CHAPTER 1

INTRODUCTION

1.1 Overview

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.

1.2 Motivation & Objective

We know that many farmers face problems to select crop for specific season, what cost it will take in whole season and how much will they earn from it. Some middle people take profit from it and make fool of poor farmers. Farmers also suicide because of the failure of crops. So, with the help of our project farmers will be able to analyze best season for cultivating specific types of crops according of their best season.

CHAPTER 2

SOFTWARE REQUIREMENT ANALYSIS

2.1 Software Requirements

2.1.1 Node.js

Node.js is an open-source, cross JavaScript code outside of a browser. Node.js lets developers use JavaScript to write command line tools and for server dynamic web page content before the page is sent Node.js represents a "JavaScript everywhere" paradigm, development around a single programming language, rather than different languages for server- and client-side scripts.

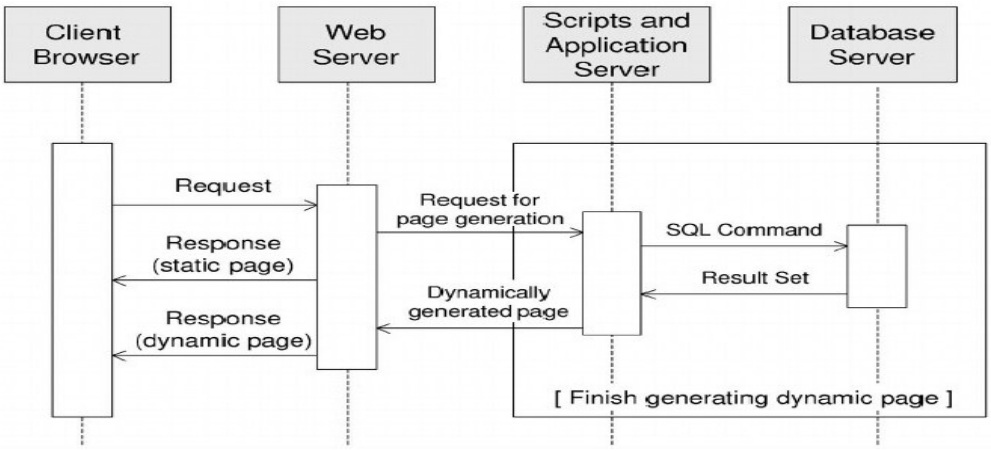

Figure 2.1: Traditional Server

2.1.2 MongoDb

MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schema. MongoDB is developed by MongoDB Inc.

MongoDB supports field, range query, and regular expression searches. Queries can return specific fields of documents and also include user-defined JavaScript functions. Queries can also be configured to return a random sample of results of a given size.

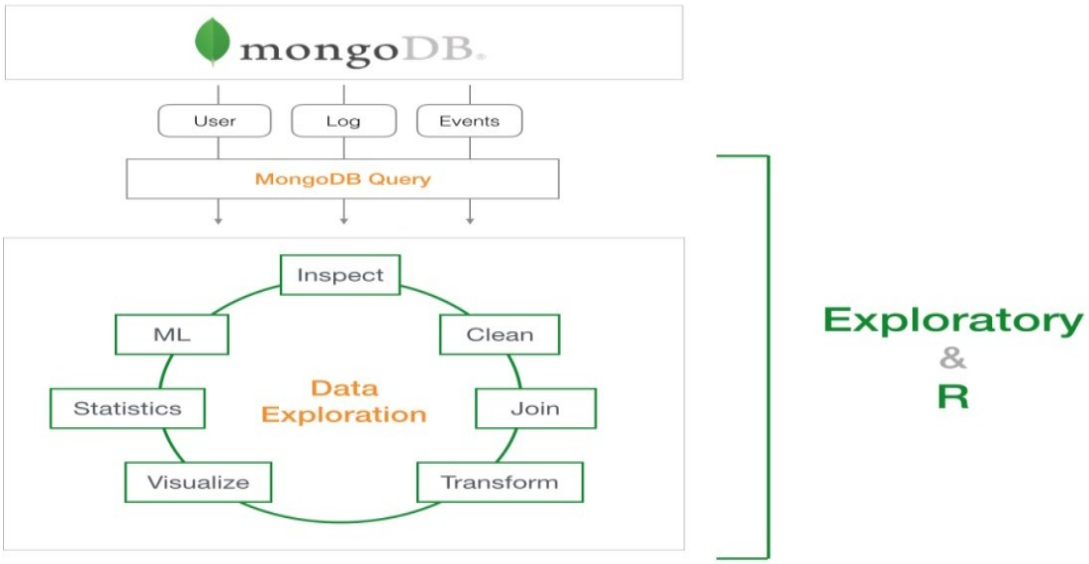

Figure 2.2 MongoDb Block Diagram

2.2.3 Web Browser

A web browser, or simply "browser," is an application used to access and view websites. Common web browsers include Microsoft Internet Explorer, Google Chrome, Mozilla Firefox, and Apple Safari.

The primary function of a web browser is to render HTML, the code used to design or "mark up" web pages. Each time a browser loads a web page, it processes the HTML, which may include text, links, and references to images and other items, such as cascading style sheets and JavaScript functions. The browser processes these items, then, renders them in the browser window.

2.2.4 Visual Studio Code

Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, MacOS and Linux. It comes with built-in support for JavaScript, Typescript and Node.js and has a rich ecosystem of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity).Visual Studio Code is an editor first and foremost, and includes the features you need for highly productive source code editing. This topic takes you through the basics of the editor and

helps you get moving with your code.

2.2 Introduction To Libraries

2.2.1 Pandas

In computer programming, pandas is a library written in python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.The name is derived from the term "panel data", and term for data sets that include observations over multiple time periods for the same individuals.

Key Features

1. Intelligent data alignment and integrated handling of missing data: gain automatic label-based alignment in computations and easily manipulate messy data into an orderly form;
2. Flexible reshaping and pivoting of data sets;
3. Intelligent label-based slicing, fancy indexing, and subsetting of large data sets;
4. Columns can be inserted and deleted from data structures for size mutability;
5. Aggregating or transforming data with a powerful group by engine allowing split-apply-combine operations on data sets;
6. High performance merging and joining of data sets;
7. Hierarchical axis indexing provides an intuitive way of working with high-dimensional data in a lower-dimensional data structure;
8. Time series-functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging. Even create domain-specific time offsets and join time series without losing data;
9. Highly optimized for performance, with critical code paths written in C.
10. Python with *pandas* is in use in a wide variety of academic and commercial domains, including Finance, Neuroscience, Economics, Statistics, Advertising, Web Analytic, and more.

2.2.2 Numpy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding.The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

Key Features

1. A powerful N-dimensional array object
2. Sophisticated (broadcasting) functions
3. Tools for integrating C/C++ and Fortran code
4. Useful linear algebra, Fourier transform, and random number capabilities

2.2.3 Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Key Features

1. Seaborn is a statistical plotting library
2. It has beautiful default styles
3. It also is designed to work very well with Pandas dataframe objects.

CHAPTER 3

DATA ANALYSIS

3.1 Dataset

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. The features of dataset are State_Name, District_Name, Season, Crop, Area, Production, P/A(production per area) and Price.

3.2 Data Cleaning

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.

3.3 Data Visualization

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.

Data Standardization is a data processing workflow that converts the structure of disparate datasets into a Common Data Format. As part of the Data Preparation field, Data Standardization deals with the transformation of datasets after the data is pulled from source systems and before

it's loaded into target systems. Because of that, Data Standardization can also be thought of as the transformation rules engine in Data Exchange operations.

We have used Min Max Scaler algorithm to stardarize our dataset. For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum.MinMaxScaler preserves the shape of the original distribution. It doesn't meaningfully change the information embedded in the original data.Note that MinMaxScaler doesn't reduce the importance of outliers.The default range for the feature returned by MinMaxScaler is 0 to 1.

We have made a bar plot with the help of seaborn library of python of State_Name Vs Production.

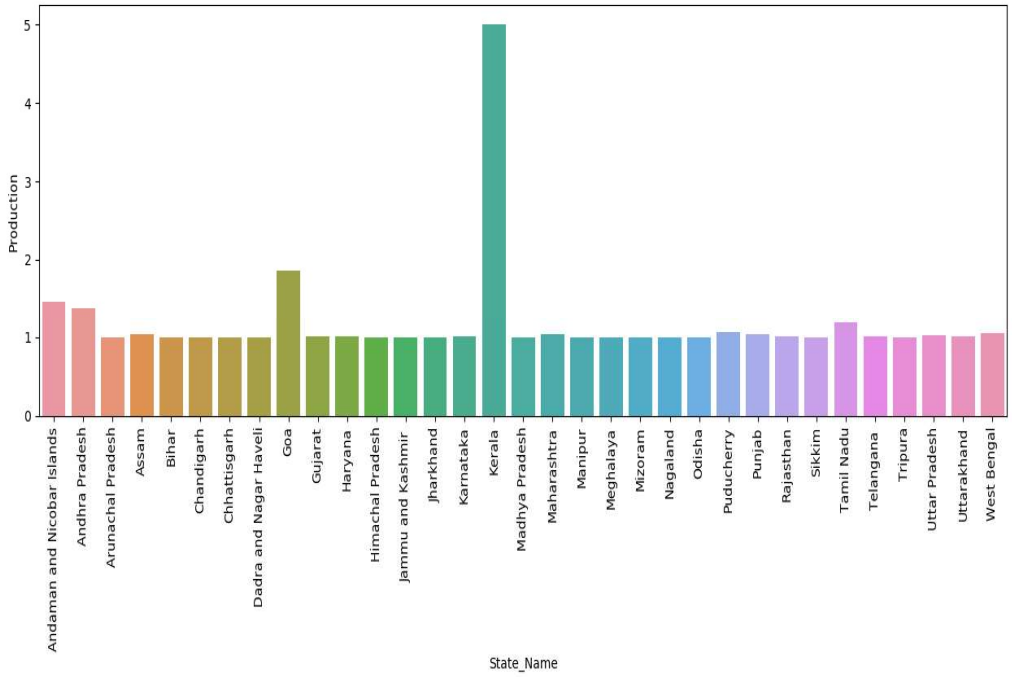

Figure 3.1 State_Name Vs Production

We have made a bar plot with the help of seaborn library of python of State_Name Vs P/A.

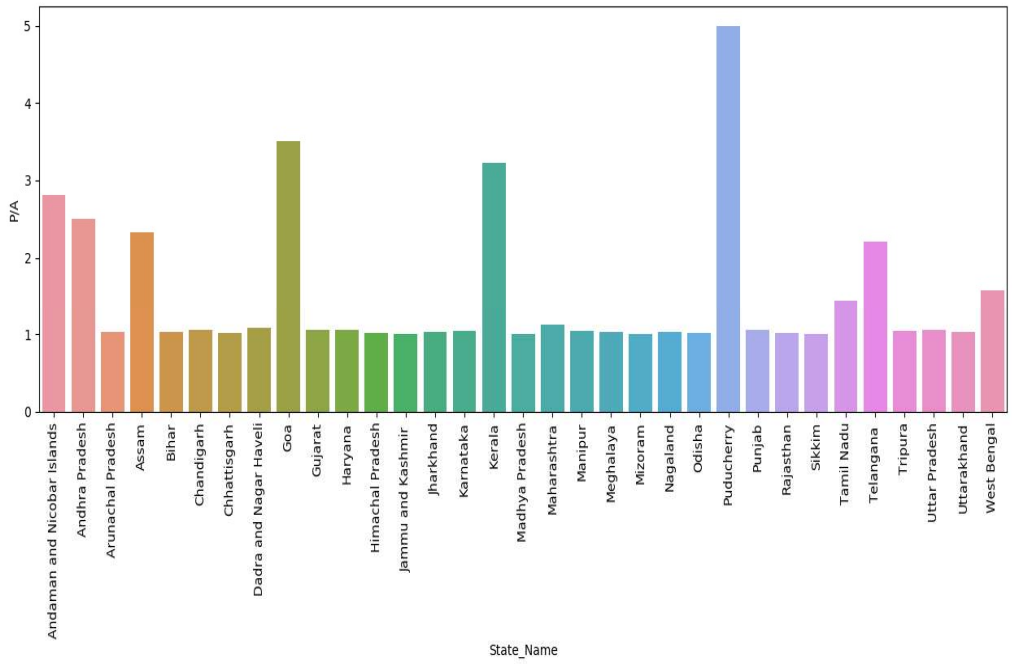

Figure 3.2 State_Name Vs P/A

CHAPTER 4

METHODOLOGY

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).

Import command in MongoDb

mongoimport –d databasename –c collection_name --type csv --file file_name.csv –headerline

Export Command in MongoDb

mongoexport –db database_name –collrction collection_name –type=csv –fields_id,State_name –out File_Name.csv

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.

CHAPTER 5

USER INTERFACE

The Main Page of user interface.

Figure 5.1 Main Page

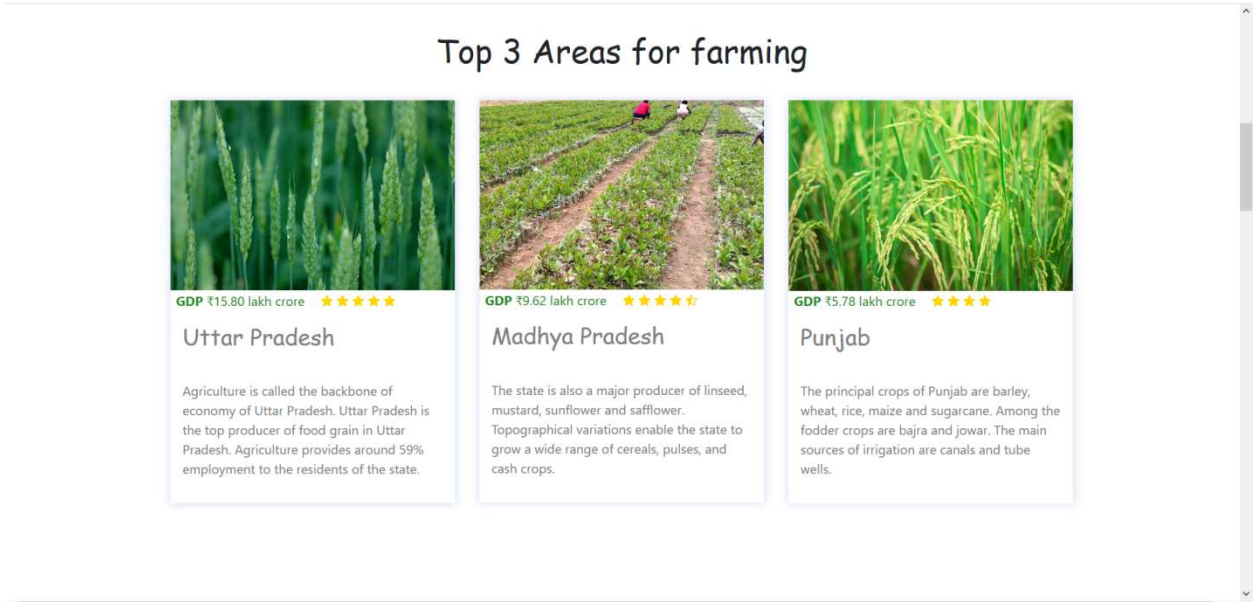

Figure 5.2 Main Page

Question Section contains question and drop down bar.

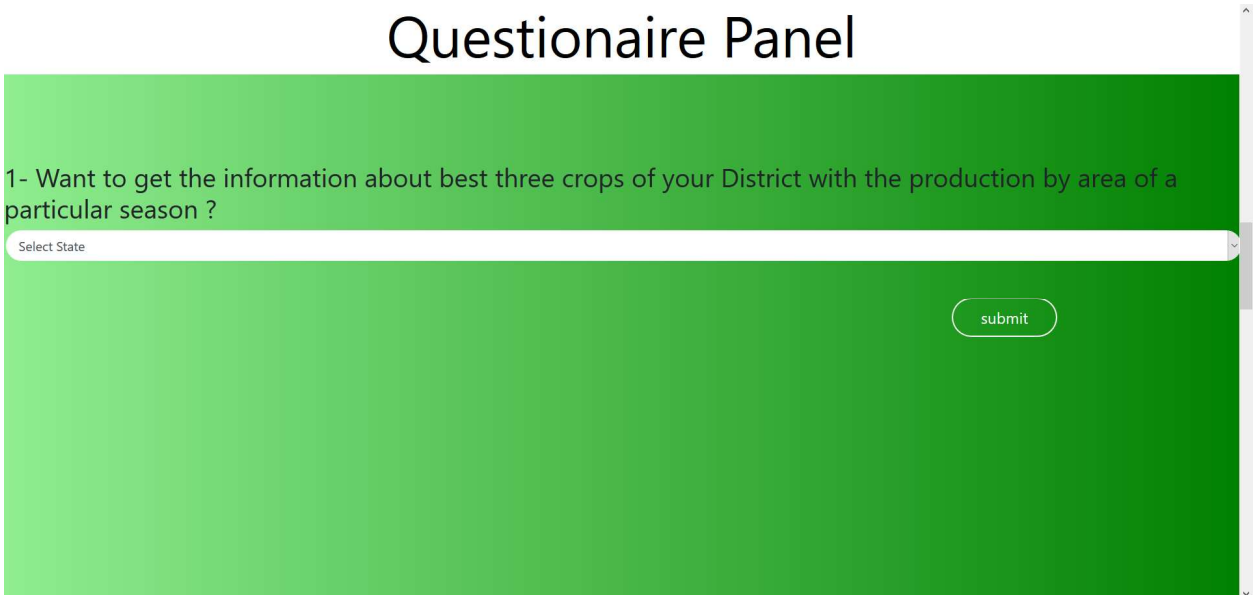

Figure 5.3 Question Page

Output comes up in new page.

Figure 5.4 Output Page

CHAPTER 6

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. We have made a project which helps to identify which crop will be helpful in which season, so as to increase the production of the farmer.

CHAPTER 6

REFERENCRS

1. data.gov.in
2. www.towardsdatascience.com
3. www.mongodb.org
4. www.nodejs.org
5. www.w3schools.com