Novel Solution for Real Time Challenges of ETL in Big Data

Vijayalakshmi M[a], Dr.R.I.Minu[b]

[a]Research Scholar,Department of Computer Science and Engineering,SRM Institute of Science and Technology,Chennai,India

[b]Associate Professor, Department of Computer Science and Engineering,SRM Institute of Science and Technology,Chennai,India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: The big data which handles large amount of data in minute to minute situation is being supported by a new technology known as Extraction, Transformation and Loading. This technique has its own challenges while extracting the data, while processing it and transforming the data. After transforming data from one form to another, it should be loaded to the particular server into a particular form and into particular size. While performing these functionalities ETL faces challenging task, this task should be combated successfully for better QOS. In this proposed work various challenges to ETL is elaborated and solutions are derived to improve the performance of ETL. An artificial intelligence algorithm based support is recommended in this work for enhancing the better ETL performance over the challenges it faces.[1],[2]

Keywords: Extraction, Transformation, Loading, Artificial intelligence, Bayesian networks, Multinomial Bayes network;

1. Introduction

Stream of Data are flowing between the time intervals nowadays .Handling of data is a really a challenging task Extraction, Transformation and Loading (ETL) is an emerging ideology that provides a support to handle the data. ETL comprises of three different types of aspects which possess their own task over the data. These aspects has set of challenges and solution for the task in data handling that is been proposed in this work.[3] [4]

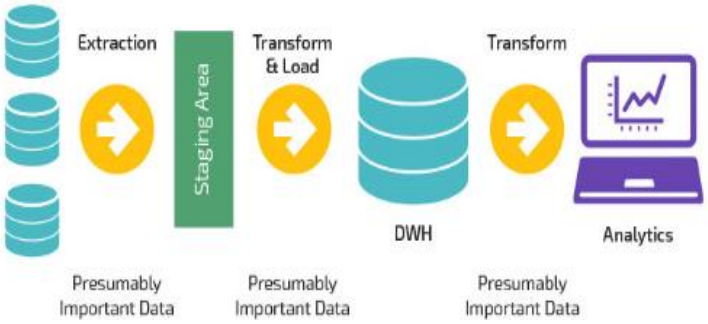

Fig 1: ETL process

1.1 Extraction

While handling mahout data in the short period of time information extraction process really faces new kind of challenges data can visualized as homogenous data, heterogeneous data , structured data, semi structured data and unstructured data. Figure 1 portrays the ETL process.

(i) Homogenous data**:** Similar kind of data are visualized as homogenous data, here the data modeled with similar kind of aspects Similarity of the data is based on percentage index , it varies point to point . Initially the a set of data can be grouped as homogenous one ,when analysis moves form one level to another further the set varies to different set of data. So identifying completely homogenous a challenging task.[5] [6]

Here the deviation of the data or the deviation point in the data should be identified through numbering or naming every point for identification. Let the number series initiates form n1,n2,n3… let n1,n2,n3 are similar homogenous type of data. From n4 the data is started to deviate it can nominated as n4m1 as the deviating point this will

differentiate a slight from the overall homogeneity of data.[7] [8]

(ii)Heterogeneous Data**:** The data with many variability such as data types as well as formats is termed as heterogeneous data. This data possess some errors due to missing values, data repetition as well as data integrity lacks in many cases. As it want to handle large amount of data, it requires components to handle the task.While collection of data happens there are the chances of noises in data that are too considered as a challenge.
Based on the nature of the data heterogeneity is classified in four types as syntactic heterogeneity which occurs while the representation of data are not from similar linguistics, conceptual heterogeneity occurs while projection of data model in different model, while in terminological heterogeneity different data sources are representing similar data in different entities and finally in semiotic heterogeneity users interpret similar data into different entities.[9][11][12]

(iii)Structured data**:** The data which possess the characters of direct entry, easy storage ,answerable for querying and also for analysis with well-organization of data is referred to as structured data in this data will in the prescribed format the structured data are represented in three formats like JSON-LD is been used as page headers ,RDFa is used to highlight the important items based on type of and properties of HTML features , Micro data is also a highlighting format based on item type and item prop HTML features.

(iv)Semi Structured Data**:** The data that can't be ordered in a particular structure is meant as semi structured data, it can't be related with particular database. It has its own differentiated structure these data will be stored in a array with different entities. it doesn't have proper meta data to make a decision making and managing is difficult to handle in direct.
(v)Unstructured Data: Data that possess many forms grouped in a unsegregated order is meant as unstructured data, it will not suit for any analysis database .These unsorted data are in many format are in mingled form.
Extraction of Data: Wresting the data from a main source server for multiple usage this is an initial process of ETL concept. As the extracting function is the time taking function it is handling the all kind of data like structured to unstructured data it is mentioned earlier. The extraction is for all formats of data too. There are many methodologies in extraction there are Logical based extraction and Physical based extraction.[13] [22] [18,64]

a. Logical based Extraction.

In Logical based extraction there are two categories complete extraction and augmented extraction. In complete extraction from the source server the completely the full data that is till on date, is extracted. The final available data to develop a particular ideology or to get the complete export file that is necessary for another source table or for the server. In augmented extraction method a specific deviating or changing data are considered for the extraction for analysis purpose. Point of changing is taken and analysis is done at point for identifying need of change in the particular point. In augmented extraction mode three mode of extraction parameters are considered they are Timestamp based, data partitioning and data trigger.

In Timestamp based data extraction taking place through time period is fixed based on date and time period .The time interval is fixed between the time slots .The data generated in between these two time slots are taken into consideration for analysis.[30][31,61,62,63]

In data partitioning the process takes place based on time slot, the time period is fixed between two date slots, the data accumulated between these two date slots is taken into consideration and analyzed.

In data trigger whenever there is a change in flow of data is considered, trigger will happen while data is updated. Through this time of change in data is easily identified.

Based on these logical based extraction is made from the main data source, for the data usage purpose.

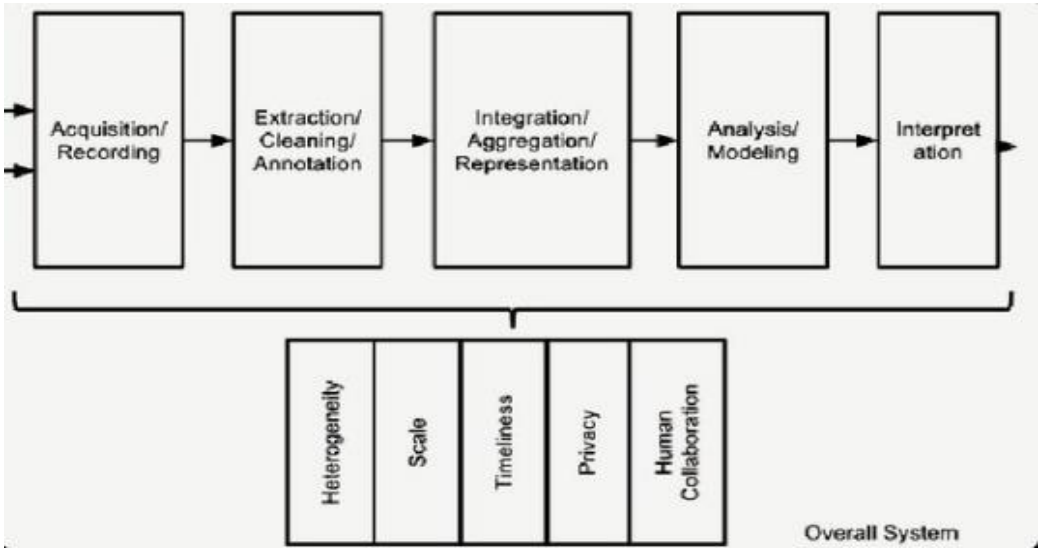

Fig 2: Data Extraction strategy in Big data

b. Physical based Extraction

Here extraction taking place physical manner based on the connectivity of the network that is online mode of extraction or else through offline mode, while the data extracted directly from the source server itself is known as online extraction .In offline extraction is done from the server which act as an intermediate one between original source system and user system.

1.1.1 Challenges of Data Extraction in Big Data[15] [17]

There are few challenges to be sorted out for providing better QOS through ETL technology. The extraction based challenges are discussed at this point.

a. Storage issues in Big Data while extraction.
Data extraction from large database is still a challenging task it needs better solution, the storage challenges are based on storing the data in a proper structure, searching the appropriate data, filtering it and transferring it is still a time consuming process. Big data storage server should be a fault tolerant, secure, flexible and scalable in nature.

b. Website issues in big data
Website accessing Big data has issues based on upgrades, these websites must be upgraded periodically in regular time interval. As it is handling huge number of data, Adjustments should be made periodically in a way of handling large data in a period of time. By that server crashing and other server based issues can be avoided.

c. Web Scarping Issues
Certain websites are maintain technologies to avoid web scarping ,this issue is a challenge for the data or information extractors .They can provide a mode of retrieving data legally ,for the users who really need of it.

d. Fake data traps

Some websites or data providers are trapped with fake data, which would old or unworthy. Fake or wrong data always possess major issues for data extractors, so proper algorithm is in need to be designed to handle this such issues. Making sure the data is updated and quality of service is maintained.

e. Extracting Noisy data

In many instances the data will be with noise and errors in that case extraction purpose will not be solved properly. Mostly unstructured data will possess this noisy data.In that case a proper solution should be sorted out to make a

clear extraction of data that is made of use to the user.

f..Large Scale Extraction.

While going for extraction of larger data, there are chances in lack of data quality and precision. Maintaining large data in a single spreadsheet of records and analysis on daily basis need more perfection this also a considered as a challenging task in extraction.

1.2 Transformation.

The structure of data is changed into various formats for the analysis purposes. The constructive nature of data transformation consists of adding, copying and making multiple copy of data as of destructive part consists of deleting the fields and records. Aesthetic or structural database are other aspects of transformation. This process consists of Data Assimilation, Data Exodus, Data Silos and Data Squabbling.

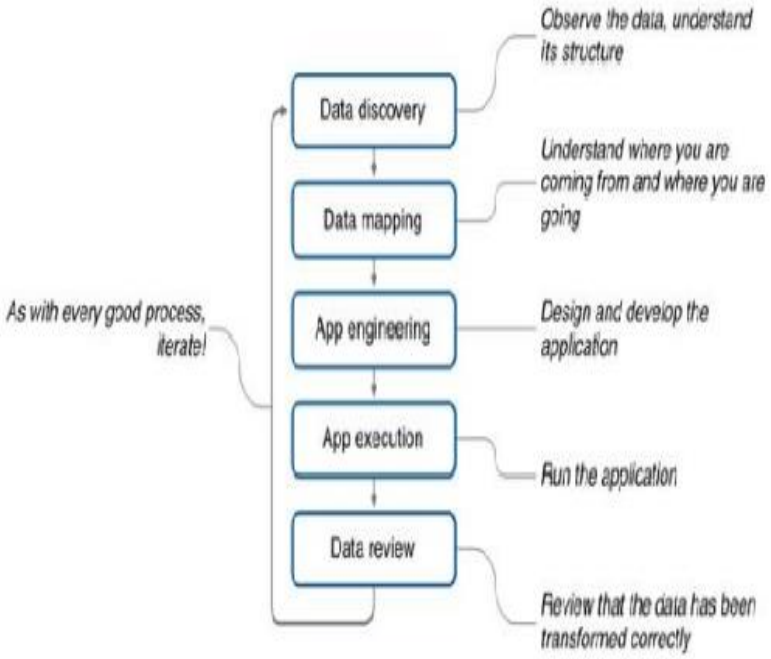

Fig :3 Data Transformation process

(i)Data Assimilation: Consolidating the data from various sources before transforming is one of the part of data transformation. Various data will be different forms these data collected and to form as integrated model for transformation in users accepting format. It unifies the data in a single format.In data integration they are certain challenges should be sorted out , the spelling issues with different spellings with same pronunciations , in banking and in other loan sections customers will possess different number in a same name ,so different records will be created for a single customers, the integration makes the data in a compatible for transformation for various application. Data assimilation supports the user in better handling data in an error free manner. With efficient time management ideology. It increases the quality and value of the data.

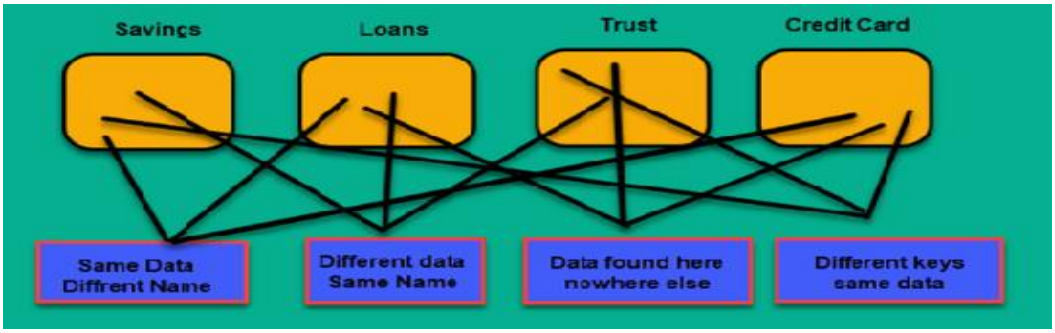

Fig :3 Challenges in Data Assimilation.

(ii)Data Exodus**:** Here the process of moving data from one system to another system is known as data exodus or migration of data. Here the extracted data which is going to transform must make a migration a system to another. While deploying is made data migration is essential along with existing one. Data exodus has four factors of processing similarly knowing about the data, tidying,
Overall data protection, managing the data is the process of a complete knowledge of the data is required. If this process is ignored unexpected issues may occur. Whereas tidying is the process of clean up or clearing unwanted noise in the data by that data truthfulness is ensured. In maintenance of data ensures the data quality. In managing the data, all aspects of data is ensured for data integrity. Overall maintenance made to improve data quality.

Data exodus has major hurdles in achieving its task. The poor knowledge about the source data generates issue in data migration duplicate data ,missing data , error spellings and error data are the factors of low knowledge of data this creates an issue in data migration. Data analysis are not being done properly there is a lack in expert people in data analysis. This leads to error data analysis and people lack correcting error data. Data processing itself a integrated work in data migration the all the data must be mapped in a order of integration to avoid unwanted errors while migration there are chances of loss of data . That too to be monitored to have better quality of service. Some specifications of the data difficult to validate some critical specifications can't be validated as of simple. This factor also impacts in data migration. Testing the data quality is an hurdle in data migration .The test will identify the nature of the data, whether it is acceptable or not in that case, the proper test will provide the proper data. In the data migration process. The test results to be provided in the right time duration in that situation the test must be made in correct duration and results also to be provided in the right time. The delayed result will delay in data propagation in turn data will be useless in delaying in the process. The data workers work with different tools for analysis, while migrating data each user try to project and provide their data in their format, this factor can be considered as a hurdle in data migration . System will not accept different kind of projection for same data. a collaborative model should be proposed for better quality exodus.

(iii)Data Silos**;** Data silo is collecting different type of data and storing it in a storage area called data silo. Data silos are storing large data in a storage area in other manner called as warehouse similar structure nowadays ,many cloud are designed to store huge amount of data as a storage as service. Nowadays data silos are maintained in the cloud environment, which provides enough spacing for storing enormous amount of data.
(iv)Data Squabbling**:** Data squabbling is cleaning the raw data into usable data, it is the part of process of data translation, It is also arranging into desired format, It is a process of structuring the data. It is the process of converting complex data in simplified format that is in user understandable format. Data squabbling comprises of six major process of discovering of data that is usable in nature for that data identification is required in that from the raw data usable data is identified with comparing with old references and confirming that user really require.

After discovering the data it should be structured in a form of user convening way and user understandable way through specific points, charts and maps. Then it is cleaned or cross checked whether the data are in proper flow, mismatches or error in the data. After cleaning it is enriched with comparisons with other relevant information to focus the data. After these process the data should be validated for consistency, quality and security of data that is taken for use. Finally the data projection in various ways like publishing and final documentation. By these ways the data squabbled for the transformation.

There are many numbers of data quality problems in the process of data squabbling ,they are unmaintained

data attributes, attributes abusing for additive information, error data ,typo mistakes, lack of accuracy in data, lack of data ,repeated data and inconsistent data , format issues, fake records and lack of updating of data.

In unmaintained data attributes the proper maintenance of data is not made while adding new data while data squabbling they wrongly managed and misplaced, while new information is required , lacking response is made available for the user. Error data and typo error is a part of whole data set in that these errors must be rectified. Lack of data accuracy and repeated data makes the database more critical to data access. Lack of consistency of data is a major issue that is noticed in the data wrangling, it can be solved through data matching algorithms and pattern matching ideologies. The format from various data source will not be in similar dimension various formats can be grouped and made into single standard format user acceptable format in the process of data squabbling. Multiple fake records which is one sort data attack, in which fake records are created. A huge form of data is shown in that original data are made to disappear .This can be solved with proper maintenance of data. In some area the data are not made updated. These issues also considered as a major issues that can be solved data squabbling.

Data wrangling should be capable of handling uncertainties in the data, it should be capable of handling such amount of data in a limited time period, this happens in the mismanagement of data. This can be resolved by proper data time slot maintenance for each data process. Error tolerance is a major support required from data wrangling system. Huge amount of data to made handle by this data squabbling system. Error tolerance algorithms must be framed to handle such data with error correction infrastructure by that quality of data service will be improved. Based on these processes the transformation of data is enabled. Transformation of data is made well organized for better humans and computer usage, the data quality is improved through formatting the data and validating it. This will develop the data quality from the well-known issues such as missing data, unpredicted replicas, unfitting directories and incompatible formats.

a. Missing Data: While data is transformed from one form to another form, there are chances of missing of data. Missing data are the values that are not available, it is comprises of sequence missing, features unavailability, missing files, incomplete data, error in data entry. Understanding the issue is an important task in data handling. Multiple approaches can be implemented to identify the missing data and rectifying the issue. There are many algorithms are proposed and derived to solve this issue.

An artificial intelligence approach is recommended for solving this kind of issue with support of hypothesis based identification of missing data with some probability support. Value of data based on probability formula outcome is replaced in the place of missing data. For this approach many artificial intelligence based algorithms like Naïve Bayes, Bayesian, and Genetic algorithm are considered for the support.

b. Unpredicted Replicas**:** In many conditions multiple copies of data is being generated by various reasons, Mismatched data, unexpected data generation happens while handling or collecting the data. This is a major issue in data analysis. The identification of this issue is one major challenge.

Multiple departments of a concern is handling the single customer's data, there are chances of double entry or multiple entry of the customer information in a single file, and this may provide a structure of bulk data.

Data that are mismatched to records makes some contradiction and complications while compiling the data, this may occur while there is no proper update in the records of data. Data formats are one of the factor that generate this kind of challenge data transformation. For an example date format is always mismatched with data and month till 12 numerical. A code is given as a format for some places .But most of the data entries don't have the knowledge of those code, there is chances wrong entry while record developing and transforming the data.

In another aspect the related data are stored in irrelevant areas in record, while obtaining those related data is considered as a challenge in obtaining related data. Collecting related data also pose a real challenge. Most of the time it is shown as a related data but while it is verified it is considered an irrelevant data. Data update is considered as both an issue as well as a solution in data transformation, in many sections lack of data update is still prevails. This poses an issue in data handling. Whatever data is feed to the system is considered as the updated to the system, in that case this update goes in many levels of record. This leads to chances of illegal activities in many areas. This can be averted by storing data time with each data and time update may solve this issue for some extent. While record analysis, mentioning of that particular data update can avoid misassumption of data update by the user.

c. Unfitting Directories**:** In most of the cases unfitted data is made to the indexing in the database in form of directories, this is one of the issues noticed here. Irrelevant data are made to enrolled and that too are indexed in the directories, this makes a mishap for the users , This issue is a common issue that occurs , This happens through human error .This issue can be solved through proper verification at each level of the database.

d. Incompatible formats**:** While storing data, the format of data will be in different manner, while processing the data needs compatibility. Most of data will be in incompatible format. These different formats makes data access a difficult .This factor is considered to be an issue, while transferring it. For example various types of format it may be in image form or it may be in document format. While transferring it is restricted to image form, then it should be transformed into image form. Sizes of the data, in certain areas the data are restricted to certain size. If it is exceeds the size the data is not accepted. These are areas the incompatible is noticed.

For this a common format is should be prescribed, or an algorithm must be proposed to handle this such issue, which would transform any form of data into acceptable formats. While ETL data handling process is enhanced to big data version, it faces the foretold techniques and issues. In this proposed work solution for the tasks are recommendations for providing quality of service.

1.3. Loading

The process of uploading the processed data for the access of users is meant as loading ideology. The loading is based on the architecture it has two step process, initial step is the process of uploading data to the system database and latter it is transferred to warehouse enable to access to the users. Initially there is two issues in while data analysis is to be avoided before complete loading because a without a complete fact and dimension the analysis will always lead to wrong decision and query enactment issue is based on queries that has been raised over dimension and fact table, during database update the query can't be answered, so enactment will be identified as an error in data loading.

The loading of data will improve access speed, level of efficiency and more flexible, More in mandatory. Data Loading comprises of insert, update, upset and bulk load operation, in insert operation new data are inserted, in update operation new data are updated over the old data as a rewriting process, the upset is a novel special operation in which the data are in manner in which where the update is required or selective update , here overall row contents that need to be update is focused. In bulk load utility if there is need of overall update a complete data file is replaced with the update data file, in the bulk of data manner. Here total update of data is replaced over old data. Two datasets is taken as input by bulk up loader in this one dataset is contains control information other is possessing data to be loaded. The control data sets identifies tables and its columns where data to be loaded. The load operation takes place through three operations called Append, Insert, Replace. In append operation rows are appended to the table, while insert the input data is filled with table. In replace operation previous data is replaced with new data.

Generally there are two types of loading system, one is history load and incremental load , in history load old and historical load is stored and loaded for the analysis process , the data are kept for long period for research and analysis process. While in incremental load are the data which are newly developed data .current or future developing data .A buffer space should be maintained in the database for processing and analyzing the data. These incremental data are based on the time slots, it starts from daily basis data to yearly based data.

2. Challenges in Data Loading:

There are tasks for loading of data to be meet for better performance of data management. Analysis of slowing down, loading likelihood data issues, knowledge about data loading.

Analysis of slowing down: while data is being uploaded, when it is large in size the uploading time is increased. This is considered as challenge. It can be rectified by increasing band width and increase in storage server size. The challenge can be resolved.

Loading likelihood data issues: Repetitive data that occur in a database in various areas, pose a chance of confusion developed as a challenge, for an example in certain database the date that is registered or entered will be repeated in the same record for some other event or for a some person, this pose a confusion sorting out the solution for this issue is the challenge for the issue. A proper segregation from initial state of data recording will tend to solve this issue. [46]

2.1. Challenges to ETL System

ETL is acting as a supporting backbone of all data based industry, it observers all data from the various users from multiple sources as a single point of source into a data warehouse. As ETL components extraction, transformation and loading characters and issues are discussed in the earlier section of this work. From here onwards the combined technology of above three techniques known as ETL's challenges are going to be discussed. Initially the challenges

of ETL systems are. Losses of data in this process, errors, incomplete as well as duplicates in data transacted, old data storage issues as well as retrievals, and testing issues in ETL system. These challenges are primary challenges is been identified other than we have secondary challenges from the end user portion like Huge loads of data, inefficient query process, poor mapping system, designing error in source and target systems.

Huge loads of data**:** Generation of data is becoming huge from day to day like heaps to bounds nowadays ETL processers are more struggling to meet the issue, error are made while loading data. They are instead of complete replacing of the data, selective replace of the data, will make chaos and lead to major issues. It should be sorted out. It can be rectified by complete replacing set up. Uploading the error data and repeated data will always be a challenge for the ETL system, proper monitoring of data with the support of artificial intelligence data check approach will ensure the stability of the data.[47]

Artificial intelligence is recommended to clear this issue, AI based approach will monitor the system. With the support of algorithm it will compare the data that are tend to be repeated one. With the support of AI algorithm flow of the data is made analyzed, if there is any difference occurs in the flow of data, AI will analyze and confirm the data is correct or error in the data, then it is rectified.

Major issues are created due to memory spaces , lack of memory space will lead to memory resource issue , if major part of memory is occupied with the duplicate data and repeated data then memory space is will be filled, to avoid this proper surveillance can be done to avoid storing of junk data. Periodical replacement of old data with new data, and storing old data in a separate server will enhance memory space capacity for handling the new data.
Data are handled in a serial manner this will lead of data collection in huge manner,In sometime it may lead to server crash instead of making data collection for the particular time level, data can be handled in parallel, to avoid bulk data process. [41]

i.Inefficient Query Process**:** basically query process is made by sql queries that are inefficiently structured, it is more structured in a manner to be running for more hours and minutes for a consideration is running in a manner in that for getting solution within two tables, it is running for whole database for this issue we can consider a solution, in that mapping and indexing can be implemented for better answering for the query answering.[35]

ii.Poor mapping system**:** The mapping system that are designed in a form is poor in nature ETL is structured in a manner irregular in a nature , it possess errors in a way that it possess error mapping ,there are data insertion and missing data in poor mapping . For this issue mapping can be done AI based algorithms, which would patch the missing data and identify the error data as it mentioned earlier.[37]

iii.Designing error in source and target systems**:** The most of ETL systems must be aligned in a possible manner, there are common design errors there are repetition and redundant errors and normalization errors in the database. This reduces the performance of the ETL system. By Artificial intelligence based algorithm like Bayesian network algorithm and naïve Bayes algorithm put in together will identify the repeated data through their probability points. These points will sort out repeated data and error data and based on that errors can be rectified.[36,63]

iv.Issues over ETL Architecture:

In any case, conventional ETL devices can't stay aware of the rapid of changes that is commanding the enormous information industry. We should investigate the deficiencies of these customary ETL apparatuses. Conventional ETL apparatuses are profoundly tedious. Handling information with ETL intends to build up a procedure in different advances each time information needs to get moved and changed. Besides, customary ETL apparatuses are rigid for changes and can't stack meaningful live-information into the BI front end. We likewise need to make reference to the way that it isn't just an exorbitant procedure yet in addition tedious. Also, we as a whole realize that time is cash. There are a few factors that impact the capacity of ETL devices and procedures. These components would be partitioned in the accompanying classifications:

2.2 Information Architecture Issues

a. Likeness of Source and Target Data Structures**:** The more the source information structure varies from the one of the objective information, the more intricate the conventional ETL preparing and upkeep exertion become. Because of the various structures, the heap procedure will normally need to parse the records, change esteems, approve values, substitute code esteems and so forth.

b. Nature of Data**:** Regular information quality issues incorporate missing qualities; code esteems not right rundown of qualities, dates and referential honesty issues. It looks bad to stack the information distribution centre with low quality information. For instance, if the information stockroom will be utilized for database showcasing, the addresses ought to be approved to maintain a strategic distance from brought email back. [51]

c. Unpredictability of the Source Data**:** Contingent upon the sourcing groups foundation, a few information sources are more unpredictable than others. Instances of complex sources may incorporate various record types, bit fields and pressed decimal fields. This sort of information will convert into necessities of the ETL instrument or exceptionally composed arrangement since they are probably not going to exist in the objective information structures. People on the sourcing group that are new to these sorts may need to do some examination in these zones.

d. Conditions in the Data**:** Conditions in the information will decide the request wherein you load tables. Conditions likewise will in general lessen equal stacking tasks, particularly if information is converged from various frameworks, which are on an alternate business cycle. Complex conditions will likewise will in general make to stack forms increasingly intricate, energize bottlenecks and make bolster progressively troublesome.

e. Meta Data: Specialized Meta information portrays not just the structure and configuration of the source and target information sources, yet in addition the planning and change rules between them. Meta information ought to be noticeable and usable to the two projects and individuals.

2.3 Application Architecture Issues

a. Logging**:** ETL procedures should log data about the information sources they read, change and compose. Key data incorporates date handled, number of lines read and composed, mistake that experienced, and leads applied. This data is basic for quality affirmation and fills in as a review trail. The logging procedure ought to be sufficiently thorough with the goal that you can follow information in the information distribution centre back to the source. Moreover, this data ought to be accessible as the procedures are racing to aid the fulfilment times.
b. Notice: The ETL necessities ought to indicate what makes an adequate burden. The ETL procedure ought to advise the fitting help individuals when a heap falls flat or has blunders. In a perfect world, the notice procedure should plug into your current mistake global positioning framework.

c. Cold beginning, warm beginning: Tragically, frameworks do crash. You should have the option to make the fitting move if the framework crashes with your ETL procedure running. Fractional burdens can be truly a torment. Contingent upon the size of your information distribution centre and volume information, you need to begin once again, known as cool beginning, or start from the last known effectively stacked records, known as warm-start. The logging procedure ought to give you data about the condition of the ETL procedure.
d. Individuals Issues the executives' solace level with innovation**:** How familiar is your administration with information warehousing design? Will you have an information distribution centre chief? Does the executive have advancement out of sight? They may recommend doing all the ETL forms with Visual Basic. Solace level is a legitimate concern, and these worries will compel your choice. [52]
e. In-House aptitude:
ETL arrangements will be drawn from current originations, abilities and toolsets. Procuring, changing and stacking the information stockroom is a progressing procedure and should be kept up and stretched out as increasingly branches of knowledge are added to the information distribution centre.
f. Backing**:** When the ETL forms have been made, support for them, in a perfect world, it should plug into a current help structures, incorporating individuals with proper aptitudes, warning instruments and mistake global positioning frameworks. On the off chance that you utilize an apparatus for ETL, the care staff may should be prepared. The ETL procedure ought to be recorded, particularly in the region of inspecting data in innovation architecture issues.
g. Interoperability between stages**:** There must be a route for frameworks on one stage to converse with frameworks on another. FTP is a typical method to move information starting with one framework then onto the next. FTP requires a physical system way starting with one framework then onto the next just as the web convention on the two frameworks. Outside information sources for the most part please a floppy tape or a web worker.
h. Volume and recurrence of burdens**:** Since the information stockroom is stacked with clump programs, a high volume of information will in general decrease the group window. The volume of information additionally influences the retreat and recuperation work. Quick burden programs decrease the time it takes to stack information into the information distribution centre.
i. Circle space: Not exclusively does the information distribution centre have necessities for a great deal of circle space, yet there is likewise a ton of concealed plate space required for arranging regions and middle of the road records. For instance, you might need to remove information from source frameworks into level documents and afterward change the information to other level records for load.

j. Planning**:** Stacking the information stockroom could include several sources documents, which begin on various frameworks, utilize distinctive innovation and created at various occasions. A month to month burden might be regular for certain parts of the stockroom and a quarterly burden for other people. A few burdens might be on request, for example, arrangements of items or outer information. Some concentrate projects might be run on an alternate sort of framework than your scheduler. [42] [43]

2.4. ETL monitoring system through AI support

These are simply many ways that during which AI and machine learning will catch ETL errors before they be converted into inaccurate analytics.

i. Discover and Alert Across ETL Metrics**:** Even though the information could be a perpetually motion-picture show, the ETL method ought to still turn out consistent values at a homogenous speed. Once this stuff amendment, it's cause for alarm. Humans will see huge swings within the information and acknowledge errors, however machine learning will acknowledge subtler faults, faster. It's doable for a machine learning system to supply time period anomaly detection and alert the IT department directly, permitting them to pause the method and remedy the difficulty while not having to discard hours of procedure effort.[50] [51]

ii. Pinpoint Specific Bottlenecks**:** Even if your results are correct, they could still start up too slowly to be of use. Gartner says that eightieth of insights derived from analytics can ne'er be controlled to form value, which is also as a result of a king can't see AN insight in time to require advantage of it. Machine learning will tell you wherever your system is speed down and supply you with answers — obtaining you higher information, faster.[53]

iii. Enumerate the Impact of amendment Supervision**:** The systems that turn out your information and analytics aren't static — they perpetually receive patches and upgrades. Sometimes, these have an effect on the approach that they turn out or interpret information — resulting in inaccurate results. Machine learning will flag results that have modified and trace them to the particular patched machine or application.[54]

Optimization of the information integration platform by uptake AI into it improve execution performance by simplifying the event lifecycle, reducing the training time for the technology, and lowering the dependency on high ability demand for ETL progress creation. Another notable advantage is millilitre will train the information set to form it apt for configuration of applied mathematics modelling on that with none manual intervention, therefore assuaging the human obligatory problems. Consecrations of AI with millilitre conjointly include: Reduction within the integration total value of possession and timeline because of a decrease within the usage quality and management of business users to perform the DI with less or no help from technical consultants. [55]

Access to a spread of pre-packaged and configurable knowledge integration templates imbibed into AI for optimized alignment alongside intuitive and self-guiding steps for straightforward readying of knowledge integration and application tasks Conversational user experiences that enhance potency through the creation of power-assisted integration procedure and querying the platform for its operational state. This assists the business leaders of various departments in a company to attach to the system and make their own knowledge structures and application severally for any of their specific knowledge curtain and analysis wants.

As missing data and error data are the major issues of the ETL system, the proposed work is based Multinomial Naïve Bayes algorithm for identifying missing and error data.

Iv. Multinomial Naïve Bayes: This algorithm comprises of two features known as data bag of assumptions and constraint independence based on this data have multi-feature analysis capability of considering multiple parameters to identify the nature of the data . Here a missing is related multiple level data and an error data occurrence is matched with aspects with prefix and suffix data to confirm the presence of data. Each data $d_i$is drawn from the multinomial distribution of data this leads to familiar ideology data bag. Here $N_{it}$is to calculate of data that is related to the similar data w_t which present in database d_i the probability of the database derived from its class form

$$\mathrm{P}(d_i|c_j;\theta) = \mathrm{P}(|d_i|)|d_i|!\prod_{t=1}^{|V|}\frac{\mathrm{P}(w_t|c_j;\theta)^{N_{it}}}{N_{it}!}$$

Here the p is the probability of data that may occur in the sequence and same probability of data that supports the data presence in that sequence ,once again Bayes optimal is calculated for the parameters that present ,the evaluation

of probability of data w_t and corresponding class of class c_j is included [58] [59]

$$\hat{\theta}_{w_t|c_j} = \mathrm{P}(w_t|c_j;\hat{\theta}_j) = \frac{1+\sum_{i=1}^{|\mathcal{D}|} N_{it}\mathrm{P}(c_j|d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|\mathcal{D}|} N_{is}\mathrm{P}(c_j|d_i)}.$$

Based on this data relativity equation the missing data is meant to identified among the various data .Comparing with other naïve Bayes algorithm Multinomial Bayes algorithm has potential to identify the data which is error and which is missing.[57] [60]

3. Conclusion and Future Enhancement

In this proposed work hurdles of ETL system is revealed in detail. As the data strength is being increased by data by data .The constraints for this system is also being increased in this work both the tasks and solution for the tasks are recommended with the support artificial intelligence algorithm, the AI algorithms proves to be a handy solution various tasks. In this work Multinomial Bayes algorithm is incorporated for identifying the missing data and error data, which is a novel approach for this challenge.

In future furthermore amalgamated algorithms can be utilized and implemented for better performing ETL system, which will enhance better solutions in future constraints

References

1. The Challenges of Extract, Transform and Loading (ETL) System Implementation For Near Real-Time Environment A Back Room Staging for Data Analytics Adilah Sabtu*1,2, Nurulhuda Firdaus Mohd Azmi1,2, Nilam Nur Amir Sjarif1,2, Saiful Adli Ismail1,2, Othman Mohd Yusop1 , Haslina Sarkan1 , Suriayati Chuprat1Advanced Informatics School (UTM AIS) 2 Machine Learning for Data Science Interest Group (MLDS) Universiti Teknologi Malaysia (UTM) Jalan Sultan Hj Yahya Petra, 54100 Kuala Lumpur, Malaysia978-1-5090-6255-3/17/$31.00 ©2017 IEEE
2. The Challenges of Extract, Transform and Load (Etl) For Data Integration In Near Realtime Environment AdilahSabtu*1,2, Nurulhuda Firdaus Mohd Azmi1,2, Nilam Nur Amir Sjarif1,2, Saiful Adli Ismail1,2, Othman Mohd Yusop1 , Haslina Sarkan1 , Suriayati Chuprat1 1 Advanced Informatics School (UTM AIS), Universiti Teknologi Malaysia (UTM), Malaysia 2Machine Learning for Data Science Interest Group (MLDS), Universiti Teknologi Malaysia (UTM), Malaysia, Journal of Theoretical and Applied Information Technology 30th November 2017. Vol.95. No 22
3. Big Data ETL Implementation Approaches: A Systematic Literature Review, Joshua C. Nwokeji∗ , Faisal Aqlan† , Anugu Apoorva∗ , and Ayodele Olagunju† ∗Comp. & Info. Sys. Dept. Gannon Uni. † Indus.Engr., Dept., Penn. State Uni. ‡ Uni., of Saskatchewan; DOI reference number: 10.18293/SEKE2018-152
4. Data quality in ETL process: A preliminary study Manel Souibguia,b,∗ , Faten Atiguib, Saloua Zammalia , Samira Cherfib, Sadok Ben Yahiaa, aUniversity of Tunis El Manar, Faculty of Sciences of Tunis LIPAH-LR11ES14, Tunis, Tunisia, bConservatoire National des Arts et M´etiers CEDRIC-CNAM, Paris, France.
5. Next-generation ETL Framework to address the challenges posed by Big Data Syed Muhammad Fawad Ali Poznan University of Technology Poznan Poland trivago N.V. Leipzig Germany, © 2018 Copyright held by the owner/author(s). Published in the Workshop Proceedings of the EDBT/ICDT 2018 Joint Conference (March 26, 2018, Vienna, Austria) on CEUR-WS.org (ISSN 1613-0073). Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.
6. A Fine-Grained Distribution Approach for ETL Processes in BigData Environments Mahfoud Balaa,∗, Omar Boussaidb, Zaia Alimazighic a Department of informatics, Saad Dahleb University, Blida 1, Blida, Algeria b University of Lyon 2, Lyon, France c Department of informatics, USTHB, Algiers, Algeria, Data & Knowledge Engineering 111 (2017) 114–136, Data & Knowledge Engineering.
7. The Opportunities and Challenges of Information Extraction Qian Zhu, Xianyi Cheng School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013,China, International Symposium on Intelligent Information Technology Application Workshops, 978-0-7695-3505-0/08 $25.00 © 2008 IEEE DOI 10.1109/IITA.Workshops.2008.165.
8. Challenges from Information Extraction to Information Fusion, Heng Ji Computer Science Department Queens College and Graduate Center City University of New York, Coling 2010: Poster Volume, pages 507–515, Beijing, August 2010.

9. Limitations of information extraction methods and techniques for heterogeneous unstructured big data, Kiran Adnan and Rehan Akbar, International Journal of Engineering Business Management Volume 11: 1–23 The Author(s) 2019 DOI: 10.1177/184797901989077.
10. Heterogeneous Data and Big Data Analytics, Lidong Wang*, Department of Engineering Technology, Mississippi Valley State University, Itta Bena, MS, USA, Automatic Control and Information Sciences, 2017, Vol. 3, No. 1, 8-15, Science and Education Publishing DOI:10.12691/acis-3-1-3.
11. https://docs.oracle.com/cd/B10501_01/server.920/a96520/extract.htm
12. https://tdan.com/extraction-transformation-and-load-issues-and-approaches/4839#
13. https://blog.datahut.co/web-scraping-at-large-data-extraction-challenges-you-must-know/
14. Problems and Available Solutions On The Stage of Extract, Transform, and Loading In Near Real-Time Data Warehousing (A Literature Study) Ardianto Wibowo Department of Informatics Engineering Politeknik Caltex Riau Pekanbaru, Indonesia, 2015 International Seminar on Intelligent Technology and Its Applications, 978-1-4799-7711-6/15/,2015 IEEE.
15. https://www.stitchdata.com/resources/data-transformation/
16. https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184
17. Don't Let Your Data Handle You: A Novel Approach to Clinical Programming, Jorine Putter, Grünenthal GmbH, Aachen, Germany Michael S Rimler, GlaxoSmithKline, Cincinnati, Ohio, US.
18. https://enterprisevisions.com/saas-presents-unexpected-data-management-challenges/
19. https://sagetech.com.hk/?page_id=374
20. https://docs.oracle.com/en/cloud/paas/bi-cloud/bilpd/troubleshooting-administration-issues.html
21. Data Transfers Between Incompatible Operating Systems Michael D. Chase Accounting Department, California State University, Long Beach, CA 90840, U.S, Computers and the Humanities 22 (1988) 153-156. a 1988 by KluwerAcademic Publishers.
22. https://www.cloverdx.com/blog/biggest-data-integration-challenges
23. https://www.talend.com/resources/what-is-data-integration/
24. Data Integration: A Theoretical Perspective Maurizio Lenzerini Dipartimento di Informatica e Sistemistica Universita di Roma "La Sapienza" ` Via Salaria 113, I-00198 Roma, Italy lenzerini@dis.uniroma1.it, Conference Paper · January 2002 DOI: 10.1145/543613.543644 · Source: DBLP
25. Data Integration - Challenges, Techniques and Future Directions: A Comprehensive Study, 1 Faculty of Computer Science and Engineering, Sathyabama University, Chennai, School of Information Technology and Engineering, VIT University, Vellore, Indian Journal of Science and Technology, Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/105314, November 2016.
26. https://blog.altran.es/telecomunicaciones-media/five-data-migration-challenges/
27. https://www.experian.co.uk/blogs/latest-thinking/data-and-innovation/8-hurdles-of-a-data-migration/
28. https://www.scnsoft.com/blog/data-warehouse-implementation
29. https://mapr.com/blog/what-future-data-warehousing/
30. Data Warehouses: Next Challenges January 2012Lecture Notes in Business Information Processing 96 DOI: 10.1007/978-3-642-27358-2,Alejandro Vaisman,Esteban Zimanyi
31. https://acadgild.com/blog/6-steps-in-data-wrangling
32. A Systematic Study of Data Wrangling Malini M. Patil, Associate Professor, Dept. of Information Science and Engineering, Basavaraj N. Hiremath, Research Scholar, Dept. of Computer Science and Engineering, JSSATE Research Centre, J J.S.S Academy of Technical Education, Bengaluru, Karnataka, I.J. Information Technology and Computer Science, 2018, 1, 32-39 Published Online January 2018 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.2018.01.04.
33. Research directions in data wrangling: Visualizations and transformations for usable and credible data Sean Kandel1 , Jeffrey Heer1 , Catherine Plaisant2 , Jessie Kennedy3 , Frank van Ham4 , Nathalie Henry Riche5 , Chris Weaver6 , Bongshin Lee5 , Dominique Brodbeck7 and Paolo Buono8, Information Visualization 0(0) 1–18 ! The Author(s) 2011 Reprints and permissions: sagepub.co.uk/journalsPermissions.nav DOI: 10.1177/1473871611415994 ivi.sagepub.com.
34. Data Wrangling: Making data useful again,FlorianEndel[a]HaraldPiringer[b] UniversityofTechnologyVienna(florian.endel@tuwien.ac.at)∗∗VRVisResearchCenter,Vienna,Austria,Science Direct ,IFAC-PapersOnLine 48-1 (2015) 111–112.
35. https://www.elderresearch.com/blog/what-is-data-wrangling
36. Data Wrangling Jeffrey Heer1, Joseph M. Hellerstein2, and Sean Kandel3 1University of Washington, Seattle, WA, USA 2University of California, Berkeley, Berkeley, CA, USA 3Trifacta. Inc, San Francisco, CA, USA,Springer International Publishing AG 2018 S. Sakr, A. Zomaya (eds.), Encyclopedia of Big Data

Technologies, https://doi.org/10.1007/978-3-319-63962-8_9-1.
37. Data Wrangling in Database Systems: Purging of Dirty Data Otmane Azeroual German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, Berlin 10117, Germany.
38. https://datadrivenperspectives.com/the-challenges-of-loading-a-data-warehouse-in-real-time-d925d8458ab5
39. https://rivery.io/7-challenges-to-load-data-to-the-cloud-and-how-to-overcome-them/
40. A Survey of Extract–Transform– Load Technology Panos Vassiliadis, University of Ioannina, Greece,International Journal of Data Warehousing & Mining, 5(3), 1-27, July-September 2009
41. https://www.networkworld.com/article/3213729/etl-is-slowing-down-real-time-data-analytics.html
42. DBMS Data Loading: An Analysis on Modern Hardware Adam Dziedzic1? , Manos KarpathiotakiIoannis Alagiannis2 , Raja Appuswamy2 , and Anastasia Ailamaki,1 University of Chicago, Ecole Polytechnique Fed´ erale de Lausanne (EPFL),RAW Labs SA.
43. Issues with data and analyses: Errors, underlying themes, and potential solutions Andrew W. Browna,1, Kathryn A. Kaisera,2, and David B. Allisona,3,4 a Office of Energetics and Nutrition Obesity Research Center, University of Alabama at Birmingham, Birmingham, AL 35294 Edited by Victoria Stodden, University of Illinois at Urbana–Champaign, Champaign, IL, and accepted by Editorial Board Member Susan T. Fiske November 27, 2017 (received for review July 5, 2017).
44. Analysis of Data Errors in Clinical Research Databases,Saveli I. Goldberg, PhD,[a] Andrzej Niemierko, PhD,[a,d] and Alexander Turchin, MD, MS[b,c,]AMIA 2008 Symposium Proceedings Page – 242
45. https://www.datavail.com/blog/4-issues-that-can-negatively-affect-your-etl-processes/
46. https://www.healthitanswers.net/etl-challenges-within-healthcare-business-intelligence/
47. https://www.glowtouch.com/etl-and-data-warehousing-challenges/
48. https://bfsi.cioreviewindia.com/cioviewpoint/challenges-in-following-etl-process-during-data-warehousing-nid-1310-cid-1.html
49. A comparative study of various ETL process and their testing techniques in data warehouse,Sonali Vyas * Pragya Vaishnav† Amity University Jaipur India,Journal of Statistics & Management Systems Vol. 20 (2017), No. 4, pp. 753–763 DOI : 10.1080/09720510.2017.1395194
50.]https://datavirtuality.com/blog-etl-tools-and-processes/
51. https://www.xplenty.com/blog/top-7-etl-tools/
52. The Process of Data Mapping for Data Integration Projects Data Mapping - A Key Work Product for Data Warehouse, Data Integration, and Data Migration Projects, Wayne Yaddow Data Quality Analyst – Consultant,Method · October 2019 DOI: 10.13140/RG.2.2.10352.81925.
53. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study Qingyu Chen*, Justin Zobel and Karin Verspoor, Department of Computing and Information Systems, The University of Melbourne, Parkville, VIC, 3010, Australia,Published by Oxford University Press.
54. Defining enterprise AI: From ETL to modern AI infrastructure, ByBob Violino
55. https://www.techopedia.com/4-ways-ai-driven-etl-monitoring-can-help-avoid-glitches/2/33969
56. https://www.wipro.com/blogs/krishna-kumar-aravamudhan/the-power-of-artificial-intelligence-in-data-integration-platforms/
57. Role of Machine Learning in ETL Automation, Kartick Chandra Mondal Jadavpur University, Neepa Biswas Jadavpur University, Department of Information Technology,Swati Saha Tata Consultancy Services Limited., ICDCN 2020, January 4–7, 2020, Kolkata, India.
58. https://www.cs.ubc.ca/~murphyk/Bayes/old.bnsoft.html
59. 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 , Patrick Hosokawa2 , Chris Uhrich6 and Lisa M. Schilling3, Ong et al. BMC Medical Informatics and Decision Making (2017) 17:134 DOI 10.1186/s12911-017-0532-3.
60. A Comparison of Event Models for Naive Bayes Text Classification, Andrew McCallum‡†Just Research, Kamal Nigam, †School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213.
61. https://towardsdatascience.com/ml-algorithms-one-sd-%CF%83-bayesian-algorithms-59785da792a
62. Nagarajan, G., Minu, R. I., & Devi, A. J. (2020). Optimal Nonparametric Bayesian Model-Based Multimodal BoVW Creation Using Multilayer pLSA. Circuits, Systems, and Signal Processing, 39(2), 1123-1132.
63. Nagarajan, G., & Minu, R. I. (2018). Wireless soil monitoring sensor for sprinkler irrigation automation system. Wireless Personal Communications, 98(2), 1835-1851.
64. Nagarajan, G., & Thyagharajan, K. K. (2012). A machine learning technique for semantic search engine. Procedia engineering, 38, 2164-2171.

65. Nagarajan, G., R. I. Minu, V. Vedanarayanan, SD Sundersingh Jebaseelan, and K. Vasanth. "CIMTEL-mining algorithm for big data in telecommunication." International Journal of Engineering and Technology (IJET) 7, no. 5 (2015): 1709-1715.