WHITE PAPER

Applications of Synthetic Data for the O&G Industry

Modelling rare events, robust test models, and closing data gaps in the operations landscape

Executive Summary

For the Oil and Gas (O&G) industry, data can be its most valuable asset and its greatest liability. This creates an 'innovation' paradox: the sensitive, fragmented, or often non-existent datasets needed for artificial intelligence (AI) development force a choice between advancing operations and protecting them.

This paper argues that this is a false choice. Synthetic data is emerging as the strategic solution—not as 'fake' data, but as engineered reality. By creating mathematically identical, privacy-compliant datasets, O&G enterprises can close critical data gaps to build, test, and deploy AI models faster and more safely. This technology enables stress testing against extreme scenarios without operational risk, secures confidentiality, and significantly reduces data-acquisition costs. This whitepaper outlines how synthetic data engines can streamline processes across the O&G value chain, unlocking the full potential of AI.

Introduction

The promise of AI in the sector often collideswith a stark operational reality: a persistent data deadlock. Across the entire value chain, from upstream exploration to downstream retail, AI models face many challenges in real-time deployment.

The reasons are fundamental. Strict confidentiality rules keep valuable data siloed. Safety protocols constrain its collection. For rare but critical 'black swan' events, like equipment failures, the necessary data may not exist at all. This scarcity starves AI of the high-quality information it needs to learn real-world patterns, delaying deployment.

Synthetic data breaks this impasse. By analyzing the statistical patterns and structure of existing information, it generates new, mathematically sound examples that mirror reality without exposing it.

Case in point, a Middle East O&G producer uses synthetic data to enhance wellbore survey datasets for rare well collision scenarios. The company strengthens its anti-collision algorithms by generating realistic drilling trajectories, improving operational reliability under diverse subsurface scenarios.

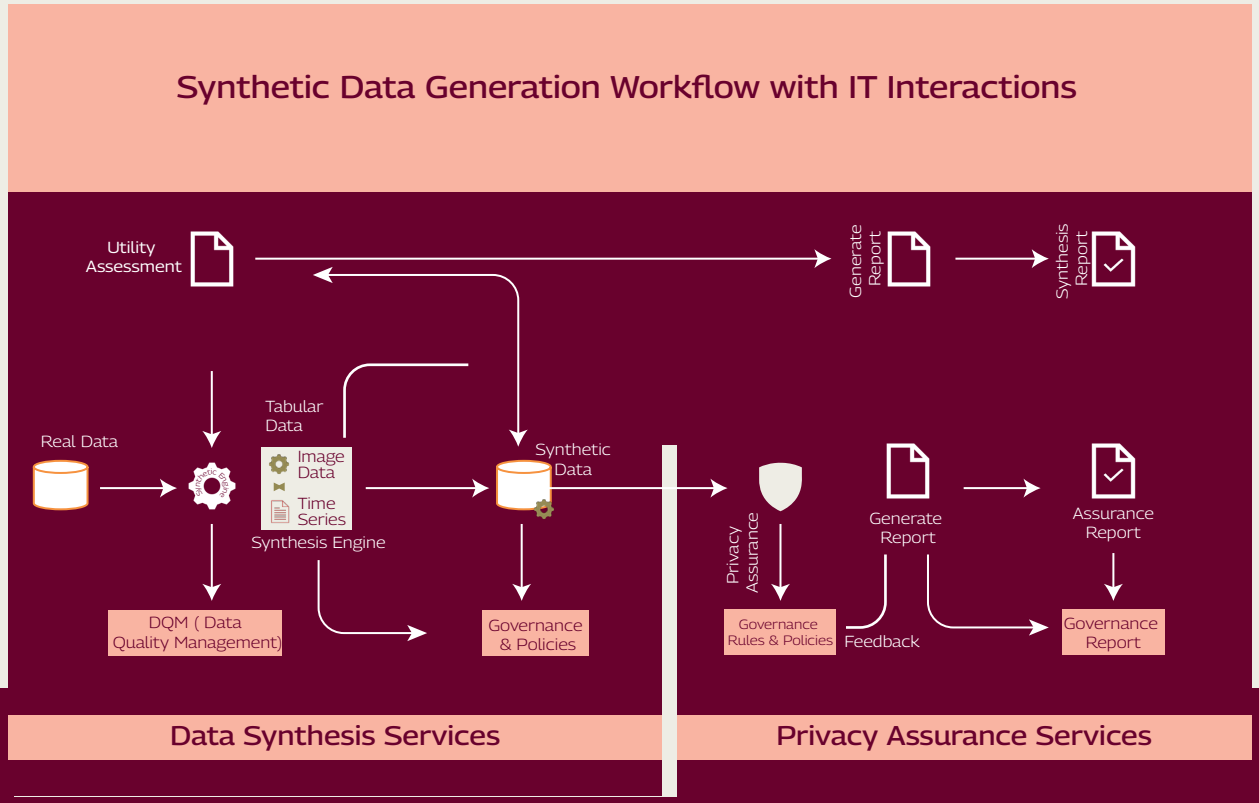


Figure 1: Workflow illustrating the synthetic data generation process

Industry Overview

Both regulators and enterprise organizations are driving this growth, recognizing synthetic data's role in enhancing privacy, scalability, and operational efficiency.

The synthetic data market is rapidly going mainstream. Coherent Market Insights reported its value at approximately USD 485.9 million in 2025. With its application surging across healthcare, automotive, and manufacturing sectors, it's poised to reach USD 3.15 billion by 2032 at a CAGR of 30%.[1]

This growth is driven by a dual push from regulators demanding privacy and enterprises seeking more scalable and efficient AI development. Unsurprisingly, GM Insights notes the AI/ML training division captured over 31% of the synthetic data market in 2024, exceeding USD 2 billion by 2034..[2]

Yet, there is a gap between this global momentum and the O&G sector's current posture. While other industries are scaling, O&G's adoption has been cautious, focused on a handful of pragmatic use cases, like predictive maintenance, Health, Safety, and Environment (HSE) data masking, and software testing. These are domains where core values can be demonstrated without jeopardizing live operations.

Synthetic Data for RPA and Analytics in Contracts

In contract automation, synthetic data addresses the core challenge: how to train systems on sensitive agreements without exposure. It creates entire libraries of realistic, anonymized contracts, and transaction samples that mirror real-world deals without revealing confidential terms or counterparties.

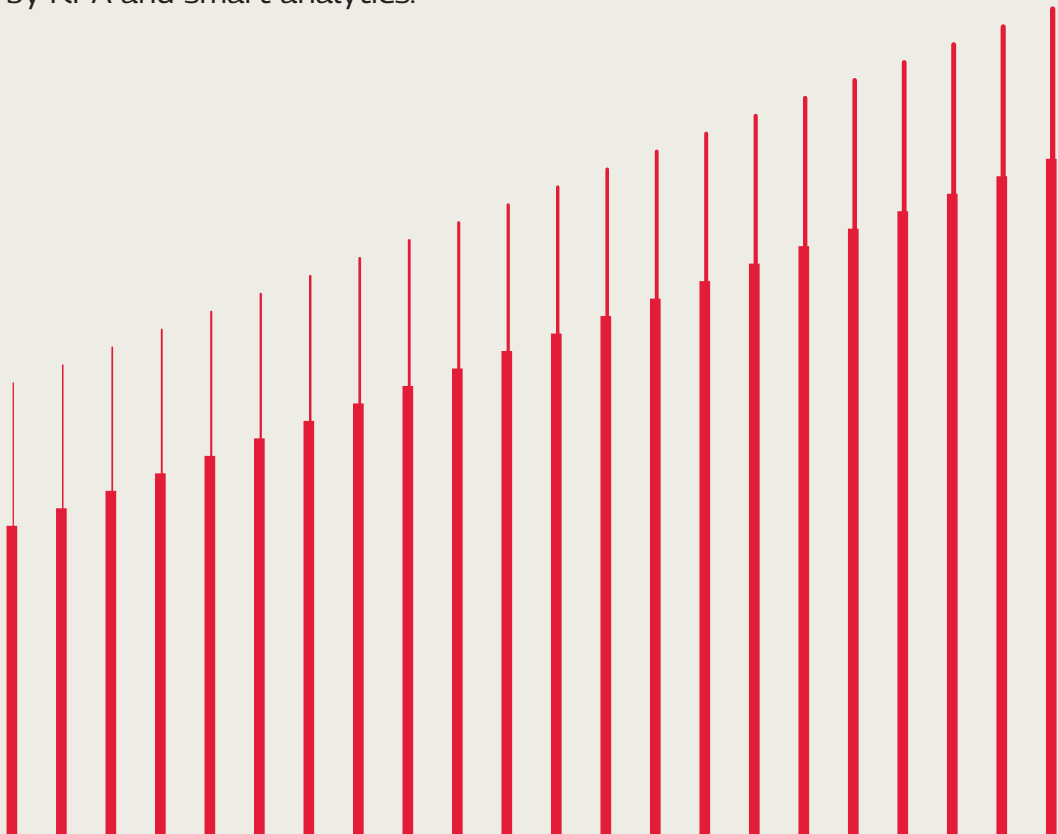
This safe, expansive dataset becomes the training ground for robotic process automation (RPA). Bots can learn to read clauses, match invoices, validate deliveries, and automatically trigger payments or penalties, delivering faster execution with fewer human errors.

The benefits extend beyond simple automation to proactive risk management. Using synthetic data, advanced analytics can simulate a barrage of 'what-if' scenarios, from price index changes and delivery delays to supply chain constraints. These simulations enable automated clauses to be tested before going live. Models learn to handle exceptions, force majeure events, demurrage disputes, ensuring stronger contract compliance from day one.

Model accuracy also improves as training data is diversified across various formats and boundary conditions, sharpening clause extraction, risk scoring, and auto-reconciliation within contract lifecycle management (CLM).

Such an approach advances contract compliance by enabling teams to rehearse regulatory checks and audit trails for synthetic clauses and workflows while maintaining the security of regulated data. Model accuracy also improves as training data is diversified across various formats and edge cases, sharpening clause extraction, risk scoring, and auto-reconciliation within contract lifecycle management (CLM).

Ultimately, synthetic data de-risks deployment, shortens cycle times, maintains contract confidentiality and compliance, and strengthens governance for automated contracts powered by RPA and smart analytics.



Primary Use Cases Across O&G Operations

- **Enhancing Safety Intelligence Without Sacrificing Privacy:** Safety data, from incident reports, safety logs, and near-miss narratives, is crucial for prediction but intensely personal. Synthetic data breaks this conflict. It creates privacy-preserving datasets for training Natural Language Processing (NLP) models and risk predictors without exposing identities or sensitive operational details, thereby directly supporting team morale. This approach, proven in healthcare sector, allows teams to prototype risk detectors and quantify incident precursors across sites, even when real-world data is restricted or fragmented.

- **De-Risking Contract Automation and Compliance:** Smart contracts and LLM review pipelines require extensive testing on realistic but confidential information. Synthetic variants of clauses, counterparty data, and transaction details provide the perfect solution. Teams can validate the accuracy of data extraction and automated controls by emulating pricing, penalties, and regulatory clauses without ever disclosing actual contract terms. This enhances compliance and dramatically speeds up iteration cycles.

- **Sharpening AI Vision for Industrial Safety:** Real-world industrial environments are messy. To make visual analytics reliable, AI models must be trained for difficult conditions. Synthetic images and video clips covering PPE detection, unsafe postures, and more can be generated under varied lighting, blocked views, and clutter. This expands the training set far beyond what's practical to collect, making real-time safety monitoring, training and inspection systems more accurate and dependable in the field.

- **Solving the 'Cold-Start' Problem for New Assets:** New plants, retrofits, and early-life assets lack the historical data needed for effective AI. Synthetic data can bootstrap this learning process. By generating time-series and tabular data that respects physical and environmental constraints, it helps tune anomaly detectors and control strategies long before sufficient real-world data accumulates. This reduces cold-start risk and accelerates the benefits of predictive maintenance and process optimization.

- **Predicting the Unthinkable:** Pipeline and Process Integrity: Major events like corrosion, fractures, or leaks are slow to capture in operational data, making them incredibly difficult to predict. Synthetic datasets can simulate these critical fault conditions. By training on this data, AI models become exponentially better at early detection and risk forecasting. This allows for scheduled predictive maintenance, reducing inspection costs while enhancing reliability across thousands of miles of pipeline.

- **Closing the Gaps in Drone Inspections:** Drone inspections are powerful, but they often leave gaps due to poor weather or equipment limits. Synthetic data fills in these blanks. It can simulate missing visual, thermal, or infrared signatures, providing AI with enriched imagery. This integration accelerates anomaly detection and automates asset monitoring for remote pipelines, tanks, and flare stacks.

- **Optimizing Downstream Processes with Digital Twins:** In downstream refining process, peak efficiency requires navigating immense complexity. Synthetic process datasets can model entire plant operations across a vast range of feedstocks and operating conditions. By training AI on countless virtual process runs, engineers can identify optimal temperature, pressure, and fluid flow regimes. This leads directly to higher yields, reduced emissions, more consistent quality, and faster detection of process deviations or bottlenecks.

Other Domain-Specific Applications

- **Predicting Corrosion in Harsh Environments:** Long-term corrosion data is difficult and costly to acquire, particularly for offshore and subsea environments. Synthetic data solves this by modeling the complex relationships between materials and environmental factors, such as coating type, salinity, temperature, and humidity. This allows AI models to classify corrosion and predict its growth, even in edge conditions rarely observed. The performance of these models is validated on real holdouts and expert-reviewed patterns. It results in smarter inspection planning and more confident lifecycle decisions.

- **Accelerating Brownfield and Greenfield Expansions:** When expanding or revamping facilities, historical data for new units simply doesn't exist. Synthetic sensor streams can simulate everything from throughput and transients to alarm bursts and startup behavior, allowing teams to prepare control logic and deployment plans offline. This digital rehearsal before migration to production significantly reduces cold-start risk and shortens commissioning timelines.

- **Optimizing Customer Demand and Retail Operations:** Understanding customer demand requires analyzing usage patterns without violating privacy. Synthetic meter profiles, conditioned by weather and customer segments, provide a privacy-safe solution. They enable routing optimization, seasonal program design, and tariff simulations while preserving privacy for benchmarking, vendor evaluation, and A/B testing (labeled ‘A’ as the control and ‘B’ as the treatment or variant; to see which performs better on selected metrics). These profiles help test demand-response triggers under atypical weather without exposing Personally Identifiable Information (PII).

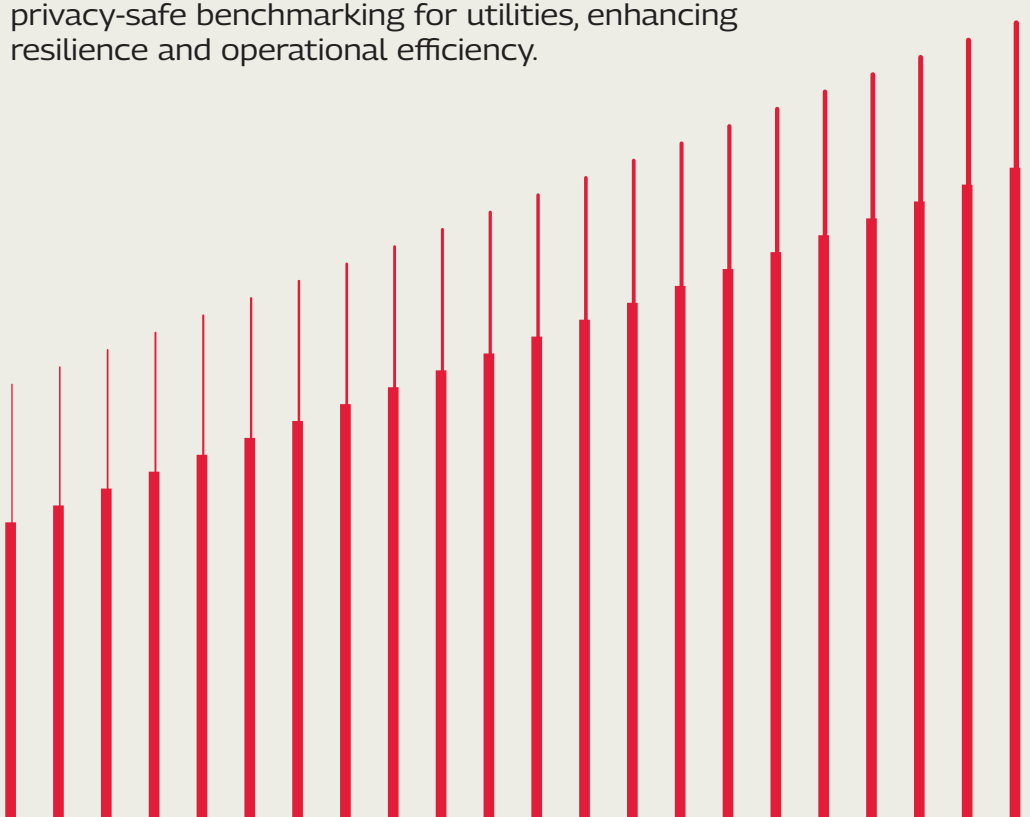
- **Building Resilient IT and AI Systems:** IT systems and AI models must be tested against worst-case scenarios, but using live production data is too risky. Synthetic datasets address this by safely mimicking operational states such as outages, latency, and burst loads. This data-driven approach allows teams to safely validate software fixes, integrations, and model behavior in a non-critical environment, significantly enhancing the robustness of end-to-end pipelines and reducing failures before changes go live.

- **Creating the Virtual Flow Meter:** Physical flow meters are expensive to install and maintain in many O&G installations. A virtual, software-driven alternative can be made using synthetic data that replicates real-time flow conditions from actual sensors and production data. These virtual meters provide continuous, accurate software-driven forecasts of Oil, Gas, Condensate, and Water output for each well. This allows operators to optimize production and reduce maintenance costs, replacing costly hardware with intelligent software.

- **Powering Digital Twins for Predictive Maintenance:** Digital twins built on synthetic operational datasets allow for thorough simulation of wear, tear, and rare fault states. They accurately predict equipment failures and recommend timely interventions before they happen, reducing downtime, extending asset lifespan, and optimizing maintenance cycles across the energy value chain.

- **Characterizing the Subsurface with Greater Certainty:** Reservoir modeling is fraught with many uncertainties. Synthetic data brings new perspectives by simulating infinite pressure changes, fluid movement, and depletion over time, even where historical field data is incomplete. This allows for a much richer characterization of complex subsurface environments, improving reservoir management strategies and more accurate long-term yield forecasts.

- **Modeling Real-Time Energy Demand:** To ensure grid resilience, utilities must accurately model real-time energy consumption without compromising customer privacy. In practice, synthetic datasets, like those generated for certain utility platforms, mirror real-time gas and electricity usage across diverse profiles and weather scenarios. This enables timely demand prediction, supports energy optimization, and provides privacy-safe benchmarking for utilities, enhancing resilience and operational efficiency.



Technical Framework and Tools

An enterprise-grade synthetic data framework is built on two pillars: a sophisticated generative toolkit and uncompromising digital rigor.

The process begins with a suite of generative models. Generative Adversarial Networks (GANs) are deployed to create high-quality images and complex tabular data. For time-series information from Supervisory Control and Data Acquisition (SCADA), telemetry, or meters, Variational Autoencoders (VAEs) and auto-regressive models excel at capturing complex patterns and representations.

But algorithms alone are not enough. Realism is enforced through lightweight simulators and rule-based refinements that reflect physical principles, safety constraints, and equipment specifications directly into the data.

Finally, the entire pipeline is governed by strict engineering practices like containerization and version control. Meanwhile, meticulous logging of seeds and parameters ensure every dataset remains fully reproducible.

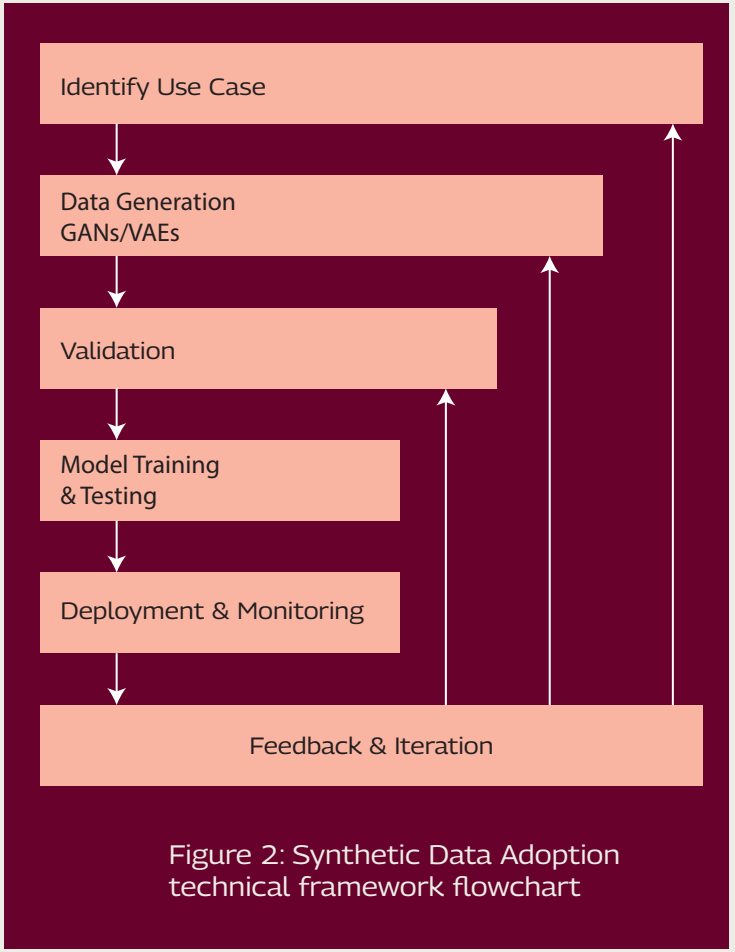


Figure 2: Synthetic Data Adoption technical framework flowchart

A rigorous verification process is critical to establishing the trustworthiness and utility of synthetic data. Building trust follows a methodical approach:

- Conduct automated assessments to evaluate data accuracy
- Compare basic statistics and patterns against the original datasets
- Perform comprehensive privacy evaluations
- Most importantly, test the usefulness of models by training them on real, synthetic, and a mix of both
- Proceed if performance on real-world test sets (for example, corrosion or fraud recall, or demand forecast error) meets agreed-upon targets

Barriers to Adoption

Despite its significant advantages, adopting synthetic data is not without its challenges. Key barriers include:

- **Difficulty with Rare Events:** A core limitation is that synthetic data cannot invent what it has never seen. It struggles to generate true 'blue moon' scenarios or rare outliers that are absent from the source data, affecting the accuracy of models built for anomaly detection or extreme event forecasting.

- **Long-Term Predictive Complexity:** In long-term predictive models, synthetic data use introduces greater uncertainty and potential error amplification. Small biases or inaccuracies in the synthetic data compound over extended time horizons, compromising forecast reliability.

- **Variable Confidence Levels:** Confidence in synthetic data is dependent on use case. It is highest for tasks like data masking and privacy protection, where the goal is structural replication. Confidence is naturally lower for complex, physics-based predictions (like reservoir simulation), where combining data sources can introduce new variables and add uncertainty.

- **Uncertain Training Data Ratios:** No universal ratio for blending synthetic and real data exists. The optimal mix must be validated for each use case to prevent model bias and ensure the integrity of the results.

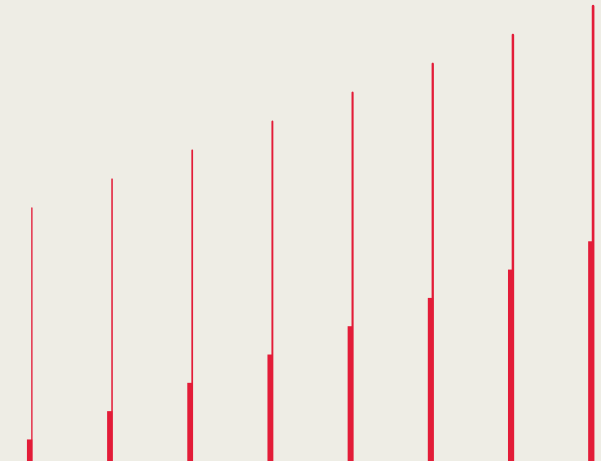
How GenAI and Agentic AI Accelerate the Loop

Generative AI and agentic AI accelerate improvement cycles across O&G operations. The former creates realistic synthetic data for complex scenarios such as extreme weather, corrosion, and supply disruptions.

Meanwhile, agentic AI plans tasks, runs experiments, and tunes the data until it performs effectively in practical tests.

This is not a one-time process; it is a continuous, closed loop. Agents constantly watch for data drift, refresh datasets, maintain libraries of proven use cases, and validate critical workflows like billing, anomaly detection and compliance against real-world benchmarks. The result is a self-improving system that cuts manual effort, reduces iteration time, and maintains model accuracy as business conditions change.

This approach is already delivering tangible value. For a leading North American gas utility, TechM delivered an institutionalized GenAI with solutions that automate complex legal and contractual compliance. The engagement drove significant productivity gains, improved compliance, and reduced operational costs, establishing a scalable platform for future AI initiatives.



Synthetic Data for ESG Outcomes

Embedding a sustainability lens in the O&G industry promotes synthetic data from an efficiency tool into a core ESG enabler. It provides a digital sandbox where teams can safely model and test scenarios for methane detection, flare reduction, and energy efficiency. This accelerates compliance with tightening ESG disclosure standards across regions without exposing live operations or sensitive data.

The real power emerges when combined with agentic automation, which enables recurring 'green regression tests.'

These automated checks validate that any software release or process change does not cause a backslide on critical targets for emissions, water, or waste. This governance extends to the entire operational ecosystem, ensuring consistent environmental performance from partners and vendors alike.

TechM-Aligned Solution to Industrialize Adoption

Tech Mahindra's solution for synthetic data is a capability built from over two decades of hands-on experience across the entire O&G value chain. Our expertise spans the entire O&G value stream, from exploration and appraisal through drilling and completions, production operations, refining, supply and distribution, and retail.

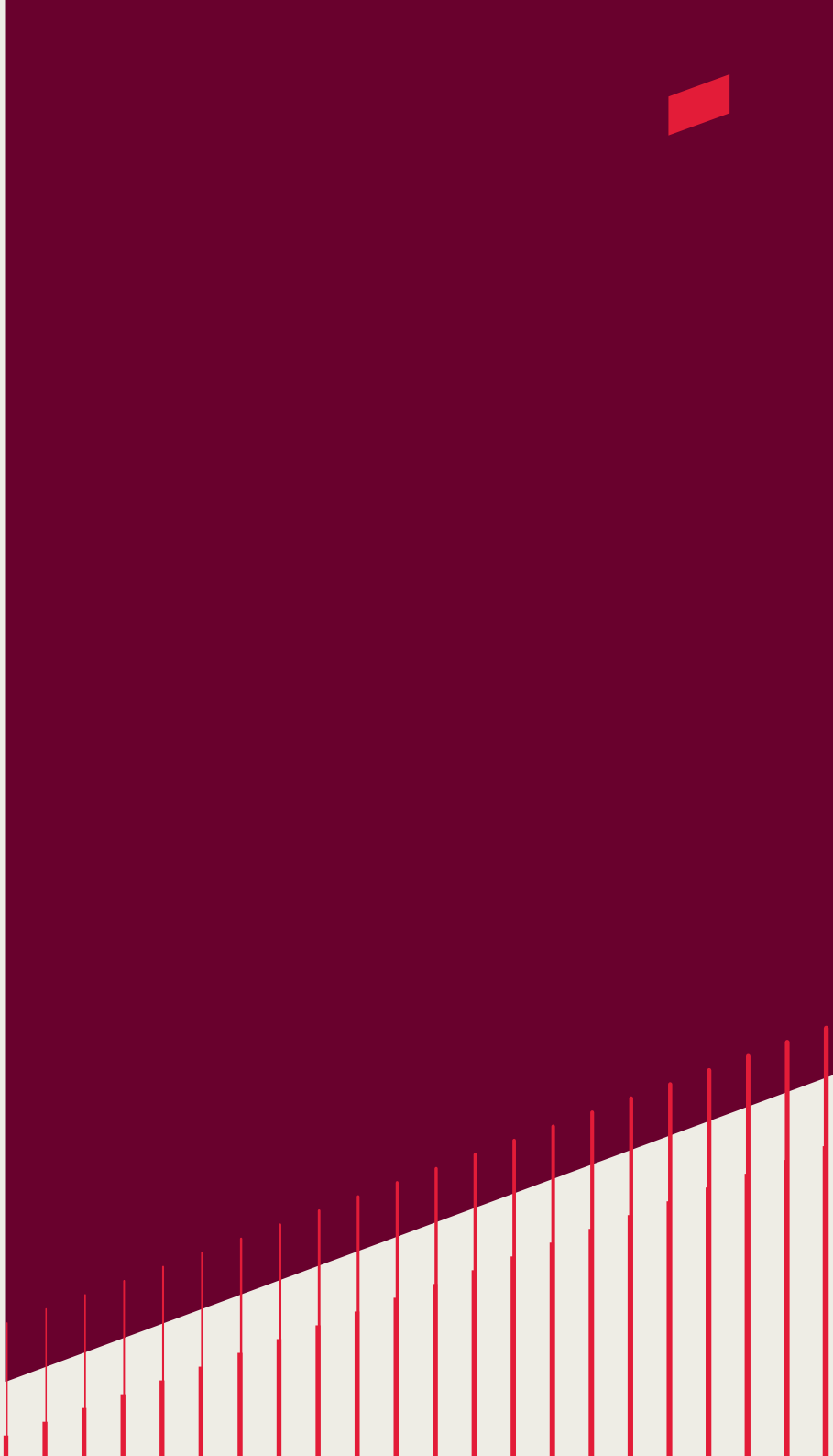
Our offerings include seismic, Production, PVT & other log data management, AI-driven drilling optimization, integrated production modeling, waterflood surveillance, and predictive maintenance, and recently to support setup, manage and run global capability centers (GCCs) for clients.

This deep domain knowledge is embedded in our AI-powered method, which creates realistic, domain-specific datasets at scale while protecting privacy and accelerating delivery. Our solution is designed to be a dependable accelerator for both engineering and AI outcomes, industrializing adoption across two critical areas:

- For Software Testing: Delivers production-like, privacy-safe test packs to validate everything from SCADA systems, master data management (MDM) to billing and analytics pipelines. This allows for rigorous functional, performance, and negative testing in a safe environment.

- For AI Model Training: Supplies balanced, condition-controlled datasets that teach models to handle rare but critical events. This includes everything from corrosion prediction and fraud detection to meter-demand forecasting. By helping models learn from edge cases and validate against real-world data, we dramatically shorten the time-to-train and approve new AI initiatives.

Our approach turns synthetic data from a novel concept into a reliable accelerator that de-risks innovation and accelerates tangible business outcomes.



The Road Forward: A Unified, Determined Strategy

Strong governance of synthetic data, guided by industry-approved standards such as the Institute of Electrical and Electronics Engineers (IEEE) Synthetic Data Activity, best practices, and recommendations from recognized bodies, is essential to ensure quality, privacy, regulatory compliance, and responsible use.

The IEEE, through its industry connections (IC) initiative, has explicitly recognized the need for synthetic data governance that addresses accuracy, privacy, and fairness. A synthetic data IC community has worked toward building the groundwork for eventual standards, particularly regarding privacy and accuracy. Their roadmap suggests creating standard project authorization requests (PARs) focused on these critical areas. As of mid-2025, this process had not yet resulted in a fully ratified, unique IEEE 'synthetic data governance' standard with a distinct number (such as IEEE 7001 or IEEE 7005) in related fields.

Despite the ongoing development of formal standards, organizations can start implementing synthetic data strategically today.

To effectively adopt synthetic data as a unified strategy:

- Start with the highest-friction areas, such as integrity, safety, or billing
- Prove gains rigorously with strict tests on real holdout data and clear success metrics
- Connect early wins by standardizing platforms and adding agent-driven automation for data refresh and testing
- Enforce governance for privacy, lineage, and model risk

This approach transforms scarce, sensitive data into a safe, renewable resource for innovation. The result is faster release cycles, safer experimentation, and more resilient AI models that perform reliably in the demanding conditions of live O&G operations.

Endnotes

1. (2025, September 22). Synthetic Data Market Size Analysis and Share Analysis (2025-2032). Coherent Market Insights. https://www.coherentmarketinsights.com/industry-reports/synthetic-data-market

2. Global Market Insights. (2025, January). Synthetic data generation market size, growth analysis 2025–2034. https://www.gminsights.com/industry-analysis/synthetic-data-generation-market

About the Author

Rajeet Jayan

Principal Consultant (O&G), Tech Mahindra

Rajeet Jayan is Principal Consultant (O&G) at Tech Mahindra. Rajeet brings over two decades of diverse, hands-on experience in the global O&G industry. His expertise was forged in demanding onshore and offshore projects, including high-pressure / high temperature (HPHT), deepwater, and E&P operations globally with leading organizations like Reliance Industries, GSPC, and the Bolloré Group.

Rajeet is a mechanical engineer and holds a master's degree in management (International Business). He has deep knowledge spanning many areas of the hydrocarbon value chain, from drilling operations and risk management to PSC and JV operations, as well as asset and key account management. He is also a passionate advocate for HSSE. Rajeet represents Tech Mahindra at key regional and national energy forums, including ADIPEC, SPE, and IDEC.

About Tech Mahindra

Tech Mahindra (NSE: TECHM) offers technology consulting and digital solutions to global enterprises across industries, enabling transformative scale at unparalleled speed. With 152,000+ professionals across 90+ countries elping 1100+ clients, Tech Mahindra provides a full spectrum of services including consulting, information technology, enterprise applications, business process services, engineering services, network services, custom experience & design, Al & analytics, and cloud & infrastructure services. It is the first Indian company in the world to have been awarded the Sustainable Markets Initiative's Terra Carta Seal, which recognizes global companies that are actively leading the charge to create a climate and nature-positive future. Tech Mahindra is part of the Mahindra Group, founded in 1945, one of the largest and most admired multinational federation of companies. For more information on how TechM can partner with you to meet your scale at speed imperatives, please visit https://www.techmahindra.com/.

www.techmahindra.com
www.linkedin.com/company/tech-mahindra
www.x.com/tech_mahindra