

Guide to Digital Transformation: CLOUD DATA MANAGEMENT



By John Ottman

Executive Chairman
Solix Technologies, Inc.

<https://www.solix.com>



A SOLIX WHITEPAPER
December 2020

Data-first Architecture

The goal of digital transformation remains the same as ever - to become more data-driven. We have learned how to gain a competitive advantage by capturing business events in data. Events are data snap-shots of complex activity sourced from the Web, customer systems, ERP transactions, social media, IoT, streaming, and even machine-generated data. By collecting and processing event data in real-time, managers gain situational awareness to make better decisions.

Data-driven applications enrich our understanding of business events because they leverage more data. To accomplish this, next generation apps that incorporate machine learning (ML) require schema flexibility and the ability to process very large amounts of data affordably. The goal is to raise the bar on a 'single version of the truth' and create advanced processes to improve business outcomes.

Event data improves visibility. Enterprise data warehouses that rely on canonical, top-down schemas often fail to describe business events adequately. For example, customer order transactions are sorted and analyzed, but what else do we know about these events? Was the customer referred and by whom? A Mobile app, Web, or retail customer? What else might the customer want to buy? Event data capture builds context and enables better decisions with more predictable results.

Event data capture may involve large scale data collection. Structured data from transaction systems provides only a partial picture. Today, up to 80% of enterprise data is unstructured or semi-structured and includes images, email, social media, audio and video. To establish a 'single version of the truth' for a particular business event, we must collect data about that event including structured data, files, streaming data, machine logs and raw files.

That sounds like a lot of data you may be thinking. Scalability is usually referred to in simple terms such as how many petabytes can we support. Yet simple bulk scaling to petabytes often results in massive systems that become so big they are less usable. Petabyte file stores become inefficient when you are seeking fine grain results. But when we scale logically to more discrete and specific namespaces, we can describe data better and processing can be optimized more effectively.

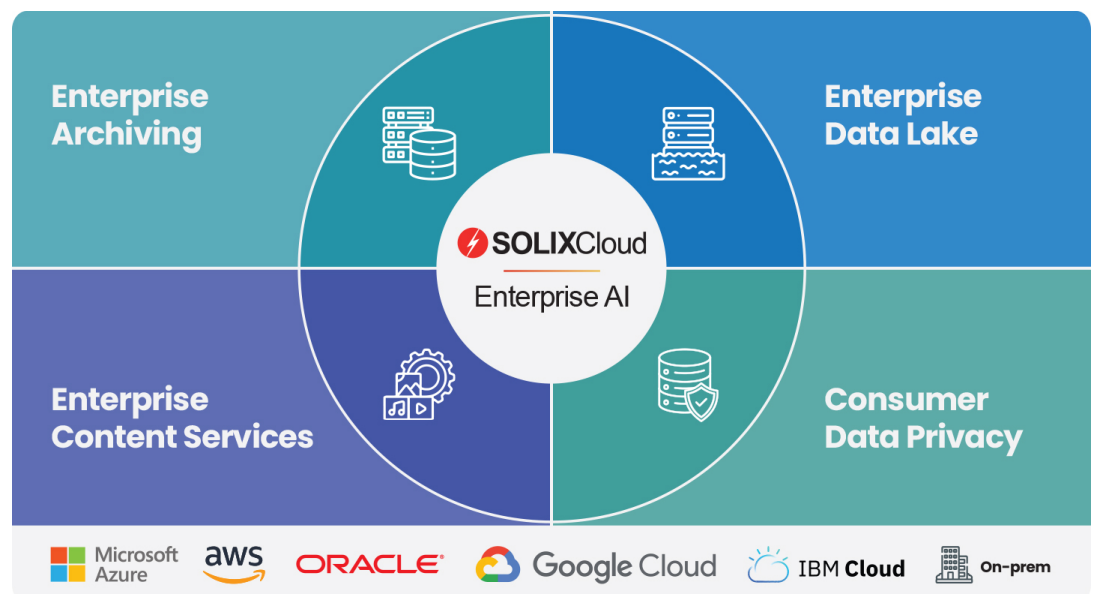
Therefore, the scalability challenge has evolved from how many petabytes can we support to how many namespaces can we manage.

With so many infrastructure requirements changing, the data-driven enterprise requires a new information architecture to achieve digital transformation. This new information architecture ingests any data, uses object storage to store bulk data at the lowest cost, and scales horizontally on clusters of commodity infrastructure. And of course, the architecture must be real-time since data loses value so fast as it ages.

Cloud Data Management

The rise of multi-cloud, data-first architecture and the broad portfolio of advanced data-driven applications that have arrived as a result require [cloud data management](#) systems to collect, manage, govern and build pipelines for enterprise data. Cloud data architectures span private, multi-cloud and hybrid cloud environments connecting with transaction systems, file servers, the Internet and multi-cloud repositories.

The scope of cloud data management includes data lakes and archives, enterprise content services and consumer data privacy solutions to manage the risk and compliance challenges of storing large amounts of data.



Cloud data platforms are the centerpiece of cloud data management programs, and they manage uniform data collection and data storage at the lowest cost.

Archives, data lakes and content services enable cloud migration projects to connect, ingest, and manage any type of data from any source including legacy systems, mainframes, ERP, CRM, file stores, relational and non-relational databases, and even SaaS environments like Salesforce or Workday which have become the new systems of record.

Data migrated to the cloud is often stored “as-is” in buckets to reduce heavy lift ETL processes. The goal is to establish real-time data pipelines to support data-driven applications. When “as-is” data will not meet application requirements, cloud services cleanse and transform raw data in preparation for future processing. Data preparation provides critical data quality measures including data profiling, data cleansing, data transformation, data enrichment and data modeling.

Data pipelines are a series of data flows where the output of one element is the input of the next one, and so on. Data lakes serve as the collection and access points in a data pipeline and are responsible for access control. As data pipelines emerge across the enterprise, enterprise data lakes become data distribution hubs with centralized controls to federate data across networks of data lakes. Data federation centralizes metadata management, data governance and compliance control while at the same time enabling decentralized data lake operations.

Cloud metadata management provides a view of the entire data landscape (including structured, semi-structured, and unstructured data) and helps users understand their data better. Analysts classify, profile and establish consistent descriptions and business context for the data.

Centralized metadata management enables users to explore their data landscape in three ways:

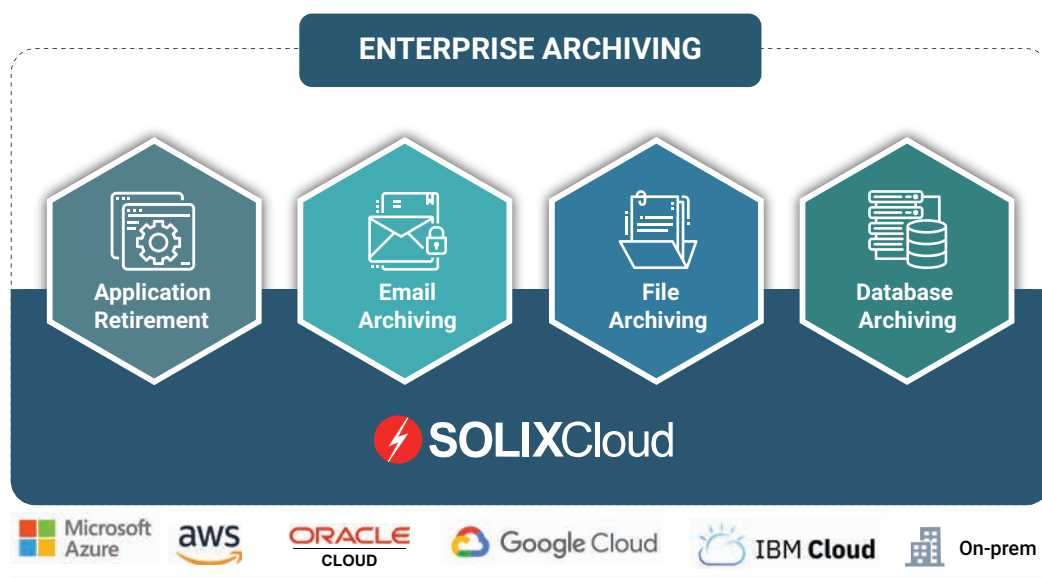
- Data lineage helps users understand the data lifecycle including a history of data movement and transformation. Data lineage simplifies root cause analysis by tracing data errors and improves confidence for processing by downstream systems.
- Data catalog is a portfolio view of data inventory and data assets. Users browse the data that they need and are able to evaluate data for intended uses.

- Business Glossary is a list of business terms with their definitions. Data governance programs require that business concepts for an organization be defined and used consistently.

Cloud data management also provides consumer data privacy and data governance controls that are essential to reduce the risks involved in handling bulk data. Information Lifecycle Management (ILM) manages data throughout its lifecycle and establishes a system of controls and business rules including data retention policies and legal holds. Security and privacy tools like data classification, data masking and sensitive data discovery help achieve compliance with data governance policies such as NIST 800-53, PCI, HIPAA, and GDPR. Consumer data privacy and data governance are not only essential for legal compliance, they improve data quality as well.

Enterprise Archiving

Studies have shown that data is accessed less frequently as it ages. Current data such as online data is accessed most frequently, but after two years, most enterprise data is hardly ever accessed. As data growth accelerates, the load on production infrastructure grows, and the challenge to maintain application performance increases. Portfolios should be screened regularly for applications that are no longer in use and those applications should be retired and decommissioned. In addition historical data should be archived from production systems to improve performance, optimize infrastructure, manage data compliance and reduce overall costs.



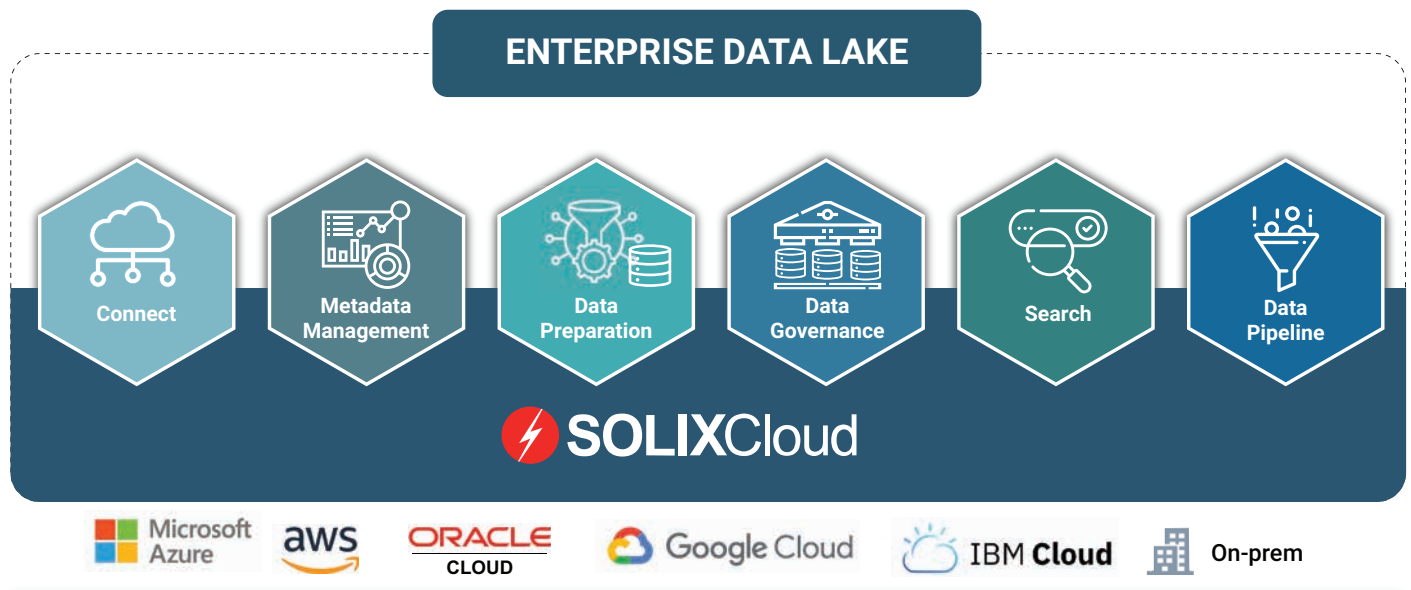
[Enterprise archiving](#) supports all enterprise data including databases, streaming data, file servers and email. Using ILM, enterprise archiving moves less frequently accessed data to nearline, real-time repositories. The archive data remains highly accessible in low cost buckets, and in some cases is even more accessible. With the data set now dramatically reduced, online application performance is improved, batch jobs complete faster and production data becomes more available. When database table sizes are reduced, production systems run faster because there is less data to handle, and the cost of special purpose infrastructure such as in-memory database processing is minimized.

And of course, the data growth challenge is even bigger with unstructured data. Overloaded file servers and email servers suffer from poor performance that frustrates end users and compliance controls become difficult to manage placing the entire organization at risk. Large organizations often operate silos of file servers across departments and divisions. Enterprise archiving consolidates these silos into a unified and compliant cloud repository.

Enterprise Data Lake

Data-driven applications leverage vast and complex networks of data and services, and [enterprise data lakes](#) deliver the connections necessary to move data from any source to any target location. Because they handle very large volumes of data and scale horizontally using commodity cloud infrastructure, enterprise data lakes are an ideal platform for cloud data migration, enterprise archiving, Operational Data Store (ODS) and to build pipelines between transaction systems and downstream analytics, SQL data warehouse, artificial intelligence (AI), machine learning (ML) applications.

Data lakes are an open source, industry standard approach to safely and securely collect and store large amounts of data. Cloud data migration projects utilize enterprise data lakes to not only collect and store data, but also to provide enterprise grade services to explore, manage, govern, prepare and provide access control to the data. Managers seeking data-driven advantage deploy enterprise data lakes to improve customer engagement or provide improved analytics based on more complete, event-driven data.



By supporting any and all file types including MS Exchange, Office 365, MS Office files, SMS, IoT, and machine logs, enterprise archiving centralizes metadata management, data governance, e-discovery, legal hold, and essential compliance controls across all historical enterprise data.

Enterprise Content Services

Corporate file shares are overflowing with files long ago abandoned. Many peg the amount of ROT (redundant, obsolete, trivial) files to be at least 80% of the total under management.

A good number of these were obsolete days after being created, over 95% within 90 days. Finding specific files or files containing certain information becomes complex when data stores are fragmented. [Enterprise Content Services](#) (ECS) collects and stores enterprise data that would otherwise be spread out in various islands of storage, on personal devices, file shares, Google Drive, Dropbox, or personal OneDrives.

Organizations planning cloud data migration to tackle content sprawl should consider ECS for secure and compliant file storage at the lowest cost. Cloud data migration with ECS consolidates enterprise data onto a single platform and unifies silos of file servers. New operations and digital processes utilize ECS in innovative ways to become more efficient and reduce costs.

Gartner estimated that the average cost to manage 1 TB of primary storage was \$3,351 annually. That number does not capture the cost of backups, data replication, managing compliance or other special requirements. The vast majority, 79% of managers and professionals in an EMC survey listed storage management as a major pain point.

ECS provides secure and compliant file storage and data services at very low cost and enables cloud data migration and modern digital transformation. ECS also provides data governance and security to protect data and make compliance easier to manage. As a secure and compliant cloud data storage and content services platform, ECS ensures IT is able to focus on critical initiatives and not bog down with data growth challenges managing enterprise content.

Consumer Data Privacy

Consumer data privacy regulations are growing with nearly 100 countries now adopting regulations. The California Consumer Privacy Act (CCPA) and Europe's General Data Protection Regulation (GDPR) are perhaps the best known laws, but new regulations are on the rise everywhere as security breaches, cyberattacks and unauthorized releases of personal information continue to grow unabated.

These new regulations mandate strict controls over the handing of personally identifiable information (PII), yet variations across geographies make legal compliance a complex requirement. Confusion exists amongst the myriad of regulations in terms of what data and entities are covered. So far, nearly \$200M in GDPR fines have been issued significantly raising the bar on the risk and expense of bulk data storage.

Consumer data privacy ensures that sensitive and personally identifiable data is handled properly and in compliance with data privacy rules and regulations. Solutions include end-to-end encryption, sensitive data discovery to ensure all PII is properly identified and classified, data masking to obfuscate data in non production environments and dynamic data masking to obfuscate data in production environments. Consumer data privacy reduces the risk of cloud data operations, improves data governance, data security and data quality, and ensures compliance with critical legal requirements.

Summary

Digital transformation requires interoperability with the cloud and its vast network of data and web services. Cloud data management connects, governs and manages data across multi-cloud landscapes and delivers the essential custody services for a data-first architecture. By providing end-to-end services such as data pipelines between OLTP systems and SQL data warehouses, archiving databases and mail servers, hosting data lakes and running NoSQL applications, cloud data management provides essential services for data-driven applications.

Data-first architectures require low-cost and efficient object storage, real-time access, data governance, metadata management, data preparation and connectivity with end-to-end data pipelines. With cloud data management any organization is able to implement these critical capabilities very quickly to achieve digital transformation and become a data-driven enterprise.

About Author:



John Ottman has over 30 years experience with enterprise applications and cloud infrastructure. He is currently the Executive Chairman of Solix Technologies, Inc. and Co-Founder and Chairman of Minds Inc.



Copyright ©2020, Solix Technologies and/or its affiliates. All rights reserved.

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.

Solix is a registered trademark of Solix Technologies and/or its affiliates. Other names may be trademarks of their respectively.

SOLIX TECHNOLOGIES, INC.

4701 Patrick Henry Dr., Bldg 20, Santa Clara,
CA 95054

Toll Free: +1.888.GO.SOLIX (+1.888.467.6549)

Telephone: +1.408.654.6400

Fax: +1.408.562.0048

URL: <https://www.solix.com>


CONNECT WITH US

 solix.com/blog

 twitter.com/SolixBigdata

 facebook.com/SolixTechnologies

 linkedin.com/company/Solix-Technologies

 info@Solix.com