ORACLE

What's the difference between a data lake, database, and a data warehouse?

There are so many buzzwords these days regarding data management. Data lakes, data warehouses, and databases – what are they? In this article, we'll walk through them and cover the definitions, the key differences, and what we see for the future.

Data lake

If you want full, in-depth information, you can read our article called, "What's a Data Lake?" But here we can tell you, "A data lake is a place to store your structured and unstructured data, as well as a method for organizing large volumes of highly diverse data from diverse sources."

The data lake tends to ingest data very quickly and prepare it later, on the fly, as people access it.

Data warehouse

A data warehouse collects data from various sources, whether internal or external, and optimizes the data for retrieval for business purposes. The data is usually structured, often from relational databases, but it can be unstructured too.

Primarily, the data warehouse is designed to gather business insights and allows businesses to integrate their data, manage it, and analyze it at many levels.

Database

Essentially, a database is an organized collection of data. Databases are classified by the way they store this data. Early databases were flat and limited to simple rows and columns. Today, the popular databases are:

- **Relational databases**, which store their data in tables
- **Object-oriented databases**, which store their data in object classes and subclasses

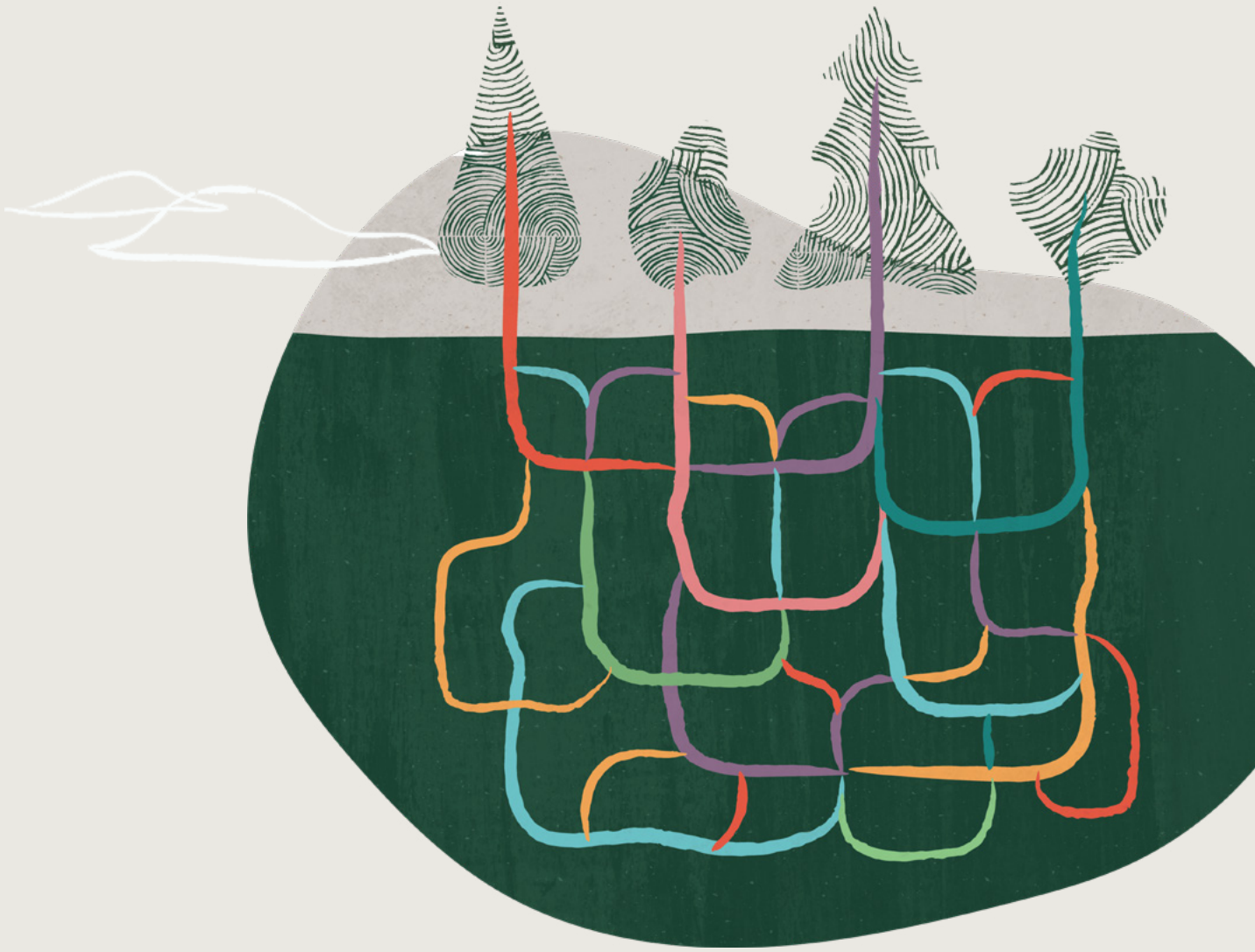
Data mart, data swamp and other terms

And, of course, there are other terms such as data mart and data swamp, which we'll cover very quickly so you can sound like a data expert.

Enterprise Data Warehouse (EDW): This is a data warehouse that serves the entire enterprise.

Data mart: A data mart is used by individual departments or groups and is intentionally limited in scope because it looks at what users need right now versus the data that already exists.

Data swamp: When your data lake gets messy and is unmanageable, it becomes a data swamp.



The differences between data lakes, data warehouses, and database

Data lakes, data warehouses and databases are all designed to store data. So why are there different ways to store data, and what's significant about them? In this section, we'll cover the significant differences, with each definition building on the last.

The database

Databases came about first, rising in the 1950s with the relational database becoming popular in the 1980s.

Databases are usually set up to monitor and update real-time structured data, and they have have only the most recent data available.

The data warehouse

But the data warehouse is a model to support the flow of data from operational systems to decision systems. What this means, essentially, is that businesses were finding that their data was coming in from multiple places—and they needed a different place to analyze it all. Hence the growth of the data warehouse.

For example, let's say you have a rewards card with a grocery chain. The database might hold your most recent purchases, with a goal to analyze current shopper trends. The data warehouse might hold a record of all of the items you've ever bought and it would be optimized so that data scientists could more easily analyze all of that data.

The data lake

Now let's throw the data lake into the mix. And because it's the newest, we'll talk about this one more in depth. The data lake really started to rise around the 2000s, as a way to store unstructured data in a more cost-effective way. The key phrase here is cost effective.

Although databases and data warehouses can handle unstructured data, they don't do so in the most efficient manner. With so much data out there, it can get expensive to store all of your data in a database or a data warehouse.

In addition, there's the time-and-effort constraint. Data that goes into databases and data warehouses needs to be cleansed and prepared before it gets stored. And with today's unstructured data, that can be a long and arduous process when you're not even completely sure that the data is going to be used.

That's why data lakes have risen to the forefront. The data lake is mainly designed to handle unstructured data in the most cost-effective manner possible. As a reminder, unstructured data can be anything from text to social media data to machine data such as log files and sensor data from IoT devices.

Data lake example

Going back to the grocery example that we used with the data warehouse, you might consider adding a data lake into the mix when you want a way to store your big data. Think about the social sentiment you're collecting, or advertising results. Anything that is unstructured but still valuable can be stored in a data lake and work with both your data warehouse and your database.

Note 1: Having a data lake doesn't mean you can just load your data however you want. That's what leads to a data swamp. But it does make the process easier, and new technologies such as having a data catalog will steadily make it simpler to find and use the data in your data lake.

Note 2: If you want more information on the ideal data lake architecture, you can read the full article we wrote on the topic. It describes why you want your data lake built on object storage and Apache Spark, versus Hadoop.

What's the future of data lakes, data warehouses, and databases?

Will one of these technologies rise to overtake the others?

We don't think so.

Here's what we see. As the value and amount of unstructured data rises, the data lake will become increasingly popular. But there will always be an essential place for databases and data warehouses.

You'll probably continue to keep your structured data in the database or data warehouse. But these days, more companies are moving their unstructured data to data lakes on the cloud, where it's more cost effective to store it and easier to move it when it's needed.

This workload that involves the database, data warehouse, and data lake in different ways is one that works, and works well. We'll continue to see more of this for the foreseeable future.

To learn more, find out about the Modern Data Warehouse, which combines a data warehouse and data lake with analytics and data science capabilities so organizations can start using their data most effectively and gain real results.

Oracle Corporation

Worldwide Headquarters
500 Oracle Parkway, Redwood Shores, CA 94065, USA

Worldwide Inquiries
Tele + 1.650.506.7000 + 1.800.ORACLE1
Fax + 1.650.506.7200

oracle.com

Connect with us

Call +1.800.ORACLE1 or visit oracle.com. Outside North America, find your local office at oracle.com/contact.

facebook.com/oracle

youtube.com/oracle

linkedin.com/company/oracle

twitter.com/oracle