**Striped Data Analysis Framework**

Oliver Gutsche (FNAL), Igor Mandrichenko (FNAL)

**Goal: Reduce Time to Insight**

Reduce analysis turn-around time

Programmatically, traditional analysis is an iterative process, repeating:

- Skimming (drop not interesting events, disk-to-disk)
- Slimming (drop unneeded attributes, disk-to-disk)
- Filtering (selectively read events into the memory)
- Pruning (selectively read attributes into the memory)

Provide access to the **needed data and only** to the needed data

- Direct, scalable, efficient

Eliminate the need to skim or slim data as a disk-to-disk operation

**Innovation: Striped data representation – move from file-based to database-based analysis paradigm**

- **Variation of columnar representation**
- **Columns are broken into stripes at the event group**
- **1K-10K events boundaries**
- **Data Stripe** – one column of data for one event group
  - **unit of representation** efficiently stored in a key/value database
- **Can be used for variety of non-HEP data**

**Implementation: cloud-friendly client/worker/database architecture**

- **Striped Data Server**
  - Distributed, scalable, redundant no-SQL key/value storage
  - Web service with simple REST interface, web cache
- **Computing Component/Workers**
  - Worker is a single-threaded stateless process with its private data cache
  - Cloud-ready: can be deployed elastically, using Docker containers
- **User code - Python, Jupyter notebook compatible**
  - ~30M events/second performance on
  - 132 core demo 11-node cluster
  - 130 TB 25-node CouchBase data cluster

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.
This project is funded by FNAL-LDRD-2016-032