Big Data's Impact on the Data Supply Chain

The proliferation of social, mobile and traditional Web computing poses numerous business opportunities and technological challenges for information services companies. By revamping their enterprise level data supply chains and IT infrastructures, as well as embracing partnerships that deliver analytical acumen, information services companies with the right organizational mindset and service delivery model can survive, if not thrive, amid the data onslaught.

Executive Summary

The information industry is in the middle of a data revolution. An unimaginably vast amount of data is growing exponentially, providing challenges and opportunities for data-centric companies to unlock new sources of economic value, take better credit and financial risks, spot business trends sooner, provide fresh insights into industry developments and create new products. At the same time, data abundance also creates daunting challenges across numerous dimensions when it comes to the ways and means of handling and analyzing data.

This white paper examines the challenges that confront the information services industry and offers guidance on ways the industry can rethink its operating models and embrace new tools and techniques to build a next-generation data supply chain (DSC). Our approach is premised on a platform for turbocharging business performance that converts massive volumes of data spawned by social, mobile and traditional Web-based computing into meaningful and engaging insights. To address the information explosion coupled with the dynamically changing data consumption patterns, this next-generation platform will also induce information providers to roll out innovative products at a faster pace to beat the competition from both existing and niche information players and meet the demand for value added, real-time data and quicker solutions for end consumers.

Information Services Trends

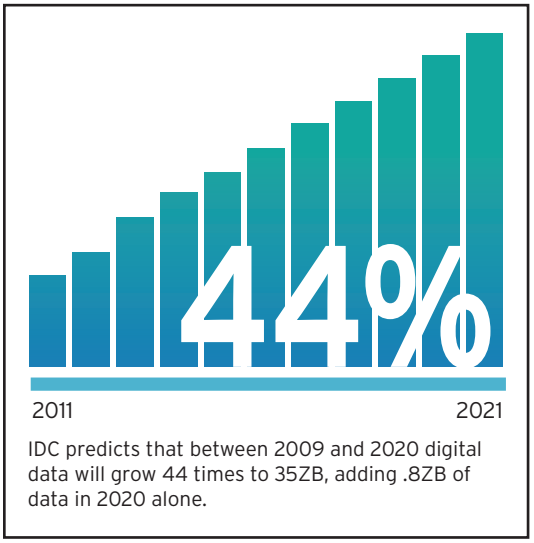
A deep-dive into today's information industry reveals the following patterns.

Continued Data Deluge

Approximately five exabytes (EB)[1] of data online in 2002 rose to 750 EB in 2009 and by 2021 it is projected to cross the 35 zettabytes (ZB)[2] level as seen in Figure 1. Statistics[3] also indicate that 90% of the data in the world was created in the last two years, a sum greater than the amount of the data generated in the last 40 years. The world's leading commercial information providers deal with more than 200 million business records, refreshing them more than 1.5 million times a day to provide accurate information to a host of businesses and consumers. They source data from various organizations in over 250 countries, 100 languages and cover around 200 currencies.

Cognizant

Data Volumes Are Growing



IDC predicts that between 2009 and 2020 digital data will grow 44 times to 35ZB, adding .8ZB of data in 2020 alone.

Figure 1

Their databases are updated every four to five seconds. With new data about companies instantaneously created and parameters of existing companies worldwide changing by the minute, the challenge will only intensify.

The world's leading providers of science and health information, for example, address the needs of over 30 million scientists, students and health and information professionals worldwide. They churn out 2,000 journals, 20,000 books and major reference works each year. Users on Twitter send more than 250 million tweets per day. Almost 50 hours of video are uploaded per minute on YouTube by hundreds of millions of users worldwide. Facebook houses more than 90 billion photos with over 200 million photos uploaded per day.[4]

The big data phenomenon threatens to break the existing data supply chain of many information providers, particularly those whose chains are neither flexible nor scalable and include too many error-prone, manual touch points.

While knowledge is hidden in these exabytes of free data, data formats and sources are proliferating. The value of data extends to analytics about the data, metadata and taxonomy constructs. With new electronic devices, technology and people churning out massive amounts of content by the fractional second, data is exploding not just in volume but also in diversity, structure and degree of authority. Figure 2 provides an indicative estimate of the data bombarding the Internet in one 60 second interval, with astounding growth forecasted.

Ever-changing data consumption patterns and the associated technology landscape raise data security and privacy issues as data multiplies and is shared freely ever more. Convergence challenges related to unstructured and structured data also add to the worries. Google, Facebook, LinkedIn and Twitter are eminent threats to established information players as are nontraditional niche information providers such as Bloomberg Law, DeepDyve, OregonLaws, OpenRegs, etc. that provide precise information targeted at specific customer groups. The big data phenomenon threatens to break the existing data supply chain (DSC) of many information providers, particularly those whose chains are neither flexible nor scalable and include too many error-prone, manual touch points. For instance, latency in processing all data updates in the existing DSC of one well-known provider currently ranges from 15 to 20 days, versus a target of 24 hours or less. That directly translates to revenue loss, customer dissatisfaction and competitive disadvantage.

These risks are real. Google reported a 20% revenue loss with the increased time to display search results by as little as 500 milliseconds. Amazon reported a 1% sales decrease for an additional delay of as little as 100 milliseconds.

"Free" Information Business Models, with Data as the New Fuel

Companies such as Duedil, Cortera and Jigsaw (recently acquired by Salesforce.com and renamed Data.com) are revolutionizing the "free" business model. The Jigsaw model, for instance, uses crowdsourcing to acquire and deliver a marketplace for users to exchange business contact information, worldwide. For sharing information on non-active prospects that these prospects would gladly do, users get new leads for free. If a user finds incorrect data, he gets points by updating the record in Jigsaw. Providing incentives to have users scrub the huge database enables Jigsaw to more easily maintain data integrity.

Essentially, users actively source and update contacts in the Jigsaw database in return for free access to the company's services. Thus, from a data quality, data entry scalability and data maintenance perspective (issues that typically plague systems such as those in the CRM space), Jigsaw is a strong tool that can be used to append incomplete records, augment leads to build highly

The Digital Data Deluge — One Minute's Worth of Data Flow

Figure 2

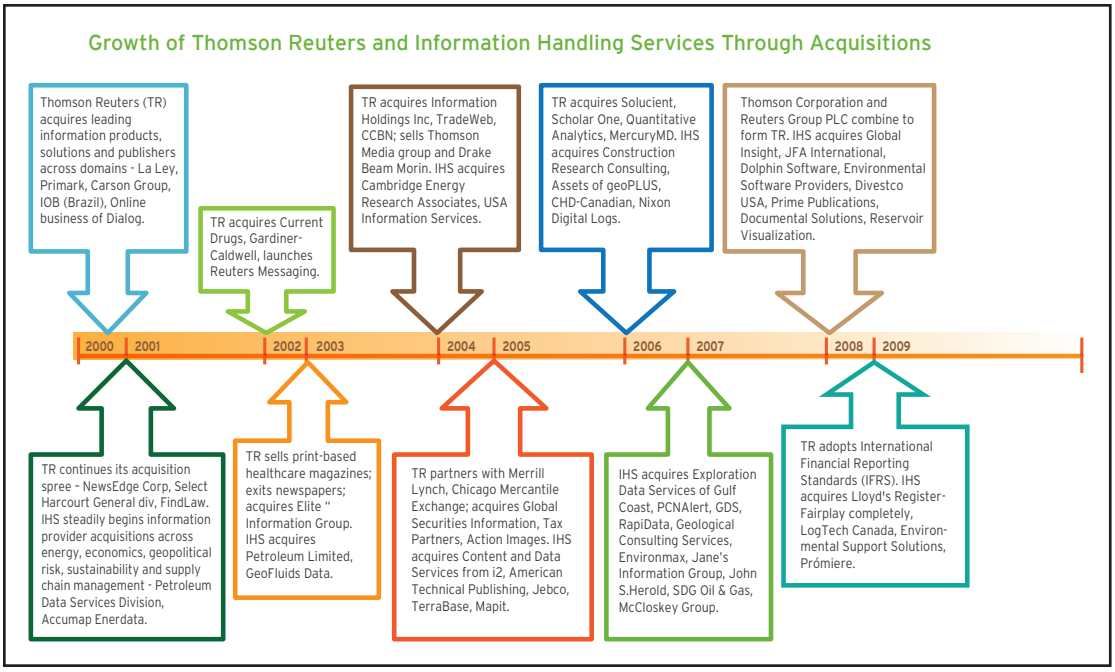
targeted lists, plan territories and gain insights on people and companies with the most complete B2B data in one place. Jigsaw has already amassed more than 30-plus million contacts and is growing. It sells this information to customers with large CRM databases who can compare their database to the Jigsaw database, identifying and cleaning up any redundant records. Jigsaw contacts then make money by offering products geared toward companies interested in increasing, updating and cleaning their contact directories. These free models intensify competition for traditional data aggregators.

In addition to Google, which operates on an ad-supported (free) model, others like WebMD (a health information provider) rely on advertising to generate a major portion of their revenue streams, enabling them to provide free services. They then make additional money from subscriptions and premium content, as well as listings from individuals who initially come to avail themselves of free services and end up paying for a listing in order to heighten awareness for existing and new customers. Such models are allowing newer entrants to underprice the competition or to offer some portions of their information portfolio for free. As such, this approach threatens traditional information providers, forcing them to step up.

How can information services companies compete and remain cost leaders? Many of their existing DSC systems provide neither enough insight nor a more robust understanding of their customers — and nor do they reveal how their end customers are interacting with their data. Their existing DSC is not built for handling big data and the corresponding big data analytics cannot be effectively applied and leveraged to shed light on what the data means or provide a pathway to reduce IT infrastructure costs to attain greener operations. Moreover, many information players and their existing DSC systems are not really leveraging social media and its related opportunities to increase customer engagement, improve content quality and provide incremental value to the ecosystem.

We are in an era where we trade our data for free goods and services. Never before have consumers wielded this much power over marketers. All the data activity on the Internet, through any device, creates click-trails, leaves digital breadcrumbs, produces data exhaust and creates metadata. There is enough economic value in this data for an entire industry to be formed around this itself. We will see a huge influx of companies dealing with the various aspects of data drilling, shipping, refining, drug discovery and so on. Hence, based

Thomson Reuters and Information Handling Services Growth via Acquisitions



Source: Cognizant analysis of Thomson Reuters and IHS data published on each company's Web site.
Figure 3

on the above precedent, large players like Exxon, Mobil, Pfizer or Merck could create large stand-alone data-slicing organizations.

Mergers and Acquisitions, Industry Consolidation and Footprint Expansion

Many information providers have expanded into new markets through M&As and with local partnerships. They seek to integrate all acquired companies into a single enterprise-level DSC. Having started in legal publishing, Thomson Reuters now has a footprint across various information domains, such as healthcare, tax and accounting, intellectual property, financial, media, risk and compliance and even science. A leading financial information provider has recently moved its data collection and storage operations in Italy to a partner. It has also bought tools for data interoperability between enterprise-level services.

Some players are also consolidating data programs to find synergies in their business line operations. They also want newer data sources to enhance data quality and variety. Figure 3 depicts how two large players, Thomson Reuters (TR) and Information Handling Services (IHS), have grown through acquisition during the last decade. This model is proving more attractive as data processing scale, distribution and brand power becomes ever more critical.

Acquisitions cause significant problems for companies' DSCs. There is almost certain to be data quality loss from disparate systems and operational inefficiencies caused by the lack of a unified view of data. DSC integration issues cause increased latency, slower time to value and customer access problems. Many existing DSCs were built as stand-alone systems with closed architectures, and have undergone many customizations. This makes integration difficult, raising costs and slowing payback time. It also increases maintenance and ongoing enhancement costs. Integrating newer functionalities developed using advanced integrated development environments (IDE), debugging and automation tools makes the development lifecycle an extremely complex task and transferring taxonomies becomes complicated. For these archaic systems, lack of productivity tools and limited hardware and software options result in greater time to market to meet dynamic business requirements or regulatory compliance.

As the industry grapples with the information explosion, the question on every CIO's mind is

how they can handle, manage and analyze this data avalanche better. From the aforementioned points, what clearly emerges is a definite need for the information providers to reexamine their existing DSC for potential solutions. They should leverage their strategic and technology partner capabilities in this discovery and eventual implementation process. Starting points include:

- Are the providers ready to take advantage of the above tipping points to emerge as lean and agile players to increase shareholder value?
- How can providers help users find relevant and compact information in a flood of big data?

To address such issues, this paper explores one key starting point, what we've termed the "next-generation data supply chain." It conceptualizes at a high level current and emerging elements embedded in a DSC that can help enable new solutions and explore opportunities, partnerships and alliances to enhance the value chain. This paper uses "data" and "information" interchangeably as data forms the foundation for any insightful information; increasingly, the two are becoming difficult to distinguish.

The Next-Generation Data Supply Chain

By reengineering the existing DSC, from data sourcing through data delivery, providers can transform their ability to ingest, process and distribute content under a wide variety of new business models. The key objective is to create a next-generation DSC that:

- Optimizes operational efficiencies.
- Reduces data latency.
- Is flexible to accommodate new data sources.
- Is scalable to handle future data volumes.
- Improves data quality while dynamically meeting consumer demands.
- Explores newer monetization models with data as an asset.
- Provides faster time to market and the potential for greater revenue recognition.

By reengineering the existing DSC, from data sourcing through data delivery, providers can transform their ability to ingest, process and distribute content under a wide variety of new business models.

Figure 4 represents paradigms that could create a truly modern DSC. The following subsections present salient thoughts around some of the prime components of such an upgraded DSC that could address current and futuristic data issues.

Data Sourcing and Collection

This category includes business process outsourcing (BPO), business process as a service (BPaaS)[5]

DSC at a Glance

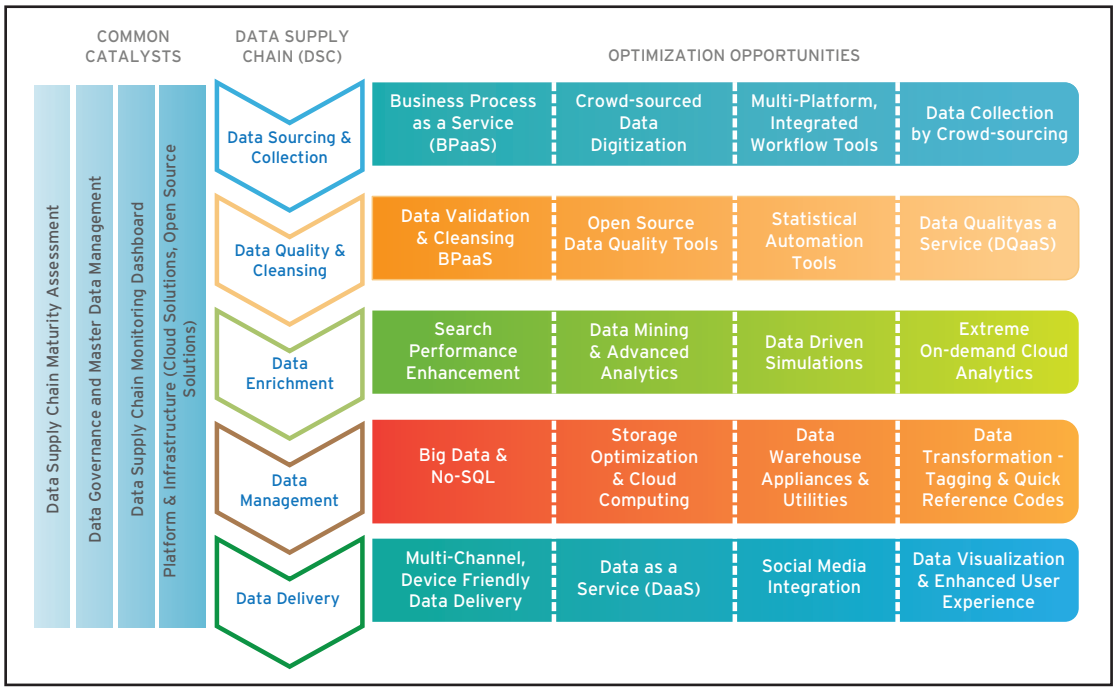


Figure 4

and crowdsourcing. Typically, organizations have treaded the traditional BPO path by focusing only on daily operational tasks to deliver the outcome. But is there year-round work for data operations to have a dedicated BPO team? Until the need for a dedicated BPO team is felt, it is tempting to have a temporary pay-per-use team but "crowdsourcing," as mentioned earlier, can be considered as a replacement for traditional outsourcing. The crowd refers to the users, who are volunteers with common, social, financial or even intellectual motivation to accomplish a task. They share solutions for mutual benefit with the crowdsourcer (organization or an individual) who is the problem owner.

With electronic devices, portables, Internet and rapidly evolving digital applications growing in importance, it becomes imperative to adopt data digitization techniques and procedures to create a truly global marketplace.

Such distributed problem-solving addresses scalability through worldwide access to people and data almost free of cost and generates innovative results by simplifying an otherwise complex task that is too difficult to handle internally. There are numerous examples of projects using the crowd to successfully collect and analyze data, some of which are noted later in this paper. Operating cost assessments for crowdsourced data using US$/full-time equivalent (FTE)/hour suggests savings of about 60% (in the U.S.) to around 65% to 70% (in India) over traditional outsourcing. But how dependable is crowdsourcing? The crowd may not work in all situations. Situations where work is intermittent is one place where crowdsourcing might not work. In this case, BPaaS could be a more practical middle road approach to integrate business, tools and information technology and to achieve greater opportunities to optimize operations and drive efficiency and flexibility. The BPaaS model retains a lightweight in-house data reporting team to interact with an outsourced team who are specialists in handling data production and validation with tools residing on the Cloud. BPaaS pushes standardization across many data sources and embodies a flexible pay-per-use pricing model.

The BPaaS proposition comes with infrastructure, platform and software "as a service" models without affecting the traditional benefits of outsourcing such as process expertise and labor arbitrage. The cloud combination provides hyper-scalability and the ability to deliver solutions affordably. Financial benefits are in tune with reduced operating costs of up to 30%, thereby cutting capital expenditures on up-front investments. Although BPaaS is not a panacea, the benefits of a variable pricing model combined with technology and business process excellence that reduce or eliminate large capital demands certainly go beyond cost savings. BPaaS needs to be embraced as a next-generation delivery model.

Data digitization entails conversion of physical or manual records such as text, images, video and audio to digital form. To address the proliferation of nonstandard data formats from multiple data sources globally, and to preserve and archive data in an orderly manner with simple access to information and its dissemination, data must be standardized. Data digitization does that. With electronic devices, portables, Internet and rapidly evolving digital applications growing in importance, it becomes imperative to adopt data digitization techniques and procedures to create a truly global marketplace. Embracing this will not only help providers to contend with the convergence of structured and unstructured data, but also enable the producer to reach the consumer directly, cutting out inefficient layers. But how you do it depends on how your DSC is organized.

For example, how do you digitize 20 years of newspaper archives in less than three months? *The New York Times* did exactly that by using the power of collective human minds. *The Times* showed archived words (as scanned images) from the newspaper archives to people who are filling out online forms across different Web sites to spell the words and help digitize the text from these archives. The subscribing Web sites (generally unrelated to the digitization process) present these images for humans to decipher and transform into text, as part of their normal validation procedures that even optical character recognition (OCR) procedures cannot interpret properly. Text is useful because scanned newspaper images are difficult to store on small devices, expensive to download and cannot be searched. The images also protect any suspicious interventions from any automated software programs or "bots," ensuring that only humans validate the words. The sites then return the results to a software service that captures and aggregates this digitized textual information. With the system reported to display over 200 million words, scalability of the crowd provides

quick results as in each case the human effort is just a few seconds and represents digitization and validation at its best.

Crowdsourced data also makes a lot of sense, particularly as business people and consumers increasingly rely on mobile devices and social media platforms to share data more freely across geographies. An excellent example of this is the GPS navigation where the foundation is the map database. Rather than rely solely on a map provider database that may not necessarily be up to date or accurate, via crowdsourcing users report map errors and new map features. Thus users can benefit immensely from each other's reports at no cost.

Crowdsourcing is a brilliant way to collect massive data as it brings down the cost of setting up a data collection unit. However, the data provided by the user community has to be made credible through data verification by the data quality tools. Although crowdsourcing might affect data quality, by looking at a large base of users the outliers in data could be easily found and eliminated.

Data Quality and Cleansing

High-quality data is a prime differentiator and it's a valuable competitive asset that increases efficiency, enhances customer service and drives profitability. The cost of poor data quality for a typical enterprise is estimated to cost 8% to 12% of revenues. British Gas lost around £180M when data quality problems caused its project to fail, resulting in degraded customer relationships and contract cancellations. ABN Amro was fined $80M for not having effective data quality compliance. Severn Trent Water was fined £26M by regulators for trying to cover up data quality issues created by its data migration project.[6]

Traditionally, companies have been shortsighted when it comes to data quality by not having a full lifecycle view. They have implemented source system quality controls that only address the point of origin, but that alone is not enough. Data quality initiatives have been one-off affairs at an IT level rather than collective efforts of both IT and the business side of the house. Failure to implement comprehensive and automated data cleansing processes that identify data quality issues on an ongoing basis results in organizations overspending on data quality and its related cleansing. These issues will only increase in number and complexity with the increasing data sources that must be integrated. A flexible data quality strategy is potentially required to tackle a broad range of generic and specific business rules and also adhere to a variety of data quality standards.

There is a need to incorporate data quality components directly into the data integration architecture that address three fundamental data quality activities: profiling, cleansing and auditing. Profiling helps resolve the data integrity issue and makes unaudited data or extracts acceptable as baseline datasets. Cleansing ensures that outlining and business rules are met. And auditing evaluates how the data meets different quality standards. To improve customer experiences, leading to higher loyalty and competitive differentiation, providers have to look beyond custom-built solutions and seek vendor assistance to maximize their results. Data quality as a service (DQaaS), should be an integral part of data quality as it allows for a centralized approach. With a single update and entry point for all data controlled by data services, quality of data automatically improves as there is a single best version that enters the DSC. DQaaS is not limited to the way data is delivered. The solution is also simple and cost conscious as data access is devoid of any need to know the complexities of the underlying data. There are also various pricing approaches that make it flexible and popular to adopt, be it quantity or subscriptions based, "pay per call to API" based or data type based. While there are a wide array of tools from AbInitio, Microsoft SSIS, IBM Ascential, Informatica, Uniserv, Oracle DW Builder, etc. to choose from, there are also open source data quality and data integration tools offered such as Google Refine. Given the size of Google and its open source reach, the company's products should be seriously considered.

Failure to implement comprehensive and automated data cleansing processes that identify data quality issues on an ongoing basis results in organizations overspending on data quality and its related cleansing.

Open source tools are powerful for working with messy datasets, including cleaning up inconsistencies, and transforming them from one format into another. Hence a combination of open source and proprietary tools will help achieve benefits of both worlds.

Data Enrichment

If we are looking at data mining and doing sentiment analysis of 120,000 Tweet feeds per

second, the enrichment components will be different than, say, handling 2,000 transactions per second as done by Visa. While data mining and search performance optimization has been an integral part of data enrichment, approaches such as sentiment or predictive analytics and data-driven simulations enable more effective extraction and analysis of data, leading to more informed decisions. That said, search is still the basis for any insightful data extraction. There is a need to evolve search by constantly integrating social data, both structured and unstructured, to result in true-to-life recommendations for smarter use. Hence efforts should continue to fine-tune search algorithms with semantic expertise focused on users to provide relevant answers (not just results) in fractions of a second. Fine-tuning the search not only increases targeted traffic and visibility, but it will also provide high return on investment. The benefits of search optimization efforts have shown increased conversion rates of over 30% for a leading American retailer in the first two weeks of use, while revenue increased over $300 million for a leading e-commerce player, to cite two examples.

As a direct impact of the various information trends adding to the data deluge, the term "big data" has come into prominence. This term is often used when referring to petabytes, exabytes and yet greater quantities of data and generally refers to the voluminous amount of structured and unstructured data that takes too much time, effort and money to handle. Since it has to do with extremely large data sets that will only increase in the future, big data and its analytics require a different treatment. Hence, there is also a need to augment platforms such as Hadoop[7] to store Web-scale data and support complex Web analytics. Since Hadoop uses frameworks to distribute the large data sets, processing loads amongst hundreds or even thousands of computers, it should be explored as an enterprise platform for extreme enterprise analytics — that is, extremely complex analytics on extremely large data volumes. Since one of the prime objectives of any supply chain is an excellent user experience, combining critical information and insight capabilities using advanced analytics is the way forward. Furthermore, to extract the best performance through hardware and software integration, tools such as Oracle Exalytics can be adopted to accelerate the speed at which analytics algorithms run. Such tools provide real-time visual analysis, and enable new types of analytic applications. They enable quicker decision-making in the context of rapidly shifting business conditions through the introduction of interactive visualization capabilities. Applications can be scaled across the enterprise with faster, more accurate planning cycles.

With big data tools like Hadoop and extreme analytics, enrichment components can crunch data faster and deliver better results. In addition, analytics can be transformed if information services with their global services partners can build a focused team of data scientists who understand cloud computing and big data and who can analyze and visualize traffic trends. They combine the skills of techie, statistician and a narrator to extract the diamonds hidden within mountains of data. All this will enhance information services providers' abilities to recommend the right data sets to the right audience in real time and hence divert more traffic, resulting in higher revenue-generating opportunities.

Data Management

Cloud computing presents a viable solution to provide the required scalability and to enhance business agility. Leading cloud storage providers such as Amazon Web Services, Google and Microsoft's Azure are steadily cutting prices for their cloud services. Instead of making millions of dollars in upfront investments on infrastructure, providers can quickly convert Cap-Ex to Op-Ex and pay for data services as they go. Furthermore, the cost for data storage is declining significantly. Providers could simply tweak their data classification schema to optimize storage for even bigger savings. For example, by moving from a three-tier classification to a six-tier one,[8] one large manufacturer cut its storage cost by almost $9 million for 180 TB of data.

Going by analysts' predictions on the impact of cloud on the business of storage, there's $50 billion of gross margin of enterprise storage in play with the transition to cloud providers. Data center optimization cum consolidation cannot be neglected either. A Fortune 50 company optimized its 100,000-square-foot data center[9] in Silicon Valley with the help of a solutions provider, resulting in $766,000 per year in annual energy savings, a $530,000 Silicon Valley power rebate and a three-month return on investment. Virtualization in all its forms, including server virtualization and storage virtualization, gets more computing power out of servers while delivering

a more adaptable data center that will enable big data to be crunched efficiently while requiring less electric power (and thus is greener). Providers should hence consider cloud and storage optimization as part of their DSC transformation.

Big Data Fundamentally Changes Data Management

Since information providers have to inevitably face the big data revolution, they need to be prepared by building the big data technology stacks. Apart from Hadoop, as discussed above, some of the necessary components in this architecture stack include Cassandra, a hybrid non-relational database that provides flexible schema, true scalability and multi-datacenter awareness; and No-SQL databases. No-SQL databases offer a next-generation environment that is non-relational, distributed, open-source and horizontally scalable. These capabilities need to be baked into the enhanced data supply value chain through a service-oriented architecture (SOA) that will enhance unstructured and structured data management and make providers future-ready by integrating widely disparate applications for a Web-based environment and using multiple implementation platforms. Data warehouse appliances and utilities will be a blessing in disguise that extends beyond a traditional data warehouse, providing robust business intelligence. They are easy, affordable, powerful and optimized for intensive analytical analysis and performance with speedy time-to-value. Open source clustered file systems like the Hadoop Distributed File Systems (HDFS) and its alternatives as well as the Google file systems also resolves some of the leading big data challenges. They are better in performance and provide cost-efficient scalability. They are hardware and operating system agnostic, making them flexible and easy to design and manage and provide industry-leading security in cloud deployments. LexisNexis High-Performance Computing Cluster (HPCC) and Appistry's cloud IQ storage Hadoop editions are examples of data supply chains built on clustered file systems storage.

Automation is an integral part of this transformation and some of the aforementioned tools use a combination of data and artificial intelligence to cut repetitive tasks of managing various aspects of the value chain. This would definitely free up personnel to focus on core strategic issues rather than on tasks that can be easily automated. Classifying databases for topicality is one of the ways to eliminate data redundancy and enhance search efficiencies; with just-in-time data updates in the production environment, overall latency is reduced. Quick response (QR) code is another optimizing measure for digitally packing more data than the erstwhile bar codes (enabling faster readability, especially through mobile devices). Modern search engines recognize these codes to determine the freshness of Web site content. To stay on top of search listings, providers need to switch to QR codes to increase revenue conversion rates as they provide quick and convenient user access to their Web sites and hence capture more of the user's attention, particularly when used in advertisements. All these have opened the door for more innovative and revenue-generating improvements and savings.

LexisNexis High-Performance Computing Cluster (HPCC) and Appistry's cloud IQ storage Hadoop editions are examples of data supply chains built on clustered file systems storage.

Data Delivery

How data is accessed, explored and visualized has been disrupted by the continued expansion and availability of mobile devices. Any information services company that has not moved away from largely static and visually uninteresting data representations, device incompatible applications or Web applications that don't support the next-age social media platforms should do so with urgency. They don't have to start by custom building from scratch. They can adopt data visualization application suites available either commercially or outsourced to their technology partners that can help build automated graphic data representation tools or revealing infographic aggregations quickly, or provide their proprietary packages at a lower cost. This will not only enhance real-time data but also seamlessly integrate with other devices, tools and software.

Classifying databases for topicality is one of the ways to eliminate data redundancy and enhance search efficiencies; with just-in-time data updates in the production environment, overall latency is reduced.

Device applications development holds a multiplicity of business opportunities to increase audience reach and customer engagement. To get there, information services companies should consider partnering with their service providers

to change the way data is viewed. This should be combined with a user-friendly interface to an external rules engine and repository that business users can use to modify or validate business rules and hence add value by further authenticating data before delivery. With the business world speeding toward mobility, the way data is presented and accessed must be altered to accommodate consumers on the move. Finally, there is tremendous opportunity to strengthen information providers' client relationships and grow their businesses, leveraging a strong social media-influenced DSC strategy. Providers will add credibility and promote attention among key user constituencies by enabling their data delivery through social aggregators since this approach enhances data and collaboration, creates a buzz by distributing and publicizing new information releases, builds virtual interactive communities and enables user-generated information to enhance core research curation. Revenue generating and increasing avenues will be the direct outcome of such a social media integration measure.

Assessment, Monitoring and Data Governance

One of the logical starting points is to perform an overall assessment of the existing DSC. An alternative approach would be that instead of considering the entire value chain in one stroke, to examine each step over time. This could take the form of an assessment of subcomponents of the DSC (as per specific pain points on a client-to-client basis) through a consulting exercise. Through such assessments, the consulting team can:

- Expose the "as-is state" of the current DSC and its capabilities; this would also include issues and potential risks.
- Arrive at a "to-be state," unlocking potential opportunities, through findings and recommendations.
- Establish a business case and ROI to quantify benefits, including justification and resourcing to realize the desired value-chain transformation.
- Road-map the data value chain implementation process and present a high-level architecture in a prioritized and time-phased fashion.

Once the DSC is baselined, the next step is to instrument the supply chain and vigorously monitor performance across the following prime dimensions: data latency, data quality, scalability, flexibility and cost of operations. A single organization-wide dashboard with logging capabilities and a user-friendly interface that reports these metrics in real time will enable effective data auditing and eventually strong data governance.

Finally, the DSC is only as good as the data which goes into it. Hence, having a systematic data governance practice with checks on data consistency, correctness and completeness will help maintain the DSC without too much effort and ensure adequate data privacy. Companies normally have data policies, but these policies are merely used to satisfy compliance obligations. Current data governance programs have not yet attained maturity but companies that have data governance typically show a 40% improvement in ROI[10] for IT investments compared to companies that don't have it.

Challenges to achieving an effective next-generation DSC include:

- The very thought of an end-to-end revamp of the entire DSC is overwhelming and the amount of time, money and effort involved is daunting to leaders.
- Without CXO-level sponsorship and motivation, there is little chance of success.
- Given the significant change management involved, and without a critical mass of catalysts from the data provider and consumer communities, frustration can ensue.
- A DSC can only be successful if the data is standardized, as otherwise the organization must write custom code to standardize, clean and integrate the data.
- Cloud computing and many of the "as-a-service" models rely on the service provider's ability to avoid service downtime.

The Path Ahead

Staying ahead of the curve and emerging victorious will require information services companies across industries and domains to embrace a next-generation DSC, selecting key elements and service delivery models to unlock productivity, seize competitive advantage and optimize the business for dynamic changes in the market. As we delve deeper to understand the architectural, infrastructural, technical and even business model implications, there is further scope for innovation.

With pressure to cut costs further while at the same time modernize, there will be an evolving need for additional models, methods, frameworks, infrastructure and techniques that allow providers to tackle a data avalanche that will only increase. Information providers should collaborate with their current or potential strategic and implementation partners to mutually explore areas in domain, products, services and technology to "wow" the end consumer.

Whether organizations expand existing architectures or start afresh by building, buying or acquiring new capabilities for a more modern and future-proof DSC, they will need to quickly make strategic decisions by carefully trading off risks and rewards that ensure coexistence with today's business constraints and tomorrow's demands for real-time, anywhere, anytime information access. Information providers that postpone the decision to face the inevitable trends in the information industry as discussed in this paper will find themselves stuck with rigid legacy environments and will be eventually overtaken by forward-looking and more innovative competitors. If the vision is to be the market leader in world-class information and to be the consumer's preferred choice for insightful information and decision-making, it is high time for information players to act now.

Footnotes

1. 1 Exabyte (EB) = 10^{18} bytes and is equal to 2^{60} in binary usage.
2. 1 Zettabyte (ZB) = 10^{21} bytes and is equal to 2^{70} in binary usage.
3. "Where angels will tread," *The Economist*, http://www.economist.com/node/21537967
4. Data points obtained directly from the following company Web sites: Elsevier, Twitter, YouTube and Facebook.
5. Business process as a service, or BPaaS, is an application delivered as a service that is used by service-provider teams that perform business tasks on behalf of the service recipient. BPaaS combines traditional business process outsourcing (BPO) and software as a service (SaaS) to optimize business processes and elevate outcomes.
6. British Gas, ABN Amro and Severn Trent Water examples: "Business Value for Data Quality," http://www.x88.com/whitepapers/x88_pandora_data_quality_management.pdf
7. Hadoop is a free (open source) Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is primarily conceived on a MapReduce paradigm that breaks applications into smaller chunks to be processed in a distributed fashion for rapid data processing.
8. "How to save millions through storage optimization," http://www.networkworld.com/supp/2007/ndc3/052107-storage-optimization-side.html
9. "SynapSense Bolsters Data Center Infrastructure Management Solutions with New Energy Management Tools," http://www.reuters.com/article/2011/06/29/idUS251193+29-Jun-2011+BW20110629
10. "IT Governance: How Top Performers Manage IT Decision Rights for Superior Results," Peter Weill and Jeanne Ross, Harvard Business School Press.

About the Author

Sethuraman M.S. is the Domain Lead for Information Services within Cognizant's Information, Media and Entertainment (IME) Consulting Practice. Apart from providing leadership to a team that consults to information services companies, he provides thought leadership built on his expertise in the information markets and global media. His extensive IT experience spans consulting, program management, software implementations and business development. Sethu has worked with clients worldwide and has won numerous awards and recognitions. He received a BTech degree in instrumentation and electronics from the National Institute of Technology (NIT), Punjab, India, an MS degree in software systems from Birla Institute of Technology and Science (BITS), Pilani, India, and an MBA from the Asian Institute of Management (AIM), Manila, Philippines. Sethu can be reached at Sethuraman.MS@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 140,500 employees as of March 31, 2012, 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
1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102
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