**Câu hỏi quyết định:** Chính sách mới (**NEW**) có **tăng hiệu quả ghép–phục vụ** so với chính sách hiện hành (**CONTROL**) không?

Dưới đây là mô tả **chi tiết cấu trúc dữ liệu** cho **Switchback (Marketplace / Network effects)**, bám đúng file switchback\_data.csv và kèm các trường dẫn xuất khuyến nghị để phân tích/triển khai.

* **Hạt dữ liệu & phạm vi**
* **Đơn vị quan sát (granularity):** theo **giờ**
* **Khoảng thời gian mẫu:** 28 ngày × 24 giờ = **672 bản ghi**
* **Thiết kế switchback:** chính sách **NEW** và **CONTROL** **luân phiên theo ngày** (mỗi ngày 1 policy)
* **Data dictionary (file switchback\_data.csv)**

Mỗi dòng là 1 giờ trong ngày.

| **Trường** | **Kiểu CSV → pandas** | **Miền giá trị** | **Ý nghĩa / Cách tính** |
| --- | --- | --- | --- |
| timestamp | string → datetime64[ns] | ISO-8601, ví dụ 2025-01-01 00:00:00 | Mốc thời gian theo **giờ địa phương**. Khuyến nghị lưu theo UTC+7 hoặc có thêm cột timezone nếu đẩy sang hệ thống khác |
| policy | string / categorical | NEW | CONTROL | Chính sách áp dụng trong **cả ngày** chứa giờ này |
| dow | int64 | 0–6 (Mon=0 … Sun=6) | Thứ trong tuần (để kiểm soát mùa vụ theo ngày) |
| hour | int64 | 0–23 | Giờ trong ngày |
| requests | int64 | ≥ 0 | Số **yêu cầu/phát sinh** trong giờ |
| conversions | int64 | 0…requests | Số **ghép thành công/đơn hoàn tất** trong giờ |
| gmv | float64 | ≥ 0 | **Gross Merchandise Value** (đơn vị tiền tệ) của các conversions trong giờ (chỉ số tham khảo ngoài CTR) |
| supply | float64 | ≥ 0 | Mức cung khả dụng (ví dụ: số driver/điểm cung). Nếu là đếm, bạn có thể ép về int; bản mẫu cho phép số thực (xấp xỉ) |
| congestion | float64 | ≥ 0 (hoặc NaN nếu supply=0) | Mức nghẽn: requests / supply. Khi supply=0, nên đặt NaN để tránh vô cực |
| match\_quality | float64 | [0, 1] | Proxy/chỉ báo nội bộ về chất lượng ghép trong giờ |
| ctr | float64 | [0, 1] | KPI chính: conversions / requests. Với requests=0, bản mẫu đặt **0**; thực tế có thể để **NaN** để tránh bias |

| **Trường** | **Kiểu** | **Công thức/Ý nghĩa** |
| --- | --- | --- |
| date | date | CAST(timestamp AS DATE) |
| is\_weekend | bit/bool | dow IN (5,6) |
| day\_index | int | **Số thứ tự ngày** trong chuỗi (1…N). Dùng cho **cluster-robust SE** theo ngày |
| switch\_block\_id | int | ID khối switchback (mỗi ngày = 1 block). Có thể trùng day\_index |
| policy\_new | 0/1 | policy='NEW' ? 1 : 0 |
| ctr\_naive | float | conversions/requests nhưng **đặt NaN** khi requests=0 (để phân tích nhạy cảm) |
| high\_congestion | 0/1 | congestion >= ngưỡng (ví dụ percentile 70–80) để kiểm tra **hiệu ứng khác biệt** |
| hour\_fe, dow\_fe | one-hot | Dummies để dùng trong hồi quy có **hiệu ứng cố định** theo giờ/thứ |

* **Ràng buộc & kiểm tra chất lượng (data quality)**
* conversions ≤ requests và requests ≥ 0 (CHECK)
* 0 ≤ ctr ≤ 1, 0 ≤ match\_quality ≤ 1
* supply ≥ 0. Nếu có giờ supply=0, **đặt congestion = NULL/NaN**
* Cân bằng thiết kế: mỗi **thứ trong tuần** xuất hiện cả ở NEW lẫn CONTROL (nếu lệch → cảnh báo)
* Không có giờ trùng timestamp (chìa khóa tự nhiên: timestamp)