INFORMATICA®

WHITE PAPER

The Safe On-Ramp to Big Data

Lower Costs, Minimize Risk, and Innovate Faster with
a Proven Approach to Big Data

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 March 2013

Table of Contents

Executive Summary 2

Perceptions of Risk and Reward in Big Data 2

A Bullish Approach to Big Data Opportunities 4

Reduce Big Data Costs by Up to 2x or More 5

Minimize Big Data Risk 6

Innovate with Big Data Up to 3x Faster 7

Conclusion 9

Executive Summary

Risk: It's the enemy of IT projects. Big data can make risk seem even bigger.

Emerging big data technologies coupled with huge spikes in data volumes, velocity, and variety put risk squarely in the spotlight, raising difficult questions of risk and reward. IT and business managers wary of spiraling costs and busted deadlines with conventional IT projects can be even more apprehensive about putting organizational resources and their own career track on the line without a proven path to big data value.

And indeed the risk in big data projects can be substantial, both in terms of project missteps if not the dreaded "epic fail," as well as lost opportunities to deliver business value and competitive advantage.

Big data risk and cost can be minimized with the right approach and proven technology to address the bulk of big data work—integrating disparate and complex information from a growing array of sources. This white paper examines the risk and reward of big data and highlights how leading organizations are utilizing Informatica technology as a safe on-ramp to big data that reduces risk and cost while speeding insights and innovation.

Perceptions of Risk and Reward in Big Data

Big data has generated big interest for organizations in virtually every industry. Most enterprises recognize the need to cost-effectively manage growing volumes of big transaction data—the conventional information in on-premise and cloud enterprise applications, data warehouses, and storage environments.

The other flavor of big data—big interaction data—opens a new frontier for business opportunity. Innovative enterprises are already leveraging vast quantities of big interaction data from social media, machines and sensors, call detail records (CDRs), geospatial systems, the web, and other sources. By combining these new data types with conventional data, leaders are increasing customer retention and acquisition, increasing operational efficiencies, improving product and service delivery, and generating breakthrough results with new insights.

Yet a sizable number of organizations are taking a conservative wait-and-see approach to big data. An Informatica survey of 589 IT and business professionals found that 31 percent had no plans to pursue big data initiatives, at least in the near term.[1] More than half (52 percent) of survey respondents identified a lack of maturity in big data tools as the top challenge in big data projects.

For instance, the open-source Hadoop framework lacks support for reusability and metadata, making it difficult to extend projects and ensure consistency. Emerging technologies such as Hadoop are still maturing with capabilities for data management and analytics in a heterogeneous environment, prompting some organizations to resort to manual scripting and introducing the risk that big data projects will end up as another data silo.

[1] Informatica, "Balancing Opportunity and Risk in Big Data: A Survey of Enterprise Priorities and Strategies for Harnessing Big Data," May 2012.

As shown in Figure 1, other key challenges cited by respondents including lack of support for real-time data and concerns over poor data quality, security, and privacy. The survey also reveals perceptions of risk in the limited availability of skilled developers to manage big data, and in potentially difficult and time-consuming development in Hadoop, NoSQL, and other new technologies.

Key Big Data Challenges

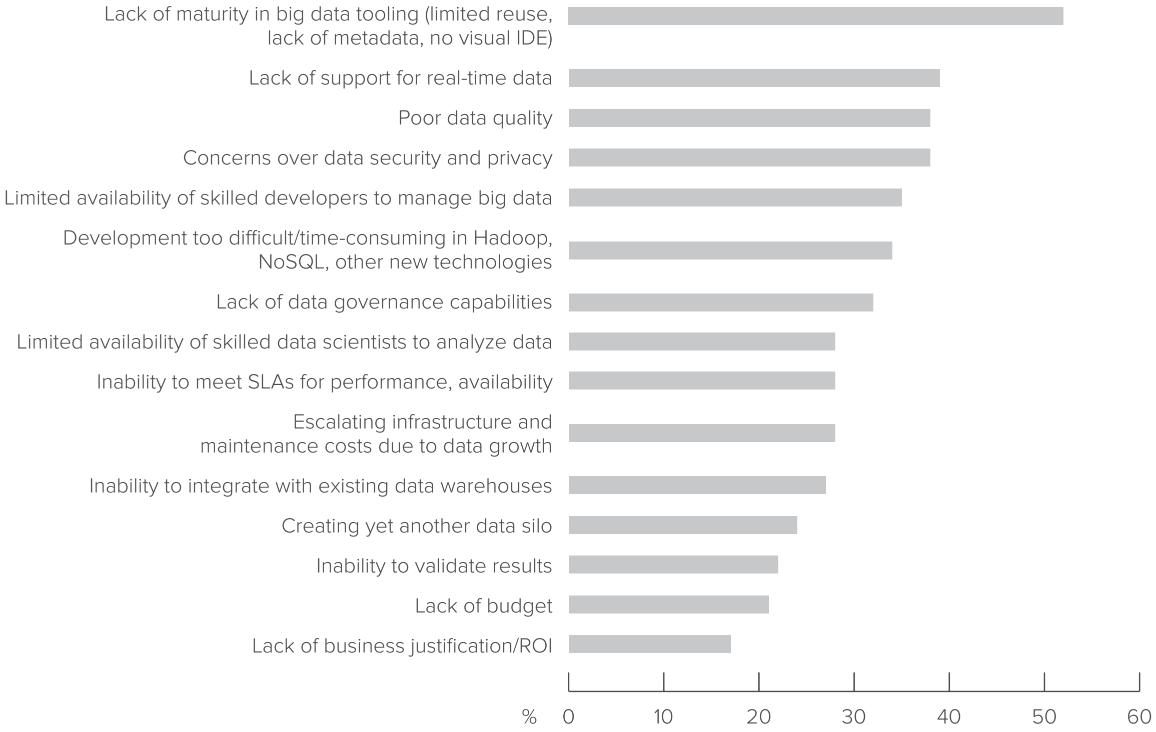

Figure 1: What key challenges do you face or foresee in managing big data?

A Bullish Approach to Big Data Opportunities

Despite those cautionary organizations on the sidelines, the majority (69 percent) is moving ahead with big data projects, with many in production or testing/pilot phases, the Informatica survey found. Overall, optimism is high—67 percent view big data as more of an opportunity than a challenge, and just 17 percent foresaw difficulties building a business case and mapping return on investment (ROI) for big data projects.

From conversations with customers and interactions at conferences and on social media, Informatica has found that those organizations most bullish on big data have typically equipped themselves with technologies and best practices that address the top challenges identified in our survey. The Informatica Platform is a key foundational element for leading big data innovators, as experts and researchers have found that roughly 80 percent of the work in big data projects focuses on data integration and quality (with the remaining 20 percent on analytics).

The Informatica Platform is the safe on-ramp to big data that works with both emerging technologies and traditional data management infrastructures to minimize risk, reduce costs, and fuel innovation. With Informatica, your IT organization can rapidly create innovative products and services by integrating and analyzing new types and sources of data. Informatica addresses the following key big data challenges so you can reap the benefits of big data:

Limited availability of skilled developers. The Informatica developers in place at an organization can readily apply their skills to big data projects without knowing Hadoop or hand-coding MapReduce scripts, while a global force of more than 100,000 developers trained on Informatica makes it easy to staff big data projects.

Lack of technological maturity. Informatica supplies a no-code, visual development environment to engineer data pipelines between all types of data sources, enterprise applications, Hadoop, data warehouses, and other sources and targets, with a metadata foundation and cross-project reusability, addressing functional limitations in many big data technologies.

Lack of real-time data. Informatica powers real-time data transformation and replication in heterogeneous environments with such technologies as data streaming, messaging, complex event processing, data virtualization, changed data capture (CDC), and high-speed data replication.

Poor data quality. Informatica enables organizations to achieve the ideal of comprehensive, trusted data across big data and traditional sources with data quality and profiling tools accessible to both business analysts and IT professionals.

Data security and privacy. Informatica technology dynamically masks production and non-production data to strengthen data privacy, comply with privacy regulations, and reduce the risk of a data breach. Data virtualization can provide a data-level security layer to Hadoop.

Leading organizations are capitalizing on best practices and proven technology to reduce big data costs by up to 2x or more, minimize risk, and innovate up to 3x faster.

Reduce Big Data Costs by Up to 2x or More

Big data projects can be a minefield of escalating costs in terms of hardware investments, poor hardware utilization, and costly manual development that needs to be repeated for each project. Innovators are reducing big data costs by 2x or more with a flexible, standardized platform that enables universal data access and integration, optimal hardware utilization, and the reusability of data integration logic across multiple projects.

Optimize hardware investments. Informatica technology enables you to identify performance bottlenecks and unused data to get more from your data infrastructure. Data architects use Informatica to deploy big data processing on the highest-performance and most cost-effective platforms, from symmetric multiprocessing (SMP) machines to distributed platforms like Hadoop, or traditional grid clusters or data warehouse appliances. For instance, moving data processing from an appliance to a grid or Hadoop can minimize loads and extend appliance capacity.

Real-time data replication and changed data capture can offload up to 60 percent of processing from source systems, while data archiving lets you offload infrequently used data from warehouses or other sources to low-cost commodity hardware or Hadoop. These approaches optimize hardware utilization and enable you to avoid investments in additional infrastructure to accommodate growing data volumes, velocity, and variety.

Increase productivity up to 5x. In an effort to jump-start big data projects, some organizations have resorted to hand-coding and manual scripting in MapReduce, Pig, and other technologies, needlessly increasing both personnel costs and the time to deploy while introducing the risk of error, troubleshooting, and delay.

Informatica's no-code, visual, metadata-driven development environment increases developer productivity by up to 5x compared to hand-coding, while an extensive library of prebuilt logic for extraction, transformation, and loading (ETL), data parsing, data quality, data matching, and more supplies a rapid and proven solution to engineer a big data environment. Developers can easily access all types of data in traditional relational, legacy mainframe, on-premise, and cloud applications, as well as complex, unstructured, multi-structured, and industry-specific data such as web logs, JSON, XML, FIX, SWIFT, ACORD, HL7, HIPAA, and more.

CUSTOMER SUCCESS SNAPSHOT

A financial services firm used Informatica technology to determine that much of the data in its 200 TB data warehouse was unused, meaning the organization bore needlessly high costs for data storage. By archiving unused data in a lower-cost platform while maintaining its accessibility to the business, the organization avoided a $20 million investment in a data warehouse appliance-based system that had been under consideration. Data archiving translated into annual savings of up to $3 million for the organization by continuously archiving infrequently used data based on retention policies, while it has kept its warehouse running without hardware upgrades.

Minimize Big Data Risk

Risk is inherent whenever a new technology is deployed, from a desktop application to an enterprise data platform. In the case of the game-changing big data phenomenon, risk attends inaction as well—innovative organizations are already gaining competitive advantage over those that choose to sit on the sidelines. Between inaction, implementation misfires, and inadequate integration, many organizations undertaking big data projects face substantial risk.

In fact, the analyst firm Gartner predicts that most large enterprises will fall short of capitalizing on big data's potential in the next several years. "Through 2015, more than 85 percent of Fortune 500 organizations will fail to effectively exploit big data for competitive advantage." Gartner said.[2] Most organizations are ill prepared to address the technical and management challenges posed by big data, Gartner has said.

A single platform that combines emerging technologies like Hadoop with your existing infrastructure helps insulate your organization from risk in two respects. One, it minimizes the risk of high costs, delays, and subpar results when undertaking a big data project. Two, it positions you to leverage existing personnel resources rather than hunting for high-priced, in-demand talent with specialized skills. The Informatica Platform supplies proven flexibility and reach in equipping your organization to rapidly tackle low-risk, high-value projects while laying the foundation for a cross-enterprise infrastructure that maximizes return from big data.

Avoid big data silos. With near-universal connectivity to virtually any data source, Informatica positions your organization to ensure that Hadoop, NoSQL databases and other big data repositories do not end up as silos. Flexibility allows developers and architects to access and exchange disparate data types across hybrid environments of Hadoop, data warehouses and appliances, legacy systems, and specialized applications.

Strengthen data security, privacy, and quality. With data virtualization, Informatica enables a secure data access layer to Hadoop, while data masking allows organizations to mask sensitive data on the fly as data is retrieved from Hadoop to meet privacy requirements and minimize the risk of a security breach. Data quality tools help ensure that data is accurate, consistent, and timely, reducing the risk of misinformed business decisions.

Avoid high-priced specialized resources. As demand for big data talent grows, developers with expertise in Hadoop, MapReduce, and other emerging technologies can be hard to find—and costly. Informatica's data-agnostic platform allows organizations to deploy in-house Informatica developers to big data projects with no special training, using the same visual development environment and "design once, deploy anywhere" principles applied to conventional data warehouse and integration projects.

If additional resources are required, more than 100,000 developers around the world are trained on Informatica, making it relatively easy to locate needed skill sets. Informatica customers that experimented with open-source data integration because of its seemingly lower cost have found that resources could be difficult to find, productivity suffered, maintenance was difficult, and subscription licensing expenses turned out to be more costly than expected.

EHARMONY SUCCESS SNAPSHOT

The online dating site eHarmony slashed development time by implementing Informatica HParser to replace manual Ruby scripting that was needed to transform hierarchical JSON data in Hadoop for use by its data warehouse. Besides the productivity gain, eHarmony increased data throughput by 4x while speeding runtime by 75 percent, gaining scalability needed for an expected 5x increase in data volumes.

> *"Through 2015, more than 85 percent of Fortune 500 organizations will fail to effectively exploit big data for competitive advantage."*
>
> *– Gartner*

[2] Gartner, "Information Economics, Big Data and the Art of the Possible with Analytics," webinar, September 2012.

Innovate with Big Data Up to 3x Faster

Analytics and business intelligence (BI) is the key focus area for big data projects, with the potential to deliver breakthrough insights that can fuel innovation. Nearly four of five (78 percent) of Informatica survey respondents cited analytics/BI as their top project, double that of the next priority (master data management, at 39 percent).

To derive these insights, organizations are beginning to cultivate the discipline known as "data science." A data scientist combines technical skills in statistical modeling, data discovery, and data visualization with strong business acumen to interrogate data and generate insights to improve business performance. Yet many data scientists invest little time in value-added analytics, instead toiling away with manual approaches to data access, parsing, extracting, cleansing, and transforming data.

As D.J. Patil, chief data scientist at Greylock Partners and former data scientist at LinkedIn, notes in his book *Data Jujitsu*, "80 percent of the work in any data project is in cleaning the data." Meanwhile, a study of 35 data scientists at 25 organizations conducted by Stanford University and University of California at Berkeley researchers found they were inordinately consumed by "time-consuming and tedious" data access, manipulation, and integration—and spent far less time on actual analysis.[3]

One data scientist told the academic researchers, "I spend more than half of my time integrating, cleansing, and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any analysis." Said another, "The biggest challenges have been making two sources work together… it would be nice to have mappings and standardizations. You end up losing the data somewhere along the path."

With rapid development and streamlined data integration, Informatica relieves data scientists and other IT personnel of time-consuming manual work, helping address a key concern of IT and business professionals surveyed by Gartner. The analyst firm found that 30 percent of respondents viewed the scarcity of analytics capabilities and skills to be the top inhibitor to realizing benefits from big data.[4] Leading organizations use Informatica to help achieve up to a 3x acceleration in analytics, insights, and innovation:

Rapidly onboard and analyze any data type to gain insights. Informatica makes it possible to onboard data 50 percent faster than alternative approaches, with flexibility to ingest data at any latency, whether in batch, real time or near-real time. Prebuilt libraries to parse and transform virtually any data and native access to a range of data types, from traditional relational and legacy information to data from social media, devices, and web sources, accelerates value from big data.

GOVERNMENT AGENCY SUCCESS SNAPSHOT

A large government agency is using Informatica to minimize risk as it deploys Hadoop on a 32-node cluster, with Hadoop envisioned to serve as a "landing pad" for all data. Informatica will manage the movement of data in and out of Hadoop to a diversity of sources and targets, including a Teradata-based enterprise data warehouse and Oracle, SAS, and MarkLogic systems.

With this initiative, the agency is extending its deployment of Informatica as the integration layer for the 500 TB Teradata warehouse, managing loads of more than 10 billion records from disparate sources. Informatica provides the agency with a flexible architecture to meet rapidly changing business needs in a heterogeneous environment and an expected 3x to 5x increase in data volumes in a matter of a couple years.

[3] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, Jeffrey Heer, "Enterprise Data Analysis and Visualization: An Interview Study," October 2012.

[4] Gartner, "Getting Value from Big Data," webinar, May 2012.

US XPRESS SUCCESS SNAPSHOT

With a far-reaching program called "No Data Left Behind," trucking company US Xpress collects 900 big data elements from tens of thousands of trucking systems—sensor data for tire and gas usage, engine operation, geospatial data for fleet tracking, and complaints posted on trucker blogs. Using Hadoop and Informatica, US Xpress processes and analyzes this Big Data to optimize fleet usage, reducing idle time and fuel consumption and saving millions of dollars a year.

Generate insights faster with collaboration, prototyping, and rapid development. Role-based data profiling, quality, and integration tools with business user self-service provide a collaborative environment that narrows the gap between business and IT and accelerates insights. Project delivery times can be shortened with Informatica's no-code development environment and rapid prototyping of data pipelines that can be automatically deployed through business-friendly views for end-user validation.

Informatica's metadata business glossary renders technical metadata in semantic terms to enable business analysts to catalog, govern, and utilize big data consistently and efficiently and adapt rapidly to changes in the data and business environment.

Operationalize big data insights to drive new revenue and cut costs. As a cross-enterprise platform suited for operational data integration, Informatica allows organizations to deliver any data anywhere, anytime, from any source, to support embedded analytics across diverse business areas driving new revenue and reducing operational costs. High availability and unlimited scalability helps ensure timely delivery of hundreds of terabytes of information.

Conclusion

Big data will only get bigger, and so will the stakes for your organization to cost-effectively manage its growth and derive business value from a growing array of sources. Instead of a wait-and-see approach, leading Informatica customers are innovating today with low-risk, high-reward big data projects that target explicit business pains and opportunities and establish repeatable processes and best practices for future deployments. The Informatica Platform supplies proven, industry-leading technology to support a start small, think big approach that minimizes risk and cost while positioning your organization to reap rewards from big data for years to come.

ABOUT INFORMATICA

Informatica Corporation (NASDAQ: INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica for maximizing return on data to drive their top business imperatives. Worldwide, over 4,630 enterprises depend on Informatica to fully leverage their information assets residing on-premise, in the Cloud and across social networks.

INFORMATICA®

Worldwide Headquarters, 100 Cardinal Way, 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/InformaticaCorp