Effective Utilization Strategies for Big Data Logs Generated from Digital Content Consumption: Analysis and Optimization Strategies

[1]Haeri An

Abstract

The big data logs generated from digital content consumption are crucial resources for companies to understand user behavior, provide personalized services, and optimize business strategies. This paper presents effective methods for collecting, storing, preprocessing, analyzing, and utilizing digital content consumption logs. Through these methods, digital content companies can improve user experience, enhance operational efficiency, and strengthen their competitiveness.

This paper analyzes various case studies involving real-time monitoring and problem-solving using log data, personalized recommendation systems, and strategies to improve user experience. Specifically, it presents concrete methods for data-driven innovation through successful log data utilization strategies employed by companies such as Netflix, YouTube, Amazon, and Instagram.

Future research should include the development of an integrated system architecture for big data log processing, encompassing data collection, storage, processing, analysis, and visualization. Such an integrated system design will enable companies to utilize log data more efficiently and respond quickly to rapidly changing market environments.

The digital content industry will continue to evolve through the analysis and optimization of big data logs. By implementing the strategies presented in this paper, companies can achieve better user experiences and higher business performance.

Keywords: *Digital content, big data logs, user behavior analysis, personalized services, business optimization, personalized recommendations, big data*

I. Introduction

Digital content exists in various forms, including text, images, audio, and video, and its consumption has surged with the development of internet and mobile technologies. Particularly, after the proliferation of various portable devices, digital content consumption has experienced rapid growth. In the past, digital content was limited to television and desktop PCs. However, nowadays, it can be consumed on laptops, tablets, mobile devices, and even on moving buses. These changes have significantly enhanced the accessibility and convenience of digital content.

Digital content refers to all forms of content that exist in digital formats, including text, images, audio, video, games, and software. It encompasses all materials accessible via the internet, playing a crucial role in modern society due to its accessibility and ease of distribution. Digital content has revolutionized the ways in which information is created, shared, and consumed [1].

Examples of digital content include e-books, online videos, streaming music, blog posts, and social media updates. This content is easily accessible through the internet and consumed via digital devices [2]. Digital content can be provided in both free and paid formats and can generate revenue through various business models such as subscription, advertising-based, and premium content sales models [3].

The vast amount of log data generated from this consumption provides valuable insights for businesses. This log data is crucial for analyzing user behavior, providing personalized services, and optimizing business operations. For example, Netflix enhances user satisfaction by offering personalized

[1]*Shingu College, Dept. of IT Software, Professor (anhr@shingu.ac.kr)*

Received: Jul. 23, 2024, Revised: Aug. 14, 2024, Accepted: Aug. 14, 2024

content recommendations based on viewing history, YouTube improves user engagement by recommending related videos based on viewing and search history, and Amazon increases sales by offering personalized product recommendations based on purchase and search history.

Big data refers to large volumes of structured and unstructured data that can be analyzed to extract meaningful insights. Big data is typically characterized by three features: Volume, Velocity, and Variety, which enable the creation of new value [4]. In the digital environment, vast amounts of data are generated anytime and anywhere through wired or wireless networks. The proliferation of smartphones, along with other devices such as laptops, tablets, and cameras, has led to a rapid increase in data traffic, contributing to the generation of unstructured data in modern society [5]. According to IT solutions company EMC [6], the total digital data generated globally is expected to reach 4.4 zettabytes by 2020. One zettabyte equals 1024 exabytes, or one trillion gigabytes (GB). This means one exabyte can hold a million times the printed materials in the U.S. Library of Congress [7]. While such vast amounts of data are often referred to as big data, its definition is not limited to size alone.

The characteristics of big data are generally summarized by the three Vs: Volume, Velocity, and Variety [8]. Additionally, the global IT company Oracle includes Value as another defining characteristic of big data [9].

Logs refer to records of events or transactions occurring within systems or applications. These log data are utilized for various purposes, such as system operations, problem-solving, and security management, and serve as important resources for big data analysis. For instance, log data can be used to monitor system status, identify security threats, and resolve issues [10].

The big data logs of digital content are emerging as key resources for enhancing corporate competitiveness. Log data can be collected and analyzed in real-time, enabling an understanding of user behavior patterns and preferences. This data is essential for providing personalized user experiences and helping companies predict market trends and respond quickly. Additionally, log data analysis plays a crucial role in monitoring and optimizing system performance and detecting and responding to security threats [10].

This paper presents effective utilization strategies for big data logs generated from digital content consumption, discussing the benefits that can be derived. It explains the importance and potential value of digital content consumption logs, introduces key analytical methodologies, and presents optimization strategies. By doing so, it aims to provide concrete methods for digital content companies to enhance user experience and business competitiveness through the effective use of log data. Furthermore, from the perspective of big data, this paper focuses on the added value of digital content consumption logs.

Effective utilization of digital content consumption logs can be a crucial factor in enhancing user satisfaction and maximizing business performance. Therefore, this study aims to explore strategies for efficiently analyzing and optimizing big data logs generated from digital content consumption, supporting companies in achieving data-driven innovation.

II. Importance of Big Data Logs in Digital Content

Big data logs of digital content hold significant value in various aspects.

First, these log data enable in-depth analysis of user behavior. By understanding which content users watch, for how long, on which devices they primarily access it, and through which paths they discover the content, companies can personalize and optimize user experiences. For example, by presenting users with their preferred content through personalized recommendation systems, viewing time can be increased, and user satisfaction can be enhanced [11].

Second, big data logs play a crucial role in real-time monitoring and problem-solving. By continuously monitoring system performance through log data, companies can quickly detect and respond to emerging issues. This is essential for maintaining the stability and reliability of services [12].

Third, log data are vital resources for security management. By detecting and responding to abnormal access attempts or attack patterns in real time, the security of systems can be enhanced. This is particularly important in industries where financial services or personal data protection is critical [10].

Fourth, big data logs provide essential insights for business intelligence (BI) and marketing strategy development. By analyzing log data, companies can define user segments and launch targeted marketing campaigns for each segment, maximizing marketing efficiency. Additionally, it helps in quickly responding to market changes and discovering new business opportunities [13].

Thus, big data logs of digital content serve as valuable assets for companies in improving user experience, optimizing system performance, enhancing security, and strengthening business intelligence. This paper aims to present specific analysis and optimization strategies for effectively utilizing these log data.

III. Collection and Storage of Digital Content Consumption Logs

To effectively utilize the vast log data generated from digital content consumption, it is essential to properly collect and store the data. During the data collection stage, data is centrally collected from various sources and stored in repositories. Examples include web server logs, application logs, and network logs.

3.1 Log Collection Tools

Representative tools for log collection include Fluentd, Logstash, and Kafka. These tools efficiently process large-scale real-time log data.

- Fluentd: Fluentd is an open-source data collection tool that can collect and integrate log data from various sources. With over 500 plugins, it can be extended to various data sources and output destinations. Fluentd structures log data in JSON format, facilitating integration with various systems [14].
- Logstash: As part of the Elastic Stack, Logstash collects logs through various input plugins, filters, and transforms them, and sends them to storage systems like Elasticsearch. Logstash is very useful for building flexible data processing pipelines [15].
- Kafka: Kafka is a tool specialized in processing large-scale real-time data streams, capable of handling large log data with high throughput and low latency. Kafka allows real-time log data streams to be distributed across multiple systems for processing [16].

3.2 Data Storage

Representative storage systems for collected log data include Elasticsearch and Hadoop HDFS. These storage systems offer features to efficiently store and search large-scale data.

- Elasticsearch: Elasticsearch is a distributed search engine capable of real-time search and analysis of large-scale data. Especially when used with Kibana, it provides visualization and analysis features to effectively utilize log data. For example, companies like Netflix and LinkedIn use Elasticsearch to analyze log data in real-time [15].
- Hadoop HDFS: HDFS is a file system suitable for distributed storage and processing of large data sets [17]. HDFS is advantageous for long-term data storage and analysis, efficiently managing large-scale log data [18].

These storage systems have distinct characteristics and advantages, allowing them to be selected based on the nature and purpose of the log data. Elasticsearch excels in real-time analysis and search, while HDFS is suitable for long-term storage and processing of large-scale data.

3.3 Log Collection Tools on Cloud Platforms

Major cloud service providers like AWS, Azure, and GCP offer their own unique log collection and management tools.

- AWS CloudWatch Logs: AWS CloudWatch Logs is a tool for monitoring and analyzing logs from AWS services and applications. CloudWatch Logs collects log data in real-time and provides search and analysis features. This allows for monitoring system performance and debugging application issues [19]
- Azure Monitor: Azure Monitor is used to collect and analyze log data from Azure resources. It provides log analytics to query, visualize, and set alerts on log data. Azure Monitor collects logs from various data sources to provide an integrated view [20].

- GCP Stackdriver Logging: Now known as Google Cloud Logging, Stackdriver Logging is a tool for collecting and managing log data from GCP resources and applications. It allows for real-time log analysis and visualization through custom dashboards. Additionally, it enhances system monitoring through log-based metrics and alerting [21]

3.4 Comparison of Log Collection Tools

Various log collection tools have different strengths and features, allowing users to select based on their needs. The following is a comparison of major log collection tools:

Table 1. Comparison of Log Collection Tools

Tool	Features	Advantages	Disadvantages
Fluentd	Supports various plugins, JSON format data structuring, low system resource usage	Scalability, flexibility, high performance	Complex configuration
Logstash	Powerful data processing pipeline, supports various input/filter/output plugins	Flexible data processing, easy integration with Elastic Stack	High memory usage
Kafka	High throughput, low latency, distributed data streaming platform	Real-time data processing, scalability	Complex configuration and operation
CloudWatch	Easy integration with AWS resources, real-time monitoring and analysis	Seamless integration with AWS services, real-time analysis	Difficult to integrate with non-AWS resources
Azure Monitor	Easy integration with various Azure services, log query and visualization	Seamless integration with Azure services, powerful visualization	Limited integration with non-Azure environments
GCP Logging	Easy integration with GCP resources, supports real-time log analysis and custom dashboards	Seamless integration with GCP services, real-time log analysis	Difficult to integrate with non-GCP resources

By utilizing various tools and storage systems to manage log data in this way, companies can improve user experience and enhance business competitiveness. It is important to understand the features and advantages of each tool and select the appropriate ones according to their needs.

IV. Log Data Preprocessing

Collected log data must undergo preprocessing before analysis. This is to enhance data quality and ensure the accuracy of the analysis. The preprocessing process includes filtering and cleaning the log data and transforming it into an analyzable structure. This process is essential to ensure data reliability and maximize analysis efficiency.

4.1 Filtering and Cleaning

The filtering and cleaning process of log data involves removing unnecessary data and converting it into a consistent format. Since log data is collected from various sources, the following steps are necessary to ensure data quality and consistency:

- Error Data Removal: Log data often includes errors or incomplete data. For example, logs may be incompletely recorded or stored in the wrong format. Identifying and removing such data enhances the accuracy of the analysis. This is the first step in ensuring data reliability [22].
- Duplicate Data Removal: Removing duplicate data improves data accuracy and efficiency. Duplicate data unnecessarily increases database size and can negatively impact analysis results. Therefore, identifying and removing duplicate log entries is important [23].
- Format Conversion: Log data can exist in various formats, so converting it into a consistent format makes it suitable for analysis. For example, unifying timestamp formats or standardizing log levels (e.g., ERROR, INFO, DEBUG) helps in the integrated analysis of log data [24]
- Unnecessary Information Removal: Log data may include information that is not needed for analysis. For example, debug information or detailed system messages may be unnecessary

for most analysis tasks. Removing such unnecessary information reduces data size and improves analysis efficiency [25]

4.2 Data Transformation

Once filtering and cleaning are complete, log data must be transformed into an analyzable structure. This process includes converting log data into table formats or extracting additional information needed for analysis:

- Structural Transformation: Convert log data into table formats or relational database formats to facilitate easy analysis using SQL queries. This involves mapping log data to database schemas. For example, web server log data can be divided into fields such as user, request time, and request URL and stored in a table format [26]
- Additional Information Extraction: Extract additional information from log data to create new data fields necessary for analysis. For example, extract user IP addresses, requested URLs, and response codes from web server logs and store them in separate fields. This adds depth to the analysis and enables various analytical possibilities [24].
- Data Integration: Integrate log data collected from multiple sources into a single data set. This is useful for providing a comprehensive view by combining log data generated from various systems. For example, integrating web server logs, database logs, and application logs allows for a holistic analysis of user behavior [27]
- Normalization and Denormalization: Apply database normalization techniques to minimize data redundancy and maintain data integrity. Denormalization can be used to optimize analysis performance when necessary. This is a crucial part of data modeling, ensuring efficient data storage and fast retrieval [28].

4.3 Data Sampling and Reduction

To efficiently analyze large-scale log data, data sampling and reduction techniques can be used. This allows analysis to be performed on representative samples without using the entire dataset:

- Random Sampling: Select random samples from the entire log data for analysis. This maintains the representativeness of the data while improving analysis efficiency. Random sampling is useful for identifying general patterns in various log data [29].
- Hierarchical Sampling: Select samples based on specific layers of log data (e.g., user groups, regions) for more detailed analysis. This can more accurately reflect the characteristics of specific layers. For example, VIP user groups and general user groups can be sampled separately for comparative analysis [30].
- Data Reduction: Reduce the size of log data by using summary statistics, aggregation, or extracting key events for analysis. For example, instead of using all logs generated throughout the day, summarized log data by hour can be used. This enables efficient management and analysis of large-scale log data [29].

By following these preprocessing steps, the quality of log data can be improved, and the accuracy and efficiency of analysis can be ensured. Preprocessing is a critical stage in data analysis, essential for producing accurate and reliable analytical results. High-quality data enhances the reliability of analysis results and improves the quality of decision-making based on these results.

V. Log Data Analysis

Based on preprocessed data, various analytical techniques can be applied to derive meaningful insights. These techniques help understand the basic characteristics of the data, visualize patterns and trends, and analyze complex user patterns using machine learning and deep learning methods.

5.1 Descriptive Analysis

Descriptive analysis involves calculating basic statistical measures to understand the overall characteristics of the data. Key statistics include mean, median, and variance, which help understand basic patterns and distributions. For example, by calculating the average response time or the frequency of errors in log data, the overall performance of the system can be evaluated. These statistics are useful for understanding general trends in the data and assessing the performance of the operating system [31].

Key indicators of descriptive analysis include:

- Mean: The central value of a data set, calculated by dividing the sum of all data by the number of data points. For example, the average number of user accesses per day can be calculated to understand traffic patterns.
- Median: The value in the middle of a data set when arranged in order, representing a measure unaffected by outliers. This is useful for understanding the distribution of log data.
- Variance and Standard Deviation: Indicators of how spread out the data values are from the mean, measuring the variability of the data. For example, calculating the standard deviation of response times can help evaluate the consistency of system performance.

5.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis involves visualizing data to explore patterns, trends, and outliers. This helps visually understand the characteristics of the data and identify potential issues. Representative EDA techniques include histograms, box plots, and scatter plots. For example, visualizing user access logs can identify traffic patterns or outliers at specific times [32].

Key techniques of exploratory data analysis include:

- Histogram: Visually represents the distribution of data, showing how data is distributed across different intervals. For example, a histogram of access times throughout the day can reveal when users primarily access the system.
- Box Plot: Visually depicts data distribution and outliers through quartiles. This helps easily identify the range and outliers in log data.
- Scatter Plot: Visualizes the relationship between two variables, helping explore patterns or correlations. For example, a scatter plot of user access times and page load times can identify performance issues.

Through exploratory data analysis, you can understand the basic structure and characteristics of the data, discover data quality issues, and establish hypotheses for further analysis. This serves as a crucial first step in data analysis, allowing you to grasp the overall state of the data and identify areas that require additional analysis.

5.3 Machine Learning and Deep Learning Techniques

Machine learning and deep learning techniques can be used to analyze complex user patterns and build predictive models. Representative techniques include collaborative filtering, content-based filtering, and hybrid methods. These techniques can predict user behavior and provide personalized services [33].

- Collaborative Filtering: This method provides recommendations based on the similarity between users, recommending content that similar users have liked. For example, Netflix recommends movies to users based on what other similar users have enjoyed [11].
- Content-Based Filtering: This method analyzes the characteristics of items that a user has liked in the past and recommends similar items. For example, it recommends news articles on similar topics to those the user has previously read.
- Hybrid Method: This approach combines collaborative filtering and content-based filtering to improve the accuracy of recommendations. By merging the strengths of both methods, it provides better recommendation performance.
-

Analyzing log data using deep learning techniques allows for the identification of more complex patterns and relationships. For example, neural networks can be used to model user behavior and predict future actions. This can be applied to personalized recommendation systems or anomaly detection systems [34]. Examples of deep learning techniques include:

- Deep Neural Networks (DNN): Models that learn and predict complex patterns through multiple layers of neurons.
- Recurrent Neural Networks (RNN): Models effective in processing sequential data over time, commonly used for time series data analysis. For example, they can learn temporal patterns in log data to predict user behavior.
- Convolutional Neural Networks (CNN): Although primarily used for image data analysis, they can also be applied to one-dimensional data. For instance, they can be used to extract and analyze patterns in log data.

These machine learning and deep learning techniques are powerful tools in log data analysis, applicable in various areas such as user behavior prediction, anomaly detection, and personalized service provision. By leveraging these techniques, companies can enhance user experience, improve operational efficiency, and detect security threats in advance.

VI. Utilization Strategies for Log Data

Effectively leveraging log data can yield various business benefits. This chapter details real-time data processing, personalization and custom recommendations, and improving user experience.

6.1 Real-Time Data Processing

Processing log data in real-time can provide immediate insights, which is especially useful in situations requiring rapid decision-making. Representative tools include Apache Kafka, Apache Storm, and Apache Flink.

- Apache Kafka: Kafka is a distributed messaging system that offers high throughput and low latency, making it suitable for processing real-time data streams. For example, Kafka can be used to collect and analyze large-scale log data in real-time, allowing for system monitoring and quick detection of anomalies [16].
- Apache Storm: Storm is a distributed computing system for real-time data processing, capable of handling large-scale data streams in real-time. It is used for log data analysis, real-time event processing, and real-time data analytics [35].
- Apache Flink: Flink is a distributed processing engine for real-time data stream processing, supporting high-performance data stream processing. Flink can be used for real-time data analysis, real-time monitoring, and building real-time data pipelines [36].

Real-time data processing allows companies to receive immediate feedback and make quick decisions. This plays a crucial role in areas such as financial transactions, security monitoring, and real-time marketing campaigns.

6.2 Personalization and Custom Recommendations

Analyzing user behavior data to recommend personalized content can enhance user satisfaction. Representative examples include Netflix, YouTube, and Amazon. These companies use personalization algorithms to optimize user experience.

- Netflix: Netflix provides personalized content recommendations based on users' viewing history and preferences. Netflix's recommendation system combines collaborative filtering, content-based filtering, and deep learning techniques to recommend tailored content to users [11]. For example, Netflix analyzes a user's viewing history to recommend content of similar genres or styles [37].

- YouTube: YouTube analyzes users' viewing history, search history, likes, and comments to recommend related videos. YouTube's recommendation algorithm uses deep learning techniques to predict users' interests and recommend relevant videos. This makes it easier for users to find videos that match their interests and increases user engagement on YouTube. For example, YouTube might recommend new music videos similar to those frequently watched by the user [38].
- Amazon: Amazon analyzes users' purchase history, search history, and shopping cart contents to recommend personalized products. Amazon's recommendation system combines collaborative filtering and content-based filtering to suggest tailored products to users, thereby increasing sales. For example, Amazon might recommend accessories related to a recently viewed product or similar items [39].

Personalization and custom recommendation systems play a crucial role in enhancing user satisfaction and increasing company revenue. These systems enable users to easily find content and products that match their preferences, thereby improving user experience. They are implemented using the following technologies:

- Collaborative Filtering: This method provides recommendations based on the similarity between users. By recommending items favored by users with similar interests, it allows for the discovery of new content or products [40].
- Content-Based Filtering: This method analyzes the characteristics of items that a user has liked in the past to recommend similar items. It finds and recommends items based on the attributes of content or products that the user has previously shown interest in.
- Deep Learning Techniques: These are used to learn complex patterns and predict user behavior. Deep learning automatically extracts meaningful features from large-scale data and provides personalized recommendations based on these features. For example, neural networks can be used to model user preferences and provide recommendations accordingly.
- By leveraging these technologies, companies can accurately understand user interests and preferences, and provide optimized personalized experiences. This not only enhances user satisfaction and encourages continuous user engagement but also contributes to maximizing business performance.

6.3 Improving User Experience

To improve user experience, feedback loops and A/B testing can be utilized. These methods allow for experimenting with various strategies and deriving the optimal solution.

- Feedback Loop: A feedback loop involves improving services based on feedback received from users. For example, analyzing feedback provided through log data can help identify issues with the service and take corrective actions. This helps in increasing user satisfaction and enhancing service quality [41].
- A/B Testing: A/B testing involves comparing two or more versions to determine which one performs better. For instance, two different designs of a website can be presented to user groups randomly, and their responses can be compared to choose the better design. This is useful for optimizing user experience and establishing effective marketing strategies [42].

By employing these techniques, companies can continuously improve user experience, thereby increasing customer satisfaction and loyalty. Feedback loops and A/B testing enable data-driven decision-making, supporting efficient and effective service improvements.

VII. Case Studies

This section examines practical applications of utilizing digital content consumption logs through case studies. It analyzes successful strategies employed by Netflix, YouTube, Amazon, and Instagram,

focusing on how they leverage digital content log data to enhance their services and achieve business success.

7.1 Netflix

Netflix analyzes user viewing data to provide personalized content recommendations, enhancing user satisfaction and reducing churn rates. Netflix's algorithm is optimized to learn user behavior patterns and recommend relevant content. Netflix implements a personalized recommendation system by combining collaborative filtering, content-based filtering, and deep learning techniques [11]. For example, if a user frequently watches movies of a particular genre, Netflix will recommend similar movies, making it easier for the user to discover new content. This personalized recommendation system improves user experience, increases viewing time, and ultimately reduces user churn.

Netflix also continuously evaluates and improves the effectiveness of its recommendation algorithms through A/B testing. By randomly presenting different algorithm versions to user groups and analyzing their responses, Netflix selects the optimal algorithm [37]. This ensures that Netflix can maintain high user satisfaction by providing optimized content recommendations.

7.2 YouTube

YouTube enhances user engagement by recommending related videos based on users' search and viewing histories. YouTube's recommendation system utilizes real-time data to provide the most relevant content to users. The algorithm analyzes the topics, length, and user reactions to the videos watched to recommend similar videos [38]. For example, if a user frequently watches music videos from a particular artist, YouTube will recommend other videos from the same artist or similar genres, encouraging the user to stay on the platform.

YouTube employs deep learning techniques to predict user behavior and offer personalized recommendations. The deep learning model learns patterns from large-scale data and uses these patterns to recommend tailored videos to users. This approach significantly increases user engagement and plays a crucial role in boosting YouTube's advertising revenue.

7.3 Amazon

Amazon increases sales by analyzing purchase and search histories to provide personalized product recommendations. Amazon's recommendation system combines collaborative filtering and content-based filtering to offer a personalized shopping experience to users [39]. For instance, if a user searches for a specific electronic product, Amazon will recommend related accessories or similar products to encourage additional purchases.

Amazon also enhances the accuracy of recommendations by analyzing user reviews, clickstream data, and shopping cart data. This contributes to higher customer satisfaction and increases the repurchase rate. Amazon's personalized recommendation system plays a crucial role in optimizing the customer shopping journey and achieving high sales performance.

7.4 Instagram

Instagram enhances user engagement by analyzing user log data to provide personalized feeds. Instagram's algorithm analyzes likes, comments, and search histories to display the most interesting content to users. For example, if a user frequently interacts with posts on a particular topic or hashtag, Instagram prioritizes related posts at the top of their feed to increase visibility [43].

Instagram uses deep learning techniques to predict user behavior and provide personalized content. This algorithm improves user satisfaction and encourages longer platform engagement, thereby increasing ad revenue.

Instagram also employs A/B testing to experiment with various feed configurations and algorithms to offer the optimal user experience. This continuous improvement helps Instagram maintain high user engagement and satisfaction

VIII. Conclusion

Big data logs generated from digital content consumption are crucial resources for companies to understand user behavior, provide personalized services, and optimize business strategies. This paper presented various methodologies and strategies for effectively collecting, storing, preprocessing, analyzing, and optimizing log data. Through these approaches, digital content companies can improve user experience, enhance operational efficiency, and strengthen their competitiveness.

This paper analyzed case studies of big data log usage in digital content and presented ways for companies to achieve data-driven innovation. Such approaches offer benefits including personalized service provision, real-time monitoring and problem-solving, enhanced security, and strengthened business intelligence.

For future research, adding a big data log processing architecture that includes the entire process from data collection to storage, processing, analysis, and visualization would be beneficial. This integrated system design would enable companies to utilize log data more efficiently and respond quickly to the rapidly changing market environment.

The digital content industry will continue to evolve through the analysis and optimization of big data logs. Companies can achieve better user experiences and higher business performance by implementing the strategies presented in this paper.

IX. References

[1] Maryville Online, "What Is Digital Media? All You Need to Know About New Media," 2020. [Online]. Available: https://online.maryville.edu/blog/what-is-digital-media/

[2] Wikipedia, "Digital content," 2020. [Online]. Available: https://en.wikipedia.org/wiki/Digital_content

[3] O. A. El Sawy, F. Pereira, "Digital Business Models: Review and Synthesis," in Business Modelling in the Dynamic Digital Space, Springer, Berlin, Heidelberg, 2013. [Online]. Available: https://doi.org/10.1007/978-3-642-31765-1_2

[4] M. Chen, S. Mao, Y. Liu, "Big Data: A Survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171-209, Apr. 2014

[5] S. Lee, D. Lee, "Current Status of Big Data Utilization," Digital Fusion Research, vol. 11, no. 2, pp. 229-233, Jun. 2013

[6] "Dell Announces New Innovations in Digital Universe," 2014. [Online]. Available: https://www.dell.com/ko-kr/dt/corporate/newsroom/announcements/2014/04/20140410.htm

[7] L. Johnston, "A 'Library of Congress' Worth of Data: It's All In How You Define It," 2012. [Online]. Available: https://blogs.loc.gov/thesignal/2012/04/a-library-of-congress-worth-of-data-its-all-in-how-you-define-it/

[8] E. Dumbill, A. Croll, J. Steele, M. Loukides, Planning for big data, Beijing: O'Reilly Media, 2012.

[9] J. Dijcks, "Big Data for the Enterprise," 2013. [Online]. Available: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

[10] K. Kent, M. Souppaya, "Guide to Computer Security Log Management," NIST Special Publication 800-92, 2006. [Online]. Available: https://csrc.nist.rip/library/NIST%20SP%20800-092%20Guide%20to%20Computer%20Security%20Log%20Management,%202006-09.pdf

[11] C. A. Gomez-Uribe, N. Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation," ACM Transactions on Management Information Systems (TMIS), vol. 6, no. 4, pp. 13, Dec. 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2843948

[12] A. Ahmad, "Monitoring Big Data Streams Using Data Stream Management Systems: Industrial Needs, Challenges, and Improvements," Hindawi Journal of Big Data, vol. 2023, art. 2596069, 2023. [Online]. Available: https://doi.org/10.1155/2023/2596069

[13] H. Chen, R. H. Chiang, V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly, vol. 36, no. 4, pp. 1165-1188, Dec. 2012. [Online]. Available: https://doi.org/10.2307/41703503

[14] "How to collect logs with Fluentd," Is It Observable. [Online]. Available: https://isitobservable.io/docs/how-to-collect-logs-with-fluentd/

[15] Elastic, "Elasticsearch: The Official Distributed Search & Analytics Engine," Elastic Official Website. [Online]. Available: https://www.elastic.co/elasticsearch/
[16] J. Kreps, N. Narkhede, J. Rao, "Kafka: A Distributed Messaging System for Log Processing," in Proceedings of the NetDB, 2011, pp. 1-7.
[17] "Using Hadoop Distributed File System (HDFS)," Apache Hadoop. [Online]. Available: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
[18] S. Ghemawat, H. Gobioff, S. T. Leung, "The Google File System," in Proceedings of the nineteenth ACM symposium on Operating systems principles, 2003, pp. 29-43.
[19] "Amazon CloudWatch Logs," AWS Documentation. [Online]. Available: https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/WhatIsCloudWatchLogs.html
[20] "Azure Monitor overview," Microsoft Azure Documentation. [Online]. Available: https://docs.microsoft.com/en-us/azure/azure-monitor/overview
[21] "Google Cloud Logging," Google Cloud Documentation. [Online]. Available: https://cloud.google.com/logging/docs
[22] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques (3rd ed.), Morgan Kaufmann, 2012, pp. 89. [Online]. Available: https://www.sciencedirect.com/science/book/9780123814791
[23] V. Ganti, R. Ramakrishnan, "Data Cleaning: A Practical Perspective," Morgan & Claypool Publishers, 2011. [Online]. Available: https://doi.org/10.2200/S00523ED1V01Y201307DTM036
[24] P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, G. Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2012, pp. 45.
[25] T. Dasu, T. Johnson, Exploratory Data Mining and Data Cleaning, Wiley-Interscience, 2003.
[26] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.), 2013, pp. 281.
[27] J. Gantz, D. Reinsel, "Extracting Value from Chaos," IDC iView, 2011. [Online]. Available: https://doi.org/10.12691/jcsa-3-6-13
[28] R. Elmasri, S. B. Navathe, Fundamentals of Database Systems (6th ed.), Addison-Wesley, 2010, pp. 547.
[29] T. Lukoianova, V. L. Rubin, "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?" Advances in Classification Research Online, vol. 24, no. 1, pp. 4-15, 2014. DOI: 10.7152/acro.v24i1.14671
[30] X. Meng, "Scalable Simple Random Sampling and Stratified Sampling," in Proceedings of the 30th International Conference on Machine Learning, PMLR vol. 28, no. 3, pp. 531-539, 2013. [Online]. Available: https://proceedings.mlr.press/v28/meng13a.html
[31] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques (3rd ed.), Morgan Kaufmann, 2012, pp. 83-85. [Online]. Available: https://www.sciencedirect.com/science/book/9780123814791
[32] E.J.M. Carranza, "Exploratory Data Analysis," in Encyclopedia of Mathematical Geosciences, Encyclopedia of Earth Sciences Series, Springer, Cham, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-26050-7_105-1
[33] K. P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012, pp. 10-15.
[34] W. Meng, Y. Liu, Y. Zhu, S. Zhang, D. Pei, Y. Liu, Y. Chen, R. Zhang, S. Tao, P. Sun, R. Zhou, "LogAnomaly: Unsupervised Detection of Sequential and Quantitative Anomalies in Unstructured Logs," in Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19), pp. 4739-4745, 2019. [Online]. Available: https://www.ijcai.org/Proceedings/2019/657
[35] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Schneider, et al., "Storm@twitter," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pp. 147-156, 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2588555.2595641
[36] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, K. Tzoumas, "Apache Flink: Stream and Batch Processing in a Single Engine," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 36, no. 4, pp. 28-38, Dec. 2015.
[37] X. Amatriain, J. Basilico, "Netflix Recommendations: Beyond the 5 stars (Part 1)," Netflix Tech Blog, 2012. [Online]. Available: https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429
[38] P. Covington, J. Adams, E. Sargin, "Deep Neural Networks for YouTube Recommendations," in Proceedings of the 10th ACM Conference on Recommender Systems, pp. 191-198, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2959100.2959190

[39] B. Smith, G. Linden, "Two Decades of Recommender Systems at Amazon.com," IEEE Internet Computing, vol. 21, no. 3, pp. 12-18, May-Jun. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7927889
[40] F. Ortega, Á. González-Prieto, "Recommender Systems and Collaborative Filtering," Applied Sciences, vol. 10, no. 20, pp. 7050, Oct. 2020.
[41] "How to Create a Customer Feedback Loop That Works," Help Scout, 2023. [Online]. Available: https://www.helpscout.com/blog/customer-feedback-loop/
[42] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: Survey and practical guide," Data Min. Knowl. Discov., vol. 18, no. 1, pp. 140-181, Feb. 2009. Available: https://doi.org/10.1007/s10618-008-0114-1
[43] Influencer Marketing Hub, "How the Instagram Algorithm Works in 2024," 2024. [Online]. Available: https://influencermarketinghub.com/instagram-algorithm/

Authors

Haeri An

2019.06 : MSC Sheffield Hallam University, UK
2024.03 ~ Present : Assistant Professor in Dept. of IT Software, Shingu College

Research Interests : Bigdata, Cloud, IT Infra, Cloud Platform, Computer Engineering