WHITE PAPER

GENROCKET'S SECURE ENTERPRISE DATA PLATFORM

Sensitive Data Never Leaves Production

A white paper for Chief Information Security Officers on eliminating production data risk across software development, testing and AI, with GenRocket's Secure Enterprise Data Platform.

Executive summary

The biggest data exposure in most enterprises sits outside the production zone, in the development, test, analytics and AI environments that hold copies of production data. Security programs protect the system of record well. The copies get much less attention, and AI is now multiplying them.

GenRocket's position is simple: **100% production data privacy and security, because sensitive data never leaves production.** The GenRocket Secure Enterprise Data Platform gets you there in three phases:

1. **Protect.** Mask a copy of the complete database or file in-place, inside the production zone. Metadata identifies sensitive data, which is replaced with deterministic synthetic values rather than algorithmically obfuscated. Nothing sensitive moves downstream. Mo.ve the fully secure masked copy from the production zone to any desired environment
2. **Expand.** Move to synthetic data that needs no production access at all. It is generated from metadata, engineered for each use case or model training objective, and saved as reusable, auditable instruction sets.
3. **Scale.** Use AI to simplify how data is requested and to orchestrate complex data flows. AI interprets requirements; deterministic generators produce the data. No hallucinated records, and no sensitive data sent to a model.

For the CISO, the result is a smaller attack surface, narrower audit scope for lower environments, governed and explainable data for AI, and one platform with role-based access and full audit trails in place of scattered scripts and copies.

The silent multiplier

One production dataset becomes many copies, and one protected environment becomes many unprotected ones. Developers, testers, analysts, data scientists and now AI platforms all ask for realistic data, and the fastest answer has long been a copy of production.

The industry data shows how routine this has become:

Production data spreads far beyond production

Share of organizations (or developers) in recent industry surveys



Another 39% of organizations hand developers subsets of production, and only 7% rely primarily on non-production data sources (Redgate, 2025). Each copy carries the full sensitivity of the original, usually with weaker controls, broader access and less monitoring. As GenRocket puts it: *"Protecting production is no longer the hard part. Protecting everywhere production data ends up is."*

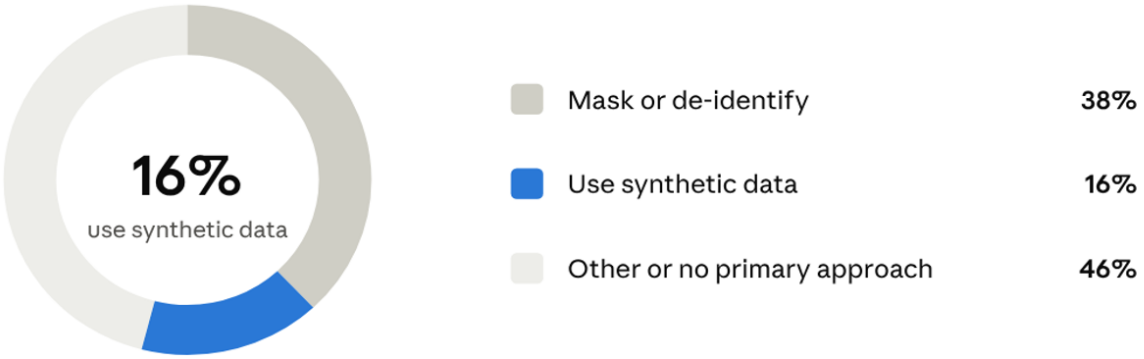
The consequences land on the security program: a larger attack surface, more systems in audit scope, more people with access to regulated data, and more places where a misconfiguration becomes a reportable breach.

Why masking copies is not enough, and why AI raises the stakes

Most organizations already mask some non-production data, but few have moved beyond it.

Only 16% use synthetic data for sensitive dev and test data

Primary approach to sensitive data in development and test, share of organizations



Even where masking is in place, the common approach has structural gaps:

Common practice	The gap for security
Copy production, then mask the copy	Sensitive data has already left production before it is protected
Algorithmic masking or obfuscation	Can break the relationships applications depend on, and reduces re-identification risk rather than removing it
Periodic refresh cycles	Every refresh repeats the exposure, and copies accumulate between cycles
Masked production as the only test data	Tests only what production already contains: no edge cases, negative paths or future states
Separate handling for unstructured data	Documents, files and images are often left out of masking entirely

AI multiplies the exposure. Coding assistants, test generators and AI agents now read and write data across the software lifecycle. About 70% of organizations see significant risk in using proprietary or customer information for AI training (Cisco 2026 Data Privacy Benchmark Study). Generative models introduce a second problem: data they generate is probabilistic. It can hallucinate values, cannot be reproduced run to run, is hard to audit, and drives unpredictable token costs.

The question for the security program is no longer only how to protect production data after it is copied. It is whether it needs to be copied at all.

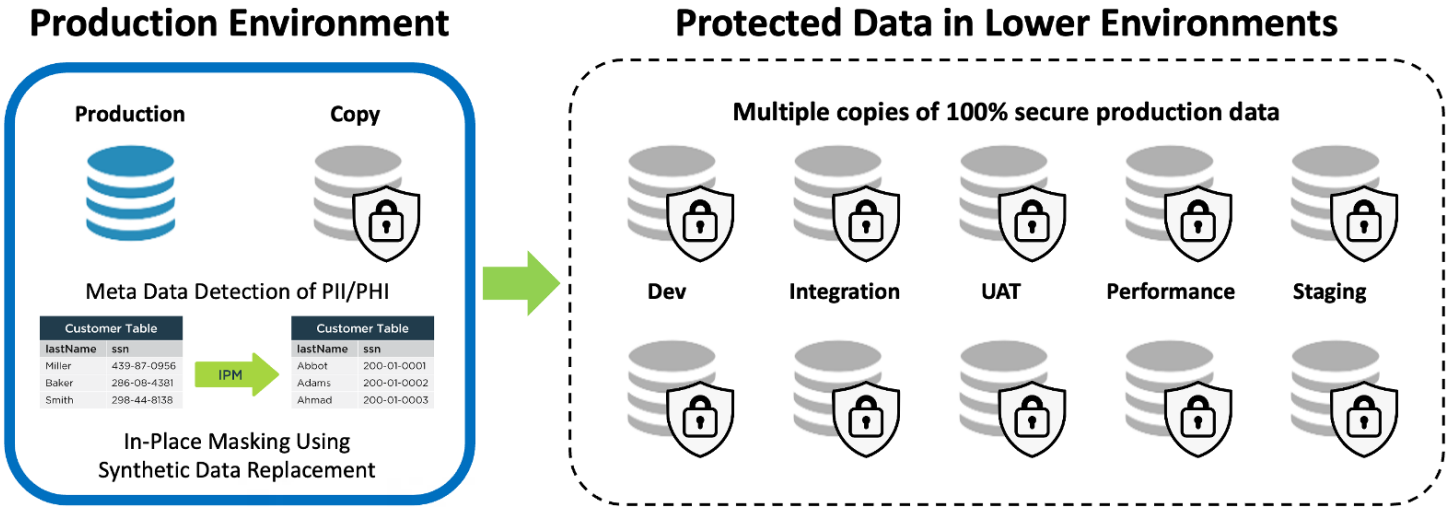
GenRocket's approach to protecting production data

GenRocket starts by securing production data where it lives, before any copy is made. A complete production data privacy plan combines four capabilities:

CAPABILITY	WHAT IT PROTECTS	HOW IT WORKS
In-Place Masking	Entire structured databases	Masks the complete database inside the production environment. Metadata identifies sensitive data across the schema, and each sensitive value is replaced with a deterministic synthetic value rather than algorithmically scrambled or obfuscated.
Data Subsetting	Data volume and spread	Creates smaller, purpose-specific datasets, so entire production databases are not distributed downstream when a slice will do.
File masking	Files (e.g., XLSX, CSV, JSON, XML)	Synthetic data replacement for sensitive data in files
PII/PHI redaction	Documents and images	Identifies and redacts sensitive content in unstructured data before it is distributed.

This approach differs from conventional masking in four ways that matter to a security program:

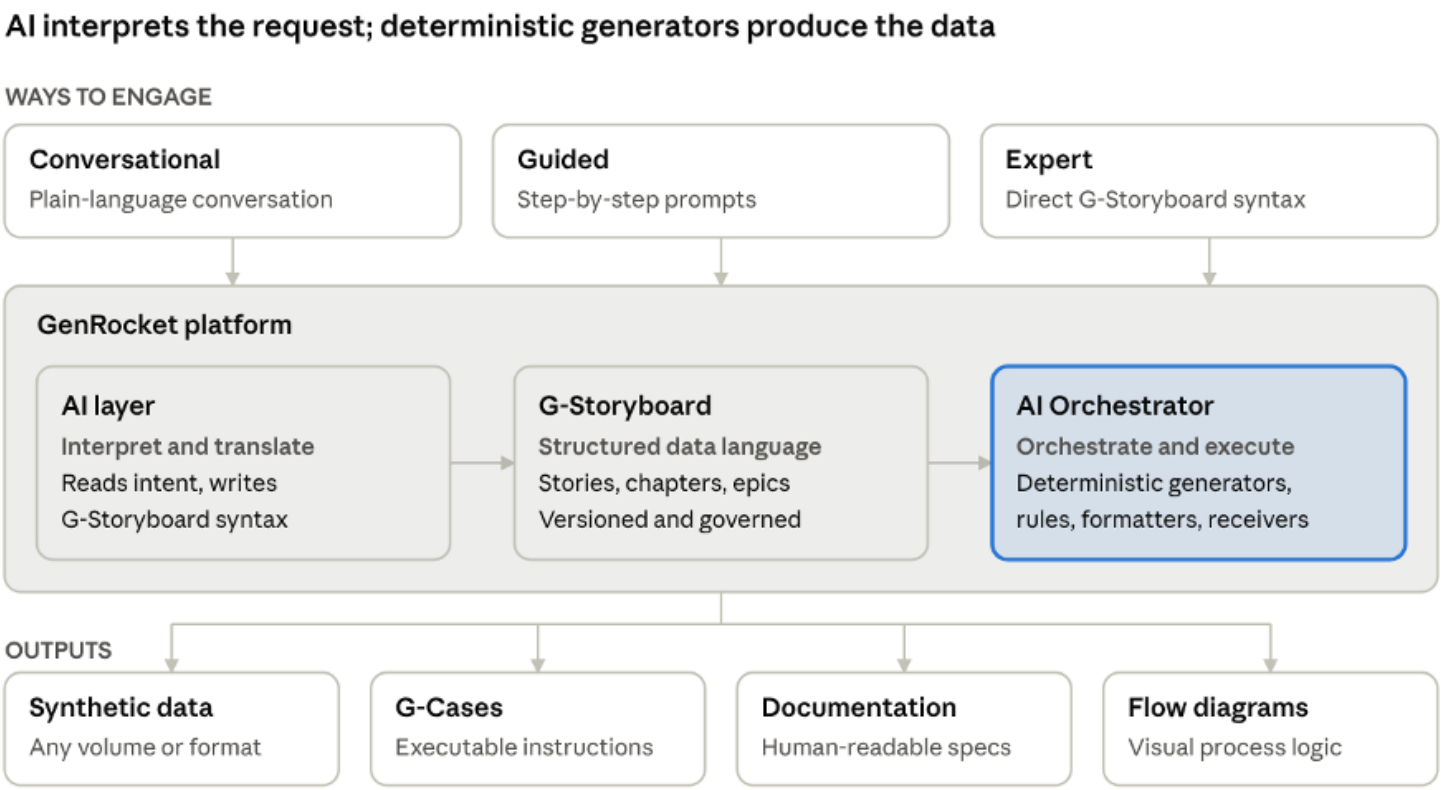
- **Protected before it moves.** Masking happens inside production, so sensitive data never reaches a lower environment unprotected, by design or by accident.
- **Replaced, not obscured.** Synthetic replacement removes the real values instead of disguising them, which reduces re-identification risk.
- **Usable by applications.** Deterministic replacement preserves the characteristics and relationships applications depend on, so teams have no reason to work around it.
- **Complete coverage.** Full databases, purpose-built subsets and structured or unstructured files are covered under one plan.



In-Place Masking solves today's exposure, and it is the first phase of GenRocket's path. But masked data is still derived from production, and it can only test what production already contains. The next step is to stop depending on production data altogether.

GenRocket's synthetic data approach - metadata first, deterministic by design

GenRocket starts from the blueprint, not the records. It uses a system's schema and metadata wherever possible to learn its structure and data profile. It reads production data only for two operations, In-Place Masking and advanced statistical profiling, and in both cases the data stays inside the production environment. **Sensitive data never moves to a lower environment.**



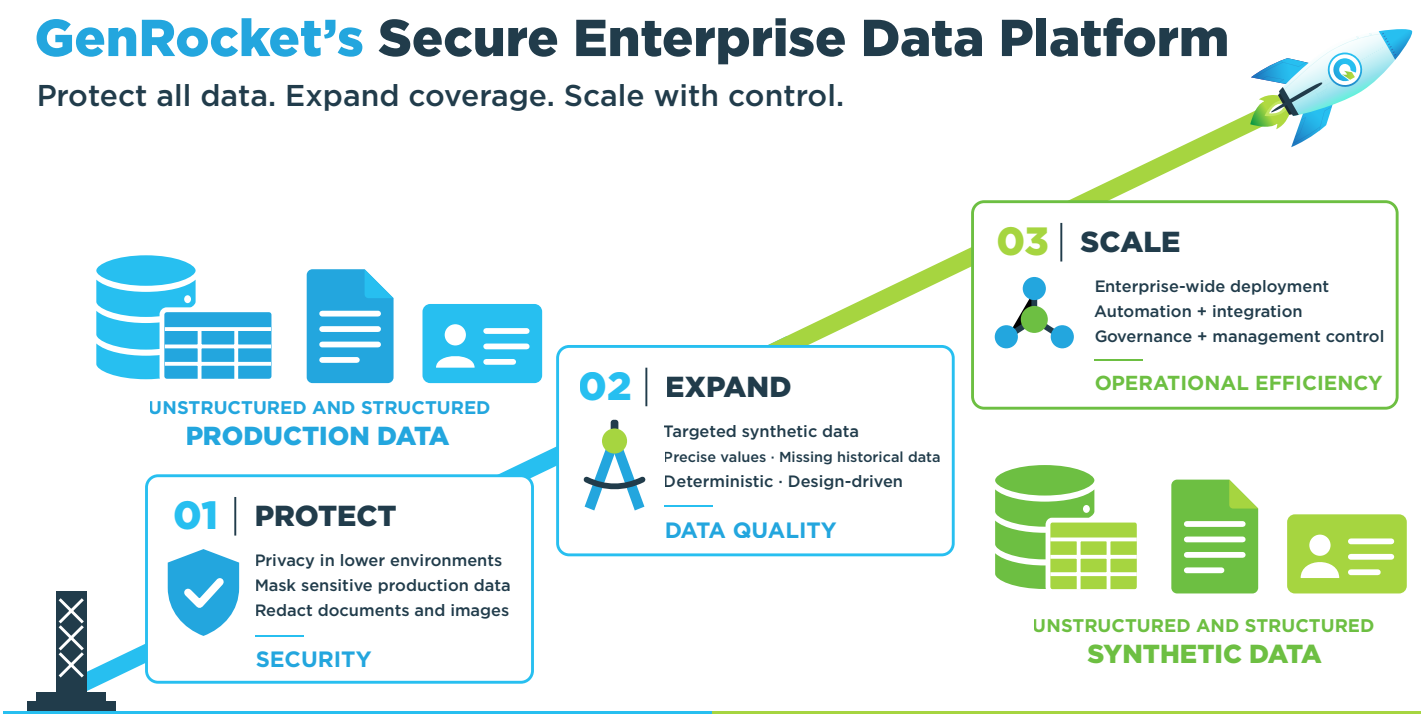
From that blueprint, deterministic generators produce data that simulates production with none of its sensitivity. Three properties matter to a security program:

- **Deterministic, not probabilistic.** The same design produces the same data every time. Every dataset traces to a defined, versioned design, so you can account for what exists, who produced it and why. GenRocket's data generation method is patented (US Patent 9,552,266 B2).
- **Engineered, not mined.** Teams design the exact conditions they need, including edge cases, negative scenarios and future states, instead of hoping production contains them.
- **Simulate or augment.** Synthetic data can stand in for production entirely or augment historical datasets to fill gaps and amplify rare or desired signals, for more complete test coverage and fairer, more accurate AI model training.

The platform supports more than 750 data generators and 125 data formats, generates millions to billions of rows of data in minutes, and is used by more than 50 Forbes Global 2000 companies.

The evolutionary path: Protect, Expand, Scale

GenRocket is a path, not a rip-and-replace. Each phase delivers security value on its own and prepares for the next. In GenRocket's words: *"**Protect** reduces risk. **Expand** improves data quality. **Scale** raises operational efficiency."*



Evolve from Production Data Protection to a fully Synthetic Enterprise.

Protect and provision any kind of data	Generate data for any test or training objective	Enable access for every kind of user
Production or synthetic · Structured or unstructured	Software testing · AI training · Model fine-tuning	Novice · Developer · Data architect

Phase 1: Protect

The first priority is to make all production data secure and private, and protected from deliberate or accidental breach. GenRocket In-Place Masking masks the complete database while the data is still in the production environment. PII profiling helps identify the sensitive data, which is replaced with deterministic synthetic values rather than algorithmically masked or obfuscated data. The relationships and characteristics applications depend on are preserved, so synthetically replaced PII data is still accurate for testing. Data Subsetting provides smaller, secure, purpose-specific datasets, and File Masking synthetically replaces PII in files while maintaining referential integrity. The Unstructured Data Accelerator (UDA-Redact) extends PII protection to documents, files and images.

For the CISO: no sensitive data leaves production, by design or by accident.

Phase 2: Expand

The Synthetic Data Generation architecture is based on metadata. Data can be engineered based on a data model and generated with referential integrity within an application or across a complete IT ecosystem. Data is designed and orchestrated for each specific data use case or model training objective. Each design is captured as an instruction set (a G-Case) that is reusable and cataloged in a data library (G-Case Library), so the same governed data can be provisioned again and again.

For the CISO: whole categories of work stop depending on production data, and every dataset is explainable and auditable.

Phase 3: Scale

Finally, AI is brought in as a vehicle to simplify data provisioning and orchestration. Users describe what they need; AI translates the request into a detailed specification, and deterministic data generators accurately and reliable deliver the data. The platform orchestrates complex data flows for systems integration testing and transaction processing test and validation, at any scale, and delivers data to pipelines and AI agents through APIs and the Model Context Protocol (MCP).

For the CISO: self-service provisioning across the enterprise, without reopening the door to production data.

The customer destination is what GenRocket calls the Synthetic Enterprise™: organizations "provision accurate, governed data that each use case requires — without ever exposing sensitive production information."



Responsible AI: the model interprets, the engine generates

GenRocket uses AI to understand what data is needed, not to generate the data itself. The design principle is *AI-Enabled. Rule-Driven. Deterministic by Design.*

AI-enabled, rule-driven, deterministic by design

However a user engages, a data request becomes a data story, a structured, human-readable and version-controlled data specification, before any synthetic data is produced. GenRocket AI Orchestrator takes the data story and orchestrates the data, on-demand, in real time.

Security concern	How the architecture answers it
Sensitive data exposed to a model	No production data ever leaves production. No customer data is stored on GenRocket systems.
Hallucinated or unexplainable data	GenRocket is a deterministic synthetic data platform which generates accurate, predicatable data.
Reproducibility and audit	Each data model is versioned and documented; the same data model specification produces the same data, with referential integrity, every time.
Uncontrolled AI use	AI features are optional; Expert users can bypass AI entirely. Every step in the system is centrally managed and tracked.
Runaway cost	No tokens are consumed to generate data, so cost does not scale with data volume.

Governance and control at enterprise scale

Scattered scripts, shadow tools and one-off copies are where data governance breaks down. GenRocket replaces them with a single, governed data provisioning platform.

Control	What GenRocket provides
Centralized management	GenRocket Organization Admins have full control and visibility into platform use
Role-based access	Defined roles for Test Data Engineer, Developer, Tester, Data Scientist and Org Admin, with project-level permissions
Versioned, cataloged assets	Data projects and designs are version-controlled, categorized and tagged; G-Cases are stored in G-Case Library for reuse. ER diagrams and flow diagrams and documentation are provided for full visibility into the structure of each data project.
Visibility and audit	G-Analytics dashboards track utilization and activity; AI-assisted design steps carry audit trails
Interfaces for every user	Data as code for engineers (APIs, CLI, CI/CD); repeatable test data cases for QA; natural language for casual users; MCP for AI agents
Performance at scale	Multi-threaded, parallel generation for millions to billions of rows of data generated in minutes, so teams don't have to wait for data when there is a slow production refresh cycle.

The security benefit compounds: when governed data is the fastest way to get data, teams stop asking for production copies.

What this means for the security program

Outcomes a CISO can expect

- A smaller attack surface: sensitive data stays in production instead of spreading across lower environments.
- Narrower audit scope for dev, test and AI environments that no longer hold real customer data.
- Fewer people with access to regulated data, and fewer access requests to approve.
- Safer AI adoption: models and agents work from requirements and governed synthetic data, not customer records.
- Full accountability: every dataset traces to a versioned design, with role-based access and audit trails.

Where to start

1. **Assess.** Map every non-production copy of production data and which environments are in audit scope.
2. **Protect.** Mask your highest-risk database in place, with sensitive data synthetically replaced inside production.
3. **Expand.** Choose a first use case or model training objective and build it as reusable, metadata-driven synthetic data.
4. **Scale.** Add AI-assisted provisioning and orchestration across teams and test environments.

About GenRocket

GenRocket provides a Secure Enterprise Data Platform: one platform to provision any kind of data for any use case, from development and testing to AI deployment and agentic workflows. More than 50 Forbes Global 2000 companies use GenRocket to protect production data in place, replace it with deterministic synthetic data, and scale governed data provisioning with AI.

Learn more at genrocket.com.

