Data Growth Challenges Demand Proactive Data Management

Merv Adrian, Principal, IT Market Strategy
www.itmarketstrategy.com

Executive Summary

Data growth is one of the most significant challenges – and opportunities - facing organizations today. New forms of data, the retention of additional detail for analysis, increasing pressure for high availability of more of the data, regulatory and security requirements – all are contributing to substantial rises in data volume. Organizations that learn to leverage the explosion of important new information sources can gain dramatic competitive advantage as others lag. Unfortunately, the budget for the acquisition of additional storage is not easy to come by, especially in the current financial environment. Worse yet, over its life, managing storage effectively may cost 3-10 times what it cost to procure. Other costs derive from data growth's impact on systems performance and organizational governance.

Regrettably, few firms have a coherent, planful approach to understanding and managing data growth. A recent IBM CIO survey showed that only 15% believed their data to be well, and comprehensively, managed. Management may be unaware of the magnitude of the volume problem, or the degree to which it will continue to accelerate. Many new hardware and software products target the issue, but without an effective assessment of information requirements, a clear understanding of costs to meet them, and a plan to assure that funding, firms are at risk of making hurried, suboptimal procurement decisions at the last minute in the most costly way.

More data – the right data - can translate to better decision making, improved processes, enhanced communication, and the rich context that enables improved analytics to drive better corporate performance – key ingredients for competitive advantage. Investing in improved exploitation of data is at the top of many companies' priority list, but rarely is that goal accompanied by a strategy and plan to identify, gather, protect and assure the availability of the right data. File-based data from applications like medical imaging, collaboration stores and routine archives is exploding and often ungoverned: IDC recently pointed out that over 50% of data creation is in file-based formats, and that in 2008, for the first time, file-based storage capacity shipped exceeded block-based capacity. Yet file-based storage is rarely included in data stewardship thinking, which remains dominated by "conventional" existing applications. In this discussion, we use "data growth" to refer not only to structured, database-resident information, but that huge volume of file-based information as well.

Executives must recognize that data growth impacts corporate performance – of people and business processes as well as technology – and require IT to deliver a clear vision

that addresses this challenge. Technology managers may be relying on traditional metrics and hardware budget processes. But CFOs are recognizing the centrality of data management in corporate risk mitigation, and expecting IT to deliver more nuanced reporting. Organizational innovations are needed, along with a process orientation that creates "conscious data management" outside of IT. A culture that inculcates this consciousness will communicate priorities, identify roles, and provide sponsorship that assures participation by stakeholders.

What Drives Data Growth

It has become a cliché that the growth of data is inevitable and unstoppable. IBM research indicates that the average company keeps 20-40 duplicates of its data. The potential payoff from managing this proliferation is substantial; many growth drivers may be obvious, while others are hidden from typical observation. Key growth drivers beyond the simple act of performing immediate, everyday business activities – transactions, business documents, emails, etc. include:

- **Application and test data proliferation**. Multiple products that serve the same purpose often have their own instances of data in specialized formats. Mergers and acquisitions with legacy products, line of business initiatives introducing new products, version migrations of existing products, new architectural initiatives, and added delivery channels such as web and mobile devices all create new, often redundant data. Some test and development data sets "never go away," even though they could be generated on demand and deleted when work is done.

- **Analytic datastore propagation**. It has become routine in many shops to create new derived data instances every time a new analytic application is built: analytic data marts, data cubes, synthetic data sets for predictive modeling, etc. Many of these are required by analytic tools' need for specific formats, but a surprising number are not.

- **Regulatory data demands.** New information requirements are driving the retention of historical data. This is hardly new – but like many other initiatives, regulatory reporting projects rarely recognize prior history. Multiple systems remain, even if inactive, and the "feeder" processes that create and store data for them are never shut off.

- **Customer and partner connections** get their own data to support self-service by customers and partners. Often the path of least resistance for meeting these needs is to create a redundant subset data store to minimize exposure, even though an effective authorization and authentication strategy might have made this unnecessary, and performance impacts were not substantial enough to require separation.

- **File system growth** represents another set of challenges; Gartner estimates that 85% of data is "unstructured." Growth of file content outside traditional databases is typically predicted to track database growth closely. In a recent Enterprise Strategy Group study, 36% of surveyed organizations predicted over 41% annual growth in email archives alone.

Data Growth Drives Cost Challenges

Simply storing “more data” is costly. As data volumes grow, costs rise year after year. Some costs of data growth are obvious: storage hardware is a perpetual, inexorable cost. Annually, external storage hardware accounts for tens of billions of dollars in capital expenditures. Andrew Bartels of Forrester estimates that storage hardware consumes about 9% of the typical hardware budget. Growth continues to outpace improvements in storage hardware. Other technology costs include power and cooling expenses, and the cost of space – collectively projected by some to increase by a factor of three in the next 5 years. An “average data center disk drive,” says IDC, drove $36.29 in power and cooling costs in 2008 – and for data centers built a decade or more ago (that is, most of them) more power, and even space, are at a premium if available at all.

Less obvious is the administrative cost. Every database license costs money. The software tools used for managing (not using) data – for backup and recovery, archiving, tuning, design, data movement, master data management, and more – are a substantial additional expense. Forrester’s Bartels estimates that software for storage management consumes 7% of the average software budget. These tools are often from multiple vendors and have overlapping functions and differing skill sets. Add to that the people costs – some of which involve scarce skills – and the picture becomes even more troubling. The larger the data volume, the larger the staff dedicated to managing it – and experienced, skilled staff are a huge cost element.

Even in the difficult technology spending environment of the 2009 recession, A Gartner study showed that 57% of surveyed companies in mature markets, and 75% of those in emerging markets, expected that their 2009 storage spending would represent an increase over 2008 levels. It’s not clear whether the larger percentage expecting lower storage spending in mature markets (43% vs. 25%) was a result of the greater impact of the recession or of better practices in data management, but little evidence for the latter has been presented anywhere.

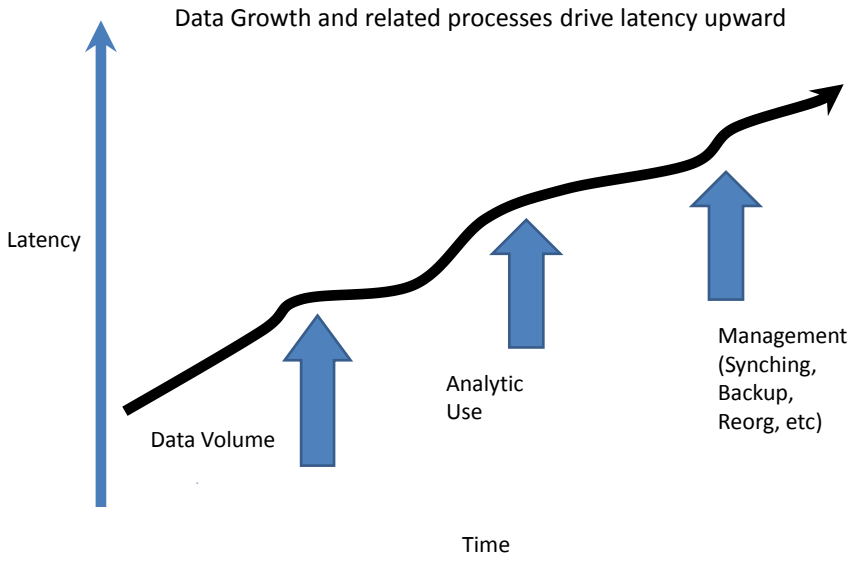
Data Growth Drives Performance Challenges

Business objectives depend on the effective performance of the applications that support business processes. And that performance is contingent upon data availability, and impacted by data volumes. It’s no accident, then, that disciplines for ensuring the availability of data and estimating its volume are always included in the design and deployment of the most significant business applications. But paradoxically, this approach has led to fragmented ownership of data management in many organizations. IT’s move to a services-based management philosophy has often failed to extend to a common set of data availability mechanisms across the application portfolio. Shared storage, shared backup and recovery management, and a vision for appropriate hierarchies of storage technology dependent on business needs are often absent – different business units, or IT functions, may have conflicting priorities.

The performance of applications themselves is impacted by the sheer volume of data being processed. Transaction processing and business analytics performance are both used as measures in industry benchmarks based on data volume. The increasing demand for analytics in real time, and for scenarios and models that span greater lengths of time to improve trending accuracy, modeling and prediction make it likely that

the volume of data to be processed will continue on its growth path far into the future. As multiple stakeholders with very different needs build multiple applications that reuse the same data, the path of least resistance has been to create redundant copies for them – which then must be kept synchronized, impacting accuracy and availability. Those processes themselves place additional stresses on performance.

Additionally, the time available to back up application data never expands as volume does, and each new application creates new data streams that must be managed. If anything, as organizations grow and become global, their available backup window often shrinks, or disappears entirely as business in every time zone requires data to be available on demand. Strategies to mitigate these challenges often have performance impact as well.



Data Growth Creates Governance Challenges

While many other IT resources can be treated primarily as physical assets, data carries risks that are profound and multifaceted. The governance of data, especially in areas such as security and compliance, is typically fragmented. It may be handled on an application-by-application basis: it may lack clear ownership even at that level. At a corporate level, policies for managing compliance and privacy consideration may exist on paper, with no clear organization chartered and empowered to administer them. A 2008 Forrester Consulting study commissioned by IBM found that in a majority of 21 companies interviewed at length, activities across the data life cycle took place in isolation, with little effective communication of design and development artifacts, policies, and standards from one stage to another and little automation planned to

improve the situation.

Many of the systems designed to support today's compliance requirements, such as those mandated by Sarbanes-Oxley or Basel II, have been implemented without a view of the bigger picture within the organization, or a plan for rolling off data no longer mandated for "near-time access" and reporting requirements into archives. Data growth compounds such problems, which occur throughout the organization, and the longer they remain unaddressed, the more challenging, and costly, it will become for an organization to tackle them.

A key missing step in tackling this is the identification of stakeholders, who must be understood, empowered and managed. People who play a role here include the players in the development, deployment and management of applications: Business Analysts, Architects, Developers, Testers, DBAs, and System Administrators. But their constituents, the application owners, must be represented as well - on an ongoing basis, not only at design time. Most important, all of their work needs to take place within a context defined by the organization's Legal staff, its Security owners, and ultimately the executives to whom they all report.

Few companies have had the foresight to design data governance processes into their portfolio; most have been too busy dealing with their immediate needs to step back and create policies and processes to guide it all. Instead, most "back into" a data governance strategy, if at all, only when repeated crises occur**.** And occur they will. But even in the absence of a more holistic data governance initiative – which is highly recommended – proactive steps may be taken to tackle the issues of data growth. In so doing, plans will be created that lay much of the groundwork, and broader governance initiatives can leverage these when the time and the funding are right.

The Five Stages of Proactive Data Management

Getting a handle on managing data growth parallels the stages of a broader data governance journey. IT Market Strategy believes that most sizable IT shops fall into 5 stages of maturity in managing data growth, and that while work often proceeds on several things at once – and usually must do so – it is critical that organizations ensure that they complete the early stage, foundational requirements if they hope to be effective in later ones.

Stage 1 – Siloed data awareness

Entering this stage, few organization-wide data priorities, for growth or anything else, have been established: data is typically in silos that correspond to applications, or business units. Analysts, programmers, and administrators "work with what's there" in a purely reactive fashion. To achieve awareness, discovery and profiling tools are used to identify and document access to and relationships across all data sources, but execution tends to be opportunistic rather than planned. Data growth may be documented at this stage, but the information necessary to determine what to do about it is missing; redundancy, for example, may not be specified, and is almost certainly not planned.

To move to the next step, a data stewardship organization must be chartered, funded, and supported, with a mandate to deliver a documented architecture for data distribution and availability, together with a metadata layer that defines common semantics. The breadth of this effort might be somewhat constrained and limited to data growth "problem areas." Even if this is the case, the skills and metadata constructed will form a base for broader efforts at data governance.

Stage 2 – Measured data value and monitored usage

Building upon the architectural definition arrived at above, the organization now has a way to measure data value in the simplest terms – by the dependence of key business processes upon it, which also helps define risk (although it is not sufficient to define all data risk). Monitoring its usage delivers a similarly rudimentary assessment of who the stakeholders are. Some design principles can now be propagated; specific data strategies can be tied to the execution of known business process and challenges in optimizing them. For example, data warehouses and data marts may have been built to enable specific BI requirements. Everything that might be useful is still being kept, and data growth remains an observed, but not managed, phenomenon, although its sources, stakeholders, and relative value of the data concerned have now been measured.

Stage 3 – Managed data quality and security

Data Quality management, information security standards for access control, privacy definitions and risk mitigation to protect data drive the final foundations for proactive data management.

With governance principles now established that conform to regulatory and compliance standards, the stakeholders identified, and policies established, data retirement and usage-based optimization can be examined as the next step in the storage strategy. Which data governance processes frequently require overriding source application activities (frequent security violations, data transfers that exceed batch windows, disks that exceed capacity standards)? Determining whether it is the rules, the data itself, or both that causes these challenges will yield key insights into data growth issues.

Stage 4 - Lifecycle awareness meets data stewardship

Distinguishing between those data assets, applications and infrastructure to be retired simply based on aging, and those which can be archived based on policies and their contribution, begins here. Policies for information collection, use, retention and deletion - the key disciplines for successful information lifecycle management (ILM) – drive many more initiatives than just the management of data growth. But it is the latter that will get the most support, be least disruptive, and deliver the most value if it is done in the broader context. At this stage, then, the data growth initiative is tightly coupled to assisting the ILM strategy. But one more step remains.

Stage 5 – Optimized data management based on business performance needs

Explicit linkage now permits a focus on what information is stored and why – based on what key business decisions it supports. Trust in the data is dramatically enhanced by the propagation, measurement and socialization of policy-based data management. Aggressive deduplication, compression and elimination strategies measure and optimize the use of data. Customized test data is generated on demand and deleted when no longer needed. Data growth is measurable and drives planned procurement strategies. Questions asked and answered include: What information is required to support the decision? How accurate does it need to be? What's the most efficient process for collecting, generating, and supplying the information? In what timeframe does the information need to be supplied?

Solution: Proactive Data Management Controls Data Growth

The complex interactions of governance, performance and costs noted above are compounded by unregulated growth. Without a plan, clear ownership milestones, and metrics, organizations will find themselves struggling as data growth overwhelms their technology, processes and people. The solution requires organizational change, policy development and a programmatic basis for enforcement. Training, including training for business unit managers who "own" the data, will provide a key way to bind the business and IT into a new model.

To take control, organizations must begin with the conscious decision not to let data growth proceed unexamined. Moving through the five stages requires three key action steps in each stage transition:

Assess – Take a clear-eyed look at the challenges by documenting the problem and its costs. A descriptive inventory is a start. Use profiling and discovery tools to identify dependencies, redundancies, ownership, stakeholders, access patterns and exposures, but don't stop there. Build a justification for the tools and practices needed. Progress will require investment, so document the costs of the current situation and the potential benefits of scenario-based reductions in data growth. Quantify the amount of reduction possible, and calculate the benefits to be achieved by eliminating unneeded and redundant data and archiving data not needed in expensive on-line and nearline facilities.

Plan – Choose from the scenarios developed and use them as an opportunity to engage the constituency you don't talk to now. Present the analysis to them and engage them in the next step. The survey of "what is", created in the Assess stage, is not sufficient. A vision of "what is needed" must accompany it, and it can only be built with their participation. The newly-identified stakeholders must be engaged in the "what should be" process.

Redefine data needs in terms of business execution and optimization requirements – alternatives will exist, such as those based on differing retrieval speeds in hierarchical storage architectures. Which data must be available immediately? Which can wait until the next day, or month- or quarter-end? Document these and create a catalog/metadata foundation – a more full governance effort will capture much richer metadata, but even this fairly limited view will be reusable when that effort begins. Establish metrics for data volume and growth and create processes for tracking them against expectations. Build and staff the function required for this effort and train the people who will be in the roles

defined. Create a “data steward certification” step for new data store creation and align it with procurement practices.

As the consensus about the scope and cost of the immediate reduction effort is built, create the tactical plan for its execution: storage virtualization launch or extension, retirement of old storage hardware where appropriate and reuse where possible, new storage procurement where needed, data deletion, data deduplication and archiving.

Execute – The rest is quite simply about getting it done, using a variety of available tools and technologies (see below, *Technologies, Techniques and Tools for Proactive Data Management*.) It is key to ensure staff is on board and fully trained, milestones have been defined and reporting to stakeholders is followed up on. Regular check-ins with stakeholders, whose data has been moved, for example, will ensure that performance expectations are being met. Discussions with budget planners will ensure no snags occur in procurement – report on savings achieved, planned purchases reduced, etc.

More effective management of data will improve corporate performance in multiple ways: reducing new expenditures; leveraging existing ones; improving the performance of systems that will become less overloaded; and mitigating security and compliance risks. Developing the process orientation and infrastructure that drives a conscious, proactive approach to managing data growth is a key step in this journey, and one that will drive benefits far beyond the investment it will require.

Technologies, Techniques and Tools for Proactive Data Management

Every data management professional needs tools. The list below covers some of the key ones, with some examples of IBM offerings.

Software Asset Management: Application instance reduction is a powerful tool; rationalizing the application inventory can reduce data. Example: IBM Optim Application Retirement Solution

Archiving Application Data: on application retirement, some archiving solutions allow applications and BI tools to continue to retrieve business objects directly from archive storage without a "restore" into an additional persistent copy. Examples: IBM Optim Data Growth Solution, IBM Smart Archive

Data Profiling and Discovery tools – can help automated discovery to uncover dependencies that may not be obvious. These tools can be the source of the metadata that forms the basis of the plan. Example: IBM InfoSphere Discovery, IBM Information Analyzer

Metadata Catalog. This drives non-application access, and in addition to serving as the source for definitions of record, can document policy for "shredding" data that can be discarded completely, as well as compliance, privacy and access. A full metadata strategy includes: collection, classification, maintenance, and may overlap significantly with MDM. Ensure that your chosen tools integrate with your chosen metadata repository, or that a clear roadmap for future integration is defined. Examples: IBM InfoSphere Metadata Workbench, IBM InfoSphere Business Glossary

Master Data Management can often provide a subset of the broader Metadata Catalog requirements; this subset may be adequate for managing data volume and reducing data growth in key application areas, but will not provide a fully comprehensive solution, since much data is outside its purview. Example: IBM InfoSphere Master Data Management Server

Consolidate ECM systems – Gartner reports companies run between 5 and 20 separate ones. Specialized ones can be subsumed; for example, IBM's ECM product line includes an SAP-specific option: IBM CommonStore for SAP.

Data Compression is used inside many database products to varying degrees. Upgrade DBMSs to utilize the latest features for compression. Example: IBM DB2 Storage Optimization for DB2 9.7. For file-based systems, data compression may be achieved with software and with storage appliances which include sophisticated data compression capabilities.

Upgrade Storage Hardware. Ensure that your data centers utilize denser, more efficient hardware – the overwhelming majority of data centers were built more than a decade ago and may be able to save space, electricity and other operating costs with an

investment in new hardware. These savings may considerably offset the new acquisition costs, especially with vendors who offer favorable upgrade trade-in policies.

Deduplication Software is often used within hardware appliances to eliminate redundancies in data sets. Deduplication ratios can be dramatic. In network-attached storage environments customers of some deduplication products claim they can achieve as much as a 7:1 reduction; in backup sets this rises to over 20:1, and in VMware and similar virtualized environments by as much as 100:1. “Single instance” strategies, used with file-based systems, eliminate additional copies. These are used in archiving and some enterprise applications support them directly. Example: IBM ProtecTIER

Hierarchical Storage Strategy: a plan to design tiers of storage based on business usage profiles for prioritization. Software for implementing this approach utilizes lower cost hardware (most often existing older storage devices) for lower priority data. Example: IBM Tivoli Storage Manager HSM

Storage Virtualization can catalyze a series of efforts to eliminate duplicates – if there is an access layer that connects users to all the data they need, additional separate copies may be eliminated, reducing duplication. Examples: IBM SAN Volume Controller, IBM Tivoli Storage Productivity Center.

Test data management solutions provide “right sized” sets of data for non-production use, to eliminate multiplying every gigabyte of data in a production database by the number of replicated copies. Example: IBM Optim Test Data Management Solution