orbus software

White Paper
Enterprise Architecture Approaches to Big Data

Key Concepts and Best EA Practices

WP0138 | March 2014

Guy B. Sereff

Guy Sereff is an author, speaker and technology practitioner. His Technology Industry experience includes Application Research and Development, Large-Scale Technology Management, and Global Enterprise Architecture.

As well as a pragmatic blend of Strategy and Tactical execution, Guy also has extensive Architectural Domain experience which covers Business Architecture, Information Architecture, Solution Architecture and Enterprise Architecture.

There is currently a tremendous level of interest in Big Data and its promise of new insights into patterns of behavior that may be difficult to detect with more traditional information resources. Many organizations are beginning to explore the potential value proposition and trying to determine how Big Data fits into their overall technology strategy. A common observation is that much of the work around Big Data is occurring outside the purview of the entity's architecture function. Considering the size of investments being made, Enterprise Architects need to ensure they are entrenched in these efforts in a way that supports growth without sacrificing important controls and safeguards.

In this discussion, we'll review a few basic Big Data concepts, such as its conceptual origins, widely accepted data characteristics and a few of the more widely used tools and techniques such as NoSQL and Hadoop. We'll also discuss how Big Data and traditional Enterprise Data Warehouse analytics differ and are more complimentary than contradictory.

From there we'll discuss recommended best practices for Enterprise Architects to guide, influence and facilitate the growth of Big Data within their organization. In this context, the recommendations for Enterprise Architects apply equally to Information Architects and Data Architects, depending upon how each organization segregates the various architectural domains. These best practices include:

- Establish Adaptive Enterprise Data Principles
- Don't Abandon the Enterprise Data Model or Data Governance
- Establish a Big Data Reference Architecture
- Clarify Big Data Roles, Accountability and Decision Rights

Basic Concepts

Overview

At the time of this publication, few things are as talked about as much in the technology industry as 'Big Data'. A simple Google search on the term 'Big Data' yields more than 1.9 Billion results, with content ranging the gamut of definitions, vendor products, case studies, and articles from zealots and naysayers alike[i]. It seems that everyone is talking about it, although not everyone can quite agree as to what it is or how best to do it.

While we won't attempt to provide the de facto definition of the term Big Data, we will spend a moment to understand its origins and some of its fundamental principles as they relate to content, tooling and utilization. Industry pundits would lead us to believe that Big Data is a very new phenomenon. However, looking at a few points in history will help us better appreciate Big Data's conceptual roots, which actually run back nearly 15 years ago, standing on the shoulders of relational data concepts that go back even more than that, better than twice that long.

One of the first identified uses of the term "Big Data" appears in the ACM digital archives, taking us back to 1997, where Cox and Ellsworth were working together on a project for the U.S. National Aeronautics and Space Agency (NASA) at the time. While addressing the problems of performing visualization of Computational Fluid Dynamics and the massive amounts of information that needed to be processed in real time, they coined the term in an IEEE publication.[ii]

> *"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data."* [iii]

This 'rocket science' concept developed within the public sector scientific community was soon recognized as having potential in the private sector. In 2000 Diebold offered the following observation as part of a presentation on advances in economics and econometrics he made at the Eight World Congress of the Economic Society:

> *"Big Data refers to the explosion in the quantity (and sometimes quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, sample sizes are no longer fruitfully measured in "number of observations," but rather in, say megabytes"* [iv]

In 2001, the META Group identified what has become heralded as the ‘Three V’s of Big Data’, although the term Big Data was not actually included in the publication:

> *“While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity, and variety.“* [v]

Now lets fast forward to a few more contemporary definitions:

Gartner *“Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.”* [vi]	**Forrester** *“[Pragmatic] Big Data is the frontier of a firm’s ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.”* [viii]
InfoWorld *“In the never-ending quest for a competitive advantage, organizations are turning to large repositories of corporate and external data to uncover trends, statistics, and other actionable information to help decide on their next move. Those data sets, along with their associated tools, platforms, and analytics, are often referred to as big data.”* [viiii]	**Mayer-Schönberger & Cukier** *“There is no rigorous definition of Big Data.* *...Big Data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value in ways that change markets, organizations, the relationship between citizens and government and more.”* [ix]

Table 1 - Various Big Data Definitions

While we see some similarities between the definitions above, we don’t find a singular authoritative description. Rather than declare a particular winner, we will instead consider three aspects of Big Data that are often clubbed together. The first aspect we’ll cover will be an overview of common characteristics of the data itself – what often distinguishes Big Data from ‘regular’ data. The second aspect we will discuss relates to enabling tools and technologies used to process information with Big Data characteristics. The final aspect we’ll review is that of Big Data analysis and utilization, based on the usage of the information derived from Big Data repositories and the supporting tools.

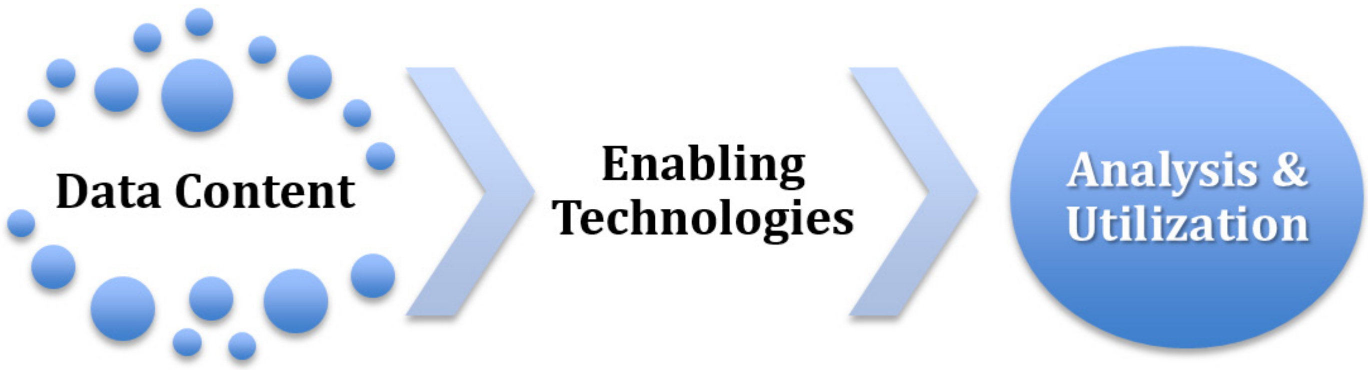


Figure 1 - Three Common Aspects Under the Banner of 'Big Data'

Discussions around Big Data are often imprecise, and you may find the term indiscriminately used to refer to any one of the aspects above, as well as commonly used as a generalized term referring them collectively as a domain, discipline or ecosystem. This isn't necessarily right or wrong, nor good or bad; simply an indicator that contextual clarification will improve conversations with Big Data suppliers and consumers alike.

Data Characteristics

There is general consensus around the use of the '3Vs of Big Data' as a means to describe the characteristics and nature of the data and what differentiates it as 'big' – Volume, Velocity and Variety. Others in the industry, be they authors, consultants, subject matter experts or product vendors have offered expanded data characteristics of Big Data to the conversation as well, using such terms as Veracity, Value, Variability and others. While not without merits, many of these additional observations regarding data characteristics are expressions of data quality, usefulness and timing anomalies (i.e. peak volume surges). For this discussion we'll stick to the core V3 and potentially explore these additional traits in future publications.

Volume

As expected, Volume refers to the shear amount of data organizations and institutions are amassing in their various systems of record, operational data stores and data warehouses. How much data is considered 'big'? Depending on the size of the institution and its functional domain, the amount of data can vary widely, from gigabytes to petabytes and beyond.

Transaction-intense organizations, such as global financial institutions and consumer facing electronic retailers generate and track massive amounts of discrete data points, not to mention additional metadata required to support secondary activities such as fulfillment and auditing. Mobile providers manage not only data related to placing voice calls, but also other electronic communications (i.e. text messages, email, video chat, etc.) plus consumer data passing through the network according to each device's data plan.

> *"IDC estimates the volume of digital data will grow 40% to 50% per year. By 2020, IDC predicts the number will have reached 40,000 EB [Exabyte], or 40 Zettabytes (ZB). The world's information is doubling every two years. By 2020 the world will generate 50 times the amount of information and 75 times the number of information containers"* [x]

The Big Data volume threshold for your organization will be driven by the amount of data required from your collective data stores needed to perform complex quantitative and analytical functions, the level of processing capability required to do so, and the pace at which the volume is growing. Keep in mind that large volumes of data are only one characteristic of Big Data and don't immediately require a data management paradigm shift from the current plans for capacity management.

Velocity

The next Big Data characteristic is Velocity, which deals with both how quickly data is generated or received, as well as how timely the assessment of the data needs to be. The distinctions between batch and real-time (or online) information processing have been with us for some time, but this goes well beyond traditional 'Transactions Per Second' measurements.

In some settings, incoming data has to be analyzed faster than it can be stored and subsequently retrieved for critical decision-making. For example, today's modern car typically has over 100 independent sensors, monitoring the vehicles performance, soundness and safety. Knowing that the brakes were good ten minutes ago when there's an emerging problem right now is not very helpful and could prove quite disastrous.

Now take that analogy forward to the work being done on Autonomous Cars, or robotic cars that do not require a human driver. The number of sensors goes up dramatically, plus we have the introduction of operational control units, sophisticated radar mechanisms, two-way communication streams to and from other devices in our immediate vicinity and the need to detect and safely react to the potentially random activities of animated life forms. All of this incoming data represents an incredibly complicated stream of data that has to be received, interpreted, decisioned and acted upon in real real time, not pseudo real time.

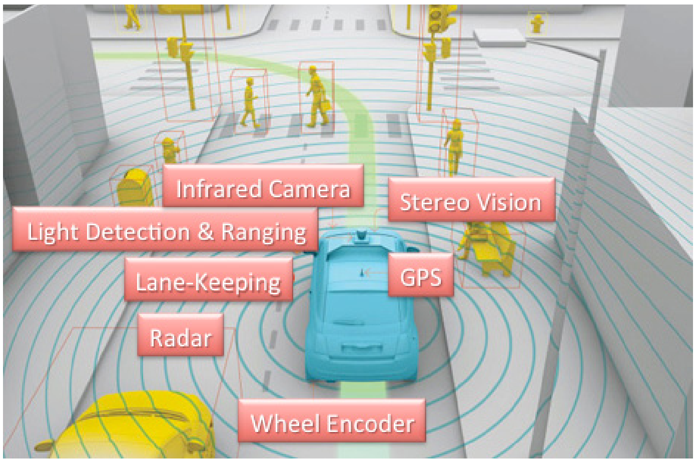


Figure 2 - Near-Future Autonomous Car Technology[xi]

Variety

The third characteristic of Big Data is Variety which refers to the ever-widening array of the types of data being processed and how they're sourced. Traditional transactional data, for example, typically adheres to a well-defined data model and corresponding database schema. Yet organizations are finding more and more value tucked away in less traditional sources of data, where name/value pairs are not so readily available. Big Data often seeks to integrate both structured and unstructured data from internal and external sources.

> *"Structured data gives names to each field in a database and defines the relationships between the fields. Unstructured data is usually not stored in a relational database (as traditionally defined) where the data model is relevant to the meaning of the data.*
>
> *The Internet of Things (equipping all objects in the world with identifying devices), blogs, videos, social media, emails, notes from call centers, and all forms of human and computer to computer communications will soon start to produce massive amounts of unstructured or semi-structured data."*[xii]

Unstructured Data	Internal Sources	*Email* *Call Center Transcripts* *Forums, Blogs*
	External Sources	*Social Networks* *Forums, Blogs* *Videos*
Structured Data	Internal Sources	*CRM* *Point of Sale* *Service Tickets*
	External Sources	*Industry Research Data* *Financial Market Data*

Enabling Technologies

Now turning our attention to technical solutions that are often applied to Big Data problems, we see a new suite of tools, both those based on vendor products and those based on open source community initiatives. These tools are designed to manage, process and manipulate large, fast-moving disparate data. For this discussion, we'll focus on MapReduce, NoSQL and Hadoop, which are three popular tools being used to address the challenges of Big Data on a broad scale.

MapReduce

MapReduce is a model for taking very large data sets, dividing and mapping those datasets into meaningful Key/Value pairs across parallel processing nodes, shuffling the data into common key groupings and then reducing the shuffled data into a manageable result set. [xiii] Programmers typically write MapReduce programs in Java, as well as other languages such as Python, Ruby and R.

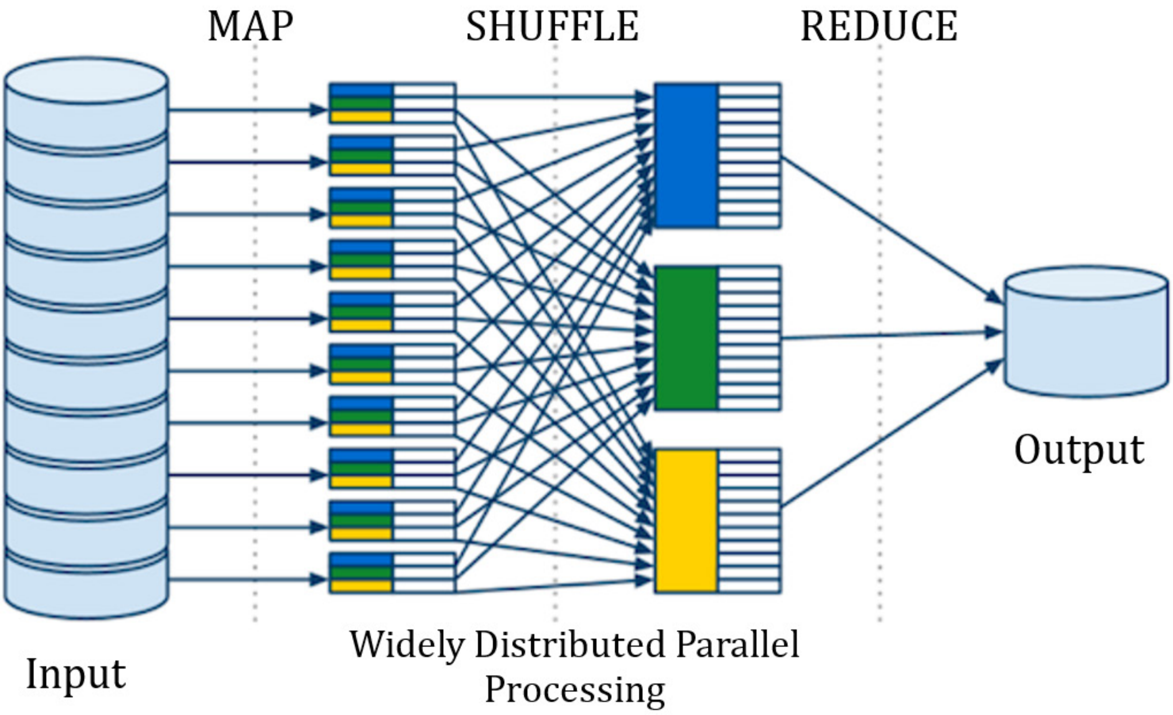


Figure 3 - MapReduce High-Level Process View[xiv]

MapReduce has been used successfully against Petabytes of input data, being mapped, shuffled and reduced across thousands of independent processing nodes. The nodes are intended to be commodity devices of nominal configuration, as compared to the dedicated data appliances we typically see in large, Enterprise Data Warehouse (EDW) configurations. It is also assumed that some number of the nodes will fail during the process, so interactions with the nodes are stateless – if a process hangs or doesn't finish in time, it is marshaled to another node until all of the processing tasks are complete. Rather than planning for traditional disaster recovery or business resumption, node failure is anticipated and processing continuity is built into the fundamental design of the environment.

NoSQL

Today, most modern organizations depend heavily on Codd's groundbreaking relational database concepts published back in the early 1970's, where information is structured and managed as a series of tables and relationships. [xv] Information is retrieved through the execution of Structured Query Language (SQL) commands against a well-defined database schema. Yet as we discussed previously, our data environs now include unstructured data from a variety of sources that must be handled, many of which are beyond our control and ill suited to traditional tabular views.

NoSQL, much like Big Data, lacks a precise definition. Instead, it is typically described as a series of attributes and characteristics that compare and contrast it to the more familiar relational database model. NoSQL, or ‘Not only SQL’, anticipates and allows for a certain lack of precision in the results, much to the uneasiness of the classically trained Data Base Analyst. In terms of a workable definition, TechTerms.com offers the following:

> *“NoSQL is a non-relational database that stores and accesses data using key-values. Instead of storing data in rows and columns like a traditional database, a NoSQL DBMS stores each item individually with a unique key. Additionally, a NoSQL database does not require a structured schema that defines each table and the related columns. This provides a much more flexible approach to storing data than a relational database.”* [xvi]

NoSQL databases generally fall into four categories: Key-Value Stores, Column Stores, Document Databases, and Graph Databases. Key-Value Stores use a hash table of key/pointer pairs to locate discrete data items. Column Stores uses key/column pairs, where the keys point to columns of data distributed across multiple servers. Document Databases are best thought of as versioned documents, each containing nested collections of key/value collections. Graph Databases move away from the concepts of tables, rows and columns, and instead support a data model based on graphing nodes across multiple environments. Monitis published a helpful visual that maps the four NoSQL categories and their relative scalability in terms of database size and complexity.

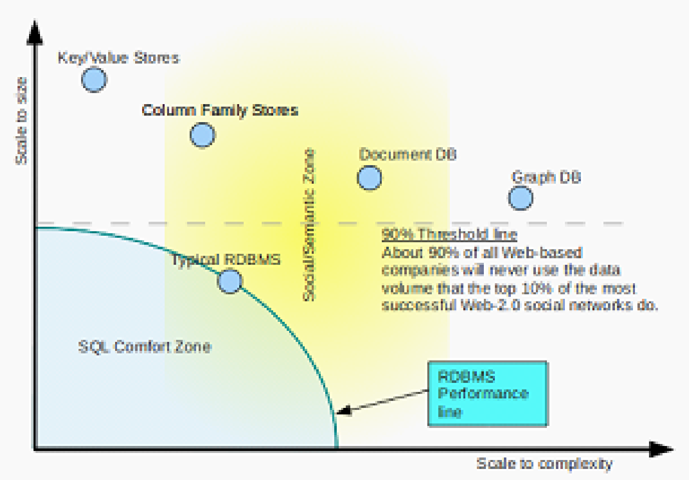


Figure 4 - Scalability of RDBMS and NoSQL Categories [xvii]

Advancements in NoSQL adaptations have led to additional database categories, including Multimodel Databases, Object Databases, Grid/Cloud Databases, XML Databases, Multidimensional Databases, Multivalue Databases, Event Sourcing and more. The NoSQL landscape is currently littered with some 150+ different database offerings.[xviii] Taking a deeper dive into which type of NoSQL database is the best choice for a particular situation is well beyond the scope of this document. In closing, we'll simply wrap up our discussion of NoSQL with the following table, which highlights key characteristics of RDBMS/SQL and NoSQL platforms.

RDBMS/SQL	NoSQL
• *Data Stored in Columns and Tables*	• *Large Data Volumes*
• *Relationships Represented by Data*	• *Scalable Replication and Distribution*
• *Data Manipulation Language*	• *Rapid Query Response Needed*
• *Data Definition Language*	• *Mostly Queries, Few Updates*
• *ACID Transactions (Atomic, Consistent, Isolated, Durable)*	• *Asynchronous Inserts/Updates*
• *Abstraction from Physical Layer*	• *Schema-less*
• *Pre-Defined Schema*	• *BASE Transactions (Basically Available, Soft state, Eventually consistent)*
• *Well-Defined Semantics*	• *CAP Theorem*
• *Standard Definitions*	• *Open Source Development*

Table 2 - Characteristics of SQL and NoSQL [xix]

Hadoop

At the time of this writing, one can hardly discuss Big Data without hearing about Hadoop. Hadoop is a popular open source framework implementation of MapReduce from the Apache Software Foundation, as well as a distributed file management system loosely based on the Google File System (GFS) specification. Hadoop was initially created in 2005, and after years of collaboration with the open source community, version 1.0.0 was released in 2011. Several leading technology vendors offer Apache Hadoop as part of their product lines, including IBM, Intel, Amazon, VMware and Hortonworks.

The Hadoop project is focused on four core modules considered to be the minimum components needed for Big Data processing. There are ten additional Apache projects underway that provide supplemental tools and utilities that further extend processing capabilities, such as environment operations, NoSQL database and data warehouse functions, and analytic tools. Self-identified organizations that are using Hadoop include well-known companies such as Adobe, Alibaba, eBay, Facebook, Google, Hulu, IBM, LinkedIn, Twitter, and Yahoo!, among many more.[xx]

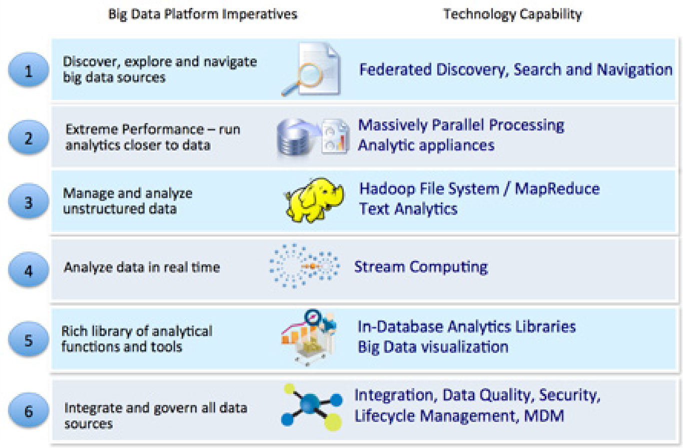
Core Hadoop Modules	
Hadoop Common The common utilities that support the other Hadoop modules.	**HDFS™** A distributed file system that provides high-throughput access to application data.
Hadoop Yarn A framework for job scheduling and cluster resource management.	**Hadoop MapReduce** A YARN-based system for parallel processing of large data sets.
Supporting Hadoop Modules/Projects	
Ambari™ A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.	**Hive™** A data warehouse infrastructure that provides data summarization and ad hoc querying.
Avro™ A data serialization system.	**Mahout™** A Scalable machine learning and data-mining library.
Cassandra™ A scalable multi-master database with no single point of failure..	**Pig™** A high-level data-flow language and execution framework for parallel computation.
Chukwa™ A data collection system for managing large distributed systems.	**Spark™** A fast and general compute engine for Hadoop data.
HBase™ A scalable, distributed database that supports structured data storage for large tables.	**ZooKeeper™** A high-performance coordination service for distributed applications.

Table 3 – Apache Foundation Hadoop Modules and Projects [xxi]

A quick review of the Hadoop modules reveals a comprehensive Big Data framework, which explains why many large organizations are making deep investments in the technology. For those that follow the Gartner Hype Cycles, Big Data is currently sitting near the top of the ‘Peak of Inflated Expectations’ (mass media hype, supplier proliferation, activity beyond early adopters) on their 2013 Emerging Technologies Hype Cycle. More specifically, Hadoop Distributions are in the early beginning of the slide into the ‘Trough of Disillusionment’ (supplier consolidation and failures, 2nd/3rd round venture capital, <5% adoption by potential audience) according to their 2013 Big Data Hype Cycle [xxii, xxiii] Only time will tell, but given the momentum and growing level of mainstream adoption, Hadoop seems well poised to make it up the ‘Slope of Enlightenment’ (2nd/3rd generation products, emerging methods/best practices) and onto the ‘Plateau of Productivity’ (high growth adoption begins) in terms of Big Data enabling technologies, indicating Hadoop’s potential for longer-term commercial survivability.

Forrester Research is also monitoring Hadoop from an enterprise solution offering perspective. In 2012 they identified Hadoop as “the nucleus of the next-generation EDW [Enterprise Data Warehouse] in the cloud”, but cautioned Big Data practitioners about the risks stemming from the relatively low level of maturity of enterprise-grade Hadoop offerings at the time.[xxiv] As Hadoop is getting well-entrenched into data centers around the globe, it is clear that it is not likely to be going away any time soon, but rather will continue to grow in terms of sophistication and adoption.

There are other technologies available in the marketplace to solve the various aspects of Big Data. In his Big Data Manifesto, Krishnan Parasuraman of IBM offered the following technology capability suggestions based on the contextual need of the problem statement or imperative. It is provided here as an example rather than as a prescriptive decision tree. The point is that your organization will likely be faced with more than one of these imperatives, ergo the solutions should not be considered to be mutually exclusive.



	Big Data Platform Imperatives	Technology Capability
1	Discover, explore and navigate big data sources	Federated Discovery, Search and Navigation
2	Extreme Performance – run analytics closer to data	Massively Parallel Processing Analytic appliances
3	Manage and analyze unstructured data	Hadoop File System / MapReduce Text Analytics
4	Analyze data in real time	Stream Computing
5	Rich library of analytical functions and tools	In-Database Analytics Libraries Big Data visualization
6	Integrate and govern all data sources	Integration, Data Quality, Security, Lifecycle Management, MDM

Figure 5 - Technology Capabilities for Big Data Imperatives [xxv]

The Changing Landscape of Analytics

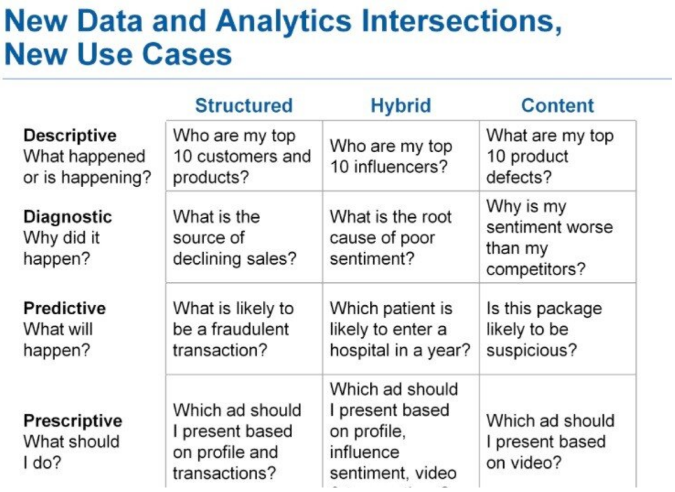
Although impressive from a design and capacity perspective, all of the aspects about Big Data we’ve discussed so far have not yet addressed a fundamental question: So what benefit do we get from all of this? The real potential value found within the massive amounts of data is the information we are able to derive from the use of Big Data, analyzing and deriving conclusions that were not otherwise discernable from traditional data viewpoints.

The field of Business Intelligence has brought us tremendous insight over the years, providing a means to analyze transactional trends and hypotheses about future activity, given a list of variables and correlations. Predictive analytics looks for important correlations between various data points from the past in order to recommend a 'next best' course of future action, whether it be prescribing a specific treatment protocol for a diagnosed disease or anticipating if a specific retail customer would have a monetary propensity to respond to a particular offer under a unique set of circumstances.

The challenge of analytics to date has been the focus on structured and semi-structured data analysis – results are partially limited in the conclusions that can be derived based on the nature of what is already known in its current form. When we consider the Big Data characteristic of Variety, we realize that there are new additional data points to be considered that may be potentially valuable to include in our assessment, given the amalgamation of internal and external unstructured data with traditional data sources. When we add the aspects of volume and velocity, we begin to appreciate how fast information needs to be processed, such as in the case of clickstream analysis, where user online behavior is being analyzed in real time from multiple data points to anticipate next-best actions. Applying analytics to Big Data may give us better insight than previously possible when trying to decipher the connection between events and behavior patterns than ever before.

Having Big Data does not guarantee we'll find the deep insights hoped for, but the adaptability and rapid cycle times allow for extensive 'what if' scenarios to be modeled in rapid succession. Depending on the nature of the organization and how quickly it can adjust its delivery channels, these models can be tested on sub-segments of the population in real time with real time analysis of the impact.

The following figure provides a few examples of the types of inquiries that can be made, driven by the structure of the data. Note that not all use cases require Big Data per se, as many of the hypothetical queries could be successfully executed through the use of traditional data management solutions.



New Data and Analytics Intersections, New Use Cases

	Structured	Hybrid	Content
Descriptive What happened or is happening?	Who are my top 10 customers and products?	Who are my top 10 influencers?	What are my top 10 product defects?
Diagnostic Why did it happen?	What is the source of declining sales?	What is the root cause of poor sentiment?	Why is my sentiment worse than my competitors?
Predictive What will happen?	What is likely to be a fraudulent transaction?	Which patient is likely to enter a hospital in a year?	Is this package likely to be suspicious?
Prescriptive What should I do?	Which ad should I present based on profile and transactions?	Which ad should I present based on profile, influence sentiment, video	Which ad should I present based on video?

Figure 6 - Comparison of Data Structure Use Cases [xxvi]

Pre-Big Data environments and data management tools are capable of handling the majority of an institution's current data needs, especially when dealing with large amounts of structured and semi-structured data. Dedicated data warehouse hardware appliances, Massively Parallel Processing (MPP) techniques and common SQL instructions are successfully processing massive amounts of data around the world every day. [xxvii]

The key is to understand the nature of both traditional data warehouse and Big Data analytic environments, their strengths and weaknesses, and when each is the appropriate solution. The lack of defined data schemas, absence of structured/formalized data access languages and the level of acceptable imprecision in both the data and analytical results are often difficult to accept for those steeped in earlier data management techniques. Not because the concepts are necessarily difficult to understand, as in many ways Big Data solutions are more flexible and less complicated in their approach.

The challenge is often due to what appears to be the contrary nature of Big Data when compared to a well-defined data management methodology with guiding principles such as data normalization, structural separation of logical/physical data constructs and general standardization. It should be comforting to know that Codd faced a lot of opposition to his 'radical' relational database model, so the pattern of resistance from the technical community in this space is not without history.

Traditional Data Warehouse	**Big Data Analytic Environment**
• *Complete Record from Transaction System*	• *Data from Many Sources Inside and Outside of Organization, Including Traditional DW*
• *All Data Centralized*	• *Data Often Physically Distributed*
• *Addition Every Month/Day of New Data*	• *Need to Iterate Solution to Test/Improve Models*
• *Analytics Designed Against Stable Environment*	• *Large-Memory Analytics also Part of Iteration*
• *Many Reports Run on a Production Basis*	• *Every Iteration Usually Requires*

Table 4 - Comparison of Traditional DW & Big Data Analytical Characteristics [xxviii]

A look at CIO Magazine's list of the top 8 emerging Big Data career opportunities reveals a nice cross-section of 'old' and 'new' roles: ETL Developers, Hadoop Developers, Visualization Tool Developers, Data Scientists, OLAP Developers, Data Warehouse Appliance Specialist, Predictive Analytics Developers, and Information Architects. [xxix] This is good new for current data analysts and engineers – there's clearly a need for a blended team with a mixture of older and newer data management skills. The key is to recognize the value of supporting both EDW and Big Data concurrently, rather than seeing them as somehow competing platforms.

Best EA Practices for Big Data

In some organizations, the excitement over Big Data, along with its positive expectations and interesting new tools and approaches has led them to usurp their typical Enterprise Architecture practices in order to move fast and avoid perceived delays due to following a rigorous assessment process. However, when we take into account that IDC anticipates the 2014 Big Data market to represent a 16B USD industry, with the average enterprise investing 8M USD in Big Data initiatives, it's clear that this level of spend needs to be taken seriously and aligned with the rest of the enterprise strategy.[xxx] Rather than struggle against Big Data, Enterprise Architects, Information Architects and Data Architects should find ways to adapt their current practices to accommodate its unique characteristics without violating the organization's fundamental precepts and principles.

Establish Adaptive Enterprise Data Principles

Big Data brings many new traits and technical challenges to the digital landscape, particularly with respect to an organization's over-arching Enterprise Data Principles. Many organizations have patterned their Enterprise Data Principles after those recommended in Section IV of TOGAF 9, suggesting that data is an asset, is shared, is accessible, has a trustee, has common vocabulary/data definitions, and should be secure. TOGAF further states that architectural principles should be understandable, robust, complete, consistent and stable.[xxxi] All of these points are consistent with sound data management and help set the guardrails for the organization when considering how data should be cared for, from acquisition through destruction.

The problem, however, is that Big Data clearly challenges several of these common principles, particularly due to its lack of traditional schema definitions, unstructured nature and obtainment from potentially unsecure external sources. Rather than assume that the well-conceived Enterprise Data Principles that are already in place are somehow irrelevant for Big Data and that Big Data should be given a 'pass' to excuse them from good data management, Enterprise Architects should work closely with their Information and Data Architects to revisit and refactor the existing principles to ensure that they address the full spectrum of data concerns. At some point, Big Data and 'regular' data are likely to intersect. Without clear principles to guide the management of both types of data, the enterprise runs the risk of having the more loosely managed data tainting or invalidating critical transactional data required to operate the organization within the constraining regulations pertinent to its industry.

Each organization will have to determine what the appropriate adaptation of their current Enterprise Data Principles should be and how much variation is tolerable. The key is to make sure that the principles reflect the existing needs of traditional data while accommodating new data requirements, as in the case of Big Data, plus whatever type of data might come after that. Don't assume this will be the last paradigm shift that will impact the data principle definitions – establish an adaptive, evergreen approach to promote flexible principles while maintaining the correct level of strategic guidance for the enterprise.

Don't Abandon the Enterprise Data Model or Data Governance

Given the nature of Big Data characteristics and rapid introduction of advanced tools, it can be easy to overlook the importance of aligning these new information flows with the organization's Enterprise Conceptual Data Model. Along those same lines, existing Enterprise Data Governance practices may not properly address the emerging needs of Big Data, either due to their lack of adaptability or responsiveness, or due to the lack of visibility of Big Data activities going on around the organization outside the purview of the architecture community.

The Enterprise Conceptual Data Model (ECDM) communicates critical information regarding key business concepts, typically represented as discrete data concepts or elements, along with primary relationships between them. The ECDM provides a valuable view as to how information flows through the organization, where it is produced, where it is consumed, and where the 'single source of truth' is, as well as stipulating enterprise-level data design patterns to be followed by solution delivery teams. Now enter the conundrum presented by Big Data, where data flows can be anything from free-form social media threads to several weeks' worth of surveillance video. Our initial reaction might be that the ECDM is not really relevant to this type of inbound data. However, we're likely analyzing these particular flows to provide additional insights about fundamental data concepts, such as customer interactions with particular products, or their propensity to respond to a particular marketing campaign. Data Scientists and Enterprise Data Architects need to work together to map and gap Big Data content to the ECDM, refactoring the model as needed.

Enterprise Data Governance provides important policies and procedures that define how data assets of the corporation are to be controlled and protected, including such aspects as data quality, data protection and data handling techniques. Once again, the nature of Big Data seems to be in direct opposition to the risk mitigation controls inherent in an effective Enterprise Data Governance policy. One approach it to totally isolate the Big Data environment from all other processing environments across the organization's network, placing it outside of the jurisdiction of Enterprise Architecture. This may provide a certain level of protection for a period of time, but the collision with the rest of the enterprise data landscape seems inevitable – eventually results from Big Data analytical operations must be paired up with other production data to support meaningful actions. Consider the potential level of financial risk based on decisions that could be made on imprecise results driven by Big Data analytics, such as a significant investment in a new product line without some level of governance or oversight around the data used to base that decision on. Hypothetical R&D lab experiments to prove out concepts and technical approaches are one thing; meeting with a regulatory auditor who wants to understand how proper data controls are used within the organization is quite another.

At some point, aspects of data governance will become clearly (and sometimes painfully) relevant. The organization must weigh the risks within the context of their operating constraints to determine what the right level of Enterprise Data Governance over Big Data should be. The goal is not to squash the advancements of Big Data within the enterprise, but rather to harness it and enable it with the appropriate level of oversight and risk mitigation. Data Architects and Data Engineers should work together to establish an appropriate level of governance that is congruent with the broader Enterprise Data Governance policy, while accommodating the unique aspects of Big Data to help its rapid progression across the organization.

Establish a Big Data Reference Architecture

Big Data, like many 'new' technologies, seems to find its way into organizations through many different paths, both formal and informal, which isn't terribly shocking. Companies are inundated with new technology offerings from vendors and requests from the technology and business communities alike. Most organizations have found ways to safely introduce new technology into the corporate ecosystem while applying some semblance of control and orchestration. Unless good portfolio management techniques are employed, however, it's not uncommon for multiple de facto standard tools to crop up with redundant capabilities, leaving an unintended number of options without establishing a clear go to solution. Many organizations have employed Reference Architectures to define strategic platforms, tools and implementation patterns, complete with roadmaps that lay out convergent adoption plans.

Given the size of the investments, as well as the wide array of data input sources and level of required network connectivity, it is imperative to establish a Big Data Reference Architecture. The organization's current reference architecture practice should be followed, including rigorous tool evaluation, selection of preferred alternative(s) and enablement of automated environment provisioning. In some cases, there may already be a prominent Big Data platform in use that simply needs to be ratified and 'productionalized' for broad consumption across the firm. In other cases, multiple solutions may be in varying states of usage in different parts of the organization, which can prove to be challenging, particularly for those organizations not currently on the 'winning' platform. The sooner the Big Data Reference Architecture can be established, the better – this enables quicker deployment of approved Big Data environments while reducing the level of variability in the environment.

Clarify Big Data Roles, Accountability and Decision Rights

The final Enterprise Architecture best practice for Big Data we'll discuss here is the recommendation to ensure that roles, accountabilities and decision rights have been clearly established and that everyone is operating under the same assumptions and expectations. While this practice may seem intuitive, informal discussions with industry peers and colleagues reveal that there is a consistent gap in this area. Rather than speculate as to why such a condition crops up, our focus is instead on how to rectify the situation and move forward.

In order to establish roles, the first step is to assemble a roster of current and prospective Big Data participants, both by individual name and the organizations they represent. A quick use case modeling exercise will reveal pertinent actors and personas currently performing the broader set of activities, which can then be grouped into more discrete generalized roles. These roles can then be compared to other data-centric disciplines such as EDW, BI and Data Analytics to leverage existing practices and promote consistency with prevailing patterns. It is important to gain consensus across the participating departments as to what the consistent Big Data roles are (or will be) and how/where those roles will be filled. Once the roles are known, levels of accountability can be established.

Many organizations go as far as defining roles and responsibilities in the context of a RACI diagram, identifying who is Responsible, Accountable, Consulted and Informed. However, application of the RACI model often stops short of clearly articulating corresponding decision rights. For example:

- Who gets to decide on tools?
- Who selects the commodity servers used for MapReduce operations?
- Who determines which external sources of unstructured data are acceptable?
- Who gets to map Big Data information to the rest of the data environment?

- Who approves deviation requests?
- Who determines the appropriate level of confidence required for various investment levels?

Even with the best of intentions, a plan that lacks clarity around decision authority is prone to conflict and unintended consequences. This is as true for the Big Data domain as it is for the rest of the organization. Enterprise Architects must be willing to slow the virtual Big Data team down long enough to help them get organized for longer term success. Each firm must determine for themselves the correct level of decision federation in a manner that is consistent to their current risk management approach, establishing what decisions can be made in the field (distributed) and what decisions need more singular execution (centralized).

Conclusion

By tracing Big Data's roots, exploring some of its core technologies and discussing relevant analytics, we've been able to establish a good baseline of knowledge. Based on the level of investment and industry attention garnered by Big Data at the time of this writing, it appears that this is more than a passing techno-trend. As the domain matures, organizations are able to choose from better solutions and are beginning to realize some of the potential value.

For those organizations with a disconnect between their Enterprise Architecture function and their current Big Data practices, there is an opportunity to bring the appropriate level of enablement and oversight. A certain amount of flexibility must be included to accommodate the unique needs of Big Data without compromising the broader needs of the organization as encapsulated in the current architecture practices.

For those Enterprise, Information and Data Architects that accept the challenge of engaging in and shaping Big Data head on within their organization, potential benefits they'll be driving include:

- Insightful and actionable analytics based on new and unique data sources
- Alignment to and consistency with the organization's data practices
- Repeatable investment criteria and deployment patterns across the Big Data spectrum
- Better internal collaboration across the Big Data communities
- Successful cohabitation of Big Data with existing data resources

Recommended Reading

A Very Short History of Big Data
Press (2013)

Big Data: A Revolution That Will Transform How We Live, Work and Think
Mayer-Schönberger and Cukier (2013)

Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses
Minelli, Chambers and Dhiraj (2013)

NoSQL Distilled: A Brief Guide to the emerging World of Polyglot Persistence
Sadalage and Fowler (2013)

Big Data Governance: An Emerging Imperative
Soares (2013)

References

[i] Google. (2014 February 8). Search "Big Data". Google.com. Retrieved February 8, 2014 from *http://www.google.com/search?q=Big+Data.*

[ii] Press, Gil. (2013 May 19). A Very Short History of Big Data. Forbes.com. Retrieved on February 9, 2014 from *http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/.*

[iii] Cox, Michael and David Ellsworth. (1997). "Application-Controlled Demand Paging for Out-of-Core Visualization." Proceedings of the 8th IEEE Visualization '97 Conference. P. 235.

[iv] Diebold, Francis X. "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting." Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society. 2000. PP. 115-122.

[v] Laney, Doug. (2001). Application Delivery Strategies: 3D Data Management: Controlling Data Volume, Velocity and Variety. February 6, 2001. Meta Group, Inc.

[vi] Gartner. (2014). IT Glossary > Big Data. Gartner.com. Retrieved February 9, 2014 from *http://www.gartner.com/it-glossary/big-data/.*

[vii] Gualtieri, Mike. (2012 December 5). Mike Gualtieri's Blog: The Pragmatic Definition of Big Data. Forrester.com. Retrieved February 9, 2014 from *http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data.*

[viii] Ohlhorst, Frank J. (2010 September 14). The Big Promise of Big Data: What You Need to Know Today. InfoWorld.com. Retrieved February 12, 2014 from *http://www.infoworld.com/d/business-intelligence/the-big-promise-big-data-what-you-need-know-today-585.*

[ix] Mayer-Schönberger, Viktor and Kenneth Cukier. (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think. P. 6. New York, NY, USA: Houghton Mifflin Harcourt Publishing Company.

[x] Walker, Michael. (2012 December 19). Structured Data vs. Unstructured Data: The Rise of Data Anarchy. Data Science Central. Retrieved February 15, 2014 from *http://www.datasciencecentral.com/profiles/blogs/structured-vs-unstructured-data-the-rise-of-data-anarchy*

[xi] Vanderbilt, Tom. (2012 January 20). Let the Robot Drive: The Autonomous Car of the Future is Here. Wired.com. Retrieved February 15, 2014 from *http://www.wired.com/magazine/2012/01/ff_autonomouscars/all/1.*

[xii] Walker.

[xiii] Dean, Jeffrey and Sanjay Ghemawat. (2004). "MapReduce: Simplified Data Processing on Large Clusters." OSDI'04: Sixth Symposium on Operating System Design and Implementation. December 2004.

[xiv] Google. (2013 November 19). MapReduce Python Overview. Developers.Google.com. Retrieved February 16, 2014 from *https://developers.google.com/appengine/docs/python/dataprocessing/.*

[xv] Codd, E. F. (1970). "A Relational Model of Data for Large Shared Data Banks." Communications of the ACM. Volume 13 (Issue 6). PP. 277-387.

[xvi] Techterms (2013 August 27). NoSQL. Techterms.com. Retrieved February 21, 2014 from *http://www.techterms.com/definition/nosql.*

[xvii] Vardanyan, Mikayel (2011 May 22). Picking the Right NoSQL Database Tool. Blog.Monitis.com. Retrieved on February 22, 2014 from *http://blog.monitis.com/index.php/2011/05/22/picking-the-right-nosql-database-tool/.*

[xviii] Edlich, Stefan. The Ultimate Reference for NoSQL Databases. NoSQL-Database.org. Retrieved on February 22, 2014 from *http://nosql-database.org/.*

[xix] Hare, Keith W. (2012 December 29). A Comparison of SQL and NoSQL Databases. Metadata Open Forum. Retrieved on February 21, 2014 from *http://www.slideshare.net/Muratakal/rdbms-vs-nosql-15797058.*

[xx] Apache (2014). Powered By. Apache.org. Retrieved on February 16, 2014 from *http://wiki.apache.org/hadoop/PoweredBy*.

[xxi] Apache Software Foundation. (2012). Welcome to Apache Hadoop. Hadoop.Apache.org. Retrieved February 15, 2014 from *http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F.*

[xxii] Rivera, Janesa and Rob van der Meulen (2013 August 19). Gartner's 2013 Hype Cycle for Emerging Technologies Maps Out Evolving Relationships Between Humans and Machines. Gartner.com. Retrieved on February 22, 2014 from *http://www.gartner.com/newsroom/id/2575515.*

[xxiii] Huedecker, Nick (2013 July 31). Hype Cycle for Big Data, 2013. Gartner.com. Retrieved on February 22, 2014 from *https://www.gartner.com/doc/2574616/hype-cycle-big-data.*

[xxiv] Kobielus, James (2012). The Forrester Wave™: Enterprise Hadoop Solutions, Q1 2012. P.2-3. Boston, MA, USA: Forrester Research, Inc.

[xxv] Parasuraman, Krishnan (2012 October 23). Part II: The Big Data Manifesto. IBMBigDataHub.com. Retrieved February 24, 2014 from *http://www.ibmbigdatahub.com/blog/part-ii-big-data-platform-manifesto.*

[xxvi] Elliot, Timo (2013 February 6). GartnerBI New Use Cases, Data and Analytics Intersection. Twitter.com. Retrieved February 21, 2014 from *https://twitter.com/timoelliott/status/299161346795241472/photo/1.*

[xxvii] Brust, Andrew. (2012 March 2). MapReduce and MPP: Two Sides of the Big Data Coin? ZDNet.com. Retrieved February 15, 2014 from *http://www.zdnet.com/blog/big-data/mapreduce-and-mpp-two-sides-of-the-big-data-coin/121.*

[xxviii] Vellante, David. (2014 January 26). Enterprise Big-data. Wikibon.org. Retrieved on February 17 from *http://wikibon.org/wiki/v/Enterprise_Big-data*.

[xxix] Hein, Richard (2014 January 15). The 8 Most In-Demand Big Data Roles. CIO.com. Retrieved on February 22, 2014 from *http://www.cio.com/slideshow/detail/135970#slide1.*

[xxx] Columbus, Louis (2014 January 12). 2014: The Year Big Data Adoption Goes Mainstream in the Enterprise. Forbes.com. Retrieved February 24, 2014 from http://www.forbes.com/sites/louiscolumbus/2014/01/12/2014-the-year-big-data-adoption-goes-mainstream-in-the-enterprise/.

[xxxi] TOGAF (2009). The Open Group Architecture Framework, Version 9. PP. 267-8, 273-277. The Open Group.

Orbus Software
3rd Floor
111 Buckingham Palace Road
London
SW1W 0SR
United Kingdom

+44 (0) 870 991 1851
enquiries@orbussoftware.com
www.orbussoftware.com