


DATA ENGINEER TEST

Câu 1: Phân biệt giữa Kimball, One Big Table và Relational Modeling. Đối với vai trò là Data Engineer, bạn sẽ lựa chọn phương pháp nào trong từng trường hợp cụ thể.

Câu 2: Trong Bigquery, tại sao chỉ partition được column dạng *INTEGER*, time-unit, ingestion time?

Câu 3: Trong Bigquery, mô tả nguyên lý hoạt động (ở cấp độ engine) của clustered table, trong trường hợp sử dụng column dạng *STRING*.

Câu 4: Trong Bigquery, so sánh hàm *APPROX_COUNT_DISTINCT* và *COUNT DISTINCT*.

Câu 5: Ứng viên được cung cấp một loạt các file dữ liệu ở dạng JSON-ND đã được nén lại ở chuẩn GZIP  MockData. Dữ liệu chứa các event thu thập được từ 1 game mobile. Ứng viên thực hiện các yêu cầu sau:

a. Import dữ liệu vào BigQuery với schema hợp lý.

b. Viết lệnh truy vấn cho các câu hỏi sau:

- Tỷ lệ chiến thắng (win) ở các level 1,5,10 của toàn bộ user?
- Tỷ lệ sử dụng skill trung bình trong 1 ván chơi của những user ở Brazil?
- Tỷ lệ user còn ở lại chơi game qua từng level? (note: giả sử rằng số lượng user chơi các level sau sẽ ít hơn level trước, vì vậy câu hỏi yêu cầu kết quả chính xác rằng sau mỗi level, số user còn lại là bao nhiêu).

Chú ý:

- Ứng viên được tùy ý thiết kế thêm các bảng, view, ... phụ trợ khác.
- Ứng viên cung cấp lại toàn bộ schema của các table.
- Ứng viên cung cấp lại các script SQL, các script phụ trợ khác.

- Ngoài SQL để truy vấn bên trong BigQuery, đối với các chức năng phụ trợ khác, ứng viên được tùy ý chọn stack và ngôn ngữ.
- Nếu giải thích được lý do lựa chọn stack, ngôn ngữ và những quyết định xung quanh giải pháp thì sẽ được đánh giá cao hơn.

Giải nghĩa các thông tin:

- event_date: ngày event được log.
- event_timestamp: thời điểm cụ thể của event được log.
- event_name: tên của event.
- event_params: các cặp key-value biểu thị dữ liệu của event.
- geo: chứa thông tin về địa lý.

Các event tiêu biểu:

- level_start: được log lúc user bắt đầu chơi 1 level, với các params sau:
 - + level: level của ván chơi.
- level_finish: được log lúc user kết thúc 1 level, với các params sau:
 - + level: level hiện tại.
 - + duration: thời gian đã chơi, tính bằng giây.
 - + result: kết quả ván chơi.
- use_skill: được log lúc user sử dụng 1 skill, với các params sau:
 - + level: level hiện tại.
 - + name: tên của skill.