We do not have Systems for Analysing IoT Big-Data

Yuya Sasaki
sasaki@ist.osaka-u.ac.jp
Osaka University, Japan

1 INTRODUCTION

Internet of Things (IoT) generates a large amount of data every second from various things such as vehicles, shops, and animals. Very large data from IoT devices, that is *IoT big-data*, will become the main line of data analysis in the near future. IoT big-data has typical characteristic of big data as follows.

Volume : The number of IoT devices will be more than 70 billions at 2025.[1] A large amount of spatio-temporal data is generated over the world.

Velocity : Many IoT devices generate data every second or less than second.

Variety : The variety of IoT devices is very wide such as environmental sensors at street lamp, GPS deployed at birds, air conditioners, connected cars, and drones. Of course, data and data format also have wide variety.

Validate :IoT data has noise and errors inherently.

Currently, many local governments and companies are collecting the IoT data for various purposes. Thus, analysing IoT big-data will be more important in near future. We are currently collaborating with city municipalities who encourage smart city initiatives and researchers in fields of ecology and urban geography. They often asked me "Do you know good systems for analysing IoT data?" and I always answered "No. You may use Spark or Hadoop for large datasets". However, since they are not familiar with such systems and/or it is not easy to use their various IoT data in the systems, they do not use the systems and take inefficient ways. So, we need new systems for efficiently analysing IoT data and handling variety of data. Furthermore, new systems are better to fix data including noise and errors automatically and to be easy for beginners without large amounts of experiments.

2 WHAT SYSTEMS DO WE NEED?

All IoT data have spatial and temporal properties. We ultimately want systems that can analyse any types of spatio-temporal data efficiently. Existing systems that can handle large spatio-temporal data are Spark-oriented (e.g., [2]) and Hadoop-oriented (e.g., [1]) systems. Since they are developed for specific types of data (e.g., trajectory without any monitoring values and stationary sensors), they do not take care for variety. And, it is hard to select optimal systems and tune configurations for their datasets. People who have IoT big-data typically want to find knowledge from diverse IoT data effectively, efficiently, and easily. Therefore, we need to develop new systems for IoT data that have five following characteristics. (1) **Efficiency**: Processing large amount of data efficiently, (2) **Diversity**: Handling several types of data devices and data with different formats, (3) **Soundness**: Providing rich analysis algorithms and user defined functions, (4) **Easy-tuning**: Providing optimal settings depending on users' datasets, and (5) **Intuitive-understanding**: Visualizing results of analysis and search intuitively and effectively.

We have systems for general purposes that probably satisfy all characteristics, but not for IoT contexts. The systems for general purposes do not work well for IoT contexts due to different properties such as data partitioning and indexing. *If you know systems that satisfy the characteristics for IoT contexts, please let me know and we should stop developing this system.*

3 CHALLENGE

For developing such systems, we are tackling to combine database and machine learning techniques, develop native database techniques, and implement many data mining algorithms on the system.

Distributed and parallel processing for diverse IoT data: We develop distributed and parallel processing for diverse IoT data. It is necessary to partition IoT data well and index them for efficient data access.

Join algorithm for diverse data: Current join algorithm for spatio-temporal data is for a single type of data. In IoT big-data, we join diverse data, for example, data on connected cars and PoI data.

Auto-configuration: We automatically configure setting of the systems, integrate data that have different formats, and fix errors of data. In more concretely, we automatically conduct noise fixing, map matching, label completion, and index selection.

Effective-visualization: We automatically select styles of visualization and learning users' preference.

Plug-in for Advanced search: In IoT data, some users would like to analyse at real-time and/or with keeping privacy. Our systems can add such advanced requirements as plug-in.

4 CONCLUSION

Analysing IoT big-data is expected at various fields. But, we do not have good systems yet. We need to develop new systems for analysing IoT big-data.

REFERENCES

[1] Ahmed Eldawy and Mohamed F Mokbel. 2015. Spatialhadoop: A mapreduce framework for spatial data. In *ICDE*. 1352–1363.

[2] Dong Xie, Feifei Li, Bin Yao, Gefei Li, Liang Zhou, and Minyi Guo. 2016. Simba: Efficient in-memory spatial analytics. In *SIGMOD*. 1071–1085.

[1] https://www.enterprise-cio.com/news/2018/jan/04/roundup-of-internet-of-things-forecasts-and-market-estimates-2018/

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn