Alexander Pitcher (100693749)

**Deliverable 4:**

**Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.**

Another processing service that is usually used in the cloud environment is Apache Spark. Spark is able to scale its usage across multiple machines to process information very quickly, it is also flexible as to which programming language you wish to use. Apache is made to be quick at processing and tolerant to faults within the structure, whereas dataflow is quite similar, but limited to Java and Python and uses Apache Beam pipelines. Some advantages of Apache Spark is that it is extremely quick, is part of one large unified engine, and it is very easy to use compared to other processing services. Some disadvantages of Apache Spark is that it can not do real time processing, it is optimized for larger files, and it is quite expensive to run. Many of the limitations of Apache spark are also its disadvantages (not good with small files, no real time processing, etc.) but it is also limited by iterative processing, and its use of fewer algorithms compared to other processing services.

**Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decide to use another dataset, It should maintain both variety and huge volume. Your report should include but not limited to**

Batch Processing is able to process a large amount of data quickly, while stream processing is able to process a non-stop data stream. An application where using both of these processing methods would be useful could be something such as a social media platform that is consistently taking in large amounts of data and needs to have it processed quickly. If you were to use a dataset such as Social Influence on Shopping(163.59 KB, using .csv), a mix of batch and stream processing could be used to process the whole dataset, as well as take on new entries to the dataset at the same time, making a very efficient processing algorithm. There are many tools available to implement these data processing techniques to this .csv file. You would be able to do so using MySql, Oracle, Microsoft Data Program, etc.