DATA STORAGE EXPLAINED: DATA LAKE VS WAREHOUSE VS DATABASE

Data storage is a big deal. Data companies are in the news a lot lately, especially as companies attempt to maximize value from big data's potential. For the lay person, data storage is usually handled in a traditional database. But for big data, companies use data warehouses and data lakes.

Data lakes are often compared to data warehouses—but they shouldn't be. Data lakes and data warehouses are very different, from the structure and processing all the way to who uses them and why. In this article, we'll:

- Define databases, warehouses, and lakes
- Summarize the big differences
- Caution the use of data lakes
- Explore the future of data storage
- And more

Defining database, warehouse, and lake

Let's start with the concepts, and we'll use an expert analogy to draw out the differences.

What’s a database?

A database is a storage location that houses structured data. We usually think of a database on a computer—holding data, easily accessible in a number of ways. Arguably, you could consider your smartphone a database on its own, thanks to all the data it stores about you.

For all organizations, the use cases for databases include:

- Creating reports for financial and other data
- Analyzing relatively small datasets
- Automating business processes
- Auditing data entry

Popular databases are:

- Oracle
- PostgreSQL
- MongoDB
- Redis
- Elasticsearch
- Apache Cassandra

(Learn more about the key difference in databases: SQL vs NoSQL.)

What’s a data warehouse?

The next step up from a database is a data warehouse. Data warehouses are large storage locations for data that you accumulate from a wide range of sources. For decades, the foundation for business intelligence and data discovery/storage rested on data warehouses. Their specific, static structures dictate what data analysis you could perform.

Data warehouses are popular with mid- and large-size businesses as a way of sharing data and content across the team- or department-siloed databases. Data warehouses help organizations become more efficient. Organizations that use data warehouses often do so to guide management decisions—all those “data-driven” decisions you always hear about.

Popular companies that offer data warehouses include:

- Snowflake
- Yellowbrick
- Teradata

What’s a data lake?

A data lake is a large storage repository that holds a huge amount of raw data in its original format until you need it. Data lakes exploit the biggest limitation of data warehouses: their ability to be more flexible.

As we’ll see below, the use cases for data lakes are generally limited to data science research and testing—so the primary users of data lakes are data scientists and engineers. For a company that actually builds data warehouses, for instance, the data lake is a place to dump and temporarily store all the data until the data warehouse is up and running. Small and medium sized organizations likely have little to no reason to use a data lake.

Popular data lake companies are:

- Hadoop
- Azure
- Amazon S3

Illustrating the differences

Lee Easton, president of data-as-a-service provider AeroVision.io, recommends a tool analogy for understanding the differences. In this, your **data** are the tools you can use.

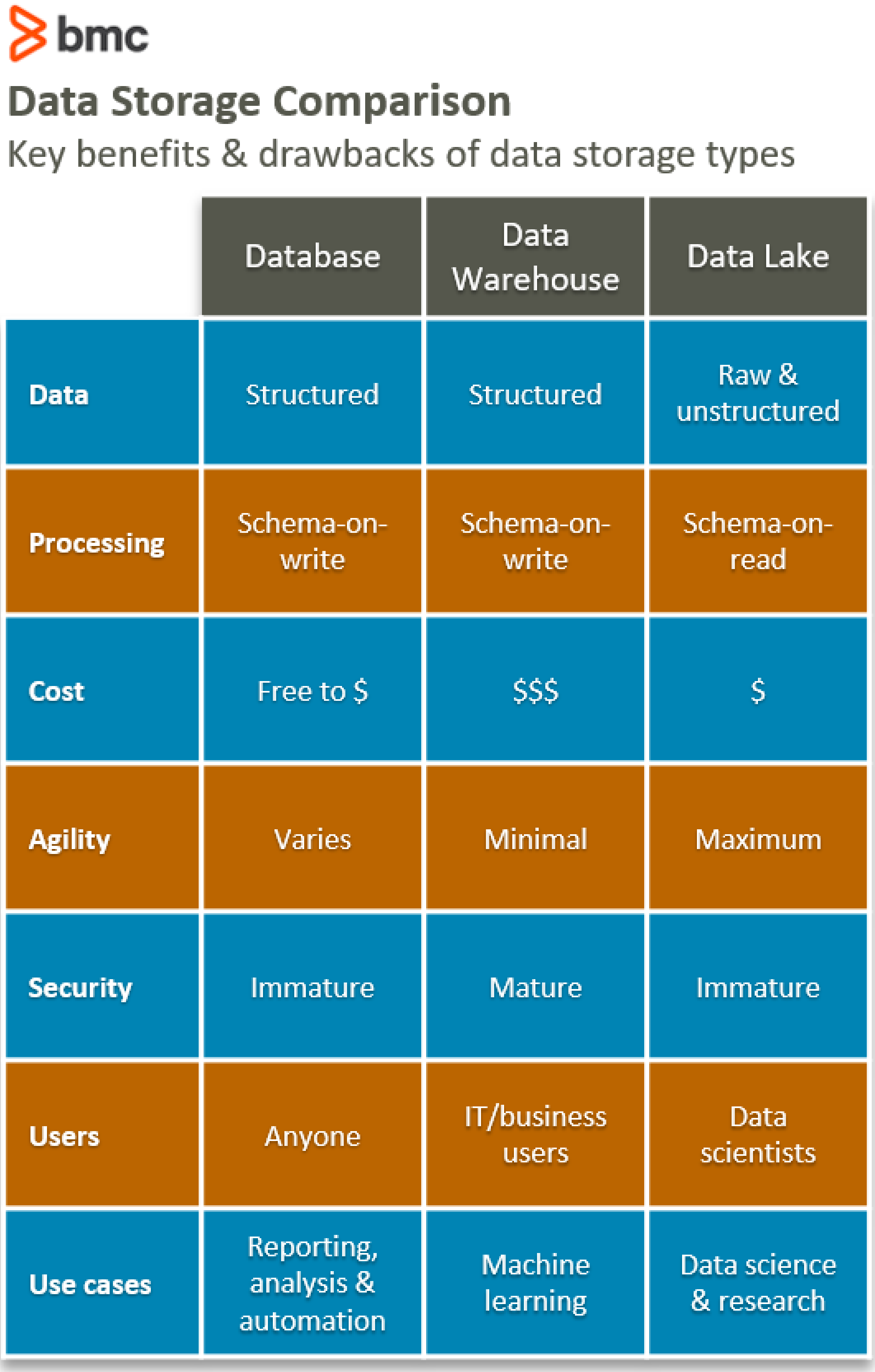
Imagine a tool shed in your backyard. You store some tools—data—in a toolbox or on (fairly) organized shelves. This specific, accessible, organized tool storage is your **database**. The tool shed, where all this is stored, is your **data warehouse**. You might have lots (and lots!) of toolboxes in the shop. Some toolboxes might be yours, but you could store toolboxes of your friends or neighbors, as long as your shed is big enough. Though you're storing their tools, your neighbors still keep them organized in their own toolboxes.

But what if your friends aren't using toolboxes to store all their tools? They've just dumped them in there, unorganized, unclear even what some tools are for—this is your **data lake**.

In a data lake, the data is raw and unorganized, likely unstructured. Any raw data from the data lake that hasn't been organized into shelves (databases) or an organized system (data warehouses) is barely even a tool—in raw form, that data isn't useful.

Comparing data storage

Now that we've got the concepts down, let's look at the differences across databases, warehouses, and data lakes in six key areas.



bmc

Data Storage Comparison

Key benefits & drawbacks of data storage types

	Database	Data Warehouse	Data Lake
Data	Structured	Structured	Raw & unstructured
Processing	Schema-on-write	Schema-on-write	Schema-on-read
Cost	Free to $	$$$	$
Agility	Varies	Minimal	Maximum
Security	Immature	Mature	Immature
Users	Anyone	IT/business users	Data scientists
Use cases	Reporting, analysis & automation	Machine learning	Data science & research

Data

Database and data warehouses can only store data that has been structured. A data lake, on the other hand, does not respect data like a data warehouse and a database. It stores all types of data: structured, semi-structured, or unstructured.

All three data storage locations can handle hot and cold data, but cold data is usually best suited in data lakes, where the latency isn't an issue. (More on latency below.)

Processing

Before data can be loaded into a data warehouse, it must have some shape and structure—in other words, a model. The process of giving data some shape and structure is called schema-on-write. A database also uses the schema-on-write approach.

A data lake, on the other hand, accepts data in its raw form. When you do need to use data, you have to give it shape and structure. This is called schema-on-read, a very different way of processing data.

Cost

One of most attractive features of big data technologies is the cost of storing data. Storing data with big data technologies is relatively cheaper than storing data in a data warehouse. This is because data technologies are often open source, so the licensing and community support is free. The data technologies are designed to be installed on low-cost commodity hardware.

Storing a data warehouse can be costly, especially if the volume of data is large. A data lake, on the other hand, is designed for low-cost storage. A database has flexible storage costs which can either be high or low depending on the needs.

Agility

A data warehouse is a highly structured data bank, with a fixed configuration and little agility. Changing the structure isn't too difficult, at least technically, but doing so is time consuming when you account for all the business processes that are already tied to the warehouse.

Likewise, databases are less agile to configure because of their structured nature.

Conversely, a data lake lacks structure. This agility makes it easy for data developers and data scientists to easily configure and reconfigure data models, queries, and applications. (That explains why data experts primarily—not lay employees—are working in data lakes: for research and testing. The lack of structure keeps non-experts away.)

Security

Data warehouse technologies, unlike big data technologies, have been around and in use for decades. Data warehouses are much more mature and secure than data lakes.

Big data technologies, which incorporate data lakes, are relatively new. Because of this, the ability to secure data in a data lake is immature. Surprisingly, databases are often less secure than warehouses. That's likely due to how databases developed for small sets of data—not the big data use cases we see today. Luckily, data security is maturing rapidly.

Users

Data warehouses, data lakes, and databases are suited for different users:

- Databases are very flexible and thus suited for any user.

- Data warehouses are used mostly in the business industry by business professionals.
- Data lakes are mostly used in scientific fields by data scientists.

Caution on data lakes

Companies are adopting data lakes, sometimes instead of data warehouses. But data lakes are not free of drawbacks and shortcomings. New technology often comes with challenges—some predictable, others not. Data lakes are no different. It isn't that data lakes are prone to errors. Instead, companies venturing into data lakes should do so with caution.

Data lakes won't solve all your data problems. In fact, they may add fuel to the fire, creating more problems than they were meant to solve. That's because data lakes tend to overlook data best practices.

- **Data lakes allow you to store anything without questioning whether you need all the data.** This approach is faulty because it makes it difficult for a data lake user to get value from the data.
- **Data lakes do not prioritize which data is going into a supply chain and how that data is beneficial.** This lack of data prioritization increases the cost of data lakes (versus data warehouses and databases) and muddies any clarity around what data is required. This slows, perhaps halts, your entire analytical process. Avoid this issue by summarizing and acting upon data *before* storing it in data lakes.
- **Data latency is higher in data lakes.** Data lakes are often used for reporting and analytics; any lag in obtaining data will affect your analysis. Latency in data slows interactive responses, and by extension, the clock speed of your organization. Your reason for that data, and the speed to access it, should determine whether data is better stored in a data warehouse or database.
- **Data lakes do not have rules overseeing what they can take in, increasing your organizational risk.** The fact that you can store all your data, regardless of the data's origins, exposes you to a host of regulatory risks. Multiply this across all users of the data lake within your organization. The lack of data prioritization further compounds your compliance risk.
- **Data lakes foster data overindulgence.** Too much unprioritized data creates complexity, which means more costs and confusion for your company—and likely little value. Organizations should not strive for data lakes on their own; instead, data lakes should be used only within an encompassing data strategy that aligns with actionable solutions.

Data is only valuable if it can be utilized to help make decisions in a timely manner. A user or a company planning to analyze data stored in a data lake will spend a lot of time finding it and preparing it for analytics—the exact opposite of data efficiency for data-driven operations.

Instead, you should always view data from a supply chain perspective: beginning, middle, and end. No matter the data, you should always plan a strategy for how you will:

- Find the data
- Bring data into organizational data storage
- Explore and transform the data

Such an approach allows optimization of value to be extracted from data.

The future is with data warehouses

If data warehouses have been neglected for data lakes, they might be making a comeback. That's for two main reasons, according to Mark Cusack, CTO of Yellowbrick:

- Data warehouse companies are improving the consumer cloud experience, making it easiest to try, buy, and expand your warehouse with little to no administrative overhead.
- Data warehousing will become crucial in machine learning and AI. That's because ML's potential relies on up-to-the-minute data, so that data is best stored in warehouses—not lakes.

When developing machine learning models, you'll spend approximately 80% of that time just preparing the data. Warehouses have built-in transformation capabilities, making this data preparation easy and quick to execute, especially at big data scale. And these warehouses can reuse features and functions across analytics projects, which means you can overlay a schema across different features. This reduces duplication and increases your data quality.

As companies embrace machine learning and data science, data warehouses will become the most valuable tool in your data tool shed.

BMC for data solutions

BMC's award-winning Control-M is an industry standard for enterprise automation and orchestration. And our brand-new SaaS solution BMC Helix Control-M gives you the same organization, control, and orchestration—in the cloud.

Additional resources

For more on this topic, explore these resources:

- BMC Machine Learning & Big Data Blog
- 3 Keys to Building Resilient Data Pipelines
- Enabling the Citizen Data Scientist
- Dark Data: The Basics and The Challenges