CO-ARCH: Methodology for COllaborative ARCHitectures for Cross-organizational Data Analysis

B. D. Van Der Waaij, E. Lazovik, T. Albers, and M. R. Vonder

***Abstract*—In modern data-driven analysis it becomes quite typical to process not only the datasets you own, but to collaborate with other organizations to receive data and analysis results from them as well. It is performed to achieve much more accurate analysis results, make better predictions, and be able to provide better decision-support mechanisms. However, to analyze data in a cross-organizational environment is not the same as to analyze your own data: there are many limitations and conditions from the collaborators to allow access to their data and/or analysis models. This paper presents a methodology called CO-ARCH dealing with the process of choosing the suitable data-driven architectures for collaboration on data analysis between different organizations having their own conditions and limitations.**

***Index Terms*—Big data, collaborative AI, data science, multi-organizational data analysis, methodology.**

I. Introduction

Data is becoming very important in our society, e.g. using sensors for measuring the environment such as bridges, arable land, traffic. Also individual items such as cars, machines, animals, etc. are getting observed and measured more frequently. The power of data really becomes unleashed when it is not only just visualized, but when it is fed to analytics algorithms/models which bring additional insights, based on the data, for many different applications.

There are many technologies nowadays which provide the means for the analytics on huge amounts of historical data as well as streaming data. Therefore, the expertise of Data Science within one organization becomes more and more common. However, in many cases the decisions require additional information from different data sources, not owned by that organization.

When multiple organizations are collaborating in the development of a new data driven analysis, things can become difficult. The hurdles can be for instance: Are all organizations willing to provide open access to their data or are there specific restrictions? Can each dataset be transported towards the analysis or are some datasets just too big, and/or the communication possibilities too limited? Is the code of analysis models open to check what it does with data?

In this paper we provide an overview of possible collaboration IT architectures for the development and operations of data-driven analytical models, when the analytical party is not necessarily the same as the data party(ies). To come to the specific choices of possible IT architectures, a methodology is presented that helps to investigate any specific case of collaboration. The steps of the methodology described in this paper should result in a list of suitable technical IT-architectures for any specific case of collaboration between different organizations with the goal to analyze data.

Manuscript received September 29, 2019; revised February 12, 2020. This work was supported and financed by TKI Jongvee project [12] in The Netherlands.

B. D. van der Waaij, E. Lazovik, T. Albers, and M. R. Vonder are with TNO (The Netherlands Organization for Applied Scientific Research), Groningen, 9727DW, Netherlands (e-mail: bram.vanderwaaij@tno.nl, elena.lazovik@tno.nl, toon.albers@tno.nl, matthijs.vonder@tno.nl).

II. State of the Art

The current method for large-scale data processing requires that all data is transferred to one central location and processed there. It concerns all MapReduce platforms (i.e., Apache Spark [1], Hadoop [2], etc.) for data-driven analysis, and all cloud-based infrastructures (i.e., Cloudera [3], TensorFlow [4], etc.). Such a data analysis platform can have a large-scale (distributed) nature where a lot of data can be processed in parallel, but it always starts from the assumption of availability of all data sources in one centralized location on the premises of that data-driven platform at the start of combined analysis. And then Big Data processing is typically done on large clusters of shared-nothing commodity machines [5]. That means that there is an assumption to always have a single point of meeting for data and analysis, and that one organization controls both data and analysis.

However, sometimes it is not possible because some data owners cannot copy their data to another place because of privacy regulations, commercial secrets, or just because they want to be in control of their data. There are some scientists, as Jesus Rodriguez, who are looking at Decentralized AI nowadays, since they are sure that "Decentralization is very likely to become one of the pillars that influences the next decade of artificial intelligence (AI). Continuing relying on centralized models is likely to increase the gap between large companies and countries with the resources to develop AI solutions and the rest of the market" [6]. There are some initiatives and projects which tackle these issues from the perspective of multi-organizational data analysis.

The health sector has many such challenges because of the issues with privacy data from patients. Therefore, it is the most advanced sector regarding multi-hospital data analysis.

There is a European initiative Personal Health Train (PHT) [7] where the aim is to create a solution for different stakeholders within the health sector. Right now there is work in progress to elaborate some standards on description of health data, and the architecture which can allow the analysis to be run at the premises of hospitals where the required data resides.

DOI: 10.7763/IJMO.2020.V10.740

For example, in the H2020 project RECAP-preterm [8] dedicated to the research on preterm-born children the main central problem is: how to analyze data from different hospitals which are in different EU-countries. They use DataShield platform [9] to be able to run the analysis from one hospital at another hospital. The idea of this platform is to support researchers from one hospital to run the same analysis with the same parameters at other hospitals. It is a good solution when all participating parties have a common agreement to share data to all participants the same kind of data with the same parameters, which is the case in hospitals. However, it works only if the data is shared to every partner, and it still lacks the solution when different organizations have the distinctive parameters on the same subject, and their data could have gaps for specific time periods. So, it still has some lacunas to address.

H2020 project Musketeer [10] addresses the challenges of privacy preservation in machine learning algorithms. One of their goals is to provide a standardized architecture for machine learning algorithms on basis of secure multi-party computations to be able to preserve privacy data. It is cross-domain project which tackles the use-cases not only from health sector, but also from industry.

One more very extensive European initiative is IDS (International Data Spaces) [11] where the solution is being built for industry companies to run their services on-demand. There the central role is given to the certification party which is responsible for accepting specific services from companies and certifying them to run specific operations on data from other companies.

From this point on, it is evident that the new forms of cross-organizational data analysis are coming, every with their own architecture and features. However, there is no solid foundation to be able to classify them and choose the suitable technical solutions for every specific use-case. We address this issue in this paper.

III. Data Across Organizations

To be able to analyze big amounts of data it is important to get a clear insight in the required datasets. To obtain such insight, it is useful to look at the data along three different axes, see Fig. 1.

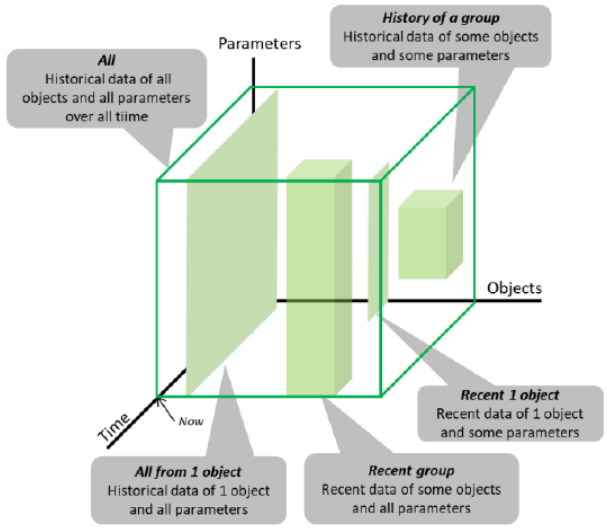

Fig. 1. The three axis of datasets.

The axis are:

- The object-axis: The physical object which has been measured, therefore, what the measurements are about: a dike, a cow, a human, etc.
- The parameter-axis: That what exactly has been measured on / about an object: temperature, color, weight, etc.
- The time-axis: The moment on which the parameter value on / about an object has been measured: timestamp.

There are different variants on distribution of data on the same physical objects:

- *All data at one organization*. The simplest situation is that all data is located at one organization. There is no distribution of the relevant data across other organizations.
- *All for themselves*. Typically organizations collect only data on their own objects and not upon the objects of other organizations. Therefore, there are no correlations on the same objects between different organizations. E.g. hospitals with their own patients.
- *Passing responsibility*. Sometimes the responsibility of collection of data is being passed from one organization to another. For instance, when the governmental obligation to collect certain data is being passed, or when a merger of two companies happens.
- *Distribution along time*. Another cause of complexity is when the data becomes also divided along the time axis. For example: in the past all data was collected by one organization (e.g. the government). At a certain point in time this is being passed to three individual organizations (commercial companies). Company 2 and 3 happen to measure partially the same parameters and therefore are now competitors.
- *Distribution along parameters*. A challenging cause of complexity is when the data about the same object becomes divided along the parameter axis. For example: the sensor data about a cow is collected and thereby divided over several organizations. Data
- about the milk (yield, fat%, etc.) is collected by the milk machine manufacturer, data about the weight of the cow by the cow-scale manufacturer, etc.

The big question is whether the way the data is distributed among the organizations is actually usable for the analysis. A problem arises when the analysis requires certain combinations / groups of data, while the different elements from the same group are stored at different locations.

IV. Collaboration Archetypes for Data Analysis

In principle there are three different ways in bringing the data and the analysis together in order to be processed. In order to be able to present standardized methodology at first we provide our classification of the collaborations which is based on where the analysis is being performed. We distinguish three principle ways by choosing to perform an analysis at:

- The **analytical party**.
- Consequences:
  - The data must be transported to the analysis party.

- the **data** party**(ies)**.
- Consequences:
  - The analysis must be transported to the data parties.
  - The results must be combined into a single end result and made available to the analysis party.
- A **trusted data** analysis **party**.
- Consequences:
  - Both the data and the analysis must be transported to the trusted data analysis party.
  - The outcome must be transported back to the analytical party.

Based on these three principle ways to analyze data at the premises of specific organization(s) we define three main collaboration archetypes for data analysis in cross-organizational context:

- **D2A**
  - Bringing the data to the analysis
- **A2D**
  - Bringing the analysis to the data
- **D&**A2L
  - Bringing the data and the analysis to a third trusted location, the lake, and perform over there the analysis.

V. General Co-arch Methodology Description

As stated in the introduction the main goal of this paper is: provide a methodology to reach suitable possible collaboration IT-architectures for data-driven analysis in cross-organizational context, when the analytical party is not necessarily the same as the data party(ies).

This section introduces a generic methodology to come to such a list of possible IT architectures in a multiparty collaboration environment.

A. Hourglass Approach

The approach to our methodology can be modelled as a hourglass (see Fig. 2 below): starting from many business collaboration options drill down to a couple of acceptable collaboration compositions. Each of these compositions will result in one or more blueprint IT architectures, finally converting into a list of suitable collaboration IT architectures.

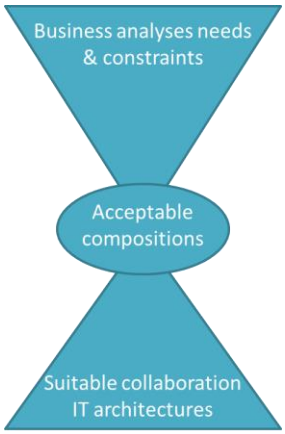

Fig. 2. Hourglass approach

At the top of the hourglass are the business needs and constraints and restrictions of each party, participating in the collaboration. In a number of steps they are converted into a limited set of workable collaboration data analysis compositions, which are called the Acceptable compositions. Along the way the continuous checking of all kinds of limitations, such as trust, disposition and volume is taking place. An impact analysis approach is used to validate each collaboration composition, and it should result in a set of acceptable compositions. For each of the acceptable compositions one or more suitable collaboration IT architectures can be chosen, using a predefined set of blueprint IT architectures.

Fig. 3 demonstrates the generic concept of going through the data-driven analysis hourglass. The methodological approach which is generally described hereinbefore, consists of going *through the hourglass* from *top* to *bottom*. At every hourglass level the user starts with checking issues relevant for this layer and defining the possible strategies to solve them. By choosing the specific strategy at upper level would influence the possible solutions underneath by restricting the specific choices.

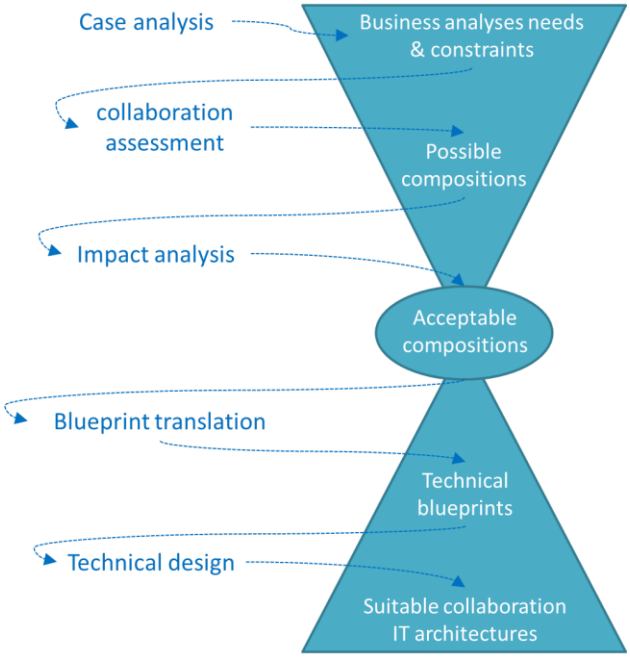

Fig. 3. "Going through hourglass" methodology.

The hourglass can be separated into two pyramids with different layers as it is demonstrated in Fig. 3. The upper pyramid contains the convergent layers that allow the organizations start with their wide and vague business goals, formulate them stricter and stricter, till making the choices between the Acceptable Compositions (i.e. the collaboration archetypes D2A, A2D, D&A2L). That decision is a convergence point for the high-level decisions based on business goals. Starting from that convergence point (the Acceptable compositions), the suitable technical choices should be made. That is the aim of the pyramid underneath.

In the next sections we describe our methodology on more details for upper pyramid to reach Acceptable compositions, and underneath pyramid to reach the overview of suitable collaboration IT architectures.

B. Convergent Layers: Steps to Reach Acceptable Compositions

The "case analysis"-step at the start of the whole process helps to identify the business needs of all participant data and analysis parties, and the limitations and restrictions per party. This step also formulates the end-goal for the collaboration between all involved parties. Starting from business needs the limitations, restrictions and wishes of every participating organization will be addressed to learn the current situation for specific use-case. It will help afterwards to achieve an agreement between all involved parties.

Studying of all possible collaboration opportunities for the specific use-case and understanding their limitations helps to extract the explicit list of the *possible* organizational *compositions* between data parties and analysis parties to achieve the desired common goal of analysis.

Hereinafter we provide a formalized description of every convergent layer from the pyramid above to reach the definitive decision on Acceptable compositions, with associated processes that should take place to resolve the problem statement of that layer.

- **Business analysis needs and constraints**

This layer provides an explicit description and formalization of the goal to achieve, the parties involved, and formulating the restrictions from every involved party.

Associated processes:

- Deciding on resulting information for achieving the business goals;
- Creating an overview of:
  - Involved parties;
  - Datasets and which party is the owner;
  - Analysis needs of each party;
  - Limitations and restrictions (on data and/or analysis) from every involved party.

The resulting information description should contain at least:

- the sharp definition of the goal and end result coming from data-driven analysis,
- goal limitations coming from business rules,
- requirements on accuracy of end result, security, and time deadline(s) for the analysis process,
- providing an overview of all suitable data attributes that are needed for the analysis.

- **Possible compositions**

This layer provides an overview of suitable compositions for collaboration between different organizations.

Associated Processes (in Mentioned Order):

- Trust:
  investigate which collaboration forms are acceptable from the point of view of Trust between the participant parties.
- Disposition:
  investigate mismatch between data arrangement over organizations versus data arrangement needs of the analysis.
- Impact:
  Verify if the analyses can deal with the collaboration architectures under investigation.

The central aspect of concluding the agreements is **Trust**. Some organizations would not mind on allowing others to use their data and/or make a copy of it. Other organizations are very protective on their datasets and want to limit the access to their data only for very specific purposes and even only to the trusted third parties. Some organizations allow you to look at the source code of their analysis, others do not allow you to know all the details. Therefore, there are quite many different choices on collaborating on data analysis compositions.

The involved organizations should evaluate the current situation from Trust perspective, and formulate the possible options for collaboration. The formulation process may be accessed in two different ways:

- The organizations reach the agreement on Trust on "one-to-one" level. That means that one specific organization would look to every other collaborating organization separately, and define the possible options per organization.
- The organizations together look on all possible variations of collaborations (Data travels to Analysis, Analysis travels to Data, Data and Analysis travel to common Data Lake), decide which level of Trust is needed from their perspective, and check whether resulting level of Trust is possible to reach on different variations of collaboration. On basis of the check results, the decision on common possible collaboration range is taken.

After tackling Trust issue, there is another point of attention to tackle: the data **Disposition** in respect to analysis. The data sources from different organizations can be structured in completely different manner, even for the same sort of data. Since every data owning organization has own business and analysis goals for that data, the data source is tuned to specific data requests to meet the needs of the analysis within data owner organization. However, when there is a need to combine the data sources from different organizations, one can expect completely different data schemas, its internal structuring and different levels of API to request the data from very simple web form till the automated complex requests.

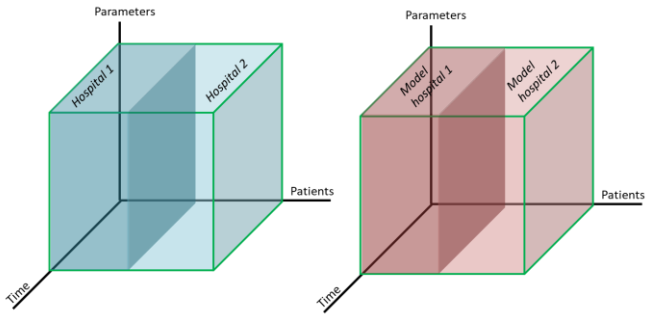

Fig. 4. Matching disposition in *hospital case* (left: data arrangement over organizations, right: analysis data needs).

From the other side, the analysis models have also their own expectations on input data arrangement and structuring. And the situations where the data structuring from different data sources do not meet the expectations of the analysis model on input are not rare. For instance, when from the

Trust perspective the analysis has to travel to the data, most distributed analysis algorithms require that the data is distributed over the data parties along the object-axis and not along the parameter-axis. For a healthcare case this is often true (see Fig. 4): each hospital can train a model using all the parameters of their patients. Models of different hospitals can then be combined into a single model.

But for a dairy case this is far more difficult (see Fig. 5) because the data of a single cow is distributed over several companies. The milk machine manufacturer has all the milk parameters of all the cows of all his clients, the cow-scale manufacturer has all the weight parameters of all the cows of their clients.

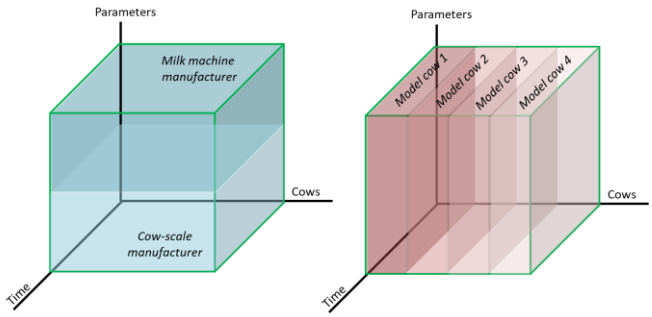

Fig. 5. Mismatching disposition in *dairy case* (left: data arrangement over organizations, right: analysis data needs).

Data distribution along parameter-axis is very difficult for algorithms to deal with.

If such mismatch exists, the analysis is not possible to run against the available data sources until the expectations of analysis are fulfilled. Therefore, immediately after Trust issues and choice of Possible collaborations the Impact of Disposition check should be performed.

If the both Trust and Disposition examinations are successfully passed through, it is time to perform a validation on **Impact** to make sure that the remaining options coming from Trust collaboration forms and Disposition solutions do still provide a valid basis to fulfil the analysis and reach the requested business goal.

Starting from the analysis needs on data and requirements on timeliness, it should be continued with the evaluation of volume parameters, and resume with the mapping of analysis needs to the results from volume evaluation.

The analysis needs often include specific requirements on the quantity of data as an input and on data quality for the reason of required accuracy, or/and timeliness-related requirements. Some examples of data quantity / volume limitations from analysis needs are:

- the requirements on data to be provided from at least so-and-so many different data sources;
- requirement for every available data source to provide at least so-and-so many data records,
- the requirements on historical data to be provided for the time period of at least, for example, 5 years, etc.

If at least one data source does not meet the given requirements, the analysis could not be performed with the requested level of accuracy, or in time. Then it may not be run at all, or the business goals should be adapted to the current situation. Such requirements are taken from Data Scientists who write the algorithm or a model, and know all needs of algorithm.

There are different costs associated with the analysis performed on basis of data from different organizations. In order to reach correct decision such costs should be addressed. Regarding costs could include:

- The cost of the resources for analysis to process data;
- The costs to access data;
- The costs to transfer data and/or analysis to different location through network, etc.

At the end of the "Impact analysis"-step, the mapping between decisions on related costs and analysis needs should be created. On basis of that mapping the Accepted collaboration compositions are chosen at Convergence point.

C. Convergence Point: Acceptable Compositions

The convergence point provides an overview of remaining suitable options for collaboration between the organizations.

Associated processes:

- Formulating Zero or more Acceptable data-analysis compositions

All the steps described above lead to a limited set of workable collaboration data analysis compositions, which are called the Acceptable compositions. After the continuous checking of business needs, goal formulating, assessing all kinds of limitations, such as Trust, Disposition and Impact the convergent decision on collaboration is taking place. All the checks are used to validate each collaboration composition, and it should result in a set of acceptable compositions.

The acceptable composition represents a specific combination based on the three possible ways of collaboration:

- **D2A (Data2Analysis)**
  all data required to perform the analysis, travels to the analysis party where the analysis is performed. Typically it means that data from every data party is queried and transferred to the analysis executing environment.
- **A2D (Analysis2Data)**

data could not be copied and/or transferred to the location of analysis, and therefore, the analysis should be performed on the premises of data parties. It can take different forms:

- The analysis model is sent to every data source in parallel, and then the results are merged by the analysis party;
- The analysis runs on the premises of first data party, and with that results travels to the premises of the second data party, etc., until it has run on the premises of all data parties and achieved all needed goals. The result is then sent to the analysis party.

- **D&A2L (Data&Analysis2Lake)**

The common data-analysis lake is created on the premises of specified Trusted Third Party (to which every data and analysis party have the highest level of trust), and both the analysis and data from all involved organizations are copied to the Lake. All the operations to reach to goal then are performed within that Lake.

It may happen that after the Trust, Disposition or Impact the remaining options for collaboration forms are reduced to zero. That means that the analysis cannot be performed, given the constraint by the different organizations. It results in the

process where the evaluation should return to the upper levels of convergent pyramid, and the organizations should relax some constraints and limitations, or increase the trust level in other organizations.

D. Divergent Layers

Afterwards the technical limitations on volume of data, frequency of analysis, etc. will lead the user to specific range of technical blueprints. In the end, after going through all levels, the user receives the resulting suitable technical architectural blueprints, suitable for the current situation for that user.

We continue with the description of the divergence layers and the processes and decisions associated with every layer from the convergence point to the pyramid underneath to reach the definitive decision on every layer.

- **Technical blueprints**

Represent the mapping of the acceptable collaboration compositions to the possible technical blueprints.

Associated processes

Map each collaboration data analyses compositions onto one or more technical blueprint solutions.

- **Suitable collaboration IT architectures**

Mapping of technical blueprints to suitable technical architecture choices.

- Associated processes

Map each technical blueprint solution into one or more suitable collaboration IT architectures.

To reach the technical blueprints the following specific ordering of different steps should be established. At first, one should look into every Acceptable composition, and look at the available **technical blueprints**, connected to that composition. Every blueprint has specific established goals, limitations and constraints. The methodology follower should decide which blueprint provides the most suitable and (sub)optimized way to reach the goal of follower.

When the blueprint is chosen, the limitations and constraints should be addressed carefully in order to evaluate the current actual situation. Sometimes the constraints could be relaxed in the case of less-demanding requirements. Sometimes extra constraints should be added in order to completely fit the actual situation.

VI. Evaluation Through Use-Cases

The proposed methodology was elaborated as a part of the TKI Jongvee project [12], where one of the main problem statements of the project is the analysis of data on calves from Dutch farmers in order to advice farmers on feeding the calves. The calf related data of the farmers is located at different parties. The partners in the project have confirmed that the problem of analysis of data from different farms exists, and that this methodology has helped to discuss and take the decisions on data exchange and its analysis. The issue of disposition of data regarding the needs of analysis is acknowledged to be one of the main matters while analyzing data coming from different organizations.

Dutch NWA Startimpuls programma VWData [13] is dedicated to the elaboration of the controlled and responsible access and usage of Big Data. This methodology was presented to the partners within this program. They have gone through the presented steps for their 7 use-cases, and confirmed that this methodology works and useful in choosing the suitable Access Compositions, and afterwards in choosing the technical blueprints for their use-cases. The use-cases for VWData program came from the health sector, telecom infrastructure monitoring, and social media bias analysis. Therefore, we can make the first conclusion that this methodology is quite generic and can be applied for cross-organizational data analysis in different domains.

VII. Conclusions and Discussion

In this paper we propose a new methodology called CO-ARCH to evaluate any specific use-case when data-driven analysis in cross-organizational context is involved. This methodology provides elaboration steps in specific order to be able to come from business needs for data analysis to suitable technical software architectures in design phase of collaborative data analysis.

Unfortunately, the existing tools and platforms cannot fully support cross-organizational data analysis since different data distributions and different organizational restrictions on access to data or analysis model make the analysis much more complex.

We have elaborated this methodology within the Dutch TKI Jongvee project for the agricultural sector. Afterwards we also have checked on how it works in different domains through Dutch National project VWData. We have received the confirmation that the methodology contains all necessary steps to be able to organize and perform the data analysis in cross-organizational context.

During the development of CO-ARCH it was a surprising fact for us that we can organize all different options for data and analysis exchange to only three main Collaboration archetypes which cover all options. We have learned that Trust plays a huge role in the organization of data analysis between different organizations, and it starts the most difficult discussions between the partners. Therefore, dealing with Trust has received the most central role in our methodology. In the process of performing data analysis it has become evident that the disposition of data regarding the needs of analysis is one of the central problems while the analysis is being performed. The fact that data is not available in the form the analysis expects it, has brought a big impact on analysis model, and has started many discussions on how to deal with such different types of data and its representation. The understanding of costs to access and transfer to data/analysis has become much more deeper. It has influenced our recognition of being able to reconsider at the layers above because of the technical limitations (for example, the need to choose another model for Trust because data is just too big to transfer to analysis premises).

We consider the main achievement of this methodology the clear separations of business analysis/data needs and requirements on it from technological specifications of architectural solutions. In convergent layers of hourglass the users can firstly concentrate on business needs and requirements until they come to convergence point. Afterwards the users can elaborate the technical architectural blueprints and come to specific suitable technological stacks, already knowing that they have the remaining Acceptable

Compositions defined.

Our future work includes the elaboration on every step of the presented methodology to demonstrate in details how to tackle the relevant issues at every hourglass layer.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

B.D. van der Waaij: is the main inventor of the hourglass model and is co-writer of the paper.

E. Lazovik: is co-inventor of the hourglass model and is the main writer of the paper.

T. Albers: is creator of the impact process and contributed to in depth discussions.

M.R. Vonder: led the project, participated in the trust elaboration, contributed to in depth discussions and reviewed the paper.

All authors approved the final version.

ACKNOWLEDGMENT

This research was supported by the Dutch Top Sector High Tech Systems and Materials program via the project TKI Jongvee (full title: “Precision feeding of young calves facilitated by IoT and Big Data technologies”). We thank our colleagues from the organizations Agrifirm, CRV and Nutrifeed who provided feedback. We thank our VWDATA P7 project colleagues from UvA and LUMC.

REFERENCES

[1] Apache Spark. [Online]. Available https://spark.apache.org/
[2] Apache Hadoop. [Online]. Available: https://hadoop.apache.org/
[3] Cloudera foundation. [Online]. Available https://www.cloudera.com/
[4] TensorFlow. [Online]. Available https://www.tensorflow.org/
[5] S. Bonner, I. Kureshi, J. Brennan, and G. Theodoropoulos, “Chapter 14 — Exploring the evolution of big data technologies, software architecture for big data and the cloud,” *Morgan Kaufmann*, 2017.
[6] J. Rodriguez. The Challenges of Centralized AI. [Online]. Available: https://towardsdatascience.com/the-challenges-of-decentralized-ai-78bb44b7b69
[7] Personal Health Train (PHT) initiative. [Online]. Available: https://www.dtls.nl/fair-data/personal-health-train/
[8] H2020 project RECAP-PRETERM. [Online]. Available: https://recap-preterm.eu/
[9] Data Shield Platform. [Online]. Available: http://www.datashield.ac.uk/
[10] H2020 project Musketeer. [Online]. Available: http://musketeer.eu/
[11] IDS (International Data Spaces). [Online]. Available: https://www.internationaldataspaces.org/
[12] TKI Jongvee: Smart Dairy Farming 3.0. [Online]. Available: https://commit2data.nl/en/commit2data-program/smart-industry/tki-ltr-information-centric-networking-1/tki-jongvee-smart-dairy-farming-3-0-1
[13] NWA startimpuls programma VWData. [Online]. Available: https://commit2data.nl/en/vwdata

B. D. (Bram) Van Der Waaij received his master in computer science in 1995 at the University of Twente (Enschede, the Netherlands).

Bram worked as research employee at the Centre for Telematics and Information Technology, as a medior innovator at the Research Department of the Dutch Telecom Operator KPN. Currently Bram works as senior research scientist at TNO, The Netherlands. Bram specializes in cloud based sensor data storage, Bigdata streaming analysis and multi-party collaboration. All with a focus on inter organizational collaboration and multi-use of sensor systems, applied in various industry sectors, among others: dike management, smart grids and agriculture.

Mr. van der Waaij Msc is member of the Smart Dairy Farming consortium, is member of the Personal Health Train Architecture group, is member of the International Data Space Architecture group.

E. (Elena) Lazovik received her master in law cum laude in 2000 at the Belarusian State University (Minsk, Belarus). She received her bachelor in informatics in 2009 at the Universitá degli Studi di Trento (Trento, Italy), and master in computing science cum laude in 2011 at the University of Groningen (Groningen, the Netherlands).

Elena worked as a researcher in the Department of methodology and digital classification of law at The National Center of Legal Information of the republic Of Belarus (2000-2003). Currently (from 2012) Elena works as a Specialist Scientist at TNO, the Netherlands. Elena specializes in adaptive distributed sensor systems, data-driven processing, Big Data Engineering and Data Science, ML, constraint-based programming, cloud-native applications.

Mrs. Lazovik is a member of the Personal Health Train Architecture group, Data Council Amsterdam, Dutch Big Data Meetup, Scala Meetup, noSQL Meetup, Big Data LinkedIn group, noSQL LinkedIn group.

T. (Toon) Albers received his master in computing science cum laude in 2018 at the University of Groningen (Groningen, the Netherlands).

Toon worked as an intern at RoQua, Anchromen and Catawiki. Currently (from 2018) Toon works as a Scientist Innovator at TNO, the Netherlands.

Toon specializes in distributed data processing, adaptive IT infrastructures, distributed sensor platforms and analysis on Big Data.

M. R. (Matthijs) Vonder Msc received his master in electrical engineering in 1992 at the University of Twente (Enschede, the Netherlands) and became mechatronic development engineer at same university in 1994.

Matthijs worked as mechatronic engineer and project leader at Buhrs Zaandam (1995-1998) and as scientist and project leader at the research department of the Dutch Telecom operator KPN (1998-2005). Currently Matthijs works as senior research scientist and senior project manager at TNO, the Netherlands.

Mr. Vonder Msc is member of the Smart Dairy Farming consortium and member of the Coordinated Innovation Network (CIN) of the Dairy Brain program (https://dairybrain.wisc.edu/).