2024 Planning Guide for Analytics and Artificial Intelligence

Published 4 October 2023 - ID G00796447 - 40 min read

By Analyst(s): Sumit Agarwal, Joe Antelmi, Christopher Long, Georgia O'Callaghan, Wilco van Ginkel, Zain Khan, Maryam Hassanlou

Initiatives: Analytics and Artificial Intelligence for Technical Professionals; Evolve Technology and Process Capabilities to Support D&A

Organizations will require a balance between establishing a solid GenAI foundation and scaling existing analytics and AI solutions. To deliver, data and analytics technical professionals must develop new skills, define a holistic architecture roadmap, enforce governance and enable trust.

Overview

Key Findings

- Generative AI (GenAI) tools, specifically those based on large language models (LLMs), are enabling enterprises to develop innovative analytics and AI solutions. However, the ever-expanding hype, risk and technology ecosystem are challenging identification and delivery of appropriate use cases, technology products and solutions.

- Although self-service analytics programs and AI model development processes are maturing, enterprises still face several obstacles to success. These include a lack of data literacy, the challenge of scaling technology and processes, and the persistent need for governance, privacy and trust.

- The growing diversity of data, platforms and deployment methods is creating traceability, reproducibility and consistency challenges with metrics, features, datasets and models.

- Democratization of data gives both technical and subject matter experts the opportunity to discover unique insights. However, successful adoption of this organizational model requires upskilling, training investments and clear expectations for the new roles.

Recommendations

Data and analytics technical professionals working to deliver analytics and AI initiatives must:

- Experiment, evaluate and adapt LLM-based solutions by conducting pilots validated by technical and business stakeholders alike. Based on these pilots, design an agile, extensible architecture to plug in new data sources, LLMs and applications.
- Drive the success of self-service and democratization initiatives by implementing data literacy programs that enable users to derive meaningful insights from data.
- Enhance the experience of AI and analytics developers by offering a variety of augmented analytics and AI platform options for different user personas. Simultaneously, provide a unified view of features, metrics and models by implementing semantic layers, feature stores and model management.
- Combine responsible AI practices across the entire advanced analytics and ML development cycle. Leverage explainability techniques and frameworks to interpret model outputs, and ensure that they can collectively address fairness, bias mitigation, ethics, risk management, privacy and regulatory compliance.

Analytics and Artificial Intelligence Trends

2023 has seen explosive interest in generative AI solutions, with LLMs garnering massive attention. While the technology is not new, the emergence of LLMs has revitalized the AI landscape. In the 2023 Gartner Voice of the Client Content Survey for Generative AI, we asked respondents about their GenAI use. More than two-thirds of the Gartner technology and business leader clients surveyed are either already using GenAI or expected to use GenAI solutions within the next three months. [1] 2024 will see a shift in focus from models to solutions and adoption.

While the enterprise spotlight is expected to be on defining and implementing GenAI use cases during 2024, data-driven decisions are still a key priority. The 2024 Gartner CIO and Technology Executive Survey highlighted increasing investments in data and analytics (D&A) and AI. [2] It asked respondents which technology areas would be receiving the largest amount of new or additional funding in 2024 compared with 2023:

- 78% selected business intelligence (BI) and data analytics
- 73% selected AI and machine learning (ML)

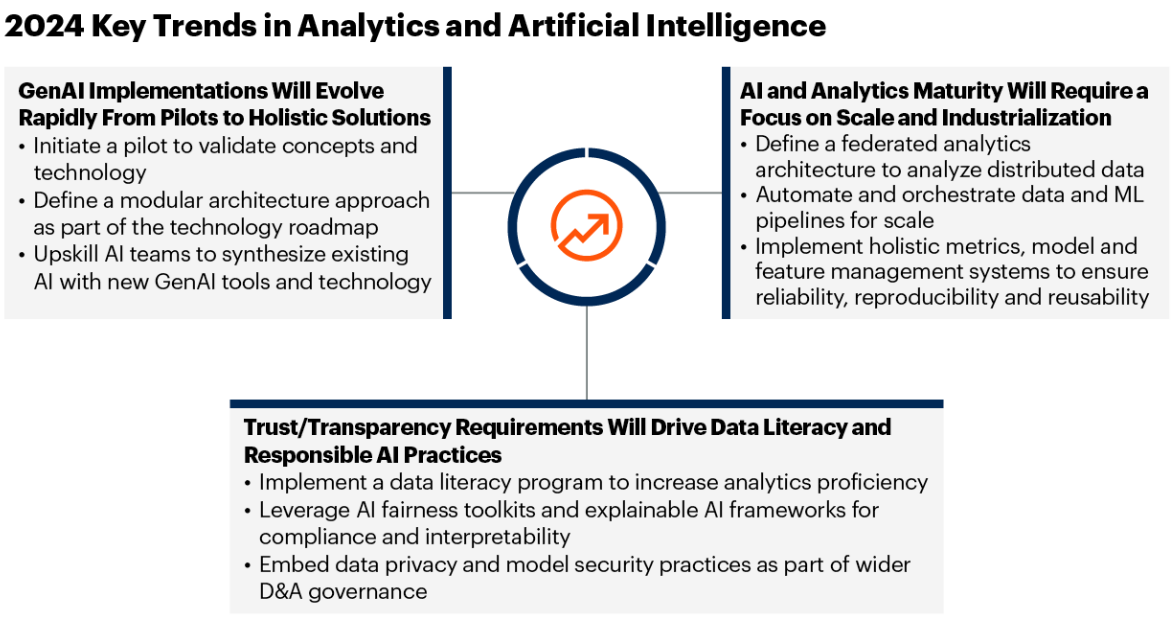
Analytics and AI are heavy data consumers. They provide a spectrum of diagnostic and predictive capabilities that help enterprises deliver personalized solutions to their customers, identify opportunities for improved efficiency within their operations, and launch differentiated products to increase revenue. They directly impact an organization's bottom line and top line. Now more than ever before, enterprises need to:

- Evaluate and assess new transformative technologies, such as GenAI. Define an initiative or a framework, such as an innovation team tasked with identifying and responding to new technology trends.
- Incrementally scale existing practices, such as self-service analytics implementation and traditional AI model development, to reflect the maturity of technology and team capabilities.
- Implement governance to ensure controls, regulatory compliance and best-practice implementation.
- Invest in upskilling teams and in augmenting tasks with available technology.
- Increase agility through a combination of automation and methodology.
- Establish guidelines and frameworks to develop standards for consistent implementations.
- Leverage all of the above to ensure business trust in the developed solutions.

As shown in Figure 1, these objectives align with key trends that organizations must adopt during 2024 (click links below to jump to trends):

- Generative AI implementations will evolve rapidly from short-term pilots to holistic enterprise solutions.
- AI and analytics maturity will require a focus on scale and industrialization.
- Trust and transparency requirements will drive data literacy and responsible AI practices.

Figure 1: 2024 Key Trends in Analytics and Artificial Intelligence



2024 Key Trends in Analytics and Artificial Intelligence

GenAI Implementations Will Evolve Rapidly From Pilots to Holistic Solutions

- Initiate a pilot to validate concepts and technology
- Define a modular architecture approach as part of the technology roadmap
- Upskill AI teams to synthesize existing AI with new GenAI tools and technology

AI and Analytics Maturity Will Require a Focus on Scale and Industrialization

- Define a federated analytics architecture to analyze distributed data
- Automate and orchestrate data and ML pipelines for scale
- Implement holistic metrics, model and feature management systems to ensure reliability, reproducibility and reusability

Trust/Transparency Requirements Will Drive Data Literacy and Responsible AI Practices

- Implement a data literacy program to increase analytics proficiency
- Leverage AI fairness toolkits and explainable AI frameworks for compliance and interpretability
- Embed data privacy and model security practices as part of wider D&A governance

Source: Gartner
796447_C

Generative AI Implementations Will Evolve Rapidly From Short-Term Pilots to Holistic Enterprise Solutions

Back to top

Generative AI has quickly emerged as a transformative technology with rapid innovation. Its timeline includes:

- The publication of a foundational paper on transformers in 2017 [3]
- The release of BERT in 2018 [4]
- The release of GPT-3 in 2021 [5]
- A paper on reinforcement learning with human feedback in 2022 [6]
- The release of ChatGPT at the end of 2022

The development of several open-source and proprietary LLMs, along with the introduction of integration components, has made solution implementations more accessible and more feasible. Several organizations are planning to implement generative AI. Surveys conducted during Gartner's well-attended Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI webinar found that: [7]

- 45% of the responding organizations have seen an increase in AI investment
- 68% of the responding executives believe that GenAI's benefits outweigh its risks
- 19% of the responding organizations are involved in solution implementations

Data and analytics technical professionals should take the following steps to get started with GenAI solution implementation:

- Initiate a pilot to validate concepts and technology.
- Define a modular architecture approach as part of the technology roadmap.
- Upskill AI teams to synthesize existing AI implementations with new GenAI tools and technology.

Planning Considerations

Initiate a Pilot to Validate Concepts and Technology

Enterprises have consistently explored and adopted new technologies as new trends emerge. Businesses and technology teams have gone through changes as they've:

- Adopted self-service analytics
- Evolved from descriptive and diagnostic analytics to predictive analytics by developing and integrating AI and ML models
- Implemented MLOps and DataOps techniques to increase the agility of solution development

A "crawl-walk-run" approach has enabled organizations to:

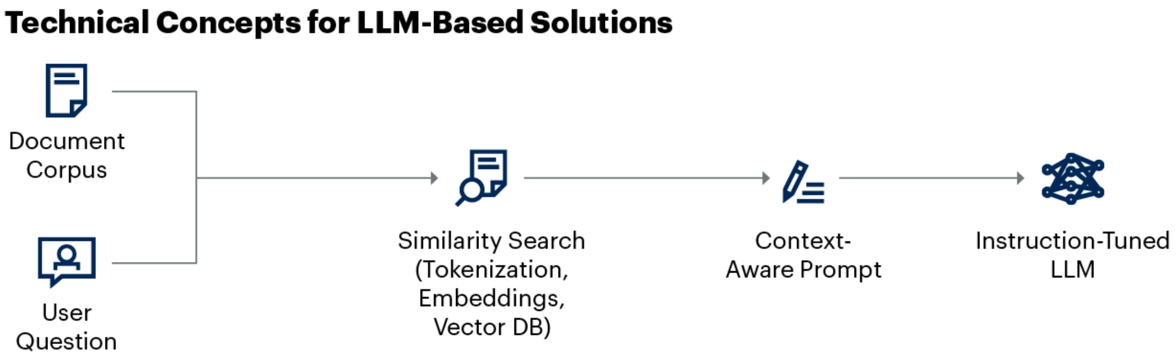
- Start small
- Incrementally improve and expand
- Deliver business value and scale

A GenAI solution implementation requires a similar approach. At the same time, GenAI has specific nuances and new technical concepts. Enterprises should start with a pilot to demonstrate a solution to various internal stakeholders. The early pilot should serve to validate the various technical concepts. The following steps outline the approach for a pilot:

1. Select an internal-facing use case.
2. Validate the various concepts and technology.
3. Select a technical ecosystem that is easily accessible.
4. Evaluate the solution outcomes.
5. Assess challenges and gaps.

As an example, Figure 2 lists the various technical concepts for an LLM solution implementation.

Figure 2: Technical Concepts for LLM-Based Solutions


Technical Concepts for LLM-Based Solutions
Document Corpus
User Question
Similarity Search (Tokenization, Embeddings, Vector DB)
Context-Aware Prompt
Instruction-Tuned LLM

Source: Gartner
796447_C

Some of the key concepts included in the Figure 2 are:

- **Tokenization:** Whereas languages use text (such as characters, words, and sentences), LLMs work with numbers. These numbers are called "tokens." For an LLM to work with tokens, conversion between text and tokens is necessary. This conversation is done by a tokenizer, which converts (maps) characters, subwords or words to predefined tokens. The tokens are entries in a token vocabulary. For instance, the English language has approximately 600,000 words. A token vocabulary is significantly smaller and contains only tokens that represent commonly used words. As an example, the token vocabulary used by OpenAI's GPT-3 has only around 50,000 tokens.

- **Embeddings:** Embeddings map the tokens into a vector space with semantic relationships across words. An embedding itself is a list of numbers for all the dimensions the embedding represents. This list is also called a "vector." The dimension space can be different, such as 300 or 768 dimensions, and depends on the embedding algorithm used.

- **Vector database:** Vector databases are specially designed to store and manage high-dimensional vector data. Enterprises need a vector database to store, manage and retrieve vector embeddings.

- **Context-aware prompts:** A prompt is simply the instructions provided to an LLM when a user instigates an interaction or makes a request. A prompt follows a certain pattern, ranging from a simple question to a large set of instructions, including examples.

- **Instruction-tuned LLM:** LLMs represent massive amounts of text data. The base model may generate an output that is biased, irrelevant or incorrect. Instruction-tuned models are refined using human feedback in a supervised learning mode to score the responses. Such models have performed better on various model quality benchmarks.

Enterprises should use the pilot implementation to:

- Demonstrate the capabilities to various technical and business stakeholders

- Define a roadmap for prioritization and implementation of new use cases

- Identify challenges related to technology, risk, compliance, privacy and security

- Define reference architectures for the various implementation patterns

- Define the roles and skills required for the team

- Integrate all of the above with their overall AI strategy

For more information, see:

- What Technical Professionals Need to Know About Large Language Models

Define a Modular Architecture Approach as Part of the Technology Roadmap

The early GenAI solution implementations revolve around a specific LLM and focus primarily on text. The solutions are limited to a specific cloud ecosystem, and integrate several cloud managed services or open-source components. Organizations across all industry verticals and functional teams are assessing use cases that would help them identify innovative insights, develop new products or improve productivity. These emerging use cases may be grouped into the following categories:

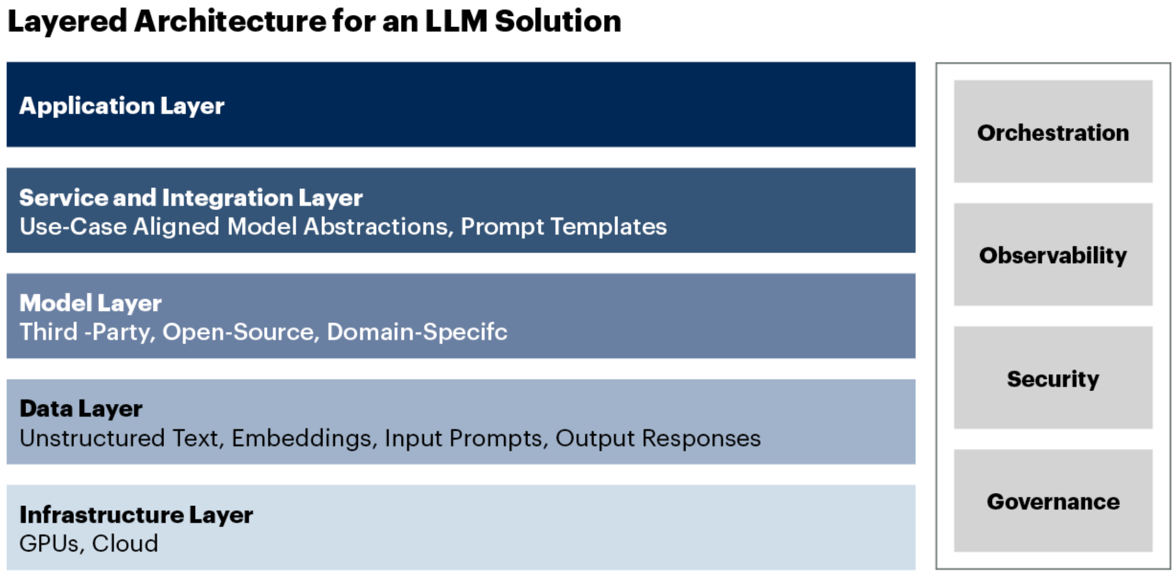
- Text summarization
- Text generation (including code)
- Sentiment analysis
- Classification
- Conversational questions and answers

The use cases have varying model accuracy requirements. For example, a sentiment model doesn't need the same predictive capabilities as a text generation model. By contrast, a text generation or text summarization model may need to be trained or customized with specific subject matter understanding.

LLM providers have developed models with different sizes and, consequently, varying accuracy to match with different use cases. However, as use-case implementations expand from technology centricity to business centricity, LLMs need more customization and specialization for the specific requirements.

As GenAI implementations mature within organizations, the requirements will expand to include a mix of general-purpose and specialized models. These models may be internally developed or sourced from different vendors of various sizes and capabilities. A holistic and modular architecture would enable a flexible framework. Figure 3 provides a layered-architecture view.

Figure 3: Layered Architecture for an LLM Solution


Layered Architecture for an LLM Solution
Application Layer
Service and Integration Layer
Use-Case Aligned Model Abstractions, Prompt Templates
Model Layer
Third -Party, Open-Source, Domain-Specifc
Data Layer
Unstructured Text, Embeddings, Input Prompts, Output Responses
Infrastructure Layer
GPUs, Cloud
Orchestration
Observability
Security
Governance

Source: Gartner
796447_C

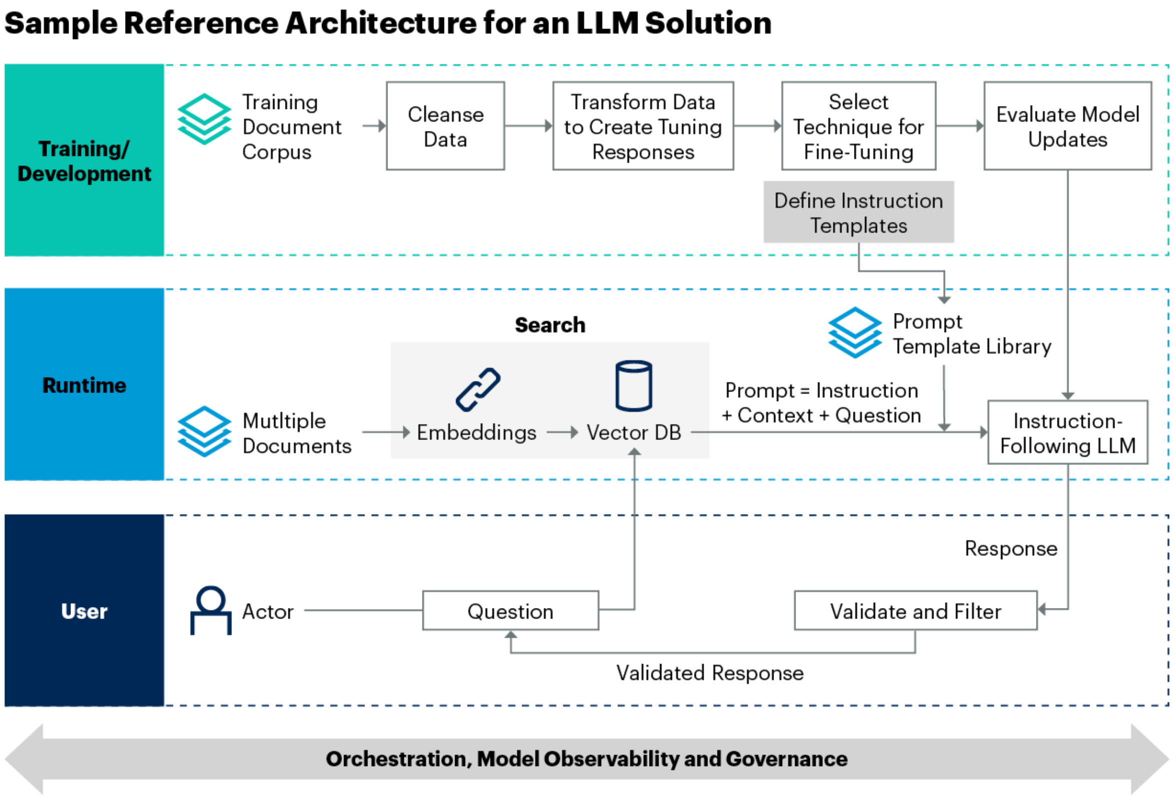
As they've evaluated ChatGPT and considered GenAI implementations, Gartner clients have shared several concerns with us. They've consistently had questions related to:

- **Data and content privacy:** Specifically, organizations are concerned about model developers using their proprietary content. The content may also include personally identifiable information (PII) about their customers or employees. Such information would need to be parsed and masked before the model could be used in production.
- **Hallucinations and outdated source data:** These create concerns related to reliability and accuracy. For example, most LLMs are autoregressive and based on transformer models. They are designed to predict the next word and, by extension, the next sentence. This approach results in very creative responses but, at times, also results in incorrect or impossible-to-validate responses, aka hallucinations. Moreover, LLMs are trained on a corpus of data from a specific point in time and, thus, lack up-to-date context. Therefore, it is necessary to look at solution patterns that provide grounded, accurate responses.
- **Access controls and potential misuse:** Access controls are required to ensure that only approved user groups have access to relevant responses. For example, a customized LLM-based solution should not provide HR-specific information to a supply chain user. Also, organizations want to ensure that the interactions are constrained to business-relevant analysis.

These challenges are spurring development of guardrails for LLMs, but such guardrails are still evolving.

Figure 4 shows a reference architecture that combines retrieval-augmented generation (RAG) with model fine-tuning. The RAG pattern combines a user prompt (the "user" swimlane in Figure 4) with the context derived from a knowledge base or a document corpus (the "operational" swimlane in Figure 4). This pattern minimizes hallucinations and provides a framework to apply access controls and content privacy. The "training" swimlane in the figure provides the steps to fine-tune a model with domain-specific content.

Figure 4: Sample Reference Architecture for an LLM Solution


Sample Reference Architecture for an LLM Solution
Training/ Development
Training Document Corpus
Cleanse Data
Transform Data to Create Tuning Responses
Select Technique for Fine-Tuning
Evaluate Model Updates
Define Instruction Templates
Runtime
Search
Prompt Template Library
Mutltiple Documents
Embeddings
Vector DB
Prompt = Instruction + Context + Question
Instruction-Following LLM
User
Response
Actor
Question
Validate and Filter
Validated Response
Orchestration, Model Observability and Governance

Source: Gartner
796447_C

AI and solution architects should use the layered architecture (Figure 3) and the reference architecture (Figure 4) to define solution blueprints and an architectural roadmap.

Upskill AI Teams to Synthesize Existing AI Implementations With New GenAI Tools and Technology

Data scientists are often at the center of the team developing AI and ML models. However, a successful technical team involves equal contribution from AI/ML engineers and data engineers. AI architects define the overall solution architecture to:

- Provision development environments
- Automate model deployment, testing, monitoring and integration into applications
- Assess technical platforms

Generative AI solutions based on LLMs have similar requirements. However, additional skills are required to handle the following:

- Additional technical components
- Scale of models and data
- Various model optimization techniques
- Use of unstructured data
- GPU dependencies and related optimization

Table 1 maps the layered architecture in Figure 3 with the various roles.

Table 1: Architecture Layers Mapped to Roles

Architecture Layer	Role
Application Layer	Application developer, UX designer
Service and Integration Layer	Software engineer
Model Layer	AI engineers, AI research engineer, data scientist
Data Layer	Data engineer, AI engineer, prompt engineer/business analyst
Infrastructure Layer	Cloud and infrastructure engineer

Source: Gartner (October 2023)

LLM solution implementations also have several additional challenges, such as data privacy, prediction reliability, IP infringement, security and transparency. These aspects require support from additional teams, including risk, compliance, security and legal, to define policies and review regulations.

For more information, see:

- Build Team Skills for LLMs and Retrieval-Augmented Generation

AI and Analytics Maturity Will Require a Focus on Scale and Industrialization

Back to top

For enterprises, the digital transformation journey includes the need to democratize and industrialize analytics and AI development and deployment. Enterprises have established KPIs based on the number of AI models or dashboards. However, the 2021 Gartner AI in Organizations Survey highlighted challenges with AI models. It found that, on average, 54% of models do not progress to production, and that it takes an average of 7.3 months to develop and deploy a model into production. [8] The 2022 Gartner AI Use-Case ROI Survey identified that surveyed organizations have deployed an average of 41 use cases, with models staying in production for 3.5 years. [9]

As their AI and analytics implementations mature, these organizations need to focus on scale and industrialization. Analytics and AI professionals, specifically technical architects, should take the following steps:

- Define a federated analytics architecture to analyze distributed data.
- Automate and orchestrate data and ML pipelines for scale.
- Implement holistic metrics, model and feature management systems to ensure reliability, reproducibility and reusability.

Planning Considerations

Define a Federated Analytics Architecture to Analyze Distributed Data

Enterprises have traditionally worked to bring all data into centralized data platforms. These centralized platforms still provide a consistent structure to data and deliver a single source of truth. However, organizations struggle with data latency, rigid structures and lack of agility to accommodate new data types, including text, images, videos and audio.

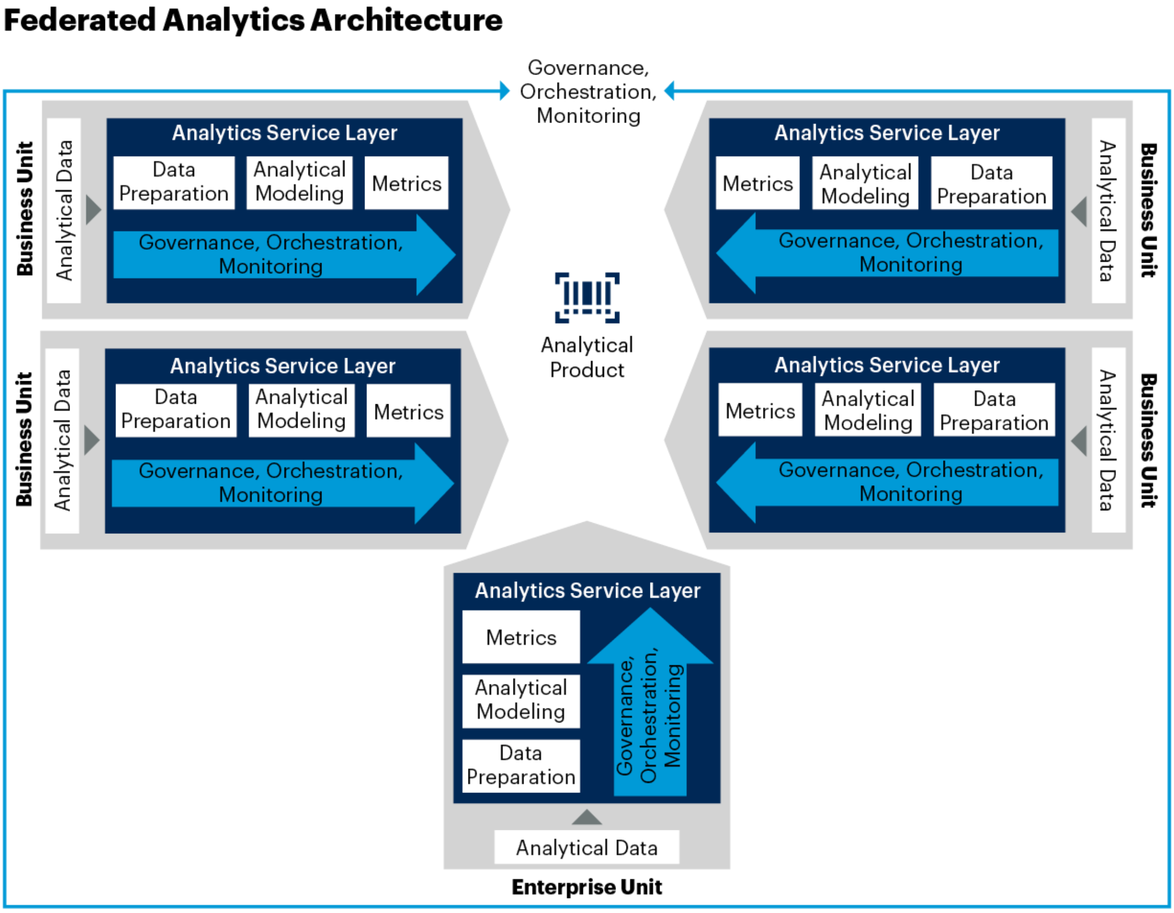
As part of a domain-centric approach to data management, data mesh is intended to enable development of distributed datasets while providing governance and reusability of data throughout the enterprise. Although this technique is helping organizations manage data across the enterprise, similar management techniques are not always applied to analytical assets. Analytics assets are often locally owned and managed without the same level of federated governance found in data mesh principles.

A federated architecture is designed to support analytics and BI. It blends both top-down and bottom-up approaches. Global guidelines are developed and enforced at a global level, but business units still have some autonomy over their local analytical assets. In short, the federated approach provides:

- A standardized scalable architecture
- Consistency in analytical data definitions
- Agility to deliver multiple end-user analytics products and services
- A strong partnership between IT and the business
- Supported and governed self-service analytics platforms

This approach centers on an analytics service layer (see Figure 5). The analytics service layer provides the ability to connect with data and to model, govern and deliver analytics products to the organization for each of the data domains.

Figure 5: Federated Analytics Architecture



Source: Gartner
786991_C

The widespread adoption of self-service analytics, coupled with the growing implementation of data mesh architectures, empowers business units to:

- Control and cultivate their domains of data and analytics
- Share resulting content as products to the wider organization

For more details, see:

- Reference Architecture for Federated Analytics
- Demystifying Semantic Layers for Self-Service Analytics

Automate and Orchestrate Data and ML Pipelines for Scale

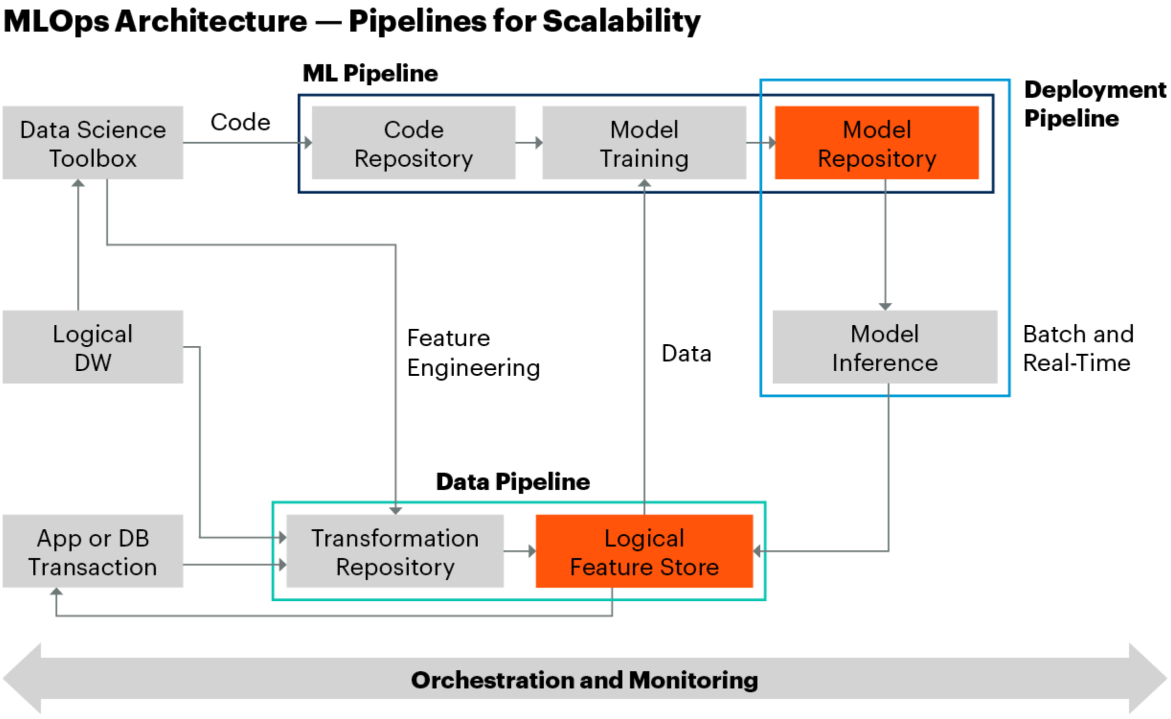
MLOps practices, along with data science and machine learning (DSML) platforms and tools, have enabled organizations to implement model quantities ranging from 10 to several thousand. At the same time, democratization of DSML platforms has allowed data scientists and citizen data scientists embedded within business teams to develop and deploy ML models. All of this expansion has resulted in enterprises leveraging multiple platforms to develop and deploy models. Analytics implementations see similar patterns, where teams have access to multiple platforms. This platform sprawl diminishes process consistency, erodes development rigor and increases technical debt.

Technical professionals, especially architects, have several implementation choices. They may leverage solutions integrated within a platform ecosystem for ease of integration. However, integrated offerings may not be sufficiently mature in all critical areas. Thus, solutions need the flexibility to fill gaps by incorporating modular, best-of-breed elements. However, as the number of implementation choices increases, a cohesive deployment of analytics and AI becomes more complicated.

Figure 6 shows an MLOps reference architecture. The architecture distinguishes between three different pipelines:

- ML pipelines
- Deployment pipelines
- Data pipelines

Figure 6: MLOps Architecture — Pipelines for Scalability



Source: Gartner
796447_C

The highlighted sections in Figure 5 ("model repository" and "logical feature store") provide the convergence or "handshake" points between the different pipelines. They enable governance and support by providing access to a consolidated enterprise view of models and features.

ML Pipelines

The architecture enables enterprises to have multiple ML model development pipelines. These pipelines may be based on different platforms or modeling technologies. For example, a business team driven by citizen data scientists may use a no-code or a low-code platform (such as AutoML or visual workflows) to develop ML models. A team of expert data scientists may prefer open-source frameworks, such as TensorFlow or PyTorch, to develop computer vision or natural language processing (NLP) models.

Deployment Pipelines

Enterprises may choose to consolidate the deployment of models for inference execution on specific targets to better manage APIs. Consolidated deployment pipelines bring consistency in security, access control and packaging. For example, all models developed in a cloud environment may be deployed to a managed Kubernetes service on a specific cloud platform. However, enterprises may require separate deployment pipelines for batch and web service deployments, or for on-premises and cloud deployments.

Data Pipelines

Development of *separate* data pipelines for every model creates an unmanageable tangle of data transformation modules. Logical feature stores provide a mechanism to consolidate transformation modules for each of the feature groups, instead of for each feature or each model. This practice also reduces the possibility of redundant and duplicate features. The feature stores may be grouped by business domains or by functional areas to create more manageable repositories. See the next section for more details on feature stores.

Implement Holistic Metrics, Model and Feature Management Systems to Ensure Reliability, Reproducibility and Reusability

The reference architectures for federated analytics and ML recognize the distributed nature of analytics and ML model development within enterprises. This development is spread across business, functional and IT teams. Thus, the following systems become critical to manage the key artifacts required for consistency and collaboration:

- Metrics management
- Model management
- Feature management

Metrics Management

Organizations that adopted self-service find themselves saddled with mounting technical debt to maintain fractured views of metrics. With fragmented metrics definitions and duplicated analytics pipelines, these organizations are looking for a means to govern self-service analytics.

In addition to the technical debt, organizations may have different calculations for the same metrics across teams. Different dashboards may show different numbers for the same metrics. Organizations looking to scale analytics struggle with a consistent approach to manage metrics. Analytics and data management teams have coded metrics within database objects, semantic layers and dashboards. A consistent metrics management solution requires a decoupled approach between metrics usage and metrics definition.

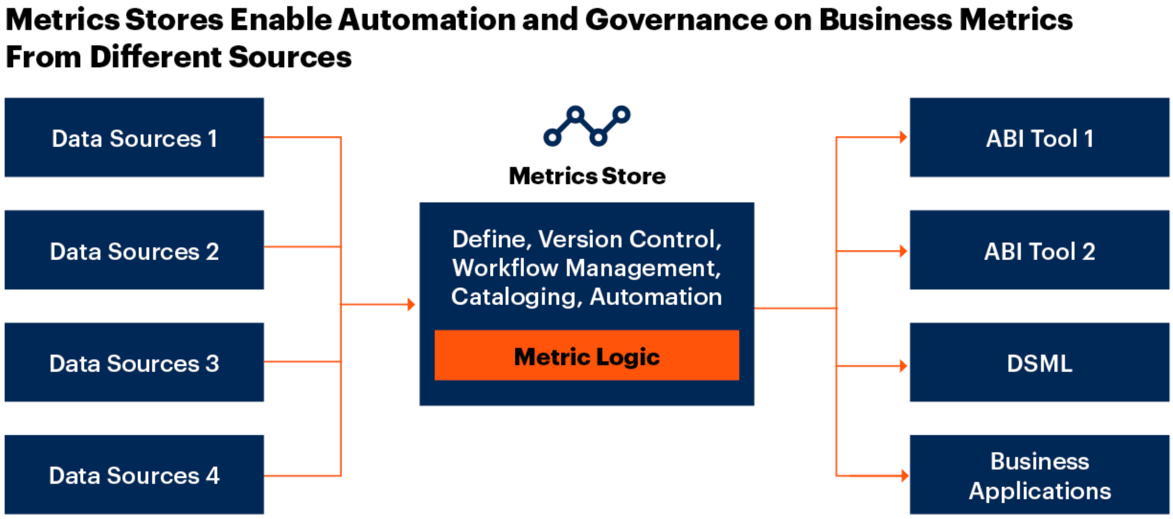
> A metrics store allows users to create and define business metrics as code, govern those metrics from data warehouses, and serve downstream analytics, data science and business applications.

A metrics store is an emergent capability within analytics platforms (see Figure 7). The primary purpose of a metrics store is to capture metrics definitions centrally and serve those metrics across any needed analytics use case. In an ideal scenario, a metrics store would allow business users to create and maintain metrics definitions, while enabling IT to act as the infrastructure custodian.

Metrics stores broaden stand-alone semantic layers by:

- Enabling business users to contribute and manage metric definitions
- Exposing metrics to use cases beyond general BI and enterprise reporting

Figure 7: Metrics Stores Enable Automation and Governance on Business Metrics From Different Sources


Metrics Stores Enable Automation and Governance on Business Metrics From Different Sources
Data Sources 1
Data Sources 2
Data Sources 3
Data Sources 4
Metrics Store
Define, Version Control, Workflow Management, Cataloging, Automation
Metric Logic
ABI Tool 1
ABI Tool 2
DSML
Business Applications

ABI = analytics and business intelligence, DSML = data science and machine learning

Source: Gartner

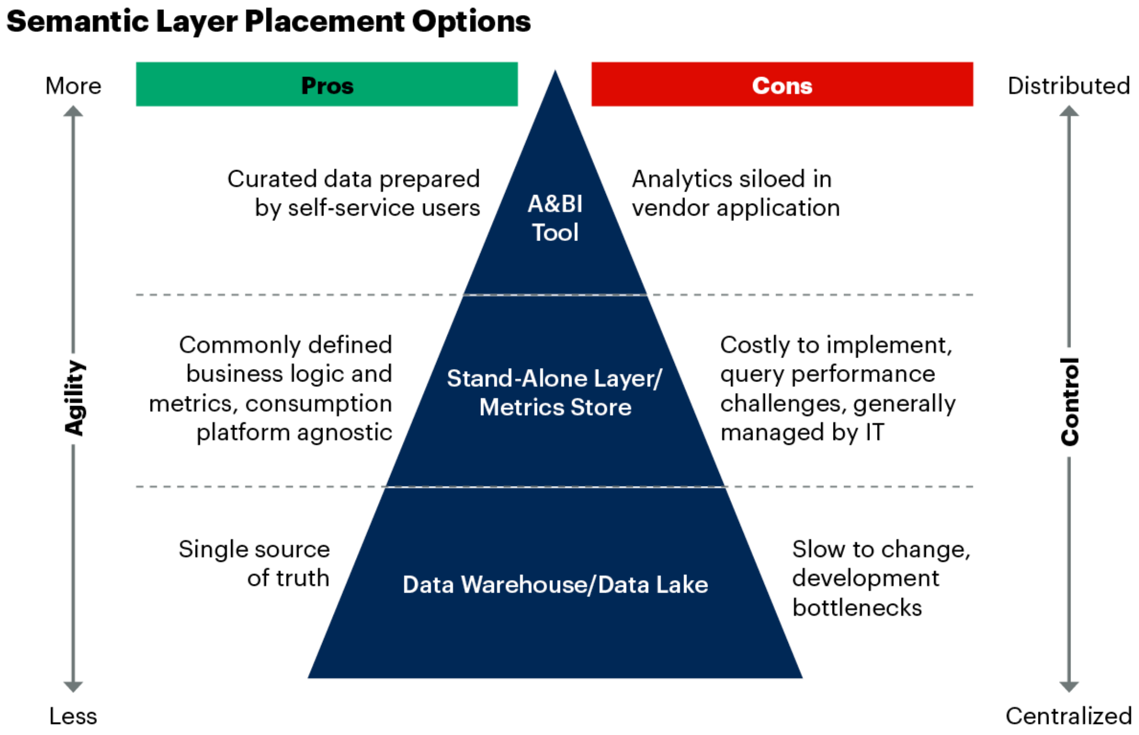
761373_C

While metrics stores provide a solution specifically for metrics, a broader approach is required to provide a consistent data model. This approach must encapsulate multiple data sources, multiple relationships across various entities, calculations, metrics, and translation from technical schema to business semantics. A *semantic layer* abstracts the underlying technology to provide a consistent and cohesive data view for business BI developers and citizen data scientists.

Semantic layers were initially implemented as monolithic enterprise wide views. However, these had maintenance and agility challenges, and they collected substantial technical debt as the core data management solutions advanced in capabilities. Vendors of analytics and BI tools have developed their own localized implementations. These provide implementation agility but bring additional governance and accessibility constraints. Nonetheless, an analytics semantic layer provides the appropriate abstraction to deliver a consistent, reusable view within its implementation scope.

Figure 8 provides a good representation of the various semantic layer placement options and their pros and cons.

Figure 8: Semantic Layer Placement Options


Semantic Layer Placement Options
More
Pros
Cons
Distributed
Curated data prepared by self-service users
A&BI Tool
Analytics siloed in vendor application
Agility
Commonly defined business logic and metrics, consumption platform agnostic
Stand-Alone Layer/ Metrics Store
Costly to implement, query performance challenges, generally managed by IT
Control
Single source of truth
Data Warehouse/Data Lake
Slow to change, development bottlenecks
Less
Centralized

Source: Gartner
783855_C

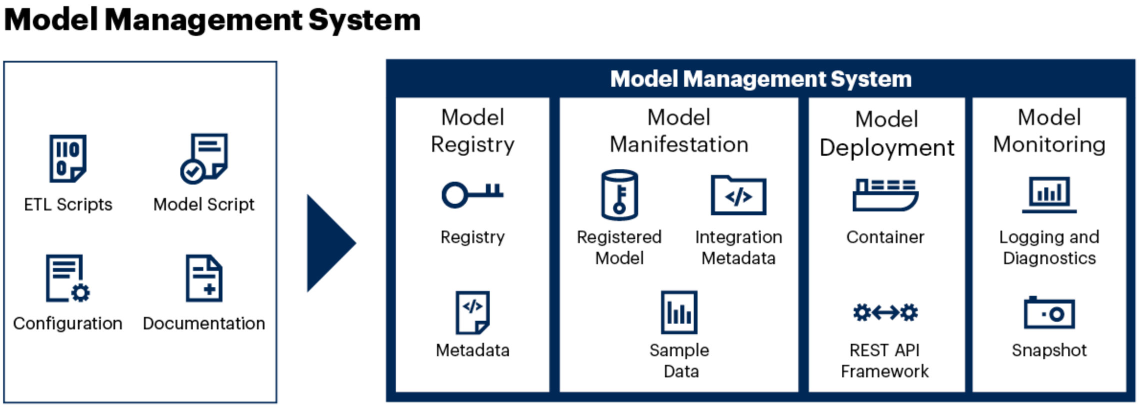
For more details on semantic layers and metrics store capabilities, see:

- Demystifying Semantic Layers for Self-Service Analytics
- Video: Demystifying the Metrics Store

Model Management

A *model management system* is essential for managing and governing the models efficiently (see Figure 9). Reproducible model inferences and reusable models and workflows help establish a baseline for audit, compliance, risk assessment and adherence to industry regulatory guidelines. They also form the basis for MLOps.

Figure 9: Model Management System



Source: Gartner
732258_C

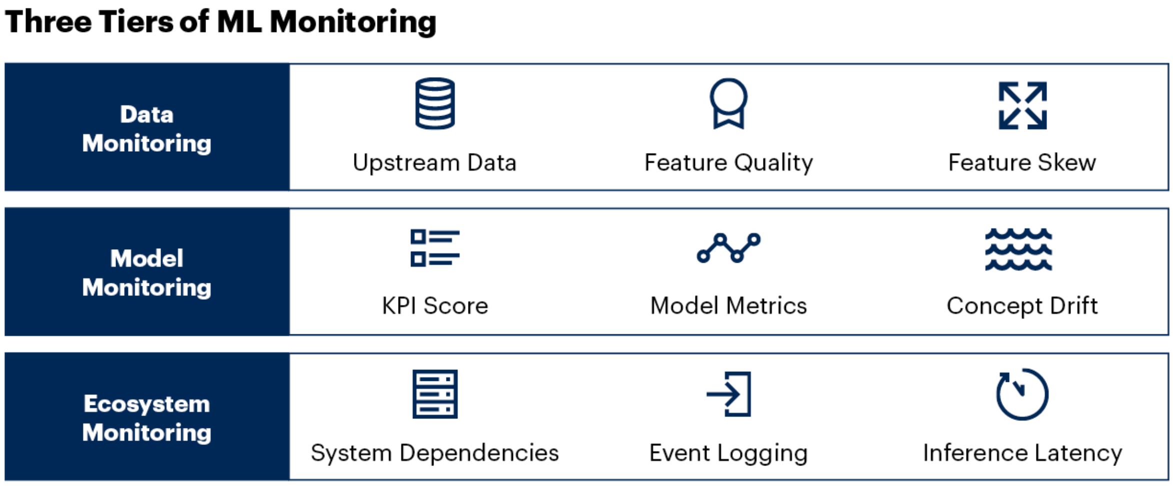
Technical professionals should implement a model management system supporting the following capabilities:

- **Model registry:** A model registry is necessary for versioning models and managing them across multiple environments — development, preproduction and production.
- **Model manifests:** Manifests are necessary for tracking the parameters and configurations of models in production.
- **Model serving:** Model serving abstracts models into common industry formats, such as Predictive Model Markup Language (PMML) and Open Neural Network Exchange (ONNX). It also deploys models for inferencing via REST APIs or containerization frameworks, such as Kubernetes.
- **Model monitoring:** Model monitoring, aka ML monitoring, aims to capture multiple dimensions across the following three tiers:
  - Data (e.g., upstream data, feature quality and feature skew)
  - Model (e.g., KPI score, model metrics and concept drift)
  - Ecosystem (e.g., system dependencies, event logging and inference latency)

ML monitoring (see Figure 10) is particularly important, as it will help interpret model inference, thereby establishing trust through explainability. The focus of ML monitoring must be on maximizing business-oriented KPI gain. Failure to demonstrate a measurable business benefit is why so many ML projects fail. The following provide general guidance on developing an effective ML monitoring system:

- **Define meaningful and measurable KPIs:** Establish a cross-domain team and combine their knowledge to design the KPIs. The team should include a data scientist, a subject matter expert, an application architect and operations personnel. The ML monitoring system should help measure and optimize KPI gain, thus facilitating the decision process.
- **Detect anomalies, and suggest the next best action:** ML monitoring must detect when there is a data pattern change that affects KPI gain (e.g., feature or concept drift), and must suggest what to do next. Most of the time, the next action will be to retrain an existing model or promote a challenger model.
- **Design a monitoring system that supports ongoing model deployment**: Rarely will an organization train and deploy a model that is never retrained or replaced. Therefore, the design must be able to simultaneously support many models that are in various stages of operation, as well as collect and update data for model retraining.

Figure 10: Three Tiers of ML Monitoring



Three Tiers of ML Monitoring

Data Monitoring	Upstream Data	Feature Quality	Feature Skew
Model Monitoring	KPI Score	Model Metrics	Concept Drift
Ecosystem Monitoring	System Dependencies	Event Logging	Inference Latency

Source: Gartner
783956_C

At the same time, model monitoring only alerts to model operational deficiencies. Enterprises should look to invest in AI observability solutions, which provide a holistic view that spans data pipelines, AI and ML models, and the interacting system components.

For more details, see:

- Launch an Effective Machine Learning Monitoring System
- Introduce AI Observability to Supervise Generative AI

Feature Management

A *logical feature store* is another key component to implement, alongside the model management system. Data scientists spend a disproportionate amount of time engineering features by finding and cleaning data, writing transformation code, and so on. This work is typically done on local machines, thus duplicating feature-engineering efforts, preventing reuse and making governance near impossible.

> **The logical feature store is a feature management solution that aims to break down silos, promote collaboration, reduce time spent on feature engineering and avoid duplication of work by sharing resources among data analysts.**

A logical feature store provides a mechanism to reuse and reproduce features and enables the various teams to collaborate. In the MLOps architecture (see Figure 6), the logical feature store provides another point of convergence where diverse modeling pipelines can come together for a unified view of features.

Technical professionals should seek to introduce the following components (see Figure 11) to their feature management solution:

- **Store:** The feature management solution can contain a database or repository that meets the enterprise's storage requirements for capacity scalability, high availability, high throughput and low latency of retrieval. Features in the store must have associated time stamps to allow dataset creation without data leakage and to recreate datasets with point-in-time correctness.

- **Access:** Features must be accessible to data scientists and other privileged people within the organization for future model development. For example, providing a UI with a searchable catalog of features allows features to be discovered, shared and reused across ML workloads.

- **Govern:** Policies and processes need to accompany this solution to govern access to features in a way that strikes a balance between democratizing and securing the data.

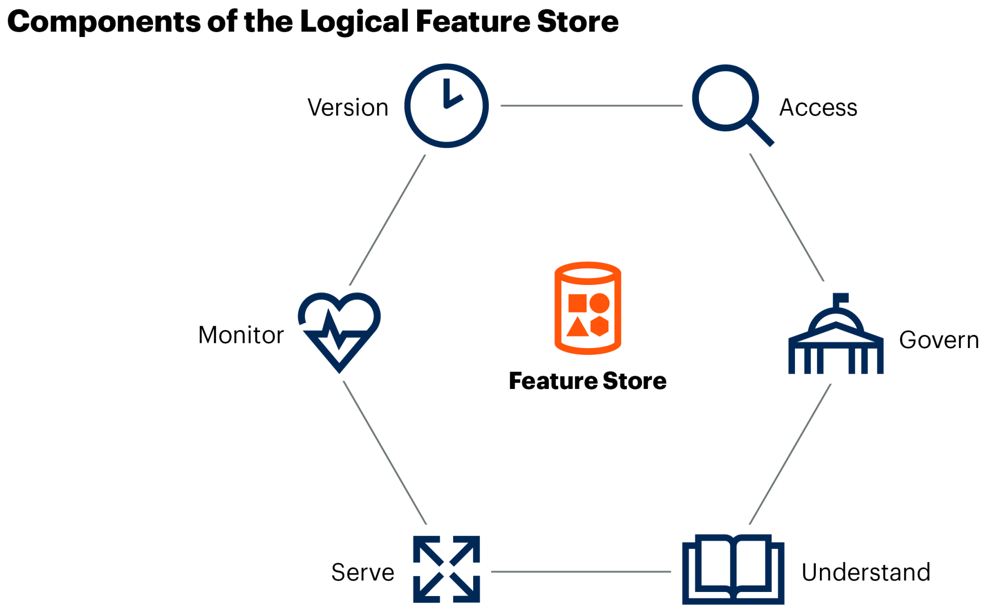
- **Understand:** To promote trust in the data, the catalog must be populated by relevant metadata that makes features understandable. Such metadata includes feature descriptions and definitions, lineage, versioned transformation code, dependencies, and feature transformation cadence.

- **Serve:** The feature management system should enable users to easily pull data to create training and test datasets for ML model development. At the same time, the system needs to serve and accommodate the organization's different ML use cases in production, which may be batch or on-demand.

- **Monitor:** Features should be constantly validated and monitored to prevent poor quality from disrupting models in production, thus decreasing accuracy and overall performance. Outgoing features serving models in production must also be monitored for data drift and training-serving skew by comparing up-to-date descriptive statistics (e.g., means, standard deviations and metrics of normality) with those collected during model training.

- **Version:** The ability to version data used for ML model development is a key component of managing the continuous cycle of ML model debugging, retraining and improvements that occur over time. The system must also have the ability to serve the right version of the right features to the right models in production.

Figure 11: Components of the Logical Feature Store



Source: Gartner
792810_C

For more details, including the architectural maturity levels of feature store capabilities, see:

- The Logical Feature Store: Data Management for Machine Learning

Trust and Transparency Requirements Will Drive Data Literacy and Responsible AI Practices

Back to top

Data governance helps ensure access to trustworthy data. While trustworthy data is foundational, analytics and AI are data consumers and support enterprises with decision making. Enterprise leaders and users must be able to trust the metrics presented in an analytics dashboard or in a prediction served by an AI model, since those metrics will drive business actions and outcomes. Data literacy and responsible AI practices bring in the structure, policies and conventions necessary for the reliability and quality of the outcomes.

These practices should apply to the range of steps that make up an analytics or AI implementation, including:

- Data wrangling and preparation
- Data source and data type integration
- Feature engineering and metrics calculation
- Dashboard and model implementation

Recently, many jurisdictions (e.g., the EU, China, the U.S. and the U.K.) globally introduced new and pending AI and data privacy regulations that challenge D&A leaders and practitioners to respond in meaningful ways. These pending regulations may conflict and lead to difficult business and technical decisions. Although traditional data governance practices are still important and useful, they must be supplemented with new approaches to accommodate the distributed nature of analytics development and the complexities of AI models.

Technical professionals looking to solve these challenges should follow these planning considerations:

- Implement a data literacy program to increase analytics proficiency.
- Leverage AI fairness toolkits and explainable AI frameworks for compliance and interpretability.
- Embed data privacy and model security practices as part of wider D&A governance.

Planning Considerations

Implement a Data Literacy Program to Increase Analytics Proficiency

A data-literate workforce is central to an effective D&A program and critical to driving measurable business outcomes. Enterprises are expanding self-service analytics initiatives and democratizing ML model development to empower business teams to analyze data and generate insights.

> **Gartner defines data literacy as the ability to read, write and communicate data in context. Data literacy includes an understanding of data sources, data constructs, and analytical methods and techniques applied, as well as the ability to describe the use case, application and resulting value.**

However, low data literacy can doom such initiatives, as it results in lower analytics adoption (due to higher barriers to entry) and less impactful analytics. The findings of the Gartner Chief Data and Analytics Officer Agenda Survey for 2023 highlighted the importance of data literacy. Facilitation of data literacy is ranked third as a critical enabler of success in the effectiveness of D&A initiatives, behind only data governance and BI and reporting. [10] The survey also highlighted trust, ethics and data privacy as the top driver for defining the organization's D&A priorities.

Therefore, enterprises striving to become "data-driven" must resolve trust and cultural challenges and provide resources that help employees achieve high levels of data literacy.

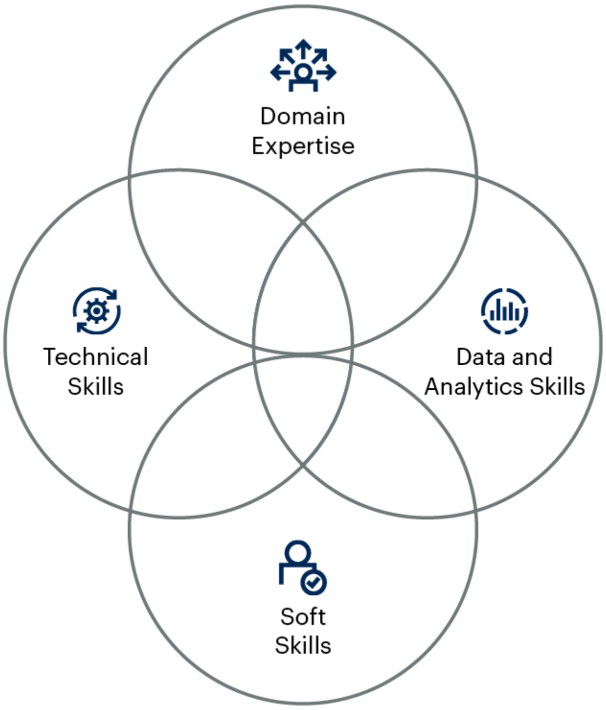
Data stewards, change management teams, and data and analytics technical professionals must collaborate to define and deliver a data literacy program. The steps for creating a data literacy program for the organization are:

- **Conduct a skills assessment:** Assess what skills users currently possess and their proficiency levels.
- **Identify core competencies:** Identify the core competencies needed by each role in IT and business user groups, especially those competencies related to data and analytics.
- **Match training with gaps in skills proficiency:** Identify the skills gaps, and match necessary training to bridge the gaps.
- **Design training curriculums:** Design training content within the subject areas.
- **Choose the right delivery format or formats:** Use multimedia or e-learning resources to disseminate information in a variety of channels.
- **Measure the effectiveness of training:** Decide on the metrics by which you will assess the effectiveness of training, and conduct this assessment periodically.

Often, data literacy curriculums put too much emphasis on developing technical skills and D&A skills, and not enough on incorporating domain knowledge. A successful data literacy curriculum needs to bring together several tenets and provide a convergence of domain-specific knowledge, technical skills, D&A skills and soft skills (see Figure 12).

Figure 12: Skills Convergence

Skills Convergence



Source: Gartner
752064_C

D&A technical professionals can support the data literacy curriculum by putting together documentation, resources and training on the organization's specific data assets, tooling and landscape. The data literacy program may be a combination of prepackaged materials (including free online materials, like YouTube videos) and companywide content that reflects domain expertise and specifics about the organization's data assets.

There are several methods for delivering training to the organization. D&A technical professionals designing upskilling paths for analytics can choose one or multiple modes of learning, such as:

- Classroom-based learning (either in-person or virtual)
- Workshops

- Self-paced online learning
- Game-based training

Successful training programs are flexible and provide a variety of options to suit user preferences and learning styles. When these techniques are used in combination, learning and retention are greatly enhanced. For example, classroom learning may be needed to introduce topics and provide a foundation for users. However, this alone is not sufficient to build proficiency in any topic. Working through generic examples in a classroom setting is unlikely to prepare users for the real challenges they may encounter when attempting to build their own analytics content.

Instead, training approaches should contextualize these skills by getting users to apply them to a specific problem that is relevant to their role within the organization. The organization may encourage users to tackle such problems on their own time, or it may guide them through the process by bringing a specific problem to a workshop. Alternatively, the organization could assign users to a working group to develop a solution as a team. Completing a project can even be made into a game or competition. There are many options.

For more information, see:

- Tackle Data Literacy Head-On to Avoid Data and Analytics Program Failure
- Data and Analytics Governance Approaches for the Technical Professional
- Self-Service Analytics Governance With Microsoft Power BI

Leverage AI Fairness Toolkits and Explainable AI Frameworks for Compliance and Interpretability

Many organizations are expecting AI to provide strategic differentiation from their competitors. The more organizations invest in AI, the more they need to understand how the AI models work and whether those models are biased. This knowledge will safeguard organizations from reputation and regulatory risks. At the same time, organizations need to ensure that solutions do not take on the bias of the designers and developers. Bias mitigation is necessary to both comply with regulations and protect the business's reputation. Explainable AI and fairness algorithms promise to resolve some of these challenges.

However, developing a fair model and explaining the rationale behind a prediction must not be an afterthought or a stand-alone activity. In order to make tasks repeatable, organizations must embed the following actions throughout the entire ML development and deployment process:

- Apply fairness metrics during data selection
- Use explainability to improve model training
- Leverage feature importance for validation, and explicitly test for bias
- Monitor the prediction, store the explainability factors and present the explanations in an interpretable way for the user

The following are key points for technical professionals to consider as they define and evaluate explainability and fairness components within the architecture:

- **“Fairness” definitions vary by use case:** Data scientists are often tasked with developing a fair model. However, defining fairness requires an understanding of the business processes and regulatory landscape. Business or product owners need to take the lead and collaborate with the compliance/risk expert and the data scientist to define what fairness means for the specific use case. They should also understand the ramifications of biased predictions and determine the guardrails required once the ML model is implemented in production.

- **Explainability tools and frameworks should align with use cases and impacted personas:** Multiple open-source and proprietary tools and frameworks provide both model-agnostic and model-specific views into the prediction model. However, there is no single solution to the problem. The process includes several decision points — algorithm, accuracy, interpretability, transparency, compute requirements, execution runtime and ease of implementation, to list a few. These decisions must be appropriate to the industry, use case and persona. For example:

    - Insurance actuaries or credit analysts would prefer accuracy and transparency over execution runtime and interpretability. They would use multiple tools — such as Shapley additive explanations (SHAP), partial dependence plots and feature importance — and analyze both local and global explanations to identify any model discrepancies.

    - Loan officers or claims processors, on the other hand, would prioritize interpretability and faster availability of results when looking to explain a decision to their customer. Local interpretable model-agnostic explanations (LIME), combined with feature importance for a specific prediction, would be more relevant for such use cases.

- **Not all business problems require explainability:** Organizations should not look to use ML explainability for all use cases. Explainability frameworks add complexity and overhead. In addition, not all models are explainable. For applications that create little or no adverse impact, the overhead associated with explainability may not be justified. ML models within games, emails, postal code sorting and movie recommendation systems are some examples.

- **The dynamic solution landscape needs proper assessment**: The landscape for explainability frameworks has been evolving rapidly. A combination of academic research, industry product extensions and open-source community contributions has provided several tools to understand ML predictions. Organizations should monitor academic research, conference presentations and publications from thought leaders to gain an understanding of the forthcoming solutions and capabilities. Most leading product vendors recognize the significance of ML explainability and are investing substantially in product development. This continuing product development, supported by research, should result in improved and more-user-friendly capabilities.

For more details, see:

- Incorporate Explainability and Fairness Within the AI Platform

Embed Data Privacy and Model Security Practices as Part of Wider D&A Governance

Enterprises are increasingly concerned about data security in several scenarios. These include:

- Collecting and retaining sensitive personal information
- Processing personal information in external environments, such as the cloud
- Sharing information internally and externally

To address this potential data security risk, consider a couple of privacy/preservation and enhancement techniques:

- **Synthetic data** is artificially generated, rather than obtained from direct observations of the real world. Data can be generated using different methods, such as statistical sampling, semantic approaches, generative adversarial networks or simulations (where models and processes interact to create completely new datasets of events). Organizations can use synthetic data to comply with data privacy regulations and to overcome data residency restrictions that prevent ML training on real data from different geographies. A synthetic dataset can be created that reflects the trends and dynamics within each individual set. Then, the synthetic dataset can be shared freely. This technique is particularly relevant in healthcare with patient data, as well as in other areas like analysis of customer data across multiple countries or multiple business units.
- **Differential privacy** is a type of data-level transformation that uses data perturbation, often in the form of adding statistical noise, to hide individual data values. Differential privacy is usually available as part of data-masking products or in the form of software components. The primary use case is analysis of “behavioral” data, where aggregating the results is sufficient for building ML models. The value of differential privacy is that it enables organizations to analyze this data without potentially exposing private information, such as specific user travel patterns, locations or web searches. Differential privacy may be applied on synthetic data for additional levels of obfuscation.

While data privacy and security challenges have been very visible, ML models are usually considered safe from a privacy standpoint because they do not contain any data. However, recent research has showcased model inversion techniques that allow adversarial attacks to infer the data used to train the models. [11] The security vulnerabilities of ML models can be broadly grouped into following categories:

- **Data poisoning attacks:** As the name of the attack implies, the goal of the attacker is to manipulate the model's output by inserting malicious ("poisoned") data into the training set. During the training phase, the model will learn from this incorrect data, potentially leading to unreliable model output. This output becomes a risk when it is used for decision making or as input for other systems.

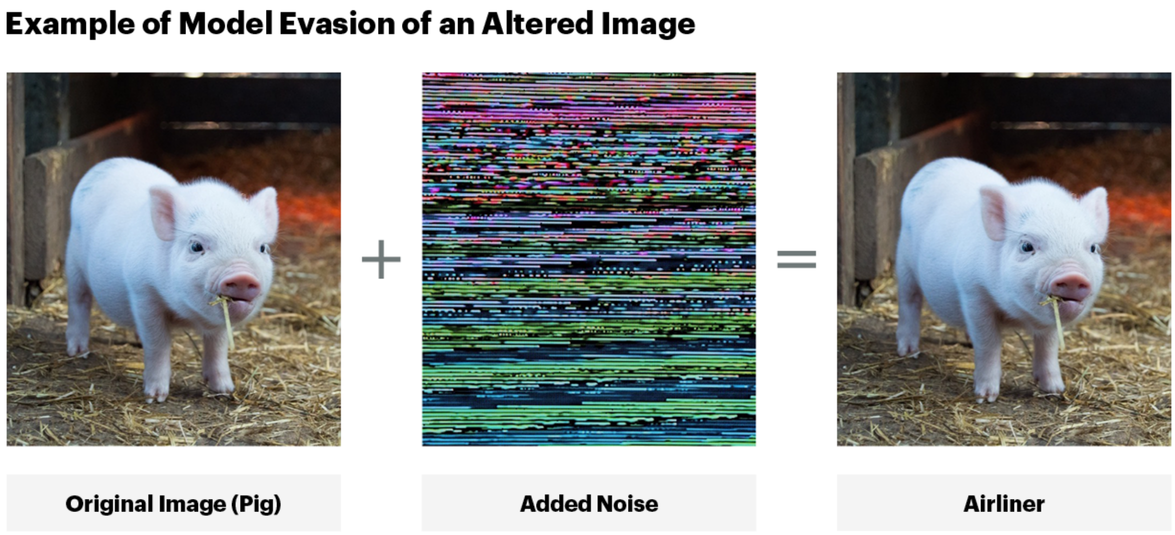
- **Membership inference attacks:** A membership inference attack tries to discover whether a certain data point was part of the training set. This attack is a risk when the training data contains sensitive, private or confidential data, such as the health information of individuals. It is an indirect attack because the data point is not revealed directly from the training data (such as a look-up of that data point in the training data), but rather, via the ML model. The attack can be a white-box or black-box attack, depending on how much the attacker knows about the underlying distribution of the training data and the target model (e.g., architecture, algorithm and hyperparameters). In the case of a black-box attack, the attacker can query only the target model to get its output.

- **Model evasion attacks:** The purpose of an evasion attack is for the attackers to evade the model's output by querying the model with manipulated input. This attack is a risk because it leads to unreliable model output. For example, in the case of a fraud detection model, the model might classify a manipulated fraudulent transaction as benign instead of fraudulent. Another example is the manipulation of physical objects, which can throw off an image classification model.

- **Model extraction attacks:** The purpose of an extraction attack is for the attackers to obtain detailed information about the target ML model. The attackers use this information to re-create a close approximation (or "copy") of the target model. This attack is a risk because the target model is copied (theft of intellectual property). Another risk is that the target model can be studied (using the copied model) as preparation for another attack, such as model evasion of the target model (two-step model attack).

As an example, Figure 13 shows how a model evasion attack can result in an incorrect prediction. The model correctly classified the original image as a pig. However, by adding noise to the picture (which is invisible to the human eye), the attacker misleads the model into incorrectly classifying the altered image as an airliner.

Figure 13: Example of Model Evasion of an Altered Image


Example of Model Evasion of an Altered Image

Source: Gartner
778428_C

Data security and data privacy (using data masking) have been integrated within data governance implementation frameworks. However, increased use of data, AI and ML models has surfaced new challenges, and enterprises need to integrate techniques such as synthetic data and adversarial testing of ML models within their technical architectures and frameworks. The increased adoption of LLMs adds greater urgency to these actions. LLMs are black-box models trained on massive datasets, and jailbreaking attempts have demonstrated examples of privacy information leakage and biased responses.

For more details, see:

- How to Securely Design and Operate Machine Learning
- Explore Data-Centric AI Solutions to Streamline AI Development

Evidence

[1] **2023 Gartner Voice of the Client Content Survey (Generative AI).** This survey was conducted online with 820 engaged Gartner clients in IT and business leader roles from 9 May through 31 May 2023. The objective of the survey was to better understand client needs, and to gauge use and expectations of generative AI in their organizations. Participants represented a wide range of industries and came from across the world: 56% from North America, 27% from EMEA, 13% from APAC and 4% from Latin America. All participants had recently engaged with Gartner's content on gartner.com (within the last 90 days). Disclaimer: Results of this survey do not represent global findings or the market as a whole but reflect the sentiments of the respondents and companies surveyed.

[2] **2024 Gartner CIO and Technology Executive Survey.** This survey was conducted online from 2 May through 27 June 2023 to help CIOs determine how to distribute digital leadership across the enterprise and to identify technology adoption and functional performance trends. Ninety-seven percent of respondents led an information technology function. In total, 2,457 CIOs and technology executives participated, with representation from all geographies, revenue bands and industry sectors (public and private). Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

[3] Attention Is All You Need, arXiv.

[4] BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, arXiv.

[5] Language Models Are Few-Shot Learners, arXiv.

[6] Training Language Models to Follow Instructions With Human Feedback, arXiv.

[7] **Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI.** This webinar was held on 30 March and 21 April 2023, for a total of 2,554 poll respondents. Results of these polls should not be taken to represent all executives as the survey responses come from a population that had expressed interest in AI by attending a Gartner webinar on the subject.

See also Executive Pulse: AI Investment Gets a Boost From ChatGPT Hype.

[8] **2021 Gartner AI in Organizations Survey.** This survey was conducted to understand the keys to successful AI implementations and the barriers to the operationalization of AI. The research was conducted online from October through December 2021 among 699 respondents from organizations in the U.S., Germany and the U.K. Quotas were established for company size and for industries to ensure a good representation across the sample. Organizations were required to have developed AI or intended to deploy AI within the next three years. Respondents were required to be part of the organization's corporate leadership or report into corporate leadership roles, and have a high level of involvement with at least one AI initiative. Respondents were also required to have one of the following roles when related to AI in their organizations: determine AI business objectives, measure the value derived from AI initiatives, or manage AI initiatives development and implementation.

Disclaimer: Results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

[9] **2022 Gartner AI Use-Case ROI Survey.** This survey sought to understand where organizations have been most successful in deploying AI use cases and figure out the most efficient indicators that they have established to measure those successes. The research was conducted online from 31 October through 19 December 2022 among 622 respondents from organizations in the U.S. (n = 304), France (n = 113), the U.K. (n = 106) and Germany (n = 99). Quotas were established for company sizes and for industries to ensure a good representation across the sample. Organizations were required to have developed AI to participate. Respondents were required to be in a manager role or above and have a high level of involvement with the measuring stage and at least one stage of the life cycle from ideating to testing AI use cases.

Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

[10] **Gartner Chief Data and Analytics Officer Agenda Survey for 2023.** This study was conducted to explore the business impact of the CDAO role and/or the Office of the CDAO and understand the leadership traits of the most successful CDAOs that distinguish them from their peers. The research was conducted online from September through November 2022 among 566 respondents from across the world. Respondents were required to be the highest-level data and analytics leader in the organization: chief data officer, chief analytics officer, chief data and analytics officer, the most senior leader in IT with data and analytics responsibilities, or a business executive such as chief digital officer or other business executive with data and analytics responsibilities.

Disclaimer: Results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

[11] For example, see:

- Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models, arXiv.
- Algorithms That Remember: Model Inversion Attacks and Data Protection Law, ResearchGate.

Document Revision History

2023 Planning Guide for Analytics and Artificial Intelligence - 10 October 2022

2022 Planning Guide for Analytics and Artificial Intelligence - 11 October 2021

2021 Planning Guide for Data Analytics and Artificial Intelligence - 9 October 2020

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Solution Path for Building Modern Analytics and BI Architectures

Roles and Skills to Support Advanced Analytics and AI Initiatives

How to Use Anomaly Detection to Implement AI-Driven Use Cases

A Guidance Framework for Deploying Data and Analytics in the Cloud

Data and Analytics Governance Approaches for the Technical Professional

Self-Service Analytics Governance With Microsoft Power BI

Table 1: Architecture Layers Mapped to Roles

Architecture Layer	Role
Application Layer	Application developer, UX designer
Service and Integration Layer	Software engineer
Model Layer	AI engineers, AI research engineer, data scientist
Data Layer	Data engineer, AI engineer, prompt engineer/business analyst
Infrastructure Layer	Cloud and infrastructure engineer

Source: Gartner (October 2023)