What is Data Lake?

Data lake defined

Here's a simple definition: 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.

Data lakes are becoming increasingly important as people, especially in business and technology, want to perform broad data exploration and discovery. Bringing data together into a single place or most of it in a single place makes that simpler.

Depending on your platform, the data lake can make that much easier. It can handle many data structures, such as unstructured and multistructured data, and it can help you get value out of your data.

Data lake versus data warehouse

The key difference between a data lake and a data warehouse is that the data lake tends to ingest data very quickly and prepare it later on the fly as people access it. With a data warehouse, on the other hand, you prepare the data very carefully upfront before you ever let it in the data warehouse.

Users tend to want to ingest data into the data lake as quickly as possible, so that companies with operational use cases, especially around operational reporting, analytics, and business monitoring, have the newest data. This enables them to have access to the latest data and see the most updated information.

With the data lake, users often ingest data in the original form without altering it. This can be for speed reasons, but can also be for other reasons including the desire to perform advanced analytics which can depend upon detailed source data. This would be analytics based on any kind of mining, whether it's:

- Text mining
- Data mining
- Statistical analysis
- Anything involving clusters
- Graph analytics

Data lake use cases

To provide all the advantages that data lakes can offer, a proper solution should be able to offer better ways to:

- **Ingest and transform:** Move and convert different kinds and formats of data
- **Persist and access:** Ensure data is secure, can be readily discovered, can easily scale as needed, and be accessed as needed across products
- **Analyze and use data science:** Uncover insights and trends within data

A data lake is more useful when it is part of a greater data management platform, and it should integrate well with existing data and tools for a more powerful data lake.

Omnichannel marketing data lake

Using the data lake to extend the data warehouse is something often seen with omnichannel marketing, sometimes called multichannel marketing. The way to think about the data ecosystem in marketing is that every channel can be its own database, and every touchpoint can be as well. And then many marketers also buy data from third parties.

For example, a marketer might want to buy data that has additional demographic and consumer preference information about customers and prospects, and that helps the marketer fill out that complete view of each customer, which in turn helps with creating more personalized and targeted marketing campaigns.

That’s a complex data ecosystem, and it’s getting bigger in volume and greater in complexity all the time. The data lake is brought in quite often to capture data that's coming in from multiple channels and touchpoints. And some of those actually are streaming data.

Companies that offer a smartphone app to its customers may be receiving that data in real time or close to it, as customers use that app. Many times, the company doesn’t really need full real time. It could be an hour or two old. But it allows the marketing department to do very granular monitoring of the business and create specials, incentives, discounts, and micro-campaigns.

Digital supply chain data lake

The digital supply chain is an equally diverse data environment and the data lake can help with that, especially when the data lake is on Hadoop. Hadoop is largely a file-based system because it was originally designed for very large and highly numerous log files that come from web servers. In the supply chain there is often a large quantity of file-based data. Think about file-based and document-based data from EDI systems, XML, and of course today JSONs coming on very strong in the digital supply chain. That's very diverse information.

There is also internal information to consider. Manufacturers often have data from the shop floor and from shipping and billing that's highly relevant to the supply chain. The lake can help manufacturers bring that data together and manage it in a file-based kind of way.

The Internet of Things data lake

The Internet of Things is creating new data sources almost daily in some companies. And of course, as those sources diversify they create even more data. Increasingly, there are more sensors on more machinery all the time. As an example, every rail freight or truck freight vehicle like that has a huge list of sensors so the company can track that vehicle through space and time, in addition to how it’s operated. Is it operated safely? Is it operated in an optimal way relative to fuel consumption? Enormous amounts of information are coming from these places, and the data lake is very popular because it provides a repository for all of that data.

A single data lake

Now, those are examples of fairly targeted uses of the data lake in certain departments or IT programs, but a different approach is for centralized IT to provide a single large data lake that is multitenant. It can be used by lots of different departments, business units, and technology programs. As people get used to the lake, they figure out how to optimize it for diverse uses and operations, analytics, and even compliance.

Different kinds of data lake platforms

The data lake can be used many ways, and it also has many platforms that can be under it. Hadoop is the most common but not the only platform.

Hadoop

Hadoop is appealing. It has proved to have linear scalability. It's a low cost for scalability compared to, say, a relational database. But Hadoop is not just cheap storage. It's also a powerful processing platform. And for those trying to do algorithmic analytics, Hadoop can be very useful.

Relational database management system

The relational database management system can also be a platform for the data lake, because some people have massive amounts of data that they want to put into the lake that is structured and also relational. So if your data is inherently relational, a DBMS approach for the data lake would make perfect sense. Also, if you have use cases where you want to do relational functionality, like SQL or complex table joins, then the RDBMS makes perfect sense.

Cloud-based storage

But the trend is toward cloud-based systems, and especially cloud-based storage. The great benefit of clouds is elastic scalability. They can marshal server resources and other resources as workloads scale up. And compared to a lot of on-premises systems, cloud can be low-cost. Part of that is because there’s no system integration.

If you want to do something on-premise, you or somebody else has to do a multi-month system integration, whereas for a lot of systems there’s a cloud provider who already has that integrated. You basically buy a license and you can be up and running within hours instead of months. In addition, the object store approach to cloud, which we mentioned in a previous post on data lake best practices, has many benefits.

And of course, you can have a hybrid mix of platforms with a data lake. If you're familiar with what we call the logical data warehouse, you can also have a similar thing like a logical data warehouse, and this is logical data lake. This is where data is physically distributed across multiple platforms. And there are some challenges to that, like needing special tools that are good with federated queries or data virtualization for far-reaching analytic queries.

But that technology is available at the tool level, and many people are using it.

Data lakehouse, the future of the data lake?

In their quest to extract more value from their data, companies are always pushing the boundaries. Enabled by cloud-based computing, they are now often combining data lake technologies and data warehouses into a single architecture referred to as “data lakehouse.” The benefits of a data lakehouse include better integration, less data movement, better data governance, and support for more use cases.

Create a data lake

The data lake is your answer to organizing all of those large volumes of diverse data from diverse sources. And if you're ready to start playing around with a data lake, we can offer you Oracle Free Tier to get started.