Seven Rules for Big Data Analytics Deployment

Guidance for Implementation Success

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) White Paper

May 2017

Prepared for:

EMA™ *IT & DATA MANAGEMENT RESEARCH, INDUSTRY ANALYSIS & CONSULTING*

Seven Rules for Big Data Analytics Deployment: Guidance for Implementation Success

Table of Contents

Seven Rules for Big Data Analytics Deployment: Guidance for Implementation Success

Executive Summary

The future of business intelligence (BI) for data-driven organizations will revolve around analytics using big data. Organizations that can use fine-grained information that either was not previously available or was not properly used will be prepared for new business models and the changing economy that data-driven strategies and culture empower. In contrast, organizations that only rely on "traditional" data sources such as the enterprise data warehouse or data marts will soon find themselves at a disadvantage to those that put insights based on big data into the hands of their business stakeholders.

To successfully implement analytics and BI on big data, organizations need to understand these seven rules:

1. **Support business stakeholders**. If analytical applications are going to impact a business's revenues, costs, and margins, those applications need to be accessible to the line of business stakeholders.
2. **Provide SQL access**. Business stakeholders like their tools, and most of those tools use the SQL (99/2003) standard to access information. Analytics on big data need to support these requirements.
3. **Deliver quick response times**. The pace of business makes business stakeholders a very impatient group. Speed of response is key to any analytical application if it is to make a true impact on the business.
4. **Provide self-service capabilities**. Making data consumers wait for IT to make changes is a poor use of their time and places IT in a difficult position. Analytics on big data environments need to have a self-service configuration component to speed implementation.
5. **Support multiple platforms**. The information that a data-driven organization requires does not reside on just one platform. Analytics and BI on big data must support the inclusion of multiple data management platforms, both on-premises and cloud, such as curated data from the EDW, master data management platforms, contextual information from operational applications, Hadoop, cloud or server-less, and external data sources.
6. **Bring data together, not duplicate it**. Integrating analytics tool access to data to minimize, or completely eliminate, the proliferation of data copies and movement around an organization will speed implementation and improve confidence.
7. **Offer a combined semantic layer**. A single point of self-service manageable metadata will break down the barriers to implementation and give business stakeholders "ownership" over their data.

Using these business and architectural guidelines, organizations can expand their big data environments from simply being "storage options" and avoid the risk of a "data swamp." EMA considers these seven rules essential to successful and effective projects using analytics on big data.

IN THE KNOW

WHO: Architects and business executives of data-driven companies.

WHEN: Implementing analytics and business intelligence on big data.

WHAT: When data-driven organizations make the decision to implement analytics on big data, they often need guidance. This white paper provides seven rules that technologists and business stakeholders should follow when implementing analytics and business intelligence on big data.

Big Data Analytics Use Cases and Applications

Analytics projects using big data come in many shapes and sizes. However, the most successful ones link the newly available data and information directly to activities that impact topline revenue, operational costs, or bottom-line profitability. These are the projects that gain the greatest adoption and receive the most support from management—and not just technical management. An important aspect of implementing analytical projects that use new big data sources is obtaining the support and sponsorship of the business stakeholders (the executives in marketing, finance, sales, and customer care) and the analysts who take that information and turn it into new prospects, new sales, lower costs, or improved customer experiences. Here are the three most common analytical projects using big data that emerged since 2012:

RULE #1

Analytics and BI on big data must support business stakeholders: If analytical applications on big data are going to impact a business's revenues, costs, and margins, those applications need to be accessible to the line of business stakeholders.

Implementation Action: Link your big data analytics project with your line of business strategic goals and prioritized use cases. This will increase visibility and impact with the business (particularly with the C-level), speed adoption, and help drive ROI across your organization.

Risk and Cost Analysis

Protection against risk and exposure to loss is key for just about every organization. For organizations with an online presence and/or an automated product sales and provisioning system, the hard dollar costs associated with fraud can be significant. Providing identification of and confirmation analysis regarding high risk or fraudulent incidents is crucial to lowering the cost of operations. Organizations using big data can manage those risks and potential losses by monitoring and managing the activities from their online and mobile app operations, as well as monitoring their automated provisioning and activation processes.

Customer and Product Intelligence

Using a big data analytics environment to access event-level information that comes from customers via online and mobile applications, as well as from product sensor data in an Internet of Things (IoT) context, provides detailed information about how customers interact with an organization or an organization's products. Likewise, a big data analytics environment with data from product sensors supplies information on how products perform and behave, helping to guide product refinement and new product development.

Marketing and Sales

To increase revenues, organizations need to have up-to-date information on their marketing and sales efforts. To this end, organizations mastered marketing activities that track outbound communications and inbound responses with astounding detail, even linking event-level touchpoint details with prospects. Understanding how those details interact with sales opportunities gives organizations the ability to use big data analytics to match marketing activities directly with opportunities and gauge the probability of those opportunities with accuracy unmatched in previous eras. This level of information and analysis can not only drive out misspent marketing costs, but motivate new areas of sales revenue.

Business Requirements for Analytics on Big Data

Business stakeholders have certain requirements for analytics on big data projects that they use on a daily basis. These requirements usually center on three common themes:

- **Access to data:** How will the business executives and analysts access the information? Forcing them to use highly technical platforms is not going to speed adoption and success.
- **Speed of results:** Business stakeholders are notorious for wanting results quickly; not because they are impatient, but because the speed of business is moving faster and faster. Analytical results must support the interactive aspects of exploration in which data consumers work through several "what-if" scenarios. Results that can provide "keystroke response" have a better chance of being adopted.
- **Self-service:** Projects in which IT must constantly make updates to align with the current state of customers, products, or sales will be too cumbersome to maintain. If they have the correct tools, business stakeholders can often "self-service" adjustments to data, schema, and implementations. These are not major architectural adjustments, but the types of changes that business users request every day.

Data Access

For any integrated environment, data access and security are important features. Traditional data management platforms based on relational database management systems (RDBMS) have robust user/group security paradigms and allow for access via SQL. However, data security on various newer platforms, such as Hadoop and NoSQL, is not at the same level as traditional platforms. There are also new groups of data consumers that need data access to API-based web services or micro-services. These data consumers include third-party data consumers, application developers creating embedded result sets, and data scientists creating advanced analytical models and machine-learning algorithms.

Moreover, today's analyzing and visualization tools speak one of two languages: SQL or MDX. As a result, data access must extend beyond just the SQL API. It is imperative that data access layers enable both SQL- and MDX-based tools so users can access underlying data sets using a variety of analysis tools.

Latency

The business requirement for faster processing response time is one of the key drivers of big data analytics projects. In big data environments, it is not sufficient to provide processing response in a batch-processing timeframe. This concept supports the premise that big data initiatives require data management platforms that go beyond the

RULE #2

Analytics and BI on big data must speak both SQL and MDX: Business stakeholders like their tools, and while most of those tools utilize the SQL (99/2003) standard to access information, some key tools only speak MDX. Analytics on big data need to support both.

Implementation Action: Choose a platform that complies with SQL and MDX standards as much as possible. Business stakeholders use analytical and visualization tools that are predominately based on the SQL data access paradigm (ANSI SQL 92, 99, 2003). Making them move to other tools slows (or even kills) adoption.

RULE #3

Analytics and BI on big data must deliver fast response times: The pace of business makes business stakeholders a very impatient group. Speed of response is key to any analytical application if it is truly going to have an impact on the business.

Implementation Action: Choose a big data analytical environment with the ability to respond shortly after a mouse click. This holds the attention of data consumers and allows them to iteratively work through analysis and "what-if" scenarios. It also creates a "sticky" application that data consumers will adopt and use.

latency often associated with Hadoop's batch response time. The result sets, whether they are operational or analytical in nature, need to show improvement over the speeds associated with existing or traditional data management platforms. The expectation is that the results will be faster than those in previous eras of computing, regardless of the processing requested.

Self-Service Agility

Organizations implementing projects using big data want "configurable application environments," as opposed to "hand-rolled technical implementations." This provides a self-service approach oriented around adjustments to their analytical environments. Environments that require customized coding or constant interfacing with technical resources from IT departments to make "minor course corrections" in the application or structure of the data leave those environments that understand the end goal out of the loop. Often, the business stakeholders will understand the adjustments that need to be made, but using a platform that requires a long and cumbersome process to implement changes will impact the success of the analytical project.

By placing self-service modeling and analytics tools in the hands of the analysts and the business, two objectives are met. Business teams attain the speed of change they are looking for, and IT teams can be seen as value-added partners instead of roadblocks to the next revenue opportunity or customer experience improvement.

Implementation Guidance for Analytics on Big Data

With the concepts above in place, the question becomes how to effectively implement your analytics on a big data environment. Aspects to consider are the types of attributes your environment will require, the integration of data between the various platforms that make up the world of big data, the power of having an integrated business-oriented semantic layer, and linking all of these new sets of information together.

Multiple Platform Attributes

In recent years, there was tremendous growth in terms of the types and number of data management platforms. Significant growth has occurred within the NoSQL community of data management platforms that includes

RULE #4

Analytics and BI on big data must provide self-service agility: Making data consumers wait for IT to make changes is a poor use of their time and places IT in a difficult position. Analytics on big data environments need to have a self-service configuration component to speed implementation.

Implementation Action: Use a solution that allows technical- and business-focused data analysts to configure their own adjustments to your big data analytics environment. Not all questions can be "foreseen," and allowing those with the knowledge of the business to configure those adjustments rather than requiring custom re-coding is important to driving the right decisions at the right time.

RULE #5

Analytics and BI on big data must support multiple platforms: The information that a data-driven organization requires does not reside on a single platform such as Hadoop. Analytics and BI on big data must support the inclusion of multiple data management platforms, such as curated data from the EDW, master data management platforms, contextual information from operational applications, and external data sources.

Implementation Action: Don't be afraid to include data from multiple platforms in your big data analytics environment. Not all information resides within any one platform. For example, event data resides in Hadoop and NoSQL. Curated business data exist in the enterprise data warehouse. Revenue and cost information come from operational platforms. All of these platforms can be considered in your big data analytics environment. Just make sure that the BI and analytics tools your business uses (remember Rule #1) can access that hybrid data without breaking Rule #6.

MongoDB, Cassandra, and Hadoop. When organizations implement their analytical big data projects, these are the some of the most popular platforms and the attributes that make them key to those analytical projects:

- **Analytical Platforms:** Analytical database platforms are used for high-volume, low-latency workloads and processing. These platforms provide the speed and the data access layers important for the iterative and interactive workloads using SQL-compliant access layers.
- **Discovery Platforms:** Data discovery platforms combine the abilities of SQL-based analytical platforms and Hadoop environments. These platforms bring together information from traditional data sources and new multi-structured datasets for exploratory workloads. Additionally, the platforms have data preparation, profiling, and analytics capabilities, which span from relatively simple counts and sums to clustering and correlation functionality. Many of these discovery platforms bring the value of self-service configuration of data layout and processing schema.
- **Hadoop:** Hadoop environments grew from the Hadoop Distributed File System (HDFS), supporting multi-structured datasets like hierarchical documents, unstructured formats like audio and video files, and the MapReduce processing engine. HDFS also supports several processing engines such as YARN and Spark. These cost-effective and flexible data storage platforms changed the way organizations look at highly-detailed event data and new data sources.

RULE #6

Connect to data, don't just duplicate it: Integrating data to minimize, or completely eliminate, the proliferation of data around an organization will speed implementation and improve confidence.

Implementation Action: Minimize or eliminate the replicas of data associated with your big data analytics environment. These copies create issues with data governance, privacy, and security.

Data Integration

Analytics on big data is not just a single-platform affair. Critical information is spread across multiple locations within any organization. Curated data in the enterprise data warehouse gives context to customer, product, and sales information. There may be event data from mobile apps, online applications, and now the Internet of Things (IoT). EMA found the following data integration techniques to be the most frequently used and important to achieving successful analytics on big data results:

- **Traditional Data Integration:** Traditional data integration techniques between data sources and data management platforms are common within a big data analytics environment. These include extract, transform, and load (ETL), which is common between operational platforms, EDW, and data marts. They also include ETL techniques common between platforms where the transformation occurs within the target system. Hadoop platforms may also offer transform, extract, and load (TEL) methodology because Hadoop has multiple processing engines and the capacity to transform data before it is extracted from the HDFS for sharing between platforms. Whichever integration process you choose, once you have the data in your chosen big data platforms, take

RULE #7

Analytics and BI on big data must offer a combined semantic layer: As a single point of self-service, manageable metadata will break down the barriers to implementation and give the business consistency in data "vocabulary" across analytics and decisions.

Implementation Action: Combine your technical and business metadata into a single repository or layer. This allows your organization to manage metadata in a single location and make updates in one place, using a single interface.

heed of Rule #6 and stop moving data back out in order to make it accessible by your BI tools. Don't copy it or move it. Though moving it is tempting since this is the traditional approach, ultimately, moving the data leads to delays in business user access (breaking Rule #1) and added workload on IT (repeated movement to refresh the data for business user analysis).

- **Streaming Integration:** Streaming integration is the latest integration strategy to come on the scene. Many enterprise architects compare streaming integration with the legacy methodology of complex event processing (CEP). However, streaming integration and processing implement the concepts of CEP with greater scale and a more flexible architecture. Kafka and Flume are popular implementations of integration between streaming data sources, such as online applications and sensor data from IoT devices. When using streaming integration into a big data platform, once again adhere to Rule #6 and don't move, copy, or duplicate data again just to make it accessible to BI users and their tools (from Rule #1). Even streaming data should be able to be analyzed in place in the big data platform you've chosen (traditional, on-premises, or cloud-based).

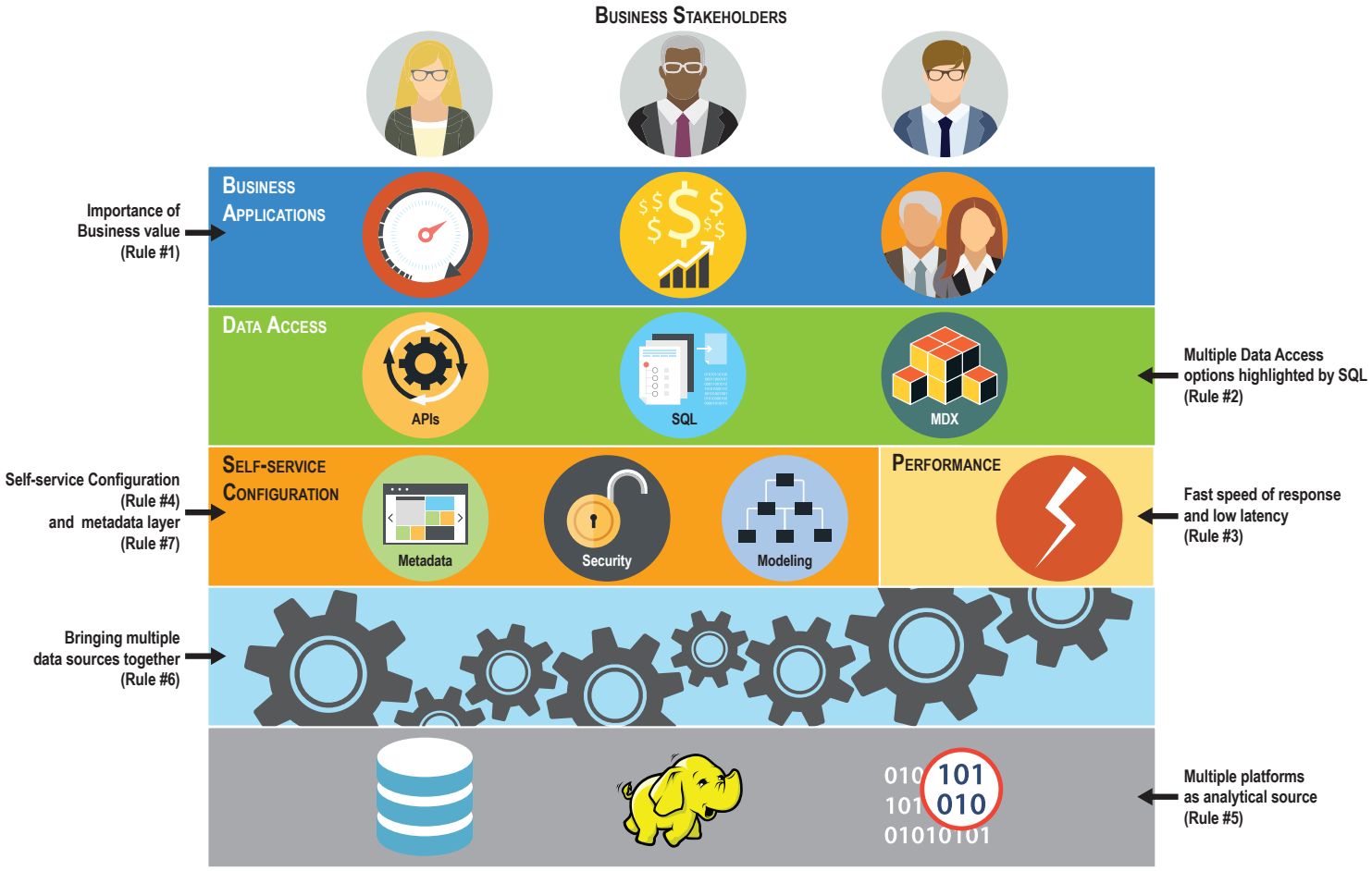
Consolidated Metadata Layer

One of the key aspects of any big data analytics environment is the ability manage metadata effectively –to store and share data about your data's "metadata" stored across your environment and understand and utilize the business value of that information. This consolidated metadata layer requires the management of both business and technical metadata. It also requires tracking the semantic metadata that defines what the information relates to. Examples are a customer, product, store, or employee. Each of these can be spread across multiple data management locations, but you need a single repository to effectively make use of it. To support self-service, the consolidated metadata layer should be frequently updated as business analysts, and other line of business stakeholders, discover and refine information about the organization's data.

EMA Perspective: How to Bring it all Together

Successfully implementing analytics and business intelligence with big data involves many moving and related parts, including technology, priorities, and processes across the business. Whether using a single platform or a series of technologies that work in concert, a functional architecture diagram can aid architects in design and inform business stakeholders and sponsors how analytics and business intelligence in big data environments can come together.

The following functional architecture diagram shows the various components involved in big data analytics and how those components relate to the seven rules covered in this paper. The purpose of the diagram is to deliver a view of how and where the components in a typical business intersect with rules for successful analytics on big data deployment. From there, you can create a similar diagram to represent your own company's architecture and goals, and more easily identify which "rules" need prioritizing and addressing to ensure the highest chance of BI on big data deployment success.



In summary, analytics and BI on big data must follow the Seven Rules of Big Data Analytics Deployment:

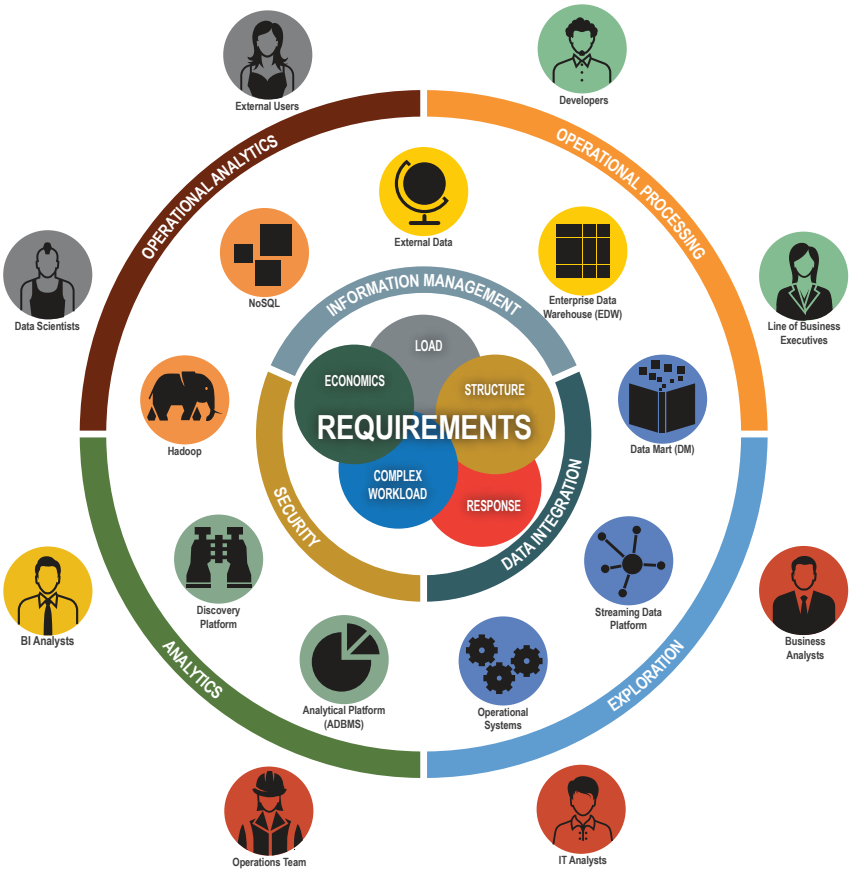
- **Rule #1 – Analytics on big data should support business stakeholders**. If big data analytics are going to impact revenues, costs, and margins, they need to be accessible to the line of business.
- **Rule #2 – Analytics on big data should provide SQL and MDX access**. Business stakeholders like their analysis and exploration tools, and most use the SQL standard to access information.
- **Rule #3 – Analytics on big data should deliver quick response times**. The pace of business makes the line of business an impatient group. Speed of response is key to adoption and impact.
- **Rule #4 – Analytics on big data should provide self-service capabilities**. Making business analysts wait for IT is a waste of time. Analytics on big data must provide self-service agility.
- **Rule #5 – Analytics on big data should support multiple platforms**. Information that data-driven organizations require does not reside on just one platform. Big data analytics must support multiple data management platforms.
- **Rule #6 – Analytics on big data should bring data together, not duplicate it**. Minimizing, or completely eliminating, the proliferation of data is core to any big data analytics environment.
- **Rule #7 – Analytics on big data should look to leverage a combined semantic layer**. As a single point of self-service, manageable metadata will break down the barriers to implementation and give the business consistency in data "vocabulary" across analytics and decisions.

In EMA's opinion, successful big data analytics deployments ultimately require the support of all seven rules of big data analytics. Realizing that every organization has its own unique characteristics, we suggest aligning your environment and processes in such a way as to heed as many of the rules as possible. Doing so will provide you the best chance for success when deploying the mission-critical analytical and business intelligence applications that will broaden and continue the adoption of big data as a valuable corporate asset.

APPENDIX

EMA Hybrid Data Ecosystem

The EMA Hybrid Data Ecosystem (HDE) is a framework that provides organizations guidance on how to implement key big application workloads for business stakeholders by utilizing existing and new data available from a wide range of sources both internal and external to our existing definition of the data center. One application of the HDE is a big data analytics environment.



The core of the EMA HDE consists of the business requirements of an organization: Economics, Data Load, Data Structure, Speed of Response, and Complex Workloads. These business requirements drive the selection and use of the appropriate data management platforms to store the information from various data sources.

This information is managed via information and metadata management, data access and security, and data integration layers between those data management platforms. This management will provide the most effective and responsive processing and presentation of that data to serve the information needs of data consumers from across the business spectrum, including business stakeholders in the line of business and operations, technical roles including data scientists, and third-party stakeholders such as customers and corporate partnerships.

For more information on EMA's Hybrid Data Ecosystem, go to: http://www.enterprisemanagement.com/

AtScale

Organizations who want to reduce the complexity of traditional approaches to accessing data within the data lake can leverage AtScale. AtScale reduces (or eliminates) the need for data movement between platforms and analytical tool-specific data models. AtScale allows organizations to access data lake information using a wide range of analytical tools. This comes from the SQL-compliant interface that allows business executives and analysts to use their analytical or data visualization tools of choice. The AtScale interface provides direct, secure, and iterative access for business analysis to important customer, product, and sales information within a data lake environment.

Business users can self-serve their modeling requirements directly with the information in their data lake by using the AtScale Design Center web application. From the AtScale interface, data analysts can design virtual OLAP cubes using familiar workflows and intuitive drag-and-drop interactions without using command-line interfaces or understanding the underlying data structure(s).

The AtScale Intelligence Platform works with enterprise data lake deployments, on-premises, or cloud to offer a data governance and security layer consistent with existing business intelligence and analytical environments. IT teams and administrators have complete control over access within the data lake, while simultaneously delivering to business users' valuable self-service access to analysis across the variety and breadth of big data.

For more information on AtScale, visit: http://www.atscale.com/technology/

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com

3572.050417