Big Data Analysis on Weather Prediction Based on Using IOT and Collected Sensor Data

Dr. Shubangi D C[1], Soumya S Nigudgi[2]

[1]*H.O.D, department Computer Science and Engineering, VTU Regional Centre, Kalaburagi, Karnataka, India*

drshubhangipatil1972@gmail.com

[2]*PG Student department of computer Science and Engineering, VTU Regional Centre, Kalaburagi, Karnataka, INDIA.,*

nigudgis@gmail.com

***Abstract*— The assets of the sensor generated data day by day, by providing high volume of massive real time data. Analysing, aggregating, storing, and collecting the sensor data is a great challenge. Metrological department use different types of sensors to find the temperature, humidity, moisture etc to get there values. Keeping these above points in mind we design a system architecture that welcomes the real time data as well as the offline data processing. The proposed architecture has three modules Data acquisition unit (DAU), Data Processing unit (DPU), Data Analysis and Decision Unit (DADU). This architecture is also capable for storing the incoming data to perform the offline analysis of the sensor based temperature data. Hadoop provides an open source framework suitable for processing large data. Map Reduce helps to process the large volume data in parallel. The main aim of the project is to build analytical engine for high velocity, huge volume temperature data from sensor using Hadoop Map Reduce and provide the predictions on the obtained results for the upcoming year's weather forecast.**

***Keywords*— Big Data, IOT, Arduino, Map Reduce, whether analytics, Temperature.**

I. INTRODUCTION

Big data has become one of the catchphrases in IT during the last couple of years. At the beginning it was shaped by the organizations for handling the speed growth rate of the data. Most of the companies now days are based on indexing and using this large amount of the data the calculation, collecting, and storing of the weather data is also a very big task. The concept of big data and IOT are related to similar areas. The one kind of a data which is similar to Big data is the data which is collected through the sensors, i.e., sensor data. The sensor data usually consist of huge amount of the data with very high velocity. The sensor generated the data for every minute and seconds in a day. Hence managing this large data is very important. The IoT mainly consists of all kind of the gadgets and devices. In those some of them are designed to sensor and storing them in the cloud, and accessing this data through internet is the work done over here.

India is a developing country. Most of the cities in India are now becoming smart cities. Varieties of sensors are being used to measure the weather parameters. Not only India many parts on the earth provide huge amount of data every second, managing this data is very big process. In our work the small city of Karnataka called Kalaburagi is considered. Here we use the sensor data which is obtained by using the IOT board, through that we connect to the internet and access this data through the website called ThingSpeak. This website will allow us to store the data in cloud which we are retrieving through the sensors. In our work the offline data is collected from the NCDC, and upon that we calculate the Max and Min for particular month of the year.

In [1] presented the article by saying that the DBMS for the both bulletin and to a great of work application workloads. It also includes the decision support systems for the more deep analytics and descriptive, which are the most dangerous part of the cloud environment. In [2] C. Eaton, D. Deroos, T. Deutsch, says that online transaction produces the huge amount of data such as email, clicks, social network data, sensor data etc. in this paper the authors in depth of understanding the big data and analysis of the hadoop and it streaming data. In [3] shows the way the data is collected is being updated by big data sensing and computer technology, processed and analysed and managed. Recently designed sensors that are used in earth observatory system are generating huge amount of data.

In[4] Hansen, James, and Sergej Lebedeff proposed the calculation of the surface air temperature. To calculate it is necessary to know the surface temperature value, which is available from the different base stations. With the help of satellite different surface value is sent to the base sation,and from there the data is accessed and processed to calculate the value of temperature. In [5] the use of map reduces technique is given to know the scientific data analysis and they are two different types, a) high energy physics data analysis b) k-means clustering. There they go to second stage of data intensive scientific analysis and it is the CGL-map reduce, a streaming based map reduce implementation. In [6] M. Sato, Hansen, J. Glascoe, and worked on the analysis of the GISS surface temperature changes research on the temperature changes over the land, areas, and seas based on measurements. In [7] Jeffery Dean

and Sanjay Ghemawat discuss the flexibility of the data processing tool in hadoop. In this paper the author specifies the map function and the basics task carried out by the amp reduce model.

In [8] some of the disadvantages of the map reduce programming module is discussed. In this article the authors firstly describe in detail the concept of map reduce. The working procedure of the MR framework, then describe the map reduce as a step backward in database accesses as the schemas are not good, high level access languages are good and many more. In [9] comparing the different approaches to handle the big data in shown in this article. A test is performed where 100 node of cluster is used to compare the analysis. To generate the data there are two steps that needed to be performed, first, find the plain text of data sets, and second is finding the hyper text mark-up language. In [10] provides a process to manage the large scale sensor network with very effectively and efficiently and also defining a cloud infrastructure that will make use of Map Reduce at the processing layer and Big table at the data layer. It also describes the basic concepts of the continuously changing data object. In [10] provides a process to manage the large scale sensor network with very effectively and efficiently and also defining a cloud infrastructure that will make use of Map Reduce at the processing layer and big table at the data layer. It also describes the basic concepts of the continuously changing data object.

II. METHODOLOGY

1. BIG DATA

The increasing the capacity of the traditional database and aggregating the information mainframe is the strategy of the Big Data. Handling the structured data is very easy but, when comes to unstructured data, the solution to it is Big Data, because it provides the ability to process the unstructured data. Big Data uses the Hadoop to store the larger datasets because to analyse these sets require new working procedure which will run in parallel and on the distributed environment. As Big Data can be defined in many ways, but we define here by using the three V's which are Volume, Variety, and Velocity.

2. Hadoop

Working with the large datasets is not quite easy hence the big data uses the Hadoop concept which is a programming ground wok which will allow operating the large volume of datasets. Hadoop works in distributed and parallel environment and for storage it uses the distributed fils systems (DFS). The distributed file systems are more complex than regular disk file systems. There are three different modules in this and they are Hadoop Distributed file system (HDFS), Hadoop Map Reduce, YARN.

3 . HDFS

This type of the file system is used to store the large amount of the data and process them on different operating environment.

4. Hadoop Map Reduce

It is also called as Map Reduc1. It consists of the four independent entities such as the client: which submits the given map reduce jobs. The Job Tracker: It wills co-ordinate the job run. The Task tracker: is responsible to run the tasks given to it. The DFS which is used for sharing job files between the other entities.

5 . YARN

Handling an array of technology is the work done by this YARN. The full form of the YARN is yet another resource negotiator. It is also called as Hadoop MR2. The YARN is responsible to manage the distributed application on the commodity hardware.

6. JAVA

The one of the most working technology which is being used now a days is java. The java programs run on the java virtual machine (VM). The advantage of using this technology is that is platform independent. Java provides many methodologies and its object oriented programming languages proved to develop the applications. Some of the features of java are listed below:

7. IOT

The internet of tings is the network of devices which are embedded with the electronics software, sensors, which also allows these object to allow and exchange the data. IOt mainly refers to devices other than the computers.

8. ARDUINO

It is a software company which provides the projects, and customer association that designs and manufactures computer open source hardware and open source software. The integrated development environment (IDE) for the Arduino is written in java and is a cross platform. It is mainly used to build the projects.

III. SYSTEM ANALYSIS

Collecting, warehousing and handling the huge amount of weather datasets using the tradition analysis requires lots of time and is very difficult the data. In the traditional approach the processing the data that is generated through the sensor is very difficult. As the number of sensors increase the amount of the data become high in volume and the sensor data usually contains the high velocity of data. Therefore we need a scalable analytic's tool to process this massive amount of the data.

Collecting only the sensor data is not important analysing these data is very important and useful. The existing system processes the sensor data but on the single hardware, in return there we get the loss of data and lots of time consumption. In our project we are mainly concentrated on the temperature of city Kalaburagi, with the existing system its it very tough to provide the maximum temperature.

Disadvantages of the existing system

The data is not processed on multiple systems, and hence. Increasing in traffic and loss of data. It takes lot of time to process the data. Only offline data is processed.

Proposed System

The propose system overcome s the issues of the existing system. In this project we use the concept of IOT and Hadoop to propose the new system for weather data analysis. The hadoop mainly consists of two methodologies which helps us to obtain the required results. The HDFS and the map reduce are the two methodologies used here. The HDFS will distribute the data on different system through which we are able to improve the efficiency and time. A new temperature data analytical engine is proposed which works with Map reduce and hadoop producing results without scalability bottleneck. Map reduce takes the input in the form of text, this input text will be online and offline. The online text data is obtained from using hardware such as Arduino mega, temperature sensors, an IOT board, Ethernet shield connected to wifi modem. Through these we gain the temperature of the city Kalaburagi. And find its MAX and MIN temperature. This is done by using the two methodologies of hadoop.

Advantages of Proposed System

Provides the maximum and minimum temperature of years and months. By using the iot online and offline data, it predicts the upcoming years weather forecast. As it runs on commodity hardware, so there is reduction in traffic and loss of data. An visualize the data for every minute.

IV. BLOCK DIAGRAM OF SYSTEM

The Flow chart provides the graphical representation to know how the data will pass in the system. Similar to this is the Data Flow Diagram which also initial step to know the environment working procedure. The Data Flow Diagram will represent process that change the data, data storage locations, and the data will flow by itself.

The figure 1 gives the flowchart for collecting the online data

And storing the data through the IOT

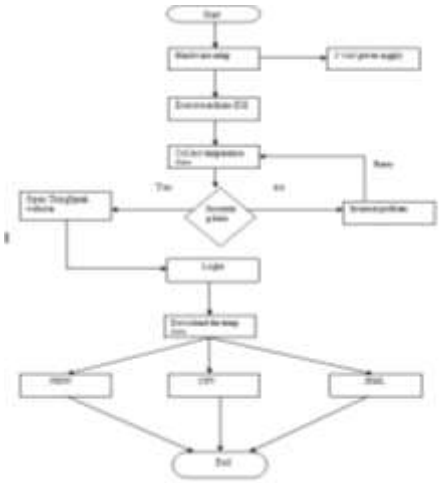

Fig 1 Data flow for collecting the online real time temperature data.

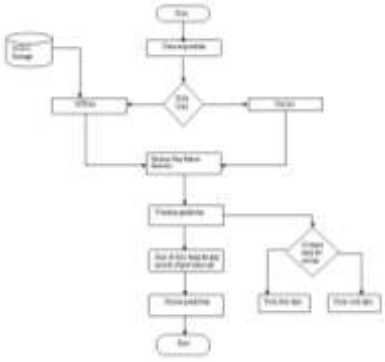

Fig 2 Flowchart for the proposed system

The figure 2 gives the flowchart for the proposed system. It gives the flow in which the system will perform the operation to do the weather analysis for offline and real time online data.

A. System Architecture

It gives the complete description of the work. The correct system architecture will help us to avoid many problems and hence it has become a very important part of a project because the architecture will provide the visualization of the project and expected output. The figure 3 shows the system architecture of our work. The infrastructure for the big data requires the special testing environment as it stores large data and files (HDFS). Some challenges occur during the system architecture and they are:

- Sampling strategy in big data is a challenge.
- Big data has no particular tools hence its range varies from programming tools like map reduce to Hive QL
- Saving a large datasets, it becomes necessary to be able to test across the multiple platforms.

In our work we have used the three modules and they are:

B. DATA ACQUISTION UNIT (DAU)

As said above this is the first module to be implemented in our work so we begin with it. In this in the data acquisition unit we collect the data so that we can get the expected output i.e., Max and Min temperature and predict the future weather forecast. As described earlier, in our work we are using the both offline data and online data.

The online data is nearly equal to real-time data. In our work, we mainly focus on the temperature data, so we use the temperature sensors. As known the sensors provide the data for every second, hence storing this data is very difficult. So we use the IOT board Arduino, which is used to transfer the data to cloud storage through the internet. The free usage for the cloud is provided by the Thing Speak. As the name its self indicates IOT i.e., Internet of Things, everything works on internet, so we need the internet connection to our device. This internet connection is made to our device by using the Ethernet shield mounted on the Arduino device and one end of LAN wire is connected to Ethernet shield the and other end is connected to the modem. If the DHFS connected then the transferring of the data will be done, if it fails then there is some error in the internet. During this, if it fails, though the sensor will display the temperature data but will not save the data anywhere.

The transfer of data from the device to ThingSpeak there is two important functions. First, when the user creates the logins Id in the ThingSpeak website then a new channel is created, after creating a channel the user is provided by an API key. This API key is very important, because this will act as the mediator to transfer the data from the device to the cloud storage. Second, the next function is to execute the IDE of the Arduino. In the IOT Arduino board there is a chip within which the code to transfer the data to cloud storage of ThinkSpeak is written.

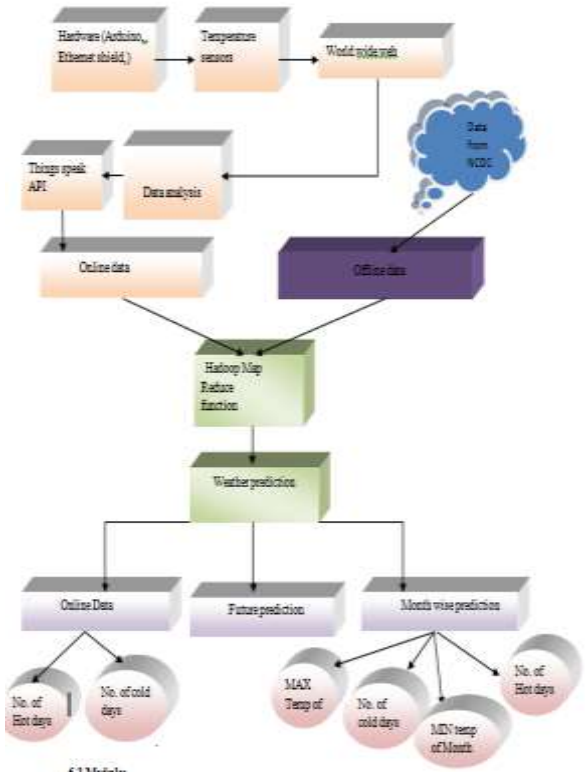

Figure 3 System Architecture for the system.

During this the API key which is got during creating a channel is also added to code. Hence then the data generated by sensor is stored in the cloud, and this can be downloaded in three form JSON, CSV, XML. In our work we are using the temperature data of city Gulbarga in Karnataka The offline data being used is taken from the National Climate Data Centre (NCDC). The NCDC provides the weather data sets. The global weather measurement is one of the biggest dataset available and is collected every day. The NCDC is located in the United States of America and is world's largest active archive for climate data. It provides the data over more than 150 year, every day 224 gigabytes of data is stored.
In the proposed system the temperature dataset of NCDC used is of city Austin and years 2013, 2014, 2015. These records are stored in HDFS.

C. DATA PROCESSING UNIT (DPU)

In the previous module there was a detail discussion about the acquiring the data, next is to know the procedure to process these data. This module mainly works on the processing the data. The main function used to process in big data is Hadoop Map Reduce.

Map Reduce works in two stage: in first stage the mapper takes the datasets and divides it into the <key, value> pairs and send this to second stage i.e, reducer. The work of the reducer is to provide the list of the <key, value> pairs of

the datasets as output by removing the unwanted data from the dataset. This will reduce the time to remove the unwanted data. In our work this is used to reduce the dataset obtained from the NCDC because it contains some unwanted temperature data. Hence it provides the correct output data for whole year. The map Reduce will work in parallel way which will reduce the user's time when compared to traditional approach. Adding more number of systems to the network will speed up the entire data process. This is the main advantage if using Hadoop Map Reduce framework. To run the hadoop in windows there are many commands to run in command prompt, some of them are:

- Creating a namenode i.e., type hadoop namenode –format in command prompt.
- Starting the Yarn and DFS, this will start the hadoop execution by typing start –yarn and start-dfs.
- Compiling the proper path to run the map reduce job.

To know that whether the executed path works properly, after compiling the path on th browser the address local host: 50070 is typed and a window occurs for hadoop showing that its working properly. Once working properly we get the reduced output in the form of part file, which is downloaded and sent to find the Max and Min temperature.

D. Data Analysis and Decision Unit (DADU)

Calculating the Max, Min and prediction of future weather forecast is the output of our work. In this module e make an analysis on data and find the maximum temperature and minimum temperature of datasets, and predict the future weather based on these. To do so we use the java classes, this helps to find an average on max and min temperature. On this we predict that whether in next year the number of hot days will be more or cool days.

V. HARDWARE IMPLEMENTATION

i. Hardware Setup

Some of the basic hardware are required. In our work the basic hardware used are:

- Board: Arduino/Genuino Mega 2560 with USB cable.
- W5100 Ethernet Shield for Arduino Uno and Mega.
- Sensors: Temperature sensor LM35.
- Connectors: Male to Male, Male to Female, Female to Male.
- Power for IOT 12v.
- Breadboard.

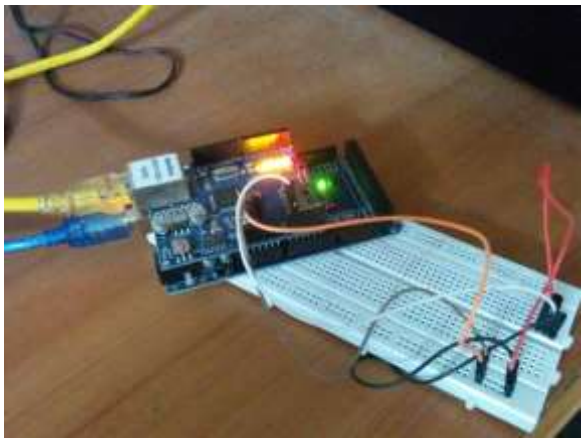

Fig 4 The Hardware set up of Arduino

ii. Configuring the Sensor with Arduino

The sensor used in our work is LM35 which is used for capturing the temperature. The sensor has three pins:

- Vcc
- Output
- GND

As shown in figure 5 the vcc is connected to the Arduino board to the place of 5v, the output is connected to the first place of the board and the GND is connected to the ground of the board. The sensor is switched to the breadboard and from there using the connectors it is connected to the Arduino.

Figure 5 Connecting the LM35 to Arduino board and the Breadboard.

iii. Starting With ThingSpeak

a. ***Channel Creation:*** The ThingSpeak can present a flawless connectivity of Arduino from internet. The next step is to create an account at Thingspeak. After having an successful creation of an chronicle, snap to the channels and then select My Channels as shown below.

Figure 6 creating a new channel at ThingSpeak.

The Thingspeak Channel is similar to the storing the data in database. We can define it as, a channel is where the data is stored or sent. Every channel in thingspeak website has maximum of eight fields, in which 1 is allotted to status field and 3 are allotted to location field. Once the channel is created, it will publish the data, process the data and also allow retrieving the data. ThingSpeak allows visualizing the live weather data, location of the particular weather.

VI. RESULTS

The experimental set-up is tested for various test cases, and different test give different results which help in improving the performance of the system. The below shows the results of our work. The proposed system uses the temperature datasets of 2013, 2014, 2015 from NCDC and the real time data which is collected through the IOT and sensors and using the thingspeak to store this collected data. The NCDC records are stored in the HDFS and perform map reduce function. Map reduce execution is shown in fig below "the results shows adding more number of systems to the network will speed up the entire data processing". This is one of the major advantage of the map reduce with hadoop frame work.

Figure 6 Visualization of sensor data

The figure 6 shows the visualization of temperature data. It gives the graph for each temperature data for every time it is collected. This data is taken from the sensor connected to Arduino device and sent to ThingSpeak which will store in the cloud and also gives the graph for each data as shown in figure.

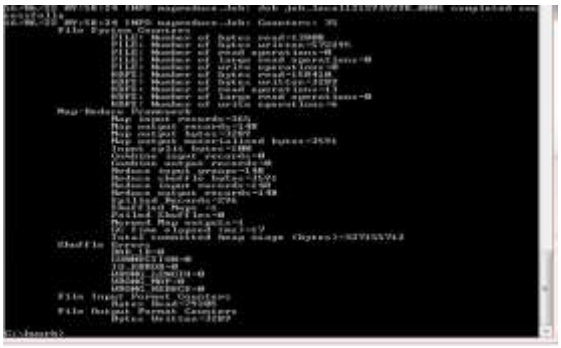

Figure 7 execution of Map Reduce on the Austin.txt data

Figure 8 weather predictions for hadoop based on month wise.

Figure 9 Number of hot days for year 2015 of city Austin.

Figure 10 predictions for upcoming year.

VII CONCLUSION

In traditional system, the processing of billions of records is quite time consuming. We live in the era of IOT where things work over the internet, to get the temperature, humidity the sensors are used by the different departments. One of the effective solutions for this is using Map Reduce

with Hadoop which will allow an easy analysis of the sensor datasets. Collection of the temperature, moisture, and humidity values is becoming quite easy by using IOT. This will provide the real time data and storing this to the cloud storage is done through the website ThingSpeak. The datasets are downloaded and is being used in the Map Reduce programming model. The hadoop Map Reduce will remove the scalability bottleneck. As it works in parallel and distributed environment it reduces time consumption. The use of such technologies for the large scale data analysis has potentially greater enhancement to the weather forecast.

Hence in our system we predict the future weather forecast for the data obtained from the NCDC, based upon the maximum and minimum temperature and number of hot days and cold days. These help the people in preplanning many outdoor events, in agriculture and provide the necessity precautions towards the weather. The use of IOT has gained lot popularity and will also gain more in future.

REFRENCES

[1] D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and cloud computing: Current state and future opportunities," in Proc. Int. Conf. Extending Database Technol. (EDBT), 2011, pp. 530–533.

[2] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: Mc Graw-Hill, 2012.

[3] Real-Time Big Data Analytical Architecture for Remote Sensing Application Muhammad Mazhar Ullah Rathore, Anand Paul, Senior Member, IEEE, Awais Ahmad, Student Member, IEEE, Bo-Wei Chen, Member, IEEE, Bormin Huang, and Wen Ji, Member, IEEE

[4] Hansen, James, and Sergej Lebedeff. "Global Trends of Measured Surface Air Temperature." J. Geophys. Res., 92,1987: 13,345-13.

[5] Jaliya Ekanayake and Shrideep Pallickara, MapReduce for Data Intensive Scientific Analyses, eScience, ESCIENCE '08 Proceedings of the 2008 Fourth IEEE International Conference on eScience, Pages 277-284, IEEE Computer Society Washington,DC, USA ©2008.

[6] Hansen, J., R. Ruedy, J. Glascoe, and M. Sato. "GISS analysis of surface temperaturechange." J. Geophys.Res.,104, 1999: 30,997-31,022.

[7] Jeffrey Dean and Sanjay Ghemawat. MapReduce: A Flexible Data Processing Tool. Communications of the ACM, 53(1):72–77,January 2010.

[8] Pavlo, A., Paulson, E., Rasin, "the disadvantages of the map reduce programming"

[9] A., Abadi, D.J., DeWitt, D.J., Madden, S., and Stonebraker, M. A comparison of approaches to large-scale data analysis. In Proceedings of the 2009 ACM SIGMOD International

[10] Improving Numerical Weather Prediction Models and Data-Access LatenciesNOAA National Climatic Data Center (NCDC),Daniel J. Crichton NASA Jet Propulsion Laboratory (JPL)

[11] Hemlata Tomer*,Kapil Mangla "Study and Development of Temperature & Humidity monitoring system through Wireless Sensor Network (WSN) using Zigbee module. Hemlata Tomer Int. Journal of Engineering Research and Application www.ijera.com ISSN : 2248-9622, Vol. 5, Issue 7, (Part -2) July 2015, pp.115-120"

