General Data Lake FAQ

What's 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.

Why a data lake?

Data lakes allow you to store relational data like operational databases and data from line of business applications, and non-relational data like mobile apps, IoT devices, and social media. They also give you the ability to understand what data is in the lake through crawling, cataloging, and indexing of data.

What is a data lake composed of?

A data lake is usually a single store of data including raw copies of source system data, sensor data, social data etc and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning.

What is the process of a data lake?

Data Lakes allow you to import any amount of data that can come in real-time. Data is collected from multiple sources, and moved into the data lake in its original format. This process allows you to scale to data of any size, while saving time of defining data structures, schema, and transformations.

What are the essential elements of a data lake and analytics solution?

As organizations are building Data Lakes and an Analytics platform, they need to consider a number of key capabilities including:

- Data movement

Data Lakes allow you to import any amount of data that can come in real-time. Data is collected from multiple sources, and moved into the data lake in its original format. This process allows you to scale to data of any size, while saving time of defining data structures, schema, and transformations.

- Securely store, and catalog data

Data Lakes allow you to store relational data like operational databases and data from line of business applications, and non-relational data like mobile apps, IoT devices, and social media. They also give you the ability to understand what data is in the lake through crawling, cataloging, and indexing of data. Finally, data must be secured to ensure your data assets are protected.

- Analytics

Data Lakes allow various roles in your organization like data scientists, data developers, and business analysts to access data with their choice of analytic tools and frameworks. This includes open source frameworks such as Apache Hadoop, Presto, and Apache Spark, and commercial offerings from data warehouse and business intelligence vendors. Data Lakes allow you to run analytics without the need to move your data to a separate analytics system.

- Machine Learning

Data Lakes will allow organizations to generate different types of insights including reporting on historical data, and doing machine learning where models are built to forecast likely outcomes, and suggest a range of prescribed actions to achieve the optimal result.

Why do you need a data lake?

Organizations that successfully generate business value from their data, will outperform their peers. An Aberdeen survey saw organizations who implemented a Data Lake outperforming similar companies by 9% in organic revenue growth. These leaders were able to do new types of analytics like machine learning over new sources like log files, data from click-streams, social media, and internet connected devices stored in the data lake. This helped them to identify, and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

What is the value of a data lake?

The ability to harness more data, from more sources, in less time, and empowering users to collaborate and analyze data in different ways leads to better, faster decision making. Examples where Data Lakes have added value include:

- improved customer interactions
- improve r & d innovation choices
- increase operational efficiencies

What are the benefits of a data lake?

The biggest advantage of data lakes is flexibility. By allowing the data to remain in its native format, a far greater—and timelier—stream of data is available for analysis. Some of the benefits of a data lake include: Ability to derive value from unlimited types of data.

What are potential disadvantages?

- Indiscriminate data hoarding, leading to stale data
- Different user/app interpretations of data may conflict
- If metatags are missing or inaccurate, it will be more difficult to find specific data
- Without initial checks, corrupt data may be ingested and used, before the problem is recognized
- The problems of stale and corrupt data can turn a data lake into a “data swamp.” Proper curation of data can prevent this from happening, although this means additional effort.

What about a data lake vs data warehouse?

Data lakes and data warehouses are both widely used for storing big data, but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose.

Data Warehouse

- A database optimized to analyze relational data coming from:
    - transactional systems
    - line of business applications.
- The data structure, and schema are defined in advance to optimize for fast SQL queries
- Results are typically used for operational reporting and analysis
- Data is cleaned, enriched, and transformed so it can act as the “single source of truth” that users can trust

Data Lake

- Stores relational data from
    - line of business applications,

    - non-relational data from mobile apps, IoT devices, and social media
- The structure of the data or schema is not defined when data is captured
- You can store all of your data without careful design or the need to know what questions you might need answers for in the future
- Different types of analytics on your data like SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights

Why does a data lake matter?

Data lakes are designed to be more open and accessible than a traditional data warehouse, widening the scope of analytics teams as well as encouraging development of line of business applications and the delivery of self-service access to valuable business insights and associated data-driven tools.