Summary of Academic Paper Number 1[1]

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

I. Summary

In this paper, the authors address a current problem in Spark Streaming: real time event processing. Spark Streaming is Apache Spark's real time streaming API[2]. The problem occurs when the amount of input data that needs to be processed changes, or becomes unstable. This causes a delay in the processing of the data and slows down the stream. The authors suggest an algorithm that when the stream encounters a dynamic shift increase in the amount of data being processed, it analyses the average amount of incoming data and either: decreases the batch interval by half to process more efficiently or if the amount of incoming data is greater than a user specified amount, even smaller batch sizes allow for more jobs to be submitted.

References

[1] X. LIao, Z. Gao, W. Ji, Y. Wang. "An Enforcement of Real Time Scheduling in Spark Streaming", School of Computer Sciene, Beijing Institute of Technology, 2015.
http://spark.apache.org/docs/latest/mllib-guide.html

[2] Apache, "Spark Streaming Programming Guide". Available:
http://spark.apache.org/docs/latest/streaming-programming-guide.html