International Journal of Advanced Science and Technology
Vol.113 (2018), pp.31-40
http://dx.doi.org/10.14257/ijast.2018.113.04

Effective use of Big Data Analytics in Crop planning to increase Agriculture Production in India

Ch. Chandra Sekhar[1], J. UdayKumar[1], B. Kishor Kumar[1] and Ch. Sekhar[2]

[1]Dept. of IT, AITAM, Srikakulam, A. P., India
[2]Dept. of CSE, Vignan's IIT, Vizag, A. P., India
[1]dlaxmics@gmail.com, [2]mail2udauykumar@gmail.com,
[3]kishorbhupathi@gmail.com, [4]sekhar1203@gmail.com

Abstract

Big Data Analytics is one of the affirmative platforms to implement a large data analytics task which comprises the way to find unidentified correlations, hidden patterns, and other essential data from an extensive distributed dataset. In this paper, applying data clustering to observe disseminated dataset of expansive crop deals for crop planning may additionally lead to the increase within the agriculture production in India. By using demand in crop kind as the clustering factor then predict the schedule of crop sowing or decide which crop should be sown in the season. Quality of inputs is vital to crop quality and yield, therefore availability and accessibility of right inputs to farmers is a key to farmer empowerment. Besides, it predicts the price of crops in further years or a season which helps farmers to adopt the crop cultivation plan. Quality of inputs is vital to crop quality and yield, therefore availability and accessibility of right inputs to farmers is a key to farmer empowerment. As a consequence, farmer will act as backbone to our nation and its economy. This paper emphasizes the usage of Big Data which enables the farmers to improve the really worth in their products thru less pesticides to be focused.

Keywords: *Big Data Analytics, Map Reduce, Clustering, Crop, ARS, and R*

1. Introduction

Farming turned to day to day activity for our farmers. This farming was inherited from our ancestors from long centuries. So the conventional methods of farming are not apt for this badly affected global warmed environment. Rains, seasons, ground water levels are out of reach for normal civilian. That's why farmer turned poorer. To resolve this, a complete 360 degree solution is required. Government is spending a lot on gathering agriculture data. Data is growing much quicker than the computation speeds. An instance of Big Data is crop sales. Crop sales data will be used to represent the crops data. Since government has actively and constantly gathering crop sales dataset but the size of dataset are considered to be a big data which are a real-world data, which is really a hard problem to analyze it. In order to analyze big data, data mining and statistical techniques can be expanded under parallel and distributed computing platform, also which consumes large amount of storage and computational time on handling massive dataset. It conforms to its name; Big Data Analytics turns out as an essential research topic. Recently, Big Data got its popularity among data scientists and business fraternity.

2. Literature Review

Big data is more real-time in nature than traditional applications. Standard architectures aren't compatible for big data applications (*e.g.*, exa-data, tera-data).

Received (October 18, 2017), Review Result (January 10, 2018), Accepted (January 17, 2018)

ISSN: 2005-4238 IJAST

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) Platform [1]. Yang CL. *et al.*, [2] proposed that the size of the dataset is very massive, so the traditional data analysis methodologies may not be sufficient to predict the crop patterns in the dataset. If the entire process is done by a single node, it usually gets exhausted and consumes time to analyze crop price and yield information. M. Moorthy *et al.*, [3], presented in their work, the data clustering will be handled under distributed Hadoop environment which serves choice in crop planning by forecast the demand in the market at the earliest. A. Pal *et al.*, [4] proposed a popular Map-Reduce concept utilized clustered file system extensively with Hadoop Distributed File System (HDFS). K. Grolinger *et al.*, [5] presented the purpose behind the Map-Reduce paradigm is high scalable which executes massively parallel and distributed over a huge number of computing nodes. The theme of the International Conference CSIBIG-2014[6], Big Data Analytics is transforming every domain and everything that in the society, including science, healthcare, government, finance, IT, *etc.*, Steve Sonka [7], proposed in their work Big Data Analytics can examine so-called all “5V”: volume, variety, velocity, veracity and value. Both organizational and technological innovation required to have impact of Big Data within the agricultural sector. Joseph O. Chan [8] in their work, analytics, the main element, exploits the values from Big Data to invent new models for business and government. The ICT Platform associates the farmers with the buyers of Agri Commodities – viz. Large Retailers, Exporters, Food Processing Units, and Mandies. G. NasrinFathima *et al.*, [9], in their work suggested data mining techniques, the expert can characterize the expansion of farming exercises to fortify different powers in existing agribusiness.

2.1. Big Data Platform

Now data is Big Data! ‘Big Data’ is comparable to ‘Small Data’, on the other hand data bigger subsequently needs completely diverse systems to solve new social issues. Big Data analytics to accomplish whose measure, range, density need novel design, procedures, and classifications to quote importance and to find unknown facts. Big data Analytics is a manner of accumulating (large volume of data), organizing and analyzing of huge set of facts to find out the patterns and alternative helpful information. The information come from several situations such as sensors used to gather climate information, posts to social media sites, digital footage also videotapes, buy business histories and mobile phone GPS signals.

Traits of Big Data: *Volume* - data at scale, quantity of data that has to be processed is growing rapidly, exponential boom in collected/generated data. *Variety* - data in lots of forms, one application is producing/gathering various kinds of forms, to extract knowledge, a majority of these kinds of facts need to relate together. *Velocity* - data in motion *i.e.*, analysis of streaming facts to make possible decisions within fractions of a second, information is being created quick plus need to be deal with quick. *Veracity* - data uncertainty, uncertainty because of data inconsistency &integrity, ambiguities, latency, deception, model approximations.

Big Data will able to handle any kind of data: *Structured Data:* Any data that can be stored, accessed and processed within the variety of mounted format is termed as a 'structured' data. *Unstructured Data:* Data that has no predefined format. Data can receive any format that need to be stored. *e.g.*, Audio, text files, social media. *Semi-Structured Data:* Any data with unknown kind or the shape is classified as unstructured data. *e.g.*, data represented in XML file.

Hadoop = HDFS + Map-Reduce: HDFS^{TM} is a Java-based Hadoop Distributed File System, stores all kinds of data and provides excessive-throughput to get application data. Donald Miner *et al.*, [10] presented in their work, Hadoop framework designed for job planning also cluster resource controlling from distributed applications. Map-Reduce plays a key role in Hadoop framework, it's a software encoding model for parallel processing of huge data sets. The role of mapper is to process input data by that generate intermediate data, then reducer acts a key role to merge all intermediate data and generates final output data.The classic Hadoop clusterdesign is shown in Figure 1.

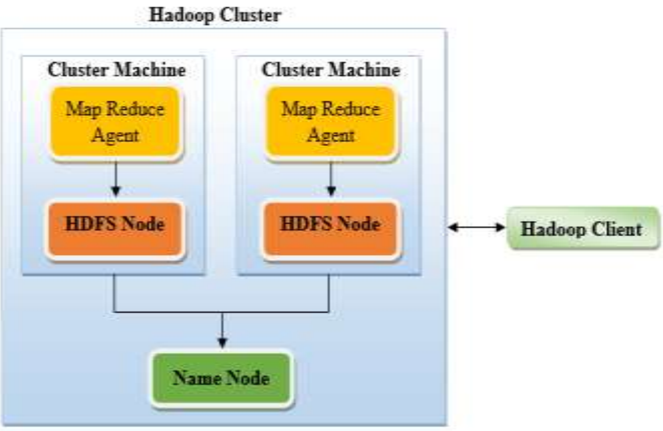


Figure 1. Classic Hadoop Cluster

2.2. State of Art

Online trading platform for Indian agriculture: State, Andhra Pradesh is called as rice bowl of India, to stand for that notation we need to work hard with good computational approach to do that inception is a key. Andhra Pradesh is at the forefront of technology-driven transformation in agriculture, and has fixed determined agricultural aims in the Sunrise Andhra Pradesh Vision 2029. To build on this momentum, the AP AgTech Summit 2017 [11] agreed to sophisticate in a manner that for all international leaders as well as business heads, start-up founders, leading policymakers, and technology consultants to debate innovative ideas for agricultural transformation. The State has recently launched a smartphone-based plant diagnostics app in Telugu in collaboration with ICRISAT.

The State's Agriculture Department is extensively using predictive analytics based on satellite data to forecast water-stressed farms crops-wise in order to provide productive irrigation through rain guns, another technology initiative introduced last year for rain-fed farming. M-trading enables small and marginal farmers to sell their fresh products directly to businesses across the country using their smartphone.

A New Age of Agriculture with Big Data: On the way to increase effectiveness of entire agriculture value chain by enabling financial inclusion of farmers through data driven insights, improving the efficiency of marketplace and the supply-chain for agri input producers, traders, and policymakers, by creating unique risk management solutions that make use of multi-sensor data such as satellites, IoT, and drones, and artificial intelligence algorithms to derive actionable intelligence.

Refined agricultural digital inputs for farmers to procure a wide choice of quality inputs that range from soil-seed-sale. The utilization digital records about farming practices show a significant part in documenting measures to figure out the factors of particular sales. As a consequence, overseas companies are greater useful in new budding trade crops to incorporate into the universal supply chain. On this scope, precision farming has become a key pattern in automated nations. As an instance, a few tasks like excessive weather conditions, unknown soil types may be done well by people.

Model Driven Solutions for Agricultural problems: The Agricultural Research Service (ARS) is the U.S. Department of Agriculture's (USDA) chief in-house scientific research agency [12]. Our individual ability to accomplish analysis that has an effect on the food we generally tend to eat, the water we generally tend to drink, and also the air we generally tend to breathe marks ARS unified analysis directing on countrywide and zone agri-business as main concern. ARS' job to conduct studies to increase and convey solutions to agricultural difficulties of excessive countrywide main concern and provide statistics right to use. The ARS visualization is to push the America in the direction of a higher destiny through rural do research and statistics. Research performed via ARS' Crop Production and Protection Program (CPP) countrywide programs can deliver science-based information and technologies to fulfill or boom crop yield. The use of effective Big Data Analytics tools to increase crop productivity and excellence needs going on:

- useful crop control techniques for innovative and traditional crops that preserve natural resources;
- well-organized integrated control techniques for multiple pests;
- mechanization of control activities to handle labor constraints; and
- advanced crop control models and choice aids.

The effect of environmental conditions on agriculture often give contradictory results at various locations because the factors such as crop, soil, water, weather, climate, and management differences. Therefore, for better to analyze the results, the Big Data approaches that enable researchers to allow new explorations to use Big Data tools to fortify the local best practices to forecast weather conditions with different agro-climatic conditions.

3. Methodology

On this, collect crop sales data from some computers to run data collection program that collect data from internet and store it to an open Agriculture database with two parameters including year and crop type. Special attributes can be used inside the clustering method, such as time (week or month) or crop type. Useful requirements are clustering files and distributing files. The relationship between some attributes within the data set, including rate and trading quantity. The process of submitting, extracting and storing can be repeated as many the numbers of crops inside the crop listing.

3.1. Big Data Analytics Framework

In this framework, crop sales will be analyzed by means of applying data clustering. The data clustering algorithm will run inside of Hadoop platform that offers parallel processing and allotted distributed computing characteristic to the process. With the aid using decision tree, can predict the grouping of the data based on the pattern that has been found previously. To speed up the analytical process, data analytics framework uses Hadoop, Map-Reduce and HDFS to save the input and the results on crop sales data as shown in Figure 2.

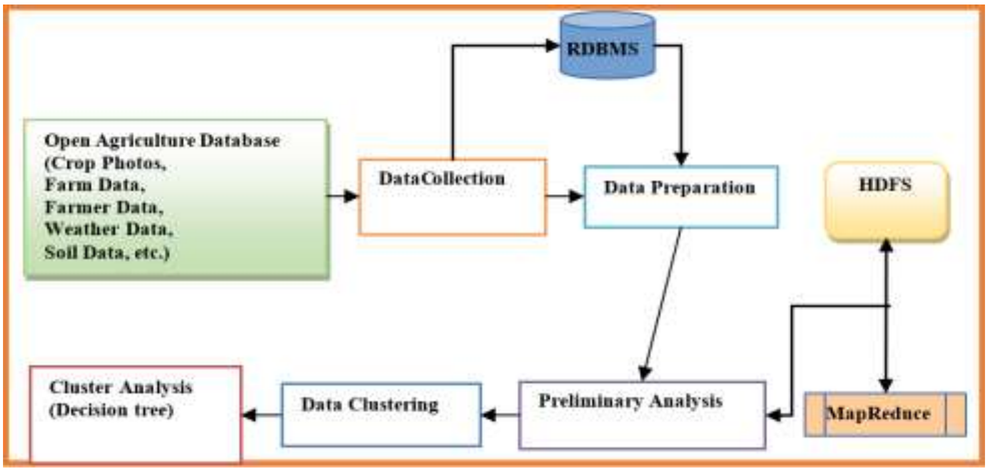

RDBMS
Open Agriculture Database (Crop Photos, Farm Data, Farmer Data, Weather Data, Soil Data, etc.)
DataCollection
Data Preparation
HDFS
Cluster Analysis (Decision tree)
Data Clustering
Preliminary Analysis
MapReduce

Figure 2. Big Data Analytics Framework on Crop Sales Data

A higher clustering technique is needed to deal with the unbalanced data, in particular when there are very few transactions in the intense high rate with very low trading quantity. The experiments of the data clustering on crop sales data may be divided into several steps as shown in Figure 3.

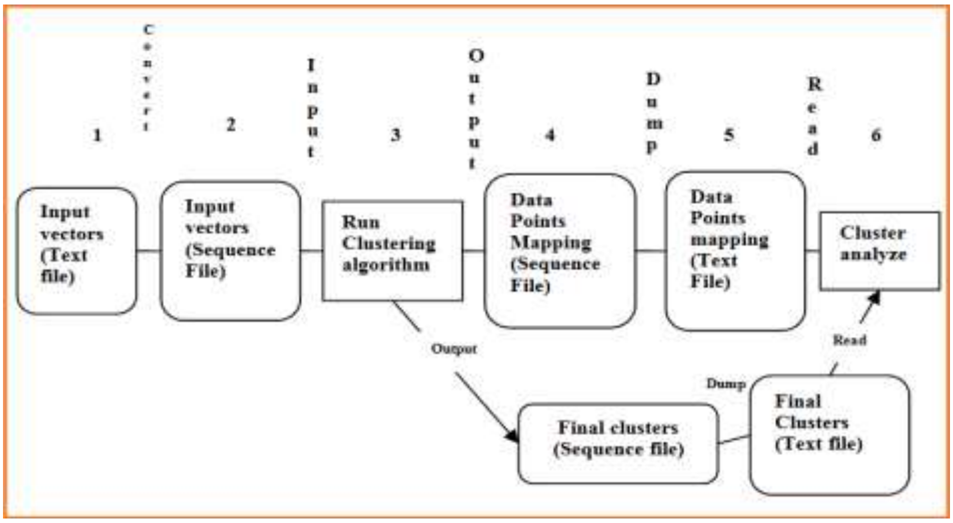

1
Convert
2
Input
3
Output
4
Dump
5
Read
6
Input vectors (Text file)
Input vectors (Sequence File)
Run Clustering algorithm
Data Points Mapping (Sequence File)
Data Points mapping (Text File)
Cluster analyze
Output
Read
Dump
Final clusters (Sequence file)
Final Clusters (Text file)

Figure 3. Crop Sales Data Analysis Steps

Create and prepare the input vectors into text file format. The input vectors are converted into sequential file format. The algorithm can be used to do data clustering on it. In the algorithm system and initial clusters will be generated and stored into a directory and final clusters will be stored into a distinct directory at the end of the data clustering. On this step, plotting the clusters into graphics can be beneficial. The crop sales data domain as the key factor, these data have to be transformed to human readable format to make it simpler to investigate the result.

3.2. Data Analytics Using R

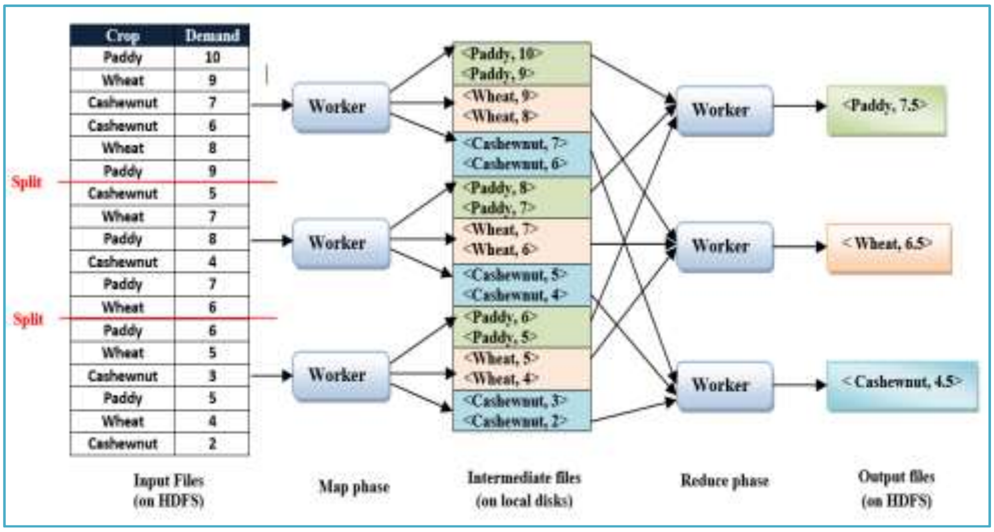
R is a free open source software environment, user interface design for statistical computing and data visualization. The R programming is widely used statisticians for emerging statistical software package and data analysis. It’s vast range of packages and built-in feature that support linear modeling, non-linear modeling, classification, clustering and more; therefore, it is implemented in huge area of financial sector, health care etc. R programming offers extensive range of statistical techniques also data visualization capabilities. Data scientists can use R to run complicated analysis on sample observations, when distinctive a significant correlation or cluster within the data, place the finding into the product through enterprise scale tools. Also, there are R packages for popular open source big data platform, as well as Hadoop and Spark. John M. Chambers [13].

4. Results

Reliable crop dataset collected from an open data ecosystem of Open Government Data (OGD) Platform India published by National Data Sharing and Accessibility Policy (NDSAP) [1]. The information is being applied to revise and have look at crop growing sample and diversification, excessive yield production, agro-climatic region wise performance and crop production contribution to district/state/country.

Case Study 1:

A sample Map-Reduce operation to search out crop demand as shown in Figure 4.



Crop	Demand
Paddy	10
Wheat	9
Cashewnut	7
Cashewnut	6
Wheat	8
Paddy	9
Cashewnut	5
Wheat	7
Paddy	8
Cashewnut	4
Paddy	7
Wheat	6
Paddy	6
Wheat	5
Cashewnut	3
Paddy	5
Wheat	4
Cashewnut	2

Figure 4. Map-Reduce Operation on Crop Demand

It is a method of calculating demand of common crops (Paddy, Wheat and Cashewnut) from input files on HDFS. The information saved in HDFS should split into varied partitions and mapped to workers. The process of mapper phase is to take the input files and map it into intermediate < key, value > pairs of shuffled and deposited into local disks, in which each file holds records with one particular key. When mapping process is over, reducer phase recover the mapping files from workers of remote device and begins the reducing process and eventually stores the top outcome to output files on HDFS.

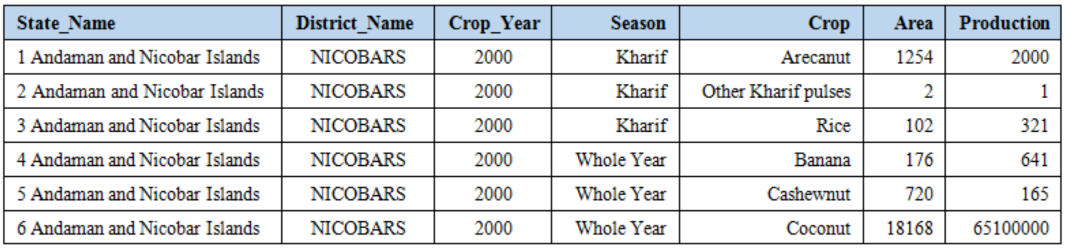
Case Study 2:

Seasons like autumn, kharif, rabi, summer, whole Year, winter. Crop contains of 124 varieties, a few necessary crops similar to Cashewnut, Coconut, Coffee, Paddy, Tobacco, Wheat, and many others.

For example: The dataset contains 2, 46,091 records with seven variables corresponding to State, District, Year, Season, Crop, Area, Production and retrieved sample state-wise crop production as shown in Table 1.

```
# Prepare and load the dataset
data<-read.csv("apy.csv")
246091 obs. of 7 variables
head(data)
```

Table 1. Sample State-Wise Crop Production for 2000

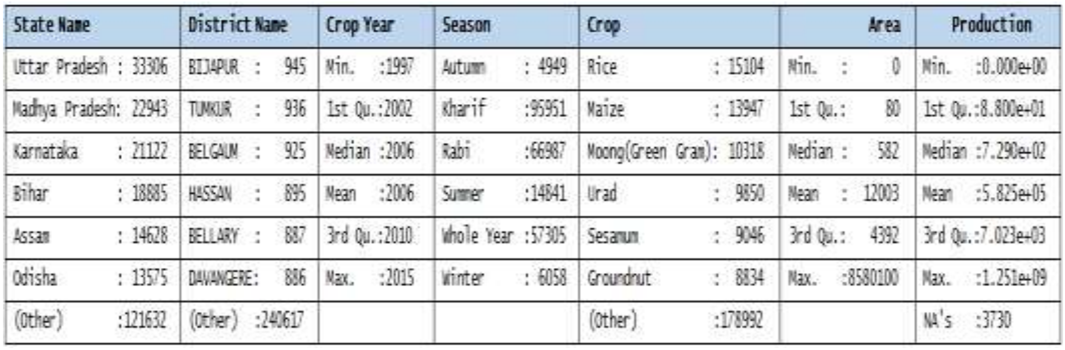


State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
1 Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254	2000
2 Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2	1
3 Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102	321
4 Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176	641
5 Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720	165
6 Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Coconut	18168	65100000

For example: Analyzed the 2 decades crop covered area summary of crop production from the year 1997 as shown in Table 2.

```
# Data Summarization
summary(data)
```

Table 2. Summary of Crop Production from 1997



State Name	District Name	Crop Year	Season	Crop	Area	Production
Uttar Pradesh : 33306	BIJAPUR : 945	Min. :1997	Autumn : 4949	Rice : 15104	Min. : 0	Min. :0.000e+00
Madhya Pradesh: 22943	TUMKUR : 936	1st Qu.:2002	Kharif :95951	Maize : 13947	1st Qu.: 80	1st Qu.:8.800e+01
Karnataka : 21122	BELGAUM : 925	Median :2006	Rabi :66987	Moong(Green Gram): 10318	Median : 582	Median :7.290e+02
Bihar : 18885	HASSAN : 895	Mean :2006	Summer :14841	Urad : 9850	Mean : 12003	Mean :5.825e+05
Assam : 14628	BELLARY : 887	3rd Qu.:2010	Whole Year :57305	Sesamum : 9046	3rd Qu.: 4392	3rd Qu.:7.023e+03
Odisha : 13575	DAVANGERE: 886	Max. :2015	Winter : 6058	Groundnut : 8834	Max. :8580100	Max. :1.251e+09
(Other) :121632	(Other) :240617			(Other) :178992		NA's :3730

For example: Sample 123 observations of 7 variables among 2, 46,091 records crop dataset.

```
# Sample observations from total dataset
index = sample(1:nrow(data), size = 0.0005*nrow(data))
data=data[index,]
123 obs. of 7 variables
# Prepare data matrix for sample dataset
index<-as.matrix(index)
```

Data preparation for cluster analysis in R: A cluster analysis are going to be conducted first of all data preparation; assessing clustering tendency; process the optimum vary of clusters; computing partitioning cluster analyses (*e.g.*: k-means), validating cluster analysis and at last visualize data in plots.

For example: K-means clustering with 6 clusters using 123 sample observations.

```
# Applying k-means clustering to sample dataset
set.seed(123)
km.res <- kmeans(index,6, nstart = 25)
km.res
```

```
K-means clustering with 6 clusters of sizes 24, 24, 25, 18, 15, 17

Cluster means:
       [,1]
1 183315.54
2 221107.88
3  18587.04
4  58790.78
5  97891.27
6 137967.59

Clustering vector:
[1] 3 3 3 2 6 2 2 6 2 3 5 4 6 6 6 2 4 3 3 5 6 1 6 2 6 6 1 5 4 6 3 6 5 6 4 2 4
[38] 3 1 4 1 6 4 6 2 1 5 1 4 2 4 1 1 3 1 2 3 3 5 2 4 1 3 1 5 3 2 1 1 4 4 3 2 3
[75] 1 4 5 2 2 1 1 3 4 4 2 2 1 1 2 2 1 6 2 1 3 1 5 4 3 3 1 3 4 5 2 5 2 3 5 3 5
[112] 1 1 3 3 5 2 6 4 3 6 5 2

Within cluster sum of squares by cluster:

[1] 3605533188 3794542051 2815230631 2425573413 1087373749 2541976512
 (between_SS / total_SS=  97.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"     "size"         "iter"         "ifault"
```

To illustrate outcomes, clustering techniques can be used to analyze the data sample, crop sample and crop sales sample. If the patterns of such crops are already identified, viable set the plan, when to harvest the product.

Find a significant sample in overall observations and also observe the demand by crop/season-wise performance and visualize the final outcomes that are graphically interpreted statistics as shown in below a, b, c, d charts.

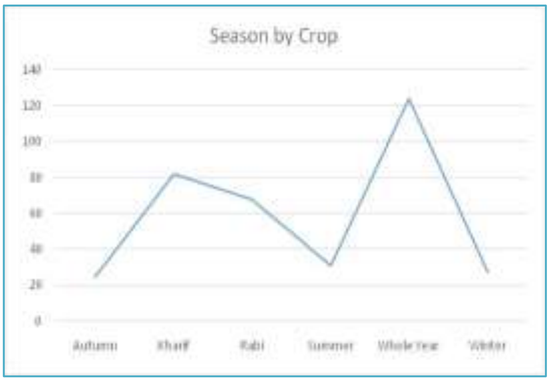


Figure a). Season by Crop Production

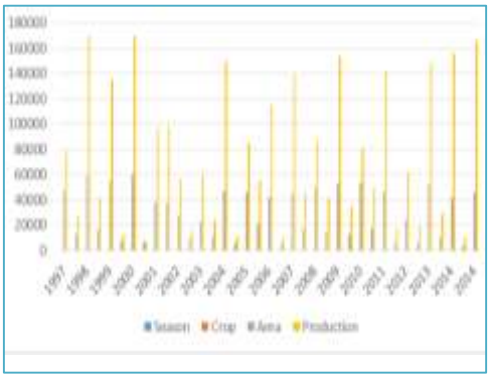


Figure b). Area by Crop Production from 1997

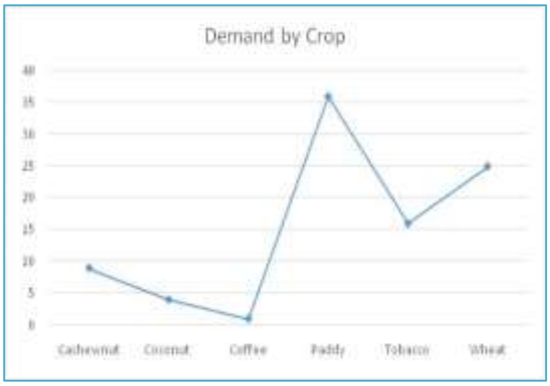


Figure c). Demand by Crop

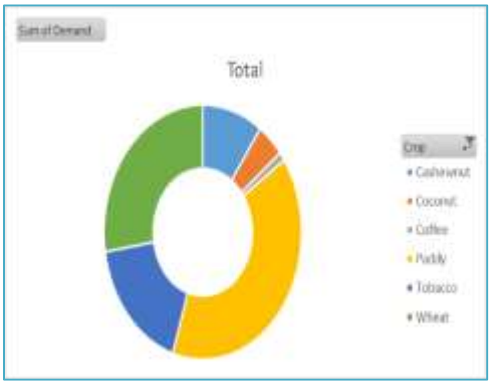


Figure d). Production by Crop

5. Conclusion

Right here people are by and large relying on cultivation in preference to jobs due to their illiteracy. Unluckily their lack of training displays on their methods of cultivation. Here our society needs a higher supervision through technology. Agriculture demonstrator states that who do well do no longer achieve true effects. So, no boom is located in lives of cultivators. Within the destiny, this study might be scaled up in terms of data size and crop variations. Apart from ICT and all, Big Data Analytics is one of the best systems for crop planning to increase agriculture productiveness. Effective use of Big Data Analytics on crop planning may be very significance work to boom agricultural manufacturing and offer the advantages of ICT& different advanced technology to the common man.

References

[1] Open data ecosystem as per National Data Sharing and Accessibility Policy (NDSAP) initiated Open Government Data (OGD) Platform, https://data.gov.in/catalogs/sector/Agriculture-9212.

[2] C. L. Yang and M. R. Nurtam, "Data Clustering on Taiwan Crop Sales under Hadoop Platform", In conference proceedings of the Institute of Industrial Engineers Asian(IIEA) 2013. Springer, Singapore, pp. 827-835.

[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.

[6] CSI on Big Data 2014 Conference on IT in Business, Industry and Government (CSIBIG), In Proceedings IEEE International Conference, Indore, India, **(2014)** March 8-9, pp. c1-c4.

[7] S. Sonka, "Big Data and the Ag Sector: More than Lots of Numbers", International Food and Agribusiness Management Association (IFAMA), vol. 17, no. 1, **(2014)**, pp. 1-20.

[8] J. O. Chan, "An Architecture for Big Data Analytics", Communications of the IIMA, vol. 13, no. 2, **(2013)**, pp. 1-14

[9] G. NasrinFathima and R. Geetha, "Agriculture Crop Pattern Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, vol. 4, no. 5, **(2014)** May, pp. 781-786.

[10] D. Miner and A. Shook, "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems", O'Reilly, **(2014)**.

[11] AP AgTech Summit 2017: Progressive Farmer, Smart Farming, 15-17 Nov, 2017,Vizag, www.apagtechsummit2017.in/.

[12] ARS within USDA - Annual Report on Science, http://www.ars.usda.gov/.

[13] J. M. Chambers, "Software for Data Analysis: Programming with R (Statistics and Computing)", Springer, **(2012)**.

Authors

Ch. Chandra Sekhar, Asst. Prof. in IT Dept. from AITAM (an autonomous institute), Tekkali, Srikakulam, A.P., India. Research interests include Big Data Analytics, IOT, R Programming, Python Programming, Cloud computing, Network Security, Image Processing and Mobile computing.

J. Uday Kumar, Asst. Prof. in IT Dept. from AITAM (an autonomous institute), Tekkali, Srikakulam, A.P., India. Research interests include IOT, Mobile computing, Wireless Sensor Networks, Big Data Analytics, Software Testing and Multimedia.

B. Kishor Kumar, Sr. Asst. Prof. in IT Dept. from AITAM (an autonomous institute), Tekkali, Srikakulam, A.P., India. Research interests include Web Technologies, Cloud Computing and Computer Networks, Network Security, Wireless Sensor Networks, Big Data Analytics.

Ch. Sekhar, Research scholar at JNTUK, Kakinada. Sr. Asst. Prof. in CSE Dept. from Vignan’s IIT (an autonomous institute), Vizag, India. Research interests include Data Mining, Big Data Analytics, Cloud Computing and Computer Networks.