Heureka Data Assessment Report

Prepared for City of Acme

November, 2018

Heureka
SOFTWARE®

CONTENTS

Quick Stats

Mounted Shares	**10**
Total File Count	**725,233**
Total File Volume	**643.48 GB**
Files > 1GB	**180**
Top File Type (volume)	**Internet**
Encrypted files	**1**
Files older than 2 years	**475,332**
Duplicative files with risk	**470**
File Share	**1**

Thank you for partnering with Heureka Software. Your personalized data assessment report contains an overview of your data with specific focus on volume, duplicative data, potential risk (PII) and risk cost.

Our Mission: Heureka's purpose is to bring order to unstructured data. We accomplish this by identifying risk while helping you realize the value of unstructured data across all endpoints.

Assessment Methodology:

Heureka's endpoint services were deployed in your environment along with the Heureka command console (cloud or on-premises). Once the indexing process completed a search for common file types was performed. For a detailed list of file types included, please see the Appendix.

Search results were analyzed to provide a "snap shot" overview of your data. Each endpoint automatically updates every 24 hours. As the unstructured data on endpoints changes, you would expect to see information in the command console, such as PII, change over time.

The endpoint service automatically classified PII residing on the endpoints and identified the following information: Social Security Numbers, Credit Cards (Visa, MasterCard, American Express, Diners Club, JCB, Discover) and Bank Routing numbers (ABA routing transit number).

Definitions

Endpoint	A computer or file share on which the Heureka endpoint service software is installed.
File	An artifact or file identified by name on the computer on which the Heureka endpoint service is installed.
Risk Score	A numeric score automatically calculated based on pattern matching by Heureka's endpoint service. Each match on a pattern =1 incremental point added to the risk score.
PII	Personally Identifiable Information This currently includes credit cards, social security numbers and bank routing numbers.
ROT	Redundant, Obsolete or Trivial Files that may be duplicative across multiple computers, obsolete (old or past data retentions), or trivial with zero or little business value.

Assessment Areas: The goal of this assessment is to provide you with specific and actionable information on areas of your data including:

1. File Age
2. Endpoint Risk
3. Duplicative Risk
4. Data Growth
5. Large Files
6. Potential Risk Cost

File Age

File Count by Category & Year

Old files represent potential risk or compliance issues. Larger volumes can lead to significant storage and backup costs. Identifying document volume by date is critical to understanding how to set or control data retention policies.

A recent survey found that 73% of users believe that half of their content probably has no business value, but they do not know which half or how to find the data.

Extension Category	Year of Doc Date												
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
3D Graphics/CAD Drawings	488	80	394	372	313	643	220	224	309	396	1,338	2,291	1,684
Adobe Acrobat Files	20,806	27,192	48,493	58,420	63,602	79,492	79,403	102,514	95,712	99,987	109,174	129,035	39,525
Audio Files	4,486	14,146	1,925	1,413	1,831	3,736	2,579	6,744	1,598	3,111	1,155	4,449	800
Batch Files	39	9	2	16	43	15	46	2	16	7	503	1,202	1,403
Bitmap Image File		1	4	57	116	146	124	120	128	113	146	181	60
Bitmap Image Files	85,329	91,895	143,558	122,729	212,718	164,013	160,258	70,379	113,728	89,418	98,526	103,131	58,885
Calendar Files	22	43	116	299	271	140	36	43	56	6	4	1	1
Compressed Files	1,341	1,373	4,251	3,952	3,426	3,418	2,666	2,364	3,870	3,277	5,147	8,951	6,105
Database Files	8,611	6,871	13,022	21,596	12,766	19,606	17,584	18,255	22,589	31,311	53,868	70,797	34,733
Document Files	4,106	3,761	5,436	5,062	8,162	9,134	9,908	10,788	12,433	10,607	10,489	10,175	4,516
Email Messages	5,939	9,469	15,639	21,821	10,961								
Font Files	256	588	412	665	450	562	661	721	299	331	403	196	153
Game Related Files	725	644	559	757	541	500	495	377	563	1,390	607	1,458	490
GIS-GPS-MAP Files		8	26	25	497	113	586	27		159	19	2	45
Graphics Files	1,260	3,229	2,619	2,206	2,303	3,786	1,693	1,885	1,936	2,317	2,498	1,750	541
Internet Related Files	84,081	66,867	146,351	90,152	141,622	195,767	65,841	68,742	66,992	83,784	66,461	85,118	106,988
Java Files	126	276	155	324	478	268	154	126	181	508	487	390	831
Load Files	92	32	33	11	79	10	11	3	6	2	4	214	836
Microsoft Excel Files	5,259	7,441	10,876	13,495	18,475	22,705	25,425	29,746	33,378	41,685	50,422	50,446	20,037
Microsoft Powerpoint Files	4,667	5,253	6,952	6,838	6,626	3,875	3,580	3,150	1,930	1,763	1,263	987	566
Microsoft System Files	41	4	4	2	876	2	5		5	1	5	39	171
Microsoft Windows Media Files	422	611	1,161	374	451	199	62	127	270	20	27	27	17
Microsoft Word Documents	106,336	127,742	161,204	161,629	172,348	143,415	113,884	116,071	116,328	118,903	118,324	117,228	44,893
Misc Windows Files	709	288	2,090	3,243	4,442	2,734	1,264	2,620	4,406	240	7,650	3,413	15,762
Mobile Phone Related Files	47	2	1	9	6	11	3	1	22	14	179	406	483
Program Executable Files	1,744	1,628	1,644	3,014	3,082	2,897	1,564	1,561	1,789	1,999	2,689	3,242	1,523
Real Media Files	68	38	2,433	1,971	504	122	6	1					
Spreadsheets	1,853	3,161	13,313	22,745	10,459	18,974	13,038	13,150	27,578	20,820	50,814	43,115	16,400
Temporary Files	673	4,982	1,972	2,820	1,565	3,743	666	772	418	265	313	475	110
Text Files	10,776	20,002	24,714	20,490	18,363	17,274	17,801	48,124	46,482	24,841	24,649	46,803	33,927
Vector Graphics Files	84	101	843	421	499	70	23	595	204	224	538	833	2,051
Video Files	958	486	821	1,832	886	1,128	1,327	738	892	855	690	1,615	4,614
Virtualization Software Files	169	3	18	1		4	2	3	240	44	286	39	25
Windows Registry Files	448	34	98	29	39	24	15	74	16	9	6	152	105

Figure 1 displays file count by the extension category and date

Heureka Date Information

Heureka's indexing engine extracts doc dates for email and loose files and displays those dates. For email files, the date shown is Sent Date. For loose files the date shown is last modified or if that is not available to the indexer, create date.

Heureka Recommendation

If a file retention policy is in place and does not include the years 2009 or older, you may have a retention policy issue. *Heureka recommends reviewing your data needs or corporate policy regarding file age and any data retention policy that may be in place.*

Endpoint Risk

PII Overview

Figure 2 shows your overall PII risk types. The majority of risk in your environment involves Social Security numbers followed by Bank Routing numbers. Risk scores are calculated by Heureka's auto-classification feature which does not deduplicate PII numbers. One positive hit = 1 point. Figure 3 shows the highest risk endpoint in the system.

According to Gartner, by 2018, regulatory disclosures that are related to a failure in the organizational information risk control environment will see a rise of 50 percent.
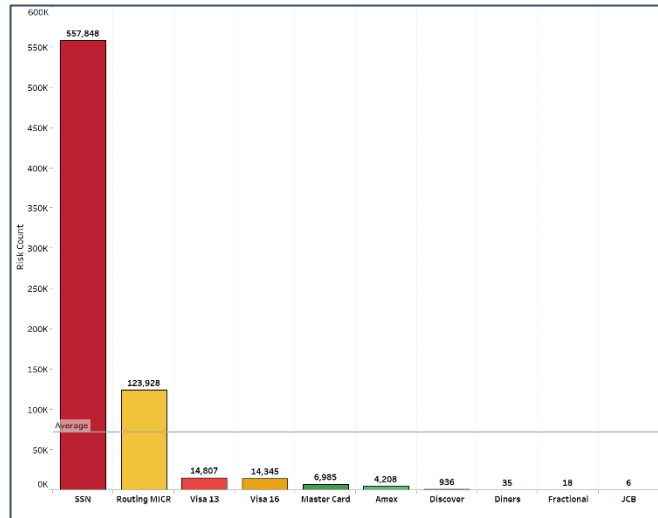

Figure 2 Displays types of risks
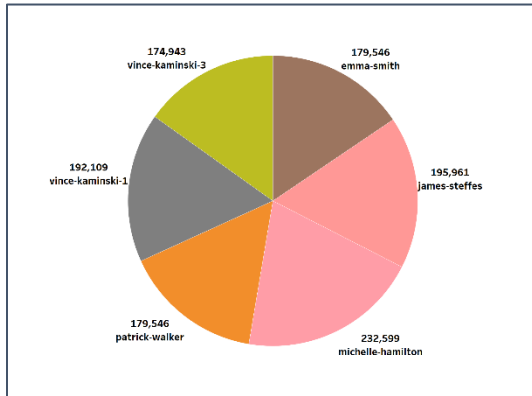
Figure 3 Displaying highest risk endpoint
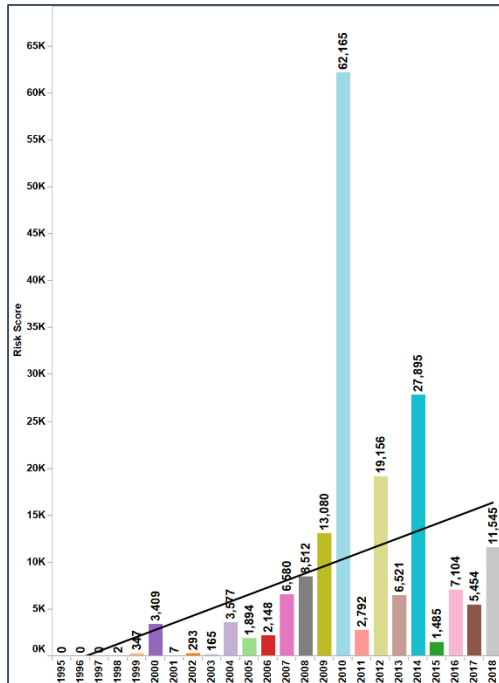
Figure 4 PII Growth over time

PII Growth Over Time

Figure 4 shows your data risk growth over time. Your risk score jumped significantly in 2018. *Heureka recommends running a PII focused search with a date filter for the year 2010 and newer. The search results will show all files in which risk was identified for 2010 to date. This allows you to further analyze and remediate your data risk at the file level.*

Heureka Recommendation

Figure 3 shows that "Folsom-PC" has the highest risk score and thus contains the largest amount of identified PII information. "Vince-Kaminski-3" contains the second-highest risk score. *Heureka recommends a file-level search for specific PII information on endpoints with high scores.*

Duplicate Risk

Amplifying your Risk

Duplicative file risk is a vital concern within an organization. Having high-risk duplicative files multiplies your risk profile and increases your targets for potential data theft. Figure 5 below shows duplicative files containing the highest risk score.

Duplicative Risk spread across an organization greatly increases your threat profile giving hackers access to sensitive data from more locations.

File Name	Computer Name	Duplicate Files	Avg. Risk Score
TinHistory.xls	vince-kaminski-2	2	112
EOLWTI_CC_simp.xls	vince-kaminski-3	3	52
EOLWTI_OC_simp.xls	vince-kaminski-3	4	52
Position_030200.xls	vince-kaminski-1	2	48
CapitalTemplate1.xls	vince-kaminski-4	2	30
Department.csv	Folsom-PC	2	29
EOLWTI_3.xls	vince-kaminski-1	2	22
	vince-kaminski-2	2	22
	vince-kaminski-3	2	22
EOLWTI_OpenClose.xls	vince-kaminski-1	2	22
	vince-kaminski-2	2	22
	vince-kaminski-3	2	22
VOD DSL.xls	vince-kaminski-3	2	17
EOLWTI_2.xls	vince-kaminski-1	4	7
	vince-kaminski-2	2	8
	vince-kaminski-3	4	7

Figure 5 Displays top 10 duplicate file names, computers containing duplicate, and number of dups. Data is sorted by highest risk score

Heureka Recommendation

Investigate duplicative file risk locations and decide if it is appropriate for each endpoint to contain the same file.

Heureka recommends eliminating or minimizing duplicate files where possible. Utilize Heureka's "Delete" quick filter search to help identify files that were moved or removed from an endpoint.

Unseen Risk

Files in Trash or Recycle Bin

It is common for users to move files to the trash or recycle bin but forgetting to empty them on a regular basis. In addition to retaining large amounts of information, files containing risk may also remain without deletion thus posing a potential risk to the organization. File risk is still present on the endpoint regardless of the location.
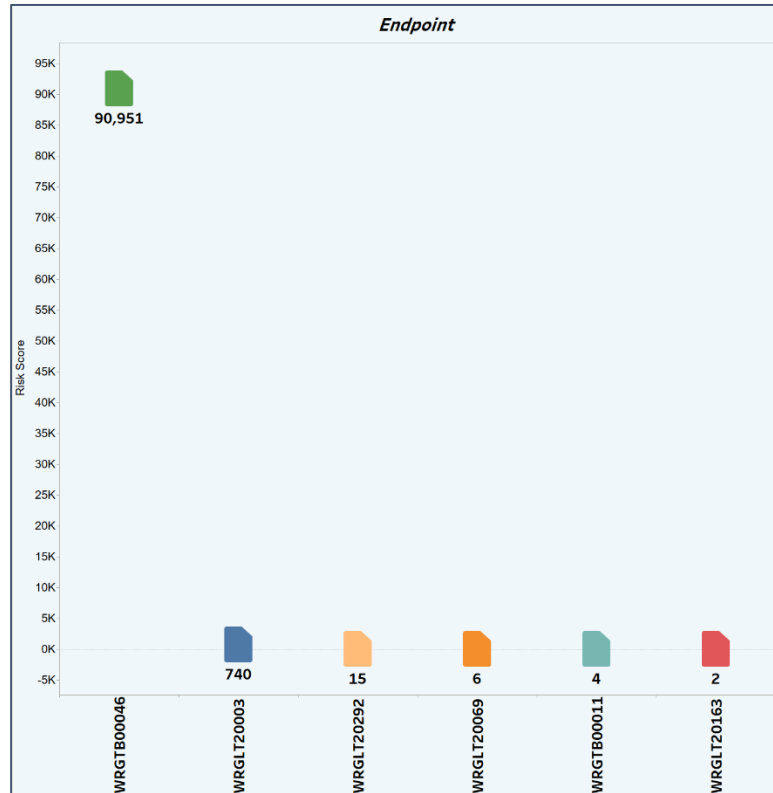
Figure 6 displays risk score of endpoints containing risk in recycle bin

Heureka Recommendation

Files that have not been completely deleted from an endpoint will continue to pose a potential risk within your environment.

It is recommended to check data retention policies and to educate endpoint users to routinely complete the emptying of trash/recycle bins. Macintosh users running MacOS Sierra have the option in Finder Preferences for removing items from the trash after 30 days. For PC users, it is recommended to set the recycle bin to remove files immediately vs moving the files to the bin. These options will help assure that build-up of potentially risky file data will not occur.

Redundant Files

Up to 69% of data in most organization is considered ROT. Duplicated files can be a burden for both storage cost as well as risk compliance.

Duplicative Files

Duplicative files fall into the category of ROT (Redundant-Trivial-Obsolete). These files may not have business value but can represent potential additional costs in storage or backup. When redundant files spread across an enterprise it greatly increases your risk profile and make for a wider attack surface should a breach occur.
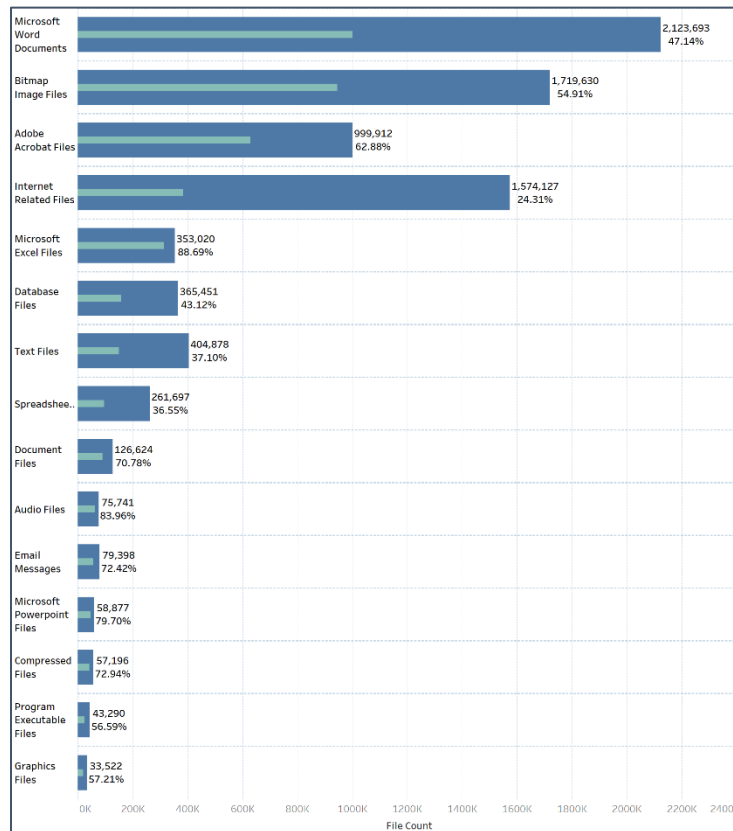
Figure 7 Showing highest duplicative file categories with total of unique and duplicative files

Heureka Recommendation

Investigate duplicative file locations and decide whether it is appropriate that each endpoint contains duplicative information.

Heureka recommends eliminating or minimizing duplicate files where possible. Utilize Heureka's "Delete" quick filter search to help identify files that were moved or removed from an endpoint.

Data Growth

Growing Data

Many organizations experience explosive electronic data growth over time. Monitoring, understanding, and controlling internal and cloud-based data growth is of major importance to an organization.

The chart to the right displays your data growth over time. It is categorized by total file count including email and loose files.

In many organizations data growth is completely natural over the normal course of business; however it is important to keep track of growth over time and to establish policies and procedures for dealing with data both internally and in the cloud.

80% of data is unstructured and grows by more than 60% each year.
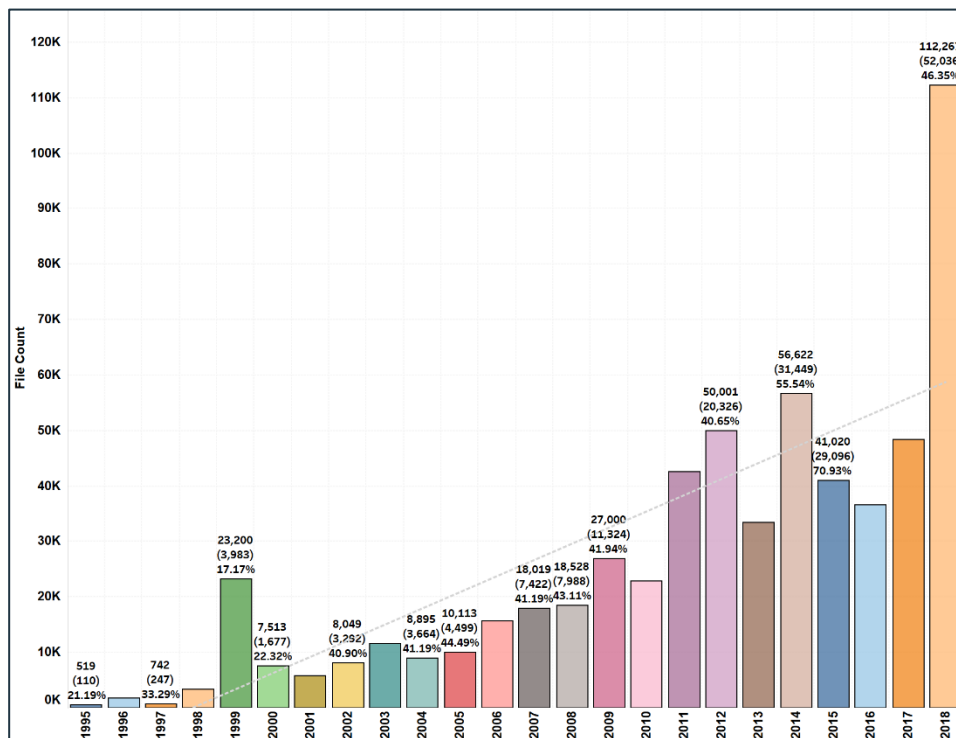
Figure 8 Shows total file count to date

Heureka Recommendation

Your data growth has grown steadily since 2006 and analyzing the age of the data suggests that you may want to evaluate or consider a data retention policy to help reduce the volume of old data.

Heureka recommends establishing or reviewing these results against any data retention policy to see if older files fall within the policy timeframe. Consider archiving files with long-term retention requirements and removing obsolete data.

Large Files

Large File Sizes

Large files represent a potential risk as they may contain a large amount of sensitive information. Larger files require larger storage and backup systems and result in additional organizational cost.

Certain file types such as email containers do not necessarily pose larger risks. However it is important to periodically scan for larger files to reduce the risk potential with retaining a high volume of sensitive information in one file.

Figure 9 displays your top files over 100MB in size.
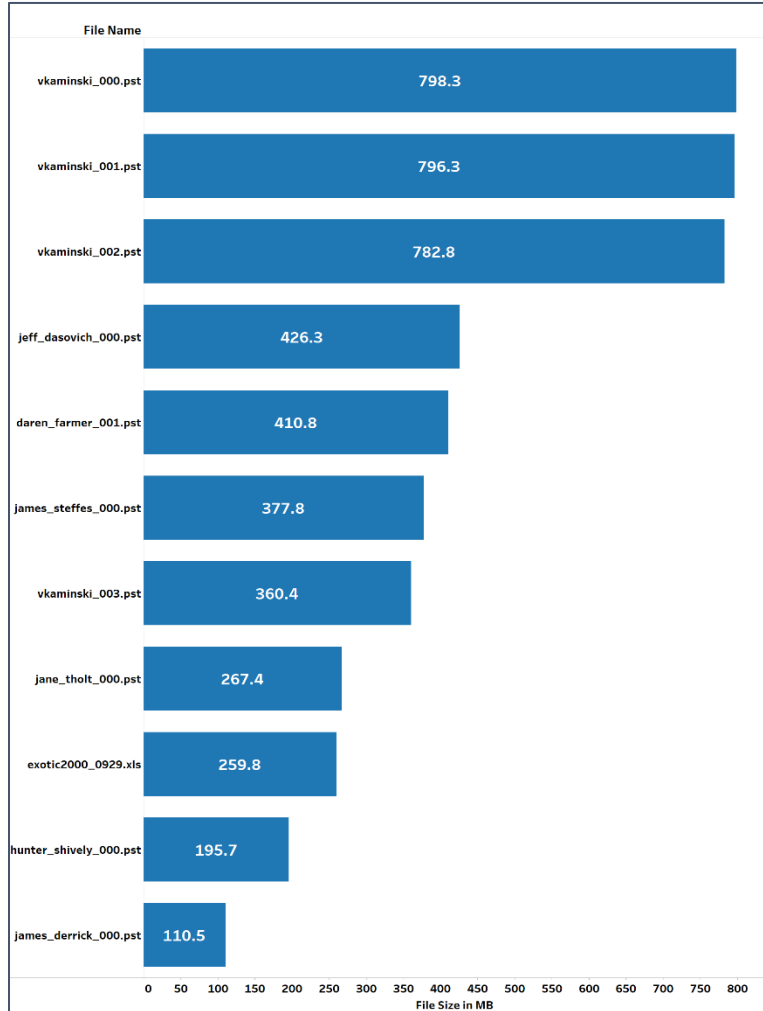
Figure 9 displays files larger than 100MB in size

Heureka Recommendation

Large file sizes may contain large amounts of information and thus may pose more risk or increase storage costs.

Your larger files consist mainly of PST (Microsoft Email) files which may not pose a significant risk. However, there is a single XLS spreadsheet which is 259.8MB in size. Heureka recommends researching this or other large files further by performing a File Name search and reviewing the returned files for risk potential.

![Heureka Software logo]

Potential Risk Cost

Risk Cost

Risk cost is one of the key factors when assessing file-level risk data. Although the overall cost can vary widely, your data can be broken down by industry and potential risk categories by performing targeted searches. This assessment applies industry standard risk cost based on the risk factors that we currently identify. To give an approximate cost, we have used the Ponemon/IBM Risk Cost analysis cost per record calculation show in the table to the right.

INDUSTRY	COST PER RECORD
HEALTH CARE	$408
FINANCIAL	$206
SERVICES	$181
PHARMACEUTICALS	$174
TECHNOLOGY	$170
ENERGY	$167
EDUCATION	$166
INDUSTRIAL	$152
ENTERTAINMENT	$145
CONSUMER	$140
MEDIA	$134
TRANSPORTATION	$128
COMMUNICATION	$128
HOSPITALITY	$120
RETAIL	$116
RESEARCH	$92
PUBLIC	$75

Your Industry: Government (Average Value Used)

Files containing Social Security Numbers: 2508

Files containing Credit Card information: 1122

Files containing Bank Acct information: 319

All Risk (excludes some duplicates):   3906

Potential Risk Cost: $578,088

Heureka Recommendation

Understanding the location and potential cost of risky information is important. Your assessment shows a significant number (85%) of scanned share folders containing PII information. This may be acceptable for your organization, but it demonstrates a wide potential breach target surface.

Heureka suggests further analysis and potential consolidation of PII related information. Additionally, encryption should be considered on extremely sensitive information on all computers. Use Heureka's collection/quarantine/delete functions to control your sensitive data availability and locations.

Conclusion

RISK

The primary objective of this report was to identify risk in your environment. Risk or potential risk on endpoints broadens the overall attack surface for your organization and offers a rich target for malicious threats or hackers. Exposing type and location of potentially risky information is of vital importance when protecting your enterprise. Visibility into the unstructured data on your endpoint helps uncover risks which may otherwise go unnoticed. Heureka allows you to pinpoint your risk down to file level with unstructured data at rest.

Pages five and six of the assessment clearly demonstrate the type of risk on your endpoints along with duplicative risk. Page ten demonstrates the potential cost (including remediation) to your organization if files were stolen or ransomware was deployed in your environment.

As stated on the individual pages, Heureka strongly recommends file-level PII searching on your highest risk computers followed by analysis of whether risk-oriented data should be encrypted, archived, transferred or deleted. Duplicative risk files should also be analyzed across all endpoints to determine if they should remain in place or be remediated.

ROT

The second goal of this assessment was to help identify two of the three components of ROT (Redundant, Obsolete, Trivial). The "trivial" component of ROT can be located using Heureka's robust endpoint search capability. Data which is considered trivial is contextual to your business. Heureka can assist you in understanding how to leverage the platform to identify your trivial data. However, this is outside the scope of this assessment report.

Page seven of the assessment demonstrates redundant files and of a larger concern, redundant files that contain risk. In some environments, it is normal for multiple owners (custodians) to possess duplicate files and even files that contain risk. This report highlights some of these files and Heureka recommends performing a more detailed analysis on the file-level for redundant risk. Where appropriate you should consider encrypting, archiving transferring or deleting redundant files.

Page four focuses on file age and helps demonstrate the obsolete portion of ROT. If there is no file retention policy in place, Heureka recommends a determination of how long files or emails remain on endpoints and remediation of any data outside of this determination.

COST

The potential cost of a data breach either by malicious outsiders or an internal employee is a complex calculation. Not all data has the same associated costs and there is a wide variation by industry type. This assessment leverages Heureka's PII classification engine to identify instances of risk, by risk type, to provide a summary level view of your potential breach cost for the subset of endpoints assessed. It is important to run a more comprehensive file-level analysis for a deeper understanding of your risk posture, across your enterprise.

This report focuses on the potential cost of PII information only including credit cards and social security numbers. As shown in the recommendation, eighty-five percent of the endpoints assessed contain this sensitive information. In some organizations, this is acceptable. However, Heureka strongly recommends further analysis across all endpoints for PII information followed by encrypting or deleting files containing risk on endpoints.

PII information alone may only account for a fraction of overall risk within your company. For example, this assessment did not include items such as personal health information, trade secrets, contracts, or sensitive information related to your organization. The Heureka platform excels in its ability to perform more targeted searches for specific information and should be leveraged to perform a deeper assessment of the data on your endpoints.

Appendix - Common File Types

The file types listed below were used when creating this report. Some file types such as multimedia files return metadata only versus full text and metadata thus factoring into size and date analytics.

Description	Kind	Extensions
3D Graphics/CAD	Drawings	3D, 3ds, c4d, dgn, dwfx, dxf, dwg, igs, plt, stp, lcf, max, skp, stl, u3d
Adobe Acrobat	Documents	pdf
Adobe Fonts	Fonts	bdf
Archive Files	Containers	mso, iso, cfs
Audio Files	Multimedia	3ga, aac, aiff, amr, ape, arf, asf, asx, cda, dvf, flac, gp3, gp4, gp5, gpx, logic, m4, m4a, m4b, m4p, m4r, m4v, midi, mp3, ogg, pcm, rec, snd, uax, wav, wma, rom
AWK Script	Scripts	awk
Batch Files	Scripts	cmd
Binary Files	Logs	dmp
Bitmap Files	Images	bmp, cpt, dds, dib, dng, dt2, emf, ico, icon, jpg, jpeg, pcx, pic, png, psd, raw, tga, thm, thmx, tif, tiff, wbmp, bdp, webp
Calendar Files	Calendars	ics
Compressed Files	Containers	001, 7z, 7zip, a00, a01, a02, a03, a04, a05, ace, air1, apk, appx, arc, arj, asec, bar, bin, c00, c01, cab, cbr, cbz, cso, deb, dlc, gz, gzip, hgx, inv, ipa, isz, jar, msu, nbh, pak, rar, prm, sis, sisx, sit, sitx, sitd, tar, tar.gx, tgx, uax, webarchive, xap, z01, z02, zip, zipx, zipp
Database Files	Databases	accdb, accdt, db, dbf, fdb, gdb, idx, mdb, dcb, mdbackup, mdf, sdf, sql, dat, sqlite, val, wdb
Document Files	Documents	abw, aww, chm, cnt, dbx, djvu, dot, docm, dotm, epub, gp4, ind, indd, key, keynote, mht, mpp, odf, ott, oxps, pages, pmd, pot, potx, pps, ppsx, pptm, pptx, prn, prproj, ps, pub, pwl, rep rtf, sdd, sdw, shs, snp, sxw, tpl, wimp, wpd, wps, wrl, xps
Digital Camera RAW	Images	arw, cr2, crw, dcr, dng, fpx, mrw, nef, ord, pcd, ptx, raf, raw, rw2
E-Book Files	E-Books	azw, azw3, cbr, cbz, epub, fb2, iba, ibooks, lit, mobi
Email Files	Email Messages	eml, emlx, msg
Encrypted Files	Encrypted	asec, crypt, ksd, pfx, pkpass, rem, tc
Financial Files	Financial	gdb, ofx, qif
Font Files	Fonts	eot, otf, ttc, ttf, woff

Description	Kind	Extensions
Game Related Files	Games	big, bik, cab, dat, dds, hi, lng, nes, pak, res, sav, save, rc2replay, sc2save, scn, scx, uax, wotreplay
GIS-GPS-MAP Files	Others	gpx, kml, kmz, map
Graphics Files	Graphics	abr, ai, ani, cdt, kjvu, eps, fla, fm, icns, ico, icon, info, mdi, odg, pic, prn, psb, psd, px, pzl, sup, vsd, vsdx, xmp
Internet Related Files	Internet	ashx, asp, aspx, atom, bc, bc!, class, crdownload, css, dlc, dtd, download, flv, gdoc, gen, gif, gsheet, gslides, htm, html, iap_xml, js, json, jsp, manifest, mf, mht, mhtml, opml, part, partial, php, png, rss, swf, torrent, url, webm, webp, xap, xhtml, xml, xul, xsd, xsl,
Java Files	Java	properties, jad, jks, jar
Image Load Files	Legal Image Loads	opt, lfp
Lotus Files	Lotus	nsf, wk1
Linux Files	Others	amx
Microsoft Access	Database	accdb, mdb
Microsoft Encrypted	Encrypted	nuspec, pfile
Microsoft Email Containers	Email Container	pst, ost
Microsoft Excel	Spreadsheets	xla, xls, xlsx, xlsm, xll, xlt, xlw
Microsoft Index Crawl	Container	crwl
Microsoft PowerPoint	Presentations	ppt, pptx
Microsoft System Files	System Files	01, 1, 2, 3, acm, ax, blf, c, cch, cdf-ms, cur, ignore, ini, lck, psmdcp, n01, n02, rdf
Microsoft Media Files	Multimedia	wmv
Microsoft Word Documents	Documents	doc, docx, dot
Microsoft vCard	Contacts	vcf
Microsoft Visio Files	Drawings	vsd
Mobile Phone Related Files	Mobile	apk, bbb, ipa, ipd, apk, ipsw, mdbackup, nbh, nbu, npf, rsc, sbf, sis, sisx, thm, xap, safariextz, xpi
Multimedia (video)	Multimedia	mpe
Program Executable Files	Executable	air, app, crx, application, appx, bat, bin, com, cpl, deb, dll, elf, exe, jar, js, lnk, msi, prg, rpm, shs, vbs, xap
Real Media Files	Multimedia	rm
Symantec License Files	License Files	slf
Spreadsheets	Spreadsheets	csv, cwss

Description	Kind	Extensions
Summation Load Files	Legal Load Files	dii
Temporary Files	Temp	tmp
Text Files	Text	rtf, txt, log, wpd
Thumbnail Image files	Images	thm
Vector Graphics Files	Graphics	ai, cdr, csh, csl, drw, emz, odg, pic, sda, svg, swf, wmf
Video Files	Multimedia	264, 3gs, 3gp, arf, asf, asx, avi, bik, dash, dvr, flv, h264, m2t, m2ts, m4, m4a, m4b, m4p, m4r, m4v, mkv, mod, mov, mp3, mp4, mpg, mpeg, mswmm, mts, ogv, prproj, rec, rmvb, swf, tod, tp, ts, vob, webm, wmv
Virtualization Software Files	Virtualization	ova, ovf, pvm, vdi, vhd, vmdk, vmem, vmwarevm, vmx
Windows Registry Files	Registry	reg
Windows Open Office Files	Open Office	rels
Windows Powershell Scripts	Scripts	ps1, psm1
Wordperfect for Windows	Documents	chk, w51

Heureka
SOFTWARE®