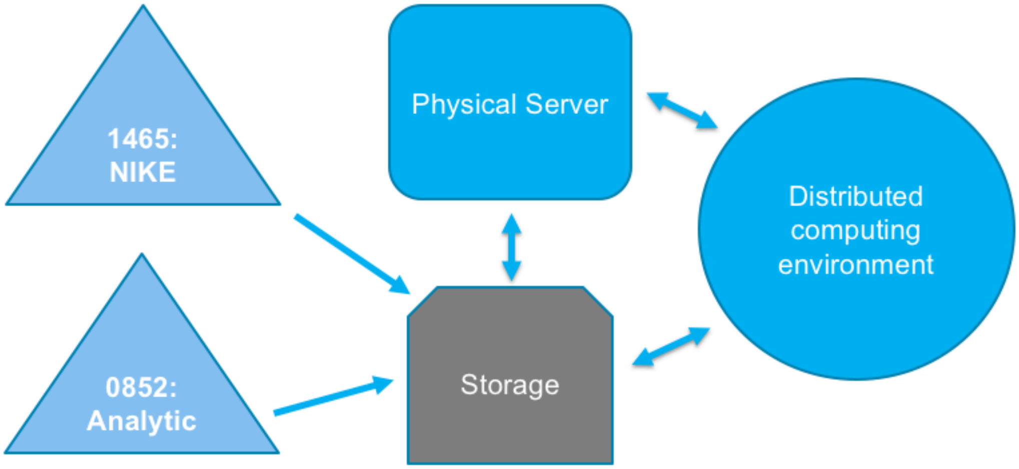
ACE Sandbox Server: Proposal

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 (special thanks to Stuart Jackson).



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:

- 1/31/2018

Timeline:

- Mid-late February

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:

- Rad hat enterprise Linux 7.3 or later

Space:

- 12TB

RAM:

- 512GB

CPU:

- 32 (suggestions of 32 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)
- R and Rstudio
- SQL (or similar tool for relational database)
- Git (or similar tool for version control)
- Jupyter Notebook

Data Access:

- Attached storage with NFTS mount to 0852 and 1465 server

Anti-crash Mechanism:

- Step 1: Create virtual machine on physical server for each user with fixed memory, session abort when memory is maxed
- Step 2: Aborted session 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 access 0852, 1465 and client data from sandbox server, through physical copy or SFTP (SSH File Transfer Protocol)
- Grant permission to reactivate, attach and utilize in-house Hadoop cluster