

Data Retention: The Unrealized Opportunity

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

Big Data retention is an inescapable, inevitable fact of life for more and more businesses today. The explosion in data volume is well documented, and is so severe that its challenge to our capacity to retain it is felt even at the macro level: in its Feb. 27, 2010 issue the Economist magazine cited IDC findings that the storage industry was not able to produce enough physical storage capacity to keep up with the anticipated volume of data. In the past, deciding what data to retain online, move to offline tape or simply delete forever was made by the business groups who generated and owned that data. Today's world requires balancing increasingly stringent compliance regulations with the need for "information at your fingertips," which translates to large online scalability and on-demand data that is retained and accessible. Offline archiving is simply not a viable option for much corporate data.

How Big Before it's a Problem?

The grandfather of all Big Data conundrums can be traced back to Google. In some way it set both business and consumer expectations, and demonstrated that large quantities of data can and should be made accessible 24x7, with no lag. When such on-demand capabilities are clearly possible, the term "history" literally refers to data that occurred mere seconds ago.

But not every business needs to have Google-size data retention needs to have a Big Data problem. More and more companies today struggle to cope with their own growing data volumes and continually revisit their infrastructure strategy to make sure they scale fast enough within cost constraints. When coupled with data retention demands driven by compliance mandates, businesses are forced to acquire more physical storage, add servers and of course, software licenses. In other words, the cost associated with keeping data online and accessible quickly become a Big Data retention problem.

Big Data – Burden and Benefit

Upon closer inspection the problems of information growth create multiple challenges. In existing transactional systems, Big Data overwhelmingly consists of historical information (on average 80 to 90% of data in production applications ends up being used for read-only historical purposes), clogging the pipes of transaction processing and analytics by overwhelming the systems that must retrieve records, update them, index them, back them up and report on them. These performance and process burdens mean that in addition to the storage acquisition costs it drives, Big Data forces spending on bigger, faster systems, more staff to manage it, and continued attention that must be paid to innovating architectural solutions.

Some organizations have understandably asked whether all the new data is truly needed, or if the old data can be purged. Many of the files growing inside and outside “production databases” contain new types of data: emails, instrument readings, weblogs, audio and video...it’s no wonder many executives ask “*Do we really need to keep all of this?*” The answer lies somewhere between complying with regulatory mandates, avoiding fines, prison time or both and being competitive in an on demand fast moving world. Uses of data are not always evident at its creation, and retaining data over a longer term provides more opportunities for business insights and trend analysis.

The Rise of the Machine Data

Automated data creation is driven by the growth of connected devices, the complexity of transactional environments, and new demands from governmental agencies for the documentation of activities in a variety of human endeavors. In telecommunications, billions of call data records (CDRs), i.e., the logging of telephone calls between parties, are created every second of the day. The growth of mobile devices has certainly contributed to the explosion of data volume but it’s the billing systems for telecom providers, and their need to create complex demand models, marketing offers and analytics assessments of their system performance that drive retention of these records for long periods of time. Compliance and other governmental requirements such as lawful intercept drive the required data retention window out even further.

Financial institutions have similar challenges: they continuously execute automated transactions, of different types, in more multi-faceted relationships with customers of increasingly complex structure. In a global environment with a bewildering and often contradictory mix of varying regulatory demands, these forces require data duplication to respond in a timely way, further compounding the problem through a simple multiplier effect.

In healthcare, patient and electronic health care records now must be retained and accessible on demand, for the life of the patient. Additionally, patient-attached instrument data is being stored and used for complex correlation analysis in real time, for modeling analytics in predictive settings to optimize health outcomes. All these retained records, it is hoped, will ultimately drive better quality care. Additionally, the healthcare industry has its own set of regulatory and financial records to manage, and process efficiencies which can be realized through effective record retention strategies. Other industries such as oil, energy, utilities, and retail all have similar patterns and their own regulatory data retention needs. The need for retrieval and re-use of historical data is clear, and the period of retention cannot be arbitrarily set without great risk. Regulatory requirement changes have extended the retention period – and increased retrieval requirements – on a regular basis over the past few decades. At the same time, as previously mentioned, analytic capabilities can make effective use of historical data for better and better predictive modeling to drive top line competitive opportunities.

Retention and Reuse Are Not Synonymous

As organizations seek out cost-effective solutions for data retention, it’s critical to ensure that older solutions be looked at with a critical eye, even if they are well-established and seem less expensive than some newer alternatives. Business and compliance needs vary now, and will in the future, for the data being stored. A “write once, read (almost) never” (WORN) medium like tape, while suitable for linear reads of data that may be needed very rarely, will quickly prove unsuitable if a new usage – whether driven by compliance or analytics – suddenly arises for

random, ad hoc usage. In such a scenario, the data will most likely still need to be read to disk so that it can be used interactively. Retention is not reuse.

Simple data duplication – into full redundant copies or subsets for specific uses – like data warehouses and data marts – is very typical. But it requires expensive hardware, at increasing rates of growth. And it often results in keeping unneeded data, far exceeding the functional requirements or lifespan of the originally developed system. Organizations must balance cost and performance with usage requirements, retaining data in a usable medium, while keeping it in a format that allows reuse without heroic measures, new tools and massive expenditures. In other words, the data needs to be *compressed, available, usable, and manageable by policy*.

The RainStor Solution

One such solution is available from Rainstor, a data retention provider that delivers an OEM software repository solution for ISVs, SIs and managed service providers (MSPs). Rainstor can achieve compression at rates ranging from 40 to 100:1, driving enormous savings in storage costs. Sophisticated de-duplication precedes compression of data values into a coded, link-based set of values that are stored once, driving huge reductions in the size of data that repeats the same value often – like common names, phone numbers, customer IDs, etc. Automation – based on configurable retention rules – is key to making all this happen in a cost-effective way, to significantly reduce the expensive people resources component, which is another large component of any retention solution.

RainStor allows seamless access to data through existing SQL statements or BI tools that rely on standard ODBC and JDBC protocols. As a result, staff assigned to work with archived data operates in a familiar environment, and use existing applications which run mostly unchanged. Data descriptions (table definitions, for example) are stored, and can reflect changes so that retrieval can be aligned with a point in time. Changes - new tables and fields in existing ones - are captured along with the dates they took place.

Finally, Rainstor does not demand specialized hardware: it works with commodity storage in low cost, network- or direct-attached storage and cloud environments. Rainstor is a compelling solution to the opportunities for Big Data retention and reuse.

Many of RainStor's ISV partners (who OEM the solution) report significant ROI to their end-customers who can free themselves from the continuous cycle of procuring physical storage, specialized hardware and specialized personnel to "baby sit" systems with massive quantities of historical data. RainStor provides a complementary repository to the production application RDBMS or data warehouse, operating side-by-side. In the case of machine generated data (e.g. CDRs, logs), where the data is immediately historical, RainStor has proven to be a better and more cost-effective primary production solution.

ISVs, SIs or MSPs that offer applications or services generating or managing large quantities of data should expect that their end-customers will be looking to them to provide a solution to their Big Data problem. For them, the Big Data retention burden presents itself as an opportunity, and a solution which can be leveraged to deploy additional revenue-generating offerings and/or to lower operating costs for a hosted or managed service.