informatica

White Paper

# How Big Data Management Turns Petabytes into Profits

## Successful Big Data Projects Begin with Three Essential Pillars

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Informatica does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice.

The incorporation of the product attributes discussed in these materials into any release or upgrade of any Informatica software product—as well as the timing of any such release or upgrade—is at the sole discretion of Informatica.

Protected by one or more of the following U.S. Patents: 6,032,158; 5,794,246; 6,014,670; 6,339,775; 6,044,374; 6,208,990; 6,208,990; 6,850,947; 6,895,471; or by the following pending U.S. Patents: 09/644,280; 10/966,046; 10/727,700.

This edition published September 2015

# Table of Contents

# Executive Summary

Now that the industry has nearly a decade of experience with big data—surrounded by far too much hype about data volume, variety, and velocity—it's time to finally catch up to the promise. After years of trial and error, businesses should now be ready to define a big data strategy that delivers tangible and real results—and ultimately achieves increased ROI.

Big data integration and management form the foundation for success. With the right approach, you can bring together disparate and complex information from multiple sources and turn it into trusted information assets at scale to drive competitive advantage.

This white paper describes a big data management platform that relies on three fundamental pillars to deliver expected business results from big data projects:

- Dynamic and optimized big data integration
- End-to-end big data governance and quality
- Risk-centric big data security

# Why Aren't We There Yet?

One of the biggest questions businesses have asked over the past year or so is: "When will the promise of big data become reality?" We should be ready to make significant steps forward, right?

Industry analysts and the press have focused on this heavily over the past couple of years. For example, a researcher at Wikibon recently wrote: "Big data has established itself as one of the key competitive differentiators in the emerging digital economy.[1]" Wikibon notes that big data is being used by a crop of new digital businesses as well as by traditional enterprises.

This should be good news—and is the typical path for new technologies (think cloud computing, in-memory analytics, data virtualization, etc.)—which are initially over-hyped and gradually mature and catch up to what has been promised.

## Pent-up demand from the C-suite

CIOs and Chief Data Officers (CDOs) we've talked with over the past few months tell us they're ready to take on big data projects and prove success, saying things like:

- "We have so much knowledge in data—it's just huge. We need data cleaning and analytics to pull this knowledge out."
- "I worry that our CEOs/CFOs will soon start complaining that the information garnered from big data didn't really make them more money. To make more money, you need to connect the dots between transactional systems, BI, and the planning systems."
- "We need to be proactive and cut the time to execution. The value we derive from our data needs to enable the enterprise to generate value different than our competitors."
- "'What are we going to do to differentiate ourselves?' is a big driver for us. Our competitors have been at this business for years. We aren't hiring two hundred people like they are. So we're differentiating ourselves by personalizing the services we offer."
- "Often the data's not available because there's no business process to capture it. Or even if there is one, the data doesn't represent what everyone thinks it does."

[1] Wikibon, "Big Data: From Promise to Reality." May 1, 2015.

Without question, the key big data challenges for most businesses today are how to get started and how to achieve quick and measurable results.

The Problem with Big Data Projects

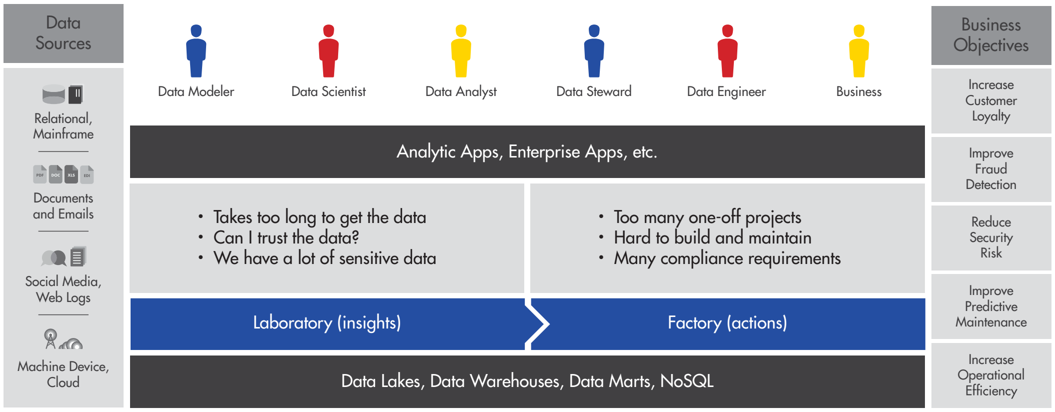
Data Sources
Relational, Mainframe
Documents and Emails
Social Media, Web Logs
Machine Device, Cloud
Data Modeler
Data Scientist
Data Analyst
Data Steward
Data Engineer
Business
Analytic Apps, Enterprise Apps, etc.
Takes too long to get the data
Can I trust the data?
We have a lot of sensitive data
Too many one-off projects
Hard to build and maintain
Many compliance requirements
Laboratory (insights)
Factory (actions)
Data Lakes, Data Warehouses, Data Marts, NoSQL
Business Objectives
Increase Customer Loyalty
Improve Fraud Detection
Reduce Security Risk
Improve Predictive Maintenance
Increase Operational Efficiency

Figure 1. Businesses are challenged by how to get started and how to achieve results with big data

# Why Do Many Big Data Efforts Fail?

Big data promises to allow new forms of analysis and insights to emerge as the volume and variety of data increases. Once you determine an insight that has value, you can make a business decision based on ROI analysis to further invest in monetizing the insight to meet a specific business outcome.

Despite the high business value that can be derived from big data, many initiatives fail for a number of reasons. Businesses keep big data in silos across separate business units, which prevent companies from seeing a total view of customer information or other assets across the company[2]. Some businesses also use traditional data technologies that were not designed to solve big data problems[3].

Big data holds a lot of promise for businesses, but enterprises are becoming disillusioned because they're not seeing immediate success. According to research by Capgemini Consulting[4], 60 percent of executives surveyed believe big data will upend their industries within three years. But a mere 8 percent describe their own projects as "very successful," and only 27 percent call their efforts "successful."

For many businesses, the sheer volume of data that is available is often incomplete, inconsistent, ungoverned, and unprotected—leading to negative or even disastrous outcomes. Traditional solutions can be expensive, manual, and complex with business analysts waiting weeks to get useful data for analytics. The time to operationalize an analytic insight into actionable results can easily exceed the time and scope of the allocated budget.

[2] InformationWeek, "8 Reasons Big Data Projects Fail." August 7, 2014

[3] Ibid

[4] Capgemini Consulting, "Cracking the Data Conundrum: How Successful Companies Make Big Data Operational."

## Big Data Success Driven by Automation and Standardization

Smart organizations are discovering that a one-off, hand-coding approach to big data management does not lead to repeatable, reliable, and maintainable production environments. On the contrary, this approach sets them up for spending most of their time rewriting and maintaining the code as things change. Organizations need help to quickly and efficiently integrate data into big data platforms. They also need tools to prepare the data using readily available staff who don't require specialized Hadoop experience.

The companies that successfully deliver business value with big data have put in place "self-service autonomy" so that data scientists and analysts don't have to wait for fully certified data to iterate through hypothesis testing and model validation. Meanwhile, automation and standardization enables scalable and flexible data pipelines to meet the SLAs of the business—delivering data directly into the business applications used by employees and customers, in real time if necessary.

> "The benefit of using standard tools for data integration was that we could leverage over a hundred of our ETL developers to build data pipelines on Hadoop without having to know anything about Hadoop. The savings were huge."
>
> – Executive Director of Data Warehouse Engineering at a large global bank

## Data Management: Two Sides of One Coin

One example of a company that is gaining business value with self-service autonomy and operational agility is at a large North American conglomerate with more than 24 business units and companies. It adopted a big data management platform to integrate, govern, and secure all data onto a central data management infrastructure.

Data comes from thousands of sources with evolving schemas and varying reliability. Each business unit produces and consumes massive amounts of data through data interchanges—with redundant efforts in gathering and using external data.

The company views big data management as two sides of one coin: on one side is the laboratory, where analysts can autonomously discover and prove value; on the other is the factory, where IT and Operations can build data products that scale. When an analyst has an idea, he or she works with IT to get the data loaded so that the analyst can do his or her own self-service discovery.

The company's business depends on delivering timely and trusted information to customers, so it's critical that its IT team can detect data quality problems and anomalies early on. The company now has the tools to profile large data sets at scale and perform continual audits using data quality scorecards and dashboards. It can use existing ETL developers to build data pipelines on Hadoop, which is much more beneficial than hand coding.

As this example demonstrates, managing big data sustainably and repeatedly for multiple projects requires an intelligent data platform that delivers clean, safe, and connected data throughout the enterprise.

"By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."

– McKinsey & Company, "Big data: The next frontier for competition."

## The Big Data "Laboratory" vs. Big Data "Factory"

Many companies starting their big data journey don't realize that big data requires people with a diverse set of skills and talent to be successful—including data scientists, data engineers, architects, and subject matter experts. The first group of talent is more scientific, primarily focused on analyzing data and discovering insights in what we'll call the "big data laboratory." Typically, they embark on a highly iterative process where nine out of ten discoveries may have no value at all.

The second group is more engineering and IT-centric, focused on architecting, building, and maintaining consistent, reliable, and trusted data pipelines that deliver actionable insights to the business. They work in what we'll call the "big data factory." Their processes can be difficult to scale and maintain in production data centers.

And both groups must adhere to security policies and industry regulations. The ability to integrate, govern, and secure data at scale gives data scientists and the business confidence in the data they are analyzing and using to drive successful business outcomes. A big data management platform is the foundation for self-service autonomy that enables data scientists to discover insights faster, and delivers operational agility that turns raw data into actionable information.

### Requirements for Successful Big Data Analytics

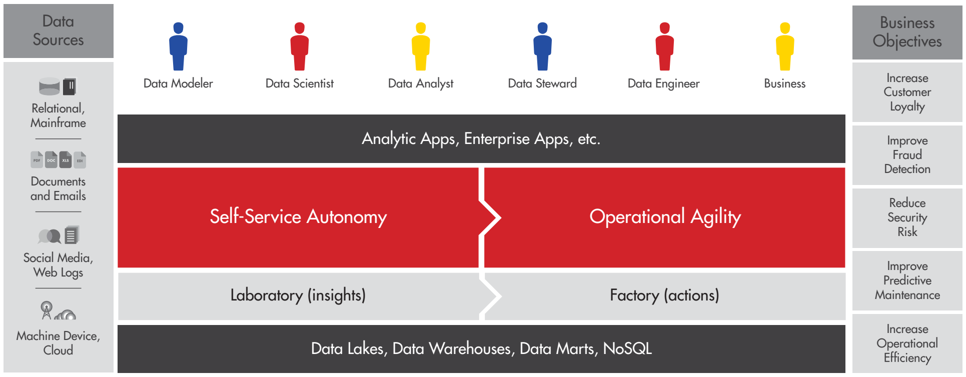

Figure 2. Requirements for success: Self-service autonomy and operational agility

# Three Essential Pillars. The Importance of Data Integration, Governance, and Security

For big data management to truly be effective, you need to start with a platform that delivers three key elements:

1. Dynamic and optimized big data **integration**
2. End-to-end big data **governance and quality**
3. Risk-centric big data **security**

In the early days of big data, most of the investment was in Hadoop and data analysis and less on data governance. But as businesses have begun building complex architectures for big data, challenges around data governance and data privacy have increased[5]. Now—as big data strategies mature—we're seeing more interest in comprehensive big data management platforms that handle data integration, data governance, and data security for multiple projects across the enterprise. And although businesses are still intrigued by Hadoop, investment in the technology remains tentative.

"Gartner's 2015 Hadoop adoption study results show that investment remains tentative in the face of sizable challenges around business value and skills."

Gartner, "Survey Analysis: Hadoop Adoption Drivers and Challenges," Nick Heudecker, Merv Adrian. May 12, 2015.

## Integration

Big data integration should deliver high-throughput data ingestion and at-scale processing so business analysts can make better decisions using next-generation analytics tools.

Big data integration helps businesses gain better insights from big data because it:

- Speeds up development, leverages existing IT skills, and simplifies maintenance through the use of a simple visual interface supported by easy-to-use templates
- Increases performance and resource utilization by optimizing data processing execution and providing  flexible, hybrid deployment across a variety of platforms
- Handles a wide variety of data sources though hundreds of pre-built transforms, connectors, and orchestrates data flows by using broker-based data ingestion

## Governance and Quality

End-to-end big data governance and quality means business and IT users can be confident with the data they're using. Look for comprehensive data governance that includes:

- Formal data quality assessments to detect data anomalies sooner
- Pre-built data quality rules to ensure data is "fit-for-purpose"
- Universal metadata catalog to facilitate search and automate data processing
- Entity matching and linking to enrich master data such as for customers
- End-to-end data lineage for data provenance, traceability, and compliance audits

The need for governance and security is greater with big data because of the larger volume and variety of data that originate from multiple sources, collected in one place (typically in Hadoop-based data lakes), and proliferated across many target systems. All this data needs to be governed and secured.

[5] EY, "Big data: Changing the way businesses compete and operate." April 2014

### Security

Risk-centric big data security analyzes all data to quickly detect and act upon risks and vulnerabilities. This requires a 360-degree view of sensitive data, supported by risk analytics and policy-based protection of data at risk. Big data security should de-identify information controlled by corporate policies and industry regulations. Risk-centric big data security must enable:

- “Single pane of glass” monitoring of sensitive data stores to provide visibility into the locations of sensitive data
- Sensitive data discovery and classification for a comprehensive 360 view of sensitive data
- Usage and proliferation analysis for a precise understanding of data risk
- Risk assessment to help prioritize investments in security programs
- Non-intrusive persistent and dynamic data masking to protect sensitive data in development and production environments to help minimize the risk of a security breach

Three Pillars of Big Data Management

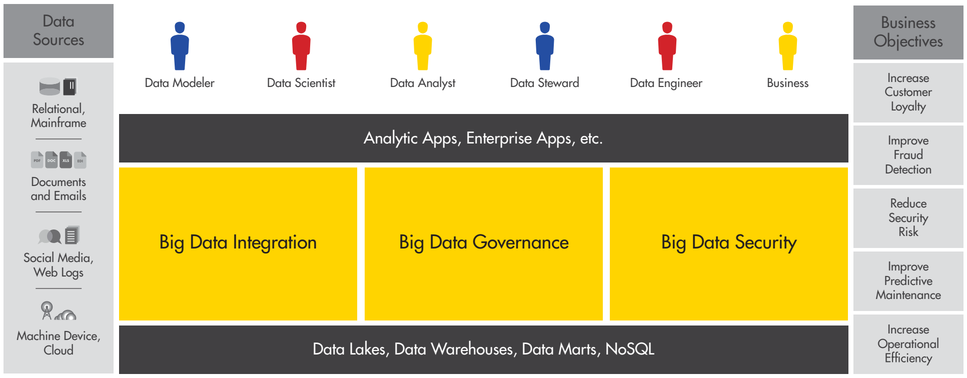

Figure 3: Data integration, governance, and security are three essential pillars of Big Data Management

## Trustworthy Data is Essential to Achieving ROI

It’s not an easy task to quantify the ROI of big data initiatives. This is largely because in the beginning, big data projects typically start off as “science experiments” in the lab.

It’s best to start small and get some quick wins—focusing on the journey to ROI without trying to do everything all at once. By gradually curating trustworthy data, you can make the right decisions at the right time, get relevant data into the hands of key decision-makers, and ultimately achieve tangible business results.

Data scientists need to explore the data for relationships, patterns, and insights. While at the same time the business expects actionable information it can use immediately to drive results. Look for a big data management platform that can take raw data and make it “fit-for-purpose” and easily accessible for data scientists and analysts to discover insights. Data engineers should be able to build scalable data pipelines that are easy to deploy and maintain and take raw data and convert it into something that is readily consumable and actionable by the business (following a process of collect, prepare, cleanse, integrate, and deliver).

6 KEY BUSINESS OBJECTIVES

Here are six of the most common business objectives for big data projects today. View the "Big Data Management in Action" section for customer examples.

- Increasing customer loyalty
- Increasing operational efficiency
- Reducing security risks
- Improving fraud detection in financial services and insurance
- Improving predictive maintenance in manufacturing
- Lowering the total cost of care in healthcare

### Increasing Customer Loyalty—A Sample ROI Scenario

Let's take a quick look at how big data management can help reduce the costs of gathering, integrating, governing, and securing data—while delivering results that can significantly increase revenue:

- You can reduce the cost of gathering massive amounts of raw data from web logs, customer transactions, marketing campaign data, and other sources by using pre-built adapters and parsers for storing data in Hadoop, thereby reducing storage costs.
- You can dramatically cut the time and cost of preparing the data for analysis by using purpose-built tools with pre-built transformations instead of hand-coding using expensive developers with specialized hand-coding skills. The cost of governance is reduced with pre-built data quality rules and the ability to match and link data sets at scale to enrich customer master data with consumer behavior.
- You can avoid the cost and bad publicity associated with security breaches by understanding where all your sensitive PII/PHI customer data resides and masking it to de-identify and de-sensitize the data.
- And finally, you can build a data pipeline that is easy to maintain and delivers information directly to consumers' mobile devices—optimizing their personal engagement and increasing loyalty thereby increasing revenue and market share.

## Big Data Management in Action: Customer Examples

Here are a few examples of how businesses are deriving value from their data by integrating, governing, and securing it at scale.

### Large Bank Significantly Enhances Fraud Detection and Improves ROI

A major bank with $403 billion in assets and 18.5 million customers wanted to significantly reduce the time it was taking to catch fraudulent events as they happened. Specifically, it wanted to bring general ledger data to the Enterprise Landing Zone (ELZ) to harmonize and standardize data before feeding it into anti-money laundering (AML) and fraud detection applications. The bank also wanted to bring other domains like customer banking, cards, mortgage, wire transfers, Salesforce, Siebel, and other ERP data to the ELZ to further improve fraud analysis.

Informatica big data management helped the company significantly reduce mainframe costs and speed up AML and fraud detection processing. The new platform helps establish and enforce AML data preparation rules, identify changes to data sets, tag log files for audit purposes, execute ETL transforms, and create aggregated data sets.

## Western Union Creates Enterprise Data Hub to Help Identify Trends and Enhance Customer Experience

Ecommerce giant and money lender Western Union wanted to develop an omni-channel marketing approach that integrated retail, web, and mobile and would help it expand into new markets with digital products. The firm sought to reach customers with a more tailored and personalized experience, while also reducing risk. Western Union built a new big data platform based on Hadoop and Informatica Big Data Management to help the company identify trends and analyze data from multiple diverse sources (legacy, online, and mobile). Western Union can now quickly evaluate customer preferences and buying patterns to enhance the overall customer experience.

Find out more about Western Union and Informatica

## Leading Insurance Firm Relies on Unified Big Data Platform to Power Marketing Campaigns

With nearly 20 million customers, a leading insurance firm wanted a 360-degree view of all consumer activity for improved marketing, planning, and analytics. It sought to discover and mine relationships and to create highly targeted and personalized campaigns.

Many data sources needed to be integrated, cleansed, and matched at scale from a variety of systems. Data sources for marketing include customer profile data, Salesforce CRM, prospect and partner data, solicitation history, web logs, and social media data.

Informatica provided a single big data management platform that delivers a consistent enterprise-wide view across all business units. The platform enables rapid intake of new data sources, both structured and unstructured, and eliminates data pipeline bottlenecks while increasing processing power for statistical analytics. Insurance is a highly regulated industry, so the platform also supports data governance with tools and processes to profile data, validate data quality, capture metadata, provide end-to-end data lineage, and ensure security.

## Healthcare Insurer Improves Patient Outcomes and Lowers Cost of Care

A North American provider of managed care services to about 4 million members uses big data management to achieve its key business goals: improve patient outcomes, lower total cost of care, increase member retention, and grow membership.

The company wanted to modernize its analytics infrastructure to help improve provider collaboration, member engagement and retention, and enhance compliance reporting and quality outcomes reporting. This was not an easy task given increasing data volumes, regulatory compliance, and data coming from various data domains ranging from claims to clinical data domains.

The insurer decided to implement a more agile and flexible logical data warehouse strategy using Informatica for Big Data Integration and Quality on Hadoop. The company expects to deliver more timely and trusted data that will help it improve provider collaboration and member engagement, provide better outcomes at lower cost to remain competitive and grow market share, increase member retention and growth, and reduce medical costs.

### Large Oil and Gas Company

A major oil and gas company is building a logical data warehouse using Informatica Big Data Management and a Hadoop-based data lake that supports critical capabilities such as data ingestion, metadata, data integration, data quality and data governance, master data management, information security, and information sharing and reusability.

The company is among the largest U.S.-based independent natural gas and oil producers. Its business users need fast, accurate data at their fingertips to make smart, quick decisions. Informatica Big Data Management makes manual data entry and reconciliation a thing of the past by providing authoritative and trusted data as it relates to wells, suppliers, and other key master and reference data.

Informatica accelerates time-to-value with readily available skills to build data pipelines in Hadoop that transform, prepare, and deliver data to the company's big data analytic applications. This improves the automation of many core business tasks such as fracking process efficiency, dispatching drivers out to locations more efficiently and improving decision support during well construction.

## Why Informatica for Big Data Management?

With 20-plus years of innovation and leadership in data management, Informatica delivers a proven platform for comprehensive big data management. We've been named a "leader" by both Gartner[6] and Forrester across multiple categories of data integration, data quality, MDM, data security, and cloud integration. By combining the best of open source technology with our own proven methodology, we provide a comprehensive intelligent data platform designed for big data management.

Backed by a strong partner ecosystem and global network of big data experts, we can help you accelerate and improve the success of your big data initiatives, gain some quick wins, and establish a solid foundation to continuously deliver business value from your data assets.

## Get Big Data Ready

Informatica can help you adopt a strategy to use big data management as part of your overall enterprise information architecture plan—to deliver results with speed, confidence, and repeatability.

We interviewed Informatica customers who have built a new data organization that turns their data into a competitive advantage. Read the eBook "How to organize the data-ready enterprise," to discover the seven key principles for building an organization that creates great data for real business value.

### About Informatica

Informatica is a leading independent software provider focused on delivering transformative innovation for the future of all things data. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. More than 5,800 enterprises depend on Informatica to fully leverage their information assets residing on-premise, in the Cloud and on the internet, including social networks.

*Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.*

[6] Gartner, 2015 Magic Quadrant for Data Integration Tools, Eric Thoo, Lakshmi Randall, 29 July 2015.

Gartner, 2014 Magic Quadrant for Data Quality Tools, Saul Judah, Ted Friedman, November 26, 2014

Gartner, 2014 Magic Quadrant for Master Data Management of Customer Data Solutions, Bill O'Kane, Saul Judah, October 30, 2014.

Gartner 2014 Magic Quadrant for Data Masking Technology, Joseph Feiman and Brian Lowans, December 10, 2014

Gartner, 2015 Magic Quadrant for Enterprise Integration Platform as a Service, Worldwide, Massimo Pazzini, et al. March 23, 2015.

informatica

Put potential to work.™

Worldwide Headquarters, 2100 Seaport Blvd, Redwood City, CA 94063, USA Phone: 650.385.5000 Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871 informatica.com linkedin.com/company/informatica twitter.com/Informatica