IBM **Analytics**
Unified Governance
& Integration

Five myths about the data lake

As enterprises become more aware of the value and importance of their data and its ability to open new opportunities and revenue models, they are being inundated with technology that offers the "best" approach to manage all that data and drive insights.

For good reason, the data lake is one such reference architecture getting a lot of attention. However, there are several myths about the data lake that cause businesses and IT managers to lose precious time as they research their options.

Five myths about the data lake

1 The data lake is a product you can buy → False

The data lake is not a product that you can just purchase. You don't just buy Hadoop or a data warehouse solution and call it a data lake.

A data lake is a reference architecture that is independent of technology. It's an approach that an organization uses to put data at the heart of its operation that includes governance, quality and management of data, thereby enabling self-service analytics to empower all consumers of data.

2 Hadoop is the only data lake → False

Even though the term data lake is often associated with Hadoop, or Hadoop-oriented object storage, a data lake could be developed and used effectively without incorporating Hadoop. For example, an effective data lake could be based on different relational database management systems.

A data lake combines a variety of technologies to establish systems of insight to provide agile data exploration for data scientists to address business needs.

3 Use data lakes to dump any data — no governance required → False

While software and hardware are key components of a data lake solution, equally important are the cataloging of data, quality of data and data governance and management processes.

Just as some data warehouses have become massive black holes from which vast amounts of data never escape, a data lake can become a data swamp if good governance policies are not applied.

In a digital business, the data in the data lake must be cataloged, accessible, trusted, and usable; active governance, quality and information management are indispensable parts of the data lake.

4 Data lake success is measured by delivering access → False

Dumping data into a central location is not a true analytics solution. The goal is to run data analyses that produce meaningful business insights; to uncover new revenue streams, customer retention models or product extensions.

But that data must be trusted, relevant, and available for all consumers of data. A data lake needs an intelligent metadata catalog that can relate to business terminology, moving cryptic-coded data and making it more understandable with context. It will also attribute to the source and quality of data from both structured and unstructured information assets and governance fabric to ensure that information is protected, standardized, efficiently managed, and trustworthy.

5 The data lake is a replacement for a data warehouse → False

The data lake can incorporate multiple enterprise data warehouses (EDW), plus other data sources such as those from social media or IoT. These all come together in the data lake where governance can be embedded, simplifying trusted discovery of data for users throughout the organization.

Therefore, a data lake augments EDW environments to allow, enable or empower data scientists and analysts to easily explore their data, and everything available to them; discovering new perspectives, insights, and accelerate innovation and business growth.

The benefits of debunking the myths

Don't get bogged down by misinformation; a governed data lake can provide access across the enterprise to a wide range of structured and unstructured data while helping ensure it is trustworthy and secure.

When optimized for a business' needs, a governed data lake can accelerate analytics and improve the accuracy of insights because:

The data sits on a secure and reliable infrastructure foundation.

The data is put in business language enabling data scientists to get to work immediately versus struggling with the meaning of cryptic terms.

Controlled data feeds populate the data lake with reliable information and then document the information assets, their metadata and business context providing real-time flow of data in to the data lake.

The data is properly classified, protected and governed.

The quality, origin and lineage of the data are well understood.

Leading organizations have sorted through the fact and fiction, and learned that a governed data lake can help them efficiently extract real business value from their data environment.

The truth is out there

The IBM® Unified Governance and Integration portfolio has successfully helped clients avoid common pitfalls and myths about the data lake; helping them through all critical steps to a successful governed data lake implementation.

Watch the video

Provide high quality trusted, business-ready data for everyone in your organization using a governed data lake.

Read the research report

Gartner outlines the most common data lake failure scenarios and how to avoid them.

Read the e-book

Explore how governed data lakes create opportunities to derive key business insights.