GATECloud.net: Cloud Infrastructure for Large-Scale, Open-Source Text Processing

Valentin Tablan Ian Roberts Hamish Cunningham
Kalina Bontcheva
University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello, S1 4DP, Sheffield, UK
[v.tablan, i.roberts, k.bontcheva, h.cunningham]@dcs.shef.ac.uk

1 Introduction

GATE[1] is a world-leading R&D platform for text processing, with over 37,000 downloads in the last 12 months, regular training events attended by researchers and companies, and widely cited in the scientific literature[2]. From its inception, over 15 years ago, one of its stated goals has been that of providing infrastructural support to natural language researchers, allowing them to focus on developing new algorithms and running new experiments, while GATE solves the software engineering issues, such as text formats, data persistence, etc.

The growth of unstructured content on the Internet has resulted in an increased need for researchers in diverse fields (e.g. humanities, social sciences, bioinformatics) to run language processing and text mining on large-scale datasets, many of which are impossible to process in reasonable time on standard desktops. An additional impetus is the availability of key datasets, e.g., Wikipedia and Freebase snapshots on Amazon S3 (500MB+). This growth in data size can be mitigated by the availability of cloud computing infrastructure which allows the deployment of significant compute power on an as-needed basis, and with no upfront costs. However, the types of systems used by language processing researchers tend to be fairly complex, which makes deployment on a cloud platform not a trivial matter.

To answer these challenges, the GATE family is being expanded with a new member: GATECloud.net - a service that deploys the GATE infrastructure (including its JAPE rule-based and machine learning engines) on the Amazon Elastic Cloud Computing platform. This enables researchers to carry out large-scale NLP experiments by harnessing on-demand compute power on the cloud. It also eliminates the overhead in implementing dedicated parallel text processing algorithms - standard GATE processing pipelines can be used unmodified and deployed in a parallel manner to process large datasets.

2 GATECloud.net

GATECloud.net is a new service that deploys GATE server software and analysis pipelines on cloud computing infrastructure. The intention behind it is to bring the advantages typically associated with cloud computing within the reach of text processing researchers and users. At the time of writing, there are two types of functionality offered: on-demand servers pre-installed with specialised software, and the execution of text analysis pipelines over large document collections, using the massive parallel compute infrastructure offered by the cloud.

On-demand Servers

The GATE family of software tools includes server-side platforms, such as GATE Teamware[3], and GATE Mímir[4], both having complex architectures and sets of dependencies that make them difficult to install and maintain. They are also the type of tools that may only be relevant during certain stages of a typical text mining project. This makes them suitable for cloud deployments, where costs are reduced both in term of admin staff time, and the purchase of server hardware that is not required on a permanent basis.

GATECloud.net offers pre-installed Teamware and Mímir servers, running as Amazon EC2 instances, based on 64 bits Ubuntu Linux, customised with the necessary additional software.

The process of reserving a new server is illustrated in

[1] http://gate.ac.uk

[2] Six key GATE papers receive over 1,700 citations on Google Scholar.

[3] GATE Teamware (http://gate.ac.uk/teamware/) is a collaborative annotation system.

[4] GATE Mímir (http://gate.ac.uk/family/mimir.html) is a multi-paradigm indexing server.

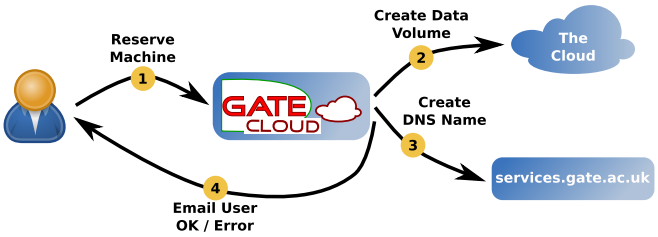

Figure 1: Reserving a Server Workflow

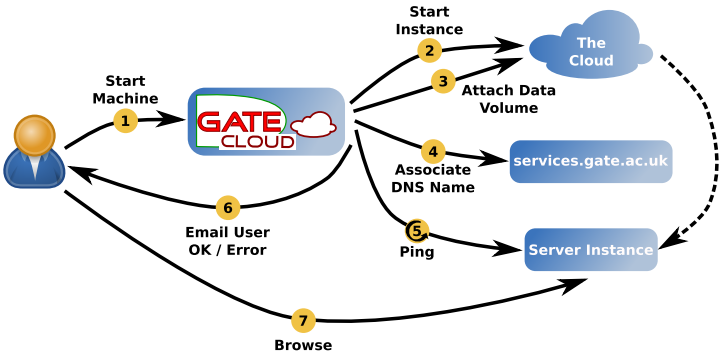

Figure 2: Starting a Server Workflow

Figure 1 and comprises four simple steps: the user requests a new reservation through the GATECloud.net web site (1), a persistent data volume is created and associated with the user's account (2), a persistent server name is reserved for the user (3), and the user is notified of success or otherwise (4).

Once reserved, a server can be started using the worflow in Figure 2, where following the user's start request (1), a new instance is started (2), the previously created data volume is attached to the new instance (3), the reserved server name is associated with the IP address of new instance (4), and, as soon as the start-up process is confirmed to have completed successfully (5), the user is notified (6). Once running, a server is fully under the control of the user, who can access its web interface, which includes support for administrative actions, such as adding user accounts and setting up security options.

On-demand Large-Scale Processing

The other half of the GATECloud.net infrastructure is the support for on-demand processing of document collections using cloud computing.

Running large text processing jobs is a task that is far from trivial. The most obvious barrier is the amount of compute power required, or, alternatively, a very large amount of time. Less obvious, but equally challenging, are tasks such as controlling the workflow that distributes the workload across several computers, dealing with errors, recovering from exceptional conditions, making the most of the available CPU power, and dealing with various input and output formats. Another difficulty is supplying sufficient storage for the input document collection, the output files, and all the temporary files – which are short-lived but can be orders of magnitude larger.

The support for *Annotation Jobs* on GATECloud.net aims to address most of these engineering issues, leaving researchers free to concentrate on their experiments. The steps for document processing on the cloud are shown in Figure 3, where three parallel workflows are present. Activities that result from interaction with the system user are on the left side of the picture, using a yellow background. The whole process starts with the user uploading the annotation pipeline and the input document collection to the storage server (1), followed by them starting the annotation job (2). This causes the GATECloud.net system to create a workflow for the new job (3), consisting of tasks that get queued for execution. As tasks are consumed, new ones are created as defined by the workflow, until all the tasks for the given job are completed. When that happens, the user is notified (4) and they can download the execution reports and annotated documents from the storage server (5).

In parallel, the GATECloud system manages groups of server nodes (referred to as *swarms*) whose job is to execute tasks from the task list. The system makes sure that the required swarm is active whenever there are tasks queued for it.

On the side of the swarms, the workflow (represented

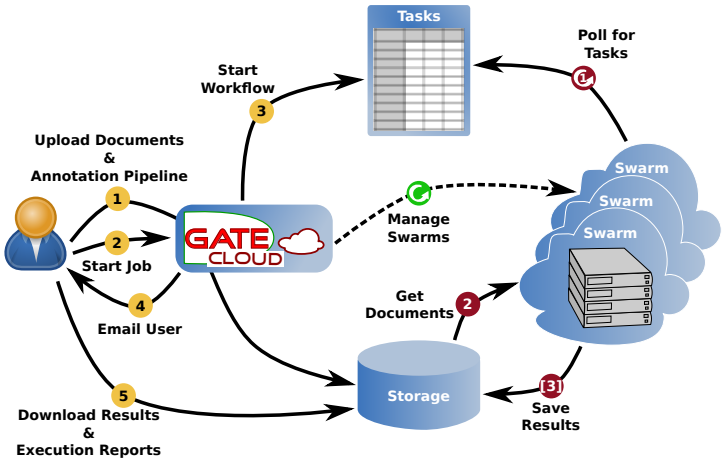

Figure 3: On-demand Document Processing Workflow

by the steps on the right side of the image, using a red background) consists of picking up tasks from the appropriate queue, executing them, saving the results produced (if any), and updating the task state in the list with the corresponding exit code.

Research Experiments

In order to quantify the performance gains from running semantic annotation experiments on large datasets on GATECloud.net, we carried out experiments on three kinds of document collections: 50 million tweets (many, very short text), 20,000 news articles (medium-sized texts, targeted by many existing semantic annotation services), and 100,000 patents (large documents, some of up to 6MB).

The news and Twitter datasets were annotated for named entities with the standard GATE components, whereas for the patents dataset we reused a pre-existing semantic annotation pipeline [Agatonovic *et al.* 08] which recognises patent-specific annotation types.

In order to carry out a cost-benefit evaluation of GATECloud.net, we ran three sets of experiments, processing the above mentioned patents, news, and Twitter datasets. Each dataset was processed on a desktop[5], a server, and the GATECloud.net system. The resulting figures are presented in Figure 4.

		Time (hh:mm:ss)			Speed (Kb/s)	Cost (GBP)	
		CPU Time	Computer Time	Clock Time		£/GB	Total
Experiment 1: Patents	Desktop	N/A	N/A	N/A			
100,000 patent documents	Server	91:42:00	18:39:54	18:39:54	85.33		
	Cloud	162:38:00	16:56:37	02:03:32	773.6	£3.08	£16.83
Experiment 2: News	Desktop	05:20:19	05:20:19	05:20:19	71.52		
20,000 documents	Server	04:43:00	03:08:00	03:08:00	121.86		
	Cloud	07:47:00	01:21:20	00:35:31	645.04	£1.51	£1.98
Experiment 3: Tweets	Desktop	32:28:46	32:28:46	32:28:46	52.80		
50,000,000 tweets	Server	22:16:15	03:19:12	03:19:12	516.53		
	Cloud	40:08:00	07:00:14	01:25:46	1199.69	£1.35	£7.92

Figure 4: Experiment Results

3 Conclusion

GATECloud.net empowers researchers from diverse fields to run experiments on large datasets, at an affordable cost, and without requiring expensive in-house compute infrastructure and system administration personnel. This cloud-based text processing infrastructure makes it possible to obtain a fully-configured web-based annotation tool and/or scalable semantic search index in a matter of minutes. The costs incurred only cover the actual number of compute hours used, thus saving researchers money by alleviating the need to purchase and maintain hardware, which will be utilised only occasionally.

4 Acknowledgements

The work on the GATECloud.net platform has been partially funded by a EPSRC/JISC grant supporting pilot projects in Cloud Computing.

References

[Agatonovic *et al.* 08]
M. Agatonovic, N. Aswani, K. Bontcheva, H. Cunningham, T. Heitz, Y. Li, I. Roberts, and V. Tablan. Large-scale, parallel automatic patent annotation. In *Proc. of 1st International CIKM Workshop on Patent Information Retrieval - PaIR'08*, Napa Valley, California, USA, October 30 2008.

[5] with the exception of the patents dataset which was too large for processing on a desktop machine.