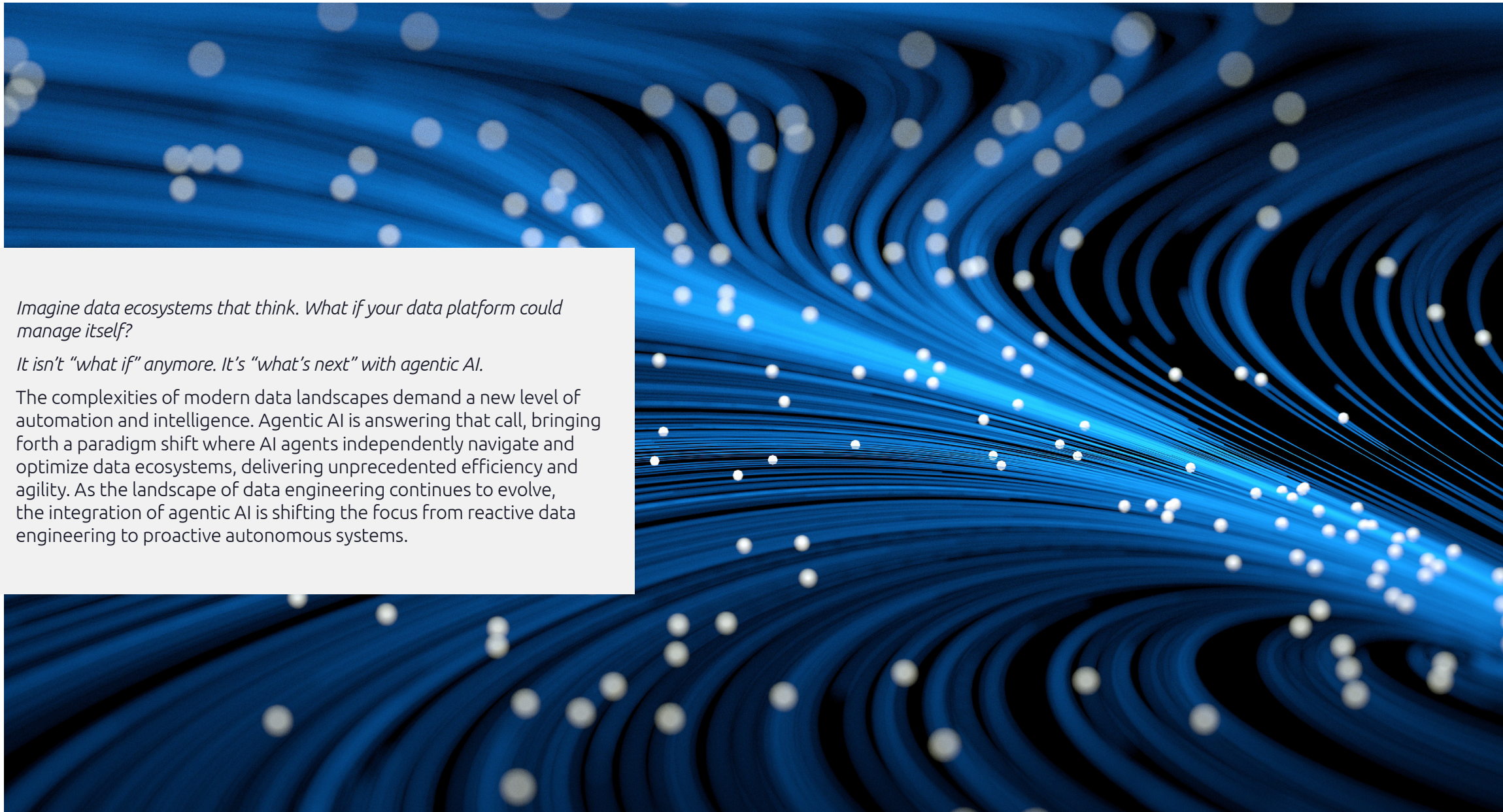
Business, *meet* agentic AI.

Agentic AI for autonomous *data engineering*

Capgemini



Imagine data ecosystems that think. What if your data platform could manage itself?

It isn't "what if" anymore. It's "what's next" with agentic AI.

The complexities of modern data landscapes demand a new level of automation and intelligence. Agentic AI is answering that call, bringing forth a paradigm shift where AI agents independently navigate and optimize data ecosystems, delivering unprecedented efficiency and agility. As the landscape of data engineering continues to evolve, the integration of agentic AI is shifting the focus from reactive data engineering to proactive autonomous systems.

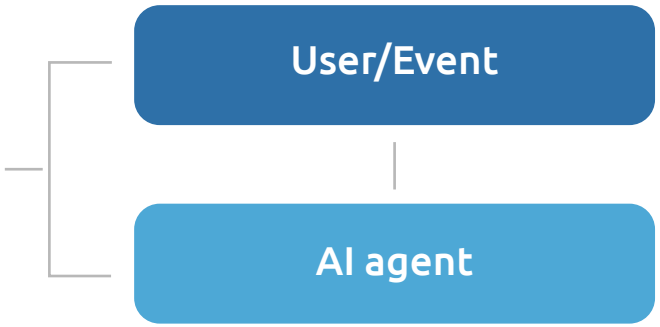
Demystifying *agentic AI*

Agentic AI refers to artificial intelligence systems that can autonomously pursue specific objectives with minimal human guidance. At its core, it comprises AI agents, which are essentially machine learning models designed to emulate human-like decision-making processes to solve problems in real time.

Agentic AI often builds upon the foundations of generative AI by utilizing large language models (LLMs) to operate effectively within dynamic environments. While generative models excel at creating new content based on learned patterns, agentic AI extends this capability by applying the outputs generated by these models toward the accomplishment of specific tasks. In a nutshell, generative AI is AI that creates, whereas agentic AI is AI that acts.

Several characteristics underpin the functionality and potential of agentic AI that includes autonomy, perception, goal orientation, learning, adaptability, reasoning, decision-making, and execution. These characteristics are enabled in the AI agent with key modular components given below.

Key components of AI agent



Perceive, reason, act, and learn

Perception module
Collect and interpret input data

Cognitive module
Agent's brain for thinking and decision-making

Action module
Translating agent's decision to execute actions

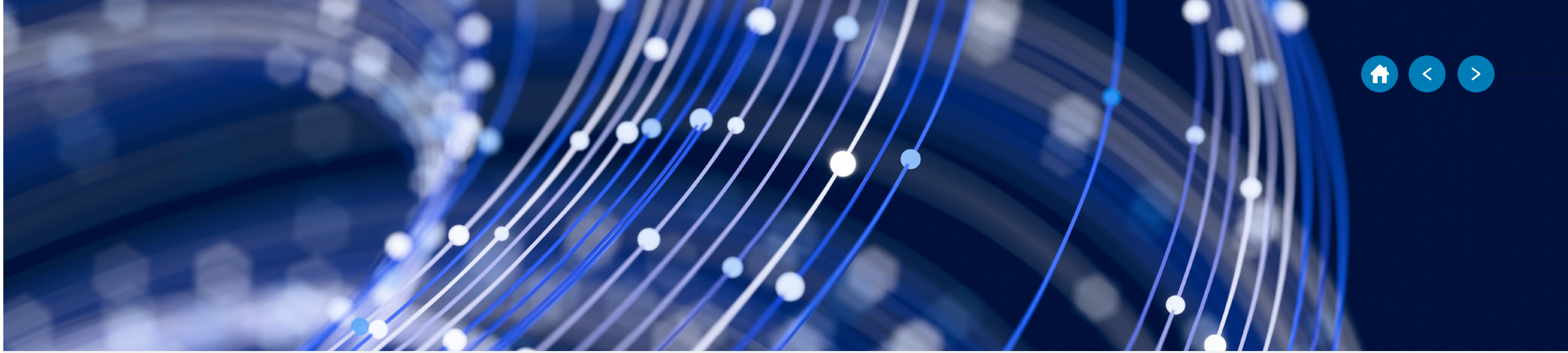
Knowledge module
Knowledge base for context-aware responses

Memory module
Retain information over time and maintain context

Learning module
Improve performance from experience and feedback

Monitoring and evaluation module
Tracks agent's performance and provides metrics for improvement

Communication module
Exchanges information with other agents, humans, or systems



Convergence of *data engineering and agentic AI*

Data engineering has emerged as a vital discipline focused on the design, construction, and maintenance of systems that enable the effective management and transformation of raw data into a usable and insightful resource.

In parallel, the field of artificial intelligence is undergoing a significant evolution with the advent of agentic AI. This paradigm represents a departure from traditional AI models that primarily react to specific inputs or follow predefined rules.

The intersection of these two dynamic fields, agentic AI and data engineering, holds immense potential for transformative synergy. The inherent capabilities of agentic AI, such as its autonomy, adaptability, and goal-oriented nature, align directly with the pressing needs within data engineering to automate complex processes, enhance efficiency, and improve the overall quality of data management practices.

Data engineering is absolutely fundamental to an organization's ability to gain insights and make sound decisions. When the data layer falters due to its inherent complexity, diverse sources, and the specialized skills required, an organization's decision-making is severely impaired. Data quality issues, pipeline failures, or slow data processing can lead to flawed insights and operational inefficiencies.

The business imperative is clear: maintain a robust data foundation without tying up an army of data engineers. This is where agentic AI becomes transformative. Agentic AI can automate routine data tasks like cleaning, transformation, and schema inference, proactively detect and resolve issues, and even enforce data governance. This convergence frees human data engineers for strategic work, accelerates time to insight, and enables scalable, reliable data operations, empowering truly data-driven organizations.

AI agents can accelerate the entire data engineering process, making it more efficient and agile and accelerating it across different stages of the SDLC.

Product backlog

Data engineering

Requirements
Requirement validators

Design
Design optimizers

Build
Code standardizers

Test
Test orchestrators

Run
Auto-remediators

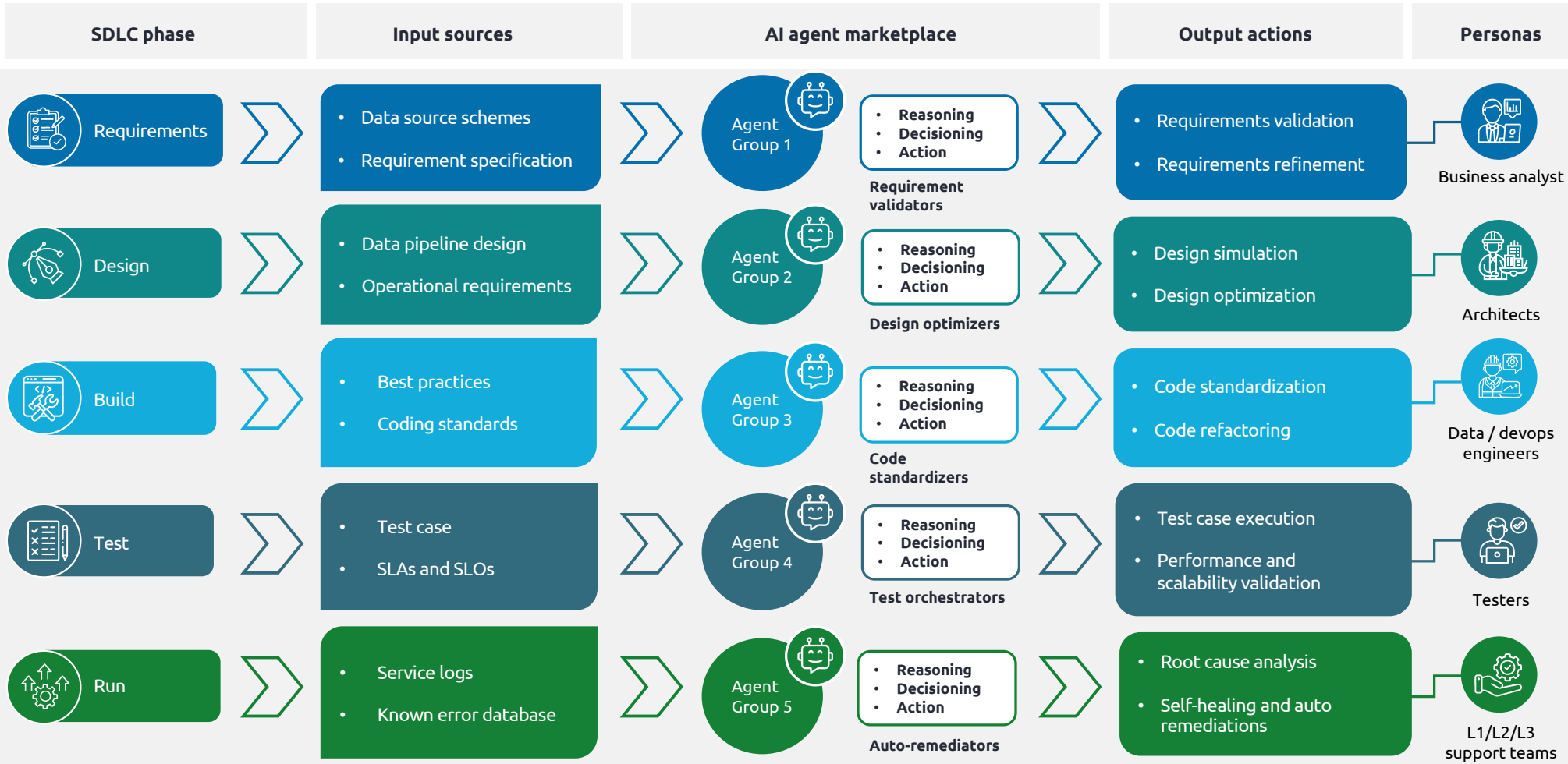
AI agent marketplace for data engineering

SDLC acceleration

The integration of AI agents into the data engineering software development life cycle (SDLC) is poised to revolutionize how data systems are built, deployed, and maintained.

Organizations stand to gain significant advantages by strategically defining and leveraging an AI agent marketplace, particularly within the context of data engineering. An AI agent marketplace would allow organizations to access a variety of specialized AI agents designed to optimize different aspects of the data engineering SDLC.

Here's a view of how an AI agent marketplace can streamline and optimize every stage of the data engineering SDLC, with acceleration examples:


SDLC phase
Input sources
AI agent marketplace
Output actions
Personas
Requirements
Data source schemes
Requirement specification
Agent Group 1
Reasoning
Decisioning
Action
Requirement validators
Requirements validation
Requirements refinement
Business analyst
Design
Data pipeline design
Operational requirements
Agent Group 2
Reasoning
Decisioning
Action
Design optimizers
Design simulation
Design optimization
Architects
Build
Best practices
Coding standards
Agent Group 3
Reasoning
Decisioning
Action
Code standardizers
Code standardization
Code refactoring
Data / devops engineers
Test
Test case
SLAs and SLOs
Agent Group 4
Reasoning
Decisioning
Action
Test orchestrators
Test case execution
Performance and scalability validation
Testers
Run
Service logs
Known error database
Agent Group 5
Reasoning
Decisioning
Action
Auto-remediators
Root cause analysis
Self-healing and auto remediations
L1/L2/L3 support teams

Each agent could be tailored to offer varying degrees of benefits and handle specific tasks across the SDLC with the autonomy to reason, decide, and act, providing targeted solutions that accelerate the end-to-end cycle and ensure smoother, more efficient operations, saving significant cost and efforts.

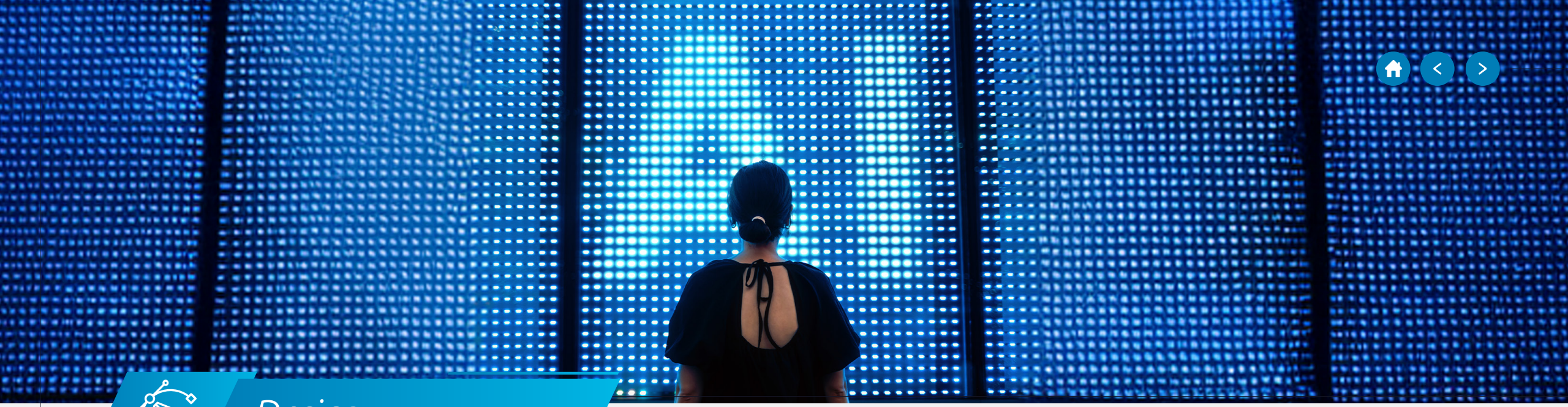
Below are a few example use case scenarios where AI agents deliver acceleration and value as a powerful assistant at each phase of the data engineering SDLC.

Requirements

Integrating new data sources poses significant, often unseen, risks to data platforms. It's tough to predict if existing infrastructure can handle new volumes or if new data will violate privacy rules or quality standards. Manually assessing these impacts during initial requirements is slow and error-prone, leading to costly surprises. Businesses frequently define new data needs, like historical transaction data for fraud or real-time social media sentiment, but these initial requirements are often vague, inconsistent, or simply unfeasible.

The main problem is the inability to accurately and efficiently validate new data requirements before development. This leads to unforeseen infrastructure strain, potential compliance and quality violations, and costly rework due to slow, error-prone manual validation. Data engineers are left to deal with vague and often unfeasible requirements.

Agentic AI steps in as a "requirement validator." It autonomously validates these requirements before any design work begins, ensuring they're feasible, consistent, and compliant. The AI cross-references new needs against existing data sources, infrastructure constraints, and critical compliance policies (like GDPR). For instance, it might identify that a request for five years of historical data is costly and raises GDPR issues, or that real-time social media sentiment is infeasible due to API limits. It also flags conflicting data definitions, prompting clarification. Ultimately, agentic AI provides proactive reports and actionable suggestions, refining requirements, highlighting feasibility issues, identifying compliance conflicts, and ensuring consistent data definitions. This critical early validation prevents expensive rework, ensuring data engineers build the right data solutions from the outset.



Design

Organizations are undergoing critical data migrations from legacy systems to modern data platforms. These legacy environments often use proprietary technologies, and expertise in them is increasingly scarce. Compounding this, designing complex ETL/ELT pipelines for optimal performance, scalability, and cost-efficiency is a significant challenge. Current manual validation frequently misses crucial bottlenecks or inefficient resource allocations, often leading to expensive re-architecting after the pipeline is already implemented.

The scarcity of specialized skills for understanding legacy schemas and implicit business logic significantly bottlenecks the migration's design phase. Manually deciphering old systems and meticulously designing ETL/ELT pipelines is slow, error-prone, and costly. The core issue is the inability to effectively validate ETL/ELT pipeline designs before implementation. This results in suboptimal performance, scalability issues, lack of resilience, and unnecessary infrastructure costs due to inefficient resource use. Discovering these problems post-implementation forces costly and time-consuming rework, delaying critical initiatives.

Agentic AI acts as an intelligent "design optimizer" during this critical phase. It autonomously analyzes legacy metadata, code, and data to infer schemas, identify data dependencies, and validate the technical architecture and components of ETL/ELT pipelines. Its job is to ensure how the actual build meets demanding performance, scalability, resilience, and cost-efficiency targets. The AI analyzes detailed design blueprints, like Spark configurations, and runs sophisticated simulations. For instance, it can simulate peak transaction bursts on a fraud detection pipeline, identifying optimal resource allocation, detecting resilience gaps, and proposing specific design modifications. Ultimately, agentic AI provides proactive reports and actionable suggestions, ensuring data pipelines are built correctly and efficiently upfront. This critical early validation prevents costly rework, accelerating the entire data migration and pipeline design process.

Build

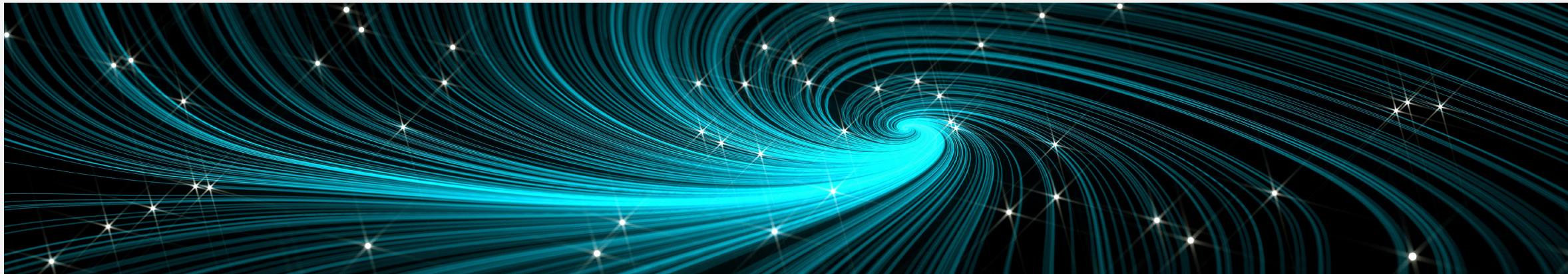
In data engineering, maintaining consistent coding standards and adhering to best practices (like PySpark optimization, Airflow DAG conventions, or proper error handling) is a constant uphill battle. Manual code reviews are slow, often missing subtle inefficiencies, resource leaks, or style deviations. This leads to accumulating technical debt and degraded pipeline performance over time. Furthermore, the migration from legacy systems to modern data platforms requires a lot of code for complex transformations and data movement.

The main problem is the difficulty in consistently applying coding standards and proactively identifying code quality issues. Manual reviews are insufficient for ensuring high-quality, performant, and maintainable data pipelines, placing a heavy burden on human reviewers and allowing issues to slip through. Also, manual coding for migration is incredibly slow, prone to errors, and expensive.

An agentic AI system acts as a continuous "code quality guardian and standardizer," directly embedded within the CI/CD pipeline. Given predefined coding standards and access to code repositories and pipeline metrics, it autonomously monitors and detects issues. AI agents analyze legacy codebases, schema definitions, and data samples to understand inherent business logic and data structures to auto-generate optimized context-aware ETL/ELT scripts and transformation routines for the target migration.

The AI continuously analyzes code changes, proactively identifying noncompliant or inefficient code blocks (e.g., suboptimal Spark operations, inadequate logging, or missing error handling). Beyond just flagging problems, the AI provides contextual refactoring suggestions, explaining why changes are needed, and can even initiate proposed fixes by opening pull requests (e.g., converting iterative processes to more efficient vectorized operations). Crucially, the AI "gatekeeps" by preventing the merge of code that fails critical quality checks, ensuring only high-standard code enters the main branch.

This significantly enhances code standardization, drives continuous code refactoring, drastically reduces manual review burdens, and ultimately ensures the development of high-quality, performant, and maintainable data pipelines with minimal human oversight.



Test

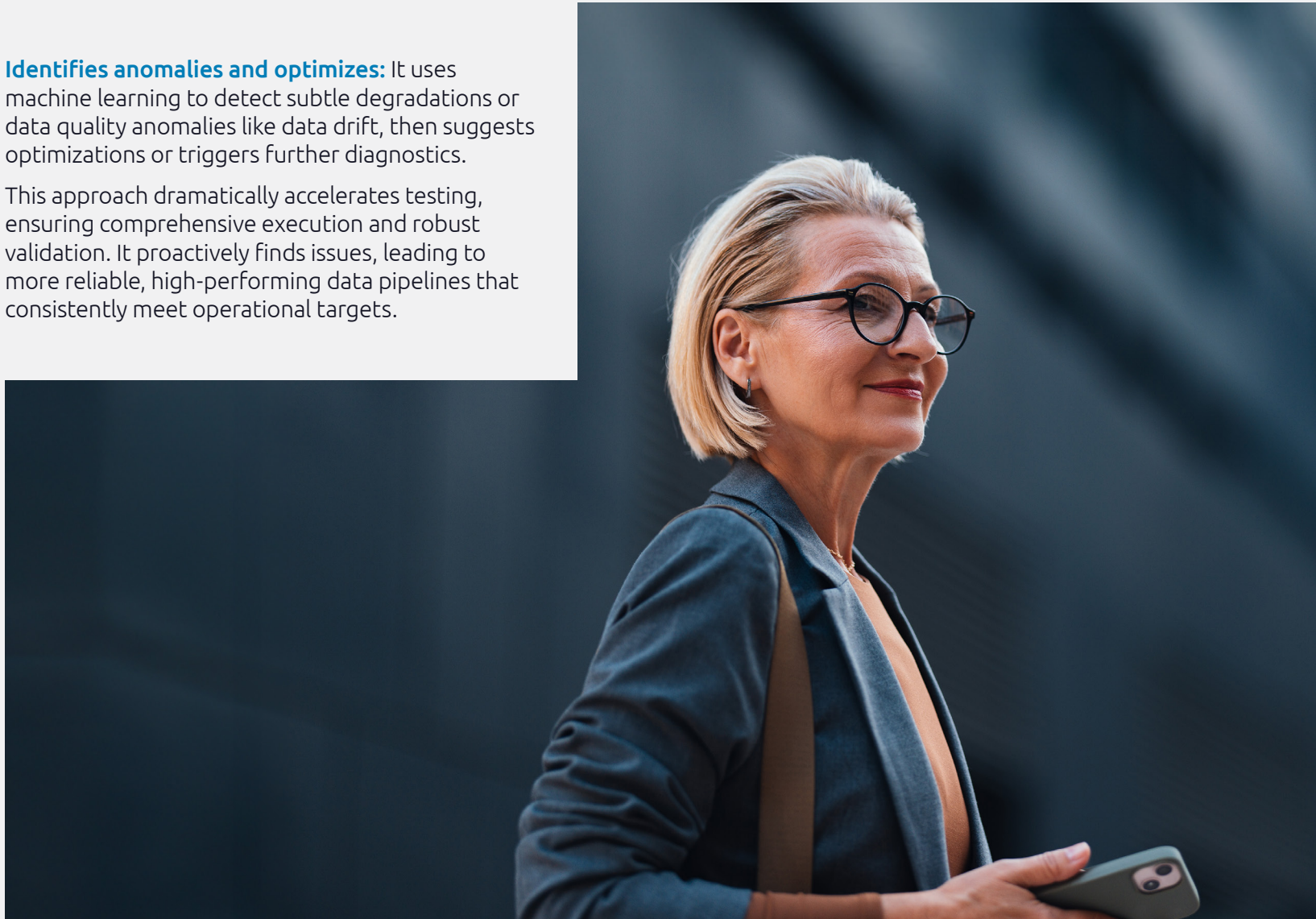
Manually testing data pipelines is inherently slow and error-prone. Trying to cover numerous test cases and validate complex SLAs/SLOs for data freshness, throughput, or latency often results in incomplete coverage and overlooked bottlenecks. This inevitably leads to poor data quality once pipelines hit production.

The core issue is the inefficiency and ineffectiveness of manual data pipeline testing. This results in incomplete test coverage, missed bottlenecks, and ultimately poor data quality in production, failing to meet crucial operational targets.

Agentic AI steps in as an autonomous "testing orchestrator." Given predefined test cases and critical SLAs/SLOs, the AI continuously:

- **Orchestrates test execution:** It intelligently schedules and runs unit, integration, and end-to-end tests across diverse environments, dynamically scaling resources as needed.
- **Validates performance and scalability:** It actively monitors pipeline metrics against SLAs, identifying if a Spark transformation misses latency targets or if a Kafka consumer group fails throughput requirements, for example. It also conducts stress tests.
- **Identifies anomalies and optimizes:** It uses machine learning to detect subtle degradations or data quality anomalies like data drift, then suggests optimizations or triggers further diagnostics.

This approach dramatically accelerates testing, ensuring comprehensive execution and robust validation. It proactively finds issues, leading to more reliable, high-performing data pipelines that consistently meet operational targets.



Run

During the “Run” phase, data pipelines constantly face threats that lead to costly downtime and data integrity problems. Manually diagnosing issues like schema drift (e.g., changes in source columns), data quality problems (e.g., sudden null spikes, duplicate records), infrastructure issues (e.g., full storage, network latency), or technical code errors (e.g., misconfigured Spark jobs) by sifting through vast service logs is inefficient and reactive.

The core problem is the inefficient and reactive manual diagnosis and resolution of operational issues in data pipelines. This leads to extended mean time to resolution (MTTR), compromising data freshness and reliability, and burdens data engineers with constant “firefighting” instead of proactive development.

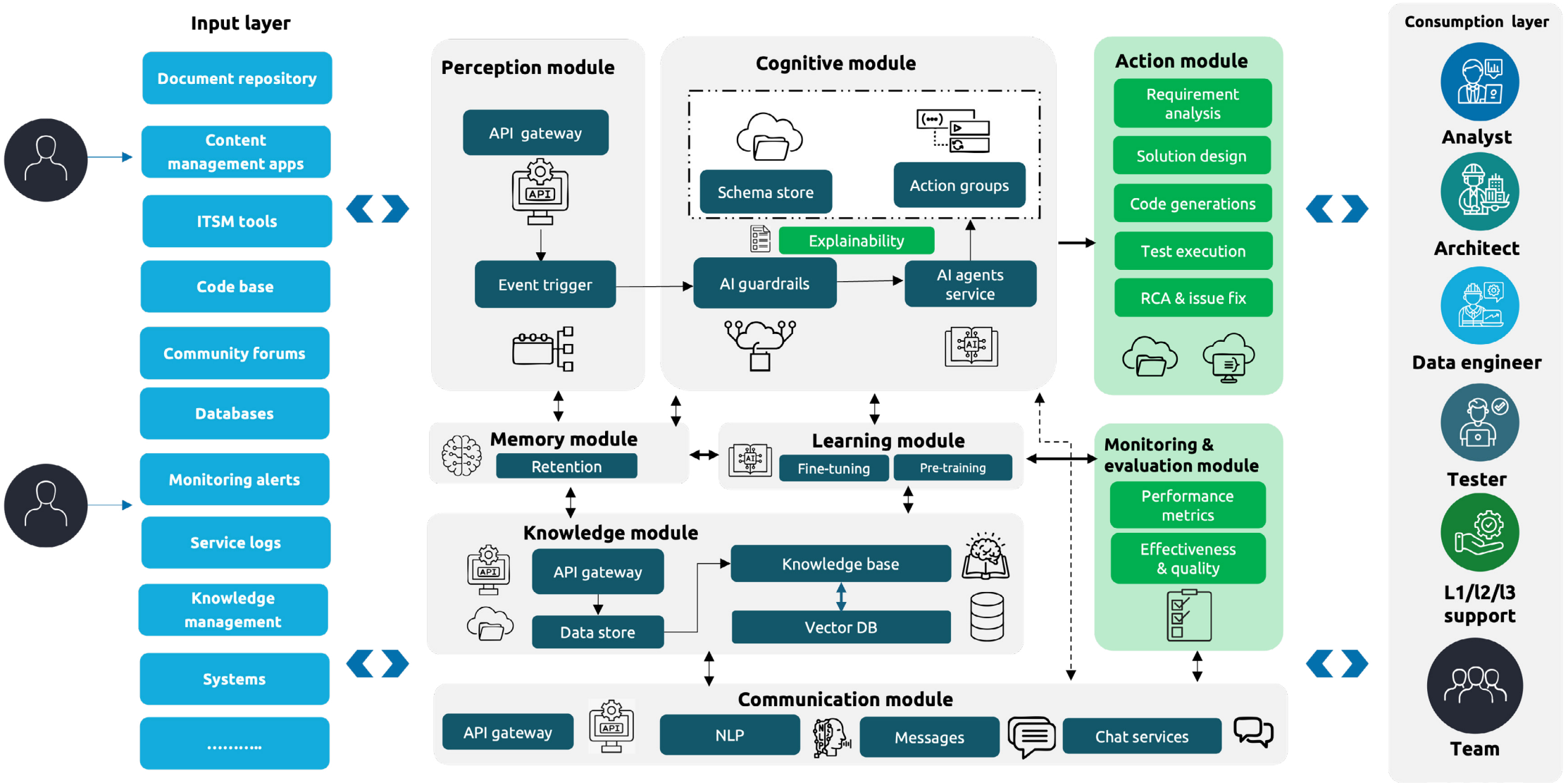
Agentic AI acts as a self-managing “automated remediator,” continuously monitoring service logs and integrating with a known error database (KEDB). It autonomously performs:

- **Intelligent root cause analysis (RCA):** For any anomaly, the AI instantly analyzes logs, identifies patterns, and cross-references the KEDB. It specifically diagnoses issues like schema drift (new/missing columns), data quality (unexpected distributions), infrastructure issues (resource saturation), and technical code errors (logic errors).
- **Self-healing and auto-remediations:** For diagnosed problems, the AI autonomously executes solutions. This includes proposing schema evolution scripts, quarantining bad data, scaling up infrastructure, retrying jobs with adjusted parameters, or deploying pre-approved code patches.
- **Contextual alerting:** If auto-remediation isn’t feasible, it escalates with precise details.

This significantly reduces MTTR, ensures data freshness and reliability, and frees data engineers from constant firefighting, even amid complex operational challenges.

AI agent enablement with *cloud AI services*

Leveraging cloud AI services is paramount for building robust and scalable agentic AI architectures. Hyperscalers like AWS, Google Cloud, and Azure provide a rich ecosystem of tools and services that can be combined to create powerful agent-based systems. Here is a reference architecture for enabling AI agents with cloud AI services:


Input layer
Document repository
Content management apps
ITSM tools
Code base
Community forums
Databases
Monitoring alerts
Service logs
Knowledge management
Systems
.....
Perception module
API gateway
Event trigger
Cognitive module
Schema store
Action groups
Explainability
AI guardrails
AI agents service
Action module
Requirement analysis
Solution design
Code generations
Test execution
RCA & issue fix
Memory module
Retention
Learning module
Fine-tuning
Pre-training
Monitoring & evaluation module
Performance metrics
Effectiveness & quality
Knowledge module
API gateway
Data store
Knowledge base
Vector DB
Communication module
API gateway
NLP
Messages
Chat services
Consumption layer
Analyst
Architect
Data engineer
Tester
L1/l2/l3 support
Team

Input layer

The input data to the agent can include organization knowledge sources like document repositories, knowledge management pages, databases, APIs, service logs, etc.

Perception module

The module focusing on collecting and interpreting the input data can be enabled with integration services such as API gateways and event triggers.

Cognitive module

The brain of the AI agent for thinking and decision-making can be enabled with responsible AI guardrails, schema store, multi-agent services, and explainability.

Knowledge module

The agent's knowledge base can be enabled with RAG and vector database services.

Memory module

The module to retain information can be enabled with short-term cache memory services and long-term semantic memory database services.

Learning module

The module to improve performance can be enabled with pre-training and fine-tuning.

Action module

The module for AI decision into actions can be enabled with compute and data store.

Monitoring and evaluation module

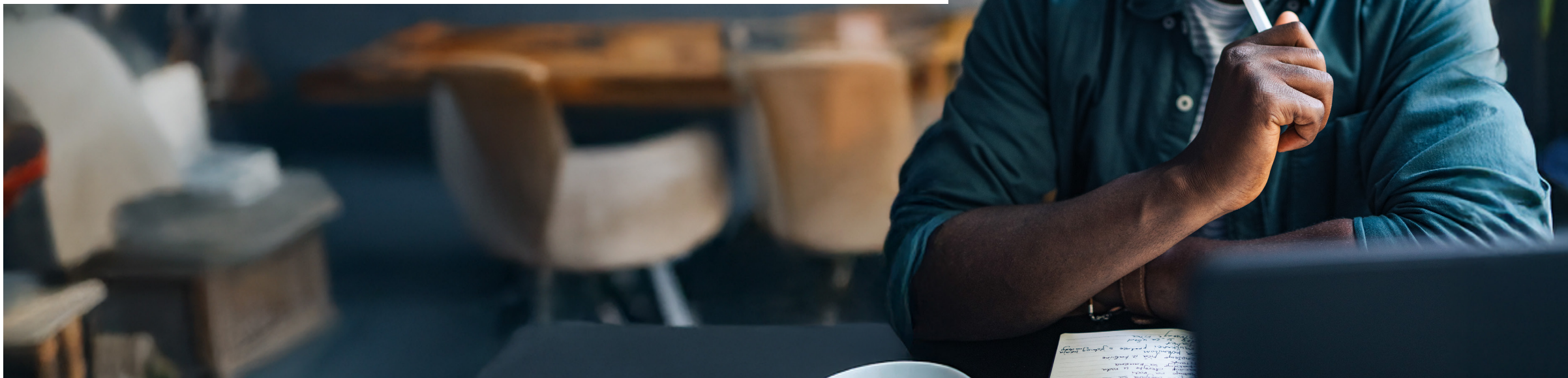
This module can be enabled with monitoring, reporting, and observability services.

Communication module

This module can be implemented with APIs, NLP, messaging, and chat services.

Consumption layer

The consumer layer of agentic AI can include humans, systems, or other agents.



Ethical AI *considerations*

The rise of agentic AI in data engineering demands a robust ethical framework throughout its life cycle. From the outset, the agent's purpose should be clearly defined to establish accountability and integrate human oversight. Proactive identification of data bias, stringent privacy measures, and regulatory compliance are non-negotiable.

Transparency in how AI agents make decisions is also crucial for building trust and ensuring accountability. Implementing explainable AI (XAI) is key for transparency with explicit human-in-the-loop mechanisms for critical data decisions, ensuring full auditability of the end-to-end processes.

Ethical considerations aren't static; the systems must adapt and evolve responsibly. Robust governance frameworks are needed to ensure that AI agents operate ethically and in alignment with business objectives. Furthermore, organizations need to manage the transition for their data engineering teams, providing adequate training and support to effectively collaborate with AI agents.

Conclusion

Agentic AI presents a transformative opportunity to make your data platform truly intelligent. Its unique combination of autonomy, adaptability, and goal-oriented behavior offers compelling solutions to many of the persistent challenges faced by data engineering teams today. Embracing agentic AI with a strategic and responsible approach focusing on AI ethics and explainability will deliver significant benefits to the organization in terms of efficiency, speed, scalability, and innovation. Capgemini helps you industrialize and modernize your data estate with agentic AI solutions, leveraging IDEA (Industrialized Data and AI Engineering Acceleration). Get in touch and let's unlock your potential as a data-powered organization.

About Capgemini

Capgemini is a global business and technology transformation partner, helping organisations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 350,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fuelled by its market leading capabilities in AI, generative AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2024 global revenues of €22.1 billion.

Get the future you want | www.capgemini.com

Capgemini