Scallop4SC: Data Platform for Storing and Processing Large-scale House Log in Smart City

Shintaro YAMAMOTO Hideharu SETO Shinsuke MATSUMOTO Masahide NAKAMURA
Kobe University, 1-1, Rokkodai-cho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
TEL: +81 78 803 6295, E-mail: { shintaro, seto }@ws.cs.kobe-u.ac.jp, { shinsuke, masa-n }@cs.kobe-u.ac.jp

Extended Abstract: *Smart City* is a next-generation city planning that aims to achieve *sustainable society* with ICT technologies. In the Smart City, a variety of data is collected from sensors and people across the city. Typical data include traffic information, energy consumption of houses and buildings, usage of home appliances, etc. These data are stored as *log*, analyzed by advanced data processing techniques, and used for value-added services for the sustainable society. We address the following three challenges to achieve the Smart City services based on large-scale log data. First, a scalable service platform is required to store and process petabyte-level of log data, because the huge number of log will be collected periodically in a wide area. Second, it is difficult to determine the data schema in advance, since various types of city data are currently expected, and even new data will be added in the future. Third, the service platform must be elastic in size and performance, in accordance to the evolution of the Smart City over years.

Our long-term goal is to establish a general-purpose service platform for collecting and using a wide variety of large-scale log data within the Smart City. In this paper, we especially focus on the *house log*, which is the log data collected from individual smart houses. Examples include energy consumption of household appliances, sensor events, device status, environment status, etc. To support storing and processing such large-scale house log, we propose a platform, called *Scallop4SC (SCALable LOgging Platform for Smart City)*. The Scallop4SC constructs a distributed multi-node cluster using Hadoop [1] and HBase [2] technologies. In the Scallop4SC, the house log data is first collected through the network, and then stored in the HBase distributed database. HBase simply stores various types of house log by a pair of key and value. It is therefore unnecessary to determine the static data schema in advance. Also, HBase can split large data to be shared by multiple nodes, which achieves high scalability. Hadoop is a distributed processing framework, where we process the large-scale house log data using multiple nodes in parallel. We propose the architecture of the Scallop4SC and present methods how to adapt the Hadoop and HBase for our purpose of house log.

In this paper, we implement a prototype of the Scallop4SC with 12 Linux servers. Using the prototype, we carried out an experimental evaluation, where we compute electricity consumption for every device in a smart home. The used data is actually collected in our smart home environment [3]. In the experiment, we use actual log data obtained every six seconds for one year from 30 kinds of appliances, comprising 180 million records. The experiments are conducted with three relevant techniques. The first technique is to use a native MapReduce program implemented in Java. The second one is to use the Pig language as high-level interface for operating the MapReduce process. The third technique is to use the conventional RDB with the MySQL for the comparison. As a result, we have found that the prototype with the native MapReduce was able to analyze hundreds of millions of data in several minutes, which was about half of the time spent by the MySQL. We also compare the three techniques from the viewpoints of scalability, performance and development effort. Based on the result, we finally estimate the processing time for several real-scale cities. Assuming a middle-scale city of 60,000 houses where every house sends log data of 30 appliances every one minute, the prototype is able to process daily data within just one hour.

Keywords: smart city, smart house, house log, service platform, Hadoop, HBase

References:

[1] Apache Hadoop, http://hadoop.apache.org/

[2] Apache HBase, http://hbase.apache.org/

[3] H. Igaki, H. Seto, M. Fukuda, S. Matsumoto and M. Nakamura, `` Implementation and Evaluation of a Looking back Service for Power Consumption Behaviors," IEICE Trans. on Information and Systems (Japanese Eds.), vol.J95-D, no.4, pp.778-789, April 2012.