Ejercicio: Data Warehousing

En Mercadolibre(MELI) navegan unos 80 millones de usuarios mensualmente, los cuáles generan un volumen de actividad y eventos enorme. Hoy vas a poder acceder a un subset del mismo.

El dataset que adjuntamos en el siguiente ejercicio, contiene porciones de la navegación de nuestros usuarios en nuestra plataforma. Cuenta con las siguiente columnas:

Nombre Columna	Descripción	Tipo
event_name	Sección del sitio navegada. Puede tomar los valores: - SEARCH (búsqueda realizada) - PRODUCT (producto visto) - CHECKOUT_1 (visito el 1er paso del checkout) - CHECKOUT_2 (visito el 2do paso del checkout) - CHECKOUT_3 (visito el 3er paso del checkout) - BUY (compra realizada)	string
item_id	item_id que se vio	númerico
timestamp	Tiempo en el cual ocurrió dicho evento	timestamp
site	Sitio/País desde el cual el usuario navegó mercadolibre	string
experiments	Listado de experimentos a los cuales fue sometido dicho usuario en esa sección del sitio	Map<exp,variant>
user_id	Usuario que navegó el sitio en cuestión	númerico

El data set se puede encontrar en
https://drive.google.com/file/d/1q-kVDe62HY-6SbLsetsi1s1vvYgzPUOi/view?usp=sharing

En el contexto de MELI, corremos cientos de AB testings para entender qué features/ideas/variantes llevan a los usuarios a comprar dentro de mercadolibre y que cosas no. La columna experiments, tiene por objeto guardar un listado de todos los experimentos a los que fue sometido un usuario a lo largo de su navegación. Nótese que la forma en la cual se guarda implica entries/tuplas del formato *experiment_name => variant_id*. Siendo estas interpretables como “al ver esta página el usuario ${user_id} participó del siguiente experimento ${*experiment_name}*, viendo la variante ${*variant_id}*”.

Nivel 1

Cargar los datos en algún storage de preferencia y responder mediante una query SQL.

a. *¿Cuál es la hora del día en que se realizan más búsquedas en MercadoLibre?*
b. *¿Cuál fue el experimento que tuvo más participantes dentro del dataset?*

Nivel 2

Pensando en la posibilidad de que sea accedido por miles y miles de usuarios…

c. ¿Qué estrategias utilizarías para asegurar que el acceso a los recursos de dicho motor sean justos y equitativos entre los distintos usuarios?
d. ¿Qué controles podríamos implementar para garantizar la auditoria y el control de acceso en dicho motor de consulta?

Pensando en la posibilidad de que el dataset crezca en 1000x su tamaño actual (llegando a los cientos de millones de registros por día)

e. ¿Qué estrategias o modificaciones le harías al guardado de la tabla en tu storage para optimizar las consultas sobre la misma?
f. ¿Qué estrategias de ahorro de costos podemos implementar para evitar el crecimiento indefinido de este storage?

Nivel 3 (opcional)

Hostear el dataset en algún motor de consulta SQL cloud de preferencia (MySQL, RedShift, BigQuery, Athena, SparkSQL, Azure SQL DataWarehouse, etc)[1] y disponibilizar instrucciones para su conexión remota y permitir ejecutar consultas SQL.

Se valora la entrega de un script/notebook/snippet para su conexión en el lenguaje de preferencia.

ENTREGABLES:

- Código fuente de la implementación del challenge (En repositorio *github privado*).
- Readme (En repositorio *github*) que incluya:
  - Instrucciones para utilizar y arrancar el repositorio
  - Respuestas del nivel 1.
  - Respuestas y consideraciones del Nivel 2 incluyendo links.
  - Documentación de acceso a la conexión y snippets/instrucciones de consulta del dataset (Nivel 3)

[1] Usar algún hosting gratuito. No vamos a estresar la solución bajo ningún punto de vista:
https://aws.amazon.com/es/free/
https://cloud.google.com/free
https://azure.microsoft.com/en-us/pricing/free-services/