USING APACHE SPARK IN THE ANALYSIS OF WIKIPEDIA PAGE VIEW DATA IN CORRELATION TO THE REAL WORLD EVENTS

Sebastian Cousins, Debzani Deb. Department of Computer Science; scousins116@rams.wssu, debd@wssu.edu

Apache Spark has emerged as the next generation big data processing engine, and businesses are adopting it at lightning speed. However, as a new framework, there are not enough resources available that explicitly illustrates the process of analyzing and querying a NOSQL (schema-less) big data store using Spark and visualizing the results in an efficient way. The objective of this study is to gain knowledge in the use and application of Spark and other newer big data tools and to identify a process of analyzing, querying and visualizing big data in an effective way. More specifically, we utilized tools such as Apache Spark, Spark SQL, Scala Programming Framework, and Apache Zeppelin to gain insight on how strongly Wikipedia page view data correlates with real-world trending topics. Wikipedia page view information is available as a collection of space delimited text files, where each file corresponds to a single hour's worth of page views. We aggregated these files to analyze recent four weeks' page view data (≈ 40GB) and performed preprocessing and combining by utilizing a Scala program in Apache Framework. We then utilized Spark SQL to run various queries such as ten most trending topics in each week, number of WSSU related Wiki search, change of trending topics in a week by week basis etc. We then linked our queries to Apache Zeppelin framework that allows us to instantly visualize the query results. The resultant visualizations reveal strong correlation between Wikipedia page search and the real-world trending topics. We also found Spark to be dramatically faster than Apache Hadoop and Hive when it comes to big data analytics.

Funding provided in part by NSF Award #1600864, WSSU RIP Award 2016-2017