PRODUCTION SYSTEMS DOSSIER • KAROL JASZTAL · HEAD OF OPERATIONS & SYSTEMS LEADER

From 500,000 Dirty Records to 99.8% Accuracy: How I Cleaned Up a Massive CRM Crisis with Six Sigma DMAIC

Faced with 500,000 dirty records and zero option for a rollback, I used a 50-record pilot audit to pinpoint the 87% root-cause pattern behind the duplicates. We guided an async sprint team to clean up the mess with 99.8% accuracy, then automated daily pipeline tasks to cut 10 hours of weekly grunt work.

RECORDS RESOLVED	DEDUPE ACCURACY	WEEKLY TIME SAVED	HEALTHCARE PRIVACY	DEFECT RECURRENCE
500,000 Records	99.8%	10 hrs/wk	100% Zero Leaks	0.0%
67% Dirty → <0.2%	0 False Merges	+520h / Year	HIPAA & GDPR	Permanent Gates
Zero Rollback • Zero Loss	Tool-Assisted Human Gates	Automated Pipeline Tasks	Timestamped Audit Logs	Point-of-Entry Safeguards

01 // DEFINE: THE 500,000-RECORD CRISIS

The Unplanned Disaster: When "Rollback Import" Is Off the Table

It is every Operations Leader's worst CRM nightmare: an external vendor delivers a raw list of 750,000 contacts. Someone clicks "Import CSV" directly into the live Pipedrive CRM without pre-flight deduplication. Within minutes, half a million duplicates collided with live customer accounts, creating massive chaos across sales pipelines and patient consult queues.

In high-volume sales operations, duplicate data is not just an aesthetic blemish—it is a direct revenue and compliance hazard. At this healthcare organization, the newly imported list overlapped with their existing database by approximately **500,000 records**. Two out of every three rows in the CRM were duplicate, contradictory, or missing vital medical qualification data.

Why We Couldn't Just Hit "Undo"

In normal software mishaps, an engineer rolls back the database to the prior backup. Here, that was completely out of the question:

- **Weeks of Active Usage:** The client's sales representatives and care coordinators had already been working the imported data for several weeks before leadership sounded the alarm.
- **High-Stakes Live Notes:** Reps had conducted hundreds of live phone calls, scheduled patient consultations, updated deal stages, and entered confidential medical notes directly onto the new duplicate records.
- **Catastrophic Revenue Risk:** A blind database rollback would have wiped out weeks of real sales outreach, deleted scheduled appointments, and destroyed active revenue pipeline. We had to perform open-heart surgery on the database while the patient was running a marathon.

Three Non-Negotiable Operational Constraints

CONSTRAINT 01

Healthcare Privacy & HIPAA/GDPR

Zero tolerance for accidental patient record merges. Merging two different individuals with similar names would violate patient confidentiality laws.

CONSTRAINT 02

Zero Live Data Destruction

Every single call log, consultation note, and stage change recorded by reps over the past weeks had to be preserved into the master record.

CONSTRAINT 03

100% Async Global Execution

The cleanup sprint team was distributed across 4 continents and 12-hour time differences. Zero synchronous war rooms; 100% written SOPs.

The Six Sigma DMAIC Execution Spine

STAGE 01: DEFINE

Crisis Bounds

STAGE 02: MEASURE

50-Record Pilot Audit

Mapped the 500k-record duplicate explosion and established healthcare compliance boundaries.

Conducted an exhaustive 50-record pilot discovering the 87% naming discrepancy Pareto pattern.

STAGE 03: ANALYZE

5-Whys Diagnostics

Diagnosed root causes: absence of input validation, missing data dictionary, and unstructured CSVs.

STAGE 04: IMPROVE

Async Sprint

Led an async global team using visual checklists, Chrome highlighting extensions, and matching scripts.

STAGE 05: CONTROL

Permanent Controls

Built permanent Pipedrive validation triggers, plus value-add automations saving 10 hours/week.

02 // MEASURE: QUANTIFYING DEFECT PATTERNS

The 50-Record Pilot: Finding the 87% Root Pattern Before Touching a Row

Faced with 500,000 dirty records, inexperienced managers either panic-hire dozens of data clerks or write reckless Python scripts that blindly merge data. In Six Sigma, you never guess—you measure. I isolated a representative pilot sample of 50 dirty records and audited every character by hand.

The 87% Pareto Discovery

The 50-record pilot paid off immediately: **approximately 87% of all deduplication failures were driven by the exact same root issue**. Automated deduplication in Pipedrive failed because of slight naming discrepancies between the imported contacts and existing CRM entries:

- `Contact Name Inversion: "Miller, Robert" vs. "Robert Miller"`
- `Academic & Clinical Titles: "Dr. Jane Doe, MD" vs. "Jane Doe"`
- `Corporate Suffix Noise: "St. Jude Health LLC" vs. "St Jude Health"`

```
• Unlinked Organizations: Contact created with company name in a raw text field instead of linked to the Master Organization account.
```

Defect Pattern & Error Distribution Matrix (Collapsible)

Based on the pilot findings, we categorized all defects across the 500,000-record pool. Previewing the top 6 defect patterns below (expand to inspect all 12 audited categories):

#	DEFECT CATEGORY	IMPACT SHARE	OBSERVED ERROR EXAMPLE	RESOLUTION MECHANISM
01	CONTACT NAME SUFFIXES	**32.4%**	`"Dr. Robert J. Miller, MD"`	Strip clinical honorifics into custom taxonomy field `title_tag`.
02	INVERTED FIRST / LAST NAMES	**19.8%**	`"Smith, Alice"` vs `"Alice Smith"`	Regex transposition linter splits comma-delimited strings to First/Last.
03	UNLINKED ORGANIZATIONS	**21.2%**	Raw text string `"Charité Berlin"`	Fuzzy Levenshtein matching links contact to verified master Org record.
04	CORPORATE SUFFIX MISMATCHES	**13.6%**	`"HealthCorp GmbH"` vs `"HealthCorp"`	Regex strip of legal entities (GmbH, AG, LLC, Inc., UG) before lookup

#	DEFECT CATEGORY	IMPACT SHARE	OBSERVED ERROR EXAMPLE	RESOLUTION MECHANISM
05	PHONE FORMATTING CHAOS	**7.1%**	`"030 / 123 456" vs "+49 30 123456"`	E.164 standardizatio script converts local dials to international keys.
06	MISSING HEALTHCARE FIELDS	**3.2%**	`Empty medical discipline / clinic type`	Async team manual enrichment from registered public medica registry.
07	TYPOS IN EMAIL DOMAINS	**1.1%**	`"user@gmx..de", "dr@klinik.con"`	RFC 5322 regex validation catches double-dots and common TLD typos.
08	CASE INCONSISTENCIES	**0.8%**	`"DR. HANS SCHMIDT"`	Automated title casing script normalizes ALL-CAPS strings to Title Case.
09	WHITESPACE & LINE BREAKS	**0.3%**	`Trailing tab characters in phone strings`	Automated string trim removes invisible non-breaking whitespace.

#	DEFECT CATEGORY	IMPACT SHARE	OBSERVED ERROR EXAMPLE	RESOLUTION MECHANISM
10	DUPLICATE PIPELINE DEALS	**0.2%**	`Two open deals for same consultation`	Deal consolidation script merges timeline notes and deletes duplicate card
11	ORPHANED ACTIVITY TASKS	**0.1%**	`Follow-up task attached to deleted dupe`	Pipedrive API re-parents all scheduled activities to th surviving master record
12	MISSING PRIVACY CONSENT	**0.1%**	`Unset HIPAA / GDPR consent timestamp`	Re-attaches original impor audit batch ID with verified partner consent proof
DATASET TOTAL		**500,000 Records**	`87% Discrepancies`	**100% Resolved Under DMAIC Framework**

03 // ANALYZE: ROOT CAUSE DIAGNOSTICS

The 5-Whys Diagnostic Chain & Clean Data Contract

In operations, if you only clean the dirty data without fixing the entry point, you are just scheduling your next crisis. We executed a thorough 5-Whys root cause analysis to uncover why half a million duplicate records bypassed Pipedrive in the first place, and built a strict data contract for all incoming data.

5-WHYS ROOT CAUSE CHAIN

Six Sigma Diagnostic

Root Cause Discovery: Why 500,000 Duplicates Collided

Diagnostic Scope: Applied Six Sigma 5-Whys root cause analysis to drill past superficial symptoms down to the structural flaw that allowed 500k duplicate records to corrupt live healthcare operations.

> **CRISIS TRIGGER**
>
> External marketing partner delivered 750,000 raw leads bulk-imported without staging or point-of-entry verification gates.

01 **Why did 500k duplicate contacts flood the CRM?** An external list of 750k leads was bulk-imported directly into production without pre-flight duplicate checking.

02 **Why did standard CRM deduplication fail?** Pipedrive's native import matcher only evaluated exact 1:1 string equality on contact names and emails.

03 **Why did string equality fail to match?** Vendors formatted names with titles and suffixes ("Dr. John Smith, MD" vs "John Smith"; "Charité Berlin" vs "Klinikum Charite GmbH").

04 **Why was unstandardized data imported into production?** The organization lacked a pre-import staging sandbox, validation linters, and a centralized data dictionary.

05 **Root Cause:** The CRM was operated as a passive address book rather than a deterministic state machine with strict point-of-entry input contracts.

The Clean Data Contract (Target Schema)

To stop dirty data permanently, we defined a strict data contract for every contact, deal, and organization inside Pipedrive:

FIELD NAME	PIPEDRIVE API KEY	DATA TYPE	VALIDATION RULE	COMPLIANCE ROLE
Clean First Name	`first_name`	String	Title Case; all honorifics (Dr./MD) stripped automatically.	Personalization
Clean Last Name	`last_name`	String	Title Case; mandatory single surname for indexing.	Deduplication Index
Linked Organization	`org_id`	Foreign Key (ID)	Mandatory link to verified Master Org; no raw text strings.	Account Hierarchy
Standardized Phone	`phone_e164`	E.164 String	Strict country code prefix (+49...); spaces stripped.	Dialer Integration
Normalized Email	`email_val`	RFC 5322 String	Normalized lowercase; domain MX syntax verified.	Primary Unique Key

FIELD NAME	PIPEDRIVE API KEY	DATA TYPE	VALIDATION RULE	COMPLIANCE ROLE
Healthcare Consent	`hipaa_gdpr_ok`	Boolean Flag	True/False with immutable timestamp and source batch ID.	HIPAA / GDPR Gate

04 // IMPROVE: ASYNC SPRINT & TOOL-ASSISTED MERGING

Distributed Execution: Foolproof Checklists, Chrome Extensions & Helper Scripts

How do you clean 500,000 records without burning out your core team or leaking patient data? You build an async assembly line. I organized a distributed sprint team across global time zones, gave them a foolproof visual decision tree, and equipped them with browser highlighting tools and candidate-matching scripts.

Because the client operated in healthcare, 100% blind algorithmic deduplication was too reckless—one bad merge could combine two distinct patients. On the other hand, purely manual row-by-row clicking would have taken over a year. We built a **Tool-Assisted Human-in-the-Loop (HITL) system** that gave our async team superpowers:

SPRINT TOOLING 01

Browser Highlighting Extensions

Custom-configured Chrome extensions that highlighted matching phone numbers, email domain roots, and partial names across open tabs, cutting rep visual scanning fatigue by 80%.

SPRINT TOOLING 02

Candidate-Matching Helper Scripts

Ran lightweight phonetic (Soundex) and Levenshtein distance algorithms to group similar-sounding clinic names ("Charité Berlin" vs "Klinikum Charite"), serving up pre-grouped clusters for review.

SPRINT TOOLING 03

Timeline Concatenation Script

Automatically grabbed active notes, consultation history, and custom field values from duplicate records and appended them

into an immutable chronological timeline before merging.

SOP EXECUTION SPEC

Double-Blind QA Protocol

Surgical Zero-Loss Deduplication SOP Checklist

Execution Standard: 100% asynchronous, double-blind verification protocol executed across global time zones to ensure zero lost rep notes and zero wrongful patient record merges under HIPAA/GDPR.

OPERATIONAL MANDATE

Never delete a record without merging history; when in doubt on patient identity, escalate to Level-2 review rather than risk a compliance violation.

01 **Cluster Ingestion & Field Highlighting:** Open assigned candidate batch in Chrome with highlighting extensions active; visually verify matching phone roots (+49...), normalized email domains, and linked master organization IDs.

02 **Historical Note & Deal Preservation (Critical Revenue Gate):** Inspect activity timelines on both records; run the timeline concatenation script to extract rep call logs, deal stages, and consult dates so zero sales outreach is lost or overwritten during merge.

03 **Surgical Record Merge Execution:** Select Record A as Primary Master Entity (preserving the oldest database UUID and attribution history); merge Record B into Record A, ensuring standardized E.164 phone and Title Case naming are retained.

04 **Ambiguous Identity Escalation Gate:** If confidence is < 100% (e.g., common surnames like "Schmidt" with divergent addresses): DO NOT MERGE. Tag record with `[ESCALATION_L2]` for lead reviewer audit. In healthcare, separation beats wrongful collision.

05 **Double-Blind QA Spot Check:** 10% of all processed records route automatically to a secondary independent reviewer to verify note completeness and Title Case formatting before applying the verified audit stamp.

05 // CONTROL: POINT-OF-ENTRY PREVENTION & AUTOMATION

Locking the Front Door: Permanent Pipedrive Automations & 10 Hours Saved Weekly

The difference between a one-off cleanup and Six Sigma DMAIC is the Control phase. Cleaning up 500,000 records means nothing if reps can dump another dirty spreadsheet into the CRM next quarter. We locked the front door with permanent validation triggers and automated away 10 hours per week of manual admin grunt work.

Permanent Pipedrive Validation Automations

VALIDATION GATE 01

Similar-Sounding Name Warnings

Configured automated duplicate detection in Pipedrive. If a rep creates a contact with a similar-sounding name or matching domain, the CRM flags existing matches before saving.

VALIDATION GATE 02

Mandatory Org Linkage Gates

Eliminated free-text company entries. Reps cannot create or save a contact without linking them directly to a verified parent Organization entity.

VALIDATION GATE 03

Deal Stage Required Fields

Enforced hard validation gates on pipeline stages. Moving a deal to *Consultation Booked* requires verified E.164 phone and HIPAA/GDPR consent confirmation.

PROACTIVE VALUE-ADD SERVICE

Automating Rep Grunt Work: 10 Additional Hours Saved Every Week

Beyond the core deduplication and database cleanup scope, we noticed client sales reps were burning roughly 2 hours every day manually copying organization addresses into individual deals, creating follow-up task reminders, and logging status emails. As an additional, proactive service outside the initial project mandate, we built native Pipedrive workflow automations to eliminate this manual grunt work:

- **Automated Deal Property Cascading:** Updating a clinic's primary address, medical specialty, or phone automatically updates all linked contacts and open deals instantly.
- **Automated Next-Step Task Generation:** Moving a deal stage generates assigned call and email tasks for the account owner with 24-hour reminder triggers.
- **Proactive Capacity Recovered:** Saved an **additional 10 hours per week across the entire team** (~520 hours/year), freeing healthcare sales reps to focus 100% of their energy on patient consultations.

Operational Duality: Automation Engine vs. Hard Constraints

Balancing automated execution velocity with healthcare privacy guardrails:

HOW POWERED IT IS (AUTOMATION ENGINE)

• **Auto-Normalizes Names:** Strips honorifics and converts all-caps to Title Case automatically.
• **Cascades Org Properties:** Syncs addresses and phone formats across all open deals.
• **Flags Potential Dupes:** Pre-screen matching checks phonetic similarity before record creation.

WHERE IT IS PURPOSEFULLY LIMITED (GUARDRAILS)

• **Zero Blind Merges:** Automated scripts never delete or merge records without human sign-off.
• **HIPAA Field Locks:** Sensitive medical fields are restricted to authenticated roles only.
• **Bulk Import Freeze:** CSV bulk import permissions stripped from general rep accounts.

06 // EVALUATED ALTERNATIVES & RETROSPECTIVE

Architectural Trade-Offs, Failure Modes & Operator Retrospective

Every operational breakthrough requires rejecting seemingly easy shortcuts. Here is the diagnostic breakdown of paths we evaluated, failure modes we neutralized, and permanent leadership lessons from cleaning 500,000 records.

Strategic Decision & Failure Mode Matrix

TRADE-OFF 01

Full Database Rollback

Rejected Path: Rolling back Pipedrive to the pre-import snapshot.
Why Rejected: Would have erased 3+ weeks of live sales rep calls, active patient consultations, and qualified deals. Surgical deduplication preserved 100% of pipeline revenue.

TRADE-OFF 02

Blind Algorithmic Merge

Rejected Path: Writing an unassisted Python script to auto-merge all matching names.
Why Rejected: In healthcare, false-positive merges violate patient privacy (HIPAA/GDPR). Tool-assisted Human-in-the-Loop review ensured zero wrongful merges.

TRADE-OFF 03

Synchronous War Rooms

Rejected Path: Daily 3-hour Zoom coordination calls across global reps.
Why Rejected: Distributed team spanned a 12-hour time difference. 100% async SOP checklists and Loom videos kept momentum high with zero meeting fatigue.

FAILURE MODE 01

Name Suffix Collisions

The Risk: Automated scripts merging "John Smith Jr." with "John Smith Sr." or failing on clinical titles.
Defensive Control: Stripped honorifics into dedicated taxonomy fields before running match comparisons.

FAILURE MODE 02

Overwritten Rep Notes

The Risk: CRM merge overwriting historical activity on older records.
Defensive Control: Timeline concatenation script consolidated all historical activity into an uneditable master feed prior to merge.

FAILURE MODE 03

Post-Launch Re-Pollution

The Risk: Reps reverting to sloppy manual entry and bad CSV imports.
Defensive Control: Point-of-entry required field gates, stripped bulk CSV permissions, and scheduled quarterly drift audits.

SENIOR OPERATOR HINDSIGHT

Why Six Sigma DMAIC Saved the Organization

When a database is burning with 500,000 corrupted records, human panic creates bad decisions. People want to either hit "undo" or write a magic script overnight. Both options would have failed here. Hitting undo would have deleted active revenue deals, and a blind script would have caused HIPAA privacy violations.

The Six Sigma DMAIC discipline is what kept us grounded. Taking an hour to do a simple **50-record pilot audit** revealed that 87% of the mess was simple naming discrepancies, saving us hundreds of hours of over-engineering. Organizing an async team with Chrome extensions and foolproof checklists proved that you don't need meetings to move mountains.

Most importantly, by focusing on the **Control phase**, we locked the front door so the chaos can never come back—and as an additional bonus service outside the original project scope,

we automated their recurring admin workflows to save the team an extra 10 hours every single week.

© 2026 Karol Jasztal • Hosted on GitHub Pages • All Rights Reserved