RUTGERS
CAMDEN

Rutgers, The State University of New Jersey
Camden Campus

Current Big Data and Advanced Cyberinfrastructure Initiatives

1) Education and Training

The Rutgers Center for Computational and Integrative Biology (CCIB) brings together leading research academics from the departments of Biology, Mathematics, Computer Science, Chemistry, and Physics. The Center is currently offering both Masters and Doctoral degrees in Computational and Integrative Biology (CIB). Students in the Graduate Program in Computational & Integrative Biology will work on the development of mathematical models for biological systems, application of the models to data from laboratory and field investigations, the adjustment of the model based on its fit to and predictive value for experimental results, and the subsequent modification of the experimental design based on the predictions of the model.

The most data-intensive portion of the CIB program involves the teaching of Genome Informatics. This includes mining the complete genomic sequences of human and other species, with the applications in evolutionary and disease biology (primarily cancer) as well as metagenomics. This requires significant data throughput and vast amounts of disk space and processing power as the number of available sequences grows exponentially. The data is produced from the Next Generation Sequencing (NGS) instruments. The students perform both coursework and research projects working with actual sequence data. Currently there is not an NGS instrument onsite. All data sets need to be produced externally and transferred over the network. These transfers will often take days for a few hundred gigabytes(GB). If the data repository host has an Aspera server then a few terabytes(TB) can be transferred in the same amount of time. Globus Online is another software based option that will be used to manage, share and reduce data transfer time.

The Business School is currently building a new degree program to offer a Master in Science in Business Analytics. This program will benefit greatly from computational and data resources. Business Analytics (also referred to as Business Intelligence) includes skills and technologies needed to examine the vast quantities of data organizations are collecting, in order to help organizations derive strategic insight and improve the quality of decision-making. Business Analytics involves knowledge of the collection and storage of data (especially from electronic sources, such as the web), methods of statistical analysis, mathematical and predictive modeling, as well as optimization techniques to improve business performance.

The Computer Science department is changing its Master's program (from Fall 2013) to focus on Computational Science (see http://cs.camden.rutgers.edu/graduate/) and will offer a Master of Science in Scientific Computing. In particular, Prof. Sunil Shende will be developing a new course in this program called "Big Data Algorithms" to study algorithmic techniques and modeling frameworks that facilitate the analysis of massively large amounts of data. The course will also include an introduction to information retrieval, streaming algorithms and analysis of web searches and crawls. It would be wonderful if the big data initiative can help us add more resources and infrastructure to the campus for supporting this and other courses that are specifically meant to train students in handling huge data sets for computationally intensive tasks.

2) Multidisciplinary Research Expertise

The research expertise of the Rutgers Camden Campus spans all of the sciences integrated within CCIB; Biology, Mathematics, Computer Science, Chemistry, and Physics. The Business School has faculty involved in the new computationally focused programs in Business Analytics and Computer Science Department has faculty expertise of those involved in the Scientific Computing degree program.

Example faculty relevant to Big Data/Cyberinfrastructure Initiatives:
Prof. Andrey Grigoriev (Biology) is focused on mining the complete genomic sequences of human and other species, with the applications in evolutionary and disease biology (primarily cancer) as well as metagenomics.
Prof. Benedetto Piccoli (Mathematics) is specialized in research about large cyber -infrastructures related to traffic monitoring. An example is the Mobile Millennium Project of UC Berkeley. Prof Michael Palis (Computer

Science) does research in parallel and distributed computing, as well as real-time computing, from the algorithmic systems design perspective. His work includes considering the systems and architectural issues - and the attendant algorithms – that need to be addressed in order to support HPC and Big Data (as well as real-time) applications. Prof. Sunil Shende (Computer Science) and Dr. Kwangwon Lee (Biology) formed a collaboration (within CCIB) that involves the analysis of large, genome-wide single-nucleotide polymorphism (SNP) sequences for various strains of N.Crassa. From the big data perspective, they are trying to develop algorithmic approaches and pipelines that could potentially require fast networks/clusters for handling and processing this data. Prof. Hao Zhu (Chemistry) is an expert of cheminformatics and computational toxicology. His research focuses on the use of the enormous volume of data obtained from the High Throughput Screening assays and the development of new tools that enable utilization of this diverse biological information of chemicals in order to generate informatics models. Prof. Grace Brannigan (Physics) uses HPC to conduct large-scale molecular dynamics simulations of biological macromolecules. Faculty in the Business School are working on problems in the following subject areas: dynamic optimization problems requiring massively parallel processing to get a solution in reasonable time; and the analysis of high frequency trading (HFT) data. The HFT is a huge issue for regulators as they have no idea on how to process/analyze these truly BIG data (dozens of terabytes of data are generated every single trading day).

3) Infrastructure

CCIB was awarded an NSF MRI grant to install a cluster to support CCIB computing. The first phase of the cluster will consist of 12 compute nodes for distributed memory computing and 1 vSMP (ScaleMP) with ~1TB of memory for large memory jobs. There will be a scratch file system using GlusterFS with an initial capacity of ~9TB. The most current release of this file system is capable of supporting Hadoop applications natively. The vSMP will have a dedicated local scratch disk as well of ~4TB. Each scratch disk can be doubled by adding disks to the existing chassis. The home directory storage consists of a single raid 6 with chassis that can hold 45 disks. The initial usable capacity will be ~25TB using 12 3TB disks. It is possible to add an additional 124TB of raw capacity using 31 additional 4TB disks. This data will be backed up to a new and relatively inexpensive LTO 5 tape library system. The network uses both FDR infiniband and Gb ethernet. All nodes and vSMP nodes can host two full size accelerator cards, Nvidia Kepler and/or Intel Phi, with 2x PCIe 3.0 x16 interfaces per node.

The cluster described above will be the only cluster located in Camden when it is installed. All computational work aside from this is done on a few individual servers, workstations or remotely using XSEDE HPC resources. In all degrees programs, M.S. in Scientific Computing, M.S. in Business Analytics, and M.S/PhD. in CIB, the students will need to be trained effectively and any additional infrastructure would be greatly beneficial for accomplishing this. An FPGA based server by Covey Computer is being considered by CCIB.

The external network connection for the campus consists of a shared 1Gb connection to Rutgers main campus. It is not known at this time when there will be an upgrade. The current network speed is expected to be a major bottleneck for data transfers where the data is generated by an NGS instrument off-campus and analyzed on-campus. The new programs in Computer Science and the Business School may require increased bandwidth too.

4) Industry Partnerships

Prof. Andrey Grigoriev is in discussions with cancer clinics outside NJ regarding analysis of patient data and this will require further hardware and network resources. Prof. Benedetto Piccoli is developing a collaboration with Octotelematics US to work on traffic data. There are a variety of possibilities for industry partnerships for both Computer Science Department and the Business School.

From a business school perspective, the biggest application area for big data would be related to customer analytics. Companies are attempting to develop an integrated view of customer behavior that includes web site visits, clickstreams, calls to customer support, purchases, social media analytics, customer complaints and customer loyalty. One potential contribution of the business school could be to help NJ firms to develop a "proof of concept" for big data applications - i.e. before a NJ company seriously invested in big data applications, it might be beneficial for them to "play out" big data analytics using Rutgers infrastructure and resources (there might be a small nominal fee for this). Once they are convinced about the ROI, they may choose to expand their own big data infrastructure.