Development of a Cube.js Dashboard for Visualization of PARADIM Instrument Streaming Data

Avery Lenihan

2021 PARADIM REU Intern @ Johns Hopkins

Intern Affiliation: Gatton Academy, Western Kentucky University

Program: 2021 Platform for the Accelerated Realization, Analysis, and Discovery of Interface Materials Research Experience for Undergraduates Program at Johns Hopkins University (PARADIM REU @ JHU)

PARADIM REU Principal Investigator: David Elbert, Hopkins Extreme Materials Institute (HEMI), Johns Hopkins University

PARADIM REU Mentor: David Elbert, Hopkins Extreme Materials Institute (HEMI), Johns Hopkins University

Primary Source of JHU REU Funding: ***Support for PARADIM is provided under NSF Grant # DMR-1539918 as part of the Materials Innovation Platform Program

Contact: avery.lenihan302@topper.wku.edu, elbert@jhu.edu

Website: ??

Primary JHU PARADIM Tools Used: Apache Kafka Streaming Platform

Program ??

Abstract

With the Materials Genome Initiative supporting the development of machine learning and predictive models regarding material properties, data needs to follow FAIR principles to be incorporated into such models. However, making data findable, accessible, interoperable, and reusable, takes either a well-developed streaming infrastructure or manual input. As most materials researchers have little interaction with data science, illustrating the streaming of data in an accessible way serves as a vehicle to acquaint them with data and possibly motivate them to build an infrastructure in their own lab. As PARADIM already has a well-established infrastructure, Cube.js was able to be connected as an intuitive frontend tool to generate useful charts and display key metrics.

Summary of Research

The Materials Genome Initiative (MGI) depends on FAIR data principles (Wilkenson *et al*) to feed data-intensive techniques, including machine learning, and accelerate the materials-by-design process. To support FAIR data sharing, PARADIM is developing cutting-edge infrastructure for streaming materials research data. Currently, seven instruments and sensors stream data through an Apache Kafka backbone. Kafka Eagle stores metrics for the data and servers in a SQL database (MySQL). A dashboard of these metrics would benefit users and staff by providing progress of data collection as well as insights into the efficiency and use patterns of the PARADIM infrastructure.

This project consisted of evaluating needs for visualization of streaming data metrics and development of a prototype for a data dashboard. Multiple options were evaluated and Cube.js was selected due to its performant backend API to link infrastructure metrics in a database to a flexible, easy frontend dashboard building. The prototype development highlighted the need to provide an accurate data schema for dashboard viability

Results and Conclusions

The need for a dashboard to visualize streaming data metrics was identified through discussion with PARADIM staff and users. Several existing dashboards were evaluated, including Kowl, Confluent Cloud, and Datadog, however, none provided the flexibility and extensibility appropriate for an instrument-based laboratory as diverse as PARADIM. Interviews of lab users revealed that flexibility was a top priority, but existing dashboards were large, overwhelming, and less modular. The possibility of developing a custom frontend in the React.js framework and Material-UI library was considered but was ruled out because of its static nature.

To meet the needs of the project, I settled on Cube.js. Cube.js can be characterized as data middleware that connects databases with a frontend that orchestrates SQL generation, caching, and security. The Cube.js frontend Dashboard App provides a way to make a modular, dynamic dashboard as a React App without requiring substantial user coding. For this project, I ran Cube.js within a Docker container which was convenient as well as promising easy replicability on other machines.

A central challenge in applying the Cube.js platform to PARADIM's infrastructure is development of a schema that includes all the metrics of value. While Cube.js will autogenerate a schema, this default failed to recognize numeric data types and did not provide all important fields. Manual schema production corrected type errors and allowed creation of cube queries to generate five charts of varied usefulness. Even with this limited schema, however, SQL queries can be autogenerated from a GUI. This provides valuable function to users interested in customizing the dashboard despite lack of SQL experience.

Manual schema editing including version control in a public GitHub repository (Lenihan *et. al*). The schema is in a specialized JSON format, but editing data types and adding new fields is possible in common text editors. The modified schema allows metrics such as lag to be easily added to charts. Further refinement of the schema is needed moving forward, but the dashboard currently could generate a number of charts for the inexperienced user to begin understanding data flow through Kafka streaming in PARADIM.

As streaming infrastructure expands to other labs, the implementation of Cube.js to read data from Kafka Eagle databases should prove valuable elsewhere. The continued need to integrate data awareness with the daily operations of a lab to create FAIR data means automated data infrastructure will become more common thereby increasing the value of solutions to

visualize operations of the infrastructure. The use of this dashboard at PARADIM will be a case study on successful integration of users in a lab with a well-maintained infrastructure. The easy replicability of this process makes it appealing for labs with a data infrastructure that often goes unnoticed.

Future Work

To improve the prototype dashboard future steps will need to focus on transferring all metrics and desired derived fields from the database to the Cube.js schema. The schema must be gone through to recharacterize data and add categories. Additionally, categories may be added that include calculations, such as average messages per topic. The dashboard could then be expanded to also look at another database that outlines specific activities of each machine.

Acknowledgements

Special thanks to David Elbert, Tyrel McQueen, James Overhiser and Darrell Schlom.
Support for PARADIM is provided under NSF Grant No. DMR-1539918 as part of the NSF Materials Innovation Platform Program.

References

- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- Lenihan, A., Elbert, D., McQueen, T.. Dashboard App https://github.com/paradimdata/dashboard.app