Facilitate Cross-Repository Big Data Discovery and Reuse

Zhiwu Xie
University Libraries, Virginia Tech

Introduction

Researchers have accumulated large amounts of observational, experimental, and simulation data. Much effort has been made to collect, curate, preserve, and provide open access to them, but putting the data online is only the start. Coined by Jim Gray as the fourth paradigm [1], the data-intensive science strives to uncover the hidden patterns and correlations across research topics and disciplines by aggregating and cross-interrogating these data silos. The productivity for e-Research may be much improved if we can provide the researchers with fast, easy, and cost-effective methods to discover and reuse these datasets in an ad-hoc and exploratory manner.

Different types of data and repositories exist. If the datasets are relatively small in size, it’s often more efficient to move the data dumps from various repositories to a centralized location for processing. When they grow beyond a certain limit, “bringing computations to the data” [2] makes more sense. A handful of top tier data repositories, usually disciplinary in nature and established to become major national and international data centers, start to design and build their hardware and software infrastructures along this line. The purpose is to offer not only sufficient data storage to aggregate many large datasets but also configurable and scalable processing power to cross-examine the datasets loaded into the same infrastructure. Similar trend is also popular in business data analytics, where MapReduce and other distributed and/or parallel processing models are prevalent. However, data held in these large-scale data infrastructures only constitute the tip of the data pyramid [3]. Vast majority of the datasets will continue to be deposited and maintained by the individual institutions. This is because for datasets generated within the same institute, the institutional repository (IR) offers the lowest cost for data publication and preservation. Moreover, due to its close integration with the existing campus IT infrastructure, the barrier to entry is also the lowest.

However, the institutional boundary also constitutes an obstacle for cross-repository data discovery and reuse, without which the potential values of the data are harder to realize. On one hand, due to the low cost of storage, IRs can easily grow up to hundreds of terabytes in size. They routinely hold large datasets with the single file size up to several gigabytes. But frequently moving data files of this size across institution boundaries is generally discouraged, especially over HTTP. On the other hand, institutional repositories do not typically provide elastic infrastructure for on-demand data processing. Indexing and searching aids typically only point back to the files. The limited metadata, mostly general purpose, descriptive in nature, and intended for human consumption, provide little help to differentiate a relevant dataset from the irrelevant ones. In other words, function wise these IRs have very limited support for big data discovery and reuse.

Scenarios

A research team A produced several terabytes of fluid dynamics simulation data. The data are generated against a 3D grid based on a new design, with each data point constitutes of a vector of temperature, pressure, and velocity, all of which change over time. The data are organized in tabular manner, compressed and then deposited in an open access institutional repository.

Now research team B finds the dataset potentially interesting. If there exists a vortex in a certain region, then the dataset is definitely relevant and worth investigations in more depth. But how can they know if a vortex exists or not without downloading the full dataset? Such data reductions and characterizations are not typically considered metadata and there is no established metadata infrastructure to convey such information.
Research team B decides to download the full dataset and conducts the initial investigation on the vortex characterization. They write open-source software to process the dataset, and produce a motion picture file with it. Then both research team C and D want to reuse the software to produce motion pictures on many other thermodynamics datasets. How can we streamline these practices and even automate the process?

From the above scenarios we can identify a few missing parts in the big data infrastructure.

**Machine Produced and Machine Actionable Metadata**

The useful metadata to describe big datasets will be varied extensively by their nature and utilities. While the community driven standardization can be useful to some extent, it’s important to recognize that data, unlike the textual documents and publications, can be several layers further from direct human consumptions. How to make use of and make sense of these data can never be predefined in a fixed set. In other words, the metadata infrastructure for datasets should be extensible and have good support for machine automation.

**The Cloud**

Besides “moving data to the computation” for small datasets and “bringing computations to the data ” for top tier national data center, institutional data repositories can also develop a new approach that pushes both the data snippets and computations to the computing cloud. While the cloud is still too expensive for long term data storage in comparison to locally built storage facilities, its elasticity and seemingly unlimited processing power can be used to generate useful metadata that facilitates big data discovery and reuse. Not only the metadata is shared, but also the environment, facilities, software, and algorithms to produce the metadata can be easily shared through the computing cloud.

**Metadata Registry for Big Data Discovery and Reuse**

The last piece to form a complete picture is a metadata registry where metadata producers can provide information on how the metadata is generated. The registry may also describe the software ad hardware environment used to produce the metadata, and possibly point to the cloud infrastructure or its replication to reproduce the process.

**Summary**

This paper introduces a framework to facilitate cross-repository big data discovery and reuse through flexible and reproducible metadata generation and sharing.

**References**

