CSSE 490 Research 1: Data processing pipelines using Falcon

Zhihao Xue, Tianjiao Mo, (Jackie) Xiangqing Zhang

Apache Falcon is a framework to simplify data pipelining and management on Hadoop clusters. In general, it eases the pain of create new workflows and/or pipelines, supports large data handling process, and provides feedback from pipelines. Since our project is about retrieving website crawl data and real-time web content from the internet, it well fits in the Falcon workflow. In fact, one of the Google Senior Developers in Google Ads team using Hadoop suggests us to “give Falcon a try.”

Here is the process of our data flow. First, raw HTML data will come in from the internet. Then, we will process the raw data and create the raw pages’ dataset. The raw data will be sorted on the cluster and page content cleanse process will in turn be performed. Clean web data that only contains valid links and other information will be generated by the cleanse process, and eventually MapReduce tasks will be performed on.