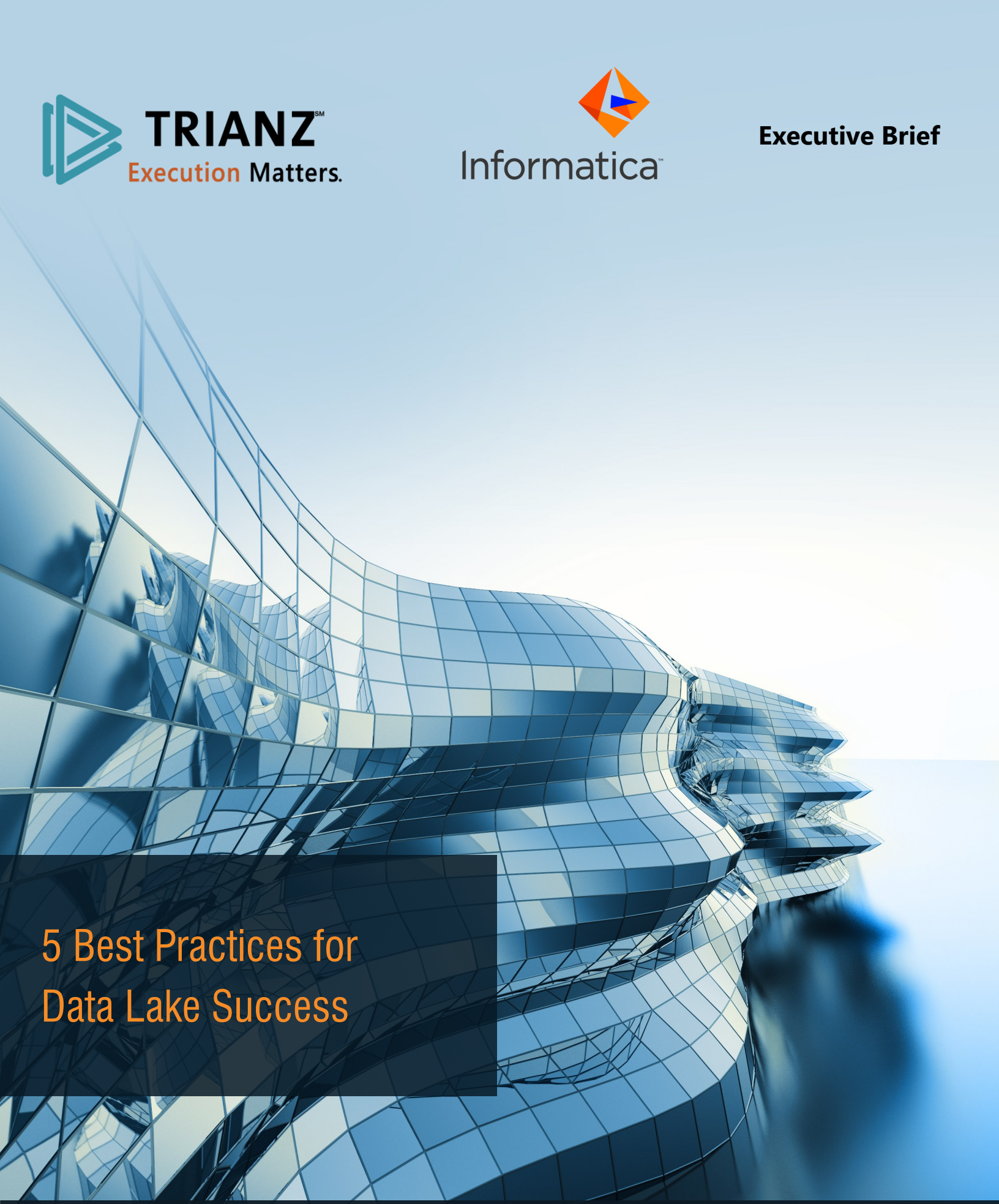
TRIANZ℠
Execution Matters.

Informatica™

Executive Brief



5 Best Practices for Data Lake Success

CLOUD • ANALYTICS • DIGITAL • SECURITY

5 Best Practices for Data Lake Success

Applying big data governance and analytics to maximize business impact

Avoiding a 'Data Swamp' Failure

The potential for business value from a data lake is immense. Data lakes equip organizations to generate data-driven insights not possible with conventional analytics, opening new avenues to transform customer experiences, supply chain efficiency, fraud detection and other areas that bear directly on bottom-line profitability.

Data lakes are distinctly different than data warehouses, which store structured, relational data from ERP, CRM, marketing, customer service and other business applications. Data warehouses have delivered value for decades, but are not well suited to waves of new information types accessible to enterprises.

In contrast, data lakes serve as a repository for structured, semi-structured and unstructured information. That includes raw data from mobile devices, social media, clickstream and server logs, text documents and a wide variety of sensor-based equipment in the Internet of Things. By blending data types, business analysts and data scientists can explore information for breakthrough insights not otherwise attainable.

But data lake initiatives are poised to fail unless they are built on a comprehensive data lake management platform. Without a comprehensive approach to data lake management, processes to ingest, prepare, access, govern and secure data can be excessively slow and labor-intensive. Data lake management best practices can also help ensure that data lakes will scale and be reusable for years to come.

With a data lake management platform and best practices, organizations are equipped to avoid the risk of a "data swamp" full of raw information that is virtually unusable.

Unleashing New Business Insights with Data Lakes

An analytic-oriented data lake is among the top uses for Apache Hadoop, the open-source framework geared for unstructured big data growing rapidly in volume, velocity and variety. Organizations have turned to Hadoop and commodity hardware as a cost-effective means of storing and archiving many terabytes or even petabytes of data.

Hadoop is also widely used for compute-intensive data processing, and to handle data staging and preprocessing for a data warehouse or data integration tool. But advanced analytics is the greatest business benefit of Hadoop, cited by 48 percent of respondents to a Data Warehousing Institute survey.[1]

Hadoop-based data lakes are still emerging, but momentum is strong. For instance, a study by Unisphere Research found that 65 percent of organizations had data lakes in production, under deployment or in exploratory research phases.[2] Data lakes complement, but do not replace, traditional data warehousing systems in offering possibilities for new insights with data science and exploratory analysis.

Across industries, organizations are taking the data lake plunge in a variety of use cases aimed at monetizing data — generating quantifiable business value from analytic insights:

- **Customer engagement:** Analysis of customers' social media activity, especially when tied to CRM or billing systems data, opens new opportunities for personalized marketing, improved retention and better customer experiences.
- **Manufacturing and logistics:** Vast volumes of sensor-based data in factories, products, trucks and more can be utilized to pinpoint problem areas and significantly improve cost efficiency.
- **Fraud and security breach detection:** Data scientists can identify indications of fraud or network security breaches by using Hadoop's massive compute power for real-time monitoring and exploration of data from multiple systems.

---

[1]The Data Warehousing Institute, "Hadoop for the Enterprise," June 2015.

[2]Unisphere Research, "Data Lake Adoption and Maturity Survey Findings Report," October 2015.

- **Defect tracking and remediation:** Capture and analysis of diverse data across a product's lifecycle — from production to customer service and warranty claims — gives manufacturers breakthrough insights to improve quality.
- **Healthcare:** Industry stakeholders have a wealth of opportunities to improve clinical outcomes, reduce costs and transition to value-based care with data from medical devices, fitness wearables, epidemiological databases, patient records and more.

Leading data lake adopters are today deriving value in dozens of use cases. With huge varieties and volumes of information at their disposal, technologists are limited only by their infrastructures and imagination in turning raw data into insights.

Key Challenges in Managing Data Lakes

Business insights don't magically bubble up from a data lake. The sheer volume of data alone poses a challenge to IT. Due to limitations in early-generation Hadoop tools, some organizations resort to manual, resource-intensive manual work and multiple point products to ingest, curate and analyze raw data.

Manual approaches to data lake management are neither sustainable nor scalable. IT and data scientists can become consumed by manual chores, while business analysts wait impatiently for usable data. As it is, data scientists can spend up to 80 percent of their time preparing data before doing any meaningful analysis.[3]

Without the right processes and tools, organizations can quickly lose control of a data lake and miss out on the opportunity to explore the complex interrelationships that may be hidden in diverse, blended data. The challenge of effective data lake management is well recognized. For instance, the analyst firm Gartner advises:

"While it is certainly true that data lakes can provide value to various parts of the organization, the proposition of enterprise-wide data management has yet to be realized. Without at least some semblance of information governance, the lake will end up being a collection of disconnected data pools or information silos all in one place."[4]

Comprehensive data lake management solutions help organizations quickly and repeatedly turn big data into useful information assets. For instance, the IT consultancy Nucleus Research found that users of Informatica data lake solutions can accelerate time to insight by 57 percent.[5] At the same time, those organizations can minimize complex manual work and flexibly scale data lakes to evolving business needs.

5 Best Practices for Data Lake Success

A sound technology platform goes hand in hand with best practices to help organizations capitalize on data lakes. Leading best practices are aimed at driving user adoption, sustainable governance and business insights through:

1. Visibility and accessibility
2. Data findability and collaboration
3. Analytic flexibility and visualization
4. Governance and security
5. Data science

1) Provide Visibility and Accessibility for Data Lake Contents

Data lakes will underperform if users lack visibility into data lake contents, which in raw form lacks structure, context and meaning. An information catalog drives user adoption by providing an intuitive framework for IT and business users to understand data lake contents and dive into data exploration.

Based on metadata services and machine learning algorithms, an information catalog collects, indexes and profiles the contents of a data lake. The catalog automatically classifies data from diverse sources by dimensions such as email address, credit card number, company name and so forth, and supplies a business glossary for data stewards to define business terms and data definitions for consistency.

---

[3]The New York Times, "For Big Data Scientists, 'Janitor Work' Is Key Hurdle to Insights," August 17, 2014.

[4]Gartner, "Gartner Says Beware of the Data Lake Fallacy," July 28, 2014.

[5]Nucleus Research, "Informatica Is the Data Lake Lifeguard," September 2016.

2) Promote Data Findability and Collaboration

Importantly, leading information catalogs also support data interaction with powerful semantic search capabilities for users to quickly find relevant data, and faceted search to filter results by specified criteria. That enables analysts to use natural language search terms to identify and narrow down data of interest. They can also discover relationships across data sets that would be extremely difficult to identify with a manual approach.

Sound data lake management also enables IT personnel, data scientists and non-technical business analysts to collaborate in delivering trusted information assets. Common project workspaces enable viewing and sharing of profile statistics, data lineage, analytic results and more. Users should also be able to assign custom attributes and annotations to data sets to enrich meaning. Role-based permissions should be implemented to limit user access and revision privileges to appropriate data sets.

3) Support Flexibility and Visualization in Analytic Tools

A few years ago, organizations faced a lack of user-friendly analytic tools that could be used with Hadoop. Many business analysts and BI professionals lacked the skills to query Hadoop-based data lakes using such native, open-source tools as Hive, HBase, Pig and MapReduce. Reliance on skilled Hadoop technologists to run analytics limited user adoption and value.

Today, leading data lake adopters are taking advantage of new Hadoop support in leading BI and advanced analytics technologies to broaden the user base. For instance, usage and insights rise when business users have self-service flexibility to utilize their BI tool of choice, be it Tableau, Qlik, IBM Cognos or others. Similarly, a data lake should be accessible to more advanced statistical and data mining tools used by data scientists, such as SAS, IBM SPSS or the R programming language.

4) Leverage Data Governance and Data Security to Ensure Data is an Asset

As a repository for huge volumes of diverse data, a data lake poses unique governance and security challenges that are particularly important for insurance, financial services, healthcare and other organizations that traffic in sensitive data. Traditionally, governance for Hadoop data lakes has not matched that of more mature data warehouses. Governance is now becoming more critical as data lakes gain a more prominent role in the technology stack.

IT and business leaders need to collaboratively devise a metadata-based governance framework that supports auditability, internal and external compliance, data lifecycle management and usage monitoring, i.e., who is accessing what data. Data lake management software supplies governance capabilities and can augment open-source governance in leading Hadoop distributions. Dedicated data stewards who oversee governance support standardized policies and sound, secure data lake evolution.

5) Empower Data Scientists to Focus on Analytics

Monetizing data depends in large part on data scientists who can uncover patterns, anomalies, interrelationships and critical indicators in massive data sets. But data scientists are often too burdened with labor-intensive processes to ingest, cleanse and prepare data that inhibits them from focusing on high-value analytical work.

An empowered culture should equip data scientists with tools that centrally and automatically provide the information assets required for statistics, predictive analytics and machine learning. By automating critical data lake management activities, data scientists can become trusted advisors for functional and line of business leadership and not be consumed by manual and time-consuming management activities.

Accelerate Your Efforts with Data Lake Management

Trusted strategic advisors help organizations quickly and securely implement data lake environments, minimizing risks and positioning IT and business users to generate groundbreaking business insights. With more than 2,000 client engagements over 15 years, Trianz partners with Informatica to accelerate your transformation into a data-driven business by turning data swamps into intelligent data lakes.

The joint Trianz-Informatica solution combines Trianz's unmatched expertise in enterprise information management with the Informatica Data Lake Management solution, which supplies industry-leading capabilities to ingest, find, prepare and protect data for analysis inside an intelligent data lake. Trianz helps guide organizations with strategic conceptualization, roadmap planning, use case development, implementation and optimization.

In tandem with Informatica Data Lake Management, Trianz also supplies expertise in Hadoop, analytic solutions and data science to give organizations a flexible, well governed data lake ecosystem accessible to business users, data scientists and IT professionals. That approach speeds time to value and equips organizations to surface the invaluable insights that lay hidden inside a data lake for years to come.

About Informatica

Informatica

Informatica is 100 percent focused on data because the world runs on data. Organizations need business solutions around data for the cloud, big data, real-time and streaming. Informatica is the world's no. 1 provider of data management solutions, in the cloud, on-premise or in a hybrid environment. More than 7,000 organizations around the world turn to Informatica for data solutions that power their businesses. For more information, visit www.informatica.com

About Trianz

TRIANZ℠ Execution Matters.

Trianz enables digital transformations through effective strategies and excellence in execution. Collaborating with business and technology leaders, we help formulate and execute operational strategies to achieve intended business outcomes by bringing the best of consulting, technology experiences and execution models. Powered by knowledge, research, and perspectives, we serve *Fortune* 1000 and emerging organizations across industries and geographies to transform their business ecosystems and achieve superior performance by leveraging Cloud, Digital, Analytics and Security paradigms. As a professional services firm, our values and culture are focused on delivering measurable business impact, predictability in execution, and a unique partnership experience.

Silicon Valley | Washington DC Metro | Jersey City | Dubai | Bengaluru | Mumbai | Delhi-NCR | Chennai | Hyderabad

www.trianz.com | sales@trianz.com | +1-408-387-5800

© 2017 Informatica LLC. All rights reserved. Informatica® is a registered trademark of Informatica in the United States and in jurisdictions throughout the world. All other company and product names may be trade names or trademarks.

@ Copyright 2017, Trianz. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the express written permission from Trianz. The information contained herein is subject to change without notice. All other trademarks mentioned herein are the property of their respective owners.