Contribution ID: **148** Type: **not specified**

Data-intensive Processing with Hadoop and friends at BiG Grid

Tuesday, 27 March 2012 15:00 (30 minutes)

Description of the Work

Hadoop is a framework built to support large-scale data-intensive processing. It is especially good at what experts call data-parallel processing, of which the most important property states that if an operation on a dataset with size n takes an amount of time t, that same operation on a dataset of size x*n will take an amount of time x*t. Because it is optimized for processing truly large amounts of data, starting at the tens of gigabytes and up to petabytes or more, it is rather complementary to than competing with existing platforms such as grids, clouds, and supercomputers.

At SARA, the Dutch center for High Performance Scientific Computing, networking, visualization, and eScience, we have piloted Hadoop since mid 2009. We have operated a small prototype cluster throughout 2010, providing six Sun Thumpers with 24 cores and 132TB of disk space. The system has been used by close to 30 users from various disciplines, ranging from the Natural Sciences (Bioinformatics) and Computer Science (Database research, Data mining), to the Social Sciences (Computational Linguistics, Information Retrieval, Biodiversity Science, and Econometry).

BiG Grid, the Dutch NGI, has invested in the operationalization of SARA's initiative. A new system is being installed at SARA, providing Dutch scientists with more than 500 cores of processing capacity, and more than 500 terabytes of data storage. The interest that several communities are showing is very promising.

We will present the package of services we will provide, focusing on i) typical use cases, ii) the user-facing services, and iii) the extensibility of the services.

Conclusions

Hadoop is a game changer. Not in the sense that it makes other systems redundant, but especially the way it handles the classes of problems it was built for - it provides large-scale processing for the masses.

At SARA we are interested in sharing our experiences with other NGIs, and partnering up to take our initiative across borders. We think communities can benefit enormously from more knowledge on this subject throughout Europe –Hadoop fills up a niche that a number of sciences up and coming in the field of large-data processing are in.

Impact

Scientific instruments are producing massive amounts of data, a relatively new trend. For some of these instruments, the brightest minds of the past decade have set up computing and data infrastructures; the Large Hadron Collider being the most spectacular example, supported by its world-wide grid infrastructure. But also others, like the operators of the radio telescope LOFAR, have set up infrastructures on larger scales than ever before.

For some sciences scale is still a struggle, albeit to different degrees. Bioinformatics has learned a lot over the past years, but is having a hard time keeping up with the amazing speed of developments in sequencing technology (amounts of data growing faster than Moore's law!). Ecoinformatics has been established as a field by now, often with sensor networks or other data producers at its heart, but is far from ready to handle the the amounts of data necessary to work on the spatial and temporal scales that ecological researchers aspire to.

Hadoop runs on commodity hardware and is being offered out-of-the-box by utility computing providers. It offers an easy to learn programming model as well as a number of interpreted languages. It is already the most widely adopted processing framework in the world –in spite of its mere four years of existence. In other words: Hadoop enables large-scale processing for the masses.

Additionally, Hadoop's impressively large community has built a number of tools around the framework to support more sophisticated types of processing, ranging from column stores, a data warehouse supporting SQL-like syntax, and an interpreter for a data-flow language, to implementations of common machine-learning algorithms and a graph processing framework.

In The Netherlands we see that because of its accessibility, its ease of use, and its rich suite of tools, Hadoop has the potential to transform the way science is done in fields that are struggling with large-data problems.

URL

http://www.sara.nl/project/hadoop

Overview (For the conference guide)

Hadoop is a framework built to support large-scale data-intensive processing. In spite of its age –it is only four years old –it has gained popularity amazingly fast and is currently deployed in both science and industry at unprecedented scales. Hadoop is an open source implementation of a system designed by Google, including the architecture of the distributed file system in use at the company, and the parallel processing model used by its in-house scientists.

In this talk we present the current Hadoop pilot service of SARA in The Netherlands, and the way we will operationalize it within the Dutch NGI, BiG Grid. We will talk about typical work being done on a Hadoop system, sciences that benefit from such a system, the types of tools it provides, and how it enables eScience. The talk is intended for both community representatives and NGI representatives. Our goal is to show communities what can be achieved, and to inspire NGI's in their approach to diversity of infrastructure.

Primary author: Mr LAMMERTS, Evert (SARA)

Presenter: Mr LAMMERTS, Evert (SARA)

Session Classification: Community-tailored Services