Prueba Técnica Data Engineer Junior

El rol de Data Engineer es crear y administrar la Arquitectura para disponibilizar procesos de Data. Esta prueba técnica evalúa tu capacidad de aprender nuevos conceptos y tecnologías, tu capacidad de leer documentación y aplicarla para lograr resultados.

Descripción: Después del análisis, el equipo de Data te envía una matriz de relaciones entre personas, un 1 cuando tienen relación y un 0 cuando no tienen ningún tipo de relación. (Archivo de Excel adjunto a este email).

		1	2	3	4	5	6	7	8	9	10	11
		A	B	C	D	E	F	G	H	I	J	K
1	A	0	0	0	0	0	1	0	0	0	1	0
2	B	0	0	0	0	0	0	0	0	0	0	0
3	C	0	0	0	1	1	1	1	1	1	1	1
4	D	0	0	1	0	1	1	1	1	1	1	1
5	E	0	0	1	1	0	1	1	1	1	1	1
6	F	0	0	1	1	1	0	1	1	1	1	1
7	G	0	0	1	1	1	1	0	1	1	1	1
8	H	0	0	1	1	1	1	1	0	1	1	1
9	I	0	0	1	1	1	1	1	1	0	1	1
10	J	0	0	1	1	1	1	1	1	1	0	1
11	K	0	0	1	1	1	1	1	1	1	1	0
12	L	0	0	1	1	1	1	1	1	1	1	1
13	M	0	0	0	0	0	0	0	0	0	0	0
14	N	0	1	1	1	1	1	1	1	1	1	1
15	Ñ	0	0	0	0	0	0	0	0	0	1	0
16	O	0	0	0	0	0	0	0	0	0	0	0
17	P	0	0	0	0	0	0	0	0	0	0	0
18	Q											

El archivo tiene información hasta la row 17, así que faltan relaciones. El equipo de Data nos enviará este archivo actualizado diariamente, y te solicitan montar un pipeline de datos usando Dagster (Orquestador de Pipelines de Datos).

Reto: Como Data Engineer debes tomar estos datos y generar un Pipeline que cargue esta info en una nueva tabla de mysql (genera una Base de Datos Mysql local) y que este proceso se ejecute cada 24h actualizando la tabla de Mysql utilizando Dagster.

Resultado esperado:

Código en Python con la configuración del Pipeline de datos, junto con los querys utilizados para generar la DB y tablas. En este código esperamos se evidencien los pasos donde se lee el excel, organizan los datos y carga a mysql en una nueva tabla junto con su respectiva automatización empleando Dagster.

Documentación Recomendada:

- Adjunto a este email encontrarás el paso a paso detallando de como configurar Dagster en local y crear tu propio ETL.
- Documentación Dagster job: https://docs.dagster.io/tutorial/intro-tutorial/single-op-job

Código de Ejemplo de Automatización:

```python
from dagster import schedule
from datetime import datetime, time, date

@schedule(
        cron_schedule= "*/20 * * * *",
        pipeline_name= "load_bigquery_table_pipeline",
        execution_timezone="America/Bogota"
)
```

Modalidad de Calificación:

Tomaremos el código y lo ejecutaremos en un nuevo entorno, ejecutaremos los querys para generar la DB y el pipeline, esperamos que el pipeline se ejecute cada 24 h y actualice la tabla en Mysql, cargaremos diferentes versiones del archivo de excel y esperamos que la tabla de la base de datos se actualice con los cambios.

Envía tu solución a camilobaquero@habi.co

Tiempo para realizar la prueba: Tienes 3 días calendario para enviar la solución de este reto después de recibir este email. Envía la prueba lo más pronto posible, se tendrá en cuenta como puntos adicionales en la calificación de la prueba entre más pronto envíes la solución.

Punto Opcional: Puede complementar su pipeline de datos utilizando otras tecnologías y librerías. Adjuntando la documentación correspondiente explicando el proceso.