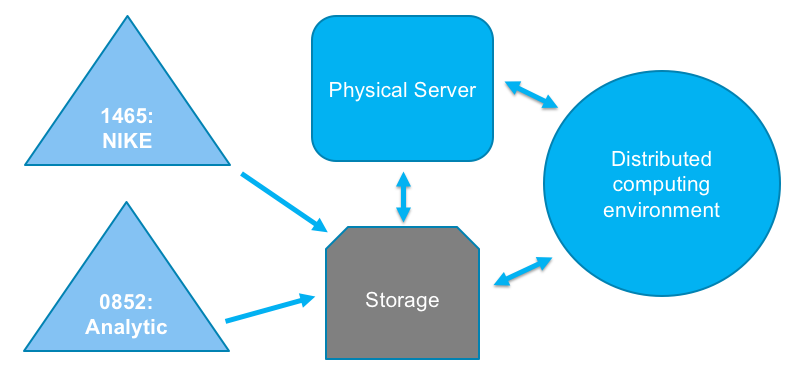
# ACE Sandbox Server: Proposal

The information below describes settings for the Data-Science-Sandbox server used by members of the Analytics Center of Excellence. These settings should facilitate numerous team projects over time. Modifications can be made incrementally; project-specific modifications should be avoided when possible.

Please contact **Yuchen Li and Yanglin Li** before considering any modifications to the server set-up.



## Objective:

* Build an in-house physical server as a team sandbox environment for model fitting and testing
* To accommodate 20-30 users on ongoing basis

## Date Requested:

* 2/1/2018

## Affiliated Projects:

* Various ACE-affiliated projects (Social Determinants of Health and etc.)

## Data Restrictions:

* General HIPAA considerations (for MarketScan initially, possible extension to other datasets)
* General guidelines for client data (Lockheed Martin and etc.)

## Operating System:

* Red hat enterprise Linux 7.3 or later

## Space:

* 5TB

## RAM:

* 512GB

## CPU:

* 24 (suggestions of 24 probably sufficient. Please use your judgement, ideally a rounded number. This should be general and future-proof (i.e., not project specific). In other words, max it out within reason (but not at the expense of other structural aspects of the set-up)

## 3rd Party Software:

* Docker(17.06.2-ee-6),
* Flexible Analytics (via API)
* Anaconda Python (>=3.6: pandas, numpy, sklearn, TensorFlow)
* R and Rstudio
* SQL (or similar tool for relational database)
* Git (or similar tool for version control)
* SPSS (and related modeling software)
* Jupyter Notebook

## Data Access:

* Attached storage with NFS/sand mount to 0852 and 1465 server

## Anti-crash Mechanism:

* Step 1: Create Docker image with a fixed amount of resources, potential crash will occur within Docker image
* Step 2: Crashed Docker image will be reassigned to distributed system for processing

# Future Proof:

Hadoop is an open-source software framework used for distributed storage and processing of datasets of big data (in TBs) using the MapReduce programming model

## Solution:

* Attach and utilize in-house Hadoop cluster, which has 30 nodes and 6TB of memory

# Management Roles:

* Grant permission to build an in-house physical sandbox server that can access 0852, 1465 and client data through physical copy or NFS/sand mount.
* Grant permission to reactivate, attach and utilize in-house Hadoop cluster