
# Deploying Big Data Analytics Applications to the Cloud: Roadmap for Success

May, 2014

# Contents

Acknowledgements ..... 3

Workgroup Leaders ..... 3

Key Contributors ..... 3

Reviewers ..... 3

Executive Overview ..... 4

Roadmap ..... 5

Step 1: Build the Business Case ..... 5

Step 2: Assess your Application Workloads ..... 8

Step 3: Develop a Technical Approach ..... 10

Step 4: Address Governance, Security, Privacy, Risk, and Accountability Requirements ..... 13

Step 5: Deploy, Integrate, and Operationalize the Infrastructure ..... 15

Works Cited ..... 16

Additional References ..... 16

Appendix A: Use Cases and Applications of Big Data in the Cloud ..... 17

Appendix B: Technical Requirements for Big Data in the Cloud ..... 18

Appendix C: Risk-Mitigation Considerations for Big Data in the Cloud ..... 20

Appendix D: Platform-choice Considerations for Big Data in the cloud ..... 20

# Acknowledgements

## Workgroup Leaders

James Kobielus, IBM; Bob Marcus, ET Strategies

## Key Contributors

James Kobielus, IBM

## Reviewers

Bob Marcus, ET Strategies; David Saul, State Street; Judith Hurwitz, Hurwitz Group; Thomas McCullough, LMCO; Jay Greenberg, Boeing; Michael Edwards, IBM

## Executive Overview

Big Data focuses on achieving deep business value from deployment of advanced analytics and trustworthy data at Internet scales.

Big Data is at the heart of many cloud services deployments. As private and public cloud deployments become more prevalent, it will be critical for end-user organizations to have a clear understanding of Big Data application requirements, tool capabilities, and best practices for implementation.

Big Data and cloud computing are complementary technological paradigms with a core focus on scalability, agility, and on-demand availability. Big Data is an approach for maximizing the linear scalability, deployment and execution flexibility, and cost-effectiveness of analytic data platforms. It relies on such underlying approaches as massively parallel processing, in-database execution, storage optimization, data virtualization, and mixed-workload management. Cloud computing complements Big Data by enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.

This paper's primary focus is to provide guidance to cloud service customer organizations in all industries on how best to deploy new and existing Big Data analytics applications to the cloud. The paper:

- Defines Big Data, analytic applications, and cloud deployment alternatives.
- Provides a description of the typical current Big Data analytics environment (i.e., what is the current progression to Big Data and/or cloud computing for the majority of enterprises and what are their primary issues).
- Provides a high-level description and decision criteria for the types of Big Data analytic applications that are appropriate to deploy to the cloud vs. the type of Big Data analytics applications which are not.
- Describes the principal use cases for Big Data in the cloud, including Big Data exploration, data warehouse augmentation, enhanced 360 degree view of the customer, operations analytics, security and intelligence analytics.
- Highlights the potential benefits of developing new and/or deploying existing Big Data analytics applications to the cloud.
- Articulates the challenges associated with deploying new and/or existing Big Data analytics applications to the cloud (challenges will vary depending on the type of Big Data analytics application which is being considered for deployment to the cloud).

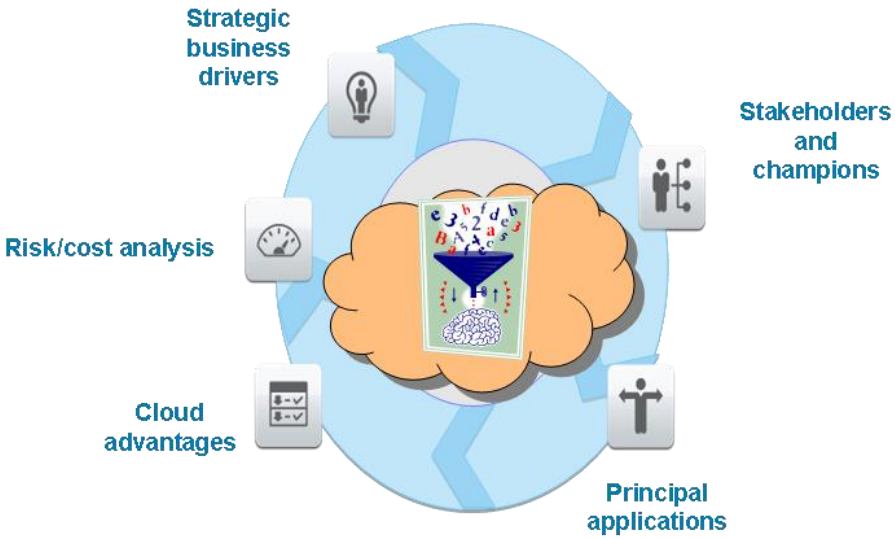
## Roadmap

This paper provides a prescriptive roadmap for justifying, deploying, and managing Big Data in the cloud. It highlights key considerations that users should follow in order to ensure successful deployment. Key steps discussed below are as follows:

- Step 1: Build the business case for Big Data in the cloud
- Step 2: Assess your Big Data application workloads for deployment into the cloud
- Step 3: Develop a technical approach for deploying and managing Big Data in the cloud
- Step 4: Address governance, security, privacy, risk, and accountability requirements
- Step 5: Deploy, integrate, and operationalize your cloud-based Big Data infrastructure

## Step 1: Build the Business Case

Building a business case for Big Data in the cloud involves the following key considerations:




Step 1: Build a business case

### Articulate strategic business drivers for Big Data in the cloud

Big Data focuses on achieving deep business value from deployment of advanced analytics and trustworthy data at Internet scales. The key business drivers for Big Data in business are the following:

- Extract greater value from enterprise data resources
- Enhance information worker and developer productivity

- Realize continued improvements in customer acquisition, loyalty, retention, upsell, and satisfaction
- Ensure more timely, agile response to business opportunities, threats, and challenges
- Manage a single view of diverse data resources
- Secure, protect, and govern data throughout its lifecycle
- Improve data/analytics platform scale, efficiency, and agility
- Optimize back-office operations

Refer to Appendix A for a thorough list of use cases and applications of Big Data in the cloud.

## Identify key stakeholders and champions for Big Data in the cloud

Every C-level business executive has strategic applications of Big Data in the cloud. The principal stakeholder/champions and their strategic imperatives are presented in Table 1:

| | |
|---|---|
| **Chief Marketing Officers** | ***CMOs*** have been the prime movers on many Big Data initiatives. Their primary applications consist of marketing campaign optimization, customer churn and loyalty, upsell and cross-sell analysis, targeted offers, behavioral targeting, social media monitoring, sentiment analysis, brand monitoring, influencer analysis, customer experience optimization, content optimization, and placement optimization |
| **Chief Information Officers** | ***CIOs*** use Big Data platforms for data discovery, data integration, business analytics, advanced analytics, exploratory data science. |
| **Chief Operations Officers** | ***COOs*** rely on Big Data for supply chain optimization, defect tracking, sensor monitoring, and smart grid, among other applications. |
| **Chief Information Security Officers** | ***CISOs*** run security incident and event management, anti-fraud detection, and other sensitive applications on Big Data. |
| **Chief Technology Officers** | ***CTOs*** do IT log analysis, event analytics, network analytics, and other systems monitoring, troubleshooting, and optimization applications on Big Data |
| **Chief Financial Officers** | ***CFOs*** run complex financial risk analysis and mitigation modeling exercises on Big Data platforms. |

***Table 1: Chief stakeholders of Big Data in the cloud***

Gaining support for your Big Data cloud business case requires that you align it with stakeholder requirements. If key business and IT stakeholders haven't clearly specified their requirements or expectations for your Big Data cloud initiative, it's not going to succeed. Make sure that you build a case that articulates business value (i.e. what are the likely or potential benefits) in business terms, such as:

- Enabling more agile business response to competitive opportunities and challenges
- Improvements in customer loyalty, engagement, satisfaction, and lifetime value
- Enhancements in operational efficiency, speed, and flexibility
- Tightened security and compliance

To gain business stakeholder support for your Big Data cloud initiative, your business case should:

- Define a business justification and return on investment for your Big Data cloud initiative
- Identify quantifiable success metrics for the project
- Measure and report project success metrics

### Emphasize the advantages of cloud-based deployment for Big Data

In building your business case for Big Data in the cloud and gaining the support of key stakeholders, you must first identify Big Data business applications where cloud approaches have an advantage that other approaches (such as software on commodity hardware or pre-integrated appliances) lack. The practical overlap between the Big Data and cloud paradigms is so extensive that you could validly claim you're doing cloud-based Big Data with an existing on-premise Hadoop, NoSQL, or enterprise data warehousing environment. Keep in mind that cloud computing includes "private cloud" deployments in addition to or in lieu of public cloud environments.

Cloud services have an important role to play in Big Data, especially for short-term jobs or applications where the bulk of data is already held in a cloud environment. Building your business case requires upfront analysis of the key value points in your business where cloud-based approaches are the right model for deploying Big Data analytics. At heart, the core principles defining what constitutes Big Data in the cloud are presented in Table 2:

| | |
|---|---|
| **On-demand self-service** | This enables a Big Data cloud service customer to self-provision cloud services--both physical and virtual--automatically and with minimal interaction involving the cloud service provider. |
| **Broad network access** | This enables Big Data cloud resources to be made available over the network and accessible through standard mechanisms by diverse client platforms. |
| **Multi-tenancy** | This enables Big Data cloud resources to be allocated so that multiple tenants and their computations and data can be guaranteed isolation from one another. |
| **Resource pooling** | This aggregates Big Data cloud resources in a location-independent fashion to serve multiple multi-tenant customers, enabling resources to be dynamically assigned and reassigned on demand and to be accessed through a simple abstraction. |
| **Rapid elasticity and scalability** | This enables Big Data cloud resources to be rapidly, elastically, and automatically scaled out up, out, and down on demand. |
| **Measured service** | This enables Big Data cloud resources to be monitored, controlled, reported, and billed transparently. |

***Table 2: Core architectural principles of cloud-based Big Data deployments***

The real questions for the architecture of Big Data and cloud computing are:

a) What is the data involved, how big is it and where is it located? One of the challenges of Big Data is that the data may be disparate and start out in a variety of locations. These locations may or may not be within a cloud service.

b) What sort of processing is going to be required on the data? Continuous or infrequent "burst" mode? How much can it be parallelized?

c) Is it practical to move the data to an environment where the processing can be performed efficiently, or is it better to think about moving the processing to the data? With the sorts of data volume involved – these are the key questions.

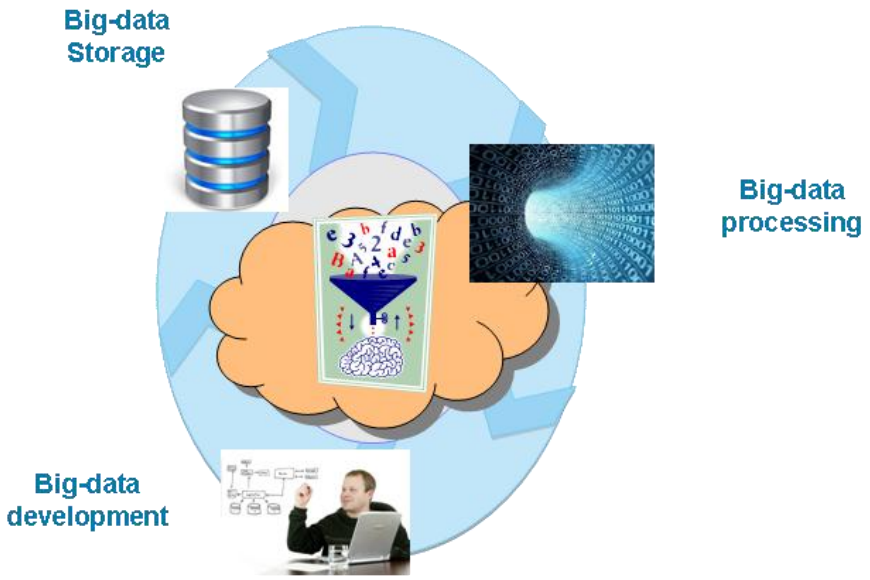
To gain IT stakeholder support for your Big Data cloud business case, you must address the following concerns:

- Technical and supportability requirements of Big Data in the cloud (refer to Appendix B for details)
- Deployment and service models of Big Data in the cloud (refer to relevant CSCC whitepapers [1], [2] for details)
- Risk-mitigation considerations for Big Data in the cloud (refer to Appendix C for details)

## Step 2: Assess your Application Workloads

Your business case must be predicated on your ability, under the proposed Big Data cloud deployment/service model and the proposed use case, to handle the expected application workloads. [1]

The principal functionality that any Big Data analytic application must support falls into the functional categories described in Table 3.




Step 2: Assess your application workloads

[1] For example, it's clear that Hadoop has already proven its core role in the emerging Big Data cloud ecosystem as a petabyte-scalable staging, transformation, pre-processing and refinery platform for unstructured content and embedded execution of advanced analytics.

| | |
|---|---|
| **Big-data storage** | A Big Data cloud must incorporate a highly scalable, efficient, low-cost data storage platform. Chief uses of the Big Data cloud storage platform are for archiving, governance and replication, as well as for discovering, acquiring, aggregating and governing multi-structured content. Typically, it would support these uses through integration with a high capacity storage area network architecture. |
| **Big-data processing** | A Big Data cloud should support massively parallel execution of advanced data processing, manipulation, analysis and access functions. It should support the full range of advanced analytics, as well as some functions traditionally associated with enterprise data warehouses, business intelligence, and online analytical processing. It should have all the metadata, models and other services needed to handle such core analytics functions as query, calculation, data loading and data integration. And it should handle a subset of these functions and interface through connectors to other data/analytic platforms. |
| **Big-data development** | A Big Data cloud should support Big Data application development which involves modeling, mining, exploration and analysis of deep data sets. The cloud should provide a scalable "sandbox" with tools that allow data scientists, predictive modelers and business analysts to interactively and collaboratively explore rich information sets. It should incorporate a high-performance analytic runtime platform where these teams can aggregate and prepare data sets, tweak segmentations and decision trees, and iterate through statistical models as they look for deep statistical patterns. It should furnish data scientists with massively parallel CPU, memory, storage and I/O capacity for tackling analytics workloads of growing complexity. And it should enable elastic scaling of sandboxes from traditional statistical analysis, data mining and predictive modeling, into new frontiers of Hadoop/MapReduce, R, geospatial, matrix manipulation, natural language processing, sentiment analysis and other resource-intensive types of Big Data processing. |

***Table 3: Principal Big Data workloads***

## Assess which application workloads are suitable for public vs. private clouds

A Big Data cloud should not be a stand-alone environment, but, instead, a standards-based interoperable environment that, when deployed in larger cloud configurations, can be rapidly optimized to new workloads as they come online. Many Big Data clouds will increasingly be configured to support mixes of two or all three of the functional requirements described above within specific cloud nodes or specific clusters. Some will handle low latency and batch jobs with equal agility. Others will be entirely specialized to a particular function that they perform with lightning speed and elastic scalability. The best Big Data analytic clouds will facilitate flexible re-optimization by streamlining the myriad deployment, configuration tuning tasks across larger, more complex deployments.

You may not be able to forecast with fine-grained precision the mix of workloads you'll need to run on your Big Data cloud in the future. But investing in the right family of Big Data cloud platforms, tools, and applications should give you confidence that, when the day comes, you'll have the foundation in place to provision resources rapidly and efficiently.

If you are considering public cloud services, you will need to identify which Big Data application workloads are best suited for public cloud deployment vs. your own private cloud deployment. Where public cloud deployment is concerned, some of the principal applications are presented in Table 4:

| | |
|---|---|
| **Enterprise applications already hosted in the cloud** | If, like many organizations -- especially small and midmarket businesses -- you use cloud-based applications from an external service provider, much of your source transactional data is already in a public cloud. If you have deep historical data in that cloud service, it might already have accumulated in Big Data magnitudes. To the extent the cloud service provider or one of its partners offers a value-added analytics service -- such as churn analysis, marketing optimization, or off-site backup and archiving of customer data -- it might make sense to leverage that rather than host it all in-house. |
| **High-volume external data sources that require considerable preprocessing** | If, for example, you're doing customer sentiment monitoring on aggregated feeds of social media data, you probably don't have the server, storage, or bandwidth capacity in-house to do it justice. That's a clear example of an application where you'd want to leverage the social media filtering service provided by a public-cloud-based, Big Data-powered service. |
| **Tactical applications beyond your on-premises, Big Data capabilities** | If you already have an on-premises Big Data platform dedicated for one application (such as a dedicated Hadoop cluster for high-volume extract-transform-load (ETL) on unstructured data sources), it might make sense to use a public cloud to address new applications (say, multichannel marketing, social media analytics, geospatial analytics, query-able archiving, elastic data-science sandboxing) for which the current platform is unsuited or for which an as-needed, on-demand service is more robust or cost effective. In fact, a public cloud offering might be the only feasible option if you need petabyte-scale, streaming, multi-structured, Big Data capability. |
| **Elastic provisioning of very large but short-lived analytic sandboxes** | If you have a short-turnaround, short-term data science project that requires an exploratory data mart (aka sandbox) that's an order-of-magnitude larger than the norm, the cloud may be your only feasible or affordable option. You can quickly spin up cloud-based storage and processing power for the duration of the project, and then rapidly de-provision it all when the project is over. One might call this the "bubble mart" deployment model, and it's tailor-made for the cloud. |

***Table 4: Principal applications of Big Data in public clouds***

## Step 3: Develop a Technical Approach

You need to perform a comprehensive technical assessment of existing and planned Big Data analytic applications and workloads to determine which can and should be deployed to the chosen cloud service. In conducting that assessment, you should identify all the Big Data, analytics, and cloud functional service layers, components, nodes, and platforms needed to support the planned

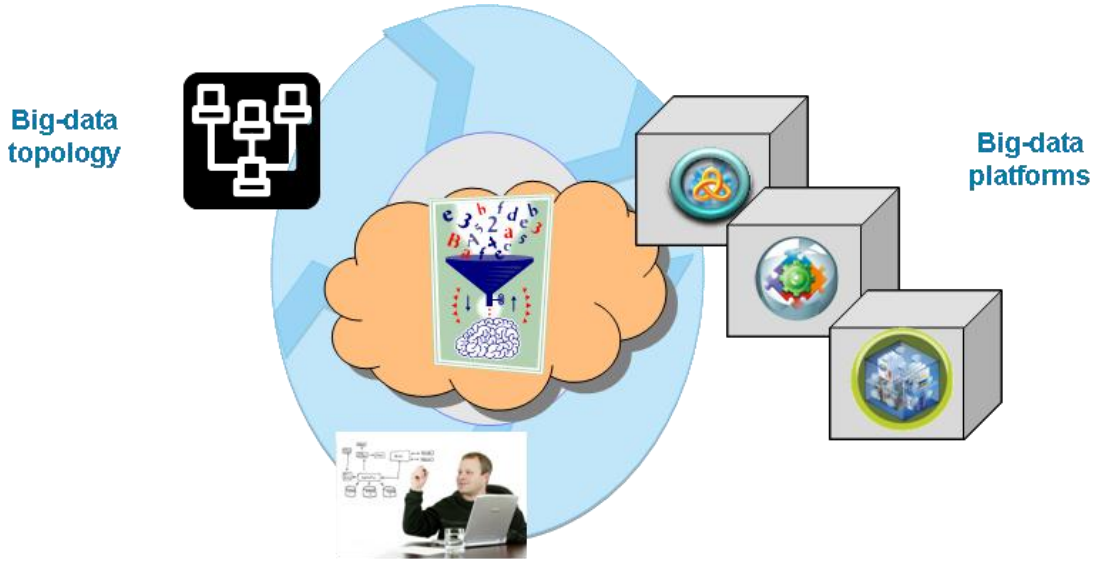
application/workloads. Just as important, you need to perform a thorough assessment of the various cloud deployment/service models to determine which are best suited to your workloads.

Migrating to the target ("to-be") Big Data cloud architecture--functional service layers and components, deployment roles, and topologies---requires that you incorporate the following tasks in your planning:

- Define the "as-is" and "to-be" Big Data analytics cloud architecture needed to support the application
- Assess the feasibility of modifying existing IT, data, and analytics investments for the to-be architecture, and/or reusing and migrating them to the planned application
- Determine the likely impact of internally hosted Big Data analytics applications on the IT infrastructure
- Calculate the additional costs that would be incurred to meet these demands
- Define the optimal phasing--staged vs. “all-at-once”--of deployment of Big Data analytics applications to the target cloud platform

When developing the technical approach and plan for your Big Data cloud initiative, pay close attention to the following critical areas:

- Topologies
- Data platforms




Step 3: Develop a technical approach

## Topologies

In defining your Big Data cloud "to-be" architecture, you will need to define the deployment topology of your environment. Big Data clouds are increasingly spanning federated private and public deployments,

encompassing at-rest and in-motion data, incorporating a growing footprint of in-memory and flash storage, and requiring on demand availability from all applications.

Consequently, your technical approach should address the distributed, multi-tier, heterogeneous, and agile nature of your Big Data environment. Optimization of your Big Data cloud topology should involve continual balancing of the topology to ensure that it has the optimal number, configuration, and size of clouds, tiers, clusters, and nodes. You will need to rebalance the topology by configuring it from end-to-end to support all anticipated workloads and meet all anticipated service level requirements. The sheer complexity, rate of change, variety of workloads, and welter of requirements on Big Data architectures can easily unbalance any fixed configuration in no time.

Virtualization is an important consideration in heterogeneous Big Data cloud topologies. It provides a unified interface to disparate resources that allows you to change, scale, and evolve the back end without disrupting interoperability with tools and applications. The degree to which you need virtualization depends in part on the complexity of your Big Data cloud requirements and technical environment, including the legacy investments with which it must all interoperate. Virtualization enables you to preserve and even expand, where necessary, the distributed, multi-tier, heterogeneous, and agile nature of your Big Data cloud environment. You can implement virtualization at the storage layer, the middleware layer, the query/access layer, the data layer, the management layer, or all of them.

One of the key enablers of Big Data cloud virtualization is the semantic abstraction layer, which enables simplified access to the disparate schemas of the RDBMS, Hadoop, NoSQL, columnar, and other data management platforms that constitute a logically unified data/analytic resource. As an integration architecture, virtualization ensures logically unified access, modeling, deployment, optimization, and management of Big Data as a heterogeneous resource.

### Data platforms

In building your Big Data cloud, you will need to determine the optimal data storage, management, and runtime platform(s) to support each deployment role, functional service layer, component, and workload, considering your target topology and service level requirements. You may require a Big Data cloud that includes different data platforms (e.g., RDBMS/row-based, Hadoop/HDFS, in-memory/columnar, etc.) fitted to different purposes but integrated into a unified architecture.

Beware of those who assert that there is only one type of cloud-oriented data platform that you should be considering for all of your requirements. Weigh your options by analyzing the different data platforms on the market, many of which have been optimized by vendors and solution providers for deployment in private, public, and other types of clouds. Rather than get mired in a "holy war" of one platform/approach vs. another, it's best to frame your platform options in terms of the principal considerations described in the corresponding table in Appendix D.

## Step 4: Address Governance, Security, Privacy, Risk, and Accountability Requirements

A Big Data cloud can be a complex, tricky thing to control as a unified business resource. The more complex and heterogeneous your Big Data cloud environment, the more difficult it will be to crack a tight whip of oversight and enforcement. Life-cycle controls, driven by explicit policies, can keep your Big Data environment from degenerating into an unmanageable mess and becoming a security and compliance vulnerability to boot.

But all things considered, Big Data cloud platforms may not be any more or less secure than established, smaller-scale databases. Where security is concerned, the sheer volume of your cloud-based Big Data is not the key factor to lose sleep over. Instead, the chief security vulnerabilities in your Big Data cloud strategy may have more to do with the unfamiliarity of the platforms and the need to harmonize disparate legacy data-security systems.

Platform heterogeneity is potentially a Big Data cloud security vulnerability. If you are simply scaling out an existing DBMS into a private cloud, your current security tools and practices will probably continue to work well. However, many Big Data deployments involve deploying a new platform–such as Hadoop, NoSQL, and in-memory databases in private or public clouds–that you have never used before and for which your existing security tools and practices are either useless or ill-suited. If the Big Data cloud offering is new to the market, you may have difficulty finding a sufficient range of commercial security tools, or, for that matter, anybody with experience using them in a demanding production setting.

In the process of amassing your Big Data repository from heterogeneous precursors, you will need to focus on the thorny issue of harmonizing disparate legacy security tools. These tools may support a wide range of security functions, including authentication, access control, encryption, intrusion detection and response, event logging and monitoring, and perimeter and application access control.

At the same time as you are consolidating into a Big Data platform, you will need to harmonize the disparate security policies and practices associated with the legacy platforms. Your Big Data consolidation plans must be overseen and vetted by security professionals at every step. And, ultimately, your consolidated Big Data platform must conform to all controlling enterprise, industry, and government mandates.

And if you want to approach Big Data security holistically, your strategy should also include tools and procedures for unified (hopefully real-time) monitoring of security events on disparate data sets. You should consider implementing a logical Big Data mart for security incident and event monitoring (SIEM) to identify threats across your disparate Big Data platforms, consolidated and otherwise.

You can control a Big Data cloud in a coherent manner, just as you oversee data and analytics investments at smaller scales. Doing so demands that you thoroughly assess your big-data infrastructure, tooling, practices, staffing, and skillsets in the following key areas:

| | |
|---|---|
| **Security controls** | These enforce policies of authentication, entitlement, encryption, time-stamping, non-repudiation, privacy, monitoring, auditing, and protection on all access, usage, manipulation, and other information management functions. |

| | |
|---|---|
| **Governance controls** | These enforce policies of discovery, profiling, definition, versioning, manipulation, cleansing, augmentation, promotion, usage, and compliance of official system-of-record data, metadata, and analytic models. |
| **Information lifecycle management (ILM) controls** | These enforce policies on the creation, updating, deleting, storage, retention, classification, tagging, distribution, discovery, utilization, preservation, purging, and archiving of information from cradle to grave. |

If you are an established data professional, this is probably all motherhood and apple pie to you. What you want to know is whether Big Data in the cloud has substantially altered best practices in any of these areas. In fact, it has. Enterprise-level controls – implemented through tools, infrastructure, and practices – are useless if they don't keep pace with the scale, complexity, and usage patterns of the resource being managed.

Where security, governance, and integrated lifecycle management (ILM) controls are concerned, Big Data innovations in several areas are raising the stakes and changing best practices in security, governance, and other key controls. You should consider the following trends as you establish your own policies, procedures, and safeguards for Big Data in the cloud:

- ***Adapt your controls to fit your new platforms for Big Data in the cloud:*** You may have mature security, governance, and ILM in your established enterprise data warehouse (DW), online transaction processing (OLTP), and other databases. But Big Data in the cloud has probably brought a menagerie of new platforms into your computing environment, including Hadoop, NoSQL, in-memory, and graph databases. The chance that your existing tools work out of the box with all of these new platforms is slim, and, to the extent that you are doing Big Data in a public cloud, you may be required to use whatever security, governance, and ILM features – strong, weak, or middling – that are native to that environment. You might be in luck if your Big Data platform vendor is also your incumbent DW/OLTP DBMS provider and if, either by itself or in conjunction with its independent software vendor partners, it has upgraded your security and other tools to work seamlessly with both old and new platforms. But don't count your proverbial chickens on that matter. In the process of amassing your end-to-end Big Data environment, you will also need to sweat over the nitty-gritty of integrating the disparate legacy security tools, policies, and practices associated with each new platform you adopt.
- ***Implement controls that address the new data domains you are managing in the cloud:*** Many of your traditional RDBMS-based data platforms are where you manage official system-of-record data on customers, finances, transactions, and other domains that demand stringent security, governance, and ILM. But these system-of-record data domains may have very little presence on your newer cloud-based Big Data platforms, many of which focus instead on handling unstructured data from social, event, sensor, clickstream, geospatial, and other new sources. The key difference from "small data" is that many of the newer data domains are disposable to varying degrees and are not linked directly to any "single version of the truth" records. Instead, your data scientists and machine-learning algorithms typically distill the unstructured feeds for patterns and subsequently discard the acquired source data (which quickly become too voluminous to retain cost-effectively anyway). Consequently, you probably won't need to apply much if any security, governance, and ILM to many new unstructured Big Data domains.
- ***Implement controls that operate at the new scales associated with Big Data in the cloud:*** Where life-cycle controls are concerned, Big Data's "bigness" is not necessarily a vulnerability.

There is no inherent trade-off between bigness – the volume, velocity, or variety of the data set – and your ability to monitor, manage, and control it. To the extent you have scaled out your Big Data initiative on a cloud-based platform with massively parallel processing (MPP), you can leverage the security, governance, and ILM tooling for that platform as you zoom into "3 Vs" territory. You might even find Big Data in the cloud to be a boon, facilitating some control-relevant workloads – such as real-time encryption, cleansing, and classification – that proved less feasible on "small-data" on-premises platforms. And in fact, the extreme scalability and agility of Big Data in the cloud may prove essential for enforcing tight security, governance, and ILM on new unstructured data types, as these needs emerge.

- ***Ensure that your controls respond to new mandates relevant to Big Data in the cloud:*** Big Data has become a culturally sensitive topic on many fronts, most notably among privacy watchdogs who point to its potential for use in intrusive target marketing and other abuses. As storage prices continue to plummet and more data is archived in Big Data clouds, the more likely those huge data stores are to be accessed for compliance, surveillance, e-discovery, intrusion detection, anti-fraud, and other applications dictated under new legal and regulatory mandates in many countries. Besides, even if no new Big Data-specific mandates hit the books, litigators will seek to subpoena these massive archives first when they're looking for evidence of corporate malfeasance. Consequently, in most industries--regulated and otherwise--Big Data archives will increasingly be managed under stringent mandates for security, governance, and ILM. This will ensure that the historical record is preserved in perpetuity, free from tampering and unauthorized access.

## Step 5: Deploy, Integrate, and Operationalize the Infrastructure

Deploying, integrating, and operationalizing your Big Data cloud environment are essential to realize the full value of this business resource over its useful life.

The criteria of Big Data cloud production-readiness must conform to what stakeholders require, and that depends greatly on the use cases and applications they have in mind for the Big Data cloud. Service-level agreements (SLAs) vary widely for Big Data deployed as an enterprise data warehouse, as opposed to an exploratory data-science sandbox, an unstructured information transformation tier, a queryable archive, or some other use. SLAs for performance, availability, security, governance, compliance, monitoring, auditing and so forth will depend on the particulars of each Big Data application, and on how your enterprise prioritizes them by criticality.

You should re-engineer your data management and analytics IT processes for seamless support in your Big Data cloud initiatives. If you can't provide trouble response, user training and other support functions in an efficient, reliable fashion that is consistent with existing operations, your Big Data platform is not production-ready. Key considerations in this respect:

- Implement continuous deployment of Big Data cloud applications, because traditional application development cycles of 6 months or more are no longer adequate;
- Facilitate improved communication, collaboration, and governance of disparate Big Data analytics development and operations organizations around your cloud initiative;

- Integrate siloed data, analytic, management, process, and operational support system infrastructures, roles, and workflows to enable deployment of a converged Big Data cloud resource;
- Provide Big Data cloud users with a "single throat to choke" for support, service and maintenance;
- Offer consulting support to users for planning, deployment, integration, optimization, customization and management of their specific Big Data cloud initiatives;
- Deliver 24x7 support with quick-turnaround on-site response on issues;
- Manage your end-to-end Big Data cloud environment with a unified system and solution management consoles; and
- Automate Big Data support functions to the maximum extent feasible.

To the extent that your enterprise already has a mature enterprise data warehousing program in production, you should use that as the template for "productionizing" your Big Data cloud deployment. There is absolutely no need to redefine "productionizing" for the sake of Big Data in the cloud.

## Works Cited

[1] Cloud Standards Customer Council. *Practical Guide to Cloud Computing Version 2.0,* April, 2014. http://www.cloudstandardscustomercouncil.org/webPG-download.htm

[2] Cloud Standards Customer Council. *Convergence of Social, Mobile and Cloud: 7 Steps to Ensure Success.* June 2013. http://www.cloudstandardscustomercouncil.org/Convergence_of_Cloud_Social%20_Mobile_Final.pdf

[3] NIST Big Data Working Group
http://bigdatawg.nist.gov/home.php

## Additional References

Report McKinsey Global Institute. *Big data: The next frontier for innovation, competition, and productivity.*
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

IBM Institute for Business Value report. *Analytics: The real-world use of big data*
https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=csuite-NA&S_PKG=Q412IBVBigData

Report to the President. *Big Data and Privacy: A Technological Perspective*
http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

## Appendix A: Use Cases and Applications of Big Data in the Cloud

Addressing the strategic imperatives of one or more business stakeholder/champions, your business case should specify the business case(s) for your proposed deployment of big data in the cloud. Some of the principal use cases for big data in the cloud include:

The following are the principal use cases and applications of Big Data in the cloud.

| | |
|---|---|
| **Big Data exploration** | Any analytic application that requires interactive statistical and visual exploration of very large, longitudinal, complex data sets. |
| **Customer experience optimization** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time customer experience data from portals, mobile devices, and other devices, applications, channels, and other touch points |
| **Data warehouse augmentation** | Any analytic application that requires a massively scalable tier to support discovery, acquisition, transformation, staging, cleansing, enhancement, and preparation of multi-structured data and subsequent loading into a downstream data warehouse, hub, or mart. |
| **Enhanced 360 degree view of the customer** | Any analytic application that requires a high-volume, scalable repository of current customer profile, sentiment, attitudinal, social, transactional, survey, and other data to supplement existing customer system of reference data. |
| **Financial risk analysis** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time data from internal and/or external sources on accounting, general ledger, budgets, and other financial sources |
| **Geospatial analytics** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time geospatial information. |
| **Influencer analysis** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time data from social media and other sources on influencers, net promoters, and brand ambassadors among customers and prospects |
| **Machine data analytics** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time data sourced from equipment, systems, devices, sensors, actuators, and other physical infrastructural components |
| **Marketing campaign optimization** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time data on the success of marketing campaigns across one or more channels or touch points |

| | |
|---|---|
| **Operations analytics** | Any analytic application that requires deep historical and/or real-time analysis of network, system, application, device, process, sensor, and other data sourced and logged from operational infrastructure. |
| **Security and intelligence analytics** | Any analytic application that requires deep historical and/or real-time analysis of security, fraud, and other event data intelligence sourced and logged from internal and/or external systems. |
| **Sensor analytics** | Any analytic application that requires automated sensors to take measurements and feed them back to centralized points at which the data are aggregated, correlated and analyzed. |
| **Sentiment analysis** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time data on customer sentiment from one or more social media or other source |
| **Social media analytics** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time data from one or more social media |
| **Supply chain optimization** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling, and visualization of historical and/or real-time data on production, equipment, materials, inventories, logistics, transport, transactions, and other supply-chain entities |
| **Targeted next best offers** | Any analytic application that requires acquisition, aggregation, correlation, processing, analysis, modeling and visualization of historical and/or real-time data required to power analytic models, business rules, and other artifacts that tailor offers made to customer on inbound and/or outbound marketing, customer service, portal, and other touch points |
| **Unstructured analytics** | Any analytic application that analyzes a deep store of data sourced from enterprise content management systems, social media, text, blogs, log data, sensor data, event data, RFID data, imaging, video, speech, geospatial and more. |

## Appendix B: Technical Requirements for Big Data in the Cloud

One key useful approach for assessing the appropriateness of cloud-based deployment for your big-data application is to describe the extent to which it is uniquely suited to handling the requisite technical and supportability requirements in these areas:

| | |
|---|---|
| **Scalability and acceleration** | Will the Big Data cloud environment support scale-out, shared-nothing massively parallel processing, storage optimization, dynamic query optimization, and mixed workload management better than alternative deployment models (e.g., on-premises appliances, software on commodity hardware)? |

| | |
|---|---|
| **Agility and elasticity** | Will the Big Data cloud environment better support on-demand provisioning, scaling, optimization, and execution of diverse data and analytic resources than alternative deployment models? |
| **Affordability and manageability** | Will the Big Data cloud environment enable lower total cost of ownership and better IT staff productivity than alternative deployment models? |
| **Availability, reliability, and consistency** | Will the Big Data cloud enable better availability, reliability, and consistency across distributed data/analytic resources that alternative deployment models? |
| **Service levels** | Will the Big Data cloud service support measurable metrics and thresholds of quality of service as agreed between the cloud service user and provider? |
| **Security** | Will it support physical, application, and data security, including such capabilities as authentication, authorization, availability, confidentiality, identity management, integrity, audit, security monitoring, incident response, and security policy management? |
| **Privacy** | Will it protect the assured, proper, and consistent collection, processing, communication, use and disposition of personally identifiable information in the relation to cloud services? |
| **Availability** | Will it meet the guaranteed uptime requirements specified in a service level agreement? |
| **Reliability** | Will it support an acceptable level of service in the face of faults (unintentional, intentional, or naturally caused) affecting normal operation? |
| **Performance** | Will it meet the latency, response time, concurrency, and throughput metrics and thresholds specified in a service level agreement? |
| **Auditability** | Will it support logging and auditing of all quality of service metrics as specified in a service level agreement |
| **Compliance** | Will it support continued, measurable, and verifiable compliance with all the relevant government, industry, and company mandates relevant to any and all aspects of operations? |
| **Interoperability** | Will it support user interaction with the service and exchange of information according to a prescribed method to obtain predictable results? |
| **Portability** | Will it allow users to move their data or their applications between multiple cloud service providers at low cost and with minimal disruption? |
| **Reversibility** | Will it allow users to retrieve their data and their application artifacts and to have strong assurance that the cloud service provider will delete all copies and not retain any materials belonging to the cloud service customer after an agreed period? |

## Appendix C: Risk-Mitigation Considerations for Big Data in the Cloud

The business case for any deployment or service model for cloud-based big data often hinges on a risk analysis that considers the following factors:

| | |
|---|---|
| **Scaling risks** | How likely is it that a particular approach will allow the solution to scale elastically in volume, velocity, and variety as Big Data analytics requirements evolve? |
| **Performance risks** | How likely is it that a particular approach will enable the solution to achieve our performance objectives for all workloads, current and anticipated, as requirements evolve? |
| **Total cost of ownership risks** | How likely is it that a particular approach will reduce the cost of deploying and managing Big Data analytics and maximize the productivity and efficiency of IT operations over the deployment's expected useful life? |
| **Availability and reliability risks** | How likely is it that this particular approach will enable us to meet our requirements for 24x7 high availability and reliability in the provisioned Big Data analytics service? |
| **Security, compliance, and governance** | How likely is it that a particular approach will meet requirements for security, compliance, and governance in Big Data analytics? |
| **Manageability** | How likely is it that a particular approach will meet requirements for administration, monitoring, and optimization of the Big Data platform/service over its expected useful life? |

## Appendix D: Platform-choice Considerations for Big Data in the cloud

These are the principal platform-choice considerations for Big Data in the Cloud:

| | |
|---|---|
| **Data architecture** | In this era of constant database innovation, the range of established and new data architectures and platforms continues to expand. The chief architectural categories include architectures that address varying degrees of "data at rest" and "data in motion." The principal market categories (which have diverse commercial and/or open-source offerings) include relational/row-based, columnar, file system, document, in-memory, graph, and key-value. A growing range of commercial and open-source offerings "hybridize" these approaches to varying degrees, and increasing numbers of them are being implemented within various public, private, and other clouds. Note the term "Hadoop" does not refer to a specific data storage architecture but instead to a range of architectures in different categories that can, to varying degrees, execute Apache MapReduce models. The same applies to the sprawling menagerie of data platforms lumped under the terms "NoSQL" and "NewSQL," which (depending on who you talk to) may include innovative databases in any of these categories, or in a narrow group of them. |

| | |
|---|---|
| **Application functionality** | The functionality of commercial and open-source data platforms and associated tooling varies widely. Every data platform has its own specific capabilities, strengths, and limitations in such key areas as query languages, data types supported, schemas and models, scalability, performance, and so forth. Frequently, you'll identify several cloud-based or cloud-enabling commercial and open-source platforms in different segments that overlap significantly in many of these features. |
| **Deployment roles** | The optimal data platform will vary depending on the cloud-based deployment role that you envision for a specific database. The range of potential cloud deployment database roles is wide, including, at the very least, specialized nodes for data acquisition, collection, preparation, cleansing, transformation, staging, warehousing, access, and other functions. For example, if you're implementing a node for acquisition, collection, and preprocessing of unstructured data at rest, Hadoop on HDFS is well-suited, or, depending on the nature of unstructured data and what you plan to do with it, various other file, document, graph, key-value, object, XML, tuple store, triple store, or multi-value databases may be more suitable. Or if you need strong transactionality and consistency on structured data in a governance hub, relational/row-based databases or so-called "NewSQL" databases may be worth looking into. Or, if real-time data-in-motion requirements predominate for data integration and delivery, you should investigate the stream-computing, in-memory and other platforms optimized for low latency. |
| **Integration requirements** | The optimal data platform(s) must integrate well with your chosen cloud-computing execution and development platforms, whether they are open source (e.g., OpenStack, Cloud Foundry) or commercial. To the extent that your data platform(s) also conform to open industry cloud-computing reference frameworks developed by the U.S. National Institute for Standards and Technologies and other groups, all the better. Ideally, the chosen Big Data and cloud-computing platforms should integrate well with your legacy or envisioned investments in analytics development and modeling tools; data discovery, acquisition, transport, integration, and so forth. |