The SIKS/BiGGrid Big Data Tutorial

Djoerd Hiemstra
U. Twente, The Netherlands
d.hiemstra@utwente.nl

Evert Lammerts
SARA, The Netherlands
evert.lammerts@sara.nl

Arjen P. de Vries
CWI, The Netherlands
arjen@acm.org

1. INTRODUCTION

The School for Information and Knowledge Systems SIKS and the Dutch e-science grid BiG Grid organized a new two-day tutorial on *Big Data* at the University of Twente on 30 November and 1 December 2011, just preceding the Dutch-Belgian Database Day. The tutorial is on top of some exciting new developments in large-scale data processing and data centers, initiated by Google, and followed by many others such as Yahoo, Amazon, Microsoft, and Facebook. The course teaches how to process terabytes of data on large clusters, and discusses several core computer science topics adapted for big data, such as new file systems (Google File System and Hadoop FS), new programming paradigms (MapReduce), new programming languages and query languages (Sawzall, Pig Latin), and new 'noSQL' databases (BigTable, Cassandra and Dynamo).

2. MAPREDUCE

The tutorial follows the University of Twente master course "Distributed Data Processing using MapReduce", that was given for the first time from November 2009 to March 2010. In the course, students performed retrieval experiments on the English part of ClueWeb09 dataset [1]. They implemented a full experimental retrieval system with little effort using Hadoop MapReduce. The system analyzes more than 12TB of data in reasonable time using a cluster of 15 low-cost machines . The code used in the experiments is open source and available to other researchers [2].

3. SIKS AND BIG GRID

The tutorial was sponsored by SIKS and BiG Grid, with the aim to bring together SIKS researchers and Ph.D. students that need to analyse large datasets. BiG Grid provides access to the Dutch e-science grid to enable researchers to analyze datasets that are beyond the scope of a single research group or university. Advanced ICT research infrastructures are crucial to scientific research communities. The BiG Grid project (led by partners NCF, Nikhef and NBIC) aims to set up a grid infrastructure for scientific research. This research infrastructure contains compute clusters, data storage, combined with specific middleware and software to enable research which needs more than just raw computing power or data storage [4].

4. KEYNOTE LECTURES BY JIMMY LIN

We are proud to have Jimmy Lin as our keynote lecturer. Lin, who holds a PhD from MIT, is associate professor in the iSchool at the University of Maryland. He also has appointments in the Institute for Advanced Computer Studies (UMIACS) and the Department of Computer Science at Maryland. Lin works at the intersection of natural language processing (NLP) and information retrieval (IR), with a recent emphasis on scalable algorithm design and large-data issues, and recently published a book about this work with Chris Dyer [3]. He directs the recently-formed Cloud Computing Center, an interdisciplinary group which explores the many aspects of cloud computing as it impacts technology, people, and society. He is also a member of both the Computational Linguistics and Information Processing Lab and the Human-Computer Interaction Lab. Lin worked on Cloudera, which aims to bring Hadoop MapReduce to the enterprise, and is currently spending a sabbatical at Twitter.

5. HANDS-ON EXPERIENCE

A major part of the tutorial consists of hand-on experience. Students will solve real large-scale data analysis problems of their choice on a cluster of machines. Students will get access to the Big Grid Hadoop test cluster, providing 20 cores for MapReduce and 100TB diskspace for HDFS. Students are encouraged to bring their own data, and present their results at the end of the second day. The organization will provide several public datasets, such as Wikipedia, the ENRON dataset, White House visitor records, Genome data, the ClueWeb09 web crawl, and more. At the Dutch Belgian Database Day, we present some of the problems that the course participants tried to solve, and the results that they achieved.

6. REFERENCES

[1] Jamie Callan. The ClueWeb09 dataset. http://boston.lti.cs.cmu.edu/clueweb09/

[2] Djoerd Hiemstra and Claudia Hauff. MapReduce for information retrieval evaluation: "Let's quickly test this on 12 TB of data". In: *Lecture Notes in Computer Science 6360. Springer Verlag.* pages 64-69, September 2010. http://mirex.sourceforge.net/

[3] Jimmy Lin and Chris Dyer. Data-Intensive Text Processing with MapReduce. *Morgan & Claypool*, 2010. http://dx.doi.org/10.2200/S00274ED1V01Y201006HLT007

[4] Arjen van Rijn and Maurice Bouwhuis. The BiG Idea. In *EU Projects Magazine*, Insight Publishers, August 2010. http://www.biggrid.nl/fileadmin/images/BiG_Grid_editorial_final_20100723.pdf