skillsoft
global knowledge™

White paper

Implementing (AWS) **data lakes** in your organisation in 5 steps

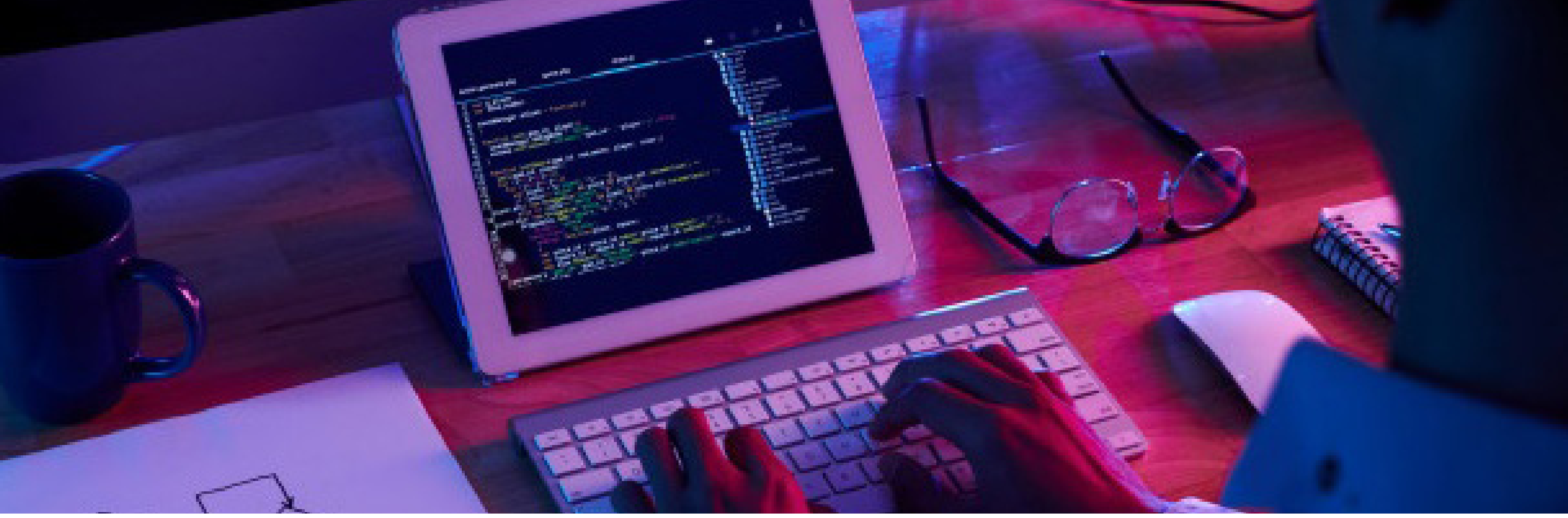
www.globalknowledge.com
info@globalknowledge.co.uk
0118 912 1929

Introduction

Modern businesses are constantly on the lookout for ways to harness the full potential of huge volumes of data without impacting their core operations. Not so strange, considering the fact that organisations that successfully generate business value from their data generally outperform their peers. Consequently, more and more organisations are exploring the possibilities of data lakes. These data lakes take the shape of inexpensive, cloud-based solutions for organisational data storage on a grand scale. An **Aberdeen survey** concluded that organisations that implemented a data lake outperformed similar companies by about 9% in organic revenue growth.

Big players like Amazon and Microsoft have done a lot to create and perfect the data lake concept. But what is a data lake? What are the main benefits of this technology? And what are the steps required to implement a data lake? This whitepaper delivers the answers to these pressing questions. We will predominantly focus on how to implement data lakes in an AWS environment.

Table of contents



What is an AWS data lake?

A data lake is a centralised repository that allows you to store all your structured and unstructured data. And here comes the big advantage: the scale doesn't matter. You can store your data 'as-is'. This means that the often tedious and time-consuming process of first structuring your data is a thing of the past. Using data lakes allows you to run different types of analytics—from dashboards and visualisations to big data processing, real-time analytics and machine learning—to guide better business decisions.

A data lake differs from a data warehouse in several ways:

- Data lakes store raw data, whilst data warehouses are geared towards storing and processing processed and refined data for specific purposes.
- The purpose of the data in a data lake is not yet defined. Data warehouse data is already in use and tailored towards specific questions, processes or business needs.
- A data lake is highly accessible and quickly updatable. Managing a data warehouse is a more complex and costly affair. Making changes requires significant technical expertise and is often expensive.
- Data lakes usually require more storage capacity than data warehouses.

> **"Using data lakes allows you to run different types of analytics—from dashboards and visualisations to big data processing, real-time analytics and machine learning—to guide better business decisions."**

The 4 main benefits of a data lake

Let's take a look at the mainbenefits of using data lake technology.

1. Improved customer interactions

Customer experience is king in modern times and has overtaken price as the prime determining factor for business success. One of the strongpoints of a data lake is that it allows you to combine customer and sales data from various sources, such as a CRM platform that harbours advanced business and social media analytics, a marketing platform that includes the complete buying history of a customer, or incident tickets that allow you to optimise your services.

2. Agile analytics

Data lakes allow you to harness the full power and diversity of data analytics. You can create various roles in your organisation, such as data scientists, data developers and business analysts, and allow them to access data with their personal choice of analytic tools and frameworks. An extra asset of data lakes is that you don't have to move your data to a separate analytics system.

3. Generate different types of insights

Dive into historical data, use machine learning models to predict future outcomes, or gather insights that allow you to formulate prescribed actions for maximum business success. Whatever your goals, data lakes give you the opportunity to generate the different insights required to achieve them. Additionally, the information comes in real time and in its original, unaltered format.

4. Operational efficiency

The Internet of Things (IoT) has increased the number of connected devices that create their own data. A data lake makes it much easier to store and run analytics on machine-generated IoT data. This allows you to discover new ways to reduce operational costs, while at the same time increasing the quality of your services, products and operations. The information boost that is often the result of using data lakes also makes it easier for R&D teams to refine their operations.

> **"A data lake makes it much easier to store and run analytics on machine-generated IoT data."**

The 5 steps for implementing a data lake

Now we know that data lakes come with many potential benefits for your organisation. But what are the steps required to implement a data lake in AWS? Read on to find out.

Step 1: Data source identification

Implementing a data lake starts with the process of data source identification. This entails analysing all the data that is destined to end up in the lake. Is the data tracked in log files? Is information coming in batches? Are there existing data stores that might be related? And who are the stewards or owners of the data origination environments?

In this first phase, you should also carefully define the data access and security policies that you want to apply to certain types and sets of data. In AWS, Lake Formation helps you with this and eases the process of collecting and cataloguing data from various databases and other object storage environments.

Step 2: Moving the data

The next step consists of moving your data to your newly created Amazon S3 data lake—a process that is often referred to as 'data ingestion'. This is a mainly technical operation that involves tasks such as:

- Setting up processes to schedule periodic file transfers or batch data extracts.
- Determining how long log data should be available.
- Setting up the storage location. In AWS, this would be a storage account with S3 buckets that serves as the data lake. From a management perspective, it is wise to establish a consistent bucket naming and storage approach.
- Setting up processes to bring in reference data (users, departments, calendar events, work project names).
- Considering other groups/departments that may be impacted by any new processes established and communicating the changes proactively.

Step 3: Data clean-up and organisation

Cleaning up and organising your data allows you to combine information in more meaningful ways. Step 3 consists of tasks such as determining common identifiers across incoming data records, identifying mappings between similar but differently named data fields, and manufacturing a global set of identifiers to unify the data across different systems.

Amazon S3 allows you to clean and classify your data using machine learning algorithms. It also gives you secure access to all of your sensitive data. Your data lake users can also access the centralised **data catalogue** AWS Glue, which describes available datasets and their appropriate usage.

Step 4: Staging data

Staging your data for targeted queries allows you to structure information for optimised usage. This step includes:
Establishing the types of queries that will be needed to draw useful insights from your data.
Setting up table layouts in your data lake.
Building a library of queries. This will prove useful for later data usage in dashboards and reports.

AWS allows you to stage and analyse datasets with your personal choice of tools and machine learning services. Amazon Redshift, Amazon Athena, Amazon ERM or Apache Spark are a couple of high-quality options. The choice is entirely yours.

Step 5: Visualising data

Accessing and visualising your data is the last step. To do this, you'll need business intelligence tools that provide you with powerful dashboards and reporting tools. Amazon Quicksight is a powerful BI tool that allows you to translate information from your data lake into insightful BI dashboards that are powered by advanced machine learning functionalities.

How Global Knowledge helps

Global Knowledge has all the necessary expertise to teach you how to implement AWS data lakes. We offer instructive courses on the subject. Our course '**Building Data Lakes on AWS**' will teach you how to build an operational data lake that supports the analysis of both structured and unstructured data.

You will learn the components and functionalities of the services involved in creating a data lake. You will also use AWS Lake Formation to build a data lake, AWS Glue to build a data catalog, and Amazon Athena to analyse data. The course lectures and labs further your learning with the in-depth exploration of several common data lake architectures.

More information

Would you like to know more about building a data lake with AWS? Or do you want to sign up for one of our courses? Then don't hesitate to contact us. Just call us at 0118 912 1929, send an email to **info@globalknowledge.co.uk** or use our live chat.

CONTACT US

www.globalknowledge.com
info@globalknowledge.co.uk
0118 912 1929