**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

# ETL in a World of Unstructured Data: Advanced Techniques for Data Integration

**Dhamotharan Seenivasan**[1*]

Abstract

The use of unstructured data in the form of text, images, audio and video is increasing making it difficult for traditional structures of data analysis. ETL, as one of the integral components of information integration, is aimed at processing unstructured data to provide it in structured forms, and therefore, this paper focuses on unstructured data ETL techniques. Based on the development of natural language processing for machine learning and deep learning, this paper reveals the most important methods from ETL for effective unstructured data analysis and decision support. The discussion also presents information on the most recent tools and applications regarding the ETL for unstructured data, in addition to presenting real-life examples of their usage. Also, it describes the difficulties and opportunities of the ETL process concerning unstructured data and discusses how organizations can benefit from its utilization and become leaders in their respective industries.

***Keywords:***

Unstructured Data,
ETL,
Data extraction,
Data Transformation,
Data Loading,
Natural Language Processing.



***Author correspondence:***

Dhamotharan Seenivasan,
Project Lead-Systems,
Mphasis, Plano, Texas, USA.

Email:
dhamotharranvs@gmail.com

[1*] Dhamotharan Seenivasan, Project Lead-Systems, Mphasis, Plano, Texas, USA.

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

## 1. Introduction

The explosive increase in the volume of unstructured data in the form of text, images, audio and videos creates a headache for traditional data processing and analysis. Unstructured data differs from structured data, which is in recognized fields of a record or file and, therefore, cannot be evaluated via common ETL procedures. Consequently, this paper reviews different ETL techniques used in the processing of unstructured data where the tools, together with methods that support the extraction, transformation as well as loading of the data to structured means, are discussed.

### 1.1 Definition and Characteristics

Unstructured Data: Not based on most familiar and widely used database systems (relational and column-oriented), no optimal schema, not friendly with SQL and most of the traditional DBMS. Inherent Organization: It is, however, unstructured most of the time, though it may have some degree of internal structure/organization.

1.1.1 Types of Unstructured Data

1.1.1.1 Textual Data:

- Examples: All messages, posts, reviews, and the content of the web page.
- Components: Sender and recipient, date, the content of the email, and even possible attachments.

1.1.1.2 Multimedia Data:

- Examples: BMP, jpg, png, GIF, MP3, WAV, MPEG, AVI, ASF.
- Processing: Commonly needs other methods, such as signal processing.

1.1.1.3 Sensor Data:

- Examples: Information from IoT devices, its surroundings: telemetry data, environmental sensors.
- Nature: Generally real-time and extremely 'unstructured'.

1.1.1.4 Implex File Types:

- Examples: Any document format, PDF, a Word document, even a PowerPoint presentation.
- Content: Text incorporating poems, photos and their descriptions in the metadata.

1.1.2 Challenges in ETL for Unstructured Data

- Extraction: Complex mainly because of the non-fixed patterns, structurally irregular data, and different sizes of data and meanings.
- Transformation: Necessitates tasks such as text analysis and analysis of audio signals.
- Loading: Problems such as propagation of storage costs and data integrity problems, as well as problems arising from difficulty in posing future queries in the indexed structure.

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

## 2. Literature Survey

The information about the ETL techniques that are related to unstructured data includes numerous strategies and instruments that are related to various types of unstructured data. NLP has been analyzed for its functions primarily in dealing with stringed data to deduce valuable information. Work done by Bird et al. (2009) and Honnibal & Montani (2017) can be referred to in proving how tokenization, named entity recognition and sentiment analysis help in transforming the text data. The study by Lecun et al. (1998) and Krizhevsky et al. (2012) show how machine learning, specifically CNN, can successfully capture features from raw image data.

The domain of audio and speech processing: Some papers by Hinton et al. (2012) and Amodei et al. (2016) speak about the progress in the sphere of speech-to-text and speaker identification. As for video data, two works by Karpathy et al. (2014) and Simonyan and Zisserman (2014) are published on action recognition and video summarization.

The features of these samples of unstructured data processing have been made possible by various tools and libraries for integration into ETL systems. Hadoop and Spark are the platforms commonly used for batch processing of big data in large volumes; tools such as Apache, NiFi, and Talend are used for data manipulation and ingestion. There is versatility in using NoSQL databases like MongoDB to store unstructured data as well as Elasticsearch.

## 3. Methodology

### 3.1. Unstructured Data Extraction Techniques

Unstructured data extraction has its importance in extracting meaningful information from data which are not characterized by a structure in any form. Such data is widespread in different forms and applications: textual, image, voice and video. Here is a look at each of the outlined techniques:

#### 3.1.1. Natural Language Processing (NLP)

NLP is a broad area of computer science whose objective is to make computers understand human language comprehensively to identify and analyze text data.

1. Tokenization: This involved segmenting the text material into more manageable components like word components or phrase components. Tokenization is the most basic feature in text processing; through it, subsequent processing is done, such as word frequency count or syntax analysis.

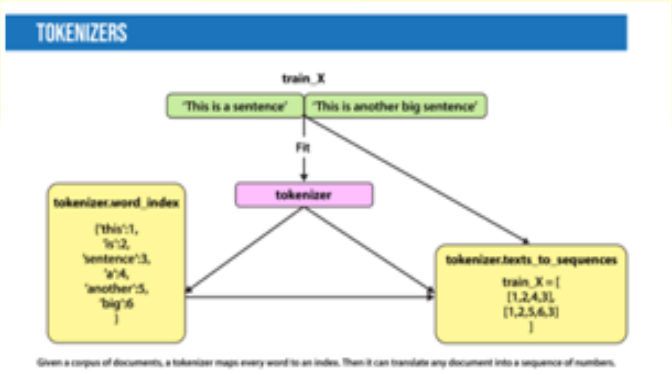


**Figure 1: Tokenization process**

**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

2. Named Entity Recognition (NER): Annotation of Named Research is the process of tagging entities that are specific, such as names, dates, and places, among others. This is useful in arriving at specific information as well as the bigger picture.
3. Part-of-Speech Tagging (POS): POS tagging is the process of tagging each word of a given text according to its grammatical classes (e.g., nouns, verbs, adjectives). This helps with syntactic as well as semantic analysis and promotes a better understanding of the texts.
4. Sentiment Analysis: Sentiment analysis defines the sentiment involved in the body of text. It can classify text as positive, negative, or neutral, which assists businesses in digesting feedback from customers and evaluating public opinion.
5. Topic Modeling: Latent Dirichlet Allocation (LDA) discovers topics which are latent in a set of documents from a collection. This is applied when there is a need to compress big data sets and identify patterns.

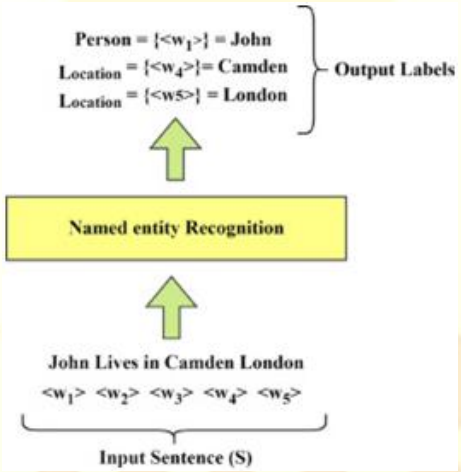


**Figure 2: Named Entity Recognition process**

**Table 1: Natural Language Processing (NLP)**

| Technique | Description | Tools/Libraries |
|---|---|---|
| Tokenization | Analyzing the text into words or phrases. | NLTK, spaCy |
| Named Entity Recognition | Annotating protocol that is aimed at defining and categorizing such items as names, dates, and places in a text. | NLTK, spaCy, Stanford NLP |
| Part-of-Speech Tagging | Tagging of the text to identify the parts of speech | NLTK, spaCy, Stanford NLP |
| Sentiment Analysis | Classification of the nature of feeling conveyed in a piece of text. | NLTK, spaCy, BERT |
| Topic Modeling | Topics from collections of documents that include abstract concepts (for example, LDA) | Gensim, MALLET |

3.1.2 Text Mining and Information Retrieval

Knowledge discovery and data mining deal with obtaining meaningful information and discovering patterns in large data sets, especially textual data. Text Mining and Information Retrieval are shown in Table 2.

**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

- Regular Expressions: Actually, regular expressions are strings of characters that determine the pattern that would be looked for. They are employed in pattern matching to enable one to extract specific information like email addresses or even dates from the text.
- Keyword Extraction: This is composed of the selection of certain significant words or phrases that can be regarded as highly significant in a text. Some of the frequently adopted methods include, for instance, the TF-IDF (Term Frequency-Inverse Document Frequency).
- Text Summarization: The generation of an abstract can entail the creation of an entire text that is much less extensive than the initial document but contains all the necessary information. Annotations, for example, can be of two types: extractive and abstractive, the former is used when only important sentences must be identified; the latter is used when new sentences should be created.
- Document Clustering: Document clustering, in turn involves the categorization of documents depending on the information contents contained in them. This can be applied in managing big datasets whereby crucial information can easily be obtained and sorted.

**Table 2: Text Mining and Information Retrieval**

| Technique | Description | Tools/Libraries |
|---|---|---|
| Regular Expressions | Pattern matching for extracting specific patterns from text | Python re |
| Keyword Extraction | Identifying the most relevant keywords in the text | RAKE, YAKE |
| Text Summarization | Creating a concise summary of a larger text document | Gensim, Sumy |
| Document Clustering | Grouping similar documents together based on content | Scikit-learn |

3.1.3 Machine Learning and Deep Learning

Artificial intelligence and deep learning approaches employ the usage of algorithms as well as artificial neural networks to learn and make decisions from data. Machine Learning and Deep Learning Table 3.

- Classification and Clustering: Classification deals with categorizing an SMS into two defined types, e.g. spam or not spam, while clustering is concerned with regrouping text with similar content, e.g. news articles with similar content.
- Sequence-to-Sequence Models: These models are commonly applied in tasks such as language translation or text summarization, where the output sequence is defined by the input sequence.
- Transformers (e. g., BERT, GPT): Auto-encoder models and transformer models currently stand out in the field of NLP because of their application of understanding and generating human language. BERT is good at addressing contextual investigation, while GPT is good at offering outcomes of the text-designing interaction.

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

**Table 3: Machine Learning and Deep Learning**

| Technique | Description | Tools/Libraries |
|---|---|---|
| Classification/Clustering | Categorizing text into predefined classes or clusters | Scikit-learn, TensorFlow, PyTorch |
| Sequence-to-Sequence Models | Used in translation and summarization tasks | TensorFlow, PyTorch |
| Transformers | Advanced models for understanding and generating human language | BERT, GPT-3, Transformers |

3.1.4 Computer Vision

Computer vision interventions are used for image and video data analysis and interpretation Table 4.

- Optical Character Recognition (OCR): OCR is useful in converting various forms of text; these are usually scanned papers or images of paper, into textual information that can be manipulated.
- Image Classification: This can be defined as the task of assigning pictures into pre-specified categories (e.g., determining whether the picture is of a cat or a dog).
- Object Detection: Object detection is all about the discovery and pinpointing of objects in images. Some of the most applied methods are YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector).
- Image Segmentation: Segmentation of images divides an image into different segments or regions to selectively or selectively extract certain objects or parts of the image that are of interest.

**Table 4: Computer Vision**

| Technique | Description | Tools/Libraries |
|---|---|---|
| Optical Character Recognition (OCR) | Extracting text from images | Tesseract, OpenCV |
| Image Classification | Categorizing images into predefined classes | TensorFlow, PyTorch |
| Object Detection | Identifying and locating objects within images | YOLO, SSD, OpenCV |

3.1.5 Audio and Speech Processing

Techniques of audio or speech processing are types of signal processing that deal with the analysis and interpretation of audio signals Table 5.

- Speech-to-Text: They facilitate converting the spoken language into written forms in speech-to-text systems. Examples are in the typist's services and the voice-command systems.
- Speaker Identification: This technique tags the speaker in an audio clip. It is used in security and clients' customized interfaces.
- Audio Classification: Audio classification means the process of distribution of sounds into different classes depending on their type (for instance, different music genres, various sounds in the environment).

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

**Table 5: Audio and Speech Processing**

| Technique | Description | Tools/Libraries |
|---|---|---|
| Speech-to-Text | Converting spoken language into written text | Google Speech API, DeepSpeech, CMU Sphinx |
| Speaker Identification | Identifying who is speaking in an audio clip | Kaldi, PyTorch |
| Audio Classification | Classifying sounds into categories | LibROSA, TensorFlow |

3.1.6 Video Processing

Abstract video analysis methods are used in pattern recognition to obtain a perfect output from imperfect videos Table 6.

- Video Summarization: Video summarization helps to use summaries for longer videos in which the important events or scenes are highlighted. This is helpful for buffering through video material.
- Action Recognition: Activity recognition, on the other hand, deals with the identification of certain activities present in a video or a scene, such as walking, running and jumping. It is used in surveillance and in recording data during sports activities.
- Object Tracking: Tracking objects enables the movement of the objects observed in the subsequent frames of the video, which remains critical in applications such as traffic surveillance and detection of people's actions.

**Table 6: Video Processing**

| Technique | Description | Tools/Libraries |
|---|---|---|
| Video Summarization | Creating short summaries of longer videos | OpenCV, PyTorch |
| Action Recognition | Identifying specific actions or activities in videos | OpenCV, TensorFlow |
| Object Tracking | Tracking the movement of objects across video frames | OpenCV, dlib |

3.17 Data Integration and Transformation

These techniques include performing operations on the data obtained from multiple sources in order to put it in a common format Table 7.

- ETL (Extract, Transform, Load): Data extraction, transformation and loading, otherwise referred to as ETL, involves pulling data from various sources, modifying it and then transferring it into a desired system. This is important in data warehousing and business intelligence processes.
- APIs and Web Scraping: APIs are methods that allow getting data from other online services with the help of programs, and web scraping is a way to gather data from web sites. Both are significant for collecting information from the World Wide Web.

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

**Table 7: Data Integration and Transformation**

| Technique | Description | Tools/Libraries |
|---|---|---|
| ETL (Extract, Transform, Load) | Processes for integrating and transforming data from various sources into a unified format | Apache Nifi, Talend |
| APIs and Web Scraping | Extracting data from websites and online services | BeautifulSoup, Scrapy |

### 3.2 Unstructured Data Transformation Techniques

Unstructured data transformation relates to the process of converting large volumes of unstructured data into forms that can be analyzed to support several uses. This process is important for expanding the possibilities of analyzing data, developing machine learning models and making better decisions. Here is an elaboration on key techniques for transforming unstructured data:

3.2.1. Text Transformation Techniques

3.2.1.1 Tokenization

- Description: Developing a method through which the text can be broken down into individual components which we may term as tokens, which can be words, phrases, or even sentences.
- Use Case: Preprocessing technique applied to texts to make them more suitable for analysis such that downstream processes such as word frequency count can be done.
- Tools/Libraries: NLTK, spaCy.

3.2.1.2 Stemming and Lemmatization

- Description: Closely related to this is a process of transforming a given word to its base or root form. Stemming reduces the word to its stem, while lemmatization, on the other hand, looks at the words' context and reduces it to the base/dictionary form.
- Use Case: Enhances decision making by standardizing words, helpful where like in searching for information.
- Tools/Libraries: NLTK, spaCy.

3.2.1.3 Stop Word Removal

- Description: The most typical presorting actions include the elimination of unimportant words that fill up a text but do not convey much information, like "the", "is".
- Use Case: Removes unwanted characters in the text data, hence making the word features defining and easily distinguishable.
- Tools/Libraries: NLTK, spaCy.

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

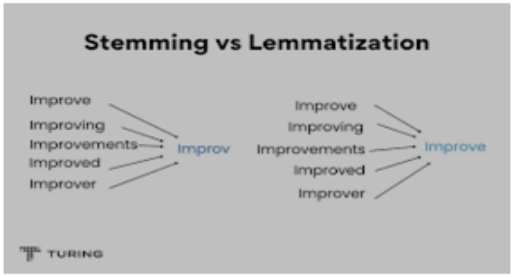


**Figure 3: Stemming and Lemmatization**

| Sample text with Stop Words | Without Stop Words |
|---|---|
| GeeksforGeeks – A Computer Science Portal for Geeks | GeeksforGeeks , Computer Science, Portal ,Geeks |
| Can listening be exhausting? | Listening, Exhausting |
| I like reading, so I read | Like, Reading, read |

**Figure 4: Stop Word Removal**

3.2.1.4 Text Normalization

- Description: Substituting the given text for a standard format, for instance, by changing super- and lowercase, eradicating punctuation marks.
- Use Case: Stabilises the possible variations in text data which must be very important to get the right results.
- Tools/Libraries: NLTK, spaCy.

3.2.1.5 N-grams

- Description: To capture the context, sequences of N words are to be developed. Some of the most typical ones are the so-called bigrams, in which there are two words, and trigrams, which consist of three words.
- Use Case: Enhances representation of context and of the associations between words, thus benefitting the ascription of language models and textual analysis.
- Tools/Libraries: NLTK, scikit-learn.

3.2.2. Feature Extraction from Text

3.2.2.1 Bag of Words (BoW)

- Description: A text representation where all the text is represented as a set of words, and no attention is paid to the grammar and the order of words.
- Use Case: One of the most straightforward and efficient ways to solve the problem of text classification and clustering.

3.2.2.2 Tools/Libraries: scikit-learn.

- Terms that are frequent in a collection of documents are assigned higher weights than those that appear in a few of the collections using TF-IDF (Term Frequency-Inverse Document Frequency).
- Description: Assigning relevance values in accordance with their patterns of occurrence and significance in documents.
- Use Case: Increases the importance of the specific words while decreasing the importance of the other frequently used words.
- Tools/Libraries: scikit-learn.

**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

3.2.2.3 Word Embeddings

- Description: Using words as vectors in a finite feature space in a way that avails semantic meaning.
- Use Case: Enhances functionalities such as determining the sentiment of the text, finding texts that are similar to each other and translation.
- Tools/Libraries: Word2Vec is a type of Word Embedding that includes training methods such as GloVe and fastText.

3.2.2.4 Document Embeddings

- Description: a basic form that amounts to storing and comparing whole documents as vectors to capture the information content.
- Use Case: Used for document classification, clustering and similarity operations are facilitated by this.
- Tools/Libraries: Doc2Vec, USE.

3.2.3 Data Transformation from Images

3.2.3.1 Optical Character Recognition (OCR)

- Description: Converting normal text in Images to Machine Readable Text.
- Use Case: Takes scanned papers and pictures and makes the data in them retrievable and analysable.
- Tools/Libraries: Tesseract OCR, OpenCV Mobile, Tesseract OCR for Android, and Tesseract OCR for iOS.

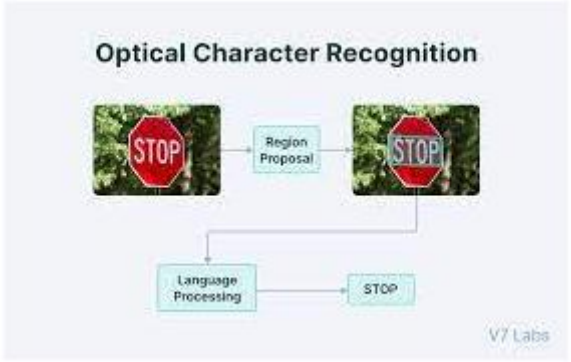


**Figure 5: Optical Character Recognition**

3.2.3.2 Image Annotation and Labeling

- Description: Associating text labels or indexing with the questioned images.
- Use Case: Among the applications using this concept, some common ones are image classification and particularly object detection.
- Tools/Libraries: LabelImg, OpenCV.

3.2.3.3 Feature Extraction

- Description: Some mirror the attributes found in other objects, while others present edges, textures or other characters in the pictures.
- Use Case: Crucial in tasks such as image classification and object recognition, among others.
- Tools/Libraries: OpenCV, TensorFlow.
-

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

3.2.4 Audio and Speech Transformation

3.2.4.1 Speech-to-Text

- Description: Transcribing of the spoken language, or, in other words, translating the speech into written language.
- Use Case: Facilitates the conversion of the recorded audio analysis for assessment and documentation.
- Tools/Libraries: Google Speech API, DeepSpeech, Chu Mey State University Sphinx.


**Figure 6: Audio and Speech Transformation**

3.2.4.2 Audio Feature Extraction

- Description: Features such as Mel-frequency cepstral coefficients (MFCCs) are to be extracted for the audio analysis that is to be done.
- Use Case: Enhance speech recognition, acoustic classification and music acoustic analysis.
- Tools/Libraries: LibROSA, PyDub.

3.2.4.3 Speaker Diarization

- Description: The method of splitting a continuous audio stream in order to differentiate between different speakers.
- Use Case: Applicable in meetings, interviews, and when there is more than one presenter.
- Tools/Libraries: Kaldi, pyAudioAnalysis.

3.2.5 Video Data Transformation

3.2.5.1 Video Captioning

- Description: Automating the process of identifying and creating textual tags for the video content.
- Use Case: It improves the availability and allows content search and summarization.
- Tools/Libraries: OpenCV, TensorFlow.

3.2.5.2 Frame Extraction

- Description: Picking out single pictures out of the video for evaluation.
- Use Case: Helps in analyzing videos frame-by-frame, segmentation of particular objects as well as identification of various actions.
- Tools/Libraries: OpenCV, FFmpeg.

**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

3.2.5.3 Action Recognition

- Description: Video methods in identification and labelling actions or events.
- Use Case: Appropriate for security systems, sports training and team analysis, and indexing of video material.
- Tools/Libraries: OpenCV, TensorFlow.

3.2.6 Structural Transformation

3.2.6.1 Schema Mapping

- Description: The conversion of data that does not have any specific structure or format into a format that is easily distinguishable and recognized.
- Use Case: This fastens the entry of the data into the structured databases while ensuring data consistency.
- Tools/Libraries: Some of the prominent big data integration tools are as follows: Tata TDWI TDM, Informatica, Teradata, Talend, and Apache NiFi.

3.2.6.2 Data Wrangling and Cleaning

- Description: Organizing clients' data to make them adhere to a predefined set of patterns.
- Use Case: Cleaning of data for analysis to correct invalid entries or to bring the data to the desired format.
- Tools/Libraries: Pandas, Dask.

3.2.7 Integration and Aggregation

3.2.7.1 Merging and Joining

- Description: It means the accumulation of multiple unstructured data into a single structure where all the data can be integrated.
- Use Case: Gathers data for cohesion to analyze them.
- Tools/Libraries: Pandas, Dask.

3.2.7.2 Aggregation

- Description: Coalescing the collected data to arrive at a generalizable mean or, in other words, capturing quantitative information.
- Use Case: Makes suggestions by means of the summarization of data.
- Tools/Libraries: Pandas, SQL.

## 4. Unstructured Data Loading Techniques

Transferring unstructured data to a system for analysis or processing involves several methods and instruments based on the nature of the data and the process that is to be performed on this data. Here is a detailed exploration of the key methods used to load unstructured data:

4.1. File-based Loading

4.1.1 Direct File Access

- Definition: Handling files through programming languages without the need to get assistance from operating systems.
- Use Cases: It is best for simple data processing activities and requires small and medium data sets at most.

**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Example Tools:

- Python: volatile variables, the open () function for textual databases, use of pandas in databases in CSV and Excel formats.
- Java: Using java. nio. file. Files for file operations.
- C++: File I/O stream, which means the input and output operation of a file.

4.1.2 Batch Processing

- Definition: To work with large amounts of data, the program should be designed to load the data in batches.
- Use Cases: High run rate, batch data feed.

Example Tools:

- Apache Hadoop: In the case of distributed batch processing of big data.
- Apache Spark: For analyzing batches of data received at the same time from users.

4.2. Database Loading

4.2.1 NoSQL Databases

- Definition: Originally dealing with unstructured data by means of using databases which are peculiar to such data.
- Use Cases: Versatility of the schema, the capability to accommodate data changes and improvements with ease, and efficient ways of accessing the data.

Example Tools:

- MongoDB: Stores data in a flexible JSON-like format that is more suitable for virtually any form of unstructured data.
- Couchbase: Provides an eventually consistent, geo-partitioned JSON document store database.
- Elasticsearch: Does offer strong indexing and searching for text data.

4.2.2 Blob Storage

- Definition: Storing large unstructured data files in binary large objects (BLOBs) into databases.
- Use Cases: Data storage and data retrieval of big files like images, videos, and documents.

Example Tools:

- Azure Blob Storage: For the large volume storage of data that is not easily categorized.
- AWS S3: For object storage, which needs high availability and durability.
- Google Cloud Storage: For large unstructured file storage with the added advantage of ease when retrieving the files.

4.3. Data Streaming and Real-time Loading

4.3.1 Stream Processing Frameworks

- Definition: For example, pulling out, preprocessing and loading data into real-time tools.
- Use Cases: Conveying analysis in real-time, data is ingested in real-time as well.

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Example Tools:

- Apache Kafka: For event-based and real-time data handling, especially in sequential data.
- Apache Flink: For processing current and streaming data and also for streaming statistical analysis.
- Spark Streaming: Real-time data processing is a term that helps define it.

4.3.2 Event-driven Architectures

- Definition: Deciding on the technique of using events to load data processes in a particular data model.
- Use Cases: Real-live data handling, no-server infrastructure.

Example Tools:

- AWS Lambda: To execute code in response to events but without prior provision of computers/servers.
- Google Cloud Functions: This is suitable for light weight functions that happen on events.

4.4. API-based Loading

4.4.1 Web APIs

- Definition: Pulling data from web services or RESTful APIs.
- Use Cases: Information integration of third-party data and real-time data acquisition capabilities.

Example Tools:

- Python's requests library: For making and performing requests to get data from the internet resources via HTTP.
- Third-party APIs: Working along with other services such as Twitter API, and Google Maps API.

4.4.2 GraphQL

- Definition: For loading data in a better and more efficient way, there is the utilization of GraphQL APIs.
- Use Cases: Filtering specific data structures, the minimization of the excessive extraction of data.

Example Tools:

- Apollo Client: Specifically, this article will apply to querying GraphQL APIs within JavaScript applications.
- Graphene: Used for constructing GraphQL APIs for your applications using the Python programming language.

4.5. Web Scraping

4.5.1 Scraping Tools and Libraries

- Definition: Transferral of web page data into a material form for utilization.
- Use Cases: Gathering data from the websites, competition analysis.

**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Example Tools:

- BeautifulSoup: For working with HTML and XML documents using a variety of operations in the Python programming language.
- Scrapy: This is an application that uses Python to scrape a website.
- Selenium: For automating web browsers for data extraction and for web scraping dynamic content.

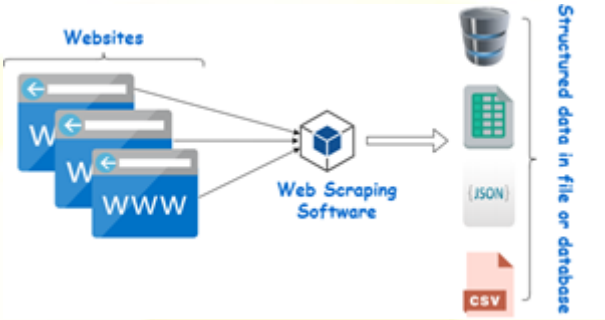


**Figure 7: Web Scraping**

4.6. Text Data Loading

4.6.1. NLP Tools

- Definition: This transformation involves loading and preparing the textual information regarding its analysis.
- Use Cases: Textraction text mining, opinion mining, and language modeling.

Example Tools:

- NLTK: For text processing and linguistic data analysis in the programming language Python.
- spaCy: Specifically, for more complex natural language processing problems such as named entity recognition and parse tree construction.
- Gensim: For topic modelling and text document similarity analysis.

4.6.2. Document Parsing

- Definition: Managing document data with different formats in extracting and loading.
- Use Cases: Creating softcopy from hardcopy, extraction of text from PDFs and Word documents.

Example Tools:

- PDFMiner, PyPDF2: To save text in a PDF file in an editable format, particularly from image-based PDFs.
- python-docx: As a result of using this technique, Word document readings and writings will be completed.

4.7. Image and Video Data Loading

4.7.1. Image Processing Libraries

- Definition: Reading in the images and initial data manipulations.
- Use Cases: Image recognition and computer vision applications.

Example Tools:

- OpenCV: In image processing and in computer vision operations where selection and recognition of objects in an image are conducted.
- Pillow: For simple and basic image processing in a Python environment.

**International Journal of Management, IT & Engineering**
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

4.7.2. Video Processing Libraries

- Definition: The loading of the video data and to some extent, the preprocessing of the video data.
- Use Cases: This includes video analysis and object detection in videos.

Example Tools:

- OpenCV: As for processing video data as well as for extracting frames from the video.
- FFmpeg: This is in consideration of video processing, conversion, and streaming.

4.8. Audio Data Loading

4.8.1 Audio Processing Libraries

- Definition: Loading and preprocessing audio data can also be represented in the given type of diagram.
- Use Cases: These are the speech and audio recognition and analysis.

Example Tools:

- librosa: For music and audio tools and libraries in Python.
- PyDub: It also contains simple audio and easily manipulated functions.

4.9. Data Integration Platforms

4.9.1 ETL Tools

- Definition: Information integration tools used in converting unstructured data into structured systems also known as ET limbering or ETL.
- Use Cases: Data conversion, data consolidation, acquiring data from sundry sources.

Example Tools:

- Apache NiFi: Used for passing data and mapping data from one application to another.
- Talend: Regarding the second component, namely data integration, data quality and big data solutions, are provided.
- Informatica: If one wants to keep a record of all his or her information needs and data requirements comprehensively correlated and systematized.

## 5. Results and Discussion

The use of ETL for unstructured data has shown positive outcomes in many fields of study. Popular AI methods that are useful in Text Mining include tokenization, named entity recognition, and sentiment analysis that has been applied to analyzing the customer's feedback and tracking the sentiments. For example, with the help of libraries such as spaCy and NLTK, the problems of text transformation have been solved well, and the extraction of keywords and topics from the given test data texts has become painless.

While in computer vision, services like OpenCV and TensorFlow have made the extraction of features from images possible, leading to object detection or even image classification. Borrow techniques like OCR have shown to be very helpful in transforming textual information within images to be more usable for operations, making image data more valuable.

International Journal of Management, IT & Engineering
Vol. 11 Issue 01, January 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Many audio and speech processing methods have proved to play a considerable role in translating spoken language into written form, which can be depicted by using Google Speech API and DeepSpeech. These tools have enhanced the performance of tasks such as transcription and identification of speakers and other related activities.

Two methods have been identified in the processing of video data, such as video summarization and action recognition which allow for the extraction of meaningful information from the videos. Software such as OpenCV and TensorFlow has been of immense help in these changes, especially in the analysis of streaming videos.

Some of these tools and libraries have come up in the past to help integrate these unstructured data transformation techniques into ETL processes. Apache Hadoop is extensively used for processing big amounts of unstructured data in batches, and Apache Spark, as an extension of it, performs the same function more efficiently; data ingestion and data transformation tools are Apache NiFi and Talend. Other databases like MongoDB and Elasticsearch are great for unstructured data storage since they are not SQL databases.

## 6. Case Studies

6.1. Finance

6.1.1. Unstructured Data Example: Social Media Sentiment Analysis

- Sources of unstructured data for finance can be social networks. As special, it can be stated that analyzing sentiment on social media can be helpful to gather information on the customer as well as on the market.
- ETL Process: People, for example, the process of extracting text data from social media platform SNS, passing this data through NLP to analyse sentiment for content that can be positive, negative or neutral after loading this data-to-data warehouse if necessary.

6.2. Healthcare

6.2.1 Unstructured Data Example: Customer Reviews and Feedback

- This involves collecting data from sources which are not formatted for easy analysis, for example, customer feedback on e-commerce sites and social networks.
- ETL Process: To extract this data, it entails extracting text data from these platforms and then preprocessing where comments feedback is categorized, the sentiment in the feedback determined, and irrelevant information removed. This transformed data is then used in analytics platforms for the analysis of customer satisfaction levels and the needed changes in the products.

**Table 8: Case Studies of ETL Implementations**

| Industry | Unstructured Data Example |
|---|---|
| Finance | Social media sentiment analysis |
| Healthcare | Clinical notes and medical images |
| Retail | Customer reviews and feedback |

## 7. Conclusion

This paper has also pointed the substantial developments, especially in the ETL processes of a cross-section of the different kinds of unstructured data. Various approaches, including NLP, machine learning and deep learning, have been useful in the conversion of texts, images, audio and especially videos into forms that can be subjected to analysis and drawing of insights. These techniques have been incorporated into ETL processes with the help of efficient tools and libraries to improve data analysis of unstructured data.

This is because the volume of unstructured data is still increasing, and hence, the future research and development of ETL processes will play a significant role in realizing the full potential of data. The directions for future work are to enhance the scalability and adequateness of such approaches and to investigate other fields and areas in which the ideas could be useful. Thus, by understanding the progress made in the application of ETL for unstructured data and its relevance to the contemporary concept of big data, organizations can open new opportunities for the acquisition of significant knowledge essential for the success of their activities.

## 8. References

[1] Bird, S., Klein, E., &Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

[2] Honnibal, M., &Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://sentometrics-research.com/publication/72/

[3] Lecun, Y., Bottou, L., Bengio, Y., &Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097-1105.

[5] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 82-97.

[6] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., & others. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. *International Conference on Machine Learning*, 173-182.

[7] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., &Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1725-1732.

[8] Simonyan, K., &Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27, 568-576.

[9] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., &Stoica, I. (2010). Spark: Cluster computing with working sets. *USENIX Conference on Hot topics in cloud computing*, 10(10-10), 95.

[10] https://medium.com/@helen.stewart_15710/deciphering-the-maze-of-etl-for-unstructured-data-97a81840ff51

[11] https://www.instill.tech/blog/unstructured-data-etl

[12] https://www.lonti.com/blog/etl-for-unstructured-data-navigating-the-complexity
[13] https://www.kadoa.com/blog/the-rise-of-unstructured-data-etl
[14] https://medium.com/@varshanayak24/overcoming-the-unstructured-data-challenge-a-guide-to-nlp-based-etl-tools-a9f6ee8ed510
[15] https://www.integrate.io/blog/processing-unstructured-data-101/