Exploratory *Big Data* Analysis for HPC Simulations at Scale

LLNL Mentors/POCs
Ming Jiang (jiang4@llnl.gov) and Brian Gallagher (gallagher23@llnl.gov)

Mentoring Strategy
We plan to guide the students through the project, providing coding and design mentoring. We plan to hold weekly telecons to address questions and discuss progress.

Abstract
The next-generation of high performance computing (HPC) scientific simulations will produce data of unprecedented size. Analyzing and visualizing this massive amount of simulation data will require fundamentally new methodologies in both software and hardware. In recent years, the advancement of *Big Data* technologies has provided the necessary means to explore massive amounts of data, especially in the commercial industry. Effectively exploiting these *Big Data* technologies for analyzing and visualizing HPC simulation data has been an ongoing challenge.

To help address this challenge, we invite BYU students to join LLNL Computer Scientists in a new research and development effort named **Alkemi**. The Alkemi framework provides a data analytics approach to improving HPC simulation workflows. HPC simulation workflows are highly complex and often require a manual tuning process that is a significant pain point for users. Developing a simulation workflow is often a trial-and-error process that can be disruptive and time consuming. In order to reduce user burden and improve efficiency, we will develop the Alkemi framework that consists of novel predictive analytics for HPC simulations and a novel infrastructure for integration of analytics. Our goal is to predict simulation failures ahead of time and proactively avoid them as much as possible.

In this project, students will develop a software library within the Alkemi framework that can provide exploratory *Big Data* analysis capabilities for HPC simulation data. This software library will integrate two key pieces of technology within the Alkemi framework: data aggregation and visual analytics. To enable exploratory analysis at scale requires a combination of: (a) heavy lifting by the back-end to aggregate, sample, or otherwise reduce the data, and (b) effective presentation of and interaction with this reduced data on the front-end.

Proposed Project Tasks
- Learn about *Big Data* tools ecosystem:
  - **Apache Spark**: a fast and general engine for large-scale data processing that can run in standalone mode as well as alongside a Hadoop cluster.
  - **Apache Zeppelin**: a web-based notebook that enables interactive data analytics and visualization using D3.js. It provides a capability to integrate with Spark.
- Software library implementation:
  - **Data ingestion**: implement code to read large-scale simulation data from LLNL in HDF5 format and convert them to RDD format suitable to cluster computing via Hadoop.
  - **Data aggregation**: use the Spark library to implement techniques for aggregating simulation data in different modalities: spatial, temporal and categorical.
  - **Data visualization**: use D3.js web-based library to implement visual analytics for visualizing and interacting with aggregated simulation data in order to glean insights.
  - **Software integration**: develop a pipeline of software modules using Zeppelin to enable exploratory *Big Data* analysis for simulation data that can be driven by users.