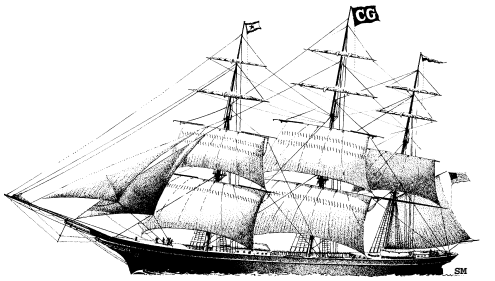
THE CLIPPER GROUP
Navigator™



Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2010031 June 28, 2010

Dealing with Cool and Cold Data — and Getting It "Just Right"

Analysts: Mike Kahn and Anne MacFarland

Management Summary

The story of *Goldilocks and the Three Bears* is known by almost everyone. Of specific importance today are *the three dishes of porridge: one "too hot", one "too cold" and one "just right"*. In today's data center, tiers of storage are analogous to those dishes of porridge.

- **Some data tends to be hot** and we don't want it to get to be too hot. It is hot because it is in high demand. For example, indices for databases tend to be very hot. If the application is mission critical, a significant effort is made to spread out the heat (so that nothing gets "too hot"), which keeps the application from overheating and possibly failing to keep up. Fortunately, the hot data tends to be less voluminous than the other kinds of data.
- **Some data tends to be very cool or even cold.** What this means is that the data is rarely being used or not being used at all. Most cool or cold data is not mission-critical, although its preservation for reuse or litigation may be important. It doesn't make sense to spend a lot of money to make this data available instantly or even to store large quantities of it in ways that are more expensive. This bulletin is about cool and cold data and how to handle it economically.
- **Some data falls in the middle.** Its heat is somewhere between hot and cold. While few would call it "just right", many do see it as "business critical".

In the enterprise data center, we have spent a decade dealing with the business critical data in the middle (where the last wave of explosive growth occurred) and, more recently, have attacked the hot end by the addition of SSD drives. What we have failed to do well is focus on the cool and cold data, which for many enterprises is now the fastest growing data segment (by volume). It may have been neglected because it was deemed to be less important (which it probably is) but it may have been neglected because the prevailing solutions were either too expensive to implement on a vast scale or were too inconvenient to use for anything but high-value activities, like litigation discovery.

What is needed for cool and cold data is a two-faced solution. First, it must be sufficiently responsive, secure, and manageable. Second, it must offer a Total Cost of Ownership (TCO) that makes economic sense for this "less valuable" or "low access" data. Remember, being less valuable is a relative statement, made with respect to what has been deemed by the enterprise to be "more valuable" or "more frequently accessed" data. In reality, the value is in the eyes of the beholders. For the folks who need current access to cold data, it probably is very important. The key here is finding a way to preserve the data for later use in an affordable way, given the large volumes of cold data that many enterprises would like to preserve (if they could afford it). This chicken and egg conundrum needs to break out of its infinite loop. We think that *IBM's Long Term File System* on *LTO-5* tape might be the way to make it happen. Read on for the details and our explanation.

IN THIS ISSUE

Too Much Digital Information

Digital information is a great simplifier. It is easy to transmit and thus easy to copy and easy to steal. These characteristics, and the endless scope of instrumentation now online and churning out data, have changed how business is done and how it is monetized. Directly or indirectly, digital information has disrupted and transformed not just business models but entire industries. The new economic models that require a free flow of digital information have a much lower cost of entry. In the long run, they may support many more ways of making a living. The wealth of digital data now available about anything from business operations to product and service experience give many ways to make goods and services more affordable to more markets. Thus, digital data is a good thing – but it comes with consequences. One consequence is that digital data has become an expensive addiction. The portion of data center budgets dedicated to storage continues to increase except, perhaps, in processing-centric clouds. Due to the relatively short lifespan of electronic media and the distributed nature of most organizations, business information must be copied and relocated several times in ways that do not compromise integrity (bit-loss), security, or the chain of custody (control). This requires strategies, processes, automation, and discipline. The data center challenge – and the business challenge – is to manage the cost of keeping information in an affordable way that supports rampant but appropriate reuse. Let us look at some of the important details of this situation.

Business Elements

- Because digital information is easy to store, it is tempting to store as much of it as possible. There are a variety of new data sources that are relevant to a business or other organization, including much real-time process data that previously was not saved or accessible at the micro-transaction level. Today, the democratization of analytics and the flexibility of modeling now can make use of this "mundane" data to optimize operations.
- The volatile nature of most markets today leads organizations to hedge their bets by second sourcing and partnering. It drives them to innovate by reaching out to new markets. Both these strategies require additional data to be stored to support increasing outward-facing business processes.

Technology Elements

- The kinds of analysis and manipulation that were once confined to structured data (in databases) are now being used as well on unstructured information, and external sources of information, such as in collections of Tweets from Twitter. Not all these information sources need traditional "mission critical" protection – but some may.
- Information management tools – such as search, indexing, data quality tools, and data transformation – have grown in sophistication to a point where most organizations can characterize the kinds of information they wish to retain for a long time and reuse. These tools produce new information assets that must be managed and protected.
- Best practices and software have been developed for situational reuse of data, including eDiscovery, content management, and master data management, where two flavors are emerging – one to manage rapidly changing data (such as product pricing and special offers) and one to assure consistency of data definitions and used for core financial and legal operations. This software has great business value, but it also requires IT capacities of all kinds – processing, network and, of course, more storage.

It's another of IT's *chicken and egg* conundrums. More sources of data create more realistic and actionable analysis – and add to the capacity of stored data. As this capacity of data grows, more technology tools are needed to manage its retention and intelligent use. The rate of growth of retained, historic data demands a pragmatic attempt to take a large bite out of the costs of the retention of structured, semi-structured, and unstructured data.[1]

Much of the unstructured and semi-structured data held by enterprises quickly cools off (i.e., is used less as time goes by) until it becomes *cold data*, ready to be warehoused away until it is needed again. This has a probabilistically thin likelihood of every happening, as only a little of what is held in cold storage is ever retrieved. It would be nice to be able to predict future data needs, but we don't see that crystal ball on the horizon. **What we do see are new ways to store data more cheaply without sending it into oblivion in a dark, cold warehouse.**

[1] *Structured data* is exemplified by databases, and storage and internal structures often at managed at the block (LUN) level by the database management software (DBMS). Email is another self-managing example of structured data. *Semi-structured data* is exemplified by the office documents that we use each day. There tends to be a lot of formatted text and numbers in semi-structured documents. *Unstructured data* consists of two types: *binary objects* (like data streams from a radio telescope or DNA sequencing) and *digitized data* (digital photos, music, photos, movies, etc.) Unstructured data tends be big is size (in comparison to text documents) and the quantities of it tend be growing rapidly, creating a problem of Malthusian proportions.

The Problems with Disk

Don't take us wrong; disk storage is wonderful and getting better every year. Whether rotating media (what most of us think of when we envision a disk) or its new, much more expensive, solid-state brethren with no moving parts (also known as *SSDs* or s*olid-state disks*), disks are great for storing hot and warm data. However, when it comes to storing cool or cold data, disks tend to be too expensive, either because of the *acquisition costs* (arrays are expensive to buy), the *operating costs* (maintenance, floor space, cooling and electricity) or the *ancillary costs* (the people and software to manage all of this, including upgrading to later generations of disk). While improvements continue for all of these costs, when multiplied times a lot of data held for a long time, disk storage still gets to be very expensive (in the aggregate). In addition, with disk, all the data is on-line (and available), which increases the risk for corruption or loss of data from attacks (from hackers, viruses, etc.) or from system errors that can occur with on-line storage.

If you have rented physical "storage space" (the kind that is measured in cubic feet and holds the stuff that you might want to use someday but no longer have any spare space at your residence), you understand the high price you are paying to hold onto your possessions of questionable future value. This is the big problem facing enterprise data centers. **Everyone seems to want to keep everything, for a long time, often approaching forever.**[2] **The cost of doing this is the constraining factor.** In the old days, we would just get rid of the "old stuff" because we didn't have any way to keep most of it. Now, we have many good business reasons to preserve it forever and it is the cost that is keeping us up at night. *How are we going to afford to do this…in the future…over the long term?*

Thus, the problem is not that disks are inherently bad (in general) but that they will eat you alive if applied as the universal solution for the Malthusian explosion of cold data to be retained. That's the problem, in a nutshell. **You need a better solution for cold data retention.**

For many of you faithful readers, you know what we are about to say, as we have said it before, but there is a new wrinkle. ***Yes, the answer is tape.*** Not your father's or your grandfather's tape, which did a good job of holding and retrieving sequential data, but a new kind of tape that acts, in many ways, like a disk! Wow!![3]

The Problems with Tape

To be fair, we need to point out that tape has its imperfections and limitations, just like disk. This list used to be longer, but the problems of longevity, care and feeding (especially in the era preceding tape automation), security, and standards have been eradicated by five generations of LTO solutions[4], virtual tape systems front-ending tape, etc. **However, tape has certain characteristics that have been its hallmark, as well as its delimiter. *Tape is, well, tape.*** It is fast streaming, off-line for data protection, portable, has a long shelf life, and is low cost. It is a serial media with a beginning and an end and a lot of stuff in between. This serial nature has defined its limits and uses over the last 60 years. Tape has its own format, which has changed over the years and through many generations of products but, by and large, it consists of a header, intermediary blocks of data, and end-of-tape marker. The intermediary blocks are measured in offsets (distances) from the beginning of the tape. This is not very user-access friendly, which is why you need special software to read from and write to tape. And that sums up the remaining problem with tape – ***tape wasn't initially designed to be user-access friendly*. Fortunately, this has been fixed, in a modest way today, and with great potential for the future.**

21st Century Tape Has Arrived

That's a pretty bold section heading, we know. Maybe it should declare the arrival of "tape with a new attitude". Whatever you want to call it, **tape**

[2] With some governmental jurisdictions requiring that medical records be held for decades beyond the death of the patient, "forever" become a tangible and real requirement.

[3] If you have been around IT for decades, you will recall that there have been several of these "Eureka" moments in the history of using magnetic devices to store information. Tape came first (the round reel kind), along with tape-based operating systems (before there were disks, mind you, operating systems were read from tape on boot up). Then there were several rotating magnetic devices (drums, with a single head for each track, and disks, with a single head for each platter). Then there was the short-lived data cell, with a retrieved length of tape wrapped around a cylinder and read via a tape head. (Really, this might have been ideal for cool and cold data but failed because too many folks used it for warm-to-hot data.) Fast forward a couple of decades, and the then common 14" platters were downsized to 8", then to 5.25", and then to 2.5", and now we have SSDs. We also have disks that act like tape (called Virtual Tape Libraries (VTLs). Each one of these changes had the potential to rewrite the rules of storage. Often, this did not happen in the ways originally envisioned. This makes the transitory possibilities much fun to predict.

[4] For more information, see *LTO Consortium Announces Next Gen Tape - LTO-5 Raises the Bar for Tier-3 Storage* in **The Clipper Group Navigator** dated January 31, 2010, and available at http://www.clipper.com/research/TCG2010002.pdf.

has gone through another genetic evolution. All tape before it must be considered "old tape". **Ignoring tape's physical formats that have evolved over the decades, some new DNA has been injected and the species has evolved.**

To most of us, what makes disk seem friendly and tape seem unfriendly? When you understand this, you will comprehend the recent genetic evolution. Disk is friendly because you can personally access and use it. You see disk drives and folders (or directories) and files within them. You may know where they physically reside (like on the *C:* drive on your laptop) or you may not know or care, because someone else has virtualized the location. *Life is good!* It's all so simple. We all like files and are comfortable with them. We don't really want to know about the physical device(s) involved or the algorithmic processes required for storing and retrieving data that has been laid down on some kind of physical media.

In order to make tape friendly, it really needs to speak the language of files that we all know how to use and manipulate. It is that simple and it has happened – **a new kind of tape has been born and it is ideal for cold data. Cool!!!**

LTO-5 Tape, with a Cool Twist

Earlier this year, the fifth generation of LTO tapes (called *LTO-5*) arrived, with several vendors announcing new drives and cartridges with capacity increased to 1.5 GBs uncompressed. IBM was among them.[5] However, only IBM took immediate advantage of the dual partitions that were part of the LTO-5 specification. (The technical explanation of the dual partitions isn't critical to our discussion that follows, but if you want to know more, read the paper referenced below in footnote #5.)

What is novel and important is what IBM did with the dual partitions. It enabled a new feature called the *Long Term File System (LTFS, based on the Linear Tape File System specification)*, where *Long Term* is another way of connoting *cold data*. LTFS is an extension to the operating system and allows files to be stored on tape and handled in the way we handle files on any file device. For now, just think of it like a giant USB-connected flash-memory drive (that looks and feels like it is part of your operating environment – because it is).[6] **LTFS allows you to store and retrieve files on off-the-shelf dual partitioned LTO-5 cartridges. It maintains an index of the files in the thinner partition and stores the files in the fatter partition.** Thus, the tape becomes *self-describing*, with no software application dependencies. It can provide faster searches, long-term archive and cross platform interchange.

Today, each tape is a standalone file system. When you insert a cartridge, your local operating system reads the index and presents a familiar folder or directory of the files residing on the cartridge. You can drag and drop files to and from the tape in the same fashion as you do with disk and other removable media.

There are several implications to this self-sufficiency. First, a tape is not under control of another piece of software, such as an archiving system. (This is both advantageous and limiting, as will be discussed below.) The tape can be moved from one system to another, cross platform, and it will work automatically, without requiring the involvement of a proprietary application, which is common to most traditional uses of tape. **By its self-descriptive nature, LTFS frees digital assets from dependency on the application that generated them.** This is important in any environment where assets are kept for decades, such as health care and animation. As discussed in the next section, this independence is valued in certain applications.

An Immediate Use Case – Media & Entertainment

There is one industry, Media and Entertainment, which has always used tape. At the recent meeting of the National Association of Broadcasters (NAB), participants were very vocal in their enthusiasm for LTO-5 and IBM's LTFS, as it is offered today (as standalone dual-partitioned cartridges with LTFS and a single tape drive).

The media and entertainment industry has always used tape and values it for its security, portability, durability, and modest cost.[7] It is a poster child for data's hyper growth. High Definition video (HD) quintuples video's storage capacity requirements, and 3D films triple the HD capacity needed – that's a twelve-fold increase, and that's just for the final cut.

However, there is much more to a movie than

[5] For fuller details about LTO-5, in general, and dual partitions and LTFS, in particular, see **The Clipper Group Navigator** entitled *IBM Enables the Data Center with LTO-5 Products – Increasing Capacity, Throughput and File Management*, dated April 12, 2010, and available at http://www.clipper.com/research/TCG2010017.pdf.

[6] No, it is not connected by USB, but by Fibre Channel or SAS. The USB reference is just a convenient analogy.

[7] 94 minutes of professional videotape costs about $200, compared to about $125 for a partitioned LTO-5 cartridge capable of holding about 4 hours of digital video.

the final cut – there are audio feeds, special effects, language translations – all of which cost a remarkable amount of human time, energy, and server cycles to generate. Producers look to recoup their investments with collateral, sequels, and other forms of asset reuse. Of course, supporting reuse means long retention of a wide variety of assets at high resolution, which is a perfect example of cold data and a perfect use of tape. With LTFS, the producer can view, access and play the video content directly from the tape. The tape can be shared with an editor who may be using MAC OS and then with a digital effects specialist, for instance, that may be using a Linux OS, thus making the interchange and portability of content. Plus, it will make migration from LTO-5 media to future LTO generations into an easy endeavor.

For long-term retention in Media and Entertainment, disk is an expensive choice. Disk arrays are more expensive, in terms of total cost of ownership, than tape.[8] For most enterprises, the useful lifetime of an array is three-to-five years. A tape library lasts much longer, and a tape cartridge can last up to 30 years in a controlled, data center environment.

The expense of the choice is also more than hardware. For the film industry, the content is intellectual as well as digital property. Any degradation of the content sharply reduces its value. Migration must be done meticulously so that all the different layers remain synced. The customer experience of viewing a film is critical to its success and ultimately to revenue. The long lifespan of tape means many fewer migrations to media during its useful life than would be needed, if stored on successive generations of disk systems. In these days of rampant use and reuse of video, migration is seldom a small-scale challenge. With the complexity of feeds involved in today's cinematic productions, it is a distinctly non-trivial event.

The portability of tape means that various program development and production teams can work on a movie without incurring the networking costs (requiring much bandwidth) and security exposure of sending it over a wire. For an industry where leaks can harm revenue, risk avoidance and a clear and enforceable chain of custody is highly valued.

The Need for More

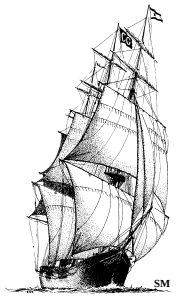
That standalone independence is a shortcoming, however, if you have petabytes of cold data to preserve for later use. Right now, IBM is only making LTFS available to those who can take advantage of a single-cartridge (standalone) solution.

To see where this might be going, you need to look into our crystal ball. Suppose you have a tape library full of these self-sufficient LTFS-based cartridges. You could quickly read each tape's index each time you want to access that tape's content, however, to find what you want across the large tape pool, it would be exponentially helpful to have a master index to this archived collection, which knows where everything is located and is able to keep that current as cartridges are added, modified, removed, and deleted. This could help address eDiscovery tasks and assist with data migration. While IBM hasn't announced this, we can't imagine that this greater capability isn't coming, especially since IBM now offers many information archiving solutions.

The need to store lots of cold data is there, for sure, and LTFS has potential across many industries and applications, like media and entertainment, video surveillance, medical images, architecture/industrial drawings, legal documents, government records, cloud storage, eDiscovery, media migration and many more. An exciting journey lies ahead!

Conclusion

Dealing with cool and cold data on a large scale – in a user-friendly way – is an important IT challenge crying out for better solutions. We think putting files on tape is that better solution. Once you stop chuckling at the idea that tape might be useful in ways that you never imagined, take a few moments to consider how you might use file-based tape, either in the standalone case of today or in the large-scale solution that we predict. **A new future for tape lays ahead … and we think it feels "just right"!**



[8] For a specific analysis of the costs of disk and tape see the February 3, 2008, issue of Clipper Notes entitled *Disk and Tape Square Off Again – Tape Remains King of the Hill with LTO-4*, which is available at http://www.clipper.com/research/TCG2008009.pdf. This is an analysis of the LTO-4 generation. The many advances of LTO-5 would make the comparative cost and energy differences even more striking.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Authors

Mike Kahn **is Managing Director and a cofounder of The Clipper Group.** Mr. Kahn is a veteran of the computer industry, having spent more than four decades working on information technology, spending the last 17 years at Clipper. For the vendor community, Mr. Kahn specializes on strategic marketing issues, especially for new and costly technologies and services, competitive analysis, and sales support. For the end-user community, he focuses on mission-critical information management decisions. Prior positions held by Mr. Kahn include: at International Data Corporation - Director of the Competitive Resource Center, Director of Consulting for the Software Research Group, and Director of the Systems Integration Program; at Power Factor Corporation, a Boston-based electronics startup – President; at Honeywell Bull - Director of International Marketing and Support; at Honeywell Information Systems - Director of Marketing and Director of Strategy, Technology and Research; at Arthur D. Little, Inc. - a consultant specializing in database management systems and information resource management; and, at Intel Corporation, Mr. Kahn served in a variety of field and home office marketing management positions. Earlier, he founded and managed PRISM Associates of Ann Arbor, Michigan, a systems consulting firm specializing in data management products and applications. Mr. Kahn also managed a relational DBMS development group at The University of Michigan, where he earned B.S.E. and M.S.E. degrees in industrial engineering.

- ***Reach Mike Kahn via e-mail at Mike.Kahn@clipper.com or via phone at (781) 235-0085 Ext. 121. (Please dial "121" when you hear the automated attendant.)***

Anne MacFarland **is a Senior Contributing Analyst for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group in 2001 after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial "128" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group *Captain's Log***, **The Clipper Group Voyager**, **Clipper Notes**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosures

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

After publication of a bulletin on *clipper.com*, The Clipper Group offers all vendors and users the opportunity to license its publications for a fee, since linking to Clipper's web pages, posting of Clipper documents on other's websites, and printing of hard-copy reprints is not allowed without payment of related fee(s). Less than half of our publications are licensed in this way. In addition, analysts regularly receive briefings from many vendors. Occasionally, Clipper analysts' travel and/or lodging expenses and/or conference fees have been subsidized by a vendor, in order to participate in briefings. The Clipper Group does not charge any professional fees to participate in these information-gathering events. In addition, some vendors sometime provide binders, USB drives containing presentations, and other conference-related paraphernalia to Clipper's analysts.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.