**Scaling Big Data Mining Infrastructure: The Twitter Experience**¹

This paper provides knowledge of big data mining infrastructure that Jimmy Lin, an Assistant Professor who spent an extended sabbatical from 2010 to 2012 at Twitter, and Dmitriy Ryaboy, who was a tech lead first then an engineering manager of infrastructure at Twitter, wrote from their Twitter experience. The two topics described were schemas and heterogeneity. Firstly, schemas alone are insufficient for data scientists to get an overall understanding of the data. Scribe is a log-transport mechanism that Twitter uses for “aggregating high volumes of streaming log data in a robust, fault-tolerant, distributed manner”¹. Another major challenge was the heterogeneity of the various components that must be integrated for a production workflow. As Jimmy Lin mentions in a talk², data cleaning and data munching takes 80% of the time, and since that takes data scientists’ majority of their time, he shares his experience for future data scientists. He says, “Schemas aren’t enough! We need a data discovery service!” Finding data easier, rather than through a complex path, is done by Data Access Layer (DAL), a loader that knows via metadata how to access it. Then, the heterogonous components are synchronized by “plumbing” so the data runs smoothly. The components are wired together via different channels, and the quality must be good to prevent clogging. A successful big data analytics platform is achieved by balancing speed, efficiency, flexibility, scalability, robustness, etc.

[1] Scaling Big Data Mining Infrastructure: The Twitter Experience - Jimmy Lin and Dmitriy Ryaboy; December 2012

[2] Jimmy Lin’s presentation of this paper:  
<https://www.youtube.com/watch?v=T5ZjSFnOxys>