Runoff prediction using Big Data analytics based on ARIMA Model

Vijay Dhote¹, Satanand Mishra²*, Jai Prakash Shukla³ & S. K. Pandey ⁴

¹, CSIR- Advanced Materials and Processes Research Institute, Bhopal 462026, India
², VNS Group of Institutions, Bhopal 462026, India

[E. Mail: snmishra07@gmail.com]

Received 20 February 2017; revised 02 June 2017

Big data Analytics is used in the study of developing forecasting models for prediction of runoff in Narmada river basin. Big data and data mining used as an advanced technique for storing and managing the large data set of runoff. Hadoop technique is used for storing and processing the large data. A new concept of big data processing is known as MapReduce it is a programming model. MapReduce is emerging as an important programming model for large-scale data-parallel applications such as web indexing, data mining, and scientific simulation. Autoregressive Integrated Moving Average (ARIMA) modelling is used for the prediction of time series runoff. Historical runoff data, which is large in size is stored in big database. The main objective of the time series modelling is to carefully collect and rigorously study the past observation of time series and to develop an appropriate model that predict the future runoff in hydrological time series.

[Key Words: Big Data, Hadoop, MapReduce, HDFS, time series analysis, ARIMA, Data mining].

Introduction

Hydrology is the scientific study and gathering of information regarding the presence and availability of water in all its forms on earth. It includes water sharing in both space and time on the earth, and the processing, storage, and retrieval of hydrological data. The goal and reason of the analyses of hydrological information is to be able to benefit people at a future time with respect to water supply, flood protection, waterpower, and coastal protection projects. The collection of hydrological data must be based on sound theoretical principles, well-coordinated, and subjected to strict quality control and error estimation and correction techniques. The detailed ways of obtaining and recording hydrological data by means of hydraulic measuring flumes and calibrated weirs are described elsewhere in the present theme and form the basis of good practice. These phenomena's are responsible for gathering the information and data in a huge quantity and born the hydrological big data.

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The volume of data with the speed, it is generated and makes it difficult for the current computing infrastructure to handle big data. To overcome this drawback, big data processing can be performed through a programming paradigm known as MapReduce. MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks¹. Hadoop is an open-source software framework for storing and processing big data in a distributed fashion on large clusters of commodity hardware. Essentially, Hadoop have two important tasks one is massive data storage and another is faster processing of data. Data mining is the process of automatically discovering useful information from large data repositories. Data mining is one of the burning research areas in computer science². Prediction is one of the basic goals of data mining. There are many types of technical methods of data mining, which mainly include: association rule mining algorithm, decision tree, classification algorithm, clustering algorithm and time series mining algorithm, etc³. The characteristics of Big Data have presented "5V" - Volume, Variety, Velocity, Veracity, and Value. The formation of scientific term big data should be attributed by Mashey, Weiss and Indurkhya, Diebold and Laney et al. In March 2012, the government of the United States start "Big Data Research and Development Plan", which marks the United States rise Big Data up to country level 5. The scale of data that one country has and the capacity to that one country maintenance data will become the important component part of comprehensive national strength and will also become the new focus among
countries and enterprises to occupy and control data. Time series data mining is to extract information and knowledge from a lot of time series data, these information and knowledge are not known in advance for people but they are potentially useful and time-related, and for short-term, medium-term or long-term forecasts, guiding people's behaviour such as society, economy, military and life. In fact, almost all data in the hydrological and meteorological field are time series data, and future in hydrological and meteorological data can be predicted better by means of the time series mining algorithm. The time series mining algorithm adopted in the article is ARIMA time series mining algorithm. Many researchers have started study in the field of data mining using big data and hadoop framework in various field, some of the study related to our work given as follows:

Suryawanshi and Wadne (2014), studied Big Data Mining using Map Reduce and introduced the challenges of Big Data Mining describes methods like cube materialization, MapReduce and MR-cube approach for the extraction of knowledge from data ware houses. Ayma V.A., et.al. (2015) given classification algorithms for Big Data Analysis and MapReduce approach. In this study, concerned about how to increase the accuracy of different classification methods, and major achievements have been made so far. In this work, a tool within the scope of Inter IMAGE Cloud Platform (ICP), which is an open-source, distributed framework for automatic image interpretation, is presented. The tool, named ICP: Data Mining Package, is able to perform supervised classification procedures on huge amounts of data, usually referred as big data, on a distributed infrastructure using Hadoop MapReduce. Leixiao et al. (2013) proposed an ARIMA algorithm based on Hadoop framework, and implement an effective weather data analysing and forecasting system. The researches on hydrological cycle effects caused by human activities are mainly focused on two aspects: one is to analyse time series of influencing factors by many statistics approaches, which mainly investigate the degree of human activities effecting on the factors; the other one is to build hydrological simulation model considering the effects of human activities. Large amounts of data from many aspects including hydrology, weather, geography and environment have been highly demanded, and the current researches are only focusing on large-scale river basins. But it does not provide sufficient information for discussing the effects of human activities on hydrological projects and real-time exertion of the engineering. Prashant Shrivastava, S. Pandiaraj and Dr. J. Jagadeesan (2014) Presented the Big Data Analytics in Forecasting Lakes Levels. Big data analytics is not only about managing large or diverse data but it is about self-questioning on the available data, deriving new hypotheses and discovering to make data-driven decisions. It helps to uncover useful information, hidden patterns, and several unknown correlations. Data Analytics can be used in the study of developing forecasting models for predicting water levels of lakes. ARIMA modelling is used for the study as it works well on time series value.

Materials and Methods

For the study, Hoshangabad site is selected on Narmada river basin. The Narmada is the largest West flowing and fifth largest river of India. It drains a large area in Madhya Pradesh besides some area in the states of Maharashtra and Gujrat. The Narmada basin line between East Longitudes 72°32’ to 81°45’ and North Latitudes 21°20’ to 23°45’. It flows through Deccan trap in between Vindhya and Satpura ranges of hills before falling into Gulf of Cambay in the Arabian Sea. The Narmada drains an area of 98796 sq. km. Out of which nearly 87% lies in Madhya Pradesh. In the Figure 1, Hoshangabad site is given. Water level, discharge and sediment data are collected from central water commission. The data is daily basis and average monthly data has been computed.

In this study, big data is used to take action – to make more accurate decisions and to do more quickly, which is called situational awareness and reduced data set in small data. It is used for the time reductions and smart decision making for the prediction of runoff in our case. For the data analysis Hadoop and MapReduce is used and for the prediction of runoff data mining techniques is used. The ARIMA Model is used for the prediction of runoff under the implicit laws of Hadoop, MapReduce and data mining. The application of this study Hadoop, MapReduce and data mining are given as under:

Transform Processing is enabling the average of the data sequence to be zero and stationarity is the premise condition of ARIMA model and which is the responsible for the model development. Figure 3 shows the algorithm flow chart of ARIMA prediction model. This diagram depicts the processes of model development.

Hadoop is an open-source software framework for storage and large-scale processing of data-sets on
clusters of commodity hardware. Hadoop framework is written in Java. At a very high level, Hadoop has two main components MapReduce and file system\(^\text{10}\). MapReduce is the processing part of Hadoop and manage the jobs. HDFS refers to Hadoop Distributed File System, which is stores all the data redundantly and it is required for computation. Processing pillar in the Hadoop ecosystem is the MapReduce framework\(^\text{11}\). Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. It is also provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high cumulative bandwidth across the cluster. Both MapReduce and the distributed file system are designed so that node failures are automatically handled by the framework.

Figure 4 shows the Hadoop architecture which consists of two main components distributed file system (HDFS) and Execution engine (MapReduce). The HDFS is inspired by the Google File System (GFS) and is designed to store very large data across machines in a large cluster. Another one MapReduce is a simple programming model by Google which divides application into many small blocks\(^\text{12}\). The MapReduce has two phases, first is a map phase which operates on input as key value pair. Other one is reduce phase, which produces output in form of key value pair. The MapReduce engine has the job
Fig. 3 — Algorithm Flow chart of ARIMA Prediction Model

Hadoop is used as a fault-tolerant storage system that is called a Hadoop Distributed File System (HDFS). HDFS is a client-server based architecture, which comprise of Name Node and many Data Nodes. HDFS is like the bucket of the Hadoop system. Hadoop Distributed File System and MapReduce are the core components of Hadoop. Hadoop Distributed file System is used for processing, storing and analyzing very large amount of unstructured data. The Hadoop Distributed File System (HDFS) is the storage component of Hadoop. It is designed to reliably store very large data sets on clusters, and to stream those data at high throughput to user applications. HDFS takes care of data storage and MapReduce handles processing. Hadoop's MapReduce and HDFS components are originally derived respectively from Google’s MapReduce and Google File System (GFS) papers. Main components of HDFS are name node, data node and secondary name node. Main components of MapReduce are job tracker and task tracker. MapReduce utilizes the Google File System (GFS) as an underlying storage layer to read input and store output. GFS is a chunk-based distributed file system that supports fault-tolerance by data partitioning and replication.

MapReduce basically used for process data on the cluster. MapReduce is currently the most well-known framework for data intensive computing. MapReduce is motivated by the demands of processing huge amounts of data from a web environment. MapReduce provides an easy parallel programming interface in a distributed computing environment. The most powerful feature of MapReduce is its high scalability that allows user to process a vast amount of data in a short time. Typical, implementation of the MapReduce paradigm requires networked attached storage and parallel processing. Hadoop and HDFS by apache are widely used for storing and managing big data. There are many fields that benefit from MapReduce, such as Bioinformatics, machine learning, scientific analysis, web data analysis, astrophysics, and security. Two execution steps are used map execution and reduce execution.

The ability to build a successful predictive model depends on past data. Data mining is subjected to learn from past success and failures and will be able to predict what will happen next (future prediction). Data mining is the process of extracting useful information or to find out hidden relationship among data. This information or knowledge is very helpful
for business organisations to grow their business as it is helpful in decision making. Data mining technology has come across several stages. In the first stage, it was a single algorithm for single machine for vector data. In the second stage, it was combined with database for multiple algorithms. In the third stage, it is where, it has provided support for grid computing. Fourth stage data mining algorithm was distributed. In the Fifth stage, parallel data mining algorithms are presented for big data and cloud services.

The ARIMA is also called Box-Jenkins technique as it is divide into three parts AR-Autoregressive process, I- Integrated and MA- Moving Average for the forecast errors. ARIMA models have 3 model parameters, one for the AR(p) process, I(d) process and one for MA(q) process, once all the parameters are combined and interacting among one another forms ARIMA(p,d,q). The analysis done through ARIMA is divide into three stage as described in Box and Jenkins. The three main steps are: Establish, Estimate, and Forecast statement. In the first step it uses the identify statement to specify the response series followed by distinctive candidate ARIMA models for it. The analysis done on the Identity statement output typically suggests one or additional ARIMA models that would be used. In the second step, It is used for checking estimation and diagnostic, it needs Estimate statement to specify the ARIMA model to fit to the variable laid out in initial stage, Identity statement to be used. It additionally estimates the parameters of that model. In the third step of statement, it uses the Forecast statement to forecast future values of the statistic and to get confidence intervals for them with facilitate from the ARIMA model created by the previous Estimate statement. ARIMA model, provides a good technique for forecasting the magnitude of any time series variable. ARIMA models provides useful tools that can be used to compare the performance of other forecasting models such as neural network, kernel regression etc.

Results
The hydrological processed data received from MapReduce is used for the development of ARIMA model. The ARIMA (1, 1, 1) (1, 1, 1) has found best fitted model. For the calibration of model, 10 years monthly data has been used and 2 years data for validation. Since data is used monthly the seasonality has been occurred. The fitted model is known as seasonality ARIMA model. In the figure, Blue line shows the observed runoff and red dotted line shows, predicted discharge. The violet line, shows validated...
runoff and green line shows, future predicted runoff. For the ARIMA model development, XLSTAT software is used.

Correlation coefficients measure the strength of association between two variables. The most common correlation coefficient, called the Pearson product-moment correlation coefficient, measures the strength of the linear association between variables. The coefficients of determination denoted by $R^2$ is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. The coefficient of determination ranges from 0 to 1. An $R^2$ between 0 and 1 indicates the extent to which the dependent variable is predictable. An $R^2$ of 0.9493 means that 95 percent of the variance in $Y$ is...
predictable from $X$ in training and 0.99 in validation. Figure 6(I) shows the Correlation between observed and computed (training) discharge in cumecs and Figure 6(II) shows the Correlation between observed and computed (validation) discharge in cumecs. The RMSE are found that 7.34 in training and 2.74 in validations. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. It means that how concentrated the data is around the line of best fit.

The difference between the observed runoff and predicted runoff is called the residual ($e$). Each data point has one residual.

$$\text{Residual} = \text{Observed runoff value} - \text{Predicted runoff value} = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero. That is, $\Sigma e = 0$ and $e = 0$. The Figure 7 shows residuals. The residual plot shows a fairly random pattern - the some residuals are positive and some residuals are negative. This random pattern...
indicates that ARIMA model provides a decent fit to the data. A random pattern of residuals supports a linear model. A non-random pattern supports a non-linear model. The sum of the residuals is always zero, whether the data set is linear or nonlinear.

Conclusions
In this paper, Hadoop data cluster, HDFS and MapReduce programming framework has been explored to provide solution to big data problem. We have examined the design and architecture of Hadoop’s MapReduce framework in great detail. Particularly, our analysis has focused on data processing. We would conclude by saying that big data is the new buzz word and Hadoop MapReduce is the best tool available for processing data and its distributed, column-oriented database, HBase which uses HDFS for its underlying storage, and support provides more efficiency to the system. In this paper, using ARIMA model for runoff prediction. The development and use of stochastic models of hydrological phenomena play an important role in many aspects of water resources engineering.

Acknowledgement
Authors would like to thank the Central Water Commission, Ministry of Water Resources, India for providing Water Level and Discharge data.

References