CITIESData: Towards Cloud Based Big Data Management for Smart Cities

Xiufeng Liu*, Alfred Heller, and Per Sieverts Nielsen
xiuli@dtu.dk

DTU Danmarks Tekniske Universitet

CITIES

Introduction

With the increasing presence of Internet of Things (IoT) and future internet technologies in smart cities, a large amount of data are generated. The data need to be properly managed and analyzed for various application using integrated ICT approach. The ICT technologies for a smart city will deal with the data from different domains, including environmental, energy, transportation and many others. We present a cloud-based ICT platform that can collect, store, share/publish, analyze, and visualize scalable data from city environment.
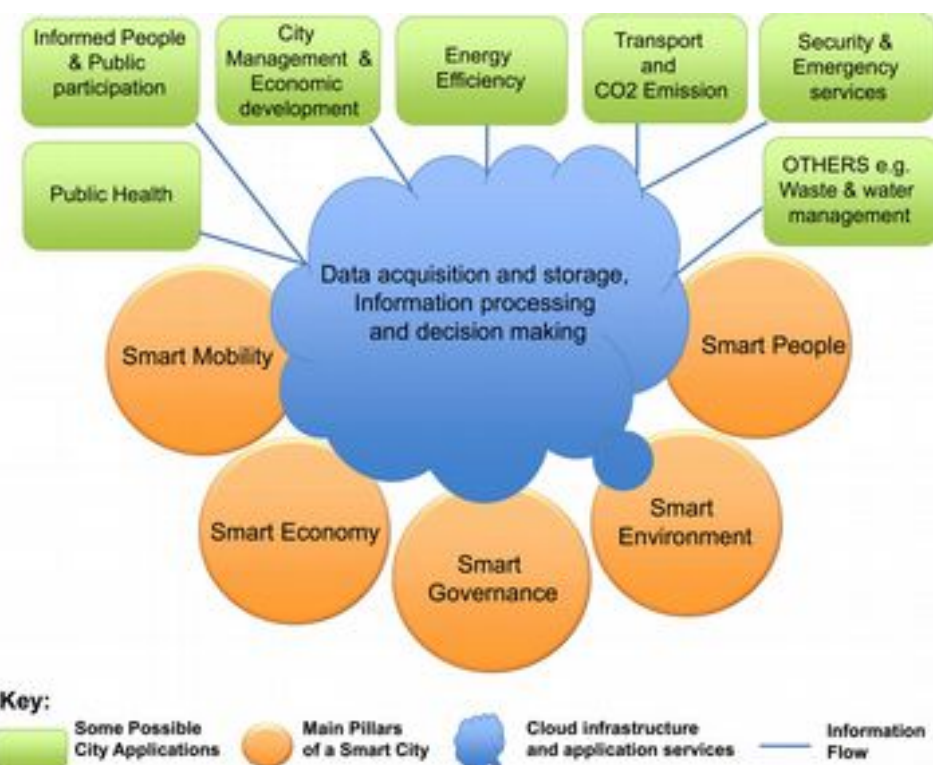
Fig 1. Using the Cloud to store data generated from different smart city components [1]

Smart city data characteristics (5Vs but more Vs are possible):
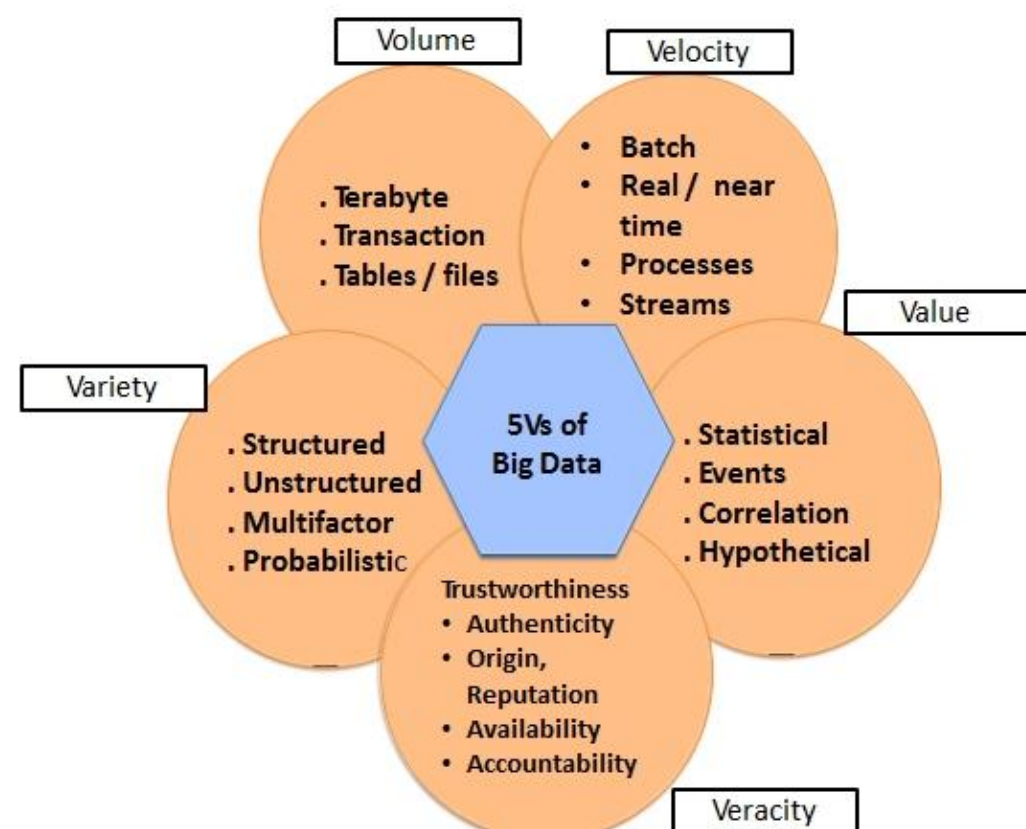
Fig. 2 The characteristics of big data

Methods

The goals:
- Data quality checking and improvement
  - Cleansing data before publishing
  - Instruction of data quality
- Security and privacy protection
  - Classify data according to different risk levels
  - Using different sharing/publishing strategies

The architecture:
- Apply the virtual machine (VM) based secured environment for using highly sensitive data
- Use the cloud-based data management system, *OwnCloud*, for semi-sensitive data sharing
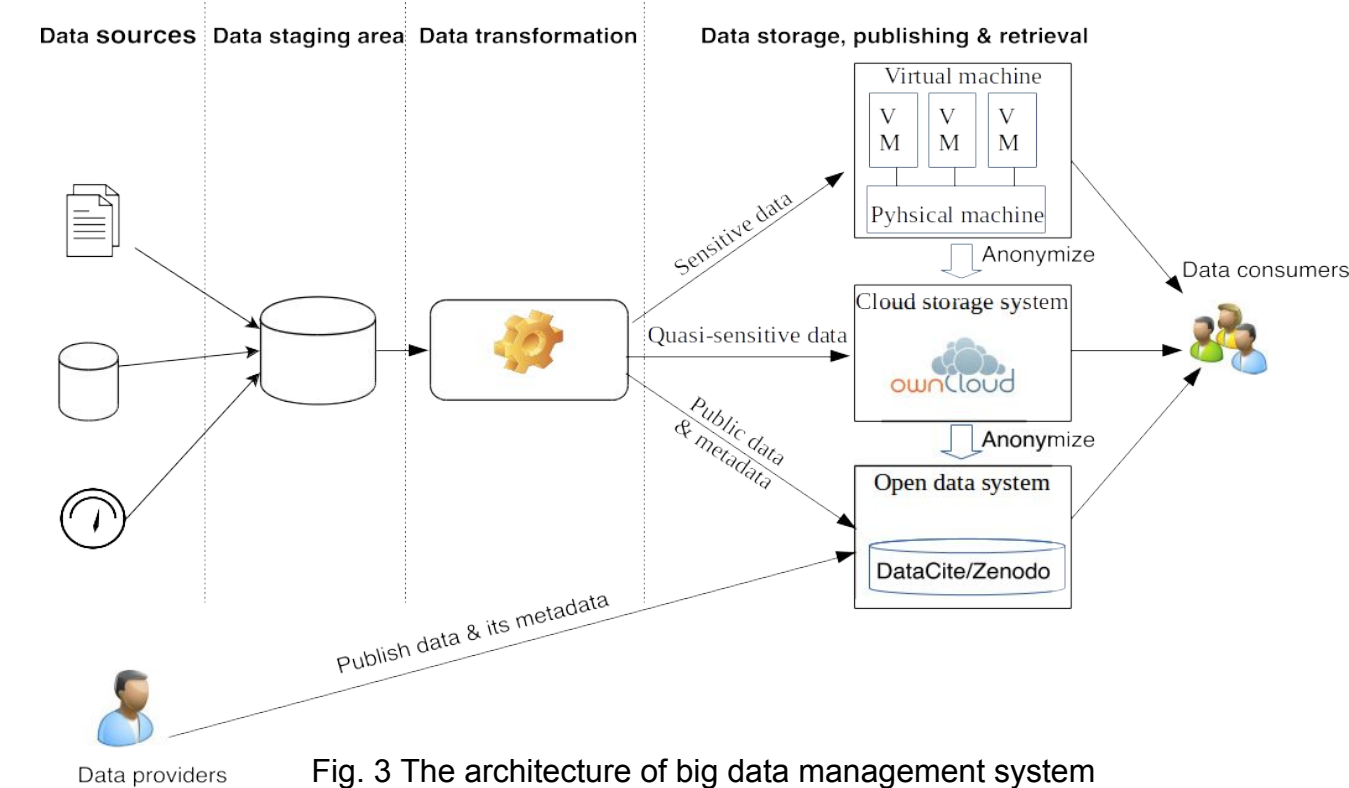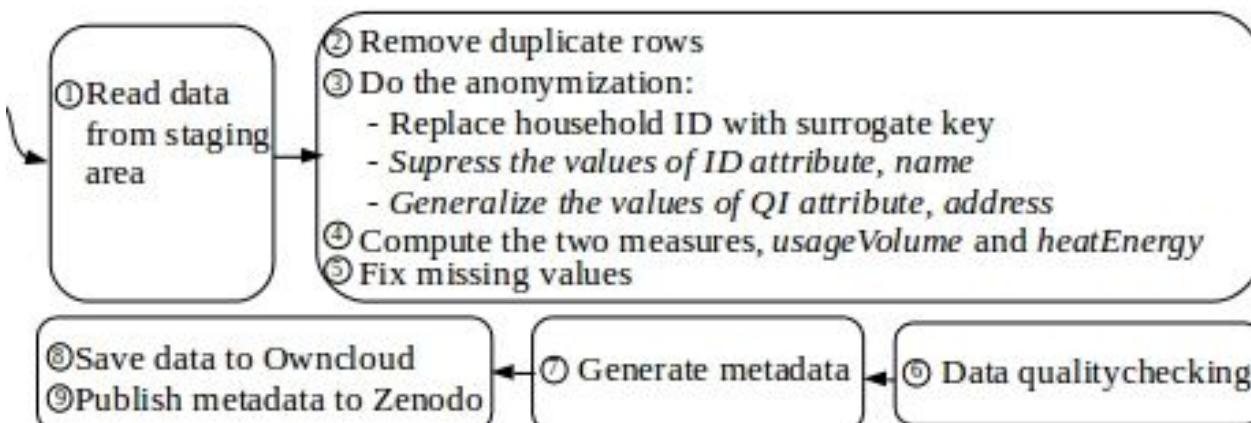- Use the open data platform, Zenodo, for indexing, and sharing

Fig. 3 The architecture of big data management system

The process:

① Read data from staging area
② Remove duplicate rows
③ Do the anonymization:
 - Replace household ID with surrogate key
 - *Supress the values of ID attribute, name*
 - *Generalize the values of QI attribute, address*
④ Compute the two measures, *usageVolume* and *heatEnergy*
⑤ Fix missing values
⑥ Data qualitychecking
⑦ Generate metadata
⑧ Save data to Owncloud
⑨ Publish metadata to Zenodo

Results

- **Data quality checking model**

$$f = \sum_{i=0}^{n-1} \omega_i * y_i, \quad \omega_0 + \omega_1 + ... + \omega_{n-1} = 1.0 \qquad (1)$$

where f is the overall data quality score, y_i is the data quality of determinate attribute i, and ω_i is its weight.

- **Anonymization methods and software package**

- **CITIES data management system**
  a. A scalable data processing platform
  b. Data cleansing, analytics and visualization

Fig. 4 CITIES data management system
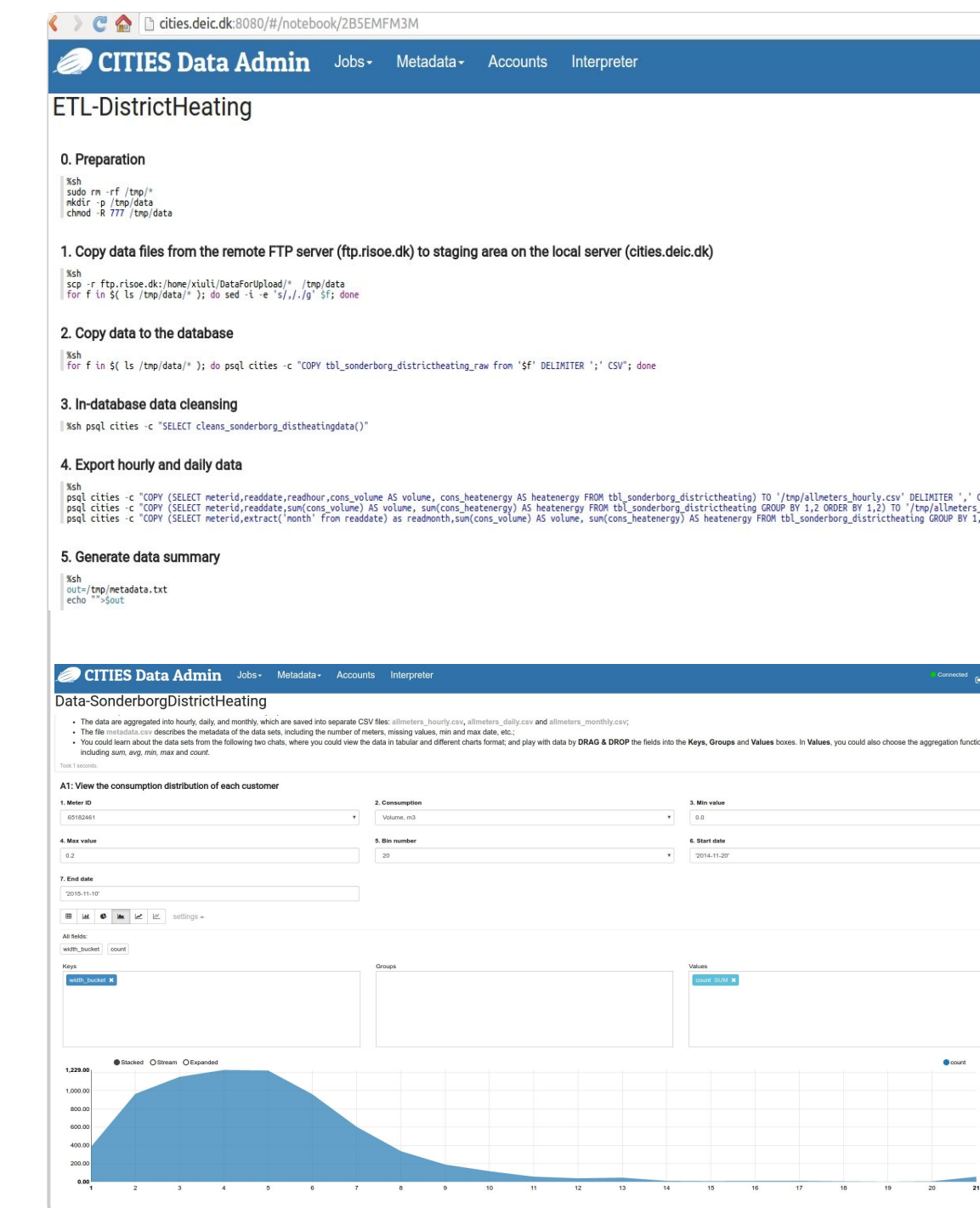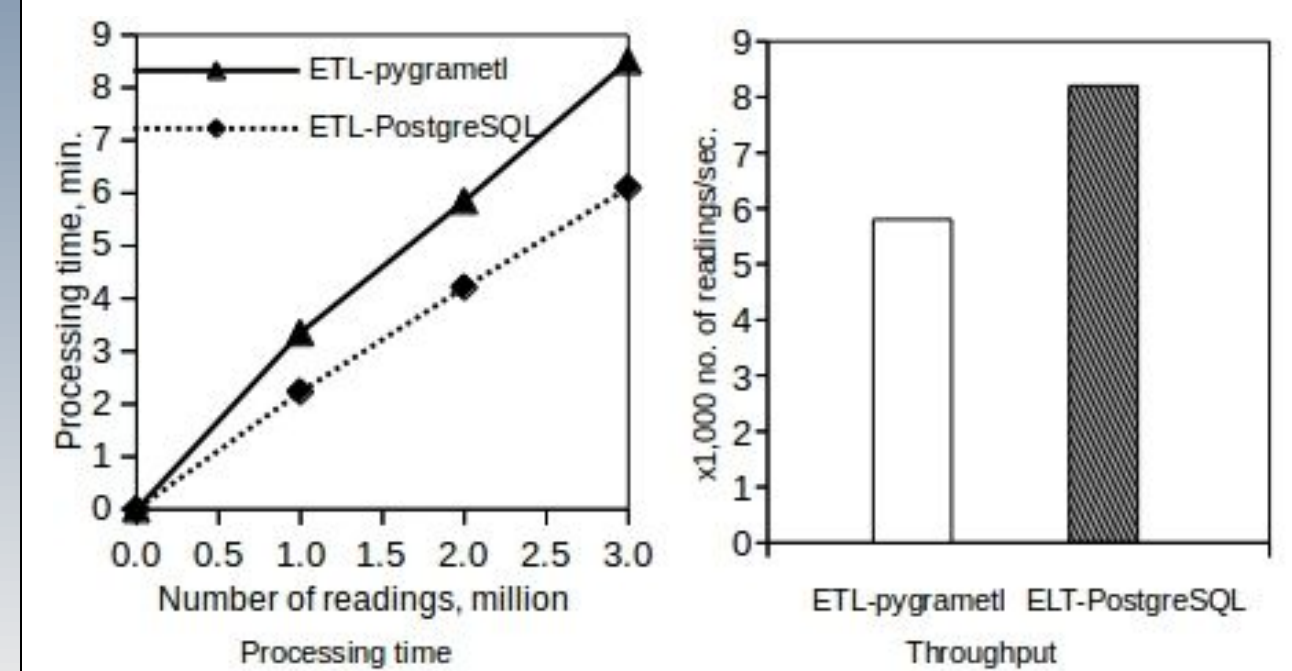
Performance
The time and throughput of processing 3 million hourly smart meter readings

Conclusions

- We have proposed a smart cities data management framework
- Proposed the method of publishing/sharing data according to different data sensitivity levels.
- Proposed linear regression based data quality checking method
- Implemented a smart cities data platform for streamlining the data management process
- The cities data platform has good performance supporting big data management towards the Cloud

Acknowledgements

References

1. X. Liu, A. Heller, and PS Nielsen. Research data management for smart cities, In submission to Journal of Information Management.
2. CITIES Data Platform. http://cities.deic.dk