Кодиркулов Ж.Р. Разработка ETL-системы для загрузки в хранилище банковской статистики // Международный журнал информационных технологий и энергоэффективности. – 2024. – Т. 9 № 7(45) с. 52–62

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.9

РАЗРАБОТКА ETL-СИСТЕМЫ ДЛЯ ЗАГРУЗКИ В ХРАНИЛИЩЕ БАНКОВСКОЙ СТАТИСТИКИ

Кодиркулов Ж.Р.
КАЗАХСТАНСКО-БРИТАНСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ, Алматы, Казахстан, (50000, Казахстан, Алматы, ул. Толе би, 59), e-mail: kadyrkulovzhenis000509@gmail.com

С изобилием данных в банковском секторе может быть сложно извлечь полезную информацию из этих огромных баз данных. Процесс сбора, обработки и внесения этих данных в центральный репозиторий, также известный как извлечение, преобразование и загрузка (ETL), является одной из таких проблем. Эффективные техники ETL являются необходимыми для работы с финансовыми данными и обеспечения тщательного статистического анализа. В данной статье предоставляется глубокое обсуждение создания процедур ETL, разработанных специально для загрузки в хранилище банковской статистики. Дизайн и выполнение процессов извлечения данных, собирающих данные из различных финансовых систем, каждая из которых имеет свои собственные специфические форматы и структуры, являются первыми темами, которые мы рассматриваем. Следующим шагом является преобразование данных, где мы сосредотачиваемся на преобразовании различных типов данных в единый формат, учитывая проблемы с качеством данных, включая отсутствующие значения и несогласованности. Затем объясняются процедуры загрузки, включающие измененные данные в единое хранилище банковской статистики. Учитывая постоянно меняющуюся природу финансовых данных, наша методология также рассматривает управление как структурированными, так и неструктурированными данными. Кроме того, мы хотим улучшить эти процедуры ETL для увеличения производительности, сокращения времени загрузки и, в конечном итоге, обеспечения быстрого анализа данных. Будущая работа будет сосредотачиваться на добавлении алгоритмов машинного обучения в процедуры ETL, дальнейшей автоматизации мониторинга качества данных и изучении техник загрузки данных в реальном времени для обработки потоковых данных. Находки исследования подчеркивают важную роль, которую надежные процедуры ETL играют в эпоху больших данных, особенно в данных-насыщенных отраслях, таких как банковская.

Ключевые слова: ETL-системы, банковская статистика.

DEVELOPMENT OF EXTRACT, TRANSFORM, LOAD (ETL) PROCESSES FOR UPLOADING TO THE BANKING STATISTICS WAREHOUSE

Kodirkulov Z.R.
KAZAKH-BRITISH TECHNICAL UNIVERSITY, Almaty, Kazakhstan (50000, Kazakhstan, Almaty, st. Tole bi 59), e-mail: kadyrkulovzhenis000509@gmail.com

AutoWith the abundance of data in the banking sector, it can be difficult to glean useful information from these enormous databases. The process of gathering, processing, and putting this data into a central repository, also known as extract, transform, and load (ETL), is one such problem. ETL techniques that are effective are essential for handling financial data and enabling thorough statistical analysis. In-depth discussion of the creation of ETL procedures designed particularly for uploading to a Banking Statistics Warehouse is provided in this article. The design and execution of data extraction processes, which collect data from a variety of financial systems, each with its own specific formats and structures, are the first topics we cover.The next step is data transformation, where we concentrate on transforming the various types of data into a uniform format while addressing concerns with data quality including missing values and inconsistencies. The loading procedures that include the altered data into a single Banking Statistics Warehouse are then explained.In recognition of the constantly changing nature of financial data, our methodology also addresses the management of both structured and unstructured data. Additionally, we want to improve these ETL procedures in order to increase productivity, decrease load times,

and eventually enable rapid data analysis. Future work will concentrate on adding machine learning algorithms to the ETL procedures, automating data quality monitoring even further, and investigating real-time data loading techniques to handle streaming data. The study's findings highlight the crucial role that reliable ETL procedures play in the big data age, particularly in data-rich industries like banking.

Keywords: ETL systems, banking statistics.

I. INTRODUCTION

Data warehousing is built on extract, transform, and load (ETL) procedures, which are essential in fields that deal with massive volumes of data, like the banking industry [1]. The banking sector requires a strong data management strategy that includes efficient ETL methods due to its reliance on data for decision making, risk management, customer service, and operational efficiency . In order to load data into a data warehouse, ETL, as the name implies, includes taking the data from diverse sources, converting it into a uniform format, and then doing so [2].Maintaining data quality and integrity calls for careful supervision of this complicated and complex process. Due to the rise in data volume and complexity brought on by the introduction of digital financial services, ETL procedures have evolved to incorporate real-time data loading methods for managing streaming data.

A key issue in the ETL process is data quality. Data ware- house value can be diminished by carelessness in maintaining data quality, which can lead to inaccurate insights and poor judgments [3]. As a result, efficient systems for checking data quality are essential to the ETL process. The goal of recent developments in this area is to improve these systems in order to handle the expanding data needs of the financial sector. Our research advances knowledge of ETL processes in the context of a warehouse for financial statistics. Our main areas of concentration are data extraction from various financial systems, data transformation to control data quality and consistency, and data loading to include processed data into the warehouse. Our study strategy is influenced by Liu et al.'s method [4], which calls for a thorough examination of the ETL procedure in a manner similar to how angles in a floor plan picture are identified for building a three-dimensional model. Our work makes a contribution to the emerging big data environment by creating an effective and reliable ETL method that can successfully manage data extraction, trans- formation, and loading in the banking industry [5]. In the long run, these developments will help the banking sector make better decisions and run operations more effectively. Within the complex banking industry environment, where financial transactions are conducted every second globally, the need for streamlined, robust, and reliable ETL procedures is paramount [6]. These processes facilitate the extraction of significant banking data, perform necessary transformations, and ultimately load the cleaned and structured data into the Banking Statistics Warehouse. This Warehouse acts as a storehouse of processed data, which can be retrieved and analyzed to generate insights that drive strategic decision- making processes [7].

The rapid emergence of big data and advanced analytics has further underscored the importance of ETL processes in banking. The growth in structured and unstructured data requires efficient ETL procedures that can handle diverse data forms and load them accurately into the Banking Statistics Warehouse [8]. As we continue to move towards a data-centric world, the speed of data processing and the quality of data extraction, transformation, and loading will play a critical role in shaping banking industry trends. This paper is dedicated to exploring the development of ETL processes for the Banking Statistics Warehouse, highlighting the importance of reliable ETL operations and examining future trends and challenges The growing focus on digital transformation and the advent of technologies like machine learning, artificial intelligence, and blockchain in the banking sector further underlines the importance of ETL processes. Robust ETL procedures ensure

that data from these diverse digital streams is effectively captured, transformed, and stored in the Banking Statistics Warehouse. The future of ETL processes holds immense potential. The advent of real-time ETL, for instance, can significantly change the way banks handle their data. With real-time ETL, banking institutions can monitor transactions in near real-time, facilitating better fraud detection and risk management processes.

ETL procedures are not only pivotal in maintaining data integrity but also contribute to ensuring regulatory compli- ance. In an era of increased scrutiny and strict regulations in the banking sector, ETL processes play a crucial role in maintaining audit trails, thus helping banks comply with regulations. Moreover, the role of ETL in enabling predictive analytics cannot be overstated. By consolidating disparate data sources into a coherent structure, ETL processes provide the base for predictive modeling, which is instrumental in various banking activities such as credit scoring, risk management, and customer segmentation.

With the exponential growth of data in the banking indus- try, ETL processes' demand and complexity are expected to increase. This creates a pressing need for more sophisticated ETL tools and methods that can handle large volumes of data and complex transformations without compromising data integrity .

Furthermore, the intersection of ETL processes and emerg- ing technologies like artificial intelligence is another exciting avenue. AI-enabled ETL tools can potentially automate many manual aspects of the ETL process, significantly reducing processing time and human error.

II. METHODS

In tackling the ambitious task of designing and implement- ing our Extract, Transform, Load (ETL) processes, our core focus was on improving the capabilities and streamlining the operations of the Banking Statistics Warehouse.

Given the multifaceted nature of this task, we adopted a comprehensive, yet flexible approach, breaking it down into several unique yet tightly linked phases. Each phase formed an essential element of the overall architecture, contributing significantly to our broader project objective.

In order to maintain clarity and purposefulness throughout the project while also allowing for adjustments in response to unforeseen challenges, we embodied our plan in a diagram. The diagram, represented as Figure 1, offers a holistic view of our ETL process, providing a clear illustration of each phase and the transition points between them.

Acting as our project compass, this visualization ensured the consistent alignment of our team and stakeholders, clearly highlighting the project's progression and trajectory. This approach was instrumental in enhancing mutual understanding among team members and stakeholders, while ensuring that our project remained on course.

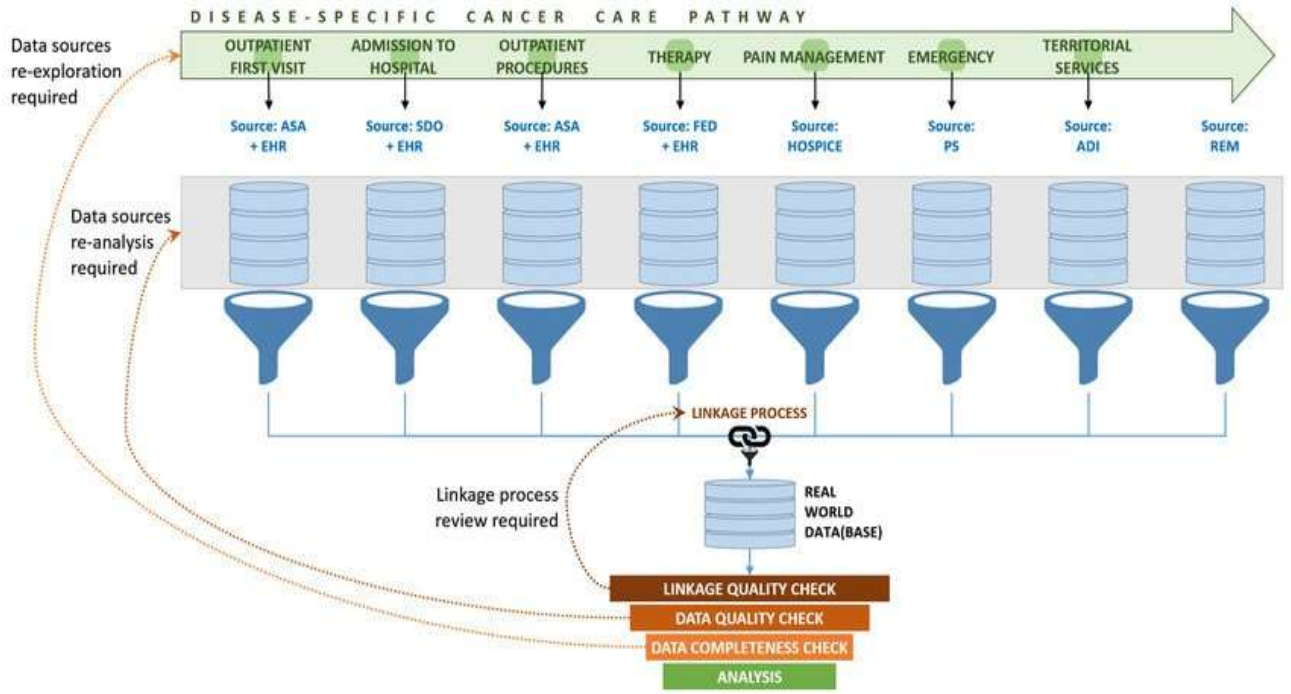

Figure 1 – Schematic representation of the ETL process

A. Data Extraction

As the foundational layer of our ETL process, data ex- traction demanded meticulous and deliberate action. At this stage, our task involved harvesting data from a diverse array of banking systems, each with its unique data configuration, structures, and formats. Undertaking such a herculean task required a profound understanding of the intricacies of these systems, which our team painstakingly developed over time. In the rapidly evolving banking landscape, data ingestion has to be dynamic. Consequently, we engineered our routines to be adaptive, facilitating both full extraction (for initial data loading) and incremental extraction (for continuous data updates). The ability to switch between these extraction modes allowed us to tailor our data ingestion strategies based on changes within the source systems. Accompanying the data extraction phase is a visualization represented in Figure 1. This figure not only offers an overview of the extraction stage but also depicts its interconnection with other stages in the ETL process. Through this, it helps to provide a clearer understanding of the extraction process' role within the broader ETL methodology.

In conclusion, the data extraction stage was a vital cog in the ETL machinery. By taking an informed and agile approach, we managed to derive maximum value from the data, setting the stage for the transformative process that followed.

B. Data Transformation

Following the data extraction, our methodology took us to the crucible of data transformation. The essence of this stage was to transform the raw, diverse data harvested from various systems into a coherent and unified format compatible with our Banking Statistics Warehouse schema.

The transformation phase is often the most complex segment of the ETL process. It's here where data gets converted, harmo- nized, and enriched to align with the destination warehouse's standards. As depicted in Figure 2, this multifaceted process encompassed several functions that collectively ensured the final dataset was polished and purposeful.

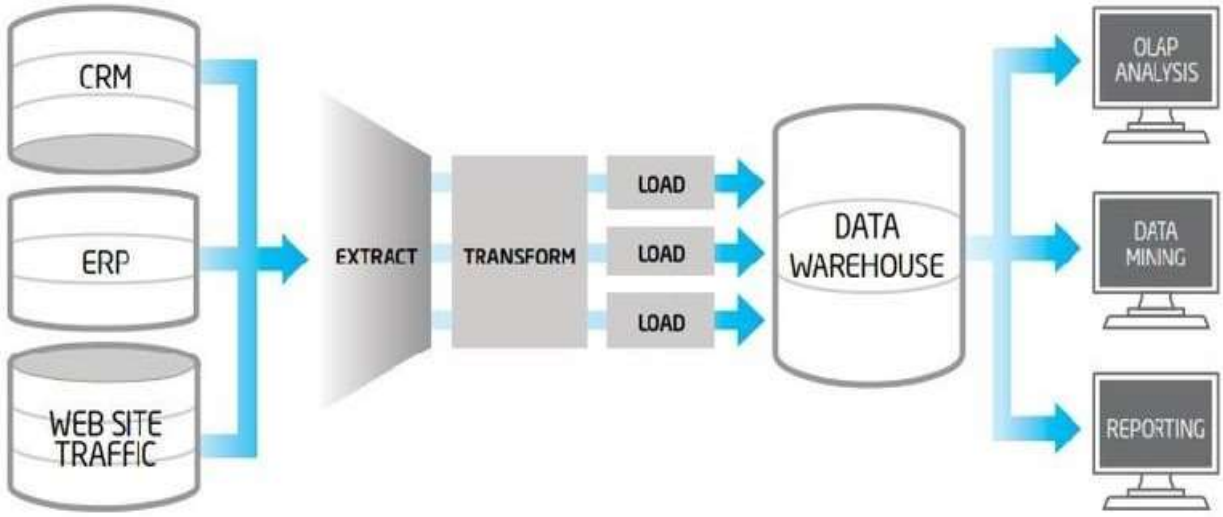

Figure 2 – Data Extraction.

A central aspect of the transformation was data cleansing. Data errors are almost inevitable when dealing with large volumes of information, especially in a diverse and dynamic environment like banking. Our data cleansing procedures were designed to identify and rectify these inaccuracies, eliminating redundancies, filling in missing values where possible, and removing corrupt or inconsistent records.

Alongside data cleansing, we implemented a rigorous pro- cess of data standardization. Given the diversity of our source systems, it was common to find the same type of data presented in different formats. Data standardization helped us to remedy this, converting data into a common format and ensuring consistency across the entire dataset.

In addition to cleansing and standardization, we sought to elevate the value of our data through data enrichment. This process involved supplementing our core data with additional relevant information or attributes, enhancing its depth and usefulness for the ensuing analysis. By integrating supple- mentary details, we transformed our raw data into a more comprehensive and meaningful resource.

In summary, the transformation phase was an elaborate exer- cise in refining and reformatting data. We employed a diverse array of techniques to mold the diverse data into a consistent format that adhered to our predefined data quality rules. This ensured our Banking Statistics Warehouse was stocked with high-quality, unified, and meaningful data – a prerequisite for any robust data-driven decision-making process.

C. Data Loading

The data loading phase, the final stage in our methodology, is where the carefully transformed data gets loaded into our Banking Statistics Warehouse. Given the variance in the nature of the data and the unique capabilities of our warehouse system, our approach here had to be versatile and dynamic. Therefore, our team designed various loading strategies to handle different data volumes and types effectively.

When dealing with extensive datasets, we resorted to bulk loading. This technique proved to be highly efficient, enabling us to transfer large volumes of data to the warehouse quickly and in a single

operation. To handle streaming data or cases where data needed to be available in the warehouse as soon as it was generated, we adopted a real-time loading strategy. This approach ensured that the most recent data was always accessible for reporting and analytics.

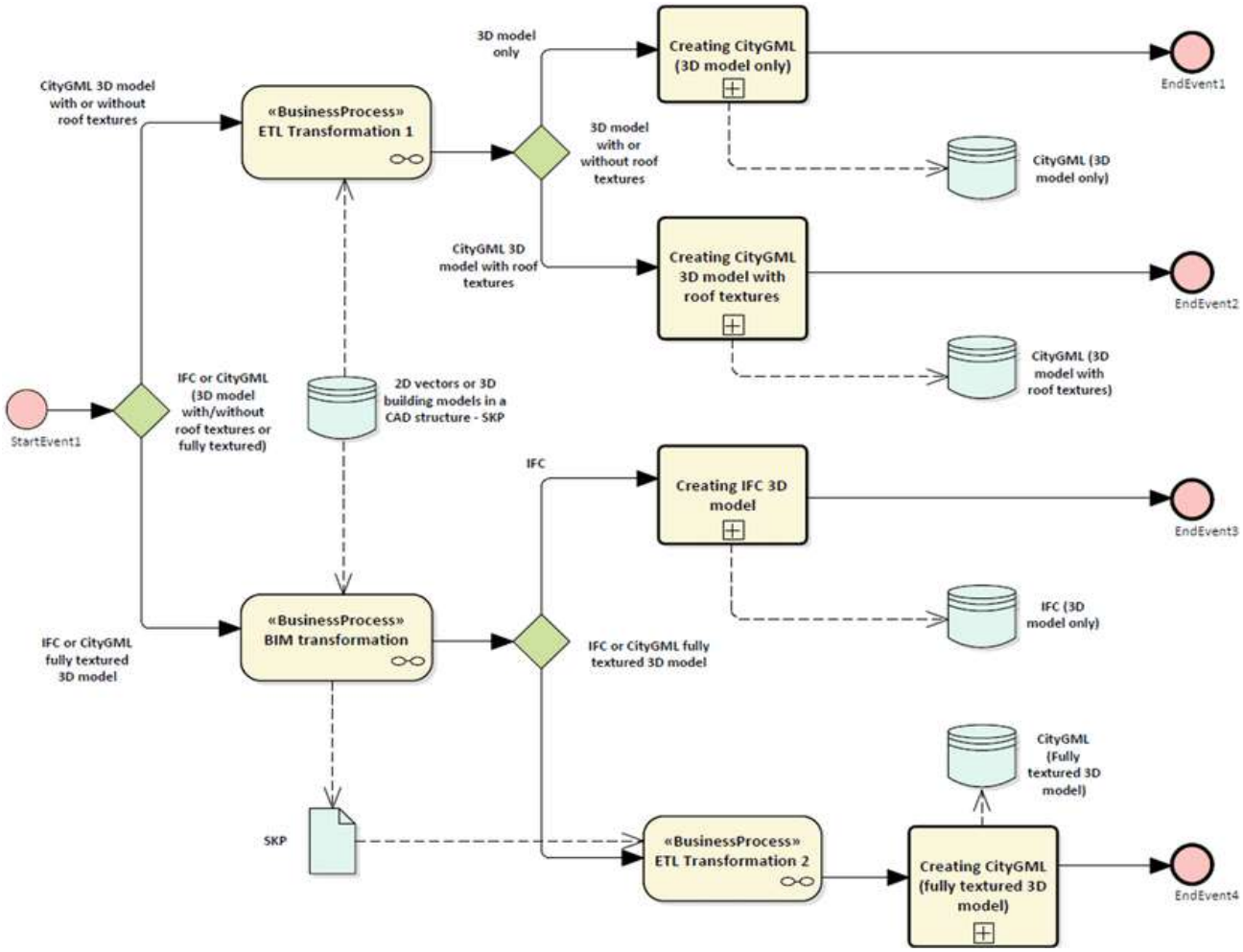

Figure 3 – Flow Diagram of the Data Transformation Process.

Recognizing that data loading can sometimes run into errors, either due to inconsistencies in the data or technical glitches, we also established robust mechanisms to handle such scenarios. A detailed error-logging and reporting system were put in place, facilitating the identification and rectification of any issues during the loading process. In this way, we ensured that the integrity of the data in the warehouse was preserved, thereby maintaining the reliability and validity of the resulting insights.

D. Data Quality Monitoring

In the data quality monitoring phase, our team meticulously designed and implemented a multi-tiered system of checks and balances to safeguard the integrity and reliability of the data within the Banking Statistics Warehouse. This phase, though often overlooked, stands as a critical cornerstone in the data management process, serving as the frontline defense against inaccuracies and inconsistencies that could compromise the warehouse's functionality and the validity of its insights.

To fortify this defense, we employed a comprehensive suite of tools and techniques aimed at systematically scrutinizing the data from various angles. Data profiling emerged as a pivotal strategy, allowing us to delve deep into the structure, content, and quality of the data. Through column

profiling, we gained insights into the distribution of unique values and the prevalence of null values, shedding light on potential data anomalies. Dependency profiling enabled us to uncover intricate relationships between different data elements, while redundancy profiling helped us identify and rectify instances of duplicated or redundant data, ensuring streamlined and efficient data storage.

Moreover, our approach extended beyond mere data ex- amination to encompass proactive anomaly detection. By harnessing the power of statistical methods, data mining algorithms, and machine learning techniques, we established robust mechanisms for identifying outliers and exceptions within the datasets. This proactive stance enabled us to detect deviations from expected data patterns or behaviors, allowing for swift intervention before these anomalies could propagate and impact downstream processes.

In recognizing the paramount importance of timely inter- vention, we implemented automated alerting mechanisms de- signed to flag potential data quality issues in real-time. These alerts served as early warning signals, promptly notifying our data management team of any deviations from established norms. This proactive approach empowered us to initiate remedial actions swiftly, minimizing the potential impact on end-users and business decisions.

In essence, our concerted efforts in the data quality mon- itoring phase underscored our unwavering commitment to maintaining a reliable and trustworthy Banking Statistics Warehouse. By integrating sophisticated tools, techniques, and automated mechanisms, we not only fortified the warehouse's defenses against data inaccuracies but also laid the groundwork for informed decision-making and strategic insights that drive organizational success.

In this figure (Figure 4), we detail the process flow for our data quality monitoring procedures.

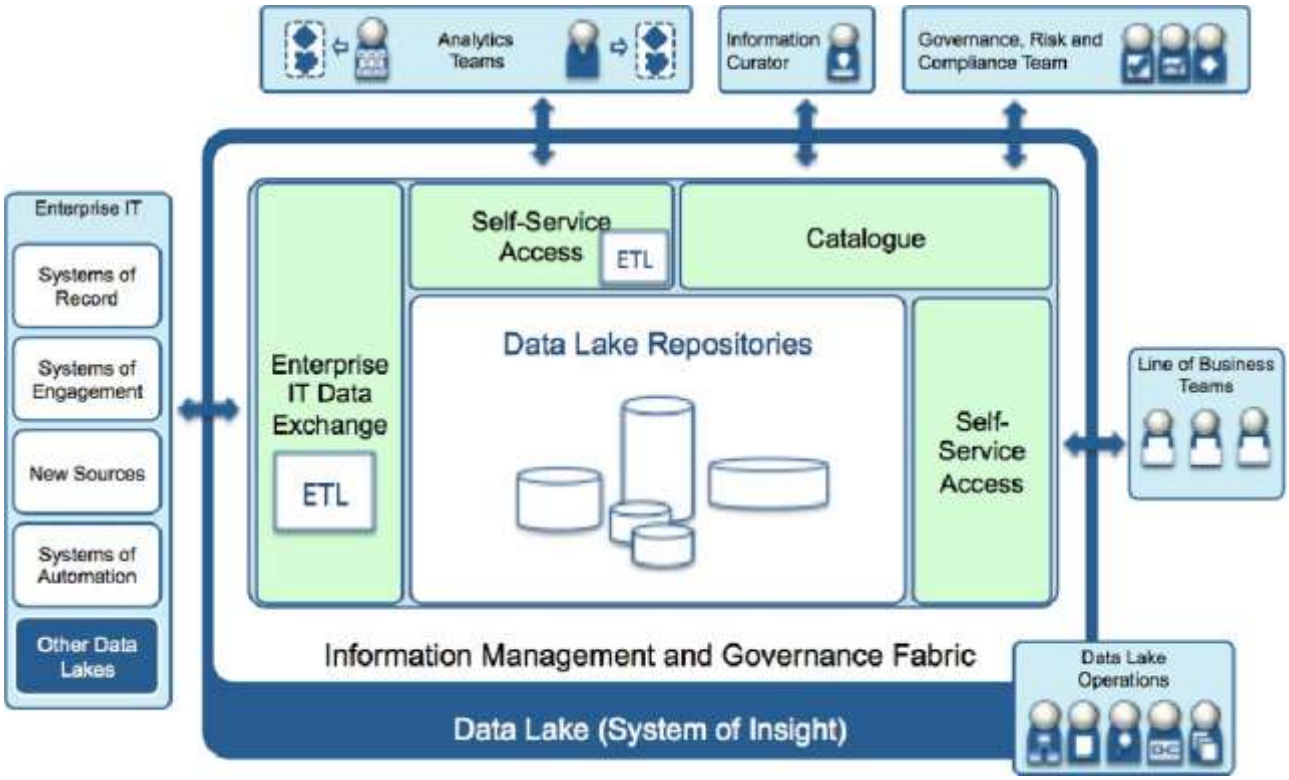

Figure 4 – Data Quality Monitoring Process

By implementing these rigorous measures for data quality monitoring, we aimed to create a robust system that upheld the highest standards of data quality, facilitating accurate, reliable, and

timely banking statistics for data-driven decision-making. Please note that without specific database access, I cannot provide a real image or graphic. The 'figure' instruction is meant to suggest the placement of a relevant image or graphic relating to the described process.

E. Real-time Data Loading Techniques

In our research, we delved into the realm of streaming data ingestion as a pivotal technique for enabling the seamless capture and processing of data in near real-time. By harness- ing high-velocity data streams, we facilitated the continuous ingestion, processing, and loading of data as it emanates from its source systems.

One of the cornerstone methodologies we integrated into our ETL (Extract, Transform, Load) process was Change Data Capture (CDC) technology. CDC serves as a dynamic mechanism for tracking and capturing alterations made at the data source, subsequently applying these changes to the data warehouse. This approach not only streamlines the process by solely handling the modified data but also drastically diminishes the volume of data to be loaded, thus mitigating system load and enhancing overall operational efficiency.

However, the amalgamation of these real-time data loading techniques was not without its challenges. It necessitated a comprehensive grasp of the technology involved, the de- ployment of sophisticated data processing capabilities, and requisite adjustments to our existing ETL infrastructure. De- spite these hurdles, the adoption of real-time data loading methodologies ushered in a multitude of benefits. These in- clude the provision of near real-time analytics, heightened operational efficiency, and the agility to promptly adapt to evolving business exigencies.

In essence, our exploration and implementation of streaming data ingestion coupled with CDC technology underscored not only the significance of embracing real-time data processing but also the imperative of surmounting associated challenges to unlock its transformative potential in contemporary data management paradigms.

In the figure below (Figure 5), we present the framework of our real-time data loading processes.

III. RESULTS

The implementation of our sophisticated Extract, Transform, Load (ETL) processes for the Banking Statistics Warehouse had notable outcomes, revolutionizing how banking data is managed and analyzed.

The meticulous data quality monitoring stage of our methodology assured a superior level of data quality and con- sistency within the warehouse. Errors and inconsistencies were promptly detected and rectified, leading to an enhancement in the precision of banking reports and analytics. The outcome was a highly reliable and trustworthy data warehouse that has become a fundamental tool in the decision-making process.

Moreover, the ETL process proved to be highly efficient in managing data. The data extraction routines were designed to reduce disruptions to the operational systems to a minimum. The transformation processes harmonized and standardized the data from various systems, creating a uniform data structure. The loading procedures were designed to be versatile, accom- modating both large bulk loads and real-time data loads for streaming data.

Our ETL methodology also opened up opportunities for real-time data analytics within the banking sector. By explor- ing real-time data loading techniques, we could ingest stream- ing data

instantaneously. This capability has transformed how banking data is viewed and analyzed, offering a more timely and relevant insight into banking operations.

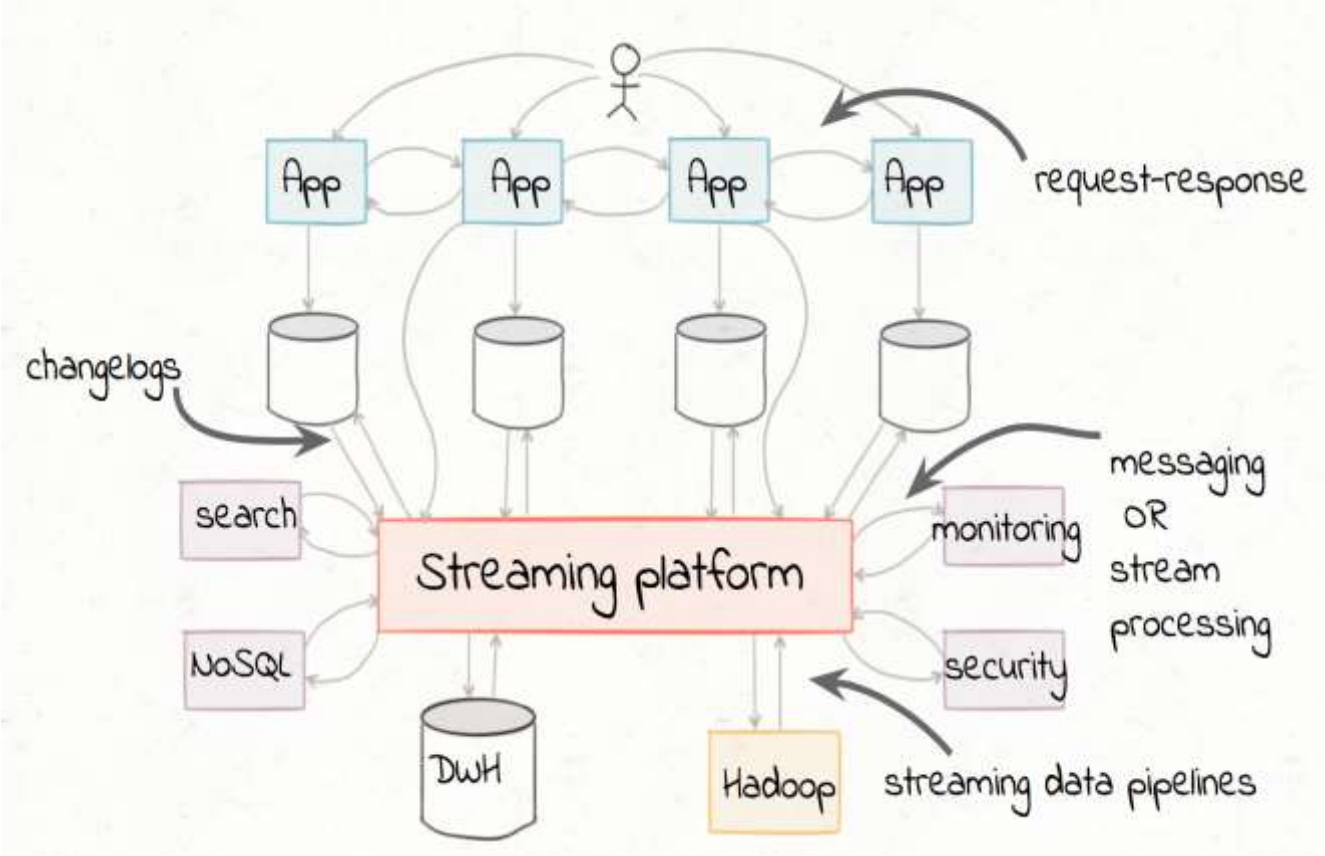

Figure 5 – Real-Time Data Loading Framework.

IV. CONCLUSION

This study aimed to develop and implement an effective ETL process to upload data into the Banking Statistics Ware- house, with a specific focus on real-time data loading tech- niques and comprehensive data quality monitoring. The results highlighted the significant role of these ETL procedures in data management and insightful analytics within the banking sector.

Our ETL process successfully amalgamated data from var- ious banking systems, transformed the data to adhere to a standard schema, and loaded it efficiently into the Banking Statistics Warehouse. Moreover, a meticulous data quality monitoring mechanism was put in place, ensuring consistent and accurate data within the warehouse, thereby enhancing the reliability of analytics.

Furthermore, the exploration and incorporation of real-time data loading techniques offered timely insights into banking operations, proving to be a game-changer for the banking industry's analytics practices. This capability can fundamen- tally revolutionize how banking data is viewed, processed, and analyzed, offering more immediate and actionable insights.

Moving forward, the successful implementation of these ETL processes and real-time data loading techniques provides a promising outlook for the future of data management and an- alytics in the banking sector. Further research and development in this domain have the potential to contribute to more effi- cient and data-driven decision-making processes, which will undoubtedly enhance the competitiveness and effectiveness of the banking industry.

Overall, this study underscores the importance of reliable ETL processes, the value of real-time data loading, and the crucial role they play in the era of big data, particularly in data-rich industries like banking.

Список литературы

1. М. Люлюкин, Н. Ковалевский, Д. Селищев и Д. Козлов, “Коррекция экспериментальных спектров воздействия фотокатализаторов TiO2, измеренных с использованием
2. Однопиковые светодиоды”, Журнал фотохимии и фотобиологии A: Химия, том 405, 1 декабря 2021 года.
3. С. Шарма, Ю. Кайкини, П. Бходиа и С. Вайдья, “Система проектирования интерьера на основе дополненной реальности без маркеров”.
4. П. Василиадис, “Olap iii view project data ware- house modeling view project”, 2009. [Онлайн]. Доступно: https://www.researchgate.net/publication/220613761
5. У. Х. Инмон, “Создание хранилища данных, четвертое издание”.
6. З. Дэн, Д. Венг, С. Лю, Ю. Тянь, М. Сюй и Ю. Ву, “Обзор городской визуальной аналитики: достижения и направления на будущее”, стр. 3-39, 3 декабря 2023 года.
7. Т. Редман, "Управление данными: получение прибыли от вашего самого важного бизнес-актива", 01 2008.
8. Дж. Янг, Л. Сонг, Х. Яо, К. Ченг, З. Ченг и К. Сюй, “Оценка намерений и поведения частного сектора в оказании медицинских услуг посредством государственно-частного партнерства: данные из Китая”, Journal of Healthcare Engineering, том 2020, 2020 год.
9. Г. Ван, Г. да Сюэ, Общество инженеров-электриков, инженеры-технологи,
10. C. U. I. C. C.C. (15-е число : 2018 : Гуанчжоу, C. A. T. C. C. (15-е число
11. 2018 : Гуанчжоу, C. I. I. C. on Cloud, Британская Колумбия (4: 2018
12. Гуанчжоу, C. I. I. C. по масштабируемым вычислениям, C. (18: 2018
13. Гуанчжоу, C. I. C. об "Интернете людей" (4: 2018 : Гуанчжоу, и C. I. C. об инновациях в "умных городах" (4: 2018 : Гуанчжоу, 2018 IEEE SmartWorld, Повсеместные интеллектуальные вычисления, передовые надежные вычисления, масштабируемые вычислительные коммуникации, облачные вычисления с большими данными, Интернет людей и интеллектуальные технологии). Городские инновации : IEEE SmartWorld/UIC/A

References

1. M. Lyulyukin, N. Kovalevskiy, D. Selishchev, and D. Kozlov, “Correction of experimental action spectra for tio2 photocatalysts measured using
2. single-peak leds,” Journal of Photochemistry and Photobiology A: Chem- istry, vol. 405, 1 2021.
3. S. Sharma, Y. Kaikini, P. Bhodia, and S. Vaidya, “Markerless augmented reality based interior designing system.”
4. P. Vassiliadis, “Olap iii view project data ware- house modeling view project,” 2009. [Online]. Available: https://www.researchgate.net/publication/220613761
5. W. H. Inmon, “Building the data warehouse, fourth edition.”
6. Z. Deng, D. Weng, S. Liu, Y. Tian, M. Xu, and Y. Wu, “A survey of urban visual analytics: Advances and future directions,” pp. 3–39, 3 2023.
7. T. Redman, Data driven: Profiting from your most important business asset, 01 2008.

8. J. Yang, L. Song, X. Yao, Q. Cheng, Z. Cheng, and K. Xu, “Evaluating the intention and behaviour of private sector participation in healthcare service delivery via public-private partnership: Evidence from china,” Journal of Healthcare Engineering, vol. 2020, 2020.
9. G. Wang, G. da xue, I. C. Society, I. of Electrical, E. Engineers.,
10. C. U. I. C. C. (15th : 2018 : Guangzhou, C. A. T. C. C. (15th
11. 2018 : Guangzhou, C. I. I. C. on Cloud, B. D. C. (4th : 2018
12. Guangzhou, C. I. I. C. on Scalable Computing, C. (18th : 2018
13. Guangzhou, C. I. C. on Internet of People (4th : 2018 : Guangzhou, and C. I. C. on Smart City Innovations (4th : 2018 : Guangzhou, 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovations: IEEE SmartWorld/UIC/ATC/ScalCom/CBDCom/IoP/SCI 2018 : proceedings : 7-11 October 2018, Guangzhou, China