

VERITAS™

DATA GENOMICS INDEX

A report on the true makeup of storage environments from
the experts in unstructured data.

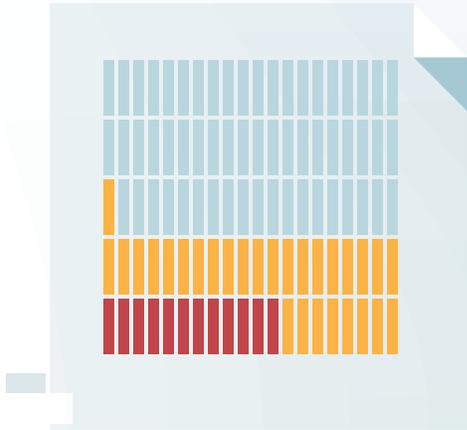
2016



The Data Genomics Project is an initiative designed to change the way we think about managing data. Veritas founded the initiative to bring a community of like-minded data scientists, industry experts, and thought leaders together with the purpose of better understanding the true nature of the unstructured data that we are creating, storing, and managing on a daily basis. Our first contribution to the project is this inaugural benchmark report on real storage environments' composition—the Data Genomics Index.

In working with 86% of the Fortune 500 and backing up, archiving, or analyzing exabytes of data for these customers, Veritas is in a unique position to glean the defining characteristics of an organization's environment. Today, the characteristic we focus on is metadata, and in leveraging this metadata aggregated from customers using our file analysis products, Veritas is able to surface accurate details on what real environments consist of.

Today's Data Environment



41%

of the total environment hasn't been modified in the past

3 years



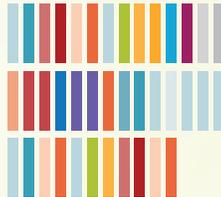
The top file types are stranger than predicted.



Traditional office file types are overly taxing to your environment.



Orphaned information



Orphaned information is disproportionately overweight and extra stale.

System archiving these reduces storage costs by



50%

Organizations should prioritize their deletion or archive strategy to gain inordinate returns on storage space.



The Inaugural Veritas Data Genomics Index

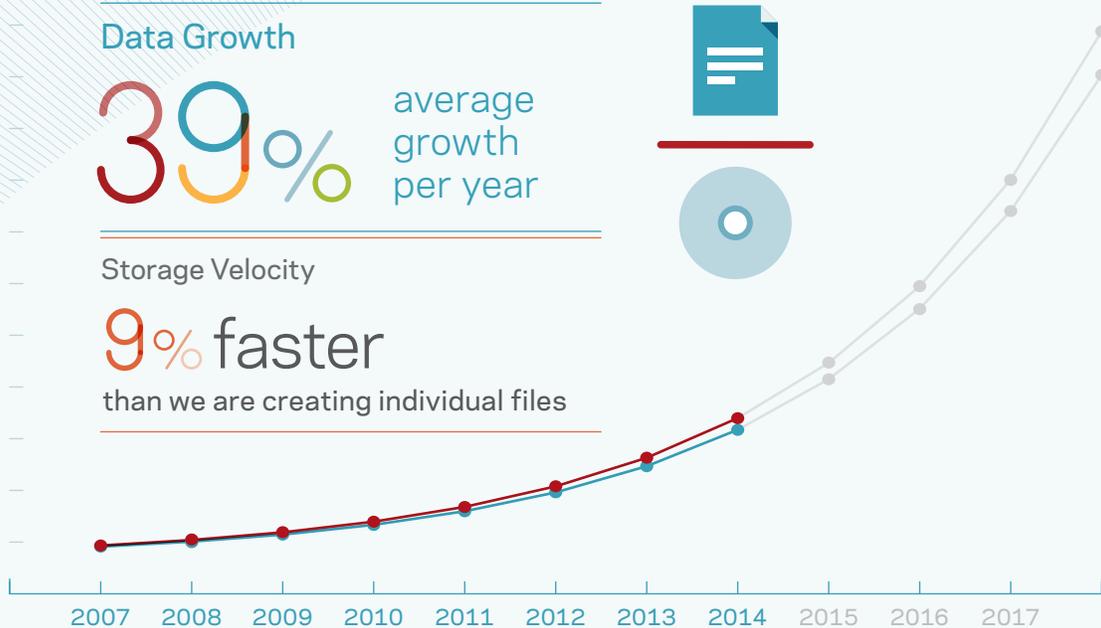
Veritas analyzed tens of billions of files and their attributes from many of our customers' unstructured data environments in 2015 to gain a better understanding of what their environments really consist of. Over 8,000 of the most popular file type extensions were considered in the analysis. Generally, this data is a representative subset of the entire file system environment of a respective customer.

Data Growth

39% average growth per year

Storage Velocity

9% faster than we are creating individual files



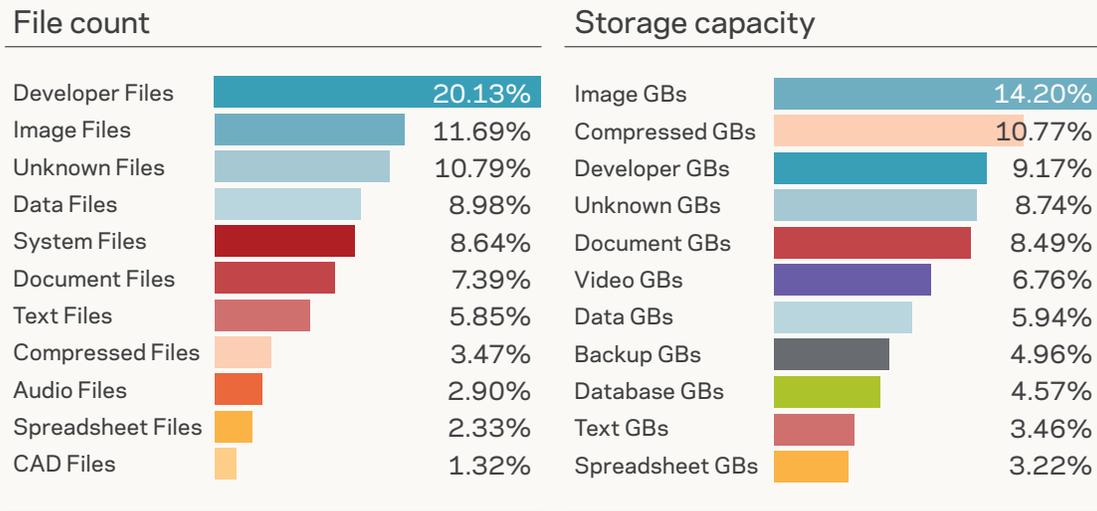
Data is exploding.

The real speed of data growth at the file level over the past seven years is 39.2481189% year-over-year. This storage capacity requirement is growing 9% faster than we are creating individual files, so while behavioral change could certainly help curb some growth, this is definitely a storage management problem.

Curbing storage capacity isn't just a storage problem. The storage environment is cluttered, where the average PB of information contains 2,312,000,000 files.

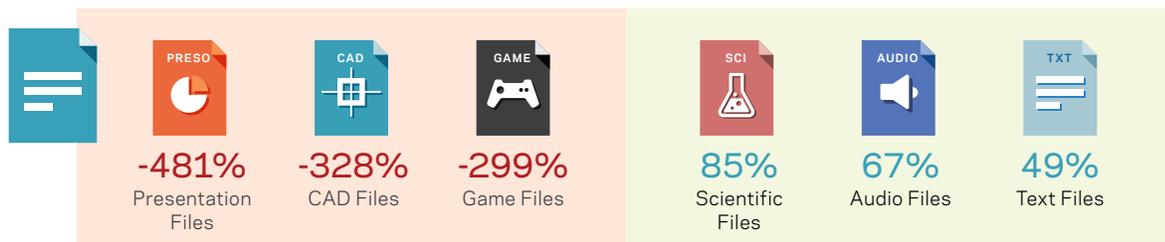
Worst offenders: Images & developer files.

Looking at the total population of data, we see striking differences between the file types that represent the highest count versus those that take up the most space. You can think of this as the clutter versus the cost of the environment.

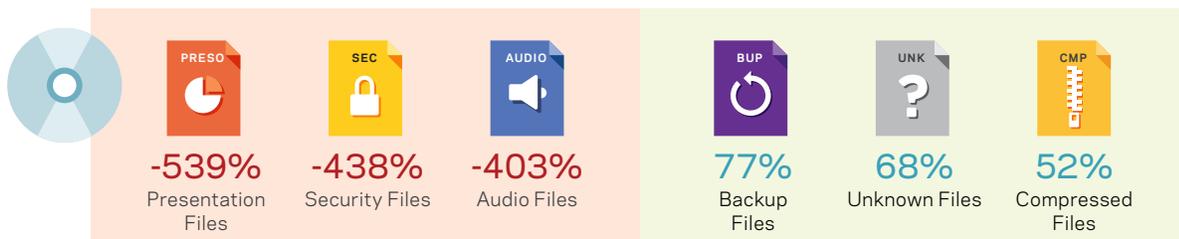


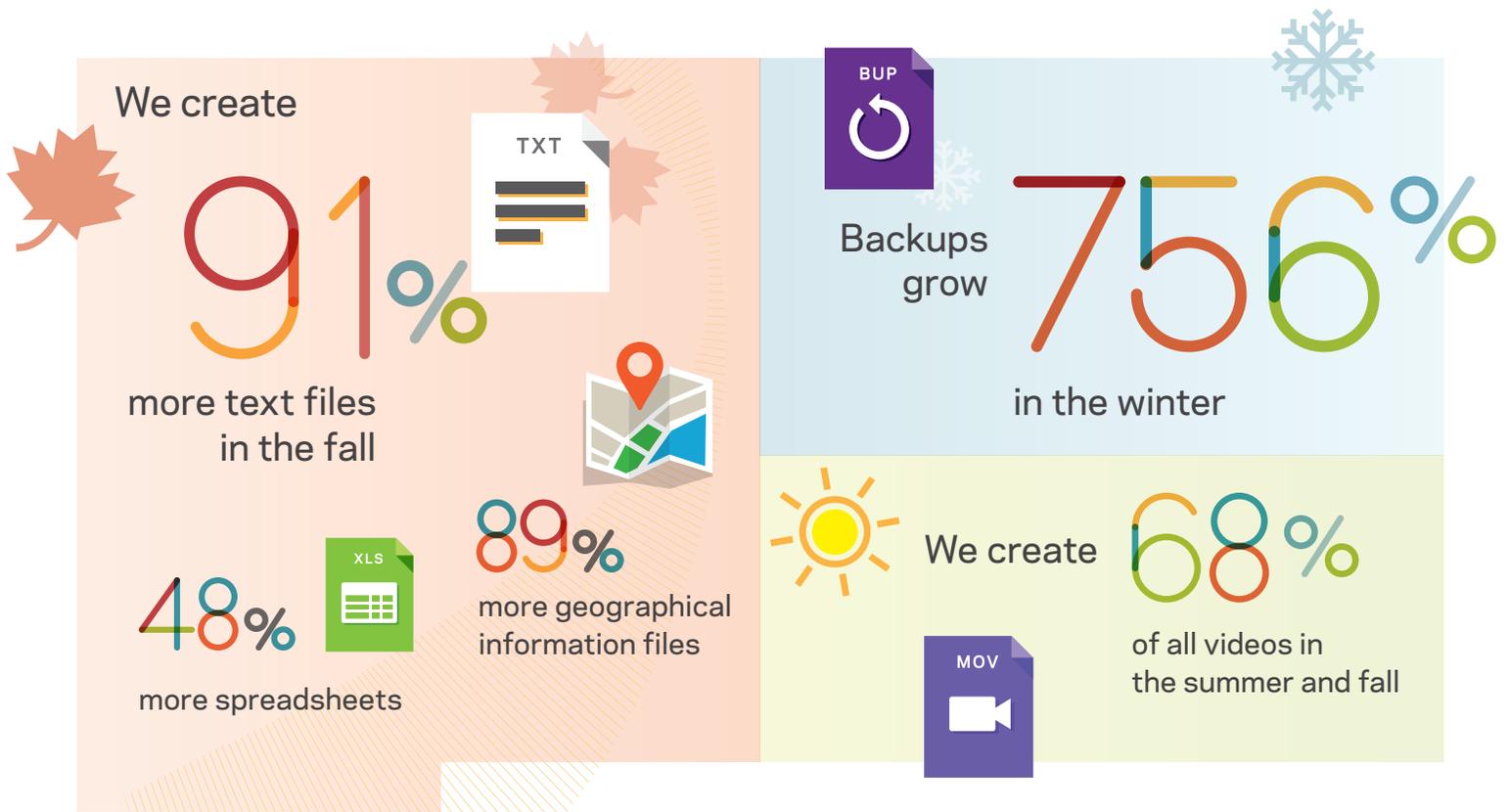
Data's biggest movers and shakers.

This composition has changed over time. The biggest movers relative to other file types over the past ten years are:



The biggest movers in terms of storage capacity are:





Data growth. Always in season.

Fall is the biggest month in terms of pure, per file creation. We create 91% more text files in fall than other seasons, 48% more spreadsheets, and 89% more geographic and information system files.

Backups and documents are the only two file types to increase in numbers from fall to winter. Backups grow 756% in terms of size with all the annual backup practices. 68% of all videos for the year are shot in summer and fall, and images fall off a winter cliff, declining 63%. Email (pst) culture is so predictable, however, that standard deviation between seasons is a mere 0.7%.

less than
3 years
old

at least
3 years
old

2015
2014
2013
2012
2011
2010
2009
2008
2007

more than
7 years
old

41%
of the total environment
has not been modified in the past

3 years

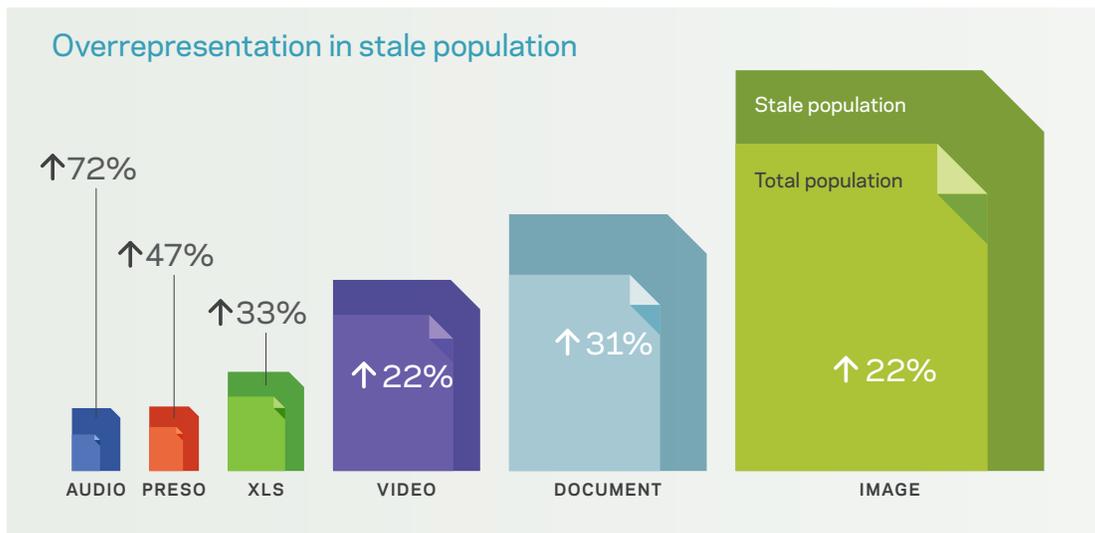
12% = ancient
(no modification in 7 years)

Why are you keeping *that*?

Information is everything when it comes to today's businesses, but it is created at such an overwhelming rate that the usefulness of an individual piece is rather fleeting.

Remediation: File types in the crosshairs.

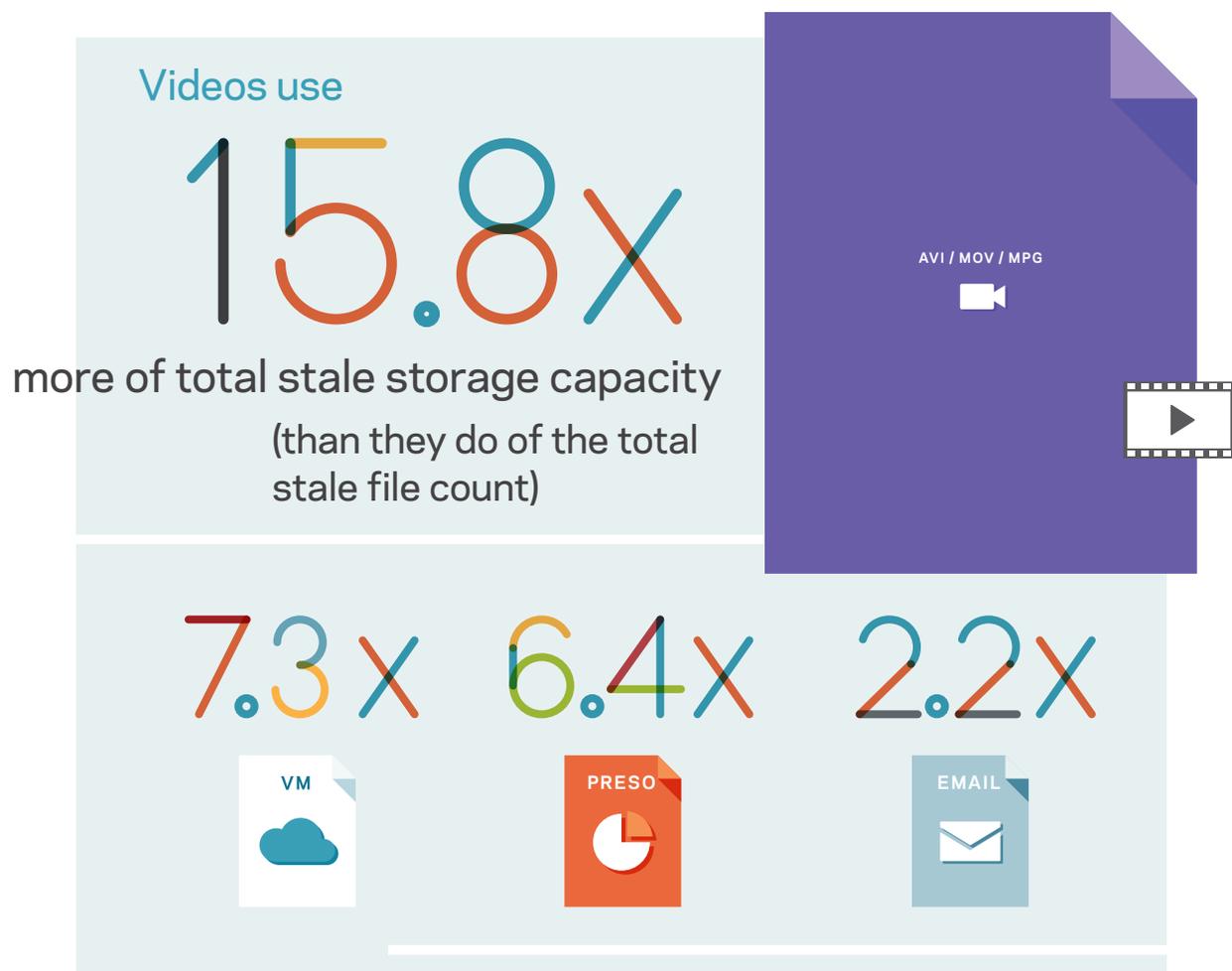
When faced with an overwhelming amount of stale data and potential remediation decisions, it helps to prioritize where your information management 'decision dollars' are best spent. Looking at what types are overrepresented in the stale data vs. total data, the traditional "office" files are a huge burden:



If you want to look for where remediation of individual files translates into the best storage space return, focus on these five formats that give you the best GB return per file:

1. Virtual Machine File Types
2. Security File Types
3. Gaming File Types
4. Scientific File Types
5. Geographic Information System File Types





Files out of proportion.

If you are willing to prioritize specific file types, look to where the number of files and percentage of the total are out of proportion. Videos, for instance, take up 15.8x more of the total stale storage capacity than they do of the total stale file count. Virtual machine files take up 7.3x more space, with presentations at 6.4x and emails at 2.2x, rounding out the best choices for file type prioritization.

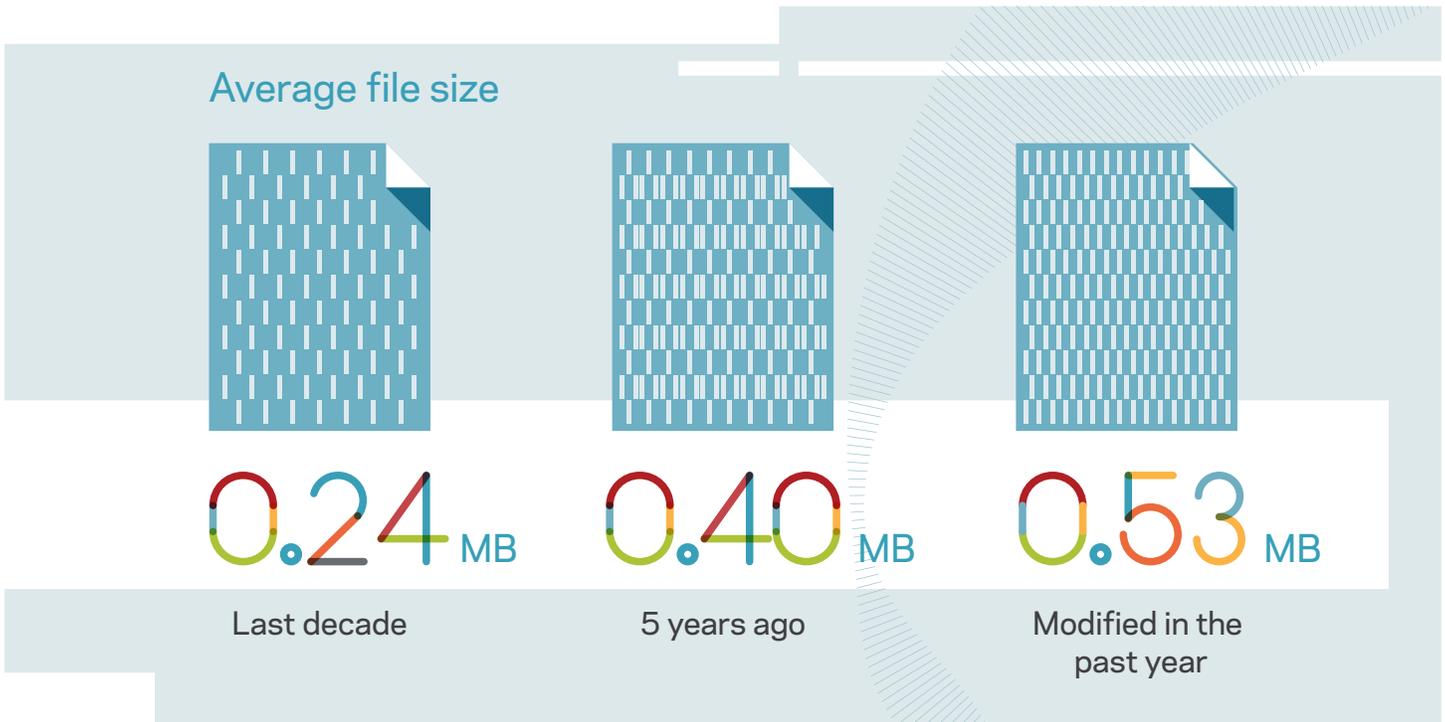


They cleaned out their desks. And that's about it.

Orphaned data is data without an associated owner. With employee turnover, role switches, and general active directory chaos, it is easy to see how the heritage of the environment is difficult to track, and that can cost organizations.

One way the cost manifests itself is in orphaned data using up a disproportionate amount of storage capacity. While orphaned data is a mere 1.6% of the total file population, it's 5.1% of the total storage capacity. Orphaned data is also disproportionately skewed towards content rich data types, with images taking up 88% more space than normal and videos and presentations at 165% and 229%, respectively.

We can also see employment tendencies potentially having an effect on storage environments. Orphan files are 222% larger than the average file. Managers may have believed that the larger the file, the more important its contents, and subsequently only kept the dense items when employees went on their way. If you want to recover storage space, focusing on content without an owner is a good place to start.



Density is a slim indicator of usefulness.

It's no surprise that we are creating denser content today, but it may surprise you that over the past 7 years, it has been a relatively slow 10.3% increase. The average size of a file:

Used last in the last decade or more:	0.24MB
Used in last 5 years:	0.40MB
Modified in the past year:	0.53MB

Files that are classified as stale are **33% smaller** than the files that have been modified in the past year.

Okay, now what do we do?

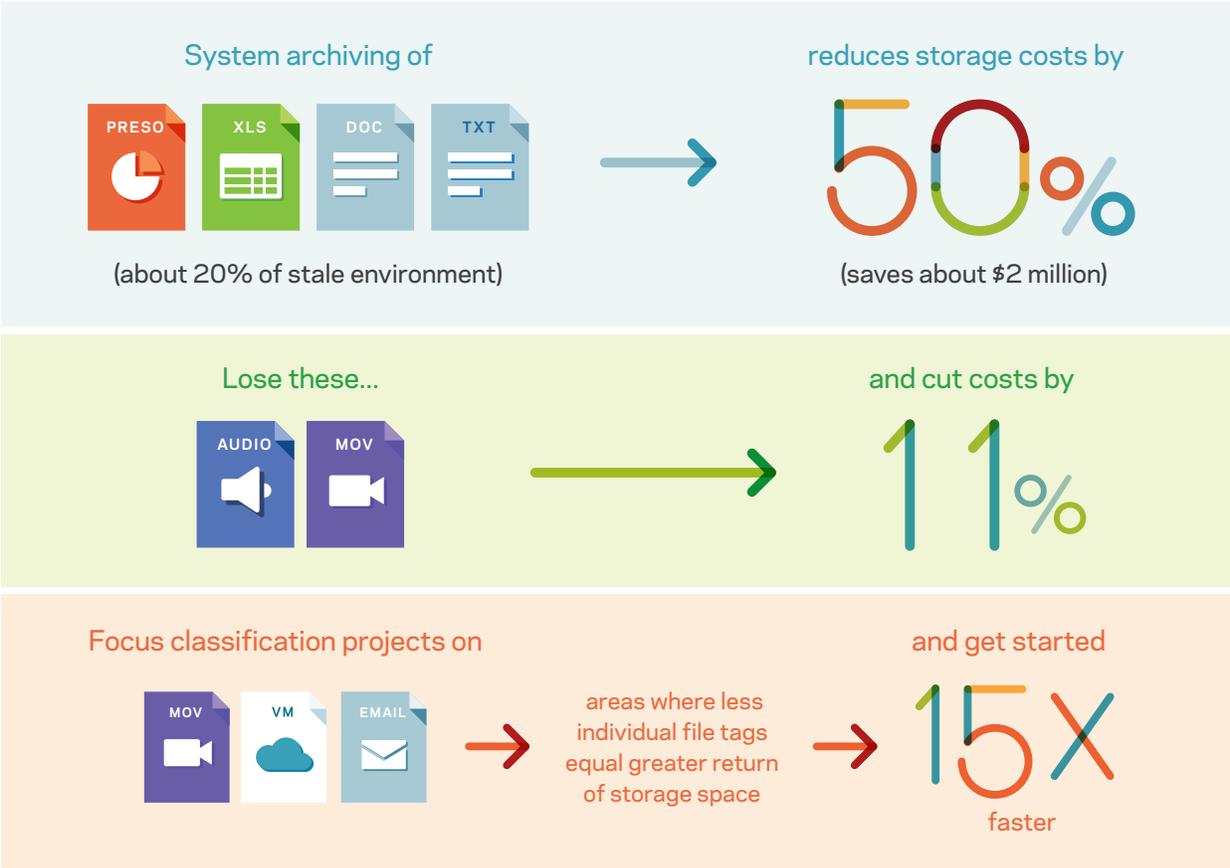
If your storage environment looks similar to the environments we analyzed here, then you have tremendous opportunity. Imagine the average 10PB environment...

41% = \$20.5 million
stale environment spending

If 41% of the environment is stale, you could be spending as much as \$20.5 million per year to manage data that hasn't been touched in three years. But cleaning it up is tough. That 4.1PB equates to 9,479,200,000 individual file decisions to classify, delete, or archive.

9,479,200,000
individual files

You have to prioritize...



Content rich files like presentations, spreadsheets, documents, and text files are 20% of the average stale environment, and a great target for file system archiving projects that can reduce storage costs by 50% or more—a return of over \$2 million. Audio and video alone could return 11%. Images dominate 18% of storage space in the ancient, seven-year or older category.

Focusing classification projects on areas where less individual file tags equal greater return of storage space, like videos, VM files, and emails, can get you started fast—15x as fast. Evaluating policies for data left behind from employee departures or role changes can get you 5%, or even a cool million...

Regardless of your situation, with insights like these, there are many opportunities to fight back against the tremendous growth curve and take back your environment.

VERITAS™

Veritas Technologies enables organizations to harness their information, bolstering business success in even the most complex environments. We serve organizations of all sizes, including 86 percent of global Fortune 500 companies. In fact, for over a decade, Veritas has been recognized as a leader in the Gartner Magic Quadrant for both Enterprise Backup Software and Integrated Appliances¹ and Enterprise Information Archiving.²

Together with our experienced partner community, we help our customers improve their data availability and unlock insights to make them more competitive. With over 7,800 employees in 58 countries around the world, Veritas is a \$2.5 billion company that partners with the largest technology leaders, including Amazon, Cisco, Fujitsu, Google, Hitachi, HP, IBM, Microsoft, NetApp, OpenStack, Symantec, and many more.

From traditional data centers to private, public, and hybrid clouds, Veritas, together with our partner community, helps enterprises—regardless of their environment—protect, identify, and manage data using intelligent information management solutions. With Veritas, enterprises have the insight and availability they need to understand what information they have, know how to keep it protected, and realize what they should delete.

To learn how we can do the same for you, visit us at veritas.com

Legal:

¹Source: Gartner, Inc., Magic Quadrant for Enterprise Backup Software and Integrated Appliances, Dave Russell, Pushan Rinnen, Robert Rhame, June 15, 2015

²Source: Gartner, Inc., Magic Quadrant for Enterprise Information Archiving, Alan Dayley, Garth Landers, Anthony Kros, October 29, 2015. Note, from 2005-2009 the research was called Magic Quadrant for Email Active Archiving.

Gartner does not endorse any vendor, product, or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

© 2016 Veritas Technologies LLC. All rights reserved. Veritas and the Veritas Logo are trademarks or registered trademarks of Veritas Technologies LLC or its affiliates in the U.S. and other countries.