**Prueba técnica ingeniero de datos NEQUI**

1. **Alcance del proyecto y captura de los datos.**

El Dataset consiste en un archivo CSV que contiene la información de casos de COVID-19 en las diferentes regiones y ciudades del territorio colombiano y también algunos casos de otros países tomado de:

<https://www.datos.gov.co/es/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr>

Estos datos serán analizados y procesados con el fin de cumplir los siguientes casos de uso:

Teniendo en cuenta que será enfocado únicamente para Colombia.

1. Análisis de casos por departamento y municipio:

Obtener el número de casos por departamento y municipio.

Identificar los municipios con mayor número de casos.

2. Análisis demográfico:

Analizar la distribución de edades de los casos

Calcular una estadística descriptiva, como la edad promedio, la mediana, etc.

3. Análisis de contagio y ubicación:

Examinar los tipos de contagio y su proporción en los casos.

4. Análisis de recuperación y fallecimiento:

Analizar el estado de los casos (recuperado, fallecido).

Calcular la tasa de recuperación y la tasa de letalidad.

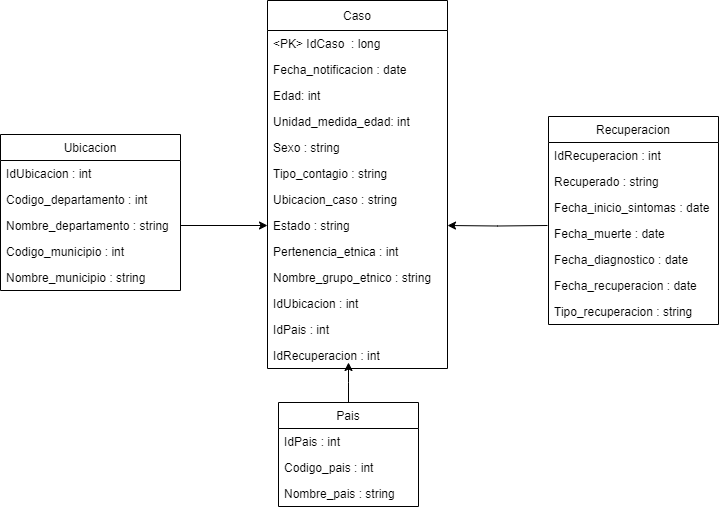
1. **Exploración y evaluación de los datos.**

Realizando la exploración la cual podremos observar más detalladamente en el archivo EDA en la carpeta sources, podemos darnos cuenta de que para la información que necesitamos para cumplir los casos de uso tenemos los siguientes problemas:

* Datos en blanco para 5 de las columnas que necesitamos
* Datos combinados con mayúsculas y minúsculas
* Datos mal registrados como ciudades en la columna departamentos

Para la limpieza de estos datos vamos a realizar los siguientes pasos:

1. Cambiar los espacios blancos en la columna país con el valor Colombia debido a que en el EDA pudimos darnos cuenta de que si corresponde.
2. Limitar el dataset únicamente a los registros correspondientes a Colombia
3. Poner todos los registros de la columna departamentos en mayúscula debido a que la mayoría tiene ese formato.
4. Poner todos los registros de la columna municipio en mayúscula debido a que la mayoría tiene ese formato.
5. Poner todos los registros de la columna sexo en mayúscula debido a que la mayoría tiene ese formato.
6. Poner todos los registros de la columna Recuperado en mayúscula debido a que la mayoría tiene ese formato
7. **Definición del modelo de datos.**



El modelo conceptual se definió de esta forma para con el objetivo de presentar los elementos más destacados de toda la data y las relaciones más relevantes presentes en los datos de forma clara y concisa, al diseñar las tablas, se buscó evitar la redundancia de los datos y su fácil comprensión.

* **Arquitectura propuesta**

A picture containing screenshot, text, diagram

Description automatically generated

La arquitectura esta compuesta de la siguiente forma:

Un AWS data pipeline que se encarga construir un flujo para orquestar cada uno de los demás servicios dentro de este, los cuales se clasifican de la siguiente forma

1. el servicio IAM: se encarga de controlar y gestionar los accesos a los demás recursos.
2. Amazon s3: este bucket se va a encargar de almacenar los archivos de la data principal, para este caso el archivo CVS.
3. Amazon EMR: es un servicio el cual se va a encargar del procesamiento de los datos, este servicio puede usar código Python para ejecutar tareas de procesamiento de datos y análisis.
4. Amazon RDS: Es el servicio de bases de datos relacionales el cual vamos a utilizar para almacenar los datos finalmente procesados.
5. Por último, tenemos el visualizador: en este caso tenemos Power bi, que se va a enlazar con la base de datos del RDS para mostrar la información de forma clara y legible

**¿Con que frecuencia actualizar la data?**

Se recomienda actualizar a diario la data, diariamente el Instituto Nacional de Salud genera reportes acerca de los casos de COVID-19 en Colombia en el siguiente enlace:

<http://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx>

1. **Complementando el proyecto**

¿Cuál es el objetivo del proyecto?

El principal objetivo de este proyecto es limpiar y procesar una data de los datos abiertos de Colombia, para poder así analizarla de manera ágil.

¿Qué preguntas quieres hacer?

- ¿En qué departamento de Colombia hay más casos de COVID-19?

- ¿En qué municipios de Colombia hay más casos de COVID-19?

- ¿Que edades son las que más casos de COVID-19 reportan?

- ¿Que tipos de contagio son los más frecuentes?

- ¿Cuál es la tasa de recuperación de los casos de COVID-19?

¿Por qué eligió el modelo que eligió?

El modelo conceptual se definió de esta forma para con el objetivo de presentar los elementos más destacados de toda la data y las relaciones más relevantes presentes en los datos de forma clara y concisa, al diseñar las tablas, se buscó evitar la redundancia de los datos y su fácil comprensión.

* **Manejo diferentes escenarios.**

1. Si los datos se incrementaran en 100x.

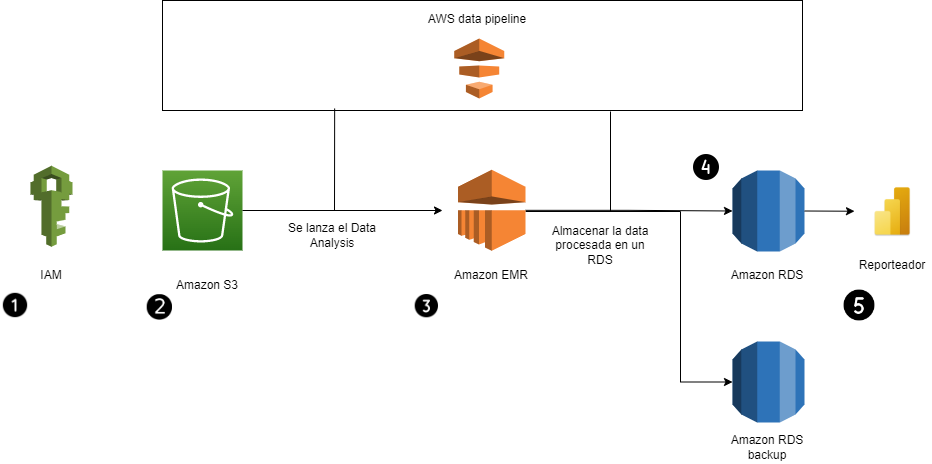
En caso de que los datos aumentaran en esa cantidad se puede inicialmente, normalizar una vez mas la base de datos, crear tablas de datos únicos como la información de la persona en el caso, y optimizar algunos atributos en la forma más atómica posible.

1. Si las tuberías se ejecutaran diariamente en una ventana de tiempo especifica.

Se deben aumentar las capacidades de procesamiento de los datos debido a que en algún momento en el tiempo la ventana no será suficiente para mover los datos existentes.

1. Si la base de datos necesitara ser accedido por mas de 100 usuarios funcionales.

En caso de que la base de datos debe ser accedida por mas de una persona se aplican las siguientes normas para mantener la integridad de los datos, primero se debe tener una alta granularidad en los permisos del servicio IAM para que cada una de las personas tenga los permisos mínimos únicamente a realizar las tareas necesarias, también es muy importante tener un respaldo de esta base de datos actualizados periódicamente y automático.



1. Si se requiere hacer analítica en tiempo real, ¿Cuáles componentes cambiarían a su arquitectura propuesta?

En caso de querer un procesamiento en tiempo real Amazon tiene un servicio que se llama Amazon Kinesis Data Stream, este servicio permite procesar los datos a medida que son recibidos donde se configuran una serie de reglas que definen en destino de la data, entonces lo que hay que modificar en la arquitectura es que ya no vamos a utilizar el data pipeline sino utilizaremos el servicio anteriormente mencionado.

A picture containing text, screenshot, diagram, plot

Description automatically generated

De esta forma mantenemos el s3 que contiene la data original o histórica y agregaremos un api Gateway para asegurar el origen de datos, y posteriormente pasa al servicio de procesamiento de datos en tiempo real, enviamos a nuestro servicio que analiza la data para enviarla a las respectivas bases de datos.