2023F_21_KUM

Research Development Fund – Fall 2023 Application

Title: Next Generation Texas Virtual Data Library (ViDaL II): A Secure & Compliant Computing and Data Infrastructure

Lead contact for RDF Application: Hye-Chung Kum
Department: Health Policy & Management (HPM)
Email address: kum@tamu.edu
Phone number: (510) 985-9306

Key Participating Units: **[SPH]** ***HPM:*** Hye-Chung Kum; Robert Ohsfeldt; Simon Haeder; Elena Andreyeva; Benjamin Ukert; Cason Schmit; ***EPIBIO:*** Eva Shipp; Qi Zheng; ***EOH:*** Itza Mendoza; ***HBEH:*** Junhyoung Kim; **[TEES/COE]** ***ISEN:*** Mark Lawley, Farzan Sasangohar; ***CSE:*** Bobak Mortazavi; Yoonsuck Choe; Jeeeun Kim; ***CVEN:*** Surya Sarat Congress; ***CRHTS*** **[CAS]** ***PSYC:*** Jessica Bernard; Sherecce Fields; ***SOCI:*** Heili Pals; Ernesto Amaral; Mary Campbell; Theresa Morris; **[TXRDC]:** Shannon Degenhart; **[HSC][COM]** ***CBG:*** Irtisha Singh; **[PHARM]:** Jongwha Chang; **[RCHI]:** Peter Yu; **[TAMIDS] [COAL]** ***BCBP:*** Paul Straight; **[Mays]** ***FINA:*** Boone Bowles; **[EHD][Huffines]*HKIN:*** Jenna Yentes; **[HAI][LAW]:** William Sage; Waren Keegan; Brian Larson; **[TTI]** Mahin Ramezani; **High Performance Research Computing (HPRC):** Honggao Liu; **Assistant Vice President and Privacy Officer:** John Pryde; **Associate Vice President & Chief Information Security Officer:** Adam Mikeal

RDF Amount Requested ($): $1,396, 029

Executive Summary

We propose the next generation Texas Virtual Data Library II (ViDaL II) infrastructure to enhance capacity for data intensive research that has legal requirements and collaboration for team science. The successful first generation ViDaL system allows faculty across units to competitively apply for grants requiring secure data environments. It also allows a space for innovative important preliminary work like PhD dissertations, Master's theses, and undergraduate research. However, it must be decommissioned before summer 2024 because the hardware is end of support life and funding is required to refresh the hardware to continue to be competitive for extramural funds in this increasingly growing field. ViDaL has become the main research hub where faculty and students can analyze sensitive data or collaborate with colleagues across campus in a comfortable easy to use Windows computing environment. It currently has over 100 users across 15 departments in 9 colleges (i.e., COE, SPH, CAS, COAL, COM, PHARM, Mayes, EHD, LAW), and 6 centers/institutes and is used in more than 20 funded projects totaling over $28M (IDC over $7.6M), 17 PhD dissertations, 4 Master's theses, and 3 undergraduate honor's theses. The ViDaL system, offers a unit membership and priority service-based free, on-site supercomputing facility with 1.5TB large memory nodes, NVIDIA V100 GPUs, and both Windows and Linux operating systems. This infrastructure is securely maintained to stay compliant with the myriad of increasing privacy and security requirements for a wide variety of research data analysis by TAMU High-Performance Research Computing (HPRC), TAMU Privacy Officer, TAMU Chief Information Security Officer, and the Population Informatics Lab. It is securely housed in the West Campus Data Center. ViDaL not only comprises the secure supercomputing environment, but also includes the policies and procedures that were established and implemented with the TAMU privacy officer to meet the approval, oversight, and training requirements when analyzing sensitive data. It meets the need of diverse research use cases that have no other good options such as projects requiring (1) safe and compliant computing (e.g., HIPAA data analysis of Medicaid claims data), (2) data intensive large memory computing (e.g., Full US census IPUMS proprietary data), (3) compute-intensive GPU computing (e.g., natural language processing of traffic notes), (4) IRB requirements for secure data management (e.g., survey data collection & analysis on sensitive topics), (5) collaborative computing environment to work across campus (e.g., analyzing telemonitoring data across faculty and students from SPH, COE, & CAS), and (6) any combination of these (e.g. brain image analysis that have legal data use restrictions that require large memory, and a team across CAS and EHD). Apart from the funds to assist with the hardware renewal, maintenance for ViDaL II has become sustainable with 10 confirmed financially contributing member units across campus (i.e., SPH, CAS, LAW, HPM, EPIBIO, EOH, TAMIDS, TXRDC, CRHTS, HAI), many priority service commitments from funded projects totaling $400K across 3 years, and the expertise and time commitments from HPRC, TAMU Privacy officer, and IT Chief Information Security Officer.