Company Brief

Informatica 9.1 and Integrating Big Data

Date: June 2011 **Author:** Julie Lockner, Vice President, Data Management

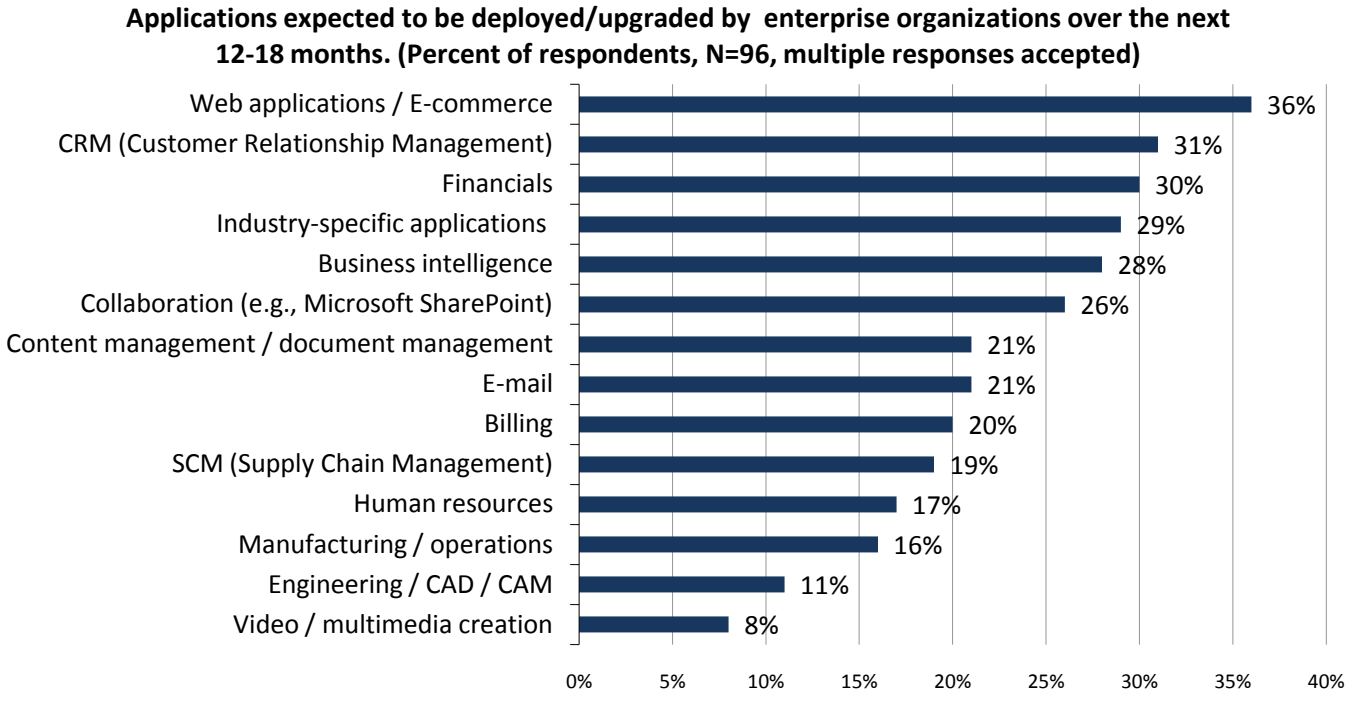
Abstract: *Organizations that have invested in enterprise data warehouses have a long history of dealing with "big data" volumes that push current infrastructure to its limits. As new data sources come online and existing data sets grow, the need for a data integration platform that is big data aware can help organizations cross over barriers to entry. Informatica customers upgrading to the latest 9.1 version have an opportunity to incorporate and streamline their data integration efforts for big data and do so while controlling data growth.*

Overview

As corporations begin to classify their large data sets as "big data," the popular label brings a big price tag if these sets are managed in silos and left uncontrolled. Organizations that have already invested in an enterprise data warehouse (EDW) with a data integration solution from Informatica are all too familiar with the challenges associated with large enterprise databases, constantly changing data integration requirements, and the need to provide integrated information to business users in real time. This brief highlights the benefits these organizations may experience if they take advantage of recent enhancements to the Informatica 9.1 release while considering an investment in an ILM solution to control big data sprawl.

Respondents to a recent ESG survey indicated that managing data growth is ranked number two in IT priorities for 2011.[1] Uncontrolled growth within database applications can negatively impact a business' ability to simultaneously control cost and maintain application performance levels. In the same survey, 28% of enterprise organizations (1,000 employees or more) planned to deploy business intelligence applications in the next 12-18 months (see Figure 1).

Figure 1. Applications Organizations Plan to Deploy or Upgrade in the Next 12-18 Months



Source: Enterprise Strategy Group, 2011.

[1] See: ESG Research Report, *2011 IT Spending Intentions Survey*, January 2011.

E-commerce and CRM applications took the lead for new deployments. With these new applications come new data sources that will most likely feed business intelligence and data warehouse solutions. The fast-paced adoption of new data management and processing frameworks, such as Hadoop, and the addition of new data sources from social networking sites, such as LinkedIn and Facebook, increase the risk of introducing silos of information. This is driving new business requirements for data integration solutions that can seamlessly operate in new analytics platforms and integrate new data types while keeping up with data growth. As these new big data sources accumulate data en masse, management challenges are compounded.

As data volumes grow organically in existing applications and new applications are added, the total amount of data replicated into an enterprise data warehouse becomes so large that it requires organizations to rethink their data economics altogether: data integration tasks and reports may take longer to complete, running complex analytics may become impractical, and data quality control may need to be extended effectively and efficiently to the new data sources.

The impact big data has on data economics is multiplicative: for every production database, IT departments maintain multiple additional copies for purposes including reporting, backup, disaster recovery, test/dev, and others. If the business needs those data copies quickly, even a slight delay could represent a major disruption to key processes. Without a sound data management strategy and an advanced data integration platform that manages growing volumes in their source applications, in the data warehouse, or during the data integration process, performance will deteriorate and businesses will soon find themselves making a significant investment in storage and compute resources.

Informatica 9.1

Advancements in Data Integration to Accommodate Big Data Sources

In its most recent release, Informatica version 9.1 can help customers leverage and find value in new data sets. Key enhancements will help integrate and process this data, transitioning it from a business disabler to a business enabler and overcoming key big data hurdles.

Support for Integrating New Big Data Sources

New data connectors for Twitter, Facebook, and LinkedIn offer the ability to integrate customer sentiment into targeted marketing campaigns. For organizations that need to incorporate smart device or sensor input as part of a comprehensive analytical platform, Informatica 9.1 announced extended support for industry-specific device output integration. New enhancements in native connectivity to OLTP and online analytical processing (OLAP) data stores will open up the ability to process large volumes (extending into petabyte range) of transactional data. In some cases, these new analytical platforms are deployed as an appliance; version 9.1 ensures integration support for appliance-based analytical databases as well.

Support for Integrating Hadoop

The latest big data trend in next generation analytics and data processing is based on the open-source project Hadoop and the MapReduce processing framework. Organizations looking to leverage a Hadoop data store for MapReduce algorithms can take advantage of native Informatica connectivity to coordinate data loads into the Hadoop Distributed File System (HDFS), available in 9.1, or Hadoop database (HBase), available in the release following 9.1. Once the data has been processed, the results can be extracted from the Hadoop platform into an existing enterprise data warehouse or operational data store (ODS). This new integration support eases some of the concerns organizations may have when deploying emerging technology.

Universal MDM

Whether new big data sources are incorporated due to new business initiatives and requirements or because of a merger or acquisition, organizations need to maintain a single set of master data to ensure business process efficiencies. Master data management (MDM) solutions need to be able to adapt and integrate new sets of customer, supplier, or product data even when those sources reside on completely different domains or architectural styles. Informatica 9.1's

enhancements promote "Universal MDM" which highlights a translation layer that allows organizations to leverage existing investments while meeting common MDM goals. One key area of promise is the ability to deploy proactive data quality assurance as new big data sources become available. Addressing data quality issues before they can hamper testing cycles and production integration plans can prevent one of the major causes for data integration project delays.

Self Service Deployment

A key concern for IT organizations is a lack of skilled application developers and database administrators. Facing a growing onslaught of data and data sources, organizations need to be able to relocate data integration and data quality processes closer to the owners of the data: the business users. The Informatica 9.1 release gives business users the ability to deploy data integration and data quality rules and processes without programming or writing database query scripts. Using wizard-driven, highly graphical, business-context aware, non-technical user interfaces, enhancements for self service data integration empower the users to quickly and efficiently integrate new data sources into existing business processes.

Self service interfaces also reduce dependence on IT, which means that organizations can get the most out existing resources using a tool such as Informatica to translate business vernacular into business process programming without code or database queries. And combining the self-service capabilities of Informatica 9.1 with application-specific accelerators for business applications (such as Oracle E-Business Suite and PeopleSoft) enables organizations to jumpstart projects with prepackaged metadata that elevates the integration conversation from a technical level to a focus on business process.

Extended Adaptive Data Service Support

While big data is one dimension of the data integration challenge, another is the need for real-time data integration. Informatica 9.1 extends support of data services for both transformational batch processes and real-time data integration requirements. The ability to reuse transformation logic independent of the application or protocol requesting the service, without the need to reprogram or redevelop code, means that organizations are not limited by existing legacy platforms as they embrace new, real-time integrated applications. When application developers and DBAs have limited bandwidth to retrofit old application integration techniques with next generation platforms, the ability to leverage existing investments without draining already taxed resources allows projects to be completed faster and within budget. Informatica 9.1's features promise both.

Managing Big Data Lifecycles

One key component that complements the Informatica 9.1 announcement was the benefit of implementing an ILM strategy with the addition of big data sources. Businesses that have invested in or plan to invest in a data integration solution should consider simultaneously deploying an ILM solution with each new big data source.

In many cases, most batches of big data used in analytical platforms are only active for a short period of time. Once analysis is complete, the raw data may be transformed and stored in a data warehouse for longer retention periods with minimal access requirements. This is a perfect opportunity to leverage Informatica Data Archive to control the cost of managing big data while maintaining access and performance service level agreements. Removing aged big data from production data stores once it is no longer needed controls storage and compute power resource requirements associated with managing larger databases. As data is removed from production databases, copies used for DR and test and development, once refreshed, correspondingly shrink in size.[2]

[2] See: ESG Market Landscape Report, *Managing Database Growth – Optimizing Database Application Lifecycles*, April 2011.

The Bigger Truth

The concept of big data is not new to those supporting an EDW deployment—a data warehouse, by its very nature, is big data. New applications supporting new business processes bring new streams of data copied into already-overburdened data warehouse and analytics platforms. Data integration platforms that process information from one source, transforming it into an acceptable format before loading into a target, need to be able to keep up with next generation analytics focused on processing massive volumes of data quickly and affordably.

As new big data sets proliferate, the notion of "herding cats" becomes less comical when viewed alongside the price tag accompanying the search for value in a sea of data assets. At some point, finance will take notice and question the ROI of investments in big data, especially as businesses request more budget to implement analytical platforms that promise answers from big data.

Business units and IT need to collaborate as new big data sources stress existing data integration processes. This includes integrating new data sources, ensuring data quality and data lineage can be tracked and maintained, preserving master data models, and ensuring real-time data integration is not impacted as data volumes explode.

As the data warehouse becomes a dumping ground, tools that are used to copy source system data into the warehouse become the proverbial plumbing. The data integration architect has a unique perspective that should be exploited once that dumping ground starts to seep to the surface. To prevent what could be considered catastrophic systemic failure as a result of too much data, data growth management, vis-à-vis ILM solutions, should be architected into the design of every application and data integration project. Vendors such as Informatica that offer both the plumbing for getting data into databases (ETL) and broader data integration as well as taking data out (ILM) offer their customers a strategic advantage that too often gets overlooked.

With the new release of Informatica 9.1 and its focus on addressing the challenges big data brings to data integration processes, customers now have an opportunity to incorporate big data more efficiently and with less risk by eliminating some of the potential barriers to entry unlocking the potential value big data promises.