A Science Cloud for Smart Cities Research

Heller, Alfred; Liu, Xiufeng; Gianniou, Panagiota

Published in:
Energy Procedia

Link to article, DOI:
10.1016/j.egypro.2017.07.369

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 122 (2017) 679–684

Energy Procedia

www.elsevier.com/locate/procedia

CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale, CISBAT 2017 6-8 September 2017, Lausanne, Switzerland

A Science Cloud for Smart Cities Research

Heller A.[a]*, Liu X.[b] and Gianniou P.[a]

[a] *Department for Civil Engineering, Brovej 1, Technical University of Denmark, 2800 Kgs.Lyngby, Denmark*
[b] *Department for Management Engineering, Produktionstorvet, Technical University of Denmark, 2800 Kgs.Lyngby, Denmark*

Abstract

Cities are densely populated and heavily equipped areas with a high level of service provision. Smart cities can use these conditions to achieve the goals of a smart society for their citizens. To facilitate such developments, the necessary IT-infrastructure has to be in place for supporting, amongst many other things, the whole lifecycle of big data management and analytics for research activities. At the Centre for IT-Intelligent Smart Energy for Cities, we have therefore been developing a flexible infrastructure, based on open sourcetechnologies. This paper presents this solution and its application in a city and building research.

Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale

Keywords: Infrastructure for smart cities; cloud computing; big data research; case examples, building application

1. Introduction

"Smart cities" is not a well-defined term, and its complexity is rather great [1]. The "smartness" of a city can stem from its citizens, organizations or technology. The last of these is the main characterization applied in this paper. The simplified idea is that the gathering of information from a city together with its intelligent handling to achieve smart decision-making and control is what constitutes a "smart city". The practical result is an innovative infrastructure.

* Corresponding author. Tel.: +45 4525 1861. *E-mail address:* alfh@byg.dtu.dk

1876-6102
Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale
10.1016/j.egypro.2017.07.369

Other definitions have been found in the literature, basically with the same vision, which apply some additional pieces to the common issue of enabling smart cities. In particular, geospatial representations have been widely applied to support urban planning, which enables information visualization on a map [2][3]. These examples show ways of integrating building information into geo-information, including all the aspects of buildings such as construction details (BIM), construction year, energy-related data, and much more [3].

This paper presents an operational example of a data management and computation infrastructure that can handle different aspects of a smart city for research. The proposed solution is different from other cloud solutions, in that it uses open source technologies and is able to handle open data and confidential data within the same infrastructure. All this is done in ways to provide researchers with high-level services of easy access.. In reality, a full-fledged city infrastructure would be much larger, but all central aspects are implemented and operational in the current setup. The goal is to develop the infrastructure that can handle the integration of building data into city systems, especially energy systems. This paper describes the infrastructure from the researchers' point of view and gives some examples of its use in the building and energy domain.

2. IT Infrastructure – An Implementation

The current infrastructure of the presented "cloud" is generic and can be applied to other subjects. However, most of the research involved in current activities is related to energy within cities. The objective of the infrastructure is to be able to do smart-city research, which is characterized by using a wide variety of data. Data can originate from sensors, legacy systems, operational databases, or other sources, with different formats and sizes. The heterogeneous data have to be handled in an efficient way by the resulting cloud. Meanwhile, generic routines will be developed to enable researchers to do the collecting work themselves and without the support from skilled-IT experts.

Adequate dealing with sensitive data is crucial. Some data are sensitive with respect to personal information. To ensure safe handling of sensitive data, the current infrastructure is designed to handle data within the system. The data are not allowed to leave the secure area, e.g., copied to personal computers and spread to elsewhere.

Open data is a highly prioritized requirement, and it is enabled via integration with international projects such as CKAN (http://ckan.org), a data publication infrastructure used by many cities in Europe to publish open data [4]. The European research repository at CERN, Zenodo (http://zenodo.org) is also supported and well-integrated. Export to Zenodo is designed into the solution by supporting the DataCite metadata (http://datacite.org), a standard that enables the description and citation of all kinds of research results, including research data.

To enable indexing and discovery of semi-sensitive and sensitive data, we support publishing metadata on the open data platforms such as CKAN and Zenodo, with a link to the data to get an authorized access. In that case that sensitive data have to be shared, anonymization will be applied.

All these services are provided to researchers within the current cloud infrastructure where the BigETL tool (see https://github.com/ xiufengliu/BigETL for more details) is used for the data transformations, such as solving missing values, removing duplicate data, and merging subsets of data into comprehensive datasets. In addition, the same package enables scheduling service for automating the transformations.

Origo is used to manage the Cloud. (https://www.origo.io). The Cloud has 18 physical servers, with 80 cores and 564 GB memory, and 4.2 TB node storage, with additional 1.2 TB network storage. Origo manages a pool of virtual machines (VMs) that can be scaled on the distributed infrastructure. The setup is composed of a core that is a centralized component managing the lifecycle of the VMs, a component for the identity and security management of users for safety of the data in the cloud, and a capacity management component that adjusts the placement of VMs based on a set of predefined policies. The default capacity scheduler implements a simple pairing policy and supports user-driven consolidation constraints.

To facilitate its use, the cloud provides Windows and Linux-based VM images with different pre-installed software packages. These include data science images with all the commonly used data analysis tools installed, such as R, Python, Pandas, Scikit-learn; and data management images with pre-installed different types of databases (e.g., PostgreSQL, MySQL, OpenTSDB, etc). Many other applications can be deployed within a VM. These settings can sufficiently satisfy different needs from our research.

The volume of data is a key value of the current infrastructure. A single fully equipped building can easily generate more than 10,000 data points ranging from temperature and movement sensors to system control sensors. Sampling rates can be almost real time, but 5- to 15-minute values are commonly applied. This means that the above-mentioned data points can easily generate 20-100 kB pr. hour. Scaling this up to a whole city, data volume and handling speed are decisive. With this in mind, the current infrastructure is designed to be scalable and flexible.

The data lifecycle describes the structure of the current infrastructure well. The data enters and is made ready to be served to the users. Then the data is used by researchers, perhaps transformed and prepared in certain ways. The intermediate and final results are stored and finally published.

Figure 1 shows the system, consisting of the following components:
1) Data ingestion layer that handles the many data sources
2) Data staging area
3) Data transformation layer
4) Data storage
5) Publication layer

Liu and Nielsen [5] present this workflow in detail, as implemented in the current setup.

In addition to this “normal data lifecycle”, there is a requirement for the archiving of research data in a university archive, in a national archive, or in Zenodo.

The infrastructure itself is analysed for efficiency, robustness and security by the system architect. More details can be found in Liu et al. [5], [6], [7] and [8].

3. Tooling for Researchers

The current cloud infrastructure aims at a high service level to researchers. In this section we elaborate on these services.

From a researcher’s perspective, the current infrastructure feels like a personal computer because of the integration in the preferred operating system in virtual machines. This way all the remote handling is hidden for the user. It is this aspect in particular, together with the computational efficiency that a cloud solution can offer, that avoids researchers copying data into their ‘unsafe’ local computers. This means data security is kept very high.

Virtualization is very important for a high service to the researchers. Therefore, a set of visualization tools are implemented.

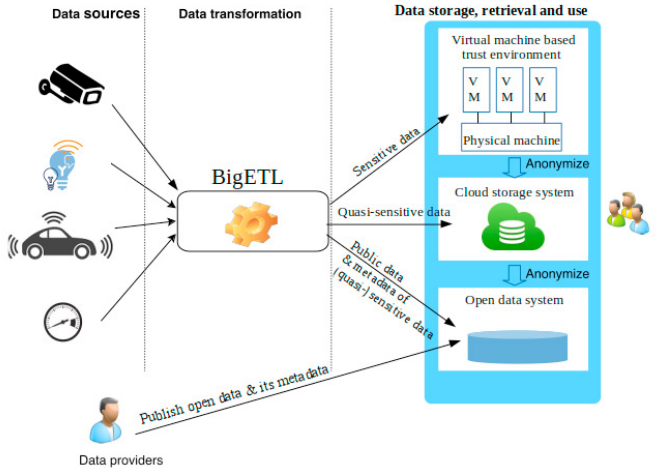

Figure 1 ICT system architecture

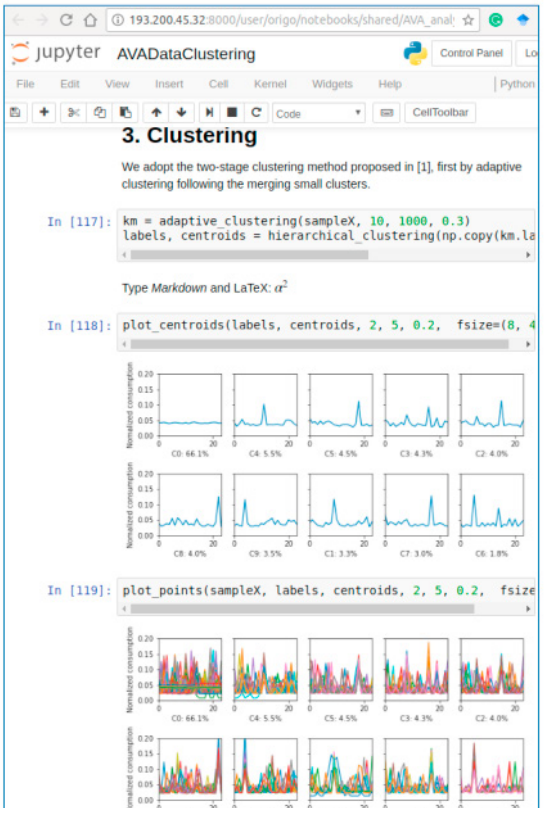

Figure 2 Data analysis by Jupyterhub

The researchers who use the current infrastructure come from a wide variety of subjects, so their traditions for methodologies and tools are also various. To enable cooperation between researchers from such different origins, a set of tools has been implemented aimed at making data management and analysis efficient, enabling reuse of work, and improving communication between researchers. Notebook tools have therefore been implemented:

JupyterHub (http://jupyter.org) enables the combination of documentation of the data handling and analytics with the coding of scripts in R and Python, two widely applied open source scripting languages where R is originated from the statistical domain, and Python is a more generic web scripting language. The scripts can be executed directly from within the notebook text, which makes it very convincing and efficient to develop the relevant data analytics routines.

In the current example, shown in Figure 2, the clustering analysis of the building dataset of a city is documented in the notebook; it enables running alternative scenarios in the same notebook environment. In such a notebook, the connection to the data, the handling of the data, the analytics and visualization are lined up for reproducibility and alternative examinations. In addition, such documents can directly be applied for educational purposes, giving students the description and their own dataset.

With this tool, the researchers can keep notes on their work in the form of codes, words and visualizations. These notes can be directly utilized for communication with peers and students, who are able to reproduce the work directly or transfer the codes to their own datasets from within a note copy. This makes it easy to share research in a dynamic and visualized form. An example is given in Figure 2, which shows documentation, codes, results and visualization in form of plots.

4. Case applications

At the research centre CITIES, the infrastructure presented here has been used for a number of smart-cities, smart grid(s), smart-buildings and smart components in the form of IoT applications. In the current paper, an example of such an application is given that can be relevant for professionals in building research and smart cities.

The city of Sønderborg in Denmark, Jutland has the ambition to be CO_2 neutral by 2029. This target drives a lot of research, development and innovation. A few years ago, the first monitoring and modelling approaches were carried out by Bacher et.al., developing grey box models for space-heating [9] and hot-water consumption [10]. They monitored and analysed a set of 16 buildings from the municipality. Before doing additional research, we first aimed at reproducing this earlier research [11]. For this purpose, we had to find data that was not publically available and had not been explicitly published. We moved the data onto the current data management system from where follow-up research can be carried out and new data can be found. This is an example of reusing, repurposing and reproducing research that is enabled and streamlined by a cloud infrastructure.

The data involved in the previous example was based on annual data. With the availability of e-meters, a follow-up study has been carried out since 2014, in which hourly values of the heat demand drawn from the district heating are being collected. 54 buildings from “Sonderborg Fjernvarme” are being monitored and the data collected in the science cloud by batch scripts. Figure 1 shows the principle of data management from collection to analysis. This case setup can be seen as a generic application of the science cloud infrastructure on typical energy and water meter data analysis for city area that can be repurposed for other cases and even applied to a collection of data between district and city cases. Hereby the infrastructure is applied in a way that differs from other implementation by the fact that it is on the one hand flexible to any application, on the other hand adapted to smart energy cities.

In the event of e.g. a loose connection to the source, warnings are sent to the relevant responsible person. In this way, losses of data are limited. Advanced automated procedures have been implemented to make data ready for use. In this case, the data is cleansed of errors and missing values, then stored in a database from where a number of predefined data packages are generated. These can be hourly, monthly or annual lists of values to be utilized by others. These data packages are then published through the channels mentioned elsewhere in this paper. This data is currently being used by the city of Sonderborg, for a series of student projects and other research. In the IEA EBC Annex 67 project, the flexibility that buildings offer to the energy system is a central aspect of the research. In this work, monitoring data and simulation data are applied to analyse these potentials. The work is still ongoing and the first publications are expected in the course of 2017. At the present time, a collection of time series data for typical Danish houses, based on the Tabula Webtool of typology for Denmark [12], has been developed and published through the current cloud infrastructure for open sharing. The published time series was generated synthetically using simulation models that will be made available parallel to the time series work to be published. This application of IT solutions is planned to be extended with the collection of other models and methodologies relevant for this research.

5. Conclusion

Smart City research is very interdisciplinary and involves huge amounts of data from various sources. The availability of powerful IT infrastructures that can handle this complexity and volume of data is vital. The current project demonstrates such an infrastructure developed using open source software only. This approach was preferred against existing large-scale infrastructures by well-known companies because of the flexibility that it gives. The drawback is that the size of the components is limited to the hardware available. We plan to overcome this problem by expanding the infrastructure with additional research projects that have the budget for this.

The solution is unique by the fact that the basic software architecture is open, flexible and extensible, and in our implementation simultaneously adjusted to the demand for energy analysis in a smart city context involving big data stemming from smart meters. This adjustment ensures secure handling of sensitive data in an environment that researchers from the relevant knowledge domain can handle without getting to be IT and data specialists.

The current infrastructure has been used for research projects, and this has shown its applicability and limitations with respect to smart-city and smart-building applications. Further development is ongoing and planned for the near future.

6. Acknowledgement

This research was supported by the CITIES project (NO. 1035-00027B) funded by Innovation Fund Denmark. The infrastructure components have been partly supported by the Danish e-Infrastructure Cooperation (DeIC) through the project "Science Cloud for Cities".

7. Literature

[1] Neirotti P, De Marco A, Cagliano AC, Mangano G, Scorrano F. Current trends in Smart City initiatives: Some stylised facts. Cities 2014;38:25–36. doi:10.1016/j.cities.2013.12.010.
[2] Kim SA, Shin D, Choe Y, Seibert T, Walz SP. Integrated energy monitoring and visualization system for Smart Green City development: Designing a spatial information integrated energy monitoring model in the context of massive data management on a web based platform. Autom Constr 2012;22:51–9. doi:10.1016/j.autcon.2011.07.004.
[3] CityZenith. CityZenith 2016. http://cityzenith.com.
[4] Open Data DK n.d. http://www.opendata.dk/ (accessed April 1, 2017).
[5] Liu X, Nielsen PS. A hybrid ICT-solution for smart meter data analytics. Energy 2016;115:1710–22. doi:10.1016/j.energy.2016.05.068.
[6] Liu X, Golab L, Golab W, Ilyas IF, Jin S. Smart Meter Data Analytics: Systems, Algorithms, and Benchmarking. ACM Trans Database Syst 2016;42:1–39. doi:10.1145/3004295.
[7] Liu X, Nielsen PS. Streamlining Smart Meter Data Analytics. Proc. 10th Conf. Sustain. Dev. Energy, Water Environ. Syst., 2015.
[8] Liu X, Heller A, Nielsen PS. CITIESData: a smart city data management framework. Knowl Inf Syst 2017:1–24. doi:10.1007/s10115-017-1051-3.
[9] Bacher P, Madsen H, Nielsen HA, Perers B. Short-term heat load forecasting for single family houses. Energy Build 2013;65:101–12. doi:10.1016/j.enbuild.2013.04.022.
[10] Bacher P, de Saint-Aubain PA, Christiansen LE, Madsen H. Non-parametric method for separating domestic hot water heating spikes and space heating. Energy Build 2016;130:107–12. doi:10.1016/j.enbuild.2016.08.037.
[11] Gianniou P, Heller A, Rode C. Building energy demand aggregation and simulation tools a Danish case study. Proceeding CISBAT 2015, 2015, p. 797–802. doi:10.5075/epfl-cisbat2015-797-802.
[12] EU project Tabula. Tabula Webtool for European Building Typology 2014. http://webtool.building-typology.eu/ (accessed June 18, 2014).