Challenges and Directions in the Infrastructure for Data-Intensive Science: A Discussion Based on CAS Data Cloud

Jianhui LI, Xiangyang HUANG, Yanfei HOU
Computer Network Information Center, Computer Network Information Center, Beijing
Email: lijh@cnic.cn, huangxy@cnic.cn, afeiisafei@cnic.cn

Science discovery has increasingly become data intensive, and it urges the infrastructure for science to be changed and optimized for adapting to this transformation. Chinese Academy of Sciences (CAS) is always promoting the infrastructure development for e-Science. Now it is building Data Cloud based on the achievements in infrastructure development got during the Eleventh Five-Years Plan of CAS(i.e., 2006-2010). CAS Data Cloud will consist of IaaS which has about 50PB data storage capability and a data-intensive computing cluster with about 10000 CPU cores, PaaS for on-line data management and publication and SaaS for data searching and on-demand service at the end of the Twelfth Five-Years Plan (i.e., 2011-2015).

Many challenges needed to cope with have been identified such as reliable and easily available storage service for big data, efficient organization and access of big data at appropriate quality, data integration and linking, interoperation between data and literature, abundant and effective data analysis and visualization tools for extracting knowledge from large amounts of data. And reasonable solutions to them are being sought and developed.

In this presentation, the Data Cloud which is being developed in CAS for the purpose of effectively supporting data-intensive science is presented. Then some of the challenges in developing the infrastructure for data-intensive science are analysed based on our above practice. Next, some of the principles and ideas for dealing with them are discussed before the presentation is concluded with a brief mention of some directions in infrastructure development.