Scaling NLP Algorithms to Meet High Demand

Connor Stokes, Anoop Kumar, Frederick Choi, Ralph Weischedel
Raytheon BBN Technologies Corp.
10 Moulton St.
Cambridge, MA, USA
cstokes@bbn.com, akumar@bbn.com, fchoi@bbn.com, weischedel@bbn.com

I. Abstract

The growth of digital information and the richness of data shared online make it increasingly valuable to be able to process large amounts of data at a very high throughput rate. At the same time, rising interest in natural language processing (NLP) has resulted in the development of a great number of algorithms designed to perform a variety of NLP tasks. There is a need for frameworks that enable multiple users and applications to run individual or a combination of NLP algorithms to derive relevant information from data [1]. In this work, we take multiple NLP algorithms that adhere to the ADEPT framework and deploy them on distributed processing architectures to satisfy the dual needs of serving a large user group and meeting high throughput standards, while reducing the time from lab to production environment.

The ADEPT framework provides a set of uniform APIs for interacting with a diverse set of NLP algorithms by defining a set of data structures for representing NLP concepts [2]. It offers multiple access points for interacting with these algorithms; a REST API, a serialized Data API, and processor components that can be used in a larger pipeline. The comprehensive ADEPT architecture can support algorithms that perform sentence-level, document-level, or corpus-level text processing, allowing a wide range of NLP algorithms to make use of the framework. ADEPT interfaces allow parallelization to occur at an optimum level for each algorithm.

Amazon Web Services (AWS) consists of a stack of technologies commonly used in the commercial sphere to host web applications designed to scale rapidly with a growing user base. The Amazon Elastic Compute Cloud (EC2) and its auto-scaling feature in particular provide a means of reliably and efficiently scaling a service to meet traffic demands. Hadoop and Spark are top level Apache projects designed to enable massive parallelization of data processing. Hadoop employs the MapReduce programming model and uses a distributed file system to store data. It is a widely used processing framework with proven potential for very high throughput. Spark is a distributed processing framework that makes use of in-memory primitives, enabling a significant performance advantage over Hadoop in certain uses at the cost of higher memory requirements [3]. Spark processes are able but not required to fit into the MapReduce model, allowing algorithms to be adapted for use in a Spark context with minimal effort.

To handle high volume of concurrent requests of DEFT algorithms, we created a mechanism to deploy the algorithms on an AWS stack [4]. We extended the ADEPT framework to create installers to deploy algorithms on an AWS EC2 node. We preserve the EC2 instance along with the algorithm and grow or shrink the number of instances to accommodate the request volume.

In summary, we demonstrate that NLP algorithms can be rapidly scaled by leveraging the ADEPT framework, parallelization models, and virtualization technologies to meet the growing demands of high volume and throughput. The ADEPT framework allows managing a diverse set of NLP algorithms. We adapt the NLP algorithms to fit into Hadoop and Spark architectures in an effort to maximize their throughput. We explore the viability of using Amazon EC2 to meet and rapidly scale based on usage demands.

II. Acknowledgement

This paper is based upon work supported by the DARPA DEFT Program. The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This research was developed with funding from the Defense Advanced Research Projects Agency. Approved for Public Release, Distribution Unlimited.

III. References

[1] Boschee, Elizabeth, et al. "Researching persons & organizations: AWAKE: From text to an entity-centric knowledge base." Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014

[2] https://github.com/BBN-E/Adept

[3] Zaharia, Matei, et al. "Spark: cluster computing with working sets." Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. Vol. 10. 2010.

[4] DEFT: http://www.darpa.mil/program/deep-exploration-and-filtering-of-text