STUDY OF BIG DATA TECHNOLOGY FOR AGRICULTURE SECTOR

AMANDEEP KAUR[1] , RAGHU GARG[2], HIMANSHU AGGARWAL[3]

[1] Department of Computer Engineering, Punjabi University, Patiala, India

[2] Research Scholer, Department of Computer Engineering, Punjabi University, Patiala, India

[3] Professor, Department of Computer Engineering, Punjabi University, Patiala, India

Abstract:-The term Big Data came into existence with the exponential rise in growth of data. Besides the conventional database management tools or data processing tools have become inefficient to handle those large and complex sets of data. 'Big data' refers to the data sets that is huge in size(volume), high in velocity and high in variety and generated from various resources and therefore it is extremely difficult to handle with traditional tools and techniques and thereby extract useful information from it. Big data analytics is the process of examining large volumes of data and extracting useful information from it for taking better decisions. In these studies developed big data analytics framework for agriculture applications, Hadoop and its tools that provide big data parallel computing compatible data management infrastructure that offers analytics solution for integration analysis of large, heterogeneous and unstructured datasets. The review also introduces the research challenges face by agriculture data scientist and data sources that can be used for big data analysis of agriculture decision making.

***Keywords:- Big Data, Big Data Analytics, Hadoop and tools,* Agriculture Big Data Sources, Agriculture Big Data Challenges.**

1. INTRODUCTION

In the 20th century the concept of relational database came into existence. Relational database was revolutionary step in the world of data where the data is stored in the form of relations or tables (rows, columns) and can be easily processed according to users need. Structured data can be easily stored into relational databases or tables. But the real problem comes into picture due to advancement of technology and the fast use of internet we got data having huge volume, high velocity, wide variety [5]. Moreover, nowadays 80% of the data is in unstructured form and comes from various resources and this data is difficult to maintain in form of table or relations.

'Big data' refers to huge volume of data that cannot be stored and processed by using traditional systems within given timeframe. 'Big data' is generated from various resources such as social networking sites like facebook, twitter, youtube, instagram, LinkedIn, Sensors etc and this data is being generated in various formats. Big data analytics refers to the process of collecting, organizing, large data set to discover different patterns and other useful information. The main goal of the big data analytics is to help organization to make better business decision. Examples of big data analytics are big online business websites like LinkedIn, Twitter, Filpkart, Snapdeal uses Facebook and Gmail data to view the customer information or behavior. Analyzing big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable.

The basic goal of the paper is to introduce the importance of big data technology in agriculture sector. In this paper discussion is related to the analytics application in agriculture domain and big data tools like RHadoop, Mahout, Hive, HBase, MapReduce, HDFS, Flume, Ooize etc that provide big data parallel computing compatible data management infrastructure and machine learning algorithms that offer analytics solution for integration analysis of large, heterogeneous and unstructured datasets on the big data. Our survey highlights the applications of big data analytics in agriculture domain. The review also brings up the research challenges, data sources that can help in agriculture big data analytics.

2. RESEARCH CHALLENGES

Data cleaning:-Data cleaning also known as data scrubbing is the process of fixing or removing inaccurate or corrupt data and removing inconsistencies from data in order to improve the quality of data [22]. It is considered as a

main challenge in the era of big data, due to the increasing volume, velocity and variety of data in many different applications. Moreover the data comes from various sources like social media sites such as facebook, twitter, LinkedIn produces data in various formats like structured, semi-structured and unstructured formats and it is very difficult to get insight [19]. Therefore a new data cleaning techniques are required to handle unstructured and semi-structured big data in efficient way.

Data protection:-Protecting big data from unauthorized access is a significant challenge for researchers. If the big data storage system will compromise, then it can be detrimental effect on commercial secrets, personal information can be disclosed. A lot of work is done in order to protect the privacy of users from data generation to data protection, but still there exist many challenges such as access control and secure end to end communication, decentralized storage etc.[10]

Data analytics:-The data generated by various heterogeneous sources are not just huge but composed of various data formats and even including streaming data [21]. So handling a massive volume of data in a limited time is a significant challenge for researchers. For handling complex data, the first big data tool which comes to mind is Apache Hadoop but it has still many challenges while handling complex data such as data management, cluster management, job scheduling, resource sharing. Furthermore, incomplete data and noisy data may also affect the performance of the data analysis [12].

Data visualization:-Data visualization is a term that is used to represent knowledge in more effective way by using different graphs such as pie charts etc. The traditional way of presenting big data have few limitations because traditional tools and techniques are inadequate to handle scalable data. In order to visualize the required information from a pool of random data, powerful algorithms are significant for accurate results. So it is challenging task to handle big data visualization limitations such as real-time scalability, perceptual scalability and interactive scalability [1].

3. BIG DATA INFRASTRUCTURE FRAMEWORK FOR AGRICULTURE SECTOR

The big data infrastructure framework for agriculture science is little bit complex as compared to the traditional analytics system as shown in Figure 2. Difference is big data technology refers to three aspects of technical innovation that cope with super-large datasets, automated parallel computation processing methods and data management schemes. In the present scenario agriculture analytics is performed with business intelligence tools like R, KNIME installed on single computer machine, but in era of big data analytics single machine can't analyze data as it will take polynomial time. Solutions that is provided is big data analytics. In big data analytics large processing is divided and executer on multiple nodes, then the combined results on are kept on the master node. For performing this task big data infrastructure for agriculture science will required.

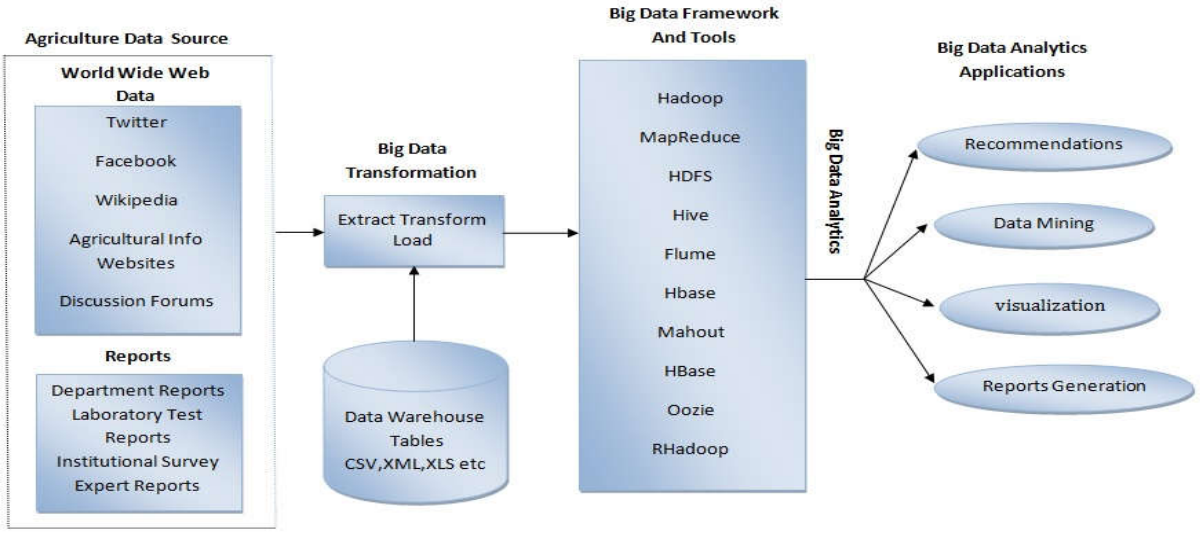

Figure 2 Big Data Infrastructure for Agriculture Sector

Agriculture big data is data which is related to agriculture production in all aspects. Agriculture data source farmer field data, social media, agricultural information websites, laboratories reports and institutional data. These data source provides statistics and real time data related to agricultural economic entities, expert discussions, investment information, laboratories recommendations and field status (field sensing device). Some agriculture data sources are discussed as below.

Web Data

Now a days social media is major source of communication and share experiences and thoughts in field of agriculture. Social network media and web services provides access to the historic data sets and also real time data access to source, possible with a 15 minutes time delay, as with Thomson Reuters and Bloomberg financial data [29]. News data access to the historic data and the real time news data sets possibly through the concepts of educational data licensing. Public data access to the scraped and archived public data of importance which is made available through the RSS feed blogs or the open government databases [30]. Researchers need access to simple programming interfaces meant for applications in order to scrape and store other available data source that cannot be automatically collected by them. Examples for social media data is Twitter, Facebook, Youtube, blogs and RSS feeds.

Farmer Field Data

Field sensing technology is a technology that is distributed globally. In field sensing technology the machine to machine communications generates large amounts of data that stored are cloud computing services. (e.g. soil moisture detection sensor, weather sensor, bio sensor etc.)()

Laboratory test reports

The laboratory test reports are very important source of data for researchers and the various types of test performed are as: Soil testing, Water testing, Plant analysis, Manure testing, Compost testing, Biosolids testing, Green roof media testing, Green house media testing etc.

Institutional Data

Institutes are responsible for conducting agricultural surveys all over world. The agricultural census may be conducted in different countries in various ways depending upon the resources. Food and agriculture organization of united nations and world bank provides these type of data**.** An agricultural census should be an essential part of an integrated system of agricultural statistics with the focus to provide primary data on the structure of the agriculture sector, such as the size of holdings, land use, land tenure etc., which do not change too quickly . Detailed data on the agricultural productions and inputs are important part of the system of agriculture statistics called the current statistics and are collected through the specialized agricultural surveys and other available sources.

Above collected data contain large number of attribute and many of these attributes are irrelevant or redundant. Their inclusion may take the learning process unstable solution of this problem is dimension reduction. Dimension reduction including feature selection and feature extraction. Feature selection is the process of selecting the subset of relevant or important features. Feature selection may take place at the data preprocessing or model learning step. When the number of features is too high, correlation analysis is often used to preselect or to screen features prior to model building [23]. Feature extraction is used to create new features by the transformation or function of raw features. One popular feature extraction procedure is principal component analysis (PCA)[28], which extracts a small set of directions to represent the data and achieves great dimension reduction.

Second problem is sometimes the data is missing or is incomplete. In machine learning process missing or incomplete data disturbs the process of decision making. Solutions of missing data problem is to delete incomplete data the reason is simple that it is not considered appropriate because ‘missing’ may itself be important information [26-27]. Mechanisms for missing data, including missing completely at random (MCAR) [25], missing at random (MAR) and not missing at random (NMAR).Solutions of incomplete data is replacing the missing data with substitute values [24].

In next component of framework take decisions regarding input, distributed design, tool selection and analytics models. Final show applications of big data analytics in agriculture sector like recommendation system, report generation, visualization and data mining. These applications will use statistics, mathematics, machine learning techniques and technology to aggregate, manipulate, analyze and visualize big data in agriculture. These applications helps farmers improve productivity, climate predicts that may damage agriculture production, control crop diseases, design future plans which will give improve production and reduce cost etc. Out of these applications next section discuss recommendation system to control crop diseases based on machine learning intelligent algorithms. Mostly use open source distributed data process platform Apache Hadoop for big data analytics, Hadoop and its tools discussed in Table 1.

Table 1 Hadoop and Tools

Platform/Tools	Description
Apache Hadoop	Apache Hadoop framework is used for storing data and running applications on a cluster of commodity hardware. 'Hadoop' architecture mainly comprises of two main components: HDFS for storing big data and Map Reduce for Big data analytics [15-16].
Hadoop Distributed File System	HDFS takes care of storing and managing very huge scale of datasets within Hadoop cluster. The Hadoop Distributed File System is designed to run on commodity hardware. It is highly fault-tolerant and provides high throughput access to application data and is suitable for applications that have large data sets. It has become a key tool for managing pool of 'Big data' and supporting 'Big data' for analytics applications [4].
MapReduce	A Programming model that enables the users to process large amounts of structured and unstructured data in parallel batches across the large cluster of machine in reliable and fault tolerant manner [9-10]
Apache Sqoop	It is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL to Hadoop HDFS and export from Hadoop file system to relational databases [2].
Apache Hive	It is designed for facebook. It is used for querying and analyzing of huge volume of datasets that are stored in the HDFS. The Hive is mainly used for data querying, summarization and analysis. It supports queries that are expressed in the language called HiveQL (Hive query language), which is capable of automatically translating SQL-like queries into MapReduce jobs for easy execution and processing of extremely large volumes of data [20].
Apache HBase	Apache HBase is an distributed, Open source, column oriented database that provides Bigtable like capabilities on top of hadoop and HDFS. HBase scales linearly to handle huge amount of data sets with billions of rows and millions of columns and it easily combines data sources that use a wide variety of different structures and schemas. It is good for semi-structured as well as structured data. Companies such as Facebook, Twitter, Yahoo, and Adobe use HBase internally [18]
Apache Mahout	An open source project that is primarily used for creating scalable machine learning algorithms. It implements popular machine learning techniques such as: recommendation, clustering, classification [8]. Companies such as Facebook, LinkedIn, Twitter, and Yahoo use Mahout internally. Yahoo uses Mahout for pattern mining.
Apache Flume	It is a distributed, highly reliable and available service for efficiently collecting, aggregating and moving large amounts of log data [7]. It is basically designed to copy streaming data or log data from various web servers to HDFS. Along with the log files, Apache Flume is also used to import huge volumes of event data produced by social networking sites like Twitter and

	Facebook and e-commerce websites like Filpkart, Amazon [17].
Apache Pig	An open source project that is being used as a platform for analyzing large data sets that consists of high level language. It enables parallel execution of data flows on Hadoop. Pig uses a language for expressing the data flows. Using Pig Latin language, programmers can perform MapReduce tasks easily without having to type complex codes in Java[6]. Apache Pig can handle structured, unstructured, and semi-structured data and stores the results in HDFS.
Apache Spark	Spark was introduced by Apache Software Foundation for speeding up the Hadoop computational. It uses Hadoop in two ways: one is storage and second is processing. Since Spark has its own cluster management computation, it uses Hadoop for storage purpose only. It is designed to handle a wide range of workloads such as batch applications, interactive queries, iterative algorithms, and streaming data [14].

Conclusion

With the advancement in technology and generation of voluminous of agriculture data there is a need for handling and managing the 'Big Data'. This Paper discussed big data analytics framework for agriculture applications, Hadoop and its tools that provide big data parallel computing compatible data management infrastructure that offers analytics solution for integration analysis of large, heterogeneous and unstructured datasets. The review also introduces the research challenges face by agriculture data scientist and data sources that can be used for big data analysis of agriculture decision making.

REFERENCES

1. Agrawal, R., Kadadi, A., Dai, X., & Andres, F. (n.d.). Challenges and Opportunities with Big Data Visualization. *7th International Conference on Management of computational and collEctive intelligence in Digital EcoSystems · MEDES 2015* , 169-173.
2. Aravinth, S. S., Begam, A. S., Shanmugapriyaa, S., & Sowmya, S. (2015). An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing. *International Journal for Innovative Research in Science & Technology* , 252-255.
3. Bhardwaj, A., Vanraj, Kumar, A., Narayan, Y., & Kumar, P. (2015). Big data emerging technologies: A Case Study with analyzing twitter data using apache hive. *2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS)* .
4. Borthakur, D. (2013). HDFS Architecture Guide [online Available at https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html]
5. Das, A. C., Mohanty, S. N., Prasad, A. G., & Swain, A. (2016). A Model for detecting and managing unrecognized data in a Big data framework. *ICEEOT* , 3517-3522.
6. Gates, A. (2010). Programming Pig .O'reilly Media, Inc. California, USA, 1-10
7. Hoffman, S. (2015). Apache Flume: Distributed Log Collection For Hadoop. p. 7-15. Packt Publishing Ltd. Birmingham, UK.
8. Ingersoll, G. (2009). Introduction of Apache Mahout: Scalable, Commercial-Friendly Machine learning for Building Intelligent Applications. *IBM Corporation* .
9. Mandal, B., Sahoo, R. K., & Sethi, S. (2015). Architecture of efficient word processing using Hadoop MapReduce for big data applications. *2015 International Conference on Man and Machine Interfacing (MAMI)* , 1-6.
10. Mehmood, A., Natgunanathan, I., Xiang, Y., Guo, S., & Hua, G. (2016). Protection of big data privacy. *IEEE Access* , 1821-1833.
11. Mishra, G., Masih, S., Tanwani, S., & Bansal, M. (2015). Glister: A Framework for Iterative MapReduce. *IEEE International Conference on Computer, Communication and Control (IC4-2015)* , 1-6.
12. Pal, K. (2015). 5 Challenges in Big Data Analytics to Watch Out For [online Available at https://www.techopedia.com/2/31258/trends/big-data/5-challenges-in-big-data-analytics-to-watch-out-for]

13. Sagiroglu, S., & Sinanc, D. (2013). Big Data: A Review . *Collaboration Technologies and Systems (CTS), 2013 International Conference IEEE* , 42-47.
14. Salloum, S., Dautov, R., Chen, X., Peng, P. X., & Huang, J. Z. (2016). Big data analytics on Apache Spark. *International Journal of Data Science and Analytics* , 145-164.
15. Singh, D., & Reddy, C. K. (2014). A survey on platforms for big data analytics. *Journal of Big Data* , 1-20.
16. Sinha, S. (2014). Hadoop Tutorial: All you need to know about Hadoop! [online Available at https://www.edureka.co/blog/hadoop-tutorial/]
17. Sinha, S. (2017). Apache Flume Tutorial : Twitter Data Streaming [online Available at https://www.edureka.co/blog/apache-flume-tutorial/]
18. Sun, J. (2010). Scalable RDF Store Based On Hbase And Mapreduce. *3rd international conference on Advanced computer theory and engineering.Chengdu, China* , 633-636.
19. Tang, N. (2014). Big Data Cleaning. *Web Technologies and Applications* , 13-24.
20. Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., et al. (2009). Hive - A Warehousing Solution Over a Map-Reduce framework. *ACM Digital library* , 1626-1629.
21. Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a Survey. *Journal of Big data* , 1-32.
22. Yousuf, F. (2015). Data Mining: Challenges in Data Cleaning. [Online Available at http://blog.appliedinformaticsinc.com/data-mining-challenges-in-data-cleaning/]
23. Saeys,. 2007. A Review Of Feature Selection Techniques In Bioinformatics. p. 2507-2517. Volume 23. No 19. Bioinformatics 23. Oxford Journal, USA.
24. Little, R.A. And Rubin. 2002. D.B. Statistical Analysis With Missing Data, John Wiley And Sons. New Jersey, USA.
25. Rja Little. 1988. A Test Of Missing Completely At Random For Multivariate Data With Missing Value. American Statistical Association. Alexandria, USA.
26. Therese D. Pigott. 2001. A Review Of Methods For Missing Data, Educational Research And Evaluation. p. 353-383. Volume 7. No 4. Taylor Francis Online .Abingdon,UK.
27. Paul D. Allison. 2001. Missing Data. A Saga University Paper. New York, USA.
28. Jon Shlens. A Tutorial On Principal Component Analysis Derivation, Discussion And Singular Value Decomposition. 2014. (Available online with update at https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf).
29. Thomson Reuters. 2014. (Available online with update at http://thomsonreuters.com/en/press-releases/2014/thomson-reuters-adds-unique-twitter-and-news-sentiment-analysis-to-thomson-reuters-eikon.html).
30. Bogdan Batrinca andPhilip C. Treleaven. 2015. Social Media Analytics: A Survey of Techniques, Tools And Platforms. p. 89-116. Volume 30. No 1. Journal of Knowledge, Culture and Communication. AI & Society, Springer.