

# Big Data and the New Storage Architecture

November 2011



**A WHITE PAPER**

by Mike Janson

*CTO Strategist, National Storage*

*OnX Enterprise Computing*

# Contents

3	Big Data
3	Roles in Big Data
4	Who Has Big Data
4	Sources of Big Data
5	Complexity of Big Data
5	New Architectures to Store Big Data
5	<i>Scale Out vs Scale Up</i>
6	<i>Benefits of Scale Out</i>
7	Summary

# Big Data

There is an emerging topic in the information technology arena that is getting a lot of press these days and that is Big Data. Big Data refers to the amount of data companies are dealing with in today's fast moving technology environments—100s of terabytes to petabytes—generated from legacy applications as well as today's newer web-based processes. This data creates new challenges for the IT groups of most all companies and has ushered in a new wave of products and technologies to deal with it. These challenges include storage, management, analysis and correlation to name a few. This paper will touch briefly on these topics but focus on the storage aspects.

## Roles in Big Data

There are several roles involved in working with big data, some are traditional IT roles from the past and some are new in an effort to deal with the amounts of data. The following is a sampling of some of these roles and the challenges they face.

**Storage Administrator** – Storage Administrators have traditionally dealt with a growth of data on the order of 20 to 30 percent per year depending on the type of company and regulations or guidelines for their industry. Many of today's storage administrators are dealing with a data explosion, information is doubling every 18 - 24 months, and it is estimated that data will grow 50 times in the next decade. This is putting a lot of stress on their ability to keep up with the demand and getting budget to grow the infrastructure to house the data.

**Data Administrators** – Data Administrators are tasked with keeping track of the data as it is being created and making sure it is moved into the analysis environment. This was traditionally done in relational databases but today's growth is outstripping the ability of standard row based databases and Structured Query Language (SQL) to process in a timely manner.

**Data Scientists** – A Data Scientist is a fairly new role defined by Hillary Mason of Bit.ly as someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning who culls information from data. These data scientists take a blend of the hackers' arts, statistics, and machine learning and apply their expertise in mathematics and understanding the domain of the data—where the data originated—to process the data into useful information. This requires the ability to make creative decisions about the data and the information created and maintaining a perspective that goes beyond ordinary scientific boundaries.



# Who Has Big Data

Most companies today are struggling with data growth that is unprecedented in previous years regardless of company type and industry. Many companies have big data but don't understand what it is or how to use it for financial gain or competitive advantage but those who are embracing it are finding new ways to create market expansion, open new markets, or move to market leadership. However, there are some industries that are seeing a more significant spike in data growth than others. A brief sampling of some of these is companies who focus on Technology, Consulting, Ad/Display Serving, Finance, Retail, Healthcare, Social Media, and Energy/Utility. Several of these are obvious where the data comes from but several others need a little more insight—which is provided in the next section.

## Sources of Big Data

There are numerous sources of big data and the types of data they create differ but fall into three generally accepted categories: structured, semi-structured, and unstructured.

**Structured Data** is the type that would fit neatly into a standard Relational Data Base Management System, RDBMS, and lend itself to that type of processing.

**Semi-structured Data** is that which has some level of commonality but does not fit the structured data type.

**Unstructured Data** is the type that varies in its content and can change from entry to entry.

Structured Data	Semi-Structured Data	Unstructured Data
<ul style="list-style-type: none"><li>▪ Customer records</li><li>▪ Point of Sale data</li><li>▪ Inventory</li><li>▪ Financial records</li></ul>	<ul style="list-style-type: none"><li>▪ Web logs</li><li>▪ Social media</li><li>▪ E-commerce</li></ul>	<ul style="list-style-type: none"><li>▪ Pictures</li><li>▪ Video editing data</li><li>▪ Productivity (office documents)</li><li>▪ Geological data</li></ul>

While the amount of data that is user generated, such as images and social media entries, is significant and growing the largest data growth is machine generated. This machine generated data is in the form of web logs, consumer behavior tracking and financial market analysis, to name a few. However, it is not the data itself that is useful but the ability to correlate the data in meaningful ways that is creating the need to keep and process this data.

# Complexity of Big Data

Due to the diversity of the data and the sources from which it comes, the patterns and associations are not clear when looking at the data. Traditional models are used to dealing with text and numbers which lend themselves to traditional database models and processing techniques. Today's big data include multiple object types as listed on page 4 and in many cases it is not the data itself but the metadata—data about the data—that is key. For instance, consider pictures on a social media site, in and of themselves there is no relevant information that a computer can pull from the data stream. But if you look at the metadata, whose account it is on, date, time and location it was taken, facial recognition patterns, etc. and you can start to put together a reference model about the image. These are the challenges facing the big data analytics engines and the data scientists who use them.

**IN MANY CASES IT IS NOT THE DATA ITSELF BUT THE METADATA—DATA ABOUT THE DATA—THAT IS KEY.**

## New Architectures to Store Big Data

These requirements have broken the traditional data storage models and created the need for new architectures to effectively store and deliver this data to the analytics systems that do the analysis. Older storage architectures couldn't scale to the size required or hold the diverse data types that are being created. Limitations on the amount of data that could be stored in an array was in the 100s of terabytes range but the file systems they provided could not scale beyond 16 terabytes. This meant that a user with 50 terabytes of data from a single source would be forced to create at least 3 file systems for the data. This created extra work on the part of storage administrators and made finding all the related data more difficult. This big data requirement has forced existing storage vendors to look at systems in a new way and provided a breeding ground for new startup companies founded by tech savvy entrepreneurs.

### Scale Out vs. Scale Up

These older architectures used the fundamental approach of scale up vs. scale out. The primary difference is how the system uses resources. Scale up system would provide a small number of access points, or data servers, that would sit in front of a set of disks protected with RAID. As these systems needed to provide more data to more users the storage administrator would add more disks to the back end but this only caused to create the data servers as a choke point. Larger and faster data servers could be created using faster processor and more memory but this architecture still had significant scalability issues.

Scale out uses the approach of more of everything—instead of adding drives behind a pair of servers, it adds servers each with processor, memory, network interfaces and storage capacity. As I need to add capacity to a grid—the scale out version of an array—I insert a new node with all the available resources. This architecture required a number of things to make it work from both a technology and financial aspect. Some of these factors include:

- **Clustered architecture** – for this model to work the entire grid needed to work as a single entity and each node in the grid would need to be able to pick up a portion of the function of any other node that may fail.
- **Distributed/parallel file system** – the file system must allow for a file to be accessed from any one or any number of nodes to be sent to the requesting system. This required different mechanisms underlying the file system: distribution of data across multiple nodes for redundancy, a distributed metadata or locking mechanism, and data scrubbing/validation routines.
- **Commodity hardware** – for these systems to be affordable they must rely on commodity hardware that is inexpensive and easily accessible instead of purpose built systems.

### Benefits of Scale Out

There are a number of significant benefits to these new scale out systems that meet the needs of big data challenges.

- **Manageability** – when data can grow in a single file system namespace the manageability of the system increases significantly and a single data administrator can now manage a petabyte or more of storage versus 50 or 100 terabytes on a scale up system.
- **Elimination of stovepipes** – since these systems scale linearly and do not have the bottlenecks that scale up systems create, all data is kept in a single file system in a single grid eliminating the stovepipes introduced by the multiple arrays and files systems required.
- **Just in time scalability** – as my storage needs grow I can add an appropriate number of nodes to meet my needs at the time I need them. With scale up arrays I would have to guess at the final size my data may grow while using that array which often led to the purchase of large data servers with only a few disks behind them initially so I would not hit bottleneck in the data server as I added disks.
- **Increased utilization rates** – since the data servers in these scale out systems can address the entire pool of storage there is no stranded capacity.

**THERE ARE A NUMBER OF SIGNIFICANT BENEFITS TO THESE NEW SCALE OUT SYSTEMS THAT MEET THE NEEDS OF BIG DATA CHALLENGES.**

# Summary

The explosion of data happening in data centers all over the world is creating unique opportunities to create new types of information for businesses and organizations to use. These new quantities of data have broken the traditional storage model and driven the need for new architectures to house this data. Using new methodologies and concepts there are a handful of companies that have created new, scalable and reliable, grid storage systems that can not only meet the needs of today's big data but have the ability to continue to scale even larger as the data explosion continues.

## ABOUT THE AUTHOR:

### Mike Janson

*CTO Strategist, National Storage  
OnX Enterprise Computing*

Mike Janson has over 25 years in the IT industry where he has held various positions in the data center from technical staff to Senior Manager for storage, databases, UNIX and Windows. His focus for the last 12 years has been on storage specializing in architecting solutions for businesses of all sizes, from SMB to enterprise, and overseeing the successful implementation of those designs. Mike holds multiple certifications in the storage industry.

## ABOUT ONX ENTERPRISE SOLUTIONS

Since 1997, OnX Consulting, Inc. has provided advisory services to Oracle, IBM, HP and Microsoft clients needing assistance in managing the fundamental changes in their software licensing assets. This highly successful consulting practice has helped hundreds of clients across Canada and the United States optimize their total cost of ownership.

[www.Onx.com](http://www.Onx.com)