File share analysis reports

xillio

making content work

Introduction

It is hard to find an organization that is not still using file shares. Nowadays, file shares are outdated, have high costs to maintain, create chaos due to the immense numbers of copies of files and insufficient ways of structuring them, and risk exposure is significant. Do you finally want to order this chaos and solve the problems it brings?

Xillio file share analysis

Xillio can perform extensive file share analysis, and generate reports in a very short time frame. You can run this process through your preferred systems integrator, or just let Xillio do the work. The analysis lets you understand which content on your file shares is redundant, obsolete and trivial. Or, gives insight in what content is located on your files shares in general.

The Xillio file share analysis package includes a number of pre-built reports and analyses. This data is all captured using the delivered scripts and robots and stored in a database for easy access and reporting. This data may be further exported to Excel or integrated with another Business Intelligence tool of choice for further analysis.

In this document Xillio explains the delivery of the file share analysis. Each report gives different insights, examples of which are also given in this document. These insights are not extensive, there are many other variations of these which can be produced to create your own new insights.

To guide you through this process, for each report the most used insights that you can get out of it, is described.

Report examples

1. Probable duplicates of content

This report shows the number of probable duplicate files based on file size and extension. The duplicates are shown with their location, name, extension and size. The duplicates are probable, if you want to be certain about their duplicity, you have to run the duplicates of content report (See section 2).

Example:

Number of duplicates	File location	Name	Extensions	Size in MB
3	G:\Data\Sales	Info	TXT	4
	G:\Data\HR	Info sales	TXT	4
	G:\Data\Operations	How to operations	DOC	15
2	G:\Data\Sales	How to OP	DOC	15
	G:\Data\HR	How to operations	DOC	15

Insights:

- **How many duplicates are there in total?**
- **Insights in number of duplicates per duplicate (order)**
- **Insights in location of interested duplicates (needed to act)**
- **Insights in total amount of storage used for duplicates**

2. Duplicates of content

This report is an extended version of the previous report and the shows the duplicates that are certain. Based on this report, you can make choices about which file to persist or what other actions to take.

3. Common file names

This reports shows the file names that are 10 times or more on your system. It also states how many times the same file name is there.

Example:

File name	Count	File size in MB
Info.txt	5	20
How to.doc	2	10
Secret.xsls	5	5

Insights:

- Are there any patterns in certain file names? For instance use of templates and not renaming
- Or are there other files you can think of needing a template.
- Are certain file name actually document types, that can be added in enriching before migration?

4. Empty folders

This report gives an overview of all the empty folder and a list of their locations.

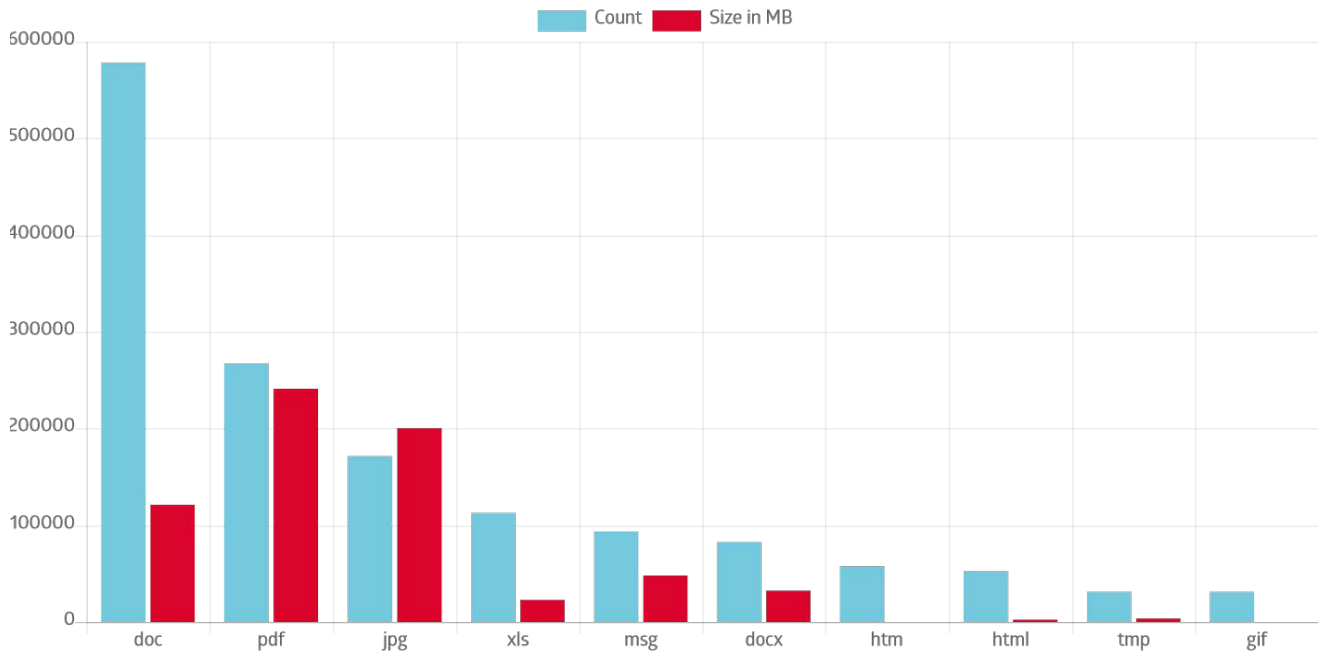
Insights:

- Where are empty folders and what are their names?
- Is it because of a default structure that must be translated to a new system?
- Are there any patterns in the naming of the folders?
- Are there certain root folders that contain most empty folders?

5. File extension and size

This report contains all the file extensions found on the file share(s), and per extension how many files there are along with their total size.

Example:



Insights:

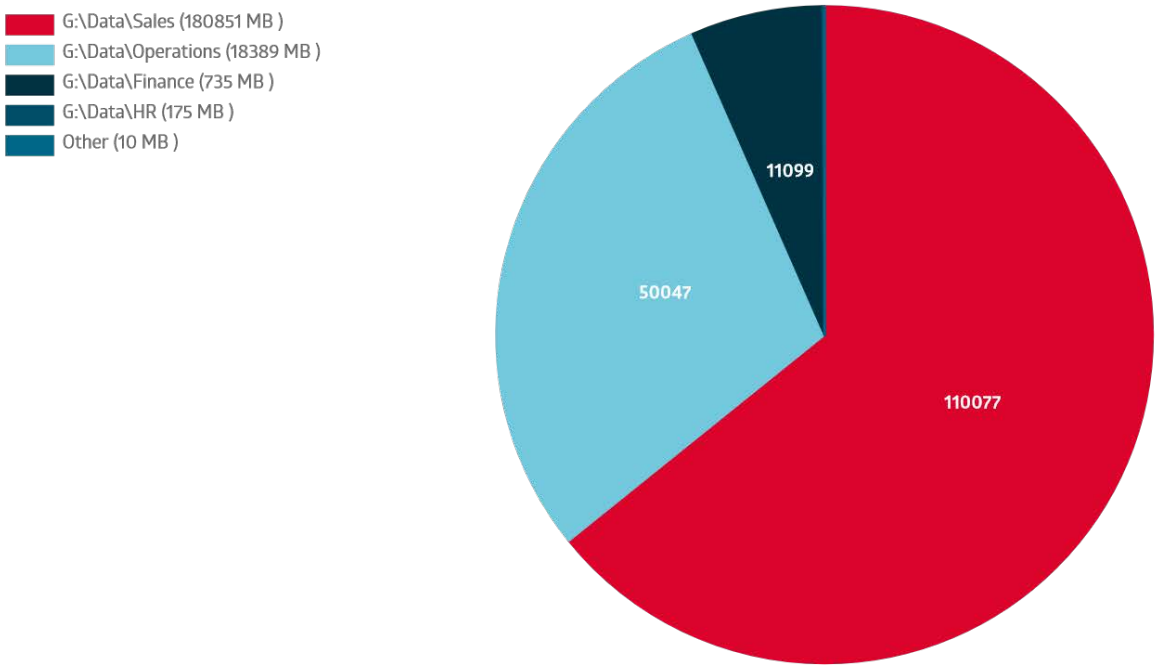
- Most used / present file extensions on total file shares , to give insights in how the network share is used
- Extensions that have biggest share in total sizing, to give input for cleaning strategy
- Amount of different extensions that are used, amount of extensions without value
- Amount of files without an extension, what is the policy on those?
- Average size per extension, are there abnormal outliers?

- What files should be updated with the correct extension
- Files with no extension, what are they and are they interesting to take action on

7. Distribution file extension per level 1 folders

This reports gives the extensions used, how many and the size per level 1 folder. Level 1 folders are usually the functional separation of departments, different projects etc.

Example:



Insights:

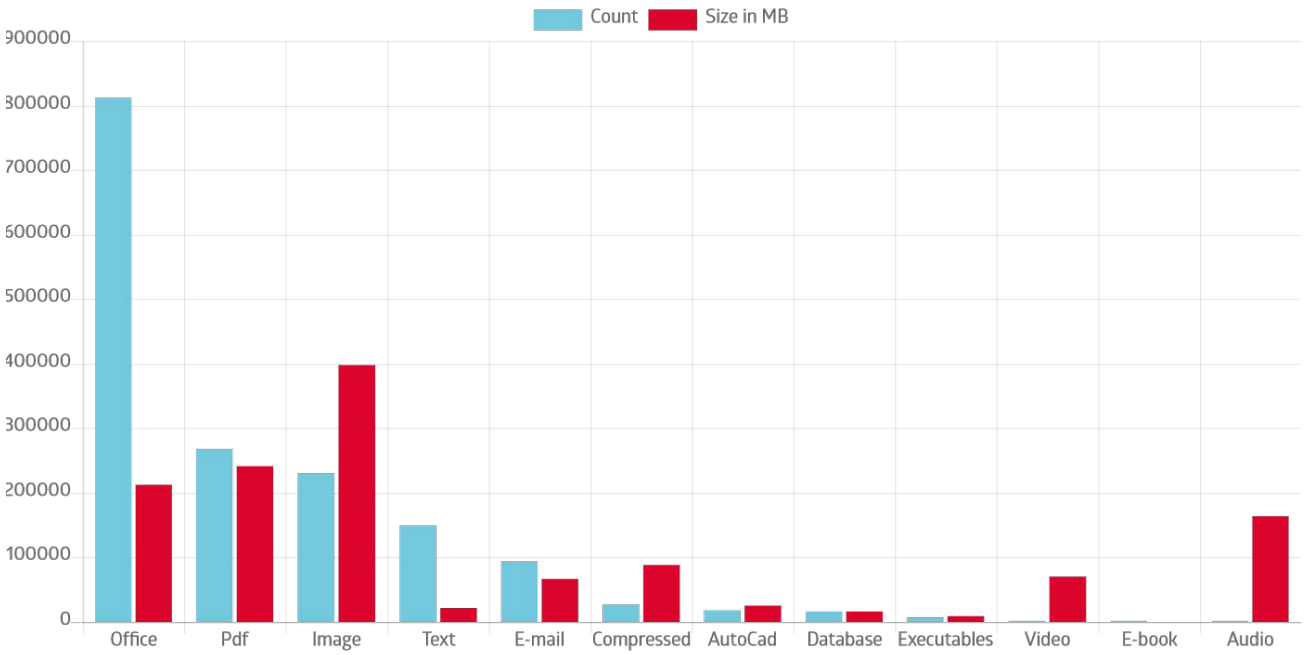
- Insights in number of specific types in different level one folders (often departments or projects within the organization).
- Distribution of specific extensions present within level 1 folder.
- Top number of certain extension and size per level 1 folder.

8. Overview default grouping in file types

This report groups the file extensions into the following logical file groups:

Office	doc dot docx docm dotx dotm docb xls xlt xlm xlsx xlsm xltx sltm xlsb xla xlam xll xlw ppt pot pps pptx pptm potx potm ppam ppsx ppsm sldx sldm
Pdf	pdf ppdf pdfx pdf_ _pdf epdf pdf xml
E-mail	vcf eml rpmsg msf mht pst vcard msg dbx emix mbox ldif rcv emlx vrge08message mso nws mab mbs oas gwi list mlm sbd wdseml mbx tbb pbx ezm box pmm boe pmi msb ldi p10 tnf agt mail bri cnm mau pmx mailhost pop pm1 maildb smd mbg mozeml pmx outlook97 smp gmp pml mmdf vrge08contact schd pmc vrge08event cnm pop pm3 idb bina asv vrge08group vrge08note bdf
Video	264 3g2 3gp asf asx avi bik dash dat dvr flv h264 m2t m2ts m4v mkv mod mov mp4 mpeg mpg mswmm mts ogv prproj rec rmvb swf tod tp ts vob webm wmv
Image	bmp dib dt2 emf gif ico icon jpeg jpg pcx pic png psd tga thm tif tiff wbmp wdp webp arw cr2 crw dcr dng fpx mrw nef orf pcd ptx raf raw rw2 ai cdr csh drw emz odg sda svg wmf
Audio	3ga aac aiff amr ape asf cda dvf flac gp4 gp5 gpx logic m4a m4b m4p midi mp3 ogg pcm snd sng uax wav wma wpl
Text	1st alx asp csv eng htm html log lrc lst nfo opml plist pts reg rep rtf srt sub tbl text txt xml xsd xsl xslt
E-book	azw azw3 cbr cbz epub fb2 iba ibooks lit mobi
Virtualization	ova ovf pvm vdi vhd vmdk vmem vmwarevm vmx
Compressed	001 002 003 004 005 006 007 008 009 010 7z 7zip a00 a01 a02 a03 a04 a05 ace apk appxbundle arc arj asec bar c00 c01 c02 c03 cab cbz cso dlc gz gzip hqx inv ipa isz msu nbh pak r00 r01 r02 r03 r04 r05 r06 r07 r08 r09 r10 rar rpm sis sisx sit sitd sitx tar tgz uax webarchive xap z01 z02 z03 z04 z05 zip
Autocad	dwfx dwg dxf 3ds dct dgn
Database	accdb accdt db dbf fdb gdb idx mdb mdf msmessagestore sdf sql sqlite wdb
Executables	air app application appx bat bin com cpl deb dll elf exe jar js lnk msi prg rpm shs vbs xap
Miscellaneous	Other extensions

Example:



Insights:
- See the total distribution of file groups for your total file share
- Get general insights in how your file share is used, what kind of files are created and used

9. Level 1 folder file type grouping

A detailed version of the previous report showing logical file groups per level 1 folder.

Insights:
- Get insights in how the file shares are used by having a distribution of the different file groups per level one folder
- Check if expected types of files are in the expected locations.

10. Size distribution

This report generates the distribution of number of files per file size range. The default ranges are shown in the example table below. The range of row N is between row N-1 and N.

Example:

Size range	Number	Size in MB
< 1 KB	2347	3
< 10 KB	10000	Etc.
< 100 KB	4563	
< 1 MB	358	
< 10 MB	9282	
< 100 MB	2021	
< 1 GB	549	
< 10 GB	83	
< 100 GB		
< 1 TB		
< 10 TB		
< 100 TB		
< 1 PB		

Example:

Insights:

- How many files of a certain range in size are there on your file share
- How many files can be excluded from for instance classification or migration because they are too small or large

11. Level 1 folder file size distribution

This report is the same as the previous one, but gives the ranges per level 1 (top level) folder. As these are often departments, or projects, this delivers insights on the size distribution per project or department.

Example:

Level 1 folder	Size range	Number	Size in MB
Sales	< 1 KB	2347	3
	< 10 KB	10000	Etc.
	< 100 KB	4563	
	< 1 MB	358	
	< 10 MB	9282	
	< 100 MB	2012	
	< 1 GB	549	
	< 10 GB	83	
	< 100 GB		
	< 1 TB		
	< 10 TB		
	< 100 TB		
	< 1 PB		
Finance	< 1 KB	2347	3
	< 10 KB	10000	Etc.
	etc		

Insights:

- How many files of a certain range in size are there per level 1 folder?
- Is there a department that has a lot of large back-up or video files?

12. Top 1000 of largest files

This report gives the list of top number of largest files on the file share.

Example:

File	Size in MB
G:\Data\Sales\Reports 1994-2017.zip	4550.998562812805
G:\Data\Sales\Images.zip	4161.388414382935
G:\Data\Sales\Customers.mdb	4049.192939758301
G:\Data\Sales\outlook\archive.pst	4040.9150390625
G:\Data\Operations\archief mail\archief.pst	3336.6337890625
G:\Data\Operations\buildings.gml	3147.735954284668
G:\Data\Operations\buildingsv2.gml	3147.735954284668
G:\Data\Documentatie\HR\who_are_whe.mpg	2850.3203125
G:\Data\Finance\arhive\before2000.zip	2582.0751094818115
G:\Data\Finance\archive2.pst	2397.4306640625

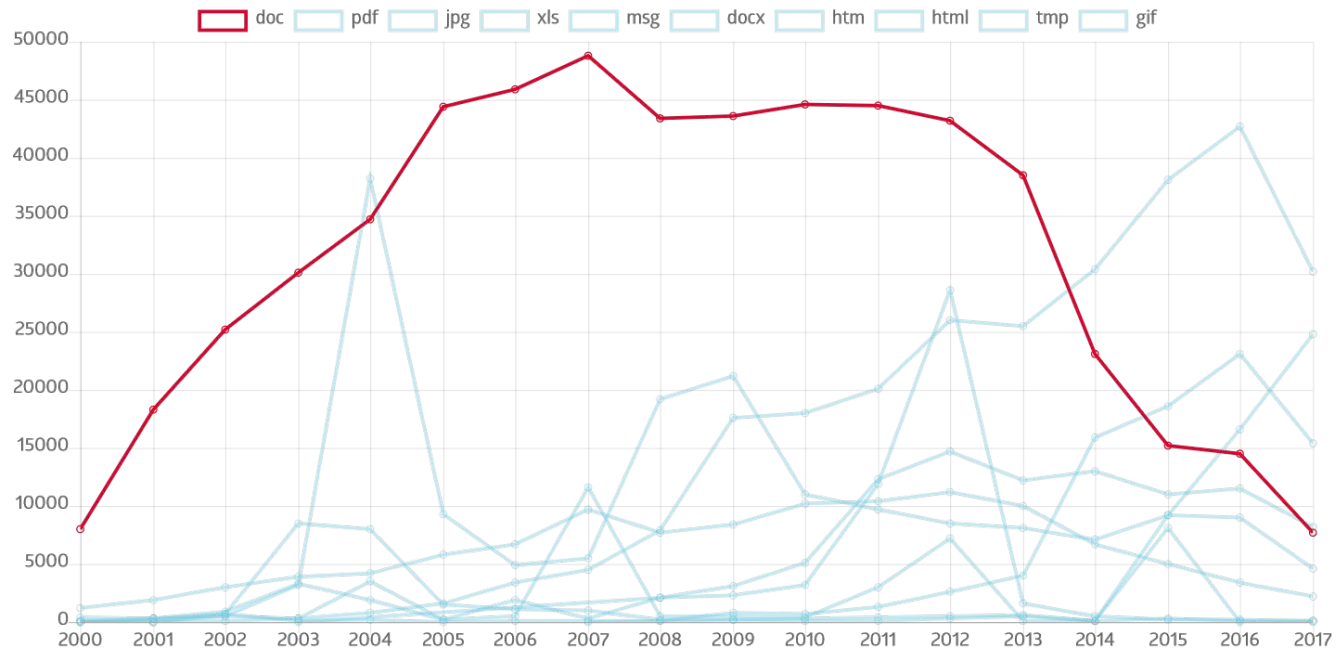
Insights:

- What are the largest files present on the file share and where are they located?
- What type of files are really large?

13. Last modified year per extensions

This report gives an overview of the modified dates. Modified dates are used to give insights in the use of the file share. This is because when copying or accessing a file, the modified date does not change. It is either really editing a document, or the creation of a document, that changes the modified date.

Example:



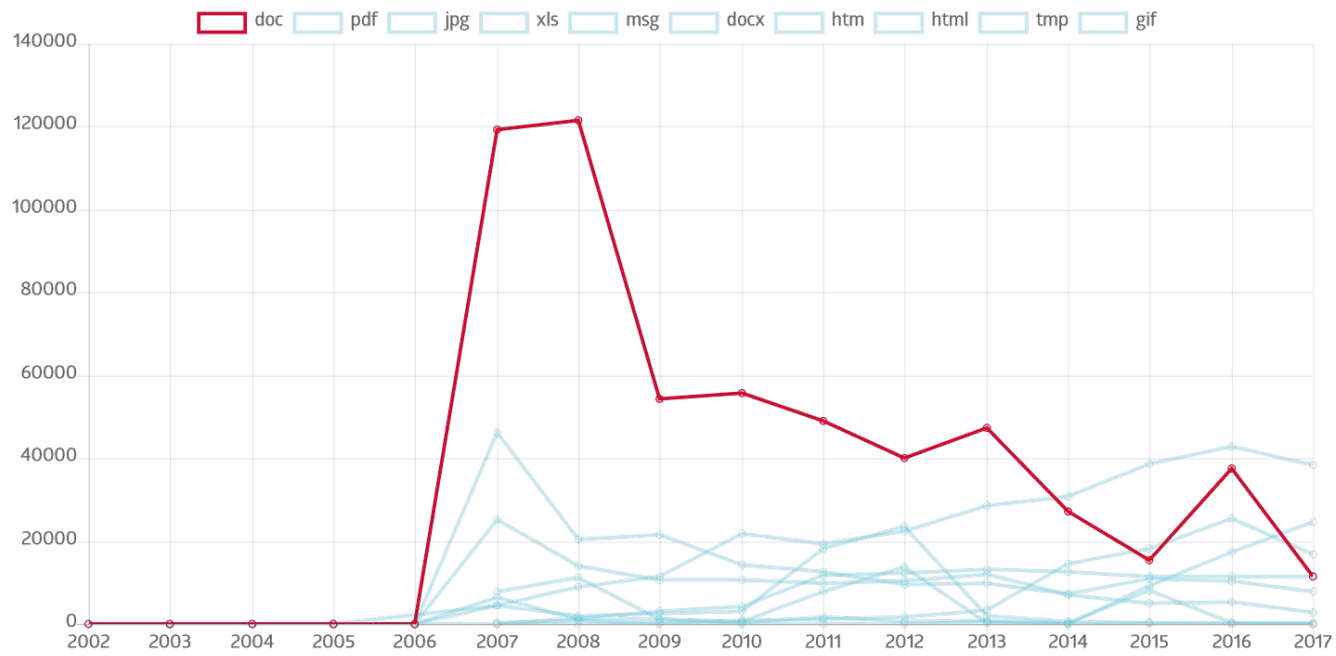
Insights:

- How is the usage of the file share growing or declining per year
- Are there different patterns per file type, for example when Office files are not edited a lot anymore, this could point to the use of SharePoint or another DMS.
- Are there any abnormal spikes per file type? A big spike of PDF for example could point to a digitization project.
- Any other peaks that stand out? Think of AutoCAD files, GIS files, other system files. This gives you information management handhelds for information management roadmap.

14. Creation year per extension

Gives an overview of the creation dates per extensions. Again in a matrix to easily analyse and see patterns.

Example:



Insights:

- Check for the most commonly used extensions how the creation dates are distributed over the years and see where creation peaks are.
- See trends in creation, that can lead to finding large amounts of copied files, legacy data etc.