INFORMATICA

WHITE PAPER

Preparing for the Big Data Journey

A Strategic Roadmap to Maximizing Your Return from Big Data

This edition published March 2013

Table of Contents

Introduction

The analyst firm Gartner offers a savvy observation on the future of big data: In less than a decade, big data won't be "big" at all—it will be "just data," an informational commodity upon which an organization's fortunes rise or fall. As Gartner put it:

"As big data's effects are pervasive, it will evolve to become a standardized requirement in leading information architectural practices and will force older practices and technologies into early obsolescence. Big data will once again become 'just data' by 2020 and architectural approaches, infrastructure and hardware/software that do not adapt to this 'new normal' will be retired or organizations resisting this change will suffer the severe economic impact."[1]

In other words, the big data challenge now unfolding will eventually give way to clear skies (or the next big technological disruption). How well your organization prepares for its journey through big data challenges will influence whether it's standing strong by the turn of the decade, leveraging big data for competitive advantage, or whether it's left battered and weakened.

With inexorable increases in big data volume, variety, and velocity, the time is now to build and refine a strategic roadmap to generate business value from big data. This white paper outlines key focus areas for big data preparation covering technology, process, and people. It draws on the deep expertise of Cognizant's Enterprise Information Management (EIM) practice in completing more than 4,900 projects for over 700 enterprises. It also leverages Informatica's insight from working with organizations worldwide to deliver the comprehensive, scalable data infrastructure needed to harness big data.

[1] Gartner, "Big Data Drives Rapid Changes in Infrastructure and $232 Billion in IT Spending Through 2016," October 12, 2012.

Turning Petabytes into Profits

We have entered the Petabyte Age. Only several years ago, a 300-TB data warehouse was considered huge. Today, leading organizations run petabyte-scale data warehouses, but traditional analytic platforms are not suited for the scale and complexity of such big interaction data as social media content, sensor and machine information, call detail records (CDRs), and Web logs. Organizations need a new approach to harness unstructured and multistructured information and to combine it with ever-growing volumes of big transaction data (from data warehouses, ERP applications, and OLTP systems).

With the right approach, big data promises to help organizations improve operational efficiency, combat fraud and customer churn, enhance products and services, and increase insights and profitability. Capitalizing on the business potential of big data is of keen interest in most industries:

- **Financial services:** Detect fraud; enable more immediate and precise marketing
- **Telecommunications:** Leverage CDRs to optimize networks; reduce customer churn
- **Manufacturing:** Streamline production; strengthen quality control
- **Healthcare:** Generate insights from clinical data; improve patient outcomes
- **Public sector:** Enhance public safety, disaster preparedness, and research
- **Consumer industries:** Deepen customer engagement with social media analytics

Size Up Your Analytics Maturity

The first step in a big data journey is to understand the maturity of your organization from an analytics perspective. The business intelligence (BI) infrastructure and processes you have in place will influence how rapidly your business and IT teams can marshal the resources needed to capitalize on the big data opportunity. The research and advisory firm Bersin & Associates offers a good analytics maturity model to benchmark big data readiness (see Figure 1).

Source: Bersin & Associates, 2012

Figure 1. Analytics maturity model for assessing readiness to handle big data

- **Level 1:** Reactive operational reporting with classic dashboards is geared to measure how an organization has performed; these lowest-level analytic processes can consume weeks and lack complete, detailed, timely data.
- **Level 2:** Proactive advanced reporting is a step up, supporting improved decision making with drill-down and segmentation of more statistically robust information, yet is still a rear-view-mirror look at performance.
- **Level 3:** Strategic analytics give organizations a forward-looking view based on advanced modeling and deeper segmentation to produce actionable insights and identify opportunities and causes of issues.
- **Level 4:** The most advanced level of predictive analytics uses predictive modeling and pattern discovery to forecast future conditions and mitigate risk, based on complete structured and unstructured data.

In consultation with many hundreds of enterprises across industries, Cognizant believes that most organizations are at Level 2 of analytics maturity. At the same time, the majority considers advancing its BI practice and data integration capabilities a high priority as big data continues to emerge.

Many enterprises are in early phases of testing and production with such big data technologies as the open-source Hadoop/MapReduce data processing framework. An Informatica® survey of 589 IT and business professionals in North America, Europe, and Asia Pacific in spring 2012 found that 70 percent of respondents had big data projects in planning or production.[2]

By sizing up your big data readiness, you position yourself to start building a foundation for leveraging new and larger data sets. It can be risky, however, to assess maturity or build a foundation in a vacuum. With its worldwide EIM practice, Cognizant has extensive experience in helping organizations benchmark their readiness and navigate what can be a confusing maze of technologies and vendor claims about big data objectives.

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

Build a Solid Foundation on a Data Integration Platform

The tip of the iceberg of most big data projects is the business insights made possible through sophisticated analytics. Below the ocean's surface is roughly 80 percent of the work in most big data projects which involves accessing, parsing, cleansing, profiling, transformation, and loading of data from disparate sources. Building a solid big data foundation requires a strategic and disciplined approach to data integration and a data integration platform engineered for big data's challenges.

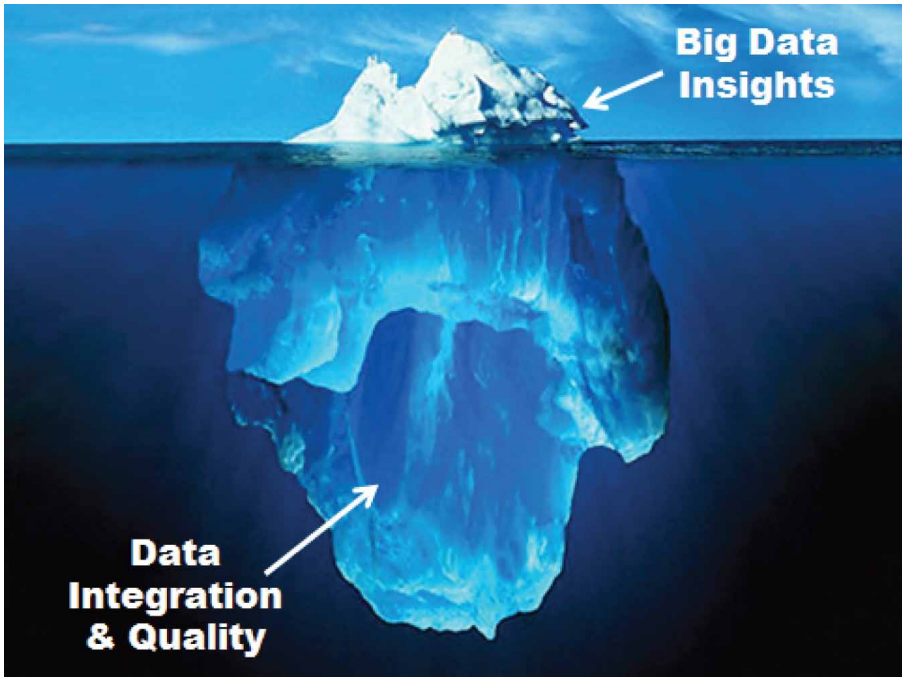


Figure 2. 80% of the work in big data projects is data integration and data quality

Some organizations have resorted to hand coding for big data extraction, transformation and loading (ETL) which is expensive and difficult to maintain. Others are increasing their productivity by up to 5x by leveraging the Informatica Platform's capabilities for data integration on Hadoop. Trained ETL developers instead of expensive MapReduce developers can build and execute on Hadoop ETL transformations, complex file parsing (e.g. web logs, JSON, XML, etc.), data quality, data profiling, and entity extraction for unstructured text from social media and documents. The Informatica Platform makes it easy to get data into and out of Hadoop with near universal data access, high-speed data replication, data streaming, and data archiving to Hadoop. Informatica brings enterprise scalability, security, and flexibility to Hadoop so you can operationalize big data insights.

Many enterprises are deploying Hadoop for economical data storage and processing, often at a cost 10 times less than traditional systems. Some are focused on experimental projects in a Hadoop sandbox. From a data integration perspective, a solid big data foundation can be built in stages with a philosophy of start small, think big, and focus on value.

- **Hadoop to augment a data warehouse.** Hadoop doesn't need to replace a data warehouse, but rather augment a data warehouse by enabling processing of unstructured or multi-structured data and pre-processing of raw data, then loading the results into the data warehouse for downstream analytics.
- **Hadoop as a distributed framework for data integration.** Hadoop's scalability and ability to handle raw, schemaless data makes it attractive for ETL preprocessing and data quality. Informatica provides ETL, complex file parsing, data quality, data profiling, and data discovery on Hadoop to cost-effectively integrate and cleanse data.

Informatica is the safe on-ramp for big data projects as new technologies emerge, such as Hadoop. The Informatica Platform maximizes your return on big data with the following benefits:

Reduces big data costs up to 2x or more by:

- Reducing infrastructure costs up to 2x with your existing analytics environment
- Increasing productivity up to 5x at half the labor cost by moving to a no-code development environment

Minimizes risk with a safe on-ramp to big data by:

- Leveraging a single platform across your existing infrastructure and new technologies
- Quickly staffing big data projects with more productive and trained data integration experts
- Providing enterprise security and protects sensitive data

Innovate faster with big data by:

- Onboarding and analyzing any type of data to gain big data insights
- Discovering insights faster through rapid development, prototyping and collaboration
- Operationalizing big data insights to generate new revenue streams

As the foundation grows, your organization is ready to align its big data capabilities to business needs, progressing from the baseline architecture to the ideal of real-time analytics on complex, fast-moving data sets (see Figure 3). Clearly defined metrics and objectives are important to success in building your big data foundation.

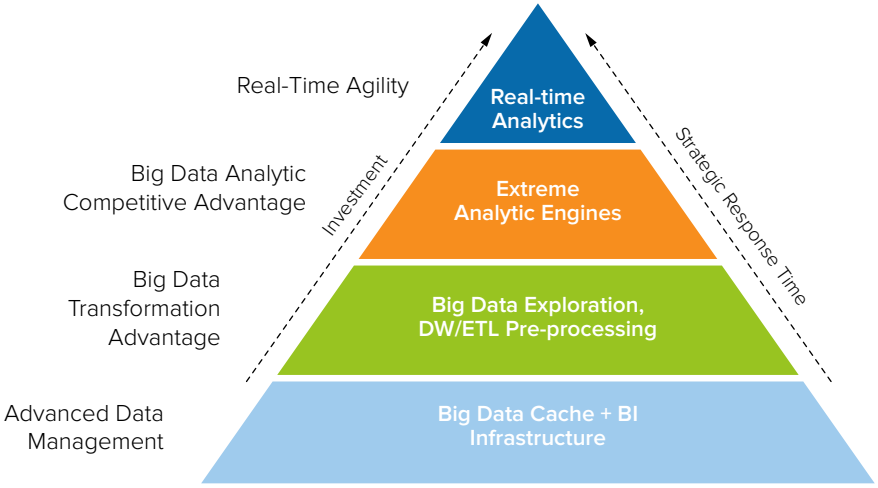


Figure 3. A big data foundation evolves toward the ideal of real-time analytics.

Cultivate a Data Science Team

Leveraging big data requires a different technological framework than traditional BI—and a different breed of practitioner. The role of data scientist is emerging as the human catalyst for business value from big data. What is a data scientist, other than being named the "Sexiest Job of the 21st Century" by the *Harvard Business Review*?[3]

A data scientist is a step above business analyst, combining technical skills in such disciplines as statistical modeling and data discovery with strong business experience and acumen to interrogate data, generate insights, and make recommendations to improve business performance. Data scientists will often leverage data visualization tools to aid their own research and promote clear visual understanding of issues to executives and other stakeholders.

[3] Harvard Business Review, "Data Scientist: The Sexiest Job of the 21st Century," October 2012.

Cultivating a data science culture is among the top initiatives your organization can take to maximize its returns from big data initiatives, but it's trickier that simply hiring a programmer. Skilled data scientists are increasingly coveted by large organizations and can command generous compensation. To create a data-driven environment attractive to top talent and to focus your resources for rapid value, consider the following:

- **Chief data scientist:** The right combination of leadership skills, technological mastery, and business focus in a chief data scientist can profoundly influence your organization's big data directions and payback.
- **Data science as a team sport:** The most effective data science programs will successfully align mathematical and statistical wizards with business managers, IT architects, developers, and others with a stake in the outcome of a big data project.
- **Independent expert perspective:** Cognizant's highly skilled analytics advisors can help jumpstart your data science program through strategic guidance, industry best practices, hands-on training, mentoring, and knowledge transfer.

Once you cultivate a data science team, you need to make sure its members are most productive and can collaborate efficiently. For example, why have your data scientists spend 80 percent of their valuable time hand-coding data integration and data quality before they can even do the analysis? Instead, use Informatica trained ETL developers to prepare the data for analysis up to 5x faster than hand-coding using a no-code visual development environment. Data integration and quality will build the foundation for successful data science. D.J. Patil, chief data scientist at Greylock Partners and former data scientist at LinkedIn, claims in his book *Data Jujitsu* that "80 percent of the work in any data project is in cleaning the data."[4]

Similarly, a recent study of 35 data scientists from 25 companies found that most of the work involved with big data analytics involves data integration. One of the data scientists in the study said, "I spend more than half 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' at all."[5]

[4] D.J. Patil, *Data Jujitsu*: The Art of Turning Data into Product (O'Reilly, 2012), 16

[5] Kandel, et al. "Enterprise Data Analysis and Visualization: An Interview Study." IEEE Visual Analytics Science and Technology (VAST), 2012.

Summary

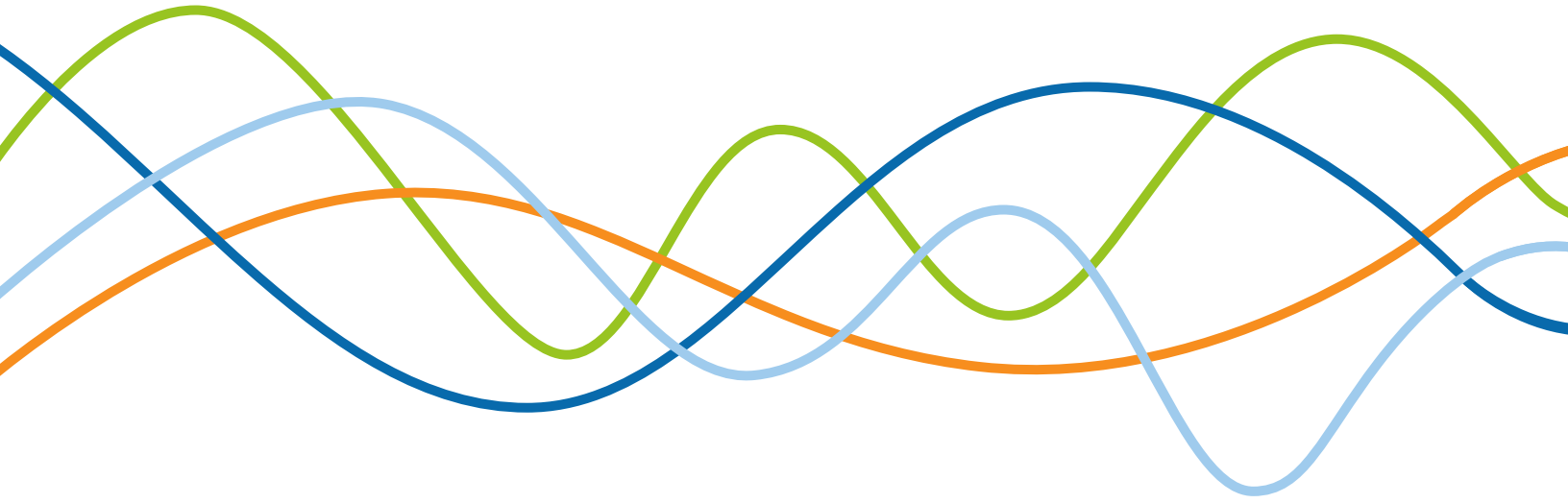
The big data storm is gathering strength, and smart organizations have it high on their radar screens. They recognize that big data represents not just an opportunity to improve the business—it also introduces substantial risk. If unaddressed, a rising tide of big, complex data can obscure visibility and degrade business and IT execution, while your rivals turn it into competitive advantage.

Cognizant and Informatica have partnered at a technological and strategic level to help enterprises prepare for their big data journey and excel at early implementations. With certified consultants around the world, Cognizant's EIM practice offers industry-leading consulting and services to guide organizations toward such big data goals as boosting sales, improving customer service, achieving new efficiencies, and increasing market share.

Cognizant's deep industry expertise can help your organization assess its analytics maturity, jumpstart a data science team, and build a solid, flexible foundation that takes advantage of Informatica's comprehensive and proven platform for big data integration. Learn more at www.cognizant.com and www.informatica.com.

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

IN09_0213_02342