**GDT – BI Dev Challenge**

**About**

As a **Business Intelligence Developer** working for the Group Data Team (GDT) at Swiss Marketplace Group (SMG), you will take part of all the data development lifecycle from their extraction, loading and transformation (ELT) using Data Build Tool (dbt) into the Data Warehouse (DWH) platform BigQuery (BQ) until their data visualization (DataViz) platform Looker (LK), all using Google Cloud Platform (GCP) environment.

This challenge has been designed to evaluate the **required** and also the **preferable** technical skills.

For the required skills, we will evaluate the knowledge and understanding of core data warehousing concepts, business intelligence techniques and quality of work. For the preferable skills, we will evaluate the knowledge about the tools we use in our stack (when possible).

The goal of this project is to provide a comprehensive BI Dev hiring challenge to our candidates, offering a standardized and well-structured case study while allowing the candidates full autonomy over the approach.

**Solution:**

* from Excel spreadsheets, SQL queries on BigQuery, to a full-blown Python environment, any approach is welcome as long as it gets the job done.

**Answers:**

* We expect written scripts and/or logical statements and for data visualizations it does not matter what tools you’ve used. They can all be added into doc, txt, pdf or sql files.
* The answers could be submitted via email in a zip file OR
* If you would like to win extra points, you can submit a pull request with your answers in your personal Github repository, in a separate branch and share with us.

There's no strict time constraint on this case study. We expect candidates to inform us when they are ready and to respect the honour system on the time invested and approaches taken.

We're estimating a maximum of 3 working hours to complete 100% of what this case study has to offer. Candidates are offered 24h to return the case study. In case of investing more (or less) than the estimate, we expect open communication on the context and welcome any approach and efforts invested.

**Getting Started**

#### **Assumptions**

* The tasks should be done in a sequential order from challenges #A to #C.
* The data model contains a table with cleaned data called “cln\_listings” with the current data state copied from the original data source.
* The data model contains a star schema and the fact table contains periodic snapshots of listings (classifieds).
* The data is a fictional dataset created specifically for this case study and does not reflect reality.
* The data or structure might be inaccurate and we will be delighted if you spot something and give us feedback.
* Platforms were renamed from Anibis.ch to Anibis and from Tutti.ch to Tutti and new names should be used in analysis in order to have normalized results.

#### **Evaluation criteria**

* The process matters more than the final result and the goal is to understand how you solve problems.
* The final output should be well-organized and easy to understand.
* The insights provided should be relevant and actionable.
* The data visualizations should be intuitive to use and provide a clear view of the data.
* The explanation of the process should be clear and concise.
* The most important thing is the level of understanding of core BI Development concepts and best practices.

#### **Scoring**

* Challenge A: 3 points (1.5 points per task).
* Challenge B: 5 points (0.5 point per component).
* Challenge C: 2 points.
* Extra 1 point: Submitting your answers via your private Github repository and sharing with us.

**Challenge A (Total: 3 Points):**

* **Scenario 1:** As a business intelligence developer, I would like to "look back in time" at previous data states of the ‘general\_marketplaces.cln\_listings’ *table.* While some source data systems are built in a way that makes accessing historical data possible, this is not the case here: this table has only current data state. In order to record changes to this mutable table over time, it is necessary to use a mechanism to do so.
* **Task 1 (1.5 Points):** Create or describe a mechanism/script to register each data change (snapshot of end of day) from the clean table ‘general\_marketplaces.cln\_listings’ *into the fact table* ‘general\_marketplaces.fct\_listings’ for each day of the interval considering validity of the records (SCD type 2).
* **Supporting Resources 1:** Kimball documentation and dbt documentation *(not mandatory to use dbt, a simple sql script can have the job done\*).*

Kimball SCD type 2

<https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/type-2/#:~:text=Slowly%20changing%20dimension%20type%202,multiple%20rows%20describing%20each%20member>.

dbt incremental model

<https://docs.getdbt.com/docs/build/incremental-models#about-incremental_strategy>

dbt snapshot

<https://docs.getdbt.com/docs/build/snapshots>

* **Scenario 2:** As a business intelligence developer, I would like to see all the listings (classifieds) changes on a daily basis for each platform in order to track its evolution and answer business questions.
* **Task 2 (1.5 Points):** Create a report for the period from Dec. 2021 to Jan. 2022 (2 months). Also consider clean-up/mapping data changes that occurred during this period, for example related to platform re-naming recorded in the dimensional table ‘general\_marketplaces.dim\_platform’.
* **Supporting Resources 2:** Spreadsheet with flat dataset to be used for the report OR CSV files with star schema dataset to check for modifications on the tables and data analysis.

**Challenge B (Total: 5 Points):**

* **Scenario:** Swiss Marketplace Group (SMG) wants to analyse its general listings (classifieds) data to determine which product types are the most popular and which ones are idle. The company wants to use this information to make decisions about which product types they should invest in different ways of promoting (idle listings).
* **Task 1 (2.5 Points):** Write a script/query that shows the following information:

a) The top 3 selling product types by platform.  
b) The bottom 3 selling product types by platform.  
c) The top 3 idle product types (amount of days).  
d) The total amount sold by product type (monetary value rounded to two decimals).  
e) Any other insights you could learn from the data that would be useful for the company to know.

* **Supporting Resources 1:** Given that you completed the Challenge A Task 2, you can use the created report output as a start.
* **Task 2 (2.5 Points):** Create data visualizations (charts or/and dashboards) that can be used to explain the results of task 1 (a,b,c,d,e).

a) The top 3 selling product types by platform.  
b) The bottom 3 selling product types by platform.  
c) The top 3 idle product types (amount of days).  
d) The total amount sold by product type (monetary value rounded to two decimals).  
e) Any other insights you could learn from the data that would be useful for the company to know.

* **Supporting Resources 2:** Given that you completed the Challenge B Task 1, you can use the results (a,b,c,d,e) as an input data for the visualizations OR you can use the CSV file (challenge\_B\_task\_02\_flat\_dataset) as a fall-back mechanism to create your data visualizations in case you could not complete Task 1 in time.

**Challenge C (Total: 2 Points):**

* **Scenario 1:** The fact table ‘general\_marketplaces.fct\_listings’is the central part of a star schema and contains periodic snapshots of listings (classifieds). This table is partitioned on the listing\_date\_key column and is clustered on the platform\_id, product\_type\_id, status\_id and user\_id columns. The dimensional table ‘general\_marketplaces.dim\_user’ contains granular information about user location. The dimensional table ‘general\_marketplaces.dim\_product\_type’ contains granular information about product type tags.
* **Task 1 (2 Points):** Please write a SQL script using BigQuery syntax that shows the top 3 countries by number of listings with the product type having the black color for product type tag.
* **Supporting Resources 1:** Entity Relationship Diagram (ERD - entity\_relationship\_diagram.jpg) and BigQuery Syntax documentation**.**
* **Pre-requisites 1:** If you have never used BigQuery, please check how to manipulate the RECORD and REPEATED columns:

For the RECORD column, a.k.a. STRUCT fields: you can query the ‘location.city’ column as follow:

SELECT location.city FROM ‘general\_marketplaces.dim\_user`

For REPEATED columns, a.k.a. ARRAY columns:

<https://cloud.google.com/bigquery/docs/reference/standard-sql/arrays>

[OPTIONAL] For an overall documentation about the BigQuery queries syntax:

<https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax>