Big Data Analytics : A Powerful Tool for Weather Forecasting

Aishwarya Chipkar, Dr. P. R. Gundalwar

MCA Scholar, Associate Professor
Master of Computer Application,
MET ICS, Mumbai, India

Abstract : The contribution of this paper is to provide an analysis of the available literature on big data analytics in weather forecasting. Accordingly, some of the various big data tools, methods, technologies and algorithms which can be applied are discussed, and their applications and opportunities provided in several decision domains are portrayed.

***IndexTerms* - Big Data Analytics, weather forecasting, map reduce algorithm, weather prediction**

I. INTRODUCTION

Every single day lots of data is produced globally and warehoused by public administration and private firms. Besides, cities are filled with sensors accumulating different types of information about weather, telephony, traffic. Big Data Analytics is an idea that groups all those technologies and mathematical developments committed to store, analyse and cross-reference all that information to try and discover behavioral patterns.

Refined software programs are used for big data analytics, but the unstructured data used in big data analytics may not be well matched to conventional data warehouses. Big data's excessive processing requirements may also make out-of-date data warehousing a poor fit. As a result, newer, bigger data analytics environments and technologies have emerged, including Hadoop, MapReduce and NoSQL databases. These technologies marked open-source software framework that's used to process huge data sets over clustered systems. [1]

II. TRADITIONAL METHODS USED IN WEATHER FORECASTING

Before the upsurge of technology, a lot of native people everywhere round the world, could precisely forecast weather and climate around the country by taking note of variations in the environment and observing animal behaviour.

It was not until in the early 1800s the electric telegraphy invention gave birth to the modern weather forecasting techniques.

After that in the 20th Century the numeric predictions were used in the prediction of the weather. Applied use of numerical weather prediction began in 1955 encouraged by the development of programmable electronic computers.

Numerical weather prediction (NWP) uses mathematical models of the atmosphere and oceans to predict the weather on the basis of current weather conditions. Though first endeavored in the 1920s, it was not until the introduction of computer simulation in the 1950s that numerical weather predictions produced convincing results. A number of global and regional forecast models are run in different countries worldwide, using current weather observations reported from radiosonde's, weather satellites and other observing systems as inputs.

Then are the Statistical methods used together with the numerical weather prediction. This method often supplements the numerical method. Statistical methods use the past records of weather data on the supposition that future will be a recurrence of the past weather.

The main purpose of studying the past weather data is to find out those characteristics of the weather that are good indicators of the future events. After establishing these relationships, correct data can be safely used to predict the future conditions.

Only overall weather can be predicted in this way. It is particularly of use in projecting only one aspect of the weather at a time. For example, it is of great value in predicting the maximum temperature for a day at a particular place.

The procedure is to compile statistical data relating temperature to wind velocity and direction, amount of cloudiness, humidity, and to the specific season of the year. Thereafter these data are depicted on charts. These charts provide an estimation of the maximum temperature for the day from the data of the current conditions. Statistical methods are of great value in long-term weather forecasts.[2]

III. LITERATURE REVIEW

Big Data pursues the conveyed methodology and various wellsprings of huge information can be given as SNS, Distributed computing, Large Processing Technology and Portable Data.

The big data architecture typically divided in three segments Storage, Processing and Analysis.

As indicated by Biswas Ranjit, to manage enormous information (growing quick in 4Vs: Volume, Variety, Speed and Veracity) he has presented 4 N's-

- New intelligent and physical capacity structures,
- New heterogeneous information structures,
- New scientific hypotheses
- New models;

According to him, existing data structures or architecture of existing distributed systems isn't equipped for taking care of this expediently expanding and evolving information.

Today, NOMADS NOAA Operational Model Archive and Distribution System) gives access to roughly 650 terabytes of model information organized in excess of 50 million documents on tape. Roughly 1.2 million of these documents are organized on disk for prompt access, and a sub-set of those are pre-amassed by factors in a by variables in a time-series to promote easier access for users. A few information need to be pre-arranged from tape before access — a very slow and monotonous procedure.

Current technology trends is focused on the utilization of distributed computing and advances such as Apache Hadoop as "big data" solutions. Quite a bit of this, nonetheless, accept specific use cases and calculations that are streamlined for these sorts of environments and technologies.

According to Bryson, "Weather is the original big data problem". Size of initial data increases, accuracy of forecasting increases. Nick Wakeman with reference to Hurricane Sandy expressed in his blog the significance of Big Data in climate prediction. With the help of available data, three-days out, forecasters predicted within 10 miles where landfall would occur.

According to Nancy Grady the velocity of weather data plays role in the development of economy as a case. This weather data can be used by combining it with other disciplines which can generate new opportunities to businesses.[3]

IV. Big Data Analytics used in Weather Forecasting

The rise of Big Data is now history. What is of importance is exactly how organizations develop the tools and resources necessary for reacting to, and exploiting the progressively available Big Data for their advantage. In line with this, notes that there is a necessity for the implementation of powerful tools such as Data Mining techniques which can support in moulding the complex relationships that are in-built in Big Data.

There exists a common belief that Big Data can support in improving forecasts only if that we can analyse and determine hidden patterns, and come to an understanding that predictions can be improved through data driven decision making.

The opportunities for growths through forecasting with Big Data are diverse. At present, there is augmented research into using Big Data for gaining precise weather forecasts and the initial results recommends that Big Data will profit weather forecasts massively.

Increasing indication of climate change worldwide is prompting governments and scientists to take action to protect people and property from its effects. But, to take effective action, they need to understand a lot more about the weather–everything from what's going to happen tomorrow to what's coming next year.

The financial potential of improved weather prediction through big data is huge. According to Weather Analytics, 33% of worldwide GDP is affected by the weather. For a whole host of business and public sector organisations, weather predictions can make the difference between profit and loss - or even between life and death. It mainly affects industries such as agriculture, tourism and fishing. So whether the issue is ice cream sales or flood defenses, the weather matters. [4]

V. ALGORITHM ROLE IN BIG DATA ANALYTICS

In the Figure 1 given below, we will see how the flow of the algorithm that should be followed for weather prediction using the Big Data Analytics tools.

Step 1: There are sensors installed everywhere in the cities which is constantly sensing and collecting data. Acquire Data component will consist of the data which will be acquired from the sensors.

Step 2: Data pre-processing generally involves the following steps:

Smoothing of noisy data.

Aggregating your data - your data will likely be collected by different recording devices simultaneously, potentially at different temporal or spatial resolutions, and will therefore need aggregating into the same tables or matrices, potentially with appropriate subsampling.

Imputing missing values - Taking the time to perform proper error handling for missing values or NaNs ("Not-a-Number") in your analysis scripts can save you hours of debugging further down your analysis pipeline.

Removing erroneous data points.

Step 3: Next step will be to develop weather prediction model and weather forecasting models. Both these models have different meaning and goal. Prediction is concerned with estimating the outcomes for unseen data. For this purpose, you fit a model to a training data set, which results in an estimator f^(x) that can make predictions for new samples x. Forecasting is a sub-discipline of prediction in which we are making predictions about the future, on the basis of time-series data. Thus, the only difference between prediction and forecasting is that we consider the temporal dimension. An estimator for forecasting has the form f^(x1,…,xt) where x1,…,x^t indicate historic measurements at time points 1,…,t while the estimate relates to time point t+1 or some other time in the future. Since the model depends on previous observations, xi, this is called an autoregressive model.

With these definitions, we can now appreciate why

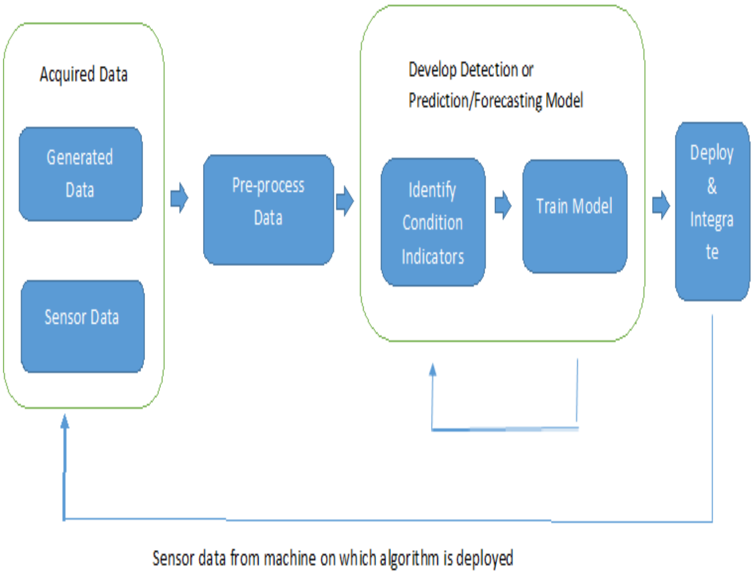

Fig. 1 Process Framework

Weather forecasting is not called weather prediction: weather forecasting predicts the weather in the future using temporal information. For example, if there is a downpour at the moment, what is the likelihood that it will still rain in five minutes? Independent of all other features that influence the weather (e.g. atmospheric pressure and temperature), the likelihood that it will still rain in five minutes will be high because it is raining at the moment.

Imagine that we would perform weather prediction rather than forecasting. This would mean we were to ignore any temporal dimension and just consider the other physical features that influence the weather. Imagine that is still raining outside. Oddly the atmospheric pressure i quite high, which is associated with clear skies. So, when you are asking your prediction model to estimate whether it is currently raining, the model would probably respond that it is unlikely to rain due to the high atmospheric pressure.

Step 4: Finally we will perform the deployment and Integration of the data that will be obtained after the processes are performed on it. For deploying objects from one environment to another, the objects must be added into containers called Applications. The application archive file can subsequently be deployed to data integration services in the same or different domain

Data integration is the process of combining data from many different sources, typically for analysis, business intelligence, reporting, or loading into an application.[5]

VI. EXPERIMENTAL SETUP USING MAP REDUCE ALGORITHM FOR WEATHER FORECASTING

MapReduce is the programming model that allows Hadoop to efficiently process large amounts of data . MapReduce breaks large data processing problems into multiple steps, namely a set of Maps and Reduces that can each be worked on at the same time (in parallel) on multiple computers.

MapReduce Algorithm uses the following three main steps Map Function, Shuffle Function and Reduce Function.

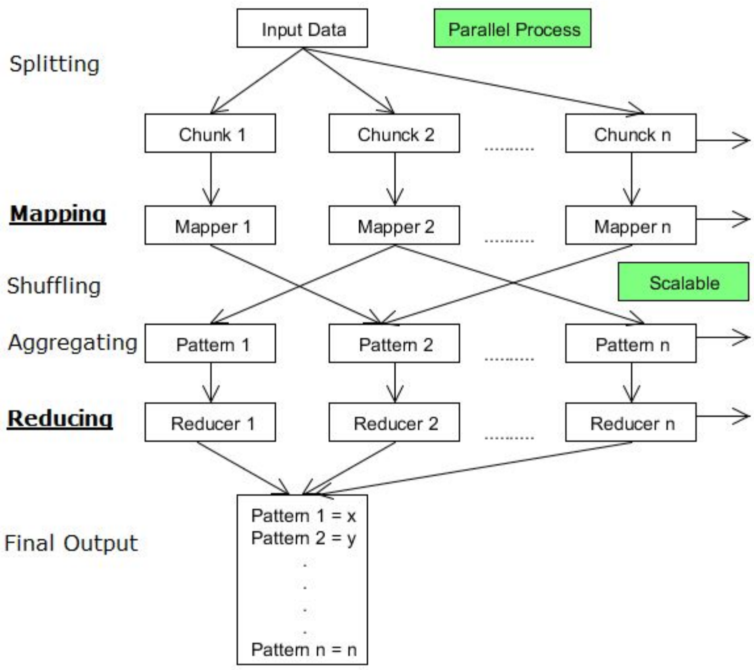

Fig 2. Map Reduce Algorithm

Map Function is the first step in MapReduce Algorithm. It takes input tasks (Datasets) and divides them into smaller sub-tasks. Then perform required computation on each sub-task in parallel. This step performs the following two sub-steps: Splitting step takes input dataset from source and divides into smaller SubDatasets. Mapping step takes those smaller (SubDatasets) and performs required action or computation on each Sub-Dataset. The output of this Map Function is a set of key and value pairs.

Shuffle Function It is the second step in MapReduce Algorithm it takes a list of outputs coming from "Map Function" and performs these two sub-steps on each and every key-value pair. Merging step combines all key value pairs which have same keys (that is grouping key value pairs by comparing "Key"). This step returns >. Sorting step takes input from merging step and sorts all key-value pairs by using Keys. This step also returns > output but with sorted key value pairs. Finally, Shuffle Function returns a list of > sorted pairs to next step.

Reduce Function it is the final step in MapReduce Algorithm. It performs only one step: Reduce step. It takes list of > sorted pairs from Shuffle Function and perform reduce operation as shown Fig.2.

The Hadoop tool 2.7.1 is used to carry out the analysis of the weather data using MapReduce algorithm. The cluster consists of three Linux machines, where the master (CPU Intel Core i7, RAM 4 GB and HD 1 TB) have two mission's management the cluster and at the same time working as a slave node (DataNode) and the other two machines (CPU Intel Core i3, RAM 4 GB and HD 1 TB).

The Master node assigns new tasks to worker nodes and reassigns tasks that take too long. A sequence overview is as the following steps. Step1: The input weather data is split into a number of pieces of a specified size (64 MB). The weather algorithm is started on all nodes. Step2: one node is set to be Master and DataNode starts delegating work to other nodes. All pieces created in Step 1 are first mapped by the mapping function. The number of reduce tasks at the start should be low. Step3: If a worker gets a map task, it runs the map task and stores the result in the memory of the machine. Step4: Periodically these stored results are written to the disk and the Master node is notified of the action. Step5: When the Master node gets notified about a location of mapped pairs, it will start a reduce test on one of the free workers. Step6: When a reduce task is called, first of all it fetches the stored results from the remote machine on which the map task has run. Secondly, these results are sorted by key. Thirdly, the results are reduced. Step7: When there are no more data to process, the Master node returns the final results to the user program. All this time the Master node has an overview of what all nodes are doing. The master will also re-assign already assigned tasks to idle nodes, because this might improve overall performance.

VII. CONCLUSION

In this study, big data analysis is used for predicting the temperature using an algorithm that can be used to predict the weather. Through the implementation of the Process framework, Map Reduce algorithm can be used as the prediction algorithm to predict the temperature. This approach is able to determine the nonlinear relationship that exists between the historical data (temperature, wind speed, humidity, etc.,) supplied to the system during the training phase and on that basis, make a prediction of what the temperature would be in future.

VIII. REFERENCES

[1] https://www.techopedia.com/definition/28659/big-data-analytics
[2] https://en.wikipedia.org/wiki/Weather_forecasting
[3] Data Structure for Representation of Big Data of Weather Forecasting: A Review Dhanashri V. Sahasrabuddhe , PallaviP.Jamsandekar
[4] https://link.springer.com/article/10.1007/s40745-015-0029-9
[5] https://www.datascienceblog.net/post/machine-learning/forecasting_vs_prediction/
[6] A Big Data Prediction Framework for Weather Forecast Using MapReduce Algorithm Khalid Adam1 , Mazlina Abdul Majid1 , Mohammed Adam Ibrahim Fakherldin2 , Jasni Mohamed Zain
[7] http://www.thisissanjeeva.com/tech/bigdata/2017/07/mapreduce-unwinding-algorithm#comments