Technical Challenge Spot (Data Engineer) (GeoData)

En spot² una de las necesidades del día a día es el procesamiento de datos crudos, definir una estrategia de extracción, procesamiento y almacenamiento que respondan a las necesidades del equipo y solucione la infraestructura de ingesta y consulta de datos para cualquier manipulación requerida por el resto de los miembros de la organización.

Para esto, un cliente ha confiado en spot² como aliado para abrir su próximo negocio, una pizzería en la ciudad de México. Se han levantado los requerimientos y tenemos una lista de spots disponibles que se ajustan a estos:

Spot	Dirección	Latitud	Longitud
423	Rio Lerma 65	19.43061268916689	-99.16507358911448
632	Zacatecas 24	19.4166999785185	-99.1556043737739
702	Oklahoma 84	19.391890374793046	-99.17759500279307
1021	Av. Universidad 707	19.375386845013207	-99.16212560238631
937	Felix Parra 117	19.366311887545983	-99.18504078433399

Por cada spot, nos gustaría saber cuántas pizzerías hay en un radio de 500 m., 1 km y 3 km, además de poder **consultar el nombre de cada una de estas pizzerías y su tamaño en fuerza laboral** (número de empleados) de los últimos 3 años. Para esto te pedimos extraer la información que ofrece el Directorio Estadístico Nacional de Unidades Económicas (DENUE) sobre el **comercio al por menor en México**, procesa, limpia y manipula la data hasta el punto que lo creas conveniente para luego implementar una función que calcule la distancia entre el spot y las pizzerías dentro de los radios solicitados.

En el siguiente link podrás acceder a los datos del DENUE:
https://www.inegi.org.mx/app/descarga/?ti=6

Estructura la salida como mejor lo veas conveniente de tal manera que la información filtrada se pueda almacenar en una base de datos y sea de fácil acceso para el resto del equipo. Toma en cuenta que a futuro se realizarán más consultas con diferentes spots y diferentes tipos de unidades económicas (no solo pizzerías), es decir, la base crecerá con el tiempo, por lo cual es importante que el acceso a la información esté optimizado.

Desarrolla y justifica el diseño de la arquitectura de este servicio y muestra un prototipo funcional de la aplicación que podría utilizar estas bases de datos para la toma de decisiones (puedes usar el framework de tu preferencia, Jupyter Notebook, AWS Sagemaker, etc) y argumenta los hallazgos obtenidos en el ejercicio.

Requerimientos:

- Tendrás **dos días** para desarrollar este desafío.
- Tech Stack sugerido: SageMaker, EC2 y RDS/RedShift. Usar otras herramientas también es posible, sin embargo en la entrevista debes justificar su uso.

Resultados esperados:

- Diseño justificado de la estructura de la base de datos.
- Pipeline ETL de los datos DENUE.

Entregables:

- Entrega código y figuras a través de tu repositorio de GitHub.
    - El README.md debe ser muy claro para poder replicar tu infraestructura.
- Muestra las visualizaciones que creas más interesantes. (Mapas, estadísticas, tablas, cuadros, matrices, diagramas, etc.)
- **BONUS:**
    - Deploya la aplicación en tu proveedor de nube de preferencia. Es importante que menciones qué servicios estás utilizando y por qué.
    - Si utilizas algún tipo de herramienta de CI/CD es importante mencionarlo.

Consejos:

- No hay respuestas incorrectas, queremos entender tu proceso creativo.
- Piensa fuera de la caja.
- ¿Por qué deberíamos adoptar tu solución o qué mejorarías?
- ¿Este sería un código que pondrías en producción?

¿Dudas? Envíame correo a dante@spot2.mx