ISSN: 2708-7123 | Volume-02, Issue Number-03 | September-2021
LC INTERNATIONAL JOURNAL OF STEM
Web: www.lcjstem.com | DOI: https://doi.org/10.47150

IMPLEMENTATION OF ETL TOOL FOR DATA WAREHOUSING FOR NON-HODGKIN LYMPHOMA (NHL) CANCER IN PUBLIC SECTOR, PAKISTAN

Anwar Ali Sathio Anwar.sathio@bbsul.edu.pk Departmnet of Computer Science and Information Technology Benazir Bhutto Shaheed University Karachi,Sindh,Pakistan Mujeeb ur Rehman Shaikh Mujeebshaikh137@gmail.com Departmnet of Computer Science and Information Technology Benazir Bhutto Shaheed University Karachi,Sindh,Pakistan Javeria Shah Departmnet of Computer Science and Information Technology Benazir Bhutto Shaheed University Karachi,Sindh,Pakistan shahmuskan074@gmail.co

ABSTRACT— The study was primarily undertaken to establish the conceptual modeling and implementation of Data Warehousing tools through existing demographic and clinicopathologic features of NHL in Pakistan. A secondary aim was to determine the applicability of the Data warehousing in the cancer domain of Lymphoma disease. In this study, we have implemented ETL tools using open source tools and technologies for making Data warehousing and it's easy to implement with low cast in the department of health for public sector hospitals in the country.

Keywords— Data warehouse, ETL, Non-Hodgkin, lymphoma, NHL, public sector

I. INTRODUCTION

This research focus on the complete implementation of data warehousing in any health sector of Pakistan. For this study we have decided to provide a solution of implementation for a single disease i.e. cancer which is very common in Pakistan now a days. In order to make it clearer and easier to understand for any doctor, physician and other related persons, this paper will perform the implementation on a single category of cancer which is lymph node also known as lymphoma cancer (Lymphoma is a cancer of the lymphatic system, which is part of the body's germ-fighting network.). As data warehousing is considered as one of essential element in any business process, it behaves as a data collector, data integrator and information supplier. Unfortunately, Majority of the government health sectors in Pakistan are lacking in technical system. Every sector keeps their record manually and there is no such system or guidance that can help them to reduce their effort. Whereas the implementation of data warehousing in Pakistan is also next to nothing. This paper will show the use of lymph-node specific clinical Datawarehouse that could be used by doctors, physicians and other health practitioners. The process will also help doctors with the clinical decision support system (DSS) so that they can formulate a suitable model for improving the quality of diagnosis, suggestions of therapy and keep the record of patient in electronic form.

II. LITERATURE REVIEW

According to Aga Khan University Hospital, Stadium Road, Karachi represents the worldwide rise in lymphoma incidence that NHL contributes predominantly to. The agestandardized incidence rate of (ASIR) was 5.3/100,000 in men and 4.1/100,000 in women in 1995, according to Karachi cancer registry. During the study period, a gradual rise in the annual incidence was noted, with the incidence of NHL rising in 2002 to 8.4/100,000 in men and 6.5/100,000 in women, almost double The Lahore cancer registry NHL is categorized as the fourth largest malignancy in all age groups and in both sexes, according to Shaukat Khanum Memorial Hospital. On a gender-based basis, NHL was the 3rd most prevalent male malignancy while it was number six in women. [1] 2. In Palaniappan and Chua Sook Ling presented a prototype clinical decision support system which combines the strengths of both OLAP and data mining. It provides an environment of rich knowledge that cannot be achieved by using OLAP or information mining alone [2]. 3. Bagdi and Patil provided a decision-making support scheme that mixed OLAP and data mining strengths. The system anticipated the future state and generated helpful data to make efficient decisions [3]. 4. Qwaider showed how the integrated approach, OLAP with data warehousing, provides advanced decision support compared to using OLAP or data warehousing alone. He listed many Questions which cannot be answered by Data Warehouse alone or OLAP alone and showed that

ISSN: 2708-7123 | Volume-02, Issue Number-03 | September-2021
LC INTERNATIONAL JOURNAL OF STEM
Web: www.lcjstem.com | DOI: https://doi.org/10.47150

combination of both OLAP and Data Warehouse can answer the complex questions [4].

III. METHODOLOGY

A. Design of DWH and Data Model We intended to set up a data warehouse based on two distinct types of data sources (SQL database file and csv format file) for Lymphoma data registry. This section showed how we split our job into measures to fully understand the process of data warehouse implementation. Figure 1 Lymphoma Cancer Types

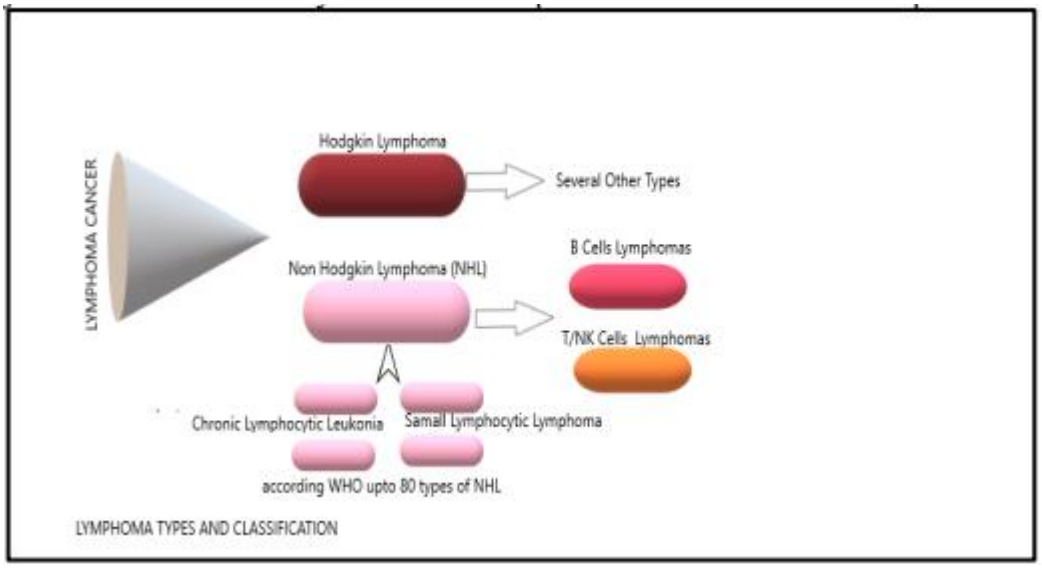

Figure 1 Lymphoma Cancer Types details

B. REQUIREMNETS GATHERING This paper mainly required to have a huge amount of data of any hospital which has the valid dataset of patients suffering from lymph node cancer. All the files are collected in .sql or .csv format files.

C. PHYSICAL ENVIRONMENT As all the process is done by using Talend for Big Data , there is no need to have a huge amount of machines or file system to keep data. The only physical device required for the software to run is a working Laptop or personal computer.

IV. DATA ANALYSIS & RESULTS

There are several factors to initiate the data modeling step following to be considered : 1. Data Labels /Attributes for ETL (Patient Profile, Hospital Info, Disease Info (Lab Findings-Attributes) 2. Data Design Schemas and Methods For ETL 3. Front End (Star Schema): 4. Backend (Data Vault Model): 5. Metadata: 6. Extract, Transform and Load (ETL) 7. Data Cube Design For OLAP 8. Reports Generating Scheme 1. DATA LABELS /ATTRIBUTES FOR ETL (PATIENT PROFILE, HOSPITAL INFO, DISEASE INFO (LAB FINDINGS-ATTRIBUTES) 2. DATA DESIGN SCHEMAS AND METHODS FOR ETL 3. FRONT END (STAR SCHEMA) First of all, we designed the data warehouse schema based on our Lymphoma data registry fields. The proposed schema is star schema which it is the most appropriate schema for our project because it can understood by the professionals and users, we can add another dimensions in the future without affect the other dimensions and make the query fast and flexible which increase the performance due to little joins between fact table and dimensions.

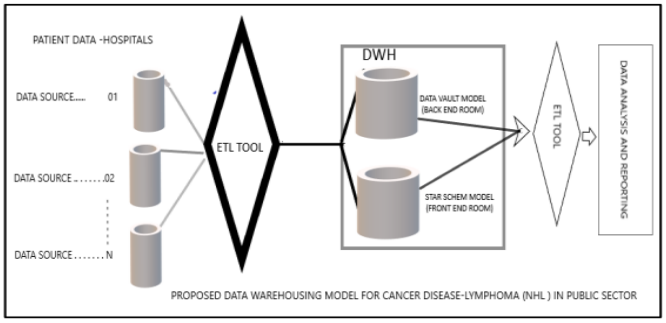

Figure 2 Proposed Data Warehousing Model

ISSN: 2708-7123 | Volume-02, Issue Number-03 | September-2021
LC INTERNATIONAL JOURNAL OF STEM
Web: www.lcjstem.com | DOI: https://doi.org/10.47150

Based on the data and by focusing on the goal of finding the relationship between the fields in the data which will benefit the clinicians and decision makers in clinical path. We designed the dimensions and fact table in the data warehouse schema using "Talend For Big Data" to the required schema.

4. BACKEND (DATA VAULT MODEL) SCHEMA:

For the backend we have used data vault model to keep long term historical data of any patient so that we can perform multiple operations on the basis of history. Another reason to select data vault as backend model is doctor or any physician can easily change the business environment whenever they want using the stored data of descriptive attributes.

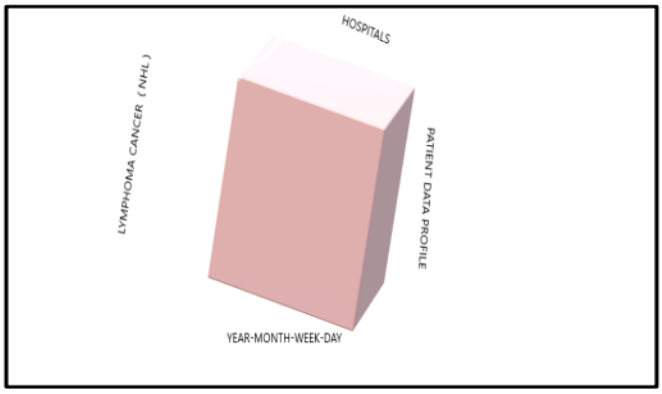

Figure 3 Data Cube Design Model

5. METADATA:

For reducing the work done by system user we added the file paths in metadata so that user should not have to add the paths again and again in order to perform Data warehousing. These metadata contain the information of path where the file is located along with the attributes the file contains

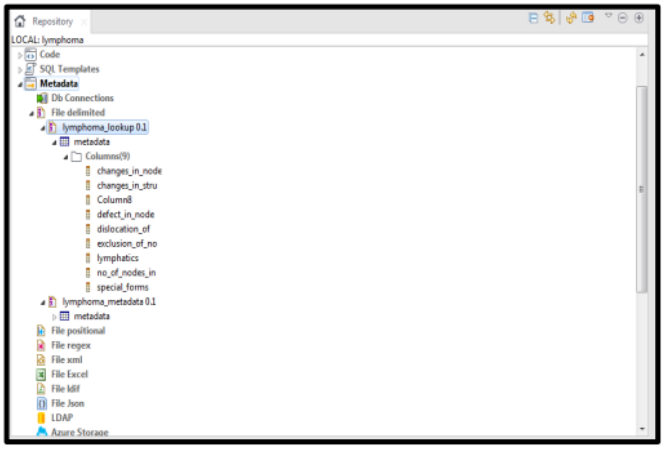

Figure 4 Implementation Meta Data overview

6. EXTRACT, TRANSFORM AND LOAD (ETL) FRAME WORK:

The structure design of our ETL based on tools from "Talend For Big Data". We used some builtin tools to create staging table or load the staging table, assign keys and some transformation processes using Talend Tools. In our project we designed ETL. The one with sequence Extract, Transform and load. In the first Package (ETL) we extract the data from a delimited(.CSV) file and work with sequence of extraction, Transformation and loading into dimensions and finally load the keys and measurements into fact table. Staging table is the intermediate step between the source and Data Warehouse tables in ETL, we loaded it with csv file data and took out the data and loaded the dimensions and fact tables from the staging table. According to [10]" our results suggest confidence in the correctness of the IDR's data, i.e. that the integrity of the EHR data was maintained during the IDR's ETL process."

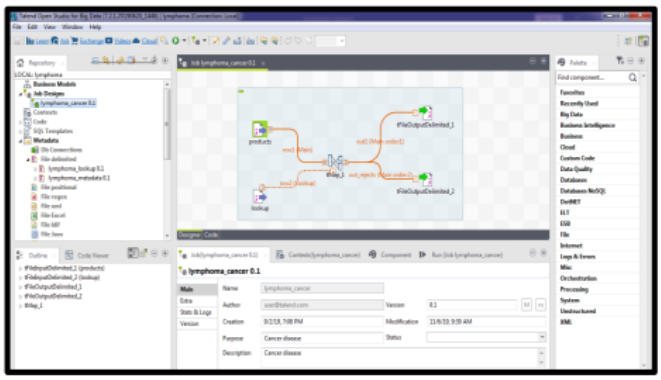

Figure 5 Overview of Meta Data table

7. DATA CUBE DESIGN FOR OLAP [8,9]

In this stage we design the cubes based on the dimensions which we created. We design many cubes, some of them based on three dimensions and some of them based on four or five dimensions. The output cubes will be used later by the reports to produce the characterized information which give full sight about the relationship between the dimension records of lymphoma registry data. The results could be based on the data of the dimension chosen, such as the count of people living in the specific province or district who were infected by lymphoma and grouped by hospital name for specific race and specific gender.

ISSN: 2708-7123 | Volume-02, Issue Number-03 | September-2021
LC INTERNATIONAL JOURNAL OF STEM
Web: www.lcjstem.com | DOI: https://doi.org/10.47150

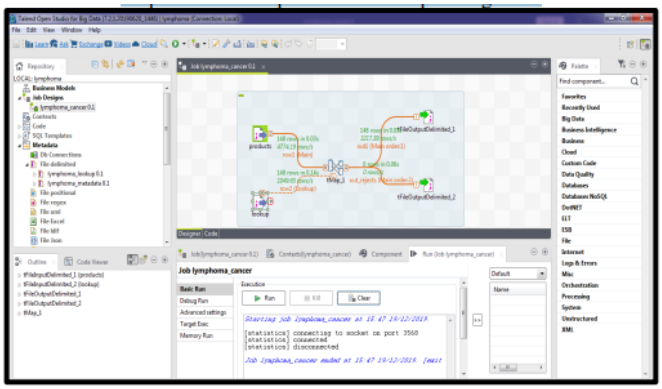

Figure 6 Data Schema and sources Overview

8. REPORTS GENERATING SCHEME:

In the final stage of the study, we prepared the reports which will be seen by the analysts and decision makers in the clinical paths. The reports are designed based on the cubes which we created in the previous step. The designing of reports is so simple so we can design a lot of reports based on our required information and with our desired chart.

V. DISCUSSION

The concept of technologies to get the low cost and maximum out and selected the open sources,

following are listed:

A. TALEND:

Talend is an application framework for open source applications. It offers various data integration, data management, enterprise application integration, data quality, cloud storage, and big data infrastructure software and services. As the first commercial open source software provider of data integration applications, Talend first came on the market in 2005. Talend released its very first software in October 2006 – Talend Open Studio, now known as Talend Open Studio for Data Integration. Since then, a wide range of products have been launched that are used in the market very favorably. Talend is considered to be the cloud and big data integration software's next-generation pioneer. This helps businesses make decisions in real time and become more data-driven. With Talend, data becomes more available, increases its accuracy, and can be easily moved to the target systems.

B. STORAGE COUNT ANALYSIS:

As mentioned above the Pakistan is still working with manual system to keep patient's record. The cost to convert this manual work into the electronic health record can vary depending on the amount of data a particular health sector. Once the data is converted to the EHR we can now find the cost of a single patient's life storage as presented In a National Library of Medicine Conference called "Long term Preservation and Management of the EHR." If we are given EHR i.e. legal record, a source of data for clinical care, and a repository of knowledge for clinical research, It was presented how do we preserve it for a sufficiently long period of time to maximize value to patient, caretaker, and scientist. The cost was generating approximately 1 terabyte of clinical text data (structured and unstructured) per year and approximately 19 terabytes of image data per year (radiology, cardiology, pathology, Gastrointestinal, Pulmonology, Ob/Gyn etc.). if there are approximately 250,000 active patients in any class /sector of patients, then the coming cost shall be as below : 20 Tb/250,000 = 80 Mb per patient per year [6]. According to Yakami et.al "The initial cost of this system was about $3,600 with an incremental storage cost of about $900 per 1 terabyte (TB). This system has been running since 7th Feb 2008 with the data stored increasing at the rate of about 1.3 TB per month. Total data stored was 21.3 TB on 23rd June 2009."[7].

C. COST COUNT ANALYSIS:

The cost count in the above conference was calculated as marginal cost of 4 megabytes of text and 76 megabytes of images for regulatory lifetimes. Every person record will add 4 megabytes per year. So the total cost will be found by adding the cost of storing old data with new data every year. So the final cost per patient was calculated as 42 cents for the first 15 years of text and $1.89 for the first 7 years of images.[5] Conversion in Pakistani rupee:According the year 2021 1 cent is 0.01868631 PKR and A dollar is 156.83 PKR. So the above cost calculations in Pakistani rupees could be: 42*0.01868631

ISSN: 2708-7123 | Volume-**02**, Issue Number-**03** | September-2021
LC INTERNATIONAL JOURNAL OF STEM
Web: www.lcjstem.com | DOI: https://doi.org/10.47150

= 0.78482502 PKR for the first 15 years the cost of text 1.89 * 156.83 = 296.4087 for first 7 years the cost of images.According to John D. Halamka, MD "In my analysis above, some may question the cost per gigabyte I used. Feel free to multiply it by 10 such that text records could be stored for $4.20 per patient for 15 years. It's still very economical" [6].

VI. CONCLUSION

Doctors, clinicians and other healthcare practitioners in Pakistani clinical institutes could implement the suggested designed strategy to support their choices. We strongly recommend the implementation of this initiative by Pakistani healthcare institutions rely on the data warehouse as a platform for their research and support their choices based on analytical information. They can view the information historically and based on place hierarchy through this project. They can adapt this CDW with Cancer data Warehouse and make them as distributed data Warehouse and after that using OLAP with data mart so they can get the valuable information from the cubes and they can Use Key Performance indicators (KPI) and performance measurements for diseases infections, indicate the critical diseases, or even watch the overall process of diseases registry systems.

REFERENCES

[1].
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3920492/#.

[2]. S. Palaniappan and C. Ling, "Clinical decision support using OLAP with data mining,"
International Journal of Computer Science and Network Security, vol. 8, pp. 290-296, 2008.

[3] R. Bagdi and P. Patil, "Diagnosis of Diabetes Using OLAP and Data Mining Integration,"
International Journal of Computer Science & Communication Networks, vol. 2, pp. 314-322,
2012.

[4]. W.Q. Qwaider, "Medicine Decision Support System Using OLAP with Data Warehousing",
The Arab Academy For Banking And Financial Sciences Faculty of Information Technology
Computer Information System Dept. JORDAN.

[5] http://geekdoctor.blogspot.com/2011/04/cost-of-storing-patient-records.html
viewed on 20th June2021

[6] https://www.massdevice.com/medical-data-storage-adding-cost-digitizing-health-records/
viewed on 20th June2021

[7] Yakami M, Ishizu K, Kubo T, Okada T, Togashi K. Development and evaluation of a low-cost and
high-capacity DICOM image data storage system for research. J Digit Imaging. 2011
Apr;24(2):190-5. doi: 10.1007/s10278-009-9267-8. Epub 2010 Feb 24. PMID: 20182765; PMCID:
PMC3056973.

[8] Parmanto, B., Scotch, M., & Ahmad, S. (2005). A framework for designing a healthcare outcome data
warehouse. Perspectives in health information management, 2, 3.

[9] Scotch M, Parmanto B. Proceedings of HICSS-38. Waikoloa, HI: IEEE; 2005. SOVAT: Spatial OLAP
Visualization and Analysis Tool. [Google Scholar] [Ref list]

[10] Denney, M. J., Long, D. M., Armistead, M. G., Anderson, J. L., & Conway, B. N. (2016). Validating the
extract, transform, load process used to populate a large clinical research database. International journal of
medical informatics, 94, 271–274. https://doi.org/10.1016/j.ijmedinf.2016.07.00