IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. VII (Mar-Apr. 2014), PP 17-20
www.iosrjournals.org

Cloud Based Software Platform for Big Data Analytics In Water Reservoir Level Forecasting

Prashant Shrivastava[1], S. Pandiaraj[2], Dr. J. Jagadeesan[3]
[1](Student M. Tech. (CSE), Department of Computer Science and Engineering, SRM University, Chennai, India)
[2](Assistant Professor, Department of Computer Science and Engineering, SRM University, Chennai, India)
[3](Head of Department, Department of Computer Science and Engineering, SRM University, Chennai, India)

Abstract: *Advanced technology solutions will help in scaling implementation, to accommodate a lot of data and data analysis capabilities while not affecting the performance. Data analytics in the field of water resource management is been seen as the new area of study that may facilitate in optimally managing the provision of water based on availability. Cloud platform will reduce the cost of maintenance singly in isolated environments and with the application of big data will help in activity the data analytics fast. Exploitation technologies to display user friendly analytics results graphically and serving to in foretelling add feather in the cap for this architecture. Reservoirs forms the rear bone of the facility inside cities will significantly help the facility department's team in formulating advanced attending to manage the facility optimally if forecasted properly. The projected solution in this article helps in building a scalable software system platform for water reservoirs levels data analysis for foretelling future levels exploitation cloud based platform through massive data technologies landscape. The aim of this study is to develop models for predicting water levels in any reservoir. It applies Autoregressive Integrated Moving Average (ARIMA) algorithmic rule for creating predictions. It stores the historic huge data set inside massive information storage and uses big data technologies to review behavior and predict the future levels by applying data-driven analytics and data mining concepts.*

Keywords: *Analytics, ARIMA, Big Data, Cloud Computing, Data Mining, Hadoop, Time Series Modeling, Water Reservoirs*

I. INTRODUCTION

In this article, we tend to describe our experiences in building a cloud-based software platform for data-driven analytics that takes a step toward the forecasting large amount of information on water levels of reservoirs. Our effort is to sample knowledge of the reservoirs of Chennai metro water system water for past 5 years and predict the long run levels. We've developed data-driven prognostic models for forecasting future levels which will help decide and set up the availability of water across city. The projected solution is generic and might be simply applied to any number of reservoirs.

Big data technologies are recent technologies that help in developing generic architectures so that various water management organizations will utilize them in cost effective manner and extract needed forecasting from very large volumes of distributed range of information by facultative quick capture, quick discovery, and analysis.

The studies captures water levels for the reservoirs and store them initial in temporary storage before transferring it into the big database primarily based on big data so that the analytics may be performed on the large knowledge set for forecasting the levels in future. The usage of cloud primarily based technologies makes the system simple to scale and out there for larger applications and usage by avoiding the additional efforts and value in maintaining the infrastructure and services. The usage of huge knowledge makes the system highly sturdy and climbable which can embody many reservoirs at city level or state level or maybe up to national levels. There are many technologies which will be integrated with the big knowledge for performing arts the analytics and displaying the ends up in simply graspable graphical form.

II. LITERATURE REVIEW

To reach to the planned resolution literature is surveyed within the domain area of setting and water resource management. The concepts utilized in energy and utilities domain that features - sensible grid operations in energy distribution, operations in water resource management/ distribution, water level prediction ways in lakes, reservoirs, statistic foretelling algorithms. Literature is also surveyed within the area of various current advanced technologies out there utilized in the world which will be utilized in the present study and provide climbable and straightforward to deploy resolution. The technologies known for the present study includes - Cloud based software system platforms, study and usage of big information in playacting information analytics, study of various statistic primarily based algorithms implementation for faster and accurate results foretelling.

III. AUTO-REGRESSIVE INTEGRATED MOVING-AVERAGE (ARIMA)

The ARIMA (also called Box-Jenkins technique) modeling analysis is employed within the study which forecasts equally spaced time series data, with transfer operate and intervention data. There are 3 parts among ARIMA – (1) AR – Autoregressive process (2) I – Integrated (3) MA – Moving Average for the forecast errors. ARIMA models have 3 model parameters, one for the AR(p) process, one for I(d) process, and one for the MA(q) process, once all the parameters are combined and interacting among one another forms ARIMA (p, d, q). The model predicts value during a latency series that is linear combination of its own past values and current and past values of alternative time series.

The analysis done through ARIMA is broadly categorized into 3 stages as described in Box and Jenkins. The 3 main steps - establish, Estimate, and Forecast statements used in ARIMA are summarized below.

1. In the first stage it uses the identify statement to specify the response series followed by distinctive candidate ARIMA models for it. The Identity statement reads time series that are going to be used in later stages, and if required differencing them, and hard numerous autocorrelations, numerous inverse autocorrelations, and other partial autocorrelations, and if any cross correlations would like for differencing are often identified by applying stationary tests. The analysis done on the Identity statement output typically suggests one or additional ARIMA models that would be used.
2. In the second stage that 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. The Estimate statement provides produces diagnostic statistics to assist in judging the correctness of the model. Significance tests are conducted for parameter estimates that indicate whether some terms in the model are necessary or not. Goodness-of-fit statistics helps in comparison this model with others.
3. In the third stage of statement it use 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 an honest technique for forecasting the magnitude of any variable. Its strength is within the indisputable fact that this method is suitable for any statistic with any pattern of modification. This can be the most necessary advantage of exploitation ARIMA model.

IV. METHODOLOGY

First, real time information has got to be inheritable from reservoirs through the file based mostly date transfer system like SFTP or manual transfer through emails. It includes daily information also as historical information aggregative over amount of time that must be loaded into the database. Behavioral research and end-use analysis require not just information but also the context for energy use. These slow dynamical datasets ought to be updated periodically, and also the sources themselves may modification over time. As a result, an automatic information ingest system has got to support discrete information acquisition at completely different rates and size, and be receptive recent information sources and operational necessities.

Data inherited is kept into the temporary storage and afterwards has got to be keep and shared with massive system. To researchers mining and exploring information for correlations and gaining information is most vital. The info collaboration has got to be balanced against the considerations of knowledge security. We tend to use private Cloud storage platforms that provide a manageable and reliable information hosting solution for distributed access and co-location with reckon resources for analytics.

Data-driven forecasting models are essential and for it data-driven models are trained exploitation historical data, and utilize massive dimensions of options that are direct and indirect indicators of water levels. in particular, our demand foretelling models use ARIMA time-series to supply high accuracy, uses the Hadoop MapReduce platform for playacting such analytics, and is tuned to scale on public cloud infrastructure.

Data-driven models have many benefits, the largest one out of that is that the convenience of automatically building a model without having deep technical information on the system. The models also can be easily well-kept up to now by grooming them over new data that is collected. Further, it permits data analysts to try completely different combos of options to find those that most significantly impact the water levels. This will facilitate outline limits for data collection, or alternatively can give insights on reducing water usage. We tend to observe that there's no "one size fits all" global model, and instead a set of models are used for various functions, however the results of exploitation ARIMA model provides most as regards to actual results.

The water level information is collected from the reservoirs that contain information of the water level in the reservoir on a specific day. Sample information is given in the table below.

Table 1: Sample Data Showing Reservoir Levels

Date	Reservoir Levels (ft.)
1-Feb-2014	48.96
2-Feb-2014	49.14
3-Feb-2014	49.31
4-Feb-2014	49.52
5-Feb-2014	49.73
6-Feb-2014	50.04
7-Feb-2014	50.30
8-Feb-2014	50.58

The level of reservoir is plotted for the info collected for the past five year and to see the patterns and behavior. Mostly it is seen that the levels volume unit additional are less consistent across last five years as shown below.

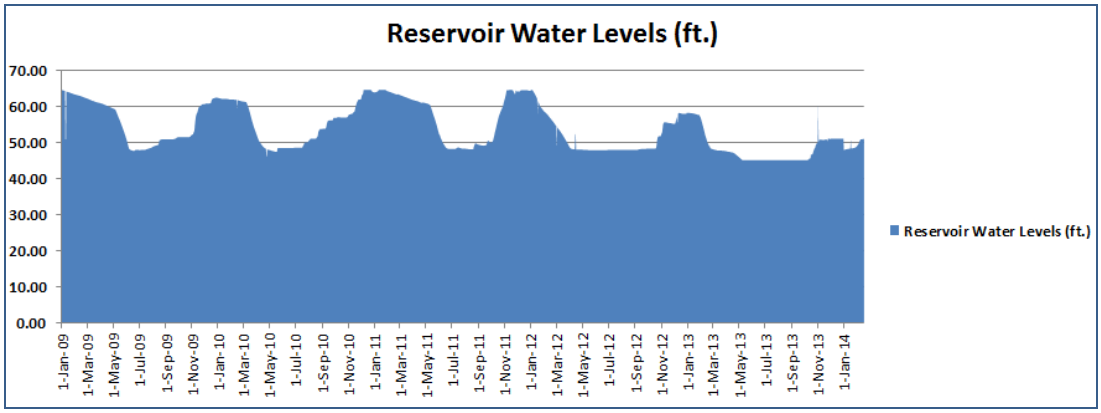

Figure 1: Sample Reservoir Levels for last 5 Years

The ARIMA model predicts the water levels that measure very close or near to the actual values levels as shown in below graph. Less variation measure seen with the forecasted and that of the actual provides the closeness of predication to the actual values. A sample forecast is show below which is close to the actual in the range to plus-minus five percent.

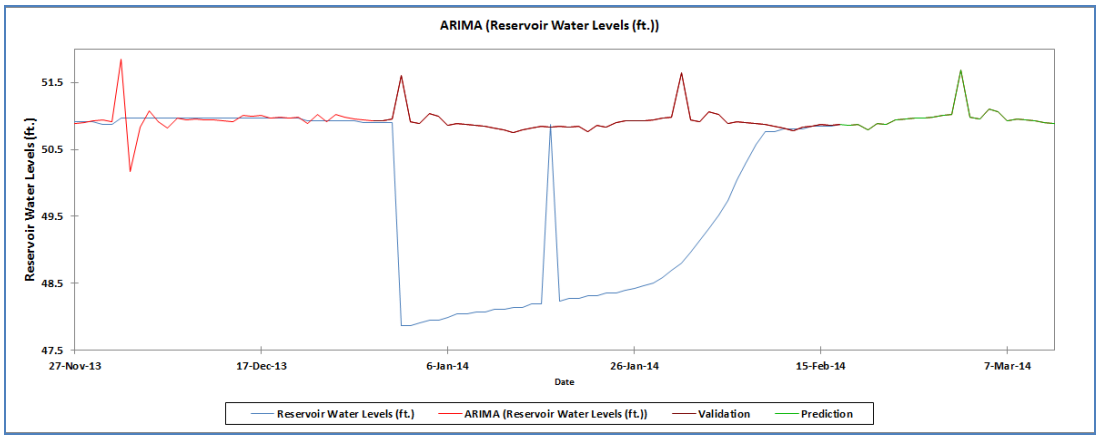

Figure 2: Sample Prediction of Water Levels through ARIMA

The software system platform that is used for implementation consists of using Ubuntu 12.04 as the base operating system on top of Oracle Virtual machine. The personal cloud is made using Ubuntu Open stack. A single node cluster is created on Ubuntu using Hadoop. A software system used for initial storage is MySQL for storing values received from the reservoir that is afterward transferred into the HBase database created on the single node cluster. The query and analysis for the info for making prediction is completed by use of Hive and R-language for forecasting the water levels for the reservoirs. The obtained predictions are often displayed in type of graphs on any user interface using Java tool for R language.

V. CONCLUSION

This effort to investigate and propose a computer platform which can be utilized in long term for reservoirs level statement comes very useful. It's established that the system is easily scalable to include additional knowledge for different reservoirs and always perform efficiently. Cloud computing has been well-tried essential in design for this system, though our expertise shows that picking the right Cloud abstraction for individual parts in the design is very vital to ensure the associated advantages and overheads. Lot of resources in distributed knowledge centers offered by commercial Cloud providers reduces the complexity of knowledge sharing and analytics on-demand however will have value associated thereto. Private Clouds definitely offers additional physical security over personal knowledge; however manageability of the hardware still remains a priority in it. Infrastructure as Service (IaaS) Clouds offer sensible controls over versatile resources, however brings with them technical challenge in planning frameworks which will efficiently utilize their capabilities. Platform as Service (PaaS) Clouds for instance Hadoop definitely scale back the time to create scalable applications however generating sensible performance still requires lot of efforts in calibration.

More study can be performed to make the system simpler by using alternative comparable technologies that are used in the present article. Results from completely different technologies can be compared and can be steered. Similarly more time-series based algorithms or models can be applied and compared and most accurate and technically viable algorithm can be suggested.

The study wiped out the sphere of water reservoirs levels management will be more applied into alternative fields not simply limiting to any particular area. The system can be very easily applied to any huge data based analytics measurement for prognostication in areas such as weather forecasting, financial markets etc.

REFERENCES

Journal Papers:

[1] Hydrological analysis for water level projections in Taihu Lake, China (Journal of Flood Risk Management)by L. Liu, Z.X. Xu, N.S. Reynard, C.W. Hu and R.G. Jones (March 2013)

[2] Stochastic modeling of Lake Van water level time series with jumps and multiple trends (Hydrology and Earth System Sciences (An Interactive Open Access Journal of the European Geosciences Union)) by H. Aksoy, N. E. Unal, E. Eris, and M. I. Yuce (February 2013)

[3] Hadoop-based ARIMA Algorithm and its Application in Weather Forecast (International Journal of Database Theory and Application) by Leixiao Li, Zhiqiang Ma, Limin Liu and Yuhong Fan(2013)

[4] Cloud-Based Software Platform for Big Data Analytics in Smart Grids (IEEE) by Yogesh Simmhan, Saima Aman, Alok Kumbhare, Rongyang Liu, Sam Stevens, Qunzhi Zhou, Viktor Prasanna

[5] Predicting Water Levels at Kainji Dam Using Artificial Neural Networks (Nigerian Journal of Technology (NIJOTECH)) by C.C. Nwobi-Okoye, A.C. Igboanugo (March 2013)

[6] New York City (NYC), government web site for the white paper on New York City's Operations Support Tool (OST) White Paper http://www.nyc.gov/html/dep/pdf/reports/ost_white_paper.pdf

[7] Chennai Metropolitan Water Supply & Sewage Board (CMWSSB), government web site for getting data http://chennaimetrowater.gov.in/public/lake.htm

[8] IEEE paper on - Cloud-Based Software Platform for Big Data Analytics in Smart Grids www.computer.org/csdl/mags/cs/2013/04/mcs2013040038.pdf

[9] Running Hadoop on Ubuntu Linux (Single-Node Cluster) http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

[10] Running Hadoop on Ubuntu Linux (Multi-Node Cluster) http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/

[11] International Journal from Science & Engineering Research Support Society (SERSC): Hadoop-based ARIMA Algorithm and its Application in Weather Forecast http://www.sersc.org/journals/IJDTA/vol6_no5/11.pdf

Books:

[12] *Book Title – Hadoop The Definitive Guide*; By Tom White; Oreilly Publication (Fourth Indian Reprint: Jun 2013)

Chapters in Books:

[1] Chapter 9 – Setting Up a Hadoop Cluster, *Hadoop The Definitive Guide*; *By Tom White; Oreilly Publication (Fourth Indian Reprint: Jun 2013)*

[2] Chapter 12 – Hive, *Hadoop The Definitive Guide*; *By Tom White; Oreilly Publication (Fourth Indian Reprint: Jun 2013)*

[3] Chapter 13 – Hbase, *Hadoop The Definitive Guide*; *By Tom White; Oreilly Publication (Fourth Indian Reprint: Jun 2013)*