Process Unstructured Data using Data Analysis: Apache Spark and Hadoop

Adjovi Irène Sokegbe [1], Ayushi Nainwal [2]
[1,2] School of advanced computing of Alakh Prakash Goyal Shimla University

Abstract— Based on the method of accessing Data there are two types of Data: Data that are already formatted called Structured Data and the non-formatted one called Unstructured Data. Many Organizations today deal with Big Data which is the combination of Structured Data and Unstructured Data. Today, methods to analyze Structured Data are well known and widely developed but for the Unstructured Data, analysis and processing methods are still immature. As observed, approaches, techniques and tools are developed, some free of charge and others chargeable to handle the complexity of the Unstructured data processing. This paper provides the summary of Unstructured Data analysis papers of the last five (5) years.

Keywords— *Unstructured Data; Structured Data; Big Data; Data analysis; Data analytics*

I. INTRODUCTION

The use of data analytics which are Data dates from 19th century while Henry Ford measures the speed of assembly lines [1]. The term Data analytics becomes very important with the introduction of Big Data concept. Hence, Organizations started giving more attention to these concepts as they helped them in decision-making. Internet is facing the personalization of contents according to the profile of each user, for instance social media like YouTube records their user's activities online and then redirect the appropriate advertisements to them. With that phenomena, information comes to you wherever you are. The term Data analytics is none other than the designation of different methods, techniques, tools used for Data analysis. Considering The four "Vs" (Velocity, Volume, Variety, Veracity) characters of Big Data, new techniques are developed to analyze them.

II. BACKGROUND STUDY ON UNSTRUCTURED DATA ANALYSIS

A. Background study on Unstructured Data analysis

This can be any kind of data that is not pre-defined. Unstructured Data doesn't have a special data schema or format. These kinds of data are more complex to handle. They are generated by human- based or machine-based. They are processed by NoSQL (Not only SQL) language and stored in Non-Relational database, data warehouse, data lakes. Examples: Data from social media, Email, text file, video, audio, web pages, sensor data, satellite imagery, scientific data, etc. The tools to manage Unstructured Data are still under development as it is a new age of such data processing in Information Technology. There is also a type of data called Semi-structured Data that can be grouped and their schema can be defined using XML (Extensible Markup Language), Open Standard JSON (JavaScript Object Notation), NoSQL languages.

B. Data analysis

Data analysis is the process of inspecting, cleaning, transforming and modelling data with the goal of discovering useful information, informing conclusion and supporting decision-making. [2] Data analysis is the process of studying data and see whether it can be useful or not for supporting the purpose of business growing. Data are analyzed to retrieve knowledge. It helps to fetch trends, patterns, relationship among data sets.

Types of Data analysis (techniques and methods):

- Text analysis
- Statistical analysis
- Predictive analysis
- Prescriptive analysis
- Diagnostic analysis

Various data analysis tools are Python, Java, SQL, MATLAB, Hadoop, Spark etc.

III. LITERATURE REVIEW

This paper provides different works done in the area of Unstructured Data analysis by various researchers as summarized below.

MadhaviLatha [3] **in 2016, "**Streaming Data Analysis using Apache Cassandra and Zeppelin'' states that Spark Streaming is more efficient than Flume Streaming. Apache Flume is used for the collection and the transportation of Data from Twitter application. Apache Spark Streaming is used for the real-time Data processing from Flume to the Database Cassandra. Apache Zeppelin (web-based solution) is used for the Data Analysis and Data validation combined with JSON. The Solution is shown on dashboard. Scala is used instead of map reduce algorithms because it is 100 times faster.

Flume and Spark Streaming are used for the input of data and the storage of data is performed by the use of Cassandra. For the output Zeppelin is used. The Twitter Application is created to get tweets. Then it generated a number of keys values during streaming into the HDFS (Hadoop Distributed File System), this is the Flume Streaming.

M. Hussain Iqbal and T. Rahim Soomro [4] **in 2015**, "Big Data Analysis: Apache Storm Perspective". Apache Storm is suited for real time Big Data processing. It minimized the time latency compared to Spark. Apache Storm is used to processed data. The Dataset that is used is the Twitter API (Application Program Interface) .Net to show the reliability and

performance of the tool. The result of this project shows the top ten words collected, top ten languages, the number of times a particular word is used. All records are fetched during last 10 minutes.

Mohammad Fikry Abdullah [5] **in 2015**, "Business Intelligence Model for Unstructured Data", proposed a Business Intelligence model for the management of Unstructured Data? This question is how to manage Unstructured Data in order to make decision. It is used in this paper Unstructured Data from NAHRIM (National Hydraulic research Institute of Malaysian). This paper emphasizes the process of transforming Unstructured Data into Structured Data: Extraction, Classification, Repositories development and Data mapping. The process of Extraction is divided into two processes: entity extraction (product, title, etc.) and fact extraction (contact, issues, content, etc.)

Classification is done by the help of meta data which concerns the type of the file. A data repository is built for each type of data (text, video, audio). Data mapping: Dublin Core Metadata element DCME is used to generate meta data then the data is mapped from the repository into its thematic topic (climate change, coastal, river, water etc.).

K. Aziz, D. Zaidouni and M. Bellafkih [6] **in 2018**, "Real-time data analysis using Spark and Hadoop" present the study of Hadoop MapReduce and Apache Spark Data analysis, and give the drawbacks of the use of Hadoop for the real time processing and then the comparison of both tools is given. The experimentation is done on virtual machine with two (2) Ubuntu Machines. Hadoop is a map reduce framework. But Spark does not use map reduce algorithm. Spark can be combined with Hadoop or another and may run with HDFS or another storage device. Spark performs in-memory operation and Hadoop perform disk-based operation. In the case of structured Data MapReduce is the suited one. But for analysing Unstructured Data Spark is the suited one. Spark supports real time batch and streaming processing.

Monica Chand et. al [7] **in 2017**, "Analysis of Big Data using Apache Spark". This paper emphasises the integration of Hadoop into Spark. The collection of data is done by various sources into HDFS and then data are filtered into Hadoop HDFS for the processing. The outcome is shown by using graphs and chart. The result of this research reveals that the processing speed was increased and the system consumes more RAM than Hadoop solution and use less space of Disk.

Ling Chen et. al [8] **in 2015,** "RAISE: A Whole Process Modeling Method for Unstructured Data Management", proposed a whole process model for Unstructured Data management. The system is built into two parts called RAISE and D-OCEAN. RAISE gathers the repository, analysis, index, search and the environment modelling. It takes an image as a data type and processes it as an object. They design also a SQL-like language called Unstructured Query Language to avoid the use of complex APIs. D-OCEAN is a distributed Unstructured Data management system. It consists of two part the infrastructure layer for storage purpose and a core layer for the central scheduling job.

Jinhua Chen [9] **in 2016,** "Study of data analysis model based on Big Data technology", presents the study of Data analysis model based on Big Data technology. This paper gives a clear definition of Data analysis, using elements such as origin, essence, method, process, result and purpose. Function and mode of Data analysis also are listed. This paper emphasises also the Big Data inside Data analysis and reveals that Big Data is based on the correlation relationship rather than causality. It states that the new trades in the development of large Data analysis refer to the Data acquisition taking into account the accurate selection of source and its automatic correction methods, the improvement of large Data processing methods, the visualization and the Big Data security.

Lukas Probst, et al. [10] **in 2019**, "Integrated Real-Time Data Stream Analysis and Sketch-Based Video Retrieval in Team Sports" present the Big Data sense in Sport field. They design a STREAMTEAM, the real time Data analysis module and SPORT SENSE, the video retrieval system. The system automatically detects events, analyses them and provides the visualization module in real time. The system is composed of Kafka as a data source, MongoDB for the Database storage. The YARN cluster is also used for the resource management. This system is a solution for Data analysis of raw position Data in real time.

Kuldeep Sambrekar et al. [11] **in 2018**, "A Proposed Technique for Conversion of Unstructured Agro-Data to Semi-Structured or Structured Data", proposes a solution to convert Unstructured Data (Agro-data) into semi-structured or Structured Data using NoSQL tool Couchbase. Map reduce algorithm is used to build the application of conversion of Unstructured Data into Structured one. Data come from Sensors, IOT, agriculture websites.

Hichem Dabbechi et al. [12] **in 2018**, "A unified multidimensional data model from social networks for unstructured data analysis", provides an extension solution for OLAP in order to deal with the complexity of Unstructured Data processing. It takes a multidimensional Data source such as Image, video, smileys, URL, Text, Hashtags and mention from Social media. The given system can work for any kind of social media dynamically. It responds to the limitation of OLAP in term of Unstructured Data processing. It can analyse the social data according to their geographic locations and temporal axes. It works with three (3) types of measures: numerical, textual and list of elements measures.

Sriraghav K. et al. [13] **in 2017**, "ScrAnViz- A tool to Scrap, Analyze and Visualize Unstructured Data using Attribute-based Opinion Mining algorithm", proposed a solution based on end-to-end processing. Input is data extracted from website on which some attribute keys are applied then the NLP module processed these elements. The output is given by a visualization tool. This is a tool for structuring and visualizing in one set.

Shruti Nair et al., in 2018 [14], "Data analysis on multivariate Image set''. The system uses MultiSciView for image visualization and the analysis module consist of three (3) modules: feature analysis to show the relationship between attributes in an image, data analysis based on global data panel

and Colorization Scheme then Image analysis that uses X-ray image analysis. The given solution allows to visualize data in 2D space.

Jian Ming et al. in 2018 [15], "Analysis Models of Technical and Economic Data Mining Enterprises Based on Big Data analysis" present a data analysis model that encompasses data analysis and mining techniques to predict the economic face of a commercial product. It is built using a neural network algorithm for the prediction, the casual prediction. Also, interpolation methods of the technical and economic data are analysed. Techniques of geostatistics and BP neural network are used for interpolation and prediction of geological missing data. The system built is strong and gives good accuracy.

Paolo Lo Giudice et al. in 2019 [16], "An approach to extracting complex knowledges patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake." In this paper, an approach of structuring unstructured sources, a logical model to construct network of Data lake sources and then the extraction of knowledge patterns from Data sources are provided. The model of extracting complex knowledge from Data sources consists in using of graph processing and semantic study by the use of dictionary. Finally, the model uses these parameters: average clustering coefficient, density and transitivity. The overall solution presents a low latency but a good correctness indicator.

Qingwu Hu et al. in 2018 [17], "An effective Selecting Approach for Social Media Big Data Analysis- Taking Commercial Hotspot Exploration with Weibo Check-in Data as An Example". A solution of clustering mining Data and analysis is proposed. It consists of an effective solution to solve the problem of limitation of Dataset volume available from social media. For that a Weibo check-in data in Wuhan between 2011 and 2015 mining is used as a data source. Then the spatial exploratory Data analysis ESDA is applied on the data sets and then the hotspot detection to reduce the area. This solution results in fourteen (14) datasets.

Amine El Haddadi et al. in 2015 [18], "Mining Unstructured Data for a competitive intelligence system XEW" present an approach of how to build an intelligence Data warehouse from web technologies sources of unstructured data. It is described the method of managing Unstructured Data.

Lovenika Kushwahain in 2016 [19], "Opinion Mining of Customer Reviews based on their Score using Machine Learning Techniques" proposed a solution for the extraction of knowledge from online shopping customer reviews. This solution aimed to provide a recommendation list based on review of the customer. Then a comparison of Naïve Bayes classifier and AdaBoost classifier was provided. The proposed solution itself implements AdaBoost classifier, and it results in better accuracy and better execution time than Naïve Bayes.

Warih Maharani, et al. in 2015 [20], "Aspect Extraction in Customer Reviews Using Syntactic Pattern". This research used reviews data sets of some electronic products. It is performed an explicit aspect of extraction from these data set using syntactic-based approach to discover new pattern.

Subramaniyaswamy V, et al. in 2015 [21], "Unstructured Data analysis on Big Data using Map Reduce", apply sentiment analysis on social media data set using Map Reduce. It is stated that Map Reduce is used to perform filtering and aggregation task. For the sentiment analysis, NLP and text analysis are required. Twitter sample is taken as Data sample. After using Map Reduce for structuring Unstructured Data, Apache Mahout is used for the collaborative filtering.

Jong-Yeol Yoo and Dongmin Yang in 2015 [22], "Classification Scheme of Unstructured Text Document using TF-IDF(Term Frequency-Inverse Document Frequency) and Naïve Bayes Classifier", in this paper Naïve Bayes and TF-IDF are implemented to classify document according to two categories: True or False, Spam or not. It is concluded that SVM (Support Vector Machine) and Word embedding will give more accuracy than Naïve Bayes.

IV. CONCLUSION

Today, Unstructured Data have taken an important place in most of organization's dataset, as they are very important in terms of volume, velocity, veracity, variety. Many techniques were developed and many studies were conducted to handle them. But for more efficiency, there remain a lot to do in order to improve the parameters such as execution time, processing time and accuracy.

ACKNOWLEDGMENT

We thank Sr. Kishore Keshav, head of the computer science department of our college and our teachers for comments that greatly improved the manuscript.

REFERENCES

[1] S. Duvvuri and B. Singhal, "*Spark for Data Science*." 2006.
[2] S. Duvvuri and B. Singhal, "*Spark for Data Science*." 2006.
[3] A. Madhavilatha and G. V. Kumar, "Streaming Data Analysis using Apache Cassandra and Zeppelin," vol. 3, no. 10, pp. 8–15, 2016.
[4] M. Hussain Iqbal and T. Rahim Soomro, "Big Data Analysis: Apache Storm Perspective," *Int. J. Comput. Trends Technol.*, vol. 19, no. 1, pp. 9–14, 2015.
[5] M. F. Abdullah and K. Ahmad, "Business intelligence model for unstructured data management," *Proc. - 5th Int. Conf. Electr. Eng. Informatics Bridg. Knowl. between Acad. Ind. Community, ICEEI 2015*, pp. 473–477, 2015.
[6] K. Aziz, D. Zaidouni, and M. Bellafkih, "Real-time data analysis using Spark and Hadoop," *Proc. 2018 Int. Conf. Optim. Appl. ICOA 2018*, pp. 1–6, 2018.
[7] M. Chand, "Analysis of Big Data using Apache Spark," pp. 1975–1980, 2017.
[8] L. Chen, J. Shao, Z. Yu, J. Sun, F. Wu, and Y. Zhuang, "RAISE: A Whole Process Modeling Method for Unstructured Data Management," *Proc. - 2015 IEEE Int. Conf. Multimed. Big Data, BigMM 2015*, pp. 9–12, 2015.
[9] J. Chen, Q. Jiang, Y. Wang, and J. Tang, "Study of data analysis model based on big data technology," *Proc. 2016 IEEE Int. Conf. Big Data Anal. ICBDA 2016*, 2016.

[10] L. Probst, F. Rauschenbach, H. Schuldt, P. Seidenschwarz, and M. Rumo, "Integrated Real-Time Data Stream Analysis and Sketch-Based Video Retrieval in Team Sports," *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 548–555, 2019.
[11] K. Sambrekar, V. S. Rajpurohit, and J. Joshi, "A Proposed Technique for Conversion of Unstructured Agro-Data to Semi-Structured or Structured Data," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–5, 2018.
[12] H. Dabbèchi, N. Haddar, M. Ben Abdallah, and K. Haddar, "A unified multidimensional data model from social networks for unstructured data analysis," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 2017-October, pp. 415–422, 2018.
[13] K. Sriraghav, S. Jayanthi, N. Vidya, and V. S. F. Enigo, "ScrAnViz-A tool to scrap, analyze and visualize unstructured-data using attribute-based opinion mining algorithm," *2017 Innov. Power Adv. Comput. Technol. i-PACT 2017*, vol. 2017-January, pp. 1–5, 2017.
[14] S. Nair, S. Ha, and W. Xu, "Data Analysis on Multivariate Image Set," 2018 New York Sci. Data Summit, NYSDS 2018 - Proc., pp. 1–3, 2018.
[15] J. Ming, L. Zhang, J. Sun, and Y. Zhang, "Analysis models of technical and economic data of mining enterprises based on big data analysis," *2018 3rd IEEE Int. Conf. Cloud Comput. Big Data Anal. ICCCBDA 2018*, pp. 224–227, 2018.
[16] P. Lo Giudice, L. Musarella, G. Sofo, and D. Ursino, "An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake," *Inf. Sci. (Ny).*, vol. 478, pp. 606–626, 2019.
[17] Q. Hu and Y. Zhang, "An effective selecting approach for social media big data analysis-Taking commercial hotspot exploration with Weibo check-in data as an example," *2018 IEEE 3rd Int. Conf. Big Data Anal. ICBDA 2018*, pp. 28–32, 2018.
[18] A. El Haddadi, A. Fennan, A. El Haddadi, Z. Boulouard, and L. Koutti, "Mining unstructured data for a competitive intelligence system XEW," SIIE 2015 - 6th Int. Conf. "Information Syst. Econ. Intell., pp. 146–149, 2015.
[19] L. Kushwaha and S. D. Rathod, "Opinion Mining of Customer Reviews based on their Score using Machine Learning Techniques," pp. 2198–2203, 2016.
[20] W. Maharani, D. H. Widyantoro, and M. L. Khodra, "Aspect Extraction in Customer Reviews Using Syntactic Pattern," Procedia Comput. Sci., vol. 59, no. Iccsci, pp. 244–253, 2015.
[21] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Unstructured data analysis on big data using map reduce," Procedia Comput. Sci., vol. 50, pp. 456–465, 2015.
[22] J.-Y. Yoo and D. Yang, "Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier," vol. 111, no. Comcoms, pp. 263–266, 2015.

I am completing my master studies in Computer Science Engineering after obtaining a professionnal Bachelor in maintenace and coputer network at the University of Lomé (TOGO). I'm deeply interrested in the field of the Data Analysis.