AWS Architecture Monthly

March 2020

Data Lakes

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers, or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Editor's Note

A data lake is the fastest way to get answers from all your data to all your users. It's 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.

A data lake is a new and increasingly popular way to store and analyze data because it allows companies to manage multiple data types from a wide variety of sources, and store this data, structured and unstructured, in a centralized repository.

We hope you'll find this edition of Architecture Monthly useful, and we'd like your feedback. Please give us a star rating and your comments on the Amazon Kindle page (https://amzn.to/Kindle-magazine). You can view past issues at https://aws.amazon.com/whitepapers/kindle/ and reach out to aws-architecture-monthly@amazon.com anytime with your questions and comments.

In February's issue:

- **Ask an Expert**: Taz Sayed, Tech Leader, AWS Analytics
- **Blog**: Kayo Sports builds real-time view of the customer on AWS
- **Case Study**: Yulu Uses a Data Lake on AWS to Pedal a Change
- **Solution**: Data Lake on AWS
- **Managed Solution**: AWS Lake Formation
- **Whitepaper**: Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility

Annik Stahl, Managing Editor

Ask an Expert:

Taz Sayed, Tech Leader, Analytics

What are the general architecture pattern trends for the Big Data/data lakes industry?

Traditionally, the Lambda architecture and its variants have been a dominant trend. More recently, we are also seeing an increase in the adoption of siloed batch and streaming analytics incorporating a combination of multi-stage decoupled data bus, serverless with event driven, and pub/sub actions.

What are some questions customers need to ask themselves before considering a data lake?

Building a data lake is a strategic move for an organization and should be looked at as more of an approach first and then as solution or a service. A few things you should initially consider:

- The type of data that the data lake will be serving — data types, data formats, data size, ingestion velocity, collection volumes, evolving over time, etc.
- Defined purpose and outcome of the data lake — what will this data drive?
- Data acquisition, unification, and metadata management.
- Organizational competency around the data lake tools and solutions.

When putting together an AWS architecture to solve business problems specifically for a company that wants a data lake, do you have to think at all differently?

Every customer engagement is unique and combined with the options available in this space as well as the answers to some of the questions above, every approach that AWS takes is tailored to the needs of the customer's workload. Differentiating factors can be as granular as a data partition strategy or as high level as the infrastructure selection approach between serverless or managed or self-managed.

Do you see different trends in data lakes in cloud versus on-premises?

On-premises data lakes present several challenges, especially when it's an undifferentiated factor for the business. Maintenance costs, scalability inhibitions, and building/managing complex pipelines are some of the experiences that differentiate a cloud deployment.

What's your outlook for data lakes, and what role will cloud play in future development efforts?

Data lakes are an essential component of a pipeline that offer organizations a solution for collecting very large volumes of data which is analyzed for valuable business insights. Building your data lake in the cloud allows customers to focus their resources on their business value and lower their engineering costs by availing of managed services to scale as per your needs, leverage automatically updated technologies and high reliability and availability factors.

Imtiaz (Taz) Sayed is the WW Tech Leader for Analytics at AWS, and drives thought leadership and global technical strategies for the Data and Analytics profiles across AWS and its customers. He champions enablement and learning initiatives for the AWS Analytics community, works with customers to architect solutions and enable adoption of Analytics and Big Data technologies. Taz is an avid reader, an outdoor sports enthusiast, and a movie buff, and you can reach him at www.linkedin.com/in/contacttaz.

By Hector Leano

Online at: *https://amzn.to/aws-DL-kayo*

Kayo Sports is Australia's dedicated multi-sports streaming service offering more than 50 live and on-demand sports streamed instantly. Kayo Sports was looking to create a unified database integrating internal and external sources of data, including customer behavior, preferences, profile information, and other interactions to provide a better experience across customer touchpoints. The company decided to build a cloud-native platform on AWS to collect, process, and manage customer engagement data in real time. This unified data platform has become a hub for machine learning and enables departments to manage their own reporting and analytics.

In this interview with Sajid Moinuddin and Narendra Bharani Keerthiseelan of Kayo Sports, we had the chance to learn about the what, how, and why his team chose to build their one-customer-view platform on AWS.

"AWS' analytics stack with native offerings like EKS, EMR, Athena, Glue, Redshift, Kinesis, and S3 helped us solely focus on the business problem while shielding us from the trivial integration and performance challenges that are typical to any on-premises Big Data ecosystem. Many operational requirements came ready out of the box with AWS' managed services, so our small engineering team was able to focus solely on the business use cases without incurring much administrative overhead to enable those use cases in three months on our production environment." – Sajid Moinuddin

Q&A with Narendra Bharani Keerthiseelan, Principal Data Engineer at Kayo Sports

Tell us a bit about the real-time streaming analytics and data lake you built on AWS. Broadly: What was your approach, how did you build it, and what are you using it for?

At Kayo Sports, we built a Customer Data Platform that provides a unified database for all customer behavior, preferences, profiles, and other interaction data, from internal and external sources. The platform stitches profiles across data sets and integrates with different

vendors, enriching the data and providing a better customer experience across all channels and customer touchpoints.

First, our approach was to build a cloud native platform which collects, processes, and manages customer data and engagements across vendors and partners in real time. This unified data platform should be a data hub for machine learning and integrations and also enable all departments with reporting/analytics needs.

As a strategy, we identified the key managed services on AWS that could be leveraged to develop our architecture around, allowing us to solely work on the business value. At the same time, it reduced the cost in time and money on integrations of vendor-based solution. On AWS, we were able to achieve this very quickly by choosing S3 as the storage for the data lake. The other AWS services seamlessly integrated with S3, on which we could build our key components like EKS and EMR for compute, Glue for data catalogue, Athena for data exploration, and Redshift and Redshift Spectrum for data warehouse.

What was the key problem you were looking to solve by implementing the streaming analytics and data lake?

Our primary objective behind building the data lake was to enable a data-driven decision-making platform at Kayo Sports. Quickly deriving value from our myriad of data sources was the key focus area. Being a greenfield solution, we wanted to avoid any architectural non-reversibility in the platform and grow organically with a rinse-and-repeat approach of build-measure-learn.

Some of the key problems addressed by the platform include:

- Eliminate data silos
- Stitching of data on the data platform instead of system to system integrations.
- Democratized access to data via a single, unified view of data across the organization
- Store high volumes of raw and transformed data in the data lake, at low cost
- Accommodate high speed data
- Secured data with governance
- Centralized catalogue
- Advanced analytics capability

What have AWS services allowed you to accomplish that you couldn’t have done without it? Put another way, why did you decide to build your data lake and streaming analytics on AWS?

Kayo was slated to grow rapidly due to the wide variety of partners and affiliate integration that it started with. So we wanted to opt for a platform that can keep up with this pace of

growth without any significant spike on infrastructure budget and engineering effort as we face the four V's of Big Data (volume, variety, velocity, and veracity).

The AWS analytics stack with native offerings like EKS, EMR, Athena, Glue, Redshift, Kinesis, and S3 helped us solely focus on the business problem while shielding us from the trivial integration and performance challenges that are typical to any on-premises Big Data ecosystem. Moreover, we were able to build a scalable, agile platform by combining AWS' full-stack support for infrastructure as code and the GitOps process that we adopted across our engineering teams.

What are the top benefits Kayo Sports has realized by building your data lake and streaming analytics platform on AWS?

Kayo is the first streaming service of its kind in Australia, and we had a very aggressive product launch roadmap. Many of the data sources and analytical use cases were only discovered during the final months before launch. AWS services provided us with the required stack to enable those use cases in three months in our production environment with a small engineering team. Currently in production, we are running 7,000 data pipelines daily with 1,500 spark jobs and 30,000 python jobs replenishing our data lake in near real time and from 125 different sources.

Also, to make the data lake production-ready, we had to sort out a lot of operational aspects of the platform like security, alerting, monitoring, availability, etc. By using mostly AWS managed services, many of these operational requirements came ready out of the box, so the team was able to focus solely on the business use cases without incurring much administrative overhead.

With EMR and Redshift Spectrum, we can scale up our capacity on demand by treating our compute nodes like "cattle." Combined with the dependable SLA provided by the AWS product offering, we were able to achieve 99.9% platform availability. Similarly, for security and compliance, we followed the shared responsibility model of AWS and built a federated security infrastructure for Kayo that is seamlessly integrated with the analytics product suite.

Real-World Example

Trends with Data Lakes & Analytics

Rahul Pathak, general manager for Amazon Athena and Amazon EMR, talks about some of the key trends we're seeing and describes how they shape the services AWS offers.

https://amzn.to/aws-dl-reinvent2019

Case Study:

Yulu Uses a Data Lake on AWS to Pedal a Change

Yulu, a shared vehicale company in India, improved service efficiency by 30–35% using its prediction model and AWS data lake.

"The biggest benefit of being on AWS is that my team completely focuses on application development and spends more time coming up with new features." — Naveen Dachuri, CTO and Cofounder, Yulu Bikes

Micro Mobility Solutions

Traffic congestion and air pollution are serious issues in India, particularly in megalopolises such as Bengaluru. Yulu's mission is to address such challenges by providing sustainable and hassle-free micro mobility solutions for commuters travelling short distances. Launched in December 2017, Yulu provides a network of over 10,000 shared vehicles, which include Yulu Move (smart bicycles) and Yulu Miracle (smart light-weight electric scooters), in Bengaluru, Pune, Mumbai, and Bhubaneswar. These vehicles can be easily rented with a user-friendly mobile app on a pay-per-use basis. But smart bicycles and electric scooters are just the beginning—various form factors catering to multiple use cases and infrastructure across India are on the horizon for Yulu's ambitious management team.

Platform of Choice

Yulu is a data-driven organization, streaming data from users' mobile phones and using bikes as Internet of Things (IoT) devices. A team of over 50 operators manage the fleet to ensure bikes are well positioned in high-demand areas—a key activity Yulu calls "Rebalancing." Profits are derived from a high utilization ratio, meaning how many times a bike is ridden in a day and how many total bikes are in service that day.

The startup selected Amazon Web Services (AWS) to launch its business for several reasons. As with most startups, speed-to-market and low upfront costs were top priorities. When recruiting staff, CTO and Cofounder Naveen Dachuri found that in the Indian market more engineers had experience with AWS than any other cloud platform, so the learning curve would be less steep. This would also allow him to launch the application quickly without heavy investments in training time or cost. Yulu did consider Google Cloud but found that AWS' 10GBps network with dedicated fiber was much better than Google's offer at the time.

The existence of an AWS data center in Mumbai was another compelling factor, as data residency requirements in India were becoming more stringent.

Data Pool to Data Lake

Dachuri has more than 18 years' experience managing databases and was careful to select the right product to fit his business needs. He chose Amazon Relational Database Service (Amazon RDS) with Amazon Aurora for its speed and demand-based scalability, particularly for real-time instances. Dachuri's team relies on such instances to feed real-time data to its proprietary prediction model, so operators can rebalance the location of its bikes when demand spikes in certain areas or at certain times. "With Amazon RDS, it was simple to move from smaller to larger instances based on the type of data we were getting," Dachuri says. Yulu uses Amazon EC2 M5 instances, which Dachuri says are well suited to analytics and IoT environments because of their storage capacity and flexibility.

Yulu spent the first six months of operations collecting data to understand usage patterns. It then began constructing its prediction model using Amazon EMR for deeper analysis. "Amazon EMR gives us a seamless integration to move our data from our transaction system to Yulu Cloud - our data lake, which runs on Amazon Simple Storage Service (Amazon S3)," Dachuri says. "We can now proactively manage our vehicles, so they are always in great condition and act quickly on vehicles that move outside our operational zone to bring them back to high demand areas."

Increased Efficiency

In the four months since launching its prediction model, Yulu has seen excellent results. "The accuracy is increasing day by day," Dachuri explains. "The interesting factor is that even on day one when we launched the model, our rebalancing efficiency increased by close to 30 or 35 percent, which is a very significant number."

The model is used in all four cities of operations, and Yulu is fine tuning it to work in new cities as the company expands across India. The model is at the heart of Yulu's business and will help the company not only grow its bike market, but also add customers when it introduces electric scooters and other vehicles. The company started with just 500 bicycles and is now approaching 7,000 bicycles and electric scooters across the four cities.

Lean Teams Need Support

Yulu can maintain a lean operation despite a rapidly growing user base. Though the IT team has increased from 2 to 15 people, they have no one dedicated to infrastructure, DevOps, or database management—and do not plan to change. Only 2–3 percent of its engineers' time

is spent on databases, a task that used to absorb at least one dedicated employee in Dachuri's past work environments.

Monitoring tools such as Amazon CloudWatch and 24/7 access to AWS Business Support enable this business model. Support tickets are usually resolved in 1 to 2 hours, Dachuri confirms. He comments that while most questions can be resolved by referencing online documentation, the ease and speed of calling someone at AWS on the phone are invaluable in fast-paced environments.

The team frequently refers to live tutorials and similar documentation when using a new AWS service such as Amazon Cognito. Yulu uses this service to authenticate and authorize users, securely store customer data, and generate verification codes. One-time passwords are then sent to riders via SMS using Amazon Simple Notification Service (Amazon SNS).

Time to Innovate

The Yulu app has improved greatly over time, as the IT team can easily deploy updates. They test and push new features in one to two weeks. "Without the AWS Cloud, that would take at least six to seven weeks," Dachuri explains. "The biggest benefit of being on AWS is that my team completely focuses on application development and spends more time coming up with new features," he adds.

In addition to development, Yulu's engineers use their time to read up on and explore new features. "My analytics team wanted to use [Apache] Kafka but figured out there's already a service called Amazon Kinesis, which they have started exploring. If this meets our requirements, we will start using it right away," Dachuri says. "With AWS overall, we get ease of use, fast time-to- market, reduced risk, stability, and, of course, cost savings because you pay for what you use."

Real-World Example

Jubilant FoodWorks: Driving a Quality Customer Experience Using Data Lake

Anand from Jubilant FoodWorks talks about the company's newly set up data lake, which is powering near real-time insights to the business, as well as CRM marketing and multiple AI models.

https://amzn.to/aws-dl-tma-jubilant

View this solution online: https://amzn.to/aws-dl-dl-solution

What does this AWS Solution do?

Many Amazon Web Services (AWS) customers require a data storage and analytics solution that offers more agility and flexibility than traditional data management systems. A data lake is a new and increasingly popular way to store and analyze data because it allows companies to manage multiple data types from a wide variety of sources, and store this data, structured and unstructured, in a centralized repository.

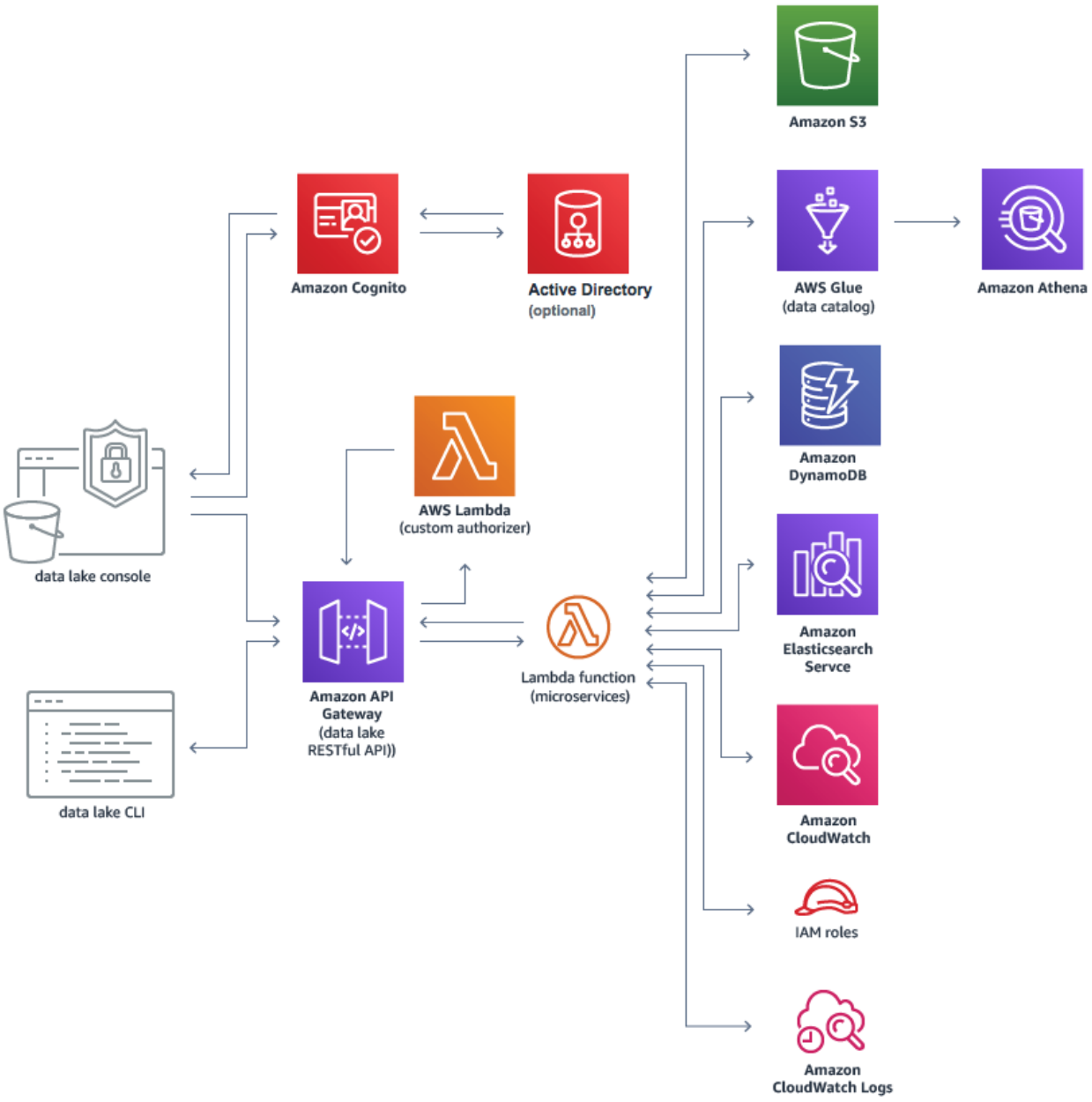
The AWS Cloud provides many of the building blocks required to help customers implement a secure, flexible, and cost-effective data lake. These include AWS managed services that help ingest, store, find, process, and analyze both structured and unstructured data. To support our customers as they build data lakes, AWS offers the data lake solution, which is an automated reference implementation that deploys a highly available, cost-effective data lake architecture on the AWS Cloud along with a user-friendly console for searching and requesting datasets.

Version 2.2 of the solution uses the most up-to-date Node.js runtime. Version 2.1 uses the Node.js 8.10 runtime, which reached end-of-life on December 31, 2019. To upgrade to version 2.2, you must deploy the solution as a new stack. For more information, see the deployment guide (https://amzn.to/aws-dl-solution-deployment).

AWS Solution overview

AWS offers a data lake solution that automatically configures the core AWS services necessary to easily tag, search, share, transform, analyze, and govern specific subsets of data across a company or with other external users. The solution deploys a console that users can access to search and browse available datasets for their business needs. The solution also includes a federated template that allows you to launch a version of the solution that is ready to integrate with Microsoft Active Directory.

The diagram below presents the data lake architecture you can deploy in minutes using the solution's implementation guide and accompanying AWS CloudFormation template.



View this solution online: https://amzn.to/aws-dl-dl-solution

Learn more about the AWS Lake Formation: https://amzn.to/aws-dl-lake-formation

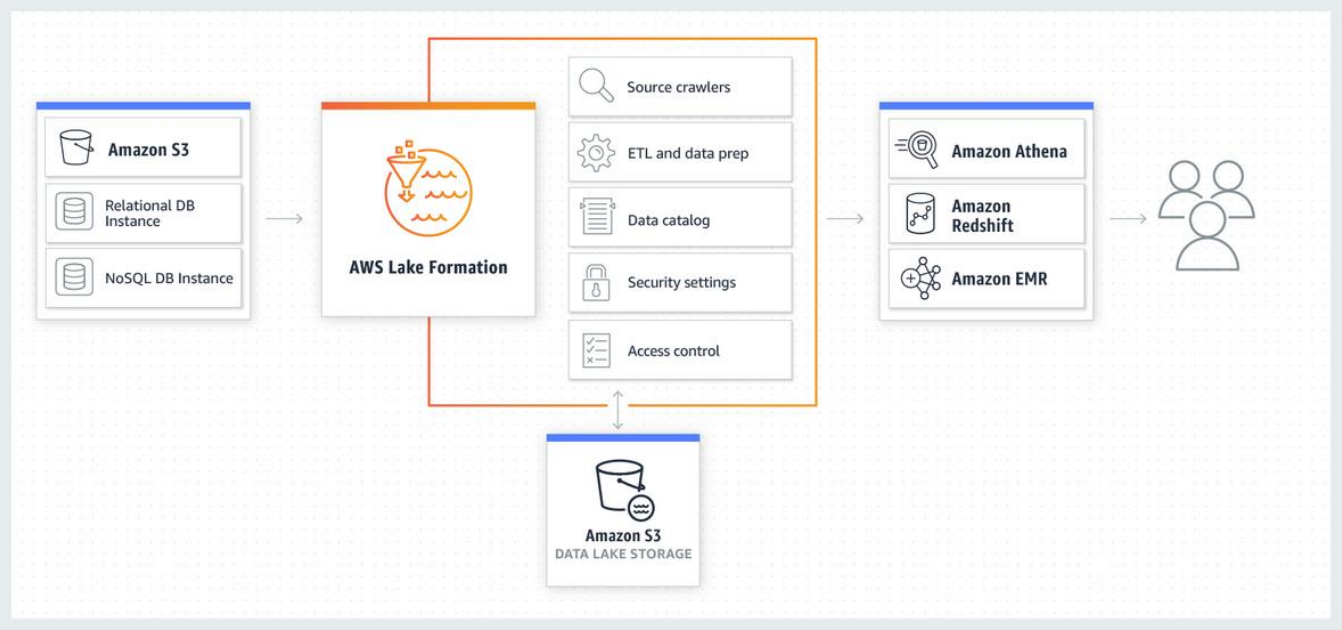
AWS Lake Formation is a service that makes it easy to set up a secure data lake in days. A data lake is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. A data lake enables you to break down data silos and combine different types of analytics to gain insights and guide better business decisions.

However, setting up and managing data lakes today involves a lot of manual, complicated, and time-consuming tasks. This work includes loading data from diverse sources, monitoring those data flows, setting up partitions, turning on encryption and managing keys, defining transformation jobs and monitoring their operation, re-organizing data into a columnar format, configuring access control settings, de-duplicating redundant data, matching linked records, granting access to data sets, and auditing access over time.

Creating a data lake with Lake Formation is as simple as defining data sources and what data access and security policies you want to apply. Lake Formation then helps you collect and catalog data from databases and object storage, move the data into your new Amazon S3 data lake, clean and classify your data using machine learning algorithms, and secure access to your sensitive data. Your users can access a centralized data catalog which describes available data sets and their appropriate usage. Your users then leverage these data sets with their choice of analytics and machine learning services, like Amazon Redshift, Amazon Athena, and (in beta) Amazon EMR for Apache Spark. Lake Formation builds on the capabilities available in AWS Glue.

How it Works

Lake Formation helps to build, secure, and manage your data lake. First, identify existing data stores in S3 or relational and NoSQL databases, and move the data into your data lake. Then crawl, catalog, and prepare the data for analytics. Then provide your users secure self-service access to the data through their choice of analytics services. Other AWS services and third-party applications can also access data through the services shown. Lake Formation manages all of the tasks in the orange box and is integrated with the data stores and services shown in the blue boxes.



Get started with AWS Lake Formation: https://console.aws.amazon.com/lakeformation/

Real-World Example

Haptik: Data Lake for Conversational AI

Learn how various AWS managed services have helped Haptik process more than 40 million messages every month, delivered by a large number of chatbots. Join in to understand Haptik's journey of real-time and batch analysis of chat sessions.

https://amzn.to/aws-dl-haptik

Read the full whitepaper online: https://amzn.to/aws-dl-flexible-storage

Abstract

Organizations are collecting and analyzing increasing amounts of data making it difficult for traditional on-premises solutions for data storage, data management, and analytics to keep pace. Amazon S3 and Amazon S3 Glacier provide an ideal storage solution for data lakes. They provide options such as a breadth and depth of integration with traditional big data analytics tools as well as innovative query-in-place analytics tools that help you eliminate costly and complex extract, transform, and load processes. This guide explains each of these options and provides best practices for building your Amazon S3-based data lakes.

Introduction

As organizations are collecting and analyzing increasing amounts of data, traditional on-premises solutions for data storage, data management, and analytics can no longer keep pace. Data siloes that aren't built to work well together make storage consolidation for more comprehensive and efficient analytics difficult. This, in turn limits an organization's agility, ability to derive more insights and value from its data, and capability to seamlessly adopt more sophisticated analytics tools and processes as its skills and needs evolve.

A data lake, which is a single platform combining storage, data governance, and analytics, is designed to address these challenges. It's a centralized, secure, and durable cloud-based storage platform that allows you to ingest and store structured and unstructured data, and transform these raw data assets as needed. You don't need an innovation-limiting pre-defined schema.You can use a complete portfolio of data exploration, reporting, analytics, machine learning, and visualization tools on the data. A data lake makes data and the optimal analytics tools available to more users, across more lines of business, allowing them to get all of the business insights they need, whenever they need them.

Unntil recently, the data lake had been more concept than reality. However, Amazon Web Services (AWS) has developed a data lake architecture that allows youto build data lake solutions cost-effectively using Amazon Simple Storage Service (Amazon S3) and other services.

Using the Amazon S3-based data lake architecture capabilities you can do the following:

- Ingest and store data from a wide variety of sources into a centralized platform.
- Build a comprehensive data catalog to find and use data assets stored in the data lake.
- Secure, protect,and manage all of the data stored in the data lake.
- Use tools and policies to monitor, analyze,and optimize infrastructure and data.
- Transform raw data assets in place into optimized usable formats.
- Query data assets in place.
- Use a broad and deep portfolio of data analytics, data science, machine learning, and visualization tools.
- Quickly integrate current and future third-party data-processing tools.
- Easily and securely share processed datasets and results.

The remainer of this paper provides more information about each of these capabilities. The figure below illustrates a smaple AWS data lake platform:

Read the full whitepaper online: https://amzn.to/aws-dl-flexible-storage

Real-World Example

HDFC Life: Event Driven Data Lake and Recovery

HDFC Life is one of the largest life insurance companies in India. For a company with more than 1.05 million individual policies, how do you create a real-time event-driven data

pipeline? Abey Alex, Solutions Architect from HDFC Life, tells us about the company's event-driven architecture and how complex aggregation and business logic can be built to process data in real time. He also shares insights on how the company's data lake has evolved over the last two years to leverage serverless frameworks, machine learning, and state machines.

https://www.youtube.com/watch?v=h0HE3bOEiMk