



Rainfall Analysis and Rainstorm Prediction using MapReduce Framework

T. Manojpraphakar

PG Scholar,

Dept of CSE, Sri Shakthi Institute of
Engineering and Technology, India

C. P Shabariram

Asst. Prof,

Dept. of CSE, Sri Shakthi Institute of
Engineering and Technology, India

Abstract—Rainfall data is collected to predict the storm warnings from the hydrological data. This is considered as a research idea as it consumes huge number of records from the distributed system. This paper describes a novel solution to manage the data based on spatial temporal characteristics using a Map Reduce Framework. The workload is classified using support vector machine(SVM). Various rainstorm concept prediction is achieved using the big raw rainfall data. The dataset impact parameters are classified into local, hourly, and overall storms. The proposed system serves as a tool for predicting rainstorm from a large amount of rainfall data in a efficient manner. The result indicates the proposed system improves the performance in terms of accuracy and efficiency.

Keywords—Storm analysis, MapReduce, Rainfall, hydrological data

I. INTRODUCTION

Big data is collection of huge volumes of data that contains both the structured and unstructured data that is difficult to store analyze process, share, visualize and manage with the traditional database and software techniques. Volume of data can be calculated by the amount of transactions. Due to growth increasing need on platforms and in many software industry applications to handle the scalability, accuracy, rate at which enterprises remain to face in a competitive global Market world .Major Big Data challenges are capturing data, storage, transfer, searching, analysis, transfer, presentation. Along with traditional transactional and analytics data stores, we now collect additional data across social media activity, web server log files, financial transactions and sensor data from equipment in the field.

II. HADOOP

Hadoop is a open source software which is developed using Java programming that helps in accessing the huge collections od datasets in a distributed manner. It is developed and managed by apache hadoop, Hadoop framework uses the MapRedue algorithm that helps the data is analysed in parallel. It stores any type of data in its own format and performs the analyses and changes on the data. Hadoop stores the information ranging from tera to even petabytes of data. It is efficient and reliable and handles the hardware failure automatically when the system malfunction occurs,with out any loss of data. The two components of hadoop are: 1)Hadoop Distributed File System(HDFS), 2) MapReduce. Both the HDFS and MapReduce are designed to continue to work in the face of System Failures. Hadoop runs code across a cluster of computers. Data are divided into directories and files. Files are divided into uniform blocks 128M and 64M. Servers can be added and removed from the cluster dynamically and hadoop operates to continue without any interruption

III. MAPREDUCE

MapReduce is a functional programming model for data processing. Hadoop can run MapReduce programs written in various languages namely Java, Ruby, Python, and C++. MapReduce processing consists of two phases: the map phase and reduce phase. Each phase of MapReduce consists of a key-value pairs as input and output. Hadoop divides the input to a MapReduce job into the fixed-size pieces called input splits

Map() Function: $(k1,v1) \rightarrow list(k2,v2)$. It performs filtering and sorting of task into queue.

Reduce() Function $(k2,list(v2) \rightarrow list(v2)$. It performs a summary operation of best candidate resource for task execution.

The MapReduce operations is based on shuffle, sort and reduce.

IV. STORM RELATED PARAMETERS

National Weather Service(NWS) Dataset:

The NWS is an agency of United States that provides the weather forecasts and other storm related warnings given to organizations for their prior purpose of protection against the disasters. They provide the detail information of data(ie) from a period of twenty years. Following are the dataset parameters that are required for predicting the storm related data. The is available from National Climate Data Center (NCDC, <http://www.ncdc.noaa.gov/>)

Local Storms:

Local storms are based on the site specific storms, that considers only the location. Local storms are the set of storms that occurred at a site location. Suppose if there are two different local storms are occurred at the same location is predicted by time[8]. It is researched by most hydrologists and provides the location -based storm functionalities.

Hourly Storms:

Formally a hourly storms is a time specific storm, which frequently analyses each hour. The number of sites covered by an hourly is said to be storm coverage [1]. And storm average is calculated by dividing the storm sites covered by the hourly storm. For example hourly storm is set occurred between 6.00 am to 10.pm. In other words local storms considers only a particular location where as a hourly storm fixes a time interval. National Weather Service (NWS) hourly rainfall data for a period of record includes number, name, latitude, longitude [8]. This hourly storms are difficult to predict, when it is based on whole year period.

Overall Storms:

Overall Storms is the combination of both the local and a hourly storms (ie) location and time when analyzing the storms. It captures the Storm-centric prediction that contains the storm center, motion and a direction. The overall storms contains storm overall depth, storm intensity, storm average [1].

Location-based Storms:

These are the storms, which is based on the locations that is based on the following parameters:

- Episode-id: Indicate whether the storm is affected or not. It may contain different events. Ex:60904(ID assigned by NWS to denote the storm episode. An Episode may contain several different events
- Event-id: The id assigned by NWS to note a single small part goes into a specific storm episode.
- Location- index Ex: 1-x a number is assigned by NWS to identify the particular locations within the specified storm event. Each event's sequentially increasing location index number will have a corresponding latitude/longitude point
- Range Ex: 0.59, 0.69, 4.84, 1.17 A hydro-meteorological event is referenced, minimally, to the nearest geographical center (not from the village/city boundaries or limits) of a particular areas indicating that the point is referenced in the Storm Data software location database.
- Azimuth Ex: E,W,N,S a 16-point compass direction from a particular areas providing that the reference point is documented in the Storm information software location database of > 150,000 locations.)
- Location Ex: PINELAND, CENTER, ORRS, RUSK the range and the position of azimuth is calculated.
- Latitude Ex: 31.25, 31.79, 32.76, 31.80 .The latitude where the event occurred
- Longitude Ex: -93.97, -94.18, -94.52, -95.13. The longitude where the event occurred

Event related storms:

These storms specify the storms that are related to events that includes as follows:

- Last_date_modified The last date of modification by NWS. Any modification to the storm episode is made solely by NWS who is authorize to modify it.
- Last_date_certified The last date of certification by NWS. Any modification to the storm episode are made solely by NWS who is authorize to modify it.
- Episode_id Ex: 61280, 62777, 63250
- The occurrence of storms and other significant weather phenomena having huge intensity to cause loss of life, injuries, significant property damage Other significant hydrological events, such as record max temperatures or precipitation that occur in connection with another event.
- Event_id Ex: 383097, 374427, 36417 ID assigned by NWS to note a single, small part that goes into a specific storm episode;
- State Ex: GEORGIA, COLORADO The state name where the event occurred.
- Year Ex: 2000, 2006, 2012 Four digit year for the event in this record
- Month_name Ex: January, February, March Name of the month for the event in this record
- Event_type Ex: Hail, Thunderstorm Wind, Snow, Ice.The chosen event name should be the one that most accurately describes the meteorological event leading to injuries, damage, etc.
- cz_type Ex: C, Z, M Indicates whether the event happened in country,zone,marine.
- cz_fips The county FIPS number is assigned to the county by NWS.
- cz_name County/Parish, Zone or Marine Name assigned to the county FIPS number or NWS Forecast Zone
- Begin_date_time Ex: 4/1/2012 20:48 M/DD/YYYY 24 hour time AM/PM
- cz_timezone Time Zone for the County/Parish, Zone
- End_date_time Ex: 4/1/2012 21:03 MM/DD/YYYY 24 hour time AM/PM

- Injuries_direct Ex: 12,18,32.The number of causes directly related to the weather event
- Injuries_indirect Ex: 0, 15, 87.The number of causes indirectly related to the weather event
- Deaths_direct The total number of deaths related to the weather causes.
- Deaths_indirect Ex: 0, 4, 6.The number of deaths indirectly related to the weather event
- Damage_property Ex: 10.00K, 0.00K, 10.00M.The total amount of damage to property incurred by the weather event.
- Damage_crops Ex: 0.00K, 500.00K, 15.00M.The total amount of damage to crops incurred by the weather event.
- Magnitude Ex: 0.75, 60, 0.88, 2.75.extent measure of the magnitude type 0.75 indicates hail speed and 60 mph indicates wind speed.
- Magnitude_type ES = Estimated Sustained Wind; MS = Measured Sustained Wind (no magnitude is included for instances of hail)
- Flood_cause Ice Jam, Heavy Rain, Heavy Rain

Measuring the Torando Wave:

The following are the characteristics that are used to measure the tornado wave.

- tor_f_scale Ex: EFS0, EFS1, EFS2, EFS3, EFS4, EFS5
- Enhanced Fujita Scale describes the strength of the tornado amount of damage caused by tornado. The F-scale of damage will vary in the destruction area; the value which has the highest value
- EFS0 – Light Damage (40 – 72 mph)
- EFS1 – Moderate Damage (73 – 112 mph) etc.
- tor_length Ex: 0.66, 1.05, 0.48. Length of the tornado or a segment.
- tor_width Ex: 25, 50, 2640, 10.Width of the tornado or tornado segment while on the ground (in feet)
- tor_other_wfo Ex: DDC, ICT, TOP, OAX. Indicates the continuation of a tornado segment
- tor_other_cz_state Ex: KS, NE, OK. The two character representation for the state name of the tornado segment as it moves from one county to another one
- tor_other_cz_fips Ex: 41, 127, 153.The FIPS number of the county assigned
- tor_other_cz_name Ex: TEXAS, HOUSTON. The FIPS name of the county entered by the continuing tornado segment as it crossed from one county to another.
- episode_title Ex: Severe weather outbreak on saturday April 14 in eastern Texas.
- event_narrative (The event narrative provides more specific details of the individual event . The event narrative is provided by NWS.) Rainfall of 2 to 3 inches fell across the area.

V. IMPLEMENTATION

OBJECTIVE

The main objective is to predict the following analysis that includes: Maximum Rainfall areas, Upcoming Rainfall, Maximum rainstorm locations based on hourly, local and overall storm parameters, Causes of Rainstorm: Light Damage, Heavy Damage, Significant Damage, Severe Damage etc.

MODULES

The following are modules that describes how to implement rainfall related datas:

ESTABLISHING DISTRIBUTED THE RESOURCE CLUSTER THROUGH HADOOP FILE SYSTEM

HDFS is installed and configured to store a very large amount of rainfall information(terabytes and petabytes) as clusters. When the data in a cluster environment breaks the HDFS into several pieces and allocates them in a different participating server machines in a hadoop cluster. Every servers, allocate a tiny segment in a entire dataset, and each data is pretend in more than one server machine. The configuration of HDFS has the following requirement.

- This requires spreading the data across a large number of machines.
- HDFS should store data reliably and retrieval of the data is flexible. If individual machines in the cluster malfunction, data should still be available.
- HDFS should provide fast, scalable access to this information. It should be possible to serve a larger number of clients by simply adding more machines to the cluster.

MODELING THE MAPREDUCE FRAMEWORK FOR TASK PROCESSING:

MapReduce is a programming model and an associated implementation for processing and generating rainfall data with provision for task classification of spatial temporal characteristics using Support Vector Machine(SVM). The computation takes a set of *input* key/value pairs, and produces a set of *output* key/value pairs.The MapReduce program consists of two functions, namely

Following architecture diagram predicts the MapReduce framework

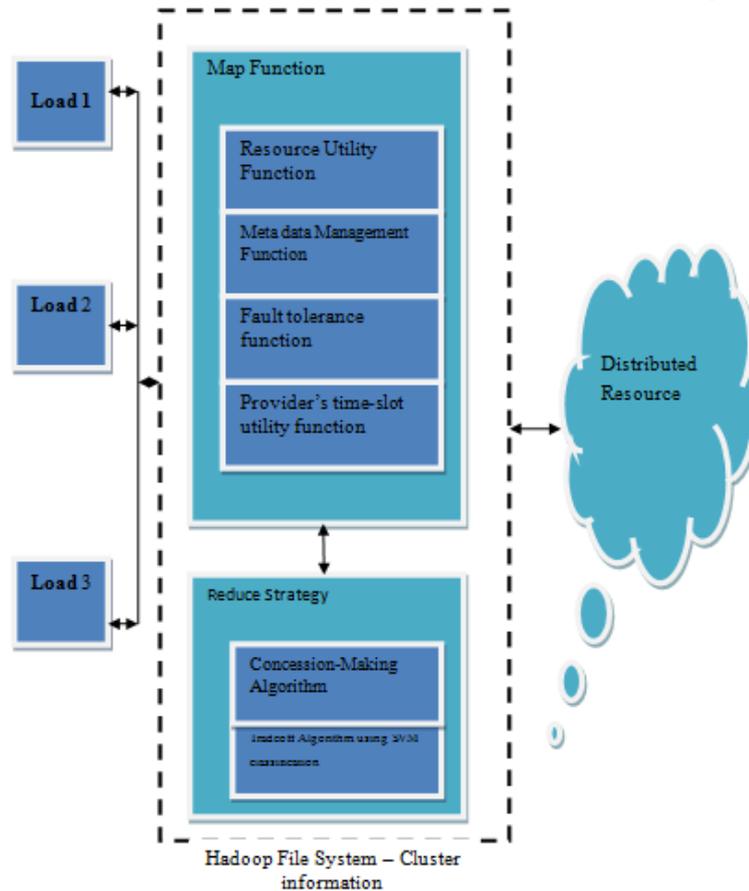


Figure 1. Architecture Diagram

A. CLASSIFICATION OF DATA TO MAPPER PHASE PROCESS BASED ON SPATIO TEMPORAL CHARACTERISTICS USING SUPPORT VECTOR MACHINE(SVM):

Map process takes carries of portioning the Spatial data of rainfall information. Spatial overlap is handled to reduce function The SVM is data mining methodology to mine the information of hydrological data based on the relational ship between the fault tolerance mechanism and a total number of support vectors involved. Various hydrological data are used to characterise the storm(flood) indicated attributes. Information mined in hydrological data were measured using various entropy. The SVM takes a input, as a set of data and predicts the possible outputs for a given input. It is the method of representing the points over a hyper plane. A separating hyper plane can be written as

$$V \cdot X + b = 0$$

Where W is a weight vector namely, $V = \{v_1, v_2, v_3, \dots, v_n\}$. Thus, any point that lies above the hyperplane satisfies the following equation

$$v_0 + v_1x_1 + v_2x_2 > 0$$

Similarly, any point that lies below the separating hyper plane satisfies

$$v_0 + v_1x_1 + v_2x_2 < 0$$

The weights can be adjusted so that the hyperplane defining the sides of the margin can be written as

$$H1: v_0 + v_1x_1 + v_2x_2 \geq 1 \text{ for } y_i = +1,$$

$$H2: v_0 + v_1x_1 + v_2x_2 \leq -1 \text{ for } y_i = -1$$

SVM belongs to a family of generalized liner and nonlinear classifiers and can be interpreted as an extension of the perception. A special property is that they simultaneously minimize the empirical classification error and maximize the geometric margin, hence they also known as maximum margin classifiers.

VI. EXPECTED RESULTS

This paper describes to predict the rainfall related areas and storm affected information based on the National Weather Service information(NWS) datasets.

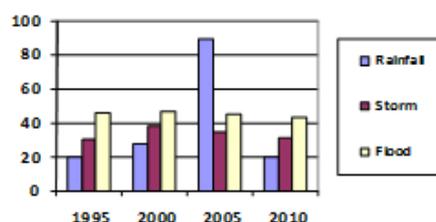


Figure 2 Showing the Rainfall and storm affected in several years.

VII. CONCLUSION

This paper shows upcoming rainfall through MapReduce and Support Vector Machine of classification that uses the Concession Making and Tradeoff algorithms. The complete storm System serves as tool or a framework that helps the big rainfall data, that is analysed by domain experts or users. The result indicates the proposed system improves the performance in terms of accuracy and efficiency.

ACKNOWLEDGMENT

I sincerely thank project guide and project coordinator for constant encouragement and support throughout and usefull suggestions given for completing work with complete guidance.

REFERENCES

- [1] K. Jitkajornwanich, R. Elmasri, C. Li, and J. McEnery, "Extracting Storm-Centric Characteristics from Raw Rainfall Data for Storm Analysis and Mining," Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (ACM SIGSPATIAL BIGSPATIAL'12), 2012, pp. 91-99.
- [2] K. Jitkajornwanich, U. Gupta, R. Elmasri, L. Fegaras, and J. McEnery, "Using MapReduce to Speed Up Storm Identification from Big Raw Rainfall Data," Proceedings of the 4th International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING'13), 2013, pp. 49-55.
- [3] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI'04), 2004.
- [4] C. Lam, Hadoop in Action. Dreamtech Press, New Delhi, 2011.
- [5] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [6] R. Elmasri and S. Navathe, Fundamentals of Database Systems, 6th ed. Pearson Education, Massachusetts, 2010.
- [7] A. Overeem, T. A. Buishand, and I. Holleman, "Rainfall Depth-Duration-Frequency Curves and Their Uncertainties," Journal of Hydrology, vol. 348, 2008, pp. 124-134.
- [8] W. H. Asquith, M. C. Roussel, T. G. Cleveland, X. Fang, and D. B. Thompson, "Statistical Characteristics of Storm Interevent Time, Depth, and Duration for Eastern New Mexico, Oklahoma, and Texas," Professional Paper 1725. U.S. Geological Survey, 2006.
- [9] W. H. Asquith, "Depth-Duration Frequency of Precipitation for Texas," Water-Resources Investigations Report 98-4044. U.S. Geological Survey (USGS), 1998.
- [10] Virginia Department of Conservation and Recreation, "Stormwater Management: Hydrologic Methods," retrieved: May 2, 2012, from: http://dcr.cache.vi.virginia.gov/stormwater_management/documents/Chapter_4.pdf.