_200902100200.BAK	.BAK	4513.991168	10-Feb-09	10-Feb-09	10-Feb-09	E:\SQL Backups\ Tracking
Tracking_db_200902110200.BAK	.BAK	4530.899456	11-Feb-09	11-Feb-09	11-Feb-09	E:\SQL Backups\ Tracking
Tracking_db_200902120200.BAK	.BAK	4542.30272	12-Feb-09	12-Feb-09	12-Feb-09	E:\SQL Backups\ Tracking

## 4.2. Overview

This section covers the highlights of each system, prominent file types and capacity utilisation of the systems.

The System Bsystem comprises of 482,000 files accounting for 220GB, (excluding the above exceptions), of stored capacity. 85% of this capacity is consumed by the top 10 file types with 7000 text files accounting for 26% which appear to be audit files. Centiq recommends a review of these audit files to confirm whether they require “weight of evidence” protection available through file fingerprinting or WORM storage devices.

The remaining capacity has a similar profile to System A with 20% of the system taken by spreadsheets.

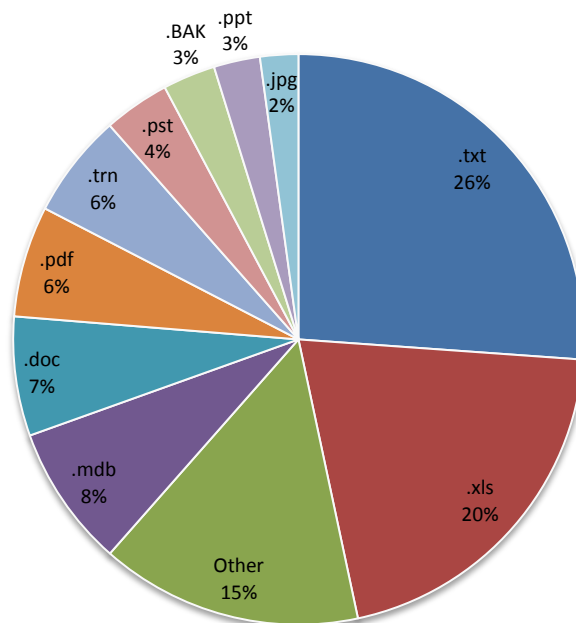
On average 240 new objects are added per day to this system and there are 2200 files which would be classed as backups, 5.5GB of media files and 8 GB of email “archives”

There are 1950 executable files (.exe) which may not be appropriately stored.

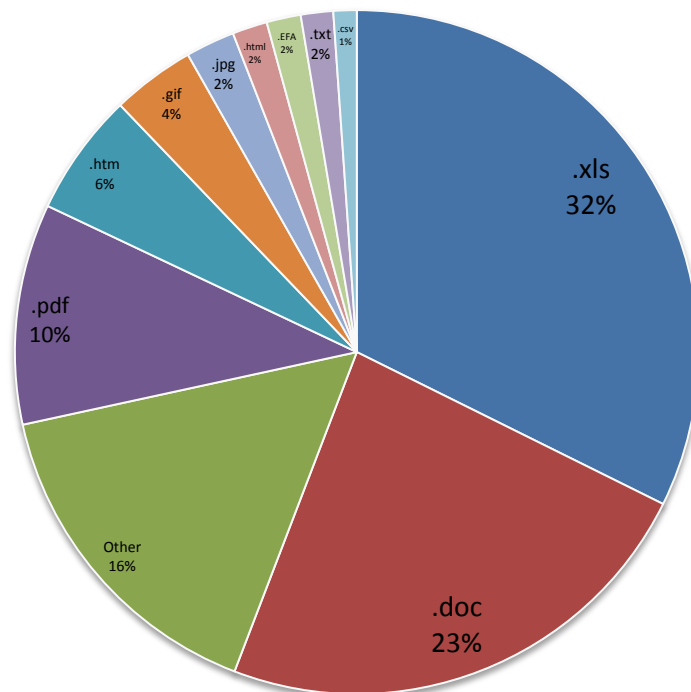
There are no inappropriate file names

## 4.3. Capacity and File Prevalence

This graph shows the capacity impact of the top 10 storage consuming file types.

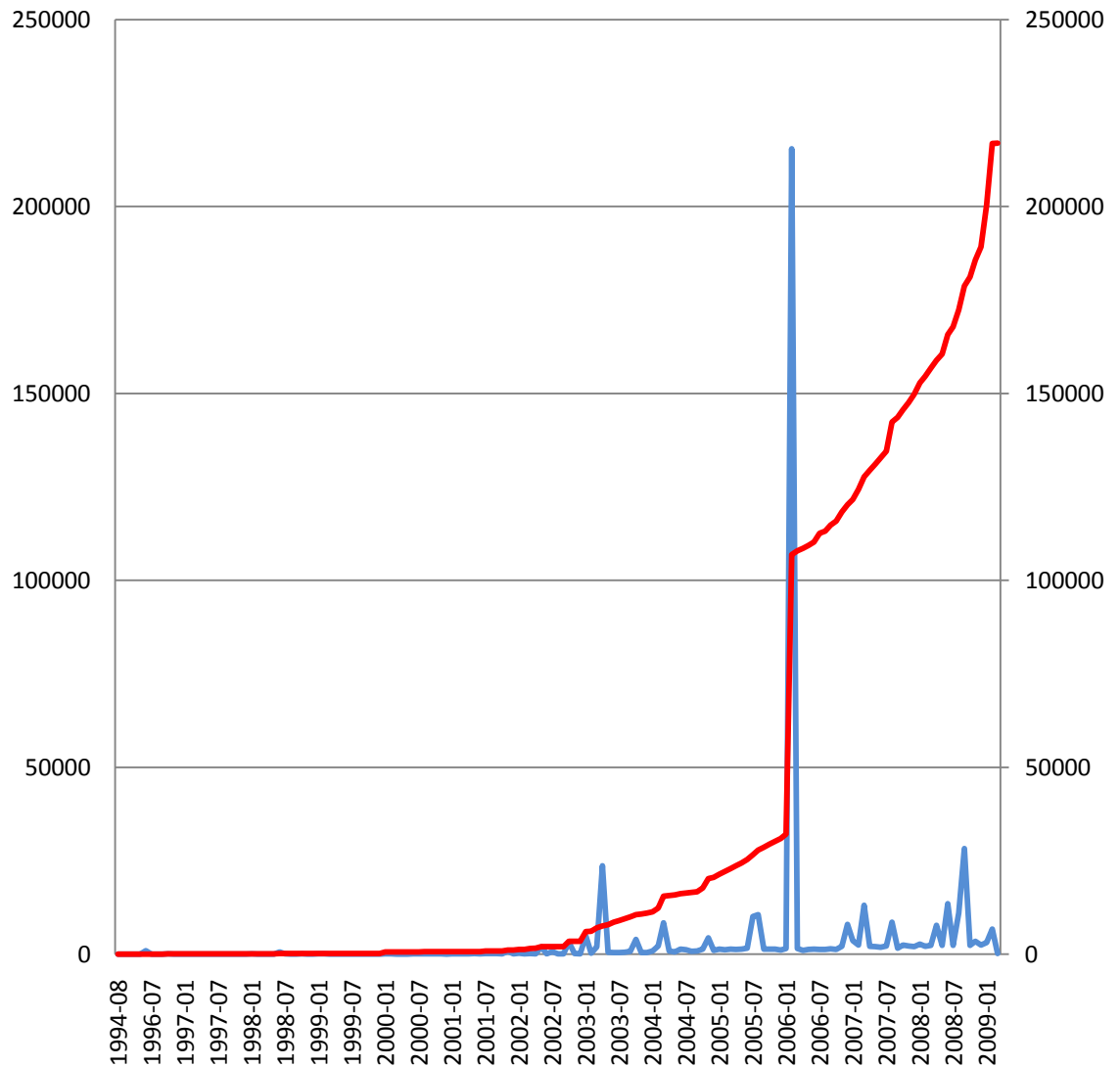


Apart from the high capacity used by the audit .txt files this graph again shows that nearly 1/3<sup>rd</sup> of the capacity is associated with tracking and report generation tools of Excel and Microsoft Access. How these spreadsheets are used cannot be determined, however again there is indication these may be static documents. Further understanding of how the business uses these tools is needed to understand whether this is the most efficient use of storage capacity and users time.



This graph shows the top ten files by count, showing where the majority of user effort and file creation occurs. This shows that over 50% of the files on the system are office documents. Additionally 10% of the files (around 50,000) are static PDF documents

#### 4.4. Capacity and File Growth



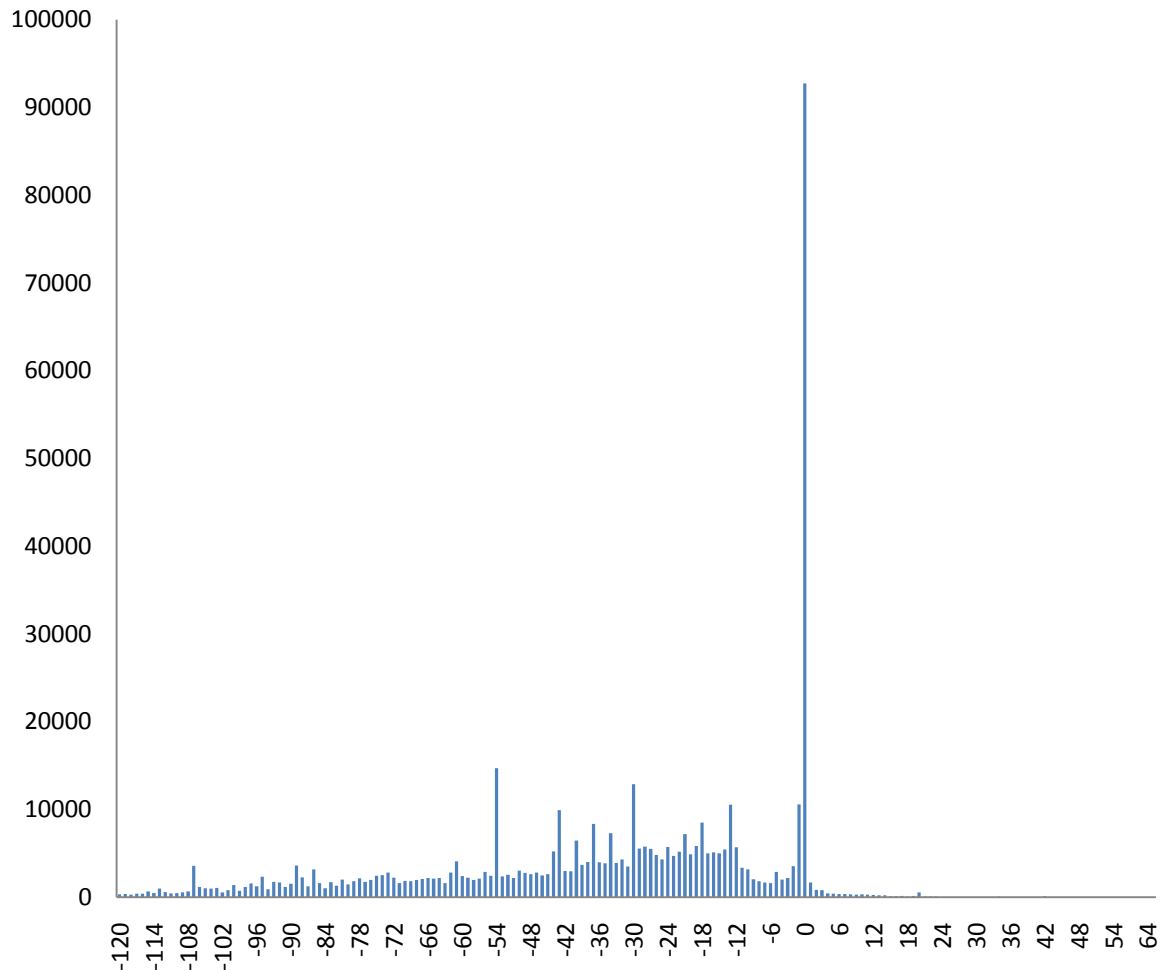
This graph shows the capacity growth (y2) and the file creation profile (y1) associated with this system. The red line indicates the capacity increase which on this system is currently running at around 40GB for the past 12 months and equates to around 18% growth per annum.

Due to a flat growth month prior to data collection actual growth rates may be higher as certain months show growth closer to 5GB.

The blue line indicates the number of files created or added to the system. High peaks typically denote file migrations or system consolidation. This information is based on the file creation date 10,000 files have been created in the past 3 months

#### 4.5. Data Persistence

This section covers how data is stored and modified on the system.

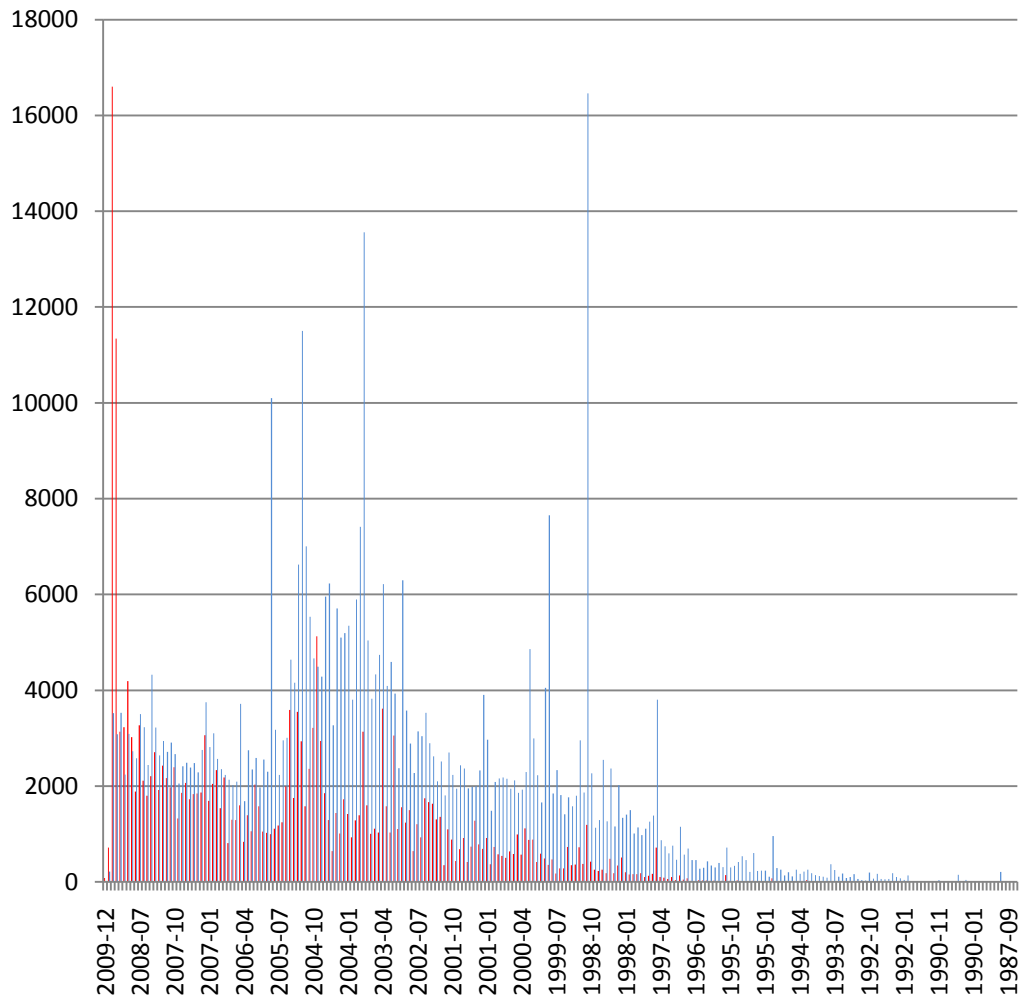


This graph shows the number of files (y) and whether they have a modified date older or younger than the creation date on the system (x) scale in months. Negative numbers indicate that the file was modified prior to creation and indicates that the file was migrated, 0 indicates that the file was created and modified within the same month whilst positive numbers indicate that the file was modified at some point after the creation. This shows that whilst some files are modified the majority of files remain passive (unchanged) within the month.

19% of files have the same creation and change dates and 79% of files indicate migration. This leaves only 2% of files modified after loading/creating on the file system. These files also account for nearly 98% of the system capacity with just 2% of capacity changing more than one month after creation.

#### 4.6. Modification Profile

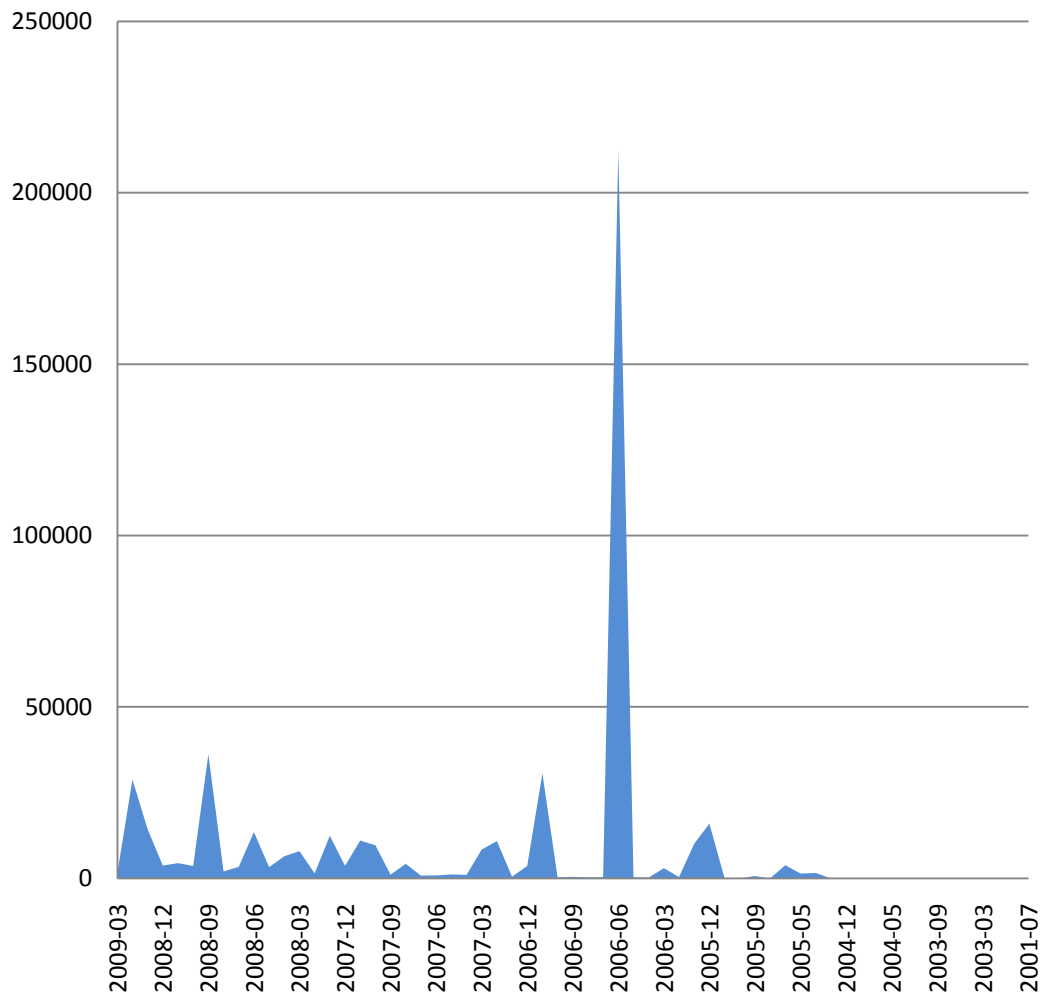
This section shows when files were last modified against the current date.



Number of Files (y1) shown in blue and capacity (y2) in red and the year month last modified.

This shows that 86.8% of the files have not been modified in the past 3 months. When reviewed the average capacity change per month excluding growth is 4GB per month and peaks at 16GB. New files account for 5GB per month. This implies an optimised backup scenario of 10-20GB per month is required.

#### 4.7. Data Dormancy

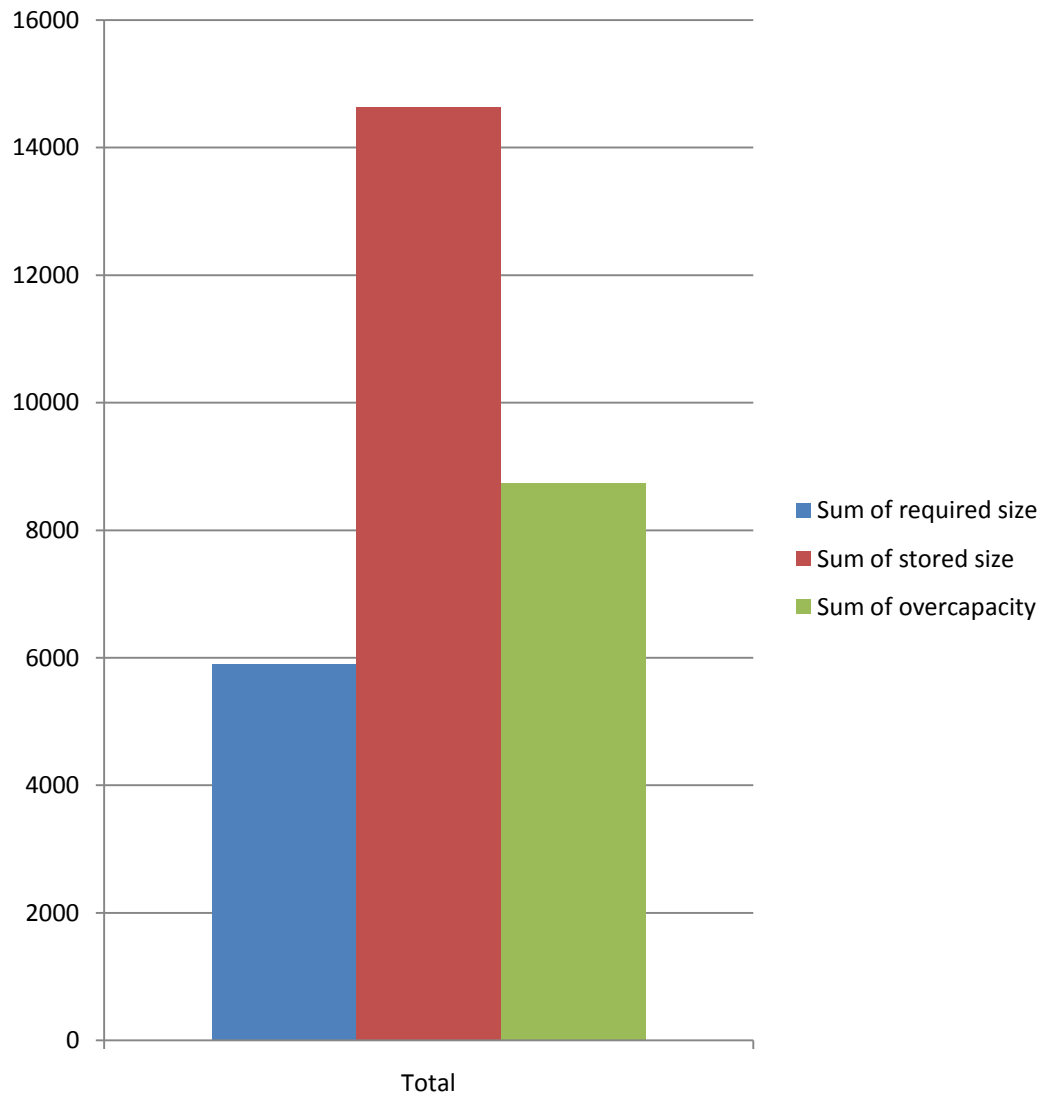


This graph shows relates to the date last access and indicates the ongoing value of the files for instant access, (as opposed to business value for compliance or data retention obligations such as Mifid).

This graph shows a more common profile to that of System A with around 90% of files not having been accessed in the previous 3 months to data collection. This equates to 77% of the system capacity (180GB).

Further analysis showed that 15% of files were not accessed after being placed on the system which equates to 12% of the capacity.

#### 4.8. Duplication Impact



This graph again shows that there is file duplication on the server and the capacity impact of those files albeit at a lower level to that seen on the System A system. These equate to 11,000 objects including 1300 spreadsheets.

Around 15,000 files had the same name but different sizes and again around half the capacity taken by these files is due to possible spreadsheet duplication. Again accurate decision making may be impacted by the ability to establish the "current" file version.

## 5. RECOMMENDATIONS

In order to address the ongoing storage management issues Centiq recommend an infrastructure discovery workshop to quantify the full business value of the following storage management alternatives.

### 5.1. Implement HSM (Hierarchical Storage Management)

After confirming the compliance requirements of the business, implement an HSM solution that migrates information to a secondary storage tier, typically slower drives or tape devices automatically. Solutions that leave stub files on the server would remain invisible to the user community and providing the secondary device is online there would be limited performance impact when opening rarely used files. When combined with de-duplication technology there is the additional benefit of reducing the storage footprint without modifying the file profiles. This will not however address the number of files users are generating

### 5.2. Introduce Version Aware Backups

By running standard weekly full backups a significant amount of data is re-saved without having changed. From the study it is estimated that backups could be reduced to a maximum of 25GB per night. Software that recognises changed data continually and can generate full backup images out of these incremental changes can significantly reduce maintenance window pressures. Often backups are taken out of cycle to provide "archives". Centiq do not recommend this approach and recommend separate archives based on business unit need.

With regards archives Centiq recommend that an email archiving tool that enables the incorporation of PST files be used. Current practice encourages user defined email archive policies which may not meet corporate needs.

### 5.3. Introduce Continuous Data Protection (CDP)

An alternative or complementary technology to version aware backups is the introduction of CDP software to "backup" new files and changes only as they occur to reduce backup windows and recovery times. As individual files within the system are unlikely to be required throughout the day, any changes can be collected and saved to the backups as they occur. This removes the need for a specific offline backup window. These systems also provide a shorter recovery point and recovery times can be reduced in the event of a file system failure as users would be able to access the information directly from the CDP server.

### 5.4. Review Business Reporting Techniques

Undertaking the full UDA (Unstructured Data Analysis) within the business units to understand the use and creation rationale of spreadsheets would allow Centiq to make additional recommendations to improve user productivity, decrease duplication of effort, automate repetitive reporting functions as well as reduce the storage management overheads of IT. The resulting map would form the basis of any archiving policy and would take into account industry regulations and obligations.