Original Research Paper Engineering

Converting Unstructured Data to Semi-Structured Data

Arup Kumar Das Student, MBA IT, Symbiosis Institute of Compute Studies and Research, Atur Centre, Model Colony, Pune-16

Dr. Pravin Metkewar Head of Department, MBA IT, Symbiosis Institute of Compute Studies and Research, Atur Centre, Model Colony, Pune-16

ABSTRACT

Unstructured data is present in huge quantity around us and processing it is very difficult. Also cost is increasing in processing information with current techniques—in turn using this huge information is complexed now. Hence there is a need to find an adjustable solution in understanding the data better and to processing the data automatically, and to remove disadvantages generated by schemas, in terms of lack of evolution & inflexibility.

Enabling unstructured information and processing it to semi-structured data is a major challenge

KEYWORDS : Semi-Structured Data, Structured Data, Conversion, ETL, Data Warehouse

1.INTRODUCTION

1.1 Understanding unstructured data & Semi-Structured Data

The expression unstructured information more often than not alludes to data that doesn't live in a conventional line segment database. As you may expect, it's the inverse of organized information — the information put away in fields in a database.

Samples of Unstructured Data:

Unstructured information records frequently incorporate content and interactive media content. Illustrations incorporate email messages, word preparing records, recordings, photographs, sound documents, presentations, pages and numerous different sorts of business archives. Note that while these sorts of records might have an inward structure, they are still viewed as "unstructured" on the grounds that the information they contain doesn't fit perfectly in a database.

Semi-structured information will be data that doesn't live in a social database however that has some authoritative properties that make it simpler to dissect. Illustrations of semi-organized information may incorporate XML reports and NoSQL databases.

The term enormous information is nearly connected with unstructured information. Enormous information alludes to a great degree extensive datasets that are hard to dissect with customary devices. Enormous information can incorporate both organized and unstructured information, however IDC gauges that 90 percent of huge information is unstructured information. A significant number of the devices intended to investigate huge information can deal with unstructured information.

1.2 Issues in mining and extracting data

Numerous associations trust that their unstructured information stores incorporate data that could offer them some assistance with making better business choices. Sadly, it's frequently extremely hard to break down unstructured information. To help with the issue, associations have swung to various diverse programming arrangements intended to look unstructured information and concentrate critical data. The essential advantage of these devices is the capacity to gather noteworthy data that can offer a business some assistance with succeeding in an aggressive domain.

Since the volume of unstructured information is developing so quickly, numerous ventures additionally swing to mechanical answers for offer them some assistance with bettering oversee and store their unstructured information. These can incorporate equipment or programming arrangements that empower them to make the most productive utilization of their accessible storage room.

2. Problem Statement

This exploration paper distinguishes the principle explanations behind issues, for example, changing unstructured information to organized information in information warehousing. Its center is to distinguish every one of the issues identified with the information change and to dispense with them for an effective and better utilization of information stockroom.

3. Literature Review

Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume.

Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume.

Non-relational databases, such as Not Only SQL (NoSQL), were developed for storing and managing unstructured, or non-relational, data. NoSQL databases aim for massive scaling, data model flexibility, and simplified application development and deployment. Contrary to relational databases, NoSQL databases separate data man-agement and data storage. Such databases rather focus on the high-performance scal-able data storage, and allow data management tasks to be written in the application layer instead of having it written in databases specific languages.

P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, 2014. © Springer International Publishing Switzerland 2014

In the recent years, XML†(eXtensible Markup Language) has reached a wide acceptance as the relevant standardization for representing semi-structured documents. XML

documents present the advantage to have an explicit structure that facilitates their presentation and their exploitation in different contexts.They are becoming more

common in various environments, permitting to represent jointly the textual information with the structure one.

A semi-structured document is a bridge between structured and unstructured data. Unstructured data (also called flat data) is data that we know neither the context

nor the way information is fixed. It includes documents of mostly natural-language text, like word-processing files, e-mail, and text fields from databases or applications.

Amina MADANIa*, Omar BOUSSAIDb, Djamel Eddine Zegour
For those thinking about Big Data - 80% of the 7 exabytes of data stored last year was unstructured meaning that using traditional analytics tools will even with the cloud's elasticity can handle

Tapping into Unstructured Data by William H.
4. Proposed concept & technique
Assortment of various programming apparatuses will be helpful to sort out and oversee unstructured information.

These can incorporate the accompanying:
1.Big data tools
Programming like Hadoop can prepare stores of both unstructured and organized information that are amazingly vast, extremely unpredictable and evolving quickly.

2. Business intelligence software
Otherwise called BI, business insight is a general class of examination, information mining, dashboards and reporting apparatuses that offer organizations some assistance with making feeling of their organized and unstructured information with the end goal of settling on better business choices.

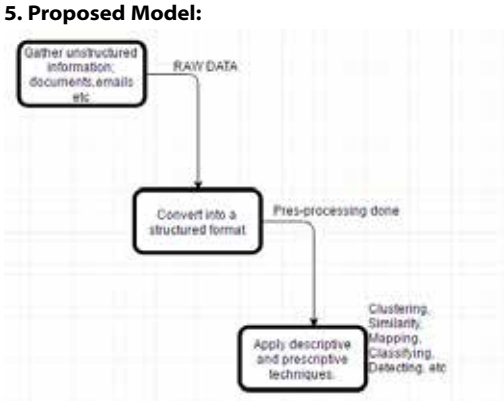
3.Data integration tools
These apparatuses consolidate information from unique sources with the goal that they can be seen or broke down from a solitary application. They once in a while incorporate the capacity to bring together organized and unstructured information.

4. Document management systems
Additionally, called venture content administration frameworks, a DMS can track, store and share unstructured information that is spared as report records.

5. Information management solutions
This kind of programming tracks organized and unstructured undertaking information all through its lifecycle.

6. Search and indexing tools
These instruments recover data from unstructured information records, for example, archives, Web pages and photographs.

5. Proposed Model:



6. Results Expected:
Semi-Structured data from a complete unstructured data.

This tedious task will provide a proper conversion of data from one format to another.

7. Conclusion:
It has been observed that conversion from unstructured data to semi-structured data is not a direct approach, however it can be made possible through various tools and methods efficiently so that data can be used in a more reliable manner.

REFERENCES

1. **Tapping into Unstructured Data by** William H. Inmon (Author)
2. SEMI-STRUCTURED DATA by MOHAMED ELTABAKH
3. P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, 2014. © Springer International Publishing Switzerland 2014
4. Amina MADANIa*, Omar BOUSSAIDb, Djamel Eddine Zegour