Analysis of Indian Crops for giving ease to Indian Farmers by using Big Data Techniques

By

Sanket Goyal,

Shivam Mehrotra,

Saurabh Singh and

Saurabh Tyagi

Abstract:

Big Data Analytics is a process to uncover some hidden patterns and information by applying big data techniques, by using this technique we can analyze all the data and can get the significant value from it. We know that in current scenario data can be in any form like in the form of some written text, in the from of audio files or in the form of images and it could be of any form so by applying modern big data techniques like hadoop, mapreduce and No SQL database we can store and process these data very efficiently. Nowdays big data is very important, it is very helpful for the business organizations because by using this technique they can understand the customer needs better, they can understand the new market strategies, by using this technique they can take help from social media to understand the customer behaviour better. Now coming to the agriculture benefit of big data, we all know that agriculture is backbone of our economy and farmer is the backbone of our agricultural practices and our most of the economy depends on farming but due to lack of knowledge they do not able decide in which season which crop should be sown and due this severe results come out which can be in the form of suicide so by keeping all these things in mind, by using hadoop, Machine learning and No SQL databases we will predict which crop should be sown in which season with the help of previous 20 years data. With this project farmers will be able to know the total cost per area for crop, which crop is best for which season, the price for the desired crop and will also be able to get the information whether the specific crop is suitable in their environment or not.

1- Introduction

Basically there are 5 foundational V's of big data which help us to understand it better -**volume**: the amount of data is stored, **velocity**: the speed of data and processing, **variety**: the different types of data like structured, unstructured and semi structured, **veracity**: This context is equivalent to quality. We have all the data, but are we missing something? is this data "clean" and accurate? Do they really have something to offer? and the last is **value**: This refers to the ability to transform a tsunami of data into business. As we know the time is changing rapidly and so as environment. Environment is badly effected due to global warming. Water bodies have become polluted. Pollution gives birth to acid rain and we know water is very essential for farming and without water farming is impossible. Therefore, in this so called modern environment we cannot use the traditional farming techniques we will have to use new techniques for farming. Government is also taking so many initiatives to improve farming. It is making data of previous farming records but we know that the previous farming record is too large because our most of the economy depends on farming and before some decades almost all the people in India were engaged in farming so the record has to be big, if record is large then the dataset will also be very large and with traditional or old techniques it is not possible to analyze the data accurately and that will take time also so, here big data comes into role, as we know big

data is a technique to process and find hidden patterns from large data set. So, with the help of this modern technique we can analyze and process large datasets very efficiently.

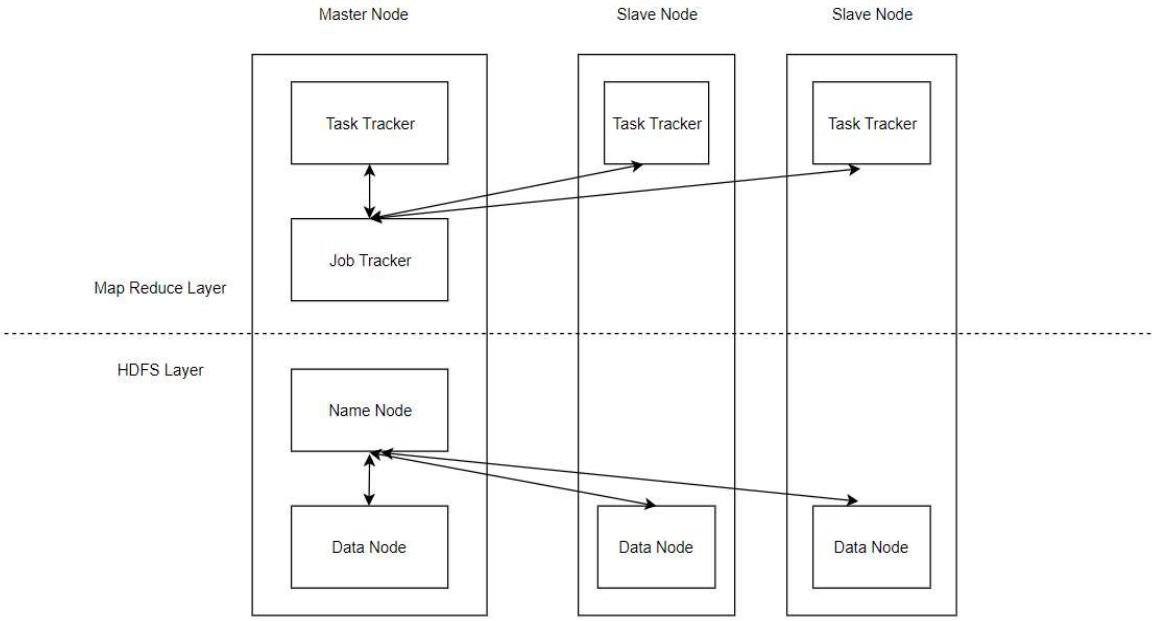

Figure:1.1 Hadoop Architecture

2-Literature Review

In modern world amount of users are very much and there are hugh amount of data exist so, normal techniques are not useful to process hugh amount of data so here bigdata comes into picture, bigdata is more real in comparison to another techniques for processing hugh data. Massively data processing, scale 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) [1].

Indian Prime Minister Narendra Modi -led government has asked Indian Council of Agriculture Research (ICAR) to prepare a comprehensive crop plan for India. A committee under the chairmanship of Praveen Rao, 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 consist of the data of all the states which consist of their districts and the production of crops with their respective seasons. The dataset further consist of area of the districts.

	A	B	C	D	E	F	G
1	State_Nan	District_Na	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 their respective districts.

1	State_Nan	District_Na	Season	Crop	Area	Productior	PA	Price
2	Andaman a	NICOBARS	Autumn	Rice	3.5	10	2.857143	19412
3	Andaman a	NICOBARS	Autumn	Sugarcane	13.4	41.75	3.115672	19886
4	Andaman a	NICOBARS	Kharif	Arecanut	1254	2030.5	1.619219	14506
5	Andaman a	NICOBARS	Kharif	Other Khar	2	1	0.5	14347
6	Andaman a	NICOBARS	Kharif	Rice	80.205	217.7733	2.715209	14489
7	Andaman a	NICOBARS	Rabi	Arecanut	944	1610	1.705508	13490
8	Andaman a	NICOBARS	Rabi	Black pepp	23	8.5	0.369565	13748
9	Andaman a	NICOBARS	Rabi	Cashewnu	1000.5	260.5	0.26037	15457
10	Andaman a	NICOBARS	Rabi	Dry chillies	12	25	2.083333	19666
11	Andaman a	NICOBARS	Rabi	Dry ginger	7	9.64	1.377143	18108
12	Andaman a	NICOBARS	Rabi	Maize	3.84	18.22	4.744792	19853
13	Andaman a	NICOBARS	Rabi	Moong(Gr	1.5	1.1	0.733333	19731
14	Andaman a	NICOBARS	Rabi	Sweet pot	22	208	9.454545	13589
15	Andaman a	NICOBARS	Rabi	Turmeric	2	0.5	0.25	14898
16	Andaman a	NICOBARS	Rabi	Urad	1.5	1.16	0.773333	12672
17	Andaman a	NICOBARS	Whole Yea	Arecanut	1095.074	1222.863	1.116694	15717
18	Andaman a	NICOBARS	Whole Yea	Banana	219.2029	1295.589	5.910455	15031

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 standardize the dataset. We plotted bar graphs using seaborn library.

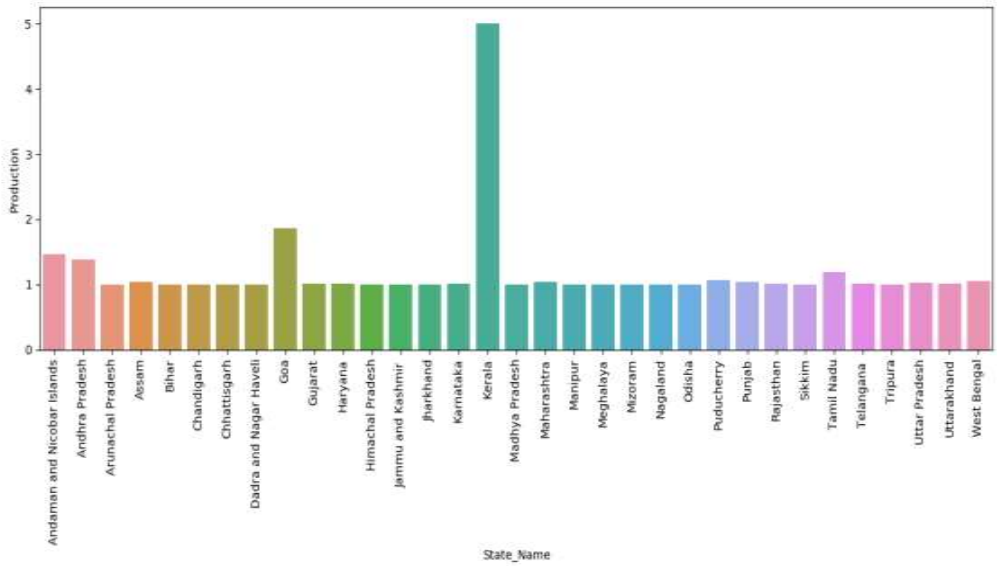

Figure 2.3 Graph of State name v/s production

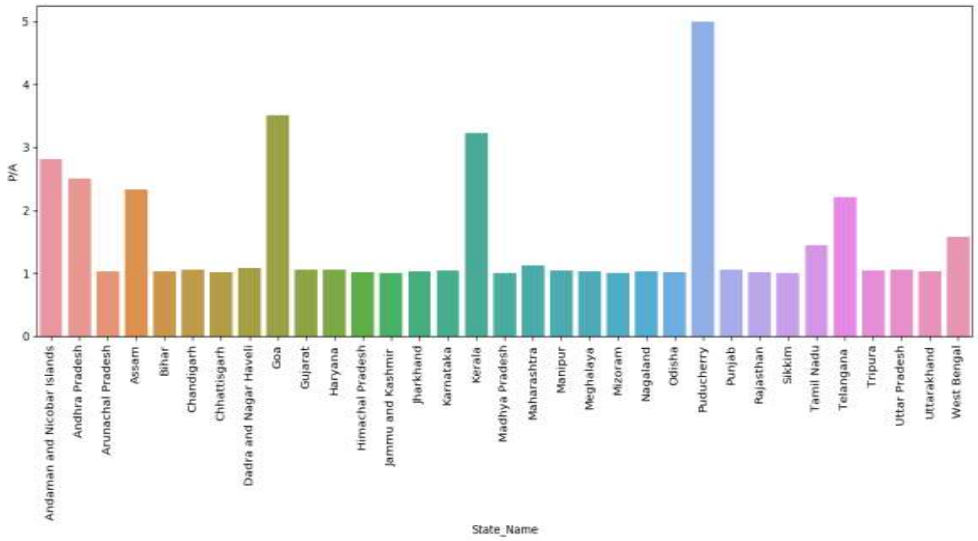

Figure 2.4 Graph of state name v/s “P/A”

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 database with node.js with the help of mongoose.js library, Mongoose is an object data 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("5de933b44c2ef12075d831f9"),
        "State_Name" : "Andaman and Nicobar Islands",
        "District_Name" : "NICOBARS",
        "Season" : "Autumn",
        "Crop" : "Rice",
        "Area" : 3.5,
        "Production" : 10,
        "PA" : 2.857142857,
        "Price" : 19412
}
{
        "_id" : ObjectId("5de933b44c2ef12075d831fa"),
        "State_Name" : "Andaman and Nicobar Islands",
        "District_Name" : "NICOBARS",
        "Season" : "Autumn",
        "Crop" : "Sugarcane",
        "Area" : 13.4,
        "Production" : 41.75,
        "PA" : 3.115671642,
        "Price" : 19886
}
{
        "_id" : ObjectId("5de933b44c2ef12075d831fb"),
        "State_Name" : "Andaman and Nicobar Islands",
        "District_Name" : "NICOBARS",
        "Season" : "Kharif",
        "Crop" : "Arecanut",
        "Area" : 1254,
        "Production" : 2030.5,
        "PA" : 1.619218501,
        "Price" : 14506
}
{
        "_id" : ObjectId("5de933b44c2ef12075d831fc"),
        "State_Name" : "Andaman and Nicobar Islands",
        "District_Name" : "NICOBARS",
        "Season" : "Kharif",
        "Crop" : "Other Kharif pulses",
        "Area" : 2,
        "Production" : 1,
        "PA" : 0.5,
        "Price" : 14347
}
{
```


Figure 2.5: Dataset in JSON format

3- Future Prospective

The time is changing, life has come to smart phones and tablets from desktops and laptops and everyone own smart phones. As of now this application will be available for farmers as a web application but the android application of this project will also be made by which some problems of farmers will get reduced. Android application should be based on location based service, by this service farmers will do 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 display.

In future, on this project MLlib will be applied which is a 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.

Refrences:

[1] Ch. Chandra Shekhar, J. Uday Kumar, B. Kishore 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 OpenGovernment Data (OGD) Platform, https://data.gov.in/catalogs/sector/Agriculture-9212.

[3] M. Moorthy, R. Baby and S. Senthamaraiselvi, "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.