Zalingo Data Refinery

Privacy-Safe Synthetic Behavioral Intelligence for AI-Ready Enterprises
Version: 3.0
Date: 4 September 2025
Contact: Joshua | joshua@alita-therapeutics.com
Website: https://zalingo.africa **Parent Company:** ALITA Therapeutics Ltd (London, United Kingdom)

Executive Summary

Zalingo Data Refinery provides high-fidelity, privacy-safe synthetic datasets that closely mirror real-world behaviors, markets, and operational signals—without using or exposing personal data. Our platform is engineered for teams that need large, reliable, and compliant training and analytics data to build and stress-test AI systems, accelerate product development, and run market-scale simulations.

What's different: multi-domain realism; micro-market segmentation at scale; built-in validation and de-duplication; and packaging that goes beyond raw data to include scoring, documentation, and ready-to-use schemas. Delivery options include S3 (Parquet/CSV) and secure file transfer. Zalingo's approach reduces data acquisition risk while increasing downstream model performance and time-to-value.

The Problem We Solve

Modern AI systems demand abundant, diverse, and representative data. Real-world datasets are costly to acquire, fragmented, and wrapped in legal and ethical risk (PII/PHI, consent, bias, and transfer restrictions). Public web data is noisy and often non-compliant for enterprise use. As a result, teams spend disproportionate effort cleaning, de-identifying, and stitching data rather than training models or shipping features.

Zalingo's answer: production-grade synthetic intelligence—data that is generated, validated, and enriched to reflect realistic structure, relationships, and dynamics, so teams can develop and evaluate models without using real user data.

What Zalingo Delivers

- **Behavioral Intelligence (B2C & B2B):** multi-event user and account journeys (views, clicks, carts, purchases, sessions, product interactions) with contextual features and intent/propensity scores.
- **Market & Trend Signals:** curated time-series with seasonality, volatility patterns, and momentum; correlation matrices that link behavioral signals with market dynamics.
- **Domain Packs:** configurable data bundles for Financial Services, E-commerce/Retail, Healthcare R&D, and AI/ML platform teams.
- **Documentation & Samples:** schemas, data dictionaries, quality reports, and quick-start notebooks.

Why Zalingo (Key Differentiators)

- **Realistic by Design:** datasets reflect cross-table and temporal relationships (e.g., actions follow funnel logic; time-of-day and cohort effects are modeled).
- **Quality Controls:** multi-stage validation; cryptographic de-duplication to minimize duplicates; distribution checks against expected ranges; drift monitoring on rolling windows.
- **Actionable Signals:** built-in scoring (intent, engagement, value estimates) enables immediate use in recommendation, targeting, fraud, and forecasting workflows.
- **Packaging for Production:** marketplace-ready deliverables with consistent schemas, sample slices, and GTM-friendly productization.
- **Flexible Delivery:** S3 (Parquet/CSV) or secure transfer; optional refresh cadences (daily/weekly/monthly) with versioned releases.

> **Note:** We avoid storing or processing real PII/PHI. Outputs are synthetic and support GDPR/HIPAA-aligned workflows; customers remain responsible for their end-use compliance.

High-Level Architecture (No Proprietary Details)

Zalingo operates a two-layer refinery model:

1) **Foundational Refinery (Generation Engine)**
Produces large-scale, multi-table relational data across user behavior, enterprise activity, financial/market signals, healthcare R&D surrogates, and macro trend scaffolds. Emphasis on relational integrity (IDs align across tables), temporal realism (seasonality, burstiness, momentum), and cohort dynamics.

2) **Premium Refinery (Intelligence & Packaging)**
Curates, scores, and assembles domain-specific products. Adds valuation heuristics, automated curation of high-intent slices, quality attestation, and enterprise documentation—ready for pilot evaluation or direct integration.

This architecture separates generation from packaging, ensuring scale and consistency at the core while keeping product variants agile and use-case-focused.

Product Catalog (Representative)

1) Behavioral Intelligence Suite
- User & Session tables (B2C) with event streams; propensity/intent scores
- Account & Seat activity (B2B SaaS) with engagement tiers
- Conversion, churn, and LTV scaffolds for supervised training

2) Market & Trend Suite
- Multi-asset time-series (indices, sectors, categories) with derived technical factors
- Behavior-to-trend correlation matrices for signal discovery
- Scenario stress packs for A/B and counterfactual testing

3) E-commerce & Retail Pack
- Product interactions, baskets, orders, returns, and CLV features
- Segments (recency/frequency/monetary, persona archetypes)
- Personalization & dynamic pricing training sets

4) Financial Services Pack
- Risk, fraud, and credit-like behavior scaffolds (synthetic)
- Transaction sequences and merchant/category features
- Anomaly and outlier challenge sets for robustness testing

5) Healthcare R&D Pack
- Synthetic patient journeys for modeling tasks (non-clinical use)
- Age-appropriate diagnosis/procedure scaffolds; intervention timelines
- Strict disclaimers: research/model development only; not clinical advice

> **Customization:** Any pack can be extended (columns, distributions, volumes) to match downstream schemas.

Data Format & Delivery

- **Formats:** Parquet, CSV, optional JSON Lines for event streams
- **Destinations:** S3 bucket or secure SFTP
- **Hosting:** AWS S3 (US East, N. Virginia – us-east-1)
- **Schemas:** normalized (relational) or denormalized (analytics-ready)
- **Refresh:** one-off drops, scheduled feeds (daily/weekly/monthly), or project-based milestones
- **Versioning:** semantic version + timestamp, with change logs and schema diffs

---

Quality, Governance & Security

- **Validation Pipeline:** schema conformance, distribution checks, cross-table referential integrity, temporal pattern audits.
- **De-duplication:** cryptographic fingerprinting to reduce duplicates across drops; dedupe reports per release.
- **Bias & Fairness:** configurable cohort distributions; optional parity checks and adverse-impact probes on request.
- **Security:** encrypted at rest and in transit; principle of least privilege for access; signed delivery artifacts.
- **Observability:** quality scorecards, drift metrics, and delivery SLIs shared with clients.

Common Use Cases & Measured Outcomes

- **Recommenders & Personalization:** faster offline training; improved top-K metrics pre-production.
- **Fraud & Risk Simulation:** expanded edge cases and tail behaviors; better recall at fixed precision in sandbox tests.
- **Churn/Propensity Modeling:** richer negative/positive samples; accelerated feature iteration.

- **A/B & Scenario Testing:** safe stress tests on pricing, demand shocks, or UI funnels without production exposure.
- **Data Engineering Readiness:** populate lower environments with production-like volumes and patterns.

Results vary by stack and objective; we provide evaluation kits to quantify uplift on your metrics.

Engagement Models

- **Starter Evaluation (2–4 weeks):** sample bundle (10–50 MB), schema docs, and benchmarks notebook; optional guided workshop.
- **Subscription Feeds:** fixed schema with scheduled refresh; SLIs for freshness and completeness.
- **Custom Programs:** collaborative design of features/distributions, co-developed scoring, and integration support.

Licensing: value-based licensing with flexible tiers (single-team to enterprise). Pricing available upon request.

Getting Started

1. **Request a sample & schema** → we'll provide a representative slice and quick-start guide.
2. **Run a focused evaluation** against your target KPI (e.g., ROC-AUC, top-K precision, recall at K, lift vs. baseline).
3. **Select delivery & cadence** (S3/SFTP; one-off or refresh).
4. **Proceed to production** with versioned feeds and quality scorecards.

Contact: joshua@alita-therapeutics.com
Website: https://zalingo.africa

Compliance & Legal Notes

- Outputs are **synthetic** and intended for research, development, analytics, and testing.
- Zalingo does **not** include real PII/PHI in delivered datasets.
- Customers are responsible for compliance in their environments and end-use cases.
- Healthcare pack is **not** clinical data or medical advice; not for diagnostic use.

About Zalingo

Zalingo is an Africa-born AI company focused on practical, high-impact data products for global enterprises. **Zalingo is a venture of ALITA Therapeutics Ltd, headquartered in London (UK).** Our Refinery platform pairs scalable generation with an intelligence-first packaging layer, enabling customers to adopt synthetic data safely and effectively for modern AI workloads.

Let's talk.
joshua@alita-therapeutics.com | https://zalingo.africa