Big Data Analytics for Logistics and Transportation

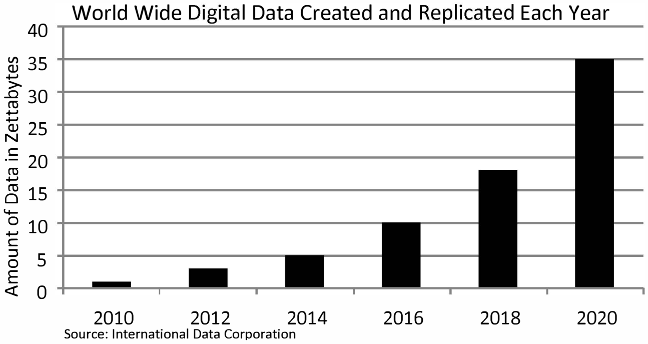
Abdelkarim Ben Ayed, Mohamed Ben Halima, Adel M. Alimi
REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax, ENIS, BP 1173, Sfax, 3038, Tunisia
{abdelkarim.benayed.tn, mohamed.benhlima, adel.alimi}@ieee.org

***Abstract*—Nowadays, there are many challenges for the logistics industry mainly with the integration of E-commerce and new sources of data such as smartphones, sensors, GPS and other devices. Those new data sources generate daily a huge quantity of unstructured data, to deal with such complex data, the use of big data analytic tools becomes an obligation. In this context, many works have been done recently in the integration of big data analytics in the logistics industry. In this paper, we propose to give a review of the latest applications of big data analytics in the field of logistics and transportation industry and to propose a novel approach to detect and recognize containers code based on a Hadoop big data analytics system.**

Keywords— logistic; transportation; big data analytics.

I. INTRODUCTION

Social networks, smart phones, tablets, GPS devices, sensors, log files, and many other devices and sources are generating every second a large quantity of unstructured data. Besides, the quantity of data created every year is much bigger than that created ever before, that is why our age is called the age of information (see Fig. 1).



Source: International Data Corporation

Fig. 1. Exponontial growth in digital data during actual decade [1], p. 3.

All these data contains treasures of valuable information that could be very useful for governments as well as private companies for making deep analysis, monitoring, taking decisions, improving their quality of services, etc. However, most of the traditional data analysis tools, such as relational databases, are unable to store and manage such very complex data, called big data.

Mainly three aspects of complexity called the 4Vs (Volume, Velocity, Variety and Veracity) characterize big data. Volume represents to the huge size of data that starts from one terabyte or more, Velocity the high speed of collecting data,Variety the high diversity of data types and formats that require to be stored and analyzed together and Veracity the uncertainty of data (see Fig. 2).

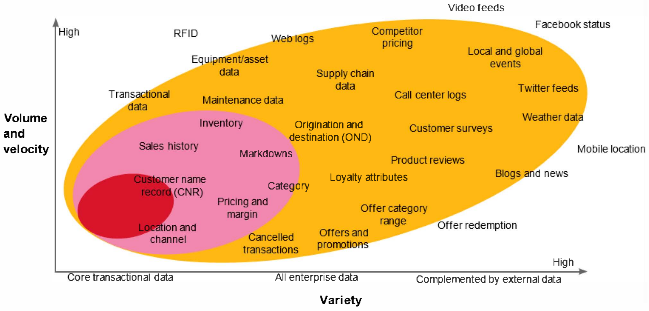


Fig. 2. Large variety of travel and transportation companies data [2], p. 3.

Big data analytics is used in many areas such as machine learning, computer vision, web statistics, medical applications, DNA analysis, data classification and clustering [3], and in public and private industry including the logistics and transportation industry.

To deal with big data, companies including the logistics and transportation industry need to use dedicated tools called big data analytics. These tools allow efficiently and easily managing and analyzing the huge data coming from roads and vehicles sensors, GPS devices, customer's applications and websites, etc.

Today, many big data analytics solutions are available, but the most used is the open source Apache Hadoop framework [4]. Hadoop uses a distributed storage and parallel computation model over a cluster of many commodity machines to easily handle big data.

The remaining sections of this paper are organized as follows. In section 2, we give a brief presentation of Apache Hadoop. In section 3, we present examples of projects using big data analytics for logistics and transportation industry. In section 4, we present our proposed system for containers code recognition using big data analytics. Finally, we present a conclusion and future perspectives in section5.

II. APACHE HADOOP: BIG DATA ANALYTICS

To deal with the increased demand on storage and computation requirements, old systems are based either on scale up solutions or scale out solutions. Scale up solutions use a classic non-parallel architecture with, however, improved resources, but it is very expensive (cost/performance) and limited by a technical barrier. Scale out solutions use parallel architectures to improve

978-1-4799-8400-8/15/$31.00 ©2015 IEEE

computation resources with a lower cost, but with a much higher engineering effort. In the other side, big data analytics solutions such Apache Hadoop are based on a framework that abstract most of the engineering effort caused by parallel architectures.

A. *Presentation of Hadoop*

Apache Hadoop is an open source framework written in Java. It is designed to deal with very large data sets using computer clusters of commodity hardware. It has two main parts, a distributed storage part: the Hadoop Distributed File System (HDFS) and a processing part: the MapReduce programming model. Doug Cutting developed Hadoop in 2005[5] based on Google File System (GFS) and Google MapReduce published papers [6], [7].

B. *Architecture of Hadoop Framework*

Hadoop is composed of two main parts, a storage part managed by HDFS and a processing part managed by MapReduce programming model or higher programming languages (see Fig. 3).

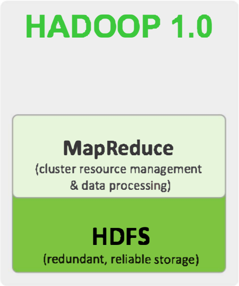

HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)

Fig. 3. Hadoop Framework main components HDFS and MapReduce [15]

A typical Hadoop system is composed by a master server (with one or two backup mirros) and a many low cost machines/slaves (thousands) running linux. Master has “Name Node” and “Job tracker” components that manage respectively “Data node” (storage task) and “Task tracker” (processing task) in other machines.

C. *Hadoop Distributed File System (HDFS)*

The Hadoop framework uses a distributed redundant storage system called HDFS that stores files in blocks replicated in multiple machines. A main server (master node) manage data splitting and replication in the other chunk servers (workers nodes) used for both data storage and processing (see Fig. 4).

D. *MapReduce programming model*

Hadoop uses mainly MapReduce as a programming model to process large datasets. MapReduce is composed of two functions: “Map” divide problems to smaller ones and “Reduce” combine results. The Map and Reduce functions are to be written by the user. MapReduce take care of all the details of distributed computation. The main server (master node) is not overloaded by computation, it is responsible only of communication with user application and managing the other workers nodes. The tasks are sent to data (not the data sent to worker machine) which improve the system performance and mainly the bandwidth (see Fig. 5).

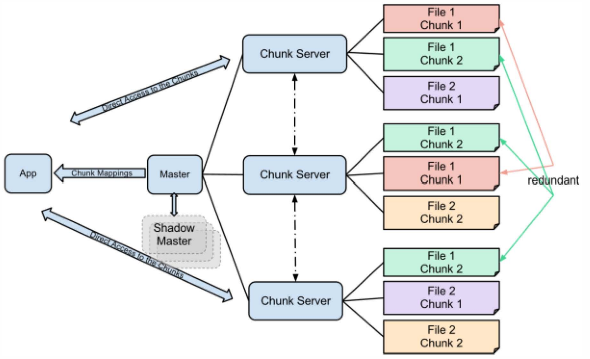


Fig. 4. Distributed file system architecture (e.g. Google File System) [17]

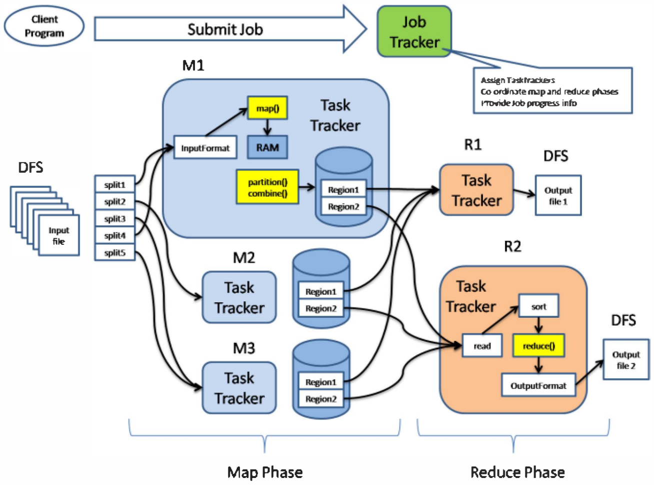


Fig. 5. MapReduce architecture [18]

E. *Other Hadoop elements*

Besides programming in MapReduce, users can develop their codes in other easier high-level programming languages that will be translated automatically to Map and Reduce tasks such as:

- Hive: data warehouse language using SQL-92 querries.
- Pig: data flows oriented language using Pig Latin programming language.
- Hbase: A sparse database for storing large quantities of data.

F. *Advantages of big data for logistics industry*

Big data analytics (e.g. Hadoop) have many advantages for logistics and transportation industry such as:

- More efficient for real-time big logistics applications collecting data from a very large network of sensors and GPS devices

- Storage and process of very big files
- Improved exploitation for both structured and unstructured logistics data
- Development of powerful logistics projects with smarter strategies based on collected and analyzed data
- Development of efficient real-time traffic monitoring applications
- Development of more accurate logistics predictions that improve service quality for customers and improve companies' revenues.

III. Big Data Analytics for Logistics and Transportation Industry

Many big data analytics research projects are realized for logistics and transportation industry to deal with the huge data coming from roads and vehicles sensors, GPS devices, customer's applications and websites, etc.

In the following, we present an overview and compare some of these projects.

A. Real-time vehicles monitoring in India

The main objective of this project is to deploy modern technologies like Big Data analytics and Hadoop in order to improve operational efficiency for logistics and transportation firms. In fact, these technologies helps firms' managers to make better business decisions.

In this project, the authors collect data from vehicles about fuel, speed, acceleration, GPS location coordinates using vehicles sensors and GPS devices with other data such date, time, driver's id, etc. and then send these information by packets over wireless communication (GPRS) to clustered servers running Hadoop.

All this unstructured data comes from hundreds of vehicles sending packets every 2 seconds to a HDFS system to store such big data. An analysis is made then weekly or monthly over these terabytes of data using Hadoop analytics system in order to improve the transportation company productivity and help reduce the costs.

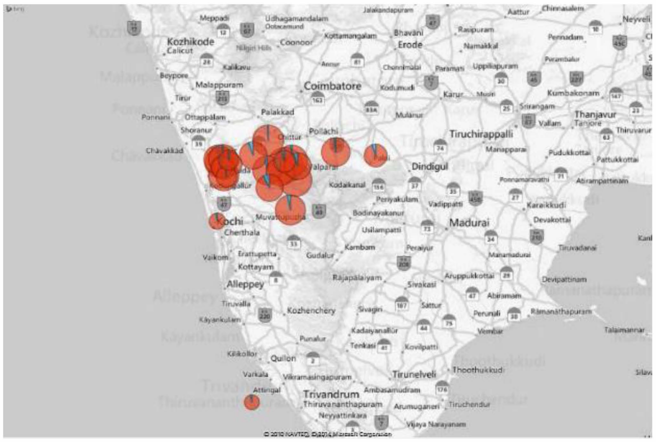


Fig. 6. Locations of trucks with risk factor [8], p. 2368.

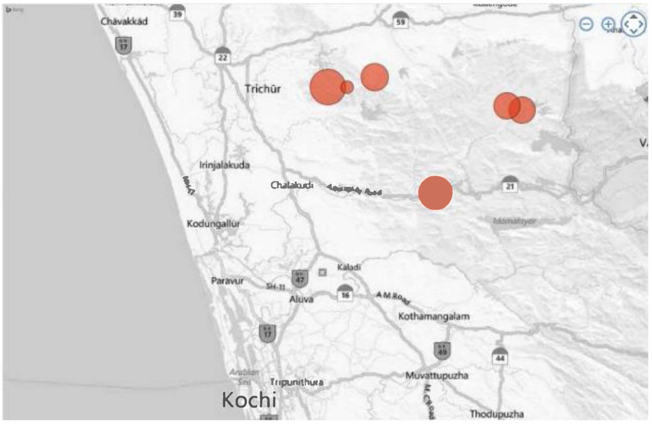


Fig. 7. Graph showing locations of idling trucks [8], p. 2368.

The analysis of data allows also monitoring driving behavior and answering to some questions like: which vehicles are wasting fuels? which drivers have the highest risks? The results of analysis are displayed by charts and graph (see Fig. 6 and Fig. 7).

B. City of Dublin, Public Transit System

This smart city project was a collaboration between IBM and Dublin City Council, it took place in Ireland from 2010 to 2013. IBM big data analytics helped the city of Dublin to improve its public bus transportation network and reduce the increasing traffic congestion problem [9], [10].

The main idea was to reduce congestion in the city of Dublin in its public bus transport network without making major modification for the city historic infrastructure.

TABLE I. Sources of Data for the Project: City of Dublin [11], p. 7.

Municipal Data – City of Dublin		
Data Source	***Description***	***Approx. Scale***
Bus System	GPS Data Speed Data Stop Data Fare Data	1000 Buses Location reported every 20 seconds
Traffic Flow Data	Traffic Light Data XML	24 Intersections
CCTV monitoring	Real-time video Stream across the city	
Road Weather Conditions	CSV	54 Stations
Road Works and Maintenance	CSV	
Dublin Event Data	Small Attendee Events Large Attendee Events	

The data is collected from many sources (see Table I) such as bus GPS devices that send data from 1,000 buses every 20 seconds, bus timetables, cameras and traffic sensors in roads, tramways, and bus lanes. Collecting such enormous amounts of multiple and fast data was not possible with traditional data sets. That is why the data is collected in clustered servers running IBM big data analytics to build a real-time digital map of Dublin city transportation network.

Advanced analytics on the collected data helped to identify traffic problems and answer questions such as the optimal time to start bus lanes, the best place to add more bus lanes, etc.

The project helped the city to better monitor and manage the traffic in real time, which accelerate decision making and improved traffic flow and mobility in the city.

As future improvement for the project, a predictive analytics solution will combine meteorological data with other data to assure good traffic flow in extreme weather conditions.

C. City of Da Nang, Vietnam, Traffic Management System

This project is made by IBM smart city technologies in order to reduce traffic congestion and pollution in city of Da Nang in Vietnam [9], [12].

The objective of the project is to build water and transportation traffic management system able to deal with the city fast-growing population of a city of more than a million population.

Sensors on Da Nang's buses, roads and highways collect data for the management system that use the collected data to optimize traffic lights synchronization and reduce traffic congestion.

Besides, data coming from ships sensors gives information about water status such as water's turbidity, salinity, pH, chlorine and conductivity levels, which allow customers to receive appropriate alerts if necessary.

All these different information are combined together in big data analytics system that provides a real-time summary of traffic status, events and incidents through maps and alerts for better management of the city transportation network.

The project reduced traffic congestion and energy consumption, created an efficient control and management system that support the high growth of the city population and resulted in a better and safer mobility for commuters.

D. British Airways' Know Me Program

A big data analytics project that started in 2012 from Opera Solutions company that aims to improve the quality of British Airways services for its customers. The objective is to understand clients' needs better than any competitive airline company [9], [13] .

The project collected different types of information about 20 million customers via websites, smart phones and tablets applications, blog sties' rating, likes on social medias, conversations with call centers, etc.

After that, a big data analytics system stores and analysis all the structured and unstructured data to identify customers' preferences, characteristics, problems, and to provide them high quality services.

E. City of Stockholm Real-Time Intelligent Transportation Services

Another project based on IBM InfoSphere big data analytics that aims to improve the quality of the transportation network in the city of Stockholm [14].

120,000 vehicles of taxis and trucks equipped with GPS devices were used to collect and send a large collection of data every second combined with a map containing over 600,000 links (see Fig. 8).

The used big data analytics system combined the collected data with past traffic data and weather forecasts to generate more accurate predictions about future traffic conditions such as shortest-time routes in real-time. The results served for publics, police officers, firemen, urban planners, etc.

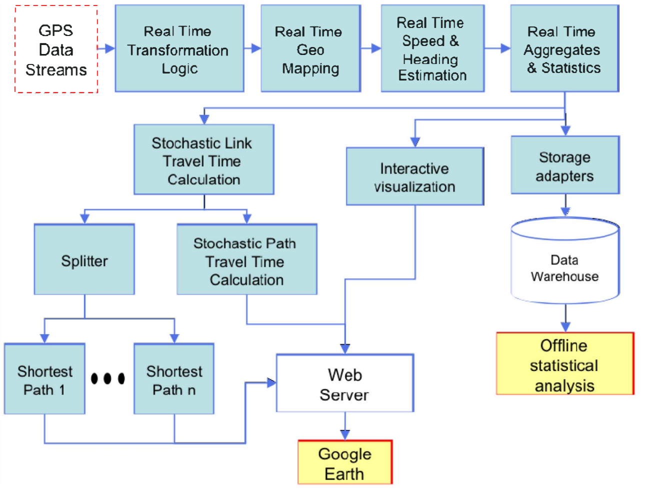

GPS Data Streams
Real Time Transformation Logic
Real Time Geo Mapping
Real Time Speed & Heading Estimation
Real Time Aggregates & Statistics
Stochastic Link Travel Time Calculation
Interactive visualization
Storage adapters
Splitter
Stochastic Path Travel Time Calculation
Data Warehouse
Offline statistical analysis
Shortest Path 1
•••
Shortest Path n
Web Server
Google Earth

Fig. 8. Stockholm real-time monitoring of transportation network flow [14], p. 4.

F. Cloud-Enhanced System Architecture for Logistics Tracking Services

The goal of this project is to build a system that combine internet of things, SaaS cloud architecture and big data analytics technologies for setting-up an efficient real-time monitoring of customers cargoes [16].

The data is collected from mobile phones: 2D QR codes, GPS locations and RFID electronic codes. Then, data is sent; using RSA encryption algorithm to protect customers' privacy; through wireless networks. A big data analytics system using HBase as databse is used for storing all these unstructured data (see Fig. 9 and Fig. 10).

Using a traditional computing system to handle such project is not possible because of the large requirements of storage, calculations and bandwidth for such lot of unstructured logistics information.

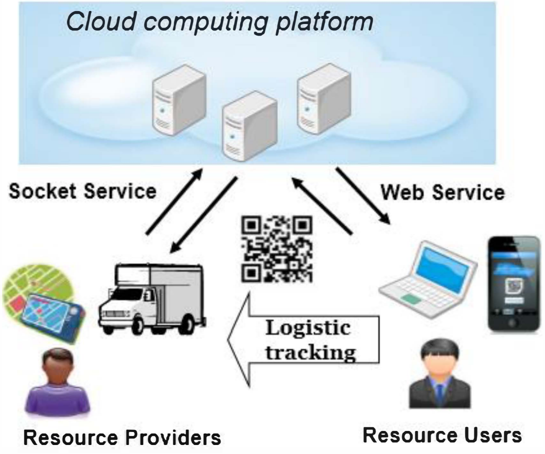


Fig. 9. Architecture overview of cloud-enhanced system architecture for logistics tracking services [16], p. 546.

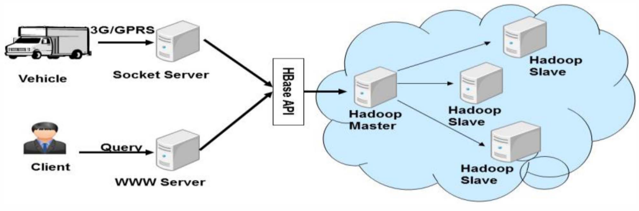


Fig. 10. The cloud computing structure of cloud-enhanced system architecture for logistics tracking services [16], p. 547.

We can classify these transportation projects in two categories, projects to improve operational efficiency and projects that improve customer experience

IV. Proposed Big data system for containers code Recognition

Ship transportation industry is nowadays a very active industry with very large number of containers to be transported every day. To supervise the delivery of all these containers, unique identifiers codes are written on the container, but manual reading of these codes include lot of problems such as slow speed and high error rate.

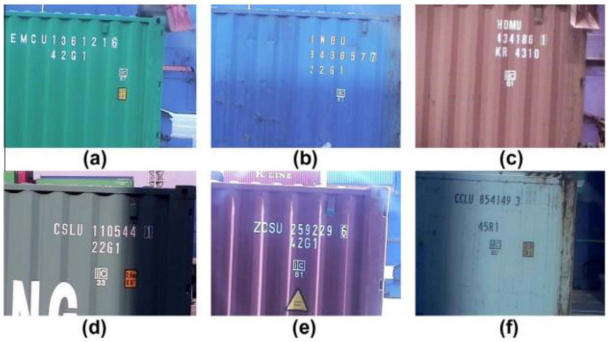


Fig. 11. Examples of containers codes [19], p. 2843.

Besides, unlike vehicle license plate recognition systems, container-code recognition systems has more challenges because of the low code/container contrast and the large varieties of sizes, colors, positions, inter-spaces and alignment of these codes (see Fig. 11).

According to [19], automatic container code recognition require three steps: text detection, characters extraction and finally text recognition. Because text detection step is very important for the other steps, we chose a robust method insensitive for code contrast and other text variable characteristics such as texture-based text detection method using Haar wavelet transform for text features extraction and Support Vector Machine (SVM) to classify these features into text and non-text regions [20].

To overcome the high computation time of the employed text detection method [20] and text recognition method [21], which is a big problem for real time and industry applications including ship transportation industry, we decided to use Hadoop MapReduce to have a parallel execution model as a solution to reduce the computation time for the proposed system.

First, container code is captured using monitoring cameras or mobile devices and stored on Hadoop distributed system file (HDFS). After that a pre-processing and graying color image steps are applied on the captured image. Next, we decompose the gray image in 20x20 pixel blocs. These blocs are analyzed and classified separately on different machines using MapReduce programming model to extract the text regions. Next, another step is used to separate code characters from the extracted text regions. Then, Optical Character Recognition (OCR) is applied on individual characters using MapReduce programming model. Finally, container code is recognized by merging these characters (see Fig. 12).

V. Conclusion

In this paper, we presented the big interest of using big data analytics for logistics companies with real project examples.

To improve their performance and assure a competitive attitude, logistics and transportation companies must shift to the big data analytics to deal quickly and efficiently with today large variety of data. Big data analytics have also lot of benefits like traffic congestion reduction, vehicles and drivers control, improvement of customer services, etc.

Besides using powerful big data analytics software enhance routing for planes, trains and trucks, which reduce energy consumption in developed countries and decrease the impact of transportation industry on the environment.

In addition, we proposed a real time system for container-code recognition based on Hadoop big data analytics solution.

As future work, we propose to apply a similar big data analytics project to our city, Sfax, in Tunisia, which suffer of a large and increasing population of more than one million habitants and dramatic pollution because of its industrial aspect and its numerous transportation vehicles. The proposed project will integrate intelligent techniques with the power of big data analytics to make new plans and strategies in the goal to reduce the impact of pollution and traffic congestion in the city of Sfax.

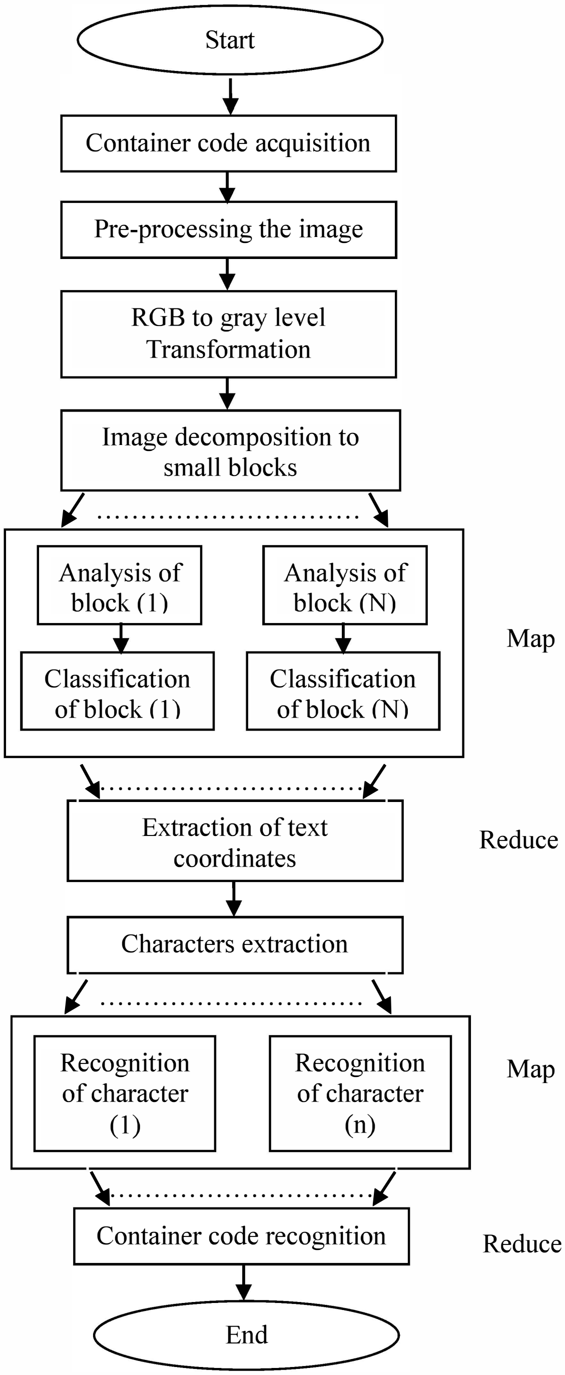

Start
Container code acquisition
Pre-processing the image
RGB to gray level Transformation
Image decomposition to small blocks
Analysis of block (1)
Analysis of block (N)
Classification of block (1)
Classification of block (N)
Map
Extraction of text coordinates
Reduce
Characters extraction
Recognition of character (1)
Recognition of character (n)
Map
Container code recognition
Reduce
End

Fig. 12. Big data system for containers code recognition

Acknowledgment

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

[1] J.M. Tien, "Big Data: Unleashing information", Journal Syst Sci Syst Eng (Jun 2013) 22(2), pp. 127-151.

[2] K. Wedgwood and R. Howard, "Big data and analytics in travel and transportation", IBM Big Data and Analytics White Paper, November 2014.

[3] A. Ben Ayed, M. Ben Halima, Adel M. Alimi, "Survey on clustering methods : Towards fuzzy clustering for big data", 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, Tunisia, 2014, pp. 331-336.

[4] http://hadoop.apache.org/

[5] http://en.wikipedia.org/wiki/Apache_Hadoop

[6] S. Ghemawat, H. Gobioff and S.T. Leung, "The Google file system", SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles, pp. 29-43, 2003.

[7] J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters", OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[8] A.P. Sivan, J. Johns and J. Venugopal, "Big Data Intelligence in Logistics Based On Hadoop And Map Reduce", International Conference on Innovations in Engineering and Technology (ICIET'14), 21-22 March, India, 2014.

[9] J. Viegas, "Big data and transport", International Transport Forum, October 2013.

[10] S.D. Galligan and J. O'Keeffe, "Big Data Helps City of Dublin Improve its Public Bus Transportation Network and Reduce Congestion", IBM press, May 2013.

[11] P. Yip, "Transform Industries, Institutions, and Societies with Watson", Smarter Business Summit, Halifax, Canada, 2014.

[12] N.Y. Armonk, "Da Nang, Vietnam Turns to IBM to Transform City Systems", IBM press, August 2013.

[13] H. Shilling, "Big Data Takes the Travel Industry in New Direction", Opera Solutions blog, june 2013.

[14] A. Biem, E. Bouillet, H. Feng, A. Ranganathan, A. Riabov, O. Verscheure, H. Koutsopoulos, C. Moran, "IBM InfoSphere Streams for Scalable, Real-Time, Intelligent Transportation Services", SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA.

[15] http://dataconomy.com/hadoop-components-need-know/

[16] X. Lin, X. Zheng, "A Cloud-Enhanced System Architecture for Logistics Tracking Services", International Conference on Computer, Networks and Communication Engineering (ICCNCE), May 2013, pp. 545-548.

[17] http://en.wikipedia.org/wiki/Google_File_System

[18] http://architects.dzone.com/articles/how-hadoop-MapReduce-works

[19] Wu W., Liu Z., Chen M., Yang X. and He X., "An automated vision system for container-code recognition", Expert Systems with Applications, 39(3), 2012, pp. 2842-2855.

[20] Sayahi S. and Ben Halima M., "An intelligent and robust multi-oriented image scene text detection", The 6th International Conference of Soft Computing and Pattern Recognition (IEEE SoCPaR'2014), pp. 418-422.

[21] Halima M. B., Karray H., and Alimi A. M. (2010), "A comprehensive method for Arabic video text detection, localization, extraction and recognition", Advances in Multimedia Information Processing-PCM 2010, Springer Berlin Heidelberg, pp. 648-659.