Database Theory Project: SystemDS

Kaleb Bishop, Nhat Phan, Jackson Clay Spieser

As a part of Database Theory (CS6051), we were tasked with studying a system within a database field. This would allow us to learn and share knowledge of the works and systems chosen. As a team, Kaleb Bishop, Nhat Phan, Jackson Clay Spieser are researching SystemDS, “An open source ML system for the end-to-end data science lifecycle” (Apache, 2024).

Our team has been preparing for the project by dividing roles among our team members to ensure that adequate work has been done on all aspects. So far, we have all spent time familiarizing ourselves with SystemDS by going through its official documentation, setting up our personal computers with SystemDS and performing small scale tasks to test its functionality.

Our primary reference for this project is the official SystemDS documentation which provides details about the platform’s features and how to use them (SystemDS Documentation, 2024). This documentation is the best way to understand the system and how to implement functionality into our project. We are also using a few academic papers to understand on a more understanding of SystemDS’s theoretical purpose. One of the main papers that we are using is *“SystemDS: A Declarative Machine Learning System for the End-to-End Data Science Lifecycle”* by Matthias Boehm and others. This paper provides an in-depth explanation of SystemDS’s architecture and its role in scaling machine learning applications. For some additional help development, we utilized community forums such as Stack Overflow for troubleshooting issues.

As we continue to work on this project, we all have gained a deeper understanding of SystemDS. We now understand its importance as an open-source platform designed for end-to-end data science. SystemDS has tools to perform large-scale machine learning tasks and can be easily integrated into any tech stack with its compiler, runtime, and API components. One of the major perks of SystemDS is its ability to run effectively on distributed systems allowing for seamless scaling.

For our demo, we plan to showcase a few scenarios that demonstrate the strengths of SystemDS as mentioned above. Our first scenario will involve implementing a large-scale linear regression model on a large dataset to demonstrate SystemDS’s ability to handle computationally demanding tasks. We will then compare the same implementation on other platforms such as TensorFlow and PyTorch to highlight their performance differences. Our second scenario will focus on SystemDS’s optimization algorithms where we will take another large dataset and compare performance enhancements. We will write test scripts to measure SystemDS’s capabilities on metrics such as execution time, memory usage, and accuracy over different conditions such as database size, computation cluster size, and machine learning model. Through these scenarios, we hope to highlight SystemDS’s place in the world of machine learning systems for data science.

In conclusion, our project is making strong progress in both implementation and our understanding of SystemDS. By focusing on the official documentation and the academic

research of Matthias Boehm and others, we will continue to push forward toward our demo scenario goals. Our scenarios will hopefully provide evidence of SystemDS's strengths in large-scale machine learning tasks.

Reference

Boehm, Matthias, et al. “SystemDS: A Declarative Machine Learning System for the End-to-End Data Science Lifecycle.” *Arxiv*, 8 Jan. 2020, Accessed 30 Sept. 2024.

SystemDS, Apache. “Apache SystemDS - an Open Source ML System for the End-to-End Data Science Lifecycle.” *Apache SystemDS - An Open Source ML System for the End-to-End Data Science Lifecycle*, systemds.apache.org/. Accessed 30 Sept. 2024.

“SystemDS Documentation.” *SystemDS Documentation - SystemDS 3.3.0-SNAPSHOT*, apache.github.io/systemds/. Accessed 30 Sept. 2024.