COGNITE
AI FOR INDUSTRY

The Cognite Atlas AI™

Definitive Guide to **Industrial Agents**

COGNITE
AI FOR INDUSTRY

The Cognite Atlas AI™

Definitive Guide to **Industrial Agents**

©Copyright, Cognite, 2024 – **www.cognite.ai** →

"According to a recent ARC Advisory Group Digital Transformation, Sustainability and Technology survey, Artificial Intelligence, AI is the most impactful technology for the next five years. Cognite Atlas AI's guide to Industrial Agents is a practical place to start for digital leaders looking to make AI work in complex industrial environments."

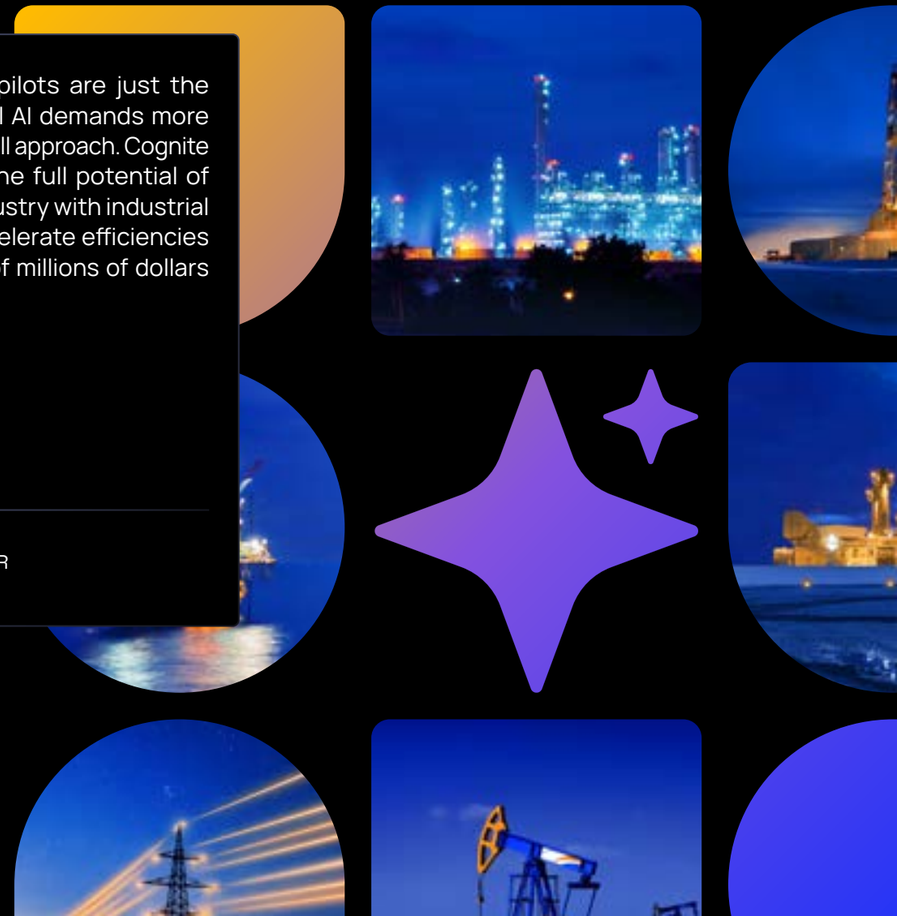
JANICE ABEL
PRINCIPAL TECHNOLOGY ANALYST
ARC ADVISORY GROUP

"Cognite Atlas AI enables us to use AI to enhance decision-making and improve efficiency, like with an industrial agent fine-tuned to understand unstructured technical documentation and Aker BP's equipment hierarchy. By implementing the Document Parser AI Agent, we are streamlining our equipment management process, saving thousands of hours of data-punching, and refocusing our experts on business problems that really matter to the short- and long-term success of our operation."

PAULA DOYLE
CHIEF DIGITAL OFFICER
AKER BP

"Data lakes and copilots are just the beginning. Industrial AI demands more than a one-size-fits-all approach. Cognite Atlas AI™ unlocks the full potential of generative AI for industry with industrial agents that can accelerate efficiencies and generate tens of millions of dollars in business impact."

PAUL GRENET
CHIEF REVENUE OFFICER
COGNITE



Contents

Section 0
Introduction 06

Executive Summary 08

The ~~Four~~ Five Things You Need to Know about Generative AI for Industry 10

Foreword 14

The Data and AI Problem 16
Demystifying Industrial AI Agents: What We Can Learn from Iron Man.....20

Section 1
Making Generative AI Work for Industry 24

Chapter 1
Industrial Agents 26

1.1 The Treacherous Path to Trustworthy Gen AI for Industry 28
1.2 The Path Forward: Industrial Agents 32
1.3 The Challenges of Implementing Industrial Agents.....36
1.4 The Applications and Benefits of Industrial Agents.....38
1.5 Agent Orchestration and Agent Ecosystems.....40
1.6 Does RAG Still Matter?.....42
1.7 From RAG to CAG..... 46

Chapter 2
Large, Small, and Custom Language Models.....48

2.1 Understanding the Difference.....52
2.2 LLMs and Their Application in Operations.....54
2.3 So Why Do We Need SLMs? 56
2.4 Custom Language Models.....58
2.5 Evaluating Large Language Models Usefulness ≠ Correctness..... 60
2.6 Choosing the Right Model with autoLLM.....68
2.7 Performance Benchmarking 70

Chapter 3
Semantic Knowledge Graphs.....72

3.1 Defining Knowledge Graphs.....76
3.2 Knowledge Graphs and Data.....78
3.3 Knowledge Graphs and AI 82

Section 2
The Business Value of AI.....88

Chapter 4
AI Is the Driving Force for Industrial Transformation 90

4.1 Verdantix View: Industrial DataOps in 2024 94
4.2 AI Will Deliver Untapped Value for Asset-Heavy Enterprises..... 102
4.3 Democratizing Data: Why AI-Infused Industrial DataOps Matters to Each Data Stakeholder 104

Chapter 5
Use Cases 108

5.1 Industrial Use Cases Require a System of Engagement.....110
5.2 Cognite Data Fusion®: An SOE to Scale Operational Use Cases..... 114
5.3 Improving RCA with AI Agents and Industrial Canvas..... 122
5.4 Examples of Industrial AI Agents..... 126

Section 3
Tools 130

Chapter 6
Tools for the Digital Maverick 132

6.1 Industrial AI & Data Management Software: How to Avoid Decision-Making Pitfalls When Purchasing..... 134
6.2 Navigating Digital Transformation: A Framework for Success..... 138
6.3 Navigating Digital Initiatives by Using Value as the North Star..... 142
6.4 Data and AI RFP Guide..... 148

Section 0

Introduction

Executive Summary 08

The ~~Four~~ Five Things You Need to Know about Generative AI for Industry 10

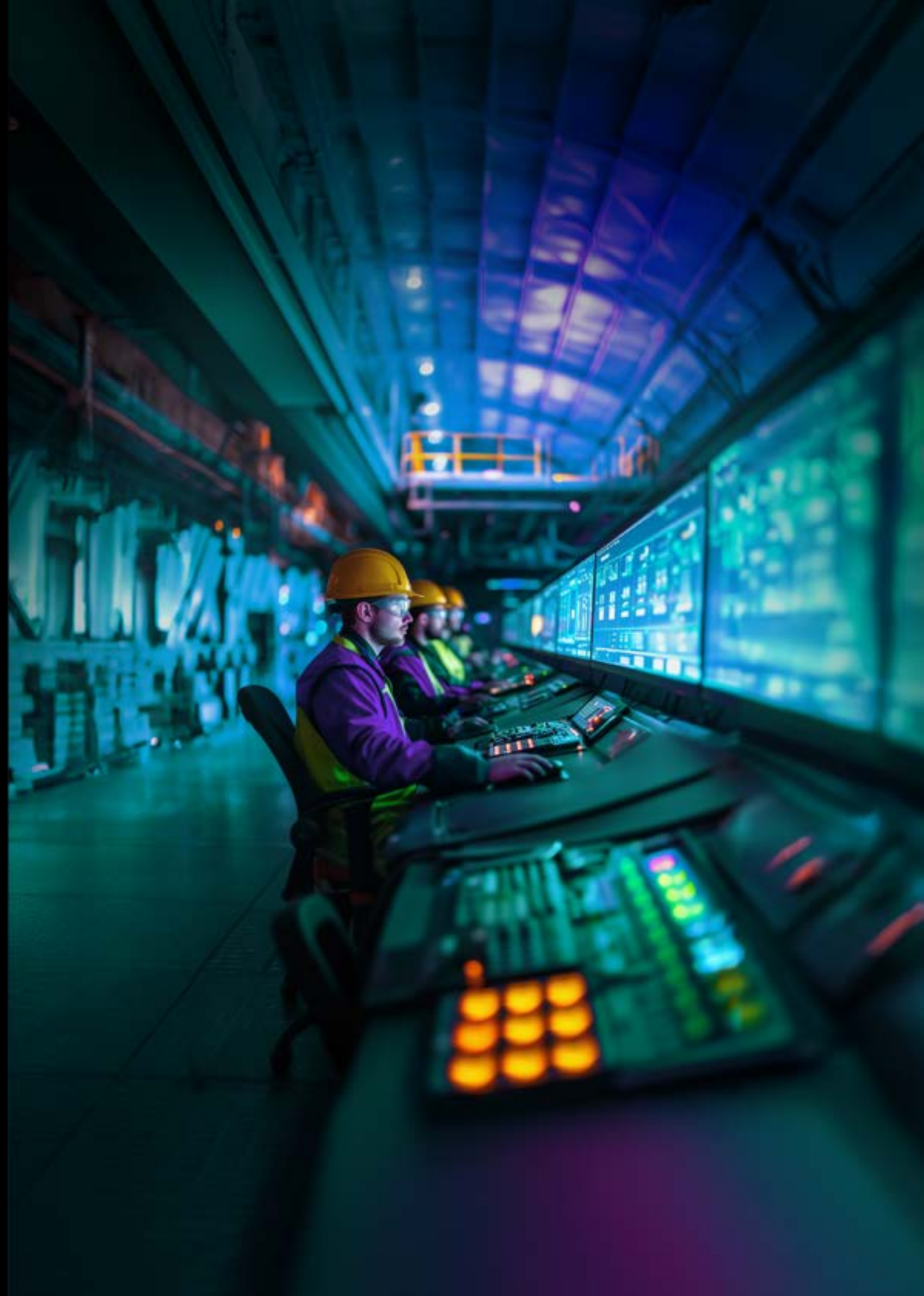
Foreword 14

The Data and AI Problem 16

Demystifying Industrial AI Agents: What We Can Learn from Iron Man.....20

Executive Summary

The ~~Four~~ Five Things You Need to Know about Generative AI for Industry



The ~~Four~~ Five Things You Need to Know

about Generative AI for Industry

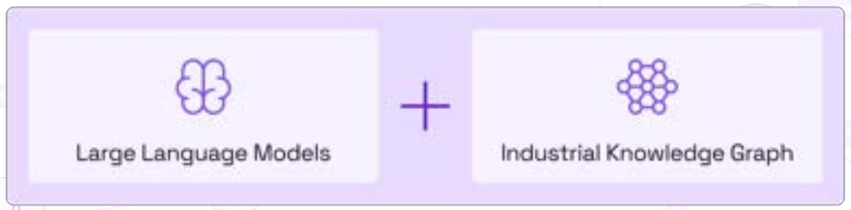
Little has changed in the four key points we've been preaching about generative AI from the beginning, but there is one notable addition in this guide: industrial value is accelerated by industrial agents.

Though really an extension of point four, industrial agents are a significant development that enables industrial organizations to use generative AI to carry out more complex operations with greater accuracy. We will dive into more details on industrial agents in the following chapters but, if you only read one part of this guide, let it be this:

1.

LLMs + Knowledge Graph = Trusted, Explainable Generative AI for Industry

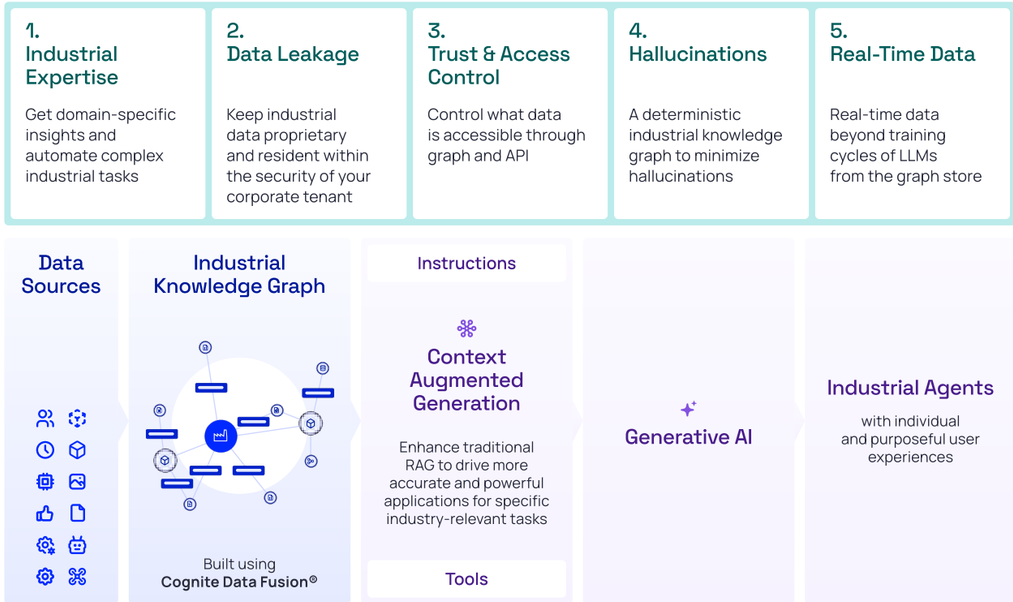
This is the simple formula for applying generative artificial intelligence (AI) in industry. Combining large language models (LLMs) with a deterministic industrial knowledge graph containing your operations data makes your asset performance management intelligent and efficient.



2.

Generative AI for Industry Needs to Be Safe, Secure, and Hallucination Free

And with the formula above, it is. You need a complete, trustworthy digital representation of your industrial reality (i.e., an industrial knowledge graph) for LLMs to understand your operations and provide deterministic responses to even the most complex questions.



3.

To Apply Generative AI in Industrial Environments, the Ability to Prompt LLMs with Your Operational Context Is Everything

This means having a deterministic industrial knowledge graph of your operations, including real-time data. You need a solution that delivers contextualized data-as-a-service with data contextualization pipelines designed for fast, continuous knowledge graph population.

4.

While Generative AI Itself Is Undeniably Transformative, Its Business Value Is in Its Application to the Real-World Needs of Process Engineers, Field Workers, Maintenance Teams, and Other Data Consumers

Innovative AI features are only valuable in a platform that also enables simple access to complex industrial data for engineers, subject matter experts, and data scientists so they can make the right decisions at the right time.

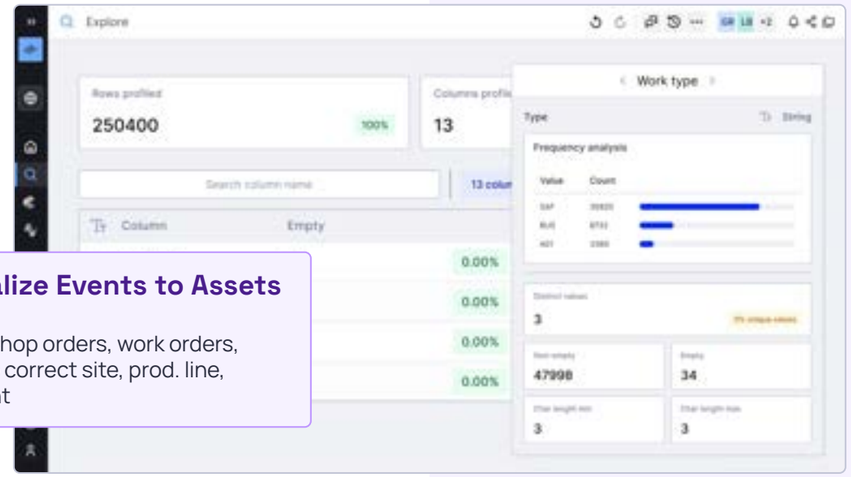
5.

Industrial Value Is Accelerated by Industrial Agents

These tailored, AI-powered applications are designed with an in-depth understanding of industry- and customer-specific terminology, processes, and standards. They utilize algorithms and data models specifically optimized for the patterns and anomalies typical in a particular domain. And, they can be customized to fit the unique workflows and requirements of different organizations.

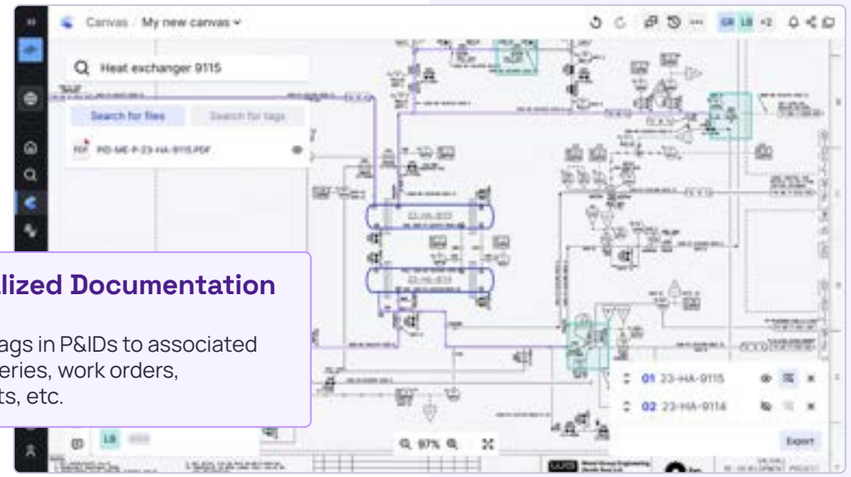
As such, industrial agents can offer more accurate and relevant guidance and can be scaled to accommodate the growing data and complexity of operations as an organization expands.

Industrial agents bring the power of AI and machine learning directly to the challenges and tasks unique to the industry **and** each unique organization. This makes them crucial in improving decision-making processes to help organizations achieve higher productivity, safety, and overall operational efficiency.



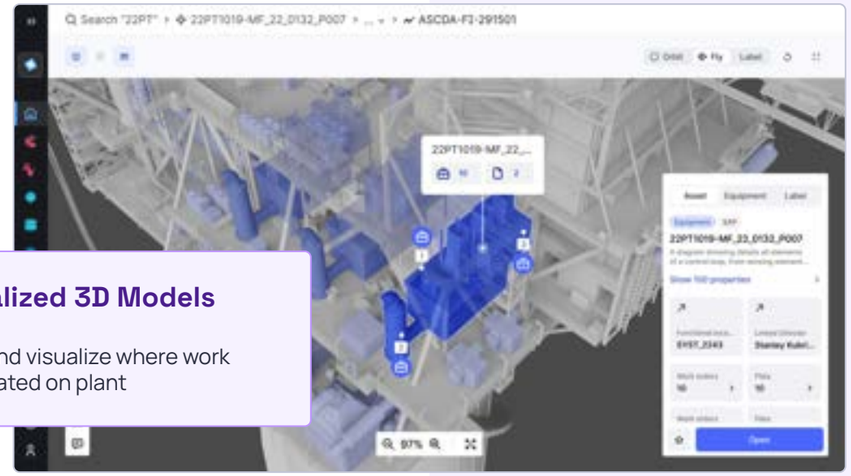
Contextualize Events to Assets

E.g. connect shop orders, work orders, and alarms, to correct site, prod. line, and equipment



Contextualized Documentation

E.g. connect tags in P&IDs to associated assets, time series, work orders, and documents, etc.

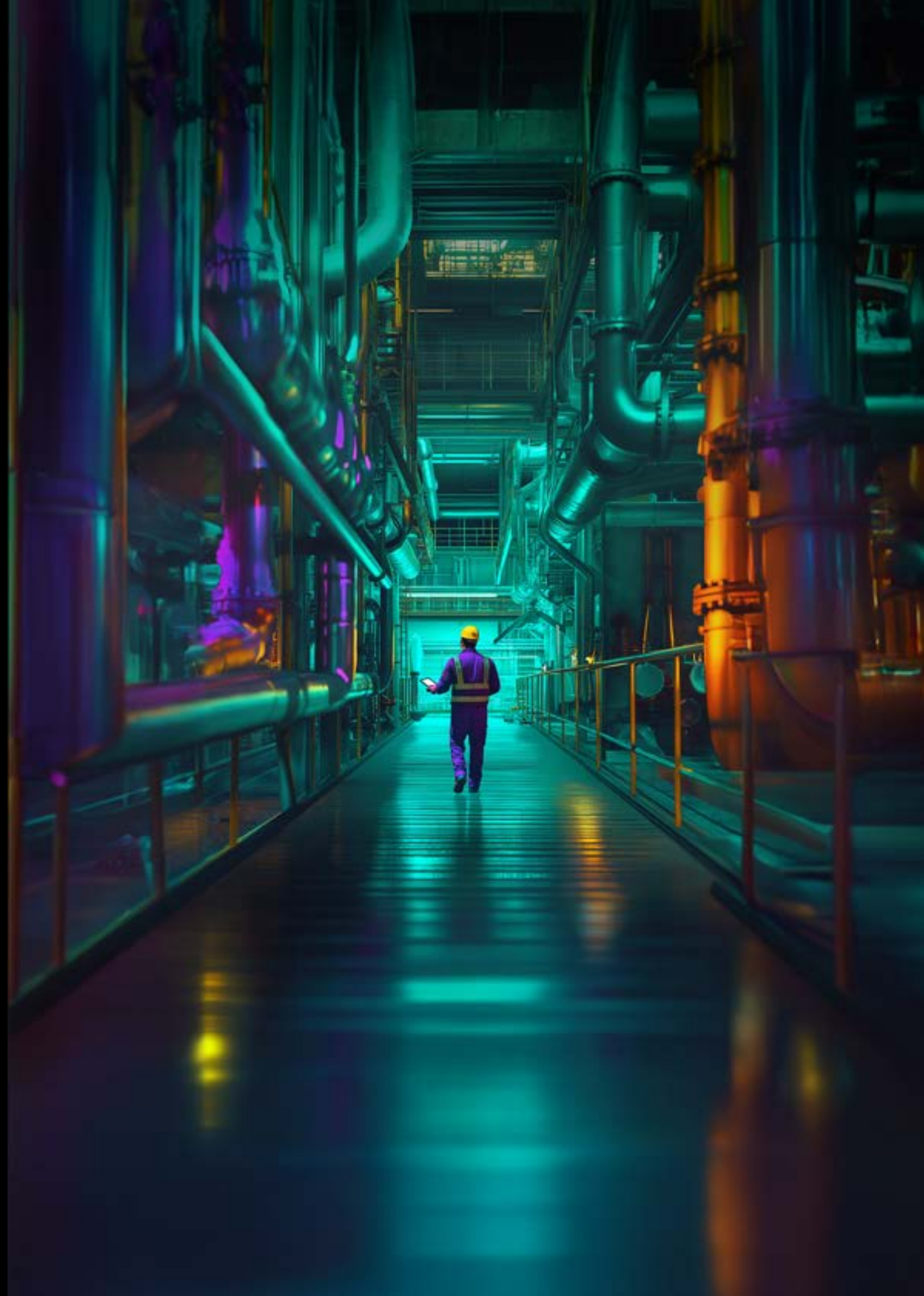


Contextualized 3D Models

E.g. filter for and visualize where work orders are located on plant

Foreword

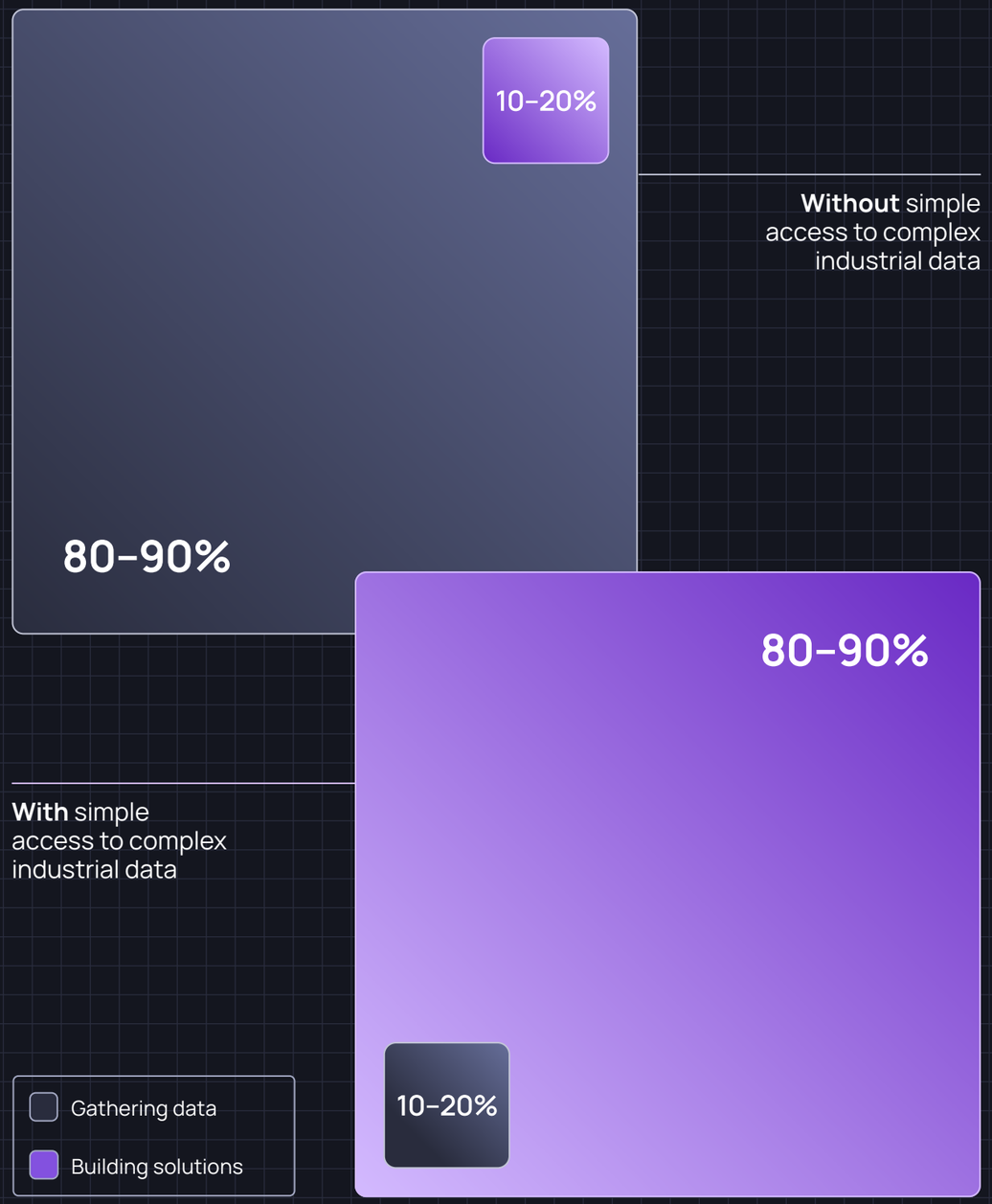
The **Data** and **AI Problem**



The Data and AI Problem

For every one person who can 'speak code,' there are hundreds of others who do not, especially in the industrial environments where there are numerous data types and source system complexity. Subject matter experts, field engineers, and data scientists deserve simple access to all industrial data. This requires a unique way to leverage and apply contextualized data (i.e. data that is enriched with relevant information and relationships, making it more meaningful and useful for analysis and decision-making).

Time Spent on Data Products



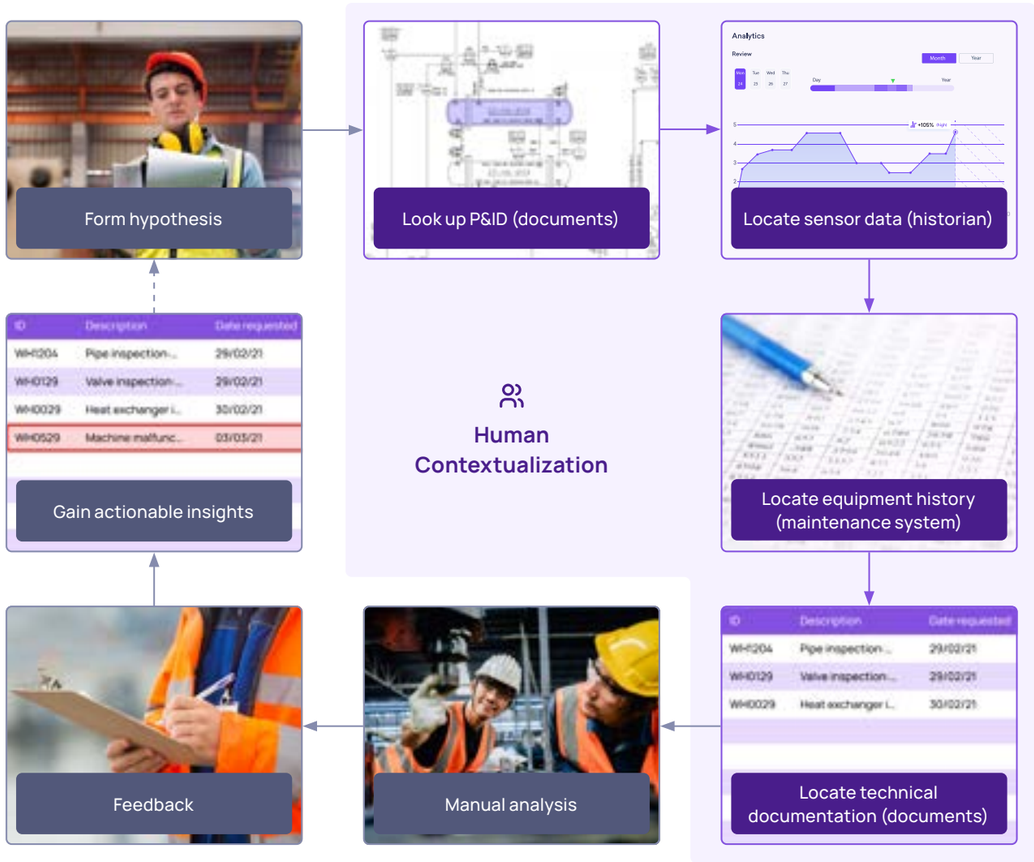
Generative AI is context dependent. While generative AI has tremendous potential to make data easier to explore, understand, and use, answers are often wrong without contextualized data. However, traditional efforts to connect data from systems are manual and time-consuming, and are not capable of managing structured data at scale, much less incorporating the growing unstructured data.

An efficient way must be found to provide generative AI solutions with more context to enable them to provide the right answers in industrial environments. Only then can it be used to optimize production, improve our asset performance, and enable AI-powered business decisions. This is where industrial agents come in.

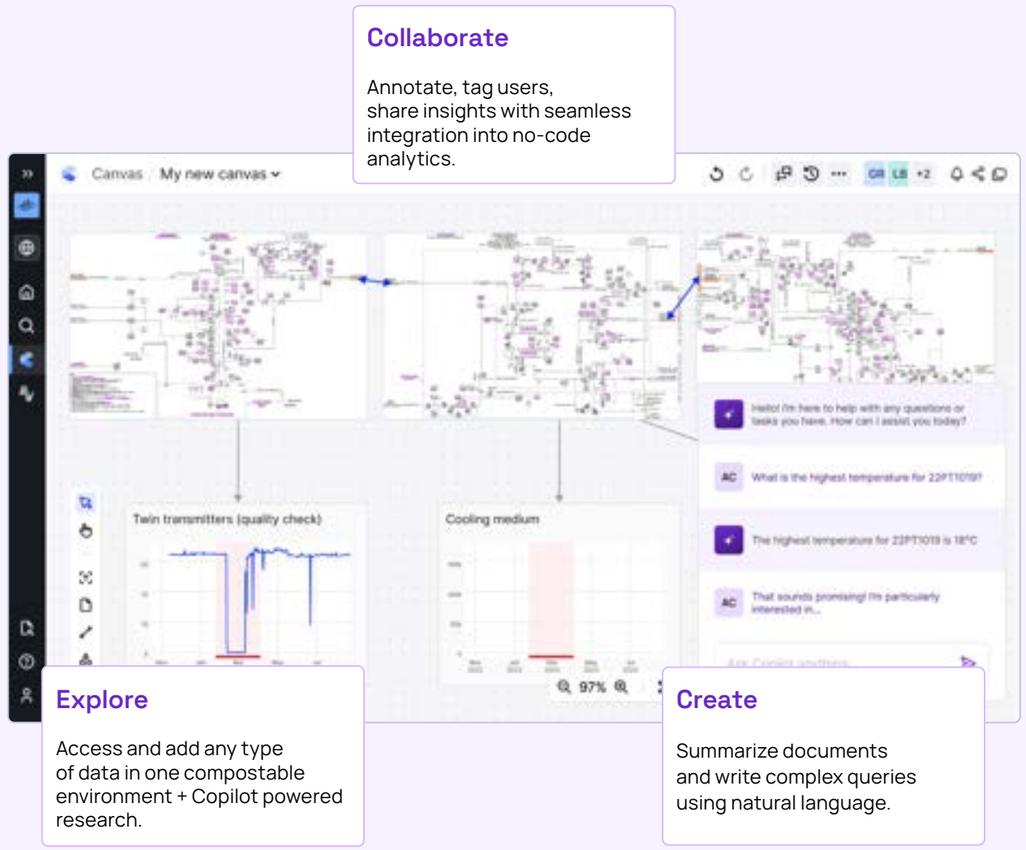
From This → to This

Manual, cumbersome industrial data workflow

No simple access to complex industrial data and insights.



Single workspace for data & analytics powered by AI



Demystifying Industrial AI Agents:

What We Can Learn from Iron Man

With all the hype around generative AI for industry, it seems as if there is a new buzzword almost every day. The latest? "Industrial Agents." But is it really a buzzword? Or is it rather a more accurate depiction of the actual endgame of data & AI for industry?

Few standard definitions of this term exist yet for industry but, put simply, industrial agents perform specific tasks in a **human-like manner** when trained with the right data and when using the right AI model and capabilities.

The operational co-pilots that everyone is talking about or the chatbot you use when trying to rebook your flight are all types of AI agents. They aim to automate or simplify a specific, constrained workflow to improve the user's productivity. However, the agents of today, which use limited pre-programmed logic, are no match for the Gen AI-based agents of the future.

If we take inspiration from the movies, we're getting closer to Iron Man's "Jarvis" assistant - a supercharged intelligent virtual agent that communicates via voice commands and helps Iron Man do his best work. While we're a far cry from this type of omniscient, cross-functional intelligence, the technical building blocks and terminology exist today to start developing specific industrial agents for particular operator domains.

For decades, industrial operators have been trying to use data and AI to optimize production, minimize outage risk, streamline production, and make smarter daily decisions. They've used physical and machine learning models to classify types of asset failure, natural language processing to search for information, and now LLMs to analyze and summarize data and make recommendations. However, with the exception of robotic process automation (RPA) for back-office functions, the **impact on factory-floor operations has been underwhelming**.

Why is this? First, as with any new paradigm, it's taken a bit of time to learn and understand what's required from a technological and process perspective. For example, in the early phases of generative AI, circa 2023, the idea was to go straight to a "Jarvis-based future" with few general agents, with broad objectives trained on large sets of data, structured and unstructured. But this made it difficult to trust and repeat the results due to the inherent hallucinations and other limitations of generative AI.

Fast forward to today —the more realistic scenario is the orchestration of many specific virtual agents, trained on smaller, secure, relevant data sets, designed with intuitive UX to improve workflows definitively.

Despite the early learnings, what's become abundantly clear is that for industrial agents to work and be trusted in industrial domains, they need three things:

1. A domain-specific task or objective
2. Secure, contextualized data for this objective
3. The most appropriate LLM for the task at hand

Until recently, industrial organizations did not prioritize the need for secure, contextualized data foundations, which are critical for training myriad LLMs on relevant data. Today, with support from boards of directors and executives, even legacy industries are investing in teams and technology to bring order and context to their vast amounts of siloed data.

Second, the way users interact with digitally enhanced industrial processes has not been intuitive, making it challenging to **actually improve** the workflow. In flight, if Iron Man was not able to speak conversationally with Jarvis and had to manually look up information with precise terms, his workflow (and mission outcomes) would not be as good. In the field, operator workflows are precise and well-established. Information must be trustworthy and accessible instantly, using handheld devices and simple commands, not lines of SQL code. Technology that doesn't offer dramatic workflow improvements does not get adopted.

Here's where things get even more interesting. Gartner predicts:

> **"Large Language Models (LLMs) will become the preferred interface to enterprise data."**

This means that the effort required for users to access and refine information (once they trust the outcomes) becomes human. So even though an operator may not be able to ask their agents the same breadth of questions as Iron Man could to Jarvis, their interface to answers becomes more human and intuitive than ever before - making it easy to adopt into a workflow.

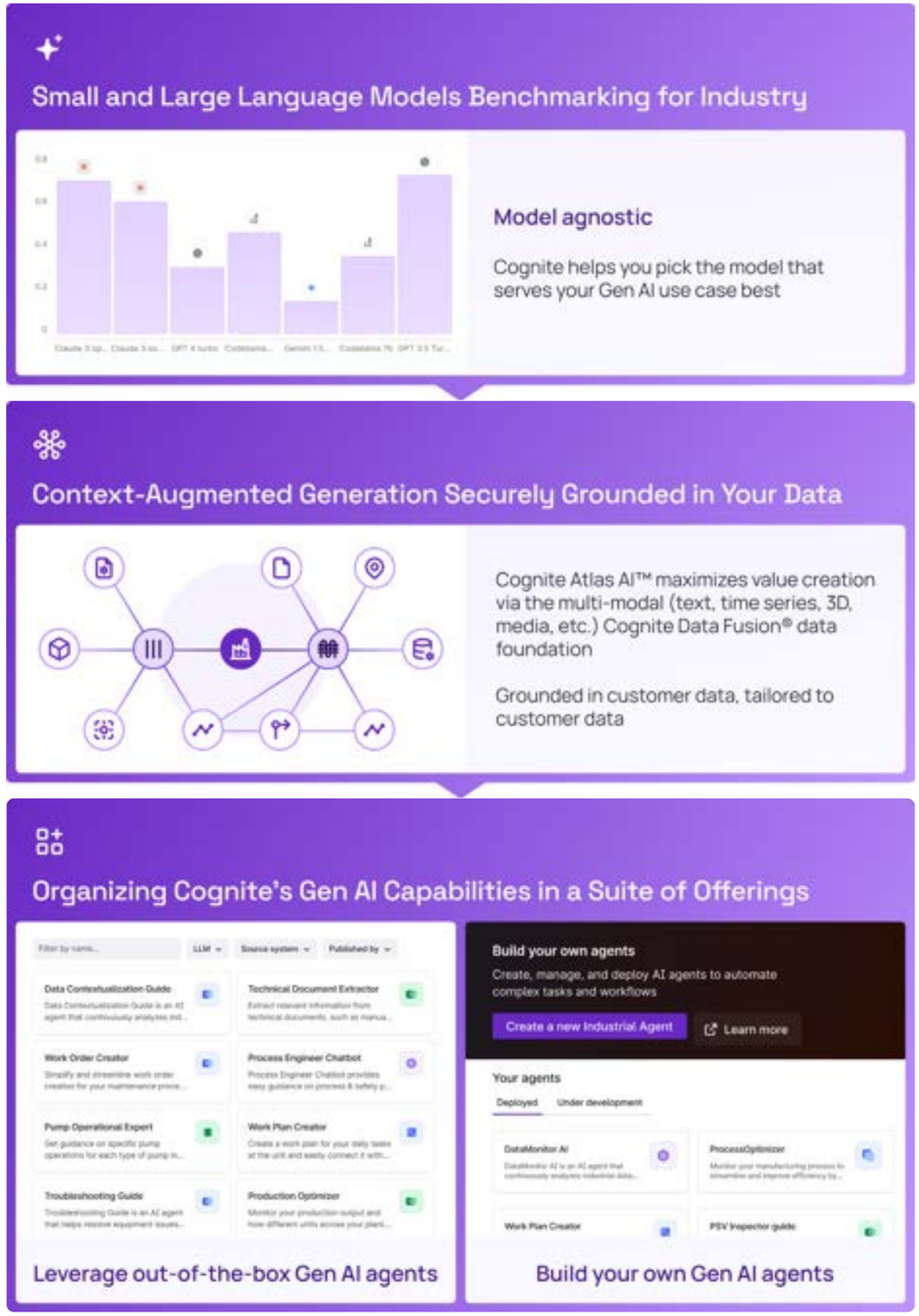
Putting these lessons into action —Iron Man didn't build Jarvis overnight, but we can make several assumptions about what it took to make this high-impact agent. The good news? The journey for industry has many parallels.

1. He started with simple access to complex data. Whether you are trying to improve operational dashboards or introduce industrial agents - both start with an industrial data foundation that uses AI to contextualize information at scale. This is the key requirement for any modern industrial digital transformation program. If you're still struggling in this area, here is a common **starting point →** that has put our customers on fast success tracks.

2. He probably used a knowledge graph to contextualize all his data. In the industrial space, LLMs depend on data in context, returning higher accuracy outputs because agents can be trained on narrow data sets based on their explicit objective. **Learn more about why this matters →**

3. He mastered model and agent orchestration: Industrial transformation has a lot of moving pieces, and proper orchestration can make or break a program.

As you can see, there are many considerations for realizing an agent-based industrial future that is reliable, useful, and valuable. Where should you start? By reading this book, designed to help you start or accelerate your Gen AI journey and get you closer to your own "Jarvis-inspired" operations.



Section 1

Making Generative AI Work for Industry

Chapter 1
Industrial Agents 26

1.1 The Treacherous Path to Trustworthy Gen AI for Industry 28
1.2 The Path Forward: Industrial Agents 32
1.3 The Challenges of Implementing Industrial Agents 36
1.4 The Applications and Benefits of Industrial Agents 38
1.5 Agent Orchestration and Agent Ecosystems 40
1.6 Does RAG Still Matter? 42
1.7 From RAG to CAG 46

Chapter 2
Large, Small, and Custom Language Models 48

2.1 Understanding the Difference 52
2.2 LLMs and Their Application in Operations 54
2.3 So Why Do We Need SLMs? 56
2.4 Custom Language Models 58
2.5 Evaluating Large Language Models Usefulness ≠ Correctness 60
2.6 Choosing the Right Model with autoLLM 68
2.7 Performance Benchmarking 70

Chapter 3
Semantic Knowledge Graphs 72

3.1 Defining Knowledge Graphs 76
3.2 Knowledge Graphs and Data 78
3.3 Knowledge Graphs and AI 82

Chapter 1

Industrial Agents

The Treacherous Path to Trustworthy Gen AI for Industry

You can't avoid the buzz and excitement. Gartner is saying, **"Large Language Models (LLMs) will become the preferred interface to enterprise data,"**[1] and almost every SaaS vendor has recently announced their Gen AI Copilot. Who wouldn't love simple access to complex industrial data and analytics – finally unlocking the data-powered enterprise?

Beyond the hype, however, those working with LLMs for search or analytical query generation are being met with real-world challenges (not scripted, cool demos with extremely limited real-world value):

1.

Generating Working Queries from Natural Language That Produce Correct Results Using LLMs Is Non-Trivial.

ChatGPT made an incredible initial impact, driving significant efficiency gains to developers with programming copilots.[2] Yet, making LLMs serve non-developers – the vast majority of the workforce, that is – is not quite so straightforward. Using LLMs to translate from natural language prompts to API or database queries, or expecting readily usable analytics outputs is challenging for three primary reasons:

- Inconsistency of prompts to completions: No deterministic reproducibility between LLM inputs and outputs.
- Nearly impossible to audit or explain LLM answers: Once trained, LLMs are black boxes.
- Coverage gap on niche domain areas: LLMs are trained on large corpora of internet data, heavily biased towards more generalist topics, not on niche domain areas that typically matter most to enterprise users.

1. **Source:** Gartner. Quick Answer: Safely Using LLMs With an Active Metadata and Data Fabric Layer. 14 August, 2023
2. **Source:** The economic impact of the AI-powered developer lifecycle and lessons from GitHub Copilot. 27 June, 2023
3. **Source:** Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. 31 July, 2023

2.

Existing LLMs Are Computationally Incredibly Expensive and Slow Compared to Database or Knowledge Graph Lookups.

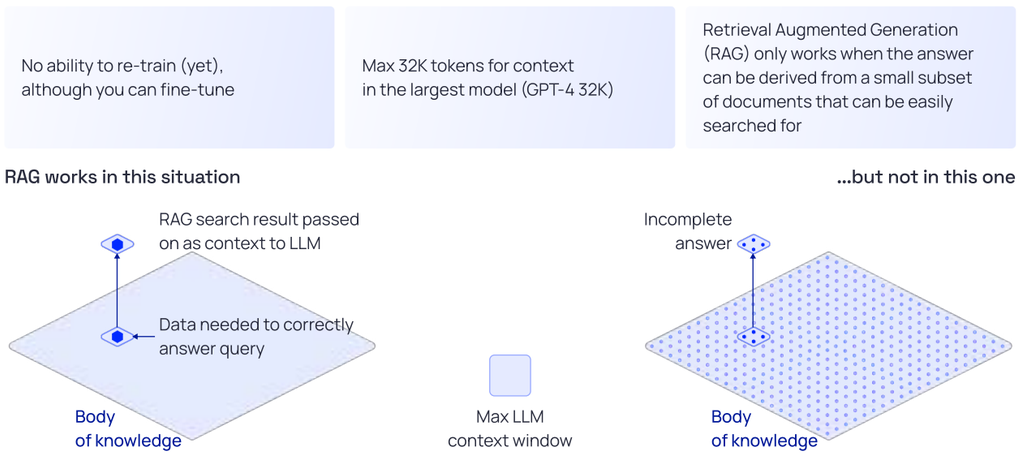
Building production solutions on LLMs will result in a very large cloud bill. For example, GPT-4o, which is a good model with reasonable cost, would cost 0.2 cents for a 100-token prompt and a 100-token answer, which is a very conservative estimate. Cost could easily end up at 1000x (three orders of magnitude) higher with more complex prompts. In addition, LLMs are also very slow compared to low-latency UX expectations in today's real-time software world. Even very large knowledge graph lookups are many orders of magnitude more efficient — and faster. In an era of focus on sustainability, driving up computation with LLMs looks unsustainable. (This is where **small language models** can be. More on that in chapter 2.)

3.

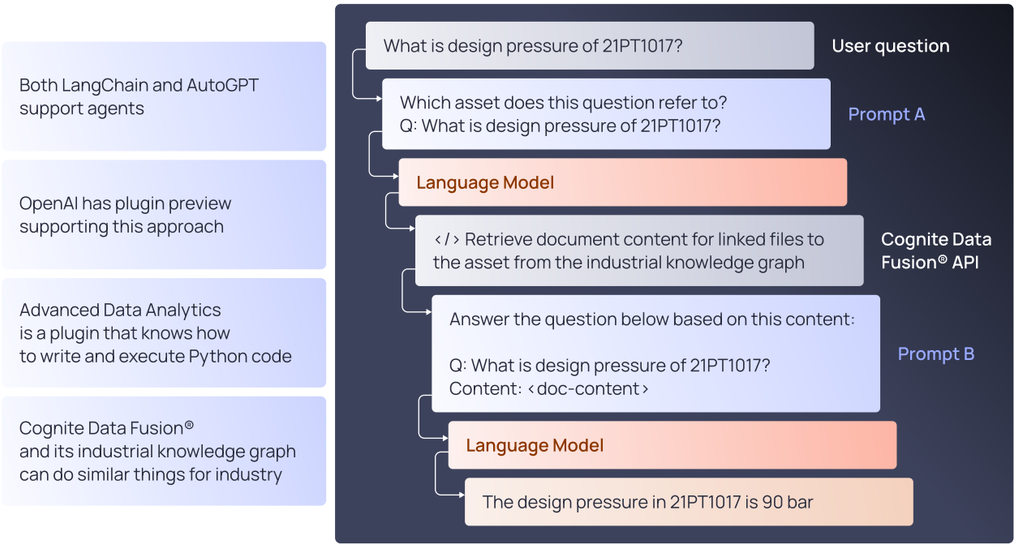
LLMs Require Context, and Using Chaining to Mitigate Context Window Limitations Can Result in Compound Probabilities and, Thus, Less Accuracy.

Without providing LLMs with context, they fail on practically all aside from creative tasks, which their original training corpus (the public internet) supports well. One way to provide this context is using LLM's **context window** for in-context learning. When comparing the context windows of popular LLMs to average enterprise data volumes, it becomes clear very quickly that the math doesn't add up. Moreover, context window limitations will persist for years, if not forever.

Cost and latency aside, growing the context window size does not linearly correlate to outcomes. Studies show[3] that the attention mechanism in LLMs works differently for various parts of a long context window. In short, the content in the middle receives less attention. On the other hand, multiple prompts with shorter context windows allow iteration and optimization of each component individually and even with different LLMs. Moreover, when designed to minimize dependencies among them, it is possible to minimize the effect of compound probabilities and even run them in parallel to reduce latency. Chaining multiple prompts, of course, adds to query volume, hence increasing cost once again.



Agents allow ChatGPT to call out the knowledge graph multiple times during its processing of a query



Lastly, certain types of queries spanning many facts are not feasible with LLMs alone. For example: **"Which assets have shown heat exchanger fouling after 2021?"** In industry domain use cases, the contrast is even more pronounced as typically, LLMs will not have been trained on any proprietary industrial data needed to answer queries, and fitting enough proprietary industrial data into the context window is impossible.

Another impossible task for LLMs is when an answer requires real-time operational data. Ask any LLM **if any condenser units in plant A have a temperature below 5 degrees C right now**, and they cannot answer. The solution architecture to answering such complex queries is to use agents.

Prompt engineering — including as an interactive model — is the Wild West of possibilities (and security risks!). But again, more focused instructions tend to work more robustly in practice than longer prompts. We're again back to chaining (see above).

On security, prompt injections can leak data unless strong data access control is in place. With Cognite, all data retrieval is done using a user's assigned credentials and thus, no user will be able to get access to unauthorized data through prompt injections any more than they would through conventional interfaces. All existing access control mechanisms in Cognite Data Fusion® apply to generative AI use as well.

4. **Source:** Gartner. Quick Answer: Safely Using LLMs With an Active Metadata and Data Fabric Layer. 14 August, 2023

4.

Understanding That LLM Solutions Are Best Assessed on Usefulness Rather Than Mathematical Truism. Gen AI Is Not a Silver Bullet, but a Terrific Pathfinder!

You usually do more than one Google search to find the result you are looking for. The same will apply to our future, wherein **"LLMs will become the preferred interface to enterprise data."**[4]

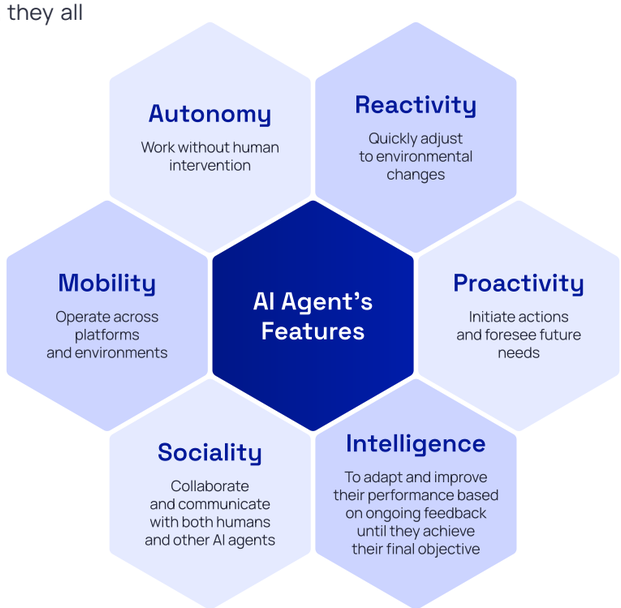
The right design approach is thereby not one that instantly produces the correct result but rather an interactive interface to facilitate the process of finding the right answer, placing the user in control, and using understandable filter inspection (as opposed to only showcasing the generated script to the non-coder user) so that users can review and adjust the suggested filters to find data of interest.

As always, the data itself needs to be provided to the LLM-enhanced interface through a deterministic knowledge graph (more on this in Chapter 3), enabling users to narrow down to relevant parts of the knowledge graph. This interface significantly helps navigate the graph to the right data, even when it might not always directly "zoom" into precisely the node/nodes that initially contain the right answer.

The Path Forward: **Industrial Agents**

Agents are designed to achieve specific goals and can perceive their environment and make decisions autonomously. Agents include chatbots, smart home devices and applications, and the programmatic trading software used in finance.

Agents are classified[5] into different types based on their characteristics but generally, they all exhibit these key attributes:

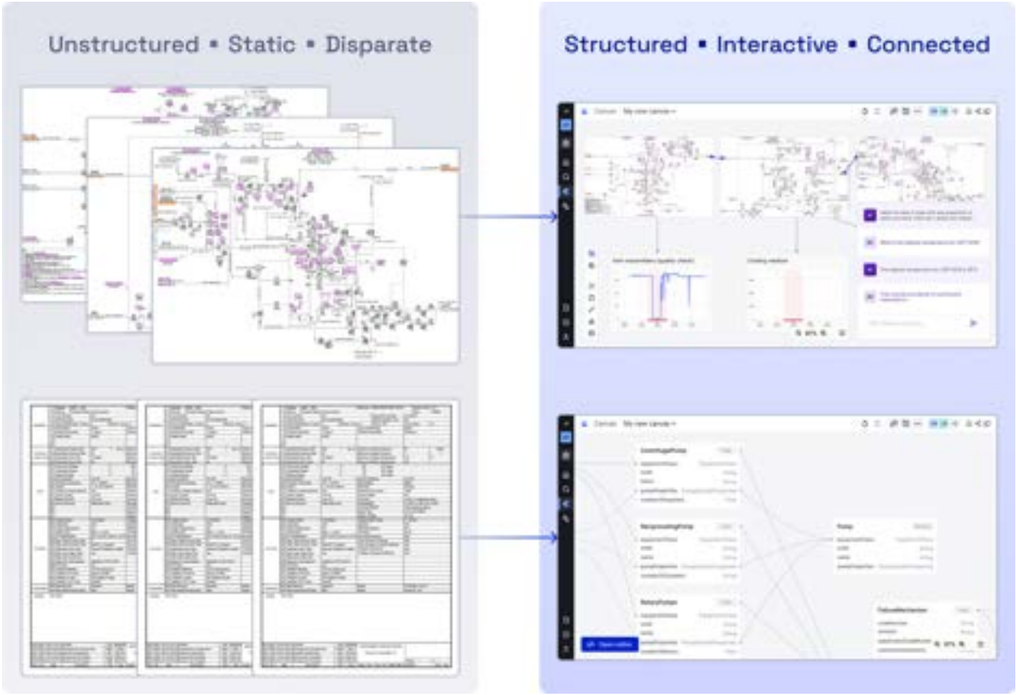


5. **Source:** https://attri.ai/blog/a-complete-guide-top-generative-ai-agents-use-cases-for-manufacturers

Industrial agents are designed to cater to the unique needs of a specific industry. They are specialized versions of general AI assistants focused on solving domain-specific problems with a deep understanding of the industry's context, terminology, and workflows. These agents leverage advanced technologies to provide expert guidance, automate tasks, and offer highly relevant insights into the industry for which they are tailored.

For example, Cognite has deployed a document parsing AI agent to automate equipment registration by reading technical documentation and unstructured input such as data sheets to find relevant information to input into a structured form. Automating data extraction in this way is projected to save more than 10,000 engineering hours.

So how do you go about building an industrial agent like our document parser?



How to Build an Industrial Agent

1 Problem Definition and Requirements Analysis:

A
Identify the specific industrial application or problem the AI agent will address.

B
Define clear objectives, performance criteria, and constraints.

C
Gather domain-specific knowledge and expertise.

2 Data Collection and Preprocessing:

A
Collect relevant data from various sources, such as sensors, historical databases, and real-time streams.

B
Clean and preprocess the data to remove noise, handle missing values, and ensure consistency.

C
Annotate and label data if supervised learning is required.

3 Model Selection and Design:

A
Choose the appropriate AI models and algorithms based on the problem requirements. Common choices include machine learning models, deep learning networks, reinforcement learning, and hybrid approaches.

B
Design the architecture of the AI agent, which may include modules for perception, decision-making, learning, and communication.

4 Training and Optimization:

A
Train the AI models using the prepared data. This involves feeding the data into the model, adjusting parameters, and iteratively improving performance.

B
Use techniques such as cross-validation, hyperparameter tuning, and regularization to optimize the model's accuracy and generalizability.

5 Integration with Industrial Systems:

A
Integrate the trained AI agent into existing industrial systems, such as manufacturing lines, robotic systems, or supply chain management platforms.

B
Ensure seamless communication and interoperability between the AI agent and other components through APIs, middleware, or custom interfaces.

6 Testing and Validation:

A
Conduct rigorous testing of the AI agent in controlled environments to evaluate its performance, robustness, and reliability.

B
Use both simulated scenarios and real-world conditions to validate the agent's effectiveness in achieving the desired outcomes.

7 Deployment and Monitoring:

A
Deploy the AI agent into the live industrial environment.

B
Continuously monitor the agent's performance and behavior to ensure it operates as expected and to identify any potential issues.

8 Maintenance and Continuous Improvement:

A
Regularly update the AI agent with new data and retrain it to adapt to changing conditions and improve performance.

B
Implement feedback mechanisms to learn from real-world experiences and incorporate improvements over time.

The Challenges of **Implementing Industrial Agents**

Developing and implementing industrial agents comes with several challenges that can affect the effectiveness and adoption of the industrial agent.

For example, industrial agents leverage natural language to understand and write code based on published API documentation and examples. This is impossible with data lakes or data warehouses where, without a contextualized industrial knowledge graph, there are no API libraries that can be used as a reliable mechanism to access rich industrial data.

A more exhaustive list of challenges involved in implementing industrial agents includes:

Data Quality and Availability

Challenge: High-quality, relevant, and sufficient data is crucial for training AI models. In many industrial settings, data may be sparse, noisy, or incomplete.

Solution: Adopt an Industrial DataOps approach that includes data validation by design and implement a robust data contextualization engine to bring together and connect industrial data from all sources (like time series, P&ID drawings, equipment logs, maintenance records, 3D models, images, and more).

Integration with Legacy Systems

Challenge: Many industrial environments rely on legacy systems that may not easily interface with modern AI technologies.

Solution: Invest in an Industrial DataOps platform with open APIs to bridge the gap between new AI systems and existing infrastructure. Incremental integration strategies can help in gradual adaptation.

Scalability and Flexibility

Challenge: AI models need to scale efficiently to handle large volumes of data and diverse industrial tasks. They also need to be flexible to adapt to changing requirements.

Solution: Design scalable architectures and use cloud-based solutions for computational flexibility. Modular designs can enhance adaptability.

Real-Time Processing and Decision-Making

Challenge: Many industrial applications require real-time data processing and decision-making, which can be computationally intensive.

Solution: Optimize AI models for speed and efficiency with performance benchmarking and autoLLM.

Interpretability and Trust

Challenge: Industrial stakeholders may be hesitant to adopt AI solutions they do not fully understand or trust.

Solution: Develop interpretable AI models and provide clear explanations for their decisions. Build user-friendly interfaces and visualization tools that will help gain trust.

Security and Privacy

Challenge: Protecting sensitive industrial data from cyber threats and ensuring privacy can be challenging.

Solution: Ensure no proprietary data is shared with third parties, and the built-in mechanisms for logging and access control remain intact. Regularly update and audit security protocols.

Workforce Impact and Change Management

Challenge: Introducing AI can disrupt existing workflows and impact the workforce, leading to resistance.

Solution: Engage with stakeholders and users early, provide training and support, and clearly communicate the benefits and changes and how they can resolve their pain points. Develop strategies for workforce transition and upskilling.

Continuous Improvement and Maintenance

Challenge: AI models require ongoing updates and maintenance to remain effective and relevant.

Solution: Establish processes for continuous monitoring, performance evaluation, and retraining of AI models. Allocate resources for ongoing maintenance and improvement.

Addressing these challenges requires a strategic approach, involving cross-functional collaboration, investment in technology and skills, and a focus on long-term benefits. Fortunately, we have just the thing—Cognite's Customer Success Framework—to help you navigate these complexities successfully (check out the Tools section at the end of this guide for more details).

The Applications and **Benefits** of Industrial Agents

Industrial agents offer significant advantages over the more generic Generative AI approaches most use today, particularly in the context of asset-heavy industries. These advantages stem from their specialized design, precision, and integration capabilities tailored to the specific needs and challenges of industrial environments.

Unlike broader LLMs, which provide general information and answers, industrial agents are equipped with a deep understanding of industry-specific processes and requirements that enables them to offer highly specialized solutions. This specialization ensures that the recommendations and optimizations provided are directly applicable and beneficial to the specific industrial context.

While LLMs can process vast amounts of pre-existing information, industrial AI agents continuously monitor and analyze live data from industrial operations. This capability allows for immediate responses to changing conditions, proactive maintenance, and optimized operational efficiency.

Additionally, industrial AI agents are tailored for predictive analytics specific to industrial applications. Their algorithms are fine-tuned to predict equipment failures, optimize maintenance schedules, and manage risks associated with complex machinery and processes. This precision in predictive capabilities is essential for minimizing downtime and enhancing the reliability and safety of operations, something broader LLMs are not equipped to handle with the same level of accuracy and relevance.

Another key advantage is the integration of industrial AI agents with existing systems and infrastructure. Industrial AI agents are designed to work seamlessly with industry-specific software, hardware, and operational protocols. This integration ensures minimal disruption and allows for smoother implementation and higher compatibility. Broader LLM searches, on the other hand, often require extensive customization to fit into industrial systems, which can be time-consuming and less efficient.

Furthermore, industrial AI agents are developed with a focus on compliance and quality control. They are equipped to monitor production standards, ensure regulatory compliance, and maintain product quality, which are critical aspects of asset-heavy industries. This level of control and assurance is beyond the scope of what broader LLM searches can offer, as LLMs are primarily geared towards providing general information rather than enforcing industry- specific standards.

The applications of targeted, real-time, highly integrated industrial agent solutions that can improve the efficiency, safety, and reliability of energy, process manufacturing, and other industrial companies are seemingly limitless. A few examples include:

- Asset Onboarding Agent
- Catalyst Management Agent
- Checklist Management Agent
- Corrosion Detection Agent
- Data Quality Assessment Agent
- Decentralized Energy Management Agent
- DrillBit Selection Agent
- Dynamic Utilities Management Agent
- Emissions Optimization Agent
- ESG Reporting Agent
- Golden Batch Analyzer Agent
- Isolation Boundary Optimizer Agent
- Kick Prevention Agent
- Leak Detector Agent
- Production Forecasting Agent
- Reservoir Intervention Agent
- Reservoir Optimization Agent
- Root Cause Analysis Agent
- Shift Handover Agent
- Welding Assurance Agent
- Work Order Optimization Agent

Industrial agents represent a significant advancement in industrial AI applications, offering tailored, efficient, and practical solutions that address the unique challenges and needs of various industries. Their ability to enhance productivity, safety, and decision-making makes them invaluable in heavy-asset industries.

Agent Orchestration and Agent Ecosystems

Agent orchestration and agent ecosystems are transformative concepts in industrial environments, offering significant benefits in terms of efficiency, flexibility, and resilience.

However, they also present challenges, particularly around complexity, integration, and security. Successfully implementing these concepts requires careful planning, robust infrastructure, and ongoing management to ensure that agents can work together effectively and adapt to changing industrial needs.

To level-set on some basic definitions, **agent orchestration** refers to the coordination and management of multiple autonomous agents to work together effectively. This involves ensuring that various agents, each with specialized functions, collaborate seamlessly to achieve common goals such as optimizing production processes, maintaining equipment, and managing supply chains.

Taking this one step further, **agent ecosystems** refer to the comprehensive network of interacting agents within a given organization. This includes not only the individual agents but also the infrastructure, protocols, and platforms that enable their interaction. In an industrial setting, an agent ecosystem encompasses all agents involved in production, maintenance, logistics, and more, working within a shared framework.

Agent Orchestration			
Coordination	**Task Allocation**	**Monitoring and Control**	**Integration**
Ensuring that agents communicate and cooperate to avoid conflicts and redundancy.	Distributing tasks among agents based on their capabilities and current workload.	Continuously overseeing agent activities and making real-time adjustments to optimize performance.	Ensuring agents can interact with existing systems and databases to access necessary data and execute actions.

Agent orchestration focuses on the coordination, task allocation, monitoring and control, and integration of individual agents. (Remember Iron Man and his many suits?)

An agent ecosystem is crucial in complex systems, particularly in industrial environments, where multiple autonomous agents must work together to achieve optimal performance.

The agent ecosystem ensures that agents do not work in isolation, but rather in a coordinated manner, leading to increased efficiency. Orchestrating multiple agents brings diverse perspectives and expertise to problem-solving, leading to more informed and effective decisions. The ecosystem leverages the strengths and capabilities of each agent to complement one another. By coordinating the activities of multiple agents, ecosystems can achieve results that are greater than the sum of individual contributions.

An agent ecosystem also allows for redundant systems that can take over tasks in case of failures, enhancing the overall resilience of the operation. AI Agents are dynamic and, thus, can adapt to changing circumstances, which more static Robotic Process Automation and custom-built automation tools are not capable of doing. As such, in case of the failure of one or more agents, ecosystems can reroute tasks and redistribute workloads among other agents, ensuring continuity and minimizing downtime.

As the number of agents increases, ecosystems help manage this complexity, ensuring that new agents are integrated smoothly, the system remains balanced, and all agents are aligned towards common objectives.

In this way, an agent ecosystem can provide a global view of the entire system, allowing for optimization across all processes and agents rather than focusing on isolated parts. This ability enables the integration of different operational aspects (e.g., production, maintenance, logistics) into a cohesive whole, enhancing overall system performance.

In a smart factory, for example, agents might control various aspects such as supply chain logistics, production line operations, and quality assurance. An agent ecosystem ensures these agents work together to optimize production schedules, minimize downtime, and ensure product quality.

Agent orchestration and agent ecosystems are vital for maximizing the potential of autonomous agents in industrial environments. It ensures that multiple agents can work together effectively to achieve superior outcomes compared to isolated agents. This holistic approach is essential for modern industrial systems aiming to leverage AI and automation to their fullest extent.

Agent Ecosystems		
Interoperability	**Scalability**	**Resilience**
Agents can interact and exchange information seamlessly, regardless of their individual roles.	The ecosystem can expand or contract as needed, adding or removing agents without disrupting overall functionality.	The ecosystem is robust, capable of adapting to changes and recovering from failures.

On the other hand, agent ecosystems focus on interoperability, scalability, and resilience of all agents.

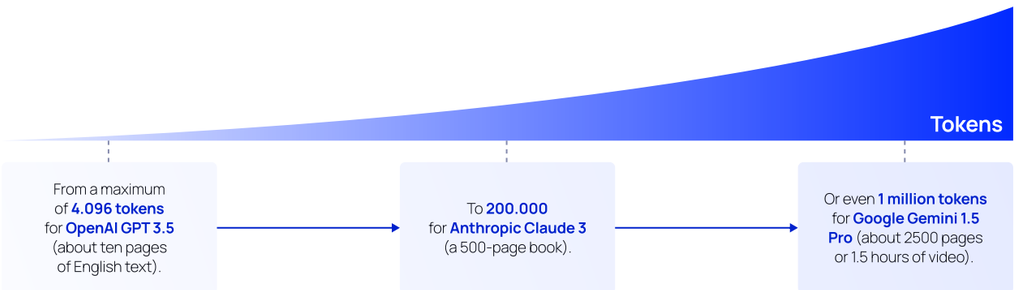
Does RAG **Still Matter?**

Spoiler: It does. While industrial agents give AI applications new capabilities, Retrieval Augmented Generation (RAG) is necessary to effectively solve hallucination and data-freshness problems.

A scalable, trustworthy, and safe generative AI implementation needs an excellent RAG solution to feed relevant, trustworthy, and up-to-date information into large language models.

Some companies are training LLMs from scratch, which requires a large investment. However, there has also been an influx of innovative methods for efficiently fine-tuning models, such as Parameter-Efficient Fine-Tuning (PEFT)[6], QLoRA[7], and prompt tuning[8], which have become popular. These techniques allow companies to "train" models on their data without large investments in data sets or hardware for model training.

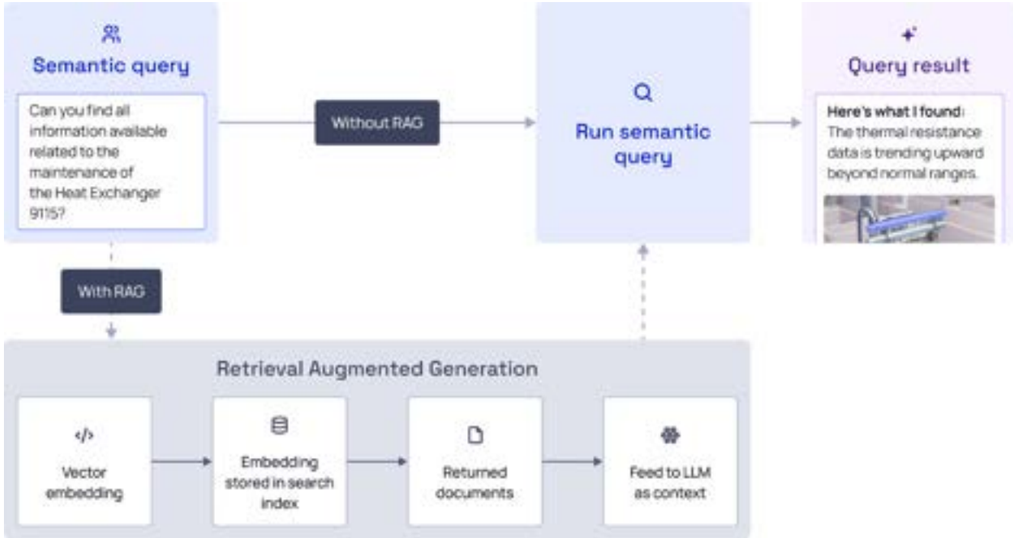
In addition, context window size (the capacity for LLMs to comprehend and retain knowledge) has greatly increased:



6. **Source:** https://www.nature.com/articles/s42256-023-00626-4
7. **Source:** https://arxiv.org/abs/2305.14314
8. **Source:** https://arxiv.org/abs/2104.08691
9. **Source:** https://medium.com/enterprise-rag/why-gemini-1-5-and-other-large-context-models-are-bullish-for-rag-ce3218930bb4

Google has reported that they are experimenting with a context window of 10 million tokens. Given this development, RAG is surely obsolete.

Or is it?



Although the new techniques open up new possibilities, RAG remains a key ingredient in an effective enterprise generative AI framework. There are several reasons why fine-tuning and prompt tuning don't replace the need for a good RAG system:

1. Challenges in Source Attribution

Determining the origin of information from the fine-tuned LLM can be difficult, making it challenging to differentiate between fabrications and factual data.

2. Incorporating New Information or Removing Outdated Information

This necessitates retraining or re-fine-tuning, translating to additional time and financial investment.

3. Issues with Information Access Control

Lacking mechanisms to manage which data is accessible by different users.

4. Ensuring Data Integrity

Compiling large, accurate data sets demands thorough data verification to guarantee the information remains current and correct.

5. Information Overload

Like humans overwhelmed by too much information, a large context window holds vast data, but LLMs might struggle to pinpoint what's relevant, increasing the risk of hallucination and inaccuracies.[9]

LLMs with a very large context window are a valuable addition to the toolbox, but it's not a universal remedy. Although these models open up a variety of solutions, they have drawbacks:

1. Performance

A long context window requires additional processing. This results in a time-to-first token (TTFT is the measurement of how long it takes before the LLM starts "typing" its response) that can be minutes. This is unacceptable for many, even most, use cases.

2. Scalability

With additional processing needs comes reduced scalability. LLM providers operate with quotas. Google allows two requests per minute for Gemini Pro 1.5 (in preview). Anthropic operates with a limit of 10,000,000 tokens per day and 40,000,000 tokens per minute for its top-tier subscription. Although there are always exceptions to such quota restrictions for large enterprises, the numbers are a clear testament to the fact that inference capacity is scarce.

3. Cost

A natural consequence of reduced performance and scalability is increased cost. At the time of writing this, a single request to Anthropic Claude 3 with 200,000 tokens costs about $3.00 USD. This might be an acceptable price for some use cases but, for automated processing pipelines triggering hundreds or thousands of requests on a daily or hourly schedule, the cost will quickly explode.

4. Environmental Impact

Training LLMs is an extremely power-intensive operation. Although energy efficiency is improving due to new hardware and algorithm improvements, interference remains a compute-intensive operation and will remain so for the foreseeable future. A recent report estimates that energy consumption associated with AI is expected to reach 0.5% of global electricity consumption by 2027.[10]

A scalable, trustworthy, and safe generative AI implementation needs an excellent RAG solution to feed relevant, trustworthy, and up-to-date information to the LLMs.

The Cognite contextualization engine and industrial knowledge graph allow accurate information to be retrieved. The contextualization engine ensures data is connected across source systems, and the new AI service for populating structured knowledge graphs from unstructured documents ensures the industrial knowledge graph is as complete, accurate, and up-to-date as possible.

Additional services for semantic search enable filtering information based on meaning rather than free-text search. This technique can even find information across multiple languages. These services make it easy to provide the best possible information to the LLMs, reducing the risk of hallucination while keeping the number of tokens required to process low.

The strategies can be integrated with fine-tuning to achieve an optimal balance of timely, relevant updates and a deep understanding of the data. This hybrid approach allows fine-tuning to adapt the LLM to the general data landscape while RAG ensures the provision of current, relevant information, thereby yielding trustworthy outputs with traceable data sources.

The key is building a representative benchmarking dataset for any generative AI capability to ensure consistent accuracy and performance across various use cases. Such data sets can also be used to compare the results from different LLMs and assess how the different methodologies mentioned above perform. More to come on fine-tuning and benchmarking language models in the coming chapters.

To conclude this section, while advancements in fine-tuning and training LLMs offer impressive capabilities, the integration of RAG remains indispensable for delivering accurate and reliable AI-driven solutions. By combining innovative fine-tuning techniques with robust RAG systems from Cognite, companies can ensure their AI implementations are powerful and practical. Such a balanced approach enhances the efficacy of generative AI and maintains trustworthiness and relevance in industrial applications.

10. **Source:** https://www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption

From RAG **to CAG**

RAG introduced the concept of enhancing generative models with retrieved information from external databases, significantly improving accuracy and relevance. However, RAG is limited by the dependency on the quality and relevance of retrieved documents. Yes, it is still important, but the next evolutionary step is Context Augmented Generation (CAG).

CAG represents a more advanced approach, where the focus shifts from merely retrieving relevant documents to deeply integrating context from multiple sources, including real-time data, sensor inputs, user interactions, and historical data. Also known as GraphRAG, this approach provides a richer and more dynamic context, enabling AI systems to generate more sophisticated and context-aware responses.

CAG is a kind of contextual synthesizer, adapting data and information from various sources within the coherent and comprehensive context of a specific situation. Using the synthesized context, the generative model produces responses that are highly relevant and tailored to the specific needs and conditions of the environment.

Key advantages to CAG include:

- **Richer contextual awareness:** CAG leverages a broader and more dynamic set of context sources, leading to more nuanced and contextually appropriate responses.
- **Real-time adaptation:** By incorporating real-time data and user interactions, CAG can adapt its responses to changing conditions and requirements more effectively.
- **Enhanced decision support:** The comprehensive context provided by CAG improves the decision-making capabilities of AI systems, making them more useful in complex and dynamic environments.

An example:

RAG System

A RAG system in a manufacturing environment might retrieve relevant maintenance manuals or past maintenance records to assist in troubleshooting equipment issues.

CAG System

A CAG system would go further by integrating real-time sensor data from the equipment, historical performance data, and current production schedules to provide a comprehensive diagnostic and maintenance plan.

This does not mean CAG is without its difficulties. Integrating and synthesizing context from multiple diverse sources can be technically challenging. A robust, scalable, and flexible Industrial DataOps platform can play a significant role in overcoming this data integration problem by delivering:

- **A unified data platform** with centralized data management and an industrial knowledge graph to simplify data management and reduce the complexity of handling disparate types of data sources (e.g. sensor data, work orders, engineering drawings and 3D models and more).
- **Data contextualization and metadata enrichment** to link data points across systems and make it easier for CAG systems to interpret and utilize the data.
- **Real-time data processing and streaming data support** to ensure that the context used in generation is current and reflective of the latest conditions and changes in the industrial environment.
- **Scalability** to ensure that, as the number of data sources and the volume of data grow, the platform can continue to integrate and process data efficiently.
- **Interoperability** to ensure that it can work seamlessly with other systems and tools used in industrial environments.

The evolution from RAG to CAG marks a significant advancement in AI capabilities, moving from a model that relies on static document retrieval to one that dynamically integrates a rich and diverse set of contextual information.

This evolution enables AI systems to generate more accurate, relevant, and contextually aware responses, making them more effective in complex and dynamic environments.

Chapter 2

Large, Small, and Custom **Language Models**

Introduction

Large, Small, and Custom Language Models

Language models are one of the main ingredients in developing AI agents, providing them with the capability to understand and generate human-like text. Different models have distinct strengths, such as handling complex language structures, specific domains, or contextual understanding. To design effective AI agents, it is crucial to carefully select and leverage the appropriate language model, ensuring its capabilities align with the intended use and objectives of the agent.

Understanding the Difference

Language models—like GPT3.5, Claude, and Gemini—come in various sizes and configurations, each suited to different types of tasks and requirements. The primary differences between large, small, and custom language models revolve around their architecture, capabilities, resource requirements, and application suitability.

In short, large models are versatile and suitable for a wide range of tasks across various domains, small models are optimal for lightweight, specific applications with limited resources, and custom models excel in specialized applications where domain-specific knowledge is paramount.

You must choose the right model for the right task. To make it easier to understand the difference, we've put together a LM Cheat Sheet.

Large Language Models (LLMs)

Examples:
GPT-3, GPT-4, BERT-Large

Characteristics

Scale

Large models typically contain billions of parameters. They are trained on vast amounts of diverse data, enabling them to understand and generate human-like text across a wide range of topics.

Capabilities

Due to their size, LLMs can perform a variety of tasks such as translation, summarization, question answering, and creative writing with high accuracy.

Performance

They achieve state-of-the-art performance on many natural language processing (NLP) benchmarks and tasks.

Resource Requirements

Training and running large models require significant computational resources, including powerful GPUs or TPUs, large amounts of memory, and substantial storage.

Applications

Best suited for applications requiring high versatility and accuracy across various domains, such as chatbots, virtual assistants, and content generation.

Small Language Models (SLMs)

Examples:
GPT-2 small, DistilBERT, ALBERT

Characteristics

Scale

Small models have fewer parameters, typically ranging from millions to a few hundred million. They are less computationally intensive to train and deploy.

Capabilities

While they may not match the performance of large models in terms of versatility and accuracy, they can still perform many NLP tasks effectively.

Performance

Adequate for many applications, especially those with specific, narrow tasks or where computational resources are limited.

Resource Requirements

More feasible for deployment on devices with limited computational power, such as mobile phones or edge devices.

Applications

Suitable for applications where quick, lightweight inference is needed, such as on-device text prediction, simple chatbots, and certain real-time applications.

Custom Language Models

Examples: **Customized versions of most open source models**

Characteristics

Tailored Training

Custom models are trained or fine-tuned on specific datasets relevant to particular tasks or industries. This training can be based on either large or small models as a starting point.

Specialization

They are optimized to perform exceptionally well on domain-specific tasks, such as medical text analysis, legal document processing, or customer service interactions.

Performance

Custom models can outperform generic large or small models in their specialized domains because they leverage domain-specific knowledge and nuances.

Resource Requirements

Vary based on the base model and the size of the dataset used for fine-tuning. Typically, they require less training time and computational resources than building a model from scratch, but require more than using pre-trained models directly.

Applications

Ideal for industry-specific applications where accuracy and relevance to the domain are crucial, such as finance, healthcare, legal, and specialized customer support systems.

LLMs and Their **Application in Operations**

The practical applications of LLMs in business operations are vast. LLMs can process vast amounts of text documents, extract relevant information, and summarize key findings. Increasingly, they are used to interpret engineering diagrams, designs, P&IDs, imagery, video recordings, voice commands, operational audio, vibration data, and other multi-modalities, all of which helps to extract insights from large volumes of unstructured data. For example, an LLM-based system can analyze maintenance reports, sensor logs, and operator notes to help operators efficiently navigate and discover relevant data, leading to better decision-making and improved operational efficiency.

LLMs can also play a vital role in industrial data analysis by assisting in critical activities such as anomaly detection and quality control. By ingesting historical data, sensor readings, and operational parameters, LLMs can learn to identify early signs of equipment failure, detect deviations from normal operating conditions, or pinpoint potential quality issues, supporting proactive maintenance strategies.

LLMs are a powerful tool for industry, improving operations in various ways that minimize downtime, reduce costs, and achieve higher overall efficiencies. With their ease of use, adaptability, and practical applications, LLMs offer a user-friendly solution that can streamline operations, automate tasks, gain valuable insights, and drive innovation in their respective industries.

So **Why Do We Need SLMs?**

SLMs, as the name implies, are a subset of machine learning models intentionally kept small in terms of their parameters and computational requirements. SLMs require less data and computational power to train and can perform inference even faster than LLMs, which is critical for real-time applications. Additionally, SLMs are often fine-tuned for specific tasks or domains, which enhances their performance in those areas despite their smaller size.

In general, SLMs are easier to maintain, update, and deploy across various systems and, due to their lower computational requirements, consume less energy, leading to lower long-term operational costs.

More specifically, SLMs provide numerous benefits to industrial organizations, including:

- **Enhanced natural language interfaces:** SLMs enhance voice-activated controls and natural language queries for industrial control systems, allowing operators to interact with machines and data systems more intuitively and efficiently.
- **Knowledge extraction:** SLMs can extract valuable insights from technical manuals, regulatory documents, and historical records, providing decision-makers with critical information that supports informed decision-making.

- **Real-time monitoring and diagnostics:** SLMs can analyze operational logs and maintenance records in real-time to detect anomalies and predict equipment failures, helping minimize downtime and optimize maintenance schedules.
- **Predictive analytics:** SLMs can process textual data, such as equipment logs and technician notes, to forecast potential issues before they escalate.
- **Edge computing:** SLMs can run on edge devices, enabling local data processing and decision making without relying on constant internet access, which is critical in heavy-asset industries in which connectivity to central servers might be limited.

SLMs are crucial in heavy-asset industries due to their efficiency, cost-effectiveness, and ability to operate in resource-constrained environments. They enable real-time monitoring, enhance decision-making, optimize resource use, and ensure safety and compliance, making them indispensable tools for modern industrial applications.

Custom Language **Models** in Industry

Custom language models are tailored AI models designed to perform specific tasks or cater to particular domains by being trained or fine-tuned on relevant—often proprietary—datasets. Custom models also incorporate nuances and specific terminologies of the targeted domain, enhancing their understanding and performance in that area.

Key benefits of custom models include:

- **Domain-specific accuracy:** Custom models provide higher accuracy and relevance in industrial applications by understanding and processing domain-specific language and context. For example, in the energy sector, a custom model can interpret and generate reports, safety guidelines, and operational procedures with higher precision.
- **Efficiency and productivity:** By automating complex tasks that require specialized knowledge, custom language models can significantly improve efficiency and productivity. In manufacturing, a custom model can optimize maintenance schedules, predict equipment failures, and streamline production processes.
- **Enhanced decision-making:** Custom models aid in better decision-making by providing insights and recommendations based on domain-specific data. For instance, in chemical processing, a custom model can analyze reaction outcomes and suggest optimal conditions, thereby improving yield and reducing waste.
- **Regulatory compliance:** In industries with strict regulatory requirements, such as pharmaceuticals or finance, custom models ensure that communications and documentations comply with industry standards and regulations, thereby minimizing legal risks.
- **Cost reduction:** By automating routine and complex tasks, custom language models reduce operational costs. For example, in supply chain management, a custom model can optimize logistics, reduce inventory costs, and improve demand forecasting.

The best language model for industrial agents depends on various factors, including the specific requirements, constraints, and goals of the industrial application. Each type of language model—large, small, or custom—offers distinct advantages and is suitable for different scenarios within industrial environments.

Evaluating Large Language Models:

Usefulness ≠ Correctness

Evaluating LLMs involves assessing their performance on specific tasks to understand how well they meet the intended use case requirements. As these models become increasingly integrated into various applications, ensuring their reliability, trustworthiness, and effectiveness is paramount. Without evaluations, you simply cannot know whether your LLM-based solution—whether it is prompt engineering, Retrieval Augmented Generation (RAG), or fine-tuning—is actually working, and neither can you improve it.

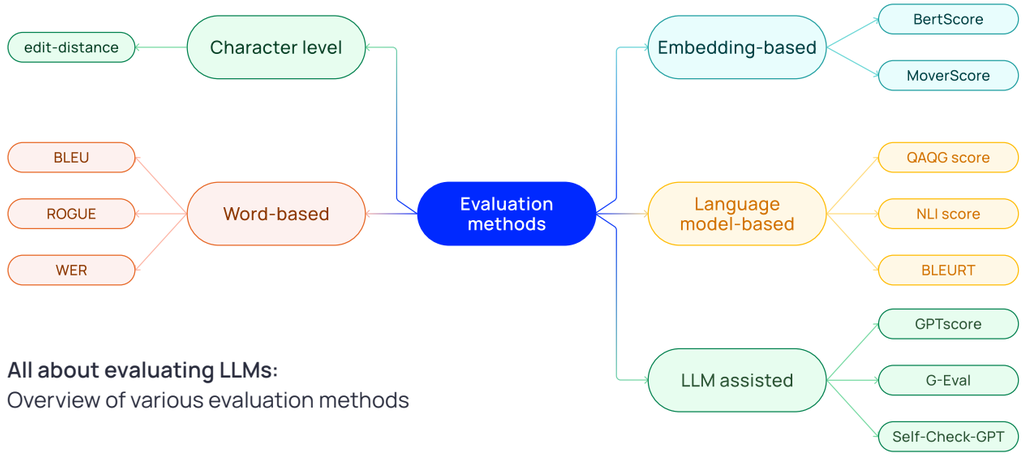
Evaluation Strategies

Evaluations are a set of measurements used to check how well a model performs a task. An evaluation consists of two main components: benchmark data and metrics.

While many benchmark data sets are available for LLMs, specialized tasks often require tailored data sets. For instance, if you want to use LLMs to generate a request body for your API service, you will need a dataset that includes examples of request bodies commonly used in your application domain.

Evaluation metrics are used to quantify the model's performance on the benchmark data set. These metrics can broadly be classified into two main groups: traditional and nontraditional.

- Traditional metrics focus on the order of words and phrases, given a reference text (ground truth) for comparison. Examples include exact string matching, string edit distance, BLEU, and ROUGE.
- Nontraditional metrics leverage language models' ability to evaluate generated text. Examples include embedding-based methods such as BERTScore, and LLM-assisted methods such as G-Eval, in which a powerful LLM is instructed to evaluate the generated text.



All about evaluating LLMs:
Overview of various evaluation methods

Methods where the performance of the system is estimated using pre-defined data sets, like the ones described above, are called **offline evaluation**. Another example of this is "sample testing", whereby answers are randomly selected and checked against a separate (more expensive) LLM or by a human reviewer (much more expensive). This is similar to how quality control is done in many manufacturing spaces. While these offline evaluation methods are essential for ensuring that an LLM-based product feature has acceptable performance before deploying to users, they have their limitations and are usually not enough.

Creating high-quality benchmark data sets takes time, and your data set can get outdated after releasing a feature and no longer represent the type of tasks users ask about. In addition, offline evaluation may not fully capture the complexity of real-world user interactions.

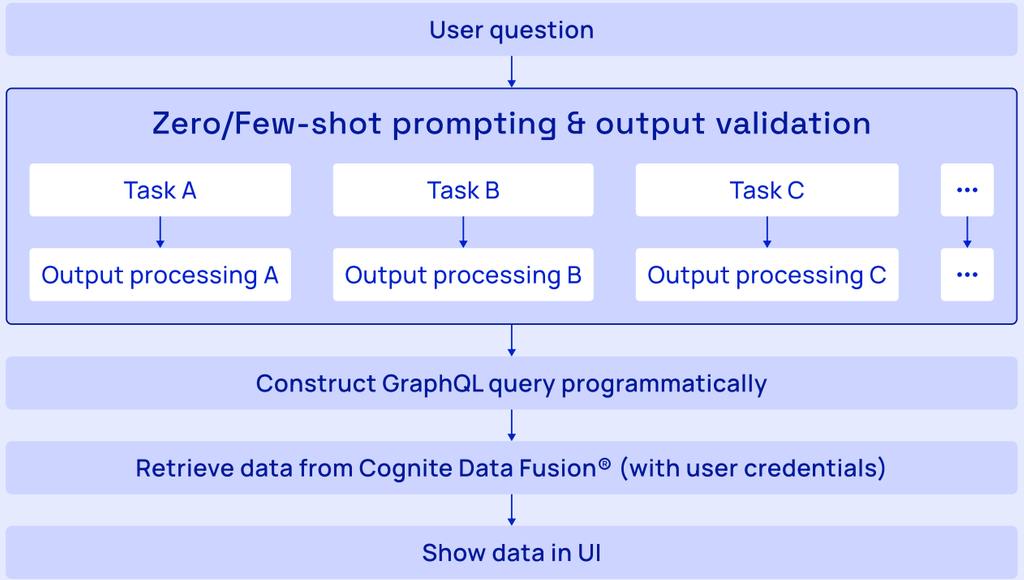
This is why offline evaluation must be complemented with **online evaluations**: the continuous evaluation of LLM-based features as they operate in a production environment—in **real** time, with **real** users and **real** data. After all, what matters most is whether the solution is actually useful for users in a production setting.

Online evaluation includes user feedback, which can either be explicit, such as having users provide ratings like thumbs up or down, or implicit, such as monitoring user engagement metrics, click-through rates, or other user behavior patterns. Note that both forms of feedback —explicit and implicit— are important. Explicit feedback is generally more accurate and less noisy than implicit feedback but tends to be much less abundant.

To summarize, offline evaluations help you decide whether your LLM-based product feature has the minimum acceptable performance before deploying to users, while online evaluations are needed to ensure that your product continues to perform well in real-time user interactions, allowing you to monitor and improve its functionality over time based on live user feedback and behavior.

So, what does this look like in practice?

Example use case:
Natural language query to GraphQL



Natural Language question to GraphQL

One of the AI features included in Cognite Data Fusion® is the ability to search for data using natural language. Described at a high level, the user input or question is converted into a (syntactically correct) GraphQL query, which is then executed (using the user's credentials) to retrieve data from Cognite Data Fusion®.

The conversion from natural language to GraphQL is done through a set of prompts—instructions for large language models to generate a response—where each analyzes different aspects of the question and returns specific components of the GraphQL query. For instance, one prompt is designed to propose a suitable query operation (e.g., get, list, or aggregate), another is designed to generate a suitable filter, and so on.

Each prompt also has a corresponding post-processing step to ensure the validity of the generated output.For instance, a simple post-processing example is to check that the suggested query operation is a valid GraphQL query method. The outputs from all prompts are combined and used to construct a valid GraphQL query programmatically.

Benchmark Data Set

To evaluate the previously described feature, multiple specialized data sets were curated with industry-relevant question-answers pairs. Our benchmark data set includes around ten different Cognite Data Fusion® data models from energy and manufacturing sectors. Each model comprises tens to hundreds of real-life question-answer pairs, allowing evaluation across diverse scenarios. Below is an example test case from an Asset Performance Management (APM) data model:

- **question:** different formulations of a relevant question, that all can be addressed using the same GraphQL query.
- **relevantTypes**: List of GraphQL types that are relevant to the question.
- **queryType:** The relevant GraphQL query and which type it should be applied on.
- **queryType:** Relevant filter.
- **properties:** Properties that are most relevant and should be returned given the context of the question.
- **summary:** A short description of the suggested GraphQL query.

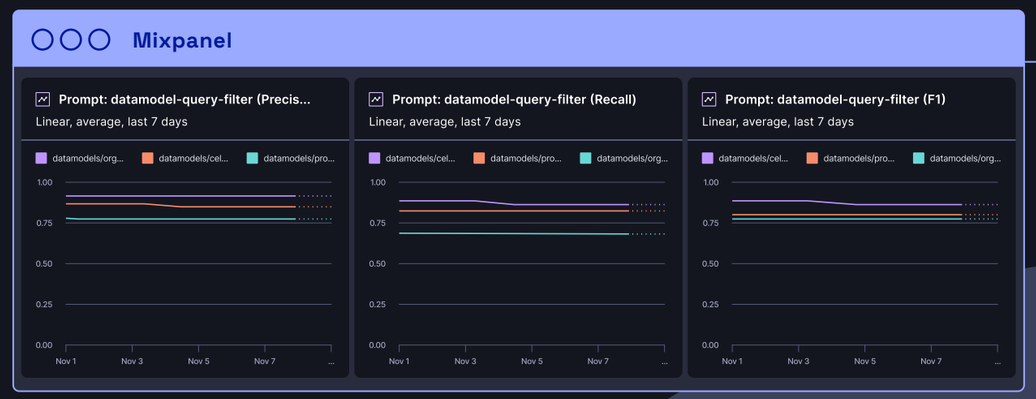
```
{
  "question": [
    "What are the details of all assets under area id 1004 "
    "Information on assets in area id 1004 and related work orders?",
    "List assets and work orders under area 1004",
    "Details of assets and respective work orders for area id 1004?"
  ],
  "relevantTypes": ["Asset", "WorkOrder"],
  "queryType": {
    "type": "Asset",
    "queryType": "list"
  },
  "filter": [
    {
      "and": [
        {
          "areaId": {
            "eq": 1004
          }
        }
      ]
    }
  ],
  "properties": {
    "Asset": ["tag", "description", "areaId", "workOrders"],
    "WorkOrder": ["title", "description", "workOrderNumber"]
  },
  "summary": "Listing assets and work orders in area 1004",
},
```

Evaluation Metrics

The prompts are evaluated using a mix of traditional and nontraditional methods. Simple string matching works for most fields but, for certain ones, like relevant properties, we calculate standard metrics (recall, precision, and F1 scores) by comparing suggested properties to the ones in the benchmark dataset. The summary field is evaluated with an LLM-assisted approach, where a powerful language model (GPT-4) grades the suggested summary, given the ground truth summary.

To monitor the model's performance, especially for changes in the underlying base model, we calculate evaluation metrics daily across all available datasets. These metrics are tracked in **Mixpanel**.

Furthermore, we created a suite of developer tools to support a rapid feedback cycle while developing prompts and refining post-processing techniques. This suite comprises a **Command-Line Interface (CLI)** for assessing one or more prompts against a dataset or various datasets. It also includes a **Continuous Integration (CI)** system that intelligently identifies the necessary evaluations to perform in response to modifications in a **pull request (PR)**. It then compiles a report directly on the PR page, offering complete insight into the impact of the changes on the evaluation metrics.



Monitoring of offline evaluation metrics in **Mixpanel**.

Report of evaluation metrics in a **Pull Request**.

CLI tool for evaluating prompts on evaluation datasets in Mixpanel.

Command-Line Interface

prompt	datasetName	question	success	truePositive	falsePositive	falseNegative	Diff
datamodel-query-filter	datamodels/apm	Equipment in functional location CG-ACE-ATQB	YES	1	0	0	None
		Equipment located at functional location CG-ACE-ATQB	YES	1	0	0	None
		Equipment at the functional location CG-ACE-ATQB	YES	1	0	0	None
		Equipment positioned in functional location CG-ACE-ATQB	YES	1	0	0	None
		Equipment changed in 2014	NO	0	2	2	{ "and": [{ "changedDate": { - "gte": "2014-01-01T00:00:00.000Z + "gte": "2013-12-31T23:00:00.000Z } }, { "changedDate": { - "lt": "2015-01-01T00:00:00.000Z + "lt": "2014-12-31T23:00:00.000Z }

Pull Request

jenkins-ci-cognitedata bot commented last week

datamodel-query-type

dataset	accuracy
datamodels/acc	1
datamodels/apm	0.795918
datamodels/apmsimple-lervik	0.917647
datamodels/c	0.962264
datamodels/fish_farming	0.769231
datamodels/manufacturing	0.769231
datamodels/movies	1
datamodels/organization	1
datamodels/process_manufacturing	1

Online Evaluation

Evaluating a proposed GraphQL query for syntactical accuracy is relatively straightforward. However, determining its semantic correctness and usefulness is **significantly** much harder. In this context, a semantically correct query retrieves relevant data based on user input or inquiry. To accurately gauge this or, more precisely, to obtain an indication of usefulness, online evaluation strategies are necessary. A few example metrics that are collected for this particular use case include:

- **Thumbs-up and thumbs-down ratings:** Users can provide feedback through simple thumbs-up or thumbs-down ratings, indicating their satisfaction or dissatisfaction with the retrieved data.
- **User modifications of suggested filters in the UI:** Users might adjust the filters suggested by the system, tailoring the query to their specific requirements. These modifications may reflect that the suggested filter did not meet their expectations.
- **User modification of the list of properties to display:** Users might adjust the list of properties displayed in the result overview. These modifications may reflect that the properties shown initially did not meet their expectations.

In addition, several other performance and utilization metrics related to latency, error responses, and wasted utilization of the LLM—due to service errors or any other unactionable response—are also collected.

The online metrics serve as a foundation for evaluating the effect of modifications to the feature through A/B testing. In such tests, various iterations of the prompt sequences are distributed across distinct user groups to determine the most effective version. Variations may include using different LLMs, alternative prompts, or varied pre- and post-processing approaches. Moreover, these online metrics are instrumental in establishing Service Level Objectives (SLOs) for response times and end-user error rates, as well as user satisfaction measured via the satisfaction metrics mentioned earlier in this section.

Usefulness ≠ Correctness

It's important to emphasize that usefulness differs significantly from correctness. While correctness can be measured explicitly and objectively, usefulness is inherently subjective. Hence, the aforementioned metrics serve as indicators of whether users find the tool valuable or not. These metrics, derived from user interactions and feedback, offer valuable insights into the practical utility of the GraphQL query and help refine the system for optimal user satisfaction.

Choosing the Right Model **Using AutoLLM**

As we mentioned before, the best language model for industrial agents depends on various factors, including the specific requirements, constraints, and goals of the industrial application. Each type of language model—large, small, or custom—offers distinct advantages and is suitable for different scenarios within industrial environments.

So how do you choose the right model for your needs if you are not an AI engineer?

AutoLLM is an automated process designed to help users select the most suitable language model for their specific needs. This process leverages various criteria, including performance, cost, and specific use case requirements, to recommend the best model.

AutoLLM refers to automated systems and frameworks designed to select, configure, and optimize language models for specific tasks or applications. It leverages various criteria, including performance, cost, and specific use case requirements to simplify the complex process of choosing the most suitable model, tailoring it to the user's needs, and ensuring optimal performance.

With autoLLM, a user provides details about their specific use case, including the type of task (e.g., text generation, summarization, classification), performance expectations, resource constraints and budget, and the system uses algorithms to match user requirements with the most suitable models from the database.

For example, models that analyze images and sensor data from production lines are better suited for quality control and defect detection, while models that analyze geological and production data are best for optimizing reservoir management strategies.

AutoLLM also automates the configuration and tuning of the recommended model as needed to optimize model performance for the specific task. Once the optimal model is selected and configured, AutoLLM can automate the deployment process, integrating the model into the user's application or workflow. Post-deployment, the system monitors model performance in real-time, providing feedback and making adjustments as necessary to maintain optimal performance.

AutoLLM simplifies the process of choosing and deploying the right LLM, making advanced AI accessible to users without deep machine-learning expertise. Not only does the system save time and resources, but it also continuously refines and adjusts models to ensure sustained performance improvements over time, ensuring better performance and more accurate results.

Additionally, autoLLM helps reduce unnecessary resource consumption and associated costs by balancing performance with computational efficiency and tailoring solutions to fit within specified budget constraints, optimizing for cost-effectiveness.

By automating the complex tasks of model selection, configuration, and optimization, AutoLLM empowers users to leverage advanced AI capabilities without requiring deep technical expertise. Thus, autoLLM makes industrial agents more accessible, efficient, and effective. And it is these agents that will enhance predictive maintenance, knowledge management, quality control, supply chain optimization, process automation and more, leading to improved efficiency, cost savings, and better decision-making.

Performance Benchmarking

Performance benchmarking for language models is the process of evaluating and comparing their efficiency, accuracy, and overall effectiveness in performing specific tasks. This process involves systematically measuring various performance metrics to ensure that the models meet the desired standards and are suitable for their intended applications.

Performance benchmarking involves assessing language models against predefined metrics and standards. This evaluation helps in understanding the strengths and weaknesses of the models, guiding improvements, and ensuring that they are fit for purpose.

- **Task-specific metrics:** Metrics such as accuracy, precision, recall, F1-score, and BLEU score, which are relevant to specific tasks like classification, translation, summarization, etc.
- **Resource metrics:** Measures of computational efficiency, such as inference time, memory usage, and processing speed.
- **Scalability metrics:** Metrics to assess how well the model handles increasing data volumes and user demands.
- **Benchmark suites:** Standardized sets of tasks and datasets are used to test models uniformly.

Benchmarking provides a quantitative basis for comparing models, ensuring that decisions are based on objective data rather than subjective judgments. It helps in identifying the most efficient models that offer the best performance relative to resource consumption.

The benchmarking process involves comparing the performance of multiple language models across a variety of standardized tasks and datasets, aiming to provide insights into how different models perform relative to each other, highlighting their strengths, weaknesses, and overall suitability for various natural language processing (NLP) tasks. We use benchmarking to determine the suitability of the model for a specific task. For example, evaluating a model's accuracy and inference time in generating text summaries and providing insights into its task-specific strengths and weaknesses.

Benchmarking models against each other provides an objective basis for comparing the strengths and weaknesses of different models, offers insights into which models are more suitable for specific tasks based on empirical data, and helps stakeholders make informed decisions about model selection and deployment strategies.

Chapter 3

Semantic Knowledge Graphs

Introduction

Semantic Knowledge Graph

Knowledge graphs are crucial to AI agents as they provide structured, interconnected information that enhances the agents' ability to understand and process complex relationships between data.

By leveraging knowledge graphs, AI agents can deliver more accurate, context-aware insights and recommendations, improving their effectiveness in tasks such as data analysis, decision-making, and problem-solving. Integrating knowledge graphs ensures AI agents can access and utilize rich, domain-specific knowledge, elevating their overall performance and utility.

Defining the **Semantic Knowledge Graph**

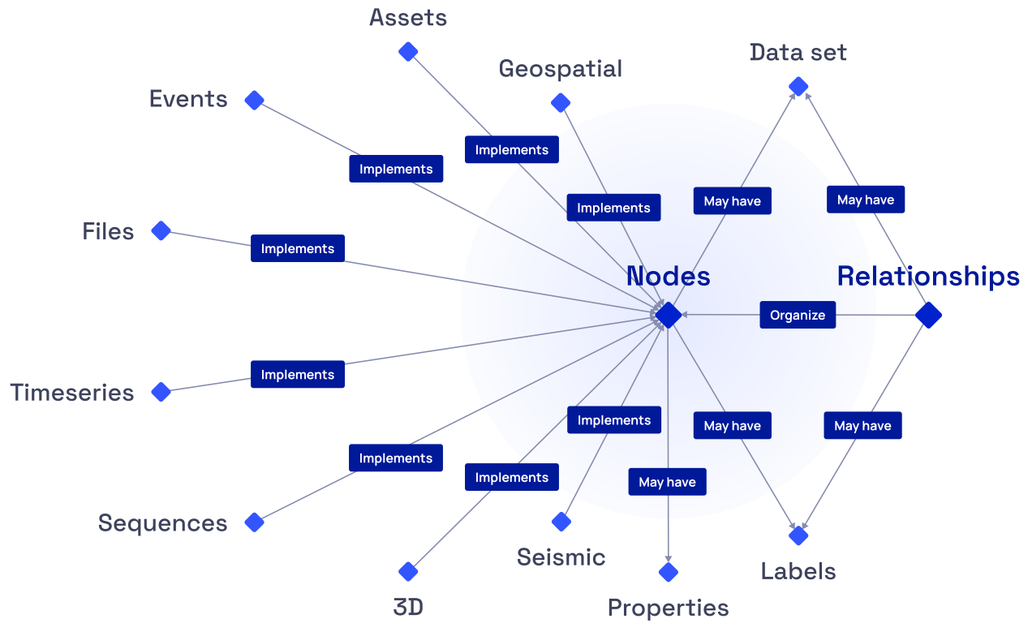
Knowledge graphs map data relationships, capture interconnections, and trace data lifecycles. They are constructed by combining data sets from diverse sources, each varying in structure. The harmony between schemas, identities, and context contributes to the coherence of this comprehensive data repository.

- Schemas establish the fundamental framework upon which the knowledge graph is built.
- Identities efficiently categorize the underlying nodes.
- Context determines the specific setting in which each piece of knowledge thrives within the graph.

Knowledge graphs use machine learning to construct a holistic representation of nodes, edges, and labels through a process known as **semantic enrichment**. Knowledge graphs can discern individual objects and comprehend their relationships by applying this process during data ingestion. This accumulated knowledge is then compared and fused with other data sets that share relevance and similarity.

You have likely heard us reference the industrial knowledge graph before. This open, flexible, labeled property graph represents your operations by focusing on the intricacies and specifics of industrial processes, machinery, operations, and related data. It liberates the data that has been locked in different systems and applications (high-frequency time-series sensor data, knowledge hidden in documents, visual data streams, and even 3D and engineering data) and makes it meaningful and manageable.

A semantic knowledge graph captures the relationships between entities in a way that humans and machines understand. It enables question-answering and search systems to retrieve comprehensive responses to specific queries. Semantic knowledge graphs are time-saving tools that streamline manual data collection and integration efforts to bolster decision-making processes.



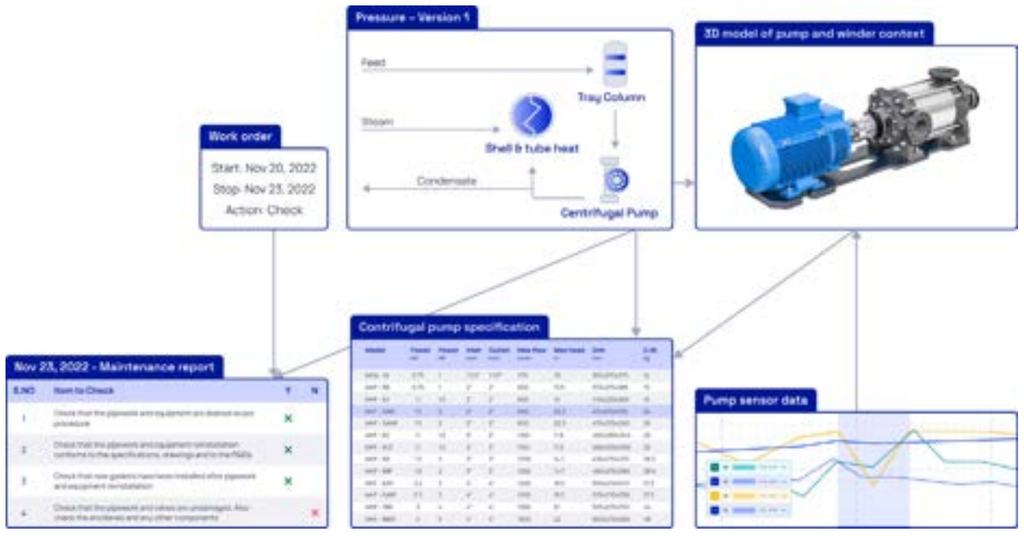
The powerful combination of industrial and semantic knowledge graphs delivers a structured representation of information that allows industrial organizations to understand the relationships and connections within complex datasets better. These data relationships are made possible through contextualization pipelines that help create and maintain a dynamic knowledge graph; thus, addressing three key challenges:

- **Overcoming data silos:** In industrial settings, data often resides in numerous silos, leading to duplication and ambiguity in meaning. Knowledge graphs play a pivotal role in breaking down these silos, providing a unified view of data, and improving the understanding of usage and consumption patterns.
- **Unleashing unstructured data:** By employing standardized metadata, the knowledge graphs allow for the categorization and management of information, enhancing the utilization of unstructured data present in documents, images, and videos (another common data silo) and turning that data into actionable insights.
- **Enhancing business insights:** The explicit contextualized knowledge, rules, and semantics embedded in knowledge graphs empower AI applications to provide high-quality, trusted insights that are absolutely necessary for working in the industrial domain. They also allow subject matter experts to make high-quality decisions, enhancing business processes, workflows, and operations.

However, **a knowledge graph is only as valuable as the data it can access.**

Knowledge Graphs and **Data**

A knowledge graph enables an industrial organization to extract value from unstructured and siloed data sources. Establishing a dynamic and interoperable industrial knowledge graph with access to high-quality contextualized data must be the first step for any organization that wants to implement generative AI initiatives that improve operations and accelerate time to value.

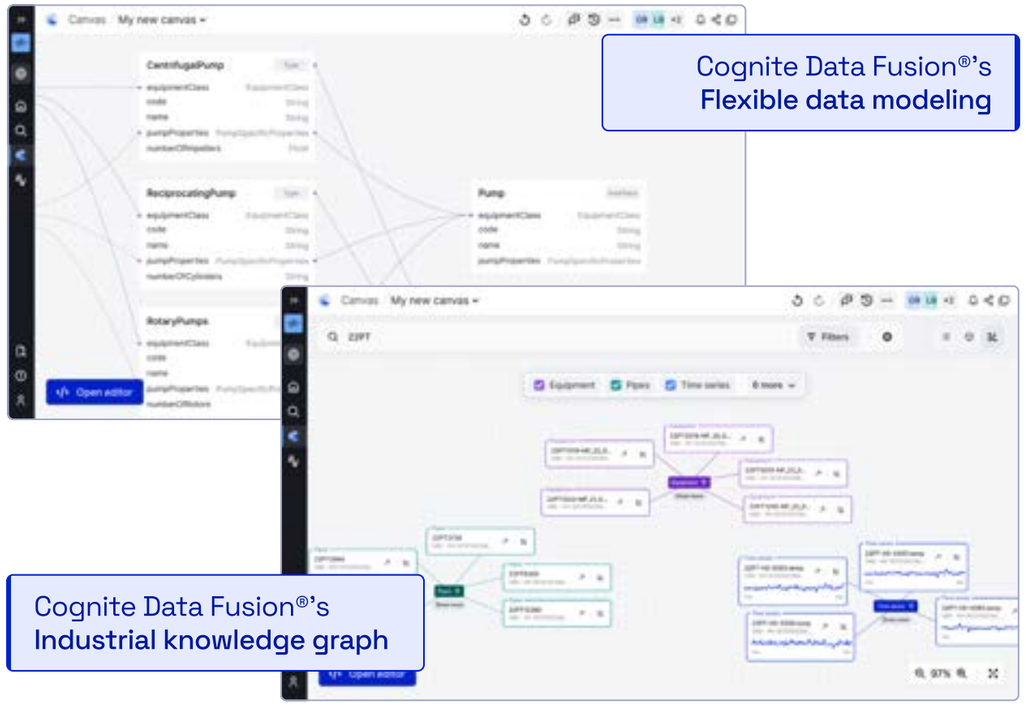


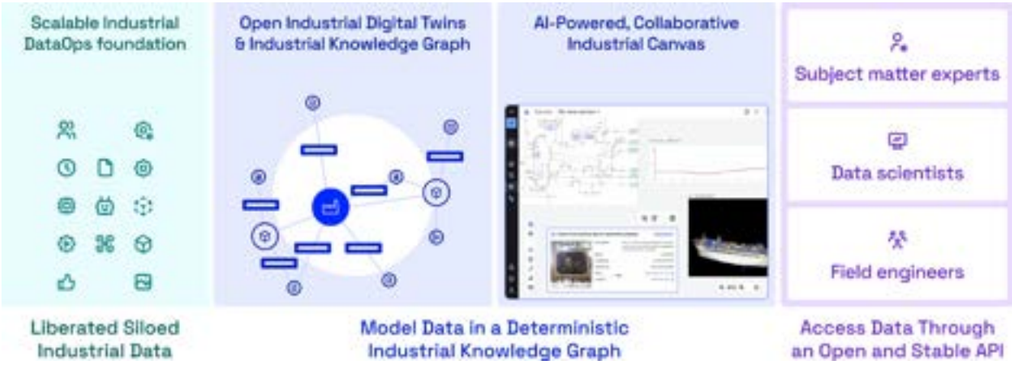
In this example, the left diagram above illustrates a simplified version of an industrial knowledge graph of a centrifugal pump. Depending on the persona, users may explore a problem with the pump from multiple entry points. Maintenance may start with the latest maintenance report, an operator may use the time series, and a remote SME may begin with the engineering diagram (e.g., P&ID). The maintenance report, the work order, the time series values, and the engineering diagrams are each in separate systems. Having all this data connected in the industrial knowledge graph creates a seamless experience, regardless of the starting point.

This simple example illustrates the importance of data contextualization across different systems. Cognite's AI-powered data contextualization capabilities power the industrial knowledge graph (as seen on the right side) to provide access to the maintenance report, work order, time series, and more in a single location.

With the industrial knowledge graph as the foundation, data is understood and structured to meet the specific needs of users or use cases, making it easier for all stakeholders to find the necessary information, and view and understand relationships between data objects.

Simple aggregation of digitized industrial data is a significant step forward from the silos and inaccessibility that often plague large enterprises. However, to provide simple access to complex data, the variety of industrial data types must be accounted for, and the semantic relationships that drive scalable utilization of this data must be incorporated to support interactive user experiences. Codifying this context as an industrial knowledge graph is vital to enabling consistent, deterministic navigation of these meaningful relationships.

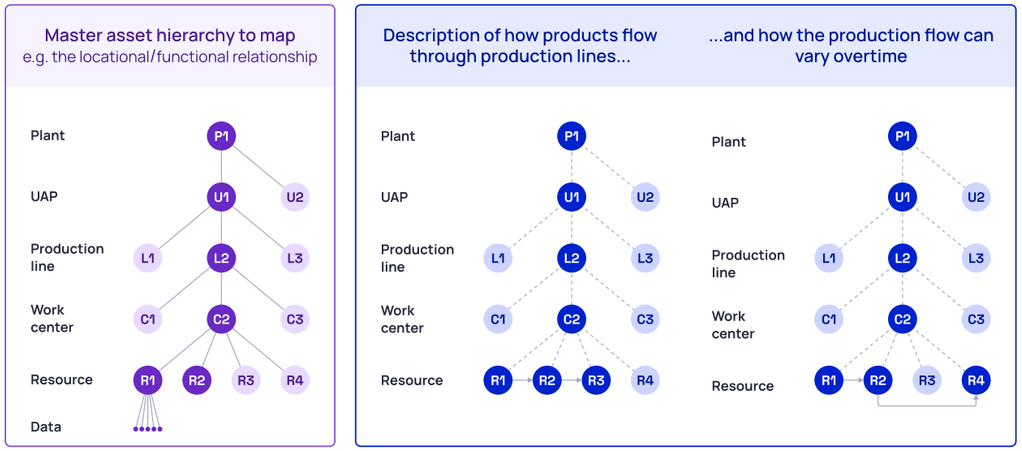




Let's give an example. An asset hierarchy is ideal for addressing use cases related to asset performance management (APM). The relationships resource type also allows the organization of assets in other structures besides the standard hierarchical asset structure. Instead, assets can be organized by their physical location, where the grouping nodes in the hierarchy are buildings and floors rather than systems and functions. Building a model with this structure opens the possibility of solving new use cases like product traceability, where the physical connections of the assets through the production process must be known.

In this way. data becomes an asset, with reusable analytics and scalable models, shareable across many users. This industrial knowledge graph encourages data reuse by creating a user-friendly architecture. By leveraging data effectively and rapidly, the organization can address business opportunities quickly and at scale.

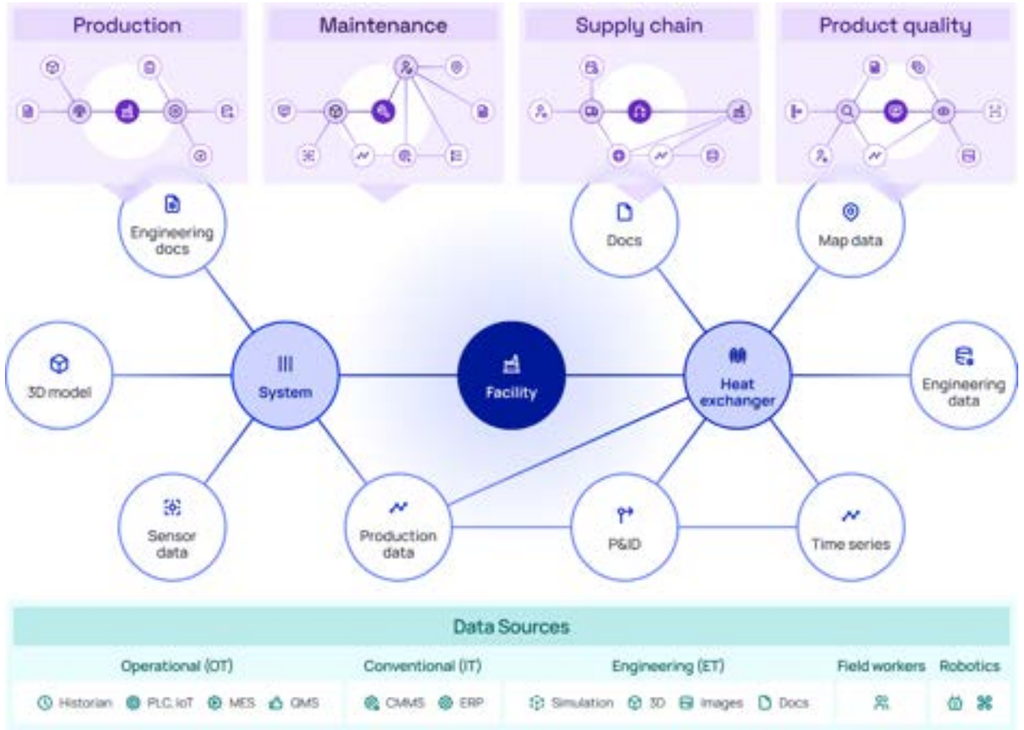
The knowledge graph enables you to capture various data perspectives



Knowledge Graphs and **Generative AI**

According to Gartner's Emerging Tech Impact Radar, Generative AI report, knowledge graph adoption has rapidly accelerated with the growing use of AI because knowledge graphs provide the explicit knowledge, rules, and semantics needed in conjunction with AI/ML methods for pattern recognition. In other words, knowledge graphs deliver trusted and verified data to LLMs and provide rules to contain the model.

The industrial knowledge graph is the foundation for a data model and provides the point of access for data discovery and application development. The most prevalent application of data modeling is to unlock the potential of industrial digital twins. The advantage of data modeling for digital twins is to avoid the singular, monolithic digital twin expected to meet the needs of all and focus on creating smaller, tailored twins designed to meet the specific needs of different teams.



The above graphic shows that a digital twin isn't a monolith but an ecosystem. What is needed is not a single digital twin that perfectly encapsulates all aspects of the physical reality it mirrors, but rather an evolving set of 'digital siblings' who share a lot of the same 'DNA' (data, tools, and practices) but are built for a specific purpose, can evolve on their own, and provide value in isolation.

Like Metcalfe's Law and the understanding of the exponential value of a network, data in the industrial knowledge graph becomes increasingly valuable as people use, leverage, and enrich that data. More useful and high-quality data leads to more trusted insights. More trusted insights lead to higher levels of adoption by subject matter experts, operations and maintenance, and data science teams. A user-friendly, AI-powered experience ensures adoption and use will grow, and this cycle repeats exponentially.

In this way, knowledge graphs are a key underlying technology and act as the backbone for generative AI solutions across business functions that will drive business impact, including:

- **Digital workplace** (e.g., collaboration, sharing and search)
- **Automation** (e.g., ingestion of data from content to robotic process automation)
- **Data exploration** (e.g., providing deeper insights into structured and unstructured data)
- **Data management** (e.g., metadata management, data cataloging, and data fabric)

Despite the undeniable benefits of knowledge graphs, Gartner identifies several challenges to successful implementation. Let's take a look at how we can address these challenges:

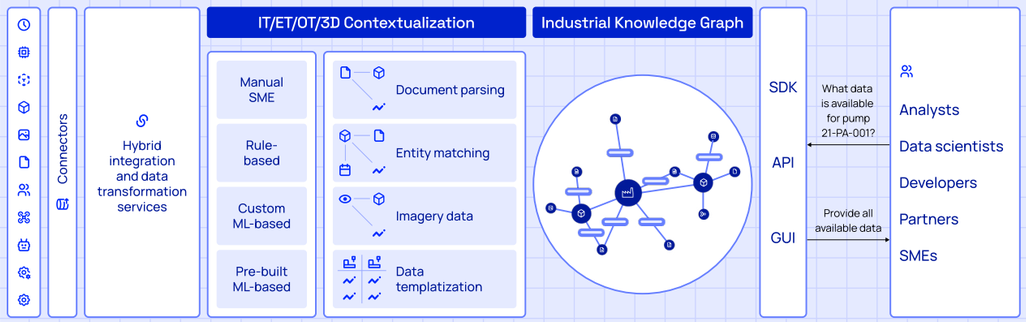
Challenge 1

Challenge
Immature Scaling Methods

As knowledge graphs transition from prototypes to production, methods to maintain their scalability while ensuring reliable performance, handling duplication, and preserving data quality are still evolving.

Solution

To provide reliable performance and scalability, organizations must ensure that their knowledge graphs are powered by contextualization services to provide high-quality, trusted insights that lead to higher levels of adoption by the teams across the enterprise.



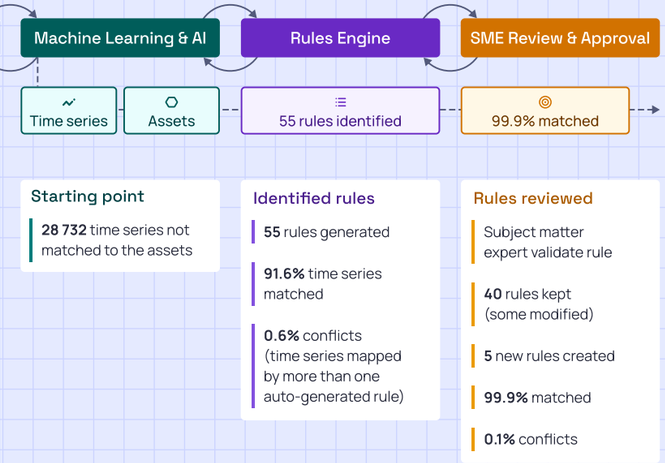
Challenge 2

Challenge
Interoperability

Enabling internal data to interact with external knowledge graphs (meaning connecting data and graphs that vary in scope, ownership, data types, etc) seamlessly remains challenging. Overcoming this hurdle is vital for the creation of a truly interconnected and interoperable industrial ecosystem.

Solution

To establish and maintain interconnection and interoperability, we must ensure access to fully documented and open APIs that help facilitate connections between different data systems, industry standards models, or third-party applications. Plus, strong contextualization capabilities ensure the necessary background for meaningful integration and interpretation of information, especially when that information is trapped in a siloed data source.



Challenge 3

Challenge
Scarcity of In-House Expertise

Particularly among small and midsize businesses, expertise in knowledge graphs is scarce. Identifying and accessing third-party providers with the necessary proficiency becomes a significant obstacle.

Solution

Working with a third-party provider with expertise in building industrial knowledge graphs and industrial data management should not be that scary, especially if you know what to avoid during decision-making when purchasing software and what software deployment type works best for your organization and goals.

To be effective in complex industrial settings, a knowledge graph must include:

- Automated population with contextualization, cross-source-system IT, operational, and engineering data;
- A robust, well-documented API integration; and
- Extremely performant, real-time, flexible data modeling.

Section 2

The Business Value of AI

Chapter 4
AI Is the Driving Force for Industrial Transformation 90

4.1 Verdantix View: Industrial DataOps in 2024 94
4.2 AI Will Deliver Untapped Value for Asset-Heavy Enterprises 102
4.3 Democratizing Data: Why AI-Infused Industrial DataOps Matters to Each Data Stakeholder 104

Chapter 5
Use Cases 108

5.1 Industrial Use Cases Require a System of Engagement 110
5.2 Cognite Data Fusion®: An SOE to Scale Operational Use Cases 114
5.3 Improving RCA with AI Agents and Industrial Canvas 122
5.4 Examples of Industrial AI Agents 126

Chapter 4

AI Is the Driving Force for **Industrial Transformation**

Introduction

AI Is the Driving Force for **Industrial Transformation**

Value Capture Reality: Despite supportive technology trends in abundance, data activation remains stuck in POC purgatory

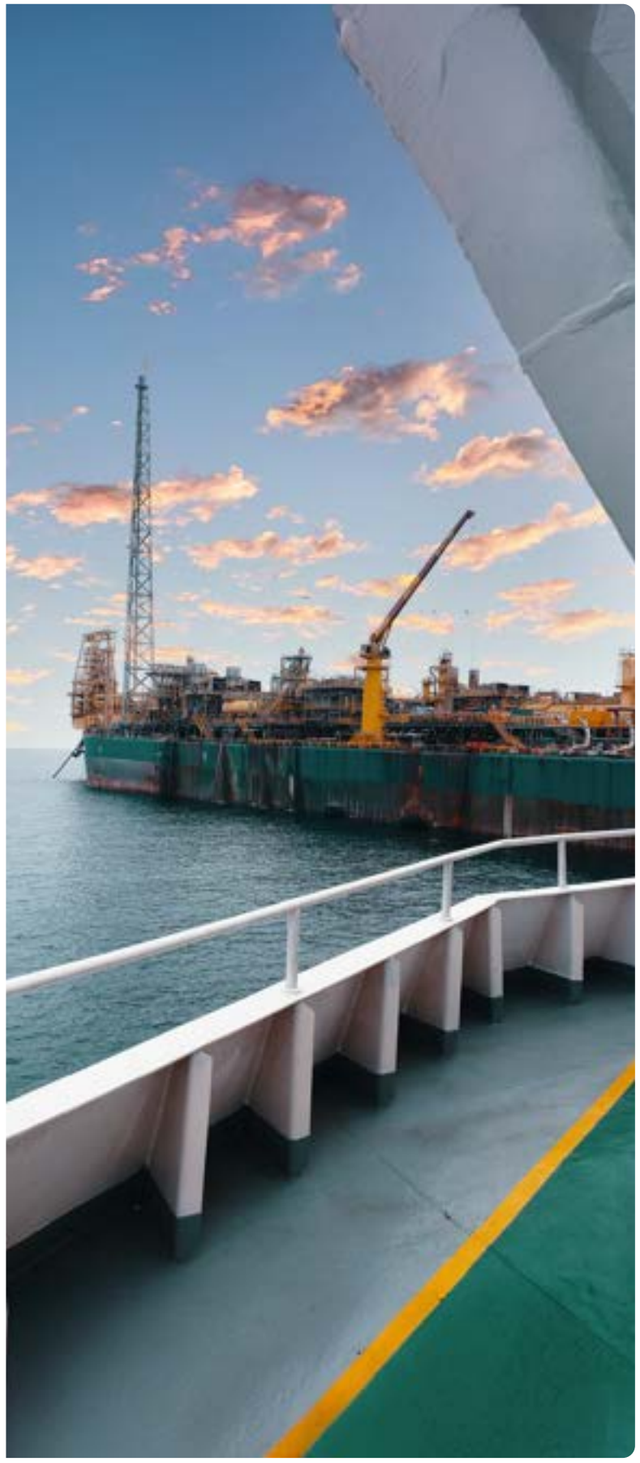
ABUNDANT digitalization ambition and necessity

MOST industrial companies have run at least a handful of AI PoCs

MANY PoCs are considered successful

FEW PoCs ever make it into production

VERY FEW PoCs end up providing substantive ROI (by becoming proofs-of-scale)



There are two discomforting truths within digital transformation across asset-heavy industries:

- **Digitalization Proof-of-Concepts (PoCs) are commonplace.** Real Return on Investment (ROI) isn't.
- **Organizations invest billions in cloud data warehouses and data lakes.** Most data ends there, unused by anyone for anything.

The Challenge

Only one in four organizations extracts value from data to a significant extent. Data dispersion and a lack of tools and processes to connect, contextualize, and govern data stand in the way of digital transformation.

The Opportunity

Industrial DataOps, infused with AI, promises to improve the time to value, quality, predictability, and scale of the operational data analytics life cycle. It provides the opportunity to offer data science liberation within any product experience while simultaneously allowing subject matter experts to acquire, liberate, and codify domain knowledge through an easily accessible and user-friendly interface. This is a stepping stone to a new way of managing data in the broader organization, enabling it to cope with growing data diversity and serve a growing population of data users.

Verdantix View: Industrial DataOps in 2024

Joe Lamming
Senior Analyst, Operational Excellence

As we enter the second half of 2024, the true potential of AI in industrial operations is becoming increasingly evident. Industrial DataOps continues to be foundational and—in many ways—at the forefront of this revolution, offering a robust framework for managing and effectively utilizing vast quantities of data generated by asset-heavy industries.

At Verdantix, we define industrial data management solutions as software that facilitates a professional approach to managing data, improving data quality, and facilitating collaboration between domain experts and data scientists—all while constructing data pipelines for tracking and verifying data origins from the moment of acquisition through usage and to eventual deletion.

The Evolution and Importance of Industrial DataOps

Verdantix has been a keen observer of the data analytics and AI landscape since our inception in 2008. We have witnessed significant shifts, from the IoT boom in the 2010s to the rise of Big Data for industrial asset management. Our focus sharpened on Industrial DataOps around 2021 and today, the drivers for industrial data management solutions are stronger than ever:

1. **Successful AI Deployments:** Effective AI models require high-quality, real-time data from diverse sources. Industrial DataOps ensures that this data is properly aggregated, cleaned, and ready for use in sophisticated AI models for anomaly detection, recommendations, and predictions.

2. **Top-Down Pressure:** Business leaders need access to both granular and big-picture data to make informed decisions quickly. Industrial DataOps provides this data in a concise, visualized format, enabling self-service analytics and real-time decision-making.

3. **Operational Efficiency:** Initiatives aimed at cost reduction, production optimization, and operational safety all benefit from robust data acquisition and seamless integration with analytics tools. Industrial DataOps plays a critical role in identifying improvement areas and boosting collaboration between departments.

4. **Sustainability Reporting:** With increasing regulatory pressure for ESG and sustainability reporting, industries need robust data management systems to comply with these requirements. Industrial DataOps ensures that all necessary data is accurately tracked and reported.

Market Trends and Challenges

The Verdantix Market Size & Forecast for industrial AI-focused analytic solutions, published in December 2023, reveals a compound annual growth rate (CAGR) of just under 24% – significant in the industrial software space. The biggest spend areas are asset condition monitoring, predictive maintenance, and product and process management, with substantial growth also seen in supply chain optimization and production and process management.

Survey data from our Operational Excellence Global Corporate Survey also highlights the importance and challenges of data aggregation, access, collaboration, and contextualization. Data aggregation remains a critical challenge for many, particularly in process manufacturing and energy sectors. Data access and collaboration between data scientists, operations, and maintenance executives are also seen as significant challenges, with a clear need for improvement in the energy sector.

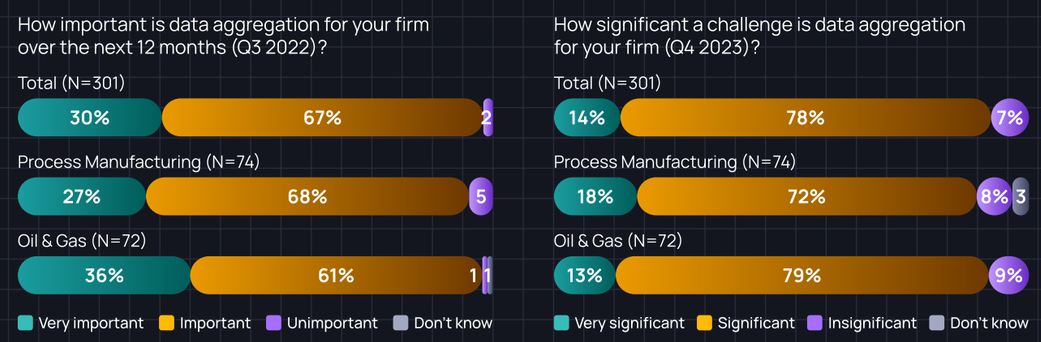
Overcoming Data Integration Challenges

A major challenge in deploying AI in industrial settings is data integration. We see a number of vendors tackle this issue with platform solutions that provide comprehensive data modeling services able to orchestrate diverse data types from disparate sources. We have witnessed Cognite Data Fusion's ability to enable the creation of detailed data models that standardize information across equipment and processes, making it easier to analyze and use.

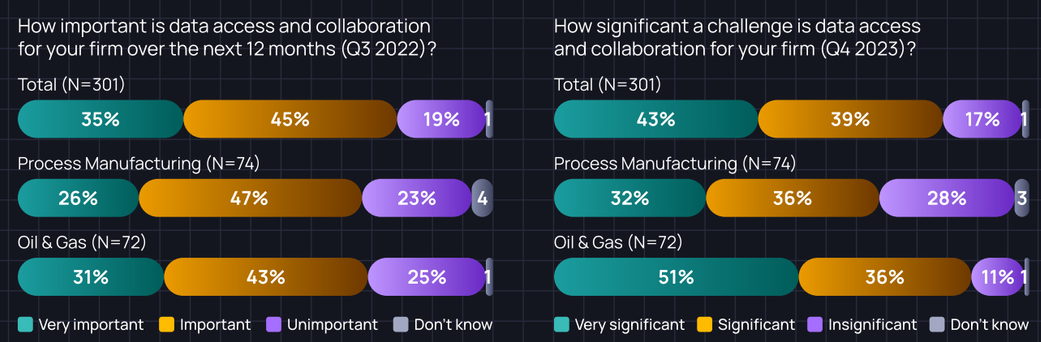
Enhancing Data Quality and Contextualization

Ensuring data quality is another critical aspect. Automated monitoring of sensor data, cross-referencing tags, and enforcing data governance frameworks are essential for maintaining the high data quality needed for trustworthy analytics. Additionally, contextualizing data – assembling it into digestible formats for data scientists and decision-makers – expedites its utility and resultant time-to-value.

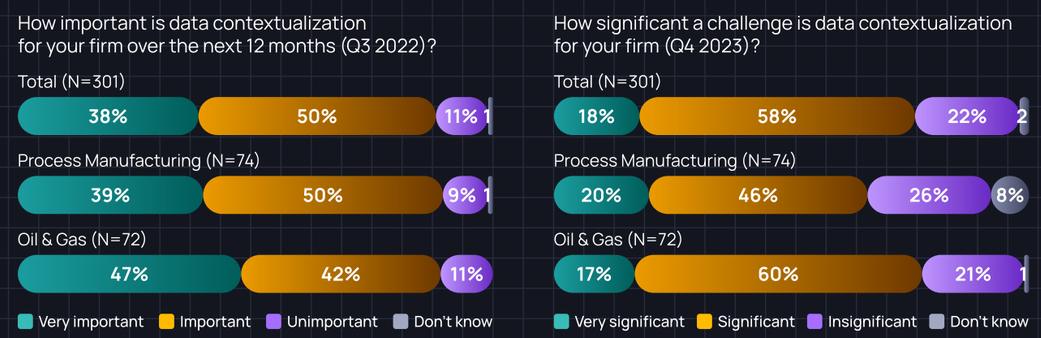
Getting Access to Data Is Industrial Firms' #1 Priority



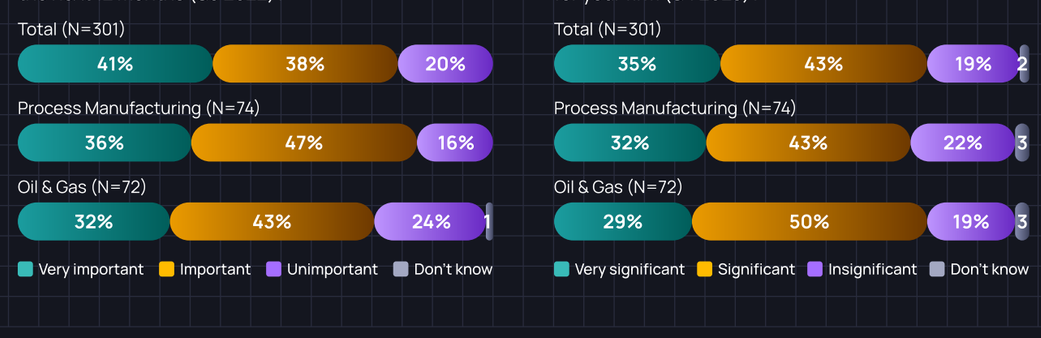
Providing Low-Code Access to Both Data Scientists and Operations and Maintenance is Top of Mind



Effective Data Contextualization Continues to Be an Important Challenge for Industrial Firms



Managing Data Quality Through Robust Governance Is a Significant Priority for Firms in 2024



Source: Verdantix Operational Excellence Global Corporate Surveys 2022 and 2023. **Notes:** Figures rounded to the nearest integer. Percentages lower than 5 are written as numbers.

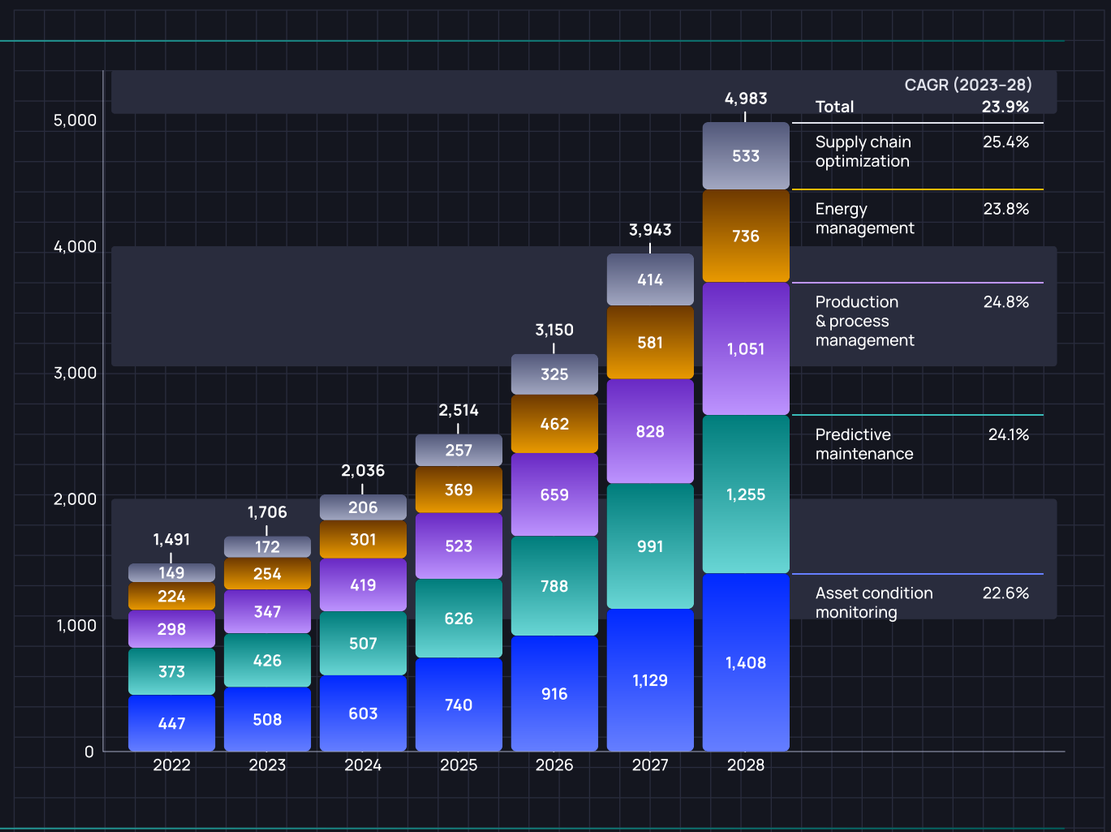
The Impact of Generative AI

AI's role in industrial transformation is multifaceted. One of its key applications is in predictive maintenance, with ML-driven anomaly detection and forecasting models analyzing data from sensors to predict equipment failures before they occur, thereby providing the opportunity to optimize production around scheduled downtime and reduce maintenance costs. AI tools are also used in asset condition monitoring, process optimization, and supply chain management, driving similar efficiency gains across operations.

Generative AI is poised to fundamentally reshape core administrative tasks within industrial operations through vastly more capable automation and improved data discoverability. Our survey data indicates that only 5% of respondents foresee no changes due to generative AI. Many expect Gen AI to automate administrative tasks, improve data access, and enable direct querying of technical documentation.

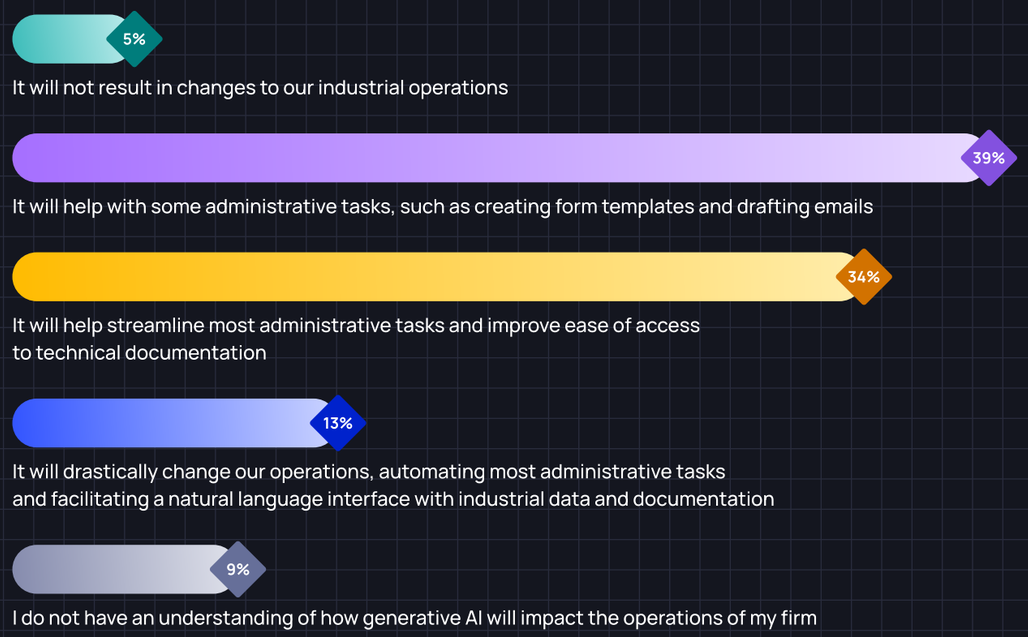
However, challenges such as regulatory compliance and the complexity of deploying AI models, especially those fine-tuned on domain-specific data, must be addressed. Partnering with data management software vendors can help overcome these hurdles and achieve successful AI implementations.

Data-Driven Decision Support Such as AI Analytics Are Expected to See Nearly 24% CAGR Until 2028

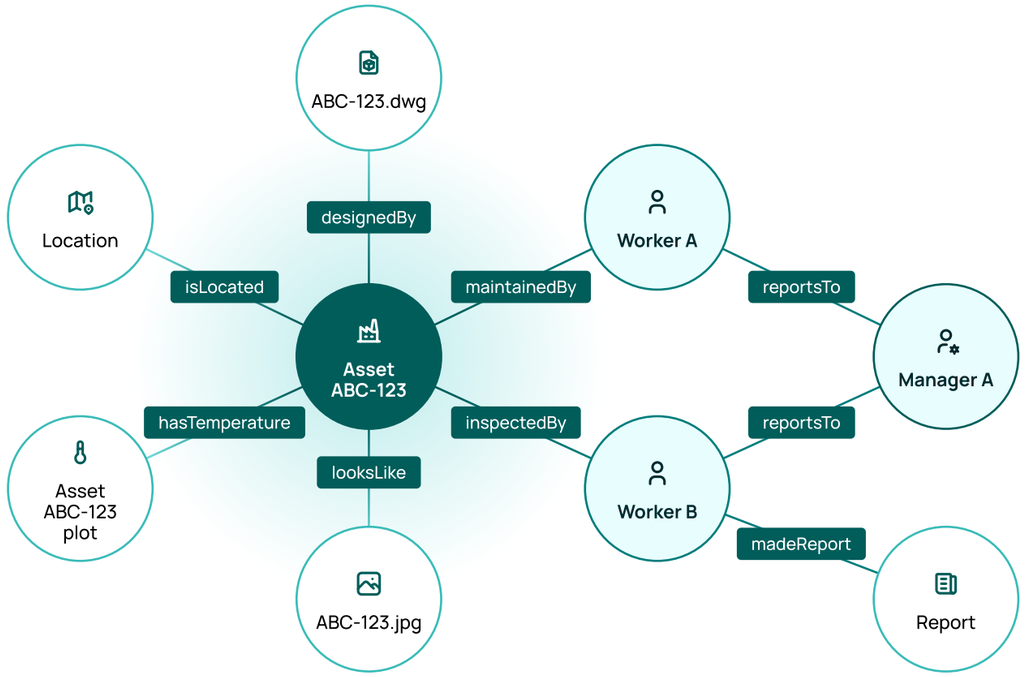


Source: Verdantix Market Size and Forecast Industrial AI Analytics 2022–2028 (Global) **Note:** Analysis published December 2023

Industrial Firms Are Interested in Generative AI – Necessitating Deeply Contextualized, Accessible Data



Source: Global Corporate Survey 2023: Operational Excellence Budgets, Priorities and Tech Preferences



Innovative Data Modeling Strategies

Key to making AI for industry work are graph databases—or industrial knowledge graphs. These databases capture and illustrate semantic relationships between entities—workers, assets, process documentation, and even 3D models. This approach reduces duplicate siloed data and enhances data discovery, improving search capabilities and enabling big-picture network analysis for critical processes. Such semantic data models also provide a solid foundation for AI agents, enterprise search engines, and AI copilots, reducing the risk of hallucinations by guardrailing outputs with clear, contextualized data.

Collaboration with data management software vendors is crucial to overcoming these hurdles and achieving successful AI implementation. The best data management solutions deliver six critical capabilities:

- Easy-to-configure data connectors
- Pre-built data quality management
- Integrations with popular BI tools
- Contextualization through both semantic and asset hierarchies
- Data discovery with knowledge graphs
- Collaborative GUIs augmented with Gen AI

As we look toward the close of 2024, it is evident that AI, supported by robust Industrial DataOps, will be the driving force behind industrial transformation. The integration of AI into data management processes not only enhances operational efficiency, but also empowers industries to navigate the complexities of modern industrial operations with greater agility and insight.

Industrial DataOps platform providers at the frontier are a critical pillar of this revolution, enabling industries to harness the full potential of AI and data-driven decision-making. The future of industrial operations is not just about managing data; it is about transforming that data into information – actionable intelligence that drives better decisions, better outcomes, and better explainability. The journey towards a data-driven future was already well underway. Now we have rocket fuel.

AI Will Deliver Untapped Value

for Asset-Heavy Enterprises

To extract the value of industrial data insights adequately, it is essential to make operationalizing data the core of your business strategy. Data must be available, useful, and valuable in the industrial context. Operational teams need a robust data foundation with a strong data context and interpretability backbone, all while applying generative AI to accelerate workflows that optimize production and make operations more efficient.

1.

Efficient Data Management and Improved Data Accessibility

A strong data foundation is required to remove the risk of 'hallucinations' and increase AI 'readiness.' An Industrial DataOps foundation maximizes the productive time of data workers with automated data provisioning, management tools, and analytic workspaces to work with and use data safely and independently within specified governance boundaries. The approach can be augmented with AI-based automation for various aspects of data management—including metadata management, unstructured data management, and data integration—enabling data workers to spend more time on use case development. Using AI to enable rapid ingestion and contextualization of large amounts of data brings a paradigm shift in how the organization accesses business-critical information, improving decision-making quality, reducing risk, and lowering the barriers to (and skills for) data innovation.

2.

Augmented Workflows and Process Improvements, Driving Innovation at Scale

Using generative AI-powered semantic search, what used to take your process engineers, maintenance workers, and data scientists hours of precious time will take only a few seconds. With the guidance of generative AI copilots, users can generate summaries of documents and diagrams, perform no-code calculations on time series data, conduct a root cause analysis of equipment, and more. Time spent gathering and understanding data goes from hours in traditional tools to seconds. Now, users can spend more time driving high-quality business decisions across production optimization, maintenance, safety, and sustainability.

3.

Rapid Development of Use Cases and Application Enablement

Too often, digital operation initiatives get trapped in 'PoC purgatory,' as scaling pilots takes too long or is too expensive. Using an AI-infused Industrial DataOps platform shortens the time to value from data by making PoCs quicker and cheaper to design and offering operationalizing and scaling tools. These copilot-based approaches leverage the power of natural language to understand and write code based on published API documentation and examples to support development processes. Generative AI further improves ML training sets of ML models by generating synthetic data, enhancing the data set used for training, enhancing process efficiency, and optimizing production. Some common use cases in asset-heavy industries are maintenance workflow optimization, engineering scenario analysis, digitization of asset process and instrumentation diagrams (P&IDs) to make them interactive and shareable, and 3D digital twin models to support asset management.

4.

Enterprise Data Governance as a By-Product and Personalized AI Tools

By having a strong Industrial DataOps foundation, you can then empower users to adapt AI models to cater to their specific requirements and tasks, using generative AI to enhance data onboarding, complete with lineage, quality assurance, and governance, while a unique generative AI architecture enables deterministic responses from a native copilot. Additionally, Industrial Canvas overcomes the challenges of other single pane-of-glass solutions, which often over-promise capabilities and are too rigid with prescribed workflows. This prevents users from working with the data how they choose by delivering the ultimate no-code experience within a free-form workspace to derive cross-data-source insights and drive high-quality production optimization, maintenance, safety, and sustainability decisions. If implemented successfully, an AI-augmented data platform provides consistency and ROI in technology, processes, and organizational structures, with better operations data quality, integration and accessibility, and stewardship. It should also enhance data security, privacy, and compliance with tracking, auditing, masking, and sanitation tools.

Democratizing Data:

Why AI-Infused Industrial DataOps Matters to Each Data Stakeholder

Extracting maximum value from data relies on applying advanced models to produce insights that inform optimal decision-making, empowering operators to take action confidently. Turning insight into action is what we mean by operationalizing data into production for value.

But for every person who can 'speak code,' hundreds cannot.

Generative AI will change how data consumers interact with data. It facilitates a more collaborative working model, in which non-professional data users can perform data management tasks and develop advanced analytics independently within specified governance boundaries. This democratization of data helps store process knowledge and maintain technical continuity so that new engineers can quickly understand, manage, and enrich existing models. It is about removing the coding and scripting and bringing the data consumption experience to the human user level.

Making the data speak human is the only way to address the Achilles heel of practically all data and analytics solutions, especially those for heavy-asset industry verticals. These organizations face many challenges: an aging workforce, extreme data type and source system complexity, and very low classical data literacy among SMEs —those needing data to inform their daily production optimization, maintenance, safety, and sustainability decisions.

Why It Matters to Executives

Priorities	Challenges	Values
Financial performance and profits	ROI on investment, cost of downtime	Enable data-driven decision-making that allows focus on the highest ROI activity at any given time
Innovation and solution delivery	The measures and KPIs used in decision-making are entrenched in showing short-term value	Serve as a bridge to increased digital maturity, as it carries forward the momentum and infrastructure to develop data analytics catalogs and libraries that can then be deployed with fewer services and at lower marginal costs
Reduce inefficient processes that waste time and effort	Scaling on assets and equipment: projects are slow to deploy and done in isolation	Align and bring together formerly isolated subject matter experts (SMEs), cultures, platforms, and data deployed by IT and OT teams to improve operational performance through unified goals and KPIs
Impact reputation, sustainability, and future viability	Poor reputation and sustainability	Empower organizations to operate with greater precision and track the correct metrics to reduce the impact on the environment
Having the best workforce	Scaling the number of people creating solutions; lack of skill set	Enable existing non-professional data users to perform some data management tasks and drive value for the enterprise while preserving their valuable accumulated knowledge and experience
Remaining relevant and competitive	Missing the digital transformation revolution	Leverage data effectively and rapidly to answer questions and gain some insulation from market volatility
Enabling digital transformation	Confusion about how to deliver digital transformation and what it means for both management and workers	Enable the organization to meet the need for fast-moving innovation by providing consistency and ROI in technology, processes, and organizational structures, with better operation data quality, integration and accessibility, and stewardship
Strategic change in culture and vision	Reactive culture is an obstacle to growth	Recognize data as an enterprise asset and build a path toward digital maturity through tools and processes so that digital ways of working become effortless across a broad range of stakeholders

Why It Matters to IT and Digital Teams

Priorities	Challenges	Values
Data preparation and integration	Legacy systems. Complex integrations and dependencies between multiple data sources	Minimize existing data silos and helps simplify the architecture to support rapid development and deployment of new analytics
Create solutions for operations by turning insights into actionable advice	There's too much data, with no context; challenge in making data usable and getting models in operation	Offer a workbench for data quality, transformation, and enrichment, as well as intelligent tools to apply industry knowledge, hierarchies, and interdependencies to contextualize and model data
Automate workflows	Scale across insights, solutions, sites	Facilitate industrial equipment and processes data models and templates—it talks domain language and scales models from one to many
Minimize the hurdles in cross-functional collaboration	Multiple hand-offs that are error-prone and increase risks	Empower organizations to operate with greater precision and track the correct metrics to reduce the impact on the environment
Driving the implementation of Industry 4.0	Difficult to explain business-wide benefits to senior decision-makers	Connecting data users with disparate operational data sources helps bridge those divides on the path to use-case operationalization

Why It Matters to Domain Experts

Priorities	Challenges	Values
Optimization of process around quality, throughput, and yield	Lack of insights or tools to make quick and correct decisions around maintenance and production	Leverage domain knowledge and human expertise to provide context and enrich data-driven insights, and further develop machine learning models that utilize data to improve planning processes and workflows
Inefficient processes waste time and effort	Work in isolation and without full possession of all the data and the facts	Provide the capabilities domain experts need to support self-service discovery and data orchestration from multiple sources
Resource management	Reluctancy and slow adoption of new tools and tech in old ways of working	Empower business functions to use data and support the digital worker
Improve planning process and workflow across assets and equipment	Manual tasks are time-consuming and prone to errors	Industrial DataOps infused with AI enables your site to solve today's challenges with a pathway towards more autonomous systems and sustainable growth

Solving the industrial data and AI problem is critical to realizing value from digitalization efforts. Benefits can be measured from streamlined APM workflows, improved SME productivity, optimized maintenance programs, and real-time data efficiencies, such as:

- **Productivity savings due to improved SME efficiency.** Industrial DataOps provides data accessibility and visibility, transforming how data scientists and SMEs collaborate.
- **Reduced shutdown time.** The opportunity cost of large industrial assets being out of production is significant. Using a digital twin and better component data visibility, SMEs are able to safely minimize shutdown periods when data anomalies arise.
- **Real-time data access enables Improvement in productivity.** Live data access enhances operational flexibility and decision-making by increasing site safety, improving predictive maintenance, and raising machine performance.
- **Optimized planned maintenance.** Cognite Data Fusion creates contextualized data to optimize planned maintenance by analyzing and interpreting available resources, workflows, and component life cycles.
- **Energy efficiency savings.** Intelligent data can be used to reduce energy use and, thus, operational costs.
- **Optimization of heavy machinery and industrial processes.**
- **Health and safety.** Reduce the amount of human movement through potentially dangerous 'hot' areas, reducing risks to employee health and safety.
- **Environmental, social, and governance (ESG) reporting.**

Chapter 5

Use Cases

Industrial Use Cases Require a **System of Engagement**

Before we dive into specific industrial agent use cases, let us take a moment to discuss a System of Engagement and why it must be paired with a System of Record to deliver an effective user experience across industrial use cases.

As discussed in detail throughout this book, the landscape of industrial data is still a black box; complex and diverse data types are difficult to unify in a way your teams can understand. Industrial data is unique in the volume of time series data, the wide spectrum of unstructured data (images, 3D models, documents, drawings, etc.), and the sheer number of sources (Historians, MES, QMS, IoT, etc.) that generate and store data.

When solving site-level and enterprise industrial use cases, breaking down data silos is not enough. Teams and people need an intuitive way to understand which time series connects to which equipment, and what is upstream and downstream of that equipment in your operation.

Unifying disparate operational data only delivers business impact when it is contextualized with industrial domain expertise, and is available through a simple, interactive user experience.

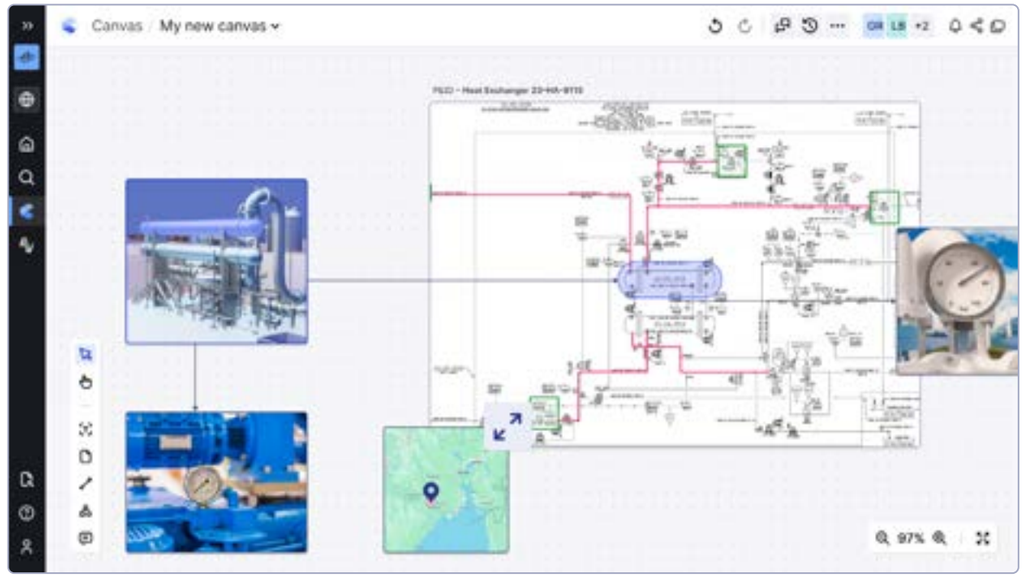
Industrial organizations have solved many enterprise-level use cases around customer experiences, finance, and supply chain with their System of Record (SoR), such as a data lakehouse, a data warehouse, or a data cloud.

The SoR already contains finance and accounting, supply chain, sales, and inventory management data; additionally, the desire to leverage the SoR to solve site-level, operational use cases such as asset performance management, production optimization, or EHS (environment, health, and safety) is logical. If the relevant operational data were to be added to the SoR, it could be extended to solve operational use cases. **If only it could be this simple!**

In reality, the site-level users of operational data—subject matter experts, production engineers, operations, maintenance, and reliability teams—have unique, real-time needs that require simple access to industrial data in a performant and interactive way.

To meet these needs, these users require a System of Engagement (SoE) to drive high-quality, AI-powered decisions that can leverage the existing SoR information and enrich it with site-specific data that is absent and difficult to incorporate into a SoR.

Before going any further, it is important to level-set the differences between a System of Record and a System of Engagement.



How Do You Define a System of Engagement and a System of Record?

While an SoR and an SoE play crucial roles in an industrial organization's technology landscape, they serve different purposes and handle different data types.

SoRs excel when working with structured data and are the trusted, authoritative source for essential business information. On the other hand, SoEs prioritize user engagement and provide dynamic interfaces for interactions between the organization and its stakeholders.

	System of Engagement for Industrial Operations	System of Record
Purpose	Designed for industrial environments where real-time data is needed for decisions, planning and operational actions. Interactive-level user-friendly experience. Self-service and automation to help users quickly access the data they need. Mobile-friendly.	Centralized repository that stores and manages primarily structured data. Authoritative source of truth for an organization's critical data, such as customer information, financial transactions, inventory levels, and other essential business data.
Industrial Use Case Examples	▪ Digital Operator Rounds ▪ Maintenance execution ▪ Root Cause Analysis ▪ Turnaround planning ▪ Production optimization ▪ EHS	▪ Supply Chain performance ▪ Production management ▪ Reporting, compliance, audits
Common Data Types	**Operational Data** ▪ Time series data (highly optimized) ▪ Documents, including interactive P&IDs ▪ Events ▪ 3D, point cloud, visual data ▪ Robotics	**IT Data** ▪ Tabular data ▪ Structured data ▪ Non-structured data
Users	▪ Subject matter experts ▪ Production engineers ▪ Operations, maintenance, and reliability teams	▪ Data Engineers ▪ Data Analysts ▪ Data Scientists
Exploring Data	Data is contextualized into an industrial knowledge graph, available for users and AI to search with natural language.	Database schemas are used to define how data is organized, stored, and related, and databases are linked through identifiers.
User Access	User-friendly and accessible through various devices, such as web browsers or mobile apps. Supports both field and remote work while still ensuring the highest level of data integrity and security.	Tightly controlled and restricted to authorized personnel to ensure data integrity and security.

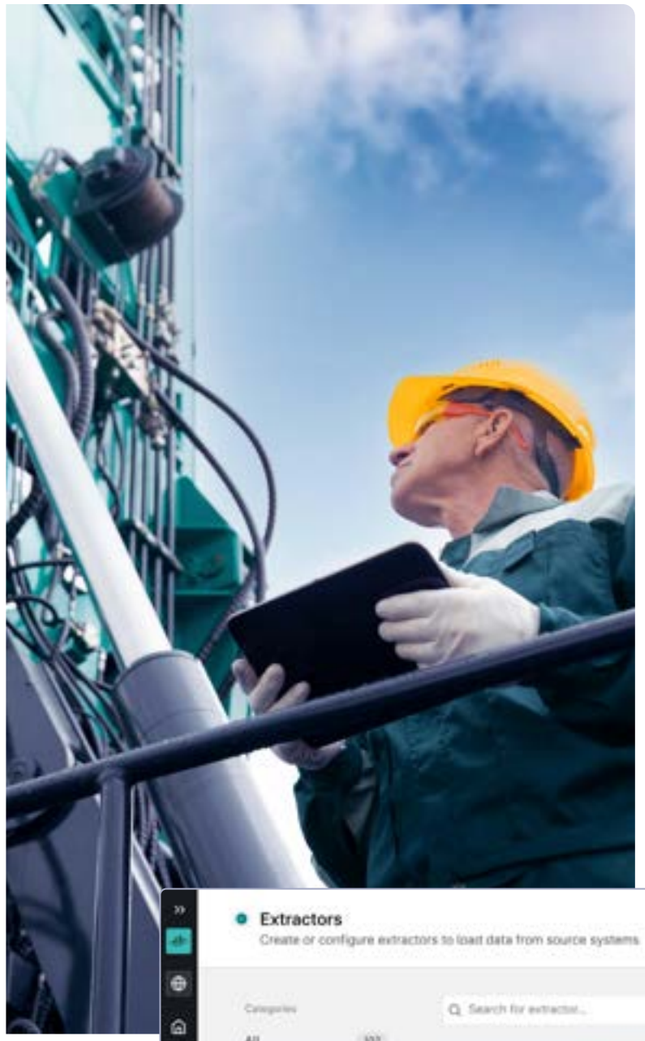
To summarize, an SoR and an SoE offer unique benefits with the clearest differences outlined when focused on the business use case and the end users. While many have already invested in an SoR, pairing this with an SoE can unlock new operational use cases that require an interactive-level user experience.

To provide specific examples of what an SoE should deliver, let's look at Cognite's core product, Cognite Data Fusion®, and how it serves as a trusted SoE for some of the largest energy, manufacturing, and power companies in the world.

Cognite Data Fusion®:

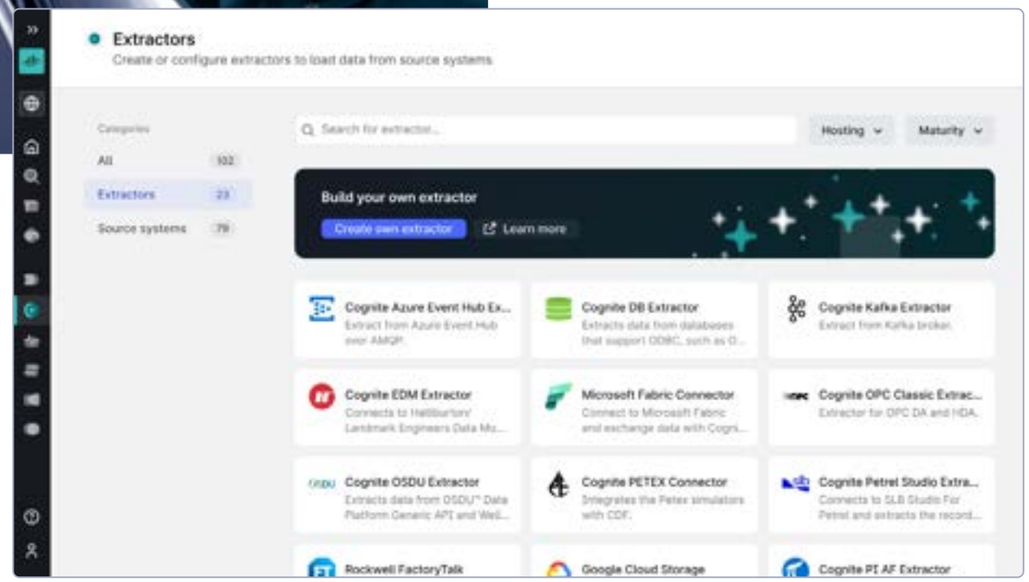
An SOE to Scale Operational Use Cases

For industrial operators to solve operational use cases using an SoE, organizations must unify disparate operational data, contextualize it with industrial domain expertise, and make it available through a simple, interactive user experience. Here are some of the key requirements to solve operational use cases at scale and how Cognite Data Fusion® addresses each:



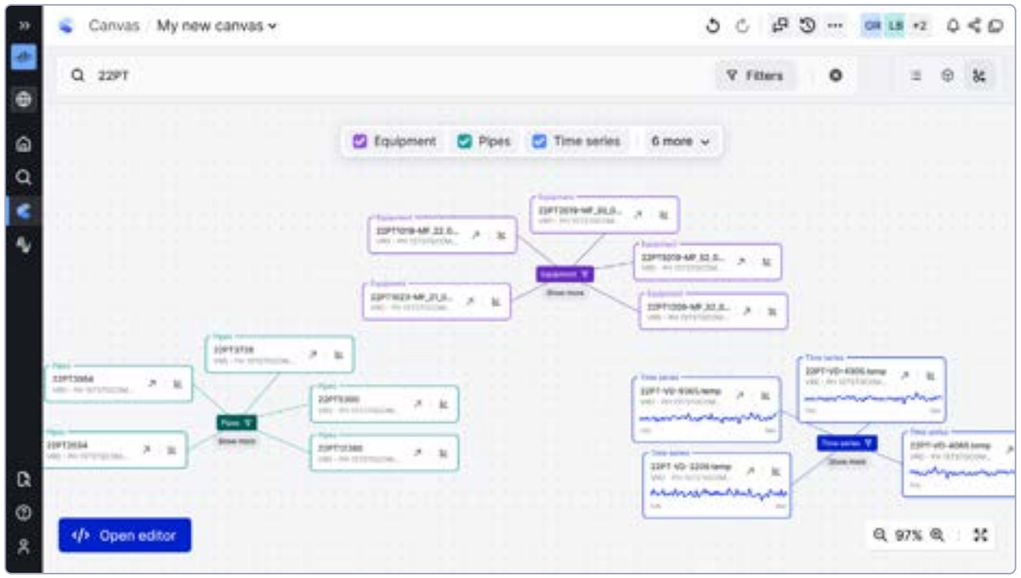
Pre-Build Data Extractors from All IT, Operations, and Engineering Data Sources

An SoE needs to access **all** industrial data. Many SoR solutions can combine IT and engineering data (think Asset Information solutions). Cognite Data Fusion® is unique in its ability to connect OT, IT, engineering and unstructured data with more than 80 pre-built extractors into common systems to reduce the effort to connect voluminous time series OT data. Cognite Data Fusion's highly optimized time series database provides performant recall for users.



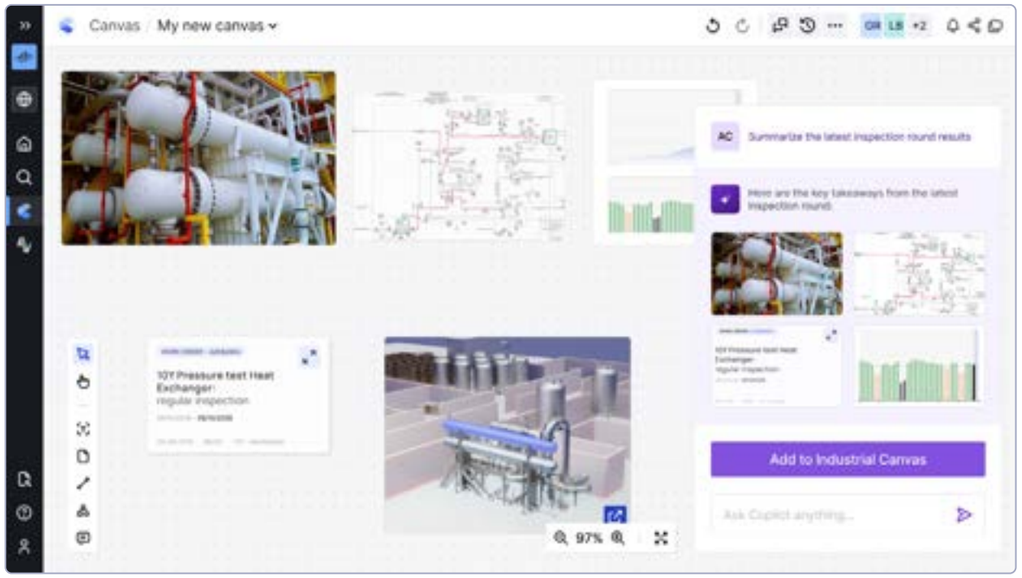
All Industrial Data Types Can Rapidly Be Put into Context

Too often, the data contextualization process is manual, requiring hundreds of hours to understand how time series data from the historian maps to the assets in the ERP asset hierarchy, how open work orders map to equipment, or where that equipment exists in the process. Cognite Data Fusion's unique ability to automate this process with AI-based algorithms can shorten the data contextualization process from months to days.



Data Relationships Are Understandable to All with an Industrial Knowledge Graph

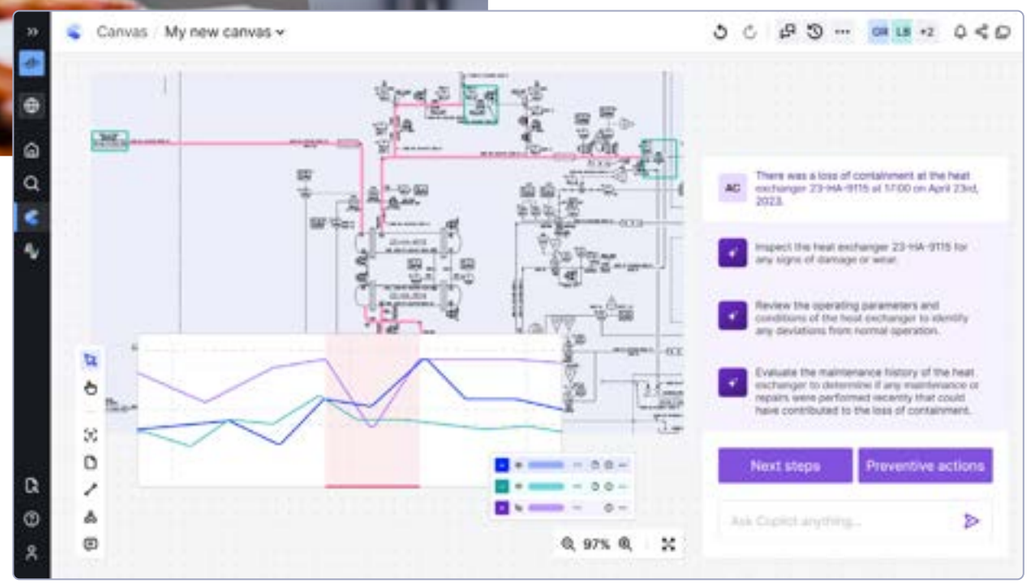
An industrial knowledge graph is the output of data contextualization and represents the connections between the many data types. This approach circumvents the effort required to create schemas when using data from a data lake.



Domain-specific agents further enhance this experience by enabling natural language search and providing recommendations and summaries of information and drawings. SMEs can gather, summarize, and generate insights orders of magnitude more efficiently. Users can rely on these specialized agents to recommend next steps for workflows like root cause analysis, tailored to specific roles within the industrial landscape.

Free-Form Data Exploration with Industrial Agents

The most promising example of an interactive user experience is Cognite Data Fusion's Industrial Canvas. Truly delivering simple access to all industrial data in a single workspace requires a unique way to leverage contextualized data. Users deserve a way to work with live sensor data, interactive engineering diagrams, images, 3D models, and more within a visual workspace to explore data in context, perform root cause analysis, and collaborate by tagging other users in an open, free-form environment.

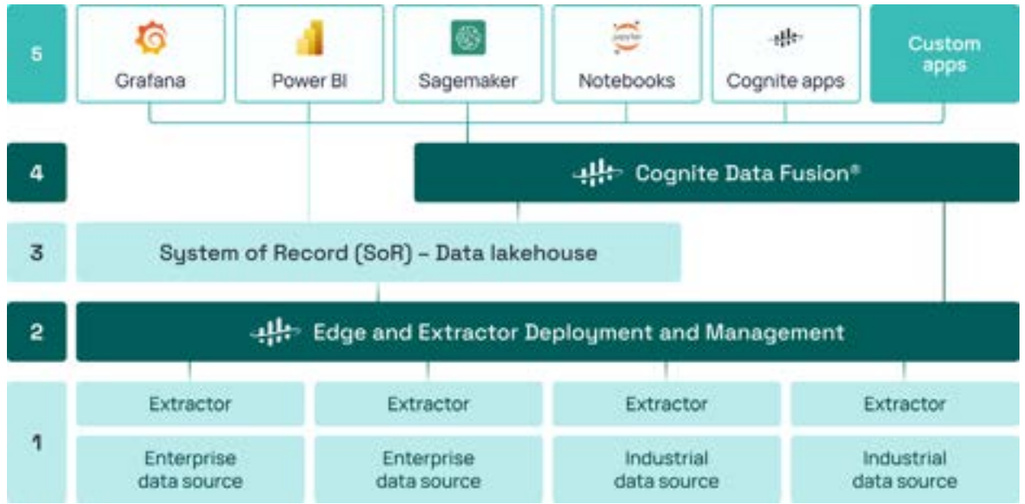


Data Can Be Analyzed with No-Code Charting and Analytics

In addition to simple access to data, users must be able to generate insights to analyze what has happened (root cause analysis) and what is currently happening (processes are running outside of steady state).

Cognite Data Fusion® can combine multiple sources of data—time series, events, and work orders—into a single view with an industrial data science library in a no-code environment to analyze underperforming assets and processes properly. This knowledge must be easy to share across the organization, to set alerts for future occurrences, and to reference. They cannot live in offline spreadsheets that require manual data dumps and updates.

While these requirements are not exhaustive, they illustrate how Cognite Data Fusion®, as an SoE, prioritizes the needs of SMEs, production engineers, operations, maintenance, and reliability teams.



1. **End-user applications:** End users can consume contextualized data through the interfaces and tools they are familiar with.
2. **System of Engagement:** Cognite Data Fusion® is a System of Engagement, providing simple, governed access to contextualized data, including non-structured data types in the data models end-users are familiar with.
3. **System of Records:** Authoritative System of Records, where customers can govern their data.
4. **Extractor orchestration:** Deployment and management of extractors and connectors are handled by the SoR, or by Cognite Extractors.
5. **Data extraction:** A combined portfolio of extractors and connectors are used to extract data from source systems.

Is Cognite Data Fusion® Intending to Replace My SoR?

Cognite Data Fusion®, and more broadly any SoE, is not intended to replace an SoR. Instead, it complements the existing technology stack.

Operational use cases require an interactive-level user experience to make timely decisions using diverse and complex operational data that doesn't reside in an enterprise SoR today. An SoE can unify complex operational data with data already in a centralized data lake, data warehouse, or data lakehouse. In addition, any data or insights generated with an SoE can be written to the SoR, providing even more value to this source of truth.

By leveraging both technologies, the SoR is enriched and maintains its authoritative, trusted role in the organization, and the SoE can unlock new, real-time operational use cases.

Now, let's look at specific examples of industrial agents in action.

Improving RCA with AI Agents and Industrial Canvas

Incident response and the ensuing root cause analysis (RCA) process are essential engineering workflows for maintaining plant operations, improving production processes, and increasing safety, reliability, and productivity over time.

But as pressure mounts for clearer answers in shorter time frames, sometimes with less human domain expertise available, real-time access to relevant data and the means to quickly mine insights are more important than ever.

RCA is Stuck in the Past

In theory, due to decades of novel instrumentation and record-keeping, engineering now has more data than ever—in historians, spreadsheets, documents, emails, manuals, work orders, etc. However, in practice, using this data translates into difficult, cumbersome processes to access what is needed and start an analysis in methodical, collaborative ways. Quick time to resolution matters, especially during every minute of an expensive outage.

What is particularly challenging about the process today?

- **Data is spread across sources, requiring significant manual work.** This means multiple systems and logins, time spent searching for the right documents, verifying potential out-of-date data, and making sense of the last inputs in the spreadsheet.
- **Collaboration is happening all over the place.** RCA is a real-time team sport happening on whiteboards, in email, across text messages, and phone calls, and there is no real, centralized place to get all the needed data and for analysis to come together in short order.
- **Past RCA work is exceedingly hard to reference.** Retaining knowledge and best practices only matter if you can reuse them, but today, it is difficult to refer to previous RCA outcomes so that users can easily learn from or compare similar incidents or rule out certain contributing factors.

Industrial RCA is Ripe for Change.

Chemical, energy, and other process industries have been trying to become less reactive and more predictive for decades, with mixed results.

By solving the fundamental challenges with data discovery, access, and collaboration, engineers can spend more time developing compelling narratives on how the root cause was identified, what evidence was used to support the findings, and how future issues can be addressed.

With simple access to complex industrial data in the frame of a more collaborative environment, teams can spend less time on analysis while still developing higher-confidence engineering outcomes and making it easy to reference in the future.

What Does a More Simple RCA Workflow Look Like in Practice?

Let's explore the three key features below, available in Cognite Data Fusion's Industrial Canvas, the digital workspace for data-driven planning, troubleshooting, and operational insights.

Feature 1:

Intuitive AI-Enabled Access to All Relevant, Contextualized Data

Start from an interactive engineering diagram to explore and source additional related data with just a few clicks.

Drag and drop time series charts, P&ID drawings, 3D models, tables, images, and more to start narrowing down your analysis quickly.

Maintain confidence in your data as it is sourced from the evergreen industrial knowledge graph, ensuring explainable and reliable results.

Feature 2:

AI-Assisted Document Summarization and Ability to Ask Questions Using Natural Language

Use a generative AI-based copilot to quickly find equipment, tags, or data points through intelligent document-based searches, saving you hours of manual effort.

When adding long-form documents, you can generate summaries or quickly search for key information using natural language. Add these summaries or insights to your canvas so others can easily access and use them.

Feature 3:

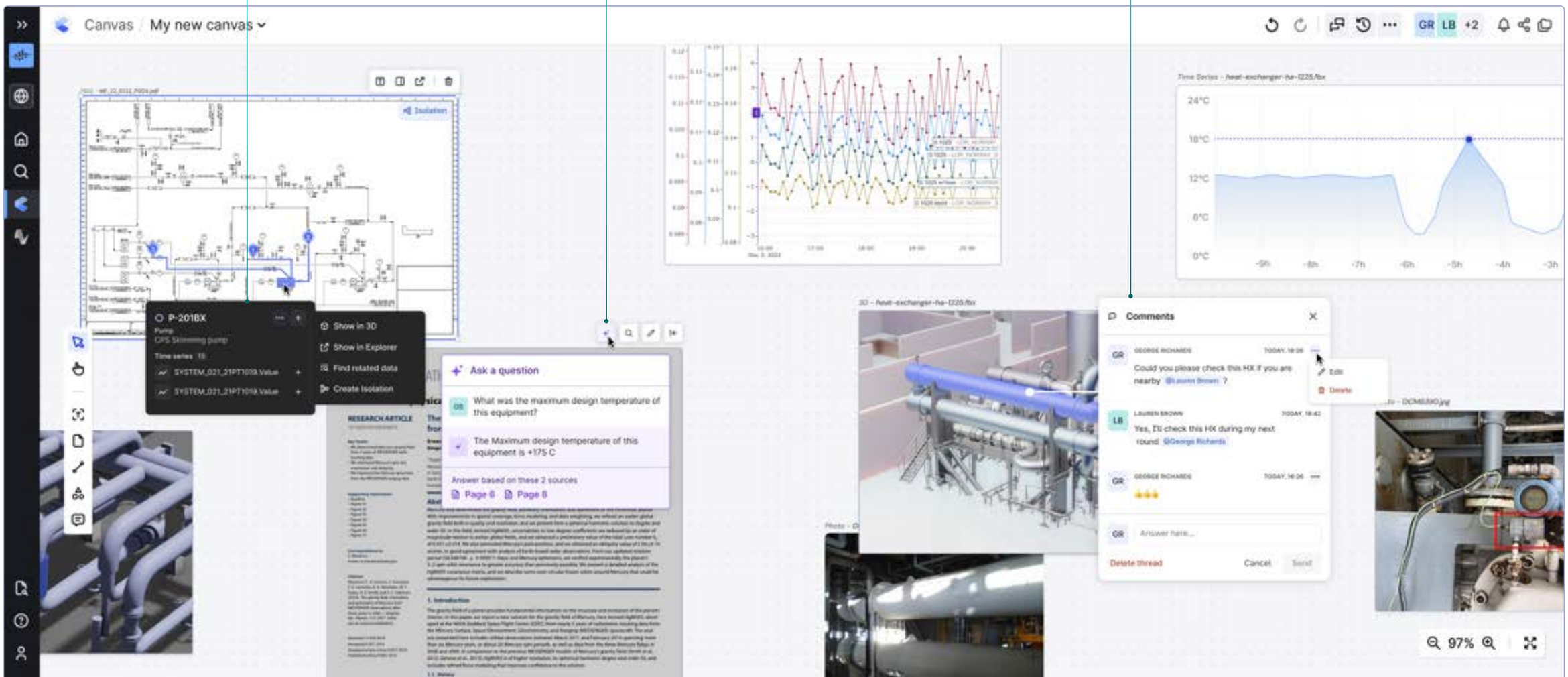
One Workspace to Collaborate on Analysis and Share Insights

Create your RCA narrative as a "storyboard" in which you can arrange and link related data and ideas to understand the flow of your analysis better.

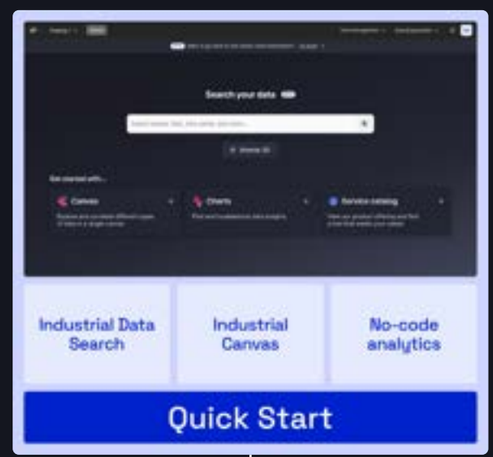
Tag colleagues in comments, annotate, share your analysis across teams, and save your investigation for future reference to help others troubleshoot the same or similar equipment or workflows more quickly.

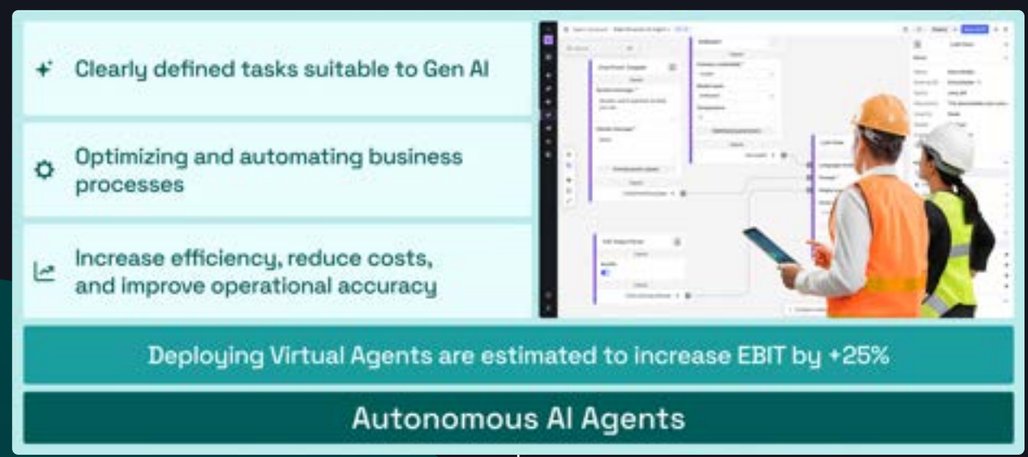
It's Time to Bring More Effective RCA Practices into the Engineering Workflow

Smart engineering thinking is critical for the modern operating floor. Here, simple access to complex industrial data allows SMEs and engineers to focus on critical thinking with all the data they need at their fingertips, so they can spend less time searching and gathering data for their RCA.



Examples of **Industrial AI Agents**





Beyond RCA, agents can be used to enhance a variety of industrial workflows. In fact, although the term agents may be the latest buzzword, Cognite has been deploying AI Agents across industries for the last 1.5 years with Cognite Data Fusion®. Our Industrial DataOps platform enables the deployment of next-generation industrial operations using a comprehensive data platform to deploy AI at scale.

Industrial agents are the future —virtual employees tailored to provide domain-specific insights and automate complex industrial tasks. Here are a few examples of agents Cognite has deployed or has in development as part of our latest offering, Cognite Atlas AI™:

Data Insights Agents	Use Case Focused Agents		Autonomous Operations Agents
Data Onboarding	Drilling and Wells	EPC Orchestration	Scaled Insights
Data Contextualization	Field Production Optimization	Flow Simulation Integration	"What-If"
Industry-Specific	Maintenance	Turnaround Planning	

Data Onboarding Agent

Enable comprehensive data insights by onboarding siloed data and building a knowledge graph that can be interrogated with natural language prompts.

Data Contextualization Agent

Leveraging AI and ML to automatically identify relationships between structured and unstructured industrial data and transform the data into an industrial knowledge graph to map those relationships, making complex data easily accessible and tailored to a user's needs and domain.

Industry-Specific Agent

Democratize data access for business users and subject matter experts without the need to know source systems or programming languages

Drilling and Wells Agents

A suite of agents to support D&W engineers by providing insights into well development from the daily drilling reports and selected WellView data, using natural language prompts and returns natural language response or summary report.

EPC Orchestration Agent

A suite of agents for efficient asset lifecycle management from automated asset onboarding to running specific scenarios and suggesting possible optimization

Field Production Optimization Agent

Support operators in querying production data, comparing and interpreting production indicators with simulation results, analyzing relevant KPIs to identify potential production deferments and suggesting possible causes.

Flow Simulation Integration Agent

Evaluate scenarios for generating insights on equipment and process operations, leveraging interpretations from first-principle-based simulators.

Maintenance Agent

Quickly find and interpret information about one or more work orders that meet certain criteria and provide predictive insights based on that information using understandable conversational language

Turnaround Planning Agent

Support turnaround planners in interacting with P&IDs and engineering documents to efficiently perform isolation planning and scoping turnarounds.

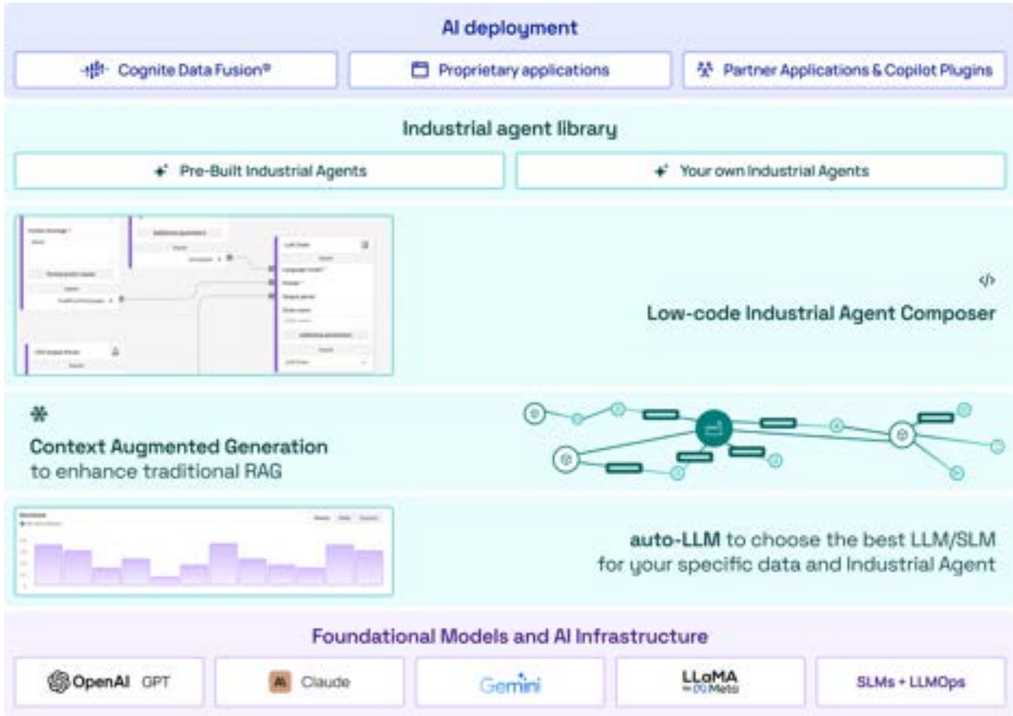
Scaled Insights Agent

Enable data discovery and consolidated insights across multiple sites to provide a more holistic view

"What-If" Agent

A suite of agents that allow you to use natural language to run simulation software and generate insights on equipment or your operations with results that are easy to understand and interpret.

Cognite Atlas AI™ is an industrial agent workbench that extends Cognite Data Fusion®. Cognite Atlas AI™ delivers everything necessary for you to build and orchestrate specialized industrial agents on top of your own data that are tailored to provide domain-specific insights and automate complex tasks.



With Cognite Atlas AI™, you can:

- Leverage a library of pre-built industrial agents or use the low-code agent builder to create your own custom industrial agents.
- Pick the best language model—be that LLM, SLM, or custom—for any given use case or industrial agent.
- Leverage your industrial knowledge graph with Context Augmented Generation for safe and dependable results.
- Get easy deployment of industrial agents into Cognite Data Fusion®, Microsoft Copilot, proprietary applications, or other 3rd-party applications.

Cognite Atlas AI™ allows you to use generative AI to carry out more complex operations with greater accuracy, including workflow automation and decision-making support, accelerating efficiencies that can generate tens of millions of dollars in business impact.

Section 3

Tools

Chapter 6
Tools for the Digital Maverick 132

6.1 Industrial AI & Data Management Software: How to Avoid Decision-Making Pitfalls When Purchasing 134
6.2 Navigating Digital Transformation: A Framework for Success 138
6.3 Navigating Digital Initiatives by Using Value as the North Star 142
6.4 Data and AI RFP Guide 148

Chapter 6

Tools for the **Digital Maverick**

Industrial AI & Data Management Software:

How to Avoid Decision-Making Pitfalls When Purchasing

Nobody wants to be in a situation where, after months of vendor meetings, internal alignment, tech reviews, and security checks for a new software purchase… your executive sponsor says, "I'm just not buying it…" But that's the reality for many complex enterprise purchases today, even with board-approved budgets opening up for AI software and an accelerating pace of procurement due to the complexity of decision-making involved.

Today's digital transformation and operations executives in chemical production, refining, and energy are under increasing pressure to deliver more productivity, reliability, and safety with fewer human resources– but more software. According to recent Gartner Maverick Research, "by 2030, 75% of operational decisions will be made within an AI-enabled application or process," demanding strategic – and rapid– realignment around technology.

However, embarking on mid-game data and technology investments can be quite complex and sensitive, especially given existing entrenched software (SAP, MAXIMO, AVEVA, etc.), siloed data management practices, and general resistance to change. Is the overarching strategy clear and communicated? Who needs to be part of the buying center? Who is personally invested in competing strategies? What impact will these decisions have on future teams and workflows?

Given Cognite's first-hand experience with a wide range of organizational buying behavior, here are two major areas that we have seen disrupt AI and data management software procurement.

1.

Fundamental Questions on Value and Strategy Remain Unanswered (or Are Answered Too Late)

Time and time again, we've seen a few common questions that, when acknowledged too late or not addressed by the right stakeholders, can either kill or delay a software purchase:

Have use cases and value been defined and mapped?

> Business value must lead the way when it comes to complex digitization projects. Rather than "eating the elephant" all at once, projects must be broken down into manageable use cases, and prioritized based on strategic near or long-term value. Companies need to combine executive-level imperatives with a true bottoms-up approach that blends the company's overall strategic direction with improving the execution ability of their frontline personnel. This action, performed with enough detail and a product roadmap-like mindset, ensures that AI-enabling software purchases are viewed by your buying center as an investment, not a cost, making it more straightforward to ensure buy-in and justify funding. More on this in our "Value as the North Star" below.

Is everyone aligned to the same software strategy?

> Often, there are competing viewpoints on do it yourself (DIY) vs software-as-a-service (SaaS) vs a hybrid approach; IT wants to build, while the business wants to buy. Compounding this misalignment is the fact that, at times, there is no clear decision-maker. This is where your digital maverick must lead and evaluate what is in the best interest of the business. Is it more important to move quickly and innovate faster than your peers? Or is it more important to build from scratch and sacrifice speed for the sake of tech stack completeness?

> It is important to note that this will be as well a good point in the conversation to align on the most suitable software deployment type for the organization (on-premise, private cloud, etc.) as it affects the financial, time, and human resource distribution, user experience, and overall software management and performance.

What's the honest state of your data and digital maturity for AI?

> We get it –Your board is pushing hard to invest in generative AI projects. However, getting buy-in, approvals, and operational value/success becomes easier when you incorporate your larger data and analytics journey. Coupled with the previous questions around use cases, evaluating your current technical strengths and weaknesses around data honestly, you can plot a more accurate roadmap with the right technical components either to build the maturity quickly, or finesse the maturity needed to deploy AI at scale.

2.

The Modern Buying Center for Data and AI Tech is Not Understood or Engaged

Modern decision-making is a team sport due to the accelerating convergence of operational technology and information technology and the business implications of any new software.

These modern buying centers can include more stakeholders than ever, such as executive sponsors, business champions, project managers, IT and security, and digital transformation leadership. Here are two ways to survive this strategic minefield:

Work as a team

> As a group, you will need to navigate key strategic questions, such as pushing enterprise-wide software decisions out to the business or using a "lighthouse strategy" to develop repeatable, scalable business capabilities built around key use cases. But all too often, there are either too few stakeholders engaged, there is an improper weighting of power in the decision, or you lack a "digital maverick" personality responsible for guiding AI and other transformation initiatives by developing a clear vision and finding creative middle ground between conflicting stakeholders.

> For example, IT teams often have the final word regarding technology decisions as they are tasked with maintaining the health of the organization's IT infrastructure and ensuring security. They are not necessarily thinking first in terms of how these types of software purchases might affect business outcomes.

Lead with business value

> Yet, IT's role is changing as this new era of technology is driving a competitive edge with flexibility and time to value and is shifting toward business and user needs. In the above-mentioned Gartner Maverick research, they make a disruptive recommendation to 'stop investing in IT skills,' given the rapid democratization of generative AI and other quickly developing technologies. While this contradicts previous generations of Gartner recommendations, it accurately reflects today's reality to lead with business value.

> Acting on this recommendation requires rebalancing the decision-making weights in your buying center to make it more attractive for IT to support a software purchase if business need and value clearly outweigh potential risks.

Software purchases around industrial AI are accelerating as boards and executives approve new budgets to capture the potential and competitive advantage. Do you have the right team in place, and have you answered the key questions that will drive a quick and successful outcome and put you on track to deliver a return on your investments in data and AI? The RFP guide at the end of this book can help.

Navigating Digital Transformation:

A Framework for Success

Digital transformation is imperative for sustained growth and competitiveness in industries like Energy and Manufacturing. However, if industrial organizations want to achieve increased efficiency, reduced costs, and enhanced decision-making capabilities promised by full-scale digital transformation, they must think critically about their goals and maturity in digital transformation initiatives and carefully assess what it would take to implement new technologies in their ecosystem.

Organizations can rely on Cognite's Customer Success Framework to navigate these complexities successfully. This framework is broken down into two parts:

Part One: the 5Ps

Purpose: Create excitement and get buy-in

Digital transformation initiatives often fail due to a lack of clear purpose and insufficient stakeholder buy-in. To counteract this risk, begin by articulating a compelling vision that resonates with employees at all levels. Then, identify your key champions, especially those looking to solve their data problems faster and more efficiently, and clearly explain the benefits of the transformation for them.

Whether it is creating digital twins, automating highly manual processes, maximizing your production, or minimizing turnaround time, highlighting user-centric success stories from similar industries (and, eventually, showcasing your own success as the program rolls out) can also serve as powerful catalysts for generating excitement and fostering a sense of purpose among teams.

> **Best practice:** Regularly communicating progress, milestones, and success stories through awareness campaigns such as newsletters, use case exploration sessions, and hands-on training to reinforce the purpose and maintain enthusiasm

People: Build your teams incrementally

Any digital transformation initiative's success relies heavily on the people involved. Cognite's Professional Services teams and certified partner networks provide the subject matter and domain expertise you'll need to quickly enable a Center of Excellence that can advise and augment your teams as you grow. Building cross-functional teams with diverse skills and perspectives is crucial, and incremental team building allows for a smoother integration of new talent and expertise. Additionally, Cognite's Academy, training, and active user community provide the latest resources to upskill and educate your organization continuously throughout the digital transformation journey.

> **Best practice:** Build a Center of Excellence and prioritize ongoing training and professional development to continuously grow your talent pool and get the most out of your digital team.

Portfolio: Be deliberate about the value you want to create

Many organizations struggle to clearly identify use cases and understand the value that can be delivered from their digital transformation. Successful organizations refrain from pursuing disparate projects simultaneously and, instead, prioritize initiatives that align with strategic objectives and deliver tangible value incrementally.

Cognite has proven paths to growing use cases and value incrementally: starting with the data foundation and unlocking generative AI-enhanced data exploration, moving that success into the field, implementing connected worker tools, and then scaling toward a more comprehensive remote operations control tower. Meticulously curating your digital portfolio, incentivizing usage, and developing a clear adoption strategy will maximize impact and ROI.

> **Best practice:** Regularly reassess the portfolio to ensure alignment with organizational goals and market dynamics, making adjustments as necessary. Explore predefined packages with Cognite where possible to demonstrate value quickly and effectively and work with your account team to ensure you have an Adoption Strategy aligned to the needs of your organization.

Process: Create a collaborative ecosystem

Collaboration is the bedrock of successful digital transformation. Breaking down silos and fostering open communication across departments is paramount. Cognite can help create a governance framework that highlights clear ownership across your organization and is aligned with all your partners. Begin with individual project execution governance and establish program governance, including change management, risk management, data management, and roadmap alignment. Ensure executive oversight through steering committees and/or quarterly executive reviews with key sponsors.

> **Best practice:** Establish clear communication channels, feedback loops, and KPIs for structure and continuous improvement to sustain a collaborative culture.

Platform: Know what powers the value you drive

As the technological backbone of digital transformation, the data platform plays a critical role. Understanding the capabilities of the chosen platform is essential for unlocking the full potential of digital initiatives. Only Cognite offers deterministic, hallucination-free, data-leakage-free, real-time, data-inclusive Generative AI solutions that cover OT, IT, engineering and unstructured industrial data. We do all this through industrial-data-specific data contextualization pipeline services that create and maintain a high-performance Industrial Knowledge Graph with flexible data modeling. And only Cognite is built API-first, allowing you to connect to any other system and develop solutions through Cognite's collaborative workspace (Industrial Canvas) or any other partner you choose.

> **Best practice:** Look for an open, extensible, and secure solution that offers robust industrial contextualization capabilities (regardless of source and type) and a prebuilt, comprehensive AI architecture.

Part Two: Strategic Evolution

Your digital transformation journey is an evolution, with a different strategy for success at each stage. Cognite calls these stages Start, Structure, and Scale, which translates to building the right foundation, developing repeatability, and then growing across assets, units, sites, and regions. Each stage is accompanied by a number of milestones we believe are necessary to track your progress.

This perspective helps you focus on key objectives throughout your digital transformation to ensure you build and grow a highly collaborative cross-functional organization that delivers value. Cognite's Customer Success team will support and help you identify where you are in your digital transformation journey and chart a plan to accomplish your end goals.

Putting the Framework into Practice

When considering the 5Ps across the Start, Structure, and Scale stages, Cognite delivers a straightforward framework that provides clear milestones throughout a digitalization journey. Below are a few examples.

Successfully navigating the complex digital transformation journey demands a holistic and strategic approach. Cognite's Customer Success Framework allows you to think critically about how you grow your vision, core skills, systems, and processes so that your organization can overcome common challenges and thrive in the data and AI age.

		First Success and Foundation	Repeatability and Predictability	Productivity and Efficiency
		• High impact use case • Create buy in - champions • Get data under control • Understand the technology	• Expand and/or replicate use case • Formalize data model (s) • Drive governance and quality • Develop Center of Excellence (CoE) • Automate processes - SOPs • Manage Change - MOCs	• Expand the CoE ecosystem • Breadth and depth of use cases • Formal adoption campaigns • Value roadmap • Secure sustainable funding
		Start →	Structure →	Scale
Purpose Create excitement and buy-in	Examples of Key Activities We Track	User centric **vision** Build your **champion**	Use wins to **evangelize** Grow network of leaders	**Champions** throughout org Create **self sustaining funding**
People Build your teams incrementally		Cognite builds – **You shadow** Activate virtual **training** programs	**We build together** Activate **community** and bootcamps	**You build** – We support Expand with **partners**
Portfolio Be deliberate about the value you want to create		**High impact** use case Grass roots **1:1 user engagement**	Use case expansion **Super users trainer program**	Defined **value roadmap** **Coordinated adoption** campaigns
Process Create a collaborative ecosystem		**Project execution** governance Communication channels	**Executive and Program** governance **SOPs, MOCs,** Documentation	Formal **KPI tracking** **Multi-tiered support program**
Platform Know what powers the value you drive		**Initial data model** Access, security, infrastructure	Formalize data models **Scalable** architecture foundation	Technical roadmap alignment **Blueprint** for scale

Navigating Digital Initiatives:

Using Value as the North Star

> "Only what gets measured gets managed." Peter Drucker

Digital initiatives have become pivotal to remaining competitive, but companies struggle to measure the potential impact of their efforts because of a lack of consistent frameworks and few benchmarks that provide an accurate apples-to-apples comparison.

For CEOs and COOs, ensuring digital initiatives contribute to shareholder value remains paramount. Having value as a North Star ensures that the stakeholders of a digital initiative are aligned and laser-focused on the highest impact and highest value business opportunities, which can create tremendous clarity and excitement throughout the journey.

However, for many companies, making a direct link between deploying a new digital program and the shareholder value that it is intended to generate can be a struggle. So, after countless hours spent with customers, here's the blueprint for success that helped Cognite uncover more than $1bn in customer value during 2023, by making it easier for executives to reach informed decisions and deliver on their initiatives.

Top Three Value Drivers

$432m
Operations Productivity

$370m
Asset Reliability

$50m
Process Optimization

Industries Value Breakdown

53%
Energy

26%
Manufacturing

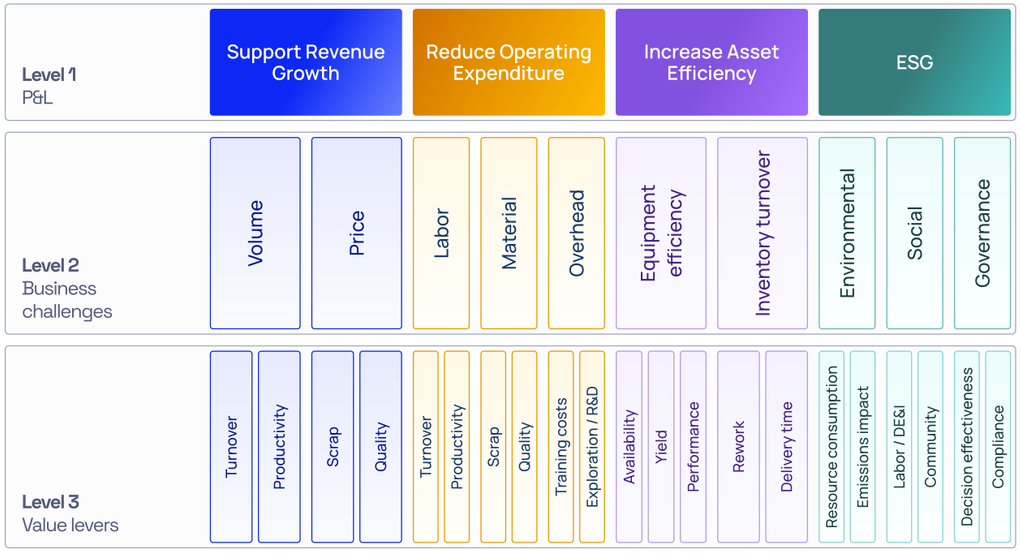
21%
Power & Utilities

Identifying Value

- The majority of companies already have a long—very long—list of potential use cases they could deploy, but rarely has it been prioritized for feasibility and business impact.
- Companies need to combine executive-level imperatives with a true bottoms-up approach that combines the company's overall strategic direction with improving the execution ability of their frontline personnel.
- At Cognite, we often see that allowing operators to tell stories about their day-to-day challenges enables them to start crafting solutions to the larger problems the company faces, and that can generate significant business impact. Using tools such as the value map (below), Cognite works with customers utilizing the value map to help articulate the underlying value drivers, tying them to shareholder value creation.

> **Customer example:** Cognite worked with the frontline operators of one of the world's largest LNG terminals to help them craft a prioritized roadmap centered around replacing a Sharepoint-based way of working with direct access to data in the field to enable more effective operations and maintenance.

Example value framework specific to manufacturing. Cognite has developed domain-specific ones across upstream and downstream, chemicals, grid, power generation, etc.



Measuring Value

- Pinpointing value is an art form, not a science, and can feel scary at times. It requires trust and a bit of a leap of faith but, by gathering a small group from operations, finance, and strategy, companies can articulate the key metrics that they are trying to improve in their business; these typically sound like "an X% increase in production, hours saved in a turnaround, or productivity for subject matter experts measured as the number of hours saved."

- With the help of benchmarks, companies can then begin to assign potential improvement potential to each initiative and ultimately arrive at an EBITDA improvement number.

- Most of the time, operators have a really solid understanding of how much they can realistically move the needle. More importantly, just doing the exercise is important because, even if you are off by a magnitude, companies can still get a feel for whether they are looking at pennies or dollars.

Customer example: Cognite worked with a large Canadian chemicals company to build a bottoms-up business plan by estimating the total EBITDA impact based on the estimate for each solution deployed. Taken against the total investment cost (Cognite Data Fusion license, implementation costs, and internal spend), they arrived at an estimated ROI of the digital initiative, which they could compare against other investment programs at the company.

Net Value

- **EBITDA increase from deployed solutions**
- Other cash flow effects (saved licenses, etc.)
- Increase in NPV from increased **Time to Value**

=

Cost of Investment

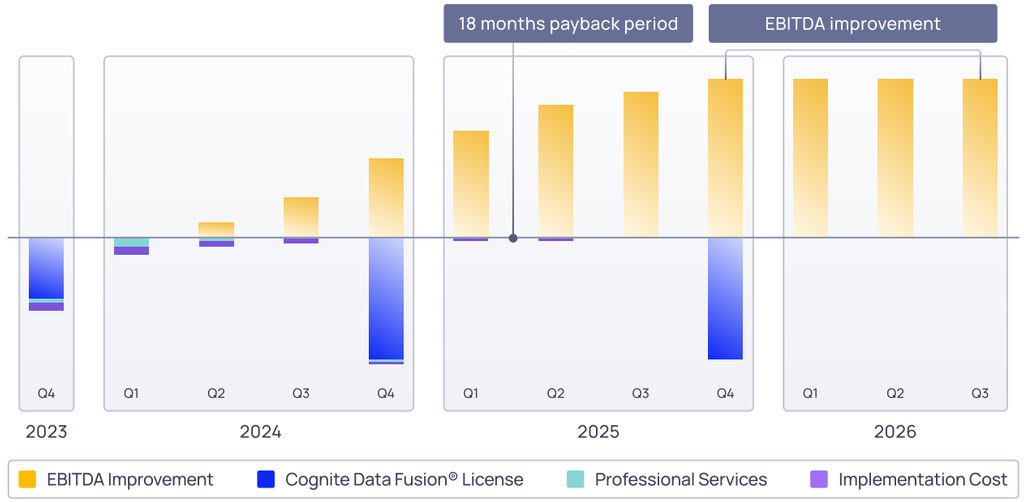
- Cost of Cognite Data Fusion® **license**
- **Professional Services** / Partner Costs
- **Internal** project implementation costs

Return on Investment

- Net Value divided by Cost of Investment
- Expressed as ROI: 0.0x

Benchmarks

Best	+10x
Good	3-10x
Reconsider	<3x



Communicating Value

- Cognite has combined a typical management consultant toolbox with our domain expertise and operational know-how to ensure that the value companies identify for their digital initiatives also gets communicated in a way that executive stakeholders are familiar with.

- Often, this can be as simple as a clear graph with a business plan breakdown in an easy-to-understand format emphasizing the EBITDA impact and ROI of the proposed digital initiatives.

- Providing clear and well-articulated ROI estimates helps the executive management team prioritize strategic initiatives across their business, whether that is investing in new machinery, hiring for a new function, or deploying digital solutions.

- By speaking the Board's language and emphasizing not only the transformative impact on operations but also the attractive returns on investment to shareholders, we have found that the likelihood of getting to yes increases dramatically.

Customer example: A forward-thinking Chief Digital Officer at a large energy company struggled to convince her CEO that the company needed to invest considerably more resources into their digital program. By putting together a succinct presentation with a clear investment ask and the expected ROI front and center, she ultimately gathered the necessary support.

Tracking Realized Value

- CContinuous monitoring and evaluation are crucial for ensuring that digital initiatives are delivered with the intended results. By tracking key performance indicators and real-time data analytics, companies can course-correct as needed and make sure they maximize value realization.

- Tracking realized value, however, is not an easy exercise and must consist of a triangulation between closely monitoring usage (measured as, e.g., monthly active users) along with interviewing frontline employees to pick up stories from the field, and surveys to hear how a broader group of employees see their day to day changed.

- While value might be hard to track directly, if you see usage increase over time, it is likely because your people are using Cognite Data Fusion® and getting value from that usage. There is a correlation between usage and value, one that companies can track in real time by looking at usage numbers.

Customer example: A US-based continuous manufacturing company struggled to reap the benefits of their digital initiative and decided to focus on increasing the amount of Monthly Active Users. As a result, over time, the company started to see concrete situations in their operations in which easy access to operational data in the field had saved them from costly incidents in the magnitude of 10s of millions of dollars.

Thus, while value can seem arbitrary, and potentially an afterthought when implementing a large-scale digital transformation program, there are three steps to ensuring a successful program:

- **First**, companies can more easily navigate the implementation of their digital programs by adopting value as the North Star to drive meaningful business impact.

- **Second**, we recommend that companies assign a person responsible for driving digital initiatives that span both the digital and business domains.

- **Third**, companies should team up with a trusted advisor who is able not only to provide the right solutions, but also guide the company during the process, from helping them articulate value to ensuring it gets realized.

Data and AI **RFP Guide**

Use Cases and Past Successes

This section should be first because Industrial DataOps must be able to deliver long-term value to your organization and you need to ensure alignment between your organizational goals and the potential solution provider. Knowing that your solution provider is competent within your domain will de-risk the probability of under-delivering on your expected ROI.

- Can you provide a brief description of your company, its industrial business areas, main products/services, relevant expertise, and business strategy?
- Are your products/services general or specific to the client's industry? Can you describe your domain expertise?
- How would you describe your key product differentiation?
- What is your experience helping clients build business cases and developing a target ROI? Can you provide examples of successful business cases delivered?

> **Expert tip:** Successful Industrial DataOps solutions should start with one to two use cases defined before any work begins and have a backlog of two to five use cases once success is achieved with initial use cases.

- How many existing customers do you have? Are there past successes you can share related to the client's industry?
- Does the proposed solution enable more effective asset management? Can you provide examples?
- What use cases have you delivered regarding unstructured data (video, 3D, etc.)?
- What are the most common types of use cases you have delivered?
- Do you have reference customers that might be available to talk with us?
- Can you provide a product demo?

Functionality

Properly assessing Industrial DataOps software requires understanding two components: foundation and connectivity. Assessing the foundation is critical to ensure that the proposed solution will support your industrial data use cases and provide the tools needed to minimize time to value, scalability, and repeatability. Connectivity has two components—data extraction and application layer. Data extraction will ensure that you can connect to both existing and future data sources, while the application layer focuses on how the solution provider will support applications on top of the foundation to deliver use cases.

Foundation

- How does the proposed solution perform data contextualization (data mapping)? Is it automatic or semi-automatic? Does the solution suggest relationships to make identification and construction easy?

> **Expert tip:** The ideal solution should automate this process as much as possible, or expanding the system to include new data sources will be extremely time consuming and hard to manage.

- How is the contextualization (data mapping) process managed? Is it easily accessible? How do users make edits?
- How is the data model created in the proposed solution? How are relationships between data sources managed?
- What types of data formats are supported in the proposed solution?
- How does the proposed solution support data visualization?
- How does the proposed solution manage data quality? Are rules pre-built? Can rules be modified? Are rules applied universally or per use case?

> **Expert tip:** Data models are designed to be reused. Data quality should have the flexibility to be applied per use case. For example, different use cases may require the same data, but using this data for remote monitoring of an asset will not require the same update rate as using this data to run an analytics model measuring performance.

- Does the proposed solution support templatization? How can applied work be reused?

> **Expert tip:** Templatization is a key component to scale solutions and ensures your organization will avoid getting trapped in Proof-of-Concept purgatory.

- How are notifications/messages supported in the proposed solution with regards to users associated with data, administrators, etc.?
- How would you describe the proposed solution's performance in regards to scalability?

> **Expert tip:** As you expand beyond initial use cases, you will want a solution that is able to scale. Industrial DataOps should be able to address scale at both the site and enterprise level.

- How does the proposed solution process large data sets? How do you ensure the proposed solution can handle peak processing?
- How does the proposed solution support trending analysis of the data? How are trends visualized and reported?

- Can the proposed solution analyze trends in data quality and predict when metrics will exceed predefined thresholds?
- How does the proposed solution document completeness (integrity) of the ingested data and ensure data is not lost in transit?
- How do you work with third-party vendors? Which have you worked with in the past?

> **Expert tip:** While many solutions talk about openness, seeing examples of proven solutions with third-party vendors will provide confidence that you will be able to connect your disparate data sources.

- Is the front-end framework of the proposed solution built on open standards? How do you support open front-end frameworks?
- How does the proposed solution ensure that data is processed quickly and readily makes available time series data?

> **Expert tip:** Access to centralized, remote relevant-time data creates opportunities for many new use cases at both the site and enterprise level.

- Does the proposed solution require plugins such as Office, Flash, etc.?
- Is the proposed solution able to ingest both tabular and graph-structured data without loss of information?
- When receiving asynchronous time series data, how does the proposed solution handle timestamping?
- Is the proposed solution able to handle data inserts, updates, and deletes by itself?
- Does the proposed solution support multiple modes of operation, such as batch and stream-based ingestion and in-memory versus persistent data storage?
- Does the proposed solution follow agile development principles? How do you ensure it is up to up-to-date on market trends and technical standards?
- How does the proposed solution support compression of data and metadata?
- Does the proposed solution report the source for a data point, event and time series, and associated metadata for users to assess the data quality?
- How are the metadata fields of existing data and metadata updated? How are updates executed and managed?
- How is the connection between data and metadata made? Are they stored or linked? Can metadata be linked to several data entries?

Connectivity

- How does the proposed solution support integration with external systems and what are the requirements of such integrations?
- What integrations are pre-built and readily available for data extraction? For the application layer?

> **Expert tip:** Pre-built data extractors should exist for many open protocols and advanced Industrial DataOps solutions will have existing extractors to individual industrial solution providers such as Siemens, ABB, and Emerson.

- What are the client's potential ways of developing their own applications on top of the product?

> **Expert tip:** Further assessment is needed when thinking about application development for data engineers and citizen data scientists. Proposed solution providers should have pre-built connections to well-adopted applications, such as PowerBI or Grafana.

- Does the proposed solution provide an associated SDK? What languages are supported?
- What types of underlying data sources do you support? What connections are most common?
- What is the proposed solution's capability to access real-time data? What are the scalability limitations to this capability?
- Does the proposed solution have connectivity and native access to relational databases?
- Does the proposed solution have connectivity and native access to non-relational structures?
- How do you ensure interfaces for data exchange (such as REST APIs) are kept stable and robust to changes?
- Does the proposed solution support versioning for continuity so that the newest version, and the previous version of data pipelines remain supported? Can versions be rolled back?
- Does the proposed solution support a layered and scalable REST API?
- Is the REST API stateless, enabling easy caching and no need for server-side state synchronization logic?
- Can underlying data be exported from the proposed solution as a CSV or XLSX? How does the proposed solution export data and metadata in standardized formats?
- Are there any limitations in the proposed solution's ability to extract historical data?

Generative AI

- How have you applied machine learning (ML) solutions to solve client use cases? Are there any use cases that utilize a hybrid AI (combination of physics and ML capabilities) solutions that you are able to share?
- How is generative AI incorporated into the proposed solution? Are AI capabilities built into the backend of the product? Is there a natural language copilot as part of the user interface?
- Can you provide details about the training data used to develop the AI model? What is your process for providing industrial context to generative AI solutions?
- How do you mitigate hallucinations within your generative AI solutions?
- How do you manage data leakage within your generative AI solutions?

- How do you manage trust and access control within your generative AI solutions?
- Is your generative AI solution compliant with relevant data protection and privacy regulations (e.g., GDPR, CCPA)?
- How frequently is the AI model updated and retrained to maintain its security and reliability?
- How does your solution handle intellectual property rights and content ownership?
- How have you applied generative AI solutions to solve client use cases? Are there any use cases that you are able to share?
- Are there any limitations or potential risks associated with using the generative AI solution?
- Can you share your generative AI roadmap and what are the most important near-term deliverables on this roadmap?

Solution Architecture

Every organization will have unique architecture requirements that should be addressed from the beginning. The key here is to ensure that the proposed solution provider is designed to meet the requirements of your existing environment.

- Can you describe the key components of the proposed solution and how they operate/interconnect?

> **Expert tip:** Any architectural requirements can be included here. Many organizations have already made investments to integrate OT/IT data silos into data lake or data warehouse solutions. Your Industrial DataOps solution should leverage the investment into the existing infrastructure.

- Is your software cloud native? Which vendors (AWS, Azure, GCP) do you support?
- Do you support hosted/private cloud or on-premise deployment?
- What is the proposed solution's ability to support real-time deployment?
- How does the proposed solution support horizontal and vertical scaling?
- How does the proposed solution offer high availability and how are failover procedures handled?
- How does the proposed solution support backup and recovery procedures?

- How does the proposed solution handle archiving?
- How do you support edge capabilities? Do you offer on-premise deployments?
- Is the proposed solution validated with the standards of W3C and HTML5 to enable browser independence?
- Does your proposed solution track the lineage of all data objects and code, showing upstream sources and downstream consumption?
- How does development occur in the proposed solution, introducing changes to its core components, adding extensions etc.?
- How is it possible to test reconfigurations, upgrades, and extensions to the proposed solution before it is put into production?
- What are the software and hardware prerequisites?

Project Execution, Services, and Support

Understanding how potential solution providers implement projects will allow you to assess time to value and a high-level roadmap for implementation. The potential solution provider should provide the resources to ensure continued success. Successful implementations require both the right technology and the right support. This section is designed to provide insights to the expected support for your team and organization when adopting an Industrial DataOps solution.

- Can you describe the 'Go Live' period between proposed solution validation/operational deployment, and final acceptance/beginning of any maintenance and support agreements?
- What maintenance and support do you offer during and after implementation?

> **Expert tip:** Proposed solution provider should have a designated customer support representative to ensure project success.

- What does a typical project implementation process look like? What support is available?
- What level of services do you typically provide?
- Please describe how your skilled experts will interact with clients' in-house experts to maximize the benefit from collaboration.
- How do you enable/support search in the proposed solution? Can you provide documentation?

> **Expert tip:** This functionality will be very valuable to save time for data engineers and make data discoverable for citizen data scientists.

- How does the proposed solution support documentation and how is it made accessible?
- What training programs are included and offered? What is typical?
- How do you ensure that competence is built within your client's organization?

> **Expert tip:** Building competence within your organization is an important trait to maturing digitally. Your solution provider should be enabling these competencies. Otherwise your organization runs the risk of being in a service-based relationship with the solution provider.

- What resources and support are provided during this period?
- What standard support do you provide in problem resolution? Do you offer varied support levels?

Security

With the importance of security always increasing, the potential solution provider must be ready to meet the needs of your organization. This is not intended to be a comprehensive security list as IT departments often have developed their own security requirements for new software products and is necessary to consider when integrating IT and OT data.

- What is your company's strategy for penetration testing and third-party assessments?
- How does the proposed solution maintain an audit trail of all data manipulation?
- How does the proposed solution offer monitoring and statistics of backbone components?
- How do you ensure that the client has access to its own data in the proposed solution?
- How is high availability maintained for security, access, and governance of the proposed solution?
- How do you support revocation of access at both user and group level?
- When and how is data encrypted in the proposed solution?
- What is the proposed solution's capability with regard to access control? What is the granularity?
- Does the proposed solution support groups for access control?
- Can authentication requirements be customized in the proposed solution?
- How does a user report suspicious activity related to data points?
- Can users be assigned special roles to fix or disapprove reported suspicious data points?
- Does the proposed solution support ISO, SOC 2 Type 2, NIST CSF, IEC, NERC CIP, or other relevant industry standards?
- How does the proposed solution track the chain of custody?

Usability

For products to be adopted by your organization, solutions must be easy to adopt by the users. Poor usability is a leading cause of poor product adoption. The potential solution provider needs to support both data scientists and citizen data scientists to truly make data usable. In order to make data discoverable and usable for both of these users, the proposed solution must deliver simple access to data and provide intuitive, well-designed user interfaces that do not require strong coding backgrounds to leverage.

- Are users able to navigate through different parts of the proposed solution without help?

> **Expert tip:** Asking for a product demo is helpful when trying to assess this topic.

- Are users able to create and edit their own dashboards to solve specific business needs, and are they able to share these dashboards to collaborate with their team or across teams?
- Do users see and feel the proposed solution responding in real time?
- How many concurrent users does the proposed solution support? Is the environment collaborative?

> **Expert tip:** As your Industrial DataOps solution gains adoption, your organization should be striving to increase user adoption to enable use case development throughout many departments.

- How does the proposed solution handle error messaging? Are errors easily interpreted by users?
- How does the proposed solution allow users to refine search results?
- Can users create data pipelines without IT assistance and without deep training in data engineering, SQL, or production processes? Do you provide a graphical user interface for pipeline creation?
- Can users execute other tasks during the execution of jobs? Are users alerted when jobs are complete?
- How do you ensure fast search results are returned to users?
- How do users report errors, bugs, lack of service, and requests for new services or extensions to existing services?

Software Maintenance

Once the solution has been implemented, this section is designed to give you an understanding of the required upkeep. Reliability is another important factor in product adoption. Improvements /enhancements to the proposed solution should not result in unexpected downtime, and the solution should not require a high level of manual support to ensure proper operation.

- How often do you release improvements to your products? Do you have major and minor release cycles?

> **Expert tip:** As your organization requires, be sure to understand the different management requirements between on-premise, private cloud, and public cloud offerings.

- Are clients entitled to all product upgrades with the base software? When are upgrades required?
- How are clients notified about both scheduled and unscheduled maintenance/downtime?
- How are new versions/updates managed?
- Do you guarantee availability and uptime of the proposed solution to be 99.5%? How do you track system uptime?

Sustainability

- Can you provide an overview of your company's sustainability mission and strategy?
- How does sustainability align with your company's core values and business objectives?
- What specific sustainability goals has your company set, and how do you measure progress towards these goals?
- In what areas does your product have a major environmental impact?

Future Development

Ensure that the potential solution provider's roadmap is aligned with your organization's goals. Seeing the top priorities of technology development will provide clarity to the product direction and the ways in which your organization will be able to grow with the Industrial DataOps software.

- Can you provide a short-term (six to 12 months) and long-term (two to five years) product roadmap?
- What is your approach to developing new products and the possibilities for developing customizations/extensions?

Pricing Model

To date, pricing has not seen convergence across the industrial software industry. Asking the high-level questions to understand the initial price (including services) required to get started will be valuable when assessing potential solution providers. In addition, Industrial DataOps solutions are designed to scale so it is also important to understand the levers of pricing when data sources, users, and use cases start to increase.

- How do you price the product? How does your pricing model support increasing use case and product adoption?
- What factors do you predict will be the main cost drivers for your product and services?

Our **Definitive Guides**

The Definitive Guide
to **Industrial DataOps** →

An essential guide to rolling out Industrial DataOps; the critical first step to implementing AI for industry and embarking on your digital transformation journey.

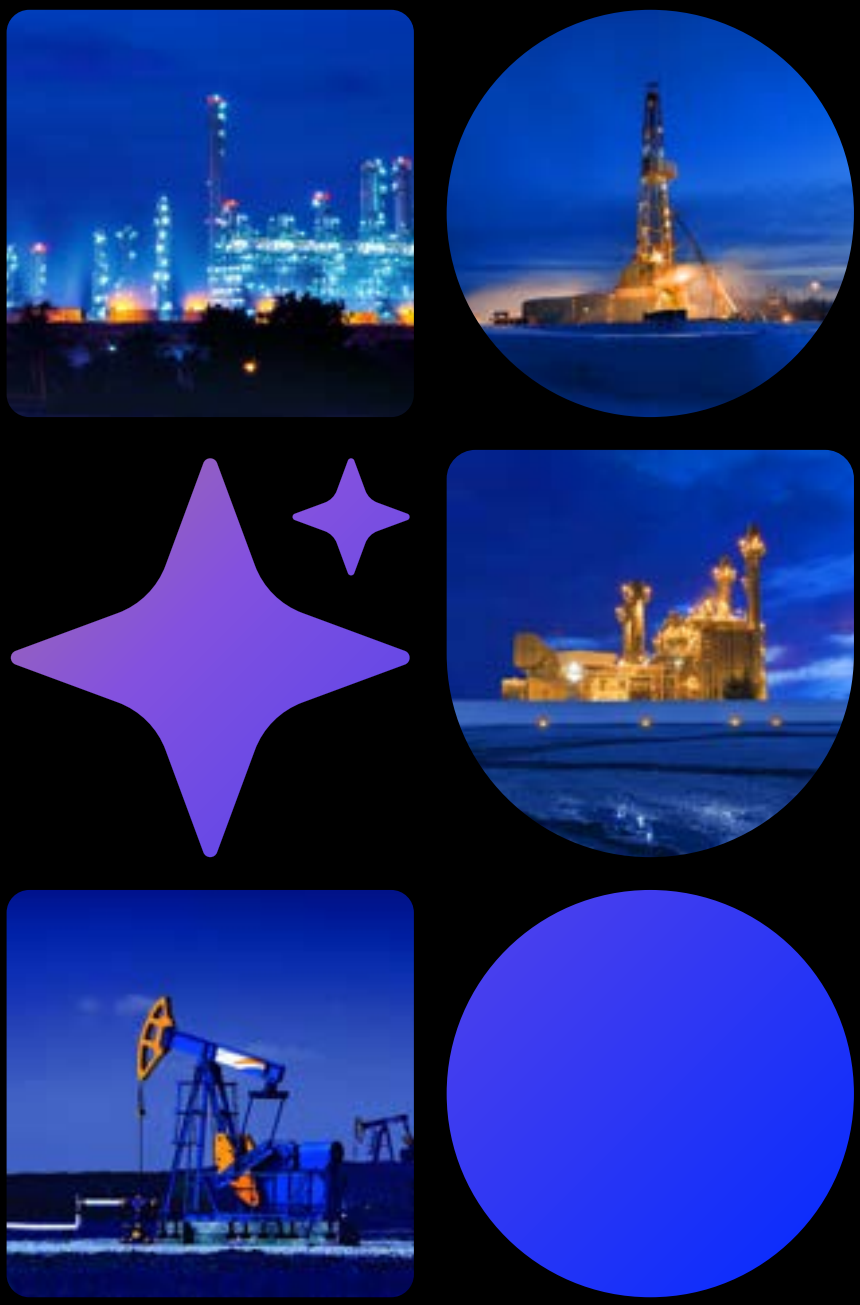
The Definitive Guide
to **Generative AI
for Industry** →

A comprehensive manual to accelerate AI innovation and reduce time to value of digital transformation programs.

COGNITE
AI FOR INDUSTRY

©Copyright, Cognite, 2024 – **www.cognite.ai** →

COGNITE
AI FOR INDUSTRY



Learn more at **cognite.ai** →