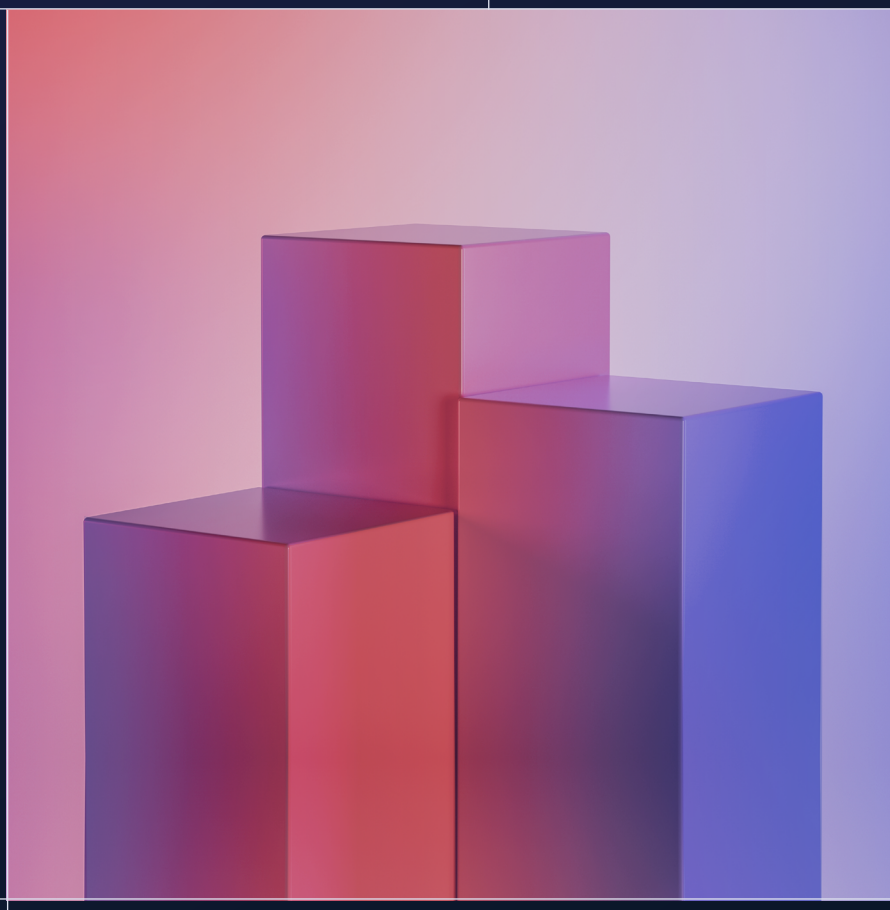
denodo

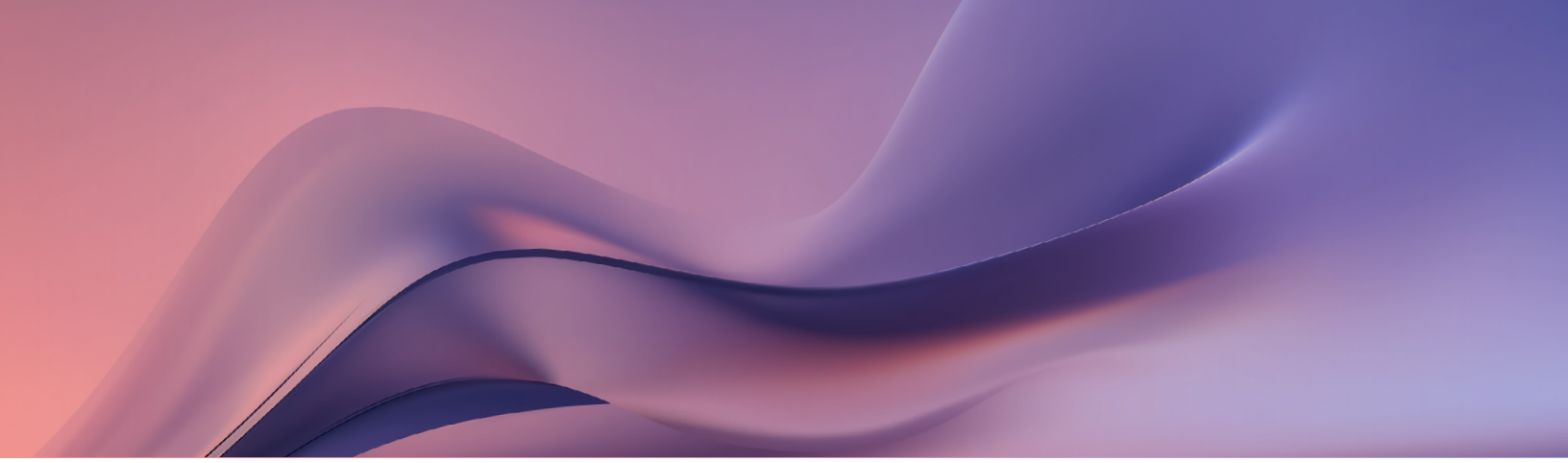
Whitepaper

Lower Cost, Higher Trust: The New Economics of Agentic AI

How the Denodo Platform optimizes agentic AI token consumption while improving trust



Contents



The Token Economics Challenge

Enterprise AI is moving from assistants that answer questions to agents that reason, retrieve, call tools, validate results, and act. That shift changes the economics. In a simple chat interaction, the model may process a prompt and produce a response. In an agentic workflow, the model may repeat that cycle many times, each time carrying forward prior context, retrieved data, tool outputs, intermediate reasoning, and error messages.

Tokens are the units AI models use to process prompts, retrieved data, tool outputs, and responses. Because many AI services are priced based on token consumption, every additional piece of context an agent carries forward can add cost, especially when that context is repeated across multiple steps in a workflow.

The result is that token usage does not merely accumulate; it compounds. When raw data is retrieved early in a multi-step task, those tokens may be re-sent on later turns. A small over-retrieval problem at the beginning of a task can become a large cost problem by the end of the task.

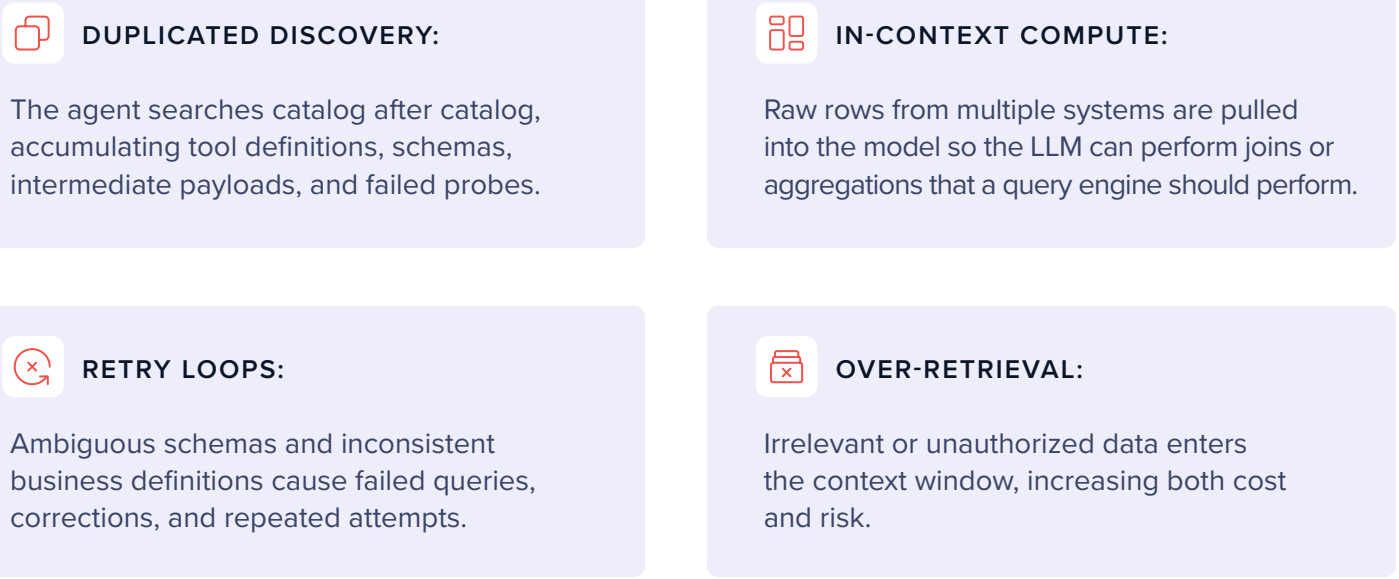
This is why AI success increasingly depends on optimized token usage. The goal is not to starve agents of context. The goal is to give agents the right context, at the right time, in the most compact and governed form possible.

For enterprise leaders, this creates a new operating challenge. AI cost is not only shaped by model pricing or infrastructure spend. It is also shaped by how agents discover data, how much context they retrieve, how often they retry failed tasks, and whether governance is applied before information reaches the model. Without control at this layer, organizations may scale AI usage while also scaling avoidable cost, complexity, and risk.

Why the Enterprise Data Reality Makes This Hard

Most enterprises do not have one clean data source for agents to use. They have lakehouses, warehouses, operational systems, SaaS applications, documents, vector stores, APIs, partner data, and on-premises systems. Each source may have different schemas, definitions, policies, access methods, and freshness characteristics. Agents that connect directly to each platform often consume unnecessary tokens before they can even answer the business question, because they must first navigate fragmented systems, inconsistent metadata, and source-specific access patterns.

This results in:



A Reason to Act Now: Avoid MCP Sprawl

Model Context Protocol (MCP) can help standardize how agents connect to tools, data, and applications. But if every data source or domain establishes its own MCP server, enterprises can quickly recreate the same integration sprawl they were trying to avoid. Each server must be secured, documented, monitored, updated, governed, and explained to agents. As the number of servers grows, so does the operational overhead, security surface area, and the amount of tool and schema context agents may need to evaluate before acting.

This creates a practical reason to establish a standard, governed data access layer early. By exposing business-ready context through one logical layer, organizations can reduce the need for source-by-source MCP patterns, limit avoidable complexity, and give agents a more stable and governed path to enterprise data.

AI Trust Gap Context

Denodo's **AI trust gap research** found that trustworthy agentic AI requires live data, the right data, and guardrails. The survey also found that enterprise AI initiatives draw on more than 400 data sources on average, while many organizations struggle with real-time data, trustworthy data identification, security/access controls, and cost/performance optimization.

Denodo: One Optimized Data and Context Delivery Layer

Denodo optimizes token consumption by changing what the agent has to bring into the context window. Instead of asking each agent to discover, retrieve, join, filter, and reconcile data entities source-by-source, Denodo provides a unified data and context layer that connects to the distributed enterprise data estate and returns business-ready answers. Also known as “active context,” this layer operationalizes trusted enterprise context for AI systems, applications, and human users alike. Unlike traditional data pipelines and static metadata catalogs, an active context layer is a unified logical tier that brings together shared business semantics, live zero-copy data access, centralized governance, and universal connectivity across all enterprise systems and data sources.

Active context also simplifies the surrounding agent architecture. Agents, copilots, and assistants can use a consistent, governed data and context delivery standard rather than relying on separate source-specific access paths for each AI use case. Agent developers get a stable, trusted interface for accessing enterprise context, while data teams continue to manage source connectivity, semantic definitions, policy enforcement, and data delivery behind the scenes.

This is not a point feature. It is the net result of adopting a complete approach that enterprises need for agentic AI: live data, the right data, and consistent guardrails, across all data sources. The middle layer in the diagram below shows the four “pillars” of active context necessary for that outcome.

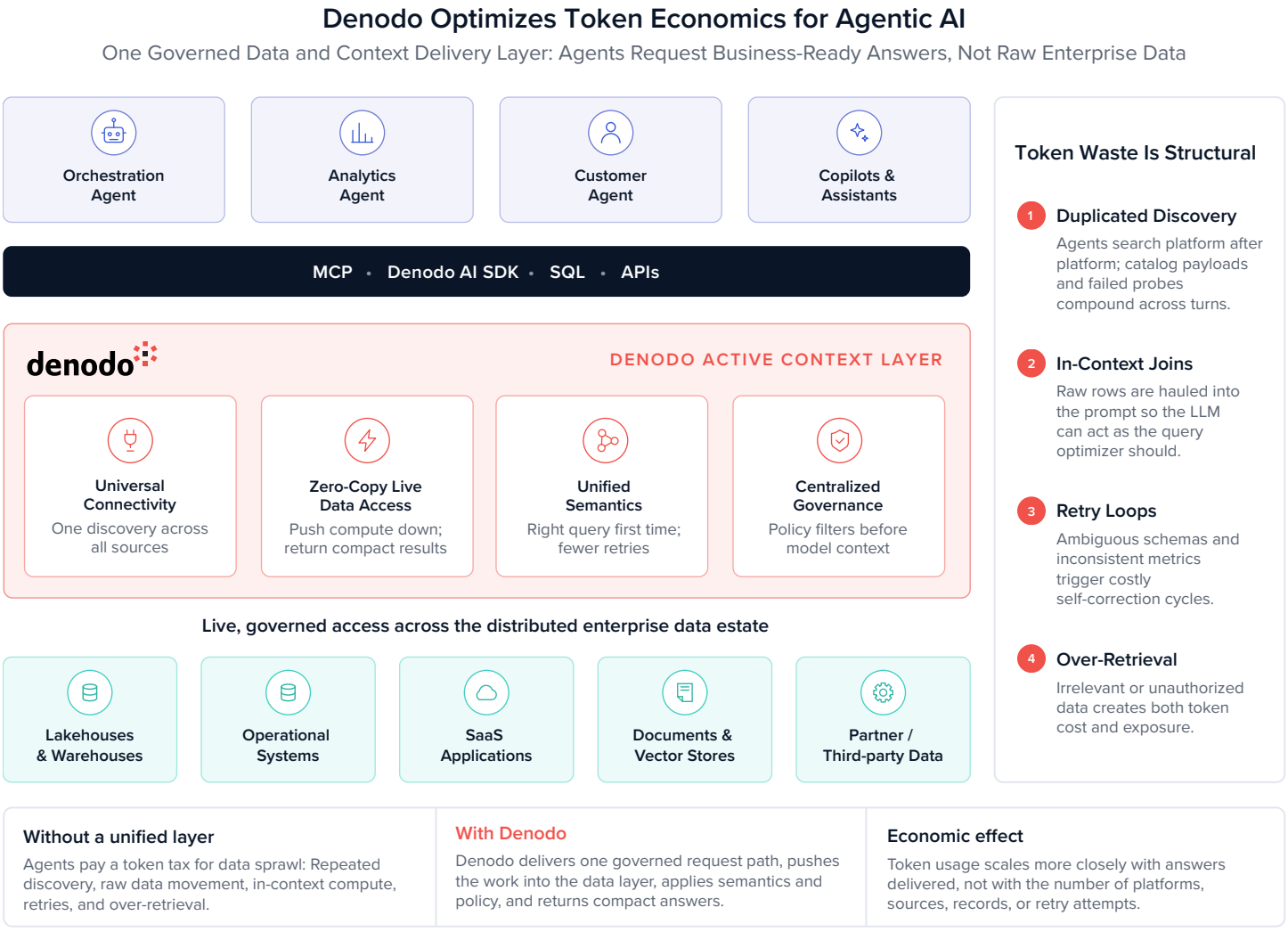


Figure 1. Denodo helps agents consume compact, governed context instead of raw data sprawl.

PILLAR	HOW IT OPTIMIZES TOKEN CONSUMPTION	WHY IT IMPROVES TRUST
Universal Connectivity	Agents search once through one logical layer instead of repeating discovery across many platforms, catalogs, APIs, and data stores.	Agents can reach the full enterprise data estate without hard-wiring inconsistent source-by-source logic.
Zero-Copy Live Data Access	Denodo pushes filters, joins, and aggregations into the data layer and returns compact result sets, not large raw payloads.	Agents act on current operational truth without waiting for every dataset to be copied into a separate AI store.
Unified Semantics	Consistent business definitions, relationships, and source-of-truth logic reduce failed queries and self-correction loops.	Agents reason with the right business meaning across systems, data products, and use cases.
Centralized Governance	Row-level security, column-level security, masking, and policy enforcement filter data before it reaches the model context.	Every answer respects entitlements, privacy rules, security policy, and business guardrails.

The Central Economic Shift

Without Denodo, token cost tends to scale with agents × tasks × sources. With Denodo, the data-access portion of token cost can scale more closely with business-ready answers delivered.

How the Four Pillars of Active Context Remove Avoidable Token Waste

UNIVERSAL CONNECTIVITY: ONE LAYER, ONE DISCOVERY

When agents use direct connectors to many systems, discovery becomes part of the token bill. The orchestration layer may need to inspect multiple catalogs, retrieve schemas, compare conflicting metadata, and handle failed probes. Those payloads can stay in context and be paid for repeatedly on later turns. Denodo provides a single discovery and access path across warehouses, lakehouses, SaaS applications, operational systems, documents, vector stores, APIs, and third-party data. The agent searches once, against a unified semantic catalog, and receives one governed answer path.

ZERO-COPY LIVE DATA ACCESS: PUSH COMPUTE DOWN, NOT DATA UP

A common source of token waste occurs when raw records are hauled into the context window so the LLM can compare, join, aggregate, or filter them. This is expensive and risky: the LLM is being asked to do probabilistic reasoning over work that should be performed by a deterministic query engine. Denodo's federated optimizer can push filters, joins, and aggregations down to source systems or execute them in the logical layer. The model receives a compact result set, not the raw material.

UNIFIED SEMANTICS: THE RIGHT QUERY, FEWER RETRIES

The hidden token killer is often the failed query. If "customer," "active account," "net revenue," or "risk exposure" means different things in different systems, the agent may query the wrong source, select the wrong metric, or attempt multiple corrections. Denodo's semantic layer gives agents consistent business meaning, relationships, and source-of-truth guidance. That improves trust and reduces the wasted tokens associated with trial-and-error retrieval.

CENTRALIZED GOVERNANCE: POLICY AS A TOKEN FILTER

Governance is often framed only as risk control. In agentic AI, it is also cost control. Every irrelevant, restricted, masked, or unauthorized field that enters the context window creates token cost and potential exposure. Denodo enforces access controls, masking, and policies before delivery, so the context contains only what is relevant and permitted. The most secure token is the one that never enters the prompt. It is also the least expensive.

Do Not Pay an LLM to Do What a Query Optimizer Does Better

The LLM should interpret the task, invoke the right tool, and explain the result. It should not be the place where enterprise joins, aggregations, filters, security checks, and source-system reconciliations are performed.

Caching and Compression Still Matter

Downstream techniques such as prompt caching, summarization, and compression can reduce model cost. Denodo makes models more effective by presenting a smaller, more stable, more reusable context interface in the first place.

Why This Matters across the Enterprise

Token economics matter to different stakeholders for different reasons, but the common concern is whether agentic AI can scale without adding unnecessary cost, complexity, or risk.

CIOs and CTOs

Need a scalable AI architecture that does not create uncontrolled data-access patterns or governance gaps.

CFOs and finance leaders

Need confidence that AI growth is tied to efficient consumption and measurable business value, not avoidable waste.

CDOs and data leaders

Need agents to use trusted, current, business-ready data rather than raw, inconsistent, or poorly governed context.

AI platform teams

Need to reduce unnecessary discovery, retrieval, retries, and tool calls so agents can operate more efficiently.

Application owners

Need AI-powered experiences that can scale economically without degrading trust, performance, or user experience.

A Four-Pillar ROI Model for Optimized Token Consumption

The following model estimates the monthly token impact of Denodo on the data-access portion of agentic AI workloads. It is intentionally directional. It does not claim a universal savings percentage, and it excludes infrastructure, vector database, model hosting, and labor costs. Its purpose is to help a data or AI leader identify where tokens are being generated unnecessarily and where Denodo can optimize consumption.

CORE WORKLOAD INPUTS

INPUT	WHAT IT REPRESENTS	EXAMPLE
T = Monthly agent tasks	Number of agent tasks that require enterprise data access	500,000
R = Context compounding multiplier	Average number of subsequent turns in which retrieved data remains in context	6
P = Token price	Blended model price per 1 million tokens	$X / 1M tokens

PILLAR-LEVEL INPUTS

PILLAR	BASELINE WITHOUT DENODO	WITH DENODO
Universal Connectivity	S = Average # of data sources searched; D = Discovery tokens per data source	d = One Denodo discovery/ context response
Zero-Copy Live Data Access	Raw = Raw tokens retrieved for joins, filters, and aggregation	Ans = Compact answer tokens returned after optimization
Unified Semantics	RetryBase = Failed-query rate × tokens per retry	RetryDenodo = Lower failed-query rate × tokens per retry
Centralized Governance	Over = Irrelevant or unauthorized tokens retrieved	Gov = Permitted, policy-filtered tokens delivered

MONTHLY TOKEN FORMULAS

Baseline data-access tokens = T × R × [(S × D) + Raw + Over] + T × RetryBase

Denodo-optimized data-access tokens = T × R × [d + Ans + Gov] + T × RetryDenodo

Estimated optimization = 1 - (Denodo-optimized tokens ÷ Baseline data-access tokens)

ILLUSTRATIVE EXAMPLE

CATEGORY	WITHOUT DENODO	WITH DENODO
Discovery tokens	4 data sources (S) × 1,000 tokens (D) = 4,000	One governed discovery response (d) = 800
Data Payload tokens	8,000 raw tokens (Raw) pulled into context	1,000 compact answer (Ans) tokens
Governance / Over-Retrieval	2,000 irrelevant or restricted tokens (Over)	200 permitted policy-filtered tokens (Gov)
Retry Overhead	20% retry rate × 6,000 tokens = 1,200 per task (RetryBase)	5% retry rate × 4,000 tokens = 200 per task (RetryDenodo)
Compounded Retrieval Tokens	Baseline data-access tokens: 500k × 6 × (4,000 + 8,000 + 2,000) + (500k x 1,200) = 85,200 per task	Denodo-optimized data-access tokens: 500k × 6 × (800 + 1,000 + 200) + (500k x 200) = 12,200 per task

Using 500,000 monthly agent tasks, the illustrative baseline would be 42.6 billion data-access tokens per month, while the Denodo-optimized pattern would be 6.1 billion. That implies an 85.7% reduction in avoidable data-access tokens in this scenario. This is not a benchmark; it is a modeling framework. Actual results depend on agent design, prompt strategy, data volumes, model choice, caching, retrieval patterns, and how much returned data is passed back to the model for final reasoning or explanation.

IN CONCLUSION: TRUST AND ECONOMICS ARE THE SAME ARCHITECTURE

AI leaders should not only ask how to manage the tokens their agents consume. They should also ask why those tokens are being consumed at all. Denodo optimizes token economics by preventing unnecessary tokens from being generated upstream: Agents discover once, compute is pushed down, semantics reduce retries, and governance filters data before it reaches the model. The same logical layer that makes agents trustworthy also makes them more economically scalable.

As organizations expand agentic AI, this same architecture can also help them avoid the long-term cost of fragmented access patterns and MCP sprawl. Establishing one governed data and context layer early gives teams a more agile foundation for new agents, new data sources, and new AI use cases.

NEXT STEP:

Contact us for a free consultation on how Denodo can optimize your AI economics.

denodo

Visit www.denodo.com | Email info@denodo.com | Discover community.denodo.com