SOLIX®
Empowering Data Management

# Enterprise Archiving with Apache Hadoop

# Executive Summary

Every CIO wants to know if their infrastructure will handle it when data growth reaches 40 zettabytes by 2020.

When data sets become too large, application performance slows and infrastructure struggles to keep up. Data growth drives increased cost and complexity everywhere, including power consumption, data center space, performance and availability.

System availability is impacted as batch processes are no longer able to meet scheduled completion times. The "outage windows" necessary to convert data during ERP upgrade cycles may extend from hours to days.

Other critical processes such as replication and disaster recovery are impacted because more data is just harder to MOVE and COPY.

Left unchecked, data growth may also create governance, risk, and compliance challenges. HIPAA, PCI DSS, FISMA, and SAS 70 mandates all require that organizations establish compliance frameworks for data security and compliance.

**According to Gartner, data growth is the No. 1 infrastructure challenge for data centers.**

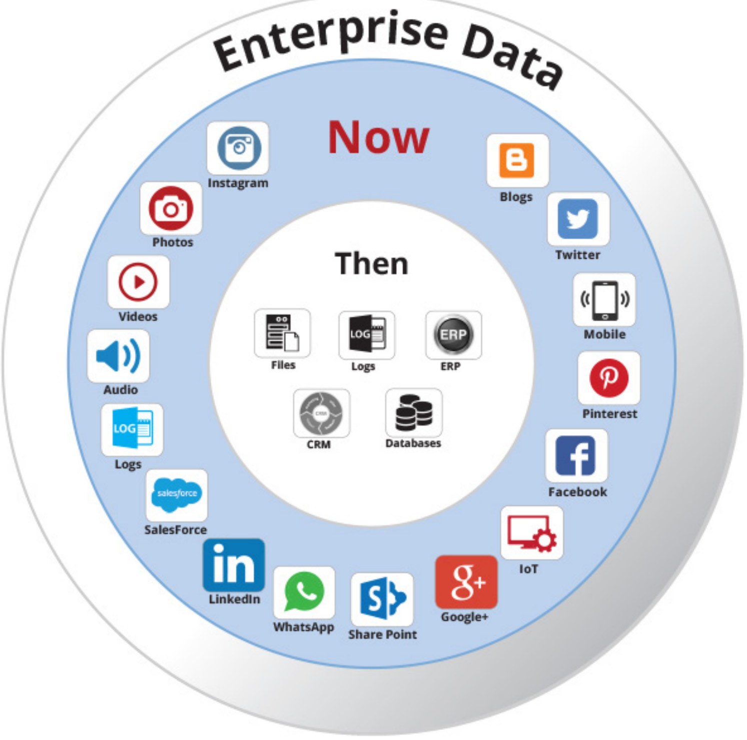
*We know the value of data declines with age because it becomes less active. Enterprise data must be managed so inactive data doesn't clog the infrastructure and impact critical processing.*

On the other hand, enterprise data is no longer confined to enterprise data centers.

Business critical data grows outside of the firewalls — with social media sites, blogs, hosted CRMs, etc. Enterprises must manage these data sources in order to stay relevant in the competitive world.

The tension between wanting more data to drive the organization successfully into the future and the need to keep infrastructure running efficiently and cost effectively has never been greater. How do organizations harness all the necessary and complex strains of data without over burdening infrastructure and personnel?

The solution is Information Lifecycle Management (ILM), which is the data management best practice to manage the lifecycle of data from creation to deletion and disposal.



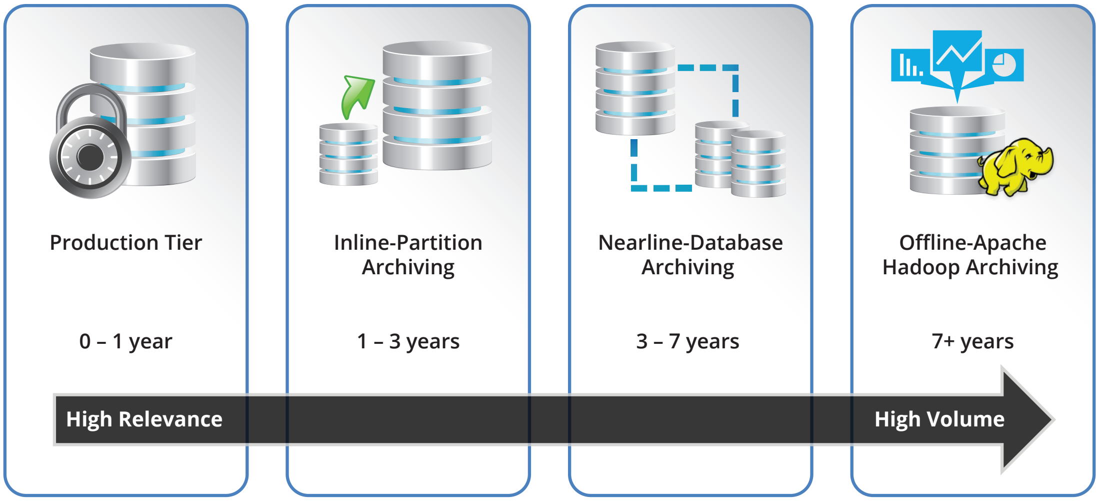
**The goals of ILM are:**

- **Optimize application performance,**
- **Manage data security, risk, compliance,**
- **Reduce infrastructure costs.**

ILM achieves these goals by assigning retention policies to data based on business rules. Solix ILM moves data to the most appropriate infrastructure tier based on retention policies such as the age of the data. Since older data is less frequently accessed; it is therefore less valuable and less deserving of limited tier one performance and capacity.

Enterprise applications such as ERP, CRM, and HCM represent an excellent opportunity for improving performance and reducing costs through application tiering with Apache Hadoop.

**FIGURE 1** Best Practice for Application Tiering


Production Tier
0 – 1 year
Inline-Partition Archiving
1 – 3 years
Nearline-Database Archiving
3 – 7 years
Offline-Apache Hadoop Archiving
7+ years
High Relevance
High Volume

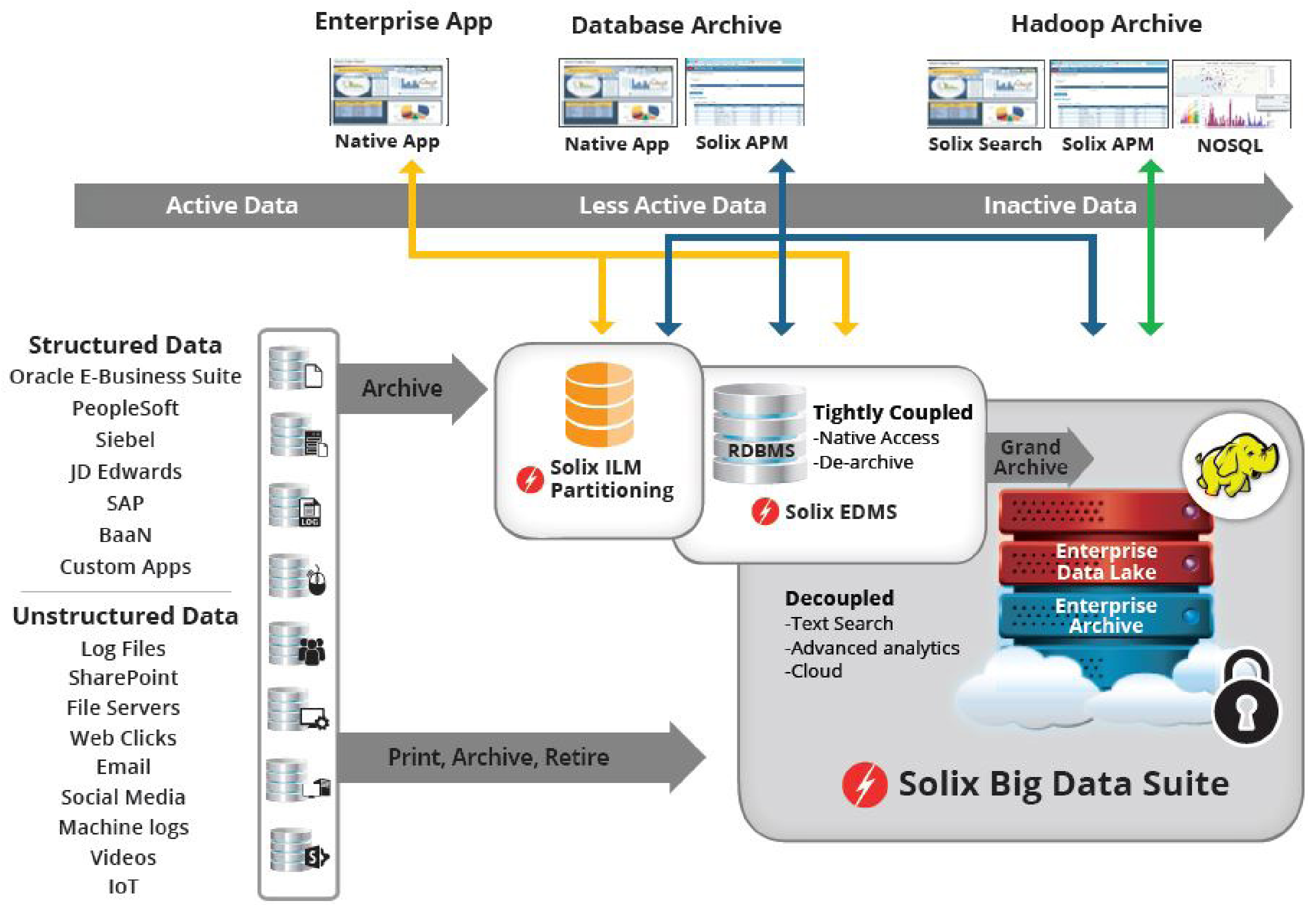
Source: Solix

**Effective use of social data is vital in delivering a high level of customer satisfaction.**

A first class passenger on an international flight tweeted about the bad food on the plane. This tweet was read by a ground crew member who relayed the feedback to the flight crew within minutes. This passenger, a high value customer with a lot of social media "followers," was pampered for the rest of the flight, prompting a glowing review of the airline to his followers.

Proactive tracking and management of social data was critical to rectifying the concerns of a high-value social influencer and improving the Net-Promoter-Score.

**FIGURE 2** Enterprise Application Tiering



Source: Solix

**Solix is the only vendor today with a solution that provides comprehensive ILM for all enterprise data — structured and unstructured.**

Source: Solix

# Best Practices for Application Archiving

The tremendous growth in volumes of data — both traditional structured data and new data types, such as Internet-of-Things (IoT) — and the advent of in-memory database technologies like SAP's HANA and NAND flash storage, which are faster but more expensive, has made data archiving mandatory. Companies simply cannot afford to operate as they once did, allowing years of data, much of it seldom used, to accumulate in single tier databases. The old data clogs systems, hurting performance, and, when that database is running on flash or in-memory, it also becomes prohibitively expensive.

For too long, organizations have debated the best way to manage the lifecycle of application data. Organizations want to implement true ILM to ensure governance, data security, and operational efficiency.

While unstructured data archiving is relatively simple as it is primarily based on age, structured data archiving is complex requiring that multiple criteria be factored into the process.

The best way to improve the management of enterprise data is to create tiers of data based on value. Our recommended ILM best practice is to leverage four processing tiers integrated with Apache Hadoop:

## Production Tier: 0 – 1 year

Highest performance infrastructure is reserved for high value, active data. Large flash arrays manage OLTP processing loads in-memory for maximum performance.

## Partition tier: 1 – 3 years

In-line ILM partitions (still running on tier one infrastructure) allow a table or index to be subdivided into ranges based on parameters such as the age of the data. Older, less valuable data may be placed in partitions to exclude them from causing processing overhead. Each partition may be assigned its own storage characteristics.

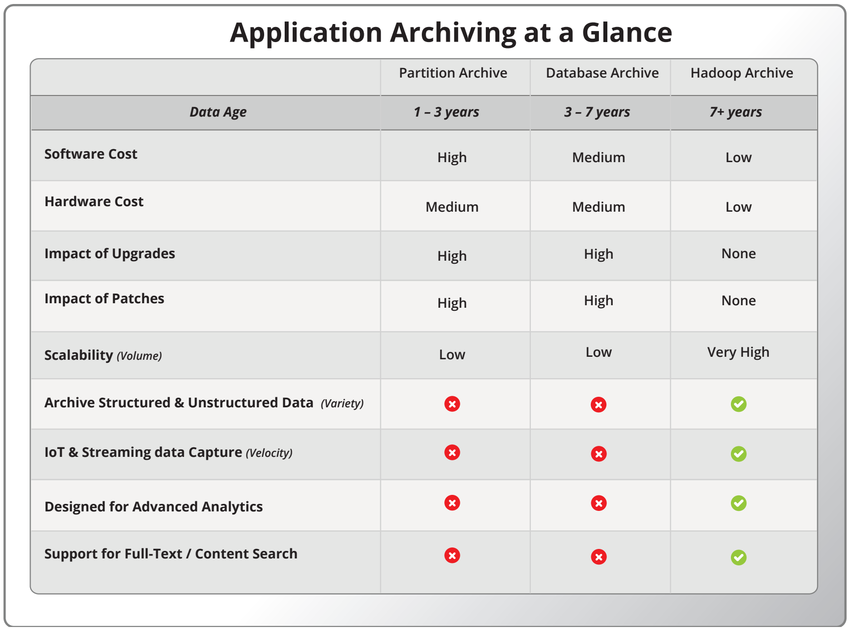
## Database archive tier: 3 – 7 years

Data which is moved and purged from the source database is called an archive. A tightly coupled archive retains native access to the application as well as the ability to de-archive back to into the source production database if necessary.

## Apache Hadoop tier 7+ years

Apache Hadoop is the ideal platform for a grand archive because it offers the lowest cost solution for bulk data storage. Hadoop provides a point-in-time snapshot of a business record. Because the data represents a complete business object decoupled from the application, data no longer must be upgraded in synch with the application. Big data analytics tools — text search as well as traditional structured query tools — provide enhanced access to the data.

Source: Solix



**Application Archiving at a Glance**

| | Partition Archive | Database Archive | Hadoop Archive |
|---|---|---|---|
| *Data Age* | *1 – 3 years* | *3 – 7 years* | *7+ years* |
| **Software Cost** | High | Medium | Low |
| **Hardware Cost** | Medium | Medium | Low |
| **Impact of Upgrades** | High | High | None |
| **Impact of Patches** | High | High | None |
| **Scalability** *(Volume)* | Low | Low | Very High |
| **Archive Structured & Unstructured Data** *(Variety)* | ✗ | ✗ | ✓ |
| **IoT & Streaming data Capture** *(Velocity)* | ✗ | ✗ | ✓ |
| **Designed for Advanced Analytics** | ✗ | ✗ | ✓ |
| **Support for Full-Text / Content Search** | ✗ | ✗ | ✓ |

Source: Solix

## Why Apache Hadoop

Apache Hadoop is a free, open source computing framework designed to operate powerful, low-cost infrastructure at a lesser tier while still delivering massive scalability and performance.

Using the MapReduce programming model to process large data sets across distributed compute nodes in parallel, Hadoop delivers highly scalable workload performance and very low-cost, bulk data storage.

All this means that Hadoop offers dramatic cost savings over traditional tier one infrastructure.

**Consider the following comparison:** According to Monash Research, the cost of tier one database infrastructure is more than $60,000 per TB.  At the same time, 1TB of S3 bucket storage at Amazon Web Services is $30 per month according to their recent price list.

**Conclusion:** Hadoop is 55.5X cheaper than tier one infrastructure.

*Recent Gartner research states that by 2017 enterprise archiving will represent 25% of the information governance efforts in enterprises. By 2016, 75% of enterprise archiving solutions will incorporate support for big data analytics.*

**The cost to store 1TB of data in Hadoop...**

... is **55.5** X cheaper than tier one infrastructure.

**$ 1,080**

The cost to store 1TB of data on Hadoop*

**VS**

**$ 60,022**

The cost of 1TB of data in tier one

Source: Solix

Source: Solix

## Solix Big Data Suite

The Solix Big Data suite provides the framework for an ILM continuum that ensures CIO's don't have to choose between application performance, operational efficiency, and cost.

> *Gartner says, "Any organization thinking of simply applying existing information governance practices to big data will likely fail — not least because much data is ungoverned; or governed by others according to a different set of objectives."*

The Solix Big Data Suite provides the first true ILM continuum that addresses the complexity of governance in the Big Data world while ensuring governance for core enterprise applications is not sacrificed.

The Solix Big Data Suite's ILM framework manages the data within HDFS and HBASE. The Solix ILM framework also provides an integrated retention-management and legal-hold capability for data within Apache Hadoop.

Structured and unstructured data from other data sources are migrated into HDFS/HBASE with full data-validation and audit reports. These reports provide the necessary defensibility and chain of custody for compliance and data governance.

This extensive ILM framework allows the Solix Big Data Suite to create a unified repository to capture all enterprise data and optimally organize it for analytics tools offered through the Solix App Store.

The suite is highly scalable, with an extensible connector framework to ingest all the enterprise data. The integrated suite allows seamless archiving, retirement, and flexible extract transform load (ETL) capabilities to improve the speed of deployment, decrease the 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 and efficient platform that creates unified and cost-effective ILM and BI infrastructures for all data, requiring smaller teams with fewer IT skills, while allowing quicker rollouts and faster results.

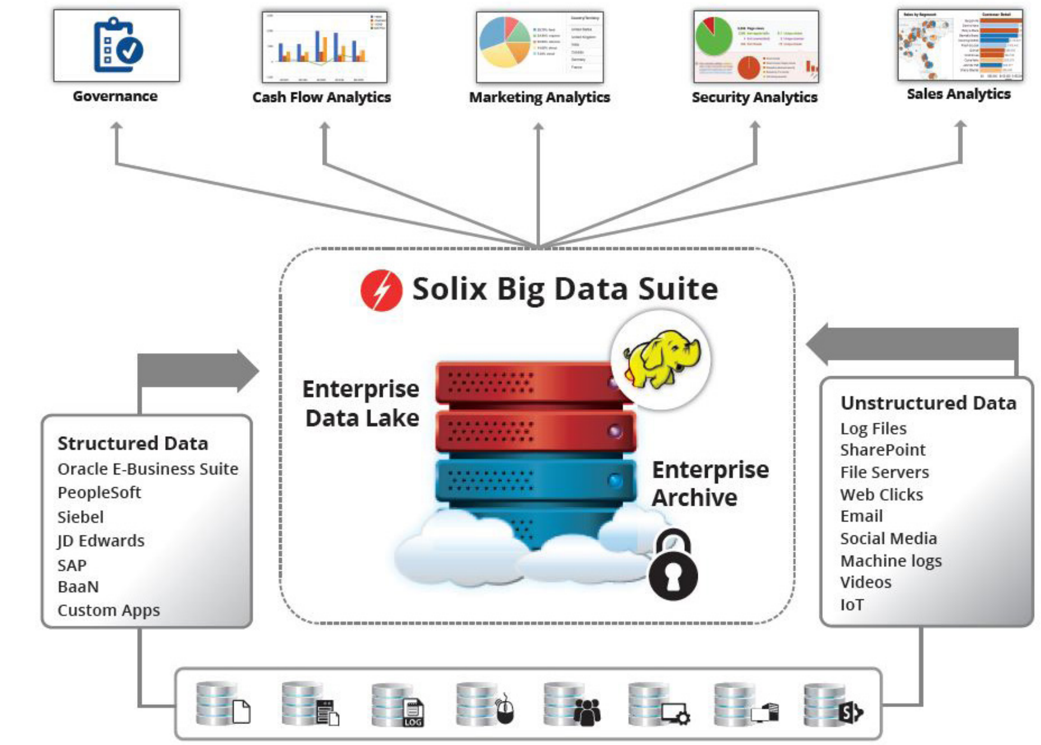
### The Solix Big Data Suite includes:

- 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, bulk 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.

**FIGURE 1** Solix Big Data Suite



Source: Solix

Source: Solix

# Conclusion

The landscape of Enterprise data is changing with the advent of Enterprise Social Data, IOT, Logs, Clicks. The reason this is called big data is because this exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your current database architectures. To gain value from this data, you need new infrastructure to manage it, and that is Apache Hadoop.

Big data technologies are being marketed to CIOs as a platform for BI and analytics. However, that is only part of the Big Data potential. With Solix Big Data Suite, CIOs can harness Apache Hadoop by using it for application archiving in addition to BI and analytics.

**Our advice to CIOs is to explore enterprise archiving on Apache Hadoop as the first step. This introduces big data technology to the enterprise, delivers immediate ROI, and can be leveraged to expand into big data analytics.**

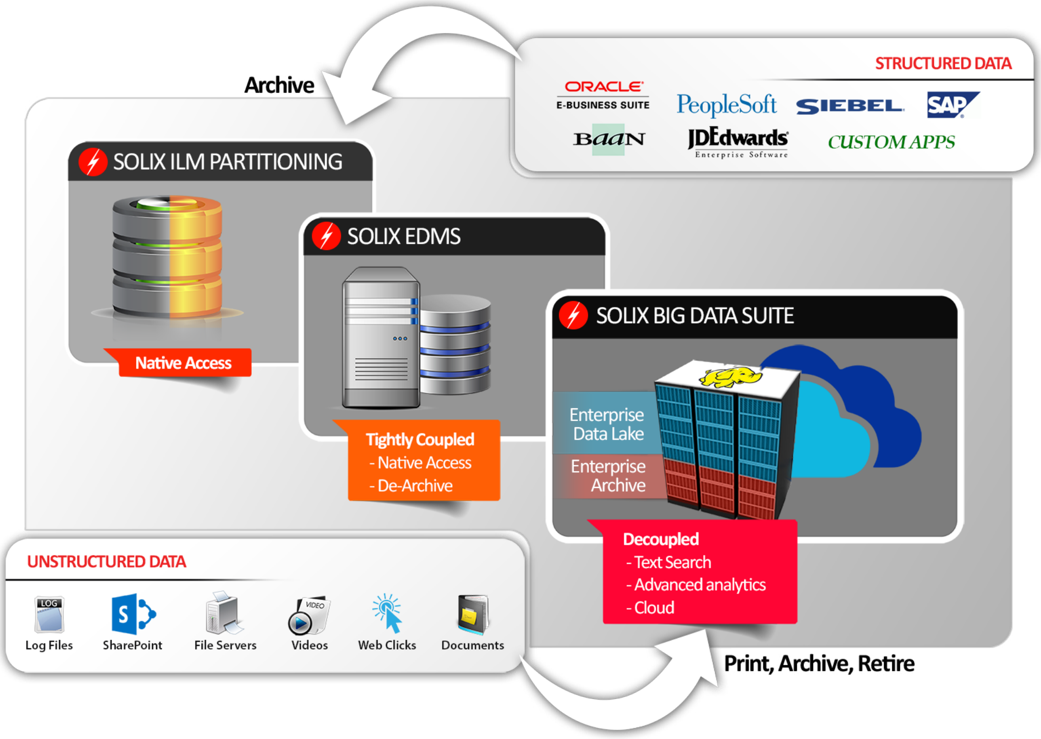
## The Benefits of
### Enterprise Archiving with Apache Hadoop

- Improve application performance
- Allows faster backups and minimizes downtime
- Eliminates infrastructure, maintenance & support costs
- Reduces operational complexity

Source: Solix

**Structured data archiving technologies help IT leaders retire legacy applications, reduce capital and operating expenses, and meet governance and compliance requirements.**

Source: Gartner, Inc. *2015 MQ for Structured Data Archiving and Application Retirement*



Source: Solix

Source: Solix

# Infographic: Solving the Data Growth Crisis with Apache Hadoop

# SOLVING THE DATA GROWTH CRISIS WITH HADOOP ENTERPRISE ARCHIVING & DATA LAKE

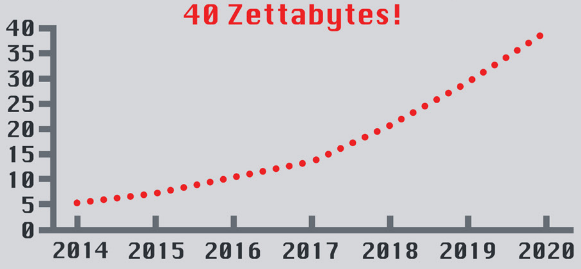
## THE CRISIS

### Did you know?

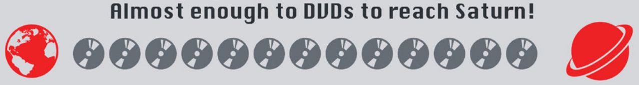
Data growth is the #1 infrastructure challenge for data centers.

### Why?

The world is experiencing so much data growth, that by 2020, the amount of generated data is expected to be **40 Zettabytes!**

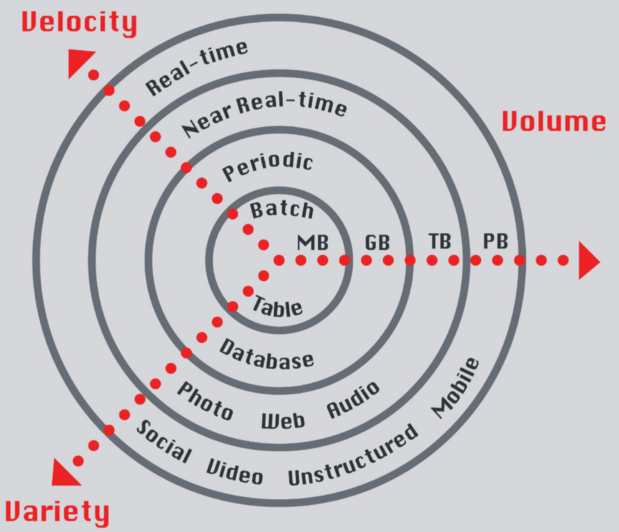


... That's roughly equivalent to **8.5 trillion DVDs.** Almost enough to DVDs to reach Saturn!



### Expansion

Big data is expanding on 3 fronts at an increasing rate.



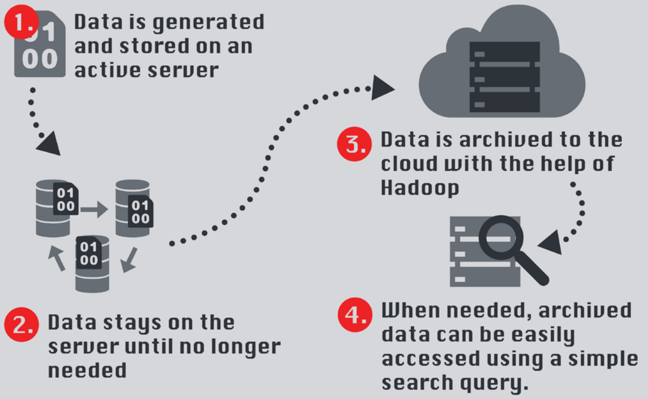
### But...

80% Up to 80% of online data is inactive.

## THE SOLUTION

### Enterprise Archiving

How does it work?



1. Data is generated and stored on an active server

2. Data stays on the server until no longer needed

3. Data is archived to the cloud with the help of Hadoop

4. When needed, archived data can be easily accessed using a simple search query.

### The benefits

Enterprise archiving...

- Improves application performance
- Allows faster backups and minimizes downtime
- Eliminates infrastructure, maintenance & support costs
- Reduces operational complexity

### The cost of archiving on Hadoop

| $1,080 | VS | $60,022 |
|---|---|---|
| The average cost to store 1TB of data on Hadoop* | | The cost of 1TB of data in a production tier |

* Based on $30/month Amazon S3 Bucket pricing in December 2014, multiplied by three years – the average lifespan of a production tier.

### The future of enterprise archiving

By 2016, 75% of enterprise archiving solutions will incorporate support for **big data analytics.**

By 2017, enterprise archiving will represent 25% of the **information governance efforts** in enterprises.

SOLIX Empowering Data Management

This Infographic was created by Solix Technologies, Inc., a leading provider of Enterprise Data Management (EDM) solutions. We help companies achieve compliance and reduce data storage costs. For more information, visit www.solix.com

Source: Solix

## About Us

Solix Technologies, Inc., the leading provider of Enterprise Data Management (EDM) solutions, is transforming information management with the first enterprise archiving and data lake application suite for big data: The Solix Big Data Suite. Solix is helping organizations learn more from their data with enterprise analytics and achieve Information Lifecycle Management (ILM) goals. The Solix Enterprise Data Management Suite (Solix EDMS) and Solix Enterprise Standard Edition (SE) enable organizations to improve application performance, meet compliance objectives and reduce the cost of data management across the enterprise. Solix Technologies, Inc. is headquartered in Santa Clara, California and operates worldwide through an established network of value added resellers (VARs) and systems integrators.

SOLIX®
Empowering Data Management