

## Kisan Call Center Data Analysis Using Hadoop and Hive

Mayank Tripathi<sup>1\*</sup> and Abhay Kumar Agarwal<sup>2</sup>

<sup>1\*</sup>Research Scholar, Computer Science & Engineering, Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India

<sup>2</sup>Assistant Professor, Computer Science & Engineering, Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India

\*[tripathimayank100@gmail.com](mailto:tripathimayank100@gmail.com), [abhay.knit08@gmail.com](mailto:abhay.knit08@gmail.com)

### Abstract

*Kisan Call Centers are operating in each and every district of states in India. Kisan Call Centers were set up by Government of India to help farmer for solving their problems related to agriculture field and allied sectors. Currently, Kisan Call Centers are generating a large amount of data related to various sectors practiced by farmers in the form of queries asked by farmers and replies given by Kisan Call Center operators. As data is growing day by day it has taken the shape of Big Data. There is a need to analyze such data so that effective utilization of data can be done in future. In this paper, we discuss how this information can be stored and processed using Hadoop. HDFS is used for storing data in a structured format, processing of data is done using MapReduce. The stored data is analyzed using Hive. The analysis of Kisan Call Center data will help farmers by presenting a solution to the problem based on specific regional condition and need of farmers. The government of India can also be acknowledged with the problems of farmer related to various states, which will help in policy making for farmers. Scientists can develop innovated methods to solve the problems of farmers.*

**Keywords:** Big Data, Hadoop, HDFS, MapReduce, Hive, Kisan Call Center (KCC).

### 1. Introduction

The challenges before Indian agriculture are immense. This sector needs to grow at a faster rate than in past to meet consumption demands. There is a need to connect our farmers through technology so that innovated methods related to agriculture field and allied sectors can transpire to them. In this measure Government of India has set up Kisan Call Center in each and every district of the state in India, to address the problems of farmers associated with agriculture field and allied sectors. These Kisan Call Center are generating an amount of data in the form of queries asked by farmers and replies given by Kisan Call Center operators using telephonic conversation or SMS. Currently, this conversation is recorded in the form of the traditional file system, i.e. MS-Excel. As over years this data is growing, it has taken the shape of Big Data.

Presently, there is a need to store this unstructured data in a structured data format. This will help to remove redundancy in the data. The processing of structured data will help to filter out refined information from data. Finally, analyzing the data will help to draw conclusive solution to problem asked by farmers related to agriculture field.

Currently, the data stored in MS-Excel is unstructured, processing of such data is a difficult task to get refined information, analyzing of data is also not possible in MS-Excel file system. Therefore, we are using Hadoop for storing the data in a structured format

---

Received (July 15, 2018), Review Result (September 20, 2018), Accepted (September 24, 2018)

Analyzing of data is done using Hive. Hadoop allows for distributed processing of large data sets across clusters of commodity computers using a simple programming model. HDFS (Hadoop Distributed File System) and MapReduce (Programming Model) are Hadoop's core component. The HDFS is the storage component and MapReduce is the processing component. HDFS signifies "Hadoop High Availability". HDFS works by declaring one of the computers as a master node and all other computers as worker nodes. This makes up HDFS cluster. HDFS becomes a virtual file system on top of windows file system. HDFS abstracts by storing data on multiple nodes in a cluster. Hadoop uses MapReduce for processing data on a distributed network of computers. It does this by being what is called "embarrassingly parallel". Hive is running on top of Hadoop to extract data out of HDFS using Key-Value pair generated after processing using MapReduce to analyze data.

In the research work, Comma Separated Value (CSV) files of Kisan Call Center are used to store onto HDFS, processed using MapReduce and finally loaded in Hive table using textfile format to analyze the data related to agriculture field of all districts of the state in India. This will help to analyze the solution given to farmer for problems asked previously so that in future better solution can be recommended to farmers which will increase the production of product related to agriculture field and allied sector. The government will be benefited by acknowledging the problems of the farmer which will help them frame policies benefiting farmers. Scientists can study the problems of farmers, as a result, new breeds of crops can be developed more resistant to diseases, and high yielding fertilizers can be developed to enhanced crop production.

Section 1 contains Introduction, Section 2 contains related work, Section 3 details about an overview of Big Data, Section 4 talks about proposed work and finally, Section 5 and Section 6 presents conclusion and future scope respectively.

## **2. Related Work**

In previous researches, Hadoop and Hive have been used to analyze data related to Stock market, Agriculture, Facebook, Twitter, Airlines, Weather, Earthquake, Medical Health Records, Internet traffic and several others. Some of them are described below:

### **2.1. Stock Market Data Analyzed using Hadoop and Hive**

In this research work, the "New York Stock Exchange" data was stored using the Hadoop framework. The Comma separated values (CSV) files that contained stock information such as stock's nominal price, stock's opening price, stock's highest price and daily quotes etc. of New York Stock Exchange was analyzed using Hive. Using Hive command, a Hive table was created. The CSV files' data was loaded into the Hive Table. By using the Hive select queries, co-variance for the supplied stock dataset for that particular year was calculated.

The co-variance results were used by stock exchange market brokers to predict the possibility of stock prices moving in the upward direction or inverse direction. [1]

### **2.2. The Identification of Crop Disease and Recommend a Solution using Big Data Analytics Framework**

Due to technological advancements data related to agriculture has moved into the era of Big Data. In this research paper, Big Data analytics framework for agriculture data was developed to identify disease based on symptoms resemblance and recommend a solution based on higher similarity with the symptom. The objective was achieved using Hadoop and Hive as tools. Datasets obtained from laboratory reports, websites *etc.*, were cleansed by extracting important information from unstructured redundant data. Next, normalization was done; to extract features from cleansed data. Finally, Normalized data was uploaded onto the Hadoop Distributed File System and saved into a file supported by

Hive. HiveQL is a SQL like a query language and used to analyze and derive results from the data. It helps by identifying disease based on crop disease symptoms similarity and recommends a solution based on historical data of treatment. Results were pictorially represented using graphs treatment based on high symptoms similarity [2].

### **2.3. Electronic Health Record's Predictive Analysis using Hadoop and Hive**

In this research paper, Electronic Health Record data management was developed to present the insights and predict outcomes from the patient record. In the paper, the author presented an EHR data management system to study and process enormous amounts of healthcare data. The systems built using Hive are dynamic and scalable compared to traditional data warehouses. Patient data was uploaded onto Hive from different sources like flat files, web pages, real-time applications and databases. The data produced during analysis used to draw graphs and chart, which helped in the easy analysis of data. The graphical charts were helpful for doctors and researchers to study and suggest medications based on evidence from a huge number of past patient records. The predictive analysis was helpful to treat patients using particular medications, based on a number of factors such as standard of living, family history, smoking practice, and health conditions such as blood pressure and diabetes [3].

### **2.4. Internet Traffic Analysis using Hadoop and Hive**

In this paper, Hadoop and Hive based traffic analysis system (Hobbits) performed analysis of large-sized Internet traffic data transmitted using Internet Protocol (IP) and Transport Control Protocol (TCP) in an easy and scalable manner. Hobbits was the first Internet traffic analysis system that integrated Hadoop and Hive to (i) provided a user friendly and easy to use query interface through Hive queries, (ii) enabled more efficient analysis by using the power of MapReduce together with other advantageous formats, and (iii) avoided the boundaries problem caused while partitioning variable length packets.

In Hobbits, traffic traces were uploaded onto HDFS and the original large file traces were split into smaller HDFS block sized files with help of packet analyzer (i.e., tcpdump), which ensured that there was no single record split across two files, thus avoided boundary issue generated by varying length records. Then the small files were uploaded to HDFS, each of which was stored in one HDFS block to be processed by a map task. Next, packet fields contained in data were extracted using the Hadoop p-cap library and then stored in one or multiple Hive tables. Several formats including Text File Format, Sequence File Format and ORC File Format were used within Hobbits. Once the data was loaded onto Hive tables, users of Hobbits have the facility to run their analysis queries by writing SQL like queries. The users were provided with an easy-to-use query interface and freed them from writing complicated and application-specific analysis programs in MapReduce. [4]

## **3. Overview of Big Data**

### **3.1. Big Data**

Big Data is much diversified in its nature. Primarily, Big Data characterize as not only a large volume of data but it basically consists of a set of unstructured, semi-structured and structured which cannot be stored in simple table formats. There are many definitions available for Big Data here in this paper we summarize definitions from attribute, comparative and architectural point of view, which play a significant role in modelling how one can view Big Data. We defined Big Data from three aspects as follows:

**Attributive Definition:** Big Data was defined by several IT companies like IBM, EMC and many more based on Big Data characteristics volume, variety, veracity, velocity and value. EMC supported IDC definition of Big Data in a report out in 2011[5] that “Big

Data technologies describe a new generation of technologies and architectures, designed to cost-effectively extracting value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and analysis."

Comparative Definition: In 2011, Mckinsey [6] elaborated Big Data as "Datasets, whose size is beyond the capacity of typical database management tools to store, capture, manage, and analyze."

Architectural Definition: The National Institute of Standards and Technology (NIST) [7] suggested that "Big Data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of major horizontal scaling for efficient processing."

### 3.2. Characteristics of Big Data

Big Data has five characteristics: Volume, Velocity, Variety, Veracity and Value, shown in Figure 1, are defined below:

Volume (a large amount of data): Volume means datasets that are huge. This data can be generated every second Ex. Images, Video, Audio, emails and sensor data share every time. We are talking about zettabytes, but yottabytes or brontobytes of data. Defined datasets those are too large and easily amassed into zettabytes, even petabytes of information. A large volume of datasets is not only an analysis issue but also a storage issue.

Velocity (fast processing velocity): It means fast dataset has been produced and data move around. For example, post comment, image, video, audio file on Facebook; watching and uploading videos on YouTube; Big Data technology now allows us to analyze the data without store data even putting it into databases.

Variety (different type of data and source): This refers to the different types of datasets that contain structured, unstructured and semi-structured data, such as emails, audio files, documents, video, images, log files, click streams, call records or financial transactions. Many different attributes in multiple dimensions in the datasets provide more and more information for traditional database management tools or application to handle.



Figure 1. Big Data Characteristics

Veracity (Correct or truthfulness): This basically refers to the trustworthiness of the data. There are many forms of data such as Facebook posts containing an asterisk, hashtag, underscore, tiled, smiles, strikers, abbreviation, typos and colloquial speech contain excellence and exactness which are less handy. Big data analytics tools and technology now allow us to work with these types dataset. The huge volumes often make up for the lack of excellence and accurateness.

Value (low-density data value): Big data tends to have a relatively low-value density, as compared to the data we manage in the traditional system. For Example, the logistics industries have the best mode to transport for goods based on weight and value or a ratio of business relevance to the size of the data.

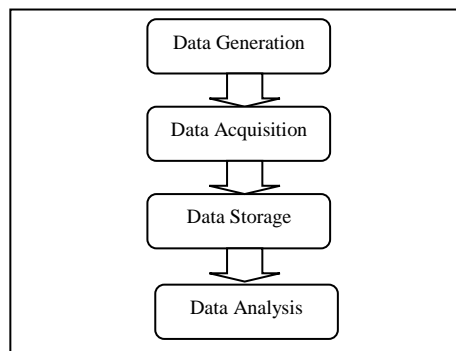
### 3.3. Big Data Architecture

Big Data system architecture provides mainly functions to deal with different phases of today's data life. The architecture of Big Data system is decomposed into four sequential modules as shown in Figure 2. It includes Data Generation, Data Acquisition, Data Storage, and Data Analytics.

**3.3.1. Data Generation:** Data generation is the first main phase of Big Data. Data sources such as sensors, social sites, health care centres, satellite, aeroplane, media, business apps, machine log data, generate large, diverse, and complex datasets. Data generation phase, which shows that the data source contains attribute values, which are mainly from the scientific field, business field and the networking field. The scientific field produces very low whereas business field produces very high attribute value and the networking field produces a very high data rate.

**3.3.2. Data Acquisition:** Data acquisition phase is divided into data collection where data is obtained from various data sources, Data transmission phase and the data pre-processing phase from which useful information is obtained.

**3.3.3. Data Storage:** The data storage is always required to keep the data needed for future use hence a data subsystem in a big data platform organizes the collected information in a format which can be used for the exploration and value abstraction purpose. The data storage consists of the two parts mainly: Hardware arrangement and for managing data: data management system is required.



**Figure 2. Big Data Architecture**

**3.3.4. Data Analysis:** Analytical methods or tools are required to inspect, transform, and model data to extract meaningful value. It has certain purposes like to understand the meaningful information from the data and what value-added functions can be added to the given data. Blackett et al. [8] classified data analytics into three levels:

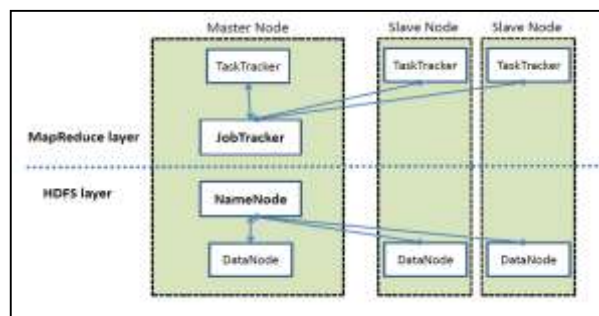
- **Descriptive Analytics:** Descriptive Analytics on the basis of historical data it is concluded that what has happened.
- **Predictive Analytics:** Predictive Analytics on the basis of current data and the testing data it predicts the future trends.
- **Prescriptive Analytics:** Prescriptive Analytics focuses on the how to make an effective decision based on the present scenario.

### 3.4. Big Data Analysis Technologies

The traditional approaches are not an appropriate solution for analysis of Big Data, as many research communities have suggested various solutions for managing various Big Data challenges. Amongst various solutions, Hadoop, MapReduce, and Hive are leading ones. Hadoop is an open source framework and provides two important facilities for Big

Data *i.e.*, storage and processing which is becoming a mainstay in handling Big Data challenges. MapReduce is a programming framework, to process the very large amount of data in parallel. It provides scalability, fault tolerance, reliability and many more. Hive turns Hadoop into a data warehouse, using Hive Query Language (HQL) data can be filtered out and analyzed. Following are some solutions for Big Data Analysis challenges:

**3.4.1. Hadoop:** The Hadoop framework provides distributed processing of large data sets across Hadoop clusters known as Nodes. It was designed to move data from single servers to thousands of machines with each providing local computation and storage. The basic aim was to allow a single query to find and collect results from all the cluster members and this model was evidently appropriate for Google's replica of search support. In a software system to provide a mechanism for storage space, treatment, and information recovery from a large amount of data is the largest technological challenge. Internet and social media today produce together a large amount of data reaching the size of petabytes daily, for example, Facebook, Twitter, Whatsapp etc. These data sometimes contain valuable information, which is not properly extracted by existing systems. Most of this data is stored in an unstructured format using different languages, which is not compatible with existing systems. Parallel and distributed computing, Figure 3, has a fundamental role in data processing and information extraction of large datasets. Hadoop framework was developed to take advantage of parallelization and distributed computing using commodity clusters for storing, processing and updating of a large amount of data. The framework was designed over the MapReduce and used HDFS as a file storage system. Hadoop has key characteristics while performing parallel and distributed computing such as data availability, data integrity, scalability, failure recovery and exception handling.

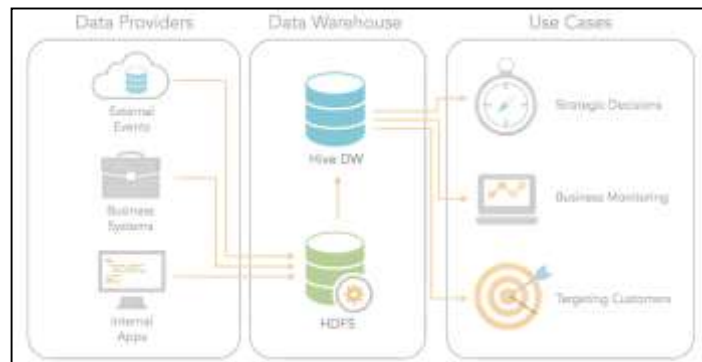


**Figure 3. Hadoop High-Level Architecture**

**3.4.2. MapReduce:** MapReduce was the model of distributed data processing introduced by Google in 2004. Algorithms are used for the processing. MapReduce is formed of Map () and Reduce () procedures. Map () process the data first, it generates the key-value pair and that output then will be sent to Reduce (). Reduce () works, process it and give the final output. All this processing is done in parallel fashion. Firstly Map function is applied to data then Reduce function can be run to combine the results of the Map phases. MapReduce is used for large-scale batch processing and high-speed data retrieval mostly common in web search. MapReduce is the fastest, most cost-effective and most scalable mechanism of returning results processed using distributed computing. Currently, most of the leading technologies for managing 'Big Data' are developed on MapReduce.

**3.4.3. Hive:** Hive was developed at Facebook in the year 2006 to handle a large amount of data which had increased from a few gigabytes to terabytes. Hive is a data warehouse system built inside the Hadoop file system. It is used to study and analyze large datasets which cannot be handled by traditional RDBMS. It provides a user-friendly platform where they can easily use queries similar to SQL but is named differently called HiveQL. HiveQL also helps in managing structured data. It hides the various complexity of

Hadoop like now there is no need to learn Map-Reduces which is very important in Hadoop. Apart from this, no need to learn Java and Hadoop APIs. All in all, it is very useful but with just one constraint that it can be just used for structured data, it cannot handle unstructured and semi-structured data. Hive can be used for log processing. In this logs get partitioned and bucketed in the forms of tables and then can be easily analyzed. Indexing of huge documents can be easily using Hive. Hive is stored inside Hadoop in the form of hive tables from which data is accessed. It is stored inside the Hadoop file system because of its properties like scalability on various type of commodity hardware. The workflow of Hive is shown in Figure 4.



**Figure 4. Hive Workflow**

## 4. Proposed Work

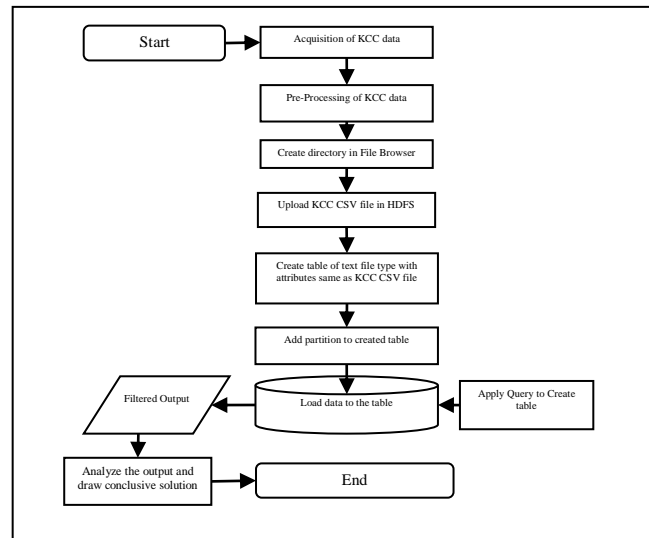
In India, the agriculture sector needs to grow at a faster rate than in the past to meet consumption demands and increase economic growth rate. As about two-thirds of the workforce is directly or indirectly dependent on agriculture. The organized and systematic cultivation will increase the crop production and as a result prosperity of the farmer.

This section details about analyzing Kisan Call Center data using Hadoop and Hive. Kisan Call Center is a service run by Government of India to provide agriculture-related information to the farming community through a toll-free number, 1800-180-1551. The Call Centers' objective is to address the need of the farming community by providing information at the farmer's doorstep, through this toll-free telephone number with help of professionals in particular agriculture field.

There is a need of an innovated method to refine the Kisan Call Center data, to store and process and draw conclusive solution to the problem by analyzing it, taking in consideration environment condition, season and region where the query is generated. The subsequent section talks about the procedure to study the problem using a flowchart, apply the algorithm to process the data and with the help of queries analyze the data.

### 4.1. Flow Chart

In this section, a flowchart is presented, shown in Figure 5, here we discuss, how data is passed through many stages to get the actual outcome required to analyze the data. The process starts with a collection of Kisan Call Center data from "data.gov.in" site [9] based on state, district and month. There need to process the data by removing unwanted data. The directory is created under File Browser section with name same as state name. Upload the file in CSV format to HDFS. Create a table with the same attributes as the KCC data file. Add the partition to table with the state name and district name as the attributes. Load table with KCC data. Apply query on create HIVE table. Get the result of query applied. Analyze the result to offer a better solution to the problem asked by the farmers related to agriculture field.



**Figure 5. Flowchart of Proposed Work**

## 4.2. Algorithm

The proposed algorithm to solve the Kisan Call Center data complexity and analyze it is as follows:

Input: Table 'T' with Set of Attributes ( $n_1, n_2, n_3 \dots n_m$ ), where  $n$  is number of attributes in table 'T'.  
 Output: Table 'T' with filter rows ( $r_1, r_2, r_3 \dots r_m$ ) and columns ( $c_1, c_2, c_3 \dots c_k$ ), where  $m$  is number of rows in filtered table and  $k$  is the number of selected attributes in column.  
 Step1: Create a table of 'n' attribute where 'n' number of attributes in KCC data table.  
 Step2: Add 'i' partition to the table where 'i' number of attributes to alter the table.  
 Step3: Load table with KCC data based on state and district name.  
 Step4: Apply query to analyze the KCC data.  
 Step5: Analyze the Output.

**4.2.1. Description of Algorithm:** A table is created with the same attributes as Kisan Call Center data file in CSV format, as shown in query Q1. The partition is added to the table, partition contains attributes to update the data related to specific state name and district name of the particular year. Load query, as shown in query Q2, is written to load the data state and district wise of the particular year onto the created table. Query to create table and load data are described in section 4.4.1 and 4.4.2.

Finally, results are drawn by applying query Q3, Q4, Q5, Q6, Q7 and Q8 detailed in section 4.5, to table and results are analyzed. This analysis will help to draw conclusive solution to problems of farmers related to agriculture field.

## 4.3. Experimental Setup

An experimental setup is prepared to perform an experiment to obtain the results to analyze the filtered output. The setup includes Hardware and Software specification each of them is as follows:

**4.3.1. Hardware Specification:** The recommended Hardware specifications for research work are as follows:



Processor(CPU)	Intel Core i3-500SU 2.0 GHz.
Operating System	Windows 10 Professional x64.
Memory	4GB
Storage	1TB
Monitor/Display	14'' LCD monitor
Network Adapter	802.11ac2.5/5 GHz wireless adaptor
Others	Lock, Carrying Case.

**4.3.2. Software Specification:** The recommended Software specifications for research work are as follows:

Oracle Virtual Box	Cloudera-Quickstart-VM-5.12.0-0-VirtualBox
CDH Image	CDH version 5.13
Exercise File	Serde Files, Kisan Call Center CSV files [9]

**4.3.3. Dataset Specification:** Dataset contains twelve attributes of which two are used for partition. The attributes of the dataset are defined below:

- Season: It contains information about the season, for example (Kharif, Rabi)
- Sector: It contains information about fields associated with farming, for example (Agriculture, Horticulture, Animal Husbandry, Fisheries, Poultry, and Bee-hiving etc.)
- Category: It contains information about a sub-category of Sector, for example  
Agriculture- Fiber crops, millets, oilseeds, pulses, sugar and starch crop etc.  
Horticulture- Condiments, Spices, Flowers, Fruits and Medicinal plant etc.  
Animal Husbandry- Animals
- Crop: It contains information about a sub-category of Category for example  
Animal- Bovine, Poultry  
Cereals- Paddy (Dhan), Wheat etc.  
Pulses- Urad Dal, Tor Dal, Moong Dal etc.  
Fruits- Mango, Apple, Banana etc.  
Flowers- Rose, Chrysanthemum etc.
- Query type: It contains information about the type of problem of the crop.  
Banana- Plant Protection, Cultural practices etc.  
Ash Gourd- Nutrient Management
- Query text: Query related to the problem  
Banana Plant Protection- Weed management and foliar nutrition in a banana plantation
- Kccans: It contains information about the solution given by Kisan Call Center Operators.  
Foliar Nutrition in Banana- Recommended for spray potassium sulphate 15 gm/lit of water.
- Month: January, February, etc.
- Year: 2013, 2014, 2015, 2016, etc.
- Block: Name of Block in particular district.
- District: Name of district in particular state.
- State: State name is stored here.

#### 4.4.1. Query to create table

```

Q1: CREATE TABLE KCC2015 (Season STRING, Sector STRING, Category STRING, Crop
    STRING, QueryType STRING, QueryText STRING, KCCAns STRING, Month STRING,
    Year STRING BlockName STRING)
    PARTITIONED BY (State STRING, District String)
    ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    WITH SERDEPROPERTIES ("separatorChar" = ",", "quoteChar" = "\"", "escapeChar" =
    "\\")
    STORED AS textfile
    LOCATION '/user/hive/warehouse/KCC2015'
    TBLPROPERTIES ('skip.header.line.count' = '1', 'transactional' = 'true');
    
```

Table 1 shows the execution time when the table is created with the same attributes as of Kisan Call Center data and adding a partition to table based on state name and district name.

**Table 1. Attempt-wise Execution Time of Query Q1**

Query Q1					
Attempts →	Attempt1 (A1)	Attempt 2 (A2)	Attempt3 (A3)	Attempt4 (A4)	Mean (A1+A2+A3+A4)/4
Time(seconds)	161sec	143sec	96sec	54sec	113.5sec

#### 4.4.2. Query to Load the table

```

Q2: LOAD DATA INPATH '/user/hive/warehouse/ kcc2015/Directoryname/file_name.csv'
    into TABLE KCC2015
    PARTITION (state= 'state_name', district= 'district_name');
    
```

Table 2 shows the execution time when table containing data of Kisan Call Center of different State and district wise is loaded onto HDFS using HIVE query Q2.

**Table 2. Attempt-wise Execution Time of Query Q2**

Query Q2					
Attempts →	Attempt1 (A1)	Attempt2 (A2)	Attempt3 (A3)	Attempt4 (A4)	Mean (A1+A2+A3+A4)/4
Time(Seconds)	153sec	55sec	90sec	76sec	93.5sec

#### 4.5.Results

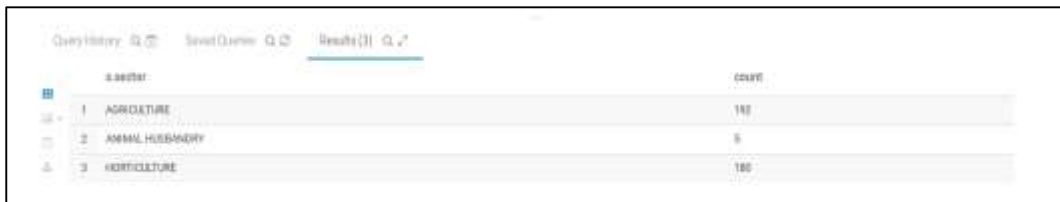
The following queries are created to execute on the Hive table to filter out the result to analyze.

#### 4.5.1. Query to Analyze which Sector is Most Practiced in Particular State.

**Q3:** Select s.sector, count (s.sector)  
From kcc2015 s  
Where lower (state) = 'state\_name'  
Group by s.sector  
Order by s.sector;

This query counts a number of problems related to the specific sector of various district of the particular state. This helps to plot the bar graph between the sector and the count of every sector practiced in that particular state, shown in Figure 6.

Total Map reduce CPU Time Spent: 16.914seconds  
Complete Execution Time Taken: 618.0675seconds



s.sector	count
1 AGRICULTURE	188
2 ANIMAL HUSBANDRY	5
3 HORTICULTURE	188

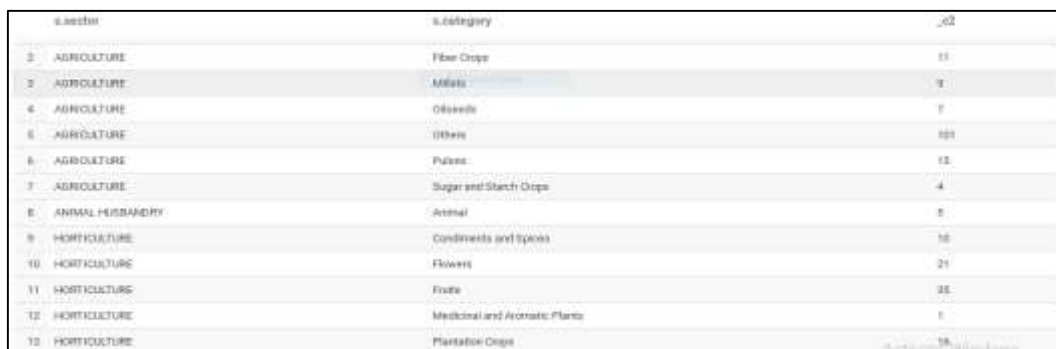
**Figure 6. Result of Query Q3**

#### 4.5.2. Query to Analyze Categories Associated with Sectors in Districts of the Particular State

**Q4:** Select s.sector, s.category, count (\*)  
From kcc2015 s  
Where state = 'state\_name'  
Group by s.sector, s.category  
Order by s.sector, s.category;

This query presents the categories associated with sectors practiced by a farmer in a particular state. This query will give the count of categories associated with a sector of the particular state. This will also help Government to establish field experts in that particular area.

Total Map reduce CPU Time Spent: 13.829seconds  
Complete Execution Time took: 893.712seconds



s.sector	s.category	_count
1 AGRICULTURE	Fiber Crops	11
1 AGRICULTURE	Mills	8
1 AGRICULTURE	Oilseeds	7
1 AGRICULTURE	Others	109
1 AGRICULTURE	Pulses	12
1 AGRICULTURE	Sugar and Starch Crops	4
2 ANIMAL HUSBANDRY	Animal	5
3 HORTICULTURE	Condiments and Spices	10
3 HORTICULTURE	Flowers	21
3 HORTICULTURE	Fruits	38
3 HORTICULTURE	Medicinal and Aromatic Plants	1
3 HORTICULTURE	Plantation Crops	19

**Figure 7. Result of Query Q4**

### 4.5.3. Query to Analyze which Crop Related to a Particular Category is Mostly Grown in that Particular Year

```

Q5: Select s.category, s.crop, count (*) as count
    From kcc2015 s
    Where state= 'tamilnadu'
    Group by s.category, s.crop
    Order by s.category, s.crop;
    
```

This query shows a count of crops grown in the particular category of a specific state. This query will help to determine which crops are sown in particular category of a specific state. The government can pre-access which crop is mostly shown in a particular state.

Total Map reduce CPU Time Spent: 9.767seconds  
 Complete Execution Time Taken: 559.665seconds

s.category	s.crop	count
1. Animal	POLUNNY MAMU	1
2. Cereals	Paddy (Praw)	26
3. Coniferous and Spices	Cardamom	3
4. Coniferous and Spices	Turmeric	3
5. Fiber Crops	Cotton (Bawal)	11
7. Foliage	Chickpea/Besan	3
8. Foliage	Coconut	3
9. Foliage	Garlic	3
10. Foliage	Mustard	1
11. Foliage	Rice	3
12. Foliage	Til/rohi	3
13. Foliage	Red Lent	1

Figure 8. Result of query Q5

### 4.5.4. Query to Study Problem Type and Associated Query of the Crop in the Specific State

```

Q6: Select s.crop, s.querytype, s.querytext
    From kcc2015 s
    Where state= 'state_name'
    Group by s.crop, s.querytype, s.querytext
    Order by s.crop, s.querytype, s.querytext;
    
```

This query shows the problem typically associated with crop and the related query. This will help to identify the problem associated with crop-based problem type in the specific state. This will help to analyze the problem beforehand so that effective solution can transpire to farmers.

Total Map reduce CPU Time Spent: 21.049seconds  
 Complete Execution Time Taken: 1157.4485seconds

s.crop	s.querytype	s.querytext
3. Acid Lime	Plant Protection	asking about sapling diseases
4. Acid Lime	Plant Protection	asking pest to clear
5. Acid Lime	Field Preparation	asking about seed line planting
6. Arroz/Arroz/Grain-Arroz/Arroz	Soeds and Planting Material	Arroz/Arroz seed type
7. Arroz/Arroz/Grain-Arroz/Arroz	Soeds and Planting Material	Arroz/Arroz seed size
8. Arroz/Arroz/Grain-Arroz/Arroz	Cultural Practices	Arroz/Arroz sowing season
9. Arroz/Arroz/Grain-Arroz/Arroz	Plant Protection	Arroz/Arroz pest management
10. Arroz	Nutrient Management	Arroz flower drops
11. Ash Gourd (Petha)	Plant Protection	asking pest in ash gourd
12. Ash Gourd (Petha)	Plant Protection	ash gourd leaf minor management
13. Ash Gourd (Petha)	Plant Protection	fruit fly in ash gourd
14. Ash Gourd (Petha)	Nutrient Management	preventive fruit drop in ash gourd
15. Ash Gourd (Petha)	Plant Protection	leaf miner control in ash gourd

Figure 9. Result of Query Q6

#### 4.5.5. Query to Study Problem Associated with Crop and Solution Told by Kisan Call Center Operator

```

Q7: Select distinct s.crop, s.querytext, s.kccans
      From kcc2015 s
      Where lower (state) = 'state_name'
      Order by s.crop;
    
```

This query studies the problems and solution given by the Kisan Call Center previously to solve any problem related to the agriculture field. This will help to study problem and solution of a specific crop in the specific region, which will be a step to develop innovate method to tackle the same problem in future.

Total Map reduce CPU Time Spent: 20.5415seconds

Complete Execution Time Taken: 987.830seconds

s.crop	s.querytext	s.kccans
Acid Lime	acid lime Fertilizer system	recommended for Fertilizer nutrition in acid/lowpH water soluble fertilizer 10-10-10/5 gm /lit
Acid Lime	sticking past to a tree	recommended for spray dimethoate 2ml/lit of water
Acid Lime	asking about sapote transmission	recommended for state of trunk should be treated with barbasol paste.
Amorpha fruticosa	Amorpha fruticosa pest management	Recommended for spray neem oil 3 ml / liter of water
Amorpha fruticosa	Amorpha fruticosa sowing season	Recommended for Amorpha fruticosa sowing season through out year
Amorpha fruticosa	Stalks are used for	Recommended for 1 kg/ha
Amorpha fruticosa	Amorpha fruticosa seed rate	Recommended for Amorpha fruticosa seed rate 1kg / ha
Apple	Apple flower drops	recommended for spray plantin 3 ml / 10 lit of water
Ashi Gourd (Petha)	ashi gourd leaf minor management	recommended for spray dimethoate 2 ml / lit of water
Ashi Gourd (Petha)	asking pain in ashi gourd	Recommended for spray Thiazethosam 4 g / 10 liter of water
Ashi Gourd (Petha)	ashi gourd sowing season and variety	Recommended for ashi gourd sowing season July and variety -D11 D42
Ashi Gourd (Petha)	ashgourd top dressing fertilizer management	Recommended for apply urea 10kg / ha
Ashi Gourd (Petha)	Water spray for ashgourd	recommended for spray ethrel 0.5 ml/10 liter

Figure 10. Result of Query Q6

#### 4.5.6. Query to Study Correlation with Crop, Season and Associated Problem in a Particular State

```

Q8: Select distinct s.season, s.crop, s.querytext
      From kcc2015 s
      Where lower (state) = 'state_name'
      Order by s.season;
    
```

This query shows the correlation among crop, season and associated problem. This will help to study the problem associated with crop based on the season in a particular state.

This will also to identify the problem more specifically based on symptom common with season, more effective measure can be taken to control the problem.

Total Map reduce CPU Time Spent: 18.248seconds

Complete Execution Time Taken: 735.9215seconds

s.season	s.crop	s.querytext
41 JAKKI	Sugarcane (Middle Cane)	Early shoot borer in Sugarcane
42 JAKKI	Tomato	Tomato Leaf Midge In tomato
43 JAKKI	Tomato	asking about tomato WPM management
44 JAKKI	Tomato	damping off in Tomato
45 JAKKI	Tomato	asking about market rate for tomato
46 KHARIP	Cross	Good planning season
47 KHARIP	Watermelon	scab on watermelon in watermelon
48 KHARIP	Tomato	tomato top dressing fertilizer management
49 KHARIP	Tomato	tomato irrigation availability
50 KHARIP	Tomato	tomato crop forecast details
51 KHARIP	Tomato	tomato basal fertilizer management
52 KHARIP	Tomato	tomato leaf blight
53 KHARIP	Tomato	market rate for tomatoes
54 KHARIP	Tomato	market rate for tomatoes

Figure 11. Result of Query Q8

The Table 3 shows that, when queries Q3, Q4, Q5, Q6, Q7 and Q8 are executed on Cloudera processing engine by varying the state name, mean time of MapReduce CPU execution time and total execution time of queries can be calculated. The execution time varies as the size of data of a particular state varies.

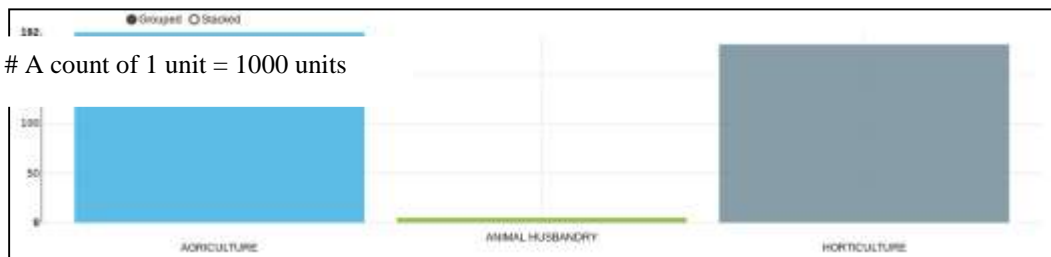
**Table 3. Execution time of MapReduce and total completion of Q3, Q4, Q5, Q6, Q7, and Q8**

Query	Time (seconds)										
	Attempt 1(a1)				Attempt 2(a2)				Mean[(a1+a2)/2]		
	MapReduce Execution Spent	CPU	Complete Time	Execution	MapReduce Execution Spent	CPU	Complete Time	Execution	MapReduce Execution Spent	CPU	Complete Execution Time
Q3	17.290sec		628.189sec		16.538sec		607.946sec		16.914sec		618.0675sec
Q4	14.860sec		911.206sec		12.798sec		876.218sec		13.829sec		893.712sec
Q5	9.659sec		554.564sec		9.876sec		564.768sec		9.767sec		559.665sec
Q6	20.800sec		1135.332sec		21.298sec		1179.565sec		21.049sec		1157.4485sec
Q7	21.390sec		1076.36sec		19.693sec		899.300sec		20.5415sec		987.830sec
Q8	18.600sec		793.296sec		17.896sec		678.547sec		18.248sec		735.9215sec

**4.6. Analysis**

The result of query Q3, shown in Figure 6, display's that in Tamil Nadu, Agriculture sector has the count of 192#, followed by Horticulture sector has a count of 180 and finally, Animal husbandry sector which has the count of 5.

The analysis shown in Figure 12, presents that Agriculture is the most practiced sector followed by Horticulture and Animal Husbandry.

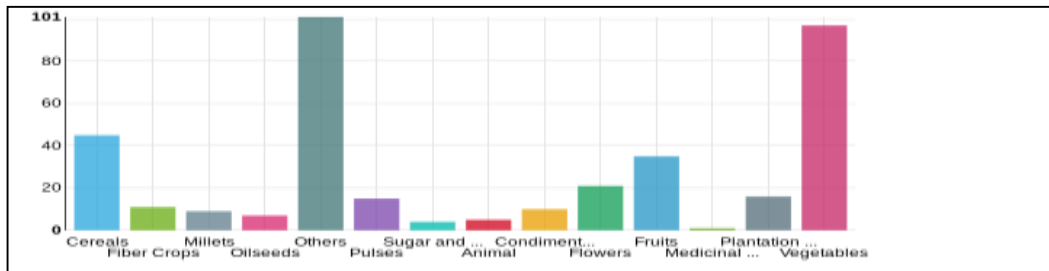


**Figure 12. Analysis of result of Query Q3**

The result of Query Q4, shown in Figure 7, display's that in Agriculture sector, Other category has a count of 101; Pulses category has a count of 15; Fibre crops has a count of 11; Millets category has a count of 9; Oilseeds category has a count of 7 and so on. Similarly, in the Horticulture sector, the Vegetable category has a count of 98; Fruits category has a count of 35; Flower category has a count of 21; Condiments and spices category has a count of 10. In Animal Husbandry sector, issues related to animal category (like bovine, Piggery, Poultry etc.) have to a count of 13.

The analysis shown in Figure 13, presents that other category like weather information, market price information of various crops are the most common issues in the agriculture sector. In the Horticulture sector, the Vegetable category has the most frequent issues of

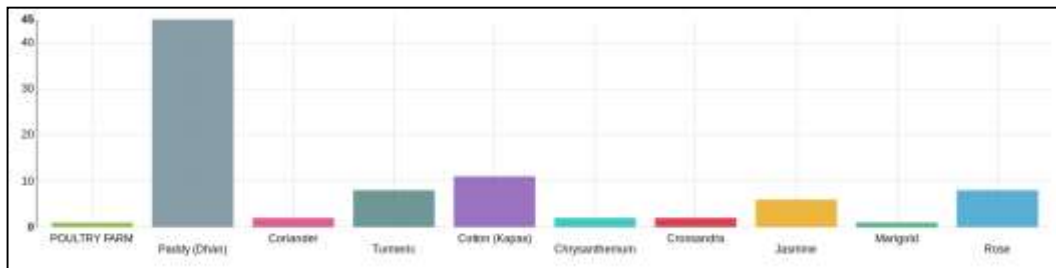
query followed by fruit category, flower category, condiment and spices category. In Animal Husbandry sector, issues related to an animal (like buffalo, cow, goat and chicken) are most common.



**Figure 13. Analysis of Result of Query Q4**

The result of Query Q5, shown in Figure 8, display's the number of crop problem of a specific type in particular category of Tamil Nadu state. In Cereal category, Paddy(dhan) crop has count of 45; In Pulse category, Black gram has the count of 36; In Fiber category, Cotton (Kapas) has count of 43; In Millet category, Jawar has count of 13; In Vegetable category Beans has count of 32; In Fruit category Mango has the count; In Flowers category Rose and Chrysanthemum has count of 23 and 15 respectively; In Spices category Coriander has the count of 10; In Animal category Poultry has count of 10, Fisheries have to count of 20 and Bovine has count of 40.

The analysis shown in Figure 14, presents that in Cereals category, Paddy (dhan) is most prevalent crop in Tamil Nadu district; Pulses category, Black gram (urad) is most sown; Fiber category cotton (kappas) plantation is most common; Millets category, Jawar is mostly sown. In Vegetable category, Beans crop plantation issues are most common; Fruit category, Mango crop is most common; Flowers category Rose, Chrysanthemum are most prevalent; Spices category Coriander is mostly sown. In Animal category Poultry farm, Fisheries and Buffalo related issues are mostly queried.



**Figure 14. Analysis of Result of Query Q5**

The result of Query Q6, shown in Figure 9, display's problem typically associated with crop and the query related to problem type. In Tamil Nadu query associated with plant protection, weed management, field preparation, varieties of seed for the crop, nutrient management, fertilizer use and availability, weather information, market information and sowing time are most common for most of the crops. Query related to plant protection asked about pest management, weed management; field preparation asked about size of farm and the quantity of seed required per hectare; Nutrient management asked about specific nutrient required by crop; seed information asked about seed varieties and suitability based on environmental condition; weather information during the sowing period; market price of specific crop etc. queries are asked by farmer related to agriculture field. Market information asked by a farmer about pricing of the cost of various pulses and vegetables in the market will help to determine which crop to sow in that season. In

the Horticulture sector query about nutrient management, when multiple crops are sown in the same farm. Government schemes, policies or any financial cover available to various crops to be grown in that season and area of that particular state. In Animal Husbandry sector query related to animal production scheme and how to increase the production of dairy products. In Fisheries, query about, how to set up fish pond, area required, costs involved and benefits in fish production. Fisheries most common problems are related to Fish Fingerling Production.

The result of Query Q7, shown in Figure 10, shows the solution given by Kisan Call Center to problem asked by farmer related to specific crop in particular district. The solution to plant protection recommended specific pesticides and weedicides to prevent diseases in plant; Nutrient management recommended specific nutrient required for crop development; Fertilizer management query recommended fertilizer to increase the production of crop; Seed variety related queries recommended solution of high productivity seed and their availability in market; weather information, information about the weather in particular district is told to farmer; market information about specific crop is detailed in market information related query. In the Horticulture sector query about vegetable seed availability, sowing time, nutrient management and multiple vegetable growing techniques at the same time in same field and pesticides to manage pest are answered by KCC operator.

The analysis helps to compare the solution given to various farmer of a different state to the same problem, this will in drawing a conclusive solution to the problem.

The result of Query Q8, shown in Figure 11, shows the mapping of a problem with a season which is affecting crops most in a particular state. This will help to increase the production and supply of pesticides, seeds, nutrients and fertilizers to solve the problem of the farmer.

## 5. Conclusion

India has showcased an impressive growth in the field of information and technology over the past several years. This technology was utilized and put forth in the form of Kisan Call Center by the Government of India to provide the farmers with such a service that would help them to reach the experts to solve their farm-related problems which would require them to travel to concerned offices. Therefore the service has proved ready reckoner for farmers to face the field level problem. Kisan Call Center is such a service that can solve almost all problems faced by the farmers with regard to queries in agriculture.

Kisan Center data can create wonders in the field of agriculture if data is studied processed and analyzed in a correct way. This research focused on how Kisan call center data can be effectively utilized to solve the problem of farmer related to agriculture field. This will help them to get updated to new farm practices, fertilizers, varieties of new seeds available. Kisan Call Center data is increasing day by day; thousands of queries are generated every day. There is a need for a suitable platform to study query in a systematic way. Hadoop and Hive are currently the best platforms to satisfy the need of Kisan Call Center. Query studied in the research will help to study, process and analyze the Kisan Call Center data. The HQL applied in the Hive table stored in HDFS produced effective and refined results which helped to study, process and analyze the result easily.

## 6. Future Scope

Prediction analysis is a key technique that will extend the Kisan Call Center data analysis. One challenge is to get access to a very large database, which is scalable to Gigabytes, to accommodate farmer's data. Hive stores data in a distributed system, easy to retrieve with simple SQL queries. Hive is a first step towards constructing an open source warehouse over a MapReduce data processing system (Hadoop). The distinct



characteristics of the underlying storage and execution engines can be utilized efficiently by writing effective Hive queries. This research work showcases the current capabilities of Hive and its integration with reporting applications to generate meaningful output from the data. There are many important avenues for future work:

Hive has Command Line Interface that accepts queries and executes them utilizing MapReduce. Hive also offers HWI (Hive Web Interface) to use the web-based schema browser. Hive allows writing our own client by allowing interaction with the database through Apache thrift server. The client can be designed in PHP or jQuery. By enhancing the thrift server connectivity with PHP client, farmer's information can be collected directly from farmers over web applications.

The text file format for storing data is used in our research work. Since the size of Kisan Call Center Data is growing day by day ORC file format can be used in future to store data in the more compressed form.

Cloudera supports Spark as the default data execution engine for analytic workloads. Apache Spark can be used as an open platform for flexible in-memory data processing that enables batch, real-time, and advanced analytics on the Apache Hadoop platform. Spark will help in real-time analytics of Kisan Call Center Data. Real-time solution to the problem of farmers can be predicted.

## References

- [1] 3pillarglobal.com, "How to Analyze Big Data with Hadoop Technologies [Online]", Available: <http://www.3pillarglobal.com/> and <http://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies>, (2018) April 11.
- [2] Er. Rupinder Kaur, Raghu Garg, Dr Himanshu Aggarwal, "Big Data Analytics Framework to Identify Crop Disease and Recommendation a Solution", IEEE, International Conference on Inventive Computation Technologies (ICICT), vol. 2, (2016).
- [3] Haritha Chennamsetty, Suresh Chalasani, Derek Riley, "Predictive Analytics on Electronic Health Records (EHRs) using Hadoop and Hive", IEEE, International Conference on Electrical, Computer and Communication Technologies (ICECCT), (2015).
- [4] Abdeltawab M. Hendawi, Fatemah Alali, Xiaoyu Wang, Yunfei Guan, Tianshu Zhou, Xiao Liu, Nada Basit, John A. Stankovic, "Hobbits: Hadoop and Hive Based Internet Traffic Analysis", IEEE, International Conference on Big Data (Big Data), (2016).
- [5] J. Gantz and D. Reinsel, "Extracting value from chaos", in Proc. IDC iView, (2012), pp. 1–12.
- [6] J. Manyika et al, "Big data: The Next Frontier for Innovation Competition, and Productivity", San Francisco, CA, USA: McKinsey Global Institute, (2011), pp. 1–37.
- [7] M. Cooper and P. Mell (2012), "Tackling Big Data [Online]", Available: [http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm\\_june2012\\_cooper\\_mell.pdf](http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf), (2018) May 13.
- [8] G. Blackett, "Analytics Network-O.R. Analytics [Online]", Available: [http://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork\\_analytics.aspx](http://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_analytics.aspx), (2018) May 13.
- [9] Agriculture data, "Kisan Call Center Data, India [Online]", [Dataset] Available: <https://data.gov.in/node/4643341>, (2018) February 14.

## Authors



**Mayank Tripathi**, He is currently pursuing M. Tech. in Computer Science and Engineering from Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India. He received B. Tech. degree in Computer Science and Engineering from Dr A.I.T.H., Awadhपुरi, Kanpur, Uttar Pradesh, India in 2016. His research interests include the study of Big Data technologies and implement it to analyze data of social relevance. The author has published a paper in Global Scientific Journals. The author has also written a paper which is under a review process in the Journal “Big Data Research, Elsevier.”



Abhay Kumar Agarwal is currently working as Assistant Professor in Computer Science & Engg. Department at Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India. He received his PhD in 2017 from Dr APJAKTU Uttar Pradesh, India. He Received his M. Tech degree in 2006 from Samrat Ashok. Technological Institute (SATI), Vidisa and B.Tech degree in 1999 from BIET Jhansi, Uttar Pradesh, India. His research interests include the study of parallel computing, data mining, data warehouses and Big Data analytics.

The author has published about 30 papers International/National Journals.