Building AWS based Enterprise Data Lakes

Generating business value from data to outperform your competition.

ANJAN BANERJEE

LEAD ARCHITECT

Contents

Introduction

Organizations collecting, processing, and analyzing large amounts of data often face two major challenges. First, data is traditionally housed on-premises which presents a scalability challenge when data volume grows rapidly. Second, data tends to be stored, processed, and analyzed in siloed data warehouses limiting accessibility and presenting scalability issues during data loading. These two challenges result in higher operational costs and lower efficiency in analyzing data, limiting the insights to be gained from the data. Forward leaning companies are harnessing the power of the cloud to consolidate these data siloes in what are called data lakes.

What is a data lake and how is it different from a data warehouse? The fundamental difference is the fact that the data stored in a data lake does not have a defined purpose. Raw data flows into a data lake, sometimes with a specific future use in mind and sometimes just to have on hand. This results in data lakes having less organization and less filtration than data warehouses. Since data warehouses only contain processed data transformed for a certain purpose, they serve only that purpose.

The distinguishing features of data lakes are:

- Storage of raw data (structured/unstructured).
- Usage of low-cost storage.
- Capability to quickly query large volumes of data.
- No fixed data model defined.
- Use by data scientists/business analysts to curate the data for profiling, data discovery and analytical models like machine learning and predictive models.
- Data dumped does not have to be used immediately and may be required for future analysis.

Amazon Web Services has multiple services that can be used individually or in combination to kick-start the building of a data lake for any organization. These include:

- **Amazon S3** (Simple Storage Service) provides object storage at a very low cost and is ideal for storing both structured and unstructured data from multiple sources.
- **Amazon Athena** is an interactive serverless query service, which works with the data stored on Amazon S3. An organization does not have to manage any infrastructure to use this.
- **Amazon Glue Data Catalog** is a service which we will be talking about which helps identify the schema of the data present in the storage automatically.
- **AWS Lake Formation**, a new service released by Amazon, makes it easy to ingest, clean, catalog and secure your data and making it available for the developers and data scientists for analysis.
- **Amazon EMR** is a cloud-based solution to run map reduced jobs on an on-demand pay as you use the Hadoop cluster.
- **Amazon Sagemaker** is a service which enabled data scientists to build, train, and deploy machine learning models quickly.

In this white paper, you'll learn technologies and AWS services to build a cloud-based data lake that's right for your needs.

Amazon Simple Storage Service

S3 has been the most reliable and versatile service offered by Amazon and is designed to be 99.999999999% (11 9's) durable. With the help of the different levels of storage options/lifecycle and multi-region availability options, this is the ideal turf to start building a data lake solution.

Loading data to S3 and Lifecycle

Amazon S3 has been around for more than a decade and almost all market-leading data transformation tools have native connectors to load data into S3. Amazon also provides AWS DataSync which has agents to transfer data sitting in on-premise NAS storage via TLS encrypted channels.

S3 Intelligent tiering provides low latency, high throughput performance, and moves the data to IA (infrequent access) if the data has not been accessed for more than 30 days. This is ideal for building a data lake, as this optimizes the cost of storing a large volume of data.

With the help of S3 Lifecycle, the storage class of data can be moved to Glacier or Glacier Deep Archive based on the utilization of historical data.

Storage Class	Data Access Strategy	Cost
S3 Intelligent Tiering	Standard Storage Frequent access	$0.023/GB
S3 Standard IA	Data used quarterly	$0.0125/GB
Glacier Storage	Data rarely used (5~10yrs historical)	$0.004/GB

Table 1. Storage Costs

Data Encryption and Security

For a data lake designed to store all the data of an organization, data encryption at rest is a key aspect with respect to security, the next being access control.

Amazon provides multiple options to encrypt data at rest, but we will focus on AWS KMS (Key Management Service). AWS-KMS provides you the option to encrypt your data with Server-side encryption with a key which has restricted access. A user may have access to list the objects stored on S3 but may not have access to decrypt the data to read the content of the objects.

Bucket policies are another key aspect to restrict access to specific users with the ability to even whitelist certain IPs to be able to access the data stored on S3.

This enables any organization to lock down their data lake as they choose and completely secure their data lake in the cloud.

Alternative Data Lake Options Comparison

Topics	Amazon Athena and S3	Amazon Aurora RDS	Redshift
Infrastructure	Serverless	AWS Managed	Customer Managed
Initialization Time	Few minutes	Few hours	Few days
Cost	Based on Bytes of Data Scanned	Server Cost Hourly	Server Cost Hourly
Partitioning for query optimization	Auto and Manual Partitioning	Manual Partitioning	Manual Partitioning – Distribution Key is important
Flat File support	CSV, TSV, JSON, Snappy, Zlib, LZO, GZIP, arrays, maps, and structs.	No Support	CSV, TSV, JSON
Storage System	S3	EBS or SSD	EBS or SSD
Query Performance Data Volume Handling	Can scale up to 50TB of data, single file size no greater than 1GB	Can scale up to a few TBs	Can scale up to a few TBs
Scalability	Auto Scalable	Auto Scalable, can have dedicated Reader nodes	Highly scalable
Maintenance	Not required	Auto maintained	Manual maintenance

Table 2. Alternative Data Lake Options Compared

Amazon Athena

For analysts, data scientists, and engineers who crunch data to derive insights, and work to continuously improve products, the performance of queries against a data warehouse is important. Being able to run more queries and get results faster improves their productivity. This quest for faster queries led Facebook to develop its Presto query engine to help their data analysts run interactive queries on its large data warehouse in Apache Hadoop.

Athena, a serverless query service by Amazon, is based on Presto, and is incredibly flexible and helpful data profiling and discovery operations on raw data being stored in a data lake.

Defining Schema

Athena provides a SQL-like query Interface on top of the data stored on S3. To be able to query the data, the schema of the data stored on S3 needs to be defined.

You can create a Glue Database, define your table schema and simply point the table to the S3 location and start querying your data. This process can also be automated, and the schema can be detected programmatically with the help of another AWS Service, Glue Crawler. A Crawler can be created by pointing it to the exact location the data stored and let the crawler create the Catalog table for you. This works the best when the data stored is structured.

The Glue Data Catalog Table is the single location where you would have to define the schema and can be used as the Hive Table metadata to be used on EMR Hadoop clusters which will be discussed below.

EMR Cluster

Hadoop clusters have traditionally been expensive and always limited when running heavy workloads due to hardware scalability. EMR (Elastic Map Reduced) Cluster is a scalable Hadoop cluster which can be easily spun up or down based on demand requirements. Apache Spark code can be run on data stored on S3 using EMR. Just like all Amazon services, EMR also provides the feature of pay for exactly what you use and EMR also has the benefit of using Spot Instances, reducing the cost of owning a cluster even further.

The data for an EMR cluster sits on S3, the Hive Catalog sits on AWS Glue. EMR provides built-in Jupyter Notebook functionalities which help developers create Spark, PySpark or Python code and to run their code on Map Reduced Clusters. With the help of a few more AWS Services the data pipeline can be automated, and daily runs can become event-based.

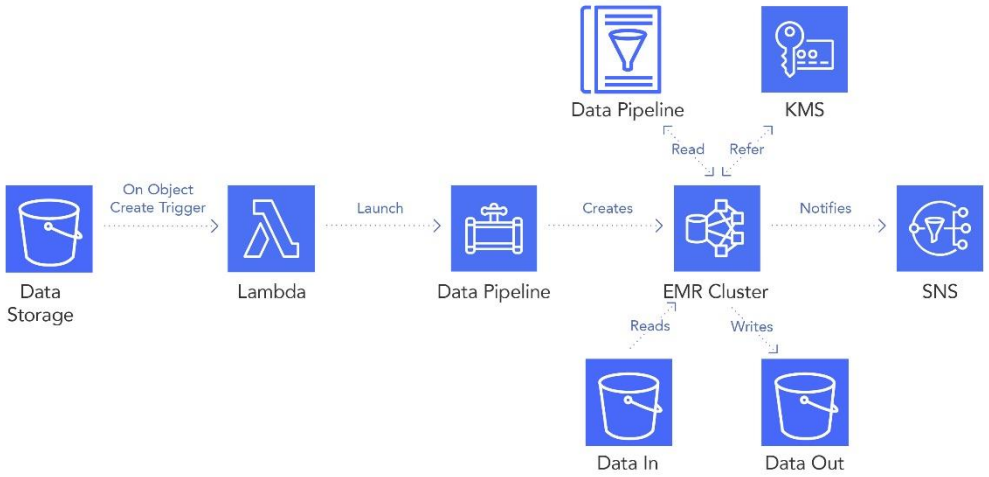

Figure 1. EMR Cluster Design

Amazon Sagemaker

Developing, training and tuning machine learning algorithms have been incredibly time-consuming if done locally. Sagemaker is a fully managed service enabling data scientists to create machine learning models for data hosted on S3. They can build, train, tune, and deploy the models without worrying about infrastructure. It also provides a catalog of commonly used machine learning algorithms which can just be fed with your cleaned data to generate insights and actionable items.

With the recent introduction of the Sagemaker Studio, the online IDE platform to create ML Models, it has never been easier to create and tune models. It also facilitates collaboration as people can work on the same code in parallel.

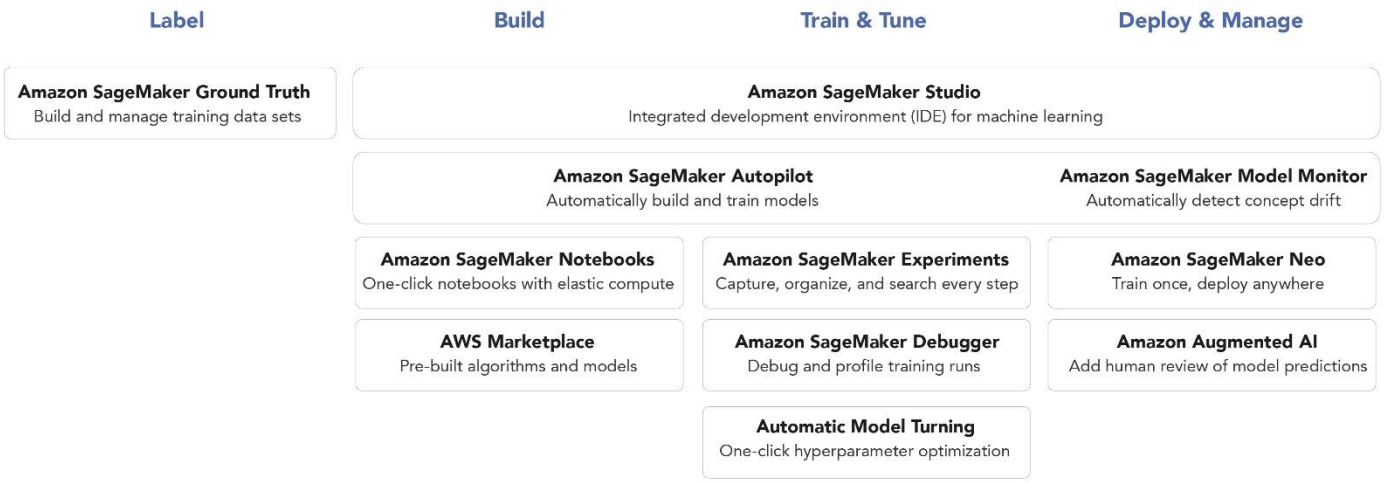

Figure 2. Sagemaker Functions

AWS Lake Formation

AWS Lake Formation integrates with underlying AWS security, storage, analysis, and machine learning services and automatically configures them to comply with your centrally defined access controls using an easy to use web console. It uses S3 as its base storage and has Glue Catalog running on top in an ad-hoc or scheduled fashion constantly keeping your data catalog up to date.

Defining access permissions, including table, row, and column level control and encryption policies for data at rest and in motion is also made easy using the services of Lake Formation. You can then use a wide variety of AWS analytic and machine learning services to access your data lake. All access is secured, governed, and auditable.

The below diagram quickly gives you the idea of the services/features provided by Lake Formation.

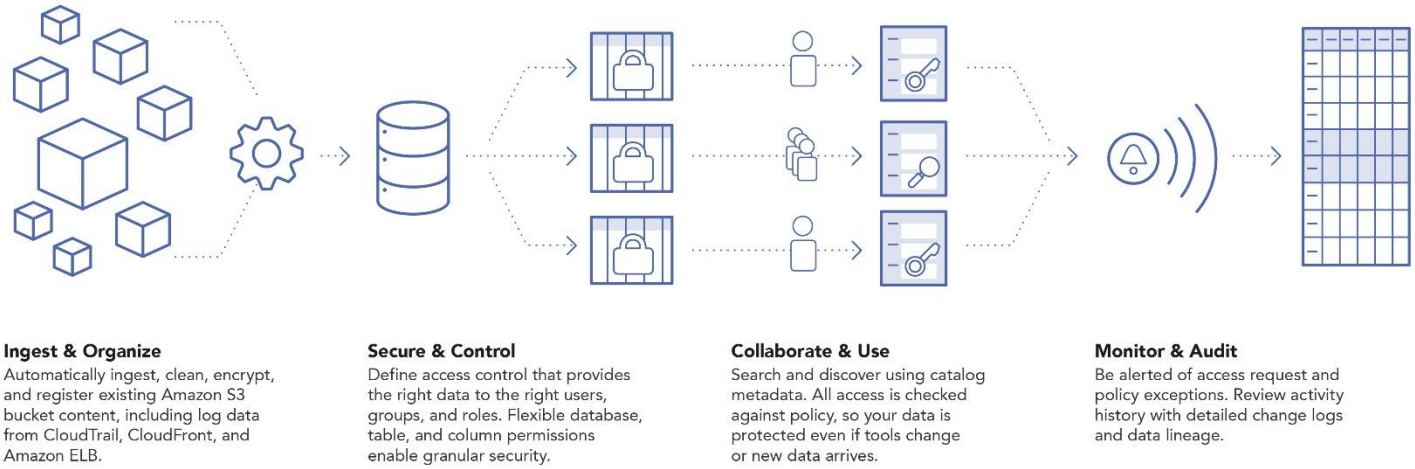

Figure 3. Amazon Lake Formation Features

No organization is the same. Based on its own needs, an organization needs to choose its ideal approach to build its data lake to gain an edge over the competition. Amazon's Lake Formation provides you the capabilities that any organization would want, based on their use case and complexity:

- Store Big Data with low cost.
- Detect schema automatically.
- Create Catalog for all your data.
- Security at row/column level with federated and non-federated users.
- SQL Layer with JDBC Connector enabling reporting tools to query the data.

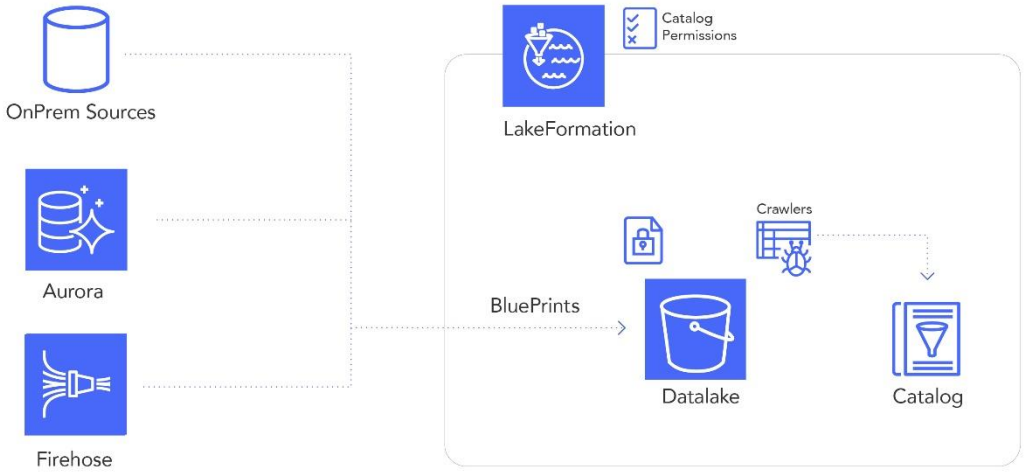

Figure 4. Amazon Lake Formation Architecture

Conclusion

With the AWS Services discussed above, organizations can design and build a data lake to meet their needs in a relatively short period of time. A data lake built on AWS provides:

- Lower operational costs, both infrastructure, and license.
- Higher efficiency in managing and analyzing data.
- Flexibility to innovate without being bound by rigid data structures.
- Secure, access to data protecting its integrity.

To begin the journey to implementing a data lake, consider the following steps:

- Define your business objectives as this will help guide your technology decisions.
- Identify your security and compliance requirements.
- Draw on external resources with experience in designing AWS-based data lakes.

About the Author

Anjan Banerjee, lead architect at Starschema, has been working with Cloud Technologies and Data warehouses for ten years. He has extensive experience in building data orchestration pipelines, designing multiple cloud-native solutions and solving business-critical problems for many multinational companies. He believes in the concept of infrastructure as code as a means to increase the speed, consistency, and accuracy of cloud deployments. Anjan holds frequent training sessions on data munging and various self-service analytics/transformation tools like Talend and Alteryx.

About Starschema

At Starschema we believe that data has the power to change the world and data-driven organizations are leading the way. We help organizations use data to make better business decisions, build smarter products, and deliver more value for their customers, employees and investors. We dig into our customers toughest business problems, design solutions and build the technology needed to compete and profit in a data-driven world.