2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.029

Big Data Instruments for Social Media Analysis

Aleksandr Blagov, Igor Rytsarev, Konstantin Strelkov, and Maximilian Khotilin

Department of Technical Cybernetics, Samara State Aerospace University, Samara, 445086, Russia

Keywords: Big data, social media, hortonworks, twitter, real-time data

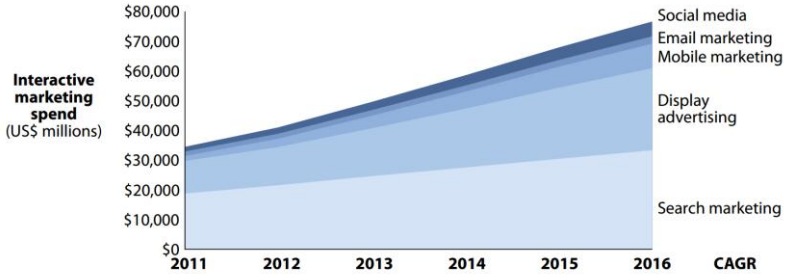
Abstract. At present, most of the researches in completely different areas are focused on big data. This paper discusses working with an important segment of big data – information from social media. There are algorithms of collecting, structuring, analyzing and visualizing collected data with a Hortonworks Sandbox instruments. The data collection is performed through given keywords. As a result there are quantitative and qualitative characteristics of tweets, collected all over the world.

Introduction

In recent years "big data" has become something of a buzzword in business, computer science, information studies, information systems, statistics, and many other fields. But what does it really mean?

«Big data» is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. In fact, the term "Big data" means working with a high-volume data, which has inhomogeneous structure and updates frequently, from different sources in order to improving efficiency of the work process, creating new products and increasing competitiveness. As technology continues to advance, we constantly generate an ever-increasing amount of data. This growth does not differentiate between individuals and businesses, private or public sectors, institutions of learning and commercial entities. It is nigh universal and we can use it almost everywhere, for example: in public and private sector, science and technology, health care and in social media. We chose social media as the aim of our research. Nowadays social networks playing an important part, being the subject of socialization of people in one case, and being one of the most available and powerful political, economical and ideological instrument in other.

In terms of marketing, social networks become the most attractive medium for different programs' realization: they are on the second place of the most quickly developing marketing channels, inferior their position only to the mobile channel [1]. You can see it on the diagram on Fig. 1.



	2011	2012	2013	2014	2015	2016	CAGR
Social media	$1,590	$2,119	$2,760	$3,453	$4,217	$4,995	26%
Email marketing	$1,510	$1,694	$1,875	$2,066	$2,262	$2,468	10%
Mobile marketing	$1,652	$2,777	$4,238	$5,697	$7,057	$8,237	38%
Display advertising	$10,949	$12,860	$16,085	$19,783	$23,919	$27,600	20%
Search marketing	$18,756	$21,553	$24,613	$27,515	$30,433	$33,319	12%
Total	**$34,457**	**$41,003**	**$49,571**	**$58,515**	**$67,888**	**$76,620**	17%
Percent of all ad spend	16%	18%	20%	22%	24%	26%	

Fig. 1. Forecast: US Interactive Marketing Spend, 2011 to 2016.

In Twitter, the social network, the specialists of marketing can communicate with their audience without Service of Public Relations. Due to this, there could be a communication with the specific person, in contrast of depersonalized companies.

Using their branded terms and hash-tags, the specialists of marketing can learn customers' opinion about their products, brands and companies.

The global attention, which Twitter uses, shows high capabilities of social network technologies for public discussion and forming the perception of brands.

The Facebook also become a universal solution for big segments of audience, which want to communicate in Internet. The Facebook, and other social networks, offers for the specialists of marketing an access to carefully segmented clients and serves as a convenient platform for running viral campaign and new forms of interactive commercials.

The direction of BigData, which related with social media is one of the most perspective and dynamically developing. [2]-[4] can be attributed to the most known works in this area.

There are many instruments for gathering, analyzing, searching, sharing and transferring big data information. For example: "BigInsights" produced by IBM, "Cloudera"produced by Cloudera, Inc, "Hortonworks Sandbox" by Hortonworks, etc.

We chose "Hortonworks Sandbox" because it is a personal, portable Hadoop environment and includes many of the most exciting developments from the latest HDP distribution, packaged up in a virtual enviroment. The Hortonworks Sandbox includes the core Hadoop components (HDFS and MapReduce), as well as all the tools needed for data ingestion and processing.

Discussed Problems

In social networks there are a lot of high-volume data types which changes very quickly. It is very important to extract useful and necessary information for specific tasks.

The analysis of this information could be used in different areas: in marketing, economy, social researches, politics, etc.

For example, in 2012 year in the United States a tool had been developed for real-time analysis of sentiment expressed through Twitter toward the incumbent President Barack Obama, and the nine republican challengers [2].With this analysis, it had been explored whether Twitter provides insights into the unfolding of the campaigns and indications of shifts in public opinion.

Twitter allows users to post tweets, messages of up to 140 characters, on its social network. Twitter usage is growing rapidly. The company reports over 100 million active users worldwide, together sending over 250 million tweets each day (Twitter, 2014).

Most work to date has focused on post-facto analysis of tweets, with results coming days or even months after the collection time. However, because tweets are short and easy to send, they lend themselves to quick and dynamic expression of instant reactions to current events. Thereby, information from Twitter is: concise, poorly structured, dynamically changed and reproducible in high volumes. To store, manage and represent such data type with traditional databases is impossible. That is why BigData instruments and methods can help us. Due to them we can organize sentiment analysis of user-generated data that can provide fast indications of changes in opinion, in spite of time.

Solutions Based on Different BigData Tools

We decided to use Hortonworks Sandbox system, because this platform provides the processing of high-volume arrays of data in a clear to every user way, and there is no need in special hardware or equipment. You just need a computer, with a technical characteristics, which are a little better than medium, and a continuous broadband to the Internet.

From the official Hortonworks website you can download the package Sandbox, which is an autonomous virtual machine. Hortonworks Sandbox built using Hortonworks Data Platform (HDP) 1.2 — an open-source program platform, which includes Apache Hadoop.

Apache Hadoop – is software for creating distributed applications for intense work with thousands of computers and petabytes of data.

The method of Hadoop functioning is in partition the whole computing task in many little fragments, which can be performed on every node of the cluster.

Moreover, after downloading Hortonworks Sandbox, users receives an access to the environment, which they can explore and evaluate the capabilities of the basic projects of ApacheHadoop family, such as Apache Pig, Apache Hive, Apache HCatalog and Apache HBase, which are included in the structure of the platform, based on BigData technology and built on open-source principles. The parts (structure) of the Hortonworks software can be seen on Fig. 2.

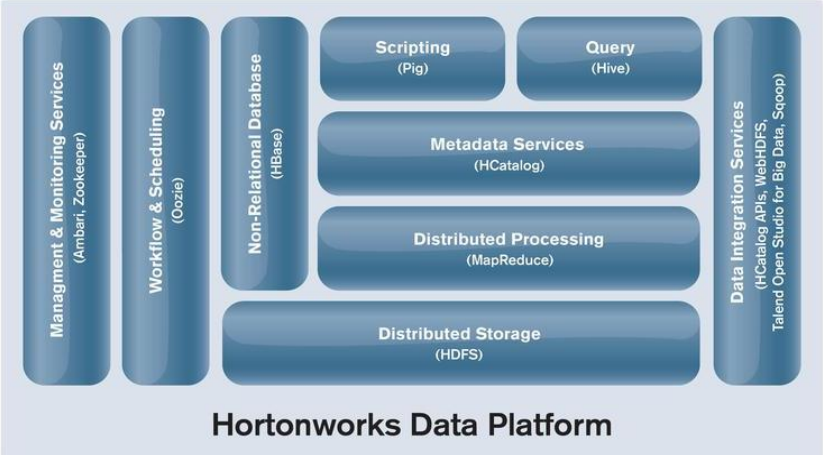


Fig. 2. Hortonworks data platform.

For distributed computing HortonWorks Sandbox uses technologies and algorithms of MapReduce, the instrument for organizing data storage - Hive and instrument for collecting data - Flume.

Algorithm of Working with Social Media

The main stages of working with social media are data collection, its' structuring and analysis, visualization.

Algorithm of collecting the data. The algorithm of gathering data using Hortonworks Sandbox consists of next few steps.

We used HortonWorks HDP 2.1 Hadoop distribution (on Windows), and Apache Flume to collect Twitter Tweets and store them in HDFS (Hadoop Distributed File System) for later analysis.

The sequence of actions looks as follows:

1. Creating Twitter application:

We have to login with our Twitter ID, and create own Twitter Application. It is necessary for access by API-keys to tweets all over the world.

2. Installing Flume

To collect information we need to use Flume. Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It has a simple and flexible architecture based on streaming data flows; and is robust and fault tolerant with tunable reliability mechanisms for failover and recovery [5]. Flume's high-level architecture is focused on delivering a streamlined codebase that is easy-to-use and easy-to-extend.

3. Installing the WinSCP which allows secure file transferring between local and remote computer. Uploading the flume.conf file, which contains our token values from twitter application. This file is necessary to set the parameters of searching: keywords and the path for saving the data.
4. Launching the data collection:

For organizing the data collection, it is necessary to define a set of keywords, which are used for data collection.

The command for this in Hortonworks Sandbox is:

nohup flume-ng agent --conf-file /etc/flume/conf/flume.conf --name TwitterAgent

Algorithm of data analysis. The collected data represented by many files without any filename extension. These files contain unstructured data.

For its structuring and analysis it is necessary:

1. To compose the dictionary, which defines positive, negative and neutral meaning of the phrases.
2. To compose the dictionary, which defines the geographical location (TimeZone).
3. To compose and perform the SQL-request, which collects information that we need in Hadoop Distributed File System tables.

The most important things at this stage are frequency algorithm and phrase meaning algorithm.

The term frequency ($fr(w, s)$) is the number of times that a word w occurs in a sourced:

$$fr\,(w,s) = |\{w \in s\}|$$

In computing the term frequency, all occurrences of a word in a document are counted. Therefore the term frequency can assume a value in the interval [0; n], where n is the total number of words in the document.

The connotation ($sp(w, d)$) only checks if a word w is negative, neutral or positive, it's define by d – sentiment dictionary:

$$sp(w,d) = \begin{cases} 0, & \text{if } w \text{ is negative,} \\ 1, & \text{if } w \text{ is neutral,} \\ 2, & \text{if } w \text{ is positive} \end{cases}$$

The obtained relation 'tweet – TimeZone' is defined by TimeZone Dictionary followed by using frequency algorithm.

1. The next step is to launch by means of an instrument, calls Hive, the processing of the SQL-request, which contains necessary algorithms.
2. To import the analyzed data the OBDC driver is needed. The OBDC driver is a program interface (API) for database access. Due to its help the Hive and importing data to Excel are able to be launched.

After collection, structuring and data analysis it is important to visualize the data.

Generally, an algorithm of working with social media could be described by the following scheme (Fig. 3).

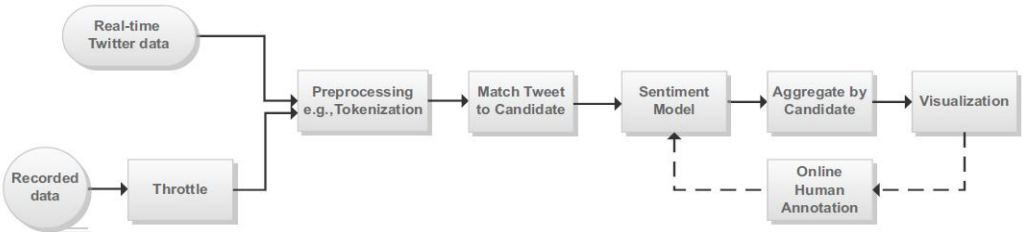


Fig. 3. The scheme for real-time processing sentiment data.

Computational examples and analysis

As an example, we collected, analyzed and visualized data from Twitter with this keywords: bigdata, hadoop, biginsights, infosphere, mapreduce, hortonworks, cloudera. We chose this keywords with the view to determine how much theme "Big Data" is popular.

Data collections were produced at workstation in Samara State Aerospace University (Samara region, Russia) from 10.12.2014 to 11.12.2014 during 24 hours with permanent access to broadband Internet connection. We collected about 30 000 records with the tool Flume – part of Hortonworks Sandbox. The algorithm is written in section 4.1.

We produced structuring and analyzing of collected data like was described in section 4.2. You can see the result of this in Fig. 4-Fig. 6.

For example, Fig. 4 we shows the world map where size of the circles is proportional to amount of the tweets from different countries.

Fig. 4. Distribution of the tweets all over the world.

In Fig. 5 you can see how the sentiment (positive, negative or neutral) is distributed in countries.

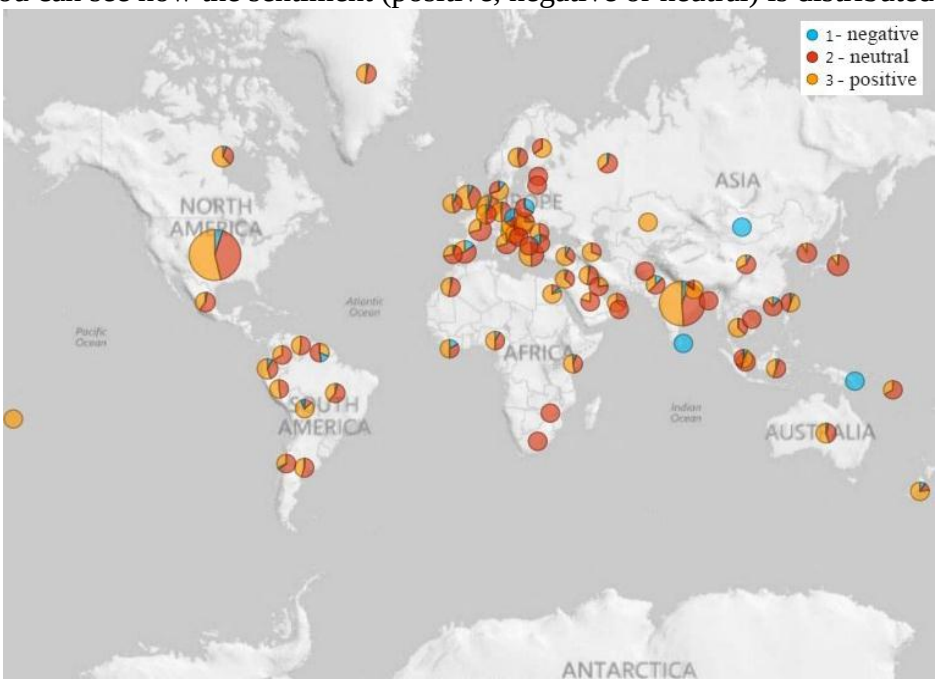


Fig. 5. Sentiment distribution of the tweets all over the world.

Fig. 6 shows the cloud of tags – the most common words in all collected data. The size of word is proportional to the amount of its usage in different tweets all over the world.

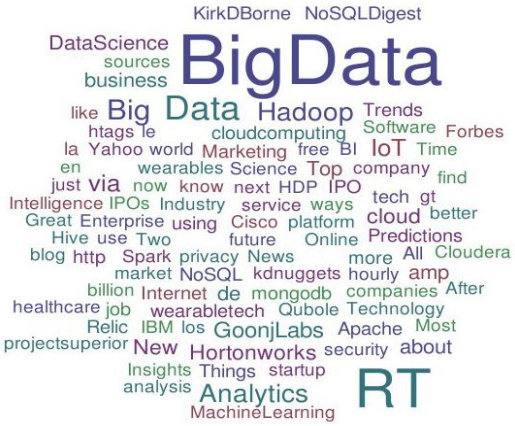


Fig. 6. Cloud of the tweets.

Based on the results of our experiments, we can say, that “Big data” technology is quite popular in whole world, especially in USA and India. People from most countries refer to this theme positive or neutral, which indicates about perspective of development tools and methods of Big Data technology

Conclusions

Nowadays, in the dynamically developing world, it is important to process and use real-time data, incoming in a high volume. Social media is one of the most striking examples of such data. The significance of such information varies from “useless” to “ represents the state's importance.”

The analysis of data helps us to evaluate the attitude of citizens of different countries to important issues, awareness, popularity.

There are a lot of instruments, based on Hadoop technology, which are flexible and customizable to achieve the goals.

One of such instruments, used by authors of this article, is Hortonworks Sandbox, which allows implementing algorithms of collecting, analyzing, visualizing data from social media.

This theme may have a development in area of forecasting and organizing social researches.

Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of SSAU among the world’s leading scientific and educational centers for 2013-2020 years.

References

[1] VanBoskirk S. US interactive marketing forecast, 2011 to 2016 //BCAMA, Marketing Association of BC. – 2011.
[2] Wang H. et al. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle //Proceedings of the ACL 2012 System Demonstrations. – Association for Computational Linguistics, 2012. – C. 115-120.
[3] Saif H., He Y., Alani H. Alleviating data sparsity for twitter sentiment analysis. – CEUR Workshop Proceedings (CEUR-WS. org), 2012.
[4] Groot R. Data mining for tweet sentiment classification. – 2012.
[5] Hortonworks Sandbox tutorial [Online]. Available: http://hortonworks.com/tutorials/