Report on Scaling Big Data Mining Infrastructure: The Twitter Experience

In this paper author discuss about the growth of infrastructure and the development of capabilities for data mining on "big data". There are two critical challenges faced in the processing for big data to generate successful insights they are the schemas that are designed to store petabytes of data and secondly the nature of data which is heterogeneous and integrating those data into deployment claims to be a tedious work. There are several stages of data cycle for a data mining process

Stages of data cycle for a data mining process

- Service Architectures and Logging – involves retrieving data from the database
- Exploratory Data Analysis – cleaning and processing the data
- Data Mining – mining the data for successful insights
- Production and Other Considerations – deployment and post production

To address the issue of Organizations seek to tackle this challenge by imposing basic, general schemas for all the data that the developers expressly desire to monitor, specifically in this article to solve the schema of twitter logs (that records the tweets and time) created for every nano second. Second, not using MySQL for log data. Instead, logs can be stored in Hadoop Distributed File System (HDFS), because logs should be viewed as immutable, append-only data sinks. In addition, logs can be stored in JSON format rather than plain text, which eliminates the need to select the appropriate delimiter.

Given all of the hurdles, establishing a successful big data analytics platform requires finding a balance between numerous criteria such as development speed, simplicity of analytics, flexibility, scalability, and resilience. A smaller team might easily access the front-end via JSON logging without having to worry about the data formats, but in terms of future scalability issues, the team will be incurring technological debt. Aside, an analytics framework would give all of the advantages of schemas, data catalogues, integration hooks, comprehensive data dependency management, workflow scheduling, and so on while needing no additional cost.

This article helps Practitioners and academics researcher with their work in a multiple ways. This document enables Practitioners to get a few giggles out of the experiences while also learning how to prevent similar blunders. A deeper grasp of the larger context of big data mining might inspire future work to expedite insight production operations for academic.