SOLIX

Empowering the Data-driven Enterprise

Logical Data Warehouse

Solix Common Data Platform For Enterprise

In this issue

EXECUTIVE SUMMARY 2

INTRODUCTION 5

SOLVING THE PROBLEMS OF THE LDW 7

- Cost Control and Future Proofing 7
- Metadata, Security, and Compliance 9
- Data Virtualization 11
- Information Governance 12
- Support for Multiple Technologies 13

CONCLUSION 14

EXECUTIVE SUMMARY

"You cannot manage what you cannot measure."

Peter Drucker's statement has never been more true. In this era of Big Data, with multiple data sources, huge data volumes and a growing multiplicity of data types from text to video, audio, still images, log files, and on and on, captured in multiple data stores in numerous locations, the biggest problem companies face is knowing what they have and beyond that, being able to combine the data in a way that allows it to be used.

Data is already remaking industries and reshaping the global economy as startups such as Uber and AirBNB disrupt traditional industries. Experts agree this is only the start of a new era for business. Every company is becoming an IT-based organization, and data has become the rocket fuel to send corporate growth "to infinity and beyond."

Companies that embrace data are finding new business opportunities and improving earnings. Organizations that ignore the promise of data will be unable to manage the constant disruption that is the new business norm and are doomed to constantly miss opportunities and fall behind. They will find it increasingly difficult to survive in the new world economy.

Enterprises today need to ingest and analyze large volumes of data from a constantly evolving set of sources:

- These include social media, IoT, weather data, security and other internally generated audio and video, as well as public audio and video from social media and other sources, still images ranging from photos to medical imaging, blueprints and other diagrams, and freeform text. In response, enterprises are adopting Big Data technology to augment their existing Enterprise Data Warehouses (EDWs).

- In an effort to control the costs of storing petabytes of data, they are also using inexpensive storage services from Amazon Web Services (AWS), Microsoft Azure, and other cloud Infrastructure-as-a-Service (IaaS) providers.

- Businesses are also using increasing numbers of Software-as-a-Service (SaaS) offerings. That data also needs to be included in business analysis.

- IoT is adding high volumes of vital data that is often stored in remote sites, such as drilling and fracking sites.

Distributed data management platforms are required to address the diverse use cases and types of data that organizations want to consume in their analytical environments.[1]

Just capturing the large volumes of data is often a challenge. But to get benefit from the data, companies need to combine these various data repositories, as well as real-time stream data, in a logical structure that can support machine learning and new-generation AI/Cognitive Computing analysis. Gartner's term for this is the Logical Data Warehouse (LDW), which it argues should supplement the EDW. This is fundamentally a data management architecture for analytics that extends over multiple data repositories of different types in multiple physical locations – in the data center, in remote sites, and in the cloud. These data repositories often include traditional data warehouses, Hadoop-based data stores, transactional data stores, content management systems, cloud-based data, columnar, in-memory data, NoSQL databases and streaming data (both structured and unstructured). The LDW sits between these and the machine learning, business intelligence and advanced AI systems that extract the value from this data, shielding them from the underlying complexity and allowing them to be treated as a single massive data store.

In theory the LDW is the perfect answer to the challenge of managing petabytes of Big Data. However, enterprises are finding that actually creating an LDW is a complex, multifaceted technical problem demanding detailed knowledge of issues, such as how to access the different underlying technologies – many of which do not have standard interfaces – how to identify the right data to address the problem and then transform that data into something meaningful.

These are massive challenges for any IT team charged with creating an LDW from ground up. Fortunately, they do not have to. Solix provides the answers to these problems in what is essentially a fully constructed and updated LDW, the Solix Common Data Platform (CDP).

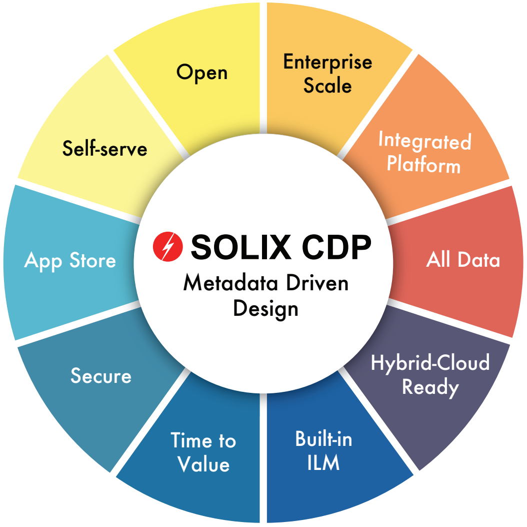
- The Solix CDP helps organizations leverage their existing technology in conjunction with low-cost mass storage in Hadoop Data Lakes, automatic archiving of older data to inexpensive lower storage tiers and deletion of data that is no longer needed.
- It allows organizations to collect, store, and analyze data from every source and store it in the most appropriate, cost-effective manner, without sacrificing governance, security, or management.
- It keeps all data in its original context and structure, allowing organizations to ask complex questions and gain deep contextual insights from the data at any point in time.

[1] Gartner Research, The Practical Logical Data Warehouse: A Strategic Plan for a Modern Data Management Solution for Analytics, March 2018

- When the needed data cannot be moved or copied – for instance when it is in transactional systems or external data sources such as weather or credit information – the Solix CDP can connect seamlessly to the data source through its data virtualization technology.
- The Solix CDP provides a la carte interfaces needed to build functional technology stacks on top of Hadoop using popular Open Source technologies, such as Spark, and to connect all the multiple data stores to the LDW.
- It fills in the inadequacies of those technologies in vital areas such as Metadata management, security and Information Lifecycle Management (ILM), including automatic archiving of older data to inexpensive lower storage tiers and deletion of data that is no longer needed.
- It creates a new paradigm, fostering a meaningful, frictionless partnership between IT departments and business leaders, with IT acting as the guardian of data, while the business users become the owners and direct consumer of that data.

Solix created the CDP to bring ILM to the Data Lake and innovation to the EDW. The Solix CDP is the next step in the evolution of the new enterprise blueprint, enhancing the LDW design concept, a.k.a. the Hybrid Data Ecosystem, by offering a secure, managed and compliant enterprise archive and compliance tools to create an advanced analytics platform with unprecedented levels of ILM in a Big Data setting.

Figure 1. Solix CDP At a Glance



Source: Solix

Solix created the CDP to bring ILM to the Data Lake and innovation to the EDW.

INTRODUCTION

Solix CDP = Enterprise Archiving + Enterprise Data Lake + Information Governance = LDW

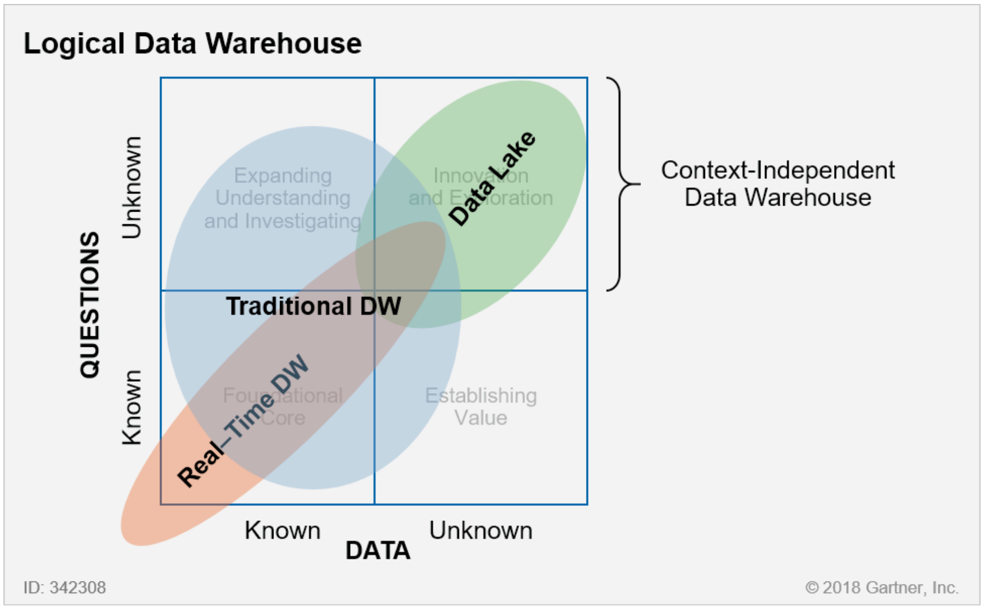
Fundamentally, the LDW is a data management architecture for analytics. It spans multiple data repositories of different types and provides associated services for management, governance and access to data. These repositories can include traditional EDWs, Hadoop-based Data Lakes, transactional system stores, content management systems, cloud-based data, columnar, in-memory stores, NoSQL and streaming data.

Solix has identified the following challenges preventing many enterprises from implementing an LDW:

The typical approach to distributed data management breaks down under the pressure of the diverse use cases that are typical in this age of disparate data types, especially in the area of analytics.

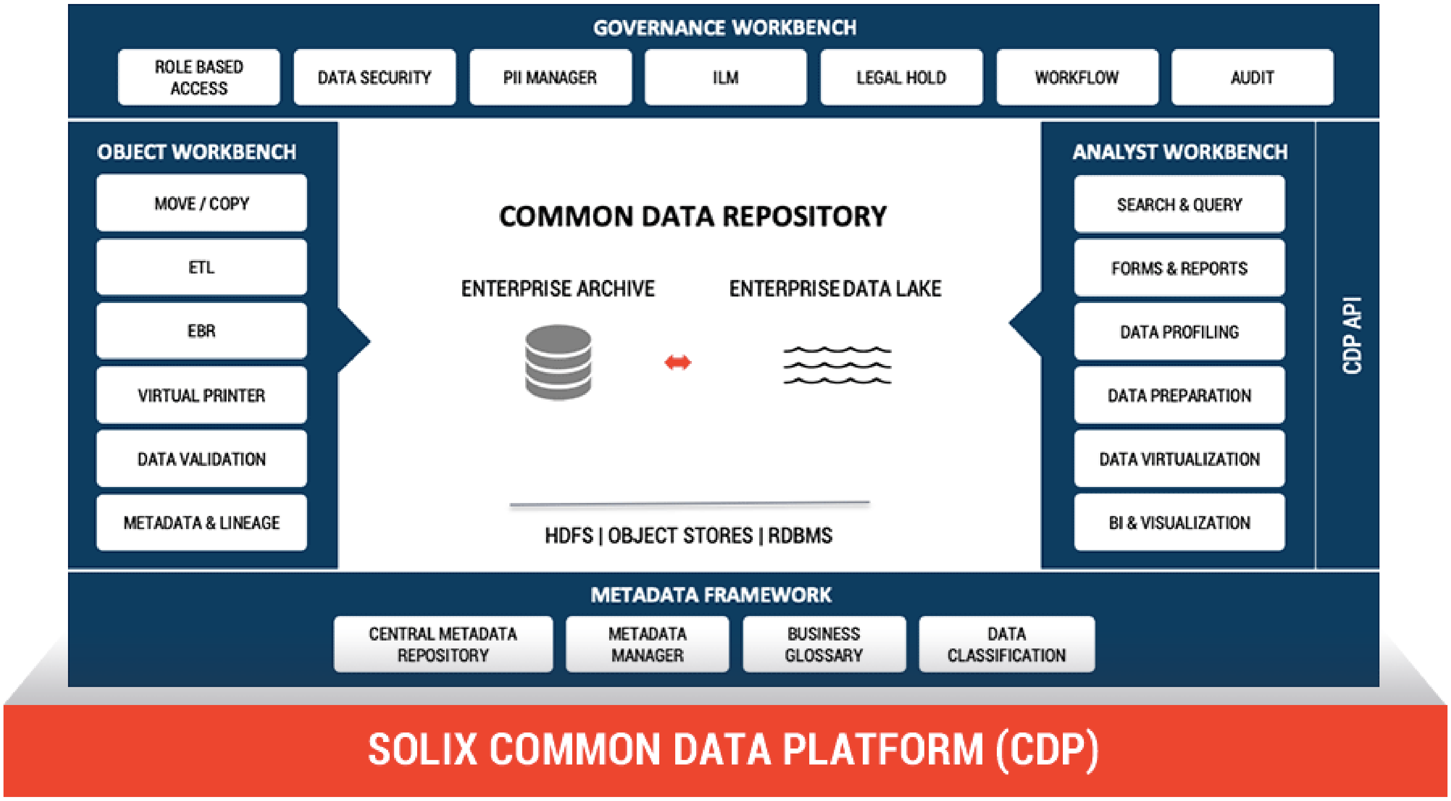
IT departments are moving away from complex infrastructure support to a more balanced and holistic view of the enterprise. This leads to challenges in implementing the complex LDW environment.

Figure 2. Logical Data Warehouse Overlay for the Data Management Infrastructure Model



DW = data warehouse
Source: Gartner (March 2018)

Figure 3. Solix Common Data Platform Architecture



Source: Solix

Where enterprises have tried to implement the LDW without taking the processes and capabilities of the technologies used to achieve their aim into consideration, they are failing due to the inability of these components to talk across the enterprise.

Metadata is the life blood of the LDW and is necessary for effective analytics. But it is difficult to track and control across disparate toolsets. Hadoop, for instance, does not maintain metadata.

The Solix CDP is a data management program designed to work across multiple, disparate data storage technologies:

- It provides all the vital management functions across multiple data stores and a robust ILM regardless of the underlying technologies.
- It brings enterprise-grade capabilities to the Hadoop framework, addressing all the issues preventing the creation of the LDW, including uniform data collection, metadata management, ILM, and secure data access for Advanced Analytics, while maximizing an organization's existing infrastructure.
- Organizations can vastly expand the reach of analytics by creating an Advanced Analytics Platform on top of the Solix CDP.
- For the CIO, the Solix CDP provides strong cost control, ensuring budgetary support from the organization, and rectifies the technical issues preventing LDW implementation.

SOLVING THE PROBLEMS OF THE LDW

The Solix CDP solves several important technical problems presented by the modern data environment, turning the LDW into a plug-and-play environment:

Cost Control and Future Proofing

Enterprises cannot afford not to get on the Big Data/ Advanced Analytics bandwagon. Doing so would sacrifice the enterprise's future. Seat-of-the-pants business decision making, with all the internal politics that too often force a company down the wrong path, simply is not good enough in the 21st Century. Companies that base business decisions on analysis of the data will consistently make better decisions and recognize opportunities sooner, avoiding wasting resources on dead-ends while realizing the benefits of new markets.

IT cannot make the mistake of the EDW again, where it spent all of its budget building the EDW and processing the data, leaving nothing for the analysis that refines that data into business insights.

Just capturing and managing the multi-petabyte-level of data strains the budget of IT shops. Hadoop promises a lower cost solution, but building and managing a Hadoop Data Lake presents its own challenges:

- The Open Source components of the Hadoop stack are incompatible with each other, with no standard interfaces.
- Thus, building and managing the stack requires new skill sets that are rare and expensive.
- Furthermore, these components are constantly evolving, with new technologies entering constantly, making future-proofing a major issue.

IT cannot make the mistake of the EDW again, where it spent all of its budget building the EDW and processing the data, leaving nothing for the analysis that refines that data into business insights.

The Solix CDP helps CIOs tame the storage cost monster in three ways:

1. The Solix CDP provides automatic data tiering and archiving.

Data archiving has emerged as an ILM best practice to meet data growth challenges and support the access requirements of the LDW while moving the 80 percent of core application production data that is inactive off expensive tier 1 storage. The Solix CDP provides automated Enterprise Archiving that will tier all the data, archiving older, little used data onto lower tiers of Hadoop infrastructure, while continuing to ensure its availability when needed.

It also provides automated deletion based on business rules, removing data the enterprise does not need automatically based on customizable business rules. This is particularly important with some forms of Big Data, such as steady-state IoT data that has virtually no value once it has been analyzed for anomalies. It also can delete data maintained for compliance, once the data ages past the legal retention requirements, both freeing storage space and protecting the organization from legal jeopardy.

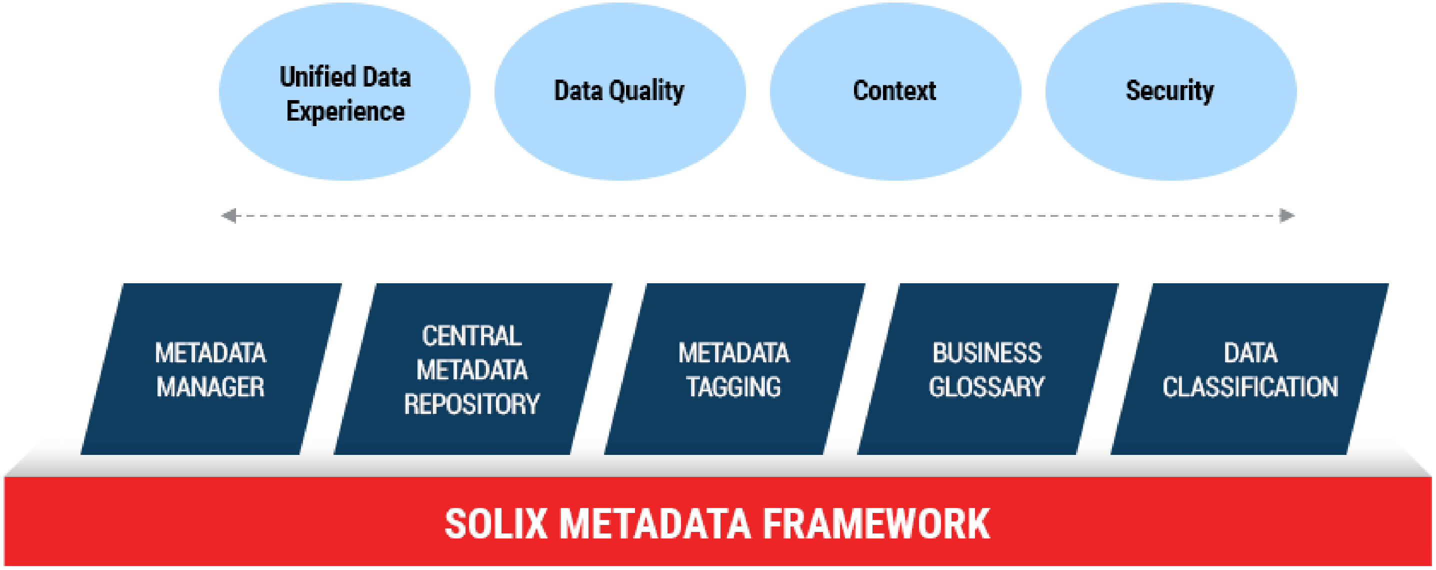
2. The Solix CDP includes all the non-standard interfaces required to build and maintain the Hadoop stack and access data on IaaS and popular SaaS platforms.

As new technologies emerge, the Solix technical team will add any new interfaces and other technology required to connect these to the stack. This turns building Hadoop Data Lake from a huge technical challenge into a plug-and-play exercise, while future-proofing the stack. The Solix CDP also includes all interfaces and other technologies required to provide full access to the different storage technologies in the cloud, such as the custom interfaces needed to access cheap storage on Amazon, again making creating an LDW a plug-and-play exercise on a higher level.

3. Finally, with no need to reboot the organization's enterprise architecture, the Solix CDP harnesses the current architecture to develop a new enterprise blueprint, establishing an LDW capable of evolving with the business requirements of the organization.

It also interconnects all the underlying databases, allowing automatic transfer of data objects from one to another, for instance from the expensive EDW to the less expensive Hadoop Data Lake.

Figure 4: Solix Metadata Framework



Source: Solix

Metadata, Security, and Compliance

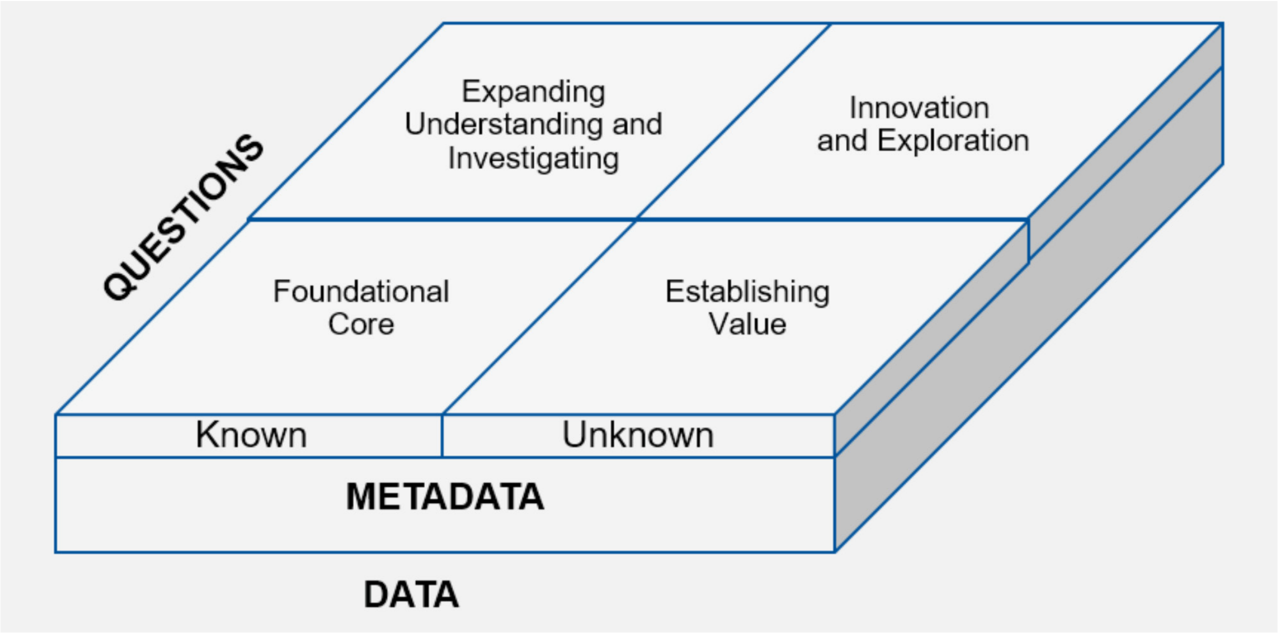
Metadata is essential for BI, AI, and Machine Learning; data management; security and compliance. Without metadata, users cannot know how old the data is or what context surrounded it when it was created. IT and the legal team cannot know who has accessed it and for what uses, vital for compliance audits and data security. No one will know how it may have been altered, again vital for data security and ensuring the data has not been sabotaged.

But maintaining metadata across multiple data types and database engine technologies is difficult at best. Hadoop, which was not designed for long-term data retention, does not manage metadata at all, meaning all data that goes into a Hadoop Data Lake loses its metadata. Without metadata, the Data Lake quickly becomes a data swamp, rending the entire system useless.

The Solix platform has, since its inception, been built on a strong and complete metadata framework (see Figure 4). This is required to govern, model and report on data successfully. Allowing data users to add custom metadata to their ingested data supports a far richer business glossary, which in turn enables a better contextual understanding of the data. Supporting automatic metadata ingestion from multiple sources enables full lineage of the data to manage compliance, validation, audit and granular security on a single pane of glass supporting all data across the LDW, whether it is physical, logical, or virtual. The Solix CDP also provides the tools and platform for comprehensive data security across all the underlying databases.

The data management infrastructure model framework may appear to overlay use cases and uses of data in a static way, but data that was unknown one day becomes known the next day, once savvy users discover how to use it meaningfully. Allowing users to 'tag' the data as they understand more about its context allows metadata derived from crowdsourcing and may increase the quality and contextual understanding of data over time. Data that is known can be used by anyone, regardless of the type of question being asked (see Figure 5). Finally, use cases may have changing physical service levels. An algorithm developed on data at rest by data scientists may need to be run in production using real-time data. Data can also flow in the other direction. For example, aging, cooler data that is being accrued for long-term trend analysis or audit purposes may be stored more economically in the Data Lake. The Logical Data Warehouse requires the ability to adapt all of these changes as they occur over time. This involves metadata.[2]

Figure 5. Metadata is Foundational for the LDW



LDW = logical data warehouse
Source: Gartner (March 2018)

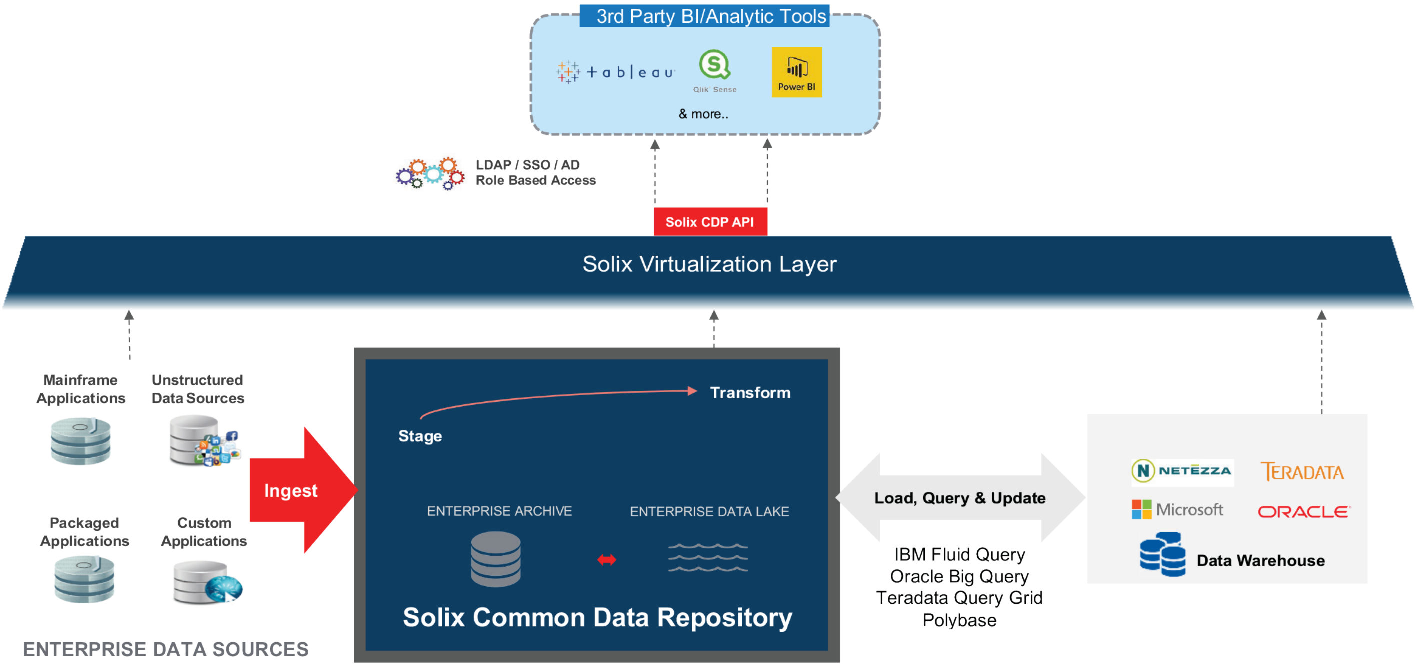
[2] Gartner Research, The Practical Logical Data Warehouse: A Strategic Plan for a Modern Data Management Solution for Analytics, March 2018

Data Virtualization

Data Virtualization is a critical part of the LDW architecture, enabling querying of data from across multiple data sources without the need to move/copy data. It can work with both traditional structured data sources, such as databases, data warehouses, etc., and less traditional data stores, such as Hadoop, NoSQL, Web Services, SaaS applications and so on, while still appearing as a single ”logical” data source to the user.

The Solix CDP’s data virtualization technology relies on its central metadata repository to provide the ability to understand the data and query it while shielding the user from the underlying complexities of querying from source systems – different technologies, formats, locations, protocols, etc., to provide a common consistent method to access data.

Figure 6: Solix LDW Architecture including Virtualization Layer



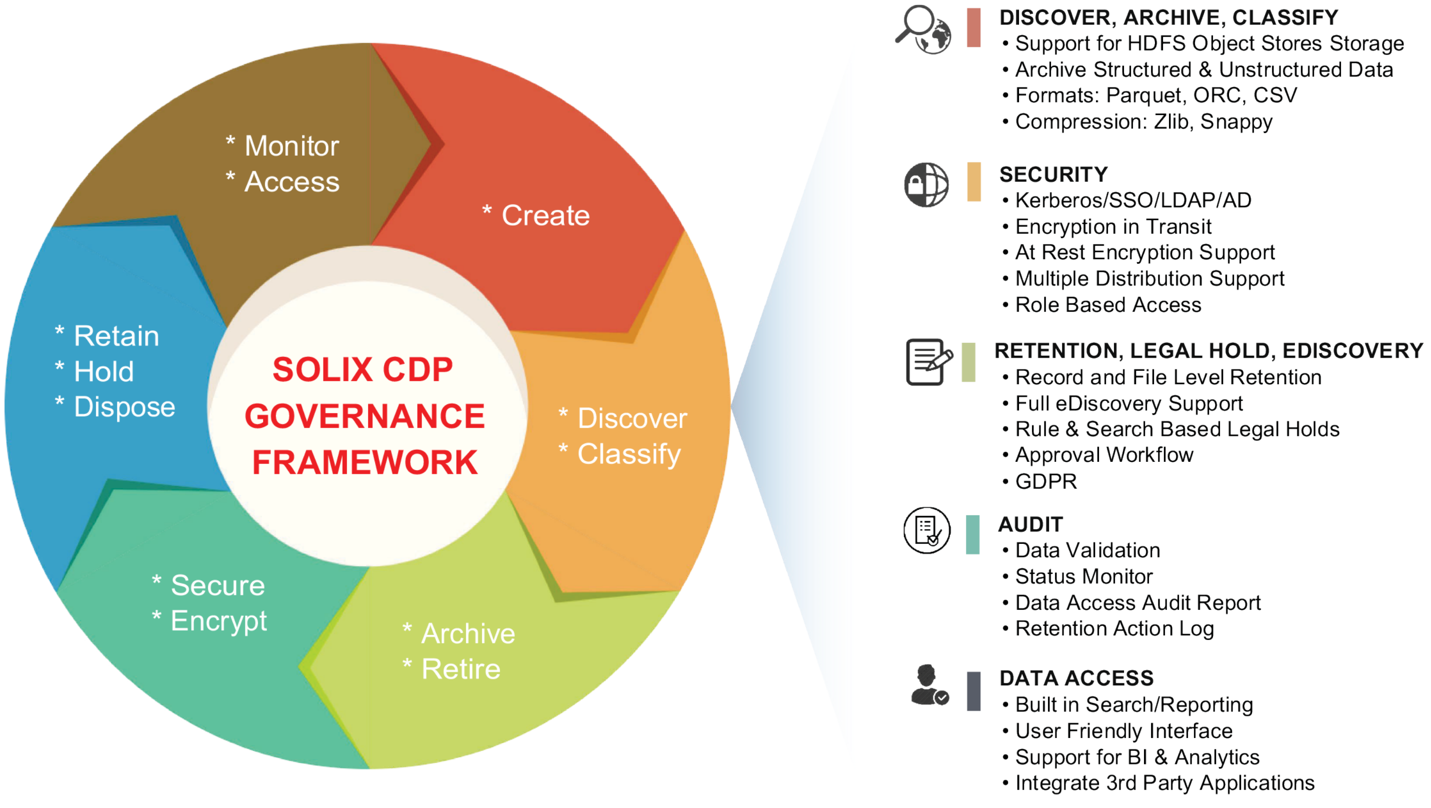
Source: Solix

Information Governance

Analysts have warned that applying existing information governance to a LDW architecture will result in failure. Comprehensive information governance provided by Solix CDP establishes the control framework necessary for proper data access control, data assessment, data discovery, data classification, data validation, retention management, legal hold, and privilege management (see Figure 6). To achieve robust ILM, new security and governance measures must be put into place to match the variety and complexity of the new data assets. The Solix CDP provides a 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 ILM framework manages the data within HDFS and provides integrated retention management and legal-hold capabilities.

Structured, semi-structured and unstructured data from various data sources are migrated into HDFS and HIVE with full data validation and audit reports. These provide the necessary defensibility and chain-of-custody for compliance and data governance. ILM policies and business rules may be pre-configured to meet industry-standard compliance objectives, such as COBIT, or custom designed objectives to meet more specific requirements.

Figure 7: Data Governance



DISCOVER, ARCHIVE, CLASSIFY
- Support for HDFS Object Stores Storage
- Archive Structured & Unstructured Data
- Formats: Parquet, ORC, CSV
- Compression: Zlib, Snappy

SECURITY
- Kerberos/SSO/LDAP/AD
- Encryption in Transit
- At Rest Encryption Support
- Multiple Distribution Support
- Role Based Access

RETENTION, LEGAL HOLD, EDISCOVERY
- Record and File Level Retention
- Full eDiscovery Support
- Rule & Search Based Legal Holds
- Approval Workflow
- GDPR

AUDIT
- Data Validation
- Status Monitor
- Data Access Audit Report
- Retention Action Log

DATA ACCESS
- Built in Search/Reporting
- User Friendly Interface
- Support for BI & Analytics
- Integrate 3rd Party Applications

Source: Solix

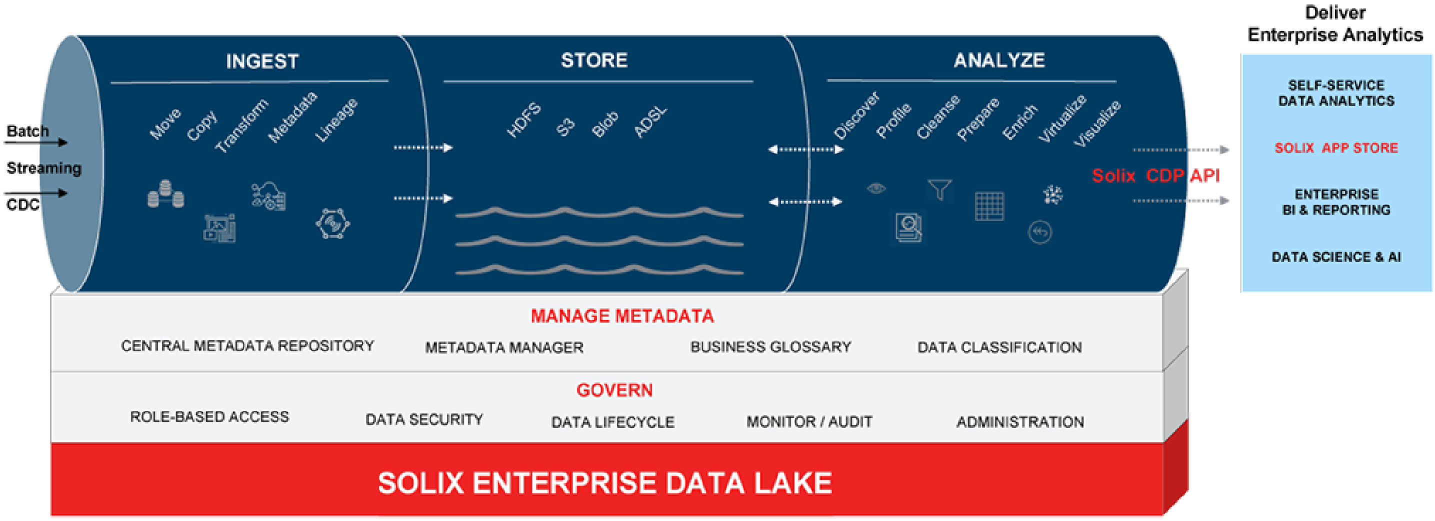
Support for Multiple Technologies

Most organizations are trying to add features and toolsets to their existing infrastructure to support implementation of a comprehensive data layer that can be accessed by all analytic and reporting tools. The implementation of non-relational technologies with cheaper, cloud-based object storage and compute is driving a paradigm shift in the way enterprises are building and architecting their infrastructure. Solix recognized the requirement for multiple technologies, including traditional relational, non-relational, and unstructured data stores, early. It built the Solix CDP to support those technologies and the varied use cases they solve.

The Solix CDP based on Apache Hadoop establishes new capabilities for Advanced Analytics applications (see Figure 8):

- It stores data "as is," also supports transformation, eliminating the need for demanding ETL processes during ingestion.
- It captures and maintains the metadata connected to each byte of data, which is half or more of the value of the data.
- The Enterprise Data Lake may then be minded for critical business insights using text search, structured query, or further processing by downstream analytics applications.
- The Solix CDP uses either Hive or Smart query frameworks, depending on user requirements.

Figure 8: Enterprise Data Lake



Source: Solix

CONCLUSION

By 2020, 10% of organizations will have a highly profitable business unit specifically for productizing and commercializing their information assets.[3]

The era of game-changing digital disruption is here. To thrive in this competitive environment, organizations need leadership who can effectively leverage all the data to derive actionable insights to fuel growth.

Reducing infrastructure costs, attaining operational efficiencies and deriving insights from BI and Advanced Analytics is the desire of many organizations. To achieve this, many are turning to a Logical Data Warehouse approach to manage, integrate and access their data silos. The Solix CDP maximizes the insights that can be achieved while reducing risk, ensuring compliance and governance to create a true ILM framework to lead organizations into the future. The Solix CDP gives organizations all the tools necessary to lower the total cost of ownership and satisfy the desire for return-on-investment.

The Solix Common Data Platform provides a single overarching layer of data management reaching across all physical data storage inside the data center, in remote sites, and in the cloud. The Solix CDP provides full integration with higher level tools for security and data analysis, allowing analysts to access data of all types across the enterprise, in multiple systems – wherever it resides – to provide better, more accurate results. The Solix CDP provides the basis for advanced security, including full metadata records of all accesses to the data, allowing the Chief Security Officer to apply measures, such as end-to-end encryption, across the entire data environment.

Therefore, the Solix CDP provides a comprehensive, fully supported, platform on which the enterprise can build its customized LDW that meets its needs and can evolve with those needs. This is the data platform on which companies need to build their future business infrastructure.

Source: Solix

3Gartner Research, For Midsize Enterprises, the Value of Data and Analytics Goes Beyond Cost Optimization, November 2017

SOLIX®

Empowering the Data-driven Enterprise

Contact us

For more information contact us at:

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

Toll Free: +1.888-GO-SOLIX, (1.888.467.6549)
Telephone: +1.408.654.6400Fax: +1.408.562.0048

info@solix.com
https://www.solix.com/

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.