BigDAWG Tutorial Proposal

for IEEE HPEC 2017

Tutorial Title: Managing Heterogenous Data with the BigDAWG Polystore System

Tutorial Presenters: Vijay Gadepally (MIT Lincoln Laboratory), Timothy Mattson (Intel Corporation), Kyle O'Brien (MIT Lincoln Laboratory)

Tutorial Description:

Modern decision support systems must integrate and synthesize a rapidly expanding collection of real-time data feeds: sensors, analyst's reports, social media, documents, logistical data, and more. A storage engine tuned to the structure of a particular data feed can deliver orders of magnitude performance improvements relative to a more general-purpose database management system. In response, a range of specialized storage engines have been crated with many modern decision support systems contain five or more distinct data storage engines. This diversity of storage engines has resulted in a shift to era of "polystore" databases characterized by the following guiding principles:

- One size does not fit all.
- Any big data application will fundamentally deal with data in multiple storage engines.
- Real-time decision support is crucial.
- The interface to big data applications will move from today's form-based approaches to interactive workflows organized around visualizations of the data.

BigDAWG is a polystore database developed by the Intel Science and Technology Center for Big Data. This tutorial will introduce participants to BigDAWG and walk them through the steps of developing a polystore database solution for a medical processing problem consisting of multiple data types each persisted in its own custom database. You will learn to handle each dataset in its own custom database and perform analytical tasks that require integration of data from each system.

Tutorial Outline:

- Introduction, Motivation
- What are Polystore Systems
- Database evolution in the Big Data era
- BigDAWG polystore system
- Hands-on: Using BigDAWG for large, heterogeneous data (using real-world medical datasets) (optional)

Potential audience:

Researchers and practitioners interested in working with heterogenous datasets. Students of all technical backgrounds will be welcome. Only minimal programming experience or background in big data will be required for hands-on exercises.

Material to be distributed to participants:

Tutorial slides will be distributed to participants. Open sourced code will be distributed for participants interested in taking part in the hands-on exercises.