Hacemos un proyecto Big Data

Estudiantes María Correas Crespo, Judith Urbina Córdoba
Profesor Marta Martín Moreno
Asignatura Análisis de Datos en entornos Big Data

1. Actividad: Definición Objetivos y KPIs

Descripción

En esta actividad, se propone que desarrolléis un pequeño proyecto de Big data utilizando las herramientas de esta plataforma. El proyecto se dividirá en dos entregas que cubrirán la definición de objetivos y KPIs y la implementación de la propuesta mediante herramientas de AWS. Finalmente, presentaréis vuestra propuesta de proyecto a vuestros compañeros del aula. La temática es libre; podéis elegir una área de interés o aplicar el aprendizaje a un problema real próximo a vosotros.

En esta primera fase del proyecto, tenéis que definir el alcance del proyecto, establecer objetivos claros y medibles, e identificar los KPI (Key Performance Indicators) que usaréis para evaluar el éxito del proyecto.

Objetivos

Se pide:

1. Selección de la temática o caso de estudio
   - ☐ área de interés o un caso de estudio que queráis abordar con el proyecto de Big data
   - ☐ temática tiene que ser relevante y aplicable a un contexto real.
2. Descripción del problema
   - ☐ Aseguraos de describir por qué este problema es importante
   - ☐ qué impacto puede tener su resolución
3. Establecimiento de objetivos. Los objetivos tienen que ser realistas y alineados con la resolución del problema identificado.
   - ☐ Definís objetivos específicos y medibles que queréis lograr con el proyecto.
   - ☐ Explicad cómo el análisis de datos contribuirá a conseguir estos objetivos.
4. Definición de KPI.
   - ☐ Definís tres KPI (Key Performance Indicators) que sean relevantes para mesurar el éxito del proyecto
   - ☐ Explicad por qué estos KPI son importantes
   - ☐ cómo se relacionan con los objetivos establecidos, y de qué manera se mesurarán.
5. Identificación de fuentes de datos.
   - ☐ Identifica y describe las fuentes de datos que has utilizado para abordar el problema.
   - ☐ Indicáis la naturaleza de los datos,
   - ☐ su procedencia
   - ☐ cómo se prevé que sean utilizadas en el análisis.
6. Asignación de tareas dentro del equipo.
   - ☐ Repartís las responsabilidades y tareas entre los miembros del equipo.
   - ☐ Aseguraos que cada miembro tiene un papel claro y definido,
   - ☐ tareas asignadas son coherentes con sus habilidades y roles dentro del proyecto.

Propuestas de datasets

Datos de buena calidad para minimizar la corrección. Datos que tengan información relevante, precisa y actualizada. Algunos enlaces que podrían ser útiles

1. Ofrecido por Amazon y con ejemplos Open data on AWS
2. estados unidos catálogo de datagov
3. europa data europa
4. Plataformas de código abierto creando cuenta: pewresearch
5. Contiene un listado mayor de datasets Open Access Directory
6. Otro listado mayor de datasets Universidad de Missouri
7. Si se sabe buscar datasets amplios Dat search
8. Muy bien documentada la manera de citar y otros detalles nature
9. Contiene lista de centros de datasets y tips para elegir sqream

No se puede abrir al entrar dentro del dataset o hay pegas

1. Migraciones de animales! movebank
2. varios re3data

2. Actividad: Implementación AWS

Descripción

En el proyecto Big Data tenemos que tener

1. Pruebas de proceso de datos.
   - ☐ Realizáis un conjunto de pruebas para aseguraros que los datos se están procesando correctamente.
   - ☐ Describís los pasos seguidos para verificar la precisión y la consistencia del proceso de datos.
2. Validación de los KPIs.
   - ☐ Validáis que los KPIs definidos están correctamente medidos y que reflejan fielmente la realidad del análisis.
   - ☐ Explicáis los métodos utilizados para asegurar la precisión y la relevancia de los KPIs.
3. Incorporación de scripts o código relevante.
   - ☐ Incluis los scripts o código más relevantes que han sido utilizados durante la implementación del proyecto.
   - ☐ Aseguraos que están suficientemente descritos para facilitar la comprensión de las operaciones realizadas.

Objetivos

Por servicio de AWS usado

1. Por qu:e hemos usado este servicio? C:omo se relaciona con nuestros objetivos de proyecto?
2. C:omo lo describimos?
3. C:omo lo hemos usado? Con qu:e tipo de datos hemos trabajado en nuestro proyecto?

En detalle describimos

1. con un diagrama de Amazon diagramas
2. c:omo realizaremos el procesamientode los datos a AWS?
   - ☐ explicación de los flujos de trabajo
   - ☐ explicación de las transformaciones de datos
   - ☐ gestión de la integración entre los diferentes servicios de AWS

Usamos

Recomendaciones

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. Maximizar el uso de los servicios PaaS:

1. servicios S3, Lambda, Kinesis, EMR, EC2, Glue, Athena o Redshift

y minimizar el uso de servicios IaaS

3. Jupyter Notebook: Actividad Cloud con AWS

3.1. Ejercicio 1: Descripción y justificación de los servicios de AWS (2,5 pts)

El proyecto tiene como objetivo manejar datos inmobiliarios de Londres almacenados en un archivo CSV, procesarlos e integrarlos en una base de datos para consultas posteriores a través de funciones Lambda. A continuación, describimos los servicios seleccionados y su justificación:

3.1.1. Amazon S3 (Simple Storage Service)

Justificación: Amazon S3 es ideal para almacenar grandes cantidades de datos estructurados o no estructurados de manera económica y escalable. En este caso, se utiliza para almacenar el archivo CSV inicial con los datos de las casas.
Rol en el flujo: Almacena el archivo datos.csv con los datos de propiedades inmobiliarias. Actúa como origen de datos para AWS Glue.

3.1.2. Amazon RDS (Relational Database Service)

Justificación: Amazon RDS ofrece una base de datos gestionada que es confiable, escalable y fácil de usar. Es perfecta para alojar los datos transformados y permitir consultas rápidas y eficientes. En este caso, se elige MySQL como motor de base de datos debido al amplio soporte de la comunidad, la buena integración con AWS Lambda y herramientas de análisis de datos y la capacidad de manejar datos estructurados como los requeridos en este proyecto.
Rol en el flujo: Almacena los datos inmobiliarios procesados de forma relacional. Sirve como origen para las consultas realizadas por las funciones Lambda.

3.1.3. AWS Lambda

Justificación: AWS Lambda permite ejecutar código en respuesta a eventos y es ideal para construir aplicaciones backend sin servidor. Es eficiente en costos y escalable.
Rol en el flujo: Una Lambda consulta la base de datos para buscar casas según filtros específicos (por ejemplo, precio, número de habitaciones, ubicación). Otra Lambda permite gestionar casas favoritas por usuario y realiza actualizaciones en la base de datos.

3.1.4. Amazon API Gateway

Justificación: API Gateway actúa como puerta de enlace para exponer las funciones Lambda como endpoints HTTP, permitiendo que los usuarios interactúen con el backend.
Rol en el flujo: Proporciona endpoints RESTful para las funcionalidades: búsqueda de casas, gestión de usuarios, casas favoritas, etc. Maneja la autenticación y autorización para garantizar la seguridad.

3.1.5. Amazon Secrets Manager

Justificación: AWS Secrets Manager es un servicio diseñado para almacenar y gestionar de forma segura secretos como credenciales de bases de datos, claves API y otros datos confidenciales. En este proyecto, se utiliza para almacenar y gestionar las credenciales de acceso a la base de datos MySQL en Amazon RDS. Esto evita la necesidad de incluir credenciales en el código fuente de las funciones Lambda, mejorando la seguridad y facilitando la rotación de credenciales.

Rol en el flujo: Se crea un secreto en Secrets Manager que almacena las credenciales de la base de datos (nombre de usuario, contraseña, endpoint, y puerto). Las funciones Lambda solicitan el secreto durante su ejecución mediante la API de Secrets Manager (GetSecretValue). Las credenciales se utilizan para establecer la conexión con MySQL en RDS.

3.1.6. AWS Identity and Access Management (IAM)

Justificación: AWS IAM proporciona control de acceso granular a los recursos y servicios de AWS. En este proyecto, IAM se utiliza para asignar los permisos mínimos necesarios a cada recurso, asegurando la seguridad del flujo de trabajo.

Rol en el flujo:

Roles para Lambda: Se crea un rol IAM asociado a las funciones Lambda con permisos específicos para: Acceder a Secrets Manager (secretsmanager:GetSecretValue). Registrar logs en CloudWatch. Este enfoque de permisos mínimos garantiza que las Lambdas solo tengan acceso a los recursos que necesitan. Permisos de Glue: Glue tiene permisos para acceder al bucket S3 donde se encuentra el archivo CSV y para cargar los datos transformados en RDS. Políticas específicas: Cada servicio (S3, RDS, Secrets Manager) tiene políticas asignadas de manera individual para minimizar riesgos.

3.1.7. Amazon Virtual Private Cloud (VPC)

Justificación: Amazon VPC proporciona una red privada y aislada dentro de la nube de AWS, lo que garantiza un entorno seguro para servicios como RDS. En este proyecto, la VPC se utiliza para proteger la base de datos y limitar el acceso desde Internet.

Rol en el flujo:

Base de datos RDS: La base de datos MySQL se encuentra dentro de una subred privada de la VPC. Esto asegura que solo los recursos autorizados dentro de la VPC (como Lambda) puedan acceder a ella. Funciones Lambda: Las funciones Lambda están configuradas para ejecutar dentro de la misma VPC, con acceso a subredes privadas que les permitan comunicarse con RDS. Configuración de seguridad: Los grupos de seguridad de la VPC permiten únicamente el tráfico necesario (por ejemplo, conexiones entrantes desde Lambda a RDS en el puerto 3306). Opcional: Endpoints de VPC: Para mejorar la seguridad y reducir la latencia, se podrían configurar endpoints de VPC para que servicios como S3 y Secrets Manager sean accesibles directamente desde la VPC, sin pasar por Internet.

3.1.8. Flujo de trabajo

A continuación, se describe el flujo de trabajo, incluyendo las transformaciones de datos y la integración entre servicios:

Ingesta de datos desde Amazon S3

El archivo `houses.csv` es subido al bucket de Amazon S3. AWS Glue utiliza un *crawler* para explorar el esquema del archivo CSV y generar una tabla en el Data Catalog. Un *job* ETL en AWS Glue procesa los datos realizando las siguientes transformaciones:

- Eliminación de duplicados y filas incompletas.
- Conversión de formatos de campos (por ejemplo, fechas y precios).
- Agregación de datos calculados, como el precio por metro cuadrado.

Almacenamiento de datos en Amazon RDS (MySQL)

Los datos transformados se cargan en una tabla de la base de datos MySQL gestionada en Amazon RDS. Se configuran relaciones entre tablas, como las que vinculan propiedades y usuarios.

Gestión segura de credenciales con AWS Secrets Manager

Las credenciales de la base de datos MySQL (nombre de usuario, contraseña, endpoint y puerto) se almacenan de forma segura en AWS Secrets Manager. Las funciones Lambda acceden a estos secretos en tiempo de ejecución mediante la API de Secrets Manager, eliminando la necesidad de incluir credenciales en el código fuente.

Exposición de funcionalidades mediante AWS Lambda y API Gateway

Las funciones AWS Lambda interactúan con la base de datos MySQL en RDS:

- Una función Lambda ejecuta consultas `SELECT` para devolver resultados basados en filtros proporcionados por los usuarios, como precio, ubicación o número de habitaciones.
- Otra función Lambda realiza operaciones de escritura, como gestionar las casas favoritas de cada usuario (por ejemplo, `INSERT` y `UPDATE`).

AWS API Gateway expone estas funciones Lambda como *endpoints* HTTP para su uso desde el frontend de la aplicación.

Gestión de accesos con AWS IAM
Se definen roles y políticas en AWS Identity and Access Management (IAM) para garantizar que cada servicio tenga acceso únicamente a los recursos necesarios:

- Las funciones Lambda tienen un rol con permisos para acceder a Secrets Manager, CloudWatch Logs y RDS.
- AWS Glue tiene un rol con permisos para leer el bucket S3 y escribir datos en RDS.

Seguridad y conectividad mediante Amazon VPC
Amazon Virtual Private Cloud (VPC) asegura la comunicación entre los servicios de forma privada:

- La base de datos MySQL en Amazon RDS está alojada en una subred privada dentro de la VPC.
- Las funciones Lambda están configuradas para ejecutarse dentro de la misma VPC, permitiendo acceso seguro a la base de datos.
- Los *security groups* controlan el tráfico permitido entre Lambda y RDS, asegurando que solo las conexiones necesarias sean aceptadas.

Opcionalmente, se pueden configurar *endpoints* de VPC para que servicios como S3 o Secrets Manager sean accesibles sin salir de la red privada.

3.1.9. Políticas y roles de IAM

AWS Glue
Se muerstra la política de IAM en formato JSON que se ha asociado para permitir a AWS Glue acceder al bucket S3, la base de datos RDS, ejecutar Glue Jobs y escribir logs en CloudWatch:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3:::<nombre-del-bucket-s3>",
        "arn:aws:s3:::<nombre-del-bucket-s3>/*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "rds:DescribeDBInstances",
        "rds:Connect"
      ],
      "Resource": "arn:aws:rds:<region>:<account-id>:db:<nombre-de-la-base-de-datos>"
    },
    {
      "Effect": "Allow",
      "Action": [
        "logs:CreateLogGroup",
        "logs:CreateLogStream",
        "logs:PutLogEvents"
      ],
      "Resource": "arn:aws:logs:<region>:<account-id>:log-group:/aws/glue/*"
    },
    {
      "Effect": "Allow",
```

```
      "Action": [
        "glue:CreateJob",
        "glue:DeleteJob",
        "glue:UpdateJob",
        "glue:StartJobRun",
        "glue:GetJob",
        "glue:GetJobRun",
        "glue:GetTable",
        "glue:GetTables",
        "glue:BatchCreatePartition",
        "glue:BatchDeletePartition",
        "glue:GetPartition",
        "glue:GetPartitions",
        "glue:UpdateTable"
      ],
      "Resource": "*"
    }
  ]
}
```

Explicación de cada sección: Acceso a S3:

Se otorgan permisos de lectura (s3:GetObject) y de listado de objetos (s3:ListBucket) sobre el bucket S3 específico. Reemplaza ¡nombre-del-bucket-s3¿ con el nombre de tu bucket. Acceso a RDS:

Se otorgan permisos para describir las instancias de RDS (rds:DescribeDBInstances) y conectar a la base de datos (rds:Connect). Reemplaza ¡nombre-de-la-base-de-datos¿ con el nombre de tu base de datos y ajusta la región y el ID de cuenta de AWS. Permisos para CloudWatch Logs:

Se permiten acciones necesarias para crear log groups, crear log streams y escribir eventos de log en CloudWatch (logs:CreateLogGroup, logs:CreateLogStream, logs:PutLogEvents). Esto es útil para que los Glue Jobs generen logs. Permisos de Glue:

Se permiten varias acciones de Glue, como crear, actualizar, ejecutar y obtener detalles de los Glue Jobs, así como trabajar con las tablas y particiones en el Glue Data Catalog.

AWS Lambda

Se muestra la política de IAM en formato JSON para permitir que una función Lambda interactúe con RDS, Secrets Manager, CloudWatch Logs y API Gateway.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "rds:DescribeDBInstances",
        "rds:Connect",
        "rds:ExecuteStatement"
      ],
      "Resource": "arn:aws:rds:<region>:<account-id>:db:<nombre-de-la-base-de-datos>"
    },
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:GetSecretValue"
      ],
      "Resource": "arn:aws:secretsmanager:<region>:<account-id>:secret:<nombre-del-secret>/*"
    },
    {
      "Effect": "Allow",
      "Action": [
```

```
        "logs:CreateLogGroup",
        "logs:CreateLogStream",
        "logs:PutLogEvents"
      ],
      "Resource": "arn:aws:logs:<region>:<account-id>:log-group:/aws/lambda/*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "apigateway:GET",
        "apigateway:POST",
        "apigateway:PUT",
        "apigateway:DELETE"
      ],
      "Resource": "arn:aws:apigateway:<region>::/restapis/*"
    }
  ]
}
```

Explicación de la política: Acceso a RDS:

Permisos para interactuar con RDS: Se otorgan permisos de descripción (rds:DescribeDBInstances), conexión (rds:Connect) y ejecución de sentencias SQL (rds:ExecuteStatement) para interactuar con la base de datos. Asegúrate de reemplazar ¡nombre-de-la-base-de-datos¿ con el nombre de tu base de datos y ajustar la región y el ID de cuenta. Acceso a Secrets Manager:

Permisos para acceder a los secretos: Se otorgan permisos de lectura (secretsmanager:GetSecretValue) para recuperar el valor de los secretos almacenados en AWS Secrets Manager. Reemplaza ¡nombre-del-secret¿ con el nombre del secret que contiene las credenciales para acceder a tu base de datos. Acceso a CloudWatch Logs:

Permisos para CloudWatch Logs: Se permiten las acciones necesarias para crear un log group, crear un log stream y escribir eventos de log en CloudWatch. Estos permisos son necesarios para que la función Lambda registre las métricas de ejecución. Se usa la ruta /aws/lambda/* para indicar que se aplica a todas las funciones Lambda en la cuenta. Acceso a API Gateway:

Permisos para interactuar con API Gateway: Se permiten las acciones para hacer peticiones HTTP (GET, POST, PUT, DELETE) a cualquier API de API Gateway en la región. Esto es útil si la Lambda va a realizar alguna operación relacionada con una API Gateway (por ejemplo, invocar API Gateway o interactuar con recursos de una API).

3.2. Ejercicio 2: Pruebas de procesamiento de datos (2,5 puntos)

3.2.1. Pruebas de procesamiento de datos

El objetivo de estas pruebas es asegurar que los datos se procesan de forma correcta y consistente desde su ingesta hasta su almacenamiento final. A continuación, se describe el conjunto de pruebas diseñadas para validar cada etapa del flujo de trabajo.

Preparación de los datos de prueba

- **Generación de datos de prueba:** Se crea un archivo CSV sintético (`houses_test.csv`) con un subconjunto de datos representativos que incluye:
  - Datos correctos.
  - Datos duplicados.
  - Filas incompletas.
  - Valores atípicos (outliers).
- **Carga en S3:** El archivo se sube manualmente al bucket de Amazon S3 utilizado por el proyecto.

Validación del proceso ETL en AWS Glue

- **Ejecución del ETL:** Se ejecuta el *job* de Glue sobre el archivo de prueba en S3.
- **Pruebas de transformación:** Los datos transformados se exportan en formato Parquet o CSV a un bucket de S3 intermedio para su validación:
  - Confirmación de la eliminación de duplicados.
  - Validación de la eliminación de filas incompletas.
  - Verificación de las transformaciones aplicadas a columnas (formato de fechas, normalización de precios, etc.).
  - Cálculo correcto de campos derivados, como el precio por metro cuadrado.

Pruebas de almacenamiento en Amazon RDS (MySQL)

- **Carga de datos en RDS:** Los datos transformados son cargados en una base de datos MySQL en Amazon RDS.
- **Consultas de validación:** Desde una instancia EC2, se realizan las siguientes verificaciones:
  - Comprobación de la cantidad de filas cargadas.
  - Validación de las relaciones entre tablas (por ejemplo, propiedades asociadas a usuarios).
  - Verificación de la integridad de los datos (sin duplicados, sin filas incompletas).

Pruebas de consulta mediante AWS Lambda

- **Pruebas de consulta:** Se invocan directamente las funciones Lambda que consultan la base de datos:
  - Verificación de que los filtros de búsqueda (precio, ubicación, número de habitaciones) devuelven resultados precisos.
  - Comprobación de la funcionalidad de gestión de favoritos (inserción y eliminación de registros).
- **Exposición de APIs:** Se validan los *endpoints* HTTP expuestos por API Gateway para asegurar que devuelven respuestas coherentes.

Monitoreo en CloudWatch Logs:
Se revisan los logs generados por las funciones Lambda y AWS Glue para identificar posibles errores o advertencias.

Validación adicional con Amazon Athena
Se configura Amazon Athena para realizar consultas SQL directamente sobre los datos procesados almacenados en S3:

- Validación de la calidad de los datos intermedios antes de su carga en RDS.
- Comparación de los resultados en S3 con los datos finales en MySQL.

Resumen del flujo y herramientas utilizadas

Etapa del flujo	Servicio utilizado	Validación realizada
Ingesta de datos	Amazon S3	Archivo CSV cargado correctamente.
ETL	AWS Glue	Transformaciones y formato validados.
Almacenamiento	Amazon RDS (MySQL)	Integridad y relaciones de datos verificadas.
Consultas	AWS Lambda + API Gateway	Respuestas correctas a las consultas HTTP.
Consulta avanzada	Amazon Athena	Validación de datos intermedios en S3.
Monitoreo	CloudWatch Logs	Identificación y resolución de errores.

Cuadro 1: Flujo de validación y servicios utilizados.

3.2.2. Ejercicio 3: Validación de los KPIs

Para asegurar que los KPIs (Key Performance Indicators) definidos están correctamente medidos y reflejan fielmente la realidad del análisis, se deben seguir los pasos y métodos descritos a continuación para validar la precisión y la relevancia de los KPIs en la versión de prueba.

1. Definición Clara y Precisa de los KPIs
En la fase de pruebas, los KPIs deben estar alineados con los objetivos específicos del proyecto. Algunos KPIs relevantes en esta fase de test son los siguientes:

- **Tiempo de procesamiento de datos (ETL):** El tiempo que toma procesar el archivo CSV a través del job de Glue.
- **Precisión de los datos cargados en RDS:** Porcentaje de datos correctos tras el procesamiento y carga en la base de datos.
- **Tiempo de respuesta de las funciones Lambda:** Tiempo medio de respuesta de las funciones Lambda al realizar consultas a la base de datos.
- **Consistencia de los datos en las funciones Lambda:** Porcentaje de consultas de Lambda que devuelven resultados correctos sin errores.
- **Disponibilidad del sistema (Uptime):** Porcentaje de tiempo en que los servicios (Lambda, RDS, Glue, etc.) están operativos durante las pruebas.

Cada KPI es fácilmente medible y cuantificable, utilizando unidades de medida estándar y asegurando que los datos se recopilan de manera coherente.

2. Recopilación de Datos Consistente
Los datos necesarios para medir estos KPIs deben ser obtenidos de manera consistente. Las fuentes de datos relevantes incluyen:

- **Logs de AWS Lambda:** Para medir el tiempo de respuesta de las funciones Lambda y verificar que las consultas se ejecutan correctamente.
- **CloudWatch:** Para analizar el rendimiento de Glue y Lambda.
- **Consultas en MySQL (RDS):** Para verificar la precisión de los datos cargados y las relaciones entre tablas.
- **AWS Glue y RDS Metrics:** Para medir el tiempo de procesamiento de los datos, la eficiencia del ETL y la carga en RDS.

Los datos se recopilan utilizando **CloudWatch** para las métricas y logs, y consultas en **MySQL** para las validaciones de precisión de datos.

3. Validación de la Precisión
Para garantizar que los datos y resultados de las pruebas sean precisos, se deben realizar las siguientes acciones:

- **Auditoría de Logs:** Revisión periódica de los logs generados por Lambda y Glue en **CloudWatch** para asegurar que no haya errores en el procesamiento de datos ni en las consultas de Lambda.
- **Comparación de Resultados:** Verificación de que los resultados obtenidos de las funciones Lambda y las consultas en MySQL sean correctos. Esto incluye la validación de los datos cargados en RDS comparándolos con los datos del archivo CSV original.
- **Comprobación de Cálculos:** Validación de cálculos derivados, como el precio por metro cuadrado, para asegurar que se realicen correctamente en el ETL.

Durante la fase de pruebas, se puede realizar una auditoría manual comparando los resultados esperados con los obtenidos de las consultas en Lambda y MySQL.

4. Análisis de la Relevancia
Para garantizar que los KPIs sean relevantes, se deben realizar las siguientes acciones:

- **Análisis de Correlación:** Realización de análisis estadísticos para determinar si los KPIs están correlacionados con los resultados deseados. Por ejemplo, se puede analizar si el tiempo de respuesta de Lambda está correlacionado con la cantidad de datos procesados en Glue.
- **Feedback Continuo:** Recopilación de retroalimentación de los stakeholders para asegurar que los KPIs sean relevantes y reflejen las preocupaciones de los usuarios del sistema.

El análisis de correlación puede incluir estudios estadísticos para validar la efectividad de los KPIs, como la relación entre el volumen de datos procesados y el tiempo de respuesta de las funciones Lambda.

5. Revisión y Ajuste Continuos
Dado que este es un proyecto en fase de pruebas, es fundamental revisar y ajustar los KPIs a medida que se identifican posibles problemas de rendimiento. Las actividades a realizar incluyen:

- **Revisión de los KPIs:** Revisión periódica de los KPIs, analizando los logs y métricas para identificar desviaciones o caídas en el rendimiento.
- **Ajustes en los KPIs:** Si se detectan problemas, como tiempos de respuesta lentos o procesamiento ineficiente, los KPIs pueden ajustarse para reflejar mejor el rendimiento del sistema en pruebas.

La revisión de los KPIs debe ser continua, ajustando las métricas según se necesite para reflejar el rendimiento real del sistema.

6. Documentación de los Métodos Utilizados
Es fundamental documentar el proceso de medición y validación de los KPIs, detallando los métodos utilizados:

- **Métodos de Recopilación de Datos:** Descripción de cómo se recopilan los datos de los KPIs a través de logs de CloudWatch, consultas MySQL y métricas de AWS.
- **Auditorías Realizadas:** Detalles de las auditorías periódicas realizadas para verificar la precisión y consistencia de los datos.
- **Análisis Estadístico:** Documentación del análisis de correlación para validar la relevancia de los KPIs.
- **Transparencia:** Mantener total transparencia en el proceso para que los stakeholders tengan confianza en los KPIs definidos.

KPIs a Medir en la Versión de Test A continuación se muestran los KPIs que se medirán en la versión de pruebas del proyecto:

Esta estructura asegura que se cubren todos los aspectos relevantes de los KPIs durante la fase de prueba y permite medir de manera precisa el rendimiento del sistema en este entorno.

KPI	Descripción	Método de Medición
Tiempo de procesamiento ETL	Tiempo para procesar el archivo CSV a través de AWS Glue.	Medición de logs de Glue y métricas de CloudWatch.
Precisión de los datos cargados en RDS	Porcentaje de datos correctos tras la carga en RDS.	Comparación de datos entre el archivo CSV y RDS.
Tiempo de respuesta Lambda	Promedio de tiempo de respuesta de las funciones Lambda.	Logs de CloudWatch de Lambda.
Consistencia de los datos en Lambda	Porcentaje de consultas Lambda que devuelven resultados correctos.	Comparación de resultados de consultas con los esperados.
Disponibilidad del sistema (Uptime)	Porcentaje de tiempo que los servicios están operativos.	Métricas de CloudWatch para Glue, Lambda y RDS.

Cuadro 2: KPIs a medir en la versión de test del proyecto.