IT’s Challenge: Bringing Structure to the Unstructured World of Big Data

Big data is a significant transformational disruptor. Savvy IT leaders will capitalize on the opportunities.
Whether or not big data has yet reached its inflection point, there’s no doubt that it is a transformational disrupter for IT organizations and the businesses they support and enable. From building new infrastructures to hiring people with new and emerging skill sets, big data will undoubtedly challenge, expand, and advance the IT function for years to come. In addition, the advanced analytics and decision support systems enabled by big data will revolutionize current ways of conducting and managing business.

As we discussed in our publication, Big Data and the Creative Destruction of Today’s Business Models, big data and business intelligence are converging: Computing resources are better equipped to handle data growth and complexity, and a clearly articulated vision for big data reaps the greatest rewards. Yet, as data volumes with varying degrees of structure explode in every dimension—the amount of business data is doubling every 1.2 years, according to one estimate—businesses are presenting their IT organizations with staggering amounts of data to organize and process. This rapid proliferation has exposed gaping shortfalls in strategy, infrastructure (hardware and software), and organization (talent and structure) required to make effective use of data.

At the same time, the recent A.T. Kearney-Carnegie Mellon University Leadership in Excellence in Analytic Practices (LEAP) study, which examined 430 companies around the world covering 11 industries, found that success in analytics is directly tied to executive sponsorship of the analytics function. For leading companies, this sponsorship is at the very top of the chart: COO, CFO, or CMO; at the laggards, analytics resides in IT. Nevertheless, the survey highlights the need for IT to be an active participant in building a successful analytics program. Taken together, these results suggest IT organizations will be under greater pressure to evolve, increase workforce skills, and collaborate in nontraditional ways across the organization (see figure 1).

Figure 1
Analytics success is directly tied to executive-level sponsorship

![Analytics success is directly tied to executive-level sponsorship](image)


In this paper, we examine the major trends that must be addressed by IT organizations as they plan for the future. Then, we break down the core architectural components of big data into eight distinct functional groups that can help lay the foundation for a comprehensive platform for companies’ effective adoption of big data.
Perspectives on a Rapidly Changing Space

Understanding the major trends in big data is central to preparing for the changes ahead. These rapidly advancing trends are challenging many of the fundamental assumptions about IT strategy and planning.

1. **Open source is the past, present, and future of big data.** Enterprises are increasingly adopting open-source technologies, as more see open-source software as a competitive differentiator that enables agility, mitigates risk, and lowers costs.

   In big data, open source, led by Hadoop, is driving the most significant innovations. With low implementation costs and high adoption levels—including direct support from trend-setting technology organizations such as Facebook, Twitter, Amazon, and LinkedIn—open-source software is spreading. Emerging open-source frameworks and technologies of particular interest in big data include Storm, Kafka, and S4 for stream processing, Drill and Dremmel for large-scale querying, and R for statistical computing.

2. **Hadoop is set to replace traditional data warehouses.** Traditional enterprise data warehouses (EDWs) are expensive and ill-equipped for solving big data problems. Typical EDWs are designed to house an enterprise’s core data. Data flows from operational systems (such as ERP, financial, and HR) into EDWs that, in turn, provide consistent and structured data for reporting and business uses.

   Hadoop on the other hand, is an open-source framework built around a high-volume distributed architecture that runs well on low-cost commodity hardware. This architecture and its associated languages and tools allow for solving complex analytics problems relatively quickly. Hadoop is the ideal platform for analyzing ERP data in conjunction with disparate data sources.

   For example, a company could combine ERP information with sensor data, weather information, and transportation rates—data sources with different structures or no structure at all, and which could be voluminous—to optimize the most cost-effective time and place to ship perishable products. All the loading, structuring, analysis, and reporting can be done directly and rapidly with Hadoop without the need to ever move data into or out of the EDW.

   If Hadoop is doing all of this, then perhaps the logical next step is reexamining the role of EDWs. Some companies are already augmenting EDWs with Hadoop, offloading traditional ETL (extract, transform, load) functions and making use of the distributed processing capability. Other forward-thinking organizations are taking that a step further, and experimenting with using Hadoop to replace EDWs altogether. EMC and Teradata, among other major vendors, have already made strong moves into the Hadoop space, so we expect it won’t be long before many more start moving in this direction.

3. **Big data and analytics will be increasingly embedded directly into devices.** Hadoop’s flexibility and open architecture make it a natural fit for embedding directly into devices—from medical equipment to drones to sensors. Future “smart” devices will be able to process data and conduct advanced analytics at the source, similar to the way in which mobile phones have transformed from simple handsets to mini computers. Embedding Hadoop in devices will accelerate and streamline the collection and processing of high-volume data, such as video and audio.

4. **The software giants are playing catch-up in big data—leaving a market vacuum waiting to be filled.** The long-viewed unequivocal leaders in analytics, the world’s leading software providers have been laggards in adopting solid big data strategies. Even database leaders

such as Microsoft and Oracle and visualization and business intelligence companies such as Tableau, TIBCO, and QlikView have been challenged to bring products to market fast enough to keep pace with the changes.

With so much big data innovation occurring in the open source space, commercial software companies need to ensure that they do not become reactive. Traditional analytics software companies will still play a role, of course, especially in large enterprises, but the nature of their leadership will take on a new meaning, and licensing costs could decrease dramatically as customers discover that they have many more low-cost options. Universities, longtime training grounds for specialized skills in data and analytics software (especially SAS and SPSS), are embracing open-source tools, which are freely available and generally a much better fit for the academic environment. If this trend holds up, it would radically change the nature of entry-level skills in the market, skills that would more closely match employers’ increasing demand for open-source experts.

Five years ago, it would have been unthinkable to question the validity of and need for traditional EDW architectures, or the ability of traditional analytics software vendors to innovate.

The challenge for traditional analytics software vendors to innovate—combined with the need to make analytics accessible to the masses—is creating a vacuum in the marketplace. If this vacuum is not addressed, IT organizations will end up procuring and supporting a vast array of complex tools ill-suited for enterprise environments.

5. Can you hear me now? Siri is just the beginning. Science fiction has long captivated audiences with sketches of the ideal computing interface—think Captain Kirk demanding answers from the USS Enterprise’s computer in Star Trek, Dave and Hal in 2001: A Space Odyssey, and Jarvis in Iron Man. Similarly for big data and real-time analytics, the tipping point—both in a corporate sense and for society in general—will be the day that a non-technical, non-mathematical, non-engineering user can ask questions of the data without typing or using a mouse. This day is not far off; Siri, the iPhone’s interactive personal assistant, was the first commercial success in this area, and its arrival in big data is fast approaching. Gartner expects that by 2016, 70 percent of analytics vendors will incorporate voice recognition into their software.²

These trends are truly disruptive and call into question the basic tenants of IT strategy and planning. A scant five years ago, it would have been unthinkable to question the validity of and need for traditional EDW architectures, or the ability of traditional analytics software vendors to innovate. Forward-looking CIOs are revisiting their big data plans with an eye toward this rapidly changing landscape.

Big Data Architecture

Big data’s full effect on business may not be completely realized yet, but it’s already a major disruptor for IT organizations. From the need to build new infrastructure to hiring people with new and different skill sets, big data is already driving a long-term transformation of the IT function.

Data is moving into, out of, and across organizations so fast that traditional data-management techniques are inadequate. As a result, data must be viewed holistically across the organization using an information management framework (see figure 2). Sensors that generate millions of terabytes of data per day are obliterating master data-management and warehousing schemes. Public external unstructured data sources, such as social media applications, are forcing IT groups to work beyond the firewall more often. Advanced analytics work with niche third-party players requires extensive data sharing, often via cloud-based storage. At the root of these changes is a new technology architecture context that requires support for new big data and analytics paradigms.

Figure 2
Big Data should be viewed in the context of an overall information management strategy

Information management framework and benefits

Information management

<table>
<thead>
<tr>
<th>Key components of information management</th>
<th>Information architecture</th>
<th>Information governance</th>
<th>Data management</th>
<th>Tools and technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information lifecycle</td>
<td>Acquire</td>
<td>Store</td>
<td>Process</td>
<td>Use</td>
</tr>
<tr>
<td></td>
<td>Vision</td>
<td>Principles and standards</td>
<td>Information domains and relationships</td>
<td></td>
</tr>
<tr>
<td>Information governance</td>
<td>Policies</td>
<td>Organization structure</td>
<td>Information stewardship</td>
<td>Performance metrics</td>
</tr>
<tr>
<td>Data management</td>
<td>Data model</td>
<td>Master data management</td>
<td>Data quality management</td>
<td>Data security</td>
</tr>
<tr>
<td>Tools and technologies</td>
<td>Application architecture</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For IT organizations to manage all of these competing priorities, they need new technologies, skills, and processes. More importantly, they must start designing and deploying full-service platforms based on big data analytics. Arguably, there is no off-the-shelf, enterprise-ready platform that can solve all of the necessary, new, and endlessly changing business requirements. For now, a variety of different options exists for each component, but a fully integrated suite is a long way off (see figure 3 on page 5).
To help sift through the big data landscape, we’ve broken the core architectural components into eight distinct functional groups. These groups can serve as a foundation for conceptualizing what will go into a comprehensive platform to enable business to adopt and make effective use of big data analytics (see figure 4 on page 6).

Accordingly, vendors and toolsets in the marketplace are closely aligned with the functional groups, which are each characterized by heightened demand, lower pricing, new players, and high acquisition and partnership activity. Both traditional and start-up vendors have launched new products, bringing dynamic evolution and consolidation of products and suites within each group. Although established players have a built-in advantage because so many firms already use their products, no one player dominates all groups. Faced with this array of choices, IT organizations often have a tendency to start their big data journeys with an assessment of the players and their products. However, evaluating as-is capabilities first allows IT to identify key gaps and prioritize them to inform its evaluation efforts.

**Analytics appstore.** Forward-looking vendors are creating pre-baked analytics apps designed to solve specific business-use cases in a particular industry. These apps (such as next best offer, fraud detection, inventory optimization, and predictive maintenance) enable companies to quickly draw upon a wealth of best practices to allow non-technical users to easily execute complex analyses. Right now, these apps are tied into individual platforms, but over time as they become more sophisticated, they likely will become more portable. Given their accessibility, these apps make an excellent starting point and increase the likelihood of success within a
broader big data strategy. A number of emerging companies are starting to build these platforms, and undoubtedly bigger companies will follow suit (see figure 5 on page 7).

**Decision support.** Decision support technologies bring together business intelligence platforms with analytics and big data technologies. This growing area will remain a top priority for CIOs for years to come, as it enables more effective and timely support for organizational decision making. The key to these products is their ability to model data and run advanced analytics techniques (such as simulation, optimization, and sentiment), and to quickly interpret and synthesize outcomes and recommendations. Solutions in this space are rapidly improving, placing real-time fact-based information in the hands of more decision makers faster than ever before. As more users across organizations gain access to these tools, identifying, procuring, and supporting the tools will remain top CIO priorities. Category leaders include traditional players—Microsoft, SAS, and IBM—as well as best-of-breed providers such as AIMMS (optimization), Arena (simulation), and Altrex (predictive modeling).

**Reporting and visualization.** As big data spreads, the number of tools that report and visualize data is, not surprisingly, expanding fast. These tools are important because they are the bridge
between complex analyses and understanding by decision makers. Rapid innovation in advanced visualization techniques is playing an integral role in how data is presented today in just about every walk of life.

Open-source products and new entrants are forcing existing business intelligence vendors to either improve or become marginalized. Tableau and TIBCO have emerged as potential leaders in the newly created domain of visualization software. Striking examples of data visualization range from the Bloom Diagram, an IBM Watson Research Center project that visualizes the contribution of individuals to open-source projects, to visualizing citation data from Thomson Reuters’ Journal Citation Reports. Traditional heavyweights Microsoft and SAS, which recently fell behind in this sector, have begun reemerging as players. In addition, increased acquisition activity by larger players in this space will close some of the current gaps in offerings.

**Analytics services.** Analytics services provide easy access to advanced analytics capabilities. Offered either as a service or as easy-to-deploy, out-of-the-box vertical solutions, this group offers access to best-of-breed analytics tools. For example, several companies offer sentiment analysis—language and text processing used to extract subjective information—which requires great skills in linguistics and natural language processing. Since few companies have these skills available in-house, the market for on-demand sentiment engines and services has emerged. Other advanced analytics techniques, such as predictive analytics, are now being offered as on-demand services and many others will likely follow. Splunk is a leader in this area, with IBM and Oracle strong contenders.

**Parallel distributed processing and storage.** This is perhaps the most disruptive to IT architecture and landscape, marked by rapid advancement in solutions built around Hadoop and distributed processing. The low cost of processing and storage combined with advanced technology developments are enabling companies to employ distributed grid computing, the foundation of a big data platform. Increasingly, vendors are manufacturing plug-and-play...
appliances that streamline acquisition and deployment—a step that could become the norm for all data processing. This step speaks to the significant consolidation in this market by traditional EDW and hardware vendors, including IBM’s acquisition of Netezza, EMC’s acquisition of Greenplum, and HP’s acquisition of Vertica (see figure 6).

Figure 6
Recent big data consolidation

Loosely structured storage. IT organizations are built to manage highly structured data. Database administrators and business analysts work to ensure that data is clean and consistent, and can serve as the single source of truth. But, with the explosion of data, IT is being forced to accommodate various types of data in semi-structured or unstructured formats. As Hadoop emerges as the de facto standard, IT organizations will need a different set of skills if they are to manage both the volume of data and the need to make the data accessible to the business. Architecturally, this segment is transitioning rapidly from its open-source origins to a mature market, with Cloudera and Teradata leading the charge. Other key notable vendors in this market segment include Hortonworks, Oracle, and MapR. Buoyed by venture capital investments, several small players will likely continue to play a role in this area.

Highly structured storage. Highly structured storage is for large volumes of hierarchical data maintained in structured formats. Oracle is the leader, ahead of rivals IBM and Microsoft. Contrary to popular belief, this segment will continue to grow due to the complementary role it plays because its growth is fueled by ever-increasing volumes of big data. Many major big data platforms on the market today combine Hadoop and other unstructured data processing systems with traditional structured relational database systems and data warehouses. After developing insights from unstructured data, typically in real time, the data is then structured and loaded into traditional structured storage for normal back-end analysis. It’s still too early
to predict if, when, and how technologies in this space will be replaced by, or will converge into, Hadoop-based or other new data management paradigms. Nevertheless, it’s clear they will continue to play an important role in the foreseeable future.

**Integration.** This group of products brings together data from diverse sources into unified systems. With more in-database analytics and technologies, the movement of data throughout an organization will change significantly. Traditional ETL (extract, transform, load) functions will be replaced by new ELT (extract, load, transform) techniques that move transformation and data analysis into the database layer, thus eliminating and simplifying data-staging efforts. Newer data-related technologies and business process management software will provide easier access to, and manipulation of, multimedia data formats. IBM, Informatica, Oracle, and TIBCO are leading in the integration domain.

---

**From the need to build new infrastructure to hiring people with new and different skill sets, big data is already driving a long-term transformation of the IT function.**

---

**The Big Data Talent Challenge**

Forward-looking IT organizations should start to assess their current strategy, infrastructure (hardware and software), and organization (talent and structure). With the emergence of cloud technologies and DIY capabilities, if IT isn’t prepared, it will be left behind.

For decades, IT organizations have employed armies of system and database administrators (DBAs) to manage different types of relational databases with highly structured data. ERP, HR, SRM, e-commerce, and just about every other IT-supported tool runs on some form of a structured database. Architecting and supporting these tools has required very specific job skills and the need to limit who can access the data as well as when and how they can access it.

Handling and using big data is the complete opposite: The goal is to give many people access to as much data as possible to improve the accuracy and breadth of analyses. Leading companies make large portions of data available with the corresponding infrastructure, so a large number of users are enabled, encouraged, or otherwise empowered to make use of the data. This data availability creates a natural tension with IT organizations that are charged with securing the enterprise’s data. On one hand, an excessively restrictive policy will force IT to become an impediment to innovation; policies that are too loose, on the other hand, will leave the data vulnerable and unmanageable.

Furthermore, conducting advanced analytics or data science efforts requires working with both vast amounts of data and a wide variety of formats that often have little or no structure. Data scientists apply a combination of programming, decision science, business expertise, analytic techniques, and creativity to generate insights. Often, they reside in their own groups—outside of IT, assigned to business units. The emergence of data scientists will
expands traditional IT roles and the skills of database administrators, programmers, and system administrators, and will lead them to be located more closely with the business (see figure 7).

Today, the pressing concern is finding these data science resources. There is a quantifiable dearth of people with these skills. University programs are working quickly to address this need, but demand far outstrips supply. According to The Wall Street Journal, nearly 80 percent of data scientist jobs in the United States go unfilled. Furthermore, the small cadre of people with these skills are often scooped up by industry heavyweights (such as Google, Facebook, Twitter, and LinkedIn) that can offer big salaries, stock options, and lavish work spaces. This leaves precious few recruits for non-tech industries. It is important to recognize that this problem is not going to be solved overnight, and there is no magic solution.

Enterprises seeking to fill these positions will need to look at cultivating existing internal employees who possess the skills and the desire to evolve into these positions after they acquire targeted training. Options for this training include data science certificate programs through university extension programs and evolving masters programs from leading educational institutions. Some enterprises may decide to recruit non-traditional resources such as PhD students from the hard sciences, social sciences, or even the humanities. These students often possess strong critical thinking, scientific methodology, or statistics backgrounds. Pairing those who have these skills with people from business can jumpstart efforts to build these teams. Naturally, companies can always fall back on hired contractors but they come

---


with a steeper price tag and the usual pitfalls of using these types of resources. By cultivating their own talent, enterprises can build data science teams with a solid foundation of dedicated employees interested in working there for the long haul.

Preparing for the Revolution

We view big data and its associated analytics as a transformational disruptor as significant as the Industrial Revolution. Rapid innovation, radical increases in sources and types of data, and the public and co-creation of analyses will enable new forms of decision support and will revolutionize how data is understood and applied.

How organizations conceive of and design big data and analytics systems will change, forcing them to consider players with products in functional categories that didn’t exist just a few years ago. The way organizations staff and source the new IT talent required to drive this revolution will also change considerably, demanding an even greater connection between the new data scientists, analytics specialists, and the business users they support.

Big data is fast becoming part of the broader analytics ecosystem, where the distinction between big and traditional data is evaporating. Soon big data will be an integral part of successful companies’ overall data collection, processing, and analytics-generation strategies. Embedded big data and analytics will open new realms of hyper-connected, “smart” devices, enabling new, dynamic capabilities.

Savvy IT leaders will recognize the coming changes and work to capitalize on the opportunities.

Authors

Christian Hagen, partner, Chicago
christian.hagen@atkearney.com

Khalid Khan, partner, Chicago

Hugo Evans, director, San Francisco
hugo.evans@atkearney.com

Bharath Thota, director, Chicago
bharath.thota@atkearney.com

Dan Wall, principal, New York
dan.wall@atkearney.com

Anand Seshadri, consultant, Chicago
anand.seshadri@atkearney.com
A.T. Kearney is a leading global management consulting firm with offices in more than 40 countries. Since 1926, we have been trusted advisors to the world’s foremost organizations. A.T. Kearney is a partner-owned firm, committed to helping clients achieve immediate impact and growing advantage on their most mission-critical issues. For more information, visit www.atkearney.com.

<table>
<thead>
<tr>
<th>Americas</th>
<th>Atlanta</th>
<th>Detroit</th>
<th>San Francisco</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bogotá</td>
<td>Houston</td>
<td>São Paulo</td>
</tr>
<tr>
<td></td>
<td>Calgary</td>
<td>Mexico City</td>
<td>Toronto</td>
</tr>
<tr>
<td></td>
<td>Chicago</td>
<td>New York</td>
<td>Washington, D.C.</td>
</tr>
<tr>
<td></td>
<td>Dallas</td>
<td>Palo Alto</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Asia Pacific</th>
<th>Bangkok</th>
<th>Melbourne</th>
<th>Singapore</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beijing</td>
<td>Mumbai</td>
<td>Sydney</td>
</tr>
<tr>
<td></td>
<td>Hong Kong</td>
<td>New Delhi</td>
<td>Tokyo</td>
</tr>
<tr>
<td></td>
<td>Jakarta</td>
<td>Seoul</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kuala Lumpur</td>
<td>Shanghai</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Europe</th>
<th>Amsterdam</th>
<th>Istanbul</th>
<th>Oslo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Berlin</td>
<td>Kiev</td>
<td>Paris</td>
</tr>
<tr>
<td></td>
<td>Brussels</td>
<td>Lisbon</td>
<td>Prague</td>
</tr>
<tr>
<td></td>
<td>Bucharest</td>
<td>Ljubljana</td>
<td>Rome</td>
</tr>
<tr>
<td></td>
<td>Budapest</td>
<td>London</td>
<td>Stockholm</td>
</tr>
<tr>
<td></td>
<td>Copenhagen</td>
<td>Madrid</td>
<td>Stuttgart</td>
</tr>
<tr>
<td></td>
<td>Düsseldorf</td>
<td>Milan</td>
<td>Vienna</td>
</tr>
<tr>
<td></td>
<td>Frankfurt</td>
<td>Moscow</td>
<td>Warsaw</td>
</tr>
<tr>
<td></td>
<td>Helsinki</td>
<td>Munich</td>
<td>Zurich</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Middle East and Africa</th>
<th>Abu Dhabi</th>
<th>Dubai</th>
<th>Manama</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Doha</td>
<td>Johannesburg</td>
<td>Riyadh</td>
</tr>
</tbody>
</table>

For more information, permission to reprint or translate this work, and all other correspondence, please email: insight@atkearney.com.

The signature of our namesake and founder, Andrew Thomas Kearney, on the cover of this document represents our pledge to live the values he instilled in our firm and uphold his commitment to ensuring “essential rightness” in all that we do.