A Profile-Based Big Data Architecture for Agricultural Context

Soumaya LAMRHARI[1], Hamid ELGHAZI[2], Tayeb SADIKI[2], and Abdellatif EL FAKER[1]

[1]ENSIAS, Mohammed V University in Rabat
soumaya_lamrhari@um5.ac.ma, abdellatif.elfaker@um5.ac.ma
[2]International University of Rabat
hamid.elghazi@uir.ac.ma, tayeb.sadiki@uir.ac.ma

***Abstract*—Bringing Big data technologies into agriculture presents a significant challenge; at the same time, this technology contributes effectively in many countries' economic and social development. In this work, we will study environmental data provided by precision agriculture information technologies, which represents a crucial source of data in need of being wisely managed and analyzed with appropriate methods and tools in order to extract the meaningful information.**

Our main purpose through this paper is to propose an effective Big data architecture based on profiling system which can assist (among others) producers, consulting companies, public bodies and research laboratories to make better decisions by providing them real time data processing, and a dynamic big data service composition method, to enhance and monitor the agricultural productivity. Thus, improve their traditional decision-making process, and allow better management of the natural resources.

***Keywords*—Big data, precision agriculture, profiling system, decision making.**

I. Introduction

In the recent years, the huge volume of real time data in the agricultural sector and its need for an efficient and effective processing, stimulate the use of novel technologies and platform to acquire, store, process, analyze and visualize large data sets for future predictions and decision making. Big Data is an evolving term given to a wide area of data-intensive technologies in which the datasets are extremely large that dealing with them become more challenging than how it was before [1].

Due to the critical challenges facing the agriculture sector [2], farmers feel more forced to adopt intensive farming practices and sustainable agricultural ones, in order to increase both economic and environmental costs.

Being able to know where and when to apply fertilizers, meeting demand for food while maintaining soil fertility, predicting future climatic conditions, controlling pests and diseases that are affecting crops and livestock, monitoring plants growth and productivity, applying efficient and sustainable techniques to crop production, all of these represent great challenges to be overcome in the near future [3].

In this context, varieties of terminologies and techniques have been done to make agricultural practices more efficient, having as purpose to increase yields and productivity while optimizing the use of natural resources and reducing the negative impacts of intensive farming on the environment. Among these techniques we mention Precision Farming [4] (PF), Smart Agriculture, Global Positioning System (GPS) [5], and Geographic Information System (GIS) [6] etc., but the underlying concept in all of them is the same.

The main goal of this paper is to provide an elastic and variable Big data architecture based on a profiling system, which aims at providing to agricultural actors a dynamic Big data service composition and an accurate analytical method to help them retrieving meaningful information, in order to enhance decision making.

Our approach responds to the situation where many users with different interest shall work in a centralized platform such as Cloud Computing. By using a dynamic and accurate selection of Big Data services, the agricultural actors can exploit data in real time and with appropriate tools.

The remainder of this paper is organized as follows. Section II presents the related work. Big data and cloud computing application in Precision Agriculture will be discussed in Section III. Section IV describes the profile based architecture for agricultural big data description, its main components, and the methodological approach used. In section V, a case study based on our approach is presented and described in detail. Finally, Section VI provides a summary and upcoming perspectives.

II. Related Work

Precision farming (PF) is simply the information technology applied to agriculture [4]. It aims to optimize yields and investments by automatic and real-time monitoring of site specific environmental and soil conditions (e.g. soil type, fertility levels, etc.) using four technologies: remote sensing [7] (RS), geographic information systems (GIS), positioning systems (GPS) and process control.

Precision farming technique was sufficient for small-scale farms, it deals with a set of data coming from sensors, GPS, GIS limited to a few hundred meters for a specific crop land area. WSN architecture was particularly well adapted to meet the needs of precision farming. This kind of network is composed of a large number of spatially distributed sensor nodes, able to cooperate with each other using wireless communication. Regarding sensing, computing, processing and communication capabilities, we can continuously sense and transmit agricultural data to a base station where data can be stored, analyzed and observed in real time [8].

978-1-4673-8469-8/16/$31.00 ©2016 IEEE

The WSN architecture adopted in PF was limited to a very small field size compared to the average of agricultural fields that exist, it focused on processing data from a local area using a remote control station in order to help farmer to make decisions related to its own farm, without carrying about the other external requirements (e.g. logistics, concurrence, market prices, etc.).

However, the implementation of these technologies has highlighted some limitations for precision farming. The emergence of new needs such as real-time processing and analysis of collected data, predicting weather and climate changes at the right time, integrating logistics requirements in the agricultural data process since the acquisition phase until the production one, has open new track of research. Hence the need to go for the new trend of information technology such as big data and cloud computing is a very prominent case study to consider. Responding to this need, we propose a novel approach which mixes and matches between Big Data technologies, cloud computing and a new profiling system, within the same architecture, in order to boost agricultural production sustainably and yields in the years ahead.

III. The Main Purpose Of Using Big Data And Cloud Computing For Precision Agriculture

With the recent technological advances in data analysis, data processing and decision making tools, information technology for agriculture becomes an evidence. Actually, many farms are using these advanced technologies to facilitate crop management, minimize losses and maximize yields.

Our approach adds a new concept to these technologies in order to build an elastic and variable Big Data architecture based on the user's needs: the profiling system.

In this section, we will start by describing the main advantages of using Big Data and Cloud Computing in the context of agriculture. We will describe the profiling system in more details in the next section.

A. *Big Data For Agriculture*

Big Data [1] can be simply defined as the huge amounts of data in need for an in-depth analysis and processing with the use of additional computational support.

Agricultural practices, in all their forms, are responsible of the considerable data quantity (yield / production, remote sensing data, etc.), and they use a large number of external data (weather data, satellite images, etc.) to guide farmer decisions. Looking behind these examples, three characteristics are typically used to describe the big agricultural data phenomenon; Volume, Variety, Velocity [9]. In other words, data originates from agriculture can be considered as a Big Data by reason of its data variety with huge volume following high velocity.

The interest of agricultural Big Data is to cross, process and analyze the data produced from distributed and heterogeneous sources in order to generate added value applications that will meet the needs and requirements of various end users.

B. *Cloud Computing For Big Data Environment*

Cloud computing [10] is the ideal model for large data because of its endless scalability and resources used on-demand. It promises reliable services based on virtualized storage technologies.

Resort to a cloud computing technology in agriculture can give considerable solutions to analysts and decision makers [11]. Using the cloud, we can benefit from special bricks in Big Data management to collect and centralize data the maximum as possible regardless their sources, to make detailed analysis in order to obtain valuable data. This can be ensured by the use of new database models including mixed approaches between relational and non-relational (*NoSQL*) databases, also by a distributed architecture at processing level of unstructured data with the aim to spread the load over a large number of servers (cluster) using a total abstraction of subjacent parallelizing mechanism (*Hadoop*[1] principle) [12].

IV. Our Profile Based Architecture For Precision Agriculture

One of the main features of our architecture is that it behaves differently with each type of user, and it uses dynamically different Big data services based on the type of information requested. As depicted on figure 1, the service composition provider represents a crucial component in the architecture, because it is responsible of creating the user profile that will query the system as well as creating the related big data services that will be used to meet the requirements requested by this user.

The creation of suitable Big data services that can be adapted to each kind of user and any data source is our ultimate goal. To this end, the methodological approach that we will follow is composed of three main steps:

1) The description of users profile;

2) The selection of accurate Big data services for each profile;

3) The composition of the Big data services for each profile.

Each step of our methodological approach is briefly described as follows.

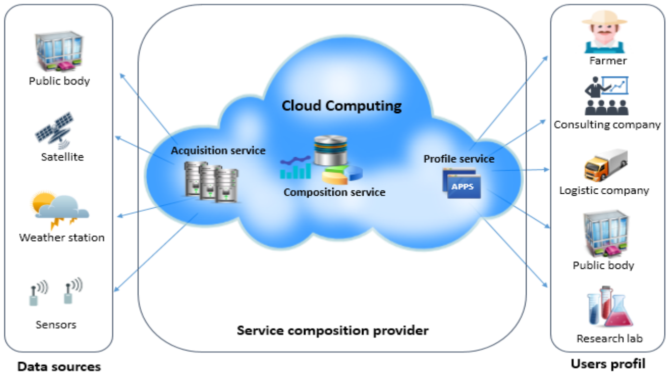


Fig. 1: a profile based architecture for agricultural big data

[1] https://hadoop.apache.org/

As shown in the figure 1, the datasets are generated from various sources such as sensors, weather stations, satellites and public bodies. The data gathering is quite a crucial task because it will help in bringing together these data sources before going through the storage process. In this work, we focus more on the Big data service composition and the profile service, while the acquisition service will be addressed in our future work.

A. *The Users Profile Description*

Due to the increasing use of ubiquitous technology including desktops, mobile devices, servers and sensor networks, data might be available at any time and from anywhere.

In this vein, we propose a profile based architecture (see figure 1) which is dedicated to manage various users according to their nature, by providing them meaningful information that can support the adoption of sustainable agricultural practices. This proposal interacts with multiple user categories depending on their access rules, and it offers a suite of services that let them benefit from real-time monitoring of data gathered from distributed fields. The users can have as profile: farmer, research laboratory, policy maker, public administration, consulting or logistic company, etc.

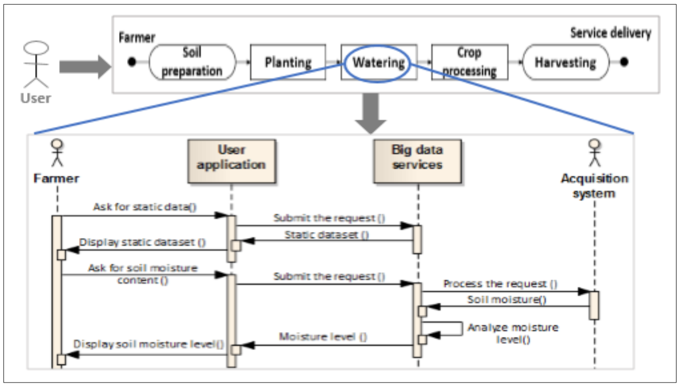


Fig. 2: The user specifications description

Within this step of our method, we shall understand the business process in a good manner and the phases through which passes the processed data flow. Because if we get any trouble describing the user needs from the beginning, the interactions conducted between actors and systems will be translated differently from what is specified by the user. Hence, the non-efficiency of the next two steps.

In the context of UML, a sequence diagram is a good way to visualize and validate various runtime interactions between actors and the system [13]. These can help to predict how the system will behave while describing the sequence of actions that will be carried out. For that reason, having a sequence diagram specification (see figure 2) allow us to know whether our system is compliant with the specification in the desired way, also the system should potentially be able to perform every behavior that the specification requires it to offer.

Among the various profiles pointed out above, we describe two types of user profile who are studied from our system:

a. Farmer Profile: Farmers are ever more users of online applications, as they are also asking for more mobility. In this vein, they can use their own devices in order to be connected to our profiling system. They can single out all relevant information related with the best crops for a given date and location, by using sensor, climate and weather data.

b. Lab Research Profile: research laboratory represents another kind of users who benefits more from our system. Generally, it is composed of a set of members like scientists, agronomists, biologists, etc. The goal of our system is to provide them a platform for high resolution crop analysis in real world growing conditions, which support their experiments. On the one hand, they can study the requirements of the cultivated plant, facing its growth and production goals, according to its vegetative stages. On the other hand, they can be able to follow weather events which vary from year to year, and pests which evolve in their location and behavior.

B. *The Selection Of Accurate Big Data Services*

Once we define the runtime actions between the users and the system, and before moving to create the correspondent big data services that provide ingestion, storage, processing, analysis, research and visualization of data, we need to identify various criteria to be taken into account, when choosing the right tools, and which meet the specifications required in the previous step.

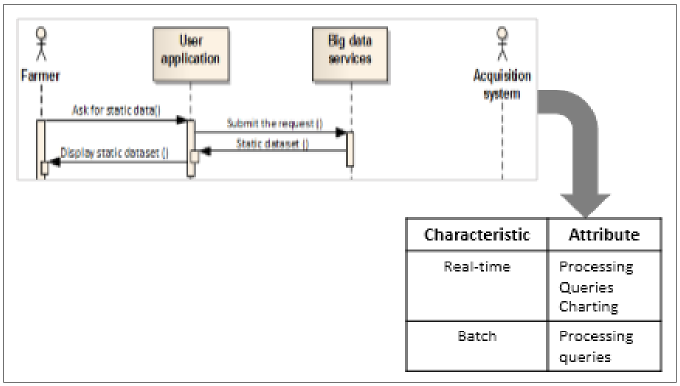
Here, we explain how we choose services from the specifications made by a user profile. In our context, we are dealing with a wide variety of agricultural data sources, so we can find: *i)* static data coming from some departments specialized in agriculture and environmental conditions like historical crop yield, soil conditions, land capability, suitability and vocation. These later are generally stored in relational databases, and they must integrate a big data system using efficient ingestion tools to be stored, analyzed and processed; *ii)* sensor, weather and image are represented as the persistent real time data, because they require real time processing in order to get better control at an appropriate time, and powerful analytics engine which leverages a distributed and parallel paradigm of Big Data processing (e.g, *MapReduce*[1], *Strom*[4]) to analyze large volumes of data, generating actionable insights rapidly. Additionally, these static and dynamic data require querying and scripting languages to simplify the use of *MapReduce* programs written in Java for a business user's perspectives. We can also note as a selection criterion, machine learning, because in some cases we will need to use a predictive model for example to manage pests and diseases. Finally, all these data types are requested to be available for exploration and visualization using flexible and real time tools.

As for the characteristics described in the table I, we note batch data processing which is an efficient way of processing high volumes of data, and it is where data could be gathered quietly over a period of time before being analyzed by packet. In contrast, real time processing is designed to act on real-time streaming data in a small time period (or near real time). Batch querying is the query process that can simultaneously analyze several queries at once. However, real time querying enables faster queries (e.g, input/output of key/value) and the operations run in real time on its database rather than *Mapreduce* jobs. Regarding the data access, it can be random read and write which enables to do faster lookups for analytics irrespective of data size. In the other hand, streaming read

TABLE I: Big data services and characteristics

Characteristic	Attribute	Sqoop	Flume	HDFS	HBase	MapReduce	Storm	Hive	Pig	Mahout	Elasticsearch	Kibana
Real-time	processing						X					
	queries				X						X	
	charting											X
Batch	processing					X						
	queries							X	X			
Data structure	Structured	X		X	X	X	X	X	X	X	X	X
	Unstructured		X	X	X	X	X		X	X	X	X
Language type	SQL							X				
	Data flow								X			
Data access	Streaming read			X								
	Random read/write				X							
Usage	Ingestion	X	X									
	Storage			X								
	Management				X							
	Processing					X	X					
	Analysis							X	X	X		
	Indexing and research										X	
	visualization											X

access provides continuous reads with a constant bitrate, as opposed to reading data as packets or chunks.



Characteristic	Attribute
Real-time	Processing Queries Charting
Batch	Processing queries

Fig. 3: Selection of accurate Big data services

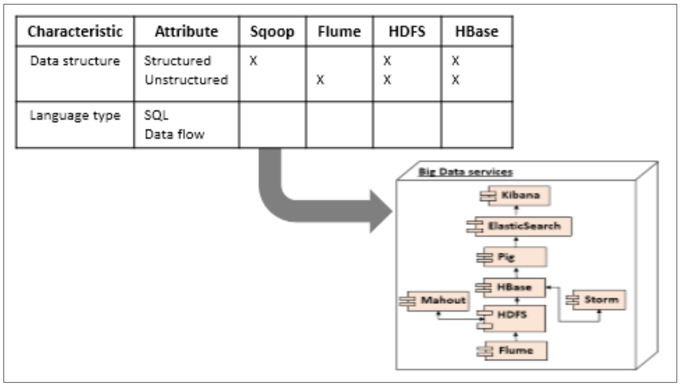
As depicted on figure 3, once we identify the needs of the user, we try to retrieve different features that correspond to the specifications given in each scenario.

C. The Composition Of Big Data Services

The data deluge that we address is likely to become even more acute as we opt for a new generation of tools. Based on the user specification, data type and its processing requirements, we propose a suite of tools in table I which contain the characteristics to be considered in order to process agricultural data in an efficient manner. And from this criteria table, we can conceive the component diagram containing the various Big data services required (see figure 4).

Among these characteristics we note real time processing, querying and charting; batch processing which is usually used for static and historical data; data flow and SQL languages; machine learning analytics; streaming reads and random access.

According to all these features, we can identify what is the most appropriate tool to a specific need, that is to say if we consider having Mahout machine learning application, it takes initiatives like real world data sets management in precision agriculture, where it can gain knowledge from historical unstructured data, uncover patterns to predict events that are hurtful for farming. Thus, these patterns and decisions may help landowners in the disaster management of their fields. Moreover, this predictive and analytics solution can ensure automatic irrigation and water management (e.g., as per soil moisture and new technology of irrigation) in real time to enhance best practices to crops.



Characteristic	Attribute	Sqoop	Flume	HDFS	HBase
Data structure	Structured Unstructured	X	X	X X	X X
Language type	SQL Data flow				

Fig. 4: Composition of big data services

V. A Case study: Farmer and Research Lab

In this section, our methodological approach will be further described and applied to two cases of study in order to show how we can provide a dynamic architecture based on big data components and tools. Farmer and research lab profiles were chosen as scenarios of profile to describe consequential operations of profiling system.

A. Farmer Scenario

Among the specific tasks done by farmers, we mention: understanding the implications of the weather and making contingency plans; buying supplies, such as fertilizer and seeds; as well as maintaining and monitoring the quality of yield, whether livestock or crops; knowing the variety of cultivated plants, conditions of its growth and its needs of seeds; choosing the type of fertilizer and pesticides, understanding their employment conditions and their impact on the climate-soil-plant; recognizing daily water needs for each kind of plant; calculating the median and mean values of yield; studying the conditions of natural environment; having knowledge about cultivation techniques employed and having the ability to estimate the financial revenue and manage the potential risks.

In order to discover the functionalities that we aim to reach using our proposal, we choose to treat as a case study the example of wheat cultivation. Generally, this later follow five steps before it becomes ready for consumption. These steps are explained in the following diagram:

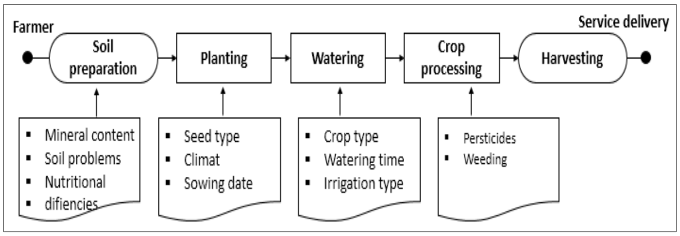


Fig. 5: Wheat cultivation steps

Taking into consideration the various steps of wheat cultivation process, we focus on one of them which is watering, and we put forward the related sequence diagram (see figure 6) to illustrate and clarify the interactions between actors and systems for a farmer profile needs.

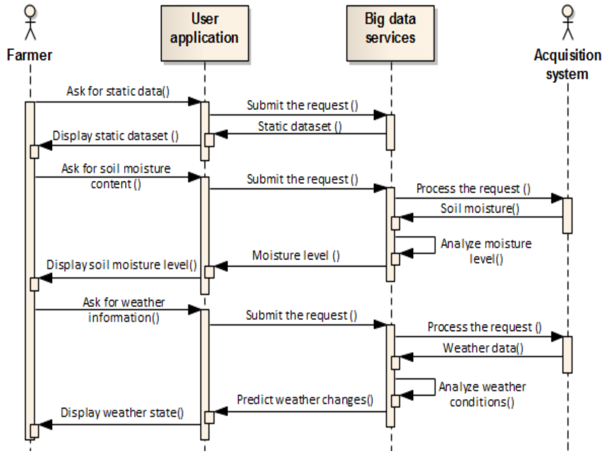


Fig. 6: Farmer sequence diagram for watering step

Based on the sequence diagram above and the characteristics mentioned in table I, the service composition provider will create the suitable big data services that respond to the farmer requests according to the selected criteria. The big data services created differ according to the nature of data processed and the type of processing and analysis required. Figure 7 shows all big data component used for a farmer profile.

In this case, the system component diagram is composed of diverse tools that deal with agricultural data in the different phases of data flow, from the acquisition step to the visualization and exploration one.

Static data requested by farmer are integrated to *HDFS* for storage using *Sqoop*[2], which is used to import data from external sources into related *Hadoop*[1] component. Unlike data coming from relational databases (*RDBMS*), sensor and weather data which are fully unstructured are collected from their origin and sent back to the storage location *HDFS*[1] through *Flume*[3].

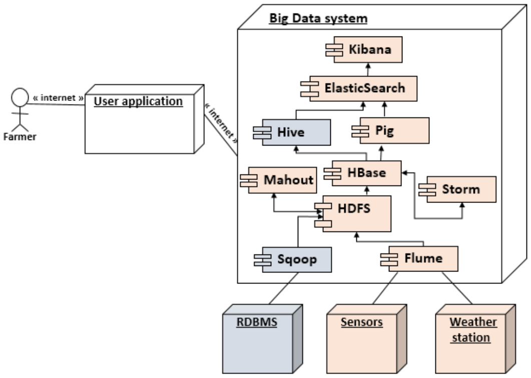


Fig. 7: System component diagram for a farmer profile

Then, once we have stored data in *HDFS* we cannot make changes. This is why we managed to work with *HBase*[1], a persistent *NoSQL* data store where we can make changes in our data also after writing it once, because *HBase* supports random 'read or write' access to big datasets.

Storm[4] is used for processing data in real time. In our case, we need to process sensor data in order to know soil moisture level, also we need to predict weather changes at the right moment, for these reasons we selected *Storm* as a real time engine.

Then, we have to query the processed data in hadoop. *Hive*[1] and *Pig*[1] can perform this function. *Hive* is used to process completely structured data as opposed to *Pig* which can be used for both structured as well as unstructured data. The difference between these two types of language is that *Pig* focuses on queries optimization aspect and provides control on the data flow more than *Hive*. *Mahout*[1] is needed to forecast weather changes (e.g. the farmer decides whether it is time to water the crop using irrigation techniques or not). Having data from various sources available through *Hive* and *Pig*, we need to get it into Elasticsearch[5] which is able to achieve fast search responses, instead of searching the text directly, it goes through an indexed search. It also holds the data, provides the analytics engine, and it is used as predecessor of *Kibana*[5] which provides in its turn the visualization rendering and the generation of queries into *Elasticsearch*.

Finally, the farmer can get clear graphs and dashboards proposed by Kibana tool to help him for making smart and accurate decision.

B. Research Lab Scenario

The present case study is related to a research lab profile. As an entity, the research lab has several activities to execute daily in an agricultural field. Among these later, we remain in the same example about wheat cultivation shown in the previous scenario, and we describe in figure 8 the different interactions performed to predict the disease X that can affect

[2] http://sqoop.apache.org/

[3] https://flume.apache.org/

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

[5] https://www.elastic.co/

the wheat crop. So, the research lab must be able to detect the right location of this disease as earlier as possible and limits its spread in the entire field [14].

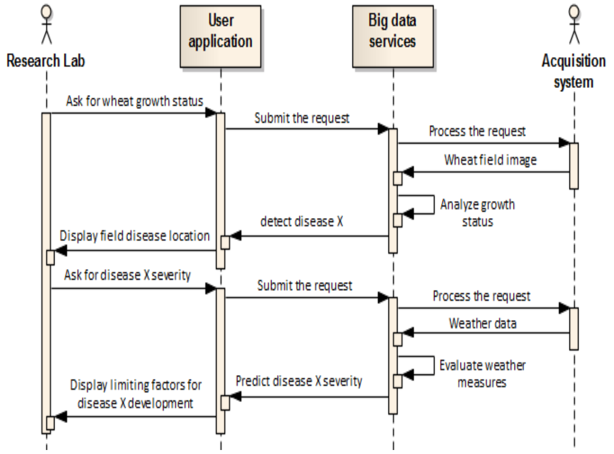


Fig. 8: Research lab sequence diagram for predicting crop disease

As mentioned previously, the big data system components required changes depending on the type of the data sources and the user's needs. The diagram below gives the composition of big data services in the case of research lab profile.

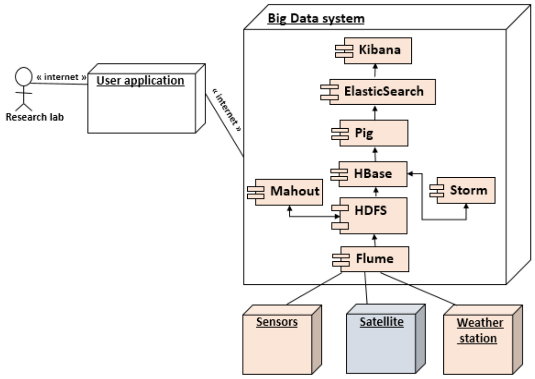


Fig. 9: System component diagram for a research lab profile

Regarding the research lab scenario, our system is faced to deal with dynamic data including satellite images, sensor and weather data. These data are invited to be stored in *HDFS* via *Flume*. Once stored, *Mahout* provides analytic methods and libraries to predict time and cause of future disease. *Storm* engine is used for real time processing. *Pig* is designed for querying the large datasets in *Hadoop*. *Elasticsearch* is used for the powerful search and analytic capabilities, while *Kibana* gives shape to any kind of data.

VI. Conclusion

In the present paper, we proposed a profile based approach to manage agricultural data within a cloud computing architecture. Our approach will guide the big data providers and the various agricultural actors to identify and select the best services adapted to their specific needs.

By using our approach, the agricultural actors can easily integrate the world of big data customers, and benefit from the advantages of big data technologies.

Actually, the work done in this paper is focused on the Big data service composition and the profile service identification and selection. In a future work, our approach will be extended by developing an application to automate the processes of selection and composition of big data services.

References

[1] R. D. Ludena, A. Ahrary et al., *Big Data approach in an ICT Agriculture project*, in Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA), 2013 International Joint Conference on, 2013, pp. 261–265.

[2] N. Alexandratos, J. Bruinsma, et al., *World agriculture towards 2030/2050: the 2012 revision*, ESA Work Pap, vol. 3, 2012.

[3] A. F. McCalla, *Challenges to world agriculture in the 21st Century*, Update Agric. Resour. Econ. Univ. Calif. Davis, vol. 4, no. 3, 2001.

[4] R. D. Grisso, M. M. Alley, P. McClellan, D. E. Brann, and S. J. Donohue, *Precision Farming. A Comprehensive Approach*, 2009.

[5] R. D. Grisso, M. M. Alley, and G. E. Groover, *Precision Farming Tools. GPS Navigation*, 2009.

[6] M. Neményi, P. á. Mesterházi, Z. Pecze, and Z. Stépán, *The role of GIS and GPS in precision farming*, Comput. Electron. Agric., vol. 40, no. 1–3, pp. 45–55, Oct. 2003.

[7] S. K. Seelan, S. Laguette, G. M. Casady, and G. A. Seielstad, *Remote sensing applications for precision agriculture: A learning community approach*, Remote Sens. Environ., vol. 88, no. 1–2, pp. 157–169, Nov. 2003.

[8] Aqeel-ur-Rehman, A. Z. Abbasi, N. Islam, and Z. A. Shaikh, *A review of wireless sensors and networks' applications in agriculture*, Comput. Stand. Interfaces, vol. 36, no. 2, pp. 263–270, Feb. 2014.

[9] S. Sonka and I. IFAMR, *Big Data and the Ag sector: More than lots of numbers*, Int. Food Agribus. Manag. Rev., vol. 17, no. 1, p. 1, 2014.

[10] S. Sakr, A. Liu, D. M. Batista, and M. Alomari, *A Survey of Large Scale Data Management Approaches in Cloud Environments*, IEEE Commun. Surv. Tutor., vol. 13, no. 3, pp. 311–336, 2011.

[11] M. Amini, N. Sadat Safavi, S. Sohaei, and S. M. Noorbakhsh, *Agricultural Development In IRAN Base On Cloud Computing Theory*, Int. J. Eng. Res. Technol. IJERT, vol. 2, no. 6, pp. 796–801, 2013.

[12] A. Fernández, S. del Río, V. López, A. Bawakid, M. J. del Jesus, J. M. Benítez, and F. Herrera, *Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks: Big Data with Cloud Computing*, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 4, no. 5, pp. 380–409, Sep. 2014.

[13] X. Li, Z. Liu, and H. Jifeng, *A formal semantics of UML sequence diagram*, in Software Engineering Conference, 2004. Proceedings. 2004 Australian, 2004, pp. 168–177.

[14] P. P. Jayaraman, D. Palmer,A. Zaslavsky, A. Salehi, and D. Georgakopoulos, *Addressing Information Processing Needs of Digital Agriculture with OpenIoT Platform*, Interoperability and Open-Source Solutions for the Internet of Things. Springer International Publishing, 2015. pp. 137-152.