SCALABLE ARCHITECTURE MODEL FOR B2B MARKETING DATA UNIFICATION: REAL-TIME SYNCHRONIZATION AND DATA QUALITY ASSURANCE ACROSS MULTIPLE THIRD-PARTY SOURCES

Victor Almeida Barros[1]

Abstract: The increasing digitalization of the Business-to-Business (B2B) market has driven the need for Marketing Analytics platforms capable of processing and analyzing data in real-time. This article presents a systematic review of recent literature (2020-2025) to address the technical challenges inherent in real-time data synchronization from multiple third-party sources (e.g., CRM, advertising platforms, product analytics). The research focuses on the areas of Data Engineering, Systems Integration, and Scalability. The main challenges identified include the heterogeneity of data schemas, the high ingestion rate, and ensuring data quality (cleansing, merging, and transformation) to build a "Single Version of the Truth" for the B2B customer. A scalable reference architecture model is proposed, highlighting the role of an asynchronous worker system (such as BullMQ) for task orchestration and the use of a high-performance columnar database, such as ClickHouse, for real-time analytical storage and processing. This model aims to provide a robust foundation for unifying marketing, sales, and product data, which is essential for predictive decision-making and personalization at scale.

Keywords: Real-Time Data Synchronization, B2B Marketing Analytics, Data Engineering, ClickHouse, Data Quality, Worker Architecture.

1 M.S. in Information Systems Management and IT Project Management, B.S. in Mechanical Engineering

Introduction

The digital transformation in the B2B sector has elevated data analysis from a support function to a strategic imperative (SHASHI et al., 2025). The modern Marketing Technology (MarTech) ecosystem is characterized by the proliferation of specialized tools, resulting in a fragmentation of critical customer and campaign data. For B2B organizations to achieve personalization at scale and real-time decision-making, the unification of this data, originating from over one hundred third-party integrations, such as CRM platforms, marketing automation, and advertising systems, becomes fundamental (PRATAMA, 2025). The central technical challenge lies in building a Data Engineering system capable of synchronizing this massive and heterogeneous volume of information with the speed and reliability required by real-time analytics (LALAOUI, 2025). The transition from batch processing to stream processing imposes rigorous requirements for low latency and high throughput, especially in scenarios where latency must be less than one second for most use cases (SCHULZE et al., 2024). This article aims to propose a scalable reference architecture model, based on a systematic literature review between 2020 and 2025, that addresses real-time data synchronization and data quality assurance in B2B Marketing Analytics platforms. The focus is on the integration of distributed worker systems and the use of state-of-the-art columnar databases.

Systematic Literature Review (2020-2025)

The Evolution of B2B Marketing Analytics and the Need for Real-Time

Recent literature demonstrates that B2B Marketing Analytics has evolved into a predictive and real-time data-driven model (MISHRA; KAUTISH, 2025). The ability to make immediate decisions, such as optimizing bids in advertising campaigns or personalizing customer journeys, is a competitive differentiator (SHASHI et al., 2025). The review by Shashi et al. (2025) highlights that digitalization requires B2B companies to move towards a real-time decision model to meet

customer needs (SHASHI et al., 2025). The complexity of the B2B decision process, with multiple **stakeholders**, makes the "Single Version of the Truth" for the customer an invaluable asset, which can only be achieved through the immediate unification and analysis of marketing, sales, and product data (PRATAMA, 2025).

Challenges of Third-Party Data Integration

The integration of third-party data is the main bottleneck in building B2B analytical platforms. The heterogeneity of APIs, formats, and data schemas across the more than 100 data sources (e.g., Salesforce, Google Ads, HubSpot) requires continuous Data Engineering effort (PRATAMA, 2025). In addition to structural diversity, data volatility and inconsistency represent a significant challenge for quality (SHAABAN, 2025). Pratama's (2025) review points out that platform and data integration is one of the main challenges in digital transformation, affecting the ability to deliver personalized experiences in real-time (PRATAMA, 2025). Ensuring that data from different sources correctly merges to form a cohesive customer profile is a **Master Data Management** (MDM) problem that needs to be solved at scale and in real-time (SHAABAN, 2025).

Columnar Databases for Real-Time OLAP

To support the high ingestion rate and low-latency analytical queries, columnar Online Analytical Processing (OLAP) databases have become the preferred solution. ClickHouse, in particular, is recognized for its ability to perform high-performance analysis on petabyte-scale datasets with high ingestion rates (SCHULZE et al., 2024). The article by Schulze et al. (2024) details that ClickHouse was designed to handle five key challenges of modern analytical data management, including **Huge data sets with high ingestion rates** and **Many simultaneous queries with an expectation of low latencies** (SCHULZE et al., 2024). Its architecture, based on **Log-Structured Merge** (LSM) trees

and vectorized processing, allows for the continuous ingestion of new data without degrading the performance of parallel reporting queries (SCHULZE et al., 2024).

Technical Challenges in Real-Time Data Synchronization

Scalability and Resilience of the Worker System

The synchronization of data from over 100 third-party sources cannot be performed by a monolithic process. A distributed and asynchronous worker architecture is necessary to manage complexity and scalability (RADLBAUER et al., 2025). The main challenge is ensuring the system's resilience against network failures, API rate limits, and third-party service unavailability. A robust queue system (such as BullMQ) is essential to decouple ingestion from processing, allowing for the re-processing of failed tasks (**at-least-once delivery**) and the prioritization of critical events (HUANG, 2024). The architecture must be able to scale horizontally to handle ingestion peaks, which are common in Marketing Analytics platforms (RADLBAUER et al., 2025).

Data Quality Assurance (Cleansing, Merging, and Transformation)

Data quality is the pillar of B2B analysis. Real-time synchronization requires that the Extract, Load, and Transform (ELT) steps be executed with rigor and speed.

- Cleansing: Involves format validation, error correction, and field standardization (e.g., country names, date formats) immediately after ingestion. Literature indicates that AI-enhanced Master Data Management is crucial for real-time data governance and quality improvement (SHAABAN, 2025).
- Merging and Deduplication: This is the most critical process for data unification. It involves identifying records that represent the same entity (customer, account) across

different sources. Probabilistic and deterministic matching techniques, based on unique identifiers (e.g., internal customer ID, email) and data enrichment, must be applied to create a unified and deduplicated profile in real-time (SHAABAN, 2025).

- Transformation: Raw data must be transformed into an analytical schema optimized for OLAP queries (e.g., star schema or data vault model). This transformation needs to be continuous and efficient, leveraging the stream processing capabilities of the data pipeline (LALAOUI, 2025).

Architectural and Technological Solutions

Distributed Worker Architecture

The proposed architectural model is based on microservices and a message queue system for orchestration. The data flow begins with event capture via webhooks or Change Data Capture (CDC) from third-party APIs (HUANG, 2024). Each event is sent to a message queue (e.g., managed by BullMQ or Kafka), which acts as a buffer and ensures message persistence. The workers are asynchronous consumers of these queues, responsible for specific tasks: a raw ingestion worker, a cleansing and merging worker, and a final transformation worker. This separation of responsibilities ensures scalability and resilience. If a worker fails (e.g., due to an API rate limit), the message is automatically resent to the queue for re-processing, without impacting the ingestion flow from other sources (RADLBAUER et al., 2025).

The Role of ClickHouse in the Analytical Layer

ClickHouse is positioned as the main analytical Data Warehouse, receiving data from the *worker pipeline*. Its columnar architecture is ideal for storing marketing and sales event data, which is typically wide and sparse. To optimize real-time performance, ClickHouse uses the MergeTree

Engines family. These engines allow data to be inserted in small batches (minimizing ingestion latency) and then merged and transformed asynchronously in the background (SCHULZE et al., 2024). Materialized Views in ClickHouse are a powerful tool for continuously pre-aggregating data, ensuring that dashboard queries (which require low latency) are executed on already summarized data, while raw data remains available for ad-hoc analysis (SCHULZE et al., 2024).

Data Quality Strategies in the Pipeline

Data quality assurance must be integrated into the worker pipeline. After raw ingestion, a validation worker applies business rules and enrichment. For example, before merging, the worker can use third-party services to enrich the record with company data (e.g., sector, size) or validate the email format. Data merging is performed by a dedicated worker that consults an internal Master Data Management (MDM) or an identity index. Upon receiving a new record, the worker attempts to match it with an existing profile. If the matching is successful, the new data is merged; otherwise, a new profile is created. This process must be fast enough for the unified profile to be available for real-time analysis (SHAABAN, 2025).

Proposed Reference Architecture Model

The reference architecture model for B2B Marketing data unification is composed of four main layers:

1. Data Sources Layer: Includes all third-party integrations (Salesforce, Google Ads, etc.) and internal sources (product database).
2. Ingestion and Orchestration Layer: This is the heart of the synchronization system. It uses distributed workers and a message queue system (e.g., BullMQ) to manage event capture via

webhooks or CDC, ensuring the resilience and scalability of ingestion.

3. Processing and Quality Layer: In this layer, dedicated workers execute the tasks of cleansing, standardization, enrichment, and, crucially, the merging and deduplication of records to create the unified B2B customer profile.

4. Analytical Storage and Consumption Layer: ClickHouse acts as the high-performance columnar Data Warehouse, optimized for real-time OLAP queries. This layer feeds dashboards, visualization tools, and real-time decision APIs.

This model, by separating the concerns of ingestion, processing, and analysis, ensures that a failure in a third-party integration does not paralyze the analytical system, and that the complex data quality logic is executed asynchronously and scalably.

Conclusion

Real-time data synchronization in B2B Marketing Analytics platforms is a Data Engineering challenge that requires a robust and scalable architecture. The systematic literature review confirms that the combination of an asynchronous worker system for task orchestration and a high-performance columnar database, such as ClickHouse, is the most promising architectural solution to handle the high ingestion rate and low latency required by the market (SCHULZE et al., 2024; RADLBAUER et al., 2025). The main contribution of this article is the proposal of a reference architecture model that integrates data quality management (cleansing, merging, and transformation) directly into the asynchronous processing pipeline. This approach ensures that the "Single Version of the Truth" for the B2B customer is built and maintained in real-time, providing the foundation for predictive marketing intelligence. For future work, research is suggested on the application of Machine Learning algorithms for the automatic detection of anomalies and data quality deviations in real-time within the worker pipeline. Furthermore, cost optimization and resource management in cloud environments for

petabyte-scale ClickHouse clusters represent an area of great practical research relevance.

References

HUANG, Z. Near-real-time data pipeline using change data capture approach. Master's thesis, Theseus, 2024.

LALAOUI, I. L. The Evolution and Challenges of Real-Time Big Data. Big Data and Cognitive Computing, v. 10, n. 1, p. 11, 2025.

MISHRA, A.; KAUTISH, P. Strategic Marketing Analytics: A Systematic Review. In: Strategy Analytics for Business Resilience Theories and Practices. Springer, p. 207-226, 2025.

PRATAMA, P. Y. A. Digital Marketing Transformation in Business: A Systematic Literature Review on Trends, Challenges, and Strategic Impacts. International Journal of Smart Business and Technology, v. 3, n. 1, p. 4-15, 2025.

RADLBAUER, E.; MOSER, T.; WAGNER, M. Designing a System Architecture for Dynamic Data Collection as a Foundation for Knowledge Modeling in Industry. Applied Sciences, v. 15, n. 9, p. 5081, 2025.

RAYHAN, M. Evaluating real-time monitoring and data-storage systems for active magnetic bearings based high-speed machines.. 2024. (Document type and details are incomplete in the original. Preserved what was extracted).

SCHULZE, R. Clickhouse-lightning fast analytics for everyone. VLDB Endowment, 2025. (Details are incomplete in the original. Preserved what was extracted).

SCHULZE, R. Optimizing Data Flows at Sc. 2025. (Details are incomplete in the original. Preserved what was extracted).

SCHULZE, R. et al. ClickHouse - Lightning Fast Analytics for Everyone. PVLDB, v. 17, n. 12, p. 3731-3744, 2024. DOI: 10.14778/3685800.3685802.

SHAABAN, M. Towards Scalable and Context-Aware Digital Twins for HRC. 2025. (The full title and publication details are incomplete in the original. Preserved what was extracted)

SHASHI, M. et al. Transforming business-to-business marketing from tradition to digitalization: a taxonomic review of current trends, methodologies and future paths. Journal of Business & Industrial Marketing, v. 40, n. 6, p. 1335-1349, 2025.