1. Lin and I have been working on setting up the new Analytics server for ACE team in pursuit of a more streamlined data science work. I want to take this opportunity to share with you guys sort of the blueprint of new Analytics sandbox and development server.
2. I assume most of you are quite familiar with the two servers we currently have and their capability. To give you a brief summary, 0852, aka, Analytics server, carries MarketScan, data mentor, client data and etc. Analytics server is very crowded because multiple scripts, including R, Python and SAS, and competing for limited resource, which I consider the major cause for rent crashes. 1465, aka, NIKE server, hosts MarketScan. It only SAS installed hence it has been behaving as usual.
3. There is no template for solving a data science problem. The roadmap changes with every new dataset and new problem. But we do see similar steps in many different projects. I wanted to make a clean workflow to serve as an example, I also wanted to give people working with data scientists an easy to understand guide to data science.
   1. The objective: What is the problem you are trying to solve? Remove modeling, evaluation metrics, and data science from the equation.
   2. The preparation: Data can come from a variety of sources. You can query SAS data from MarketScan, import CSV file from your local machine, or use a web scraper to strip data from the internet. Once data is loaded, you should start exploring variables, develop hypotheses and patterns
   3. During the analysis and reflection phase, you will use a variety of algorithms to perform a wide range of tasks. You will begin with intuitive models such as regression and tree, and gradually advance to black box model such as neural network
   4. In my mind there are two directions your data science project can go, a report and, ultimately, a product.
4. We used data science workflow as a reference, compared it with what we currently have, and set up the objectives that seek to solve current problems and future challenges.
   1. Build a development server that run final models, archive final code, metadata, and interface with production engineering team.
   2. Build a sandbox server that provide space for model fitting and testing, archive code, metadata and interface with development server
5. The two servers will be built on the 4 build blocks: capacity, connectivity, compatibility and compliance.
   1. …
   2. …
6. In terms of capacity that plays an integral part in Preparation, and Analysis phase, we will set a new physical server with two attachment, storage and distributed computing environment.
7. In terms of connectivity, storage will communicate with 1465 and 0852 servers through NFS/Sand mount. NFS provides network-wise communication while sand mount is connected through fiber.
8. Git and Docker seek to provide compatibility issue by providing means of version control and standardization, while Docker also plays an important role in alleviating potential crashes
9. Docker is a realization of container technology. You can think of Docker image as shipping container for codes, system tools, system libraries, settings.
10. Docker provides standardization and resource control. Rather than building a new environment for every analysis, you can put the tools and packages required for certain types of analyses (e.g., SAS, python, etc.) into a container, create an image of that container, and have every user boot up an isolated, standardized environment from that image. On the other hand, although a Docker image, by default, has no resource constraints and can use as much of a given resource as the host’s kernel scheduler allows. Docker provides ways to control how much memory, CPU, or block IO an image can use, errors will crash and abort the image, hence protecting the server.
11. Last but not least, the new servers will follow HIPPA and general client data guidelines. Since we are working with PHI, confidentiality and integrity will always be our top priorities.
12. The last slide has detailed server spec for development and sandbox server side by side, any questions?