International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/30509246/IJETCSIT-V3I4P102
Eureka Vision Publication | Volume 3, Issue 4, 8-16, 2022

Original Article

ETL Techniques for Structured and Unstructured Data

Dr. Hana Kimura
Tokyo International University, Japan.

Abstract - ETL concepts are among the most crucial in data management since as organizations continue to grow more especially in the numbers and types of data required for operation. Therefore, the purpose of this paper is to comparatively consider ETL methodologies for both, structured and unstructured data with regards to the differences, problems, and suggestions. This is because the first kind of data in data mining which is structured data proves to be comprised in a tabular structure with patterns like rows and columns while the other kind of data known as unstructured data lacks such format. Particularly in the initiatives of the evaluation, there is emphasis on issues such as the tools required in the extraction of data from several sources, transformation of data so that it is quality and consistent and the methods used in loading it in the target system. Alongside with the cases and examples from the real live, commenting on the tendencies for evolution of ETL and perspectives for further advancements the paper attempts to describe the situation comprehensively. Managers, data analysts, engineers and IT specialists concerned with extending the use of various data and its function for business will find this research useful.

Keywords - ETL (Extract, Transform, Load), Structured Data, Unstructured Data, Data Integration, Data Transformation, Data Extraction, Data Loading.

1. Introduction

The uniqueness of data is beyond measure and with the advancement of technology, data is generated and multiplied at a very fast pace, and it has become very important to come up with good techniques of handling data. Thus, the ETL process plays a critical role in the successful handling of data from various sources to be integrated, transformed and made available for use by the analysts and decision-makers. This paper deals with methods of structuring data handling in ETL processes as well as analysing their importance, approaches and difficulties of realization**.**

1.1 Definition and Importance of ETL

ETL is an acronym that is commonly used in data processing, where E represents Extract, T stands for Transform and the last one, L stands for Load. It is a data management technique that entails pulling data from sources, cleaning it, and then fixing it into a database or data warehouse that can be useful to a firm's operations. This is an important process of ensuring uniformity from one system to the other and a key process in data warehousing and business intelligence.

1.2 Structured vs. Unstructured Data

This kind of data is well categorized and is often in a format that can readily be searched, normally coming from databases or spread sheets complete with respective columns. While structured data has a neat format that can be put into a table and sorted through, unstructured data does not includes text documents, images, videos and others. These different data types pose issues and concern specific ETL approaches that need to be applied.

1. **Structured Data**: Refers to data that is structured in a way that is predefined often in the context of databases in rows and columns. These include SQL databases; excel spreadsheet databases, and CSV database.
2. **Unstructured Data**: A term used to depict information that does not have a fixed format, hence proving difficult to analyze using old-school instruments. Some of them are emails, posts on social networks, video and audio messages.

Table 1. Comparison between Structured and Unstructured Data

Aspect	Structured Data	Unstructured Data
Format	Rows and columns	Text, images, videos, etc.
Storage	Databases	Data lakes, NoSQL databases
Accessibility	Easily searchable and analyzable	Requires advanced processing techniques
Examples	SQL databases, spread sheets	Emails, social media posts, multimedia files

1.2.1 Challenges in ETL for Structured Data

Some of the problems that arise when working with structured data include data quality problems, problems of grouping data from different data sources and problems of ensuring that the set data is consistent. Such challenges are usually solved by applying the data profiling, cleansing, and normalization techniques.

1.2.2 Challenges in ETL for Unstructured Data

However, unstructured data available in various formats has entered the big data environment as the complete opposite from the structured data type. Texts mining some of the techniques that are applied on unstructured data include the text mining Natural language processing and Image recognition. Also, the management of unstructured data and their indexing need specific solutions like NoSQL DBs and data lakes.

1.3 ETL Process Overview

Figure 1. ETL steps

1. **Extraction**: The first of them is the extraction of data from the identified primary sources. For structured data, this may refer to selecting data through querying of data bases or from files. In the case of unstructured data, the extraction can be relatively harder and might include process like web scraping API, NLP techniques.
2. **Transformation**: When data is extracted, then it has to be prepose and prepared in a manner that the specific analytics tool can understand. This entails data cleaning, data normalization as well as data enrichment. Regarding structured data, it can involve type conversion and aggregation among others. Transformation in dealing with unstructured data may include processes like text parsing, sentiment analysis, development of features and more.
3. **Loading**: The last operation is the transferring of the transformed data into a destination system like a data warehouse or database or data lake. This step must guarantee the data quality and coherence, sometimes including ways of managing exceptions as well as improving the efficiency.

1.3.1 Importance of ETL Processes

Essentially, ETL is a core component of data warehousing and business intelligence. The involves the process of obtaining data from diverse sources, cleaning the data and then moving it to the desired location often a data warehouse of a data lake. ETL basically refers to the process of extracting data from different sources, transforming this data to fit the required format, and then loading this data into the required locations, and solid ETL processes help to provide the correct data at the right time to aid analysis.

2. Literature Survey

ETL processes for structured and unstructured data are widely discussed in literature across multiple disciplines such as database administration, big data, and machine learning. This section provides an overview of prior studies, and it divides them into focus areas the researchers examined.

2.1 Structured Data ETL

ETL processes for structured data have long been researched and mainly dealt with DBMS, though more recent ideas apply to other environments as well. Some of the works that belong to this tradition include

i Kimball and Ross (2013): Their book is “The Data Warehouse Toolkit” which will give an idea of method of creation of data warehouses. It has rich methodological guidelines regarding ETL activities, paying attention to data quality and the speed of the process.

ii Inmon (2005): The directive works by Inmon is often called a father of the data warehouse; however, in fact, the work is dedicated to the architecture and design of the data warehouses. The specifics of method that Stolze applies to the ETL process demonstrate how critical the data integration is, as well as how metadata can be beneficial for data consistency.

2.2 Unstructured Data ETL

This is especially applied in the rising field of big data to consider the ETL processes of unstructured data. Key contributions include

i Bello-Orgaz et al. (2016): Also in their survey, they talk about big data technologies and uses, including the issues concerning the utilisation of unstructured data. They discuss different methods of harvesting, cleansing and processing big data, text data and social media data, through natural language processing and other related artificial learning methods.

ii Gandomi and Haider (2015): Their paper on big data analytics has given an understanding in to the approaches to analyzing and handling of large volumes of and format data. They talk about the Hadoop and MapReduce for dealing with volumes of unstructured data and the problem of ETL scalability.

2.3 Hybrid Approaches

Many authors have attempted to describe methods of combining elements of both structured and unstructured data.

i Chen et al. (2014): They integrate three different hybrid ETL models that consist of several approaches to ETL and include big data processing. This strategy uses the features of relational databases and distributed computing paradigms to process different data forms.

ii Sakr et al. (2011): Their work is concerned with one aspect of cloud computing, ETL solutions; the trend shown here is the notion of elasticity and scalability to structure data as well as unstructured.

2.4 ETL Process

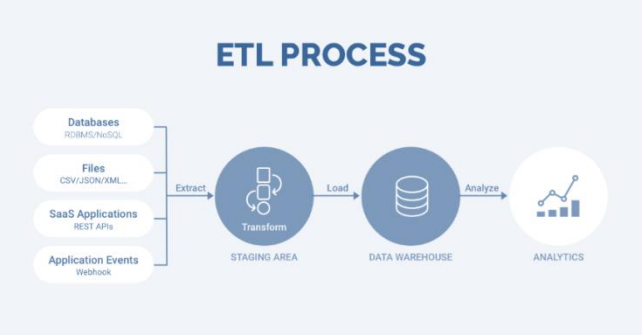

Figure 2. ETL Process

1. Extract:

i Databases (RDBMS/NoSQL): Data is scraped from both Relational database management systems (RDBMS) such as MySQL, PostgreSQL and NoSQL such as MongoDB, Cassandra.

ii Files (CSV/JSON/XML): It is also obtained from various file types including CSV files, JSON files, and XML files.

iii SaaS Applications (REST APIs): SaaS applications deliver data using REST APIs and here we need to download the data of users. Such applications could be Salesforce, Google Analytics or any other Software as a Service application.

iv Application Events (Webhook): It collects one or multiple events because of an application, often in the form of a webhook. Such activities may include user activities, system events and the likes.

2. Transform:

Staging Area: Data extracted is then relocated to a more convenient area usually referred to as a staging area where the data is customized. Data transformation works comprise data cleansing, data filtering, data transformation where the data must be adapted to be in a format that allows it to be used for analysis. This might include:

- **Cleaning**: Dealing with the uniqueness of the values, deletion of erroneous data points and processing the cases of missing data-points.
- **Filtering**: Choosing only the variables that must be analyzed during the examination of the research topic.
- **Modifying**: Converting the data, combining the existing fields, and making new Fields with own algorithms.

3. Load:

Data Warehouse: The raw form of data is converted and then it is stored in a data warehouse. Data warehouse is distinct data repository that contains the consolidated data from the different sources. However, it is mainly designed for query and analysis and not for day-to-day business transactions. Some of the familiar options for data warehouses are Amazon Redshift, Google Big Query and Snow Machine.

4. Analyze:

Analytics: Once data is updated in the data warehouse, then it is in a form that can be used for analysis. This can involve:

i **Business Intelligence (BI)**: Creating and designing dashboards and reports using BI tool solutions such as Tableau, Power BI, Looker, etc.

ii **Data Mining**: Estimation of relationships between the variables and interpretation of the data using statistics and machine learning.

iii **Reporting**: Copy the developing of periodical reports for business members.

3. Methodology

ETL Process from Data Extraction to Data Loading is mentioned and with figure 3 are mentioned below.

1. **Data Extraction**: This initial step is concerned with collecting the crude information from disparate sources. The goal is to guarantee that all information required for the next stage of the analysis is recorded and stored.

2. **Data Transformation**: The extracted data is usually unprocessed and possesses large variations. Transformation includes various operations such as

i **Data Cleaning**: In this case, the ability to eliminate errors and inconsistencies from the data.
ii **Data Integration**: Accumulating structured data within an integrated dataset of an organization.
iii **Data Conversion**: Fluctuating data types or structures whenever necessary.
iv **Data Aggregation**: After collecting data, it is summarized to get a general view of the research problem and to make conclusions about it.

3. **Data Loading**: This last process is where the transformed data is then transferred to the target system (data warehouse). The data which has been loaded into the target tables is now available for querying and analysing of complex information.

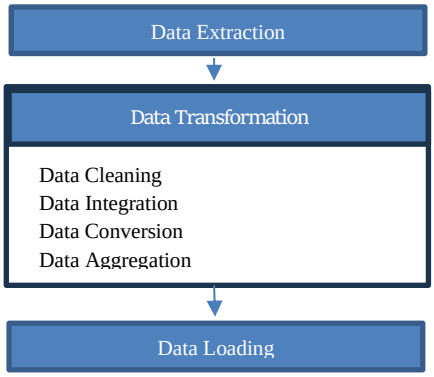

Figure 3. ETL Process: From Data Extraction to Data Loading

3.1 Extraction Techniques

The extraction stage focuses on the gathering of data from different source. In the case of structure data, it may mean issuing relational databases in an instance using the structured query language. For unstructured data, extraction methods include, web data extraction, API conversation and file and log data intake. A typical data extraction process is depicted in the following Figure 4.

Figure 4. Data Extraction Process

1. Start

This is the start of the data extraction process where all the data that is relevant is gathered and extracted from various sources. It is occasionally called as the starting part where the process is begun, though it may be started manually by an operator or may be started by the running of a scheduled job or by a trigger event.

2. Identify Data Source

As a result, the specific data sources, which must be extracted, are defined here. These sources could be databases, Web services, files, APIs or any other data resources. It is therefore pivotal to correctly identify the data source to improve on the probability of extraction.

i **Structured Data Sources**: Some of the data structures that are included here are the database structures which include the SQL and NoSQL databases, spreadsheet data structures as well as any structures that are in the form of a table.
ii **Unstructured Data Sources**: This covers text data such as word documents, emails, Tweets, Facebook posts, Instagram posts, logos, movies, songs and any other data format that is not a table.

3. Extract Data

After that, the next activity is to pull out the data from these sources identified with the system Main data sources identified for this system includes: The method of extraction varies depending on the type of data source:

i **Structured Data Extraction**: This mainly involves executing SQL statements to databases to get the required data.
ii **Unstructured Data Extraction**: It can be done, for example, through web scraping opening links to web pages, parsing text, using APIs or tools such as BeautifulSoup, Scrapy or any other custom scripts.

4. Validate Data

The next step involves the verification of the data that has been extracted to confirm on the quality of the data extracted inclusive of accuracy, completeness, and consistency. This involves several sub-steps:

i **Data Profiling**: To apply data digestion techniques this includes understanding the structure, content and quality of the data.
ii **Data Cleaning**: Deleting redundancy, data cleaning performing operations, dealing with missing data, and checking whether the data fits into some pre-set format or not.
iii **Schema Validation**: Verifying that the extracted data is in a format or has a schema that was anticipated.

5. Store Extracted Data

After validation, the extracted data is saved into Intermediate or target data store. This can be a staging database, a data lake, or a data warehouse based on the overall ETL topology.

i **Staging Area**: The staging of data before the transformation and loading activities take place in the destination.
ii **Data Warehouse**: Centralized database for storing structured data optimized for query and report generation.
iii **Data Lake**: Large collection of data in its raw form or native data format where the unstructured and semi structured data is stored.

6. End

This is where the process of data extraction is finished. Here, the data extraction, data validation, and storing can be said to be complete, and proceeding to the next step, which is the data transformation step in ETL process.

3.2 Data Transformation Process

Figure 4. Data Transformation Process

1. Raw Data Input:

Description: It is important to note that raw data is gathered from a variety of sources these include; databases, flat files, web services and other sources.
Purpose: Gathering raw data which require some form of conversion to be in a suitable form for analysis.

2. Data Cleansing:

Description: This sub-process concern involves the detection of errors or inconsistency of data and correction is made to these errors.
Tasks:

i **Removing duplicates**: Validating that each record is unique, and free from duplication.
ii **Correcting errors**: Correcting all the typographical errors, wrong figures, or any other irregularities.
iii **Handling missing values**: Imputing of the missing data by applying proper means of imputations, such as Mean imputations or applying some kind of predictive models.

3. Data Normalization:

Description: Grouping of data so that one piece of data does not depend on other piece of data in a file.
Tasks:

i **Schema alignment**: Organizing information in the form of tabular form or columns and rows.
ii **Standardization**: Referencing (formatting of date formats, currency etc)
iii **Scaling**: Standardizing numerical variables to bring the variables into the same range of values in preparation for analysis.

4. Data Aggregation:
Description: Taking specific information and condensing it which affords an overview.
Tasks:

i **Summation**: Also, you are recalling the sums by deriving totalities from numerical values.
ii **Averaging**: Calculating mean values.
iii **Grouping**: The process of grouping data into categories or groups depending on the set criteria.

5. Data Enrichment:
Description: By the process of including other details into data that has been collected, the quality and value of the data is improved.
Tasks:

i **Adding metadata**: Adding qualifiers with the data to have more detail.
ii **Data matching and merging**: Forming a dataset from several data sources together to create a great data set.
iii **Incorporating external data**: Categorizing of the additional data to improve the given dataset.

6. Transformed Data Output:
Description: The final output of the transformation process that means after data transformation activity it is in the form of clean, enriched data and arranged for loading into the target system.
Purpose: Preparation of quality data to be used in analysis, reporting and decision-making.

3.3 Loading Techniques

The loading stage requires the writing of the transformed data into the target system. With structured data, this likely means that it is loaded into a data warehouse or a database. For unstructured data, loading might mean, placing the data in a data lake or NoSQL database. Figure 5 depict one data loading process, though, which is common in data integration practice.

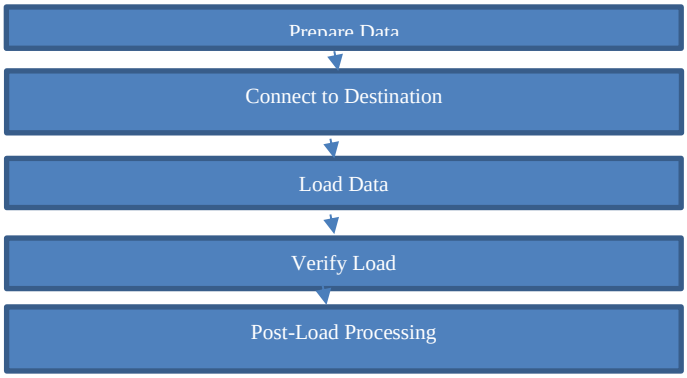

Figure 5. Data Loading Process

1. **Prepare Data**: This step requires prepping of the data; this simply means that one needs to make sure that the data is ready to be loaded. This encompasses the process of translating the data to fit into the specified and required format of the destination system and it equally checks if there are any form of preprocessing required on the data such as cleaning or normalization.
2. **Connect to Destination**: Specify a link to the output database or data warehouse. This includes identifying the database connection details, the security authentication, and security authorization to enable the data to be loaded.
3. **Load Data**: Move the data into the target system from the consolidation area or the original system. This step may include loading in large totals, increments, or while the stream is running based on the destination system utilized.
4. **Verify Load**: Never have we to check if the load operation has been done successfully, once the data have been loaded into the new servers. This may involve checking row counts, data integrity and as simple as ascertaining that there were no errors in the load process or data was not lost.
5. **Post-Load Processing**: What actions are taken after the data has been loaded if the actions are an option. This may involve indexing so that queries on the structures run faster, sub partitioning for better manageability, or altering metadata and statistics in the target database.

3.4 Tools and Technologies

3.4.1 Traditional ETL Tools

Most commercial ETL tools that are currently on the markets including Informatica, Talend, and Microsoft SSIS are complete ETL solutions that enhance processes such as profiling, transformation, and loading. They are commonly used in the scenarios of the classic data warehousing architecture.

i **Informatica**: It is famous for the effective data transformation qualities.
ii **Talend**: Fully open source with an impressively extendable set of connectors.
iii **Microsoft SSIS**: Support of MS SQL Server environment.

3.4.2 Big Data ETL Tools

Apache Nifi, Apache Spark, and Hadoop are large data ETL tools that can also handle semi and unstructured data. These tools have support for distributed processing and thus are beneficial in the realm of big data.

i **Apache Nifi**: Real-time data integration and the flow of data from one application to another.
ii **Apache Spark**: Data processing and machine learning at rates near real time.
iii **Hadoop**: Architecture for distributed computing model of storage and computation.

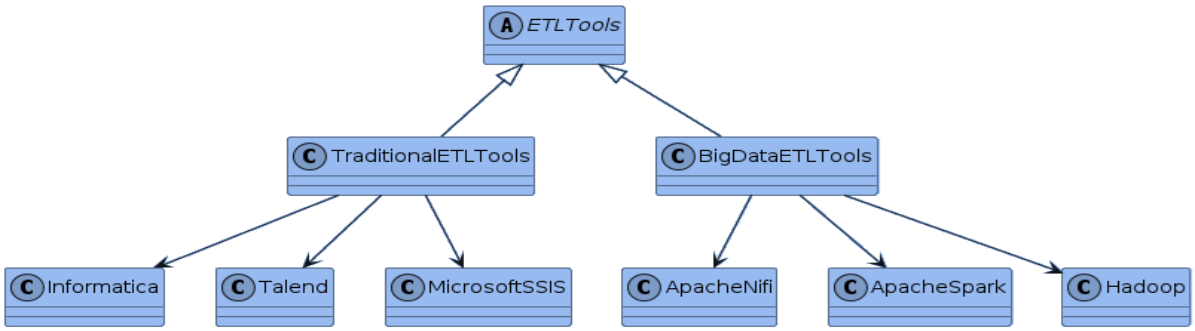

Figure 6. ETL Tools for Structured and Unstructured Data

1. Informatica
Description: Informatica is an ETL tool that is widely accepted and implemented in large organizations to handle and transform data.
Features:
i Supports communication to numerous data sources that could include the database, cloud applications, and flat files.
ii Enhances more sophisticated data profiling, cleaning and data transformation features.
iii Contains features to manage ETL processes and their schedules.
Use Cases:
i B2B audiences that are large enterprises in need of optimal data integration capabilities.
ii Connection to the data mart and business intelligence software.

2. Talend
Description: Talend is an open source ETL tool that contains all the features and functionalities for integrating and managing data.
Features:
i It covers a variety of connectors and ingredients to accomplish the processes of data extraction, transformation, and loading.
ii Offers a user interface for defining ETL jobs and their logistics.
iii Specifications include Data Enrichment, Data Confirmation, Data Cleansing, Data Suppression, Data Encryption, Data Profiling, Data Auditing, and Reporting.
Use Cases:
i Small and medium-sized enterprises that wish to integrate data without having to spending a huge sum of money.
ii Cloud platform and Big Data compatibility.

3. Microsoft SSIS (SQL Server Integration Services)
Description: SSIS stands for SQL Server Integration Service that is a part of the Microsoft SQL Server System and is widely used as an ETL tool.
Features:
i Chemical integration with Microsoft sql server and other Microsoft related products.
ii It offers data profiling, and cleaning or transformation of data as per the requirements.
Use Cases:
i Companies or institutions that depend on Microsoft SQL Server as their main means of data management.
ii Microsoft Azure Integration for cloud-based data integration.
iii Big Data ETL Tools

4. Apache Nifi
Description: Apache Nifi is a real time system used in integration and distribution of data flow and help in the automation of the flow.
Features:
i Supports data acquisition from different sources such as the IoT devices, databases, and APIs.
ii Delivers a browser-accessible GUI for information design and tracking of the data streams.

iii Provides facilities for routing, transforming and enriching customer's data.

Use Cases:

i Firms and institutions that require live data streaming and incorporation into the large volumes of data.

ii Interoperability with Apache Kafka and other streaming technologies.

5. Apache Spark

Description: Apache Spark is, in fact, an analytics framework designed for the massive scale data processing, including the ETL operations.

Features:

i Offers high speed data processing using in-memory computing.

ii Supports both, the processing of large sets of data at once, processing data streams in real time, machine learning, and graph processing.

iii APIs in Java Scala Python or R for ETL pipelines.

Use Cases:

i Business communication and information management; statistics and data analysis; information science.

ii Sorting big amounts of information quickly in shared environments.

6. Hadoop

Description: Apache Hadoop is used to store large datasets on multiple components of computers known as clusters.

Features:

i Hadoop has Hadoop Distributed File System (HDFS) for distributed Storage.

ii These are components of Hadoop, and some are MapReduce, which is the programming model of hadoop that enables parallel data processing.

iii Some of the tools that it supports are hive which supports data processing, queries and analysis, pig which is a data flow language and HBase which supports big data storage.

Use Cases:

i Scalability in the processing of large volumes of disparate and semi-structured data and being able to do so in a fault-tolerant environment.

ii Processing and analysis of massive data in various business and research organizations.

4. Results and Discussion

4.1 Comparative Analysis of ETL Techniques

Although the ETL of structured data differs significantly from that of unstructured data, there is also a degree of contrast. Unlike structured data ETL's main concern is on mapping of schemata and data standardization, unstructured data ETL has uses approach like text mining and NLP for data extraction as well as transformation.

4.2 Best Practices in ETL

Typically, it is possible to consider such measures as profiling and cleansing of the data to be loaded into the data warehouse, using appropriate tools and technologies, and the availability of proper monitoring and error-handling procedures. Furthermore, it is relevant to select an appropriate ETL tool depending on many requirements and peculiarities of the data of an organization.

4.3 Future Trends in ETL

Possible developments of ETL in the future are as follows: growing utilization of AI and machine learning to work with data extraction and transformation, integration of ETL with data warehouses based on the cloud computing concept and more improved instruments for structuring unstructured information.

Table 2. Future Trends in ETL

Trend	Description
AI and ML Integration	Using AI and ML for automated data processing
Cloud-based ETL	ETL processes integrated with cloud data storage
Advanced Unstructured Data Handling	New tools for unstructured data ETL

5. Case Studies

1. Finance

Structured Data Example: Customer Transaction Records

i In the finance industry structured data refers to much more formalized data sets that can easily be indexed. Customer transaction records are one of the examples of such data sources. They stand as electronic versions of past documents and often contain elements such as: transaction amounts, dates and numbers, accounts, and many others all in forms of tables and databases.

ii ETL Process: This data becomes extracted from the transaction systems, transformed to effectively standardize and conform to the correct format often by conversion of currency, date, etc and then is loaded into the data warehouses for analysis and reporting.

Unstructured Data Example: Social Media Sentiment Analysis

i 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.

ii 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 positively, negatively or neutral, after that loading this data-to-data warehouse if necessary.

2. Healthcare

Structured Data Example: Sales Data

i Structured data in the context of retail entails sales data, which must provide details concerning the product code, quantity sold, cost, and time of sale. This data is most kept in POS systems, also known as spoke databases.

ii ETL Process: In the ETL process, point of sale or POS systems extract sales data, transform the same data to eliminate any disparities, such as combining data from different outlets, and load the data into data warehouses which are useful in sales analysis and inventory.

Unstructured Data Example: Customer Reviews and Feedback

i This involves collecting data from sources which are not formatted for easy analysis, for example, the customer feedback on e-commerce sites and social networks.

ii **ETL Process**: To extract this data, it entails extracting text data from these platforms then pre-processing 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 the customer satisfaction levels and the needed changes in the products.

Table 3. Case Studies of ETL Implementations

Industry	Structured Data Example	Unstructured Data Example
Finance	Customer transaction records	Social media sentiment analysis
Healthcare	Patient demographics	Clinical notes and medical images
Retail	Sales data	Customer reviews and feedback

6. Conclusion

As such, ETL holds enormous value as part of the basic data management strategy since the amount of data today has exponentially increased and the number of different types of data sources continues to grow. It is therefore only imperative that by understanding the various issues relating to structured and unstructured data one can effectively be able to formulate strategies that relate to ETL that are elastic to the need of the organization. New trends, technologies, and strategies such as AI and cloud integration also pose to improve the effectiveness and efficiency of ETL tools soon.

References

[1] Ralph Kimball, and Margy Ross, The Data Warehouse Toolkit the Definitive Guide to Dimensional Modeling, Wiley, pp. 1-608, 2013. https://www.google.co.in/books/edition/The_Data_Warehouse_Toolkit/4rFXzk8wAB8C?hl=en&gbpv=0

[2] W. H. Inmon, Building the Data Warehouse, John Wiley & Sons, pp. 576, 2005. https://books.google.co.in/books/about/Building_the_Data_Warehouse.html?id=QFKTmh5IFS4C&redir_esc=y

[3] Bello-Orgaz G, Jung JJ, Camacho D. Social big data: Recent achievements and new challenges. Inf Fusion. 2016. doi: 10.1016/j.inffus.2015.08.005.

[4] Amir Gandomi, Murtaza Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 137-144, 2015. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[5] What Is ETL Process, Medium. https://medium.com/@datadrix/what-is-etl-process-in-data-science-4249745453bd