

Data-Quality Impact Report — before vs after automated checks

Project: Automated Data-Quality framework for the **HEXA VPN** app · **Built at:** Hazel Mobile
· **Data shown:** synthetic / illustrative

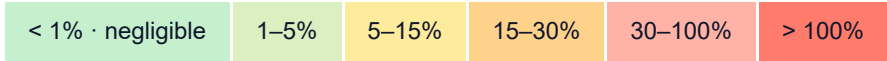
Data-quality framework I built at Hazel Mobile for the HEXA VPN app. The figures below are **synthetic / illustrative** for portfolio use — they show the *method* (an automated data-quality before/after model), **not real client, user, or production data.**

What this shows: an automated data-quality framework that reclassifies telemetry to reveal the *true* health of a metric. The checks **never delete events** — a caught event is **reclassified** success → failure, and users left with no clean success drop out. **Affected users are split three ways:**

- **One and done (Type A)** — a single session ever.
- **One and done (Type B)** — several sessions but active on one calendar day.
- **Retained** — active on 2+ days (the users worth acting on).

Self-validation (self-check statement): all cross-checks PASS.

Heatmap legend (shading = size of the change / contribution)



1 — Headline: before vs after, with the lost-user breakdown

Raw success is **92.00%**; after the checks, true clean success is **68.00%** — a **24-point** drop. The checks leave **60,000 users** with no clean success, of whom **~83% are one-and-done** (34,000 single-session + 16,000 single-day) while only **10,000 were genuinely Retained**.

Metric	Before	After	Change
Events / connections	500,000	500,000	0
Unique users (≥1 clean success)	150,000	90,000	-60,000
Successes	460,000	340,000	-120,000
Failures	40,000	160,000	+120,000
Success %	92.00%	68.00%	-24.00 pts
Failure %	8.00%	32.00%	+24.00 pts

2 — What the checks found, with per-check user split

Row-level data quality is strong — **~96%** of rows pass all row-level checks. The erosion is behavioural, dominated by **inactivity churn (B5)**. The two newest checks behave as designed: **#11 self-replace** flags a small, precise set, and **B6 bot/burst deep-dive** flags only a handful — i.e. **no bots**, just a few borderline accounts.

Tier	ID	Check	Flagged	Users	Type A	Type B	Retained	%	Status
Row-level	#1	Timestamp validity	3,000	1,200	600	300	300	0.60	FLAGGED
Row-level	#2	Valid country code	0	0	0	0	0	0.00	OK
Row-level	#3	Country present	0	0	0	0	0	0.00	OK
Row-level	#4	Non-negative duration	0	0	0	0	0	0.00	OK
Row-level	#5	Non-zero duration	2	0	0	0	0	0.00	FLAGGED
Row-level	#6	Handshake completed	0	0	0	0	0	0.00	OK
Row-level	#7	Valid server IP	5	3	1	1	1	0.00	FLAGGED
Row-level	#8	Collection continuity	3	3	1	0	2	0.00	FLAGGED
Row-level	#9	Plausible session length	12,000	3,500	1,200	1,300	1,000	2.40	FLAGGED
Row-level	#10	Duration present	0	0	0	0	0	0.00	OK
Row-level	#11	Self-replace bug (Q2)	600	400	0	400	0	0.12	FLAGGED
Behavioural	B1	Return / repeat	45,000	45,000	0	27,000	18,000	30.00	ALERT
Behavioural	B2	One-session share	105,000	105,000	105,000	0	0	70.00	ALERT
Behavioural	B3	Tenure success-floor	500	500	0	300	200	0.33	FAIL
Behavioural	B4	Very-churny countries	4,000	1,500	900	400	200	1.30	FLAGGED
Behavioural	B5	Inactivity churn 7d	100,000	50,000	30,000	12,000	8,000	38.90	FAIL
Behavioural	B6	Bot / burst deep-dive (Q4)	10	10	0	4	6	0.01	FAIL
Behavioural	B7	Country hopping	0	0	0	0	0	0.00	OK
Behavioural	B8	user_type validity	0	0	0	0	0	0.00	OK

Behavioural	B9	Failure with no reason	0	0	0	0	0	0.00	OK
Behavioural	B10	Churn-after-switch	0	0	0	0	0	0.00	OK
Journey	T1	Flap (success <30s)	41,400	25,000	12,000	8,000	5,000	9.00	OK
Journey	T2a	Activation (first session)	6,000	6,000	4,000	1,200	800	4.00	OK
Journey	T2b	D1 next-day return	—	—	—	—	—	0.00	INFO
Journey	T3	Churn-risk users	100	100	0	30	70	0.07	FAIL

3 — Which check caused how much of the drop (one-owner attribution)

Each lost success / lost user is attributed to **one** check (first in priority order), so the contributions **sum** to the overall change. **Inactivity churn (B5) is the single biggest driver** by far; the newer checks add precision with minimal impact.

Rank	ID	Check	Successes flipped	Users lost	Contribution to success %
1	B5	Inactivity churn 7d ★ biggest driver	110,000	50,000	-22.00 pts
2	#9	Plausible session length	6,000	3,000	-1.20 pts
3	#1	Timestamp validity	3,000	1,000	-0.60 pts
4	B4	Very-churny countries	500	300	-0.10 pts
5	#11	Self-replace bug (Q2)	300	0	-0.06 pts
6	B3	Tenure success-floor	150	150	-0.03 pts
7	T3	Churn-risk users	30	30	-0.01 pts
8	B6	Bot / burst deep-dive (Q4)	10	10	+0.00 pts
9	#7	Valid server IP	5	3	+0.00 pts
10	#8	Collection continuity	3	3	+0.00 pts
11	#5	Non-zero duration	2	0	+0.00 pts

4 — Bottom line

1. **The drop is churn, not corruption.** ~96% of rows pass all row-level checks; the 92% → 68% fall is driven almost entirely by **inactivity churn (B5)**. 2. **The newest checks add precision, not noise.** #11 self-replace and the B6 deep-dive each reclassify a small, well-targeted set without punishing real users. 3. **Most lost users barely engaged.** Of 60,000 dropped users,

~83% were one-and-done; only ~10,000 were retained. 4. **Every number is produced by SQL and self-validated**, so the before/after is exact and reproducible. *Method demonstration on illustrative data. No real client, product, or production figures are shown.*