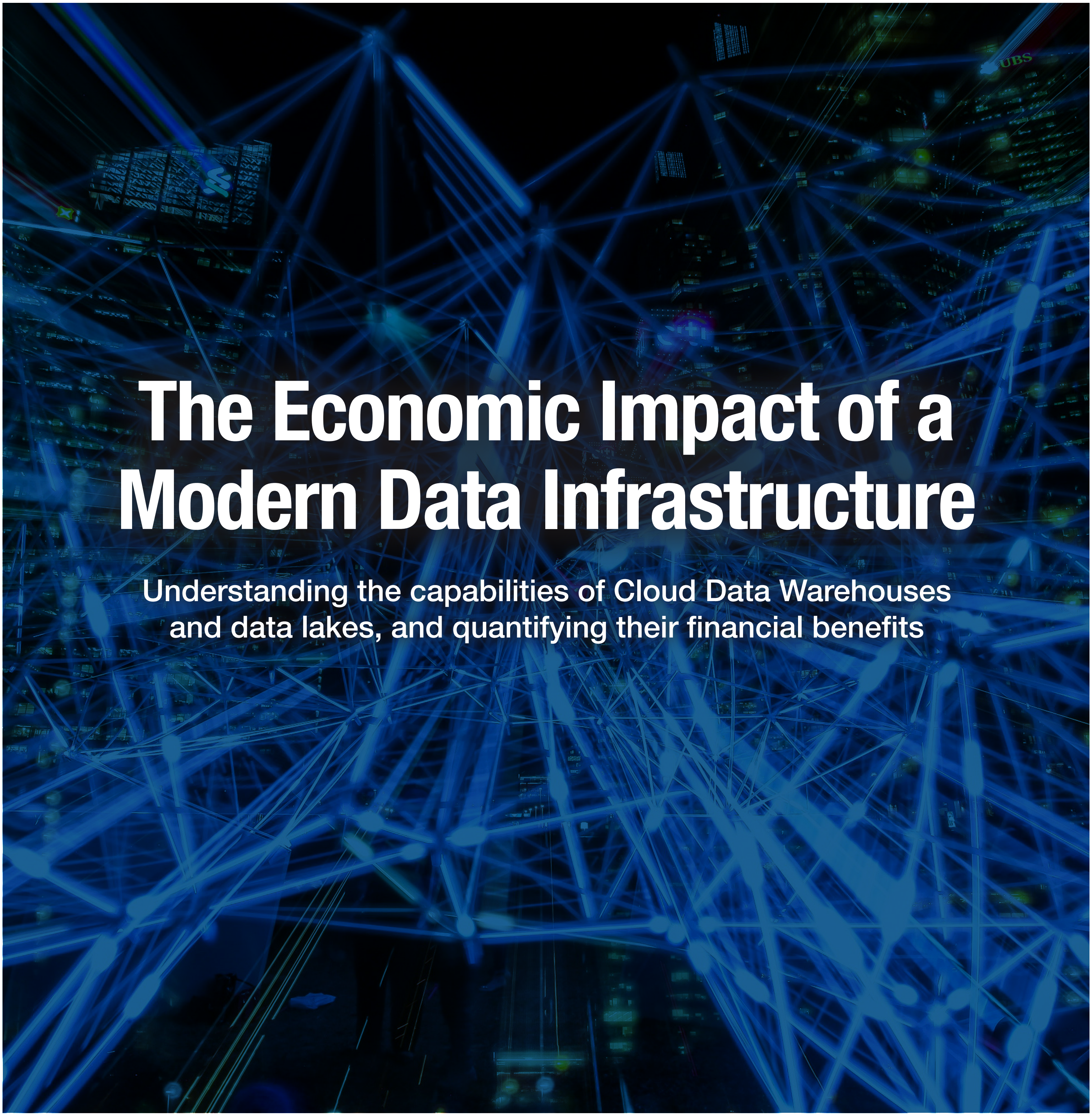


The Economic Impact of a Modern Data Infrastructure

Understanding the capabilities of Cloud Data Warehouses and data lakes, and quantifying their financial benefits

MarTech Review ™

What is a Modern Data Infrastructure?

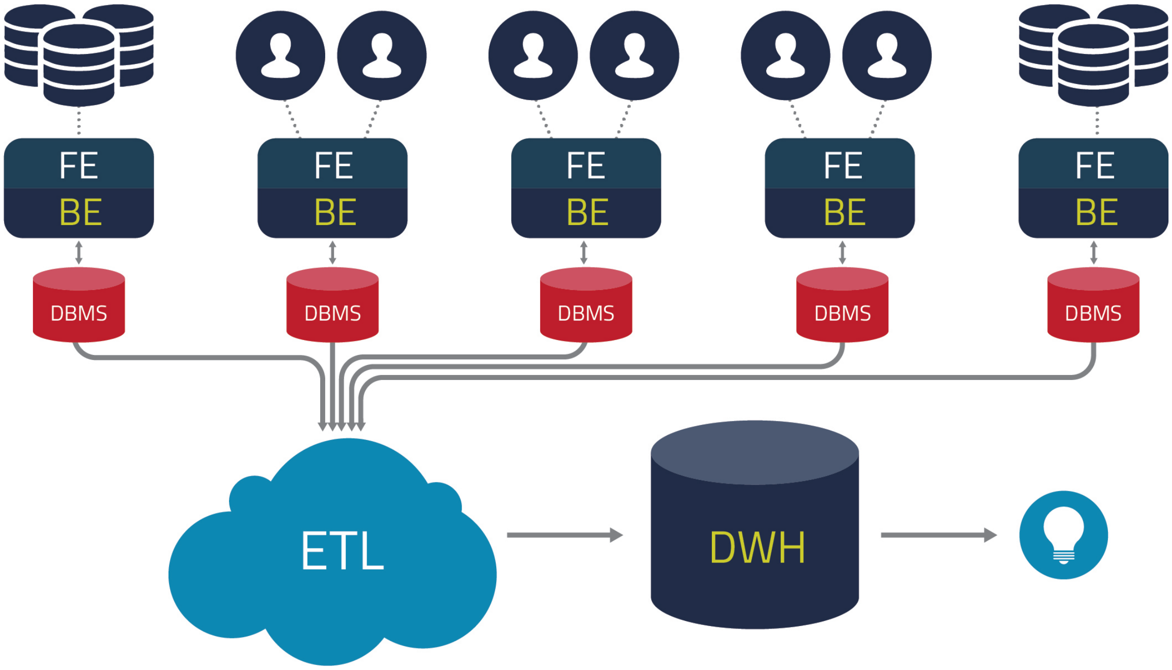
Modern Data Infrastructures are cloud-based, and offer a pay-as-you-go, on-demand, and elastic scalability model that can provide significant benefits for both a business and IT. Compared to an on-premises IT environment, cloud computing reduces up-front project costs and enables organizations to scale their applications as required while paying only for the resources they use. Modern Data Infrastructures typically separate spend on compute and storage, so analytic activity does not need to be constrained. The cloud is an ideal environment for data warehousing and data storage projects, given large data volumes and the unpredictable nature of the analytic workloads involved.

History of Data Infrastructures

Twenty years ago, data were typically stored and analyzed in a highly structured way, within a preconceived schema. SQL, Relational Databases, Data Warehouses, and ETL (Extract, Transform, Load) tools were used widely in the industry to handle data. But then JSON burst onto the scene, and unstructured Web, document, and log data demanded new approaches. New technologies, such as Hadoop / HDFS, NoSQL, and Columnar databases rose to the occasion. Cloud technologies, such as Cloud storage / AWS S3, Cloud Data Warehouses / AWS Redshift, and real-time infrastructure / Kafka / Webhooks, have recently become go-to solutions as the volume of data continues to explode. Leadership is increasingly asking for more analytics based on data, and users are demanding more real-time interactions. This paper will focus primarily on data warehouses and data lake storage.

Traditional On-Premises Data Warehouses

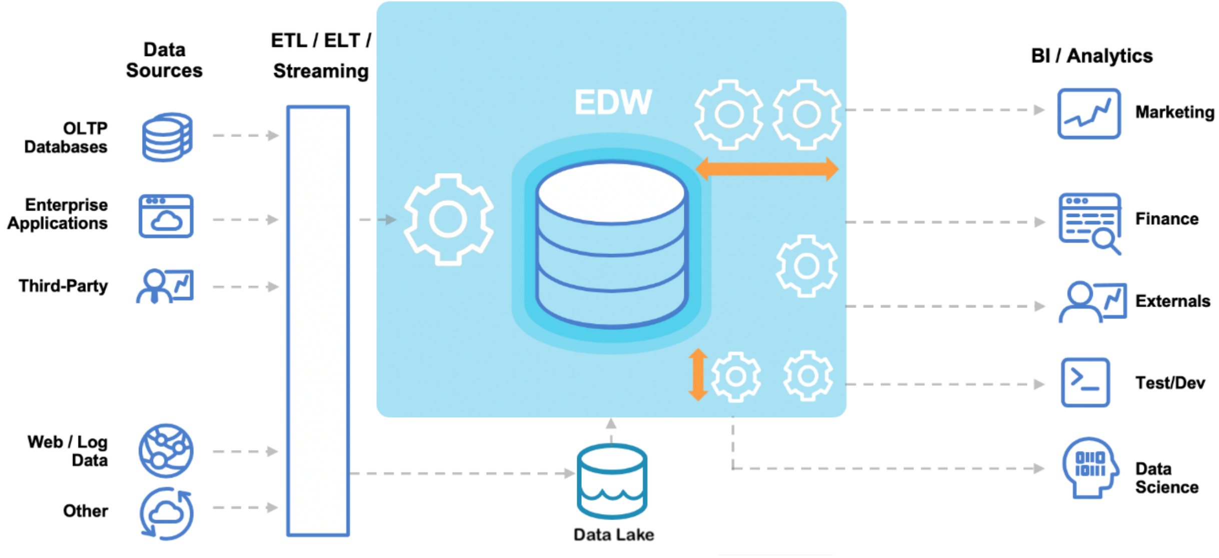
Traditionally, data have been generated separately by both Front End (FE) and Back End (BE) systems. To tackle the load from dual data sources, databases were split based on functional divisions. But this created data integrity problems as different databases reported different figures, and load also increased with ingestion from multiple divisions. This led to the introduction of data warehouses and different data marts to cater to different reporting needs. Functional teams such as Marketing, Sales, Product Development, and Finance, as well as users who held technical, line-of-business, or executive roles could extract actionable insight from data warehouses. Data warehousing technologies, like IBM, Oracle, and Teradata became very prominent.



Modern Cloud Data Warehouses

New, cloud-based data warehouse technology provides a means to use more types of data and data sources. Several Cloud Data Warehouses and broader Cloud Services, like Snowflake and AWS Redshift Spectrum, enable analytics with semi-structured and even unstructured data from data lakes. OLX Group, a global online marketplace, has unlocked an entire petabyte of its data for analytic workstreams with Redshift Spectrum.

Cloud Data Warehouses offer several other advantages beyond on-premises data warehouses, such as elastic scalability, reliable and inexpensive data storage, and much less time and resources required for maintenance. Compute resources do not need to be managed as tightly, as many Cloud Data Warehouses can separate storage and compute, so more business analysts at a company can perform more queries without putting a burden on IT.

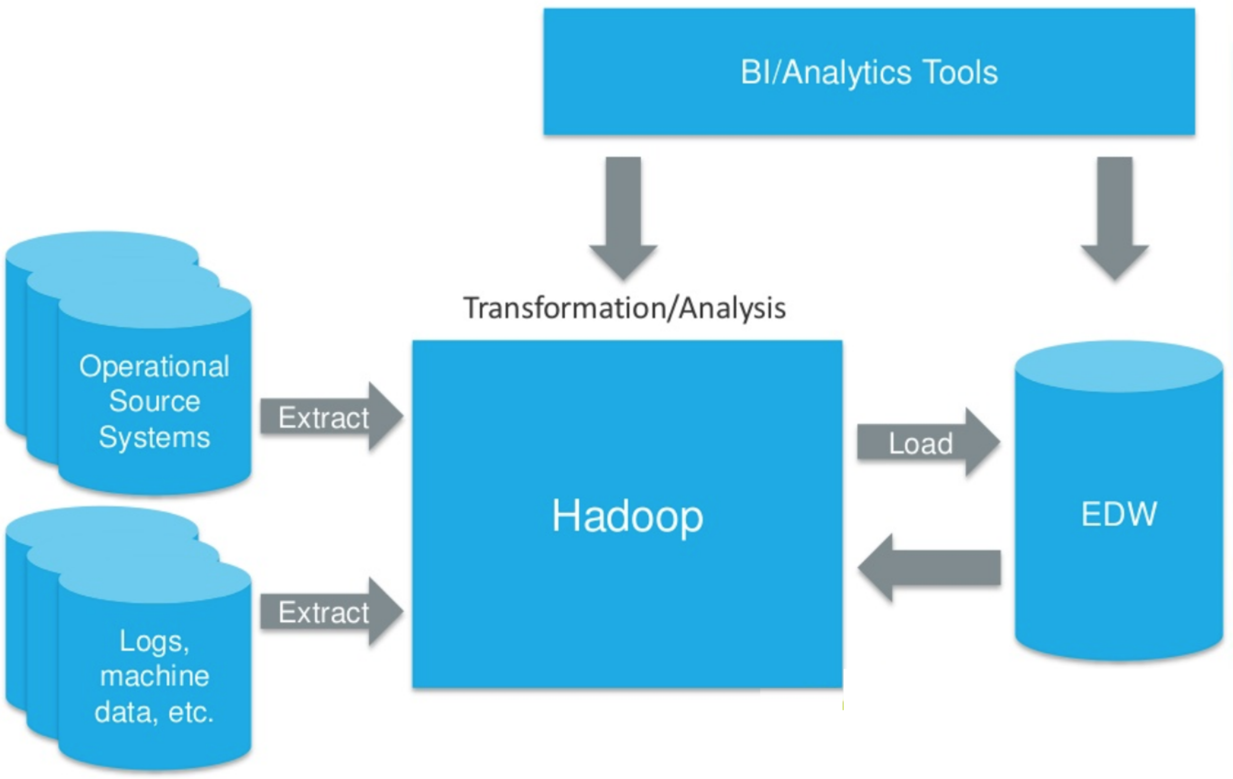


Source: AWS

On-Premises Data Lakes

A data lake is a scalable, centralized repository that can store raw data. Data lakes differ from older data warehouses in that they can store both structured and unstructured data, which you can process and analyze later. Furthermore, a data lake can store any type of data in its native format, ignoring size limits. Data lakes were developed primarily to handle large volumes of data, and thus they excel at handling unstructured data. You typically move all the data into a data lake without transforming it. Each data element in a lake is assigned a unique identifier, and is extensively tagged so that you can later find it via a query. The benefit of this is that you never lose data. It can be available for extensive periods of time, and it's very flexible because it need not adhere to a particular schema before it is stored.

For most of the late 2000s and early 2010s, data lakes were built on HDFS (Hadoop) clusters on premises.



There are numerous challenges to setting up an on-premises data lake:

- Complexity of building data pipelines: When you build your own on-premises infrastructure, you commonly need to manage both the hardware infrastructure - spinning up servers, orchestrating batch ETL jobs, and dealing with outages and downtime - as well as the software side, which requires data engineers to integrate a wide range of tools used to ingest, organize, pre-process and query the data stored in the lake.
- Maintenance Costs: Aside from the upfront investment needed to purchase servers and storage equipment, there are ongoing management and operating costs when operating an on-premises data lake, mostly resulting in IT and engineering costs.
- Scalability: If you want to scale up your data lake to support more users or bigger data, you'll need to manually add and configure servers. You need to keep a close eye on resource utilization, and any additional servers create additional maintenance and operating costs.

Finally, the storage and compute are the same nodes in on-premises data lakes. If you have a 100-node cluster, it stores the data and performs the compute. In the cloud, you can have separate storage and compute services.

The schema-on-read data model, on the other hand, allows you to structure data when you retrieve it from storage. This provides a higher level of flexibility in data analysis and exploration while enabling organizations to easily store massive volumes of data.

Modern Cloud Data Lakes

With new cloud technologies over the last few years - such as AWS S3 and Athena, and Google Cloud Storage - it's becoming increasingly practical and inexpensive to build a cloud-based data lake. Moving your data lake to the cloud has numerous advantages:

- Focus on business value, not infrastructure: Use the cloud to store big data in the cloud and eliminate the need to build and maintain infrastructure, so you can use engineering resources to develop new functionality, which you can connect to business value.
- Lower engineering costs: You can build data pipelines more efficiently with cloud-based tools. The data pipeline is often pre-integrated, so you can get a working solution without investing hundreds of hours in data engineering.
- Agile infrastructure: Cloud services are flexible and offer on-demand infrastructure. If new use cases come up for your data lake, you can re-think, re-engineer and re-architect your data lake more easily.
- Reliability and availability: Cloud providers work to prevent service interruptions, storing redundant copies of data on different servers. Availability spans several data centers. Amazon S3, for example, promises "11 nines" of durability for your data.

The primary downside of moving your data lake to the cloud is storage costs. In the cloud, you pay for storage by the hour. Providers like Amazon offer multiple options for storing your data with variable per-hour costs, so it's possible to optimize, but the fact remains that storage will become an ongoing and growing expense, given expanding data volumes. It will always be more cost effective to buy local storage once and store your data there, although those physical storage costs are dwarfed by ongoing engineering and IT maintenance costs. Some companies in healthcare, finance, and other regulated industries also have data security, compliance and access concerns around storing sensitive data in the cloud.

Consider Hybrid Deployments to Maintain Full Control Over Data

Many organizations managing huge data volumes are exploring hybrid cloud strategies to enable them to keep some storage on premises, while keeping other data that requires more frequent analysis in the cloud.

With an on-premises deployment, enterprises have full control over data security, data access, and data governance. While enterprises are moving to the cloud for its flexibility and scalability, there is still the need for the robust security and control inherent in an on-premises deployment.

Due to these needs, hybrid cloud data lakes have emerged as a logical middle ground between these two consumption models. In a hybrid model, there is more flexibility as to where workloads can run and how much and how fast to scale. Data that needs to be tightly controlled (e.g., customer data) can be stored in an on-premises system, while data that doesn't need to be as tightly controlled can be stored in the cloud. This enables flexibility and cost-effectiveness, while still maintaining proper security controls.

For example, a risk insurance company that has sensitive customer information and transactional data can store that information in an on-premises system. The cloud could then be leveraged for burst-out scenarios, such as processing and adjusting risk policies around a real-time event (e.g. earthquake, flood, or fire), where the data collected does not need to be as tightly controlled. Since an earthquake event can generate gigabytes of data, a company can spin up extra computing nodes, process the data, and spin down the nodes once the processing is complete.

Signs when it's Time to Move to a Modern Data Infrastructure

Some of the major signs that highlight it's time to move to a Modern Data Infrastructure include:

- When you need more expensive hardware or need to set up yet another cluster to keep up with demands.
- When you don't have enough technical staff to keep up with the requirements on your data infrastructure.
- Similarly, when the IT team is severely limiting query access compared to requests from analysts across lines of business.
- When your systems prohibit you from accessing enough data and/or recent enough data.

Technologies to Consider for a Modern Data Infrastructure

All the major cloud vendors, as well as Snowflake, are offering an array of technologies to support modern cloud data lakes and Cloud Data Warehouses. For example:

- AWS supports multiple relevant technologies, including Redshift, S3, and Athena;
- GCP likewise supports BigQuery, Google Cloud Storage, Dataproc, and other technologies;
- Azure supports Azure Synapse Analytics, Azure Data Lake, and other relevant technologies; and
- Snowflake is a unique specialist in this space. Snowflake prefers companies use AWS S3 or other Cloud providers for the raw storage, but it offers a range of Cloud Data Warehouse and data pipeline technologies.

Each of these reputable vendors has its own cost structure, as well as strengths and weaknesses, so it's important to assess the best vendor for your specific requirements.

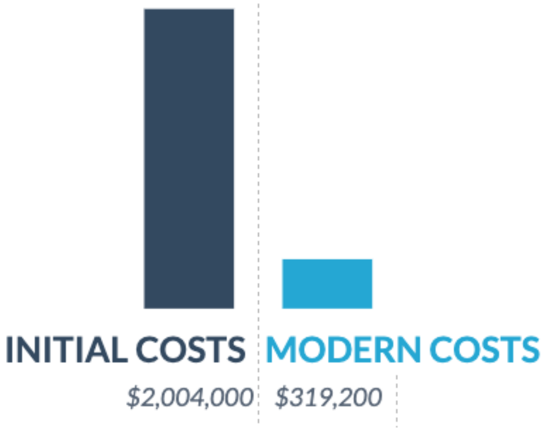
Assessing the Economic Impact of a Modern Data Infrastructure

Modern Data Infrastructures help reduce costs, increase flexibility, and speed up deployment. The right data warehouse will quickly deliver a better ROI by consistently improving the speed, efficiency, and accuracy of data-driven action. It will also reduce the workload on IT and dedicated BI teams, and enable business analysts distributed throughout a company to access important data without fear of bringing down critical systems.

We gathered data from companies with Cloud Data Warehouses to quantify the financial benefits of a Modern Data Infrastructure. Here are some of the major benefits, grouped by the warning signs of an outdated infrastructure above. You can see the detailed model here, and modify it for your own company's circumstances.

Avoid Expensive Hardware or the Need to Set Up Yet Another Cluster to Keep Up With Demands

This group of benefits is typically very impactful, and can add up to well over $1 million in benefits in a single year. It comprises of:



Legacy storage costs saved: High and growing costs for data storage were hindering companies from expanding their data for analytic workstreams until significantly cheaper storage technologies were introduced. Data lake-type approaches and simple data compression is worth over $1 million per year to the average Global 2000 company.

Legacy compute costs saved: With on-premises infrastructures, companies typically over-provision compute resources to ensure they have enough. Cloud Data Warehouses' responsive, elastic provisioning enables new projects and handles seasonal or time of day/week surges in compute usage, which is worth nearly $250,000 to the average customer.

Reduced need for capacity planning: Modern Data Infrastructures eliminate the need for capacity planning by providing scalable infrastructure, which is worth about $200,000 annually in senior DBAs' time.

Free Up Your Technical Staff From Many Data Infrastructure Requirements

This group of benefits is also very impactful, and can add up to about $1 million in benefits in a single year to the average Global 2000 company. It comprises of:

ETL labor cost avoided: When Data Engineering teams no longer need to build pipelines and move data from transactional systems to operations systems to data warehouses, and can leverage modern ELT tools, the average Global 2000 company realizes labor savings around $600,000 per year.

Reduced maintenance costs: A big advantage to a cloud-based solution is that, as a managed solution, tasks like sharding, replication, and scaling are done for you—with many even happening automatically in the background. Managed Cloud Data Warehouses require minimal IT or DBA involvement for upkeep, and you can eliminate backup scripts and many of your previous data retention processes. Also, no IT resources are required to manage the underlying hardware. Reassigning database administrators (DBAs) and hardware support engineers is typically worth over $200,000 per year.

Savings from automatic upgrades: When organizations migrate off their on-premises data infrastructure, they get to skip the cost—and the system downtime—of the next needed upgrade or re-investment in infrastructure. This is worth about $100,000 per year in IT time.

Democratize Query and Data Access Across Your Business

One of the biggest benefits of moving to a Modern Data Infrastructure is **greater employee satisfaction and capabilities**. Capacity limitations, concurrency bottlenecks, and saying “No” to the business analysts because of resource constraints make for frustrated IT staff. Also, as these capacity limits are removed, many more people in a company can access analytic workstreams and data, and add tremendous value to the company.

Modern Data Infrastructures lead to query “democratization;” business analysts no longer need to wait weeks for IT to create a requested query while working through data query backlogs. Also, companies typically see **much faster queries** with a burst-supporting Cloud infrastructure. Kalibri Labs, for example, saw some queries drop down to seconds from 10 hours before.

Gain Access to More Data and More Real-time Data; Move in a More Agile Way With Advanced Analytics Projects

With significantly lower storage costs and simpler access to unstructured data, companies can **drastically expand the data available** with a Modern Data Infrastructure. Equinox Fitness, for example, works with a large amount of data including data on customer visits to clubs, performance data from connected equipment, and digital behavior.

Similarly, companies can remove existing latencies in data ingestion and availability. Photobox had a legacy system that could only ingest data daily from source systems. Analysts could only access data that was at least a day - and sometimes even more than a week - old. They couldn’t make decisions in a timely manner. With a Modern Data Infrastructure, Photobox can now **access data in almost real-time** from users’ browsers and internal systems. While we have yet to model the financial impact of real-time data access, Photobox is realizing many more business insights.

Faster time-to-production: When a company can cut the time to launch new advanced analytics projects, for either the benefit of the organization or for customers, the value is around $720,000 to the average Global 2000 company. This value is highly dependent on actual company investment in new analytics capabilities.

Summary

Of course, there are new costs involved when moving to a Modern Data Infrastructure. The typical cost of using a Cloud Data Warehouse and data lake within a Global 2000 company is around $400,000 per year, assuming about 1 petabyte of data and 8000 hours of compute each month. Additionally, there are implementation, training, and data migration costs, which will run around $500,000 in the first year. Still, the economic benefits far outweigh the costs:

Benefit:	**Annual Savings:**
Avoid Expensive Hardware	$1,680,000
Free Up Technical Staff	$920,000
Democratize Data Access	$600,000
Open up data & Faster time-to-production	$720,000
Total Annual Savings	**$3,920,000**

Moving to a Modern Data Infrastructure will realize a payback period in just 2 months or so on average, and the 3-year return on investment can easily be above 1000%! That does not even include some monetary benefits, like democratizing query access for business analysts.

Resources and Next Steps

You can use this ROI model in Google Sheets (see http://bit.ly/clouddataroi) that can help you persuade others in your company about the financial benefits of a Cloud Data Warehouse and a Modern Data Infrastructure. If you want professional help customizing this financial model to your company, please contact MarTech Review at cloudroi@martechreview.com.