Data Lake Foundation on the AWS Cloud

with AWS Services

Quick Start Reference Deployment

September 2017
Last updated: June 2018 (see revisions)

47Lining Team
AWS Quick Start Reference Team

AWS can provide you with AWS credits for this deployment. Please fill out our form and we will reach out to you.

Contents

Overview....2
Costs and Licenses....5
Architecture....5
Prerequisites....9
Specialized Knowledge....9
Quick Start Dataset....10
Deployment Options....11
Deployment Steps....12
Step 1. Prepare Your AWS Account....12
Step 2. Launch the Quick Start....12
Step 3. Test the Deployment....22
Step 4: Use the Wizard to Explore Data Lake Features....23

Optional: Using Your Own Dataset 25
Optional: Adding VPC Definitions 26
Troubleshooting and FAQ 27
Additional Resources 28
GitHub Repository 29
Document Revisions 29

This Quick Start deployment guide was created by 47Lining (a REAN Cloud Company) in collaboration with Amazon Web Services (AWS). 47Lining is an AWS Premier Consulting Partner specializing in big data.

Quick Starts are automated reference deployments that use AWS CloudFormation templates to deploy a specific workload on AWS, following AWS best practices.

Overview

This Quick Start reference deployment guide provides step-by-step instructions for deploying a data lake foundation on the AWS Cloud.

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, 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 Quick Start 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 software as a service (SaaS) tools.

The Quick Start 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.

The usage model diagram in Figure 1 illustrates key actors and use cases that the data lake enables, in context with the key component areas that comprise the data lake. This Quick Start provisions foundational data lake capabilities and optionally demonstrates key use cases for each type of actor in the usage model.

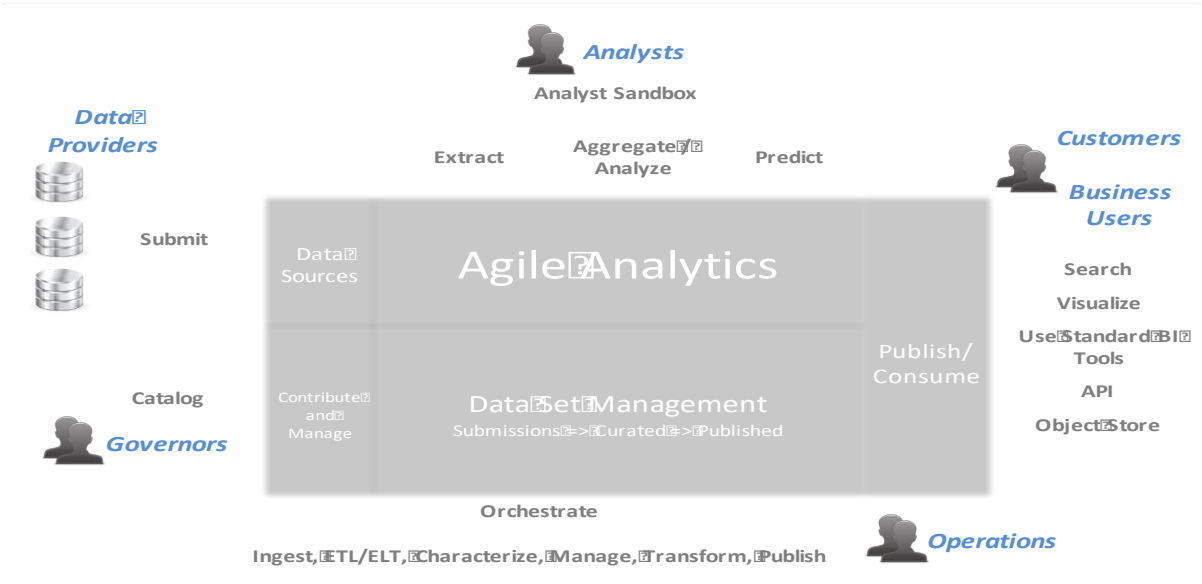


Figure 1: Usage model for Data Lake Foundation Quick Start

Figure 2 illustrates the foundational solution components of the data lake and how they relate to the usage model. The solution components interact through recurring and repeatable data lake patterns using your data and business flow.

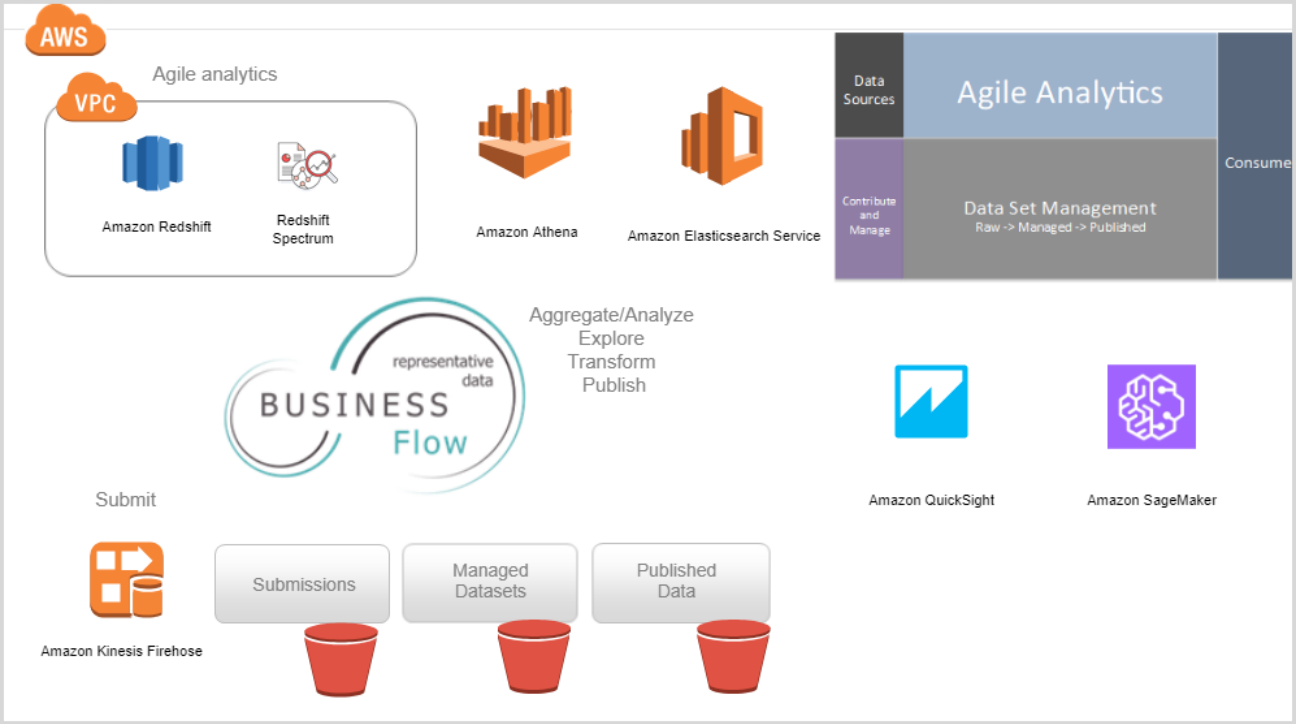


Figure 2: Capabilities and solution components in the Data Lake foundation Quick Start

Figure 2 illustrates these key capabilities:

- Submit full-snapshot and streaming data via Amazon S3 and Kinesis Data Firehose.
- Ingest and validate submissions, creating an initial curated dataset.
- Transform, aggregate, and analyze curated datasets with Athena, Amazon Redshift, and Amazon Redshift Spectrum.
- Search data lake metadata and view data lake statistics using Amazon ES with Kibana.
- Build and deploy machine learning models with Amazon SageMaker.
- Publish data via Amazon S3.
- Copy results into Amazon Redshift.
- Visualize data in Amazon QuickSight.

The Quick Start also deploys an optional wizard and a sample dataset. You can use the wizard after deployment to explore the architecture and functionality of the data lake

foundation and understand how they relate to repeatable data lake patterns. For more information, see step 4 in the deployment instructions.

Whether or not you choose to deploy the wizard and sample dataset, the Quick Start implementation is consistent with foundational data lake concepts that span physical architecture, data flow and orchestration, governance, and data lake usage and operations.

To learn more about 47Lining, foundational data lake concepts and reference architecture, and how to extend your data lake beyond this Quick Start implementation, see the 47Lining data lake resources page.

Costs and Licenses

You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using the Quick Start.

The AWS CloudFormation template for this Quick Start includes configuration parameters that you can customize. Some of these settings, such as instance type, will affect the cost of deployment. For cost estimates, see the pricing pages for each AWS service you will be using. Prices are subject to change.

Because this Quick Start uses AWS-native solution components, there are no costs or license requirements beyond AWS infrastructure costs. This Quick Start also deploys Kibana, which is an open-source tool that's included with Amazon ES.

Architecture

Deploying this Quick Start for a new virtual private cloud (VPC) with **default parameters** builds the following data lake environment in the AWS Cloud.

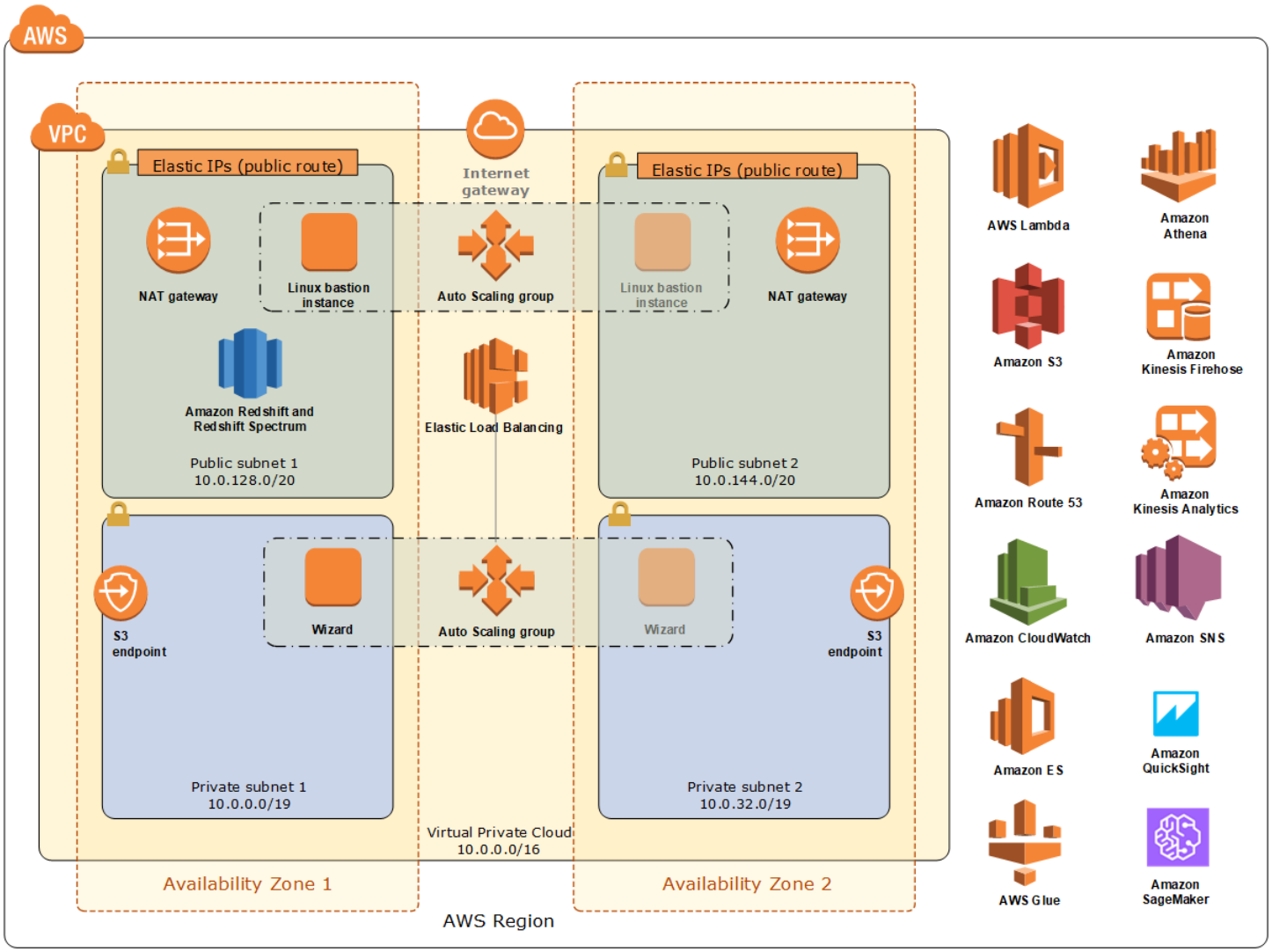


Figure 3: Quick Start architecture for data lake foundation on the AWS Cloud

The Quick Start sets up the following:

- A virtual private cloud (VPC) that spans two Availability Zones and includes two public and two private subnets.*
- An internet gateway to allow access to the internet.*
- In the public subnets, managed NAT gateways to allow outbound internet access for resources in the private subnets.*
- In the public subnets, Linux bastion hosts in an Auto Scaling group to allow inbound Secure Shell (SSH) access to EC2 instances in public and private subnets.*
- In a private subnet, a web application instance that hosts an optional wizard, which guides you through the data lake architecture and functionality.

- IAM roles to provide permissions to access AWS resources; for example, to permit Amazon Redshift and Amazon Athena to read and write curated datasets.
- In the private subnets, Amazon Redshift for data aggregation, analysis, transformation, and creation of curated and published datasets. When you launch the Quick Start with **Create Demonstration** set to **yes,** Amazon Redshift is launched in a public subnet.
- An Amazon SageMaker instance, which you can access by using AWS authentication. This instance is created only if the **Create Demonstration** parameter is set to **yes**.
- Integration with other Amazon services such as Amazon S3, Amazon Athena, AWS Glue, AWS Lambda, Amazon ES with Kibana, Amazon Kinesis, and Amazon QuickSight.
- Your choice to create a new VPC or deploy the data lake components into your existing VPC on AWS. The template that deploys the Quick Start into an existing VPC skips the components marked by asterisks above.

Figure 4 shows how these components work together in a typical end-to-end process flow. If you choose to deploy the Quick Start with the wizard and sample dataset, the wizard will guide you through this process flow and core data lake concepts using sample data, which is described in the Quick Start Dataset section.

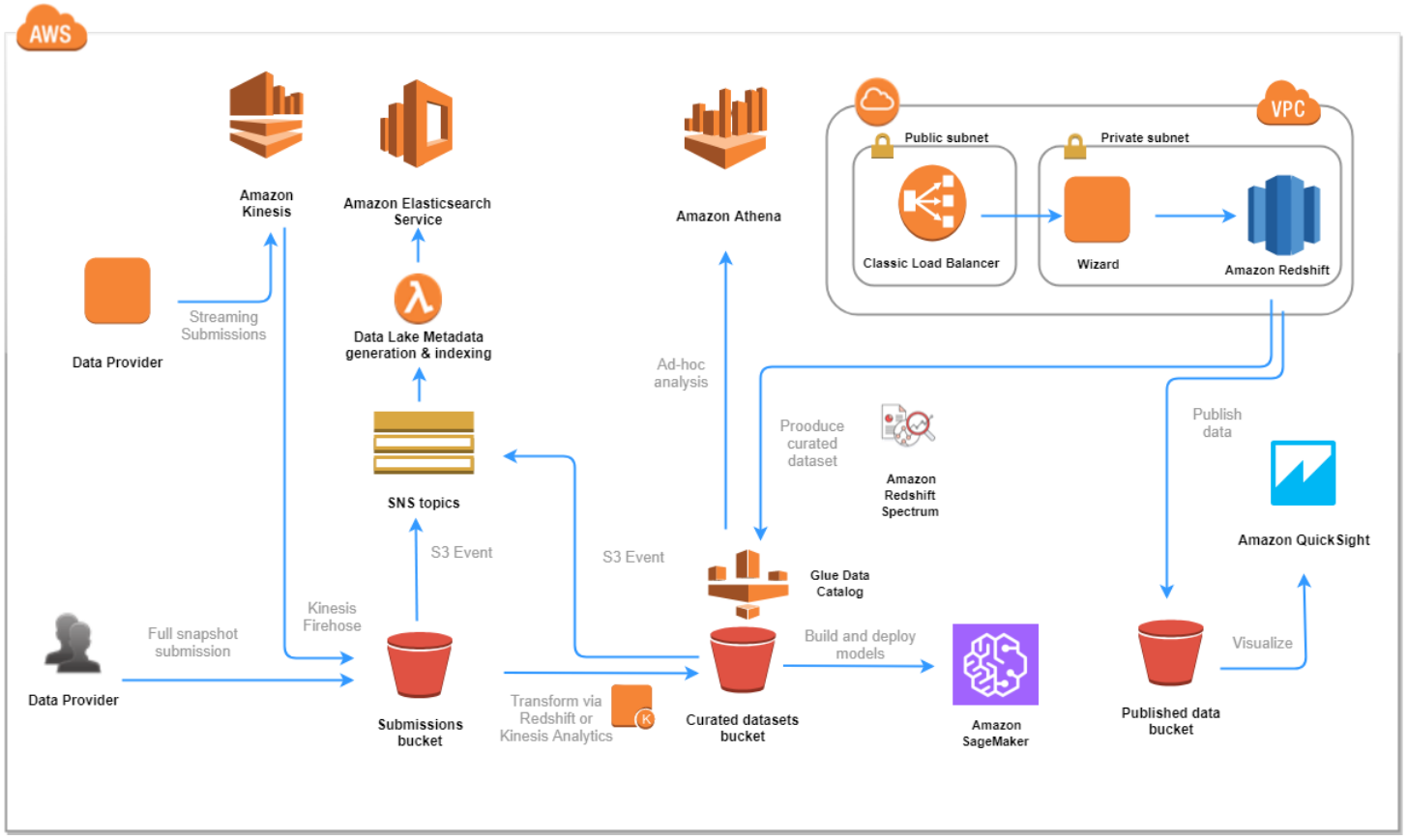


Figure 4: Data lake foundation process flow

The process flow consists of the following:

- **Submit**. Distinct submission processes accept both batch submissions to an S3 submissions bucket and streaming submissions via Amazon Kinesis Data Firehose.

  All data lake submissions are indexed in Amazon ES, triggered by Amazon S3 events.

- **Ingest**. Ingest processes validate submissions, create characterization metadata, and, in some cases, transform submissions as they are accepted into the data lake and maintained as curated datasets.

- **Curated datasets**. Curated datasets provide the foundation for a value-tiering strategy within the data lake. The simplest curated datasets minimally transform submissions, so that historic submissions can be "replayed" to obtain a correct, full dataset representation. Higher-value curated datasets combine multiple input datasets using an agile analytic transform. Amazon Redshift, AWS Glue, or Amazon Kinesis Analytics implements these transforms and creates the resulting curated datasets. Curated datasets reside in a dedicated S3 bucket. They are also indexed in Amazon ES.

- **Agile analytics to transform, aggregate, analyze**. Amazon Athena performs ad-hoc analyses on the curated datasets, and Amazon Redshift Spectrum helps join dimensional data with facts. AWS Glue auto-discovers datasets and transforms datasets with ETL jobs. You can choose the right analytics engine for the job to create and maintain each curated dataset, based on your data and the requirements and preferences of your analysts.
- **Build and deploy machine learning models**. Amazon SageMaker uses transformed data (stored in an S3 bucket as curated datasets) to train a machine learning model. The model is then deployed into the Amazon SageMaker hosting service and used for real-time inference.
- **Search**. Search is enabled by indexing metadata in Amazon ES and exposing it through Kibana dashboards. When new data is added to Amazon S3, it triggers events, which are published to Amazon SNS. Amazon SNS triggers AWS Lambda functions to index the metadata in Amazon ES.
- **Publish results**. The publishing process moves and transforms data from the S3 curated datasets bucket to the published results buckets for downstream consumers like Amazon QuickSight or a dedicated analyst sandbox.
- **Visualize**. Published results are visualized with Amazon QuickSight.

Prerequisites

Specialized Knowledge

Before you deploy this Quick Start, we recommend that you become familiar with the following AWS services. (If you are new to AWS, see the Getting Started Resource Center.)

- Amazon Athena
- Amazon EC2
- Amazon ES
- Amazon Kinesis
- Amazon QuickSight
- Amazon Redshift
- Amazon Redshift Spectrum
- Amazon S3
- Amazon SageMaker
- Amazon VPC
- AWS Glue

Quick Start Dataset

The Quick Start includes an optional sample dataset, which it loads into the Amazon Redshift cluster and Kinesis data streams. The data lake wizard uses this dataset to demonstrate foundational data lake capabilities such as search, transforms, queries, analytics, and visualization. You can customize the parameter settings when you launch the Quick Start to replace this dataset as needed for your use case; see the section Optional: Using Your Own Dataset for details.

The sample data set is from ECommCo, a fictional company that sells products in multiple categories through its ecommerce website, ECommCo.com. The following diagram summarizes the requirements of ECommCo's business users.

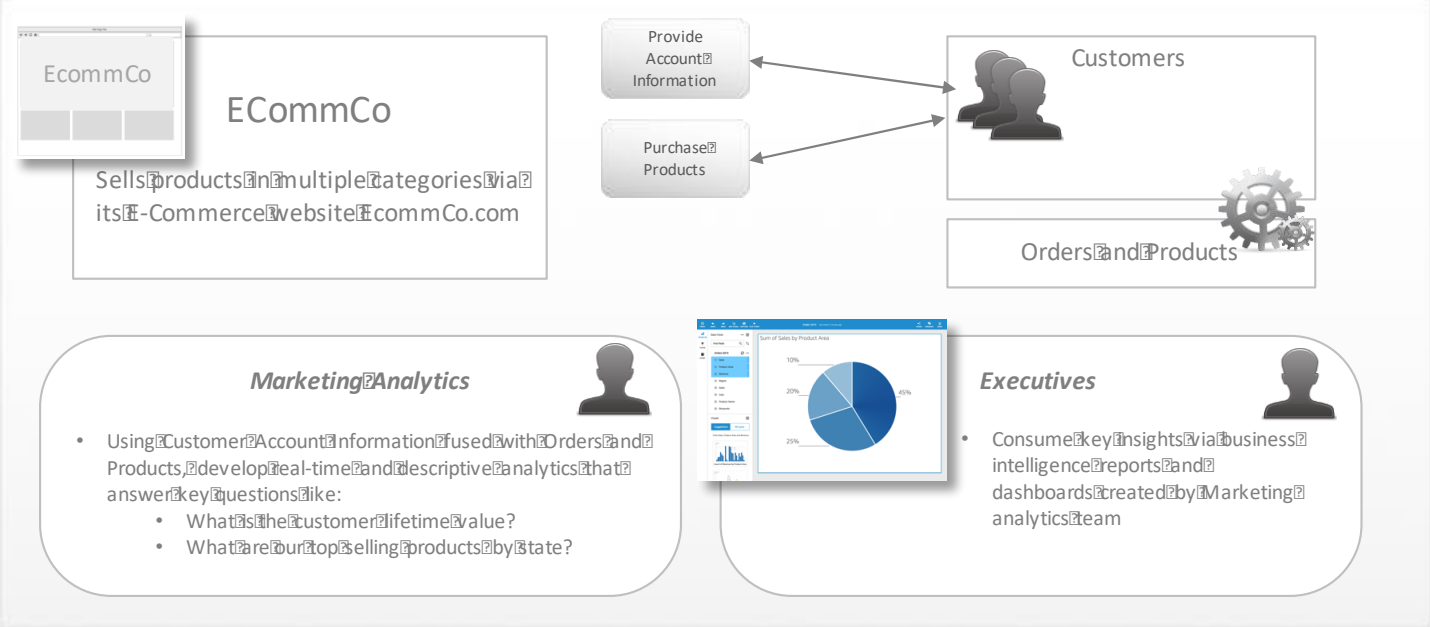


Figure 5: ECommCo at a high level

The Quick Start dataset includes representative full-snapshot and streaming data that demonstrate how data is submitted to, and ingested by, the data lake. This data can then be used in descriptive, predictive, and real-time analytics to answer ECommCo's most pressing business questions. The Quick Start data is summarized in Figure 6.

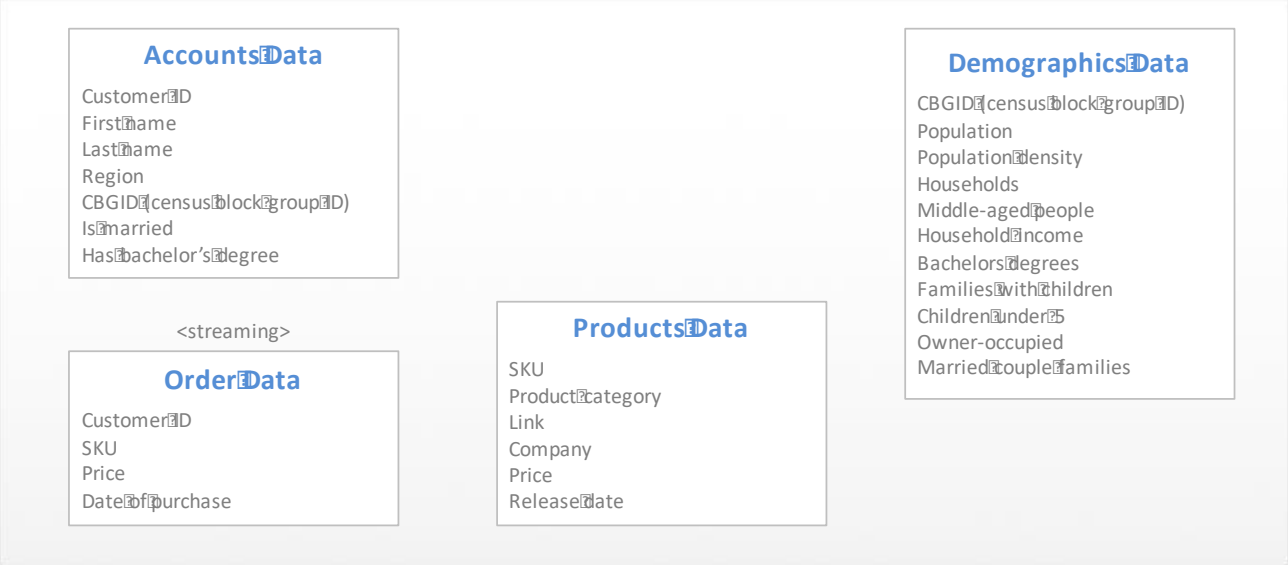


Figure 6: Quick Start sample data

For more information about the Quick Start dataset and the demonstration analytics performed in the Quick Start environment, see the 47Lining Data Lake Quick Start Sample Dataset Description.

Deployment Options

This Quick Start provides two deployment options:

- **Deploy the Quick Start into a new VPC** (end-to-end deployment). This option builds a new AWS environment consisting of the VPC, subnets, NAT gateways, bastion hosts, security groups, and other infrastructure components, and then deploys the data lake services and components into this new VPC.
- **Deploy the Quick Start into an existing VPC**. This option deploys the data lake services and components in your existing AWS infrastructure.

The Quick Start provides separate templates for these options. It also lets you configure CIDR blocks, instance types, and data lake settings, as discussed later in this guide.

Deployment Steps

Step 1. Prepare Your AWS Account

1. If you don't already have an AWS account, create one at https://aws.amazon.com by following the on-screen instructions.
2. Use the region selector in the navigation bar to choose the AWS Region where you want to deploy the data lake foundation on AWS.

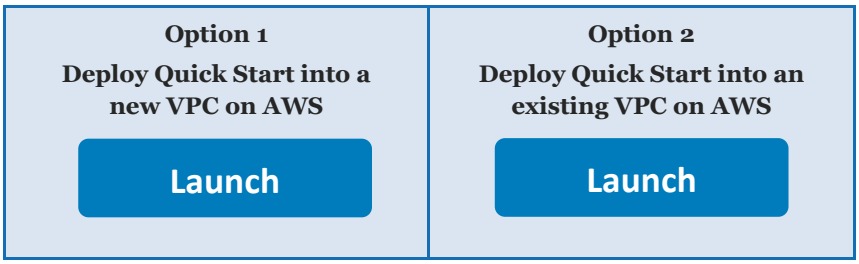
> **Important** This Quick Start includes services that aren't supported in all AWS Regions. For a list of supported regions, see the pages for Amazon Kinesis Firehose, AWS Glue, Amazon Redshift Spectrum, and Amazon SageMaker on the AWS website.

3. Create a key pair in your preferred region.
4. If necessary, request a service limit increase for the Amazon EC2 **t2.micro** instance type. You might need to do this if you already have an existing deployment that uses this instance type, and you think you might exceed the default limit with this reference deployment.
5. If necessary, request a service limit increase for AWS CloudFormation stacks. The Quick Start will create up to eleven (11) stacks. You may need to request a service limit increase if you already have existing deployments that use AWS CloudFormation stacks.

Step 2. Launch the Quick Start

> **Note** You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using this Quick Start. For full details, see the pricing pages for each AWS service you will be using in this Quick Start. Prices are subject to change.

1. Choose one of the following options to launch the AWS CloudFormation template into your AWS account. For help choosing an option, see deployment options earlier in this guide.



> **Important** If you're deploying the Quick Start into an existing VPC, make sure that your VPC has two private subnets in different Availability Zones for the database instances. These subnets require NAT gateways or NAT instances in their route tables, to allow the instances to download packages and software without exposing them to the internet. You'll also need the domain name option configured in the DHCP options, as explained in the Amazon VPC documentation. You'll be prompted for your VPC settings when you launch the Quick Start.

   Each deployment takes about 50 minutes to complete.

2. Check the region that's displayed in the upper-right corner of the navigation bar, and change it if necessary. This is where the network infrastructure for the data lake will be built. The template is launched in the US West (Oregon) Region by default.

> **Important** This Quick Start includes services that aren't supported in all AWS Regions. For a list of supported regions, see the pages for Amazon Kinesis Firehose, AWS Glue, Amazon Redshift Spectrum, and Amazon SageMaker on the AWS website.

3. On the **Select Template** page, keep the default setting for the template URL, and then choose **Next**.
4. On the **Specify Details** page, change the stack name if needed. Review the parameters for the template. Provide values for the parameters that require input. For all other parameters, review the default settings and customize them as necessary. When you finish reviewing and customizing the parameters, choose **Next**.

   In the following tables, parameters are listed by category and described separately for the two deployment options:

- Parameters for deploying the Quick Start into a new VPC
- Parameters for deploying the Quick Start into an existing VPC

- **Option 1: Parameters for deploying the Quick Start into a new VPC**

View template

Network Configuration:

Parameter label (name)	Default	Description
Availability Zones (AvailabilityZones)	*Requires input*	The list of Availability Zones to use for the subnets in the VPC. You must specify two Availability Zones. By default, the Quick Start preserves the logical order you specify.
VPC Definition (VPCDefinition)	QuickstartDefault	VPC definition name from the `Mappings` section of the template. Each definition specifies a VPC configuration, including the number of Availability Zones to be used for the deployment and the CIDR blocks for the VPC, public subnets, and private subnets. You can support multiple VPC configurations by extending the map with additional definitions and choosing the appropriate name. If you don't want to change the VPC configuration, keep the default setting. For more information, see the Adding VPC Definitions section.

Demonstration Configuration:

Parameter label (name)	Default	Description
Create Demonstration (CreateDemonstration)	yes	Set this parameter to **no** if you don't want the Quick Start to deploy the data lake wizard and load sample data into the Amazon Redshift cluster and Kinesis streams. For more information about the wizard, see step 4.
The following five parameters are used only if **Create Demonstration** is set to **yes**.		
Wizard Instance Type (WizardInstanceType)	t2.micro	The EC2 instance type for the data lake wizard.
Wizard User Name (WizardUsername)	DataLakeUser	The user name for the wizard, consisting of 1-64 ASCII characters.
Wizard Password (WizardPassword)	*Requires input*	The password for the wizard, consisting of 8-64 ASCII characters. The password must contain one uppercase letter, one lowercase letter, and one number. This password is required, but it will be used only when you launch the Quick Start with **Create Demonstration** set to **yes**.
Dataset S3 Bucket Name (DatasetS3BucketName)	aws-quickstart-datasets	S3 bucket where the sample dataset is installed. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen. Keep the default setting to use the sample dataset

Parameter label (name)	Default	Description
		included with the Quick Start. If you decide to use a different dataset, or if you decide to customize or extend the Quick Start dataset, use this parameter to specify the S3 bucket name that you would like the Quick Start to load. (For more information, see Using Your Own Dataset.)
Dataset S3 Key Prefix (DatasetS3KeyPrefix)	quickstart-datalake-47lining/ecommco/v2/	S3 key prefix where the sample dataset is installed. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes, but should not start with a forward slash, which is automatically added. Keep the default setting to use the sample dataset included with the Quick Start. If you decide to use a different dataset, or if you decide to customize or extend the Quick Start dataset, use this parameter to specify the location for the dataset you would like the Quick Start to load. (For more information, see Using Your Own Dataset.)

Elasticsearch Configuration:

Parameter label (name)	Default	Description
Remote Access CIDR (RemoteAccessCIDR)	*Requires input*	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. For example, you might want to grant only your corporate network access to the software. You can use http://checkip.amazonaws.com/ 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).
Elasticsearch Node Type (ElasticsearchNodeType)	t2.small.elasticsearch	EC2 instance type for the Elasticsearch cluster.
Elasticsearch Node Count (ElasticsearchNodeCount)	1	The number of nodes in the Elasticsearch cluster. For guidance, see the Amazon ES documentation.

Redshift Configuration:

Parameter label (name)	Default	Description
Enable Redshift (EnableRedshift)	yes	Specifies whether Amazon Redshift will be provisioned when the **Create Demonstration** parameter is set to **no**. This parameter is ignored when **Create Demonstration** is set to **yes** (in that case, Amazon Redshift is always provisioned). Set to **no** if you've set the **Create Demonstration** parameter to **no**, and you don't want to provision the Amazon Redshift cluster.

Parameter label (name)	Default	Description
Redshift User Name (RedshiftUsername)	datalake	The user name that is associated with the master user account for the Amazon Redshift cluster. The user name must contain fewer than 128 alphanumeric characters or underscores, and must be lowercase and begin with a letter.
Redshift Password (RedshiftPassword)	*Requires input*	The password that is associated with the master user account for the Amazon Redshift cluster. The password must contain 8-64 printable ASCII characters, excluding: /, ", \', \ and @. It must contain one uppercase letter, one lowercase letter, and one number.
Redshift Number of Nodes (RedshiftNumberOfNodes)	1	The number of nodes in the Amazon Redshift cluster. If you specify a number that's larger than 1, the Quick Start will launch a multi-node cluster.
Redshift Node Type (RedshiftNodeType)	dc1.large	Instance type for the nodes in the Amazon Redshift cluster.
Redshift Database Name (RedshiftDatabaseName)	quickstart	The name of the first database to be created when the Amazon Redshift cluster is provisioned.
Redshift Database Port (RedshiftDatabasePort)	5439	The port that Amazon Redshift will listen on, which will be allowed through the security group.

Kinesis Configuration:

Parameter label (name)	Default	Description
Kinesis Data Stream Name (KinesisDataStreamName)	streaming-submissions	Name of the Kinesis data stream. Change this parameter only if you've set the **Create Demonstration** parameter to **no**. Keep the default setting to use the sample dataset included with the Quick Start.
Kinesis Data Stream S3 Prefix (KinesisDataStreamS3 Prefix)	streaming-submissions	S3 key prefix for your streaming data stored in the S3 submissions bucket. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes, but should not start with a forward slash, which is automatically added. Use this parameter to specify the location for the streaming data you'd like to load. Change this parameter only if you've set the **Create Demonstration** parameter to **no**. Keep the default setting to use the sample dataset included with the Quick Start.

SageMaker Configuration:

The Quick Start creates an Amazon SageMaker instance and uses the parameters in this section **only** when the **Create Demonstration** parameter is set to **yes**.

Parameter label (name)	Default	Description
Notebook Instance Name (NotebookInstanceName)	SageMaker Notebook	Name of the Amazon SageMaker Notebook instance.
Notebook Instance Type (NotebookInstanceType)	ml.t2.medium	The EC2 instance type for the data lake Amazon SageMaker Notebook instance.
Notebook Training Instance Type (NotebookTraining InstanceType)	ml.m5.xlarge	The EC2 instance type for the Amazon SageMaker training instance, which will be used for model training.

AWS Quick Start Configuration:

Parameter label (name)	Default	Description
Quick Start S3 Bucket Name (QSS3BucketName)	aws-quickstart	S3 bucket where the Quick Start templates and scripts are installed. Use this parameter to specify the S3 bucket name you've created for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen.
Quick Start S3 Key Prefix (QSS3KeyPrefix)	quickstart-datalake-47lining/	S3 key prefix used to simulate a folder for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes.

- **Option 2: Parameters for deploying the Quick Start into an existing VPC**

View template

Network Configuration:

Parameter label (name)	Default	Description
Availability Zones (AvailabilityZones)	*Requires input*	The list of Availability Zones to use for the subnets in the VPC. You must specify two Availability Zones. By default, the Quick Start preserves the logical order you specify.
Existing VPC ID (VPCID)	*Requires input*	ID of your existing VPC (e.g., vpc-0343606e).
Existing VPC CIDR (VPCCIDR)	*Requires input*	CIDR block for the VPC.
Existing VPC Private Subnet 1 ID (PrivateSubnet1ID)	*Requires input*	ID of the private subnet in Availability Zone 1 in your existing VPC (e.g., subnet-a0246dcd).
Existing VPC Private Subnet 2 ID (PrivateSubnet2ID)	*Requires input*	ID of the private subnet in Availability Zone 2 in your existing VPC (e.g., subnet-b1f432cd).
Existing VPC Public Subnet 1 ID (PublicSubnet1ID)	*Requires input*	ID of the public subnet in Availability Zone 1 in your existing VPC (e.g., subnet-9bc642ac).
Existing VPC Public Subnet 2 ID (PublicSubnet2ID)	*Requires input*	ID of the public subnet in Availability Zone 2 in your existing VPC (e.g., subnet-e3246d8e).
NAT 1 IP address (NAT1ElasticIP)	*Requires input*	Elastic IP address for the first NAT gateway instance that will be allowed access to Amazon ES.
NAT 2 IP address (NAT2ElasticIP)	*Requires input*	Elastic IP address for the second NAT gateway instance that will be allowed access to Amazon ES.

Demonstration Configuration:

Parameter label (name)	Default	Description
Create Demonstration (CreateDemonstration)	yes	Set this parameter to **no** if you don't want the Quick Start to deploy the data lake wizard and load sample data into the Amazon Redshift cluster and Kinesis streams. For more information about the wizard, see step 4.
The following five parameters are used only if **Create Demonstration** is set to **yes**.		
Wizard Instance Type (InstanceType)	t2.micro	The EC2 instance type for the data lake wizard.

Parameter label (name)	Default	Description
Wizard User Name (WizardUsername)	DataLakeUser	The user name for the wizard, consisting of 1-64 ASCII characters.
Wizard Password (WizardPassword)	*Requires input*	The password for the wizard, consisting of 8-64 ASCII characters. The password must contain one uppercase letter, one lowercase letter, and one number. This password is required, but it will be used only when you launch the Quick Start with **Create Demonstration** set to **yes**.
Dataset S3 Bucket Name (DatasetS3BucketName)	aws-quickstart-datasets	S3 bucket where the sample dataset is installed. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen. Keep the default setting to use the sample dataset included with the Quick Start. If you decide to use a different dataset, or if you decide to customize or extend the Quick Start dataset, use this parameter to specify the S3 bucket name that you would like the Quick Start to load. (For more information, see Using Your Own Dataset.)
Dataset S3 Key Prefix (DatasetS3KeyPrefix)	quickstart-datalake-47lining/ ecommco/v2/	S3 key prefix where the sample dataset is installed. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes, but should not start with a forward slash, which is automatically added. Keep the default setting to use the sample dataset included with the Quick Start. If you decide to use a different dataset, or if you decide to customize or extend the Quick Start dataset, use this parameter to specify the location for the dataset you would like the Quick Start to load. (For more information, see Using Your Own Dataset.)

Elasticsearch Configuration:

Parameter label (name)	Default	Description
Remote Access CIDR (RemoteAccessCIDR)	*Requires input*	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. For example, you might want to grant only your corporate network access to the software. You can use http://checkip.amazonaws.com/ 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).
Elasticsearch Node Type (ElasticsearchNodeType)	t2.small. elasticsearch	EC2 instance type for the Elasticsearch cluster.
Elasticsearch Node Count (ElasticsearchNodeCount)	1	The number of nodes in the Elasticsearch cluster. For guidance, see the Amazon ES documentation.

Redshift Configuration:

Parameter label (name)	Default	Description
Enable Redshift (EnableRedshift)	yes	Specifies whether Amazon Redshift will be provisioned when the **Create Demonstration** parameter is set to **no**. This parameter is ignored when **Create Demonstration** is set to **yes** (in that case, Amazon Redshift is always provisioned). Set to **no** if you've set the **Create Demonstration** parameter to **no**, and you don't want to provision the Amazon Redshift cluster.
Redshift User Name (RedshiftUsername)	datalake	The user name that is associated with the master user account for the Amazon Redshift cluster. The user name must contain fewer than 128 alphanumeric characters or underscores, and must be lowercase and begin with a letter.
Redshift Password (RedshiftPassword)	*Requires input*	The password that is associated with the master user account for the Amazon Redshift cluster. The password must contain 8-64 printable ASCII characters, excluding: /, ", \', \ and @. It must contain one uppercase letter, one lowercase letter, and one number.
Redshift Number of Nodes (RedshiftNumberOfNodes)	1	The number of nodes in the Amazon Redshift cluster. If you specify a number that's larger than 1, the Quick Start will launch a multi-node cluster.
Redshift Node Type (RedshiftNodeType)	dc1.large	Instance type for the nodes in the Amazon Redshift cluster.
Redshift Database Name (RedshiftDatabaseName)	quickstart	The name of the first database to be created when the Amazon Redshift cluster is provisioned.
Redshift Database Port (RedshiftDatabasePort)	5439	The port that Amazon Redshift will listen on, which will be allowed through the security group.

Kinesis Configuration:

Parameter label (name)	Default	Description
Kinesis Data Stream Name (KinesisDataStreamName)	streaming-submissions	Name of the Kinesis data stream. Change this parameter only if you've set the **Create Demonstration** parameter to **no**. Keep the default setting to use the sample dataset included with the Quick Start.
Kinesis Data Stream S3 Prefix (KinesisDataStreamS3 Prefix)	streaming-submissions	S3 key prefix for your streaming data stored in the S3 submissions bucket. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes, but should not start with a forward slash, which is automatically added. Use this parameter to specify the location for the streaming data you'd like to load.

Parameter label (name)	Default	Description
		Change this parameter only if you've set the **Create Demonstration** parameter to **no**. Keep the default setting to use the sample dataset included with the Quick Start.

SageMaker Configuration:

The Quick Start creates an Amazon SageMaker instance and uses the parameters in this section **only** when the **Create Demonstration** parameter is set to **yes**.

Parameter label (name)	Default	Description
Notebook Instance Name (NotebookInstanceName)	SageMaker Notebook	Name of the Amazon SageMaker Notebook instance.
Notebook Instance Type (NotebookInstanceType)	ml.t2.medium	The EC2 instance type for the data lake Amazon SageMaker Notebook instance.
Notebook Training Instance Type (NotebookTraining InstanceType)	ml.m5.xlarge	The EC2 instance type for the Amazon SageMaker training instance, which will be used for model training.

AWS Quick Start Configuration:

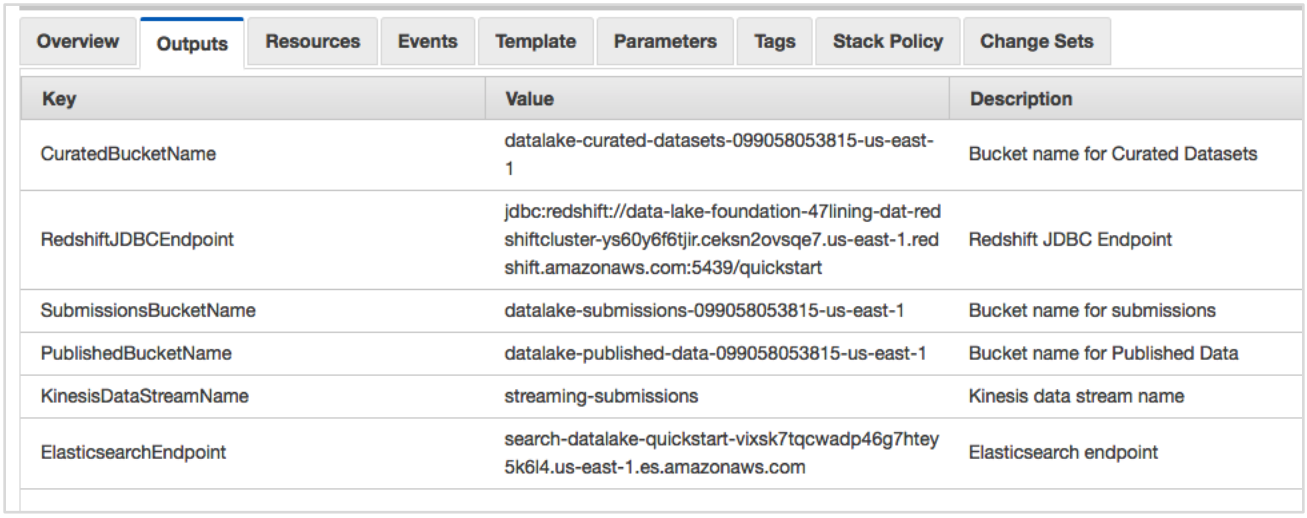
Parameter label (name)	Default	Description
Quick Start S3 Bucket Name (QSS3BucketName)	aws-quickstart	S3 bucket where the Quick Start templates and scripts are installed. You can specify the S3 bucket name you've created for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen.
Quick Start S3 Key Prefix (QSS3KeyPrefix)	quickstart-datalake-47lining/	S3 key prefix for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes.

5. On the **Options** page, you can specify tags (key-value pairs) for resources in your stack and set advanced options. When you're done, choose **Next**.
6. On the **Review** page, review and confirm the template settings. Under **Capabilities**, select the check box to acknowledge that the template will create IAM resources.

7. Choose **Create** to deploy the stack.
8. Monitor the status of the stack. When the status is **CREATE_COMPLETE**, the data lake cluster is ready.
9. You can use the information displayed in the **Outputs** tab for the stack to view the resources that were created and to verify the deployment, as discussed in the next step.

Step 3. Test the Deployment

When you launch the Quick Start with **Create Demonstration** set to **no**, you can validate and test the deployment by checking the resources in the **Outputs** tab.



Key	Value	Description
CuratedBucketName	datalake-curated-datasets-099058053815-us-east-1	Bucket name for Curated Datasets
RedshiftJDBCEndpoint	jdbc:redshift://data-lake-foundation-47lining-dat-redshiftcluster-ys60y6f6tjir.ceksn2ovsqe7.us-east-1.redshift.amazonaws.com:5439/quickstart	Redshift JDBC Endpoint
SubmissionsBucketName	datalake-submissions-099058053815-us-east-1	Bucket name for submissions
PublishedBucketName	datalake-published-data-099058053815-us-east-1	Bucket name for Published Data
KinesisDataStreamName	streaming-submissions	Kinesis data stream name
ElasticsearchEndpoint	search-datalake-quickstart-vixsk7tqcwadp46g7htey5k6l4.us-east-1.es.amazonaws.com	Elasticsearch endpoint

Figure 7: Quick Start outputs

You should confirm the following:

- The S3 buckets listed on the **Outputs** tab for the stack are available in the Amazon S3 console. The Quick Start provisions distinct S3 buckets for submissions, curated datasets, and published results.
- If you launched the Quick Start with **Enable Redshift** set to **yes**, Amazon Redshift is accessible at the Java Database Connectivity (JDBC) endpoint specified on the **Outputs** tab for the stack, using the **Redshift User Name** and **Redshift Password** that you specified when you launched the Quick Start.
- The Kinesis stream for streaming submissions listed on the **Outputs** tab for the stack is available in the Kinesis console.
- The ElasticSearch cluster listed on the **Outputs** tab for the stack is available in the Amazon ES console, and the Kibana endpoint listed on the **Outputs** tab is accessible

from a web browser client within the **Remote Access CIDR** that you specified when launching the Quick Start.

Step 4: Use the Wizard to Explore Data Lake Features

If the **Create Demonstration** parameter is set to **yes** (its default setting), you'll see a URL for the wizard in the **Outputs** tab, and you can use the wizard to explore the data lake architecture and the AWS services used within this Quick Start. The wizard includes eight 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.

1. Choose the URL for **DataLakeWizardURL** in the **Outputs** tab, and open it in a web browser.
2. Log in to the wizard by using the parameters you specified in step 2: Use the value of **Wizard User Name** as your login name, and **Wizard Password** as your password.

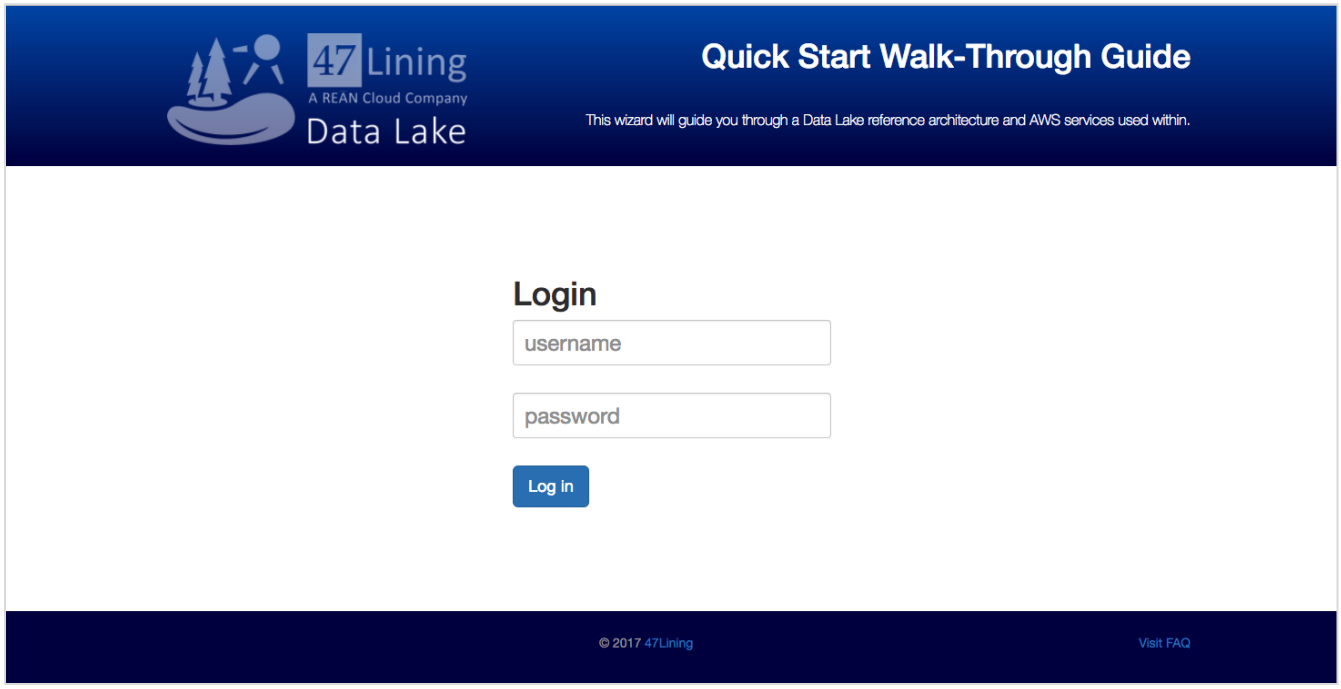


Figure 8: Login page of wizard

3. On the **Get Started** screen, read the directions carefully to learn how to step through the path from initial data submission to transformations, to analytics, and finally to visualizations.

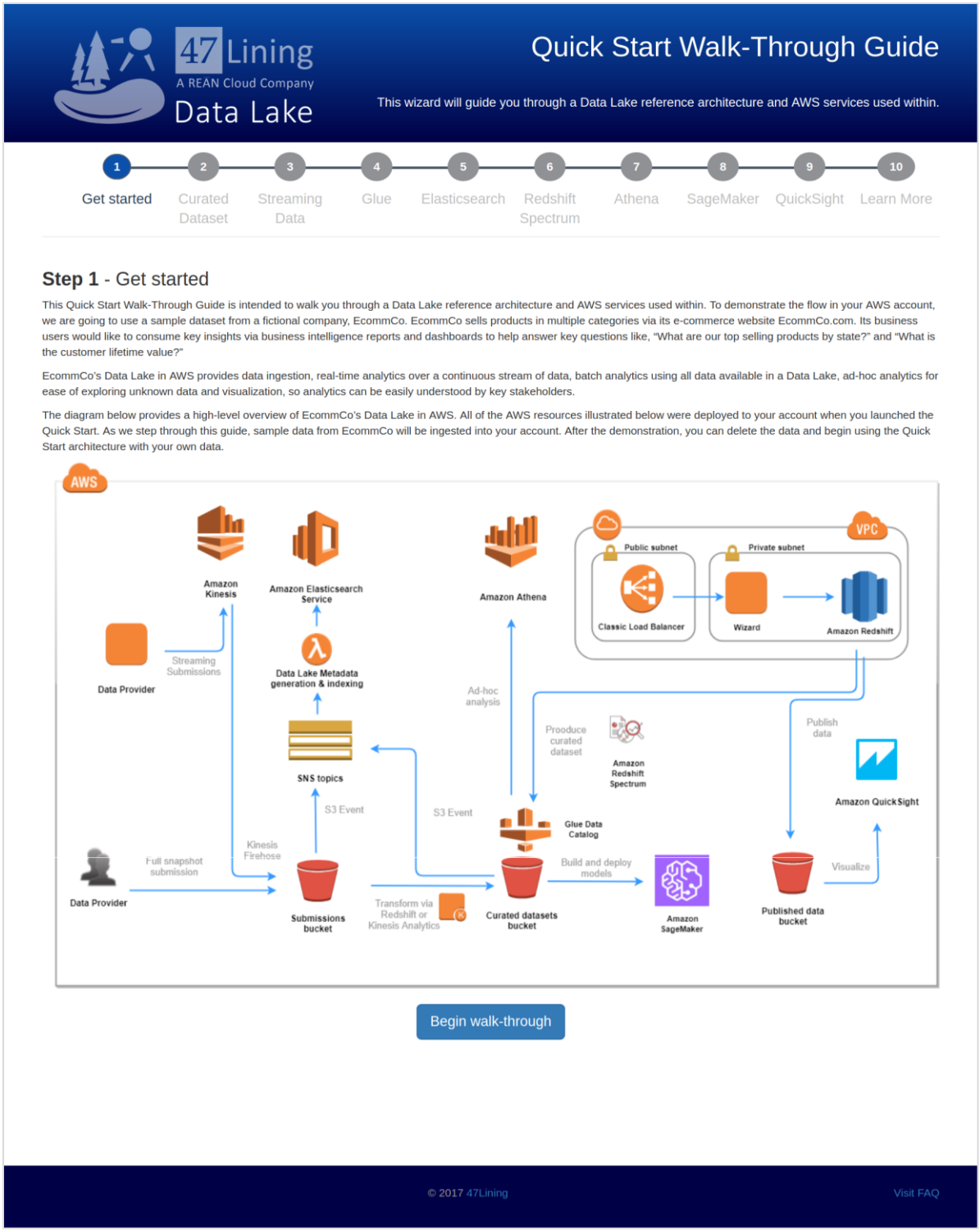


Figure 9: Getting started with the wizard

Optional: Using Your Own Dataset

You can deploy this Quick Start without the sample dataset and wizard, and extend it with your own dataset instead. To do so, set the **Create Demonstration** parameter to **no**. You can then use the following infrastructure, which the Quick Start sets up:

- Kinesis Firehose endpoint, which accepts streaming submissions into the S3 submissions bucket
- Amazon Redshift, which is optionally deployed into a private subnet, to ingest data from Amazon S3 or through JDBC, and analyze it
- Elasticsearch cluster, which provides you with a data lake dashboard of all S3 objects that were placed in the S3 submissions bucket, curated datasets bucket, and published datasets bucket

The data lake foundation provides a solid base for your processes. Using this infrastructure, you can:

- Ingest batch submissions, resulting in curated datasets in Amazon S3. You can then use your own SQL scripts to load curated datasets to Amazon Redshift.
- Ingest streaming submissions provided through Kinesis Firehose.
- Auto-discover curated datasets using AWS Glue crawlers, and transform curated datasets with AWS Glue jobs.
- Analyze the data with Amazon Redshift, using your own SQL queries.
- Analyze the data with Kinesis Analytics, by creating your own applications that read streaming data from Kinesis Firehose.
- Publish the results of analytics to the published datasets bucket.
- Get a high-level picture of your data lake by using Amazon ES, which indexes the metadata of S3 objects.
- Use Amazon Athena to run ad-hoc analytics over your curated datasets, and Amazon QuickSight to visualize the datasets in the published datasets bucket. You can also use Amazon Athena or Amazon Redshift as data sources in AWS QuickSight.

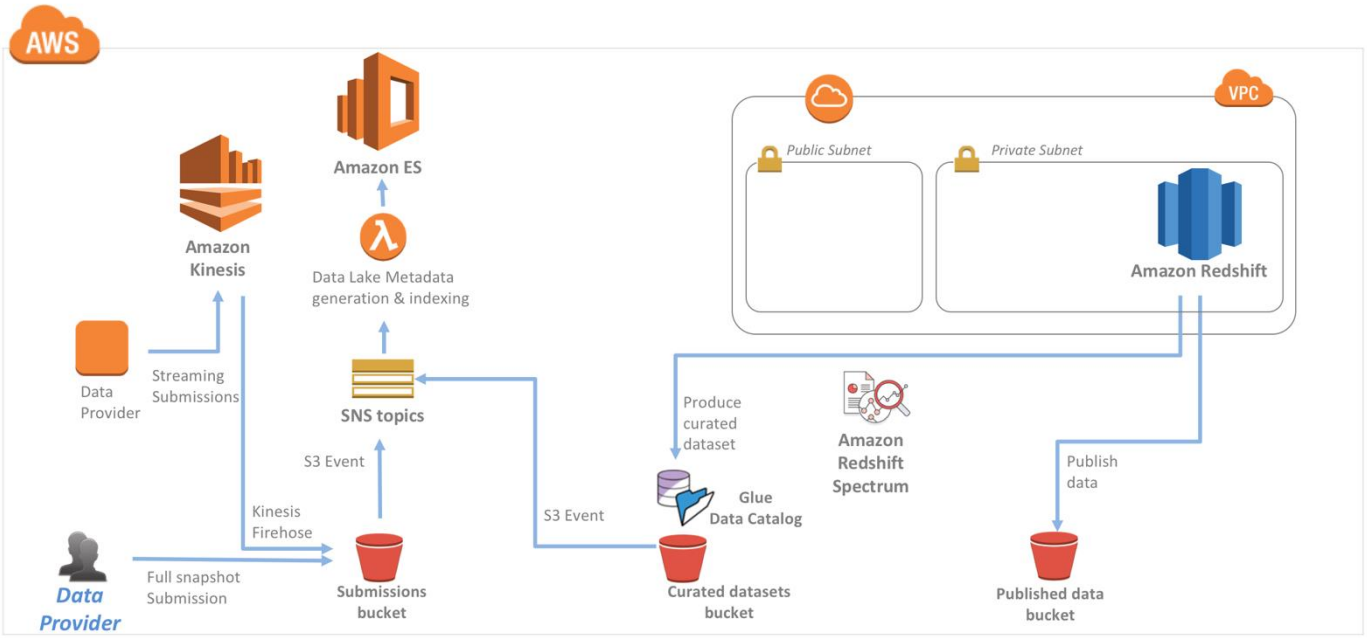


Figure 10: Infrastructure deployed when launching Quick Start without demonstration

Optional: Adding VPC Definitions

When you launch the Quick Start in the mode where a new VPC is created, the Quick Start uses VPC parameters that are defined in a mapping within the Quick Start templates. If you choose to download the templates from the GitHub repository, you can add new named VPC definitions to the mapping, and choose one of these named VPC definitions when you launch the Quick Start.

The following table shows the parameters within each VPC definition. You can create as many VPC definitions as you need within your environments. When you deploy the Quick Start, use the **VPC Definition** parameter to specify the configuration you want to use.

Parameter	Default	Description
NumberOfAZs	2	Number of Availability Zones to use in the VPC.
PublicSubnet1CIDR	10.0.1.0/24	CIDR block for the public (DMZ) subnet 1 located in Availability Zone 1.
PrivateSubnet1CIDR	10.0.2.0/24	CIDR block for private subnet 1 located in Availability Zone 1.
PublicSubnet2CIDR	10.0.3.0/24	CIDR block for the public (DMZ) subnet 2 located in Availability Zone 2.
PrivateSubnet2CIDR	10.0.4.0/24	CIDR block for private subnet 2 located in Availability Zone 2.
VPCCIDR	10.0.0.0/16	CIDR block for the VPC.

Troubleshooting and 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.

Q. I encountered a size limitation error when I deployed the AWS CloudFormation templates.

A. We recommend that you launch the Quick Start templates from the location we've provided or from another S3 bucket. If you deploy the templates from a local copy on your computer, you might encounter template size limitations when you create the stack. For more information about AWS CloudFormation limits, see the AWS documentation.

Q. I deployed the Quick Start in the EU (London) Region, but it didn't work.

A. This Quick Start includes services that aren't supported in all regions. See the pages for Amazon Kinesis Firehose, AWS Glue, Amazon SageMaker, and Amazon Redshift Spectrum on the AWS website for a list of supported regions.

Q. Can I use the QuickStart with my own data?

A. Yes, you can. See the section Optional: Using Your Own Dataset.

Q. I encountered a problem accessing the Kibana dashboard in Amazon ES.

A. Amazon ES is protected from public access. Make sure that your IP matches the input parameter **Remote Access CIDR**, which is white-listed for Amazon ES.

Additional Resources

AWS services

- Amazon Athena
  https://aws.amazon.com/documentation/athena/
- Amazon EBS
  https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AmazonEBS.html
- Amazon EC2
  https://aws.amazon.com/documentation/ec2/
- Amazon ES
  https://aws.amazon.com/documentation/elasticsearch-service/
- Amazon Kinesis
  https://aws.amazon.com/documentation/kinesis/
- Amazon QuickSight
  https://aws.amazon.com/documentation/quicksight/
- Amazon Redshift
  https://aws.amazon.com/documentation/redshift/
- Amazon Redshift Spectrum
  http://docs.aws.amazon.com/redshift/latest/dg/c-using-spectrum.html
- Amazon S3
  https://aws.amazon.com/documentation/s3/
- Amazon SageMaker
  https://aws.amazon.com/documentation/sagemaker/
- Amazon VPC
  https://aws.amazon.com/documentation/vpc/
- AWS CloudFormation
  https://aws.amazon.com/documentation/cloudformation/
- AWS Glue
  https://aws.amazon.com/documentation/glue/
- Kibana plug-in
  https://aws.amazon.com/elasticsearch-service/kibana/

47Lining Data Lake Resources

- Data lake foundational concepts
  http://www.47lining.com/datalake/foundational-concepts/
- Data lake reference architecture
  http://www.47lining.com/datalake/reference-architecture/
- Data lake sample dataset details
  http://www.47lining.com/datalake/quickstart-sample-dataset/

Quick Start reference deployments

- AWS Quick Start home page
  https://aws.amazon.com/quickstart/

GitHub Repository

You can visit our GitHub repository to download the templates and scripts for this Quick Start, to post your comments, and to share your customizations with others.

Document Revisions

Date	Change	In sections
June 2018	Added support for Amazon SageMaker	Changes in templates and throughout guide
December 2017	Added integration with AWS Glue	Changes in AWS CloudFormation templates and throughout guide
September 2017	Initial publication	—

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/ 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.