Technical assignment: Data Quality Analyst

Our technical assignment consists of three tasks -

- first being a practical task where you can showcase your SQL knowledge
- second is more of a theoretical one where you have to describe your thought process,
- and the last one is a practical task with a real dataset which you can do in Excel, SQL or other tool of your liking.

As we understand that some of the tasks can be time consuming, in cases where a similar task has been done before (for instance, a personal pet-project that shows off your technical skill set in a similar matter) we accept the respective project as a substitute.

// Submission

Solution can be submitted in a form of your choice - might that be a link to a google docs, a git repository, a notebook, a presentation or a text file. Most importantly is to include items described in each task.

In case you substitute some parts of the assignment with some other project or task, please clearly describe how to access the respective task, what was the goal of that task and which part of the task was done by you independently.

Submit your task by replying to the original technical assignment email. We'll be happy to hear from you within the agreed time frame, - and if we don't, we will automatically accept your submission as cancelled.

In case you have any questions or technical issues, feel free to contact us.

// Task #1: SQL

Couple of months ago[1] Nordigen officially introduced the work-from-anywhere policy, and you have volunteered to write a couple of queries that would help the Operations team to get an overall feeling whether employees are using the option.

Overview of tables are given in subsections below. The **task** is to write SQL queries to perform the following actions:

1. Write CREATE TABLE statement for table `EMPLOYEE` based on sample data;
2. Employee named *Valentīna Konfekte (140200-22221)* is starting next monday. Insert this information into `EMPLOYEE` table;
3. Calculate the average request (of traveling) count per employee since the policy has been introduced;
4. Calculate how often an employee got rejected for a request to work from elsewhere during up until now;
5. Calculate *(what currently looks like)* top 10 countries by all employees in year 2022; *(note: "currently" as year is not yet over)*
6. Calculate average length of approved travel in each country;
7. Find all employees who haven't used the opportunity to work from another country and currently also haven't requested any travel dates;
8. List all employees who have approved travel during the same time to the same destination;
9. List each employee and their location on their birthday. If the birthday falls on a weekend, the value should be just "*Weekend*".
10. List all employees with their preferred method of work - from the office, from home or work from another country.

Note: If you get stuck in a query - describe with words what you wanted to do, how much you were able to figure out and and what was the part you got stuck in.

// Descriptions

- All employee data is saved in the `EMPLOYEE` table. Each employee has a unique identifier saved in `employeeId`.
- Every workday employees add information about their whereabouts in the system. For the Operations team this information is saved in the `ATTENDANCE` table.
  - If employee works from home `office` columns will have value `0`;
  - Whereas if an employee will work from the office, the value is going to be `1` together with additional information in columns - `floor` and `table`.

[1] November 30, 2021

- If an employee decides to work from another country, (s)he adds a request about dates and country in the system that adds an entry in the `TRAVELS` table.
  - Direct team lead or HR team can either approve or reject this request (marked in column `process`)
  - Column `byEmployee` saves information about which employee (`EMPLOYEE.employeeId`) made the final decision for that request.

// Sample data

EMPLOYEE

employeeId	name	surname	personalCode	startDate
1	Alberts	Keda	310172-11223	2021-01-14
2	Sniedze	Ieviņa	280282-22133	2021-02-10
3	Tīna	Zibens	310392-33211	2021-02-11
...	...	...	...	...

TRAVELS

travelId	employeeId	process	byEmployee	country	startDate	endDate
1	2	approved	1	Estonia	2022-01-02	2022-01-31
2	3	approved	1	Spain	2021-12-12	2022-01-12
3	3	rejected	1	Greece	2022-01-30	2022-02-14
...	...	....	...	...	...	...

ATTENDANCE

attendanceId	employeeId	date	office	floor	table
1	1	2022-01-03	1	3	G1
2	1	2022-01-04	1	3	G1
3	1	2022-01-05	0		
...	...		...	...	...

// Task #2: Potential issues

Let's continue with the idea of the previous task. You are already familiar with the process of how tables are populated by the system as well as how they are related to each other.

The task is to think about potential data quality issues. Please **briefly describe** your thought process - your assumptions about the data and potential issues.

// Task #3: Data profiling

Nordigen's business team has had a couple of brainstorming sessions, where one of the ideas is to look into youtube and its video trends and patterns. Data team quickly jumped into possible data sources, set up ETL jobs and - *before a data analyst starts to extract insights* - you are asked to double-check data quality of **this export**.

Data source is fairly new, and not everything is completely clear. The team is still learning how to read this data and how columns are correlated to each other. All information the team knows is described below.

The **task** is to perform data profiling for the dataset. Expected output of the task is:

1. **brief description** of how you checked the data. If needed screenshot(s), code/query snippets, excel formulas can be added and explained;
2. **brief description** of *non-trivial* data quality issues you noticed if any;
3. **brief description** of *non-trivial* improvements in data processing you would suggest if any;
4. your final conclusion whether this data set is good enough for further analysis.

The aim of the task is to demonstrate your ability to notice data quality issues, raise questions and make sure an engineer can fix all bugs or the team is aware of respective problems in data before they invest their time in doing in-depth analysis.

// Data

This dataset includes several months of data on daily trending videos. Data is included for the GB (Great Britain) region only, with up to 200 listed trending videos per day.

Data includes information like the video title, channel title, publish time, tags (comma separated), views, likes and dislikes, description, and comment count. Every video should also have a `category_id` that can be mapped to a supplementary dataset.

// Categories

Categories are assigned based on a tree-based structure. It has three-level depth, where the highest level - named *lvl_0* - can have *lvl_1* sub-category, and *lvl_1* level can have *lvl_2* sub-category. Each transaction is described with at least one level and, at most, three level granularity; and the final category for each transaction is always assigned to the lowest possible level.

For instance, let's assume that we have four video observations. One observation has been assigned to the category ID 600, two transactions have been assigned to the categories ID 101 and ID 801, respectively, and the last transaction is assigned to the category ID 495. The full category tree for all observations is listed below.

	category_id	category_id_lvl_0	category_id_lv_1	category_id_lvl_2	category_title_lvl_0	category_title_lvl_1	category_title_lvl_2
#1	600	600			Music		
#2	101	100	101		Politics & Nonprofits	Nonprofits & Activism	
#3	801	800	801		Entertain-ment	Auto & Vehicles	
#4	495	400	404	495	DIY & Beauty	Beauty	Chloe Morello

Category titles describe the category itself very well, however if you need more information on some parent or child categories, **here** you can find Youtube's guide on how to assign categories based on video content.