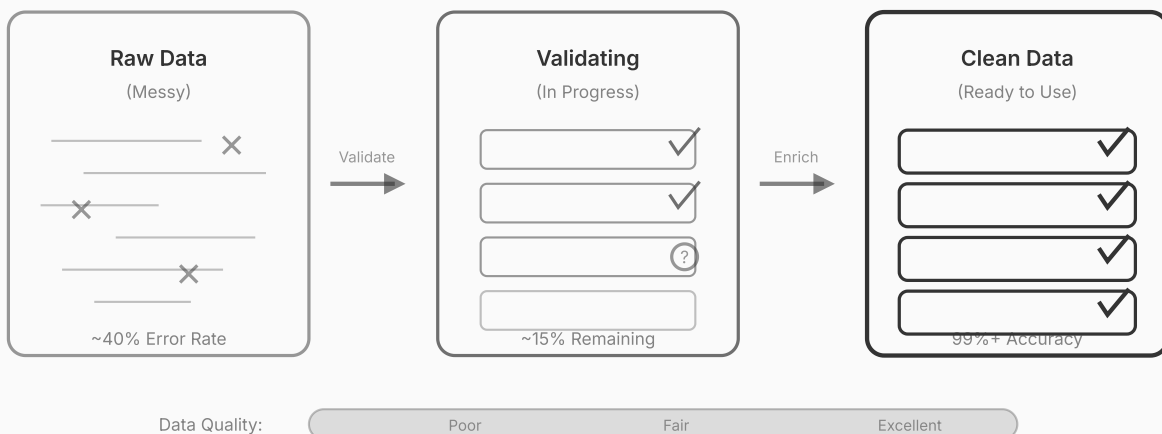**Databender**

Data Cleanup for Manufacturing

Fix Your Data Foundation First

databender.co | Data & AI Consulting for Growing Businesses

80-90%

Cost Reduction

AI-powered data cleanup costs 80-90% less than traditional methods.

Data Quality Transformation Journey

Raw Data
(Messy)

Validating
(In Progress)

Clean Data
(Ready to Use)

Validate → Enrich →

~40% Error Rate

~15% Remaining

99%+ Accuracy

Data Quality: Poor · Fair · Excellent

Systematic cleanup transforms chaotic data into reliable business intelligence

Your CRM has 47,000 contacts. How many are duplicates? How many are the same company spelled three different ways? How many phone numbers go nowhere, how many emails bounce?

Nobody knows. And that uncertainty costs real money.

We cleaned 1.69 million ownership records for a client last year. The same person was listed ten different ways across their database. John Smith at ABC Industries. J. Smith at ABC. Jonathan Smith at A.B.C. Industries Inc. Ten records. One customer. The sales team was calling the same leads multiple

times, embarrassing themselves and annoying prospects who'd already said no.

The traditional approach would have taken a team of contractors six months and cost hundreds of thousands of dollars. We did it with AI agents at 125x less cost than doing it by hand. Not because the technology is magic. Because the methodology is right.

Why Data Decays

Clean data doesn't stay clean. This isn't a failure of discipline. It's physics.

People change jobs. Companies merge and split. Addresses update. Phone numbers transfer. Your CRM captures a moment in time. That moment was accurate once. Now it's stale.

Meanwhile, data enters from multiple sources. Website forms. Trade show badge scans. Sales rep notes typed on phones. Each source has its own format, its own level of completeness, its own quirks. Marketing imports a list from a webinar. Sales adds contacts from business cards. Customer service creates records during support calls. None of them coordinate.

The result is predictable. The same customer exists five times with slight variations. Their company name is spelled differently in each record. Their contact info reflects different moments in their career. Your database doesn't know these are the same person.

Data quality isn't something you achieve once. It's something you maintain continuously.

The True Cost of Dirty Data

Sales teams waste time calling bad numbers and chasing duplicates. Marketing sends campaigns to addresses that bounce, damaging sender reputation and email deliverability. Customer service can't find the complete picture of a client relationship because the history is scattered across multiple records.

But the visible costs are the small part. The invisible costs hurt more.

Bad data leads to bad decisions. If your CRM shows 50,000 prospects in the manufacturing sector, but 30% are duplicates and 20% are defunct companies, your market size estimate is off by half. Pipeline forecasts built on dirty data are fiction. Territory assignments based on flawed customer counts create imbalances nobody understands.

One industrial equipment distributor discovered that 40% of their "active prospects" hadn't existed for over two years. Dead companies. Merged entities. Changed industries. The sales team had been working a territory map based on ghosts.

When they cleaned the data, they didn't just remove bad records. They found opportunities they'd missed. Contacts they thought were unaffiliated were actually at the same company. Companies they'd ignored were actually perfect fits once the duplicate records merged and revealed the true relationship history.

Deduplication That Actually Works

Most deduplication tools match on exact fields. Same email? Duplicate. Same phone number? Duplicate. This catches the easy cases and misses everything else.

Real-world duplicates don't match exactly. The same person might be listed as:

- Mike Johnson at Precision Manufacturing LLC

- Michael Johnson at Precision Mfg

- M. Johnson at Precision Manufacturing

- Mike Johnson at Precision (no company suffix)

- Michael K. Johnson at Precision Manufacturing LLC

Exact matching catches none of these. Fuzzy matching catches some but creates false positives. The Precision Manufacturing in Ohio isn't the same as Precision Manufacturing in California, even though the names match.

The solution is probabilistic matching with human-informed rules. The system calculates confidence scores based on multiple factors: name similarity, address proximity, phone area codes, email domains, industry classifications. High-confidence matches merge automatically. Medium-confidence matches flag for review. Low-confidence matches stay separate.

We build these systems to learn from corrections. When a reviewer says "these are actually different people," the model updates. When they say "these should have matched," it adjusts. Over time, accuracy improves without manual intervention.

Record Matching Across Systems

Your CRM isn't the only place customer data lives. The ERP has it too. So does the billing system. The marketing automation platform. The customer service database. The accounting software.

Each system has its own version of the truth. Sometimes they match. Usually they don't.

Cross-system record matching is where data cleanup gets interesting. It's not enough to dedupe within one system. You need to establish which records across all systems refer to the same real-world entity. The customer in Salesforce, the account in SAP, the contact in Hubspot, the payer in QuickBooks. Four records, one customer, four different IDs, four slightly different versions of the facts.

The goal is a master data hub. A single source of truth for customer identity that all systems can reference. When someone updates a phone number, it updates everywhere. When a company changes its name, all records reflect the change.

Building this hub requires mapping fields across systems (what Salesforce calls "Company" might be "Account Name" in SAP and "Organization" in the marketing platform), establishing hierarchy rules (which system wins when they conflict), and creating sync mechanisms (how changes propagate).

It's not glamorous work. It's the foundation that everything else depends on.

Data Quality Beyond Deduplication

Removing duplicates is only part of the problem. The records that remain need to be complete and accurate.

Standardization. Addresses should follow consistent formatting. Company names should use consistent suffixes. Phone numbers should include area codes. States should be abbreviations or full names, not a mix. Standardization makes matching easier and reporting more reliable.

Validation. Email addresses should be valid formats that don't bounce. Phone numbers should have the right number of digits for their country code. ZIP codes should match cities and states. Invalid data should be flagged or removed, not left to cause problems later.

Enrichment. Missing fields should be filled where possible. If you have an email domain, you can often determine the company. If you have a company name, you can look up industry, size, and location. Enrichment fills gaps that reduce record value.

Freshness. Data ages. Contact info from 2019 may be useless now. Decay detection identifies records that haven't been updated or verified recently. Some should be re-verified. Some should be removed. Stale data is worse than no data because it creates false confidence.

The Cleanup Process

A realistic data cleanup project for a mid-sized manufacturer follows this pattern.

Week 1: Assessment. Export data from all relevant systems. Run automated analysis to quantify the problem: duplicate rates, invalid formats, missing fields, cross-system mismatches. Build the business case with real numbers.

Weeks 2-3: Rule development. Define matching logic based on your specific data patterns. What constitutes a duplicate in your context? How should company name variations be handled? What's your tolerance for false positives versus false negatives?

Weeks 4-6: Processing. Run the deduplication and standardization algorithms. Review edge cases. Tune the rules based on results. Iterate until quality meets your threshold.

Weeks 7-8: Merge and load. Execute the merges in production systems. Map cleaned data back to source systems. Establish the master record relationships.

Ongoing: Maintenance. New data arrives daily. The same problems that created your mess will create it again without prevention. Add validation rules on data entry, scheduled cleanup runs, and monitoring applications.

Eight weeks to fix a problem that's been building for years. Not months. Not quarters. Weeks.

Common Mistakes

We've seen data cleanup projects fail the same ways repeatedly.

Starting with tools instead of goals. Software vendors will happily sell you data quality platforms. But if you don't know what "clean" means for your business, no tool will help. Define success criteria first. What questions should you be able to answer? What processes should work better?

Cleaning once without maintaining. A one-time cleanup provides temporary relief. Without prevention and maintenance, you're back where you started in 18 months. Budget for ongoing work, not just the initial fix.

Ignoring the source of the problem. If duplicates keep appearing, something in your process is creating them. Fix the leak, not just the flood. Validate at data entry. Train users on proper record creation. Integrate systems that should share data instead of duplicating it.

Over-automating decisions. Aggressive auto-merge settings save time but destroy good data when they're wrong. Start conservative. Trust the system more as it proves itself. Prefer false negatives (missing a merge) to false positives (merging records that shouldn't merge) early on.

What Changes After Cleanup

The industrial distributor we mentioned saw immediate improvements. Sales call efficiency increased because reps weren't wasting time on dead numbers. Email campaign deliverability improved because they weren't hitting invalid addresses. Territory planning became possible because they finally knew how many real prospects existed in each region.

The longer-term changes mattered more. Marketing could run accurate attribution analysis because customer journeys weren't split across duplicate records. Finance could trust AR aging reports because customer accounts weren't duplicated. Customer service could see complete relationship histories instead of fragments.

And the sales team stopped embarrassing themselves. No more calling the same prospect twice in a week. No more arguing about who owns a lead that's actually the same person. No more lost deals because critical context was attached to a duplicate record nobody found.

Clean data isn't exciting. It's the foundation that makes everything else possible. Skip it, and every system you build wobbles. Invest in it, and everything you do works better.

Ready to stop fighting your data? Schedule a conversation about what cleanup could look like for your organization, or explore our full manufacturing solutions.

databender.co