Big Data Spatial Analytics in Social Networks using Hadoop

Sultan Alenezi
College of Computing & IT,
Arab Academy for Science,
Technology and Maritime Transport,
Alexandria, Egypt.

Saleh Mesbah
College of Computing & IT,
Arab Academy for Science,
Technology and MaritimeTransport,
Alexandria, Egypt

ABSTRACT
A lot of applications in different fields generate huge data streams, known as Big Data. These kinds of applications needs special systems of data analytics in order to collect, store and process these big data files, like Hadoop framework. This paper aims to analyze social media big data to identify the widespread of certain keywords. A Java-Hadoop application is developed to analyze data obtained from Twitter social network. The application is used to identify number of people (Tweets) who mentioned specific medical keywords (e.g. Cancer) classified by location. The application is developed using Java Eclipse on CentOS Linux operating System and runs on Oracle Virtual Machine. The analysis aims to help in decision making according to the number of people tweeting about cancer or any related word (like tumor) and analyze them according to their cities all over the world. The results are used to create a GIS layer to spatially enhance the visualization of the obtained results.

General Terms
Big Data Analytics, Hadoop

Keywords
Big Data Analytics, Twitter, Hadoop, Social Network, Cancer, CentOS.

1. INTRODUCTION
Big data is created from accelerating plurality of sources, including Internet clicks, mobile transactions, user-generated data, especially in the social media and the content that is generated through sensor networks or business transactions such as bank transactions and market demand. Moreover, genomics, health care, engineering, operations management, and commerce; all add more to big data pervasiveness. These data require the use of powerful computational applications in order to view the trends within or in between these extremely large socioeconomic datasets. New insights created from this data-information can be very beneficial as added value in official statistics, or surveys source that are found in large vast amounts, though this will add more clearer information from collective experiences and doing so in real time, thereby narrowing both information and time gaps [1].

Recently, the technological improvements have introduced a very huge amount of data from distinctive domains. This includes social media, health care, information systems, and anonymous user generated data as in the internet, and industries systems [2]. The term "Big Data" was created to cover the meaning of this kind of data, as in Twitter. The huge volume of big data in Twitter or in general had other unique characteristics compared to the traditional data. Big data can be structured or unstructured or semi-structured and thus requires a very sophisticated techniques for analysis. This new advances in technology requires the creation of new systems and new architectures to enable data collection, storage, and processing algorithms mechanisms, which happens to be in the Hadoop software framework. Twitter is an online social networking space that is very popular since 2006, where the registered persons do share or post messages less than or equal to 140 characters which is called tweets. It is a social network for people from all over the world to communicate and stay connected by tweeting or exchanging of fast, continuous, and frequent messages. People on Twitter share information, opinions, local and international news with another people who are called followers and seek knowledge and expertise in these public tweets. Furthermore, the Tweets have been created on day to day basis of conversations, news, comments on some events, movies, politics, life, catastrophic events, wars, diseases. Eric Schmidt who is Google CEO stated that "There was 5 Exabyte of information created between the dawns of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing. People aren't ready for the technology revolution that's going to happen to them." All these accumulated tweets results in creation of many statements of huge volume of data, if this kind of data is analyzed in an intelligent way, it can be of big value, as it would help us in a variety of useful information that will be used to make decisions [2].

Through the era of continuing increase of the social media which is creating heavy quantities of new digital info about appropriates, organizations, also institutions that is instantly usually labeled Big Social Data. Social network analytics is a designate shared to relate to the collection, storage, analysis, furthermore reporting of these new data. These social data sets behave useful info if analyzed using proper routines, procedures [3].This paper aims to present a Java-Hadoop application developed for big data analytics. The definition of the big data is first introduced followed by a discussion of some attributes regarding big data analytics. A systematic framework is presented to extract important information from big dataset using Hadoop Java Eclipse application on an Oracle CentOS.

The rest of the paper is organized as the follows: Section (II) presents the literature review related to the big data analysis past researchers. Section (III) contains the proposed framework, while the results, and the discussion are in the subject of Section (IV). Finally, section (V) concludes the paper and directions for further work.

2. LITERATURE REVIEW
Huddar and Ramannavar (2013) presented different analytical tools comparison for the big data. First, they argued about the definition of the big data, it appeared in year 2005 by Roger Magoulas. The term big data is used to describe large size of data that is not organized and used worldwide, cannot be saved or query using the ordinary database techniques. Big data might be a combination of many different languages and

structures of data (Unstructured NoSQL). The main characteristics of big data can be expressed as: Volume, Velocity, Variety and Veracity. The "Volume" is considered the first and most notorious feature.

In 2000, the data that were stored is about 800,000 petabytes in all over the world. This number is expected to reach 35 zeta bytes by 2020. Twitter and Facebook generate around 7 TB and 10 TB of data every day respectively. Companies might get use of the big data analytics tools in order to understand and build statistics on these big data. "Variety" is referred to different types of data. "Velocity" is referred to how can the researchers can get the data and store it. The fourth vector of big data is the "Veracity" which is referred to how much the data acquired is trusted for the decision makers to make decision upon its statistics. Users from different social media sites, or micro blogging, or blogs create content like blog posts, tweets, photos. Servers continuously log messages about what these users are posting. Fine detailed procedures and measurements done for companies' record information about sales, suppliers, operations, customers, etc. [4].

More than 4 billion persons (60% of the world's population) in 2010 use the mobile phones, and about 12% of those people had smartphones, whose generation is growing more than 20% a year. More than 30 million networked sensor nodes are now in effective work all over the transportation, and automobiles, and especially in industries. The number of these sensors is increasing at a rate of 30% a year. The ultimate source of data is the Internet therefore incomprehensibly is large. A study showed that every year the amount of data will grow by 40 percent to 2020. Big data represents large sets of data, but the notion of volume is not the only one to consider. A high velocity and variety are generally also associated with big data. Velocity of the big data is much more concerned with how fast the data is gathered.

The most widely used tool for processing large data sets is Hadoop and more precisely the MapReduce framework and using Hive Hadoop. The platform Hadoop processes data by batch, so when the process started, new incoming data will not be taking into account in this batch but only in the next batch. Meanwhile, the organizations and companies deal with different data types and sources and are not dealing with only their own data. Moreover, this happens only to make sure that perfectly understand third parties outside of the main company or organization generate the market. As a conclusion, the data comes from many different sources in various types (variation in type and volume). It could be text from social network, image data, geo location, logs, and sensors data and so on [5].

Users increasingly count on company-sourced info, such as criticisms on Yap and Amazon, liked posts furthermore ads on Facebook. This has led to a mart for black hat ballyhoo styles way false (e.g., Sybil) besides compromised accounts, furthermore conspiracy networks. Extant approaches to notice such bearing relies mostly on supervised (or semi-supervised) culture atop notorious (or hypothesized) assaults. They are powerless to discern assails missed by the agent though labeling, or whereas the attacker modulates policy. Such information considered big or huge data [6].

The Twitter microblogging site is increasingly shared to explore latest info on various matters of United's stake. Numerous personalized inquire besides recommender computers developed to contribute Twitter users find content that is of intrigue to them. These computers wear an assortment of fashions ranging from established collaborative filtering (or variants such as co-factorization robots) to extra new sociable advices, where the components to be suggested to a user u are worn from her sociable network [7].

When researchers are discussing the Term big data, they are classifying the kind of data that exceeds the overall capacity of any conventional or ordinary database. This kind of data is too huge, fast, and doesn't work into the structures of traditional database architectures. The Big data also refers to the large, complex volume of data captured from multiple sources or single source like social media site and cannot be processed using the traditional ways plus the fact that the Big Data holds massive raw data which can be explored through great efforts. Taking into consideration that the impact of the Big Data despite providing large or vast potential in competition and growth for developers, decision makers and individual companies, but also the good and right use of the Big Data can be useful in terms of performance, efficiency, innovation, and competitiveness for entire public and private sectors, economies and governments [8].

3. PROPOSED FRAMEWORK

The proposed model presented in this paper works on retrieving useful statistical information out from big unorganized data. The researchers develop Java-Hadoop application to help in identifying areas according to percentage of twitter user's tweets, as shown in the Figure (1).

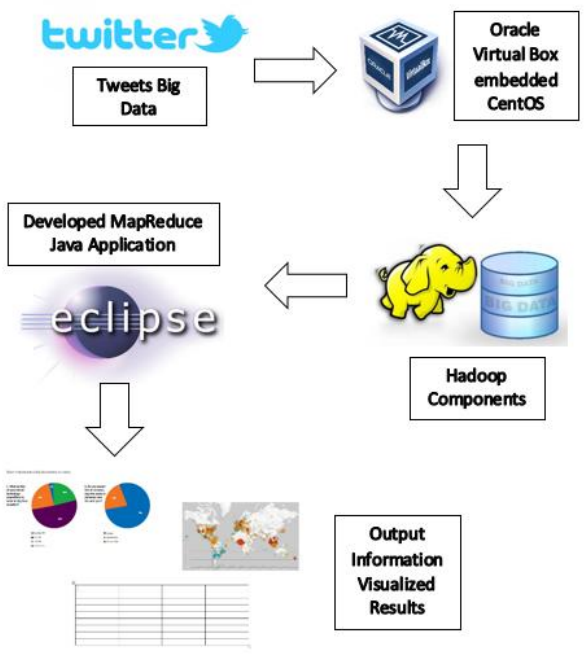

Fig. 1. Proposed Framework Model

A communication channel has been created with different research centers in order to capture big data released from twitter tweets. Then researchers have installed Oracle Virtual Machine from Cloudera Big Data Analytics into the operating system. Then researchers ran CentOS 6.4 operating system on it to hold the JDK 7, Hadoop Components 2.5.0 and Eclipse 4.2.6. In order to make sure that the Hadoop components are running perfectly, the researchers have to run some tests. To do that the researchers open an internet browser to the URL: http://localhost:8088, researchers should see the resource manager UI. The VM uses port forwarding for the common Hadoop ports, so when the VM is running, those ports on localhost will redirect to the VM. The result should be as

shown in Fig. 2 to prove that the virtual machine is working perfectly and Hadoop components are working perfectly.

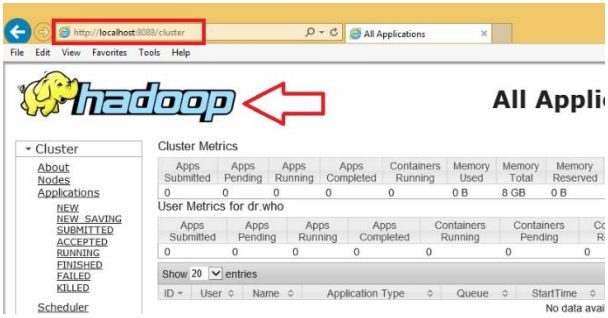

Fig. 2. Proof Hadoop working perfectly

As shown in Fig. 2, the localhost link shows that Hadoop is installed perfectly, otherwise the webpage will not display anything

3.1 Dataset Information

The dataset [9] is from a running conferences within the current 9 years and is concerned in internet and especially in social media sector. As per the researchers request to them through email they told us that it's okay to use the dataset for academic use, which is anonymous twitter dataset includes full description of tweets including the tweet, location, id, gender, DOB, followers id, and date/time values. The dataset is composed of more than 14,000,000 tweets, where to obtain these data for academic use the researchers had to sign on an agreement and they accept it and send us credentials to download the data.The researchers have extracted a sample composed of almost 1 million tweets, exactly 973,916 tweets information including the tweet statement and the location from where it was tweeted in a separate file to do the researchers big data analytics on.

3.2 Oracle Virtual Box embedded CentOs

Cloudera offers a great solution based on Oracle Virtual Box which includes CentOS Linux operating system, where the Eclipse, JDK & Hadoop Components are compatible with Java Eclipse.

3.3 Hadoop Components & Eclipse

The researchers have developed a Java application that runs with the aid of Hadoop Libraries in order to deal with the big data files, ensured that the researchers Twitter data file is read successfully into the developed application to analyze it.

The application goal was to analyze the big data file and search for the tweets that the users tweeted or retweeted that includes the word cancer. In addition to ensure that the strategy is complete the researchers included in the algorithm to capture in the same instance and at the same time all the words that means cancer as tumor, malignant … etc., as shown in the Table 1 [10], in order to include in the search results all the words that meant to be cancer.

Table 1. Cancer Synonyms

Cancer	Malignant	Bone Cancer	Tumor
Carcinoma	Prostate Cancer	Melanoma	Lung Cancer
Lymphoma	Endocrine Caner	Neoplasm	Metastasis
Neurofibroma	Teratoma	Breast Cancer	Meningioma
Testicular Cancer	Ordeal	Tribulation	Affliction
Pestilence	Cancerous	Sarcoma	Malignancy
Thyroid Cancer	Scourage	Colorectal Cancer	Fibroadenoma
Myeloma	Brain Tumor	Childhood Cancer	Melanoma

Thus as illustrated before the developed application:

- Search for Cancer word and its many synonyms
- Covers the case issues (small, capital & combinations)

The developed application is written in Java, where the researchers took the help from the Hadoop components that are built in the Eclipse or plugged in. The code is multi-threading parallel programming, where the researchers faced a huge problem trying to avoid the access of each single thread to the data. From that problem, the concept of the functional programming comes where the researchers actually passed the data between the Map/reduce methods as parameters.

Map/reduce is considered to be a special form of such data accessed and of which it is applicable in a wide range of use cases. It is organized as a map function where a method that transforms a data piece into some number of key/value pairs. Each of these elements discussed previously is sorted by their key and connect to the same node, where there is a reduce method is used in order to merge these values (of the same key) into a single result, as shown in the Fig. 3.

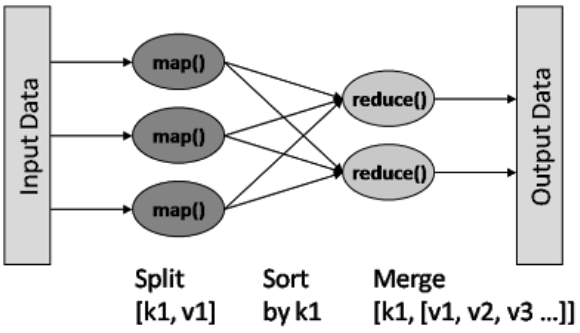

Fig. 3. Map/Reduce [11]

The researchers first step in the paper was loading the data file to be analyzed into the Hadoop components in the developed application that is built over map reduce algorithms. This happens in order to enable users of the application generate map reduce jobs using the queries of hive (SQL-like queries). Moreover, the researchers need to analyze the Big Data and create some graphical charts to provide useful insights and help in decision making process. Taking into consideration that the data are stored and accessed through the distributed file system which handles Big Data files in Gigabytes with sequential read/write operations as described in Map/Reduce procedures. Moreover, there is a master "NameNode" to keep track of overall file directory structure and storing the data chunks. The NameNode is the central control point and may be redistributed or replicated as needed. On the other hand, the DataNode reports all its chunks to the NameNode at boot up, every single chunk have got version number.

After the code is ran over the sample big data file, the researchers application outputs two files, one file is the results file which contains Tab delimited information about the cities where the users tweeted something about cancer or it's synonyms with counters, and the other file ensures that the application ran successfully. This file is then used to create charts and data into maps to visualize the results in a way convenient for the user. Also there're two log files generated collecting information about the node itself and the other about the job that occurred in the developed application.

This log file that is generated directly from the application is the one concerned with detailed information about the run operation including information about the data chunks as the map reduce operations plus the number of bytes the application read and the number of bytes the application outputs, and the other is about the Job tracker details information. Finally, the researcher's goal was to retrieve useful statistical information out from big unorganized data. The researchers developed Java-Hadoop application will help in identifying areas according to percentage of twitter users tweeting. The developed application log files were as shown in the Fig. 4

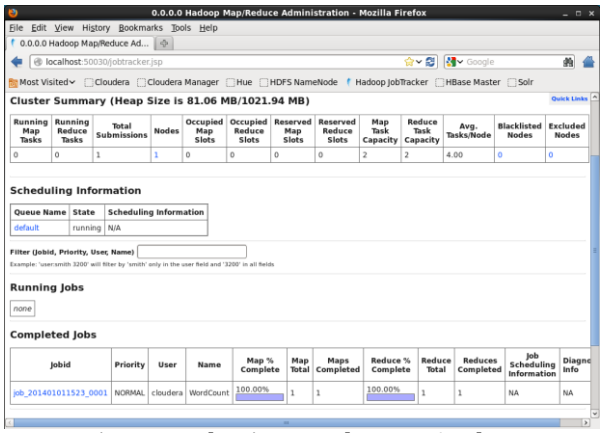

Fig. 4. Log showing complete Map/Reduce

4. RESULTS & DISCUSSIONS

Results are as shown in the Table 3

Table 3. Results

City	Counter	City	Counter
Addis Ababa	6787	Lahore	27147
Alexandria	30541	London	108588
Ankara	33934	Los Angeles	6787
Baghdad	20361	Madrid	16967
Bangkok	40722	Melbourne	6787
Beijing	27148	Mexico	20361
Berlin	6787	Moscow	30540
Cairo	40720	Mumbai	27147
California	33935	Nairobi	6787
Cape Town	6787	New York	84835
Casablanca	6787	Rio de Janeiro	13574
Chengdu	20361	Riyadh	13574
Delhi	47508	Santiago	6787
Durban	33934	Sao Paulo	27147
Guangzhou	33935	Seoul	13574
Hong Kong	27147	Shanghai	33935
Istanbul	40722	Singapore	13574
Jeddah	6787	Sydney	6787
Kabul	16967	Tokyo	13574
Karachi	13574		

Results are used to create a visualizing histogram chart and Map using Microsoft Excel to make it easier for the user to see the differences and cities where there is more users tweeting about cancer or synonyms, as shown in Fig.5, 6:

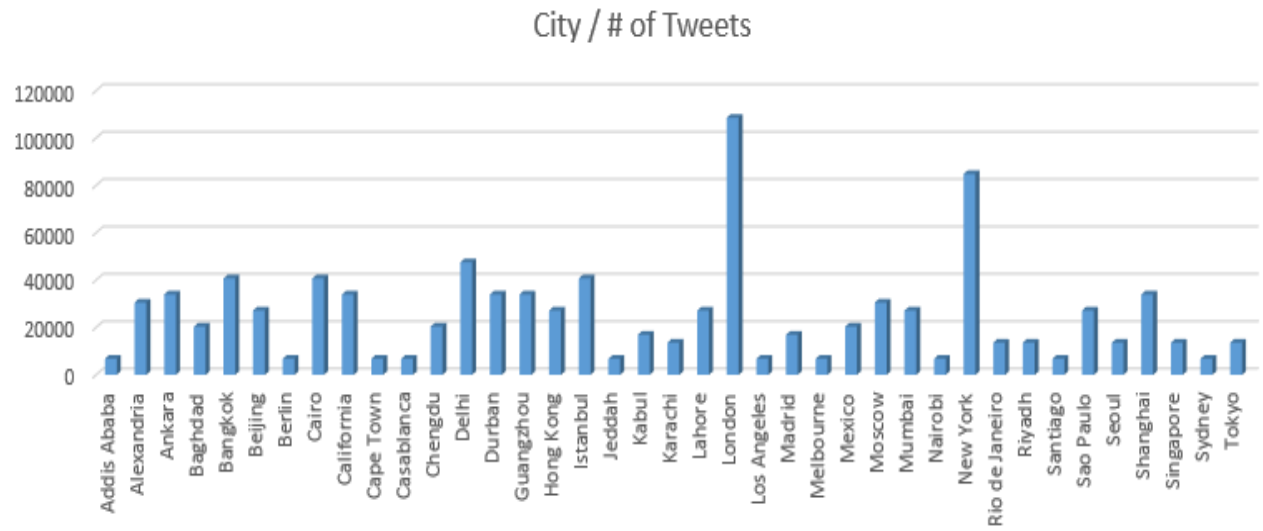

Fig. 5: Histogram of the cities against the number of Tweets

The above histogram shows the cities and the corresponding number of tweets.

Fig. 6. Map cities\tweets

As shown in the above figure on the map each of the cities and the dot represents the number of tweets. Taking into consideration that the application can work offline it does not have to be online to generate the results.

After the researchers checked the log file generated from the developed application, found the following details:

Map 100% reduce 100%

FILE: Number of bytes read=667069108

FILE: Number of bytes written=73374759

Job job local1584409728_0001 completed successfully

Map input records=1048576

Map output records=973916

Map output bytes=1189889

5. CONCLUSION

New era of big data is upon us now, its bringing with it the need for sophisticated advanced data acquisition, and analysis techniques. In this paper, the researchers have presented the concept of big data and presented the framework, which is working using java Hadoop components. The big data is either structured, semi-structured or unstructured where its cycle is composed of, are the data generation, Map/Reduce function, and data analysis. Moreover, the researchers have provided a literature review the related work and techniques in different big data phases. Firstly, the researchers have acquired rich big data source and discussed the data attributes. Secondly, the researchers have worked on the proposed framework using the Java Hadoop technology. Thirdly In the big data analysis the researchers have produced results that will be in benefit of decision makers, in other words produced useful information about Twitter tweets concerning Cancer and its synonyms. As the programming model is accompanied with data storage approach it played a role in big data analytics.

6. REFERENCES

[1] Bae,Y. and Lee, H., Sentiment Analysis of Twitter Audiences: Measuring the Positive or Negative Influence of Popular Twitterers, Journal of the American Society for Information Science and Technology, vol. 63, issue 12, December 2012, pp. 2521–2535.

[2] Lima, A.C.E.S., de Castro, L.N., Automatic Sentiment Analysis of Twitter Messages, 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), November 2012, pp. 52-57.

[3] Mukkamala, R.R.; Hussain, A. ; Vatrapu, R. , Towards a Set Theoretical Approach to Big Data Analytics, IT University of Copenhagen, 2014 IEEE International Congress on Big Data

[4] Mahesh G Huddar; Manjula M Ramannavar, A Survey on Big Data Analytical Tools, International Journal of Latest Trends in Engineering and Technology (IJLTET), Special Issue – IDEAS 2013.

[5] Chardonnens, T.; Cudre-Mauroux, P. ; Grund, M.; Perroud, B., Big Data analytics on high velocity streams, Department of Informatics University of Fribourg (Switzerland), Masters Thesis, June 2013.

[6] Bimal Viswanath; M. Ahmad Bashir; Mark Crovella; Saikat Guha, Towards Detecting Anomalous User Behavior in Online Social Networks, Boston University, 2013

[7] Parantapa Bhattacharya; Muhammad Bilal Zafar; Niloy Ganguly; Saptarshi Ghosh; Krishna P. Gummadi, Inferring User Interests in the Twitter Social Network, IIT Kharagpur, India, 2013

[8] Etpo Staus, Networked European Software and Services Initiative (NESSI) Big Data White Paper, "Big Data A New World of Opportunities", December, 2012

[9] ICWSM Academic published social media datasets

[10] http://www.icwsm.org/2015/datasets/datasets

[11] http://www.thesaurus.com/browse/dictionary, September 2015: words that have the same meaning (synonyms)

[12] http://www.ibm.com/developerworks/cloud/library/cl-openstack-deployhadoop/, Fig. 3, September 2015