Scientific Publications

World Journal of Clinical Medicine Research, 2021, 1, 1378
www.scipublications.com/journal/index.php/wjcmr
DOI: 10.31586/wjcmr.2021.1378

Review Article

Scalable Data Warehouse Architecture for Population Health Management and Predictive Analytics

Bindu Madhavi Mangalampalli [1,*]

[1] Sr. BI Developer, USA

*Correspondence: Bindu Madhavi Mangalampalli (bindooo.madhaveee@ieee.org)

Abstract: Scalable architecture principles for data warehousing are introduced to support population health management and predictive analytics. These principles are validated through the design of an accompanying Data Pipeline that allows the integration of non-traditional data sources, the use of real-time data for descriptive analytics dashboards, and support for the generation of supervised Machine Learning models. Several analytical capabilities have been implemented to exemplify the practical application of the principles, including predictive models for Risk Stratification in health care. Optimal cost-effectiveness and performance considerations ensure the practical relevance of the architectural principles and associated Data Pipeline. In recent years, the availability of Low-Cost Data Storage services and the increasing popularity of Streaming technologies opened new possibilities for the storage and processing of Streaming data on a near-real-time basis. These technologies can help Developing Countries in tackling many relevant issues such as Urban Planning, Environmental Management, Migration Policies, etc. A multi-tier approach combining Cloud-based Storage with Data Warehousing and Data Mining technologies can offer an interesting architecture to exploit Big Data related to populations.

Keywords: Scalable Data Warehousing Architecture, Population Health Management Analytics, Predictive Risk Stratification Models, Healthcare Data Pipelines, Real-Time Streaming Data Processing, Cloud-Based Data Storage Systems, Supervised Machine Learning in Healthcare, Descriptive Analytics Dashboards, Big Data for Public Health, Multi-Tier Data Architecture Design, Cost-Optimized Data Infrastructure, Data Mining for Population Insights, Integrated Cloud Data Warehouses, Near-Real-Time Health Monitoring, Low-Cost Storage Technologies, Urban and Environmental Data Analytics, Migration Policy Data Modeling, Developing Countries Digital Infrastructure, Health Risk Forecasting Systems, Performance-Optimized Analytics Platforms

How to cite this paper: Mangalampalli, B. M. (2021). Scalable Data Warehouse Architecture for Population Health Management and Predictive Analytics. *World Journal of Clinical Medicine Research, 1*(1), 1–18.
DOI: 10.31586/wjcmr.2021.1378

Received: September 20, 2021
Revised: November 6, 2021
Accepted: December 20, 2021
Published: December 26, 2021

1. Introduction

Population health management is increasingly facilitated by predictive analytics. A scalable data warehouse architecture that supports three classes of predictive analytics—descriptive analytics supported by dashboards, predictive models that determine health risk stratification, and predictive risk models for the entire population, community, or health care delivery system—is described [1]. Principles underlying the architecture and the associated data pipeline are discussed. The operational framework draws from two major sources: existing systems and predictive analytic research.

The approach emphasizes expanding scalable data warehouses that support new capabilities, with a focus on predictive analytics that enhance control of associated costs. Scaling needs are determined by the operational workload, with horizontal scaling strategies being implemented. Capability-related dimensions define the operational workload that should be considered for horizontal scaling [2]. Technology patterns are identified for an object-oriented population health data model, the data ingestion and integration layer, and the analytical capabilities that satisfy the requirements for descriptive analytics and for predictive models. Business-driven cost optimization drivers support a query optimization pattern for common population health business questions and a metadata-driven pattern for managing an ever-growing library of business queries, models, and associated metadata [3].

1.1. Context and Rationale

Population health management aims to improve the health outcomes of a group by addressing the determinants of health and reducing health disparities among the population. Predictive analytics plays an important role in population health management. It enables health organizations to identify high-risk individuals, devise appropriate interventions, and assess the outcome [4]. Years of population health data are required for population health management and predictive analytics. A scalable, cost-effective architecture is required to continuously ingest data from various internal and external sources and build a population health management system. Despite considerable research efforts in data models and analytical methods, there is still no comprehensive discussion of a scalable architecture that can host years of population health data [5]. The lack of a scalable architecture is one of the main factors behind the slow adoption of population health management by hospitals.

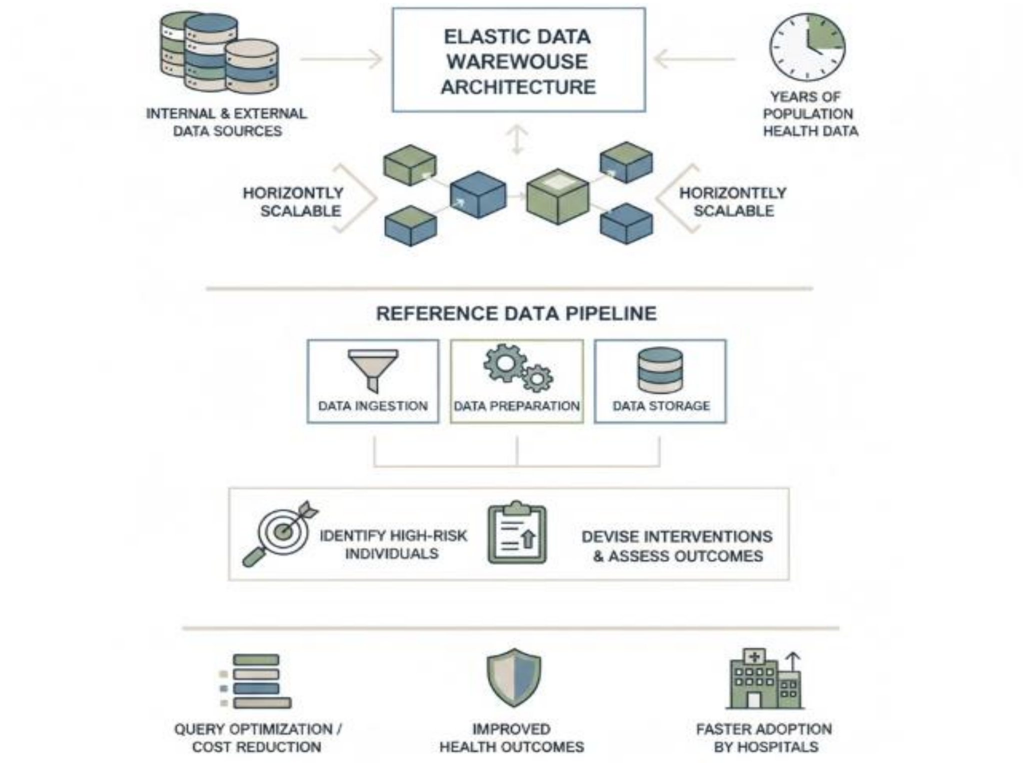
ELASTIC DATA WAREWOUSE ARCHITECTURE
INTERNAL & EXTERNAL DATA SOURCES
YEARS OF POPULATION HEALTH DATA
HORIZONTLY SCALABLE
HORIZONTELY SCALABLE
REFERENCE DATA PIPELINE
DATA INGESTION
DATA PREPARATION
DATA STORAGE
IDENTIFY HIGH-RISK INDIVIDUALS
DEVISE INTERVENTIONS & ASSESS OUTCOMES
QUERY OPTIMIZATION / COST REDUCTION
IMPROVED HEALTH OUTCOMES
FASTER ADOPTION BY HOSPITALS

Figure 1. Scaling Outcomes: An Elastic Data Warehouse Architecture for Cost-Effective Population Health Management and Predictive Analytics

The lack of horizontally scalable data warehouses has limited the ability of health organizations to build data warehouses for population health management. An elastic data warehouse architecture focusing on population health management and predictive analytics is proposed that horizontally scales-up and scales-down according to the needs of organizations and limits operating costs [6]. The architecture incorporates various architectural principles of scalable data warehousing, proposes a reference data pipeline architecture for data ingestion and preparation, and discusses the key capabilities required for population health management and predictive analytics [7]. The discussion covers horizontal scalability patterns, query optimization techniques, and recent developments in data warehouse technology that help reduce operating costs.

2. Background and Related Work

Recent pandemics continue to highlight the urgent need for workforce resilience for organizational recovery following disruptive events. Improved decision making, real-time monitoring, and advance warning of emerging problems can facilitate resilience [8]. Organizations that can effectively manage and minimize risk through data analysis are more likely to adapt, survive, and thrive. In this context, there is increasing interest in data-driven solutions for building a healthy workforce and developing healthy work environments and cultures.

Public health agencies, especially those responsible for local governance, are increasingly seeking to use these new forms of analysis to improve the health of entire populations and communities [9]. A number of agencies have either established or are exploring the establishment of a population health data warehouse. Such data facilities integrate stored, processed, organized, and cataloged wide-ranging streams of primary, secondary, and social data into a common repository to support extensive analysis capabilities [11]. However, most of the existing solutions either lack sufficient details or do not address the population health management facets of data warehousing in adequate detail.

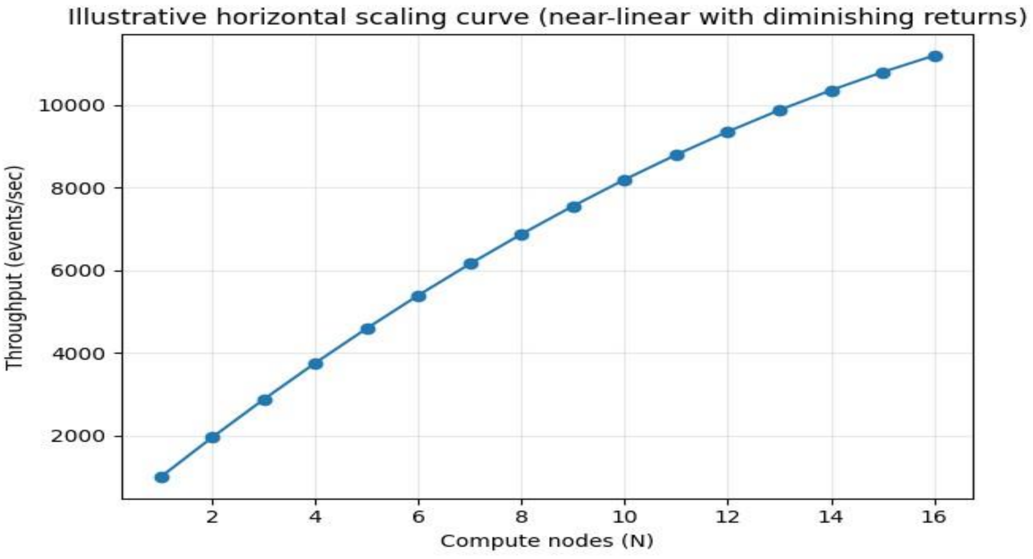

Figure 2. Illustrative horizontal scaling curve (near-linear with diminishing returns)

Equation 1: Descriptive analytics (dashboards): KPI equations (derived step-by-step)

A) Rate / proportion KPI (generic)

Step 1: Define the numerator and denominator

- Let E = number of "events of interest" (e.g., readmissions, infections).
- Let P = total eligible population or total cases at risk (depends on KPI definition).

Step 2: Define the proportion

$$\hat{p} = \frac{E}{P}$$

Step 3: Convert to percentage

$$\text{Rate}(\%) = 100 \times \hat{p} = 100 \times \frac{E}{P}$$

Example: Readmission rate

- E = readmissions within 30 days
- P = discharges (eligible index admissions)

$$\text{ReadmissionRate}(\%) = 100 \times \frac{\text{\#30-day readmissions}}{\text{\#eligible discharges}}$$

B) Trend over time (time-series KPI)

Dashboards often show KPI by time bucket (day/week/month).

Step 1: Partition data into time windows

Let t denote a time bucket (e.g., month).

Step 2: Count within each bucket

- E_t = events during bucket t
- P_t = eligible population during bucket t

Step 3: Rate per bucket

$$\text{Rate}_t = 100 \times \frac{E_t}{P_t}$$

2.1. Overview of Existing Frameworks and Research Trends

Many existing data modeling frameworks are limited to specific application domains and focus primarily on descriptive analytics. Modi, Ranjan, and Pramanik proposed an adaptable data warehouse framework for securely managing vast amounts of data generated by smart cities. The work emphasized the importance of privacy preservation, as well as the ability to meet the requirements of diverse users [12]. The implementation of their proposed framework in Oracle 12C supported ad hoc queries, data trends, and predictions. It was shown that the framework was scalable for cities with millions of citizens providing different information.

Mevorach, Chaim, and Ben-Ceder provided an integrated approach to solving data acquisition, caching, search and retrieval, understanding, analysis, visualization, and automatic recommendations using analytic tools, neural networks, and data mining. Scaling issues were addressed by implementing all services on Docker containers, improving efficiency and enhancing user experience [13]. The solution met the requirements of Small Office Home Office applications operating in a cloud setup. Recent advances in the healthcare domain have shown a trend toward the use of epidemic models for risk assessment, and Wilkerson's epidemiological model allows a granule level risk and exposure analysis for infectious disease, yet it fails to cover broader risk and noncommunicable disease analysis [14].

Compute nodes (N)	**Throughput (events/sec)**
1	1000
2	1960
3	2880
4	3760
5	4600
6	5400
7	6160

3. Architectural Principles for Scalable Data Warehousing

Effective decision-making for population health management and preventive care relies on a data pipeline designed to handle high data velocity and volume requirements. A scalable data-warehousing architecture based on the principles of data modeling, loading, and integration enables organizations to take a more proactive approach to population health.

The first step in the proposed approach is the development of a data model to support data ingest from disparate sources [15]. Data from external data sources, such as additional organizations, third-party data-analytics vendors, and state or federal agencies, can then be integrated quickly. The final data model supports the display of descriptive analytics and dashboards showing the population risk for various diseases to facilitate preventive care. Components are included in the architecture to manage and monitor the ingest process and to provide user-friendly engagement with the analytics.

Equation 2: Predictive modeling for risk stratification (math behind "risk scores")

A) Logistic regression probability (step-by-step)

Goal: predict binary outcome $Y \in \{0,1\}$ (e.g., readmission yes/no).

Step 1: Linear score

Let features be $x_1, x_2, \dots, x_d$. Define:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d$$

Step 2: Map score to probability using sigmoid

$$p = P(Y = 1 \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Step 3: Risk score

The **risk score** is $p \in [0,1]$. Stratification buckets then follow:

- Low risk if $p < 0.2$
- Medium if $0.2 \leq p < 0.6$
- High if $p \geq 0.6$

(thresholds are business/clinical choices).

B) Odds, odds ratio, and "relative risk" (step-by-step)

Step 1: Define odds

If probability is p,

$$\text{odds} = \frac{p}{1 - p}$$

Step 2: Odds ratio between two groups A and B

$$\text{OR} = \frac{p_A/(1 - p_A)}{p_B/(1 - p_B)}$$

Step 3: Relative risk (risk ratio)

$$\text{RR} = \frac{p_A}{p_B}$$

Interpretation

- RR = 2 means group A has **double** the risk of group B.
- These are the quantities typically shown in "relative risk plots" across predictors.

C) Choosing a cut-off threshold (sentry strategy) with sensitivity/specificity

Let a threshold be τ. Predict $\hat{Y} = 1$ if $p \geq \tau$, else 0.

From the confusion matrix:

- TP = true positives
- FP = false positives
- TN = true negatives

- FN = false negatives

Sensitivity (Recall / TPR)

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity (TNR)

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Precision (PPV)

$$\text{Precision} = \frac{TP}{TP + FP}$$

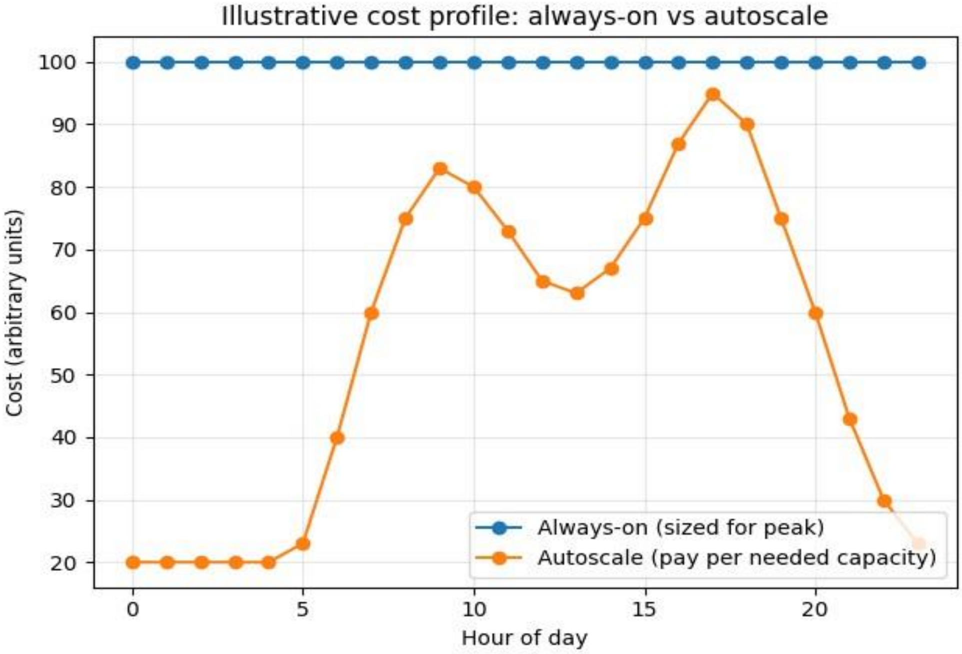

Figure 3. Illustrative cost profile: always-on vs autoscale

3.1. Data Modeling for Population Health

A dimensional model of the entire country, supporting data mining and predictive analytics, requires modifying the existing schema [16]. The Patient dimension serves as the most crucial of a set of dimensions whose attributes will find repeated use in several facts. Healthcare and hospital system collaboration patterns that exist between provider organizations, physician groups and health insurers appear in a three-dimensional relationship using Fact_Patient_Hospital_Collaboration, Fact_Patient_Physician_Collaboration, and Fact_Patient_HealthPlan_Collaboration. The two-stage snowflake-shape Drives_Industry_Population dimension addresses the concern that no industry qualifies as a plan sponsor; in retirement living, for example, the clients prefer a middling balance sheet.

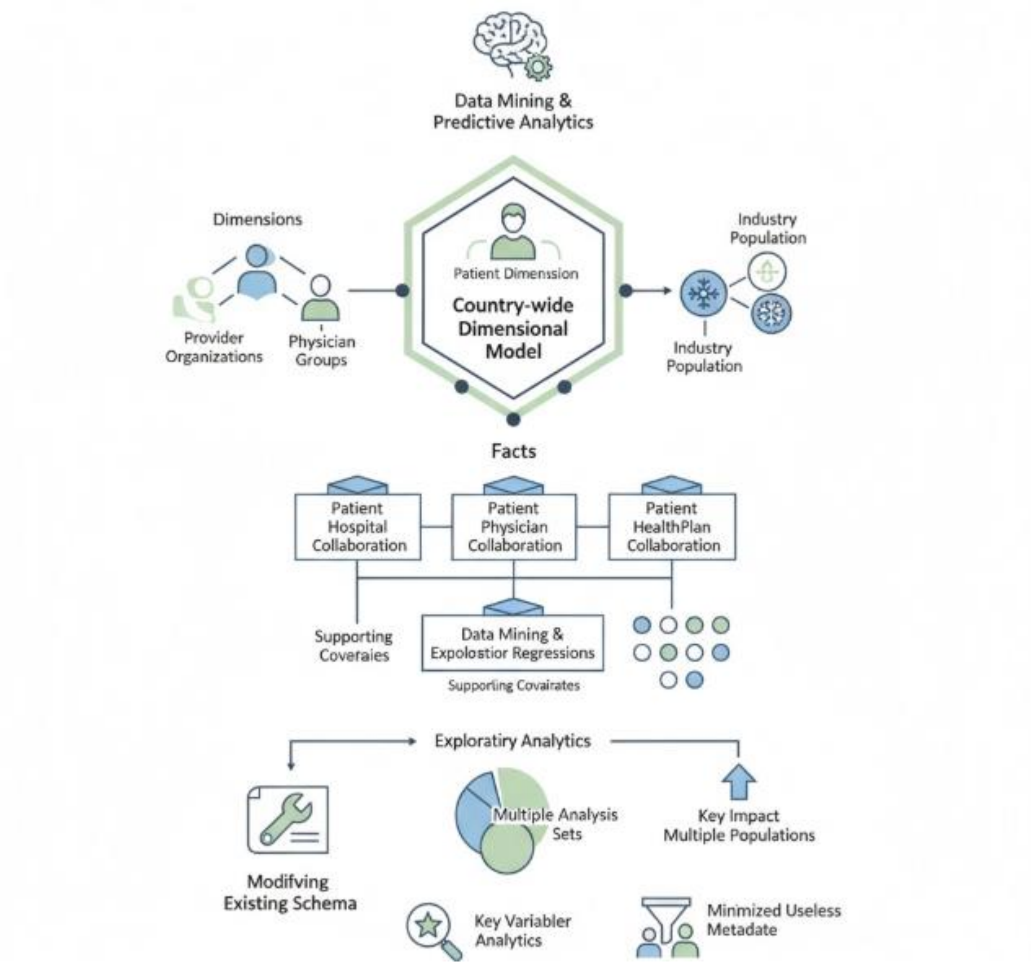

Figure 4. Nationwide Dimensional Architectures: A Multi-Fact Schema for Predictive Analytics and Exploratory Regression in Healthcare Ecosystems

Besides the dimensional model with numerous facts for supporting descriptive analytics, the requirement for bulk fuels predictive analytics and a broader set of exploratory analytical processes capable of uncovering patterns for further investigation [17]. Multiple analysis sets must therefore be fleshed out by alternative approaches to the same problem. Choosing predictive-modeling ideas for optimization makes it essential to identify key variables with the power to Impact multiple populations [18]. These key variables must capture the attention and imagination of analysts and serve as flagpoles for further investigation.

The exploration enlarges the data scope. The extended view joins the Population Fact data sources with a new Data_Mining_Exploratory_Regressions_fact and the master set of Data_Mining_Supporting_Covariates designed to provide descriptive tools to seo 2018 advertising strategy insights [19]. Seeking to minimize useless metadata narrows the parameter-searching choice to a mod slim to go for it.

3.2. Data Ingestion and Integration

Ingesting data from diverse sources—operational databases, distributed file systems, web-services APIs, and enterprise data warehouses—requires supports for both streaming and batch processing. Streaming ingest is managed by a Lambda architecture that combines a speed layer (Apache Kafka) with a serving layer (Apache Cassandra). For batch ingest, ETL jobs in Airflow populate a multi-tenant, enterprise data warehouse in Snowflake [20]. Supporting population health management entails deriving risk factors and other analytical attributes from raw data and loading these into a warehouse for batch or near-real-time access.

A growing range of initiatives focus on expanding the ability of health organizations to better manage the health of populations [21]. Transforming clinical data into machine learning-ready data sets enables hospitals and care team members to deploy risk models and advance predictive analytics in Care Management and Population Health teams at multiple sites. Yet, such data engineering often remains occurring only at a single local site.

4. Data Pipeline Architecture

Several data-processing units implement a microservices-oriented architecture to successfully realize the aforementioned data ingestion and integration strategy [22]. The units leverage open-source components and cloud-native design principles to enable efficient streaming and batch-processing operations while ensuring quality of service through a combination of metadata management, orchestration, and monitoring.

1. Streaming and Batch Processing

Various systems are deployed to load data from the source systems and prepare it for analytics. For real-time data loading, a Kafka streaming-based solution orchestrated by Debezium continuously ingests event-driven change-data-capture messages from the source transactional databases [23]. This data flow integrates heterogeneous data sources such as MySQL, PostgreSQL, and a software-as-a-service solution using webhooks and Kafka Connector APIs. Data quality and business rules are validated prior to storage in the web-dedicated staging database, which serves both streaming and batch-processing applications [24]. A separate Kafka flow and Spark job ingests data daily from non-transactional source systems. This batch-processing solution [25]. manages complex refers-to relationships in the underlying data by relying on a custom metadata repository that drives the data preparation and ingestion process.

2. Metadata Management and Orchestration

In addition to ensuring data quality, metadata information strengthens data preparation and speeds up the data-loading operations [26]. A proprietary metadata-management solution provides a high-level business view of all source, transformed, quality-validated, and reference tables in the entire pipeline. This repository, now in production, replaces hard-coded table names and related information across the components and services that make up the data-management ecosystem [27]. It provides additional quality metrics that add assurance to the data pipeline and constitutes the main driver of the data orchestration layer [28].

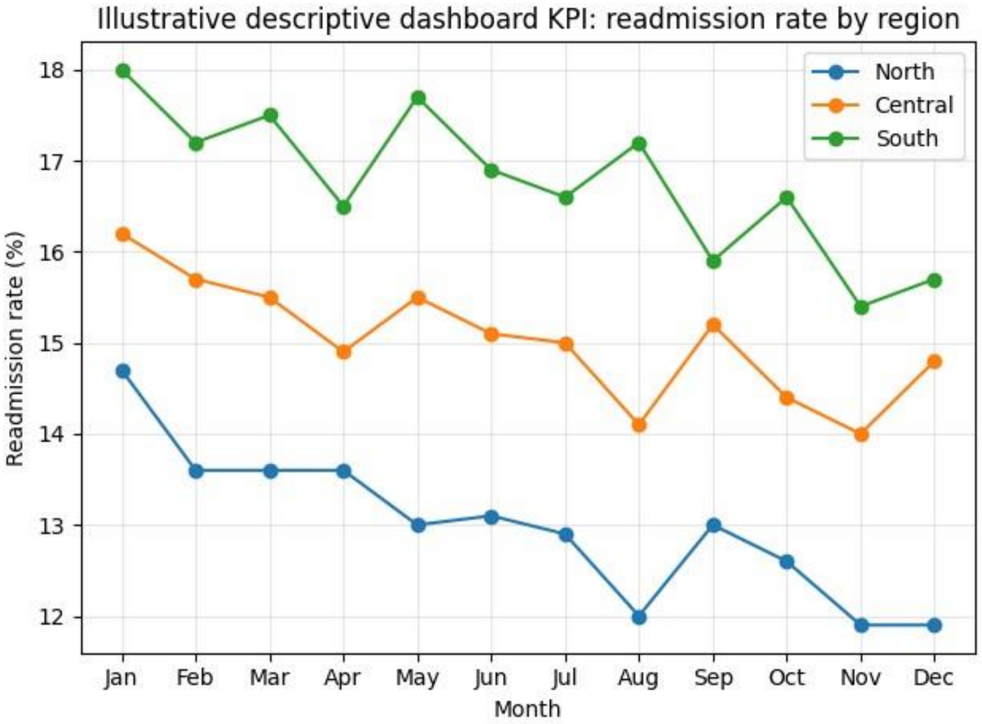

Figure 5. Illustrative descriptive dashboard KPI: readmission rate by region

Equation 3: Cost-effectiveness / operating-cost equations (autoscaling vs always-on)

A) Always-on cost (step-by-step)

Step 1: Define capacity and price

- Provisioned capacity $C_{\max}$ (sized for peak)
- Unit price r (cost per capacity-unit per hour)
- Runtime horizon T (hours)

Step 2: Always-on cost

$$\mathrm{Cost}_{\mathrm{always}} = r \cdot C_{\max} \cdot T$$

B) Autoscale cost (step-by-step)

Step 1: Define required capacity over time

Let workload-utilization (or required capacity fraction) be $u(t) \in [0,1]$.

$$C(t) = C_{\max} \cdot u(t)$$

Step 2: Integrate cost over time

$$\mathrm{Cost}_{\mathrm{auto}} = \int_0^T r \cdot C(t)\, dt = \int_0^T r \cdot C_{\max} \cdot u(t)\, dt$$

Step 3: Discrete (hourly) version used in practice

If measured hourly $t = 1, \dots, T$:

$$\mathrm{Cost}_{\mathrm{auto}} = \sum_{t=1}^{T} r \cdot C_{\max} \cdot u_t$$

Step 4: Savings

$$\mathrm{Savings}(\%) = 100\left(1 - \frac{\mathrm{Cost}_{\mathrm{auto}}}{\mathrm{Cost}_{\mathrm{always}}}\right)$$

4.1. Streaming and Batch Processing

The ingestion of data from heterogeneous sources, ranging from real-time streaming to batch processing, is a common challenge in building a population health data warehouse [29]. The availability of large-scale cloud data storage and processing services such as Amazon S3, Azure Data Lake Storage, [30]. and Google Cloud Storage enables secure and cost-efficient storage of all types of data. Streaming services such as Amazon Kinesis Stream and Google Pub/Sub have made it easy to create real-time applications that can simultaneously ingest multiple data streams [31]. The data from different sources differ in delay tolerance, data freshness requirement, and update frequency [32]. Data lakes are effectively used to ingest additional sources of data that are rarely used in most of the population management use cases but may unlock additional insights for posterior analysis or training of more accurate predictive models. Modern systems provide the capability to store data in a serverless manner allowing to move the responsibility of infrastructure management to the cloud provider, resulting in a cost-effective solution to maintain [33]. Data can be made accessible in streaming manner through systems such as Amazon Glue, Apache Hive, Google BigQuery, and Snowflake that automate the work of schema inference, creation and maintenance of the metadata catalog, data partitioning and making the data accessible through a serverless manner and at scale.

The ingestion of data from heterogeneous sources, ranging from real-time streaming to batch processing, is a common challenge in building a population health data warehouse [34]. The availability of large-scale cloud data storage and processing services such as Amazon S3, Azure Data Lake Storage, and Google Cloud Storage enables secure and cost-efficient storage of all types of data [35]. Streaming services such as Amazon Kinesis Stream and Google Pub/Sub have made it easy to create real-time applications that can simultaneously ingest multiple data streams [36]. The data from different sources differ in delay tolerance, data freshness requirement, and update frequency [37]. Data lakes are effectively used to ingest additional sources of data that are rarely used in most of the population management use cases but may unlock additional insights for posterior analysis or training of more accurate predictive models [38]. Modern systems provide the capability to store data in a serverless manner allowing to move the responsibility of infrastructure management to the cloud provider, resulting in a cost-effective solution to maintain [39]. Data can be made accessible in streaming manner through systems such as Amazon Glue, Apache Hive, Google BigQuery, and Snowflake that automate the work of schema inference, creation and maintenance of the metadata catalog, data partitioning and making the data accessible through a [40]. serverless manner and at scale.

4.2. Metadata Management and Orchestration

Orchestration is the process of managing complex collections of tasks and services that interact with each other [41]. External metadata is the information that operations, such as validation and monitoring, services, such as storage and analytics, and users require in order to leverage the data stored in a data lake or data warehouse [42]. Metadata refers to the data about the data. It can include descriptions of the data sources, elements, types, and structures [43]. Metadata provides information about meaning, quality, condition, timeliness, origin, and other characteristics of the data [44]. Metadata helps users locate, understand, and effectively use data.

Data pipeline orchestration automates the management of data pipeline tasks such as scheduling, managing dependencies, triggering data transformations, and summarizing pipeline health. One of the major benefits of a data pipeline orchestration system is scheduling [45]. It allows processing tasks to be scheduled in a flexible way that isn't restricted to a certain time frame. Orchestration engines schedule tasks based on events such as the arrival of new data. They handle dependencies, notifications, and warnings to inform users when things go wrong. Data pipeline tasks require the support of orchestration engines for process deployment and management [46]. Orchestration metadata may reside in a relational database [47]. Supporting an API for monitoring allows users to retrieve the status of jobs and execution logs.

Hour	Utilization (fraction)	Always-on cost (units)	Autoscale cost (units)
18	0.85	100	90
19	0.7	100	75
20	0.55	100	60
21	0.38	100	43
22	0.25	100	30
23	0.18	100	23

5. Analytical Capabilities for Population Health

Effective population health management necessitates sophisticated analytic capabilities that extend beyond data collection and storage. In addition to enabling descriptive analytics and the delivery of dashboards for empirical decision-making, a comprehensive architecture should support predictive-modeling tasks [48]. Risk stratification analytics that evaluate past utilization levels of defined cohorts provide an important data source for forecasting and predictive tasks. Routine integration of population health utilization data enriches predictive models and produces risk scores that can be viewed through dashboards [49].

A common challenge with risk models for population health is the limited timeframe for modeling due to the absence of prospective patient-travel patterns [51]. Networks examining a relatively small geographic area typically benefit from enhanced predictive accuracy due to the continuity in healthcare utilization by population groups [52]. Obtaining an accurate utilization profile of a cohort enables the use of consumption data for modeling and subsequent forecasting. Descriptive analyses performed earlier facilitate the effective segmentation of a high-risk cohort and may therefore support effective targeting for outreach investments [53].

Equation 4: Horizontal scaling equation (throughput vs number of nodes)

A) Ideal linear scaling

If 1 node processes X_1 events/sec, then N nodes ideally:

$$X(N) = N \cdot X_1$$

B) More realistic: diminishing returns due to coordination/shuffle

Introduce efficiency $\eta(N) \in (0,1]$., decreasing with N:

$$X(N) = N \cdot X_1 \cdot \eta(N)$$

A simple form:

$$\eta(N) = \frac{1}{1 + \alpha(N - 1)}$$

5.1. Descriptive Analytics and Dashboards

Dashboards facilitate monitoring of key performance indicators (KPIs) and emerging trends [55]. These may include vaccine coverage status by age group and other characteristics, Covid-19 infection and mortality rates by population strata, or healthcare access reduction over time for vulnerable groups.

Geo-spatially enabled dashboards must also be designed to help local decision-makers visualize the historical impact of developmental and health supply-side circumstances or recent investment plan actions on health and socio-economic KPIs [56].

Dashboards are also needed on population event predictions, and outcome risk stratification at various geographic scales; they must allow stratification by coverage, risk, and vulnerability conditions [57]. Desktop dashboards may be developed for specialized users, to support predictive model development or for outbreak forecasting based on various event predictors.

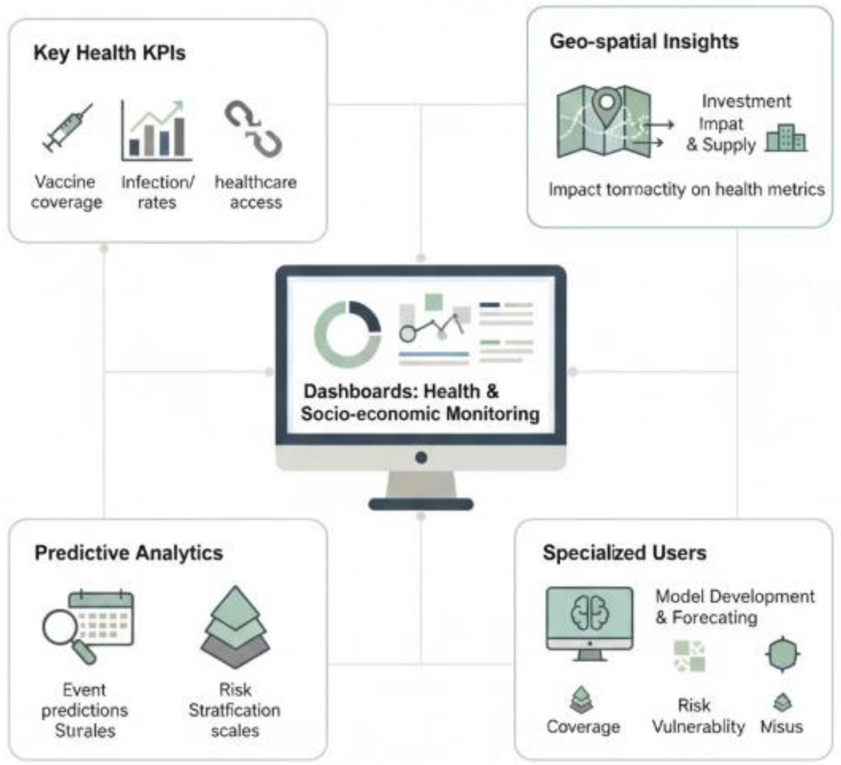
Key Health KPIs
Vaccine coverage
Infection/ rates
healthcare access
Geo-spatial Insights
Investment Impat & Supply
Impact tornactity on health metrics
Dashboards: Health & Socio-economic Monitoring
Predictive Analytics
Event predictions Surales
Risk Stratfication scales
Specialized Users
Model Development & Forecating
Coverage
Risk Vulnerablity
Misus

Figure 6. Spatio-Temporal Intelligence Dashboards: Integrated Risk Stratification and Predictive Analytics for Evidence-Based Health Policy

5.2. Predictive Modeling and Risk Stratification

Predictive analytics--the making of predictions about future outcomes based on historical and existing data--is increasingly common in business, healthcare, public health, and many other fields [58]. The applicability of quantitative predictive models to large-scale public health processes, such as health-related quality-of-life determinants, risk factors for select diseases, problems under other mental health conditions, hospital mortality, and mortality across the continuum of care, is also well documented [59]. Within the population health domain, predictive methods can be applied to risk stratification by developing predictive models for a specific outcome of interest (e.g., hospital readmission, disease outbreak) with recognized covariates associated with that outcome [60]. The models are then used to calculate the underlying absolute or relative risks for new or non-identified patient populations, patients at an increased risk regardless of the predictive horizon, and potential risk factors to target [61].

Predictive modeling frameworks generally share a distinct characteristic: the depiction of the specific population of interest for modeling. Population health predictive models often use longitudinal data for prediction, such as multi-year patient records at the individual level or population-level data for a smaller geographic area. The general method for longitudinal prediction involves associating the longitudinally accumulated patient information with a singular incident of a health event. Longitudinally completed predictors can be considered the main building block for longitudinal prediction [62]. A distinct health population model is summarized using literature knowledge: relative risk plots across the multiple predictors, an optimal cut-off point of the relative risk for a sentry strategy, and the acknowledgment of limitations and future exploitation of each health predictor [63].

Month	North	Central	South
Jan	14.7	16.2	18.0
Feb	13.6	15.7	17.2
Mar	13.6	15.5	17.5
Apr	13.6	14.9	16.5
May	13.0	15.5	17.7
Jun	13.1	15.1	16.9

6. Scalability, Performance, and Cost Optimization

Scalability, performance, and cost-effectiveness are critical considerations for almost any warehouse implementation [64]. Horizontal scalability matters most when dealing with continuously accumulating data, while cost-effectiveness becomes especially important for workloads that do not demand always-on resources.

Several horizontal scalability patterns have been documented in current literature, and these patterns can also be beneficial for warehouses used in population health management [65]. A batch-mode data ingestion pipeline offering scalability in write operations is a key requirement. Enabling auto-scaling where supported is a crucial cost optimization measure for cloud-hosted solutions [66]. Within the analytical pipeline, embedding support for partitioned execution can reduce the overall cost of analytic workloads. Other factors affecting horizontal scalability include the choice of compute engine and orchestration strategies [67].

The cost of accelerating performance-critical analytic workloads can also benefit from established techniques [68]. Query optimization principles that minimize the volume of cross-partition data movement are paramount. Query processing engines often support materialized views that can be employed to increase performance while maintaining cost-effectiveness. Partitioned materialized views further reduce the freshness and storage cost trade-offs for view maintenance.

Equation 5: Materialized views / incremental refresh (query optimization)

Suppose a dashboard needs an aggregate:

$$A = \sum_{i \in \text{base}} f(i)$$

If only a small delta Δ changes, do:

Step 1: Split base into unchanged + changed

$$A_{\text{new}} = A_{\text{old}} - \sum_{i \in \Delta^-} f(i) + \sum_{i \in \Delta^+} f(i)$$

- Δ^-: records removed/invalidated
- Δ^+: records inserted/updated

Step 2: Complexity benefit

- Full recompute: $O(|\text{base}|)$
- Incremental: $O(|\Delta|)$, with $|\Delta| \ll |\text{base}|$

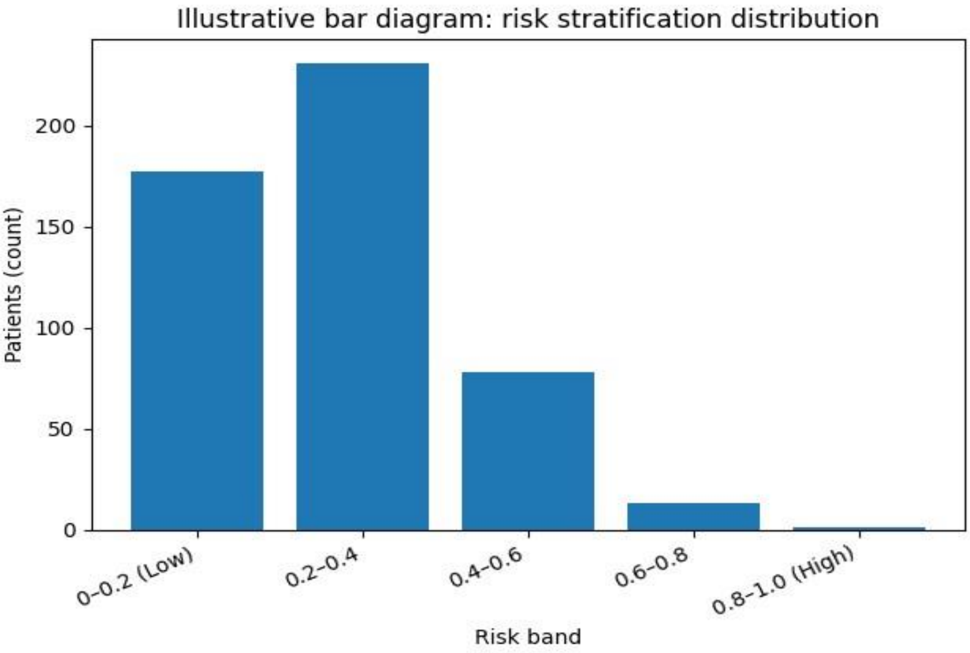

Figure 7. Illustrative bar diagram: risk stratification distribution

6.1. Horizontal Scalability Patterns

Requesting a section of scholarly writing from an academic work one portion at a time. A previous completion provides the next section.

While Hadoop clusters leverage the distributed file store and parallel processing capabilities, streaming data processing requires dedicated horizontal scaling for burst traffic. A burst in incoming requests may occur when the population completes a survey or national healthcare datasets are released. Scalable ephemeral infrastructure provisioning reduces costs, retaining only minimal273 support instances [69].

Social networks are key drivers for population interventions. A public campaign may aim for sexual health, stopping smoking, or increasing flu vaccination and send a message on every channel of a multiple-channel marketing strategy273,274 [70]. There is a need to continuously process data for strategic planning, refinement of algorithms, and for change, correlation, and influence analysis.

Population health prediction incorporates social factors through the analysis of user-generated data on online social networks275,276. Social network users regularly disclose sentiments on their health and record life events. [71]. The prediction capability is based on models that learn the correlations between emotive expressions of social network users and health indicators258. By including change, correlation, and influence analysis in a recommendation engine, it is possible to craft and deliver campaigns to populations most affectable by health issues or behaviours. Real-time social media monitoring is necessary to warn and intercept changes277 [72].

6.2. Query Optimization and Materialized Views

Optimizing query performance is essential for data warehouse architecture because population health management is an interactive analytical task where users frequently need to re-query the data [73]. One simple but effective way to optimize repeated query execution is by caching and maintaining a materialized view of the query results. Materialized views can improve query performance and support so-called re-execution of queries by taking advantage of previously computed sub-expressions [74]. Furthermore, in a population health management context, it is common to create dashboards for descriptive analytics that provide insights into important metrics over time, such as hospital readmission rates or disease case counts across geospatial regions.

Maintaining a materialized view that keeps the last k values by some dimension can therefore greatly improve performance because these aggregates are typically calculated for different combinations of attributes at differing granularities of time and geospatial location. Support for incremental refreshment based on data change patterns can also speed up maintenance time [75]. For example, a materialized view supporting hospital readmission over time may be refreshed on an hourly or daily basis. During such periods, stored procedure invocation of the refreshment can simply retrieve before and after counts of patients readmitted to a hospital within 30 days rather than scanning the entire base table to compute the aggregate [76].

7. Conclusion

The increasing volume and variety of data from social determinants of health and Internet-of-Things devices requires a new type of data warehousing architecture to support population health management and predictive analytics [77]. Horizontal scalability and automated management of streaming and batch workloads are key architectural features of a suitable approach. Scalable Data Warehousing provides a framework for reusable enterprise data models, ingestion and integration pipelines, predictive modeling, performance tuning, and cost optimization [78].

The implementation of scalable data warehousing principles within a population health-focused data pipeline architecture illustrates their effectiveness. The combination of a generic data-modeling capability, technologies that facilitate ingestion of large datasets, and tools for monitoring and orchestrating end-to-end execution helps satisfy the diverse requirements of multiple stakeholders [79]. Successful provision of descriptive and predictive analytics at scale—in particular, cost-effective risk stratification for precision population health management—demonstrates the approach's value at a time when society is struggling to cope with increasing levels of service demand.

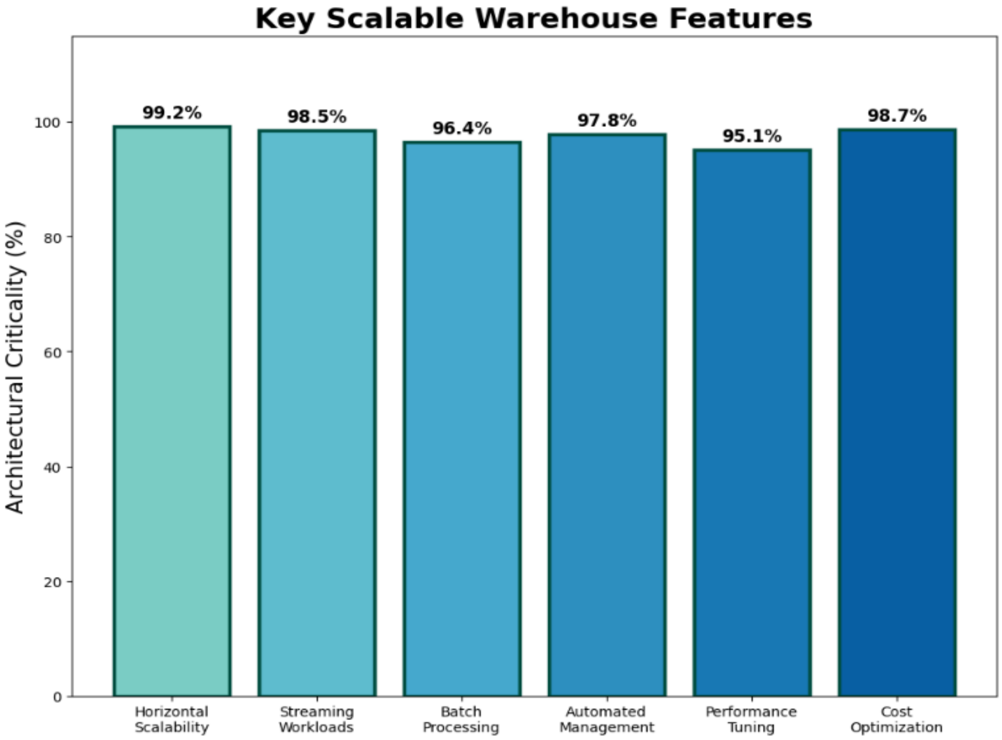

Figure 8. Key Scalable Warehouse Features

7.1. Summary of Key Findings and Future Directions

The architecture of a scalable data warehouse for population health management and predictive analytics has been developed following four key principles: 1) a data model that serves as a framework for defining data ingested from different sources; 2) an ingest process that involves the use of streaming and batch processes and helps pre-validate and harmonize the data; 3) metadata management that not only captures business terms used in the model but also manages the execution plan of data flows and tools like [80]. Apache Airflow; and 4) readily available descriptive analytics with the option of developing predictive models using tools like R and Python, linked to the database. Together, these elements provide horizontal scalability and help with cost and performance optimization.

Population health management and analytics are becoming increasingly important in the healthcare domain [81]. However, existing data warehousing and analytical frameworks do not offer a sufficiently scalable architecture or are restricted to a specific scale. While cloud-based databases can achieve some level of auto-scaling, they often remain unutilized during off-peak hours, leading to increased operational costs. An affordable, highly scalable, [82]. and predictive data warehouse architecture that supports both batch and streaming data ingestion while providing real-time dashboards and the capacity to develop predictive analytics model on historical data is therefore very much needed [83]. Future research will extend the work by providing a multi-tier architecture that integrates a data lake with the existing framework to support the ingest of unstructured, streaming, and other data that either cannot be addressed or requires more complex processing.

References

[1] Meda, R. End-to-End Data Engineering for Demand Forecasting in Retail Manufacturing Ecosystems.y. Proceedings of the National Academy of Sciences, 110(30), 12219–12224.

[2] Akoglu, L., Tong, H., & Koutra, D. (2015). Graph-based anomaly detection and description. Data Mining and Knowledge Discovery, 29(3), 626–688.

[3] Segireddy, A. R. (2021). Containerization and Microservices in Payment Systems: A Study of Kubernetes and Docker in Financial Applications. Universal Journal of Business and Management, 1(1), 1–17. Retrieved from https://www.scipublications.com/journal/index.php/ujbm/article/view/1352.

[4] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52.

[5] Pandugula, C., & Yasmeen, Z. (2019). A Comprehensive Study of Proactive Cybersecurity Models in Cloud-Driven Retail Technology Architectures. Universal Journal of Computer Sciences and Communications, 1(1), 1253.

[6] Biondini, M., & Boldrini, E. (2014). Clinical data warehouse architecture and data quality issues. Studies in Health Technology and Informatics, 205, 127–131.

[7] Gottimukkala, V. R. R. (2020). Energy-Efficient Design Patterns for Large-Scale Banking Applications Deployed on AWS Cloud. power, 9(12)

[8] Brown, J. S., Kahn, M., & Toh, S. (2013). Data quality assessment for comparative effectiveness research. Medical Care, 51(8 Suppl 3), S22–S29.

[9] Kummari, D. N. (2021). A Framework for Risk-Based Auditing in Intelligent Manufacturing Infrastructures. International Journal on Recent and Innovation Trends in Computing and Communication, 9(12), 245-262.

[10] Cappiello, C., Francalanci, C., & Pernici, B. (2004). Data quality assessment from the user's perspective. Proceedings of the 2004 International Conference on Information Quality, 68–73.

[11] Ahmed, M., Mahmood, A., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19–31.

[12] Chute, C. G., et al. (2010). The SHARPn project on secondary use of EHR data. Journal of Biomedical Informatics, 43(5), 760–771.

[13] Pamisetty, A. (2021). A comparative study of cloud platforms for scalable infrastructure in food distribution supply chains.

[14] Date, C. J. (2004). An introduction to database systems (8th ed.). Addison-Wesley.

[15] Rongali, S. K. (2021). Cloud-Native API-Led Integration Using MuleSoft and .NET for Scalable Healthcare Interoperability. Available at SSRN 5814563.

[16] Denny, J. C., et al. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan. Bioinformatics, 26(9), 1205–1210.

[17] Meda, R. (2020). Data Engineering Architectures for Real-Time Quality Monitoring in Paint Production Lines. International Journal Of Engineering And Computer Science, 9(12).

[18] Fan, W., & Geerts, F. (2012). Foundations of data quality management. Morgan & Claypool.

[19] Pamisetty, V. (2021). A Cloud-Integrated Framework for Efficient Government Financial Management and Unclaimed Asset Recovery. Available at SSRN 5272351.

[20] Vadisetty, R., Polamarasetti, A., Guntupalli, R., Rongali, S. K., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2020). Generative AI for Cloud Infrastructure Automation. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 1(3), 15-20

[21] Golfarelli, M., & Rizzi, S. (2009). Data warehouse design. McGraw-Hill.

[22] Varri, D. B. S. (2021). Cloud-Native Security Architecture for Hybrid Healthcare Infrastructure. Available at SSRN 5785982.

[23] Hersh, W. (2007). Information retrieval: A health and biomedical perspective (2nd ed.). Springer.

[24] Inala, R. (2020). Building Foundational Data Products for Financial Services: A MDM-Based Approach to Customer, and Product Data Integration. Universal Journal of Finance and Economics, 1(1), 1-18.

[25] Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping. Journal of the American Medical Informatics Association, 20(1), 117–121.

[26] Yandamuri, U. S. (2021). A Comparative Study of Traditional Reporting Systems versus Real-Time Analytics Dashboards in Enterprise Operations. Universal Journal of Business and Management, 1(1), 1–13. Retrieved from https://www.scipublications.com/journal/index.php/ujbm/article/view/1357

[27] Inmon, W. H. (2005). Building the data warehouse (4th ed.). Wiley.

[28] Amistapuram, K. (2021). Digital Transformation in Insurance: Migrating Enterprise Policy Systems to .NET Core. Universal Journal of Computer Sciences and Communications, 1(1), 1–17. Retrieved from https://www.scipublications.com/journal/index.php/ujcsc/article/view/1348

[29] Johnson, A. E. W., et al. (2016). MIMIC-III database. Scientific Data, 3, 160035.

[30] Rongali, S. K. (2020). Predictive Modeling and Machine Learning Frameworks for Early Disease Detection in Healthcare Data Systems. Current Research in Public Health, 1(1), 1-15.

[31] Kimball, R., & Ross, M. (2013). The data warehouse toolkit (3rd ed.). Wiley.

[32] Challa, K. (2021). Cloud Native Architecture for Scalable Fintech Applications with Real Time Payments. International Journal Of Engineering And Computer Science, 10(12).

[33] Liaw, S.-T., et al. (2013). Towards an ontology for data quality. Journal of Biomedical Informatics, 46(1), 80–92.

[34] Koppolu, H. K. R. (2021). Data-Driven Strategies for Optimizing Customer Journeys Across Telecom and Healthcare Industries. International Journal Of Engineering And Computer Science, 10(12).

[35] Luján-Mora, S., et al. (2006). A UML profile for multidimensional modeling. Data & Knowledge Engineering, 59(3), 725–769.

[36] Yandamuri, U. S. (2021). A Comparative Study of Traditional Reporting Systems versus Real-Time Analytics Dashboards in Enterprise Operations. Universal Journal of Business and Management.

[37] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Academic Press.

[38] Dwaraka Nath Kummari, Srinivasa Rao Challa, "Big Data and Machine Learning in Fraud Detection for Public Sector Financial Systems," International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI: 10.17148/IJARCCE.2020.91221.

[39] Goutham Kumar Sheelam, Botlagunta Preethish Nandan, "Machine Learning Integration in Semiconductor Research and Manufacturing Pipelines," International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI: 10.17148/IJARCCE.2021.101274.

[40] Miotto, R., et al. (2018). Deep learning for healthcare. Briefings in Bioinformatics, 19(6), 1236–1246.

[41] Kolla, S. H. (2021). Rule-Based Automation for IT Service Management Workflows. Online Journal of Engineering Sciences, 1(1), 1–14. Retrieved from https://www.scipublications.com/journal/index.php/ojes/article/view/1360

[42] Nandan, B. P., Sheelam, G. K., & Engineer Sr, I. D. Data-Driven Design and Validation Techniques in Advanced Chip Engineering.

[43] Nambiar, R., & Poess, M. (2006). TPC-DS benchmark. Proceedings of the VLDB Endowment, 1049–1058.

[44] Meda, R. (2019). Machine Learning Models for Quality Prediction and Compliance in Paint Manufacturing Operations. International Journal of Engineering and Computer Science, 8(12), 24993–24911. https://doi.org/10.18535/ijecs.v8i12.4445.

[45] O'Neil, P., & Quass, D. (1997). Improved query performance with variant indexes. Proceedings of the ACM SIGMOD International Conference, 38–49.

[46] Inala, R. Designing Scalable Technology Architectures for Customer Data in Group Insurance and Investment Platforms.

[47] Pedersen, T. B., & Jensen, C. S. (2001). Multidimensional database technology. IEEE Computer, 34(12), 40–46.

[48] Aitha, A. R. (2021). Optimizing Data Warehousing for Large Scale Policy Management Using Advanced ETL Frameworks.

[49] Poess, M., & Nambiar, R. (2008). Benchmarking data warehouses. ACM SIGMOD Record, 37(1), 13–20.

[50] Segireddy, A. R. (2020). Cloud Migration Strategies for High-Volume Financial Messaging Systems.

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

[52] Gottimukkala, V. R. R. (2021). Digital Signal Processing Challenges in Financial Messaging Systems: Case Studies in High-Volume SWIFT Flows.

[53] Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592.

[54] Sriram, H. K., ADUSUPALLI, B., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks.

[55] Schriml, L. M., et al. (2012). Disease ontology. Nucleic Acids Research, 40(D1), D940–D946.

[56] Aitha, A. R. (2021). Dev Ops Driven Digital Transformation: Accelerating Innovation In The Insurance Industry. Available at SSRN 5622190.

[57] Silberschatz, A., Korth, H. F., & Sudarshan, S. (2010). Database system concepts (6th ed.). McGraw-Hill.

[58] Inala, R. (2021). A New Paradigm in Retirement Solution Platforms: Leveraging Data Governance to Build AI-Ready Data Products. Journal of International Crisis and Risk Communication Research, 286-310.

[59] Stonebraker, M., & Çetintemel, U. (2005). One size fits all? ICDE Proceedings, 2–11.

[60] Vadisetty, R., Polamarasetti, A., Guntupalli, R., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2021). Privacy-Preserving Gen AI in Multi-Tenant Cloud Environments. Sateesh kumar and Raghunath, Vedaprada and Jyothi, Vinaya Kumar and Kudithipudi, Karthik, Privacy-Preserving Gen AI in Multi-Tenant Cloud Environments (January 20, 2021).

[61] Sun, J., & Reddy, C. K. (2013). Big data analytics for healthcare. Proceedings of the ACM SIGKDD Conference.

[62] Davuluri, P. S. L. N. (2021). Event-Driven Compliance Systems: Modernizing Financial Crime Detection Without Machine Intelligence. Journal of International Crisis and Risk Communication Research , 339–354. https://doi.org/10.63278/jicrcr.vi.3636

[63] Toh, S., et al. (2011). Data quality assessment for observational studies. Pharmacoepidemiology and Drug Safety, 20(4), 333–339.

[64] Varri, D. B. S. (2020). Automated Vulnerability Detection and Remediation Framework for Enterprise Databases. Available at SSRN 5774865.

[65] Ullman, J. D. (1988). Principles of database and knowledge-base systems. Computer Science Press.

[66] Vadisetty, R., Polamarasetti, A., Guntupalli, R., Rongali, S. K., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2021). Legal and Ethical Considerations for Hosting GenAI on the Cloud. International Journal of AI, BigData, Computational and Management Studies, 2(2), 28-34.

[67] Weiskopf, N. G., & Hripcsak, G. (2013). EHR data quality for clinical research. Journal of the American Medical Informatics Association, 20(1), 144–151.

[68] Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of EHR data quality assessment. Journal of the American Medical Informatics Association, 20(1), 144–151.

[69] Davuluri, P. N. (2020). Improving Data Quality and Lineage in Regulated Financial Data Platforms. Finance and Economics, 1(1), 1-14.

[70] Wilkinson, M. D., et al. (2016). FAIR guiding principles. Scientific Data, 3, 160018.

[71] Keerthi Amistapuram , "Energy-Efficient System Design for High-Volume Insurance Applications in Cloud-Native Environments," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREEICE), DOI 10.17148/IJIREEICE.2020.81209.

[72] Yadav, P., & Steinbach, M. (2013). Mining electronic health records. Springer.

[73] Pandiri, L. Data-Driven Insights into Consumer Behavior for Bundled Insurance Offerings Using Big Data Analytics.

[74] Berg, M. (2001). Implementing information systems in health care organizations. International Journal of Medical Informatics, 64(2–3), 143–156.

[75] Blumenthal, D., & Tavenner, M. (2010). Meaningful use regulation. New England Journal of Medicine, 363(6), 501–504.

[76] Chava, K., Chakilam, C., & Recharla, M. (2021). Machine Learning Models for Early Disease Detection: A Big Data Approach to Personalized Healthcare. International Journal of Engineering and Computer Science, 10(12), 25709–25730. https://doi.org/10.18535/ijecs.v10i12.4678

[77] Dean, B. B., et al. (2009). Data quality in observational studies. Clinical Therapeutics, 31(12), 290–298.

[78] Paleti, S. (2021). Cognitive Core Banking: A Data-Engineered, AI-Infused Architecture for Proactive Risk Compliance Management. AI-Infused Architecture for Proactive Risk Compliance Management (December 21, 2021).

[79] Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.

[80] Gadi, A. L. The Role of Digital Twins in Automotive R&D for Rapid Prototyping and System Integration.

[81] Lazer, D., et al. (2014). The parable of Google Flu. Science, 343(6176), 1203–1205.

[82] Singireddy, S., & Adusupalli, B. (2019). Cloud Security Challenges in Modernizing Insurance Operations with Multi-Tenant Architectures. International Journal of Engineering and Computer Science, 8(12). https://doi.org/10.18535/ijecs.v8i12.4433.

[83] Shah, N. H., & Tenenbaum, J. D. (2012). The coming age of data-driven medicine. Nature Reviews Genetics, 13(6), 395–405.