Data Economies and Cultural Incentives

Daniel A. Reed
dan-reed@uiowa.edu
University of Iowa
April 2013

It has been well-documented that the volume of scientific data is growing exponentially. Large-scale instruments and scientific simulations; ubiquitous, wireless sensors that continuously capture data from the environment, flora and fauna, and critical national infrastructure; and social and economic data derived from our daily interactions constitute a data torrent of unprecedented scale. In almost every research discipline, instruments and computational models produce now more data each day than was previously created via decades of often-arduous experimentation and data capture.

Any change brings new opportunities and new challenges. Many have hailed this new world of big data and data-intensive science as the Fourth Paradigm[1], where researchers pose and answer questions from extant data, rather than capturing and analyzing data from specific experiments. Machine learning expertise and data analytics are now *de rigeuer,* for the power of big data is enabling new insights into large, complex systems and changing the sociology and culture of research laboratories.

Not all the news is happy in the world of big data. The explosive growth of scientific data is outstripping traditional approaches to data management, analysis and curation, both socially and technically. The cost of some scientific instrumentation long ago outstripped the resources of institutions and even some countries, leading to national and international consortia. Today, data volumes now exceed the capabilities of many individual research teams to host and manage, and they are increasingly challenging the capabilities and resources of research institutions and even national research agencies.

In addition, the desktop tools and techniques familiar to most researchers do not support analysis of tens to hundreds of terabytes and increasingly, petabytes of structured and unstructured data. This is especially true when teams must correlate data across disciplinary boundaries. In domains as diverse as astrophysics, biomedicine, environmental science, urban planning and public policy, the needs of disciplinary and multidisciplinary data analytics are driving development of new tools and techniques.

It is painfully clear that we need a new approach, one that can scale adaptively and dynamically from research group capabilities to institutional resources, and from there to national resources, both government supported and privately funded. Second, this approach should hide unnecessary details and allow data fusion and assimilation, while supporting the same tools and software regardless of the data's location. In short, we need a data marketplace, one that is broadly accessible and extensible, with well-defined mechanisms for resource allocation and management.

Such a marketplace would have several salutatory attributes, including:

- metadata data standards and interoperable data access protocols,

[1] The Fourth Paradigm: Data-Intensive Scientific Discovery, T. Hey, S. Tansley and K. Tolle (eds), Microsoft Research, 2009, http://research.microsoft.com/en-us/collaboration/fourthparadigm

- multiple data valuation metrics and pricing models that encompass both subsidies and cost recovery, encouraging groups to contribute data,
- "pay" by the use access model to highlight and track data preservation costs,
- differential usage rights that include commercial, non-commercial, exclusive and non-exclusive, anonymous and non-anonymous,
- triage and deletion mechanisms to determine the relative value and benefits of preserving data, for not everything can be saved, nor should it be,
- participation by private, government and non-profit entities,
- interoperable, industry-compatible hardware and software infrastructure,
- high-level services and data collection management mechanisms, and
- institutional, corporate, national and international data sites for redundancy and sharing.

There are several possible implementation mechanisms for such a market place. One could be based on a combination of (a) the open source Eucalyptus toolkit,[2] (b) a set of standard Linux virtual machines, (c) a set of data processing and analysis tools within the virtual machines, including Hadoop[3] and its in-memory variant Spark,[4], RDF[5] and Sparql, OWL2[6] and Pregel, and (d) high-level data management services as Dataverse.[7] Eucalyptus is API-compatible with Amazon Web Services (AWS) and supports Amazon Machine Images (virtual machines). An alternative would be to build on infrastructure such as Facebook's Open Compute hardware[8] or Netflix' Open Connect appliance[9] and leverage Netflix open source infrastructure[10].

This combination would allow groups to deploy local AWS compute (EC2) and storage (S3) capabilities and "cloud burst" (i.e., transfer workloads on demand) national research infrastructure or to the public cloud. A generalization could encompass Windows virtual machines, Microsoft's Azure cloud service and its data marketplace[11] as well.

In all cases, the key will be drawing lessons, experiences and infrastructure from the commercial cloud and Internet services world. There is little value in recreating research-specific versions of well-tested commercial infrastructure. Indeed, there are strong practical and economic reasons not to create new infrastructure. Rather, limited resources and attention should be focused on the differentiating, higher-level services and capabilities that would add research value to standard infrastructure and capabilities. Equally important is shifting the culture and psychology of academic research service providers to embrace and adopt new technologies, service models and approaches.

Acknowledgments

An earlier version a few of these ideas originated in discussions with Dennis Gannon and Daron Green while at Microsoft

[2] Eucalyptus, http://www.eucalyptus.com/
[3] Hadoop, http://hadoop.apache.org/
[4] Spark data analysis software, http://spark-project.org
[5] RDF, http://www.w3.org/TandS/QL/QL98/pp/rdfquery.html
[6] OWL2, http://www.w3.org/TR/owl2-primer/
[7] Dataverse, http://en.wikipedia.org/wiki/Dataverse
[8] Facebook Open Compute, http://www.opencompute.org/
[9] Netflix Open Connect https://signup.netflix.com/openconnect/hardware
[10] Netflix Open Source Center, http://netflix.github.io/#repo
[11] Windows Azure data marketplace, http://datamarket.azure.com/