



A Toshiba Group Company

White Paper

Supporting Big Data Applications with Flash-Based Storage

Part 1: Providing a Better Understanding of the Big Data Storage Opportunity

Scott Harlin

| Contents | | Page |
|----------|---|------|
| 1 | Introduction | 2 |
| 2 | Big Data Overview | 3 |
| 3 | Big Data Use Case Examples | 5 |
| 4 | Big Data Analytics | 6 |
| 5 | Benefits Of Flash-Based SSDs For Big Data | 8 |
| 6 | Summary | 9 |



1 Introduction

As IT managers continue to unlock opportunities for flash-based solid-state storage adoption in the enterprise, one of the more prominent application areas centers on Big Data. This application represents a large volume of both **structured** and **unstructured** data that is too big, moves too fast, or exceeds current processing capabilities using traditional database and software techniques¹. It is a massive collection of data from conventional and digital sources, internal and external to a company's enterprise that provides an ongoing source of data discovery and analysis². Big Data represents the voluminous amount of data a company creates but does not refer to a specific quantity. However, tens of terabytes and petabytes are typical for this application.

From an analytical perspective, the ability to examine large amounts of data, as well as a variety of data types, enables companies to uncover hidden data patterns, unknown correlations and other useful information that invariably provide competitive advantages, enable better business decisions and can result in more effective marketing and increased revenue. According to a McKinsey Global Institute (MGI) report, there are five broad ways in which Big Data creates corporate value³:

- It enables information to be transparent and usable at a much higher frequency
- It enables the collection of more accurate and detailed performance information to expose variables and boost performance
- It allows narrower customer segmentation that results in more precisely tailored products or services
- It enables the use of sophisticated analytics to substantially improve business decision-making
- It improves the development of next generation products and services

The first part of this white paper introduces key concepts and characteristics associated with Big Data applications to provide a better understanding of this enterprise storage opportunity. Part 1 also addresses how flash-based solid-state storage fits into the Big Data model.

In a separate document, Part 2 of this white paper includes an overview of OCZ enterprise SSD and software solutions that best address Big Data applications and the ability to deliver ultra-fast processing of large datasets that enable data-driven analytics. The OCZ solutions covered in Part 2 include:

- Intrepid 3000 SATA SSD Series
- Z-Drive 4500 PCIe SSD Series
- Windows Acceleration (WXL) Software
- VXL Virtualization Software
- ZD-XL SQL Accelerator
- StoragePeak 1000 Central Management


2 Big Data Overview



According to IDC market research⁴, Big Data will continue to represent a fast-growing multi-billion dollar worldwide opportunity for the next five (5) years. More importantly, it will transform businesses globally making them Big Data-driven in the process. IDC's definition of Big Data describes a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery and/or analysis.

According to IDC, the Big Data market is comprised of three primary segments:

1. **Infrastructure:** includes external storage systems such as SSDs available from OCZ, as well as server components (internal storage, memory, network interface cards, etc.), data center networking infrastructure components (switches, network controllers, physical layer devices, etc.) and cloud infrastructure services;
2. **Software:** includes information management software, analytics and discovery software, and application software specific to Big Data;



IDC expects the Big Data technology and services market to grow from \$9.8 billion in 2012 to \$32.4 billion in 2017, representing a CAGR of 27%.

3. Services: includes business consultation, integration services, storage services, security services, hardware and software support, training, outsourcing and invariably any services related to Big Data implementations;

IDC expects the Big Data technology and services market to grow from \$9.8 billion in 2012 to \$32.4 billion in 2017, representing a compound annual growth rate (CAGR) of 27% or about 6 times that of the overall Information Technology (IT) market.

When defining Big Data, it is important to understand the mix of unstructured and multi-structured data that comprises this volume of information⁵:

- **Unstructured Data:** information that is not organized or easily interpreted by traditional databases or data models, and is typically text-heavy such as metadata, twitter tweets and other social media posts. The majority of unstructured data resides in text files that accounts for at least 80% of an organization's data, and if left unmanaged, the sheer volume it generates annually can be costly in terms of storage and can pose a liability to the company or business operation if information cannot be located.
- **Multi-Structured Data:** a variety of data formats and data types typically derived from interactions between people and computing systems, such as web applications/transactions or social networks. Web log data is a good example of multi-structured data as it includes a combination of text, visual images and transactional information.

Within this data lies valuable patterns and information that in the past had been previously hidden because of the amount of work required and costs associated to extract them. In today's data center, Big Data has become viable as cost-effective approaches have emerged to address the **3Vs of Big Data – volume, velocity and variability**, as briefly described below⁶:

- **Volume:** represents the amount of data
- **Velocity:** represents the speed of data in and out of a system, server and/or storage device (or the real-time processing of a data stream)
- **Variability:** represents the varying range of data types and sources

3 Big Data Use Case Examples

Big Data can represent chatter from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast streams (audio, video, both), financial transactions, internet downloads, document retrieval, document scans, engineering designs, GPS trails, automobile telemetry, market data, analytical data, or literally any project by which data is too big, moves too fast, or exceeds current processing capacities. The following represents examples of what Big Data looks like and how it is used in real-world sectors⁷:



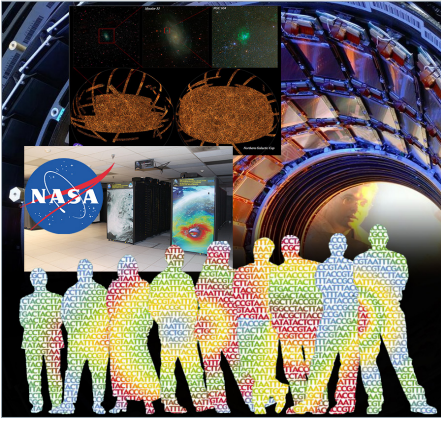
Private Sector

- Facebook handles 50 billion photos from its user base daily
- Walmart handles more than one million customer transactions every hour
- Amazon handles millions of back-end operations per day as well as queries from more than half a million third-party sellers
- eBay uses two (2) data warehouses at 7.5 and 40 petabytes respectively to handle product searches, consumer recommendations and merchandising



Government Sector

- The Big Data Research and Development Initiative announced by the Obama administration in 2012 explores how Big Data can be used to address important issues faced by the government
- Big Data analysis played a large role in Barack Obama's successful 2012 re-election campaign
- The U.S. federal government owns six of the ten most powerful supercomputers in the world
- The National Security Agency (NSA) is developing a data center to handle the large amount of information collected by the NSA over the Internet




Science & Research Sector

- The world's most powerful particle accelerator supporting about 150 million sensors uses Big Data analytics to test theories and predictions associated with particle physics and high-energy physics (called the Large Hadron Collider project) delivering data at 40 million times per second
- The Sloan Digital Sky Survey (SDSS) collects astronomical data at a rate of about 200GB per night amassing over 140 terabytes of information since 2000
- Decoding the human genome originally took 10 years to process, however, it can now be achieved in less than a week through Big Data analytics
- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations

4 Big Data Analytics

Advancements in Big Data analytics offer alternative opportunities to improve decision-making in critical development areas that touch our lives daily or eventually will affect our lives in the future. In his article entitled, "The Awesome Ways Big Data is Used Today to Change Our World," Bernard Marr presents a variety of uses for Big Data analytics as follows:

- **Understanding and Targeting Customers:** Big Data analytics are used to better understand customers, their behaviors and buying preferences, and create predictive models that better gauge purchasing trends. This is the most common and pervasive use for Big Data analytics that can lead to new high-growth market segments and the development of new product features, capabilities and technologies.
- **Understanding and Optimizing Business Processes:** Big Data analytics optimizes business processes such as analyzing/monitoring product stock for retailers based on predictive models. Another example includes delivery route optimization as geographic positioning and radio frequency identification sensors track goods and delivery vehicles, optimizing routes while expediting delivery times through the integration of live traffic data.
- **Optimizing and Quantifying Individual Performance:** Big Data analytics optimizes and quantifies personal requirements and capable of collecting data from wearable devices such as smart watches or bracelets. The



analytics can determine daily calorie consumption, activity levels, sleep patterns, etc. In addition, most online dating sites apply Big Data analytics to find appropriate matches for subscribers.

10 Uses of Big Data Analytics:

- *Understand/Target Customers*
- *Optimize Business Processes*
- *Quantify Individual Performance*
- *Improve Healthcare*
- *Improve Sports Performance*
- *Improve Science/Research*
- *Optimize Machine Performance*
- *Improve Security/Law Enforcement*
- *Improve Cities/Countries*
- *Optimize Financial Trading*

- **Improving Healthcare and Public Health:** Big Data analytics can decode entire DNA strings in minutes, find new cures for diseases and challenging medical conditions, and better understand and predict disease patterns. One example monitors premature babies by recording and analyzing heart beats and breathing patterns, which in turn, can predict infections before physical symptoms appear. Another example monitors flu outbreaks in real-time.
- **Improving Sports Performance:** Utilizing video, Big Data analytics tracks athletic performances that helps determine trends and athletic consistency. Sensor technology in the equipment itself (balls, golf clubs, tennis rackets, etc.) provides another level of analysis that can result in improvements to the game. Many elite sports teams use Big Data analytics to monitor their athletes for such items as nutrition, sleep and emotional well-being.
- **Improving Science and Research:** The Large Hadron Collider project discussed earlier is an example for how Big Data analytics improves science and research. The CERN nuclear physics lab in Switzerland is another example as Big Data analytics uses 65,000 processors in its data center to analyze 30 petabytes of data it stores from thousands of computers distributed across 150 data centers worldwide. Another example performs seismic image analysis that locates ideal places to drill for oil and gas as part of an exploratory process.
- **Optimizing Machine and Device Performance:** Machines and devices become smarter and operate more autonomous using Big Data analytics. For example, Google's self-driving Toyota Prius is fitted with cameras, GPS, powerful computers and sensors so the car can be safely driven without human intervention.
- **Improving Security and Law Enforcement:** The National Security Agency (NSA) uses Big Data analytics to detect and prevent cyber-attacks and to help foil terrorist plots before they occur. These tools can not only predict criminal activity but can catch criminals. Credit card companies use Big Data analytics to detect fraudulent activities.
- **Improving and Optimizing Cities and Countries:** Traffic flow optimization is a good example for how a city or municipality utilizes Big Data analytics



Big Data applications use mixed read and write workloads that require very low latency and significant IOPS performance which is not a good match for HDD storage but is ideally suited for enterprise-class SSDs.

and is based on real-time traffic and weather data as well as social media. The transport infrastructure and utility processes are joined together so invariably a bus could be waiting for a delayed train or traffic signals operate to minimize traffic jams by predicting traffic flows.

- **Optimizing Financial Trades:** The majority of equity trading is performed using Big Data analytics that utilize social media networks and news websites so that buy and sell decisions can be made in split seconds.

Though these categories represent the areas in which Big Data is presently applied the most, with so many potential Big Data applications on the horizon, more innovative tools will become widespread, which in turn, will spawn new Big Data categories.

5 Benefits Of Flash-Based SSDs For Big Data

To gain value from Big Data and achieve a significant return on investment (ROI), IT departments must choose alternate ways to process and analyze data since conventionally the data is too large, moves too quickly or doesn't fit the database architecture structures. Big Data challenges today's enterprises and IT infrastructures as the application requires ultra-fast processing of large datasets, which in turn expedites data-driven analytics.

Big Data applications use mixed read and write workloads that require very low latency and significant input/output operations per second (IOPS) performance which is not a good match for hard disk drive (HDD) storage but is ideally suited for enterprise-class solid-state drives (SSDs). HDDs have performance and physical limitations that prevent them from keeping pace with Big Data applications and with growing server workloads in general. While basic servers can handle hundreds of thousands of IOPS, a traditional HDD can only deliver between 100 and 300 IOPS performance typically causing a huge performance disparity. For every instance that data is requested from a different location in HDD storage, the mechanical head of the hard drive needs to move, limiting its physical ability to quickly read random data.

HDDs are designed for straightforward data streams, handling sequential reads and writes that are physically located on the same track. As modern operating systems have become more capable of multiprocessing complex data, more random reads and writes are occurring that HDDs simply cannot keep pace with. In comparison to traditional hard drives, the NAND flash cells within an SSD are much denser and do not use rotating disks or magnetic

Contact us for more information

OCZ Storage Solutions
6373 San Ignacio Avenue
San Jose, CA 95119 USA

P 408.733.8400
E sales@oczenterprise.com
W ocz.com/enterprise

EMAIL SALES TEAM >

VISIT OCZ ENTERPRISE >

heads to search for a specific location to process and access data. As such, the controller already has the required data locations available which translate to faster read and write access times, no moving parts that can break or malfunction, and effortless I/O access of random data with low latency.

Flash-based SSDs have become the popular choice for Big Data applications as they provide faster I/O performance than HDD storage, support large storage capacities and a variety of form factors and interfaces, consume less power and retain data when power is removed.

6 Summary

As forecast by leading market research firms that follow the storage industry, Big Data will continue to represent a fast-growing multi-billion dollar worldwide opportunity over the next five years. It represents a new generation of technologies and architectures designed to economically extract value from very large data volumes covering a wide variety of data for the intent of enabling high-velocity capture, discovery and/or analysis.

Flash-based SSDs have become the popular choice for Big Data applications as they provide faster I/O performance (than HDD storage), support large storage capacities and a variety of form factors and interfaces, consume less power, and retain data when power is removed. Big Data requires ultra-fast processing of large datasets which in turn expedites data-driven analytics. As Big Data represents a large volume of both structured and unstructured data that is too big, moves too fast, or exceeds current processing capabilities, the ability to manage and monitor the data activity and flash resources remotely provides a major benefit to this application.

OCZ provides a complete portfolio of SSD hardware and storage solutions targeted toward Big Data applications as outlined in Part 2 of the white paper entitled, "[Driving Big Data Applications with OCZ Solid State Solutions.](#)"

References and Industry Sources

- ¹ “What is Big Data” – a definition from Webopedia.com.
- ² “What is Big Data,” by Lisa Arthur, Forbes Magazine, August 15, 2013.
- ³ “Big Data: The next frontier for innovation, competition, and productivity,” by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, McKinsey Global Institute (MGI), McKinsey & Company, May 2011.
- ⁴ “Worldwide Big Data Technology and Services 2013-2017 Forecast,” by Dan Vessel, Rob Brothers, Steve Conway, Matthew Eastwood, John Grady, Brian McDonough, Henry D. Morris, David Schubmehl, Mary Johnston Turner, Melissa Webster, Ashish Nadkarni, Christian Christiansen, Mukesh Dialani, Maureen Fleming, Tim Grieser, Rohit Mehra, Carl W. Olofson, Kuba Stolarski, Mary Wardley, Ali Zaidi, IDC, Report #244979, December 2013.
- ⁵ “What is Big Data,” by Lisa Arthur, Forbes Magazine, August 15, 2013.
- ⁶ “What is Big Data,” by Edd Dumbil, Strata-O'Reilly Media, January 11, 2012.
- ⁷ “What is Big Data,” – a definition from Wikipedia.com.
- ⁸ “The Awesome Ways Big Data Is Used Today To Change Our World,” by Bernard Marr, LinkedIn.com, November 13, 2013.

Disclaimer

OCZ may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of OCZ products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and OCZ does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. OCZ assumes no obligation to update or otherwise correct or revise this information.

OCZ MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

OCZ SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL OCZ BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF OCZ IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2014 OCZ Storage Solutions, Inc. – A Toshiba Group Company. All rights reserved.

OCZ, the OCZ logo, OCZ XXXX, OCZ XXXXX, [Product name] and combinations thereof, are trademarks of OCZ Storage Solutions, Inc. – A Toshiba Group Company. All other products names and logos are for reference only and may be trademarks of their respective owners.