Summary of Academic Paper Number 3[1]

Andrew K Boles
University of Texas at San Antonio
Email: ckj771@my.utsa.edu

I. Summary

In this paper, the authors discuss a method to analyze traffic data in real-time using Apache Spark[2] and specifically Spark Streaming[3]. The main problem being discussed is the heterogeniety of the incoming data, or the variety in the type and amount of data. The authors present their methods used to create this system, including using Apache Flume[4] to control the movement of the large amounts of data to work alongside Spark Streaming. It was found that the best real-time analysis was done while using the Java Messaging Service[5] to avoid overflow. Additionally, the authors noted that the biggest contributor to the latency of their analysis was the HDFS data aggregation and that it can create a bottleneck if the configuration of Flume and Spark Streaming are not set up ideally.

References

[1] A. I. Maarala, M. Rautiainen, M. Salmi, S. Pirttikangas, J. Riekki. "Low latency analytics for streaming traffic data with Apache Spark", Department of Computer Science and Engineering, University of Oulu, Finland, 2015.
[2] Apache, "Apache Spark". Available: *http://spark.apache.org/*
[3] Apache, "Apache Spark Streaming". Available: *http://spark.apache.org/streaming/*
[4] Apache, "Apache Flume". Available: *https://flume.apache.org/*
[5] Java, "Java Messaging Service Tutorial". Available: *http://docs.oracle.com/javaee/6/tutorial/doc/bncdq.html*