



AHEAD

DATA LAKES VS. DATA WAREHOUSES:

WHAT YOU NEED TO KNOW AS YOU OPTIMIZE STORAGE FOR BIG DATA

The rise of big data is driving organizations across industries to reevaluate their storage strategies. Faced with an array of data storage choices, businesses are examining which option enables them to best leverage massive amounts of data for business insights.

Data lakes and data warehouses are both beneficial — albeit different — storage repositories. Businesses have historically relied on warehouses to manage their data, while data lakes are a newer technology that enables high volumes of diverse file storage.

Which solution is best for your business? We explore your data storage by looking at how to:

1. **STRUCTURE STORAGE WITH DATA WAREHOUSES**
2. **ACCESS DIVERSE STORAGE CAPABILITIES WITH DATA LAKES**
3. **LEVERAGE DATA WAREHOUSES AND DATA LAKES**
4. **CREATE A ROADMAP TOWARD INNOVATIVE STORAGE SOLUTIONS**

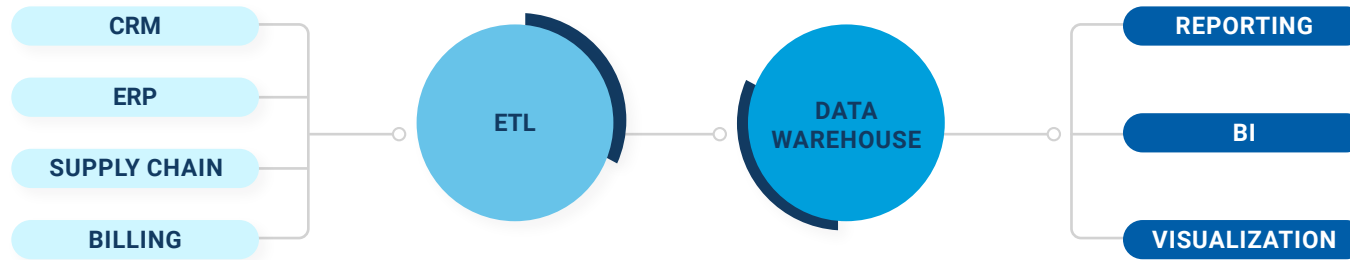
1

STRUCTURE STORAGE WITH DATA WAREHOUSES

A data warehouse is a large collection of raw data in a scalable environment that supports a variety of workloads. Data warehouses require you to purpose-build your solution to deliver maximum end-user value.

The data stored in a warehouse is in a structured format schema — which is preprocessed, formatted, indexed, and designed for performance. It's a non-transactional system that's optimized for reads in a column-oriented format and is queried via SQL. Due to its schema, you'll want to carefully think through your warehouse storage design before leveraging it for insights.

Data warehouse organization



OLTP VS. OLAP

As you consider your data warehouse options, it's important to note the difference between Online Analytical Processing (OLAP) and Online Transactional Processing (OLTP). OLAP is a category of software tools used for data retrieval and analysis in warehouses. Redshift, an OLAP system, is optimized for reads and formatted to read select columns in multiple rows.

Alternatively, OLTP is a transactional system that includes sales orders, product inventory, online messaging, and financial transaction systems. It's typically used by core business functions to administer their organization's various transactions. An example of an OLTP system is MySQL, which simplifies processes but reads or writes one record at a time. A primary difference between OLAP and OLTP is ACID transactions.

In short, OLAP is used for data analysis, and OLTP is used for data processing.

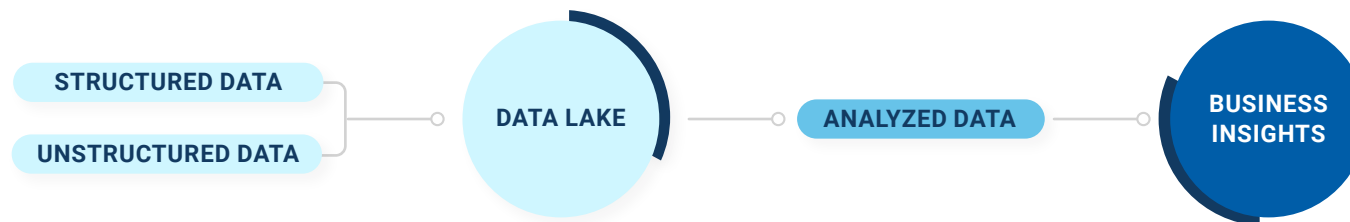
OLAP and OLTP capabilities



ACCESS DIVERSE STORAGE CAPABILITIES WITH DATA LAKES

A data lake is a vast collection of raw data in a scalable environment that supports a variety of workloads. This data is made up of mixed structure data — sometimes unknown schema, flat files, and multiple different formats — along with raw data with unknown value. It's accessed through a variety of languages including SQL, Python, R, and Spark.

Data lake inputs and outputs



Both data lakes and data warehouses are designed to be non-transactional systems for analytics that offer SQL interfaces into the data. Data lakes tend to consist of data that seems to have no value to end users. However, this type of data delivers high value if you correlate it with other data sources.

Data often originates in data lakes. Once it's placed into a pipeline and processed to extract the necessary fields, it's promoted to a data warehouse.

LEVERAGE DATA WAREHOUSES AND DATA LAKES

A closer look at your data storage options

	Data warehouse	Data lake
Schema	Schema-on-white	Schema-on-read
Processing	Highly preprocessed	Totally raw to some processing
Tools	SQL	SQL, Python, Hadoop, R, etc
Value	High value	Unknown, low to high
Ease of use	Simple	Hard
Access patterns	Scheduled jobs and ad hoc analysis	Ad hoc analysis
Scale	Terabytes	Petabytes
Formatting	Optimized, partitioned, indexed	Flat files, mixed formats, text
Documentation	Well-defined data model	Often unknown data model
Users	Data analysis, business users, reporting	Data engineers, data scientists

The ideal data storage strategy uses warehouses and lakes in tandem.

Data lakes allow you to store massive amounts of all data types in a low-cost storage location, like Amazon S3 (and archived in S3 Glacier). Data can be maintained and recovered at a later date, so valuable information is available for migration to data warehouses.

After you determine business-critical assets in your data lake, you're able to efficiently process it in a warehouse for end-user access. By using both data warehouses and lakes, you can sift through vast amounts of data, ensuring valuable information is easily leveraged for insights.

4

CREATE A ROADMAP TOWARD INNOVATIVE STORAGE SOLUTIONS

Without an effective data storage strategy, you lose access to the insights you need to power business success. AHEAD supports your business-critical storage needs and helps you deliver on the promise of digital transformation. The result? You build the foundation you need to rethink your strategy, accelerate migration, and drive innovation.

Ready to improve your data storage strategy? [Let's talk.](#)

AHEAD