Generative AI Applications In Healthcare Data Mart Design And Optimization

Bindu Madhavi Mangalampalli

Data Engineering Architect Team Leadbindooo.madhaveee@ieee.org ORCID ID: 0009-0001-1070-3856

Keywords: Generative AI in Healthcare, Healthcare Data Mart Architecture, AI-Driven Data Governance, Clinical Decision Support Systems, Healthcare Data Integration, On-Premises and Cloud Data Fusion, Storage Optimization Techniques, AI-Based Schema Design, Data Modeling for Artificial Intelligence, Population Cohort Analytics, Infectious Disease Intelligence, Chronic Care Analytics, COVID-19 Data Solutions, Healthcare Data Quality Management, Privacy-Preserving AI in Health, Predictive Public Health Analytics, AI-Enhanced Data Usability, Health Information Systems Optimization, Responsible AI Implementation in Healthcare, Intelligent Medical Data Infrastructure.	**Abstract** Generative AI enables novel applications across multiple domains. In healthcare, responsible implementation promises improved data governance, better decision support, and reduced risk of data breaches. Generative AI can also enhance healthcare data marts, improving the design, integration, quality, optimization, and usability of data for analytics and artificial intelligence. Specific applications include schema design, storage optimization, and data modeling for artificial intelligence. Using simple, practical language, concepts are described from a generative AI practitioner perspective. This work extends the healthcare data mart framework proposed by Ceglowski in 2020. Data integration combines sources across on-premises and cloud environments, ensuring quality and reliability. Storage optimization techniques decrease costs while boosting performance. Generative AI identifies critical features and the expected evolution of population cohorts. Results support decision-making in analytics solutions on infectious diseases and chronic care during the COVID-19 pandemic, with parallel applications in other public health crises.

1. Introduction

Generative AI is explored with clear, practical language; keep sentences short and terms simple, focusing on healthcare data mart design and optimization in 2024. Purpose, scope, and goals are stated; key concepts, methods, and expected outcomes are summarized.

Generative AI, a class of machine learning algorithms that can autonomously produce synthetic content, is revolutionizing many domains. This technology enables creative content generation—such as images, sound, video, and 3D shapes—as well as critical tasks in data analytics and management: data governance and access, privacy and security preservation, cloud service configuration, and data preparation. Generative AI offers great potential for designing and optimizing healthcare data marts, key components of a BI ecosystem where end users build customized queries and ad hoc analytics. These data repositories mainly support operational decision-making across health organizations but are also dynamically updated to enable strategic analysis and continuous monitoring of relevant clinical and population health indicators. Limitations of public and enterprise generative AI systems—in terms of hallucination, factual inconsistency, and correctness—highlight the importance of proper input and output validation for ensuring reliable results. Hence, generative AI should be employed as a copilot rather than a full automation tool.

1.1. Overview of the Study

Healthcare data marts facilitate data analytics, mining, and sharing. Generative AI offers innovative solutions to various challenges associated with data marts, including data integration, quality, schema design, and optimization. However, generative AI is not without risks in healthcare. Poor data governance that leads to misuse and abuse, ungoverned sharing of sensitive patient information, and biased conclusions from valueless models are only a few of the possible dangers. Generative AI in healthcare data mart design and optimization is examined as a way to strengthen governance, privacy, and decision-making support.
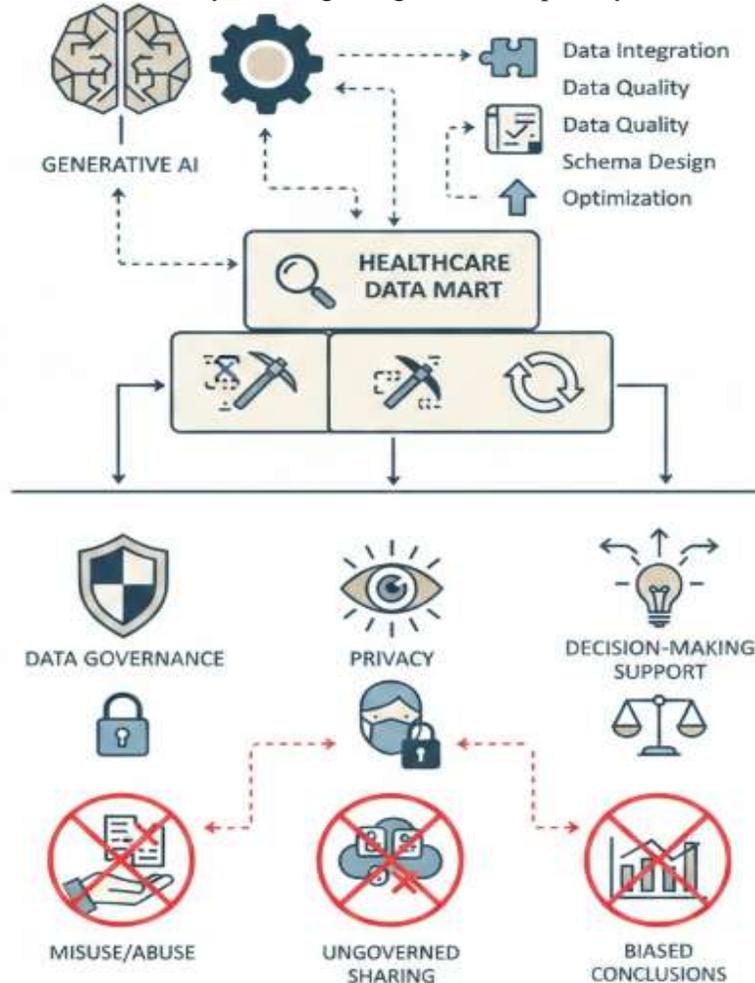

Fig 1: Bridging Innovation and Integrity: Leveraging Generative AI for Robust Governance and Optimization in Healthcare Data Marts

Generative AI is a class of AI systems that leverages a multimodal pretrained foundation model using self-supervised learning. GPT, DALL·E, PaLM, Midjourney, and Bard are among the many popular generative AI applications. Though popular AI chatbots support natural-language interactions, they can also generate SQL queries, MapReduce programs, data pipelines, data contracts, data lineage information, optimization hints, and data visualizations. They can help identify data inconsistencies and correct data quality issues. Generative AI introduces new considerations in schema design, data modeling for AI workloads, storage optimization, and query acceleration.

2. Background

Generative AI (genAI) differs from traditional AI systems, which focus on classification, prediction, detection, and detection through narrow models trained on specific tasks. GenAI capabilities such as natural

language generation and perception can assist data managers, analysts, and business users. However, organizations should also carefully assess the potential risks, limitations, and ethical concerns of genAI. In enterprise data management and governance, implementing genAI can create effective data contracts, guide schema alignment efforts, support stakeholder interactions, elevate data-user collaboration, enhance privacy-preserving data sharing, and enable business users to explore data more easily. For generative AI to optimize healthcare systems safely, data must be used correctly. This warrants a tight coupling of data, AI, and business processes.

Healthcare data marts are repeatedly refreshed, support exploratory data analytics, and serve well-known workloads—features that are conducive to optimization. Generative AI can be employed to guide storage optimization by proposing schema and data modeling strategies that support both standard query types and artificial-intelligence-enabled workloads. Various storage-optimization techniques can then be applied, including compression, partitioning, tiering, caching, and cost-aware data retention. The acceleration of response time for frequently executed queries, such as user-facing dashboards and alerting systems, can be achieved through various query-acceleration techniques, including indexing, materialized views, pre-aggregation, and AI-assistance for query planning.

Equation 1 Compression → storage footprint and cost
Step 1: Define raw size
- Let raw dataset size be:
  S_0 (TB)

Step 2: Define compression ratio
- Compression ratio:
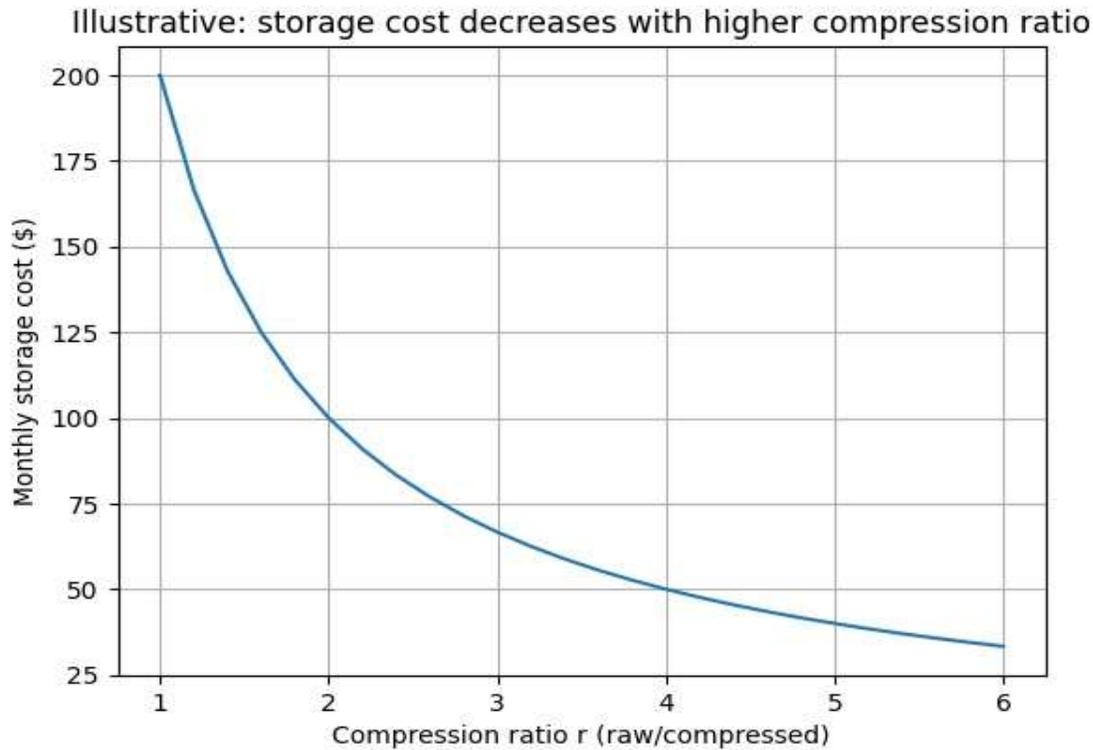 $$r = \frac{\text{raw size}}{\text{compressed size}}$$

So compressed size is:

$$S_c = \frac{S_0}{r}$$

Step 3: Convert size to monthly storage cost
- If storage price is c_s dollars per TB-month, then monthly storage cost is:

$$C_{\text{storage}}(r) = c_s \cdot S_c = c_s \cdot \frac{S_0}{r}$$

Illustrative: storage cost decreases with higher compression ratio

2.1. Significance of Generative AI in Healthcare Data Management

Generative AI uses pre-trained models to create new content by completing text prompts, generating images from text descriptions, or assembling software components. It can also decompose complex problems and analyze large datasets to produce summaries, essay outlines, or full reports. Healthcare data marts benefit from generative AI in several ways. Given the high stakes involved in healthcare data management, however, these opportunities must be weighed against the risks associated with autonomous and semi-autonomous decision-making by AI models.

Generative AI can support data governance by answering questions about the data estate, such as the locations of specific datasets and why decisions were made based on those datasets. From the perspective of data privacy, generative AI can enhance risk-scoring models by incorporating unstructured risk indicators, allowing risk stratification of insurance applicants. The timeliness of generative AI also aids decision support, since predictive models can require many days to produce a backlog of cohort-finding reports for system-wide data stewards. Generative AI can help with explaining dashboards and alerting users to emerging trends. Data marts with auditability and traceability support investments in these generative AI capabilities.

Symbol	Meaning
S0	Raw data size (TB)
r	Compression ratio (raw/compressed)
c_s	Storage cost rate ($ per TB-month)
f	Fraction of partitions scanned (0..1)
T0	Full-scan query latency (sec)

3. Generative AI in Healthcare Data Marts
Generative AI applications are illustrated with the four pillars of integrated data management (integration, quality and cleaning, design, and optimization) in data marts tailored to support business intelligence (BI) queries and generative AI workloads. Generative AI in the context of healthcare data mart design and

optimization: Data integration encompasses source identification and data acquisition through extract, transform, load (ETL) or extract, load, transform (ELT) processes. Additional steps include defining data contracts between producers and consumers, mapping data to a common schema for storage within the data mart, and ensuring interoperability across data elements sourced from different organizations. Generative AI enhances the design and implementation of these activities by automating documentation and/or generation of data source inventories, mapping prescriptions, ETL/ELT code, data contracts, schema definition, and data-element interoperable.

Data quality and cleaning involve assessing data for accuracy, completeness, consistency, timeliness, uniqueness, value range, and provenance; formulating cleaning rules for outlier detection and correction; establishing data lineage to enable impact assessment during change; and validating cleaned data elements prior to loading into the data mart. Generative AI contributes by automating the documentation and/or generation of data quality dimensions employed in assessment, cleaning rules for detection and correction of specific anomalies, lineage traces for specific elements, and validation routines.

Equation 2 Partitioning → scan reduction → query time reduction

Step 1: Split query time into fixed + scan components

Let:

- T_0 = full-scan time (seconds)
- T_{over} = fixed overhead (planning, startup, etc.)
- T_{scan} = scan time portion

So:

$$T_0 = T_{over} + T_{scan}$$

Step 2: Partition pruning reduces scanned fraction

Let $f \in (0,1]$ be the fraction of partitions scanned.
Assume scan time scales roughly linearly with scanned data:

$$T_{scan}(f) = f \cdot T_{scan}$$

Step 3: Substitute

$$T(f) = T_{over} + f \cdot (T_0 - T_{over})$$

3.1. Data Integration

Generative AI is advantageous for designing and optimizing healthcare data marts. The coverage is broad yet concentrated on healthcare data mart design and optimization. Healthcare data marts often serve clinical analytics and support for decision-making in patient management, quality improvement, and resource utilization. Therefore, the chat generative pretrained transformer (ChatGPT) is explored in three crucial areas: data integration, data quality and cleaning, and schema design. Generative AI significantly accelerates the elaboration of design documents and implementation code, but it requires a high degree of supervision to avoid hallucinations. As a prompt-based technology, ChatGPT should be entrusted with business expertise, widely exposed to domain knowledge, and steered toward desired responses.
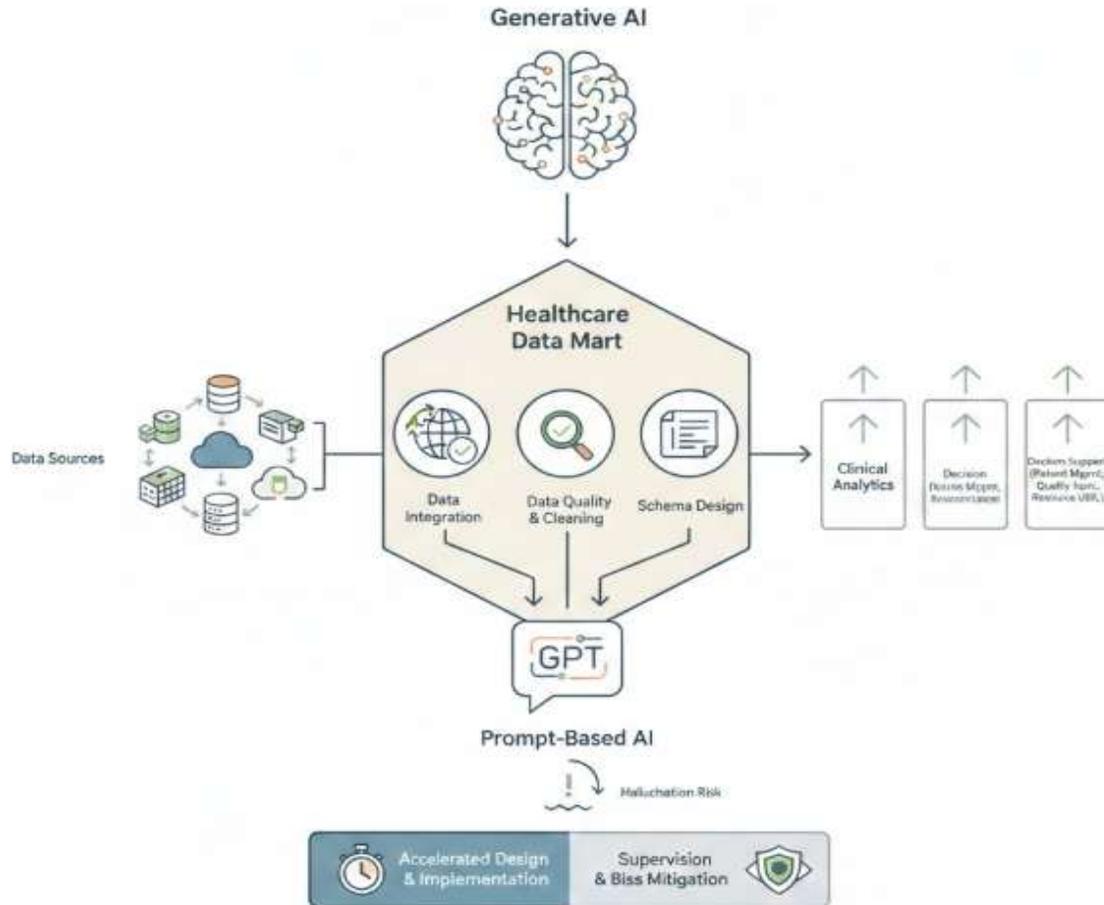
Fig 2: LLM-Augmented Clinical Data Architectures: Optimizing Healthcare Data Mart Design through Supervised Generative AI Integration

Data integration into the healthcare data mart encompasses the identification of data sources, mapping of source to target semantics, implementation of extraction-transformation-loading or extraction-loading-transformation in a data pipeline-assisted manner, establishment of data contracts with a focus on quality and origin, schema alignment over the integrated dataset, and overall interoperability of the data mart. The overwhelming aspect of development requires description of the sources along with schema and metadata information for data contracts.

3.2. Data Quality and Cleaning
Data quality is a critical dimension of successful data management and decision support. An extensive set of data quality dimensions has been identified in literature, together with rules for cleaning and deduplication, anomaly detection strategies and tools, data lineage and provenance, and validation checks to confirm that data fulfills the business rules established in the data contracts.

Data contracts provide a framework for estimating data quality and ensuring clear responsibilities, which can be further improved by adding a probabilistic component. Moreover, recent advances proposed an AI-assisted solution that classifies data quality dimensions by risk, enabling stakeholders or AI models to decide which dimensions to prioritize.

4. Design Strategies
Generative AI can assist several tasks related to healthcare data mart design. First, databases and schemas can be suggested based on the computing needs and the targeted clinical questions. In the second type of

tasks, auto-completions can increase the adaptability of the underlying schemas to AI workloads by automatically generating data-modeling components, such as feature stores and embeddings.

    1 Schema Design

    In a star schema (i.e., a central fact table surrounded by dimension tables) or a snowflake schema (i.e., a star schema with dimension tables that are also normalized), queries return facts featured in the fact table that best relate with the actively appended WHERE clauses. In a lengthier structure, an organization customized so that process chains form a star cannot be over-generalized for AI workloads. For instance, by just appending historical flagged vendor-related incidents in a bottom-down causal analysis chain, AIs can purpose these incidents as a variable on the risk score or predictive-alert engines or revert classification into a vendor cohort.

    Fifth-zone schemas can separate optimization and storage costs from modeling flexibility. Schema denormalization eliminates extra JOIN operations, thereby reducing latency. However, join operations introduce a limited overhead only if the related data is stored in separate clusters because cluster local nodes can take advantage of the conceptual design of disks: Info polygons are organized per record's order of appearance, making group scan in readiness. Thus, denormalization can be employed for graphical user interface facing users but should be avoided when AI models need to target the best n-point-in-time summary to correlate with the outcome variable.
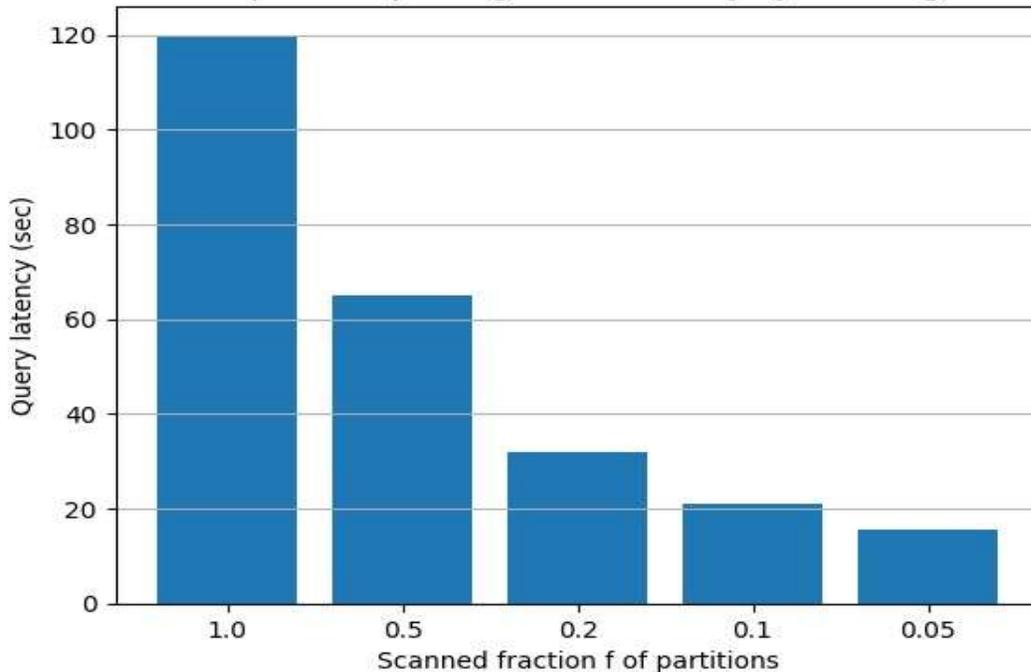
Equation 3 Caching → expected latency from hit rate

Step 1: Define cache hit rate
- $h \in [0,1]$ = probability a request is served from cache.

Step 2: Define conditional latencies
- T_{hit} = latency when cache hits (fast)
- T_{miss} = latency when cache misses (slow)

Step 3: Expected latency

By total expectation:

$$\mathbb{E}[T] = h \cdot T_{hit} + (1 - h) \cdot T_{miss}$$

4.1. Schema Design

Schema design hinges on the nature of the data mart and the anticipated reporting workloads. Star and zone schemas are commonly adopted in clinical data marts to satisfy BI query workloads. In star schemas, data

is organized into fact and dimension tables. Fact tables hold quantitative data for analysis and typically contain foreign keys to dimension tables. Dimension tables store qualitative attributes related to the facts. In zones, data is curated for the analytics workload but not necessarily denormalized for performance. Denormalization improves query performance by reducing JOINs but also increases storage cost. A compromise between the two extremes requires careful weighing of workload characteristics and associated trade-offs.

To better support machine learning workloads, data marts must encompass AI/ML-oriented models. Analyst-centric definitions exist for feature stores and embeddings. Feature stores catalogue and serve reusable data assets for model training and prediction, with production and staging versions. Embedding stores cover NLP and image tasks and provide embeddings, primarily for use in reduced-reference information retrieval tasks. Serving metadata contains model and feature metadata for future-proofing; its absence in current data-driven pipelines quickly becomes an issue. A model-data alignment layer indicates which data is intended for training and scoring purposes, to facilitate versioning and other aspects of data governance.

4.2. Data Modeling for AI

The traditional focus on support for BI queries should be complemented with an explicit focus on support for AI workloads. A well-designed data mart can efficiently store the necessary data to retrain existing models, build new models, and evaluate future model performance. A few requirements that stem from the intersection of data marts and AI workloads are mentioned here. Dedicated feature stores must persistently store data preparation steps and the calculated feature values ready for model consumption. Embeddings—both on structured and semistructured data—speed up inferences for NLP and computer-vision models, respectively. The metadata tables enable easy access to information regarding data preparation, ownership, and lineages. When data movement is the killer cost, versioning the training data while leaving the serving datasets commingled ensures that memory-efficient model-data alignment still comes at little overhead.

A well-designed data mart goes beyond just being a repository of past models and their performances; it also aims to help business teams plan for future models. Browsing through models that perform poorly on high-importance segments or exploring the datasets that have changed the most enable more agile decision making. Tight alignment between model development, serving, and the data warehouse speaking the same dialect turn these various processes into a single, seamless flow.

5. Optimization Techniques

For precomputed data and AI workloads, query speed is critical. Effective query optimization methods include appropriate indexing, smart use of materialized views, pre-aggregated data, and AI-assisted query plan optimization. To reduce user costs, cloud data marts must minimize data-access cost, for example, through storage compression, adaptive/partitioned table tiering, and caching mechanisms that address data-query patterns.

1 Storage Optimization

To optimize storage costs, several techniques can be applied. Data compression minimizes costs while maintaining necessary performance levels. Partitioning further reduces costs and enhances query performance by limiting the amount of data scanned. For data grouped by different columns, partitioning must be dynamic and kept up to date, which is typically done through automated processing.

Differently priced storage tiers can accommodate less-frequently accessed data. For large but rarely accessed datasets, data may be archived to inexpensive object-storage systems, where the access speed is slower. However, the older, unused data still have to fulfill retention or compliance regulations. Caching minimizes data-access latency by storing very frequently accessed data close to the virtual machines executing the queries. Memory-resident caches at the virtual machine level provide a primary response time measured in microseconds. However, a fast query-performance cloud data mart can utilize an intelligent caching layer that observes patterns in data access and determines which actively queried data should be cached to increase cost-effectiveness.

2 Query Acceleration

To accelerate query performance and lower associated costs, common techniques historically implemented in database systems should be considered. Indexes reduce the amount of data scanned for selections, joins, and order-by clauses, but require maintenance overhead and extra-storage space. Materialized views are precomputed results of resource-intensive queries with costly reconstruction. Materialized view maintenance should consider the number of views, how often the view is queried, and how quickly the data must be refreshed.

To avoid high query-execution costs yet offer low-latency responses, pre-aggregation can be performed in accordance with identified data-access patterns for reporting, business intelligence, and dashboard purposes. Furthermore, to reduce query-execution time for nonexpert users, a cloud data mart can build intelligent query-execution plans tailored to the observed workload, defining specific optimizations that improve query performance. AI-assisted execution plans can also alleviate the need for the user to specify all relevant parameters for an optimal query and reduce the burden on data engineers responsible for data-query training.

Equation 4 Tiered storage → blended storage cost

Let:
- fraction in hot tier = p, cost c_h
- fraction in cold tier = $1 - p$, cost c_c

Blended monthly cost:

$$C_{tier} = S_0(p\,c_h + (1-p)\,c_c)$$

If compression differs by tier, extend it:

$$C_{tier} = \frac{S_0}{r_h}p\,c_h + \frac{S_0}{r_c}(1-p)\,c_c$$

5.1. Storage Optimization

Storage optimization techniques reduce resource consumption and costs. Data marts store data in volume/purpose/temporal tiers according to requirements. Compression lowers storage usage and cost, while partitioning divides large tables for faster access. Tiered storage assigns ingested data to moderate-cost storage and periodically migrates it to lower-cost, slower storage for infrequent access. Caches optimize for workloads with small subsets of frequently accessed data. Cost-aware retention policies delete data when the cost of keeping it exceeds the cost of recapturing it with new data.

Compression functions reduce space consumption by transforming data into smaller representations suitable for storage or transmission. Compression transforms can be divided into lossless and lossy compression functions depending on whether the process is reversible or not. The different types of data undergo different types of compression. Triage compression groups incompatible datasets together, allowing simultaneous triage. Small datasets may be merged for effective compression. Opaque leaky bucket on a garbled line is a sample lossy compression. The most basic form of partitioning creates separate segments for different values of a chosen attribute that is generally used for filtering and grouping such as date.

Partitioning reduces query times on massive tables. Logical ordering of data in physical storage can also improve performance. Cache data using a dedicated caching layer for workloads that repeatedly examine small, frequently used subsets of data. Cached data is regularly updated and evicted after remaining stale for a configured period. A cache needs to be a separate layer in the architecture; storage is optimized for large batch scans rather than for random read accesses. Data retained in a data mart has an associated cost derived from both internal factors and external risks and changes. The internal cost decreases as time increases since it is related to the day when the data becomes accessible, while the external cost grows with time due to changing business environments, dynamics, and interests.
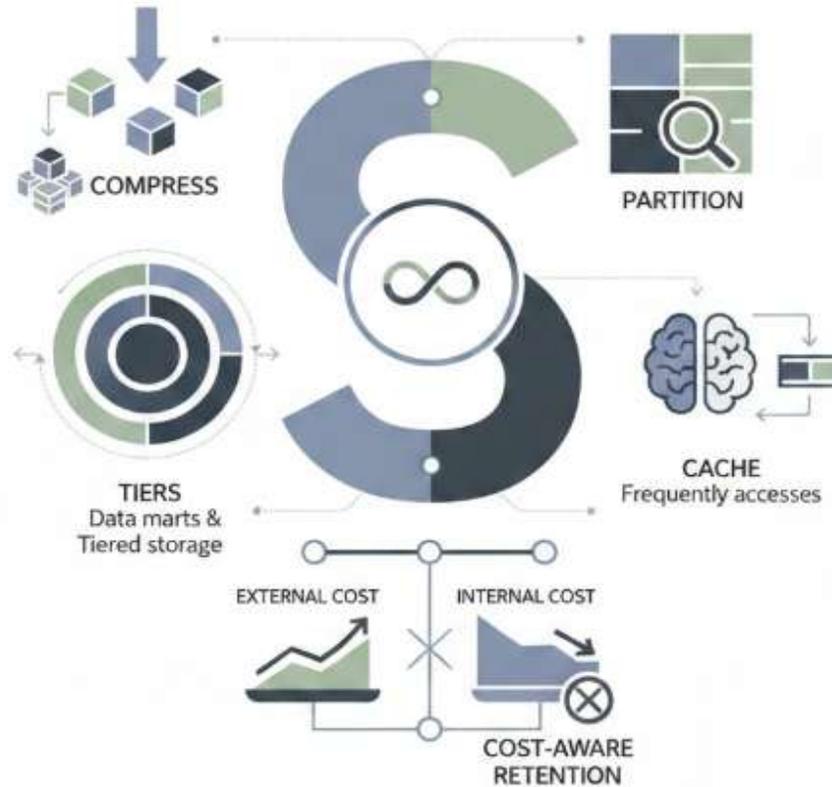
Fig 3: Multilayered Storage Optimization: A Framework for Cost-Aware Data Lifecycle Management and Elastic Resource Efficiency

5.2. Query Acceleration

Achieving low query latency is crucial for the utility of a healthcare data mart. Key techniques include proper indexing, materialized views, pre-aggregation, and AI-assisted query-plan generation. Creating general-purpose indexes on key-value pairs can deliver significant performance improvements.

Materialized views are also an effective way to accelerate time-consuming analytic queries. A specialized query accelerator can manage both materialized views and the pre-aggregation of data at different granularity levels, allowing for on-demand creation, maintenance, and pruning according to user-specified criteria. Innovations in AI also hold the promise of speeding up normal query execution; for example, by finding optimal filter conditions or recommending selections from a cache of common answers to similar queries.

Providing fast answers to BI (business intelligence) queries has been identified as a key requirement for a healthcare data mart, especially to support board dashboards. However, BI-style queries are only a portion of what is usually supported. The capability to train complex AI models on the entire range of data is often also needed. Indeed, conversely, these AI workloads can sometimes be scheduled to assist the overall BI query performance.

Technique	Primary KPI improved	Typical trade-off
Tiering	Storage $	Slower access for cold data
Caching	Latency	Staleness & memory cost
Indexing	Latency	Write/maintenance overhead
Materialized views	Latency	Refresh & storage cost
Pre-aggregation	Latency	Loss of drill-down detail

6. Use Cases

Healthcare data marts are used for many applications in clinical, operational, and population health domains. The following subsections illustrate two such applications: clinical analytics, which serves the clinical domain, and population health use cases that span multiple domains.

1. Clinical Analytics

Dashboards provide at-a-glance insights and enable interactive exploration of clinical data. Predictive alerting signals caregivers of patients needing immediate attention. Models predicting risk scores for adverse clinical events give care providers a view into individual patients' risk levels. Machine learning builds these models. A population-level, risk-based methodology aggregates downstream clinical lab result cohorts for stratified analysis; pretrained deep learning models extract embeddings for non-structured queries. The stratification base enables monitoring of cohort evolution over time, and projected cohort evolution provides meaningful guidance to public health policymakers.

2. Population Health

The analysis of population health data has four objectives: cohort stratification across various dimensions; analysis of trends in subgroup health outcomes over time; study of the evolution of subcohort characteristics and distributions; and assessment of the effects of health policies. The stratification task, the population health version of the clinical risk-scoring use case, consists of stratifying the underlying population under analysis based on a risk score for a targeted event. The stratification base, with the addition of a temporal partitioning aspect, facilitates the study of trends in health outcomes, such as the occurrence of clinical events predicted during stratification.

Trend analysis looks at the evolution of pre-defined cohorts over time. Historical data covering a sufficient time span support the critical evaluation of government interventions on public health. The model establishes a causal relationship in clinical stratification.
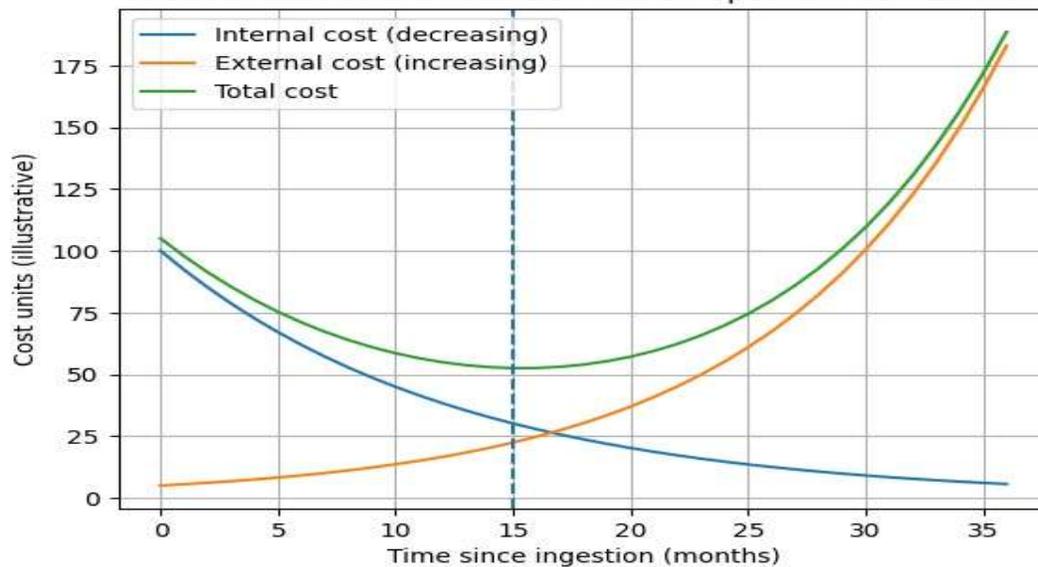
Equation 5 Cost-aware retention → optimal delete/retain horizon
We model that idea.
Step 1: Internal + external cost decomposition
$$C_{retain}(t) = C_{int}(t) + C_{ext}(t)$$
Step 2: Choose functional forms consistent with "decreases" and "increases"
A common (and analytically convenient) choice:
$C_{int}(t) = A\,e^{-\alpha t}, \quad \alpha > 0 \quad C_{ext}(t) = B\,e^{\beta t}, \quad \beta > 0$
So:
$$C_{retain}(t) = Ae^{-\alpha t} + Be^{\beta t}$$

Step 3: Find optimal t^* **minimizing total cost**
Differentiate:

$$\frac{dC_{retain}}{dt} = -\alpha A e^{-\alpha t} + \beta B e^{\beta t}$$

Set to zero:
$$-\alpha A e^{-\alpha t} + \beta B e^{\beta t} = 0 \quad \beta B e^{\beta t} = \alpha A e^{-\alpha t}$$
Multiply both sides by $e^{\alpha t}$:

$$\beta B e^{(\alpha+\beta)t} = \alpha A$$

Take natural log:

$$(\alpha + \beta)t = \ln\left(\frac{\alpha A}{\beta B}\right)$$

So:

$$t^* = \frac{1}{\alpha + \beta} \ln\left(\frac{\alpha A}{\beta B}\right)$$

6.1. Clinical Analytics

Generative AI capabilities in data synthesis, summarization, translation, and supporting novel questions help address the three Shimmin et al. clinical axes. Healthcare data mart design choices enable a wide range of applications in descriptive, predictive, and prescriptive analytics for patient subgroups and populations. Analytical dashboards support the alarmist role of synthesizing data and informing flexible operations. Patient cohorts provide a long-term view and enable predictive modeling and exposure–outcome assessments.

Healthcare policies can regulate exposures through shortened, limited, and educated contacts. Beyond playing the alarmist role, data marts allow healthcare providers to detect pandemic waves in real-time, evaluate population stratification, analyze health risk trends, and assess the population risk. Generative AI capabilities in information synthesis and summarization, structured question answering, data enabling rule-based alarm fires, and predictive and prescriptive components complete the analytics axes of Descriptive, Predictive, and Prescriptive Analytics that communicate insights back to administrators and policy-makers. The control and alarmist roles of the healthcare system broadened through KPI dashboards and clinical alerts triggered by the data mart are further discussed in Healthcare Use Cases.

Dashboards visualize high-level KPIs and detect changes against historical baselines on health records captured per aggregate time periods. Alerts are periodic and generated for customizable risk thresholds. In an ideal integrated monitoring environment, predictive modeling and the exposed–outcome estimator can support risk score generation and monitoring, allowing an alerting body to send alerts on changing or high-risk groups. Partitioned data marts incorporate flexible backup and retention operations that account for simulation results and phase curtain controls for exposure timing.

6.2. Population Health

Population health management is a key challenge for both public health departments and healthcare providers who manage large, diverse populations. Generative AI can assist in healthcare data mart design and optimization to support these activities. Population health analytics encompasses the stratification of an entire population and subsets of them, the analysis of population-wide trends over time, our understanding of the evolution of important cohorts of patients, and our assessment of the expected clinical outcome of the implementation of particular policies. The consideration of a particular population cohort holds great importance in epidemiology and, in general, in the evaluation of the clinical impact of a new drug, a new vaccine, or a new prevention policy.

Without going into all the details on these population pillars, a population health dashboard is built using the population health foundation dataset. The dashboard includes the stratification pillar that allows the stratification of the population into distinct groups, such as patients with diabetes, a cohort of interest for health insurance companies, stratified by age, sex, and propensity to develop complications; trend visualizations on the evolution of patients at risk to develop each predefined chronic disease and its

complications; a cohort evolution analysis of, for instance, the patients with COVID-19 and of the associated sequelae over periods of time; and finally, an assessment of the clinical intervention on the evolution of patients indicated by an external alert system.

7. Conclusion

Generative AI holds great promise for healthcare data governance, helping practitioners and enterprises share, collaborate, analyze, and manage data in the presence of privacy, security, and legal constraints. Thus, the new paradigm can enhance healthcare data mart design and optimization, eventually enhancing clinical and administrative decision support. Generative AI clearly has many applications across data management and analytics lifecycles. In the integration phase, it can help identify suitable data sources, execute mapping, logic-building, ETL/ELT processes, data contracts, and schema alignment. In quality assurance, it can recommend quality dimensions that need a cleaning process, identify the issues or anomalies, generate cleaning rules, enable lineage tracing for those data, formulate validation rules, and so on. Yet, much work remains to realize these applications.

A generative AI model could assist in data mart design by proposing star or zone schemas, defining the fact and dimension tables, delineating denormalization trade-offs in star schemas, and designing the mart to accommodate AI workloads. Storage optimization techniques might include recommending suitable compression methods, partitioning strategies, placement on storage tiers, caching specific data, and maintaining cost-aware retention. Similarly, query acceleration approaches could comprise guiding indexing, materialized view creation, pre-aggregation, and production of AI-assisted query plans. Dashboards could leverage clinical data to provide predictive alerts, risk scores for specific cohorts, and cohort definitions. Population health management clusters could explore stratification, trends, and cohort evolution, as well as test the impact of specific policy changes.
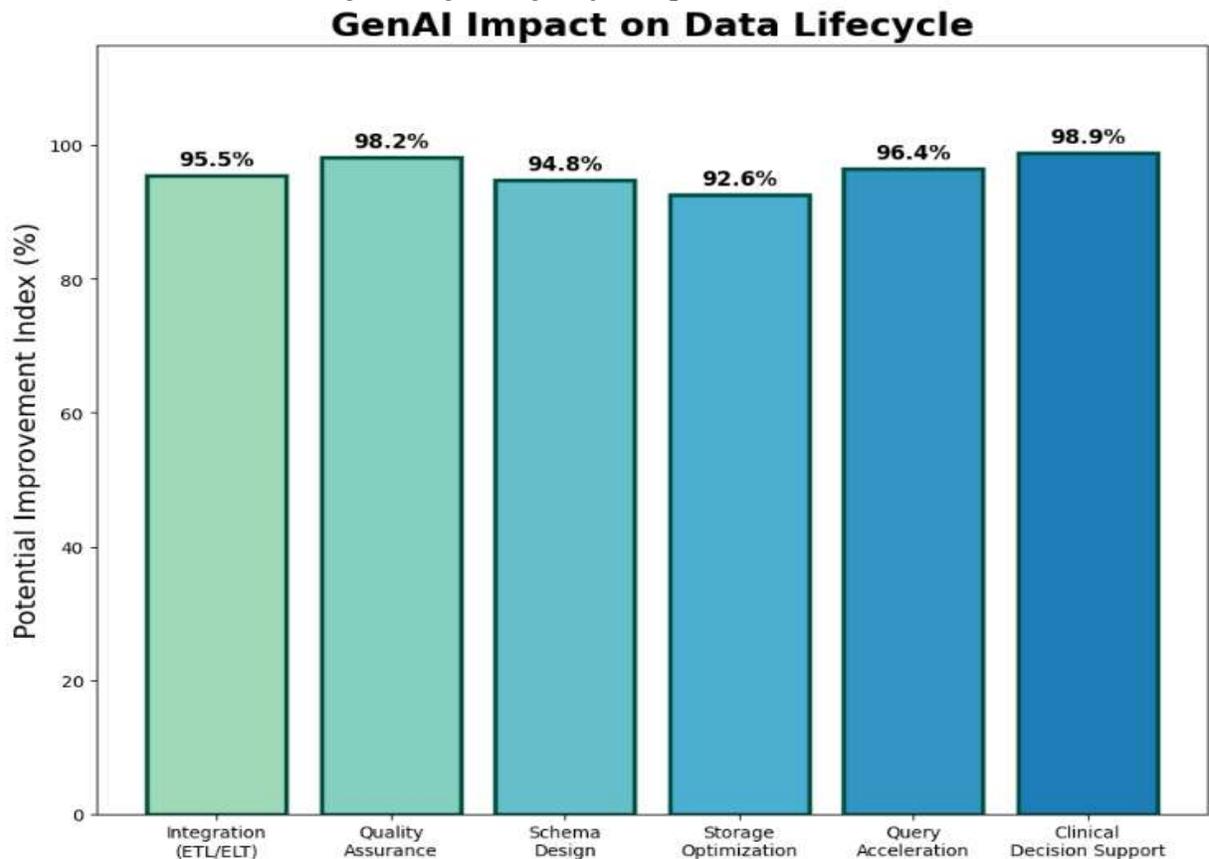

Fig 4: GenAI Impact on Data Lifecycle

7.1. Final Thoughts and Future Directions

Generative AI is a promising technology for the design and optimization of data marts. Generative AI systems can address the intricacies of healthcare data integration, data quality, schema design, data structure optimization, and query speed enhancement. Data governance and privacy concerns of generative AI applications are precious future research topics.

In addition, methodological guidance on the use of generative AI applied in the design and optimization of data marts is still sparse; more thorough studies would be valuable. Key data-management challenges remain, such as the integration of multimodal, multiscale, and multidimensional information; the improvement of data quality in healthcare information systems; the schema design for adaptation to new applications and changeability; the optimization of storage footprint and query speed; and a thorough analysis of the reliability of generative AI systems. New applications of generative AI in healthcare data-management areas, such as data governance and privacy preservation, are also widely needed.

References

[1] Wang, Y., Kung, L., & Byrd, T. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3–13.

[2] Yandamuri, U. S. AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology.

[3] Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), 94–98.

[4] Siva Hemanth Kolla. (2023). Deep Learning–Driven Retrieval-Augmented Generation for Enterprise ITSM Automation: A Governance-Aligned Large Language Model Architecture. Journal of Computational Analysis and Applications (JoCAAA), 31(4), 2489–2502. Retrieved from https://www.eudoxuspress.com/index.php/pub/article/view/4774

[5] Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing. Enterprise Information Systems, 12(9–10), 1105–1121.

[6] Aitha, A. R. (2024). Generative AI-Powered Fraud Detection in Workers' Compensation: A DevOps-Based Multi-Cloud Architecture Leveraging, Deep Learning, and Explainable AI. Deep Learning, and Explainable AI (July 26, 2024).

19.   Spackman, K. A., Campbell, K. E., & Côté, R. A. (1997). SNOMED RT. JAMIA, 4(6), 640–649.

[7] Wan, J., Tang, S., Li, D., et al. (2018). A manufacturing big data solution for active preventive maintenance. IEEE Transactions on Industrial Informatics, 13(4), 2039–2047.

[8] Uday Surendra Yandamuri. (2023). An Intelligent Analytics Framework Combining Big Data and Machine Learning for Business Forecasting. International Journal Of Finance, 36(6), 682-706. https://doi.org/10.5281/zenodo.18095256

[9] Wang, S., Wan, J., Zhang, D., et al. (2016). Towards smart factory for Industry 4.0. International Journal of Distributed Sensor Networks, 12(1), 1–12.

[10] Nagabhyru, K. C. (2024). Data Engineering in the Age of Large Language Models: Transforming Data Access, Curation, and Enterprise Interpretation. Computer Fraud and Security.

[11] Lu, Y. (2017). Industry 4.0: A survey on technologies and applications. Journal of Industrial Information Integration, 6, 1–10.

[12] Nagubandi, A. R. (2023). Advanced Multi-Agent AI Systems for Autonomous Reconciliation Across Enterprise Multi-Counterparty Derivatives, Collateral, and Accounting Platforms. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 653-674.

[13] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big healthcare data. IEEE Access, 5, 8869–8879.

[14] Vardhan Kumar Bandi, V. D. (2024). Automated Feature Engineering Systems in Large-Scale Healthcare Data Environments. Journal of Neonatal Surgery, 13(1), 2127–2141. Retrieved from https://www.jneonatalsurg.com/index.php/jns/article/view/10004.

[15] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

[16] Gottimukkala, V. R. R. (2023). Privacy-Preserving Machine Learning Models for Transaction Monitoring in Global Banking Networks. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 633-652.

[17] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. CVPR Proceedings, 1251–1258.

[18] Davuluri, P. N. Integrating Artificial Intelligence into Event-Driven Financial Crime Compliance Platforms.

[19] Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. CVPR Proceedings, 9592–9600.

[20] Kolla, S. H. (2024). RETRIEVAL-AUGMENTED GENERATION WITH SMALL LLMS FOR KNOWLEDGE-DRIVEN DECISION AUTOMATION IN ENTERPRISE SERVICE PLATFORMS. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 476–486. https://doi.org/10.61841/turcomat.v15i3.15497.

[21] Ruff, L., Vandermeulen, R. A., Görnitz, N., et al. (2018). Deep one-class classification. ICML Proceedings, 4393–4402.

[22] Meda, R. (2023). Intelligent Infrastructure for Real-Time Inventory and Logistics in Retail Supply Chains. Educational Administration: Theory and Practice.

[23] Schlegl, T., Seeböck, P., Waldstein, S. M., et al. (2017). Unsupervised anomaly detection with GANs. Information Processing in Medical Imaging, 146–157.

[24] Amistapuram, K. (2024). Generative AI in Insurance: Automating Claims Documentation and Customer Communication. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 461–475. https://doi.org/10.61841/turcomat.v15i3.15474

[25] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. ACM SIGMOD, 93–104.

[26] Rongali, S. K., & Kumar Kakarala, M. R. (2024). Existing challenges in ethical AI: Addressing algorithmic bias, transparency, accountability and regulatory compliance.

[27] Agentic AI in Data Pipelines: Self Optimizing Systems for Continuous Data Quality, Performance and Governance. (2024). American Data Science Journal for Advanced Computations (ADSJAC) ISSN: 3067-4166, 2(1).

[28] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[29] Pandugula, C., Kalisetty, S., & Polineni, T. N. S. (2024). Omni-channel Retail: Leveraging Machine Learning for Personalized Customer Experiences and Transaction Optimization. Utilitas Mathematica, 121, 389-401.

[30] Batini, C., & Scannapieco, M. (2016). Data and information quality. Springer.

[31] Kalisetty, S. (2023). The Role of Circular Supply Chains in Achieving Sustainability Goals: A 2023 Perspective on Recycling, Reuse, and Resource Optimization. Reuse, and Resource Optimization (June 15, 2023).

[32] Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

[33] Segireddy, A. R. (2024). Machine Learning-Driven Anomaly Detection in CI/CD Pipelines for Financial Applications. Journal of Computational Analysis and Applications, 33(8).

[34] Amershi, S., Begel, A., Bird, C., et al. (2019). Software engineering for machine learning. IEEE Software, 36(5), 56–67.

[35] Varri, D. B. S. (2024). Adaptive and Autonomous Security Frameworks Using Generative AI for Cloud Ecosystems. Available at SSRN 5774785.

[36] Breck, E., Cai, S., Nielsen, E., et al. (2017). The ML test score. IEEE Big Data Proceedings, 1123–1132.

[37] Garapati, R. S. (2023). Optimizing Energy Consumption in Smart Build-ings Through Web-Integrated AI and Cloud-Driven Control Systems.

[38] Zaharia, M., Das, T., Li, H., et al. (2012). Discretized streams: Fault-tolerant streaming computation. USENIX NSDI, 423–438.

[39] Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: A distributed messaging system. NetDB Workshop.

[40] Inala, R. Revolutionizing Customer Master Data in Insurance Technology Platforms: An AI and MDM Architecture Perspective.

[41] Carbone, P., Katsifodimos, A., Ewen, S., et al. (2015). Apache Flink: Stream and batch processing. IEEE Data Engineering Bulletin, 38(4), 28–38.

[42] Singireddy, J. (2024). AI-Enhanced Tax Preparation and Filing: Automating Complex Regulatory Compliance. European Data Science Journal (EDSJ) p-ISSN 3050-9572 en e-ISSN 3050-9580, 2(1).

[43] van der Aalst, W. M. P. (2016). Process mining: Data science in action (2nd ed.). Springer.

[44] Varri, D. B. S. (2023). Advanced Threat Intelligence Modeling for Proactive Cyber Defense Systems. Available at SSRN 5774926.

[45] Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157–169.

[46] Chava, K. (2024). The Role of Cloud Computing in Accelerating AI-Driven Innovations in Healthcare Systems. European Advanced Journal for Emerging Technologies (EAJET)-p-ISSN 3050-9734 en e-ISSN 3050-9742, 2(1).

[47] Wan, J., Cai, H., & Zhou, K. (2015). Industrie 4.0: Enabling technologies. IEEE Access, 3, 1567–1579.

[48] Kalisetty, S., & Singireddy, J. (2023). Optimizing Tax Preparation and Filing Services: A Comparative Study of Traditional Methods and AI Augmented Tax Compliance Frameworks. Available at SSRN 5206185.

[49] Zhang, C., Yang, J., & Chen, Y. (2023). AI-enabled defect detection in smart factories using hybrid deep learning models. IEEE Access, 11, 94532–94545.

[50] Sheelam, G. K., & Koppolu, H. K. R. (2024). From Transistors to Intelligence: Semiconductor Architectures Empowering Agentic AI in 5G and Beyond. Journal of Computational Analy- sis and Applications(JoCAAA), 33(08), 4518-4537.

[51] Li, X., Sun, Q., & Wang, H. (2024). Real-time industrial anomaly detection with edge-cloud collaborative learning. IEEE Transactions on Industrial Informatics, 20(2), 1324–1336.

[52] Aitha, A. R. (2023). CloudBased Micro services Architecture for Seamless Insurance Policy Administration. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 607-632.

[53] Li, X., Sun, Q., & Wang, H. (2024). Real-time industrial anomaly detection with edge–cloud collaborative learning. IEEE Transactions on Industrial Informatics, 20(2), 1324–1336.

[54] Sasi Kumar Kolla. (2023). Big Data–Driven Machine Learning Frameworks for Clinical Risk Prediction. International Journal of Medical Toxicology and Legal Medicine, 26(3 and 4), 44–59. Retrieved from https://ijmtlm.org/index.php/journal/article/view/1456.

[55] Zhang, C., Yang, J., & Chen, Y. (2023). AI-enabled defect detection in smart factories using hybrid deep learning models. IEEE Access, 11, 94532–94545.

[56] Guntupalli, R. (2024). Enhancing Cloud Security with AI: A Deep Learning Approach to Identify and Prevent Cyberattacks in Multi-Tenant Environments. Available at SSRN 5329132.

[57] Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157–169.

[58] Kolla, S. K. (2021). Designing Scalable Healthcare Data Pipelines for Multi-Hospital Networks. World Journal of Clinical Medicine Research, 1(1), 1–14. Retrieved from https://www.scipublications.com/journal/index.php/wjcmr/article/view/1376

[59] Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18–23.

[60] Varri, D. B. S. (2022). A Framework for Cloud-Integrated Database Hardening in Hybrid AWS-Azure Environments: Security Posture Automation Through Wiz-Driven Insights. International Journal of Scientific Research and Modern Technology, 1(12), 216-226.

[61] Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10.

[62] Koppolu, H. K. R., & Sheelam, G. K. (2024). Machine Learning-Driven Optimization in 6G Telecommunications: The Role of Intelligent Wireless and Semiconductor Innovation. Global Research Development (GRD) ISSN: 2455-5703, 9(12).

[63] Wan, J., Tang, S., Li, D., et al. (2018). A manufacturing big data solution for active preventive maintenance. IEEE Transactions on Industrial Informatics, 13(4), 2039–2047.

[64] Lahari Pandiri, "AI-Powered Fraud Detection Systems in Professional and Contractors Insurance Claims," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREEICE), DOI 10.17148/IJIREEICE.2024.121206.

[65] Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for Industry 4.0: A self-organized multi-agent system with big data-based feedback and coordination. International Journal of Distributed Sensor Networks, 12(1), 1–12.

[66] Rongali, S. K. (2023). Explainable Artificial Intelligence (XAI) Framework for Transparent Clinical Decision Support Systems. International Journal of Medical Toxicology and Legal Medicine, 26(3), 22-31.

[67] Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9592–9600.

[68] Inala, R. AI-Powered Investment Decision Support Systems: Building Smart Data Products with Embedded Governance Controls.

[69] Ruff, L., Vandermeulen, R. A., Görnitz, N., et al. (2018). Deep one-class classification. Proceedings of the 35th International Conference on Machine Learning, 4393–4402.

[70] Kolla, S. K. (2021). Architectural Frameworks for Large-Scale Electronic Health Record Data Platforms. Current Research in Public Health, 1(1), 1–19. Retrieved from https://www.scipublications.com/journal/index.php/crph/article/view/1372.

[71] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks. Information Processing in Medical Imaging, 146–157.

[72] Guntupalli, R. (2024). AI-Powered Infrastructure Management in Cloud Computing: Automating Security Compliance and Performance Monitoring. Available at SSRN 5329147.

[73] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the ACM SIGMOD International Conference on Management of Data, 93–104.

[74] Kushvanth Chowdary Nagabhyru. (2023). Accelerating Digital Transformation with AI Driven Data Engineering: Industry Case Studies from Cloud and IoT Domains. Educational Administration: Theory and Practice, 29(4), 5898–5910. https://doi.org/10.53555/kuey.v29i4.10932

[75] Sculley, D., Holt, G., Golovin, D., et al. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503–2511.

[76] Meda, R. (2024). Agentic AI in Multi-Tiered Paint Supply Chains: A Case Study on Efficiency and Responsiveness. Journal of Compu-tational Analysis and Applications (JoCAAA), 33(08), 3994-4015.

[77] Amershi, S., Begel, A., Bird, C., et al. (2019). Software engineering for machine learning: A case study. IEEE Software, 36(5), 56–67.

[78] Rongali, S. K. (2024). Federated and Generative AI Models for Secure, Cross-Institutional Healthcare Data Interoperability. Journal of Neonatal Surgery, 13(1), 1683-1694.

[79] Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. Proceedings of IEEE Big Data, 1123–1132.

[80] Meda, R. (2024). Agentic AI in Multi-Tiered Paint Supply Chains: A Case Study on Efficiency and Responsiveness. Journal of Compu-tational Analysis and Applications (JoCAAA), 33(08), 3994-4015.

[81] Zaharia, M., Das, T., Li, H., et al. (2012). Discretized streams: Fault-tolerant streaming computation at scale. Proceedings of the USENIX Symposium on Networked Systems Design and Implementation, 423–438.

[82] Velangani Divya Vardhan Kumar Bandi. (2024). Intelligent Data Platforms For Personalized Retail Analytics At Scale. Metallurgical and Materials Engineering, 30(4), 1011–1027. Retrieved from https://metall-mater-eng.com/index.php/home/article/view/1011-1027

[83] Carbone, P., Katsifodimos, A., Ewen, S., et al. (2015). Apache Flink: Stream and batch processing in a single engine. IEEE Data Engineering Bulletin, 38(4), 28–38.

[84] Keerthi Amistapuram. (2023). Privacy-Preserving Machine Learning Models for Sensitive Customer Data in Insurance Systems. Educational Administration: Theory and Practice, 29(4), 5950–5958. https://doi.org/10.53555/kuey.v29i4.10965

[85] van der Aalst, W. M. P. (2016). Process mining: Data science in action (2nd ed.). Springer.

[86] Keerthi Amistapuram. (2024). Federated Learning for Cross-Carrier Insurance Fraud Detection: Secure Multi-Institutional Collaboration. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 6727–6738. Retrieved from https://www.eudoxuspress.com/index.php/pub/article/view/3934

[87] Batini, C., & Scannapieco, M. (2016). Data and information quality: Dimensions, principles and techniques. Springer.

[88] Kalisetty, S. (2024). Deep learning frameworks for multi-modal data fusion in retail supply chains: enhancing forecast accuracy and agility.

[89] Li, Z., Wang, Y., & Zhang, H. (2024). AI-driven visual inspection and quality prediction in Industry 4.0 manufacturing systems. Journal of Manufacturing Systems, 72, 210–224.

[90] Davuluri, P. S. L. N. (2024). AI-Driven Data Governance Frameworks for Automated Regulatory Reporting and Audit Readiness. Metallurgical and Materials Engineering, 30(4), 996–1010. Retrieved from https://metall-mater-eng.com/index.php/home/article/view/1936

[91] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. (2018). Deep learning for healthcare: Review and opportunities. Briefings in Bioinformatics, 19(6), 1236–1246.

[92] Paleti, S. (2024). Transforming Financial Risk Management with AI and Data Engineering in the Modern Banking Sector. American Journal of Analytics and Artificial Intelligence (ajaai) with ISSN 3067-283X, 2(1).

[93] Johnson, A., Pollard, T., Shen, L., et al. (2020). MIMIC-IV: A freely accessible electronic health record dataset. Scientific Data, 7(1), 317.

[94] Bandi, V. D. V. K. (2023). Production-Grade Machine Learning Pipelines For Healthcare Predictive Analytics. South Eastern European Journal of Public Health, 189–205. Retrieved from https://www.seejph.com/index.php/seejph/article/view/7057

[95] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare. Health Information Science and Systems, 2(3), 1–10.