SOLIX®
Empowering Data Management

BIG DATA AND THE NEW ENTERPRISE BLUEPRINT

JOHN B. OTTMAN, JR.
Executive Chairman
Solix Technologies, Inc.
http://www.solix.com
http://twitter.com/johnottman

A SOLIX WHITEPAPER
JULY 2014

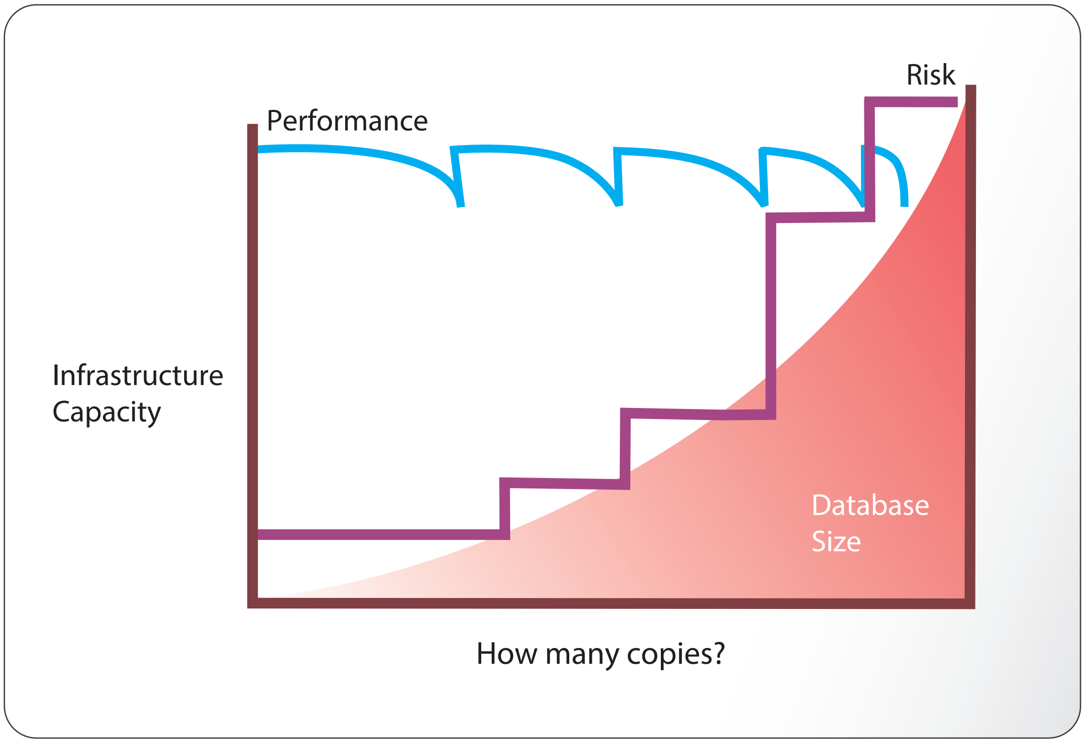
Data Growth Crisis

We now understand that the world is drowning in data. It is estimated that over 15 petabytes of new information is created every day, which is eight times more than the information in all the libraries in the United States. This year, the amount of digital information generated is expected to reach 988 exabytes. This is equivalent to the amount of information if books were stacked from the Sun to Pluto and back.[1]

Gartner agrees data growth is now the leading data center infrastructure challenge. In a recent survey, 47 percent of Gartner survey respondents ranked data growth as their number one challenge. [2]

Data growth is capable of stripping entire data centers of cooling and power capacity. System availability is impacted as batch processes are no longer able to meet scheduled completion times. The "outage windows" necessary to convert data during ERP upgrade cycles may extend from hours to days, and other critical processes like replication and disaster recovery are impacted since more and more data is harder and harder to move.

Additionally, data growth creates governance, risk and compliance challenges. HIPAA, PCI DSS, FISMA and SAS 70 mandates all require that organizations establish frameworks for data security and compliance. Information Lifecycle Management (ILM) programs are required to meet compliance objectives throughout the data lifecycle.



The Paradox of Moore's Law

Alongside the data growth explosion, Moore's Law continues to astound as processor and integrated circuit performance doubles every few years, just as the visionary Intel founder predicted. The benefits of such dramatic technological advance may not be overstated. Today, more processing power is packed into a smartphone than yesterday's mainframe, and the cost to store a terabyte of data in the cloud has fallen to as low as $10 per month.

But rather than saving, organizations are spending more and gaining improved business value as a result by processing mission-critical enterprise data faster and faster. Until recently, few ERP users would ever have imagined paying the high cost to process enterprise data using full flash arrays, yet today, most have either already installed or are evaluating premium performance enterprise platforms with solid state disk (SSD) such as SAP Hana and Oracle Exadata.

Despite such spectacular gains in semiconductor price/performance, the overall cost of IT continues to rise, largely because we are now processing so much more data. Challenged by this dilemma, CIOs must continually find new ways to reduce the cost of data growth, so they may afford to fund more mission-critical applications that improve business results.

A New Enterprise Blueprint

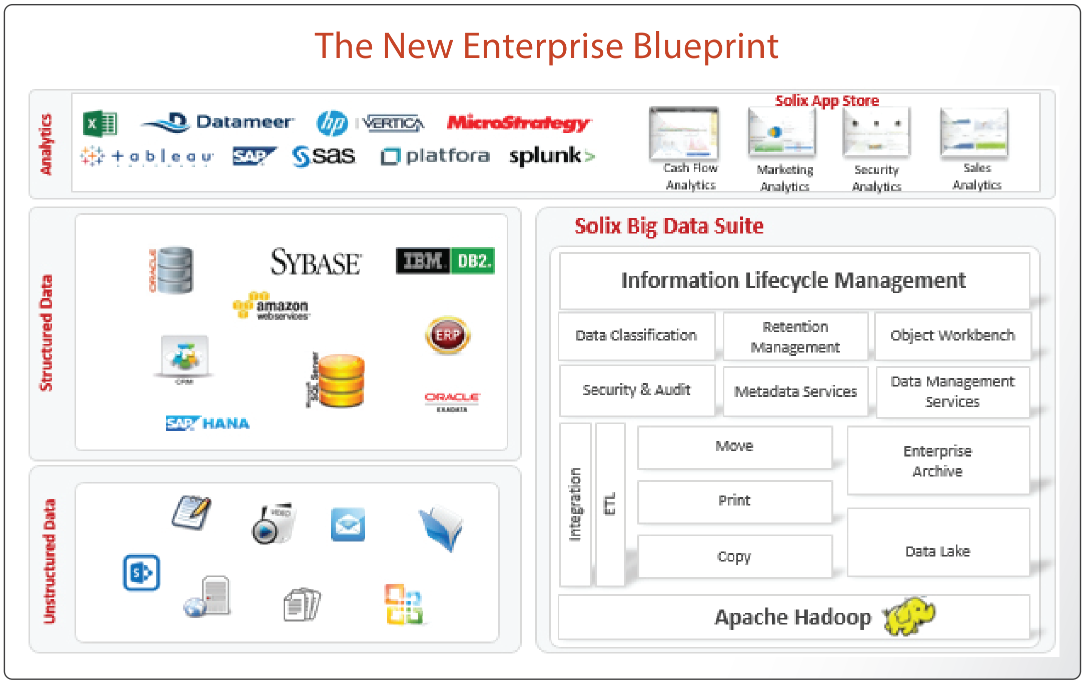
Advances in semiconductor technology have indeed enabled "commodity" hardware to process and store extraordinary amounts of data at lower unit costs. Through virtualization, this low-cost infrastructure may now be utilized with extraordinary efficiency.

Apache Hadoop is designed to leverage commodity infrastructure to deliver massive scalability. Using the MapReduce programming model to process large data sets across distributed compute nodes in parallel, Hadoop provides the most efficient and cost-effective bulk data storage solution available. Such capabilities enable compelling new big data applications, such as Enterprise Archiving and Data Lake, and establish a new enterprise blueprint for data management on a petabyte scale.

Experts agree that as much as 80 percent of production data in ERP, CRM, file servers and other mission-critical applications may not be in active use, and both structured and unstructured data becomes less active as they age. Large amounts of inactive data stored online for too long reduces the performance of production applications, increases costs and creates compliance challenges.

Big data offers a low-cost bulk storage alternative for storing inactive enterprise data. By moving inactive data to nearline storage, application performance is improved and costs are reduced as data sets are smaller and workloads are more manageable. Even more important, universal data access is enhanced through analytics applications, structured query and reporting and simple text search.

Big data and the new enterprise blueprint enable organizations to gain improved value from their data. Enterprise data warehouse (EDW) and analytics applications leverage big data for better described views of critical information. As a low-cost data repository to store copies of enterprise data, big data is an ideal platform to stage critical enterprise data for later use by EDW and analytics applications.



Enterprise Archiving

Organizations continually demand improved performance from their mission-critical online systems, but ultra-high performance infrastructure costs more to deliver such results. How much more depends on how much data will be processed online using high performance compute nodes and full-flash memory arrays.

Production application performance is improved and infrastructure costs are reduced by moving less frequently accessed data to nearline, bulk storage. Enterprise archiving with Apache Hadoop moves older, inactive data from production data storage to a nearline repository for ongoing data access by end users. Because online data sets are reduced, enterprise applications run faster and with higher availability.

Enterprise Archiving involves a coordinated MOVE of online application data to nearline storage, and then a PURGE process deletes data from the production location to free infrastructure resources. ILM policies control the ingestion processes and ensure proper governance, risk and compliance objectives are met.

End user access to the enterprise archive is available through structured query, reporting or full text search.

Data Lake

Apache Hadoop represents a significant opportunity for enterprise data warehouse (EDW) and analytics applications.

Data warehouse users continually seek better ways to describe data, challenging EDW platforms to deliver highly specific data views that meet end user requirements. Canonical views of data delivered top-down may not satisfy end user requirements for more specifically-described data. Data lake applications leverage big data technology to store copies of production data "as is," reducing the complexity to stage EDW and analytics applications.

Big data enhances traditional EDW strategies because Apache Hadoop stores and processes structured and unstructured enterprise data in bulk and at a low cost. Data lake applications utilize a simple copy process to eliminate the need for heavy extract, transform and load (ETL) processes during ingestion. Once resident within the Hadoop Distributed File System (HDFS), enterprise data may be more easily distilled by analytics applications and mined for critical insights at a petabyte scale.

By storing data "as is," the cost and complexity for ETL processes to stage EDW and analytics applications is reduced. Low cost and flexibility make the data lake a highly efficient storage solution for data until it is needed later by analytics applications.

Information Lifecycle Management

Data is the life blood of any organization, and big data applications require ILM solutions for effective enterprise data management. Too much data stored online slows application performance and taxes IT resources. Lack of proper data access and availability leads to poor decision making and missed business results. It is the job of ILM to control data growth and provide a framework for data governance and compliance.

ILM frameworks classify enterprise data and ensure that best practices, data retention policies and business rules are deployed to meet compliance objectives, such as legal hold and COBIT. Enterprise archive and data lake applications store vital information, and compliance mandates for data governance and security must always be met.

Enterprise archiving and data lake applications use ILM to:

- Classify enterprise data
- Manage the cost and performance impact of data growth
- Align system performance and service levels to business goals
- Establish retention and compliance policies for enterprise data
- Enable proper retirement and decommissioning of obsolete enterprise applications and data

Summary

As Gartner reports, data growth has emerged as the biggest data center and hardware infrastructure challenge, and is "particularly associated with increased costs relative to hardware, software, associated maintenance, administration and services."[3] As more and more data is processed and stored, system performance is impacted, costs increase and compliance objectives become harder to meet.

At the same time, demand for improved access to data through EDW and enterprise analytics applications has never been higher. Organizations are seeking new, more efficient ways to gain value and competitiveness by mining enterprise data.

A new enterprise blueprint for data management has emerged to meet these challenges. Highly scalable, big data technology stores less frequently accessed data in bulk and at the lowest possible cost. Enterprise archiving and data lake applications improve the performance of online systems, reduce infrastructure costs, improve compliance and help organizations gain improved business value from their enterprise data.

Biography

John Ottman is Executive Chairman of Solix Technologies, Inc. and has over 25 years of enterprise software experience. He is also Chairman and Co-Founder of Minds.com, a leader in open source social media.

Previously, Mr. Ottman was President and CEO of Application Security, Inc., President of Princeton Softech, Inc., and Executive Vice President, Worldwide Markets at Corio, Inc.. He also has held key senior management roles at Oracle and IBM.

Mr. Ottman is author of the database security book, *Save the Database, Save the World!* and holds a B.A. from Denison University.

Footnotes:

1. http://www.enterprisestorageforum.com/management/features/article.php/3911686/CIOs-Struggling-With-Data-Growth

2. http://www.gartner.com/it/page.jsp?id=1460213

3. http://www.gartner.com/it/page.jsp?id=1460213

SOLIX®
Empowering Data Management

Solix Technologies, Inc.
4701 Patrick Henry Dr.,
Building 20
Santa Clara, CA 95054
Phone: 1.888.GO.SOLIX (1.888.467.6549)
1.408.654.6400
Fax: 1.408.562.0048
URL: http://www.solix.com

This document is provided for information purposes only and the contents hereof are subject to change without notice.

This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchant- ability or fitness for a particular purpose.

We specially disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.