A New Approach to Use Big Data Tools to Substitute Unstructured Data Warehouse

Oras Baker
School of Computing,
Southern Institute of Technology
Invercargill, New Zealand
oras.baker@sit.ac.nz

Chuong Nguyen Thien
School of Computing
Southern Institute of Technology
Invercargill, New Zealand
chuongit@gmail.com

***Abstract*— Data warehouse and big data have become the trend to help organise data effectively. Business data are originating in various kinds of sources with different forms from conventional structured data to unstructured data, it is the input for producing useful information essential for business sustainability. This research will navigate through the complicated designs of the common big data and data warehousing technologies to propose an effective approach to use these technologies for designing and building an unstructured textual data warehouse, a crucial and essential tool for most enterprises nowadays for decision making and gaining business competitive advantages. In this research, we utilised the IBM BigInsights Text Analytics, PostgreSQL, and Pentaho tools, an unstructured data warehouse is implemented and worked excellently with the unstructured text from Amazon review datasets, the new proposed approach creates a practical solution for building an unstructured data warehouse.**

***Keywords*— *Data Mining, Big Data tools, Data warehouse, Text Analytics, PostgreSQL*.**

I. INTRODUCTION:

Business data is originating in various kinds of sources with different forms from conventional structured data like the details of employees stored in relational databases to unstructured data like emails or text documents, these components are the inputs for producing useful information essential for businesses in decision making and earning competitive advantages.

Data warehouses typically store structured business data. However, most data in organisations is unstructured data, which occupies about 80 per cent of an enterprise's data [1], [2]. Therefore, it is vital to transform the traditional data warehouse into an efficient unstructured data warehouse. Big data refers to huge data sets characterised by larger volumes and greater variety and complexity, generated at a higher velocity than the normal operational data that an organisation has handled before. As more and more enterprises recognise the values and advantages associated with big data insights, the adoption of big data tools like Hadoop ecosystem is growing. Hence, utilising big data tools as an enhancement to the data warehouse to handle unstructured data besides structured one is a feasible and practical approach to resolve the limitation of the traditional data warehouse and potentially expand its adoption in organisations.

The aim of this research is to propose a solution for building an unstructured data warehouse with big data tools and to demonstrate that traditional data warehouses can support unstructured data by an intermediate phase to convert unstructured data to structured data before it is loaded into data warehouses. with the help of this method, organisations can still use their current data warehouses with an additional component to support unstructured data.

II. RELATED WORK:

There are various techniques implemented by scholars to support unstructured data for data warehouse systems. Gupta and Rathore [3] summarised the challenges to deal with unstructured data such as: Extracting the right information from it, rebuilding it into knowledge, analysing it to discover its patterns and trends, storing information for effective access, controlling the workflow and making beneficial business intelligence reports.

Text documents as unstructured data can be processed and analysed then integrated with the conventional structured data for better business utilisation. In textual document warehouse, information from text data is extracted by 3 main techniques: Text mining, information retrieval, and information extraction [4]. Text mining and natural language processing (NLP) are two techniques for knowledge discovery from textual context [5]. In general, text mining is a popular technique to extract beneficial patterns in text data while NLP is a technique for information discovery.

Prasad and Ramakrishna [6] examined various text analytics techniques to process text documents. Text analytics techniques, a superset of text mining and cooperated with linguistic and statistical techniques, are the solutions to tame unstructured data as the structured one to be able to extract facts, answers to questions, and knowledge for decision making support.

The limitation of text analytics techniques is that extracting text data from various sources needs a lot of programming effort.

Text tagging and annotation technique, also known as named entity extraction, one of the text analytics techniques based on NLP and machine learning, is a proposed solution to analyse text data and determine names related to domain-specific entities. Entities and features are similar to dimensions in a conventional decision support model.

978-1-7281-9246-8/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: National Sun Yat Sen Univ.. Downloaded on April 10,2023 at 08:19:56 UTC from IEEE Xplore. Restrictions apply.

Sukumaran and Sureka [7] indicated that the named entity extraction technique has been a research topic for many years and is integrated into many open-source or commercial systems in various domains with a good level of accuracy. This technique is used to add structure or meaningful information to unstructured data to make it ready to be integrated with other structured data sources.

Text documents as unstructured data can be processed and analysed then integrated with the conventional structured data for better business utilisation. Gupta [8] discussed the deriving facts and dimensions from unstructured data, he described that although text tagging and annotation techniques with XML documents are used, some answers for the sample queries cannot be extracted from a single file directly but large groups of unstructured data from various input content need to be explored to recognise the relationships in data for the answers. Besides, some knowledge cannot be extracted from the input explicitly but can be derived from an outside knowledge base or prior learning.

Alqarni and Pardede [9] proposed using WordNet, a large lexical database of English, and they pointed out that domain-specific knowledge is not required when utilising WordNet. In this approach, multi-layer schema and linguistical matching using WordNet are employed for mapping different structured data and unstructured data into data warehouse. The multi-layer schema task enables the mapping by determining corresponding data and connecting them using XML schema matching based on semantic relations. Wordnet supports semantic meaning and matching. Semantic hierarchy levels are determined by the relation of words, which are used to measure the relevance of the input concepts. The authors implemented a prototype based on WordNet Similarity, a free open-source software package that supports measuring the semantic similarity and relationship between pair concepts, and the result showed that semantic matching is crucial to compute the equivalence of the schemas, but the processing time is high [9].

Tekadpande and Deshpande [10] proposed a system with the ETL process in Hive and traditional dimensional modelling in Hadoop. It is compared with Pentaho Data Integration tool, an ETL tool that can integrate with various data sources including relational databases. Developing ETL tasks in Hive needs a good knowledge of HiveQL and programming, and Hive shell is used to run all ETL operations. Meanwhile, Pentaho Data Integration tool, also known as Spoon, provides a straightforward user interface with drag-and-drop support to develop and execute ETL tasks.

Tekadpande and Deshpande [10] used the star schema for multidimensional modelling and different transformations such as filter, aggregations, and joins were carried out. Pentaho is better than the Hive in multipoint transformation, and Pentaho also supports SCDs type 1 and type 2 while Hive does not support SCDs. Data modelling in Hive is done by coding while Pentaho can integrate with the database and provide direct modelling on its straightforward user interface. The result shows that the handling of unstructured data in Hive has more benefits than in Pentaho, and Hive runs better than Pentaho regarding the processing time.

NoSQL or "Non-relational" databases as their names are non-relational DBMS which are different from the standard RDBMS. They are suitable for big data such as unstructured, semi-structured data, or frequently changed data while the regular RDBMS do not support scaling and are not effective in processing big data.

Sahiet and Asanka [11], proposed an approach to construct a data warehouse for unstructured data stored in NoSQL databases by supporting an ETL framework to extract, transform and load unstructured data from NoSQL databases to the traditional data warehouses. This approach conforms with the contemporary trend of enterprise systems that can be integrated with NoSQL databases to support unstructured document data. Hence, NoSQL systems should not be the substitute for enterprise systems such as data warehouse or business intelligence systems, but they can be integrated together to have additional enhanced capabilities, especially the ability to handle the unstructured or semi-structured data besides the structured data. However, other researchers implemented a different solution for building an unstructured data warehouse that uses NoSQL databases directly as the data marts of the data warehouses [12].

III. PROPOSED SYSTEM:

The proposed unstructured data warehouse follows the three-tier architecture. In this design, IBM BigInsights Text Analytics platform does not substitute but augment the existing traditional data warehouses of organisations by enabling them to receive data from unstructured textual sources.

IBM BigInsights contain a text processing engine and a library of predefined extractors that help identify and extract relevant structured items from textual documents [13].

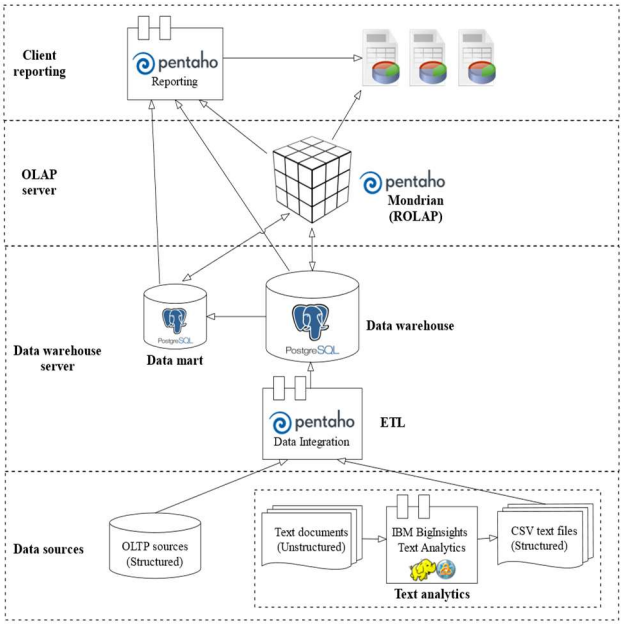


Fig. 1: Unstructured textual data warehouse architecture

As illustrated in Fig. 1. Architecture Design, unstructured textual data will be handled first by the Text Analytics platform before its structured results are combined with other structured data sources to be loaded into data warehouse

Authorized licensed use limited to: National Sun Yat Sen Univ.. Downloaded on April 10,2023 at 08:19:56 UTC from IEEE Xplore. Restrictions apply.

servers in the bottom tier by Pentaho Data Integration tool in ETL process. ETL tasks will help transform the input structured text to match with the multidimensional data design.

Online analytical processing (OLAP) server is implemented with Pentaho Mondrian located at the middle layer to support reporting and querying the data warehouse. The top tier is the frontend layer with Pentaho reporting tools integrated with the OLAP server or data warehouse servers to help users make various analytics reports.

In order to import text documents into data warehouse, Text Analytics tool is used first to extract structured information from text documents to a Comma Separated Value (CSV) files, then Pentaho Data Integration tool is used to transform and load dimension data into data warehouse before fact data is transformed, mapped with dimension data and loaded into data warehouse.

For new data import, new data tables for dimensions and facts will be created. For incremental data import, existing data tables of dimensions and facts will be updated or refreshed as shown in the activity diagram Fig. 2.

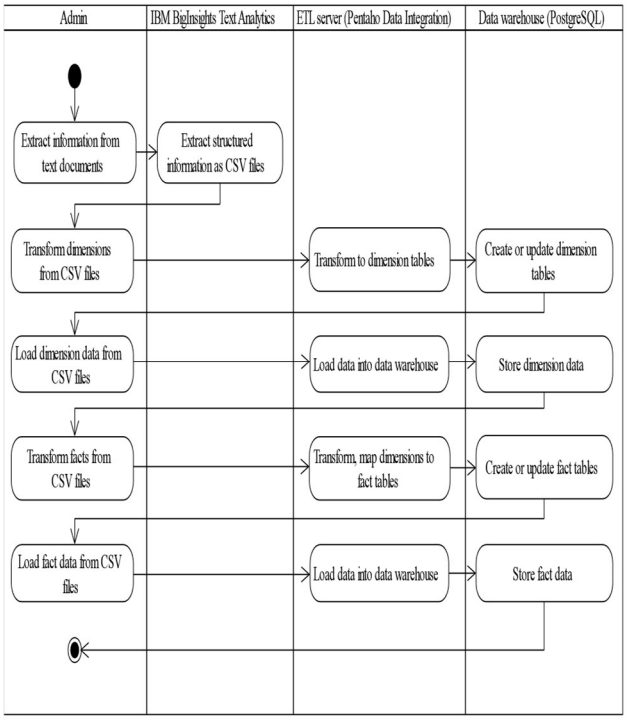


Fig 2 . Activity diagram for importing text documents into data warehouse

The process of extracting new text documents with incremental illustrated in Fig. 3. below show the detailed steps to extract new text documents that contain either new or existing types of information. The input files can be imported into Text Analytics web tool or put in HDFS servers. Users identify all the information to extract.

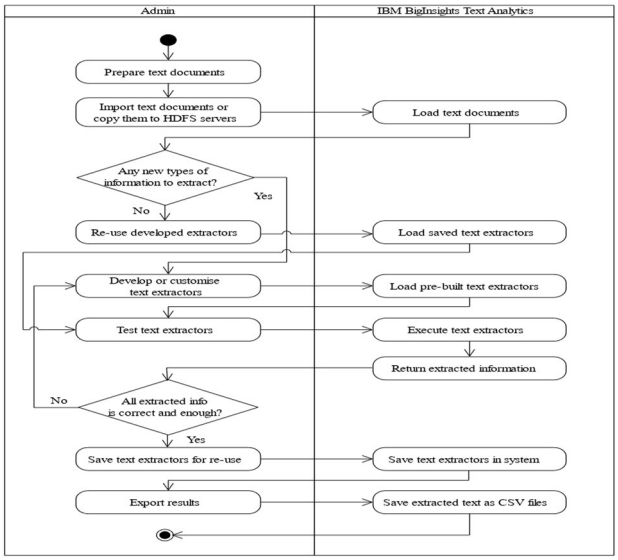


Fig. 3: Activity diagram for incrementally extracting information from text documents.

Amazon customer reviews dataset will be used for this experiment. As introduced in its documentation page [14] this public dataset is available for academic researches, especially with millions of customer reviews as a rich source of unstructured text data for text analytics in this research.

IV. DATABASE DESIGN:

Data is designed with a multidimensional star schema as shown in Fig. 4. to support the multidimensional analytics reporting. Fact data is retrieved from the Text Analytics extractor results, and data for dimensions are extracted from the selected Amazon datasets. The review date is extracted along with the review data, then it is modelled in a separate time dimension with additional fields generated at the transformation step to help analyse review data with this time dimension from different perspectives.

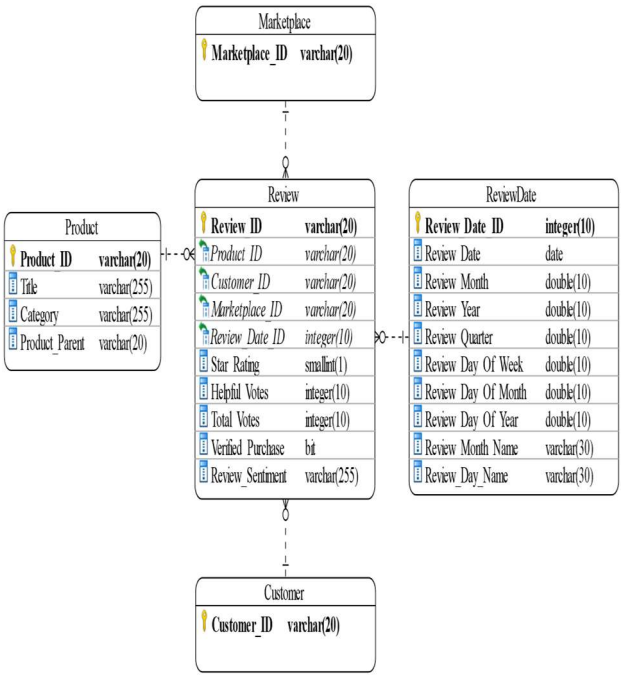


Fig. 4: Star schema design for Amazon review datasets

Authorized licensed use limited to: National Sun Yat Sen Univ.. Downloaded on April 10,2023 at 08:19:56 UTC from IEEE Xplore. Restrictions apply.

V. TEST METHOD:

Testing will be conducted manually along with the implementation by verifying each phase output with the acceptance criteria:

- Data extraction: Sample dataset files will be tamed by Text Analytics tool.
- ETL process: Extracted text as CSV files will be transformed, cleansed and loaded into data warehouse.
- Analytics reporting: OLAP cube can be created from data warehouse

VI. IMPLEMENTATION AND RESULT:

Following the proposed architecture and database designs an unstructured data warehouse has been implemented and tested successfully with the Amazon product review datasets. Sentiment information as structured data could be extracted from the unstructured review text by utilising the text extractor for sentiment analysis in IBM BigInsights Text Analytics. ETL process was completed without issues with Pentaho Integration tool to extract, transform, load and refresh the data of the review fact and its dimensions from the CSV files including the extracted review sentiment into data warehouse in the multidimensional star schema design.

The developed extractor is executed and tested on the web tool. Results are shown and highlighted on the tool as in Fig. 5. For some results that are not exact yet, an extractor is reworked by updating the dictionaries for better sentiment analysis. The values in the polarity column, either positive or negative, are the sentiment of product review that will be used in data warehouse for analytics reporting. Other fields in results such as text, pattern name, clue, and target are the extra information to explain the extractor rules and how they run to determine the polarity results. They can be omitted in the exported results.

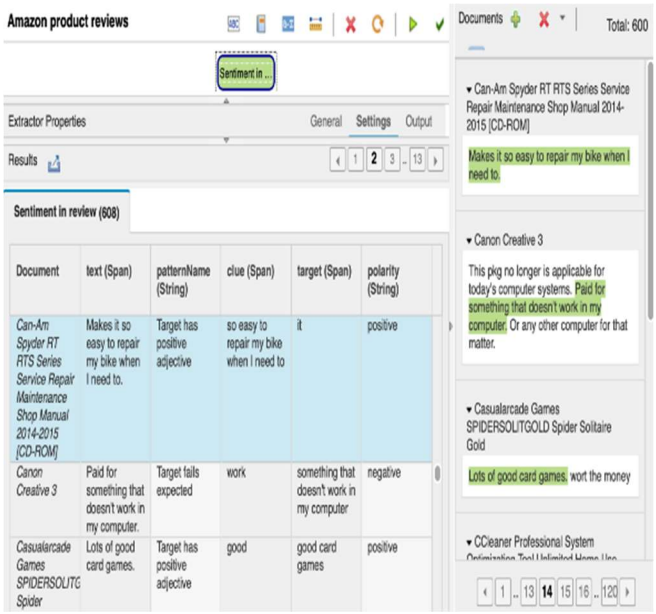


Fig. 5: Results of sentiment analysis for Amazon product reviews.

The final extractor is saved and executed on a Hadoop cluster against the full Amazon datasets stored in HDFS server, that can be invoked right in the Text Analytics web tool. Then the result files are retrieved from HDFS server as the input for ETL process to be loaded into data warehouse.

The transformation for Review fact data is rather the same as the transformation for dimensions except fact data needs to be linked with its dimensions already transformed and loaded into database in previous steps. Some calculated fields for fact can also be created. As the steps in Fig. 6. This transformation starts with loading CSV file of the Review dataset including the review sentiment information extracted by Text Analytics before. Then the properties such as data type, format, length of all loaded fields need to be defined or adjusted accordingly to match with the datasets and designed database.

For mapping fact data with its dimensions in database, each related dimension such as product, customer, marketplace and review date are looked up and their keys are mapped. At the last step, the suitable fields in the processing stream are selected for fact table before it is created in database by executing the generated SQL scripts. Then the transformation is ready to run against different Review datasets of various product categories and marketplaces to load them into the database.

Fig. 6: Transform Review fact

There are few errors with some Amazon datasets when loading them into the data warehouse, such as invalid values in the marketplace, product title or review date, and some rows with missing columns. The transformations would throw errors in these cases that could be fixed by correcting or cleaning the error data or updating the transformations like changing the mapping fields or fixing the data length. Pentaho tool supports regenerating SQL update scripts as a hotfix for any database changes.

After Amazon datasets are loaded into data warehouse successfully, an OLAP cube is created for some analytics to demonstrate that this unstructured data warehouse is working with the new data updated from the unstructured data source.

Authorized licensed use limited to: National Sun Yat Sen Univ.. Downloaded on April 10,2023 at 08:19:56 UTC from IEEE Xplore. Restrictions apply.

For instance, the product review sentiment extracted by Text Analytics is analysed from different dimensions and perspectives, useful for the management team to understand and improve their businesses. As in Fig. 7. Pentaho Schema Workbench tool is used for designing OLAP schema with its fact and dimensions similar to the schema in the data warehouse.

The OLAP schema is a multi-level hierarchy in which the fact contains its dimensions. The review fact table is designed with 2 measures for positive and negative sentiments that will be aggregated in the reports. The schema defines the fields that can appear in the OLAP cube.

The review date dimension needs to be correctly marked as a time dimension. After the schema is completed, it is published to Pentaho Business Analytics server as a data source for making analytics reports.

Saiku Analytics tool, a popular plugin in Pentaho Business Analytics, is utilised in this experiment for working with OLAP reports. It can help create dashboards and MDX queries to display OLAP cube in tables or charts.

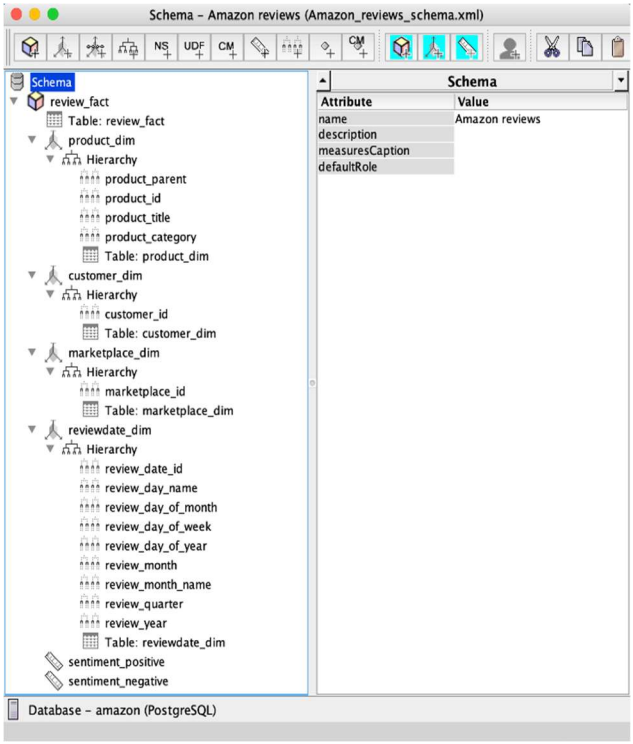
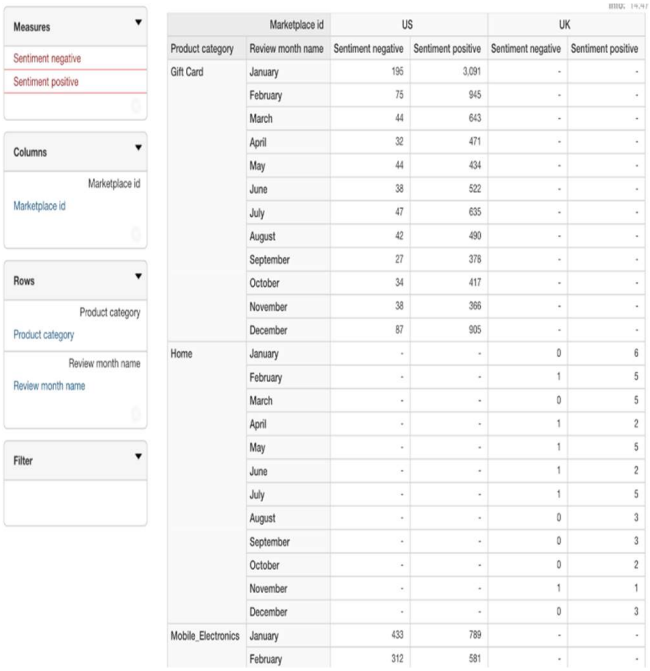


Fig. 7. OLAP schema design for Amazon review datasets

After the schema is completed, it is published to Pentaho Business Analytics server as a data source for making analytics reports.

Using Saiko analytics tools enabled creating dashboards and MDX queries to display OLAP cube in tables or charts. The OLAP cube for Amazon review data in Fig 8. shows the sentiment measures with marketplace dimension in columns, and product category and month dimensions in rows.



	Marketplace id	US		UK	
Product category	Review month name	Sentiment negative	Sentiment positive	Sentiment negative	Sentiment positive
Gift Card	January	195	3,091	-	-
	February	75	945	-	-
	March	44	643	-	-
	April	32	471	-	-
	May	44	434	-	-
	June	38	522	-	-
	July	47	635	-	-
	August	42	490	-	-
	September	27	378	-	-
	October	34	417	-	-
	November	38	366	-	-
	December	87	905	-	-
Home	January	-	-	0	6
	February	-	-	1	5
	March	-	-	0	5
	April	-	-	1	2
	May	-	-	1	5
	June	-	-	1	2
	July	-	-	1	5
	August	-	-	0	3
	September	-	-	0	3
	October	-	-	0	2
	November	-	-	1	1
	December	-	-	0	3
Mobile_Electronics	January	433	789	-	-
	February	312	581	-	-

Fig. 8. An OLAP cube for Amazon review data.

VII. CONCLUSION:

The proposed solution is similar to the total data warehouse solution recommended by [3]. However, the suggested design of this research is further advanced by utilising IBM BigInsights Text Analytics tool with hundreds of pre-built text annotators or extractors that make the text analytics work much simpler. In the experiment, the extractor with sentiment analysis returned all correct results after being customised to extract review sentiment from Amazon datasets. Overall, although this solution is convenient and flexible, it is still using RDBMS to store the information extracted from unstructured sources while these sources can grow much bigger later, which can impact the performance of the data warehouses. A forthcoming design should use big data tools to process, store and analyse the unstructured data completely as a data warehouse, so that the advanced design of big data tools can be utilised in the right place with greater outcomes. Apache Hive can be an unstructured data warehouse with similar design that requires transforming the unstructured data to structured data to be stored in it. NoSQL can be a suitable data warehouse for unstructured data once the support tools like ETL, analytics reporting is ready. However, it is challenging and takes significant effort to design and implement big data tools as a complete replacement for a data warehouse. Using them together like the current design is still a practical solution.

REFERENCES

[1] Kelly, J. E. (2015). Computing, cognition and the future of knowing. IBM.

[2] Roberts, P. (2010). Corralling Unstructured Data for Data Warehouses. Business Intelligence Journal, 15(4), 50-55.

[3] Gupta, V., & Rathore, N. (2013). Deriving Business Intelligence from Unstructured Data. International Journal of Information and Computation Technology, 3(9), 971-976.

[4] Gonzalez, S. M., & Berbel, T. d. (2014). Considering unstructured data for OLAP: a feasibility study using a systematic review. Salesian Journal on Information Systems, 14, 26-35.

Authorized licensed use limited to: National Sun Yat Sen Univ.. Downloaded on April 10,2023 at 08:19:56 UTC from IEEE Xplore. Restrictions apply.

[5] Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing. 2011 5th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE.

[6] Prasad, K. S., & Ramakrishna, S. (2010). Text Analytics to Data Warehousing. International Journal on Computer Science and Engineering (IJCSE), 02(06), 2201-2207.

[7] Sukumaran, S., & Sureka, A. (2006). Integrating structured and unstructured data using text tagging and annotation. Business Intelligence Journal, 11(2), 8-17.

[8] Gupta, V. (2013). Extracting Facts And Dimensions From Unstructured Data For Business Intelligence. International Journal of Engineering Research & Technology (IJERT), 2(7), 2602-2606.

[9] Alqarni, A. A., & Pardede, E. (2012). Integration of data warehouse and unstructured business documents. 2012 15th International Conference on Network-Based Information Systems (pp. 32-37). IEEE.

[10] Tekadpande, S., & Deshpande, L. (2015). Analysis and Design of ETL process using Hadoop. International Journal of Engineering and Innovative Technology (IJEIT), 4(12), 171-174.

[11] Sahiet, D., & Asanka, P. D. (2015). ETL framework design for NoSQL databases in dataware housing. International Journal of Research in Computer Applications and Robotics, 3(11), 67-75.

[12] Bicevska, Z., & Oditis, I. (2017). Towards NoSQL-based Data Warehouse Solutions. Procedia Computer Science, 104, 104 – 111.

[13] IBM Corporation. (2018, June). IBM BigInsights 4.1 documentation.

[14] Amazon. (n.d.). Amazon Customer Reviews Dataset.

Authorized licensed use limited to: National Sun Yat Sen Univ.. Downloaded on April 10,2023 at 08:19:56 UTC from IEEE Xplore. Restrictions apply.