



Making Sense of Big Data in the Petabyte Age

Executive Summary

The concept of “big data” is gaining attention across industries and the globe, thanks to the growth in social media (Twitter, Facebook, blogs, etc.) and the explosion of rich content from other information sources (activity logs from the Web, proximity and wireless sources, etc.). The desire to create actionable insights from the ever-increasing volumes of unstructured and structured data sets is forcing enterprises to rethink their approaches to big data, particularly as traditional approaches have proved difficult – if even possible – to apply to structured data sets.

While data volume proliferates, the knowledge it creates has not kept pace. For example, the sheer complexity of how to store and index large data stores, as well as the information models required to access them, have made it difficult for organizations to convert captured data into insight.

The media appears obsessed with how today's leading companies are dealing with big data, a phenomenon known as living in the “Petabyte Age.” However, coverage often focuses on the technology aspects of big data, leaving such concerns as usability largely untouched.

For years, the accelerating data deluge has also received significant attention from industry pundits and researchers. What is new is the threshold that has been crossed as the onslaught

continues to accelerate in terms of volume, complexity and formats.²

A traditional approach to handling big data is to replace SQL with tools like MapReduce.³ However, the sheer volume of data contained in a data set that is routinely analyzed does not solve the more pressing issue – that people have difficulty focusing on massive amounts of tables, files, Web sites and data marts, all of which are candidates for analysis. It's not all about just data warehouses, anymore.

Usability is a factor that will overshadow the technical characteristics of big data analysis for at least the next five years. This paper focuses specifically on the roadmap organizations must create and follow to survive the Petabyte Age.

Big Data = Big Challenges

The Petabyte Age is creating a multitude of challenges for organizations. The accelerating deluge of data is problematic to all, for within the massive array of data sources – including data warehouses, data marts, portals, Web sites, social media, files and more – is the information required to make the smartest strategic business decisions. Many enterprises are facing the dilemma that the systems and processes devised specifically to integrate all this information lack the required responsiveness to place the information into a neatly organized warehouse in

time to be used at the current speed of business. The heightened use of Excel and other desktop tools to integrate needed information in the most expedient way possible only adds complexity to the problem enterprises face.

There are a number of factors at the heart of big data that make analysis difficult. For example, the timing of data, the complexity of data, the complexity of the synthesized enterprise warehouse and the identification of the most appropriate data are equal if not larger challenges than dealing with large data sets, themselves.

The increased complexity of the data available for generating insights is a direct consequence of the following:

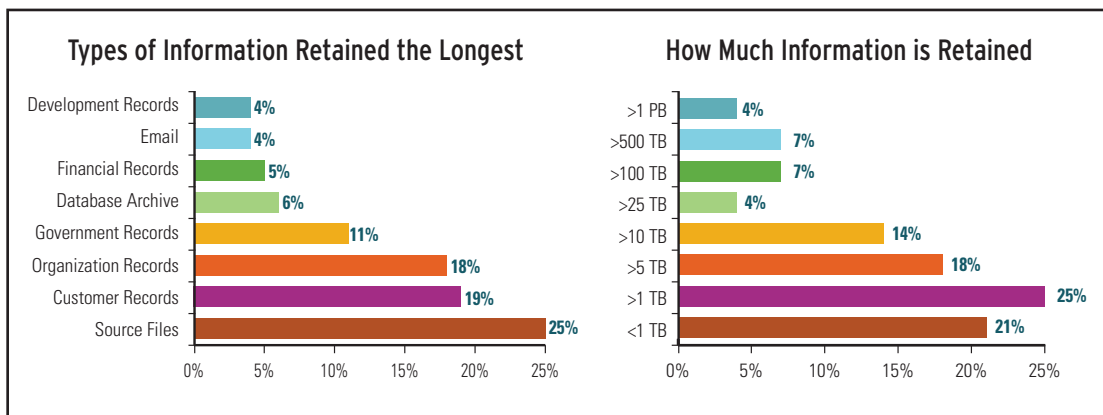
- **The highly communicative and integrated global economy.** Enterprises across industry are increasingly seeking more granular insight into market forces that ultimately shape their success and failure. Data generated by the “always-on economy” is the impetus, in many cases, for the keen interest in implementing so-called insight facilitation appliances on mobile devices - smartphones, iPads, Android and other tablets – throughout the enterprise.
- **The enlightened consumer.** Given the explosion in social media and smart devices, many consumers have more information at their fingertips than ever before (often more so than businesses) and are becoming increasingly sophisticated in how they gather and apply such information.
- **The global regulatory climate.** New regulatory mandates – covering financial transactions, corporate health, food-borne illnesses, energy

usage and availability and other issues – require businesses to store increasingly greater volumes of information for much longer time frames.

As a result of these factors, enterprises worldwide have been rapidly increasing the amount of data housed for analysis to compete in the Petabyte Age. Many have responded by embracing new technologies to help manage the sheer volume of data. However, these new toys also introduce data usability issues that will not be solved by new technology but rather will require some rethinking of the business consequences of big data. Among these challenges:

- **Big data is not only tabular; it also includes documents, e-mails, pictures, videos, sound bites, social media extracts, logs** and other forms of information that is difficult to fit into the nicely organized world of traditional database tables (rows and columns).
- **Companies that tackle big data as a technology-only initiative** will only solve a single dimension of the big data mandate.
- **There are sheer volumetric issues, such as billions of rows of data, that need to be solved.** While tried-and-true technologies (partitioning) and newer technologies (MapReduce, etc.) permit organizations to segment data into more manageable chunks, such an approach does not deal with the issue that rarely used information is clogging the pathway to necessary information. Traditional lifecycle management technologies will alleviate many of the volumetric issues, but they will do little to solve the non-technical issues associated with volumetrics.

Tabulating the Information Lifecycle



Source: The 100 Year Archive Task Force, SNIA Data Management Forum
Figure 1

As a result of mergers and acquisitions, global sourcing, data types and other issues, the sheer number of tables and objects available for access has mushroomed. This increase in the volume of objects has made access schemes for big data overly complex, and it has made finding needed data akin to finding a needle in a haystack.

- **The information lifecycle management⁴ considerations of data have not received the attention they deserve.** Information lifecycle management should not be limited to partitioning schemes and the physical location of data. (Much attention is being given to cloud and virtualized storage, which presumes a process has been devised for rationalizing the fact that always-on information made available in the cloud is worthy of this heightened availability.) Information lifecycle management is the process that traditionally stratifies the physical layout for technical performance. In the Petabyte Age, where the amount of information available for analysis is increasing at an accelerating rate, the information lifecycle management process should also ensure a heightened focus on information that matters. This stratification should categorize information into the following groups:
 - Information directly related to the creation, extraction or capture of value.
 - Supporting information that could be referred to when devising a strategy to create, extract or capture value.
 - Information required for the operations of the enterprise but not necessarily related to the creation, extraction or capture of value.

- Information required for regulatory activities but not necessarily related to the creation, extraction or capture of value.
- Historical supporting information.
- Historical information that was once aligned with value, regulatory or other purposes but is now kept because it might be useful at some future date.

- **Much of the complexity of information made available for deriving insight is a complex weave of multiple versions of the truth, data organized for different purposes, data of different vintages and similar data obtained from different sources.** This is a phenomenon that many organizational stakeholders describe as a “black box of information” into which they have no insight into its lineage. This adds delay to the use of insight gained from such information, a development caused by the obvious necessity of validating information prior to using it for anything out of the ordinary. Much of the data available for analysis results from the conventional wisdom that winning organizations are “data pack rats” and that information that arrives on their doorsteps tends to stay as a permanent artifact of the business. Interestingly, according to the 100-Year Archive Task Force,⁵ source files are the most frequently identified source of data retained as part of the “100-Year Archive.”
- **A sizable amount of operational information is not housed in official systems administered by enterprise IT groups but instead is stored on desktops throughout the organization.** Much of these local data stores were created with good intentions; people respon-

Many Sources of Data

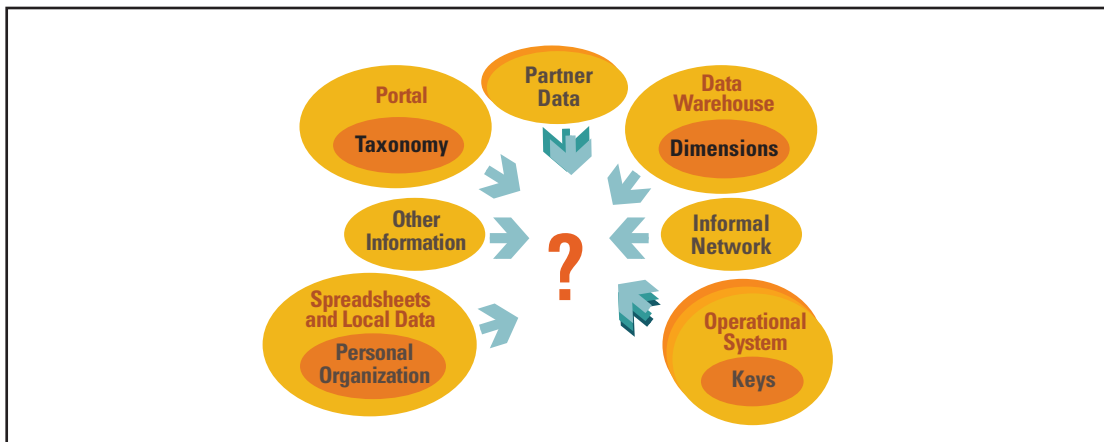


Figure 2

sible for business functions had no other way of gaining control over the information they needed. As a result, these desktop-based systems have created a different form of big data – a weave of Microsoft Access, Excel and other desktop tool-based sources that are just as critical in running enterprises. The contents of these individually small to medium-size sources of information collectively add up to a sizable number of sources. One large enterprise was found to have close to 1,000 operational metrics managed in desktop applications (i.e., Excel) – which is not an uncommon situation. These sources never make it to the production environment and downstream data warehouses.

- **Much of the big data housed in organizations results from regulatory requirements that necessitate storing large amounts of historical data.** This large volume of historical data hides the data required for insight. While it is important to retain this information, the necessity to house it with the same priority as information used for deriving insight is both expensive and unnecessary.

The Case for Horizontal Partitioning

Horizontal partitioning is the process defined as segmenting data in such a way that prioritizes information required for value extraction, origination and capture.⁶ This partitioning process should result in a process that tiers information along the traditional dimensions of a business information model. Such a model enhances the focus of information that supports the extraction, origination and capture of value.

The Roadmap to Managing Big Data

Big data will be solved through a combination of enhancements to the people, process and technology strengths of an enterprise.

People-based initiatives that will impact big data programs employed at companies include the following:

- **Managing the information lifecycle employed at the organizations.** For good reason (glaring privacy and security concerns, among them), organizations have placed significant focus on information governance. The mandate for determining which data deserves focus should be part of the overall governance charter.
- **Ensuring a sufficient skill set** to tackle the issues introduced as a consequence of big data.
- **Developing a series of metrics to manage the effectiveness of the big data program.** This includes:
 - Average, minimum and maximum time required to turn data into insight.
 - Average, minimum and maximum time required to integrate new information sources.
 - Average, minimum and maximum time required to integrate existing information sources.
 - Time required for the management process.
 - Percentage of people associated with the program participating in the management process.
 - Value achieved from the program.

Converting Big Data Into Value

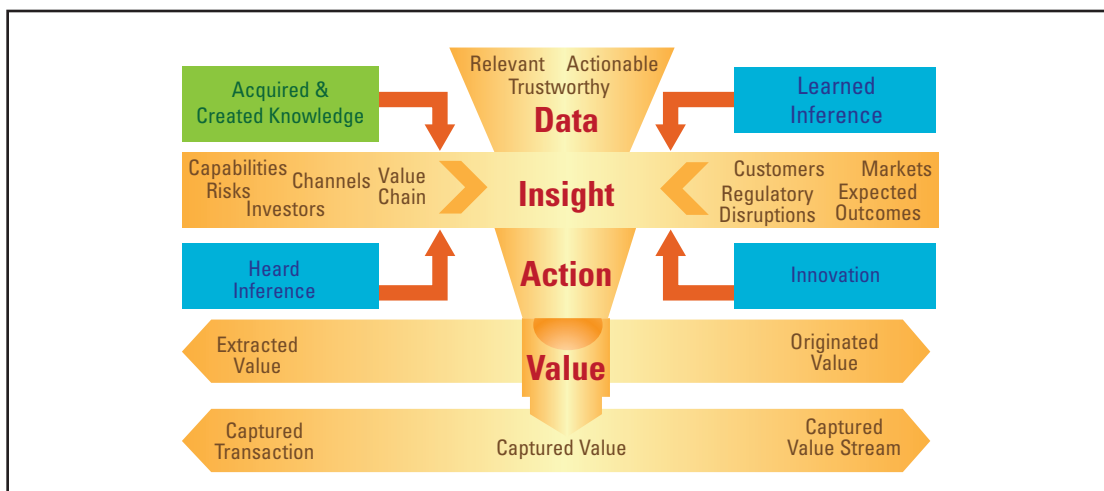


Figure 3

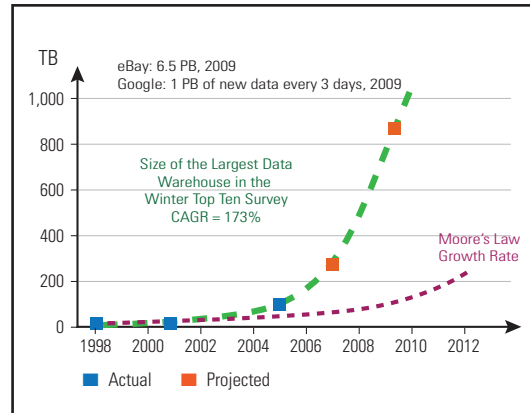
According to McKinsey,⁷ the activities of people steering big data will include:

- Ensuring that big data is more accessible and timely.
- Measuring the value achieved by unlocking big data.
- Embracing experimentation through data to expose variability and raise performance.
- Enabling the customization of populations through segmentation.
- Facilitating the use of automated algorithms to replace and support human decision-making, thereby improving decisions, minimizing risks and unearthing valuable insights that would otherwise remain hidden.
- Facilitating innovation programs that use new business models, products and services.

Process-based initiatives that will impact big data programs are best enabled as augmentations of a company's governance activities. These augmentations include:

- **Ensuring sufficient focus on information that will drive value within the organization.** These processes are best employed as linkages between corporate strategy and information lifecycle management programs.
 - It is important to note that information lifecycle management is defined in many organizations as a program to manage hierarchies and business continuity. For the purposes of this paper, this definition is extended to include promotion and demotion of data items as aligned with the business information model (how information is used to support enterprise strategies and tactics) of the organization.
 - The process used to govern big data and its information lifecycle should continually refine and prioritize the benefits, constraints, priorities and risks associated with the timely publication and use of relevant, focused, actionable and trustworthy information published under big data initiatives.
- **Ensuring the metrics that drive proper adhesion and use of big data are developed.** This should cover the following topics:
 - Governing big data.
 - Big data lifecycle.
 - Big data use and adoption.
 - Big data publication metrics.

Big Data Getting Bigger



Source: Winter Corp.

Figure 4

Technology-based initiatives that will impact big data programs employed at companies include:

- **Ensuring that the specialized skills required to administer and use the fruits of the big data initiative are present.** Some of these include the databases, the ontologies used to navigate big data and the MapReduce concepts that augment or fully replace SQL access techniques.
- **Ensuring that tools introduced to navigate big data are usable by the intended audience** without upsetting self-service paradigms that have slowly gained traction during the past several years.
- **Ensuring that the architecture and supporting network, technology and software infrastructures** are capable of supporting big data.

It is safe to state that if history is any prediction of the future, the sheer volume of data that organizations will need to deal with will outstrip our collective imaginations for how much data will be available for generating insights.. Only eight years ago, a 300 to 400 terabyte data warehouse was considered an outlier. Today, multi-petabyte warehouses are easily found. Failure to take action to manage the usability of information pouring into the enterprise is (and will be) a competitive disadvantage (see Figure 4).

Recommendations

Big data is a reality in most enterprises. However, companies that tackle big data as merely a technology imperative will solve a less important dimension of their big data challenges. Big data is much more than an extension of the technolo-

Storage Definitions

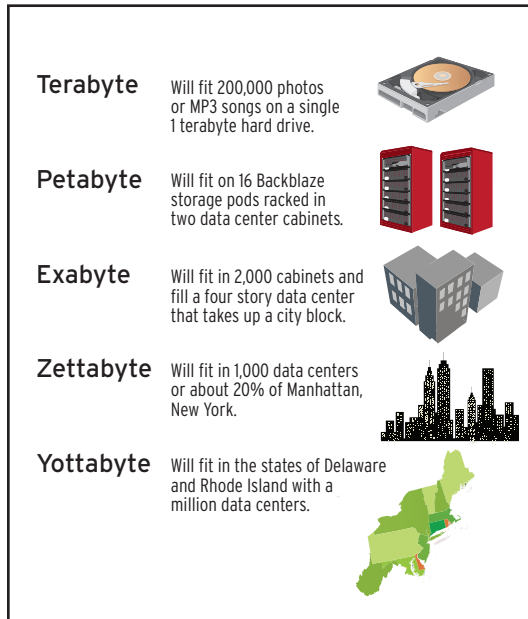


Figure 5

gies used in partitioning strategies employed at enterprises.

Companies have proved that they are pack rats. They need to house large amounts of history for advanced analytics, and regulatory pressures influence them to just store everything. The reduced cost of storage has allowed companies to turn their data warehousing environments into data dumps, which has added both a complexity

to the models (making it difficult for knowledge workers to navigate the data needed for analysis) and added an analytic plaque, which makes finding the required information for analysis more akin to finding a needle in a haystack.

In many organizations, information lifecycle initiatives are mandated that too often focus on the optimal partitioning and archiving of the enterprise (i.e., vertical partitioning). Largely a technology focus, this thread of the information lifecycle overlooks data that is no longer aligned with the strategies, tactics and intentions of the organization. The scope and breadth of information housed by enterprises in the Petabyte Age mandates that data be stratified according to its usefulness in the organizational value creation (i.e., horizontal partitioning). In today's organizations, the only cross-functional bodies with the ability to perform this horizontal partitioning are virtual organizations employed to govern the enterprise information asset.

It was only a few years ago that a data warehouse requiring a terabyte of storage was the exception. As we embrace the Petabyte Age, companies are entering an era where they will need to be capable of handling and analyzing much larger populations of information. Regardless of the processes put in place, ever-increasing volumes of structured and unstructured data will only proliferate, challenging companies to quickly and effectively convert raw data into insight in ways that stand the test of time.

Footnotes

- ¹ Big data refers to data sets that grow so large that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analytics and visualizing.
- ² "The Toxic Terabyte" (IBM Research Labs, July 2006) provides a thorough analysis of how companies had to get their houses in order to deal with a terabyte of data. If this authoritative work were rewritten today, it would be called "The Problematic Petabyte" and in five years, most probably, "The Exhaustive Exabyte."
- ³ MapReduce is a Google-inspired framework specifically devised for processing large amounts of data optimized across a grid of computing power.
- ⁴ Information lifecycle management is a process used to improve the usefulness of data by moving lesser used data into segments. Information lifecycle management is most commonly interested in moving data from always-needed partitions to rarely needed partitions and, finally, into archives.
- ⁵ SNIA Data Management Forum's 100 Year Archive Task Force, 2007.
- ⁶ Horizontal partitioning is a term created by the author. It describes the application of generally accepted techniques of gaining performance by segmenting data into partitions (vertical partitioning) to segmenting groups of data by the likelihood of it achieving organizational value.
- ⁷ "Big Data, The Next Frontier for Innovation, Competition and Productivity," McKinsey & Company, May 2011.

About the Author

Mark Albala is Director of Cognizant's North American Business Intelligence and Data Warehousing Consulting Practice. This practice provides solution architecture, information governance, information strategy and program governance services to companies across industries and supports Cognizant's business intelligence and data warehouse delivery capabilities. A graduate of Syracuse University, Mark has held senior thought leadership, advanced technical and trusted advisory roles for organizations focused on the disciplines of information management for over 20 years. He can be reached at Mark.Albala@cognizant.com.

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world's leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 111,000 employees as of March 31, 2011, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.



World Headquarters

500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277
Email: inquiry@cognizant.com

European Headquarters

Haymarket House
28-29 Haymarket
London SW1Y 4SP UK
Phone: +44 (0) 20 7321 4888
Fax: +44 (0) 20 7321 4890
Email: infouk@cognizant.com

India Operations Headquarters

#5/535, Old Mahabalipuram Road
Okkiyam Pettai, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060
Email: inquiryindia@cognizant.com