Test data subsetting and synthetic generation

Test data that looks real. Because it isn't.

DataFactory gives every test environment production-shaped data, without a single real customer in it. Cut a huge database down to size with every relationship intact, or generate realistic data from nothing but a schema.

PRUNE
Subset real data

BLOOM
Generate synthetic

Self-hosted on your own server with Docker. Nothing is sent to Enov8.

Virtual data set · banking	Target 5%
customers anchor	120,000
accounts	195,000
cards	105,000
addresses	130,000
transactions	9,200,000
statements	1,550,000

0 orphaned rows · CSV / JSON per table · Schema DDL

Source row counts shown. Example plan only.

250+
data skills with real check digits

12
industry templates

10M
rows per run by default

0
real customers in synthetic data

HOW IT WORKS

From source to test environment in four steps.

1. **Connect**
Point it at a source with a read-only account, or start from a schema or template.

2. **Define**
Pick anchor tables. Every related table is mapped into a reusable virtual data set.

3. **Prune or Bloom**
Cut a subset to size, or generate synthetic rows with your skills and rules.

4. **Deliver**
Download files, load a test database, or let pipelines and AI agents do it.

QUALITY YOU CAN PROVE

Realistic shape. Enforced rules. Checked every run.

A quality score on every run

✓ Keys ✓ Links ✓ Constraints ✓ Formats ✓ Rules

Realistic spread

Set child rows per parent and how they're distributed.

Uniform · Poisson · Long-tail

Business rules that hold

`claim.date ≥ policy.start` date order

`lines.amount sums to total` parent total

Same plan + same seed = the same data, every time. Reproduce a failing test months later, row for row.

TWO ENGINES

Shrink what you have. Grow what you don't.

Prune cuts real databases to size. Bloom builds realistic data from scratch. Use either, or both.

PRUNE

Subsetting

Cut a large database down to anywhere from 1% to 100%, with every foreign-key relationship intact. Nothing orphaned, nothing broken.

- ✓ **Any target size**, plus optional row filters by value, text, year or range
- ✓ **Read-only on the source.** Prune never changes it
- ✓ **Reduction script** that shrinks a cloned copy in place

Sources · MySQL · SQL Server · Oracle · PostgreSQL · Db2 · SAP HANA

Output · CSV or JSON per table · schema DDL · reduction script

BLOOM

Synthetic data

Generate realistic data from a schema. No source rows needed, so no real customer is ever involved.

- ✓ **250+ skills** with real check digits, from cards to Medicare and NHS numbers
- ✓ **Relationships and business rules** applied while generating, then checked
- ✓ **"Learn from source"** copies only table sizes and value mixes, never personal values

Schemas · PostgreSQL · MySQL · SQL Server · Oracle · Snowflake · BigQuery and more

Output · CSV · JSON · SQL inserts · direct load

250+ SKILLS

Fake data that passes real validation.

Identifiers come out with real check digits, so they pass the same validation your applications run.

Card number	Global	Luhn
ABN	AU	Mod 89
TFN	AU	Weighted mod 11
Medicare	AU	Weighted mod 10
NHS number	UK	Mod 11
NPI	US	Luhn, 80840 prefix
IMEI	Telco	Luhn
ACN	AU	Weighted mod 10

Also `IHI` `MBI` `ICCID` `NMI` `MIRN` `USI` `IBAN` `BSB` `VIN` `MSISDN` `IMSI` and more

INDUSTRY PACKS

Twelve templates that speak your industry.

Banking & finance · 1
BSB · IBAN · Cards

Health insurance · 4
IHI · Medicare · NHS · NPI · MBI

General insurance · 1
VIN · Rego · ABN

Telecommunications · 1
MSISDN · IMSI · IMEI · ICCID

Government · 1
CRN · TFN · Driver licences

Energy & utilities · 1
NMI · MIRN · NEM12 reads

Super & wealth · 1
USI · Contributions · Rollovers

Retail & e-commerce · 1
Orders · Products · Customers

Salesforce · 1
Sales Cloud · 18-char IDs

IN THE PRODUCT

Real screens. Real data sets. No real customers.

Every prune and every Bloom plan runs through the same guided flow, with results you can check.

Prune result

FIG. 1 · PRUNE

RESULT

98% smaller

212,074 rows from 9,580,550

2% of source footprint

ROWS REMOVED **9,368,476**

DURATION **26s**

Source
Sales DB (production copy)

Project name
Orders & payments

Anchor tables
public `order_items` `payments`

Target percentage
1%

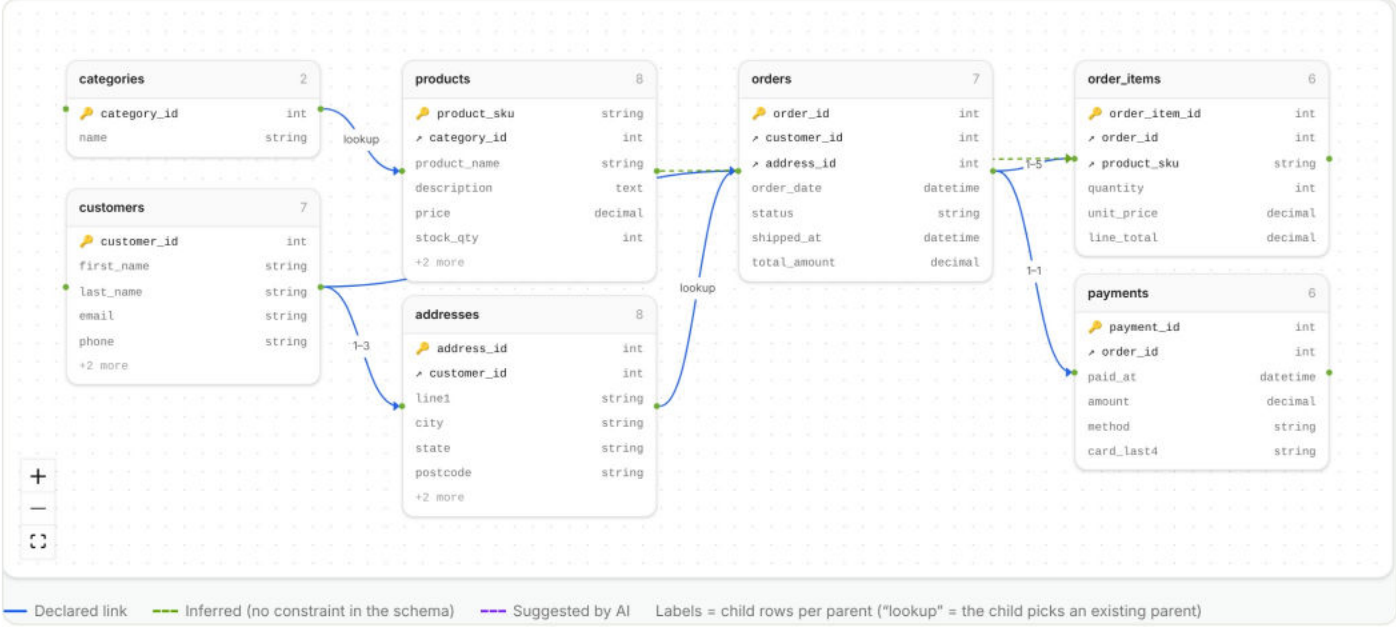
File type
CSV

Created
October 15, 2026 9:00 a.m.

1. **98% smaller** in 26 seconds, from a 9.5 million row production copy.
2. **Anchor tables** drive the subset; related rows follow, nothing orphaned.
3. **Every run is recorded** with source, target and output type.

Bloom relationships

FIG. 2 · BLOOM



4. **Seven tables, seven links** mapped from the schema, including inferred ones.
5. **Rows per parent** set on each link, such as 1–3 addresses per customer.
6. **AI suggestions** are optional and shown separately from declared links.

Screens shown with demonstration data.

AI & AUTOMATION

AI if you want it. Never your data.

DataFactory works fully without AI. Switch it on to describe columns, rules and changes in plain English. Only the structure goes to the model.

✓ Sent to your AI

Table and column names

Plan settings

The project brief

✕ Never sent

Rows of data

Source values

Database passwords

Built for pipelines

A REST API covers prunes, plans, runs and loads, so fresh data becomes a CI/CD step.

Ready for AI agents

The built-in MCP server lets agents work as a signed-in user. No passwords pass through.

Use the model you trust: Anthropic Claude · OpenAI · Azure OpenAI · Ollama (local)

SECURITY

Runs inside your walls. Nothing is sent to Enov8.

Self-hosted
Runs on your own server with Docker. Data never leaves your network.

Read-only on sources
DataFactory only reads from source databases. It never changes them.

Secrets encrypted
Passwords and AI keys are encrypted at rest and never shown again.

Single sign-on
Microsoft Entra ID, OpenID Connect or SAML 2.0.

Private by default
Admin and Member roles. You see only your own work unless shared.

A built-in safety catch
Refuses to load a plan into the database it was learned from.

BETTER TOGETHER

Part of the Enov8 platform.

DF + Test Data Manager

Mask what you keep. Replace what you don't.

TDM masks the real copies you keep. DataFactory replaces the ones you don't.

DF + VirtualizeMe

Clone in seconds, then shrink.

Clone with vME, then run the reduction script to shrink the clone.

DF + Env & Release Manager

Book it. Then fill it.

Book the environment, then fill it with exactly the right data.

Fill your next test environment. Leave your customers out of it.

See Prune and Bloom on your own schema in a 30-minute session with our team.

Book a demo · enov8.com/contact-us

IN THE SESSION

1 · Prune a subset of your schema to size.

2 · Generate a synthetic data set with Bloom.

3 · Review quality checks and delivery options.