Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Extract, Transform, And Load (ETL) Technique for Pre-Processing of Agricultural Pest Dataset

J. Cruz Antony [1], J. Refonaa [2], S. L. Jany Shabu [3], S. Dhamodaran [4], P. Asha [5]

[1]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
jcruzantony@gmail.com

[2]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
refonna.cse@sathyabama.ac.in

[3]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
janyshabu.cse@sathyabama.ac.in

[4]Department of Information Technology,
Sathyabama Institute of Science and Technology, Chennai, India
dhamodaran.cse@sathyabama.ac.in

[5]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
asha.cse@sathyabama.ac.in

Article Info
Page Number: 595 - 604
Publication Issue:
Vol 71 No. 3s2 (2022)

Abstract
Data pre-processing the preliminary data mining procedure is taking a new dimension, the Extract, Transform, and Load (ETL) based data pre-processing. Data pre-processing is a crucial step in knowledge discovery through data mining which shapes the raw data by cleaning, integrating, and transforming using predefined techniques. Studying the pest population dynamics was never easy, as the pest population was mostly correlated with the abiotic and biotic features for knowledge discovery and forecasting the occurrence of the pest in crops. ETL technique, a Data Warehouse concept was chosen for data pre-processing, because of its fast and efficient handling in regarding with the huge dataset, also the dataset is heterogeneous from five major districts of Maharashtra which includes the pest population on the crop for five years along with its respective abiotic features. Talend Open Studio (TOS) was used to design an ETL job for performing data extraction, integration, and discretization. The designed ETL job has exhibited good performance and accuracy in pre-processing the pest population dataset. This paper will provide an insight into building an ETL tool using Talend Open Studio and review the issues of data pre-processing in the field of agricultural pest management.

Article History
Article Received*: 28 April 2022*
Revised*: 15 May 2022*
Accepted*: 20 June 2022*
Publication*: 21 July 2022*

Keywords: - Data Warehouse, ETL, Pest population dataset, Pre-processing, Talend Open Studio

INTRODUCTION

Data pre-processing is often given less importance, but an important preliminary procedure in the data mining process. Heterogeneous raw data are highly susceptible to noise and inconsistency, thus resulting in a poor quality of data mining results [1]. To improve the data quality and consequently increase the quality of the results, the raw data must be pre-processed to make it simple and efficient for further analysis. Data Warehouse (DW) tools are considered, devoted to analytical processing and report generation. These tools are employed to assist the activities of decision-making in modern business settings [2]. Several modern business firms are having a rich number of data, but relatively poor meaningful information for business strategy [3]. A well-developed DW tool can steadily increase any firm's ability for decision-making. It holds consolidated historical data as a warehouse that supports a firm to understand different business scenarios and take decisions accordingly [4]. In earlier days during the introduction of technology, the DW is considered a critical process particularly in constructing the model, and also the cost for implementation of DW was very expensive [5]. However, nowadays the cost has gradually become less and has become a tool that is a mandate for business firms [6].

Building a data warehouse for precision agriculture has become the most important foundation for developing a crop intelligence platform that will facilitate resource-effective agronomy recommendations and decision-making [7]. Data are identified from different sources, extracted, transformed, and loaded into the DW. After this process, the dataset is mined using OLAP (online analytical processing) tools for providing insights about the dataset [8]. As a result of technological development, farmers can gain knowledge on pest precautionary control measures, trends of decline in post-yield losses, identifying enhanced ways to access markets, etc.

In the DW environment, the process of ETL-Extract, Transform, and Load retrieves data from various sources (extract), changing it in accordance with the pre-processing technique (transform) and uploading it into a data warehouse system (load). The ETL tools are employed to pre-process voluminous and heterogeneous raw information and mainly focus on the performance [9] [10]. Initially, the first software of ETL was crude, but it has matured rapidly and can pre-process almost all types of data [6] [11]. Nowadays, ETL tools can perform fast and efficiently on complex datasets in a hassle-free manner. Many open source ETL tools are now available, which provides easy to use graphical user interface (GUI) which is an advantage for non-programmers [12]. Talend Open Studio (TOS) is the first software with open-source data integration, which can carry out the process of extracting, transforming, and loading data warehouses [13]. Talend Open Studio opens up the marketplace for integration and transformation to the entire consumers, regardless of the needs of data integration and size [14]. The ETL process for the planned study is depicted in "Fig. 1".

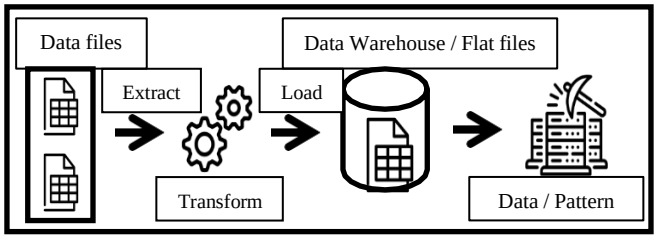

Fig. 1. Proposed Extract, Transform, and Load process

The "Extract" segment performs extracting data from multiple sources such as the pest population data from the field crop and the abiotic features of the respective field crop region. Both these datasets have been extracted and placed in a single warehouse using built-in components from the extract module. The "Transform" segment is very crucial since it performs data pre-processing by transforming the continuous abiotic features into categorical counterparts and also classifying pest incidence data as low, high, and medium based on the Economic Threshold Level. Finally, the "Load" segment performs the loading of the processed dataset into a flat file for further mining of data through prediction models for obtaining a better decision-making system for Integrated Pest Management.

II. Literature Review

Agriculture is the backbone of the economy for various developing countries. It is the principal income source especially in rural places of a country [15]. Agricultural crops are often attracted by various insect pests resulting in crop damage and loss. These insect pests' population has to be evaluated to manage the infestation on the crops [16]. Predicting the outbreaks of the insect pest population is the primary challenge for agriculture. Study on population dynamics of individual pest species using ecology model is mandatory [17]. Due to the global warming and climate change scenarios, abiotic features like rainfall, temperature, and humidity play a vital role in governing the pest population dynamics [18], [19]. A pest population model is a powerful tool in evaluating the correlation between the pest population and the influencing abiotic features which is a part of integrated pest management (IPM) [16].

A data warehouse is an archive of data retrieved from various sources such as the abiotic features from the weather monitoring station and the pest population from the agricultural land, accumulated and compiled to provide a decision support system via Online Analytical Processing (OLAP) [20]. Nowadays the growth of data especially in the field of agriculture, industry, and transport is increasing and the need for the application to manage data pre-processing and integration of big data is very challenging. Extract transform load (ETL) tool is the solution to efficiently process and extract data with high precision and low storage space in a moderate execution time [21]. ETL tool is used to integrate heterogeneous data and construct a data warehouse to provide an exhaustive data foundation for data analysis [22].

Sitanggang et al. built a SOLAP-Spatial Online Analytical Processing for Indonesia Agricultural Commodity. The SOLAP system is connected to a data warehouse that governs past data of agricultural commodities including food crops, horticulture, livestock, and plantations. The ETL module in the SOLAP system was developed to provide information on agricultural commodities to the users. Pentaho Kettle was used to develop the proposed ETL technique to fetch data from the Agricultural Statistics Database governed by the Ministry of Agriculture, Indonesia. The proposed ETL module has been efficient and all the jobs developed within the module were successfully implemented [23].

Binmonte et al. analyzed spatial datasets of agricultural farm energy use performance using technologies viz., SOLAP-Spatial Online Analytical Processing, SDW-Spatial Data Warehouse, and geo business intelligence (GeoBI). The SOLAP system uses the Spatial ETL tool for extracting data from spatial RDBMS, transformed and loaded into the spatial data warehouse. The ETL tool used here was Talend Open Studio with a geospatial facility for better performance [24].

Xu et al. proposed an approach for establishing an AEEIS- Agricultural Ecosystem Enterprise Information System based on integrated information systems. The proposed system extracts data on the

plantation, terrain, etc., and accommodates it for ecosystem management in agriculture. The ETL sub-system in AEEIS was designed to extract data from multiple sources, transform the original data structure to analytical data, and then integrated it into the data warehouse. The users are allowed to perform ETL operations using ETL dynamic link lib (DLL) and also can query the result of the ETL process in graphical output [25].

Abdullah used Agriculture Decision Support System (ADSS) which was developed as an Online Analytical Processing tool shortly called ADSS-OLAP, to understand the mealybug population on cotton. The manual ETL process was used to extract, transform and load into the data warehouse [26].

Tripathy et al. proposed a framework for the pest management system using geospatial data mining techniques integrated with various agricultural parameters to provide pesticide usage and pest management information. In the ETL module, both aspatial and spatial data were extracted and transformed to match operational needs and finally loaded the processed data into the data warehouse. A manual approach using JAVA programming language and PostgreSQL for the database was made to design and develop the ETL module [27].

The ETL tools using GUI are popular in the market due to their flexibility in user interface and have improved performance compared to the other manual ETL tools [28]. These modern ETL tools emerge as the need for business intelligence is growing and the ETL framework are capable of integrating and consolidating data with a good performance index [29].

The ETL process performed for agricultural data especially in the field of pest population dynamics is scarce, as per Sitanggang et al and Binmonte et al the GUI based open-source ETL tools *viz*., Pentaho Kettle, and Talend Open Studio (TOS) has always been performed well when compared to the traditional manual ETL method. Hence the open-source TOS is adopted in this study.

III. Materials and Methods

The dataset includes field pest scout of *Gesonia gemma* Swinhoe incidence on soybean crop from various districts in Maharashtra namely Nagpur, Akola, Amravati, Wardha, and Yavatmal along with its respective abiotic features like rainfall (mm), minimum temperature (◦c), relative humidity (%), maximum temperature (◦c), soil moisture index (%), moisture adequacy index (%), and the number of rainy days in a week. The sample records from the dataset are displayed in Table 1. *G. gemma* is a regular defoliator of soybean crop that causes huge loss to the farmers by reducing the grain weight and yield [30]. Population dynamics study of *G. gemma* was made with the aid of data warehouse and machine learning techniques.

Table. 1. Sample records from the dataset

Pest Incidence	Week	Crop Stage	MaxT (°C)	MinT (°C)	RH (%)	MAI (%)	SMI (%)	RF (mm)	RFD
0.0	27	1	32.8	24.8	90.8	100	67	80	4
0.0	28	1	30.0	24.0	91.5	100	100	224	5
1.1	29	1	27.3	23.8	91.3	100	100	106	5
1.3	30	1	30.5	23.5	86.1	100	100	38	2
2.9	31	2	33.2	24.0	76.7	95	89	0	0
7.7	32	2	32.5	24.8	80.8	100	90	3	1
13.7	33	3	31.9	24.7	82.9	94	85	20	1
11.6	34	3	31.0	23.0	93.0	100	100	37	2
6.1	35	4	31.6	23.7	91.4	100	100	167	4
1.9	36	4	33.0	23.8	91.7	100	100	117	5
0.5	37	4	32.7	23.0	82.9	100	98	21	1
0.0	38	4	35.3	22.9	78.6	89	83	0	0
0.0	39	4	35.8	23.7	77.1	92	78	18	2
0.0	40	4	32.4	24.2	90.8	100	94	28	1
0.0	41	4	33.6	19.7	77.1	88	82	0	0

Talend Open Studio (TOS) is a handy tool that shortens the time required to pre-process and integrate the data [31]. The jobs are created by using the predefined components available, instead of coding individually. TOS comes with over 800+ pre-built components, a component is a functional piece that performs a single operation. A component consists of an XML (Extensible Markup Language) descriptor file which contains the component definition information i.e., the function of the component and the way of interaction with other components, etc; a message properties file which contains the information about the component label displayed in component properties; a java template file which contains templates that generate output from the model such as Structured Query Language (SQL), XML, Text, etc; and a component icon which contains a 32*32 size Portable Network Graphics (PNG) image displayed in the palette to represent the component.

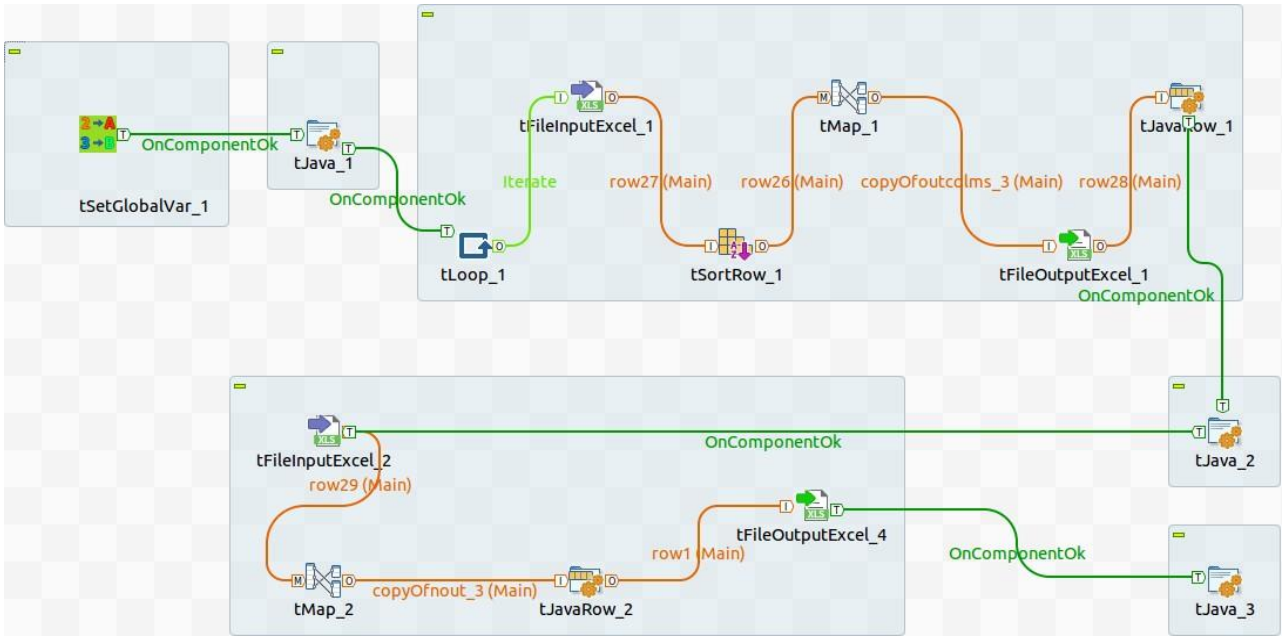

Fig. 2. Talend Open Studio job for performing ETL job

The primary objective of TOS here is to extract heterogeneous data *viz., G. gemma* incidence on soybean from five major districts of Maharashtra in spreadsheet file (xlsx) and the abiotic features of the respective districts in another xlsx file; then integrate it into a single xlsx file and pre-process the data by applying discretization method and classification. The Max-diff method was proposed to convert the continuous abiotic features into discrete categories. Usually, the traditional binning methods like equal frequency/equal width are used for converting the continuous abiotic features, but the Max-diff discretization method was chosen over traditional binning methods because it categorizes datasets with non-superfluous in nature. Max-diff performs better in predictive accuracy terms when applied to any kind of evaluation. It could also provide certain data that the usual econometric examination is not capable of managing [32]. The classification of pest population data as low, high, and medium were made based upon the Economic Threshold Level using standard week data of the soybean crop [33]. The ETL job designed in this study using various built-in components and custom java code is displayed in "Fig. 2".

The tSetGlobalVar_1 is a pre-built component used to set global variables used across the job, it consists of two columns, the key column which contains the name of the variables, and the value column which contains the value assigned to these variables. The tJava_1 is a custom component that contains java code that was used to extract xlsx file sheets and write the sheet names into a list. The tLoop_1 component sends the sheets available in the list one by one to the tFileInputExcel_1 component, which reads the data from the xlsx sheet and stores it into a separate xlsx sheet. Then the tSortRow_1 component was used to sort the columns which are to be transformed using the Max-diff discretization method in ascending order and the tMap_1 applies the first step (i.e. finding the difference between consecutive tuples) of the Max-diff discretization method for the columns. The tFileOutputExcel_1 component outputs the data from the tMap_1 component into a new xlsx file with the current sheet name. The tJavaRow_1 is a built-in component that allows entering customized code which was used to add the difference between consecutive tuples columns into a separate list. Again tJava_2 component was used to sort the difference between consecutive tuples columns and according to the Max-diff discretization method, five maximum values among the column were chosen. Again the modified data was brought to a new xlsx file using the tFileInputExcel_2 component. The tMap_2 component here assigns the transformation values over the columns such as A1, A2, A3, A4, and A5. Thus the Max-diff discretization method was successfully applied to convert the continuous abiotic feature columns into categorical counterparts. Next, the tMap_2 component creates a separate column for pest incidence data and classifies the data as low, high, and medium in support of the Economic Threshold Level and finally updates into the new column. After completing all the transformation procedures, the tJavaRow_2 component reads the transformed data as a separate file and then hands it over to the tFileOutputExcel_4 component which writes the data into a sheet of the xlsx output file. Finally, the tJava_3 components reset all the global variables assigned to the job.

The xlsx (pre-processed) file obtained from the designed ETL job is ready to be loaded into the Naïve Bayesian classification model. The Naïve Bayes model is easy to build especially for large datasets and is said to outperform other advanced machine learning models [34]. The Naïve Bayesian classification model outperforms the Decision tree model to study the frequency characterization of Mango pulp weevil in different adult stages [35]. An expert knowledge system

to detect papaya disease was developed using Fuzzy reasoning and the Naïve Bayesian classifier was used in the expert system for disease classification which displayed better accuracy [36]. Mostly, the Naïve Bayesian model pursues two stages. The first stage is the training stage, where the training dataset was used to construct the NB machine learning model using the concept of probability. In the next stage, the trained ML model was used to interpret the target feature of the new test dataset. Thus, the dataset was split into 80% of training data and 20% of test data. The successfully built Naïve Bayesian ML model for predicting *G. gemma* incidence on soybean is evaluated and discussed in the next section.

IV. Results and Discussion

Talend Open Studio is a freeware data integration software used to perform the Extract, Transform, and Load job for transforming and integrating the dataset especially the abiotic features using discretization techniques. To examine the performance of the pre-processed dataset, the dataset was classified into training and test datasets then Naïve Bayesian classification was used on the training dataset and thereafter predicted its respective test dataset using the trained Naïve Bayesian model for unprocessed, Equal frequency, Equal width, and Max-diff models [37]. To verify the performance of the classification models, a confusion matrix was generated and from which the classification accuracy was estimated.

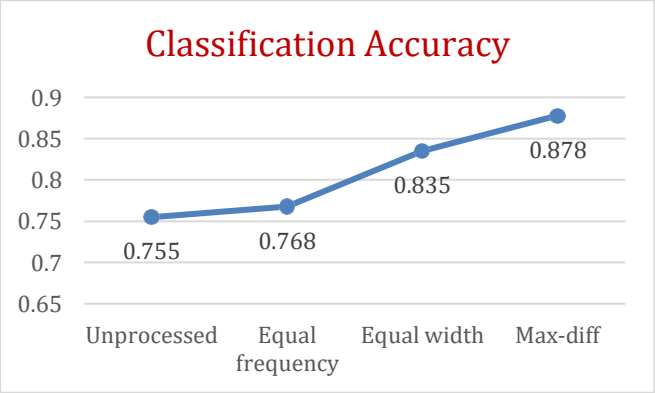

Fig. 3. Performance of unprocessed and pre-processed classification models

The graph displayed in "Fig. 3" makes it clear that the Naïve Bayesian model which engaged pre-processed dataset using the Max-diff discretization method outruns the other three classification models which engaged unprocessed dataset, Equal frequency and Equal width discretized dataset. Thus, it is evident that Max-diff has several methodological benefits when compared to other traditional methods.

IV. Conclusion

In the study of *Gesonia gemma* population dynamics on soybean crops, data quality plays an important role as better data quality improves the accuracy of decision-making capabilities. Data pre-processing techniques were used over the dataset to remove inconsistent and noisy data, thereby increasing the data efficiency for further mining of information. Once the incoming dataset is retrieved from different sources, it is transformed using a proper technique that can be loaded to a non-volatile medium referred to as the data warehouse. The process of Extract, Transform, and Load (ETL) is considered a major hub for a data warehouse. The ETL process has the power to

transform the non-trivial process of retrieval of hidden, essentially useful, and previously unknown data from a huge dataset. The proposed ETL job developed using Talend Open Studio has exhibited good performance and also offers ease of use. The classification model which uses the pre-processed dataset of the Max-diff discretization method has shown better accuracy when compared with the unprocessed dataset model and traditional equal frequency model in classification and to identify the population dynamics of the *G. gemma* incidence on soybean crop.

V. Acknowledgement

The dataset utilized in this work was collected from the CROPSAP programme, Commissionerate of Agriculture, Government of Maharashtra, India.

References

[1] J. Han, and M. Kamber, “Data mining: concepts and techniques,” 2nd edn. Morgan Kauffman Publishers, San Francisco 2006.

[2] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, S. Paraboschi, and P. Milano, “Designing data marts for data warehouses,” ACM Transactions on Software Engineering and Methodology, vol. 10, no.4, pp.452–483, 2001, doi:10.1145/384189.384190

[3] J.A. Hoffer, M.B. Prescott, and F.R. McFadden, “Modern Database Management,” Pearson Education, Inc., New Jersey 2007.

[4] S. Chaudhuri, U. Dayal, and V. Ganti, “Database technology for decision support systems,” Computer (Long Beach, Calif.), vol. 34, no.12, pp. 48–55, 2001, doi:10.1109/2.970575

[5] N. Vijayendra, and M. Lu, “A web-based ETL tool for data integration process,” In: 6th International Conference on Human System Interaction (HIS), pp. 434–438. IEEE, Poland 2013, doi:10.1109/HSI.2013.6577861

[6] W.H. Inmon, “Building the data warehouse,” Wiley Computer Publishing, New York 2005.

[7] V.M. Ngo, N.A. Le-Khac, and M.T. Kechadi, “An Efficient Data Warehouse for Crop Yield Prediction,” in Proceedings of the 14th International Conference on Precision Agriculture, pp. 1-12. Montreal, Quebec, Canada 2018, https://arxiv.org/abs/1807.00035v1

[8] S. Nilakanta, and K.P. Scheibe, “The Digital Persona and Trust Bank: A Privacy Management Framework,” Journal of Information Privacy and Security, vol.1, no.4, pp. 3-21, 2005, doi: 10.1080/15536548.2005. 10855777

[9] Gowri, S., J. Jenila, Bathula Sowmya Reddy, and M. Antony Sheela. "Scrutinizing of Fake News using Machine Learning Techniques." In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 223-227. IEEE, 2021.

[10] J. Li, J. Wang, and H. Pei, “Data Cleaning of Medical Data for Knowledge Mining,” Journal of Networks, vol. 8, no.11, pp. 2663–2670, 2013, http://dx.doi.org/10.4304/jnw.8.11.2663-2670

[11] T.A. Majchrzak, T. Jansen, and H. Kuchen, “Efficiency evaluation of open source ETL – tools,” in Proceedings of 2011 ACM Symposium on Applied Computing, pp. 287-294. Association for Computing Machinery, New York, US 2011, doi:10.1145/1982185.1982251

[12] T. Ghadiyali, K. Lad, and B. Patel, “ETL techniques and challenges in agriculture intelligence,” in Proceedings of Agro-Informatics and Precision Agriculture (AIPA), pp. 85-91. Allied Publishers, New Delhi, India 2012, ISBN: 9788184247725, 8184247729

[13] Narmatha, P., and S. Gowri. "Detection of Human Facial Expression Using CNN and

Deployment in Desktop Application." In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1187-1192. IEEE, 2021.

[14] N. Biswas, A. Sarkar, and K.C. Mondal, “Efficient incremental loading in ETL processing for real-time data integration,” Innovations in Systems and Software Engineering, vol. 16, pp. 53–61, 2020, doi:10.1007/s11334-019-00344-4

[15] S. Ranganathan, “Loading Data Using Talend as an ETL,” OneGlobe LLC, blog, August 20, 2019 [Online]. Available: https://www. oneglobesystems.com/blog/loading-data-using-talend-as-an-etl

[16] W.H. Inmon, “The Evolution of Integration,” [White Paper]. Available: https://www.keyinfo.com/wp-content/uploads/2015/04/ The-Evolution-of-Integration-by-W.-H.-Inmon.pdf

[17] V.K. Shankarnarayan, and H. Ramakrishna, "Paradigm change in Indian agricultural practices using Big Data: Challenges and opportunities from field to plate," Information Processing in Agriculture, Vol. 7, No. 3, pp. 355-368, 2020, doi: 10.1016/j.inpa.2020.01.001

[18] T.A. Rand, C.E. Richmond, and E.T. Dougherty, "Modeling the combined impacts of host plant resistance and biological control on the population dynamics of a major pest of wheat," Pest management science, Vol. 76, No. 8, pp. 2818-2828, 2020, doi:10.1002/ps.5830

[19] J.G. Illán et al, "Landscape structure and climate drive population dynamics of an insect vector within intensely managed agroecosystems," Ecological Applications, Vol. 30, No. 5, e02109, 2020, doi:10.1002/eap.2109

[20] U. Naeem-Ullah et al, “Insect pests of cotton crop and management under climate change scenarios,” in Environment, climate, plant and vegetation growth, S. Fahad et al., Eds. Cham, Switzerland: Springer, 2020, pp. 367-396.

[21] M.M. Phophi, P. Mafongoya, and S. Lottering, "Perceptions of climate change and drivers of insect pest outbreaks in vegetable crops in Limpopo province of South Africa," Climate, Vol. 8, No. 2, 27, 2020, doi:10.3390/cli8020027

[22] S. Nilakanta, K. Scheibe, and A. Rai, "Dimensional issues in agricultural data warehouse designs," Computers and Electronics in Agriculture, Vol. 60, pp. 263-278, 2008, doi: 10.1016/j.compag.2007.09.009

[23] B.B. Semlali, C.E. Amrani, and G. Ortiz, "SAT-ETL-Integrator: an extract-transform-load software for satellite big data ingestion," Journal of Applied Remote Sensing, Vol. 14, No. 1, 018501, 2020, doi: 10.1117/1.JRS.14.018501

[24] J. Chen, S. He, and X. Li, "A Study of Big Data Application in Agriculture," Journal of Physics: Conference Series, Vol. 1757, 012107, 2021, doi:10.1088/1742-6596/1757/1/012107

[25] S. Sitanggang, R. Trisminingsih, F. Fuady and H. Khotimah, "Extract, Transform, Load Module in SOLAP for Indonesia Agricultural Commodity," 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 2018, pp.101-105, doi:10.1109/SIET.2018.8693152.

[26] S. Bimonte, M. Pradel, D. Boffety, A. Tailleur, G. André, R. Bzikha, and J. Chanet, “A New Sensor-Based Spatial OLAP Architecture Centered on an Agricultural Farm Energy-Use Diagnosis Tool,” Int. J. Decis. Support Syst. Technol., vol. 5, no. 4, pp.1–20, 2013, doi:10.4018/ijdsst.2013100101

[27] Ancy S, Kumar R, Ashokan R, Subhashini R ,"Prediction of onset of south west monsoon using multiple regression",Proceedings of ICCCS 2014 - IEEE International Conference on

Computer Communication and Systems, 2014, pp. 170–175, 7068188

[28] Sindhu K, Subhashini R, Gowri S, Vimali JS, "A Women Safety Portable Hidden camera detector and jammer", Proceedings of the 3rd International Conference on Communication and Electronics Systems, ICCES 2018, 2018, pp. 1187–1189, 8724066

[29] L. Xu, N. Liang and Q. Gao, "An Integrated Approach for Agricultural Ecosystem Management," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 4, pp. 590-599, July 2008, doi: 10.1109/TSMCC.2007.913894.

[30] A. Abdullah, "Analysis of mealybug incidence on the cotton crop using ADSS-OLAP (Online Analytical Processing) tool," Comput. Electron. Agric., vol. 69, no. 1, pp. 59–72, 2009, doi: 10.1016/j.compag. 2009.07.003

[31] Ajitha, P., Lavanya Chowdary, J., Joshika, K., Sivasangari, A., Gomathi, R.M., "Third Vision for Women Using Deep Learning Techniques", 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315196

[32] A Sivasangari, P Ajitha, RM Gomathi, "Light weight security scheme in wireless body area sensor network using logistic chaotic scheme", International Journal of Networking and Virtual Organisations, 22(4), PP.433-444, 2020

[33] RM Gomathi, JML Manickam, A Sivasangari, P Ajitha,"Energy efficient dynamic clustering routing protocol in underwater wireless sensor networks", International Journal of Networking and Virtual Organisations, Vol.22,4 pp. 415-432

[34] Kanyadara Saakshara, Kandula Pranathi, R.M. Gomathi, A. Sivasangari, P. Ajitha, T. Anandhi, "Speaker Recognition System using Gaussian Mixture Model", 2020 International Conference on Communication and Signal Processing (ICCSP), pp.1041-1044, July 28 - 30, 2020

[35] A. K. Tripathy, J. Adinarayana and D. Sudharsan, "Geospatial data mining for Agriculture pest management - a framework," 17th International Conference on Geoinformatics, 2009, pp. 1-6, doi: 10.1109/GEOINFORMATICS.2009.5293296.

[36] Patel M., Patel D.B. (2021) Progressive Growth of ETL Tools: A Literature Review of Past to Equip Future. In: Rathore V.S., Dey N., Piuri V., Babo R., Polkowski Z., Tavares J.M.R.S. (eds) Rising Threats in Expert Applications and Solutions. Advances in Intelligent Systems and Computing, vol 1187. Springer, Singapore. https://doi.org/10.1007/978-981-15-6014-9_45

[37] C. Xavier and F. Moreira, "Agile ETL," Procedia Technology, Vol. 9, pp. 381-387, 2013.

[38] H.K. Verma, R. Verma, A.R. Azad, and R. Aggarwal, "Impact of climatic factors on the incidence of soybean leaf feeders," Journal of Pharmacognosy and Phytochemistry, Vol. 9, No. 3, pp. 577-580, 2020.

[39] J. Bowen, "Getting started with Talend Open Studio for Data Integration," Packt Publishing Ltd, Open Source 2012. ISBN: 9781849514729

[40] Sawtooth Software, Inc., "The MaxDiff System Technical Paper," October 2020 [Online]. Available: https://sawtoothsoftware.com/ resources/technical-papers/maxdiff-technical-paper

[41] J.C. Antony, M. Pratheepa, "Study of population dynamics of soybean semi-looper *Gesonia gemma* Swinhoe by using rule induction model in Maharashtra, India," Legume Research – An International Journal, Vol. 40, No. 2, pp. 369-373, 2017, doi:10.18805/lr.v0i0.7297

[42] Sivasangari, A., Krishna Reddy, B.J., Kiran, A., Ajitha, P.(2020), " Diagnosis of liver disease using machine learning models", ISMAC 2020, 2020, pp. 627–630, 9243375

[43] K. Robindro, K. Nilakanta, A. Das, and D. Naorem, "Decision Support System for Rice Plant

Disease Diagnosis using Naïve Bayes' Algorithm," International Journal of Scientific Research in Multidisciplinary Studies, Vol. 3, No. 7, pp. 23-27, 2017.

[44] I. A. P. Banlawe, J. C. Dela Cruz, J. C. P. Gaspar and E. J. I. Gutierrez, "Decision Tree Learning Algorithm and Naïve Bayes Classifier Algorithm Comparative Classification for Mango Pulp Weevil Mating Activity," 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS), 2021, pp. 317-322, doi:10.1109/I2CACIS52118.2021.9495863.

[45] W. E. Sari, Y. E. Kurniawati and P. I. Santosa, "Papaya Disease Detection Using Fuzzy Naïve Bayes Classifier," 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 42-47, doi:10.1109/ISRITI51436.2020.9315497.

[46] J.C. Antony, M. Pratheepa, "A Bayesian classification approach for predicting *Gesonia gemma* Swinhoe population on soybean crop in relation to abiotic factors based on economic threshold level," Journal of Biological Control, vol.32, no.1, pp.68-73, 2018, doi:10.18311/jbc/2018/16309