A New Big Data Architecture for Analysis: The Challenges on Social Media

Abdessamad Essaidi, Mostafa Bellafkih
RAISS Laboratory-Department of Mathematics and Computer Science,
National Institute of Posts and Telecommunications (INPT), Rabat, Morocco

***Abstract*—The streams of social media big data are now becoming an important issue. But the analytics method and tools for this data may not be able to find the useful information from this massive amount of data. The question then becomes: how do we create a high-performance platform and a method to efficiently analyse social networks' big data; how to develop a suitable mining algorithm for finding useful information from social media big data. In this work, we propose a new hierarchical big data analysis for understanding human interaction, and we present a new method to measure the useful tweets of Twitter users based on the three factors of tweet texts. Finally, we use this test implementation score, in order to detect useful and classification tweets by interested degree.**

Keywords*—Social media; useful information; big data analysis; stream processing; classification tweets

I. Introduction

Big data on social networks and the progress of tools for calculating and analyzing these data have occupied the first place of investment in the last few years [1]. As a result, much of the literature devoted to Big Data in social media oscillates between two approaches. That is why, in recent years, there is so much interest in public debate on social media [2]. For this, an analysis of the information induced by the use of this type of data was strongly linked by social data. Also, and above all, the appearance of a new form of society "led by the analysis of Big Data." Following this proliferation of information on social media, it seemed useful to us to propose a new analysis method based on big data to draw up an initial assessment of the effects of big data on its practices [3], its objects and its results.

To this end, we analyzed the social network community around two main questions that we felt should not be separated: how does Big Data transform society? [4], How does this data affect personnes practice (Covid-19 use case)?, This double approach – which we wanted to keep in the preparation of this article – shows the fact that the proliferation of information in social networks is today confronted with a real change in the processes of transformation of societies which affects it so directly.

The main objective of this article is how big data in social networks affect the practice of society based on the new architecture of Big Data. The development and analysis of this massive data would thus provide those who use it with new knowledge. This new knowledge will affect the behavior of individuals, their interactions, their uses. We also provide a detailed comparison of the various tools and architectures used to process data stream [5]. Lastly, we offer our own architecture based on comparisons made. So how does big data in social media affect society practice?

To talk about this issue in depth, this paper starts with a review of previous social media research on human interaction data and social big data stream processing architectures on social media platforms, followed by the discussions of our proposed big data analysis architecture, and we suggest a new method with new measuring process based on important aspects, such as number of likes, comments and retweets on the Twitter platform that can be used to calculate the LCR score for tweet text. Finally, we use our data mining algorithm to predict the tweet texts (useful or not), and as the key to crucial insights on human behavior.

II. Research Background

A. Existing Systems

Most works focus on how big data affect the practice of society focused on the concept of analyzing content and patterns of interaction in social networks.

Felt [6], this qualitative research is a study on the analysis of social media data in addition to traditional, qualitative methods to analysis of big data.

Ghani et al. [7] based their study on an overview of recent works and provided a general perspective on the subject of social media big data analytics research. This study also provides a comparison of possible big data analysis techniques and their quality attributes.

Bello-Orgaz et al. [8] have worked this study on the revising new methodologies designed to enable efficient data mining and merging information from social media and new apps and frameworks that are now emerging under the social media "umbrella", social media as well as big data paradigms.

Immonen et al. [9] based their work on the introducing a new framework to assessment and manage quality of social networks data at each treatment phase of the Big Data. The proposed solution delivers real-time validated data to social network users, which results in better decision-making. This data is extracted from social media and used to determine customer satisfaction with the quality of a product.

B. Social Media

The social media platform has become the primary contact with our family, friends and colleagues. People share all the information in a lot of different forms, like messages, video as well as audio on the social networks to shares their feeling

distinguished or bad. The world average for Internet users is 2.5 hours a day on platforms social media. Social networking platform users produce a significant amount of information and data that cannot be processed through traditional data analytics. Therefore, social networking platforms are making substantial investments in this information and big data because it provides insights into the analysis of media, which shapes public opinion, etc. [10].

Social media messages are rich with possibilities for data mining and analysis. This shift on social media provides new challenges for researchers interested in analyzing public publications and messages of social networking user platforms to better understand human interaction and improve the human condition [6].

Social media platforms like Twitter were viewed as a critical source of useful information.

During Covid-19, people post updates regarding their statuses on social media platforms like Twitter, request assistance and other relevant information, report damage to infrastructure, injured people, etc.

Information posted on social media during crises, their worth varies considerably. Most publications do not contain any useful information nor are they useful for disaster response. Most of the posted messages contain a deluge of noise that is of a personal nature, do not contain any useful information, or are irrelevant [11].

The proposed model for obtaining information on social media, which is shown in Fig. 1, aims to collect clean data consisting of posts containing potentially useful information. This information can be used for various purposes, such as helping injured people affected by Covid-19.

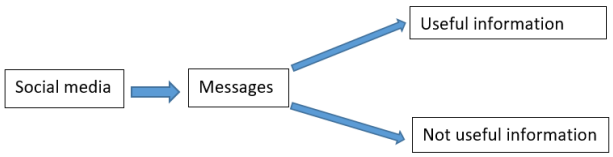

Fig. 1. Information on social media.

C. Big Data

Big data is a term that aims to describe massive volumes of data and offer an alternative to conventional solutions of analyse databases in order to make sense of it and make the most of it.

The Data mining on social media and efficient utilisation of this Big data became a significant issue for numerous research fields such as massive data processing, big data analytics techniques, machine learning, semantics of the data, information fusion, computational intelligence, and social networks.

The big data concept has been specified by the three essential dimensions 3V: volume, variety, and velocity that were set out in 2001 by Laney [12] as: “high-volume, high-variety and high-velocity information, new approaches to information processing and use this information for various purposes such decision making”.

This set of Big Data 3V models, which is shown in Fig. 2, provides a direct and broad definition related to what constitutes a big data-based problem, software, application, or framework, as well as what does not. In this context, it may be described in brief as follows:

- Volume: Designates big amounts of different data and from various sources, including mobile digital data development devices as well as digital devices. The benefit of collecting, processing and analyzing these large quantities of data creates a number of difficulties in obtaining for knowledge human interaction and enterprises [8-13].
- Variety: The variety of data refers to different types of structural heterogeneities within a dataset. These various types of information collected via social networks, smartphones or sensors, such as text, data logs, pictures, videos, audio, and the like. Additionally, these data is structured (such as relational databases data), semi-structured, or of an unstructured form [8-14].
- Velocity: Refers to the rate at which are generated and speed of data transfers. Proliferation rapid and growth of digital devices technologies as smartphones and sensors have resulted in record levels of data creation, and the various forms of streamed data from variety sources. As a result, velocity of big data proliferation at the time of the development process should be taken into consideration [14, 15].

Big Data represents the central set of technologies and parts for processing user-generated data used on social media platforms. Without regard on the type of data (structured, informal or semi-structured), they can be containing useful information.

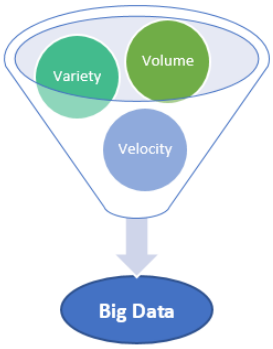

Fig. 2. The 3V model of big data.

III. Social Big Data Stream Processing Architectures

In this section, we provide a brief summary of Big Data characteristics generated by social media platforms, and we will look at the two major flow processing architectures, Kappa and Lambda—that are adopted for executing analytics.

When looking at popular online social media containing stacks of big data, makes the underlying processing of this data becomes challenging, and requires the implementation of an application-specific.

Big data technology influences current data management, and makes comparative analysis of these data essential to academic and industry communities.

Given the huge interest in social media platforms big data by universities, industry, a variety of solutions and techniques have been published over the past few years. With their progressive maturity, there is an increasing need to evaluate and compare such solutions. As a result, the science community was particularly interested in big data techniques. In fact, comparative analysis greatly facilitates the comparison of performances and provides useful information [16].

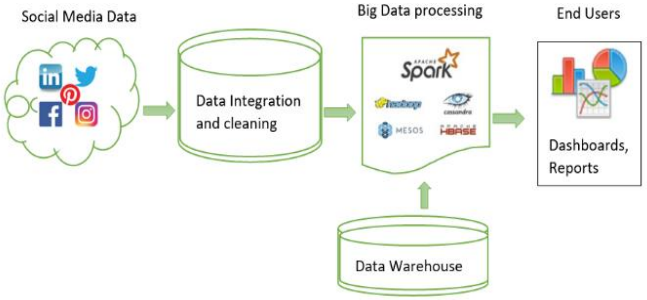

Fig. 3. Social media big data processing.

The social media businesses want to adapt real-time advertising placement algorithms analytics for insight generation, and Fig. 3 is shown the model of Social Media Big Data processing.

The analyzing big data continuously in real time is not an easy task.

There are many architectural propositions for analyzing big data in real-time streaming, but the most interesting one is Lambda and Kappa architecture.

A. Lambda Architecture

Lambda Architecture, suggested in 2011 by Marz and his team [17], offer the option to calculate arbitrary functions on an arbitrary dataset in real time by breaking down the problem into three layers (batch layer, speed layer and serving layer) [18], and provides a series of architectural principles that can be integrated and treatment of flow and batch data, in a low latency and a single Big Data architecture.

The Lambda Architecture consists of three main layers that interact with incoming data: the batch layer, serving layer, and speed layer. Fig. 4 illustrates the core technologies, components, processes, and responsibilities that constitute each layer of the Lambda Architecture.

The batch layer contains ever-increasing master dataset stored and pre-computes batch views on a distributed filesystem (such as HDFS) and produce batch views. MapReduce is used for processing the batch data, which is the Hadoop programming model. The serving layer indexes the batch views and does not require random writes, but must random reads. To implement the serving layer, usually technologies such as HBase, Storm and Cassandra are used. The last layer is the speed layer which only handles new data and uses an incremental model whereby the real-time views are incremented. At this layer stream processing system is generally based on Strom technology.

B. Kappa Architecture

The Kappa architecture, which was first described by Kreps, is perfectly suitable for processing data flows [19]. In this architecture, the key thought is to use one simple layer in real time for processing data flow and batch processing [16].

In the kappa architecture, everything is a stream [20], and you need is a stream processing engine data. So, what we would traditionally call batch processing of data is simply streaming through bounded datasets. However, Kappa is only focused and offers the option of creating a streaming and data batch treatment system based on the same technology. Similarly, queries search for a single location of the Kappa architecture instead of two of the Lambda architecture.

As shown in Fig. 5, the Kappa architecture is composed of the streaming layer module, responsible for orderly data events and one Serving layer, manages query results.

Summarizing, the kappa architecture focuses on the processing of data streams more than storage and this architecture is better adapted for cases where there is no need to permanently store the data.

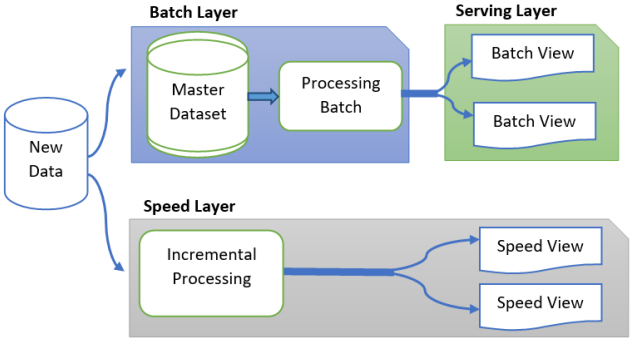

Fig. 4. Lambda architecture.

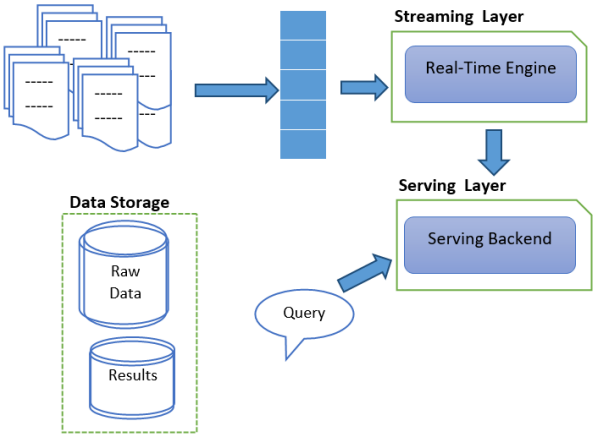

Fig. 5. Kappa architecture.

In the present paper, we have reviewed studies on the processing of big data mining on social media and focused on the concept of analyzing, we present our analysis processing, and we suggest a new method with new measuring process based on important aspects, such as number of likes, comments and retweets on the Twitter platform that can be used to predict the tweet texts (useful or not). Finally, we use our data mining algorithm of the tweet texts is used to calculate the LCR score for tweet text.

IV. Our Proposed Big Data Analysis Architecture for Understanding Human Interaction

Definition of the environment in which Big Data is processed continuously in real time, regardless of the type of data is no easy feat. There are much architectural proposals for real-time big data analysis, but the most interesting thing about our problem is developing architecture for extracting high-quality information about relevant messages [21]. In this section, we are presenting motivation to offer the architecture that best suits our use case and how it works.

The wide-spread popularity of social media has grown considerably the ratio of various streamed data, no matter what type of information or data (structured, semi-structured or non-structured). This information comes from regular people and this data may be in any of the next two statuses: data Useful (in motion) and not Useful (at rest) [22]. Therefore, various data quality distributions of generated by people using social networks are naturally fuzzy and non-structured [23], range from high-grade to low-grade. All this information can include the personal opinion of social media users [24], behaviors and thoughts, which makes it more and more important to extract useful information. Due to the availability of user-generated content can encapsulate helpful and high quality. Such increased in data generation has brought attention to the needs of analyzing big data in real time [25].

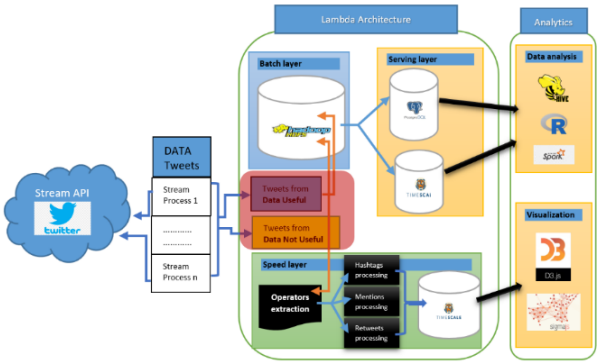

Fig. 6. Stream analytics system architecture for analyzing twitter text data.

As shown in Fig. 6, the practice example demonstrates the analysis process in Hadoop for analysis Twitter text data.

A researcher can retrieve data stored on Twitter through the public API provided by the social networking service Twitter to process specific information. However, in many cases, they do not provide data required by researchers. For example, may be needed for some additional data or perform cleaning and filtering operations to obtain useful data, such as the number of Retweets of user content or their popularity or reputation. Therefore, collecting useful data is a set of skills and methods and discipline of users in order to capture all the big data without any restrictions.

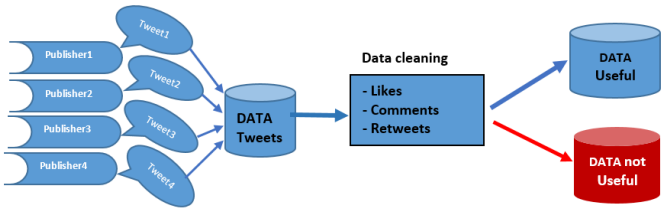

Fig. 7. Our proposed analysis procedures model.

As shown in Fig. 7, the Our Proposed Analysis Procedures to select the most relevant and interesting Tweets by cleaning and filtering using factors based on overall number of likes, overall number of comments and overall number of tweets. However, we need a more advanced method that can detect useful tweets.

LCR score, which is shown in equation (1), is our method used to measure the performance of the model.

$$f(L, C, R) = C1 \times Likes\_count + C2 * Comments\_count + C3 * Retweets\_count \quad (1)$$

Where f(L, C, R) is LCR score, C1, C2 and C3 is coefficients parameters and Likes_count, Comments_count and Retweets_count is the number of (Likes, Comments, Retweets).

The measuring process is shown in equation (1) that combines likes, comments, and retweets (LCR Score) for a tweet. And we rank those tweets based on the LCR score for each one. It should be noted that news tweets and it contains important information generally have a high number of retweets and that this type of score will be useful for our analysis of the information. We therefore assigned a high C3 coefficient for retweets to this type of content and a relatively low C2 coefficient for comments and a very low C1 coefficient for likes when calculating the LCR score.

In Table I, the three factors of tweet texts are shown, and we note that the Tweet texts not Useful and is not taken into account since the LCR Score is equal to zero.

TABLE I. Three Factors of Tweet Texts

Retweets_ count	Comment_cou nt	Likes_cou nt	LCR Score	Tweet texts
R > 0	C >= 0	L >=0	LCR> 0	Useful tweet
R= 0	C= 0	L= 0	LCR= 0	Not Useful tweet

Data mining of the tweet texts algorithm is used to calculate the LCR score for tweet text, based on the overall number of likes L, the overall number of comments C and the overall number retweets R for a tweet texts in D (D represents the raw data). The algorithm LCR begins with an initial score equal to zero. Thus, that the Tweet text Not Useful and is not taken into account. then, the LCR score is computed step by step for a tweet text, and sequential patterns that combines likes count, comments count, and retweets count— is going to use these operators to find the raw data tweet useful.

Algorithm 1: Data mining of the tweet texts

Input:

D (The raw of dataset)
L (The likes count)
C (The comments count)
R (The retweets count)
LCR (Score of tweet text)
Output:
LCR (Score of tweet text)
Initialize Tweet_text = Scan(D)
  If Tweet_text contains content
    L= Likes_count
    C = Comments_count
    R = Retweets_count
    LCR = Score(L,C,R)
  End

V. Results and Discussion

Big data in social networks occupy a big place in our lives, and during the COVID-19 pandemic, social media have become essential to be in contact with our loved ones.

The tweets data analyzed in Table II are pre-defined data (not in real-time). We use this tweets text to test implementation of the LCR score, in order to detect useful and to classification tweets by interested degree.

TABLE II. Results Values Calculated by LCR Score For Pre-stored Tweets Texts

Tweet text	Retweets count	Comments count	Likes count	LCR Score
The EU says it may have a digital identity portfolio by 24, regardless of the challenges	106	11	89	429
WHO: COVID-19 - we all need to remain vigilant.	142	333	270	1362
Swedish company DSruptive Subdermal that specializes in microchip implants has created a way to store COVID-19 passport data. Here's how it works.	340	70	265	1425
Andrew Tate nails the covid narrative in 90 Seconds!	1013	53	2292	5437
UNICEF vaccine ad	1245	547	1345	6174

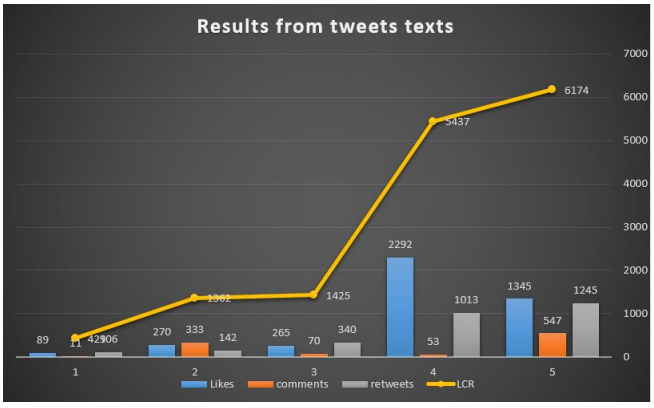

Fig. 8. Procedures a part of result from tweets texts.

The results in Fig. 8 show that an important Tweets stimulate the users to retweet, where the important information generally have a high number of retweets. The high precision of the LCR Score manifests in the form of a large number of retweets, which brings useful information. Furthermore, the results indicate that the number of likes and the user's ability to comment on the tweet are relatively low factors in detecting useful tweets. The experiments carried out allow us to know the most useful tweet from which the important information is disseminated and retweeted in a major time. To conclude, the more Retweets we have, the stronger to detect useful tweets are and the more significant the calculated LCR score is.

The work presented in this document represents our efforts towards creating a truly representative and comprehensive Big Data benchmark suite, using a high volume of data collected on Twitter by viral tweets related to COVID-19. We suggest a new approach with new measures that can be used for prediction the tweet texts (useful or not), then we make several interesting observations throughout diffusion graph and an analysis in detail using factors based on number of likes, number of comments and number of tweets on the Twitter platform.

We have compared of the two most known architectures for analyzing streamed big data in Section III. We're talking about the criteria here of choice of our architecture for analyzing Twitter text data and how well it meets our problem's requirements. The proposes of this work is about building a new Big Data architecture and advanced method capable to analysis the tweets texts and detecting useful tweets. By two solutions main challenges for analysis of social media data. First, how to detect not useful and irrelevant messages from big data analysis and second, categorization of this informative tweet into different degree of interest. By utilizing data from past detectoing, we show the performance of LCR score algorithm for multi-class classification factors. We observe the tweet texts to contains useful information always help improve the classification accuracy.

To evaluate the performance, the proposed method and obtained results show that the method achieves good results to detect useful and to classification tweets by interested degree.

VI. Conclusion

In this article, we reviewed studies on the procesures of big data mining on social media and focused on the concept of analysing. By the main stream processing and our proposed big data analytics architecture for understanding human interaction, the LCR measuring process method provides a framework for such research and is summarized in three sections: tweet texts input, tweet texts analysis (LCR score), and tweet output (tweet containing useful information or not useful). From a big data analytical framework perspective and our method, the discussions are focused on the analysis-oriented, and results-oriented of tweet texts information. It has to do with the perspective of data mining, this document gives a technologies of data mining algorithm of the tweet texts information and classification.

In the future phase of research, we're going to be implementing additional elements of our Architecture for

analyzing Twitter texts data in order to analyze the batch data and real-time data. We shall test the speed layer with social networking platforms, by combining technologies such as Hadoop, Storm and Kafka.

REFERENCES

[1] Fu, Weina, Shuai Liu, and Gautam Srivastava. "Optimization of big data scheduling in social networks." Entropy 21.9 (2019): 902.

[2] Kay, Samantha, Rory Mulcahy, and Joy Parkinson. "When less is more: the impact of macro and micro social media influencers' disclosure." Journal of Marketing Management 36.3-4 (2020): 248-278.

[3] Sivarajah, Uthayasankar, et al. "Critical analysis of Big Data challenges and analytical methods." Journal of business research 70 (2017): 263-286.

[4] Loebbecke, Claudia, and Arnold Picot. "Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda." The Journal of Strategic Information Systems 24.3 (2015): 149-157.

[5] Nirmal, V. Jude, and DI George Amalarethinam. "Parallel implementation of big data pre-processing algorithms for sentiment analysis of social networking data." International journal of fuzzy mathematical archive 6.2 (2015): 149-159.

[6] FELT, Mylynn. Social media and the social sciences: How researchers employ Big Data analytics. Big Data & Society, 2016, vol. 3, no 1, p. 2053951716645828.

[7] GHANI, Norjihan Abdul, HAMID, Suraya, HASHEM, Ibrahim Abaker Targio, et al. Social media big data analytics: A survey. Computers in Human Behavior, 2019, vol. 101, p. 417-428.

[8] Bello-Orgaz, Gema, Jason J. Jung, and David Camacho. "Social big data: Recent achievements and new challenges." Information Fusion 28 (2016): 45-59.

[9] IMMONEN, Anne, PÄÄKKÖNEN, Pekka, et OVASKA, Eila. Evaluating the quality of social media data in big data architecture. Ieee Access, 2015, vol. 3, p. 2028-2043.

[10] KUMARI, Savita. Impact of big data and social media on society. Global Journal for Research Analysis, 2016, vol. 5, p. 437-438.

[11] Nguyen, Dat Tien, et al. "Applications of online deep learning for crisis response using social media information." arXiv preprint arXiv:1610.01030 (2016).

[12] Laney, Doug. "3D data management: Controlling data volume, velocity and variety." META group research note 6.70 (2001): 1.

[13] Hashem, Ibrahim Abaker Targio, et al. "The rise of "big data" on cloud computing: Review and open research issues." Information systems 47 (2015): 98-115.

[14] Ghani, Norjihan Abdul, et al. "Social media big data analytics: A survey." Computers in Human Behavior 101 (2019): 417-428.

[15] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International journal of information management 35.2 (2015): 137-144.

[16] Persico, V., Pescapé, A., Picariello, A., & Sperlí, G. (2018). Benchmarking big data architectures for social networks data processing using public cloud platforms. Future Generation Computer Systems, 89, 98-109.

[17] N. Marz and J. Warren, Big Data: Principles and best practices of scalable real-time data systems. New York; Manning Publications Co., 2015.

[18] Hasani, Z., Kon-Popovska, M., & Velinov, G. (2014). Lambda architecture for real time big data analytic. ICT Innovations, 133-143.

[19] J. Kreps, "Questioning the lambda architecture," Online Artic. July, p. 205, 2014.

[20] Lin, Jimmy. "The lambda and the kappa." IEEE Internet Computing 21.05 (2017): 60-66.

[21] El Marrakchi, M., Bensaid, H., & Bellafkih, M. (2017). E-reputation prediction model in online social networks. International Journal of Intelligent Systems and Applications, 9(11), 17.

[22] Arnaboldi, Valerio, et al. "Online social networks and information diffusion: The role of ego networks." Online Social Networks and Media 1 (2017): 44-55

[23] S. Samanta, V.K. Dubey and B. Sarkar, Measure of influences in social networks. Appl. Soft Comput. 99 (2021) 106858.

[24] Essaidi, Abdessamad, Dounia Zaidouni, and Mostafa Bellafkih. "New method to measure the influence of Twitter users." 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS). IEEE, 2020.

[25] Jesmeen, M. Z. H., et al. "A survey on cleaning dirty data using machine learning paradigm for big data analytics." Indonesian Journal of Electrical Engineering and Computer Science 10.3 (2018): 1234-1243.