Crop Data Analysis of Indian Region using Big Data Techniques for aiding Indian farmers

By

Ashutosh Anand,

Vineet Rathore and

Aman Sharma

Abstract:

Agriculture or farming is the largest contributor to the GDP of our country. Approximately 70 percent of rural household's primary source of income is agriculture. But still farmers don't get results worth their efforts. As a matter of fact, the condition of farmers in our country is horrible. The agriculture supports nearly 50 percent of the employment but contributes only 15 percent to the Gross Domestic Product (GDP). Due to variations in climatic conditions, improper crop selection or irrigation often leads to less yield than expected. On the other note, in the field of Information Technology, Big Data has come forth as a blazing topic. So by using Big Data approach in analysing the crop data over various factors such as soil type, temperature, water level, humidity, soil pH, fertilizers, we will devise an algorithm that will predict which crop should be sown in which season that both the yield and the profit can be maximized. For fulfilling the above task, we have used Hadoop framework, Machine Learning and NoSQL Database.

1. Introduction

The Big Data Analysis is process of inspecting, cleansing and modelling the data with the purpose of discovering useful information and conclusions. It is a process of analysing, extracting and predicting the meaningful information from huge data in order to gain some pattern. This process is used by companies to turn the raw data of their customer to useful information. This analysis can also be used in the field of Agriculture. Most farmers generally rely on their previous experiences on a particular crop to expect a higher yield of that crop in the next harvesting season. But still the they don't get result/income worth their efforts. The above listed problems are the recurring problems that all the farmers have to face. But we also know that during modern times, our environment is affected by global warming, pollution, acid rain, bio-waste and e-waste, so conditions keep changing drastically, and therefore traditional farming techniques and only past experiences are not sufficient. Hence, there is a need new and more reliable farming techniques that can meet the requirements of the country. As a result, Government is also taking many initiatives to improve farming. One of them is it is making data of previous farming records. But the problem with it is that the record or the dataset has become very large and can not be handled or analysed accurately with old methods. Furthermore doing so is also a very time consuming process. That's where Big Data comes into play. By using Big Data Analysis, we can analyse large amount of data and find the underlying pattern within it precisely and efficiently.

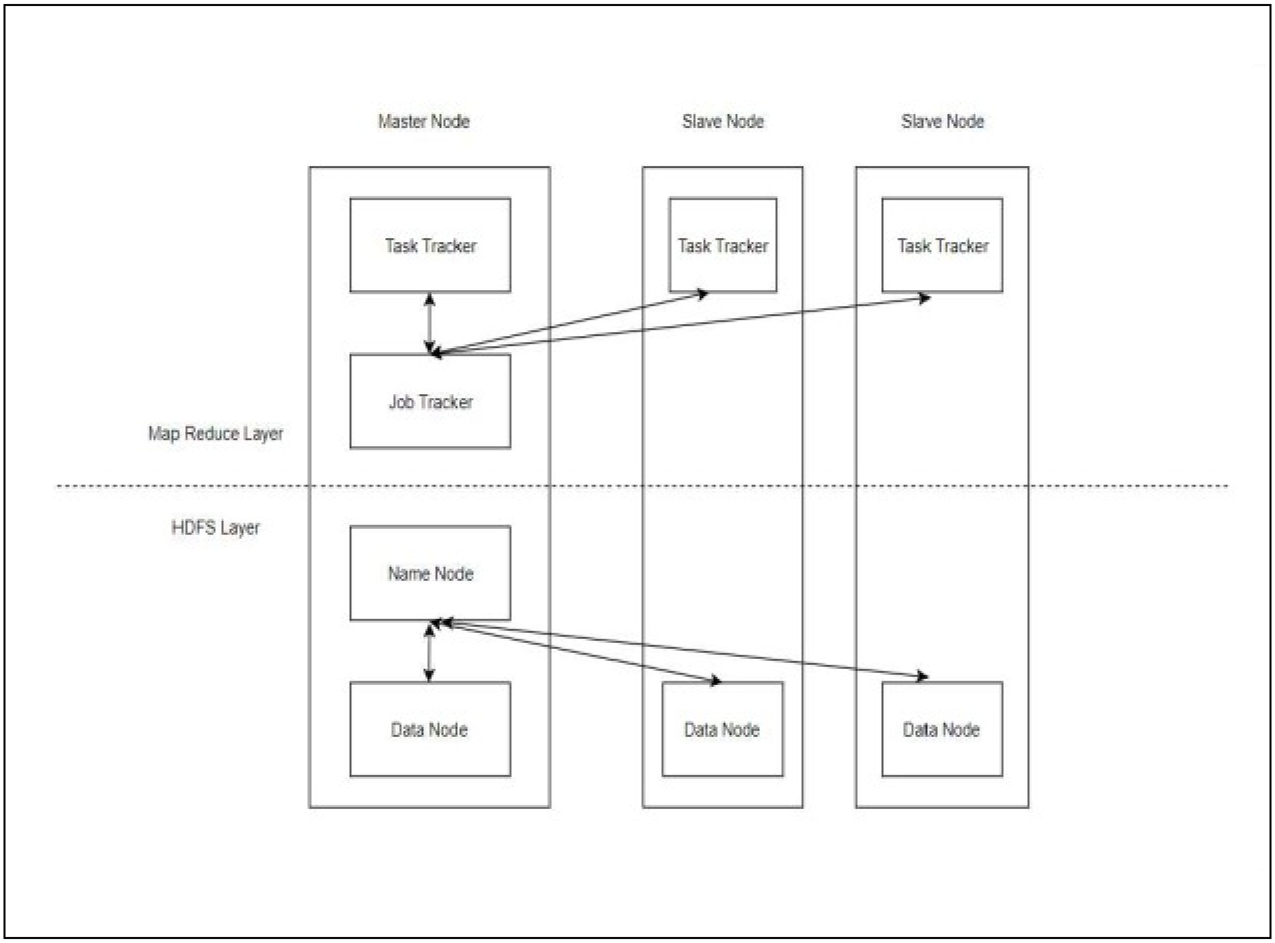


Figure:1.1 Hadoop Architecture

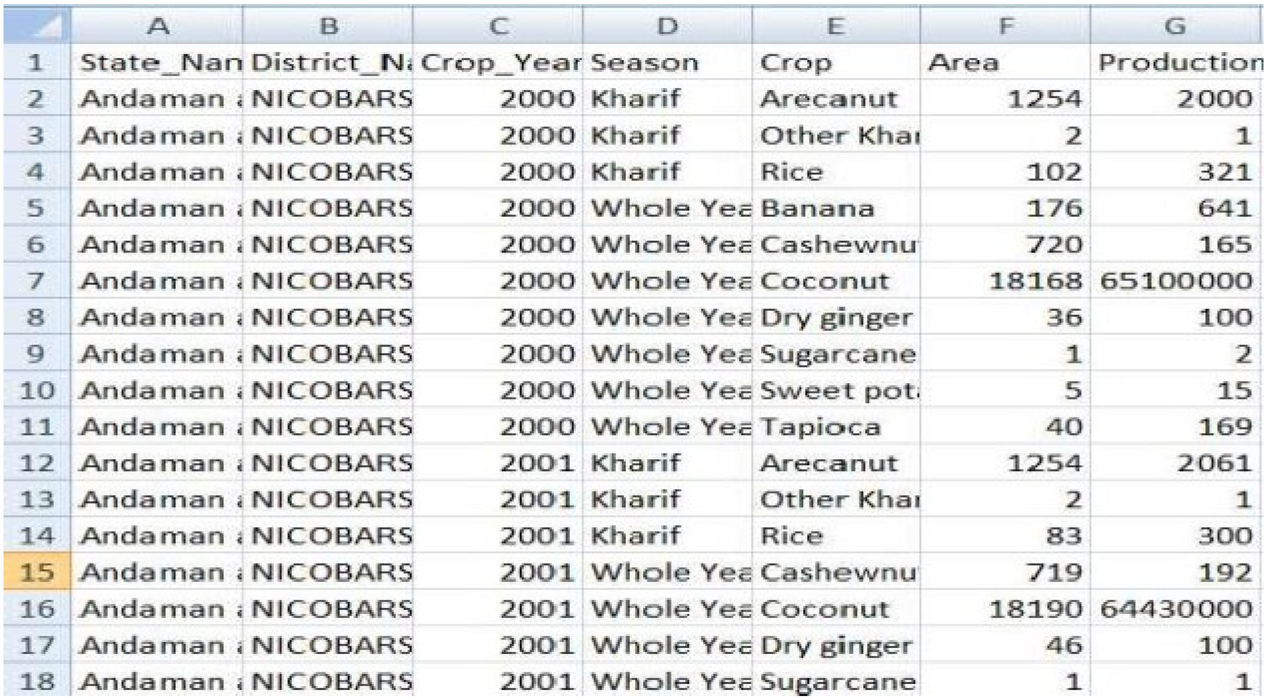
2. Literature Review

In today's world, more and more people are using electronic devices, surfing the web and using the social medias, thereby generating vast amount of data which mainly compromises of semi-structured and unstructured data. But the problem is that the data is highly unstructured and is present in large amount, so analyzing them by traditional means is a tedious and time consuming task without high chances of errors in outcomes. Fortunately, this problem in modern world can be solved to a great extent by using Big Data. Big Data is more real in comparison to another techniques for processing huge data. Massively data processing, scaling out architectures are unit compatible for big data applications.

Govt. of India created an open data ecosystem for the motive of sharing crop dataset as per National Data Sharing and Accessibility Policy (NDSAP) initiated Open Government Data (OGD).

Indian Prime Minister Narendra Modi -led government has asked Indian Council of Agriculture Research (ICAR) to prepare a chairmanship of Praveen Rao, comprehensive crop plan for India. A committee under the vice chancellor, Prof Jayashankar Telangana State Agriculture University, has been constituted for this purpose, Trilochan Mohapatra, Director General,

ICAR said, “The committee is expected to come out with a set of crop recommendations for each geography after considering the climatic conditions, soil health, water stress and the estimated short-term as well as long term demand for the produce within the country and globally”. The need to reorient India's agriculture practices and systems was one of the key recommendations of the high-level committee that was tasked with the job of finding ways to double farmers' income. The technical, scientific and economical feasibility of crops will be looked at a granular level due to the agro-climatic diversity of the country, Mohapatra said. According to him, water, the most critical element for any farming exercise, is fast becoming a scarce resource, and hence calls for urgent interventions at every level of agriculture. We have taken the dataset from the government website data.gov.in/crop. Our dataset consists of the data of all the states which consist of their districts and the production of crops with their respective seasons. The dataset further consists of area of the districts.

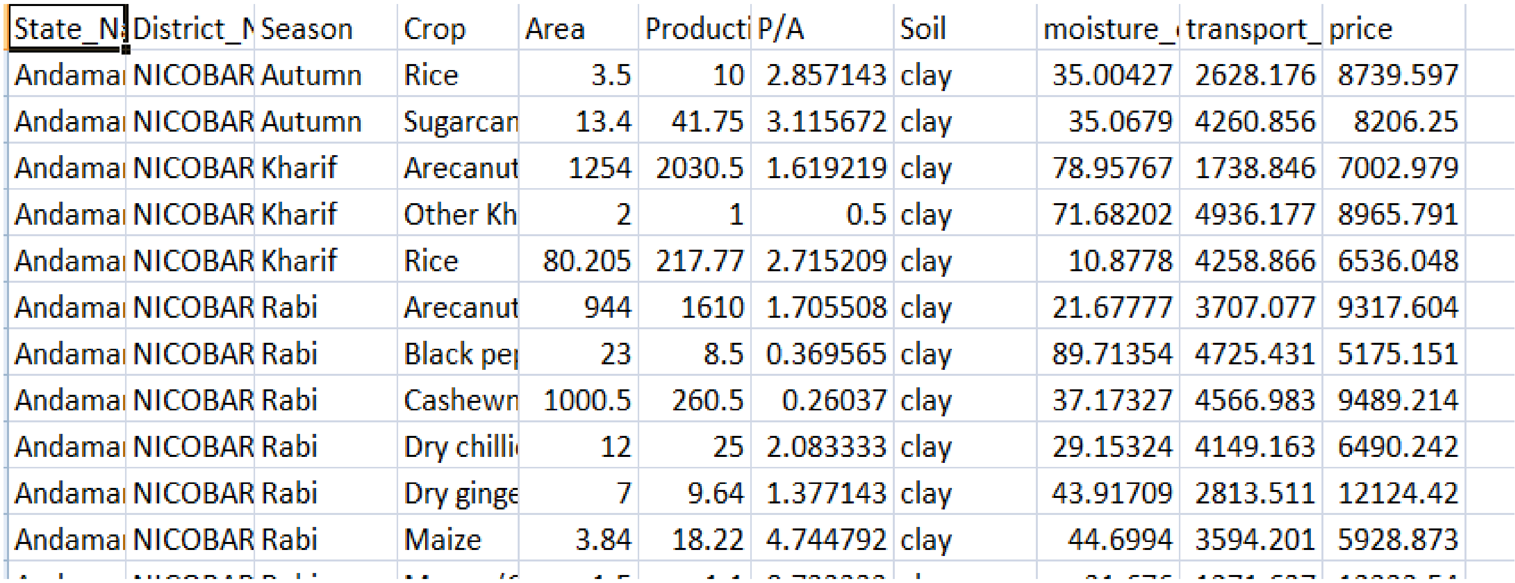


	A	B	C	D	E	F	G
1	State_Nan	District_N	Crop_Year	Season	Crop	Area	Production
2	Andaman a	NICOBARS	2000	Kharif	Arecanut	1254	2000
3	Andaman a	NICOBARS	2000	Kharif	Other Khar	2	1
4	Andaman a	NICOBARS	2000	Kharif	Rice	102	321
5	Andaman a	NICOBARS	2000	Whole Yea	Banana	176	641
6	Andaman a	NICOBARS	2000	Whole Yea	Cashewnu	720	165
7	Andaman a	NICOBARS	2000	Whole Yea	Coconut	18168	65100000
8	Andaman a	NICOBARS	2000	Whole Yea	Dry ginger	36	100
9	Andaman a	NICOBARS	2000	Whole Yea	Sugarcane	1	2
10	Andaman a	NICOBARS	2000	Whole Yea	Sweet pot	5	15
11	Andaman a	NICOBARS	2000	Whole Yea	Tapioca	40	169
12	Andaman a	NICOBARS	2001	Kharif	Arecanut	1254	2061
13	Andaman a	NICOBARS	2001	Kharif	Other Khar	2	1
14	Andaman a	NICOBARS	2001	Kharif	Rice	83	300
15	Andaman a	NICOBARS	2001	Whole Yea	Cashewnu	719	192
16	Andaman a	NICOBARS	2001	Whole Yea	Coconut	18190	64430000
17	Andaman a	NICOBARS	2001	Whole Yea	Dry ginger	46	100
18	Andaman a	NICOBARS	2001	Whole Yea	Sugarcane	1	1

Figure: 2.1 Initial Dataset

One of the most important aspect of analyzing a data requires cleaning of dataset. We removed ”crop year” column because of some redundancy. We used “pandas” which is a famous library of python that takes data (like a CSV or TSV file, or a SQL database) and creates a python object with rows and columns called data frame that looks very similar to table in a statistical software. It is mainly used for data manipulation and data analysis.

We used group by function of pandas to group our states with their respective districts and seasons. We used aggregate function to retrieve the average of the area and the production of the crop in a particular season of a district. Further we added a new column named as "P_A" which shows about production per area of a particular of a respective season with districts.



State_N	District_N	Season	Crop	Area	Producti	P/A	Soil	moisture_	transport_	price
Andamai	NICOBAR	Autumn	Rice	3.5	10	2.857143	clay	35.00427	2628.176	8739.597
Andamai	NICOBAR	Autumn	Sugarcan	13.4	41.75	3.115672	clay	35.0679	4260.856	8206.25
Andamai	NICOBAR	Kharif	Arecanut	1254	2030.5	1.619219	clay	78.95767	1738.846	7002.979
Andamai	NICOBAR	Kharif	Other Kh	2	1	0.5	clay	71.68202	4936.177	8965.791
Andamai	NICOBAR	Kharif	Rice	80.205	217.77	2.715209	clay	10.8778	4258.866	6536.048
Andamai	NICOBAR	Rabi	Arecanut	944	1610	1.705508	clay	21.67777	3707.077	9317.604
Andamai	NICOBAR	Rabi	Black pe	23	8.5	0.369565	clay	89.71354	4725.431	5175.151
Andamai	NICOBAR	Rabi	Cashewn	1000.5	260.5	0.26037	clay	37.17327	4566.983	9489.214
Andamai	NICOBAR	Rabi	Dry chilli	12	25	2.083333	clay	29.15324	4149.163	6490.242
Andamai	NICOBAR	Rabi	Dry ginge	7	9.64	1.377143	clay	43.91709	2813.511	12124.42
Andamai	NICOBAR	Rabi	Maize	3.84	18.22	4.744792	clay	44.6994	3594.201	5928.873

Figure 2.2 Updated Dataset

Data visualization is very important to represent the features of data in graphical form to understand complicated relationship in data. Standardizing the data is essential need before
visualization. We used MinMaxScaler of sci-kit learn library to the dataset. We
standardize plotted bar graphs using seaborn library.

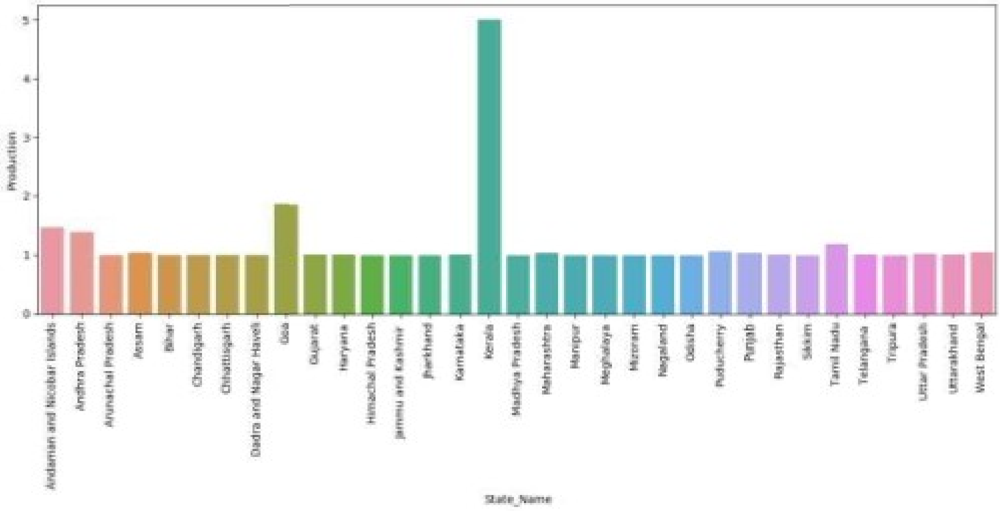


Figure 2.3 Graph of State name v/s production

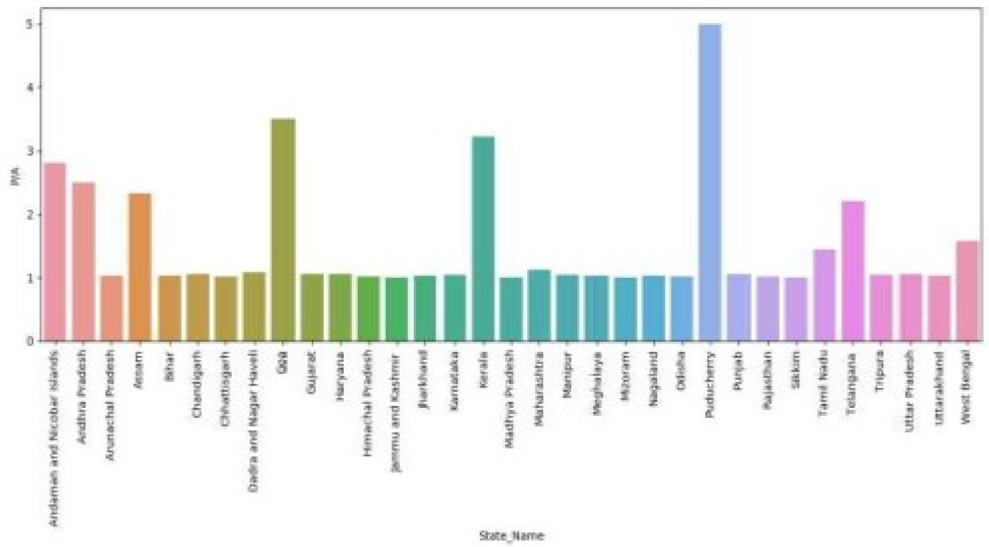


Figure 2.4 Graph of state name v/s “P_A”

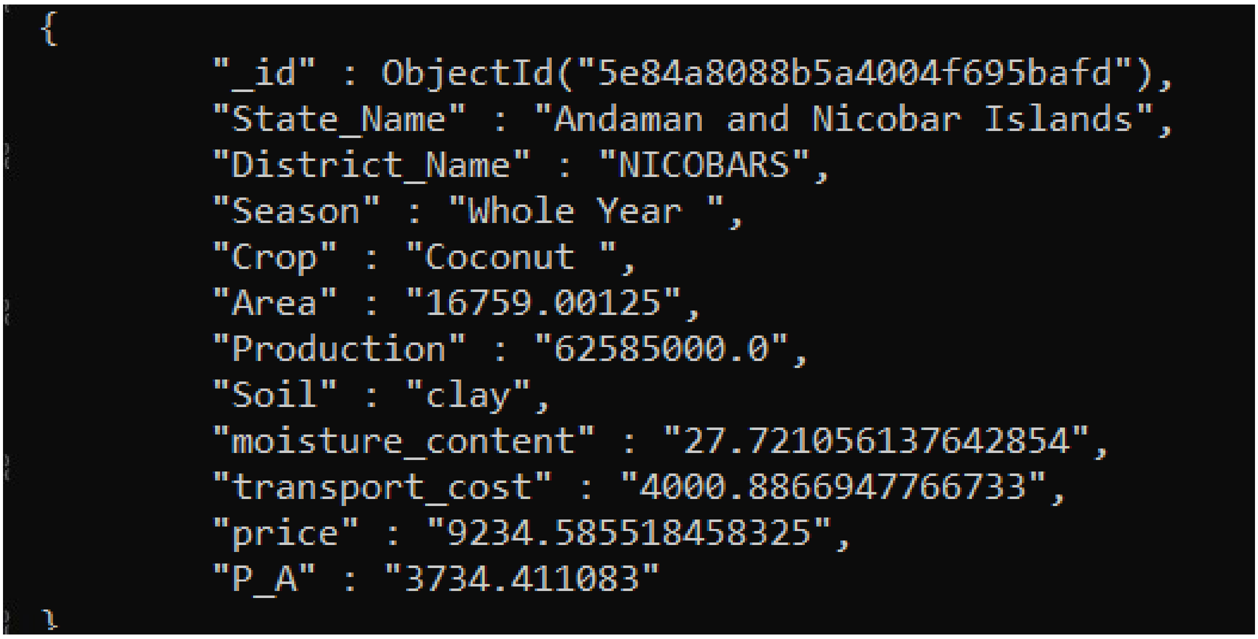
The trend of social media is increasing rapidly. Almost all the peoples are active on social media platforms like facebook, instagram, twitter etc. On these platforms we post text messages, images, audio files, video files. More than millions of people use these platforms it means more than billon of messages and to store these messages there is a database which is present. These messages can be stored in RDBMS but RDBMS is not efficient to store and process the data which is more than 1 million and this type of data is a unstructured data and RDBMS is not able to store unstructured data so here NOSQL databases comes into picture. NOSQL databases are efficient to store and process the unstructured, semi structured as well as structured data also.

In this project first of all we imported CSV file into mongoDB by using mongo-import command and then it got converted into JSON format(JSON is a format in which all the data in MongoDB is stored).

Here, we are connecting our Mongoose is an object data

database with node.js with the help of mongoose.js library, modeling (ODM) library that provides a rigorous modeling

environment for your data, enforcing structure as needed while still maintaining the flexibility that makes MongoDB powerful. After that we are embedding mongodb queries in nodejs by which it will process the input data given by the user and will print the desired output from database by taking the given input into account.



```
{
        "_id" : ObjectId("5e84a8088b5a4004f695bafd"),
        "State_Name" : "Andaman and Nicobar Islands",
        "District_Name" : "NICOBARS",
        "Season" : "Whole Year ",
        "Crop" : "Coconut ",
        "Area" : "16759.00125",
        "Production" : "62585000.0",
        "Soil" : "clay",
        "moisture_content" : "27.721056137642854",
        "transport_cost" : "4000.8866947766733",
        "price" : "9234.585518458325",
        "P_A" : "3734.411083"
}
```

Figure 2.5: Dataset in JSON format

3. Future Prospective

The technology in our time is changing and improving fast. Life has come to smart phones and tablets from desktops and laptops and everyone own smart phones. As of now this application is only available for farmers as a web application but the android application of this project will also be made by which some more problems of farmers will get reduced. Android application will be based on location-based services, by which farmers will have to make less efforts, they won't need to type or select anything, they will just open up the application and on the basis of their location and choices, results will get displayed.

In future, on this project ML lib will be applied which is an apache spark's scalable machine learning library by using this library many algorithms like regression, clustering, collaborative and classification will get implemented on this application by which it will give better results and will predict more results.

We will also use Google Maps API's to show the Maps of the specific region to the farmer and will increase the dataset to give more information to the farmers.

References:

[1] Ch. Chandra Shekhar, J. Uday Kumar, B. Kishor Kumar, Ch. Shekhar, Effective use of Big Data Analytics in Crop planning to increase Agriculture Production in India.

[2] Open data ecosystem as per National Data Sharing and Accessibility Policy (NDSAP) initiated Open Government Data (OGD) Platform, https://data.gov.in/catalogs/sector/Agriculture9212.

[3] M. Moorthy, R. Baby and S. Senthamaraislvi, "An Analysis for Big Data and its Technologies", International Journal of Computer Science Engineering and Technology (IJCSET), vol. 4, no. 12, (2014) December, pp. 412-418.

[4] A. Pal, K. Jain, P. Agrawal and S. Agrawal, "A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop", 4th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, doi: 10.1109/CSNT.2014.124, (2014), pp. 587-591.

[5] K. Grolinger, M. Hayes, M. Hayesm, A. L'Heureux and D. S. Allison, "Challenges for MapReduce in Big Data", 2014 IEEE World Congress on Services, June 27-July 2, IEEE Computer Society Washington, DC, USA © 2014, pp. 182-189.