FORRESTER®

For: Enterprise Architecture Professionals

Market Overview: Big Data Integration

by Noel Yuhanna, December 5, 2014

KEY TAKEAWAYS

Big Data Creates New Data Challenges

Today, most big data deployments are built in silos, largely to address specific business needs: collecting sensor data to support smart metering, web clickstream data to support customer analytics, and geolocation data to support customer personalization. These silos create a major challenge -- especially when it's time to integrate them.

Big Data Integration Delivers A Unified View Of The Business And Its Customers

Big data integration supports integration with Hadoop and NoSQL; data warehouses and data marts; packaged and custom apps; sensors and devices in the Internet of Things; web logs and clickstream data; social platforms like Facebook, LinkedIn, and Twitter; and software-as-a-service applications like salesforce.com.

Big Data Integration Should Be Part Of Your Big Data Strategy Going Forward

Although just about every enterprise can benefit from big data integration, it's not a silver bullet that will fix all of your big data issues. Leverage big data integration for small to medium-size projects first to establish the process, approach, and architecture for a strong foundation before tackling larger ones.

Market Overview: Big Data Integration

Your Big Data Strategy Is Not Complete Without Big Data Integration

by Noel Yuhanna
with Gene Leganza, Leslie Owens, and Elizabeth Cullen

WHY READ THIS REPORT

Enterprise support of big data strategies has increased data management challenges, especially around integration, security, quality, and governance. Business users want to integrate real-time data from multiple sources, including big data, to enable accurate business decisions; developers want to integrate all kinds of data to support next-generation mobile, web, and interactive applications. But organizations are realizing that simply putting lots of diverse data into Hadoop doesn't always magically result in the ability to gain insights from that data without further integration, transformation, and enrichment. Big data integration brings together disparate big data sources to deliver a unified, comprehensive view of the business and its customers, employees, and products. It supports integration with Hadoop and NoSQL; data warehouses and data marts; packaged and custom apps; Internet of Things (IoT) sensors and devices; web logs and clickstream data; social platforms like Facebook, LinkedIn, and Twitter; and software-as-a-service (SaaS) applications like salesforce.com. Enterprise architects should consider making big data integration part of their big data strategy and focus on integrating diverse data sources to support an agile big data platform.

Table Of Contents

Notes & Resources

Forrester interviewed 12 vendor companies, including Cisco Systems (Composite Software), Denodo Technologies, IBM, Informatica, Microsoft, Oracle, Pentaho, SAP, SAS Institute, SnapLogic, Syncsort, and Talend.

Related Research Documents

TechRadar™: Big Data, Q3 2014
September 10, 2014

The Forrester Wave™: Big Data Hadoop Solutions, Q1 2014
February 27, 2014

The Forrester Wave™: Enterprise ETL, Q1 2012
February 27, 2012

BIG DATA CREATES NEW DATA CHALLENGES

Big data is about gaining business insights from very large data sets — insights that were previously impossible due to prohibitive infrastructure costs and a lack of scalable solutions. However, the evolution of massively parallel processing software solutions such as Hadoop, flexible on-demand infrastructure platforms such as cloud, and the lower cost of computing and storage have made big data architectures affordable enough to support pursuing these new insights. Enterprise architects are leveraging Hadoop for customer intelligence, fraud detection, asset and inventory management, advanced analytics, and predictive analytics. Although Hadoop is one of the main technologies helping organizations bridge the gap between the available data and the ability to turn into that data into business insights, it's not the only one. Technologies like NoSQL, distributed in-memory, integrated data management appliances, and elastic cloud are also helping to store, process, and manage big data. Forrester defines big data as:

> *The practices, processes, and technologies that close the gap between the data available and the ability to turn it into business insights.*

Today, most big data deployments are built in silos, largely to address specific business needs: collecting sensor data for smart metering, web clickstream data for customer analytics, or geolocation data for customer personalization. As a result, most enterprises struggle to integrate their data (see Figure 1). But big data is often stored in a Hadoop cluster or a NoSQL scale-out platform with its own metadata structure and can't support broader business needs, as it lacks the integration with master data, reference data, or customer data necessary to provide a complete view of the business or customer. For example, a retailer that uses customer intelligence to create a multidimensional view of the customer requires data from all kinds of sources, including Hadoop, salesforce.com, web logs, data warehouses, and master data management (MDM) systems. Delivering such connected data across on-premises and cloud sources is not trivial, especially when the task involves large data volumes and complex data models.

Figure 1 Big Data Means Dealing With A Variety Of Data Types

Structured data	• Data described by a schema • Relational databases, XML, delimited flat files, system events, IoT device data
Semistructured data	• Has some structure • Emails, documents, tweets, blog comments, Facebook statuses, genomes
Unstructured data	• Audio, images, and video • Surveillance footage, geological survey maps, Siri

117834 Source: Forrester Research, Inc. Unauthorized reproduction or distribution prohibited.

New Big Data Challenges Emerge Around Integration, Security, And Real-Time Support

Big data marketing hype is starting to fade away, and the reality is kicking in. Organizations are realizing that simply putting lots of diverse data into Hadoop doesn't always magically result in the ability to gain insights from that data without further integration, transformation, and enrichment. Traditional data integration solutions — extract, transform, and load (ETL), data integration, replication, enterprise information integration, and enterprise application integration — are failing to meet big data project expectations. They can't keep up with petabyte-scale data volumes; they need tighter integration with semistructured and unstructured data; they deal with data models that might not be readily available, like Hadoop's schema on read; they have unpredictable data frequencies; they include data fields with unpredictable sizes; they need to support missing data without problems; and they deal with real-time, near real-time, or batch data.[1] Top challenges include:

- **Too many silos of Hadoop clusters.** Today, many enterprises are building silos of Hadoop clusters to support specific business needs. Some store clickstream data; others store data from web logs, social networks, the Internet of Things (IoT), or geolocation apps. Each cluster is optimized to support a specific domain, creating further silos that complicate the support of broader business initiatives. While organizations can use traditional ETL or replication solutions to move big data, these technologies do not scale to the necessary degree and don't support optimized processing for unstructured and semistructured data sets.

- **Fragmented data.** Forrester estimates that the average Hadoop repository doubles in size every year; some implementations double in volume every month. Many large enterprises already have petabytes of data stored in various big data repositories, and this is likely to grow to exabytes in coming years. However, not all of this data is in Hadoop; enterprises still store high-value data in databases, data warehouses, and legacy systems to support their low-latency data platform for MDM, transactional applications, and other critical real-time applications. This hybrid data management architecture leaves data spread across many repositories, creating silos that become difficult to integrate.

- **Greater compliance and security risks.** Tougher global compliance regulations and increased internal and external data theft are driving the need for enhanced big data security measures. Today, most big data implementations are vulnerable due to poor authentication, weak access controls, and highly visible sensitive data in Hadoop and NoSQL platforms. Enterprises also face pressure to deal with data privacy laws that require stronger big data security measures.

- **Real-time data support.** As more big data projects are deployed, it becomes increasingly complex to integrate them in real time, especially as each Hadoop and NoSQL repository often requires data from other data sources to make it complete. With increased big data volumes comes a bigger challenge: knowing what the necessary data is and where to look for it. Hadoop was originally designed for processing large amounts of data in batch using MapReduce.

Organizations are realizing that, while MapReduce helps with aggregation and transformation, it falls short when it comes to real-time analytics. However, recent evolution of in-memory technologies such as Apache Spark promises faster access to big data platforms.

- **The complexities of providing self-service big data platforms.** More enterprises are starting to implement a self-service big data platform to empower business users to make faster, more accurate decisions. Self-service allows business users and consumers to directly access the big data they need when they need it, reducing the involvement of technology management in supporting them. However, delivering a self-service data platform is not trivial; it requires more sophisticated solutions to integrate all big data elements and create a business view of data.

- **The need for new skills and more resources.** Enterprises want simplified big data management solutions that enable them to focus on business issues rather than dealing with technology challenges. Although SQL integration with Hadoop has simplified access to Hadoop distributions, organizations still require new skills on Hadoop, HBase, Cassandra, Hive, Pig, Yarn, and HSQLDB to support a comprehensive, enterprisewide big data strategy. Big data complexity is delaying and deferring the expansion of big data initiatives in organizations.

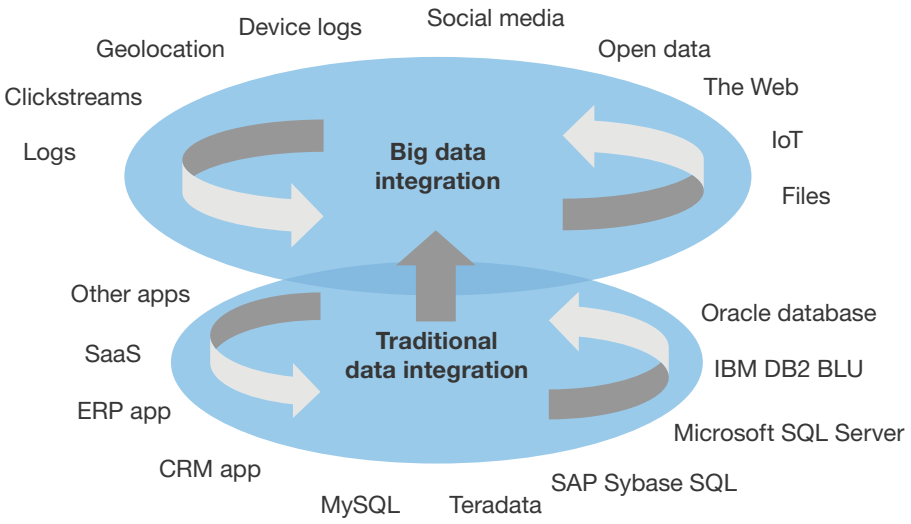
BIG DATA INTEGRATION DELIVERS A UNIFIED VIEW OF BUSINESS AND CUSTOMERS

Big data integration is much broader than traditional data integration; it integrates Hadoop, NoSQL, files, and traditional database sources such as databases, data warehouses, and applications (see Figure 2). Solutions have evolved; traditional integration vendors have extended their coverage to support big data sources such as Hadoop, NoSQL, and IoT, while new niche vendors leverage Hadoop as the foundation for integration with other sources. In addition, machine learning is helping to parse large volumes of data through automation and data intelligence. Forrester defines big data integration as:

> *The integration of data from disparate big data sources including Hadoop, NoSQL, IoT, cloud, data warehouses, files, and databases, whether on-premises or cloud, structured or unstructured, to support big data analytics, predictive analytics, and other workload patterns.*

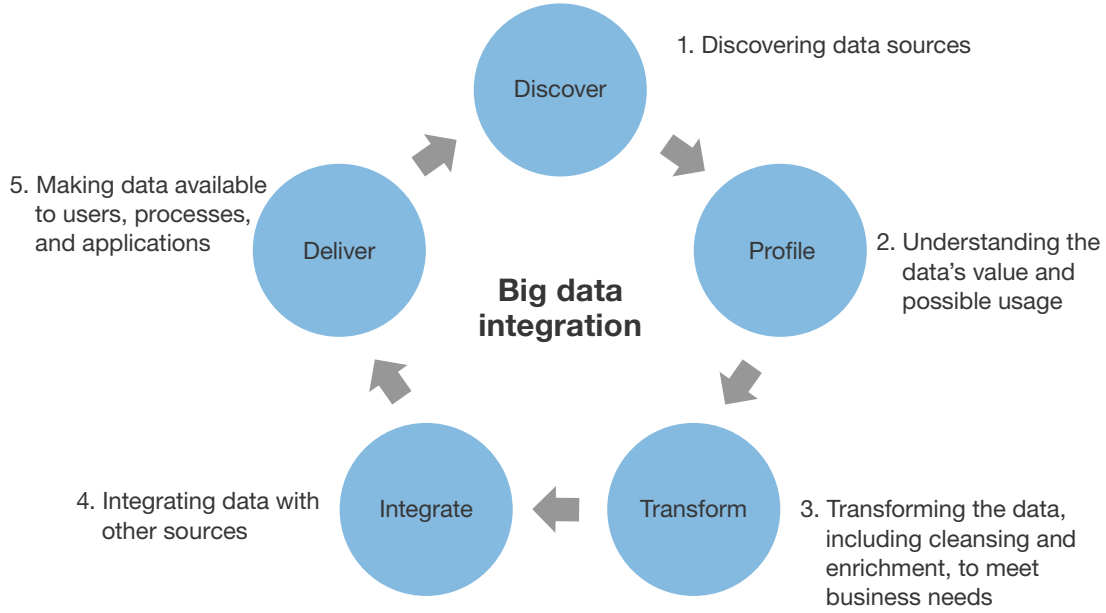
Big data integration occurs in five phases: 1) Discovering data sources; 2) understanding the data's value and possible usage through profiling; 3) transforming the data, including cleansing and enrichment, to meet business needs; 4) integrating data with other sources in Hadoop, NoSQL, data warehouses, and databases; and 5) making the data available to users, processes, and applications to support business intelligence (BI), analytics, and predictive analytics (see Figure 3). Big data integration architecture is flexible; it supports some level of transformation and integration within the Hadoop platform, as well as with non-Hadoop sources (see Figure 4).

Figure 2 Big Data Integration Is Much Broader Than Traditional Data Integration



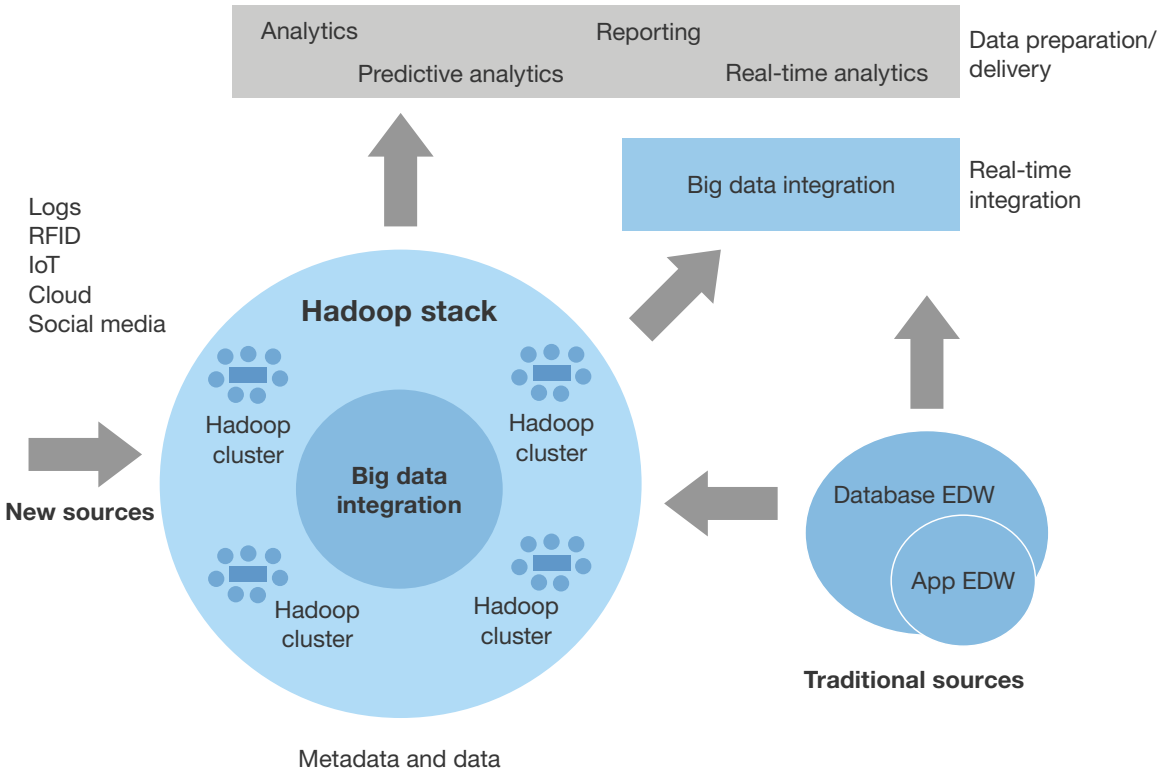
117834 Source: Forrester Research, Inc. Unauthorized reproduction or distribution prohibited.

Figure 3 The Big Data Integration Framework



117834 Source: Forrester Research, Inc. Unauthorized reproduction or distribution prohibited.

Figure 4 Big Data Integration Architecture



117834 Source: Forrester Research, Inc. Unauthorized reproduction or distribution prohibited.

Big Data Integration Saves Money And Minimizes Complexity

Why implement big data integration solutions? Because they automate and simplify the integration of big data, eliminating complexity from the process. These solutions also deliver faster time-to-value — a critical factor in today's competitive world. Big data integration offers several key benefits, including:

- **Reducing technology management costs.** Briefings and inquiries with Forrester clients indicate that the average Hadoop cluster is only 50% utilized; most are dedicated to a single use case, which is often a waste of resources. Big data integration focuses on reducing costs by leveraging underutilized Hadoop clusters to support more use cases. It also automates data integration, security access, and management, requiring less developer and administrator time.

- **Minimizing complexity.** All enterprises deal with heterogeneous platforms and software. Some have thousands of servers and applications, as well as various big data repositories containing petabytes of data. This diverse infrastructure and application landscape increases complexity, making management a major challenge. Big data integration minimizes complexity by automating processes, generating MapReduce code, and integrating workflows to simplify them.

- **Providing a complete view of business information.** Big data integration offers the ability to aggregate data from multiple sources, which can then be presented in dashboards, reporting tools, and web applications. Examples of aggregated data use include hourly sales data across regions, mobile phone activation across cities, and network usage across states.

BIG DATA INTEGRATION USE CASES GO BEYOND OFFLOADING DATA TO HADOOP

Big data integration helps enterprises process large amounts of data on big data platforms quickly to support numerous use cases. It offers the ability to orchestrate data flow and processing across various big data platforms to support a single version of the truth, consumer personalization, and data aggregation. The top big data integration use cases include:

- **A 360-degree view of the business.** Big data integration processes large amounts of data quickly from internal and external big data sources to deliver a 360-degree view of products, employees, and the business. For example, a company may want to learn about customer reaction to a product launch and how it compares with previous launches. Big data integration can extract and transform data from social networks such as Facebook and Twitter and process them in Hadoop; it can then further integrate this data with web logs and clickstreams from another Hadoop cluster that determines the interactions, and finally integrate it with data warehouse data containing historical information about previous products. Big data integration delivers all kinds of unified data views to deliver new insights and predictive analytics.

- **A multidimensional view of the customer.** Customer data stored in standalone Hadoop clusters delivers less business value than if it is integrated with other customer sources. Big data integration helps create opportunities to upsell and cross-sell new products to customers based on their likes, dislikes, circles of friends, buying patterns, and past orders based on data from Hadoop, NoSQL, in-memory, and traditional databases and data warehouses. When a customer speaks with a support specialist or salesperson on the phone, being able to zero in on a product or service that the customer would find attractive becomes critical.

- **Data warehouse optimization and offloading.** Big data integration helps augment existing data warehouses with Hadoop and NoSQL by offloading low-value and cold data. This helps save money by using lower-cost commodity servers with Hadoop and NoSQL in a scale-out architecture. Although organizations could write scripts to move data from data warehouses to Hadoop, big data integration speeds development and deployment by automating the process, eliminating the need for coding on both the source and destination targets.

- **Data staging on Hadoop to support data virtualization.** Data virtualization is the process of integrating disparate data sources in real time or near real time.[2] Although data virtualization can integrate with Hadoop to extract data, it does not integrate very large volumes of data residing in multiple Hadoop and NoSQL clusters. On the other hand, big data integration offers

the ability to store, transform, and process big data in Hadoop and NoSQL platforms, creating a staging area for data virtualization for use with real-time analytics or predictive analytics to produce business insights.

- **Big data analytics processing.** Hadoop handles complex data sets — such as time series data, semistructured and unstructured data, log files, and streaming data — with ease. It can also handle complex algorithms using R and further process data via complex transformations. This data can then be further integrated with structured data from data warehouses and operational stores to deliver big data analytics. For example, finding a cure for a new disease requires looking at an array of patient data across geographies and determining suitable medications and dosages based on age and sex. This complex scenario requires gathering millions of data points, storing them in Hadoop, performing complex algorithms with R, and finally combining this data with historical data for use with BI and analytical tools.

- **Deeper IoT analytics.** Hadoop offers the ability to store, process, and access large volumes of IoT data from sensors, devices, and switches more efficiently and at a lower cost. However, unless such data is integrated with a machine's description, location, history of issues, manufacturer, and acquisition and replacement costs, it provides little value. Big data integration helps deliver IoT analytics by integrating Hadoop data with master data and other transactional and operational data in real time or near real time.

THE BIG DATA INTEGRATION MARKET IS STARTING TO HEAT UP

Forrester expects that big data integration software will have significant success in the coming years, leveraging existing and new big data sources to deliver broader end-to-end business insights. Increased adoption of the Hadoop ecosystem, NoSQL, and streaming and in-memory technologies and the need to have a 360-degree view of the business, customers, employees, and partners will fuel enterprise demand for big data integration. Currently, a combination of traditional and specialized integration vendors serve this market — but we're likely to see startups delivering innovative solutions to the market in the coming years. Although the big data integration market is relatively small compared with the overall data management market and the traditional data integration market, it promises to be bigger than the latter within three years. Among the top big data integration vendors:

- **Acquisition by Cisco helped advance Composite Software's solution.** Composite has done extremely well since it was acquired by Cisco Systems; it has increased adoption of its software, added advanced big data integration capabilities, and delivered a more highly scalable product. Cisco has provided an SQL approach to augment specialized MapReduce coding of Hadoop queries and simplify integration with major Hadoop distributions. It supports access to Hadoop through Hive and Cloudera's Impala. Cisco can also integrate with NoSQL, in-memory, cloud, SaaS, and traditional big data sources. Customers use Cisco's solution to support a 360-degree view

of customer data and as a big data repository to ingest large volumes of data, archive big data, and replatform existing databases and data warehouses to Hadoop to save money. One of the largest deployments deals with more than a petabyte of data to support data warehouse augmentation.

- **Denodo's easy-to-use data virtualization platform extends into big data integration.** Denodo Technologies supports integration of unstructured and structured information to support real-time and near real-time big data requirements. In addition to enterprise apps, databases, and data warehouse products, Denodo Platform also integrates with multiple NoSQL databases — such as MongoDB, CouchDB, Cassandra, Neo Technology (Neo4J), and RDF repositories — and in-memory databases such as Oracle TimesTen and SAP Hana. It can natively integrate with a Hadoop distribution and supports multiple interfaces including Hive, Hadoop Distributed File System (HDFS), HBase, and HCatalog. Denodo supports vendor-specific interfaces including Cloudera's Impala, IBM's Big SQL and BigInsights, and Pivotal Software's Hawq. Denodo Platform integrates directly with cloud and SaaS platforms such as Salesforce, Google Apps, and NetSuite and can deal with semistructured data from websites, social media, and IoT sources. Customers use Denodo's data virtualization platform to create a logical data layer that supports virtual big data marts, big data analytics, and IoT data processing. Denodo's key strengths are delivering unified and centralized data services with security and real-time integration across multiple traditional and big data sources including Hadoop, NoSQL, cloud, and SaaS.

- **IBM leverages its mature and scalable integration solution.** IBM continues to be one of the most aggressive vendors in the data integration space, continually raising the bar on scale and performance while also supporting big data platforms. Common use cases seen with IBM's big data integration initiative include moving data into Hadoop and preparing it for big data analytics. IBM provides its own connectors to support various Hadoop distributions; it integrates with NoSQL databases including MongoDB, HBase, and Cassandra and in-memory databases such as IBM DB2 BLU, Oracle TimesTen, and SAP Hana. One of its largest installations — which has been in production for more than a decade — includes a grid deployment of more than 500 nodes that processes more than 3 petabytes of data per month. The client uses it for market analysis and integrates a variety of data including customer data, market demographics, and production information. To make client's projects simpler and more flexible, IBM offers Hadoop distribution entitlement with its data integration solution; a beta program is also underway to allow comprehensive data integration and quality capabilities to bypass pushdown and MapReduce layers to run natively under Yarn resource management. IBM's key strengths are around high-end scalability dealing with data integration at petabyte scale while allowing customers to process data inside or outside of Hadoop; end-to-end big data governance to provide a complete lineage; integrated metadata; and data quality that also supports privacy, security, MDM, and reference data management.

- **Informatica innovates and expands its solution to support big data.** More than 5,000 firms use Informatica for their information management initiatives; its integration technology is proven and mature. Informatica provides scalable integration with Hadoop, NoSQL, cloud, SaaS, in-memory,

and other big data sources. In 2011, it became one of the first vendors to deliver native integration with Hadoop distributions. Enterprises use Informatica's big data integration solutions for data warehouse optimization and offloading, managed data lakes, IoT initiatives, and real-time operational intelligence like fraud detection and proactive customer engagement. One of the vendor's largest deployments is an enterprise running a 300-node Hadoop cluster processing tens of billions of rows of data to support cross-company data analytics. Informatica's strength lies in increasing developer productivity via its intuitive visual and metadata-driven development environment that existing ETL developers can leverage for big data sources and prebuilt parsers, transforms, and connectors that help parse, integrate, cleanse, mask, and match data natively on Hadoop. It also supports the reuse of big data integration pipelines to support other infrastructures such as NoSQL, cloud, traditional grid, or symmetric multiprocessing platforms.

- **Microsoft delivers a viable Windows big data integration solution.** Microsoft's Azure HDInsight is based on the Hortonworks Data Platform (HDP) and is exclusively designed for and offered on Microsoft Azure. Microsoft has worked closely with the Hadoop community and Hortonworks to release Hadoop for Windows. Microsoft also offers PolyBase to allow SQL Server customers to execute queries that also include data stored in Hadoop. Microsoft's significant presence in the database, data warehouse, cloud, OLAP, BI, spreadsheet (PowerPivot), collaboration, and development tool markets offers an advantage when it comes to big data integration, especially on the Windows platform.

- **Oracle's big data integration solution leverages its own stack, partners, and appliances.** Oracle Big Data SQL provides data federation with Hadoop; Oracle Big Data Connectors delivers a high-performance Hadoop to Oracle Database loader and enables optimized analysis using Oracle's distribution of open source R directly on Hadoop data. Big Data SQL capability also allows the Oracle data security model to govern access to data residing in Hadoop. Oracle's big data integration software natively integrates with Oracle Big Data Appliance, Cloudera's CDH, HDP, and the Apache distribution. Oracle supports batch or real-time data movement from and to big data sources such as Hadoop, NoSQL databases, in-memory platforms, and cloud-based data repositories. Oracle's GoldenGate replication solution provides real-time capabilities, integrating with Oracle Data Integrator tools for a unified development experience; it also supports real-time big data integration to dynamically push data into the HDFS, HBase, Hive, Flume, Storm, and Kafka big data frameworks. The Oracle Enterprise Metadata Management tool can harvest metadata from HCatalog, Hive, Cloudera Impala, and any files that reside on HDFS. Oracle's key strengths lies in its security and governance capabilities, highly scalable data movement and transformations, and tight integration with Oracle Big Data Appliance.

- **Pentaho's tight Hadoop integration delivers a scalable platform.** Pentaho is an open source data integration and analytics vendor founded in 2004. The top reasons why customers use Pentaho for big data integration include a 360-degree view of the customer, augmentation of existing data warehouses with Hadoop and NoSQL, and a data refinery that streamlines the process of blending and refining diverse data inside Hadoop to support fast interactive analytics. Like most other integration vendors, Pentaho provides native integration with common Hadoop

distributions, NoSQL databases, and some in-memory technologies. One of the largest big data integration deployments uses several petabytes of data; it captures millions of stock trades and performs advanced algorithms on the data. Although Pentaho runs on the public cloud, it does not have a preconfigured installation. Pentaho's strengths are its ability to build and execute MapReduce jobs visually through the Pentaho data integration tool; the job can be executed outside or inside Hadoop. Pentaho can also run its data integration in a clustered format within Yarn, delivering elasticity to run on many nodes depending on the data and analytical needs. Pentaho also provides ease of integration through visualization.

- **SAP leverages Data Services and Hana for real-time integration.** SAP has built a strong data management framework to support data access, data movement, data quality, transformation, and integration. SAP Data Services integrates disparate data sources, including data warehouses, databases, applications, and Hadoop distributions such as those from Hortonworks, MapR Technologies, and Cloudera. SAP Data Services combines with SAP Hana to store and process huge amounts of data in real time. Customers using SAP big data sources will find that SAP Data Services offers a strong framework for big data integration.

- **SAS Institute delivers a scalable, integrated big data solution.** SAS integrates with NoSQL, in-memory databases, and most Hadoop distributions using native access engines to SAP Hana, Hadoop, and Cloudera Impala. It also allows access to HBase via Hive connectivity. SAS uses SAS/Access engines to integrate with cloud, SaaS, and traditional data sources with a combination of native access and embedded open database connectivity drivers. One of SAS's largest customers, a financial services firm, uses SAS in a variety of business lines including capital markets trading, risk management, and transaction monitoring. It processes well over a petabyte of data annually using the SAS data integration platform. SAS's solution runs on Hadoop appliances and can be deployed on SAS cloud or Elastic Compute Cloud (EC2) from Amazon Web Services (AWS). Customers use SAS's big data integration solution for scaling ETL processing, data analysis, data warehouse augmentation, and supporting archival and analytical sandboxes. SAS's top strengths are support for distributed in-memory analytics, push-down processing in Hadoop, in-database processing for databases and data warehouses using the SAS Embedded Process, and its ease of use.

- **SnapLogic leverages an extensive array of robust connectors.** Customers use SnapLogic's big data integration solution to support data preparation, data wrangling for analytics, data warehouse augmentation, multisource data aggregation for exploration, and delivery of filtered data. SnapLogic integrates bidirectionally with NoSQL databases such as MongoDB and HDFS. It also integrates with cloud, SaaS, and traditional databases and data warehouses such as Teradata, Pivotal Software, Oracle, IBM DB2, and SQL Server. Currently, SnapLogic does not run on an appliance, but its Elastic Integration Platform is a multitenant cloud service that runs on AWS. SnapLogic runs natively on Hadoop clusters including Apache Hadoop, Cloudera, and Hortonworks using Yarn and MapReduce. It turns a Hadoop cluster into a SnapLogic Hadooplex that allows it to create pipelines to execute side by side with other Hadoop jobs. Through Yarn, it can scale in or out depending on resource utilization and constraints.

- **Syncsort's contribution to Hadoop makes it a unique integration vendor.** Syncsort is the only ETL engine that runs natively within each node of a Hadoop cluster via its contribution to Apache Hadoop. This allows MapReduce to call out external sort or even suppress sort during MapReduce processing, increasing performance by a factor of two or more. The software ships with Hadoop releases that incorporate Apache Hadoop 2.0, including Cloudera, Hortonworks, MapR Technologies, and Pivotal HD. Syncsort's top use cases for big data integration include offloading heavy ELT workloads and associated data from data warehouses and mainframes into Hadoop and using an agile data platform to quickly collect, process, and distribute data from new and existing big data sources. It can integrate big data sources such as Hadoop, NoSQL, and legacy data warehouses. Syncsort runs natively within Yarn and MapReduce for scale and high performance. One of the vendor's largest big data integration deployments consists of a 400-node Hadoop cluster where Syncsort is used to accelerate the development and execution of Hadoop jobs. Recently, Syncsort introduced Ironcluster Hadoop ETL, which allows enterprises to deploy a full-featured ETL environment on AWS EC2 and Amazon Elastic MapReduce.

- **Talend generates native Hadoop code for scalable big data.** Talend is a commercial open source data integration vendor founded in 2005. It was one of the first vendors to support Yarn and MapReduce and has just made technical previews of Apache Spark and Storm available to enable real-time and operational big data. Talend provides a unified platform that supports ETL, extract, load, and transform (ELT), ELT-like Hadoop loading and transformation (Pig and Hive), log- and trigger-based change data capture, data quality, MDM, and Java Message Service-compliant/enterprise service bus message queues. Talend Platform for Big Data simplifies the process of working with Hadoop distributions: You can move data into HDFS, Hive, or NoSQL databases, perform operations on it inside Hadoop, and extract it without having to do any coding. In the Eclipse-based Talend user interface, you can drag, drop, and configure graphical components representing Hadoop-related data transformation and data quality operations and natively connect to applications, databases, NoSQL, and IoT. Talend automatically generates the corresponding native MapReduce code for transforming data using the Hadoop cluster. One of the vendor's largest projects is a 700-node cluster deployment for a major financial institution that processes 3 to 4 terabytes of new data per day.

RECOMMENDATIONS

BIG DATA INTEGRATION SHOULD BE PART OF YOUR BIG DATA STRATEGY

Although just about every enterprise can benefit from big data integration, it's not a silver bullet that will fix all of your data integration issues. Leverage big data integration for small to medium-size projects before tackling bigger ones, establishing the process, approach, and architecture. Enterprise architects must ensure that their organization's big data integration strategy:

- **Looks beyond traditional data integration solutions.** Big data integration is more than just integrating with Hadoop — it's about dealing with extreme data volumes, leveraging the Hadoop ecosystem (e.g., Yarn, Hive, HBase, and Pig) to store, process, transform, and enrich data, and natively integrating with other big data sources such as NoSQL, files, and streaming and in-memory data.

- **Clearly defines the initial big data integration use cases.** Big data integration projects can be complex, making it difficult to know what approach to take, especially when dealing with extreme data volumes. Ensure that your use cases are well-defined. Are you pursuing data warehouse augmentation, customer analytics, IoT machine or device analysis, or a 360-degree view of the employee? How and where will you need to integrate data for each scenario? This approach helps to speed up big data projects and ensure that the right big data integration tools are used.[3]

- **Looks at solutions that can automate the big data integration process.** Many vendors, such as Cisco Systems, IBM, SAP, and SAS Institute, now offer integrated appliances that help accelerate big data integration deployments. Look for vendors that can minimize big data application disruption when upgrading appliances; consider how modular the appliances are in extending the platform to support additional processors and memory.

- **Leverages Hadoop to support big data integration.** The majority of big data being stored today in Hadoop and NoSQL is semistructured and unstructured data. Look for solutions that can scale to petabytes, integrate Hadoop clusters concurrently, and integrate with large data warehouses, databases, and files.

- **Includes asking vendors to provide a big data integration road map.** If you're looking at traditional data integration solutions, ask the vendor how it plans to extend its solution to support big data sources. Look at the various components that the vendor has integrated and how it supports integration with Hadoop, HBase, NoSQL, cloud, and traditional relational database and file sources.

- **Leverages Hadoop to scale big data integration.** Hadoop can be a data source, but it can also be a compute grid to help store, process, and transform big data. Look for solutions that can scale to petabytes, using Hadoop clusters to process large amounts of structured and unstructured data sets.

- **Standardizes on one big data integration solution initially.** Big data integration solutions are still evolving to integrate more big data sources and tighter integration with the Hadoop ecosystem, NoSQL, cloud, and IoT sources. Unlike traditional structured data sources, where enterprises typically use two or even three different ETL tools to lower costs and

meet specific data integration requirements, look at just one big data integration solution for now. Choosing one will help reduce risk and time-to-value and will spur staff to become knowledgeable about the big data integration approach and architecture.

WHAT IT MEANS

ACCELERATE BIG DATA PROJECTS BY LEVERAGING BIG DATA INTEGRATION

Enterprises are spending more time integrating big data sources than on any other big data management function, including loading, transforming, processing, or securing. Integrating a dozen structured, unstructured, and semistructured big data sources with varying velocities and extreme volumes can be very challenging. Although enterprise architects and their companies can integrate big data sources by writing code and leveraging big data application programming interfaces and connectors, such an approach is time-consuming and requires skilled developers. The bottom line? In order to succeed in big data initiatives, look at big data integration solutions that can help automate and simplify the integration process, reducing the time-to-value.

ENDNOTES

[1] Sensor, geolocation, or Internet of Things data can be infrequent, having no fixed internal schedule. New data from some sources may arrive every minute, while for others it may be every 5, 10, or 60 minutes. It's difficult to know what, how, and when to integrate data that is received infrequently.

While Apache open source products have significant maturity for "schema on read" use cases, innovation is occurring in automated ingestion and flexible schema creation on capture. Forrester expects continued growth in this area as big data solutions move from scientific experimentation to production; however, we think that the total business value in the equilibrium phase will be moderate.

[2] Forrester defines data virtualization and the information fabric as the integration of any data in real time or near real time from disparate data sources, whether internal or external, into coherent data services that business transactions, analytics, predictive analytics, and other workloads and patterns.

Enterprises face growing challenges in bridging disparate sources of data to fuel analytics, predictive analytics, real-time insights, and applications. For more information about guiding your data virtualization strategy, see the August 8, 2013, "Information Fabric 3.0" report.

[3] Big data patterns are still evolving, but look at analytics, predictive, and new insights as the top use cases. See the June 11, 2013, "The Patterns Of Big Data" toolkit.

FORRESTER®

About Forrester

A global research and advisory firm, Forrester inspires leaders, informs better decisions, and helps the world's top companies turn the complexity of change into business advantage. Our research-based insight and objective advice enable IT professionals to lead more successfully within IT and extend their impact beyond the traditional IT organization. Tailored to your individual role, our resources allow you to focus on important business issues — margin, speed, growth — first, technology second.

FOR MORE INFORMATION

To find out how Forrester Research can help you be successful every day, please contact the office nearest you, or visit us at www.forrester.com. For a complete list of worldwide locations, visit www.forrester.com/about.

CLIENT SUPPORT

For information on hard-copy or electronic reprints, please contact Client Support at +1 866.367.7378, +1 617.613.5730, or clientsupport@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

Forrester Focuses On Enterprise Architecture Professionals

By strengthening communication and collaboration across business lines and building a robust, forward-looking EA program, you help transform your organization's business technology strategies to drive innovation and flexibility for the future. Forrester's subject-matter expertise and deep understanding of your role will help you create forward-thinking strategies; weigh opportunity against risk; justify decisions; and optimize your individual, team, and corporate performance.

« **ERIC ADAMS,** client persona representing Enterprise Architecture Professionals