**Research and Creativity Statement**

*Judy Qiu, Indiana University, February 5 2015*

**Research Overview:** My research is motivated by the exciting challenges that come from the growing importance of data and computing in scientific discovery. My approach to large-scale data analysis is based on the integration of Cloud and High Performance Computing (HPC) concepts. I am still following the broad plan identified in my NSF CAREER award on **D**ata-**E**nabled **D**iscovery **E**nvironments for **S**cience and **E**ngineering (DEDESE) which includes linkage of research, education and outreach. The CAREER plan identified 7 thrust areas where I have made good progress with the 5 research activities:

1. ***Research a next generation Iterative MapReduce*** using a Map-Collective model where independent iterative maps are followed by user customizable collective operations;
2. ***Research a*** ***higher-level programming model*** that compiles to an iterative MapReduce runtime;
3. ***Research a*** ***scalable NoSQL storage model*** with compute-data affinity optimized for data processing
4. ***Research a*** ***fault tolerance model*** that supports checkpointing between iterations for robustness and individual node failure without compromising performance by configurable settings;
5. **Evaluation of my research** with real workloads, applications, and simulations.

My major technical achievements in these goals are centered on five software systems: Twister, Twister4Azure, IndexedHBase, Pig and Harp. **Twister** was the forerunner of several parallel runtime engines that extend the MapReduce concept into the realm of iterative computations, whose importance is now generally recognized across the Big Data processing landscape. The seminal paper “Twister: A Runtime for Iterative MapReduce” was published in 2010 and has been cited 507 times as of February 2015. **Twister4Azure** is a Map-Collective framework for Microsoft Azure Cloud. The open source **Harp** project builds on Twister and Twister4Azure. It supports general data abstractions with high performance collective communication, giving synchronized data movement among computer nodes. We implemented it as a library that plugs into Hadoop and enables users to run complex data analysis algorithms on both clouds and supercomputers. **IndexedHBase** is a scalable, fault-tolerant, and indexed NoSQL table storage system. It provides a scalable storage substrate for query, batch and streaming data analysis. Our research led to some unexpected discoveries. One was the identification of the importance of the HPC-ABDS (High Performance Computing - Apache Big Data Stack) software stack which is illustrated by Hadoop (with our Harp plugin) which can run Kmeans, Graph Layout, and Multidimensional Scaling algorithms with realistic application datasets over 4096 cores on the IU BigRed II Supercomputer (Cray/Gemini) and achieves linear speedup. This demonstrates the portability of HPC-ABDS to HPC and eventually Exascale systems. A second discovery was the Map-Streaming concept and our collaboration with the IU Network Science group on parallel Twitter feed clustering using Apache Storm in HPC-ABDS.

The final two CAREER goals are given below and described in the following sections.

1. ***Integrate research with education*** from undergraduate students to graduate students;
2. ***Develop new curricula*** in cloud and distributed computing and proactively support adoption of such curricula in other universities with a cloud repository and classes, workshops and tutorials.

**Future Research:** New funded projects will allow me to further develop our Map-Collective and Map-Streaming programming models with new driving applications and integration with data systems where I am starting a collaboration with IU database group. Attention to fault tolerance will be focus as well as building on our exciting results on streaming data analysis by integrating Harp with Apache Storm. We will research extending Map-Collective (Harp) research to give a high performance machine learning library.

**Collaboration and Funding:** I have been on leadership team of 10 research and 2 education funded projects since 2009 and I am PI of 6 of these. Newly funded projects include NSF *DIBBs project on middleware and high performance analytics libraries* in collaboration with Rutgers, Virginia Tech, Utah, Arizona, Kansas and StonyBrook universities; NSF *Rapid Prototyping HPC Environment for Deep Learning* in collaboration with Stanford and Tennessee; NIH *Open Resource for Collaborative Biomedical Big Data Training* in collaboration with UCSD. Other outside collaborators include Jülich, Seattle Children’s Hospital and Microsoft while I have strong collaborations at Indiana University in Network Science (Menczer), Bioinformatics (Tang) and Cheminformatics (Wild).

**Teaching:** I have introduced two new curricula at Indiana University, CSCI-649 and CSCI-534, ‘Cloud Computing for Data Intensive Sciences” and “Distributed Systems” for Computer Science and Data Science graduate students in both online and regular modes. I taught CSCI-343 “Data Structures” and CSCI-P434 “Distributed Systems”, which are undergraduate courses. These courses have been taken by over 470 PhD, Masters, and undergraduate students in Computer and Data Science. In 2013 I received an Indiana University Trustee Teaching Award. I have been an advisor to 4 Ph.D. students, one of whom graduated this year, and served on the research committee of 12 Ph.D. candidates. So far I have supervised 42 Ph.D. and Master’s students for independent study and provided 58 student recommendation letters.

My teaching was included in NSF CAREER award but additionally funded by two dedicated grants. I extended the Google Course Builder Massive Open Online Course (MOOC) framework to allow customization of other MOOC structured courses for use on campuses, in minority serving institutions, or distance education in general. This Google-funded framework was used to successfully host my B649 “Cloud Computing” online course, which is part of the new Data Science program at IU. This framework also formed the basis of a collaboration with University of San Diego in a new NIH funded project on Big Data training. We are exploring these features in our massive online course platform where instructors can conveniently assemble course materials (slides, videos and virtual machines (VMs)) and students or researchers can do customized experiments from the repository. I have developed hands-on tutorials around data and Clouds for two national workshops, [Big Data for Science](http://salsahpc.indiana.edu/tutorial/index.html) 2010 and [Cloud Computing for Science](http://sciencecloudsummer2012.tumblr.com) 2012, hosted at Indiana University as part of the Great Lakes Consortium Virtual School of Computational Science and Engineering (VSCSE) that reached over 500 participants across the nation.

**Publications:** From Google Scholar, I have 1674 citations and an hindex of 19. Since 2010 I have published 6 book chapters, 7 Journal papers, 33 conference or workshop papers, 5 editorials, 2 position papers, 3 open-source software releases and 9 technical reports. I co-edited the new book “Data Intensive Application for Cloud Computing,” which was published in 2014 and is the first comprehensive account of data intensive scientific applications on Clouds. I contributed to a 2011 textbook “Distributed and Cloud Computing”, which was one of 6 Elsevier books on the American Library Association 2012 *Choice Outstanding Academic Titles List*. I was honored by the “Many-Task Computing on Grids and Supercomputers” workshop at the SC12 supercomputing conference. My conference paper with follow up journal article was cited as paper with greatest impact in cloud area and I gave an invited keynote talk.

**Professional Activities including Diversity Outreach**: I have been part of 45 conference committees since 2011 as well as organizing several workshops and journal special issues in multicore computing, cloud computing, and data enabled science. I was the founder and the chair of ECMLS2012, ECMLS2011, ECMLS2010, Multicore2010, and Big Data for Science 2010 Workshops. The ACM/IEEE Multicore workshop was co-located with CCGrid2010 conference and the ACM Emerging Computational Methods for the Life Sciences (EMCLS 2012, 2011 and 2010) workshops were co-located with the ACM HPDC conference. I was Co-Program chair of the IEEE CloudCom2010 conference. I am on the editorial board of International Journal of Cloud Computing (IJCC). Considerable education and outreach activity has been achieved through my development of a strong minority-oriented Research Experience for Undergraduates (REU) program. For the past four years, I have hosted a total of 27 HBCU (Historically Black Colleges and Universities) REU students as part of Indiana University HBCU STEM (science, technology, engineering and math) Summer Scholar Institute. They have actively engaged in faculty-mentored research with my group on life science, polar science and cyberinfrastructure projects.