Data Lake Foundation on the AWS Cloud

Demo and Walkthrough

September 2019

47Lining, a Hitachi Vantara Company

AWS Quick Start Team

What you’ll accomplish: This demo helps you explore foundational data lake capabilities such as search, transforms, queries, analytics, and visualization. The demo is based on a sample data set from ECommCo, a fictional company that sells multiple categories of products through its ecommerce website, ECommCo.com.

Estimated time: 50 minutes for deployment, 20 minutes for walkthrough

Cost: You are responsible for the cost of the AWS services used while running this demo. There is no additional cost for using the demo.

Experience with AWS: Advanced

Experience with data lakes: Beginner

Prerequisites: An AWS account (sign up at https://aws.amazon.com) and a key pair in your preferred region.

Contents

Overview...........................................................................................................................................3

Demo components and flow .............................................................................................................4

Technical specifications....................................................................................................................5
Deploy the demo .......................................................................................................................... 6

Explore the data lake on AWS .................................................................................................. 7

Access the demo wizard ............................................................................................................. 8

Create curated datasets ............................................................................................................. 10

Start the flow from a streaming data provider. Stream data ..................................................... 10

Start an AWS Glue ETL job and observe crawlers and databases ......................................... 11

Use Elasticsearch to index data ............................................................................................... 11

Query and analyze data with Amazon Redshift Spectrum ....................................................... 11

Retrieve data with Athena queries ............................................................................................ 12

Explore machine learning with Amazon SageMaker ................................................................. 12

Gain business insights with Amazon QuickSight ....................................................................... 12

Delete the demo ......................................................................................................................... 12

Next steps .................................................................................................................................. 13

FAQ .......................................................................................................................................... 13

Additional resources .................................................................................................................. 13

Document revisions .................................................................................................................... 14

This demo was created by 47Lining in collaboration with solutions architects at Amazon Web Services (AWS) for evaluation or proof-of-concept (POC) purposes on the AWS Cloud.

The demo automatically deploys a data lake into your AWS account with sample data. After the demo is up and running, you can use the walkthrough in the following pages for a guided tour of product features.

**Quick Start option**  This demo sets up an evaluation environment. For production-ready deployments, use the Data Lake Foundation on the AWS Cloud with AWS Services Quick Start.
Overview

A data lake is a repository that holds a large amount of raw data in its native (structured or unstructured) format until the data is needed. Storing data in its native format enables you to accommodate any future schema requirements or design changes.

Increasingly, valuable customer data sources are dispersed among on-premises data centers, software as a service (SaaS) providers, partners, third-party data providers, and public datasets. Building a data lake on AWS offers a foundation for storing on-premises, third-party, and public datasets at low prices and high performance. A portfolio of descriptive, predictive, and real-time agile analytics built on this foundation can help answer your most important business questions, such as predicting customer churn and propensity to buy, detecting fraud, optimizing industrial processes, and content recommendations.

This demo is for users who want to get started with AWS-native components for a data lake in the AWS Cloud. When this foundational layer is in place, you may choose to augment the data lake with ISV and SaaS tools.

The demo builds a data lake foundation that integrates AWS services such as Amazon Simple Storage Service (Amazon S3), Amazon Redshift, Amazon Kinesis, Amazon Athena, AWS Glue, Amazon Elasticsearch Service (Amazon ES), Amazon SageMaker, and Amazon QuickSight. The data lake foundation provides these features:

- **Data submission**, including batch submissions to Amazon S3 and streaming submissions via Amazon Kinesis Data Firehose.
- **Ingest processing**, including data validation, metadata extraction, and indexing via Amazon S3 events, Amazon Simple Notification Service (Amazon SNS), AWS Lambda, Amazon Kinesis Data Analytics, and Amazon ES.
- **Dataset management** through Amazon Redshift transformations and Kinesis Data Analytics.
- **Data transformation, aggregation, and analysis** through Amazon Athena, Amazon Redshift Spectrum, and AWS Glue.
- **Building and deploying machine learning models** with Amazon SageMaker.
- **Search**, by indexing metadata in Amazon ES and exposing it through Kibana dashboards.
- **Publishing** into an S3 bucket for use by visualization tools.
- **Visualization** with Amazon QuickSight.
Demo components and flow

This demo builds a simplified version of the Quick Start architecture for a foundational data lake that includes these components:

- An AWS Cloud environment that spans two Availability Zones.
- A virtual private cloud (VPC) configured with public and private subnets according to AWS best practices, to provide you with your own virtual network on AWS.
- An internet gateway to allow access to the internet. This gateway is used by the bastion host to send and receive traffic.
- In the public subnets, managed NAT gateways to allow outbound internet access for resources in the private subnets.
- In the public subnets, a Linux bastion host in an Auto Scaling group to allow inbound Secure Shell (SSH) access to EC2 instances in public and private subnets.
- In the private subnets, Amazon Redshift for data aggregation, analysis, transformation, and creation of curated and published datasets.
- AWS Identity and Access Management (IAM) roles to provide permissions to access AWS resources; for example, to permit Amazon Redshift and Amazon Athena to read and write curated datasets.
- An Amazon SageMaker instance, which you can access by using AWS authentication.
- Integration with other AWS services such as Amazon S3, Amazon Athena, AWS Glue, AWS Lambda, Amazon ES with Kibana, Amazon Kinesis, and Amazon QuickSight.
- A sample dataset from ECommCo, a fictional company that sells products in multiple categories through its ecommerce website, ECommCo.com.
Figure 1 illustrates the process flow and features you’ll explore with this walkthrough.

![Diagram of process flow for a data lake foundation on AWS]

**Figure 1: Process flow for a data lake foundation on AWS**

**Technical specifications**

The demo provisions and configures the following resources. You won’t be able to customize these resources in the demo environment; for customization options, launch the Quick Start instead.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kinesis data stream</strong></td>
<td>The demo provisions a Kinesis data stream names streaming-submissions.</td>
</tr>
<tr>
<td><strong>S3 prefix for the Kinesis data stream</strong></td>
<td>The demo points to the S3 prefix streaming-submissions.</td>
</tr>
<tr>
<td><strong>Web server instances</strong></td>
<td>The demo provisions only one instance for the demo walkthrough.</td>
</tr>
</tbody>
</table>
### Deploy the demo

**Note** These instructions reflect the older version of the AWS CloudFormation console. If you’re using the redesigned console, some of the user interface elements might be different.

1. **Create an AWS account, or sign into your existing account.**
   
   If you don’t already have an AWS account, create one at [https://aws.amazon.com](https://aws.amazon.com) by following the on-screen instructions.

2. **Sign up to launch the demo.**
   
   After you answer a few questions and submit the sign-up form, you will be redirected to the AWS CloudFormation console. In the console, you will need to provide the requested information to launch the demo.

3. **In the Parameters section, complete the following:**

<table>
<thead>
<tr>
<th>For parameter</th>
<th>Select or enter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability Zones</td>
<td>Select two Availability Zones.</td>
</tr>
<tr>
<td>Wizard Password</td>
<td>Enter a password for the wizard that will guide you through the demo, as explained in the next section. The password consists of 8-64 ASCII characters and must contain one uppercase letter, one lowercase letter, and one number.</td>
</tr>
<tr>
<td>Remote Access CIDR</td>
<td>Enter the CIDR IP range that is permitted to SSH into the bastion host instance and access Amazon ES. We recommend that you set this value to a trusted IP range. You can use <a href="http://checkip.amazonaws.com/">http://checkip.amazonaws.com/</a> to check your IP address. This parameter must be in the form x.x.x.x/x (e.g., 96.127.8.12/32, YOUR_IP/32).</td>
</tr>
<tr>
<td>Redshift Password</td>
<td>Enter a password for the master user account for the Amazon Redshift cluster. The password must contain 8-64 printable ASCII characters, excluding: /, &quot;, , , and @. It must contain one uppercase letter, one lowercase letter, and one number.</td>
</tr>
</tbody>
</table>

   **Note** If you want to customize the configuration further, please deploy the [Quick Start](https://aws.amazon.com/quickstart) instead.
4. On the **Create Stack** page, under **Capabilities**, select the two check boxes, to acknowledge that the template will create IAM resources and might require the capability to auto-expand macros.

5. Choose **Create**.

   The demo takes about 50 minutes to deploy.

6. Your stack appears in the list of AWS CloudFormation stacks, with a status of **CREATE_IN_PROGRESS**. When the status is **CREATE_COMPLETE**, the demo is ready to use.

7. Follow the step-by-step instructions in the next section to explore the features of the demo.

### Explore the data lake on AWS

**Time to complete:** 20 minutes

This demo provides a wizard for exploring the architecture and features of the data lake on AWS. The wizard includes ten steps, each of which demonstrates and explains a particular data lake feature. For example, step 2 of the wizard walks you through the process for promoting data from the S3 submissions bucket to the curated datasets bucket, step 3 demonstrates how to start the flow from a streaming data provider, and so on, all within your AWS account.

**Steps**

- [Access the demo wizard](#)
- [Curated Dataset](#)
- [Streaming Data](#)
- [Glue](#)
- [Elasticsearch](#)
- [Redshift Spectrum](#)
- [Athena](#)
- [SageMaker](#)
- [QuickSight](#)
- [Learn More](#)
Access the demo wizard

1. Open the AWS CloudFormation console at

2. In the Outputs tab, choose the URL for DataLakeWizardURL, and open it in a web browser. This will open the login page for the wizard.

3. Log in to the wizard by entering the user name DataLakeUser and the password you specified when you deployed the demo.

4. Read the brief introduction, and then choose **Begin walk-through** to start your exploration of the data lake.
Step 1 - Get started

This Quick Start Walk-Through Guide is intended to walk you through a Data Lake reference architecture and AWS services used within. To demonstrate the flow in your AWS account, we are going to use a sample dataset from a fictional company, EcommCo. EcommCo sells products in multiple categories via its e-commerce website ecommco.com. Its business users would like to consume key insights via business intelligence reports and dashboards to help answer key questions like, “What are our top selling products by store?” and “What is the customer lifetime value?”

EcommCo’s Data Lake in AWS provides data ingestion, real-time analytics over a continuous stream of data, batch analytics using all data available in a Data Lake, ad-hoc analytics for ease of exploring unknown data and visualization, so analytics can be easily understood by key stakeholders.

The diagram below provides a high-level overview of EcommCo's Data Lake in AWS. All of the AWS resources illustrated below were deployed to your account when you launched the Quick Start. As we step through this guide, sample data from EcommCo will be ingested into your account. After the demonstration, you can delete the data and begin using the Quick Start architecture with your own data.

Figure 2: Getting started with the wizard
Create curated datasets

In this step, you’ll create curated datasets by copying data submissions (Orders, Customers, Products) into a curated dataset bucket. You’ll transform the Demographics data submission to the curated dataset bucket, and load Orders to the orders table in Amazon Redshift, and Customers to the customers table in Amazon Redshift.

The wizard guides you through these steps:

1. Understand the principles of data organization within EcommCo’s data lake.
2. Review the diagram that describes datasets contributed by data providers.
3. To transform submissions into curated datasets, choose Create Curated Datasets.
4. To see the curated datasets in Amazon S3, choose the link S3 in your AWS Management Console and review the buckets listed in the web application.

Start the flow from a streaming data provider. Stream data

In this step, you’ll stream sample data and observe the flow of this data throughout the data lake. You’ll create two Amazon Kinesis Data Analytics applications: clean-order-app and aggregate-orders-app. The enabled data provider streams Orders data to Kinesis Data Firehose. Streamed orders are cleaned and analyzed, and results are published to Amazon ES, persisted in the S3 submissions bucket, and validated and loaded to Amazon Redshift.

In the wizard:

1. Demonstrate support for streaming of sample data using a Kinesis data stream – review the diagram.
2. Run streaming submissions by choosing Create Kinesis applications and start data stream.
3. Check that Kinesis data analytics applications are running by choosing Visit Kinesis Analytics in your AWS Management Console.
4. Observe how the dashboards are updated in real time with analytics results by choosing Visit Orders analysis dashboard in Elasticsearch Kibana.
Start an AWS Glue ETL job and observe crawlers and databases

In this step, you will create and start an extract, transform, and load (ETL) job on the sample dataset to perform JSON-to-Parquet conversion, and observe how AWS Glue components, such as crawlers and databases, behave.

In the wizard:

1. Review the diagram, which illustrates the ingest process (ETL) and automatic schema discovery with AWS Glue.
2. Start the ETL job on the curated datasets by choosing **Start Curated Datasets ETL**.
3. Observe AWS Glue crawlers, and databases created by the crawlers, by choosing **AWS Glue crawler in your AWS Management Console** and **AWS Glue databases in your AWS Management Console**.
4. Observe the AWS Glue ETL job by choosing **AWS Glue jobs in your AWS Management Console**.

Use Elasticsearch to index data

Before you search data, you must index it. In this step, you will see how streaming data and EcommCo’s batch data is indexed in Amazon ES.

In the wizard:

1. Review the diagram that demonstrate data governance.
2. Observe data being indexed by **Visit Elasticsearch UI**.

Query and analyze data with Amazon Redshift Spectrum

In this step, you’ll explore how the data warehousing solution can help you gain insights about data. You’ll use Amazon Redshift Spectrum to query and analyze data, and you’ll store the output as new curated datasets: Customer Lifetime Value, Spend Distribution, SKU Distribution, Product Category Distribution, State Population, and Orders by Month.

In the wizard:

1. Review the diagram that illustrates how to use Amazon Redshift Spectrum for queries.
2. Run analytics using Amazon Redshift Spectrum by choosing **Run analytics with Spectrum**.
3. Observe analytics output stores as curated datasets in S3.
Retrieve data with Athena queries
In this step, you will explore the curated datasets database and learn how to use SQL to retrieve data from the database.

In the wizard:
1. Inspect the tables in data catalog by choosing **Visit Amazon Athena Console**.
2. Run Athena queries by trying the listed queries in the Athena console.

Explore machine learning with Amazon SageMaker
In this step, you will learn how you can use machine learning with Amazon SageMaker to gain additional insights about the data.

In the wizard:
1. Run an Amazon SageMaker notebook, which is a machine learning compute instance that runs the Jupyter Notebook App. Choose **Visit SageMaker Notebook**, explore the notebook, and then run it.
2. Review the created artifacts by choosing **Trained model** and **Deployed endpoint**.
3. Obtain inferences from the deployed machine learning model by choosing the appropriate date and then choosing **Infer sales**.

Gain business insights with Amazon QuickSight
In this step, you will learn how to use business intelligence (BI) to analyze data with Amazon QuickSight.

In the wizard:
- Review the charts and follow the story.

Delete the demo
The demo provisions a collection of AWS resources in your AWS account. When you complete your pilot or POC, you can delete these resources by deleting the demo stack.

To delete the demo
1. Open the AWS CloudFormation console at
   
   https://console.aws.amazon.com/cloudformation/

   You will see a list of the currently running stacks in your account.
2. Select the stack named **qs-dl-foundation-demo**.

3. Choose **Actions**, and then choose **Delete Stack**.

4. When prompted, choose **Yes, Delete**.

**Next steps**

Try the production-ready [Data Lake Foundation Quick Start](#).

**FAQ**

**Q.** I encountered a CREATE_FAILED error when I launched the Quick Start.

**A.** If AWS CloudFormation fails to create the stack, we recommend that you relaunch the template with **Rollback on failure** set to **No**. (This setting is under **Advanced** in the AWS CloudFormation console, **Options** page.) With this setting, the stack’s state will be retained and the instance will be left running, so you can troubleshoot the issue. (You’ll want to look at the log files in `%ProgramFiles%\Amazon\EC2ConfigService` and `C:\cfn\log`.)

```
Important When you set Rollback on failure to No, you’ll continue to incur AWS charges for this stack. Please make sure to delete the stack when you’ve finished troubleshooting.
```

For additional information, see [Troubleshooting AWS CloudFormation](#) on the AWS website.

**Additional resources**

**AWS services**

- [Amazon Athena](#)
- [Amazon EBS](#)
- [Amazon EC2](#)
- [Amazon ES](#)
- [Amazon Kinesis](#)
- [Amazon QuickSight](#)
- [Amazon Redshift](#)
• Amazon Redshift Spectrum
• Amazon S3
• Amazon SageMaker
• Amazon VPC
• AWS CloudFormation
• AWS Glue
• Kibana plug-in

47 Lining data lake resources
• Data lake foundational concepts
• Data lake reference architecture
• Data lake sample dataset details

Document revisions

<table>
<thead>
<tr>
<th>Date</th>
<th>Change</th>
<th>In sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 2019</td>
<td>Initial publication</td>
<td></td>
</tr>
</tbody>
</table>
© 2019, Amazon Web Services, Inc. or its affiliates, and 47Lining, Inc. All rights reserved.

**Notices**

This document is provided for informational purposes only. It represents AWS’s current product offerings and practices as of the date of issue of this document, which are subject to change without notice. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS’s products or services, each of which is provided “as is” without warranty of any kind, whether express or implied. This document does not create any warranties, representations, contractual commitments, conditions or assurances from AWS, its affiliates, suppliers or licensors. 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.

The software included with this paper is licensed under the Apache License, Version 2.0 (the "License"). You may not use this file except in compliance with the License. A copy of the License is located at [http://aws.amazon.com/apache2.0/](http://aws.amazon.com/apache2.0/) or in the "license" file accompanying this file. This code is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.