IDC
Analyze the Future

I D C I V I E W

THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East

December 2012

By John Gantz and David Reinsel

Sponsored by EMC Corporation

Content for this paper is excerpted directly from the IDC iView "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," December 2012, sponsored by EMC. The multimedia content can be viewed at ***www.emc.com/leadership/digital-universe/index.htm****.*

Executive Summary: A Universe of Opportunities and Challenges

Welcome to the "digital universe" — a measure of all the digital data created, replicated, and consumed in a single year. It's also a projection of the size of that universe to the end of the decade. The digital universe is made up of images and videos on mobile phones uploaded to YouTube, digital movies populating the pixels of our high-definition TVs, banking data swiped in an ATM, security footage at airports and major events such as the Olympic Games, subatomic collisions recorded by the Large Hadron Collider at CERN, transponders recording highway tolls, voice calls zipping through digital phone lines, and texting as a widespread means of communications.

With the rise of Big Data awareness and analytics technology, the digital universe in 2012 has taken on the feel of a tangible geography — a vast, barely charted place full of promise and danger. The digital universe lives increasingly in a computing cloud, above terra firma of vast hardware datacenters linked to billions of distributed devices, all governed and defined by increasingly intelligent software.

In this context, at the midpoint of a longitudinal study starting with data collected in 2005[1] and extending to 2020, our analysis shows a continuously expanding, increasingly complex, and ever more interesting digital universe. This is IDC's sixth annual study of the digital universe, and it's chock-full of new findings:

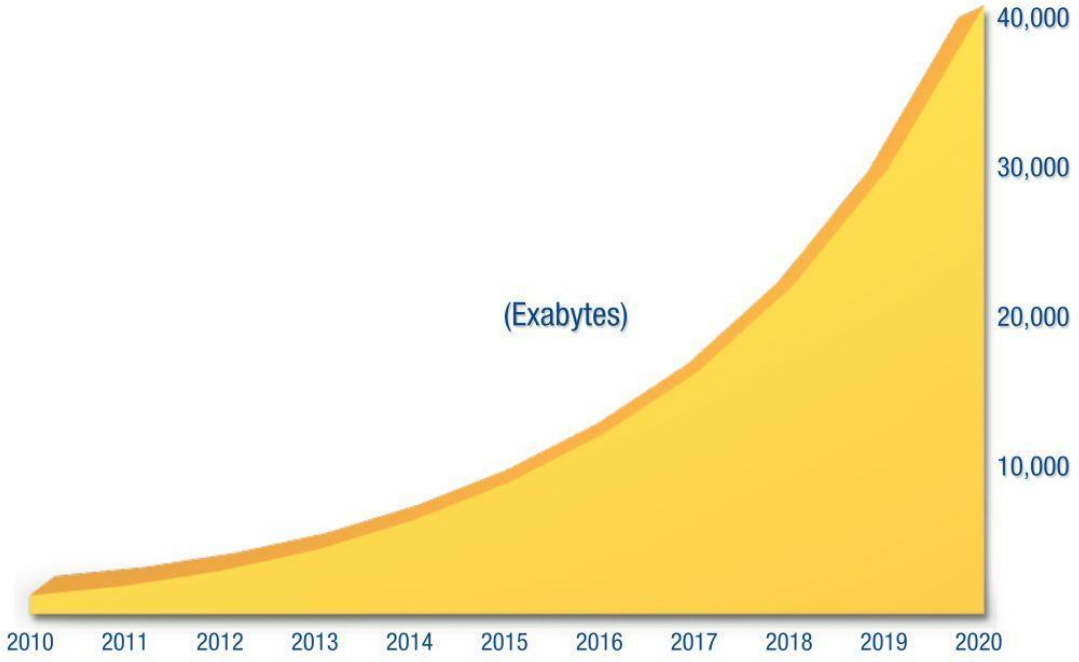
- From 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes (more than 5,200 gigabytes for every man, woman, and child in 2020). From now until 2020, the digital universe will about double every two years.
- The investment in spending on IT hardware, software, services, telecommunications and staff that could be considered the "infrastructure" of the digital universe and telecommunications will grow by 40% between 2012 and 2020. As a result, the investment per gigabyte (GB) during that same period will drop from $2.00 to $0.20. Of course, investment in targeted areas like storage management, security, big data, and cloud computing will grow considerably faster.

[1] The first *Digital Universe Study* was published in 2007 (see http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf). At that time, IDC's forecast for the digital universe in 2010 was 988 exabytes. Based on actuals, it was later revised to 1,227 exabytes.

- Between 2012 and 2020, emerging markets' share of the expanding digital universe will grow from 36% to 62%.
- A majority of the information in the digital universe, 68% in 2012, is created and consumed by consumers — watching digital TV, interacting with social media, sending camera phone images and videos between devices and around the Internet, and so on. Yet enterprises have liability or responsibility for nearly 80% of the information in the digital universe. They deal with issues of copyright, privacy, and compliance with regulations even when the data zipping through their networks and server farms is created and consumed by consumers.
- Only a tiny fraction of the digital universe has been explored for analytic value. IDC estimates that by 2020, as much as 33% of the digital universe will contain information that might be valuable if analyzed.
- By 2020, nearly 40% of the information in the digital universe will be "touched" by cloud computing providers — meaning that a byte will be stored or processed in a cloud somewhere in its journey from originator to disposal.
- The proportion of data in the digital universe that requires protection is growing faster than the digital universe itself, from less than a third in 2010 to more than 40% in 2020.
- The amount of information individuals create themselves — writing documents, taking pictures, downloading music, etc. — is far less than the amount of information being created *about them* in the digital universe.
- Much of the digital universe is transient — phone calls that are not recorded, digital TV images that are watched (or "consumed") that are not saved, packets temporarily stored in routers, digital surveillance images purged from memory when new images come in, and so on. Unused storage bits installed throughout the digital universe will grow by a factor of 8 between 2012 and 2020 but will still be less than a quarter of the total digital universe in 2020.

Figure 1

50-Fold Growth from the Beginning of 2010 to the end of 2020

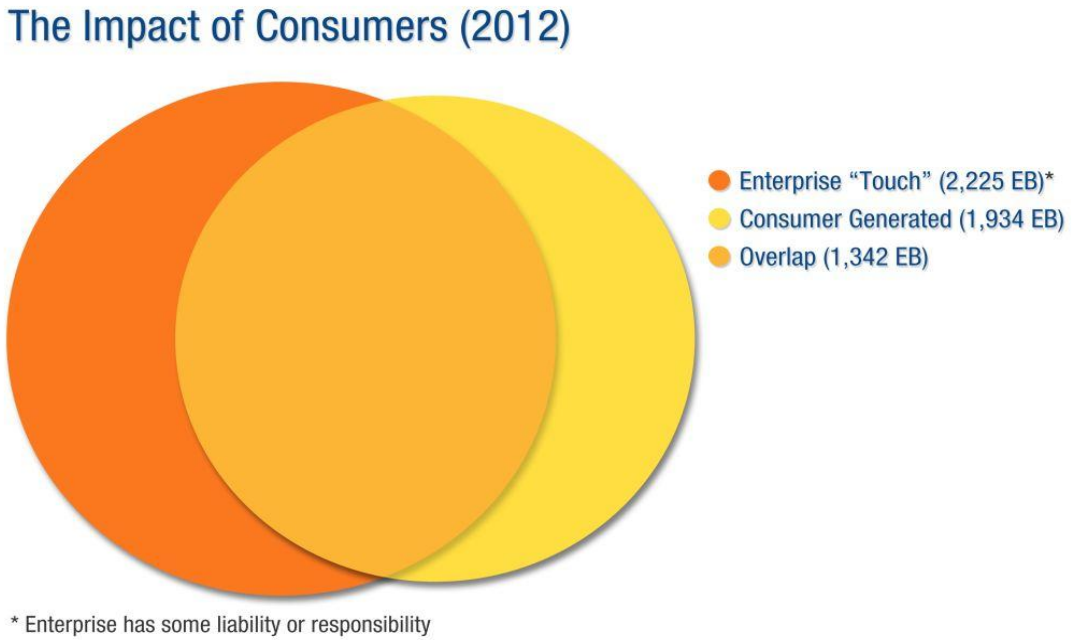


Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Within these broad outlines of the digital universe are some singularities worth noting.

First, while the portion of the digital universe holding potential analytic value is growing, only a tiny fraction of territory has been explored. IDC estimates that by 2020, as much as 33% of the digital universe will contain information that might be valuable if analyzed, compared with 25% today. This untapped value could be found in patterns in social media usage, correlations in scientific data from discrete studies, medical information intersected with sociological data, faces in security footage, and so on. However, even with a generous estimate, the amount of information in the digital universe that is "tagged" accounts for only about 3% of the digital universe in 2012, and that which is analyzed is half a percent of the digital universe. Herein is the promise of "Big Data" technology — the extraction of value from the large untapped pools of data in the digital universe.

Figure 2

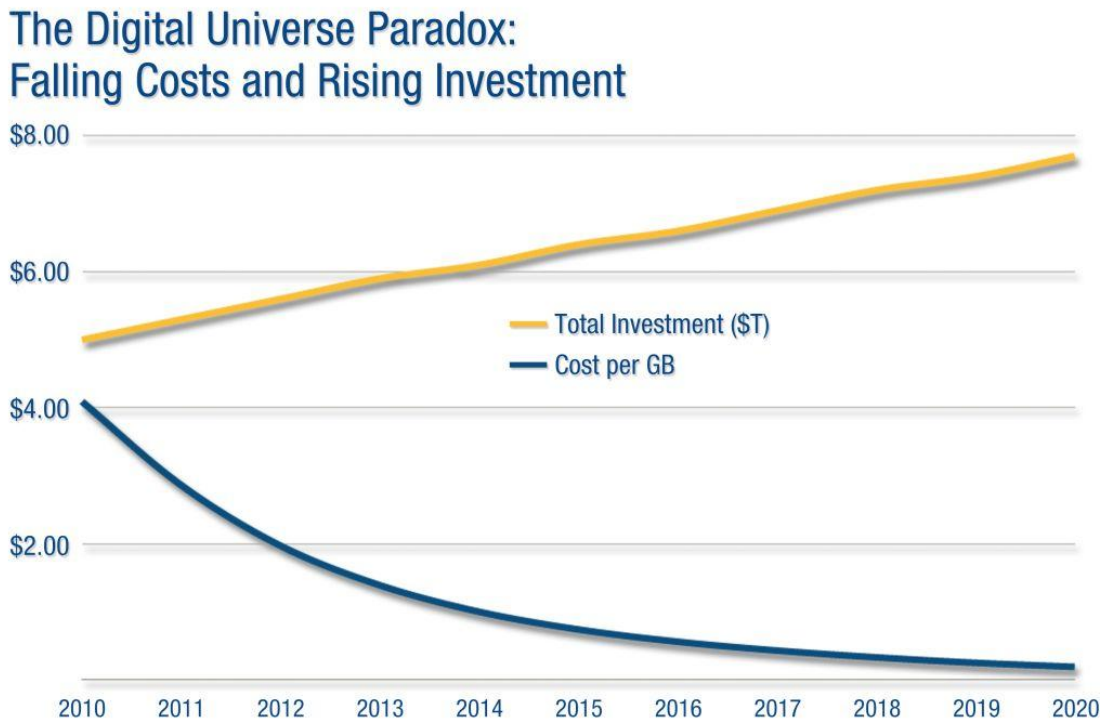


Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Moreover, IDC believes that much of the digital universe is unprotected. Our estimate is that about a third of the data in the digital universe requires some type of protection — to protect privacy, adhere to regulations, or prevent digital snooping or theft. However, currently, only about 20% of the digital universe actually has these protections. The level of protection varies by region, with much less protection in emerging markets.

Therefore, like our own physical universe, the digital universe is rapidly expanding and incredibly diverse, with vast regions that are unexplored and some that are, frankly, scary.

Figure 3



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

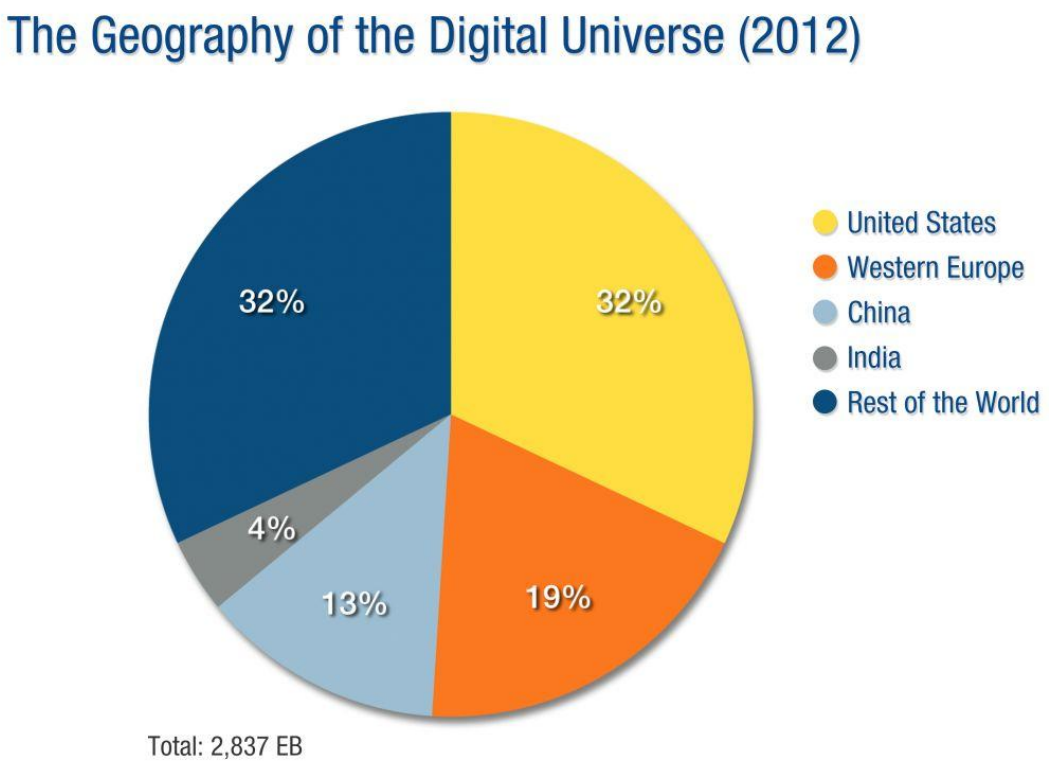
However, the digital universe astronauts among us — the CIOs, data scientists, digital entrepreneurs — already know the value that can be found in this ever-expanding collection of digital bits. Hence, there is excitement about Big Data technologies, automatic tagging algorithms, real-time analytics, social media data mining, and myriad new storage technologies.

The Geography of the Digital Universe

Although the bits of the digital universe may travel at Internet speeds around the globe, it is possible to assign a place of origin to them and chart the map of the digital universe.

In this year's study, for the first time, we have managed to determine where the information in the digital universe was either generated, first captured, or consumed. This geography of the digital universe maps to the users of the devices or applications that pump bits into the digital universe or pull bits into one's own personal digital solar system for the purpose of consuming information — Internet users, digital TV watchers, structures hosting surveillance cameras, sensors on plant floors, and so on.

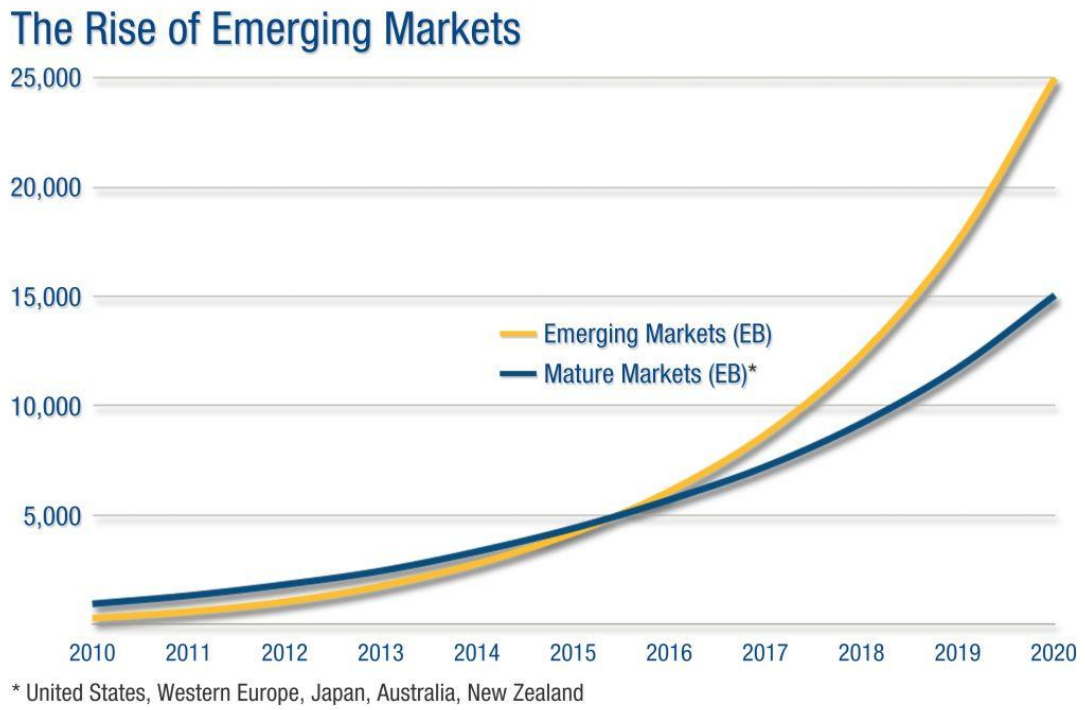
Figure 4



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

In the early days, the digital universe was a developed world phenomenon, with 48% of the digital universe in 2005 springing forth from just the United States and Western Europe. Emerging markets accounted for less than 20%. However, the share of the digital universe attributable to emerging markets is up to 36% in 2012 and will be 62% by 2020. By then, China alone will generate 21% of the bit stream entering the digital universe.

Figure 5



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

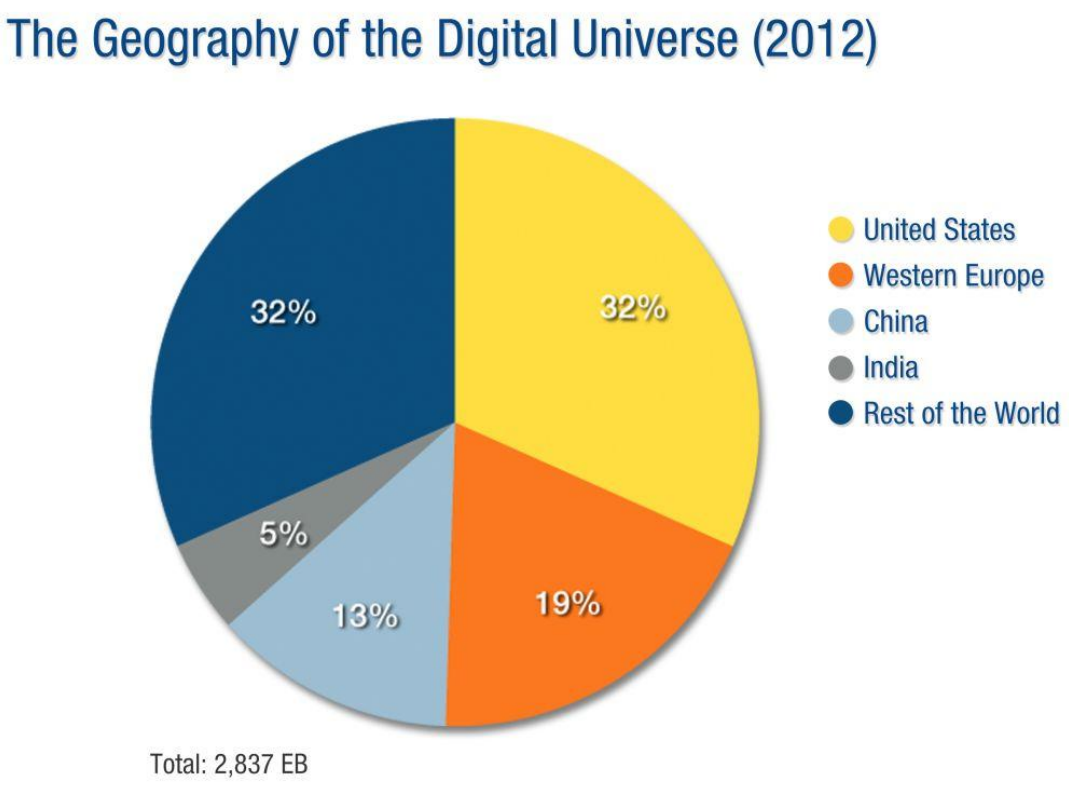
It stands to reason. Even though China accounts for only 11% of global GDP today, by 2020 it will account for 40% of the PCs, nearly 30% of smartphones, and nearly 30% of Internet users on the planet — not to mention 20% of the world population.

At the same time, the money invested by the regions in creating, managing, and storing their portions of the digital universe will vary wildly — in real dollar terms and as a cost per gigabyte.

This disparity in investment per gigabyte represents to some extent differing economic conditions — such as the cost of labor — and to some extent a difference in the types of information created, replicated, or consumed. The cost per gigabyte from bits generated by surveillance cameras will be different from the cost per gigabyte from bits generated by camera phones.

However, to *another* extent, this disparity also represents differences in the sophistication of the underlying IT, content, and information industries — and may represent a challenge for emerging markets when it comes to managing, securing, and analyzing their respective portions of the digital universe.

Figure 6



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

This might not be a major issue if the geography of the digital universe were as stable and fixed as, say, the geography of countries. However, bits created in one part of the physical world can easily find themselves elsewhere, and if they come with malware attached or leaky privacy protections, it's a problem. The digital universe is like a digital commons, with all countries sharing some responsibility for it.

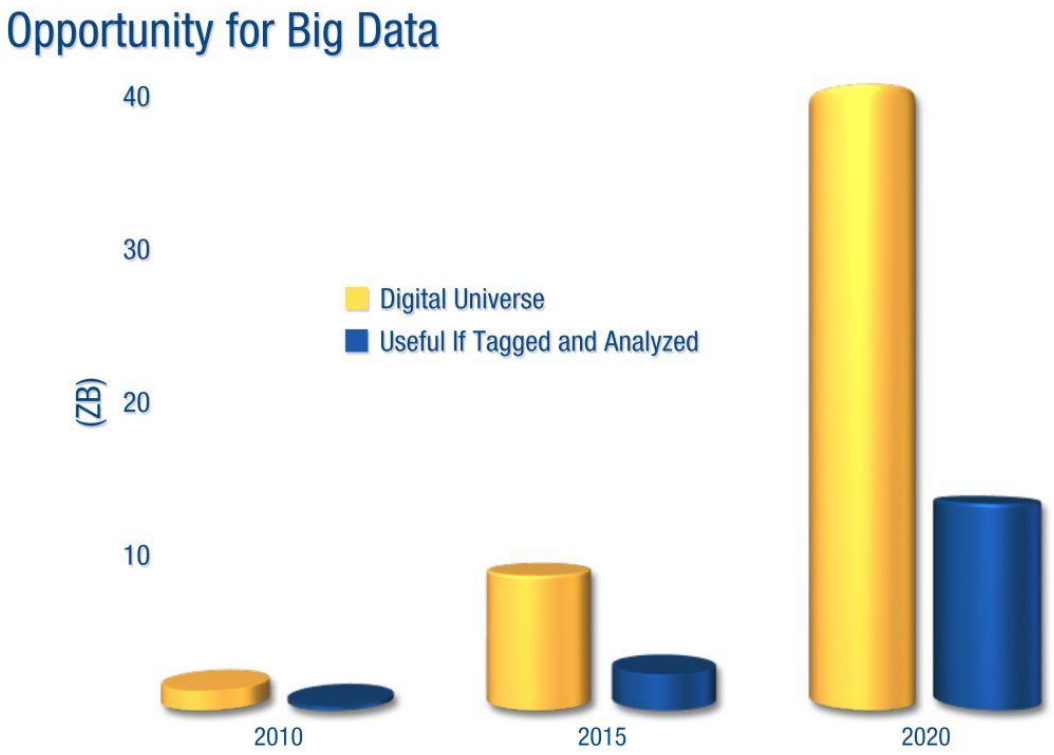
The installed base of unused storage bits introduces an interesting geographic twist that establishes a new dynamic by which to understand our digital universe. While emerging markets may indeed grow as a percentage of the digital universe, remember that much of the digital universe is a result of massive consumption on mobile and personal devices, digital televisions, and cloud-connected applications on PCs. As ownership of smartphones and tablets (that have relatively low internal storage and rely heavily on consuming information from "the cloud") increases exponentially within emerging markets, information consumption grows at an even faster pace. Given the connected infrastructure of our digital universe, information does not need to (and in fact will not) reside within the region where the information is consumed. Hence, today's well-running datacenters will continue to expand and to fulfill an increasing number of requests — both local and from halfway across the globe — for information.

Big Data in 2020

Last year, Big Data became a big topic across nearly every area of IT. IDC defines Big Data technologies as *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*. There are three main characteristics of Big Data: the data itself, the analytics of the data, and the presentation of the results of the analytics. Then there are the products and services that can be wrapped around one or all of these Big Data elements.

The digital universe itself, of course, comprises data — all kinds of data. However, the vast majority of new data being generated is unstructured. This means that more often than not, we know little about the data, unless it is somehow characterized or tagged — a practice that results in metadata. Metadata is one of the fastest-growing subsegments of the digital universe (though metadata itself is a small part of the digital universe overall). We believe that by 2020, a third of the data in the digital universe (more than 13,000 exabytes) will have Big Data value, but only if it is tagged and analyzed (see "Opportunity for Big Data).

Figure 7



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Not all data is necessarily useful for Big Data analytics. However, some data types are particularly ripe for analysis, such as:

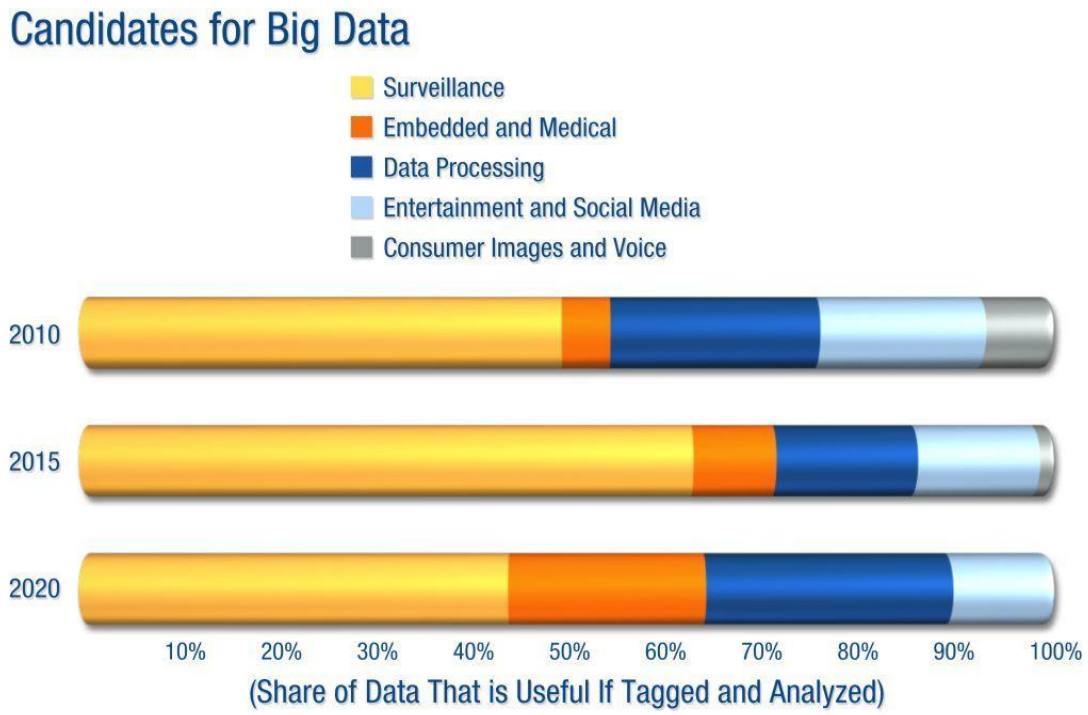
- **Surveillance footage.** Typically, generic metadata (date, time, location, etc.) is automatically attached to a video file. However, as IP cameras continue to proliferate, there is greater opportunity to embed more intelligence into the camera (on the edge) so that footage can be

captured, analyzed, and tagged in real time. This type of tagging can expedite crime investigations, enhance retail analytics for consumer traffic patterns, and, of course, improve military intelligence as videos from drones across multiple geographies are compared for pattern correlations, crowd emergence and response, or measuring the effectiveness of counterinsurgency.

- **Embedded and medical devices.** In the future, sensors of all types (including those that may be implanted into the body) will capture vital and nonvital biometrics, track medicine effectiveness, correlate bodily activity with health, monitor potential outbreaks of viruses, etc. — all in real time.
- **Entertainment and social media.** Trends based on crowds or massive groups of individuals can be a great source of Big Data to help bring to market the "next big thing," help pick winners and losers in the stock market, and yes, even predict the outcome of elections — all based on information users freely publish through social outlets.
- **Consumer images.** We say a lot about ourselves when we post pictures of ourselves or our families or friends. A picture used to be worth a thousand words, but the advent of Big Data has introduced a significant multiplier. The key will be the introduction of sophisticated tagging algorithms that can analyze images either in real time when pictures are taken or uploaded or en masse after they are aggregated from various Web sites.

These are in addition, of course, to the normal transactional data running through enterprise computers in the course of normal data processing today. "Candidates for Big Data" illustrates the opportunity for Big Data analytics in just these areas alone.

Figure 8

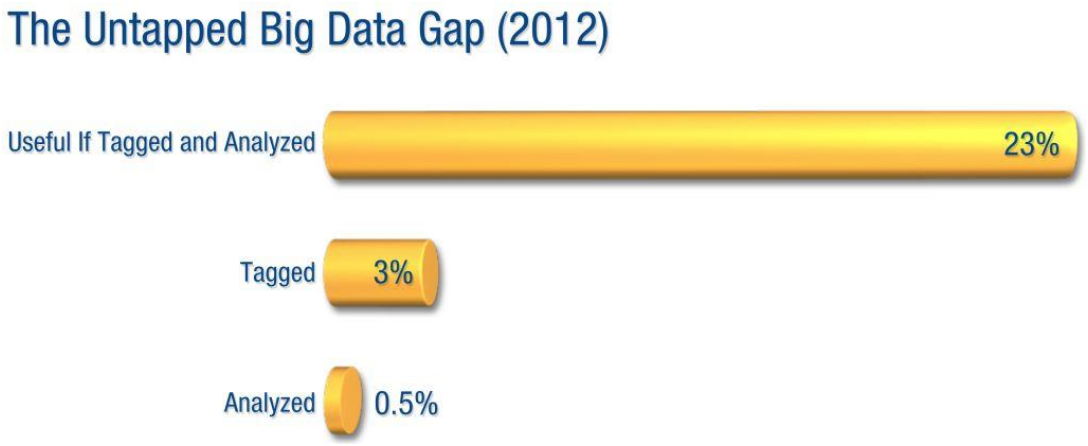


Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

All in all, in 2012, we believe 23% of the information in the digital universe (or 643 exabytes) would be useful for Big Data if it were tagged and analyzed. However, technology is far from where it needs to be, and in practice, we think only 3% of the potentially useful data is tagged, and even less is analyzed.

Call this the Big Data gap — information that is untapped, ready for enterprising digital explorers to extract the hidden value in the data. The bad news: This will take hard work and significant investment. The good news: As the digital universe expands, so does the amount of useful data within it.

Figure 9



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Information Security in 2020

The rise in mobility and participation in social networks, the increasing willingness to share more and more data, new technology that captures more data about data, and the growing business around Big Data all have at least one assured outcome — the need for information security.

However, the news from the digital universe is as follows:

- The proportion of data in the digital universe that requires protection is growing faster than the digital universe itself, from less than a third in 2010 to more than 40% in 2020.
- Only about half the information that *needs* protection *has* protection. That may improve slightly by 2020, as some of the better-secured information categories will grow faster than the digital universe itself, but it still means that the amount of unprotected data will grow by a factor of 26.
- Emerging markets have even less protection than mature markets.

In our annual studies, we have defined, for the sake of analysis, five levels of security that can be associated with data having some level of sensitivity:

1. **Privacy** only — an email address on a YouTube upload
2. **Compliance** driven — emails that might be discoverable in litigation or subject to retention rules
3. **Custodial** — account information, a breach of which could lead to or aid in identity theft

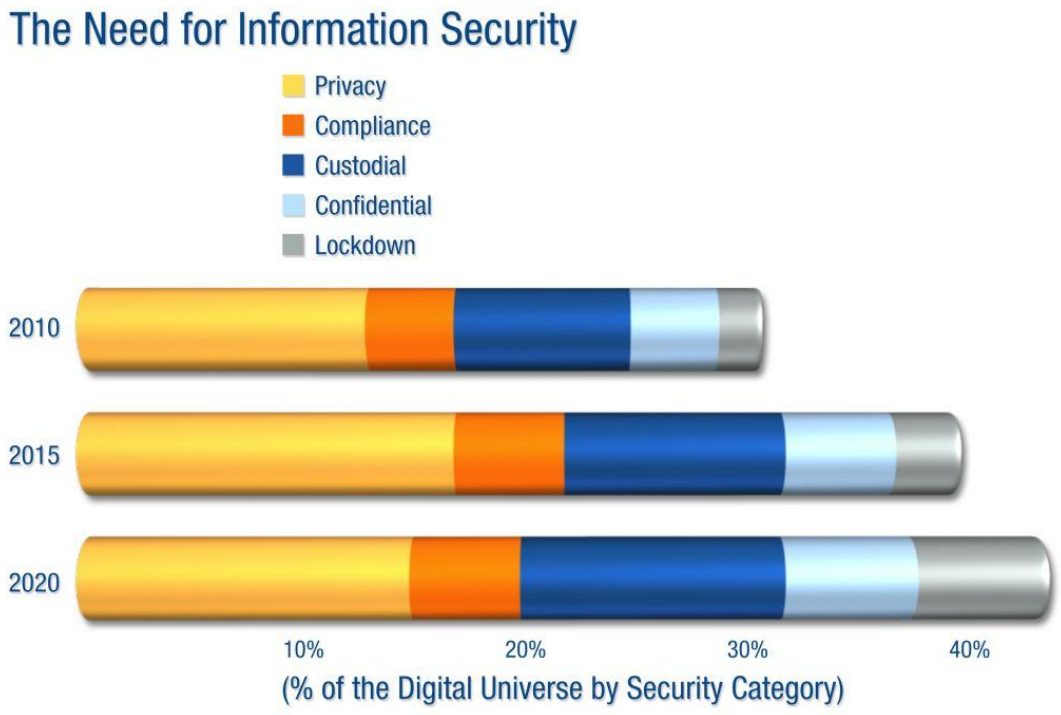
4. **Confidential** — information the originator wants to protect, such as trade secrets, customer lists, confidential memos, etc.

5. **Lockdown** — information requiring the highest security, such as financial transactions, personnel files, medical records, military intelligence, etc.

The tables and charts illustrate the scope of the security challenge but not the solution. While information security technology keeps getting better, so do the skills and tools of those trying to circumvent these protections. Just follow the news on groups such as Anonymous and the discussions of cyberwarfare.

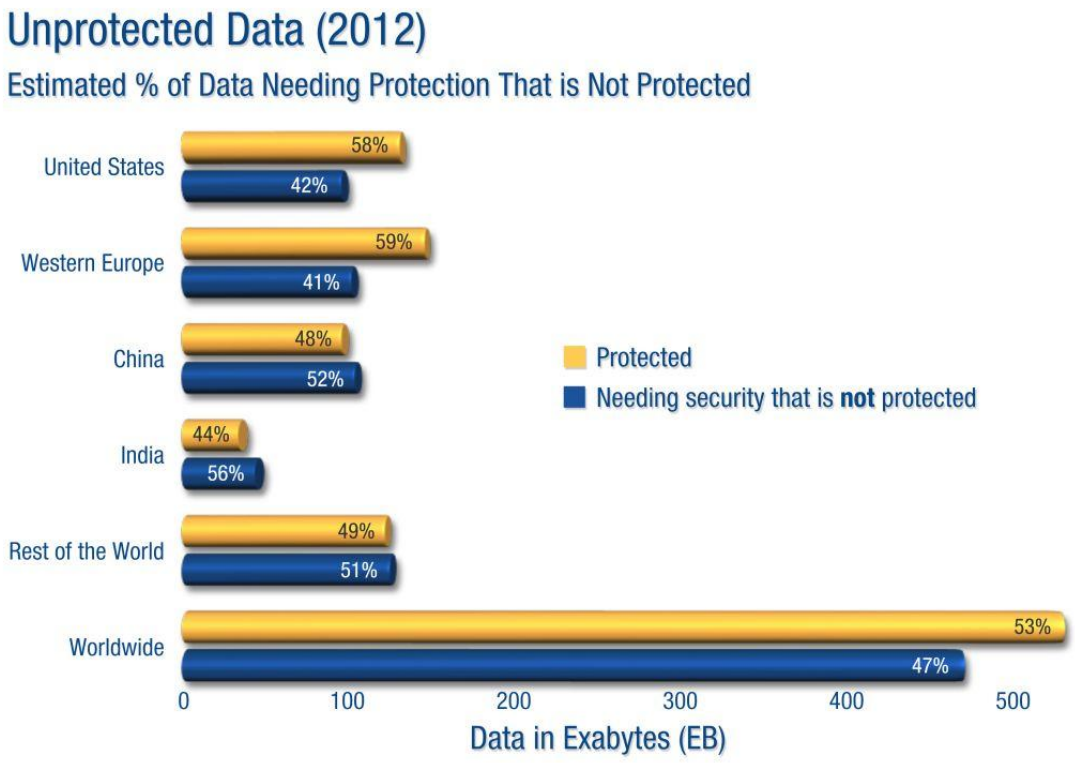
However, for enterprises and, for that matter, consumers, the issues may be more sociological or organizational than technological — data that is not backed up, two-phase security that is ignored, and corporate policies that are overlooked. Technological solutions will improve, but they will be ineffective if consumer and corporate behavior doesn't change.

Figure 10



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Figure 11



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Big Data is of particular concern when it comes to information security. The lack of standards among ecommerce sites, the openness of customers, the sophistication of phishers, and the tenacity of hackers place considerable private information at risk. For example, what one retailer may keep private about your purchase, such as your transaction and customer profile data, another company may not and instead may have other data hidden. Yet intersecting these data sets with other seemingly disparate data sets may open up wide security holes and make public what should be private information.

There is a huge need for standardization among retail and financial Web sites as well as any other type of Web site that may save, collect, and gather private information so that individuals' private information is kept that way.

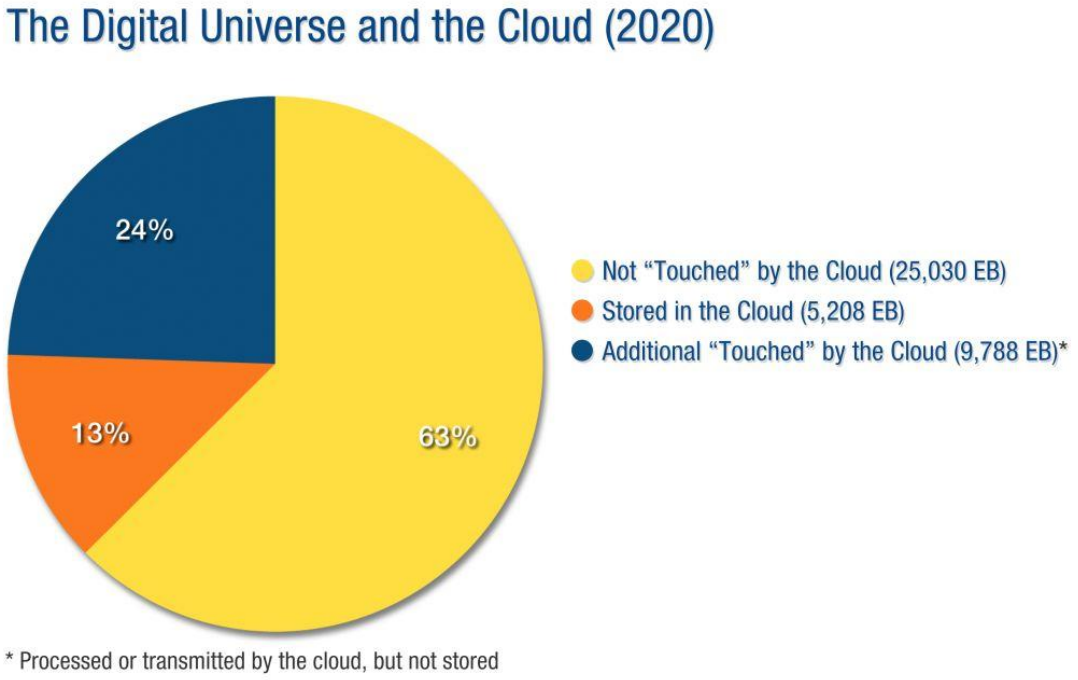
Cloud Computing in 2020

Between 2012 and 2020, the patch of the digital universe that CIOs and their IT staffs need to manage will become not just bigger but also more complex. The skills, experience, and resources to manage all these bits of data will become scarcer and more specialized, requiring a new, flexible, and scalable IT infrastructure that extends beyond the enterprise: cloud computing.

To this end, the number of servers (virtual and physical) worldwide will grow by a factor of 10 and the amount of information managed directly by enterprise datacenters will grow by a factor of 14. Meanwhile, the number of IT professionals in the world will grow by less than a factor of 1.5.

In addition, while spending on public and private cloud computing accounts for less than 5% of total IT spending today, IDC estimates that by 2020, nearly 40% of the information in the digital universe will be "touched" by cloud computing — meaning that a byte will be stored or processed in a cloud somewhere in its journey from originator to disposal. Perhaps as much as 15% will be *maintained* in a cloud.

Figure 12

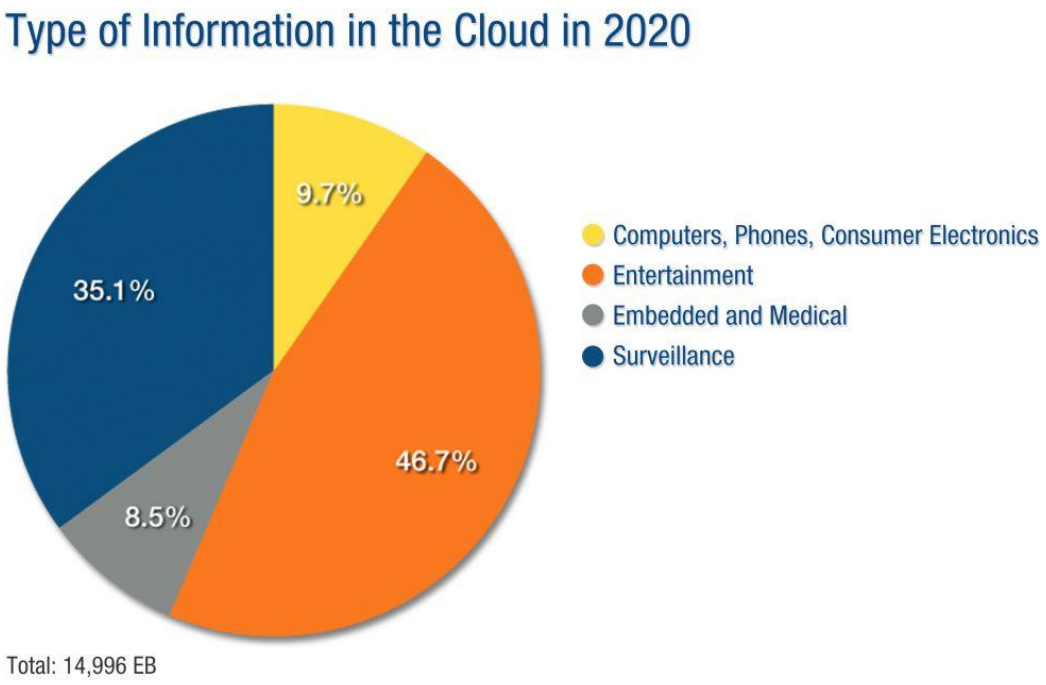


Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Of course, cloud services come in various flavors — public, private, and hybrid. For organizations to offer their own cloud services, they have to do more than just run virtual servers. They must also allow for virtualized storage and networking, self-provisioning, and self-service and provide information security and billing.

Part of the real genesis of this conversion to the cloud will be a migration to converged infrastructures, where servers, storage, and networks are integrated together, sold, and installed as a unit of IT infrastructure. Few enterprises are at this point yet, so the impact of private clouds in the digital universe today is small.

Figure 13



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

However, by 2020, it seems likely that private clouds and public clouds will be commonplace, exchanging data seamlessly. There won't be one cloud; rather, there will be many clouds, bounded by geography, technology, different standards, industry, and perhaps even vendor. We may still call it cloud computing, but it will be an interconnected ether, easy to traverse but difficult to protect or manage.

Call to Action

Our digital universe in 2020 will be bigger than ever, more valuable than ever, and more volatile than ever.

By 2020, we'll also be storing a smaller and smaller percentage of our expanding digital universe; yet our digital shadows will be larger than life and on the move given the increase in mobility, and they will require more protection than ever before. IT managers will be responsible not only for ensuring that proper security surrounds our digital lives but also for managing the storing, analyzing, and delivery of zettabytes of content …no easy task.

Requests for data could come from a faraway jungle, across a mashup of connected devices and network points, to a device that has an obtuse screen. The delivery of the requested data must happen in an acceptable amount of time, guaranteeing that it is consumed flawlessly; if not, then a business may lose a customer. Consider this:

- The network is growing in importance. Latencies must get shorter, not longer. Data must be analyzed, security applied, and authentication verified — all in real time and in levels yet to be seen. Network infrastructure must be a key investment to prepare for our 2020 digital universe.
- Big Data is going to be a big boon for the IT industry. Web sites that gather significant data need to find ways to monetize this asset. Data scientists must be absolutely sure that the intersection of disparate data sets yields repeatable results if new businesses are going to emerge and thrive. Further, companies that deliver the most creative and meaningful ways to display the results of Big Data analytics will be coveted and sought after.
- The laws and regulations governing information security must harmonize around the globe, though differences (or absences) will certainly exist. IT managers must realize that data will be requested outside geographic boundaries, and a global knowledge of information security may be the difference between approval and denial of a data request.
- IT managers must find ways to drive more efficiency in their infrastructures so that IT administrators can focus on more value-add initiatives such as "bring your own device" (BYOD) policies, Big Data analytics, customer onboarding efficiency, security, etc. One way this is likely to happen is through converged infrastructures, which integrate storage, servers, and networks.

Are you ready to create, consume, and manage 40 trillion gigabytes of data?
