Indian Journal of Science and Technology, *Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/100728, August 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Big Data Analytics Recommendation Solutions for Crop Disease using Hive and Hadoop Platform

Raghu Garg* **and Himanshu Aggarwal**

Department of Computer Engineering, Punjabi University, Patiala - 147002, Punjab, India; raghugarg@hotmail.com, himanshu.pup@gmail.com

Abstract

Objective: With the digital advancements in the field of agriculture, a large amount of data is being produced constantly as a result agriculture data has entered the world of big data. **Statistical Analysis:** As the development processes, the requirement of the parallel computing, compatible data management infrastructure and novel analytics paradigm to extract information from huge amounts are also increases. A single machine cannot store and analyze this large amount of data. The polynomial time required to access this kind of large data. **Findings:** The solution to store and analyze such massive amounts of data is big data analytics. In this paper a big data analytics recommendation framework is developed for providing solutions of crop disease based on historical data using Hive and Hadoop. The data is collected from various sources like laboratory reports, agriculture information web pages, and expert recommendation for the developed framework. After the collection of raw data, the irrelevant or the redundant data that is also known as the noise, should be removed. The next step is to extract the features from cleaned data, normalization of data is done in order to remove the technical variations. Once normalization is complete the data is uploaded on HDFS and save in a file that is supported by Hive. Thus classified data is finally located on the specific place. In the next step HiveQL is used to analyze agriculture data based on features and then prioritize the outcome based on crop disease symptoms and in the last a high priority solution is recommended. **Application/Improvements:** In the paper prioritize outcomes are useful for agriculture officers, researchers to easily understand, and helpful for recommending a solution based on evidence from historical data.

Keywords: Agriculture, Big Data, Expert Recommendation, Hadoop, Hive

1. Introduction

Digital technology is transforming agriculture into an intelligent world. With the rapid increase of data there arises the need of innovative technical and analytical strategies that are capable of handling the complex data structures. As a result with the increasing impact of world wide web (agriculture information websites, social media etc.,) and internet of things (radio equipments), agriculture has entered into the world of big data. Big data is a heterogeneous mixture of structured and unstructured data that is growing at an astonishing rate. Big data refers to the digital large scale data that is difficult to manage and analyze using traditional software tools and technologies[1–3]. Agriculture science researchers are discovering novel solutions for three major challenges of agriculture big data that are scalable infrastructure, management schemes and data mining analysis methods for large datasets. The focus of the paper is to develop recommendation system for crop disease control that will help researchers and agriculture officers in decision making from historical data with help of big data analytics. Discussed that if big data analytics is used in agriculture it will not only be a great innovation in the history of human agriculture as well as a pioneering work in human history[4]. The Apache Hadoop provides tools and libraries for distributed data storage and management that has been discussed in the Table 1. Apache Hadoop platform solves the problems of agriculture big data that are of scalable infrastructure and management schemes[5–6]. Machine learning is used for solving the problem of big data analytics as machine learning is a multidisciplinary field of computer science,

**Author for correspondence*

Table 1. Hadoop and tools

Platform/Tools	Description
Apache Hadoop	An open source software framework developed for distributed storage and processing of large amounts of data within a single platform from the Apache Software Foundation[10].
MapReduce	A Programming paradigm that easily writes applications that have the capabilities to process huge amounts of data in parallel[11].
HDFS	HDFS as the name suggests is a distributed file system. It is the sub project of Apache Hadoop and provides the access to large amounts of data on Hadoop clusters[12].
Apache Mahout	Apache Mahout is a open source project developed by the Apache Software Foundation (ASF) for producing scalable machine-learning algorithms[13].
RHadoop	R is widely accepted programming language that is accepted worldwide for data analysis, statistical computing and data visualization[14].
Apache Hive	It is an open source data warehouse software that resides on the top of the Hadoop .It serves the purpose of easy analysis, querying and summarization of big data[15,16].
Apache HBase	HBase is a Hadoop database a distributed, column oriented DBMS. Useful when one wants realtime read/write access for the big data[17].
Apache Pig	Apache Pig is a open source high level platform that is being for parallel execution of data flows on Hadoop[18].
Zookeeper	Zookeeper provides simple interface and operational services for the hadoop clusters with the key benefits of simplicity, reliability, fast and ordered accessibility[19].
Oozie	Oozie is a java web application that acts as a workflow scheduler for Hadoop[20].

artificial intelligence and statistics. Recently, it has been used by data scientists for exploiting the information hidden in big data by discovering the associations and understanding the patterns and trends of the big data[7]. In agriculture, most of the analytical methods are statistical based and are designed for analyzing single experimental dataset. Here one needs to rethink about the data analytics strategies to develop powerful tools with the capabilities that can analyze the big data in a better way. Researchers are trying to develop a large scale data analytics tool using machine learning. The big data machine learning analytics tools Mahout and Rhadoop are already being used in healthcare sector, telecom industry, education system, banking sector, e-commerce etc. IBM introduced the term precision agriculture like healthcare last year[8]. Precision agriculture analytics can be used to make intelligent decisions in farming by collecting data on weather, soil, air from data sources like social media etc. and machine to machine communications generates data that can be used for making decisions[9]. Till date single digit research articles are published in this field that are considered as the opinion articles but few agriculture based companies are currently working on machine learning analytics techniques based on single data set that are discussed in Table 2. The aim of the current research work is to develop a big data analytics recommendation framework to control crop disease that helps agriculture officers and researchers. The proposed work is easily understandable and helpful for recommending a solution based on evidence from historical data. In the next sections Hadoop and its tool that are used in the research work, and methodology of developed framework will be discussed in detailed. In the last section the developed framework for recommending solutions for the paddy leaf blast in Punjab (state) region will be demonstrated.

2. Framework Methodology

The purpose of the framework is to help agriculture officers and researchers to understand and recommend a solution for crop disease based on evidence from historical data. Developed big data analytics recommendation framework for providing solution for crop disease is little bit complex as compared to the traditional analytics system as shown in Figure 1. Agriculture data source provides information related to agriculture production, agriculture industry, discussion and recommendation related on agriculture problems. In world of technology uncountable agriculture data sources are available. In this framework data is collected from agriculture laboratory, agriculture department and agriculture info websites.

Table 2. Machine learning in agriculture

Project	Description
MySmartFarm[21]	MySmartFarm is Software as a Service platform that enables the farmers to work with live data. The farmer’s data and technology services that are essential for them are unified at one place for the betterment .
Awhere Weather[22]	Awhere is a SaaS platform that provides the location intelligence for global development in the field of agriculture. The objective of the platform is to provide visibility to the data and enable the evidence based decision making .
Phenonet[23]	Phenonet is world s first ever high resolution crop selection platform that is invaluable because of its unique capabilities to collect data from remote locations and send it data back to the laboratories in real time.
FarmLog[24]	FarmLog is software platform used by 20% USA farmers in form of web application or mobile application. Farmlog application helps the farmers to collect and retrieve data, helps in decision making.
Datafloq[25]	MyJonDeere is a software platform as well as mobile application that uses R programming analysis tool for decision making. It acts as a mentor for the farmers and guides them in making decisions regarding which crop to plant where and when. It uses historical as well as the real time data related to the weather predictions, soil conditions, water reports etc..
Farmeron[26]	Farmeron is a first cloud based dairy farm business software that uses the SaaS cloud services. Farmeron helps farmers to manage their farming data online. The farm performance analysis is done by using the existing statistics.

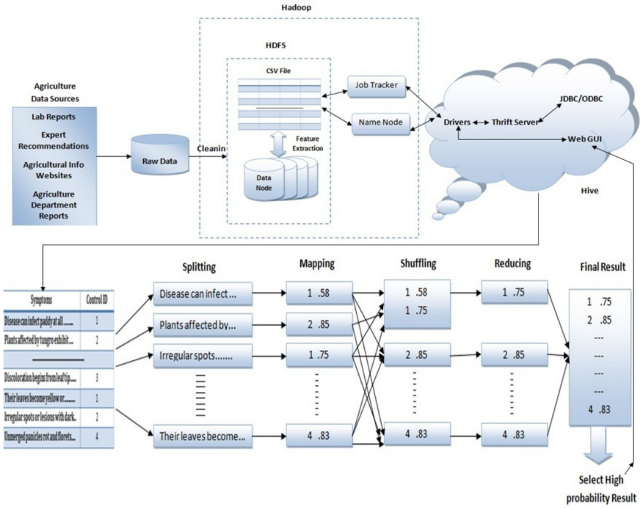


Figure 1. Big data analytics framework for recommendation solution of crop disease.

2.1 Laboratory Test Reports

The laboratory test reports are very important source of data for researchers and the various types of test performed. The tests that are conducted are soil, water, plant analysis, manure, compost, biosolids, green roof media and green house media testing etc.

2.2 Agriculture Info Websites

Agriculture info websites provides statistics, real time data related to agricultural economic entities, expert discussion and articles on specific problems. Aim of agriculture info websites is to provide information to everyone, anywhere and at anytime that empowers farmers to go for new techniques. Flume is used to download web pages from agriculture info websites and are save on Hadoop distributed file system.

2.3 Agriculture Department Reports

Agriculture department Reports provide information related to individual field of specific geographically area. For example effects of high temperature on paddy and how to control it in specific district. Agriculture department Reports very helpful for decision making of specific geographical area. In collected data is stored on Hadoop distributed file system in raw format. Row data contain large number of attributes and intolerance noise that make data meaningless. First step is to remove noise from data for example web pages contain hyperlinks of other web pages, advertisement etc. then dimension reduction including feature selection and extraction. Feature selection can be described as the process of selecting the subsets of relevant or important features. Feature selection take place either at the data preprocessing or the model learning. After the process of feature extraction normalization of data is done in order to remove technical variations by replacing values because missing or incomplete data disturbs the process of decision making. Solutions of the missing data problem are to delete incomplete data.

Next step is to run SQL query on Hive. Hive is open source data warehouse software that resides on the top of the Hadoop. It serves the purpose of easy analysis, querying and summarization of big data. Hive handle SQL query using HiveQL. HiveQL run SQL query on distributed environment. Query is submitted to distributed system via command line interface, Web User Interface or using external Application Programming Interface (API) like iReport, JDBC or ODBC using thrift server. Thrift server is a framework for cross language services where client and server communicate with each other using different languages. After submitting query masternode optimizes the query and submitted to map-reduce that implemented query on distributed environment. Hive server returns Solution ID, Symptoms and location and save into text file on HDFS. The output file that is saved on HDFS is again submitted to masternode for map-reduce task to calculate similarity between crop disease symptoms. Masternode splits file on hadoop environment. Mapping calculate the similarity of each record, similarity check gives the information like how much given symptoms match with disease symptoms based on adjacent characters that ignores minor spelling mistakes and symptoms word ordering. For example check similarity between AGRICULTURE (s1) and AGRICLTURE (s2).

$$Similarity\,(s1, s2) = \frac{\text{Pairs(s1)U Pairs(s2)}}{Pairs(s1)}$$

Paris s1 = AG, GR, RI, IC, CU,UL, LT, TU, UR,RE
Paris s2 = AG, GR, RI, IC, CL, LT, TU, UR,RE

$$Similarity\,(s1, s2) = \frac{\text{AG, GR, RI, IC, LT, TU, UR, RE}}{\text{AG, GR, RI, IC, CU, UL, TU, UR, RE}}$$
$$= 8/10 = .80$$

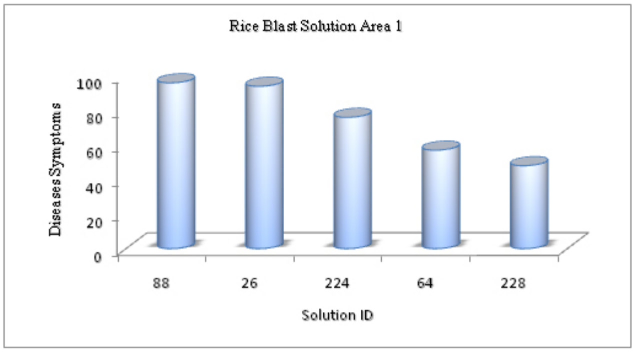
After mapping, reducer combine results based on solution, select highest matching symptoms where solution is same and results are saved in text file. Now sorting results based on similarity and select top (5 to 10) solution from each category, categories the data based on location from same state. All categories recommended solutions and common solution from all categories are display in graphically chart format using freeChart library that helps to recommend good solution.

3. Experimental Setup and Results

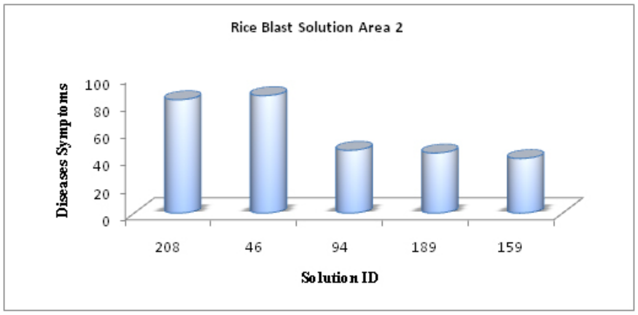
In this experiment setup private Hadoop 2.6.4 infrastructure with configuration master node has 8GB RAM and i7 CPU and two datanodes have 3 GB RAM and i3 CPU has been established. Data is collected from TATA chemicals laboratory and agriculture websites. After feature extraction dataset contain proximate five thousands records related to paddy, wheat and cotton related diseases. Data includes crop name, disease name, diseases type, disease detail, location, disease symptoms, solution id etc Eclipse IDE used to develop an application using java programming language. The developed framework for recommending solutions for the paddy leaf blast in Punjab (state) region will be demonstrated now. Now take a SQL query to select region specific leaf blast related data from Hive server and save into text file on Hadoop server for further use.

```
Insert Overwrite Directory
‘/http://localhost:9000/DataWareHouse/’
ROW FORMAT DELIMITED FIELDS TERMINATED
BY ‘;’
SELECT Field-Location, Disease-Symptoms, Solution-ID
FROM Crop-Disease-Table C
WHERE
C. Crop-Name=”Paddy” AND
C.Disease-Name =”Leaf Blast” AND
 C. Disease-Type =”Main Field Disease” AND
C.Field-State =”Punjab”
ORDER
BY location;
```

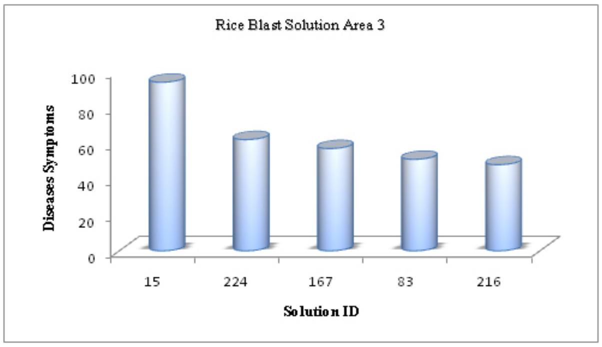
Now Hive convert this query to HiveQL query that is implemented using map-reduce and return results that save into a file. Above query result returns 55 out of 1094 records fall in this category. Result file contain Solution ID, Field-Location and Disease symptoms detail. Now mapper split a file into parts on different host, individual host evaluate probability based on similarity between crop diseases symptoms that insert by user and historical dataset crop symptoms and reducing shuffling the results based on Solution-ID, select highest matching symptoms where solution is same. After check crop disease Symptoms similarity select top 5 high probability records from each category show in Graph 1to 4. Now analysis common solution from all categories shows in Graph 5 and Graph 6. In solution 216 and 224 are mostly used solution for leaf blast in multiple regions. Now expert recommend solution from these solutions or proposed a new solution and data insert into hive database file for next time recommendation. Expert opinions on the recommendation increase the efficiency of system.



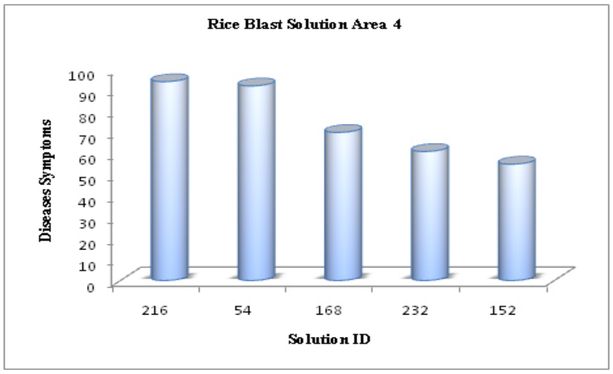
Graph 1. Rice leaf blast solution area 1.



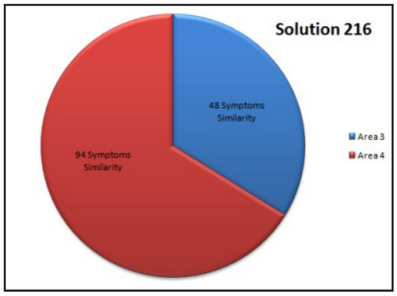
Graph 2. Rice leaf blast solution area 2.



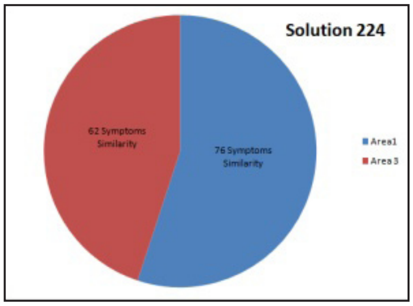
Graph 3. Rice leaf blast solution area 3.



Graph 4. Rice leaf blast solution area 4.



Graph 5. Rice leaf blast solution.



Graph 6. Rice leaf blast solution.

4. Conclusion

Agriculture has been transformed into an intelligent world as a result agriculture data has entered the world of big data. Big data analytics is magical technique that will help us in decision making. To handle agriculture crop disease problems big data analytics framework has been developed that provides recommendation solutions by using HIVE and Hadoop. The proposed framework is quite helpful for researchers and officers for recommending solutions based on evidence from historical data. The proposed work will help farmers to increase the yield of their crops. The future work involves the following:

- The developed framework is implemented using map-reduce and the retrieved data is stored on the masternode. The data retrieved with map reduce amounts in GB that masternode can't handle. The solution of this problem is spark. The future work will be to use Spark to store data at cluster node.
- The developed framework is location specific but the next step will be to develop a framework that will be dynamic in nature i.e., longitudinal and latitudinal parameters will be used.
- Agriculture ontology will be used to remove the language variations.

5. References

1. Laney D. 3D data management: Controlling data volume, velocity and variety. Meta Group Inc Application Delivery Strategies; 2001 Feb. p. 1–4.
2. Chen X-W, Lin X. Big data deep learning challenges and perspective. IEEE Access. 2014 May; 2:514–22.
3. Marx V. Biology: The big challenges of big data. Nature. 2013 Jan; 498(7453):255–60.
4. Zhang H, Wei X, Zou T, Li Z, Yang G. Agriculture big data: Research status, challenges and countermeasures. Proceedings of Computer and Computing Technologies in Agriculture; China. 2014 Sep. p. 137–43.
5. Schumacher A, Pireddu L, Niemenmaa M, Kallio A, Korpelainen E, Zanetti G, Heljanko K. Simple and scalable scripting for large sequencing data sets in hadoop. Bioinformatics. 2014 Jan; 30(1):119–20.
6. Nordberg H, Bhatia K, Wang K, Wang Z. Biopic: A hadoop-based analytic toolkit for large-scale sequence data. Oxford Bioinformatics. 2013 Sep; 29(33):3014–9.
7. Ratner B. Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data. 2nd ed. CRC Press Taylor and Francis Group; 2011.
8. Precision agriculture. Available from: http://ilsi.org/wp-content/uploads/2016/05/Robin-Lougee-2016-ILSI-Open-Data.pdf
9. Big data in agriculture. Available from: http://www.citethisforme.com/topic-ideas/technology/'Big%20Data'-6678234
10. Lam C, Warren J. Hadoop in action. 1st ed. Greenwich: Manning Publications; 2010 Dec.
11. Xu Y, Zhou W, Cui B, Lu L. Research on performance optimization and visualization tool of hadoop. Proceedings of 10th International Conference on Computer Science and Education; Cambridge University, USA. 2015 Jul. p. 149–53.
12. Shvachko K, Kuang H, Radia S, Chansler R. The hadoop distributed file system. Proceedings of IEEE 26th Symposium on Mass Storage Systems and Technologies; Incline Village, USA. 2010. p. 1–10.
13. Introducing apache mahout: Scalable, commercial-friendly machine learning for building intelligent applications. Available from: http://www.ibm.com/developerworks/library/j-mahout/
14. Song Y. Storage mining: Where it management meets big data analytics. Proceedings of IEEE International Congress on Big Data; New York, USA. 2013 Jun-Jul. p. 421–2.
15. Thusoo A. Hive- A warehousing solution over a map-reduce framework. ACM Digital Library. 2009 Aug; 2(2):1626–9.
16. Hive performance benchmark. Available from: https://issues.apache.org/jira/browse/HIVE-396
17. Sun J. Scalable RDF store based on Hbase and Mapreduce. Proceedings of 3rd International Conference on Advanced Computer Theory and Engineering; Chengdu, China. 2010 Aug. p. 633–6.
18. Gates A. Programming pig. 1st ed. California, USA: O'reilly Media Inc; 2011.
19. Hunt P, Flavio P. ZooKeeper: Wait-free coordination for internet-scale systems. Proceedings of Usenix Annual Technical Conference; MA. 2010 Jun. p. 11–29.
20. Lu Q, Li Z, Kihl M, Zhu L, Zhang W. A conceptual framework for big data analytics applications in the cloud. IEEE Access. 2015 Oct; 3(1):944–52.
21. My smart farm. Available from: https://www.kickstarter.com/projects/1911579744/my-smart-farm-active-food-producer
22. Awhere weather. Available from: http://www.awhere.com/about/news/awhere-weather-access-weather-data-at-the-click-of
23. Phenonet. Available from: http://www.csiro.au/en/Research/D61/Areas/Robotics-and-autonomous-systems/Internet-of-Things/Phenonet
24. Farm log. Available from: http://farmindustrynews.com/precision-farming/farmlogs-adds-rainfall-monitoring-new-smartphone-apps
25. Datafloq. Available from: https://www.linkedin.com/pulse/iot-big-data-needs-ceo-boardroom-agenda-praveen-senadheera
26. Farmer on. Available from: https://datafloq.com/read/john-deere-revolutionizing-farming-big-data/511