Archiving and Analyzing Tweets and Webpages with the DLRL Hadoop Cluster

Sunshin Lee
Dept. of Computer Science, Virginia Tech
Blacksburg, VA 24061 USA
sslee777@vt.edu

Edward A. Fox
Dept. of Computer Science, Virginia Tech
Blacksburg, VA 24061 USA
fox@vt.edu

ABSTRACT

In the Integrated Digital Event Archive and Library (IDEAL) [1] project we research the next generation integration of digital libraries and event archiving. The project team has been collecting Internet information such as tweets and webpages related to crises or tragedies in addition to recovery and government/community events. This poster is about the Hadoop cluster in the Digital Library Research Laboratory (DLRL) of the Department of Computer Science, Virginia Tech, along with its use in archiving and analyzing tweets and webpages.

1. HADOOP CLUSTER

To archive and analyze tweets and webpages, we built a 20 node Hadoop cluster (20 x i5 quad core CPUs, 704GB RAM, 160TB HDDs), which is good for archiving and analyzing bigdata.

2. ARCHIVING TWEETS AND WEBPAGES

We have archived 1.1 billion tweets and roughly 11TB of webpages. We used yourTwapperKeeper (yTK)[1] and DMI-TCAT[2] to collect tweets. Table 1 shows the list of tweet collections. We also used a focused crawler and Internet Archive (IA) services to collect webpages. Archiving efforts are expanding as interested researchers launch diverse types of collaborations.

Table 1. List of tweet collections [1]

Project	Tools	# of tweets	Started at
IDEAL	yTK	1,082,970,034	2012
GETAR	yTK	37,984,627	2015
IDEAL	DMI-TCAT	53,693,045	2015

We archived our data in HDFS and HBase. To store both data and schema in HDFS, we used the AVRO file format. Sqoop is used for moving data from the yTK DB to HDFS and a Pig script is used for loading data into HBase from HDFS.

3. ANALYZING AND VISUALIZING USING HADOOP

To analyze bigdata, we use Hadoop tools including Mahout and Spark for machine learning, Solr for indexing and supporting a search interface, and both the Natural Language Toolkit (NLTK)[3] and Stanford NER for natural language processing.

Figure 1 shows the system architecture for the IDEAL project. The left summarizes data sources, the center shows the Hadoop-based processing framework, and the right shows (user) services.

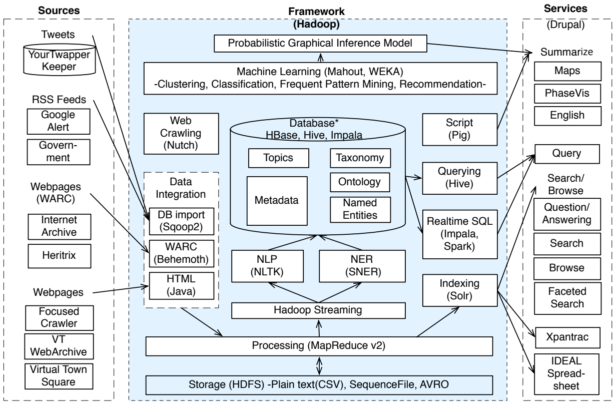
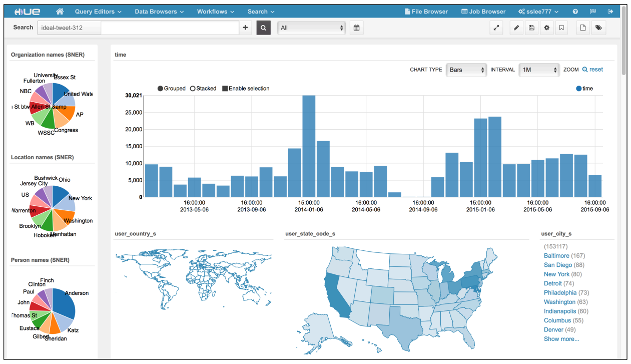


Figure 1. System Architecture for IDEAL project

We support searching, browsing, summarizing, analyzing, and visualizing services through the HUE[7] interface. Figure 2 shows the Hue interactive dashboard applied to analyzing and visualizing tweets mentioning "water main break" [2, 3].



Search: ideal-tweet-312 · All

CHART TYPE: Bars · INTERVAL: 1M

(153117)
Baltimore (167)
San Diego (88)
New York (80)
Detroit (74)
Philadelphia (73)
Washington (63)
Indianapolis (60)
Columbus (55)
Denver (49)
Show more...

Figure 2. Analyzing / Visualizing Water Main Break Tweets

4. ACKNOWLEDGEMENTS

This material is based upon work supported by the US National Science Foundation under Grant No. IIS-1319578.

5. REFERENCES

[1] Integrated Digital Event Archive and Library (IDEAL), http://eventsarchive.org

[2] S. Lee, N. Elsherbiny, and E. A. Fox, "A digital library for water main break identification and visualization," JCDL '12: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, Washington, D.C., USA, 2012, pp. 335–336.

[3] S. Lee, M. Farag, T. Kanan, and E. A. Fox, "Read between the lines: A Machine Learning Approach for Disambiguating the Geo-location of Tweets," JCDL '15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, Knoxville, TN, USA 2015, pp. 273–274.

[1] https://github.com/540co/yourTwapperKeeper

[2] https://github.com/digitalmethodsinitiative/dmi-tcat

[3] http://www.nltk.org, http://gethue.com