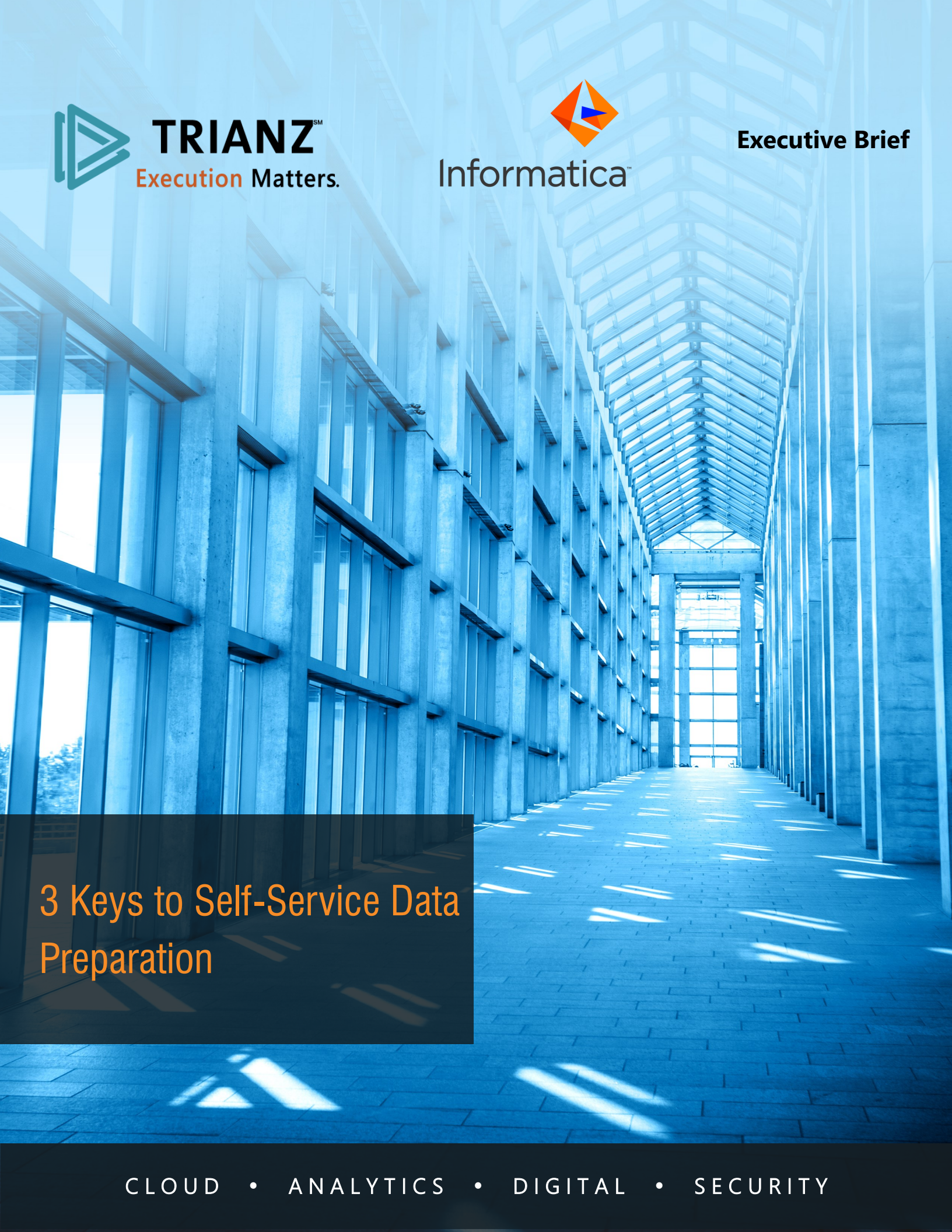

TRIANZ℠
Execution Matters.

Informatica™

Executive Brief

3 Keys to Self-Service Data Preparation

CLOUD • ANALYTICS • DIGITAL • SECURITY

3 Keys to Self-Service Data Preparation

Liberate resources from time-consuming data chores to focus on high-impact data lake analytics

The Black Hole of Data Preparation

Suppose you spent 80% of your time having your car repaired, and just 20% of your time driving it? Or, if it took you six hours each workday to get your laptop up and running, leaving you just two hours to actually use it?

That sort of time disparity is not uncommon with analytic data lakes. Data analysts and data scientists spend up to 80% of their time on essential data preparation — working to generate accurate, complete and standardized data that's ready for analysis in a data lake. Alternatively, data analysts may offload data preparation to an already overworked IT team.

But that introduces lengthy delays and additional costs. In either case, data preparation amounts to a black hole that consumes time and resources. It's a major obstacle for organizations that aim to use data lakes for data-driven insights that can transform customer experiences, supply chain efficiency, profitability analysis, fraud detection and more.

To combat the problem, leading organizations are transitioning toward data preparation methodologies distinguished by self-service, collaboration and sound governance. The objective is to liberate resources from the time-consuming chore of data preparation to focus on value-add analytics, while improving the reliability and quality of data for analysis.

Addressing the Growing Complexity of Data Preparation

Data preparation comprises the nuts-and-bolts processes needed to get data ready for analytics in a data lake, typically based on the open source Apache Hadoop framework. The data preparation task is complicated by the fact that data lakes can include both structured and unstructured data from multiple sources.

Those sources can range from conventional ERP, CRM, Excel and other business applications to mobile devices, social media, clickstream and server logs, and sensor-based equipment from the Internet of Things. According to Gartner, the challenges of larger and more diverse data "have made data preparation one of the biggest roadblocks to pervasive and trusted modern analytics."[1]

As data volumes and complexity grow, data analysts, data scientists and IT spend more time on such data preparation tasks as deduplicating records, addressing empty fields, incorporating metadata and converting currencies. They also need to standardize different treatments of dates, addresses and names, and transform data from multiple sources into a uniform format.

Conventional manual approaches to data preparation through spreadsheets, custom coding and scripting are time-consuming and not scalable amid growing business demand for timely analytics. As it is, data analysts using traditional methods spend between four and six hours a day on data preparation, a survey by the analyst firm Blue Hill Research found.[2]

Beside inordinate time and cost, manual data preparation can result in such problems as:

- Inability to derive analytic value from data
- Errors and inconsistencies resulting from manual work
- Poor accessibility and visibility into disparate data
- No trusted, authoritative set of data lake assets
- No reusability of manual, one-off data preparation efforts

[1] *Gartner, "Market Guide for Self-Service Data Preparation," August 25, 2016.*

[2] *Blue Hill Research, "Quantifying the Case for Enhanced Data Preparation," February 2016.*

Ultimately, organizations run the risk of turning data lakes into "data swamps" full of information that's incomplete, inaccurate and lacks context.

Embracing Best-Practice Models and Technologies

The barriers imposed by traditional approaches have given rise to a new strategic focus on more efficient and collaborative data preparation. Leading organizations are embracing best-practice data preparation models and a relatively new class of self-service data preparation tools distinguished by straightforward ease of use in a cloud-based framework.

As Gartner put it, "The escalating challenges associated with larger and more diverse data are contributing to the demand for adaptive and easy-to-use data preparation tools. Self-service data preparation addresses these challenges by accelerating the way users find, access, clean, combine, model and transform, and collaborate with data in an agile yet trusted way."

Technology-driven self-service data preparation is increasingly in use at leading organizations, and delivering significant benefits. For instance:

- 25% of organizations now use self-service data preparation tools[3]
- By 2020, more than 50% of new data integration efforts for analytics will utilize self-service data preparation tools[4]
- Business benefits increase 2x at organizations that provide agile, curated datasets for a range of content authors[5]
- 79% of those using self-service data preparation tools accelerate their processes[6]

Another key benefit is reducing the burden that data preparation imposes on IT. When data analysts are set up for self-service data preparation, IT can devote more time and resources to value-add initiatives. Organizations also eliminate the days or weeks that IT may require for data preparation that analysts were unable to perform themselves.

3 Key Focus Areas for Effective Data Preparation

A robust data strategy goes hand in hand with purpose-built data preparation technology for organizations to eliminate the struggles of conventional data preparation and help maximize value from a data lake.

A data strategy typically has two phases. The first is an evaluation and gap analysis of current internal and external data assets needed for a data lake. The second is planning and designing for a data lake that meets business objectives from a near-term, one-year, three-year and longer term perspective.

This effort should involve stakeholders in both business and IT in assessing the overall data landscape and refining goals for the data lake. It should be anchored by sound data governance and include evaluation of data management approaches, such as data catalogs, data lineage, data relationships, business glossaries and data protection.

Organizations should baseline present-day data preparation efforts and quantify time investments by data analysts, data scientists and IT personnel. Identifying critical pain points and limitations is instrumental in mapping out a new approach, and tracking benefits accrued once the initiative is in place.

[3] *Blue Hill Research, "Quantifying the Case for Enhanced Data Preparation," February 2016.*
[4] *Gartner, "Market Guide for Self-Service Data Preparation," August 25, 2016.*
[5] *Gartner, "Market Guide for Self-Service Data Preparation," August 25, 2016.*
[6] *Blue Hill Research, "Quantifying the Case for Enhanced Data Preparation," February 2016.*

In both modeling data preparation processes and evaluating potential use of standalone data preparation technology, three key focus areas are:

1) **Self-service**
2) **Collaboration**
3) **Governance**

Self-Service

Equipping data analysts and scientists with self-service capabilities can dramatically reduce the time spent on data preparation by both end-users and IT professionals. That frees resources to focus time and energy on analytics that impact business performance. Self-service data preparation is fast becoming a requirement for data lake analytics at speed and scale.

Today's best data preparation tools provide an Excel-like, browser-based interface that combines ease of use with robust functionality. Data analysts can explore, blend, enrich and standardize data. They can readily search and discover data, understand its provenance, and find related data. They can also identify data attributes and lineage, and correct errors, resolve duplicates, and validate names and addresses.

Data may then be moved into a Hadoop-based data lake, or made accessible to business analytics tools like Qlik or Tableau. The result is more time for analytics, and information that is more complete, accurate and reliable.

Collaboration

Conventional data preparation often consists of ad hoc, tactical efforts by data analysts, data scientists and IT. These efforts are rarely well coordinated across the larger enterprise, and may rely on the email exchange of spreadsheets. The results can include duplication of data preparation efforts, and teams starting from scratch without the benefit of best practices and lessons learned by other practitioners.

A collaborative environment for users to prepare, publish, iterate and share data should be a key element in self-service data preparation. All stakeholders in data preparation benefit with a community-oriented platform geared for knowledge and resource sharing. That can include common project workspaces that enable viewing and sharing of data assets, sources and data relationships across multiple domains and dimensions.

Governance

Data preparation governance is an essential focus area that rewards stakeholders with greater visibility, control and trust in data. Governance depends on establishing roles, rules and operationalized processes to streamline and continuously optimize data preparation, from start to finish. Governance in the context of data preparation should work in tandem with a broader governance framework in place for a data lake.

Done right, data preparation governance surfaces and enriches the metadata of data assets for greater transparency. Data lineage will illustrate source, time of origin, revisions by whom and more. Governance will also reflect data usage and user activity, and will highlight gaps that should be addressed.

In a governance framework, an information catalog can drive user adoption by enabling IT and business users to understand data lake contents and dive into data exploration. These insights are important to building confidence that data is accurate and ready for analytics.

Transform Data Preparation with Trusted Guidance and Technology

Trusted strategy advisors and implementation partners help organizations transform resource-intensive data preparation into a straightforward self-service discipline that provides a foundation for effective data lake analytics. With more than 2,000 client engagements over 15 years, Trianz partners with Informatica to provide proven best practices and market-leading technology to accelerate and enrich data preparation.

Trianz helps organizations develop data strategy (capability assessment, roadmap planning, big data

and cloud readiness, operating model design) and implement data management capabilities that are foundational for self-service data preparation. The joint Trianz-Informatica solution combines Trianz's expertise in enterprise information management with Informatica Intelligent Data Lake, a self-service data prep tool, which supplies industry-leading capabilities to ingest, find, prepare and protect data for analysis inside an intelligent data lake.

With Trianz and Informatica, your organization can eliminate the time and headaches of data preparation and speed your evolution into a data-driven business.

About 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 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 | info@trianz.com | +1-408-387-5800