Information Logistics and Fog Computing: The DITAS* Approach

Pierluigi Plebani[1], David Garcia-Perez[2], Maya Anderson[4], David Bermbach[3], Cinzia Cappiello[1], Ronen I. Kat[4], Frank Pallas[3], Barbara Pernici[1], Stefan Tai[3], and Monica Vitali[1]

[1] Politecnico di Milano – Dipartimento di Elettronica, Informazione e Bioingegneria
Piazza Leonardo da Vinci, 32 - 20133 Milan, Italy
[firstname].[lastname]@polimi.it
[2] Atos Spain SA
Pere IV, 08018 Barcelona, Spain
david.garciaperez@atos.net
[3] TU Berlin – Information Systems Engineering Research Group
Einsteinufer 17 - 10587 Berlin, Germany
{db,fp,st}@ise.tu-berlin.de
[4] IBM Research – Haifa
Haifa University Campus, Mount Carmel, Haifa, 3498825, Israel
{ronenkat,mayaa}@il.ibm.com

Abstract. Data-intensive applications are usually developed based on Cloud resources whose service delivery model helps towards building reliable and scalable solutions. However, especially in the context of Internet of Things-based applications, Cloud Computing comes with some limitations as data, generated at the edge of the network, are processed at the core of the network producing security, privacy, and latency issues. On the other side, Fog Computing is emerging as an extension of Cloud Computing, where resources located at the edge of the network are used in combination with cloud services.

The goal of this paper is to present the approach adopted in the recently started DITAS project: the design of a Cloud platform is proposed to optimize the development of data-intensive applications providing information logistics tools that are able to deliver information and computation resources at the right time, right place, with the right quality. Applications that will be developed with DITAS tools live in a Fog Computing environment, where data move from the cloud to the edge and vice versa to provide secure, reliable, and scalable solutions with excellent performance.

Keywords: Fog Computing, Edge Computing, Data movement, Cloud Computing

*http://www.ditas-project.eu/

X. Franch, J. Ralyté, R. Matulevičius, C. Salinesi, and R. Wieringa (Eds.):
CAiSE 2017 Forum and Doctoral Consortium Papers, pp. 129-136, 2017.

1 Introduction

Fog Computing [9], often also referred to as Edge Computing [12], is an emerging paradigm aiming to extend Cloud Computing capabilities to fully exploit the potential of the edge of the network where traditional devices as well as new generations of smart devices, wearables, and mobile devices – the so-called Internet of Things (IoT) – are considered. Especially for data-intensive applications, since IoT is a source for enormous amounts of data that can be exploited to provide support in a multitude of different domains (e.g., predictive machinery maintenance, patient monitoring), this new paradigm has opened new frontiers.

Nowadays, the typical approach for data processing relies on cloud-based applications: data are collected and pre-processed on the edge, then they are moved to the cloud, where a more scalable and reliable processing environment is provided. Finally, the output of the analysis is sent to the end user, whose devices are again on the edge. While the cloud offers virtually unlimited computational resources making data processing extremely efficient, the resulting data movement could have an impact on the application performance due to possibly significant latencies of the transmission. To improve this type of applications, proper information logistics become fundamental for delivering information at the right time, the right place, and with the right quality [10]. However, developing applications that are able to deal with these issues can be really challenging for several reasons: heterogeneity of devices at the edge, privacy and security issues when moving data from the edge to the cloud, or limited bandwidth.

The goal of this paper is to introduce the recently started DITAS project which aims to improve, through a cloud platform, the development of data-intensive application by enabling information logistics in Fog environments where both resources belonging to the cloud and the edge are combined. The resulting data movement is enabled by Virtual Data Containers which provide an abstraction layer hiding the underlying complexity of an infrastructure made of heterogeneous devices. Applications developed using the DITAS toolkit will be able to exploit the advantages of both cloud-based solutions about reliability and scalability, and edge-based solutions with respect to latency and privacy.

To properly discuss the relevant aspects of information logistics, Fog computing as well as challenges that need to be faced, the rest of the paper is organized as follow: Section 2 proposes the DITAS vision on Cloud, Fog, and Edge computing. Details on the importance of information logistics in general, as well as in Fog environments in particular, are discussed in Section 3. Section 4 gives an overview of the DITAS approach with emphasis on the architectural description of the proposed cloud platform.

2 On the Fog, Edge, and Cloud Computing

Cloud computing [7] has been widely adopted as a model for developing and providing scalable and reliable applications in a cost-efficient way with minimal infrastructure management efforts for developers. In such an environment, computational resources as well as storage capacity can be considered unlimited,

which has boosted the proliferation of applications able to manage huge amount of data. Nevertheless, relying only on Cloud computing or even on federated clouds [6] could have some drawbacks especially when the data to be processed are generated by devices located at the edge of the network. If we consider that predictions for IoT estimate about 32 billion connected devices in 2020 that will be producing about 440 Petabytes per year (accounting for 10% of the world data) [13], the impacts of data movement between edge-located resources – where data are produced – and the cloud-located resources – where data are processed and stored – can be significant. In the telecommunication sector, when

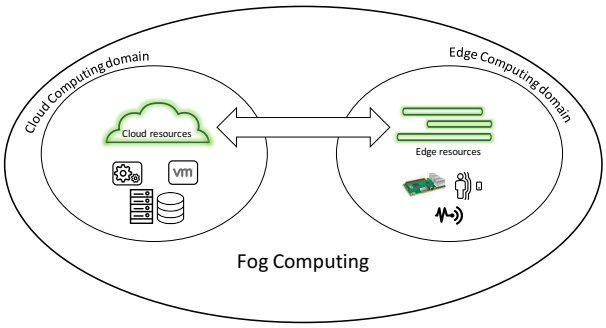

Fig. 1. Fog Computing environment

mobile devices moved from being simple consumers of contents to being producers of contents, the term Fog Computing has been coined to identify a platform able *"to provide compute, storage, and networking services between Cloud data centers and devices at the edge of the network"* [3]. The term Edge Computing has been instead proposed by information systems engineering researchers when IoT was recognized as a great opportunity for providing new services. It identifies the *"technologies allowing the computation to be performed at the edge of the network, on downstream data on behalf of cloud services and upstream data on behalf of IoT services"* [11].

Based on these definitions, Cloud Computing is mainly related to the core of the network whereas Edge Computing is focused on providing to the owner of resources the local 'in-situ' means for collecting and preprocessing data before sending it to cloud resources (for further utilization), thus addressing typical constraints of sensor-to-cloud scenarios like limited bandwidth and strict latency requirements. For this reason, also in the light of the definition proposed by the OpenFog Consortium [9], we here after consider Fog Computing as the sum of Cloud and Edge Computing. As shown in Figure 1, cloud resources include physical and virtual machines which are capable of processing and storing data. On the other side, smart devices, wearables, or smartphones belong to the set of edge-located sources. While Cloud Computing is devoted to efficiently managing capabilities and data in the cloud, Edge Computing is focused on providing the

means for collecting data from the environment (which will then be processed by cloud resources) to the owner of the available resources. On this basis, Cloud and Edge Computing are usually seen as two distinct and independent environments that, based on the specific needs, are connected to each other to move data usually from the edge to the cloud.

Exploiting the Fog Computing paradigm, in DITAS these two environments seamlessly interoperate to provide a platform where both computation and data can be exchanged in both downstream and upstream direction. For instance, when data cannot be moved from the edge to the cloud, e.g., due to privacy issues, then the computation is moved to the edge. Similarly, when there are insufficient resources at the edge data are moved to the cloud for further processing. With DITAS, we want to provide tools, specifically designed for developers of data-intensive applications, which are able to autonomously decide where to move data and computation resources based on information about the type of the data, the characteristics of applications as well as the available resources at both cloud and edge locations and applicable constraints such as EU GDPR privacay regulation [5].

3 Information Logistics

Since the 1990s, when interconnection of heterogeneous information systems managed by different owners became easier and when the Web started managing significant amounts of information, the problem of delivering such information has become more and more relevant: the more the data are distributed, the more difficult is it to find the information needed. Thus, tools are required to guide the users in this task. In this scenario, Information Logistics[6] has emerged as a research field for optimizing the information provision and information flow especially in networked organizations [10].

As discussed in [8], Information Logistics can be studied from different perspectives: e.g., how to exploit the data collected and managed inside an organization for changing the strategy of such organizations, how to deliver the right information to decision makers in a process, or how to support supply chain management. In our case, according to the classification proposed in [8], we are interested in user-oriented Information Logistics: i.e., the delivery of information at the right time, the right place, and with the right quality and format to the user [4]. As a consequence, user requirements can be defined in terms of functional aspects, i.e., content, and non-functional ones, i.e., time, location, representation, and quality.

Especially in the recent years, where data deluge has been ever increasing, providing solutions that are able to satisfy these types of requests becomes more

[5] http://http://www.eugdpr.org/

[6] Actually, according to [5] the Information Logistics term appeared the very first time in *Wormley P. W., Information logistics: Local distribution (delivery) of information, Journalism Quarterly, Vol. 55 Issue 3, 1/9p, 1978*, but a real interest on this field started only more recently.

and more challenging. Different sources available on the Web could provide data of interest to the user and selecting the best source depends on non-functional requirements. Data quality dimensions, such as timeliness or accuracy [1], can be used to measure whether data are useful or not. Location of data sources and the performance of the system offering these data can influence data quality: e.g., data stored in the cloud will take more time to be obtained than data on the premises of the user. Privacy aspects are also fundamental as some data must not leave the boundaries of the organization which holds these data, be used for specific purposes as defined by GDPR regulations, or may only do so in preprocessed (pseudonymized, aggregated, etc.) form. Finally, dealing with edge-located resources significantly increases the complexity of the resulting system for many reasons. These resources are often very heterogeneous in terms of software platforms, storage capabilities, computational resources, and data formats. The network connection may vary in bandwidth and, due to their nomadic nature, there is a high rate of churn among edge-located resources.

When developing data-intensive applications in a Fog Computing environment, information logistics holds a central role and DITAS wants to simplify the work of developers providing tools that are able to enact the proper data and computation movements so as to satisfy user requirements. In DITAS, the processing tasks composing the data-intensive application specify not only the content of required data, but also the data utility which subsumes all non-functional requirements that may vary with respect to the status and the context of the application. In contrast to the typical approaches where data move to the processing modules, in DITAS, computational tasks may also move to where the data are stored whenever it is neither feasible, e.g., due to privacy issues, nor acceptable, e.g., due to high latency, to move the data. Thus, DITAS is studying data and computation movement strategies to decide where, when, and how to persist data – on the cloud or on the edge of the network – and where, when, and how to compute part of the tasks composing the application to create a combination of edge and cloud that offers a good balance between reliability, security, sustainability, and cost. To ascertain whether such DITAS-based applications actually meet their quality goals monitoring, but also experiment-driven cloud service benchmarking [2] will be used.

4 DITAS approach

The objectives of DITAS discussed above will be offered through a *cloud platform* where developers can design their data-intensive applications where tasks are linked to *Virtual Data Containers* that satisfy the data and application requirements by enacting the proper information logistics.

4.1 Virtual Data Container

The concept of a Virtual Data Container (VDC) represents one of the key elements in the DITAS proposal. Generally speaking a VDC embeds the logic to

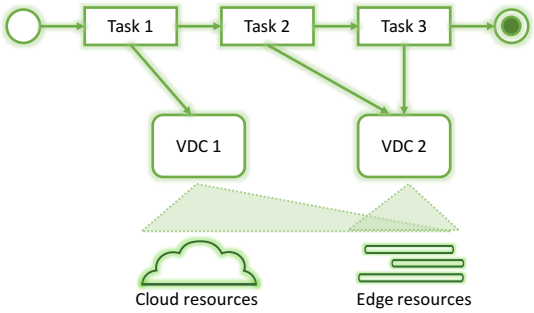

Fig. 2. DITAS Virtual Data Container

enable a proper information logistics depending on both the application to be executed and the available data sources. On the one hand, a VDC is linked to one or more tasks composing the data-intensive applications. Along with these links the developers specify the needs in terms of data, including both functional (i.e., the content) and non-functional aspects (i.e., data quality). On the other hand, a VDC is connected to a set of data sources offering a given information.

Thus, the goal of a VDC is to provide a single virtual layer, which can be built upon the principles of Service Oriented Computing, which abstracts the connected data sources. This virtual layer hides from the developer the intricacies of the underlying complex infrastructure composed by smart devices, sensors, as well as traditional computing nodes located in the cloud. The VDC also embeds the capabilities not only for moving the data from the data sources to the tasks operating on them, but also between the data sources to make the data provisioning effective taking into account data access policies and regulations. Finally, the definition of data movement techniques implies the definition of data transmission (e.g., data stream, bulk) and data transformation (e.g., encryption, compression) mechanisms, provided by a VDC, to define which will be the destination of the data, which is the data format to be adopted during the transmission, and how data will be stored in the destination.

Once the application is running, the application interacts only with the VDC as designed, which, in turn, enact all the data movement strategies needed to satisfy the posed requirements.

4.2 DITAS platform

The DITAS platform is a cloud-based solution which offers to the data-intensive application developers a set of tools for improving the design and the execution of the applications exploiting the functionalities provided by VDCs.

As shown in Figure 3 the functionalities offered by the DITAS cloud platform are composed of two main blocks: (i) an SDK in charge of supporting the developers in design and deployment of the applications, and (ii) a distributed execution environment responsible for running and controlling the behavior of the

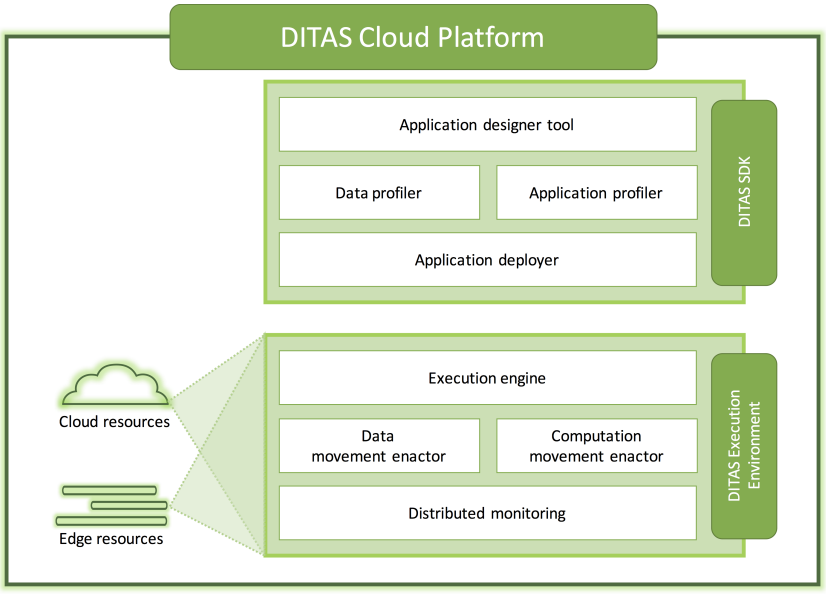

Fig. 3. DITAS Cloud Platform Architecture

application. About the SDK, extension of popular tools (e.g. Node-RED), will be proposed to define applications in which data processing is a central element. The key element of this tool is to allow the developer to design the applications by specifying the VDC and constraints/preferences about the resources to be exploited. For this reason, a communication with the cloud platform infrastructure is required to have a complete picture about the available resources both on the cloud and the edge. Based on the developer's instructions, and leveraging on the degree of freedom given by the VDC while satisfying all the constraints, the application is deployed among the selected resources.

The deployment of a DITAS-enabled data-intensive application implies that the resources selected to be involved in the execution embed a set of modules supporting the data and computation movement. For this reason, the execution environment is based on a distributed architecture where the execution engine is in charge of executing the tasks assigned to the resources on which the engine is running and of maintaining coordination with the other resources involved in the same application. Moreover, a monitoring system is able to check the status of the execution, track data movement, and collect all data necessary for understanding the behaviour of the application. In case of deviation with respect to the expected behavior, data and/or computation movements can be enacted to move the application to an acceptable state. Finally, data collected by all involved monitoring systems can be further analysed to provide advice to the developers for improving the design of their application.

5 Conclusions

This paper has presented the vision of DITAS, a recently started EU project, which aims to simplify the development of data-intensive applications where data and computation can be moved among resources belonging to both the core and the edge of the network, i.e., both federated clouds and edge environments.

As the project is in its infancy, future work is mainly devoted to implementing all the concepts expressed in this paper. Two case studies, one about an Industry 4.0 scenario, and another about an e-Health scenario will be used for testing and validating the proposed approach.

Acknowledgments

DITAS project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement RIA 731945.

References

1. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41(3), 16:1–16:52 (2009)
2. Bermbach, D., Wittern, E., Tai, S.: Cloud Service Benchmarking: Measuring Quality of Cloud Services from a Client Perspective. Springer Int.l Publishing (2017)
3. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the internet of things. In: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing. pp. 13–16. MCC '12 (2012)
4. D'Andria, F., Field, D., Kopaneli, A., Kousiouris, G., Garcia-Perez, D., Pernici, B., Plebani, P.: Data Movement in the Internet of Things Domain. In Service Oriented and Cloud Computing: 4th European Conf., ESOCC 2015, Taormina, Italy, September 15-17, 2015, Proc. Springer Int.l Publishing (2015)
5. Haftor, D.M., Kajtazi, M., Mirijamdotter, A.: A Review of Information Logistics Research Publications. Springer Berlin Heidelberg (2011)
6. Kurze, T., Klems, M., Bermbach, D., Lenk, A., Tai, S., Kunze, M.: Cloud federation. In: Proceedings of the International Conference on Clouds, Grids, and Virtualization (CLOUD COMPUTING). vol. 2011, pp. 32–38 (2011)
7. Liu, F., et al.: NIST Cloud Computing Reference Architecture: Recommendations of the National Institute of Standards and Technology (Special Publication 500-292). CreateSpace Independent Publishing Platform, USA (2012)
8. Michelberger, B., Andris, R.J., Girit, H., Mutschler, B.: A Literature Survey on Information Logistics. In Business Information Systems: 16th Int.l Conf., BIS 2013, Poznań, Poland, June 19-21, 2013. Proc. Springer Berlin Heidelberg (2013)
9. OpenFog Consortium Architecture Working Group: OpenFog Architecture Overview (February 2016), http://www.openfogconsortium.org/ra
10. Sandkuhl, K.: Information Logistics in Networked Organizations: Selected Concepts and Applications. Springer Berlin Heidelberg (2008)
11. Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L.: Edge computing: Vision and challenges. IEEE Internet of Things Journal 3(5), 637–646 (Oct 2016)
12. Shi, W., Dustdar, S.: The Promise of Edge Computing. Computer 49(5), 78–81 (May 2016)
13. Turner, V., Reinsel, D., Gatz, J.F., Minton, S.: The Digital Universe of Opportunities. IDC White Paper (April 2014)