Streaming Data Platform on the AWS Cloud

Quick Start Reference Deployment

February 2020

German Osin, Provectus Inc.
Shivansh Singh, AWS Quick Start team

Visit our GitHub repository for source files and to share feedback, report errors, or submit feature ideas for this Quick Start.

Contents

This Quick Start was created by Provectus Inc. in collaboration with Amazon Web Services (AWS).

Quick Starts are automated reference deployments that use AWS CloudFormation templates to deploy key technologies on AWS, following AWS best practices.

Overview

This Quick Start reference deployment guide provides step-by-step instructions for deploying Streaming Data Platform on the AWS Cloud.

This Quick Start is for users who are interested in enabling real-time analytics and want to explore capabilities of a streaming-first data platform.

Streaming Data Platform on AWS

Streaming Data Platform is a unified solution that enables real-time data analytics and serves as a foundational service for AI solutions. Provectus's streaming-first architecture powers and provides governance for a data lake ecosystem. This solution consolidates data pipelines and improves scalability in the cloud for real-time analysis. It accelerates time-to-market and mitigates technology risks.

Cost 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.

For cost estimates, see the pricing pages for each AWS service you will be using. Prices are subject to change.

> **Tip** After you deploy the Quick Start, we recommend that you enable the AWS Cost and Usage Report to track costs associated with the Quick Start. This report delivers billing metrics to an Amazon Simple Storage Service (Amazon S3) bucket in your account. It provides cost estimates based on usage throughout each month and finalizes the data at the end of the month. For more information about the report, see the AWS documentation.

Because this Quick Start uses AWS native solution components, there are no costs or license requirements beyond AWS infrastructure costs.

Architecture

Deploying this Quick Start with **default parameters** builds the following Streaming Data Platform environment in the AWS Cloud.

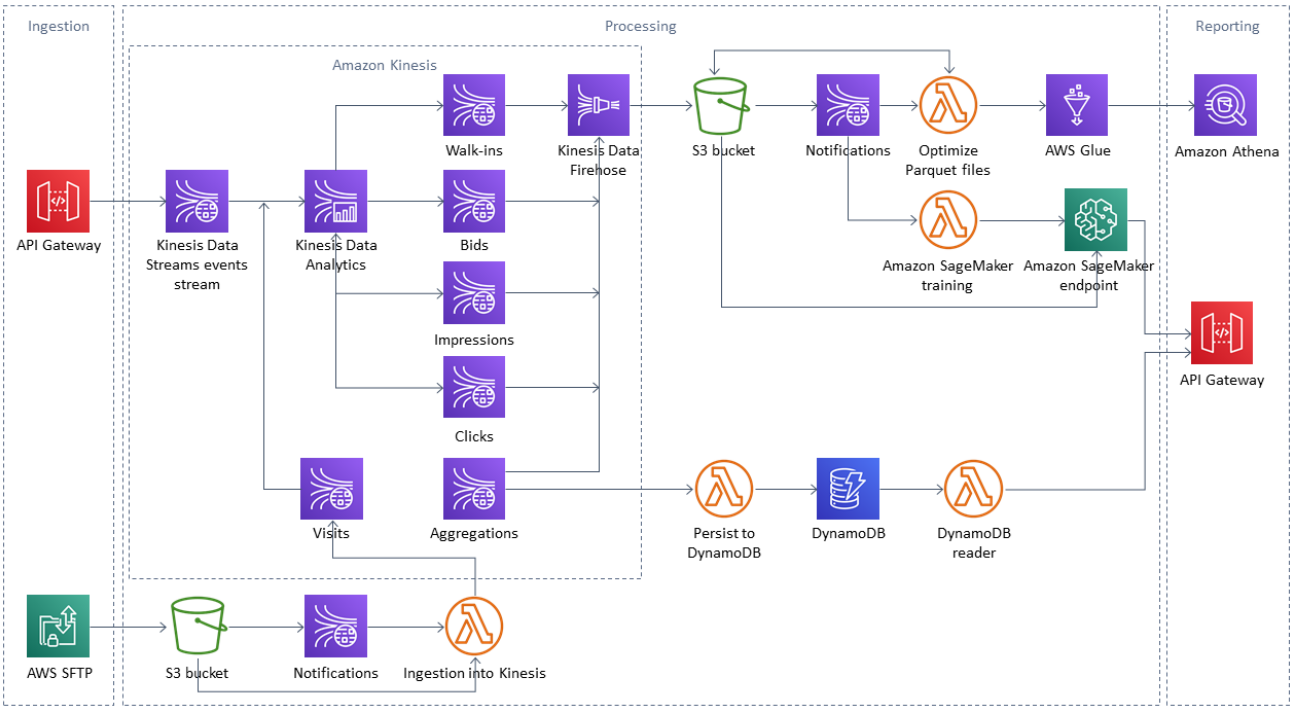


Figure 1: Quick Start architecture for Streaming Data Platform on AWS

The Quick Start sets up and configures the following:

- Amazon API Gateways to consume click stream by HTTP endpoints and for near real-time reports
- AWS Transfer for SFTP (AWS SFTP) for data flow

- Amazon Kinesis Data Streams for each type of event and aggregated message (a total of five data streams)
- Amazon Kinesis data analytics application for event enrichment and aggregation
- Amazon DynamoDB database to persist aggregated messages and provide reporting
- AWS Glue database with tables for each type of message
- AWS Lambda functions for file ingestion, data optimization, and continuous machine learning training
- Amazon SageMaker endpoint for predictions
- S3 bucket to store data in columnar format

Planning the deployment

Specialized knowledge

If you're new to AWS, visit the Getting Started Resource Center and the AWS Training and Certification website.

This deployment guide requires a moderate level of familiarity with the following AWS services.

- Amazon API Gateway
- Amazon Athena
- Amazon DynamoDB
- AWS Glue
- Amazon Kinesis
- Amazon S3
- Amazon SageMaker
- AWS SFTP
- Amazon VPC

AWS account

If you don't already have an AWS account, create one at https://aws.amazon.com by following the on-screen instructions. Part of the sign-up process involves receiving a phone call and entering a PIN using the phone keypad.

Your AWS account is automatically signed up for all AWS services. You are charged only for the services you use.

Technical requirements

Before you launch the Quick Start, your account must be configured as specified in the following table. Otherwise, deployment might fail.

Resources	If necessary, request service quota increases for the following resources. You might need to do this if an existing deployment uses these resources, and you might exceed the default quotas with this deployment. The Service Quotas console displays your usage and quotas for some aspects of some services. For more information, see the AWS documentation.

Resource	This deployment uses
AWS Identity and Access Management (IAM) roles	15
API Gateway account	1
API Gateway deployments	2
API Gateway Rest APIs	2
DynamoDB table	1
Glue database	1
Glue tables	8
Kinesis data analytics application	1
Kinesis Data Firehose delivery streams	4
Kinesis data streams	9
Lambda functions	9
Lambda permissions	2
Amazon CloudWatch Logs groups	12
CloudWatch Logs streams	2
S3 bucket	1
SageMaker model	1
SageMaker endpoint	1

Regions	This deployment includes Kinesis Data Analytics and Amazon Athena, which aren't currently supported in all AWS Regions. For a current list of supported Regions, see Service Endpoints and Quotas in the AWS documentation.
Key pair	Make sure that at least one Amazon EC2 key pair exists in your AWS account in the Region where you plan to deploy the Quick Start. Make note of the key pair name. You need it during deployment. To create a key pair, follow the instructions in the AWS documentation. For testing or proof-of-concept purposes, we recommend creating a new key pair instead of using one that's already being used by a production instance.
IAM permissions	Before launching the Quick Start, you must log in to the AWS Management Console with IAM permissions for the resources and actions the templates deploy. The *AdministratorAccess* managed policy within IAM provides sufficient permissions, although your organization may choose to use a custom policy with more restrictions.

Deployment steps

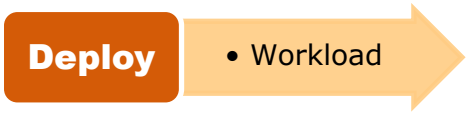
Step 1. Sign in to your AWS account

1. Sign in to your AWS account at https://aws.amazon.com with an IAM user role with the necessary permissions. For details, see Planning the deployment featured earlier in this guide.
2. Make sure that your AWS account is configured correctly, as discussed in the Technical requirements section of this guide.

Step 2. Launch the Quick Start

> **Notes** The instructions in this section 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.
>
> 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. Sign in to your AWS account and choose one of the following options to launch the AWS CloudFormation template. For help with choosing an option, see deployment options earlier in this guide.



Deploy Streaming Data Platform on AWS

Each deployment takes about 20 minutes to complete.

2. Check the AWS Region that is displayed in the upper right corner of the navigation bar and change it if necessary. The network infrastructure for Streaming Data Platform will be built accordingly. The template is launched in the US West (Oregon) Region by default.

> **Note:** This deployment includes Kinesis Data Analytics and Amazon Athena, which aren't currently supported in all AWS Regions. For a current list of supported Regions, see Service Endpoints and Quotas.

1. On the **Select Template** page, use the default settings for the template URL, and then choose **Next**.
2. 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.

   In the following tables, parameters are listed by category.

   When you finish reviewing and customizing the parameters, choose **Next**.

   View template

   AWS Quick Start Configuration:

> **Note** We recommend keeping the default settings for the following two parameters, unless you are customizing the Quick Start templates for your own deployment projects. Changing these parameter settings automatically updates code references to point to a new Quick Start location. For additional details, see the AWS Quick Start Contributor's Guide.

Parameter label (name)	Default	Description
Prefix for the stack resources (`ServicePrefix`)	*Requires input.*	Prefix for the stack resources.

Parameter label (name)	Default	Description
Quick Start S3 bucket name (QSS3BucketName)	aws-quickstart	The S3 bucket you created for your copy of Quick Start assets, if you decide to customize 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-provectus-streaming-data-platform/	The S3 key name prefix used to simulate a folder for your copy of Quick Start assets. You need to use this if you want to customize the Quick Start for your own use. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes.

General Configuration:

Parameter label (name)	Default	Description
Buffering interval (BufInterval)	60	Amazon Kinesis Data Firehose buffering interval, in seconds.
Buffering size (BufSize)	50	Amazon Kinesis Data Firehose buffer size.
Period of triggering aggregation function (AggregationPeriod)	10	Aggregation period to be collected and persisted on Amazon S3 and DynamoDB.
Session length (in minutes) for joining bids with impressions (BidsSessionTimeout)	5	Session length (in minutes) for joining bids with impressions.
Number of Kinesis streams shards (ShardCount)	2	Number of shards to be created in all data streams.
Session length (in minutes) for joining impressions with clicks (ClicksSessionTimeout)	5	Session length (in minutes) for joining impressions with clicks.
Session length (in minutes) for joining locations with clicks and impressions (LocationsSessionTimeout)	10	Session length (in minutes) for joining clicks and impressions.
Default log level (LogLevel)	debug	Default Lambda log levels.

3. On the **Options** page, you can specify tags (key-value pairs) for resources in your stack and set advanced options. When you've finished doing that, choose **Next**.
4. On the **Review** page, review and confirm the template settings. Under **Capabilities**, select two check boxes to acknowledge that the template will create IAM resources and that it may require the capability to auto-expand macros.
5. Choose **Create** to start deploying the stack.
6. Monitor the status of the stack. When the status is **CREATE_COMPLETE**, the Streaming Data Platform cluster is ready.

- Use the URLs displayed in the **Outputs** tab for the stack to view the resources that have been created.

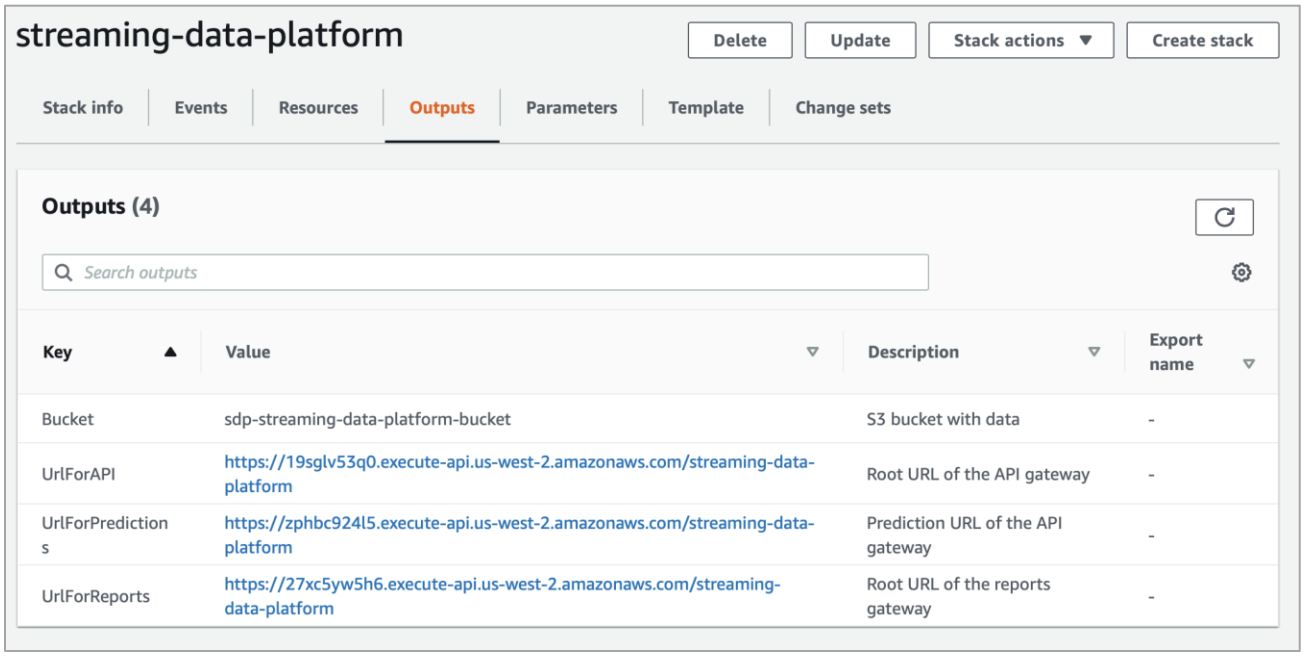


Figure 2: Streaming Data Platform outputs after successful deployment

Step 3. Using the deployment

To ingest data and get results, use the guidelines at https://github.com/provectus/streaming-data-platform#user-guide.

Best practices for using Streaming Data Platform on AWS

Start by listing data sources and consumers, and prioritize batch workloads that will benefit from migration into streaming.

Use the provided connectors and transformers as templates or blueprints for your own implementation.

Streaming Data Platform is built entirely on Kinesis Data Streams and Apache Flink, so it doesn't require a separate Apache Spark (Amazon EMR) cluster for stream processing. This allows to you to reduce the cost and complexity of processing in a stream upon data ingestion into Amazon Kinesis.

Use Apache Parquet as the default data format of choice.

For Apache Flink transformers and Parquet access, consider using more native Java/Scala implementations instead of Python SDK.

Troubleshooting

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

A. If AWS CloudFormation fails to create the stack, relaunch the template with **Rollback on failure** set to **No**. (This setting is under **Advanced** in the AWS CloudFormation console, the **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.

> **Important:** When you set **Rollback on failure** to **No**, you will continue to incur AWS charges for this stack. Please make sure to delete the stack when you finish 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. Launch the Quick Start templates using the links featured in this guide or from another Amazon S3 bucket. If you deploy the templates from a local copy on your computer or from a non-S3 location, you might encounter template size limitations when you create the stack. For more information about AWS CloudFormation limits, see the AWS documentation.

Send us feedback

To share feedback, submit feature ideas, or report errors, please use the **Issues** section of the GitHub repository for this Quick Start. If you'd like to submit code, please review the Quick Start Contributor's Guide.

Additional resources

AWS resources

- Getting Started Resource Center
- AWS General Reference
- AWS Glossary

AWS services

- Amazon API Gateway
- Amazon Athena
- Amazon DynamoDB
- AWS Glue
- IAM
- Amazon Kinesis
- AWS Lambda
- Amazon S3
- Amazon SageMaker
- AWS SFTP
- Amazon VPC

Provectus documentation

- Streaming Data Platform

Other Quick Start reference deployments

- AWS Quick Start home page

Document revisions

Date	Change	In sections
February 2020	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.