Supercell Puzzle Arena
Activation, Retention & Monetization

Báo cáo phân tích dữ liệu người dùng mới, retention, chất lượng phiên bản và hiệu quả thu hút người dùng cho một game mobile giả định.

THỰC HIỆN	KỲ DỮ LIỆU	CÔNG CỤ	QUY MÔ
Nguyễn Hữu Triệu	01/04 – 30/06/2026 (UTC)	BigQuery (SQL), Power BI	9 bảng · 701,652 dòng

Tóm tắt

Puzzle Arena là game hư cấu trong bối cảnh giả định của Supercell; toàn bộ dữ liệu là dữ liệu giả lập phục vụ học tập. Báo cáo trả lời 9 câu hỏi cho Product Manager, Game Designer và User Acquisition về 15,000 user mới cài game trong 91 ngày. Mỗi user được theo dõi tối đa 28 ngày sau khi cài.

User mới	Doanh thu quan sát	D7 / D28 retention	D7 / D28 ROAS
15,000	**$19,315**	**24.4% / 7.4%**	**65.5% / 127.7%**
install trong kỳ	IAP net 78.4% · quảng cáo 21.6%	exact-day, cohort đủ tuổi	user có campaign

Phát hiện chính

- **Tăng trưởng đến từ engagement, không từ acquisition:** user quay lại tăng từ 511 lên 867/ngày (T4 → T5), install mới đi ngang ~165/ngày.
- **Điểm mất user duy nhất trong ngày đầu là bước hoàn thành tutorial:** 30.7% user (4,608) bắt đầu nhưng không hoàn thành; tỷ lệ gần như như nhau ở mọi device tier, platform, version và kênh.
- **Onboarding mới (ONBOARDING_01) tốt hơn rõ rệt:** tutorial D0 80.5% vs 66.6%, D7 retention +6.02 pp (CI 95% +3.0 → +9.0 pp), không guardrail nào xấu đi.
- **Crash tăng vọt do v1.1.0 lỗi trên máy low-tier:** 18.1% so với ~1.4–1.7% ở các tổ hợp khác; không phải do cơ cấu thiết bị. v1.2.0 đã sửa.
- **Level 7 khó bất thường:** 53.0% lượt chơi thất bại, gấp ~3 lần các level khác; chỉ 58.9% user chơi level 7 vượt qua được.
- **CPI thấp chưa chắc tốt:** GC09, GC12 (Search_Ads US/JP) giữ chân tốt nhất nhưng là 2 campaign duy nhất có D28 ROAS dưới 100% (82%, 86%) và chiếm 39.9% spend.

Đề xuất

- Roll out onboarding mới cho 100% user mới.
- Thêm bước kiểm thử trên máy low-tier và phát hành theo từng đợt, theo dõi crash theo device tier.
- Giảm dần ngân sách GC09/GC12, chuyển sang campaign US/JP có D28 ROAS cao hơn; đánh giá campaign bằng ROAS cohort.
- Thử nghiệm tiếp theo: **LEVEL7_TUNING_01** – giảm độ khó level 7, metric chính là % user vượt level 7 trong 7 ngày.

1. Bài toán và phạm vi

Studio cần đánh giá trải nghiệm người dùng mới, lỗi theo phiên bản, điểm nghẽn progression và hiệu quả thu hút người dùng. Một thử nghiệm onboarding (ONBOARDING_01) đã được phân nhóm ngẫu nhiên cho user mới từ 01/05 đến trước 21/05. Dashboard là phương tiện trình bày; mục tiêu là hỗ trợ 4 quyết định dưới đây.

Quyết định	Người quyết định	Câu hỏi	Kết quả phân tích đổi quyết định thế nào
Phân bổ lại ngân sách UA giữa 12 campaign	User Acquisition	Q3, Q7, Q8	Campaign có ROAS cohort thấp hơn mặt bằng là ứng viên giảm ngân sách; campaign CPI cao nhưng ROAS cohort tốt có thể giữ hoặc tăng.
Roll out onboarding mới của ONBOARDING_01	PM, Game Designer	Q2, Q6	Roll out khi không có SRM, tutorial completion và D7 của variant cao hơn control với CI 95% không chứa 0, và guardrail không xấu đi.
Thứ tự ưu tiên sửa lỗi theo phiên bản và thiết bị	Product Manager	Q4	Crash vẫn cao khi so trong cùng device tier/platform → hotfix phiên bản; nếu chênh lệch do cơ cấu user → tối ưu cho phân khúc thiết bị.
Điều chỉnh độ khó level	Game Designer	Q5	Level có fail rate theo lượt chơi cao và tỷ lệ user vượt qua thấp là ứng viên giảm độ khó hoặc thêm gợi ý.

#	Câu hỏi phân tích	Trang dashboard
Q1	DAU/WAU/MAU và số install thay đổi ra sao? Tăng trưởng đến từ acquisition hay engagement?	Overview
Q2	Funnel first_open → tutorial_begin → tutorial_complete → level_start mất user ở bước nào?	Onboarding
Q3	Retention D1/D7/D14/D28 khác nhau theo install cohort, country, platform, campaign thế nào?	Retention & UA
Q4	Phiên bản và loại thiết bị nào có crash rate cao; có phải do cơ cấu user không?	Release & A/B
Q5	Level nào có fail rate cao? Tách thất bại theo lượt chơi và tỷ lệ user vượt qua.	Onboarding
Q6	A/B ONBOARDING_01 có phân bổ hợp lý không? Variant thay đổi tutorial, D7 và guardrail thế nào?	Release & A/B
Q7	Campaign nào có CPI thấp nhưng retention hoặc revenue/user thấp (so cohort cùng tuổi)?	Retention & UA
Q8	D7/D28 ROAS có ghép đúng doanh thu cohort với spend của ngày install không?	Retention & UA
Q9	Đề xuất 3 action cho Product/UA và một thiết kế thử nghiệm tiếp theo.	Action

2. Dữ liệu và cách xử lý

Bảng	Số dòng	Grain
`users`	15,000	1 user; install time, country, platform, device tier, campaign first-touch, install version
`sessions`	81,634	1 phiên chơi; app_version của phiên, crashed
`events`	468,105	1 bản ghi ingest; event_id có thể lặp do gửi lại
`purchases`	2,708	1 giao dịch IAP; refund tích lũy đến snapshot
`ad_impressions`	129,758	1 lượt quảng cáo có doanh thu
`campaign_daily`	1,092	1 campaign × 1 ngày; spend, attributed installs
`ab_assignments`	3,340	Tối đa 1 phân nhóm ONBOARDING_01 mỗi user
`campaigns`	12	1 campaign; ad network, country, platform
`app_versions`	3	1 phiên bản; 1.0.0 (01/04), 1.1.0 (12/05), 1.2.0 (05/06)

Quy trình

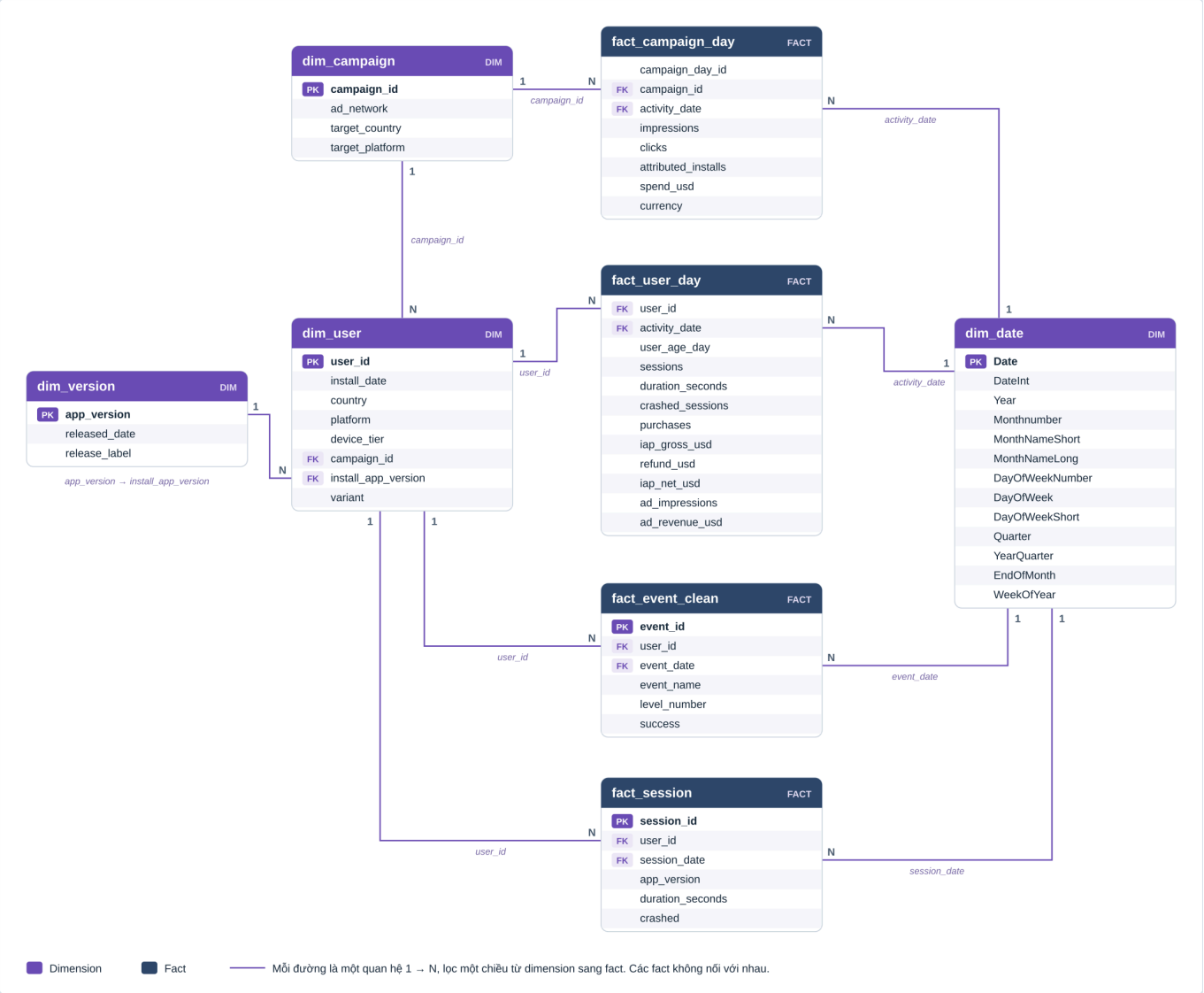
9 file CSV được nạp vào một dataset BigQuery. Trước khi dựng model, dữ liệu được kiểm tra bằng SQL theo 7 nhóm: số dòng và khóa chính, giá trị thiếu, khóa ngoại mồ côi, phạm vi giá trị, thứ tự thời gian và số dòng trước/sau join (Phụ lục B). Mỗi bảng của model Power BI được tạo bằng một câu SQL riêng (import, Phụ lục A), sau đó tạo quan hệ trong Power BI.

Quy tắc làm sạch

- **Event lặp:** dedup theo `event_id`; 250 bản ghi gửi lại bị loại (468,105 → 467,855 event).
- **Organic:** `campaign_id` rỗng (4,516 user) là organic có chủ đích, gán `campaign_id = 'ORGANIC'` và thêm một dòng tương ứng vào `dim_campaign`.
- **Phiên bản:** crash tính theo `app_version` của từng phiên, không dùng install version (22% phiên chạy version khác version lúc cài).
- **Cohort chưa đủ tuổi:** retention DN và ROAS DN chỉ tính trên cohort đã quan sát đủ N ngày trước snapshot 01/07.
- **Doanh thu:** IAP net = gross − refund (73 giao dịch, $445 hoàn); chưa trừ phí store/thuế nên không gọi là lợi nhuận.
- **Kỳ quan sát:** đã kiểm tra mọi timestamp nằm trong 01/04 – 30/06/2026; không có bản ghi từ snapshot trở đi.

Mô hình dữ liệu

Star schema với 4 dimension và 4 fact. Quan hệ một chiều từ dimension sang fact; các fact không nối trực tiếp với nhau. ROAS ghép doanh thu ở `fact_user_day` với spend ở `fact_campaign_day` qua `dim_campaign`. Measure theo cohort dùng `TREATAS` để coi ngày đang chọn là ngày install.



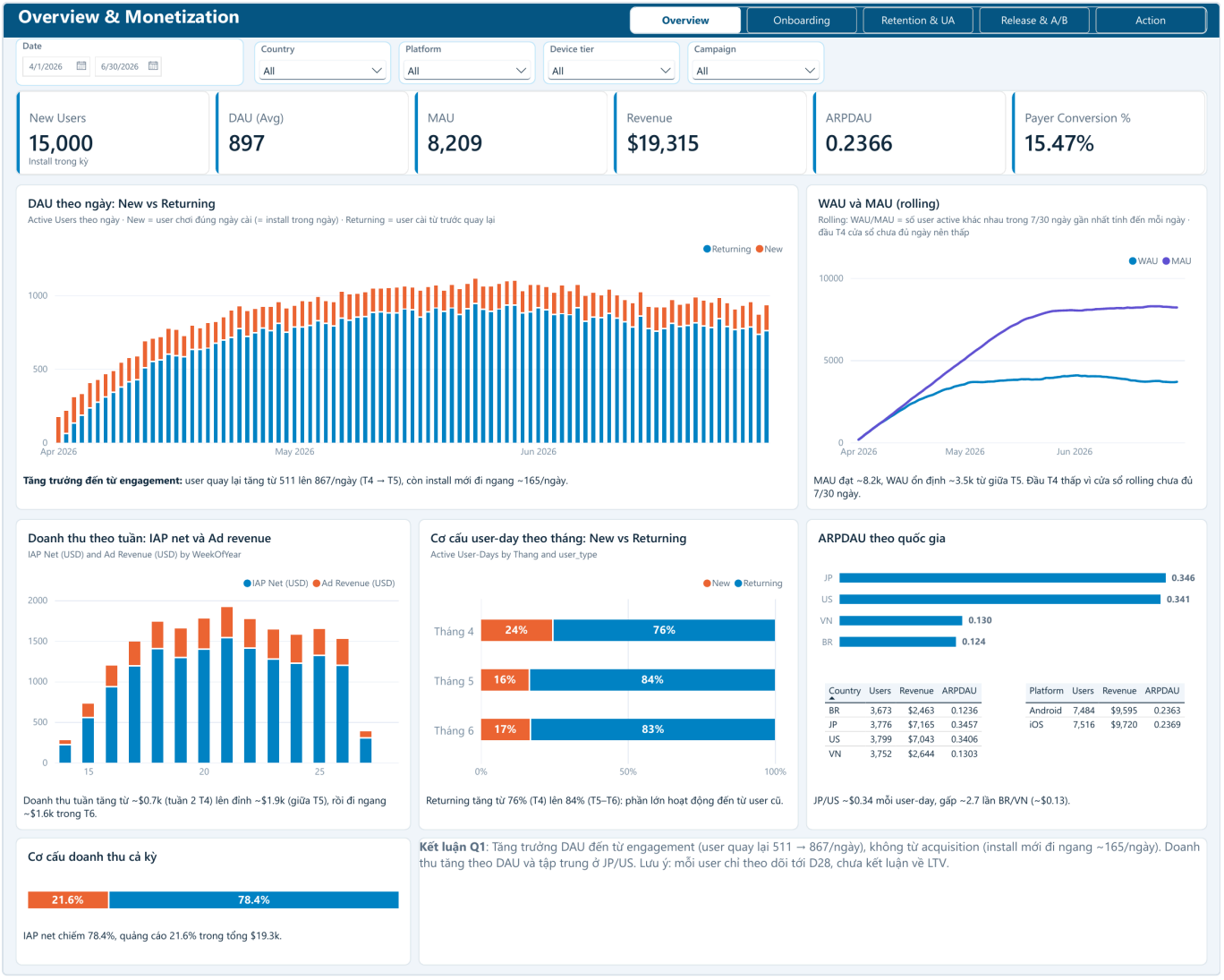
Hình 1. Model phân tích trong Power BI.

Định nghĩa KPI chính

KPI	Cách tính	Lưu ý
DAU / WAU / MAU	Số user khác nhau có session trong ngày / 7 / 30 ngày gần nhất	Không cộng dồn DAU; đầu T4 cửa sổ rolling chưa đủ ngày
Tutorial completion D0	User có tutorial_complete trong ngày cài / user có first_open	Đếm user, không đếm event retry
D1/D7/D14/D28 retention	User có session đúng ngày N sau install / user đã đủ N ngày	Exact-day, không phải rolling
Crash session rate	Session crashed / tổng session	Theo version của phiên
Level fail rate	level_end thất bại / tổng level_end	Theo lượt chơi; một user có thể thử nhiều lần
ARPDAU	(IAP net + ad revenue) / DAU	USD mỗi user-day
CPI	Spend / attributed installs	Organic không có CPI
D7/D28 ROAS	Doanh thu tuổi 0..N của cohort / spend để có đúng cohort đó	Không chia doanh thu tháng cho spend tháng
A/B D7 uplift	D7 retention variant − control, tính ITT	Đơn vị pp (điểm %); kèm CI 95%, sample size, SRM

3. Kết quả

3.1 Q1 – Tăng trưởng đến từ đâu?



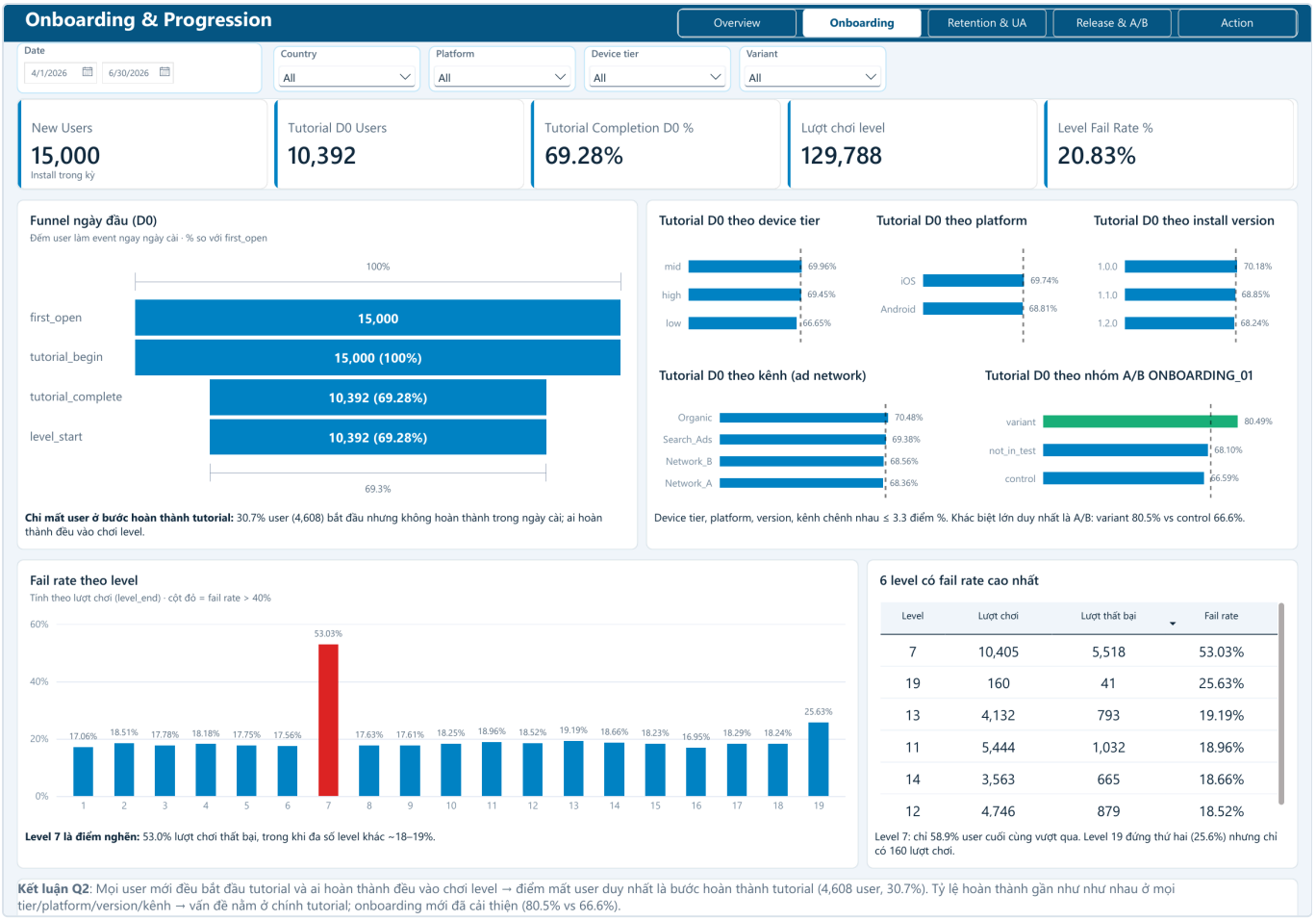
Hình 2. Trang Overview.

Q1 – Tăng trưởng và doanh thu

- User quay lại tăng từ 511 lên 867/ngày (T4 → T5), trong khi install mới đi ngang ~165/ngày. Tỷ trọng returning trong user-day tăng từ 76% (T4) lên 84% (T5–T6).
- MAU đạt ~8.2k, WAU ổn định ~3.5k từ giữa T5. DAU trung bình 897.
- Doanh thu tuần tăng từ ~$0.7k lên đỉnh ~$1.9k (giữa T5), rồi đi ngang ~$1.6k trong T6. IAP net chiếm 78.4%, quảng cáo 21.6% trong tổng $19.3k.
- ARPDAU ở JP/US ~$0.34, gấp ~2.7 lần BR/VN (~$0.13); Android và iOS gần như bằng nhau.

Kết luận: tăng trưởng DAU đến từ engagement, không từ acquisition. Mỗi user chỉ được theo dõi tới D28 nên chưa kết luận về LTV.

3.2 Q2 & Q5 – Onboarding và progression



Hình 3. Trang Onboarding & Progression.

Q2 – Funnel ngày đầu

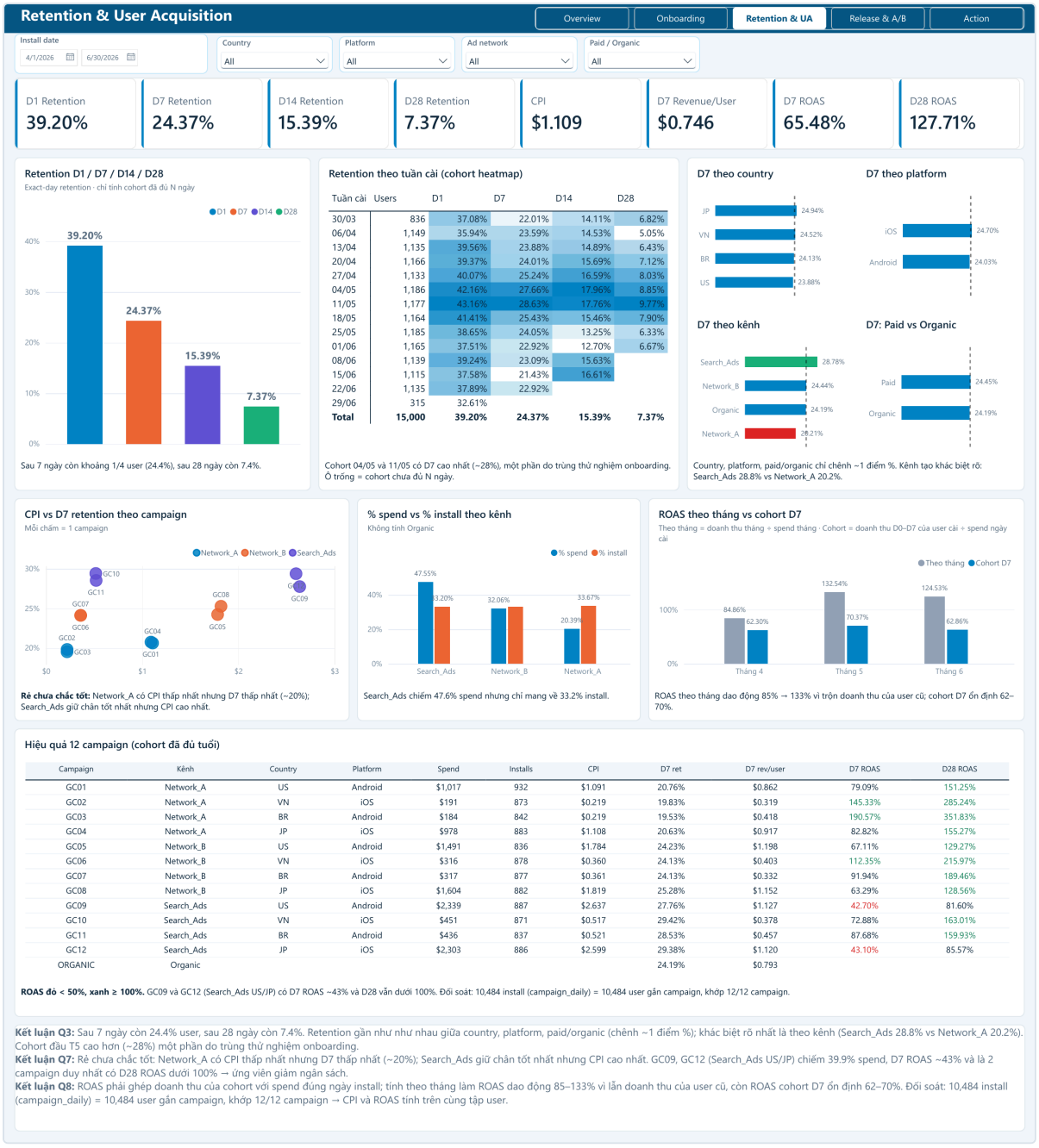
- 15,000 first_open → 15,000 tutorial_begin → 10,392 tutorial_complete (69.3%) → 10,392 level_start.
- Mất user chỉ ở bước hoàn thành tutorial: 4,608 user (30.7%). Ai hoàn thành đều vào chơi level.
- Device tier, platform, install version, kênh chênh nhau ≤ 3.3 điểm % → vấn đề nằm ở chính tutorial.
- Khác biệt lớn duy nhất là nhóm A/B: variant 80.5% vs control 66.6%.

Q5 – Fail rate theo level

- Level 7: 53.0% lượt chơi thất bại (5,518 / 10,405), trong khi đa số level khác ~17–19%.
- Theo user: chỉ 58.9% user đã chơi level 7 vượt qua được; các level khác ~83–92%.
- Level 19 đứng thứ hai (25.6%) nhưng chỉ có 160 lượt chơi, chưa đủ để kết luận.

Kết luận: level 7 là ứng viên điều chỉnh độ khó.

3.3 Q3, Q7 & Q8 – Retention và User Acquisition



Hình 4. Trang Retention & User Acquisition.

Q3 – Retention

- D1 39.2%, D7 24.4%, D14 15.4%, D28 7.4%: sau 7 ngày còn khoảng 1/4 user.
- Country, platform, paid/organic chỉ chênh ~1 điểm % ở D7. Kênh tạo khác biệt rõ: Search_Ads 28.8% vs Network_A 20.2%.
- Cohort 04/05 và 11/05 có D7 cao nhất (~28%), một phần do trùng thời gian chạy ONBOARDING_01.

Q7 – CPI và chất lượng user

- Network_A có CPI thấp nhất ($0.22–1.11) nhưng D7 thấp nhất (~20%). Search_Ads giữ chân tốt nhất (D7 ~28–29%) nhưng CPI cao nhất ($0.52–2.64).
- Search_Ads chiếm 47.6% spend nhưng chỉ mang về 33.2% install; Network_A ngược lại (20.4% spend, 33.7% install).
- GC09, GC12 (Search_Ads US/JP): D7 revenue/user ~$1.12–1.13, không cao hơn Network_B cùng thị trường ($1.15–1.20), nhưng CPI ~$2.6 nên D7 ROAS chỉ ~43%.

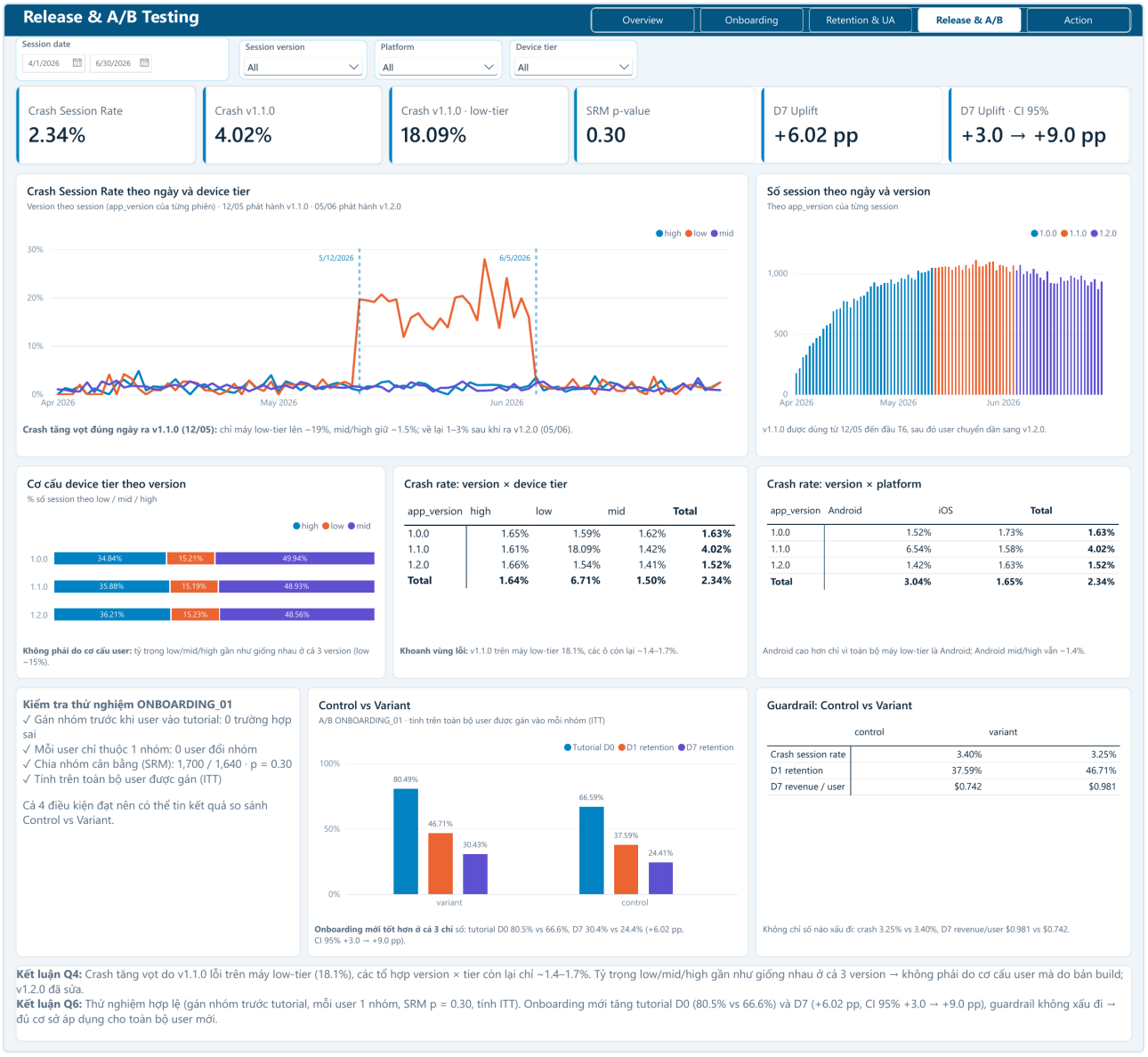
Q8 – ROAS cohort

- ROAS tính theo tháng (doanh thu tháng ÷ spend tháng) dao động 85% → 133% vì trộn doanh thu của user cũ. ROAS cohort D7 ổn định 62–70%.
- Toàn bộ user có campaign: D7 ROAS 65.5%, D28 ROAS 127.7%.
- Đối soát: 10,484 install trong `campaign_daily` = 10,484 user gắn campaign, khớp 12/12 campaign → CPI và ROAS tính trên cùng tập user.

Campaign	Kênh	Thị trường	Spend	CPI	D7 ret	D7 ROAS	D28 ROAS
GC01	Network_A	US Android	$1,017	$1.09	20.8%	79.1%	151.3%
GC04	Network_A	JP iOS	$978	$1.11	20.6%	82.8%	155.3%
GC05	Network_B	US Android	$1,491	$1.78	24.2%	67.1%	129.3%
GC08	Network_B	JP iOS	$1,604	$1.82	25.3%	63.3%	128.6%
GC09	Search_Ads	US Android	$2,339	$2.64	27.8%	42.7%	81.6%
GC12	Search_Ads	JP iOS	$2,303	$2.60	29.4%	43.1%	85.6%

Bảng 1. Các campaign US/JP (6/12 campaign). D7 ret, D7 ROAS tính trên cohort đủ 7 ngày; D28 ROAS trên cohort đủ 28 ngày. Bảng đầy đủ 12 campaign có trong Hình 4.

3.4 Q4 & Q6 – Chất lượng phiên bản và A/B ONBOARDING_01



Hình 5. Trang Release & A/B Testing.

Q4 – Crash theo phiên bản và thiết bị

- Crash session rate cả kỳ 2.34%. Crash tăng vọt đúng ngày ra v1.1.0 (12/05): chỉ máy low-tier lên ~19%, mid/high giữ ~1.5%; về lại 1–3% sau khi ra v1.2.0 (05/06).
- Khoanh vùng: v1.1.0 trên máy low-tier 18.1%; mọi tổ hợp version × tier còn lại ~1.4–1.7%.
- Không phải do cơ cấu user: tỷ trọng low/mid/high gần như giống nhau ở cả 3 version (low ~15%).
- Android cao hơn (6.5% ở v1.1.0) chỉ vì toàn bộ máy low-tier là Android; Android mid/high vẫn ~1.4%.

Q6 – A/B ONBOARDING_01

- Kiểm tra hợp lệ: gán nhóm trước khi user vào tutorial (0 trường hợp sai), mỗi user chỉ 1 nhóm, SRM 1,700 / 1,640 với p = 0.30, tính trên toàn bộ user được gán (ITT).

Chỉ số	Control	Variant	Chênh lệch
Tutorial completion D0	66.6%	80.5%	**+13.9 pp**
D1 retention	37.6%	46.7%	**+9.1 pp**
D7 retention	24.4%	30.4%	**+6.02 pp (CI 95% +3.0 → +9.0 pp)**
Crash session rate (guardrail)	3.40%	3.25%	không xấu đi
D7 revenue/user (guardrail)	$0.742	$0.981	không xấu đi

Kết luận: thử nghiệm hợp lệ; onboarding mới tăng tutorial completion và D7 retention, guardrail không xấu đi → đủ cơ sở áp dụng.

4. Đề xuất hành động (Q9)

Mỗi đề xuất viết theo khung **Số liệu → Vấn đề → Action**, kèm đối tượng và người phụ trách. Số liệu ghi rõ kỳ và mẫu số.

1 · Roll out onboarding mới cho 100% user mới

Số liệu

- 30.7% user mới (4,608 / 15,000, install 01/04–30/06) không hoàn thành tutorial ngày đầu.
- A/B (cohort 01/05–20/05): tutorial D0 variant 80.5% (n = 1,640) vs control 66.6% (n = 1,700); D7 +6.02 pp (CI 95% +3.0 → +9.0 pp); crash và D7 revenue/user không xấu đi.

Vấn đề

Gần 1/3 user mới rời game ngay ở bước tutorial.

Action

Áp dụng onboarding của nhóm variant cho 100% user mới; theo dõi tutorial D0 và D7 của các cohort sau roll-out so với 80.5% và 30.4%.

Đối tượng: toàn bộ user mới · **Người phụ trách:** Product Manager, Game Designer

2 · Kiểm thử máy low-tier trước khi phát hành

Số liệu

- Crash v1.1.0 trên máy low-tier 18.1% (699 / 3,865 session, 12/05–04/06), so với v1.0.0 cùng tier 1.6% (75 / 4,731).
- Các tổ hợp version × tier khác ~1.4–1.7%; cơ cấu thiết bị gần như giống nhau giữa 3 version.

Vấn đề

Bản build lỗi trên máy yếu, kéo dài 24 ngày mới có bản sửa (v1.2.0, 05/06).

Action

Thêm test trên máy low-tier vào checklist phát hành; phát hành theo từng đợt và theo dõi crash session rate theo version × device tier trong những ngày đầu.

Đối tượng: user Android máy low-tier (2,258 user) · **Người phụ trách:** Product Manager, QA

3 · Giảm ngân sách GC09, GC12 (Search_Ads US/JP)

Số liệu

- Chiếm 39.9% spend ($4,642 / $11,628, 01/04–30/06), CPI ~$2.6.
- D28 ROAS 82% và 86% (cohort đủ 28 ngày); các campaign US/JP khác đạt 129–155% (GC01, GC04 Network_A ~151–155%; GC05, GC08 Network_B ~129%).

Vấn đề

Gần 40% ngân sách đổ vào 2 campaign chưa hoàn vốn sau 28 ngày.

Action

Giảm dần ngân sách GC09/GC12, chuyển sang GC01, GC04, GC05, GC08; tăng từng bước vì CPI có thể tăng khi tăng chi, theo dõi CPI và ROAS theo cohort sau mỗi lần điều chỉnh.

Đối tượng: campaign Search_Ads US/JP · **Người phụ trách:** User Acquisition

Thử nghiệm tiếp theo · LEVEL7_TUNING_01

Số liệu

53.0% lượt chơi level 7 thất bại (5,518 / 10,405 lượt, 01/04–30/06), các level khác 17–19%; chỉ 58.9% user đã chơi level 7 vượt qua (4,026 / 6,841).

Vấn đề

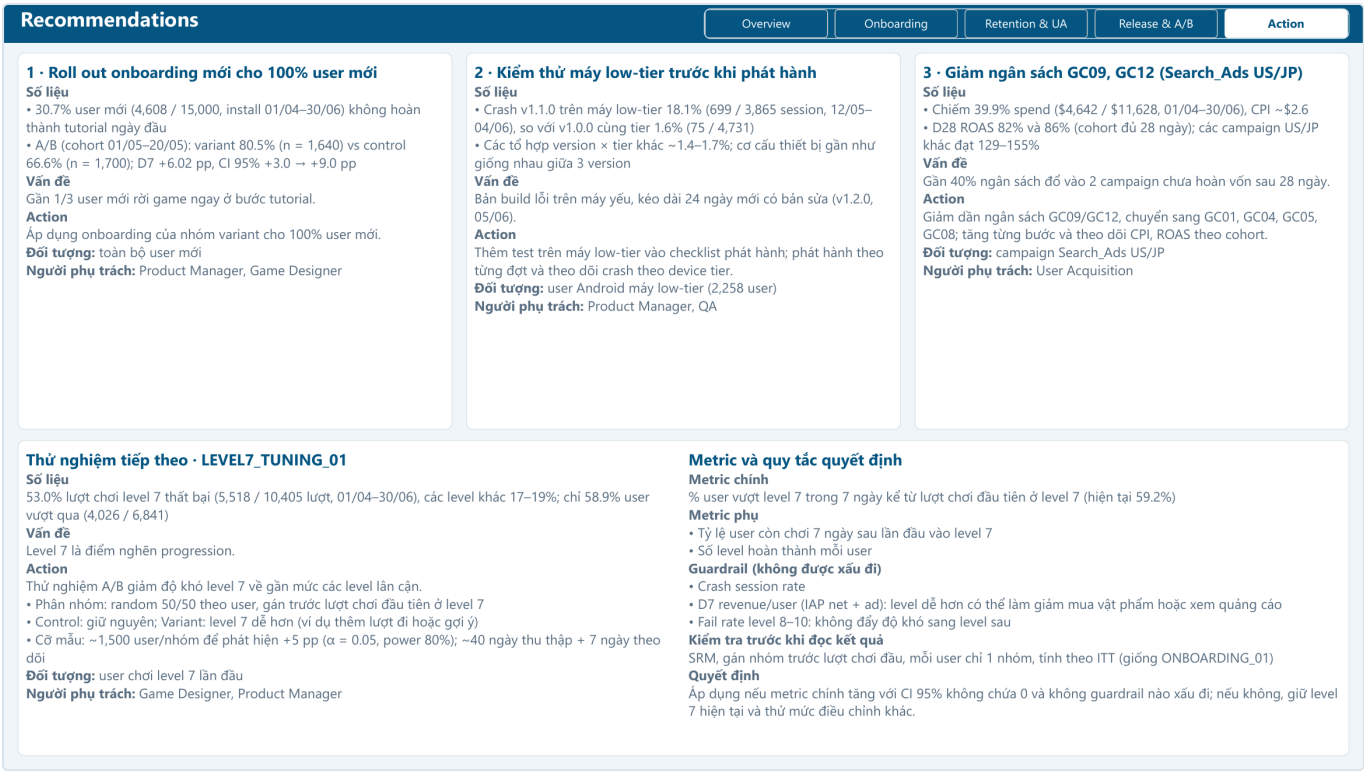
Level 7 là điểm nghẽn progression.

Action

Thử nghiệm A/B giảm độ khó level 7. Giả thuyết: đưa độ khó về gần mức các level lân cận sẽ tăng tỷ lệ user vượt qua mà không làm giảm doanh thu.

Đối tượng: user chơi level 7 lần đầu · **Người phụ trách:** Game Designer, Product Manager

Hạng mục	Thiết kế
Đối tượng & phân nhóm	User chơi level 7 lần đầu sau ngày bắt đầu; random 50/50 theo user, gán trước lượt chơi đầu tiên ở level 7.
Control / Variant	Giữ nguyên level 7 / level 7 dễ hơn (ví dụ thêm lượt đi hoặc gợi ý).
Metric chính	% user vượt level 7 trong 7 ngày kể từ lượt chơi đầu tiên ở level 7 (hiện tại 59.2%).
Metric phụ	Tỷ lệ user còn chơi 7 ngày sau lần đầu vào level 7; số level hoàn thành mỗi user.
Guardrail	Crash session rate; D7 revenue/user (IAP net + ad), vì level dễ hơn có thể làm giảm mua vật phẩm hoặc xem quảng cáo; fail rate level 8–10.
Cỡ mẫu & thời gian	~1,500 user/nhóm để phát hiện mức tăng 5 điểm % từ 59% (α = 0.05 hai phía, power 80%). Khoảng 75 user/ngày chơi level 7 lần đầu → ~40 ngày thu thập + 7 ngày theo dõi.
Kiểm tra & quyết định	SRM, gán nhóm trước lượt chơi đầu, mỗi user 1 nhóm, ITT. Áp dụng nếu metric chính tăng với CI 95% không chứa 0 và không guardrail nào xấu đi.



Hình 6. Trang Recommendations (tab Action) trong dashboard.

5. Hạn chế

- **Dữ liệu mô phỏng:** các con số minh họa cách phân tích, không phản ánh sản phẩm hay dữ liệu thật của Supercell.
- **Cửa sổ quan sát D28:** mỗi user chỉ được theo dõi tối đa 28 ngày nên doanh thu là observed revenue, không phải lifetime LTV; ROAS D28 chỉ tính trên cohort đủ 28 ngày.
- **Doanh thu chưa trừ phí:** IAP net = gross − refund, chưa trừ phí store và thuế; không đủ để tính lợi nhuận.
- **So sánh kênh không phải thử nghiệm:** campaign và country không được phân ngẫu nhiên, nên chênh lệch giữa kênh là tương quan. Việc chuyển ngân sách nên làm từng bước và đo lại.
- **Thứ tự level:** dữ liệu level không đi tuần tự (số user chơi level 8 lớn hơn số user vượt level 7), nên báo cáo không kết luận level 7 làm user rời game; chỉ kết luận level 7 khó bất thường.
- **Cohort trùng thử nghiệm:** cohort đầu T5 có D7 cao hơn một phần do trùng ONBOARDING_01, cần tách khi so sánh retention theo thời gian.

Phụ lục A. SQL tạo bảng model (BigQuery)

Mỗi khối là một query import vào Power BI. `dim_date` tạo bằng DAX trong Power BI.

```
-- dim_user
select u.user_id
  , date(u.installed_at) as install_date
  , u.country
  , u.platform
  , u.device_tier
  , coalesce(u.campaign_id, 'ORGANIC') as campaign_id
  , u.install_app_version
  , coalesce(a.variant, 'not_in_test') as variant
from `jda-k1.Trieu_BigProject1.users` u
left join `jda-k1.Trieu_BigProject1.ab_assignments` a
on u.user_id = a.user_id


-- dim_campaign
select campaign_id
  , ad_network
  , target_country
  , target_platform
from `jda-k1.Trieu_BigProject1.campaigns`
union all
select 'ORGANIC'
  , 'Organic'
  , null
  , null


-- dim_version
select app_version
  , released_date
  , release_label
from `jda-k1.Trieu_BigProject1.app_versions`


-- fact_campaign_day
select campaign_day_id
  , campaign_id
  , activity_date
  , impressions
  , clicks
  , attributed_installs
  , spend_usd
  , currency
from `jda-k1.Trieu_BigProject1.campaign_daily`


-- fact_user_day
with s as (
  select user_id
    , date(started_at) as activity_date
    , count(*) as sessions
    , sum(duration_seconds) as duration_seconds
    , sum(cast(crashed as int64)) as crashed_sessions
  from `jda-k1.Trieu_BigProject1.sessions`
  group by 1, 2
)
, p as (
  select user_id
    , date(purchased_at) as activity_date
    , count(*) as purchases
    , sum(gross_usd) as iap_gross_usd
    , sum(refund_usd) as refund_usd
    , sum(gross_usd - refund_usd) as iap_net_usd
  from `jda-k1.Trieu_BigProject1.purchases`
  group by 1, 2
)
, a as (
  select user_id
    , date(impression_at) as activity_date
    , count(*) as ad_impressions
    , sum(revenue_usd) as ad_revenue_usd
  from `jda-k1.Trieu_BigProject1.ad_impressions`
  group by 1, 2
)
select s.user_id
  , s.activity_date
  , date_diff(s.activity_date, date(u.installed_at),
day) as user_age_day
  , case when s.activity_date = date(u.installed_at)
then 'New' else 'Returning' end as user_type
  , s.sessions
  , s.duration_seconds
  , s.crashed_sessions
  , coalesce(p.purchases, 0) as purchases
  , coalesce(p.iap_gross_usd, 0) as iap_gross_usd
  , coalesce(p.refund_usd, 0) as refund_usd
  , coalesce(p.iap_net_usd, 0) as iap_net_usd
  , coalesce(a.ad_impressions, 0) as ad_impressions
  , coalesce(a.ad_revenue_usd, 0) as ad_revenue_usd
from s
join `jda-k1.Trieu_BigProject1.users` u on s.user_id
= u.user_id
left join p on s.user_id = p.user_id and
s.activity_date = p.activity_date
left join a on s.user_id = a.user_id and
s.activity_date = a.activity_date


-- fact_event_clean
select event_id
  , user_id
  , date(event_at) as event_date
  , event_name
  , cast(json_value(event_params_json,
'$.level_number') as int64) as level_number
  , cast(json_value(event_params_json, '$.success')
as int64) as success
from (
  select *
    , row_number() over (partition by event_id order
by event_row_id) as rn
  from `jda-k1.Trieu_BigProject1.events`
)
where rn = 1

-- fact_session
select session_id
  , user_id
  , date(started_at) as session_date
  , app_version
  , duration_seconds
  , cast(crashed as int64) as crashed
from `jda-k1.Trieu_BigProject1.sessions`
```

Phụ lục B. Báo cáo chất lượng dữ liệu

Kết quả từ bước kiểm tra bằng SQL trước khi dựng model (số dòng và khóa, giá trị thiếu, khóa ngoại, phạm vi giá trị, thứ tự thời gian, số dòng trước/sau join). Số liệu kỳ 01/04 – 30/06/2026, snapshot 01/07.

Issue	Tác động nếu không xử lý	Xử lý	Số dòng ảnh hưởng
`event_id` lặp do gửi lại (retry)	Funnel, số lượt chơi và fail rate bị đếm dư	Dedup theo `event_id`, giữ bản ghi có `event_row_id` nhỏ nhất	250 / 468,105 bản ghi events
`campaign_id` rỗng	Join với campaign làm mất user organic; dễ hiểu nhầm là thiếu attribution	Gán `'ORGANIC'`, thêm dòng tương ứng vào `dim_campaign`; organic không tính CPI	4,516 / 15,000 user
`refunded_at` rỗng	Không phải lỗi: rỗng khi giao dịch không bị hoàn	Giữ nguyên null, không thay bằng 0	2,635 / 2,708 giao dịch
Giao dịch bị refund	Doanh thu IAP cao hơn tiền thực nhận	Dùng IAP net = gross − refund (chưa trừ phí store/thuế)	73 giao dịch, $445
Version của phiên khác version lúc install	Crash bị gán sai phiên bản	Tính crash theo `app_version` của từng session	17,996 / 81,634 session (22%)
Join purchases với ad_impressions qua `session_id`	Tiền bị nhân dòng: tổng IAP gross tăng từ $15,596 lên $31,334	Gộp từng nguồn theo user-ngày rồi mới join (`fact_user_day`)	Toàn bộ 2,708 giao dịch IAP nếu join sai
Cohort chưa đủ tuổi tại snapshot	Retention và ROAS bị kéo thấp vì user chưa có đủ N ngày	Chỉ tính DN trên cohort đã đủ N ngày	Chưa đủ D7: 1,101 user; chưa đủ D28: 4,554 user
Doanh thu quảng cáo rất nhỏ mỗi lượt	Làm tròn sớm làm lệch tổng ad revenue	Giữ kiểu DECIMAL 4 chữ số, chỉ làm tròn khi hiển thị	129,758 impression (tổng $4,164.07)

Các kiểm tra không phát hiện lỗi

- Khóa chính duy nhất ở cả 9 bảng; mỗi user tối đa 1 nhóm A/B (3,340 dòng = 3,340 user).
- 0 bản ghi mồ côi ở 12 cặp khóa ngoại (users, sessions, events, purchases, ad_impressions, campaign_daily, ab_assignments).
- 0 lỗi thứ tự thời gian: không có phiên kết thúc trước khi bắt đầu, session/event trước install, refund trước khi mua; mọi mốc thời gian nằm trong kỳ.
- Không có refund lớn hơn gross; giá trị phân loại (country, platform, tier, variant, event_name, SKU, ad_format) đúng danh mục.
- sessions → users giữ nguyên số dòng (81,634 = 81,634).

Ví dụ: kiểm tra số dòng trước và sau join

```
-- join đúng: sessions -> users, số dòng giữ nguyên
select count(*)
from `jda-k1.Trieu_BigProject1.sessions` s
join `jda-k1.Trieu_BigProject1.users` u on s.user_id = u.user_id

-- join sai: purchases với ad_impressions qua session_id, tiền IAP bị nhân
select sum(p.gross_usd)
from `jda-k1.Trieu_BigProject1.purchases` p
left join `jda-k1.Trieu_BigProject1.ad_impressions` a on p.session_id = a.session_id
```