UBISOFT

Data Science Test (1h30)

1 –SQL Table Quiz

Consider the following table definitions:

DIM_PLAYER		
Field	***Type***	***Key***
PLAYER_KEY	Integer	PRI
CREATION_DATE	Date	
LAST_UPDATE	Date	
COUNTRY_CD	Varchar(5)	

DIM_INSTALLMENT		
Field	***Type***	***Key***
INSTALLMENT_KEY	Integer	PRI
INSTALLMENT_DESCRIPTION	Varchar(50)	

FT_TRAFFIC		
Field	***Type***	***Key***
ID_TRAFFIC	Integer	PRI
PLAYER_KEY	Integer	
INSTALLMENT_KEY	Integer	
ACTION_DATE_KEY	Integer	
PLAYTIME_VALUE	Integer	

We have two dimension tables (DIM_PLAYER and DIM_INSTALLMENT) and one fact table (FT_TRAFFIC).
Notes regarding the **FT_TRAFFIC** table:

- Contains info about the all games played by players on PC platform day by day.
- ID_TRAFFIC is unique for PLAYER_KEY, INSTALLMENT_KEY and ACTION_DATE_KEY
- In the field ACTION_DATE_KEY we keep the date in the format YYYYMMDD.
- Playtime value :
    - o In minutes. If the playtime is 0: the player owned the game in his library but didn't play with it.
    - o For a given date and a given player, the playtime value is the total playtime for the game until this day.
    - o The playtime value can be the same for two different dates for a player: this player didn't play to this game between the two dates.

1. **Make a query (or many queries) that will return the monthly top games by country according to the playtime, in the format:**

MONTH *(YYYY-MM)*	COUNTRY_CD	INSTALLMENT_ DESCRIPTION	PLAYTIME

while taking into account the following conditions:

- Only games that meet all these two conditions are considered:
  - Players played a maximum of 10 hours per day during the month.
  - The installment name doesn't contain 'beta' or 'test'
- Only the top 100 must be displayed
- The results must be ordered by playtime, in descending order

2. **Make a query (or many queries) that will return the cross-ownership for the top 100 games played during the last 2 weeks.**

2 -Data Science Quiz

1. **How would you explain the bias-variance tradeoff in simple terms?**
2. **How would you predict revenue tomorrow given daily revenue from the past few years?**
3. **What is R^2 ? What are some other metrics that could be better than R^2 and why?**
4. The '*matrix_crossownership.csv*' file is the result of the query (*subtask 2*) in Part 1 of the test.
   How to represent the information in this matrix in a simple way that can be communicated to a non-technical audience?
   a. **Make a data visualization proposal.**
   b. **Explain the technique used.**
   c. **If you had the time to go even further in the analysis, based on data available in Part 1 or others, what would you like to do?**

5. **How would you build a model to predict when a player will churn (stop playing a given game)? How would you define this flag? Which features do you expect to have and would you build? Which models would you try training? What are the expected business use cases of such a model? You can make extensive use of the results of previous questions.**