Qubole®

Modern Unified Data Lake and Data Warehouse Architecture

The 2018 Guide to Modernizing Your Data Warehouse

Authored by Steve Dille

Introduction

Data warehouses are one of the most important IT assets today. They serve as the basis for vital analytics necessary to run today's hyper competitive, fast moving businesses. But data warehouses were built on top of traditional SQL databases that have not evolved well to handle new requirements. They were not designed to handle the volume, variety and velocity of todays data. Nor can they support todays new applications such as machine learning or real-time predictive analytics that move beyond retrospective reporting and basic querying. Over the years, more and more data has been added to warehouses, while the BI query load has also grown exponentially. Consequently, many data warehouses are now over burdened, lack agility and do not meet today's analytic needs.

Fortunately, new big data cloud services will allow organizations to continue to get value from their data warehouse and meet today's business requirements without starting over. Today, new object data storage and management technologies in Amazon S3 or Microsoft Azure can be used to construct a data lake to support the rapidly increasing volumes, variety and velocity of data modern enterprises need. A data lake, when paired with an existing data warehouse to form a unified cloud data platform, will allow today's businesses to manage massive data sets, integrate structured and unstructured data, redesign ETL and data preparation processes for scale and perform machine learning and predictive analytics. Those following this approach can expect lower CAPex, and OPex while improving performance, agility and scalability from using optimized technologies for the work at hand.

1. Why Do Data Warehouse Modernization

Today's data warehousing teams are probably already aware that the requests they receive to support business initiatives are not going to be met with a traditional relational database management system. According to studies from Gartner and TDWI, the leading drivers for data warehouse modernization include:

- Realigning the data warehouse to reduce the incremental costs of supporting today's increased data volumes while providing the agility to support new business goals. Today's businesses run by the numbers and modern analytics more than ever are required.
- Increasing data warehouse scale for ingestion, transformation and processing of big data while continuing to support growing volumes of queries and reports from ERP and CRM data.
- Moving from retrospective reporting to machine learning and real-time predictive analytics. IT teams today must enable new analytics like 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 everything from risk, fraud, personalized engagement to driverless cars. But, these new techniques require new tools which are not SQL-based and are not supported by a data warehouse.
- Embracing new data sources and types, particularly unstructured and semi-structured data from web, social and IOT devices and sensors.

Underpinning these drivers is a movement to the cloud, an IT-wide trend towards building platforms that are elastically scalable at low cost which is also now a possibility in data warehousing with new cloud platforms like Snowflake, Microsoft Azure SQL Data Warehouse or Amazon Redshift.

Data Lakes Modernize the Data Warehouse

Adding a new data lake as a complement to the data warehouse has emerged as the answer to supporting these modernization desires for many organizations. Data lakes address the shortcomings of data warehouses in four ways. First, in data lakes, the data can be stored in structured, semi-structured, or unstructured formats. Second, the data schema is decided upon reading, rather than loading, or writing, the data. You can thus always change the schema if there is extra information or structures that you need from the raw data, leading to greater organizational agility. This also means that the data is quickly available because it does not have to be curated before it can be consumed by the processing engines. Third, data lakes on cloud services

> *"…. data warehousing has reached the most significant tipping point since its inception. The biggest, possibly most elaborate data management system in IT is changing".*
>
> **Gartner**
> The State of Data Warehousing Report

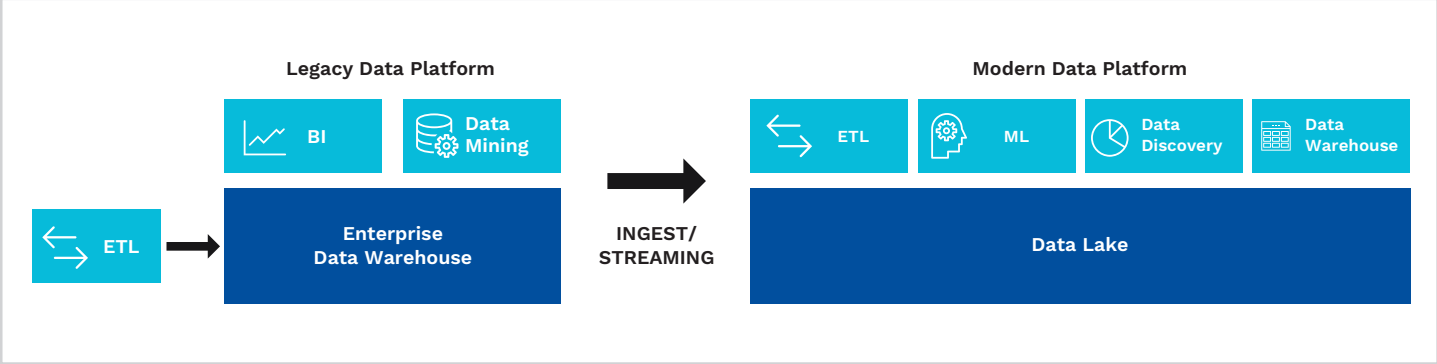
are cost effective due to the separation of compute and storage and the pay per use model. They can also scale with massive compute power available at a moments notice. Last, these object stores support the best in class tools like Spark or Tensorflow used by data scientists in building machine learning and predictive analytics.

Because of the cost effectiveness of data lakes, there is never any need to throw away or archive the raw data. It is always there should any of your users want to revisit it. All of these points—cost effective storage of all content, different types of processing abilities and engines on both structured and unstructured data, fast availability of data, agility, and flexibility—are essential as organizations strive to meet today's needs.

However, what many organizations have found, is that building a data lake to capture all of these new data types alone is not the final solution. What results is two separate platforms - the data lake which supports ML/AI and non-structured data types and the data warehouse for reporting and business intelligence. What is needed is a way to unify the data warehouse and the data lake to coordinate all of the enterprise data ingestion and make it available to the right tools. As a way forward, to enable optimized support of traditional data warehouses and the new data lakes, the concept of a unified cloud data platform has become the prefered path to modernization.

A unified cloud data platform 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 is still the best platform for the aggregated, standardized, and documented data that goes into standard reports, dashboards and OLAP. So instead of replacing the data warehouse, a unified cloud data platform consists of multiple specialized platforms that are optimized for workloads to manage, process, and analyze data that's new, big, unstructured, or real time. It serves as the single ingestion point for all data. Transformations and processing of data for the data warehouse can be performed or data can be accessed with a schema on read approach directly from the data lake.

Breakdown of Data Warehouse-Emergence of Data Lake

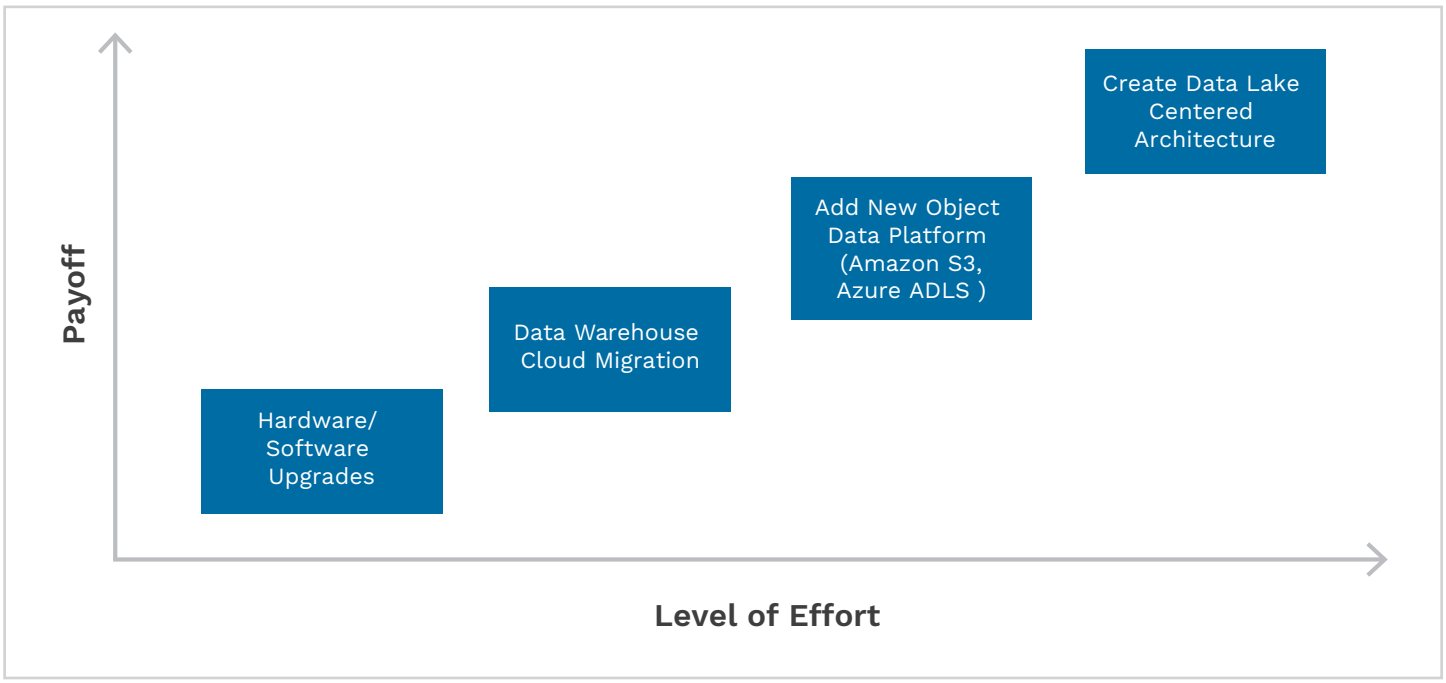


A major benefit to this approach is that it gives users options: they can match a given data set or workload with a platform that's the best technical fit or the most cost-effective. When a unified data platform is placed in the cloud, companies literally have a best of all worlds approach to data warehousing from a cost, performance and flexibility perspective.

2. Data Warehouse Modernization Approaches

No matter the vintage or sophistication of your data warehouse, it probably needs to be modernized. This can take many forms. At the most basic level, there are hardware and software upgrades and the addition of new data subjects and dimensions. This can help with increasing the scalability and performance of your data warehouse and support similar new relational data. Moving the data warehouse to the cloud is another viable option today. Again this will help with scalability, cost and the adaptability to change for structured data, but will not allow you to meet the emerging needs for new data types, faster velocity or support machine learning. Only integrating a new object store data platform, designed for massive scale, rapid data collection, unstructured data and machine learning, with your existing data warehouse will allow organizations to be truly modern in their approach.

Data Warehouse Modernization Strategies



Beyond the simple hardware and software upgrades, there are 3 stages you will most-likely pass through to create a modern unified cloud data platform. First, a migration to the cloud of your data warehouse. Second, the addition of unstructured data and or ML capabilities via a new object store data lake platform. Last, the transformation of all data flows to use the data lake as the new ingestion area for all data warehouses, data marts, analytic sand boxes etc.. in the organization. Each of these approaches has merit on the path to a unified data architecture.

> *"Modernization is more about integrating new platforms than replacing old ones"*
>
> **The Data Warehouse Institute**

Data Warehouse Modernization Approaches Compared

	Migrate to a Cloud Native Data Warehouse	Add a New Object Store Platform (Data Lake)	Unified Cloud Data Platform
Analytic Needs	Retrospective Analytics, Traditional BI	Machine Learning, Predictive Analytics, Artificial Intelligence	Varied, rapidly evolving analytic requirements
Impetus to Modernize	Cost, Supporting Growing Scale of Structured Data	Supporting Unstructured Data, Faster Access to Data, Model & Engine Diversity	Higher velocity of analytics and insights
New Advantages	Lower Cost from Separation of Compute and Storage, Auto Scaling, Control	Analytic Flexibility from Schema on Read, Fast Ingestion, Raw Data Available, Spark and Other Engines Available	No limits on analytic capabilities

I.) Migration to a Cloud Native Data Warehouse

While not an absolute requirement for modernizing your data warehouse, migrating to a cloud native data warehouse such as Snowflake, Microsoft Azure SQL Data Warehouse or Amazon Redshift is one of the easiest ways to achieve better scalability, lower costs and flexibility. By separating compute and storage in the underlying platform, a cloud migration will accomplish some of your goals, but an organization stopping there still won't have support for all of the emerging unstructured and semi-structured data types, real-time ingestion at scale, machine learning/predictive analytics or analytics sandboxes. But, the RDBMS options in the cloud will be very similar to what you may be supporting on-premises today so you will be able to migrate SQL queries, continue using major BI tools and reports largely unchanged.

Whether migrated to the cloud or not, maintaining the traditional data warehouse, which supports the operational and management reporting with respect to understanding "what happened?" types of business questions today is important. Businesses run on this information and augmenting this environment rather than replacing it is the right approach for many.

II.) Add a New Cloud Object Store Data Lake to Support New "Big Data"

A founding principle of 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 for business advantage.

Big data' could be found in three forms:

1. **Structured**
2. **Unstructured**
3. **Semistructured**

Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. This is what data warehouses support well today in an RDBMS, although the scale of this data is pushing the growth limits for many.

Unstructured

Any data with unknown form is classified as unstructured data. In addition to the size being huge, unstructured data poses multiple challenges in terms of its processing for deriving value out of it. Typical examples of unstructured data would be: a heterogeneous data source containing a combination of simple text files, images, videos etc.

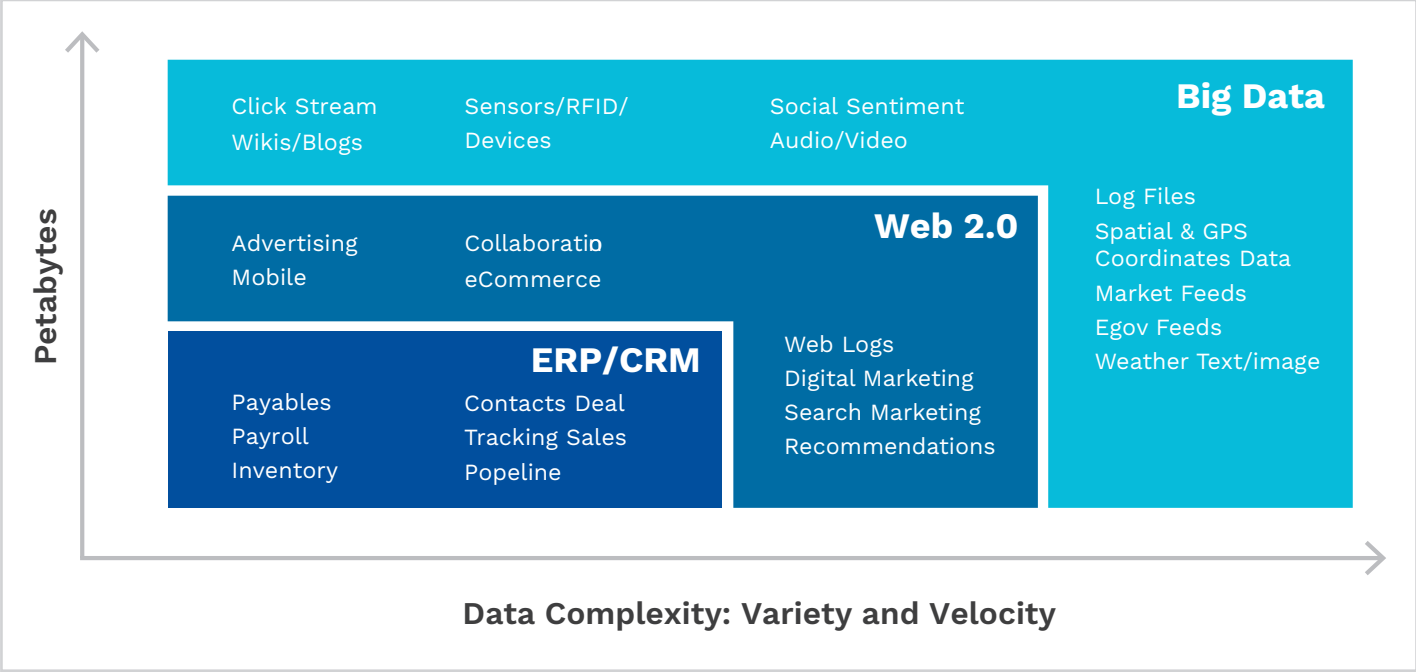
Analyzing unstructured data is useful. Tools based on natural language processing, Google search results, and text analytics provide visibility into text heavy business practices, such as insurance claims, medical records, or call center/ help desk agent notes. Sentiment analysis and voice of the customer, based on human language social media data has also become very common in customer oriented businesses. 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 they don't know how to derive value out of it.

Semistructured

Semistructured data can contain both the forms of data. We can see semistructured data as structured in form but it is actually not defined with a table definition in an RDBMS. An example of semistructured data would be data represented in XML file 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 partners and inefficiencies.

> *"Gartner continues to find through client interactions, that organizations identify variety of data as their biggest challenge with big data compared to velocity or volume."*
>
> **Gartner Group**

What is Big Data?



Managing and leveraging these new data types and sources is worthwhile because of their business value. However, IT teams are challenged by the newness of the data types, the massive volumes, the wide range of data structures, and the streaming nature of some sources. The problem is further compounded because **most traditional data warehouses were designed for structured or relational data alone.**

In order to preserve their 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 or Microsoft Azure built for new data types and then integrate them with the core data warehouse.

Data Lakes and Data Warehouses Compared

Data Lake	vs	Data Warehouse
Semi-structured / unstructured / structured / raw	**DATA**	Structured data
SQL / Machine Learning / ETL / Graph Analytics etc.	**ANALYTICS FLEXIBILITY**	SQL
Cheap storage for large volumes of data	**VOLUME**	Expensive at large volumes of data
High agility with ability to quickly reconfigure for new workloads	**AGILITY**	Fixed configuration and limited agility
Data Engineers / Data Scientists / Analysts	**USERS**	Analysts / Business Users

Object Storage Data Platforms Underlying the—Data Lake

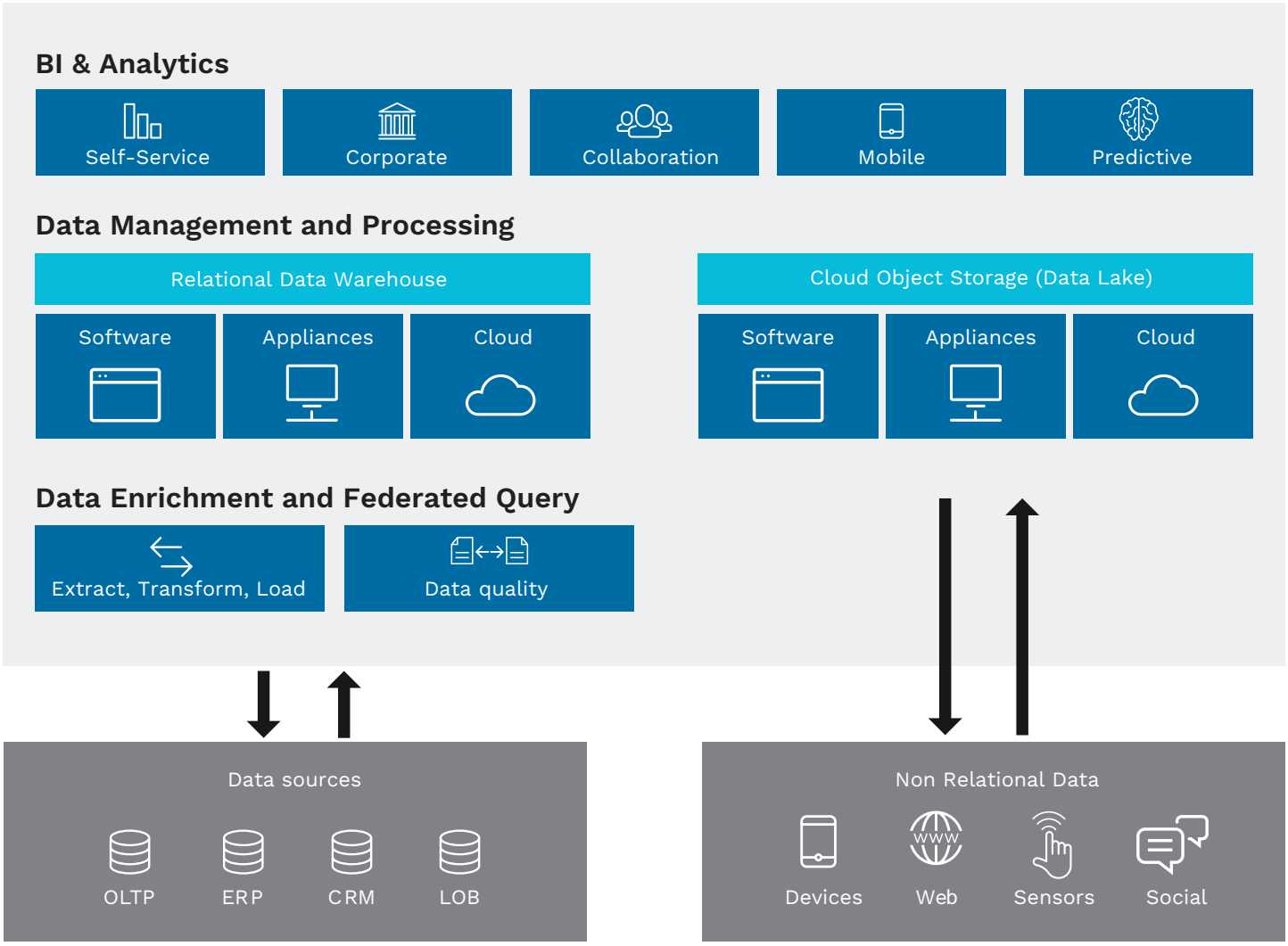
Nothing will have as big a positive impact on your long-term data storage, management and analysis capabilities as ingesting your new big data on Amazon S3 or Microsoft Azure in the cloud. For many companies, this adoption begins with the establishment of the Data Lake - storage repository that holds a vast amount of raw data in its native (as-is) format until it is needed.

These platforms are proliferating for several reasons:

Cost and performance. These object store platforms have the functions required for enterprise use (security, administration, maintenance, high availability, disaster recovery, query, etc.) but the pay per usage is more affordable than comparable enterprise software licenses. Furthermore, they are proven to perform massive analytics and scale linearly.

Data-type diversity. Theoretically, any data you can put in a file can be handled by an object store platform. This empowers user organizations to finally get full business value from structured, unstructured (e.g., social media, image, video, audio) and semi-structured data (e.g., web logs, sensor feeds).

The Modern Data Warehouse Unifies Structured and Unstructured Data



Big Data Object Store Use Cases

The new object store platforms complement and extend your data warehouse without replacing it. This adds years of productive use, new functionality, and greater scale to traditional investments in data warehouses, reporting tools, analytic tools, and data integration tools. Here are some use cases:

Data staging

Data lakes are designed for "early ingestion, later processing" of data. Hence, it works well for data landing, data staging, and the data transformation that usually accompanies such practices.

Source data archiving

It's impossible to foresee all the ways that source data will be repurposed in the future. The current best practice is to retain raw data with all of its original detail. Using up the storage capacity of a traditional data warehouse with all of this source data would be very expensive. A data lake can store and process this data as well, but at a fraction of the cost. In addition, this data archive is online, queryable, and searchable, so users get daily business value from it without time-consuming data-restore processes from tape or optical.

Computational analytics at scale

Valuable computational analytics performed on object stores today include website behavior analysis, sentiment analysis, customer cluster analysis, and statistical modelling or data mining with large volumes of diverse data. Real-time and streaming data analytics are easier in a cloud object store as well.

Analytic Sandboxes for Data Exploration and Discovery

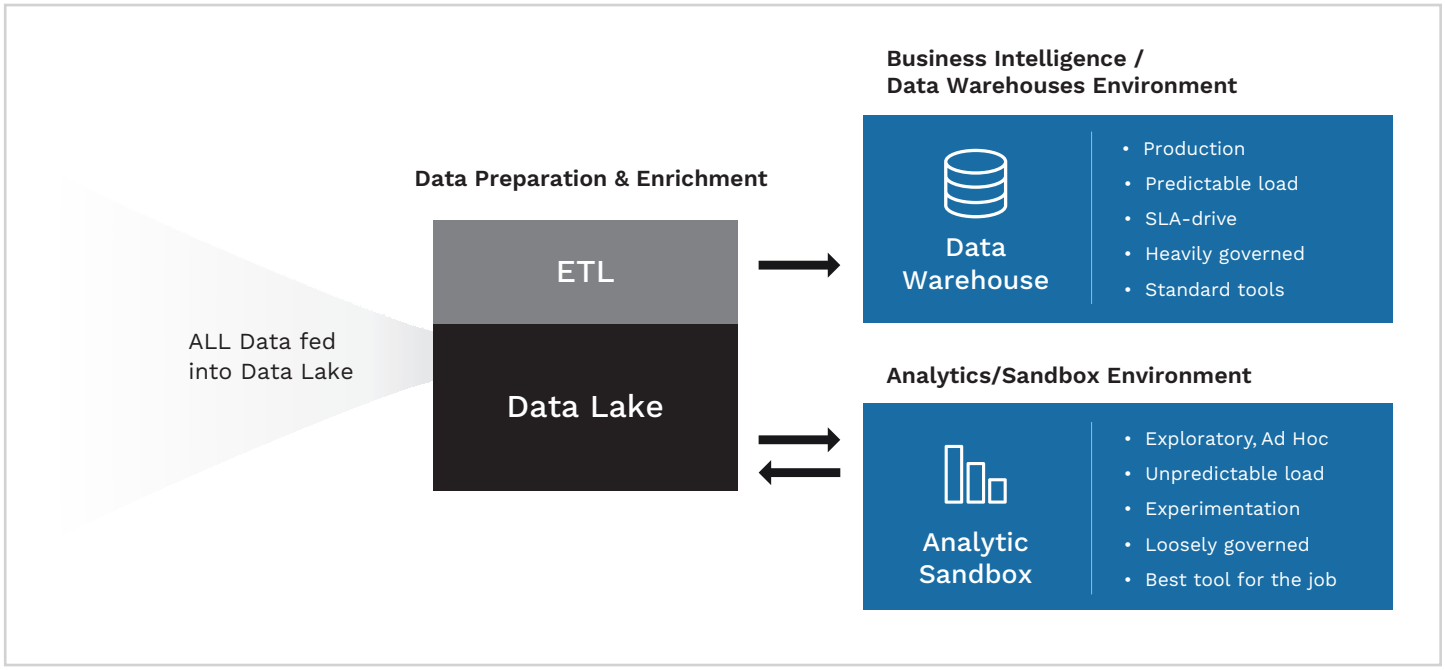
Many analytic methods begin with a data analyst exploring data as a prelude to analysis, reporting, and visualization. Some of these analytic methods for data science work best with full sets of many terabytes or petabytes and require massive computational power. With data samples of this size, your data science team needs to provisions compute environments and desired data sources in order to uncover new customer, product and operational insights.

These sandboxes are exploratory environment with very unpredictable load and usage patterns that would not work on a data warehouse with performance SLAs. The data science team needs to be free to experiment with new data sources, new data transformations and enrichment algorithms, and new analytic models in order to uncover insights buried in the data and build predictive models of an organization's key business process.

III.) Make The Data Lake Central to All Analytics

The last step to a full modernization would be to re-architect the data ingestion and transformation processes to revolve around the data lake. This is the most ambitious approach but ultimately how organizations achieve the highest business value. In this scenario, the data lake becomes the central repository for all the organization's data ingestion (absent the burden of predefining your data schemas). The Data Lake can feed both the production BI/Data Warehouse environment with processed and curated data and the exploratory analytics sandbox as necessary.

Modern Data Lake Environment



One immediate modernization opportunity is off-loading the ETL (extract, transform, and load) routines from the expensive data warehouse to the Data Lake. These existing ETL routines can be dramatically accelerated by taking advantage of the elastic resources of the cloud and access to scale-out big data processing technologies.

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, e-mail, doctors' notes, or claims descriptions to create new metrics that may be better predictors of behavior that would not have been possible before the data lake. These new metrics can then be easily integrated into an organization's existing business intelligence queries, reports, dashboards, and analyses combining structured and unstructured data.

Conclusion

Traditional data warehouses are unable to meet the growing need 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 a modernized data warehouse architecture. Each of the tactics described delivers its own business benefits. Organizations who employ these tactics should see improved CAPex and OPex costs through decreasing data acquisition, maintenance and administrative costs, while improving overall performance, agility and scalability.

Modernizing your data warehouse with the addition of a data lake for new predictive and unstructured data 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 that need to be supported in today's business. It leverages the past investments and supports modern scale in a practical fashion.

For more information:

Contact:
sales@qubole.com

Try QDS for Free:
https://www.qubole.com/products/pricing/

469 El Camino Real, Suite 205
Santa Clara, CA 95050
(855) 423-6674 | info@qubole.com

WWW.QUBOLE.COM