

Data Migration Practices and Tiered Storage Management: Challenges and Opportunities

August 31, 2010

Abstract

Firms continue to face multiple challenges in managing large amounts of data across a tiered storage infrastructure. Data migration between storage tiers is recognized as one of the most complex challenges. We review factors that can cause storage managers to develop ineffective data migration practices. We describe a prototype decision support tool that managers can use to better understand when to migrate data to different tiers. Effective data migration practices are critical to balancing information value, risk, and information management costs.

Prof. Paul Tallon
Lattanze Center for Information Value
Loyola University
Baltimore, MD
Phone: 410-617-5614
Email: pptallon@loyola.edu

Dr. Jim Short
Global Information Industry Center
University of California
San Diego, CA
Phone: 858-534-5014
Email: jshort@ucsd.edu

1. Introduction

Businesses are experiencing explosive rates of data growth. Based on current trends, the total amount of data under management is expected to double in size every other year. Firms in sectors such as healthcare, financial services, telecommunications, and retail are especially susceptible to extreme data growth as new regulation and investment in business analytics calls for more transaction-specific data to be captured and retained for longer periods of time.

Many firms have adopted a tiered approach to storage management. Although there is no agreement on the optimal number of tiers, most firms have adopted a three-tier structure, placing highly valuable data in tier 1 and least valuable data in tier 3. Tier 2 is typically reserved for frequently accessed data of moderate to high value. As noted in earlier research, information value is subject to the vagaries of time.¹ Value can increase as usage expands but information value can also decay as data ages and becomes less relevant to future decision making. In principle, when value changes, data managers can elect to migrate the data to a storage tier that is more reflective of its new value. The reality, however, is that IT managers often lack an agreed-upon means for quantifying information value. The absence of specific metrics has meant that data migration practices have become largely ad hoc and ineffective at managing exposure to data risk factors.

The structure of this paper is as follows: first, we present a conceptual model of data migration practices that address technical, economic, and business policy factors. Second, we review findings from interviews with IT managers on the effectiveness of data migration practices. Third, we present a prototype decision support tool to model the risk-adjusted cost of data migration. A future, second phase of research plans to further refine and validate the tool for use in field settings.

Below we describe current data migration practices, outline our risk and value based conceptual model, and describe the data migration tool. We conclude with observations on current practices and describe future planned research.

2. Triggers for Data Migration

Firms migrate data for many reasons, some of which are not directly related to changes in information value. For example, a standard practice is to free up additional storage capacity, presenting managers with an opportunity to migrate data across storage tiers. Data de-duplication projects can also deliver migration opportunities, as can equipment refreshes. Our analysis found that many if not most migration exercises are usually triggered by factors that are not directly tied to policy-based changes in information value. As most firms have yet to implement a systematic way to calculate information value or to assess changes in value, there will rarely be a mechanism in place to automatically migrate data whenever its value changes.

3. Best Practices

Storage hardware vendors have tackled the question of best practices for data migration as part of their product solution strategies.² Similar practices have also been described in the trade press, most notably in Storage Magazine and by industry services organizations such as Gartner and IDC.³ A common theme in migration planning is a discussion of business and technical risk. Business risk is typically defined as the cost of business disruptions if users cannot access data during migration. Technical risk is usually defined

¹ See Information Lifecycle Management, 2007, P. Tallon, Communications of the ACM, 11 (7), 65-69.

² For examples of “Best Practices for Data Migration” discussion documents and whitepapers by storage vendors, see IBM Global Technology Services (<http://www-935.ibm.com/services/us/gts/pdf/softek-best-practices-data-migration.pdf>) and Network Appliances or NetApp (http://partners.netapp.com/go/techontap/NGS_migration.pdf). See also: Gartner Research (Ted Friedman), 2009, Best Practices: Mitigate Data Migration Risks and Challenges.

³ Damoulakis, J., Nov 2007, “Best Practices: Tackling Data Migration” Storage Magazine.

as technical disruptions that can occur in the data migration process. Validation checks are recommended as part of data migration to ensure that data migrates correctly and is fully accessible at its destination tier. In summary, a review of industry practices finds the following steps:

- Planning:
 - evaluate storage or other IT practices that lead to the discover of migration opportunities
 - identify data to be moved
 - set migration parameters (migration paths, timing, capacity requirements)
 - consider data de-duplication measures prior to migration
- Data Migration (technical):
 - monitor data migration
- Data Validation (technical):
 - verify completion, accuracy, and accessibility
 - ensure migration log is updated
 - remove data from source disk

Notwithstanding industry best practices for data migration, several important factors are often omitted from best practice procedures. For example, factors including *information value* (what data can be moved and when), *cost* (what do we stand to gain or lose economically from migrating data between tiers), and *risk* (what technical, organizational and economic risks are associated with data migration) tend to receive minimal attention in published procedures. Current best practices assume that firms know in advance the trigger points leading up to when data will be migrated and where it needs to go. Vendor models typically assume that customers have stable and reliable processes in place to identify changes in information value prompting decisions to move data to lower or higher tiers. Vendor models may also assume that customers have policies defining the risks of migrating data across tiers, especially when migrating data from higher to lower tiers of service. In this instance, cost savings in the form of lower total cost of ownership (TCO) may need to be scaled back to account for an increase in business or technical risk from equipment failure, slower access or reduced back up frequency. In summary, current best practices assume that firms will directly account for the different components of cost and risk in moving data between tiers.

Our research suggests that best practices can be difficult to apply in operational settings. Generic practices minimally account for the complexities of reliably migrating large quantities of data between tiers. In case study interviews, data migration was described as more *art* than *science* as data managers devised ad-hoc procedures to assure themselves that data will not be lost during migration. In previous studies, we found best-of-breed firms in different sectors that have not yet fully socialized or implemented standard policies to define information value and business risk across organizational units. In these firms, IT departments operate according to their own definitions of value and risk, and as such local practices can be incomplete and ineffective. For example, risk may be defined as systems reliability by one IT department. Another IT group may see systems reliability as a part of risk determination that includes other factors such as speed of access, recoverability, and maintainability.⁴

⁴ An ideal scenario that would help firms to understand risk and to factor it into their storage decisions would be to collapse all risk characteristics into a single number rather than talking about separate measures of reliability, up-time, access, etc. This has not yet occurred. Elsewhere in the literature, researchers have succeeded in developing a single resource metering measure of server capability based on CPU, memory, disk storage, and I/O. A case in point is the resource unit metric developed by Provment (later acquired by Satori Technologies) – see presentation at http://intervirt.maphis.homeip.net/training/VMWare/VMWorld-2007/Sessions/BM/PS_BM18_288968_166-1_FIN_v3.pdf or HP's Computon metric using in utility pricing (see <http://en.wikipedia.org/wiki/Computon>) and http://www.computerworld.com/s/article/81522/HP_takes_new_pricing_path_for_utility_based_computing). No such equivalent risk metric or score currently exists to succinctly capture a measure of risk for storage technology.

4. Operational Measures of Information Value

In this section, we review current industry practices for defining information value. We begin by outlining the typical association between value and frequency of data access or use. While frequency of use is often used as the proxy for value – low frequency of use means low value, frequent use means high value – the value of information is more easily understood as how much impact a piece of information has in making a specific decision. By corollary, information value is a function of how much value a firm fails to realize (an opportunity cost) in the event that information is unavailable when called upon by the decision maker.

The theoretical cap on information value is the market value of the corporation, assuming there are no other physical assets other than the information held within the firm's computer systems. The maximum amount that a company stands to lose if it loses all of its data is the market value of the firm, even in cases where data loss leads to punitive damages that surpass the entire market value of the firm. Naturally, the minimum cap on information value is zero. Even if archival data is seen as having no clear future use and, therefore, is seen as having no value to the firm, a decline in value could reverse if the data was needed by managers at a later time or if, for legal purposes, the data were required as part of an e-discovery process. Failure to hand over data – even zero-value data – could lead to significant court-imposed financial fines.

After frequency of use, information value can be defined as the information input into business impact analyses. If the information is important, it is included; if not, it is excluded. The problem with defining information value as input into business impact analyses, however, is the reliance on a single time period estimate of value. Typically, business impact analyses are not ongoing, whereas in a constantly evolving world, information value is likely to change based on the information lifecycle. If the BI analysis is conducted at an early stages in the information lifecycle, that value may not be accurate for a later stage in the lifecycle. Current generation BI software has only rudimentary ways to account for lifecycle changes. The point is that imprecision in either underestimating or overestimating value has cost implications.

Firms may also try to identify information value from the perspective of the end user. However, users are prone to bias and error since they typically view data in local usage terms. Data that is seen as necessary to their jobs is considered valuable. This *could* correlate with the value of the information to the business process or functional activity of the firm, but it does not *have to*. Of course, users typically are in a better position to define information value than IT / storage managers. However, in practice few firms have attempted to inform end users of the number of technical and business factors involved in data migration. A few firms boil complexity down to the cost of service. In cases where users are assessed a chargeback fee for storage usage, there is greater motivation for users to weigh costs against their perceptions of value.⁵

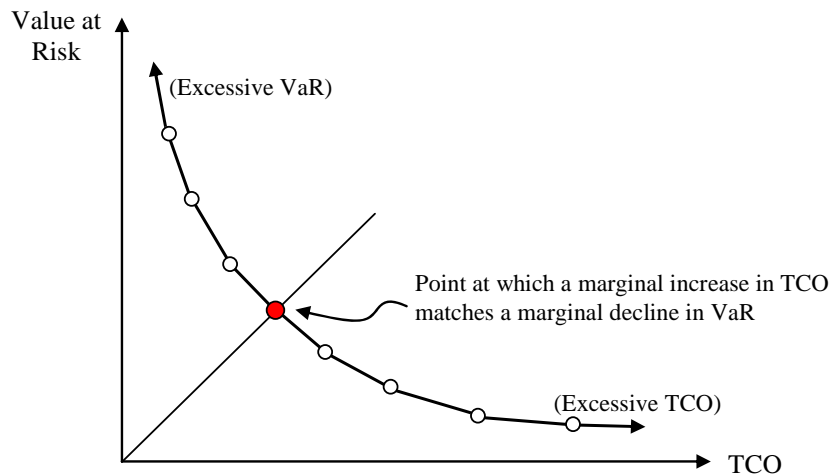
Without an underlying value to risk model, however, it can be difficult for users to compare cost to value. Value can be in the *eye of the beholder* while cost (for example, if charged on the basis of gigabytes of stored data per month) is tangible.⁶ The bridge between value and chargebacks (or TCO if chargebacks are not used) is for users to see what they are getting for their chargeback. As we discuss in the next section, the issue is less one of aligning value with TCO than it is to align value-at-risk (VaR)⁷ with storage

⁵ Chargebacks, to the extent that they are accurate, fair, and traceable, can change user behavior. In a data storage setting, chargebacks can motivate users to scale back requests for storage space that might not be used. Users are prone to requesting storage just in case it is needed. Even if this storage is free from a user's perspective, it is still a charge to the corporation.

⁶ Chargebacks can also be problematic if data ownership is unknown or if data are shared between different groups. A user group may feel a certain inequity if they are assessed a fee for data that they rarely used. This could force the group to move data that they perceive to be of minimal value onto a lower cost storage tier (or to delete data in some instances) even though the true value of that information from the corporation's perspective remains high.

⁷ Value-at-risk (VaR) is a standard metric used by portfolio managers to monitor their exposure to market volatility. VaR represents the maximum amount of value that managers stand to lose, within a given confidence interval, in a

costs within each tier. As illustrated in the figure below, changes in value (with no changes in TCO) will prompt VaR to rise or fall. If information value rises and VaR climbs to an uncomfortable level, data can be migrated to a higher (more expensive) storage tier where improved storage performance and reliability will help to reduce VaR to a more reasonable level. What this means is that users need to be aware of risk to their data (what do we stand to lose if we cannot access our data), not necessarily on a minute-by-minute basis but often enough that they can decide if risk is excessive (or if they have overly insured themselves against risk) and whether data need to be migrated to a different storage tier to bring VaR more into line with storage TCO. It must also be noted that equipment refreshes will cause a shift in the relationship between TCO and VaR insofar as the same level of TCO on a newer, better, faster, and cheaper storage platform is likely to give rise to a reduced level of risk.



5. Risk Analysis and Mitigation⁸

The discussion above leads naturally to a review of risk at the level of each tier and how a closer understanding of risk may affect policies for data migration.

IT risk is generally defined as, “the potential for an unplanned event involving a failure or misuse of IT to threaten an enterprise objective”.⁹ Risk is the product of a probability of a damaging event and the cost of that event, summed over all such events. Risk can be assessed at the level of an individual storage tier by examining logs of past outages to identify the incident type and the frequency of each incident. Users may then be able to assign a dollar value based on downtime or recovery costs to each event. It is then possible to compute risk from the probability of various events and their costs. A limitation of this approach is that it does not include all possible events. Extreme events could be omitted if they have never occurred in the past meaning that whatever risk values are determined will probably understate the true actual risk levels.

particular trading period. For example, a portfolio manager may say, based on an analysis of past market data that, with 95% certainty, their mutual fund will not lose any more than 5% of its value in any trading session. VaR can be represented as an absolute dollar value or a percentage of some underlying portfolio value. If projected market losses exceed VaR, portfolio managers may invest in hedging strategies to limit their losses. Equally, if VaR falls, managers may decide to limit their hedging expenditure in the hope of earning a higher return. Data environments can be considered using the same logic. In this instance, information VaR is the expected cost associated with the risk of data loss, slower access, and system outages. IT managers can hedge against an increase in VaR by moving their data to a more expensive storage tier where there is a lower risk of system outage or data loss.

⁸ Risk is part of ITIL (see http://www.nysforum.org/documents/html/itil-6-6-06/itil_files/800x600/slide32.html).

⁹ IT Risk, 2007, G. Westerman, R. Hunter, Harvard Business School Press.

To the extent that these risk values are communicable to users, it may still not be sufficient to enable users to make sense of the link between information value, risk, and cost. What users ideally need to see is how changes in TCO relate to changes in value-at-risk. If firms see a way to reduce VaR by migrating data to a higher storage tier, they may be better able to perceptually cost-justify the extra expense of paying for the data to be retained on a higher tier. If they believe that information value has fallen and this is verified by a decline in VaR, they may be able to justify archiving or migrating the data to a lower cost storage tier.¹⁰

Risk mitigation may result from greater awareness of information value and how storage costs can act like an insurance premium to protect information from adverse events. Yet risk mitigation can also result from policies that do not necessarily drive up cost but that force users to adhere to policies meant to backup and protect data. Risk mitigation can also result from being smart about how data are managed. For instance, data de-duplication may reduce the length of time needed to backup data while also allowing e-discovery to occur at a faster pace. Standards-based data migration policies can incorporate consideration of risk as part of a comprehensive effort to bring costs into line with information value and value-at-risk.

6. Information Migration Drivers and Execution: A Generalized View

Our interviews reinforced the view that data migration in many organizations is an ad-hoc process where data is moved only in response to noted capacity shortages or extreme utilization peaks. Data is typically migrated as part of a housekeeping exercise, enacted in response to a sudden shift in data growth rates. If there is ample storage capacity available, organizations may not feel pressure to move data around from tier to tier according to the logic of the information lifecycle curve. Paradoxically, if data are moved to an excessive degree (say, if an automated tool is employed), there may be higher costs from inconveniencing the user than there are cost savings. Tools or users may opt to migrate data on the basis of various rules: date of last access (data that have not been accessed in a prescribed number of days may be moved) or data of creation (permits an aged data analysis). Regardless of the rules that are used to move data, it is beneficial to maintain an audit log (showing what data was moved, when, why, and the migration path) so that in the event that data need to be restored, an audit trail will be available.

Respondents noted that the creation of a rules-based data migration policy, while useful, may not work in each part of a business. If a single rule applies across the business (for example, data on tier 1 with date of last access greater than 90 days must be moved to tier 2), it may work for some users but not for others. In these cases, some users may be inclined to circumvent the system if they consider the rule to be unfair or damaging to their work. Consequently, while it may be considerably more complex, individual rules may need to be set for different departments or potentially for different users. As previously noted, if users are conscious of their costs through a chargeback fee, they may be better able to justify in their own minds why it is appropriate to retain data on a premier storage tier for longer periods than are permissible under a corporate standard.

Industry best practices also report that prior to data migration, data can be de-duplicated. Risk can also be reviewed on the originating and destination tiers once the migration has ended. If, for example, utilization rates fall on the origination tier (the source point) after data has moved, risk can also decline. Similarly, if utilization rates increase on the destination tier after data has migrated to it, risk can rise to an appreciable and perhaps even an uncomfortable degree. The concern, of course, is that if data migration is continuous (as is likely to occur once an automated data migration tool is instituted), this risk re-assessment exercise is also continuous. It would be useful if this risk monitoring exercise could be automated as a manual step is likely to be time consuming, repetitive, and highly complex. Errors are also likely if data managers are unable to accurately determine information value or the risks of information loss within an individual tier.

¹⁰ See Information Lifecycle Management, 2007, P. Tallon, Communications of the ACM, 11 (7), 65-69.

Unstructured data represent specific challenge for migration purposes. Unlike structured data in financial accounting applications or data residing in ERP system tables, unstructured data reside in standalone files. Email is among the best known examples of unstructured data since email is seldom tagged to reveal its contents. Desktop files such as spreadsheets, word documents, and presentation files are another example. Oftentimes, users will retain copies of earlier draft versions of documents, long after the final version has been produced. While these data can be migrated – just as email can be migrated once it passes a specific age – it may be preferable for users to have visibility into their data (even if only to metadata) in order to clean up or remove unnecessary files prior to a migration exercise.

7. Hardware Refresh Decisions

Respondents also stated that data migration can be triggered by the need to refresh hardware at the end of its useful economic life. Firms can decide to move data en masse to new hardware within the same tier or to systematically identify information value for various data types on the old hardware before moving that data to the new storage hardware within the same tier or moving the data to other tiers. While hardware refresh decisions can be timed to occur in three-year cycles, for example, there may be different motivating factors behind hardware refresh decisions within the highest tiers versus the lowest tiers.¹¹ A decision to replace equipment in the highest tiers may be motivated by the desire to reduce risk through a better infrastructure solution. Even if existing hardware is not fully depreciated and new hardware is more costly, it may still be appropriate to consider equipment replacement as a way to bring risk down to more tolerable levels. For lower tiers, risk may not be as important as cost. In these instances, the appearance of better, faster, and cheaper storage hardware may accelerate a decision to replacement lower tier hardware as a way to reduce cost. Naturally, the complexity of large scale data migration and the disruption to users shows the risk involved in data migration exercises. Such exercises need considerable planning and, in the expectation of technical failures, a tested recovery plan that ensures data integrity and business continuity.

8. Changing User Behavior

An insight gained from our interviews was the view that problems associated with escalating storage costs can be traced to user perceptions that storage was cheap and becoming even cheaper over time. Users also falsely believed that improvements in the price performance ratio of storage equipment more than offset a planned increase in the rate of information retention. Of course, for most users, there is no cost to keeping data forever or to requesting that their data be placed on the highest possible tier. Respondents stated that changing user behavior is critical to the success of any data migration policy. Besides signaling the actual cost of data storage through chargeback fees, users may also be motivated to consider information value if policies compel deletion of data after a specified number of days. Although this would force users to rate information value on an ongoing basis, it may also impose an overhead cost on users each time they try to assess the value of their data against the cost of retaining this data.

9. Data Migration Decision Support Tool

As part of our research, we created a spreadsheet-based decision support tool to help gauge the impact of data migration decisions. Initially, our intent was to use the tool in conjunction with interviews of storage managers. However, we determined that the tool could be refined and distributed as a prototype tool to help inform cost and risk trade-off decisions during the data migration process.

The objective of the tool is to simulate the financial impact of moving a specific quantity of data between any two storage tiers. The tool reports cost savings if data are moved to a lower tier or the additional costs

¹¹ Tallon, P. Surviving the Data Deluge: A Framework for Understanding the Dynamics of Information Management Costs, Communications of the ACM, February 2010.

if data migrate to a more reliable or expensive tier. In each case, costs are automatically adjusted to reflect an increase or decrease in risk. Moving data to a higher, more expensive tier will increase costs but it will also reduce risk. Moving data to a less expensive tier can create cost savings but it can increase risk. The key to implementing the tool in practice is the ability to associate a specific risk value with each tier (for the sake of illustration, in the figure below, we label tier 1 as having a risk level of 100; lower risk values signify higher risk although lower tiers also exhibit lower data management costs on a per gigabyte basis).

For the purposes of this paper, we do not present a complete description of the design and use of the tool. We are completing a full exposition in a companion document that will include a copy of the tool (Excel spreadsheet) and a sample case. Interested readers should contact the authors for more details. This is part of an ongoing research activity for which we plan to provide periodic updates. Future research will help to develop a more precise quantification of risk that will lead to better risk-adjusted cost estimates within the model. Future research may also help to automate key aspects of the model, allowing managers to see at a glance what data needs to be moved and the ideal destination tier for that data.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|----|--|----------------------|-------------------------------|--------------|------------|-------------------------------|---------------------------------|----|--------------------|---------------------------|---|---|
| 1 | NOTE ALL COST AND UTILIZATION DATA ARE PURELY FOR ILLUSTRATIVE PURPOSES ONLY | | | | | | | | | | | |
| 2 | | | | | | | | | | | | |
| 3 | | | Volume of data to move (GB) = | | 500 | Impact on storage costs: | | \$ | 44,737 | | | |
| 4 | | | From what tier? | | 1 | Risk adjusted impact on cost: | | \$ | 25,139 | | | |
| 5 | | | To what tier? | | 2 | | | | | | | |
| 6 | | | | | | | | | | | | |
| 7 | | Baseline TCO metrics | | | | | | | | | | |
| 8 | Risk Unit | Service Tier | Utilization (GB) | Costs | TCO per GB | | Projected Costs after migration | | Cost per Risk Unit | Cost per Risk Unit per GB | | |
| 9 | 100 | 1 | 10000 | \$ 3,000,000 | \$ 300 | | \$ 2,850,000 | | \$ 30,000.00 | \$ 3.00 | | |
| 10 | 95 | 2 | 9500 | \$ 2,000,000 | \$ 211 | | \$ 2,105,263 | | \$ 21,052.63 | \$ 2.22 | | |
| 11 | 92 | 3 | 8000 | \$ 1,500,000 | \$ 188 | | \$ 1,500,000 | | \$ 16,304.35 | \$ 2.04 | | |
| 12 | 90 | 4 | 10000 | \$ 1,800,000 | \$ 180 | | \$ 1,800,000 | | \$ 20,000.00 | \$ 2.00 | | |
| 13 | 88 | 5 | 20000 | \$ 2,500,000 | \$ 125 | | \$ 2,500,000 | | \$ 28,409.09 | \$ 1.42 | | |
| 14 | | | 57500 | \$10,800,000 | \$ 188 | | \$ 10,755,263 | | | | | |

Note: the above example assumed a five-tier storage infrastructure. The model can be adjusted to reflect any number of tiers including tier zero (solid state storage) and the possible additional of a storage cloud.

10. Conclusion

As firms continue to experience a near-exponential increase in their rates of data collection, retention, and processing, vendors have moved quickly to bring a stream of hardware and software innovation to bear on IT management problems posed by this data explosion. One approach introduced by the storage industry, information lifecycle management (ILM), is a way to manage data over its projected useful economic life.

The challenge of ILM, as many firms found, is that information value and the time period over which data must be retained are often indeterminate. Yet firms must make key decisions on when, where, and how to move data across storage tiers to better control cost. Moving data economically through its lifecycle is the main objective of data migration. As outlined in this brief report, data migration is technically addressable but the business rules and economics governing migration are far from settled. The key to any successful data migration exercise, according to our research, is a careful examination of risk, costs, and information value (value-at-risk). Information is increasingly seen by firms as a strategic asset. Managing that data is no less strategic and yet data migration has long been regarded as a non-strategic tactical task. Increasing rates of data growth are forcing firms to confront the challenge of data migration and how to optimize the risks and costs of both migrating and retaining that data. On the basis of our interviews and case analyses, we believe that data migration presents an opportunity for firms to realign their data risk and management costs in a more effective manner.