OPEN DATA CENTER ALLIANCE℠: BIG DATA CONSUMER GUIDE

TABLE OF CONTENTS

LEGAL NOTICE

OPEN DATA CENTER ALLIANCE℠: BIG DATA CONSUMER GUIDE

EXECUTIVE SUMMARY

According to IDC, unstructured data accounts for more than 90 percent of the data in today's organizations. The McKinsey Global Institute says that 15 out of 17 U.S. business sectors have more data stored per company than the U.S. Library of Congress, while a 2010 Forrester report states that 62 percent of organizations use between 100 terabytes (TB) and 9 petabytes (PB) of data. IDC predicts that the volume of data will double every 18 months. Volume is only part of the picture, however. Data is being generated at a staggering velocity and from a growing variety of disparate sources. Twitter receives 200 million tweets per day—that's 46 megabytes per second; Facebook collects an average of 15 TB every day. The sheer volume and complexity of data is overwhelming traditional database software and is demanding a new approach to data analysis that can deal with Big Data. In 2012, *The Economist* stated that Big Data, data analytics, and smart systems are among the three highest-impact technologies of the next decade.

Although it presents significant challenges, Big Data also presents significant opportunities to those enterprises that can take advantage of it by analyzing Big Data to drive organizational efficiency and competitive advantage. According to Gartner, by 2015 only 10 to 15 percent of businesses will fully take advantage of Big Data, and they'll outperform their unprepared competitors by 20 percent in financial metrics.

In this document, the Open Data Center Alliance (ODCA) Data Services (DS) Working Group provides an introduction to what Big Data is and how it differs from traditional data, surveys existing Big Data technology solutions, describes potential Big Data use cases, and identifies the challenges and considerations (both technical and organizational) associated with leveraging Big Data.

By providing an introduction to the Big Data landscape, this paper lays the foundation for future work by the ODCA DS Working Group, which will include bringing solution providers and Big Data consumers (enterprises) together to help drive the creation of Big Data technologies that are open and standards-based, with a greater degree of interoperability and cost effectiveness for large enterprises.

"There is opportunity for us to come together to start solving the issues that this degree of change brings about," said Matt Estes, Chairman of the ODCA DS Working Group.

This document serves as the catalyst for further defining the open specifications, formal or de facto standards, and common intellectual property-free (IP-free) solution designs that will help accelerate adoption of Big Data technologies and create an open, interoperable, and thriving market of Big Data solutions.

INTRODUCTION

We are drowning in data—structured and unstructured, human-generated, and machine-generated. Digital data is being created at almost unimaginable rates, and the floodgates continue to open wider. We are creating oceans of data as businesses, government agencies, and individuals interact across public and private networks around the globe. Over the next few years, another billion users will be connecting to the Internet with more and smarter devices, driving online transactions—and the data they generate—to ever-higher levels. The flow of digital information within and between businesses is also growing rapidly. Many companies are integrating sensors into their products and processes, creating new sources of high-volume data flow.

Cloud computing deployment models are reducing the time it takes to deploy products to market and decreasing the cost required to provide services to consumers over the Internet. As such, this is increasing the degree to which businesses are pursuing Internet commerce models—in turn, adding to the explosion of data.

Nevertheless, we are still starved for knowledge and intelligence. In many cases, the ability to collect data outpaces the ability to derive meaning and actions out of it.

Some of the world's most successful companies owe their success, in part, to the innovative strategies they have developed for accessing, managing, and using select portions of that data to identify opportunities, make better business decisions more quickly—sometimes in near-real-time—and deliver personalized customer experiences. According to Gartner, by 2015 only 10 to 15 percent of businesses will fully take advantage of Big Data, and they'll outperform their unprepared competitors by 20 percent in financial metrics.

Objective

To date, the ODCA has focused the majority of its efforts on Cloud Computing. This focus has been due to the profound impact that Cloud Computing has had on the marketplace. Big Data is shaping up to be another such profound change. As such, the ODCA sees this area as a logical next step for its members to take action on. Further, there is an intersection between Big Data, analytics, and Cloud Computing deployment that the ODCA plans to address in future work. The scope of this paper is focused around Big Data only. This has been done to lay a foundation upon which future work, including the intersection of Big Data and Cloud Computing, can be explored.

The objective of this document is to promote Big Data and determine how the ODCA DS Working Group can provide concrete recommendations and insight that can benefit both Big Data solution providers and enterprise consumers of Big Data. By providing an introduction to the Big Data landscape and illustrating the various use cases for Big Data, this paper lays the foundation for a set of future work by the ODCA DS Working Group, which will include bringing solution providers and Big Data consumers (enterprises) together to help drive the creation of Big Data technologies that are open and standards-based, with a greater degree of interoperability and cost effectiveness for large enterprises.

Enterprises can use the information in this document to better understand Big Data and achieve a balance between existing investments and new ones that best address the exponentially increasing volume, velocity, and variety of enterprise data.

BIG DATA 101

Before enterprises can determine why and how to use Big Data, and before solution providers can begin to devise technology that meet enterprise needs, it is important to understand what Big Data is, how it came about, and why it is important.

Defining Big Data

Big Data refers to massive amounts of data, the size and variety of which are beyond the processing capabilities of traditional data management tools to capture, manage, and analyze in a timely manner. Big Data comes from everywhere. Common sources include:

- Machine-generated data from sensors, devices, RFID, machine logs, cell phone GPS signals, and more
- Digital media proliferation (both online and off-line) and social media sites
- Sub-transactional records of online transactions

According to IDC, unstructured data accounts for more than 90 percent of the data in today's organizations, stored in email messages, documents, notes fields, and Web content. According to Gartner, unstructured data doubles every three months and seven million Web pages are added every day. Big Data also includes traditional structured data that exists in massive quantities. Wal-Mart is a good example: More than 1 million customer transactions occur every hour, generating more than 2.5 PB of data—equivalent to 167 times the information contained in all the books in the Library of Congress.

Big Data has inspired new and complementary approaches to storing, querying, and analyzing both structured and unstructured data. NoSQL databases are useful for working with huge quantities of data—structured or unstructured—when what really matters is the ability to store and retrieve vast quantities of data, not the ability to examine the relationships between the data elements. NewSQL is a new category of relational databases that improves transaction speed and scalability. MapReduce is a newly developed programming model for processing large data sets.

All these new tools and approaches embody a common definition of Big Data as a combination of three Vs: Volume, Velocity, and Variety.

- **Volume.** As the name Big Data suggests, its volume can take up terabytes and petabytes of storage space. It has arisen as a result of an increasing enterprise demand to use and analyze more types of structured and unstructured data that do not fit into existing operational and analytic business systems. Data is growing at an exponential rate, so much that 90 percent of the data in the world today has been created in the last two years alone.
- **Velocity.** Increasingly, enterprises need answers not next week or next month, but right now. Nightly batch loading is poorly suited for e-commerce, multimedia content delivery, ad targeting, and other real-time applications. This puts pressure on accelerating data loading at the same time that data volumes are skyrocketing. Data streaming, complex event processing, and related technologies, once mostly prevalent in financial services and government, are now emerging as enterprise data architecture requirements in multiple industries. Likewise, as more enterprises engage in social media and the Web, responding in real-time or in near real-time becomes significantly more necessary.
- **Variety.** Variety relates to the complexity of data types and data sources. Also, much of today's data is unstructured or semi-unstructured. This means that it doesn't fit into neat rows and columns of the traditional relational database management systems (DBMS).

Note: Other "Vs" can also be used to discuss Big Data, such as Variability and Value; however, the three described above are the ones most commonly discussed in the industry.

Figure 1 shows how data storage has grown since 1996.

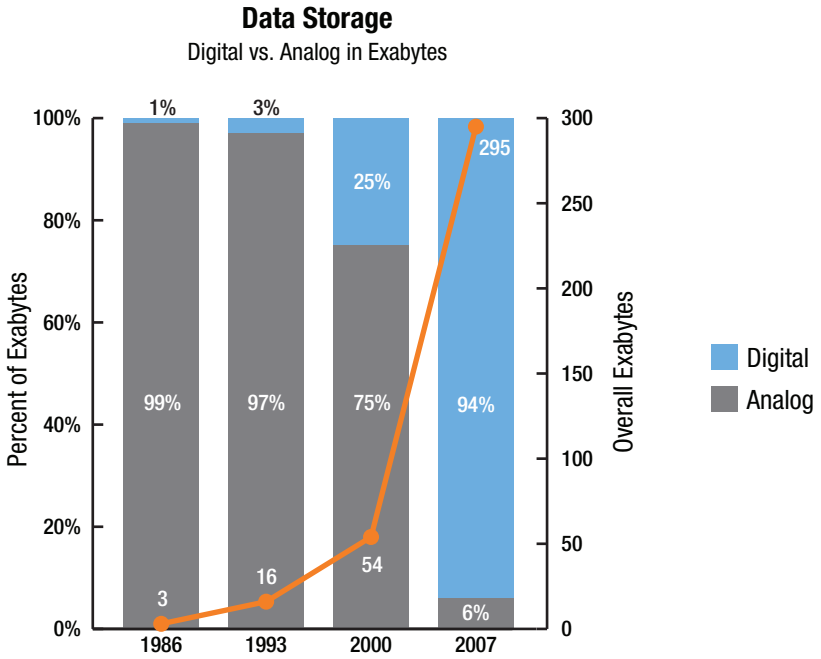

FIGURE 1: DATA STORAGE HAS GROWN SIGNIFICANTLY, SHIFTING MARKEDLY FROM ANALOG TO DIGITAL AFTER 2000.

Source: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," Science, 2011. Numbers may not sum due to rounding.

Big Data Evolution

Big Data has emerged because we are living in a society that makes increasing use of data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide and between 1 billion and 2 billion people accessing the Internet. Basically, there are more people interacting with each other, and with information, than ever before. The actions of each user result in a cascade of subsequent actions, each of which is now logged, creating additional data. This is sometimes called "data exhaust."

Although Big Data has some new—and fairly disruptive—characteristics, it is simply the next step in a long evolution of enterprise reliance on data. In the early 1980s RDBMs were fledgling systems, and then grew into billion-dollar enterprises such as Oracle and SAP. With the growth of the Internet, it wasn't long before enterprises turned to online transaction processing (OLTP) databases, then to dimensional data warehouses (DWs) to meet their data storage and analytic needs.

Today we stand at the threshold of yet another transformation, where those who "get it" will continue to thrive and grow, and those who remain lodged in outdated technologies will fall by the wayside. What used to be considered a storage problem is now a strategic asset.

Why Big Data Is Important

Tackling Big Data using traditional approaches to data management and analysis may not always be a viable option. For example, a company may decide the return on investment (ROI) associated with scaling up its RDBMS is not sufficient and may decide a different approach will be more cost effective. Or, a company may be sensitive to latency issues and can't afford to wait three days for data to be processed. The bottom line is that enterprises cannot afford to ignore Big Data, as it contains compelling and powerful information about evolving customer needs, product pain points, and recurring service issues.

Using new technologies that go beyond RDBMS and that enable new types of data aggregation and analysis, enterprises can obtain deeper and richer insights, thereby accelerating decision making and the pace of innovation, as well as increasing business value through significant cost savings and increased revenue.

Here are just two examples where the velocity and volume of incoming data is simply too large to fit into a traditional RDBMS. Prior to Big Data, these types of scenarios were solved using sampling and aggregation.

- **Network Operations.** Mobile, wireline, and cable service providers need a holistic view of network, applications, devices, and subscriber data to gain insights for improved network planning and optimization that affect the company's bottom line. Big Data analytics can help answer questions such as "What is the traffic usage in every data plan?" and "How do we create the right pricing plan for our customers?" Enterprises can use Big Data to mine details about network, device, subscribers, and applications to identify the most popular devices or applications in a particular location, create traffic profiles, and identify the top subscriber usage patterns.
- **Utility Usage.** Imagine fusing data from a home energy management system and external utility data sources. This would make it possible to perform usage pattern and what-if analyses that could help detect possible degradation of appliances over time, correlate weather data with energy spending to understand comfort and cost, and provide a better understanding of what drives energy consumption and how those drivers change over time.

Although many of the use cases that can benefit from Big Data have existed for quite some time, they now are characterized by the ability to leverage new data obtained from unstructured data types. Previously, unstructured data was either ignored or, at best, used inefficiently. By combining the new data sources with traditional sources of data, enterprises can achieve new and more valuable insights at a more granular level. Prior to Big Data, most insights were generalized for a group or segment.

Cost is also a big differentiator for Big Data use cases. Historically, companies needed to spend a significant amount of money on hardware, software, and custom application development to obtain similar results. With the power of today's commodity servers and open-source solutions, companies can now implement these use cases at a fraction of the cost, and with far less effort. Before open-source solutions such as Apache™ Hadoop® arrived on the scene, parallel programing was very difficult, especially in situations where SQL queries were not sufficient to express the necessary analysis. In these cases, custom programing was required, which is expensive. Hadoop now enables cost-effective parallel processing.

However, many businesses currently don't understand the importance of Big Data or how to go about beginning to leverage it. At the Gartner Catalyst 2012 Conference in San Diego, it was stated that "An understanding of when to use Big Data is lacking right now." The remainder of this paper identifies a variety of potential use cases, describes the existing technologies, and outlines some planning considerations businesses should keep in mind as they form their Big Data strategies. As shown in Figure 2, the right Big Data solution for a particular enterprise depends on choosing the right use cases, tools, and staff, as well as making high-level decisions about investment and infrastructure.

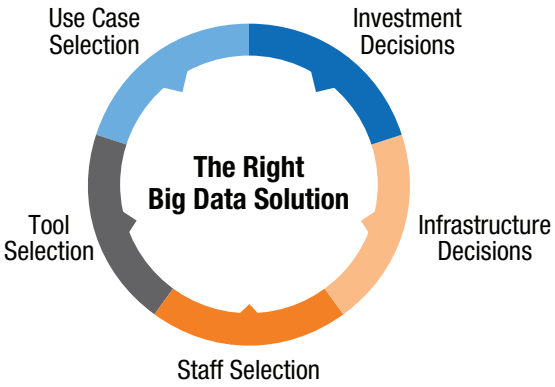

FIGURE 2: CHOOSING THE RIGHT BIG DATA SOLUTION INVOLVES MANY DECISION POINTS.

BIG DATA USE CASES

One of the advantages of Big Data is that it can bring value to almost any industry sector, enabling new insights and enhancing decision support for a wide variety of projects. The following are examples of the industry sectors that can benefit from Big Data.

- Social media and search engines
- Insurance, banking, and finance
- Telecommunication and Internet service providers
- Mobile service providers
- Retail and point-of-sale analytics
- Manufacturing optimization
- Utility and energy
- Healthcare
- IT operations
- Research and development
- Transportation
- Cloud computing
- Marketing

Consumers can benefit by understanding the use cases that span industries or that can be used for industry-specific purposes to extract actionable value from analyzed data sets. Conversely, solution providers can design better solutions if they understand enterprise needs. This section provides some examples of use cases that span multiple industry sectors. It is not intended to be an exhaustive list; the application of Big Data is an emerging field, and new use cases arise on a regular basis. Appendix A: Use Case Details contains a table that provides more guidance on which use cases are relevant to particular industry sectors.

Cross-industry Examples

Common cross-industry uses for Big Data include, but are not limited to, compute-intensive data science and cost-effective storage. A particular enterprise may use Big Data in any of the following use cases:

- **Data accumulation and archiving.** Big Data technologies are being used to preserve and archive data. The redundant nature of Hadoop, coupled with the fact that it is open source and an easy way to access file systems, has compelled several organizations to use Hadoop as an archival solution. In some ways, with commodity servers bringing the cost of storage down, Big Data has become the "new tape backup." Archiving huge amounts of data can help enterprises meet regulatory compliance regulations, as well as preserve data, even if the enterprise isn't quite sure what to do with that data yet.
- **Research and development.** Many companies, such as pharmaceutical manufacturers, use Hadoop to comb through enormous volumes of text-based research and other historical data to assist in the development of new products.
- **Predictive analysis.** Analysts have used advanced algorithms for correlations and probability calculations against current and historical data to predict markets as standard practice. The large amounts of historical market data and the speed at which new data needs to be evaluated make this an excellent application of Big Data technology. The ability to perform these calculations faster and on commodity hardware makes Big Data a reliable substitute for the relatively slow and expensive legacy approach.
- **Network optimization to prevent failure.** Big Data technologies are used to analyze networks of any type. Networks, such as the transportation network, the communications network, the police protection network and even a local office network all can benefit from better analysis. Consider a local area network. With these new technologies, massive amounts of data are collected from servers, network devices, and other IT hardware. Patterns leading up to network issues can be identified so that bottlenecks and other issues can be prevented before they have an adverse effect on productivity.
- **Real-time decision making and scenario tuning.** Increasingly, OEMs are adding sensors to platforms so they can be used for payment, authorization, and identity. These include sensors for many other factors beyond traditional location and connectivity, such as humidity, temperature, and ambient light. Sensors can generate a large amount of data, and companies can use this data to adapt an environment, product, or service to more accurately meet the needs of consumers (or providers) in real-time.
- **System usage.** Monitoring system usage—whether the resource is servers, memory, storage, network, or common services such as Lightweight Directory Access Protocol (LDAP)—generates large amounts of data that can illuminate trends and enable enterprises to plan better. Using this data, operations staff can monitor how the subsystems are behaving, and establish rules and policies to respond to usage thresholds. Further, operations staff can optimize system utilization by tracking spikes and troughs of utilization, helping to prevent both over-allocation (wasteful) and under-allocation (potentially catastrophic).
- **Root cause analysis.** In cases where there is a large system failure, the root cause could be unclear. Often a cascade of events occurs, and the history is contained in the full aggregate of log files and monitoring data collected across the data center. Therefore, finding the root cause may involve analyzing very large data sets, searching for a specific data point in a massive collection of data, and correlating data from disparate sources.
- **Sentiment analysis and customer churn management.** Companies can use voice analytics and text analytics, based on voice modulation and keyword analysis, to better understand customer sentiments. Getting timely actionable insights about customer sentiment can enable organizations to improve customer satisfaction in a timely manner and manage churn appropriately.
- **Data preservation.** Finding the right information and discovering trends
- **Data movement.** Extract, transform, and load (ETL) offload
- **Marketing funnel analysis (conversion analysis)**
- **Information security.** Malware detection and fraud detection
- **Recommendation engine.** App store and eCommerce
- **A/B testing of web pages**
- **Cohort analysis and social graphs**

BIG DATA TECHNOLOGIES

Big data platforms pursue a simple approach of divide and conquer. They achieve massive parallelism by distributing the data and the processing effort, as well as data storage, between many worker nodes and then aggregating the results to return the answer. Big Data technologies offer massive parallelism, scalability, availability, and "smart" storage that enables the execution of code and algorithms where the data resides. Although this sounds conceptually simple, the staggering diversity of data means that one application cannot "do it all."

When surveying the landscape of available Big Data technologies, it is helpful to understand some of the terms commonly used when discussing Big Data. An alphabetical list of common terms is provided in Table 1.

TABLE 1. COMMON BIG DATA TERMINOLOGY.

Term	Definition
Apache™ Hadoop®	An open-source storage and processing framework based on MapReduce, using a distributed file system.
BigTable	A type of NoSQL database that is based on Google's BigTable paper from 2006. In essence it is a highly scalable, sparse, distributed, persistent multidimensional sorted map.
distributed file system	A file system that allows access to files from multiple hosts sharing a computer network. Hadoop and other Big Data technologies use this approach to implement parallel processing and improve availability and performance. Distributed file systems often imply replication of data and fault tolerance.
document store	A type of NoSQL database that stores entire documents.
graph database	A type of NoSQL database that uses graph structures with nodes, edges, and properties to represent and store data.
key-value store	A type of NoSQL storage that enables storage of arbitrary data (a value) using a unique identifier (key).
machine learning	A branch of artificial intelligence concerned with the development of algorithms that take as input empirical data, such as from sensors or databases. The algorithm is designed to identify complex relationships thought to be features of the underlying mechanism that generated the data and employ these identified patterns to make predictions based on new data.
MapReduce	A programming paradigm that enables parallel processing.
massively parallel processing (MPP)	The coordination of a large number of processors (or separate computers) to perform computations, where a processor or group of processors works on different parts of the program.
NewSQL	A category of databases that uses new approaches to modify the underlying architecture of SQL databases so that they can scale similar to many NoSQL technologies.
NoSQL (not only SQL)	A broad class of non-relational, non-SQL databases that often does not offer ACID guarantees. This class of databases encompasses document store, key-value store, BigTable, and graph databases. This class of databases is useful for working with huge quantities of data—structured or unstructured—when the ability to store and retrieve vast quantities of data is more important than the ability to examine the relationships between the data elements.
polystructured data	Data that is in various formats, and those formats may change over time.
streaming analytics	Analysis of data as it is generated–data in motion. To be compared to the analysis of data after persistence–data at rest.
structured data	Data that resides in fixed fields within a record or file such as a relational database or a spreadsheet.
unstructured data	Information that either does not have a pre-defined data model or does not fit well into predefined attributes or row/column formats.

Big Data Ecosystem

The Big Data ecosystem consists of hundreds of solution providers that concentrate on one or more focus areas. In general, solution providers can be categorized as developing solutions in the areas of storage, integration, and analytics. However, some cross categories, and some don't fit easily into any one category. In addition, it seems a new solution provider or new product arrives in the marketplace every week. Therefore, any snapshot of the Big Data ecosystem is bound to be out-of-date within a few months after it is published.

There are a couple of ways of classifying the Big Data ecosystem. One approach is to look at the data and how it is processed. Another is to look at who is supplying the solution—open source or third party.

Data Structure and Latency

The emerging Big Data capabilities enable the value of data to be maximized at an extreme scale, in an affordable manner. Figure 3 shows a graphical view of Big Data capability options driven by the data structure and whether the data is processed in batch or in real-time.

"Real-time" can have different meanings, depending on the context. Often, it refers to data stream processing or querying data streams, such as that provided by StreamBase, HStreaming, and Splunk. In other contexts, "real-time" can refer simply to low-latency access to data that is already stored on disk. Sample applications that concentrate on reduced latency include Impala, Drill, Dremel, Spire, Apache HBase, and Splunk, as well as various in-memory RDBMSs such as SAP HANA.

The following list describes each of the four quadrants of the figure in more detail.

- **Batch and highly structured.** These solutions use a massively parallel architecture and a highly scalable and virtual infrastructure. This approach significantly reduces the cost of storage and uplifts processing efficiency for traditional DWs. (Examples solutions include Oracle Exadata, Netezza, and Teradata.)
- **Real-time and highly structured.** In these solutions, there is a shift toward stronger online analytic processing and data mining capabilities. In-memory databases store the data in the main memory instead of using a traditional disk storage mechanism, which reduces the I/O response time when querying data, providing faster and more predictable performance than disk storage. (Aster, Greenplum, Oracle TimesTen, and Hana are examples of these types of solutions.)
- **Batch and polystructured.** These solutions provide a software framework for Big Data, which may include a distributed file system, an engine that can sift large volumes of raw data, and management applications. (Cloudera, Hortonworks, and MapR are examples of vendors developing these types of solutions.)
- **Real-time and polystructured.** These solutions use Event Stream Processing to process multiple streams of events and identify meaningful insights by employing techniques such as detection of complex patterns of many events, event correlation and abstraction, and event hierarchies and relationships. (Sample vendors for this type of solution include Cassandra, Storm, and Splunk.)

Refer to Solution Sources later in this document for further discussion of how to choose the right Big Data tool for the right job.

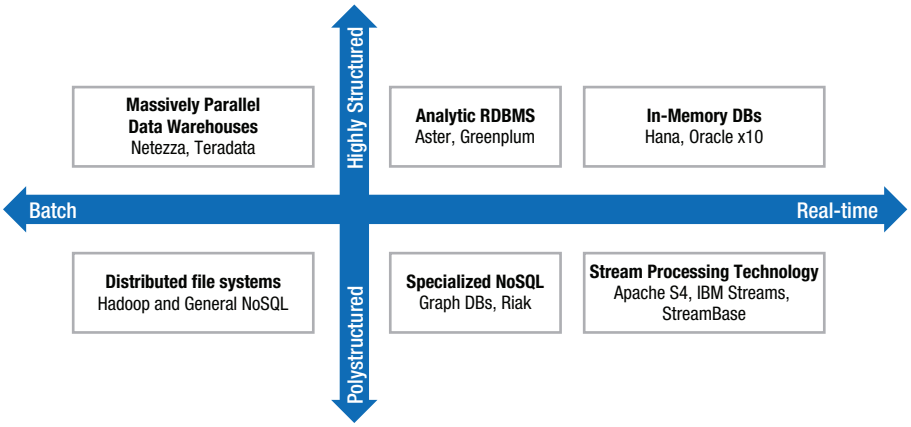

FIGURE 3: DATA STRUCTURE AND LATENCY REQUIREMENTS CAN GUIDE TOOL CHOICE.

Source: Forrester Leadership Boards, EA Council

Technology Gaps and Marketplace Maturity

Although the open-source community and proprietary vendors have made great strides in meeting marketplace needs for Big Data tools and technologies, the marketplace is still immature and evolving, for both consumers and solution providers.

- Many enterprises have not yet fully implemented a Big Data platform and are not yet familiar with the available choices, or even their own needs.
- No pool of experts exists for these emerging technologies that can help optimize and tune systems, compared to the availability of experts for traditional data systems such as MySQL or Oracle.
- There is currently a lack of training materials, online courses, and books relating to Big Data technologies. These materials are growing, but do not yet match the multi-year depth that a technology such as relational databases have.

Figure 4 shows the gaps that will need to be addressed before Big Data can mature and provide the most value.

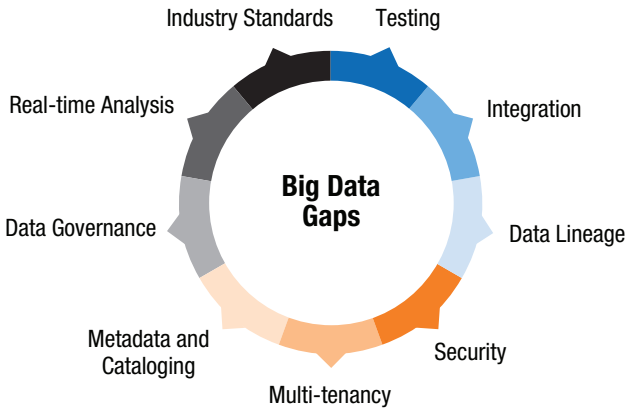

FIGURE 4: GAPS IN BIG DATA TECHNOLOGIES.

- **Development of industry standards.** Although Hadoop Distributed File System (HDFS) is a de facto standard of sorts, and SQL is a standard (although most Big Data solution providers do not claim to be compliant with actual SQL standards), the development of industry standards that govern various aspects of Big Data technology is perhaps the most important task facing both consumers and solution providers. Without standards, the marketplace will move forward only by fits and starts, with no cohesive goals or means to measure success.
- **Support for real-time analysis.** The ultimate goal of Big Data is extracting value from enterprise data assets. Therefore, enterprises need to be able to harvest data, perform compute operations, and utilize the intelligence gleaned. And, because the velocity of business is steadily increasing, these goals must be met on shorter and shorter timelines. Batch processing of data warehouses is no longer sufficient. Instead, enterprises are under pressure to move to micro batch processing, and to real-time data warehousing. Many Big Data technologies, Hadoop in particular, are batch-oriented. This factor limits what enterprises can do with the technology.
- **Support for metadata and cataloging.** Most Big Data solutions do not come with built-in data catalog support; adding customized support is time-consuming and expensive. Platforms that include add-on storage and processing engines require even further customizations and extensions to existing catalog services. Metadata and cataloging are important, because they provide the ability to track incoming and stored data at several layers: the data element, the data structure, and the file construct. With metadata, enterprises can define schema, thus laying a foundation for schema enforcement, conditional data quality workflow, and other conditional processing based on the data's adherence to the metadata definition in the catalog. Apache Hcatalog is one such tool, but wider adoption and development is necessary.

- **Support for data governance.** Data is a first-class resource; as such, it needs to be secure and to be subject to policies such as retention, data recovery, replication, and auditing. These policies need to be enforced at different levels of granularity. For example, HDFS provides file-level granularity, while Accumulo provides cell-level granularity.
- **Support for multi-tenancy.** When several resources are sharing data, security and resource (such as network, compute, and storage) governance are paramount.
- **Support for security.** Closely related to multi-tenancy and data governance, security is a critical aspect of a successful Big Data platform. Specifically, applications and tools need pluggable authentication and authorization, such as the LDAP. Customer data privacy and the confidentiality of an enterprise's commercial data becomes a real challenge with the growth of Big Data and new capabilities.
- **Support for data lineage.** Data is derived from other data and is used to derive more data. Currently, there is no way to centrally interrogate a single location for all of this data. This kind of a service needs to exist independently from a specific data source and connect to multiple data sources, such as Hadoop, HBase, Riak, or an enterprise DW.
- **Support for integration.** Established architectural patterns do not yet exist, and systems for tying things together are still in the design phase. Compare this situation to the integration that exists between enterprise application integration, enterprise service bus, and ETL technology, and it is obvious there is significant room for improvement in this area of Big Data technology.
- **Support for testing.** When Big Data volume, variety, and velocity are introduced to an existing information architecture ecosystem, testing the quality of information delivery is a major challenge. Currently, there are no mature tools available in the market to address this issue. Without appropriate quality control, the integration of traditional and Big Data will result in data quality issues.

Several of these areas are closely tied together. For example, security partially enables multi-tenancy and governance. Governance and data lineage are implemented as policies, which end up looking like metadata. Services that read those policies enforce them, and policies, like security, need to be centralized and integrated, so they are applicable to multiple kinds of storage.

Additional technology gaps that need to be addressed by the marketplace include:

- **Improved monitoring and management tools** to keep systems up and running.
- **Development tools** that enable customization and integration. While some exist already, such as Datameer and Karmasphere, this area is still emerging.
- **User interface (UI) frameworks,** similar to those available with traditional database solutions. Big Data systems don't have an ecosystem of UI tools. Most solutions do include Open Database Connectivity, Java Database Connectivity, or SQL to allow existing tools to work with Big Data technologies. But this may not be the best long-term solution, and Big Data application frameworks may become an area of interest.

PLANNING A BIG DATA STRATEGY

As with any major technological shift, an enterprise that wants to leverage Big Data to drive competitiveness and the pace of innovation needs to plan carefully. Much is at stake, both in terms of capital and time spent. Decisions made at the upper executive level about approach and investment, as well as those made at lower levels about staffing and infrastructure needs, can significantly affect the success of the project. Because Big Data represents a new way of interacting with and using data, a mindset shift as well as a structural shift is required, from the database administrators to the top-level executives. Implementing a Big Data solution is not just about recruiting data scientists, but represents a paradigm shift for the entire organization.

High-level Stakeholder Considerations

Before delving into the details of staffing and infrastructure, high-level enterprise executives should determine the ROI for a Big Data project. That is, what is the value the enterprise will gain, and how will it be measured? While IT staff focus on technological aspects, executives should focus on determining how the project will make money and/or save money. Obviously, there are a lot of factors to consider, including staffing and technical resources; license fees; deployment, monitoring, and maintenance costs; and business-related issues such as how the data will be used, project and budget considerations, and compliance and security requirements.

During this phase executives should decide whether, once data is gathered, analytics will be performed in-house or outsourced. Either approach can generate actionable insights, but the answer will differ from enterprise to enterprise. A few service providers are available now; as Big Data becomes even more prevalent across all industries, the number of analytics service providers will undoubtedly increase.

Another example of a decision point is how to distribute Big Data investment in the organization. Should solution investments be based on individual department needs (allowing the departments to make decisions about Big Data tools), or should investments be made at the enterprise level, made centrally through pooled investment? This impacts overall license, software upgrade, and operational costs of the organization. At a more granular level, should each business unit have their own specialized data scientist, or should business units share a pool of such experts who can perform analytics for various business units such as information security, marketing, and IT? The answers to these questions will vary from enterprise to enterprise, and depend on how much demand exists and how much the company wants to invest.

Executives should also be aware of the effect that gathering and using Big Data can have on customers in certain industries. For example, in the telecommunications, retail, and financial industries, it is possible to gather a large amount of data about what customers are doing and when they are doing it. Companies can then use this information to provide, for example, targeted promotions. However, there is often a certain point at which many customers sense an invasion of privacy, and this can cause customers to look for another provider of services. Executives need to understand the potential for customers' concern, and define how far the enterprise will take Big Data.

Answering these and similar questions will help the enterprise develop an operating model that encourages consistency and cost control across the organization, as well as a definitive roadmap for staffing, application selection, and infrastructure build-out.

Anatomy of a Typical Big Data Project

The following list describes how a typical Big Data project might flow at an enterprise that already has a Big Data platform in place. (If the task is to choose and deploy a Big Data platform, the steps would be much different. For example, refer to the discussion in Infrastructure Considerations later in this document.)

- **Understand new methodologies.** As stated in earlier sections of this document, it is critical to understand what Big Data is, where it can be applied, and what it takes to operate Big Data solutions. By understanding the differences between Big Data and traditional approaches to acquiring and using information to drive business decisions, enterprises can select an approach with the highest chance of succeeding. For example, the skills required for Big Data solutions are critical to success. Big Data skills are still relatively new—new ways of thinking and new tools are creating a void in the industry. According to the McKinsey Global Institute report cited earlier, by 2018 the United States alone could face shortages of from 140,000 to 190,000 people with deep analytical skills and a shortage of 1.5 million managers and analysts who know how to use the analysis of Big Data to make effective decisions. Existing experts come at a premium. Some enterprises may choose to consult a business advisory service to help jump-start a project.
- **Establish credibility for the project.** Because Big Data projects are using techniques that are unfamiliar to many accustomed to traditional data solutions, it is important to determine how the rest of the enterprise—especially top-level executives—perceive the project, in order to establish consistent messaging for it. It can also be helpful to establish clear criteria for success and how that success will be measured, gain executive sponsorship, and engage in regular outreach both upstream and downstream to show-and-tell and get feedback.
- **Understand the data and data processing requirements.** Understanding the data involved in a project is not unique to Big Data projects. However, because Big Data technologies and methodologies are new to many organizations, it is important to isolate variables in the project to as few elements as possible. By understanding the data, teams reduce the possibility that new processing paradigms, such as MapReduce, will be inaccurately diagnosed as a problem when the real problem may be the underlying data quality. Additionally, because Big Data solutions allow greater flexibility in the structure of the data, project teams cannot rely on data constraints and built-in quality checks that are inherent to many relational database systems; instead, this must be explicitly programmed into the data processing. Teams must also remember that Big Data, like other data solutions, must take into account compliance and security considerations.
- **Determine project management needs.** Dealing with new technologies and methods increases the likelihood of problems. To address this issue it is critical that proper project structure be established. Having a clear understanding of the various roles and responsibilities, as well as when and what the output of a task will be, will increase the chance of project success. For additional information on the staff and resources needs of the project, see the discussions of Staffing Considerations and Infrastructure Considerations, later in this document.

- **Identify and test use cases.** To ensure the project provides the best possible return on investment, project leaders should identify high-value use cases, ensuring that those use cases are applicable to Big Data solutions. For example, selecting a project that requires analysis that cannot be executed in parallel defeats one of the primary benefits of Big Data technologies. Once the use case has been selected, the project team should conduct pilots and proofs of concept. Many of the organizations that the ODCA has come into contact with have been using a model of "failing fast." Teams test often, try, fail, try again, and eventually determine the optimal mix of data, analysis, and treatment of the output for maximum success. By doing this, teams avoid the scenario where they have invested the entire project budget only to find an issue at the very end that diminishes or invalidates the value that could have been realized from the project.
- **Operationalize for full rollout.** Many of the early adopters of Big Data technologies shared with us that they underestimated the amount of operationalization that was required for Big Data technologies. It is critical to understanding the amount of staff required to support acquiring and processing data, creating and distributing output, and responding to system alerts and issues. Failure to properly operationalize a Big Data project can result in invalidating the value of the project, which in turn can cause the company to have a negative view of Big Data. Because Big Data technologies have not had the years of development and refinement that traditional database and business intelligence technologies have, extra effort is required to operate Big Data solutions with the same degree of availability, precision, and excellence. As these technologies mature, better graphical user interfaces, tools, and interfaces for integrating with existing operations infrastructure will emerge.
- **Operate and perform ongoing evaluation and cost analysis.** Having completed the preceding tasks, it's time to operate the solution and drive value for the organization. But this is just the beginning. It is highly recommended that teams perform ongoing evaluation, tracking the degree to which the solution drives value for the organization. Projects that are extremely successful require factual data to prove their success. And even if the project is only marginally successful, if the team has been deliberate about understanding the cost to operate and value derived from the solution, executives are more likely to approve future uses of Big Data solutions for company projects.

When to Use Big Data

Enterprises should perform a detailed needs analysis before diving into a Big Data project. For example, is the business using a public or private cloud? Is trending and after-the-fact analysis sufficient, or does the business require real-time analytics (sometimes referred to as "analytics at the edge")? What sorts of data will be stored and analyzed? (Answers could include social media site blog posts, network traffic flows, customer call data, intrusion detection logs, RFID data, sensor data, and more.)

As a best practice, enterprises should develop a decision framework that helps engineers and operations staff decide what the right data technology is on a use-case basis. Big Data tools are only one data technology and are not the appropriate fit for all data analysis projects. The framework should include all the data technologies available, which might include Big Data tools, relational databases, and transactional data stores. For example, if an enterprise wants to take third-party-measured web analytics and click logs in-house, the volume of data is significant, the velocity is high, and the row structure is variable—a Big Data perfect candidate—the data can be carved up and distributed across a cluster of machines working in parallel.

Traditional business intelligence, data warehousing, and online transaction processing solutions still have a role to play in the enterprise. For example, even among Big Data advocates, people acknowledge that online analytical processing (OLAP) has its advantages: It's optimized for a certain data-size range, works well when the structure of the data is well known, and is useful when data consistency and data quality need to remain high. Enterprises will continue to leverage the investments they have made in tools such as Business Objects, Microstrategy, Cognos; data warehousing platforms such as Netezza and Teradata; and general database platforms such as Oracle, SQL Server, and MySQL. Knowing when not to use Big Data solutions is equally as important as knowing when to use them. Understanding the target business problem, understanding the nature of the data, and taking into consideration the architecture and existing systems can help guide organizations to implementing Big Data when and where it will best complement existing data capabilities in the enterprise. Many organizations will have a mix of traditional databases as well as Big Data technologies.

In summary, as discussed earlier, it is important to remember that Big Data isn't something that necessarily supplants incumbent technologies, except in situations where scale and structure is an issue. Where the data still fits the existing solution, there's no need to change just for change's sake. Thus, enterprise resource planning applications can still effectively run on a relational database, but it doesn't necessarily make sense to use a relational database for importing and analyzing clickstream data.

Solution Sources

Whatever their focus, Big Data solutions fall into three types: pure open-source, third-party distributions of open-source code, and proprietary solutions. Each type has its advantages and disadvantages, shown in Figure 5, and an enterprise's Big Data ecosystem may contain elements from all three types, depending on enterprise needs. This is especially true for larger companies, which typically pursue initiatives for a unified data platform across the organization.

When choosing solutions, it is important to consider how the components integrate and act in concert with each other. To choose the right combination, an enterprise must evaluate their needs (see Infrastructure Considerations later in this document), the technical capabilities of staff (see Staffing Considerations later in this document), and the solutions themselves.

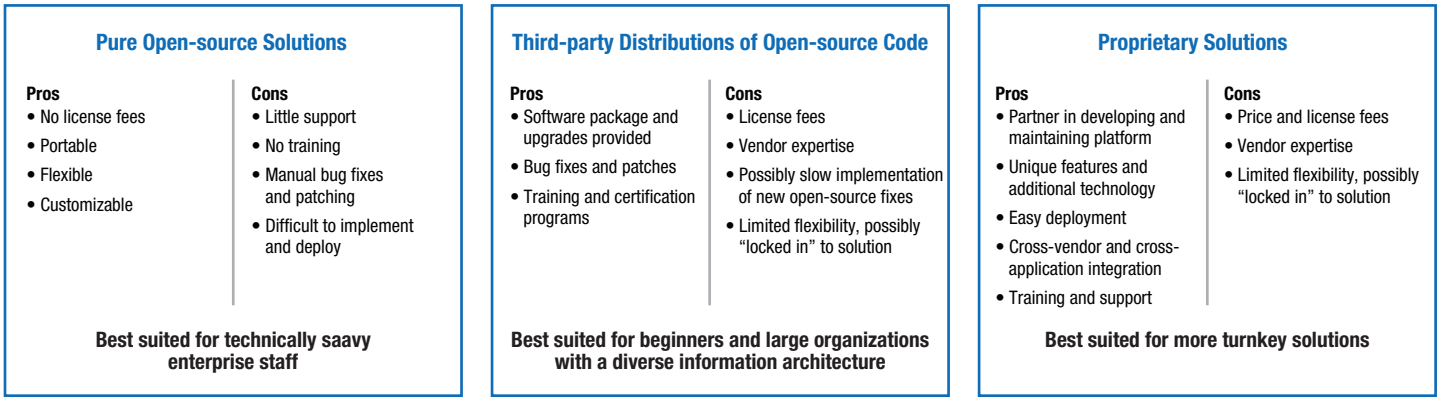

FIGURE 5: BIG DATA SOLUTION ADVANTAGES AND DISADVANTAGES.

Pure Open-source Solutions

Apache is the predominant open-source player; Apache Hadoop is an open-source implementation of the MapReduce algorithm introduced by Google in 2004. One of the leading and most well-known Big Data technologies, Hadoop is a distributed platform for parallel batch processing that can employ hundreds or thousands of nodes of commodity hardware on top of a distributed file system. Features such as fault tolerance, self-healing, and redundancy enable Hadoop to be highly available and reliable.

Although Hadoop is a leading Big Data technology, there are a number of other open-source Big Data projects underway, in areas including machine learning, cloud deployment, workflow, statistical tools, real-time analytics, data access, and data query flow. For example, Apache also spearheads the Cassandra Project, another distributed DBMS, as well as the Apache Accumulo Project, which is a sorted, distributed key/value store based on Google's BigTable design. Other notable projects not associated with Apache include Riak, MongoDB, CouchDB, Redis, Hypertable, Storm, Spark, and High-Performance Computing Cluster (HPCC). Some of these other open-source projects add value to Hadoop, while others may be used instead of Hadoop due to certain features that may be of interest to specific groups.

Open-source solutions offer several benefits. These include no license fees and portability. In addition, open-source solutions offer more flexibility. Enterprises are not locked into a vendor's release and patch schedule, and can customize the solution without violating patents or copyrights. However, open-source solutions can be daunting to implement for the less experienced, and except for online discussion groups, offer very little in the way of training and support.

Open-source solutions are typically best suited for an enterprise with a highly technical staff that can figure things out on their own, or know someone who can. When choosing an open-source solution, it is important to evaluate the maturity of the project. A project that has been underway for a considerable length of time and enjoys significant industry support may be a better choice than a project that is relatively new and untried. On the other hand, enterprises can help drive the marketplace forward by adopting such immature projects and helping guide them to completion and maturity.

Third-party Distributions of Open-source Code

For those just getting their feet wet in the Big Data space, it may be worthwhile to investigate one of the multiple commercial third-party distributions of Hadoop. Similar to third-party distributions of open-source operating systems (such as Linux), a third-party distribution can make the solution easier to adopt by enterprises. For example, the vendor usually provides packaging, bug fixes and patches, support, upgrades, and training. These vendors also form partnerships and alliances with hardware and software vendors, and may offer certification programs.

When evaluating such a solution, enterprises should determine how close the distributed code is to the open source: how much is changed and whether the solution offers anything above and beyond the open source, such as better installation or monitoring tools. Enterprises should also consider how "locked in" they may become to a particular solution. For example, it may be relatively easy, or fairly difficult, to switch from one third-party distribution to another.

Enterprises should also evaluate whether a particular third-party distribution has open APIs that allow other tools to be easily integrated and used, such as business analytics or management software. Since many enterprises already have established business tools, the ability to use those with the chosen Big Data solution is critical in terms of minimizing training and financial investments.

Another consideration is whether the third-party vendor is well connected to the open-source community. When the vendor makes an improvement, do they share it with the open-source project? Conversely, when there are new fixes in the open-source community, does the vendor have an orderly process for quickly incorporating them, based on the priority of the change? It is also important to determine what level of expertise the vendor has in-house, their support policies, and license fees. In general, third-party distributions of open-source software are best suited to large organizations with a diverse information architecture.

Proprietary Solutions

Although it is possible to use only open-source and internally developed solutions to build a Big Data platform, many enterprises may prefer to build their platform using a combination of these and out-of-the-box solutions. One advantage of choosing a proprietary solution is that the vendor can serve as a partner in developing and maintaining the Big Data platform, sometimes even providing developer support and training. It would be impossible to list all the available proprietary Big Data solutions available in the marketplace, and the list changes daily. See the Resources section later in this paper for some links to web sites that can help identify what solutions are currently available.

Enterprises may be more comfortable having a well-defined path to access support personnel, and proprietary vendors tend to be very familiar with their own product; therefore, getting answers to specific questions may be easier than with open-source solutions. Another possible advantage of proprietary solutions concerns the likelihood of cross-vendor and cross-application integration, compared to open-source solutions which generally place more of the burden of integration on the enterprise.

When evaluating a proprietary solution, enterprises should consider pricing and licensing fees. But more importantly, they should determine what's distinctively different from the open-source offerings: How is the proprietary solution unique, and what value does that add? For example, a solution may have additional technology or may be extremely easy to deploy and maintain. As with choosing any other type of solution, other considerations include interoperability, comprehensiveness, and extensibility.

Also, given that proprietary solutions are dependent on a single vendor, enterprises should consider the reputation or proven track-record of the vendor to provide what is needed for organizations today—and for the forseeable future. For example, if new enhancements are made in open-source models, these may or may not be incorporated into the proprietary code. The choice to the enterprise is if the unique value offered is worth that risk, and for each organization this answer may be different.

Staffing Considerations

Big Data represents both a steep technological learning curve as well as a major shift in how people interact with and use data. The combination makes staffing one of the major challenges associated with ramping up a Big Data project. In fact, acquiring the new skills can be more difficult than implementing the technology. This includes skills for stakeholders and business units, operations and infrastructure, scientists and analysts, and developers. Each tier must have a greater understanding of Big Data and shift away from the traditional way of performing data collection and analysis.

Overall, enterprises interested in Big Data should hire people with specific blended skills, a demonstrated intense curiosity, and a track record of driving business results. The most successful candidates—especially team leads—should possess multiple skills, including the following:

- **Business acumen.** Knowledge of the business
- **Data algorithms.** Turn the business problem into a set of rules/steps
- **Software.** Encode those rules in the software

Traditional DW and business intelligence (BI) focuses on extracting data from a given system, bringing it into a transformation area, and loading it into a DW or DW appliance. This process requires modeling, transformation, and integration techniques, all of which are different for Big Data, although there are commonalities, such as acquire, transform, store (in one order or another), analyze, and query.

The following quote summarizes the staffing needs for Big Data:

> "The ability to extract analysis from oceans of unstructured data requires a unique toolset, beginning with coding skills and ending with an intense curiosity… Think of Big Data as an epic wave gathering now, starting to crest—if you want to catch it, you need people who can surf."
>
> Harvard Business Review's Thomas H. Davenport,
> in "Data Scientist: The Sexiest Job of the 21st Century"

By overlaying project requirements (needs) with a team's strengths (roles), enterprises can quickly determine where they need to focus on adding staff.

Where to Find the Experts

Based on availability and budget, enterprises have three choices for staffing their Big Data projects:

- **Re-skill internally.** If a good programmer is available who has strong Java skills and understands the ideas behind distributed computing, it may be possible to train him or her. He or she could attend external training classes and participate in online resources such as developer days and Hadoop user groups.
- **Hire consultants.** Taking advantage of external capability is a possibility. Enterprises may choose this approach for areas that require specific skills that may change over time because Big Data is new and tools will continuously change.
- **Hire direct.** Enterprises are more likely to make a long-term investment in capabilities that represent more stable skills that will last for a long time.
- **Outsourced analytics.** Enterprises can form partnerships with information service providers. In this scenario, company data is sent to the service provider (for example, Quantium). The service provider analyzes the data on behalf of the enterprise, along with wider external and social media data, and provides the nuggets of actionable insights back to the organization. This approach reduces the staffing, storage, and tool costs for an organization that may be hesitant to implement a Big Data solution in-house, considering the overall immaturity of Big Data capability in the industry.

Stakeholders and Decision Makers

Big Data enables business decision makers to think about their business in different ways, to pose questions in different ways, and to use data in different ways. Big Data technologies have opened up capabilities that were not previously available, either due to cost, performance, time to market, or other reasons. But although now analysts can provide a greater array of analysis to business decision makers, it is often the case that they are not familiar with some of the techniques or how to ask questions relevant to the analysis techniques. Therefore, it is important to help them make the mindset shift: to "train" them to think in new ways. If the business decision makers know what insights are possible, the science/analyst team can know better what to focus on, creating a collaborative synergy between the two realms.

Enterprise and Solution Architects

In the past few years, the information technology industry has made a significant shift toward information delivery. As a result, there has been an exponentially increasing demand for data and information architects to define data architecture (at the enterprise level and at the individual solution level). These architectures provide an end-to-end data flow by spanning a rapidly growing variety of distributed open-source, proprietary, and cloud solutions integrated with in-house legacy applications.

The new Big Data concepts and capabilities will further add to the specialized capabilities required of data and information architects. Solution and enterprise architects need to be able to determine whether continuing to develop with traditional data architecture is appropriate and where Big Data is more appropriate, and how to integrate the two environments. They need to be trained to have a good understanding of where and how Big Data solutions fit in the overall information architecture ecosystem of the organization to meet new business demands. Two examples include:

- Real-time decisions connecting the organization's front-line processing and head office with back-end calculations such as real-time credit decisions
- Business analytics of unstructured data such as sentiment analysis, customer interaction path finding, patterns, and correlation

Operations and Database Administrators

Database administrators need to know how to plug in the new Big Data infrastructure and how to get the best use of it—it's not just another relational database. It is also important not to underestimate the amount of operationalization that is needed to support a Big Data project.

If the enterprise is using a third-party product (either distributed open source or proprietary), the vendor may be able to help determine how much operations support will be necessary. In particular, Big Data solutions do not typically include the ancillary management tools and helper processes that accompany many of the vendor-packaged traditional data applications. Therefore, operations staff is required to integrate the Big Data solution into existing monitoring and reporting systems.

Scientists and Analysts

Analysts need to know how to use Big Data tools and how to move beyond answering merely already known questions to finding unasked questions. This process requires an open mind to identify nuggets of insight as they appear. In addition, Big Data analysts need to be more tool agnostic and be willing to change skill sets when necessary, as available tools emerge and evolve. It is also important that the analysts be able to apply business judgment to the information; that is, someone who has enough knowledge about the business who can ask the right questions, find the patterns, and turn them into business value.

Software Developers

While a background in prior data systems, such as RDBMs, is useful, software developers need to experience a paradigm shift from being RDMS-oriented to leveraging a distributed file system. To achieve this shift, developers should work closely with data scientists to co-evolve a solution. For example, data scientists need to understand the capabilities of various tools, while developers need to understand the needs of various analytic techniques. Developers also should build the technical expertise to work with the new class of Big Data tools.

More importantly, developers need to learn to think differently. For instance, NoSQL systems often have no transactional guarantees, so any application that relies on such guarantees needs to be re-tooled. Also, instead of thinking about a single storage path with a relational database with support for transactions and CRUD (create, read, update, and delete), developers need to work with distributed file systems, sets of documents, key-value stores, graph databases, or one of the other NoSQL approaches.

When working with a distributed file system, developers must think about data processing, analysis, and how analysis can be broken down and disseminated across a cluster of distributed processing nodes. Not all analysis techniques can be broken down this way. Additionally, developers or other staff must be trained to be able to test the information delivery quality in Big Data solutions.

Software developers within the solution provider community, as well as product development managers, need to think of innovative ways to integrate Big Data with new and existing user interfaces, application code, development and testing platforms, SDKs, APIs, and integrated development environments (IDEs).

Database Developers

While software developers often adjust easily to Big Data-style processing because they already think in recursive patterns and algorithms, database developers may have more difficulty making the mental shift; these developers' mindset is usually set-based not algorithm-based. An enterprise may need to invest in additional training for database developers if re-skilling internally (see the Training section that follows.) Enterprises should leverage database developers' data modeling skills, although these skills may need to be adapted to potentially different technologies. In comparison, software developers typically do not focus on data modeling.

Training

Technologists tend to be particularly fond of one tool or another. To help staff make the shift from traditional ways of thinking about data to the Big Data paradigm, it is helpful to find a way to meet the existing DW/BI staff on their own ground and bring them along through somewhat of a social change, as they learn to think differently, use different tools, and find different ways of interacting and working with data.

For example, the open-source product Hive (HQL) looks a lot like a database, with tables and queries. Similarly, the Pig language for defining and processing data flows is based on concepts similar to how DW engineers think about ETL. Choosing Big Data tools that look and work similar to what existing staff are used to working with can help ease the transition. Training can provide information about what is going on behind the scenes and what makes these tools different from the traditional ones. This approach helps bridge the gap between where staff are and where they need to be.

Traditional DW, BI, and information management (IM) teams are often rooted in large vendor apps, which are significantly different from open-source solutions. To ease the transition, enterprises can hire a subject matter expert (SME) who is familiar with Hadoop, distributed data processing, textual search, distributed computing clusters, and code customization. Pairing a SME with a traditional DW staff member and having them work on projects together can help bring the DW/BI/IM teams up to speed on Big Data. Expect some friction—two experts who may see solutions to a problem in two different ways—but this approach can create a bridge and pull existing staff along the learning curve.

Vendors can help provide training programs. Enterprises can choose to purchase such support at the beginning of a Big Data project and may continue to use that support as the project progresses. Once a few key personnel have attended vendor classes, enterprises can also develop their own training videos and presentations and code examples as an extension for training additional staff.

Evangelization is the next step. It is important to make people across the entire enterprise aware of what Big Data is and what it can do. But be careful: Evangelization can snowball. For example, an enterprise may construct a hosted development/testing/production cluster in place, thinking to insulate developers from setup so they can focus on simply using the new tools. However, developers might want to "get their hands dirty" and actually set up the entire environment. Also, when a team is identified to work on the new technology, there may be some people who wanted to be included that weren't (for example, because there was no room on the team, they didn't have the right skill set, or there wasn't enough time). If there is a lot of buzz in the organization about Big Data, upper-level management should determine the messaging to the rest of the staff about how they can get involved and how they can participate in the future.

One final note about training: Big Data is an area that is rapidly evolving. Therefore training isn't a one-time occurrence. Also, some staff like to learn new things and spend time reading the Big Data news. These individuals stay abreast of change. But other individuals aren't like that. Technology isn't their hobby. By partnering with the internal learning organization, the gurus can share their learnings with less passionate but equally talented individuals.

Infrastructure Considerations

When deciding on the tool set to use to tackle Big Data challenges, a good place to start is with the requirements. The aspects to consider at a high level are the "three Vs" mentioned earlier:

- **Volume.** What is the expected size of data that will be analyzed?
- **Velocity.** What is the expected velocity of data? What are the needs for processing in real time or near-real time? What are the latency requirements?
- **Variety.** How varied is the data? How much is it subject to change? How diverse are the sources?

In addition, enterprises should also determine what type of value is to be derived from the data and how it is to be derived.

At a lower level, enterprises should consider the following:

- Which components to replace and which to complement
- Capacity planning
- Trade-offs
- Latency requirements
- Solution complexity

Replace, Complement, and Close Gap

Of course, enterprises already have data systems, with associated tools and infrastructure. For example, Oracle or MySQL may be in use, with preferred report builders, data exchange protocols, and analytic tools. When beginning a Big Data project, it is important to understand that the Big Data infrastructure is not going to replace all these existing systems. Instead, it will integrate with much of the existing systems, and replace one or more components.

For example, an enterprise may be using an SQL-based database, but has found that it doesn't scale and isn't flexible enough to meet evolving data volumes. However, the executives appreciate the reports, graphics, and charts and the fact that they are easy to publish to a PDF and put on the company web site. In this situation, the data system on the back-end isn't keeping up with the enterprise's needs, so new technology is needed to solve the problem. Part of solving the problem will be determining how hard it is to mesh with existing components, and how expensive it will be to re-tool. By taking a replace/complement/close gap approach, enterprises can gain a holistic benefit from Big Data.

The concept of replace/complement/close gap is also applicable within a single Big Data solution. Although Hadoop is very common, an enterprise may mix and match at several different layers of the solution, because some vendors or other open-source projects offer a unique feature or improve on a certain function. For example, it is possible to replace the HDFS with something else. Or a NoSQL database vendor may use HDFS as a back-end. In choosing overall solutions and layers within solutions, enterprises should make decisions based on their requirements for performance, data access, and data resiliency.

Capacity Planning

In addition to the existing infrastructure, enterprises should consider the capacity of existing and new infrastructure. Certain data systems will require more hardware, different storage types, and different network configurations. If the Big Data solution is cloud-based, it is also important to determine how well the tools will work in a cloud infrastructure as opposed to in a data center under direct enterprise control.

It is easy to forget that Big Data is called that for a reason—it comes in only one size (big). When estimating storage requirements, for example, a traditional DW solution may require GBs of storage. But a Big Data solution may require terabytes, or even petabytes of storage space. When estimating infrastructure requirements, the Big Data team must keep in mind that the paradigm of how data is collected, processed, stored, queried, and iterated over records is very different compared to a traditional DW solution.

Trade-offs

Any Big Data solution is going to have varying levels of several attributes. While there are several ways to categorize these attributes, the main idea is that the Big Data team needs to think about what is most important to the project, and choose tools and approaches that make the appropriate trade-offs to support the chosen attributes.

One way to think about project attributes is called the "CAP Theorem." In this paradigm, consistency, availability, and partition tolerance are balanced against each other. The theorem states that you can optimize for any two—for example, consistency and availability, or availability and partition tolerance—but pursuing all three can be very expensive, or even impossible. For example, NoSQL systems are very fast but are not always consistent; instead, they use the concept of "eventual consistency." For certain projects, such as dashboard reporting, availability is more important than consistency. In other situations, such as financial transactions, consistency is very important, and availability less so.

Another model is called "ACID"—atomicity, consistency, isolation, and durability. In this model, the data system must offer guarantees in each of these areas in order to be considered reliable. Atomicity refers to an "all or nothing" approach to transactions; if one part of the transaction fails, the entire transaction fails. Consistency indicates that the data system ensures that only valid data is written to storage. Isolation implies that concurrent transactions do not affect each other. Durability means backups or other recovery mechanisms exist that ensure that data is not lost. This model is appropriate for evaluating many NewSQL applications, in order to choose the tools that offer the best ACID guarantees.

Yet another set of attributes to consider when building a Big Data platform is elasticity, availability, and scalability. Elasticity means that as workloads grow and shrink, the system can dynamically follow suit, providing efficient use of resources. Elasticity is particularly important in a virtualized system, keeping as few machines idle as possible. Cyclic data workloads, such as end-of-the-month financial reporting, are one example where elasticity is important. Cloud-based services and eCommerce, where demand for an application could suddenly spike, provide another example. In this model, availability refers to redundancy. If a portion of the infrastructure fails, does the solution degrade gracefully, or, better yet, self-heal? To enhance availability, Big Data teams should identify single points of failure (SPOFs) and mitigate the risk by using a backup system. Scalability relates to how easy it is to add resources and whether the workload scales linearly with additional resources.

Latency

Big Data systems fall into two main categories: batch processing of a huge amount of data overnight and interactive queries through live dashboards and analysts, based on real-time. The tools appropriate to each scenario are quite different and require significantly different infrastructure.

Complexity

Sometimes Big Data teams become so involved in choosing tools based on their capabilities—or cost— that it is easy to overlook the complexity of deployment, maintenance, monitoring, and management. The technology press tends to contribute to the problem by reporting on new features, but making little mention of how difficult these features are to implement or maintain. Enterprises should not be lulled into ignoring these important questions.

- **Deployment.** How difficult is it to set up? How do I configure the system? How do I change the configuration?
- **Maintenance.** How do I keep it running? How do I handle updates? How do I control the system admin access versus user access? Does everything have to go through admin? Is there too much privilege granted to users?
- **Monitoring.** Is there a robust set of tools to track the health of the system? Are these tools included, or do I need to obtain them separately? Are there threshold alerts for critical resources? Does it support activity monitoring to tell me what the system is being used for and by whom? Does it support license management (using only 20 percent of the licenses)?
- **Management.** How easy is it to perform ongoing changes to configurations? Is automation available? In a large enterprise, a good management tool that can automatically execute changes is far preferable to having an admin change 1,000 nodes manually.

SUMMARY RECOMMENDATIONS

This document has explained what enterprises can do with Big Data and how solution providers can help meet these needs and opportunities. Using the information in this document, with special attention to the discussion of use cases, Big Data solution providers can explore industry-specific solutions, providing differentiation and value above and beyond their standard offerings.

When starting to explore Big Data, enterprises can engage in the following best practices:

- Understand what Big Data is and how it differs from traditional databases and analysis techniques, and what value it can bring to the enterprise.
- Examine the enterprise's needs, identify possible use cases, and test them.
- Carefully consider the enterprise's requirements and choose Big Data tools and technologies that meet these requirements.
- Develop a Big Data strategy that includes paying careful attention to staff skills and infrastructure needs.
- Work with other enterprises and solution providers through the ODCA to help drive the Big Data market toward maturity and standards.

ODCA CALL TO ACTION

In the interest of following through on the mission of the Open Data Center Alliance, this document is an introduction to the Big Data landscape. By providing an introduction to the topic, it lays the foundation for future work by the ODCA DS Working Group. In addition to Big Data, the Working Group will assist other ODCA working groups by providing data expertise to the cloud computing topics those groups are addressing. ODCA solution providers and ODCA consumers must work together to further define the open specifications, formal or de facto standards, and common IP-free solution designs if we are to realize an open, interoperable, and thriving market of data solutions.

The following ODCA projects are under consideration for the Data Services Working Group to contribute to in 2013:

- Revision of previously published ODCA usage models to revise the data perspective of each usage model. These revisions may include usage models for interoperability, management, licensing, orchestration, automation, and other published topics.
- "Data Management Features of Big Data Solutions" Usage Model. This project will describe additional features that large enterprises require of Big Data solutions to maintain feature parity with other data solutions in the enterprise. Metadata, master data management, support for data governance, and data lineage are examples of data management features that enterprises have come to rely upon.
- "Exchanging Data between SaaS Providers and the Enterprise" Usage Model. This project will delineate the requirements that enterprise consumers of SaaS have for exchanging data with the enterprise.
- "Data Encryption" Usage Model. This project will describe the data encryption requirements that large enterprise consumers of cloud services require.
- "Data as a Service" Usage Model. This project will describe the requirements that cloud service providers need to meet as a full-service PaaS offering.
- "Management and Operations of Big Data Solutions" Usage Model. This project will describe the requirements that Big Data solution providers need to meet to ensure that consumers can integrate Big Data solutions into their operational environment with adequate management tooling; interfaces for integrating with existing services management, monitoring, and performance management systems; and other support for operations staff.

RESOURCES

(links are clickable)

- Big Data landscape chart created by Matt Turck
- Building data science teams
- "Data Scientist: The Sexiest Job of the 21st Century"
- "Filtering the Digital Exhaust"
- Open Data Center Alliance
- Apache Hadoop Foundation

APPENDIX A: USE CASE DETAILS

The following table provides detailed information about various use cases and in which industry sectors they are most relevant. All of the use cases take advantage of the ability to parallelize data access and computation.

USE CASE	DESCRIPTION	Finance	Retail	Telecom	Utilities & Energy	Manufacturing	Healthcare	Transportation	Government	Mobile	Marketing
Risk Management	Finance organizations seek to mitigate risk with continuous risk management and broader analysis of risk factors across wider sets of data. These organizations need to integrate growing amounts of data from multiple, standalone departments across the firm and analyze this data in near real time.	•									
Risk Profiles	The ability to accurately assess the risk profile of an individual or a loan is critical to offering (or denying) services to a customer. Getting this right protects the bank and can create a satisfied customer.	•	•								
Dynamic Pricing	By tracking users' history, businesses can offer dynamic prices at the point of sale to match what people are willing and able to pay, according to their prior behaviors.	•	•							•	•
Rogue Trading	Deep, near real-time analytics that correlate accounting data with position tracking and order management systems can provide valuable insights that are not available using traditional data management tools because the data sources are multiple and inconsistent.	•							•		
Fraud Detection	Correlating data from multiple, unrelated sources has the potential to identify fraudulent activities. For example, correlating point-of-sale data (available to a credit card issuer) with web behavior analysis (either on the bank's site or externally) and cross-examining it with other financial institutions or service providers can improve fraud detection.	•	•						•		
Companion to Real-Time Engines	Many companies can benefit by combining Big Data streaming solutions with large batch-oriented data transfers to a distributed file system such as Hadoop.	•	•	•	•	•		•			
Sentiment Analysis	Whether looking for broad economic indicators, specific market indicators, or sentiments concerning a specific company, there is a wealth of data available to analyze in both traditional and new media sources. While basic keyword analysis and entity extraction have been in use for years, the combination of this old data with new sources, such as Twitter and other social media sources, provide great detail about public opinion, in near real time.	•	•	•	•	•	•	•	•	•	•
Predictive Analytics	Within capital markets, analysts have typically used advanced algorithms for correlations and probability calculations against current and historical data to predict markets—a slow and expensive approach. The large amounts of historical market data, the speed at which new data needs to be evaluated, and the availability of commodity hardware make this a good Big Data use case.	•	•	•	•	•	•	•	•		
Network Monitoring	Networks, such as transportation, communications, police protection, and local office networks, can benefit from better analysis. Massive amounts of data collected from servers, network devices, and other IT hardware can enable administrators to monitor network activity and diagnose issues before they adversely affect productivity.	•	•	•	•	•	•	•	•	•	•
Action Tracking Analysis	The ability to process, track, and analyze individual actions (such as financial trades) lets businesses quickly change strategy with greater insight.	•	•	•	•	•			•		
Data Integration	Businesses often have data coming from several different sources. Big Data technologies allow the integration and efficient processing of this disparate data.	•				•	•				
Social Graph Analysis	Big Data technologies can mine social networking data to identify the participants that pose the most influence over others inside social networks. This helps enterprises ascertain the "most important" customers, who may or may not be the customers with the most products or who spend the most.		•				•		•	•	•
Marketing Campaign Analysis	The more information made available to a marketer the more granular targets can be identified and messaged. Big data is used to analyze massive amounts of data, such as click-stream data and call detail records that are impossible to analyze with traditional relational solutions.	•	•	•			•	•	•		
Recommendation Engine	Recommendation engines match products, people, and ads based on analysis of user profile and behavioral data. This was one of the first use cases for Big Data and has helped develop the technology into what it is today.	•	•	•			•			•	•

USE CASE	DESCRIPTION	Finance	Retail	Telecom	Utilities & Energy	Manufacturing	Healthcare	Transportation	Government	Mobile	Marketing
Customer Retention and Churn Analysis	An increase in products per customer typically equates to reduced churn. However, analysis of customers and products across lines of business is often difficult because formats and governance issues restrict these efforts. Big Data technologies can help perform wide-scale analysis and identify patterns that indicate which customers are more likely to expand their relationship with the company and which ones are most likely to switch to a competing vendor. Actions can then be taken to save or motivate these customers.	•	•	•	•		•			•	•
User Experience Tracking	Business can gain tremendous insight into how customers interact with products by using measurement tools that collect interaction data. Online, heatmaps and other data collection agents can be used to track mouse clicks; brick-and-mortar stores can use cameras and sales data in order to discern how customers move in the store.	•	•	•	•		•			•	•
Behavior Modeling	Using Big Data technologies, businesses can observe and build profiles based on customer interactions. These profiles can help the business to more intelligently offer the products and services that match how customers have interacted with the business in the past.	•	•	•	•	•	•	•	•		
Manufacturing	In manufacturing, machines create a massive amount of data that operators can use to track production and performance. This is increasingly important in vulnerable industries such as oil production, where sensors can serve as an early indication of potentially catastrophic problems.		•	•	•	•		•	•		
Exploration/ Visualization	Receiving information visually, instead of through spreadsheet data, can improve the understanding of larger trends. Software companies have created applications to build such interfaces, tailored to various industries that allow people to get greater value more quickly from their Big Data.	•			•				•		
Machine Learning	Statisticians use Big Data technologies to run computationally intense algorithms designed to enhance their understanding of complex mathematical problems.	•	•			•					
Technical Product Optimization	To aid in product development, scientists use Big Data as a critical part of the information they gather to obtain better insights into how to improve the product. Data is collected using sensors and is then processed using Big Data platforms such as Hadoop.					•					
Wireless Reporting and Monitoring	Organizations may need to monitor sites remotely with wireless devices in order to ensure operations are running smoothly. A device transmits the status of a particular location as well as any other metric important for operations. Large numbers of these devices working together generate a vast amount of data, which can be used to optimize operations.			•	•	•		•	•		
Observational Research	Researchers examine massive amounts of data in order to detect significant items of interest among the large amount of noise in data sets. Examples include oceanography or astronomical data, where small changes can be important signs of potentially major findings.	•							•		
Capacity Forecasting	Processing Big Data can provide significant insights into managing current inventory and making future projections.		•			•	•	•		•	
Delivery and Logistics	Logisticians and delivery service monitor packages traveling from one location to another, using a combination of GPS and RFID tags. This enables logisticians to know if deliveries haven't arrived or when complications occur in the delivery process.				•	•	•	•			
Archiving	Big Data technologies can archive large amounts of data, acting as the "new tape backup." With commodity hardware available, organizations can archive data even when they are not sure what to do with it yet.	•	•	•	•	•	•	•	•	•	•
Imaging Analytics	Energy and petroleum companies measure and collect massive amounts of data to gain a better understanding of the Earth. Using sonar and other similar technologies for geospatial and acoustic analysis, companies can measure acoustic waves to construct maps that show the location of underground resources.				•				•		
Research and Development	The development of new products, such as those from pharmaceutical manufacturers, can use Big Data technologies to comb through enormous volumes of text-based research and other historical data to assist in the product development process.	•	•	•	•	•	•		•		
IT Operations	Big Data technologies can help optimize network performance, perform network capacity planning, help with incident management and system usage monitoring, enhance extract-transform-load operations, and perform root cause analysis	•	•	•	•	•	•	•	•	•	•
Cloud Computing	Cloud-hosted solutions can generate a large volume of a variety of data, including clickstream data, network traffic, application data, and more. Big Data technologies can help integrate and analyze this data to manage capacity, predict user behavior, and develop new offerings.	•	•	•	•	•	•	•	•	•	•