DTU Management Engineering

DeiC and DEFF
sekretariat@deic.dk

2. september 2018
PSN

Final report: Science Cloud 4 CITIES

Please find enclosed the final report (including financial report) for the Science Cloud 4 Cities project. Please don't hesitate to get back to me for further information. We have prepared a draft power point presentation for the meeting on October 2, 2018. Please let us know if you want us to forward the presentation.

Kind regards

Per Sieverts Nielsen
Seniorresearcher
DTU Management Engineering

CVR-nr. DK 30 06 09 46

Scientific Cloud Project Report

1. The project and the infrastructure

The current trend of data processing is increasingly transferring traditional centralized solutions to cloud environments. Within our Centre of IT-Intelligent Energy System in Cities [1], a private cloud platform, *SciCloud*, was created to support the scientific research in smart cities. The purpose of implementing this private cloud is to integrate infrastructure, platform and resources; and to provide a consolidated infrastructure for scientific experiments, software development, provision of analytics services and educational purposes. The current infrastructure of the SciCloud consists of two major parts: one is the *cloud infrastructure*, which includes the physical infrastructure, network, security and cloud management platform; and the other is the *data infrastructure* that supports the scientific research for smart city projects. They are applied in a number of research on PhD level and student work on Bachelor and Master level, described in the following.

1.1 Cloud Infrastructure

Figure 1 shows the cloud infrastructure of SciCloud. SciCloud has 17 identical physical servers (1 master server and 16 slave server), each of which is configured with a 4-core CPU and 8GB memory, and 256 GB Hard Drive. Besides, there is an additional 1.2 TB network storage for backup. The servers are installed with Ubuntu 18.04 (4.15.0-32 Linux kernel), and use PXE booting. The open source cloud computing management platform, OpenNebula (https://opennebula.org/), was deployed to manage the computing and storage resource management. OpenNebula is one of the most popular cloud management platforms, which has a large community. OpenNebula can efficiently manage virtual machines (VMs) that are scaled on a distributed infrastructure, and it is easy to deploy, monitor and control VMs on a pool of distributed physical resources. To support the use of this cloud, SciCloud offers Windows based and Linux-based VM images with different pre-installed customized software packages. These include data science images with all the commonly used data analysis tools, such as R, Python, Pandas, Scikit-learn; and data management images with pre-installed different data management systems (e.g., PostgreSQL, MySQL, MongoDB, Influxdb, etc). But, note that like a physical computing node, any applications can be run on VM instance, and the dependent software packages can be install within the VM.

1.2 Data Infrastructure

Figure 2 shows the data infrastructure deployed on SciCloud. The data infrastructure is an open source software stack deployed for facilitating the data management, processing, and analysis on the cloud. More specifically, the data infrastructure consists of two layer, data layer and analytic layer. The data layer is responsible for data collection, data transformation (cleansing), and storage. The open source tools including Apache Zeppelin, OwnCloud, and Zenodo were deployed in the data layer. Besides, the open source data management system, PostgreSQL, is deployed for managing data. The analytic layer has deployed the online analytic platform, Jupyterhub and RStudio, where researchers can directly analyze the date using web-based user interface.

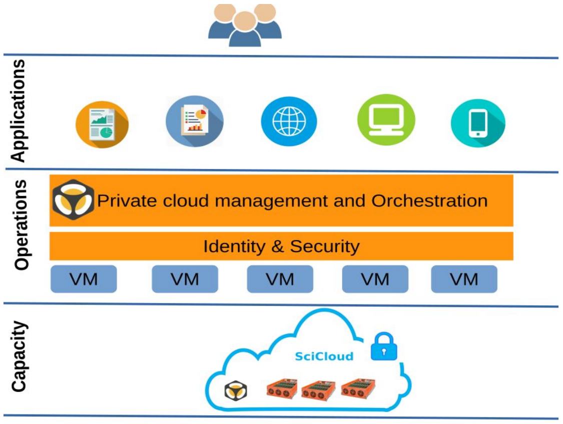

Figure 1. Cloud infrastructure

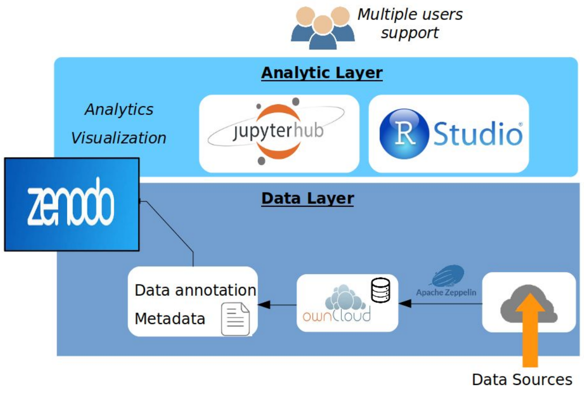

Figure 2. Data infrastructure

1.3 Security and privacy

Security and privacy protection is the first priority of using the cloud. The following measures were adopted for cloud platform: First, the hard drive for saving images uses the RAID 5, which can ensure the performance, reliability, and safety of the data storage. Second, strict firewall rules are applied to the external access to the internal network of the clusters, where only the necessary ports, such as 22 for SSH, are open to the access of the cluster from subnet of DTU building 424 and 426. In order to access the VM instance from outside DTU firewall, public IP addresses are required. It is only the VM that can be accessed externally while the physical cluster is shielded from attacks. The firewall rules for the VM can be set on OpenNebular. Third, the VM images are backuped everyday by a schedule job. For the data infrastructure, the following measures were adopted: First, a two-layer data infrastructure is used to obtain better safety, where the data layer is only open to the data administrator, while the analysis layer is open to data users through the WAYF single sign-on service. Second, the data saved in the data layer are the “cleansed data”, which were pre-processed before stored in the cloud, such anonymized, fixing missing values, and outliers, etc. The pre-processing of the data can release users for tedious data pre-processing work. Fourth, A role-based model is applied to protect the data, which a user can only access belonging to his/her project/research group. For example, the researchers of the CITIES project can only access the data that are granted the permission to this project. An account is created for each user in the analysis layer, and the data will automatically synchronized from data layer to his/her home directory in the analytics layer. To avoid creating data copies, the linux soft-link to the data is created to the copy that was synchronized from the data layer. Fifth, users run their analysis models on the cloud, instead of downloading the data to their local machines. The benefit is not only for the data protection, but also for taking advantage of the power of the cloud, such as computing and storage. Sixth, although Jupyterhub and RStudio support multiple online users, actually the users have their own working environments, and they can install additional software packages and manipulate their data. Multiple users can share their work and work together, e.g., on the same Jupyter notebook.

1.4 Cloud administration, Source code and Open data

- The administration console of SciCloud is ***https://192.38.83.152***, which is only accessible within DTU internal network (note: A new setup of the SciCloud is on the way, and this link is not accessible yet).
- The source code is available at ***https://github.com/xiufengliu/scicloud***, which include the program of uploading data into ownCloud, the synchronizing program of the data between data layer and analytics layer, and the notebooks (Apache Zeppelin) for data processing.
- The open data is a heating data and BBR data from Sonderborg. The data are anonymized and published at Zenodo open data platform, which can be accessed by ***https://doi.org/10.5281/zenodo.1300308***

2. Achievements

The following achievements were made in this project:

- The contracting between partners went initially well, but getting the signature of DTU was delayed. The project coordination has been moved from DTU Civil Engineering to DTU Management Engineering – but the DTU institutes involved still consists of DTU Management Engineering, DTU Civil Engineering and DTU Compute.
- Software is integrated into the national university network, and can be accessed through single sign-on service WAYF.
- The data included into the data infrastructure has been applied in various research and study project by the students of Århus University and DTU.
- The infrastructure is developed in cooperation with the CITIES, a strategic research project, used by Smart Cities Accelerator (EU project) and proposed for various upcoming research projects, among others including REBUS (EUDP project) and Danish innovation Fund project proposals. The infrastructure is also suggested in a number of EU research project proposals as the tool for data management and analytics.
- The infrastructure closely integrates the data from data.deic.dk (now sciencedata.dk), which has its own version of Owncloud. A data set was published in Zenodo open source data platform.
- SDU has hired relevant specialist for supporting the project (Bertil Dorch), AU has Elyzabeth, AAU is involved via the DyCips center.
- The Science Cloud has been described in two publications (see Ref [3,4]), and the cloud has also been used for the data analytics for a number of papers (incls Ref [1-6]). One paper specifically describes the Science Cloud and data management platform (see Ref [4]).
- A data set is published on Zenodo open data platform, and a data paper [6] is in submission.

3. Sustainability

Some challenges have been identified, which are described as follows:

- Host challenge with the accounting department due to the fact that the grant comes without overhead. This has been solved.

- The General Data Protection Regulation (GDPR) (adopted by EC and EP after all deadline) enters into application by 25th of May 2018, and this means a lot of extra work and unanswered questions, because e.g.:
    - The GDPR widens the definition of personal data, i.e. handling of data that as of now is unaffected by data protection rules, must by May 18, 2018 comply with the new rules (retroactive effect).
    - The GDPR requires that data controllers make privacy impact assessments (PIAs)
    - All organisations working with/providing personal data must comply with rules such as data minimisation, anonymization and encryption.
- PI foresees that some deliverables may have to be dropped to secure means to pay for legal assistance – either that, or extra funding needs to be found. It was discussed if DM Forum could investigate possible answers.
- The sustainability of the infrastructure is ensured for the upcoming years through the CITIES project and investment by DTU Management. It is expected that the Science Cloud will be applied for research and education through these and additional, applied fundings. Copies of the cloud are expected to be applied in commercial activities by, among others NIRAS A/S, consultant company.

References:

1. X. Liu and P. S. Nielsen. Scalable Prediction-based Anomaly Detection on Smart Meter Data. Journal of Information Systems, vol 77, pp. 34-47, 2018.
2. P. Gianniou, X. Liu, A. Heller, P. S. Nielsen and C. Rode. Clustering-based Analysis for Residential District Heating Data. Journal of Energy Conversion & Management, vol. 165, pp. 840-850, 2018.
3. A. Heller, X. Liu, and P. Gianniou. Science Cloud for Smart Cities Research. Accepted by the international conference – future buildings & districts – energy efficiency from nano to urban scale (CISBAT), 2017.
4. X. Liu, P. S. Nielsen, A. Heller, and P. Gianniou. SciCloud: A Scientific Cloud and Management Platform for Smart City Data. In Proc. of DEXA Workshop, pp. 27-31, 2017.
5. X. Liu, P. S. Nielsen, and A. Heller. CITIESData: A Framework for Research Data Management for Smart CITIES. Journal of Knowledge and Information Systems (KAIS), 2017.
6. P. Gianniou, X. Liu, A. Heller, P. S. Nielsen, and C. Rode. A heating load measurements dataset of Danish households, In submission, 2018.

Regnskabsskema - Forskningsinfrastruktur

Grundoplysninger

1. Bevillingens akronym og titel: Science Cloud for Cities ☑
2. Bevillingshavers navn, institution og email-adresse: Danmarks Tekniske Universitet, samlet for konsortium ☑
3. Faglig kontakts navn institution og email-adresse: Per Sieverts Nielsen, pernn@dtu.dk ☑
4. Økonomimedarbejders navn og email-adresse: Svetlana Sokolska, sveso@adm.dtu.dk ☑
5. Sagsnr. (f.eks. 5000-01234B): □
6. Regnskab for perioden (dd-mm-åååå): Fra 01-10-2016 Til 30-06-2018 ☑
7. Type af regnskab - års eller slut (sæt kryds): Års □ Slut ☒ ☑

Regnskab

8. Udgifter sammenlignet med seneste godkendte budget:

	Bevilling		Egenfinansiering		Anden finansiering		Afvigelser ift bevilling				
	Budget	Udgifter	Budget	Udgifter	Budget	Udgifter	Afvigelser	Afvigelser i %			
VIP-løn	1.192.000,00	996.865,81	1.192.000,00	1.371.471,17			195.134,19	16,37%	☑	☑	□
TAP-løn	66.000,00		66.000,00						□	□	□
Instrumenter og udstyr	120.000,00	353.000,00	120.000,00				-233.000,00	-194,17%	☑	□	□
Internationale medlemsbidrag									□	□	□
Andet	22.500,00	24.500,00	22.500,00	2.894,65			-2.000,00	-8,89%	☑	☑	□
I alt	1.400.500,00	1.374.365,81	1.400.500,00	1.374.365,82	0,00	0,00	-39.865,81				

9. Den samlede bevilling (alle år): 1.400.500,00 ☑
10. Bevilget beløb for regnskabsperioden: 1.400.500,00 ☑
11. Bogført udbetalt beløb fra FI i regnskabsperioden: 0,00 □
12. Evt. overført uforbrugt/merforbrug fra foregående år: 0,00 □
13. Anden indtægt (renter mv.) 0,00 □
14. Indtægter i alt jf. ovenstående: 0,00
15. Udgifter i alt: 1.374.365,81
16. Total uforbrugt/merforbrug (punkt 13 fratrukket punkt 14): -1.374.365,81
17. Medsend ny udbetalingsprofil, hvis boksen er afkrydset: □
18. Erklæringer vedr. medfinansiering (sæt kryds): ☒ Medfinansiering □ Andre kilder ☑ □
19. Medfinansiering i alt 1.374.365,82
20. Graden af medfinansiering ift bevilling 100,00%
21. Evt. kommentarer til regnskabet.

Underskrift og påtegning

22. Underskrift bekræfter, at bevillingen er anvendt indenfor bevillingsformålet og i overensstemmelse med bevillingsgrundlaget (sæt kryds): ☒ ☑
23. Dato og bevillingshavers underskrift: 22-08-2018 □
24. Påtegning af regnskabschef eller bemyndiget medarbejder:
    Navn: Torsten Foersom ☑
    Virksomhed/institution: Danmarks Tekniske Universitet ☑
    Stilling: Økonomichef ☑
25. EAN-nummer EAN:5798000430426 ☑
    CVR/CPR-nummer for adm. DK30060946 ☑

30/8-18

DTU
Technical University of Denmark
Department of Finance and Accounting
Finance Manager Torsten Foersom
Lundtoftevej 150, Building 266
DK-2800 Kgs. Lyngby