A Survey on Big Data in Agricultural Farming

E. Banu[1]
[1]Ph.D Research Scholar, PG and Research Department of Computer Science,
Chikkanna Government Arts College, Tirupur - 641602. Tamil Nadu, India.

Dr. A. Geetha[2]
[2]Assistant Professor and Head, PG and Research Department of Computer Science,
Chikkanna Government Arts College, Tirupur - 641602. Tamil Nadu, India.

Abstract

The agricultural farming are rapidly increasing the challenges due to various environment, climate change and complex agricultural ecosystems. Nowadays, the cutting edge technologies are used to monitor the environment continuously and send the data to the cloud server in large volume to make further study. The international Data Corporation (IDC) forecast the total amount of data worldwide around 175 Zettabytes by 2025. The internet, sensor devices and data analytics paves the way for better performance, advanced resources planning and enhanced production in all industries. The agriculture is the most important and essential need for the world to invest and make research to strengthen the yield, crop disease control, farming community and advance agricultural production. The recent reports from world food and agricultural organization indicates t hat there has been an increase of global investment in agri-food technology. The main objective of this paper is to carry out a survey on research and recent trends in agriculture using big data analysis to solve various relevant problems and increase the production. The wide openings of big data in agriculture towards the smart agricultural farming, various techniques and methods for big data analysis, other latest mutual technologies shall encourage more academic research, government bodies and giant agribusiness industries.

Keywords: Big Data, Big Data technologies, Smart Farming, Big data in Agriculture, Recent trends in Agriculture.

I. Introduction

The rapid growth in the world population and the climatic changes leads to heavy food shortage in the coming future. In December 2018, the United Nations General Assembly adopted a Resolution declaring 2020 as the International Year of Plant Health[1]. This year is treated as good chance to create a universal awareness to know the health of the crops and plants. It can help to address the issues of poverty, environmental impacts and agricultural economic development for most of the countries. The UN predicts there will be population growth in Africa and Asia during 2020-2050 creating a huge shortage for food[2]. Therefore the world should look into the recent technologies to boost the agriculture activities and increase the production of food to be achieved. To persuade these increasing needs, various initiatives and research have been started by various world organization from the last decades.

Smart Farming is the classical word in the recent modern agriculture to utilize the cutting edge technologies like Remote sensing, IoT, Cloud Computing, etc., The implementation of new digital technologies for the plant level to farm management broadens the agriculture accuracy by improving the existing tasks and decision making. Smart farming is significant for handling the challenges of real time agricultural activities in terms of crop growth monitoring, yield prediction, weather impact, disease control and sustainability[3]. It is implicit that the agriculture integrates biological, chemical, physical, ecological and economical factors to build up a secure and smart farming. It should not affect the society and environment by increasing the yield/ production and maximize to give nutritious, healthy food to the society.

The recent innovative technologies gives more confident to understand simply by supervising and evaluating continuous aspects of the physical environment from lot of sensors which generates huge volume of data. It raise the need for extensive data collection, storage, pre-processing, data modeling and finally the analysis of collected data to provide meaningful information to the farmers or the management. This "Big Data" comes into real time practice which helps farmers and related organizations can pull out needed information from a huge volumes of a wide range of data by facilitating high velocity data capture, search and analysis.

Big data analysis is effectively used in different industries like social media, banking, share market, insurance, customer behavior analysis, online ecommerce platforms as well as in ecological and environmental research[5]. This survey clearly explains the role of big data technologies in agriculture with recent advancements and applications Thus the major part of this survey is that it presents a more focused overview of the various factors which leads to failure in agriculture yield and production by keeping the impact of environment. The survey highlights the Big data techniques used to find the solutions gives technical perspective with the potential of recent technologies.

II. Big Data

The word "Big Data" was first introduced by Roger Magoulas in 2005 to define a large amount of data that traditional application cannot handle and process due to its complexity and size. It is the collection of various structured, semi-structured and unstructured data sets with huge volume in Terabytes and complex in nature. A recent survey says that 80% of the data sets are unstructured[6]. The biggest real challenge in this domain are storage and unstructured data can be structured before analysis. The big data 5Vs are diagrammatically shown in the following figure1,

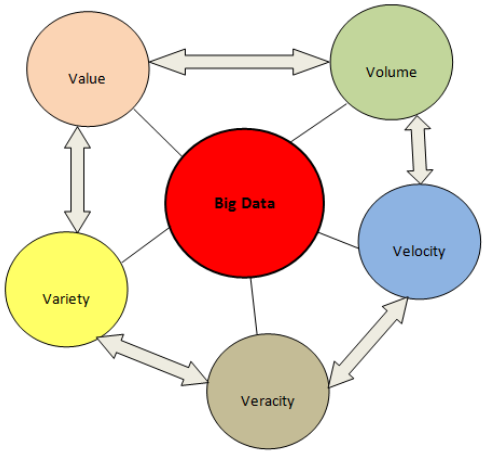

Figure 1: Characteristics of Big data

The Figure 1 shows the Volume, Velocity, Variety, Veracity and Value of the big data[6]. It collect, store, manage and process huge amount of data with speed and time to get proper information to make right decisions.

Volume: It defines the data capturing that are being generated from different data sources in large volumes and are in the very huge form usually measured in Exabyte and Zettabytes. The big data starts with minimum data size as Terabytes[5][6]. This is the quantity of data collected that enclosed the important information for its own purpose of collection to process and analyze with their own criteria.

Velocity: The huge volume of different types of data generated from various sources has to be process quick and speed. The velocity deals with the speed that data collected from different sources to be gathered and processed quickly which is not possible with existing traditional applications[6]. It is the time in which data can be processed and analyzed fast in order to avoid fraudulent activities and sensitive information leakage.

Variety: It refers to the various types of data captured. This data may be structured or unstructured or semi structured. Big data consists of any combination of data, including structured and unstructured data such as sensor outputs, simple text, video, audio, images, likes, feedback, comments and files in various formats[6]. The structured data can be processed quick and easily when compared to the other formats which make more complex and difficult to process and analyze.

Veracity: The data collected and stored from different data sources in various formats and types frequently are not perfect and accurate. The poor quality of data in large volumes are not accurate and doubtful. Even though the data is not precise, the big data provides ways to work with these types of data[6]. Because of these, establishing the trust is a very big challenge in big data when the number of data points and types increases.

Value: It refers to the significant value of the collected data which is small or large in volume irrespective of the type. As per the validness of the data, it can be utilized in a right way to the planned action of process and data analysis[6]. Depending on the data value and importance, it can be collected completely and stored for a lengthy duration for future use. The data volume and variety are interconnected with the value of the collected data.

III. Big Data Technologies

The Big data technology is nothing but the software function developed and designed to meet the requirements to store, process, analyze and extract the information from an very huge various data sets with more complexity[7]. This cannot be done with the existing software application in traditional way of data processing. The Big data gather and process huge amount of real time data from various sources with its analyzing tools to show the possible predictions in perfect time to reduce the risks and fraudulent acts quick and fast[7]. The Big data technologies are widely classified into two types as per their properties are,

- Operational
- Analytical

The operational one is the simple data sources for the collection of huge data volume in our day to day activities. Social media, ecommerce and the various online booking data are the example for the operational big data technologies. It is mainly deal with the raw data containing the details of the individual needs and requirements. The Analytical big data technologies take the data from the operational with huge complexity and analyse the data. It will conclude and predict about the user specifically and show what the user needs exactly based on their searches and critical business decisions for the organisations[8]. Weather forecast for smart agriculture, share market predictions and ecommerce particular product sales forecast are the finest examples of the analytical big data technologies.

The big data technologies consist of mainly four major functionality domains in order to achieve the final end results. The software utilities and various tools are also developed and designed to meet the requirements of the each domain to have its own functions and objectives. The figure 2, shows the four division of the big data technologies as follows,

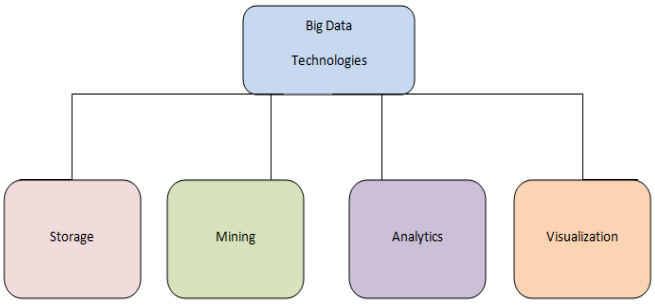

Figure 2. Core Divisions of Big Data Technologies

The above said core divisions of the Big data technologies are widely used in many organizations and government enterprises partially or completely with the software utilities as per their own requirements and need. The top technologies are discussed below with their ability and specifics in order to fit with the objectives of the companies[8][9].

A) Hadoop

Hadoop is mainly designed to store and process the large set of data in a Distributed Data Processing Environment with commodity hardware with a simple programming model. It can store and analyse the data present in different machines with high speeds and low costs. It is designed using JAVA language and developed by Apache Software in December, 2011. It is used in many popular companies like IBM, Intel, Microsoft, Cloudera, Hortonworks and MapReduce. The Hadoop ecosystem consist of both the apache projects and various types of commercial tools and solutions[10][14]. The Spark, Hive, Pig, Sqoop and Oozie are the few examples of apache open source. It is mainly used for the data storage when compared with the other software application tools.

B) NoSQL

The traditional relational database management system gathers and collect information in structured format like database tables (Rows and Columns). The application developers and database administrators use query language called SQL to manipulate and manage the data stored in the database. The NoSQL[14] mainly deals with storing unstructured data and providing fast performance even though they are not reliable like other traditional databases[11]. It stores data in non tabular for and different than the relational database The MongoDB, Cassandra, Couchbase are the popular known NoSQL database in the market as the big data technologies grows many will arise.

The NoSQL MongoDB provides an option to the inflexible representation of data used in the traditional databases. It is a document database which makes it greater flexible over handling a wide variety of Datatypes with huge volumes like JSON documents in a distributed environment. The different data types widely used are document, key value, wide column and graph. It provides flexible schemas, scale easily with huge amounts of data and large user loads. It is written in C++, Go, JavaScript, Python and released in the Feb, 2009. The popular companies using the technology are MySQL, Microsoft SQL Server, Google, Cisco and SAP.

C) RAINSTOR

RainStor, a software company developed Rainstor database management system in 2004 for the purpose of storing, processing and analyzing huge amount of data for giant organization and companies. It uses deduplication method to manage the process of storing huge data in a highly compressed way by its own file formats. The data can be accessed and retrieved through the interactive SQL instead of the MapR layer [12]. The MapReduce can access the Rainstor database by using the HCatalog. It has higher level of controlling mechanism and data security such as data masking, encryption keys, kerberos and light weight directory access protocol authentication, access rights policies, version control, data retention and replication rules. Its mainly known and popular for its storage facilities and techniques for structured and semi structured types of big data for big organization. Credit Suisse and Barclays are using RainStor in banking

industry and quite used in Airport maintenance and services industry, marketing, advertising and other industries too.

D) HUNK

Hunk is a integrated analytics platform with complete features to search, process, analyze and visualize data from a huge data sets from Hadoop and NoSQL databases. It is very fast and quick tool to visualize the data so the dashboards and report sharing making it popular among others[10][13]. It can search and analyze data from a hadoop cluster and deliver results quickly right after linking without any data configurations. The interactive search of unstructured data, analysis, identifying samples and inconsistency from a huge volume of data without any training dataset makes it more powerful tool. It has good data visualizations options, various task and role based dashboards, interactive charts and graphs. The complete option loaded environment makes developers to build their own way of big data applications to meet the requirements. It is fully written in Java and developed by Splunk in 2013.

E) SPARK

Apache Spark[14] is the popular framework and unified analytics engine for huge SQL queries, machine learning, stream and batch processing. It offers in-memory computing facility in order to execute fast and deliver results in a speedy manner with wide range of flexibility. The environment is rich, friendly, flexible for the developers and the spark API integration is very easy and can start distributing the data processing among other machines and big data processing[10]. The first version released in 2014 and designed using various languages such as Java, Scala, Python and R. It constructs the users data processing commands into a Directed Acyclic Graph (DAG) which schedule the task among various nodes with the corresponding sequence. It uses the Resilient Distributed Dataset (RDD) concept to manage distributed processing of data and the transformation of that data. The RDD process can be divided among the clusters and execute in a parallel batch processing which results in fast and scalable parallel processing. The Spark consist of Core, SQL, Streaming, GraphX and MLlib (Machine Learning library) as components to process, analyze and visualize the data in various formats. Amazon, Oracle and Cisco are the giant companies using the spark for their operations.

F) RAPIDMINER

It is a centralized and powerful solution having graphical user interface which facilitates the developers and users starting from the data preparation, machine learning algorithms to predictive analytics[15]. It is completely written in Java language and released in 2001. The Rapidminer studio provides a visual workflow designer environment which speed up the machine learning end to end process for the enhanced output. It is used to analysis the organizational risk management and avoid many fraudulent activities well before it happen. Intel, GE, BMW, HP and Samsung are the known organization using it for their business. The various Rapidminer tools are widely used in all industries starting from customer retention to the products selling and price optimization.

G) PRESTO

It is an open source data mining tool with distributed SQL query engine to run interactive data analytics from a huge volume of data[16]. It is not restricted only to Hadoop and HDFS allow the users to work with various types of data sources such as Traditional Relational Databases Hive, Cassandra and Proprietary Data Stores. The two servers namely coordinator (planning queries and managing worker nodes) and the worker(executing task and processing data) communicate each other through RESTAPI and finally send the result to the coordinator and then to the real client machine. It is designed in JAVA language in the year 2013 to manage huge volume of data sets and analytics effectively. Facebook, Airbnb and Netflix are using Presto for analyzing the interactive user queries against multiple data sources within their organization to deliver fast response results.

H) KAFKA

It is a open source and distributed streaming platform having key capabilities such as publisher, subscriber and consumer. It publish and subscribe streams of huge records quite similar to a Message Queue or an Enterprise Messaging System [10] [17]. It works as a cluster to store and process the real time streaming of records with the application reliably with different data sources. The messaging, storage and streaming of huge data records using the five APIs such as, Producer, Consumer, Streams, Connector and Admin. The communication between the server and client are using TCP protocol and the client application are available in various language but the JAVA client is the default one. The scaling up of the data in storage and streaming is very easy without affecting its performance and suitable for real time processing. It is written in Scala and Java in the year 2011 by Apache software foundation. Netflix, Yahoo, Twitter and Linkedin are the big organization using it for the messaging, storage and processing of stream of real time data.

I) SPLUNK

Splunk is a analytical software tool in big data to search and analyze machine generated data like data from different sensors, manual user input data, devices and web data[14]. It explore and analyze the logs created from different processes and structured or semi-structured data with its built in features. It collects data from real time environment, setting corresponding indexes and compare it with various data sources and visualize through dashboards, notifications, graphs and reports form. It is released in 2014 by Splunk Inc, America and written in AJAX, C++, Python and XML. Easy to deploy, high scalability with data security and real time alerts are the key things to make it strong and popular. Trustwave, Qlabs and Splunk Inc are the organization are using this software for their operations.

J) KNIME

KNIME (Konstanz Information Miner) is open source platform for data analytics and reporting. It provides graphical user interface to build and integrates with different database native connectors or through the JDBC connectivity[18]. It is used to recognize the data in a better way design and execute the data science workflows without having programming language knowledge. The data can be collected from different sources, databases and various formats to make it ready by cleaning and generate log details. After the process, the machine learning models can be build, optimize, validate and predictions obtained. The data visualization can be shown in dashboards, reports and stored for future use. It is written in JAVA and based on eclipse, released in the year 2008. Continental, Siemens, DAIMLER and Seagate are the corporate companies using this for their business operations.

K) DATA LAKES

It is a centralized storage repository that hold a huge amount of raw data in binary, structured, semi-structured and unstructured format. It can allow to store the data in its native form from the real time without its structure and run various types of data analytics[14]. It saves huge time to arrive the data structures, schema and data transformation while scaling up the data. It offers to store relational data from various databases, business application and non relational data from sensors, smart devices and social media applications. Machine learning models can be built and analytical tools can be integrated to visualize the data without moving data to those tools. It is named by James Dixon, CTO, Pentaho in 2010 to differentiate with datamart and used nowadays in various cloud platform. The Netflix, NASDAQ, iRobot, and FINRA are using data lakes in cloud for their critical business operations.

L) MAP REDUCE

MapReduce is a programming software framework to carry out the parallel and distributed processing on very large record sets in a cluster[19]. The name itself explain the task as Map and Reduce. The task Map will split and map the dataset from various datasets, the Reduce task will shuffle and reduce the large dataset into specific to give the output[10]. Hadoop can execute its program in Java, Ruby, Python, and C++. The programs written are parallel in nature so it can execute large scale data analysis using several nodes in the cluster. Parallel processing and the processing unit to the data nodes are the key points make it more efficient for the operations. It save huge time and the cost of the traditional approach. It was introduced by Google through a paper in 2004 to work with large clusters of commodity hardware and replace in 2014 with new technologies.

M) BLOCKCHAIN

Blockchain is used in crucial dealings such as financial transactions, smart legal agreements, crypto currencies and digital asset contracts to reduce the fraudulent actions, globalization, secured and safe transactions. Blocks, Miners and the Nodes are the key concepts which used to record the transaction and validate. The decentralization, that is the distributed ledger connected like chain over the nodes make it more scalable in trust[14]. Nowadays the developed countries and banks are implementing this technology to secure their transaction, privacy and crypto currencies. It can be integrated in areas like videogames, energy trading, supply chain and domain names for the security and safe transactions. It is developed by Satoshi Nakamoto in 2008 and written in Python, C++ and JavaScript. Bitcoin network, IBM, Facebook, Oracle and Alibaba.com are the players using this technologies for their business operations.

N) R PROGRAMMING

It is a statistics programming software language for statistical computing, reporting and graphical representation of large datasets[20]. The R Studio is the best integrated development environment for the R language and the language can be extended using any number of packages both the system and user created. There are more than 15000 packages with data storage and manipulation potential makes the R as powerful language to a wide variety of statistical computing such as linear modelling, nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering and various graphical techniques. It is written in language C, Fortran and released by R Foundation in the year 2000. Bank of America, Barclays and American express are using the technology for their complex operations.

O) TABLEAU

It is an interactive, secure, flexible and powerful data visualization and fast analysis software tool in business intelligence with rich dashboard and worksheets. Tableau[21] becomes fast growing popular tool in the industry for the data discovery and its governance. It can be integrated with the cloud as well as the available infrastructure in the organizations. The desktop version extracts a huge volume of data from in-memory or offline with graphics to analyze in the personal system. It can share the data securely with online and server version to collaborate the organizational data to get various possible insights. The device designer tool integrates with any version to publish the dashboards and reports on any device like personal computer, tablet, smart mobile phones with proper alignments. It is written in C, C++, Java and Python. The TableAU, Amercian based company has developed it and released in the year 2013. Oracle - Hyperion, Ford Cars, Qlik and COGNOS are the multinational companies using this tool for their data analytics and visualization.

P) PLOTLY

It is the fastest and efficient tool used for the creating of interactive graphs with data analytics. It creates various analytic reports and data visualizations to take wise decisions possibly in all industry. Plotly Dash become the popular open source having python library to create interactive analytical web based application and bring the Artificial Intelligence, Machine Learning to the business analysts for visualizing the data[22]. The python interface in dash have the plenty of interactive web components which facilitate easy to build the complex applications having lot of interactive elements without using Javascript and HTML. The libraries like plotly python, plotly open source graphing, plotly javascript and support other API libraries for languages R, MATLAB, Node.js, Julia and Arduino makes plotly as. The Plotly computing and analytical company based in Canada, developed and released in the year 2012. It is fully written in Javascript. BASF, Tesla, Cisco and Air Canada are using plotly applications for their data analytics and visualization.

Q) TENSORFLOW

It is a open source software library for developing and implementing complex Machine Learning models for various application. It have complete set of libraries, tools and community resources in machine learning especially with neural networks. Tensorflow[23] core is the complete machine learning with python coding and Keras based API. Tensorflow.js library provides to build models in Javascript and published in browser and Node.js. Tensorflow Lite provides a framework to build and implement deep learning models in smart phones and IoT devices. Tensorflow Swift provides a programming model which unites the graph performance, efficiency, usability and flexibility to build and deploy machine learning models. TFX is the Tensorflow extended platform provides to implementing the machine learning model to the production pipelines. Simply, implementing the models from the research level to the actual production. It is mainly used in research purpose and nowadays to solve various complex issues. The Google brain team used internally from 2015 for their own operations and released full version in 2019. It is fully developed using C++, CUDA and Python. The giant companies like Google, Airbnb, Coco-Cola, ebay and Intel are using for their core operations and develop various machine learning models.

R) BEAM

It is a free, open source, portable and unified model for the streaming and batch data parallel processing pipelines. The Apache Beam[24] provides same development model for on-premise and cloud infrastructure. It supports the distributed data processing tools like Google cloud dataflow, Apache spark, Samza and Flink to execute the pipelines. The huge task will be divided into smaller pieces and process in parallel without linking other task. It can be used for data abstraction, data integration, data extraction, data transformation and data loading operations. The Java, Python and Go software development kit are used to create programs for data processing. It is developed by Apache Software foundation in the year 2016 and fully written in Java and Python. Amazon, Oracle and Verizon wireless are using this tool for their data extraction, integration and the transformation of data in various format.

S) Airflow

It is an open source workflow engine for the workflow automation, running various task with its schedules, setting task priorities, tracking task status, identifying dependencies, allocation of resources and recovering the task from failure. It uses Directed Acyclic Graphs (DAGs) to create workflows for the given task[25]. The code level of defining workflow for various task makes it easy for the data processing, collaboration, data maintenance, testing and version upgrading. The web server, scheduler, executor and metadata database are the building blocks of its basic architecture for defining the work flow. It can be integrated with various data sources like file system, databases etc with fully scalable to infinity. It is developed by Apache software foundation in the year 2019 and fully written in python

language. The companies Airbnb, Checkr, Walmart and Robinhood are using this tool for the solution of their organization complex work flows and critical business requirements.

IV. Role of Big Data in Agriculture

Nowadays various Bigdata technologies help to enhance and secure the agriculture process which protects plant growth and free from plant killing diseases. The Big data supports the plant growers and farmers from small scale to large scale farming by giving appropriate predictions to take decisions as per the real time data analysis[26]. After the social media booming in the world, data has grown in a unpredicted manner almost in all industry. Every second the data is being created with different sources and collected in servers, cloud and stored for future use. The Big data technologies having the facility to collect huge volume of datasets, storage, access, analyze with different algorithms and handle terabytes of data easily to address the enhancement of agricultural process and business. It is a very critical issue to manage the agricultural operation with farmers to make it successful in their yield and production to meet the market demand[27].

It helps to improve the crop yields, soil information, fertilizers, pesticides, rain water storage and production forecasting to optimize and plan for their labourers, agri equipments with respect to the size of the farming area. The agricultural operation data, agri business data, social media, weather data, sensors, smart devices, machineries and tractors data are huge volume while collecting and very difficult to manage. Big data can collect, store all the data from different sources and visualize the forecasting to make appropriate decision making[28]. The soil, climatic conditions, seeds, plants, soil supplements, water irrigation and the yield are the major things which influence the profitability of the farming business. These technologies facilitate the farmers in a safe way to analyze a wide range of data sources for better decisions to avoid huge loss. The figure 3 shows the role and the other influencing criteria of Bigdata in agriculture,

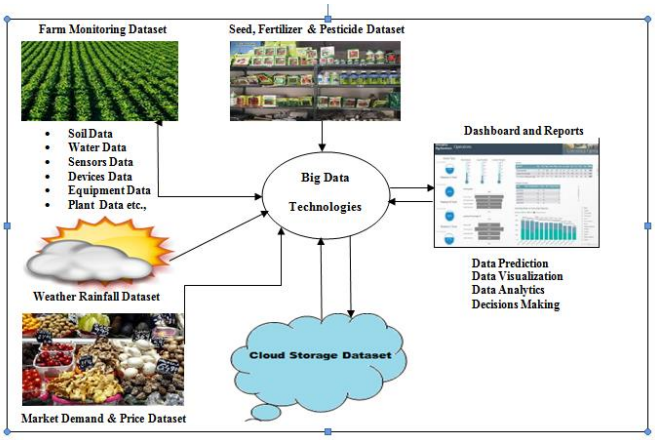

Figure 3: Bigdata in Agriculture

In the above figure, the market demand and the price data set helps the farmers to know which to cultivate during the appropriate season. The availability of seeds, fertilizers and the pesticides are also needed for planning of the procurement to protect crop health[29]. The dashboards, reports and the other analytical information are the outcome that farmer can use, make decisions and utilize for farm operation[30]. Big data technologies provides interactive graphs, dashboards and charts in easy understandable manner to make proper insights of their farming requirements.

V. Challenges of Big Data in Agriculture

There are lot of challenges in natural and practical to handle the agricultural processes and farm operations. The changing climatic conditions by the global warming is a big challenge for the world to handle natural disasters, food security, agricultural production and flood. The technical challenges are the collection of huge dataset from different sources, data storage, choosing technologies to analyze and the data security[7][9]. The result and output of the every query requested by the farmers to be handled individually and inform them with timely prediction is the major challenge of Bigdata[9]. Most of the farmers are older age, trust the traditional way of farming, not fond of using latest technologies and not aware to use the smart devices to improve the yield and profit.

The policies of the government regarding the environment and farming operations are also difficult to input in the technologies. The rural areas are not having speed internet facilities in order to collect the data from different farms to the centralized servers. Data from the sensors, equipments, tractors, farmers manual input data and other sources are carefully handled by the application developers to integrate with the technologies with proper security[31].

Adapting the newer technologies and the funding are the major challenges of the modern precision agriculture implementation.

Technically, the data management, security and the efficient scalable data storage are the critical challenges[10] to implement the Big data in agricultural farms and business. Most of the huge volumes of data in agriculture are unstructured and semi structured, converting them to understandable manner in real time for analyzing with privacy is a complex task. The hardware infrastructure including internet, awareness of using smart devices for their farming operations and choosing the appropriate big data technologies and software tools are the real challenges to be addressed as per their requirements. The technical issues can be solved or upgraded with newer technologies update in order to maintain the proper guidelines to the famers for the better operations and profits.

VI. Discussions

There are numerous research work going on to protect the agricultural activities and business by implementing cutting edge technologies. The recent advancement in agricultural equipments, internet connections and smart devices made the new ways of doing the farming operations and business with proper guidelines from data experts[32]. The main drawbacks are choosing the right technologies and updating to the latest trends wisely. The agricultural community issues will also make things complicated to implement new technologies instead they trust traditional way of doing agriculture globally. There are a lot of initiatives from the government to make the awareness among the farmers regarding the technology and modern farm operations. This research area need more attention because of the future food demand as well as the social economic issues for the majority of the population in the world. The environmental impact is also a big challenge for implementing the new technologies in agricultural farm operations[33]. In this survey, the Big data and its technologies are discussed clearly to build various applications and use to maximize the production, profit and wise decision making in the agriculture[28].

The internet advancement, smart agricultural equipments and smart devices in the farming operations makes significant changes in the agricultural activities and opens a broad area of research to address the data storage, data analytics and prediction to make the real time decisions. There are a lot of challenges in the data storage, selection of appropriate technologies, data governance and data security in the real world scenario of agriculture[34]. The research gap became very wide, more needed to prevent and protect the farming industry to the next generation is very important than any other industry in the world. Ensuring the modern farming operations with Big data technologies to be successful is a very significant research domain.

VII. Conclusions

Big data can be the biggest revolution in the agricultural sector with real time insights to support the farmers for their operation and make the agriculture business and farming more profitable. The world population will cross 10 billion by 2050, the food demand to feed such huge population is difficult without using new techniques in agriculture in order to increase the production of yield and food. The main goal of the smart farming using Big data techniques is to ensure planning of efficient farming operation, smart real time and immediate decision making, high yield, profitability and sustainability. According to the research report by market and markets, the smart agriculture market is estimated to grow from USD 13.8 billion in 2020 to USD 22.0 billion by 2025. The Indian government has introduced recently various incentive and subsidiary schemes to support new startups in agricultural sector like NEEDS, ASPIRE and NSTEDB to attract creative young generation to bring innovations in the agriculture.

Therefore, there is a biggest need for future research to meet the food demand and maintaining a good healthy agricultural environment around the globe. The farmers, agriculture equipment manufacturers, government, research universities, weather stations, technological companies, research scholars, young innovators and other related bodies can be joined as team to solve the critical issues in agriculture by using Big Data technologies. In this survey, the big data technologies, roles and challenges of big data in agriculture are discussed briefly and the future need of research are also suggested as per the growing demands.

VIII. References

[1] The Food and Agriculture Organization of the United Nations Homepage, Available online: http://www.fao.org/plant-health-2020 (accessed on May 18, 2020).

[2] Daniel Frona, Janos Szenderak and Monika Harangi-Rakos, University of Debrecen, 4032 Debrecen, Hungary. "The Challenge of Feeding the World", Published in MDPI Journal , 2019, 11, 5816; doi:10.3390/su11205816

[3] Sjaak Wolfert, Lan Ge, Cor Verdouw and Marc - Jeroen Bogaardt, "Big Data in Smart Farming - A review", published in ScienceDirect Agricultural Systems,

Vol 153, May 2017, Pages: 69-80.
https://doi.org/10.1016/j.agsy.2017.01.023

[4] Miguel A Zamora-lzquierdo, Jose Santa, Juan A Martinez, Vicente Martinez and Antonio F.Skarmeta, " Smart farming IoT platform based on edge and cloud computing", published in Elsevier, BioSystem Engineering, Vol:177, Jan 2019, Pages:4-17.
https://doi.org/10.1016/j.biosystemseng.2018.10.014

[5] Saeid Sadeghi, Darvazeh, Iman Raeesi Vanani and Farzaneh Mansouri Musolu, "Big Data Analytics and Its Applications in Supply Chain Management", published in "New Trends in the Use of Artificial Intelligence for the Industry 4.0", Mar 2020, DOI:10.5772/intechopen.89426

[6] Reihaneh H. Hariri, Erik M. Fredericks and Kate M. Bowers," Uncertainity in big data analytics: survey, opportunities and challenges", published in Springer Open, Journal of Big data, Vol 6, Article no:44(2019).

[7] Ahmed Oussous, Fatima -Zahra Benjelloun, Ayoub Ait Lahcen and Samir Belfkih, "Big Data Technologies - A Survey", published in Journal of King Saud University - Computer and Information Sciences. Vol:30, Issue:4, October 2018, Pages: 431-448.
https://doi.org/10.1016/j.jksuci.2017.06.001

[8] Orange-Rogla, Sergio, Chalmeta and Ricardo, "Framework for implementing a big data ecosystem in organizations" published in Communications of the ACM Journal, 2019, 62(1).

[9] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz and Abdullah Gani," Big Data: Survey, Technologies, Opportunities and Challenges", published in Hindawi The Scientific world journal, Vol:2014, Article Id:712826,
https://doi.org/10.1155/2014/712826.

[10] Anurag Agrahari, Prof D.T.V. Dharmaji Rao, "A Review paper on Big data: Technologies, tools and Trends",published in International Research Journal of Engineering and Technology Vol: 04 Issue: 10, 2017.

[11] Ifeyinwa Angela Ajah and Henry Friday Nweke, " A Review on Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications", published in MDPI big data and cognitive computing journal 2019, Vol:3,32; doi:10.3390/bdcc3020032

[12] PAT Research, Dark data analytics Inc, Canada company website, www.predictiveanalyticstoday.com/teradata - rainstor-analytics-with-archived-data(May 21,2020).

[13] Toby Wolpe, Big data Analytics, Article published on 2013, Available online:
https://www.zdnet.com/article/splunks-big-data-hunk-gives-hadoop-muscle-to-non-techies(May 29, 2020)

[14] Neelam Tyagi, Big Data, Article published on Mar 2020 as Top 10 Big data Technologies in 2020, Available online: https://www.analyticssteps.com/blogs/top-10-big-data-technologies-2020 (accessed May 29,2020)

[15] Varsha C. Pande and Dr. Abha S.Khandelwal, "Clustering And Classification Evaluation Using Rapid Miner", International Journal of Emerging Technologies and Innovative Research,Vol.5, Issue 9, page no.936-939, Sept 2018, http://www.jetir.org/papers/JETIR1809788.pdf

[16] The Presto Homepage "Distributed Query Engine for Big data", Available online: https://prestodb.io (accessed on Jun 18, 2020)

[17] Han Wu, Shang Zhihao and KatinkaWolter,"Performance Prediction for the Apache Kafka Messaging System", Published in IEEE 5th International Conference on Data Science and Systems on Aug 2019, DOI: 10.1109/HPCC/SmartCity/DSS.2019.00036

[18] The KNIME Homepage, https://www.knime.com (accessed on Jun 18, 2020)

[19] Dr. Madhavi Vaidya, "Big Data Storage Mechanisms and Survey of MapReduce Paradigms", Article published on Apr 2020, Available online:https://analyticsindiamag.com /big-data-storage-mechanisms-and-survey-of-mapreduce-paradigms/

[20] Vlod Krotov, "A Quick Introduction to R and R Studio", technical report published in Research Gate on Nov 2017, DOI: 10.13140/RG.2.2.10401.92009

[21]Steven Batt, Oskar R.Harmon and Paul Tomolonis,"Tableau - A Data Visualization Tool", published in Elsevier SSRN Jounal on Aug 2019, http://dx.doi.org/10.2139/ssrn.3438993

[22] The Plotly Website, https://plotly.com/graphing-libraries (accessed on Jun 22, 2020)

[23]The Tensorflow Website, https://www.tensorflow.org (accessed on Jun 22, 2020)

[24] The Apache Website," An Advanced unified Programming Model", https://beam.apache.org (accessed on Jun 22 ,2020)

[25] The Airflow Apache Website, https://airflow.apache.org/docs/stable (accessed on Jul 01, 2020)

[26] Van Evert FK, Fountas S, Jakovetic D, Crnojevic V, Travlos I and Kempenaar C. (2017). "Big Data for weed control and crop protection" published in Weed Research, https://doi.org/10.1111/wre.12255

[27] Hofmann, E. and Rutschmann, E., 2018. Big data analytics and demand forecasting in supply chains: a conceptual analysis.. The International Journal of Logistics Management, Volume 29, pp. 739-766.

[28] Desamparados Blazquez and Josep Domenech, "Big Data sources and methods for social and economic analyses", published in Elsevier Journal Technical forecasting and social change, 2018, Vol:130, Page:99-113. https://doi.org/10.1016/j.techfore.2017.07.027

[29] C. Sekhar, J.U. Kumar, B.K. Kumar, C. SekharBig data analytics on indian crop planning to increase agricultural production Adv Sci Technol Lett, Vol:147 (2018), pp. 211-216

[30] The Talend Website, Topic: Big data and Agriculture https://www.talend.com/resources/big-data-agriculture (accesed on Jul 13, 2020)

[31] Andres Villa, Henriksen, Gareth T.C. Edwards Lissa A. Pesonen, Ole Green and Claus Aage Gron Sorensena, "Internet of Things in arable farming: Implementation, applications, challenges and potential", published in Biosystem Engineering on Mar 2020, Vol:191, P:60-84.

[32] Veronica Saiz-Rubio and Francisco Rovira-Mas, "Review on From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management", published on MDPI Agronomy journal, Feb 2020, Vol: 10, 207. https://doi.org/10.3390/agronomy10020207

[33] Matthieu De Clercq, Anshu Vats and Alvaro Biel, "The Future of Farming Technology", 2018 published article on World Government Summit. Page no: 9 https://www.worldgovernmentsummit.org/api/publications/document?id=95df8ac4-e97c-6578-b2f8-ff0000a7ddb6

[34]Vinay Kellengere Shankarnarayan and Hombaliah Ramakrishna," Paradigm change in Indian agricultural practices using Big Data: Challenges and opportunities from field to plate", published in Information Processing in Agriculture, Vol:7, Issue:3, Sep 2020, Pages:355-368. https://doi.org/10.1016/j.inpa.2020.01.001

AUTHORS

First Author

E. Banu., MCA., M.Phil

Ph.D Research Scholar, PG and Research Department of Computer Science, Chikkanna Government Arts College, Tirupur - 641602. Tamil Nadu, India. Mob: +91- 9942044189
Email: banustephen4@gmail.com

Second Author

Dr. A. Geetha., M.Sc., M.Phil, Ph.D

Assistant Professor and Head, PG and Research Dept of Computer Science, Chikkanna Government Arts College, Tirupur - 641602. Tamil Nadu, India.
Email:gee_sam@yahoo. com.

Correspondence Author

E. Banu., MCA., M.Phil

Ph.D Research Scholar, PG and Research Department of Computer Science, Chikkanna Government Arts College, Tirupur - 641602. Tamil Nadu, India. Mob: +91- 9942044189
Email: banustephen4@gmail.com