Economic and Technology Management Review, Vol. 14(2019): 79 – 87

Repurposing the utilisation of data in Agrobiodiversity Information System

(Pengubahan kegunaan khusus data dalam Sistem Maklumat Agrobiodiversiti)

Muhammad Izzat Farid Musaddin*, Azuan Amron**, Elmaliana Albahari*, Mohamad Zulkifly Zakaria @ Mustafa*, Mohd Shukri Mat Ali*** and Faizah Patahol Rahman*

Keywords: data transformation, extract, transform, load, talend, column, row

Abstract

AgrobIS or Agrobiodiversity Information System is a repository system that was developed to store and manage data on genetic resources generated by studies conducted in MARDI. The repository contains data on PGRFA, livestock, biotechnology, arthropods and microbes. These data are not only important for conservation purposes and as a reference for future generations but also essential for developing or producing other systems such as dashboards. Expanding the use of these data to be implemented and integrated in other systems is important as it would highly benefit MARDI in the future. However, repurposing the data for newer decision making information system was difficult and problematic as the data in the database were not properly recorded, formatted and collated which impedes and delays the database querying and retrieval of required data during the data transformation process. Thus, this paper describes the steps taken to enhance the database query and retrieval times during the repurposing of data available in the AgrobIS system which includes the Extract, Transfer and Load (ETL) process and the use of a tool to accommodate the ETL process known as Talend Open Studio for Data Integration. Paddy data was specifically chosen for data transformation as it covered the most accessions available in the AgrobIS database compared to other categories of genetic resources.

Introduction

The Agrobiodiversity Information System, AgrobIS, is a repository system which was successfully developed in 2016 to solve the issues of data preservation and management of genetic resources which were collected based on research projects conducted at Malaysian Agricultural Research and Development Institute (MARDI). MARDI was given the task by the Ministry of Agriculture (MOA) to conduct research on most aspects of agriculture except for rubber, oil palm, cocoa, sago and aquaculture. This includes the conservation of local genetic resources which are made up of the vast biodiversity of plants, animals and microorganisms (McGuire and Qaulset 1986). The AgrobIS system was developed to assist in the conservation of biodiversity to store data on genetic resources from different categories namely, Plant Genetic Resources for Food and Agriculture (PGRFA), livestock, biotechnology, arthropods and microbes (Azuan et al. 2016). Each genetic resource residing in the database contains information that describes the accession numbers which are unique identifiers given to a material as it enters the

*ICT Management Centre, MARDI Headquarters, Persiaran MARDI-UPM, 43400 Serdang, Selangor
**Agrobiodiversity and Environment Research Centre, MARDI Headquarters, Persiaran MARDI-UPM, 43400 Serdang, Selangor
***Industrial Crop Research Centre, MARDI Headquarters, Persiaran MARDI-UPM, 43400 Serdang, Selangor
E-mail: izzatf@mardi.gov.my

collection (Matija 2015). For instance, users are able to view passport data of paddy accessions such as variety name, variety group, collection number and location of collecting site.

There are many reasons why it is important to conserve these genetic resources. One of the main reasons is to enable future generations to understand and acknowledge the fact that these genetic resources existed and were once discovered by researchers during their collection activities across the country. For instance, many local underutilised fruits are currently unknown to many Malaysians let alone by people around the world. One good example is the *ceri Terengganu* which is scientifically named *Lepisanthes fruticosa*. With the implementation of multi-disciplinary research approaches, data on new findings can be gathered and stored in the AgrobIS database where it can provide visibility to these rare and newly discovered species. Besides, the preservation of these genetic resources data will prevent institutions like MARDI from repeating the same field works or expeditions to gather accessions that have already been collected from past research activities. Thus, this research approach will save time and research funding when researchers want to decide on new project proposals and let them focus on areas that are yet to be explored.

A newer decision making information system can be developed from data available in the AgrobIS database. However, problems arise when trying to utilise the available data in the database to produce a newer decision-making information system such as the dashboard. One of the main issues with the AgrobIS database was its complex architecture containing 62 tables where data were not properly recorded, formatted and collated (Muhammad Izzat Farid et al. 2018). This condition impedes and delays the database querying and retrieval times of required data during the data transformation process. The challenges faced in repurposing the data in the AgrobIS database were to enhance database query and retrieval times. This paper describes the steps taken to enhance the database query and retrieval times during the repurposing of data available in the AgrobIS system in order to develop newer and improved decision-making information systems.

The AgrobIS database

A database is a storage system that encompasses a collection of information which is organised in order for easy access, management and update as defined by Margaret (2006). A database is an important medium to store information in relation to any organisation's scope of business as it is made traceable so that security measures can be applied to prevent any events of breach when data are stored on online servers. Many different techniques are used when designing and modeling a database. They differ from one another depending on the complexity of the data and the relationship between the data involved. The AgrobIS database stores data on different types of genetic resources in a unique manner. Each of these data are linked together by categorising them in a hierarchical approach known as a Tree Data Structure. As in other data structure such as the Maps and Sets, it forms a core element which is required in databases for quick search capability (Adrian 2018). By implementing this approach, a system will be able to retrieve data in a quick demeanour compared to a normal database structure. This is convenient for system users who could just browse through the database when acquiring specific data.

The AgrobIS database was designed as a tree data structure that only covers the tables storing the list of accessions rather than the entirety of the database including traits and characteristics of the record. The tables containing the accession figures are isolated from the characteristics table which impedes the data querying process and delays the retrieval of the data from the AgrobIS database. In view of repurposing the AgrobIS data to produce dashboards or

other similar systems, data transformation is required. The challenge was to enhance the speed of querying and retrieval of data from the AgrobIS database.

Data transformation is a process of converting the format or structure of data to a different format or structure as defined by Garrett (2018). Utilising the original database structure of the AgrobIS system would be a hassle for data retrieval activity which is affected by slow execution of the database query. Database volume also affects the query loading time as it takes time for the database engine to search for each preferred data. The completion of query time to access data for each genetic resource category differs from each other. The paddy data which has the highest amount of accessions stored in the AgrobIS database, requires a longer query time compared to other genetic resources. Applying direct query to the database including the Join statements for the merging of relevant tables will prolong the period of obtaining preferred data. In order to reduce the execution time to fetch the data, a new database structure should be imposed by disregarding the joining and the merging of the tables. Besides, the data transformation feature is not feasible enough to be used in a normal database management system such as the Navicat software. This is due to the method used to transform the data which involves the usage of database query script containing Join statements when merging the tables.

The Database Management System (DBMS) is the technology solution to cater for optimisation and management of storage and data retrieval from databases (Muhammad 2018). The Database Management System allows database management to be operated systematically as it provides user friendly interfaces which makes it easier for users when managing the databases. Storage optimisation and database management are activities that support the process of organising data that is stored in the databases. The more organised the data in a database, the faster the performance of the database which includes data retrieval from database to system applications. Minor tweaks can be conducted on selected database tables such as configuring the data type of any column from one type to another but it is insufficient to transform the data when dealing with a large set of data or complex database structures.

Methodology

The Extract, Transform and Load (ETL) process was used in repurposing the utilisation of data in the AgrobIS system. Another tool known as the Talend Open Studio for Data Integration was utilised to assist in the ETL process.

ETL process

Extract, Transform and Load (ETL) process is a backend process that is operated in a data warehouse environment (Aman and Jaiteg 2016). The main goal of the ETL process is often to produce a data warehouse after going through various phases of data extraction, data transformation and data loading. The structure and architecture of the developed data warehouse is in accordance with predefined standards and format during the early stages of an ETL process. Another reason for implementing the ETL process is to make sure that the data is well formatted and cleaned as there are existing tools which are available for download or purchase to support operations related to the ETL process. Data cleaning or cleansing is an important process which is a subset of the ETL process. Its goal is to make sure that data is correct, consistent and useable by identifying any errors or corruptions in the data, prior to error correction or deletion and, or manually processing these errors when needed to prevent its recurrence as suggested by Leo (2018).

The Extraction phase of the ETL commences once data from the multiple sources are predetermined and selected. *Figure 1* shows that data source is not only referring to operational databases but also

flat files such as Excel files or even plain text that contains data that are separated with commas or delimiters which are specifically called comma-separated values (CSV) files. After retrieving all data from the selected sources, transformation activity will ensue. Data transformation is important where data is configured according to a proper standard of database structure. It involves activities such as data cleansing and data massaging which in the end prepares for the loading of transformed data into a predefined database.

After completion of the transformation phase, the data is then loaded into targeted databases such as data warehouses and data marts. A data warehouses is a repository of enterprise or business databases which provide a clear picture of current historical operations of an organisation where all the transformed data resides in and the data volume is huge (Date 2003). A data mart on the other hand, focuses on a more specific approach where dimensions and measures represented as table columns are selected from a data warehouse or any other databases. Dimension is the term used to define the method of categorising data whereas measures are numerical values that quantifies the dataset as mentioned by Arden (2016). Moreover, a data mart can be a subset of data warehouse but not vice versa as there is a difference in the data volume.

Talend Open Studio for data integration

When performing an ETL process, it is advisable to use an ETL tool because it can help smoothen the extraction, transformation and data loading activities rather than using queries in a database management system directly to a specified database. The Talend Open Studio for Data Integration is a software tool which supports the execution of the ETL process. This tool was chosen in the present experiment because a community version of the tool is a free-to-use software that is easily accessible and downloaded via the web. To run an ETL process, some configurations must be predefined which includes a connection to the database server containing data source tables. This activity is necessary to notify the tool where the data resides in the server and which tables are to be used for data extraction. Furthermore, the input scheme of the source table must be set by specifying the database query based on the required data. *Figure 2* depicts a window

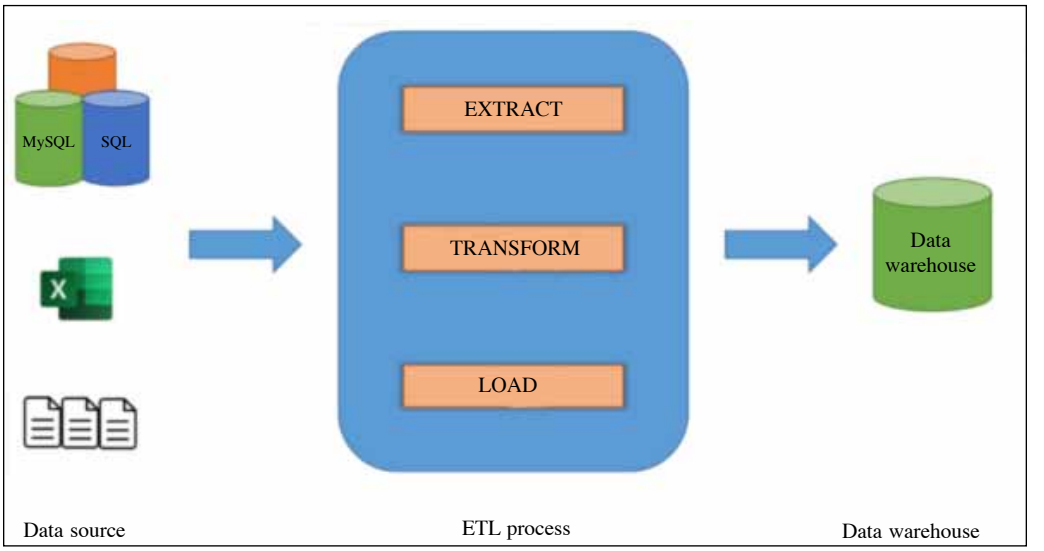

Figure 1. The ETL process

to key in compulsory information in order to connect the Talend Open Studio to a MySQL database.

Secondly, after connecting to the preferred source of database tables, a job must be designed. In Talend Open Studio, any activity from initialising a database connection to producing a well-designed data warehouse is considered a Talend Open Studio job. A Job can contain many components which interconnects with each other in order to complete a cycle of activities. In performing the ETL on data in a database, a Talend component called tMap is used to transform the data obtained once the connection is established. The tMap helps in removing any unwanted outliers from the database source. Besides, it can also automatically fill in table fields with default data. For instance, if there is no data stored in any field of a table, then any default data such as "Not Available" can be inserted straight away when executing a designed Talend job.

The AgrobIS database stores field values, characteristics and the accession numbers in separate tables. Besides, the data specific to field values are stored in different tables based on data types but having the same table fields. There are six database tables that are used to store the six types of values which includes data for dates, links, lists, numbers, pictures and texts. *Table 1* depicts the original database table structure which stores the data for the six different types of values. Each table has the same column name with the exact same data type for each column. To smoothen the data transformation process, data from the six database tables are consolidated into a new database table. This allows data querying from only one database table which results in a faster query time compared to queries with join statement. *Figure 3* shows the Talend Job that merges the data from dates, links, lists, numbers, pictures and texts tables. Upon completion of the consolidation

Table 1. Structure of the six database tables which stores data based on different value types

Table column
ID
Object_ID
Descriptor_ID
Value

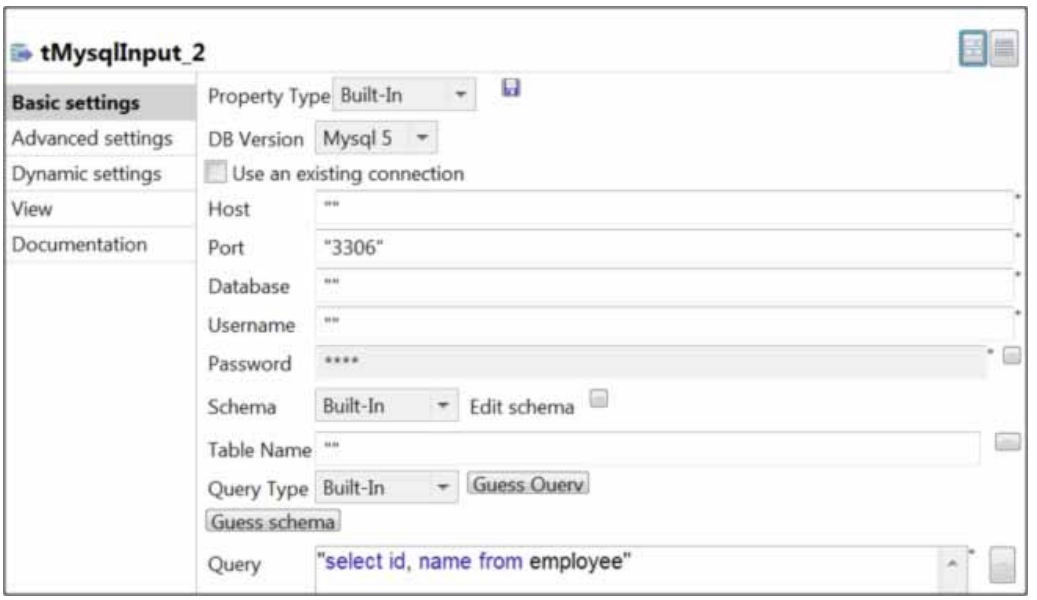

Figure 2. Form used to define connection to a database

activity, the newly created table is then further integrated with descriptor tables in AgrobIS database in order to map the paddy descriptor IDs with the paddy descriptor values prior to using it as a data source in a new Talend Job in order to fulfil the transformation process. This can be viewed in the Talend Job as shown in *Figure 4*.

The data transformation process does not end here, as the resulting data are illegible when viewed. Non-IT users will have a hard time in understanding when browsing through the data as the characteristics data are stored in a row format. Thus, a transformation of the database table structure to a columnar data format is required to solve the issue of readability.

The difference between row store and column store is that storing data in a row format enables easier transaction processing where it accelerates the process of writing data to databases. Storing data in a columnar data format on the other hand would be ideal for accumulating large volumes of data for a subset of columns as suggested by Rick (2016). The structure makes it difficult for users to comprehend, as it is hard to read them directly. Subsequent to the transformation of data and database table structure from a row format to a columnar approach, it becomes more readable to users where records targeted for transformation have a column on its own. In order to accomplish this activity, a Talend Job is designed to transform the table structure and data of paddy as shown in *Figure 5*.

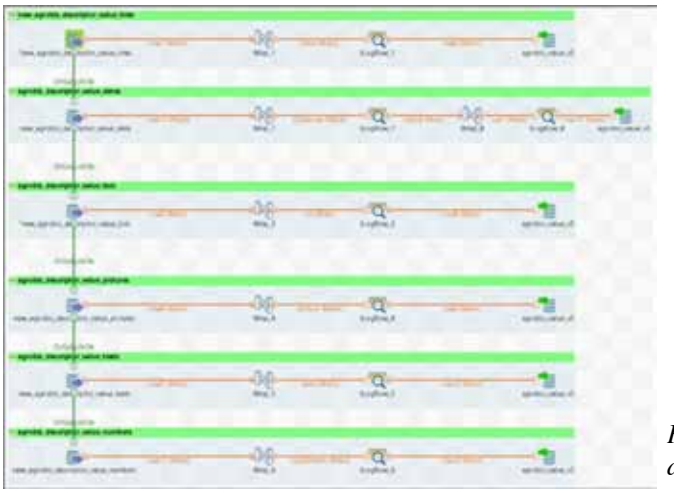

Figure 3. Talend Job to merge data of multiple tables

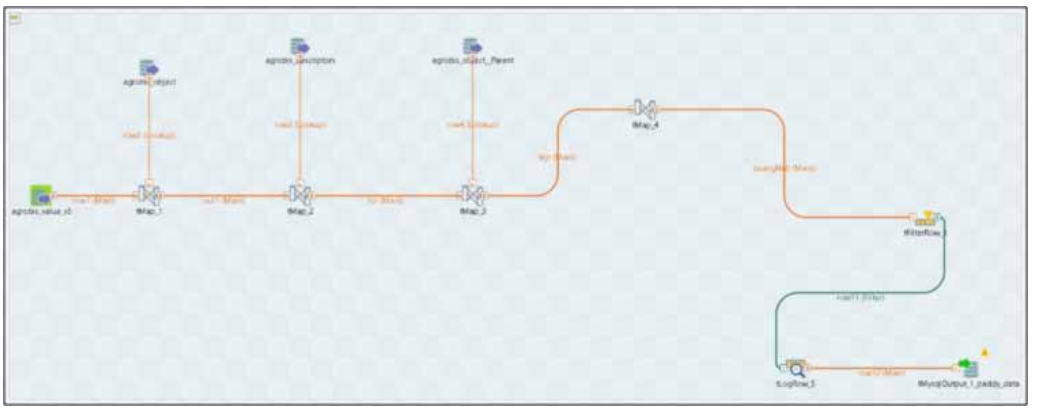

Figure 4. Talend Job to integrate newly created table with tables in AgrobIS database

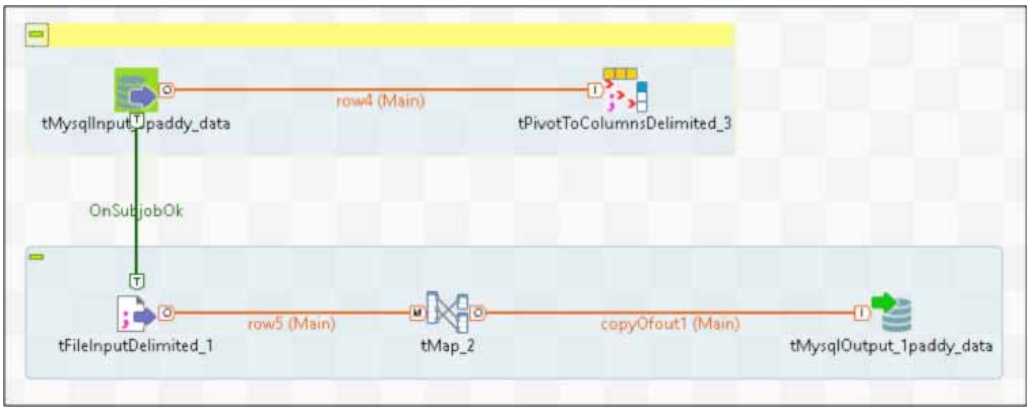

Figure 5. Talend Job to transform table structure and data of paddy

The data source, tMysqlInput used in the Talend Job is configured to filter the paddy data prior to utilising the tPivotToColumnDelimited component. This component enables the conversion of table structure from row format to a columnar view. The source data will be transmitted to a newly created Comma Separated Values (CSV) file according to the converted table based on columnar data format following execution of the pre-configured tPivotToColumnDelimited component. A CSV file is a plain text file that contains a list of data as suggested by Chris (2018). Subsequently, a metadata which is a prerequisite for implementing the tFileInputDelimited component is defined and established by taking the generated CSV file as data source to further support the transformation process. The full process can be seen in *Figure 5* which consequently creates a new MySQL table based on the standards and formats defined in the tMap component of the Talend job. Once the new table is generated, it can be used for multiple purposes like development of newer information systems such as the dashboards.

Results and discussion

Challenges arose when creating Talend Jobs involving the ETL process during data transformation. Utilising Talend components such as tPivotToColumnDelimited and tMap required repeated trials on how to implement the features in a one-job cycle. Many trials were conducted to remove errors that occurred during the process of building up the job, which required a large amount of time and effort to fix the issues, and errors encountered. Besides, the amount of records to be converted to the columnar data format for every genus data were big which averaged up to 130 fields. The conversion process became hectic where configuration is required for each field. The configuration was carried out on each field involving activities like removing outliers, converting data format to a new format and replacing null values with default values. These activities required the use of Talend components such as tMap. For instance, when using a tMap component, formulas were designed for the filtering process which were necessary for the enforcement on each record in order for data transformation to be successful.

The repeated trials conducted during the Extract, Transform and Load (ETL) process showed that it had the ability to transform data and produced new database tables that were useful for future systems especially dashboard systems. The database table can be directly used as a source table in the form of data mart for the development of many new analysis-based dashboard systems. Following the transformation of the database table structure into a columnar data format, the number of records

decreased by a huge amount compared to the original database in AgrobIS. The new columnar format enhanced the speed of querying process of the newly created database. A sample list of records that was successfully transformed to a columnar data format is shown in *Table 2*.

The newly transformed data were of better quality due to the fact that proper formats and standards were implemented on the data and all outliers had been removed. Besides, all table fields or columns contained data without leaving any rows with null values or the state of being empty.

Table 2. Sample list of columns generated after data transformation

Table columns
Accession Number
Acquisition Date (ACQDATE)
Age of Seeds Collection
Alkali Tolerance (ALK)
Elevation/Altitude (m)
Amylose (AMY)
Ancestral Data (ANCEST) - pedigree
Collecting/Acquisition Source (COLLSRC)
Collecting Date of Sample (COLLDATE)
Collecting Institute Address (COLLINSTADDRESS)
Collecting Institute Name (COLLNAME)
Collection Number (COLLNUMB)
Collector Name (COLLECTOR)
Latitude
Location of Collecting Site (COLLSITE)
Longitude
MLS Status of the Accession
Remarks
Registration Date
Seed Source
Variety Group (VG)
Variety Name

Conclusion

In conclusion, the Extract, Transform and Load (ETL) process with the help of Talend Open Studio for Data Integration was shown to successfully transform the data in AgrobIS database by producing new database tables that would be useful for future systems especially the dashboard systems. Consequently, this has significantly enhanced the speed of querying process for the newly created database which follows a new columnar format. Thus, newer decision making information systems such as dashboards can be developed in the future.

References

Adrian M. (2018). Tree data structures for beginners. Retrieved on 9 May 2019 from https://adrianmejia.com/blog/2018/06/11/data-structures-for-beginners-trees-binary-search-tree-tutorial/

Aman, P.S.P. and Jaiteg, S. (2016). ETL methodologies, limitations and framework for the selection and development of an ETL tool. *International Journal of Research in Engineering and Applied Sciences* 6(5)

Arden, M. (2016). Dimension vs. measure: What's the difference? Retrieved on 1 July 2019 from https://yseop.com/blog/dimension-vs-measure-whats-the-difference/

Azuan, A., Muhammad Izzat Farid, M., Mohd Shukri, M.A., Faizah, P.R., Muhammad Luqman Hakim, M.F., Mohd Saifuddin, I. and Rusli, A. (2016). Sistem Pengkalan Data Agrobiodiversiti (AgrobIS). Ver. 2: Sejarah dan Sumbangan National Agrobiodiversity Conference 2016, Kuala Terengganu.

Chris, H. (2018). What is a CSV file and how do I open it? Retrieved on 12 June 2019 from https://www.howtogeek.com/348960/what-is-a-csv-file-and-how-do-i-open-it/

Date, C. (2003). Introduction to database systems, 8th ed., Upper Saddle River, N.J.: Pearson Addison Wesley

Garrett, A. (2018). What is Data transformation? Retrieved on 16 May 2019 from https://www.alooma.com/blog/what-is-data-transformation

Leo, G. (2018). Six steps for data cleaning and why it matters. Retrieved on 12 June 2019 from https://www.geotab.com/blog/data-cleaning/

Margaret, R. (2006). Database (DB). Retrieved on16 May 2019 from https://searchsqlserver.techtarget.com/definition/database

Matija, O. (2015). Accession Passport Data Basics. Retrieved on15 May 2019 from https://www.genesys-pgr.org/doc/0/basics

McGuire, P.E. and Qaulset, C.O. (1986). Excerpt from "Genetic Resources Conservation Programme Annual Report 1985 – 1986. Report No. 1. Unversity of California Genetic Resources Conservation Programme, Davis, Ca. 30 p.

Muhammad Izzat Farid M., Faizah P.R., Site N.A.R., Mohd Shukri M.A., Azuan A. and Mohd Shafiq, A. (2018). Development of dashboard to identify rice germplasm in MARDI Genebank. *Economic and Technology Management Review*. 13: 111 – 120

Muhammad, R. (2018). Introduction to Database Management Systems (DBMS). Retrieved on 16 May 2019 from https://www.bmc.com/blogs/dbms-database-management-systems/

Rick, G. (2016). Row store and column store databases. Retrieved on 11 June 2019 from https://www.percona.com/blog/2016/12/14/row-store-and-column-store-databases/

Abstrak

AgrobIS atau Sistem Maklumat Agrobiodiversiti merupakan satu sistem repositori yang dibangunkan untuk menyimpan dan mengurus data sumber genetik yang dihasilkan daripada penyelidikan yang dijalankan di MARDI. Sistem repositori tersebut mengandungi data berkaitan PGRFA, ternakan, bioteknologi, artropod dan mikrob. Data ini bukan sahaja penting untuk pemuliharaan dan sebagai rujukan kepada generasi akan datang tetapi juga penting untuk membangunkan atau menghasilkan sistem pembuat keputusan yang baharu. Perluasan penggunaan data ini untuk dilaksanakan dan diintegrasikan ke dalam sistem lain adalah penting kerana ia akan memberi manfaat kepada MARDI pada masa hadapan. Namun, penggunaan semula data sedia ada yang terdapat dalam sistem AgrobIS bagi tujuan pembangunan sistem pembuat keputusan baharu yang lebih kuasa adalah susah dan bermasalah kerana data dalam sistem AgrobIS tidak direkod, diformat atau dikumpul semak dengan sempurna. Keadaan ini akan melengah dan menghalang pertanyaan pangkalan data dan tempoh penerimaan kembali data yang diperlukan dalam proses transformasi data. Artikel ini menerangkan langkah-langkah yang diambil bagi memperkasakan pertanyaan pangkalan data dan tempoh penerimaan kembali data yang diperlukan semasa proses penggunaan semula data dalam sistem AgrobIS seperti peggunaan proses *Extract, Transfer and Load (ETL)* dan alat *Talend Open Studio for Data Integration*. Data padi dipilih secara khusus untuk aktiviti transformasi data kerana ia mengandungi paling banyak aksesi dalam pangkalan data AgrobIS berbanding dengan kategori sumber genetik yang lain.