Webology (ISSN: 1735-188X)
Volume 18, Number 2, 2021

Big Data Analytical Architecture For The Agricultural Sector

Pramod Sunagar[1] , S. Rajarajeswari[2] , Anita Kanavalli[3]

[1,2,3]Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bengaluru, India.

Abstract
Agriculture sector in India consists of one of the large work force and also contributes to the GDP of the country. Many sectors in India have implemented modern tools and techniques like data mining, machine learning etc to increase the productivity. Unfortunately the usage of modern tools and techniques in agriculture sector is low compared to other sectors. The major goal is to design data analytics architecture, construct a data analytic framework for large data analysis, and implement a prototype that uses this framework to estimate crop output. The data is gathered from diverse sources and is stored in the NoSQL database. The data collected is unstructured in nature. The R programming language is implemented on the dataset collected to predict the yield of the crops. There are many other factors which can create an influence on the yield of the crop.

Keywords Big data, Data analytics, MongoDB, Spark, NoSQL, Raw data, Pre-processing, R Studio.

Introduction

Big data refers to the large volumes of data created by various applications, which can be structured, unstructured, or semi-structured. The traditional file system is inefficient when it comes to collecting, computation, and examining the generated data [Marjani M, et al., 2017]. The main three aspects of data analytics are first the source of the data, next is the data analysis, and finally, the presentation. Big data can be mainly categorized on three core facets namely volume, variety, and velocity. Over the period of time the other V's like Variability, Veracity, Visualization and Value have been introduced in big data. Gartner established these elements into big data.

This is essential for emerging web-based applications, significantly generating a colossal amount of data. The significant challenge in such cases will be to extract the required information via analysis. The principal research done here is on the factors affecting the yield, various dependencies, and constraints that directly impact productivity. Hence these analyses are essential in making decisions.

Big Data Analytics

Big data analytics is about analyzing the vast data and making decisions from this analysis. The analysis of this data is very much essential, and some of the existing systems for this purpose are:

1. Real-time analysis

This analysis is performed on the real-time data generated from the sensors. The challenge here is the data has to be captured immediately in a short period. Therefore, parallel processing clusters like Hana and Greenplum are used here.

2. Offline analysis

It is usually used when the analysis result can be delayed, like in an enterprise application. The Hadoop-based analysis is used in such cases.

3. Memory-level analysis

This is used when the size of obtained data is smaller and is more efficient for analyzing the accurate data using a non-relational database like Mongo DB.

4. BI Analysis

Analyzing the massive amount of data in terms of terabytes to discover the business planning and according have various strategies for the organization's long-term success.

5. Massive Analysis

It is mainly for analyzing massive data, which is much greater than BI Analysis.

Table 1: Different types of data analysis

Types of Analysis	Type of data	Architecture used	Category
Real-time	Data from sensors	Greenplum HANA	Parallel processing
Offline	Less response time	Scribe	Efficient data
Memory level	Memory for cluster	MongoDB	Real-Time
Business intelligence level	Data more than memory level	Data analysis plan	Both offline and online
Massive level	Massive data	Map Reduce	Mostly Offline

Agriculture and Big Data Analytics

Agriculture is celebrated as the backbone of India and is one of the foremost sectors of the economy. Agriculture in India continues to face several problems as a result of a variety of reasons, including a rising population. The goal is to maximize the use of technology in order to enhance the agriculture industry. Rainfall, soil nutrients, seed data, weather, pests, and a variety of other variables may all impact yield. As a consequence, if we can concentrate on giving insightful details about all of these elements to the various stakeholders, an analysis can be carried out and subsequently supplied to the end-user, mostly farmers, with which he may draw significant conclusions and therefore boost production.

For this purpose, we can make use of Big Data as the data generated here from various sources like the weather forecasting department, agricultural department, etc. on the soil, weather, pest control, crop productivity prediction, irrigation, disease prediction models, and other factors are enormous in terms of TB/PB and may be structured or unstructured. By analyzing this data, the weather analysis, crop prediction, climate change, crop patterns, etc. can be generated and fed to an agro advisory system in the form of text, image, voice, or video so that it can be easily understood by the end-user to make effective decisions.

The primary purpose is to propose a data analytics architecture, build a data analytic framework for big data analysis and implement a prototype that can predict the crop yield using this framework.

Literature Survey

The agriculture field is tremendously benefited by the advances in big data, especially in India, because more research is necessary for organizations to meet big data demand in agriculture. However, the Government of India has three award-winning projects: e-SAGU [Dey U, et al., 2017], Esagu [5], BHOOMI [6], and Sampark, which have played vital task in the development process between agriculture and other modern technologies. Agro system created by IIT Bombay that has assisted farmers in providing valuable inputs to their queries is aqua [8]. It is a multimedia-based tool that delivers crop disease prevention. m-Krishi [7] is a mobile-based agricultural system that provides information based on soil and crops developed by TCS. This application also includes advice from experts to the farmers in audio and video.

Krishimitra is an Android-based application in Gujarat that predicts cotton crop yields using RESTful services. ICRISAT's programmes include Acropodia, Green Phablet, and Agrovoc. Kisan Yojana is a scheme started for farmers that provides the necessary crops information. The mKisan is an android application available on the Google play store which offers various features like SMS, IVRS, etc. There are various other applications like Kisan Suvidha, Kisan Yojana, etc., which try to meet the multiple needs of farmers.

The most crucial is soil classification. It has an impact on many other qualities as well as the importance of land management practices. Soil quality is an essential factor in agricultural soil categorization. It has an impact on soil fertility, water holding capacity, drainage, aeration, tillage, and strength. To categories agricultural soil, we used data mining methods such as GA Tree, Fuzzy classification rules, and the Fuzzy C-Means algorithm. These procedures were applied to soil data gathered [Shah P, et al., 2016], and the results were compared and evaluated. The GA Tree and Fuzzy Classification rules were utilized for supervised learning. However, categorization based on fuzzy rules outperforms GA Tree. For Unsupervised Learning, the Fuzzy C-Means method was employed to categories the soil data. Fuzzy C-Means and K-Means algorithms are used to classify the plants with diseases [Shedthi B S, et al., 2017].

Soil classification is the systematic categorizing of soils based on differentiating properties and criteria that guide usage decisions. Soil categorization is a lively topic, beginning with the system's structure and progressing to class definitions and, eventually, applicable in the field. Soil categorization may be considered from the standpoint of a resource [Bhargavi P & Jyothi S, 2011]. Agriculture forecasting is useful in evaluating risk, deciding on stockpiling, transit, and commercialization. However, rain and weather conditions are very unpredictable and need analysis. Data sets are retrieved and evaluated in order to forecast the yield based on rainfall patterns, wind speed, humidity, and temperature. The basic concept here is to gather data with multiple characteristics impacting yield, categories the data using KNN, forecast the yield using Apriori algorithm, and monitor productivity to aid in marketing and risk decision making [Aishwarya S P, et al., 2019].In this work, the authors have introduced an agricultural intelligent system [Zhao J. C & Guo J. X, 2018] which will be beneficial to improve the yield of the crops. To enhance the performance of such a system big data analysis is utilized.

Big data analysis can be used in animal agriculture to ensure the health and safety of the farm animals [Morota G, et al., 2018]. Tools can be used to monitor the health and movement of the animals, to inspect the farm's surroundings for any threats to the animals, and to keep the farm owner informed of any threats. Big data analysis and machine learning techniques aid in the discovery of hidden patterns in agricultural datasets, allowing researchers to provide valuable information about farming, crop yield, and other factors [Tantalaki N, et al., 2019]. According to the authors, many of the practices listed in the work have not been implemented or have failed to produce valuable results. Many of the most advanced information and communication technologies are being used in precision agriculture to extract valuable information that will aid farmers in employing the best practices [Bhat, S. A & Huang, N. F, 2021].In India, the use of modern tools in agriculture is less prevalent than in other fields. As a result, farmers will suffer losses as well as crop yield [Vandana, B & Kumar, S. S, 2018]. The most recent techniques in big data analysis can be used to forecast crop yield, crop quality, and crop price. These factors will aid farmers in crop planning for the coming seasons. In this work, the authors have combined the data mining techniques and latest information of agriculture to eradicate the problems capturing the complete data [Rao, Z & Yuan, J., 2021]. The authors have used the time series management and other methods to improve the prediction of the agriculture related data.

Technology Landscape

Cassandra

Apache Cassandra is a distributed NoSQL database management system that is open-source. It is meant to manage colossal volumes of data from several sources while maintaining availability of the data 24X7 without any failure. It is highly scalable and can handle massive data which is structured. It is robust and can span many different data centers. In addition, NoSQL database management provides security via data protection, replication of multiple data centers where failure detection is easy, and data can be recovered.

Spark

Apache Spark is cluster computing framework which is also open-source. It provides data parallelism and fault tolerance. It is used for processing a large amount of data and is better than Map reducing when the latency of data is low. It provides APIs that support various programming languages and integrates well with other data sources.

MongoDB

MongoDB is mainly used for the NoSQL database. It is highly flexible open-source, and most suitable for distributed databases, but Cassandra is more durable and can handle failures.

Architecture

The system is interacted by four categories of users: 1) the system developer, 2) the data scientist, 3) the domain expert/agricultural scientist, and 4) the farmers. The following diagram depicts the suggested BDA architecture for the agricultural advisory system.

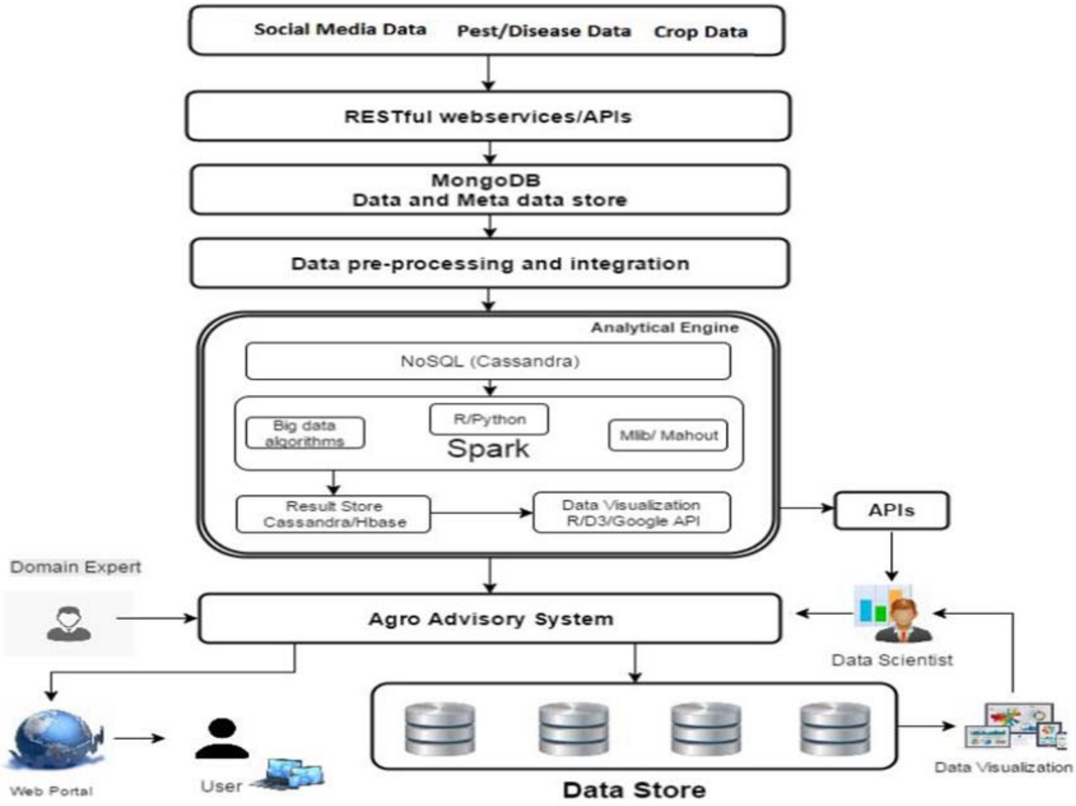


Figure 1. Proposed Architecture

Proposed System

The system proposed here is a significant data analytics architecture with three components for the agricultural sector. Initially a data pipeline, then an advisory system, i.e., the analytical section, and finally a repository, a data store. The actors here are the system developer, the data scientist, the agricultural expert, and the end-user, i.e., the farmer.

Big Data Pipeline

The difficulty with data analytics is in streamlining the data and constructing a pipeline. The architecture proposed is broken down into five stages:

- Extracting Data
- Storing Raw Data
- Pre-processing and Integration of Data
- Analytical Engine
- Data Visualization

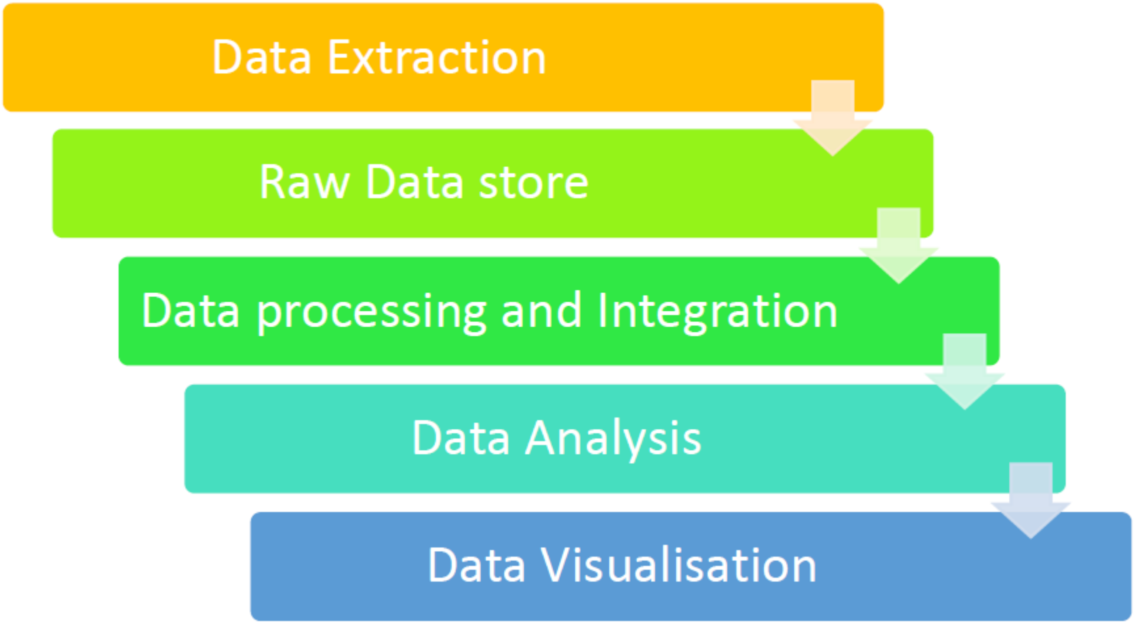


Figure 2. Stage of big data pipeline

Data Extraction

In the agricultural sector, a vast amount of data will be generated based on different factors influencing the yield; they may be weather reports, the water content of the soil, moisture, nutrients, climate, rainfall, pest control, irrigation, etc., and also characterize the yield and the market conditions for the produce. The data necessary for all this is provided by various state and central government organizations like ISRO, SAC, ICRISAT, meteorological department, weather stations, etc. The data from this heterogeneous environment are fetched, processed, and stored as raw data.

Raw Datastore

The data from various sources are collected and stored in the repository, which is highly heterogeneous and unstructured. The data is in the form of XML files or JSON API. The database used here for this unstructured data is MongoDB, a NoSQL database best used where the constraints on the data are not very rigid. There will be multiple data centers where each information center will be a replica of the other, and the data will be stored in this repository. Grid FS is used for storing the specification of data that is the metadata. R libraries are used for extracting the data or importing it to MongoDB.

Data processing and Integration

The raw data obtained from various sources are stored on multiple tables, which need to be integrated for processing. Only if these different tuples are merged can they be processed and analyzed. Hence in this stage, the raw data is unified into one format, which would be easy for querying and processing.

Analytical engine

The data analysis is the main objective here because simply processing and storing it on a repository is of no use. The real-time data can be managed, and effective decisions can be taken when the data is analyzed. Various analytical frameworks are available, but the best suited for this type is Spark. Here Spark is used for analyzing the data.

Data Visualization

Once the data is analyzed, visualizing it is crucial to make effective decisions. The visualization can be in different forms like patterns, graphs, histograms, etc., which will be easy to understand and make decisions. R libraries, D3, Google API, etc., can be used for the same.

Agricultural Sector

Once after the above five stages of streamlining the data, the analytical report, which has information about the rainfall, crop disease, fertilizers, irrigation, soil, etc., is to be examined by the data scientist and the domain expert or agricultural expert to generate recommendations and reports for the end-user.

Data Store

A datastore is where the final analyzed data is stored, and there may be multiple data centers where the replica of the information is stored. Web applications for the end-users can also fetch this. Cassandra database is used to represent the Raw or integrated data and resulting data to allow for efficient retrieval and processing. Cassandra is a highly scalable NoSQL database with exceptional speed and increased availability. The architecture of Cassandra is peer-to-peer, with no system failure. The Cassandra is the only NoSQL database that performs persistent write operations, as opposed to Couchbase, HBase, and MongoDB. The Cassandra is the go-to platform for architects and developers creating large data applications. Structured data is stored in the Cassandra database.

Big data analytics relies heavily on data processing. There are several frameworks available for developing applications which are not only large-scale but also data-intensive on commodity clusters, such as MapReduce and its derivatives. Spark is the ideal tool for managing iterative machine learning workloads and interactive analytics when compared to Hadoop/MapReduce. Spark can beat Hadoop in disc analytics by a factor of ten and in memory analytics by a factor of one hundred. Despite having its own storage system, Spark can access data from a wide range of sources, including HDFS, S3, Cassandra, HBase, and any Hadoop data source. Scala, Python, Java, and R may all be used to create Spark applications. Mila and Mahout Libraries are used to develop machine learning algorithms. The Spark-Cassandra-Connector from DataStax is an open-source project written in Scala. Spark can now read and write data to and from Cassandra. It enables the Cassandra table to be used in Spark RDD. To connect Spark to the Cassandra database, use the Spark-Cassandra Connector. Cassandra is used to save the resulting data for further analysis and visualization. R is a free and open-source statistical programming language that offers data visualization tools and frameworks (ggplot2, shine, and so on). R libraries are used to create dashboard apps.

Implementation

The raw data stored is fed into Cassandra as it is a NoSQL scalable database to process the heterogeneous information efficiently. Since it is highly durable, it is better than Mongo DB, HBase and Couchbase. This is now fed to Spark, which can be implemented using Java, Scala, or Python. Apache Spark is an open-source cluster computing framework. It provides data parallelism and fault tolerance. It is used for processing a large amount of data andis better than Map reducing when the latency of data is low. It provides APIs that support various programming languages and integrates well with other data sources. Other frameworks like Hadoop, MapReduce but Spark has better performance and storage. The connector used for Spark and Cassandra is DataStax, and the project here is implemented using Scala. The resultant data is now to analyzed and visualized. R libraries are used for this, which helps develop the dashboard API, and hence the analyzed data can be imagined.

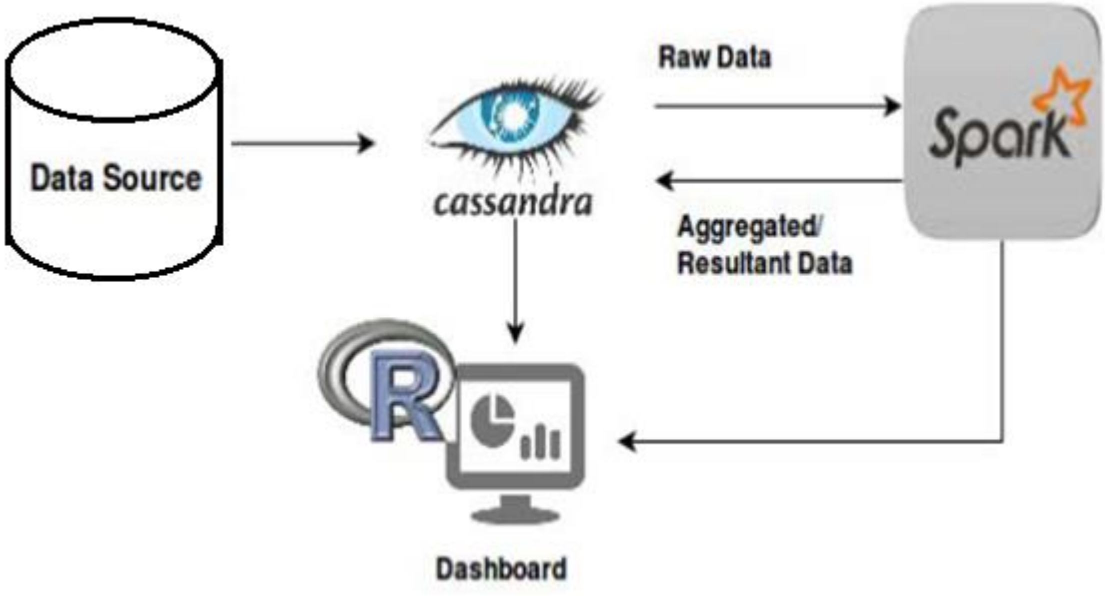


Figure 3. Proposed analytical framework

Results

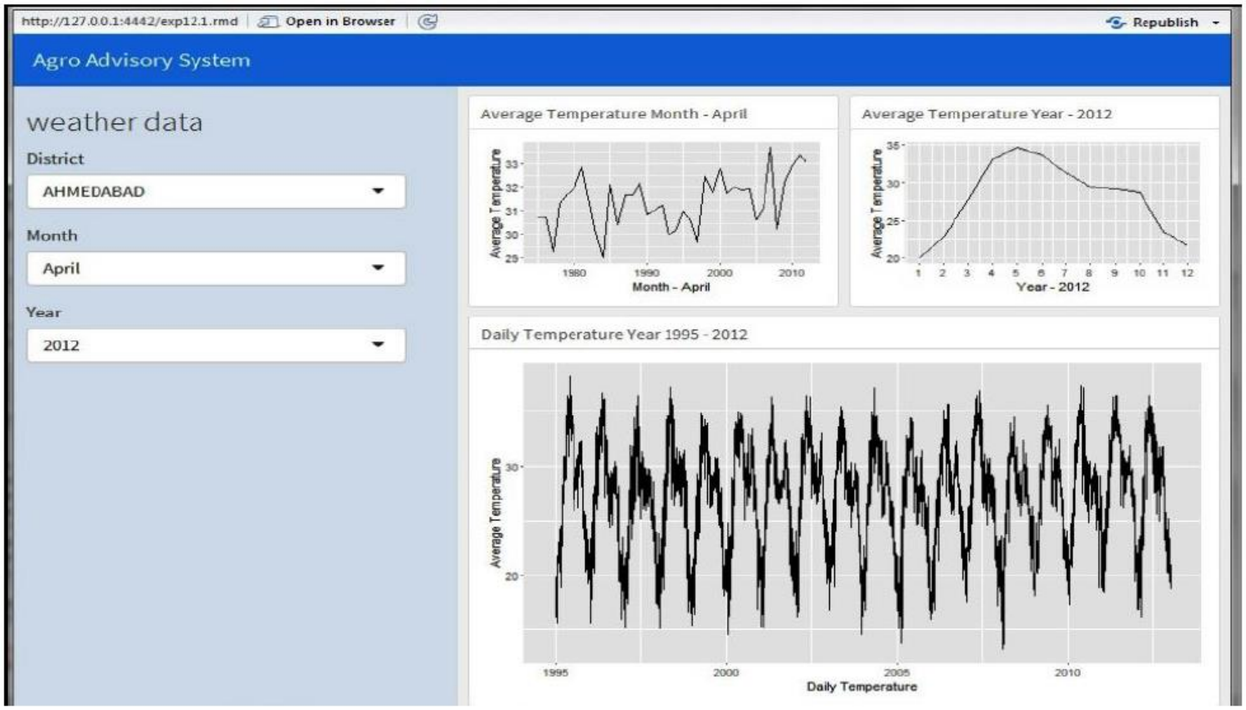
A prototype of this application is being implemented on cotton crops in Ahmedabad district, Gujarat, India, on an Intel machine with 4GB Ram and 3.40GHz processor.

Data Collection

The digital data is available, which is open to the users in India. However, this data set is very few. Hence, collecting this sensitive big data is one of the significant issues. The data ranging from 1995 to 2011 was compiled based on only the yearly crop production. For this implementation purpose, seven months of data is considered based on yield production. Cotton crops usually start from early June to late July and end by January or late February. Therefore, the daily data is converted into a monthly report.

Results

The R libraries are used to develop the dashboard application. It has the analysis made on the crop prediction shown in the following figures. The figure 4 depicts that the average temperature in the Ahmedabad region is 30 to 32 in the month of April and the minimum temperature is around 29 and maximum being at 35 degree Celsius.


Agro Advisory System

weather data

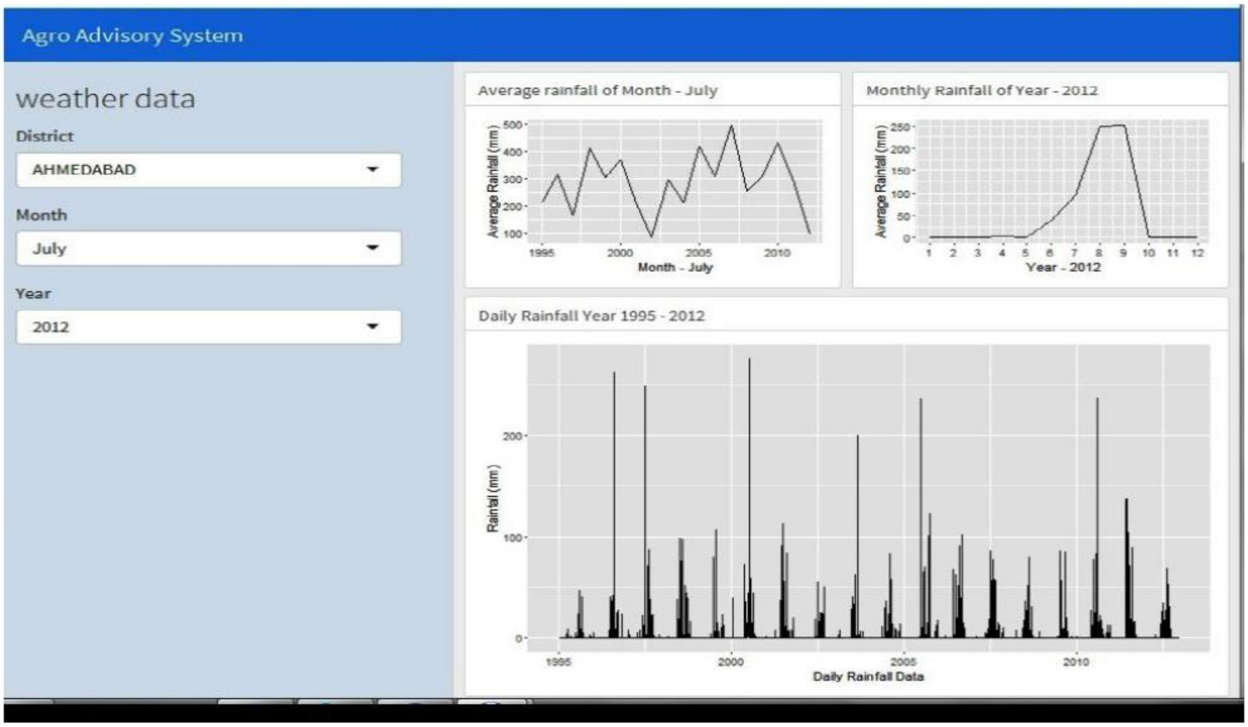
District: AHMEDABAD

Month: April

Year: 2012

Figure 4. Analysis of the temperature data for the Ahmedabad district

The figure 5 depicts that the rainfall data in the Ahmedabad region. The rainfall varies between 100mm to 475 mm. The dataset consists of daily rainfall data from 1995 to 2012. The figure 6 consists of weather trend from 1995 to 2012. The temperature analysis shows that the trend has been almost the same. The rainfall trend is almost identical except some high rain in one instance.



Agro Advisory System

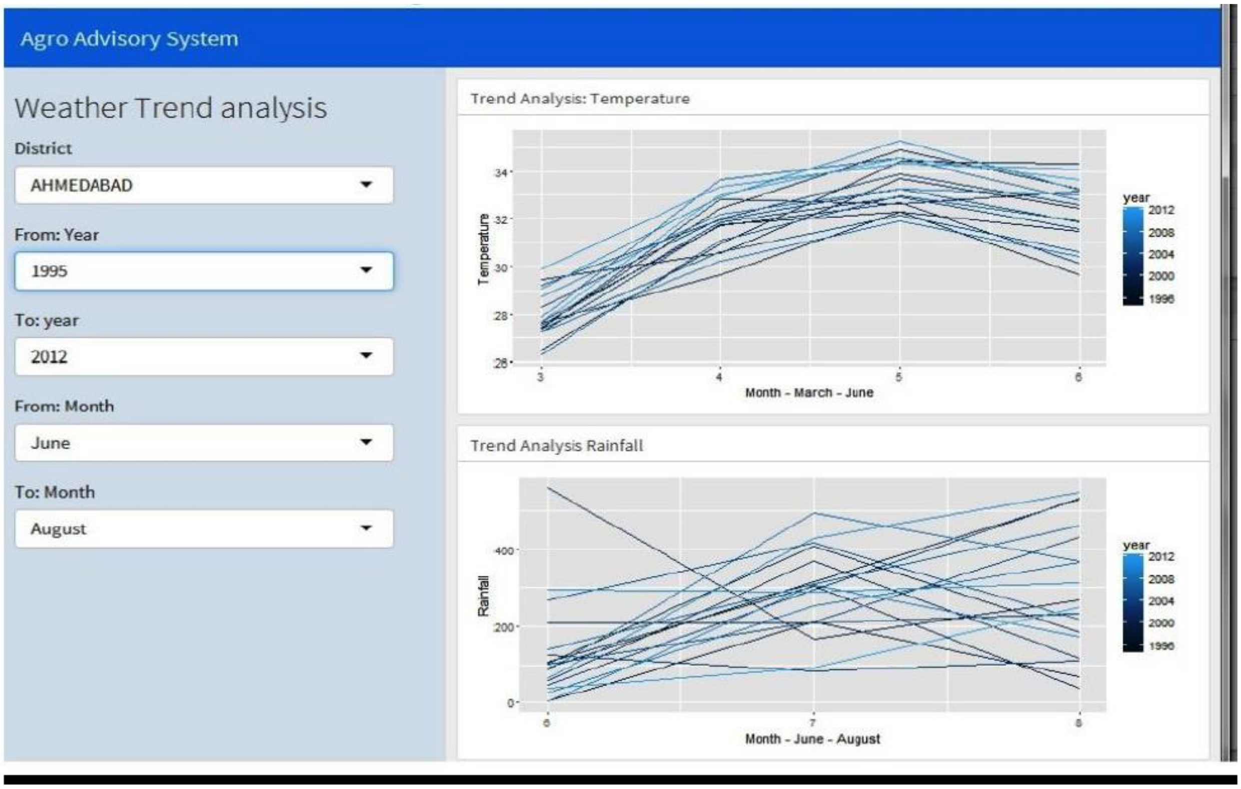
weather data

District: AHMEDABAD

Month: July

Year: 2012

Figure 5. Analysis of the rainfall data for the Ahmedabad district



Agro Advisory System

Weather Trend analysis

District: AHMEDABAD

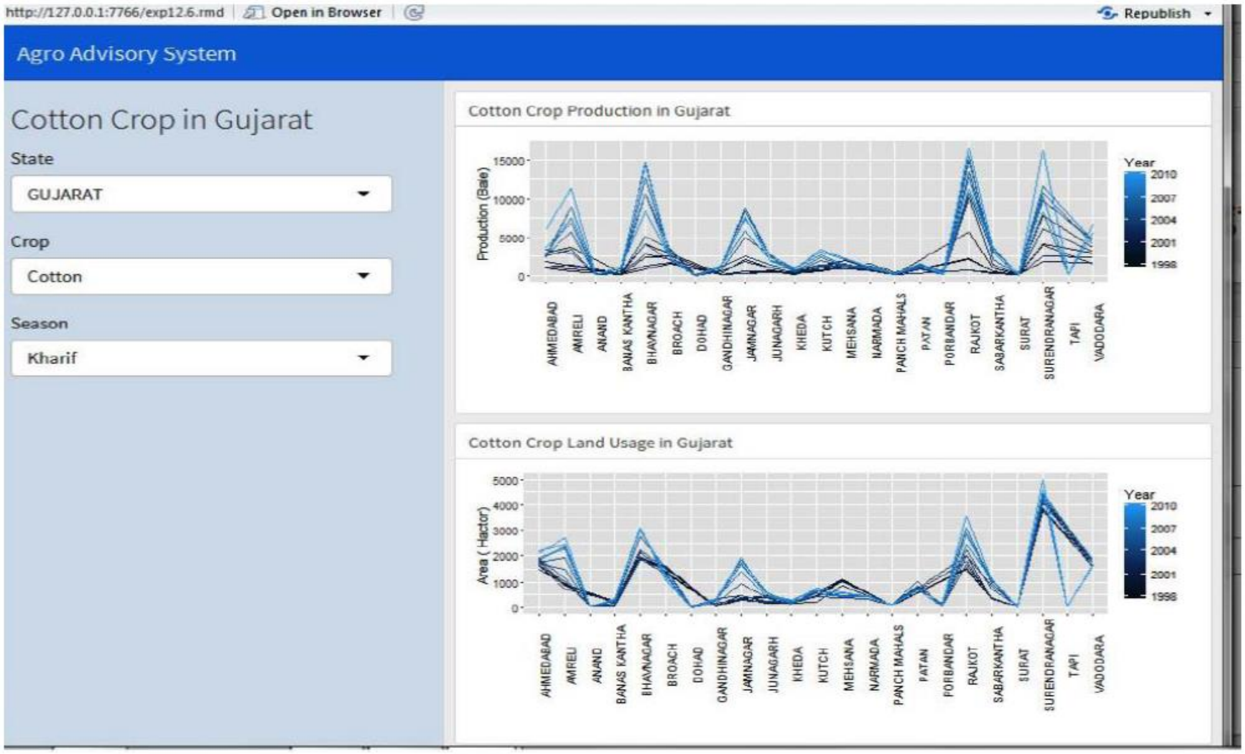
From: Year: 1995

To: year: 2012

From: Month: June

To: Month: August

Figure 6. Analysis of the weather trend for the Ahmedabad district



Agro Advisory System

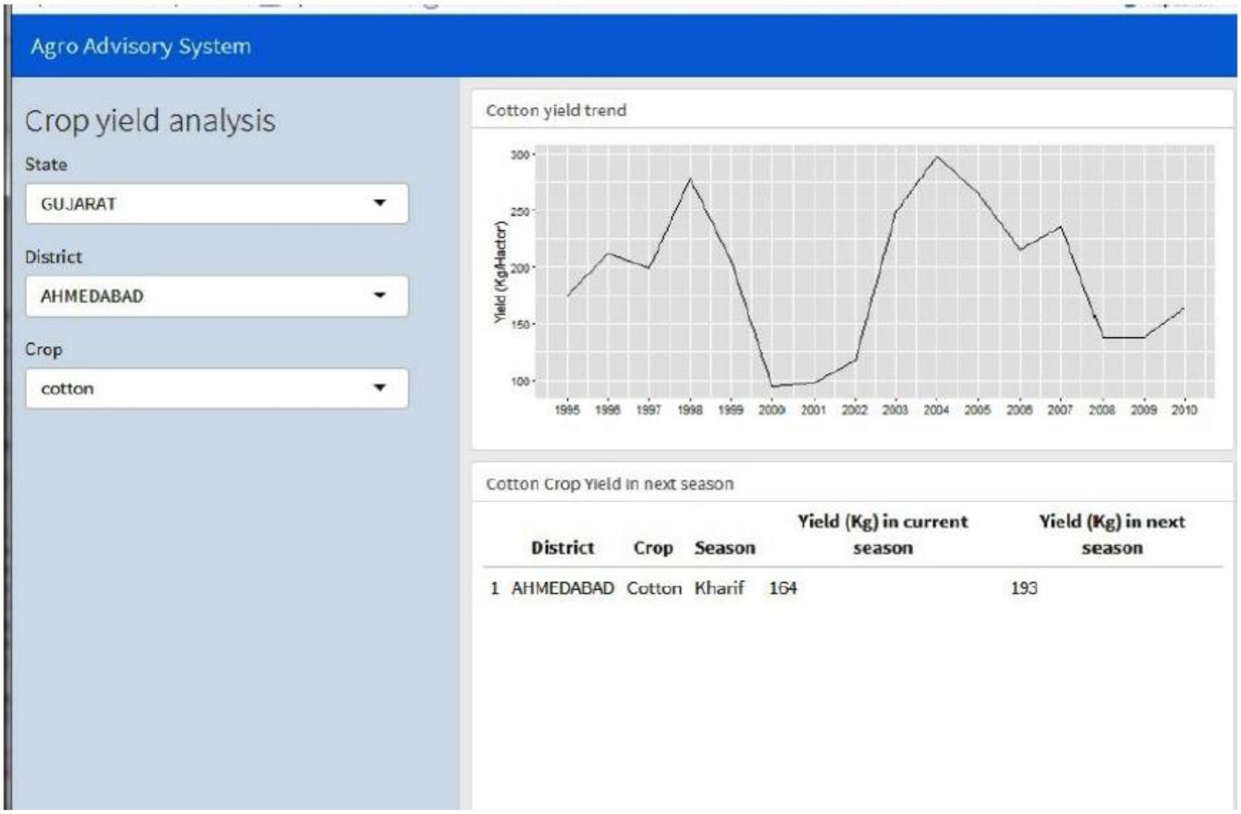
Cotton Crop in Gujarat

State: GUJARAT

Crop: Cotton

Season: Kharif

Figure 7. Production of Cotton crop and usage of land in Gujarat



Agro Advisory System

Crop yield analysis

State: GUJARAT

District: AHMEDABAD

Crop: cotton

Cotton Crop Yield in next season

	District	Crop	Season	Yield (Kg) in current season	Yield (Kg) in next season
1	AHMEDABAD	Cotton	Kharif	164	193

Figure 8. Prediction of next season's Cotton yield in Ahmedabad

The figure 7 depicts cotton crop production and the usage of land in that state during the kharif season. The production of the crops are directly related to the usage of the lands. More lands are used in the cultivation, more is the yield. The figure 8 shows the forecasting of the yield for the upcoming season. Based on the historical data and other factors the big data analysis can help in predicting the yield for the upcoming seasons.

Conclusion

A Big data analytical architecture for the agricultural sector is proposed here. The raw data is collected from various sources, integrated, and processed. This processed data is analyzed, and the data scientist and agricultural experts study the analysis to report to the end-user to be easily understood. This system is being built and implemented in Gujarat for cotton crop yield prediction. The method implemented here can only help the farmers in predicting the yield. Still, we can also include rainfall prediction, crop disease detection, crop disease prediction, and pest detection as future works which can help in better analysis.

References

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., &Yaqoob, I. (2017). Big IoT data analytics: architecture, opportunities, and open research challenges. ieee access, 5, 5247-5261.

Shah, P., Hiremath, D., & Chaudhary, S. (2016, December). Big data analytics architecture for agro advisory system. In 2016 IEEE 23rd International Conference on High Performance Computing Workshops (Hi PCW) (pp. 43-49). IEEE.

Kumar, S. (2022). A quest for sustainium (sustainability Premium): review of sustainable bonds. Academy of Accounting and Financial Studies Journal, Vol. 26, no.2, pp. 1-18

Dr. Ritika Malik, Dr. Aarushi Kataria and Dr. Naveen Nandal, Analysis of Digital Wallets for Sustainability: A Comparative Analysis between Retailers and Customers, International Journal of Management, 11(7), 2020, pp. 358-370.

Bhargavi, P., &Jyothi, S. (2011). Soil classification using data mining techniques: a comparative study. International Journal of Engineering Trends and Technology, 2(1), 55-59.

Dey, U. K., Masud, A. H., & Uddin, M. N. (2017, February). Rice yield prediction model using data mining. In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 321-326). IEEE.

www.esagu.in

http://www.bhoomi.karnataka.gov.in/

http://www.tcs.com/offerings/technologyproducts/mKRISHI/Pages/default.aspx

http://agropedia.iitk.ac.in/

Aishwarya S P, Sunagar, Pramod. Kanavalli, Anita. (2019). Yield Prediction of Paddy based on Temperature and Rain Fall Using Data Mining Techniques. International Journal of Recent Technology and Engineering (IJRTE), 8(2S11), 65-70

Zhao, J. C., &Guo, J. X. (2018, April). Big data analysis technology application in agricultural intelligence decision system. In 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 209-212). IEEE.

Shedthi, B. S., Shetty, S., &Siddappa, M. (2017, March). Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data. In 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 105-108). IEEE.

Morota, G., Ventura, R. V., Silva, F. F., Koyama, M., & Fernando, S. C. (2018). Machine learning and data mining advance predictive big data analysis in precision animal agriculture. Journal of Animal Science.

Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2019). Data-driven decision making in precision agriculture: the rise of big data in agricultural systems. Journal of Agricultural & Food Information, 20(4), 344-380.

Bhat, S. A., & Huang, N. F. (2021). Big Data and AI Revolution in Precision Agriculture: Survey and Challenges. IEEE Access, 9, 110209-110222.

Vandana, B., & Kumar, S. S. (2018, May). A novel approach using big data analytics to improve the crop yield in precision agriculture. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 824-827). IEEE.

Rao, Z., & Yuan, J. (2021). Data mining and statistics issues of precision and intelligent agriculture based on big data analysis. Acta Agriculturae Scandinavica, Section B—Soil & Plant Science, 1-14.