THE DEFINITIVE GUIDE TO THE DATA LAKE

Author: John O'Brien, CEO, Radiant Advisors
Editor: Lindy Ryan, Research Director, Radiant Advisors

UNISPHERE RESEARCH
database TRENDS AND APPLICATIONS
Radiant Advisors

THE DEFINITIVE GUIDE TO THE DATA LAKE

TABLE OF CONTENTS

Introduction **1**

Data Lake Definitions and Perspectives **2**

Clarifying the Data Lake's Organization **4**

The Data Lake Unifies Data Discovery, Data Science, and Enterprise BI **6**

Critical Data Lake Success Factors **8**

Conclusion **11**

Sponsors **12**

The Definitive Guide to the Data Lake

Author: John O'Brien, CEO, Radiant Advisors
Editor: Lindy Ryan, Research Director, Radiant Advisors

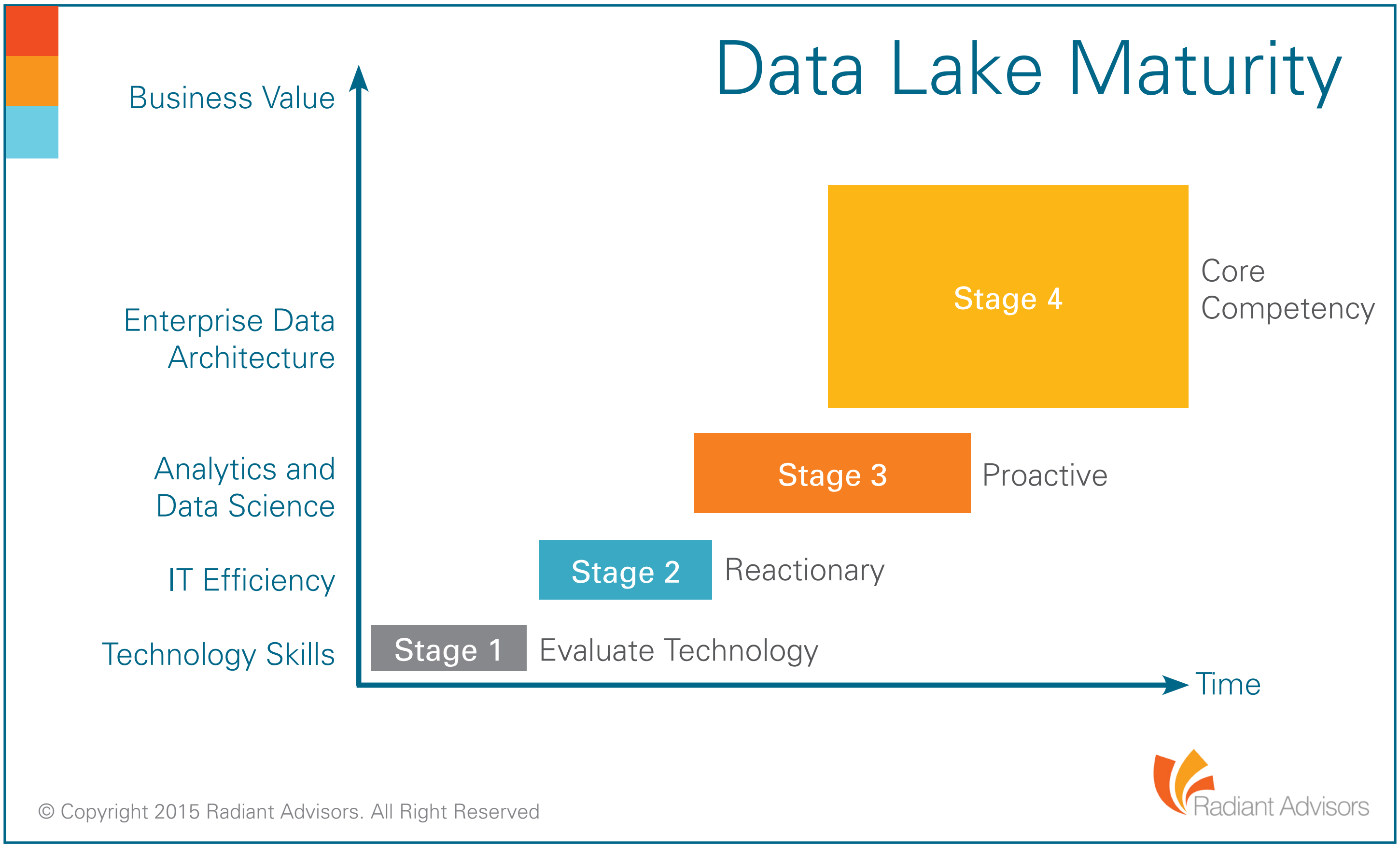
It would be an understatement to say that the hype surrounding the data lake is causing confusion in the industry. Perhaps, this is an inherent consequence of the data industry's need for buzzwords: it's not uncommon for a term to rise to popularity long before there is clear definition and repeatable business value. We have seen this phenomena many times when concepts including "big data," "data reservoir," and even the "data warehouse" first emerged in the industry. Today's newcomer to the data world vernacular—the "data lake"—is a term that has endured both the scrutiny of pundits who harp on the risk of digging a data swamp and, likewise, the vision of those who see the potential of the concept to have a profound impact on enterprise data architecture. As the data lake term begins to come off its hype cycle and face the pressures of pragmatic IT and business stakeholders, the demand for clear data lake definitions, use cases, and best practices continues to grow.

This paper aims to clarify the data lake concept by combining fundamental data and information management principles with the experiences of existing implementations to explain how current data architectures will transform into a modern data architecture. The modern data architecture includes Hadoop and its surrounding ecosystem, integrated alongside the data warehouse, discovery-oriented environments, and highly-specialized analytic or operational data technologies. Therefore, the data lake has become the metaphor for the transformation of enterprise data management, and will continue to evolve the data lake definition according to established principles, drivers, and best practices that will quickly emerge as hindsight is applied at companies.

Data Lake Definitions and Perspectives

As the data lake concept becomes a part of the core data architecture, the question is often whether the data lake is an architectural strategy or an architectural destination. The answer is both. By defining the data lake as a centralized repository of enterprise data for various data workloads (within and a part of), we address both the end state architecture as well as establish a guiding light for data architecture-related decisions on the journey to achieving critical mass for the data lake. Many companies wisely apply the data lake concept to the Ralph Waldo Emerson adage that "life is a journey, not a destination," and recognize that the data lake is the ultimate destination of a unique journey that follows its own pace and direction based on the drivers and priorities of the adopting company. Based on the architectural adoption patterns that we see happening in companies today, we can distill the journey to the data lake into four stages of business adoption.

Stage 1 comprises "kick-the-tire" big data pilot projects that focus on particular business outcomes that serve as an introduction to using and managing an Apache Hadoop environment. Stage 2 is a more reactionary approach as companies begin to concentrate on leveraging Hadoop's strong points to tackle existing architecture inefficiencies and isolate clear, quick, and measurable business value opportunities. One such pattern is to leverage Hadoop's scalable, low-cost persistence layer or its ability to perform big data processing and analytics. An example is the recent acceptance of relocating historical data warehouse data into the



data lake as an online storage extension with lower service-level agreements to reduce the overall size and management of the data warehouse. Other deployments can provide potentially long-term value, such as relocating a data warehouse staging area of operational source data acquisitions into the data lake so that data warehouse integrations can source and persist acquired data as long-term raw data (and additional data not used by the data warehouse) to enable the affordability and data processing power needed by data scientists and analytics development.

In Stage 3, organizations move from a reactionary approach to a proactive one. This stage includes initiatives to further consolidate data for big data and analytics projects and utilize Hadoop's affordable scalability to enable vast amounts of social networking platforms, emerging data in the Internet of Things, and unstructured data in a single, centralized repository that focuses on and encourages reuse while avoiding data duplication. Last, Stage 4 is that of continuous optimization. Inevitably, there comes a tipping point when a data lake becomes a core part of IT's strategy and planning, and walls between operational applications and analytic application silos are rethought as a single enterprise platform that leverages the strengths of each architecture component. Consider Stage 4 to be the level at which the journey to the data lake has Hadoop fulfilling a foundational component of the enterprise data architecture strategy, and supporting more of the operational, analytic, and big data workloads with both persistence and data engine layers.

Ultimately, the data lake destination is about building an efficient enterprise data architecture with a single repository of enterprise data that can meet the needs of various enterprise application workloads that gain efficiency through the reusability and consistency of data. The data lake will continue to grow over the maturity stages through consolidation (to avoid having the multiple Hadoop cluster silos typically found in Stage 1), and strive for a critical mass momentum. However, the rate at which any company can accelerate data lake adoption will be tempered by the establishment of data governance, security, roles, and access.

The benefits of having an efficient modern data architecture must align with the risks associated. Therefore, the data lake must meet each one of these data management rigors if it is destined to be a part of a company's core data architecture. The data lake strategy should extend data governance to include big data, data discovery, and data science use cases and roles. Driving the business use case, Hadoop as a discovery platform requires access to new and existing data sources. However, data discovery via unfettered access to all data is difficult for data governance programs to accept. Therefore, trusted and reliable security and monitoring approach requires a relationship between privileges, roles, and data groups that follow the rules of data owners and accountability. And, enterprises need to recognize that their own principle-based approaches to data governance and security will require more work to maintain until widespread enterprise adoption begins to establish best practices and successful case studies. This is where data lake leaders will create best practices for data lake followers as the technology continues to rapidly evolve to address enterprise needs.

Clarifying the Data Lake's Organization

The flexibility and scalability of Hadoop powers the data lake concept to be the next-generation, centralized data repository that can be leveraged by the enterprise. As a persistence layer in data architecture, all forms of data can be maintained, cataloged, explored, and utilized together through a common access layer of Hadoop's data engines powered by Hadoop's interactive and multi-tenancy Yet Another Resource Negotiator (or, YARN). What separates data lakes from becoming data swamps is data governance (with data ownership policies supported by metadata and lineage); what balances data discovery from operational usage is data management. Organizing data and enforcing security is argued to be the antithesis of agility, freedom, and data discovery. However, striking a balance between risk, reward, and optimization is an individual decision for every company. Further strengthening core data management principles (such as minimizing data duplication and enabling data reusability), the data lake must embrace multi-tenancy and overall resource management that can be logically approached by business priority—including data classification, various data application types, and additional special considerations. Over time the data lake will move beyond the initial analytics and data science workloads to encompass operational systems workloads and thereby begin minimizing movement of data out of the data lake and instead moving the work to the data in the data lake. An enterprise challenge will be to ensure that the data lake remains singular, and peripheral Hadoop clusters and data stores don't undermine the benefit of full multi-tenancy and overall resource management of a robust, single data lake environment.

Organize by Data Classification

First and foremost, all data within the data lake will fall into one of three basic classifications: raw, derived, or aggregated. Raw data can be used for initial projects; however, the future needs of the persisted history from this same raw data cannot always be anticipated. Often, acquired raw data is extended with further classification, derivations, or enhancements that benefit follow on activities, like easing data integration or enhancing discovery or data science. For that reason, always strive to ingest raw data at a detailed level for as much (chronological) history as makes sense. (For example, ingesting operational system data into Hadoop can benefit data warehouses and data marts as a staging area; can be a fallback data set if data integration code is found to be flawed ▶

Striking a balance between risk, reward, and optimization is an individual decision for every company.

or changes; and, the same raw data can be used differently by data scientists for discovery projects and other needs.) The second classification, derived data, involves any data based on data integration, cleansing, calculations, or business rules that benefit others in the enterprise from reusability. Finally, aggregated data represents result sets and consumable information and context from a consumers' perspective—business users, customers, and applications.

Organize by Usage

The primary distinction of usage is between discovery environments and production environments within the data lake. Following basic data classifications, the next sub-organization becomes new data sources of raw data for discovery and evaluation purposes. These new sub-classifications can be used for individual or collaborative team analysis; this is where new data sources are explored and then, if deemed valuable, considered for overall data lake inclusion. Other use cases include operational and analytical systems where data access and optimizations are already well understood. (For example, derived or aggregated data warehouse historical data can be stored within optimized schemas for data access.) Ultimately, governance determines what data can be brought into the data lake as well as what data can be combined for discovery and analysis. Again, modern data architectures will reduce the need to replicate and move data as operational applications adopt YARN as the data operating system and singular data domains rise to meet the operational and analytical applications workloads within the data lake. The arrival and maturity of YARN engines (like Storm and Spark) support the adoption of this paradigm shift.

Clear, logical data organizations will benefit Hadoop engineers' technical decisions regarding physical partitioning, compacting, or compressing data files and file types to align with usage patterns.

Special Considerations That Influence Data Lake Organization

Special considerations that can influence the organization of the data lake will include proximity, compliance, and security. The concept of proximity is related to minimizing data movement across the network; if the majority of the data ingestion is cloud-based public data, some enterprises may consider logically splitting the data lake into a cloud-public portion and an internal-enterprise portion. Equally critical are enterprises' compliance requirements. These may force segregation of private data with security requirements on a portion of the data lake that is required to be maintained on-premises. Further requirements for data security may not only include access and monitoring, but also data encryption and transport between persisted data and end users.

Whether in-cloud or on-premise, discovery environments can be used for evaluation of new or external data that is later governed to be incorporated into the data lake. In some cases, companies find on-demand Hadoop environments an easier way to explore new data for consideration before it is incorporated into the larger data lake. Hadoop's strength as a low-barrier to entry of data acquisition is a precursor to data lake adoption. The trade-off is that these on-demand environments don't benefit from the full data prolificacy of having all of the data available in the data lake.

The Data Lake Unifies Data Discovery, Data Science, and Enterprise BI

One of the driving benefits of the data lake is the fact that the three data classifications outlined above are able to service multiple data application types based on a centralized data repository. Besides Hadoop meeting the needs of big data applications, the data lake meets the needs of enterprise data architecture to support enterprise BI needs (whether through persistent staging with either a data warehouse that is internal or external dependent to Hadoop) and data discovery and data science with the same raw data for statistical functions and machine learning algorithms (as well as the obvious facilitation of big data applications and unstructured analytics). Of particular interest for enterprises will be the relationship between enterprise BI and data discovery.

> Discovery hinges on the ability to explore in a fail-fast, iterative process and thus requires an environment where activities that slow down the process are as minimal —or, frictionless—as possible.

Proven BI methodology is based on metrics for performance management and context of business subject areas for consistency already pre-defined by the business. Once these definitions are converted to data models and schema, then the transformation of operational data into these metrics can be developed. The past decade saw the rise of agile BI methodologies as a response to the business not always being able to fully pre-define or understand the quality of operational source data. Agile BI methodologies offer a more responsive approach to navigating business volatility and agility. Leveraging the data lake enables data warehouse staging areas to persist detailed, raw transaction history of operational systems that can be extracted and transformed into detailed business transactions, measures, and metrics for independent enterprise data warehouses and data marts with the added benefit of new projects having historical, raw data to source from. Further, the data lake serves as a location for historical, rarely accessed data warehouse data to be persisted in addition to big data subject areas. Enterprise BI ▶

> ***Agile BI methodologies offer a more responsive approach to navigating business volatility and agility.***

methodologies further benefit from the data lake when quick data exploration for discovery, profiling, and feasibility are part of the BI requirements analysis phase. Finally, for use cases where the raw, derived, and aggregated data falls into the realm of big data, the data lake also easily facilitates those production environments.

With the fast-paced change in competitive business environments today, companies are realizing the critical role of discovery, and that uncovering answers to questions they don't already know are equally as (if not more) important as the pre-defined metrics, business rules, and definitions that currently exist in the enterprise. This prompts the need for analysts to work directly with the data to discover definitions that are meaningful to the business. These can then be moved to eventually live in the Hadoop environment, in the enterprise data warehouse, or multiple data warehouses or data marts. The data lake enables user and collaborative discovery by making the ability to work with data more accessible and by minimizing the friction incurred through incremental IT activities. Robust and efficient discovery environments within the data lake facilitate the discovery of new business definitions, insights, and analytics that can be institutionalized through governance as enterprise or departmental perspectives.

Vastly improving SQL access to the data lake has improved the compatibility of data warehouse and discovery processes with the data lake in a modern data architecture.

Previously, when physically independent analytic sandboxes attempted to meet the needs of self-service data discovery by providing a stand-alone environment, these environments were at a disadvantage: they did not have the benefit of the data lake's overall easy access to all raw, historical data, nor the ability to explore unstructured or yet-to-be-defined-schema data. Meanwhile, the virtual analytic sandbox approach will continue to provide value with an abstraction of available data assets and computing resources, and will continue to become the standard approach for managing agility and discovery within the much larger data lake. As discovery environments become increasingly critical, business intelligence and business analytics are becoming more valuable when coupled with robust discovery environments.

Critical Data Lake Success Factors

A data lake produces compounding value for the business as it grows and becomes part of the overall enterprise data strategy. However, current data management mindsets require justification of the value of data instead of quantifying the potential value of data. In other words, a more long-term approach to data storage dictates a new mindset to achieve critical mass with the data lake. Data- and analytic-centric companies are driven by an obsession with acquiring and analyzing as much data as possible.

Rethink Data for the Long-Term

Several tenets for achieving data lake critical mass can be applied to every data project. To begin, every data project should consider and recognize its data for reusability in future applications and discovery, while understanding that upcoming data needs will be unknown—this is balancing the decision of data lake retention after the current process has consumed the raw data. Second, given the data lake's lower barriers to data ingestion, consider acquiring new data as it is identified so that more history will be readily available later for projects that are working down the coming pipeline that would benefit from historical data. Third, as part of a centralization strategy, an embrace of the concept of the data lake will avoid the costly duplication and movement of data throughout the enterprise. The arrival of Hadoop 2.0 YARN, fundamentally changed Hadoop from an exclusive, batch-oriented jobs system to being able to manage overall resources for many data processing engines accessing data within HDFS (supporting batch, interactive, and streaming applications). While robust resource management was much needed for cluster management, the significance of YARN is its facilitation of current and future data application engines in a single environment. Continual investment into growing the data lake fosters the compute and memory resources to successfully manage all users and applications. The benefits of data engines and single data persistence are what enable the concept of a data operating system that reduces data duplication and movement across the enterprise.

Establish Data Governance First

The need for data governance has always reinforced a framework for the business drivers and risk management related to all data and information policies in the enterprise through the definition and assignment of data owners, data stewards, and specialists. These authorities and delegations should now be extended to the data lake. While the data lake is compounded by increased diversity of data usage as well as its intrinsic capability to house many varieties of data, the data governance framework itself should remain fundamentally unchanged. Thus,

▶

A more long-term approach to data storage dictates a new mindset to achieve critical mass with the data lake.

the second largest contributor to successful data lake growth is establishment of data governance at the onset of the data lake strategy. As outlined in the earlier four-stage maturity model, companies typically begin with a pilot project and then respond to reactionary use cases before moving ahead to proactive activities. If data governance is not established at the start of the data lake strategy, major complications can arise when data governance has to be retrofitted—if that is even possible. In worst-case scenarios, companies must create a second data lake to address data governance and security, and destroy the original attempt that missed data security at ingestion, before data is replicated throughout the HDFS where file deletions are more difficult.

Governance establishes a common understanding for how everyone will work with data in the data lake and mitigates the doubts and second-guessing that is harmful to data lake momentum. Established data governance programs will need to extend current definitions surrounding data owners, stewards, roles, and delegation rights to support big data, data discovery, and data science all contained in a single data lake. New data governance policies should focus heavily on data ingestion, covering the evaluation of data sources (internal and external), acquired third-party data, and specialized user data sets. Discovery-oriented policies will address data accessibility for the iterative discovery process and the institutionalization of discovered context and analytic models. Policies related to data integration and the distribution of analytic models and insights should also be given special consideration in data lake governance. Finally, data governance will inherently define other key technology requirements, like access controls, mobility, and security.

Tackle Security Needs Up-Front

A data-centric security approach provides a broad perspective to think about data from creation to consumption. The cornerstones of data security will be authentication first and then authorization. Authentication must first verify that the user is actually who they claim to be through the use of strong passwords, two-pass authentication, or security tokens. Once verified, a secure, centralized repository of data access credentials are made available to the governing application to invoke. Therefore, the key factor in a data-centric security approach is to understand the data at each data element level to define access rights for what will be brought into the data lake. Taking this further, determining the point which the data may need to be originally encrypted—and to what degree of encryption—will dictate how security technologies will work together. The degree of decryption needed will be defined by data governance policies; however, security technology is responsible for the consistent enforcement of those policies across all forms of data consumption from the data lake. While encryption for data at rest is a common tactic for data security, the reality is that breaches

Data governance will inherently define other key technology requirements, like access controls, mobility, and security.

in security occur between the various application layers of access to the data. Production data lakes must pay very close attention to managing the physical entry points, such as command line and network access of users to the cluster.

Equally important in the security approach is to understand data usage patterns to recognize how to properly secure and encrypt data in the data lake. For analytics, data result sets typically do not include secure data elements because they are descriptive statistics about the result data set. However, where an encryption technology can preserve the format and logical value of the data and data relationships (such as date ranges), the analytic itself can be based on information that is encrypted without requiring the analyst to have security access. This will vary for data that is accessed for data science, for data discovery, for business intelligence, and for operational applications. Therefore, understanding a data element's usage by the enterprise enables optimized security controls and the ability to set the proper level of data encryption (or masking) to avoid unnecessary performance degradation.

Some companies will incorporate the concept of security zones in the data lake. Hybrid cloud data lake architectures are an example of secure data on-premises and non-secure public data in the cloud (such as data acquired from social network platforms) with careful restrictions on how secure and non-secure data can be joined. Another data lake architecture may include separate Hadoop clusters for highly secure data, or a landing zone for encrypting data prior to ingestion into the primary data lake cluster.

Conclusion

While the data lake is new to the data world vernacular, it is ultimately a concept that allows companies to have conversations with a common lexicon. These conversations drive clarity and reduce the risk of data lake initiatives becoming the very data swamps about which industry pundits warn. While there is reason for data management and data governance professionals to be concerned when adopting the data lake, the existing lack of definition within an enterprise should not condemn the data lake itself.

The data lake strategy is a part of an enterprise IT application and data strategy that is architecturally sound and goes well beyond the conversation of cheap commodity infrastructure and data science or machine-learning analytics. Instead, the right conversation should provide the foundation for the data lake to realize its full potential to have profound impact on enterprise data architecture and become the next-generation operational system. This replatforming of existing IT applications and data management may take years to achieve, but the arrival, rapid maturation, and adoption of Hadoop's data persistence layer and application engines layer is more aligned to the decades of service-oriented architecture or computing systems architecture than most realize. This paper provides the definitive guide on the critical areas of importance to bring data lake organization, governance, and security to the forefront of the conversation to ensure a successful —and efficient—journey to the data lake.

Sponsors

Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities.

The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications.

Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop.

www.hortonworks.com

TERADATA

Teradata helps companies get more value from data than any other company. Our big data analytic solutions, integrated marketing applications, and team of experts can help your company gain a sustainable competitive advantage with data. Teradata helps organizations leverage all their data so they can know more about their customers and business and do more of what's really important.

www.teradata.com

THINKBIG

A TERADATA COMPANY

Think Big, a Teradata company, offers big data roadmap, architecture, engineering and ongoing support services for data lake and analytic solutions.

http://thinkbig.teradata.com/

MAPR

MapR delivers on the promise of Hadoop with a proven, enterprise-grade platform that supports a broad set of mission-critical and real-time production uses. MapR brings unprecedented dependability, ease-of-use and world-record speed to Hadoop, NoSQL, database and streaming applications in one unified distribution for Hadoop.

MapR is used by more than 700 customers across financial services, government, healthcare, internet, manufacturing, media, retail and telecommunications as well as by leading Global 2000 and Web 2.0 companies. Amazon, Cisco, Google, Teradata and HP are part of the broad MapR partner ecosystem.

www.mapr.com

Voltage security

Voltage Security®, Inc. is the world leader in data-centric encryption and tokenization. Voltage provides trusted data security that scales to deliver cost-effective PCI compliance, scope reduction and secure analytics. Voltage solutions are used by leading enterprises worldwide, reducing risk and protecting brand while enabling business.

www.voltage.com/hadoop