Introducing the Data Lake

What is a data lake

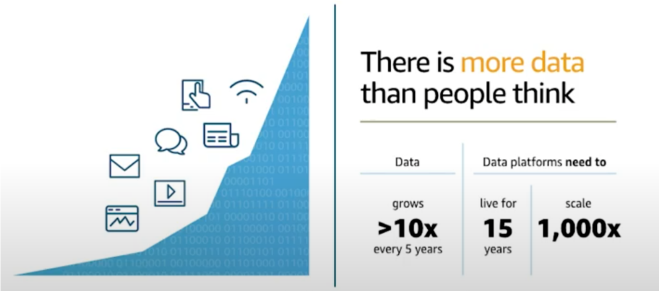
A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

What makes a data lake useful

- Data can be moved into the data lake on premise.
- Data can be moved into the data lake through batch or real time streams.
- Data is securely stored, regardless of size.
- Data can be analyzed directly from the data lake.
- Machine learning can be implemented directly from the data lake.
- Data can be in several different forms but can be accessed through out several different mediums with in the data lake without being altered (although an automated data lake prefers data to be altered). Some of these mediums can be:
  - Databases
  - Cloud networks
  - Dashboards, etc.

Why Do We Need Data Lakes, Where Does AWS Fit In

Exponential increase in data



There is more data than people think

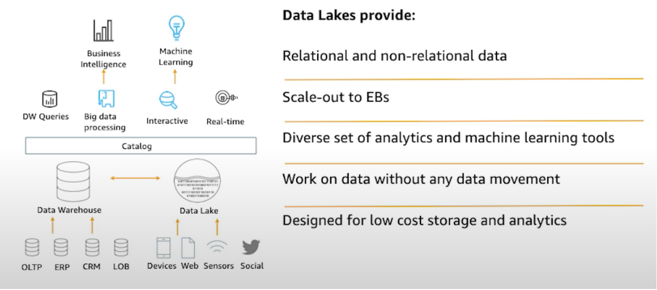
Data	Data platforms **need to**	
grows	live for	scale
>10x	**15**	**1,000x**
every 5 years	years	

The variety, speed and amount of data increases from day-to-day exponentially, and with the new technologies being created; it will continue to grow and be available through new and existing avenues. This is leading to larger and more diverse data sets coming from ERPs, CRMs, IoTs and many other types of data sources.

Also, the volume of data that we deal with is also growing exponentially. Machine generated data is growing much faster than business data because of the networks encompassing smart devices, micro services and the growth in development.

Because of this, data platforms need to live for at least 15 years and when building a data platform we need to account for the vast increase in data over those years, along with the increase of avenues ingesting them. The amount of data most companies deal

The need for centrality



Data Lakes provide:

Relational and non-relational data

Scale-out to EBs

Diverse set of analytics and machine learning tools

Work on data without any data movement

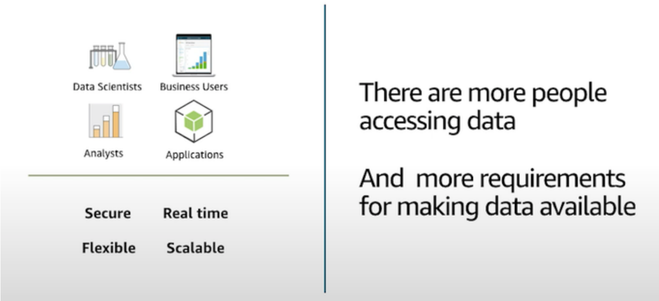
Designed for low cost storage and analytics

Data lakes allow us to break down data and place it into a single, central repository. We can store a wide variety of data formats at any scale at low costs. We also work with structured, unstructured or even incomplete data with in a data lake. With data lineage we can track where data comes from and use it with many types of analytical and machine learning tools. A well designed data lake is clean, searchable, low cost and easy to maintain.

Use of a unified set of tools

with multiplies 10 times every 5 years and so our platforms need to be able to scale 1000 times from what it is at the time of creation.

More collaboration with in an organization

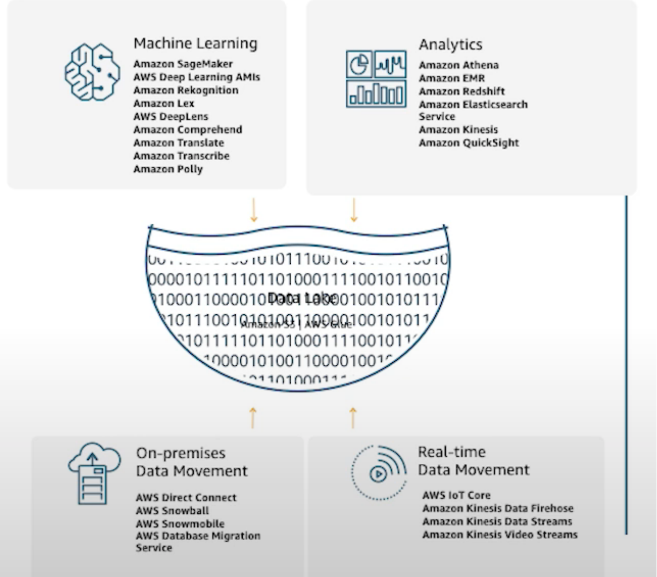


Secure
Real time
Flexible
Scalable

There are more people accessing data

And more requirements for making data available

Today, more and more teams with in a business need to access data collaboratively in order to analyze reports using business intelligence and gain insights through machine learning. These teams need flexibility to load and share this data with each other in real time and at any scale. This needs to be done while also adhering to company compliance and security requirements.



Machine Learning
Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly

Analytics
Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch Service
Amazon Kinesis
Amazon QuickSight

Data Lake
Amazon S3 | AWS Glue

On-premises Data Movement
AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service

Real-time Data Movement
AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

The architecture of a data lake allows us to manage multiple data types and structures from wide ranging sources using a unified set of tools. Using these tools makes data easily available to be categorized, processed, shared, analyzed and consumed by several different groups with in an organization.

Because there are many forms and structures, predefined schemas are no longer needed and we may figure out what questions to ask after the ingestion and curation process has completed.

Where AWS comes in

AWS makes it easy and cost effective to collect, store, share and analyze data using a data lake. Many of its services help with the needs of engineers, architects, analysts and data scientists. Some of these services are:

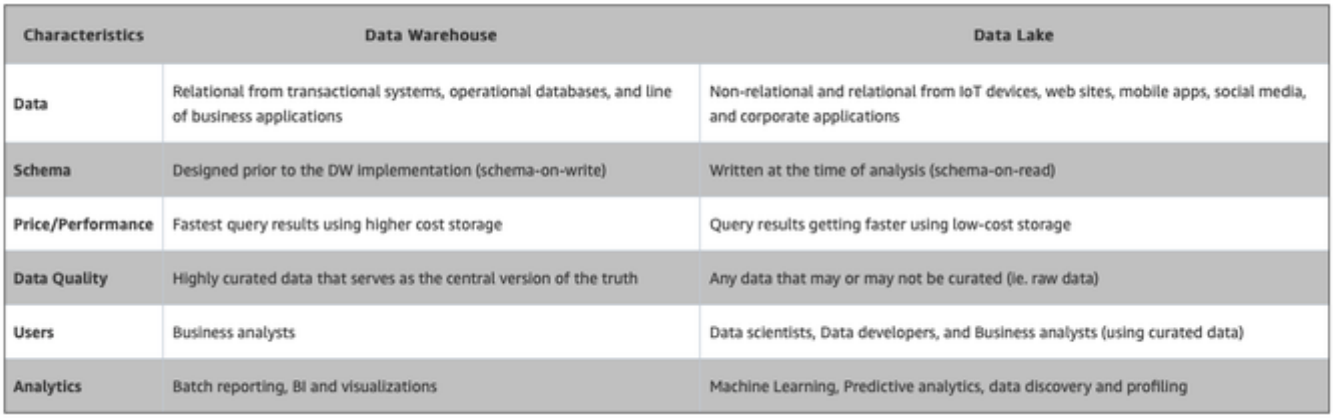
- Amazon Athena for unstructured and ad-hoc querying
- Amazon Glue for transforming and moving data
- Amazon EMR for processing vast amounts of unstructured data from open source frameworks
- Amazon redshift for data warehousing
- Amazon Kinesis for processing streaming data
- Amazon Quicksight for building BI dashboards and reports
- Amazon Elasticsearch Service for running elasticsearch clusters
- Amazon s3 for vast amounts of data storage

Why Not Create a Data Warehouse Instead

Data lakes and data warehouses are both widely used for storing large amounts of data, but they are not interchangeable terms.

There are several challenges in creating a data warehouse during the early stages of building a business or moving into a new revenue stream. It all basically comes down to what questions to ask and what business problems need to be solved. In creating a data warehouse, it is important to know how data sets should be cleaned, enriched and transformed to solve these problems. In essence, there isn’t a clearly defined goal with clearly defined boundaries at the beginning.

Also, a data warehouse can live with in a data lake after a clear direction has been decided.



Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

8 Myths Surrounding Data Lakes - Architecture, Strategy & Analytics

1. You can only choose one - data lake vs. data warehouse

People often think that we have to choose between a data lake and a data warehouse. In reality, a data lake should be what is created first and eventually we incorporate a warehouse with in the lake.

2. Data warehouses are data lakes

The idea behind this is that we can forego building a data lake and just dump all of our data into a data warehouse because it is clean and relational. The problem with this, as stated above, is we might not always know what we're going to do with our data.

3. Data lakes are just Hadoop solutions

A lot of examples of data lakes are synonymous with Hadoop. In reality a well-abstracted lake architecture, although working with Hadoop, will bring much more to the business and be able to use many more versatile tools.

4. Data lakes are just for storage in the cloud

Data lakes have an almost infinite storage space. But there is so much more to it. We store, process, transform, catalog and analyze data with in a data lake. We can also automate processes all the way up to performing business intelligence.

5. Data lakes are just for dumping raw data

Although it may appear this way, not all data that is ingested into a data lake is raw. Depending on the architecture and source, a large amount of (or even most) data can be processed and/or transformed during ingestion.

6. Data lakes are only for big data

Most data lake solutions are geared towards dealing with big data. But businesses come in different shapes and sizes and have different needs. In reality, data lakes are a central repository for storing data.

7. Data lakes have no security

AWS provides a system for creating specific access to certain types of data for people, while also providing encryption methods to anonymize sensitive data during movement.

8. Data lakes will always eventually become a swamp

With proper data governance (direction, management, curation, automation), this can easily be avoided.

Building a Data Lake on AWS

We have included a little more information on what a data lake is and what an AWS data lake can provide. Below, is the introduction to 'Building a Data Lake on AWS' by the AWS team as well as the full PDF to read. It is a short 30-45 minute read and provides a comprehensive idea on how a data lake works through some of AWS' product line:

As organizations are collecting and analyzing increasing amounts of data, traditional on-premises solutions for data storage, data management, and analytics can no longer keep pace. Data silos that aren't built to work well together make storage consolidation for more comprehensive and efficient analytics difficult. This, in turn, limits an organization's agility, ability to derive more insights and value from its data, and capability to seamlessly adopt more sophisticated analytics tools and processes as its skills and needs evolve.

A *data lake*, which is a single platform combining storage, data governance, and analytics, is designed to address these challenges. It's a centralized, secure, and durable cloud-based storage platform that allows you to ingest and store structured and unstructured data, and transform these raw data assets as needed. You don't need an innovation-limiting pre-defined schema. You can use a complete portfolio of data exploration, reporting, analytics, machine learning, and visualization tools on the data. A data lake makes data and the optimal analytics tools available to more users, across more lines of business, allowing them to get all of the business insights they need, whenever they need them.

Sources:

Building Data Lake on AWS

Data Lakes? Big Mythos About Architecture, Strategy and Analytics

What is a Data Lake and How to Create One for Your Business

Data Lakes and Analytics on AWS

What is a Data Lake on AWS