SOLIX®
Empowering Data Management

THE CIO GUIDE TO BIG DATA ARCHIVING

How to pick the right product?

Includes Forrester Market Overview for Big Data Archiving

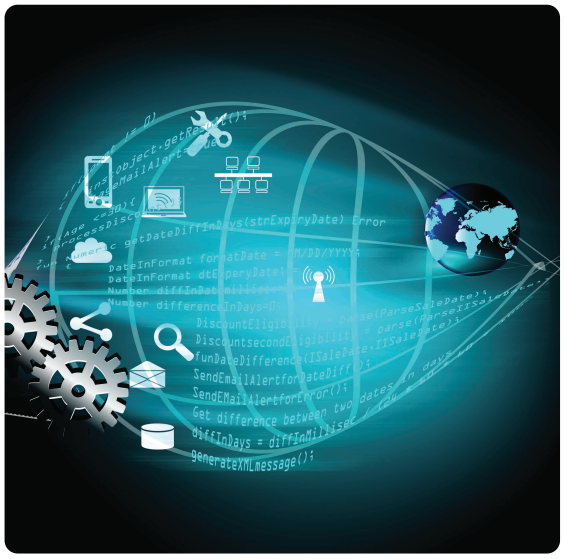
FORRESTER®

All Forrester references that appear in this document are from the August 2015 Forrester Research report, 'Market Overview: Big Data Archiving.

The landscape of enterprise data is changing with the advent of enterprise social data, IoT, logs and click-streams. The data is too big, moves too fast, or doesn't fit the structures of current database architectures.

As Forrester points out, "with growing data volume, increasing compliance pressure, and the evolution of Big Data, enterprise architect (EA) professionals should review their archiving strategies, leveraging new technologies and approaches."

> *"In The Era Of Big Data, Archiving Is A No-Brainer Investment."*
>
> – Forrester Research.



In recent years, we have clearly seen a trend towards Big Data technology adoption, and this adoption can be accelerated by simplifying the ingestion, organization, and security of enterprise data within Big Data platforms.

As Forrester points out, Big Data technologies open up new possibilities for data archiving, "by leveraging open standards, integration, consolidation and scale-out platforms. Hadoop and NoSQL can store and process very large volumes of structured, unstructured, and semi-structured data and enable search, discovery, and exploration for both compliance and analytical purposes."

Archiving is an important first step towards Big Data adoption that allows organizations an opportunity to create a simpler, scalable, and economical data management strategy.

Furthermore, a consolidated Big Data archive makes analytics, machine learning, search, and predictive analytics more straightforward than storage of data in multiple repositories. The value of enterprise data can be maximized through analytics, and the Big Data archive must provide the flexibility to integrate with variety of industry-specific and function-specific analytic tools.

Merely dumping data into an Apache Hadoop repository is not going to provide any insight. Plus, most companies use ETL tools or custom scripts to copy data into Big Data repositories. Not only does this create risk through proliferation, it potentially violates compliance and regulatory guidelines. Picking the big data vendor requires some careful consideration.

> *Big data archiving uses Hadoop and NoSQL technologies to store, process, and access information that helps deliver a 360-degree view of the business, product, and customer as well as meeting compliance and governance requirements.*
>
> – Forrester Research.

All big data products claim to be scalable, high performance and low cost. So, how does a CIO pick the right product for BIG DATA archiving?

HERE ARE 12 QUESTIONS TO ASK WHEN SELECTING A BIG DATA ARCHIVING VENDOR:

1 Does the product have prebuilt archiving rules and templates for structured data applications like ERP (Oracle EBS, SAP, PeopleSoft and CRM (Seibel, Salesforce)?

Note: *Most vendors provide generic RDBMS connectors or open source connectors like Sqoop. These connectors allow only table data copy, with no application awareness, requiring expensive service engagements to customize the archiving.*

WHY IS THIS IMPORTANT:

- Enterprise Applications have large, complex schemas.
- Master data and transactions span 1000s of tables with referential constraints.
- Context of application data must be maintained upon archival.
- Out-of-the-box knowledge base is required for repeatable, efficient archiving and for maintaining data integrity.
- Without application specific templates, service engagements can run into months.

2 Can the product archive unstructured data from SharePoint, File shares, Desktops, Laptops, FTP, websites, etc.?

Note: *Most vendors use third party products and connectors to archive file shares and SharePoint data which may be difficult to integrate, error prone and might not preserve all necessary information across the vendor products.*

WHY IS THIS IMPORTANT:

- ~80% of enterprise data is unstructured; and 60% is stale or not business related.
- Purpose-built connectors can preserve context and metadata along with the original source files.
- Cheap storage is ideal to archive fast growing data within SharePoint, File Shares, etc.
- Consolidating all unstructured data into one repository will simplify information governance and compliance.

3 Can the product MOVE data into the archive versus COPY?

> ***Note:*** *Most vendor products merely COPY data using tools like Sqoop, Flume, or Talend, etc.. A COPY function does not also PURGE the source data to reduce data load on the source database to improve application performance.*

WHY IS THIS IMPORTANT:

- Purging less active data from the source application improves performance and lowers cost.
- Atomicity of move is essential to data consistency.
- Purge routines are high risk, complicated and may cause compliance issue.
- Copying the data only results in more data growth and does not contribute to improved application performance which is a primary goal of database archiving.

4 Does the product employ data validation algorithms like MDS, SHA-1 and summation on structured and unstructured data?

> ***Note:*** *Most vendor products do not provide these capabilities, leaving the customer exposed to non-compliance and risk.*

WHY IS THIS IMPORTANT:

- Archived data must be identical to the original source data.
- Product should provide necessary data validation reports for regulatory purposes.
- For unstructured data, validation algorithm like MDS or SHA-1 are required to validate the accuracy of the archived file.
- For structured data, algorithms like checksums, column summations, etc. should be used to validate the archived data.

5 Does the product provide Information Lifecycle Management (ILM) for retention management, legal hold, eDiscovery of structured and unstructured data?

Note: *ETL tools, Sqoop, and other scripts used to copy data typically save the data as delimited (CSV) files within HDFS. These products lack retention policies and legal holds at a record level. Plus, no audit trail is maintained for the COPY / MOVE operation, which is a compliance gap.*

WHY IS THIS IMPORTANT:

- Archived data must be purged in a compliant manner based on retention policies and business rules.
- Archive must support "Legal Hold" on the data at record-level and file level.
- Effective ILM policies can prevent data proliferation ongoing through policy based, active archiving.
- Archive must track all data and provide an audit trail of all information.

6 Does the product provide secure, role based access for all the archived data?

Note: *Most products will integrate with Active Directory to allow user logins, but they will not limit user access at an application or defined role level.*

WHY IS THIS IMPORTANT:

- Archives contain data from multiple applications.
- Each application's data can be accessed and viewed by granular roles or groups.
- Active Directory I Kerberos allow the necessary mapping between users and their group roles.
- Users should be able to use their existing credentials to access data based on role based privilege.
- For compliance purpose all access must be tracked with an audit trail.

7 Does the product support text search and SQL queries or reports for all the archived data?

Note: *Most products do not provide out of the box text searching. Custom data de-normalization and text search tools are required for structured data to become text searchable.*

WHY IS THIS IMPORTANT:

- Archived data should be easily retrievable by the users using simple query, text search and reporting.
- For eDiscovery and compliance use cases, all content should be searchable.
- For trend reporting and case assessment, metadata from unstructured files should be made query-able through Hive.
- Text searching should be available for most popular file formats within SharePoint, and file shares, as well as BLOBS and CLOBS attachments or formatted columns within databases.
- Archived data must be accessible for administration, analysis, and compliance.
- An integrated, browser based, user-friendly interface is required for the business user.

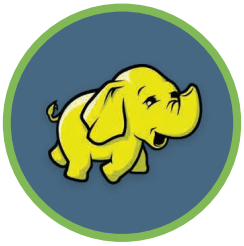
8 Does the product support print stream archiving?

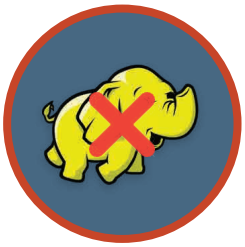
Note: *If available at all, most companies use isolated infrastructure for print stream capture, with no searching and reporting capabilities available. Captured data is maintained in a silo without any integration with other applications.*

WHY IS THIS IMPORTANT:

- Most enterprises are required to preserve print and fax data for compliance purposes.
- Reports from archived and retired applications should be preserved in their original format.
- Correlation of print stream with other enterprise apps is necessary for business intelligence.
- Print stream archiving is often the fastest, cheapest and effective solution.

9 Does the product support different HDFS file formats and compression algorithms for archiving (Parquet, ORC, Avro, CSV, snappy, zlib) for archiving?





Note: *CSV and other text formats are not optimized for queries. Most archiving products use Sqoop, custom scripts, or 3rd party tools to integrate with Apache Hadoop. These integrations have to be manually modified when a new file format is required. This is typically time consuming, error prone, and expensive.*

WHY IS IT IMPORTANT:

- Parquet and ORC are columnar file formats that significantly improve query performance for large data volumes.
- Avro file format supports dynamic typing and schema evolutions.
- For large archives, compression optimizes storage utilization, relieves IO bottlenecks, and improves performance.

10 Is the product certified on Cloudera and Hortonworks?

Note: *Many vendors claim to have Big Data support without Cloudera or Hortonworks certification. Furthermore, some vendors claim to have Big Data support if they can write to a NAS device that frontends Hadoop. Using a NAS device can be expensive and it leverages none of the real capabilities of an Apache Hadoop stack.*

WHY IS THIS IMPORTANT:

- Cloudera and Hortonworks provide the necessary enterprise support for Apache Hadoop.
- Certified solutions are tested and verified on a stable release of Apache Hadoop.
- Security fixes, patches, and upgrades are delivered sooner on supported distributions such as Cloudera and Hortonworks.

11 Can the product keep the archive in sync with the source application after upgrades and patches?

Note: *Most vendors require archive data to be upgraded along with the original application. Other vendors use a proprietary format for the archive, making access to the data complicated.*

WHY IS THIS IMPORTANT:

- Enterprise applications are periodically upgraded for business and technical reasons.
- Upgrades introduce changes to schema and data organization.
- The archiving engine should work seamlessly without any impact to the archive.

12 Does the product support de-archiving?

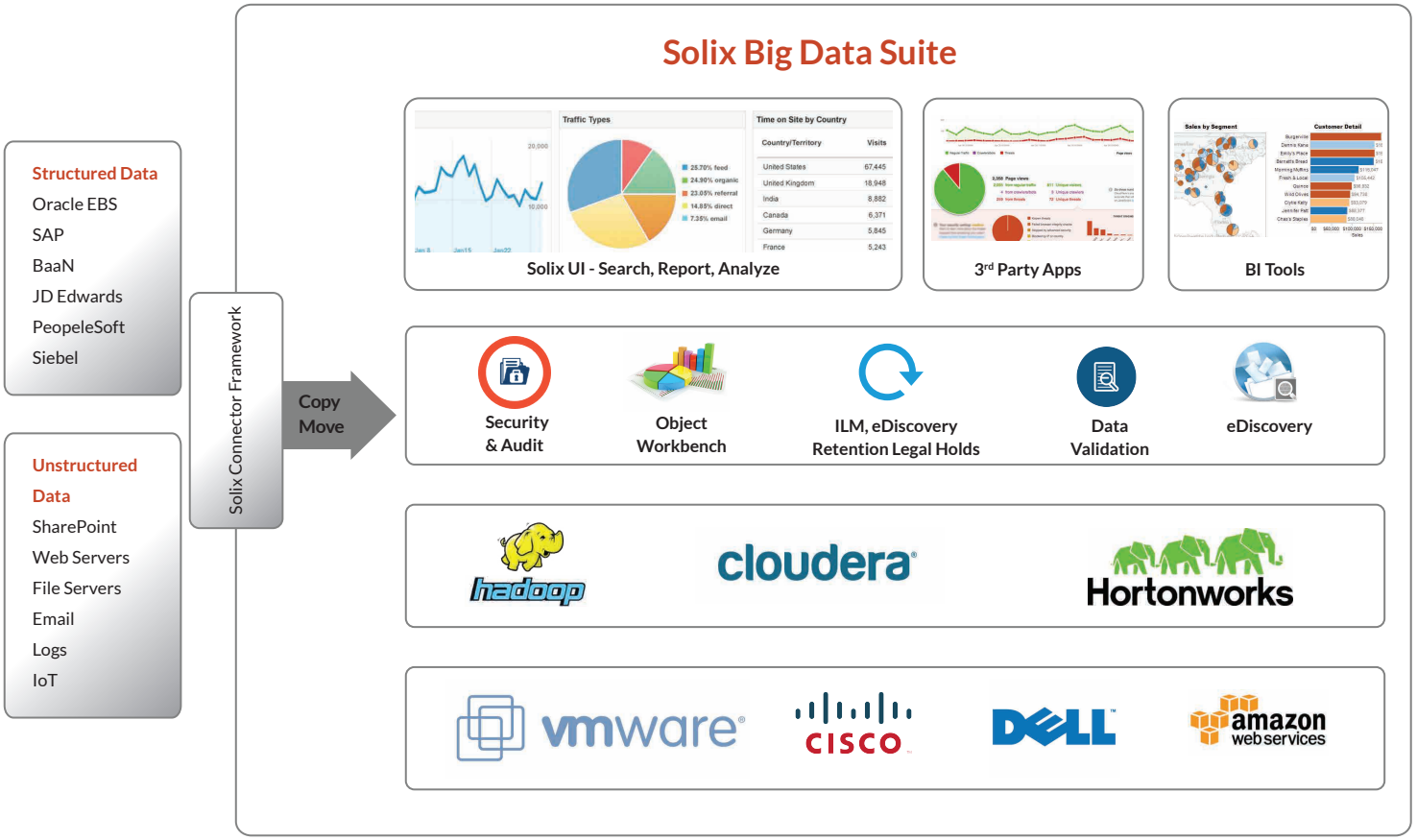
Note: *Most archiving tools do not support de-archiving operations. Can source data be manually moved back into the source database?*

WHY IS IT IMPORTANT :

- For business or compliance reasons, sometimes its required to move data back from the archive into the live application repository.
- De-archiving is a complex process that requires reversing the archival process.
- De-archiving must ensure the data integrity of the original applications for compliance reasons.

SOLIX BIG DATA SUITE

Solix Big Data Suite is a comprehensive enterprise archiving solution for Apache Hadoop. Solix is certified on Cloudera CDH and Hortonworks and provides an out of the box solution to accelerate enterprise archiving and enterprise data lake projects.



As shown in the figure above, Solix Big Data Suite holds the key to enterprise archiving:

1. Structured Data Archiving	Solix supports an integrated knowledge base for enterprise applications like Oracle EBS, SAP, PeopleSoft, Seibel, Baan, etc. for Enterprise Archiving or Data Lake.
2. Unstructured Data Archiving	Solix unstructured data connectors can archive data from SharePoint, file shares, or FTP with no additional development.

3. Move and Copy Operations	Solix can perform atomic MOVE or COPY of data from source applications.
4. Data Validation	Solix provides full validation for structured and unstructured data operations using algorithms like SHA1, MDS, etc.
5. Integrated ILM & Governance	Solix provides Information Lifecycle Management (ILM) capability for the entire archive with retention management, legal hold, eDiscovery, etc.
6. Enterprise Security & Role Based Security	Solix integrates with Active Directory and Kerberos to provide secure role-based access to all the data.
7. Integrated UI for Searching & Reporting	All data within Solix is automatically text search, query and reportable on content and metadata. No additional products or licenses are required.
8. Print Stream Capture	Solix Virtual Printer can capture print streams into a PDFIA format with the necessary ILM and GRC capabilities.
9. HDFS file formats and compression	Solix support HDFS formats like Parquet, ORC, Avro, etc. and compressions like snappy, zlib, etc.
10. Certified on Apache Hadoop distributions	Solix is currently certified on Cloudera and Hortonworks. Roadmap includes other distributions like MapR, Amazon EMR.
11. Support for Application Upgrades	Solix Object Workbench provides an efficient application decoupling methodology for enterprise archiving which enables upgrades with no impact to the archiving and minimizing IT down time.
12. De-archiving	Solix supports de-archiving and can move archived data back into the original source database in a compliant manner.

The Solix Big Data Suite provides an extensive ILM framework to create a unified repository to capture and analyze all enterprise data with analytics tools. The suite is highly scalable with an extensible connector framework to ingest all enterprise data. The integrated suite allows seamless archiving, application retirement, and flexible extract – transform – load (ETL) capabilities to improve the speed of deployment, decrease cost, and optimize infrastructure. Solix also supports on premise and cloud based deployment on a variety of Hadoop distributions.

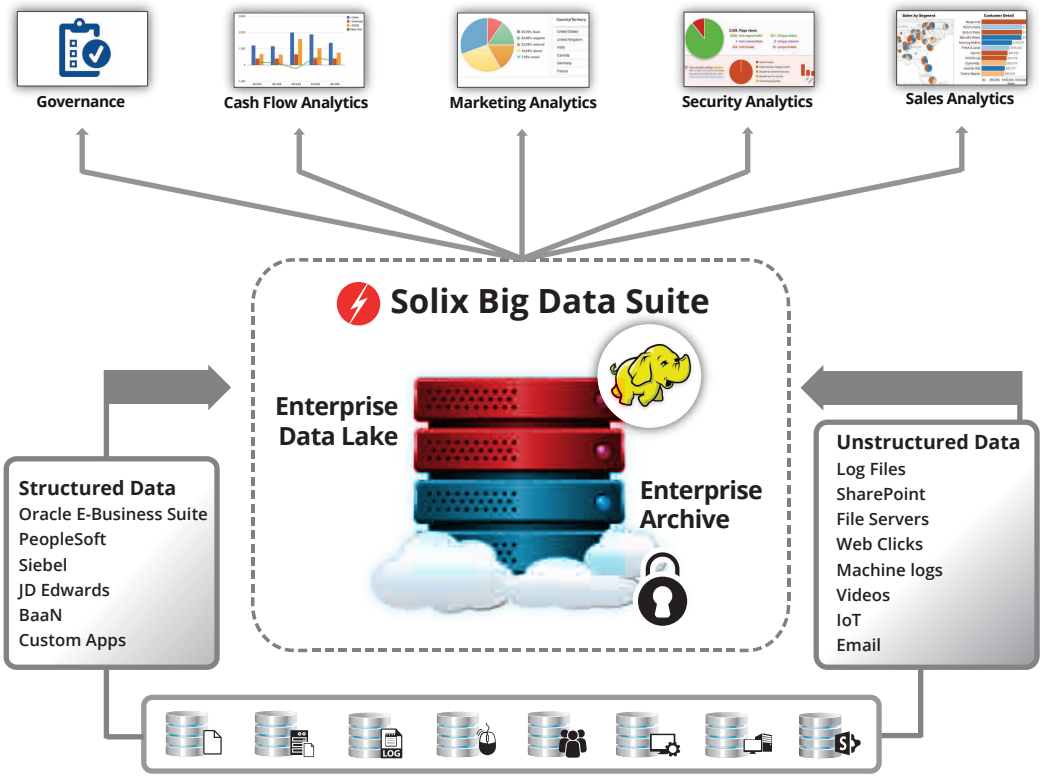
The Solix Big Data Suite harnesses the capabilities of Hadoop to create a comprehensive, efficient, unified and cost-effective ILM repository for all data.

THE SOLIX BIG DATA SUITE INCLUDES:

- Solix Enterprise Archiving to improve enterprise application performance and reduce infrastructure costs. Enterprise application data is first moved and then purged from its source location according to ILM policies to ensure governance, risk, and compliance objectives are met.

- The Solix Enterprise Data Lake reduces the complexity and processing burden to stage enterprise data warehouse (EDW) and analytics applications and provides highly efficient, low-cost storage of enterprise data for later use when it is needed. Solix Data Lake provides a copy of production data and stores it "as is" in bulk for later use.

- The Solix App Store offers pre-integrated analytics tools for data within Enterprise Archiving and the Enterprise Data Lake.

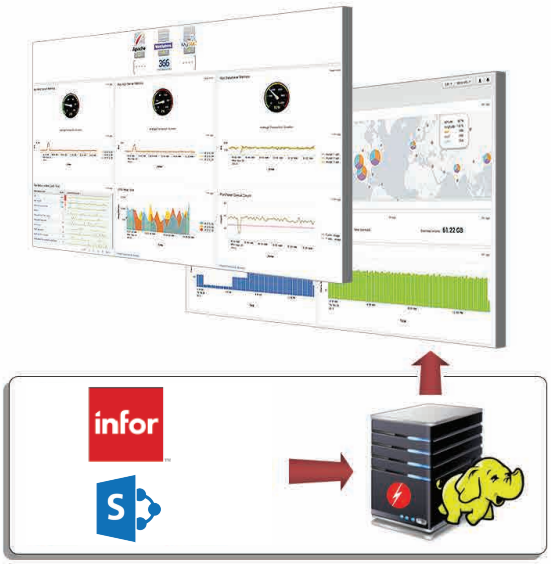


BIG DATA CUSTOMER SUCCESS STORIES

Consolidated Enterprise Archive for Manufacturing Application and SharePoint

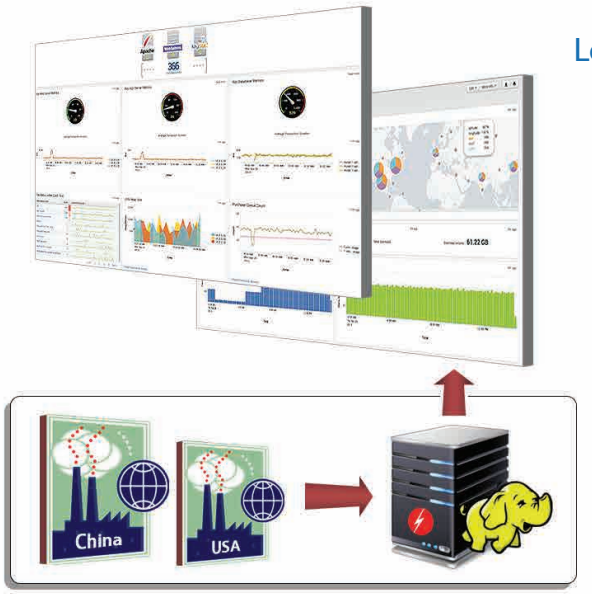
Customer: National Building Supplies Distributor

Solix Big Data Suite was used to retire a legacy manufacturing application into Apache Hadoop.



SOLUTION BENEFITS:

- Preserve original application data for compliance
- Use Solix UI for reporting I searching
- Generate necessary audit reports for GRC
- Save license cost for the manufacturing application
- Apply ILM, Retention Management, Legal Hold
- Search, Report on all archived data through Solix



Log File Archiving for Threat Detection and Security Analytics

Customer: Publically Traded Technology Company

Solix Big Data Suite was used to archive network and security logs. Consolidated archive was used to build a dashboard for threat analysis.

SOLUTION BENEFITS:

- Single repository to aggregate all security logs
- Support for structured and unstructured data
- Ability to process large data set for better analysis
- User friendly dashboard for visualization and event correlation

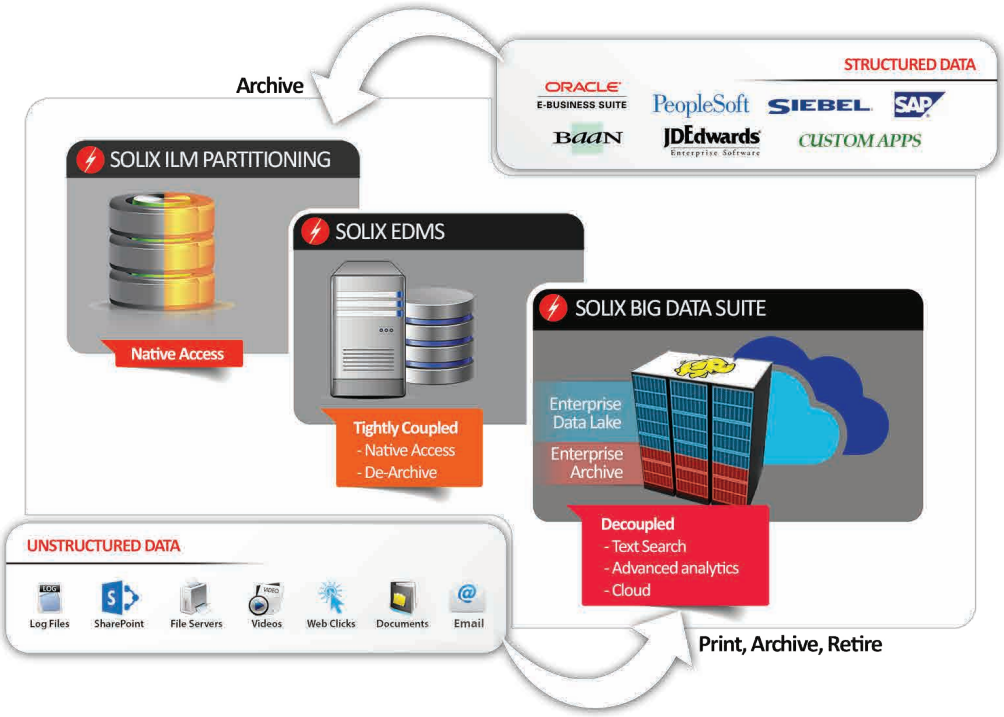
CONCLUSION

As Forrester stated in their recent research report, "In The Era Of Big Data, Archiving Is A No-Brainer Investment."

However, enterprise archiving is not trivial, and selecting the right product requires careful consideration. The archiving product must be a purpose-built platform with the necessary security and ILM framework in place.

The platform must be built on open standards, integrate with analytics tools, and offer APIs to meet the needs of different industries-it must leverage the best-of-breed technologies for both, enterprise archiving and analytics.

Solix Big Data Suite delivers that ideal platform that can offer immediate ROI and help the CIO maximize the value of enterprise data.



Hortonworks CERTIFIED

cloudera CERTIFIED

CIO's Guide to Big Data Archiving

12 questions to ask when selecting the Big Data archiving vendor:

1. Does the product have prebuilt archiving rules and templates for structured data applications such as ERP (Oracle EBS, SAP, PeopleSoft) and CRM (Seibel, Salesforce)?
2. Can the product archive unstructured data from SharePoint, File shares, Desktops, Laptops, FTP, websites, etc.?
3. Can the product "move" the data into the archive as opposed to copying and proliferating the data?
4. Does the product employ data validation algorithms such as MD5, SHA-1, summations, etc. on structured and unstructured data?
5. Does the product provide ILM (retention management, legal hold, eDiscovery) for structured and unstructured data?
6. Does the product provide secure, role based access for all the archived data?
7. Does the product support keyword searches and SQL queries of all the archived data?
8. Does the product support print stream archiving?
9. Does the product support different HDFS file formats (Parquet, ORC, Avro, CSV) and compression algorithms (snappy, zlib) for archiving?
10. Is the product certified on Cloudera and Hortonworks?
11. Can the product keep the archive in sync with the source application after upgrades and patches?
12. Does the product support de-archiving?

Solix Technologies, Inc.
4701 Patrick Henry Dr., Bldg 20
Santa Clara, CA 95054
Phone: 1.888.GO.SOLIX (1.888.467.6549)
1.408.654.6400
Fax: 1.408.562.0048
URL: http://www.solix.com

FORRESTER®

FOR ENTERPRISE ARCHITECTURE PROFESSIONALS

Market Overview: Big Data Archiving

In The Era Of Big Data, Archiving Is A No-Brainer Investment

by Noel Yuhanna
August 11, 2015

Why Read This Report

Many organizations fail to see data archiving as part of their business technology agenda because they perceive archiving as highly complex and delivering low business value. However, with growing data volume, increasing compliance pressure, and the evolution of big data, enterprise architect (EA) professionals should review their archiving strategies, leveraging new technologies and approaches. Big data archiving uses Hadoop and NoSQL technologies to store, process, and access information that helps deliver a 360-degree view of the business, product, and customer as well as meeting compliance and governance requirements. This report profiles 12 vendors of big data archiving solutions and describes technology enhancements and use cases.

Key Takeaways

Inactive Data Volume Continues To Grow, Creating New Data Challenges

Many enterprises treat inactive data as an overhead because of its low business value and the significant cost associated with storing and managing it for compliance and other business needs. However, with the growing volume of inactive data, organizations are looking at ways to store, process, and access it.

Big Data Technologies Open Up New Possibilities For Data Archiving

Big data technologies help support new approaches to data archiving by leveraging open standards, integration, consolidation, and scale-out platform. Hadoop and NoSQL can store and process very large volumes of structured, unstructured, and semi-structured data and enable search, discovery, and exploration for both compliance and analytical purposes.

The Big Data Archiving Vendor Landscape Consists Of New And Traditional Vendors

This report highlights 12 big data archiving vendors and open source projects. Related Forrester reports cover comprehensive archive platforms and archive solutions for content, social media, and email and messaging.

Market Overview: Big Data Archiving

In The Era Of Big Data, Archiving Is A No-Brainer Investment

by Noel Yuhanna
with Leslie Owens, Cheryl McKinnon, Elizabeth Cullen, and Diane Lynch

August 11, 2015

Table Of Contents

2 **Big Data Archiving Is A New Approach That Creates New Possibilities**

The Benefits Of Hadoop And NoSQL For Data Archiving Are Enormous

Keep In Mind That Big Data Archive Integration And Security Still Need Attention

5 **Big Data Archive Market Is Growing Rapidly**

What It Means

10 **Rethink Your Data Archive Strategy With A Big Data Approach**

Notes & Resources

Forrester interviewed 12 vendor companies, including Cloudera, Commvault, EMC, Hortonworks, HP Inc., IBM, Informatica, MapR Technologies, SAP, Solix, Teradata, and ZL Technologies.

Related Research Documents

Digital Insights Are The New Currency Of Business

Market Overview: Information Archiving, Q2 2015

TechRadar™: Big Data, Q3 2014

FORRESTER®

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
+1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

Big Data Archiving Is A New Approach That Creates New Possibilities

Many enterprises store inactive data in their respective applications or in a siloed, archived data repository such as a file or database; such data is not further integrated to provide a broader context of the customer, business, or product (see Figure 1). Forrester believes that inactive data accounts for nearly two-thirds of all enterprise data, which various business applications, systems, and devices create. Big data archiving is a new approach to data archiving that leverages recent advancements in distributed data technologies such as Hadoop and NoSQL and allows organizations to store various data types and sources in a centralized data repository (see Figure 2). Besides supporting the need for compliance and data governance, big data archiving can help enterprise architects discover new patterns, trends, and actionable insights to deliver innovative products and services and predict the future.[1] Big data archiving allows for the consolidation of archived data from multiple applications such as CRM, enterprise resource planning (ERP), supply chain management (SCM), and custom applications in a centralized repository that is more scalable and economical.

Forrester defines big data archiving as:

> *"The capability to store, process, and access inactive structured, unstructured, and semi-structured data in a big data platform such as Hadoop and NoSQL for long-term compliance and analytical purposes."*

A retailer was able to use big data archiving, along with active data, to understand product sales trends across countries, which resulted in optimizing inventory, saving over $5 million annually, and improving customer satisfaction by ensuring that products were always available to meet demands. In addition, the archived data was available for reporting, search, and legal discovery for compliance purposes. As a technology management manager in a large North American retail company recently stated:

> "Without using the archived data in Hadoop, we would not have an accurate inventory prediction for our product line. The archived data also serves us for our compliance requirements — it's a dual strategy, and it's a win-win for various teams involved. Previously, our archived data was used only for compliance, and that, too, spanned many files and repositories, which was problematic."

FIGURE 1 The Traditional Data Archiving Framework

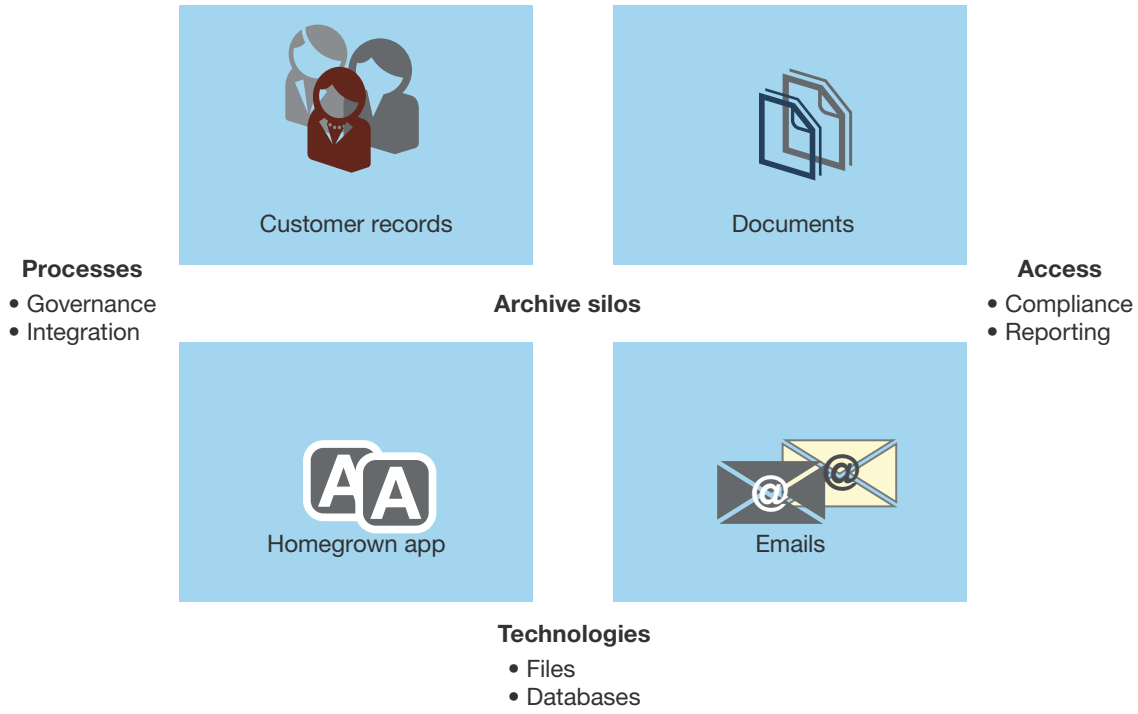
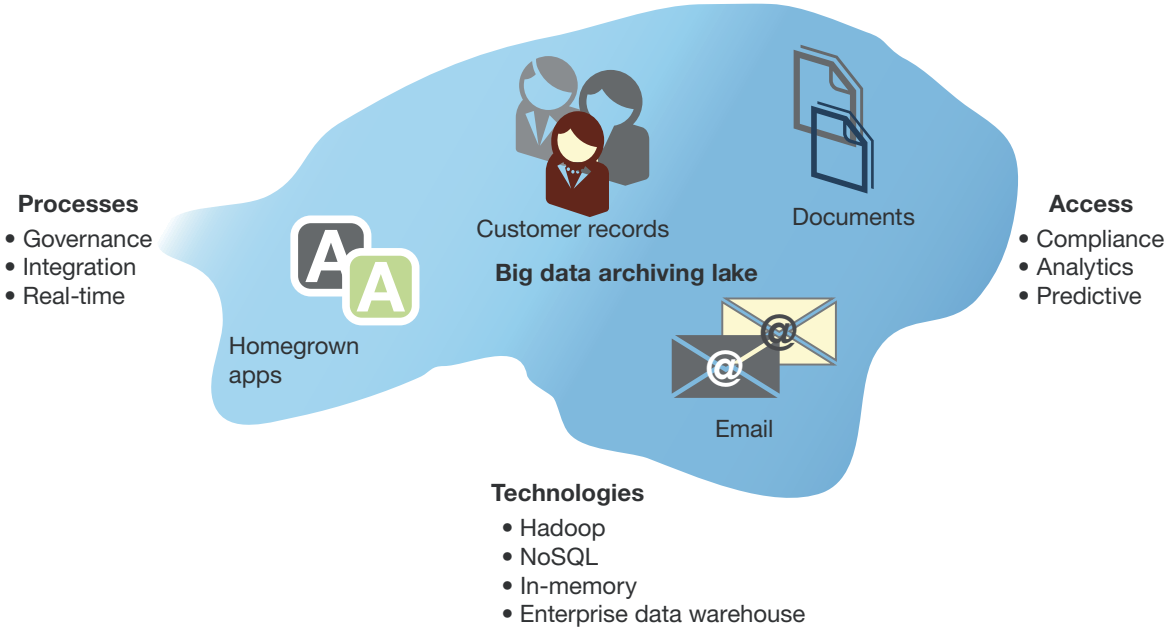


FIGURE 2 The Big Data Archiving Framework



The Benefits Of Hadoop And NoSQL For Data Archiving Are Enormous

Organizations have been archiving data for decades, so why transition to the big data archive? First, the emergence of big data technologies such as Hadoop and NoSQL offers new, next-generation data management features and functionality that make archiving data simpler, scalable, and economical. Second, big data archiving minimizes complexity, especially when dealing with all kinds of structured, unstructured, and semi-structured data. For example, schema-on-read enables storage of data in Hadoop in raw format without a formal data structure. Finally, big data archiving makes analytics, machine learning, search, and predictive analytics more straightforward than storage of data in multiple repositories.

Top benefits of using Hadoop and NoSQL for big data archiving include:

- **They're built on open standards.** Apache Hadoop and many NoSQL solutions are open source projects that continue to enhance and improve through community contributions. Big data Apache projects such as Apache Flume, Apache Hadoop, Apache Hive, Apache Mahout, Apache Sqoop, and other projects are built on open standards that help store, process, and access data in a centralized repository. These big data solutions are ideal for long-term data requirements, such as archiving, without worrying about vendor lock-in, discontinued products, or a vendor changing its product strategy or commitment.
- **They support on-demand elastic scale.** Hadoop and NoSQL offer the ability to use commodity servers to scale out horizontally based on the business requirements. Hadoop can handle petabytes of information in a data lake environment, ingesting billions of events a second, and can support all kinds of data for retention, analytics, and compliance needs.
- **They're more economical.** Unlike traditional data repositories, Hadoop and NoSQL are more economical and don't require specialized appliances or hardware architectures to run. As a result, we see many organizations offloading inactive data from expensive servers and storage to Hadoop and NoSQL for long-term retention.
- **They support all your data.** With traditional data archiving approaches, most organizations kept structured data separate from unstructured data, creating many archival silos. Hadoop and NoSQL can support any type of data — structured, semi-structured, and unstructured.
- **They're ready for analytics.** With all data centralized in a big data archive lake, you have the flexibility to run all kinds of analytics for your business, including analytics that weren't possible without centralization, such as customer analytics that require inactive and active data from log files, clickstream, past orders, and CRM application.

Keep In Mind That Big Data Archive Integration And Security Still Need Attention

Hadoop and NoSQL deliver an agile, open, scalable, and economical platform that makes archiving an attractive use case for all enterprises. However, big data archiving is still evolving, and gaps still exist in data management. Note that:

- **Integration of diverse data is complex.** Simply putting lots of diverse data into Hadoop or NoSQL isn't going to create new insights, or even meet compliance requirements, without further integration, transformation, and enrichment. You need context to tie the various data elements together, such as linking archived customers' CRM data with historic emails. While vendors have tried to integrate silos of structured, semi-structured, and unstructured data for long-term retention for years, the process remains complex and requires a lot of manual effort. However, Forrester expects that in the coming years, we are likely to see improved archiving solutions that will be able to associate data across various sources to support enterprise-level search, legal hold, analytics, and predictive analytics in a context-driven and automated manner.[2] Apache projects such as Flume, HCatalog, Kafka, Mahout, Spark, Sqoop, and Storm help in storing, processing, and accessing archive data more efficiently in a Hadoop environment.
- **Data security needs to ramp up beyond the basic features.** Although Hadoop and NoSQL offers basic data security features and functionality, such as data encryption, masking, access control, and authentication, security gaps still exist. The platform lacks comprehensive data security features such as native data masking, end-to-end auditing, vulnerability assessment, and real-time data protection that are often found in traditional platforms such as data warehouses and databases. Enterprise architects need to enforce stronger security measures, especially when dealing with sensitive data.[3]
- **Companies need to clearly define retention rules.** Whether using big data technologies or traditional files or databases to store archived data, organizations need to define retention rules about when to dispose of obsolete data and when to remove it from the archival system.

Big Data Archive Market Is Growing Rapidly

Forrester expects that the big data archive market will see significant momentum in the coming years as organizations combine people, process, and technology to close the gap between insights and action.[4] Here are the top big data archive vendors:

- **Cloudera is a viable archival platform with several large implementations.** Cloudera's Enterprise Data Hub (EDH) is a data architecture based on Hadoop that can store data, economically and reliably, for long-term and diverse use cases. One unified platform can collect, store, process, explore, model, and serve data. It connects directly to existing analytical and operational systems and can be deployed on-premises or in the cloud. Top big data archiving use cases include active archive, data warehouse offload, and compliance archive. Cloudera's key

strengths lie in its scalable Hadoop platform, integrated search, SQL access, automation level, and security. Although Cloudera lacks a complete end-to-end data archiving solution to automate the process of data retention, search, legal hold, and eDiscovery, enterprises can still leverage Cloudera's platform to support a big data archive initiative by using Apache tools such as Flume, Hive, Kafka, and Sqoop. For enterprises that need end-to-end archiving strategy, Cloudera partners with Solix to support such implementation. One of Cloudera's customers stores more than 33 petabytes of raw data online for querying, and another archives more than 2 petabytes of data in Cloudera. Customers leveraging Cloudera for big data archiving include Allstate, Camstar, Cerner, eBay, Epsilon, NetApp, RelayHealth, and YP.

- **Commvault's big data archiving is ramping up to support broader use cases.** Commvault was founded in 1996 and has been offering big data archiving since 2011. The top use cases for big data archiving include information governance, storage management, and information visualization to keep track of and review growth of machine- and user-generated data and make predictions, thus improving awareness and business value. Commvault offers scalable, unified data and information management software design that uses a single management console to provide backup, recovery, archive, replication, search, and analytics. The company plans to support native integration with Hadoop in its next release. Its key differentiators are its support for big data, with converged backup and archive, and the ability to optimize big data environments using content-based retention. Commvault provides a single platform for structured and unstructured data. One of its largest big data file archiving deployments is more than 20 petabytes. Its largest file analysis production customer is more than 500 million objects/files.
- **EMC ramps up its big data archive solution with Hadoop and NoSQL integration.** EMC's data archiving solution has offered data archiving for a decade, but its big data archive is a relatively new offering, similar to those of other traditional vendors. It provides native integration with Hadoop Distributed File System (HDFS) through the storage layer. The top use cases for big data archiving include application infrastructure optimization, application rationalization, and data center consolidation. EMC's big data archiving solution, which includes EMC InfoArchive, is application and technology agnostic; it works with various extract, transform, and load (ETL) and data integration vendors for data ingestion into the Hadoop and NoSQL platform. EMC's core strength is in its ability to unify structured and unstructured data into a single repository, combine multiple data types into a single business record, and optimize the archiving approach based on data elements. The Isilon Archive solution for big data, part of EMC's Isilon scale-out storage, lets customers scale to 50 petabytes in a single cluster. Combined, InfoArchive and Isilon provide a compliant data source for big data archiving and analytics.
- **HP Inc.'s mature and scalable data archiving solution now integrates with Hadoop.** HP Inc. has been offering a data archiving solution for more than a decade and is now integrating its offering with Hadoop. The top use cases for HP Inc. include: 1) analytics and storage of emails, social media, and documents; 2) retention policy enforcement for all data, including production for regulatory and governance; and 3) corporate policy enforcement and application retirement.

HP Inc.'s key strengths lie in the ability to scale on a technology platform that provides advanced analytics/search combined with a complete information governance portfolio and services. HP Inc. has many customers doing big data archiving, including some large enterprises, such as Datang, HP IT, Morgan Stanley, and SanDisk. One of its largest deployments includes a customer that has archived more than 100 billion rows, which amounts to more than 83 terabytes of compressed archived data.

- **Hortonworks offers big data archiving based on open source and partner products.** Hortonworks has been offering a big data archiving solution for more than three years and has several large customer deployments that run into petabytes. Hortonworks Data Platform (HDP) is an enterprise-grade data management platform that is built on Apache Hadoop, powered by YARN, and supported by a comprehensive set of capabilities that address the core requirements of security, operations, and data governance. HortonWorks' key strength lies in its ability to centralize data storage, processing, and access, delivering 100% open source software components, and in its joint engineering with various partners such as EMC, HP, Informatica, Oracle, SAP, Teradata, and others. One of the largest deployments of Hortonworks is more than 400 petabytes, where some of the data is archival data for long-term retention. Hortonworks' recent release of HDP 2.3 offers new capabilities around security, governance, and compliance and is part of the Apache Atlas, which extends industry and enterprise governance from traditional systems to Hadoop.
- **IBM leverages Hadoop for analytics, with plans to use it for archival repository.** IBM has been offering a structured data archiving solution for more than 17 years and is now integrating with the IBM BigInsights Hadoop platform and other distributions. IBM's integration of archive with Hadoop is primarily to support scalable big data analytics. The archival repository in Hadoop complements the governance capabilities of the current archive platform. Today, the key use cases for big data archiving include consolidation as well as retiring applications, archiving data to reduce cost, and improving application efficiency. IBM's core strength is in its governance and compliance solution, a comprehensive archive solution with support for eDiscovery, litigation hold, and defensible delete, proven with hundreds of deployments across various industries. In one deployment, a large telco provider archives about 350 terabytes of structured data a year and, to date, has archived more than a petabyte of data. IBM's big data archive solution is viable for moderate to large implementations; however, for smaller implementations with fewer than 10 terabytes of data, the solution might be overkill.
- **Informatica provides flexibility in archiving to and within Hadoop.** Informatica is one of the leading data archiving vendors, with thousands of customers that use it to store archive data and manage compliance and long-term data retention. Informatica products for big data archiving include Big Data Edition, Data Archive, and Vibe Data Stream. Vibe Data Stream streams data in real time (e.g., machine data and log files) to the archival data store, while Big Data Edition copies and moves data from/to Hadoop natively and prepares (e.g., parses, integrates, and cleanses) data once in Hadoop. Informatica archives data to Hadoop in popular native formats such as Optimized Row Columnar (ORC) and Parquet. In addition, Data Vault, Informatica's long-term data

store for highly compressed archive data, runs on HDFS to leverage low-cost infrastructure and built-in redundancies. Informatica Data Archive delivers traditional archiving capabilities to manage inactive or infrequently used data optimally and delivers secure, immutable, compressed archival format for retention management, legal hold, and activity audits. Top use cases for Informatica's big data archive include retention management within Hadoop, using HDFS as low-cost storage for archiving, offloading EDW data, and archiving for compliance. Informatica's key strengths lie in its comprehensive high-performance connectivity to various data sources, integrated built-in compliance features, and high degree of data compression ratio. Organizations can still leverage Hadoop for analytics across all supported use cases.

- **MapR Distribution delivers a scalable archiving platform.** MapR has been offering a big data archiving solution since 2011, when the MapR Distribution was first available. MapR has archiving as a native capability in the platform, which it has architected to optimize for data resilience for production workloads as well as respective enterprise requirements for archiving. MapR has unique archiving advantages as a Hadoop distribution, with features such as volumes, multitenancy, data placement control, transparent compression, consistent snapshots, and block-level intercluster mirroring. Top use cases include archiving for disaster recovery, capacity management, machine learning, and regulatory compliance. MapR has reference architectures with Cisco, HP Inc., and IBM, and MapR appliances are available via several resellers. Some of the large customers using MapR for archiving include American Express, Cisco, Complete Genomics, comScore, Experian, MediaHub, Novartis, Pontis, and Solutionary. One of the largest deployments is a large financial services firm that has more than 20 petabytes of archived data in the MapR Distribution.
- **SAP's big data archiving strategy integrates with Hadoop and Hana.** SAP has been offering archiving solutions for more than two decades and has been working closely with Hadoop partners for more than two years. With the SAP Information Lifecycle Management (ILM) solution, users can segregate data in accordance with legal/country-specific requirements and store the data compliantly. eDiscovery functionality provides the ability to find, isolate, and preserve data at the hardware level while compliantly destroying all other data that has fulfilled its long-term legal requirement. Customers also receive a full audit trail for the process to ensure they can document defensible deletion. Some of the more recent innovations that SAP has delivered to the market in the context of managing big data are Data Warehousing Foundation and Dynamic Tiering. SAP has a variety of seamless integration methods for a variety of its products to a Hadoop framework, e.g., SAP Hana's smart data access (SDA) capability, which allows the consumption of data stored in Hadoop while circumventing the need to move or persist the data in another location. The SAP archive platform integrates with SAP Hana to support low-latency access to sensitive archived data, delivering useful insights and compliance requirements. SAP has customers in various vertical industries, including banks, retail, oil and gas, and utilities. Some of the largest deployments are in the petabyte range.

- **Solix leverages Apache Hadoop for Enterprise Archiving.** In 2014, Solix launched Solix Big Data Suite, which provides a solution for enterprise application archiving, data lake, and analytics on Hadoop. Currently, Solix supports Cloudera and HortonWorks, with plans for other distributions in the near future. Solix Enterprise Archiving provides the ability to archive and retire structured, unstructured, streaming, and email data within a Hadoop repository. The solution enforces ILM policies to ensure that it meets governance, risk, and compliance objectives. Solix Data Lake and Solix Enterprise Archiving applications utilize best-practice ILM processes to ingest and store both structured and unstructured enterprise data. Data retention is based on policies and business rules to ensure proper compliance and control. The solution maintains universal data access that provides reporting and full-text searching capabilities.
- **Teradata RainStor leverages analytical solutions to support big data archiving.** Companies use Teradata's big data archiving to support regulatory compliance, analytical archiving with Hadoop, and data warehouse augmentation. Its key strengths lie in deep compression, built-in auditing and security features, and standards-based access to archive repositories such as BI Tools, Hive, Map Reduce, and SQL. Teradata RainStor archive applications run natively on both Cloudera CDH and Hortonworks HDP Hadoop distributions. Teradata's portfolio also includes combined hardware and software big data archiving solutions, including the Teradata Appliance for Hadoop platform and the Teradata RainStor archive application. One customer has a deployment of 3 petabytes with Teradata, and many others run hundreds of terabytes of archived data.
- **ZL Technologies delivers a credible big data archive platform.** ZL Technologies, founded in 1999, offers an archiving solution to support various enterprise needs across a range of applications, including emails, files, instant messaging, and structured data. The ZL Unified Archive addresses eDiscovery, compliance, enterprise search, records management, storage optimization, and analytics. With its latest enhancements to version 8.0.1, ZL has introduced ZL NoSQL DB — a new class of NoSQL analytical persistence engine to handle billions of rows and thousands of columns of data. The latest ZL platform is Hadoop-compatible across various distributions. New enhancements allow data movement in and out of Hadoop-based data stores and the ability to leverage HDFS. ZL UA also provides open data access where data in ZL UA can be accessed with ZL Data API or data can be exported to Hadoop NoSQL stores like Apache Cassandra, Apache Hbase, and MongoDB. ZL's key differentiators include unified architecture that focuses on consolidating all applications and documents under one platform to support search, archive, and control point for retention policies; analytics across human, business, and machine data; and unified applications to support information governance. ZL offers its big data archive solutions via software or hosted services but not yet in an appliance format. ZL is cloud-ready to support Amazon AWS or Rackspace providers. One of the largest deployments of ZL has more than 200,000 mailboxes/user accounts, with a document count exceeding 12 billion.

What It Means

Rethink Your Data Archive Strategy With A Big Data Approach

Many companies don't want to archive data and do it only because of compliance requirements. This is why the adoption of application- and database-level archiving still remains low, as most organizations don't find the ROI or budget to support such a strategy. Enter Hadoop and NoSQL — and that changes everything. They can help you with customer analytics, fraud detection, customer risk assessment, or supporting a 360-degree view of your customer, business, or product. Big data archiving based on Hadoop and NoSQL technologies allow organizations to think differently and create new, intelligence-based business opportunities that weren't possible before. The bottom line? Enterprise architects with investments in big data archiving will be ready to use data in new and creative ways to support new customer insights, respond quickly to changing market demands and competitive threats, and deliver new and innovative products and services.

Endnotes

[1] For more on vendors that support the shift for archiving use cases from compliance to become sources of insight and corporate memory, see the "Archiving Platforms Evolve Into Sources Of Insights And Corporate Memory" Forrester report.

[2] Enterprise architecture professionals with requirements to improve archive access to a broad swath of enterprise content while meeting life-cycle, security, and discovery requirements should see the "Market Overview: Information Archiving, Q2 2015" Forrester report.

[3] Traditional perimeter-based approaches to security are insufficient. Security and risk professionals must take a data-centric approach that ensures security travels with the data regardless of user population, location, or even hosting model. To set a vision for data security, see the "The Future Of Data Security And Privacy: Growth And Competitive Differentiation" Forrester report.

[4] To address the gap between data and business actions, digital startups and advanced firms alike must take a new approach that Forrester calls "systems of insight." Specifically, this is a business discipline and technology approach that harnesses digital insights and consistently turns data into action. Key to this approach are multidisciplinary teams and an insights-to-execution process to embed insights in software, digital experiences platforms, and everyday work. See the "Digital Insights Are The New Currency Of Business" Forrester report.

FORRESTER®
CHALLENGE THINKING. LEAD CHANGE.

We work with business and technology leaders to develop customer-obsessed strategies that drive growth.

PRODUCTS AND SERVICES

- Core research and tools
- Data and analytics
- Peer collaboration
- Analyst engagement
- Consulting
- Events

Forrester's research and insights are tailored to your role and critical business initiatives.

ROLES WE SERVE

Marketing & Strategy Professionals

CMO
B2B Marketing
B2C Marketing
Customer Experience
Customer Insights
eBusiness & Channel Strategy

Technology Management Professionals

CIO
Application Development & Delivery
- Enterprise Architecture

Infrastructure & Operations
Security & Risk
Sourcing & Vendor Management

Technology Industry Professionals

Analyst Relations

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.