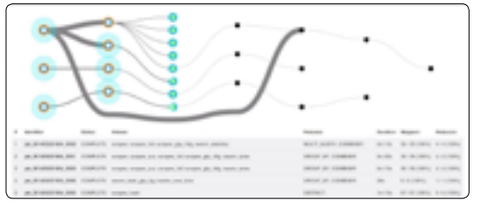
19. BTWorld: A Large-scale Experiment in Time-Based Analytics



These days, large amounts of data are collected about the operation of many important systems, for instance, traffic systems and the financial system. Extracting meaningful information is very challenging: big data must be processed in time and without error.

At TU Delft, for the last four years, we have been collecting data about BitTorrent, a system used by hundreds of millions of people worldwide for sharing videos and other files. For example, musicians use it for the distribution of their work and software developers for the distribution of open source software.

Looking at the collected data from all the BitTorrent servers in the world, we can understand BitTorrent. To do so, we have created a workflow of big data queries that give answers to such questions as “How many videos are shared?” or “What is the location of the most used servers?” As our key innovation, we have designed an efficient iterative method to optimize big data workflows.

ICT science question

Despite a large number of empirical and theoretical studies, observing the state of the global information networks remains a grand challenge. The main question we set out to answer was how to reliably analyze large scale time based datasets through different types of queries.

Other questions we address are: What are the programming models to be used? Only the MapReduce programming model or more? To what dataset sizes we can push our analysis? Whereas in 2013 we could only process 100 GB, we can now process 1.5 TB, and we are working on processing all of the 15 TB we collected. With that size, we are among the largest publicly reported Big Data experiments.

Application

We apply our network analysis to the global BitTorrent network. We have a unique 15 TB dataset

obtained from monitoring BitTorrent, which we analyze in one of the largest publicly reported Big Data experiments. This leads to both the design of a workflow of queries with wide applicability, and an understanding of the monitored system. The dataset has been obtained by us by periodically contacting many so called BitTorrent servers worldwide over a period of more than four years.

Alternative Application

Our design and experimentation with the BTWorld workflow is applicable to many other time-based datasets obtained from monitoring large scale distributed systems, e.g., financial systems, traffic systems and sensor networks.

Quote

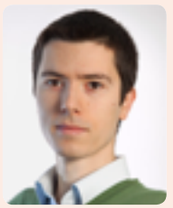
“This is a great example of a well-researched and well-engineered real use case of Big Data processing.” - Douglas Thain, chairman of the jury for the SCALE Challenge of the IEEE/ACM CCGrid 2014 conference in May 2014. We won this Challenge with BTWorld.

Productivity is increasingly associated with big data. Our innovation allows companies to pursue ambitious big data projects with complex workflows. E-governance processes may benefit to the same degree.

We enable SMEs and research labs with little technical expertise to process big data in innovative and creative workflows, helping to overcome the data deluge they face.

By 2020, companies will have access to over 40 ZB of data per year. It constitutes as much a business opportunity as it is a technical challenge. Our TRL 4 technology offers a critical advantage.

How to enable innovative but non-IT research labs and SMEs to process data with large volume, high velocity, and significant variety? We offer an efficient, iterative, flexible method.



Alexandru Iosup
A.Iosup@tudelft.nl
Mihai Capotă
mihai@mihaic.ro
www.pds.ewi.tudelft.nl, www.btworld.nl.

COMMIT/ project
IV-e e-Infrastructure Virtualization for e-Science Applications

This research was supported by the Dutch national program COMMIT/