Big Data Trainwreck Ahead!

GETTING A GRIP ON THE BIG DATA PROTECTION AND MANAGEMENT CHALLENGE

The Big Data train has already left the station and enterprises, seeking higher meaning and competitive intelligence from their data investment, are jumping on board. As data sets continue to grow and the value of extracting meaning from Big Data is proven further, IT and business analysts might start to feel they have Big Data handled. But there is a train wreck looming and it comes down to this: Big Data is not being protected with the same careful processes put into place for ‘regular’ data sets and Big Data presents data management and other operational challenges never experienced before. Add to that the need to execute on rules and governance at the same level as regular data compliance, and it’s no wonder the Big Data train is headed for trouble.
PBS series host Robert X. Cringely, in detailing the tech evolution leading up to Big Data, assigns its contribution to everything from self-driving cars to the success of Amazon, a pioneer in collecting consumer behavioral data and capturing consumer intelligence via the Internet. As Cringely puts it, Amazon’s ability to track a consumer’s recent activity was the beginning of Big Data. When you visit a shopping site today and they ‘know’ your preferences you can thank Big Data.

At the heart of Big Data is intelligence, or as Cringely describes it, “the accumulation and analysis of information to extract meaning.” Gartner defines Big Data as “…high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing for enhancing insight, decision-making, and process optimisation.”

Enterprises today commonly use Big Data for business intelligence initiatives and Big Data shows no signs of slowing down, with estimates that data production will be 44 times greater in 2020 than 2009. Other estimates show an additional 2.5 quintillion bytes of data is being generated daily.

The inherent challenges of Big Data are only going to escalate as enterprise investment continues and venture capital deals are still funding Big Data plays, if at a slower pace. Gartner predicts the advanced analytics market to grow 14 percent this year to drive $1.5 billion in spending, part of a two-year surge in which the firm expects 75 percent of enterprises to invest in Big Data. E&Y estimated Big Data analytics drove an estimated 70 percent of merger and acquisition deals during Q1 2016. Clearly, the Big Data train has already left the station.

To avoid being derailed by this train, gathering momentum, enterprises need to tackle the issues of data protection and recovery now, and the issues of data management and operation. A key point is that Big Data does not get a pass on compliance. Big Data is being subjected to the same rules and governance needs as more traditional data loads, making secure protection and recovery even more critical. This is applying new pressures to find solutions that are Big Data aware so that automated disaster recovery and enhanced levels of visibility can be achieved for platforms including Hadoop, Greenplum and GPFS.

DATA PROTECTION AND RECOVERY

Enterprises have avoided applying data protection and disaster recovery to their Big Data environments due to the sheer size and complexity, not to mention the cost. But if $1.5 billion is being spent on Big Data analytics it makes good business sense to securely protect these massive volumes of information. Even in the general data picture, backup is inadequate in most organizations. Compliance is another factor. Lack of Big Data protection can put enterprises at risk of failing to meet data governance requirements.

---

1 PBS, “Thinking about Big Data – Part One,” July 2016
4 CloudTweaks, “Facts and Stats about the Big Data Industry,” March 17, 2015
5 Gartner, “Predicts 2016: Advanced Analytics are at the Beating Heart of Algorithmic Business,” January 28, 2016
6 EY, “Technology M&A stabilizes in Q1 after reaching record-high value in 2015”, May 2016
7 ibid
In planning to secure, protect and recover Big Data, here are a few things to consider:

- **Big Data Aware:** Integrating the protection of large data sets into your existing data protection infrastructure will require a solution that is truly Big Data aware, that can provide automated disaster recovery and enhanced levels of visibility into the leading common Big Data tools including Hadoop, Greenplum and GPFS.

- **Prioritization:** Similar to more traditional data backup and restore, IT needs to scrutinize what information is contained within Big Data volumes and what should be protected, what is mission critical, and what has compliance impact. Machine generated data, for example, can be reproduced and may not need to be backed up and recovered.

- **Recovery Objectives:** For those Big Data sets deemed a priority to protect and restore, IT needs to set a Recovery Point Objective (RPO) and a Recovery Time Objective (RTO), the point in time at which data was last backed up, and the amount of time you’re willing to wait for restore.

- **Scan Time:** Scanning files every time IT runs a backup is an insurmountable task in a Big Data environment. One solution is an object-level converged process for collecting backup, archiving and reporting data. The data is collected and moved off the primary system to a virtual repository for completing the data protection operations. Once the scan is completed, an agent can be placed on the file system to report on incremental backups, making the process even more efficient.

**DATA MANAGEMENT: GETTING A GRIP**

Enterprises are starting to see the need to do a better job of Big Data backup and recovery to realize the full potential of their investment. Right now, what’s also keeping Big Data from full utility is the management and operational challenges of harnessing and organizing these large volumes of data.

Big Data can be managed thanks to technological advancements and strategic practices designed to bring efficiency to what otherwise truly would be a train off the rails. Here are a few key ideas:

- **Parsing the Nodes:** Big Data is not a unilateral entity. To make use of it, the data is structured into nodes, and enterprises are running multi-nodal systems as a regular practice. But not all nodes are created equal. There are solutions available to query these nodes, in essence find what information they house, then make smart decisions, per node, on retention, recovery and disaster recovery. This capability is key to efficient use and storage of Big Data and enables enterprises to assign levels of importance to Big Data just as they would to regular data sets.

- **Vendor Sprawl:** Scattered data protection and a data management architecture made up of disparate systems are inefficient in a traditional system. Add managing Big Data multi-node systems to the mix and you have a brand new level of inefficiency and budget waste.

Estimates report that data production will be 44 times greater in 2020 than 2009.

*Forbes, February 2015*
IT can look at the areas of possible consolidation, including what is needed to meet Big Data service level objectives, and determine if there are operations for a leaner, more streamlined data management infrastructure.

- **Data Portability**: Big Data that is deemed mission critical needs rapid recovery but with seamless data portability, to the same or new infrastructure – whether it be cloud, on-premises, virtualized, traditional or converged.

- **Managing at Scale**: While Big Data is big, it’s getting bigger. Enterprises need a solution that can scale with a changing environment, one that can support hundreds to thousands of nodes, and offers smart integrated data protection, rapid recovery and parallel performance. Big Data sets can also be replicated into a public cloud for cost effective storage at scale and for business continuity in the event of a disaster.

**BE THE CONDUCTOR OF YOUR DATA**

Big Data is truly a train that appears unstoppable. It is up to enterprises and IT professionals to budget some time and resources now to examine the data protection and data management systems in place to ensure that data’s inevitable growth does not plunge the enterprise further into inefficiency. Most enterprises are more interested in the accessibility of actionable data rather than the extent of the datasets. Proof, once again, that enterprises must have an infrastructure in place that indeed parses the nodes, delivers data portability and simplifies the use of data. Once these have been achieved, you can conduct your Big Data enterprise, full steam ahead.

To learn more about how Commvault® will help you protect and manage Big Data, visit [commvault.com/solutions/by-topic/protect-big-data](http://commvault.com/solutions/by-topic/protect-big-data).