DATA LAKES

Facts, Figures, and Forecast

INTRODUCTION: WHAT'S THE DATA LAKE TREND ALL ABOUT?

Picture this: you've been super busy lately and the house is kind of a mess. Paperwork is piling up, there's mail you haven't dealt with, photos you meant to sort into albums, stuff the kids have drawn or made that you need to decide whether to keep, and the list goes on. One free afternoon you say to yourself, right: time to get this stuff in order.

You allocate drawers for bills, for important documents, for family stuff, and so on, and you start sorting. Only, it's not a perfect system; some things don't fit neatly into any drawer. Perhaps it's a different category, perhaps it's too big for the drawer. Oh well, you think, I'll put all the miscellaneous stuff into a big box and figure out what to do with it later. And pretty soon, that miscellaneous box is full to the brim, you can't find anything in it, and you're back to square one.

Data lakes are like that box. They're the place you put all the data that doesn't slot neatly into the rows and categories of a relational database. This makes them great for storing valuable data that enriches your analysis but confounds a typical data warehouse. You can throw all kinds of data in there - text and image-based, product logs, IoT data, you name it - and bring all this high-volume, high-velocity data into a single stream.

This flexibility, combined with the ever-expanding nature of Big Data drawn from disparate and non-traditional sources, means that demand for data lakes is growing exponentially. This is especially true of cloud data lakes, which come with the additional benefits of being easy to use and infinitely scalable without expensive infrastructure, as well as offering flexible pricing and plenty of options for combining with other tools and services.

But while the flexibility to store data in a raw, unstructured format is a plus in the short term, it's also a challenge. Like that "miscellaneous" box, without an effective way to search through the contents of your data lake, the important stuff can end up buried in the mess. You need a clear, reliable, effective way to cut through the chaos and extract your valuable data for analysis.

WHY USE DATA LAKES?

Data lakes can provide huge amounts of storage of different types of data with high availability and greater flexibility and at a lower cost than data warehouses. With cloud managed services offerings, you can also shift your infrastructure, power and maintenance requirements to the provider, reducing the total cost of ownership.

The benefits come from having a schema-on-read rather than schema-on-write approach.

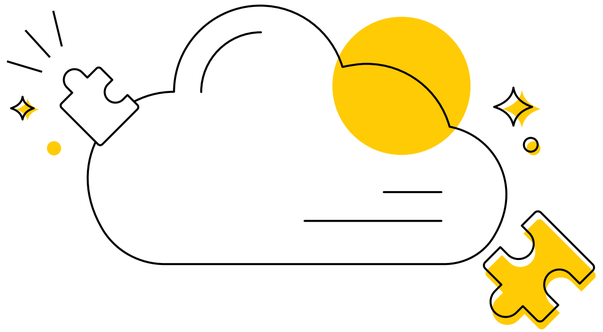
Schema-on-write is used by data warehouses: in essence, it means that data is structured and organized when it is first entered into the relational database. This makes it far simpler to find, retrieve and query that information when you need it for analysis. However, it also restricts the kind of data you can store in the database. If a particular item doesn't fit neatly into the schema of the data warehouse, there's no way of entering it - or, at least, no way of preserving all the details you might need later on, within the confines of the database.

The thing is, much of the data we use today isn't nice and tidy and easy to organize into the predefined schema. Google was the first to realize this; during attempts to index the entire internet, the company needed a smart way to label and index things like images and video so that these would show up in searches. Attempts to find a way to accommodate all that unstructured and semi-structured data led to a new approach: schema-on-read.

With schema-on-read, you don't have to predefine your schema. You preserve all the messy details from all kinds of different data sources and then impose a schema at the time of running analysis. This also means that you get to keep both the original and transformed versions to track how data has been manipulated.

And this is where data lakes come in handy. Unlike data warehouses, these are schema-on-read. You don't have to squish your data into a strict schema when you first write it to the database.

That said, you do need a way to organize it when you come to actually run queries - and the fact that all the cleaning and organizing work happens at the time of querying obviously puts a lot of pressure on the system. That means you need to make sure that the tools you use to do this are powerful and effective enough to sweep through your unstructured data quickly and efficiently.



USING YOUR DATA LAKE AS A DATA WAREHOUSE

On their own, data lakes are solely used for storage.
As we've seen, in order to make all the unstructured data in your data lake useful, when it comes to querying, you will ultimately need to replicate the way a data warehouse behaves by organizing it in a way that can be used for analysis.

As such, it's important to recognize that the question **is not**: should I opt for a data warehouse or a data lake?

Rather, the question is: should I opt for a traditional data warehouse, or opt for a data lake combined with a query engine that organizes the data into an easily searchable schema?

However, there is one more layer to this. When you're connecting a data lake to a BI platform, you also need to use a processing engine on top of your data lake. This processing engine runs between the query generated from your BI platform and the huge mass of data in the lake.

Think of it like this: a data warehouse is like a neatly organized library. There's a clear Dewey Decimal system in place based on predictable and standardized details (title, author name, the area of interest, etc). You can pop your query directly into the library search function, find out exactly where the book you need is, and go get it off the shelf yourself.

A data lake is more like a big, chaotic bookshop that has all kinds of experimental literature as well as different formats thrown in - audio recordings, videos of live productions, comic books, installations, you name it. It's impossible to standardize the way this stuff is organized, certainly no nice query tool built in to help you out, and it will take you forever to trawl through it all to find what you're looking for. Instead, you ask the helpful assistant, who knows every item in the shop and can run off and find what you need in the blink of an eye. That assistant is your processing engine.

Processing engines organize data for you and save you the time and hassle of rummaging around looking for it yourself. Examples include Athena, Presto, HIVE, Impala, and Drill.

You might be thinking: Can't I use both? Absolutely.

For example, take Spectrum, Amazon's new product feature for Redshift. This connects your Redshift data warehouse with a data lake. As a result, you get all the scalability and economical benefits of storing historical data in a data lake, while continuing to use Redshift for structured, immediately available, current data.... And when you need to use both at once, you can load the historical data from the data lake into the database to work with.

MANAGED DATA SERVICES VS SELF-DEPLOYED DATA SERVICES

If you decide to opt for a data lake, then you have two broad options to choose from. Ultimately it comes down to: should I build or should I buy?

Your first option is a managed data service, also called a data-lake-as-a-service. These are cloud-only solutions. Examples include Amazon S3, Azure Blob Storage and Google Cloud Storage.

Your second option is to go for a self-deployed data lake, such as Hadoop HDFS. This can be deployed on-premises or in the cloud.

The key difference between these two is that, with a managed service, you don't incur all the overheads, resource requirements, infrastructure demands and other costs that come with setting up and running a data lake. The vendor handles all that. You just pay-as-you-go to use your data lake.

If you choose to self-deploy a Hadoop cluster on your own data center, you're responsible for buying and maintaining servers and network equipment as well as hiring people to install, update and generally take care of the system. In some cases, at least for smaller companies and use cases, this can work out a lot more costly and involved, but the payoff is that you're totally in control of that system.

For most organizations, a managed data service will be just fine. For others, such as banks or healthcare, fierce regulations mean that putting data in the cloud is just not an option. Some industries have strict rules about where cloud data can be stored - for example, that it can't be transferred to servers outside the US or the EU - which may pose a challenge to finding a suitable data-lake-as-a-service. In these situations, it's advisable, even necessary, to choose a self-deployed data service. That said, some providers, such as AWS, are catching on to this issue by providing specialized cloud-hosted services designed to meet the stringent government regulations of certain countries and regions.

FORECAST: WHAT'S NEW FOR DATA LAKES?

Data lakes are taking off in a big way. That much is clear. Much of that growth is dominated by cloud-based data lakes, which are fast stealing market share from relational databases when it comes to storing data for analysis.

All of this has a big knock-on effect for BI platforms and tools. These are having to adapt fast, getting smarter and more powerful in order to handle huge volumes of data from data lakes and other sources, as well as increasing demand for self-service tools that make data and insights available to non-technical users across the business.

One major trend emerging now is Smart Data Discovery or Augmented Analytics. This draws on artificial intelligence and machine learning to automate many of the processes involved in tracking down data in data lakes, preparing it for analysis, and extracting insights.

What's more, as data analytics and business intelligence become increasingly important to teams and departments all across the organization, they will demand specialized tools, products, data models, and streams to meet their needs.

This wide range of ad hoc requirements, especially those involving huge, diverse data streams, is one reason cloud data lakes are taking off. It also makes it crucial for BI platforms to have fast and effective ways to sort through that data - for example, by using in-memory columnar technology to relieve pressure on infrastructure when scanning and retrieving data sets.

MAXIMIZING THE VALUE OF YOUR DATA LAKE

Data lakes are awesome when you want to store all kinds of data that doesn't fit neatly into other categories - but as we've seen, you need a way to make sense of that data before you can use it. That means you need a way to interface your data lake with powerful tools to make it accessible.

Sisense does just that. We add value to your data lake by giving you a way to extract, clean, and prepare your data for analysis. You can use connectors to work with any kind of data from any type of data warehouse or data lake, and we can even connect Live (through redshift), allowing you to load data from S3 into a schema-on-write ElastiCube in order to begin processing and working with it.

In other words, with Sisense, you actually maximize the value of your data lake, overcoming its intrinsic setbacks while taking full advantage of its flexibility and freedom.

See how Sisense makes it easy to instantly reveal business insights from complex data.

Watch a demo here

SISENSE