Big Data, Big Deal!

Challenges and opportunities for NREN's

Harold W.A. Teunissen
SURFnet B.V.
Utrecht, The Netherlands
harold.teunissen@surfnet.nl

Peter Hinrich, PhD
SURFnet B.V.
Utrecht, The Netherlands
peter.hinrich@surfnet.nl

***Abstract*—Big data is all about comprehending the data sets. It is a term applied to data sets whose size or complexity is beyond the ability of commonly used software tools to capture, manage, and process the data within a reasonable elapsed time [1]. In this extended abstract we will presents different perspectives on how National Research and Educational Networks (NREN) can benefit and accommodate the big data revolution that goes beyond the obvious constituency of high-energy particle physics and astronomy.**

Keywords—Big data; Data Deluge; Research and Education Networks.

I. Introduction

Big data is not about size or scale of data sets, it is about the comprehension of an organization for it to deal with these data sets. A big data set is just too big to be handled and analyzed by traditional methods and protocols of an organization; a data set of 250 megabyte can already push the limits of the data capabilities because people are unable to retrieve and analyze the data using tools like spreadsheet applications or even a tablet computer. In any case, there is a limit to how much data leads to more knowledge (see figure below).

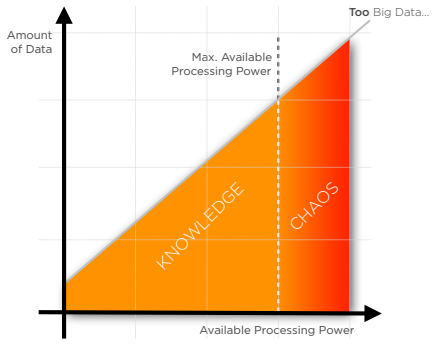

Figure 1. *Too big data*

Small and medium science is also suffering from big data problems since most of the National Research and Educational Networks (NREN) activities are focused on the on top 1% producers and consumers of scientific data, i.e. high-energy particle physics, astronomy, and in the recent years also genomics. To clarify matters, the three V of *volume*, *velocity* and *variety* [2] are commonly used to characterize different aspects of big data. Others talk about adding even a fourth *variability* [3].

II. Volume

It is evident that big data is about ever increasing volume. According to [4], in 2015 the digital universe will expand almost 6.5 times compared to 2010 as shown in figure 2.

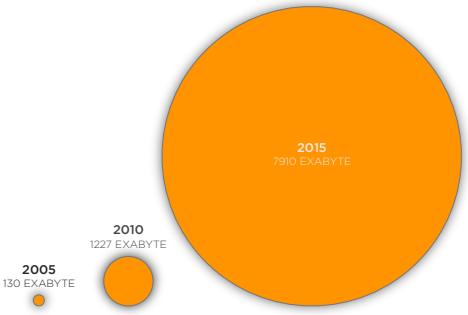

Figure 2. *Expanse of the Digital Universe*

All Internet traffic entering and leaving the European NREN's accounted for 1000 petabyte [5] or 0.3% of the total Internet traffic in 2011 [6]. However this ignores the scientific traffic on dynamic light paths connected through Netherlight and GLIF. At the moment it is difficult to estimate how much of this data is in the domain of the NREN's. Clearly the amount of data produced will continue to grow. It is expected that the Square Kilometer Array to produce a few Exabyte[1] of data per day in 2024 for a single beam per one square kilometer. After processing this data the expectation is that per year between 300 and 1500 Petabytes of data need to be stored. In comparison, the large hadron collider at CERN produces approximately 15 petabytes in 2012 [7].

Scientists and researchers are clearly overwhelmed by the sheer amount of data even if it is in the gigabyte range. The challenges are the transport, storage, providing meta-data, securing of the data. Since the data is seen as a tangible asset for (research) organizations, they become reluctant to discard it. NREN's ideally positioned to provide for big data solutions for its constituency, both nationally and internationally. Currently focused on data transport, but the next step could be the offering of cloud HPC and storage.

III. Velocity

Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, finding a super nova, or correlating

[1] One exabyte is 10^{18} or 1 000 000 000 000 000 000 bytes.

social media streams, big data must be used as it streams into the organization in order to maximize its value [8]. The data comes in large quantities, from multiple sources at a stunning rate. To collect, process, combine, correlate, and to analyze this data is a huge challenge. Furthermore, the application of the data can change over time, where new insights may lead perspectives on the data at hand. For NREN's it is essential to understand where does the data come from, what kind of data is it, where is it stored, where is analyzed, who needs access to it and for how long? NREN's typically have an overview of the *research supply chain*, so it is a small step to become big data resource broker for their constituency.

IV. Variety

Big data maybe any type of data - structured and unstructured like text, sensor data, audio, video, twitter streams, log files, etc. [8]. The strength of big data is that new insights can be found when analyzing these data types together. By 2015, organizations integrating high-value, diverse, new date types and sources into a coherent information management infrastructure will outperform their industry peers financially by more than 20% according to [9]. However for big data to become valuable a researcher first needs to find the best algorithms to combine, process and analyzes the data. MapReduce seems the de facto big data crunching method where commercial cloud providers see a steep increase in the use of these methods [10]. NREN's should focus on establishing a peering relationship with these providers.

V. Sustainability

As science is becoming more data driven, hence becomes limited by factors such as energy consumption and sutainability. Scientists, computing centers and NREN's have to join forces and develop strategies and architectures for handling of data. Distributed versus centralistic approaches have to be evaluated, and since energy consumption of data transport and optical networks is almost independent of distance [11] [12], smart solutions for storing and or processing of the data can be found by using locations where sustainable energy is abundant. Nevertheless sustainability and big data remains a controversial topic.

VI. Outline of Presentation

In the presentation we will give a short overview of big data and how it impacts SURFnet, and NREN's in general. We will discuss the challenges and opportunities and how NREN's need to collaborate to serve both its top 1% and top 20% constituency of big data producers and consumers. We will conclude with an outlook for 2020.

About the Authors

Harold W.A. Teunissen is department head of the middleware services and security group of SURFnet, responsible for leveraging the first class research network by exploiting and developing innovative services. These middleware building blocks allow the knowledge infrastructure in the Netherlands and beyond to collaborate, share and perform their research and experiments.

Peter Hinrich studied Physical Chemistry at the University of Amsterdam. After receiving his degree on the subject of laser spectroscopy of supercooled molecules in 1989 he moved to Leiden University where he received his PhD in Theoretical Chemistry in 1995. He joined SURFnet in 1996. Currently, Peter is Community Manager Science and is responsible for the communication between SURFnet and the scientific user community, informing them about possibilities new technology offers and identifying their requirements.

Acknowledgment

This project was made possible by the support of SURF, the collaborative organization for higher education institutes and research institutes aimed at breakthrough innovations in ICT.

References

[1] Wikipedia, "Big Data", http://en.wikipedia.org/wiki/Big_data, Nov. 2012.

[2] M. Slocum, "Big Data Now", O'Reilly, 2012.

[3] B. Hopkins, "Big Data, Brewer, And A Couple Of Webinars", http://blogs.forrester.com/brian_hopkins/11-08-29-big_data_brewer_and_a_couple_of_webinars, Nov. 2012.

[4] IDC, "Digital Universe Study 2011", http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf, 2011.

[5] Terena, "TERENA Compendium of National Research and Education Networks In Europe 2012", Amsterdam, 2012.

[6] Wikipedia, "Internet Traffic", http://en.wikipedia.org/wiki/Internet_traffic, Nov. 2012.

[7] Astron, "Astron & IBM Collaborate to explore orgins of the universe", http://www.astron.nl/about-astron/press-public/news/astron-and-ibm-collaborate-explore-origins-universe/astron-and-ibm-co, Nov. 2012.

[8] IBM, "What is Big Data", http://www-01.ibm.com/software/data/bigdata/, Nov. 2012.

[9] Gartner, "Information Mgmt in the 21st Century", Sep. 2011.

[10] A. Jassy, keynote presentation, AWS re:Invent Developer Conference 2012, Nov. 2012.

[11] L. M. Hilty, "Green ICT and Sustainability: A Critical Perspective", EU-RESPONDER Workshop. Vienna University of Economics and Business. Vienna, Austria, May 30, 2012.

[12] J. Baliga et al., "Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport", Proc. of IEEE, Vol 99, No. 1, Jan. 2011.