Data Lake vs. Data Warehouse - Working Together in the Cloud

Data warehouses and data lakes are two different types of data storage repositories. Find the right cloud solution for your business.
Organizations use data warehouses and data lakes to store, manage and analyze data. Data warehouses have a long history as an enterprise technology used to store structured data, cleaned up and organized for specific business purposes, and serve it to reporting or BI tools. Data lake is a newer technology, made popular by Hadoop and its open source ecosystem. A data lake enables storing both structured and unstructured data in its original form, and processing later when analysis is needed.

In this page we'll define these strategies, explain the differences, and show that "data warehouse vs. data lake" is no longer the question. The two technologies go hand in hand, especially as many move to cloud-native data infrastructure.

What is a data lake?

A data lake is a highly scalable storage system that holds structured and unstructured data in its original form and format. A data lake does not require planning or prior knowledge of the data analysis needed - it assumes that analysis will happen later, on-demand.

What is a data warehouse?

Data warehouse solutions are designed to hold summarized data from many applications and data sources, usually organized by business function. Typical data sources are Online Transaction Processing (OLTP) databases that store transaction data, customer relationship management (CRM), and Enterprise Resources Planning (ERP).

Traditional data warehouses use a process called Extract Transform Load (ETL). Data is meticulously mapped from the original data sources to tables in the data warehouse, and undergoes transformations to achieve a structured format, to enable reporting and BI analysis.

There are several types of data warehouses, including Enterprise Data Warehouse (EDW) which provides decision support for an entire organization, an Operational Data Store (ODS), used for routine activities like transaction recording or employee data reporting, and Data Marts, smaller data warehouses for specific business functions.

Key differences: data warehouse vs. data lake

The following table summarizes the differences between a data warehouse and data lake:

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Image Source

Data types

Data warehouses store structured organizational data such as financial transactions, CRM and ERP data. Other data sources such as social media, web server logs, and sensor data, not to mention documents and rich media, are not stored because they are more difficult to model, and their sheer volume make them expensive and difficult to manage. These types of data are considered more appropriate for a data lake.

Processing

In a data warehouse, data is organized, defined, and metadata is applied before the data is written and stored. This process is called ‘schema on write’.

A data lake consumes everything, including data types considered inappropriate for a data warehouse. Data is stored in raw form; information is saved to the schema as data is pulled from the data source, not when written to storage. This is known as a ‘schema on read’.

Storage and data retention

Before data can be loaded to a data warehouse, data engineers work hard to analyze the data and how to use it for business analysis. They design transformations to summarize and transform the data to enable extraction of relevant insights. Data that doesn’t answer concrete business questions is not included in the data warehouse, in order to reduce storage space and improve performance - a traditional data warehouse is an expensive and scarce enterprise resource.

In a data lake, data retention is less complex, because it retains all data – raw, structured, and unstructured. Data is never deleted, permitting analysis of past, current and future information.

Data lakes can easily be created and scaled to Petabytes. They run on commodity servers using inexpensive storage devices, removing storage limitations.

Agility

Data warehouses store historical data. Incoming data conforms to a predefined structure. This is useful for answering specific business questions, such as “what is our revenue and profitability across all 124 stores over the past week”.

However, if business questions are evolving, or the business wants to retain all data to enable in-depth analysis, data warehouses are insufficient. The development effort to adapt the data warehouse and ETL process to new business questions is a huge burden.

A data lake stores data in its original format, so it is immediately accessible for any type of analysis. Information can be retrieved and reused - a user can apply a formalized schema to the data, store it, and share it with others. If the information is not useful, the copy can be discarded without affecting the data stored in the data lake. All this is done with no development effort.

Security, maturity and usage

Data warehouses have been around for two decades and are a secure, enterprise-ready technology. Data lakes are getting there, but are newer and have a shorter enterprise track record. A large enterprise cannot buy and implement a data lake like it would a data warehouse - it must consider which tools to use, open source or commercial, and how to piece them together to meet requirements.

The end users of each technology are different: a data warehouse is used by business analysts, who query the data via pre-integrated reporting and BI. Business users cannot use a data lake as easily, because data requires processing and analysis to be useful. Data scientists, data engineers, or sophisticated business users, can extract insights from massive volumes of data in the data lake.

Cloud data warehousing and data lakes in the cloud

There are many robust data warehouse tools offered today on cloud-based infrastructure, including:

- **Amazon Redshift** - a fully-managed, analytical data warehouse that can handle petabyte-scale data, and enable querying it in seconds.
- **Google BigQuery** - an enterprise-grade cloud-native data warehouse, which runs fast interactive and ad-hoc queries on datasets of petabyte-scale.
- **Panoply** - the world's first smart data warehouse, which is cloud-based, scalable and performant, and also able to automatically transform data to analytics in minutes.

There are also a number of options for running data lakes in the cloud, including:

- **Amazon S3** - an object storage platform built to store and retrieve any amount of data from any data source, and designed for 99.999999999% durability.
- **Azure Blob Storage** - stores billions of objects in hot, cool, or archive tiers, depending on how often data is accessed. Data ranges from structured (converted to object form) to any unstructured format - images, videos, audio, documents.

In the cloud, data warehouse and data lake strategies go hand in hand

Here are two examples of how cloud-based infrastructure enables data warehouses and data lakes to play together. This allows you to enjoy the unlimited low-cost storage and flexibility of a data lake, together with the high performance and analytical capabilities of a data warehouse.

Amazon Redshift Spectrum

This solution from Amazon extends the analytic capabilities of Redshift beyond the data stored on its local disks. It can directly query unstructured data in an Amazon S3 data lake, data warehouse style, without having to load or transform it. Redshift Spectrum optimizes queries on the fly, and scales up processing transparently to return results quickly, regardless of the scale of data being processed.

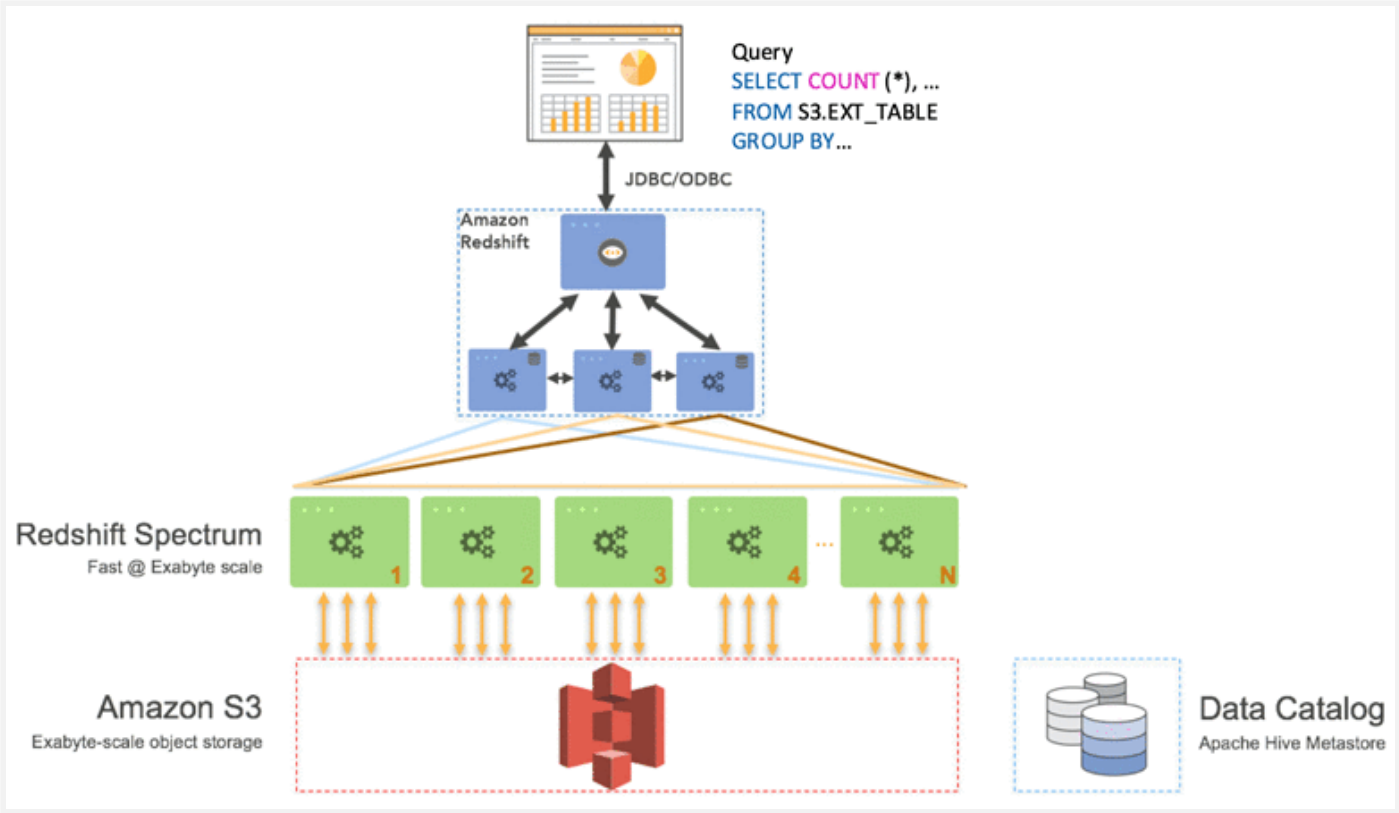


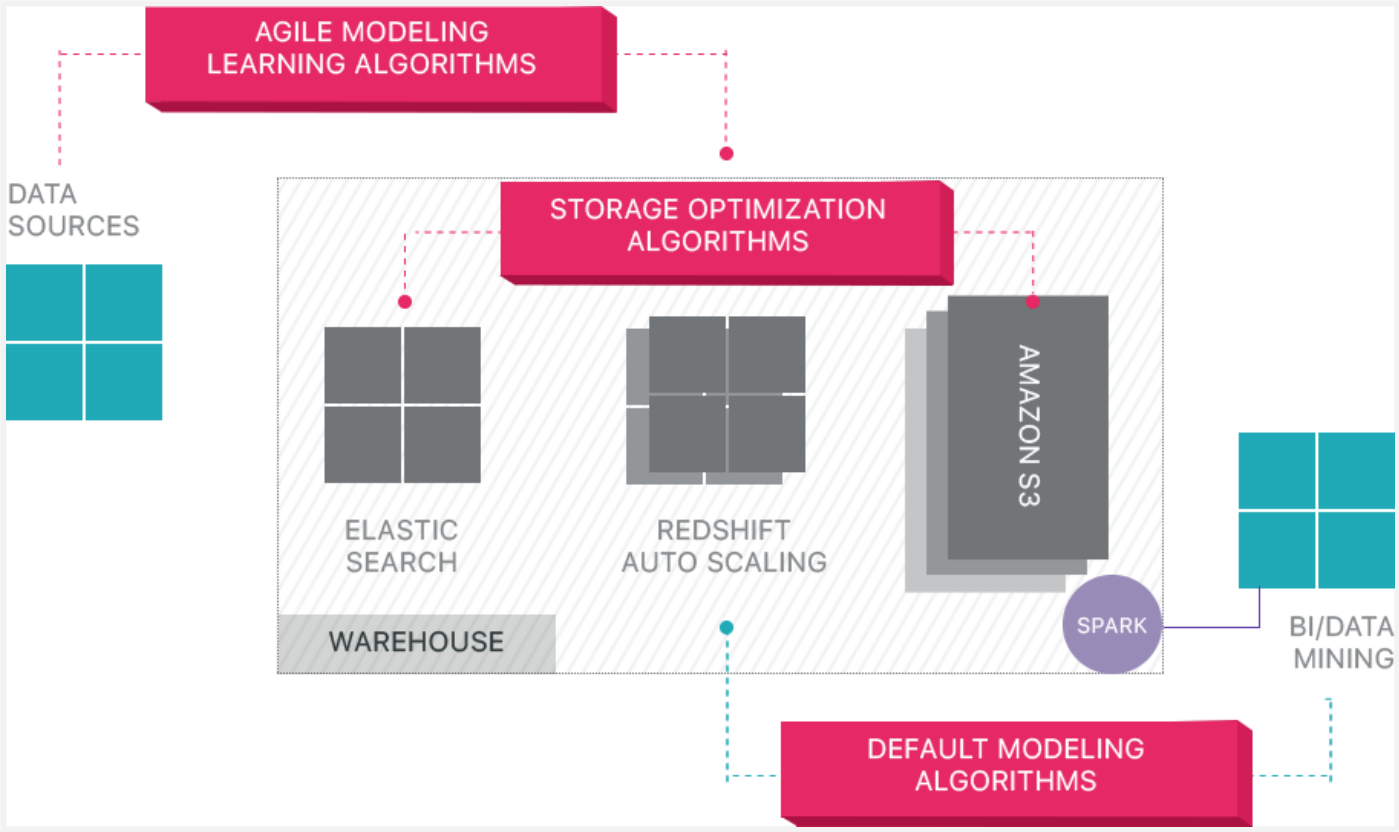
Image Source

Panoply - automated data preparation

Panoply is a cloud-based data warehouse which integrates with S3 data lakes and many other data sources. Panoply is a pioneer of data warehouse automation, offering a self-optimizing

architecture, which uses machine learning and natural language processing (NLP) to model the data journey from source to analysis.

Panoply allows you to pull large volumes of data from a cloud-based data lake like S3, without having an ETL process in place. Once the data is in Panoply it is automatically treated, prepared, and optimized for fast analysis - you can immediately start running analytical queries.



Which strategy is best for your data?

As organizations move data infrastructure to the cloud, the choice of data warehouse vs. data lake, or the need for complex integrations between the two, is less of an issue. It is becoming natural for organizations to have both, and move data flexibly from lakes to warehouses to enable business analysis.

In this article, we defined the two data storage paradigms - data lake vs. data warehouse - and provided two examples of cloud-based solutions that allow almost effortless integration

between data warehouses and data lakes - Amazon Redshift Spectrum and Panoply's automated data warehouse.

In the cloud - and only in the cloud - you can connect a data lake to a data warehouse and start analyzing data in minutes, without laborious data preparation and complex ETL processes.

Built by Panoply