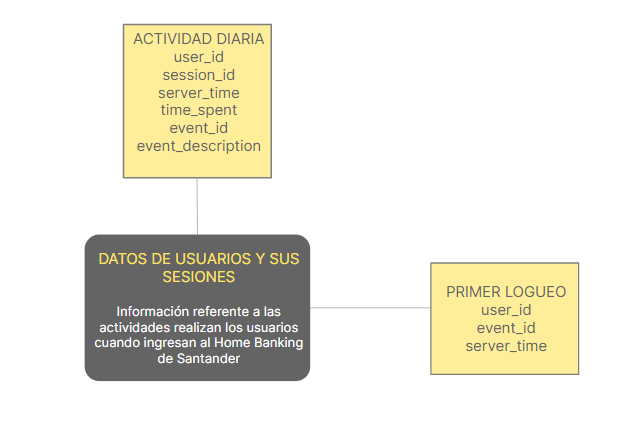
# Data Engineer Programming Test

**Pregunta 1**

Como resolvería este tipo de petición? Explique detalladamente el proceso de limpieza y transformación del modelo inicial. Que tecnologías utilizaría y por qué?

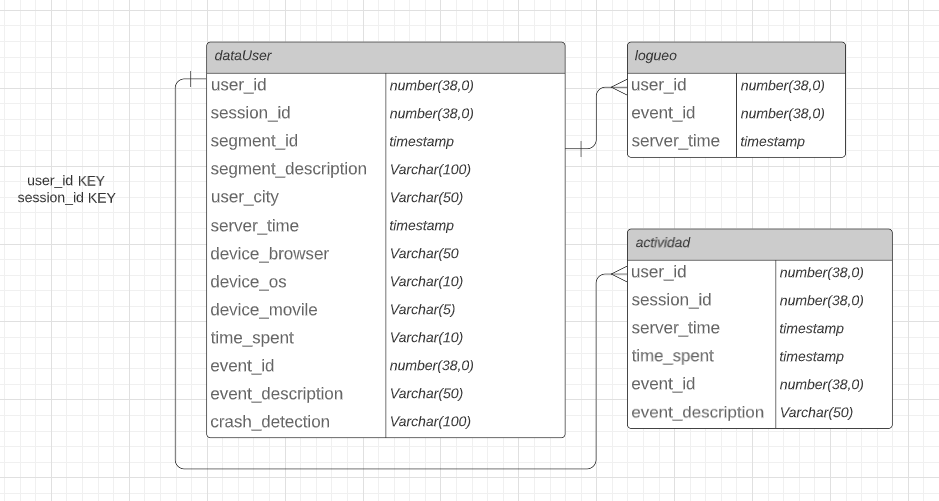
*Usaría el siguiente modelo dimensional*

*Extraería los datos de Big Query y dependiendo de lo que necesite:*

* *En caso de no necesitar hacer trasformación a los datos, armaría solo un ETL con NIFI, extraería los datos de BigQuery con algún procesador como* **“ExecuteSQL”** *(dejando los datos en formato avro), posteriormente lo alojaría en HDFS con un procesador* **“PutHDFS”** *con formato parquet.*
* *Si por ejemplo se necesita alguna transformación en especial deberíamos usar algún/algunos procesadores además de los dos procesadores mencionados anteriormente.*
* *Y sino para obtener KPI’s directamente de la base de datos se podrían construir y visualizar en Power BI (Elijo este porque tiene una forma más fácil e intuitiva de conectarse a Big Query).*

**Ejercicio 1**

Realice el DER que de soporte al modelo dimensional solicitado por la banca privada.



**Ejercicio 2**

Escriba las queries necesarias partiendo de la tabla inicial y que de como resultado el modelo planteado en el ejercicio anterior.

*create table actividad as (select user\_id,SESSION\_ID ,SERVER\_TIME , TIME\_SPENT , EVENT\_ID , EVENT\_DESCRIPTION from santander.dataUser);  
  
  
create table logueo as (select USER\_ID ,EVENT\_ID ,SERVER\_TIME from santander. dataUser);*

**Ejercicio 3**

Escriba la consulta necesaria para obtener el KPI de retención de clientes para los 10 clientes que más veces se hayan logueado en el último mes.

*select u.user\_id,u.EVENT\_ID ,u.EVENT\_DESCRIPTION ,count(u.USER\_ID) as CANTIDADLOGUEOS from userhb u where u.EVENT\_DESCRIPTION = 'logueo' and month(sysdate()) = month (u.SERVER\_TIME) group by u.USER\_ID order by CANTIDADLOGUEOS desc limit 10;*

### Pregunta 2

Suponga que la ingesta de estos datos se realiza utilizando Apache Spark debido a que la tabla cruda tiene billones de registros. Que parametros de spark tendría en cuenta a la hora de realizar dicha ingesta? Explique brevemente en que consta cada uno de ellos. En que formato de compresión escribiría los resultados? Por qué?

**spark.app.name** *nombre de la aplicación. Esto aparece en la interfaz de usuario y los datos de registro.*

**spark.driver.cores** *El número de núcleos de CPU necesarios para que se ejecute el programa del controlador.*

**spark.driver.memory** *La cantidad de memoria utilizada por el proceso del controlador.*

**spark.executor.memory** *La cantidad de memoria utilizada por cada proceso ejecutor. Tiene el mismo formato que la cadena de memoria JVM (por ejemplo, 512 m, 2 g).*

**spark.logConf** *Cuando se inicie SparkContext, registre el SparkConf válido como INFO.*

**spark.master** *Donde se conecta el administrador de clúster.*

**spark.hadoop.parquet.block.size** *Cantidad mínima de datos que puede leer de un archivo de parquet que es lógicamente legible por lo que no puede tener más hilos de lectura que los bloques de entrada,* e*l valor predeterminado es 128 Mb por bloque.*

**spark.sql.files.maxPartitionBytes** *El número máximo de bytes para empaquetar en una sola partición al leer archivos. Esta configuración solo es efectiva cuando se utilizan fuentes basadas en archivos, como Parquet, JSON y ORC.*

**spark.sql.files.openCostInBytes** *El costo estimado para abrir un archivo, medido por la cantidad de bytes que se pueden escanear al mismo tiempo. Esto se usa cuando se colocan varios archivos en una partición. Es mejor sobreestimar, entonces las particiones con archivos pequeños serán más rápidas que las particiones con archivos más grandes (que se programan primero). Esta configuración solo es efectiva cuando se utilizan fuentes basadas en archivos, como Parquet, JSON y ORC.*

* *Elegiría el formato de comprensión parquet porque es el mejor formato para el rendimiento, ya que tiene una compresión eficiente, almacena datos en formato de columnas y está muy optimizado en Spark.*

### Pregunta 3

Existen varios problemas en cuanto a la calidad de datos de la tablas que consultan los usuarios de la banca privada, se está investigando como mejorar y prevenir estos incidentes. Describa brevemente que implementaría para garantizar la confiabilidad de los datos.

*Optaría por importar la base de datos a HDF, o utilizar una base de datos NoSQL (por ejemplo, oMongoDB) como back-end.* *Realizaría auditorías periódicas, podría crear diferentes capas/fases como staging y verificar los datos desde su fase inicial, cuidando el proceso de recolección, y transformación de los mismos que nos permita descubrir fallos en el proceso de ingestas, tener métricas para detectar posibles fallos que puedan estar surgiendo. De esta forma es más probable que podamos tener confiabilidad sobre los datos (****Consistencia****,* ***Exactitud****,* ***Integridad****,* ***Completitud****) y evitar errores que nos lleven a tomar malas decisiones.*