International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://theaspd.com/index.php

Effective Distributed Based Decision-Making Approach Using ETL Data Warehousing Based On Smart Business Intelligence Technology

G. Jagan Naik[1], Dr. Sunita M[2], Vankudoth Ramesh[3], Jayaram Boga[4], A.Vijendar[5]

[1]Associate Professor, Department of CSE (DS), CMR Institute of Technology, Hyderabad-501401, India

[2]Associate Professor, Computer science and Engineering (Data Science), Marri Laxman Reddy Institute of technology and Management, Hyderabad- 500043, India

[3]Associate Professor, Department of CSE-DS, CVR College of engineering, Hyderabad- 501510, India

[4]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Hyderabad-500043, India

[5]Associate Professor, Department of CSE (AI&ML), CMR Engineering College, Hyderabad-501401, India

*Corresponding Author: A.Vijendar, E-mail: amgothu.vijendar@gmail.com

Abstract

In advancement of recent era, business organization is developing drastically as there is an increase number of information Technology (IT) as they give wide impact on the national and international development as they all managing the vast and complex amount of data. In order to analyses the performance of data, those vast data have to be processed and analysed. In managing the data, Extract, Transform and Load (ETL) process is applied and stored in the data warehouse as repository in order to take effective distributed based decision as it faces the problem of time-consuming process. As the data warehouse takes the data source in the distributed manner as it is difficult to integrate those data.

To overcome the above challenges, the Modified ETL based Data warehouse is proposed as it is applied with modified multi-dimensional based bottom-up approach as it carried out various activities to deeply analysis the data based on content profiling. In the process of cleaning, confirming and delivery of data depends on the various data sources. Then algorithm makes three various processes of data extraction as it provides data cleaning and conforming to create the conform steps based on the source data analysis using data hierarchy structure. Until the data sources gets integrated based on the various distributed database, ETL steps are performed. In the analysis, modified ETL is applied on any real time organization as it takes various source table to compare the actual and expected results. Then various metadata testing is performed on various documentation to makes the process of transformation effective.

Keywords: *ETL process, Data extraction, IT organization, Transformation, data warehouse, distributed data.*

1. INTRODCUTION

In the CoVID-19 pandemic era, it makes the target on various organization to surviving as this IT business is growing drastically. The business activity makes various complex survival as it cannot be detached from the IT as it plays a key role in their organizational development [1]. Even though IT organization is growing drastically, it solely depends on the business activity as it gives greater impact. Here big data plays a vital role as the business activity has generated and collected vast amount of data as it needs to be processed and analysed. This big data helps to generate and collect the real time data and the business process is used to perform the large data and determine the resource efficiency, which can be improved and it helps to evaluate the business activity in terms of marketing [2]. The process of big data helps to make the business to run better in terms of optimizing the architecture of data warehouse as it makes the data to more accurate and secure.

Here the data warehouse is the storage repository as it integrates the process of data to be more knowledge as it helps to access the business information and make better decision. This helps to construct the new data warehouse as it integrates the data to be manipulated using Extract Transform Loading (ETL) process [3]. This process of ETL makes the raw data to be analysed and extracted into the storage of data warehouse. This ETL process makes the data to be manipulated from various sources, which includes the improper data elimination, filtering the duplication, data filtering, data cleaning and formatting. The

data's considered will be of various formats. The ETL is defined as the process sets as it gets those data from various Online Transaction Processing (OLTP) into data warehouse repository. In ETL initial steps, final data being collected from various heterogeneous data sources as it requires quality data and reliable one. This traditional ETL process requires more execution time as more developed project requires 80% time increase in execution [4]. The data warehouse helps to store the data records as it makes the effective decision process based on the business process in Distributed environment.

Mainly the data warehouse takes the system to perform data extract, data cleaning and deliver to data dimensional from various data sources and able to determine the effective decision making. Then makes data warehouse operates on the source data, presenting the data which depends on various data format [5]. In this research some datasets are considered as it takes academic information along with the faculties and the department. Based on the various departments and faculties who are working will gives which kind of services are offered in the university or institution? Generally, data is arranged and collected in the distributed fashion and initiate the system information [6] and [7]. Then the distributed database is deployed on single or multiple source system.

The advantage of data configuration is to represents the structure organization, autonomy the improved data and their performance; the limitations faced during the configuration is the data standard, complexity and data integrity. The data warehouse process using ETL process is represented in figure 1.

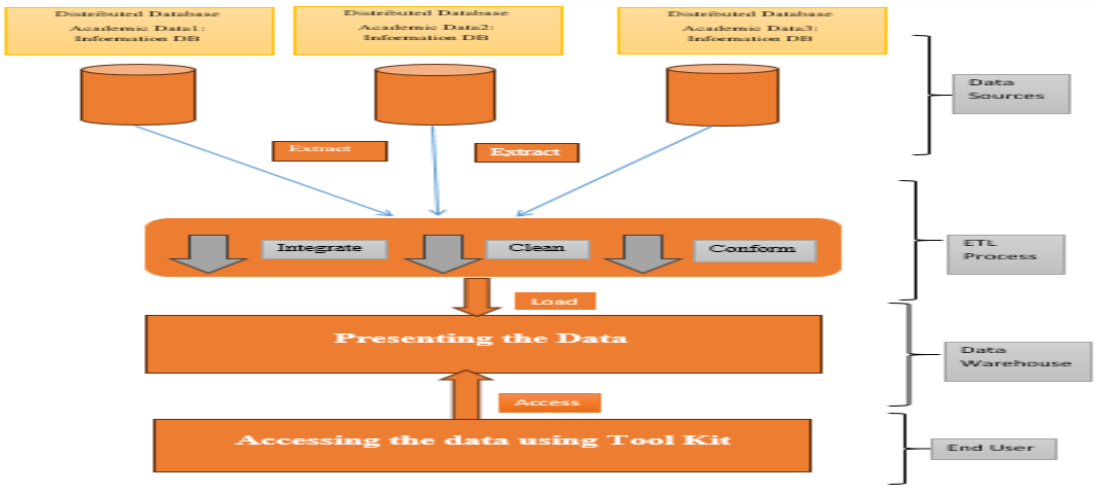

Figure 1: Data warehouse Process using ETL

In this paper, based on the distributed database, the standard ETL process is more complicated as it does not contain a central mechanism for each database. Based on the data fact of the table, structuring the data help to clean the data with various dimension and then construct the data warehouse based on a certain process. the objective is organized as it gets initiated with the modified multi-dimensional based bottom-up approach as it carried out various activities to deeply analyse the data based on content profiling. The business requirements are provided based on the multi-dimensional representation of the data table. Initially, the particular business process is selected, which is to be modelled and represented as a table and do certain measurements. Data grain is made to represent the individual representation. Then the data dimension is chosen based on the set of data attributes and identify the numeric facts.

2. LITERATURE SURVEY

Assessing the performance can be achieved based on the organization's achievement with certain tasks and activities. The organization's achievement and performance can be calculated and analyzed to represent the operational impact, which may be positive or negative [8]. Here comes the role of big data technology as it plays a challenging role while the top companies like Google, Facebook, etc. consider the unstructured data as it needs to be managed and utilized based on vast amount of large data [9].

This big data takes the major role in the field of information technology as it makes the process of managing and analysing the data. As the large volume of data to be process will be much more complex with the standard technologies, big data takes the process of managing and analysing those data [10 - 14]. The information technology purely depends on data utilization and managing them as the entire

organization takes the lead as it dependence on the data. Those organization takes the role of organizing the tool to managing the data through application and their features are represented in Figure 2.

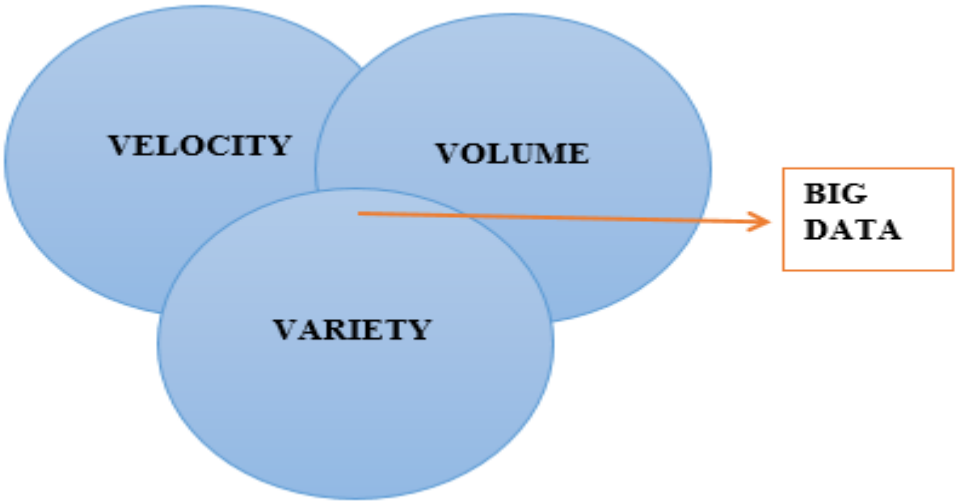

Figure 2. Features of Big Data

Based on the business organization, data warehouse plays a significant role as it makes the various processes, technologies and tools are performed along with the data as it gathers certain knowledge and useful data as it results in the profitable information based on certain organization activities. In order to utilize the emerging opportunities, there are various integration of data warehouse, mining data, analysing the data in multi dimension with certain graphical representation [15]. Then these data takes certain analyses to determine the growth and enhancement of the organization. Here the business intelligence includes, data warehouse, data mining, analysis the multi dimension, etc. In this data warehouse technology, data can be managed from various multiple sources as it gives improved insights on the business. It is the repository to store large volumes of information especially as the data gets generated by the business organization. In this data are processed, transformed, and absorbed by the users based on the tools being used by the business organization.

The ETL process is being studied based on the data warehouse on sales as it is represented using pentaho tools [16]. In this study, the sales in the data warehouse are processed to form useful information as it gets analysed and processed to determine the performance of the sales to make effective decision making. [17] has analyzed the data store of the United States from 2014-2017 as it takes the representation of Microsoft office doc as it transforms the data into MYSQL database. The data warehouse is analyzed on certain applications of PHI Minimart using pentaho tools to analyze the data warehouse sales with the help of an ETL system with information sales.

The data warehouse is analyzed using OLAP process of data creation as schema. The data sales of minimart sales 2008 [18] as it is represented as sales chart and their total sales for each department in the organization [19][20] has consider the online market datasets and based on that the data warehouse is designed and it is represented as the schema of snowflake schema. The data modeling of this snakeflake schema takes various 9 steps of methodology to model the data [21]. The data platform is used to process the data in order to represent the data, which is informative are useful. Business intelligence is applied to determine the effectiveness of decision-making and improved data quality.

3. problem statement

Based on the performance of ETL Process, [4] has proposed the ETL process with three processes of data changer, cleaning the data and load the data to determine the data robustness. In this work, segment the data and modification are made, which can handle the data to be represented as the data tables among the various users but those tables are not associated with each other users. Then ETL performance is further modified based on query cache technique as it helps to reduce the response time. Then [5] has modelled the ETL based data warehouse as the managing the organization, data warehouse and users acts an intermediate as here we have considered the dataset of clinical based data warehouse. [6] has consider the development of data warehouse with respect to distributed environment. [7] has discussed various

ETL process and their approaches being used in order to determine the effective ETL techniques being selected. The challenges and open issues being addressed is the analyse of the data source in the distributed database and specific data case study has to be considered by deploying the modified ETL process in order to analyse the data being used.

4. RESEARCH METHODOLOGY

In the proposed approach, the multi-dimensional model is deployed as it helps to construct the data warehouse as the modelling process takes certain stages as mentioned below,

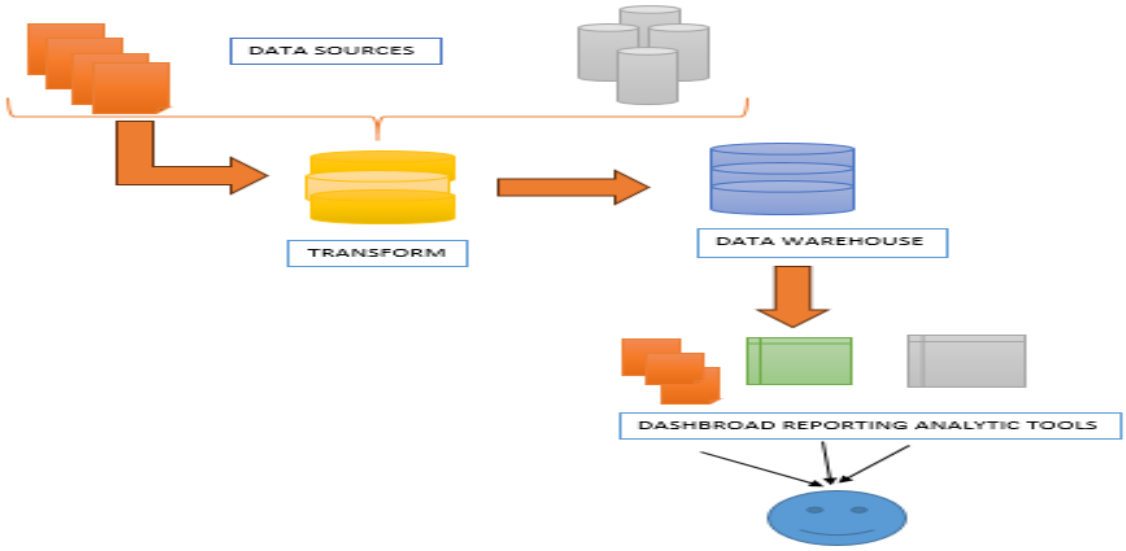

Figure 3: Data Warehouse Architecture using ETL Process

A. Multi-Dimensional Model:

The data are modelled based on the process of data warehouse with kimbell's approach as, Initial step helps to identify the business process as analyse the students entry, their registering for the new course, analyse the student grade obtained, student payment and their graduation count of students. Then data can be grained and declared based on the data measure as it is integrated with the data granularity. The data measurement is based on the student information with every year, information of the student who is registering, distribution of grade of the student, graduation data and payment done by the student. The data dimension is identified based on role associated with the student.

"Rajesh **(Gender)**, Semester, Adgami **(Religion)**, Mahesh **(Student)**, Mahesh Reg **(Student Registration)**, Model, ... Bank Name **(Name of the Bank)**, Delhi **(Payment)**, STAS **(High School)**,"
Finally, the facts are identified based data tables being deployed by the individual user as represented as,
Student Table | Student Count | Facts
Table Student Registration | Student Count
Grade table Student | CGPA | GPA
Payment Student Table | Payment Student | Status

B. Modified ETL Process:

In the ETL process, the server, mainframe, source production is taken as the input as it get the initial process of extraction as takes the unstructured data and perform extraction on the data stage and generalize the data as structured one. Then it performs data cleaning as it cleans the unwanted data into useful one. Then data conforms and deliver make the data to be analysed and processed. Finally it is sent back to the use application by the end user as it operates several processes such as, scheduling the data, handling the data exception, data recovery, and data restart, checking the data quality and reliability with certain user support.

Algorithm 1: Modified ETL Process for Data Integration

Input: Server; Mainframe; Data Source
Output: Data Integration

1. Identify the data source to perform data extraction
2. The data tables are structured and their explanation

- Using a merging strategy, create the master table

a. Table 1 → DB1
b. Table 1 → DB2
c. Merger (Table 1 from different DB1 and 2)

- Using Merge Union Strategy, Create the master table

a. Table 1 → DB1
b. Table 1 → DB2
c. Table 1 not influence on Table 2
d. Apply Merge strategy with Union Action

- Using Union Strategy, create a transaction table

a. Large table → Target Table
b. Target Table = {Pilot Table}

3. Heterogeneous Data Source
4. DB = DB {Merge and Union Strategy}
5. Perform Data identification and analyze the data source

a. hostname (Domain name or IP address of the database server)
b. Database name (The schema or other database identifiers)
c. Port Username and password to access the data source)

6. IFNULL () and NULL Values Expression
7. Creating the data dimension tables.

The process of data integration is performed and represented in Algorithm 1 and figurate in figure 4.

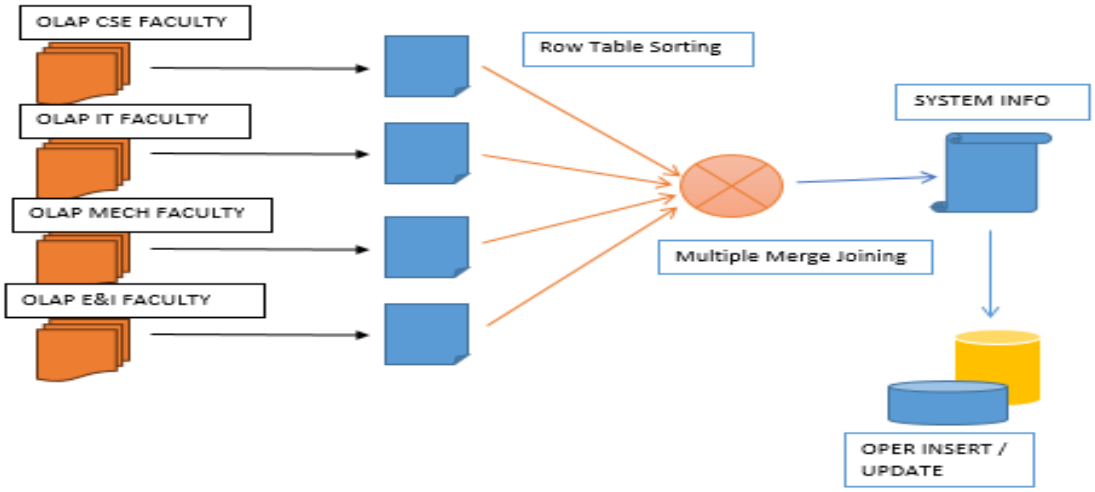

Figure 4: Data Integration Using ETL Process

C. Data Cleaning and Conforming:

In the cleaning process, the data errors are detected and removed as inconsistent data's are identified as it results in enhancing the data quality.

- Analysis the data
- Refining the data
- Verify the data

Based on the data rule and value, re-analyse the data based on ETL process to perform data confirmation, data join and association. Then the data helps to create conform data dimension as it takes data hierarchy and dimension and it is represented in Figure 5.

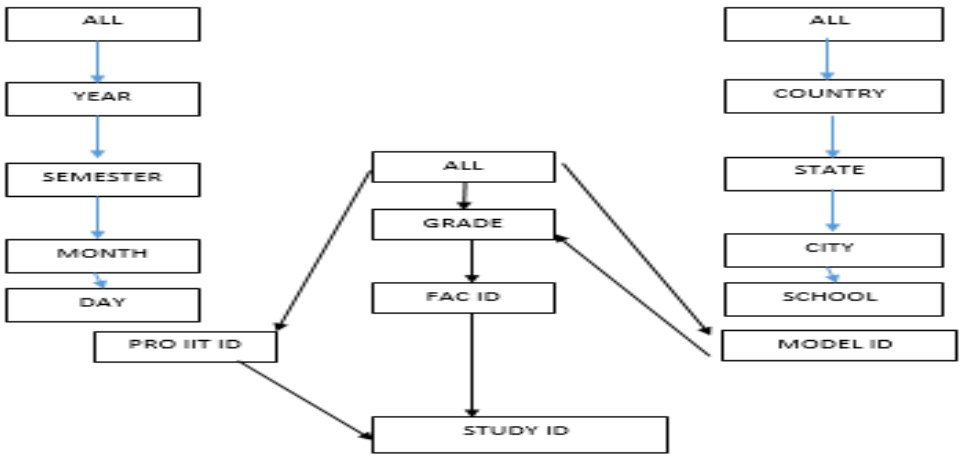

Figure 5: Data Conforming

D. Data Delivery and Loading:

In this data table, only user analysis data are present and for each data table, the key is generated. Here the start schema being used as it helps to de normalize the data and then deliver the data as the targeted data table.

- Slow Changing Dimension (SCD)
- Loading the Fact Table

E. ETL Process Testing:

The main objective of ETL process is to test the data by identifying and collecting the data errors as it occurs before the process of data analytical and reporting. Various data testing is performed i.e. validate the data completion, testing the Meta data and ETL test incrementing. All loaded data are measured as it helps to validate the objective if data completion. The table defined are verified while performing the Meta data testing. Data document can be mapped based on checking the data type, length of the data, checking the data index between the source and destination table. Then the unwanted data duplication can be determined based on the incremented ETL testing.

5. PERFORMANCE ANALYSIS

Based on ETL strategy, data warehouse makes the total cost construction as contains Data Dimension Cost (DDC) and Import Data Cost (IDC).

Total Cost (TC) = Data Dimension Cost (DDC) + Import Data Cost (IDC)

A. Analysis of the OLAP Process:

As there are two OLAP query requests, various data dimension is required based on value query of fact information as it gets H Base get( ) operation and Hbase Scan ( ) as represented in Table 1 and 2.

Table 1: Time cost based on Get ( ) operation

No. of Rows	HBase	Oracle
1000	4	22
10000	3.6	34
100000	2.8	45

Table 2: Time cost based on Scan ( ) operation

No. of Rows	HBase	Oracle
1000	20	25.4
10000	12.5	24.6
100000	19.2	26.6

B. Analysis of ETL process:

Here 1000, 10,000 and 1,00,000 data entries are considered as the HBase and traditional oracle are analysed as represented in Table 3, 4 and 5.

Table 3: ETL Process based on 1000 Datasets (ms)

No. of Rows	HBase	Oracle
1	104	34
2	109	16

Table 4: ETL Process based on 10,000 Datasets

3	100	15
4	116	12
5	109	13
6	111	25
7	116	12
8	128	13.5
9	120	13.8
10	98	14.2

No. of Rows	HBase	Oracle
1	650	190
2	600	96
3	625	60
4	680	45
5	800	48
6	790	54
7	765	51
8	600	53
9	605	170
10	790	68

Table 5: ETL Process based on 1,00,000 Datasets (ms)

No. of Rows	HBase	Oracle
1	2900	509
2	2850	450
3	2950	465
4	2980	445
5	2450	451
6	2800	458
7	2750	462
8	2950	710
9	3450	462
10	2800	491

6. CONCLUSION

Based on the data heterogeneity problem, ETL process is deployed as it performs certain process such as, data extraction, cleaning the data, data conforming and data delivery & loading. The data content profiling and analysing the data source are performed in the data extraction process.

The data are integrated in the distributed environment as it performs the database with different strategy i.e. merge, merge-union and union. The data conformation is created based on the analysing the data source, data refinement based on structure hierarchy. The data dimension and fact tables are associated in the process of ETL loading. In the analysis, Hbase data warehouse takes reduce time in the process of ETL than the traditional oracle. Ithe HBase takes faster query response by varying the datasets 1000, 10000, 100000 data entries. If the data value is larger, memory consumption is high and the query optimization gets reduced as the data value is smaller and it needs some better data partition in the data value.

REFERENCES

1. Loshin, David. Business Intelligence: The Savvy Manager's Guide. 2nd Editio, Elsevier Science,2012,https://www.google.co.id/books/edition/Business_Intelligence/L7SLNlS1

2. Qalam, Yance Ibnu. "Hubungan Data Warehouse Dengan Business Intel-ligence Dan ETL." Kepo.Co, 2020, https://kepo.co/hubungan-data-ware-house-dengan-business-intelligence-dan-etl/.

3. B. Ravi, Vijendar Amgothu," A Robust Noise Reduction Strategy in Magnetic Resonance Images", Annals of R.S.C.B.,ISSN:1583-6258, Vol. 25, Issue 6, 2021, Accepted 08 May 2021.

4. B. Ravi, Vijendar Amgothu," Feature Selection using Multi-Verse Optimization for Brain Tumour Classification", Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Accepted 08 May 2021.

5. Nordeen, Alex. Learn Data Warehousing in 24 Hours. Guru99, 2020, https://www.google.co.id/books/edition/Learn_Data_Warehousing_in_24_Hours/wgf9DwAAQBAJ? hl=id&gbpv=0.

6. Prasetia, I. Putu Widia, and I. Nyoman Hary Kurniawan. "Implementasi ETL (Extract, Transform, Load) Pada Data Warehouse Penjualan Menggunakan Tools Pentaho." TIERS Information Technology Journal, vol. 2, no. 1, 2021, pp. 39–47, doi:10.38043/tiers.v2i1.2844.

7. Udayana, Gede Acintia, I Made Yoga Mahendra, I Kadek Anom Suka-wirasa, Gde Deva Dimastawan Saputra, and Ida Bagus Made Mahendra. "Implementasi Data Warehouse Dan Penerapannya Pada PHI-Minimart Dengan Menggunakan Tools Pentaho

Dan Power BI." JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), vol. 10, no. 1, 2021, p. 163, doi:10.24843/jlk.2021.v10.i01.p19.
8. Marbun, Ivan Rivaldo, and Ramos Somya. "Perancangan Data Ware-house Untuk Data Transaksi Penjualan Menggunakan Schema Snowflake Studi Kasus : Online Market Dataset." Universitas Kristen Satya Wacana, vol. 5, no. 1, 2021, pp. 87–91.
9. Saeed K Rahimi, F.S Haug. Distributed Database Management System–A Practical Approach. New Jersey: John Wiley & Sons Inc. 2010:1.
10. M.T Ozsu, P. Valduriez. Principles of Distributed Database–Third Edition. New York: Springer. 2011
11. T.Connolly, C. Begg. Database System. A Practical Approach to Design, Implementation and Management. Fourth Edition. Essex: Pearson Education. 2005: 695.
12. Kimball, Ralph., Ross, Margy. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition. Indianapolis: John Wiley & Sons Inc. 2013:38.
13. Igor Mekterovic, Ljiljana Brkic, and Mirta Baranovic. Improving the ETL process of higher education information system data warehouse. Proceedings of the 9th WSEAS International Conference on Applied Informatics and Communications (AIC'09). Moscow.2009: 265-270.
14. Vishal Gour, S.S. Sarangdevot, G.S. Tanwar, A. Sharma. Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse. Int. Journal on Comp. Sci. and Eng. 2010; 2(3):786-789
15. Abid Ahmad, Muhammad Zubair. Using Distributed Database Technology to simplify the ETL Component of Data Warehouse. Proceedings of WSEAS International Conference on Applied Computer Science (ACS'10), Iwate. 2010; 61-65.
16. Tute E, Steiner J. Modeling of ETL-Processes and Processed Information in Clinical Data Warehousing. Stud Health Technol Inform. 2018; 248 204-211. PMID: 29726438.
17. Sonali Vyas & Pragya Vaishnav. A comparative study of various ETL process and their testing techniques in data warehouse, Journal of Statistics and Management Systems. 2017; 20(4): 753-763.
18. C. Adamson. Mastering Data Warehouse Aggregates. Solutions for Star Schema Performance. Indianapolis: Wiley Publishing Inc. 2006:20.
19. W.D Back, N. Goodman,J Hyde. Mondrian in Action. Open Source Business Analytics. New York: Manning Publications Co. 2014:195.
20. Meadows, A.S. Pulvirenti, M.C. Roldan. Pentaho Data Integration Cookbook. Birmingham: Packt Publishing, 2013:11.
21. Rahm, E., H. H. Do, Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 2000; 23(4): 313.
22. R. Bouman, J.V. Dongen. Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL. Indianapolis: Wiley Publishing, Inc. 2009:160.
23. G. Jagan Naik, V. PC Rao, A. Govardhan. title: The Data Warehouse Design Problem through a Schema Transformation Method, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.
24. Mr. G. Jagan Naik, Dr. P C Rao, Dr. A. Govardhan. FULLY AUTOMATED DATA WAREHOUSE FRAMEWORK USING ETL PROCESS FOR DECISION SUPPORT SYSTEM, Turkish Journal of Physiotherapy and Rehabilitation; 32(3), ISSN 2651-4451 | e-ISSN 2651-446X
25. V. Amgothu, P. R. Kumar, R. Boda and B. R. Naik, "Image compression using Adaptively Scanned Wavelet Difference Reduction Technique (ASWDRT)," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2017, pp. 1131-1134.
26. Ravi, B., et al. "A Robust Noise Reduction Strategy in Magnetic Resonance Images." Annals of the Romanian Society for Cell Biology 25.6 (2021): 3938-3952.