A Windows-Based Framework for Enhancing Scalability and Reproducibility of Large-scale Research Data

Nasir A. Boakye-Boateng and Raga Ahmed
University of Tennessee at Chattanooga/University of Tennessee at Chattanooga

Abstract

Graduate and undergraduate students involved in research projects that generate or analyze extensive datasets use several software applications for data input and processing subject to guidelines for ensuring data quality and availability. Data management guidelines are based on existing practices of the associated academic or funding institutions and may be automated to minimize human error and maintenance overhead. This paper presents a framework for automating data management processes, and it details the flow of data from generation/acquisition through processing to the output of final reports. It is designed to adapt to changing requirements and limit overhead costs. The paper also presents a representative case study applying the framework to the finite element characterization of the magnetically coupled linear variable reluctance motor. It utilizes modern widely available scripting tools particularly Windows PowerShell® to automate workflows. This task requires generating motor characteristics for several thousands of operating conditions using finite element analysis.

Keywords

Workflow automation, data management, ArchiMate

Framework Description

The primary objective of research projects is to develop or identify information products using specified inputs, tools, and processes. In addition to the final product, large volumes of data may be generated as inputs to the process or as intermediate data between processes. Due to the need to share data generated through research, many funding and archiving institutions have recommended standards for creating and storing research data. Graduate and undergraduate may use template-based guidelines for making data management plans for research projects[1].

Figure 1 shows the hierarchical view of the data management framework in ArchiMate 3.0 notation. Archimate[2] is an open standard by the Open Group, and it enables broad access to the information presented here. The framework uses four process interfaces to control all data management tasks from configuration to archiving.

Business Layer

At the top level is the business layer which shows the research data pipeline. It starts with research workflow information and outputs the research results. Research workflow information contains a comprehensive description of all data sources, external application interfaces, and data processing tasks undertaken in the research. The workflow configuration interface generates it in the Generate Workflow Specification task. It is implemented at the application layer through

another workflow configuration interface. Based on the workflow specification, intermediate data are acquired, generated and analyzed by the data preprocessor, data generator, and data postprocessor interfaces respectively. The workflow automation application interface is responsible for implementing the data preprocessor, data generator, and data postprocessor.

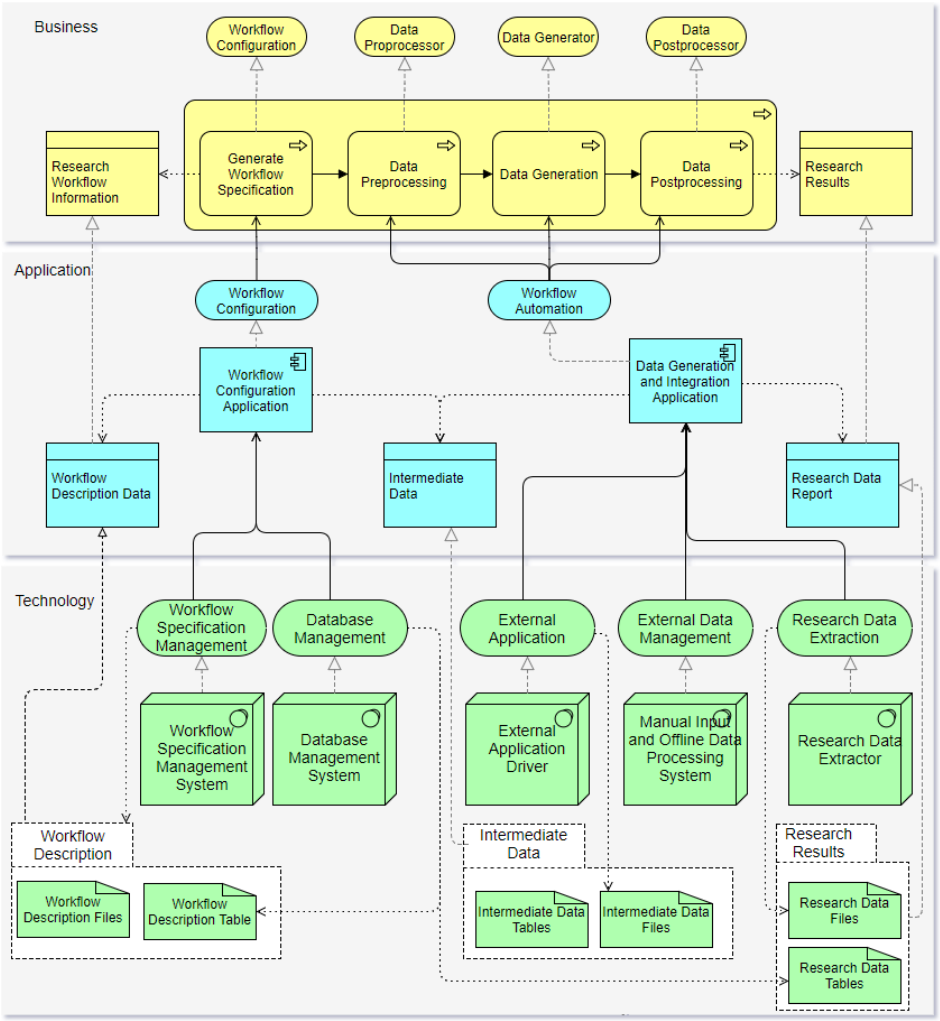

Figure 1 Hierarchical view of the Research Data Management Framework

Application Layer

The application layer has two applications namely the workflow configuration application and the data generation and integration application. The former converts workflow description data to project-specific intermediate data and the latter uses the intermediate data to produce the needed research data report. Intermediate data includes manually inputted data, field/offline generated data and application-specific files required to be converted to match the project's data archiving requirements.

Technology Layer

At the technology layer, the workflow specification management and the database management interfaces implement the workflow configuration application and the external application

interface, external data manager and research data extractor interfaces respectively implement the data generation and integration application.

Application to Ongoing Research

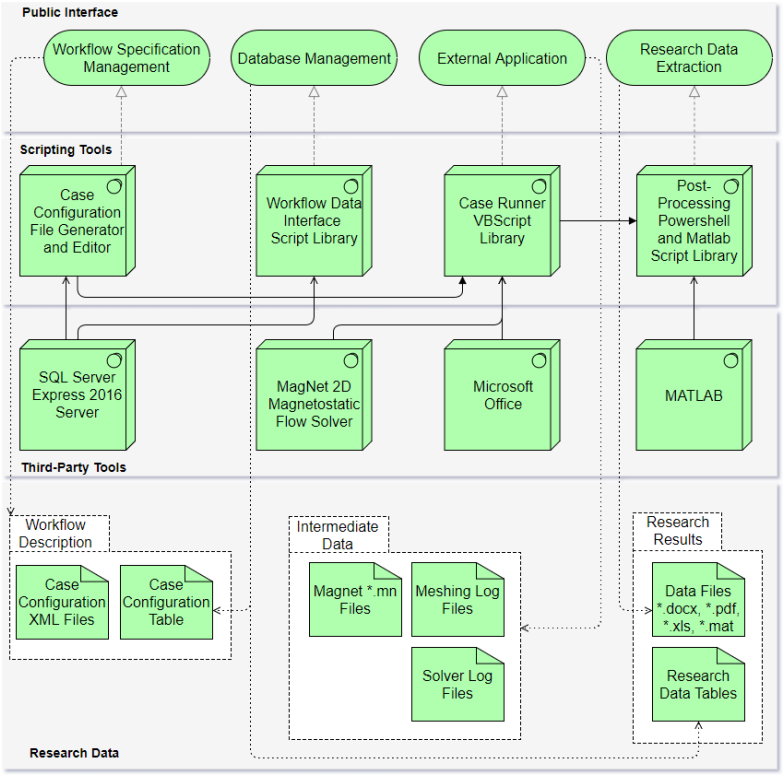

Figure 2 Research-Specific Technology Layer for Case Study

Figure 2 shows a representative implementation of the framework's technology layer as applied to a specific research project. The original plan without the framework required configuration of more than 121 finite element simulation cases each requiring at least ten minutes to set up and another ten minutes to retrieve and track the simulation results that take about 10 minutes to generate. Finally, collating all results into a single relational database required another 5 minutes. The entire process was error-prone and would have taken about 72 manhours to complete. An additional 408 hours were needed to manually create the detailed simulation reports that include field plots if required.

The layer's interfaces are implemented using PowerShell® and VBScript® script libraries that interface with underlying data processing applications to perform finite element simulations as specified by the case configuration file generator and editor. Firstly, the scripts were used to generate case configuration XML files that contained all the input parameters of the system. Based on these data, another script was used to drive the simulation software to create all the meshes for finite element simulation, configure the flow solver to the settings specified in the XML file then run the flow solver and save the results to a text file. Finally, some scripts automatically aggregate all the simulation data into a relational database and generate detailed reports including field plots of each simulation in Microsoft Word®, Microsoft Excel®, and PDF files. Files created in this application are automatically backed up locally and remotely using a

cloud backup service. Applying the framework cut the time to perform each simulation to about 10 minutes and the time to generate each detailed report to about 15 minutes. All the simulations therefore required about 21 hours to run and document and the detailed reports needed about 31 hours to generate.

Conclusion

This paper has presented a hierarchical framework (with an application-specified technology layer) for automating data management processes in a research project, and it has also shown a case study that applied the framework to reduce the length of the data generation and analysis from an estimated 480 hours down to about 62 hours. Although it uses windows scripting tools at its technology layer to implement application layer requirements, it is possible to apply this framework to a non-windows environment by modifying the technology layer while leaving the upper layers intact.

References

1 California Digital Library University of California Curation Center. Data management general guidance, 2017.

2 The Open Group. ArchiMate 3.0.1 specification, 2017.

Nasir Abdulai Boakye-Boateng

Nasir Abdulai Boakye-Boateng is a Ph.D. Candidate in the SimCenter – Center of Excellence in Applied Computational Science and Engineering at the University of Tennessee at Chattanooga (UTC). He received his MS. (2013) in Electrical Engineering from the University of Tennessee at Chattanooga and his CEMBA (2012) in Strategic and Quality Management from Kwame Nkrumah University of Science and Technology in Kumasi, Ghana. He earned his Project Management Professional (PMP) from the Project Management Institute in 2009. His research interests include motion control and variable reluctance motor technology and real-time simulation of power systems. He is a member of the Institute of Electrical and Electronic Engineers, IEEE and the Project Management Institute, PMI.

Raga Ahmed

Dr. Raga Ahmed is an Assistant Professor in the Electrical Engineering Department at the College of Engineering and Computer Science at the University Tennessee at Chattanooga (UTC). She received her Ph.D. (2013) in Electrical Engineering from the Georgia Institute of Technology, Atlanta, GA. Her research interests include control systems and embedded controls, motion control and variable reluctance motor technology in addition to engineering and K-12 STEM education. She is a member of the Institute of Electrical and Electronic Engineers, IEEE.