CONTINUA PROCESS SYSTEMS

August 2026

WHITE PAPER

Building the Semantic Data Foundation and Industrial LLM AI Harness

A Controlled Architecture for Reliable LLM in Chemical, Pharmaceutical, and Petrochemical Manufacturing

Authors:

Paul Brodbeck, Anthony DeBiase and Tony DeJohn
Continua Process Systems

Reliable industrial AI depends on two durable capabilities: a semantic foundation that gives plant data operational meaning, and a controlled harness that governs how models and agents use that information.

1. Executive Summary

Continua ProcessIQ™ is an industrial AI platform that combines structured plant data, Knowledge Graphs, large language models, analytical models, and purpose-built AI agents to identify hidden production capacity, reduce downtime, improve process performance, and support better operational decisions.

The most important requirement for successfully deploying AI in chemical, pharmaceutical, and petrochemical facilities is not selecting the largest or most advanced language model. It is creating a reliable semantic data layer that allows AI systems to understand plant information in its proper operational context.

These facilities generate large volumes of data across distributed control systems, process and batch historians, alarm databases, maintenance systems, laboratory applications, workflow tools, quality systems, and production systems. Although this information may already exist electronically, it is often distributed across disconnected platforms, described using inconsistent naming conventions, and missing the relationships required for meaningful analysis.

ProcessIQ addresses this challenge through a data-first architecture built on a Unified Namespace and a Knowledge Graph. Together, these components organize, standardize, and connect plant information before it is made available to models and agents.

ProcessIQ uses this semantic foundation through an industrial AI harness that surrounds the selected language model with controlled retrieval, authorized tools, purpose-built agents, validation rules, source citations, audit trails, and human-review workflows. The harness converts a general-purpose language model into a controlled industrial application.

This combination of a semantic data foundation and controlled industrial AI harness reduces hallucination risk, improves accuracy, preserves traceability to source data, and enables multiple high-value applications to operate from the same trusted architecture.

2. The Industrial AI Challenge

Industrial AI has often struggled to deliver consistent value because plant data is difficult to access, poorly contextualized, and distributed across vendor-specific systems.

A historian may contain process values, an alarm database may contain event sequences, a maintenance system may contain equipment records, a laboratory system may contain quality results, and an operations application may contain workflow information. However, these systems typically do not describe how their data relates to the same asset, production batch, process step, material, product, or operating event.

Without those relationships, a large language model may retrieve individual data points but cannot reliably determine their operational meaning. It may also attempt to infer missing relationships, increasing the risk of incomplete, inaccurate, or hallucinated responses.

These limitations are particularly important in chemical, pharmaceutical, and petrochemical operations, where AI-generated insights must be evaluated against process constraints, safety requirements, operating procedures, quality requirements, regulatory obligations, and engineering judgment.

These requirements cannot be addressed by the language model alone. They require an industrial AI harness that controls what information the model receives, which tools it may use, what actions it may recommend or execute, and how its outputs are verified, documented, and approved. ProcessIQ therefore treats data organization, contextualization, and governed model operation as the first steps in industrial AI deployment.

3. The Semantic Data Layer

The semantic data layer is the core information foundation of the ProcessIQ architecture. It creates a consistent representation of plant assets, processes, materials, batches, events, procedures, quality records, and operating activities.

The architecture is built around two primary semantic components: a Unified Namespace and a Knowledge Graph. The Unified Namespace organizes operational information into a standardized and accessible structure. The Knowledge Graph connects that information through explicitly defined relationships.

Together, these components allow AI agents to follow defined plant relationships and retrieve relevant information with less reliance on probabilistic inference. The semantic data layer supplies trusted context; the industrial AI harness governs how that context is retrieved, presented to models, validated, and used by agents.

Architecture layer	Primary purpose	Representative capabilities
Semantic data foundation	Give plant data consistent operational meaning.	Unified Namespace, Knowledge Graph, historian references, asset and batch relationships.
Runtime harness	Govern how models and agents use plant information.	Identity, access, context assembly, tools, validation, citations, audit records, human review.
Evaluation harness	Test the system before release and after changes.	Versioned test cases, expected evidence, regression tests, authorization tests, accuracy and abstention metrics.

Table 1. Distinct roles within the ProcessIQ™ architecture.

4. Unified Namespace

The Unified Namespace provides a centralized, real-time data layer for plant operations. Selected data from control systems, historians, databases, alarm systems, laboratory systems, workflow applications, and other operational systems is published through connectors into a common structure using consistent naming conventions and contextual models.

The Unified Namespace reduces data silos, standardizes asset and process naming, improves accessibility, simplifies integration, and separates systems that produce data from applications that consume it. This allows individual systems to evolve without requiring every downstream application to be redesigned.

A proof-of-concept implementation may use industrial data-modeling technologies, MQTT, OPC UA, APIs, database interfaces, and historian connections.

The Unified Namespace complements rather than replaces the existing process historian. Historians continue to provide detailed real-time and historical time-series data, while the Unified Namespace provides the structure and context needed to locate and interpret that data.

5. Knowledge Graph

The Knowledge Graph extends the Unified Namespace by representing plant information as a network of interconnected entities and relationships. It can connect equipment, instruments, control modules, process units, materials, batches, recipes, alarms, events, procedures, workflows, maintenance activities, laboratory results, quality records, and production information.

For example, an AI agent investigating a reactor alarm could use the Knowledge Graph to identify the affected equipment, determine which batch or production campaign was running, locate related process data in the historian, retrieve associated operator actions, review laboratory or quality results, and identify relevant maintenance or workflow records.

The Knowledge Graph improves AI performance in three important ways:

- It reduces the need for the model to infer relationships that should be explicitly defined.
- It enables traceability back to authoritative source data and documents.
- It allows multiple AI applications to use the same operational context.

The Knowledge Graph therefore becomes a long-term plant information asset that can support AI agents, analytics, digital twins, optimization tools, regulatory reporting, and future applications.

6. Industrial AI Harness, Tools, and Agents

Definition The ProcessIQ™ industrial AI harness is the controlled software and governance layer surrounding the language model. It assembles relevant semantic context, manages retrieval and tool use, enforces identity and access rules, validates and cites results, maintains audit records, and routes consequential recommendations through appropriate human-review workflows.

Once plant information has been organized within the Unified Namespace and Knowledge Graph, the runtime harness provides the controlled environment through which language models and agents interact with that information. It connects the semantic foundation to authorized retrieval tools, analytical models, workflows, validation services, and user applications.

6.1 Context and retrieval

The harness assembles the context needed to answer a request. It can identify the applicable asset, batch, recipe phase, material, product, time period, and governing procedure before retrieving supporting information. This limits unnecessary data exposure and reduces the likelihood that unrelated records will influence the response.

6.2 Identity, access, and tool control

Controlled software tools may provide access to the Unified Namespace, Knowledge Graph, process historian, batch historian, alarm and event databases, workflow applications, maintenance systems, laboratory systems, quality systems, and other approved sources. The harness determines which tools a user or agent may invoke, what information may be retrieved, whether access is read-only or action-enabled, and how each result is traced to its origin.

6.3 Evidence, validation, and human oversight

The harness can require source citations, distinguish observations from hypotheses, perform data-quality and consistency checks, identify missing evidence, and direct the model to abstain when available information is insufficient. It can also route consequential recommendations for engineering, operations, Quality, or regulatory review before they are accepted or acted upon.

Versioned records of the model, prompt, retrieved context, tool calls, calculations, generated response, and reviewer disposition provide the auditability required to investigate performance and manage change.

6.4 Purpose-built agents

Purpose-built AI agents operate through the harness to support defined use cases. A root-cause-analysis agent, for example, may retrieve process trends, alarm sequences, batch information, operating records, laboratory results, and maintenance history before generating a structured investigation summary.

The architecture can support multiple agents simultaneously, allowing several use cases to operate from the same semantic data layer while applying consistent security, validation, citation, monitoring, and human-review controls.

7. Large Language Model Deployment

The large language model used for a ProcessIQ proof of concept would typically run locally on a dedicated GPU within the plant firewall. For sites requiring a fully local solution, the system can operate without internet access. Plant data, user prompts, retrieved information, and generated responses can remain within the controlled site environment.

Several open-source models may be evaluated across a range of sizes to determine the best balance of accuracy, reasoning capability, response quality, response time, GPU memory, computing requirements, and model-management complexity.

Selected local models may also be benchmarked against approved cloud-based models under appropriate cybersecurity, confidentiality, data-governance, and regulatory controls.

However, the model is only one component of the solution. A smaller model operating within a strong industrial AI harness—supported by a semantic data layer, controlled retrieval tools, validation services, and purpose-built agents—may outperform a larger model that lacks access to structured, contextualized, and governed plant information.

8. Proof-of-Concept Scope

A typical ProcessIQ proof of concept focuses on a selected production unit, process area, or manufacturing operation and integrates data from several operational systems. These may include distributed control systems, process historians, batch historians, alarm and event systems, workflow applications, operations-tracking systems, maintenance systems, quality systems, laboratory systems, manufacturing execution systems, and production databases.

The initial project would typically select two or three high-value use cases that can demonstrate measurable operational benefits. A disciplined proof of concept would proceed through four stages:

1. Select the production scope, users, use cases, success measures, and operating constraints.
2. Build the Unified Namespace and Knowledge Graph needed to supply reliable context.
3. Configure the runtime harness, language models, retrieval tools, analytical services, and purpose-built agents.
4. Use the evaluation harness to test accuracy, evidence quality, authorization, response time, failure behavior, and measurable operational value.

The proof of concept should demonstrate not only that the system can generate useful answers, but that it can do so with controlled access, traceable evidence, appropriate uncertainty handling, and defined human responsibility.

9. Potential Industrial Use Cases

9.1 Real-Time Operational Insight

ProcessIQ allows operators, engineers, scientists, and production personnel to ask natural-language questions about plant operations rather than manually searching dashboards, historian displays, alarm summaries, batch records, laboratory systems, and workflow applications. The platform can retrieve and consolidate process data, alarms, equipment history, operating actions, production context, and quality information into a single traceable response.

9.2 Knowledge Capture and Retrieval

ProcessIQ can organize procedures, historical operating information, prior investigations, shift records, batch documentation, laboratory information, and process data into a searchable knowledge foundation. This helps users identify how similar situations were handled, which procedures applied, what actions were taken, and what outcomes followed.

9.3 Hidden Plant Capacity

ProcessIQ can identify lost production capacity caused by workflow delays, scheduling conflicts, extended equipment hold times, delayed approvals, laboratory-release delays, shift handoffs, and coordination problems rather than physical equipment limitations. The objective is to increase effective plant capacity without significant capital investment.

9.4 Process Optimization

ProcessIQ can combine historical and real-time data, process constraints, equipment relationships, product requirements, and operating conditions to identify opportunities to improve yield, cycle time, equipment utilization, sequencing, resource consumption, and operating consistency. Recommendations remain subject to engineering review, approved procedures, safety requirements, quality requirements, and established operating limits.

9.5 Batch and Campaign Performance

For batch-oriented chemical and pharmaceutical operations, ProcessIQ can compare batches, identify deviations from expected performance, evaluate differences in raw materials or operating conditions, and determine which factors contributed to cycle-time, quality, or yield variations. For

continuous petrochemical operations, the same architecture can compare production campaigns, operating grades, feedstocks, equipment states, and changing process conditions.

9.6 Autoclave Venting

An autoclave venting application may analyze venting duration, pressure, temperature, gas flow, valve operation, batch phase, operator actions, and differences among batches. The objective is to reduce variability, improve visibility, and shorten cycle time while maintaining process, safety, quality, and regulatory requirements.

9.7 Root-Cause Analysis

A root-cause-analysis agent can assemble process trends, alarm sequences, equipment information, operating history, maintenance records, laboratory results, quality data, and previous occurrences. The industrial AI harness can require each potential contributing factor to be linked to supporting evidence, distinguish observations from hypotheses, identify missing information, and preserve human responsibility for the final root-cause determination.

9.8 Process Completion Prediction

ProcessIQ can use current operating conditions and historical batch or campaign performance to estimate process completion times. These predictions may improve scheduling, staffing, downstream coordination, equipment utilization, laboratory planning, material movement, and product-release activities.

9.9 Predictive Maintenance and Anomaly Investigation

ProcessIQ can correlate equipment condition information with process variables, production state, load, operating mode, product grade, feedstock characteristics, and maintenance history. Specialized analytical or machine-learning models may perform the underlying anomaly detection, while the language model retrieves supporting context and explains the findings.

9.10 Downtime Reduction

ProcessIQ can combine process deviations, alarm activity, equipment condition, maintenance history, workflow delays, quality holds, and operator observations to identify developing issues earlier. The objective is to reduce the duration or frequency of unplanned production interruptions.

9.11 Quality and Regulatory Support

In pharmaceutical and regulated chemical manufacturing, ProcessIQ can help retrieve and organize batch records, process deviations, laboratory results, corrective actions, procedures, and equipment histories. The system can support investigations, continued process verification, annual product reviews, process-performance assessments, and preparation of supporting documentation while maintaining required human review and approval. For regulated workflows, the harness can enforce approved-source retrieval, role-based access, document-version checks, citation requirements, audit logging, electronic-record controls, and human review before generated content is accepted or entered into a quality record.

10. Architecture and Data Flow

The ProcessIQ architecture separates the plant's semantic data foundation from the industrial AI harness that governs model and agent interactions. This separation allows language models and analytical methods to change without rebuilding the underlying data context or compromising established controls.

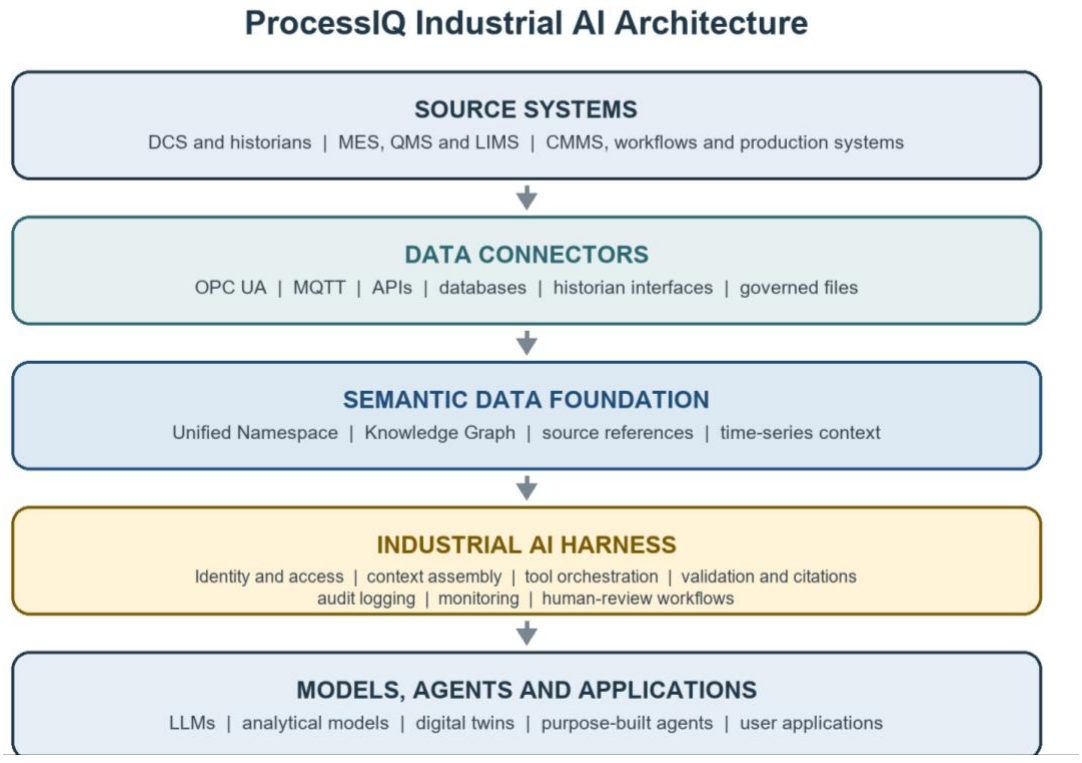

Figure 1. High-level ProcessIQ™ architecture and governed data flow.

Source systems provide operational, production, maintenance, laboratory, and quality data. Connectors publish or retrieve this information through industrial protocols, MQTT, OPC UA, APIs, databases, historian interfaces, or governed file-based integrations.

The Unified Namespace standardizes and contextualizes real-time plant information. The process historian provides detailed current and historical time-series data. The Knowledge Graph represents relationships among assets, materials, batches, campaigns, products, events, procedures, laboratory results, and production activities.

For each user request, the runtime harness can perform the following controlled sequence:

1. Authenticate the user and determine applicable permissions.
2. Identify the relevant asset, batch, process phase, material, time period, and governing context.
3. Select authorized retrieval, analytical, simulation, or workflow tools.
4. Retrieve evidence from the semantic foundation and authoritative source systems.
5. Invoke the selected language model, analytical model, digital twin, or purpose-built agent.
6. Validate the response, attach source citations, identify uncertainty, and record the execution trace.
7. Present the result to the user or route it through the required human-review and approval workflow.

11. Proof-of-Concept Deliverables

A ProcessIQ proof of concept should deliver the following components:

- Integrated and documented connections to selected source systems.
- A defined Unified Namespace model and Knowledge Graph covering the selected scope.
- Historian, document, database, and semantic-retrieval tools.
- A configured industrial AI runtime harness supporting identity, authorization, context assembly, tool orchestration, validation, citations, audit logging, monitoring, and human-review routing.
- An evaluation harness containing versioned test questions, expected sources, acceptance criteria, adversarial cases, authorization tests, and regression tests.
- A local or otherwise approved language-model deployment, selected analytical services, and purpose-built agents.

- Use-case demonstrations, performance benchmarks, cybersecurity and governance documentation, and a final value-assessment report.

The final report should document the use-case objectives, data sources, architecture, model and harness versions, response accuracy, source fidelity, citation completeness, appropriate abstention, response time, computing requirements, user feedback, identified operational improvements, estimated financial or productivity value, limitations, risks, and recommendations for production deployment.

12. Expected Outcome

The objective of the proof of concept is to determine whether ProcessIQ can convert existing plant data into reliable, accessible, and measurable operational intelligence.

The intended result is not simply a conversational interface. It is a structured industrial intelligence platform that allows users to access trusted information, investigate problems more efficiently, identify improvement opportunities, and support better-informed operating decisions.

The Unified Namespace and Knowledge Graph form the durable semantic foundation of this platform. The industrial AI harness provides the durable control framework for model access, tool use, validation, evidence, auditability, monitoring, and human oversight.

The harness also provides model independence. Language models, embedding models, analytical services, digital twins, and agents can be replaced or upgraded while retaining the same semantic context, security policies, validation rules, audit requirements, and human-approval workflows.

For chemical, pharmaceutical, and petrochemical facilities, the combination of a semantic data foundation and governed industrial AI harness is the most important architectural requirement for an effective, scalable, and trusted industrial AI deployment.