International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com
Volume 1, Issue 12, 2022
DOI: https://doi.org/10.38124/ijsrmt.v1i12.1189

Data Engineering Techniques for Retail Customer Behavior Analysis

Dileep Valiki[1]

[1]Independent Researcher

Publication Date 2022/12/12

Abstract

The retail sector is one of the most data-intensive industries in the world. Retailers collect a wide variety of data from various sources across the enterprise, from the e-commerce transaction logs to supply chain activities, customer service conversations, and marketing campaigns. The availability of big data in retail is a double-edged sword—while the amount of data presents opportunities, the effectiveness of customer analytics initiatives also relies heavily on data quality and governance. Reliable, governed, and proactive customer data engineering is essential for operational and analytical workloads across the retail ecosystem since analytics-based decisions undertaken by marketers, customer service, and product development teams impact customer experience and ultimately the bottom line. This requires an engineering approach to customer data and a focus on how it is governed, processed, segmented, transformed into features, and served to data scientists, visualization tools, and digital platforms.

The data landscape within retail can be categorized based on the stages of the data life cycle starting from data sources and ingest to data quality and governance, data modeling for Analytics, data storage, pipeline orchestration, and finally feature engineering of customer behavior catered for modeling and analysis. Customer-related data used for data sciences use cases can be broadly divided into two categories: identity data responsible for a single customer's identity and behavioral data that enables analysis of the customers' actions, including visits, purchases, interactions, and conversations, over time and in response to various marketing push messages.

Keywords: *Retail Big Data Analytics, Customer Data Engineering, Retail Data Governance, Data Quality Management, Omnichannel Retail Data, E-Commerce Transaction Logs, Supply Chain Data Integration, Customer Service Analytics, Marketing Data Pipelines, Retail Data Lifecycle, Data Ingestion And Processing, Retail Data Modeling, Analytics-Oriented Data Storage, Pipeline Orchestration, Feature Engineering In Retail, Customer Identity Data, Customer Behavioral Data, Customer Journey Analytics, Personalization And Segmentation, Data-Driven Retail Decisions.*

I. INTRODUCTION

The emergence of e-commerce has made possible the collection of detailed data on the behavior of buyers, including the times, types, characteristics, and quantities of items of interest, as well as sorting, rating, and purchasing decisions. In addition to data from customer interactions with the company's website, data from other social networks and platforms have begun to provide even more insight into customer behavior. Both buzz and sales data have led to changes in the way product suppliers are chosen, prices set, advertising and marketing campaigns designed, and suppliers identified and managed. Tracking the activity of customers during their visits to retail stores has also become possible, with comment cards and redemption notes allowing for more direct involvement and feedback from customers, and a reward system providing incentives for customers to remain loyal to a supermarket or shopping center. As personal data becomes increasingly available and companies help create "digital twins" in the digital world, the ability to analyze customer behavior and preferences will be expanded.

Despite the variety of data generated, identifying important trends or interesting facts is not simply an exercise in sophisticated statistics. Many behaviors cannot be explained simply by observing changes in details over time. Abnormally high sales figures captured by cluster detection algorithms can be further investigated through

Valiki, D. (2022). Data Engineering Techniques for Retail Customer Behavior Analysis. *International Journal of Scientific Research and Modern Technology*, *1*(12), 266–280. https://doi.org/10.38124/ijsrmt.v1i12.1189

statistical techniques to identify surprises. Evaluations of the effectiveness of marketing campaigns can combine statistical hypothesis testing with graphical display tools. Emerging data, mining, analytics, and visualization techniques can help organizations remain profitable in an environment where product life cycles are constantly shortened and competition becomes fiercer. Information systems and data management systems have matured to be able to support the integration of diverse types of data in homogeneous formats, facilitating the use of user-friendly, yet high-performance data mining tools.

- *Overview and Objectives*

Data is an important landscape for many organizations including retailers willing to analyze customer behavior and learn new patterns. Retailers accumulate data from multiple sources such as web applications, point of sales, warehouse management systems, delivery services, etc. The actual challenge is not data storage, but data quality and data governance, ensuring the datasets ingested meet compliance and privacy regulations before and after processing. When analyzing data it is critical to have high-performance models ready to speed up the analysis cycle. Data Engineering delivers the right models to make data-driven decisions in near real-time, such as offering new promotions to specific customer segments or analyzing customers most likely to churn.

Behavioral analysis is an important research area in casinos, e-commerce, and finance. Behavior is usually modeled through metrics based on customer purchases, such as RFM (Recency/Frequency/Monetary Value) analysis. However, these criteria do not take into account all the customer possible interactions with the business. For Retailers, further patterns found in the customers' journey through the data architecture can boost the quality of the analysis. Recognizing that RFM is a poor metric to explain customer behavior in a Casino (L. Viana et al., 2017), a Customer Behavior Analysis based on Events is proposed. The foundation for this analysis is data engineering, and in detail, a set of principles regarding data sources, quality, modeling, processing, and features driving the final Customer Behavior Analysis is presented. The proposal demonstrates how behavior can be modeled through Events and not restricted to purchases.

This section uses data engineering to provide RFM-like models and events to classify different groups of Retail customers on their behavior, improve strategies to increase sales and reduce churn, and decrease the number of customers re-activated on a campaign.

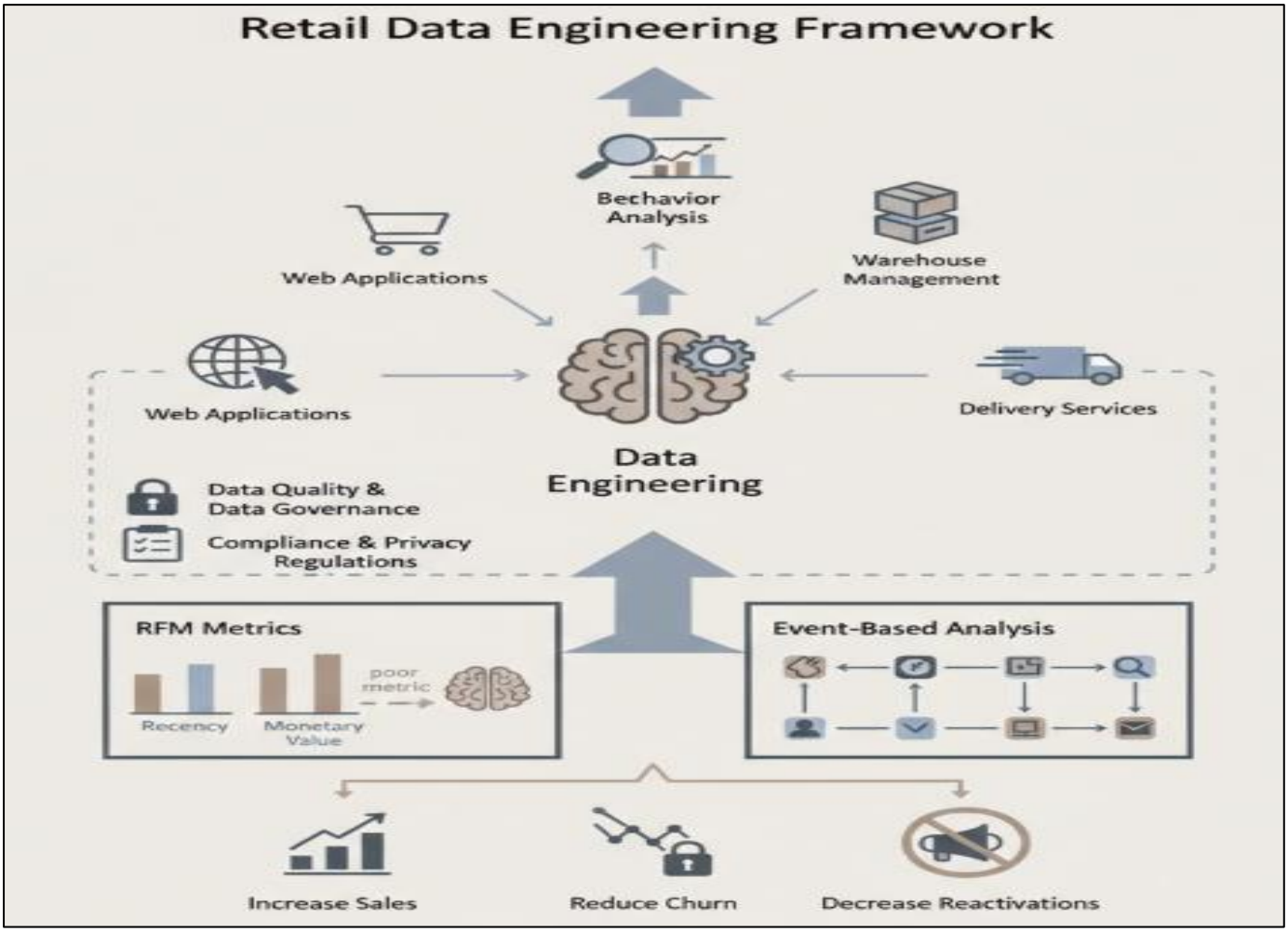
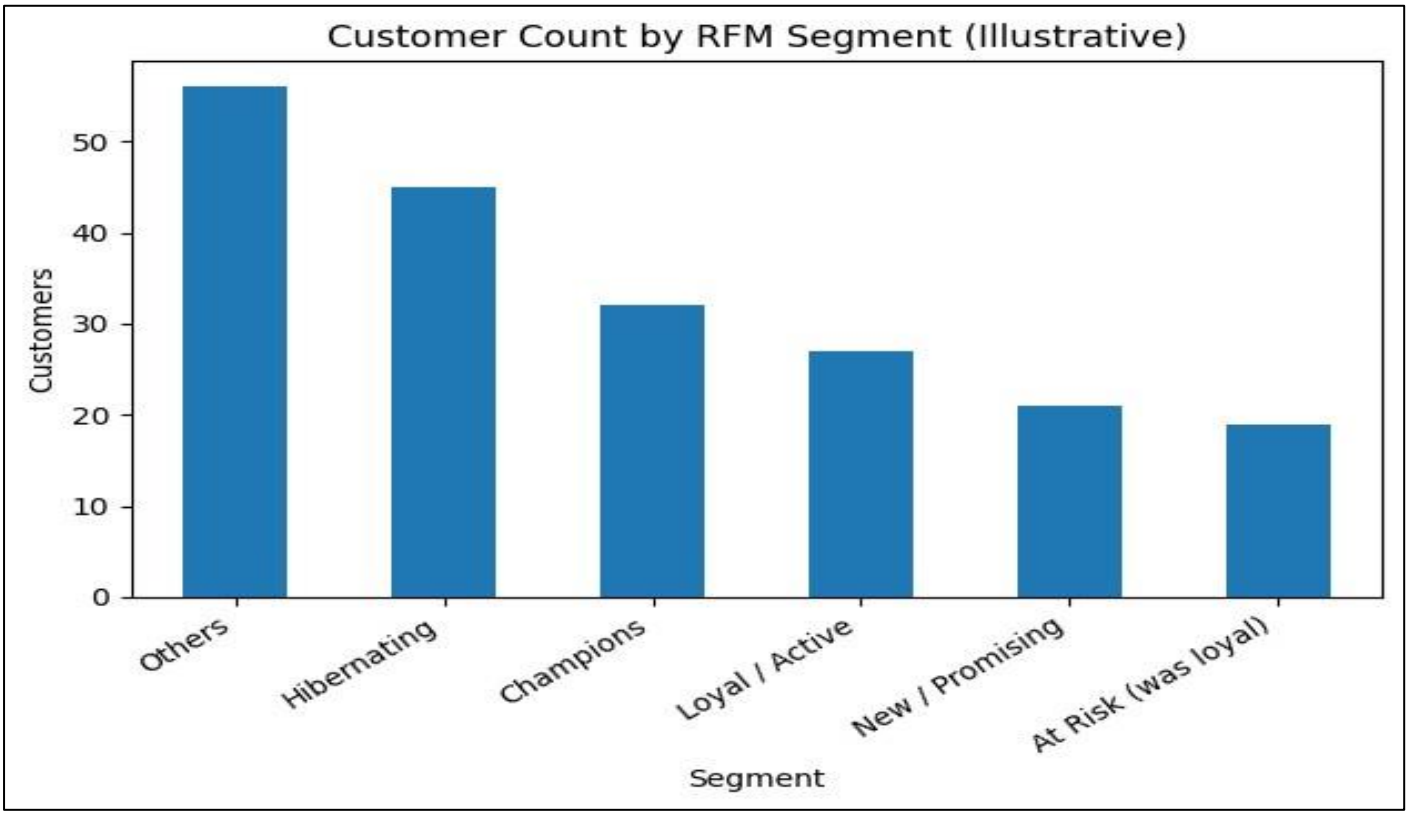


Fig 1 Beyond Transactional Metrics: An Event-Driven Data Engineering Framework for Advanced Customer Behavioral Analysis and Churn Mitigation in Retail Ecosystems

II. DATA LANDSCAPE IN RETAIL

In retail, the need for data is ubiquitous. Data for behavioral analysis typically exists in multiple sources, including structured transactions databases, semi-structured web and mobile platforms, and unstructured social media. Getting the needed data into a data analytics platform is essential for retail companies wishing to gain experience-based customer insights. Data quality, integrity, availability, and governance for the analysis of customer behavior are critical to gaining trust and adoption by analytics and business teams.

Retailers examine customer behavior to identify opportunities and assist in long-term planning. Analytics requires a foundation of shared data from multiple sources, spanning holidays, seasons, and years, and is typically built from the details of every transaction and customer event, so elements of the entire customer journey can be explored. Even with data at scale in a data warehouse or lake, it needs to have the right format, cleansing, accessibility, and depth to meet performance benchmarks for batch and real-time processing.



Graph 1 Equation A) RFM constructs (Recency, Frequency, Monetary)

The frames RFM as: customers who purchased *recently*, *frequently*, and spend *more* are more valuable; it also notes typical 1–5 scoring per dimension.

➢ *Define the raw transactional data*
Assume a transaction table with:

- customer i
- transaction times $t_{i1}, t_{i2}, \dots, t_{in_i}$
- transaction amounts $a_{i1}, a_{i2}, \dots, a_{in_i}$
- analysis "as-of" (reference) date T

Let the **window** be $[T - W, T]$ (e.g., last 180 days).

➢ *Recency (days since last purchase)*

- *Last transaction time for customer i:*

$$t_i^{\text{last}} = \max\{t_{ij}\}$$

- *Recency (time gap from reference date):*

$$R_i = T - t_i^{\text{last}}$$

If measured in days:

$$R_i(\text{days}) = \text{days}(T) - \text{days}\left(t_i^{\text{last}}\right)$$

✓ *Interpretation:*
smaller R_i is "better" (more recent).

➢ *Frequency (Purchase Count in the Window)*

- *Count transactions in the window:*

$$F_i = \sum_{j=1}^{n_i} \mathbf{1}\left[t_{ij} \in [T - W, T]\right]$$

where $\mathbf{1}[\cdot]$ is an indicator (1 if true, else 0).

- *Monetary (total spend in the window)*

$$M_i = \sum_{j=1}^{n_i} a_{ij} \cdot \mathbf{1}\left[t_{ij} \in [T - W, T]\right]$$

- *Data Sources and Ingest*

From a data engineering perspective, retail customer behavior analysis encompasses the techniques and frameworks that facilitate data ingestion, quality assurance and governance, storage, processing, feature engineering and modeling—all with the end goal of producing a set of well-engineered features that support customer-vicinity behavior-based machine-learning models in a cloud platform environment. Many of these features support a round-trip paradigm in which machine-learning models iteratively segment customers into target groups for customer relationship management programs; a second set of features support dedicated churn prediction and predictive lifetime-value modeling projects.

Data from multiple sources are needed to analyze customer behavior. RFM models require transaction data, while churn-prediction models similarly benefit from the consideration of customers' recency, frequency and monetary activity characteristics. Marketing campaigns, events surrounding the campaigns, and other contextual factors are also important sources of information. Attribute information from both internal and external exploration also supplements the behavioral view of the customers and their surroundings. To support this comprehensive collection of sources, a data pipeline collects the data in a "best-available quality" mode, reshaping the data along the way to fit a behavioral-event schema.

- *Data Quality and Governance*

Reliable data is a prerequisite for insights about customer behavior. Yet ensuring high data quality is a major challenge because data is often ingested into the data platform in raw form and can contain errors due to incorrect schemas, inconsistencies, repetition, and other issues (e.g., missing values, outliers). Data quality assurance should be built into data pipelines along with the ETL/ELT processes that prepare the data for analytics. Such validation uses a set of wholesome rules to automatically assess quality at different stages within the pipeline, applying remediation where possible and alerting users as needed. Performance can be monitored to track execution time and resource consumption across different data sources, pipelines, or components.

Many organizations have established enterprise data governance frameworks. These policies cover not only the availability of relevant, reliable, and high-quality data but also how data can be shared and used. Data catalogs act as central repositories for organizing metadata, including data definitions, schemas, lineage details, and contact points for different data domains that users can consult for clarification on ambiguous data elements—adopting a common language makes collaboration easier. Automated data lineage tracking provides insight into data transformations, helping users trace the steps that a particular attribute has gone through, enabling auditing efforts, and allowing them to trust the quality of the combined dataset for subsequent analytics.

III. DATA MODELING FOR BEHAVIORAL INSIGHTS

RFM, or recency/frequency/monetary analysis, is a classic marketing analysis technique that works on the principle that customers who recently purchased, who purchase frequently, and who spend money contribute the most to a company's bottom line. Typically, scores between one and five are assigned to recency, frequency, and monetary dimensions, with higher values representing a smaller customer base but higher business impact. External tools such as DataRobot or Alteryx can be used to create radial plots of the derived RFM scores. RFM analysis can be conceptualized in terms of an intermediate, higher-level feature for customers in a data lake, rather than requiring raw transactional data to be available for each new analysis.

The fundamental idea behind behavioral personas is that a company should dedicate specific resources to developing strategies around certain customer groups. Instead of creating different departments for each persona, companies may require a central analytics function to uncover the personas, often based on data from existing resources. Recent work on behavioral personas employs a persona identification framework that provides an intuitive methodological description along with the technical motivations and considerations that underpin the proposed solutions.

- *Entity Resolution and Customer Identity*

Although every customer interaction is captured in a unique session entry, once customer identity is established, behaviors must be aggregated into customer records for analysis. In the presence of multiple identity spaces, implementing a customer-centric identity store and entity resolution procedure is essential. The identity store maintains identities and links across identity spaces (e.g. credit/debit card transactions, ecommerce login name, loyalty identifiers), while the entity resolution process seeks to resolve these identities into a unique customer identity for each behavioral record.

The identity store supports two types of identity links: 1) deterministic links, which are created when a customer logs in to make an ecommerce purchase, and 2) probabilistic links that are based on a combination of credit/debit card transactions (same geographical location, common time, common point of sale, etc). These deterministic and probabilistic identity links are used when an identity needs to be resolved to a customer identity for the features being generated, thus allowing for customers to be identified even when they have not logged in to the website for the last transaction or are non-registered users.

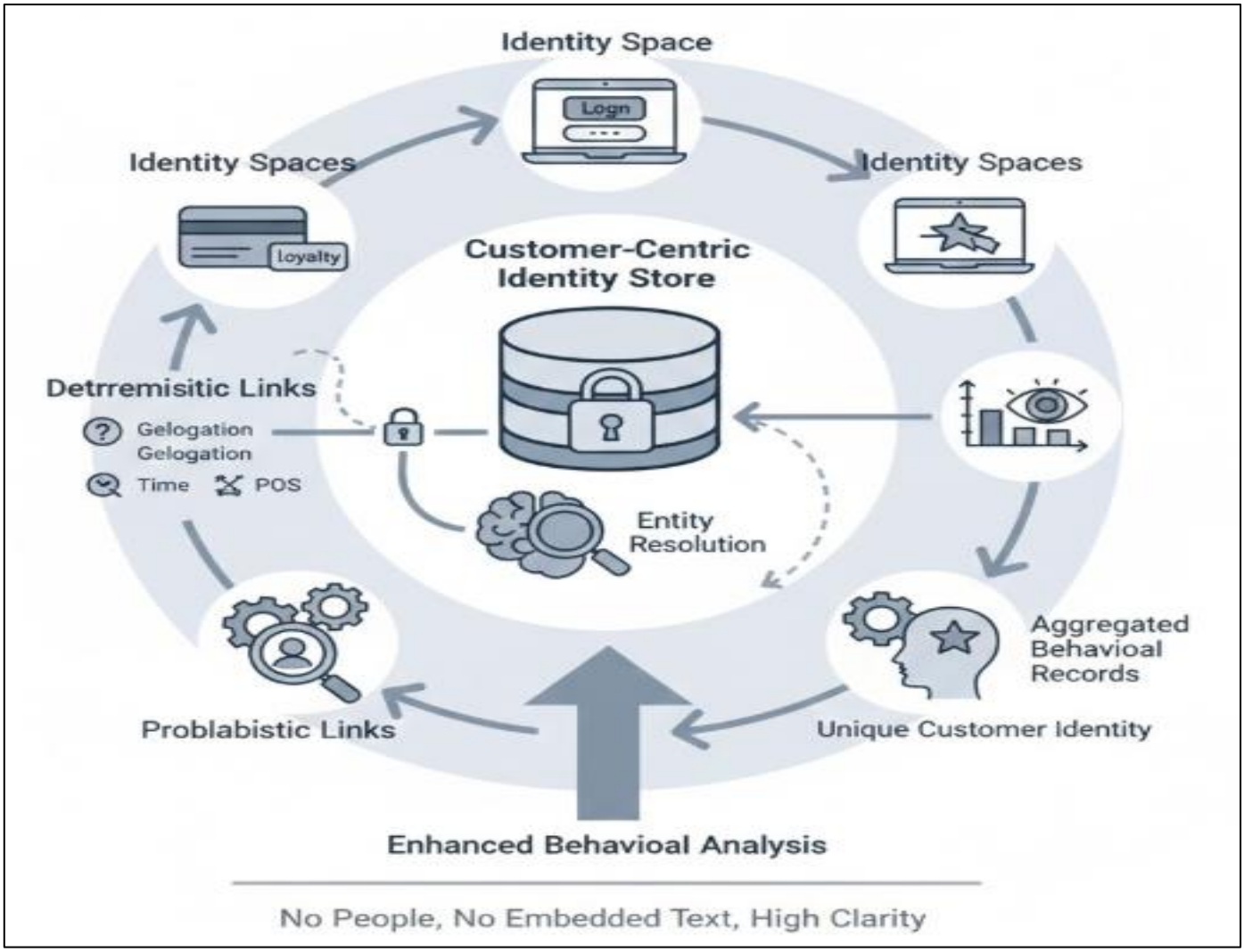


Fig 2 Customer-Centric Entity Resolution: A Hybrid Deterministic-Probabilistic Identity Framework for Cross-Channel Behavioral Aggregation

➢ *Behavioral event schemas*

Behavioral event schemas address the conceptual design of the information captured. When modelling web and app interactions in a detailed way, the informational foundation also commonly follows the behavior of the clickstream, using a session as the core concept. A session consists of a set of ordered browser or application events associated with the same user in a specific timeframe. When analyzing web interactions, the data is typically organized in the form of sessions. Each session captures the interactions of a specific user in a timeframe where interactions to other sites outside the main site do not take place.

Combining the notion of session with the concept of replaying a session enables web event schemas to naturally cover a wide variety of topics relating to customer behavior analysis and simulation. The model contains all individual session interactions, enabling analysis of patterns that predate transactions, bounce rates, conversion rates, page exits, flow patterns, touch points, and more. All web event captures are identified by their respective session IDs, allowing the replaying of a session to show a customer navigation as a video.

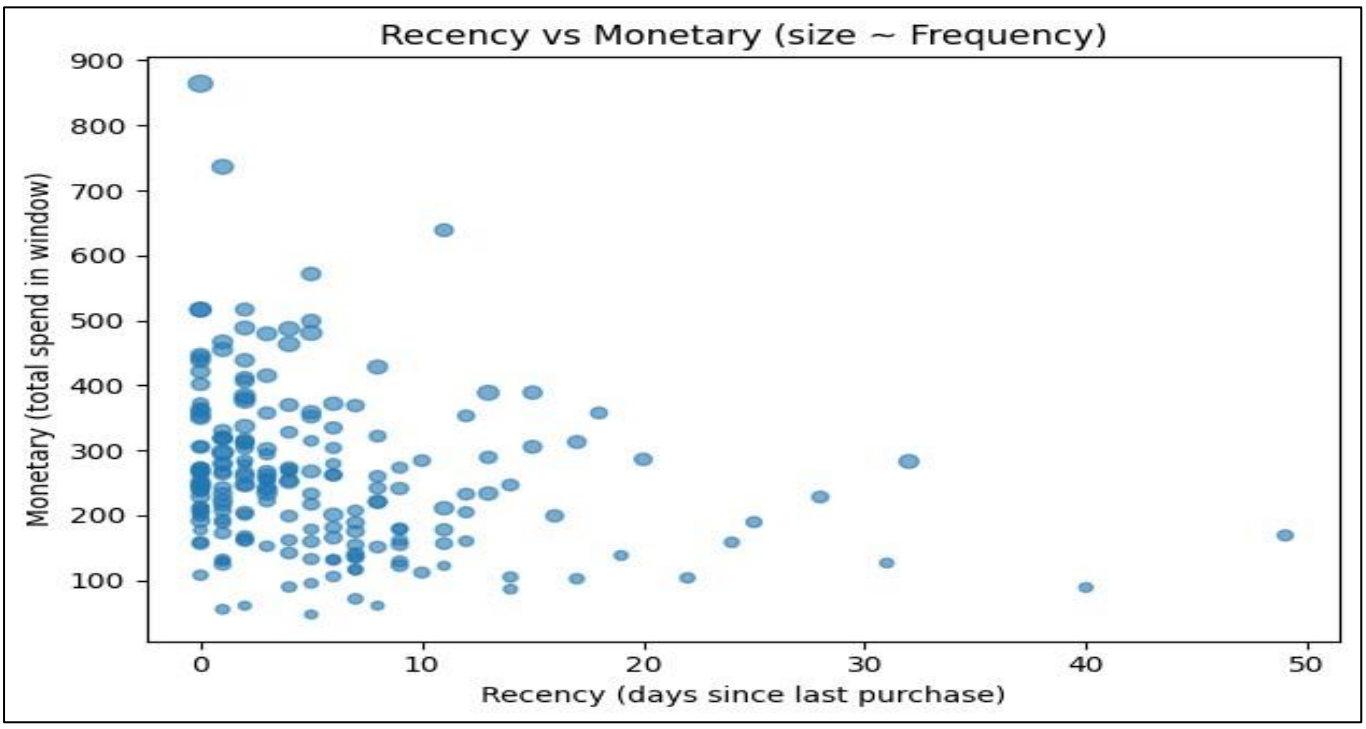
Apps also track similar interactions, but often using different types of events than browser styles. The paradigm behind analyzing these interactions is typically not identifying sessions but observing the sequence of actions and types of actions taken by the user. Therefore, app event schemas capture the individual event patterns for each user. Although not structured as sessions, tools for sequence mining associate visits to different categories with specific user types.

IV. DATA STORAGE ARCHITECTURES

When discussing data storage architecture, one often hears about the conflict between data lakes and data warehouses. Data lakes accept structured and unstructured content and provide limited governance and quality assurance capabilities. Data from a data lake is often replicated into a data warehouse or similar structure, from which production workloads are run. The data warehouse's primary demand is for well-structured data. The lake and warehouse models thus serve opposite purposes, amplifying the demand for both. The concept of a lakehouse derives from the observation that the two needs co-exist in almost all enterprises. By applying carefully considered governance, quality, and optimization to a data lake, it is possible to support both consumption models

with a single data architecture. The lakehouse principle allows analysts, data scientists, and data engineers to draw upon one set of shared assets, while reducing the risk of introducing defective or inconsistent data into an enterprise's marketing mix.

Most retail enterprises combine both batch and streaming storage models, while few rely predominantly on just one. Although streaming storage systems can be very lightweight, such as Apache Kafka, more frequently they are implemented using the additional capabilities incorporated into distributed file systems from cloud hyperscalers like Microsoft, Amazon, and Google. It is probable that in most non-bank financial enterprises a purpose-built streaming data store would be defined as a secondary store and only created when needed—indeed, beyond necessary disaster recovery and security requirements, the only other redundancy normally considered is to ensure that a suitable streaming store is provisioned and available for near-real-time applications.



Graph 2 Equation B) Converting Raw R, F, M into 1–5 Scores (Quintile Scoring)

The assigning scores 1–5 per dimension.

➢ *Rank-Based Binning (Quintiles)*

Let $Q_k(\cdot)$ denote the k-th quintile threshold (20%, 40%, 60%, 80%).

For Frequency (higher is better):

$$F_i^{\text{score}} = \begin{cases} 1 & F_i \le Q_{20}(F) \\ 2 & Q_{20}(F) < F_i \le Q_{40}(F) \\ 3 & Q_{40}(F) < F_i \le Q_{60}(F) \\ 4 & Q_{60}(F) < F_i \le Q_{80}(F) \\ 5 & F_i > Q_{80}(F) \end{cases}$$

For Monetary (higher is better) same form:

$$M_i^{\text{score}} = \text{QuintileBin}(M_i)$$

For Recency, because *lower* recency is better, we reverse it:

- First compute the quintile bin the same way (lower recency → lower bin).
- Then invert:

$$R_i^{\text{score}} = 6 - \text{QuintileBin}(R_i)$$

So the most recent customers end up with score 5.

➢ *RFM code and/or composite score*

A common representation:

$$\text{RFM\_code}_i = (R_i^{\text{score}}, F_i^{\text{score}}, M_i^{\text{score}})$$

Sometimes collapsed to a single number:

$$\text{RFM\_sum}_i = R_i^{\text{score}} + F_i^{\text{score}} + M_i^{\text{score}}$$

Or weighted (if a business values one dimension more):

$$\text{RFM\_weighted}_i = w_R R_i^{\text{score}} + w_F F_i^{\text{score}} + w_M M_i^{\text{score}}, \quad w_R + w_F + w_M = 1$$

➢ *Data Lakes versus Data Warehouses*

Data Lakes enable storing structured and unstructured data with minimal data quality control, usually leveraging distributed file systems. In combination with cloud computing, they enable very low ingestion costs and often serve as the one and only data storage of an organization because they can fulfill multiple data

storage roles (operational, warehouse, etc.) and allow a Data-as-a-Service architecture.

These rhetorical advantages have made lakes popular in retail customer behavior analytics. Yet, they only reduce upfront costs and management complexity at the cost of making data harder and thus costlier to consume, risking a bottleneck for analytics. They are nevertheless optimal for Audio-Video Processing or Internet-of-Things Data. But preventing more than three lake usage patterns helps preserve the Data Warehouse's advantages over lakes as a petabyte-scale repository for analysis-ready data. A Lakehouse Storage Structure, combining a lake and a warehouse, may fulfill these patterns.

➢ *Lakehouse Principles for Retail Analytics*

An architectural evolution has combined the advantages of data lakes and data warehouses in the "lakehouse," a data management technology that seeks to harness the advantages of each model while eliminating their major drawbacks. At a general level, lakehouses are a set of integrated data management processes and services that enable the following:

- Unified storage for all data assets within a single system that is used as a central repository by different teams: analytics, machine learning, artificial intelligence, reporting, and business intelligence.
- Direct access to data from a wide range of analytics/microservices tools for use cases involving numerous ad hoc queries.
- Support for data sharing and collaborative development processes that span different business domains.
- Support for schema evolution and governance capabilities that last while the data is available.

Numerous embellishments on this general story have been created by the industry players that are involved in the space, including technology providers (such as Databricks), large cloud providers (such as Microsoft Azure and Amazon Web Services), and data management companies (such as Cloudera). All of these companies have invested massively in customer-facing lakehouse products. Major success stories have already surfaced: for example, notable customers have been identified for Databricks, and Microsoft Azure Synapse has been deployed across many organizations.

In the case of retail organizations, the major principles associated with a lakehouse definition can be tailored specifically to fit retail business activities, data requirements, and data strategy concepts.

V. DATA PROCESSING AND PIPELINE ORCHESTRATION

Data processing in data engineering comprises a variety of techniques designed to convert ingested raw input data into useful derived data for consumption—whether for analysis, data science, machine learning, or as input to business processes. Two major considerations are whether the processing is performed in batch mode or in streaming mode, and therefore incurring latency; and whether the processing is performed via Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT) patterns. The choice between the two patterns arises from the architectural layout, whether data ingestion proceeds from an operational environment to a separate data model designed for consumption or insight generation, or rather a central store serves as a consolidating repository for raw data supporting diverse use cases.

Batch processing creates sets of derived values from longer-lived raw data stores in response to schedule-based orchestration. Event-driven streaming processing reacts dynamically to changes in close proximity, leveraging a real-time or near real-time capability to produce small lateness. It can either retract old data values or append new data values as they changes occur.

➢ *Batch and Streaming Processing*

Data processing in a retail context can be broadly categorized into batch and streaming processing. Many retail customer behavior analytical workloads are inherently batch driven, utilizing data stored in a data warehouse. In practice, however, data is often generated in real-time, and business stakeholders envision real-time down-streaming use cases that are partly realized through batch processing or by making data available sooner in a streaming format and iteratively refining it with successive batch processing cycles. Customer behavior analytics continues to be an active area of research. Results are used to define customer segments, generate customer personas, and create audience segments for marketing campaigns in retail business functions. Retail customer behavior analytical workloads can be performed using traditional batch processing and simple tools, but business stakeholders also require quick-and-dirty analyses in response to business queries to further define behavioral segments for various business functions. A specific group of behavioral segments may also change rapidly, raising the need and importance of supporting business queries with data stored in a streaming format.

Load patterns of different retail business functions are different, with the retail sales function generating the hottest data and other business functions supporting real-time and near-real-time use cases with some delay. A significant portion of business functions is also supported by batch processing, leveraging existing historical data to respond to queries but with delays beyond business tolerance. Although the batch processing approach satisfies business needs in many cases, the evolving needs of some business functions are being enabled by streaming, including key functions of generating marketing audience segments using customer behavior segments and personas. A small library of reusable components for customer behavior analysis is available for business stakeholders to consume directly to support business queries in a timely manner while enabling data engineers to refine these analyses incrementally in a batch-processing manner.

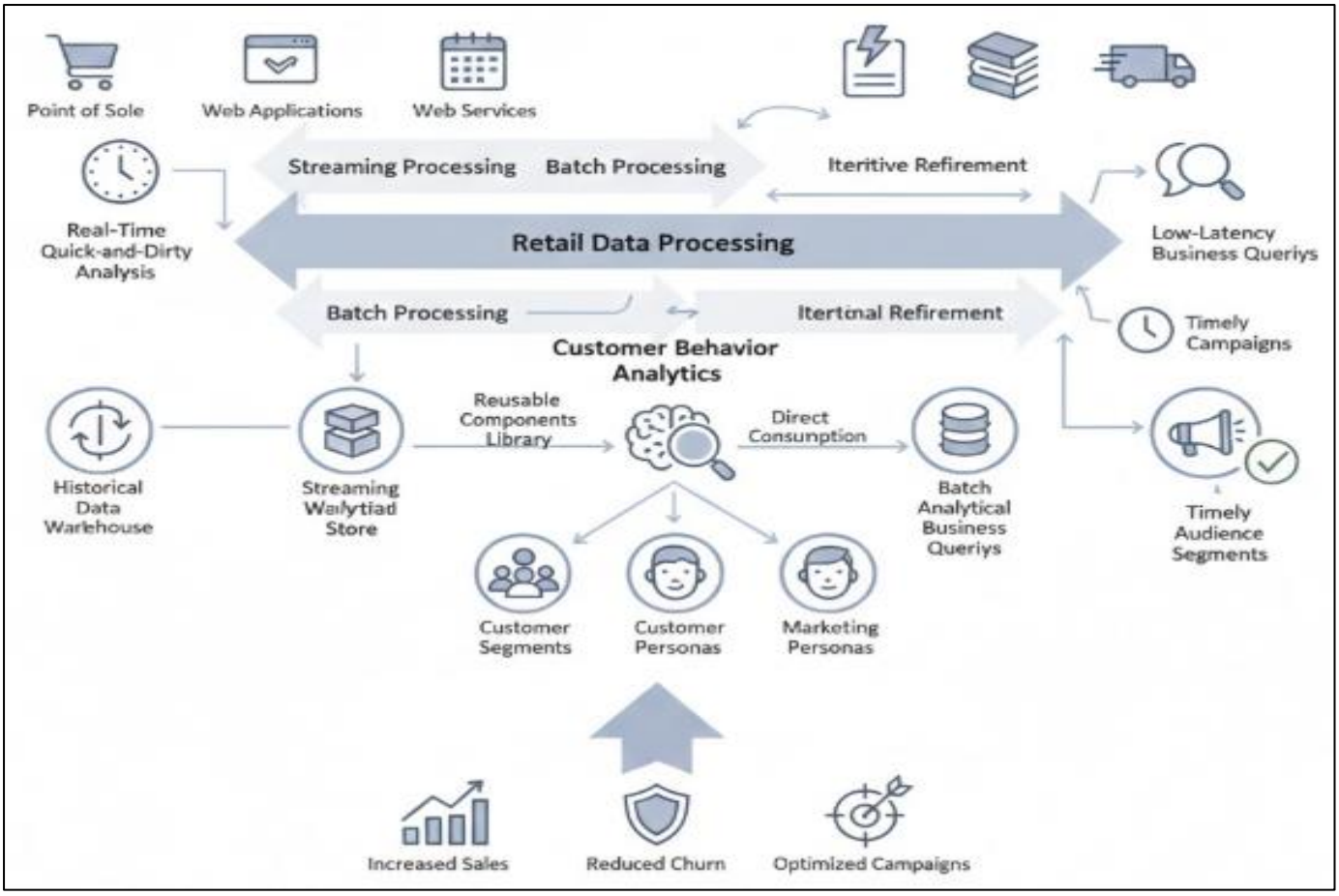


Fig 3 Hybrid Analytical Architectures in Retail: Integrating Stream-Batch Convergence and Reusable Component Libraries for Real-Time Customer Persona Synthesis

➢ *ETL/ELT Patterns and Tools*

ETL (extract, transform, load) and ELT (extract, load, transform) pipelines flow through retail data landscapes to automate the ingestion and preparation of data for analysis. Deciding between the contrasting paradigms depends upon the architecture, purpose, resources, and tools being used. For example, a classic data warehouse consuming structured sources through scheduled jobs would favour ETL. In contrast, trivial data lakes using ungoverned diverse sources might exploit ELT, although this is usually a bad idea due to performance, quality, and maintenance implications. Retail use cases across both batch and streaming process engines could benefit from a hybrid approach, implemented in Apache Spark or Kafka.

The top level of a lakehouse architecture runs a query layer that encapsulates data in the underlying lake and serves many types of analysis. Data in the query layer is not directly written to; instead it is built through periodic jobs in batch or streaming. Within a batch job, data engineers create, refresh, materialise, or clean tables. Within a streaming job, data is continuously pushed into a table or view that can be used for visualisation or machine learning. Storing feature sets in these ways simplifies consumption and makes data engineering more scalable.

VI. FEATURE ENGINEERING FOR CUSTOMER BEHAVIOR

Feature engineering maps data into features suitable for particular modeling tasks. In the context of customer behavior, various feature-generation exercises shape behavioral signal features that inform interventions, campaign target selection, and handwritten digital-signage messages. Event data schemas described in the previous section can be mined, scored, and combined into constructs focused on customer Recency, Frequency, and Monetary (RFM) activity as a basic proxy for brand engagement. Multiple options exist for scoring RFM activity schemas—time since last purchase, purchase count, and dollar value—but it's advisable to experiment with brute-force or linear regression curves to identify the combination that drives ROI over test-and-learn campaigns. RFM constructs can also be leveraged for modeling-based lookalike audience building and persona definition.

Loyalty schemes exist to create customer profiles in retail transaction data and stimulate customer frequency by offering rewards for purchase commitment. An analysis-driven or rules-driven approach can identify brand categories that strongly influence customer loyalty: "buy a certain amount of beer or freshwater fish" might drive "buy beer" versus "buy beer and drink-label whisky" messages. For high-frequency categories, the challenge is to reduce churn and/or maximize value for the retailer. Minority-class classifiers could support a second or

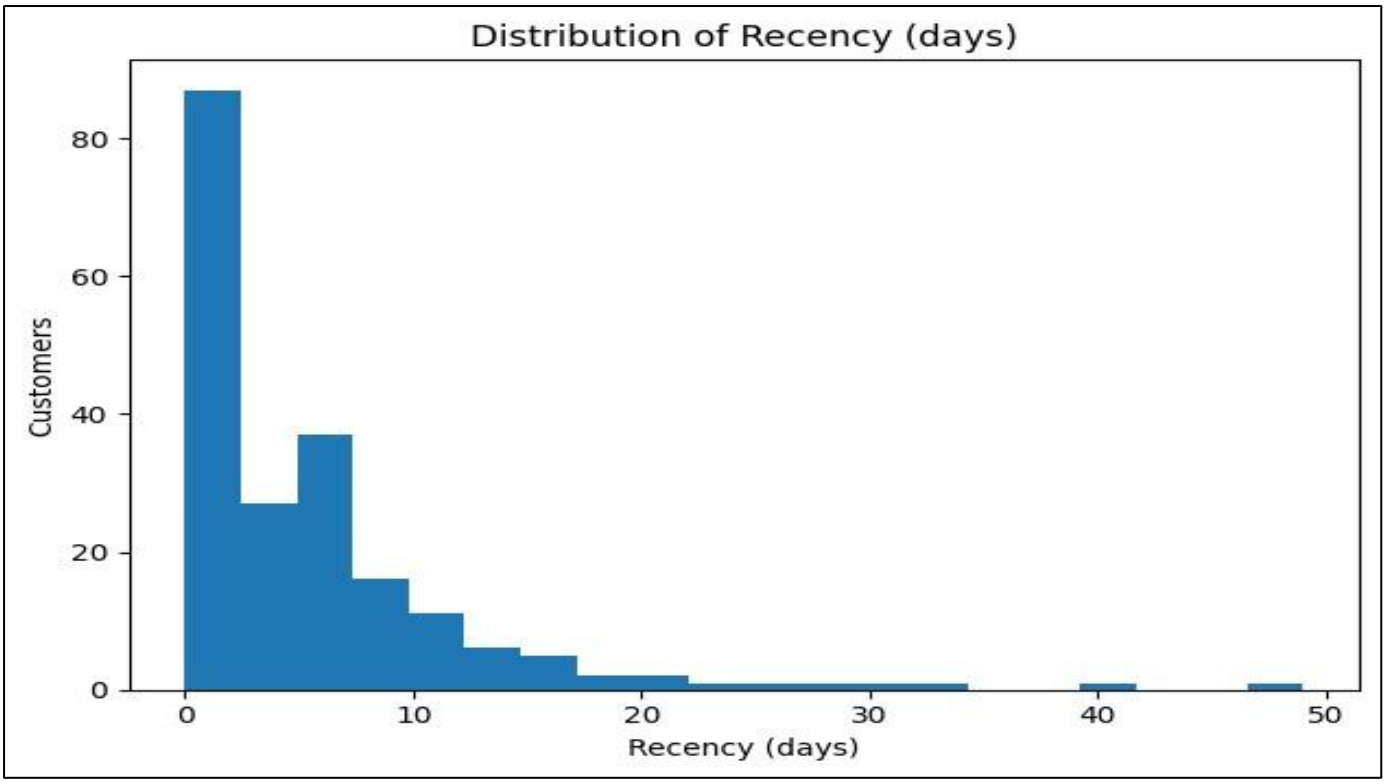
tertiary segmentation strategy for retention messages—for example, differentiating non-KYC-ing customers with an acute reactivation signal ("last purchase OOL")—or control experiment insights.

➢ *RFM and Recency/Frequency/Monetary Constructs*

Definition and operationalization of the Recency, Frequency, and Monetary (RFM) constructs can embellish the complete data documentation of retail datasets. A concept heavily based on the following two questions, can also be discussed through RFM analysis - (a) how often customers visit the store, (b) how much is the total money spent per customer.

From the marketing perspective it can be concluded that, improving retention rate with current customers is easier than attempting to acquire new customers, thus the marketing resources are usually more effective in targeting existing customers. Intrigued by this concept, RFM is an essential tool in marketing strategy. RFM segmentation is used worldwide because it can help retailers forming a proper marketing campaign such as allocating marketing budgets and resources effectively. RFM is a powerful and widely used model that utilizes customer transaction data for behavioral modeling and offers critical information regarding the strategy.

RFM analysis and modeling keeps track of customer behavior (the recency, frequency, and monetary spent by the customers) and identifies those who deliver (or are very likely to) the most value to the business. RFM is mainly used by companies focused on retaining existing customers or upselling, with main goal of driving individuals from light spenders or no-spending customers to heavy spenders.



Graph 3 Equation C) Event-driven behavioral features (sessions + journey metrics)

The emphasizes event schemas, often via sessions for web clickstream, and sequences for app behavior.

➢ *Sessionization from Raw Events*

Suppose we have events for user$i$:

$$(e_{i1}, e_{i2}, \dots), \quad \text{with timestamps } (\tau_{i1} \le \tau_{i2} \le \cdots)$$

Pick an inactivity threshold Δ (e.g., 30 minutes).

Define a new session whenever the time gap is large:

$$\text{new\_session}(k) = \mathbf{1}\left[\tau_{ik} - \tau_{i(k-1)} > \Delta\right]$$

Then session id increments cumulatively:

$$s_{ik} = 1 + \sum_{m=2}^{k} \mathbf{1}\left[\tau_{im} - \tau_{i(m-1)} > \Delta\right]$$

➢ *Session-level features*

For a session s of user i, with events E_{is}:

- *Session Length:*

$$\text{dur}_{is} = \max(\tau) - \min(\tau), \quad \tau \in E_{is}$$

- *Page/Event Count:*

$$\text{events}_{is} = |E_{is}|$$

➢ *Common Journey Metrics the Alludes to (Bounce, Conversion, Exits, Touchpoints)*

The lists metrics like bounce rates and conversion rates.

Let:

- S = set of sessions in period
- $\text{purchase}(s) = 1$ if session s contains a purchase event
- $\text{singlepage}(s) = 1$ if only 1 page view (or 1 event of certain types)

Then:

- *Conversion Rate*

$$\text{CR} = \frac{\sum_{s \in S} \text{purchase}\,(s)}{|S|}$$

- *Bounce Rate*

$$\text{BR} = \frac{\sum_{s \in S} \text{singlepage}\,(s)}{|S|}$$

You can then aggregate per customer:

$$\text{CR}_i = \frac{\sum_{s \in S_i} \text{purchase}\,(s)}{|S_i|} \quad , \quad \overline{\text{dur}_i} = \frac{1}{|S_i|} \sum_{s \in S_i} \text{dur}_{is}$$

➢ *Behavioral Segmentation and Personas*

A behavioral segmentation is a knowledge structure that ties intents and needs to a set of option attributes. It provides a perspective to interpret a broader array of other data. For example, retailers typically also collect product categories, local stores, and salespersons involved in the transaction. All those attributes are seen as options to satisfy the underlying intent or need. Moreover, profiles built from a single segmentation are unlikely to tie those attributes together with meaningful insights. Collecting multiple such segments for the same individuals enables intents and needs to emerge and become the explanatory structure behind the other data.

Retail analytics is a prime use-case of personas. Here, the underlying segmentation represents a frontline service delivered to all the clients. One of the missions of retail analytics is to infer how to fulfill that need more effectively. In this context each of the product catalog items is managed in relation to that segmentation. Using those records, it is expected that a reasonably complete sample is observed. Behavioral data is often detailed enough to let the model represent other data only loosely related with the primary need.

VII. CONCLUSION

Combining a succinct summary of key arguments with forward-looking remarks, the conclusion synthesizes core messages and highlights points for future consideration.

The data landscape in retail is often characterized by stagnation, as companies frequently focus on tactical uses for data in support of sales, marketing, finance, and inventory processes. These stand-alone tactical patterns yield no lessons for the future. To leverage the full breadth of data resources available for deep same-store sales growth, companies must consider data engineering outside of these tactical silos. The answer to this dilemma lies in enterprise data engineering for data analytics and machine learning, supporting resumed same-store sales growth through customer behavior.

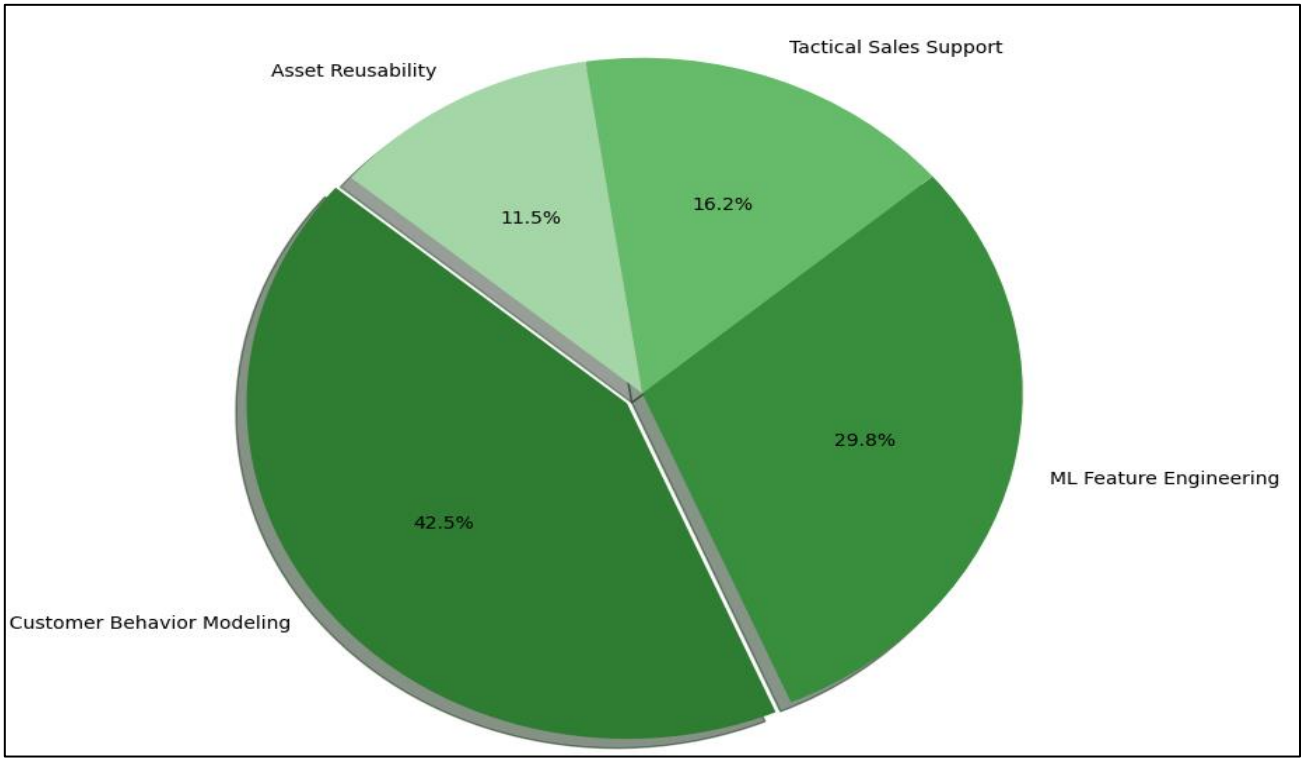


Fig 4 Growth Driver Prioritization

Engineers create scalable pipelines that fuse real-time and day-end batch data streams into unified models of customer behavior. Retrospective investigation using these data pipelines provides insights, while also producing analytic and ML feature assets. Building features from these assets allows marketers and analysts to conduct instant segmentation, filtering data selected by multiple behavioral dimensions using common analytic tools of their choice.

- *Key Takeaways and Future Directions*

The overarching goal of effective data engineering is to regularly present the right analytical data to power management decisions and enable data product features that enhance profitability and customer satisfaction. Without data engineering, analytic development teams would have to manually gather, join, cleanse, and prepare the data for their analysis with every new analytic request. Furthermore, without thoughtful data governance, the required datasets can quickly degrade in quality as source systems change, new data sources are added, and knowledge about how the analytic datasets are created is lost over time.

In retail, the volume of data to manage is particularly large, and there is a constant requirement for new customer behavior analyses, such as studies to guide promotions, loyalty programs, store design, and customer experience improvements. Data management teams must therefore provide a well-governed reservoir of high-quality data to meet ever-growing demand. To do so they consume data not only from traditional enterprise systems—transactional, distribution, inventory, customer relationship management, enterprise resource planning systems—but also from an increasing number of emerging sources, such as social media, website clickstream, mobility, and sensor data.

REFERENCES

[1]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, 265–283.

[2]. Kalisetty, S. Leveraging Cloud Computing and Big Data Analytics for Resilient Supply Chain Optimization in Retail and Manufacturing: A Framework for Disruption Management.

[3]. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734–749.

[4]. Kothapalli Sondinti, L. R., & Syed, S. (2022). The Impact of Instant Credit Card Issuance and Personalized Financial Solutions on Enhancing Customer Experience in the Digital Banking Era. Universal Journal of Finance and Economics, 1(1), 1223. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1223

[5]. Arasu, A., & Kaushik, R. (2014). Data cleansing: A context dependent approach. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, 135–146.

[6]. Annapareddy, V. N. (2022). Integrating AI, Machine Learning, and Cloud Computing to Drive Innovation in Renewable Energy Systems and Education Technology Solutions. Available at SSRN 5240116.

[7]. Armbrust, M., Das, T., Davidson, A., Ghodsi, A., Or, A., Rosen, J., Stoica, I., Wendell, P., Xin, R., & Zaharia, M. (2021). Delta Lake: High-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411–3424.

[8]. Kommaragiri, V. B., Gadi, A. L., Kannan, S., & Preethish Nanan, B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.

[9]. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50–58.

[10]. 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

[11]. Babcock, J., Chaudhuri, S., & Das, G. (2004). Dynamic sample selection for approximate query processing. Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, 539–550.

[12]. Sriram, H. K. (2022). Advancements in Credit Score Analytics using Deep Learning and Predictive Modeling Techniques. *Available at SSRN 5255128*.

[13]. Bifet, A., & Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. Proceedings of the 2007 SIAM International Conference on Data Mining, 443–448.

[14]. Muthusamy, S., Kannan, S., Lee, M., Sanjairaj, V., Lu, W. F., Fuh, J. Y., ... & Cao, T. (2021). Cover Image, Volume 118, Number 8, August 2021. *Biotechnology and Bioengineering*, *118*(8), i-i.

[15]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[16]. Annapareddy, V. N. (2022). AI-Driven Optimization of Solar Power Generation Systems Through Predictive Weather and Load Modeling. *Available at SSRN 5265881*.

[17]. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209.

[18]. Chakilam, C., Suura, S. R., Koppolu, H. K. R., & Recharla, M. (2022). From Data to Cure: Leveraging Artificial Intelligence and Big Data Analytics in Accelerating Disease Research and Treatment Development. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v9i3.3619

[19]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

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

[21]. Das, T., Zhu, A., Li, S., Narayanamurthy, S., & Bhat, P. (2013). Distributed and fault-tolerant streaming computation in Spark. Proceedings of the ACM Symposium on Cloud Computing, 1–12.

[22]. *Pallav Kumar Kaulwar, "Designing Secure Data Pipelines for Regulatory Compliance in Cross-Border Tax Consulting," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREEICE), DOI 10.17148/IJIREEICE.2020.81208*

[23]. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113.

[24]. *Paleti, S. (2022). Financial Innovation through AI and Data Engineering: Rethinking Risk and Compliance in the Banking Industry. Available at SSRN 5250726.*

[25]. DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., & Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. Proceedings of the 21st ACM Symposium on Operating Systems Principles, 205–220.

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

[27]. Dwork, C. (2008). Differential privacy: A survey of results. Proceedings of the 5th International Conference on Theory and Applications of Models of Computation, 1–19.

[28]. *Gadi, A. L., Kannan, S., Nandan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87–100. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1296*

[29]. Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 1–16.

[30]. Koppolu, H. K. R., Recharla, M., & Chakilam, C. Revolutionizing Patient Care with AI and Cloud Computing: A Framework for Scalable and Predictive Healthcare Solutions.

[31]. Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. Marketing Science, 24(2), 275–284.

[32]. Pandiri, L. The Future of Commercial Insurance: Integrating AI Technologies for Small Business Risk Profiling. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI, 10.*

[33]. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

[34]. Meda, R. Enabling Sustainable Manufacturing Through AI-Optimized Supply Chains.

[35]. Ghemawat, S., Gobioff, H., & Leung, S. T. (2003). The Google file system. Proceedings of the 19th ACM Symposium on Operating Systems Principles, 29–43.

[36]. 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.

[37]. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

[38]. Inala, R. Advancing Group Insurance Solutions Through Ai-Enhanced Technology Architectures And Big Data Insights.

[39]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[40]. Vadisetty, R., Polamarasetti, A., Guntupalli, R., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2022). AI-Driven Cybersecurity: Enhancing Cloud Security with Machine Learning and AI Agents. *Sateesh kumar and Raghunath, Vedaprada and Jyothi, Vinaya Kumar and Kudithipudi, Karthik, AI-Driven Cybersecurity: Enhancing Cloud Security with Machine Learning and AI Agents (February 07, 2022)*.

[41]. Hellerstein, J. M., Haas, P. J., & Wang, H. J. (1997). Online aggregation. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, 171–182.

[42]. Garapati, R. S. (2022). Web-Centric Cloud Framework for Real-Time Monitoring and Risk Prediction in Clinical Trials Using Machine Learning. Current Research in Public Health, 2, 1346.

[43]. Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. Proceedings of the 2008 IEEE International Conference on Data Mining, 263–272.

[44]. Amistapuram, K. (2022). Fraud Detection and Risk Modeling in Insurance: Early Adoption of Machine

Learning in Claims Processing. *Available at SSRN 5741982*.
[45]. Inmon, W. H. (2005). Building the data warehouse (4th ed.). Wiley.
[46]. Meda, R. (2022). Integrating Edge AI in Smart Factories: A Case Study from the Paint Manufacturing Industry. International Journal of Science and Research (IJSR), 1473-1489.
[47]. Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86–94.
[48]. Segireddy, A. R. (2020). Cloud Migration Strategies for High-Volume Financial Messaging Systems.
[49]. Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148–152.
[50]. 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.
[51]. Kleppmann, M. (2017). Designing data-intensive applications. O'Reilly Media.
[52]. Nagabhyru, K. C. (2022). Bridging Traditional ETL Pipelines with AI Enhanced Data Workflows: Foundations of Intelligent Automation in Data Engineering. Available at SSRN 5505199.
[53]. Lahiri, M., & Venkatasubramanian, S. (2013). Robust record linkage. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 101–112.
[54]. Avinash Reddy Aitha. (2022). Deep Neural Networks for Property Risk Prediction Leveraging Aerial and Satellite Imaging. International Journal of Communication Networks and Information Security (IJCNIS), 14(3), 1308–1318. Retrieved from https://www.ijcnis.org/index.php/ijcnis/article/view/8609
[55]. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets (2nd ed.). Cambridge University Press.
[56]. Rongali, S. K. (2022). AI-Driven Automation in Healthcare Claims and EHR Processing Using MuleSoft and Machine Learning Pipelines. *Available at SSRN 5763022*.
[57]. Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76–80.
[58]. Meda, R. (2021). Digital Infrastructure for Predictive Inventory Management in Retail Using Machine Learning. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI, 10.
[59]. Lin, J., Kolcz, A., & Szymanski, B. K. (2012). Large-scale machine learning at Twitter. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 793–804.
[60]. Sheelam, G. K. Power-Efficient Semiconductors for AI at the Edge: Enabling Scalable Intelligence in Wireless Systems. International Journal of Innovative Research in Electrical, Elec-tronics, Instrumentation and Control Engineering (IJIREEICE), DOI, 10.
[61]. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.
[62]. 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.
[63]. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations, 1–12.
[64]. Ramesh Inala. (2022). Cross-Domain MDM Integration Using AI-Driven Data Governance: A Case Study In Financial Technology Architecture. Migration Letters, 19(2), 280–304. Retrieved from https://migrationletters.com/index.php/ml/article/view/11982
[65]. Montoya, D. Y., Neto, A. M., & da Silva, A. S. (2016). A survey of entity resolution in big data. Journal of Big Data, 3(1), 1–22.
[66]. Aitha, A. R. (2021). Optimizing Data Warehousing for Large Scale Policy Management Using Advanced ETL Frameworks.
[67]. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 1–7.
[68]. Varri, D. B. S. (2022). AI-Driven Risk Assessment and Compliance Automation in Multi-Cloud Environments. *Available at SSRN 5774924*.
[69]. Zaharia, M., Das, T., Li, H., Shenker, S., & Stoica, I. (2012). Discretized streams: Fault-tolerant streaming computation at scale. Proceedings of the 24th ACM Symposium on Operating Systems Principles, 423–438.
[70]. 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.
[71]. Zhai, C., & Massung, S. (2016). Text data management and analysis: A practical introduction to information retrieval and text mining. ACM & Morgan Claypool.
[72]. Goutham Kumar Sheelam, "Semiconductor Innovation for Edge AI: Enabling Ultra-Low Latency in Next-Gen Wireless Networks," International Journal of Advanced Research in Computer and Communication Engineering

(IJARCCE), DOI: 10.17148/IJARCCE.2022.111258
[73]. Abedjan, Z., Golab, L., & Naumann, F. (2016). Profiling relational data: A survey. The VLDB Journal, 24(4), 557–581.
[74]. Yandamuri, U. S. (2022). Big Data Pipelines for Cross-Domain Decision Support: A Cloud-Centric Approach. *International Journal of Scientific Research and Modern Technology*, *1*(12), 227–237. https://doi.org/10.38124/ijsrmt.v1i12.1111
[75]. Aggarwal, C. C. (2017). Outlier analysis (2nd ed.). Springer.
[76]. Gottimukkala, V. R. R. (2022). Licensing Innovation in the Financial Messaging Ecosystem: Business Models and Global Compliance Impact. International Journal of Scientific Research and Modern Technology, 1(12), 177-186.
[77]. Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2021). Fraud analytics using descriptive, predictive, and social network techniques: A guide to data science for fraud detection (2nd ed.). Wiley.
[78]. 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)*.
[79]. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org (Book manuscript).
[80]. Garapati, R. S. (2022). AI-Augmented Virtual Health Assistant: A Web-Based Solution for Personalized Medication Management and Patient Engagement. Available at SSRN 5639650.
[81]. Batarseh, F. A., & Yang, R. (2019). Federal data science: Transforming government and society. Academic Press.
[82]. Gottimukkala, V. R. R. (2020). Energy-Efficient Design Patterns for Large-Scale Banking Applications Deployed on AWS Cloud. power, 9(12).
[83]. Bhasin, H., & Bhatia, P. (2020). Clickstream data mining for web analytics and customer behavior modeling: A review. ACM Computing Surveys, 53(6), 1–34.
[84]. Rongali, S. K. (2021). Cloud-Native API-Led Integration Using MuleSoft and .NET for Scalable Healthcare Interoperability. *Available at SSRN 5814563*.
[85]. Böhm, M., Koleva, G., Leimeister, J. M., Riedl, C., & Krcmar, H. (2017). Towards a generic value network for cloud computing. Future Generation Computer Systems, 72, 286–297.
[86]. Goutham Kumar Sheelam. (2022). Reconfigurable Semiconductor Architectures For AI-Enhanced Wireless Communication Networks. Kurdish Studies, 10(2), 1027–1040. https://doi.org/10.53555/ks.v10i2.3867
[87]. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.
[88]. 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
[89]. Cao, L. (2020). Data science: A comprehensive overview. ACM Computing Surveys, 52(3), 1–42.
[90]. Uday Surendra Yandamuri. (2022). Cloud-Based Data Integration Architectures for Scalable Enterprise Analytics. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(3s), 472–483. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/8005
[91]. Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2019). Big data challenge: A data management perspective. Frontiers of Computer Science, 13(1), 1–17.
[92]. Dwaraka Nath Kummari,. (2022). Machine Learning Approaches to Real-Time Quality Control in Automotive Assembly Lines. Mathematical Statistician and Engineering Applications, 71(4), 16801–16820. Retrieved from https://philstat.org/index.php/MSEA/article/view/2972
[93]. Chen, T., Qin, Z., & Wang, J. (2020). A survey on deep learning for customer churn prediction. IEEE Access, 8, 172–187.
[94]. Inala, R. (2022). Engineering Data Products for Investment Analytics: The Role of Product Master Data and Scalable Big Data Solutions. International Journal of Scientific Research and Modern Technology, 155-171.
[95]. Cretu, C., et al. (2020). The modern data warehouse ecosystem: Architectures and best practices. IEEE Software, 37(6), 78–85.
[96]. Varri, D. B. S. (2021). Cloud-Native Security Architecture for Hybrid Healthcare Infrastructure. *Available at SSRN 5785982*.
[97]. Damji, J., Wenig, B., Das, T., & Lee, D. (2020). Learning Spark: Lightning-fast data analytics (2nd ed.). O'Reilly Media.
[98]. Dwaraka Nath Kummari. (2022). Fiscal Policy Simulation Using AI And Big Data: Improving Government Financial Planning. Kurdish Studies, 10(2), 934–945. https://doi.org/10.53555/ks.v10i2.3855
[99]. Dehghani, M. (2019). Data mesh: Delivering data-driven value at scale. O'Reilly Media.
[100]. Aitha, A. R. (2022). Cloud Native ETL Pipelines for Real Time Claims Processing in Large Scale Insurers. Available at SSRN 5532601.
[101]. Demchenko, Y., Grosso, P., de Laat, C., & Membrey, P. (2017). Addressing big data issues in scientific data infrastructure. Journal of Grid Computing, 15(1), 1–9.
[102]. Avinash Reddy Segireddy. (2022). Terraform and Ansible in Building Resilient Cloud-Native

Payment Architectures. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 444–455. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/7905
[103]. Doan, A., Halevy, A., & Ives, Z. (2012). Principles of data integration. Morgan Kaufmann.
[104]. 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
[105]. Dutta, S., & Bose, I. (2015). Managing a big data project: The case of Ramco Cements Limited. International Journal of Production Economics, 165, 293–306.
[106]. 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.
[107]. Eckerson, W. W. (2020). The future of data management: From data warehouses to data fabrics. TDWI Research.
[108]. Gottimukkala, V. R. R. (2021). Digital Signal Processing Challenges in Financial Messaging Systems: Case Studies in High-Volume SWIFT Flows.
[109]. Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897–904.
[110]. Dwaraka Nath Kummari. (2022). AI-Driven Audit Frameworks For Enhancing Compliance In Modern Manufacturing Systems. Migration Letters, 19(S8), 2150–2177. Retrieved from https://migrationletters.com/index.php/ml/article/view/11912
[111]. Fan, W., & Bifet, A. (2013). Mining big data: Current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, 14(2), 1–5.