Dealing with Big data

Introduction

Every day, trillions of bytes of data are generated by organizations and individuals around the world. Companies know there are valuable insights buried within those mountains of bits—and are looking to their CIOs to find ways to use them. This paper shares Pythian’s practical perspective on how to approach the big data challenge organizationally and technically, and now to sift information of value out of the ever-growing background of digital noise.

Marketing and sales staff. R&D teams. CXOs. They have targets to hit, questions to answer, strategies to build. All of which demand information.

The crucial insights are out there, somewhere—buried under mountains of data: in web usage stats and customer help center records, e-commerce logs, and social media archives. The vast majority of this information is unsorted and uncategorized. It’s chaotic and unwieldy. In most organizations, it falls to the CIO to figure out how to sift through all this big data.

So where do you start? What do you need to do—organizationally and technically—to distill all those masses of unstructured data into kernels of actionable wisdom that will drive informed, effective business decisions?

It requires new processes, and a variety of skills, including analytical skills that lie outside the traditional IT domain. This paper offers some practical approaches to dealing with big data—and how to quickly start realizing its value to your organization.
Solutions for a Datafied Planet
The ‘datafication’ phenomenon shows no signs of slowing down. Companies create trillions of bytes of data every day tracking customers, purchases, suppliers, and operations. More than 30 billion pieces of content are shared on Facebook each month.

Data is also created passively by millions of network sensors in everything from cars and computers to mobile phones, tablets, energy meters, and elevators. Even coffee machines and refrigerators track and create data. If you paid for parking downtown or had your picture taken on a roller coaster, chances are that data is captured somewhere.

The ability to analyze this data can be a great asset to organizations, yielding valuable advantages in marketing and customer relations, driving growth, and pointing the way to future markets.

Large private- and public-sector organizations have always generated, stored, processed, and analyzed high volumes of data. In other words, big data isn’t new, though the formats and scale have changed.

What’s really new is that gathering insights from big data is more economical and attainable today than ever before. Any organization can acquire the tools to do so.

It’s important to be clear that not all data is big data. Bad schema design is not big data. If you lack a purging policy and run an unoptimized system on 10-year-old hardware, that’s not big data. Improperly configured databases and operating systems are not big data, nor is poor data filtering.

big data generally refers to datasets that are too large and volatile—due to the amount of unstructured data constantly streaming in—for typical database software tools to capture, store, manage, and analyze.

How to Organize Big Data
Making the most of big data demands certain organizational approaches—ways of optimizing access to data sources, capitalizing on best practices, and achieving efficiencies so you can concentrate on getting value out of the data rather than simply storing it.

The six key principles are:

1. STREAMLINE
Many organizations are still structured functionally into silos, but big data doesn’t respect organizational boundaries. Collaborative internal processes are crucial for streamlining the ways you gather, organize, and analyze data.

2. DIVERSIFY
Dealing with big data demands specific knowledge and skills that often come from multiple sources, both in-house and external. The broader your knowledge of the technology and techniques related to big data, the easier it will be to solve problems that may have once seemed impossible.

3. DON’T REINVENT THE WHEEL
Someone else has almost always encountered a challenge similar to the one you’re facing—and implemented a successful solution. It is important to keep in touch with the industry through blogs, articles, and conferences to be aware of how various solutions are used—and to trust the experts, who have seen it all and know which
solutions will best solve specific problems.

4. ACQUIRE THE RIGHT SKILLS
To handle big data, it’s not enough to simply understand technology. You need to understand your business and have a good command of how statistical and analytical techniques distill value from raw data. This means seeking out and bringing people with advanced, specialized skills (e.g., a PhD in statistics or related field) into the IT fold and creating a structure that gives them access to products, processes, marketing, and sales information.

5. AUTOMATE
The sheer scale of big data demands task automation: it’s the only way small, focused teams can efficiently and effectively oversee huge data centers that process terabytes of data daily. Automation is also an important part of the analysis process. Data arrives quickly, and the ability to automatically analyze it and notify the right people of interesting trends can give organizations a competitive edge.

6. ADAPT
Organizations that can respond rapidly to change have a distinct advantage. This involves an institutional agility and flexibility when it comes to processes, data access, standards, and change control. At the same time, you don’t want to sacrifice reliability and repeatability of the data management process.

It is unlikely that all but the largest corporations will have all of the necessary big data capabilities in house. Accessing external big data expertise is a cost-effective and practical way to advance toward your goals without taking on more burden than you need to.

Key Big Data Skill Sets
With all its complexity and nuance, dealing with big data requires patience and practice. Consider the volume of multimedia available: each second of high-definition video generates more than 2,000 times as many bytes as required to store a single page of data.

Big data requires the right team for the job. Assembling a group of specific, expert skill sets related to aspects of gathering, organizing, analyzing, and implementing big data will help maximize organizational potential. The following roles represent archetypes of those skills sets related to working with big data:

Data Scientists – Identify what analysis makes sense in context. Typically have backgrounds in math and statistics, as well as artificial intelligence and natural language processing.

Data Architects – Create data models and identify required data sources and analytical tools.

Data Visualizers – Explore what data means using visualization and present how it will affect the organization.

Data Change Agents – Good communicators with a Six Sigma background, who understand how to apply statistics.

Data Engineers/Operators – Run big data infrastructure operations, develop architectures that support analytics and supply data in ways the organization needs, and make sure systems are performing smoothly.
**Data Stewards** – Ensure data sources are properly accounted for, and may also maintain a central repository as part of a master data management approach, in which there is one gold copy of enterprise data to be referenced.

**Data Virtualization/Cloud Specialists** – Build and maintain a virtualized data service layer that can draw data from any source and make it available across organizations in a consistent, easy-to-access manner.

**System Administrators** – Ensure overall operations and maintenance of IT and communications infrastructure.

Again, acquiring and retaining all of this expertise internally—especially when its only application may be to your big data strategies and not other aspects of IT—is an expensive and impractical proposition. Building the right team to achieve big data objectives for most organizations will involve partnering with expert organizations whose primary business is to develop and maintain these skills.

**Working with Big Data**

**Six Common Scenarios**

There are many different approaches to deriving value from big data. From the questions you ask to the way you analyze the answers, each scenario requires a different approach to mining big data for vital intelligence. It depends on the nature of the data you’re presented and what you want to get out of it.

Overall, there are six typical big data scenarios. They include:

1. *Bringing unstructured data into structured environments*
2. *Combining different kinds of structured data by passing it through an unstructured layer*
3. *Gathering marketing intelligence from social media*
4. *Analyzing log files in equipment, such as routers, web servers, switches, and sensors*
5. *Fueling effective consumer recommendation systems by assessing purchase patterns on a massive scale*
6. *Conducting market analyses to tailor offerings, going beyond one’s own data stores to pull from web logs and others’ sites*

What’s common in every case is that the practical insights lead to better business decisions—insights that would otherwise have been completely invisible or overlooked. Effective big data mining accelerates time to insight, leading to more precise decisions and greater returns.

**Where Do You Begin?**

One of the most basic, yet overlooked, places to begin is with a simple question: What do you want to use this information for? Once you determine answers to this question, you can start to prepare a plan.

It may not be clear at the outset exactly what value can be mined from a specific dataset. In those cases, rather than risk losing data of value, a smart approach is to store it and sort it later.
Modern data storage and hosting capabilities are ubiquitous and cheap enough to allow companies to store petabytes of data for a fraction of what it used to cost. Once the data is stored, you can unleash your data scientists and let them work their magic.

In other situations, it may be wisest to process and filter data at the front end of the process. This applies when data is coming from outside sources—usually not in any useful condition and with the potential that the majority of it may be irrelevant anyway. Good processing and filtering can reduce the volume and variety of data. This requires massive computing power, but it will cut down on the amount of storage you need if you keep only the useful one percent.

For many enterprises, it’s not just the scale of the data that’s daunting; it’s also the scope—the extent to which it touches the organization. Trying to do all things big data at once can overwhelm your resources or take too long to realize results. It’s often a good idea to carve off a small initial project to pilot your big data approach and scale incrementally—fueled by the achievement of quick, clear, obvious progress.

There are tools to do this, like Oracle, MySQL, or Hadoop, a platform that allows huge quantities of unstructured data to be parsed and queried, but you always want the right tool for the job. Oracle and MySQL are useful for structured data and repeatable reports, while Hadoop is best at scanning large and unstructured datasets, looking for unexpected insights and cheaper scalability. Integrating various data processing systems into your data-value chain is the key to distilling new insights from your data.

Once you have a set of results from your analysis, implementing it effectively is critical and this often involves trial and error. When you’re applying the results of data analyses, whether they deal with customer service, purchasing or supply lines, ecommerce or recommendations, it is critical to be able to measure the results of the changes you’ve made based on your conclusions. This brings everything back full circle, because the measurements you collect are then contributed to your volumes of big data and stored for future analysis.

The Power of Big Data

We live in an age where almost everything is captured and cached, pinged and processed, and it all leaves a digital trail. From baseball teams to politics to retailers, data-driven decision-making has become mainstream. Online media and social sites mine user-behavior data to determine what interests whom—when, why, and how. Online retailers sift through big data to predict consumer behavior, identify purchase preferences and design high-impact offers that will drive up sales per session. Insurance companies leverage big data to improve their overall performance through more accurate pricing, deeper customer relationships, and more effective and efficient loss prevention.

Insights into customer and consumer trends worth billions of dollars, avenues for innovation, and strategies for streamlining operations—they’re all available in the digital bytes created and stored in data centers around the world, every instant. Distilling the useful information out of all of that
digital noise is the next frontier in marketing and customer service, as well as streamlining and optimizing business practices. The companies that are best at mining big data will be the ones that stay ahead of the curve.
About Pythian

Pythian is a global data management consulting company specializing in planning, deploying, and managing mission-critical data infrastructures. Since 1997, Pythian has optimized business-critical databases to help companies strategically align their IT and business goals. Learn more about Pythian and its elite data experts at http://www.pythian.com, follow @Pythian, or find Pythian on LinkedIn at http://linkd.in/pythian.

Produced in Canada by The Pythian Group Inc
April 2013

Pythian, The Pythian Group, “love your data”, and pythian.com are trademarks of The Pythian Group Inc.