SOLIX®
Empowering Data Management

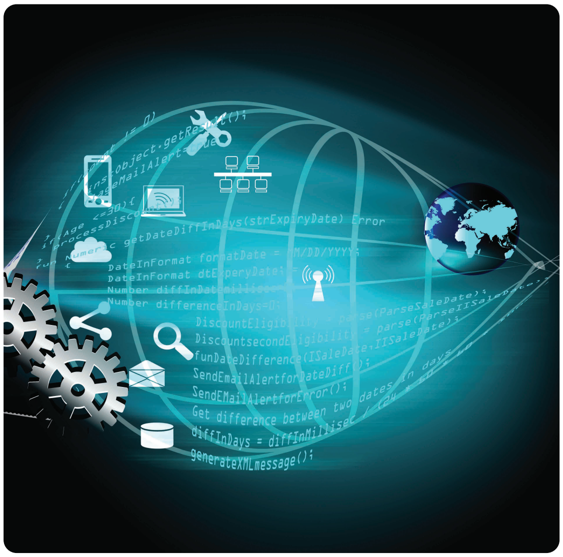
THE CIO GUIDE TO BIG DATA ARCHIVING

How to pick the right product?

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.


DateInFormat formatDate = MM/DD/YYYY
SendEmailAlertforDateDiff();
SendEMailAlertforError();
Get difference between two dates
generateXMLmessage();

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.

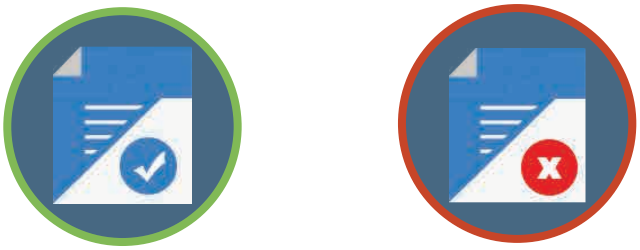
3 Can the product MOVE data into the archive versus COPY?

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.

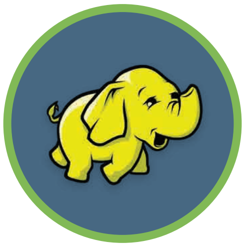
8 Does the product support print stream archiving?

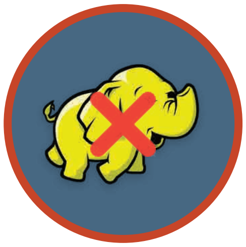
Note: *If available at all, most companies use isolated infrastructure for print stream capture, with no searching and reporting capabilities available. Captured data is maintained in a silo without any integration with other applications.*

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?

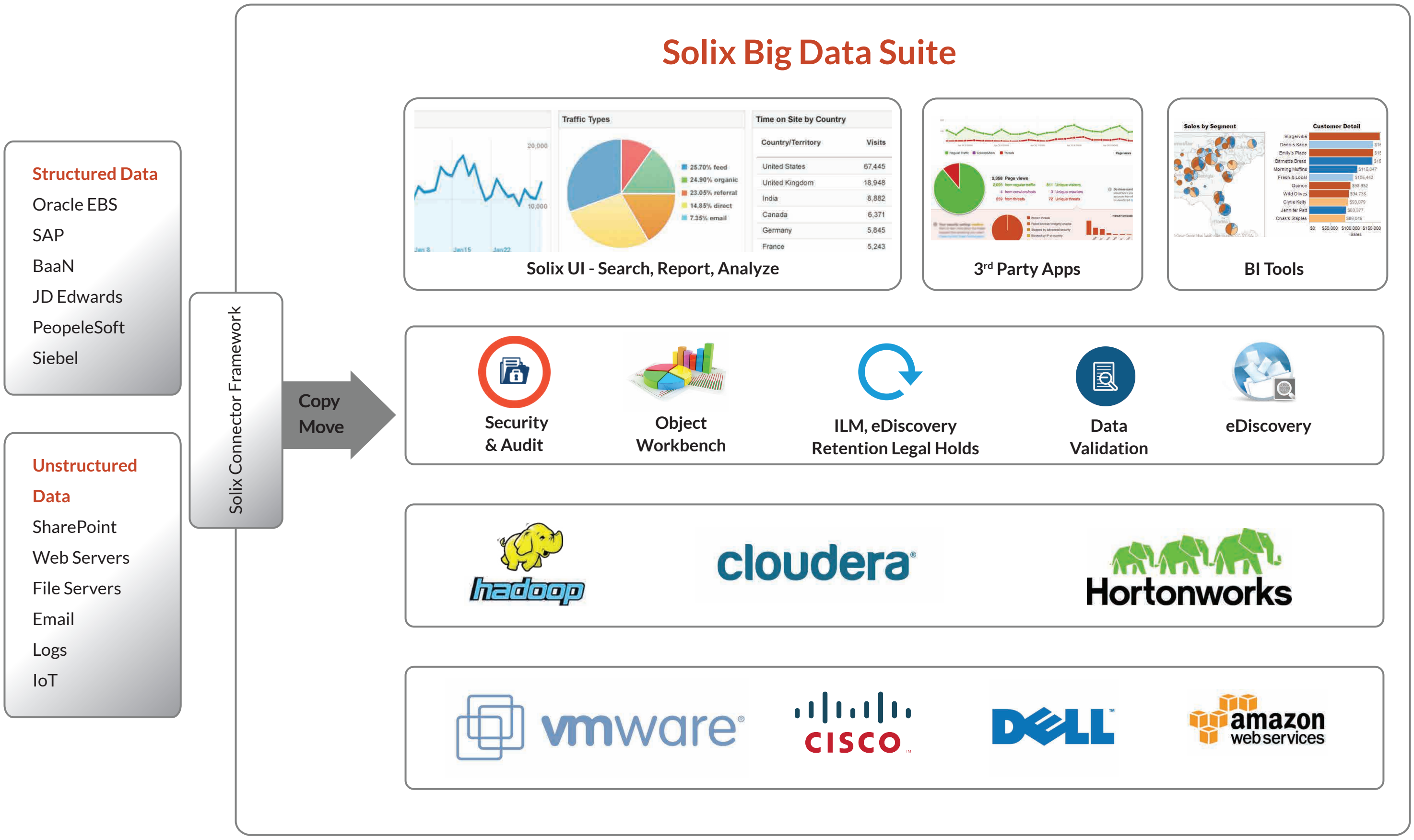
Note: *Most archiving tools do not support de-archiving operations. Can source data be manually moved back into the source database?*

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.

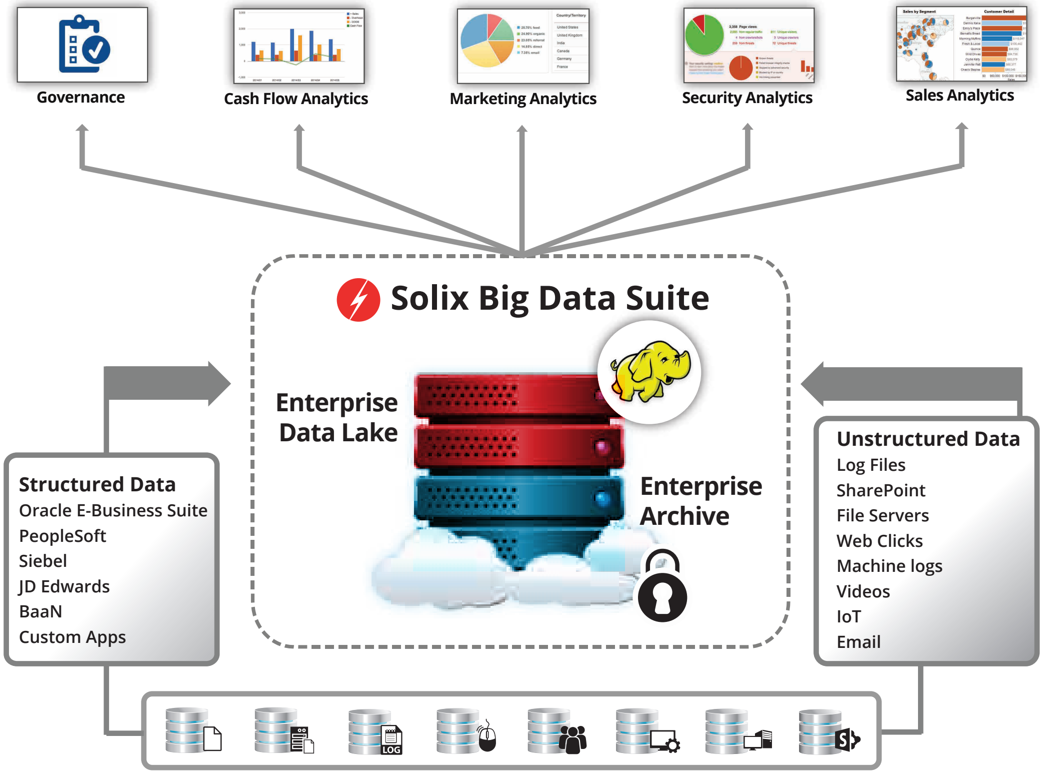
3. Move and Copy Operations	Solix can perform atomic MOVE or COPY of data from source applications.
4. Data Validation	Solix provides full validation for structured and unstructured data operations using algorithms like SHA1, MDS, etc.
5. Integrated ILM & Governance	Solix provides Information Lifecycle 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.

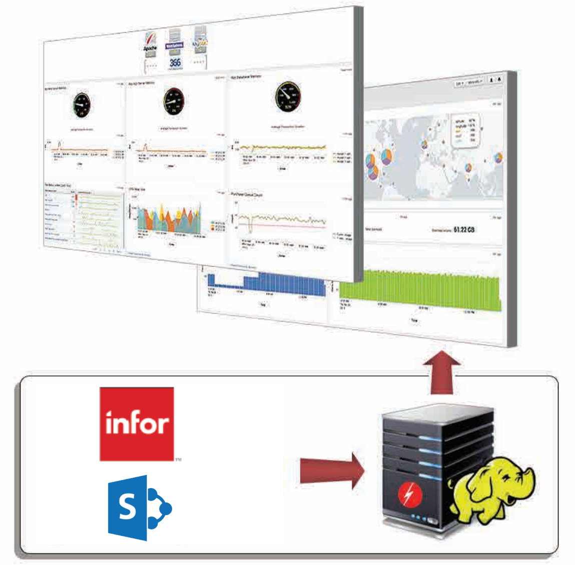
THE SOLIX BIG DATA SUITE INCLUDES:

- 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

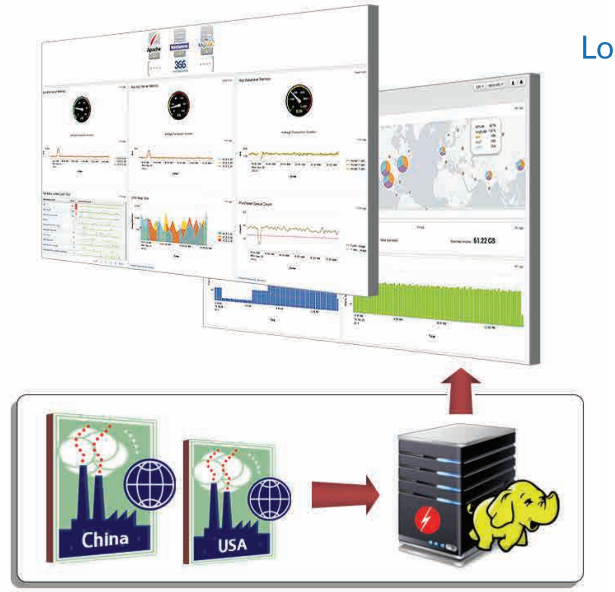
Consolidated Enterprise Archive for Manufacturing Application and SharePoint



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

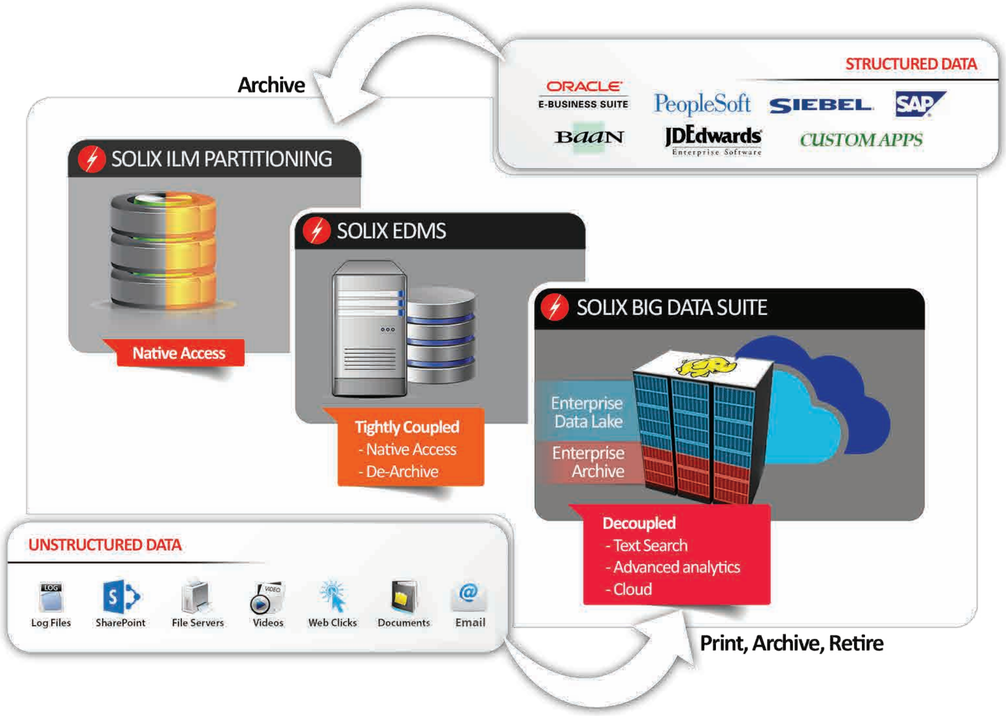
CONCLUSION

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®
Empowering Data Management

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

This document is provided for information purposes only and the contents hereof are subject to change without notice.

This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchant- ability or fitness for a particular purpose.

We specially disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.