Big Data Analytics on Weather Data: Predictive Analysis Using Multi Node Cluster Architecture

Vikash Kumar
Department of ISE, SDMCET (Affiliated to VTU, Belagavi), Dharwad, Karnataka,

Abhijeet D, Anurag Gupta
Department of ISE, SDMCET (Affiliated to VTU, Belagavi), Dharwad, Karnataka,

Rajashekarappa
PhD, Department of ISE, SDMCET (Affiliated to VTU, Belagavi), Dharwad, Karnataka, India

Parameshachari B D
PhD, Department of TCEE, GSSSIETW (Affiliated to VTU, Belagavi), Mysuru

ABSTRACT
Data is growing at a tremendous rate with an increase in digital universe. However increase in data itself is a minor problem, but the increase in percentage of unstructured data in the overall data volume is matter of concern. Big Data Analytics surpasses conventional business intelligence programs through enabling users to analyze larger amounts of data, including unstructured data which is usually left redundant. Big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data challenges. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues. This paper, describes the formatting guidelines for IJCA Journal Submission.

Keywords
Big data analytics; Hadoop; multi node cluster architecture; Map reduce

1. INTRODUCTION
Big Data has bought a paradigm in the way industry looks at storing and consuming data. The weather is awash in data, much of it is unstructured. The weather datasets cannot be processed by using traditional database system. Weather sector is constantly adapting to use technology in new ways in order to get some useful information. The driving force behind an implementation of Big Data is Hadoop. The project Report starts with the introduction to Big Data which describes the characteristics of Big Data and its need in the Weather Analysis technology.

Typical attributes of Weather data are as follows:

1. Volume – Weather data is massive. National Oceanic and Atmospheric Administration (NOAA) have data measured in terabytes.

2. Velocity – Weather data is not a real-time and require prediction.

3. Variety – Ad-hoc systems traditionally have different structures and shapes. Data analysis on such data is difficult because of rigid schemas

4. Value – Weather data on its own is low value until it is rolled up and analyzed. At this point data becomes information which, in turn, becomes knowledge for the broader market.

Characteristics of big data is shown in figure 1. The big data trend in climate department promises that harnessing the wealth and volume and information in enterprise leads to better prediction, operational efficiency, and futuristic benefits.

2. PROBLEM STATEMENT
The purpose of this application is to process and analyze large data sets of Weather in order to get some useful information which will help the users to improve their business intelligence. Data is available in abundant form, using HDFS to retrieve the weather data for the further analysis. Using multi node cluster architecture for fast retrieval of uploaded dataset. To find solution from analysis which can be useful for many and using big Data Approach. Big data and predictive analytics can potentially provide accurate analytics.

3. OBJECTIVES
Weather data set is loaded on to HDFS for the further usage. The data will be put into the MAPPER function which will map the data with respect to the key values such as temperature**.** The Mapped valued are then shuffled and sorted and given to the REDUCE Function as an input. Useful values out of reducer are taken for evaluation and plotting the Graph.

This Project Mainly does the analysis on Weather data set, the purpose of a weather forecast is to provide as accurate as possible prediction of what the weather will be like in the near future. Weather has adverse effect on all community. So taking into consideration analysis on weather data will benefits in various aspects. They are important to most aspects of day to day life, including aviation, boating, other modes of transportation, farming, tourism, sports, etc. With forecasting methods, companies can get better outcomes with the help of accurate prediction.

Big data and predictive analytics can potentially provide accurate, real-time or near real-time analytics. Big data analysis can provide enough information to do research work. Big data and predictive analytics technologies will provide stakeholders to process huge volumes of data fast and generate accurate insights.

People will be warned earlier of what the weather will be like for that particular day. With forecasting methods, companies can get better outcomes with the help of accurate prediction**.**

4. METHODOLOGY

Big data sources refer to data collection or data gathering. The data can be collected from public domain sites. In this project Weather datasets have been collected from https://data.gov.in For analyzing datasets of Weather Hadoop Clustering Technique have been used which uses Map Reduce programming model on Cygwin terminal on Windows operating system. Hadoop is a popular choice needed to filter, sort or pre-process large amounts of new data in place and distill it to generate denser data that theoretically contains more "information". The core of Hadoop consists of a storage part (Hadoop Distributed File System) and a processing part (MapReduce).

Let us discuss how the Hadoop Multi-node Cluster works, how the data is stored in HDFS, and lastly how the data is processed using the Map-Reduce programming model.

4.1 Multi-Node Cluster Architecture

To start with multi-node cluster each machine which is considered as master and slaves nodes must have Hadoop installed on Cygwin terminal on Windows operating system. Before starting the cluster step it is needed to configure files such as core-site.xml, mapred-site.xml, hdfs-site.xml and assign the static IP address to each of the nodes in cluster. There should be at least on master node and any number of slave nodes. The master node will assign the job to slave nodes; the slave nodes are the actual ones that execute the job. jobs running in multi-node cluster can be seen through browser. In figure 2, Let us see how the multi-node cluster looks with 1 master node and 2 slave nodes.

The master node will run the "master" daemons for each layer. Name node for the HDFS storage layer, and Job Tracker for the MapReduce processing layer. Both machines will run the "slave" daemons: Data Node for the HDFS layer and Task Tracker for MapReduce processing layer. Basically, the "master" daemons are responsible for coordination and management of the "slave" daemons while the latter will do the actual data storage and data processing work.

4.2 Hadoop Distributed File System (HDFS)

Hadoop provides a distributed file system and a framework for the analysis and transformation of weather datasets using the MapReduce paradigm. For maximum portability, HDFS is implemented as a user-level filesystem in Java which exploits the native filesystem on each node, such as NTFS, to store data. Files in HDFS are divided into large blocks, typically 64MB, and each block is stored as a separate file in the local filesystem [3]

HDFS is implemented by two services:

•Name Node

•Data Node

The name node manages the filesystem namespace. It maintains the filesystem tree and the metadata for all the files and directories in the tree. This information is stored persistently on the local disk in the form of two files: the namespace image and the edit log. The name node also knows the data nodes on which all the blocks for a given file are located, however, it does not store block locations persistently, since this information is reconstructed from data nodes when the system starts.

A client accesses the file system on behalf of the user by communicating with the name node and data nodes. The client presents a POSIX-like filesystem interface, so the user code does not need to know about the name node and data node to function.

Data nodes are the workhorses of the filesystem. 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.

Without the name node, the filesystem cannot be used. In fact, if the machine running the name node was obliterated, all the files on the filesystem would be lost since there would be no way of knowing how to reconstruct the files from the blocks on the data nodes. For this reason, it is important to make the name node resilient to failure, and Hadoop provides two mechanisms for this.

The first way is to back up the files that make up the persistent state of the filesystem metadata. Hadoop can be configured so that the name node writes its persistent state to multiple filesystems. These writes are synchronous and atomic. The usual configuration choice is to write to local disk as well as a remote NFS mount.

It is also possible to run a secondary name node, which despite its name does not act as a name node. Its main role is to periodically merge the namespace image with the edit log to prevent the edit log from becoming too large. The secondary name node usually runs on a separate physical machine, since it requires plenty of CPU and as much memory as the name node to perform the merge. It keeps a copy of the merged namespace image, which can be used in the event of the name node failing. However, the state of the secondary name node lags that of the primary, so in the event of total failure of the primary, data loss is almost certain. The usual course of action in this case is to copy the name node's metadata files that are on NFS to the secondary and run it as the new primary.

4.3 Map Reduce

For processing of Weather datasets MapReduce [4] programming model is used. MapReduce works by breaking the processing into two phases: the map phase and the reduce phase. Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer. The programmer also specifies two functions: the map function and the reduce function. A text input format is taken that gives us each line in the dataset as a text value. Our map function is simple. The MapReduce function works on key value pair.

The mapper takes a list of key-value pairs, and applies some operation to each element independently. The reducer takes a single key and a list of values for that key, and outputs a new list of values with the same key. The reducer takes a single key and a list of values for that key, and outputs a new list of values with the same key. Let us consider the logical data flow of MapReduce in figure 4.

Here input is the frequent change in temperature. Map function takes the frequent temperature as the key and date as the value. The reducer function takes the frequent temperature as the key and date as values. In the output the average will be displayed.

5. LITERATURE SURVEY

5.1 Need For Research

This research has been done because of the rapid increasing of Weather data set. From scientific discovery to business intelligence, "big data" is changing our world. The dissemination of nearly all information in digital form, the proliferation of sensors, breakthroughs in machine learning and visualization, and dramatic improvements in cost, bandwidth, and scalability are combining to create enormous opportunity. The field also presents enormous challenges, in the volume, velocity, and variety of information ripe for mining and analysis.

5.2 Existing System

The Weather organization uses Data mining and Data Warehouse techniques for analyzing the weather and climate conditions. Files are designed by using programs written in programming languages such as COBOL, C, C++, and DBMS. Existing system uses traditional RDBMS to retrieve data from large dataset. Therefore the techniques to be used for very large databases will have to be different. RDBMS lacks in high velocity because it‘s designed for steady data retention rather than rapid growth. Even if RDBMS is used to handle and store ―big data, it will turn out to be very expensive. Relational databases, with their limitations in handling ―big data, aren‘t much help either. As a result, ―big data‖ is sometimes considered to be the data that can‘t be analyzed in a traditional database. As a result, the inability of relational databases to handle ―big data‖ led to the emergence of new technologies.

5.3 Proposed System

In this system, input data is stored in the HDFS of the Hadoop. MapReduce algorithm is applied. Output is stored in files. Bar chart is generated from the output. Predictive analysis is done to get the predicted weather conditions.

6. TECHNOLOGY USED

6.1 Hardware Requirements

- Minimum 50GB Hard disk free.
- Minimum 4GB DDR3 memory.
- A screen to display the content.

6.2 Software Requirements

- Platform : Windows operating system.
- Language : Python, Java
- Software : Cygwin terminal, Hadoop 2.7.0

6.3 Windows

Windows, which supports large, commoditized, distributed environments, is a good choice for Big Data scenarios. Perfect for in-house environments and for cloud-based systems.

Windows offers a number of distinct advantages for organizations doing Big Data analytics. The worldwide Windows developer community has optimized a large number of Big Data applications to run on Window Server, including Cygwin and Hadoop. Windows offers all the latest developer tools in this fast-moving market – ensuring it remain the platform of choice for Big Data development. Big Data applications are deployed to provide focused results or reports, the infrastructure that supports them must be fast and easy to deploy. To meet this need, Windows comes with a range of tools for rapid infrastructure and service deployment.

7. APPLICATIONS

- The Application of Big Data analytics on Weather data can be used by the Weather related organization in order to improve their business intelligence [5].
- Big data and predictive analytics can potentially provide accurate and near real-time analytics.
- Big data and predictive analytics technologies have enabled stakeholders to process huge volumes of data fast and generate accurate insights.
- People get warned earlier of what the weather will be like for that particular day.
- With forecasting methods, companies can get better outcomes with the help of accurate prediction.
- It provides the business with valuable information that the business can use to make decisions about future business strategies.
- Delivers visual forecasts by methods most companies prefer.
- Helps agricultural organizations in buying/selling livestock.
- Helps farming industry in planting crops, pastures, water supplies.

8. CONCLUSION

Big data isn’t just hype – and it’s much more than a buzz frame. Today weather department are finding that they not only need to manage increasingly large data volumes in their real time systems, but also analyze that information so they can make the right decisions –fast-to compete effectively in the market. Big Data Analytics on weather data helps in providing the correct conclusion from huge mass of data outliers. This in turn will help in creating proper business model which is used for intelligence business decision and it also highlights where they can do better so that it increases the better prediction techniques.

9. FUTURE SCOPE

Weather data comes in a variety of forms from a number of Forecasters use meteorological data to support a number of programs including public, aviation, fire and marine. Weather warnings are important forecasts because they are used to protect life and property. Forecasts based on temperature and precipitation are important to agriculture, and therefore to traders within commodity markets. People get warned earlier of what the weather will be like for that particular day. With forecasting methods, companies can get better outcomes with the help of accurate prediction.

10. ACKNOWLEDGEMENT

The sense of contentment and elation that accompanies the successful completion of the task would be incomplete without mentioning the names of the people who helped in accomplishment of this project, whose constant guidance, support and encouragement resulted in its realization. Special Thanks to the principal of SDMCET Dharwad, Dr. S.B.VANAKUDRE for providing the serene and healthy environment within college, which helped in concentrating on the work. We are grateful to Dr. RAJASHEKARAPPA for his valuable guidance and encouragement.

Lastly it is much indebted to the parents and friends for their unquestioning cooperation and help.

11. REFERENCES

[1] International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016, A Survey on Big Data Analytics.

[2] Michael Minelli, Michele Chambers, Ambiga Dhiraj published in the year March 2013. Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses.

[3] Oreilly.Hadoop.3rd.Edition.Jan.2012, BIGDATA: The Definitive Guide

[4] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters.

[5] Big Data Executive Survey 2017, Big Data Business Impact: Achieving Business Results through Innovation and Disruption.

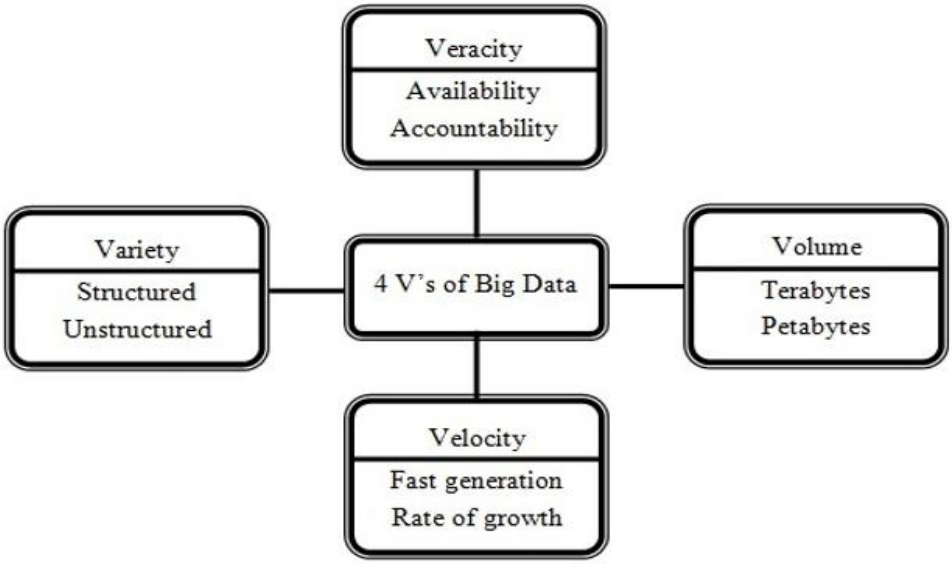

Fig.1: Characteristics of Big Data [1]

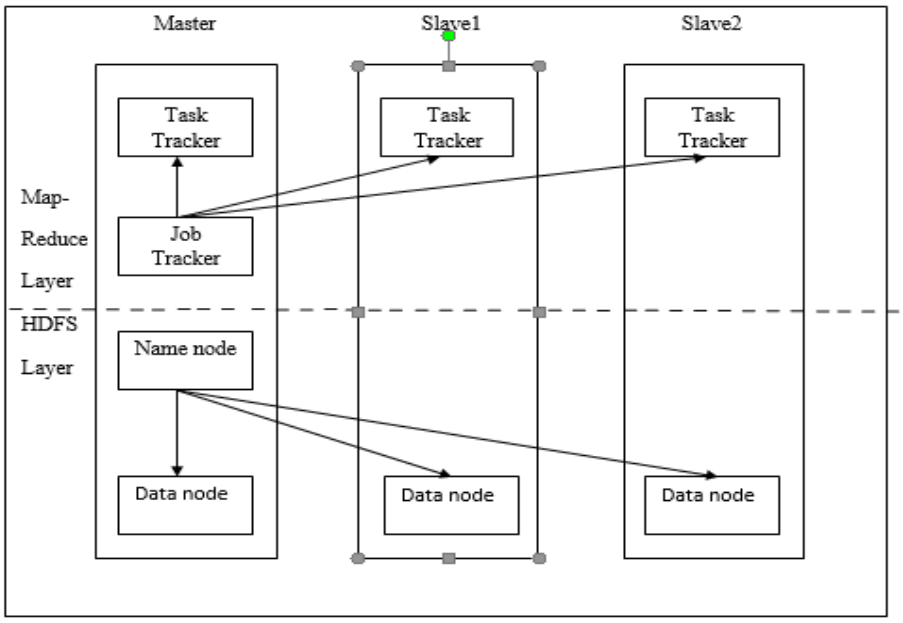

Fig. 2: Multi-node Cluster

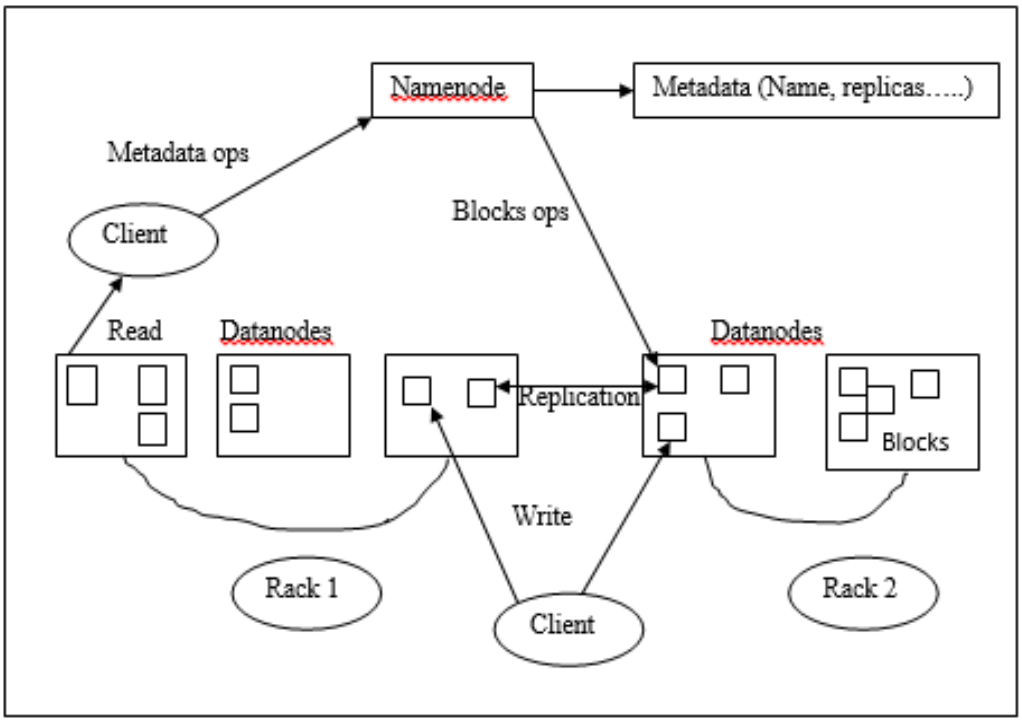

Fig.3: Architecture of HDFS

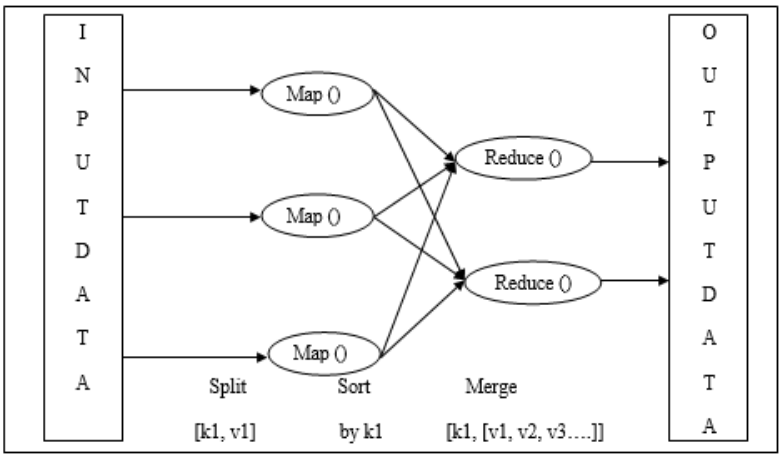

Fig.4: MapReduce Processing

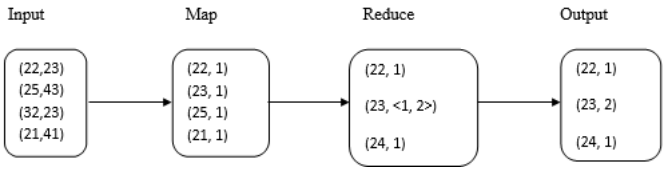

Fig. 5: Logical data flow of MapReduce