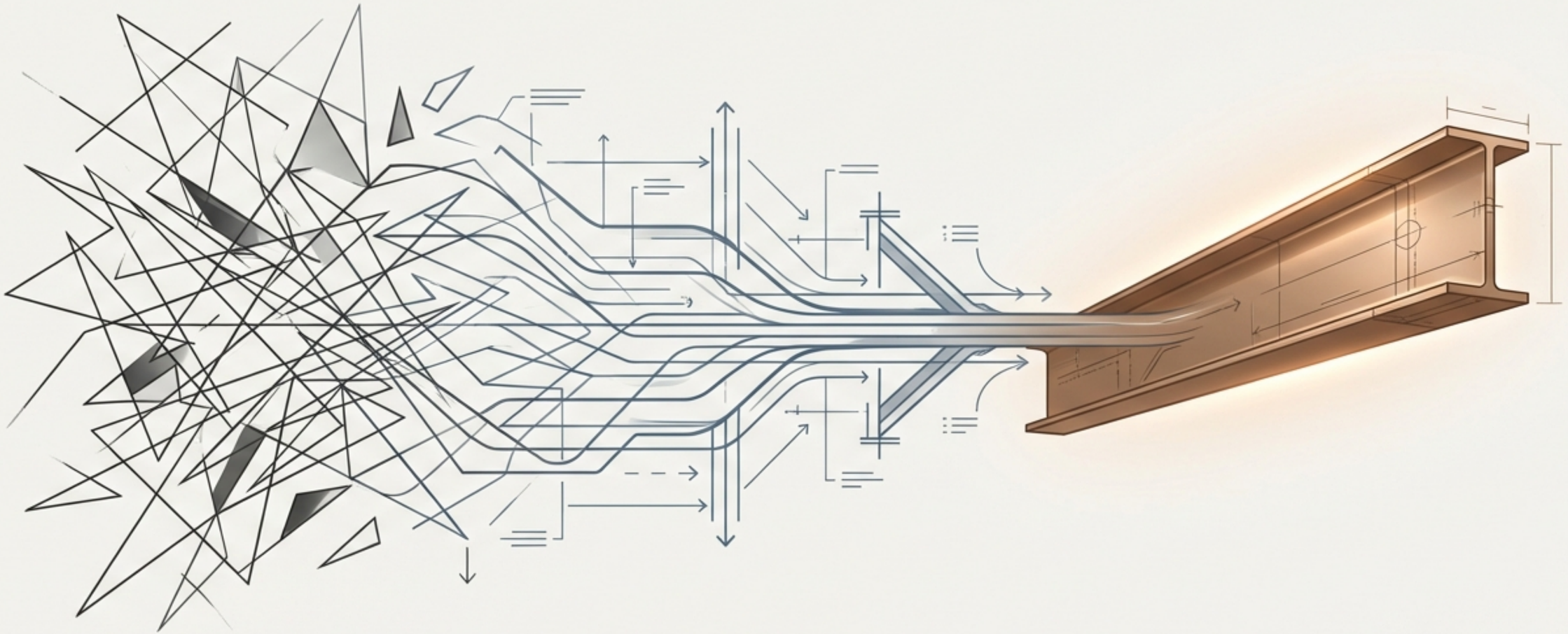
Foundry IQ: The Future of Enterprise RAG

From Fragmented Data to Intelligent Retrieval



The AI Expectation Gap: User Queries Have Evolved

Yesterday's Query

"Find maintenance procedure CTL-11-R2 for the XJ-400 compressor."

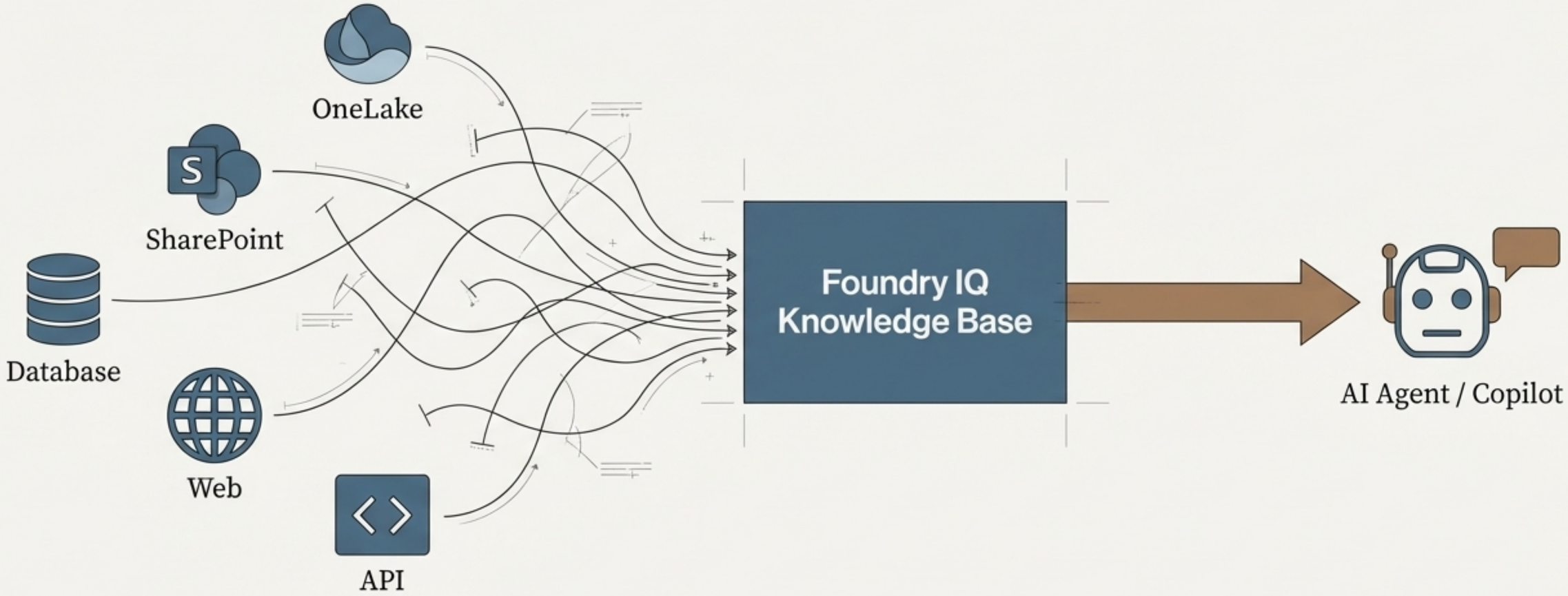
Today's Reality

"I'm looking at a Gen-4 compressor with the label 'CTL-11-R2'. The status light is red. I think it's a power issue, the cord is labeled 'UL 817'. I don't know what that means. Should I replace the unit? What's the policy?"

- "CTL-11-R2" — *Specific identifier*
- "status light is red" — *Observed state*
- "UL 817" — *External knowledge required*
- "replace the unit? What's the policy?" — *Seeks procedural guidance*

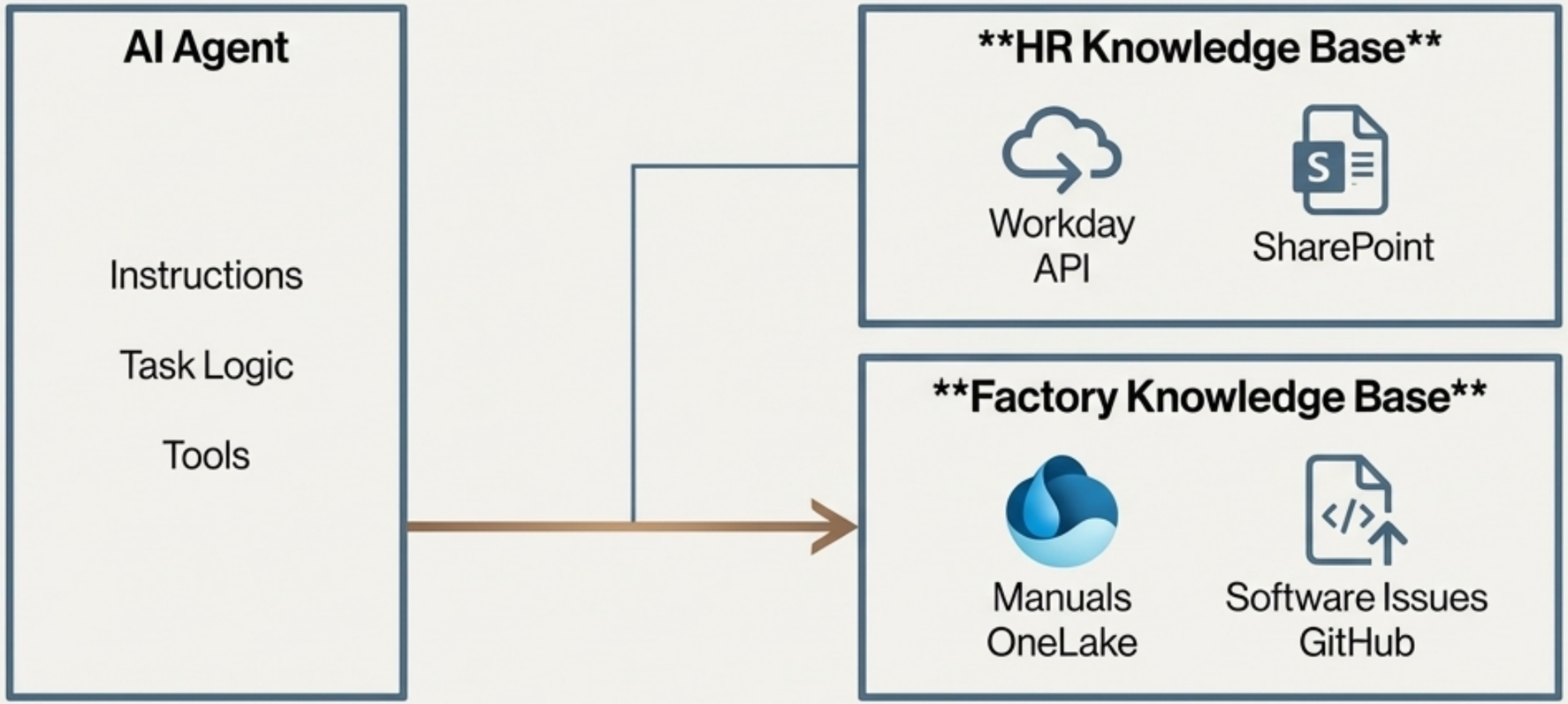
Today's users don't craft perfect queries; they have complex, multi-part problems. Standard RAG systems break because the required data is fragmented and the user's intent is nuanced.

Foundry IQ Unifies Your Entire Knowledge Landscape



Foundry IQ connects agents to your entire scope of knowledge, abstracting away the complexity of how, when, and where to retrieve data.

The Core Abstraction: A Knowledge Base for Each Domain



A Knowledge Base is a semantic container for a domain. Instead of burdening the agent with data source details, you create focused agents and focused knowledge bases. This makes both more reusable and maintainable.

Knowledge Sources Connect to Any Any Data, Static or Live

Indexed Sources

For Data You Control

Content is ingested, chunked, and vectorized into a managed index for maximum performance and relevance.

Best For: Static or semi-static content like PDFs, Markdown files, or document repositories.

Blob Storage

OneLake

SharePoint
(can be indexed or remote)

Remote Sources

For Data You Don’t

Queries are federated in real-time to external systems, ensuring information is always current.

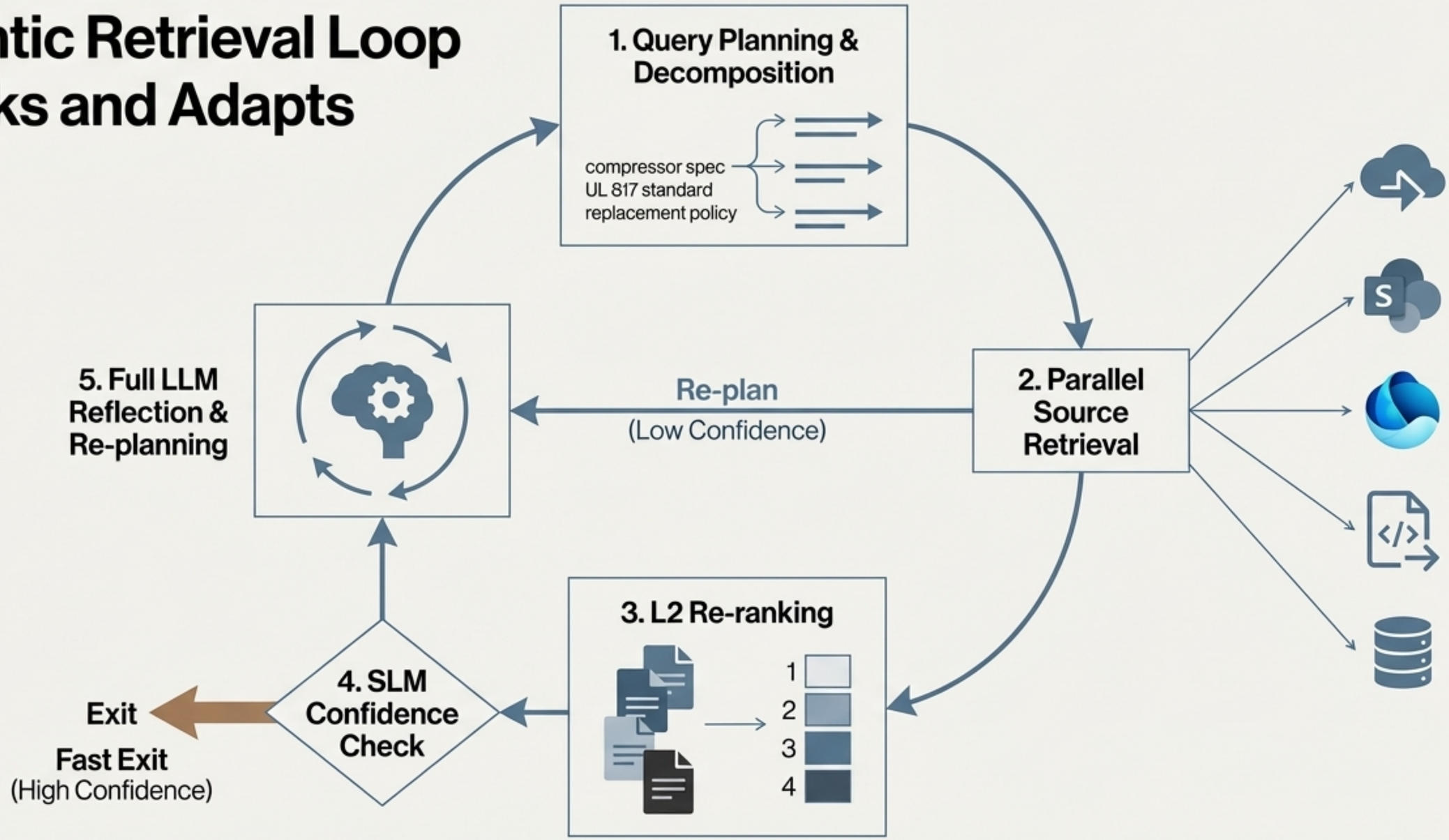
Best For: Live applications, public web grounding, or systems with their own search endpoints.

Web Search

GitHub

API

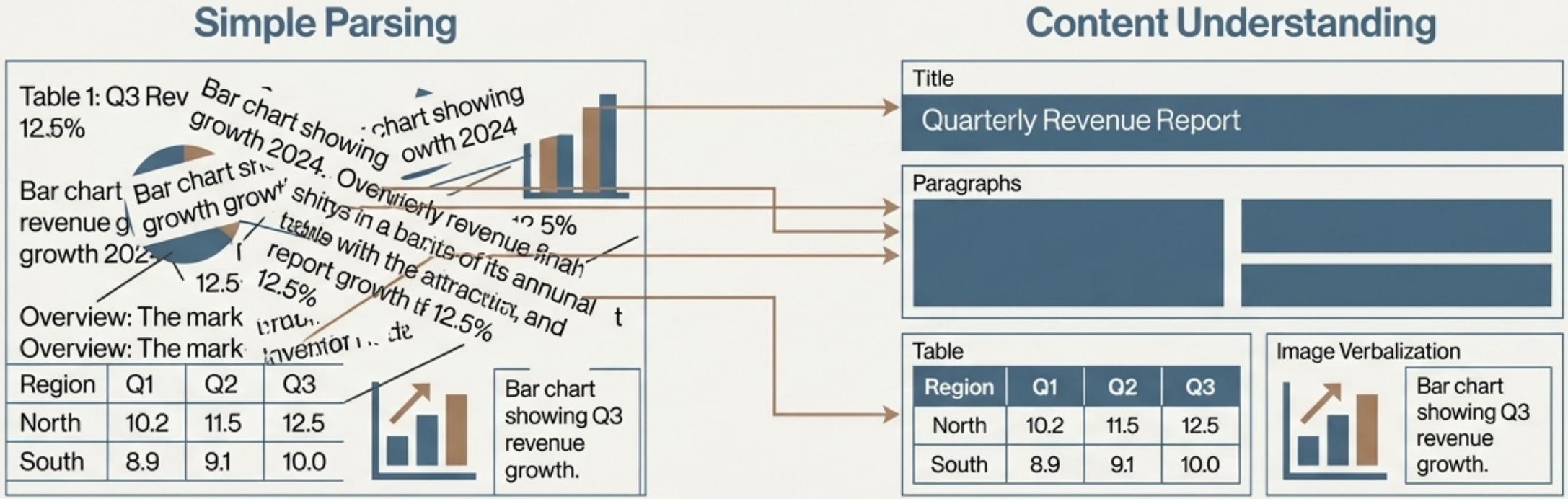
Beyond Search: The Agentic Retrieval Loop Thinks and Adapts



Dialing In Performance: Tune for Latency vs. Quality

Mode	Goal	Key Characteristics	Best For
Minimal	Sub-second Latency	No planning, no iterations. Behaves like a standard search query.	Fast lookups, simple chatbots.
Low	Balanced	Enables the full agentic loop but is tuned to exit early if possible.	General purpose, complex Q&A.
Medium	Maximum Quality	The system is encouraged to take more time and perform more iterations to find the best possible answer.	High-stakes research, deep analysis.

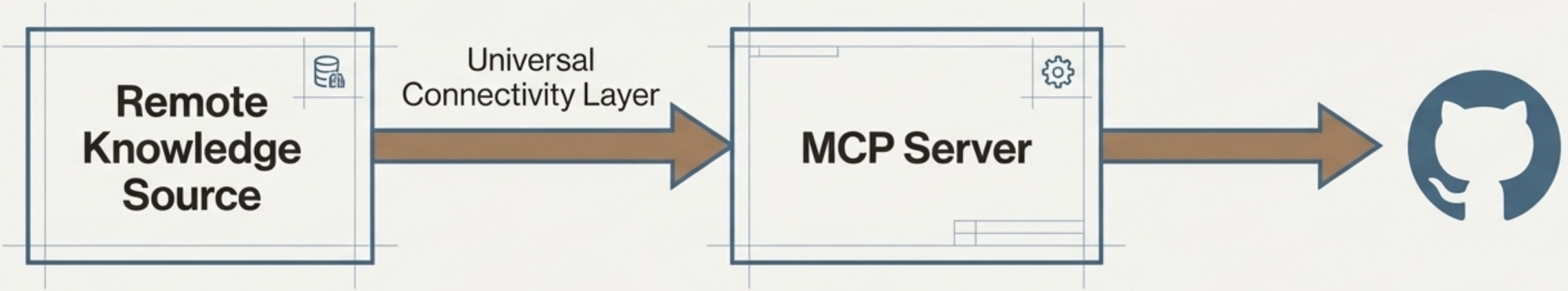
High-Fidelity Ingestion with Azure Content Understanding



Key Features

- Advanced OCR and PDF parsing
- Layout understanding to preserve document structure
- Accurate table extraction
- Image verbalization for charts and schematics

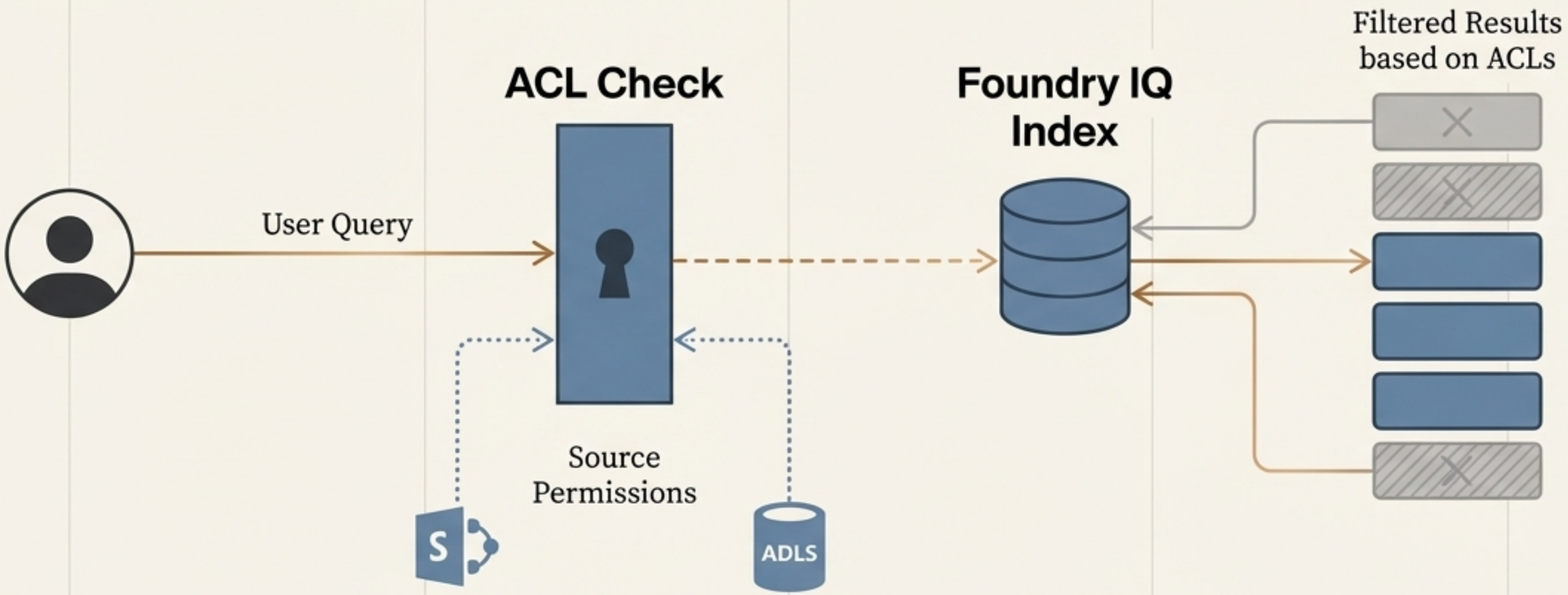
Connect to Any Application via Microsoft Copilot Platform Servers



Foundry IQ uses MCP servers as a universal connectivity layer. If an application has an MCP server—or you can build one for it—it can become a knowledge source.

> *"If you have a third-party application... very likely it has an MCP server."*

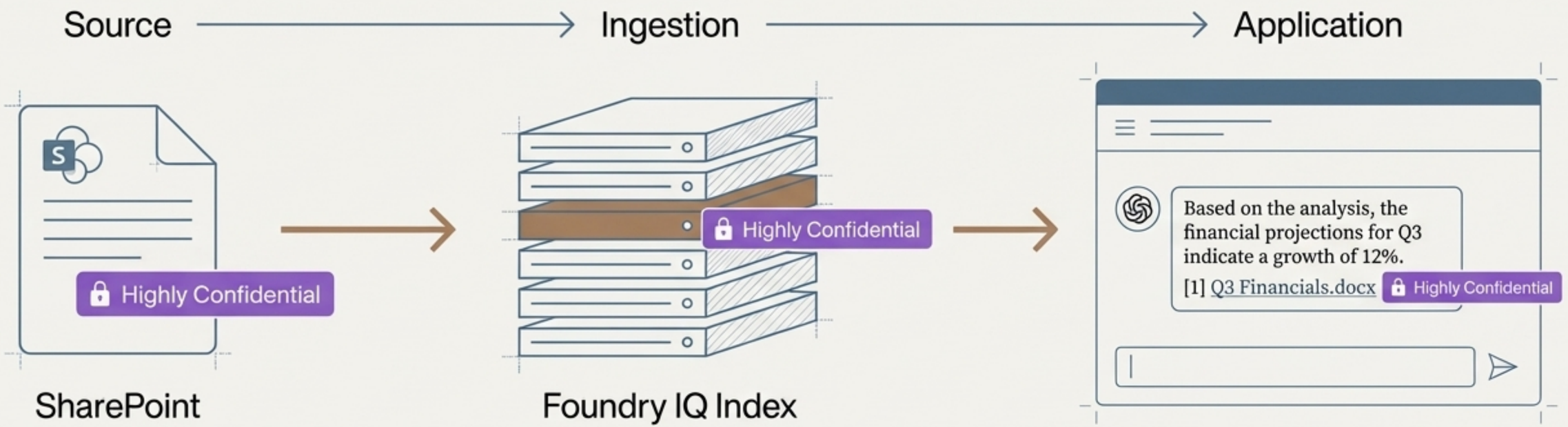
Security That Follows the Data: Access Control is Built-In



Users only see the data they are authorized to see.

Foundry IQ propagates and enforces source-level permissions automatically, ensuring queries are grounded only on accessible information.

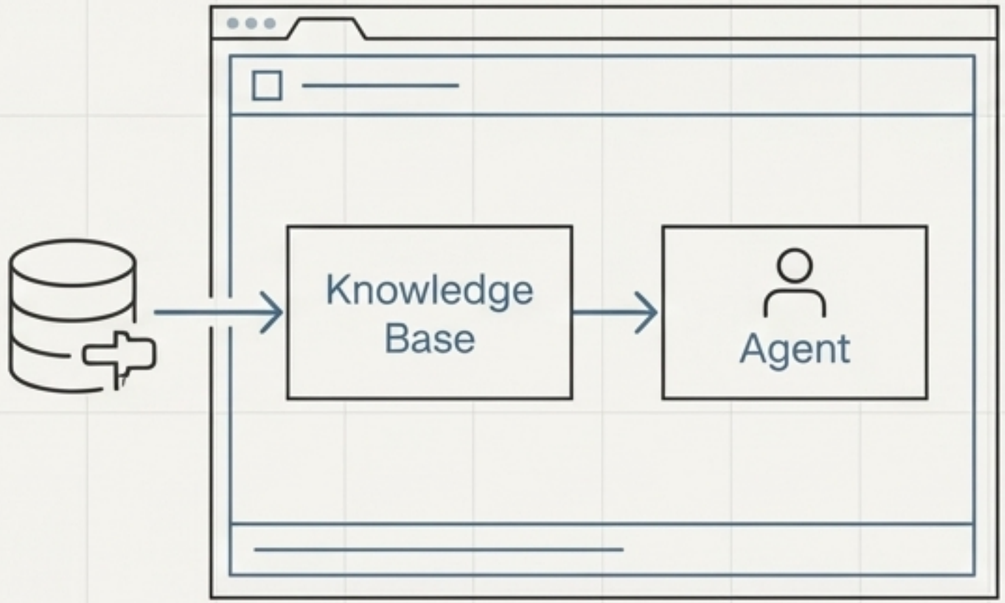
Enforce Governance with Native Microsoft Purview Integration



Handles encrypted documents, propagates sensitivity labels, and enables enforcement and display of labels in the end-user application.

A No-Cliffs Experience for Every Developer

Build in Foundry



Use Directly in Azure AI Search

```
from azure.search.documents import
  KnowledgeBaseRetrievalClient

client = KnowledgeBaseRetrievalClient(...)

results = client.retrieve(
    knowledge_base_name="demo-kb",
    query="What is the replacement policy
      for the CTL-11-R2 compressor?"
)
```

The Proof is in the Performance: Measurable Gains in Answer Quality

36%

Boost in Answer Score on hard questions when using Agentic Retrieval versus brute-force search across all sources.

+34 pts

Answer Score recovered on a fragmented finance dataset by enabling Web Grounding to find missing public information.

95%

Exit accuracy of the “fast exit check” on easy questions, demonstrating efficiency by avoiding unnecessary LLM iterations, saving time and tokens.

Foundry IQ: The Enterprise-Grade Retrieval System

1. Unified Architecture

Tame data complexity with the Knowledge Base abstraction. Connect to any source, indexed or remote.

2. Intelligent Engine

Achieve superior answer quality with the Agentic Retrieval loop. Adapt performance to your needs.

3. Enterprise-Ready Foundation

Build with confidence on a platform with built-in security, governance, and a no-cliffs developer experience.

Dive Deeper and Get Started

Read the Foundry IQ Deep-Dive

A detailed technical blog post explaining the architecture and capabilities.

blogs.microsoft.com/ai/foundry-iq-details

Neue Haas Grotesk Display Pro Medium

Explore the Evaluation Whitepaper

The full methodology, data sets, and results from our internal quality and performance testing.

AKA.ms/KB-evals

Scan to read the evaluation whitepaper.