Scaling Big Data Mining Infrastructure: The Twitter Experience

Data mining tasks are often complex and sophisticated. The authors of the paper suggest the best practices for data mining tasks .These practices are explained through their experience in scaling Twitter’s analytics infrastructure. Additionally, the evolution of big data infrastructure and the development of data mining on big data are being discussed thoroughly in the paper. Analytics at Twitter lies at the intersection of data explosion, online analytical processing and open source ecosystem that range widely from simple aggregations to training machine-learned models which makes it a perfect case study. This complex system is supported by Hadoop open source implementation of MapReduce along with systems such as HBase, ZooKeeper, Pig, Hive, and many others.

The authors focus on two challenges faced while building a data analytics platform: data mining and plumbing. They believe successful data mining requires thorough data preparation before any algorithms are applied. Plumbing refers to integration of various components into the production workflow.

Before beginning a data cycle with a given vague goal, data scientists must perform exploratory data analysis (EDA) to understand and what the data is about.In EDA the data scientists can take care of duplicate values, outliers and corrupt records in general must perform sanity checks to obtain meaningful insights from the data which helps in setting metrics for the mining. It must be taken care that the log data should be designed to behave independently of each other, which will result in a large number of isolated data stores that must be compiled to reconstruct a complete picture of what happened.

Although the process appears simple there are inherent limitations in practice and requires finding the right balance between the development speed and ease of analysis.

The tools described in the paper would greatly simplify the data mining process for data scientists and would be a great help to them in their work. For example using HDFS and other tools such as Apache, HBase, Pig, and HIVE instead of traditional MySQL databases. Furthermore, it highlights the importance of plumbing and how heterogeneity of different components can be deployed to create data analytics platforms within production workflows when integrated together.

Lastly, despite the fact that the insights provided within this paper are not a complete solution for every scaling issue in big data, the paper nevertheless paves the way for a smoother workflow and easier data-driven analytics and production environment.