MIT LINCOLN LABORATORY

The Bulletin

News, events, and notices transmitted weekly

April 14, 2017

Staff Release Open Source Software for BigDAWG Polystore System

Cyber Security and Information Sciences | Lincoln Laboratory

Working with a vast variety of data is often a challenge for organizations. Imagine, for example, all of the different types of medical data that hospitals record—such as vital sign measurements, results of laboratory tests, descriptions of patients' symptoms, etc. It would be difficult to organize such diverse data into a single database engine. But, the ability to connect multiple engines through a single user interface would greatly improve an organization's ability to access all the data, and help them analyze it without having to know which specific database to probe.

The Big Data Analytics Working Group (BigDAWG) polystore system, developed in part by Laboratory staff, is simplifying the integration and analytics of data. The system is a prototype of a "polystore" concept—a database management system that connects multiple, heterogeneous database engines behind a single programming interface. The interface allows users to enter queries, which BigDAWG automatically optimizes and responds to by moving data across database engines without user intervention. The team recently released the first open source version of the BigDAWG software.

Kyle O'Brien (left), Group 104, and Dr. Vijay Gadepally, Supercomputing Center, are part of the team behind a new database management system called BigDAWG.

"BigDAWG is beneficial to anyone seeking a simpler way to use data that spans multiple data models," said Dr. Vijay Gadepally, Supercomputing Center. He has led the BigDAWG effort for close to two years, collaborating with Kyle O'Brien, Intelligence and Decision Technologies, Group 104; Dr. Jeremy Kepner, Laboratory Fellow, Supercomputing Center; and a team of researchers from a dozen universities around the country. The collaboration is fostered through the Intel Science and Technology Center for Big Data—a series of research collaborations that Intel is establishing with U.S. universities to prototype revolutionary technologies. "I'm very excited that we have essentially opened up a new research area with BigDAWG," Gadepally said.

Traditionally, if a user wants to access all of their data that is split into separate database engines, they

Staff Release Open Source Software for BigDAWG Polystore System (continued)

have to code multiple connectors to link the data into a single data store. The BigDAWG architecture greatly reduces the need to write single connectors between each database the user wants to connect. The architecture is constructed in layers. At the top, the interface receives a user query, and passes it below to the appropriate "island" for execution. An island consists of a query language and a specific data model, and is associated with one or more database engines. A "shim" then connects an island to the engines below. "A shim basically serves as a translator that maps queries, expressed in terms of the operations defined by an island, into the native query language of a particular storage engine," Gadepally explained. Software components called "casts" move data across database engines as needed, and users are presented with the results of their query.

So far, the team has tested BigDAWG on two real-life use cases. The first was in collaboration with MIT's Chisholm Lab, whose datasets encompass ocean metagenomic data, including millions of bacterial DNA samples and associated data like the depth, salinity, and temperature of water the samples were collected from. The second was on an intensive-care unit dataset named MIMIC II, which comprises de-identified health data collected from about 40,000 critical care patients from Beth Israel Deaconess Hospital. In both cases, BigDAWG provided researchers with real-time support for streams of diverse data. "These two examples do a great job of simulating the volume, velocity, and variety of real applications. Working on these problems let us work through real issues such as messy and missing data, unknown or unquantifiable performance metrics, and large-scale data," Gadepally said.

The open source release allows users to download and begin testing the BigDAWG polystore system with real data. Scripts included in the release allow users to download publicly available parts of the MIMIC II dataset and load them into three database engines. "Patient history data is inserted into PostgreSQL, physiologic data is inserted to SciDB, and free-form text data is inserted into Accumulo. In a few minutes, users can have three databases running and issue cross-engine queries to them without having to install the databases permanently," O'Brien said.

The team hopes that research analysts, data scientists, and database administrators will try BigDAWG with MIMIC II and share feedback. Their goal is to continue developing the system to run as simply and automated as possible. "We recently organized a workshop at the IEEE Big Data conference, and it was amazing to see our work being followed by top researchers in three continents," Gadepally said. The team's broader goal is to further exploration and innovation in the field.

"This release was just our first and has opened up a number of fascinating problems," Gadepally said. "We hope to have many more releases in the future. We want to continue expanding the system based on user feedback, and apply our work to new and challenging problems." BigDAWG is available for download at bigdawg.mit.edu.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.