Building a data lake on Amazon Web Services (AWS)

Contents

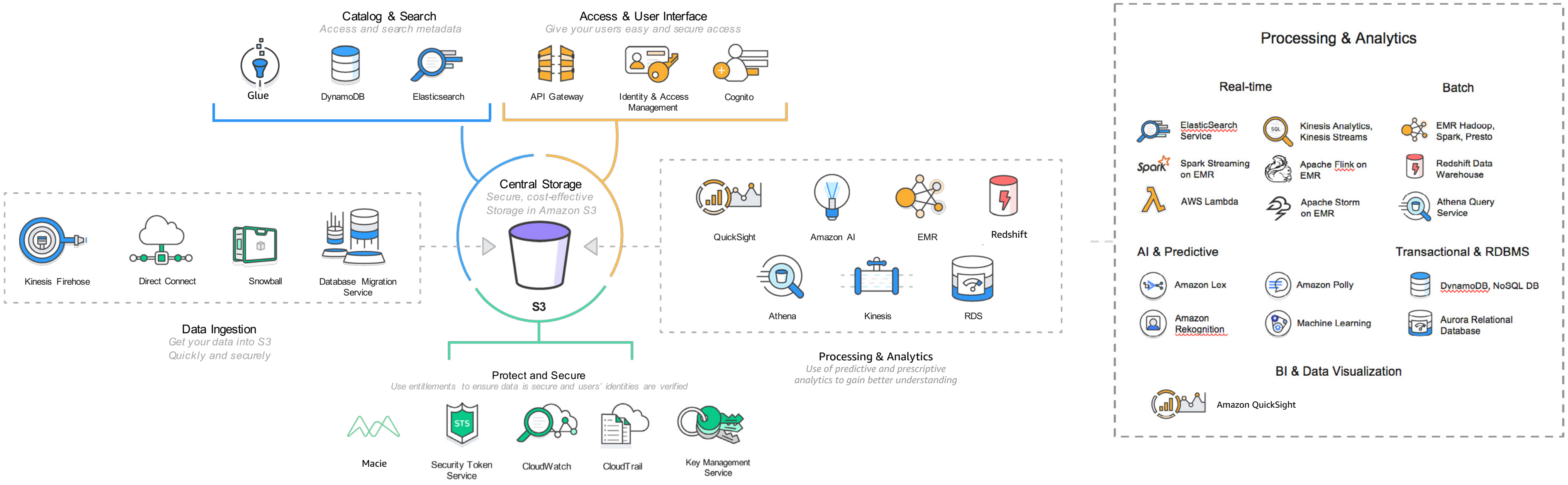
What is a data lake?

Today's organizations are tasked with managing multiple data types, coming from a wide variety of sources. Faced with massive volumes and heterogeneous types of data, organizations are finding that in order to deliver insights in a timely manner, they need a data storage and analytics solution that offers more agility and flexibility than traditional data management systems. A data lake is an architectural approach that allows you to store massive amounts of data into a central location, so it's readily available to be categorized, processed, analyzed, and consumed by diverse groups within an organization. Since data - structured and unstructured - can be stored as-is, there's no need to convert it to a predefined schema and you no longer need to know what questions you want to ask of your data beforehand.

A data lake should support the following capabilities:

- Collecting and storing any type of data, at any scale and at a low cost
- Securing and protecting all of data stored in the central repository
- Searching and finding the relevant data in the central repository
- Management frameworks to govern the data, including moving, transforming, and cataloging data
- Quickly and easily performing new types of data analysis on datasets
- Advanced engines to query and analyze data; and build, test, and run models in a variety of ways, including machine learning and AI

A data lake isn't meant to replace your existing data warehouses, rather complement them. If you're already using a data warehouse, or are looking to implement one, a data lake can be used as a source for both structured and unstructured data, which can be easily converted into a well-defined schema before being ingested into your data warehouse.



Benefits of a data lake on AWS

According to an Aberdeen survey[1], organizations that implement data lakes successfully generate business value from their data and outperform their peers by nine percent in organic revenue growth.

To help organizations get the most value from their data, AWS offers a portfolio of services to build and host data lakes in the cloud. Data lakes on AWS can help you securely store data from a variety of new sources like log files, click-streams, social media, and IoT devices.

The benefits of hosting a data lake on AWS, include:

Cost-effective data storage

You can build a centralized repository for all of your data using Amazon Simple Storage Service (Amazon S3), which provides cost-effective and durable storage that allows you to store nearly unlimited amounts of data of any type, from any source. You also gain the flexibility to apply schemas for data analysis on demand because storing data in Amazon S3 doesn't require upfront transformations. This enables you to more easily answer new questions as they come up and improve the time-to-value.

Security and compliance

When you host your data lake on AWS, you gain access to a highly secure cloud infrastructure and a deep suite of security offerings like Amazon Macie, a security service that uses machine learning to automatically discover, classify, and protect sensitive data. As an AWS customer, you will benefit from a data center and network architecture built to meet the requirements of the most security-sensitive organizations. AWS also actively manages dozens of compliance programs in its infrastructure, helping organizations to easily meet compliance standards such as PCI DSS, HIPAA, and FedRAMP.

Easy data collection and ingestion

There are a variety of ways to ingest data into your data lake, including services such as Amazon Kinesis, which enables you to ingest data in real time; AWS Snowball, a secure appliance AWS sends you for ingesting data in batches; AWS Storage Gateway, which enables you to connect on-premises software appliances with your AWS storage; or through AWS Direct Connect, which gives you dedicated network connectivity between your data center and AWS.

Categorizing and managing your data

As data is used in different platforms, ETL (Extract, Transform, Load) is an important function to ensure that it is moved and understood properly. AWS Glue is an ETL engine you can use to understand data sources, prepare data, and load it reliably to data stores. AWS Glue discovers your data and stores the associated metadata (e.g. table definition and schema) in the AWS Glue Data Catalog. Once cataloged, your data is immediately searchable, can be queried, and is available for ETL.

Insight into analytics

AWS offers a comprehensive suite of cloud-native solutions to query and analyze your data. By using advanced analytical approaches on data from new sources such as IoT devices stored in your data lake, you can identify business opportunities at an accelerated pace and make informed decisions.

Common analytics approaches are covered, including Amazon Redshift for data warehousing; Amazon Athena for SQL querying on demand; Amazon EMR for running popular open source frameworks such as Hadoop, Spark, Presto, Flink, and HBase; Amazon QuickSight for business intelligence; and Amazon Elasticsearch Service for logs and text. There is no need to migrate data to different operating environments, allowing you to avoid the accompanying overhead, costs, effort, or delays.

Amazon EMR, Amazon Redshift Spectrum, and Amazon Athena allow you to query data directly in Amazon S3 and, with AWS Glue, they all can share the same data catalog.

Additionally, artificial intelligence (AI) and machine learning are becoming increasingly popular tools for building smart applications such as predictive analytics and deep learning. To make it more accessible, Amazon Machine Learning abstracts away from the algorithms with wizards, APIs, and guidance. AI services include Amazon Polly for text-to-speech, Amazon Lex for natural language processing and conversational bots, and Amazon Rekognition for image identification and classification.

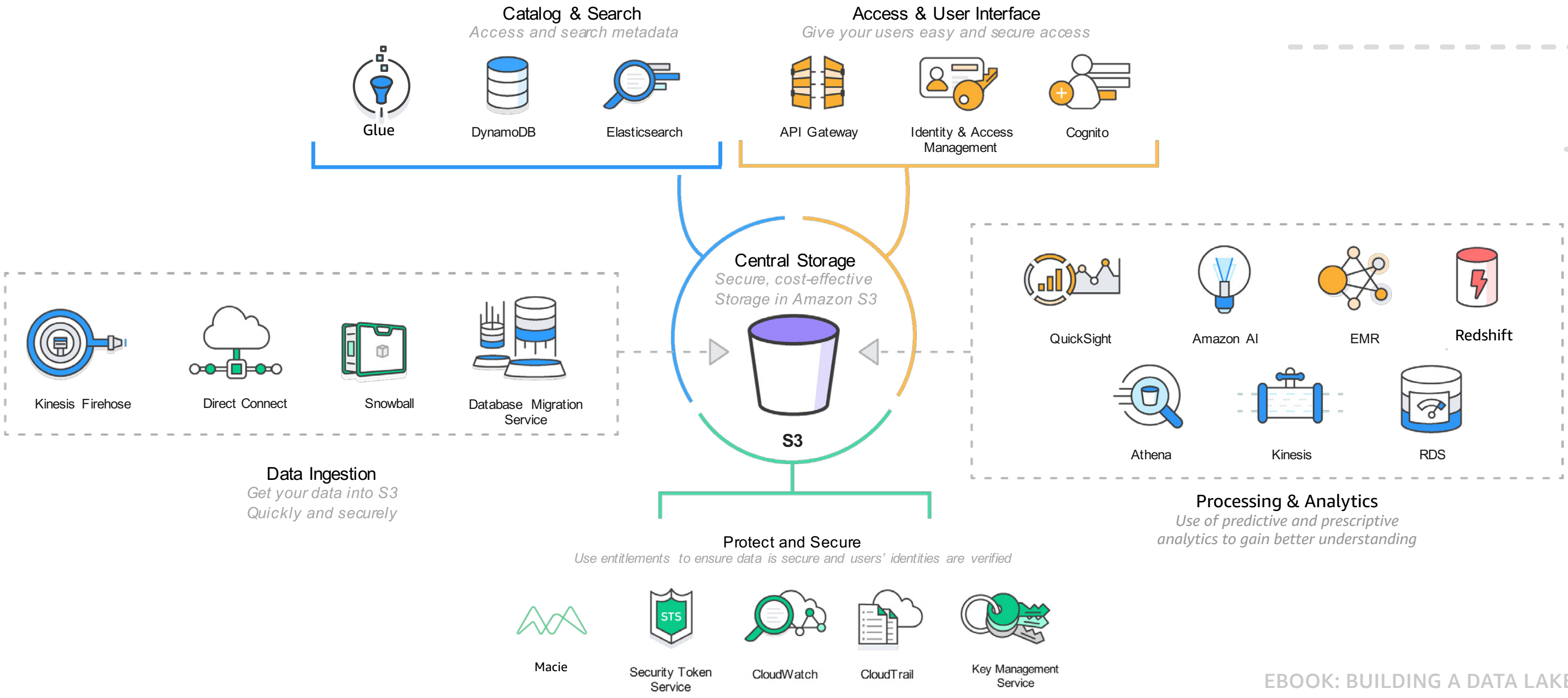
[1] Michael Lock, Aberdeen Group, *Angling for Insight in Today's Data Lake*, October 2017

Building a data lake on AWS

A data lake solution on AWS, at its core, leverages Amazon S3 for secure, cost-effective, durable, and scalable storage. You can quickly and easily collect data into Amazon S3 from a wide variety of sources by using services like AWS Snowball (ideal for migrating/moving data in bulk) or Amazon Kinesis (an excellent solution for ongoing collection/ ingestion of data). Amazon S3 also offers an extensive set of features to help you provide strong security for your data lake, including access controls and policies, data transfer over SSL, encryption for data at rest and in motion, logging and monitoring, and more.

For the management of the data, you can leverage services such as AWS Glue, Amazon DynamoDB, and Amazon ElasticSearch to catalog and index the data in Amazon S3. Using AWS Lambda functions that are directly triggered by Amazon S3 in response to events such as new data being uploaded, you easily can keep your catalog up-to-date. With Amazon API Gateway, you can create an API that acts as a “front door” for applications to access data quickly and securely by authorizing access via AWS Identity and Access Management (IAM) and Amazon Cognito.

For analyzing and accessing the data stored in Amazon S3, AWS provides fast access to flexible and low-cost services, like Amazon EMR, Amazon Redshift with Redshift Spectrum, Amazon Athena, and Amazon AI services, so you can rapidly scale any analytical solution. Example solutions include data warehousing, clickstream analytics, fraud detection, recommendation engines, event-driven ETL, and internet-of-things processing. By leveraging AWS, you can easily provision exactly the resources and scale you need to power any big data applications, meet demand, and improve innovation.


Catalog & Search
Access and search metadata
Glue
DynamoDB
Elasticsearch
Access & User Interface
Give your users easy and secure access
API Gateway
Identity & Access Management
Cognito
Central Storage
Secure, cost-effective Storage in Amazon S3
S3
Kinesis Firehose
Direct Connect
Snowball
Database Migration Service
Data Ingestion
Get your data into S3 Quickly and securely
QuickSight
Amazon AI
EMR
Redshift
Athena
Kinesis
RDS
Processing & Analytics
Use of predictive and prescriptive analytics to gain better understanding
Protect and Secure
Use entitlements to ensure data is secure and users' identities are verified
Macie
STS
Security Token Service
CloudWatch
CloudTrail
Key Management Service

Featured data lake Partners

Qubole

Qubole is passionate about making data-driven insights easily accessible to anyone. Its solution is natively designed for AWS and compatible with its storage, compute, security, and other key architectural elements. Qubole Data Service helps optimize data lakes on AWS for a streamlined analytics approach that allows data teams to focus on business outcomes rather than on managing the platform.

Cognizant

Cognizant Out of the Box Data Lake for the Enterprise streamlines the process of deploying a cloud-based data lake by combining AWS elastic infrastructure, Talend big data integration software, and pre-configurations from Cognizant. Deployed as an enterprise-grade data lake on AWS, Cognizant's solutions provide a unified platform to consume and process historical and real-time data for analytics.

ATTUNITY

Attunity offers data integration and a big data management platform that helps accelerate data delivery and availability, automate data readiness for analytics, and optimize data management. Attunity offers heterogeneous data availability solutions that streamline and accelerate the process of ingesting both streaming and historical data to data lakes on AWS. Attunity solutions also address legacy mainframe systems, modern databases, data lakes on AWS, SAP, Hadoop, and real-time messaging systems on AWS such as Amazon Simple Notification Service (SNS) and Amazon Simple Queue Service (SQS), as well as messaging systems on Kafka and on-premises solutions.

databricks

Databricks Unified Analytics Platform, powered by Apache Spark, enables collaboration between data engineering and business intelligence teams, allowing them to work efficiently with streaming data and historical data stored in data lakes on Amazon S3. Users achieve fast time-to-value with Databricks by creating analytic workflows that go from ETL and interactive exploration to production. The company makes it easier for its users to focus on their data by providing a fully managed, scalable, and secure cloud infrastructure that reduces operational complexity and total cost of ownership.

AWS Case Study: FINRA

FINRA, the Financial Industry Regulatory Authority, is dedicated to investor protection and market integrity. It regulates one critical part of the securities industry: the brokerage firms doing business with the public in the United States. To respond to rapidly changing market dynamics, FINRA moved about 90 percent of its data volumes to AWS, capturing, storing, and analyzing a daily influx of 37 billion records.

FINRA hosts its data lake and analytics platform on AWS, allowing its analysts to perform discovery on petabytes of trade data and identify instances of potential fraud, market manipulation, and insider trading. The company centralizes all its data in Amazon S3 and uses analytics services such as Amazon EMR and Amazon Redshift, which enable its data scientists to apply machine learning techniques to develop and test new surveillance patterns and improve investor protection. By migrating to AWS, FINRA has created a flexible platform that can adapt to changing market dynamics while providing its analysts with the tools to interactively query multi-petabyte data sets.

To learn more about how FINRA performs trade surveillance and analytics on its multi-petabyte data lake on AWS, watch the recorded AWS re:Invent session.

Explore FINRA's cloud journey on AWS >>

> “At FINRA, we chose AWS because we wanted to be able to deliver innovation at a much larger scale and much more rapidly to our core business.”

Saman Michael Far
Senior Vice President of Technology
FINRA

Getting started

For more information about data lakes on AWS, visit:

> [Data lakes on AWS]

Learn more:

> Big data on AWS

About AWS

To learn more about AWS, visit:

> http://aws.amazon.com