Efficient Management of Unstructured Data in Big Data Environment

T.Jhansirani, D.Satyanarayana
Kurnool. A.P,INDIA. satyadotcom@gmail.com

***Abstract*— Unstructured data is commonly referred to as information that does not have a predefined data model or is not organized in a predefined way. The world creates 2.5 quintillion bytes of data per day from unstructured data sources like sensors, social media posts and digital photos, resulting in irregularities that make this type of data more difficult to understand compared with data stored in a fielded form in databases. Unstructured data accounts for at least half of all data stored by organizations, and with the growth of social-media sites and the use of online video and images for business applications, the volume of unstructured data is likely to grow significantly. This paper exposes storage issues, real cost of storing and file archiving solution for efficiently managing unstructured data in Big data environment.**

***Keywords*— Big data, unstructured, sensors, digital photos.**

I. INTRODUCTION

Unstructured data is data that does not follow a specified format for big data. If 20 percent of the data available to enterprises is structured data, the other 80 percent is unstructured. Unstructured data is really most of the data that you will encounter. Until recently, however, the technology didn't really support doing much with it except storing it or analyzing it manually.

Sources of unstructured big data

Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine generated or human generated.Here are some examples of machine-generated unstructured data:

- **Satellite images:** This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
- **Scientific data:** This includes seismic imagery, atmospheric data, and high energy physics.
- **Photographs and video:** This includes security, surveillance, and traffic video.
- **Radar or sonar data:** This includes vehicular, meteorological, and oceanographic seismic profiles.
-

The following list shows a few examples of human-generated unstructured data:

- **Text internal to your company:** Think of all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today.
- **Social media data:** This data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr.
- **Mobile data:** This includes data such as text messages and location information.
- **website content:** This comes from any site delivering unstructured content, like YouTube, Flickr, or Instagram.

Some people believe that the term unstructured data is misleading because each document may contain its own specific structure or formatting based on the software that created it. However, what is internal to the document is truly unstructured.

By far, unstructured data is the largest piece of the data equation, and the use cases for unstructured data are rapidly expanding. On the text side alone, text analytics can be used to analyze unstructured text and to extract relevant data and transform that data into structured information that can be used in various ways.

For example, a popular big data use case is social media analytics for use with high-volume customer conversations. In addition, unstructured data from call center notes, e-mails, written comments in a survey, and other documents is analyzed to understand customer behavior. This can be combined with social media from tens of millions of sources to understand the customer experience.

Key Issues for Storage Today

1. The bottom line on storage cost is increasing:
Storage is typically in the top 3 spends for IT infrastructure. As a result, it is often a target for cost cutting. IT departments find it challenging to address mounting pressures to lower the total cost of ownership (TCO) of their storage. Data growth and the increasing importance of storage to enterprises are chief contributors to these challenges.
2. Storage growth can overwhelm the capabilities of backup systems:

As data proliferates and IT adds capacity, backup devices often cannot collect all data within the available backup window. This failure to achieve full protection can put service level agreements (SLAs) regarding RPOs at risk, result in lost data sets, and add weeks to a critical project. Unless an organization invests in a massive upgrade to its backup system, its choices are to stagger backups, extend the backup window, or risk a backup failure. Extending backup time or staggering backups risks availability issues and subsequent business

disruption. On the other hand, if you continually add capacity, you'll have no choice but to continually add backup resources.
3. Storage growth slows disaster recovery:
Without a way to easily locate data and assess its value, a company is storing inactive data on primary disks. In the event of an outage, the company must restore massive amounts of content that may not all be mission critical. As a result, RTOs of an SLA may be at risk, which is a serious threat to business continuity.
4. Unstructured data complicates regulatory compliance and legal discovery.
Corporate and government regulations often require IT to store data for many years, to provide particular data on request, and to prove the authenticity of data. Unstructured data in particular can be difficult to categorize and locate for archive or retrieval. When terabytes of unstructured data are involved, manual methods of searching can cripple an organization.
Let's take a look at the real cost to an organization of maintaining data the old way.

The Real Cost of Unstructured Data
In a July 2005 Scientific American magazine article, Matt Kryder, then chief technology officer of Seagate Technologies, noted that storage was outpacing the processor and memory innovation noted in Moore's Law1. "Kryder's Law" has held true. In effect, storage is increasing exponentially as measured in bytes stored per squareinch. At the same time, the cost of storage devices and media is dropping. While the industry does a phenomenal job of storing ever-greater data quantities, the real cost of storing unstructured data is not media or devices, it is data management and protection.

Unstructured Data Usage Patterns
Businesses with retention policies frequently adopt a "save everything" mentality. When business units bring in new sources of unstructured data and utilize productivity applications to dynamically create data, it becomes IT's responsibility to store and protect these files. However, usage of any given file from these new sources of data tends to drop off steeply, leaving a significant amount of fixed data on Tier 1 disk. Just as quickly as a file or data collection becomes important, it can become obsolete and ignored.
In the hundreds of data audits conducted by a 3rd party, 90% of clients found that 50% to 80% of their data had not been accessed or modified in more than a year. In large organizations, this figure can run even higher. One large company found that in 50TB of unstructured data, 60% had not been accessed in 3 years. That amounts to 30TB of unused data that IT had to store, protect and tag for Tier 1 backup.

Existing Treatment of Unstructured Data
An IT department often has no visibility into the status of a given file. Since storage, as measured by Kryder's Law, is perceived as cheap, users and applications do not habitually delete unused files. As a result, the amount of data stored only grows, inevitably outgrowing the capacity of direct attached, NAS, SAN or any other model of storage. Companies must either add new servers and their accompanying maintenance overhead or "forklift" the old device in favor of a new device with increased capacity.

Without a change in the treatment of unstructured data, the enterprise has no choice but to continue to pay to upgrade its storage as its data grows at an exponential rate. Supporting unstructured data costs money. The past decade has proven that simply chasing data growth with additional storage and backup horsepower is not sustainable. What IT organizations need most is the ability to add intelligence to the data management process.

Identify the Need
Cheap-to-purchase drives and user-managed file systems are business enablers, but they weigh heavily on the shoulders of the IT budget and resources. Organizations need to lower the cost of storage going forward while maintaining their previous investments in storage and backup systems. They can accomplish both objectives with a solution that gives them visibility into unstructured data and the ability to automatically sort unstructured data by criteria. New file archiving technologies provide the ability to intelligently reduce storage management cost. The techniques include:

- Automated storage tiering.
- Object stores that utilize metadata to categorize and apply polices to unstructured data.
- New backup strategies that do not rely on removable media and manual processes.

File Archiving Solution
The file archiving solution from Hitachi Data Systems provides all of the functionality required to implement an intelligent data management strategy that aligns data activity with storage and backup infrastructure. The solution takes advantage of file discovery, contextual classification and automated actions to drive data to a high-performance object archive with built-in data protection capabilities. Companies can reclaim their primary storage and run fullsystem backups on only the most active and valuable data. They can also identify and store data cost-effectively, without impacting the ability of an end user or application to access data in the usual ways.

By implementing this file archiving solution, an IT department can immediately reduce the cost of storage. It can slow the growth of Tier 1 data while continuing to provide the highest level of service to business units. This solution offers the following benefits:

- Automatically move inactive data to a long-term object store to reduce the real cost of managing storage and optimize the backup and recovery process.
- Apply consistent policies across the enterprise while simplifying the application of those policies.
- Provide granular policies to all data and make that data available to the right business unit.

- Reduce the total cost of maintaining data by gathering information to help make optimal use of storage tiers.
- Provide business units with transparent access to archived data.
- Expedite discovery by searching across all sources at once.
- Pick the best storage and operating system platform to meet business requirements. Our solutions support Microsoft Windows®, UNIX/Linux and any flavor of NAS.
- Archive unstructured data from any storage platform to a best-in-class object store.

The power and completeness of the file archiving solution from system that contributes discovery, classification, reporting and automated data actions functionality and Content Platform that adds scalability, accessibility and content preservation capability for managing all your unstructured data. This potent combination optimizes storage assets and builds resilience into the IT infrastructure. Applying policies to unstructured data based on usage offers several advantages.

You can:

1) Reclaim your primary storage.

2) Create storage policies that reduce backup overhead.

3) Run full-system backups on only the most active and valuable data without impacting users and applications.

The file archiving solution gives IT the ability to continually provide the services businesses need without a corresponding growth in the storage budget.

With this solution, we can:

- Adhere to compliance and retention policies.
- Lower the per-terabyte cost of backup.
- Consistently complete backup of mission-critical data.
- Improve disaster recovery capabilities.
- Provide secure, high-performance services to business units.
- Stop putting a bandage on your storage problem every year.
- Implement a solution that provides best-in-class ROI for the management, protection and retention of unstructured data

The speed of business and the prevalence of inexpensive drives and user-managed file systems indicate that IT needs to prepare for a future of ongoing storage expansion. IT also needs the ability to deliver high-quality services to meet the expectations of SLAs for recovery and performance. This means not just adding more drives, but also providing backup, data protection, compliance, long-term retention, oversight and eventual destruction of the accumulated data.

To continually reduce storage on primary drives and manage data throughout its lifecycle, you need a solution that goes beyond deduplication and compression. The file archiving solution from Hitachi Data Systems offers comprehensive data management to ensure data is available, protected and automatically relegated to the right storage drive based on its value to your business.

II. CONCLUSIONS

IT professionals realize that the cost of maintaining storage is growing and threatens to consume their budgets. The core problem driving this data growth is unstructured data remaining on expensive Tier 1 drives even after sitting untouched for years. This data places a burden on backup systems and adds complexity to meeting disaster recovery windows.

Unstructured data will continue to grow at a high rate, which means IT must find cost-effective, competitive ways to provide the data capacity business units need while ensuring complete and consistent backup of business-critical data. It's time to stop treating all data as equal, because it's not. An organization that can automate the archiving of infrequently used content can resolve several sticky issues that haunt backup and storage planning.

REFERENCES

1. Doan, AnHai, et al. "The case for a structured approach to managing unstructured data." arXiv preprint arXiv:0909.1783 (2009).
2. Blumberg, Robert, and Shaku Atre. "The problem with unstructured data." DM REVIEW 13.42-49 (2003): 62.
3. Baars, Henning, and Hans-George Kemper. "Management support with structured and unstructured data—an integrated business intelligence framework." Information Systems Management 25.2 (2008): 132-148.
4. Carnes, David, and Nicholas Longtin. "Method and device for managing unstructured data." U.S. Patent Application No. 11/148,757.
5. Narancic, Perry J., and Paul Krneta. "Method and system for processing structured data and unstructured data." U.S. Patent No. 7,668,849. 23 Feb. 2010.
6. Corti, Louise, and Paul Thompson. "Secondary analysis of archived data." Qualitative Research Practice: Concise Paperback Edition, SAGE Publications Ltd, London (2006): 297-313.

