SPATIO-TEMPORAL STREAM ANALYSIS BASED ON HADOOP AND R COOPERATING VIA RELATIONAL DATABASE

[1]JUNGHOON LEE, [2]JIWON JUNG, [3]DONGHYUN KIM, [4]SUNA KIM, [5]GYUNG-LEEN PARK

[1,2,3,4]Dept. of Computer Science and Statistics, Jeju National University, Republic of Korea
E-mail: [1]jhlee@jejunu.ac.kr, [2]jjy1809@naver.com, 3dh0806@naver.com, 4musiqline@naver.com [5]glpark@jejunu.ac.kr,

Abstract – This paper presents how Hadoop, MySQL, and R cooperate on different machines for the analysis of diverse data streams generated in smart grids, particularly integrating geographic information. Our analysis scenarios benefit from big data processing, information sharing, and abundant statistical analysis, respectively. Specifically, 1) R solely plots the geographic information obtained from an ESRI shape file not so intensively shared. 2) R combines the location information with the district-by-district statistics downloaded from MySQL. 3) Hadoop first summarizes the daily report from a potentially massive amount of data and stores in MySQL. Then, R downloads and conducts the trend pattern analysis using its neural network library. In such procedures, R installs the libraries of RMySQL, GISTools, and neuralnet. This framework can host additional components for more sophisticated analysis for ever-growing smart grid entities such as electric vehicles.

Keywords— Smart Grid, Massive Data Stream, Hadoop, R Package, MySQL Database.

I. INTRODUCTION*

In modern power networks, called smart grids, are ceaselessly generating a massive amount of real-time data streams, originally implemented for the purpose of keeping the safe operation of the energy system [1]. A single stream consists of a series of status records, each of which contains a timestamp, or sometimes a location stamp for mobile object such as EVs (Electric Vehicles). Those objects range from power generation to power consumption, and they are geographically distributed over a wide area. Hence, the stream processing must integrate the geographic information and geo-computing analysis mechanisms [2]. Moreover, the massiveness of streams requires parallel and distributed computing to proactively detect abnormal operation signals, find some trends, and the like. Nowadays, many software tools are available even for free, making it possible to build a systematic analysis framework for massive and heterogeneous data sets. In this regard, this paper builds a stream analysis framework and demonstrates preliminary results from the stored and managed data stream.

II. FRAMEWORK ARCHITECTURE

Figure 1 illustrates the architecture of our system, which mainly includes Apache Hadoop, R, and MySQL. First, the stream will be stored either on MySQL or HDFS (Hadoop Distributed File System) in a Linux machine. In case a stream too large to be stored in a database, it is reasonable to store in text files. Actually, there is little advantage of relational real-time stream, as Hadoop can efficiently load complex data types from text files [3]. On the contrary, for some static information such as locations of facilities, MySQL provides well-defined and standardized query interfaces. Hadoop performs on the big data and generates the result in a text file usually in its data-bag style. Here, other programs stuffed with JDBC or ODBC (Java or Open Java DataBase Connectivity) drivers can also co-work with abovementioned modules, retrieving and uploading relevant data.

Next, we install the R statistical package on Windows machines. R provides the RMySQL extension package to interact with MySQL even running on remote machines. That is, an R script can retrieve the information of its interest from a database table as well as upload the analysis result to share with other analysis tasks. This architecture allows multiple analyzers to customize data, goal, and so on. In addition, R can read ESRI shape files which are widely used for geographic object representation [4]. As a result, analysis results and geographic objects meet in R, taking advantage of abundant graphic capabilities. More impressively, more R extension packages are being added, enriching our data processing work.

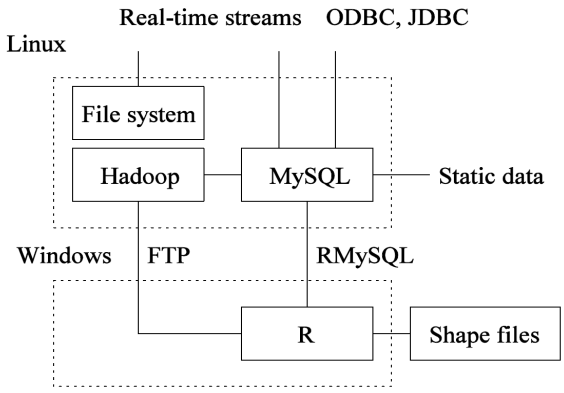


Fig. 1. Analysis framework architecture

* Following are results of a study on the ”Leaders INdustryuniversity Cooperation” Project, supported by the Ministry of Education, (MOE). database for a

* Following are results of a study on the “Leaders Industry university Cooperation” Project, supported by the Ministry of Education, (MOE).

III. ANALYSIS WORK

This section exhibits some analysis work from our framework. First, Figure 2 illustrates the contour map of Jeju Island, which is surrounded by a coastline stretching about 200 km. The contour map is stored in a shape file, while this display is obtained solely by an R script importing the GISTools library. This area has a big mountain as tall as 1,950 m in its center, hence this area is quite steep as can be inferred from the contour map. The shading color is automatically selected by the library, the script just specifying the green tones.

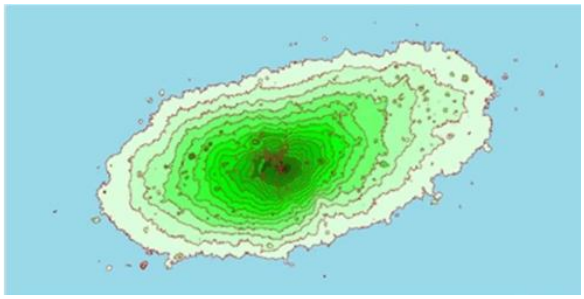


Fig. 2. Contour map plotting in R

Figure 3 shows the administrative districts of Jeju Island and the number of tour spots registered in the Jeju local government. Here, the district map is read from a shape file, while the per-district statistics data is downloaded from the MySQL database. To this end, R installs the RMySQL library. With this library, it doesn't matter whether an R script runs on the same machine as the MySQL server. In our system, the R script is in operation in a Windows machine, while the MySQL server runs on a Linux machine. A table in MySQL is downloaded as an array in R. Many other applications can upload the data to be shared to MySQL, possibly via xDBC according to the program language by which the application is implemented.



Fig. 3. District-by-district statistics

Finally, Figure 4 shows the trend analysis result of the number of records in a charger status monitoring system, by means of Hadoop, MySQL,and R, altogether. The monitoring system keeps collecting a series of periodical status reports from chargers. The records of the stream are sequentially stored in a file system according to the arriving order. To cope with the large volume of the auto-generated reports, Hadoop calculates a daily number of records and stores the result in the predefined MySQL table. This operation can be executed on daily basis and the result can be stored incrementally, which is very simple in database tables. Now, an R script downloads the daily statistics just as in the previous example, and then conducts a trend analysis using artificial neural networks [5]. It is supported by the neuralnet library, which is widely used for nonlinear series modeling.

In Figure 4, we can see two curves. The first one looks like a porcupine style having many irregular peaks, and it is the real change in the number of records. The second curve smoothly proceeds roughly in the middle of the first one. This curve traces the trend pattern estimated by the neural network. The neuralnet APIs allow us to build a neural network and make it learn from (day, number of records) tuples. It is also possible to trace the trend value by feeding the day field to the model and getting the computed value. This figure is directly exported from R studio as a form of the postscript file.

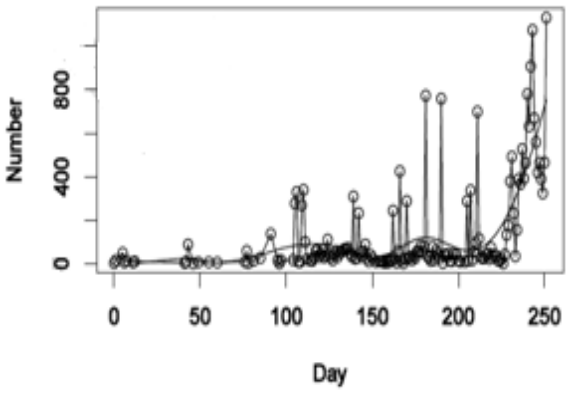


Fig. 4. Neural network-based trend analysis

CONCLUDING REMARKS

Smart grids are generating a massive amount of data streams for digital and intelligent management. In this paper, we have presented an analysis framework in which R and Hadoop cooperate mainly with the intervention of MySQL. MySQL stores static information, and partial results created from Hadoop, while R scripts download those tables they want on necessary basis. Three analysis models are demonstrated, including solely R, R and MySQL, and all of them, adding more intelligence. Here, for the analysis of trend pattern in a time series, a neural network model is built by making it learn from the Hadoop-generated summary. As future work, we are planning to acquire more stream data from the smart grid in Jeju city and try to search for meaningful analysis targets which will lead to a promising business model.

REFERENCES

[1] N. Balac, "Green machine, Intelligence: Greening and sustaining smart grids", IEEE Intelligent Systems, pp. 50-55, 2013.

[2] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. Saltz, "Hadoop-GIS: A High performance spatial data warehousing system over MapReduce", International Conference on Very Large Data Bases, pp. 1009-1020, 2012.

[3] J. Lee, and et al., "Data processing framework for electric vehicle generated streams using Hadoop," Lecture Notes in Electric Engineering, Vol. 330, pp. 777-781, 2014.

[4] C. Brunsdon and L. Comber, An Introduction to R for Spatial Analysis & Mapping, SAGE Publication Ltd., 2015.

[5] J. Lee, M. Kang, and G. Park, "Battery consumption modeling for electric vehicles based on artificial neural networks", Lecture Notes in Computer Science, Vol. 8582, pp. 733-742, 2014.

★★★