Big Data Analytics and Visualization Techniques: A Case Study from Agriculture Domain

Van-Quyet Nguyen, Giang-Truong Nguyen, Sinh-Ngoc Nguyen, Jintae Choi, Kyungbaek Kim
Dept. Electronics and Computer Engineering, Chonnam National University
E-mail: quyetict@utehy.edu.vn, truongnguyengiang.bk@gmail.com, sinhngoc.nguyen@gmail.com,
jefron1100@gmail.com, kyungbaekkim@jnu.ac.kr

Abstract

In modern agriculture development, a huge amount of data is generated from many devices (e.g., sensors, robotic drones, etc.) and services (e.g., weather service, market price service, etc.). Leveraging these multiple forms of data might provide many benefits to farmers and agribusiness. However, these massive data are often complex and heterogeneity (e.g., structured, semi-structured, and unstructured), so it is very difficult to comprehend them. Due to these factors, big data analytics and visualization techniques play an important role in human understanding. This paper describes how we propose that big data can be efficiently used for creating agriculture services and supporting decision-making systems. We first discuss the challenges in agriculture big data. We then present a comprehensive study related to the techniques for collecting, storing, analyzing, and visualizing of big data. We show that our proposed techniques can be used to answer important questions in agriculture domain.

1. Introduction

Big data is a term that describes data sets with characteristics: high volume, high velocity, high variety, and high value [1]. It comes from many areas in our life such as public health, social networks, and agriculture. For example, in agriculture domain, a huge amount data is generated from many kinds of devices (e.g., temperature sensors, soil sensors, robotic drones, etc.) or agriculture services (e.g., weather forecast, market price, etc.). Besides, farmers can generate data during their growing products or manage their income by using traditional systems on a relational database. Thus, agricultural data becomes more and more complex with heterogeneity data structures (e.g., structured, semi-structured, and unstructured). It is necessary to study advanced techniques and technologies to enable the capture, storage, analysis, and visualization of the information.

There are a number of researches focusing on data analytics to provide the end users and experts with vision of the information contents. Jie Wang et al. in [2] has designed and implemented a platform for crawling and analyzing of agricultural product big data based on Jsoup. However, their platform only supported deploying on a single computer which encounters of the challenges of big data problem. Chen et al. in [3] proposed a crop breeding data analysis platform based on Hadoop [4] and Spark [5]. The platform consists of Hadoop distributed file system (HDFS) and cluster based on memory iterative components. A big data platform for collecting and analyzing agricultural big data has been proposed in [6]. The authors presented multiple choices of each phase of handling big data. But, this platform was lack of big data visualization module which represents data in human understanding.

Big data visualization plays an important role in creating a complete view of data and discover data values. Visualization techniques are used to create maps, tables, charts, and other forms to represent data. Most traditional data visualization approaches and tools are often inadequate to handle big data [7]. They are challenging to solve the limitations such as perceptual scalability, real-time scalability, and interactive scalability. In recent years, there are several researchers focusing on large-scale data visualization. Liu et al. in [8] proposed *imMens*, which is a browser-based system using WebGL for data processing and rendering on the GPU. Some tools with the functions of visualization and interaction for visualizing data are presented in [9]. Also, the extension of some conventional methods (e.g., Treemap, Streamgraph, etc.) to big data visualization is presented in [10].

In this paper, we present a comprehensive study related to the techniques for collecting, storing, analyzing, and visualizing of big data in the agriculture domain. We show that our proposed techniques can be used to answer important questions in agriculture, which provide a promising solution for developing strategies and managing services in agriculture based on big data.

2. Challenges in Agricultural Big Data

Agricultural big data is also characterized by 4Vs as mentioned above. These four characteristics cause many of challenges that organizations/farmers encounter in their data.

2.1 Dealing with data growth

Massive data in agriculture is generated in every minute through multiple kinds of devices and services such as sensors and agricultural web markets. For instance, a huge amount of images is generated in real-time through multi-devices during monitoring plant growth, which leads to the challenges in storing and analyzing all that information. Besides, much of data in agriculture is unstructured (e.g., images, documents, etc.) or semi-structured (e.g., JSON from weather services, HTML files from agriculture web markets), which is unsuitable to reside in a traditional relational database (e.g., MySQL, SQL Server). Managing such data is considered as a challenge.

2.2 Real-time scalability

It is important to provide users visual information in a real-time manner, which make their decisions to be faster and more efficiency. This requires agricultural big data platform need to support real-time handling data in all phases including collecting, analyzing and visualizing. For example, in recent years, e-government and e-commerce are in the early stages of development in many countries. In which, the market price is one of the public services that provide information about transactions (e.g., product name, price, amount, etc.) to the citizens or businesses. However, collecting such data from the web faces some challenges such as the limit number of HTTP requests for crawling simultaneously.

2.3 Performance guarantee

The huge volume and high variety associated with agricultural big data lead to challenges in performance guarantee. Two common metrics should be considered are speed and accuracy. To achieve high speed, we need to develop a new generation of ETL (i.e. Extract, Transform, Load) and analytics tools that dramatically reduce the time needed to generate reports. For obtaining high accuracy in the analysis results, identifying big data analytics techniques being suitable for solving a specific problem is very important.

2.4 Interactive scalability in visualization

Interactive visualization often helps us understand the insight of data faster and better. However, it takes a long time for processing and analyzing data before visualizing, especially in case of huge amount of data. For example, the users can make an action like “SELECT ALL” in the web interface, therein, output data is made by querying with “JOIN” operation from a few “big tables”, it could disrupt interaction. Thus, although there are already some existed data reduction strategies and query optimization techniques, the data for visualization could be still too large. Scaling interactive visualization is a major challenge in big data research.

3. Big Data Analytics Techniques

To explore how agricultural big data can be leveraged to bring benefits for farmers/agribusiness, we worked on a case study of agriculture in South Korea. Three different datasets are integrated to discover the interesting knowledge. The first dataset contains approximate 10GB of agricultural data of 16 regions in South Korea from 2015 to 2016, which includes information about farms, income, soil data, production, and other data. Two other datasets (weather data and market price data) were collected by our system. We will describe handling these datasets as follows.

3.1 Combining data collecting and storing methods

We separate input data into two kinds, the first one is real-time data from web pages (e.g., weather data, market price) and the second is historical data from archives (*.csv and *.xls files). For real-time data, we use Flume to collect them into HDFS. For historical data, whose volume is huge, we develop a MapReduce [11] based module to increase the speed of collecting data by gathering them in a parallel manner. Also, we use Sqoop to import data from *.csv and *.xls files into HDFS and export the results after analysis back to MySQL.

To provide quick random access to huge amounts of structured data, we propose using HBase, which is a distributed column-oriented database built on top of HDFS. Moreover, to support SQL command which is a common type of data analysis, we use Hive [12] which is a data warehouse infrastructure tool to process structured data in Hadoop.

3.2 Large-scale statistical techniques to the huge volume of data.

Traditional agriculture systems do not scale to the huge volume of data. For instance, to generate a report to answer the question: *who are the top 100 farmers obtaining the* *highest of income from growing onion*, we need to join information from three “big tables” including farmers, products, incomes. This work might take long response time. To solve such a problem, we propose using Hapdoop-based programmings such as MapReduce, MapReduce with Hbase or Hive. Also, Spark-based programming is a good choice for analyzing data which fits into distributed memory.

3.3 Machine learning-based big data analysis

In this section, we present major tasks in precision agriculture which can be enhanced by using machine learning techniques.

Using clustering and classification techniques to group and monitor quality of soil data. Soil quality plays an important role in determining how well plants grow. In fact, each type of plant adapts to different types of soil. Understanding the different properties of soil helps farmers to select the best plants for growing on their own farms. There are three main characteristics of soil: texture, structure, and chemistry. The values of them are changed by the time, and they can be captured by dedicated sensors. However, how farmers recognize these changes and distinguishable the types of soil are important questions. To answer them, in our study, we utilized clustering and classification techniques to categorize soil data. Specifically, we use Spark Mllib which contains many algorithms and utilities including classification: logistic regression, and naive Bayes, and clustering: K-means, Gaussian mixtures (GMMs). Of course, the results could be visualized for better understanding.

Using distributed parallel association rule mining to find out important associations related plants growing. Choosing which plants should be grown in a season is a very crucial question to farmers because it is a factor for manipulating their income and productivity. Usually, each farmer would choose a few main plants which they think those are most suitable for them. However, in order to get most benefits, they should also choose other types to grow with their chosen ones. This is a big question because among many plants (around 730 types of plants in our first dataset), finding the suitable one requires a lot of conditions such as high frequency growing by other farms and high production.

To support farmers do so, we used distributed parallel association rule mining techniques. First of all, in the pre-processing step, from farms’ information in the dataset, we get the so-called “transactions” which includes each farms’ corresponding grown plants collection, then frequent itemsets (plants) would be found. Then, we implement SON-based algorithm [13] based on Hadoop/MapReduce to find the association rules about the grown plants. In other cases, if the dataset is fit enough into distributed memory we can also use Spark MLlib with FP-growth algorithm to speed up the processing.

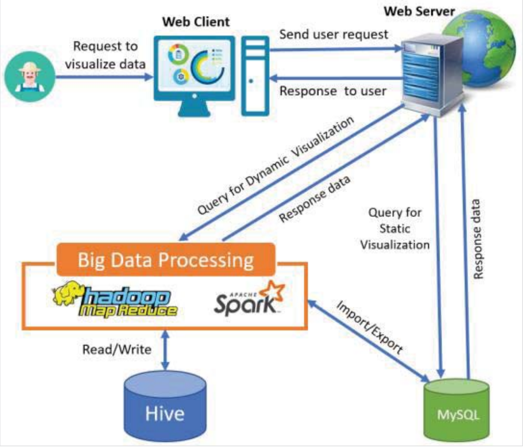
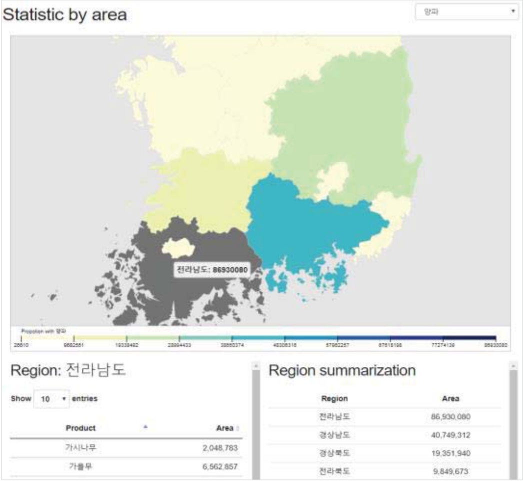


Figure 1. Overview of big data visualization platform


Statistic by area

Region: 전라남도

Show 10 entries

Product	Area
가시나무	2,048,783
가을무	6,562,857

Region summarization

Region	Area
전라남도	86,930,080
경상남도	40,749,312
경상북도	19,351,940
전라북도	9,849,673

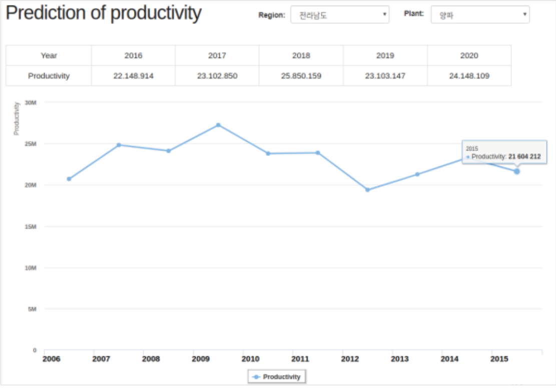
Figure 2. Statistic of the grown plants with Korea map visualization

4. Agricultural Big Data Visualization

An overview of the architecture of our big data visualization platform is illustrated in Figure 1. Firstly, users can make a request to see data visualization on the client web interface, then the request is submitted to the Web server. Secondly, the web server will check: if the request is a static visualization, it will query data on MySQL Server and show the result immediately; otherwise, it will be sent to Hive Server to get dynamic data from Hive warehouse. Here, data processing is handled by MapReduce or Spark engine. Finally, when web server gets the result from the static or dynamic query, it will respond to the web client and the response will be shown in a visual view.

Next, we give a brief overview of the tasks which a(n) user/expert can carry out using the big data visualization platform and describe how it was used in our case study.

Explore growing plants by region. When agriculture researchers encounter the integrated data for the first time, one of the important tasks they want is taking a look at an overview of growing plants by region. To accomplish that, we have created a region map (Korea Map in our case) as shown in Figure 2. It provides a quick overview of the growing plants of the region, such as types of plants and planted area.



Prediction of productivity

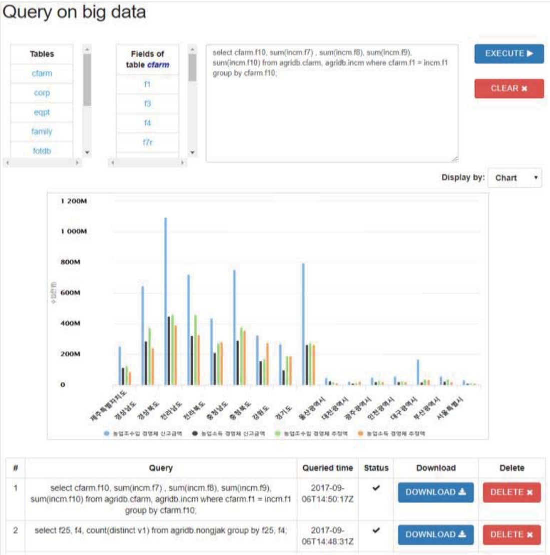
Region: 전라남도

Plant: 양파

Year	2016	2017	2018	2019	2020
Productivity	22.148.914	23.102.850	25.850.159	23.103.147	24.148.109

Figure 3. Prediction of plant production

Predict plant production. In fact, every farmer and agriculture manager would like to predict their plant production for the current season as well as the next seasons. To support that, we design a visualization module for predicting the production of each plant in each region. Also, a farmer can forecast the plant production on their farms. Figure 3 illustrates prediction of onion productivity of South Jeolla Province in South Korea.



Query on big data

Tables
cfarm
corp
eqpt
family
fotdb

Fields of table *cfarm*
f1
f3
f4
f7r

```
select cfarm.f10, sum(incm.f7) , sum(incm.f8), sum(incm.f9),
sum(incm.f10) from agridb.cfarm, agridb.incm where cfarm.f1 = incm.f1
group by cfarm.f10;
```

Display by: Chart

#	Query	Queried time	Status	Download	Delete
1	select cfarm.f10, sum(incm.f7) , sum(incm.f8), sum(incm.f9), sum(incm.f10) from agridb.cfarm, agridb.incm where cfarm.f1 = incm.f1 group by cfarm.f10;	2017-09-06T14:50:17Z	✓	DOWNLOAD	DELETE ✖
2	select f25, f4, count(distinct v1) from agridb.nongjak group by f25, f4;	2017-09-06T14:48:31Z	✓	DOWNLOAD	DELETE ✖

Figure 4. HiveQL-based big data visualization on the web

Explore agricultural big data by using dynamic visualization with Hive query. For experts researching on agricultural big data, dynamic visualization is one of the best ways to discovery great insight from big data. In order to do that, we develop a web-based visualization as shown in Figure 4. In this function, users could make a Hive query, then the query will be sent to Hive Server to be executed by MapReduce to get the result. Finally, the result is plotted in multiple views, such as chart, table, or file depend on its size.

5. Conclusion

This paper presented a comprehensive study related to the techniques for collecting, storing, analyzing, and visualizing of big data. We have shown that our proposed techniques can be used to answer important questions in agriculture domain. Our findings are based on an exploratory agricultural big data in South Korea. These findings provide a promising solution for developing strategies and managing services in agriculture based on big data.

Acknowledgement

This work was carried out with the support of "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ01182302)" Rural Development Administration, Republic of Korea. This work was supported by the National Research Foundation of Korea Grant funded by the korean Government(NRF-2014R1A1A1007734).

References

[1] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management35.2 (2015): 137-144.

[2] Wang, Jie, et al. "The crawling and analysis of agricultural products big data based on Jsoup." Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on. IEEE, 2015.

[3] Chen, Shuangxi, et al. "Analysis of plant breeding on hadoop and spark." Advances in Agriculture 2016 (2016).

[4] Hadoop, Apache. "Apache Hadoop". http://hadoop.apache.org (2017).

[5] Zaharia, Matei, et al. "Spark: Cluster Computing with Working Sets." HotCloud 10.10-10 (2010): 95.

[6] Van-Quyet, Nguyen, et al. “Design of a Platform for Collecting and Analyzing Agricultural Big Data.” JDCS vol. 18, no.1, pp. 149-158, 2017

[7] Agrawal, Rajeev, et al. "Challenges and opportunities with big data visualization." Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems. ACM, 2015.

[8] Liu, Zhicheng, et al. "imMens: Real- time Visual Querying of Big Data." Computer Graphics Forum. Vol. 32. No. 3pt4. Blackwell Publishing Ltd, 2013.

[9] Sucharitha, V., et al. "Visualization of big data: its tools and challenges." International Journal of Applied Engineering Research9.18 (2014): 5277-5290.

[10] Wang, Lidong, et al. "Big data and visualization: methods, challenges and technology progress." Digital Technologies 1.1 (2015): 33-38.

[11] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters. " Communications of the ACM 51.1 (2008): 107-113.

[12] Thusoo, Ashish, et al. "Hive: a warehousing solution over a map-reduce framework." Proceedings of the VLDB Endowment 2.2 (2009): 1626-1629.

[13] Savasere, A., Omiecinski, E. R., & Navathe, S. B. (1995). “An efficient algorithm for mining association rules in large databases”. Georgia Institute of Technology.