DLM: A Common Data Process Pipeline for Large Scale Scientific Dataset based on Duckling Collaboration Platform

Kai Nan, Jianjun Yu

Computer Network Information Center, Chinese Academy of Sciences, No.4, 4th South Street, Zhongguancun, Haidian District, Beijing, China, e-mail: nankai@cnic.ac.cn, yujj@cnic.ac.cn

Nowadays how to process large scale scientific data from diverse disciplines has become the research focus in the e-Science field. The scientific datasets are from different field stations and institutes with heterogeneous formats, whereas they have the common requirements for data collecting, transferring, processing, storage and visualization though they have different purposes for data usage. We have provided an e-Science platform named Duckling for scientists' collaboration and application development. In this paper, we promote a novel architecture named DLM based on Duckling collaboration platform to balance the differences between the common data processing workflow and diverse data applications. When scientists want to generate a data application, they can generate a data processing workflow, selecting data transferring type, processing handle, visualization modes, and then they would get a new workflow for data application. The only thing one should do is to develop his own data processing handles or new visualization mode. DLM would provide the transferring tool to ftp server, website and email account with the help of wireless technologies, such as GPRS (General Packet Radio Service), TD-SCDMA (Time Division-Synchronous Code Division Multiple Access) and WIFI. When the dataset stream is transferred into the server, DLM would search for the data processing handles that the user configured previously. Finally the end user would enjoy the visualization of data with several preassembled visualization tools.