Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Effectiveness of Digital Transformation in Data Warehouse Technology

Siti Rohajawati
Email siti.rohajawati@bakrie.ac.id
Sistem Informasi, Universitas Bakrie

Article Information

Submitted : 10 Feb 2024
Reviewed: 19 Feb 2024
Accepted : 25 Feb 2024

Keywords

Data Warehouse, ETL
DataStage, DB2, Oracle

Abstract

Banking performance must always be maintained in providing fast and accurate services based on data processing results. PT Bank ABC has several data warehouse from various sources. The technology used makes the reporting process more complex, and maintenance costs are high. This research presents the data warehouse system transformation process using a single source of truth approach and the use of Extract, Transform, and Load (ETL) tools from IBM InfoSphere DataStage version 11.5. ETL testing is carried out on the Performance Measurement System (PMS) process by business units to monitor, identify, and analyze performance in each business unit. The results obtained proved to be faster than DB2 and Oracle database queries. It is depicted on technical, testing, and validation flowchart. Using the DataStage platform provides advantages in time efficiency, which is a crucial factor in overall data processing. In addition, DataStage Integration can have a positive impact on overall system performance, support smooth operations, and minimize the potential for database server overload.

A. Introduction

The amount of bank data is growing exponentially due to developments in information technology (IT). This size of data, which is impossible to manage in traditional databases, can help in the decision-making process. Data warehouse (DWH) and data mining (DM)-based systems are increasingly developing rapidly and are used in the management decision-making process. But traditional DWH measures make queries very complex, which can cause delays for decision support [1]. Several industries have adopted technological transformation in the digital era. One of them is PT ABC Banking. Bank ABC is a leading state financial and banking institution in Indonesia. Since carrying out the transformation in the merger process with another bank, Bank ABC has faced significant challenges related to data management efficiency due to the use of various applications and features in the data warehouse (DWH) system staging environment, resulting in delays in data processing. Apart from that, it also hampers the decision-making process and has the potential to reduce service quality.

In the context of data staging, this process becomes crucial for DWH processing as a storage area with special functions such as cleaning, changing, combining, duplicating, and preparing data sources. Problems faced by DWH practitioners include big data, changing business requirements, performance problems, complexity, auditability, checkpoint restarts, and team member fluctuations [2]. According to [3] the main stages in this process are extraction, transformation, and loading (ETL), and the advantages of DWH and ETL are: 1) providing one reliable data source for the business; 2) providing accurate, relevant, and timely information to make effective business decisions; 3) architecting scalable and scalable DWH solutions across the enterprise; 4) identifying and solving data quality and data cleansing issues. However, companies are increasingly interested in unstructured data access and its integration with structured data. Unstructured data is a significant challenge for scientists because it often requires a lot of time to organize and prepare a challenge. For example, in this process, striking a balance between writing data to stage tables and storing it in memory during the ETL process is key to creating an optimal process [4].

In an effort to overcome these challenges, research was carried out by designing a transformation model and implementing DWH. The leading ETL platform, namely IBM InfoSphere DataStage, is the main target. The platform leverages a high-performance parallel framework, is available both on-premises and in the cloud, and is scalable. Additionally, its key features include extended metadata management and enterprise connectivity, as well as supporting heterogeneous data integration. In the context of data storage, DB2, AS/400, and Oracle are used. The DB2 family of database management systems manages a wide range of hardware platforms [5]. Meanwhile, Oracle, as a relational database management system (RDBMS), is recognized as very efficient in managing information in a comprehensive, integrated, and open manner [6]. The ETL process is the key to the process of collecting, filtering, processing, and combining relevant data from various sources to be collected and stored in a data warehouse. The results of the ETL process are data with criteria including integrated, historical, uniform, summarized, static, and a structure in accordance with the needs of the work unit analysis process in the organization [6] [7].

A data warehouse is a collection of data with the characteristics of being integrated, subject-oriented, time-variant, and non-volatile [8] [9] [10]. Processed data can support the decision-making process for management. The Online Analytical Processing (OLAP) method is a merging process based on dynamic analysis and consolidation for large multidimensional data so that it can support interactive examination and manipulate data from various perspectives [1]. Several DWH studies have been carried out with different approaches in design and implementation, from specific aspects such as the use of the Nine Step Design Methodology to application for multidimensional analysis and business intelligence [11] [12] [13].

The problem with Bank ABC is that there is a need for data processing for integration, scalability, upgrading, and updating the DWH system. Therefore, research was carried out to design a transformation model and test its effectiveness on DWH technology. This process is carried out in a single staging environment so that data management becomes integrated, processing is efficient, and the decision-making support process is fast and accurate. Apart from that, it can also increase customer satisfaction [3] [14]. Therefore, improving the DWH consolidation process with a Performance Measurement System (PMS) can have a positive impact on banking operations and customer service [10] [15] [16].

B. Research Method

This research begins with the stages of problem analysis and identification, determining research objectives, preparing raw materials in the form of the old ETL process, working on the ETL process, testing and validation, and reporting (Figure 1). The improvement process approach is carried out using the convert method. The conversion mechanism for a PMS job is to switch the source that previously used AS400 to an Oracle database, as shown in Figures 2 and 3.

The process of completing the DWH design to increase the efficiency of the PMS process (Figure 3) begins with a pre-assessment, namely. At this stage, a list of ETL jobs that use AS400 is created, along with a list of tables sourced from AS400 in the PMS process. This information will be the basis for the next steps in changing connections to the PMS ETL job, especially in importing data from the AS400 to the systems reference code (SRC) scheme in the PMS database.

The next step is to backup all existing jobs on the DataStage v11.5 server. Backup was also carried out on all open database connectivity (ODBC), TNS, and DB2 catalogs on the DataStage 11.5 server. This step is taken to ensure data security and integrity before making significant connection changes to the system.

After the backup is complete, connection changes to the ETL PMS job are continued using the transparent network substrate (TNS), which is already available in the database configuration. This stage is very important to ensure the smooth operation of all ETL jobs, including verification regarding whether the output results are in accordance with the previous method and system environment. The implementation of this change is intended so that the system can run more efficiently and in accordance with the latest needs of the organization.

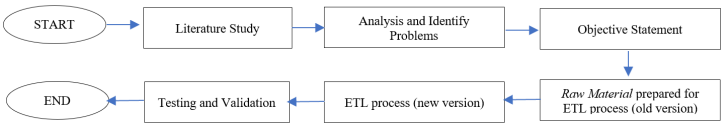

Figure 1. Research Stages

Figure 2. Short process for updating ETL jobs

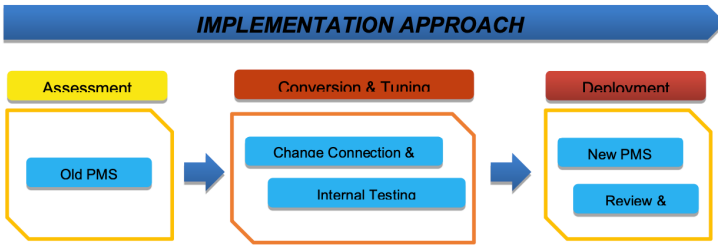

Figure 3. Implementation of a single source of truth

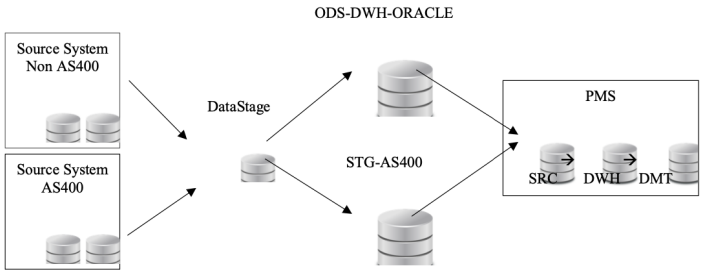

Figure 4. Existing Condition

Currently, the operational system describes the running conditions as shown in Figure 4. There are two main system sources, namely non-AS400 systems and AS400 systems, which contribute data to the PMS process. The process of withdrawing data from non-AS400 systems is carried out through the ETL mechanism, while data originating from text files is retrieved using the secure file transfer protocol (SFTP) mechanism (G et al., 2020).

On the other hand, retrieving data from the AS400 system is divided into two methods. First, the data is retrieved and synchronized into Oracle's operational data store-data warehouse (ODS-DWH) via the ETL mechanism. Second, there is a special process using the VTL (Virtual Tape Library) mechanism (Yan et al., 2015) (Iliadis, Jordan, Lantz, & Sarafijanovic, 2022), namely that data from the AS400 system is processed by backing up and restoring.

Furthermore, in the context of data processing, it can be seen that the process of pulling and processing data from ODS-DWH Oracle to Data Mart generally uses the DataStage platform. Likewise with the process of withdrawing and processing data from the STG-AS400 to Data Mart, which also relies on the DataStage platform. Figure 6 provides a visual overview of the architecture and current

process flow. The use of DataStage as a platform for data processing is characteristic of this process, demonstrating good integration in managing and transforming data from various sources for PMS needs.

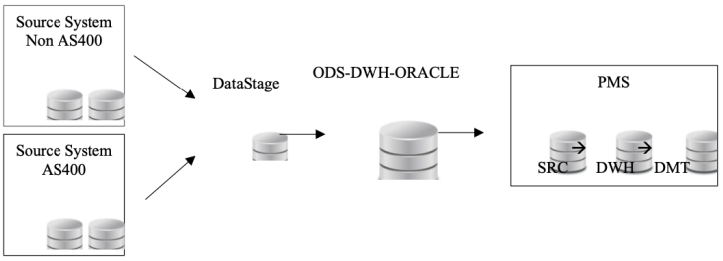

Figure 5. Proposed Conditions

The PMS process includes three main stages. The first stage is to withdraw data from a non-AS400 system using the ETL mechanism. This step ensures efficient integration of data into the PMS through data extraction, transformation, and movement. The second stage involves pulling data from the AS400 system to ODS-DWH Oracle via an ETL mechanism, ensuring synchronization and integration of data from both systems. The final stage is pulling and processing data from ODS-DWH Oracle to Data Mart using the DataStage platform. This process is illustrated in Figure 5, providing a visual representation of the proposed system architecture and workflow. Technically, this process is described in a flowchart and also for the testing and validation process (Figure 6).

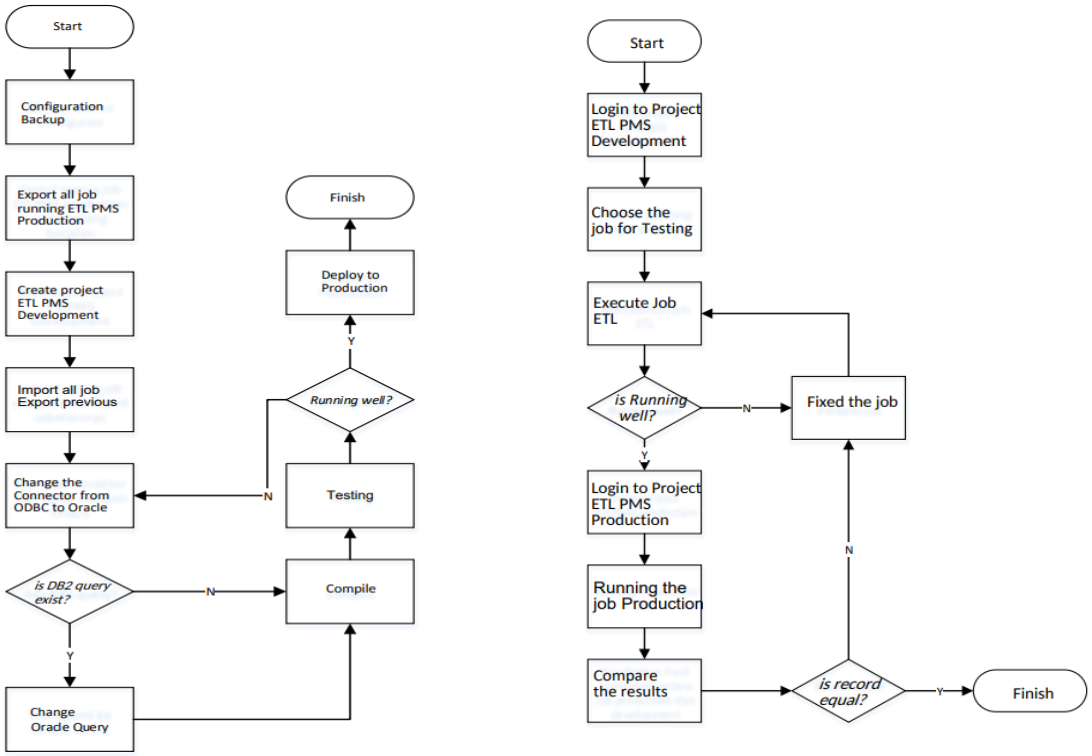

Figure 6. Technical, Testing and Validation

C. Result and Discussion

Data Warehouse Update. The first stage in the ETL process is data extraction from the newly defined ODS-DWH-Oracle database. In this context, it is important to create a database structure that suits your needs. To simplify integration, a schema mapping was carried out from DB2 to Oracle, and the relevant schema changes can be seen in Table 1.

Table 1. Schema Changes

No	**Staging AS400**	***ODS-DWH***
1	DMSTGPDAY.CDDHIS	BMSTGCBS.CDDHIS
2	BMMASSINA. CDMAST	BMSTGCBS.CDMAST
3	BMMASSIND. CDPAR2	BMSTGCBS. CDPAR2
4	BMCNUSDD. CFMAST	BMSTGCBS. CFMAST
5	BMHKUSDD. DDMAST	BMSTGCBS. DDMAST
6	BMSTGPDAY.DDDHIS	BMSTGCBS.DDDHIS
7	BMSTGPDWH.BIPARA	BMSTGCBS.BIPARA

The next step is the transformation stage (Figure 7). There are two approaches used in this process. First, the transformation was performed using the DataStage ETL tool, and second, using queries to handle differences in language syntax between DB2 and Oracle.

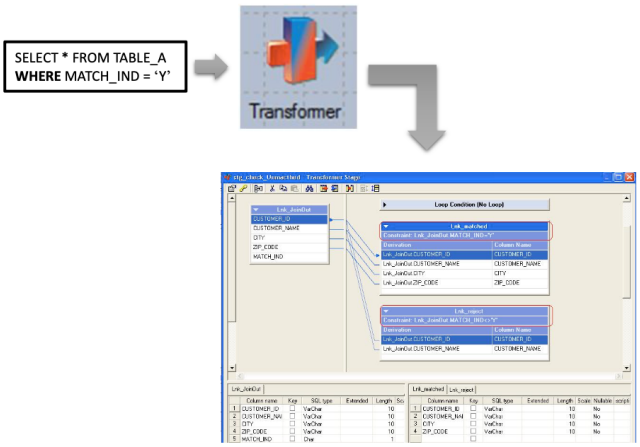

Figure 7. Transformation process

The final stage of the ETL process is loading, i.e., the data that has been extracted and transformed is placed in DWH as the final destination. This data can then be used by PMS, including for reporting purposes to Bank Indonesia. The ETL process involves the steps of data extraction from sources, transformation to maintain data consistency and clarity, and data placement in DWH.

First, the extraction stage in the ETL PMS process focuses on extracting data from the old database (DB2) to the new database (Oracle) so that it is integrated in ODS-DWH-Oracle. This process requires detailed mapping of schema changes to ensure accurate data transitions. Table 1 provides an overview of the schema changes from the AS400 database to Oracle and also provides an important basis for creating more efficient and tailored ETL jobs.

Second, in the transformation stage, two approaches are used: 1) direct transformation is carried out in the DataStage ETL tool, and 2) there is a change in language syntax between DB2 and Oracle. This transformation process, illustrated in Figure 8, is a crucial step to ensure the extracted data is converted to the correct format before being loaded into the DWH.

The final stage is loading, i.e., the data that has been transformed is placed in DDWH. This step is very important because it ensures that the data is available for use by the Performance Measurement System (PMS) for reporting purposes, including reports that must be submitted to Bank Indonesia. Thus, the entire ETL process carries out a series of integrated stages, from extraction to data loading, to effectively support the functions and needs of the PMS.

Next, for the work stage, the first step taken in implementing changes to the ETL system is Backup Job Existing. This stage involves creating backup copies of all existing ETL jobs on the DataStage v11.5 server and ODBC, TNS, and DB2 catalog configurations. The existence of this backup copy is crucial before moving on to changes, as a start in the backup process (Figure 8).

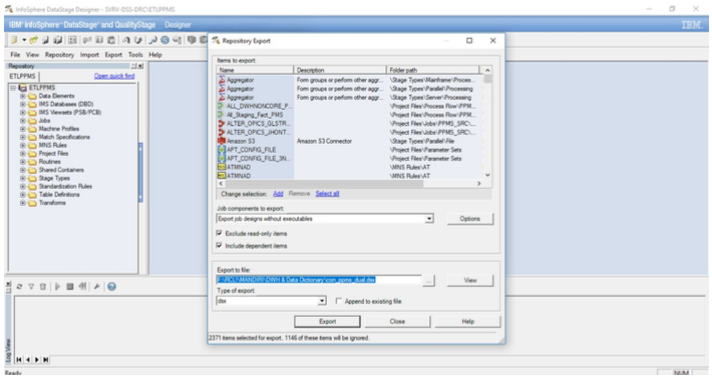

Figure 8. Step 1 Backup Job

The logic and configuration stage focuses on aspects of connections to source and target systems, including database and file connections. A detailed explanation is provided regarding the use of ODBC by IBM DataStage to connect to SQL Server, DB2, and various other databases. This information is an important foundation for ETL jobs to be connected correctly to the desired data sources, according to the source documentation.

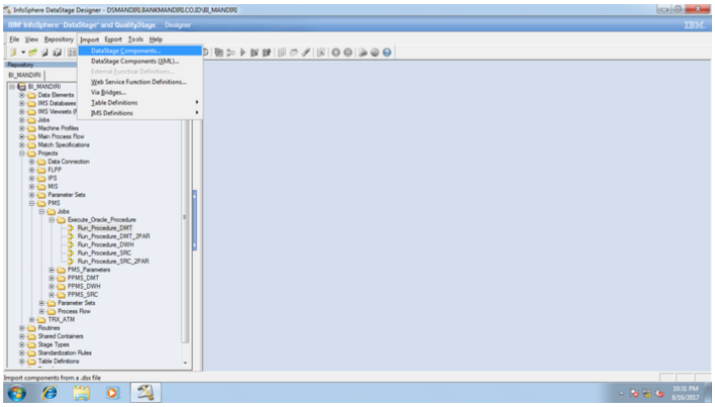

Figure 9. Step 1 Import Job

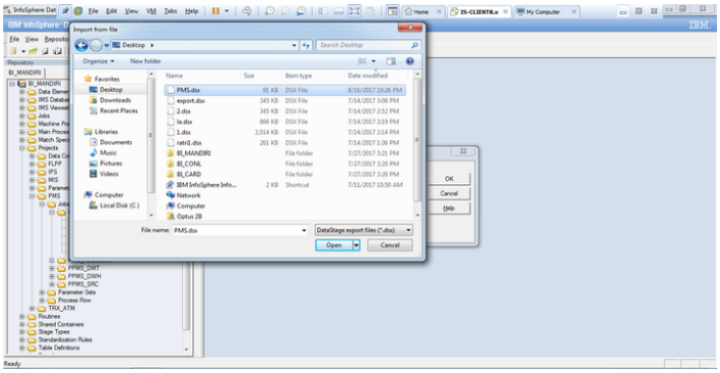

Figure 10. Step 2 Import Job

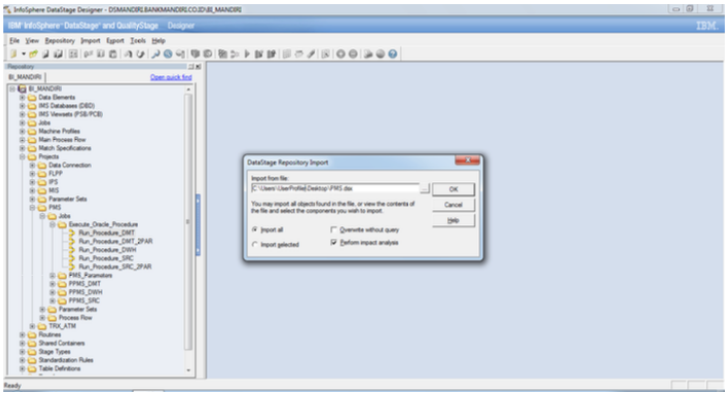

Figure 11. Step 3 Import Job

The ETL job testing process is a critical step before full implementation into the production environment. Steps in the job import process that have been tested (Figures 9, 10, and 11). Verifying the quality and success of previous ETL jobs executed in a production environment is essential to avoid potential problems.

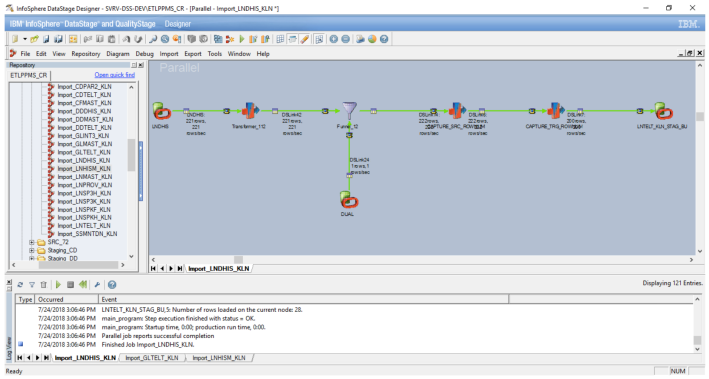

Figure 12. Running an ETL Job

After the testing process is complete, the Running Job ETL stage is carried out to confirm the results of the ETL job, which has been successfully executed (Figure 12). This stage is important to ensure that the converted ETL job runs successfully and produces output according to expectations. Finally, the Data Validation Test stage involves examining the output of each ETL stage, including sums (SUM) and calculations (COUNT) on specific data. Success criteria are determined based on data validation that has been determined by Bank ABC. This stage confirms that

the data produced by the ETL system not only works well but is also accurate and consistent with established standards.

DataStage Director Client. The features in this software are used to run, set validation schedules, and monitor ETL jobs. The ability to run and monitor ETL jobs effectively is critical to ensuring the integrity and success of the entire data warehouse update process. Figure 13 shows the user interface of this tool.

DataStage Operations Console. This feature provides visualization related to monitoring job activity, system resources, and workload management queues. The steps for using it are important to ensure that all ETL jobs run smoothly and efficiently. Figure 13 shows the interface of the DataStage Operations Console.

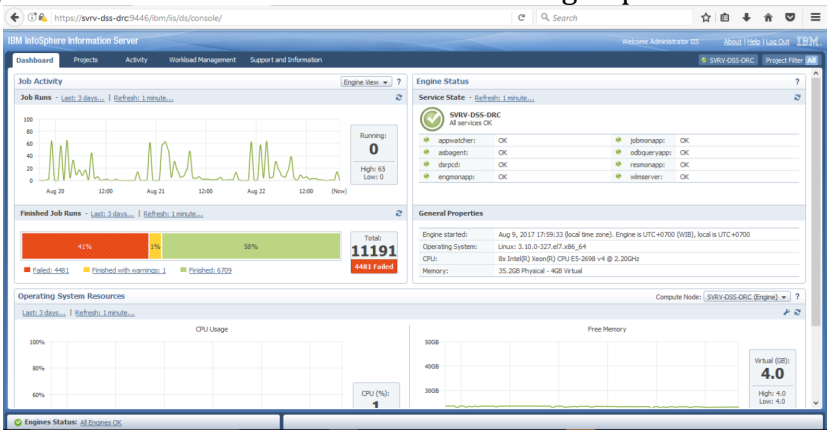

Figure 13. Testing Tool

Test Running. This stage tests the effectiveness of using the IBM DataStage ETL Tools with example queries. Figure 14 shows the query used for testing. In Table 2, we present the results of the effectiveness of IBM DataStage in various environments (DB2, Oracle, and running via DataStage) to evaluate the performance and effectiveness of the tool in different environments. The results of the analysis show that the DWH renewal process was carried out effectively. The changes implemented, including database migration and conversion of ETL jobs from DB2/AS400 to Oracle, have proven effective and efficient. This has been proven through various stages of testing and validation that have been carried out.

```
CREATE PROCEDURE BMDWHAML.PDCUSEML ()
    LANGUAGE SQL
    DYNAMIC RESULT SETS 0
    MODIFIES SQL DATA
    DISALLOW DEBUG MODE
    OLD SAVEPOINT LEVEL
    COMMIT ON RETURN NO
BEGIN
    CALL BMDWHAML . PCRTVIEW ;
DELETE
    FROM BMDWHAML . TCUSEML ;
INSERT INTO BMDWHAML . TCUSEML (
    CIFNO
    , EMAIL
    , EMRCDNO
    )
SELECT CIFNO
    , EMAIL
    , EMRCDNO
    FROM (
SELECT CC . *
    , ROW_NUMBER ( ) OVER (
        PARTITION BY CIFNO
        , EMAIL
        ) AS SEQ2
FROM (
    SELECT CCO . CFCIF AS CIFNO
        , REPLACE ( TRIM ( TRANSLATE ( CCO . CFEADD , ' ' , '!#$%^&*()_-=+/\{}[];:,<>?' )
        ) , ' ' , '' ) AS EMAIL
        , COALESCE ( CCO . SEQ , 0 ) AS EMRCDNO
    FROM (
        SELECT *
        FROM (
            SELECT CFCIF
                , ACCTNO
                , SRCSYS
                , ROW_NUMBER ( ) OVER ( PARTITION BY CFCIF ) AS SEQ
            FROM BMDWHAML . DDWHSRC DWH
            WHERE DWH . SRCSYS IN ( 'CUSEML' )
            ) X
        ) X
    WHERE SEQ = 1
    ) DWH
  INNER JOIN (
    SELECT CFCIF
        , REPLACE ( REPLACE ( TRIM ( CFEADD ) ,
        CHR ( 13 ) , '' ) , CHR ( 10 ) , '' ) AS CFEADD
        , SEQ
    FROM BMDWHAML . CFCONN2
    WHERE CFCIF > 0
        AND CFEADC = 'EM'
    ) CCO ON CCO . CFCIF = DWH . CFCIF
  ) CC
INNER JOIN BMDWHAML . CFMAST2 CFT ON CC . CIFNO = CFT . CFCIF
WHERE CC . EMAIL <> ''
) X WHERE SEQ2 = 1 ;
END
GO
```

Figure 14. Query testing

Table 2. IBM DataStage Effectiveness Results

No.	Running place	*Running Start*	*Finished*	*Total*
1	DB2 Query	10:15:19	10:15:40	21 sec
2	Oracle	10:17:23	10:17:59	36 sec
3	Datastage	10:20:29	10:20:40	11 sec

From Table 2 and Fig. 13, we conducted a test running (using query Fig. 14) data source with 11191 records. It finished with 6709 records correctly and 1 of warning data. The time processing was calculated from running start until finished. The best time is covered by processing DataStage with 11 seconds, DB2 Query with 21 seconds, and Oracle with 36 seconds.

D. Conclusion

The process of updating the Data Warehouse configuration at Bank ABC succeeded in changing the initial configuration involving two sources, namely DB2 and ODS-DWH-Oracle, into a single source of truth by only utilizing the ODS-DWH-Oracle database. This clarifies and simplifies the system configuration structure, increasing clarity and efficiency in data management. Proving the effectiveness of the ETL process using DataStage is important in this update. The time required to run the ETL process is proven to be faster compared to DB2 and Oracle database queries. This conclusion shows that using the DataStage platform provides an advantage in time efficiency of 11 seconds compared to others of 36 and 21

seconds, which is a crucial factor in overall data processing. The IBM DataStage Server role helps optimize the workload on the Bank ABC database. By dividing the workload on the DataStage Server, database performance optimization is achieved, providing better support for data management and consolidation. Thus, this confirms that DataStage integration in a data warehouse environment can have a positive impact on overall system performance, support smooth operations, and minimize the potential for overload on the database server.

The upgrading and updating process of ODS-DWH Oracle with the implementation of Single Environment Staging. opens up the potential for developing better management both technically and conceptually in ETL. First, the size of the DataStage server needs to be adjusted to the workload in order to avoid the potential for server downtime. This is an important key to maintaining smooth operations and system availability. Further development of server infrastructure will support system scalability according to needs that may develop in the future. Furthermore, applying parameterization to ETL jobs is an aspect that can be optimized. This can reduce repetition in the same work process, increase efficiency, and minimize the potential for errors. By using parameterization, ETL jobs can be managed more flexibly and easily adjusted, providing the adaptability needed to address changing needs or specific scenarios. This improvement will support the overall effectiveness and efficiency of the data warehouse system at Bank ABC.

E. Acknowledgment

The author would like to thank to Universitas Bakrie for the support during the research reported in this paper.

F. References

[1] J. Luo, J. Xu, O. Aldosari, S. A. Althubiti, and W. Deebani, "Design and Implementation of an Efficient Electronic Bank Management Information System Based Data Warehouse and Data Mining Processing," *Inf. Process. Manag.*, vol. 59, no. 6, p. 103086, 2022, doi: https://doi.org/10.1016/j.ipm.2022.103086.

[2] D. Linstedt and M. Olschimke, "Chapter 1 - Introduction to Data Warehousing," in *Building a Scalable Data Warehouse with Data Vault 2.0*, D. Linstedt and M. Olschimke, Eds. Boston: Morgan Kaufmann, 2016, pp. 1–15. doi: https://doi.org/10.1016/B978-0-12-802510-9.00001-5.

[3] A. Khan, N. Ehsan, E. Mirza, and S. Z. Sarwar, "Integration between Customer Relationship Management (CRM) and Data Warehousing," *Procedia Technol.*, vol. 1, pp. 239–249, 2012, doi: 10.1016/j.protcy.2012.02.050.

[4] S. Bouaziz, A. Nabli, and F. Gargouri, "Design a data warehouse schema from document-oriented database," *Procedia Comput. Sci.*, vol. 159, pp. 221–230, 2019, doi: 10.1016/j.procs.2019.09.177.

[5] R. Hughes, "Chapter 15 - Fully Agile EDW with Hyper Generalization," in *Agile Data Warehousing for the Enterprise*, R. Hughes, Ed. Boston: Morgan Kaufmann, 2016, pp. 375–420. doi: https://doi.org/10.1016/B978-0-12-396464-9.00015-1.

[6] G. Garani, D. Tolis, and I. K. Savvas, "A trajectory data warehouse solution for

workforce management decision-making," *Data Sci. Manag.*, vol. 6, no. March, pp. 88–97, 2023, doi: https://doi.org/10.1016/j.dsm.2023.03.002.

[7] Y. K. Meena and D. Gopalani, "Evolutionary algorithms for extractive automatic text summarization," *Procedia Comput. Sci.*, vol. 48, no. C, pp. 244–249, 2015, doi: 10.1016/j.procs.2015.04.177.

[8] R. Hughes, "Essential DW/BI Background and Definitions," *Agil. Data Warehous. Enterp.*, pp. 59–84, 2016, doi: 10.1016/b978-0-12-396464-9.00004-7.

[9] A. K. Agarwal and N. Badal, "A novel approach for intelligent distribution of data warehouses," *Egypt. Informatics J.*, vol. 17, no. 2, pp. 147–159, 2016, doi: 10.1016/j.eij.2015.10.002.

[10] V. Guerola-Navarro, R. Oltra-Badenes, H. Gil-Gomez, and A. Iturricha Fernández, "Customer relationship management (CRM) and Innovation: A qualitative comparative analysis (QCA) in the search for improvements on the firm performance in winery sector," *Technol. Forecast. Soc. Change*, vol. 169, 2021, doi: 10.1016/j.techfore.2021.120838.

[11] M. Akbar and Y. Rahmanto, "Desain Data Warehouse Penjualan Menggunakan Nine Step Methodology Untuk Business Intelegency Pada Pt Bangun Mitra Makmur," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 1, no. 2, pp. 137–146, 2020, doi: 10.33365/jatika.v1i2.331.

[12] T. R. Priono, W. Purnomo, and N. Y. Setiawan, "Pengembangan Data Warehouse menggunakan Metode Kimball (Studi Kasus : Ekspor & Impor Fauna dan Flora Hias Air Laut)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 8, p. 1, 2021.

[13] D. J. C. Sihombing, "Academic Data Warehouse Modeling in Higher Education Using Nine-Step Design Methodology," *J. Inf. Syst. Informatics*, vol. 4, no. 4, pp. 1126–1134, 2022, doi: 10.51519/journalisi.v4i4.399.

[14] F. A. Khan, A. Ahmad, M. Imran, M. Alharbi, Mujeeb-ur-rehman, and B. Jan, "Efficient data access and performance improvement model for virtual data warehouse," *Sustain. Cities Soc.*, vol. 35, pp. 232–240, 2017, doi: 10.1016/j.scs.2017.08.003.

[15] M. Kubina and V. Lendel, "Successful Application of Social CRM in The Company," *Procedia Econ. Financ.*, vol. 23, no. October 2014, pp. 1190–1194, 2015, doi: 10.1016/s2212-5671(15)00487-6.

[16] S. Yasiukovich and M. Haddara, "Social CRM in SMEs: A systematic literature review," *Procedia Comput. Sci.*, vol. 181, pp. 535–544, 2021, doi: 10.1016/j.procs.2021.01.200.