awsmarketplace

Data Integration
Data Lakes

Challenge

Ingesting data into a data warehouse can be slow, because the data must fit the database's schema. Also, before any data analytics can be done, the data must first be extracted from the database.

Solution

A data lake, which complements your data warehouse, allows you to store data in its native format. Data lakes also allow analytic tools to work with the data directly, rather than having to first extract it from the database.

Data Lakes Efficiently Consolidate Your Data

Data lakes, which complement rather than replacing data warehouses, provide a cost-effective way for large organizations to store, process, analyze, and efficiently utilize large amounts of data in many different shapes and levels of structure. A data lake is an architectural approach that allows users to store massive amounts of data in a central location, so it's readily available to be categorized, processed, analyzed, and consumed by diverse groups within an organization. For instance, an organization can place internal data, external data, partner data, competitor data, business process data, social data, and people data in a data lake so that it is at the ready for analysis and integration for diverse use.

Unlike a data warehouse, the data lake excels at utilizing the availability of large quantities of coherent data, along with deep learning algorithms to recognize items of interest that will power real-time decision analytics. To get data into the data warehouse, it is usually extracted, transformed, and then loaded (ETL) from the data source into the data warehouse, which has a highly-structured data model. By contrast, a data lake uses a flat architecture to store data.

In a data lake, the data is cataloged. One way to do this is with metadata—data that describes the data— in this case, using tags. With metadata tags, each data element in a data lake is assigned a unique identifier, and then tagged with a set of extended metadata tags. Metadata is maintained so that users have access to and can understand the data. When a business question arises, the data lake can be queried for relevant data, and that smaller set of data can then be analyzed to help answer the question.

Other Data Lake Uses

Data lakes help organizations more easily tackle big data challenges. Data lakes help organizations because they are cost-effective and flexible. They allow you to store massive amounts—exabytes—of data in extremely low-cost storage and still have it readily available for analytics. Also, because you don't have to extract, transform, and load (ETL) the data into a specific format, you can run a broader set of analytics. Data going into a data lake today might consist of machine-generated logs and sensor data (Internet of Things), low-level customer behavior (website clickstreams), social media, collections of documents (e-mail and customer files), and geo-location trails. There could also be images, video and audio, and even the more structured enterprise or customer resource management and other online transaction processing data useful for integrated analysis. No matter the data type, the data size, structure or format, data lakes are fast becoming standard as data repositories and processing systems.

Data Lake Core Requirements and Capabilities

A data lake should be able to:

- Store an index of the data and metadata that is available, including sources, versioning, veracity, and accuracy.
- Secure data at scale for operational availability, and business continuity and disaster recovery (BC/DR) requirements. Authorize, audit, and grant access to subsets of data safely and securely.
- Enable IT authority of what is in the data lake, enforcing policies for retention and disposition, and importantly, tracking personally identifiable information (PII) and its pre-cursors.
- Provide agile analytics into and from the data lake using multiple analytical approaches and data workflows.
- Scale to accommodate growing amounts of data, data systems, networks and processes.
- Converge all data sources including log files, XML files, multimedia, sensor data, binary data, social data, chat data, and people data.
- Accommodate high-speed data and integrate with the historical data to have its fullest insights.

How Data Lakes Works with Big Data

Available data grows by the minute, and useful data comes in many different shapes and levels of structure. Instead of forcing data into a static schema and loading an extract, transform, and load (ETL) set into a structured database, essentially filtering, aggregating, and in general, losing detail, a data lake approach enhances analysis agility by enabling the analyst to create new views of the data, or new data schemas on demand.

The key difference between data warehousing and a data lake, is that with a data lake, the data is at the conceptual center. Data is stored in its original format in the data lake, and you have a wide variety of tools that you can use to analyze the data. These are likely tools familiar to you, but with much more data available, the tools become much more powerful. For example, with a data warehouse, you can run a visualization tool on only a small subset of the data from the data warehouse, but with a data lake, you can run these tools on terabytes of data in Amazon S3 or through Amazon Elastic MapReduce (EMR).

1

The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.

2

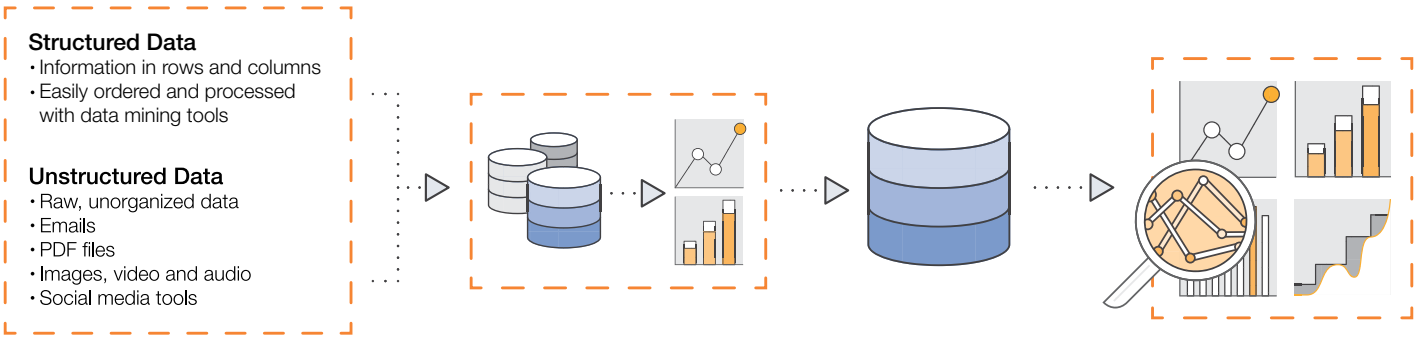
The reservoir of water is a dataset where you run analytics on all the data.

3

The outflow of water is the analyzed data.

4

Through this process, you are able to "sift" through all the data quickly to gain key business insights.



The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out

The data lake's schema-on-read allows organizations to draw full value from their data, no matter how large it grows. Data can be loaded first, then transformed and indexed in an iterative methodology as organizational understanding of data improve. With a data lake, all relevant information associated with a piece of data—whether it be data like log files, .csv files, .json files, data in columnar formats, or image data—is stored with that item in the form of metadata tags. These tags make it possible to store and manage vast amounts of data of all types and have it immediately available for analysis. This ability, coupled with the data lake's inexpensive storage running on commodity hardware, enables organizations to add a virtually unlimited number of new data sources at minimal risk.

Conclusion

By incorporating a data lake into your overall data architecture, you can increase the agility of your business, while at the same time lowering your costs. With more sophisticated data lake strategies, you can combine raw storage, SQL and NoSQL database approaches, and even meld online analytics processing and online transaction processing capabilities, which help you to discover new and actionable insights. Leveraging a data lake also means administrators can better provide and widely share not only the data, but also an optimized infrastructure with simpler management overhead.

When actively used, the data lake becomes much more than a large collection of data. It can become the central repository that all applications and analysis can reference, and from which your company's key insights are derived. Visit http://aws.amazon.com/mp/datalakes for more information about Data Lakes on AWS.

Get Started with Bi-Data Analytics at AWS Marketplace

Find and deploy the solution you need in minutes

Save money with pay-as-you-go pricing

Scale globally across all AWS regions