Wyatt Blair

EMT-678-WS: Big Data Technologies

9/16/2024

**Class 1-- Scaling Big Data Mining Infrastructure Report**

This paper, written by Jimmy Lin and Dmitriy Ryaboy, lays out several of the lessons learned in the development of the social media platform Twitter (now known as X). Social media platforms often deal with a great deal of data flowing into the company through many different services at once. As such, the company must make sure it can collect, aggregate, and analyze this data at a rate sufficient to keep up with the amount flowing in.

Throughout the paper, the authors use an analogy to plumbing which is helpful when attempting to picture the scale of the data science these engineers are working on. If one were to imagine the phone or computer with which they access Twitter as a sink, then the words they publish to the platform can be thought of as water. The amount of water which goes down the sink drain is limited by the size of the pipe and the flow of the entire apartment buildings’ is limited by the size of the waterways which connect it to the larger municipal sewer system, and so on. By leveraging open-source services like Hadoop and through careful construction of these pipelines, Twitter has managed to squeeze every bit of value it can out of the flow of data into its warehouse.

Along with many organizational tips related to company structure, database best-practices such as not using SQL or JSON, the paper discusses a number of interesting models and data-visualization tools which can be situated on some portion of this pipeline. In so doing, these ML models and associated charts can be updated in real time as new data flows into the company. All of these decisions have helped Twitter scale at remarkable speeds and maintain their composure as 12 TB of data streams into the company every day.