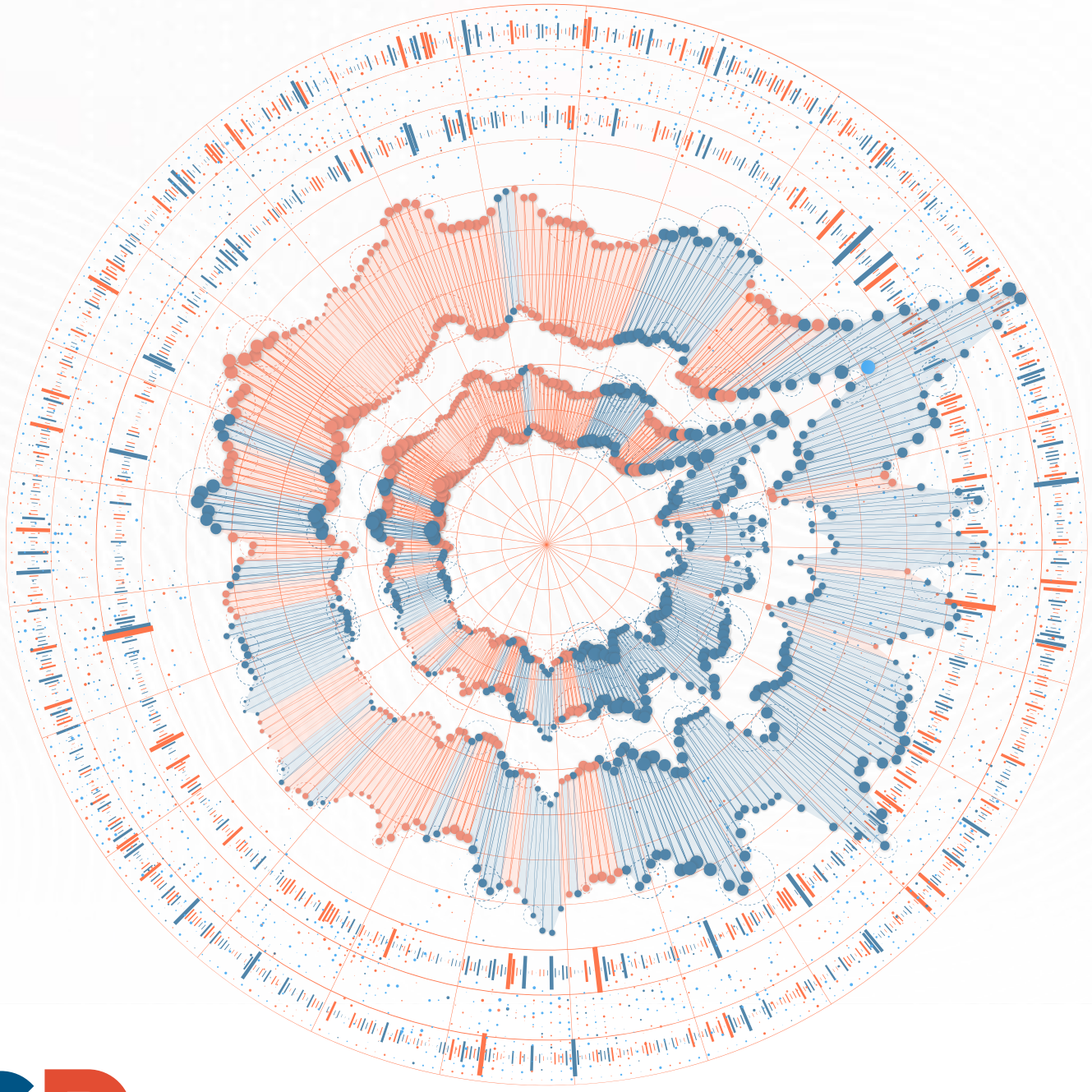

GoDgtl
by PruTech

AWS DATA LAKES & BEST PRACTICES

Table Of Contents

GoDgtl understands how cloud computing - and the benefits of flexibility, scalability, security, and agility enabled by **cloud computing** - can transform organizations.

Introduction

A Data Lake provides you with a centralized repository for a wide variety of data forms in a central platform. It supports structured, semi-structured, and unstructured data types. With Data Lakes, you can break down data silos and support a wide range of applications across analytics and machine learning use cases. Moreover, you can achieve all these capabilities without moving or duplicating data or interfering with different use cases.

To break it down, imagine structured, semi-structured, and unstructured data from various forms of documents, databases, text, JSON, and much more. How can an organization place all this data into a repository to go through the process of ETL and convert it into normalized data? Through Data Lakes.

If your organization collects and depends on data-driven decisions, there are several reasons to ingest all your data into a Data Lake. Think of all the data in a structured database. Everything ranging from clickstream data, IoT sensor data to network device data could be aggregated into a centralized repository to perform actions like training machine learning models on the data or running predictive analytics. Structured data can help you gain deeper insights, drive greater efficiencies, and generate meaningful experiences for better business outcomes.

This white paper sheds light on the importance of Data Lakes, their benefits, and how your business can build an effective Data Lake by following best practices to drive meaningful insights from your data.

Why Use Data Lakes?

The reason why so many customers are building and moving to Data Lakes is that it provides a way to store relational and non-relational data at a massive scale. They also support various tools that help you analyze this data and gain deeper insights. Moreover, you get a central data catalog that can provide you with an insight into what you own. Additionally, it can help you run services like EMR for your Big Data applications or Amazon Athena for ad-hoc, real-time interactive analysis.

You can also use Amazon Redshift for your Data Warehouse and Redshift Spectrum to run scale-out exabyte queries across data stored in your Data Lake in S3 or Redshift. Organizations need to have dashboards and visualizations to view their real-time analytics and gain better insights into their current organization to make better decisions leading to improved outcomes.
And that is where Data Lakes help.

Building Out a Data Lake

- **Set up the storage:** S3 is a very cost-effective option, and with its 9.9999999999s of availability, it provides a great storage layer for the Data Lake.
- **Move raw data:** You must move your storage from on-premises (or from various sources) into the Data Lake in its raw form.
- **Organize the data:** Once the data is ingested into the storage in its raw form, the data needs to be cleaned, prepped, and cataloged to make it readily discoverable and available for analytics.
- **Encrypt the data:** The data must then be encrypted with the appropriate security policies specified on the data, ensuring only authorized users can access the data and keep it in compliance.
- **Make the data readily available:** Finally, make the data available for a wide variety of use cases within your organization.

Essential Elements to Consider when Building Data Lakes:

Following are some of the vital elements that you must consider when building data lakes:

Data movement: Data movement is a process of importing any amount of real-time data from multiple sources and moving it into the Data Lake. It also allows you to scale to data of any size, defining structures, schema, and transformations.

Securely store and catalog data: It allows you to store relational and non-relational data. This process enables you to understand data through crawling, cataloging, and indexing. Finally, you must secure it to ensure that your data assets are protected.

Analytics: It allows data scientists to access data with their choice of analytic tools and frameworks

Machine Learning: It allows organizations to generate insights with the help of machine learning models, predictions, and recommendations to achieve optimal results.

Don’t Lose Sight of the Important Details

If your data lake is poorly organized or contains too much “junk,” it is no longer a data lake; instead, it is referred to as a “data swamp.” As you can guess, aside from other issues that may arise, data swamps can be unnecessarily costly. To ensure that your data lake remains “clean,” there are a few things you need to be mindful of.

First, as a business, reduce the collection of useless data as much as possible. With access to limitless storage, it has become easy to store each data point, and this freedom to keep everything has put companies in a disadvantageous position. It allows them to hoard information that serves no purpose other than to increase costs and render their data lake ineffective. Also, it is crucial to keep the lifecycle of data in mind. All the data stored should be used for a purpose and then either archived or destroyed (unless you need it for other purposes). Automation comes in very handy here, and you should try to implement it as early as possible.

Why Data Lakes **Fail**

There are several reasons why Data Lakes fail. The first reason is because of the data swamps issue discussed above. After unnecessary hoarding occurs and all structures and organizations are lost, a data lake becomes much less practical and reliable, and users eventually stop using it. Data volumes are another issue. While data lakes are supposed to contain large amounts of information, having to parse through all of it is a challenge — and for some, it is a challenge they cannot handle.

Another important reason behind data-lake failure is that businesses fail to utilize the data for analytical purposes effectively. This often happens when data becomes stale, thanks to the slow nature of business processes, and is no longer valuable. In many cases, this leads to the analytics produced by the Data Lake not having the expected impact, causing businesses to re-evaluate the use of data lakes altogether.

AWS Data Lake Best Practices

Here are some of the best practices you should follow to ensure success when building a Data Lake for your business.

1 Capture and Store Raw Data in its Source Format

Before any cleaning, processing, or data transformation takes place, your AWS data lake should be configured to ingest and store raw data in its source format. Storing data in its raw format allows analysts and data scientists to query the data in innovative ways, ask new questions, and generate novel use cases for enterprise data. The on-demand scalability and cost-effectiveness of Amazon S3 data storage mean that organizations can retain their data in the cloud for more extended periods and use data from today to answer questions that pop up months or years down the road.

Storing everything in its raw format also means that nothing is lost. As a result, your AWS Data Lake becomes the single source of truth for all the raw data you ingest.

2 Leverage Amazon S3 Storage Classes to Optimize Costs

Amazon S3 offers multiple classes of cloud storage, each cost-optimized for a specific access frequency or use case. Amazon S3 Standard is a solid option for your data ingest bucket, where you'll be sending raw structured and unstructured data from your cloud and on-prem applications.

Remember, data that is accessed less frequently costs less to store. Amazon S3 Intelligent Tiering saves you money by automatically moving objects between four access tiers (frequent, infrequent, archive, and deep archive). Intelligent tiering is the most cost-effective option for storing processed data with unpredictable access patterns in your data lake.

You can also leverage Amazon S3 Glacier for long-term storage of historical data assets or to minimize the cost of data retention for compliance/audit purposes.

3 Implement Data Lifecycle Policies

Data lifecycle policies allow your cloud DevOps team to manage and control the flow of data through your AWS data lake during its entire lifecycle.

They can include policies for what happens to objects when they enter S3. In addition to that, there can be specific policies for transferring objects to more cost-effective storage classes and also policies for archiving or deleting data that outlived its usefulness.

While S3 Intelligent Tiering can help with triaging your AWS Data Lake objects to cost-effective storage classes, this service uses pre-configured policies that may not suit your business needs. With S3 lifecycle management, you can create customized S3 lifecycle configurations and apply them to groups of objects, giving you total control over where and when data is stored, moved, or deleted.

4 Utilize Amazon S3 Object Tagging

Object tagging is a useful way to mark and categorize objects in your AWS Data Lake. Object tags are often described as “key-value pairs” because each tag includes a key (up to 128 characters) and a value (up to 256 characters). The “key” component usually defines a specific attribute of the object, while the “value” component assigns a value for that attribute.

Objects in your Data Lake can be assigned up to 10 tags, and each tag associated with an object must be unique. However, many different objects may share the same tag.

There are several use cases for object tagging in S3 storage. For example, it allows you to replicate data across regions using object tags, filter objects with the same tag for analysis, apply data lifecycle rules to objects with a specific tag, or grant users permission to access data lake objects with a specific tag.

5 Manage Objects at Scale with S3 Batch Operations

With S3 Batch Operations, you will be able to execute operations on large numbers of objects in your AWS data lake with a single request. This feature is especially useful when your AWS Data Lake grows in size, and it becomes more repetitive and time-consuming to run operations on individual objects.

Batch Operations can be applied to existing objects or new objects entering your Data Lake. You can also use batch operations to copy data, restore it, apply an AWS Lambda function,replace or delete object tags, and more.

AWS Lake Formation

AWS Lake Formation is a service that allows you to get a Data Lake up and running in the Amazon cloud. It organizes various AWS tools (such as AWS Cloud Backup) into one orchestrated service. This means AWS Lake Formation is a wrapper that glues many other services together to present you with a functional data lake. This service isn't necessary (as you can do all this by yourself), but it certainly helps you remove the massive overhead required for this process. For example, creating a data lake involves running services like IAM, S3, SQS, and SNS, and configuring all of these takes up your valuable time.

AWS Lake Formation works by utilizing a pre-configured set of templates, which are used to bring up all the AWS services discussed above quickly and coherently. You can also modify these templates to tailor them to your specific needs. To create a data lake using AWS Lake Formation, you need to define the data sources and the security policies to be applied. Then, the service collects all the existing data for you and moves it to your new data lake stored in S3.

But while AWS Lake Formation does a great job of creating a functional data lake for you, it does only that—and nothing else. To have an actually useful Data Lake, you need to have an entire pipeline in place, including active data ingestion and data analytics, to produce some value. None of this will be created for you, so there is still some manual work that has to be done. How you set up your data ingestion and whether you will rely on machine learning, Athena, Amazon Redshift, Amazon EMR, or something else is entirely up to you

AWS Lake Formation itself comes at no additional cost—being a wrapper service, there is nothing to charge. **But you will be paying for all the benefits brought up using AWS Lake Formation, so keep that in mind.**

Solving Your Big Data Challenges with AWS Data Lakes

As is evident, there are numerous benefits to deploying AWS Data Lakes in the cloud. Improved elasticity, security, deployment time, availability, and cost-effective storage growth are some of the notable advantages. However, there are also a few downsides, particularly if your Data Lakes are poorly organized.

With this white paper, we also reviewed AWS Lake Formation, an AWS managed service that takes all the necessary services to run a Data Lake. In addition to running a Data Lake, the service also packages and configures them for you. While not a complete solution, AWS Lake Formation is a great place to start, and with a bit of additional work, you can have your Data Lake environment up and running fairly quickly. If you are running your business on the AWS cloud and if Data Lakes provide value to your company, we encourage you to experiment with AWS Lake Formation.

As valuable as Data Lakes can be, it is crucial to remember that **their value can decrease very quickly if not utilized correctly.**

How Does GoDgtl Collaborate With AWS?

GoDgtl brings a team of experienced cloud experts who work directly with AWS to bring value and real solutions for your cloud projects. With direct access to AWS resources and in-house cloud consulting talent, GoDgtl is ready to guide you through your cloud journey, regardless of where you are on that path. Whether it is more knowledge-based information on cloud topics such as security, governance, and compliance, or basic cloud migration aspects, or even if an assessment is needed, GoDgtl can provide a roadmap for your path to project completion and success.

Sources

https://aws.amazon.com/s3/features/batch-operations/

https://dev.to/awsmenacommunity/amazon-connect-data-lake-best-practices-aws-whitepaper-summary-3b9i

https://www.chaossearch.io/blog/data-lake-best-practices

https://d1.awsstatic.com/analyst-reports/idc-bv-datalakes-analytics-ml-2020.pdf

https://info.convergeone.com/hubfs/C1-AWS-Data-Lakes-White-Paper.pdf

https://aws.amazon.com/products/storage/data-lake-storage/

https://aws.amazon.com/s3/

GoDgtl by PruTech

Our mission is to **help client organizations like yours access the latest resources and make their DX goals a reality.** Connect with our teams at Go-Dgtl to embrace new ideas and key enablers. **We promise to make your digital acceleration journey a success.**

go-dgtl.com/contact-us

OUR LOCATIONS // Charlotte | Bangalore | Hyderabad | Mexico City | New Jersey (Iselin) | New York | Washington DC

CONTACT US // info@go-dgtl.com | (646) 536-7777 | go-dgtl.com