



Unified Data Fabrics for Multi-Sector AI Forecasting

1st Jai kiran reddy burugula , 2nd Thomas Müller

Jai Kiran Reddy Burugulla, Senior Engineer , jaikirrann@gmail.com
Thomas Müller , Junior Researcher

Abstract

This study assembles and evaluates intelligent big-data architectures that support AI-driven predictive analytics across multiple domains, using a common use-case structure in which predictive models are trained in one domain and evaluated in others. This approach surfaces architectural characteristics and patterns that improve the quality and appeal of predictive-analytics solutions within multi-industry digital ecosystems. A proposed predictive-analytics maturity model provides a roadmap for aligning data strategy, technological capability, and organizational readiness and investment to build more intelligent architectures and deliver more successful predictive-analytics outcomes. Key contributors to predictive-analytics-ready architectures include attention to data quality, governance, and lineage; well-designed feature stores; robust ingestion, storage, and management of data pipelines; and support for diverse data consumers—modelers, business-expert validators, and domain-specific data scientists and engineers. Together, these elements enable architectures capable of exploiting cross-domain transfer learning for both training and evaluation.

keywords: Big Data Architectures, Artificial Intelligence, Predictive Analytics, Data Engineering, Machine Learning Models, Cloud Computing Platforms, Data Lakes and Warehousing, Real-Time Data Processing, Digital Ecosystems, Multi-Industry Integration, Scalable Data Infrastructure, Edge Computing, Data Governance, Intelligent Decision Systems, Advanced Analytics.

1. Introduction

Organizations are digitalizing their processes and products to provide more competitive offerings. To support rhetorical and operational decision-making during the digitalization lifecycle of products and services, digital ecosystems are emerging in various industry sectors. Realized as partnerships across industry boundaries, digital ecosystems allow organizations to access new markets and business models through sharing and monetizing data and services. However, the heterogeneous and distributed data environment of a digital ecosystem makes operationalizing AI-driven predictive analytics challenging.

Yet, AI has been identified as a key technology to test ideas quickly, with applications emerging in virtual assistants, autonomous vehicles, and image and face recognition. Organizations across industries are seeking AI as a service, enabling predictive analytics solutions and establishing the need for intelligent big data architectures that allow deeply trained skills to be reused. By leveraging AI-driven predictive analytics service offerings and assessing investments in AI, organizations can identify specific solutions and services that demonstrate immediate value.



1.1. Data Quality, Governance, and Lineage

No predictive analytics system can outperform subpar data quality. Therefore, for environments inhabited by multiple organizations, establishing and maintaining data quality is an ongoing process requiring adequate investment in data governance. High data quality is facilitated by having a well-defined data lineage, which reveals not only where data is stored but also how it was created, updated, and transformed throughout the analytics life cycle. Data lineage directly affects explainability: when a model produces unexpected results, having a clear trace of how the data was handled allows analysts to focus on possible error sources. For assets generated externally, data provenance reveals the original source, enhancing trust.

Two established frameworks, the Data Quality Framework and the Data Management Body of Knowledge, can be adapted for data quality and governance in predictive analytics systems spanning multiple industries. The approach considers the functional areas of data management, which comprises data governance, data quality management, data integration, data architecture and design, data storage and operations, data security management, reference and master data management, and data warehousing. When cross-domain predictive analytics are required, the influence of data quality, governance, and lineage expands beyond the organization that generates the data to include other collaborators.

For inference, the feature store serves multiple models. Both operational and retrieval performance are crucial. Integrating multi-domain or multi-system model-building requirements in a feature store enables cross-domain modeling and related transfer-learning scenarios.

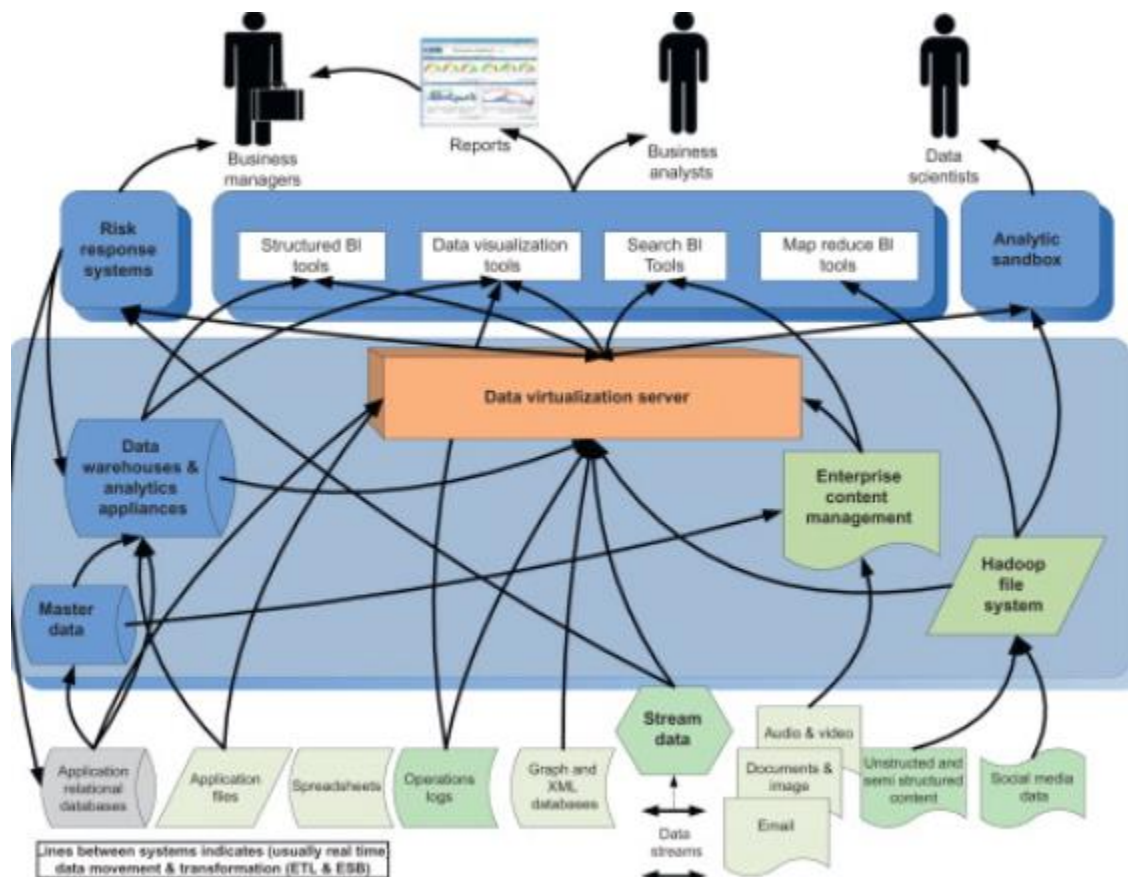


Fig 1: Big Data Architecture - an overview

1.2. Feature Store Design for AI-Driven Insights

AI-driven predictive analytics focus on data preparation and on the model training and deployment cycle. In this process, the data serving system feeds a data analytics engine, serving prepared datasets to fit model training or inference requirements. The choice of interaction pattern with the serving system – whether a streaming or batch approach, real-time or offline serving, etc. – depends on the task and on its parameters.

Serving features to models during training and inference calls for a dedicated feature store. This is a centralized repository designed to store features, serving them reliably, accurately, and efficiently for both model training and inference. Historical feature values are stored up to the time of model training, powered by a batch processing engine.



Equation 1: Predictive model equation

$$\hat{y} = f(\mathbf{x})$$

Step-by-step derivation

Let the input feature vector be

$$\mathbf{x} = [x_1, x_2, x_3, \dots, x_n]^T$$

A predictive model is a mapping from inputs to outputs:

$$f: \mathbb{R}^n \rightarrow \mathbb{R}$$

So when we feed the feature vector $\mathbf{x}$ into the model f , the predicted output is

$$\hat{y} = f(\mathbf{x})$$

2. Frameworks for Intelligent Big Data Architectures

Intelligent big data architectures for AI-driven predictive analytics must integrate frameworks for the different facets that affect data ingestion, storage, and management. Data can be stored and managed by means of traditional relational or columnar data warehouses or databases; NoSQL databases optimized to store semi-structured data; graph, time-series or geo-distance databases; or low-cost object storage solutions. Regardless of where data is ultimately stored, data-lake-pattern digital ecosystems gather, store, and provision sources from as many of these heterogeneous storage solutions as possible.

Various architectural approaches may be employed when integrating AI-driven predictive analytics with additional services offered in the context of a digital ecosystem. Microservices or serverless architectures can ease the development and deployment of the analytics services by decoupling them from legacy systems and operational processes. A data-mesh pattern may be employed to build the analytics services in conjunction with domain teams. These services can generate enriched prediction data that is then shared, exposed, and made available for exploration in a federated-analytics pattern by downstream data-and-analytics teams developing solutions in other domains and business lines.

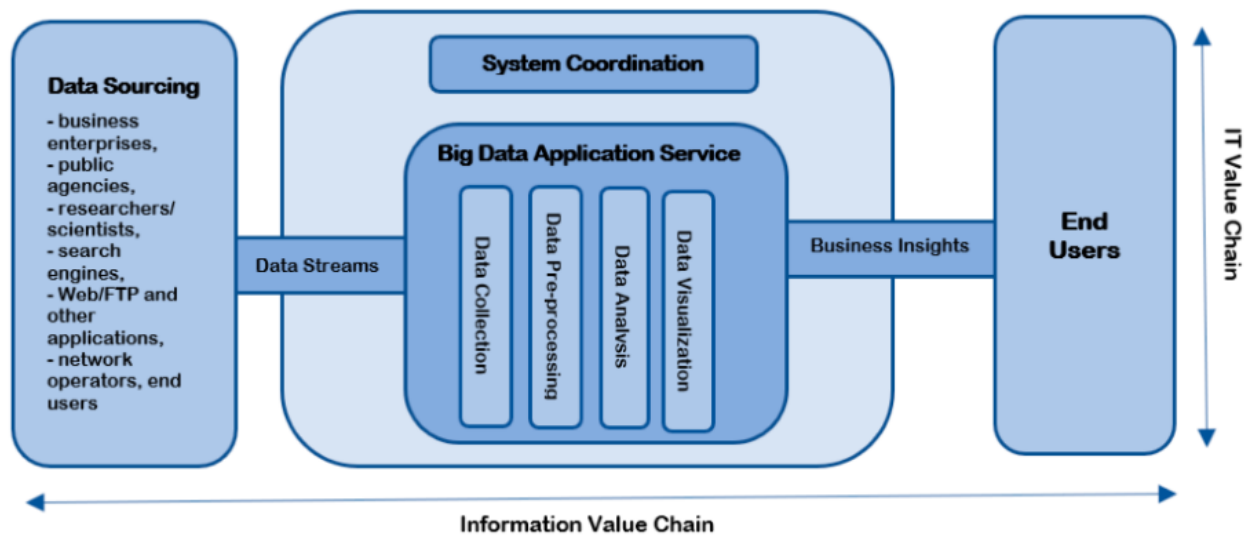
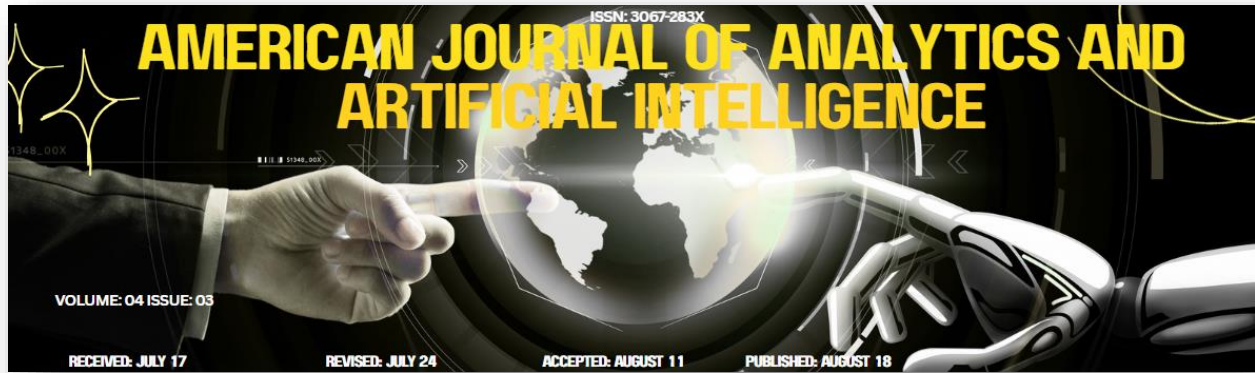


Fig 2: Big Data Analytic Framework

2.1. Data Ingestion and Integration in Heterogeneous Environments

This section discusses data ingestion and integration in heterogeneous environments, particularly in the context of multi-industry digital ecosystems. Modern enterprises typically require intelligent big data pipelines that facilitate the collection and preparation of data from different sources, including third-party partners, for predictive analytics. The data can be structured or unstructured, representing a single domain or different domains. Ingesting, enriching, and preparing the raw data are resource-consuming processes, so batch processing is often preferable, though real-time capabilities may be required for specific cases.

Data ingestion is the process of collecting and transporting data from one or more sources to a datastore, often a data lake, for farther processing. A subset of processes related to data ingestions—capturing actual changes from data sources—is commonly referred to as change data capture (CDC). In the data pipeline context, data ingestion is part of the extract step in the extract, transform, load (ETL) or extract, load, transform (ELT) paradigms. Data pipelines need to support multiple ingestion mechanisms beyond traditional APIs or file transfers, such as data as a service (DaaS) provided by third-party suppliers or integrations with other data pipelines. A detailed overview of the latest developments in data ingestion is provided in Chen et al. Data ingestion from streaming and batch sources, such as files or relational database management systems, into cloud-based data lakes is described in Malavika et al.



Architecture Layer / Component	Main Purpose	Key Functions in the Article	Expected Benefit
Data Sources	Collect multi-domain data from heterogeneous ecosystems	Structured, unstructured, batch, streaming, third-party, and cross-industry sources	Richer data foundation for predictive analytics
Data Ingestion Layer	Bring data from sources into the analytics platform	ETL/ELT, change data capture, API/file/DaaS ingestion, batch and real-time pipelines	Reliable data acquisition across heterogeneous environments
Data Storage Layer	Persist raw and processed data at scale	Data lakes, warehouses, relational/columnar stores, NoSQL, graph, time-series, geo-distance databases, object storage	Scalable and flexible storage for varied workloads

Table 1: Core Intelligent Big Data Architecture for AI-Driven Predictive Analytics

2.2. Data Storage and Management for Scalable Analytics

The semantic processing stage (7) aims to design a common, unified representation for multi-domain applications. Semantic technology is different from traditional data management approaches in that the focus is built around the definition of semantic metadata linking. The semantic metadata creation and maintenance is the main activity in the semantic processing layer. The information comes from the data sources participate in the data integration process. The data repository keeps the semantic metadata skeleton and provides a graphical interface for administrators. Applications implement the semantic approach through the interface provided by the repositories.

Performance is a major concern for high volume data storage architecture and scalable data processing. The analysis and comparison of the performance is performed in two directions: first, traffic of data storage and the purpose of each memory level; and second, data scheduling and control of data processes distribution. Efficient traffic flow avoids data blockage and ensures that the speed of the slowest link does not limit the performance of the entire application. Control of data topic ensures dynamic assignment of processing resource to the data in the application. Both are major issues for the performance contract.

Equation 2: Linear predictive form from stored features

$$\hat{y} = w_1x_1 + w_2x_2 + \dots + w_nx_n + b$$

Step-by-step derivation

Start from the general prediction rule:

$$\hat{y} = f(\mathbf{x})$$



Assume the model is linear. Then each feature contributes proportionally to prediction:

$$f(\mathbf{x}) = w_1x_1 + w_2x_2 + \dots + w_nx_n + b$$

Using vector notation:

$$\mathbf{w} = [w_1, w_2, \dots, w_n]^T$$

$$\hat{y} = \mathbf{w}^T \mathbf{x} + b$$

Because

$$\mathbf{w}^T \mathbf{x} = [w_1, w_2, \dots, w_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = w_1x_1 + w_2x_2 + \dots + w_nx_n$$

Hence,

$$\hat{y} = \mathbf{w}^T \mathbf{x} + b$$

3. AI-Driven Predictive Analytics in Multi-Industry Contexts

Multi-domain correlations can be identified and subsequently exploited via cross-domain predictive modeling. The core task is to develop a model for a given target domain and predict values in that domain on the basis of models that have been learned on other source domains, thus allowing it to benefit from empirical data that ensure high prediction accuracy in those domains. Since predictive models for a target domain cannot be learned in a direct fashion (e.g., a lack of a sufficient amount of reliable data), multi-domain transfer learning techniques can be applied to reduce the noise that can be caused by a poor-quality model in the source domain chosen to transfer knowledge.

Time-series forecasting aims at predicting future values of a variable in a time-ordered series of observations. Anomalies, that is, observations occurring at a different frequency than the data norm, can have significant effects on system performance and therefore should be detected in advance. Such predictions are enabled by techniques involving anomaly detection, dynamics extraction, temporal pattern mining, and temporal representation learning. Artificial neural networks are used for multivariate time-series forecasting and for a novel unbalanced multivariate time-series forecasting and anomaly-detection technique.

3.1. Cross-Domain Modeling and Transfer Learning

Recently popularized, the concept of cross-domain modeling enables the definition of predictive models in a domain where there



is a lack of proper information or data, but it is still related to other domains supporting the knowledge transfer through learning across domains, known as transfer learning. Cross-domain modeling has been successfully applied to the healthcare domain where the prediction of the number of meningeal carcinomatosis cases in the Brazilian Federal District was defined without the need for proper domain-expert knowledge. The model was built by transferring the learned patterns from the more abundant volume of historical data from the neighboring state of Goiás, for which the proper prediction of the number of cases is feasible.

A recent survey on cross-domain modeling in the area of predictive analytics has identified domain recommendation as an important research opportunity. This demand emerges from the fact that most of the studies in this area applied the concept to related domains in a multi-domain operational context, characterized by inter-domain relationships that enable the cross-domain approach. As a response to this opportunity, a successive research work proposed a new direction for cross-domain modeling by applying the approach in a context in which the set of domains is not governed by reciprocal relationships, enabling the recommendation of a source domain able to support the definition of a predictive model in a target destination domain. Recent advances in predictive analytics have highlighted cross-domain modeling as a promising yet underexplored area, particularly in the context of domain recommendation. Traditionally, research in cross-domain modeling has focused on multi-domain environments where domains share explicit or implicit relationships, allowing knowledge transfer based on their interdependencies. However, a recent line of work shifts this perspective by addressing scenarios in which such reciprocal relationships do not exist. In these cases, the challenge lies in identifying a suitable source domain that can effectively inform model development in a distinct target domain, despite the absence of direct connections. This emerging direction expands the applicability of cross-domain approaches, positioning domain recommendation as a critical mechanism for enabling predictive modeling across heterogeneous and loosely related domains, and opening new avenues for research in adaptive and transferable learning systems.

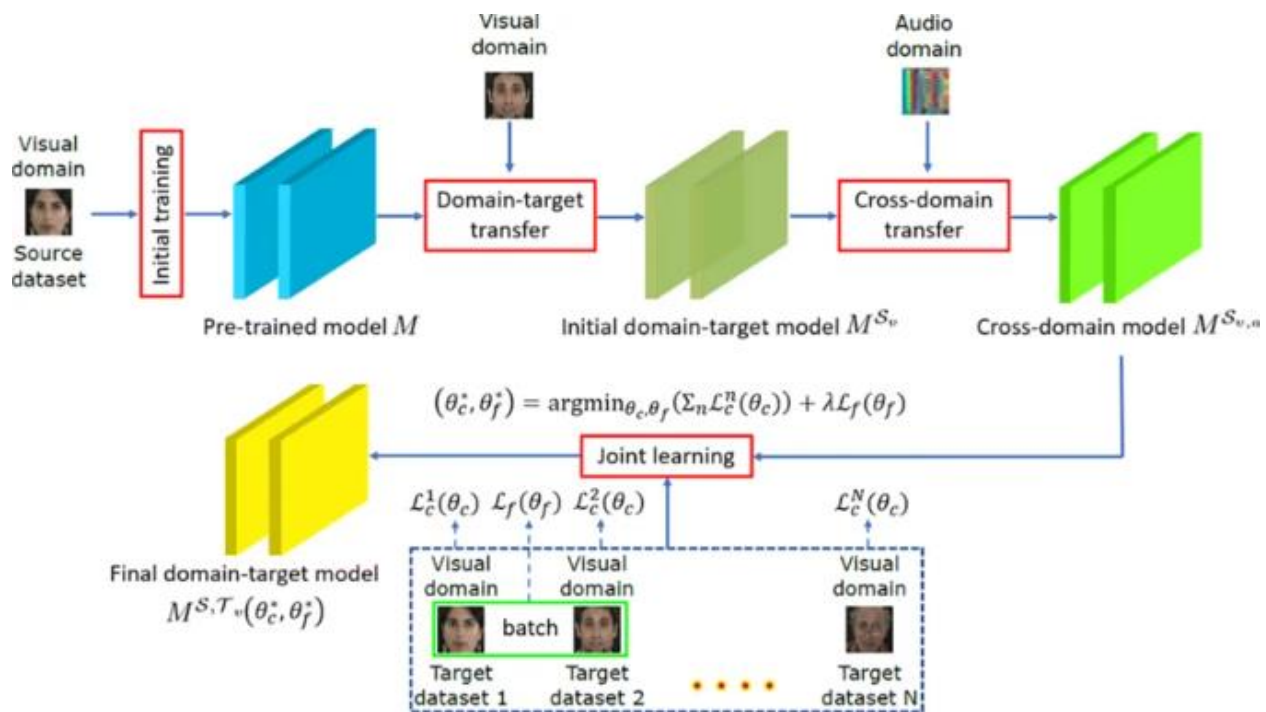


Fig 3: Our cross-domain transfer learning

3.2. Time-Series Forecasting and Anomaly Detection

Time-series forecasting and anomaly detection are predominant use cases for AI-driven predictive analytics, incorporating TensorFlow, PyTorch, and AutoML frameworks for building custom solutions, as well as open-source and commercial time-series-specific tooling. Key implementations—forecasting electricity consumption and wind power generation, detecting anomalies in the operation of critical infrastructures, and detecting abnormal transaction patterns to flag Financial Action Task Force compliance concerns—serve as industry exemplars, with mature data products, design patterns, and implementation roadmaps. These examples showcase a controlled transfer learning approach across domains where data-rich sectors assist data-scarce counterparts.

Time-series forecasting encompasses multiple subcategories: general forecasting, traffic forecasting, demand forecasting, wind power generation forecasting, and electricity load demand forecasting. The growing interest in time-series has resulted in specialized libraries that exploit the unique characteristics of temporal data, such as NeuralTST, NeuralForecast, and PyTorch Forecasting. Transfer learning for time-series comprises several variants, including pretraining on auxiliary tasks and distinct domains. Anomaly detection for temporal data is supported by a growing number of specific solutions.



Equation 3: Cross-domain transfer learning objective

$$\theta_T = \theta_S + \Delta\theta$$

Step-by-step derivation

Let

- θ_S = parameters learned in the **source** domain
- θ_T = parameters needed in the **target** domain

If there were no transfer, we would train θ_T completely from scratch.

But transfer learning says the source model already contains reusable knowledge, so the target model can be written as:

$$\theta_T = \text{source knowledge} + \text{adaptation}$$

Replace “source knowledge” with θ_S , and “adaptation” with $\Delta\theta$:

$$\theta_T = \theta_S + \Delta\theta$$

Why this makes sense

- If the domains are very similar, adaptation is small:

$$\Delta\theta \approx 0$$

so

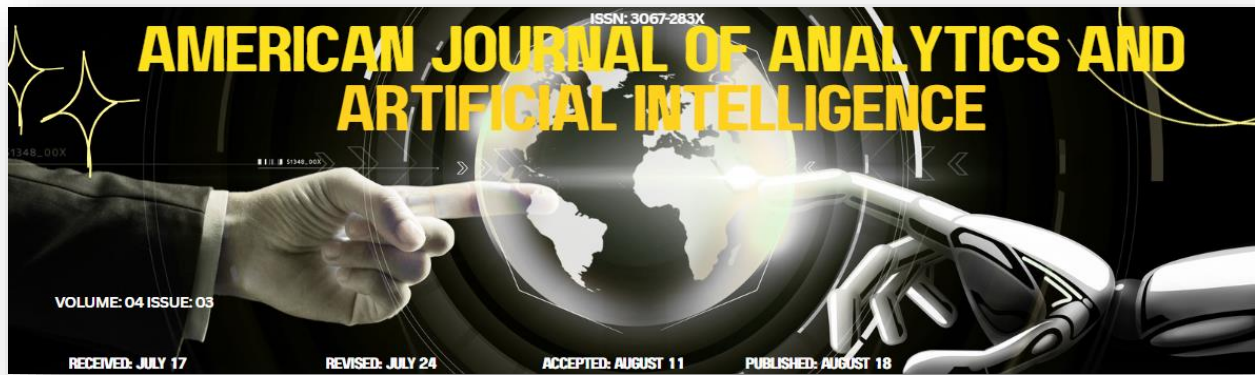
$$\theta_T \approx \theta_S$$

- If the domains differ more, a larger correction is required:

$$|\Delta\theta| \text{ is larger}$$

4. Architectural Patterns for Digital Ecosystems

Intelligent big data architectures support AI-driven predictive analytics across multiple industries, with shared business logic deployed in either microservices or serverless architectures. A data mesh enables borderless data access and sharing, while



federated analytics distributes the governance of federated analysis across the ecosystems. Digital ecosystems are typically driven by AI, with predictive analytics designed to anticipate behaviors such as future resource demand, maintenance needs, and anomalous events. AI-based predictions enable proactive measures that minimize adverse effects and optimize the allocation of costs and resources.

4.1. Microservices and Serverless Architectures

The microservices architectural style offers flexibility and agility in deploying cloud-based applications. Each microservice executes an isolated task on a lightweight server and interacts with other microservices as needed. The key idea is to break the application down into self-contained, distributed components that achieve single functionality. Microservices are often jointly designed with continuous development and delivery principles in mind. They may be hosted in virtual machines or containers on an Infrastructure-as-a-Service platform. The behavior and deployment of services are usually governed by an orchestration engine.

The serverless model builds upon the microservices paradigm and is optimized for sporadic use cases involving peaks/dips in the load profile. With the serverless approach, execution costs correspond to the actual execution time of the function rather than the resources provisioned for the whole microservice. Developers write the application code and deploy individual functions that respond to particular events. In the cloud, only the function code and runtime environment are bundled together and all aspects of runtime management—including provisioning, scaling, scheduling, and execution of resources—are entirely hidden from the developer. The transaction is thus “serverless” for the developer, although it incurs a cold-start latency.

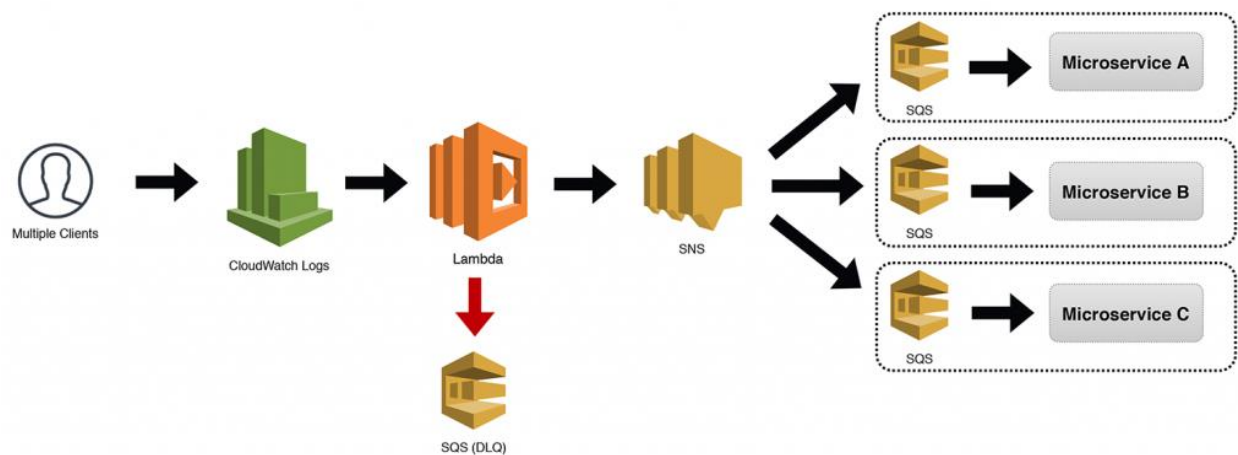


Fig 4: Microservices VS Serverless

4.2. Data Mesh and Federated Analytics

These architectural patterns support multi-domain AI-driven analytics systems within the digital ecosystem of organizations and



their partners. De-centralizing responsibility for the data and creating APIs that federate access to it introduces ownership and speed of development whilst allowing privacy and compliance requirements to be met.

In a data mesh approach, every part of an organization supporting a particular business domain is accountable for its data products. Using software and governance rules that enable interoperability across domains, a domain's data are created and consumed internally and by others using data-as-a-product principles. Each domain deploys a scalable data platform component to support use cases that require low-latency or high-volume access, while data pipelines supporting ad-hoc analytics and supporting models are deployed using a federated architecture, with instantiation driven by demand for specialized hardware or by the scale of use.

With a federated analytics approach, organizations combine data without leaving their own environments. Formal arrangements between partners enable data sharing across institutional boundaries, addressing enterprise data sovereignty and regulatory requirements. Data owners decide how their data are accessed, using APIs to remove the need for copies. Risk hierarchies underpin federated learning and other decentralized forms of collaboration.

Industry / Context	Main Predictive Use Case	Data / Modeling Approach	Architecture Enablers	Business Value
Healthcare & Life Sciences	Disease trend prediction, patient/resource planning, medical analytics	Time-series analysis, anomaly detection, predictive modeling, transfer learning	Data pipelines, dashboards, feature stores, governed sharing	Better resource allocation, improved care delivery
Epidemiology / Public Health	Outbreak forecasting and impact assessment	Multi-domain public data, online prediction, anomaly detection	Cross-domain ecosystem, integrated ingestion, scalable analytics	Earlier response to health crises
Financial Services	Risk management, fraud/anomaly detection, compliance monitoring	Predictive, cognitive, and interactive AI; anomaly detection	Security-by-design pipelines, access control, masking, governed data usage	Reduced fraud, stronger compliance, better response times

Table 2: Cross-Industry Predictive Analytics Use Cases and Enablers

5. Governance, Security, and Trust in AI-Driven Analytics

Governance, security, privacy, and trust are of paramount importance in AI-driven predictive analytics systems. Effective technologies, architectures, and processes for such analytics must comply with regulatory requirements and industry guidelines for data governance and risk management. Security at all system levels is crucial since the applications involve sensitive data in a multi-cloud environment, covering domains such as finance, healthcare, and energy. Security measures deal with data



provenance and lineage throughout the entire lifecycle: both the flow of data during ingestion and integration, as well as the masking of sensitive features used in predictive modeling for risk assessment. Trust creates a virtuous cycle of more data feeding better models, leading to improved predictions and further confidence in the governance framework because of the safeguard of bias monitoring.

Access control mechanisms, identity management, and data sovereignty requirements are complemented by data masking techniques in the data factory and in the delivery of predictions to consumers. Although federated analytics obviates the need to share training data, it is still necessary to define the circumstances in which a predictive model can be deployed in a consuming domain—a challenge given that data leaks can occur even with well-trusted third parties. Security-by-design principles guide the construction of all analytical and predictive models, enabling the secure execution of scripts by all users of the analytics framework.

Equation 4: Time-series forecasting equation

$$\hat{y}_{t+1} = f(y_t, y_{t-1}, \dots, y_{t-k+1})$$

Step-by-step derivation

Let the time-series be

$$y_1, y_2, y_3, \dots, y_t$$

To forecast the next point, use the previous k observations:

$$y_t, y_{t-1}, \dots, y_{t-k+1}$$

Define the input vector for forecasting as

$$\mathbf{x}_t = [y_t, y_{t-1}, \dots, y_{t-k+1}]$$

From the general prediction rule:

$$\hat{y} = f(\mathbf{x})$$

Substitute the time-series input vector:

$$\hat{y}_{t+1} = f(\mathbf{x}_t)$$

Expanding $\mathbf{x}_t$:

$$\hat{y}_{t+1} = f(y_t, y_{t-1}, \dots, y_{t-k+1})$$



5.1. Access Control, Identity, and Data Sovereignty

Fundamentally, access to data must be restricted to those authorized to use it for the intended purpose, taking into account the principle of least privilege, with access dynamically granted and revoked as appropriate. Protecting data from unauthorized use also requires addressing identity and data sovereignty considerations. Organizations should define policies governing their users' identities, whether based on location or organization, and should support national and regional differences in regulation and compliance.

Data that is shared across borders must also align with the sovereignty requirements of the different parties involved, and a federated architecture can help organizations address these questions. Granting those developing or maintaining an analytic deployment access to the data being analyzed for quality control and maintenance must be handled in compliance with access and sovereignty regulations. Security-by-design considerations during data pipeline development assist in safeguarding data quality while enabling access where appropriate.

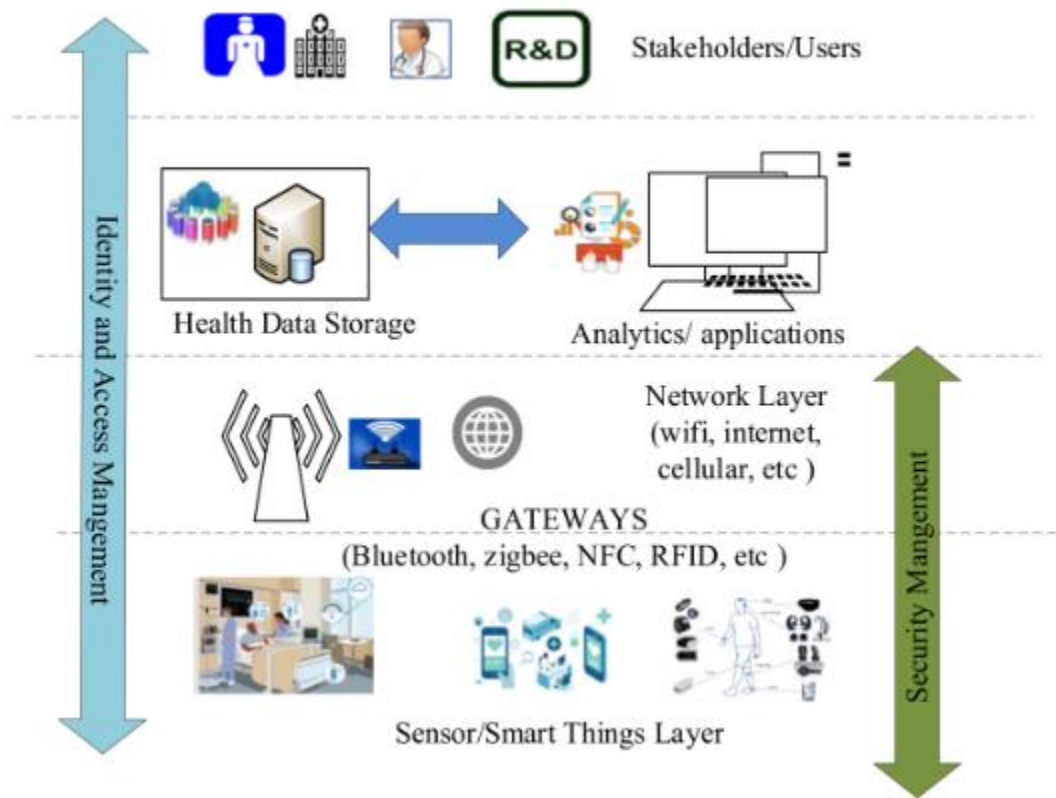
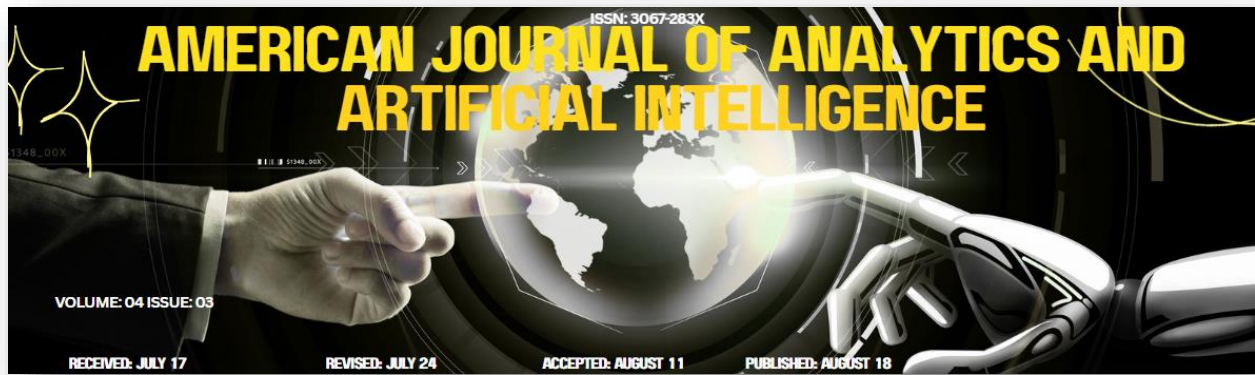


Fig 5: Access Control, Identity, and Data Sovereignty

5.2. Security-by-Design in Data Pipelines

Data-driven AI systems operating in the cloud and edge are susceptible to numerous attacks. Prioritizing security requires adopting a security-by-design prospect for development and operation. Integrating security protocols in the model lifecycle is essential to identify and mitigate threats. Data used in analytics, such as feature stores and training data, should also be protected. Secure data pipelines prevent unauthorized access and the insertion of corrupted data during ingestion, training, and testing phases. Data integrity measures ensure that label flipping, data stuffing, and poisoning attacks are minimized.

Strategies for securing data pipelines include combining security requirements with the architecture and functional requirements of the machine learning workflow. A defense-in-depth approach involves security mechanisms at various pipeline locations. Streamlined processes allow instant detection and response during the pipeline operation phase. Security and privacy measures can also be associated with the individual building blocks of a data pipeline. Incorporation into the design and planning phases



enhances prevention and response capabilities, satisfying compliance and security assessments. Segregating controls and provisioning data masking further reduce the attack surface and enhance data protection.

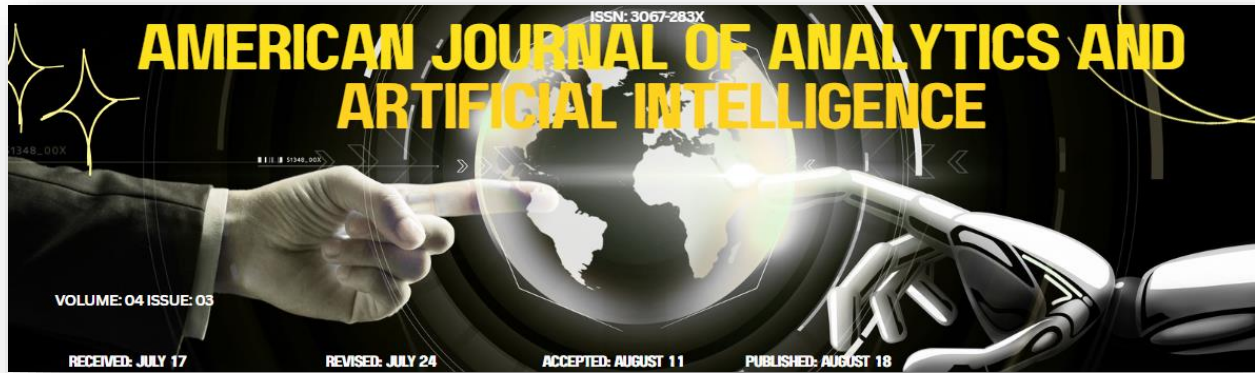
6. Evaluation of Predictive Analytics Systems

Evaluating predictive analytics systems entails a number of considerations. A collection of model assessment metrics for measuring quality and performance is beneficial to determine the applicability of different systems across datasets and industry verticals. Such metrics help anticipate the transferability of the models and provide guidance for reassessing models in a new domain. Multiple tasks might need to be considered depending on the large variety of industries adopting AI-driven predictive analytics, such as forecasting future values of time-series data and detecting anomalies in time-series data and multidimensional data.

The evaluation framework thus covers model performance estimation by using various well-known forecasting quality metrics for the forecasting task and combination of the quality score of one-class classification-based anomaly detection methods with decision-supporting metrics such as ROC-AUC for the anomaly detection task. These measures can be used for both quality assessment and benchmarking of predictive analytics systems touching on different tasks in various contexts across industries, aiding further development and deployment of AI-driven analytics in both research and industry.

Dimension	What the Article Recommends	Typical Actions	Desired Outcome
Data Quality	Treat quality as foundational	Define quality rules, monitor defects, improve completeness/consistency	More reliable predictions
Governance	Build governance across organizational boundaries	Establish policies, stewardship, compliance controls, decision rights	Trusted multi-party analytics
Lineage & Provenance	Maintain full traceability of data and transformations	Capture origin, movement, updates, transformations, model input history	Better explainability and auditability
Technology Architecture	Choose scalable and interoperable architecture patterns	Use data lakes, warehouses, feature stores, microservices/serverless, cloud platforms	Flexible and scalable predictive platform
Security & Privacy	Apply security-by-design throughout pipelines	Identity management, access control, masking, defense-in-depth, sovereignty controls	Lower risk and stronger regulatory compliance

Table 3: Implementation Roadmap and Maturity Dimensions



6.1. Evaluation Metrics and Validation Protocols

Rigorous evaluation of AI-driven analytics systems is essential for an informed adoption. Evaluation protocols should tailor the general recommendations for supervised learning, consider the significance of data quality and governance, and extend to the various AI scenarios supported by the system. Standard single-model accuracy measures fail to capture the predictive performance in both cross-domain and time-series settings and should be complemented with specialized metrics. Ex-post validation provides a useful complement to the traditional ex-ante assessment. Centralized benchmarking efforts accelerate the assessment while making it more affordable for individual organizations.

The standard metric spreadsheet of a predictive modeling toolbox offers a basis for defining a comprehensive set of predictive-performance measures. For ML models, the spreadsheet also specifies appropriate data-splitting and cross-validation protocols. Ultimately, however, testing remains a task that should specifically receive dedicated attention for each new implemented ML model to avoid misuse of these guidelines. More pressing than ever, the pressing need for trusted predictive insights relevant to a wide range of industry sectors has been underscored by the COVID pandemic. Forward-looking modelling is a powerful approach for satisfying this need, as it enables exploration of not only the future course of an indicator but also the drivers shaping that course. The ability to harness data spanning diverse countries and industry sectors enables modelling that captures broader patterns, imbuing it with a higher order of generalization power. Indeed, for certain indicators and data domains, more reliable predictive models may reside in a domain orthogonal to that of direct interest, allowing transfer learning methodologies to be exploited. The case for cross-domain predictive modelling approaches is bolstered by the breadth of data-sharing technology now available, which permits timely sourcing of a vast range of data beyond traditional industry silos.

Equation 5: Anomaly detection score and decision rule

$$s(\mathbf{x}) = \|\mathbf{x} - \boldsymbol{\mu}\| \text{ and Anomaly if } s(\mathbf{x}) > \tau$$

Step-by-step derivation

Assume normal behavior has average feature vector

$$\boldsymbol{\mu}$$

For a new observation $\mathbf{x}$, measure how far it is from normal.

Using Euclidean distance:

$$s(\mathbf{x}) = \|\mathbf{x} - \boldsymbol{\mu}\|$$

Expand the norm:

$$\|\mathbf{x} - \boldsymbol{\mu}\| = \sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + \dots + (x_n - \mu_n)^2}$$

This gives an anomaly score.



Now define a threshold τ :

- if score is small, point is normal
- if score is large, point is anomalous

So the decision rule becomes

If $s(\mathbf{x}) > \tau$, $\mathbf{x}$ is anomalous

6.2. Benchmarking Across Industries

The practical utility of AI-driven predictive analytics can be demonstrated through benchmarking forward-looking predictive models of business performance based on transversal data domains against various contextualized industry-specific applications. Published literature is mined to identify indicators and data sources for industry-focused domains, confirming that services are a prominent contributor to industry value added in developed economies and reliable data sources are available for key services sectors including healthcare, education and financial services. The results make the case for creating Portals of Predictability that distil the collective wisdom residing within, and across, industry ecosystems. A structure is proposed that integrates performance indicators, explanatory drivers, data pipelines, building blocks and portal links, thereby enabling users to contextualize and operationalize the insights from transfer-learning forward-looking forecasting models, and adapt them for practical use. The overall endeavour contributes to addressing the pressing need for insightful, cross-industry, transparent and trustworthy predictive models that help predict, prepare for and mitigate the impact of unforeseen events.

AI-driven predictive analytics offers significant practical value by enabling organizations to forecast business performance through the integration of diverse, cross-domain data. By benchmarking forward-looking models against industry-specific contexts, it becomes possible to identify meaningful indicators and leverage reliable data sources—particularly in service-driven sectors such as healthcare, education, and financial services, which are central to value creation in developed economies. Building on insights from existing literature, the concept of “Portals of Predictability” emerges as a powerful framework to consolidate and operationalize collective industry knowledge. These portals combine performance metrics, underlying drivers, data pipelines, and modular analytical components, allowing users to apply advanced forecasting models enhanced through transfer learning. Such a structured and transparent approach not only improves decision-making but also strengthens trust in predictive systems, ultimately helping organizations anticipate disruptions, prepare proactively, and mitigate the effects of unforeseen events across industries.

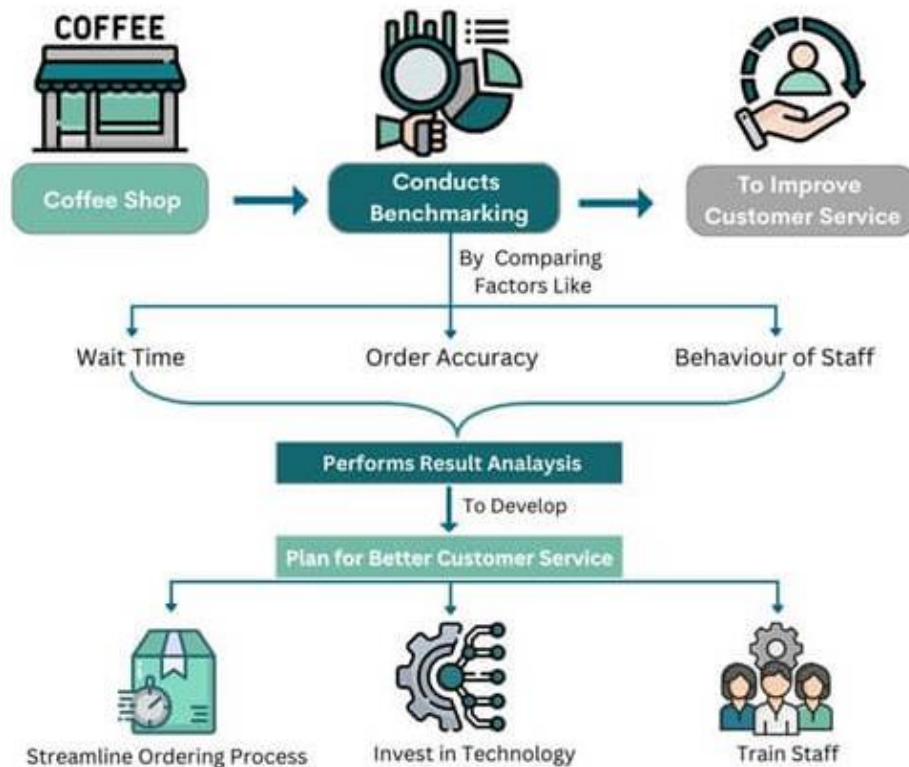


Fig 6: Benchmarking

7. Case Studies Across Industries

Cross-domain predictive analytics systems applied to data from multiple domains enhance interpretability and accelerate development for new domains. Studies across different industries, including healthcare, epidemiology, financial services, and manufacturing, showcase the feasibility of applying a predictive analytics system to diverse use cases. The same approach yields time-series predictions, anomaly detection, and depth effect prediction in healthcare, risk assessment in service industries, and remaining useful life predictions in non-availability of asset use cases. The domain-agnostic nature supports transfer learning across domains and tasks, provided data from one domain is suitably sufficient and representative.

Healthcare Systems – Severe Acute Respiratory Syndrome (SARS-CoV) is a single-stranded positive-sense RNA virus responsible for the respiratory illness known as SARS, first reported in Asia, Europe, and North America in 2002 and subsequently spread globally. Data on the outbreak and development of the disease and the responsiveness of health systems are



rarely collected in a timely and coherent manner across many countries and regions. To address these challenges, the Data4COVID project was launched as a multi-continental, multi-domain collaboration with the aim of supporting the healthcare infrastructure by providing predictive analytics necessary for allocating the right resources to the right places at the right times. A basic predictive analytics ecosystem with an interactive data exploration and prediction dashboard, utilizing time-series analysis, online anomaly detection, and online prediction of the wealth effect with the underlying depth effect of COVID-19 on different wealth groups, was developed from publicly available social, economic, and epidemiological data. Validation of the adopted approaches demonstrated their feasibility and provided initial support to the project.

Flow chart :

flowchart TD

A[Multi-industry data sources] --> B[Data ingestion layer]

B --> C[Raw data lake / landing zone]

C --> D[Data cleaning and transformation]

D --> E[Semantic integration / unified representation]

E --> F[Feature engineering]

F --> G[Centralized feature store]

G --> H[Model training environment]

H --> I[Validation and benchmarking]

I --> J[Model deployment]

J --> K[Prediction services / APIs]

K --> L[Business apps / dashboards / decision systems]

L --> M[Monitoring and feedback]

M --> D

M --> F



M --> H

7.1. Healthcare and Life Sciences

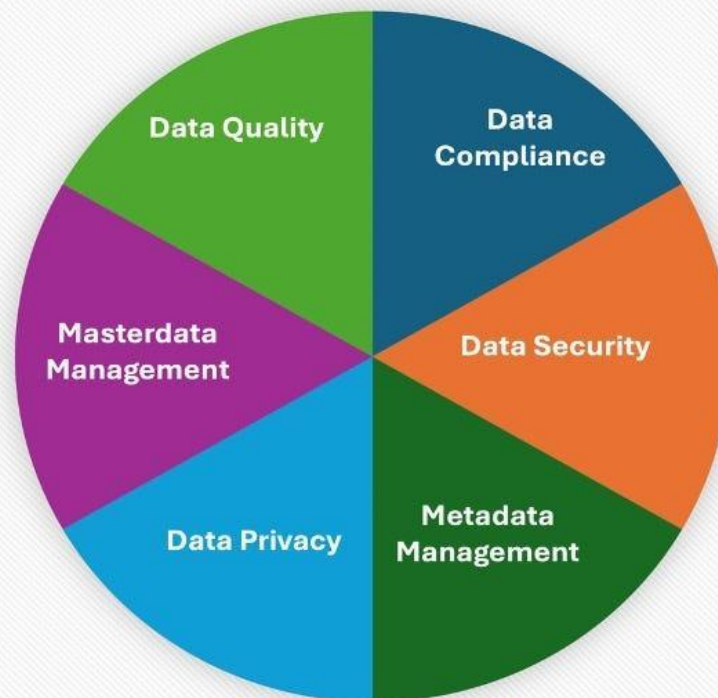
Disruptive technologies change the world of healthcare and life sciences. Technical advances lead to improved gathering of data, more precise dissemination and making decisions based on the knowledge of all available information. Examples are improvement in genetics for a more personal treatment of patients or Artificial Intelligence especially image recognition in medical scans. Yet in these rapidly changing worlds, the need for software applications and for new models still exists.

The healthcare industry has been at the center of technological and socio-political discussions since the COVID-19 pandemic began. In spring 2020, companies began to reinvent themselves to find solutions to improve service efficiency or manage their internal processes in times of crisis. Unlike other industries, for example, the retail or travel industries, it was not only the damage to sales but also the image loss that forced change for survival. Digital transformation had already begun years before, but COVID-19 acted as a catalyst while creating new challenges. The healthcare industry had to rethink how to support patients and, at the same time, manage staff, real estate and supplies in a resource-constrained environment.

Some companies converted their factories to produce masks or vaccine and medicine distributors changed their logistics. The proximity to the customer has been maintained through the use of Apps and data analytics. But these newly implemented IT infrastructures should also meet the needs of Business Intelligence requirements by managing data both for operations and strategy. Despite the development of new production and distribution processes, in the medium to long term the focus should return to the consumer, his satisfaction and his experience.



Pillars of Data Governance



7.2. Financial Services and Risk Management

Financial institutions, including banks, insurance companies, and other investment firms, are primarily concerned with the trade, investment, or transference of money and financial instruments. These institutions seek to maximize profit through an efficient allocation of capital, while managing and controlling risk. Risk management encompasses the identification, assessment, monitoring, and response to risk exposures. Risk-sensitive investment management requires that risk be measured, controlled, and reported. The industry has a recognized demand for these activities in both the content area and for the enabling technology.

Artificial intelligence can support transformations across all lines of business; its predictive, cognitive, and interactive capabilities make it ideal for driving more proactive risk management, enabling seamless customer experience, reducing costs, and supporting compliance. Nevertheless, implementing AI is no small undertaking. AI requires a new way of thinking and practicing risk and security analytics. Beyond dealing with scale and complexity, organizations must contend with critical human-centric issues: skills and talent shortages, culture and change management, and unintended bias and ethical security concerns. Questions also linger around the best way to deliver AI capabilities, especially for complex, risky, security-related use



cases that demand advanced analytics. Organizations are not investing in AI just for the sake of it. They are chasing hard benefits such as improved response times and reduced fraud, enhanced customer experience, and, in some cases, cost reductions.

8. Roadmap for Implementation

Although several Intelligent Big Data Architecture benchmarks are articulated, considerable gaps still exist when applying the guidelines to practical, real-world implementations. Consequently, considerations are presented to guide successful implementations, such as adoption journey, maturity models, and the three strategic pillars — business, data, and technology — that must be addressed in parallel across their dimensions. This roadmap can underpin the implementation of a multi-industry common code base and corresponding library/framework for deploying these AI-driven predictive analytics systems into production.

Developing a data strategy is a complex endeavor due to the diversity of technology choices and data management challenges across organizations. The roadmap identifies the core decisions that affect the success of a data strategy in its organizational, data, technology, and analytical-dimension components. An assessment of organizational readiness estimates whether a company has the right level of motivation, resources, assets, alignment, and expertise to be ready for transformation. The readiness statement provides recommendations specifically tailored toward increasing readiness so that successful Digital Transformation can occur and the data strategy can be executed effectively.

Flow chart :

A[Data source created or received] --> B[Register provenance]

B --> C[Apply governance rules]

C --> D[Assess data quality]

D --> E{Quality acceptable?}

E -- No --> F[Clean / enrich / correct data]

F --> D

E -- Yes --> G[Track lineage through transformations]

G --> H[Store governed data]

H --> I[Use in feature store / models]

I --> J[Explain results using lineage]



8.1. Maturity Models and Roadmapping

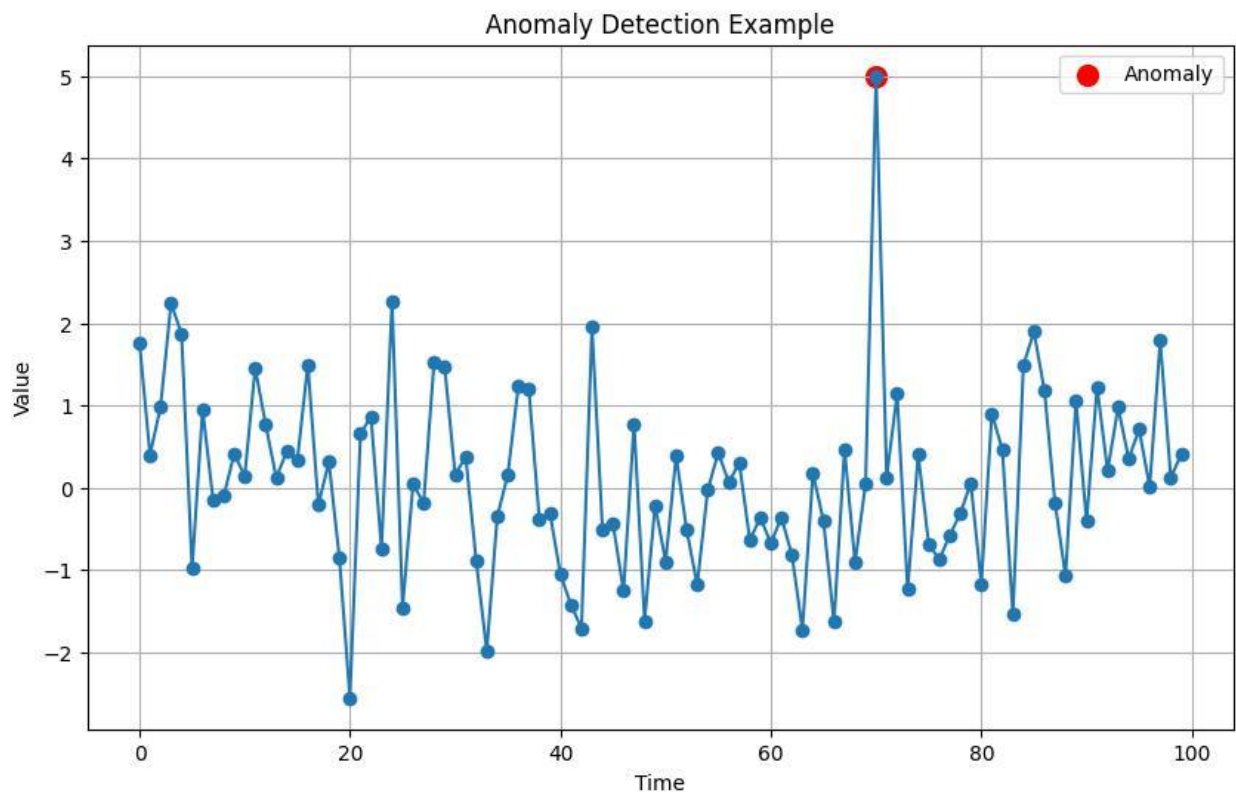
Holistic roadmaps guide the development of AI-driven predictive analytics platforms. The discussion follows an industry-agnostic maturity model encompassing an intelligent big data architecture and AI-powered predictive analytics stack. It concentrates on roadmap items common to many industries and sectors—in particular, predictive analytics systems fulfilling governance by design and security by design principles.

The analytics maturity model encompasses six dimensions: analytics capabilities, technology architecture, governance, model quality, user experience, and ecosystem interconnectivity. High degrees of interconnectivity and external-facing components differentiate operating systems for industry ecosystems. Roadmap directions concentrated on the governance and model quality dimensions, which align with the challenges and opportunities for many industries and sectors.

8.2. Data Strategy and Organizational Readiness

Digital ecosystem projects are typically complex and high-stakes initiatives that demand commitment across all levels of an organization. To strike the right balance of challenges and opportunities and deliver real business returns from digital transformation, organizations should assess their ability to take data-driven decisions. A maturity model helps evaluate data strategy across levels of operationalization of common themes such as: vision, governance, readiness and capabilities.

The nature of AI-based predictive analytics projects underlines that the success of new machine learning solutions hinges more on data than on algorithms. Therefore, data strategy should be tailored to improvement areas that maximize the quality of data that feeds these solutions. Data and analytics have traditionally been seen as a by-product of business processes, managed by a central team. The emergence of cloud, open-source technologies and second-generation AI has shifted that paradigm, enabling a decentralized model for improving data quality and availability, while implementing and monitoring lineages across the ecosystem. Thus, organizations should invest not only in engineering and operational processes but also on corporate data strategy, governance and the scarce role of data steward in the AI-driven era.



9. Conclusion

Digital ecosystems can benefit from several emerging main architectural patterns, such as microservices and serverless. Their design enables sharing and reuse of advanced AI capabilities, while meeting strict sovereignty requirements. The federated data ecosystem natively supports data-sharing among autonomous sources, the data mesh takes the further step of completely decentralizing the technical governance of analytics workloads, and the client-dedicated semantic layer enables cross-organization analytics with minimized data movement. To properly assess the solution's maturity, various dimensions can be defined and leveraged to build roadmaps that global enterprises can follow for consistent implementation.

AI-driven predictive analytics and decision-support systems can be deployed across domains and industries. By employing transfer learning, cross-domain modeling approaches tackle key problems ranging from energy consumption and wastewater quality to risk and price prediction. Time-series forecasting approaches address a wide range of applications—from cyber-security to traffic and utilities—together with dedicated anomaly detection solutions. Finally, the design and development of AI-



driven workloads in any sector or domain should adhere to security and privacy-by-design principles, with implicit authentication and authorization integrated within the data pipeline.

9.1. Future Directions

Current research gaps comprise the integration of causal reasoning at large scale, real-time time-series forecasting and anomaly detection on fee-based cloud services, systems for elastic AI and resource multitenantage, inter-system level performance evaluation, pragmatic microservices evaluation and selection, and qualitative benchmarking of AI explainable systems.

Cloud-based hyperscale computing infrastructure plays a crucial role in the currently prevalent setup of distributed artificial intelligence (AI) and intelligent big data processing by providing fee-based services ranging from GPU-based elastic AI services to automated ML-2 through ML-7. The latter enables the creation and deployment of multitenant systems that serve multiple organizations on a single set of model instances. Such services, although fee-based of negligible cost, require real-time requirements to be fulfilled. Data pipelines enabling such cost-effective resources should thereby be operated under an explicit premise of security—privacy—trust for becoming a preferred choice for enterprises.

The continuing research in intelligent big data analytics and the development of associated intelligent big data architectures must therefore now focus on practical and realistic topics to add contribution and value to science and society beyond cloud CO2 savings, reduced energy consumption, exploitability of digital natives created digital monopolies, and such already attained societal benefits.

10. References

1. Alsharef, A., Aggarwal, K., Sonia, Mishra, A., & Kumar, M. (2022). Review of ML and AutoML solutions to forecast time-series data. *Archives of Computational Methods in Engineering*, 29, 5297–5311.
2. Araújo Machado, I., Costa, C., & Santos, M. Y. (2022). Data mesh: Concepts and principles of a paradigm shift in data architectures. *Procedia Computer Science*, 196, 263–271.
3. Sukumar, S., Kumar, S., Yandamuri, U. S., & KV, M. N. (2026). *Business Ethics and Corporate Sustainability*. BR Publications.
4. Blohm, I., Wortmann, F., Legner, C., & Köbler, F. (2024). Data products, data mesh, and data fabric: New paradigm(s) for data and analytics? *Business & Information Systems Engineering*, 66, 643–652.
5. Donta, P. K., Dehury, C. K., & Hu, Y.-C. (2024). Learning-driven data fabric trends and challenges for cloud-to-thing continuum. *Journal of King Saud University Computer and Information Sciences*, 36, Article 102145.
6. Gieß, A., & Hutterer, A. (2025). The future of data management: A delimitation of data platforms, data spaces, data meshes, and data fabrics. *Information Systems and e-Business Management*, 23, 971–997.
7. Sharma, C. K., Gorle, R., Mangala, N., Jindal, A., J. A., & Sundram, S. (2026). Blockchain-Based Human Resource Management System for Transparent Workforce Governance. In *2026 7th International Conference On Computational Vision and Bio Inspired Computing (ICCVBIC)* (pp. 1741–1745). IEEE. 2026 7th International Conference On Computational Vision and Bio Inspired Computing (ICCVBIC). <https://doi.org/10.1109/iccvbic71195.2026.11689175>



8. Gu, Z., Corcoglioniti, F., Lanti, D., Mosca, A., Xiao, G., Xiong, J., Calvanese, D., & others. (2024). A systematic overview of data federation systems. *Semantic Web*, 15(1), 107–165.
9. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), Article 71.
10. Kanagarla, K. P. B. (2024). Data fabric: A new approach to data integration. *International Journal of Innovative Engineering and Management Research*, 13(10).
11. Penaka, B. (2026). AUTONOMOUS MULTI-AGENT AI FRAMEWORK FOR END-TO-END CLINICAL DECISION SUPPORT AND CARE COORDINATION. *Genetics and Molecular Research*.
12. Kim, J., Kim, H., Kim, H., Lee, D., & Yoon, S. (2025). A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges. *Artificial Intelligence Review*, 58, Article 216.
13. Kong, X., Chen, Z., Liu, W., Ning, K., Zhang, L., Marier, S. M., Liu, Y., Chen, Y., & others. (2025). Deep learning for time series forecasting: A survey. *International Journal of Machine Learning and Cybernetics*, 16, 5079–5112.
14. Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*, 11, 31866–31879.
15. Kuftinova, N. G., Ostroukh, A. V., Filippova, N. A., Gaevskii, V. V., & Podgorny, V. V. (2022). Integration of scalable IT architectures on the basis of data fabric technology. *Russian Engineering Research*, 42(11), 1199–1202.
16. Kuftinova, N. G., Maksimych, O. I., Ostroukh, A. V., Volosova, A. A., & Matukhina, E. N. (2022). Data fabric as an effective method of data management in traffic and road systems. In *2022 Systems of Signals Generating and Processing in the Field of Onboard Communications* (pp. 1–4). IEEE.
17. Arunakumari, M., Mattaparthi, R., Kothuri, S., Prakash, V. D., & Prasad, E. S. (2026, July). Data-Driven Framework for Real-Time Carbon Accounting and Sustainability Monitoring in Off-Grid Urban Residential Microgrids. In *2026 14th International Conference on Smart Grid (icSmartGrid)* (pp. 493–500). IEEE.
18. Legner, C., Pentek, T., & Otto, B. (2020). Accumulating design knowledge with reference models: Insights from 12 years' research into data management. *Journal of the Association for Information Systems*, 21(3).
19. Li, W., & Law, K. L. E. (2024). Deep learning models for time series forecasting: A review. *IEEE Access*, 12, 92306–92327.
20. Bandi, V. D. V. K. Autonomous Data Platforms: Converging AI, MLOps, and Cloud Engineering for Digital.
21. Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), Article 20200209.
22. Liu, K., Yang, M., Li, X., Zhang, K., Xia, X., & Yan, H. (2022). M-data-fabric: A data fabric system based on metadata. In *2022 5th International Conference on Big Data and Artificial Intelligence* (pp. 57–62). IEEE.
23. Macías, A., Muñoz, D., Navarro, E., & González, P. (2024). Data fabric and digital twins: An integrated approach for data fusion design and evaluation of pervasive systems. *Information Fusion*, 103, Article 102139.
24. Potturi, P. (2026). Agentic Business Intelligence Architecture for Autonomous Data Exploration, Insight Generation, and Strategic Decision Orchestration. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(22s), 536–542.
25. Rieyan, S. A., News, M. R. K., Rahman, A. B. M. M., Khan, S. A., Zaarif, S. T. J., Alam, M. G. R., Hassan, M. M., Ianni, M., & Fortino, G. (2024). An advanced data fabric architecture leveraging homomorphic encryption and federated learning. *Information Fusion*, 102, Article 102004.



26. Song, X., Deng, L., Wang, H., Zhang, Y., He, Y., & others. (2025). Deep learning-based time series forecasting. *Artificial Intelligence Review*, 58, Article 23.
27. Mahadevan, S. (2026). Governed Serverless Automation Ecosystem for AI-Driven Enterprise Integration and Sustainable Cloud Operations. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(16s), 1561-1570.
28. Su, L., Zuo, X., Li, R., Wang, X., Zhao, H., & others. (2025). A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review*, 58, Article 80.
29. Hall, T., & Rasheed, K. (2025). A survey of machine learning methods for time series prediction. *Applied Sciences*, 15(11), Article 5957.
30. Abu Rumman, A., & Al-Abbadi, L. (2023). Structural equation modeling for impact of data fabric framework on business decision-making and risk management. *Cogent Business & Management*, 10(2), Article 2215060.
31. Bode, J., Kühn, N., Kreuzberger, D., & Hirschl, S. (2023). Data mesh: Motivational factors, challenges, and best practices. *arXiv preprint*.
32. Movva, V. R., & Pareek, P. K. (2026). Multi-Agent AI Architecture for Autonomous Retail Operations: Coordinating Intelligent Agents Across Cloud, DevOps, and Enterprise Platform. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(23s), 925-932.
33. Wider, A., Verma, S., & Akhtar, A. (2023). Decentralized data governance as part of a data mesh platform: Concepts and approaches. In *2023 IEEE International Conference on Web Services*.
34. Ye, H., Zhang, N., Chen, H., & Chen, H. (2022). Generative knowledge graph construction: A review. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
35. Alsharef, A., Aggarwal, K., Sonia, Mishra, A., & Kumar, M. (2022). Review of machine learning and automated machine learning solutions for time-series forecasting. *Archives of Computational Methods in Engineering*, 29, 5297–5311.
36. Song, H., Rajan, D., Thiagarajan, J. J., & Spanias, A. (2020). Attend and diagnose: Clinical time series analysis using attention models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 4091–4098.
37. Loganathan, R. (2026). SaaS Entitlement Governance: A Reproducible Model for Detecting and Reclaiming Over-Provisioned Access at Enterprise Scale. *Journal of Intelligent Decision Making and Information Science*, 3(7s), 2519-2530.
38. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764.
39. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). Transformers in time series: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (pp. 6778–6786).