Using ETL framework for managing data in the Covid-19 Laboratory Network-Project Management Unit of the Department of Health

Dwight D. Sabio
**Corresponding Author: dwight_sabio@dlsu.edu.ph*

Abstract: Way back in January 2020, the Covid-19 testing capacity in the Philippines was very limited. The Department of Health(DOH) had to send its samples to Australia for testing. By February of the same year, DOH through its National Reference Laboratory, the Research Institute for Tropical Medicine(RITM), was the first laboratory to perform Covid-19 testing. This is followed by its 7 sub-national laboratories. By the end of 2020, the DOH has licensed 199 Covid-19 PCR-based laboratories nationwide. As of June 10, 2022, we have 340 labs, with 238 in Luzon, 46 in Visayas, and 56 in Mindanao.

This paper created a framework on how data can be organized using concepts of extraction, transformation and loading of information products into a knowledge-based repository. As the network of Covid-19 testing laboratories keep growing, and increases the testing capacity of the country, comparing the daily testing capacity output between public and private laboratories and among plate-based laboratories and cartridge-based laboratories become a problem partly due to inconsistent data in the DOH Covid-19 Data Drop.

The study seeks to address the problem by adopting ICT to organize the data as knowledge resource for the Covid-19 Laboratory Network-Project Management Unit(CLN-PMU). In creating the framework, process models were used to analyze sample datasets from the Covid-19 Data Drop, particularly the Testing Aggregates (daily output samples tested dataset). Based on the analysis, the framework which contains three phases was implemented with ICT following technology review on Talend Open Studio for data integration, Tableau Public for visualization and Wordpress for content management. To validate the framework, an informal demonstration was presented to the Covid-19 Lab Network to gather feedback and validation to its personnel.

Key Words: Testing aggregates, plate-based, cartridge-based, Covid-19 Data Drop, ETL

1. INTRODUCTION

The importance of information in decision making processes cannot be overemphasized. Successful outcome is the result of decisions made with the right information. Having data to produce verifiable information is therefore critical in major decision undertakings.

One of the technology's contributions to today's decision making processes lies in its capacity to transform data from their basic forms to sources of information and knowledge. From source data to information products, technology operates on intermediate stages of data processing which include extraction, consolidation, presentation, storage and distribution. There is a need to understand the role of technology in the broad spectrum of data processing which starts from collection of data up to consumption of processed data and information.

Knowledge creation is the primary reason why we transform data into useful information. Understanding the role of information and communication technology (ICT) is a key to address the challenges in providing the right information to decision makers. With the contribution of ICT in

governance reforms now widely recognized, one of the ways to ensure availability of processed data and information products to decision makers is through ICT. Finding the right intervention to society's problems includes the effective use of ICT in decision making processes.

Organizing datasets from the Covid-19 data drop is a process of extracting data on the Testing Aggregates, transforming them into information products and presenting the results as comprehensive as possible. Accordingly, this study aimed to develop a framework on data organization which outlines the processes of transforming the Covid-19 data drop to information products for decision support to Covid-19 Lab Network planning. The proponent believes that ICT is capable of organizing data in ways that generate knowledge which can in turn influence decisions on Covid-19 Lab Network programs.

Objective of the study

This study aimed to develop a framework to organize data with the Covid-19 Data Drop using ICT. To implement data organization, the Testing Aggregates was collected from the Covid-19 Data Drop and standardized from various sources prior to processing into information products. The process of data organization includes consolidating the Testing Aggregates data, transforming the data into tables and graphs using business intelligence tools, and loading them into a repository for presentation and storage.

In developing the framework, ICT tools for organizing the data were evaluated. The researcher conducted review of software products and synthesized ICT procedures on extracting and transforming data prior to loading to the repository. The ways ICT organized the Covid-19 Data Drop was explored using the concepts derived from Extract, Transform and Load(ETL) technology.

The researcher was guided by the objectives as follows

a. To develop a framework which describes how ICT organizes the Covid-19 Data Drop into a knowledge resource for the Covid-19 Laboratory Network under the Department of Health.

b. To describe the processes of transforming the Covid-19 Data Drop into information products.

c. To describe how the Covid-19 Data Drop information products can be presented, stored and disseminated to the Covid Laboratory Network decision makers.

2. METHODOLOGY

This study falls under the qualitative method of research on the data analysis of the framework which were gathered from top officials and the staff through informal questionnaire and interview.

The main respondents are the top officials of the Covid Laboratory Network.

Data Collection

The collection of data took 11 months since I worked in the Covid-19 Laboratory Network as the Data Manager. My work allowed me to interact closely with the top officials and staff as well as employees of the population. Part of my work as the Data Manager was attending staff meetings, attending Covid-19 Laboratory Network meeting, National Health sector meeting, join field visit in RITM, scanned data files and review publications. Essentially, my data gathering involved informal interviews and discussions with top officials and staff of the Covid-19 Lab Network Project Management Unit.

3. RESULTS AND DISCUSSION

The management of datasets from the Covid-19 Laboratory Network will be carried out using Talend Open Studio to extract the data, Tableau to transform the data and Wordpress to load information products into the Content Management System.

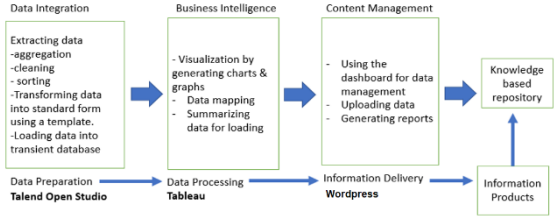

Fig. 1. ETL based framework for Managing Data in CLN-PMU.

ETL Technology (Extract, Transform and Load Framework)

-it is a data warehousing process of extracting data from source systems, applying set of rules to transform the data, and loading processed data into the website.

Process Models on Data Management	Description	ICT tools for managing Testing Aggregates dataset
Extraction (Data Integration)	- extracting datasets by aggregating, cleaning & sorting include de-duplication, removal of empty records(rows) or extraneous values in the columns. -Retrieving data from different source systems for eventual transformation using set of rules	Talend Open Studio
Transformation (Data Visualization/Business Intelligence)	- It helps in simplifying raw data in a very easily understandable format. - Tableau helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboards.	Tableau
Loading (Data Presentation/Content Management)	- Writing the data into the target database(website) - The final phase of data organization is presentation of information products. This area uses the content management software(Wordpress)	Wordpress

Fig. 2. Process Models on Data Management

Data Integration Software

	Pros	Cons
Talend Open Studio	-Huge amount of data can be cleaned and formatted within a few seconds -Very easy to handle data since you can visually setup jobs and see their progress during execution -The free version lets you write custom codes; built in modules are easy to configure to connect all kinds of database and major servers	- Although its tools can be learned easily, they can be quite complex for first-time users. - Knowing Java is pre-requisite for advanced coding; however, Java integration lets you write down custom Java routines for all sorts of transformations to your data.
Azure Biz Talk	-Simple architecture and user-friendly development environment with minimal hardware requirements. -Lots of support from Microsoft community. -Ease of deployment using Microsoft's .NET platform but can be limiting for developers working on non-Windows environment	- expensive for developing products for long term - cloud based service requires dedicated services to run on per pay basis; overkill for simple application system requirement

Informatica	-Supports integration across all different platforms such as different hardware/OS and multiple database. -Easy data mapping with visual reinforcement	-Tools can be costly in maintenance and support base -high performance tools for enterprise-wide processes but too complex to manage

Fig. 3. Data Integration Software

Data Integration Software

-For integration of sample datasets, the researcher use Talend Open Studio primarily for ease of use and having a fast and powerful open source ETL engine with free version readily available for download on its website. Although the proponent was completely new to Talend, the software was easy to learn given its simple interface and straightforward data integration procedures contained in its modules.

-Despite Informatica's edge on Gartner's Magic Quadrant in leader's category, Talend was chosen for its narrower learning curve in addition to having an open source version which immediately gets you started through an ample set of connectors and hosts of components & transformation tools. The paid version of Talend has data quality and management & monitoring tools which are necessary for high level projects.

Business Intelligence Software

	Pros	Cons
Tableau	-enables users the ability to build data visualization in a matter of minutes with capability to incorporate several data sources -with drag and drop functionality	-Can be very slow when data sources become large -visualizations and customizations of text fields are limited compared to some other BI solutions -needs improvement in
	easy for any user to understand and doesn't require technical training sessions to build a dashboard -the user is technical and the ability to incorporate free form queries or advanced join statements is an option -the location of the map can be shown easily given the latitude and longitude values	user interface and better data connection process.
QlikView	-Filter once and have it applied to all the visuals with graph and tables rendering quickly with good-looking user interface -ability to load from many data sources(flat files, direct queries form database) -easy to build basic filters, charts, tables and can load and process millions of rows quickly - the user community is excellent where you can usually get useful replies to most	- no ability for non-technical people to self-service e.g. can alone an object and tailor it but the end users struggle enough with the concept of filtering that they usually can't handle creating their own views. - some of the syntax is really strange in the expression formulas so it takes a long time to figure out how to do some basic function. - there are some basic features that is really complicated to program e.g. the ability to filter an expression value.

	of the questions asked.	
Microstrategy	-one of the most consistent product architecture(across version) and scalable products in the market -at the forefront of Business Intelligence design and development on the mobile platform -under the hood, the query engine can write some really complex logic and churn out impressive results. - you can literally control every setting with in Microstrategy but they are way too many. -can easily write custom SQL for a report without setting up the tables, attribute and dimensions	- has a steep learning curve, needs significant time and resources to build up expertise and has a lot of options to choose from which is a bit of overkill. -expects your development/working environment in a certain format hence has right requirements in terms of how it expects to be structured which may add more time in terms of ETL design time. - it is arduous to setup tables, attributes, and metrics with too much clicking.

Fig. 3. Business Intelligence Software

CMS Software

	Pros	Cons
Drupal	-One size fits all in terms of usability mainly for websites with multiple users. -For complex, advanced and versatile websites, Drupal has the flexibility and complex functionality without necessarily knowing advanced Php and JavaScript	- requires complex data organization - requires the most technical expertise among the three CMS.
Joomla	-Best used for e-commerce, social networking sites with more content and structure flexibility than WordPress but fairly easier to use than Drupal -With relatively uncomplicated installation and setup and small investment in effort to understanding structure and terminology.	- Not as customizable as Drupal nor as user friendly as WordPress which makes it a middle-of-the-road CMS. - Joomla easily gets too complicated for sample CMS needs(sites with few pages and fewer changes of content over period of time)
WordPress	-Ideal for fairly simple websites such as blogs and news sites -Easy to manage with add-ons that's easy to apply to extend functionality. -Technical expertise is not necessary due to its intuitive interface compared to Drupal and Joomal; copy-pasting text from MS Word document is more straight forward	- finding the best plugins and theme is time consuming, good free theme are especially hard to find without links back to the developer site. - Only MySQL is supported as backend database-neither Postgre SQL nor Oracle are supported.

Fig. 4. CMS Software

Content Management System

Of the three CMS software, Wordpress is the most popular due to ease of use and uncomplicated set of features which makes it particularly ideal for beginners-WordPress provides automatic installation which enables you to create a website in five minutes or less than an hour through manual installation. Due to its popularity, it also has the largest modules and plugins available compared to Joomla and Drupal. With millions of websites running in WordPress, it also has the largest community of support form its users. The downside to its popularity, however, is the lack of advanced development features and its significant use of server resources which often times results to slowdown in website performance.

On the part of Joomla, it is still considered as a compromise between WordPress and Drupal. Although it doesn't require the same level of expertise as Drupal, it requires less resources than WordPress to run smoothly on most web servers without any problems. Another strength of Joomla is its native support on setting up commercial websites such as online stores which also possible with WordPress and Drupal, using Joomla makes it faster and easier. The middle ground between Joomla's ease of use and Drupals's more powerful features makes Joomla ideal for developers seeking for something easier to manage yet contains enough features for more advanced website functionality.

Among the three CMS, Drupal is considered the most difficult to master. This is partly due to its more advanced capabilities. At the cost of its powerful features is a steep learning curve with requires a higher level of understanding of its complicated paradigm on website development. Despite higher investment on skilled manpower and development time, Drupal offers a more robust development platform for more advanced web applications. The researcher then chooses Wordpress as the content management system to use since it is easy to manipulate for everybody, especially the non-technical people can create content on it by drag and drop the user interface.

4. CONCLUSIONS

Based upon the findings in the Covid-19 Laboratory Network, I found out that the personnel/staff lacked technical training and too little time to train using Talend Open Studio, Tableau and Wordpress. Majority of the personnel/staff in the CLN are familiar with Microsoft/Google applications so the CLN decided to use the technology that everyone knows. Therefore, a serverless system with no setup cost was created. The personnel/staff of the CLN tend to maximize the utilization of Microsoft Excel and build an entire system based on Excel, Google Spreadsheet, Google Cloud free-tier products(Google Sheets, Google Forms, Google Site, Google Apps Script, Big Query and Cloud Storage). By maximizing the google technology, it allows data synchronization in real time including the 340 covid-19 licensed laboratories. The personnel/staff only needs to be familiar with Excel, Spreadsheets and web forms.

The CLN-PMU has not yet developed up until now but has plans in buying a Laboratory Information System(LIS) which include features and tools for data analytics and visualization.

5. ACKNOWLEDGMENTS

I would like to acknowledge my father(Eduardo A. Sabio) and my family for the inspiration and the never ending support they gave me in writing this paper. My parents took care of my 2 daughters while I was busy working on my job and writing this paper.

6. REFERENCES (use APA style for citations)

Vo, H. T. A covid-19 data management system in ho chi minh city, vietnam[Google slides]. The City College of New York. Google Drive. https://drive.google.com/file/d/1p8RORhkAwFETihxEV2V5-x89IFQkS6KN/view?usp=sharing

COVID-19 Lab Network Commodities 2021[Google slides]. Department of Health. Google Drive. https://docs.google.com/presentation/d/19NkAETGFsj8rUGJhNtvHZyz2WjQ37-Pi/edit?usp=sharing&ouid=107544296634563766729&rtpof=true&sd=true

Gomez, E. B. (2016). ETL-based framework for organizing data on maternal and child mortality and related indicators. Retrieved from https://animorepository.dlsu.edu.ph/etd_masteral/5136