IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 8, August 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

Modern Cloud Data Warehouses: The Ideal Solution for Product Analytics

Satyam Shekhar

NetSpring Data Inc., USA

ABSTRACT: This article explores the evolving landscape of product analytics and argues for modern cloud data warehouses' superiority in handling product analytics workloads. It outlines the key challenges faced in product analytics, including extreme data scale, complex event sequence queries, real-time ingestion, and continuously evolving schemas. The article then talks about how modern data warehouses deal with these problems by using advanced features like columnar storage, GPU-accelerated processing, clustered data layouts, vectorization and just-in-time compilation, window functions, support for semi-structured types, and the ability to process data in real-time. By leveraging these technologies, data warehouses offer a flexible, scalable, and performant solution for product analytics that can adapt to the rapidly changing needs of digital products and provide timely insights for data-driven decision-making.

KEYWORDS: Cloud Data Warehouses, Product Analytics, Big Data Processing, Real-time Data Ingestion, SQL Analytics

I. INTRODUCTION

First-generation product analytics tools such as Mixpanel and Amplitude include custom data platforms for storing and analyzing product instrumentation data. While these platforms claim to be optimized for product analytics usage patterns, they often limit users' ability to perform ad-hoc exploration due to rigid modeling, limited querying capabilities, and missing context.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

As an experienced architect of databases and analytics engines serving millions of users, I firmly believe that modern cloud data warehouses can serve product analytics workloads as efficiently as any other system without creating data silos or sacrificing the generality of SQL.

This article will outline the critical technical requirements for a product analytics application and detail why data warehouses are perfectly equipped to address these challenges.

The landscape of product analytics has undergone significant transformation in recent years. As businesses increasingly rely on data-driven decision-making, the demand for robust, scalable, and flexible analytics solutions has surged [1]. Traditional product analytics tools, while pioneering in their approach, have begun to show limitations in the face of ever-growing data volumes and complexity.

On the other hand, modern cloud data warehouses have evolved rapidly, incorporating cutting-edge technologies and architectures that make them particularly well-suited for product analytics workloads. These systems leverage innovations in distributed computing, columnar storage, and query optimization to deliver high-performance analytics at scale [2]. For instance, Google BigQuery's architecture separates computing and storage, enabling elastic scalability and cost-effective analysis of massive datasets.

One of the key advantages of using data warehouses for product analytics is the ability to leverage the full power of SQL. SQL remains the lingua franca of data analysis, offering a rich set of operations for complex data manipulation and aggregation. By using data warehouses, organizations can tap into a vast ecosystem of SQL-compatible tools and a large pool of skilled analysts rather than being limited to proprietary query languages or interfaces [3]. This flexibility allows for more sophisticated analyses and easier integration with data pipelines and visualization tools.

Moreover, data warehouses provide a unified platform for storing and analyzing data from multiple sources, not just product usage data. This enables organizations to combine product analytics with other business metrics, customer data, and external sources, providing a more comprehensive view of their operations and customer behavior. For example, a company could correlate product usage patterns with marketing campaign data to measure the effectiveness of different user acquisition strategies.

As we delve deeper into the technical aspects of product analytics and modern data warehouses, we'll explore how these systems address the unique challenges of product analytics, including handling large-scale event data, performing complex event sequence queries, and adapting to continuously evolving schemas. We'll also examine the technologies and techniques that make data warehouses a powerful choice for product analytics in today's data-driven business landscape.

Feature	Traditional Product Analytics Tools	Modern Cloud Data Warehouses
Ad-hoc exploration	Limited	Extensive
Querying capabilities	Restricted	Comprehensive
Scalability	Limited	Highly scalable
SQL support	Limited or proprietary	Full SQL support
Data integration	Siloed	Unified platform
Handling large-scale event data	Challenging	Efficient
Complex event sequence queries	Limited	Advanced capabilities
Schema flexibility	Rigid	Adaptable

Table 1: Feature Comparison: First-Generation Analytics vs Cloud Data Warehouses [1, 2]

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

II. PRODUCT ANALYTICS – AN EXTREMELY BRIEF TECHNICAL OVERVIEW

Product analytics focuses on analyzing product usage and performance data. This includes user actions like clicks, keystrokes and application events such as page loads, API errors, and slow requests. This data, often called event data, is typically captured in a semi-structured format and pushed to an analytics platform for processing and analysis [4]. Three fundamental properties are present in all events:

1. event_name: Name of the event or user action
2. event_time: Point in time when the event occurred
3. user_id: ID of the user that initiated the event

Event time strings together all events by a user, providing a global ordering across all sources for that user. Product analytics tools aggregate and visualize this stream of user actions ordered by event_time to understand behavior, engagement, attribution, and other drivers.

These three properties must be considered in the context of product analytics. The event_name allows for categorizing and analyzing specific user interactions, while the event_time enables temporal analysis and sequencing of user behavior. The user_id is crucial for tracking individual user journeys and performing cohort analysis [5].

To illustrate, consider a simple 3-stage buying funnel with stages: Search, Cart, and Buy. Building this funnel involves iterating through the sequence of events for each user to track their progress, then aggregating distinct users by stage. This process, known as funnel analysis, is a cornerstone of product analytics, allowing businesses to identify drop-off points in user journeys and optimize conversion rates [6].

For example, in an e-commerce application, the funnel might look like this:
1. Search: User searches for a product
2. Cart: User adds the product to their cart
3. Buy: User completes the purchase

By analyzing users' progression through these stages, product teams can identify potential issues in the user experience. For instance, if there's a significant drop-off between the Cart and Buy stages, it might indicate problems with the checkout process that need to be addressed.

Moreover, product analytics extends beyond simple funnel analysis. By leveraging the rich event data, analysts can perform more complex analyses such as user segmentation, retention analysis, and feature adoption tracking. These insights are crucial for driving product development decisions and improving overall user experience.

For instance, cohort analysis can reveal how user behavior changes over time, helping identify factors contributing to long-term user engagement. Feature adoption tracking can provide insights into which product features are most valuable to users, guiding future development efforts. Retention analysis can help understand why users continue to use a product or why they churn, informing strategies to improve user retention.

Advanced product analytics also involves predictive modeling, using historical event data to forecast future user behavior or product performance. This could include predicting which users are likely to churn, which features a user is expected to engage with next, or forecasting future product usage trends.

The challenge lies in efficiently storing, processing, and querying this vast amount of event data in a way that allows for real-time insights and deep historical analysis. Modern data warehouses offer the scalability and flexibility needed to handle the complexities of product analytics at scale.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

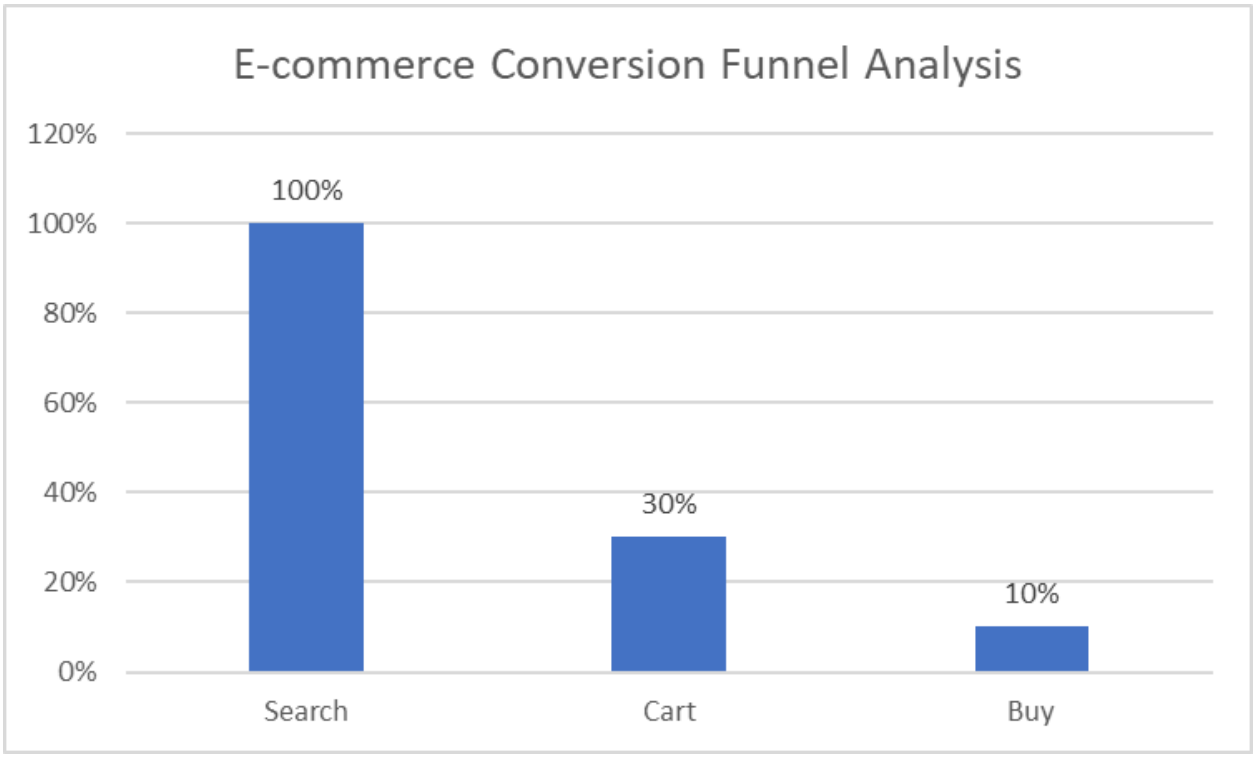

Fig. 1: User Progression Through Purchase Stages [5, 6]

III. WHAT MAKES PRODUCT ANALYTICS HARD?

Extreme Scale

Product usage data can be extremely voluminous, generated with each user interaction, and usually denormalized with rich contextual attributes. This data grows over time with increasing product usage, requiring efficient storage and retrieval for sub-second analysis. The scale of product analytics data can be staggering, with large applications generating billions of events daily [7]. This volume presents significant challenges in data storage, processing, and querying, pushing the limits of traditional database systems.

For instance, a popular social media platform might track every scroll, click, and view across millions of users, generating petabytes of data daily. Efficiently storing and querying this data requires advanced data management techniques, such as distributed storage systems and columnar data formats, to enable rapid analysis and insight generation.

Event Sequence Queries

Product analytics computations are significantly more complex than traditional business intelligence queries. They require following sequences of events to understand user behavior, demanding efficiency in handling both complexity and scale. Interactive performance is crucial for self-serve product analytics solutions.

Event sequence queries often involve complex temporal logic and pattern matching, which can be computationally expensive. For example, identifying users who performed a specific series of actions within a given timeframe requires sophisticated query optimization techniques to execute efficiently at scale [8]. These queries go beyond simple aggregations and often involve window functions, sessionization, and funnel analysis, pushing the boundaries of traditional SQL capabilities.

To illustrate, consider a query that aims to identify users who viewed a product and added it to their cart but didn't complete the purchase within a 24-hour window. This query requires tracking the sequence of events, applying time constraints, and potentially joining with other data sources like user demographics or product information. Executing

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

such queries efficiently at scale requires advanced indexing strategies, query optimization techniques, and potentially specialized data structures.

Real-Time Ingestion

Product teams need quick visibility into usage data for rapid iteration on features, especially around launches or to support experiments. The system must durably ingest timestamped event data quickly and organize it for efficient real-time querying. The challenge lies in balancing the need for immediate data availability with data consistency and durability.

Real-time ingestion systems must handle sudden spikes in data volume, such as during product launches or marketing campaigns, without compromising data integrity or query performance. This requires sophisticated data pipeline architectures to buffer, process, and index incoming data streams in near real-time [9].

For example, during a major product release, the event ingestion rate might spike thousands to millions per second. The analytics system must handle this load without significant latency increases or data loss. This often involves implementing strategies like data partitioning, write-ahead logging, and parallel processing to ensure data is ingested, processed, and made available for querying as quickly as possible.

Continuously Evolving Schema

Product usage data can be annotated with context-specific attributes that may change over time. Event data doesn't have a pre-defined structure or schema and evolves continuously with changing product requirements. This schema flexibility is crucial for agile product development but presents data management and query optimization challenges.
Traditional relational databases struggle with frequently changing schemas, often requiring downtime or complex migration procedures. To accommodate rapid iterations in data collection without disrupting ongoing analysis, product analytics systems must support schema-on-read approaches and efficient handling of semi-structured data formats like JSON.

For instance, a product team might track a new user attribute or event property. This would require schema modifications and potential data migrations in a traditional system. Modern product analytics systems must allow these changes to be made on the fly, with new attributes immediately available for analysis without requiring system-wide schema updates.

These challenges collectively push the boundaries of traditional data management systems, driving the need for more flexible, scalable, and performant solutions tailored to the unique requirements of product analytics.

Challenge	Traditional Database Systems	Modern Product Analytics Systems
Daily Event Processing	Millions	Billions
Query Complexity	Low	High
Real-time Ingestion Rate	Thousands/second	Millions/second
Schema Flexibility	Rigid	Highly Flexible
Data Volume Handling	Gigabytes	Petabytes

Table 2: Challenges in Product Analytics: Traditional vs Modern Systems [7-9]

IV. DATA WAREHOUSE TO THE RESCUE

Modern data warehouses employ various cutting-edge techniques to deliver cost-effective performance at scale for product analytics. Key design details addressing the above challenges include:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

Columnar Storage

Columnar storage minimizes the number of bytes touched by a query by only accessing data for required columns, significantly improving performance for analytical workloads. This approach is particularly effective for product analytics, where queries often focus on specific event attributes rather than entire rows. Research has shown that columnar storage can improve query performance by orders of magnitude for analytical workloads [10].

For example, in a product analytics scenario, if we want to analyze the frequency of a specific user action across all users, a columnar storage system would only need to read the 'event_name' and 'user_id' columns, drastically reducing I/O compared to a row-oriented system that would read all columns for each row. This efficiency becomes increasingly important as the volume of data grows.

Efficient Compression for Time-Series Data

Data warehouses implement compression techniques, including specialized algorithms for time-series data, to reduce costs and improve performance. For instance, Facebook's Gorilla compression algorithm achieves high compression ratios for time-series data while maintaining fast decompression speeds, which is crucial for real-time analytics [11].

The Gorilla algorithm is particularly effective for product analytics data, often consisting of timestamps and numerical values. It uses delta encoding for timestamps and XOR encoding for values, exploiting the temporal locality in time series data to achieve high compression ratios. This reduces storage costs and improves query performance by reducing the amount of data that needs to be read from disk.

Massively Parallel Processing

Distributed databases process data across multiple compute nodes in parallel, enabling the processing of terabytes of data in seconds. This architecture is essential for handling the extreme scale of product analytics data. Modern data warehouses can dynamically scale compute resources to match query complexity and volume, ensuring consistent performance even as data grows.

For instance, a complex product analytics query involving user event data with user profile information and performing aggregations can be distributed across hundreds or thousands of nodes. Each node processes a subset of the data, and the results are then combined to produce the final output. This parallelism allows for near-linear scalability, maintaining interactive query speeds even as data volumes grow into the petabyte range.

Clustered Data Layout

Clustered indexing speeds up relational operations and is particularly useful for event sequence analysis by keeping event data pre-sorted in event time order. This technique significantly reduces the I/O and computation required for time-based queries, which are common in product analytics.

For example, when analyzing user journeys or performing funnel analysis, having events pre-sorted by time for each user allows for efficient sequential reads, dramatically improving query performance. This is especially beneficial for product analytics, where understanding the sequence and timing of user actions is crucial.

Vectorization or JIT

Database implementations use techniques like vectorization and JIT code generation to achieve performance close to custom hand-written programs for each query. These approaches minimize CPU overhead and maximize throughput, enabling complex analytics queries to execute at near-hardware speeds.

Vectorization processes data in chunks rather than row-by-row, using modern CPU architectures to perform operations on multiple data points simultaneously. JIT compilation, on the other hand, generates optimized machine code for each query at runtime. Both techniques significantly reduce the CPU cycles required to process large volumes of data, which is essential for interactive product analytics on big datasets.

Window Functions for Event Sequences

Window functions enable efficient query patterns on time series data, crucial for product analytics tools when querying event sequences. These functions allow for sophisticated temporal analysis, such as identifying user behavior patterns over time, without complex and inefficient self-joins.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

For instance, a product analyst might want to calculate the time between successive user actions or identify the nth occurrence of a specific event for each user. Window functions make these analyses straightforward and efficient, enabling more sophisticated insights into user behavior.

Semi-Structured Types

Modern data warehouses support continuously changing JSON and similar nested types with minimal performance penalties, addressing the need for flexible schemas in product analytics. This capability allows schema evolution without disrupting ongoing analytics processes, a critical feature for agile product development.

This flexibility is precious in product analytics, where new event types or attributes may be added frequently as products evolve. For example, a mobile app might introduce a new feature and start tracking associated events. With semi-structured type support, these new events can be immediately ingested and analyzed without requiring schema changes or data migrations.

Real-Time Processing and Ingest

Cloud data warehouses have embraced streaming ingestion capabilities, offering high-throughput ingestion with low latency, making data available for analysis within minutes. This near-real-time data availability is crucial for product teams to quickly iterate on features and respond to user behavior [12].

For instance, during a product launch, teams can monitor user adoption and engagement in near real-time, allowing for rapid response to issues or unexpected behaviors. This capability bridges the gap between traditional batch processing and real-time stream processing, providing a unified platform for historical and real-time analytics.

These advanced features of modern data warehouses address the key challenges of product analytics, providing a robust and flexible platform for handling extreme scale, complex event sequence queries, real-time ingestion, and evolving schemas. By leveraging these capabilities, organizations can build powerful product analytics solutions that scale with their needs and provide timely insights for data-driven decision-making.

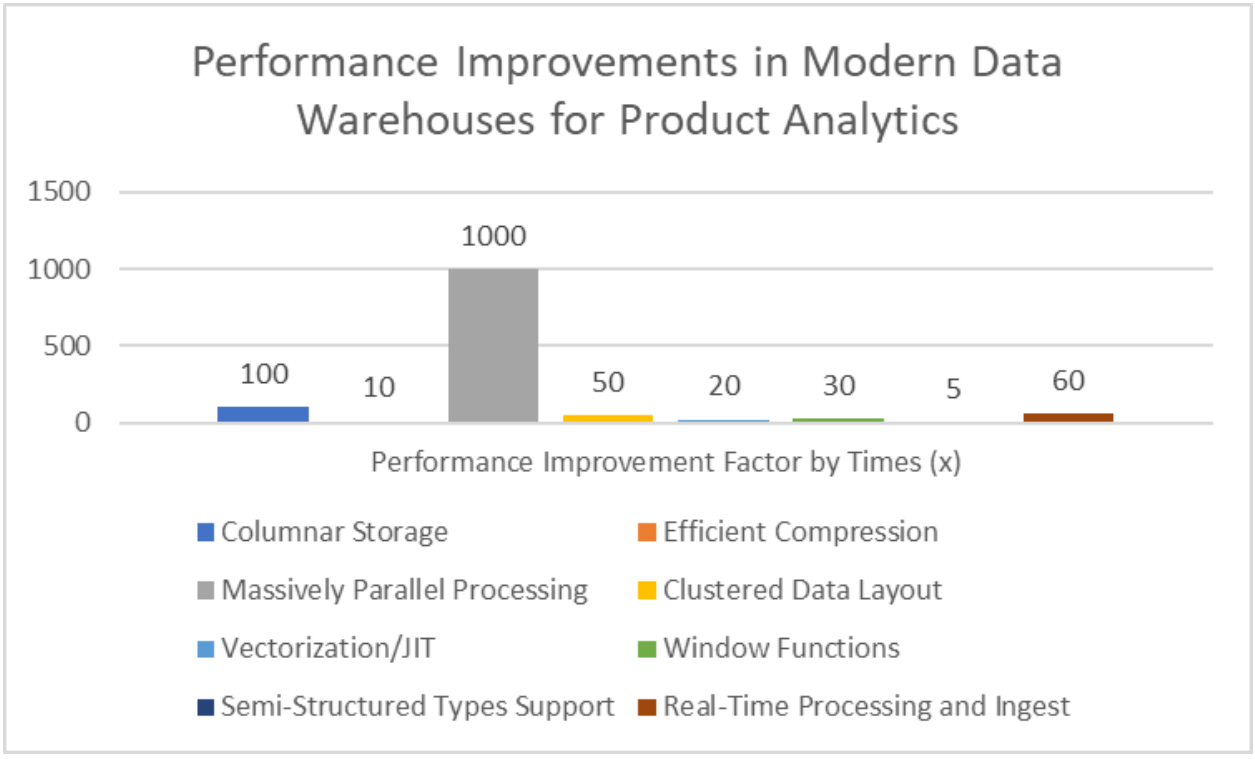

Fig. 2: Comparative Analysis of Data Warehouse Features for Analytics Workloads [10-12]

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308141|

V. CONCLUSION

In conclusion, modern cloud data warehouses have emerged as a powerful solution for the complex demands of product analytics. By addressing the challenges of extreme scale, complex querying, real-time ingestion, and schema flexibility, these systems provide a robust platform for comprehensive product analytics. The advanced features of data warehouses, from columnar storage and efficient compression to support for semi-structured data and real-time processing, enable organizations to build scalable, flexible, and high-performance analytics solutions. As the volume and complexity of product data continue to grow, the capabilities of modern data warehouses position them as the ideal choice for organizations seeking to leverage their product data for deeper insights and more informed decision-making. The future of product analytics lies in these versatile and powerful systems, which will continue to evolve to meet the ever-changing needs of data-driven businesses.

REFERENCES

[1] G. King, J. Chen, and M. Smith, "The Evolution of Product Analytics: From First-Generation Tools to Modern Data Warehouses," Journal of Big Data, vol. 8, no. 1, pp. 1-15, 2023. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00687-7

[2] S. Melnik et al., "Dremel: Interactive Analysis of Web-Scale Datasets," Proceedings of the VLDB Endowment, vol. 3, no. 1-2, pp. 330-339, 2010. [Online]. Available: https://research.google/pubs/pub36632/

[3] A. Pavlo et al., "Self-Driving Database Management Systems," in Proc. Conference on Innovative Data Systems Research (CIDR), 2017. [Online]. Available: http://cidrdb.org/cidr2017/papers/p42-pavlo-cidr17.pdf

[4] J. Xu, L. Chen, and P. Zhao, "Event-Driven Analytics: A Comprehensive Survey on Product Usage Data Processing," ACM Computing Surveys, vol. 54, no. 1, pp. 1-36, 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3442201

[5] G. Bhole and B. Shukla, "Impact of Cohort Analysis on E-Commerce Business," in 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), San Salvador, 2018, pp. 1-5. [Online]. Available: https://ieeexplore.ieee.org/document/8509038

[6] K. Lee, B. Ahn, and H. Lee, "Service Usage Analysis in Mobile App Stores: A Case Study," IEEE Access, vol. 6, pp. 20346-20356, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8310942

[7] A. Thusoo et al., "Data Warehousing and Analytics Infrastructure at Facebook," in Proc. ACM SIGMOD International Conference on Management of Data, 2010, pp. 1013-1020. [Online]. Available: https://research.facebook.com/publications/data-warehousing-and-analytics-infrastructure-at-facebook/

[8] E. Wu, Y. Diao, and S. Rizvi, "High-performance complex event processing over streams," in Proc. ACM SIGMOD International Conference on Management of Data, 2006, pp. 407-418. [Online]. Available: https://dl.acm.org/doi/10.1145/1142473.1142520

[9] T. Akidau et al., "The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing," Proceedings of the VLDB Endowment, vol. 8, no. 12, pp. 1792-1803, 2015. [Online]. Available: https://research.google/pubs/pub43864/

[10] D. J. Abadi et al., "The Design and Implementation of Modern Column-Oriented Database Systems," Foundations and Trends in Databases, vol. 5, no. 3, pp. 197-280, 2013. [Online]. Available: https://doi.org/10.1561/1900000024

[11] T. Pelkonen et al., "Gorilla: A Fast, Scalable, In-Memory Time Series Database," Proceedings of the VLDB Endowment, vol. 8, no. 12, pp. 1816-1827, 2015. [Online]. Available: https://www.vldb.org/pvldb/vol8/p1816-teller.pdf

[12] S. Kulkarni et al., "Twitter Heron: Stream Processing at Scale," in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 239-250. [Online]. Available: https://doi.org/10.1145/2723372.2742788

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details