Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 141 (2018) 112–118

Procedia Computer Science

www.elsevier.com/locate/procedia

The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018)

Cloud Platform using Big Data and HPC Technologies for Distributed and Parallels Treatments

Olivier Debauche[a,b,*], Sidi Ahmed Mahmoudi[a], Saïd Mahmoudi[a], Pierre Manneback[a]

[a] *Computer Science Unit, Faculty of Engineering, University of Mons, 20 Place du Parc, B-7000 Mons, Belgium*

[b] *Biosystems Dynamics and Exchanges Axis, Biosystem Engineering Department, ULiège Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés 2, B-5030 Gembloux, Belgium*

Abstract

Smart agriculture is one of the most diverse research. In addition, the quantity of data to be stored and the choice of the most efficient algorithms to process are significant elements in this field. The storage of collecting data from Internet of Things (IoT), existing on distributed, local databases and open data need a particular infrastructure to federate all these data to make complex treatments. The storage of this wide range of data that comes at high frequency and variable throughput is particularly difficult. In this paper, we propose the use of distributed databases and high-performance computing architecture in order to exploit multiple re-configurable computing and application specific processing such as CPUs, GPUs, TPUs and FPGAs efficiently. This exploitation allows an accurate training for an application to machine learning, deep learning and unsupervised modeling algorithms. The last ones are used for training supervised algorithms on images when it labels a set of images and unsupervised algorithms on IoT data which are unlabeled with variable qualities. The processing of data is based on Hadoop 3.1 MapReduce to achieve parallel processing and use containerization technologies to distribute treatments on Multi GPU, MIC and FPGA. This architecture allows efficient treatments of data coming from several sources with a cloud high-performance heterogeneous architecture. The proposed 4 layers infrastructure can also implement FPGA and MIC which are now natively supported by recent version of Hadoop. Moreover, with the advent of new technologies like Intel® Movidius™; it is now possible to deploy CNN at the Fog level in the IoT network and to make inference with the cloud and therefore limit significantly the network traffic that result in reducing the move of large amounts of data to the cloud.

Keywords: GPU; FPGA; MIC; CPU; TPU, Cloud; Big Data; parallel and distributed processing; heterogeneous cloud architecture

* Corresponding author. Tel.: +32(0)65-374-059 ; fax: +32(0)71-140-095.
E-mail address: olivier.debauche@umons.ac.be

10.1016/j.procs.2018.10.156

1. Introduction

Agriculture provides vital resources such as food, fiber and energy [6]. With the global population growth, the need for crop production and raw fiber also increases. Indeed, the Food and Agricultural Organization of the United Nation (FAO) predicts that the global population will reach 8 billion people by 2025 and 9.6 billion people by 2050. This means in particular, that an increase of 70% in food production must be achieved by 2050 worldwide. The great increase in global population and the rising demand for high-quality products create the need for the modernization and intensification of agricultural practices. At the same time, there is a need for high efficiency use of water and other resources. To meet this challenge, the agricultural industry uses technologies such as the Internet of Things that produce huge amounts of raw data that need to be processed with high performance computing architectures.

According to Superuser [13], 75% of High Performance Computing (*HPC*) centers worldwide are used to perform deep learning and artificial intelligence. Nowadays thanks to the high-performance networks and the use of Single Root I/O virtualization (*SR-IOV*), HPC is now available in a cloud environment [5]. The actual trend is the shifting to hyperscale [13]. Data on the Internet of Things are dynamic: very diverse in term of origin, amount, quality and speed. The throughput and the quality of data depend on the use case which are unlabeled [1]. Indeed, data transmission mode can be continuous, at certain frequencies or by burst that applies large and fast throughput.

For example, in phenotyping and 3D images are particularly consuming in processing and storage [6]; in smart farming, data can arrive at very high speeds close to real time (100 Hz) [7] [10]; or arrive massively from connected objects that transmit them at a fixed frequency [8] [9]. In the case of monitoring landslides, the data burst as soon as anomalies are detected [11]. Moreover, users of cloud infrastructures are increasingly asking for response time to request made on data close to real time. Scientists need to process data such as images arriving with velocity (tens of MHz) which must be treated at real time [2].

Data is not available under the form of data feeds but also in local and distributed databases, open data, firehose, and structured files in which are stored raw data and results of treatments from external processes, and applications, etc. These information are under construction; because, they are not in relation to other data to achieve more relevant analysis. Crossing this data is necessary to handle complex events and make better decisions. For this reason, the data must be previously processed and stored in a form that ensures their crossing, aggregation and exploitation. Moreover, deep learning models are traditionally trained in the cloud with a supervised method which requires a tremendous amount of training data labeled by humans. However, in the case of IoT, raw data are coming from a large number of nodes. All these IoT big data are difficult to label; hence, the traditional supervised training is not suitable, which require an unsupervised method to really exploit the potential of IoT raw data with reduced data movement [1].

This result leads to think about a new way to design cloud computing architectures for the Internet of Things. It becomes crucial to reconsider how data is stored and how it is processed to maintain performance regardless of the increase of data volume to ensure response timing of less than one second. The use of high-performance computing (*HPC*) architectures for the distribution of treatments coupled with the use of Many Integrated Core (*MIC*), Field-Programmable Gate Array (*FPGA*) and multi Graphics Processing Unit (*GPU*) allow today to process a tremendous amount of raw data quickly. In addition to the data from the connected objects the processing infrastructures are also brought from other data sources, private distributed database and open database.

Our goal is to develop a solution that composed mainly of two main parts: (1) An heterogeneous high-performance cloud architecture able in combination with fog computing to process rapidly large amount of data in quasi real time and store data in a distributed database. This approach allows to exploit the benefits of a heterogenous cloud architecture and the combination with the fog [20]. (2) A platform managed by mean of REST APIs hosts the researchers' model and external applications which exploit data stored in the distributed database and allow to visualize statistical data calculated on basis of raw data.

In this paper, we propose a cloud architecture which combines the use high-performance computing, FPGA, MIC, TPU, and Multi GPU to enhance the speed of treatment of distributed big data coming from multitenancy and multi-sources. With this approach, treatments are distributed between GPU / FPGA and parallelized in each GPU / FPGA to ensure a high efficiency.

2. Related Works

Currently, three trends emerge: (1) the increase of the power and complexity of modern HPC systems in order to build exascale class machines, (2) the increase use and sophistication of commercial and open cloud infrastructure, (3) the increase functionality and the use of Big Data in conjunction of HPC [3]. Several authors have already used heterogeneous architecture based on GPU and/or FPGA to accelerate the processing of large amount of data. Among these authors, we cite the most important contributions.

Fox et al., 2017 [3] suggested the integration of HPC and Apache Big Data Stack (*ADBS*) to offer usability, functionality, and sustainability that is not available in the HPC ecosystem. They mention also that an implementation of HPC-ABDS is provided in the SPIDAL project [12].

Napoli et al., 2014 [2] have developed a GPU Architecture using parallel and distributed treatments to process and interpret tremendous amount of data in real-time of tens of millions of raw images. They use dynamic adjusting of number of hosts, because the performance of data processing that cannot be predicted and the throughput of data can overheat while time goes. The use of solutions such as MPI cannot adapt dynamically the number of processes once the execution has begun.

Song et al. 2018 [1] proposed a novel framework and an architecture based on the principle of Fog computing to train the Deep Learning locally with the aim of reduce data movement, speedup model update, and by consequently contribute to the energy saving. This approach addresses the problem of the transfer of all data on the cloud needed to train Deep Learning statically models. However, they cannot handle with high accuracy raw IoT data which are dynamic and unlabeled.

Lu et al. 2016 [5] studied the impact of choosing network technologies on the HPC Cloud. They proposed an architecture based on Hadoop multi-protocol aware to take advantage of Reliable Connection (*RC*), Unreliable Datagram (*UD*), hybrid protocols for InfiniBand (*IB*) and RDMA over Converged Ethernet (*RoCE*) that leads to Remote Direct Memory Access (*RDMA*) to provide high bandwidth, latency and the throughput for Hadoop RPC and HBase communication.

Salaria et al, 2017 [4] have compared performance between the latest generation of HPC-like cloud and a HPC for Graph500[1] which is a well-known Big Data benchmark and show that Cloud HPC can provide good compute performance with low variability.

Sood et al., 2017 [14] proposed an architecture in 4 layers from bottom to the top: (1) The IoT layer (*IL*), (2) The Fog Computing Layer (*FCL*), (3) The Data Analysis Layer (*DAL*) and (4) The Presentation Layer (*PL*). The IL organizes the social collaboration and energy saving of IoT devices. The FCL achieves on one hand the routing of data from IoT devices to cloud computing using multiple network devices and on the other hand the pre-processing or the prediction of data on nodes or gateways. The aim of this layer is the reduction of the latency of the system by sending calculated values on the cloud computing. The DAL contain any big data based on smart system. It is also responsible for collecting, storing, mining and data analyzing to obtain results. Finally, the PL is the user views of the system in which results can be attained after processing of all the information.

Bojan et al., 2015 [17] have described a solution for large scale time series visualization and showed that approach based on statistical data has better performances in terms of consuming time up to 10 times and data traffic up to 271 times.

Mocanu et al., 2015 [16] proposed a SOA architecture based on two controllers isolated and interconnected: one local and the second in the cloud. The local Farm Controller (*LFC*) provides access to more recent and historical data and execute preconfigured farm workflows. While the Cloud Farm Controller (*CFC*) provides data aggregation from farms, farm control and external sources.

Musat et al., 2018 [15] propose an integrated platform coupling a smart platform and a social network. This platform is based on two independent application (frontend and backend) using REST services or web sockets to communicate together. The backend architecture is developed in Java with Spring Framework, MySQL and Liquidbase. While, the fronted architecture is built with AngularJS framework, HTML 5 and CSS 3. This software architecture offers a wide range of services and gathering large amount of data produced by connected things to notify farmers in case of problem. The platform also offers forum, groups, store, tendencies and correlations services.

[1] http://www.graph500.org/

Ruy et al., 2015 [18] proposed a connected farm system composed of sensing and actuating connected IoT devices, IoT gateway and IoT service platform where all interfaces are REST APIs.

Ramirez et al., 2017 [20] has showed the benefits of combined and continuous Fog and cloud computing (*F2C*) architectures with over 50% reduction in terms of power consumption.

3. Proposed Architecture

Our architecture is composed of 4 layers as proposed by Sood et al. [14] from the bottom to the top: (1) The sensing layer (*SL*) is constituted by sensors and micro-controllers which acquire physical measurements of their environment. A primal treatment of data (Edge computing) is also processed on capable sensors to send only valuable data. (2) The Fog Computing Layer (*FCL*) is composed of nodes, gateways able to achieved more important treatments than the SL, and mobile GPU and FPGA used for the incremental deep learning training [1], (3) the Data Analysis Layer (*DAL*) aim to collect data from IoT sensor on one hand and from external sources such as local and distributed databases, and open data on the other hand. Finally, (4) the Presentation layer allows to users to view results of treatments (Fig. 1).

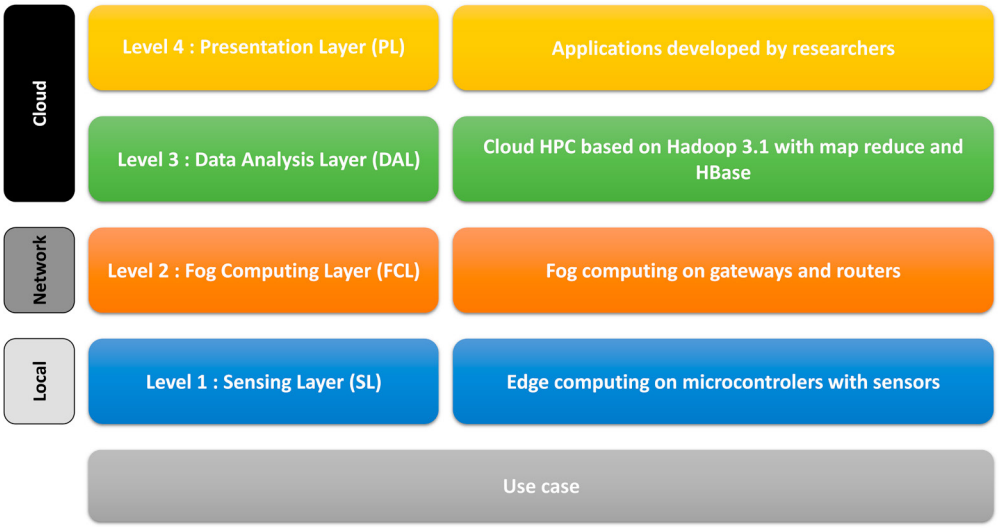

Fig. 1. Proposed Architecture

3.1. Data collection

The collect of data is operated at the sensing layer by means of sensors connected to a node. A node generally controls several sensors which make the physical measure of their environment. The Edge computing is used to preprocess data and to send only relevant data to limit the data sending that is energy consuming and thus improve the life of the nodes that control sensors. However, the use of edge computing cannot be done on the older nodes that are more limited in memory, processing capacity, and storage than the most recent nodes. The data are then sent to gateways and hubs that are capable of performing the heavier processing (Fog Computing) and therefore limiting data sending and processing in the cloud. This process limits the use of bandwidth and brings the management of privacy issues back to the data producers. The Edge and the Fog computing are respectively implemented in the first two layers.

3.2. Data management

Data coming from IoT sensors via the second layer (data stream) and other external sources such as local, distributed database, CSV and TSV files, open data, and messaging services, etc., are collected and treated before to be

stored in the distributed database (HBase). A specific framework has been developed to retrieve open data in XML, RDF and CSV file format and store them in HBase. The centralization of all these data is needed to enrich analytic and make better decisions.

3.3. Data quality

The data quality (*DQ*) is important for the data mining process. Indeed, a low DQ impact directly the validity of the results and their interpretation. The quality of a sensor data stream can be evaluated with five dimensions: accuracy, confidence, completeness, data volume, and timeliness. In addition, easiness of retrieving data, access, security, and interpretability are also describing the DQ for IoT. Numerous factors influence the DQ: (1) The huge number of devices and resources constrained increase the error occurrence, the ratio packet loss and tradeoffs between quality and cleanliness with battery life; (2) Sensor precision and loss of calibration, the lack of maintenance, the vandalism both from humans and animals or a defective node which send erroneous data that affect the DQ; (3) DQ can also be altered by privacy processing, security attacks or applying of certain operators of data stream processing [19].

Raw data are the base of the context-awareness which is composed of four phases: acquisition, modeling, reasoning and dissemination of the context and are achieved at application or middleware level. Major DQ enhancement are outlier detection, interpolation, data integration, data deduplication, and data cleaning.

Outlier detection is the enhancement of the consistence by elimination of discovered outliers or highlighting of rare events or patterns underlying in a dataset.

According to Karkouch et al [19], outlier are events with extremely small probabilities of occurrence and are classified as below: (1) **error** due to node failure; (2) **event** caused by a sudden or extreme change; (3) point anomaly that differs greatly from other values of the dataset; (4) **contextual anomaly** is a value considered abnormal in a determine context; (5) **collective anomaly** is a collection of data that differs greatly of the rest of the dataset.

Interpolation infers missing data due especially to sensor dysfunctions or loss of connections, on the base of available data using methods such as linear or polynomial interpolation. The choose of an interpolation method must be done accordingly to the accuracy of the interpolation.

Data integration is ensured on one hand by a suite of components and services specifying standardized and efficient interoperability of sensor data and on the other hand by the Resources Description Framework (*RDF*), Web Ontology Language (*WOL*) which provides mechanism to describe data. Moreover, the Linked Data is also an approach to ease data integration and retrieval.

Data deduplication is a compression mechanism to reduce the amount of data by removing of duplicate data and replacing with a pointer to the unique remaining copy.

Data cleaning begin by the determination of error types followed by the identification of potential errors and finally the correction of identified errors. The interested reader can find more details on the techniques of data cleaning in Karkouch et al. [19].

3.4. High Performance computing

The Stored data on Apache HBase are processed with Apache Hadoop 3.1, which support natively GPU and FPGA. We have modified the support of GPU for the use of Multi GPU by using Map Reduce to ensure the distribution between GPU where data are processed in parallel using tools like CUDA[2], OpenCL[3], etc. In this new release of Hadoop, Erasure Coding which provides significant improvement in data access speed on HDFS. The exploitation of GPU allows also to train Neural Networks in order to use Machine Learning and Deep Learning using tools like TensorFlow[4], Keras[5], OpenAi[6], etc. on stored data.

[2] https://developer.nvidia.com/cuda-downloads
[3] https://www.khronos.org/opencl
[4] https://www.tensorflow.org
[5] https://keras.io
[6] https://openai.com

3.5. Exploitation of data

Finally, applications and models developed by researchers exploit the results analysis achieved by the heterogeneous cloud HPC on base of data stored in the big data. A web interface using REST APIs allows on one hand to host and monetize models and applications developed by researchers on the platform. Users of the platform can use the visualization of statistical data or treat data with the application hosted on the market place of the platform which reaches the high performance of the cloud architecture to treat raw data.

4. Conclusion and future work

In this paper, we propose a versatile architecture in 4 layers for Smart Farming which is able to collect, store, and treat data coming from IoT nodes and integrate external data from other sources such as local and distributed data, messaging services, and open data, etc. Our architecture aims to improve the quality of data by means of outliers detection, data cleaning and interpolation of missing data. In addition, our architecture offers at the same time the possibilities to achieve using different kinds of Deep Learning supervised and unsupervised algorithms using a combined heterogeneous cloud architecture with the fog computing. The main novelty of our approach is to combine at same time heterogeneous high performance cloud computing, distributed databases and fog computing.

In future work, we will collect real data and test this architecture on real data in a future research project that will be start in January 2019. Other format of open data will be implemented in order to increase the amount of data which can be matching with IoT data.

Acknowledgements

The authors would especially like to thank Mr Adriano Guttadauria for his technical support and for setting up all the electronic systems and computing systems necessary for carrying out this research and Mrs. Meryem Elmoulat for the English editing of this paper.

References

[1] Song, Mingcong, Zhong Kan, Zhang, Jiaqi, Hu, Yang, Liu, Duo, Zhang, Weigong, Wang, Jing and Tao Li (2018) “In-situ AI: Towards Autonomous and Incremental Deep Learning for IoT Systems.”, in 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), 92–103. doi: 10.1109/HPCA.2018.00018

[2] Napoli, Christian, Pappalardo, Giuseppe, Tramontana, Emiliano and Gaetano Zappal‘a (2014) “A Cloud-Distributed GPU Architecture for Pattern Identification in Segmented Detectors Big-Data Surveys.” *The Computer Journal* **59(3)**, 339–352. doi: 10.1093/comjnl/bxu147

[3] Fox, Geoffrey C. and Jha Shantenu (2017) “Conceptualizing a Computing Platform for Science Beyond 2020: To Cloudify HPC, or HPCify Clouds?”, in 2017 IEEE 10th International Conference on Cloud Computing (CLOUD), 808–810. doi: 10.1109/CLOUD.2017.120

[4] Salaria, Shweta, Brown, Kevin, Jitsumoto, Hideyuki and Satoshi Matsuoka (2017) “Evaluating of HPC-Big Data Applications Using Cloud Platforms.”’, in 2017 17h IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 1053–1061. doi: 10.1109/CCGRID.2017.143

[5] Lu, Xiaoyi, Shankar, Dipti, Gugnani, Shashank, Subramoni, Hari and Panda Dhabaleswar K. (2016) “Impact of HPC Cloud Networking Technologies on Accelerating Hadoop RPC and HBase.”, in 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), 310–317. doi: 10.1109/CloudCom.2016.0057

[6] Debauche, Olivier, Mahmoudi, Saïd, Manneback, Pierre, Massinon, Matthieu, Tadrist, Nassima, Lebeau, Frédéric and Sidi Ahmed Mahmoudi. (2017) “Cloud architecture for digital phenotyping and automation.”, in 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), 1–9. doi:10.1109/CloudTech.2017.8284718

[7] Debauche, Olivier, Mahmoudi, Saïd, Andriamandroso, A.L.H., Manneback, Pierre, Bindelle, Jérôme and Frédéric Lebeau. (2017) “Web-based cattle behavior service for researchers based on the smartphone inertial central.”, in 14th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2017) / 12th International Conference on Future Networks and Communications (FNC 2017) / Affiliated Workshops. *Procedia Computer Science* **110**, 110–116. doi: 10.1016/j.procs.2017.06.127

[8] Debauche, Olivier, El Moulat, Meryem, Mahmoudi, Saïd, Manneback Pierre and Frédéric Lebeau. (2018) “Irrigation pivot-center connected at low cost for the reduction of crop water requirements”, in 2018 International Conference on Advanced Communication Technologies and Networking (CommNet), 1–9. doi: 10.1109/COMMNET.2018.8360259

[9] Debauche, Olivier, El Moulat, Meryem, Mahmoudi, Saïd, Boukraa, Slimane, Manneback, Pierre and Frédéric Lebaeau (2018) “Web Monitoring of Bee Health for Researchers and Beekeepers Based on the Internet of Things.”, in The 9th International Conference on Ambient Systems,

Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops *Procedia Computer Science* **130**, 991–998. doi: 10.1016/j.procs.2018.04.103

[10] Debauche, Olivier, Mahmoudi, Saïd, Andriamandroso, A.L.H., Manneback, Pierre, Bindelle, Jérôme and Frédéric Lebeau. (2018) "Cloud services integration for farm animals' behavior studies based on smartphones as activity sensors." *Journal of Ambient Intelligence and Humanized Computing*, 1–12. doi: 10.1007/s12652-018-0845-9

[11] El Moulat, Meryem, Debauche, Olivier, Mahmoudi, Saïd, Brahim, Lashen Aït, Manneback, Pierre and Frédéric Lebeau. (2018) "Monitoring System Using Internet of Things For Potential Landslides.", in 15th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2018) *Procedia Computer Science* **134**, 26–34. doi: 10.1016/j.procs.2018.07.140

[12] Fox, Geoffrey C., Qiu, Judy, Kamburugamuve, Supun, Shantenu, Jha and Andre Luckow (2015) "HPC-ABDS High Performance Computing Enhanced Apache Big Data Stack.", in: 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 1057–1066. doi: 10.1109/CCGrid.2015.122

[13] Martinelli, Nicole (2018) "Trends to watch in high performance computing.", Superuser. Online: http://superuser.openstack.org/articles/trends-high-performance-computing/ (10/07/2018)

[14] Sood, Sandeep K., Sandhu, Rajinder, Singla, Karan and Victor Chang (2017) "IoT, big data and HPC based smart flood management framework.", *Sustainable Computing Informatics and Systems*. doi: 10.1016/j.suscom.2017.12.001

[15] Musat, George-Alexandru, Colezea Mădălin, Pop, Florin, Negru, Catalin, Mocanu, Mariana, Exposito, Christian, Castiglione, Aniello. (2018) "Advanced services for efficient management of smart farms." *Journal of Parallel Distributed Computing* **116**, 3–17. doi: 10.1016/j.jpdc.2017.10.017

[16] Mocanu, Mariana, Cristea, Valentin, Negru, Catalin, Pop, Florin, Ciobanu, Vlad, Dobre, Ciprian (2015) "Cloud-Based Architecture for Farm Management", in 2015 20th International Conference on Control Systems and Science, 814–819 doi: 10.1109/CSCS.2015.55

[17] Bojan Valentina-Camelia, Raducu, Ionut-Gabriel, Pop, Florin, Mocanu, Mariana, Cristea, Valentin (2015) "Cloud-based Service for Time Series Analysis and Visualisation in Farm Management System.", in 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), 425–432. doi: 10.1109/ICCP.2015.7312697

[18] Ryu, M., Yun, J., Miao, T., Ahn, I., Choi, S., Kim, J. (2015) "Design and implementation of a connected farm for smart farming system.", in: 2015 IEEE Sensors, 1–4. doi: 10.1109/ICSENS.2015.7370624

[19] Karkouch, Aimad, Mousannif, Hajar, Al Moatassime Hassan, Noel Thomas (2016) "Data quality in internet of things: A state-of-the-art survey." *Journal of Network and Computer Applications* **73**, 57–81. doi: 10.1016/j.jnca.2016.08.002

[20] Ramirez, W., Masip-Bruin, X., Martin-Tordera, E., Souza, V.B.C., Jukan, A., Ren, G.-J., Gonzalez de Dios, O. (2017) "Evaluating the benefits of combined and continuous Fog-to-Cloud architectures." *Computer Communications* **113**, 43–52. doi: 10.1016/j.comcom.2017.09.011