SOLIX™
Empowering the Data-driven Enterprise

SOLIXCloud Enterprise Data Lake

A Third-Generation Cloud Data Platform

SOLIXCloud Enterprise Data Lake responds to a critical challenge presented by digital transformation. With data volumes projected to reach 175 zettabytes by 2025 - enough to stack Blu-ray discs to the moon 23 times - IT organizations are under immense pressure to manage, govern, and monetize their growing volumes of structured, semi-structured, and unstructured data. And now, demand is growing for real-time enterprise data to power advanced analytics, generative AI, and enterprise intelligence applications.

In his Market Study: *2023 Modern Data Architecture Trends*, researcher John O'Brien notes that 53% of respondents consider modernizing to Cloud Data Warehouse, and 55% "understand well… its compelling business value." And at the same time, CISOs and CDOs are calling for these next-generation cloud data platforms to support advanced metadata management and data governance frameworks to manage all this data throughout its lifecycle.

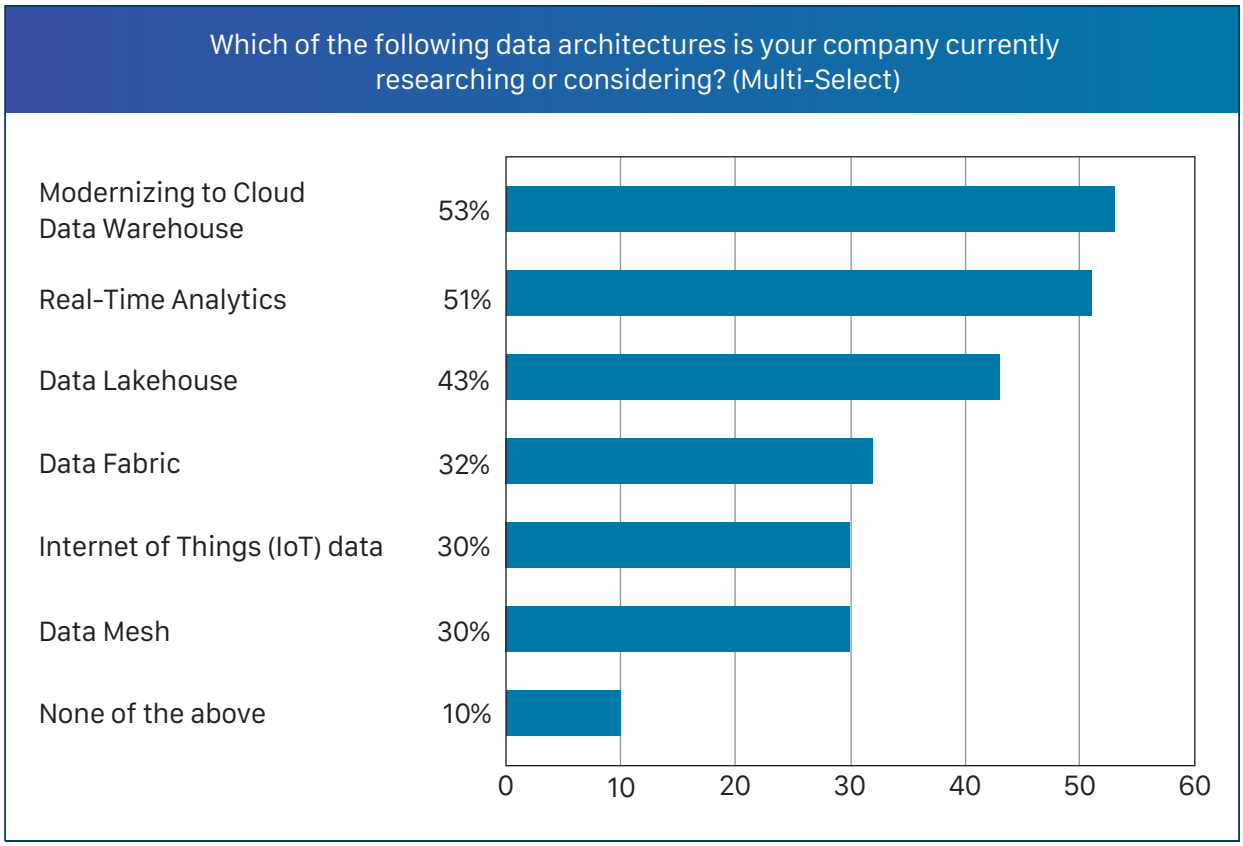

Figure 1. *Market Study: 2023 Modern Data Architecture Trends, John O'brien, Principal Advisor, Industry Analyst, Radiant Advisors, visit www.unisphere.com. Unisphere Media,*

First-generation data warehouse designs processed structured data only and featured fixed, canonical schemas. Under such fixed schema models, data is often prepared first and then bulk-loaded into the database. While this traditional approach is effective to optimize performance for well-organized queries and reporting, the ETL requirements are expensive and slow. Perhaps more important, end-users and application owners still called out for flexibility to use more specific schemas that describe their data better. Data quality also suffered from the lack of fresh data as batch updates were expensive.

Sadly, the vast majority of enterprise data would never make it into the data warehouse, and therefore was never analyzed or monetized.

Second-generation data platforms, often running on Apache Hadoop, introduced the data lake as a practical place to store the tsunami of data piling up in the enterprise. Hadoop offered the major advantage of being able to store not just structured data, but unstructured and semi-structured files as well enabling all enterprise data to be stored in a common repository in low-cost S3 buckets.

These new data lake architectures featured schema-on-read. Instead of an upfront, heavy-lift ETL process, second-generation data warehouses offered the ability to directly query the data and reduce ETL requirements overall.

But still the data lake was not easy to manage or govern, and SQL performance on these platforms also proved a stumbling block. While data lake users could load large volumes of data, analysts were quick to criticize second generation systems as a 'data swamp' over the lack of adequate metadata management and data governance controls.

As a third-generation cloud data platform, SOLIXCloud Enterprise Data Lake shares many of the same features of earlier generation platforms, like separation of storage and compute and schema-on-read, but more significantly, the SOLIXCloud Enterprise Data Lake adds a metadata layer for version control, caching and indexing, and advanced transaction management. Solix Enterprise Data Lake is a lakehouse architecture that supports Parquet files with rich metadata and auxiliary data structures for statistics and data layouts that can be queried by a broad range of APIs and compute engines. And it is from this metadata layer that critical data governance, data security and access controls may be applied to ensure safe, secure and compliant data management with advanced data privacy, legal hold and Information Lifecycle Management (ILM) features.

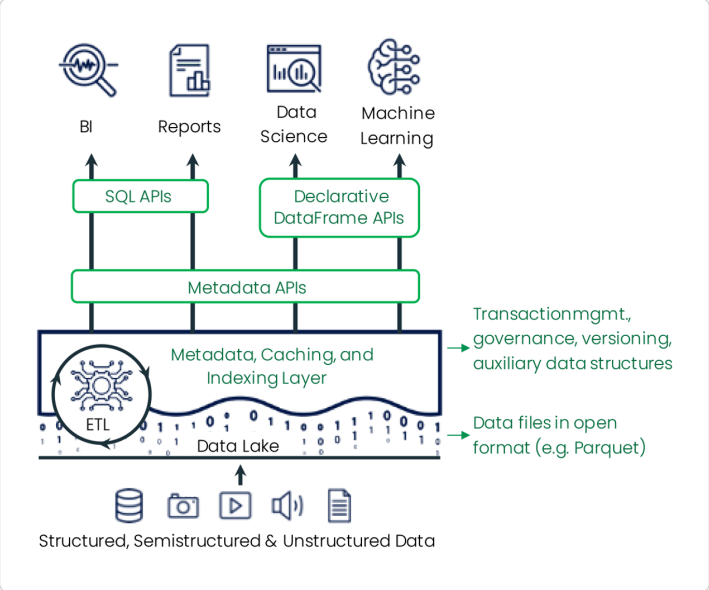

Figure 2: *Example of a Data Lakehouse system design with key components shown in green. The system centers around a metadata layer such as the SOLIXCloud Enterprise Data Lake that adds transactions, versioning and auxiliary data structures over files in an open format, and can be queried by diverse APIs and engines.*

The Solix Enterprise Data Lake leverages open file formats for data that are key to delivering higher performance and robust data governance. Parquet is an open-source, column-oriented file format that supports complex data types and advanced nested data structures.

By converting CSV files into Parquet, many-fold performance improvements may be achieved to reduce scan and deserialization time, and also storage requirements by up to one-third for large datasets of unstructured and semi-structured data.

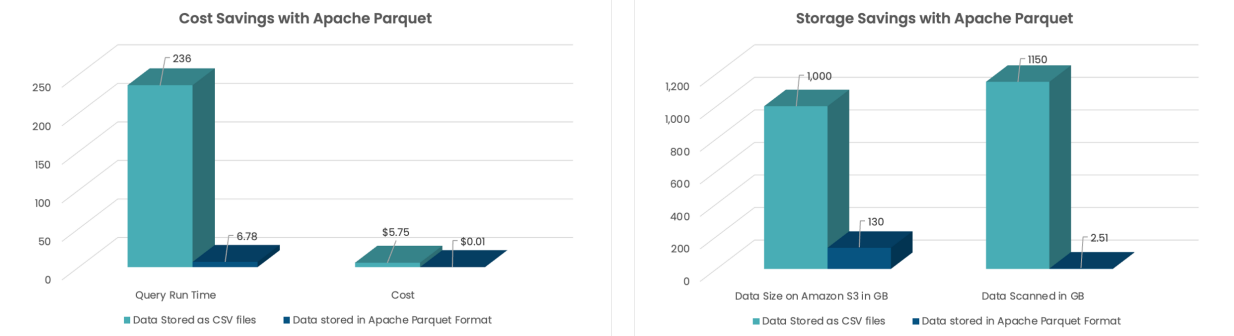

Figure 3: *Difference Between Parquet and CSV*

Running on the cloud-native Solix Common Data Platform (CDP), SOLIXCloud Enterprise Data Lake is a third-generation, transactional streaming data lake that brings core data warehouse and database functionality directly to a data lake. Designed for high-performance, real-time cloud database workloads, the SOLIXCloud Enterprise Data Lake supports ACID transactions and delivers transactional guarantees to the data lake with consistent atomic writes and concurrency controls tailored for longer-running data lake transactions.

To ensure your data infrastructure is not tied to any one vendor, the SOLIXCloud Enterprise Data Lake supports Apache Hudi Open Table Format at customer early access, and Open Table Formats for Apache Iceberg and Delta are planned to follow. SOLIXCloud Enterprise Data Lake is also a multi-cloud solution and an ideal platform for data pipelining, data integration, data engineering, machine learning, advanced analytics, generative AI, real-time data warehouse and low latency, transaction processing applications.

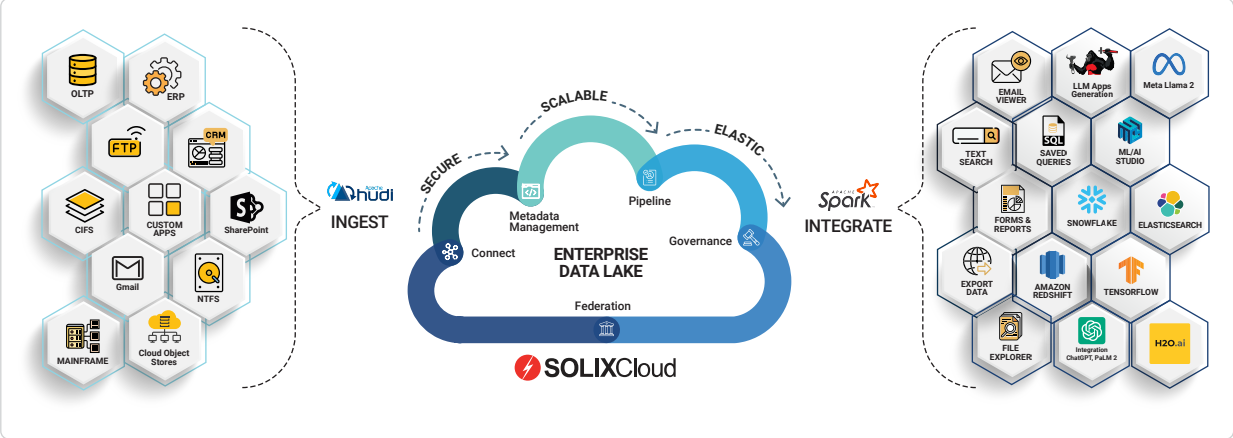

Figure 4: *SOLIXCloud Enterprise Data Lake - A Third-Generation Cloud Data Platform*

The SOLIXCloud Enterprise Data Lake featuring Apache Hudi (Hadoop-Upserts-Deletes-Incrementals) delivers powerful cloud database management features including:

ACID Transactions - The Apache Hudi data lake framework provides real-time, ACID transactional guarantees to your data lake with consistent Atomic Writes and Isolated Reads for Concurrency control tailored to longer-running data lake transactions. These features include Tables, Transactions, Upserts/Deletes, Advanced Indexing methods to manage and query large datasets, Clustering/Compaction, Performance Optimizations to scale writes and reads independently and optimize infrastructure, Bulk Inserts and Transactional Writes, Snapshots so readers don't block writers and writers don't block readers, and Time Travel to enable querying past versions of the dataset useful for audit trails or rollbacks.

Fast Data Processing - Reimagine slow, batch data processing jobs with a powerful new incremental approach to reading and writing data using Streaming Ingestion. Fast Data Processing runs alongside batch data processing and is ideal for incremental data ingestion into HDFS or to re-think and re-engineer ETL processes for Hive and Spark jobs which are running too slow and taking up a lot of resources. Incremental data processing facilitates the processing of only new or updated data since the last batch, enhancing efficiency in data pipelines.

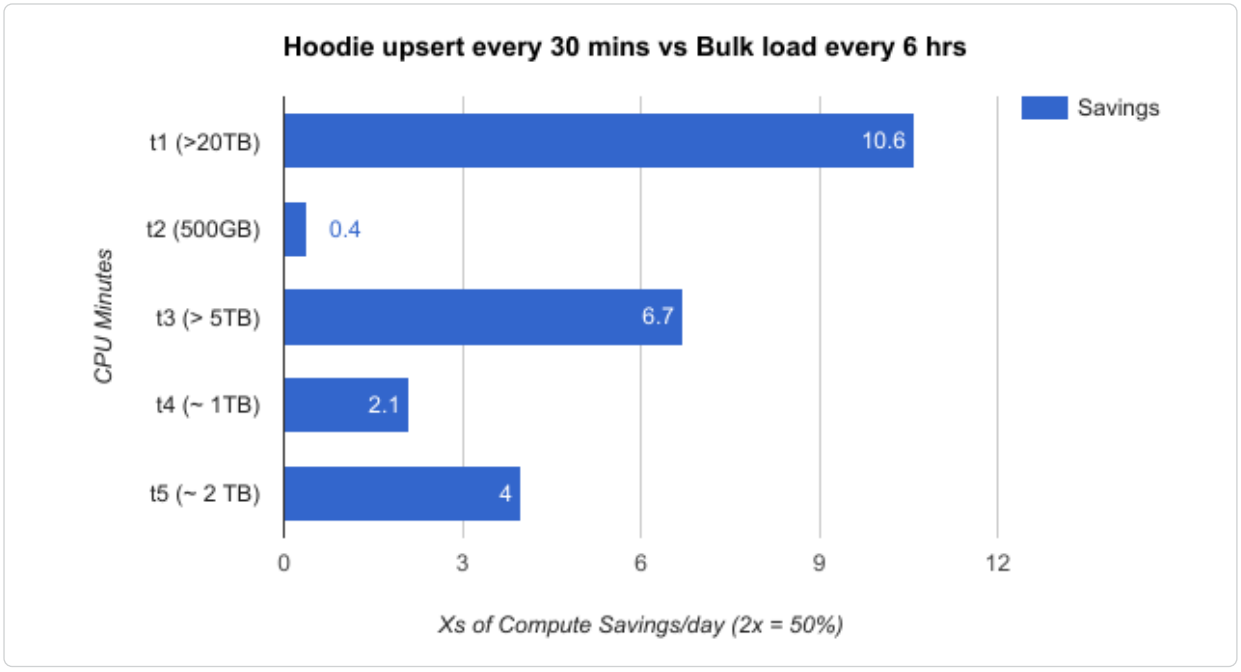

Figure 5: *HUDI Upsert every 30 minutes vs bulk load every 6 hours*

High-performance Loading - Even moderately big NoSQL database installations store billions of rows, making full bulk loads infeasible and a more efficient approach necessary to ingest such data volume. Replace costly and inefficient bulk loads with managed ingestion via Upserts and incremental streaming to keep your data up to date.

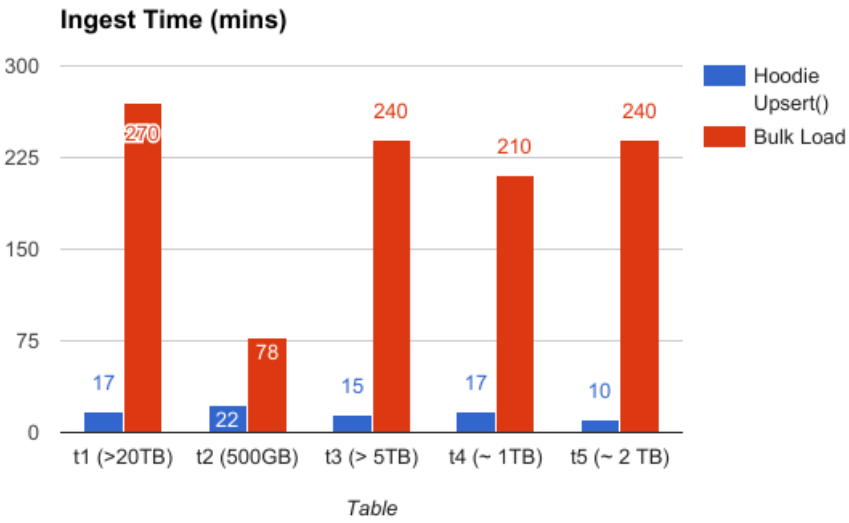

Figure 6: *Graph shows the speed up obtained for NoSQL database ingestion, from incrementally upserting on a Hudi table on the copy-on-write storage, on 5 tables ranging from small to huge (as opposed to bulk loading the tables)*

Competitive Benchmarking is generally not the real world, and the only true tests are from your own data and environment, but a clear pattern emerges when comparing Hudi, Iceberg, and Delta in the TPC-DS decision support performance benchmark. Brooklyn Data, Databricks, and Onehouse performed TPC-DS performance benchmarks between Hudi, Delta, and Iceberg, and while Delta and Hudi are comparable, Apache Iceberg consistently trails behind as the slowest of the projects. Of course, performance isn't the only factor in a database decision, but high performance matters to reduce costs and speed up the results of transaction processing and data pipeline operations.

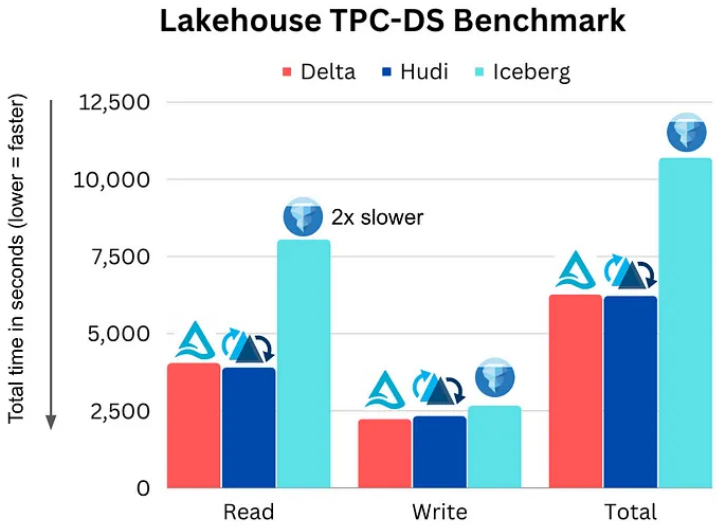

Figure 7: *A clear pattern emerges from these benchmarks, Delta and Hudi are comparable, while Apache Iceberg consistently trails behind as the slowest of the projects.*

Additional SOLIXCloud Enterprise Data Lake features include:

Data Catalog - Data scientists and engineers require a detailed inventory of all data assets and metadata in order to manage data governance and discover data for any analytical or enterprise intelligence purpose. Features include role-based access control, business glossary, data classification, metadata repository, data lineage, and support for Enterprise Business Records, which are complex business objects containing combinations of structured, unstructured, and semi-structured data.

Low-code, Incremental Data Pipelines - To create real-time, incremental data pipelines from source to target that are fit-for-use by artificial intelligence (AI), machine learning (ML) and advanced analytics, data engineers must collect data from any source and apply data cleansing, data enrichment or other transformations. By transforming files, removing erroneous records, masking sensitive data, tagging, labeling, or combining data objects, and delivering incremental, real-time updates, Solix Data Pipeline improves data quality for machine learning and advanced analytics applications.

A **continuous data delivery processing framework is ideal for low latency, minute-level analytics, and changing** data capture workloads. Create declarative templates for incremental ingestion and transformation and provision of continuous data delivery pipelines for machine learning operations. Automate the operational burdens of scheduling, monitoring, and moving enterprise data.

Apache Spark - Apache Spark's parallel in-memory data processing is the world's most widely used engine for scalable computations against structured and unstructured data. Thousands of companies, including 80% of the Fortune 500, use Apache Spark™ today.

Change Data Capture (CDC) - Change data capture enables seamless, efficient database ingestion into your data lake. SOLIXCloud Enterprise Data Lake supports fast upserts and deletions of data suitable for CDC and streaming use cases like Operational Data Store (ODS).

Federated Data Governance - Federated Data Governance provides a centralized control framework to deliver Information Lifecycle Management (ILM) and enforce policy-driven compliance for when several groups have authorities over the data. Through delegated authorities, virtual policy enforcement, and audit management, Federated Data Governance enables compliance control over remote tables and data, reducing risk and improving security for decentralized, multi-cloud data operations.

Cloud data platforms are a cornerstone to any digital transformation strategy and the third-generation SOLIXCloud Enterprise Data Lake brings powerful data governance, data pipelining and high performance, streaming transactions to data lakes. Collect, store and govern all of your enterprise data with a unified data catalog. Provide data warehouse, self-service analytics, and end-to-end data fabrics. Pipeline and publish high-quality datasets to improve the results of machine learning, generative AI, advanced analytics and data warehouse applications at scale.

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: +1.408.654.6400
Fax: +1.408.562.0048
Email: info@solix.com
URL: http://www.solix.com

About Solix Technologies, Inc.
Solix Technologies, Inc. is a leading provider of information architecture and data fabric solutions and is trusted by Fortune 2000 companies for digital transformation and data-driven operations. The Solix Common Data Platform (CDP) is a cloud native, enterprise data management solution for cloud data management applications including Enterprise Data Lake, Enterprise Archiving, Enterprise Security and Compliance and Enterprise AI. Solix is headquartered in Santa Clara, California, and operates worldwide through direct sales and an established network of value-added resellers (VARs) and systems integrators.

References

Figure1. *Market Study: 2023 Modern Data Architecture Trends, John O'brien, Principal Advisor, Industry Analyst, Radiant Advisors, visit www.unisphere.com. Unisphere Media*

Figure 2. *Evolution of data platform architectures to today's two-tier model (a-b) and the new Lakehouse model © https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf*

Figure 3. *https://www.databricks.com/glossary/what-is-parquet*

Figure 5. *https://hudi.apache.org/docs/next/performance/*

Figure 6. *https://hudi.apache.org/docs/next/performance/*

Figure 7. *https://medium.com/@kywe665/delta-hudi-iceberg-a-benchmark-compilation-a5630c69cffc*