Guía de implementación

Sala de espera virtual en AWS

Sala de espera virtual en AWS: Guía de implementación

Table of Contents

Absorba grandes ráfagas de tráfico a su sitio web con la sala de espera virtual activada AWS

Fecha de publicación: noviembre de 2021 (última actualización: junio de 2024)

La AWS solución Virtual Waiting Room on ayuda a controlar las solicitudes de los usuarios entrantes a su sitio web durante las grandes ráfagas de tráfico. Crea una infraestructura en la nube diseñada para descargar temporalmente el tráfico entrante a su sitio web y ofrece opciones para personalizar e integrar una sala de espera virtual. Esta solución se puede integrar con sitios web nuevos o existentes para escalar sin problemas y gestionar los picos repentinos de tráfico.

Algunos ejemplos de eventos a gran escala que podrían provocar un aumento en el tráfico del sitio web incluyen:

- Inicio de la venta de entradas para conciertos o eventos deportivos
- Venta al por menor u otra venta minorista importante, como el Black Friday
- Lanzamiento de nuevos productos con amplios anuncios de marketing
- Acceso a los exámenes y asistencia a clases para las pruebas y las clases en línea
- Liberación de las franjas horarias para citas médicas
- Lanzamiento de un nuevo direct-to-customer servicio que requiere la creación de cuentas y los pagos

La solución actúa como un área de espera para los visitantes de su sitio web y permite el paso del tráfico cuando hay suficiente capacidad. El software cliente utilizado por los visitantes se puede configurar para permitir el tráfico de forma transparente a través de la sala de espera hasta que el sitio web esté al máximo de su capacidad, momento en el que la sala de espera frena a los visitantes. Cuando su sitio web tiene capacidad para recibir más tráfico, la solución genera JSON Web Tokens (JWT) que permiten a los usuarios acceder al sitio web. Por ejemplo, si tiene un evento que dura dos horas y su sitio web puede procesar 50 usuarios por segundo, pero espera un volumen de 250 por segundo, puede utilizar esta solución para regular el tráfico y, al mismo tiempo, permitir que los usuarios mantengan su posición en la cola.

Esta solución ofrece las siguientes funciones clave:

- Colocación estructurada de los usuarios en su sitio web

- Escalabilidad para controlar el tráfico de eventos de gran tamaño
- Generación de un token web JSON para permitir la entrada al sitio de destino
- Toda la funcionalidad se controla mediante las API REST
- Autorizador API Gateway listo para usar para soluciones de clientes
- Integración independiente o uso con OpenID

Esta guía de implementación describe las consideraciones arquitectónicas y los pasos de configuración para implementar Virtual Waiting Room AWS en la nube de Amazon Web Services (AWS). Incluye enlaces a AWS CloudFormationplantillas que lanzan y configuran los AWS servicios necesarios para implementar esta solución utilizando las AWS mejores prácticas de seguridad y disponibilidad.

La guía está dirigida a arquitectos de TI, desarrolladores, DevOps personal, analistas de datos y profesionales de la tecnología de marketing que tengan experiencia práctica en la arquitectura en la AWS nube.

Costo

Usted es responsable del coste de los AWS servicios utilizados durante la ejecución de esta solución. A partir de esta revisión, el coste de ejecutar esta solución con la configuración predeterminada en la región EE.UU. Este (Virginia del Norte) es de aproximadamente 10 USD al día por pila, más los cargos por las solicitudes de API y el tráfico de datos en función del tamaño del evento.

Coste diario de mantenimiento de la solución sin ningún incidente

AWS service	Solicitudes/tiempo	Costo [USD]
Amazon API Gateway	0	0,00$
Amazon CloudFront	0	0,00$
Amazon CloudWatch	0	0,00$
Amazon DynamoDB	0	0,00$
Amazon ElastiCache	Horas de nodo de cómputo (Redis)	~6,00 $
AWS Lambda	Nivel gratis*	0,00$
AWS Secrets Manager	Nivel gratuito*	0,00$
Amazon Simple Storage Service (Amazon S3)	Nivel gratuito*	0,00$
Amazon Virtual Private Cloud (Amazon VPC)	Horas de punto final de VPC Horario de la pasarela NAT	~5,00 $
TOTAL:		~11,00 $

*La estimación del costo se basa en un entorno limpio. Si utiliza este servicio de AWS fuera de esta solución, es posible que supere la cuota de la capa gratuita.

En las tablas siguientes se muestran los costes estimados de una sala de espera para 50 000 usuarios y 100 000 usuarios, con una duración del evento de 2 a 4 horas, con 500 usuarios por segundo entrantes y 1000 usuarios/minuto salientes. Los precios están sujetos a cambios. Para obtener más información, consulte la página web de precios de cada servicio utilizado en esta solución. AWS

Coste estimado para 50 000 usuarios de la sala de espera durante un evento de 2 horas

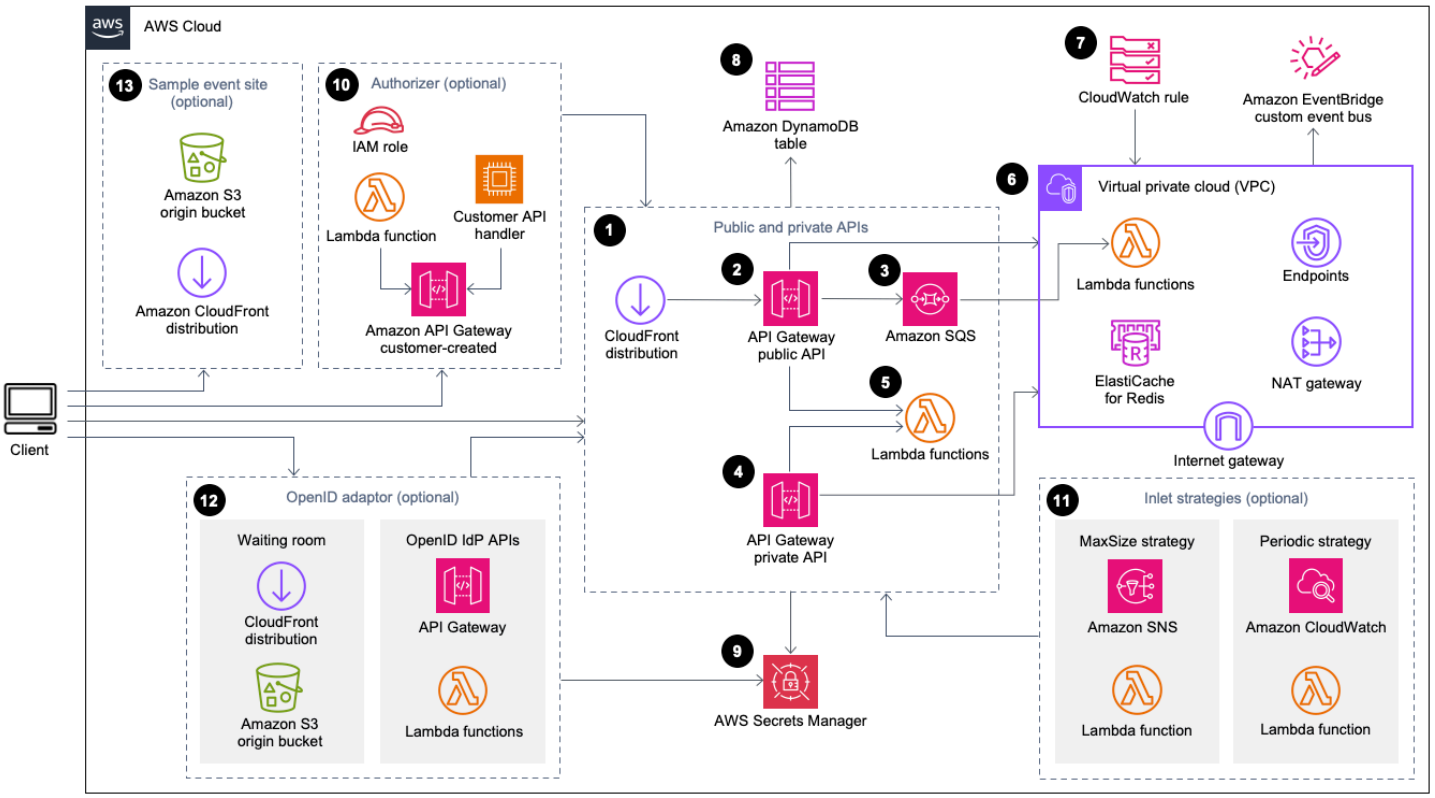
AWS service	Dimensiones	Coste [USD]
Amazon API Gateway	Solicitudes	2,00$
CloudFront	Solicitudes, ancho de banda	75,00$
CloudWatch	Métricas, alarmas, almacenamiento	1,00$
CloudWatch Eventos de Amazon	Eventos	1,00$
DynamoDB	Unidades de lectura/escritura, almacenamiento	1,00$
ElastiCache	Horas de nodo	8,00 DÓLARES
Lambda	Solicitudes, tiempo de cómputo	1,00$
AWS Secrets Manager	Secretos, solicitudes	1,00$
Amazon S3	Solicitudes, almacenamiento	1,00$
Amazon VPC	Transferencia de datos, hora de finalización	2,00$
TOTAL		94,00 DÓLARES

Coste estimado para 100 000 usuarios de la sala de espera durante un evento de 2 horas

AWS service	Dimensiones	Coste [USD]
Amazon API Gateway	Solicitudes	4,00$
CloudFront	Solicitudes, ancho de banda	296,00 DÓLARES
CloudWatch	Métricas, alarmas, almacenamiento	1,00$
CloudWatch Eventos	Eventos	1,00$
DynamoDB	Unidades de lectura/escritura, almacenamiento	4,00$
ElastiCache	Horas de nodo	32,00 DÓLARES
Lambda	Solicitudes, tiempo de cómputo	1,00$
AWS Secrets Manager	Secretos, solicitudes	1,00$
Amazon Simple Queue Service (Amazon SQS)	Solicitudes	1,00$
Amazon S3	Solicitudes, almacenamiento	1,00$
Amazon VPC	Transferencia de datos, hora de finalización	6,00$
TOTAL		348,00 DÓLARES

Información general de la arquitectura

Al implementar esta solución con las plantillas necesarias y opcionales, utilizando los parámetros predeterminados, se crea el siguiente entorno en la AWS nube.



Sala de espera virtual sobre AWS arquitectura

Las AWS CloudFormation plantillas implementan la siguiente infraestructura:

1. Una CloudFront distribución de Amazon para entregar llamadas de API públicas para el cliente.
2. Recursos de la API pública de Amazon API Gateway para procesar las solicitudes de cola desde la sala de espera virtual, realizar un seguimiento de la posición de la cola y admitir la validación de los tokens que permiten el acceso al sitio web de destino.
3. Una cola de Amazon Simple Queue Service (Amazon SQS) para regular el tráfico a AWS Lambdala función que procesa los mensajes de la cola. En lugar de invocar la función Lambda para cada solicitud, la cola de SQS agrupa por lotes las ráfagas de solicitudes entrantes.
4. Recursos de API privados de API Gateway para respaldar las funciones administrativas.
5. Lambda funciona para validar y procesar las solicitudes de API públicas y privadas y devolver las respuestas adecuadas.

6. Amazon Virtual Private Cloud (VPC) para alojar las funciones de Lambda que interactúan directamente con el clúster de Amazon ElastiCache for Redis. Los puntos finales de la VPC permiten que las funciones de Lambda de la VPC se comuniquen con los servicios de la solución. Además, la puerta de enlace NAT permite que las funciones Lambda de la VPC conecten los CloudFront puntos finales e invaliden la caché según sea necesario.
7. Una CloudWatch regla de Amazon para invocar una función de Lambda que funciona con un bus de EventBridgeAmazon personalizado para transmitir periódicamente actualizaciones de estado.
8. Tablas de Amazon DynamoDB para almacenar datos de token, posición de cola y contador de entrega.
9. AWS Secrets Manager para almacenar claves para operaciones de token y otros datos confidenciales.
10. (Opcional) Componente de autorización compuesto por una función AWS Identity and Access Management (de IAM) y una función de autorización de Lambda para su uso con API Gateway.
11. (Opcional) Amazon Simple Notification Service (Amazon SNS) y Lambda funcionan para admitir dos estrategias de entrada. CloudWatch
12. (Opcional) Componente adaptador OpenID con funciones API Gateway y Lambda para permitir que un proveedor de OpenID autentique a los usuarios en su sitio web. CloudFront distribución con un bucket de Amazon Simple Storage Service (Amazon S3) para la página de la sala de espera de este componente.
13. CloudFront Distribución (opcional) con el depósito de origen de Amazon S3 para la aplicación web de sala de espera de muestra.

Cómo funciona la solución

En esta sección se describen los pasos del flujo de trabajo de una sala de espera AWS virtual a un alto nivel. Consulte la Guía para desarrolladores GitHub para obtener más información sobre cómo crear, personalizar e integrar una sala de espera para su sitio web.

La API pública de la sala de espera puede estar ubicada detrás del perímetro de seguridad del sitio o puede estar disponible sin autorización alguna. Según el enfoque que utilices para integrar la sala de espera con el sitio web, es posible que el usuario deba autenticarse primero en el sitio web antes de poder acceder a la sala de espera y obtener un puesto en la cola.

El software cliente debe tener el identificador del evento para entrar en la sala de espera y realizar otras solicitudes. Un ID de evento es un identificador único que se requiere para la mayoría de las solicitudes relacionadas con las API públicas y privadas. El ID de evento se establece durante la instalación de la pila de API principal. Durante el funcionamiento, el ID del evento se puede proporcionar como parámetro de URL o cookie a través de la página de la sala de espera; se puede proporcionar como parte de las solicitudes de autenticación o se puede distribuir a los clientes a través de una ruta de datos diferente.

Hay casos en los que el cliente necesita tanto el ID del evento como el ID de la solicitud para realizar determinadas llamadas a la API. El ID de solicitud es un identificador único emitido desde la sala de espera que representa a un cliente específico en la fila.

En los siguientes pasos se describe el flujo de solicitudes de la API para entrar en una cola, esperar a que la cola avance y salir de la sala de espera con un token de acceso al sitio web.

El usuario entra en la sala de espera:

1. Al usuario se le presenta una pantalla o página que representa el punto de entrada a la sala de espera. Eligen entrar en la cola y el software del cliente (navegador, móvil, dispositivo) llama a la API `assign_queue_num` pública para solicitar una posición en la cola.
2. API Gateway envía inmediatamente la solicitud de API a la cola de Amazon SQS.
3. La llamada a la `assign_queue_num` API se devuelve cuando la solicitud se coloca en la cola. El cliente recibe un identificador de solicitud único que se puede utilizar más adelante para recuperar la posición de la cola, la hora de la solicitud y un token de acceso.
4. La función `AssignQueueNum` Lambda recibe lotes de hasta diez solicitudes de la cola de SQS. El servicio Lambda distribuye las invocaciones para procesar varios lotes de solicitudes.

5. La función `AssignQueueNum` Lambda valida cada mensaje de su lote, incrementa el contador de colas ElastiCache para Redis y almacena cada solicitud en Redis con su posición de cola ElastiCache asociada.
6. Cada mensaje se elimina a medida que se procesa correctamente. Los mensajes que presentan una condición de error se vuelven a procesar una vez en un lote posterior. Tras un segundo error, se envían a una CloudWatchalarma dead-letter-queue conectada.
7. El cliente puede empezar a sondear la `queue_num` API después de recibir el ID de solicitud de la `assign_queue_num` llamada. El cliente envía el ID de evento y el ID de solicitud a la `queue_num` API y recibe una posición numérica en la cola o una respuesta que indica que la solicitud aún no se ha procesado. Es posible que el cliente necesite realizar esta llamada más de una vez durante eventos grandes. API Gateway invoca la función `GetQueueNum` Lambda y devuelve la posición numérica del cliente en la cola desde DynamoDB.

El usuario espera en la sala de espera:

8. Una vez que el cliente ocupe su posición en la cola, puede empezar a sondear la `serving_num` API a intervalos regulares. Se llama a la `serving_num` API con el ID del evento y devuelve la posición de servicio actual de la cola. La respuesta de la `serving_num` API indica al cliente cuándo puede pasar de la sala de espera al sitio de destino real, donde se puede realizar la transacción final. La función `GetServingNum` Lambda devuelve la posición de servicio actual de la sala de espera.
9. Cuando la posición de servicio es igual o superior a la posición en la cola (solicitud) del cliente, el cliente puede solicitar un token web JSON (JWT) desde la API pública. El token se puede usar con el sitio de destino para finalizar la transacción. Se llama a la `generate_token` API con el ID de evento y el ID de solicitud. API Gateway invoca la función `GenerateToken` Lambda con los parámetros.
10. La función `GenerateToken` Lambda valida la solicitud y comprueba si este token se ha generado previamente. La función Lambda consulta la tabla de DynamoDB en busca de un token coincidente. Si se encuentra, ese token se devuelve a la persona que llama y no se regenera. Este proceso evita que se utilice un único identificador de solicitud para generar varios tokens diferentes con nuevos tiempos de caducidad.
11. Si el token no se encuentra en DynamoDB, la función Lambda recupera las claves para crear el token y lo guarda en DynamoDB con el ID de evento y el ID de solicitud del cliente. La función Lambda escribe un evento EventBridge para indicar que se ha generado un nuevo token. La

función Lambda incrementa un contador de Redis que realiza un seguimiento del número de fichas generadas ElastiCache para el evento.

12. Si `queue_pos_expiry` está activada, el cliente puede consultar el tiempo restante antes de que caduque llamando a la `queue_pos_expiry` API que invoca la función Lambda`GetQueuePositionExpiryTime`.

El usuario sale de la sala de espera:

13. Cuando el cliente recibe su token, entra en el sitio de destino para comenzar su transacción. En función de cómo su infraestructura soporte la integración con JWT, es posible que el cliente deba presentar el token en un encabezado de solicitud, en una cookie o de otra forma. El autorizador de API Gateway se puede utilizar para validar el token incluido en la solicitud de un cliente. Se puede utilizar cualquier biblioteca comercial o de código abierto para validar y gestionar los JWT con Virtual Waiting Room en los tokens. AWS Si el token es válido, el cliente puede continuar con la transacción.
14. Una vez que el cliente completa la transacción, se llama a una API privada para actualizar el estado del token del cliente y se completa en DynamoDB.

Caducidad de las posiciones de cola:

15. Cuando esta función está activada, el ID de solicitud correspondiente a una posición de cola concreta solo puede generar un token durante un intervalo de tiempo específico.

Aumente el contador de servicio al expirar la posición de la cola:

16. Cuando se activa esta función, el contador de servicio se incrementa automáticamente en función de las posiciones de cola caducadas que no pudieron generar fichas.

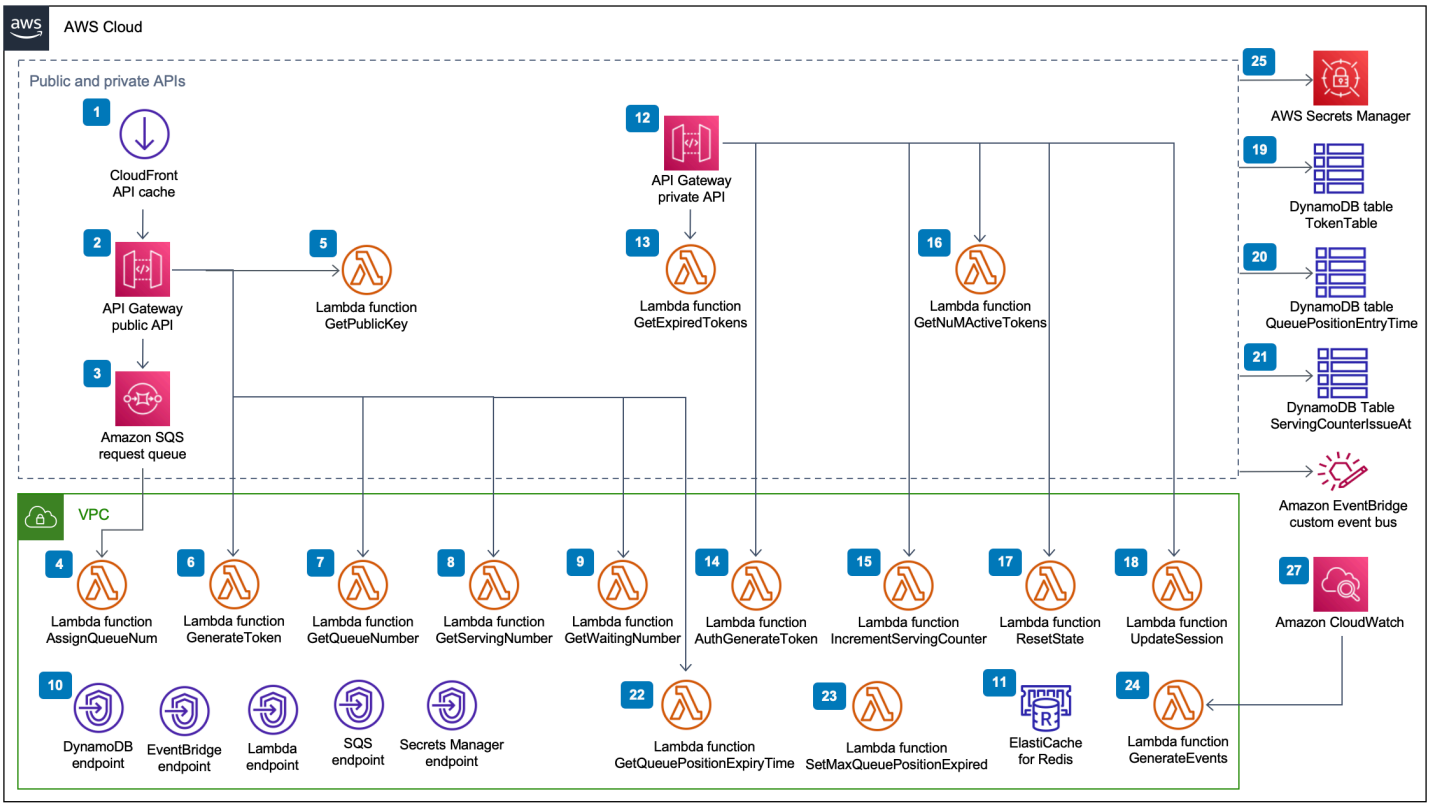
Componentes de la solución

API públicas y privadas de la sala de espera

El objetivo principal de la AWS solución Virtual Waiting Room on es controlar la generación de tokens web JSON (JWT) para los clientes de forma controlada, a fin de evitar las oleadas de nuevos usuarios que puedan sobrecargar el sitio web de destino. Los JWT se pueden utilizar para proteger el sitio, impidiendo el acceso a las páginas web hasta que se obtenga el token de la sala de espera, y también para autorizar el acceso a la API.

La plantilla principal instala una API pública y una API privada (autorizada por la IAM) que se utilizan en la mayoría de las operaciones de la sala de espera virtual. AWS La API pública se configura con una CloudFront distribución con varias políticas de almacenamiento en caché según la ruta de la API. Se crean una tabla EventBridge y un bus de eventos de DynamoDB. La plantilla añade una nueva VPC con dos zonas de disponibilidad (AZ), un clúster ElastiCache de Redis en ambas AZ y varias funciones Lambda. Las funciones de Lambda con ElastiCache las que interactúa Redis tienen interfaces de red dentro de la VPC y todas las demás funciones de Lambda tienen conectividad de red predeterminada. Las API principales son el nivel más bajo de interacción con la solución. Otras funciones de Lambda, la instancia de Amazon Elastic Compute Cloud (Amazon EC2) y los contenedores pueden actuar como extensiones y llamar a las API principales para crear salas de espera, controlar el tráfico entrante y reaccionar ante los eventos generados por la solución.

Además, la pila principal crea una alarma para todos sus errores de función Lambda y condiciones de aceleración, así como alarmas para cada implementación de API Gateway para los códigos de estado 4XX y 5XX.



Sala de espera virtual en el componente de API públicas y privadas de AWS

1. CloudFront la distribución ofrece llamadas a la API pública para el cliente y almacena en caché los resultados cuando corresponde.
2. La API pública de Amazon API Gateway procesa las solicitudes de cola de la sala de espera virtual, rastrea la posición de la cola y admite la validación de los tokens que permiten el acceso al sitio web de destino.
3. La cola SQS regula el tráfico a la AWS Lambda función que procesa los mensajes de la cola.
4. La función `AssignQueueNum` Lambda valida cada mensaje del lote recibido, incrementa el contador de colas ElastiCache para Redis y almacena cada solicitud en Redis con su posición de cola ElastiCache asociada.
5. La función `GetPublicKey` Lambda recupera el valor de la clave pública de Secrets Manager.
6. La función `GenerateToken` Lambda genera un JWT para una solicitud válida a la que se le ha permitido completar su transacción en el sitio de destino. Escribe un evento en el bus de eventos personalizado de la sala de espera en el que se indica que se ha generado un token. Si anteriormente se generó un token para esta solicitud, no se generará ningún token nuevo.

7. La función `GetQueueNumber` Lambda recupera y devuelve la posición numérica del cliente en la cola de Redis. ElastiCache
8. La función `GetServingNumber` Lambda recupera y devuelve el número que atiende actualmente la sala de espera de ElastiCache Redis.
9. La función `GetWaitingNum` Lambda devuelve el número actualmente en cola en la sala de espera y al que aún no se le ha emitido un token.
10. Los puntos finales de la VPC permiten que las funciones de Lambda de la VPC se comuniquen con los servicios de la solución.
11. ElastiCache en el caso de Redis, el clúster almacena todas las solicitudes de entrada a la sala de espera con un identificador de evento válido. También almacena varios contadores, como el número de solicitudes en cola, el número que se están atendiendo actualmente, el número de tokens generados, el número de sesiones completadas y el número de sesiones abandonadas.
12. Recursos de API privados de API Gateway para respaldar las funciones administrativas. Las API privadas están autenticadas por AWS IAM.
13. La función `GetExpiredTokens` Lambda devuelve una lista de identificadores de solicitud con tokens caducados.
14. La función `AuthGenerateToken` Lambda genera un token para una solicitud válida a la que se le ha permitido completar su transacción en el sitio de destino. Se pueden anular el emisor y el período de validez de un token establecidos inicialmente durante la implementación del core stack. Escribe un evento en el bus de eventos personalizado de la sala de espera en el que se indica que se ha generado un token. Si se ha generado previamente un token para esta solicitud, no se generará ningún token nuevo.
15. La función `IncrementServingCounter` Lambda incrementa el mostrador de servicio de la sala de espera almacenado en Redis dado un incremento ElastiCache por valor.
16. La función `GetNumActiveTokens` Lambda consulta a DynamoDB el número de tokens que aún no han caducado, que no se han utilizado para completar la transacción y que no se han marcado como abandonados.
17. La función `ResetState` Lambda restablece todos los contadores almacenados en ElastiCache Redis. También elimina y vuelve a crear las tablas `TokenTableQueuePositionEntryTime`, y DynamoDB`ServingCounterIssuedAt`. Además, invalida la caché. CloudFront
18. La función `UpdateSession` Lambda actualiza el estado de una sesión (token) almacenada en la tabla de DynamoDB`TokenTable`. El estado de la sesión se indica mediante un número entero. Las sesiones configuradas en un estado de `1` indican finalizadas e `-1` indican abandonadas.

Escribe un evento en el bus de eventos personalizado de la sala de espera en el que se indica que se ha actualizado una sesión.

19 La tabla de `TokenTable` DynamoDB almacena datos de token.

20 La tabla `QueuePositionEntryTime` DynamoDB almacena los datos de posición de cola y tiempo de entrada.

21 La tabla de `ServingCounterIssuedAt` DynamoDB almacena las actualizaciones del contador de servicio.

22 La función `GetQueuePositionExpireTime` Lambda se invoca cuando el cliente solicita el tiempo de caducidad de la posición de cola restante.

23 La función `SetMaxQueuePositionExpired` Lambda establece la posición máxima de la cola que ha caducado correspondiente a los valores de la `ServingCounterIssuedAt` tabla. Se ejecuta cada minuto si el `IncrSvcOnQueuePositionExpiry` parámetro se establece `true` durante la implementación de la pila principal.

24 La función `GenerateEvents` Lambda escribe varias métricas de la sala de espera en el bus de eventos personalizado de la sala de espera. Se ejecuta cada minuto si el parámetro Habilitar la generación de eventos está establecido `true` durante la implementación del core stack.

25 AWS Secrets Manager almacena las claves para las operaciones de los tokens y otros datos confidenciales.

26 El bus de eventos EventBridge personalizado de Amazon recibe un evento cada vez que se genera un token y se actualiza una sesión en la tabla de `TokenTable` DynamoDB. También recibe eventos cuando el contador de servicio se mueve en la `SetMaxQueuePositionExpired` Lambda. Se escribe con varias métricas de sala de espera, si se activa durante la implementación del core stack.

27 La regla de CloudWatch eventos de Amazon se crea si el parámetro Enable Events Generation se establece en true durante la implementación del core stack. Esta regla de eventos inicia la función `GenerateEvents` Lambda cada minuto.

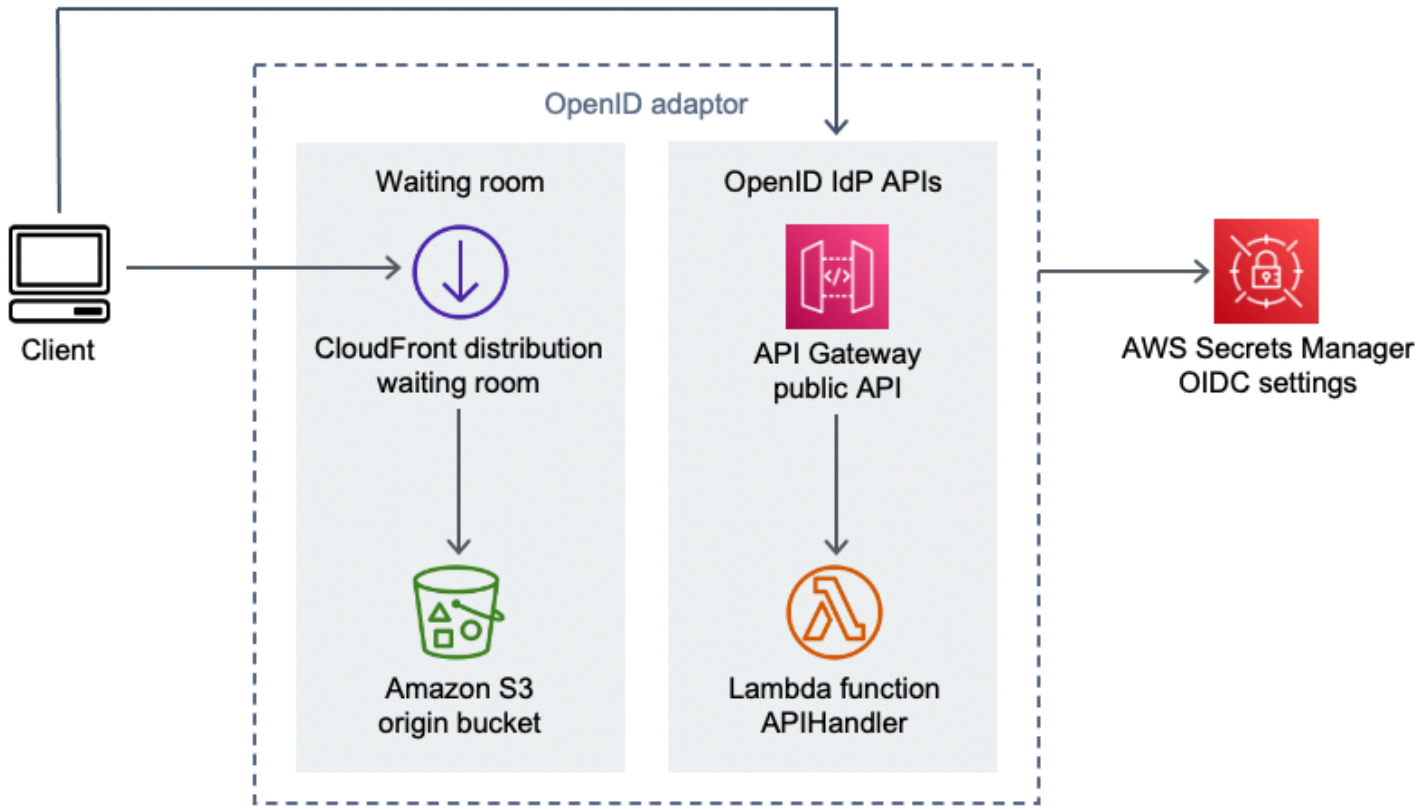
Autorizadores

La solución incluye una pila de autorizadores Lambda de API Gateway. La pila consta de una función de IAM y una función de Lambda. La función `APIGatewayAuthorizer` Lambda es un autorizador de API Gateway que puede validar la firma y las reclamaciones de un token emitido por la sala de espera virtual de la API. AWS La función Lambda que se suministra con la pila se puede utilizar para proteger las API de la nube hasta que el usuario haya pasado por la sala de espera y reciba un

token de acceso. El autorizador recupera y almacena en caché automáticamente la clave pública y la configuración de la API principal para verificar el token. Se puede usar sin modificaciones y se puede instalar en cualquier AWS región compatible. AWS Lambda

Adaptador OpenID

La pila de adaptadores OpenID implementa una API Gateway y funciones Lambda que actúan como un proveedor de identidad de OpenID. El adaptador OpenID proporciona un conjunto de API compatibles con OIDC que se pueden usar con el software de alojamiento web existente que admite los proveedores de identidad de OIDC, como AWS Elastic Load Balancers WordPress, o como proveedor de identidad federado para Amazon Cognito o un servicio similar. El adaptador permite al cliente utilizar la sala de espera en el flujo de Authn/Authz cuando utiliza un software de alojamiento off-the-shelf web con opciones de integración limitadas. La pila también instala una CloudFront distribución con un bucket de Amazon S3 como origen y otro bucket de S3 para registrar las solicitudes. El adaptador OpenID muestra una página de sala de espera, similar a la que se proporciona en la pila de salas de espera de muestra, pero diseñada para un flujo de autenticación OpenID. El proceso de autenticación implica conseguir un puesto en la cola de la sala de espera y esperar hasta que el puesto de servicio sea igual o mayor que el puesto de la cola del cliente. La página de la sala de espera de OpenID redirige de nuevo al sitio de destino, que utiliza la API OpenID para completar la adquisición del token y la configuración de la sesión para el cliente. Los puntos finales de la API de esta solución se asignan directamente a la especificación de flujo name-for-name oficial de OpenID Connect 1.0,. Consulte Autenticación OpenID Connect Core 1.0 para obtener más información.



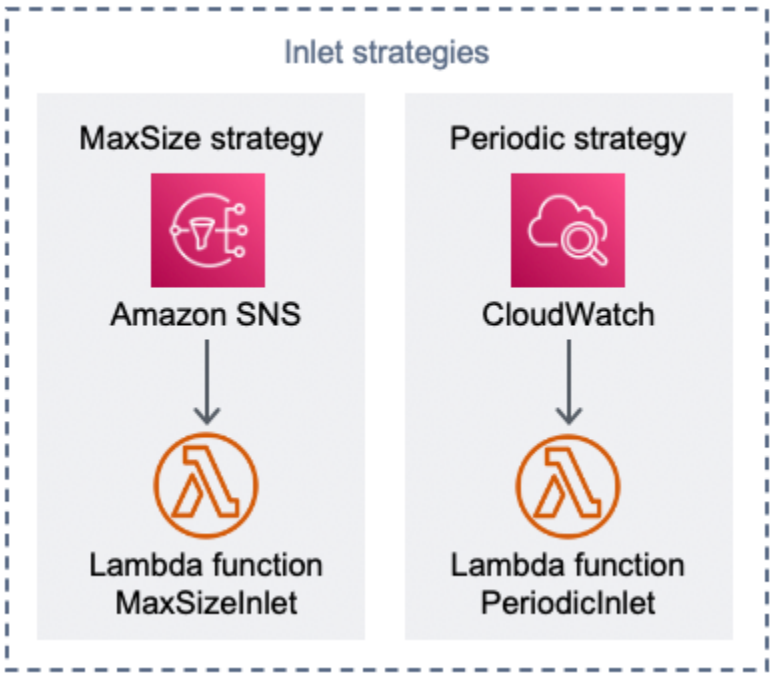
Sala de espera virtual en el AWS componente adaptador OpenID

1. CloudFront la distribución entrega el contenido del bucket S3 al usuario.
2. El bucket S3 aloja ejemplos de páginas de sala de espera.
3. La API Amazon API Gateway proporciona un conjunto de API compatibles con OIDC que se pueden utilizar con el software de alojamiento web existente que admite la función de autorización Lambda del proveedor de identidades OIDC.
4. La función `APIHandler` Lambda gestiona las solicitudes de todas las rutas de recursos de API Gateway. Se asignan diferentes funciones de Python dentro del mismo módulo a cada ruta de API. Por ejemplo, la ruta de `/authorize` recursos de API Gateway se invoca `authorize()` dentro de la función Lambda.
5. La configuración del OIDC se guarda en Secrets Manager.

Ejemplos de estrategias de entrada

Las estrategias de entrada determinan cuándo debe avanzar el mostrador de servicio de la solución para dar cabida a más usuarios en el sitio de destino. Para obtener más información conceptual sobre las estrategias de entrada a las salas de espera, consulte Consideraciones de diseño.

La solución ofrece dos ejemplos de estrategias de entrada: MaxSizey Periodic.



Componente de estrategias de sala de espera virtual en AWS Inlet

Opción de estrategia de entrada de tamaño máximo:

1. Un cliente emite una notificación de Amazon SNS que invoca la función `MaxSizeInlet` Lambda para aumentar el contador de servidores en función de la carga útil del mensaje.
2. La función `MaxSizeInlet` Lambda espera recibir un mensaje para determinar cuánto incrementar el contador de servicio.

Opción de estrategia de entrada periódica:

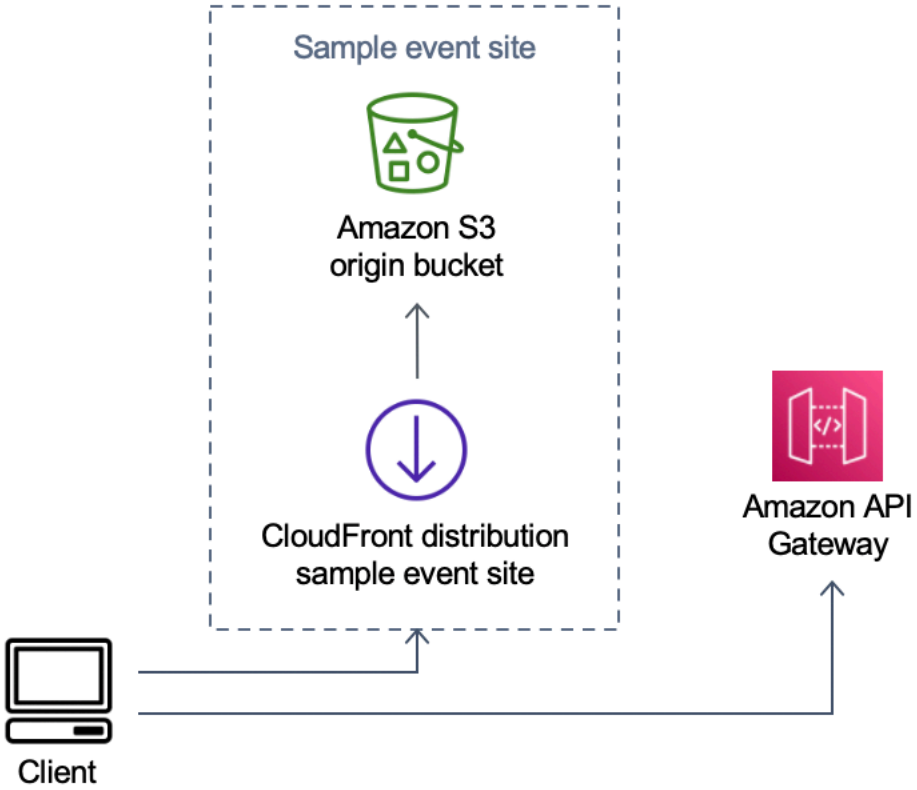
3. Una CloudWatch regla invoca una función Lambda cada minuto para aumentar el contador de porciones en una cantidad fija.
4. La función `PeriodicInlet` Lambda incrementa el contador de servicio en el tamaño indicado si el tiempo está comprendido entre la hora de inicio y la hora de finalización indicada.

Opcionalmente, comprueba una CloudWatch alarma y, si la alarma está activa, realiza el incremento; de lo contrario, lo omite. 0K

Ejemplo de sala de espera

El ejemplo de sala de espera se integra con las API públicas y privadas, además del autorizador personalizado, para demostrar una solución de sala de end-to-end espera mínima. La página web principal se almacena en un depósito de S3 y se utiliza como origen para CloudFront ella. Guía al usuario a través de los siguientes pasos:

1. Haga cola en la sala de espera para ingresar al sitio.
2. Obtenga la posición del cliente en la fila.
3. Obtenga la posición de servicio de la sala de espera.
4. Obtenga un juego de fichas una vez que la posición de servicio sea igual o superior a la posición del cliente.
5. Use el token para llamar a una API protegida por el autorizador Lambda.



AWS Ejemplo de componente de sala de espera virtual en el sitio del evento

1. El depósito S3 aloja el contenido de muestra para la sala de espera y el panel de control.
2. CloudFront la distribución entrega el contenido del bucket de S3 al usuario.
3. Ejemplo de implementación de API Gateway con rutas de recursos similares a las de compras, como y. `/search` `/checkout` Esta API se instala por pila y se configura con el autorizador de tokens. Pretende ser un ejemplo de una forma sencilla de proteger una API en la sala de espera. Las solicitudes que presentan un token válido se reenvían a la Lambda; de lo contrario, se devuelve un error. La API no tiene ninguna funcionalidad aparte de la respuesta de la función Lambda adjunta.

Seguridad

Cuando crea sistemas en una AWS infraestructura, las responsabilidades de seguridad se comparten entre usted y AWS. Este modelo compartido reduce la carga operativa porque AWS opera, administra y controla los componentes, incluidos el sistema operativo anfitrión, la capa de virtualización y la seguridad física de las instalaciones en las que operan los servicios. Para obtener más información sobre AWS la seguridad, visite AWS Cloud Security.

ElastiCache para Redis se le asigna una interfaz de red dentro de la VPC privada. A las funciones de Lambda con ElastiCache las que interactúa Redis también se les asignan interfaces de red dentro de una VPC. Todos los demás recursos tienen conectividad de red en el espacio de red compartido AWS . Las funciones Lambda con interfaces de VPC que interactúan con otros servicios AWS utilizan puntos finales de VPC para conectarse a estos servicios.

Las claves públicas y privadas que se utilizan para crear y validar los tokens web JSON se generan en el momento de la implementación y se almacenan en Secrets Manager. La contraseña utilizada para conectarse a ElastiCache Redis también se genera en el momento de la implementación y se almacena en Secrets Manager. No se puede acceder a la clave privada ni a la contraseña de Redis a través de ninguna API de solución. ElastiCache

Se debe acceder a la API pública a través de CloudFront. La solución genera una clave de API para API Gateway, que se utiliza como valor de un encabezado personalizado,`x-api-key`, en CloudFront. CloudFront incluye este encabezado al realizar solicitudes de origen. Para obtener más información, consulta Cómo añadir encabezados personalizados a las solicitudes de origen en la Guía para CloudFront desarrolladores de Amazon.

Las API privadas están configuradas para requerir la autorización de AWS IAM para su invocación. La solución crea el grupo de usuarios de `ProtectedAPIGroup` IAM con los permisos adecuados para invocar las API privadas. Un usuario de IAM agregado a este grupo está autorizado a invocar las API privadas.

Las políticas de IAM utilizadas en los roles y permisos asociados a varios recursos creados por la solución otorgan solo los permisos necesarios para realizar las tareas necesarias.

En el caso de recursos como los depósitos de S3, las colas de SQS y los temas de SNS generados por la solución, el cifrado en reposo y durante el tránsito se activa siempre que es posible.

Monitorización

La pila principal de API incluye varias CloudWatch alarmas que se pueden monitorear para detectar problemas mientras la solución está en funcionamiento. La pila crea una alarma para los errores de la función Lambda y las condiciones del acelerador, y cambia el estado de la alarma de `OK` a `ALARM` si se produce un error o una condición de aceleración en un período de un minuto.

La pila también crea alarmas para cada implementación de API Gateway para los códigos de estado 4XX y 5XX. La alarma cambia de estado `OK` a `ALARM` si la API devuelve un código de estado 4XX o 5XX en un período de un minuto.

Estas alarmas vuelven a un `OK` estado después de un minuto sin errores ni interrupciones.

Roles de IAM

AWS Identity and Access Management Las funciones (IAM) permiten a los clientes asignar políticas y permisos de acceso detallados a los servicios y usuarios de la nube. AWS Esta solución crea funciones de IAM que permiten a las AWS Lambda funciones de la solución acceder a crear recursos regionales.

Amazon CloudFront

La `virtual-waiting-room-on-aws.template` CloudFormation plantilla, que crea las principales API públicas y privadas de la sala de espera, también implementa una CloudFront distribución para la API pública. CloudFront almacena en caché las respuestas de la API pública, lo que reduce la carga en API Gateway y en las funciones de Lambda que realizan su trabajo.

Esta solución también incluye un ejemplo opcional de plantilla de sala de espera que despliega una aplicación web sencilla alojada en un bucket de Amazon Simple Storage Service (Amazon S3). Para ayudar a reducir la latencia y mejorar la seguridad, se implementa una CloudFront distribución de Amazon con una identidad de acceso de origen, que es un CloudFront usuario que proporciona acceso público al contenido del bucket del sitio web de la solución. Para obtener más información, consulte Restringir el acceso al contenido de Amazon S3 mediante una identidad de acceso de origen en la Guía para CloudFront desarrolladores de Amazon.

Grupos de seguridad

Los grupos de seguridad de VPC creados en esta solución están diseñados para controlar y aislar el tráfico de red dirigido a Redis ElastiCache . Las lambdas que necesitan comunicarse con las ElastiCache de Redis se colocan en el mismo grupo de seguridad que las de Redis. ElastiCache Le recomendamos que revise los grupos de seguridad y restrinja aún más el acceso según sea necesario una vez que la implementación esté en marcha.

Consideraciones sobre el diseño

Opciones de implementación

Si es la primera vez que la instala, o si no está seguro de qué instalar, implemente la CloudFormation plantilla `virtual-waiting-room-on-aws-getting-started.template` anidada, que instala el núcleo, los autorizadores y las plantillas de sala de espera de muestra. Esto le proporciona una sala de espera mínima con un flujo sencillo.

Protocolos admitidos

La AWS solución Virtual Waiting Room On se puede integrar con lo siguiente:

- Bibliotecas y herramientas de verificación de JSON Web Token
- Implementaciones de API Gateway existentes
- Clientes de API REST
- Clientes y proveedores de OpenID

Estrategias de acceso a las salas de espera

Las estrategias de entrada encapsulan la lógica y los datos necesarios para trasladar a los clientes de la sala de espera al sitio web. Una estrategia de entrada se puede implementar como una función Lambda, un contenedor, una instancia de Amazon EC2 o cualquier otro recurso informático. No es necesario que sea un recurso en la nube siempre que pueda llamar a las API públicas y privadas de la sala de espera. La estrategia de entrada recibe eventos sobre la sala de espera, el sitio web u otros indicadores externos que le ayudan a decidir cuándo más clientes pueden solicitar la emisión de fichas y entrar en el sitio. Existen varios enfoques para las estrategias de entrada. El que se adopte dependerá de los recursos de los que disponga y de las limitaciones del diseño del sitio web que se va a proteger.

La acción principal que lleva a cabo la estrategia de entrada es llamar a la API privada de `increment_serving_num` Amazon API Gateway con un valor relativo que indique cuántos clientes más pueden entrar en el sitio. En esta sección se describen dos ejemplos de estrategias de entrada. Se pueden usar tal cual, personalizarlas o se pueden emplear un enfoque completamente diferente.

MaxSize

Con la MaxSize estrategia, la función `MaxSizeInlet` Lambda se configura con el número máximo de clientes que pueden utilizar el sitio web de forma simultánea. Se trata de un valor fijo. Un cliente emite una notificación de Amazon SNS que invoca la función `MaxSizeInlet` Lambda para aumentar el contador de servidores en función de la carga útil del mensaje. La fuente del mensaje de SNS puede provenir de cualquier parte, incluido el código del sitio web o una herramienta de monitoreo que observe el nivel de utilización del sitio.

La función `MaxSizeInlet` Lambda espera recibir un mensaje que puede incluir:

- `exited :`número de transacciones que se han completado
- lista de identificadores de solicitud que deben marcarse como finalizados
- lista de identificadores de solicitud que deben marcarse como abandonados

Estos datos se utilizan para determinar cuánto se debe incrementar el contador de servicio. Puede haber casos en los que no haya capacidad adicional para incrementar el contador, en función del número actual de clientes.

Periódico

Cuando se utiliza la estrategia periódica, una CloudWatch regla invoca la función `PeriodicInlet` Lambda cada minuto para aumentar el contador de porciones en una cantidad fija. La entrada periódica se parametriza con la hora de inicio, la hora de finalización y la cantidad de incremento del evento. Opcionalmente, esta estrategia también comprueba una CloudWatch alarma y, si la alarma está activa, realiza el incremento; de lo contrario, lo omite. `OK` Los integradores del sitio pueden conectar una métrica de utilización a una alarma y utilizar esa alarma para pausar la entrada periódica. Esta estrategia solo cambia la posición de servicio mientras la hora actual esté entre la hora de inicio y la hora de finalización y, opcionalmente, la alarma especificada esté en ese `OK` estado.

Personalización y ampliación de la solución

El administrador del sitio de su organización debe decidir los métodos de integración que se utilizarán con la sala de espera. Dispone de dos opciones:

1. Integración básica directamente mediante API y autorizadores de API Gateway.

2. Integración de OpenID a través de un proveedor de identidad.

Además de la integración anterior, es posible que tengas que configurar la redirección de nombres de dominio. También es responsable de implementar una página de sitio de sala de espera personalizada.

La AWS solución Virtual Waiting Room on está diseñada para ampliarse mediante dos mecanismos: EventBridge para la notificación unidireccional de eventos y las API REST para la comunicación bidireccional.

Cuotas

La principal limitación de escala para Virtual Waiting Room on AWS es el límite de aceleración Lambda para la región instalada. AWS Cuando se instala en una AWS cuenta con la cuota de ejecución simultánea predeterminada de Lambda, la AWS solución Virtual Waiting Room On puede gestionar hasta 500 clientes por segundo que soliciten un puesto en la cola. La tasa de 500 clientes por segundo se basa en que la solución tenga disponibles exclusivamente todos los límites de cuota simultánea de la función Lambda. Si la región de la cuenta se comparte con otras soluciones que invocan funciones Lambda, la sala AWS de espera virtual de la solución debe tener al menos 1000 invocaciones simultáneas disponibles. Puede usar CloudWatch métricas para graficar las invocaciones simultáneas de Lambda en su cuenta a lo largo del tiempo para tomar una decisión. Puede utilizar la consola Service Quotas para solicitar aumentos. El aumento del límite de regulación de Lambda solo aumenta los cargos mensuales de la cuenta si realmente se producen invocaciones adicionales.

Por cada 500 clientes adicionales por segundo, aumente el límite máximo en 1000.

Se esperan usuarios entrantes por segundo	Cuota de ejecución simultánea recomendada
0-500	1000 (predeterminado)
501 a 1000	2,000
1.001-1,500	3000

Lambda tiene un límite de ráfagas fijo de 3000 invocaciones simultáneas. Para obtener más información, consulte Escalado de funciones Lambda. El código del cliente debe esperar algunas

llamadas a la API y volver a intentarlo si se devuelve un código de error que indique una situación de aceleración temporal. El ejemplo de cliente de sala de espera incluye este código como ejemplo de cómo diseñar clientes que se utilicen en eventos de alta capacidad y ráfagas altas.

Esta solución también es compatible con la simultaneidad reservada y aprovisionada de Lambda con pasos de configuración personalizados. Para obtener más información, consulte Administración de la simultaneidad reservada de Lambda.

El límite máximo de usuarios que pueden entrar en la sala de espera, recibir un token y continuar con una transacción está limitado por el límite superior de los mostradores de ElastiCache Redis. Los mostradores se utilizan para ver el puesto de servicio en la sala de espera y para hacer un seguimiento resumido del estado de la solución. Los contadores utilizados en ElastiCache Redis tienen un límite máximo de 9.223.372.036.854.775.807. Se utiliza una tabla de DynamoDB para almacenar una copia de cada token emitido a un usuario de la sala de espera. DynamoDB no tiene ningún límite práctico en cuanto al tamaño de las tablas.

Implementaciones regionales

Los servicios que utiliza esta solución son compatibles en todas AWS las regiones. Para obtener la disponibilidad más actualizada de AWS los servicios por región, consulte la Lista de servicios AWS regionales.

AWS CloudFormation plantillas

Para automatizar la implementación, esta solución utiliza las siguientes AWS CloudFormation plantillas, que puede descargar antes de la implementación.

Si es la primera vez que realiza la instalación, o si no sabe qué instalar, implemente la `virtual-waiting-room-on-aws-getting-started.template` AWS CloudFormation plantilla, que instala el núcleo, los autorizadores y los ejemplos de plantillas de códigos de sala de espera. Esto le permite probar una sala de espera que funcione con un flujo sencillo.

virtual-waiting-room-on- aws-api-gateway-cw -logs-role.template: usa esta plantilla para agregar un ARN de rol predeterminado a API Gateway a nivel de cuenta para los permisos de registro. CloudWatch Consulte los requisitos previos para obtener más información sobre si su cuenta requiere el despliegue de esta plantilla o no.

virtual-waiting-room-on- aws-getting-started .template: utilice esta plantilla anidada para instalar el núcleo, los autorizadores y las pilas de salas de espera de muestra.

virtual-waiting-room-on-aws.template: utilice esta plantilla básica para instalar las principales API REST públicas y privadas y los servicios en la nube para crear eventos en las salas de espera. Instale esta plantilla en la cuenta y la región en las que necesite las API REST de la sala de espera, ElastiCache para Redis y la tabla DynamoDB.

virtual-waiting-room-on-aws-authorizers.template: utilice esta plantilla para instalar el autorizador Lambda diseñado para verificar los tokens emitidos por las salas de espera y destinado a proteger las API de los usuarios finales. Requiere la pila principal. Algunas salidas de la pila principal son necesarias como parámetros para implementar esta pila. Se trata de una plantilla opcional.

View template

virtual-waiting-room-on-aws-openid.template: utilice esta plantilla para instalar un proveedor de identidad OpenID para la integración de la sala de espera con las interfaces de autorización. Requiere la pila principal. Se necesitan algunos resultados de la pila principal para implementar esta pila. Se trata de una plantilla opcional.

View template

virtual-waiting-room-on- aws-sample-inlet-strategy .template: utilice esta plantilla para instalar estrategias de admisión de muestras diseñadas para usarse entre un sitio objetivo y la sala de espera. Las estrategias de entrada ayudan a encapsular la lógica para determinar cuándo permitir la entrada de más usuarios al sitio de destino. Requiere la pila principal. Los resultados de la pila principal son necesarios para implementar esta pila. Se trata de una plantilla opcional.

View template

virtual-waiting-room-on-aws-sample.template: utilice esta plantilla para instalar un ejemplo de configuración mínima de web y API Gateway para una sala de espera y un sitio de destino. Requiere la pila principal y la pila de autorizadores. Los resultados de las pilas principal y de autorizadores son necesarios como parámetros para implementar esta pila. Se trata de una plantilla opcional.

Implementación automatizada

Antes de lanzar la solución, revise el costo, la arquitectura, la seguridad de la red y otras consideraciones que se describen en esta guía. Siga las step-by-step instrucciones de esta sección para configurar e implementar la solución en su cuenta.

Tiempo de implementación: aproximadamente 30 minutos (solo para los primeros pasos)

Requisitos previos

- AWS permisos de consola de cuentas equivalentes al acceso de administrador.
- Active el CloudWatch registro desde API Gateway:
  - Inicie sesión en la consola de API Gateway y seleccione la región en la que planea instalar las pilas.

    Si tienes API existentes definidas en esta región:

    1. Seleccione cualquier API.
    2. En el menú de navegación de la izquierda, selecciona Configuración.
    3. Compruebe si hay un valor en el campo CloudWatch ARN del rol de registro.
    - Si no hay ningún ARN, instale el. `virtual-waiting-room-on-aws-api-gateway-cw-logs-role.template`
    - Si hay un ARN, comience por lanzar la pila de introducción.

    Si no hay ninguna API existente definida en esta región, instale la. `virtual-waiting-room-on-aws-api-gateway-cw-logs-role.template`
- Conocimiento de la arquitectura y los detalles de implementación del sitio de destino que se va a proteger.

Descripción general de la implementación

Siga los siguientes pasos para implementar esta solución en AWS. Para obtener instrucciones detalladas, siga los enlaces de cada paso.

Paso 1. Lance la pila de introducción

- Abre la AWS CloudFormation plantilla en tu cuenta. AWS

- Revise los parámetros de la plantilla e introduzca o ajuste los valores predeterminados según sea necesario.

Paso 2. (Opcional) Pruebe la sala de espera

- Genere AWS claves para llamar a las API seguras de IAM.
- Abre el panel de control de la sala de espera de muestras.
- Pruebe la sala de espera de muestras.

Paso 1. Lanza la pila de introducción

Esta AWS CloudFormation plantilla automatizada incluye las plantillas principales, las de autorización y las de ejemplo para salas de espera, lo que le permite ver y probar una sala de espera que funcione. Debe leer y comprender los requisitos previos antes de lanzar la pila.

> **Note**
>
> Usted es responsable del coste de los AWS servicios utilizados al ejecutar esta solución. Para obtener más información, visite la sección de costos de esta guía y consulte la página web de precios de cada AWS servicio utilizado en esta solución.

1. Inicie sesión en AWS Management Consoley seleccione el botón para lanzar la `virtual-waiting-room-on-aws-getting-started.template` AWS CloudFormation plantilla.

   Launch solution Como alternativa, puede descargar la plantilla como punto de partida para su propia implementación.
2. La plantilla se lanza en la región Este de EE. UU. (Norte de Virginia) de forma predeterminada. Para lanzar la solución en una AWS región diferente, utilice el selector de regiones de la barra de navegación de la consola.
3. En la página Crear pila, compruebe que la URL de la plantilla correcta esté en el cuadro de texto URL de Amazon S3 y seleccione Siguiente.
4. En la página Especificar los detalles de la pila, especifique un nombre para la pila. Para obtener información sobre las limitaciones de nombres de caracteres, consulte los límites de IAM y STS en la Guía del AWS Identity and Access Management usuario.

5. En Parámetros, revise los parámetros de esta plantilla de solución y modifíquelos según sea necesario. Esta solución utiliza los siguientes valores predeterminados.

Parámetro	Predeterminado	Descripción
ID del evento	`Sample`	ID único para esta instancia de la sala de espera, se sugiere el formato GUID.
Periodo de validez	`3600`	Periodo de validez del token en segundos.
Habilite la generación de eventos	`false`	Si se establece en`true`, las métricas relacionadas con la sala de espera se escriben en su bus de eventos cada minuto
Puerto Redis	`1785`	El número de puerto que se utilizará para conectarse al ElastiCache servidor Redis. Se recomienda no utilizar el puerto predeterminado ElastiCache para Redis de. `6379`
EnableQueuePositionExpiry	`true`	Si se establece en`false`, no se aplica el período de caducidad de las posiciones de cola.
QueuePositionExpiryPeriod	`900`	Es el intervalo de tiempo en segundos después del cual una posición de cola no es apta para generar un token.

Parámetro	Predeterminado	Descripción
IncrSvcOnQueuePosi tionExpiry	`false`	Si se establece en`true`, el contador de servicio avanza automáticamente en función de las posiciones de cola caducadas que no generaron fichas correctamente.

6. Elija Siguiente.
7. En la página Configurar opciones de pila, elija Siguiente.
8. En la página Revisar, revise y confirme la configuración. Marque la casilla para confirmar que la plantilla crea recursos AWS Identity and Access Management (IAM).
9. Elija Create stack (Crear pila) para implementar la pila.

   Puedes ver el estado de la pila en la AWS CloudFormation consola, en la columna Estado. Deberías recibir el estado CREATE_COMPLETE en aproximadamente 30 minutos.

Paso 2. (Opcional) Pruebe la sala de espera

Si implementaste la pila de introducción, los siguientes pasos te ayudarán a probar la funcionalidad de la sala de espera. Para completar las pruebas, necesita AWS claves con permisos para llamar a las API seguras de IAM de la pila principal.

Genere AWS claves para llamar a las API seguras de IAM

1. Cree o utilice un usuario de IAM en la AWS cuenta en la que se implementó la `aws-virtual-waiting-room-getting-started.template` CloudFormation plantilla.
2. Conceda al usuario de IAM acceso mediante programación. Al crear un nuevo conjunto de claves de acceso para el usuario de IAM, descargue el archivo de claves cuando se presente. Necesita la clave de acceso y la clave de acceso secreta del usuario de IAM para probar la sala de espera.
3. Añada el usuario de IAM al grupo de usuarios de IAM de ProtectedApiGroup creado por la plantilla.

Abra el panel de control de la sala de espera de muestras

1. Inicie sesión en la AWS CloudFormation consola y seleccione la pila de introducción de la solución.
2. Elija la pestaña Salidas.
3. En la columna Clave, busque ControlPanella URL y seleccione el valor correspondiente.
4. Abre el panel de control en una nueva pestaña o ventana del navegador.
5. En el panel de control, expanda la sección Configuración.
6. Introduzca el ID de la clave de acceso y la clave de acceso secreta que ha obtenido en Generar AWS claves para llamar a las API protegidas de IAM. Los puntos finales y el identificador del evento se rellenan a partir de los parámetros de la URL.
7. Selecciona Usar. El botón se activa después de haber proporcionado las credenciales.

Pruebe el ejemplo de sala de espera

1. En la AWS CloudFormation consola, seleccione la pila de introducción de la solución.
2. Elija la pestaña Salidas.
3. En la columna Clave, busque WaitingRoomla URL y seleccione el valor correspondiente.
4. Abre la sala de espera y, a continuación, selecciona Reservar para entrar en la sala de espera.
5. Vuelve a la pestaña del navegador que contiene el panel de control.
6. En Contador de porciones incrementadas, selecciona Cambiar. Esto permite que 100 usuarios pasen de la sala de espera al sitio de destino.
7. Vuelve a la sala de espera y selecciona ¡Préstalo ahora! Ahora serás redirigido al sitio de destino.
8. Selecciona Comprar ahora para finalizar la transacción en el sitio de destino.

Implementación de pilas separadas

La pila principal es la única pila necesaria para utilizar la funcionalidad principal de la sala de espera. Todas las demás pilas son opcionales. Abre la pila de autorizadores si aún no tienes una forma de validar los tokens emitidos por las salas de espera o de proteger las API que ya tengas. Inicie la pila OpenID si necesita un proveedor de identidad OpenID para la integración de la sala de espera con las interfaces de autorización. El ejemplo de una pila de estrategias de entrada proporciona un par de ejemplos sobre cómo y cuándo permitir la entrada de más usuarios al sitio que está intentando proteger.

1. Lanza la pila principal

Tiempo de implementación: aproximadamente 20 minutos

Esta AWS CloudFormation plantilla automatizada despliega una sala de espera virtual AWS en la AWS nube. Debe completar los requisitos previos antes de lanzar la pila.

> **Note**
>
> Usted es responsable del coste de los AWS servicios utilizados al ejecutar esta solución. Para obtener más información, visite la sección de costos de esta guía y consulte la página web de precios de cada AWS servicio utilizado en esta solución.

1. Inicie sesión en AWS Management Consoley seleccione el botón para lanzar la `aws-virtual-waiting-room-on-aws.template` AWS CloudFormation plantilla.

   Launch solution

   Como alternativa, puede descargar la plantilla como punto de partida para su propia implementación.
2. La plantilla se lanza en la región Este de EE. UU. (Norte de Virginia) de forma predeterminada. Para lanzar la solución en una AWS región diferente, utilice el selector de regiones de la barra de navegación de la consola.
3. En la página Crear pila, compruebe que la URL de la plantilla correcta esté en el cuadro de texto URL de Amazon S3 y seleccione Siguiente.

4. En la página Especificar los detalles de la pila, especifique un nombre para la pila. Para obtener información sobre las limitaciones de nombres de caracteres, consulte los límites de IAM y STS en la Guía del AWS Identity and Access Management usuario.
5. En Parámetros, revise los parámetros de esta plantilla de solución y modifíquelos según sea necesario. Esta solución utiliza los siguientes valores predeterminados.

Parámetro	Predeterminado	Descripción
ID del evento	`Sample`	ID único para esta instancia de la sala de espera, se sugiere el formato GUID.
Periodo de validez	`3600`	Periodo de validez del token en segundos.
Habilite la generación de eventos	`false`	Si se establece en`true`, las métricas relacionadas con la sala de espera se escriben en su bus de eventos cada minuto.
Puerto Redis	`1785`	El número de puerto que se utilizará para conectarse al ElastiCache servidor Redis. Se recomienda no utilizar el puerto predeterminado ElastiCache para Redis de. `6379`
EnableQueuePositionExpiry	`true`	Si se establece en`false`, no se aplica el período de caducidad de las posiciones de cola.
QueuePositionExpiryPeriod	`900`	Es el intervalo de tiempo en segundos después del cual

Parámetro	Predeterminado	Descripción
		una posición de cola no es apta para generar un token.
IncrSvcOnQueuePosi tionExpiry	`false`	Si se establece en`true`, el contador de servicio avanza automáticamente en función de las posiciones de cola caducadas que no generaron fichas correctamente.

6. Elija Siguiente.
7. En la página Configurar opciones de pila, elija Siguiente.
8. En la página Revisar, revise y confirme la configuración. Marque la casilla para confirmar que la plantilla crea recursos AWS Identity and Access Management (IAM).
9. Elija Create stack (Crear pila) para implementar la pila.

   Puedes ver el estado de la pila en la AWS CloudFormation consola, en la columna Estado. Debería recibir el estado CREATE_COMPLETE en aproximadamente 20 minutos.

2. (Opcional) Inicie la pila de autorizadores

Tiempo de implementación: aproximadamente cinco minutos

1. Inicie sesión AWS Management Consoley seleccione el botón para lanzar la `aws-virtual-waiting-room-on-aws-authorizers.template` AWS CloudFormation plantilla.

   Como alternativa, puede descargar la plantilla como punto de partida para su propia implementación.
2. La plantilla se lanza en la región Este de EE. UU. (Norte de Virginia) de forma predeterminada. Para lanzar la solución en una AWS región diferente, utilice el selector de regiones de la barra de navegación de la consola.
3. En la página Crear pila, compruebe que la URL de la plantilla correcta esté en el cuadro de texto URL de Amazon S3 y seleccione Siguiente.

4. En la página Especificar los detalles de la pila, especifique un nombre para la pila. Para obtener información sobre las limitaciones de nombres de caracteres, consulte los límites de IAM y STS en la Guía del AWS Identity and Access Management usuario.
5. En Parámetros, revise los parámetros de esta plantilla de solución y modifíquelos según sea necesario. Esta solución utiliza los siguientes valores predeterminados.

Parámetro	Predeterminado	Descripción
Punto final de la API pública	*<Requiere que se introduzcan datos>*	Punto final público para las API de la sala de espera virtual.
ID del evento de la sala de espera	`Sample`	ID del evento de la sala de espera.
URI del emisor	*<Requiere que se introduzcan datos>*	URI del emisor de las claves y los tokens públicos.

6. Elija Siguiente.
7. En la página Configurar opciones de pila, elija Siguiente.
8. En la página Revisar, revise y confirme la configuración. Marque la casilla para confirmar que la plantilla crea recursos AWS Identity and Access Management (IAM).
9. Elija Create stack (Crear pila) para implementar la pila.

   Puedes ver el estado de la pila en la AWS CloudFormation consola, en la columna Estado. Deberías recibir el estado CREATE_COMPLETE en aproximadamente cinco minutos.

3. (Opcional) Inicie la pila OpenID

Tiempo de implementación: aproximadamente cinco minutos

1. Inicie sesión en AWS Management Consoley seleccione el botón para lanzar la `aws-virtual-waiting-room-on-aws-openid.template` AWS CloudFormation plantilla.

   Como alternativa, puede descargar la plantilla como punto de partida para su propia implementación.

2. La plantilla se lanza en la región Este de EE. UU. (Norte de Virginia) de forma predeterminada. Para lanzar la solución en una AWS región diferente, utilice el selector de regiones de la barra de navegación de la consola.
3. En la página Crear pila, compruebe que la URL de la plantilla correcta esté en el cuadro de texto URL de Amazon S3 y seleccione Siguiente.
4. En la página Especificar los detalles de la pila, especifique un nombre para la pila. Para obtener información sobre las limitaciones de nombres de caracteres, consulte los límites de IAM y STS en la Guía del AWS Identity and Access Management usuario.
5. En Parámetros, revise los parámetros de esta plantilla de solución y modifíquelos según sea necesario. Esta solución utiliza los siguientes valores predeterminados.

Parámetro	Predeterminado	Descripción
Punto final de la API pública	*<Requiere que se introduzcan datos>*	URL del punto final público para las API de la sala de espera virtual.
Punto final de la API privada	*<Requiere que se introduzcan datos>*	URL del punto final privado para las API de la sala de espera virtual.
Región de la API	*<Requiere que se introduzcan datos>*	AWS nombre de la región para las API de las salas de espera públicas y privadas.
ID del evento	`Sample`	ID del evento de la sala de espera.

6. Elija Siguiente.
7. En la página Configurar opciones de pila, elija Siguiente.
8. En la página Revisar, revise y confirme la configuración. Marque la casilla para confirmar que la plantilla crea recursos AWS Identity and Access Management (IAM).
9. Elija Create stack (Crear pila) para implementar la pila.

   Puedes ver el estado de la pila en la AWS CloudFormation consola, en la columna Estado. Deberías recibir el estado CREATE_COMPLETE en aproximadamente cinco minutos.

4. (Opcional) Inicie la pila de estrategias de entrada de muestras

Tiempo de despliegue: aproximadamente dos minutos

1. Inicie sesión en AWS Management Consoley seleccione el botón para lanzar la `aws-virtual-waiting-room-sample-inlet-strategy.template` AWS CloudFormation plantilla.

   Como alternativa, puede descargar la plantilla como punto de partida para su propia implementación.
2. La plantilla se lanza en la región Este de EE. UU. (Norte de Virginia) de forma predeterminada. Para lanzar la solución en una AWS región diferente, utilice el selector de regiones de la barra de navegación de la consola.
3. En la página Crear pila, compruebe que la URL de la plantilla correcta esté en el cuadro de texto URL de Amazon S3 y seleccione Siguiente.
4. En la página Especificar los detalles de la pila, especifique un nombre para la pila. Para obtener información sobre las limitaciones de nombres de caracteres, consulte los límites de IAM y STS en la Guía del AWS Identity and Access Management usuario.
5. En Parámetros, revise los parámetros de esta plantilla de solución y modifíquelos según sea necesario. Esta solución utiliza los siguientes valores predeterminados.

Parámetro	Predeterminado	Descripción
ID del evento	`Sample`	ID del evento de la sala de espera.
Punto final de la API principal privada	*<Requiere que se introduzcan datos>*	URL del punto final privado para las API de la sala de espera virtual.
Región de la API principal	*<Requiere que se introduzcan datos>*	AWS Región en la que está instalada la API principal.
Estrategia de entrada	`Periodic`	Se va a implementar la estrategia de entrada. `Periodic`aumenta el número de raciones cada

Parámetro	Predeterminado	Descripción
		minuto. `MaxSize`increment a el número de servicio en función del número máximo de transacciones que el sitio de destino intermedio puede gestionar en un momento dado.
Incrementar en	*<Requiere que se introduzcan datos>*	Cuánto debe incrementarse el contador de porciones por minuto. Necesario si se selecciona una estrategia de entrada periódica.
Hora de inicio	*<Requiere que se introduzcan datos>*	Marca la hora en la que se debe empezar a increment ar el número de raciones (tiempo de la época en segundos). Obligatorio si se selecciona una estrategia de entrada periódica.
End Time (Hora de finalizac ión)	*<Requiere que se introduzcan datos>*	Marca el tiempo que indica cuándo dejar de aumentar el número de raciones (tiempo de la época en segundos) . Si se deja 0, el número de raciones se incrementa indefinidamente. Obligatorio si se selecciona una estrategi a de ingesta periódica.

Parámetro	Predeterminado	Descripción
CloudWatch Nombre de la alarma	*<Requiere que se introduzcan datos>*	Nombre CloudWatch de alarma opcional que se asociará a la estrategia de entrada periódica. Si se proporciona y está en estado alarmante, el número de raciones no se increment a. Aplicable únicamente a la estrategia de entrada periódica.
Tamaño máximo	*<Requiere que se introduzcan datos>*	El número máximo de transacciones que el sitio de destino descendente puede procesar a la vez (MaxSize estrategia).

6. Elija Siguiente.
7. En la página Configurar opciones de pila, elija Siguiente.
8. En la página Revisar, revise y confirme la configuración. Marque la casilla para confirmar que la plantilla crea recursos AWS Identity and Access Management (IAM).
9. Elija Create stack (Crear pila) para implementar la pila.

   Puedes ver el estado de la pila en la AWS CloudFormation consola, en la columna Estado. Debería recibir el estado CREATE_COMPLETE en aproximadamente dos minutos.

5. (Opcional) Inicie la pila de ejemplos de salas de espera

Tiempo de implementación: aproximadamente cinco minutos

1. Inicie sesión en AWS Management Consoley seleccione el botón para lanzar la `aws-virtual-waiting-room-sample.template` AWS CloudFormation plantilla.

Launch solution Como alternativa, puede descargar la plantilla como punto de partida para su propia implementación.

2. La plantilla se lanza en la región Este de EE. UU. (Norte de Virginia) de forma predeterminada. Para lanzar la solución en una AWS región diferente, utilice el selector de regiones de la barra de navegación de la consola.
3. En la página Crear pila, compruebe que la URL de la plantilla correcta esté en el cuadro de texto URL de Amazon S3 y seleccione Siguiente.
4. En la página Especificar los detalles de la pila, especifique un nombre para la pila. Para obtener información sobre las limitaciones de nombres de caracteres, consulte los límites de IAM y STS en la Guía del AWS Identity and Access Management usuario.
5. En Parámetros, revise los parámetros de esta plantilla de solución y modifíquelos según sea necesario. Esta solución utiliza los siguientes valores predeterminados.

Parámetro	Predeterminado	Descripción
Región API Gateway	*<Requiere que se introduzcan datos>*	AWS Nombre de la región de la API Gateway.
ARN del autorizador	*<Requiere que se introduzcan datos>*	ARN del autorizador Lambda de API Gateway.
ID del evento	`Sample`	ID del evento de la sala de espera.
Punto final de API privado	*<Requiere que se introduzcan datos>*	URL del punto final privado para las API de la sala de espera virtual.
Punto final de la API pública	*<Requiere que se introduzcan datos>*	URL del punto final público para las API de la sala de espera virtual.

6. Elija Siguiente.
7. En la página Configurar opciones de pila, elija Siguiente.

8. En la página Revisar, revise y confirme la configuración. Marque la casilla para confirmar que la plantilla crea recursos AWS Identity and Access Management (IAM).
9. Elija Create stack (Crear pila) para implementar la pila.

   Puedes ver el estado de la pila en la AWS CloudFormation consola, en la columna Estado. Deberías recibir el estado CREATE_COMPLETE en aproximadamente cinco minutos.

Actualización de la pila desde una versión anterior

Recomendamos eliminar la pila y crear una nueva para la nueva versión. Actualmente, no se admite la migración a la versión más reciente mediante la actualización de la CloudFormation pila. ConsulteDesinstalar la solución, a continuación, Lanzar la pila de introducción.

> **Note**
>
> Recomendamos migrar a una versión más reciente cuando no esté utilizando la solución de forma activa para soportar un evento en curso.

Datos de rendimiento

Se AWS ha realizado una prueba de carga de Virtual Waiting Room on con una herramienta llamada Locust. Los tamaños de los eventos simulados oscilaron entre 10 000 y 100 000 clientes. El entorno de pruebas de carga constaba de la siguiente configuración:

- Locust 2.x con personalizaciones para despliegues en la nube AWS
- Cuatro AWS regiones (,,,) `us-west-1` `us-west-2` `us-east-1` `us-east-2`
- 10 hosts `c5.4xlarge` Amazon EC2 por región (40 en total)
- 32 procesos de Locust por huésped
- Los usuarios simulados se distribuyeron uniformemente entre los 1280 procesos

Los pasos de prueba de la end-to-end API para cada proceso de usuario:

1. Llama `assign_queue_num` y recibe un identificador de solicitud.
2. Sigue `queue_num` el identificador de la solicitud hasta que devuelva la posición del usuario en la cola (poco tiempo).
3. Realice un bucle `serving_num` hasta que el valor devuelto sea >= posición en la cola del usuario (tiempo prolongado).
4. Llame con poca frecuencia `waiting_room_size` para recuperar el número de usuarios en espera.
5. Llame `generate_token` y reciba un JWT para usarlo en el sitio de destino.

Resultados

No existe un límite máximo práctico para el número de clientes que pueden tramitarse en la sala de espera.

La velocidad a la que los usuarios entran en la sala de espera afecta a las cuotas de ejecución simultánea de la función Lambda en la región en la que está desplegada.

La prueba de carga no pudo superar el límite de solicitudes predeterminado de API Gateway de 10 000 solicitudes por segundo con las políticas de almacenamiento en caché utilizadas con CloudFront.

La función `get_queue_num` Lambda tiene una tasa de invocación cercana a la proporción de usuarios entrantes a la sala de espera. Esta función Lambda puede estar limitada cuando hay

altas tasas de usuarios entrantes debido a los límites de simultaneidad o de ráfaga. La limitación provocada por un gran número de invocaciones a funciones de `get_queue_num` Lambda puede afectar a otras funciones de Lambda como efecto secundario. El sistema en general sigue funcionando si el software cliente puede responder adecuadamente a este tipo de error de escalado temporal con una lógica de retracción o retroceso.

La CloudFront distribución configurada por el núcleo con una configuración de cuota predeterminada puede gestionar una sala de espera con capacidad para 250 000 usuarios, y cada usuario consulta la `serving_num` API al menos cada segundo.

Solución de problemas

En esta sección se proporciona información sobre la solución de problemas de esta solución.

Si en esta sección no se aborda el problema, Contact AWS Support proporciona instrucciones para abrir un caso de AWS Support para esta solución.

4 veces el estado de respuesta de las API

- Esto puede deberse a un ID de evento o de solicitud incorrectos, o a ambos. Esto ocurre en los CloudWatch registros de la función Lambda relacionada.
- Las API privadas están autenticadas por IAM y el cliente necesita AWS claves con derechos para invocarlas. Esto ocurre en los CloudWatch registros de API Gateway.

5xx estados de respuesta de las API

- Respuesta de Lambda o API Gateway regulados, compruebe la alarma. *<LambdaFunctionName>*`ThrottlesAlarm` CloudWatch
- Configuración incorrecta en el back-end. Compruebe *<LambdaFunctionName>*`ErrorsAlarm` CloudWatch la alarma y los registros para obtener más información. CloudWatch

5XX/ErrorPublicPrivateApiAlarm

- Este estado de alarma se produce `ALARM` cuando la API devuelve un estado 5XX a la persona que llama en un período de 60 segundos.
- Esta alarma vuelve a activarse `OK` cuando no se devuelve el estado 5xx durante 60 segundos.
- Esta alarma se puede iniciar mediante una función de Lambda o un motor de ejecución de Lambda que devuelva un error a API Gateway.

4XX/ErrorPublicPrivateApiAlarm

- Este estado de alarma se produce `ALARM` cuando la API devuelve un estado 4XX a la persona que llama en un período de 60 segundos.
- Esta alarma vuelve `OK` al estado 4XX durante 60 segundos.
- Esta alarma se puede iniciar con una URL de API incorrecta.

*<LambdaFunctionName>*ThrottlesAlarm

- Este estado de alarma es ALARMA cuando la Lambda indicada encuentra un límite de ejecución simultánea en un período de 60 segundos.
- Esta alarma vuelve a activarse OK si no se detecta ningún acelerador durante 60 segundos.
- Es posible que tengas que aumentar el límite de simultaneidad de la región de tu cuenta.
- Es posible que se esté encontrando con el límite de ráfagas de Lambda, lo que requiere cierta lógica de reintento en el cliente.

*<LambdaFunctionName>*ErrorsAlarm

- Este estado de alarma se produce ALARM cuando la Lambda nombrada encuentra un error de ejecución en un período de 60 segundos.
- Esta alarma vuelve a activarse OK si no se detecta ningún error durante 60 segundos.
- Esto puede deberse a un error de configuración en el backend.
- Esto puede deberse a un error en el código de Lambda.

Contacto AWS Support

Si cuenta con AWS Developer Support, AWS Business Support o AWS Enterprise Support, puede utilizar el Centro de soporte para obtener asistencia de expertos con esta solución. En las siguientes secciones, encontrará instrucciones.

Cree un caso

1. Inicie sesión en Support Center.
2. Elija Crear caso.

¿Cómo podemos ayudar?

1. Elija Técnico.
2. Para el servicio, seleccione Soluciones.
3. Para la categoría, seleccione Otras soluciones.
4. En Gravedad, seleccione la opción que mejor se adapte a su caso de uso.

5. Al introducir el servicio, la categoría y la gravedad, la interfaz rellena los enlaces a las preguntas de solución de problemas más frecuentes. Si no puede resolver su pregunta con estos enlaces, seleccione Siguiente paso: información adicional.

Información adicional

1. En Asunto, introduce un texto que resuma tu pregunta o problema.
2. En Descripción, describe el problema en detalle.
3. Selecciona Adjuntar archivos.
4. Adjunta la información AWS Support necesaria para procesar la solicitud.

Ayúdenos a resolver su caso más rápido

1. Introduzca la información solicitada.
2. Elija Siguiente paso: Resuelva ahora o póngase en contacto con nosotros.

Resuelva ahora o póngase en contacto con nosotros

1. Revise las soluciones Solve now.
2. Si no puede resolver su problema con estas soluciones, elija Contactar con nosotros, introduzca la información solicitada y pulse Enviar.

Recursos adicionales de

AWS servicios	
• AWS CloudFormation	• Amazon DynamoDB
• Amazon Simple Storage Service	• Amazon API Gateway
• AWS Lambda	• AWS Secrets Manager
• Amazon CloudFront	• Amazon Simple Queue Service
• Amazon EventBridge	• Amazon CloudWatch
• Amazon ElastiCache para Redis	• Amazon Comprehend
• Amazon Virtual Private Cloud	• AWS Identity and Access Management

Desinstalar la solución

Puede desinstalar la AWS solución Virtual Waiting Room on de la AWS Management Console o utilizando la AWS Command Line Interface. Debe eliminar manualmente los depósitos de S3 que se utilizan para almacenar los registros de los distintos recursos creados por esta solución. AWS Las implementaciones de soluciones no eliminan automáticamente estos depósitos de S3, por lo que puede seguir revisando los registros una vez que se haya eliminado la solución.

Si ha agregado manualmente un usuario de IAM al grupo de usuarios de `ProtectedAPIGroup` IAM creado por la solución, elimine al usuario de IAM del grupo de usuarios de IAM antes de desinstalar la solución. De lo contrario, el grupo de usuarios de IAM y la política de IAM asociada no se eliminarán.

Para cada una de las pilas implementadas, siga las instrucciones que se indican a continuación.

Usando el AWS Management Console

1. Inicie sesión en la consola de AWS CloudFormation.
2. En la página Pilas, seleccione la pila de instalación de esta solución.
3. Elija Eliminar.

Usando AWS Command Line Interface

Determine si el AWS Command Line Interface (AWS CLI) está disponible en su entorno. Para obtener instrucciones de instalación, consulte ¿Qué es AWS Command Line Interface? en la Guía AWS CLI del usuario. Tras confirmar que AWS CLI está disponible, ejecute el siguiente comando.

```
$ aws cloudformation delete-stack --stack-name <installation-stack-name>
```

Eliminar los buckets de Amazon S3

Esta solución está configurada para conservar el bucket de Amazon S3 creado por la solución (para implementarlo en una región opcional) si decide eliminar la AWS CloudFormation pila para evitar la pérdida accidental de datos. Tras desinstalar la solución, puede eliminar manualmente este depósito de S3 si no necesita conservar los datos. Siga estos pasos para eliminar el bucket de Amazon S3.

1. Inicie sesión en la consola de Amazon S3.
2. En el panel de navegación izquierdo, elija Buckets.
3. Localice los buckets *<stack-name>* de S3.
4. Selecciona el bucket de S3 y selecciona Eliminar.

Para eliminar el depósito de S3 mediante AWS CLI, ejecute el siguiente comando:

```
$ aws s3 rb s3://<bucket-name> --force
```

Código fuente

Visite nuestro GitHubrepositorio para descargar los archivos fuente de esta solución y compartir sus personalizaciones con otras personas.

Colaboradores

- Jim Thario
- Thyag Ramachandran
- Joan Morgan
- Justin Pirtle
- Allen Moheimani
- Garvit Singh
- Bassem Wanis

Revisiones

Date	Cambio
Noviembre de 2021	Versión inicial
Septiembre de 2022	Versión 1.1: incremento automático del contador de servicio en función de las posiciones de cola caducadas. Traslade parte del uso de Redis a DynamoDB. Punto final de la API pública para obtener el tiempo de caducidad de las posiciones de cola restantes . Para obtener más información, consulta el archivo ChangeLog.md del repositorio. GitHub
Abril de 2023	Versión 1.1.1: Se ha mitigado el impacto provocado por la nueva configuración predeterminada de propiedad de objetos de S3 (las ACL están deshabilitadas) en todos los buckets de S3 nuevos. Para obtener más información, consulte el archivo ChangeLog.md del repositorio. GitHub
Noviembre de 2023	Versión 1.1.2: versiones de paquetes actualizadas para resolver las vulnerabilidades de seguridad. Para obtener más información, consulte el archivo ChangeLog.md del repositorio. GitHub
Marzo de 2024	Versión 1.1.3: se solucionaron tres problemas : las posiciones de cola caducadas que persistían en el tamaño de la sala de espera, la `queue_num` API devolvía los resultados antiguos incluso después de un restablecimiento y los fallos intermitentes en la API del adaptador OpenID. `/userInfo` Para

Date	Cambio
	obtener más información, consulta el archivo ChangeLog.md del repositorio. GitHub
Abril de 2024	Versión 1.1.4: versiones de paquetes actualizadas para resolver las vulnerabilidades de seguridad. Para obtener más información, consulte el archivo ChangeLog.md del repositorio. GitHub
Junio de 2024	Versión 1.1.5: versiones de paquetes actualizadas para resolver las vulnerabilidades de seguridad. Para obtener más información, consulte el archivo ChangeLog.md del repositorio. GitHub

Avisos

Es responsabilidad de los clientes realizar su propia evaluación independiente de la información que contiene este documento. Este documento: (a) tiene únicamente fines informativos, (b) representa las ofertas y prácticas de productos AWS actuales, que están sujetas a cambios sin previo aviso, y (c) no implica ningún compromiso ni garantía por parte de AWS sus filiales, proveedores o licenciantes. AWS los productos o servicios se proporcionan «tal cual» sin garantías, representaciones o condiciones de ningún tipo, ya sean expresas o implícitas. AWS Las responsabilidades y obligaciones con sus clientes están reguladas por AWS acuerdos, y este documento no forma parte de ningún acuerdo entre sus clientes AWS y sus clientes ni lo modifica.

AWS Glosario

Para obtener la AWS terminología más reciente, consulte el AWS glosario de la Glosario de AWS Referencia.

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.