We do not have Systems for Analysing IoT Big-Data

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

1. INTRODUCTION

Internet of Things (IoT) generates a large amount of data every second from various things such as vehicles, shops, and animals. Large and various IoT data, that is *IoT big-data*, will become the main line of data analysis in the near future. IoT big-data has the typical characteristic of big data: (1) **Volume**: The number of IoT devices will be more than 70 billions at 2025. A large amount of spatio-temporal data is generated over the world; (2) **Velocity**: Many IoT devices generate data every second or less than one second; (3) **Variety**: The variety of IoT devices is very wide such as environmental sensors at street lamp, air conditioners, connected cars, and drones. Of course, data and data format also have wide variety; and, (4) **Validity**: 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 the 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 do 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 spatial-temporal data are Spark-based and Hadoop-based systems. Since they are developed for specific types of data (e.g., trajectory and stationary sensors), they do not support analysis for various IoT data. In addition, it is hard to choose optimal systems and configure their parameters 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 big-data that have the following five characteristics. (1) **Efficiency**: Processing large amounts of data efficiently, (2) **Variety**: Supporting several types of devices and data with different formats, (3) **Usefulness**: Providing rich analytic algorithms and user defined functions, (4) **Easy-tuning**: Tuning optimal settings automatically, 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.

3. CHALLENGE AND GOAL

For new systems that satisfy the above characteristics, we need develop core database techniques (e.g., partition and query plan), combine database and machine learning techniques (e.g., learned index selection), and implement new query operators on the system. Our challenges and goals are as follows:

Distributed and parallel processing: We need partitioning, indexing, and query optimization techniques on the system for efficient access and retrieval.

Query operators: We need compare and join various IoT data such as kNN and spatial join on drone and PoI data. We need support searching, joining, and aggregating various IoT data with their heterogeneous relationships.

Auto-configuration: Various IoT data makes hard optimal settings (e.g., index) and preprocessing (e.g., data integration). We try to mitigate such concerns by both online and offline learning from user's data and workloads.

Visualization: We automatically select effective styles of visualization based on users' preference.

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. Therefore, we are trying to develop IoT big-data analytic systems that can process large and various IoT data efficiently, support rich query operators, and provide optimal setting and preprocess automatically, with effective visualization.

10th Annual Conference on Innovative Data Systems Research (CIDR '20)
January 12-15, 2020, Amsterdam, Netherlands.