42nd IEEE/ACM International Conference on Software Engineering - Poster Track, Seoul, Korea, May 22-29 2020

A Practical, Collaborative Approach for Modeling Big Data Analytics Application Requirements

Hourieh Khalajzadeh
Faculty of IT
Monash University
Melbourne VIC Australia
hourieh.khalajzadeh@monash.edu

Andrew Simmons
School of IT
Deakin University
Melbourne VIC Australia
a.simmons@deakin.edu.au

Mohamed Abdelrazek
School of IT
Deakin University
Melbourne VIC Australia
mohamed.abdelrazek@deakin.edu.au

John Grundy
Faculty of IT
Monash University
Melbourne VIC Australia
john.grundy@monash.edu

John Hosking
Faculty of Science
University of Auckland
Auckland New Zealand
j.hosking@auckland.ac.nz

Qiang He
School of Software and Electrical Engineering
Swinburne University
Melbourne VIC Australia
qhe@swin.edu.au

Prasanna Ratnakanthan, Adil Zia, Meng Law
Department of Radiology
The Alfred Hospital
Melbourne VIC Australia
{P.Ratnakanthan, A.Zia, Meng.Law}@alfred.org.au

ABSTRACT

Data analytics application development introduces many challenges including: new roles not in traditional software engineering practices – e.g. data scientists and data engineers; use of sophisticated machine learning (ML) model-based approaches replacing many programming tasks; uncertainty inherent in the models; interfacing with models to fulfill software functionalities; as well as deploying models at scale and undergo rapid evolution, as business goals change and new data sources become available. We describe our Big Data Analytics Modeling Languages (BiDaML) toolset to bring all stakeholders around one tool to specify, model and document big data applications. We report on our experience applying BiDaML to three real-world large-scale applications. Our approach successfully supports complex data analytics application development in industrial settings.

1 Introduction

In order to successfully develop such big data analytics applications, a range of perspectives, roles, tasks and interactions need to be taken into consideration: Business perspective, including management need for the solution; Domain experts, who understand the various datasets available and how analysis of these can lead to useable value; Target end-users of the data analytics solution, i.e. the data visualizations produced - sometimes this is business management and/or domain experts, and sometimes other end users e.g. business staff, planners, customers and/or suppliers; Data analysts who have deep knowledge of available analytics toolsets to integrate, harmonize, analyze and visualize complex data; Data scientists or ML experts who have the expertise to deploy sophisticated ML software solutions; Software engineers with expertise to deploy solutions on large scale hardware for data management and computation, and end-user devices for data presentation; and Cloud computing architects who deploy and maintain large-scale solutions and datasets.

Existing ML-oriented tools only cover the technical ML and data science part of such problems, i.e. a very small part of the data analytics engineering life cycle. As identified in our recent work on end-user support for big data application development [2], while many techniques and tools exist to support the development of such solutions, they have many limitations. In general, developing big data applications suffers from several key challenges: 1) Domain experts, business analysts and business managers do not have a background in data science and programming; 2) Data analysts, data scientists and software engineers do not have domain knowledge; 3) Data scientists lack software engineering expertise; 4) Lack of a common language between team members; 5) Poorly support of solution evolution; 6) Simply re-using existing solutions is not feasible. It is not easy to choose **one** language that is **understandable among such** diverse **teams**.

2 Our Approach

BiDaML [1] is a suite of domain-specific visual languages (DSVL) to support interdisciplinary teams through the development of data analytics systems. It helps stakeholders to collaborate in specifying, modeling and documenting what and how the software should perform. BiDaML visual languages support modeling of complex, big data software at differing levels of abstraction, using big data analytics domain constructs, and can be translated into big data solutions using Model-Driven Engineering (MDE)-based partial code generation. An overview of BiDaML approach is shown in Fig 1(a). It comprises five diagram types at different levels of abstraction to cover the whole data analytics software development life cycle, from high-level requirement analysis and problem definition through the low-level deployment of the final product. The five diagrammatic types are: - Brainstorming diagram, which provides an overview of a data analytics project and all the tasks and sub-tasks involved in designing the solution at a very high-level; - Process diagram that specifies the analytics processes/steps including key details related to the participants (individuals and organizations), operations, and data items in a data analytics project; - Technique diagrams show what techniques have been used or are planned to be used for each of the tasks in the brainstorming and process diagrams and whether they were successful or there were any issues; - Data diagrams document the data and artifacts that are produced in each of the above diagrams and the outputs associated with different tasks in a low-level.; - Deployment diagram depicts the run-time configuration, i.e. the system software, hardware, and the middleware.

We report our experience in applying BiDaML to three real-world large-scale applications with teams from: realas.com — a property price prediction website for home buyers; Monash/VicRoads — a project seeking to build a digital twin (simulated model) of Victoria's transport network updated in real-time by a stream of sensor data from inductive loop detectors at traffic intersections; and the Alfred Hospital — Intracranial hemorrhage (ICH) prediction through Computed Tomography (CT) Scans.

3 BiDaML in Industry Practice

***ANZ REALas*:** A complex new model needed to be developed to improve accuracy and coverage of the property price prediction model. The project team originally comprised a project leader, a business manager, a product owner, three software engineers, and a data scientist. There was an existing working website and ML model, and a dataset purchased from a third party. Two new data analysts/scientists were appointed to the project to create new models and integrate them with the existing website. The solution had initially been developed without the use of BiDaML tool, and the challenges the team faced to communicate and collaborate through the process, was a key motivation of our research. New data scientists initially lacked an understanding of the existing dataset and solution as well as domain knowledge. Moreover, communicating progress to the business manager and other members of the team was another challenge. Together with the REALas team, we have used BiDaML to document the process from business analysis and domain knowledge collection through to the deployment of the final models in software applications. Using BiDaML was efficient in the way it took less time for the stakeholders to communicate and collaborate, and the step-by-step automatic documentation made the solution reusable for future reference. Based on the product owner's feedback "*this tool would have been helpful to understand and communicate the complexity of a new ML project within an organisation. It would assist the wider team to collaborate with data scientists and improve the outputs of the process*".

***Monash/VicRoads*:** The Civil Engineering department at Monash University sought to build a traffic data platform to ingest a real-time feed of the Sydney Coordinated Adaptive Traffic System (SCATS) data and integrate it with other transport datasets. Initially, the Civil Engineering department consulted with a software outsourcing company, who proposed a platform composed of industry standard big data tools but was unable to begin work on the project due to lacked understanding of the datasets and intended use of the platform. Furthermore, responsibility of maintaining the computing infrastructure, monitoring data quality, and integrating new data sources was unclear. We worked with transport researchers and used BiDaML to document the intended software solution workflow from data ingestion through to traffic simulation and visualization. This assisted us in formation of an alternative software solution that made better use of the systems and services already available. The BiDaML tool supported live corrections to the diagrams such as creation, modification or re-assignment of tasks. Feedback from the expert was: "*I think you have a good understanding of the business... how do you know about all of this? I think this is very interesting, very impressive what you are proposing. It covers a lot of work that needs to be done.*" While the expert stated that the BiDaML diagrams were helpful to "*figure out all the processes and what tasks need to be done*" they were reluctant to use BiDaML to communicate with external stakeholders in other organizations: "*to use this tool, it will be likely not possible, because they [the other organizations involved] have their own process, they don't want to follow a new one*." The project leader noted that an adaptation of the BiDaML data diagrams as a means to document data provenance (i.e. the ability to trace the origins of data analysis results back to the raw data used) would be "*very useful*".

***The Alfred Hospital*:** A group of radiologists, researchers and executives from the Alfred Hospital planned to use AI for predicting Intracranial hemorrhage (ICH) through CT Scans, work traditionally done by radiologists. The AI platform would enable them to prioritize the CT Scans based on the results and forward them to the radiologist for an urgent double check and follow up. Hence, a CT Scan with positive outcome could be reported in a few minutes instead of a few days. The team wanted to analyze the data before and after using the AI platform and based on the turnaround time (TAT) and cost analysis decide whether to continue using the AI platform or not. However, due to the diversity of the team it was difficult to communicate the medical terms to the data analysts and software engineers, and the analysis methods and software requirements and solution choices to the radiologists and the executive team. Thus, we used BiDaML with clinical, data science and software team members to model and document the steps and further plan for the next stages of the project. Feedback from the team was that *"BiDaML offered a simplified visual on different components of the project. These diagrams could be circulated to the project team and would clarify the workflow, requirements, aims and endpoints of each role and the entire project. In large-scale projects, BiDaML would be of even greater benefit, with involvement of multiple teams all working towards a common goal."* Although, *"The user interface seemed quite challenging to navigate. However, this could be easily negated with appropriate training and instructional material*." Fig 1(b) shows some of the diagrams for the Alfred Hospital Use-case. Full list of the diagrams and the generated reports for all the projects are available in [3].

REFERENCES

[1] H. Khalajzadeh, M. Abdelrazek, J. Grundy, J. Hosking, and Q. He, "BiDaML: A Suite of Visual Languages for Supporting End-user Data Analytics," in IEEE Big Data Congress, Milan, Italy, 2019, pp. 93-97.

[2] H. Khalajzadeh, M. Abdelrazek, J. Grundy, J. Hosking, and Q. He, "Survey and Analysis of Current End-user Data Analytics Tool Support," IEEE Transactions on Big Data, vol. 5, 2019.

[3] BiDaML Case Studies [Online]. Available: http://bidaml.visualmodel.org

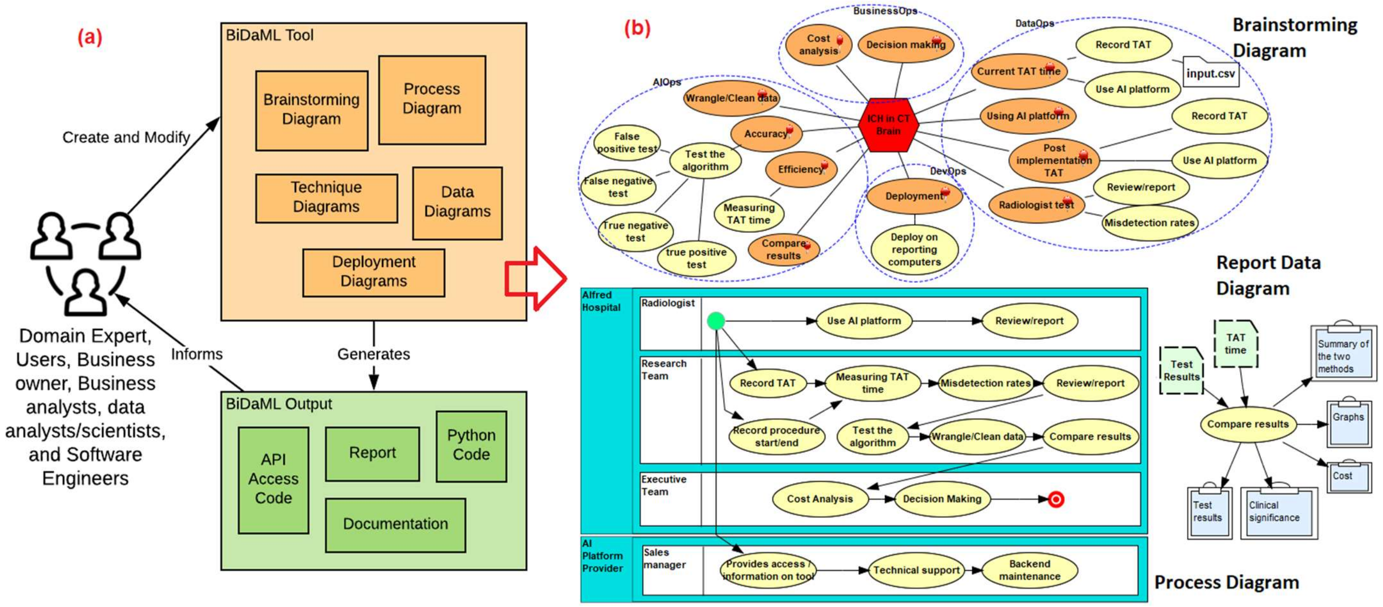

Figure 1. (a) Overview for BiDaML and (b) Examples of BiDaML Diagrams for the Alfred Hospital Use-case.