Volume 1, 2025, Pages 12-22 *Journal of Smart and Sustainable Farming*
| Regular Article | Open Access

Foundations and Challenges of Big Data Analytics for Agricultural Systems

Dong Samuel Taye Galu
Department of Forestry, School of Agriculture and Natural Resources Management, Mattu University, Metu, Ethiopia.
samuel.taye@mau.edu.et

Article Info
Journal of Smart and Sustainable Farming
https://www.ansispublications.com/journals/jssf/jssf.html

© The Author(s), 2025.
https://doi.org/10.64026/JSSF/2025002

Received 15 November 2024
Revised from 30 December 2024
Accepted 28 December 2024
Available online 08 February 2025
Published by Ansis Publications

Corresponding author(s):
Dong Samuel Taye Galu, Department of Forestry, School of Agriculture and Natural Resources Management, Mattu University, Metu, Ethiopia.
Email: samuel.taye@mau.edu.et

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract – Various methodologies have been used by agriculturists, agribusiness entities, organizations, and scholars to gather and consolidate such data. Subsequently, the collected data undergoes modification, often transitioning from a quantitative to a qualitative form. The primary objective is to obtain valuable insights from it, which could be utilized by end users and farmers to enhance their operations and enhance their likelihood of achieving success. The aforementioned factors include precise crop forecasting, precise farming techniques, intelligent agricultural practices, cultivation of superior quality seeds, and accurate meteorological and environmental predictions. To succeed in these specialized markets, it is essential to acquire proficiency in various big data analytic methodologies, such as machine learning, clustering and classification, predictive analytics, time series analytics, recommendation systems, data mining, and regression analytics. The aforementioned issues have been the subject of discourse. Furthermore, a comprehensive integration of several big data analytic approach and their application in the sector of agriculture has been accomplished. However, novel technology often come with significant challenges. The present study has investigated the challenges associated with the implementation of big data analytics within the agricultural industry, as well as strategies for further enhancing its use in this domain.

Keywords – Big Data Analytics, Information Extraction, Knowledge Management, Data Acquisition, Predictive Analytics, Intelligent Crop Recommendation System.

I. INTRODUCTION

Big data analytics is used to analyze vast amounts of agricultural data in order to improve decision-making and optimize farming techniques. The increasing popularity of data analysis in precision agriculture is mostly attributed to its ability to swiftly and consistently discover patterns and trends. This analytical approach facilitates the estimation of crop yields, determination of optimal growth conditions, and development of more accurate management plans. Precision agriculture plays a crucial role in the agricultural sector's shift towards sustainability, since it heavily relies on technology to implement focused and data-driven interventions that enhance crop output while reducing resource wastage. The use of big data analytics in precision agriculture has the potential to enhance the economic efficiency and improve the overall quality of crop yields for farmers. Ultimately, this initiative facilitates the augmentation of farmers' revenues and engenders a sense of confidence among inhabitants about their ability to get nourishing sustenance.

Agriculture is often seen as a fundamental pillar of a nation's economy due to its crucial role in sustaining and bolstering other sectors. It is also among the first occupations that individuals have pursued. Farmers largely depend on their cognitive abilities and the knowledge acquired through personal experience while making judgments. Consequently, the region experiences a series of unforeseen natural calamities, including climate change, insufficient or delayed monsoons, droughts, and floods. The insufficiency of support from institutions and governments in terms of enhanced agricultural programs, lending facilities, and other forms of aid for the agricultural sector. However, the current juncture presents an opportune time for technology to assume control of the transformation and provide a viable resolution.

The perception of "Big data analytics" is widely recognized in the field. It refers to the intricate procedure of analyzing extensive and diverse data sets, commonly referred to as big data. The primary objective of this process is to unearth valuable

data, such as unfamiliar correlations, customer preferences, market trends, and concealed patterns. These insights play a crucial role in enabling organizations to make well-informed decisions pertaining to their business operations. A shift may be seen among agriculturalists as they increasingly choose for advanced analytic modeling and decision support tools, which are guided by up-to-date scientific research, therefore progressively abandoning traditional approaches. Farmers possess a comprehensive understanding of the huge quantities of information readily accessible to them. Consequently, farmers and the agricultural enterprises catering to their needs are confronted with a substantial volume of data, including both structured and unstructured formats.

The objective of this study is to investigate the utilization of Big data analytic techniques in the evaluation of farm information, including information from various sources such as corporate agronomical data, satellite-drone imagery, public data from government agencies, field-level sensors, prevailing consumer-based data, weather station data, and historical data from various growers and conditions or factors of growing. The aim is to enhance farmers' decision-making processes by providing them with more comprehensive and informed insights. The remainder of the article has been organized as follows: Section II discusses the concept of agricultural data. Section III reflects on big data analytic techniques within the agricultural sector. Section IV reviews the implementation data analytics techniques used in the sector. Section V presents the challenges of big data within the agricultural field. Section VI draws final remarks to the article, as well as directions for future research.

II. CONCEPT OF AGRICULTURE DATA

The concept of the value chain was first introduced by Tardi [1]. The fundamental concept is that a corporation enhances the worth of a commodity via the acquisition of resources and their transformation into a finalized product. The financial performance of a corporation is positively correlated with the extent to which it generates value. Furthermore, a firm might get a competitive advantage in the market by prioritizing the happiness of its customers. Kien NGUYEN et al. [2] have posited a similar viewpoint, contending that the data-value chain serves as a conceptual framework delineating the sequential processes involved in extracting value from prior data. **Fig. 1** illustrates the segmentation of this framework into its component elements, namely data collection, extraction, business insights, and management, analysis.

Data Acquisition

The initial phase in data analytics is collecting the necessary information. It is made to gather information from a broad variety of sources, both structured and unstructured, such as soil conditions, weather reports, satellite photos, and more. Therefore, filters must be programmed to collect just the relevant data sets and to discard all other information. In addition, appropriate metadata must be generated for every data collection in order to clearly outline how the data may be further analyzed and rendered. Using the constructs "date, "time, "file-name," and "geolocation," for instance, one may create 2019 weather report information.

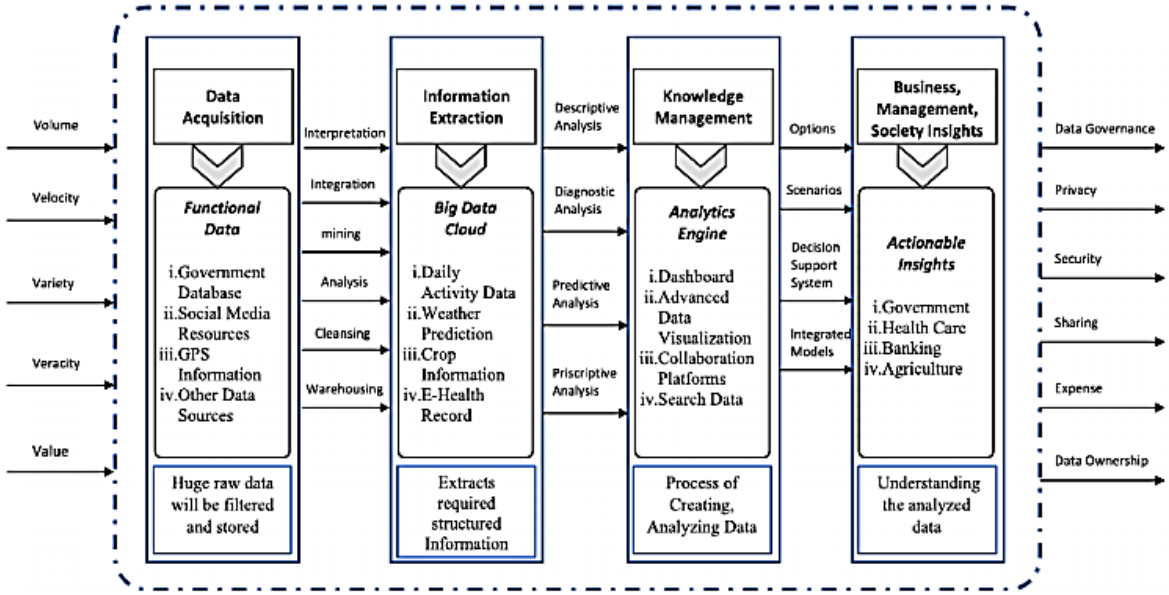

Fig 1. Frame Work Segmentation of The Components of Agricultural Data.

Information Extraction

In the majority of instances, the data that has been gathered is not yet prepared for the process of analysis. The data required for analysis is meticulously chosen throughout the process of information extraction. Filatov et al. [3] presented instances of information extraction in practical use. The information extraction process involves many essential processes, including data interpretation, integration, mining, analysis, cleaning, and warehousing.

Knowledge Management

The use of data analytics contributes to the improvement of decision-making processes. Hughes [4] asserts that advanced analytical techniques such as Online Analytics Processing (OLAP) and Map Reduce are utilized in the big data analytics domain. Extensive statistical samples, such as those facilitated by Big Data, contribute to the attainment of more precise and reliable results. In other words, it may be said that the reliability of the findings increases proportionally with the size of the sample. This research may use descriptive, diagnostic, predictive, and prescriptive analytics techniques in **Table 1**.

Table 1. Knowledge Management Techniques

Descriptive Analytics	The primary objective of descriptive analytics is to provide an account of the events or occurrences that transpired. Al Qundus, Gupta, Abusaimeh, Peikert, and Paschke [5] contributed to the elucidation of a prescriptive analytics-driven automated system in the field of agriculture, shedding light on the interconnections between agricultural practices, laws, and the realm of data science. Descriptive analytics is used to characterize the accuracy or inaccuracy of the result.
Diagnostic Analytics	The inquiry on the cause or rationale behind a certain phenomenon may be addressed via the implementation of a diagnostic investigation. Tang, Lenzen, McBratney, and Maggi [6] examined the many variables that lead to the extensive dispersion of pesticides among agricultural regions in Canada. The primary objective of this kind of analytics is to assess the possible risk associated with a dataset.
Predictive Analytics	Predictive analytics has the capability to forecast future events or outcomes. Cleary et al. [7] provided a description of an outbreak involving Salmonella Bareilly, a food-borne pathogen. The study elucidated the considerable association between the consumption of bean sprouts and the occurrence of this disease.
Prescriptive Analysis	A prescriptive analysis is advised for identifying actions that may be taken to mitigate future concerns. According to the research conducted by Martins, Opedal, Armbruster, and Pélabon [8], it is recommended to plant seeds in alignment with the prevailing rainfall patterns. The study also established a correlation between monsoon and post-monsoon precipitation.

Business Insights

The last stage involves the conversion of unprocessed data into valuable insights. In this context, the use of sensor data on the farm serves to optimize feed efficiency and mitigate the risk of crop failure in real-time. The use of Big Data enables the acquisition of predictive insights pertaining to future results in the field of agriculture.

The use of sensors in modern agricultural practices yields a substantial volume of data pertaining to many aspects such as soil composition, crop growth, intercultural management techniques, crop distribution patterns, and harvesting methodologies. Official organizations collect and administer databases containing a diverse range of data, including longitudinal census records, weather patterns, and other forms of information. AgMES, farmOS, agribusiness data, and other similar resources serve as only a few examples. The definition of the data infrastructure should be such that it has the capability to accommodate several independent modules, each of which may use a distinct set of application programming interfaces (APIs) and services. The system should possess sufficient flexibility to accommodate the integration of new modules without necessitating substantial adjustments to the foundational instructions. The agricultural sector use the tools and databases indicated above to collect data, which may be further utilized by the sector's data infrastructure to identify trends, abnormalities, and other deviations.

The use of big data within the agriculture industry has significant potential. The use of big data analysis has the capacity to enhance crop quality and output via the identification of patterns and trends that may remain undetected by traditional agricultural approaches. Hence, this technology may be utilized to analyze and gather massive quantities of data from many agricultural procedures. The integration of big data into the field of agriculture has yielded significant benefits across several domains, including benchmarking, analytics, model prediction, visualization, marketing, and management. Additionally, there are two noteworthy elements, **Table 2**, that need consideration.

The topic of precision agriculture has many prospects for visualization. Charvat et al. [9] present two methods: a 3D cooperative visualization of different zones of yield productivity and a visualization of agricultural equipment, both using the data modeling ideas. Both techniques of visualization use the newly released HS Layers NG tool, which may be accessed at https://github.com/hslayers/hslayers-ng. This particular software application enables users to see data in a three-dimensional manner, while also offering functionalities for searching and visualizing RDF (Rich Description Framework) data. By monitoring the spatial and temporal distribution of machines, valuable insights may be gained on the consequential outcomes resulting from human activities. The consideration of a farmer's fuel consumption, flight efficiency, and other economic indicators has significance in the management of tax revenue and government subsidies. In contrast, ecological initiatives aim to mitigate the environmental impact caused by human actions, such as the production of excessive carbon

dioxide resulting from inefficient transportation systems or the contamination of water sources due to excessive fertilizer leading to nitrogen pollution.

Table 2. Elements Business Insights to Consider	
Push factor	The use of smart devices throughout the agricultural value chain is a significant driving factor. The seamless integration of various components inside the system ensures constant accessibility to data, enabling its comprehensive analysis for diverse insights. In summary, this process enables more effective decision-making.
Pull factor	The primary emphasis is on offering business-oriented solutions. One potential strategy for achieving this goal is to allocate resources towards the adoption of advanced big data technology. The use of "smart agriculture" has the potential to significantly reduce costs and enhance the efficiency of agricultural operations. The use of big data has been employed to enhance several domains, including security, safety, productivity, and quality.

III. BIG DATA ANALYTIC TECHNIQUES IN AGRICULTURE

The research inquiries in this investigation were addressed by an exhaustive examination of existing scholarly literature. This examination of the literature is structured around two primary areas. The first stage involves doing a thorough examination of the prevailing analytical methodologies and technologies that have shown efficacy in the agricultural domain. The terminology used in the field includes terms such as precision agriculture, recommendation systems, spatial analytics, smart agriculture, and data-driven agriculture. On the other hand, the second stage prioritizes novel approaches for resolving its challenges and broadening the scope of prospective investigations.

Predictive Analytics

Within the realm of big data analytics, this particular approach is used to provide predictions about future events via the examination and analysis of historical data. This breakthrough has the potential to significantly transform several industries. The primary components of the methodology consist of three prominent approaches, including regression analysis, decision trees, and neural networks. In addition to the aforementioned methodologies, several additional techniques are used in this context, such as random forests and ensemble models. The domain of predictive analytics highlights the need of generating accurate fit statistics throughout the process of constructing models. This particular field of statistics offers a viable approach to resolving intricate business difficulties. Agricultural applications include the prediction of crop yields and analysis of customer buying behaviors. **Table 3** presents the forms of predictive analytics, exemplifying a few instances.

Table 3. Different Forms of Predictive Analytics	
Decision trees	This approach use categorization algorithm to modify the anticipated risks or rewards associated with certain activities, hence impacting the graphical representation of outcomes intended for human comprehension. A decision tree is comprised of a root node, which serves as the initial point, a series of branching questions, and a collection of leaf nodes representing several potential outcomes.
Simple Statistical Modelling	Statistical modeling, ranging from rudimentary mathematical models to intricate deep learning models, is elucidated and implemented in the realm of predictive analytics. However, within the field of predictive analytics, the approach that is predominantly used is the multiple linear regression model. Multiple linear regression may be used to develop models that anticipate future trends and estimate the effect of future changes in dependent variables.

Recommendation System

The informational system offers outcomes derived from the analysis of behavioral patterns and other functional data. The suggestions provided by a recommender system are often the outcome of its methodology and classifications. This software may be conceptualized as an intelligent tool designed to enhance decision-making processes. Recommendation systems used within the agricultural sector leverage advanced data techniques such as hybrid filtering, content-oriented filtering, Apriori with association rule, and collaborative filtering. These algorithms are employed to analyze extensive datasets and provide crop recommendations by considering many criteria such as weather conditions, soil characteristics, and product demand. Recommendation systems may be classified into three main categories as shown in **Table 4**.

Table 4. Classification of Recommendation Systems	
Content based	When formulating suggestions, the content-based method primarily depends on the analysis of similarities and shared attributes. In this scenario, the extent of resemblance is influenced by preexisting preferences or heuristics. The primary emphasis of a content-based recommendation system is in providing recommendations that are derived from the intrinsic characteristics of the textual material. Various techniques may be utilized to determine the utility forecast according to the present data, such as Bayesian classifiers, clustering, and neural networks.

Collaborative Filtering (CF)	This system may be classified as a recommendation system that utilizes user reviews to evaluate and prioritize different attributes of a product or service. In essence, it might be characterized as a kind of societal filtering. The fundamental concept is on assessing the likelihood of repeat approval from a cohort of customers who have previously expressed satisfaction with a certain product or goods. By using trends in users' evaluations, collaborative filtering has the potential to provide more precise outcomes. Both model-based and memory-based techniques are considered to be viable alternatives. Two examples of techniques used in collaborative filtering (CF) include matrix factorization and neural networks, namely the K-nearest neighbor (clustering) approach.
Hybrid	The proposed approach might be characterized as a fusion of collaborative filtering, and content-based filtering techniques. These factors mitigate the limitations associated with both approaches and provide more optimal outcomes for the recommendation system. Various choices are available, including the integration of features, the introduction of novel elements, the use of a combination of approaches, the implementation of cascading techniques, and the adoption of switching strategies.

Data Mining

The application of data mining is significant in the field of agriculture. There are different data mining approaches, which may be effectively used within the agriculture industry. Specifically, techniques such as pattern mining may be used to discover patterns within extensive datasets. The examination and evaluation of soil characterisation, for example, involves the use of K-means clustering and GPS-based technologies in the context of precision agriculture and farm management. **Table 5** provides a more comprehensive explanation of several strategies.

Table 5. Strategies of Data Mining

Association	The association mining approach is widely used as a tool for mining data. This technique of data mining discovers a pattern by exploring connections between actions or occurrences. This approach is employed in the analysis of basket market to investigate the habits of shoppers.
Classification	Classification is a prevalent approach in the field of machine learning that is often used for data mining purposes. The data was subjected to mathematical analysis in order to classify it into several groups. Examples of methods used in several academic disciplines include decision trees, linear regression, and statistical analysis.
Clustering	Clustering may be seen as a data mining technique that has its roots in the field of machine learning. In this particular instance, a collection of objects with like characteristics has emerged. In the context of data analysis, clustering involves the process of categorizing objects based on their inherent characteristics and then assigning them to appropriate groups. On the other hand, classification entails the assignment of items to preset categories or classes.
Prediction	Prediction data mining is a technique used to reveal the interconnectedness between both known and unknown variables within a certain context. Various sorts of data, such as sentiment analysis and text mining, may be used to make predictions.
Sequential Patterns	Sequential pattern mining, sometimes referred to as pattern mining, is a prominent data mining technique used for the purpose of uncovering recurrent structures, patterns, or transactions within a given dataset. Data mining jobs may be classified into several categories, and sequential pattern mining is one such type.

Spike and Slab Regression Analytic Technique

The Bayesian approach is used to describe the probability distribution of the regression coefficient in extensive datasets of linear regression. The discussion is on the regression coefficients associated with the "spike" and "slab" types. The use of the analytic technique offers some notable advantages for addressing issues in three-dimensional contexts. In the field of agriculture, it is often used for the purpose of selecting tools and models.

Time Series Analytics Approach Based on Big Data

The forecasted outcome of a time series analytical model is determined by using past observed data. A time series alludes to a gathering of different points of data, which are organized in a chronological order. In this context, time is seen as an independent variable with the aim of forecasting future events. This tool serves as a valuable instrument for forecasting fluctuations in market prices and the movements of crop and plant prices.

IV. IMPLEMENTING DATA ANALYTIC TECHNIQUES INAGRICULTURE

The afore mentioned technologies possess several potential applications within agricultural contexts, whereby they might be used to enhance productivity and reduce labor requirements. The afore mentioned examples have been cited.

Intelligent Crop Recommendation System

This big data analytics-based system and machine learning has a wide range of applications, including decision-making, crop recommendation, and forecasting. This approach takes into account many factors such as soil quality, precipitation, latitude, and temperature. The recommendation model is often categorized into two interrelated components. The crop predictor and the rainfall predictor are components within a broader system.

Crop Predictor

The use of data mining techniques has potential in the examination of agricultural productivity. Crop yield forecasting plays a crucial role in the field of agriculture as it enables farmers to make informed decisions about their plantings by providing them with valuable insights into what they may expect to harvest. Historically, cultivar prediction techniques were mostly based on the experiential knowledge of farmers on specific parcels of land and the crops that exhibited optimal adaptation to those environments. Various data mining techniques are used and evaluated in order to provide anticipatory forecasts on crop output in the field of agriculture.

The suggestion subsystem, which is the central component of the recommendation system, has immense value for farmers. The first stage in achieving optimal performance of a machine learning system is to collect high-quality and precise training data. The schema of the produced dataset may include variables such as aquifer thickness, soil type, soil pH, precipitation, topsoil thickness, temperature, and location, among others. The subsequent phase in the crop forecast process involves the collection of preliminary data. There are two subordinate processes within this process. The original training dataset may have had missing values that required removal and subsequent replacement. The removal and replacement of these entities is necessary due to their potential to undermine the values and significant impacts the performance of machine learning techniques. Subsequently, it is essential to construct class tags for the sets of data prior to the use of the method.

Rainfall Prediction System

This subsystem has the capability to predict the occurrence of precipitation, as suggested by its nomenclature. It is well acknowledged that various crops exhibit distinct requirements for rainfall, and failure to meet these requirements may result in diminished agricultural yields. However, an abundance of precipitation may also lead to unforeseen repercussions. Hence, precipitation emerges as a vital factor within the realm of agricultural analytics. Nevertheless, accurately predicting rainfall throughout the periods of planting and harvesting may provide a challenge. Therefore, this particular element of the system is responsible for generating yearly precipitation projections.

The dataset consists of monthly rainfall data obtained from the official website of the government meteorological agency. The dimensions are measured in millimeters. In order to mitigate further deterioration, comparable preprocessing techniques are used on the earlier subsystem data, whereby missing values are substituted with negative values. In this study, Vogelstein et al. [10] use linear regression, a supervised learning technique, to estimate the numerical value of rainfall in the given geographical region. The parameters under analysis are X, representing location, and Y, representing precipitation levels. In summary, this recommendation system demonstrates a high level of execution and intelligence, offering valuable benefits to agriculturists and farmers as they evaluate their alternatives.

Precision Agriculture Using Map-Reduce

In the field of distributed computing, the term "map-reduce" encompasses both a processing approach and a programming paradigm. The data is structured so that the value and key components are distinct and independent from each other. The first step involves the mapping of data using a relative key value, followed by the reduction process where the data is distributed among several nodes. Any nodes that are linked to the central ones have the potential to function as data repositories. This particular approach is widely recognized as one of the most efficient methods for obtaining a prompt and reliable three-dimensional data viewpoint. This capability significantly enhances the process of decision-making and reporting. In the field of agriculture, this approach is used to get a deeper comprehension of the temporal and geographical fluctuations in soil and plant components via the strategic administration of inputs and variable rates. This technology has the potential to provide valuable insights via its capacity to do comprehensive data analysis, benefiting professionals in the agricultural industry, including agriculturists, farmers, and computer scientists specializing in agriculture. Therefore, this concept is often referred to as "Precision agriculture".

The acquisition of a dataset serves as the first step in any processing methodology. The data is obtained from official sources such as government and meteorological organizations' websites. The datasets that have been obtained are of considerable magnitude, sometimes surpassing petabytes in size for a single research endeavor. The information encompasses several parameters, such as daily market use, FCI capacity and storage, soil analysis, and water level. Given that the data has been gathered and prepared. In order to incorporate big data into distributed computing, it is essential to use the map-reduce approach, which facilitates the aggregation of data from several sources into a unified report. The report often includes data pertaining to market circumstances and weather patterns. Furthermore, an integrated business analytical tool such as PowerBI generates a 3D visual representation that combines several aspects with charts and graphs. This technology assists farmers in making informed choices. The cognitive processing of visual stimuli and visual representations

is much faster and more efficient compared to textual information, thereby explaining the widespread appeal and popularity of visualization techniques.

Crop Prediction Using Various Machine Learningapproaches

A continuous flow of agricultural data, including both homogenous and diversified types, is continuously being generated. Common datasets include a range of sources, including streaming data, historical data, social and web-based data, official institution websites and agricultural sensor data. Consequently, the accessibility of substantial quantities of agricultural data, along with the emergence of machine learning methodologies, has facilitated the extraction of patterns and predictive insights from this data. Therefore, an examination of the Machine Learning methods now used will be presented in the next section.

Grey Wolf Optimization (GWO) Technique

The use of optimization techniques in machine learning involves the process of feature selection, which aims to reduce the complexity of the classification task by identifying a more manageable subset of features. The procedure of preparing the data required for training the algorithm is executed in a manner that is comparable to previous methodologies. The GWO technique has a discerning focus on prioritizing aspects that are deemed desirable. During the optimization phase, the features are allocated a weight relative to each other depending on the training error and the relative information. The afore mentioned method demonstrates a superior ability to identify semantic associations as opposed to geometric coherence. The map-reduce framework is used for the management and analysis of datasets obtained from esteemed governmental entities, hence enhancing the efficiency of tasks like as feature selection and classification. The results may be effectively presented via the use of a visualization tool, hence facilitating decision-making processes. The use of the grey wolf optimization strategy effectively reduces the training error of the model, while also establishing causal relationships among its characteristics. This enables farmers to get more precise forecasts via the use of crop prediction models.

K-Means Clustering

The K-means technique is classified as an unsupervised machine learning technique. The method partitions objects into clusters based on the average values they exhibit across several iterations. The K-Means algorithm is a non-hierarchical clustering approach that partitions data into two or more groups. Data sets exhibiting similar qualities will be grouped together, and data sets demonstrating distinct characteristics will be segregated into separate groups.

The primary objective of grouping is to optimize diversity within groups while simultaneously decreasing the target functions specified inside the grouping procedure. The K-Means clustering technique is anticipated to facilitate the partitioning of harvest areas into distinct clusters, using variables such as harvest area (Ha), production (tons), and harvest year. The K-Means algorithm was used to categorize potential areas for maize cultivation in this study. The purpose of utilizing the K-Means technique is to facilitate the partitioning of areas depending on their maize production levels. The final outcome is a cartographic representation that delineates prospective regions for corn cultivation.

The cultivation of crops plays a crucial role in the agricultural industry. The agriculture industry provides employment to over 70% of the population. The decline in crop yield resulting from infection has led to a reduction in production, and the allocation of inefficient food materials has occurred. The use of pesticides in agricultural practices is a contributing factor to the dissemination of plant diseases. Sturgis [11] conduct research on leaf detection and the identification of disease prevalence. This first measure is crucial in mitigating the continued transmission of the illness. The use of enhanced k-means clustering (EKMC) is proposed as a method for segmenting plant pictures and isolating the green pixels, with the aim of facilitating early crop prediction.

Apriori Algorithm

The apriori approach is used for the purpose of extracting regularly repeating information from databases. First, a K-itemset is formed. Next, the generated itemset is trimmed. Finally, the pruned itemset is categorized under a similar data item. Sets of often seen items are characterized by the phenomenon where the item with the least amount of support is the most prevalent. After the process of data collecting and data cleansing, the factors chosen for analysis include site characteristics, area size, crop production, soil type, and rainfall patterns. The subsequent phase involves the training of the model, followed by the use of a tool used for visual analytics such as Power BI or Tableau to present the results. This presentation facilitates decision-making processes and the prediction of crop yields. Within this particular setting, it is plausible to extract valuable resources from this source and use them to make approximations about agricultural production. The use of the Naive Bayes approach might be considered in order to assess and appreciate a categorization strategy. The Bayes classification approach encompasses the use of historical information with the outcomes derived by functional learning algorithms. By using the available data, explicit probabilities may be computed.

The objective of Supriyono, Ferine, Puspitasari, Rulinawaty, and Timotius in reference [12] was to use the Apriori sales data approach in order to leverage data mining methods in the context of agricultural equipment. The presence of constraints on the use of technology within Indonesia's highly prospective agricultural sector accounts for this phenomenon. The study was conducted at a retail establishment specializing in agricultural supplies situated in the Simalungun locality. The continuous sales activities conducted by the store are generating a consistent flow of novel information. To ensure the use of the resulting data, it is essential to subject them to an algorithmic processing that presents many benefits, particularly in

terms of augmenting the income derived from the agricultural commodities' sales. The primary objective of data mining is to discover consistent patterns, regularities, or associations among extensive datasets. The Apriori Algorithm is a prominent technique used in the field of data mining. There are many potential advantages for future decision-makers in using the algorithm's results. One such advantage is to the reorganization of the product layout, wherein the most popular items are strategically positioned in prominent locations, hence increasing their likelihood of being sold.

Smart Farming

Internet of Things (IoT) and Artificial intelligence (AI) play a significant role in facilitating the integration of cyber-physical farm management within the context of smart agriculture. The use of advanced technologies in agriculture, such as the monitoring of weather patterns, soil conditions, moisture levels, and several other factors, plays a crucial role in addressing a diverse array of crop-related challenges within the context of smart agriculture. The Internet of Things (IoT) technology facilitates the interconnection of various devices over the internet, enabling their autonomous operation. This connectivity extends to a range of remote sensors, such as robots, ground sensors, and drones. The primary objective of precision agriculture is to optimize crop productivity by minimizing the inefficient use of chemical inputs, such as fertilizers and pesticides. The use of IoT-based "smart agriculture" technology may provide several advantages to various real-time agricultural processes and practices, such as irrigation and plant protection, enhancement of product quality, regulation of the fertilization process, and anticipation of disease outbreaks, among others.

Here are the advantages of using "smart agriculture". Enhancing the volume of real-time data pertaining to crops; Facilitating remote monitoring and regulation of farmers; Enhancing the efficient use of water and other natural resources; Optimizing livestock management practices; Conducting accurate assessments of soil and crop conditions; Enhancing agricultural productivity. **Table 6** presents the relevant technologies associated with smart farming.

Table 6. Technologies Integrated in Smart Farming

Internet of Things (IoT)	The use of sensors in agricultural environments for data collection is facilitated by Internet of Things (IoT) technology. Wireless sensors may be utilized to evaluate various agricultural features like soil quality and temperature data, among others.
Cloud Computing	Similarly, it is essential that all data be ultimately consolidated into a single area that is easily accessible. Cloud computing is a viable option that enables the consolidation and distribution of computer resources at a reduced cost. The storing of agricultural data is a prominent application that is well integrated with the Internet of Things.
Mobile Computing	The agriculture industry is susceptible to the pervasive use of mobile computing technologies. This technology has the potential to provide regular updates to farmers and agriculturalists on weather conditions, market trends, and the status of their crops, on a daily, weekly, or seasonal basis. Ultimately, the integration of the aforementioned technologies with analytical methodologies leads to the discovery of novel approaches in design, evaluation, ideation, and development.

The smart farming strategy utilizes the Map-reduce technique for data analytics and consists of a combination of integrated technologies such as the Internet of Things (IoT), Cloud computing, and Mobile devices. Managing data with several nodes is straightforward in this particular context. The process is divided into two distinct phases: the mapping stage and the reducing stage. One may use a map to apply filtering and sorting operations. In contrast, the reduction function has the capability to do summarization. Therefore, this kind of study is often used in the domain of predictive analysis. Given the substantial volume of data generated by Internet of Things (IoT) sensors, there exists the possibility of using this data for the purpose of data mining. The sensor data is sent to a cloud-based storage provider for archival purposes. The first step in the data analysis process involves the cleaning and organization of the data. Subsequently, the data is categorized and characteristics are carefully selected. Following this, algorithms are used to analyze the data, and the obtained findings are thoroughly examined. Only after these steps have been completed can patterns be anticipated. The use of a Business analytics tool might potentially enhance decision-making processes pertaining to factors such as fertilizer application, crop rotation, weather patterns, and other related variables.

Crop Analysis Based on Data MiningThe objective of this project is to conduct an analysis of greenhouse crops via the use of a data mining methodology in order to get insights pertaining to their growth and development. Crop growth, the prediction of soil moisture, and the influence of other greenhouse factors such as temperature, light, and humidity all play significant roles. The inclusion of an objective variable and the development of an analysis system are crucial stages within the data mining analysis process. Data is gathered inside greenhouses via the use of wireless networks and embedded systems, including various devices and sensors associated with the Internet of Things. Variables such as soil moisture, carbon dioxide content, humidity, and luminosity are among the components used in the construction of the databases.

The process of knowledge discovery in databases (KDD) [13] involves the collection and preparation of data, followed by the use of pattern mining techniques such as decision trees and classification rules to analyze the dataset. Based on the collected data, this approach produces a prognostic model. In contrast, data mining is implemented with the purpose of exploring previously unidentified patterns within extensive datasets. Once the final interpretation has been produced, the output is evaluated and presented utilizing the graphical user interface's inherent functionalities. The graphical user interface

in question encompasses a variety of icon sets and algorithm libraries. Hence, the user interface and the choice of particular greenhouse factors would enable farmers to accurately forecast production patterns and crop patterns, empowering them to make informed decisions.

Spark-oriented Agricultural Data System

Geo-spatial information is used across many disciplines and applications, yielding significant advantages. Similarly, the use of geographic data might prove to be advantageous inside the agriculture industry. Various technological advancements have emerged in the field, including remote sensing, global positioning systems, and devices designed to gather geo-spatial data of high resolution, which may afterwards be used in analytical modeling. This article discusses a system built on spark technology that is capable of collecting, acquiring knowledge from, training, validating, and presenting globally distributed datasets. Big data from the agricultural sector is gathered from a variety of sources, including institutions, open internet data repositories, agricultural schools and government official websites. The data that has been gathered is in its original, unprocessed form and contains many sources of interference, necessitating the need for annotation and refinement prior to its suitability for training a model.

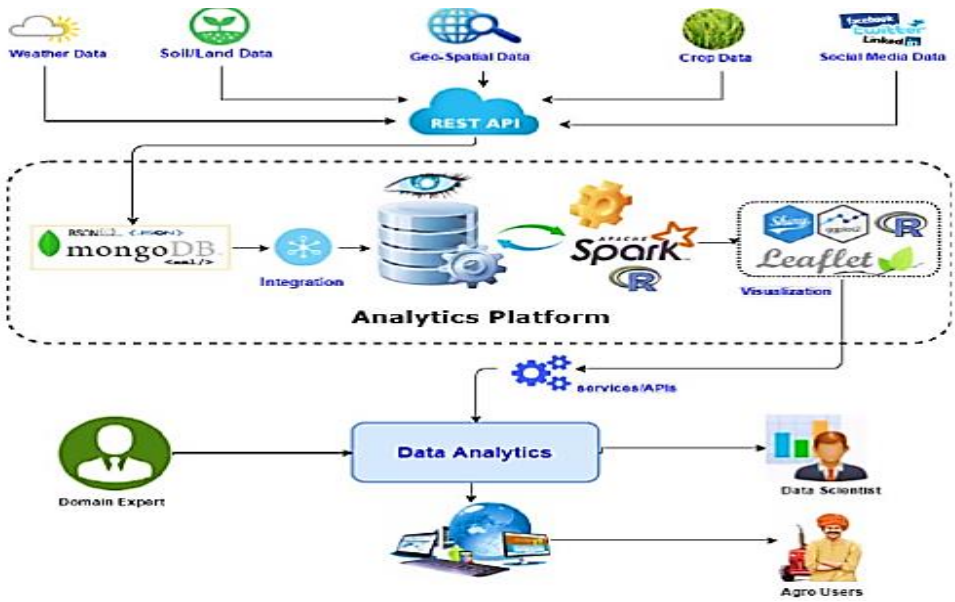

Fig 2. An Agricultural Information System Based on Spark Technology.

Cassandra is a database known for its exceptional availability and performance, serving as the repository for the presently stored preprocessed data. Once the data has been effectively stored, it is combined with the Apachespark model, which is an accessible cluster computing environment that provides support for a diverse range of programming languages and libraries (such as Java, Scala, Python, among others) via its application programming interfaces (APIs). The Geospark spark plugin facilitates the seamless integration of Cassandra data with the remaining spark data. The aforementioned software is a freely available and open-source extension designed for the purpose of doing geographical analysis. The present study used the Multiple Linear Regression approach, which necessitates extensive data modification. Once the essential model is generated, the outcomes are presented using web-based, interactive JSON maps. This kind of data analysis may provide valuable information on the current weather patterns, projections for agricultural output, and other significant insights. The agricultural market data is available for analysis.

There exists a diverse array of databases that may be used in the creation of big data solutions, including stream processing engines, analytical databases, and non-relational data management systems such as column databases. The objective at hand is the development and implementation of advanced big data systems and tools to effectively manage substantial volumes of data, including geographical data. In contrast to commercial and proprietary alternatives, open-source big data tools exhibit a notable deficiency in terms of functionalities. Data architects and engineers have a considerable dilemma when it comes to determining the most suitable big data open-source frameworks for efficiently managing large volumes of data. In order to address the disparity in information technology and provide support for the agricultural sector, **Fig. 2** illustrates a spark-based agricultural information system.

V. CHALLENGES OF BIG DATA IN AGRICULTURE

The challenges pertaining to big data analytics within the agricultural sector may be categorized into two distinct groups: technological challenges and organizational challenges. Technical hurdles include a range of challenges, including the installation of equipment, implementation of information technology (IT) frameworks, demonstration of technical capability, troubleshooting, ensuring security measures, and maintaining a smooth and uninterrupted operation of setup. On the other

hand, companies have issues that are mostly associated with business operations, including matters such as financial performance, investment decisions, security concerns, the maintenance of a proficient workforce, the acquisition of new talent, and overall organizational management. The primary areas of emphasis are revenue generation and investments, with a particular emphasis on maintaining high standards of quality, yield, safety, and security. It is evident that the integration of advanced technology requires a significant financial investment and meticulous strategic planning, both of which are unlikely to be feasible for farmers in prosperous nations.

Nevertheless, developing nations have challenges in meeting the financial demands associated with infrastructure development and its associated expenditures. From a sociopolitical perspective, an additional challenge arises from the dependence of farmers on the monopoly held by the agrifood business, which has expanded in recent years due to the sector's use of more sophisticated methodologies. The concentration of big data technology inside these groups is attributable to its use in manners that exclusively favor the individual, so imposing constraints on its progress and autonomy. Nevertheless, several apprehensions have been expressed about issues such as the methodologies used for data collection, the safeguarding of data, and the tactics employed for capitalizing on data. Farmers exhibit caution in disclosing sensitive information due to concerns over potential access by their competitors. Therefore, the aforementioned subjects have shed light on many unresolved inquiries about the use big data analytics within the agricultural sector.

VI. CONCLUSION AND FUTURE RESEARCH

This research examines the concept of integrating big data analytics within the agricultural sector, exploring the underlying motivations driving this integration and the concrete and beneficial advancements that have been seen after its implementation. In addition, we explore various sources, methodologies, and strategies used in the acquisition of diverse datasets essential for the exploration of big data. A framework rooted in the domain of agronomy has also been delineated for doing agricultural analytics. Furthermore, there has been a proliferation of novel technology infrastructures and resources. This paper provides a comprehensive analysis of the integration of big data analytics with various techniques, including description, explanation, and in-depth examination, focusing on their use in the agriculture sector. In conclusion, a thorough examination of various challenges and potential strategies for surmounting them has been undertaken.

Subsequently, a tabular representation will be shown, delineating diverse agricultural locations, their respective sources of data, and the corresponding methodologies that might potentially be used within each region. The broad adoption of open standards and the increasing prevalence of big data and its analytics have significant potential for the advancement of smart agriculture. However, it is equally important to consider the potential negative consequences of compelling farmers and agriculturists to relinquish their existing markets. Conversely, it should serve as a source of motivation and enhancement for the existing protocols aimed at enhancing product quality, generating more financial gains for the organization, and formulating sustainable strategies that minimize the depletion of natural resources.

The potential of big data in the agricultural sector is quite attractive. The biological manufacturing system may be characterized as a complex system that relies on the interaction of several factors such as human involvement, machinery, natural systems, chemical processes, biological elements, meteorological conditions, and environmental factors in order to achieve optimal performance. The rapid rate at which the shift from the "green revolution" to the "evergreen revolution" in contemporary agriculture is occurring is noteworthy. Currently, the agricultural sector has the challenge of providing sustenance for an expanding global population, while also adhering to ecologically sustainable practices amidst the prevalence of chemicalization and industrialization.

Additionally, it must adapt to the fast advancements in technology and navigate the uncertainties posed by unexpected climatic fluctuations. Gaining a deeper understanding of the intricate interrelationships among biological processes occurring within the networks of genes, proteins, and metabolites that govern the phenotypic characteristics of organisms would be advantageous in elucidating the underlying biological mechanisms responsible for the physiological traits exhibited by both the organism itself and its offspring. Efforts in environmental processes encompass various areas, including energy generation, next-generation commercial crops and biofuels, global carbon management, and remediation of contaminants. These endeavors play a crucial role in enhancing agricultural productivity, mitigating the effects of climate change on crops, facilitating adaptations in organisms to combat biotic and abiotic stresses, and elucidating the interactions between pathogens and host plants and animals.

Acknowledgement
I would like to thank Mattu University for their extraordinary support in this research.

CRediT Author Statement
The author reviewed the results and approved the final version of the manuscript.

Data Availability
The datasets generated during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interests
The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding
No funding was received for conducting this research.

Competing Interests
The authors declare no competing interests.

References
[1]. C. Tardi, "Value chain: Definition, model, analysis, and example," Investopedia, 20-Nov-2003. [Online]. Available: https://www.investopedia.com/terms/v/valuechain.asp. [Accessed: 30-Jul-2023].
[2]. The Kien NGUYEN et al., "Influence of macro environment on tourism value chain in Vietnam: Case of Daklak province," GeoJ. Tour. Geosites, vol. 41, no. 2, pp. 400–407, 2022.
[3]. V. O. Filatov et al., "Information space model in tasks of distributed mobile objects managing," Otbor i peredača informacii, vol. 2019, no. 47, pp. 80–86, 2019.
[4]. R. C. Hughes, "Online analytical processing and the OLAP cube multidimensional database," in Human Capital Systems, Analytics, and Data Mining, Boca Raton : CRC Press, 2018. | Series: Chapman & Hall/CRC data mining & knowledge discovery series ; 46: Chapman and Hall/CRC, 2018, pp. 103–126.
[5]. J. Al Qundus, S. Gupta, H. Abusaimeh, S. Peikert, and A. Paschke, "Prescriptive analytics-based SIRM model for predicting Covid-19 outbreak," Glob. J. Flex. Syst. Manag., vol. 24, no. 2, pp. 235–246, 2023.
[6]. F. H. M. Tang, M. Lenzen, A. McBratney, and F. Maggi, "Risk of pesticide pollution at the global scale," Nat. Geosci., vol. 14, no. 4, pp. 206–210, 2021.
[7]. P. Cleary et al., "A foodborne outbreak of Salmonella Bareilly in the United Kingdom, 2010," Euro Surveill., vol. 15, no. 48, 2010.
[8]. A. Martins, Ø. H. Opedal, W. S. Armbruster, and C. Pélabon, "Rainfall seasonality predicts the germination behavior of a tropical dry-forest vine," Ecol. Evol., vol. 9, no. 9, pp. 5196–5205, 2019.
[9]. K. Charvat et al., "Advanced visualisation of big data for agriculture as part of databio development," in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018.
[10]. J. T. Vogelstein et al., "Supervised dimensionality reduction for big data," Nat. Commun., vol. 12, no. 1, p. 2872, 2021.
[11]. W. C. Sturgis, "On some aspects of vegetable pathology and the conditions which influence the dissemination of plant diseases," Bot. Gaz., vol. 25, no. 3, pp. 187–194, 1898.
[12]. Supriyono, K. F. Ferine, D. Puspitasari, Rulinawaty, and E. Timotius, "Implementation of data mining with Apriori techniques to determine the pattern of purchasing of agricultural equipment (Case Study: XYZ Store)," J. Phys. Conf. Ser., vol. 1933, no. 1, p. 012029, 2021.
[13]. J.-M. Guldmann, "Analytical strategies for estimating suppressed and missing data in large regional and local employment, population, and transportation databases: Estimating suppressed and missing data," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 3, no. 4, pp. 280–289, 2013.

Publisher's note: The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The content is solely the responsibility of the authors and does not necessarily reflect the views of the publisher.

ISSN: 3104-4654