A Review on Big Data Analytics in the field of Agriculture

Harish Kumar M
Department. of Computer Science and Engineering
Adhiyamaan College of Engineering,Hosur, Tamilnadu, India

Dr. T Menakadevi
Dept. of Electronics and Communication Engineering
Adhiyamaan College of Engineering,Hosur, Tamilnadu, India

Abstract- Big Data Analytics is a Data-Driven technology useful in generating significant productivity improvement in various industries by collecting, storing, managing, processing and analyzing various kinds of structured and unstructured data. The role of big data in Agriculture provides an opportunity to increase economic gain of the farmers by undergoing digital revolution in this aspect we examine through precision agriculture schemas equipped in many countries. This paper reviews the applications of big data to support agriculture. In addition it attempts to identify the tools that support the implementation of big data applications for agriculture services. The review reveals that several opportunities are available for utilizing big data in agriculture; however, there are still many issues and challenges to be addressed to achieve better utilization of this technology.

Keywords—Agriculture, Big data Analytics, Hadoop, HDFS, Farmers

I. INTRODUCTION

The technologies employed are exciting, involve analysis of mind-numbing amounts of data and require fundamental rethinking as to what constitutes data. Big data is a collecting raw data which undergoes various phases like Classification, Processing and organizing into meaningful information. Raw information cannot be consumed directly for any form of analysis. It's a process of examining uncover patterns, finding unknown correlation and finding useful information which are adopted for decision making analysis. Big data supports the public and private sectors in providing the discovered knowledge patterns which are later used for future predictions. It creates an opportunity to extract and process the valuable data which is a valuable asset of an organization. Data analytics and storage tools enables to refine the data for future prediction and decision making analysis. This will create an economic or societal value in the society.

II. WHY BIG DATA ANALYTICS

Big data analytics are important to the core of various applications since data is the raw material which is fed as the input for processing. In major it supports Business Intelligence by allowing the business to grow rapidly and provide better solution at the right time by providing the right format of data for this we need to deal with the large datasets .Text mining, data mining and statistical analysis are some practices were big data are in use.

Parallel processing frame work are adapted by Big data analytics developed applications are deployed in the cloud environment.

III. VALUE OF DATA

The amount of data is growing exponentially. Estimates suggest that at least 2.5 quintillion bytes of data is produced every day. Table: 1 provides the live statistics of data processed in a second [1].

Table 1. Live Statistics of Data in A Second.

Live Statistics of data processed in a second	
749	Instagram uploads
1,173	Tumblr posts
2,328	Skype calls
7,393	Tweets in twitter
56,992	Google searches
1,34,394	You tube videos
2,535,672	Emails sent

III. CHARACTERISTICS OF DATA

As a result of Digitalization organizations are seeking a new approach to enhance their business aspects and implement various technologies for their growth.

A. Volume:

The quantity of data generated as Big Data ranges from Terabytes to Exabytes and Zettabytes of data. The volume has been increasing exponentially: up to 2.5 Exabyte of data is already generated and stored every day. This is expected to double by the end of 2015 [2].

B. Velocity:

Big data is growing rapidly, generating a bizarre of quantities needed to be stored, transmitted, and processed quickly. It refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development [3].

C. Variety:

This refers to the inconsistency which can be shown by data at times. In Big data, the variety and heterogeneity of data sources and storage has increased, fuelled by the use of cloud, web & online computing [4], various formats, types and structure.

D. Veracity:

Big Data Veracity refers to the biases, noise and abnormality in data. Accuracy of analysis depends on the veracity of the source data. In comparison to Big Data's volume and velocity, veracity is the most challenging characteristic in data analysis [2].

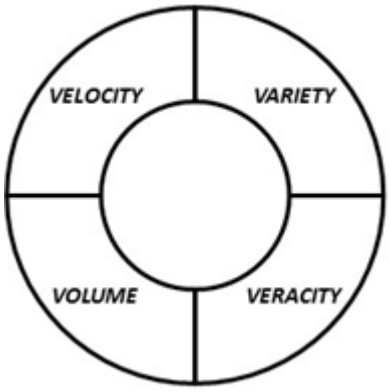

Figure 1.Characteristics of Big Data-4V's

IV. AGRICULTURE IN INDIA

Agriculture plays a vital role in India, At present India is the 2nd largest agricultural land in the world with 157.35 million hectares certainly with 20 Agri-climatic regions, all 15 major climates in the world exist in India and also possess 40 to 60 soil types in the world. According to the survey total food grain production in India reached an all-time high of 252.68 million tonnes FY15.As on August 17, 2015 Rice and Wheat production in the country stood at 104.84 and 88.94 Million tonnes respectively. In addition of these various Farm equipment's like tractors, harvesters and tillers India holds largest manufacturers [5].

A. ROBUST DEMAND.

Population growth is the key factor for the demand of agricultural and its products. Rising urban and ruler incomes also have an impact on the growth of agriculture needs this also aided a demand on growth.

B. INDIA'S GROWTH IN AGRICULTURE.

1. GDP of Agriculture and allied sectors in India has recorded at USD 259.23 billion in FY15.
2. According to the advanced estimates of Central Statistical organization agriculture and allied sector recorded a growth of 8.5% in FY15.Figure 2 projects the Gross Domestic product survey from the period of 2007 to 2015 [6].
3. As a whole Agriculture is the primary source of livelihood for about 58% of India's population.

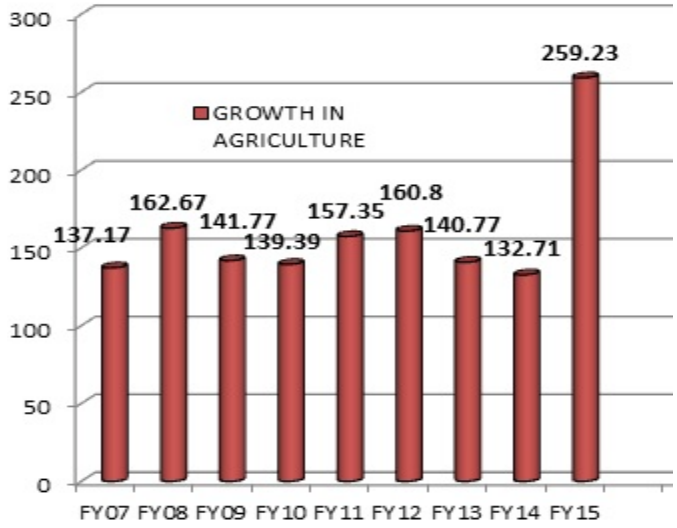

Figure 2.GDP by value added-size of agriculture from 2007 to 2015 (in USD billion).

C. TWO MAJOR AGRICULTURAL SEASONS IN INDIA.

Kharif and Rabi are two majors crops cultivated in India. Kharif season lasts from April to September (summer), Rice (Paddy), is the seasons main crop, Rabi season lasts from October to March (winter) Wheat is the seasons main crop. According to the 4th advance estimates for FY15 by the "Ministry of Agriculture", the total Production is estimated to be around 252.88 million tonnes [7].

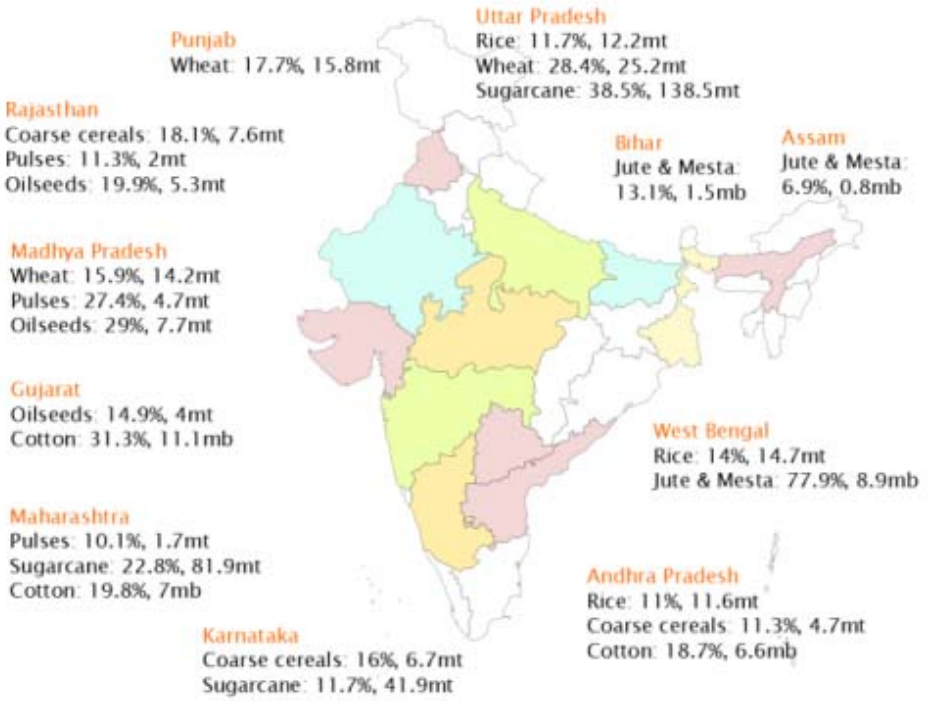

Figure 3.States production of major crops in India.

Table 2 Areas Production and yield of Major Crops.

Crops	Area (Lakh Hectare)			Production (Million Tonnes)			Yield (kg/hectare)		
	2012 to 2013	2013 to 2014	2014 to 2015	2012 to 2013	2013 to 2014	2014 to 2015	2012 to 2013	2013 to 2014	2014 to 2015
Rice	427.54	441.36	438.56	105.24	106.65	104.80	2461	2416	2390
Wheat	300.03	304.73	309.69	9.51	95.85	88.94	3117	3145	2872
Coarse Cereals	247.57	252.20	241.49	40.04	43.29	41.75	1617	1717	1729
Pulses	232.56	252.13	230.98	18.34	19.25	17.20	789	764	744
Food Grains	1207.7	1250.4	1220.72	257.13	265.04	252.68	2129	2120	2070
Oil seeds	264.84	280.51	257.27	30.94	32.74	26.68	1168	1168	1037
Sugarcane	49.99	49.93	51.44	341.20	352.14	359.83	68254	70522	69857
Cotton	119.77	119.60	130.83	34.22	35.90	35.48	486	510	461
Jute & Mesta	8.63	8.38	808	10.93	11.69	11.45	2281	2512	2550

D. MAJOR CHALLENGES FACED BY INDIAN AGRICULTURE

- Stagnation in Production of major crops
- High cost of farm inputs
- Soil Exhaustion
- Depletion of fresh ground water
- Adverse impact of Global climatic changes
- Impact of Globalization
- Providing Food Security
- Farmer's suicide.

E. FARMING IN PRACTISE

India has diversified topography we are following different kinds of farming methods like

1) Subsistence and Commercial Farming
2) Intensive and Extensive Farming
3) Plantation Farming
4) Mixed Farming

V. PRECISION AGRICULTURE

Precision agriculture is a precise farm management technology it can be implemented in various regions where there is a shortage of rainfall, soil is not fertile, temperature is worse. In precision agriculture technique actual crop inputs which are adopted to the available land, were cultivation is progressed this method is not new ,But the input data's are collected and digitalized in the form of datasets and later it is implemented via various techniques which are precisely suitable for increasing economic growth and yields high production[8].

In this technique it is mandatory to note the condition of soil and crop for certain period of time to adhere the spatial variability , The data's collected are mapped with the management system using Geographical information system and advance sensor equipment's by doing these it will maximize the sustainable productivity of the crop with more profit.

A. Why Precision Agriculture?

Due to globalization we are unaware to predict future climatic conditions, so farmers are looking for new techniques to increase the efficiency and reduce the cost [9], so Precision farming method would be an alternate method to utilize new techniques in this modern era.

Techniques behind Precision Farming:

1) Remote sensing
2) Satellite Navigation system
3) Geographical Information system
4) Automatic yield recording system
5) Automatic soil sensor
6) Variable rate technology
7) Advanced Farm Management

Precision farming requires data collection, analysis and processing the information gathered. In addition to these PF requires some important aspects like Mapping. Remote sensing, Geographical Information system, Investigation during the field operation

B. Mapping

Soil is the main asset of farming, mapping will measure and control spatial variability .It is need to collect the details of crop like sampling of the crops like nutrients, strength of the soil and pH of water before and after production so that it is used for future prediction this mapping process can be done by utilizing Remote sensing, Satellite Navigation system and Geographical Information system which is recorded during field operation.

C. Remote Sensing

We make use of remote sensing to monitor the visible and invisible areas of land cultivated, it will convert these data into spatial information which is rather and is sent to GIS the generated images will allow mapping. In addition to these it will supply variability management through decision support system.

D. Geographical Information system

GIS is used to integrate spatial data collected from various sources it is used to develop decision environment and makes weed control, pest control, and fertilizer application rather than it gives information regarding drought, yield information and weather forecasting as a whole it is integrated with GPS for location tracking.

E. Investigation

Manual field operations are required to monitor the quality and quantity of the crop yielded, manual sample test are done for soil to know its fertility in order to measure more accurate reliable sensors and advanced GPS devices are required, but practically farmers cannot afford such an expensive equipment's.

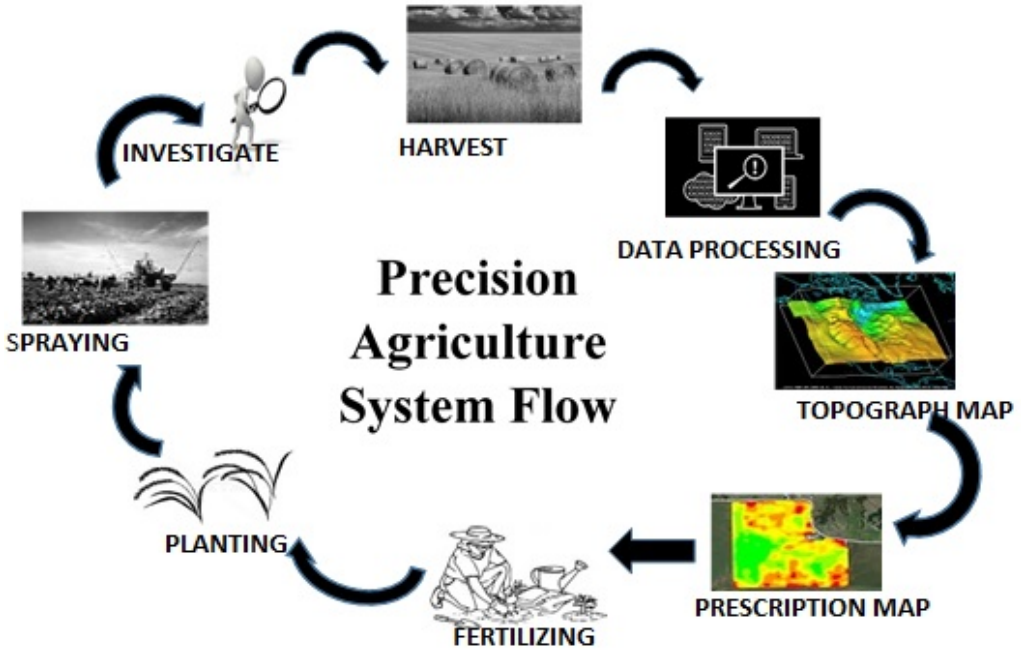

Figure 4 Precision Agriculture System flow.

VI. HADOOP

Hadoop stores and processes massive volume of data. Hadoop is an open-source framework and uses commodity hardware to store colossal quantity of data, which is based on distributed computing model.
Hadoop protects data and executing applications against hardware failure .If a node fails ,it will automatically redirects the job that has been assigned to this node to the other functional node which is available. It stores multiple copies of data on different clusters. Since RDBMS is not suited for Process large files and Data sets we need to use Hadoop Frame work. The table projects the difference between RDBMS and Hadoop.

Table 3 RDBMS Vs Hadoop.

PARAMETERS	RDBMS	HADOOP
System	Relational Database Management System	Node based Flat Structure
Data	Suitable for Structured Data	Suitable for structured, unstructured and supports files of all formats
Processing	OLTP	Analytical Big data Processing
Processor	Requires High end-Expensive Hardware	In Hadoop Cluster a node requires only a processor, a network card and hard drive
Cost	Around $10,000 to $14,000 per Terabytes of storage	Cost around $4,000 per Terabyte of storage

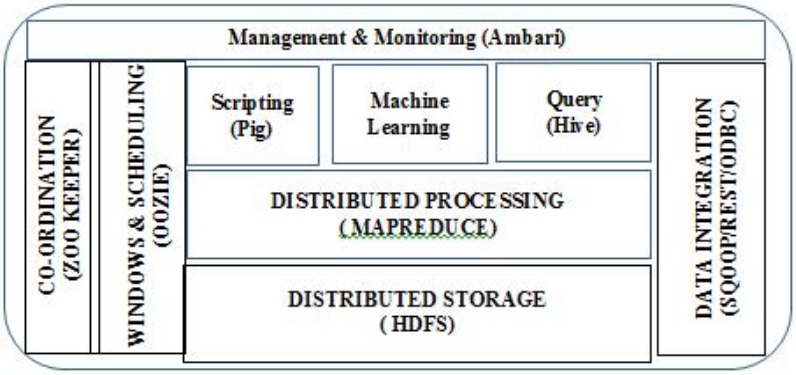

Figure 5.Hadoop Components

Hadoop Framework Description: The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models and it is a part of Apache project sponsored by Apache Software Foundation. Doug Cutting is the creator of Hadoop named the framework after his child`s stuffed toy elephant. Hadoop is used by many companies like Google, IBM and Yahoo.

For our proposed Work Apache Hadoop Framework will be used to implement. This paper reports the core elements of Hadoop and its usage. Fig.5 represents arrangements of components in Hadoop Eco System [2].
Hadoop [3] [4] is a framework which has two main components.

1) Hadoop Distributed File System (HDFS)
2) Hadoop MapReduce

HDFS

Distributed File System handles large files and performs Read/Write operation sequentially each large file is broken into frames and stored across multiple data nodes. In HDFS Environment "NameNode" will keep track of overall file directory structure and it monitors the frames .This NameNode acts as a central point if it is required is distributed into many copies. DataNodes reports all frames to the NameNode at boot up. Each frame will have its unique version number and it's not dominant after its updating.

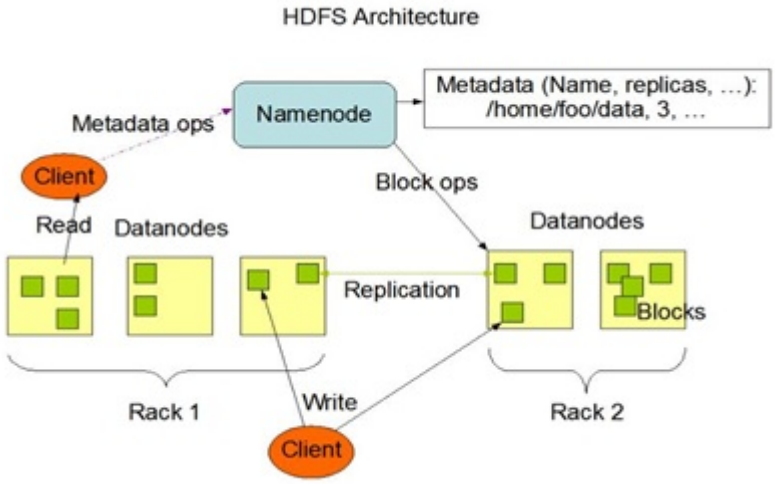

Figure 6 HDFS Architecture.

To read a file the client API will calculate the frame index of the file pointer and it sends a request to the NameNode .As a result NameNode will reply which DataNode has a replica of that frame. To write a file, Client API will communicate the NameNode who is responsible for granting a replica which acts as primary. The client updates its changes to all DataNodes, but this updates are stored in a buffer of individual DataNode .After all the changes in DataNodes, then client sends a "Commit" request to the primary which determines an order to update all secondary Nodes. Once secondary Nodes complete the commit an operation then primary will send a response to the client about its success. All updations of frame distribution and metadata will be written to an operation log file at NameNode. The objective of this log file is to maintain the order list of operation which is important for the NameNode to recover its view after crash. If NameNode is crashed a new NameNode will be restoring to start from the save point file and replay the log operation.

MAPREDUCE:

The Execution of a Job starts when the client program submits a job configuration to the Job Tracker which specifies the map, combine and reduce function.

The role of JobTracker is to determine the number of frames are separated from the input path and select some TaskTracker based and their network, then JobTracker will send the task request to those TaskTracker which are selected. Now TaskTracker will start the Map processing by extracting the input data from the splits. Each record parsed by the "Input Format", it invokes the user provided " Map" function, which emits a number of key pair in the memory the "Combine" function. The key pairs are sorted into one of the R local file maintain the order list of operation which is important for the NameNode to recover its view after crash.

Once the Map task completes the TaskTracker will notify the JobTracker when all the TaskTracker are done, then JobTracker will notify the selected TaskTracker for the reduce phase. Each TaskTracker will read the region files remotely. Map/Reduce framework rebounds to crash of any components. JobTracker keeps tracks of the progress of each phase and periodically pings the status of TaskTracker. When any one of the Map phase and TaskTracker crashes the JobTracker will reassign the Map task to some other TaskTracker node. Similarly if reduce phase crashes the JobTracker will return the reduce phase to different TaskTracker. If both Map and Reduce phase completes, the JobTracker will unblock the Client Program.

Hadoop Ecosystem is categorized based on the following parameters:

1. Filesystem
2. Programming models
3. Databases
4. Data Ingestion
5. Security
6. Scheduling & DR
7. Machine Learning
8. Bench Marking & QA Tools
9. Meta Data Management
10. System Deployment
11. Development Framework

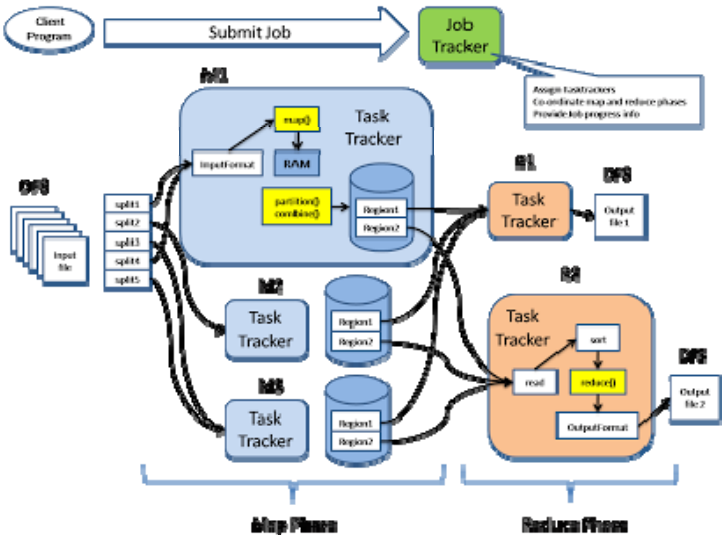

Figure 7.MapReduce Phase.

Table 3 shows the list of tools available in Hadoop Ecosystem.

<table>
<tr><th colspan="2">DISTRIBUTED</th><th colspan="2">NOSQL DATABASE</th></tr>
<tr><th>FILE SYSTEM</th><th>PROGRAMMING</th><th>COLUMN-DATA MODEL</th><th>DOCUMENT-DATA MODEL</th></tr>
<tr><td>-Apache HDFS
-RedhatGlusterFS
-Quantcast File System
-Ceph Filesystem
-Lustre File System
-Alluxio
-GridGain
-XtreemFS</td><td rowspan="3">-Apache Ignite
-Apache MapReduce
-Apache Pig
-Pachyderm Map
-Apache Beam
-Apache Storm
-Apache Flink
-Apache Apex
-Netflix PigPen
-Amplab SMR
-Facebook Corona
-Apache REEF
-Apache Twill
-Damballa Parkour
-Apache Hama
-DatasaltPangool
-Apache DataFu
-Pydoop
-Kangaroo
-Tinker Pop</td><td>-Apache HBase
-Apache Cassandra
-Hyper table
-Apache Accumulo
-Apache Kudu
-Apache Parquet</td><td>-MongoDB
-RethinkDB
-ArangoDB</td></tr>
<tr><th>DATA INGESTION</th><th>STREAM-DATA MODEL</th><th>KEY VALUED-DATA MODEL</th></tr>
<tr><td>-Apache Flume
-Apache Sqoop
-Apache Chukwa
-Facebook Scribe
-Apache Kafka
-Netflix Suro
-Apache Samza
-Cloudera Morphline
-HIHO
-Apache NiFi
-Apache Manifold CF</td><td>Evenstore</td><td>-Redis Database
-Linkedin Voldemort
-RocksDB
-OpenTSDB</td></tr>
<tr><th colspan="2">SERVICE PROGRAMMING</th><th colspan="2">GRAPH-DATA MODEL</th></tr>
<tr><td colspan="2">-Apache Thrift
-Apache ZooKeeper
-Apache Avro
-Apache Curator
-Apache Karaf
-Twin Elephant Bird</td><td colspan="2">-ArangoDB
-Neo4j
-TitanDB</td></tr>
<tr><th colspan="2">SECURITY</th><th colspan="2">NEW SQL DATABASES</th></tr>
<tr><td colspan="2">-Apache Sentry
-Apache Knox Gateway
-Apache Ranger</td><td colspan="2">-TokuDB
-Handler Socket
-Akiban Server
-Drizzle
-SenseiDB
-Sky
-Bages DB
-Influx DB
-Haeinsa</td></tr>
<tr><th colspan="2">MACHINE LEARNING</th><th colspan="2">SQL ON HADOOP</th></tr>
<tr><td colspan="2">-Apache Mahout
-WEKA
-Cloudera Orynx
-Deep learning 4j
-MADLib
-H2O
-Sparkling Water
-Apache SystemML</td><td colspan="2">-Apache Hive
-Apache HCatalog
-Apache Trafodion
-Apache HAWQ
-Apache Drill
-Cloudera Impala
-DatasaltSplout SQL
-Apache Tajo
-Apache Pheonix
-Apache MRQL</td></tr>
<tr><th colspan="2">SYSTEM DEPLOYMENT</th><th colspan="2">SCHEDULING & DR</th></tr>
<tr><td colspan="2">-Apache Ambari
-Cloudera HUE
-Apache Mesos
-Myriad
-Marathon
-Apache Helix
-Apache Bigtop
-Buildoop
-Deploop
-Apache Eagle</td><td colspan="2">-Apache Oozie
-Linkedin Azkaban
-Apache Fakon
-Schedoscope</td></tr>
<tr><th colspan="2"></th><th colspan="2">BENCHMARKING & QA TOOLS</th></tr>
<tr><td colspan="2"></td><td colspan="2">-Apache Hadoop Benchmarking
-Yahoo GridMix 3
-PUMA Benchmarking
-Intel HiBnch
-Apache Yetus
-Berkeley Swim Benchmark</td></tr>
<tr><th colspan="2"></th><th colspan="2">DEVELOPMENT FRAMEWORK</th></tr>
<tr><td colspan="2"></td><td colspan="2">-Jumbume
-Spring XD
-Cask Data Application Platform</td></tr>
<tr><th colspan="2"></th><th colspan="2">METADATA MANAGEMENT</th></tr>
<tr><td colspan="2"></td><td colspan="2">-Metascope</td></tr>
</table>

VII. CONCLUSION.

The role of Bigdata analytics in the field of agriculture are explored .Agriculture will face affordable challenges to provide sufficient nutrients. We have reviewed latest tools and technology which will afford to increase the productivity of crops. In this paper we have discussed about Agriculture statistics of India and its traditional farming methods. In addition Bigdata paves a major role by introducing Precision Agriculture techniques which is already initiated in many developing countries which makes the farmers to integrate traditional farming methods. Detailed architecture of Hadoop Ecosystem and its components has been listed along with their framework; these tools are used to predict future analysis by using the collected Datasets .In future direction of Agriculture will provide technical backup support to the farmers to implement innovative models which are replicated in large scale. We plan to work on precision agriculture techniques.

REFERENCES

[1] www.internetlivestars.com
[2] Dilpreet Singh and Chandan K Reddy, 2014. "A survey on platforms for big data analytics" Journal of Big Data,1:1,8.
[3] Tom White,2015."Hadoop The Definite Guide", O'Reilley Publications, 4th Edition April2015.
[4] htttp://en.wikipedia.org/wiki/Big Data.
[5] http://www.ibef.org/industry/agriculture-india.aspx.
[6] http://www.agricrop.nic.in.
[7] http://www.makanaka.wordpress.com/tas/foodgrains.
[8] Rupika yadhav, Jhalak Rathod, Vaishnavi Nair, "Big Data meets Small sensors in precision Agriculture", International Journal of computer Applications(0975-8887)Applications of Computer and Electronics for the Welfare of Rural Masses(ACE WRM)2015.
[9] Tekiner, f: Keare,J.A,"Big Data Frame Fork", systems, Man, and cybernetics (SMC),2013 IEEE International conference on vol., no., pp.1494, 1499,13-16 oct.2013
[10] http://www.hadoopecosystemtable.github.io.