SOLIXCloud

Guide to Digital Transformation:

ENTERPRISE DATA LAKE

By John Ottman
Executive Chairman
Solix Technologies, Inc.
https://www.solix.com

SOLIX™
Empowering the Data-driven Enterprise

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.

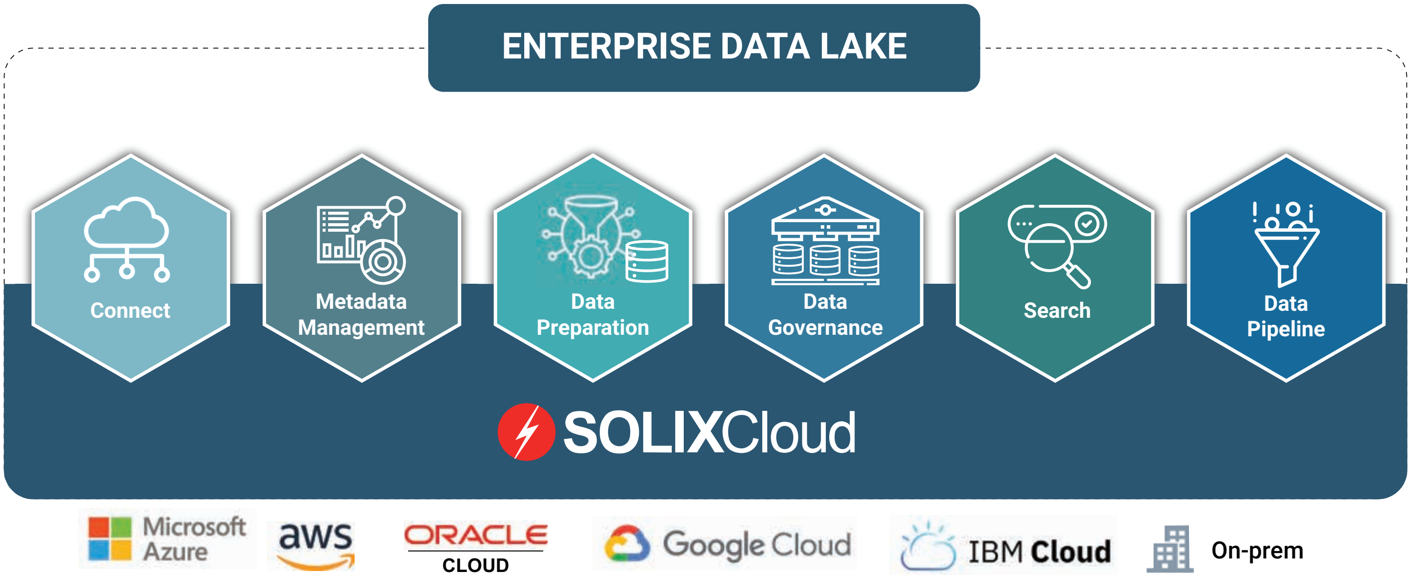
Data-driven applications are hungry for more data and event data capture may involve large scale data collection.

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, data can be described 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.

Enterprise data lakes are the centerpiece of cloud data management programs because they collect, manage, govern and prepare the data pipelines that feed data-driven applications. Enterprise data lakes 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 rapidly become the new systems of record.



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.

Data migrated to the cloud is often stored "as-is" in buckets to reduce heavy lift ETL processes. The goal is to enable real-time data-driven applications.

When “as-is” data will not meet application requirements, enterprise data lakes 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.

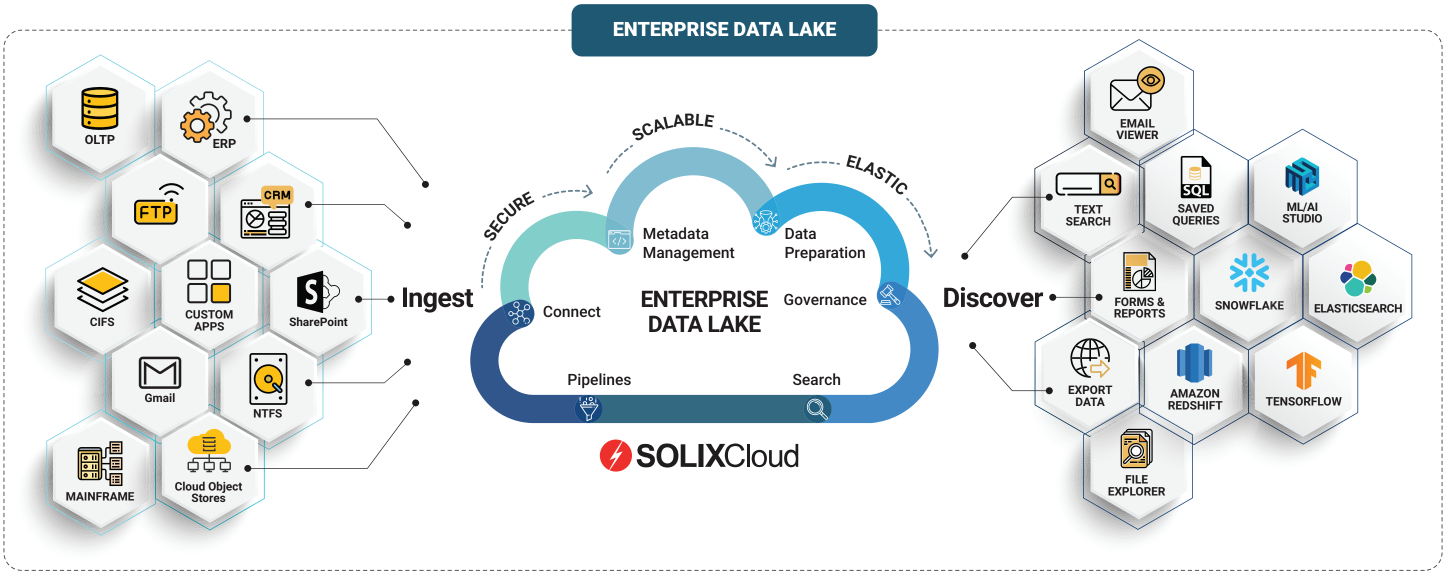
Data pipelines are a series of data flows where the output of one element is the input of the next one, and so on. Enterprise data lakes serve as the collection and access points in a data pipeline and are responsible for access control. As data pipelines emerge across the enterprise, enterprise data 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.

Enterprise data lakes need metadata management to view 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.

Enterprise data lakes also provide 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.

Cloud data migration is accelerating because it offers a safe and secure approach to low-cost bulk data storage. Online transaction processing (OLTP) systems use data lakes as Operational Data Stores (ODS) to improve performance, optimize infrastructure and simplify maintenance. An ODS is not only an active archive for historical data, it is also a real-time replicated copy of the current online database. With a replicated copy of all current and historical data, and a choice of powerful downstream SQL engines, ODS becomes an ideal solution to offload decision support processing from high cost, in-memory OLTP database servers to low cost, commodity, cloud infrastructure.

ODS also relieves upgrade pressure on production systems since customized interfaces are decoupled from the OLTP system and moved to the ODS API. Customized point-to-point interfaces often require significant maintenance in support of database and application upgrades. In some cases the level of effort is substantial. By rearchitecting external interfaces from the OLTP system to the ODS, maintenance issues are eliminated and the architecture is simplified making everything easier to manage.

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 warehouses, artificial intelligence (AI) and machine learning (ML) applications.

Digital transformation requires interoperability with the cloud and its vast network of data and web services. Data lakes are an open source, industry standard approach to safely and securely collect and store large amounts of data. Enterprise data lakes 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.

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

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.

SOLIX
Empowering the Data-driven Enterprise

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