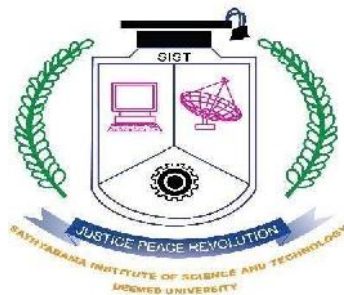


BIG DATA IN SMART FARMING

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering Degree in
Computer Science and Engineering

By

MEDULLA SAMHITHA (Roll. No. 17SCS0268)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600 119**

MARCH 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of **MEDULLA SAMHITHA (Roll. No. 17SCS0268)** who underwent the professional training in "**BIG DATA IN SMART FARMING**" under our supervision from Jan 2019 to April 2019.

Internal Guide

Mrs. SELVI, B.E, M.E., (Ph.D.)

Head of the Department

Dr. S. VIGNESHWARI, M.E., Ph.D.

Dr. L. Lakshmanan, M.E., Ph.D.

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **MEDULLA SAMHITHA (Roll. No. 17SCS0268)** hereby declare that the Project Report on “**BIG DATA IN SMART FARMING**” done by us under the guidance of guide **Mrs. SELVI, B.E, M.E., Ph.D.** Department of CSE at Sathyabama Institute of Science and Technology is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala, M.E., Ph.D. Dean, School of computing** and **Dr. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D. Head of the Department, Dept. of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Mrs. Selvi, B.E, M.E, Ph.D.** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of **Computer Science and Engineering** who were helpful in many ways for the completion of the project.

Abstract

Agriculture is the art and science of cultivating the soil, growing crops and rising livestock. It plays a key role in increasing India's Economy. Big data analysis is used to discover solutions which acts as means for analyzing data set, so that it plays a significant role in decision making in specific field such as agriculture. Harvesting yields relies upon different components like soil, climate and pesticides. The research paper facilitates the farmers in selecting the right crops and appropriate cropping pattern for a particular locality. This helps the farmers in selecting the appropriate crop. The prediction can be done using three machine learning algorithms like K-means, Random Forest and Apriori Algorithms.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	v
	LIST OF FIGURES	viii
1	INTRODUCTION	1
	1.1 SCOPE OF THE PROJECT	3
2	LITERATURE SURVEY	4
3	METHODOLOGY	11
	3.1 EXISTING SYSYTEM	11
	3.2 DATA MINING TECHNIQUES	11
	3.3 PROPOSED SYSTEM	12
	3.4 SYSTEM REQUIREMENTS	12
	3.4.1 HARDWARE REQUIREMENTS	12
	3.4.2 SOFTWARE REQUIREMENTS	13
	3.5 SYSTEM ARCHITECTURE	13
	3.6 MODULES	16
	3.6.1 DATA COLLECTION	16
	3.6.2 DATA PREPROCESSING	17
	3.6.3 HADOOP FRAMEWORK	18
	3.6.4 FINAL PREDICTION	18

	3.7 UML DESIGN	19
	3.7.1 USE CASE DIAGRAM	19
	3.7.2SEQUENCE DIAGRAM	20
	3.7.3 COLLABORATION DIAGRAM	21
	3.7.4 STATE CHART DIAGRAM	22
	3.7.5 DEPLOYMENT DIAGRAM	23
	3.8 FUNCTIONAL REQUIREMENTS	23
	3.9 SYSTEM STUDY	24
4	RESULTS AND DISCUSSION	37
5	SUMMARY AND CONCLUSION	39
6	REFERNCES	40
	APPENDIX	
	A. SOURCE CODE	41
	B. SCREEN SHOTS	50
	C. PAPER WORK	52

LIST OF FIGURES

FIGURE No.	FIGURE NAME	PAGE No.
3.1	System architecture	13
3.2	Use case diagram	20
3.3	Sequence diagram	21
3.4	Collaboration diagram	22
3.5	State chat diagram	22
3.6	Deployment diagram	23
4.1	3D Pie Chart Prediction	37
6.1	Representation of categorical diagram	50
6.2	Prediction and enhancement of crop yield by big data analysis.	50
6.3	Map reduce frame work	51

CHAPTER 1

INTRODUCTION

INDIA is a highly populated country and randomly change in the climatic conditions need to secure the world food resources. Farmers face serious problems in drought conditions. Type of soil plays a major role in the crop yield. Suggesting the use of fertilizers may help the farmers to make the best decision for their cropping situation. The number of studies Information and Communication Technology (ICT) can be applied for prediction of crop yield. By the use of Data Mining, we can also predict the crop yield. By fully analyze the previous data we can suggest the farmer for a better crop for the better yield.

For the better yield we need to consider soil type and soil fertility things and also one of the major factors rainfall and groundwater availability if it is dry land it is better to go for cash crops and if it is wetland it is better to go for wheat and sugarcane. There are 15 agro-climatic regions in India these regions are divided on the bases of a type of the land. Each agro-climatic region can grow some specific crops. Based on that we need to suggest the farmer that which crop is best among those crops which belong to those climatic regions. Achieving the maximum crop at minimum yield is the ultimate Aim of the project. Early detection of problems and management of that problems can help the farmers for better crop yield. Crop yield prediction is the important research which helps to secure food. For the better understanding of the crop yield, we need to study of the huge data with the help of machine learning algorithm so it will give the accurate yield for that crop and suggest the farmer for a better crop.

Improving the quantity of the crop is the key goal of precision agriculture means obtaining a better understanding of the crop using the information technology methods. The main goal of precision agriculture is profitability and sustainability.

From ancient times agriculture has become the backbone of our country. Nowadays climatic conditions vary very often. So, it is hard to grow crops by understanding weather conditions. We need to use some technology to find or understand the crop details and guide the farmers to grow crops accordingly and moreover fertilizer also one of the major factors to grow crops accordingly. If fertilizer is used more or less in the field the soil may lose its fertility and crop may not give the expected yield. so,

fertilizer also becomes the major factor in it. mostly understanding the temperature conditions is much necessary for India because we can improve the Indian economy with the help of the crop prediction because it plays a major role in the Indian economy. Generally, machine learning algorithms will predict the most efficient output of the yield. Previously yield is predicted on the bases of the farmers prior experience but now weather conditions may change drastically so they cannot guess the yield.so, technology can help them to predict the yield of the crop weather to go for that crop or no. machine learning model will understand the pattern of the crop and yield based on the several conditions and predicts the yield of the area in which he is going to crop. The challenge in it is to build the efficient model to predict the most efficient model to predict the output of the crop so try with the different algorithms and compare all the algorithms and which one has the less error and loss chose that model and predict the yield of that particular crop.

The human population is increasing exponentially which indicates that there is going to be higher demand for food and resources that are needed for human basic needs (www.worldback.org). Farmers in the future will need to produce more food to satisfy the peoples' and food corporations' demands. Our project was initially about finding the best combination of irrigation system and fertilizers that allows for the maximization of corn crop yields. Since every external input was considered to be constant, same irrigation system, quantity and quality of water, soil water capacity, temperature, humidity, and rainfall, we focused more on fertilizer use. The synthetic fertilizers used in Russell Ranch varied but the ones used the most were NPK, Urea, and combinations of NPK and Urea and Ammonium-sulfate (A-S) and Urea.

Fertilizers provide nutrients to plants that allow them grow faster and stronger resulting into higher yields. Fertilizers have a great impact in conventional agriculture and often they are ratified for high crop yields (Sepirl 77). On the other hand, if bad managing of fertilizers is practiced, they will have negative impacts on the environment and society. They are often blamed for polluting surface and groundwater, emitting Green House Gasses, degrading soil fertility, and are linked to human respiratory problems. The problem is not the fertilizer itself but rather its application on the fields, remember, all things in moderation. Better fertilizers management and decision-making techniques

can lead to a better relationship between fertilizers and the environment. This manifesting that negative impacts by synthetic fertilizers can be reduced and finding out which fertilizer provides the necessary nutrients for a strong plant growth and mastering its management in a more environmentally sound manner, could result into less contaminated planet.

There are other practices and disciplines in agriculture, such as the use of organic fertilizers, water efficient irrigation systems, crop rotation, biological and cultural pest control, polyculture, small scale farms, community supported agriculture, community foods, and direct sale, which leads to a sustainable agricultural system. Moving towards sustainable agriculture is something that could be considered when trying to reduce the contamination due to synthetic fertilizers. In the same way, more research should be done towards sustainable agriculture for findings about sustainable agriculture being capable to provide the same crop yields to support the planet's fast-growing population.

1.1 SCOPE OF THE PROJECT

The prediction will also aid the farmers to make decisions such as the choice of alternative crops or to discard a crop at an early stage in case of critical situations. Further, predicting crop yield can facilitate the farmers to have a better vision on cultivation of seasonal crop and its scheduling. Thus, it is necessary to simulate & predict the crop yield before cultivation for efficient crop management and expected outcome. As there exists a non-linear relationship between crop yield and the factors influencing crop, Big Data techniques might be efficient for yield predictions.

CHAPTER 2

LITERATURE SURVEY

2.1 SURVEY 1

Suvidha Jambekar ; Shikha Nema ; Zia Saquib “Prediction of Crop Production in India Using Data Mining Techniques” 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)

Agriculture sector is the primary source of food and plays a major role for the Indian economy and employment. India is the top producer of major crops such as Rice, Wheat and Maize. Data mining is the study of extracting useful information from the data sets. The present study focuses on the application of data mining techniques to predict future production of crops such as Rice, Wheat and Maize with respect to various parameters observed during the period (1950-2013). The parameters considered for the study were rainfall, mean temperature, area under irrigation, area, production and yield. The regression algorithms used in the study were Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth). The experimental results show that the performance of Multivariate Adaptive Regression Splines (Earth) is better than Multiple Linear Regression and Random Forest Regression for Rice and Wheat dataset and the performance of Multiple Linear Regression is better than Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) for Maize dataset.

2.2 SURVEY 2

Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review , Computers and Electronics in Agriculture 151 (2018) 61–69,Elisver,2018.

Accurate yield estimation and optimised nitrogen management is essential in agriculture. Remote sensing (RS) systems are being more widely used in

building decision support tool us for contemporary farming systems to improve yield production and nitrogen management while reducing operating costs and environmental impact. However, RS based approaches require processing of enormous amounts of remotely sensed data from different platforms and, therefore, greater attention is currently being devoted to machine learning (ML) methods. This is due to the capability of machine learning based systems to process a large number of inputs and handle non-linear tasks. This paper discusses research developments conducted within the last 15 years on machine learning based techniques for accurate crop yield prediction and nitrogen status estimation. The paper concludes that the rapid advances in sensing technologies and ML techniques will provide cost-effective and comprehensive solutions for better crop and environment state estimation and decision making. More targeted application of the sensor platforms and ML techniques, the fusion of different sensor modalities and expert knowledge, and the development of hybrid systems combining different ML and signal processing techniques are all likely to be part of precision agricultural (PA) in the near future.

2.3 SURVEY 3

Prakash Kumar, Anil Kumar, Sanjeev Panwar, Sukanta Dash, Kanchan Sinha, Vipin Kumar Chaudhary and Mrinmoy Ray “Role of big data in agriculture- A statistical prospective” Agric. Res. New Series Vol. 39 (2) : 210-215 (2018)

Data are playing an important role making good planning and policies for agricultural growth and development. Population growth and climate change are worldwide trends that are increasing the importance of using big data science to improve agriculture. Add to that land degradation increasing marginal land and loss of biodiversity are better deals with study of big data science. Crop data can be break down into bits and bytes it will give better study about the crop development by using advance data analytics tools for betterment of agriculture. Here, talk about some important tools and techniques to handle and study the big data.

2.4 SURVEY 4

Shriya Sahu et al," An Efficient Analysis of Crop Yield Prediction Using Hadoop framework based on random Forest approach, International Conference on Computing, Communication and Automation , 2017.

In the growth of Information technology, Big data come forth as a blazing topic. The main source of human survival depends on agriculture; where it needs a key contribution in the field of crop data analysis. This paper gives a purpose about how to find experiences from accuracy agriculture information through big data approach. In this way, gathering the valuable data in an effective way drives a framework towards major computational challenges in crop analysis where information is remotely gathered. For the storage purpose of huge data availability in agriculture, we are intending Hadoop framework for our work to store a huge volume of crop data. This work gives a better prediction for the farmers to plant which kind of crops to their farm field based on their soil content to improve the productivity. The random forest algorithm is integrated with the MapReduce programming model in Hadoop framework.

2.5 SURVEY 5

Ramya M, Chetan Balaji, Girish L ,” Environment Change Prediction to Adapt Climate-Smart Agriculture Using Big Data Analytics”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 5, May 2015 .

The increased climate variability is challenging to farmers. It will expect to influence on crop and livestock productivity. So new approaches are required to farmers with updated and relevant information so as to support them in the decision-making process, and to make them more resilient to climate change. The main objective of this project is to develop an application that predicts the environmental changes and to assist farmers in adapting to climate-smart Agriculture systems using big data Approach which increase the income and productivity of farmers. This paper is aiming to find Climate-Smart, Site-Specific Agriculture Decision-Making to problems, such as climate change-induced global food insecurity, to predicting and mitigating the impact of

extreme weather events on global finance and also we can predict the Economic Dimensions of Climate Change around the world using Data and Analytics.

2.4 SURVEY 4

Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh,” Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique”,Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.138-145

Agriculture planning plays a significant role in economic growth and food security of agro-based country. Selection of crop(s) is an important issue for agriculture planning. It depends on various parameters such as production rate, market price and government policies. Many researchers studied prediction of yield rate of crop, prediction of weather, soil classification and crop classification for agriculture planning using statistics methods or machine learning techniques. If there is more than one option to plant a crop at a time using limited land resource, then selection of crop is a puzzle. This paper proposed a method named Crop Selection Method (CSM) to solve crop selection problem, and maximize net yield rate of crop over season and subsequently achieves maximum economic growth of the country. The proposed method may improve net yield rate of crops.

2.5 SURVEY 5

Veenadhari, S., Bharat Misra, D Singh, “Data mining Techniques for Predicting Crop Productivity – A review article”, IJCST, International Journal of Computer Science and technology march 2011.

The data mining extraction of unseen predictive information from huge records, is a powerful new technology with great potential to help companies focus on the most significant data in their Data warehouses. Data mining tools predict upcoming trends and performance and growth, allowing businesses to make proactive, knowledge driven decisions. Though these methods are plausible, theoretically well created, and perform

well on extra or less artificial test data sets, they depend on their skill to make sense of real-world data. This article gave us a detail project that is smearing a range of machine learning plans to problems in agriculture and horticulture. They briefly surveyed some of the techniques emerging from machine learning study, define a software workbench for testing with a variability of methods on real-world data sets, and a learning of dairy herd management in that culling rules were inferred from a medium-sized record of herd information. They also defined a range of machine learning plans to problems in agriculture and horticulture. There is a rising number of applications of data mining methods in agriculture and a rising amount of data that are presently available from several resources. This is relatively a novel research field and it is expected to grow in the upcoming. There is a lot of effort to be done on this emerging and interesting study field. The multidisciplinary method of integrating computer science with agriculture will help in predicting managing agricultural crops effectively.

2.6 SURVEY 6

M.S.PrasadBabu, N.V.Ramana Murty, S.V.N.L.Narayana, “A Web Based Tomato Crop Expert Information System Based on Artificial Intelligence and Machine Learning Algorithms”, IJCSIT, Vol. 1 (1), 2010, 6-15.

The tomato is now the most widely grown vegetable crop in World. It is grown up through the world in farm gardens, small homegardens, and by market gardeners for fresh consumption as well as for processing purposes. This Tomato crop skilful advisory system is intended at a collaborative venture with eminent Agriculture Scientist and Experts in the area of Tomato Plantation with an excellent team of computer Engineers, computer programmer and creators. This Expert System contains two main parts one is Tomato Info System and the other is Tomato Crop Expert System where in Data system, the user can get all the static information about different species, Illnesses, Symptoms, chemical controls, Preventions, Pests, Virus of Tomato fruits and plants. In Advisory System, the user is having a communication with the expert system online; the user has to answer the questions asked by the Expert System. Depends on the reply by the operator the expert system decides the disease and shows its control

measure of disease. This Tomato Crop Data Expert System deals with different varieties of Tomato Crop, Identification of various diseases usually chances to tomato crop based on the symptoms.

2.7 SURVEY 7

Leemans V, M F Destain, "A Real Time Grading Method of Apples Based on Features Extracted from Defects", J. Jood Eng., 2004, pages: 83-89.

The suggested Fresh market fruits like apples are graded into quality groups according to their size, color and shape and to the attendance of defects. The two first quality criteria are actually automatic on industrial graders, then fruits classifying according to the presence of faults is not yet efficient and so remains a manual operation, repetitive, luxurious and not reliable. The classifying of apples using machine vision can be arbitrarily separated into four steps: the images acquisition, their segmentation, their interpretation and in conclusion the fruit classification. This paper presents the three former points on the basis of a literature review, the research outcomes being absorbed on the last point: having extracted data from images acquired on fruits, the paper defines a classifying technique which was implemented on an existing machine and tested on Jon gold apples (bi-color fruits). The first step consists of acquiring images of the surface of the fruit, though it goes through the classifying machine. In order to grade apples, two necessities have to be met: the images should cover the entire surface of the fruit; a high contrast has to be created among the defects and the healthy tissue, while maintaining a low variability for the healthy tissue.

2.8 SURVEY 8

R J Brooks, "Simplifying Sirius: Sensitivity Analysis and Development of a Meta-Model for Wheat Yield Prediction", European Journal of Agronomy, vol. 14, 2001

A simpler meta-model, which produced very similar yield predictions to Sirius of potential and water-limited yields at two locations in the UK, Roth Amsted and Edinburgh. This greatly increases the understanding of the nature and consequences

of the relationships implicit within Sirius. The study showed that the reply of wheat crops to climate could be explain using a few simple relationships. The meta-model aggregates the three main Sirius components, the computation of leaf area index, the soil water balance model and the evapotranspiration calculation, into simpler equations. These results in a obligation for calibration of fewer model parameters and means that weather variables can be provide on a monthly rather than a daily time-step, because the meta-model can use cumulative values of weather variables. As a result the meta-model is a valuable tool for regional impact assessments when detailed input data are usually not available. As the meta-model was developed from the analysis of Sirius, rather than from statistical fitting of yield to weather data, it must do well for other locations in Great Britain and with different management scenarios.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

In the existing studies, the variety of methods is implemented to classify the agriculture data. In the existing system, only a few of many factors available are considered for the agricultural data classification. Many of the existing systems are implemented using Data Mining techniques where the accuracy of classifying the data is lower when compared to the big data analytics. The existing systems do not consider the dependency between the various factors affecting the crop production. And more over the existing does use the old data which is not up to date. Thus, the classification and prediction of data can't be true and accurate. Thus, using up to date data with the "Big Data analytics" helps in accurate classification and prediction of data.

3.2 DATA MINING TECHNIQUES

Regression analysis is a predictive modeling technique which estimates linear relationship between dependent variable and one or more independent variables. The dependent variable is also called as predictand and independent variables is also called as predictors. Considering production as dependent (target) variable and area, area under irrigation, monthly rainfall and monthly mean temperature as independent (feature) variables, regression analysis was conducted using 1950-2013 (64 years) data for Rice, Wheat and Maize. Regression algorithms used in this research were Multiple Linear Regression and Multivariate Adaptive Regression Splines (Earth).

3.1.1 DISADVANTAGES OF THE EXISTING SYSTEM

-) Access to quality and mechanized inputs is a major challenge that has affected yield, profitability and the development of the agricultural sector.
-) Limited storage facilities, high cost of processing equipment and power generation are some of the core challenges that face the transformation phase of the agricultural value chain.

3.3 PROPOSED SYSTEM

The proposed work is intended to help the farmers and researchers to understand the crop field and to choose a suitable crop for it. Various parameters are considered from soil to atmosphere for predicting the suitable crop. Soil parameters such as type, ph level, iron, copper, manganese, sulphur, organic carbon, potassium, phosphate, nitrogen. By considering the entire parameter in agricultural field benefits the prediction process in a better way. The K-Means, Random Tree and Apriori algorithm is used to classify the dataset which provides result in good accuracy with poor error rate. Since this framework can handle large dataset by processing it in MapReduce programming model.

3.3.1 ADVANTAGES OF THE PROPOSED SYSTEM

- ☐ Faster and Better decision making by using distribute solutions.
- ☐ Extensibility and More Reliable way of analyzing data.

3.4. SYSTEM REQUIREMENTS

3.4.1 HARDWARE REQUIREMENTS

- System: I3 PROCESSOR
- Hard Disk: 300 GB.
- Floppy Drive: 1.44 Mb.
- Monitor: 15 VGA Colour.
- Mouse : Logitech.
- Ram: 4 gb.

3.4.2 SOFTWARE REQUIREMENTS

- Operating system: - Windows 10/Ubuntu
- Coding Language: java for Mapper and Reducer
- Front End: Php, JavaScript (Intelligent Graph)
- Back End: Hadoop Cluster
- Tool: Virtual Box Oracle tool

3.5 SYSTEM ARCHITECTURE

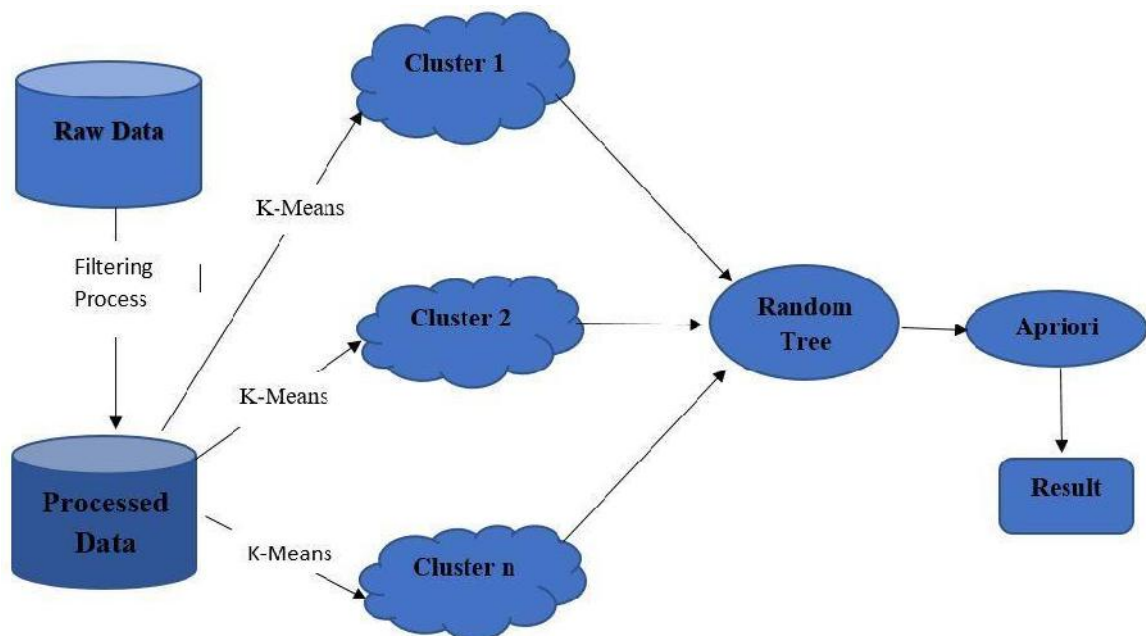


FIG 3.1 ARCHITECTURE OF THE SYSTEM

we collect all the required dataset. Regarding the datasets the initial description and the point to be remembered is about the attributes. The dataset regarding the attributes which suits the project must be analysed such that the entire results depend on the dataset collected and attributes containing in it. The data set collected are crops vs. seasons, crops vs. price(various years), seasons vs. temperature(various years). In datasets collected may contain various null values, inconsistent values and datasets may also be in different formats. In this process of the pre-processing, we will remove null values and inconsistent values. The null values reduce the accuracy of prediction

of the data. And inconsistent data reduces and confuses the algorithm. In order to pre-process the data, we will be using the Hadoop Tool. The pre-processed data helps in more accurate prediction of the data. Map reduction technique in the Hadoop tool is used to get the data that's only required. Map reduction technique reduces the amount of data needed to be processed by the classification algorithm is reduced. Data classification and clustering: Data classification is the process (or) procedure of sorting and categorizing data into various types, forms which will help us more when we are analysing the data. Data classification helps us a lot to do the separation of data and classification of data according to data set requirements for various business motives or personal objectives. It is mainly a data management process. Data classification helps in the prediction of the data.

A. Analysis Algorithm Get Data from DFS

```
Varval, total.avg  
For each val then  
Begin  
  ReadFile()  
End For each total then  
Begin  
  Total+=val  
End  
Avg=total/N;
```

We get the data from the Hadoop Distributed File System (HDFS) and read it line by line, and perform normalization for the data by taking statistical average mean of data.

B. Classification Algorithm

```
Varmean,sum  
Var pi= 1068966896 / 340262731;
```

Get Normalized Data from DFS

For each row then

Begin

Sum+=value

End

Return $\sqrt{(\text{sumSq} - ((\text{sum} * \text{sum}) / n)) / (n - 1)}$

Our classification algorithm used is a statistical based learning system having the following qualities:

- It works astonishingly well
- Fast learning and fast training
- Small runtime memory required
- Simple and perfect

The learning can be completed via statistical modeling. Their attributes are similarly important whereas statistically independent. This implies that learning about the estimation of a specific trait doesn't let us know anything about the estimation of another quality.

C. Prediction Algorithm

Var like, what, k class

Get Classified Data from DFS

If like then

Begin

Likelihood (like, k class)

End

Function likelihood (like, k class)

Begin

Case CROP:

put In Crop(like);

break;

Case SOIL:

put In Soil(like);

break;

Case DISEASE:

put In Disease(like);

break;

default:

return k class;

End

Here we classify data based on disease and the data extracted from the classifier and this helps to predict the soil and crop.

3.6 MODULES

The phase of the proposed work

1. Data Collection
2. Data Classification (K-Means, Random Tree and Apriori algorithm)
3. Hadoop Framework – MapReduce programming model
4. Final Prediction

3.6.1. Data Collection

The dataset can be gathered in two ways: One is fetching data from the online resource and the other is from the sensor node data. The online datasets are readily available in the form of data matrix or single database table. The dataset may contain more data which are in structured or unstructured formats. Since some dataset is highly secured need some initial registration and authentication process to access. The other form of data is sensor datasets, which mainly used by an organization/environment. The organization/environment has a sensor node infrastructure where the data are collected as per their programming framework and store it in the backend (Cloud Storage).

3.6.2 Dataset pre-processing

Dataset pre-processing was done to collect all the datasets in Microsoft Office Excel. For every crop it consists of following columns: year, area, production, yield, area under irrigation, monthly rainfall and monthly mean temperature. When we divide Production by Area, we get values very close to Yield. we should take either production or yield column. Considering Production and removing Yield the file was saved as .csv.

1. K-means Clustering Algorithm

K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached.

Input: $D = \{t_1, t_2, t_3, \dots, t_n\}$ //Set of elements

k //Number of desired clusters

Output: K //set of clusters.

Algorithm:

assign initial values for means $m_1, m_2, \dots, m_k$;

repeat

assign each item t_i to the cluster which has closest mean;

calculate new mean for each cluster;

until convergence criteria is met.

2. Random Tree

The algorithm can deal with both classification and regression problems. A random tree is a collection (ensemble) of tree predictors that is called forest further in this section (the term has been also introduced by L.Breiman). The classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". The soil presented in various places of district and these places are classified under different soil series based on the soil physical and chemical factors of soil using Random Tree Classifier. It shows the different types of Agri block and these blocks are splitted according to the soil factors. The soil samples collected from various places of district are analyzed and the factors of the soils are similar which were found in same series

3. Apriori

The apriori Mining process is composed of two major steps. The first one (generating frequent item sets) of the apriori algorithm simply counts item occurrences to determine the large 1-itemsets. This step can be seen as support based pruning, because only item sets with at least minimum support were considered. The second step is the generation of rules out of the frequent item sets.

In the Datasets nearly two hundred and fifty soil samples were collected from various places to predict which soil is suitable for selected crop. Association rules are used to find out elements that co-occur repeatedly within a data set consisting of many independent selections of elements (such as purchasing transactions), and to discover rules. Finally, ARM is used to find the suitable crop for appropriate soil.

3.6.2 Hadoop Framework

A Classical approach in Hadoop framework consists of loading data into the HDFS and the MapReduce programming model is used for processing the dataset. The MapReduce programming model works on both the structured and the unstructured data. The map task is done by mapper class and the reduce task is done the reducer class. The mapper class generally tokenizes the input data to sort it and this will act as the input for the reducer task. Reducer class removes the duplicate entries from the key-value pair. The primary characteristics of the MapReduce is to merge the input <key, value> pair with another list of <key, value> pair. The input is processed by mapper function to produce the intermediate set of <key, value> pair. The reducer merger the output produced from mapper to form a smaller set of values.

3.6.4 Final Prediction

The final stage is the suitable crop based on the prediction obtained from the dataset containing the parameter such as Land type, Temperature, Nitrogen and Crop.

3.7 SYSTEM DESIGN

UML DIAGRAM OF THE PROJECT

UML is a standard language for specifying, visualizing, and documenting of software systems and created by Object Management Group (OMG) in 1997. There are three important type of UML modelling are Structural model, Behavioural model, and Architecture model. To model a system the most important aspect is to capture the dynamic behaviour which has some internal or external factors for making the interaction. These internal or external agents are known as actors. It consists of actors, use cases and their relationships. In this fig we represent the Use Case diagram for our project.

3.7.1 Use case diagram

A use case is a set of scenarios that describing an interaction between a user and a system. A use case diagram displays the relationship among actors and use cases. The two main components a user or another system that will interact with the system modelled. A use case is an external view of the system that represents some action the user might perform in order to complete a task.

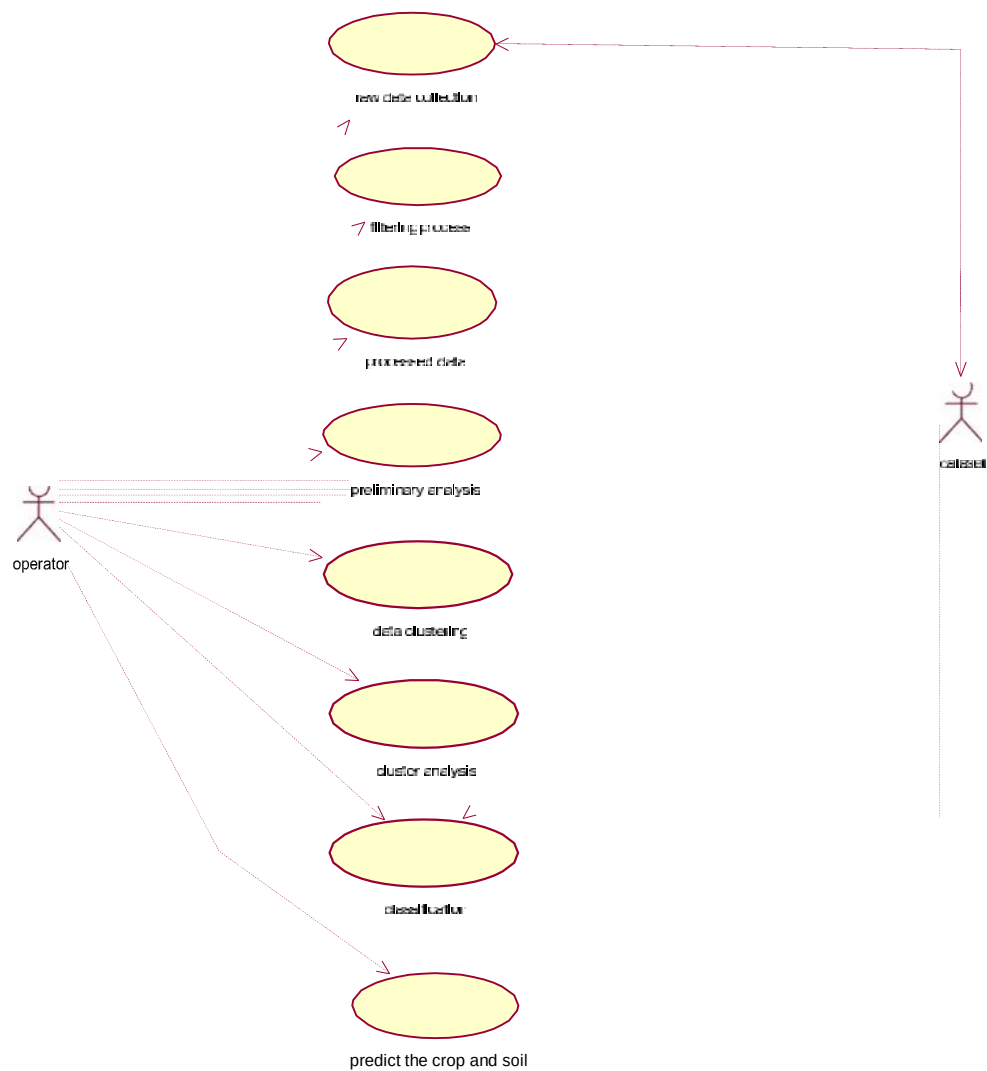


FIG 3.2 USE CASE DIAGRAM

3.7.2 SEQUENCE DIAGRAM

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

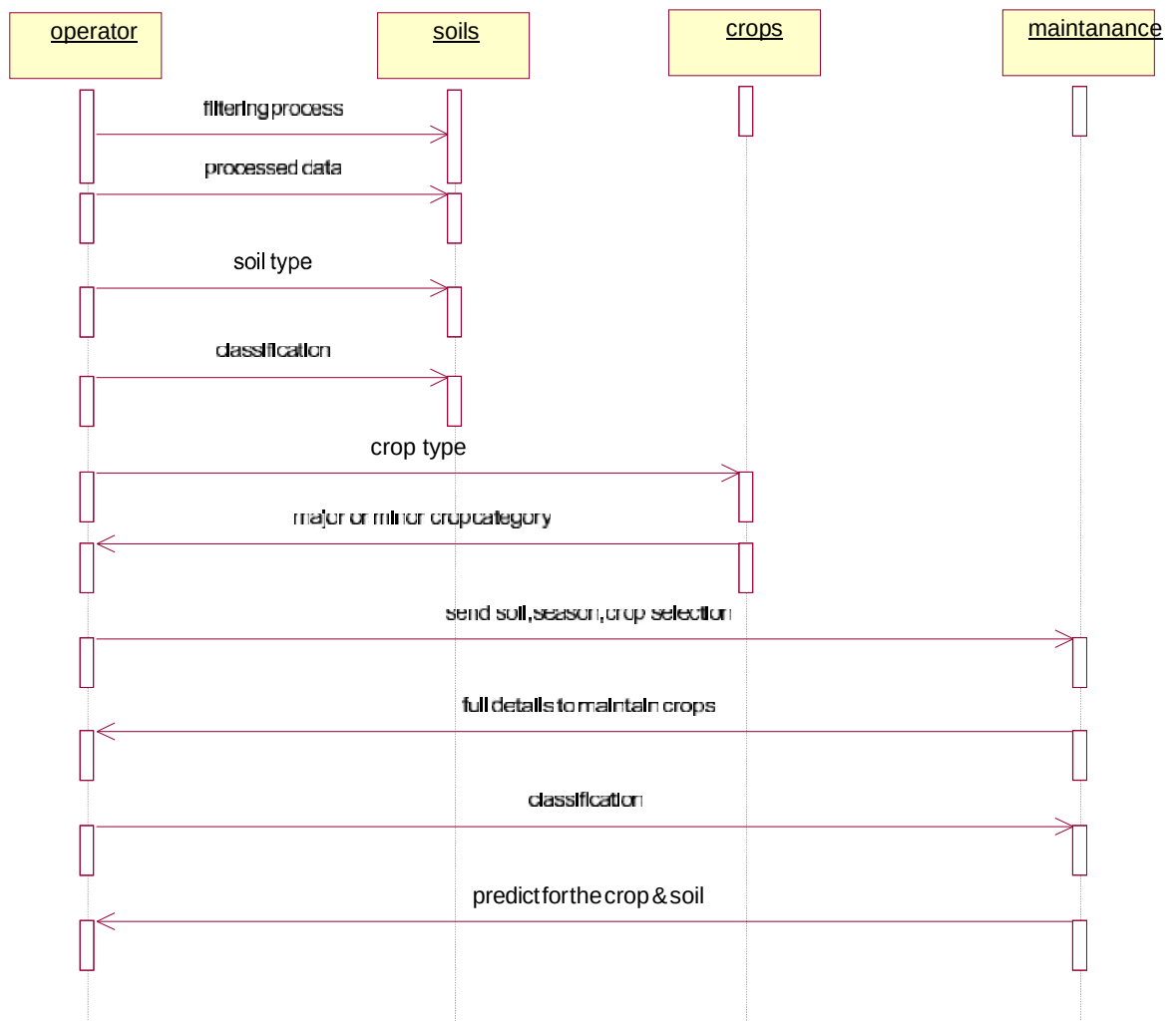


FIG 3.3 SEQUENCE DIAGRAM

3.7.3 COLLABORATION DIAGRAM

A collaboration diagrams belongs to a group of UML diagrams called interaction diagrams. Collaboration diagrams, like sequence diagrams, show how objects interact over the course of time. However, instead of showing the sequence by numbering the messages of the diagram. This makes it easier to show how the objects are linked together, but harder to see the sequence at a glance.

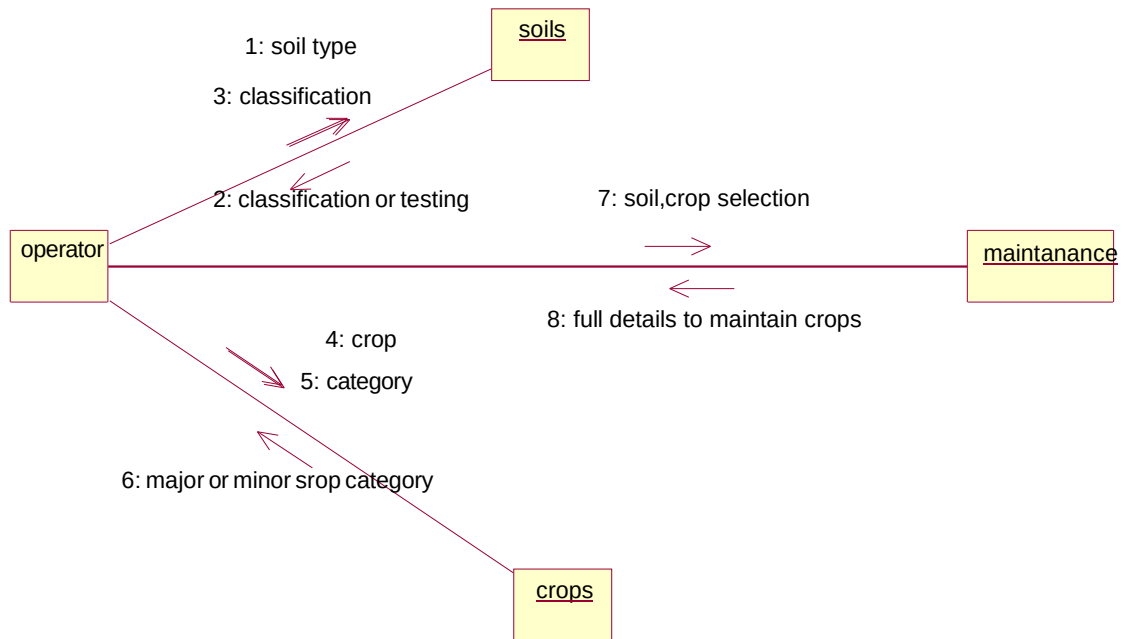


FIG 3.4 COLLABORATION DIAGRAM

3.7.4 STATECHART DIAGRAM

State chart diagrams provide us an efficient way to model the interactions or communication that occur within the external entities and a system. These diagrams are used to model the event-based system. A state of an object is controlled with the help of an event.

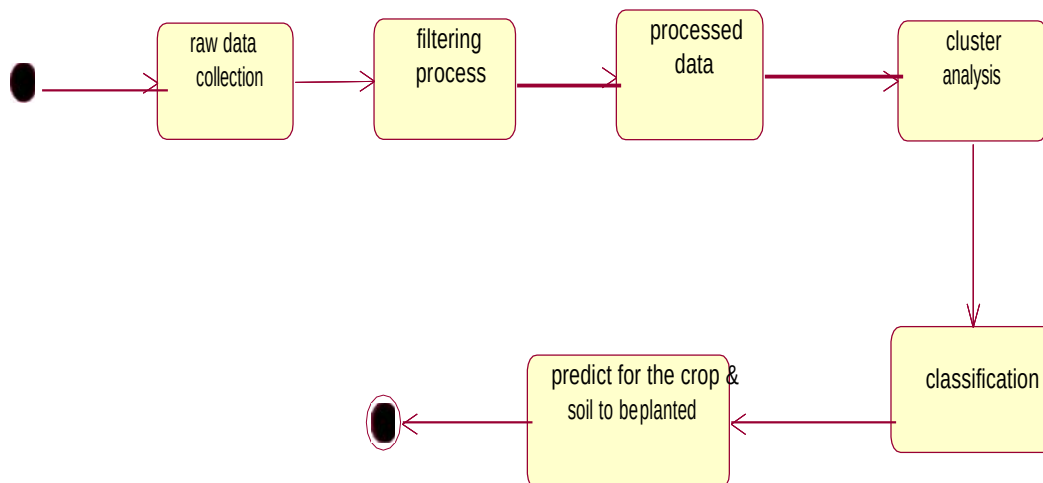


FIG 3.5 STATECHART DIAGRAM

3.7.5 DEPLOYMENT DIAGRAM

Deployment Diagram is a type of diagram that specifies the physical hardware on which the software system will execute. It also determines how the software is deployed on the underlying hardware. It maps software pieces of a system to the device that are going to execute it.

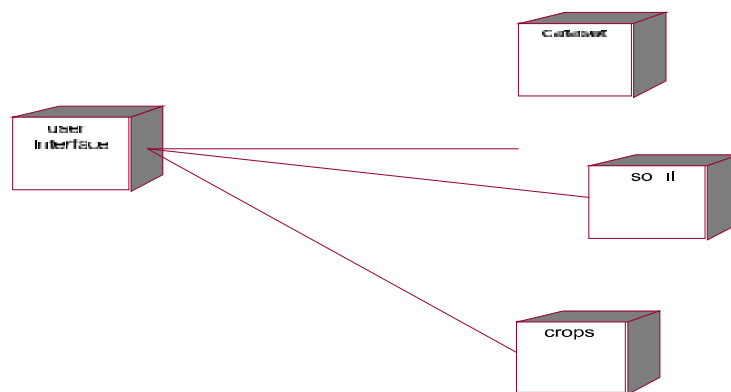


FIG 3.6 DEPLOYMENT DIAGRAM

3.8 FUNCTIONAL REQUIREMENTS

A functional requirement defines a function of a software-system or its component. A function is described as a set of inputs, the behavior, and outputs. Our system requires minimum three systems to achieve this concept.

OTHER FUNCTIONAL ATTRIBUTES

Modifiability

Requirements about the effort required to make changes in the software. Often, the measurement is personnel effort.

Portability

The effort required to move the software to a different target platform.

Security

One or more requirements about protection of your system and its data. The measurement can be expressed in a variety of ways to break into the system. Do not discuss solutions (e.g. passwords) in a requirements document.

Usability

Requirements about how difficult it will be to learn and operate the system. The requirements are often expressed in learning time or similar metrics.

3.9 SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- © ECONOMICAL FEASIBILITY
- © TECHNICAL FEASIBILITY
- © SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased. This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client.

The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

OBSTACLES

There are numerous issues adopting quality farming in all the countries dependent on farming across the globe. These issues differ in terms of regions but some issues are specific to the Indian conditions as mentioned below.

- Institutional constraints, land possession and infrastructure
- Observations and Culture of the users
- Technical gaps and knowledge
- Small farm size are smaller
- Lack of Success stories are very less
- Local mechanical expertise are very few
- Market failures and Heterogeneity of harvesting systems
- Local mechanical expertise are very few
- Quality, costs and data availability.

3.10 SOFTWARE ENVIRONMENT

BIG DATA PLATFORM

PURPOSE

-) The concept of big data has been around for more than a decade – but while its potential to transform the effectiveness, efficiency, and profitability of virtually

any enterprise has been well documented, the means to effectively leverage big data and realize its promised benefits still eludes some organizations. Ultimately, there are two main hurdles to tackle when it comes to realizing these benefits.

- J The first is realizing that the real purpose of leveraging big data is to take action – to make more accurate decisions and to do so quickly. We call this situational awareness. Regardless of industry or environment, situational awareness means having an understanding of what you need to know, what you have control of, and conducting analysis in real time to identify anomalies in normal patterns or behaviors that can affect the outcome of a business or process. If you have these things, making the right decision within the right amount of time in any context becomes much easier.
- J Defining these parameters for any industry is not simple, and thus surmounting Big Data's other remaining challenge of creating new approaches to data management and analysis is also no small feat. Achieving situational awareness used to be much easier because data volumes were smaller, and new data was created at a slower rate, which meant our world was defined by a much smaller amount of information. But new data is now created at a hugely exponential rate, and therefore any data management and analysis system that is built to provide situational awareness today must also be able to do so tomorrow. So, the imperative for any enterprise is not to just create systems that manage Big Data and provide situational awareness, but to build systems that provide *scalable* situational awareness.
- J Take, for instance, the utilities industry. This space is in particular need of scalable situational awareness so that they can realize benefits for a wide range of important functions critical for enabling Smart Grid paradigms. A properly-functioning power grid network shifts power around to where it is needed. Scalable situational awareness for utilities then means knowing where power is needed, and where it can be taken from, to keep the grid stable. When power flow is not well understood its direction will start changing rapidly, moving energy around like a power hurricane.

- J As with any hurricane, at the middle there is an eye that is totally quiet and dark (a fitting, although ironic, analogy considering the goal of awareness). This is what happened in 2003, during one of the worst blackouts to hit the Northeast. The various power companies involved were quickly analyzing all of the information but, despite the fact that they were all communicating, they didn't know what exactly to do to alleviate the drastic shift in power flow and, thus, ended up making the wrong decisions that resulted in the blackout.
- J If situational awareness had been present, the blackout could have been prevented. This seems especially relevant given the recent blackout in India, and begs the question: is the U.S. aware enough of the potential dangers that we have taken steps to enable our smart grid to respond in the correct way to avoid such outages?
- J Utilities in the U.S. and beyond can learn much about how to achieve scalable situational awareness from other industries, most notably building management and telecommunications, which have learned to deal with Big Data's complexity and scale well. For industries like utilities to achieve scalable situational awareness, it requires building standards-based, interoperable, and scalable data management systems.

3.11 SCOPE

The Scope of the Big data and Analytics are:

1. Visual data discovery tools will be growing 2.5 times faster than rest of the Business Intelligence (BI) market. By 2018, investing in this enabler of end-user self-service will become a requirement for all enterprises.
2. Over the next five years spending on cloud-based Big Data and analytics (BDA) solutions will grow three times faster than spending for on-premise solutions. Hybrid on/off premise deployments will become a requirement.
3. Shortage of skilled staff will persist. In the U.S. alone there will be 181,000 deep analytics roles in 2018 and five times that many positions requiring related skills in data management and interpretation.

4. By 2017 unified data platform architecture will become the foundation of BDA strategy. The unification will occur across information management, analysis, and search technology.
5. Growth in applications incorporating advanced and predictive analytics, including machine learning, will accelerate in 2015. These apps will grow 65% faster than apps without predictive functionality.
6. 70% of large organizations already purchase external data and 100% will do so by 2019. In parallel more organizations will begin to monetize their data by selling them or providing value-added content.
7. Adoption of technology to continuously analyze streams of events will accelerate in 2015 as it is applied to Internet of Things (IoT) analytics, which is expected to grow at a five-year compound annual growth rate (CAGR) of 30%.
8. Decision management platforms will expand at a CAGR of 60% through 2019 in response to the need for greater consistency in decision making and decision-making process knowledge retention.
9. Rich media (video, audio, image) analytics will at least triple in 2015 and emerge as the key driver for BDA technology investment.
10. By 2018 half of all consumers will interact with services based on cognitive computing on a regular basis.

GENERAL DESCRIPTION OF HADOOP:

Hadoop is an Apache Software Foundation project that importantly provides two things:

A distributed filesystem called HDFS (Hadoop Distributed File System)

A framework and API for building and running MapReduce jobs

HDFS

HDFS is structured similarly to a regular Unix file system except that data storage is distributed across several machines. It is not intended as a replacement to a regular file system, but rather as a file system-like layer for large distributed systems to use. It has in built mechanisms to handle machine outages, and is optimized for throughput rather than latency.

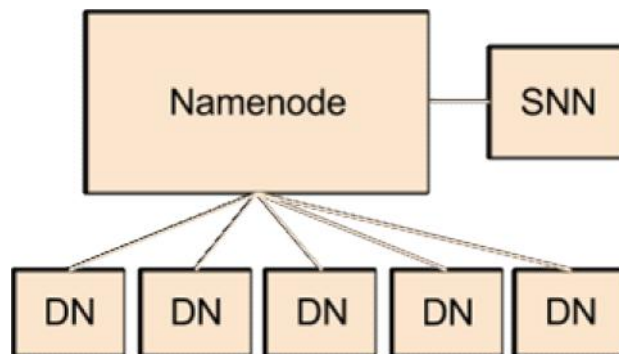
There are two and a half types of machine in a HDFS cluster:

Data node - where HDFS actually stores the data, there are usually quite a few of these.

Name node - the 'master' machine. It controls all the meta data for the cluster. Eg - what blocks make up a file, and what data nodes those blocks are stored on.

Secondary Name node - this is NOT a backup name node, but is a separate service that keeps a copy of both the edit logs, and file system image, merging them periodically to keep the size reasonable.

this is soon being deprecated in favor of the backup node and the checkpoint node, but the functionality remains similar (if not the same)



Data can be accessed using either the Java API, or the Hadoop command line client. Many operations are similar to their Unix counterparts. Check out the documentation page for the full list, but here are some simple examples:

list files in the root directory

```
hadoop fs -ls /
```

list files in my home directory

```
hadoop fs -ls ./
```

cat a file (decompressing if needed)

```
hadoop fs -text ./file.txt.gz
```

upload and retrieve a file

```
hadoop fs -put ./localfile.txt /home/matthew/remotefile.txt
```

```
hadoop fs -get /home/matthew/remotefile.txt ./local/file/path/file.txt
```

Note that HDFS is optimized differently than a regular file system. It is designed for non-realtime applications demanding high throughput instead of online applications demanding low latency. For example, files cannot be modified once written, and the latency of reads/writes is really bad by filesystem standards. On the flip side, throughput

scales fairly linearly with the number of datanodes in a cluster, so it can handle workloads no single machine would ever be able to.

HDFS also has a bunch of unique features that make it ideal for distributed systems:

Failure tolerant - data can be duplicated across multiple datanodes to protect against machine failures. The industry standard seems to be a replication factor of 3 (everything is stored on three machines).

Scalability - data transfers happen directly with the datanodes so your read/write capacity scales fairly well with the number of datanodes

Space - need more disk space? Just add more datanodes and re-balance

Industry standard - Lots of other distributed applications build on top of HDFS (HBase, Map-Reduce)

Pairs well with MapReduce - As we shall learn

HDFS Resources

For more information about the design of HDFS, you should read through [apache documentation page](#). In particular the streaming and data access section has some really simple and informative diagrams on how data read/writes actually happen.

MapReduce

The second fundamental part of Hadoop is the MapReduce layer. This is made up of two sub components:

An API for writing MapReduce workflows in Java.

A set of services for managing the execution of these workflows.

THE MAP AND REDUCE APIS

The basic premise is this:

Map tasks perform a transformation.

Reduce tasks perform an aggregation.

In scala, a simplified version of a MapReduce job might look like this:

```
def map(lineNumber: Long, sentence: String) = {  
  val words = sentence.split()  
  words.foreach{word =>  
    output(word, 1)  
  }  
}
```

```
}
```

```
def reduce(word: String, counts: Iterable[Long]) = {  
  var total = 0l  
  counts.foreach{count =>  
    total += count  
  }  
  output(word, total)  
}
```

Notice that the output to a map and reduce task is always a KEY, VALUE pair. You always output exactly one key, and one value. The input to a reduce is KEY, ITERABLE[VALUE]. Reduce is called exactly once for each key output by the map phase. The ITERABLE[VALUE] is the set of all values output by the map phase for that key.

So if you had map tasks that output

map1: key: foo, value: 1

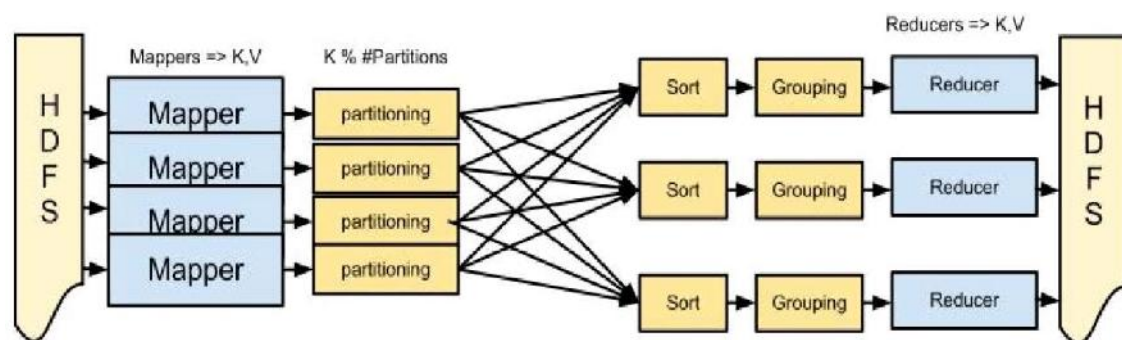
map2: key: foo, value: 32

Your reducer would receive:

key: foo, values: [1, 32]

Counter intuitively, one of the most important parts of a MapReduce job is what happens between map and reduce, there are 3 other stages; Partitioning, Sorting, and Grouping. In the default configuration, the goal of these intermediate steps is to ensure this behavior; that the values for each key are grouped together ready for the reduce() function. APIs are also provided if you want to tweak how these stages work (like if you want to perform a secondary sort).

Here's a diagram of the full work flow to try and demonstrate how these pieces all fit together, but really at this stage it's more important to understand how map and reduce interact rather than understanding all the specifics of how that is implemented.



The MapReduce Pipeline

A mapper receives (Key, Value) & outputs (Key, Value)
 A reducer receives (Key, Iterable[Value]) and outputs (Key, Value)
 Partitioning / Sorting / Grouping provides the Iterable[Value] & Scaling

What's really powerful about this API is that there is no dependency between any two of the same tasks. To do its job a map () task does not need to know about other map task, and similarly a single reduce task has all the context it needs to aggregate for any particular key, it does not share any state with other reduce tasks.

Taken as a whole, this design means that the stages of the pipeline can be easily distributed to an arbitrary number of machines. Workflows requiring massive datasets can be easily distributed across hundreds of machines because there are no inherent dependencies between the tasks requiring them to be on the same machine.

MapReduce API Resources

If you want to learn more about MapReduce (generally, and within Hadoop) I recommend you read the Google MapReduce paper, the Apache MapReduce documentation, or maybe even the hadoop book. Performing a web search for MapReduce tutorials also offers a lot of useful information.

To make things more interesting, many projects have been built on top of the MapReduce API to ease the development of MapReduce workflows. For example, Hive lets you write SQL to query data on HDFS instead of Java. There are many more examples, so if you're interested in learning more about these frameworks, I've written a separate article about the most common ones.

THE HADOOP SERVICES FOR EXECUTING MAPREDUCE JOBS

Hadoop MapReduce comes with two primary services for scheduling and running MapReduce jobs. They are the Job Tracker (JT) and the Task Tracker (TT). Broadly speaking the JT is the master and is in charge of allocating tasks to task trackers and scheduling these tasks globally. A TT is in charge of running the Map and Reduce tasks themselves.

When running, each TT registers itself with the JT and reports the number of 'map' and 'reduce' slots it has available, the JT keeps a central registry of these across all TTs and allocates them to jobs as required. When a task is completed, the TT re-registers that slot with the JT and the process repeats.

Many things can go wrong in a big distributed system, so these services have some clever tricks to ensure that your job finishes successfully:

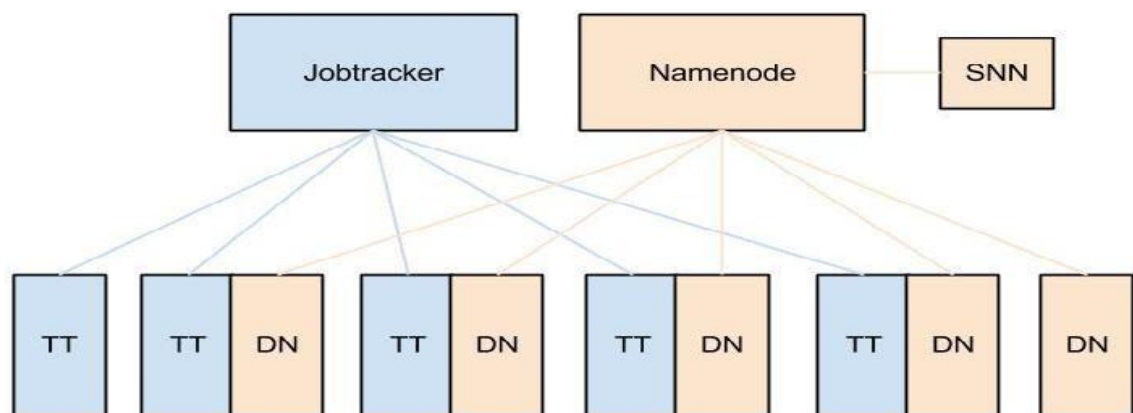
Automatic retries - if a task fails, it is retried N times (usually 3) on different task trackers.

Data locality optimizations - if you co-locate a TT with a HDFS Datanode (which you should) it will take advantage of data locality to make reading the data faster

Blacklisting a bad TT - if the JT detects that a TT has too many failed tasks, it will blacklist it. No tasks will then be scheduled on this task tracker.

Speculative Execution - the JT can schedule the same task to run on several machines at the same time, just in case some machines are slower than others. When one version finishes, the others are killed.

Here's a simple diagram of a typical deployment with TTs deployed alongside datanodes.



MapReduce Service Resources

For more reading on the Job Tracker and Task Tracker check out Wikipedia or the Hadoop book. I find the apache documentation pretty confusing when just trying to understand these things at a high level, so again doing a web-search can be pretty useful.

Wrap Up

I hope this introduction to Hadoop was useful. There is a lot of information on-line, but I didn't feel like anything described Hadoop at a high-level for beginners.

The Hadoop project is a good deal more complex and deeper than I have represented and is changing rapidly. For example, an initiative called MapReduce 2.0 provides a more general purpose job scheduling and resource management layer called YARN, and there is an ever growing range of non-MapReduce applications that run on top of HDFS, such as Cloudera Impala.

CONSTRAINTS:

A big data processing constraints on a low-power Hadoop cluster. Big Data processing with Hadoop has been emerging recently, both on the computing cloud and enterprise deployment. However, wide-spread security exploits may hurt the reputation of public clouds. If Hadoop on the cloud is not an option, an organization has to build its own Hadoop clusters. But having a data center is not worth for a small organization both in terms of building and operating costs. Another viable solution is to build a cluster with low-cost ARM system-on-chip boards.

BIGDATA APPLICATIONS:

Big Data is slowly becoming ubiquitous. Every arena of business, health or general living standards now can implement big data analytics. To put simply, Big Data is a field which can be used in any zone whatsoever given that this large quantity of data can be harnessed to one's advantage. The major scope of Big Data have been listed below.

1 THE THIRD EYE VISUALIZATION

Organizations worldwide are slowly and perpetually recognizing the importance of big data analytics. From predicting customer purchasing behavior patterns to influencing them to make purchases to detecting fraud and misuse which until very recently used to be an incomprehensible task for most companies big data analytics is a one-stop solution. Business experts should have the opportunity to question and interpret data according to their business requirements irrespective of the complexity and volume of the data. In order to achieve this requirement, data scientists need to efficiently visualize and present this data in a comprehensible manner. Giants like Google, Facebook, Twitter, EBay, Wal-Mart etc., adopted data visualization to ease complexity of handling data. Data visualization has shown immense positive outcomes in such business organizations. Implementing data analytics and data visualization, enterprises can finally begin to tap into the immense potential that Big data possesses and ensure greater return on investments and business stability.

2 INTEGRATION – AN EXIGENCY OF THE 21 ST CENTURY

Integrating digital capabilities in decision-making of an organization is transforming enterprises. By transforming the processes, such companies are developing agility, flexibility and precision that enables new growth. Gartner described the confluence of mobile devices, social networks, cloud services and big data analytics as the as nexus of forces. Using social and mobile technologies to alter the way people connect and interact with the organizations and incorporating big data analytics in this process is proving to be a boon for organizations implementing it. Using this concept, enterprises are finding ways to leverage the data better either to increase revenues or to cut costs even if most of it is still focused on customer-centric outcomes. Such customer-centric objectives may still be the primary concern of most companies, a gradual shift to integrating big data technologies into the background operations and internal processes

3 BIG DATA IN HEALTH CARE

Healthcare is one of those arenas in which Big Data ought to have the maximum social impact. Right from the diagnosis of potential health hazards in an individual to

complex medical research, big data is present in all aspects of it. Devices such as the Fitbit, Jawbone and the Samsung Gear Fit allow the user to track and upload data. Soon enough such data will be compiled and made available to doctors, which will aid them in the diagnosis.

CHAPTER 4

RESULT AND DISCUSSION

This system how to make agriculture well organized by predicting and thus improve the crop yields by using soil information. The paper introduces a new Agro algorithm which is used to predict the suitability of a crop for a particular soil type and enhances the overall quality of agricultural production. This also helps the farmers to select a particular crop to sow depending on the climatic condition and provides necessary information to choose the best weather to do quality farming. The paper uses big data using Hadoop platform which helps to deal with the large number of datasets in agricultural domain.

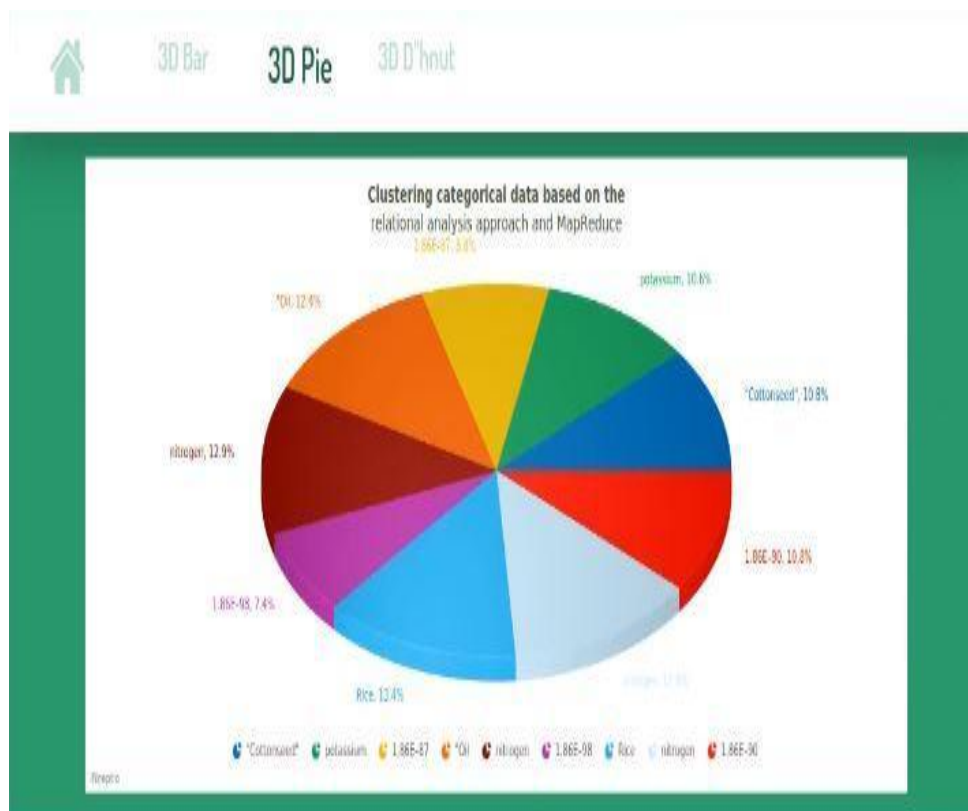


Fig 4.1 3D Pie Chart Prediction

The prediction soil and crop are predicted with the help of historical data. In above result the crop is predicted by the help of phosphorous, potassium, nitrogen and rust present in soil. Our prediction is for rice, sunflower and cotton seed. It helps to grow more production in agriculture. To predict these crops we used huge number of data collection which gives us the accurate prediction of the crop yield.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This framework the best approach to make farming efficient by anticipating and consequently improve the harvest yields by utilizing soil data. The paper presents a substitution Argo calculation which is utilized to foresee the appropriateness of a harvest for a particular soil type and upgrades the general nature of farming creation. This additionally helps the ranchers to pick a particular harvest to plant relying on the climatic condition and gives essential data to choose the least complex climate to attempt to quality cultivating. The paper utilizes huge information utilizing Hadoop stage which assists with influencing the monstrous measure of datasets in rural space.

The future work on this project can be adding more options such as Crop disease detection and prevention, crop price prediction. Current Market status and analysis for getting the information about market rates of crop, production cost, fertilizers. Government schemes option to know new government schemes related with loans, fertilizers, and crop.

REFERENCES

- [1] Suvidha Jambekar ; Shikha Nema ; Zia Saquib “Prediction of Crop Production in India Using Data Mining Techniques” 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
- [2] Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review , Computers and Electronics in Agriculture 151 (2018) 61–69,Elisver,2018.
- [3] Prakash Kumar, Anil Kumar, Sanjeev Panwar, Sukanta Dash,Kanchan Sinha,Vipin Kumar Chaudhary and Mrinmoy Ray “Role of big data in agriculture- A statistical prospective” Agric. Res. New Series Vol. 39 (2) : 210-215 (2018).
- [4] Shriya Sahu et al, " An Efficient Analysis of Crop Yield Prediction Using Hadoop framework based on random Forest approach, International Conference on Computing, Communication and Automation , 2017.
- [5] Ramya M, Chetan Balaji, Girish L ,” Environment Change Prediction to Adapt Climate-Smart Agriculture Using Big Data Analytics”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET).
- [6] Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh,” Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique”,Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.138-145
- [7] Veenadhari, S., Bharat Misra, D Singh, “Data mining Techniques for Predicting Crop Productivity – A review article”, IJCST, International Journal of Computer Science and technology march 2011.
- [8] M.S.PrasadBabu, N.V.Ramana Murty, S.V.N.L.Narayana, “A Web Based Tomato Crop Expert Information System Based on Artificial Intelligence and Machine Learning Algorithms”, IJCSIT, Vol. 1 (1), 2010, 6-15.
- [9] Leemans V, M F Destain, "A Real Time Grading Method of Apples Based on Features Extracted from Defects", J. Jood Eng., 2004, pages: 83-89.
- [10] R J Brooks, “Simplifying Sirius: Sensitivity Analysis and Development of a Meta-Model for Wheat Yield Prediction”, European Journal of Agronomy, vol. 14, 200

APPENDIX

A. SOURCE CODE:

```
import java.io.FileNotFoundException;
import java.util.Scanner;
import java.io.File;

//
// Decompiled by Procyon v0.5.36
//

public class Crop
{
    public static void main(final String[] args) {
        final String fileName = "/home/user/Desktop/Agricultural/Crop/data.csv";
        final File file = new File(fileName);
        try {
            final Scanner inputStream = new Scanner(file);
            double sum = 0.0;
            double numberOfMonths = 0.0;
            double sum2 = 1.0;
            double numberOfMonths2 = 1.0;
            double sum3 = 2.0;
            double numberOfMonths3 = 2.0;
            double sum4 = 4.0;
            double numberOfMonths4 = 4.0;
            double sum5 = 5.0;
            double numberOfMonths5 = 5.0;
            double sum6 = 6.0;
            double numberOfMonths6 = 6.0;
            double sum7 = 4.0;
            double numberOfMonths7 = 10.0;
```

```

double sum8 = 5.0;
double numberOfMonths8 = 9.0;
double sum9 = 6.0;
double numberOfMonths9 = 8.0;
while (inputStream.hasNext()) {
    final String data = inputStream.next();
    final String[] Values = data.split(",");
    final double avMonth = Double.parseDouble(Values[1]);
    sum += avMonth;
    ++numberOfMonths;
    sum2 += avMonth;
    ++numberOfMonths2;
    sum3 += avMonth;
    ++numberOfMonths3;
    sum4 += avMonth;
    ++numberOfMonths4;
    sum5 += avMonth;
    ++numberOfMonths5;
    sum6 += avMonth;
    ++numberOfMonths6;
    sum7 += avMonth;
    ++numberOfMonths7;
    sum8 += avMonth;
    ++numberOfMonths8;
    sum9 += avMonth;
    ++numberOfMonths9;
}
inputStream.close();
System.out.println("Rice Total nutrients Average ");
System.out.print("\n");
System.out.println("Black-Soil Total nutrients Average ");

```

```

        System.out.println("Adequate values of Nitrogen in Rice : 2.6 - 5.2");
        System.out.println("Nitrogen Total Average : " + sum / numberOfMonths);
        System.out.println("Adequate values of Phosphorus in Rice : 1,000-
1,800");
        System.out.println("Phosphorus Total Average : " + sum2 /
numberOfMonths2);
        System.out.println("Adequate values of Potassium in Rice : 1.4 - 2.8");
        System.out.println("Potassium Total Average : " + sum3 /
numberOfMonths3);
        System.out.print("\n");
        System.out.println("Rice Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Red-Soil Total nutrients Average ");
        System.out.println("Adequate values of Nitrogen in Rice : 2.6 - 5.2");
        System.out.println("Nitrogen Total Average : " + sum /
numberOfMonths3);
        System.out.println("Adequate values of Phosphorus in Rice : 1,000-
1,800");
        System.out.println("Phosphorus Total Average : " + sum2 /
numberOfMonths);
        System.out.println("Adequate values of Potassium in Rice : 1.4 - 2.8");
        System.out.println("Potassium Total Average : " + sum3 /
numberOfMonths2);
        System.out.print("\n");
        System.out.println("Rice Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Sandy-Soil Total nutrients Average ");
        System.out.println("Adequate values of Nitrogen in Rice : 2.6 - 5.2");
        System.out.println("Nitrogen Total Average : " + sum /
numberOfMonths2);

```

```

        System.out.println("Adequate values of Phosphorus in Rice : 1,000-
1,800");
        System.out.println("Phosphorus   Total   Average   :   "   +   sum2   /
numberofMonths3);
        System.out.println("Adequate values of Potassium in Rice : 1.4 - 2.8");
        System.out.println("Potassium   Total   Average   :   "   +   sum3   /
numberofMonths);
        System.out.print("\n");
        System.out.println("CottonSeeds Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Black-Soil Total nutrients Average ");
        System.out.println("Adequate values of Nitrogen in CottonSeed : 2.2 -
3.2");
        System.out.println("Nitrogen   Total   Average   :   "   +   sum4   /
numberofMonths4);
        System.out.println("Adequate values of Phosphorus in CottonSeed :
1,000 - 1,500");
        System.out.println("Phosphorus   Total   Average   "   +   sum5   /
numberofMonths5);
        System.out.println("Adequate values of Potassium in CottonSeed : 1.5 -
2.2");
        System.out.println("Potassium   Total   Average   :   "   +   sum6   /
numberofMonths6);
        System.out.print("\n");
        System.out.println("CottonSeeds Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Red-Soil Total nutrients Average ");
        System.out.println("Adequate values of Nitrogen in CottonSeed : 2.2 -
3.2");
        System.out.println("Nitrogen   Total   Average   :   "   +   sum4   /
numberofMonths6);

```

```

        System.out.println("Adequate values of Phosphorus in CottonSeed :
1,000 - 1,500");
        System.out.println("Phosphorus    Total    Average    " + sum5 /
numberOfMonths4);
        System.out.println("Adequate values of Potassium in CottonSeed : 1.5 -
2.2");
        System.out.println("Potassium    Total    Average    :    " + sum6 /
numberOfMonths5);
        System.out.print("\n");
        System.out.println("CottonSeeds Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Sandy-Soil Total nutrients Average ");
        System.out.println("Adequate values of Nitrogen in CottonSeed : 2.2 -
3.2");
        System.out.println("Nitrogen    Total    Average    :    " + sum4 /
numberOfMonths3);
        System.out.println("Adequate values of Phosphorus in CottonSeed :
1,000 - 1,500");
        System.out.println("Phosphorus    Total    Average    " + sum5 /
numberOfMonths2);
        System.out.println("Adequate values of Potassium in CottonSeed : 1.5 -
2.2");
        System.out.println("Potassium    Total    Average    :    " + sum6 /
numberOfMonths7);
        System.out.print("\n");
        System.out.println("Sunflower Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Black-Soil Total nutrients Average ");
        System.out.println("Adequate values of Nitrogen in Sunflower : 1.2 -
2.2");

```

```

        System.out.println("Nitrogen    Total    Average    :" + sum9 /
numberofMonths9);
        System.out.println("Adequate values of Phosphorus in Sunflower : 1.8 -
2.3");
        System.out.println("Phosphorus    Total    Average    : " + sum7 /
numberofMonths7);
        System.out.println("Adequate values of Potassium in Sunflower : 0.5 -
1.2");
        System.out.println("Potassium    Total    Average    : " + sum8 /
numberofMonths8);
        System.out.print("\n");
        System.out.println("Sunflower Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Red-Soil Total nutrients Average ");
        System.out.println("Adequate values of Nitrogen in Sunflower : 1.2 -
2.2");
        System.out.println("Nitrogen    Total    Average    :" + sum9 /
numberofMonths5);
        System.out.println("Adequate values of Phosphorus in Sunflower : 1.8 -
2.3");
        System.out.println("Phosphorus    Total    Average    : " + sum7 /
numberofMonths);
        System.out.println("Adequate values of Potassium in Sunflower : 0.5 -
1.2");
        System.out.println("Potassium    Total    Average    : " + sum8 /
numberofMonths3);
        System.out.print("\n");
        System.out.println("Sunflower Total nutrients Average ");
        System.out.print("\n");
        System.out.println("Sandy-Soil Total nutrients Average ");

```

```

        System.out.println("Adequate values of Nitrogen in Sunflower : 1.2 -
2.2");
        System.out.println("Nitrogen    Total    Average    :"    +    sum9    /
numberOfMonths6);
        System.out.println("Adequate values of Phosphorus in Sunflower : 1.8 -
2.3");
        System.out.println("Phosphorus    Total    Average    :    "    +    sum7    /
numberOfMonths4);
        System.out.println("Adequate values of Potassium in Sunflower : 0.5 -
1.2");
        System.out.println("Potassium    Total    Average    :    "    +    sum8    /
numberOfMonths3);
        System.out.print("\n");
        System.out.println("_____
_____");
        System.out.println("Prediction and enhancement of crop yield by big
data analysis");
        System.out.println("_____
_____");
        System.out.printf("%10s %20s %30s %20s %20s", "Crop", "Land",
"Nitrogen", "Phosphorus", "Potassium");
        System.out.println();
        System.out.println("_____
_____");
        System.out.format("%10s %20s %30s %20s %20s", "Rice", "Black-Soil",
"2.1", "2.081967213114754", "2.064516129032258");
        System.out.println();
        System.out.format("%10s %20s %30s %20s %20s", "", "Red-Soil",
"2.032258064516129", "2.1166666666666667", "2.098360655737705");
        System.out.println();

```

```

        System.out.format("%10s %20s %30s %20s %20s", "", "Sandy-Soil",
"2.0655737704918034", "2.0483870967741935", "2.1333333333333333");
        System.out.println();
        System.out.println("_____");
        _____);
        System.out.format("%10s %20s %30s %20s %20s", "CottonSeed",
"Black-Soil", "2.03125", "2.0153846153846153", "2.0");
        System.out.println();
        System.out.format("%10s %20s %30s %20s %20s", "", "Red-Soil",
"1.9696969696969697", "2.046875", "2.0307692307692307");
        System.out.println();
        System.out.format("%10s %20s %30s %20s %20s", "", "Sandy-Soil",
"2.096774193548387", "2.1475409836065573", "1.8857142857142857");
        System.out.println();
        System.out.println();
        System.out.println("_____");
        _____);
        System.out.format("%10s %20s %30s %20s %20s", "Sunflower",
"Black-Soil", "1.9411764705882353", "1.8571428571428572",
"1.8985507246376812");
        System.out.println();
        System.out.format("%10s %20s %30s %20s %20s", "", "Red-Soil",
"2.0307692307692307", "2.1666666666666665", "2.1129032258064515");
        System.out.println();
        System.out.format("%10s %20s %30s %20s %20s", "", "Sandy-Soil",
"2.0", "2.03125", "2.1129032258064515");
        System.out.println();
        System.out.println("_____");
        _____);
    }
    catch (FileNotFoundException e) {

```

```
        e.printStackTrace();
    }
}
}
```

B. SCREEN SHOTS:

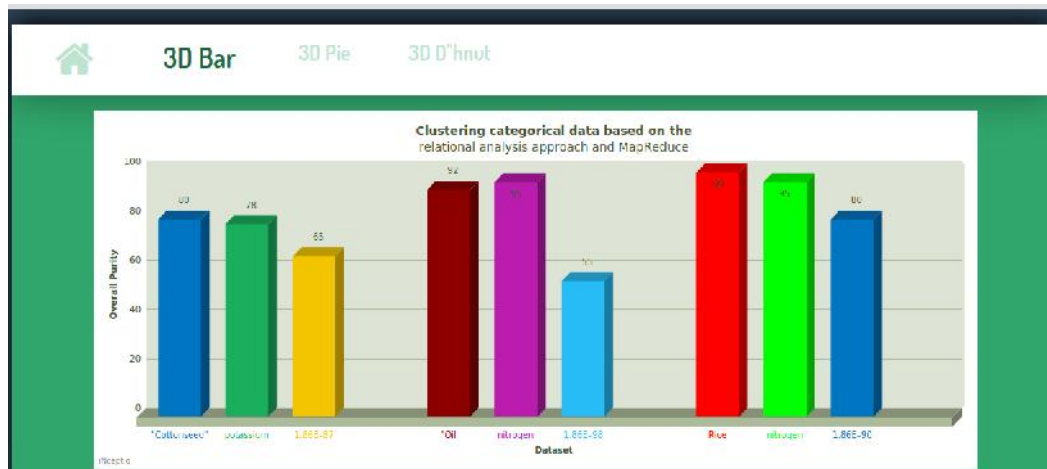


Fig 6.1: Representation of categorical data

```

-----
Prediction and enhancement of crop yield by big data analysis
-----

```

	Crop	Potassium	Land	Nitrogen	Phospho
754	Rice	2.064516129032258	Black-Soil	2.1	2.081967213114
667		2.098360655737705	Red-Soil	2.032258064516129	2.116666666666666
935		2.133333333333333	Sandy-Soil	2.0655737704918034	2.0483870967741
153	CottonSeed	2.0	Black-Soil	2.03125	2.0153846153846
875		2.0307692307692307	Red-Soil	1.9696969696969697	2.046
573		1.8857142857142857	Sandy-Soil	2.096774193548387	2.1475409836065
572	Sunflower	1.8985507246376812	Black-Soil	1.9411764705882353	1.8571428571428
665		2.1129032258064515	Red-Soil	2.0307692307692307	2.166666666666666
125		2.1129032258064515	Sandy-Soil	2.0	2.03

```

-----
user@node:~/Desktop/Agricultural/Crop$

```

Fig 6.2: Prediction and enhancement of crop yield by big data analysis.

```
ser@node: ~/Desktop/Agricultural/Project
FILE: Number of bytes written=3069163
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=4136153
HDFS: Number of bytes written=173
HDFS: Number of read operations=9
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=35714
  Total time spent by all reduces in occupied slots (ms)=12650
  Total time spent by all map tasks (ms)=35714
  Total time spent by all reduce tasks (ms)=12650
  Total vcore-seconds taken by all map tasks=35714
  Total vcore-seconds taken by all reduce tasks=12650
  Total megabyte-seconds taken by all map tasks=36571136
  Total megabyte-seconds taken by all reduce tasks=12953600
Map-Reduce Framework
  Map input records=88627
  Map output records=88627
  Map output bytes=1198217
  Map output materialized bytes=1375483
  Input split bytes=174
  Combine input records=0
  Combine output records=0
  Reduce input groups=10
  Reduce shuffle bytes=1375483
  Reduce input records=88627
  Reduce output records=10
  Spilled Records=177254
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=593
  CPU time spent (ms)=8580
  Physical memory (bytes) snapshot=705421312
  Virtual memory (bytes) snapshot=2510303232
  Total committed heap usage (bytes)=494927872
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=4135979
File Output Format Counters
```

Fig 6.3: Map reduce framework

C. PAPER PUBLICATION

IJSRD - International Journal for Scientific Research & Development | Vol. 8, Issue 1, 2021 | ISSN (online): 2321-0613

Big Data in Smart Farming

¹Bhawana Kumari, ²Medulla Samhitha, ³Mrs. Selvi

^{1,2}UG Student

^{1,2}Department of Computer Science and Engineering

^{1,2}Sathyabama Institute of Science and Technology, Chennai, India

Abstract: Farming is the culture, science and art of cultivating the soil, growing crops and rising livestock. It plays a key role in increasing India's Economy. Big data analysis is being used for discovering solutions which acts as the means for analyzing the dataset that plays a significant role in making a right decision in that specific field such as agriculture. Harvesting yields relies upon different components like soil, climate and pesticides. This research paper facilitates the farmers in selecting the right crops and appropriate cropping pattern for a particular locality. This helps farmers in selecting the most suitable crop. The prediction can be done using three algorithms like K-means, Random Forest and Apriori Algorithms.

Keywords: Agriculture, K-means, Random Forest, Prediction, Apriori, etc.

I. INTRODUCTION

India is one of the most populated countries in the world. It has many random changes in its weather and it is necessary to have a secured food resources around the world. Our country has to face different types of serious problems during drought, floods and uneven rainfall conditions. Different categories of soils imports a major role in the crop yield. Crop production may be a complicated development that's influenced by soil and environmental conditions with a set of input/output parameters. Agronomical input/output parameters varies from pasture to pasture and planter to planter. Collection of such info on a bigger scale may be a discouraging task. However, the environmental condition info collected in Republic of India at each 2sq.m space in numerous components of the neighborhood is monitored by Indian meteoric Department. This excessive variations of expertise sets maybe used for the projection of its impact on the most needed saplings of that individual neighborhood, area, place, etc. Some are completely discrete foretelling methodologies erupted and examined by the researchers out in the globe within the field of agribusiness or similar agro sciences. A number of such studies are: Agronomics research panels in alternative countries have observed that efforts in crop yield

policies have resulted in hazardously increase in high synthetically outbreak. These studies have outlined many association between biochemical usage and plantation yields. Agriculture is associate trade sector that's benefiting powerfully from the event of detector technology, knowledge science, and machine learning (ML) techniques within the latest years.

Normally, ML algorithms can result in the most stable and substantial way for the précised output of the yield. Beforehand, these types of yield were projected on the opinions and experience of the farmer.

But nowadays weather forecasts changes unexpectedly which makes it difficult to make a guess at the rates of yields. This is the period when technology comes into the picture for helping the farmer's to make a better decision during the cultivation mode. ML model are quite efficient in capturing the patterns of the crop and yield and bid data is quite useful in storing the major datasets of agriculture for later analysis. The below figure Fig.1 shows the architecture of the proposed system.

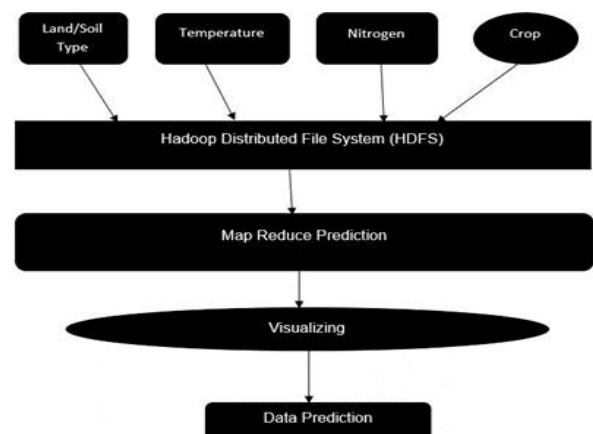


Fig.1

I. RELATED WORK

improvisation and growth through pro-pesticide district

[1] Data mining is the investigation of extricating valuable data from the informational collections. At that point

present examination centers around the use of Data mining procedures to foresee future creation of yields, for example, Rice, Wheat and Maize as for different parameters saw during the period (1950-2013). The parameters considered for the examination were precipitation, mean temperature, territory under water system, region, creation and yield. The relapse calculations utilized in the examination were Various Straight Relapse, Arbitrary Timberland Relapse and Multivariate Versatile Relapse Splines (Earth).

[2] The paper examines and explore progress took place inside the timespan of around 16 years on Artificial Intelligence based arrangements for précised harvesting yields announcements and nitrogen status approximations. The paper also states that the speedy modifications in detection of innovations and ML procedures will give practical and complete answers for better harvest and condition state estimation and dynamic. Quite determined use of the sensor segments and Machine Learning methods, the formation of variety of sensor devices and master details, and the improvisation of interrelated framework joining distinctive.

[3] With this paper, an attempt is tried to make for focusing on the use of knowledge mining techniques in the field of agriculture. Calculation technique J48from C4.5 is used. Currently, information mining strategies are being developed which utilize past data such as a specific soil type, soil pH, ESP, EC.

[4] Locale to give better harvest yield estimation for that district. This model can be utilized to choose the most magnificent harvests for the locale and furthermore its yield there by improving the qualities and increase of cultivating too. This guides ranchers to settle on the yield they might want to plant for the inevitable year. Forecast will help the related enterprises for arranging the coordination of their business.

[5] This paper gives a reason about how to discover encounters from exactness horticulture data through huge information approach. Right now, the important information in a successful way drives a structure turning to big computing difficulties in crop examination where data is remotely accumulated. For the capacity motivation behind colossal information accessibility in agribusiness, we are planning Hadoop structure in store for our jobs a gigantic volume of yield information. That job gives you a superior forecast for the ranchers to plant which sort of harvests to their homestead field dependent on their dirt substance to improve the efficiency. The arbitrary woods calculation is incorporated with the MapReduce programming model in Hadoop structure.

Information are assuming a significant job making great arranging and approaches for agrarian development and advancement.

Add to that land debasement expanding peripheral land and loss of biodiversity are better arrangements with investigation of huge information science. Harvest information can be separate into bits and bytes it will give better investigation about the yield improvement by utilizing advance information examination instruments for improvement of farming. Here, talk about some significant apparatuses and strategies to deal with and study the enormous information.

II. METHODOLOGY

A. Overview

For our proposed system, we are aiming for predicting the correct crop choice and to increase the yield of the production with the power of big data analysis. To predict according to the current scenario, we'll collect data from the past growth and fails of production. This process is designed to provide a better accessibility to our farmers to decide upon the crop choice and soil type. The output will be in a simple and user-friendly format which will allow the user analyze correctly.

For this proposed system, there are two routes like analyzing and then classifying the data. First route is to filter the raw data by the filtering process and then to convert it into a form of processed data with the help of K-means clustering. And later for classifying these clusters by using some classification algorithm like random tree and apriori algorithm. The below figure Fig.2 shows working of the three algorithms used.

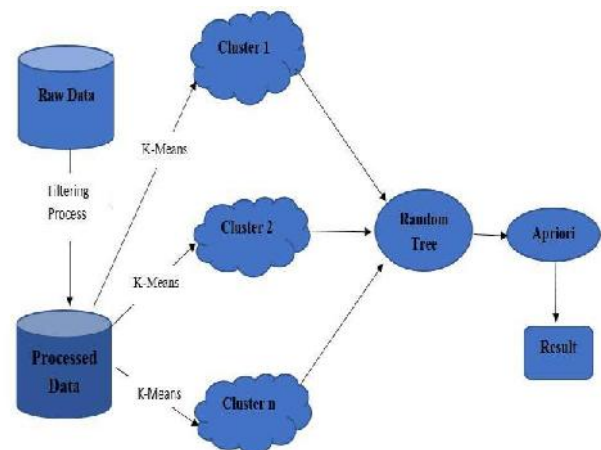


Fig. 2

B. Data Collection

The raw data is collected from data.gov and Kaggle.com. These raw data are gathered from online sources namely weather, soil type, land data, temperature datasets and different types of crop data. The data is stored in a format of .csv file in excel sheet.

[illegible]

C. Data Preprocessing

D. K-means Clustering

Input: $D = \{h_1, h_2, h_3, \dots, h_n\}$ /; Set of segments h ; No. of preferred cluster.

Algorithm: allot initial values for means $\mu_1, \mu_2 \dots \mu_k$

Calculate new mean for each cluster; until convergence criteria is met.

```

graph TD
    Start([Start]) --> K[/Number of cluster k/]
    K --> Centroid[Centroid]
    Centroid --> Distance[Distance object to centroid]
    Distance --> Grouping[Grouping of minimum distance]
    Grouping --> Decision{No object move group}
    Decision --> End([End])
    Decision -.-> Distance
  
```

Fig.4

The random tree is helpful in making our prediction more accurate as it classifies and completes the regression work. It is used to provide more accuracy for our project. The data that is clustered using K- means clustering allows to remove any set of data that is missing or unwanted, then it will handle these types of values. The datasets are in large amount so the random tree makes more decisions and try to merge them for obtaining a next level accurate prediction.

Require: Input: Q dataset – features with a target class
for $\forall$ features do

Execute the Decision Tree algorithm end for

Obtain the total number of leaf nodes $t_1, t_2, t_3, \dots, t_n$ with its constraints (t_0)

Output: Partition datasets $p_1, p_2, p_3, \dots, p_i$

F. Apriori Algorithm:

The apriori algorithm is used to make the project more reliable and efficient by the classification algorithm which is used for data mining different item sets. It is associated with the frequent datasets and are considered for the least minimum support of a particular item. It also is providing an additional benefit of maintaining quite good accuracy for a ginormous amount of datasets. The below figure Fig.5 shows working of the apriori algorithm used.

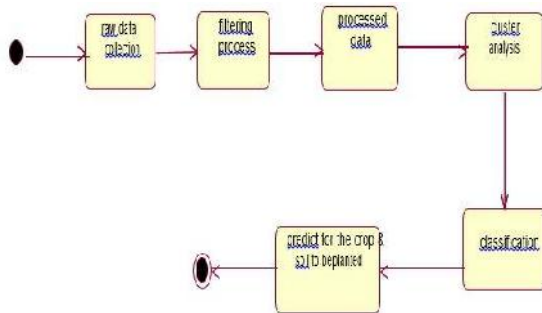


Fig. 5

G. Hadoop Framework:

The data is uploaded into the HDFS and then processed for the dataset by the Mapreducing technique model. It is capable of handling all types of structured and unstructured data. The mapping part is done by the mapper class and the reducing part is completed by the reducer class. We are using a ginormous amount of dataset so it helps in processing this type of huge amount of data in our code and later we can implement in our driver. These data are generally sorted by the mapper class and used as input data for the reducer class. The duplicate data is removed by reducer from the key value pair.

The essential qualities of the MapReduce is to blend the information <key,value> pair with another rundown of <key, value> pair.

The information is handled by mapping capacity to create the halfway arrangement of <key,value> pair.

The reducer merger the yield made from mapper to outline a smaller game plan of characteristics.

The figure Fig.6 mainly contains two tasks first is mapping and second is reducing first it takes a set of data and it changes to another set of data and then this set of reduce in the key value pair. The file input counter reads the bytes is 4135979 after map and reducing file output format counters read the bytes is 173.

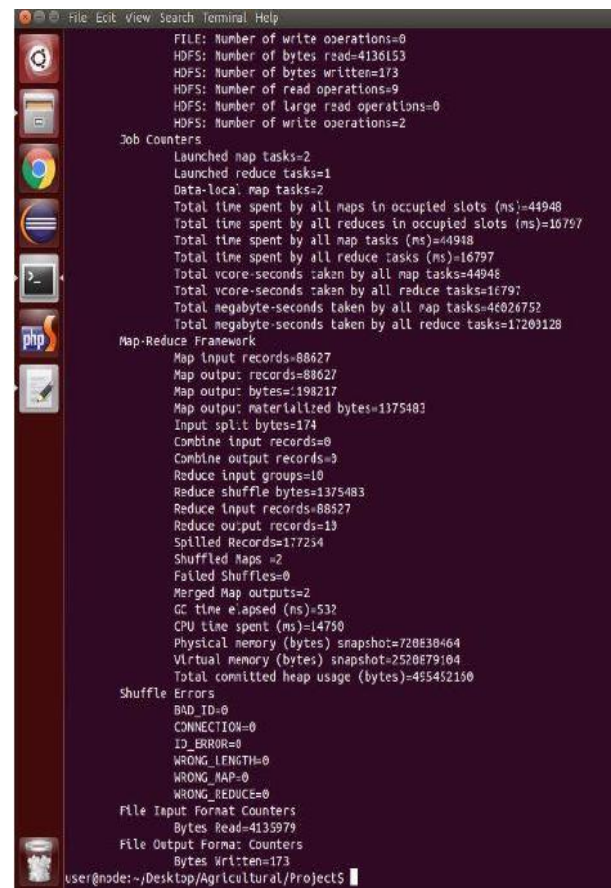


Fig. 6

III. RESULTS AND DISCUSSION

This proposed system shows how to create a horticulture efficient by just anticipating it and consequently improving the harvest yields by utilizing soil data types. The paper also tells us about how the yield can be more productive for the crops based on the historical/past datasets of different soil types. It is also helpful for the farmers to make a smart decision on cultivating which crop is suitable in accordance to the climatic conditions and gives the data to cultivate best crop. With reference to climatic conditions it will also consider the content of nitrogen, phosphorus and carbon dioxide. The paper facilitates quite enormous information based on the utilizing of Hadoop stage in accordance with managing the huge measure of datasets in rural regions.

In figure Fig.7 the prediction of yield the crop is shown using big data analysis. According to the given crop it predicts the soil and nutrition.

In figure Fig.8 the x axis represents datasets and y axis represents overall purity. Where it tells about the crop, nutrition and temperature of the soil.

ACKNOWLEDGMENT

Sathyabama institute of science and technology

REFERENCES

- [1] Suvidha Jambekar; Shikha Nema; Zia Saquib "Prediction of Crop Production in India Using Data Mining Techniques" 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
- [2] Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Computers and Electronics in Agriculture 151 (2018) 61–69, Elsevier, 2018.
- [3] Prakash Kumar, Anil Kumar, Sanjeev Panwar, Sukanta Dash, Kanchan Sinha, Vipin Kumar Chaudhary and Mrinmoy Ray "Role of big data in agriculture- A statistical prospective" Agric. Res. New Series Vol. 39 (2): 210-215 (2018).
- [4] Pooja M C et al, —Implementation of Crop Yield Forecasting Using Data Mining, International Research Journal of Engineering and Technology (IRJET), 2018, p-ISSN: 2395-0072
- [5] Shriya Sahu et al, "An Efficient Analysis of Crop Yield Prediction Using Hadoop framework based on random Forest approach, International Conference on Computing, Communication and Automation, 2017.
- [6] Bask J., Sudarshan, A., Trivedi D., M.S. Santhanam. 2004. "Weather Data Mining Using Independent Component Analysis". J. of Machine Learning Research 5: pp. 239- 253
- [7] P. Surya, Dr. I. Laurence and M. Ashok Kumar, "The role of big data analytics in agriculture sector: A survey", March 2016
- [8] Ms. Kanaan Devi, "Enhanced Crop Yield Prediction and Soil Data Analysis Using Data Mining", International Journal of Modern Computer Science, agriculture: A review, Computers and Electronics in Agriculture 151 (2018) 61–69, Elsevier, 2018.
- [9] Prakash Kumar, Anil Kumar, Sanjeev Panwar, Sukanta Dash, Kanchan Sinha, Vipin Kumar Chaudhary and Mrinmoy Ray "Role of big data in agriculture- A statistical prospective" Agric. Res. New Series Vol. 39 (2): 210-215 (2018).

Crop	Land	Nitrogen	Phosphorus	Potassium
Rice	Black-Soil	2.1	2.081267213114754	2.064516122032258
	Red-Soil	2.032258064516129	2.116656666666667	2.098330655737705
	Sandy-Soil	2.0655737704910334	2.3483378967741935	2.133333333333333
CottonSeed	Black-Soil	2.03125	2.3153346153346153	2.0
	Red-Soil	1.96969969969997	2.046875	2.0107692307692307
	Sandy-Soil	2.096774191548387	2.1475109836063573	1.8357112057142357
Sunflower	Black-Soil	1.9411794701882353	1.3571428571428572	1.89859724376312
	Red-Soil	2.0107692307692307	2.166666666666667	2.1129032253054515
	Sandy-Soil	2.0	2.03125	2.1129032253054515

Fig. 7

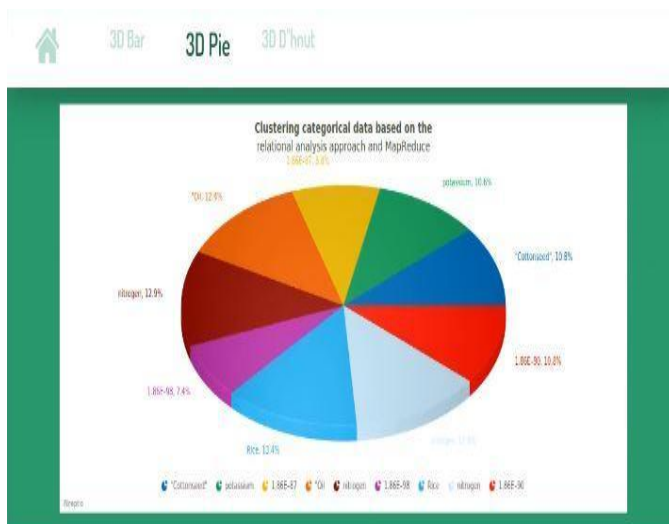


Fig. 8

IV. CONCLUSION

This framework the best approach to make farming efficient by anticipating and consequently improve the harvest yields by utilizing soil data. The paper presents a substitution Argo calculation which is utilized to foresee the appropriateness of a harvest for a particular soil type and upgrades the general nature of farming creation. This additionally helps the ranchers to pick a particular harvest to plant relying on the climatic condition and gives essential data to choose the least complex climate to attempt to quality cultivating. The paper utilizes huge information utilizing Hadoop stage which assists with influencing the monstrous measure of datasets in rural space.

- [10] Pooja M C et al, —Implementation of Crop Yield Forecasting Using Data Mining, International Research Journal of Engineering and Technology (IRJET), 2018, p-ISSN: 2395-0072
- [11] Shriya Sahu et al," An Efficient Analysis of Crop Yield Prediction Using Hadoop framework based on random Forest approach, International Conference on Computing, Communication and Automation, 2017.
- [12] Bask J., Sudarshan, A., Trivedi D., M.S. Santhanam. 2004. "Weather Data Mining Using Independent Component Analysis". J. of Machine Learning Research 5: pp. 239- 253
- [13] P. Surya, Dr. I. Laurence and M. Ashok Kumar, "The role of big data analytics in agriculture sector: A survey", March 2016
- [14] Ms. Kanaan Devi, "Enhanced Crop Yield Prediction and Soil Data Analysis Using Data Mining", International Journal of Modern Computer Science, Volume 4, Issue 6, December 2016
- [15] Anon (2010). Agricultural Statistics at a Glance 2010, Ministry of Agriculture Govt. of India
- [16] Soil, Big Data and the Future of Agriculture Conference, Canberra, 25 June, 2015.
- [17] Venkata eddy Konasani, Mukul Biswas and Praveen Krishnan Koleth, "Fraud Management using Big Data Analytics", A Whitepaper by Trend wise Analytics.
- [18] Radha Krishna Murthy "Crop Growth Modelling and its Applications in Ag in Agricultural Methodology"
- [19] J. O. Chan, "An Architecture for Big Data Analytics", Communications of the IIMA, vol. 13, no. 2, (2013), pp. 1-14
- [20] J. M. Chambers, "Software for Data Analysis: Programming with R (Statistics and Computing)", Springer, (2012).
- AP AgTech Summit 2017: Progressive Farmer, Smart Farming, 15-17 Nov, 2017,Vizag, www.apagtechsummit2017.in/

