IJETIE International Journal of Emerging Technology and Innovative Engineering
Volume 4, Issue 5, August 2018 (ISSN: 2394 – 6598)

DATA ANALYTICS IN AGRICULTURE

Priyanka .G

Assistant Professor, Sri Krishna College of Technology

Coimbatore, India

ABSTRACT

Big data analytics is the process of collecting, organizing and analysing large sets of data to discover patterns and other useful information. Big Data is usually defined by 3V's. They are Volume, Variety and Velocity. The main aim of this paper is to analyze the given soil and its corresponding properties. This information can be used to determine the soil fertility which is essential for agricultural purposes. The dataset contains details about various soil types of the world and its composition. This dataset is the World Harmonized Soil Database and it is taken from FAO. In this project there are two analyses. First analysis is to determine the draining capacity of the soil based on the various levels of sand, silt, clay and gravel given in the dataset. The second analysis is to take the organic carbon content and pH level values from the dataset and determine which crop can be cultivated from the given soil sample.

Keywords- Big data, Hadoop, Map Reduce.

I. INTRODUCTION

Agriculture involves the domestication of plants. While soil is frequently referred to as the fertile substrate, not all soils are suitable for growing crops. Ideal soils for agriculture are balanced in contributions from mineral components (sand, silt, and clay), soil organic carbon, pH, air and water. The balanced contributions of these components allow for water retention and drainage, oxygen in the root zone, nutrients to facilitate crop growth and they provide physical support for plants. The soil properties dataset provides us with information about various percentages of sand, gravel, silt, clay, and pH and soil organic carbon. Analyzing the data can help us to gain information about which crop can be grown for given levels of pH and organic carbon. It also helps us to analyze whether agriculture is possible or not by predicting the level of draining capacity in the given soil.

Big data analytics is the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the knowledge that comes from analyzing the data. To analyze such a large volume of data, big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, and forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics. Using big data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions in the future.Big Data analytics can be applied in the field of agriculture as it helps to analyze the large amount of soil properties data and predictions made from the analysis can be used to bring general awareness to the farmers about how the soil can be used for agricultural purposes.

II. RELATED WORK

The use of standard statistical analysis techniques is both time consuming and expensive. Efficient techniques can be developed and tailored for solving complex soil data sets using data mining to improve the effectiveness and accuracy of the Classification of large soil data sets [14]. A soil test is the analysis of a soil sample to determine nutrient content, composition and other characteristics. Tests are usually performed to measure fertility and indicate deficiencies that need to be remedied [15]. The soil testing laboratories are provided with suitable technical literature on various aspects of soil testing, including testing methods and formulations of fertilizer recommendations [17]. It helps farmers to decide the extent of fertilizer and farm yard manure to be applied at various stages of the growth cycle of the crop. In a research carried out by Leisa J. Armstrong, comparative study of current data mining techniques such as cluster analysis and statistical methods was carried out to establish the most effective technique. They used a large data set extracted from the Western Australia Department of Agriculture and Food (AGRIC) soils database to conduct this research. The experiments analyzed a small number of traits contained within the dataset to determine their effectiveness when compared with standard statistical techniques [16].

III. SYSTEM ANALYSIS

Application development can be generally being thought of having two major components: analysis and design. In analysis more emphasis is given to understanding the details of an existing system or a proposed one and then deciding whether the proposed system is desirable or not and whether the existing system needs improvement. Thus, analysis is the process of investigating a system, identifying problems, and using the information to recommended improvements to the system.

1) Existing system

Existing system analyzes the data with less performance and also huge amount of data cannot be analyzed efficiently. The database size is normally limited to 10GB. The existing system has some drawbacks. Parallel processing of data is not supported and the system consumes more time to retrieve the required data. There is no assurance to effectively analyze the large amount of data.

2) Proposed system

Big Data analytics using hadoop framework enables to analyze even terabytes to petabytes of data. MapReduce algorithm enables parallel processing of the large amount of data. The prediction can be made effectively from the large amount of data. This framework enhances the speed and performance to retrieve the required data. It also enables the parallel processing of data. Large amount of data can be analyzed effectively. The prediction can be used to bring general awareness about diabetes to public.

IV. IMPLEMENTATION DETAILS

1) Basic system setup-Hadoop

In this module it contains the installation of hypervisor. A new guest operating system (Ubuntu 14.04.4) is created. After this process Hadoop is installed.

A) Hadoop

Hadoop is a batch processing system for a cluster of nodes that provides the underpinnings of most Big Data analytic activities because it bundle two sets of functionality most needed to deal with large unstructured datasets namely, Distributed file system and Map Reduce processing. It is a project from the Apache Software Foundation written in Java to support data intensive distributed applications. Hadoop enables applications to work with thousands of nodes and petabytes of data.

Architecture of Hadoop

Hadoop is a Map/Reduce framework that works on HDFS or on HBase. The main idea is to decompose a job into several and identical tasks that can be executed closer to the data (on Data Node). In addition, each task is parallelized: The Map phase. Then all these intermediate results are merged into one result: the Reduce phase. In Hadoop, The Job Tracker (a java process) is responsible for monitoring the job, managing the Map/Reduce phase, managing the retries in case of errors. The Task Trackers (Java process) are running on the different Data Nodes. Each Task Tracker executes the tasks of the job on the locally stored data.

The core of the Hadoop Cluster Architecture is given below.

HDFS:

HDFS is the basic file storage, capable of storing a large number of large files. An HDFS cluster has two types of node operating in a master-worker pattern: a Name Node (the master) and a number of Data Nodes (workers).

MapReduce:

MapReduce is the programming model by which data is analyzed using the processing resources within the cluster.

Name Node:

Manages file system metadata and access control. There is exactly one Name Node in each cluster. It maintains the file system tree and the metadata for all the files and directories in the tree. The name node also knows the data nodes on which all the blocks for a given file are located. Name Node decides about replication of data blocks.

Secondary Name Node:

It downloads periodic checkpoints from the name Node for fault-tolerance. There is exactly one Secondary Name Node in each cluster.

Data Node:

Data nodes are the workhorses of the file system. Holds file system data. Each data node manages its own locally-attached storage and stores a copy of some or all blocks in the file system. There are one or more Data Nodes in each cluster. They store and retrieve blocks when they are told to (by clients or Name node), and they report back to the name node periodically with lists of blocks that they are storing.

Job Tracker:

It hands out tasks to the slave nodes. There is exactly one Job Tracker in each cluster.

Task Tracker:

Task Trackers are slaves that carry out map and reduce tasks. There are one or more Task Trackers in each cluster.

Hadoop distributed file system (HDFS)

An HDFS cluster has two types of node operating in a master-worker pattern: a Name Node (the master) and a number of Data Nodes (workers). The name node manages the file system namespace. It maintains the file system tree and the metadata for all the files and directories in the tree. The name node also knows the data nodes on which all the blocks for a given file are located. Data nodes are the workhorses of the file system. They store and retrieve blocks when they are told to (by clients or the name node), and they report back to the name node periodically with lists of blocks that they are storing. Name Node decides about replication of data blocks. In a typical HDFS, block size is 64MB and replication factor is 3 .The Figure1 shown architecture distributed file system HDFS.

Hadoop MapReduce applications use storage in a manner that is different from general-purpose computing. To read an HDFS file, the client applications simply use a standard Java file input stream, as if the file was in the native file system. Behind the scenes, however, this stream is manipulated to retrieve data from HDFS instead. First, the Name Node is contacted to request permission.

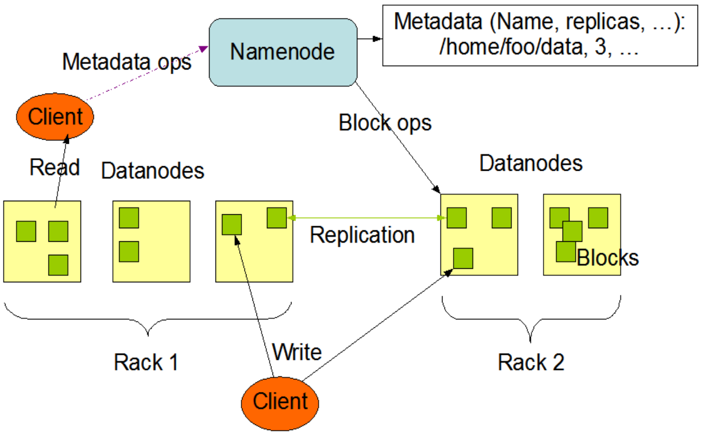

Fig 1: HDFS Architecture

If granted, the Name Node will translate the HDFS filename into a list of the HDFS block IDs comprising that file and a list of Data Nodes that store each block, and return the lists to the client. Next, the client opens a connection to the “closest” Data Node and requests a specific block ID. That HDFS block is returned over the same connection, and the data delivered to the application. To write data to HDFS, client applications see the HDFS file as a standard output stream. Internally, however, stream data is first fragmented into HDFS-sized blocks (64MB) and then smaller packets (64kB) by the client thread. Each packet is enqueued into a FIFO that can hold up to 5MB of data, thus decoupling the application thread from storage system latency during normal operation. A second thread is responsible for dequeuing packets from the FIFO, coordinating with the Name Node to assign HDFS block IDs and destinations, and transmitting blocks to the Data Nodes (either local or remote) for storage. A third thread manages acknowledgements from the Data Nodes that data has been committed to disk. To support this throughput HDFS leverages unusually large (for a file system) block sizes and data locality optimizations to reduce network input/output (I/O).

B) Map-reduce

MapReduce is a data processing or parallel programming model introduced by Google. In this model, a user specifies the computation by two functions, Map and Reduce. In the mapping phase, MapReduce takes the input data and feeds each data element to the mapper. In the reducing phase, the reducer processes all the outputs from the mapper and arrives at a final result.

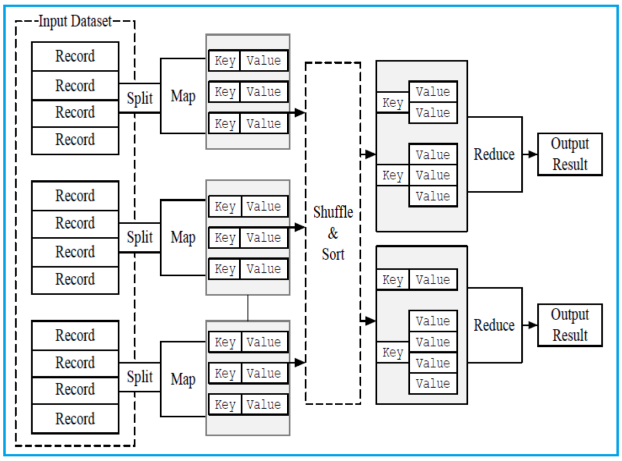

Fig 2: MapReduce Execution

In simple terms, the mapper is meant to filter and transform the input into something that the reducer can aggregate over. The underlying MapReduce library automatically parallelizes the computation, and handles complicated issues like data distribution, load balancing and fault tolerance. Massive input, spread across many machines, need to parallelize. Moves the data, and provides scheduling, fault tolerance. The original MapReduce implementation by Google, as well as its open-source counterpart, Hadoop, is aimed for parallelizing computing in large clusters of commodity machines. Map Reduce has gained a great popularity as it gracefully and automatically achieves fault-tolerance. It automatically handles the gathering of results across the multiple nodes and returns a single result or set. MapReduce model advantage is the easy scaling of data processing over multiple computing nodes.

2) Data Collection

Database for agriculture data analysis is obtained from the world harmonized soil database. The data is stored in CSV file format. Initially it is stored in local system. After that it is copied to HDFS. The data includes the soil properties such as draining capacity, levels of gravel, sand, silt, clay, pH and organic carbon .The analysis on the data is performed using Map reduce.

3) Data aggregation and storage

In this module, the soil properties such as draining capacity, levels of gravel, sand, silt, clay, pH and organic carbon is aggregated from the soil properties database. The aggregated data is stored in the local system. By using command the aggregated data is copied to HDFS. The aggregated data size is 2GB. The data is then chunked based on DFS block size specified in hdfs-site.xml.

4) Data analytics

The analysis is performed by using map reduce algorithm in java. The output of the analysis will be in key value pair.

Analysis 1:

The output of this analysis is the draining capacity of the given soil sample. Here the input to the map phase is the CSV file and the output is key value pair like (1 1, 1, 1,), which is given as input to reduce phase. The output of the reduce phase is the draining capacity.

Analysis 2:

The output of this analysis is the type of crop grown for the given pH values in the dataset.

Analysis 3:

The output of this analysis is the type of crop grown for the given organic carbon values in the dataset.

V. CONCLUSION AND FUTURE WORK

The first analysis determines the draining capacity of the given soil which helps us to predict whether agriculture is possible or not. The analysis made on the pH and organic carbon content of the soil helps us to determine the percentage of particular crop grown. As future enhancement more crops can be added to the dataset that has been analysed. This application can further be enhanced using some other classification algorithms like Bayesian classification algorithm etc. Data visualization can also be done using other data visualization tools like Tableau, Sap Lumira, etc.

REFERENCES

[1] ‘Apache hadoop, http://hadoop.apache.org/.’

[2] A. R. Chaaitupe, Prof. S. A. Joshi (1984), ‘Data Classification Algorithm Using k-Nearest Neighbour Method Applied to ECG Data’, IOSR Journal of Computer Engineering, Vol. 14, pp. 13-21.

[3] http://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies

[4] http://hadooping.weebly.com/blog/mapreduce-to-calculate-max-temp

[5] Jay Gholap, Anurag Ingole, Jayesh Gohil, Shailesh Gargade, Vahida Attar, “Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction”, Vol. 12, pp. 12-20, 2004.

[8] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on Large clusters”, Commun. ACM, vol. 51, pp. 107-113, 2008.

[9] J. S. Raikwal, Kanak Saxena, “Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set”, International Journal of Computer Applications, Vol. 50, pp. 141-150, 2012.

[10] Khaled Alsabti, Sanjay Ranka, Vineet Singh, “An Efficient K-Means Clustering Algorithm: Analysis and Implementation”, Vol.4, pp. 8-10, 2001.

[11] Rabi Prasad Padhy, “Big Data Processing with Hadoop-MapReduce in Cloud Systems”,International Journal of Cloud Computing and Services Science, Vol.2, pp. 16-27, 2013.

[12] Sangeeta Bansal, Dr. Ajay Rana, “Transitioning from Relational Databases to Big Data”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, pp.23-30, 2014.

[13] Shahrukh Teli, Prashasti Kanikar, ‘A Survey on Decision Tree Based Approaches in Data Mining’, International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 5, pp. 56-60,2015.

[14] A. Kumar & N. Kannathasan, (2011), “A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining “, IJCSI International Journal of Computer Science Issues,Vol. 8, Issue 3,

[15] “Soil test”, Wikipedia, February 2012

[16] L. Armstrong, D. Diepeveen & R. Maddern, **“**The application of data mining techniques to characterize agricultural soil profiles”,2004,

[17] “Methods Manual-Soil Testing in India”, Department of Agriculture & Cooperation Ministry of Agriculture Government of India, 2011.