A HADOOP-BASED DATA PPROCESSING PLATFORM FOR FRESH AGRO-PRODUCTS TRACEABILITY

Mark Xu[1], Sajid Siraj[1] and Lin Qi[2]

[1]Portsmouth Business School, University of Portsmouth, Portsmouth, UK PO1 3DE

[2]School of Economics and Management, , Beijing Information Science and Technology University, NO.12 Xiaoying East Road, Qinghe, Haidian District, Beijing, 100192,China

ABSTRACT

Wireless sensors on cold chain containers generate huge volume of real-time data that needs efficient processing for decision making, e.g. shelf life prediction due to quality decay. This paper introduces a Cloud Computing Traceability Platform (CCTP) which is constructed on top of Hadoop cloud computing framework. The user requirements for CCTP are gathered through various means including literature review, case study, brainstorming and expert questionnaire. The system is developed in Java and is evaluated on its effectiveness. CCTP provides a cloud-based unified data mining and decision support model for enterprises to achieve fresh agro-product supply chain management optimization.

KEYWORDS

Big Data; Hadoop; Traceability; Shelf-life prediction; Cold Supply Chain.

1. INTRODUCTION

Food traceability is a mechanism to keep a track of food products right from its origins to the final destination. It is a necessary measure to ensure quality and safety in the supply chain for fresh agricultural products (Kher et al., 2010, Saltini and Akkerman, 2012). Wireless sensor networks (WSN) and radio frequency identification (RFID) have been widely used for acquiring important information on food items and thereby generating continuous stream of data related to those items (Qi et al., 2011; Papetti et al., 2012, Azuara et al., 2012, Feng et al., 2013, Qi, 2014). The big data generated, on one hand, reduce the computational efficiency of traceability system, while on the other hand, provides an opportunity to exploit better data mining techniques and decision support models (Chen et al., 2013, Galvão et al., 2010, Saltini et al., 2013). It is suggested that fresh agro-product supply chain needs a new platform for traceability with massive storage capacity and highly parallel processing ability (Bo and Wang, 2011).

Cloud computing is a demand-oriented resource allocation method that has potential for massive storage, powerful processing, and rapid response ability with affordable cost. Applying cloud computing to agriculture data shows potential advantages (Zhang et al., 2010), for example, (1) an cloud-based traceability platform can serve a large number of enterprises so as to save costs on hardware, software, technicians and maintenance, (2) the unified data mining and decision support model can be shared by enterprises to achieve supply chain management optimization, and (3) increased credibility of traceability data so that food quality and safety information can be effectively shared by the public. However, the benefits of applying Hadoop parallel processing architecture to live data generated from the cold chain has not been adequately explored. This paper reports the design, development and evaluation of such a system - cloud computing based traceability platform (CCTP). The system is based on Hadoop cloud computing technology and Map/Reduce programming framework.

In the following section, the relevant literature on big data and cloud computing are reviewed. Section 3 describes the CCTP system analysis including survey design and users requirements. The system architecture and model design are reported in section 4. The system implementation and evaluation in test beds are provided in section 5. This is followed by discussion and conclusion of this research.

2. HADOOP BIG DATA PROCESSING FRAMEWORK

The concept of “big data” has emerged from three major recent developments in volume, velocity, and variety of data, commonly referred to as the three V’s. The first development is quite obvious, that is, data has now gained huge volume crossing the threshold of 2.5 exabytes per day and is doubling almost every three years (McAfee & Brynjolfsson, 2012). Velocity is the second very important development which is, in some cases, more important than the volume of data, for example, multimedia applications involving streaming of data and/or remote sensors monitoring. The third development is the variety in data formats that includes several structured and unstructured forms of data. For example, digital images saved in several different formats, different types of sensors readings in real-time, web server logs, unstructured text from social networks and website blogs. Big data has another attribute – veracity, that refers to the reliability of acquired data. Big data poses potential for business because standard computational tools and procedures are not designed to analyze such massive, multi-dimensional and dynamic data sets. In a traditional approach of extracting, transforming, and loading (ETL) of data, the data is often stored in one place, in table format with rows representing each record and columns as the different variables or parameters. In order to perform data analysis, data needs to be moved from hard drives to system memory (or cache) near to processing units. Whilst many of these techniques are still relevant in many applications, new approaches are required to process big data without losing time and energy in the ETL procedure.

One of the new approaches is Hadoop which is a collection of tools, architectures and systems. Unlike the traditional architectures, Hadoop implements a distributed file system (DFS) that does not recommend the ETL approach of combining different sources into a single database. Although there exist several components of Hadoop, the framework primarily consists of Hadoop Distributed File System (HDFS), Map/Reduce programming framework and Hive data warehouse (see Dean and Ghemawat 2008, and http://hadoop.apache.org/ for the full list of components). HDFS realizes redundant and reliable data storage on clusters consisting of low-cost machines/commodities. In a Hadoop world, big data is divided into small manageable blocks that are then stored at different places across the cluster. In order to avoid any loss of information, HDFS replicates these blocks as clones onto multiple machines. The number of these clones is usually configured according to the reliability requirements, and how much space is available for replication. This breaking up of big data into smaller blocks also opens up another possibility i.e. to divide computational resources into smaller chunks for processing – hence the capability of the Map/Reduce programming model, which provides a way of performing computations on data without moving it to central location or central database (Vaquero et al., 2008, Grossman et al., 2009). In other words, it breaks the ETL paradigm by moving computational resources near persistent data (also known as *data locality*). The principle of Map/Reduce programming framework is shown in Figure 1. In the Map stage, one or more data blocks are processed in a parallel way and intermediate results are stored in a distributed manner. In the Reduce stage, intermediate results are then processed for summarizing and sorting, and finally giving an output (Wu et al., 2009). Generally, the final output is also written back into HDFS along with the original data. The Map and Reduce functions are designed by application while the input and output of each function is described using <key, value> data format.

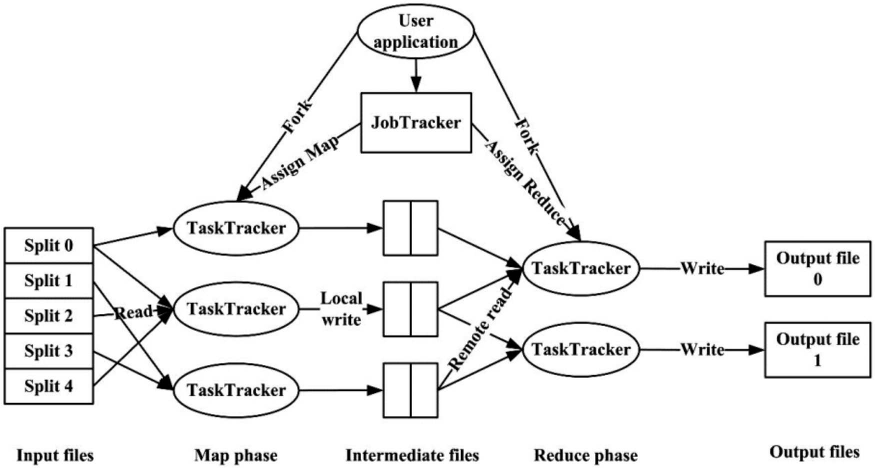

Figure 1. The principle of Map/Reduce programming framework

The third main component of Hadoop is HIVE – a data warehouse that is designed to achieve SQL-like data query on HDFS.HIVE data storage model supports high-efficiency access of massive data. Hadoop-aware applications exchange data via HiveQL language which is another SQL-like language designed specifically for HIVE. The significant advantage of HiveQL over SQL lies in massive data processing in On-Line analytical processing (OLAP) applications.

3. CCTP USER REQUIREMENTS

The CCTP system requirements were captured in two stages. The first stage was to identify the potential data fields for traceability from related literature, case studies and brainstorming methods. The second stage was to verify the results of the first stage through expert questionnaires (using Delphi method). The fields that were finally selected in the second stage became user requirements for the CCTP system. In the first stage, 14 papers (published between 2008 and 2013) on traceability data mining were reviewed. The agro-products described in those papers covered cod, tilapia, cocoa beans, pork and vegetables. The case study covered the cultivation, breeding, processing, logistics, distribution and consumption stages in fresh agro-product supply chain. A brainstorming was conducted with a group of 3 professors, 3 associate professors, 1 lecturer and 17 doctoral and master students in food safety, supply chain management and agriculture informatics fields. In the second stage, Delphi method was used with three rounds of questionnaires with selected experts. The final round converged to consensual decisions where experts agreed on the final set of user requirements.

From fresh agro-product supply chain management perspective, traceability data is useful not only for operational efficiency, but also for managerial and strategic decision making. Hence, user requirements for CCTP include three levels of data classification as shown in Figure 2.

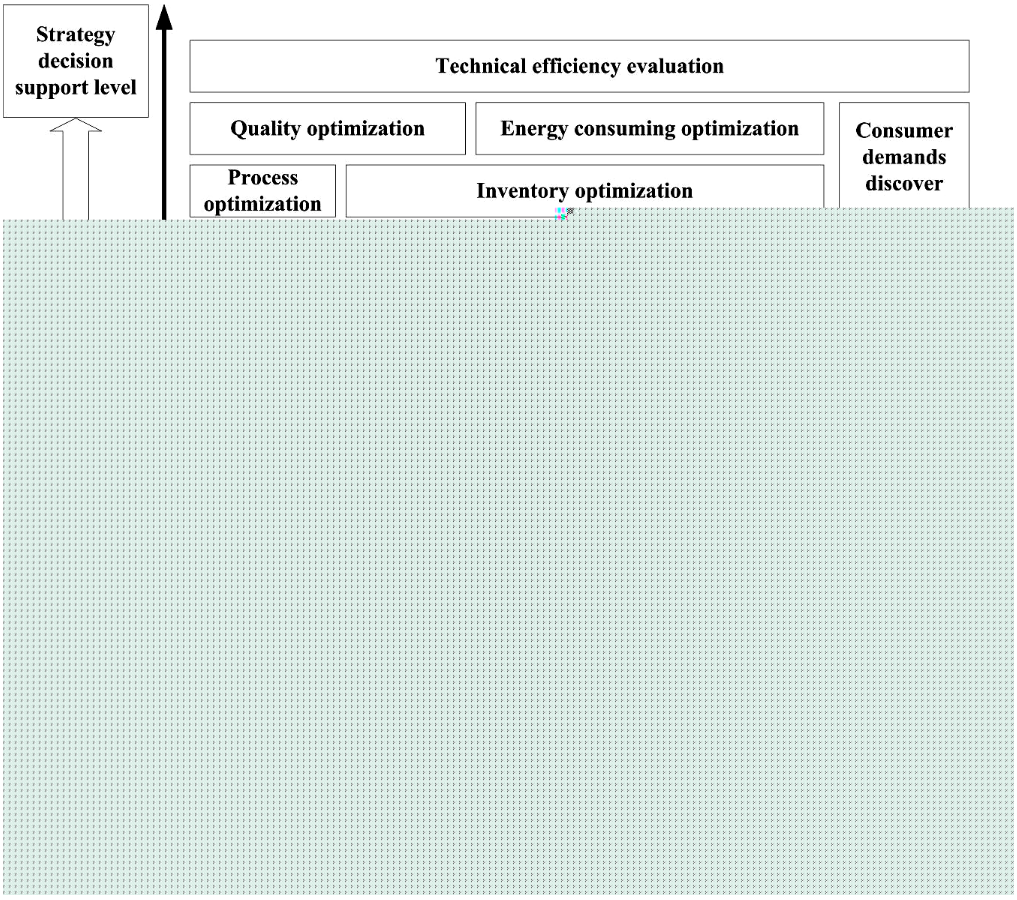
Strategy decision support level
Technical efficiency evaluation
Quality optimization
Energy consuming optimization
Consumer demands discover
Process optimization
Inventory optimization

Figure 2. Potential value and classification of massive traceability data

At the transaction level, the acquisition and application of traceability data can improve document standardization, hazard trace and precision recall. At the management level, the traceability data works together with food quality safety management model e.g. HACCP (Hazard Analysis and critical Control Point) to achieve quality control and early warning during cultivation, breeding, processing, logistic and inventory, and consumption phases, and at the strategic decision level, traceability data can support quality loss function and shelf life modelling that enables shelf life prediction and Least shelf-life first out (LSFO) inventory optimisation. The CCTP user requirements were finally summarised as documentation standardization, hazard trace, precision recall, logistics monitoring, critical point early warning, quality prediction, shelf life management and inventory optimization.

4. CCTP SYSTEM ARCHITECTURE AND MODEL DESIGN

4.1. CCTP System Architecture Design

The CCTP system is developed with three-layer architecture i.e. user interface layer (UIL), logic layer (LL) and data storage layer (DSL), as shown in Figure 3. The design realized high polymerization and low coupling between system modules and functions.

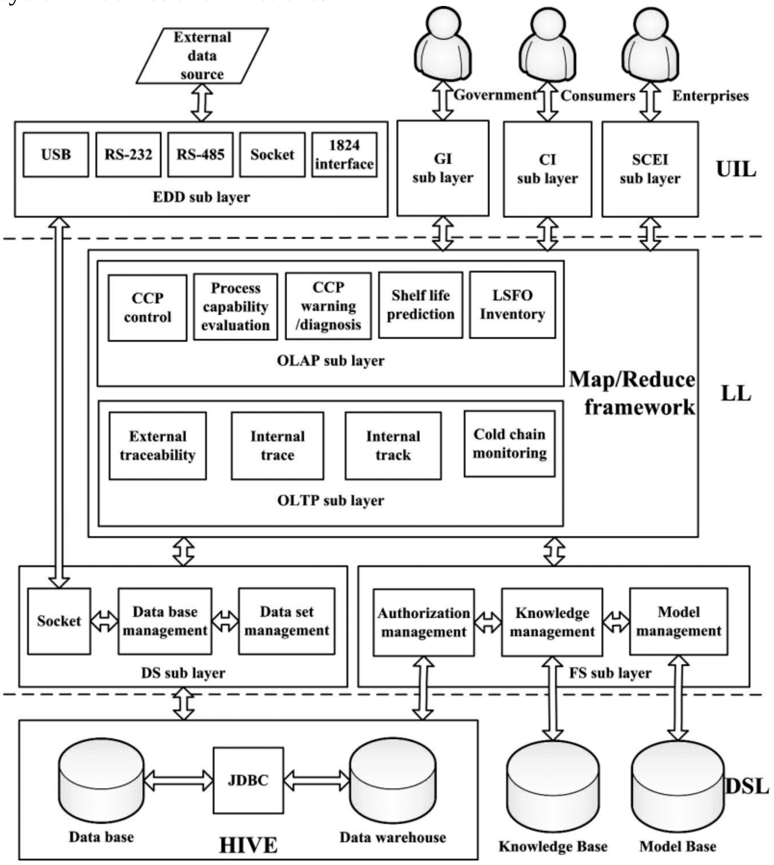

Figure 3. The CCTP system architecture design

User interface layer (UIL)

UIL provides the interface for data interaction between users and external data source and the logic layer. It is divided into external data driving sub layer (EDD), government interaction sub layer (GI), consumer interaction sub layer (CI) and supply chain enterprises interaction sub layer (SCEI). The EDD includes common external equipment driver to grab the external data (e.g. from USB, RS-232, RS-485, and Ethernet), and relay it to the logic layer. The GI provides rough granularity traceability data which contain certain data fields while meet the specific criteria paradigm which a government may demand. The CI gives consumers a querying interface for fresh agro-product traceability information. The information comes from the logic layer’s output. The SCEI layer provides interfaces, data and information output of functions within logic layer to all supply chain enterprises. The UIL also completes the user input data verification and validation based on pre-designed regular expressions.

Logic layer (LL)

LL consists of online transaction processing (OLTP) sub layer, online analysis process (OLAP) sub layer, function support (FS) sub layer and data support (DS) sub layer. LL is the core of the CCTP system’s business logic and functions. The OLTP functions include external traceability, internal traceability and cold chain real-time monitoring. The data source of OLTP is system database and real-time external data source (sensor generated real-time live data - temperature reading). The OLAP functions include SPC (Statistical Process Control) based critical point control and early warning, critical point’s process capability evaluation, FTA (Fault Tree Analysis) based critical point fault diagnosis, shelf life prediction and shelf life based LSFO inventory decision support. The data source of OLAP is CCTP system data base and data warehouse.

The FS and DS provide the system’s OLAP and OLTP maintenance and functional support. The FS includes modules for system authorization, and management business models and knowledge-base. The role oriented system authorization management module maintains an access control matrix which provides the system user-function authorization. The model management module calls the trace algorithm, track algorithm, SPC model, process capability evaluation model, FTA model, shelf life equations and LSFO inventory model in case of the CCTP system needs during user operation. The knowledge base management module maintains and calls the knowledge stored in the format of knowledge and rules in knowledge base such as SPC control charts, SPC defect rules and shelf life equation selection rules and so forth. The DS includes Socket interface module, data management module and dataset management module. The Socket interface module receives real time monitoring data from the EDD and then connects with the data management module so as to store it into database.

Data storage layer (DSL)

DSL provides real time data, historical data, metadata, knowledge and models for CCTP. The database stores real time data and metadata for OLTP functions while the data warehouse maintains the massive historical data for data mining and analysis usages in OLAP sub layer. The database and data warehouse are designed based on HIVE data storage structure. The knowledge base works with decision models for CCTP in the FS sub layer.

4.2 CCTP Decision Model

The fresh agro-product quality safety evaluation and shelf life prediction function process is shown in Figure 4. At the beginning, the CCTP system selects key indicators of quality loss from knowledge base. By reading the continuous temperature data reported from the cold chain sensors, the system determines the temperature condition according to the SPC defect rules and control chart. Once the key indicators and temperature conditions are calculated, the CCTP system continuously monitors the serial data received from temperature sensors. If there is a possibility to calculate quality decay in parallel using Map/Reduce framework, the calculation procedure is divided into the Map and Reduce phases. However, if not possible, the equation will be solved directly without using the distributed computing approach. This study simulated tilapia cold chain logistics where the key indicators: Total Viable Count (TVC) and Total Volatile Basic Nitrogen (TVB-N) were used to measure quality loss. The fresh agro-product quality decay results are comprehensively evaluated after solving all the equations and the shelf-life for each product is calculated based on these results.

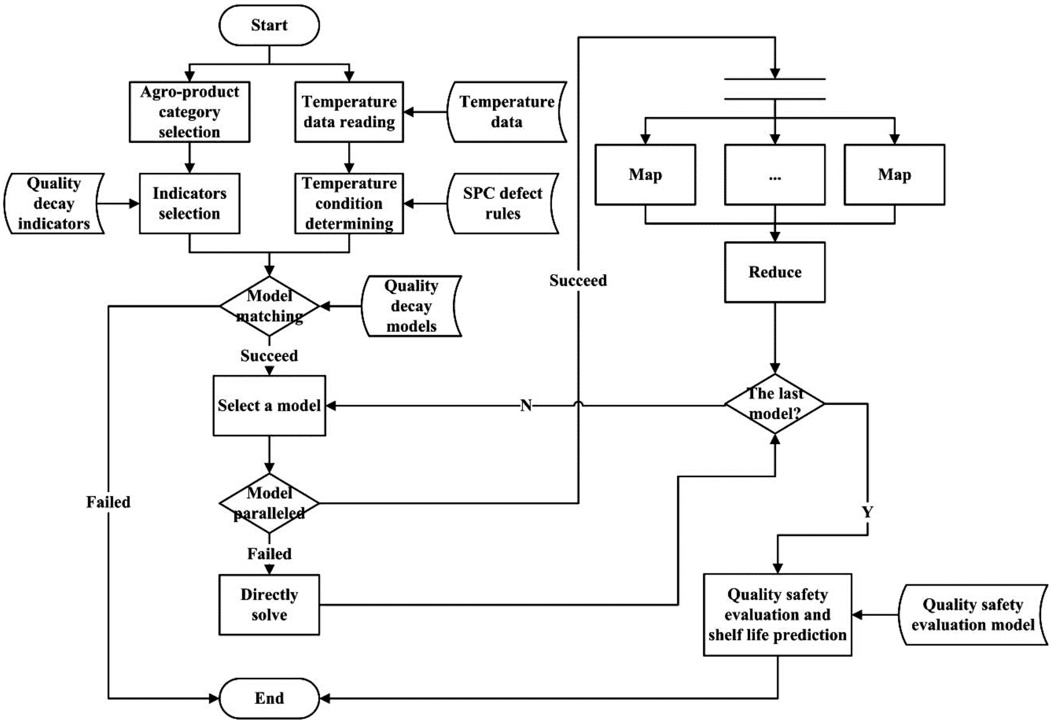

Figure 4. The fresh agro-product quality safety evaluation and shelf life prediction process

5. CCTP SYSTEM IMPLEMENTATION AND EVALUATION

5.1 CCTP System Implementation

The CCTP system was implemented and tested on three Linux-based machines which were connected with standard 100Mbps Ethernet link. The hardware specification of these machines is shown in Table 1. The system was built on top of Hadoop 0.20.0 parallel computing platform. The framework was developed using Java Development Kit version 1.6

Table 1. The specification for the prototype implementation of CCTP

No.	Host type	OS	IP address	Node type	Processer	Memory	Hard drive
1	ThinkPad X220	Ubuntu10.0.4	192.168.0.1	Host	Intel Core i7	4G	250G
2	Lenovo Yangtian M400	Ubuntu10.0.4	192.168.0.101	Slave	Intel Core 2	2G	160G
3	Lenovo Yangtian M400	Ubuntu10.0.4	192.168.0.102	Slave	Intel Core 2	2G	160G

The real-time monitoring functions were based on TCP/IP protocol in order to access fresh agro-product supply chain sensor data via EDD sub layer directly. This interface provides batch information and port information configuration functions. User can also access historical data that is stored in HIVE in a structured format.

The shelf life prediction function introduces the environmental sensor data from fresh agro-product batch information at the beginning and then sets the quality decay key indicator from the knowledge base according to sensor data distribution. The quality indicators and shelf life model are matched so as to determine the configuration of the shelf life model. The configuration parameter is set from the knowledge base directly or modified by users. Once finishing the model configuration, the shelf life is calculated by Map/Reduce framework and prediction result is outputted.

5.2 CCTP System Evaluation

The CCTP system is evaluated by 11 professionals who specialize in agricultural systems and knowledge engineering, agricultural informatics and enterprise management fields. Aquatic products cold chain logistics data is used for the evaluation.

The result shows that the CCTP is capable to handle large real-time data to support three levels requirements. On the transaction level, the CCTP system enables WSN based real time temperature data acquisition and monitoring during logistics and platform based batch information recording. On the management level, the batch records make it possible to trace and track information on the total aquatic products cold chain. On the strategic decision support level, the quality decay, shelf life prediction and LSFO inventory strategy are achieved based on the temperature data on the entire product chain. Meanwhile, testing the combination of temperature data and SPC and FTA model shows capability in evaluation and fault diagnosis on critical control points. The effect of the CCTP system in aquatic products cold chain logistics is depicted in Table 2.

Table 2. The effect of the CCTP system in aquatic products cold chain logistics

No.	Function Level	Function Point	Before Implementation		After Implementation	
			Method	Timeliness	Method	Timeliness
1	Transaction	cold chain temperature monitoring	Paper based	Time lag	WSN	Real time
2		cold chain batch management	Paper based	Time lag	Platform based	Quasi real time
3	Management	Total traceability	None	-	Platform based	-
4	Strategy decision support	quality evaluation and shelf life prediction	None	-	Model based	Real time
5		Inventory optimization based on shelf life	FIFO	-	LSFO	-
6		CCP process capability evaluation	None	-	SPC	Real time
7		Cold chain equipment fault diagnosis	None	None	FTA	Real time

Note: “-“means no information or not applicable

6. CONCLUSION

This study applies big data architecture to the field of fresh agriproduct supply chain. A Hadoop-based cloud computing architecture for real-time monitoring of food quality and shelf-life prediction has been designed, configured and tested. The conceptualized CCTP system achieved the functionality of receiving and storing massive fresh agro-product data (environmental and quality monitoring) acquired by WSN, and establishing a knowledge and model base for unified data mining and decision support, e.g. agri-product critical point control, and shelf-life prediction.

There are challenges in developing a full functional system of this type. Limitations must be noted, for example, the live temperature data is captured for a short period of time and is based on one sensor reading, only two quality decay key indicators TVC and TVB-N are selected for modelling shelf-life predication. In future, it is possible to improve this by introducing additional measures and more data from multiple sensors. Nevertheless, the system demonstrates the whole process of developing a Hadoop based big data processing and the knowledge bases and models relevant to fresh agro-product supply chain management.

REFERENCES

ABADI, D. J. (2009) Data Management in the Cloud: Limitations and Opportunities. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,* 32**,** 3-12.

AZUARA, G., TORNOS, J. L. & SALAZAR, J. L. (2012) Improving RFID traceability systems with verifiable quality. *Industrial Management & Data Systems,* 112**,** 340-359.

BO, Y. & WANG, H. (2011) The application of cloud computing and the internet of things in agriculture and forestry. *Service Sciences (IJCSS), 2011 International Joint Conference.* Taipei, IEEE.

CHEN, L., HU, Y., ZHANG, F., DUAN, W. & YU, P. (2013) Performance improving design on cloud computing for agricultural products safety traceability system. *Transactions of the Chinese Society of Agricultural Engineering,* 29**,** 268-274.

DEAN and GHEMAWAT (2008) MapReduce: simplified data processing on large clusters, Seminar paper, Google, Inc.

FENG, J., FU, Z., WANG, Z., XU, M. & ZHANG, X. (2013) Development and evaluation on a RFID-based traceability system for cattle/beef quality safety in China. *Food Control,* 31**,** 314-325.

GALVÃO, J. A., MARGEIRSSON, S., GARATE, C., VIÐARSSON, J. R. & OETTERER, M. (2010) Traceability system in cod fishing. *Food Control,* 21**,** 1360-1366.

GROSSMAN, R. L., GU, Y., SABALA, M. & ZHANG, W. (2009) Compute and storage clouds using wide area high performance networks. *Future Generation Computer Systems,* 25**,** 179-183.

IBM (2013). What is Big Data?, [Online]. Available from: http://www-01.ibm.com/software/data/bigdata/, [accessed March 2013].

KHER, S. V., FREWER, L. J., De JONGE, J., WENTHOLT, M., DAVIES, O. H., LUIJCKX, N. B. L. & CNOSSEN, H. J. (2010) Experts' perspectives on the implementation of traceability in Europe. *British Food Journal,* 112**,** 261-274.

McAfee, Andrew and Brynjolfsson, Erik (2012) Big Data: The Management Revolution, *Harvard Business Review,* OCTOBER

PAPETTI, P., COSTA, C., ANTONUCCI, F., FIGORILLI, S., SOLAINI, S. & MENESATTI, P. (2012) A RFID web-based infotracing system for the artisanal Italian cheese quality traceability. *Food Control,* 27**,** 234 – 241.

QI, L. (2014) A Study on the Traceability Oriented IoT’s Data Acquisition and Modeling Methods. Beijing, China Agricultural University.

QI, L., ZHANG, J., XU, M., FU, Z., CHEN, W. & ZHANG, X. (2011) Developing WSN-based traceability system for recirculation aquaculture. *Mathematical and Computer Modelling,* 53**,** 2162-2172.

SALTINI, R. & AKKERMAN, R. (2012) Testing improvements in the chocolate traceability system: Impact on product recalls and production efficiency. *Food Control,* 23**,** 221-226.

SALTINI, R., AKKERMAN, R. & FROSCH, S. (2013) Optimizing chocolate production through traceability: A review of the influence of farming practices on cocoa bean quality. *Food Control,* 29**,** 167-187.

VAQUERO, L. M., RODERO-MERINO, L., CACERES, J. & LINDNER, M. (2008) A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review,* 39**,** 50-55.

WU, J., PING, L., PAN, X. & LI, Z. (2009) Cloud computing: concept and platform. *Telecommunications Science,* 12**,** 23-30.

ZHANG, J., GU, Z. & ZHENG, C. (2010) Survey of research progress on cloud computing. *Application Research of Computers,* 27**,** 429-433.