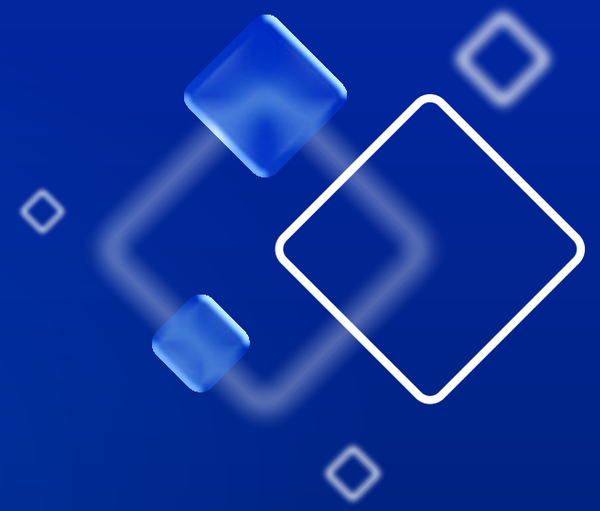
Informatica
from salesforce

How to Operationalize Data Science and Machine Learning with Trusted Data



informatica.com

E-BOOK

Table of Contents

From Insight Team to Innovation Engine: How Data Science Powers the AI Enterprise 3

Why Data Science Only Delivers Business Value If the Data Is Ready 5

6 Common Challenges Slowing Down Your Data Science Programs 7

How a Unified Data Foundation Elevates Data Science Outcomes Across the Model Lifecycle 9

From Trusted Data Science to Intelligent Action 12

As Algorithms Evolve and Models Improve, Trusted Data Makes the Difference 14

About Us 15

From Insight Team to Innovation Engine: How Data Science Powers the AI Enterprise

AI has moved from the periphery to center stage in the boardroom, placing unprecedented pressure on data science teams to operationalize reliable models at scale. In 2025, CEOs ranked enhanced decision-making and data analysis among the top three benefits expected from AI investments.[1] And Gartner® predicts that by 2027, 50% of business decisions will be augmented or automated by AI agents for decision intelligence.[2]

Data science is no longer a sophisticated but siloed function. Its role continues to evolve and has now become firmly entrenched in enterprise strategy (as shown in Figure 1). The mandate is clear and flows directly to the teams responsible for building the models that power modern decision-making.

However, ambition and outcomes are two different things. Despite record executive interest – with 99% of CXOs citing data and AI as a top organizational priority[3] – 60% of companies still struggle to scale AI or achieve material value from their investment.[4] This gap is not driven by a shortage of computing capabilities, algorithmic sophistication or boardroom buy-in.

It comes down to something more fundamental: the data underneath the models.

Trustworthy AI doesn't begin with the sophistication of the model; it starts with the data feeding it. Clean, governed and integrated data providing essential, trusted context is the defining line between models that deliver reliable, explainable intelligence and those that drift, hallucinate and quietly erode confidence in AI across the organization.

Currently, most enterprises are asking their data science programs to perform at a level their data foundation was never built to support. Reliable AI at scale cannot run on infrastructure that was assembled incrementally, built around point solutions and never designed for the scale of enterprise-wide operation.

This fundamental mismatch is where most enterprise AI programs stall. For data engineers, analytics architects and data science leaders tasked with delivering faster, more accurate predictive intelligence than ever before, this eBook provides three critical perspectives:

1. **Why operationalizing data science at scale remains so difficult,** and the specific friction points your peers navigate daily
2. **How future-ready organizations are leveraging trusted context** to take data science and machine learning (ML) from isolated experimentation into a genuine competitive advantage
3. **What you can do next** to build the unified data foundation your models need to deliver on enterprise AI ambitions at scale

[1] KPMG, KPMG US CEO Outlook 2025.

[2] Gartner Press Release, "Gartner Announces the Top Data & Analytics Predictions," June 17, 2025. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved

[3] Data and Leadership Exchange, 2026 AI & Data Leadership Executive Benchmark Survey.

[4] BCG, The Widening AI Value Gap, 2025.

From Insight Team to Innovation Engine: How Data Science Powers the AI Enterprise (continued)

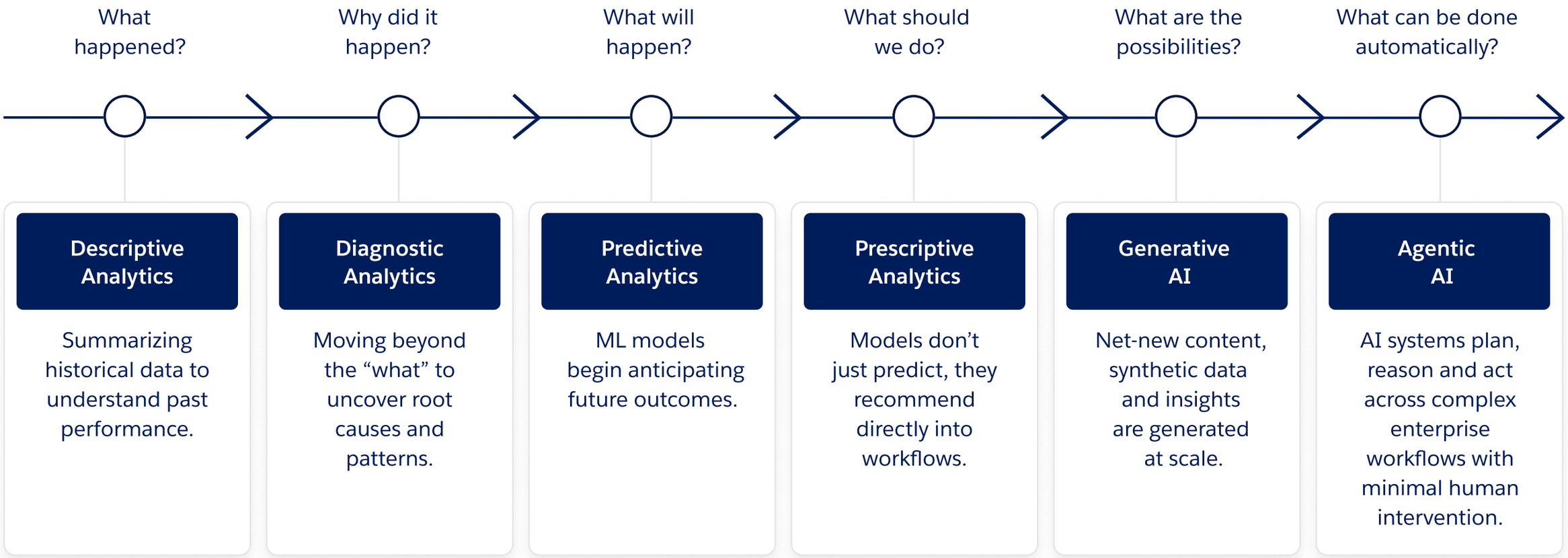


Figure 1. From hindsight to autonomous intelligence, how data science has evolved into enterprise strategy

Why Data Science Only Delivers Business Value If the Data Is Ready

Data readiness has evolved from a technical requirement into a critical business metric. 61% of CDOs cite better data, including higher data quality and completeness, as a factor making it easier to successfully transition more of their GenAI pilots into production.[5]

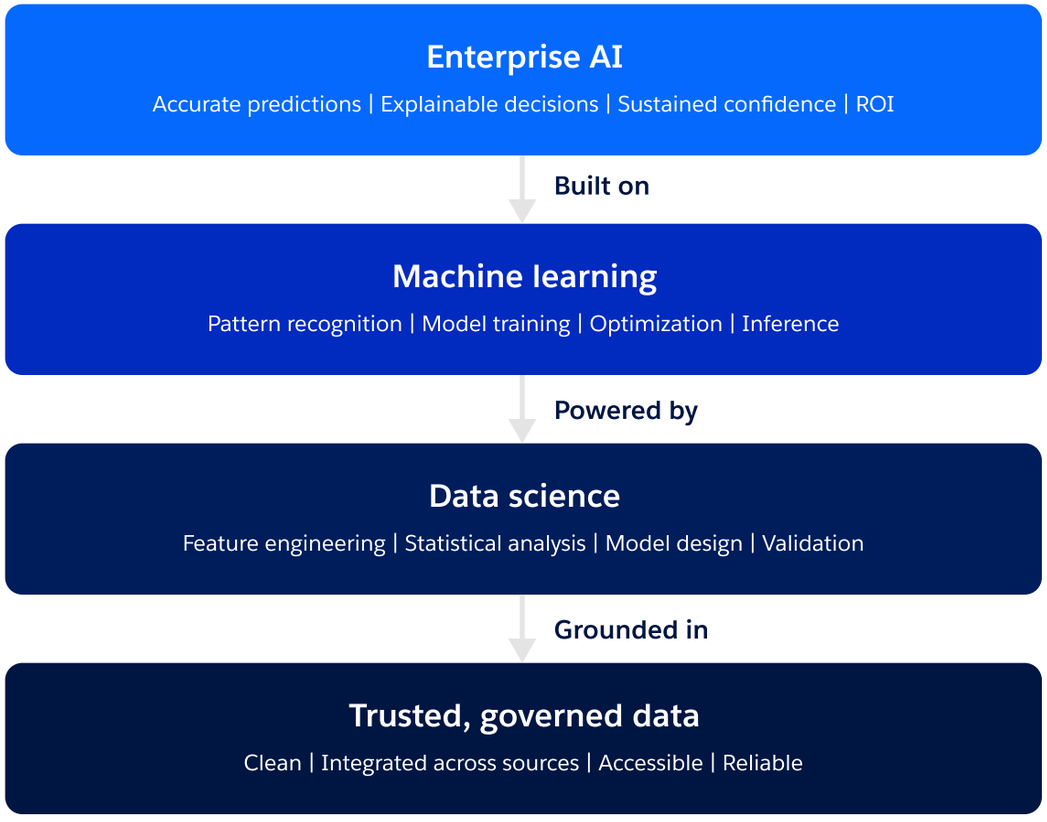


Figure 2. How a trusted data foundation powers enterprise AI

Yet, significant hurdles remain. Among companies adopting or planning to adopt agentic AI, 50% of data leaders still identify data quality and retrieval concerns as a top challenge in moving AI agents into production.[6]

The conclusion is clear: without a foundation of trusted, governed data, operationalizing data science and ML at scale remains an insurmountable task.

The Hidden Cost of the Unreliable Data Foundation

Data science teams are in a race to shrink the data-to-insight lifecycle. Retailers, for example, must mine streams of customer behavior, product data and transactional history to personalize experiences in real time. This urgency is reflected globally: 'real-time data' has emerged as the second highest data priority for enterprises, jumping seven places in just two years.[7]

However, in the context of AI, speed without trust is a liability. When models draw on unreliable data, the consequences extend far beyond a flawed dashboard. They compound into erroneous predictions, misguided decisions and potential compliance gaps – ultimately eroding confidence in the very AI programs that took years of investment to build.

[5] Informatica, CDO Insights 2026.
[6] Ibid.
[7] Salesforce, State of Data and Analytics, 2025.

Why Data Science Only Delivers Business Value If the Data Is Ready (continued)

Industry feedback validates this reality. 90% of data leaders are concerned about new AI pilots moving forward without having addressed previously uncovered reliability issues, while 57% identify data reliability as a top barrier to moving AI projects from pilots to production.[8]

The consequences are tangible and widespread. 49% of data and analytics leaders say their companies occasionally or frequently draw incorrect conclusions from data that misses or misunderstands business context. And 55% of data and analytics leaders at companies training or fine-tuning their own models say they have wasted significant resources doing so with bad data.[9]

This reflects a structural problem sitting squarely at the intersection of data science ambition and data foundation reality. The question is no longer whether poor data quality affects your models. It is how much it is costing you in wasted time, drained resources and the hidden opportunity costs of misguided decisions.

The Cost of Poor Quality, Fragmented Data Across Industries

Financial Services

- Fraud detection models depend on continuous, real-time feeds spanning transactional systems, behavioral signals and external risk data.
- A model trained on stale, incomplete or poorly governed data doesn't simply underperform; it misses fraud in motion. Even worse, it may flag legitimate transactions and damage customer trust.

Healthcare and Pharmaceutical

- Feeding unreliable or outdated data into models can cause misleading or erroneous results in clinical trials.
- Degraded model accuracy can directly affect patient outcomes and expose organizations to significant regulatory risk.

Retail

- Demand forecasting and dynamic pricing engines require clean, unified signals across supply chain, point-of-sale and customer behavior data, often in real time.
- Fragmented, inaccurate or untrustworthy data can erode margins and break down personalization exactly when the customer expects it most.

[8] Informatica, CDO Insights 2026.
[9] Salesforce, State of Data and Analytics, 2025.

6 Common Challenges Slowing Down Your Data Science Programs

The root cause of most enterprise data science challenges isn't just a technology gap, but instead legacy "point solution" thinking. Years of fragmented investment have left organizations with data silos scattered across dozens of disconnected systems, each governed by its own standards, formats, processes and governance models.

These legacy tools, which were never designed with AI in mind, are now being forced to serve as the foundation for programs that demand consistency, quality and trust at scale. This misalignment creates friction that only compounds as AI ambitions grow. In the US, CEOs cited data readiness – defined as clean, complete, contextual and continuous data – as the top challenge to successful AI implementation.[10] Such readiness can never be achieved through a fragmented toolset.

1. Data Fragmentation Compromises Model Development

Enterprise data is vast, fragmented and ever-expanding. It resides distributed across dozens of systems in formats ranging from structured records to unstructured text, sensor feeds and images. When these source systems remain isolated, training data reflects those gaps – and the models learn them, too.

A model trained on an incomplete picture of your enterprise will inevitably produce incomplete, biased or out-of-context intelligence. Comprehensive data, integrated across a diverse array of sources and formats, is the non-negotiable prerequisite for intelligence you can act on with confidence, free of systemic bias and hallucinations.

2. Ingestion and Integration Bottlenecks Steal Engineering Time

Bridging the gap between raw data diversity and model-ready pipelines remains a persistent bottleneck in data science. Most data scientists spend more time wrestling data into shape than actually performing analysis. 95% of IT leaders say integration issues are actively impeding AI adoption.[11] Every new data source added without a coherent strategy only exacerbates the problem.

The result is a compounding "data preparation tax": countless hours lost to manual wrangling, pipelines stitched together under pressure and the inevitable security and quality gaps that follow.

3. Inconsistent Data Quality Amplifies Errors

In a fragmented ecosystem, every system operates under its own isolated quality standards. A customer record missing a postal code might be acceptable for a sales system, but it renders a logistics model unusable. Multiply that discrepancy across hundreds of fields and millions of records, and the errors compound quickly.

Worse, data sources that appear clean today can degrade quietly as pipelines shift or new sources are added. Without continuous **observability**, teams often discover quality issues only after they have corrupted model outputs, at which point the damage is difficult to trace and even harder to undo.

[10] KPMG, KPMG US CEO Outlook 2025.

[11] Salesforce, State of Data and Analytics, 2025.

6 Common Challenges Slowing Down Your Data Science Programs (continued)

4. Gaps in Data Governance, Privacy and Access Accelerate Risks

More than three-quarters (76%) of data leaders say their governance has not kept pace with employee use of AI technology,[12] increasing exposure to vulnerabilities around privacy, security and ethical use. Operationalizing data science requires embedding robust **governance** directly into data pipelines from the outset, rather than attempting to "bolt it on" after problems surface.

When thousands of employees are expected to act on AI-driven insights in their daily work, the underlying data must be more than just technically accurate; it needs to be safe. Unless data is demonstrably trustworthy, governed, contextualized and compliant, enterprise-wide adoption stalls.

5. Without Lineage and Explainability, You Can't Diagnose Failure

The stakes are high: 89% of data and analytics leaders have encountered inaccurate or misleading AI outputs.[13] In these instances, lineage and explainability are the only tools that can give you a clear diagnosis.

- **Lineage** tracks the data's "biography": its origin, transformations and relationships.
- **Explainability** reveals the "logic": how the model used that data to reach a conclusion.

Without both, diagnosing and correcting anomalous outputs becomes an expensive guessing game.

6. Siloed Teams Duplicate Efforts and Fragment the Truth

When "customer" carries a different definition across your customer relationship management (CRM), enterprise resource planning (ERP) and supply chain systems, every team has a different version of reality. Pipelines are duplicated, datasets are rebuilt from scratch and outputs cannot be reconciled across the organization.

The result is a triple threat: wasted engineering capacity, slower delivery speeds and AI initiatives that never scale beyond isolated pockets of the business.

[12] Informatica, CDO Insights 2026.

[13] Salesforce, State of Data and Analytics, 2025.

How a Unified Data Foundation Elevates Data Science Outcomes Across the Model Lifecycle

The case for complete, connected and traceable data goes beyond being a risk mitigation strategy; it is a performance mandate. Over the next 12 months, more than a quarter of CDOs (29%) will leverage AI to enhance customer experience and loyalty (29%), while a similar percentage (28%) are looking at AI initiatives to improve business intelligence, analytics and decision making.[14]

Organizations prioritizing the data quality, governance and accessibility powering their data science can answer fundamental questions about data comprehensiveness, accuracy, lineage and governance mandates with confidence. This trust is the defining line between programs that achieve enterprise scale and those that stall.

A unified, AI-powered data foundation transforms the entire model lifecycle, driving superior AI outcomes from development to deployment.

- **Accelerated model development:** Teams move faster when they can discover and trust datasets instantly, eliminating weeks of manual source validation.
- **Enhanced model training:** High-quality, consistent data provides the "clean" inputs necessary for models to learn accurately. At scale, this consistency ensures models identify meaningful patterns rather than learning from noise or errors, turning raw data into genuine predictive power.
- **Manageable model deployment:** Integrated governance and data lineage allow teams to catch issues early, respond rapidly, and maintain accountability.

As a result, data science moves beyond isolated experiments into trustworthy, operationalized AI that the business can scale with confidence.

Faster Model Development

Model development stalls when data is difficult to find, understand or trust. A unified platform directly addresses all three of these issues with:

- **Streamlined data integration:** Automated **extract, transform, load (ETL)** tools simplify ingesting and transforming data across sources. By fast-tracking the journey from raw data to model-ready inputs, you can reduce the custom engineering burden on overstretched teams and eliminate a primary drag on timelines.
- **Centralized discovery:** A single, searchable catalog transforms siloed data into a shared resource. Instead of spending weeks tracking down owners, teams can instantly locate and access datasets from across the organization, shifting the focus from searching to building.
- **Contextual clarity:** Rich metadata and business context allow data scientists to understand what a dataset means, its history and its relationships. Identifying the right inputs for a model becomes a matter of hours rather than days.
- **End-to-end lineage:** Full visibility into data origins and transformations slashes investigative overhead. Teams no longer waste time manually tracing pipelines to understand why a dataset looks the way it does.

In practice: A team building a "next-best-offer" engine no longer wastes time chasing multiple owners for customer, transaction and inventory data. With a unified catalog, these datasets are instantly discoverable and contextually understood, accelerating development from day one.

[14] Salesforce, State of Data and Analytics, 2025.

How a Unified Data Foundation Elevates Data Science Outcomes Across the Model Lifecycle (continued)

Model Training at Scale

Reliable data science and ML models require large, clean and diverse datasets. Each of those requirements places distinct demands on the underlying data foundation, including:

1. **Depth and breadth of training data:** Volume alone is insufficient; models need quality to avoid learning the wrong patterns and diversity to handle real-world variation. Integrating representative data from broad sources prevents bias and "silent failures" in production, ensuring models remain robust and accurate.
2. **Systematic validation:** A governed foundation replaces manual checks with continuous, automated profiling and monitoring. By standardizing data as it moves through the pipeline, datasets become trustworthy by default, eliminating the days of investigation typically required before every training run.
3. **Reusability of data assets to train models:** Certified, vetted, purpose-built data products are easily accessible to train models. This reduces the time needed to start training models and encourages cross-team collaboration during the training phases.

In practice: In a healthcare setting, training a patient deterioration model requires years of standardized records from diverse populations. Gaps — like missing codes or underrepresented groups — directly degrade predictive accuracy in critical moments. A governed data foundation identifies and fixes these inconsistencies before they reach the model for clinical reliability.

How a Unified Data Foundation Elevates Data Science Outcomes Across the Model Lifecycle (continued)

Model Deployment and Monitoring in Production

Deployment is not the finish line; it marks the transition from development to long-term governance. Because models are living systems exposed to shifting real-world data, a robust foundation is required to maintain performance and accountability. This includes:

- **Observability for continuous monitoring:** To prevent performance degradation, continuous monitoring must be baked into your infrastructure. Automated observability detects model drift and emerging bias by flagging when production data diverges from training patterns, ensuring reliability without relying on manual reviews.
- **Guardrails for data use:** A unified foundation enforces privacy, protection and access controls for data at rest and in motion. This is critical for agentic AI. Because autonomous agents act faster than human review, guardrails must be embedded in the data layer to strictly define what an agent can see, do and decide.
- **Continuous audit-readiness:** In regulated industries, the obligation to explain model outputs never ends. A unified foundation ensures every decision remains traceable and defensible, providing the lineage necessary to satisfy compliance audits long after a model goes live.

In practice: In financial services and the dynamic world of fraud detection, shifting transaction patterns can quickly turn a static model into a liability. Without continuous visibility into data quality and drift, recalibration is a slow, reactive response to failure. A governed foundation enables proactive adjustments, ensuring the model remains accurate, auditable and trustworthy even as fraud vectors evolve.

From Trusted Data Science to Intelligent Action

The ultimate test of data science is its ability to guide enterprise decisions at scale. By anchoring models and agents in a governed foundation, organizations transform raw information into “decision-grade” fuel.

Without a unified foundation, every stage of the lifecycle becomes slower, riskier and more expensive. Fixing a failing model in production requires retracing the entire pipeline; if integration, governance and lineage aren’t baked in from the start, this process becomes a costly manual burden rather than a routine update.

Informatica Intelligent Data Management Cloud™ (IDMC) is a trusted cloud-native, AI-powered platform that turns fragmented enterprise data into context, so your agents can reason, act and deliver outcomes with precision.

Its rich data catalog, integration, governance, quality and privacy, metadata management and **master data management** (MDM) services help establish a unified and comprehensive data foundation.

Integrated with the **Salesforce Agentforce 360 platform**, businesses can drive innovation, efficiency and growth by leveraging a single platform that unites applications, trusted data and AI agents.

Combined with Tableau’s intuitive AI-powered visualizations, Informatica’s enterprise-grade data management empowers teams to move from raw data to refined data science insights at scale. Models fueled by compliant, high-quality enterprise data and high-integrity context allow data scientists to spend less time on data cleansing and more time delivering trusted, actionable outcomes. As this intelligence reaches the experience layer, it helps ensure your AI interactions are reliable and compliant.

Together, Informatica and Salesforce ecosystem capabilities make trusted data science the central engine of the agentic enterprise, eliminating the “trust gap” and empowering autonomous agents to act as precise, data-driven partners across every business channel.



From Trusted Data Science to Intelligent Action (continued)

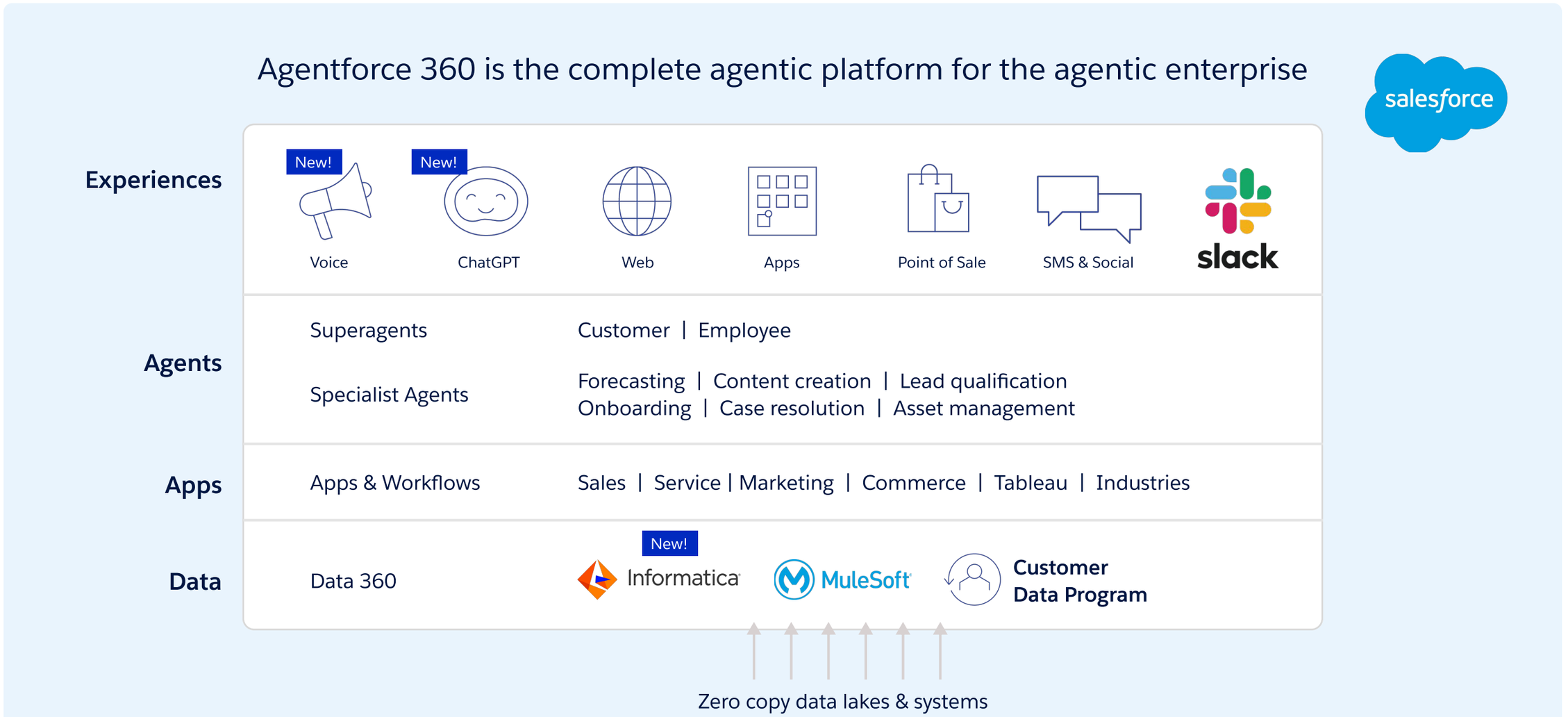


Figure 3: The architecture of the agentic enterprise.

As Algorithms Evolve and Models Improve, Trusted Data Makes the Difference

Data science will continue its rapid, relentless advancement. The models available today are more capable than anything that existed five years ago, and the models five years from now will make the current models look rudimentary.

But effective data science will always depend on the quality, trustworthiness and governance of the data supporting it. Sophisticated models trained on fragmented, ungoverned data will not produce sophisticated outcomes. No wonder 89% of data and analytics leaders believe a strong data foundation is the most critical factor for successful AI deployment.[15]

The organizations that have invested in trusted, contextual data know it unlocks:

- Faster model development
- More reliable training to mitigate drift and bias
- Production deployments that hold up under real-world conditions
- The infrastructure needed to power agentic AI systems that can plan, reason and act autonomously, at enterprise scale

Ready to start your journey, too? Discover how your data science programs can excel when your foundation is no longer a bottleneck. **Start building your business case for advanced data management and analytics here.**

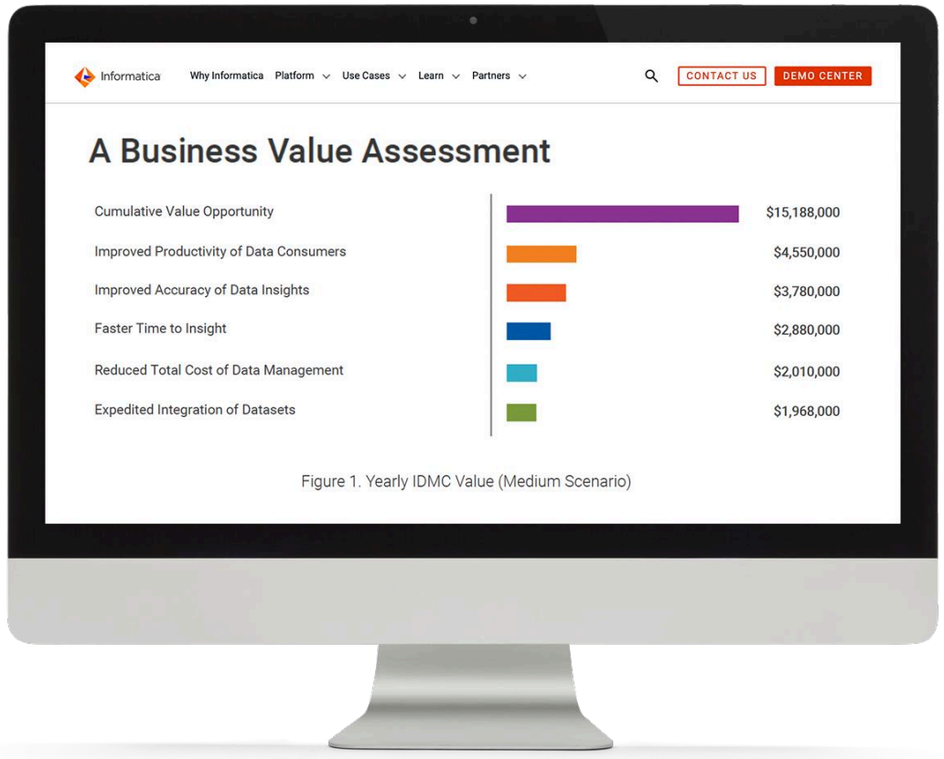


Figure 1. Yearly IDMC Value (Medium Scenario)

[15] Salesforce, State of Data and Analytics, 2025.

About Us

About Informatica

Informatica from Salesforce is a leader in AI-powered enterprise cloud data management. Its Intelligent Data Management Cloud (IDMC) platform enables organizations to connect, manage and unify AI-ready data across the enterprise. With capabilities spanning data cataloging, integration, governance, quality, privacy, metadata management and master data management, Informatica supports a broad partner ecosystem and helps customers unlock the full value of their data and AI initiatives.

About Salesforce

Salesforce is the #1 AI CRM, empowering companies to connect with their customers in a whole new way through the power of artificial intelligence, data, and trust. For more information about Salesforce (NYSE: CRM), visit: www.salesforce.com.

Worldwide Headquarters
2100 Seaport Blvd.
Redwood City, CA 94063, USA
Phone: 650.385.5000
Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871

informatica.com
linkedin.com/company/informatica
x.com/Informatica

Contact us



The information provided in this report is strictly for the convenience of our customers and is for general informational purposes only. Publication by Salesforce, Inc. does not constitute an endorsement. Salesforce.com does not warrant the accuracy or completeness of any information, text, graphics, links, or other items contained within this guide. Salesforce.com does not guarantee you will achieve any specific results if you follow any advice in the report. It may be advisable for you to consult with a professional such as a lawyer, accountant, architect, business advisor, or professional engineer to get specific advice that applies to your specific situation.

© Copyright 2026, Salesforce, Inc. All rights reserved.

IN19-5344-0926