Integration Process and Analytic Interface of different Environmental Open Data Sets with Java/Oracle and R

Authors : Pavel H. Llamocca, Victoria Lopez

Abstract : The main objective of our work is the comparative analysis of environmental data from Open Data bases, belonging to different governments. This means that you have to integrate data from various different sources. Nowadays, many governments have the intention of publishing thousands of data sets for people and organizations to use them. In this way, the quantity of applications based on Open Data is increasing. However each government has its own procedures to publish its data, and it causes a variety of formats of data sets because there are no international standards to specify the formats of the data sets from Open Data bases. Due to this variety of formats, we must build a data integration process that is able to put together all kind of formats. There are some software tools developed in order to give support to the integration process, e.g. Data Tamer, Data Wrangler. The problem with these tools is that they need data scientist interaction to take part in the integration process as a final step. In our case we don't want to depend on a data scientist, because environmental data are usually similar and these processes can be automated by programming. The main idea of our tool is to build Hadoop procedures adapted to data sources per each government in order to achieve an automated integration. Our work focus in environment data like temperature, energy consumption, air quality, solar radiation, speeds of wind, etc. Since 2 years, the government of Madrid is publishing its Open Data bases relative to environment indicators in real time. In the same way, other governments have published Open Data sets relative to the environment (like Andalucia or Bilbao). But all of those data sets have different formats and our solution is able to integrate all of them, furthermore it allows the user to make and visualize some analysis over the real-time data. Once the integration task is done, all the data from any government has the same format and the analysis process can be initiated in a computational better way. So the tool presented in this work has two goals: 1. Integration process; and 2. Graphic and analytic interface. As a first approach, the integration process was developed using Java and Oracle and the graphic and analytic interface with Java (jsp). However, in order to open our software tool, as second approach, we also developed an implementation with R language as mature open source technology. R is a really powerful open source programming language that allows us to process and analyze a huge amount of data with high performance. There are also some R libraries for the building of a graphic interface like shiny. A performance comparison between both implementations was made and no significant differences were found. In addition, our work provides with an Official Real-Time Integrated Data Set about Environment Data in Spain to any developer in order that they can build their own applications.

Keywords : open data, R language, data integration, environmental data
Conference Title : ICDMBDDDE 2016 : International Conference on Data Mining, Big Data, Database and Data Engineering
Conference Location : Amsterdam, Netherlands
Conference Dates : May 12-13, 2016

Open Science Index, Computer and Information Engineering Vol:10, No:05, 2016 publications.waset.org/abstracts/46932.pdf