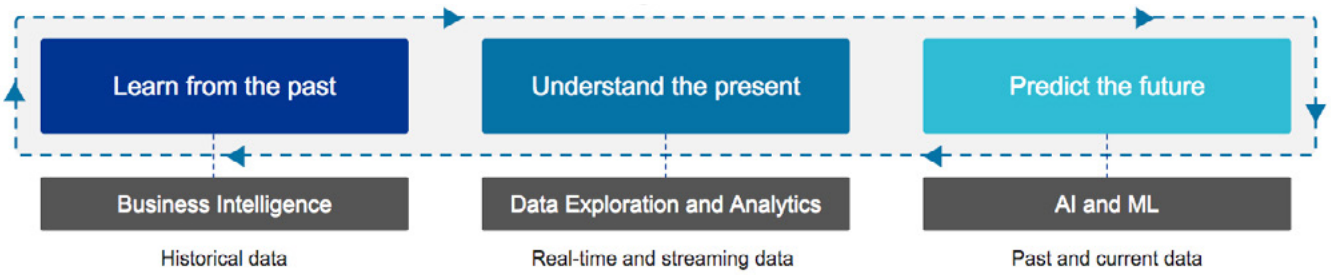
BUILDING A MODERN DATA PLATFORM

Jorge Villamariona
Qubole Product Marketing

Introduction

All data-driven organizations use data in three ways:

- To report on the past
- To understand the present
- To predict the future



Data warehouses are one of today's most important IT assets. They serve as the basis for vital analytics necessary to run today's hyper-competitive, fast-moving businesses. But traditional data warehouses were built on top of traditional Structured Query Language (SQL) based systems that have not fully evolved to handle the amount and diversity of today's data.

Fortunately, newer cloud-based data warehouse services allow organizations to continue to maximize value from their data warehouse and meet today's business requirements of scale and reliability for traditional BI applications.

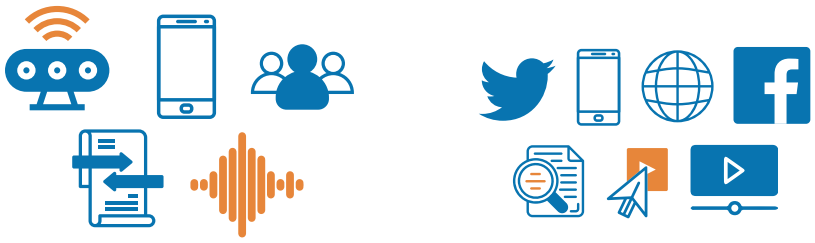
However newer applications like machine learning and real-time predictive analytics move beyond retrospective reporting and basic querying, and require a different set of capabilities from the underlying data platform—delivered by data lakes.

A data warehouse, when enhanced with a data lake to form a comprehensive cloud data environment, will allow today's businesses to manage massive data sets, integrate structured and unstructured data, and redesign data preparation processes for greater scale.

Industry Leaders Win with Data

Searching for a common denominator amongst the world's most successful companies, we see that they all have figured out innovative ways to exploit and take advantage of their data. They take full advantage of their traditional structured data by leveraging data warehouses1 while leveraging data lakes for their semi-structured and unstructured data sets. Today's data warehousing teams are fully aware that the requests they receive to support business initiatives will not be met with a traditional relational database management system. According to studies from Gartner and TDWI[2], leading data platform characteristics include:

1. **Data volumes:** Supporting today's increased data volumes while providing the agility to support new business goals.
2. **Data variety:** Embracing new data sources and types, particularly unstructured and semi-structured data from web, social, and IoT devices and sensors.
3. **Scalability:** Increasing scale for ingestion, transformation, and processing of data.
4. **Agility:** Moving from retrospective reporting to machine learning and real-time predictive analytics, IT teams today must enable advanced analytics with machine learning and artificial intelligence (ML/AI) that utilize massive volumes of data and require high degrees of parallelism for performance. These ML/AI models make software smarter than ever and are becoming critical to compete in all categories, from risk management and fraud detection to personalized engagement and driverless cars. However, these new techniques require new tools that are not SQL-based and are not supported by a data warehouse. [1]



Sensors/ Click Streams/ Social Media/ Transactions/ Market Prices

Streaming data sources | IOT Data Sources | Web Data Sources

Diagram 1 - New Data Sources

Underpinning these requirements is a movement to the cloud, an IT-wide trend toward building platforms that are elastically scalable at a low cost.

[1] Throughout this document we use the term data warehouse to represent a traditional Enterprise data warehouse anchored by a relational database.

[2] Transforming Data with Intelligence — formerly The Data Warehousing Institute

What Industry Leaders do Differently

Industry leaders are able to leverage all of their relevant data regardless of format. Modern organizations are being asked to deal with ever increasing volumes of data, whether **Structured, Unstructured and Semistructured**. Let's take a brief look at each one.

Structured: Any data that is stored, accessed, and processed in a fixed format is termed as 'structured' data. Data warehouses are designed to support structured data through the use of a relational database, although the scale of data today is pushing the growth limits for many of these solutions.

Unstructured: Any data with unknown form is classified as unstructured data. Please refer to **Diagram 1** above. Examples of unstructured data include: a heterogeneous data source that contains a combination of simple text files, images, and videos; data from tools based on natural language processing; search results from google; and text analytics that provide visibility into text-laden business processes (such as the insurance claims process, medical records, or call center and helpdesk agent notes). Sentiment analysis and voice of the customer, based on natural language data, have also become very common in customer-oriented businesses that use social media data. Sensor data from robots in manufacturing and vehicles are often seen today as well. Organizations have a wealth of unstructured data available to them, but unfortunately, because of multiple processing challenges, they are unable to derive value out of it. Traditional data warehouses do not fully support unstructured data because their underlying relational database systems do not fully support these ever growing data formats.

Semi-structured: Semi-structured data can contain both structured and unstructured forms of data. We can see semi-structured data as structured in form, but not necessarily in a table in a relational database. An example of semi-structured data would be data represented in an XML file, containing for example personal data about a user. Partnering firms that work together through a supply chain often exchange information via XML and JSON documents, which include a mixture of structured data, hierarchies, text, and other elements. Analysis of this data can help quantify profitable partnerships and supply chain management inefficiencies. Traditional data warehouses can support semi-structured data but only after it has been transformed into a structured data format.

Foundations of a Modern Data Platform Architecture

Data-driven industry leaders leverage a data warehouse, however, their data warehouse is complemented with a cloud-based data lake to provide the future scalability and agility that a traditional data warehouse simply can't give them. They don't just focus on what has happened in the past. They also focus on what is happening TODAY using real-time and streaming data sources that they can combine with historical batch-type data sources. But even this is not enough to make them true leaders. They also use all this current and historical data to build analytical models that allow them to predict, with a great degree of confidence, what is likely to happen so that they can plan better and innovate or serve their customers in a better way.

All of these organizations have realized that they require a modern data platform architecture in order to accomplish their goals. This architecture continues to have a role for traditional technologies such as data warehouses, BI tools, and standardized ETL (Extraction, Transformation and Loading) processes and tools, but it also leverages data lake platforms in order to work with their streaming, unstructured and semi-structured datasets. They want to increasingly use this data lake platform in a more efficient and strategic way in conjunction with their more traditional structured data warehouse sources. The whole purpose is to provide data exploration and ad-hoc analytics capabilities to everyone directly from the trusted datasets in the data lake platform. The difference is the speed and agility with which users get access to and can gain insights from the data.

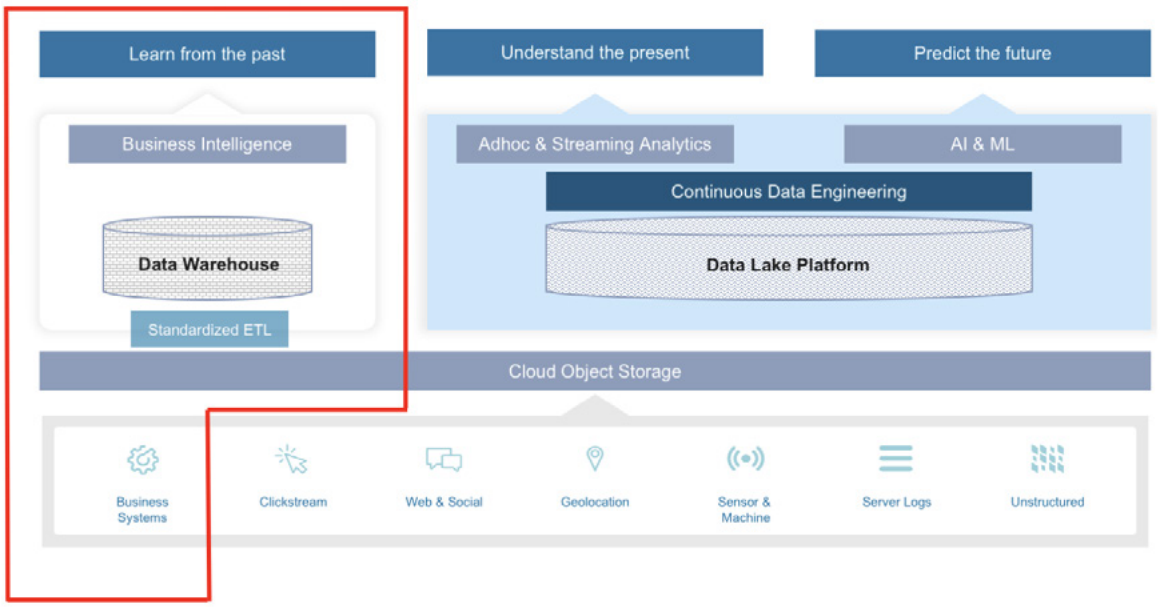


Diagram 2. Modern Data Platform Architecture

With traditional data warehouse platforms, data has to go through a standardized ETL process to load the data and make it available to the users – which takes time, whereas with a modern data lake platform, it is not required to do this ETL up front. What we think of as ETL happens at the time the data is read or consumed by the user. Through a process of continuous data engineering, and a-cyclical process, data engineers make data available in the most efficient way allowing users to get access to the data directly from the data lake. This is critical to today's analytics needs because data today can become stale in a matter of hours, or in some cases minutes, and often, organizations don't have the luxury of going through long standard ETL processes to make the data available to their users.

A Modern Data Platform also accelerates data warehouse initiatives and ROI by combining traditional and non-traditional approaches to federate relational and non-relational data stores into one cohesive architecture. This enables new practices that complement the core data warehouse without replacing it, because the warehouse remains the best platform for the aggregated, standardized, and documented data used for standard reports, dashboards, and OLAP.

This Modern Data Platform consists of multiple specialized components that are optimized for workloads to manage, process, and analyze new types of data — whether larger volumes, unstructured formats, or in real time. Such a Platform can serve as the single ingestion point for all new data. Organizations transform and process data, formerly destined only for the data warehouse, with a schema-on-read approach directly from the data lake. **Diagram 2** above shows a modern data platform architecture where a cloud data lake enhances the traditional data warehouse. Note the smaller footprint of a traditional data warehouse centric architecture.

This approach of enhancing your existing data warehouse with a cloud data lake provides freedom of choice. You will no longer need to compromise between technical fit or cost reduction for managing all types of data. When an integrated data environment leverages a cloud-native data platform, companies get the best approach to managing all types of data from the perspective of cost, performance, and flexibility. This integrated data environment is able to provide **Data Engineering, Data Analytics and Machine Learning services which will ultimately yield trusted data sets to the business.**

Building a Modern Data Platform

The most common data environments today are anchored by a data warehouse. Regardless of the sophistication of your current data environment, it can be enhanced. At the data warehouse basic level, there may be hardware and software upgrades as well as the addition of new data subjects and dimensions. This will help with increasing the scalability and performance of your data warehouse but it will continue to only support relational data without the ability to support semi-structured or unstructured data for Machine Learning (ML) Applications.

Getting your data warehouse ready to support current Analytical and future (or present) ML needs requires that you go through three **(3)** well defined steps. First, **migrating your data warehouse to the cloud**. Second, **adding a new cloud object data store to support unstructured data**. Third, **creating an integrated cloud data environment**. These steps will allow you to obtain the maximum payoff and prepare your environment for machine learning and data science. Let's take a closer look at each one of these steps:

1. Ingest data into an object store in the cloud

A founding principle behind data warehousing is that user organizations should repurpose data from the enterprise and other sources to gain additional insights and guide decisions. In that spirit, organizations are grappling with new data types and sources (big data) and how to capture and manage this information in a way that is advantageous to the business.

Managing and leveraging new data types and sources such as semi-structured and unstructured format is worthwhile because of their business value. However, IT teams face challenges around the newness of the data types, the massive volumes, the wide range of data structures, and the streaming nature of some sources. The variety of data further compounds the problem, because **most traditional data warehouses were designed for structured or relational data alone.**

In order to preserve an existing data warehouse investment and support new types and sources of data, many companies choose to reserve their core data warehouse for the relational data that goes into standard reports, dashboards, and analytics. For new big data, these companies are deploying specialized object store platforms like Amazon S3, Google Cloud Storage, or Microsoft Azure Storage built for new data types and then integrating them with the core data warehouse.

These cloud-based storage platforms provide better cost performance and support greater diversity of data type over their on-premises counterparts and that makes them well suited to support a number of data science and ML use cases such as: data staging, archiving, computational analytics, and analytical sandboxes for exploration and discovery. **It is important to note that adding a cloud object data store is the initial step towards a data lake but that's only the physical storage -- and leveraging a cloud object data store by itself doesn't mean you have a data lake. The greater value will be derived from building an integrated cloud data environment supported by a cloud data lake. (most companies today use a data store, even for temporary storage on the cloud as a staging area)**

2. Integrate your Data Platform

In parallel to the above efforts, you will update the data ingestion and transformation processes to account for migration to the cloud and the addition of a data lake. As you evolve these pipelines, the data lake becomes the central repository for all unstructured and semi-structured data (absent the burden of predefining your data schemas). Eventually, the data lake could feed both the production BI/data warehouse environment and the exploratory analytics sandbox with processed and curated data.

One immediate enhancement opportunity is to load data sources that are currently read through ETL routines directly into the data lake. This data can start feeding data science and machine learning models immediately thanks to the schema-on-read capabilities of the data lake, without impacting the functions underpinned by the traditional database.

An easy way to start building experience with using the data lake for ETL is to parse data to create new metrics from an unstructured data source that can be fed into the existing data warehouse. This provides the ability to leverage data such as social, mobile, consumer comments, email, doctors' notes, or claims descriptions to create new metrics that better predict behavior than would be possible without the data lake. Organizations can easily integrate these new metrics into their existing business intelligence queries, reports, dashboards, and analyses by combining structured and unstructured data.

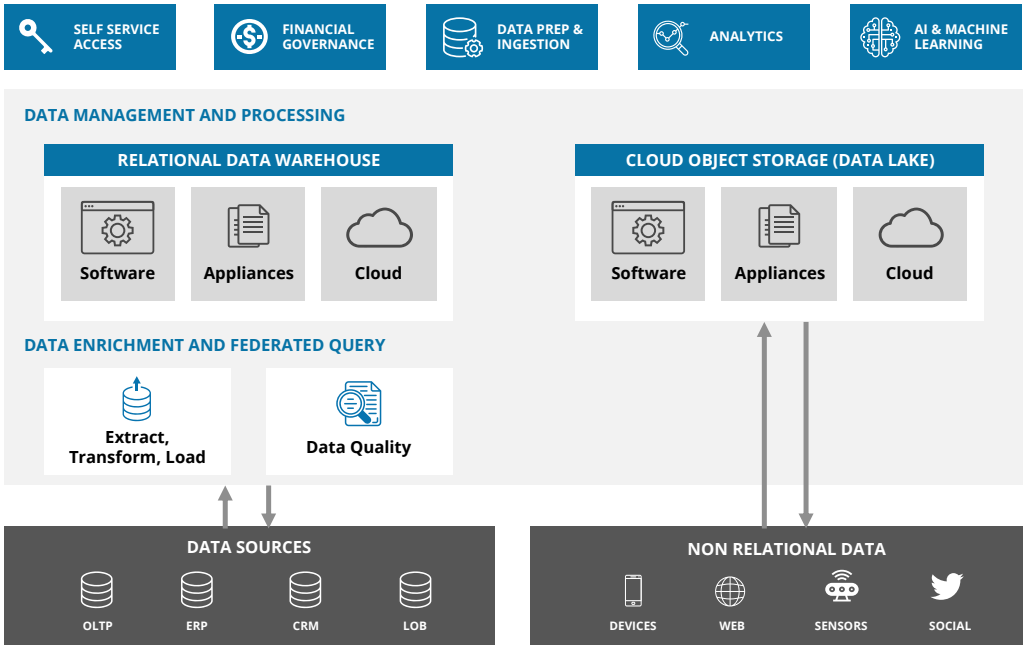


Diagram 3. Integrated Data Environment

This new cloud-based integrated data environment (Diagram 3) will give your organization the functionality you previously had with your data warehouse, but now with greater flexibility and scalability thanks to the cloud. Your organization will now be able to store and take advantage of semi-structured and unstructured data to support ML applications. Finally, blending unstructured and structured data will provide your organization with greater and richer analytical insights into the operations of the business and the entities your organization tracks through various data sets.

Data Lake Best Practices

Once your organization has chosen to modernize your data environment by adding a data lake, it is important to keep some data lake best practices in mind. There are three key areas that require particular attention from data teams in order to build effective data lakes: data ingestion, data layout, and data governance. Let's take a closer look at each one of these areas.

1. Data Ingestion Best Practices

Data must be ingested into the lake before it becomes useful for analysts and data scientists. Data ingestion lays the foundation for the effectiveness of downstream analytics. Architectural best practices for data ingestion include:

Managing Batch and Streaming Data Ingestion. Data flows into a data lake with variable velocity and in various types. Whether ingesting batch or streaming data, the data lake must ensure zero data loss and write exactly-once or at-least-once; handle schema variability; write in the most optimized data format into the right partitions; and provide the ability to re-ingest data when needed. The data lake must support continuous optimizations that create aggregates and curated datasets for downstream analytics.

Source-to-Target Schema Conversion. Data engineers will benefit from the ability to intelligently detect source schema and create logical tables on the fly, and flatten semi-structured JSON, XML or CSV into columnar file formats. The schema of these tables should be kept in sync to support continuous integration.

Monitoring Movement of Data. DData ingestion pipelines sometimes fail for reasons that are difficult to control, such as erroneous data, schema drift, intermittent platform glitches, and more. That's why it's important to connect pipelines and the underlying infrastructure to rich monitoring and alerting tools that shorten time to recovery after a failure. As a best practice, erroneous records should be sent to an error stream for root cause analysis.

2. Optimize your data layout for Analytics and Machine Learning

Data generation, and by extension data collection, is bursty and continuous. Machine data is often sparse and Data generation, and by extension data collection, is bursty and continuous. Machine data is often sparse and collected in semi-structured formats such as JSON or CSV. Attempting to inspect, explore and analyze these datasets in their raw form is arduous and slow, because the analytical engines scan the entire data set across multiple files. We recommend planning ahead and building ETL pipelines that reflect a well-defined layout strategy for frequently accessed data. The key is to reduce data scanned and query overheads using different techniques.

Use Columnar Data Formats for Read Analytics. Publish data in open source columnar formats such as ORC and Parquet to reduce data scans. These machine-readable, binary file formats are optimized for read analytics. Proactively avoid queries that need to parse JSON with functions such as json_parse and json_extract as they scan the entire dataset. Flatten frequently accessed JSON values by cleaning them, casting to a data type and storing in a columnar format.

Partition Data. Partition data by frequently used predicates (SQL WHERE clause) such as time, geography and line of business to reduce unnecessary data scans. Since partitions add metadata to the metastore and result

in lookup overheads in query execution, it is important to tune the partition granularity based on the dataset under consideration. A typical requirement is to partition by year, month, week, day or hour — but not by minute or second.

Use Compaction to Chunk Up Small Files. The bursty arrival of data, along with real-time stream ingestion, results in data written into multiple files of different sizes in the cloud object store. Compaction is necessary and a better strategy than attempting to tune ingestion systems based on data arrival patterns, which are unpredictable or at least hard to predict.

Collect Statistics for Cost-Based Optimization. Collect and maintain statistics of the dataset such as file size, rows and a histogram of values. The cost-based optimizer in the analytical engine's runtime can significantly improve performance by using the available statistics to optimize queries through techniques such as join reordering.

Z-Order Indexed Materialized View for Cost-Based Optimization. A materialized view (MV) is an ordered copy of data based on a particular sort key. Analytical engine runtimes can use materialized views to selectively scan data for query plans with filter conditions. Indexes should be maintained not just at the table-level but also the partition-level. While more than one materialized view can be configured for a given table, that requires additional storage and computing to keep it updated. Z-order indexing on a single materialized view helps solve this problem. A z-order index serves queries with multiple columns in any combination and not just data sorted on a single column.

3. Data Governance

When data ingestion and data layout are implemented well, data can be made widely available to users in a democratized fashion. When multiple teams start accessing data, data architects need to exercise oversight and manage outcomes. Enterprise-grade data platforms that serve customers well and deliver meaningful experiences need to blend the best of innovation with oversight, regulatory compliance and role-based access controls.

Discover Your Data. Data itself is hard to find and comprehend and not always trustworthy. Users need the ability to discover and profile datasets for integrity before they can trust them for their use case. A data catalog enriches metadata through different mechanisms, uses it to document datasets, and supports a search interface to aid discovery.

Keep Regulatory and Compliance Needs in Mind. New or expanded data privacy regulations, such as GDPR and CCPA, have created new requirements around Right to Erasure and Right to Be Forgotten. These govern consumers' rights about their data and involve stiff financial penalties for non-compliance, so they must never be overlooked. The ability to delete specific subsets of data without disrupting a data management process is essential.

Permissioning and Financial Governance. Cloud data lakes facilitate instant access to data and avoid long procurement cycles. For example, a deep integration with the Apache Ranger open source framework facilitates table, row and column level granular access. With wide-ranging usage, monitoring and audit capabilities are essential to detect access violations, flag adversarial queries, and more. While the cloud offers agility, it can come at a high price if you take your eyes off of cost or don't forecast computing needs. A serverless approach is optimal for cost controls.

Your chosen data platform to power your cloud data lake should offer: Intelligent Query Management - Reliable query execution through a smart, patent-pending algorithm to estimate and allocate resources; Support for Custom Metastores - So there should be no need to recreate schemas; Unified Query Experience - The ability to Seamlessly switch between managed and serverless mode as needed; Autoscaling - The ability to scale up, down, or out to match your workload and SLA requirements; Flexibility of access - UI, REST API, Scheduler, and ODBC/JDBC access.

Conclusion

Traditional data warehouses are unable to meet the growing needs of the modern enterprise to integrate and analyze a wide variety of data being generated from social, mobile and sensor sources. More importantly, these data warehouses struggle to answer the forward looking, predictive questions necessary to run the business at the required levels of granularity or in a timely manner to remain competitive.

There are multiple ways for organizations to begin to benefit from the advantages of an integrated cloud data environment. Each of the steps described delivers its own business benefits. Organizations that employ all of these steps will see lower costs through decreasing data acquisition, maintenance and administration, while improving overall performance, agility and scalability.

Enhancing your data warehouse with the addition of a cloud data lake for new predictive and machine learning use cases and then creating a single logical view which combines both data platforms is a best-of-all worlds approach. This allows organizations to keep up with the growing volumes of data, the velocity of data and the variety of data types required in today's business environment. It leverages past investments and supports today's scale with a practical approach.

For further reading on data lakes best practices, please refer to the Gartner Report entitled Building Data Lakes Successfully by Sumit Pal.

Qubole is a single multi-cloud platform that integrates your cloud data warehouse with your cloud data lake providing your data teams an easier path to machine learning, analytics, data preparation and ingestion while lowering your computing costs by over 50% and enabling self-service which helps you deploy in days instead of weeks.

Qubole is passionate about making data-driven insights easily accessible to anyone. Qubole customers currently process nearly an exabyte of data every month, making us the leading, and industry first cloud-agnostic Open Data Lake Platform. Qubole's Open Data Lake Platform self-manages, self-optimizes and learns to improve automatically and as a result delivers unbeatable agility, flexibility, and TCO. Qubole customers focus on their data, not their data platform. Qubole investors include CRV, Lightspeed Venture Partners, Norwest Venture Partners and IVP. For more information visit www.qubole.com