Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2015) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2015 October 7-9, 2015

Health Twitter Big Bata Management with Hadoop Framework

João Cunha[1], Catarina Silva[1,2], Mário Antunes[1,3]
[1]School of Technology and Management, Polytechnic Institute of Leiria, Portugal
[2]Center for Informatics and Systems of the University of Coimbra, Portugal
[3]Center for Research in Advanced Computing Systems, INESC-TEC, University of Porto, Portugal
2121591@my.ipleiria.pt, {catarina,mario.antunes}@ipleiria.pt,

Abstract

Social media advancements and the rapid increase in volume and complexity of data generated by Internet services are becoming challenging not only technologically, but also in terms of application areas. Performance and availability of data processing are critical factors that need to be evaluated since conventional data processing mechanisms may not provide adequate support.
Apache Hadoop with Mahout is a framework to storage and process data at large-scale, including different tools to distribute processing. It has been considered an effective tool currently used by both small and large businesses and corporations, like Google and Facebook, but also public and private healthcare institutions. Given its recent emergence and the increasing complexity of the associated technological issues, a variety of holistic framework solutions have been put forward for each specific application.
In this work, we propose a generic functional architecture with Apache Hadoop framework and Mahout for handling, storing and analyzing big data that can be used in different scenarios. To demonstrate its value, we will show its features, advantages and applications on health Twitter data. We show that big health social data can generate important information, valuable both for common users and practitioners. Preliminary results of data analysis on Twitter health data using Apache Hadoop demonstrate the potential of the combination of these technologies.

Peer-review under responsibility of SciKA - Association for Promotion and Dissemination of Scientific Knowledge.

Keywords: Big Data, Apache Hadoop, Mahout, Healthcare, Twitter

1. Introduction

We live in a society where the textual data on the Internet is growing at a rapid pace and many companies are trying to use this deluge of data to extract people's views towards their products. A great source of unstructured text information is included in social networks, where it is unfeasible to manually analyze such amounts of data. There is a large number of social networks websites that enable users to contribute, modify and grade the content, as well as to express their personal opinions about specific topics. Some examples include blogs, forums, product reviews sites,

Peer-review under responsibility of SciKA - Association for Promotion and Dissemination of Scientific Knowledge.

and social networks, like Twitter (`http://twitter.com/`). These examples are presented as the future trend in mining evolving data streams [1].

The tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. To build classifiers for tweets processing we need to collect training data so that we can apply proper learning algorithms. Twitter data is used as source, for example, in sentiment analysis where the task is to classify messages into two categories depending on whether they convey positive or negative feelings, since labeling tweets manually as positive or negative is a laborious and expensive task. However, a significant advantage of Twitter data is that many tweets have author provided sentiment indicators: changing sentiment is implicit in the use of various types of emoticons. Smileys or emoticons are visual cues that are associated with emotional states [2]. When the author of a tweet uses an emoticon he is annotating his own text with an emotional state. Such annotated tweets can be used to train a sentiment classifier.

Streaming algorithms use probabilistic data structures with algorithms that may give fast approximated answers. However, sequential online algorithms are limited by the memory and bandwidth of a single machine. Achieving results faster and scaling to larger data streams usually requires the use of parallel and distributed computing. Map Reduce algorithm is currently a standard programming paradigm in this area, partly due to the popularity of Apache Hadoop, an open source implementation [2].

Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from various sources [3]: sensor networks, measurements in network monitoring and traffic management, log records or clickstreams in web exploring, manufacturing processes, call detail records, email, blogging, Twitter posts, etc. In fact, all data generated can be considered as streaming data or as a snapshot of streaming data, since it is obtained during a time interval. These data, being not structured and generated in large quantities are referenced as big data.

By definition, big data in healthcare refers to electronic large and complex health data sets, very difficult to manage with traditional algorithms [3, 4]. Big data in healthcare is overwhelming not only because of its volume, but also due to the diversity of data types and the speed at which it must be processed.

Data related with patient healthcare and well-being is making up "big data" in the healthcare industry. It includes clinical data from CPOE (Computerized Physician Order Entry) and clinical decision support systems, which includes physician's written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative information. It also includes the patient data in electronic patient records (EPRs), sensor data, live details coming from monitoring vital signs and the social media posts associated with this subject, including Twitter feeds (tweets), blogs, status updates on Facebook and other platforms and web pages [4, 5].

In this paper we propose the use of Twitter health big data as source for further processing and information retrieval. Sites like Twitter contain prevalently short comments, like status messages. Additionally many web sites allow rating the popularity of the messages, which can be related to the opinion expressed by the author [4]. One of our goals is to analyze and assign a sentiment to a tweet, i.e. whether the author expresses positive or negative opinions on a specific health topic, e.g. diabetes or dyslexia. Having such information, one can analyze the data obtained to reach results that help on the prevention and decision support for better health.

We propose a generic functional architecture with Apache Hadoop framework and Mahout for handling, storing and analyzing big data that can be used in different scenarios that involve health big data. To demonstrate its value we will analyze its features, advantages and applications on health Twitter data. This work is driven by the potential to improve the quality of healthcare delivery, while reducing costs and increasing speed of response. Big data holds the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management.

The rest of the paper is organized as follows. Section 2 presents related work and introduces the frameworks that underpin our proposal. Section 3 details the proposed architecture and Section 4 includes the implementation and evaluation. Section 5 concludes the paper and presents future lines of research.

2. Background and Related Work

Nowadays, there are many services of text analysis architectures, namely sentiment analysis and text classification. We introduce below the frameworks that support such services. When, as often occurs, text is the base of information,

various text classification algorithms are available, e.g. Naïve-Bayes or deep learning, in Mahout, `https://mahout.apache.org/`, within a distributed processing Hadoop cluster for text data from Twitter [6].

Given the unstructured data from Twitter, solutions usually require "NoSQL" management systems, which model data without using tabular relations, contrasting relational databases. A commonly used NoSQL database is MongoDB that supports text formats, and conforms to the NoSQL paradigm, i.e. non-relational, distributed horizontally scalable. MongoDB is specially designed to provide scalable and high performance data storage solutions and uses a versatile query language with syntax somewhat similar to object-oriented query languages.

Hadoop platform, `https://hadoop.apache.org/`, was designed to deal with huge amounts of data. Hadoop technology uses a divide-and-conquer methodology for processing, by handling large complex unstructured data that usually does not fit into regular relational tables. In our scenario the data source will be the public Twitter stream.

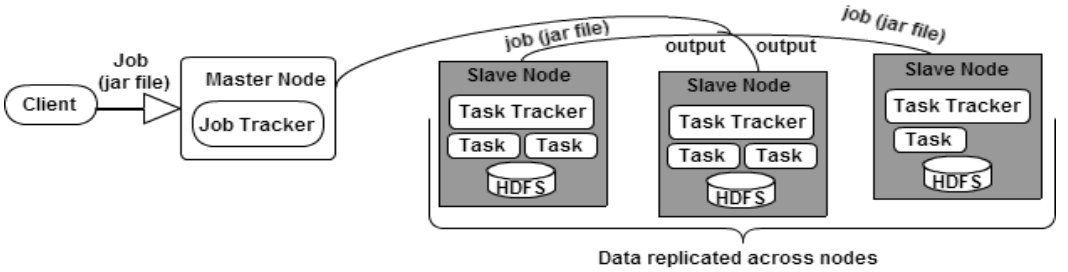

Fig. 1. Hadoop Architecture [12]

In Fig. 1 we depict the Apache Hadoop general architecture we use to process health big data. Data is stored in a distributed file system architecture (HDFS), in which each data file is spread by several nodes, connected through a high-speed network. The model has a Master Server "Master Node" that regulates the distribution of the information handled to the data Nodes "Slave Nodes". The Slave nodes are responsible for task operations in the file system (reads/writes) like block creation, deletion, data replication and data integrity check. There are several "Task Trackers" that report and aid the progress of the task to the Master node "Job Tracker", making the system tolerant to failures and reducing the data loss in task operations. Map operations, like Map Reduce, are applied in a distributed way to the data and the results are further merged and reduced [7].

In our work, we have managed the entire project using the HDFS, which was designed to deal with very large data sets and to stream those data sets at high bandwidth to user applications. By distributing storage and computation across many servers, the resources may grow on demand, while remaining economical even for large volumes. For example, the Yahoo has managing of 40 petabytes of enterprise data [8].

A Hadoop cluster divides the data into small parts and distributes them across the various servers/nodes. The data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout the cluster. In this way, the map and reduce functions can be executed on smaller subsets of larger data sets, providing the scalability that is needed for big data processing. HDFS uses a master-slave architecture, with each cluster consisting of a single NameNode that manages file system operations and supporting DataNodes that manage data storage on individual compute nodes, making it the more useful choice in managing big data and Twitter [9].

Hadoop is a de facto standard of big-data analytics. It is focused on batch processing and can analyze huge big data sets. Mahout, along with the Hadoop platform is a promising technology to analyze and solve data intensive problems [10]. It has various clustering implementations like K-means, fuzzy K-means, Dirichlet and many others.

A framework for processing health data in Twitter big data (Twitter users generate 140 million tweets per day) [11] needs to be a robust framework that can handle failures at the application layer and must have clustering classification. In Mahout, firstly, preprocessing needs to be done if the data is not numeric and only then the data is converted to Hadoop format that is then converted to vector form on which clustering is executed [10]. In this work, we developed an architecture with Mahout and Hadoop using big data from Twitter as described in the next section.

3. Proposed Twitter Data Processing Framework

In this section, we present an architecture consisting of a web application that will feed a NoSQL database, MongoDB, with data coming from the Twitter API. These data (tweets) are further processed to be able to perform various analysis and operations.

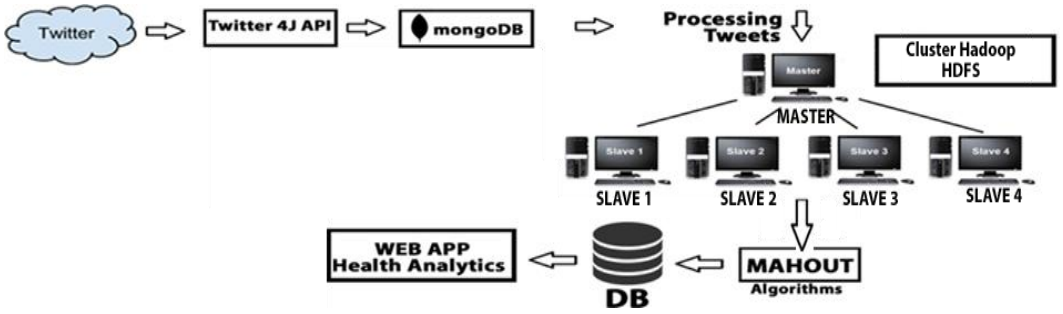

Fig. 2. Proposed Architecture

Fig. 2 illustrates the overall architecture used for the proposed solution for a health care system based on Twitter health data. Our architecture can also be applied to health care systems, making it a reference in healthcare big data analysis. By classifying both social media health data and healthcare specific systems data together, we can greatly improve the accuracy of the results.

The health care system has a large volume of unstructured data, so it is difficult to do research and diagnoses without an appropriate tool or technique. Hadoop is a tool designed to process huge volumes of data based on map-reduce concept. Map Reduce divides the data set into multiple chunks, which will be processed in parallel among multiple nodes.

In this framework we define some health keywords to be searched in big data store through a web application. We then use those keywords to classify and analyze results. We have chosen the Twitter4j library in our architecture because it is one of the most used open source libraries in Twitter data management and has frequent updates, making it a reliable choice. The API extracts Twitter data at intervals of 30 minutes [12]. When these requests to the API are overused, Twitter is blocked temporarily and only comes back to work after no less than 15 minutes, and may, in extreme cases, be blocked for exceeding "calls" to the API [12]. To avoid this problem, the tweets are retrieved in 30 minutes intervals.

We use Twitter4j with a parameter containing a set of health-related hashtags, extracting all tweets based on those hashtags. Since every 30 minutes a call is made to the API, the latest tweets are read by "id" (each tweet) larger and the last "ID" read previously, so there is no duplication or missing records and tweets. Because of the hard and cost effective process of analyse the Twitter data and retrieve them from the Twitter platform, it is mandatory to have a tool to aid and simplify the process of save data to a database. We chose the MongoDB database system, which is based on NoSQL standards and simplifies the process of saving the tweets without further processing. The MongoDB is our preferable choice because when retrieving the tweets from Twitter big data, the information isn't structured and can't be stored in relational database quickly. With MongoDB the info is saved in database files rather than in tables. The MongoDB files are furthermore similar with json files, making it simpler to analyse. The objects use key-value pairs that simplify the information management and classification implementation. MongoDB acts as the tweeter big data repository and pre-processing unit, saving all the information returned by the Twitter4j API.

After the MongoDB storage processing, the Hadoop will then process the data using the Map Reduce feature. The Map Reduce is a processing information algorithm, which handles the data and applies several calculations to it in order to organize and structure it in key value representations, storing it in the HDFS. The HDFS, which is the main storage system of the any Hadoop processing unit, generates several clones of the tweets and distributes those blocks by the cluster nodes, in order to handle fail safe recovery and distributed analysis. After the data distribution, the Hadoop system will apply Map Reduce functions and reorganize the key-value data blocks in order to quickly prepare itself for search and processing treatment. When the Map Reduce finishes the processing, the Hadoop Mahout will search the data and apply specific learning and classification algorithms, as mentioned in the previous section, in order

to further prepare the data to final result sets. The Mahout has several classification algorithms like hash count and word count, but others can be applied in order to increase its classification performance.

The final processing, the resulting data will then be saved to a traditional relational SQL becoming available to any application as a web service. These resulting data are then stored logically in a relational database ready to be used by a data analysis web application (see Fig. 2). In our sample web application, we show several statistics and results, the retrieved results change according to the inputted hashtag, and also information about a specific disease, habit changing and disease prevention, among others. The displayed results can then be used in several decision support applications as well as to improve or create new health decision support systems that can handle reliable data in a way where it once were considered improbable.

The applications of the resulted data are endless and can certainly improve the healthcare information systems and therefore people lives as well.

4. Implementation and evaluation

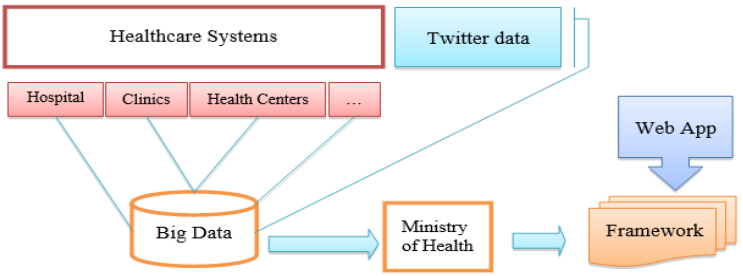

Fig. 3. Setup Architecture

4.1. Setup

Fig. 3 illustrates our setup architecture built upon our analysis framework described in the previous section. It is applied to a hybrid approach using the health systems data and the large Twitter health data.

The analyzed data are therefore complemented with classification from the retrieved and processed Twitter data, making our recommendations the more accurate we can. The process relies on our classifications algorithms and therefore we obtain a more reliable system than if the analysis were being processed separately.

The proposed setup (Fig. 2) was developed to make a simple interface to both users and gathering all the information from the framework main components. We have chosen a web application because it can work in most devices (pc, tablets, smartphones) and can rapidly be adapted in future versions, with less work at the user interface.

4.2. Evaluation

Mahout has built-in libraries used to apply multiple algorithms and Hadoop is able to provide parallelism between them. In this framework we have used Mahout, but other clustering algorithms can be used to improve results.

The semantics of text analysis is an important problem to address since, for instance, sarcastic comments are very difficult to identify. Tweets containing sarcastic comments often give exactly opposite results given the mindset of the author. These are extremely difficult to track. Also, depending on the context in which a word is used, the interpretation may also change. In a rapid manual review of the Twitter data, we have identified several sentences that are difficult to analyze. For instance the sentence *"...world politics is sick..."* or *"...eat candies can make diabetes..."* are two samples which can make the information less coherent than initially expected. Even the sentences punctuation like exclamation sentences or citations can change the classification of the sentence, disrupting the effectiveness of the classification algorithms we use.

Mahout does not really provide any mechanism to store the top terms in any database, which motivated us to extend Mahout and provide a custom program that writes the top terms to the database for each run. For Twitter API queries this implementation can be limited due to complexity and, since search does not support authentication, all queries

are made anonymously. Search is then focused on relevance and not completeness and this means that some tweets and users may be missing from search results.

4.3. Advantages

The analysis of the large tweet datasets may allow health care organizations, hospital networks and many others to attain significant improvements. Potential benefits include detecting diseases at earlier stages making it possible to be treated more easily and effectively based on prevention.

Certain developments or outcomes may be predicted or estimated based on vast amounts of historical data and based on trends of people. That analytics could help reduce waste and inefficiency in different health areas.

5. Conclusions and Future Work

In this paper we have proposed a Hadoop-based architecture to manage Twitter health big data. The proposed architecture is a promising platform to handle big data from Twitter and to perform clustering, given their inexpensive and scalable features. We conclude that the coming Twitter data have to be analyzed carefully, because the semantics of sentences may change the meaning of words. Data must be serialized before being converted in structured data. The tweets analysis in healthcare has the potential to transform the way people and health care providers use sophisticated technologies to gain different clinical insight to make informed decisions.

In future work, datasets from different social media can be taken to further research on clustering with Hadoop or Apache Storm. The criteria for platform evaluation can also evolve and may include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance. Mahout algorithms will be applied to information filtering along a web application to visualize results.

In another approach the architecture can be scalable to include data from other means, such as hospital aggregates centers to mobile applications. The future holds access to important data sources, such as smart watches and wearable devices to help our investigation in health sector. We have important information to combat the disease and allow monitoring and predicting the evolution of epidemics and disease outbreaks.

References

[1] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, “Get your jokes right: ask the crowd,” *Model Data Eng.*, vol. 6918 LNCS, pp. 178–185, 2011.

[2] A. Agarwal and J. Sabharwal, “End-to-End Sentiment Analysis of Twitter Data,” *24th Int. Conf. Comput. Linguist.*, vol. 2, no. December, pp. 39–44, 2012.

[3] K. Priyanka and N. Kulennavar, “A Survey On Big Data Analytics In Health Care,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5865–5868, 2014.

[4] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.

[5] J. Bian, U. Topaloglu, and F. Yu, “Towards large-scale twitter mining for drug-related adverse events,” *Proc. 2012 Int. Work. Smart Heal. wellbeing*, pp. 25–32, Oct. 2012.

[6] S. Mane, Y. Sawant, S. Kazi, and V. Shinde, “Real Time Sentiment Analysis of Twitter Data Using Hadoop,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 3098–3100, 2014.

[7] M. Stonebraker, U. Çetintemel, and S. Zdonik, “The 8 requirements of real-time stream processing,” *ACM SIGMOD Rec.*, 2005.

[8] V. Prajapati, *Big data analytics with R and Hadoop*. O’Reilly, 2013.

[9] T. White, *Hadoop: the definitive guide: the definitive guide*. O’Reilly, 2009.

[10] E. Jain and S. Jain, “Categorizing Twitter users on the basis of their interests using Hadoop/Mahout platform,” *9th Int. Conf. Ind. Inf. Syst.*, 2014.

[11] M. T. Jones, “Process real-time big data with Twitter Storm An introduction to streaming big data,” pp. 1–9, 2013.

[12] A. Gopal, “Enhanced Clustering of Technology Tweets,” San Jose State University, 2013.