IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. V (Sep. - Oct. 2016), PP 27-31
www.iosrjournals.org

Big Data on Content Credibility of Social Networking Sites and Instant Messaging Applications

[1]Anurag Kumar, [2]Monica Mishra, [3]Rinkita Mittal
[1]*(Department of Computer Science and Engineering, Sir MVIT, VTU, India)*
[2]*(Department of Computer Science and Engineering, UVCE, BU, India)*
[3]*(Department of Computer Science and Engineering, Sir MVIT, VTU, India)*

Abstract: *The quantity of unstructured and structured data from social networking platforms and mobile phone instant messaging applications is massive and is produced at an exponential rate yet there is no mechanism to verify the content's truthfulness and trustworthiness. In this paper we have proposed a theory of how Big Data technology can be employed to validate the credibility of the vastly diffused data. Using technologies like Hadoop to analytically process the fast paced incoming data and measure the reliability of the content. To verify the integrity, the processed data must be inspected against entrusted sources. These sources must be accessible and should have trustworthy data that can assist in measuring the authenticity of contents from various sources. While privacy concerns are often dismissed when data is scraped from public-facing platforms such as Facebook, the need for these sites to validate the data posted on their site becomes prudent. Posting false rumors devalues the extent to which social networking acts as an effective method of spreading true information. In this paper we provide a brief exploration on Big Data, Hadoop and influence of unsolicited messages propagated from social networking websites and instant messaging applications.*

Keywords: *Big Data Definition, Content Credibility, Hadoop, Instant Messaging Applications, Social Networking Sites*

I. Introduction

Social networking websites is an online platform that enables its user to interact with each other using requite User Generated Content such as text posts, digital photos and videos. Instant messaging applications are real time text transmission internet services for fast and effective communication. According to Global social media research as of July 2015, total worldwide population is 7.3 billion and the internet has 3.17 billion users, in that there are 2.3 billion active social media users. Another astonishing fact is 1 million new active mobile social users are added every day that's 12 each second. Facebook Messenger and WhatsApp handle 60 billion messages a day. These facts can help to understand the magnitude of the data flow which occurs between Social media platforms and mobile phone IM (Instant Messaging) applications. The data exchanged could be textual or multimedia, produced at an extensive rate that may not be accounted or validated. Though the SNS (Social Networking Sites) and IM help people to communicate and share information, the content's integrity remains uncertain. The truthfulness of data is unresolved until it is exclusively inspected by the content viewer.

BIG DATA The term Big Data is often used to denote a storage system where different types of data in different formats can be stored for analysis and driving business decisions. Big Data is an assortment of such a huge and complex data that it becomes very tedious to capture, store, process, retrieve and analyze it with the help of traditional RDBMS databases or traditional data processing techniques [1].
Clearly, size is the first characteristic that comes to mind while considering 'What Is Big Data?' However, there are other characteristics that have emerged recently:
1. Volume: Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.
2. Velocity: Data streams in at an unprecedented speed and must be dealt with in a timely manner. For an example, an organization may need to analyze 2 million records each day to identify the reason for losses. Companies like Facebook and Google analyze much bigger data sets every day for their data processing needs.
3. Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio and financial transactions.

We also consider two additional dimensions when it comes to Big Data:
Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage.

Complexity: Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

[2] **HADOOP** is a framework that has the ability to store and analyze data present in different machines at different locations very quickly and in a very cost effective manner. It uses the concept of Map Reduce which enables it to divide the query into small parts and process them in parallel. Because of its power of distributed processing, Hadoop can handle large volumes of structured and unstructured data more efficiently than the traditional enterprise data warehouse. Hadoop is open source and therefore, it can run on commodity hardware. That means the initial cost savings are dramatic with Hadoop while it can continue to grow as your organizational data grows.

Hadoop has two main sub features:

1. Hadoop Distributed File System (HDFS) [3] is a file system that provides reliable data storage and access across all the nodes in a Hadoop cluster. It links together the file systems on many local nodes to create a single file system. Data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout various nodes in the cluster. This way, the map and reduce functions can be executed on smaller subsets of your larger data sets, and this provides the scalability that is needed for Big Data processing. This powerful feature is made possible through the HDFS of Hadoop.

2. Map Reduce is a programming framework of Hadoop suitable for writing applications that process large amounts of structured and unstructured data in parallel across a cluster of thousands of machines, in a reliable, fault-tolerant manner. Map Reduce is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The Map Reduce concept is fairly simple to understand for those who are familiar with clustered scale-out data processing solutions.

YARN (Yet Another Resource Negotiator) is a cluster management technology. It is one of the key features in second-generation Hadoop. It is the next-generation Map Reduce, which assigns CPU, memory and storage to applications running on a Hadoop cluster. It enables application frameworks other than Map Reduce to run on Hadoop, opening up a wealth of possibilities. Part of the core Hadoop project, YARN is the architectural center of Hadoop that allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform.

II. Literature Review

[4] The increasing use of social networks and instant messaging applications, such as Facebook, Twitter, WhatsApp, Line and many other has produced and is producing huge volume of data in the form of posts and text messages. Several organizations are interested in discovering new business insight to increase business performance. By using advanced analytics, enterprises can analyze big data to learn about social leaders who influence the behavior of others in the network, and on the other hand, to determine which people are most affected by other network participants. There has been study on modeling the knowledge diffusion in social networks on a directed, scale-free network environment. Also, the content of big data is known to possess different characteristics, which affect its quality. [5] Another advantage of Big Data in the field of social networking platform is the ability to extract from the mass of incoming information what is important for situational awareness during mass emergencies. Researchers have presented a model which describes four novel approaches using focused twitter crawling, trustworthiness analysis, geo-parsing, and multilingual tweet classification in the context of how they could be used for monitoring crises. [6] Big data can have ambiguities and inaccuracies which needs to be identified and accounted for to reduce inference errors and improve the accuracy of generated insights from a business perspective. The veracity or accuracy of data is identified from a veracity index which provides a useful way of assessing systematic variations in big data quality across datasets with textual information. In this paper we discuss about how the reliability of the content posted in social media plays a crucial role that impacts the lives of people and how to approach this scenario.

III. Problem Statement

The need for validation of the content of social media and instant messages becomes imperative when these posts start to influence the rational thinking of individual. The instant messaging applications are widely used for swift communication and at times, the messages in these applications are forwarded without any content verification. There are websites available to assist people to use internet wisely by keeping up-to-date articles on Internet scams, hoaxes and rumors that circulate the Internet. For instance www.ThatsNonsense.com and www.ThatsFake.com serves the internet user to carefully share information but there is no assurance that uncertain content doesn't reach a larger pool of audience.With the ability to communicate with people so quickly and efficiently, regardless of the distance spreading misinformation, half-truths and hearsay has rather become easier. The problem with this is that whilst most rumors may appear harmless – and can be passed along "just in case" they're true – they often have unseen or unpredictable consequences, not only for the sender, but for others, or even the social media community in general. Law-enforcement agencies are increasingly facing law and order problems triggered by sensitive or controversial content on the instant messaging application. Thus, the need to have data inspected for its dependability before it hits a wider range of public is desirable. We present an approach where technologies like Big Data can be used to manage the structured and unstructured data emerging from various sources [7].

IV. Proposed Solution

[8] When data from various social networking applications is available for reference, the content must be verified against a trusted authority. The entrusted authority could be websites issued by government or any other official media forum, which could ensure credibility of the content. Here we go forward with an assumption that these valid sources have enough information about the recent events. The key authority is not absolute, it is subjected to vary based on the assurance credibility level of these sites which comprise the verified source. The following illustration Fig 1 displays the verified source (trusted authority) as a repository which accumulates data from diverse websites and web pages.

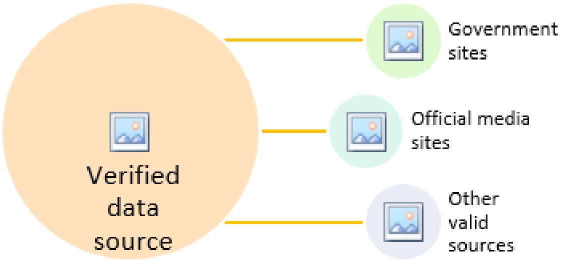

Fig 1

There is a great amount of data generated from social networking platforms every minute for instance, according to a business intelligence company DOMO the leading social networking site is Facebook as of year 2015, with over 1.4 billion active monthly users. These users generates the most amount of social data over 4 million posts every minute which adds up to 250 million posts per hour. Instagram, with 300 million monthly users, comes in second with over 100 million photos uploaded per hour. The number of IM applications user's data exchange is even astounding, WhatsApp users barter around 42 billion text messages per day. The next question would be how do we recognize which content needs to be verified and how to examine it? The most proficient platform to achieve this is, Big Data analytics which is a process of collecting, organizing and analyzing large sets of data (Big Data) to discover useful information. Big Data uses technologies like Hadoop. As explained earlier, Hadoop is the core platform for structuring Big Data. It also solves the problem of formatting it for analytic purposes. Hadoop uses a distributed computing architecture consisting of many servers using commodity hardware. This in turn makes it inexpensive to scale and support massive data stores for analyzing it based upon the algorithms used. The Fig 2 displays the content from SNS to be processed and analyzed by Big Data platform.

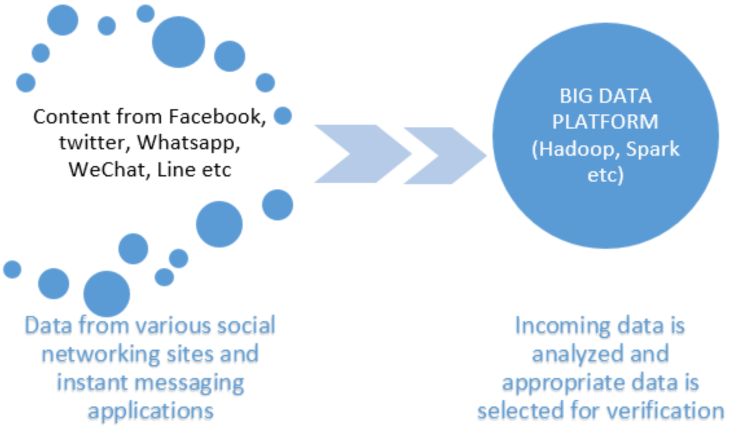
Content from Facebook, twitter, Whatsapp, WeChat, Line etc
BIG DATA PLATFORM (Hadoop, Spark etc)
Data from various social networking sites and instant messaging applications
Incoming data is analyzed and appropriate data is selected for verification

Fig 2

The content from verified source serves as a reference to measure the trustworthiness of data gathered from networking sites and IM applications. Big Data is employed to perform this analysis and validation at a cost effective rate given the fast pace of incoming data. Fig 3 shows the advantage of using Big Data as it enables processing of data from any source and the result is delivered to sites requesting for the integrity check of their content.

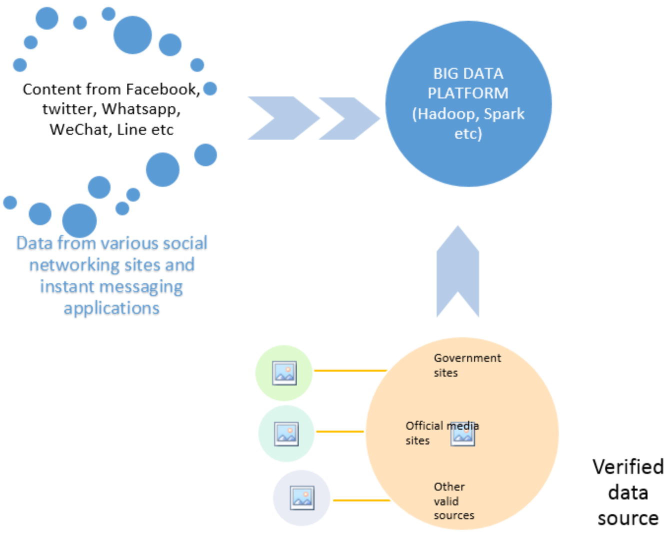
Content from Facebook, twitter, Whatsapp, WeChat, Line etc
Data from various social networking sites and instant messaging applications
BIG DATA PLATFORM (Hadoop, Spark etc)
Government sites
Official media sites
Other valid sources
Verified data source

Fig 3

The flowchart for this approach is illustrated in the Fig 4.

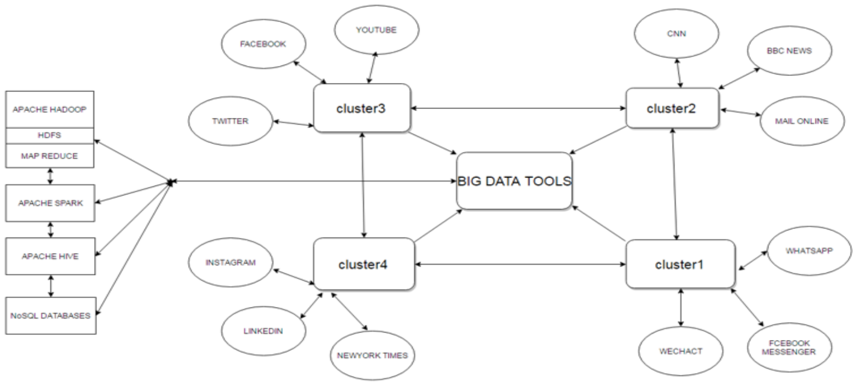

Fig 4

V. Challenges

1. The economical view of data denotes that the value of data does not grow as the magnitude of its volume increases. Although this view holds good when we pursue the analytics for value of data but it does not account for the fact that increasing volume and variety creates opportunities to extract added value.
2. Data Integration - The ability to combine data that is not similar in structure or source and to do so quickly and at reasonable cost.
3. Data volume – The ability to process the volume at an acceptable speed so that the information is available to organization when they require.
4. Solution cost – It is crucial to reduce the cost of the solutions as it is never profitable to have a higher solution cost than the problem at hand.
5. Complexity - Between social media and traditional internet information outlets, keeping a track of sources becomes cumbersome.
6. Data access – Data ingress and connectivity can be an obstacle. A majority of data points are not yet connected today and companies may not have the right platforms to aggregate and manage the data across the enterprise.

Fig 5 exhibits other challenges in the form of a bar chart in the field of Big Data processing.

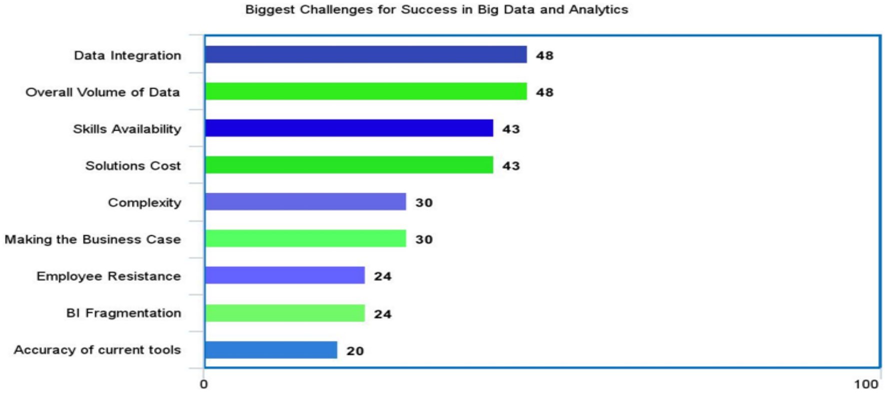

Fig 5

VI. Conclusion

This paper is a proposition to empower the users of social networking sites and various instant messaging applications to avail the content with reliability. This analytics is achieved with the help of Big Data which is low-cost commodity hardware that has expanded exponentially in the last few years and is expected to grow more. The convergence of these trends in Big Data management means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively for the first time in history. The end user would develop deeper conviction on the credibility of data which is profitable for all the parties involved.

References

[1] Meira, "*Exploring Big Data in Social Networks*" INWEB – National Science and Technology Institute for Web Federal University of Minas Gerais – UFMG May 2013

[2] R.D.Schneider, "*Hadoop for Dummies*", John Wiley, Mississauga, 2012

[3] K, Chitharanjan, and Kala Karun A. *"A review on Hadoop — HDFS infrastructure extensions."* JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013

[4] Lukoianova T., & Rubin, V. "*Veracity Roadmap: Is Big Data Objective, Truthful and Credible?*" Advances In Classification Research Online, 2014

[5] O. Liu, K.L. Man, W. Chong, and C.O. Chan. "*Social Network Analysis Using Big Data*". *Proceedings of the International MultiConference of Engineers and Computer Scientists 2016 Vol II*, IMECS 2016, March 16-18,2016

[6] Zielinski, "*Social Media Text Mining and Network Analysis for Decision Support in Natural Crisis Management*" Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May 2013

[7] A. Nagy and J. Stamberger, *"Crowd sentiment detection during disasters and crises"*, Proceedings of the 9th International ISCRAM Conference, Vancouver, Canada, 2012, April 22-25

[8] F. H. Khan, S. Bashir and U. Qamar, *"TOM: Twitter opinion mining framework using hybrid classification scheme"*, Decision support systems. *vol. 57, 2014, pp. 245 – 257*

[9] L. de Vries, S. Gensler and P.S.H, Leeflang, *"Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing",* Journal of interactive marketing, *vol.26, no. 2, 2012, pp. 83-91*