Comparison of Different Extraction Transformation and Loading Tools for Data Warehousing

Md. Badiuzzaman Biplob
Department of CSE
International Islamic University
Chittagong
mdbadiuzzamanbiplob@gmail.com

Galib Ahasan Sheraji
Department of CSE
International Islamic University
Chittagong
galibahasan@yahoo.com

Shahidul Islam Khan
Department of CSE
International Islamic University
Chittagong
nayeemkh@gmail.com

***Abstract*— Data Warehouses (DW) are database implementations that supports the storage and analysis of historical data. The key components of DWs are known as Extraction, Transformation and Loading (ETL). Since wrong or misleading data may deliver the wrong results. Suitable ETL Tools are necessary for a DW to enhance data quality. The choice of ETL tools is a difficult as well as important issue in data warehousing. This paper first describes the ETL procedure in brief and compare the features of the ETL tools. In this paper, we have compared the existing ETL tools to choose the best option in different situations. From a current industrial market, we collected feedback from the industry professional and documented it to establish the relevance of the data warehouse. We have implemented the available popular ETL tools to compare their strengths and weaknesses to choose the best among them for National Health Data Warehouse of Bangladesh.**

Keywords— ETL, Data warehouse, Database, Data Integration;

I. INTRODUCTION

A data warehouse is a large records repository that consolidates diverse kinds of data converted into a single appropriate format. Relying on particular business desires it can be architecture in another way. However in general data stored in operational databases is transferred to a data warehouse the pre-processing platform also is referred to as the staging area, then after processing into the data warehouse and lastly is transformed into sets of confirmed data marts. **E**xtract, Transform and Load (ETL), is an important element of the data warehousing structure. The method consists of the extraction of data from numerous data sources, the transformation of extracted data consistent with business necessities and loading of that data into the warehouse [1]- [6].

Any programming language may be used to make an ETL technique, but, making it from bits and portions is pretty complicated. Various ETL tools are available in the marketplace easing an organization to select one primarily based on its requirements & needs. With the passage of time, those tools have matured and now provide must more than just extraction, transformation, and loading of data. The improvements consist of capabilities together with “data profiling, data high-quality manage, tracking and cleansing, actual-time and on-demand statistics integration in a provider-orientated structure, and metadata control” [7]-[12]. Furthermore, ETL tools are now customizable according to the functional necessities of an enterprise data warehouse.

Extraction:

In this step, we extract data from different internal and external sources, dependent and/or unstructured. Simple queries are sent to the source structures, using native connections, message queuing, ODBC or OLE-DB middleware. The records could be installed a so-known as a staging area (SA), generally with the same structure as the source. In a few instances we need best the data this is new or has been modified, the queries will simplest go back the adjustments. Some ETL equipment can do this mechanically, providing a changed data capture (CDC) mechanism.

Transformation:

The transformation section guarantees the data consistency and executes data cleansing earlier than loading data into the data warehouse . With a purpose to transform the data properly, some of the guidelines and business calculations are carried out to the extracted data in order that different data formats are mapped into a single format. The transformation may be integrated with the extraction or loading section depending upon when it is performed.

Loading:

Loading the data into a data warehouse or data mart or data repository other reporting applications that house data [13] – [20].

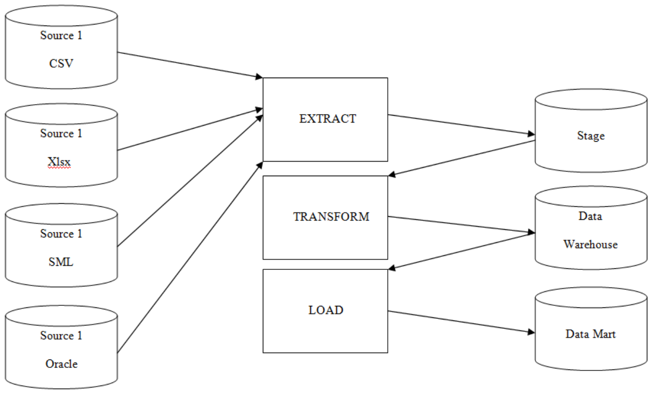


Fig. 1 ETL Workflow

To be used for various analytical purposes. It is carried out often to keep away from data stacks to get piled up. It is able to be required in one of the two conditions:

- Load the new data that is currently contained inside the operational database.
- Load the updates similar to the adjustments happened within the operational database.

The paper gives a knowledge of the ETL tools within the following segments. Section II presents the necessary features of the ETL tools. In Section III, we presented a brief overview of popular ETL tools. In the Section IV, we give a comparative analysis of various existing ETL tools that we have implemented. The paper is ended with a conclusion of the general observations in the segment V.

II. Main Features of ETL Tools

1. Easy to use

   - Easy administration of data maps
   - The capacity to control data either inside or outside the objective database
   - Extensive client manuals and demos

2. Multi-role Team Collaboration

Collaboration software leverages existing technologies to enable groups to communicate, share, coordinate, cooperate, solve problems, negotiate, or even compete for the purpose of completing a task.

3. Large Volume Performance

ETL has its place in the data movement and mixes world. ETL extract, transform (in the stream), at that point load to the object database. Be that as it may, ETL is demonstrating to have issues with super-tremendous volume sets (like 800 million columns in a solitary table) once the information has achieved the objective database, and still needs a change.

4. Clustering and Job Distribution

   - Parallel execution of change over numerous servers or hubs.
   - Load adjusting in view of individual hub use.
   - The group can be powerfully reconfigured by including or clear hubs.
   - Clustering is scaling out. You isolate the worldwide workload and appropriate it crosswise over many hubs, these littler undertakings will be handled in parallel. The worldwide execution approaches the slowest hub of your bunch.

5. Data Partitioning

The related rationale to data partitioning isn't extremely insecure whichever way it requires a polite information stream arranging and a profound comprehension of the required number of queries per each stage. Above all else, We should say that in a few devices this data dividing stream is straightforward to definite clients while in others it must be accomplished physically. As regular one should know exceptionally well the apparatuses' capacities. We originate from IBM and SSIS world, a few years prior IBM Datastage required this way to deal with be accomplished physically utilizing something many refer to as shared holders since variant 8 IBM DataStage handles information distribution on a considerably more easy to use way (yet out of sight, it utilizes this kind of rationale).

6. Automatic Recovery of Flows or Workflow

This is another maximum crucial function is to create workflows that arrange and connect those responsibilities and transformations. This includes the constraint (standards), looping (repeating), branching, and grouping of obligations. Let me provide you with a few examples (the great way to examine is by way of examples).

   - If project A and venture B are a success, do task C, otherwise, do task D.
   - If criteria1 is genuine do challenge A, else if criteria2 is real to do challenge B.
   - Repeat the execution of undertaking a for 10x, or until the quantity of rows is 0.
   - The method the rejected rows from a project the usage of venture B, however the precise rows using venture C.

7. Functionality

Two fundamental angles identifying with the usefulness of an ETL apparatus are essential i.e. the metadata bolster and in general, the usefulness gave by the instrument.The principal usefulness centers around whether the instrument is data purging focused or data change arranged, or it performs both similarly. This one gets an unmistakable picture of what instrument to choose unsure upon the idea of the data that should be put into the instrument. Additionally, the help of guide association with the data hotspot for input is additionally an essential part of the usefulness.

8. Usability

The ease of use is one of the necessary variables of any instrument. In this manner focuses to consider are that the apparatus function to be anything but difficult to utilize, understand and quick to become acclimated to. In such manner angles of concern are that apparatus ought to have a very much adjusted interface also, must help the regular assignments succession starting at any ETL utilization.

9. Reusability

The reusability relies upon that the parts of a data warehouse design, which is developed utilizing the ETL tools and the instrument must be reusable and can deal with parameters. The devices ought to be equipped for partitioning the procedure into little building squares, enable the client to make client characterized capacities and enabling these capacities to be utilized as a part of the procedure stream.

10. Connections

The most critical component of an ETL instrument is the Connections. It must have the capacity to interface with Exceed expectations, FTP, Thomson Reuter, FIX, Salesforce, SAP, Cloud, Hadoop, MQ, LDAP, and web administrations (finish list beneath). In the event that it can't interface with the data source or target framework, at that point it doesn't make a difference what preparing capacity the ETL device has, it can't be utilized. A few instruments can't associate with Exceed expectations or SharePoint or associate with a database situated on various servers, not to mention be interfacing with web administrations, MQ. All devices can associate with a database/RDBMS, however, a few apparatuses have "local customer" drivers, giving a much better execution and control contrasted with ODBC.

11. Performance

- A few tasks just process 1 million lines. In any case, a few tasks process 10 billion lines and it should be done in 60 minutes, not 10 hours. For these organizations, execution is extremely imperative.
- Run the procedure as multi-strings. Screen the string measurements, for example, go ahead without moving time, work time and occupied time. Permit information A to be done in 10 strings, however assignment B just in a solitary string.
- In the event that it the execution is better, given the database a chance to do the joins, separating, arranging, accumulations, and counts. The ETL device can judge in the event that it is fast to do it in the ETL apparatus, or in the RDBMS, in view of the size and kind of the data.
- While doing a query, store the query table (1000 lines) in memory, at that point do the queries 1 million times without contacting the circle. Consequently, choose whether to do full reserve, no store, or incomplete store, conditional upon the data volume and sort.

III. Popular ETL Tools

Some well-known ETL tools accessible in the market are as per the following:.

- *Informatica PowerCenter* is the ETL tools presented by Informatica Corporation, which has the solid client base of more than 4500 organizations. The fundamental parts of PowerCenter are its customer's devices and storehouse instruments and servers. PowerCenter begins the execution procedure as per the Work Flow of the customer service.
- *IBM DataStage* is the main ETL stage that incorporates data over various endeavor frameworks. It uses a prime parallel system, accessible on-premises or in the cloud. The adaptable stage gives broadened metadata administration and the Endeavor network. It coordinates heterogeneous data, including huge information very still (Hadoop-based) or huge information in movement (stream-based), on both circulated and centralized computer stages. It bolsters IBM Db2 Z and Db2 for z/OS applies workload and business governs and incorporates continuous information in a simple to send stage.
- *Microsoft SSIS* Microsoft SQL Server Joining Administrations (MS SSIS) [4] permits run-time data exchange and administration. Intended for big business-wide application bolster, it gives a stage for performing ETL works and making and controlling data bundles. It permits the arrangement of content the application utilizing .net stage bolster expanded adaptability with string pooling, and a further developed import and fare wizard. It likewise permits customization of the bundle suiting particular association needs, utilization of the computerized signature for security also, bolsters benefit situated design.
- *Pervasive Data Integrator* is the earth which creates usefulness for data total and data development computerizing. It could stack superior data distribution center and also littler data parts. Bound together toolset lets deal with the incorporation of numerous applications and operational data stores. Unavoidable Incorporation Motor backings planning multi-step forms, that can be executed consequently. In the event that any mistakes will show up between the procedure, there are logging or rollback tasks accessible.
- *Talend Integration Suite* is a standout amongst the most ground-breaking information Incorporation ETL tool in the market. Begun in 2006, has a less network of devotees yet at the same time has a significant piece of the overall industry as 2 supporters are fund organizations. As opposed to metadata driven it utilizes a code-driven approach and has a GUI for the client connection. The code age property permits creating an executable code of Java and Perl that can be run later on a server.

Using Talend Open Studio (Fig. 2) convert one format (XML) from multiple CSV file. For this purpose, first input multiple CSV files in Talend Open Studio and after the full process the output is only one format (XML) (see Fig. 3)

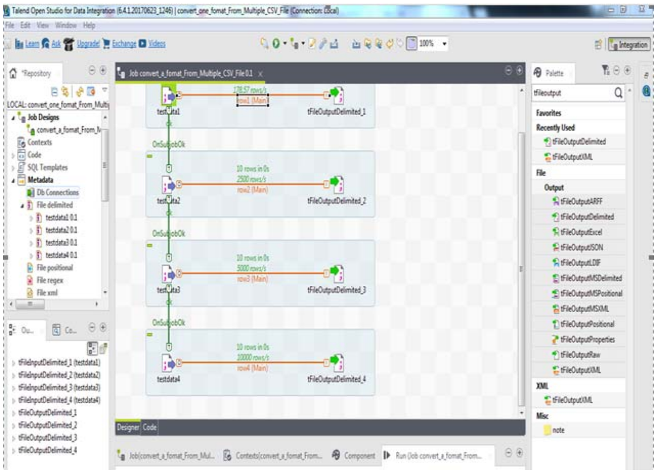


Fig.2. CSV to XML format Talend Open Studio

Using Talend open Studio (Fig. 3) convert Xlsx and CSV to CSV. For this purpose, first input Xlsx and CSV file in Talend Open Studio and after the full process the output is only one format (CSV).

Xlsx and CSV to CSV:

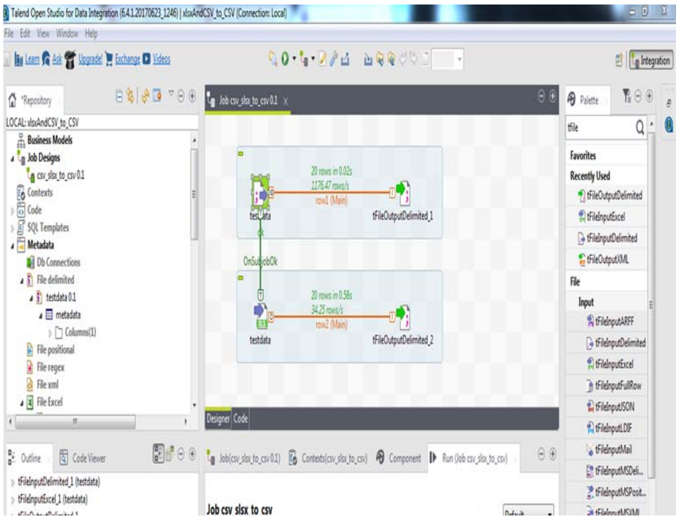
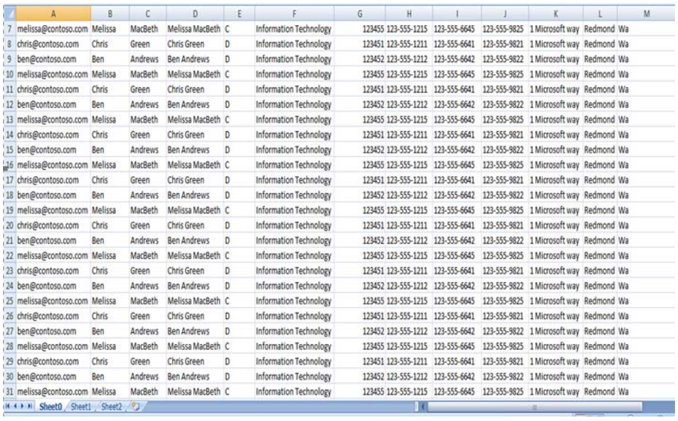


Fig.3. Xlsx and CSV to CSV

- *Pentaho Kettle Enterprise* Pentaho data Combination is an ETL tool keep running by Pot runtime. Here, the techniques are spared in XML documents and deciphered in Java records while changing the data. Utilizing the imaginative meta-driven approach, it is quick having a simple to utilize GUI. Having begun in 2001 it has developed and today it has a solid network of 13,500 enrolled clients. It likewise underpins multi-arrange data and permits data development between a wide range of databases and records.
- *CloverETL* is the data Combination portfolio presented by Javlin Inc in 2002, in light of Java Stage, for the most part, intended to change and scrub the data from one source Database to required target Database.

Using CloverETL easily filter and write the data from the dataset. In the dataset filter and write the data in Job Title attribute. In Job Title attribute filter all 'A' value from the dataset (Fig. 4).



	A	B	C	D	E	F	G	H	I	J	K	L	M
7	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
8	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
9	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
10	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
11	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
12	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
13	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
14	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
15	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
16	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
17	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
18	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
19	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
20	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
21	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
22	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
23	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
24	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
25	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
26	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
27	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
28	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa
29	chris@contoso.com	Chris	Green	Chris Green	D	Information Technology	123451	123-555-1211	123-555-6641	123-555-9821	1 Microsoft way	Redmond	Wa
30	ben@contoso.com	Ben	Andrews	Ben Andrews	D	Information Technology	123452	123-555-1212	123-555-6642	123-555-9822	1 Microsoft way	Redmond	Wa
31	melissa@contoso.com	Melissa	MacBeth	Melissa MacBeth	C	Information Technology	123455	123-555-1215	123-555-6645	123-555-9825	1 Microsoft way	Redmond	Wa

Fig.4 CloverETL Filtering and Writing Data Output

IV. COMPARATIVE ANALYSIS OF VARIOUS ETL TOOLS

With every one of the angles, as talked about in segment 4, as a primary concern an examination of the administrations gave by the apparatuses is talked about from this point forward. In this manner, in picking any instruments its separate perspectives ought to be considered. Following chart based examination offers help for the basic leadership. For this investigation, different sites, merchant's white papers, web journals, correlations, and past overviews were counseled and consequently in view of the fundamental arrangement of highlights examined in area 4 the investigation was directed. Every one of the previously mentioned ETL instruments, as examined in segment 3, is evaluated based on indicates concurring the level of administrations upheld while the merchants are delineated by the acronyms in the chart.

There are different kinds of ETL Tools. Here we compared 7 ETL Tools. They are Informatica Power Center, IBM Data Stage, Microsoft SSIS, Pervasive Data indicator, Talend Integration Suite, Pentaho Kettle Enterprise and Clover ETL. There are different kinds of features available in this comparison Table I which compares all of the ETL tools to find out which feature is available and other doesn't have that feature. In this Table, 'Yes' indicates It is available in the ETL tool and 'No' indicates It is not available in the ETL tool. Talend Integration Suite, Pentaho Kettle Enterprise, and Clover ETL are easy to use but Informatica Power Center, IBM Data Stage, Microsoft SSIS, Pervasive Data indicator are not easy for a general user to use. web based UI is available in Informatica Power Center, IBM Data Stage, and Talend Integration Suite and other doesn't have this feature. Automatic Recovery of flows are available in Informatica Power Center, IBM Data Stage and other doesn't have this feature. In this similar way, We can compare all the ETL tools and find out which feature is available or not.

TABLE I. Comparative analysis of various ETL tools

Feature	Informatica Power Center	IBM Data Stage	Microsoft SSIS	Pervasive Data indicator	Talend integration Suite	Pentaho Kettle Enterprise	Clover ETL
Easy to use	No	No	No	No	Yes	Yes	Yes
Web-based UI	Yes	Yes	No	No	Yes	No	No
Multi role Team collaboration	No	No	No	No	No	No	No
Process Centric Approach	No	No	No	No	No	No	No
Enables SOA	No	No	No	No	No	No	No
Reusable service Repository	No	No	No	No	No	No	No
Single install light footprint	No	No	No	Yes	Yes	No	Yes
Large volume Performance	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clustering And job distribution	Yes	Yes	No	No	Yes	Yes	Yes
Data partitioning	Yes	Yes	No	Yes	Yes	No	Yes
Automatic Recovery of flows	Yes	Yes	No	No	No	No	No
XA-Transaction Rollbacks	Yes	Yes	No	No	No	No	No
Meta driven approach vs. code	Yes	No	No	No	No	Yes	Yes
Build-in scheduler	Yes	Yes	No	No	Yes	No	No
Real Time Triggers	Yes	Yes	No	No	No	No	No
Non-RDMBS collections	Yes	Yes	No	Yes	Yes	Yes	Yes
Web services clients	Yes	Yes	No	Yes	Yes	Yes	No
Publishes flow as Web Services	Yes	Yes	No	No	No	No	No
join multiple sources	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sprite Datastreams	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Graphical Data Mapper	Yes	Yes	Yes	Yes	Yes	No	Yes
Complex transformation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data validations	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Previous Sources Data in design	Yes	Yes	No	Yes	Yes	Yes	Yes
Running Mapping Rules in design	Yes	Yes	No	No	Yes	Yes	Yes
Library of new Mapping functions	Yes	Yes	No	No	No	No	No
Complex Lockup	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Automatic Documentation	Yes	Yes	No	No	No	No	No
Human workflow of Error Handling	No	No	No	No	No	No	No
Plug-in external programs	Yes	Yes	No	Yes	No	No	No
Alert and notifications	Yes	Yes	No	Yes	Yes	Yes	Yes
Version control	Yes	Yes	No	No	Yes	No	No
Deploy in cloud option	No	No	No	Yes	No	Yes	No

V. CONCLUSION

Important data in the majority of the organizations are underused in light of the fact that it exists around in various formats and in different resources. Data warehouses (DWs) are complex processes having combined data with the principal goal to help the information workers in the decision-making system. The key parts of DWs are the Extraction, Transformation and Loading (ETL). It is important to find out the proper ETL tool according to business necessities and investments. The objective of this paper is to explain ETL process, its significance important to the data warehouse and give a comparison with implementation in view of some summed up criteria to discover appropriateness of a tool for a certain class of consumers. We have compared the existing ETL tools to choose the best option in different situations. From a current industrial market, we collected feedback from the industry professional and documented it to establish the relevance of the data warehouse. We have implemented the available popular ETL tools to compare their strengths and weaknesses in an aim to choose the best among them for National Health Data Warehouse of Bangladesh.

References

[1] A. Kabiri, F. Wadjinny, and D. Chiadmi, "Towards a Framework for Conceptual Modelling of ETL Processes", Proceedings of The first international conference on Innovative Computing Technology (INCT 2011), Communications in Computer and Information Science Volume 241, pp 146-160.

[2] Ahmed Kabiri, Dalila Chiadmi, ―Survey on ETL Processes‖, Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645 Volume 54, No. 2, pp – 219 – 229, 20th August 2013.

[3] T. Jaorg, S. Dessloch, Near-Real-Time Data Warehousing Using State-of-the-Art ETL Tools, University of Kaiserslautern, 67653 Kaiserslautern, Germany, 2009.

[4] Passionned Group Stichtse Rotonde, 'The BI Tool survey report", 2008 (2018) https://www.passionned.com/kb/business-intelligence-tools-survey/background.[accessed 21 February 2018].

[5] Adeptia Adeptia incorporation, ETL Vendors Comparison(2018) http://www.adeptia.com/products/etl_vendor_comparison.html.[accessed 25 January 2018].

[6] ETL data ware house concepts (2018) http://etl-information.blogspot.com/2007_07_01_archive.htm.[accessed 19 January 2018].

[7] Guide to Data Warehousing and Business Intelligence (2018) https://data-warehouses.net/architecture/etlprocess.html.[accessed 1 January 2018].

[8] 2018 ETL Tools Comparison (2018) https://www.alooma.com/blog/etl-tools-comparison.[accessed 12 December 2017].

[9] ETL tools Survey (2018) http://www.etltool.com/what-is-etl.htm. [accessed 12 December 2017].

[10] Oracle Ware house builder 11g, A technical overview (2018) http://www.oracle.com/technology/products/warehouse/index.html. [accessed 9 January 2018].

Dr. R. Chillar; B. Kochar; Extraction Transformation Loading –A Road to Data warehouse, 2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries.

[11] Pervasive Systems, Extraordinarily Flexible ETL Platform (2018) http://www.pervasiveintegration.com/scenarios/Pages/etl_tools_data_aggregation.aspx [accessed 15 November 2017].

[12] Guide to Data Warehousing and Business Intelligence (2018) http://data-warehouses.net/architecture/etlprocess.html. [accessed 15 November 2017] .

[13] A. Binstock and J. Rex, Practical Algorithms for Programmers. Addison-Wesley, Reading, Mass., pp. 158-160, 1995.

[14] The best fit for your purpose (2018) http://hosteddocs.ittoolbox.com/ab100104.pdf. [accessed 15 November 2017].

[15] Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading Toan C. Ong1*, Michael G. Kahn1,4, Bethany M. Kwan2, Traci Yamashita3, Elias Brandt5, PatrickHosokawa2, Chris Uhrich6, and Lisa M. Schilling3. Ong et al. BMC Medical Informatics and DecisionMaking (2017) 17:134DOI 10.1186/s12911-017-0532-3.

[16] A Hybrid Feature Selection Method for Classification Purposes. Silvia Cateni, Valentina Colla, MarcoVannucciScuolaSuperiore S. AnnaTeCIPPERCROPisa, Italys.cateni{colla,mvannucci}@sssup.it.

[17] ETL process modeling in DWH using Enhanced Quality Techniques, kushanoorAkbar; Dr.S.Murali Krishna And T. Vidya Sagar Reddy

[18] International Journal of Database Theory and Application Vol. 6, No. 4, August 2013 179 ETL Process Modeling In DWH Using Enhanced Quality Techniques.

[19] Data Warehouse, *http://datawarehouse4u.*info accessed on September 10, 2014.

[20] ETL Tools information, *https://etl-tools.info/en/bi/etl_process.htm* accessed on September 12, 2014.

[21] T.Y. Wah, H. Peng, and C.S. Hok, "Building Data Warehouse," Proc. 24th South East Asia Regional Computer Conference, November 18- 19, 2007, Bangkok, Thailand.

[22] Dr. R. Chillar; B. Kochar; Extraction Transformation Loading –A Road to Data warehouse, 2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries.

[23] Muhammed Arif, Ghulam Mujtaba, ―A Survey: Data Warehouse Architecture‖, International Journal of Hybrid Information Technology, ISSN: 1738-9968 IJIT Vol. 8, No.5, pp. 349-356, 2015.

[24] S. I. Khan and A.S.M.L. Hoque, "Towards Development of National Health Data Warehouse for Knowledge Discovery", Intelligent Systems Technologies and Applications, Springer, Vol. 385 No.2, pp.413-421, 2016

[25] S. I. Khan and A.S.M.L. Hoque, ""Development of national health data warehouse Bangladesh: Privacy issues and a practical solution," 18th International Conference on Computer and Information Technology (ICCIT), IEEE, 2015.

[26] Khan, S.I. and Hoque, A.S.M.L., 2010. A new technique for database fragmentation in distributed systems. International Journal of Computer Applications, 5(9), pp.20-24.

[27] S. I. Khan and A. S. M. L. Hoque, "Health data integration with Secured Record Linkage: A practical solution for Bangladesh and other developing countries," 2017 International Conference on Networking, Systems and Security (NSysS), Dhaka, 2017, pp. 156-161.

[28] S. I. Khan and A.S.M.L. Hoque, "Privacy and security problems of national health data warehouse: a convenient solution for developing countries," In Proc. of the International Conference on Networking Systems and Security (NSysS). IEEE, 2016.

[29] Khan, S.I., 2016. Efficient Partitioning of Large Databases without Query Statistics. Database, 1, p.2.

[30] S.I. Khan and A.S.M.L. Hoque "Development of National Health Data Warehouse for Data Mining," Database Systems Journal, Vol. VI, No. 1, 2015

[31] Khan, S.I., Hoque, A.S.M.L. and Ullah, M., 2016, January. National Health Data Warehouse Bangladesh for Remote Health Monitoring: Features, Problems and Privacy Issues. In Remote Health Monitoring Workshop.

[32] Khan, S.I. and Hoque, A.S.M.L., 2015, May. Towards development of health data warehouse: Bangladesh perspective. In Electrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference on (pp. 1-6). IEEE.

[33] Chaturvedi, Alok R., Ashok K. Choubey, and Jinsheng Roan. "Scheduling the allocation of data fragments in a distributed database environment: a machine learning approach." IEEE Transactions on Engineering Management 41, no. 2 (1994): 194-207.