Primary Health Care: Open Access

Research Article Open Access

Data Analytics and Operational Data Integration to reach out to Rural Masses for Early Detection of Non-communicable Diseases

Vanishri Arun[1]*, Shyam V[2] and Padma SK[1]

[1]Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore– 570 006, India
[2]Forus Health Private Ltd., Bengaluru – 560070, India

Abstract

This paper demonstrates the analysis of healthcare data and integration of operational data to abate the prevalence incidence of non-communicable diseases (NCD). Pilot experiments have been carried out in Suttur village, by screening the masses for early detection of NCDs. An app has been developed to record patient profile onto a tablet. The record is synchronized to update the database on the cloud where the repository is maintained. This provides an efficient way of analysis and statistics to the huge amount of health data. ETL (Extract, Transform and Load) is a process used to extract data from data repository and transform the data based on user requirements and store in a target database as a single repository which helps to achieve goals proactively and on time. The main aim of this paper is to generate reports from the collection of health data by using tools like QlikView for Business Objects as a front end tool for generation of reports and charts, Oracle 11g as backend tool for creation of data repository and Talend Open Studio 5.4 as an ETL tool. We conclude that ETL systems enable a smooth migration from one system to another. By creating an ETL script for each system, data can be stored in a consistent format in the repository. The source system can then be changed, without any impact on the repository or the reporting/analysis systems. Therefore there are phenomenal improvements in turnaround time for data access and reporting. The entire health data can be standardized as there will be one view of information. Health data synced by various sources from different places can be merged to create a more comprehensive information source. This leads to reduction in costs to create and distribute information and reports and also helps in reduction of prevalence incidence of NCDs.

Keywords: Healthcare; Non-Communicable diseases; ETL; Talend open studio; Qlikview

Introduction

The mortality rate due to Non-communicable diseases (NCDs) is increasing and 75% of Indians in rural villages are the victims of NCDs such as Diabetes, Hypertension and Obesity which represent global burden of diseases and cause deaths each year mainly in rural areas [1]. This is in turn affecting the economic conditions of the people. The health-related datasets provide access to patient information to study, analyze and report the health conditions and outcomes for making decisions, policies and to develop processes [2,3].

Data warehouse has been defined by Bill Inmon as a subject-oriented which can be used to analyze a particular subject area, integrated which integrates data from multiple sources, time-variant which enables to retrieve historical data and non-volatile collection of data in support of management's decision making process [4]. Ralph Kimball defines data warehouse as a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision makes [5]. In this paper, data warehouse has been modeled to analyze to improve outcomes, safety and patient satisfaction. And accordingly the database is structured. The data warehouse consists of tools to extract data from the multiple operational databases and other external sources, to clean, transform and integrate these data and to load into the data warehouse [6].The health data collected are stored and managed in the warehouse. Data marts present multidimensional views of data to a variety of front end tools like query tools, report writers, analysis tools, and data mining tools [7]. Data warehouse provides an effective way of analysis and statistics to the huge amount of data and helps in decision making. The concept of ETL is followed which means Extract, Transform and Load where data from different sources are extracted, stored in a single repository and analyzed for supporting decision making. Data warehouse enables extraction of information from more than one area and have a single view of information used for decision making. In healthcare sector, in order to store all the patient data, a tool is used which is a comprehensive Adverse Event (AE) tracking and reporting system which will be used as a source for any ETL tool for generating the report in order to investigate the results and provide regulatory authorities to take necessary actions. Available data sets from the authorities as a source for data analysis and report generation may also be used. Early diagnosis of NCDs is beneficial in enabling less invasive diagnostic evaluation and treatment. It also accelerates the timeline for intervention, reduces the scale and costs of medication and other interventions and May also increases mortality rate or at least delay the onset of symptoms which helps to maximize patient quality of life. Therefore there is a burning need for an effective screening system for early detection, analysis and reporting of NCD in rural areas. Many programs have been launched by the government and Non-governmental organizations to screen the rural masses to identify the individuals at risk and take necessary actions. But the bottleneck for all these programs has been the acute shortage of specialists and experts to screen analyze and report [8].The analysis of the data collected is manually done by Community medicine personnel and reports are generated.

The effectiveness of NCDs screening and utilization of experts time rely on the platform which is offering screening, management of referrals and deskilling. The objective of this work is to develop and

***Corresponding author:** Vanishri Arun, Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore–570006, India, Tel: +919480564806; E-mail: vanishriarun@gmail.com

Received October 06, 2015; **Accepted** November 06, 2015; **Published** November 13, 2015

Citation: Arun V, Shyam V, Padma SK (2015) Data Analytics and Operational Data Integration to reach out to Rural Masses for Early Detection of Non-communicable Diseases. Primary Health Care 5: 210. doi:10.4172/2167-1079.1000210

measure the efficiency of screening the rural masses for NCDs and the analysis of the information flow, patient referrals and follow-ups providing an intelligent, affordable and accessible system that involves deskilling and optimization.

This study is done in association with the Department of Community Medicine, JSS Hospital, Mysuru, India, which is a 1800-bedded hospital offering comprehensive medical services. JSS Hospital has taken up several projects involving community health survey, health education & treatment for various ailments in Sri Kshethra Suttur village which is situtated in Nanjangud Taluk, Mysuru District, Karnataka ,India. It has a population of nearly 5,000. This is an on-going project in which charts and reports are generated to serve various departments of the hospital to take necessary actions or conduct various programs related to the patients' health in the 13 villages of Sri Kshethra Suttur, Mysuru, India. As per guidelines issued by JSS Medical College, screening involves simple clinical examination comprising of relevant questions and easily conducted physical measurements such as demographic details, history of diabetes, history of tobacco and alcohol consumption and measurement of blood pressure, blood sugar, height, abdominal circumference etc. These data stored in the warehouse are then used to periodically generate reports and charts based on queries.

Existing system

Mart is designed for analytical and reporting purposes and also as a data provider for the downstream applications in Reporting and Analytics such as Dashboards/Graphs. Due to the nature of data in the source application there is a possibility of sparsely populated record set in the source application. And as per the regulatory guidelines certain information need to be captured as reported without any corrections to the data [9].

Owing to the above, it is understood that the source data can be incomplete in many instances. This key consideration is being taken into account while populating and accessing the data using ETL process. Existing data (legacy and data that is available on the initial run) will be loaded on the first run. Thereafter, only incremental data will be captured and processed during every subsequent run. Any updates within the case version will be captured and maintained as Slowly Changing Dimensions (SCD) [9].

Fetching the data from various sources requires sequential query fetching and output will be generated as excel sheet which has several drawbacks in which user should know the concept of Database queries and identification of required data in millions of records which is tedious. Since excel sheets are used, it has many limitations which mean it stores only 256 columns. Adding such kind of sheets in millions reduces the performance to fetch the data. Hence to improve this, database concept is used where sequential query fetching for the end users is a tedious job.

Therefore limitations of the existing system are sequential query fetching, redundancy in data and requirement of query language by experts to fetch the data from data base. To overcome these limitations, a system is modeled to develop and measure the efficiency of screening and the analysis of the records using Oracle Schema creation, DDL statements creation, Talend software to create and run jobs and Qlikview for analysis and reporting.

Methodology

This is an ongoing project undertaken in association with the Department of Community medicine, JSS Medical College, Mysuru and Primary Healthcare Centre, Sri Kshethra Suttur, Nanjangud Taluk, Mysuru, to reach out to the rural masses. Pilot experiments have been carried out in Suttur Mysuru, by screening the masses for early detection and analysis and reporting of non-communicable diseases. An app has been developed to record patient demographics, vitals and other details onto tablets. These tabs are given to Community health workers who covered 13 villages of Suttur, go door-to-door and recorded patient demographics by screening for vitals like BP, Sugar, abdominal circumference, etc., [10]. A total of 150 cases were screened and values were recorded on the tab.

These records are synchronized to update the data warehouse on the cloud where the repository is maintained. The cloud is maintained at Forus Health Pvt. Ltd., Bengaluru. The data warehouse is in a denormalized form with duplications or incomplete entries in the source layer. Then the warehouse is normalized in the staging area where duplicate, redundant and incomplete records are removed or filtered. This normalized and cleansed database (target database) is connected to reporting layer which displays only required data based on the selection criteria of the user which are used for analysis and reporting. Data warehouse are refreshed periodically by extracting, transforming, cleaning and consolidating data from several operational data sources. The data in the warehouse is then used to periodically generate reports, or to rebuild multidimensional views of the data for on-line querying and analysis.

Proposed system

The overall system architecture of the proposed system is as Shown in Figure 1

The system is grouped into the following areas or layers of the warehouse.

Source layer: Source layer contains data from different databases across Suttur villages which are in de-normalized form i.e., databasing with redundant data [11]. Due to some reason a patient record could be entered more than once by the community health worker. This layer is not accessible to end user for report generation.

Staging area: The information in this layer is typically a copy of the source system (or operational system) data along with some cleaning (filtering out error records), control and audit events. Staging area is strictly a Data Warehouse layer and no access will be given to the business users. Therefore reports cannot be run against this area. This area ensures that the window of access to the source system is minimal. At the same time, it is ensured that all the data required for running the daily cycle is available in the Data Warehouse. The frequency of the load is configurable. These processes may apply consistency, standardization of structure, data cleansing, business rule transformation etc. The Centralized Data Warehouse (CDW) layer is generally in the Third Normal Form (3NF) i.e., reduction of duplicate or redundant data [11], in which all the duplicate and redundant records are removed, which can also contain de-normalized tables.

Data mart: It's a subset of Data Warehouse normally referred to as target database. The Data Mart is based on Dimensional modeling which uses the concepts of facts (measures), and dimensions (context) [11]. This is the ideal place for analytical reports and dash boards to be generated. This contains aggregated tables at various levels. This is fully normalized and cleansed required data schema and the reporting layer is directly connected to this database. All the operations performed in the dashboard of the reports will be directly handled by the Data Mart.

Report layer: Finally report layer is the front end of the Data Mart

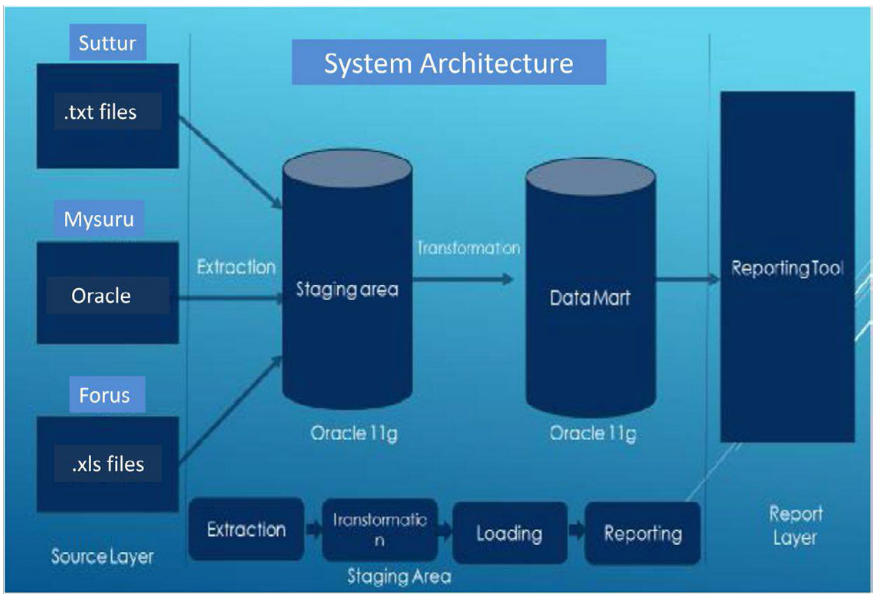
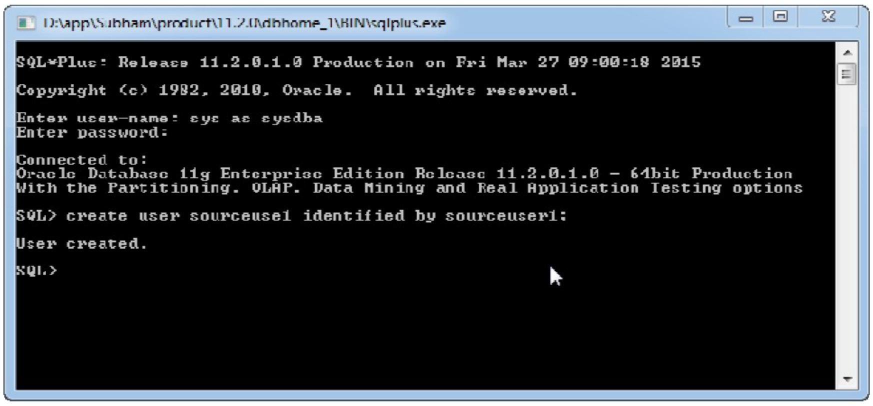


Figure 1: System architecture of Data Warehouse application.



```
SQL*Plus: Release 11.2.0.1.0 Production on Fri Mar 27 09:00:18 2015

Copyright (c) 1982, 2010, Oracle.  All rights reserved.

Enter user-name: sys as sysdba
Enter password:

Connected to:
Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 - 64bit Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options

SQL> create user sourceuse1 identified by sourceuser1:

User created.

SQL>
```

Figure 2: User creation.

for user friendliness which displays only required data based on the selection criteria of the user [12].

Implementation

The following 4 modules are implemented in this application:

1. Oracle Schema creation.

2. Creation of DDL statements.

3. Create and Run jobs using Talend software.

4. Reporting layer

Oracle schema creation

This section explains the creation of Oracle users in order to load the data at various levels (i.e. source, staging and target).

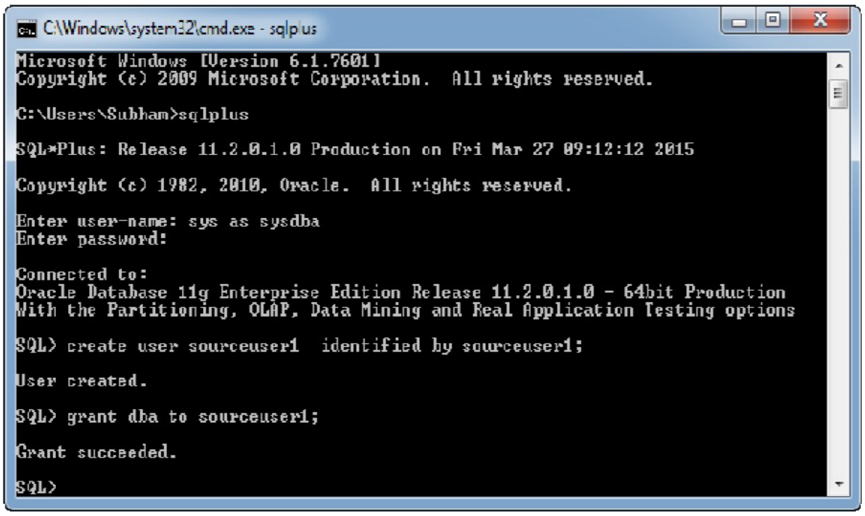
Open the Oracle installed and login to Oracle as system user and create Oracle user by using the syntax below. Use the CREATE USER statement to create and configure a database user, which is an account through which you can log in to the database, and to establish the means by which Oracle Database permits access by the user as Shown in Figure 2.

Syntax: Create user USERNAME identified by Password;

Once user is created, grant DBA privileges to the user so that user will have the respective privileges to create table and structures.

Syntax: Grant DBA to USERNAME; Shown in Figure 2-4. Data are collected from various community health workers going door-to-door in all the 13 villages of Suttur.The files in different tabs of source will be migrated to staging area with one to one mapping without any change but in single format (Oracle 11g) this staging area is exact replica of source in single format as shown in Fig 4. All the data in staging area are in denormalized form and cleaning/filtering is required to remove duplicate/incomplete records. Some of the fields are Age_group, Hyper_tension, Family_income, etc., which are migrated to staging area without any change.

Staging to target: The data flow between staging and target is depicted in Fig 5 where business logic is applied and all the tables' primary keys of staging area will be moved to the fact table where they



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\Subham>sqlplus

SQL*Plus: Release 11.2.0.1.0 Production on Fri Mar 27 09:12:12 2015

Copyright (c) 1982, 2010, Oracle.  All rights reserved.

Enter user-name: sys as sysdba
Enter password:

Connected to:
Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 - 64bit Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options

SQL> create user sourceuser1  identified by sourceuser1;

User created.

SQL> grant dba to sourceuser1;

Grant succeeded.

SQL>
```

Figure 3: Grant DBA to User.

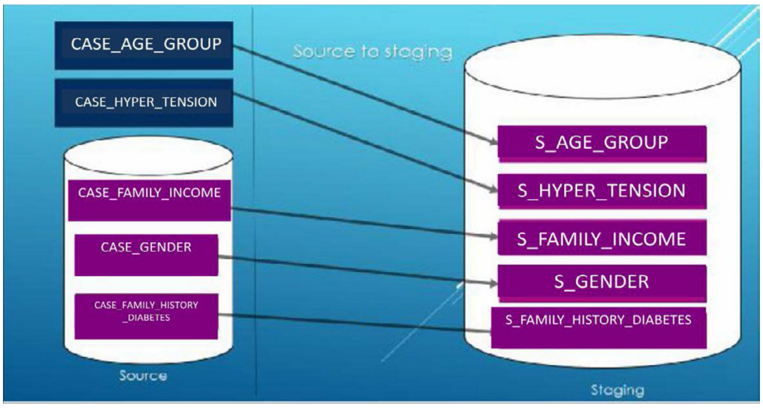


Figure 4: Source to staging data flow.

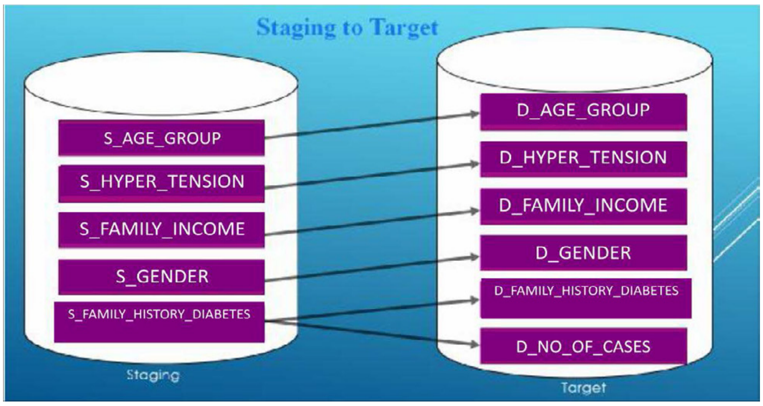


Figure 5: Staging to target dataflow.

act as foreign keys. All cleaned data in normalized form after joining all the required tables will be obtained as star schema structure in target database. For e.g., source S_Age_group which is in denormalized form is transformed into normalized D_Age_group (Figure 5).

Target to reporting: The data flow between target and report layer is depicted in Fig 6 where databases are fetched on the selection criteria with the help of Business Objects (BO) tools and it will internally fetch the required columns and display the data as an output on the dashboard as per the requirements. For e.g., the normalized fields D_Age_group, D_Hyper_tension, D_Family_income, etc., are ready to be presented to the Report layer. Based on the user query, fields are fetched and related charts and reports are generated (Figure 6).

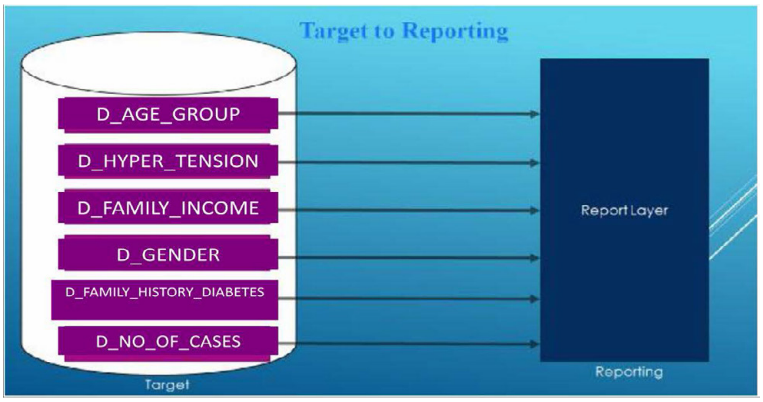


Figure 6: Target to reporting data flow.

Creation of DDL statements: Creation of tables/structures for the Oracle users based on the data modeling design document is done in this step.

Create/import jobs using talent software

Stage jobs creation: Once the source data preparation is done, read data from source and load to the stage. The sources may be database files/text files/excel files. One-to-one mapping from source to staging and no cleaning of data are required. In the query section the data from the source is read by selecting all the columns [13].

Various components are used to read the data from different sources namely:

a. t Oracle input: To reads the data from the data base.

b. t File list: To read the data from textfiles.

c. t File Input Excel: To read the data from excel files.

Once it is all set click the run button in order to read the data from source to staging. After successful run the stage jobs verify the data loaded properly in the stage schema. Similarly load the data from different sources. Once one to one mapping of data loading from source to staging apply the business logics in order to clean the data by selecting only the relevant records in the query section and load into the target database. After reading data from stage schema store data into the target database using Oracle SCD components which allow to change the dimension of the table before loading into the target. After successful run of the target data, verify whether relevant data are loaded by logging into the target schema database. Ensure that all the data are loaded properly and perform unit testing for the data loaded. Then this should act as a source to the report layer to fetch data. The data loaded successfully can be verified by cross checking the target database details which ensures that Data cleansing and Data denormalizations are done.

Old	M	N	Y	Y	N	Extreme lower	1
Old	M	Y	N	Unknown	Y	Extremelower	1
Old	M	Y	N	Y	Y	Extremelower	1
Old	M	Y	Y	Unknown	Y	Extremelower	1
Old	M	Y	Y	Y	Y	Extremelower	1
Young	F	N	N	N	Y	Extremelower	1
Young	M	N	N	N	Y	Extremelower	1
Young	M	Y	Y	Y	Y	ExtremeLower	1

Table 1: Report in the form of Excel sheet generated for the criteria above.

ODBC connections from target to report layer

Open Database Connectivity (ODBC) is a protocol that can be used to connect a Microsoft Access database to an external data source such as Oracle/server.

Reporting layer

Once the data source is added, load the data into the front end by performing the following;

1. Open the Qlikview.

2. Design the Qlikview.

3. Reload the configuration.

While reloading the process, it will fetch the data from the target database and display in the front end. The data display can be modified with various selection criteria.

Ensure that Oracle Data Base Connection (ODBC) connection is successful and select the data source of target ODBC and click on load by selecting the various chart types and selection criteria. The data are displayed in report format on the front end as shown in Figures 7-9, Table 1.

Table 1 is the report generated in tabular form for the selected criteria. Totally 150 cases were imported to source layer in denormalized form. After normalization, the total cases were filtered to 145.

Result and Discussion

Main aim of our project is to generate reports by collecting data from various data sources by transforming data from sources and loading in a Data Warehouse repository. Different data, based on various selection criteria can be selected. After data are loaded, verify for the correctness of data by querying in the backend. Thereby the count from target to

AGE_GROUP	GENDER	DRINKING HISTORY	SMOKER	HYPER_ TENSION	FAMILYHISTORY DIABETES	FAMILY INCOME	CasesCount
Middle	F	N	N	Y	Y	Extreme Lower	1

Table 2: Case Count =1 with the selected criteria.

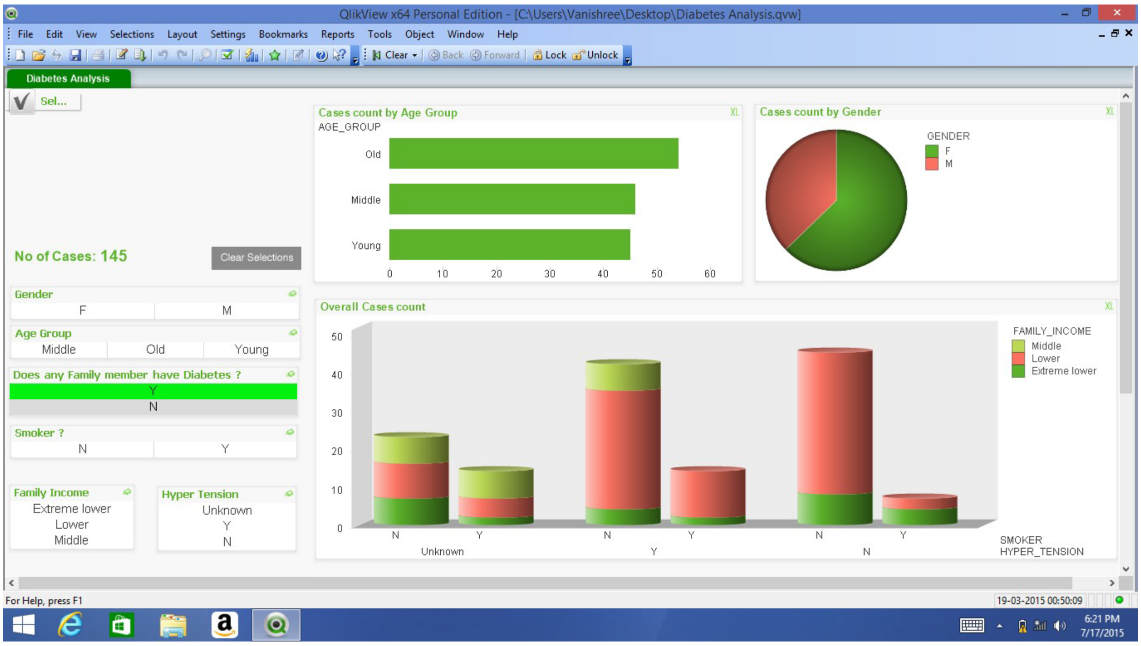


Figure 7: Chart generated without any Criterion with No. of cases = 145.

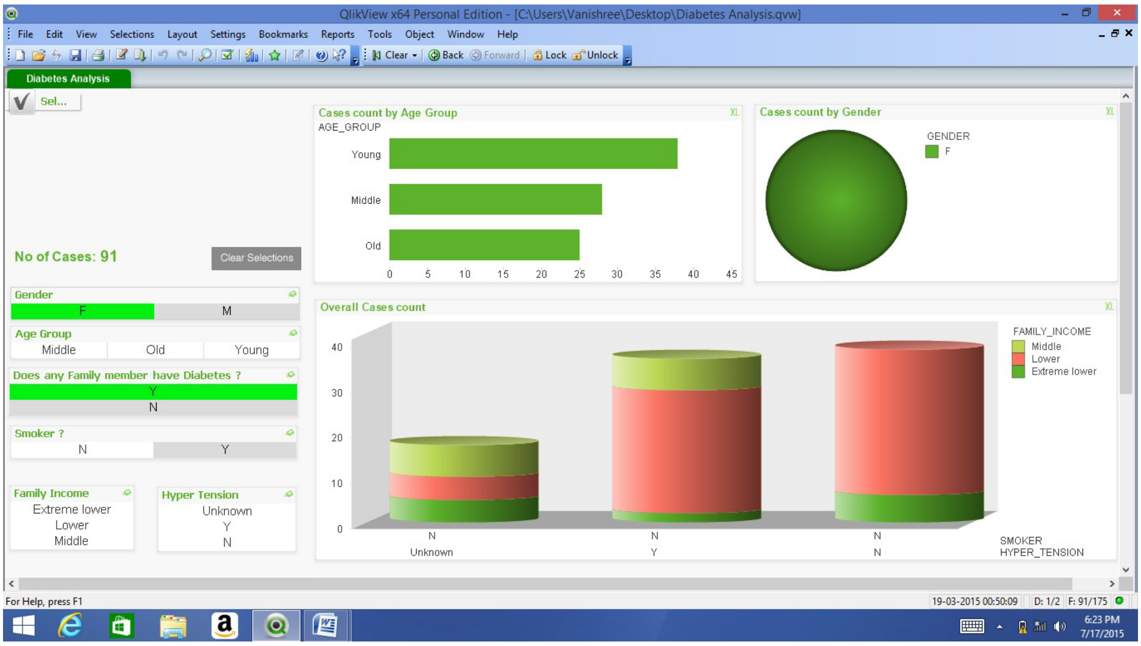


Figure 8: Cases by GENDER="Female".

reporting can be matched. Reports are generated in the form of charts and tables. It is also easy to analyze the report and check the severity of the cases (Table 2).

Table 2 depicts the number of case count with the criteria, Age_group between 30 and 50 years, Gender being female, with no alcohol and smoking history, BP > 120/80, with a family member being diabetic and economically poor background. This helps the patient to know her health status, the health worker alerts the patient of the complications, clinical intervention and treatment required. It also helps the Department of Community Medicine to have a statistical and disease analysis of these rural areas to take necessary actions.

Conclusion and Future Enhancement

A key use for ETL systems is to enable a smooth migration from one system to another. By creating an ETL script for each system, data can be stored in a consistent format in the data warehouse. The source system can then be changed, without any impact on the data warehouse or the reporting/analysis systems. It has the following advantages:

1. Phenomenal improvements in turnaround time for data access and reporting.

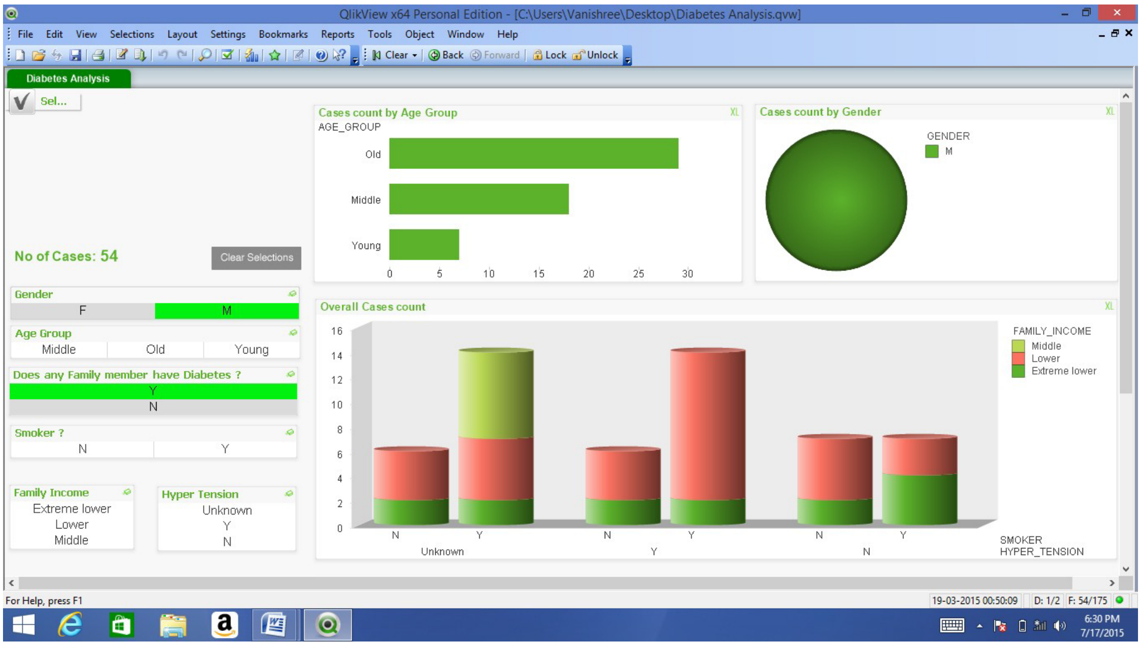


Figure 9: Chart generated for a combination of criteria.

2. Standardizing data across the organization so that there will be one view of information.
3. Merging data from various source systems to create a more comprehensive information source.
4. Reduction in costs to create and distribute information and reports.
5. Helps in informing the public health system of the prevalence of various routine disease conditions.
6. Prepares the health system to respond to unforeseen epidemics so that effective therapy is given at the right time and place.

Since open source software's are used in developing this system, it is sufficient to build the databases through the app which can be installed on tabs. People with minimum computer literacy can use the app and fill in the information where most of the fields have drop-down menus to select options.

As future enhancement, it is intended to port the app on to smart phones which are handy and using the fields available in the data warehouse, predict related diseases to abate the prevalence incidence of various diseases.

References

1. Olivia Namusisi, Juliet N Sekandi, Simon Kasasa, Peter Wasswa, Nicholas T Kamara, et al. (2011) Risk factors for non-communicable diseases in rural Uganda: a pilot surveillance project among diabetes patients at a referral hospital clinic. Pan Afr Med J 10:47.
2. Stead WW, Searle JR, Fessler HE, Smith JW, Shortliffe EH (2011) Biomedical informatics: changing what physicians need to know and how they learn. Acad Med 86: 429–434.
3. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA 309: 1351–1352.
4. W. H. Inmon (2002) “Building The Data Warehouse”.Wiley Computer Publishing.
5. Ralph Kimball, Joe Casertam (2004) The Data Warehouse ETL Toolkit.
6. Laura Hadley (2002) Developing a Data Warehouse Architecture.
7. Surajit, Chaudhuri, Umeshwar Dayal (1997) An Overview of Data Warehousing and OLAP Technology.
8. M. G. Deo (2013) “Doctor population ratio for India - The reality”. Indian J Med Res 137: 632–635.
9. Marek Wancerz, Paweł Wancerz , “ History management of data – slowly changing dimensions”.
10. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui (2013) “Application of data mining: Diabetes health care in young and old patients”, Journal of King Saud University – Computer and Information Sciences 25: 127–136.
11. https://www.en.wikipedia.org
12. Dario Antonellia, Elena Baralisb, Giulia Brunoa, Tania Cerquitellib, Silvia Chiusanob, et al. (2013) “Analysis of diabetic patients through their examination history”. Expert Systems with Applications 40: 4672–4678.
13. Talend Infosense Solution Brief Master Data Management for Health Care Reference Data: White Paper.