

## Evening lecture: The future of Big Data: Polystore, specialized storage engines, and embedded analytics.

*Friday, 28 October 2016 18:30 (1 hour)*

Theory is nice when trying to understand Big Data systems, but nothing beats experience with real data. Working with the MIMIC II data set (data from an intensive care unit) we've concluded that:

1. Data must match the storage engine if you care about performance
2. Data in flat files is almost equivalent to deleting it.

Or, turning these conclusions into slogans, "one size does not fit all" and "we need to bring the power of a database to all data". In this talk we describe our ongoing work to create a system that responds to these slogans. We call this the BigDAWG Polystore system. A Polystore system contains multiple storage engines integrated behind a common API but exposing features of individual storage engines as needed.

We are also working on a new storage engine tuned to the needs of sparse array data called TileDB. TileDB has entered production usage at the Broad Genomics institute. Our continuing work with TileDB is to extend it to dense arrays (thereby competing with HDF5).

Finally, we believe that key analytics functions need to be integrated into the storage engines. We'll describe our early efforts to create GraphBLAS routines and other machine learning primitives integrated into our Polystore system.

**Presenter:** Dr MATTSON, Tim (Intel)

**Session Classification:** Sessione di venerdì'