International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2538-2544 www.ijaem.net ISSN: 2395-5252

Structured and Unstructured Data in Data Warehouse: An Extensive Survey

LS Abinash Nayak , Kaberi Das, Debahuti Mishra, Srilekha Hota,
Institute of Technical Education and Research, Odisha
Institute of Technical Education and Research, Odisha
Institute of Technical Education and Research, Odisha
Cognizant Technology Solutions

Submitted: 10-06-2021 Revised: 20-06-2021 Accepted: 23-06-2021

ABSTRACT. IT industries or organizations are continuously growing day by day where Data Warehousing Technology plays an important role. A Data Warehouse has the characteristics such as subject-oriented, integrated, time-variant, and nonvolatile which supports the management to take necessary decisions. Data warehousing can be defined as a data hub or repository, where a huge and large amount of meaningful data are stored. It gives a clear visualization for analysis of the huge data and helps to take the decisions for business growth. Data warehouse stores historical information and makes it simple for business analysts to analyze data. This technology enables the analysts to extract and view business data from different sources. Here a discussion has been made about the challenges faced during collection of the data from different sources, and making a survey about those data and finally integrating them to take smart decisions. While extracting the data from the sources, the data is unstructured, less potential and contains error. This paper gives an overall review of the research work done by the authors regarding how to handle the structured and un-structured data sources inside a Data Warehouse.

Keywords: Data warehousing; OLAP system; Meta Data; Time-Variant.

I. INTRODUCTION

Data Warehouse (DWH) satisfies the properties such as subject oriented, integrated, time variant and non-volatile [1]. Subject oriented means it allows the user s to directly access the subject database. DWH collects data from different sources but the data available in DWH is in the form of integrated form. The time variant property explains the data that are available in the DWH is extensive in nature means maintained historically. Non-volatile property explains that we are able to access the new data along with the existed data in a DWH. It gives an effective visualization to the management, CEO, Business head and the analysts to take smart and effective decision for the organization. DWH is nothing but an environment which is designed to be used by the experts for writing the query to get the desired results which helps to analyze the data in different perceptions [2]. In a Data Warehouse the data is obtained in the form of structured and un-structured data format. Structured data is considered as the well-organized data which maintains proper documentation in order to inspect easily. In case of unstructured data it's size is huge, available in different types of configurations such as e-mails, memos, thesis, web content, excel spread sheet and in the forms of images etc. It is difficult to write query for analysis. Within the Data warehouse the structured data is treated as well formatted data structure. This data can be made to measure the importance of the available data in a complex graphical presentation which includes facts & data-set table. Facts are the parameters that are studied from various point of views to take necessary decisions for the businesses and Dimensions are the background for the facts to describe. But unstructured data is used intelligently for many important purposes which is related to the decisions taken by the analysts for the business growth. As the size of the online analytical processing (OLAP) server is very huge it is used for data warehousing purpose for storing historical data for a long term period . It helps the analysts to extract the structured data from the server to handle the challenges in the real time environment. As the amount of data is growing rapidly in the real world in a scattered form, we need to handle those data by rearranging the data in Data Warehouse, as a result of which we can retrieve efficient data from the unknown source of data. Many researchers and authors have used the concepts such as integrating different types of data and many authors have extracted important information from the unorganized raw data.

The remainder of this paper is as follows Section II, describes about the procedures to handle the unstructured data in the data warehouse. Section III, gives a tabular representation of overall ideas about correlative surveys. Then we have discussed about the conclusion and future work in section IV.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2538-2544 www.ijaem.net ISSN: 2395-5252

II. STATE OF THE ART

Data Warehousing models are of 3-types:
a) Enterprise Data Warehousing
b) Operational Data Store
c) Data Mart

a) Enterprise Data Warehousing:

It is treated as the consolidated warehouse which aims to provide decision support services (DSS) over the enterprises. It is a consolidated proposal for the construction and representation of data.

b) Operational Data Store:

In this model the real time data is refreshed. So, it is preferable mostly for activities like storing detailed record of Student.

c) Data Mart:

This model is used by various enterprises/organizations simultaneously for a certain business areas such as ERP, HR, Finance, HR or Finance. The unique data mart directly takes data from the origins. This model may be ruled by one or more organizations.

Different approaches or procedures have been adopted to handle the source or fuzzy data available in different forms in various locations. According to B. Rieger et al., [3] it has been introduced that a meta database by integrating quantitative and qualitative information by using associate management functionalities. The authors M. Z. Bleyberg et al., proposed [4] to build text warehouse by using the concept of dynamic multi-dimensional model. The index table of a snowflake schema is the origin of any model. The information are extracted from the warehouse with respect to their syntactic groups which is committed to those data. Maria Zamfir et al., [5] have discussed about the multidimensional warehouse structure where queries are written to get the output based on the documents from a Data Warehouse.

A document based Data Warehouse system is object oriented multimedia-database system where the data related to the multimedia are analyzed for the reusability purpose. The authors Hiroshi Ishikawa et al., [6].According to J. M. Perez et al., [7] a data warehouse is developed in such a way that it integrates all the traditional data warehouse in a consolidated database which is represented in forms of cubes. Authors Tseng et al., [8] have given a description about the three procedures of integrating the structured and unstructured data inside a data ware house to make an efficient business system. Data warehousing technique enhances the quality of decision making by integrating data from various source. An organization maintains its data sources in a form of functional or departmental databases that have developed to provide services to localize the requirements.

F. Ravat et al., [9] have focused on the OLAP concept for aggregation function, that is for the textual data warehouse. OLAP technology makes it easy for the user to organizing the multi-dimensional data. The analytical calculation of an OLAP system is very fast. After writing a query the analysts immediately get the response from the OLAP system that is nothing but speedup of thoughts analysis. The authors Cindy Xide Lin et al., [10] have tried to propose a text cube model concept of OLAP to analyze textual data in terms of dimension hierarchy. Henning Baars & Hans-George Kemper [11] have focused on the different methods of integrating the structured and unstructured data for the analysis purpose by the business analysts. The efficiency and challenges are discussed .The authors D. Zhang et al., tried to present [12] a model of data that brings together OLAP and topic models to expand the OLAP to help the analysts or experts to take crucial decision in a multi-dimensional data base. According to Kalli Srinivasa Nageswara Prasad and Prof. S. Ramakrishna [13] the text-data-analysis frame work is used to handle the structured and unstructured data in the Data Warehouse for future analysis.

The authors M.Thenmozhi and K.Vivekanandan [14] have discussed about creating a multidimensional schema to derive the element automatically. According to Ahmad Abdullah Alqarni et al., [15] the structured-data and unstructured data are depicted in a multi-layer schema. To integrate various types of unstructured-documents, the mechanism of Linguistic match and word-net are used to identify the similarity among the data between both of the data sources. According to the authors Y. Sharma et al., Askand [16] a service oriented architecture is provided for the integration of data sources from different areas which has the ability to develop composite applications. G. Kassem et al., [17] authors have proposed about the conceptual generic data model for matching the data models in a standardize business process. Authors J. Sreemathy et al., [18] have used ETL tools like Talend to integrate data from different data sources. N. El Moukhi et al., [19] have discussed about the state of art and the future challenges of data warehousing technologies. The authors G. Garani et al., [20] have tried to give an idea of integrating star and snowflake schemas concept in an effective data warehouse. According to the authors P. S. Diouf et al., [21] it has been discussed how verities of data are extracted from the various sources and integrated in ETL model process by using some ETL tools. The authors S. Mandal et al., [22] have proposed how the raw data is extracted from the different operational data bases and integrated to develop a central Data Ware house for

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2538-2544 www.ijaem.net ISSN: 2395-5252

effective business analysis.

To create a successful Data Warehouse Environment, ETL Tool along with the process to extract, is the pre-requisite. **G**athering the scattered data from the real time environment is not possible without ETL tool. So by using some ETL tool it's easy to collect the data from the sources and maintain a repository system. ETL tool is very much needed for an organization to consolidate data from heterogeneous sources to develop an effective data warehousing environment. ETL tools gather data, understand it, and integrate large volumes of raw data from multiple data sources platforms. All the collected data are loaded into a single database to access easily. Then the data is processed and converted to make it significant by joining, reformatting, filtering, merging, and aggregation. Then we can visualize the data in a graphical representation to understand easily.

III. EXTENSIVE ANALYSIS

We have analyzed different types of survey works done by several authors and constructed an outlook of differentiation in the below table.
NA'- Indicates doesn't exist '-' indicates no mentions of the parameter in the related paper.

Table 1 : Survey Table

Authors	Basic Approaches	Propositions	Structure Derived	Implementation /Tools	Benefits & Limitations
B. Rieger, A. Kleber, E. von Maur (2000)	Metadata, Keyword Based	Integration of quantitative + qualitative information sources	NA	Web based Tools + ETL Tools	Benefits: 1.automaticaly catch contextual links 2.dynamic acquisition of supplementary information
M.Z.Bleyberg, K.Ganesh (2000)	Metadata	A model of text warehouse capturing dimension dynamics	Meta-snowflake schema	Oracle DBMS + Java + embedded SQL	Benefits: 1. scalablear chitecture 2. Well-organized and inspected information extraction.
F. Ravat, O. Teste, R. Tournier (2008)	Domain Ontology	An aggregation function AVG_KW that aggregates keywords into more general ones using domain knowledge.	Raw Textual Measure + Keyword Measure + Formula AVG_KW	OLAP analysis tool: Graphic Olap SQL	Benefits: 1. supports non-numeric textual measures 2. Exploits domain knowledge
C. X. Lin, B. Ding, J. Han, F. Zhu, B. Zhao (2008)	Information Retrieval	A text-cube model on multidimensional text database to study OLAP over it, by processing in full cube, partially	Term Hierarchy + Term Frequency measure + Inverted Index measure	C++(Visual Studio 2005) + SQL(MS SQL Server 2005)	Benefits: 1. Bidirectional analytical navigation of unstructured data. 2. Efficient execution of IR applications Limits: 1. High cost to

		materialized cube, and optimizing cube materializati on with Bounded Query Processing			maintain large data base
Henning Baars ,Hans-George Kemper (2008)	Integrated User Interface + Metadata + Knowledge Management	Three approaches for integrating structured and unstructured data. The approaches are mapped to a three layer BI framework.	Facts + Dimensions	NA	Benefits: 1. Applied in domains of CI & CRM to explore hidden interrelations Limits: 1. Higher cost for the end users
Zhang, Zhai, Han, Srivastava,O za (2009)	Text OLAP	A data model(Topic Cube) to combine OLAP and Topic models, to extend OLAP to text dimensions	Topic dimension + Context Dimension + text content measures	NA	Benefits: 1. flexibily explore contents in text documents
K. Srinivasa, N. Prasad, S.Ramakrish n a (2010)	Text Analytics-Text Tagging and Annotation using XML	A framework for handling unstructured data to fit it into business applications like Data Warehouses.	Entity dimension+ Feature Dimension + Keyword Attributes	NA	Limits: 1. Volumino us programming for extracting textual data from varioussources. 2. Managing integrated unstructured data.
M.Thenmoz hi, K.Vivekanan d an (2012)	Source and Tagged Ontology	A framework is designed to demonstrate the business requirements by using natural language process	Facts+ Measures+ Dimensions	RDBtoOnto+ JXML2OWL + OntoLT + PROMPT	Benefits: Completely an automatic process, measures the facts and recognizes functional dependencies
G. Garani, and S. Helmer	Integrating approach	Data integration from	Star schema+ Snowflake	OLAP Server, ETL Tool	Simple and fst query ,

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2538-2544 www.ijaem.net ISSN: 2395-5252

(2012)		different data sources	schemas		Easier to maintain in a data warehouse
Y. Sharma, R. Nasri and K. Askand (2012)	Data Ware house architecture based on service oriented architecture	Ability to develop composite applications	Service oriented architecture	ETL tools	Helps to analyze bulk of data in a middle level organization
S. Mandal, and G. Maji (2016)	Data integration from various Operational databases	Create a centralized data warehouse to analyze business status	Integrating clients data to the centralized warehouse	ETL tool, MS Sql Server	Flexibility of storing a huge amount of data into the data warehouse to take necessary and effective decision for future business analysis
G. Kassem and K. Turowsk (2018)	Matching of business data in a generic business process warehousing	Provides a generic solution for analyzing the business status using data warehousing technology	Developed generic solution using warehouse data	Multidimensional model ETL	Analyze business data for business process using warehouse data
P. S. Diouf, A. Boly and S. Ndiaye (2018)	Variety of data in ETL Process	Data Validation	Meta Data	ETL Tool	Easier implementation of the framework so that to access the required data easily
N. El Moukhi, I. El Azami and A. Mouloudi (2019)	Future challenges	Necessary terminologies for understanding Data Warehousig mechanism	OLAP+ Data Mining+ Multidimensional data bases	Operational data bases, ETL	Provides an effective way to maintain efficient databases like Business,Medical, Scientific and Research analysis fields
J. Sreemathy, I. Joseph V., S. Nisha, C. Prabha I. and G. Priya R.M., (2020)	Data integration in ETL process	General ETL process framework using Talend tool	Extract+ Transformation+ Load	Talend Tool	Gathers the scattered data to make an effective decision support system

IV. CONCLUSION & FUTURE WORK

Data warehousing is one of the key technology for any organizations or enterprise. It helps to handle the analytical terms such as CRM, ERP, Supply chain, Products, and Customers. According to some authors, Unstructured data has grown extensively to be treated as an important part of industrial decision making process and it needs to be incorporated into the warehouse and structures derived from it for information retrieval, knowledge discovery

and business intelligence. In this paper we have presented a survey of different approaches being proposed by various authors to deal with the unstructured data in the data warehouse. We have studied all the possible and related works done by the different authors and presented a relative analysis in a tabular form. In future, we'll be trying to develop a frame-work for manipulating the un-structured data to get the better result from the on-going research works.

The future will give the real time data warehouse updates with the ability to give the organizations a view of ongoing projects and take action either manually or through a condition triggered by the data warehouse data. It will increase the productiveness of the decisions made by the decision-makers or analysts by designing a structured warehouse of consistent, subject oriented and historical data. It combines data from different inconsistent sources which gives a compatible view of the organization. It converts the raw data into a meaningful information which allows the business analysts to execute more substantive, accurate, and consistent analysis.

REFERENCES

[1] W.H. Inmon. "Building the Data Warehouse". John Wiley, 1993.

[2] Kimball, Ralph; Margy Ross, Warren Thornthwaite, Joy Mundy, Bob Becker . The Data Warehouse Lifecycle Toolkit (2nd ed.). Wiley, ISBN 978-0-470-14977-5, 2008

[3] B. Rieger, A. Kleber, E. von Maur. "Metadata-Based Integration of Qualitative and Quantitative Information Resources Approaching Knowledge Management" ECIS 2000

[4] M. Z. Bleyberg and K. Ganesh, "Dynamic Multi-Dimensional Models for Text Warehouses", Proc. IEEE International Conference on Systems, Man, and Cybernetics, Nashville, Tennessee, 2000

[5] Maria Zamfir, Bleyberg and Pallavi S Paranjape, "A Content Delivery Strategy for Text Warehouses", IEEE International Conference on Systems, Man and Cybernetics, vol.4, Page (s) 2322 - 2325, 2001

[6] Hiroshi Ishikawa, Manabu Ohta, Koki Kato . "Document Warehousing: A Document-Intensive Application of A Multimedia Database", Eleventh International Workshop on Research Issues in Data Engineering, pp.25,31, 2001

[7] J. M. Perez, R. Berlanga, M. J. Aramburu, T. B. Pedersen, "A relevance-extended multi-dimensional model for a data warehouse contextualized with documents," Proc. 8th ACM Intl. Workshop Data Warehousing and OLAP, pp. 19 - 28, 2005

[8] Tseng, A.Y.H. Chou, "The concept of Document Warehousing for multidimensional modeling of textual-based Business Intelligence", Decision Support System, vol.42, pp.727- 744, 2005

[9] F. Ravat, O. Teste, R. Tournier, "OLAP aggregation function for textual Data Warehouse", Proc. Int. Conf. on Enterprise Information Systems (ICEIS 2007), Vol. DISI, INSTICC Press, pp. 151-156, 2007

[10] Cindy Xide Lin, Bolin Ding, Jiawei Han, Feida Zhu, Bo Zhao, "Text Cube: Computing IR Measures for Multidimensional Text Database Analysis". Eighth IEEE International Conference on Data Mining . ICDM '08, 2008

[11] Henning Baars & Hans-George Kemper Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework, Information Systems Management, 25:2, pp.132-148, 2008

[12] D. Zhang, C. Zhai, J. Han, "Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases", Proc. of the 9th SIAM Int. Conf. on Data Mining, pp.1123-1134, 2009

[13] Kalli Srinivasa Nageswara Prasad, Prof. S. Ramakrishna, "Text Analytics to Data Warehousing", (IJCSE) International Journal on Computer Science and Engineering ,Vol. 02, No.06, Pg 2201-2207, 2010

[14] M.Thenmozhi, K.Vivekanandan. "An Ontology based Hybrid Approach to Derive Multidimensional Schema for Data warehouse", International Journal of Computer Applications (0975 – 8887) Volume 54– No.8, 2012

[15] Ahmad Abdullah Alqarni, Eric Pardede . "Integration of Data Warehouse and Unstructured Business Documents", 15th International Conference on Network-Based Information Systems, 2012

[16] Y. Sharma, R. Nasri and K. Askand, "Building a data warehousing infrastructure based on service oriented architecture," 2012 International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM), Dubai, pp. 82-87, 2012

[17] G. Kassem and K. Turowski, "Matching of Business Data in a Generic Business Process Warehousing," International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, pp. 284-289, 2018

[18] J. Sreemathy, I. Joseph V., S. Nisha, C. Prabha I. and G. Priya R.M., "Data Integration in ETL Using TALEND," 6th International Conference

on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, pp. 1444-1448, 2020

[19] N. El Moukhi, I. El Azami and A. Mouloudi, "Data warehouse state of the art and future challenges" International Conference on Cloud Technologies and Applications (CloudTech), Marrakech, pp. 1-6, 2015

[20] G. Garani, and S. Helmer, "Integrating Star and Snowflake Schemas in Data Warehouses", International Journal of Data Warehousing and Mining, Vol. 8, No. 4, pp. 22-40, 2012

[21] P. S. Diouf, A. Boly and S. Ndiaye, "Variety of data in the ETL processes in the cloud migration and validation : State of the art" IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, pp. 1-5, 2018

[22] S. Mandal, and G. Maji, "Integrating Telecom CDR and Customer Data from Different Operational databases and Data warehouses into a Central Data Warehouse for Business Analysis", International Journal of Engineering and Technical Research, Vol. 5, pp. 516- 523, 2016