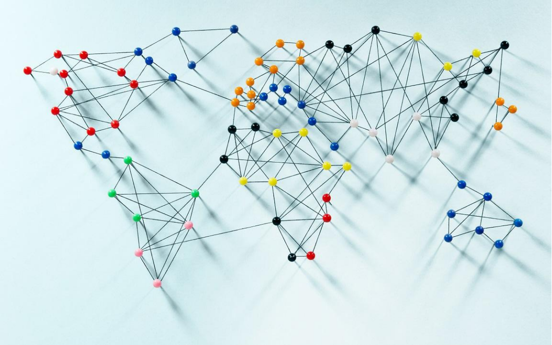
Comparing the performance of Graph Analysis algorithms using Apache Flink and Apache Spark graph processing libraries



Mohamed Gabr & Óttar Guðmundsson

2018.09.12

Aims

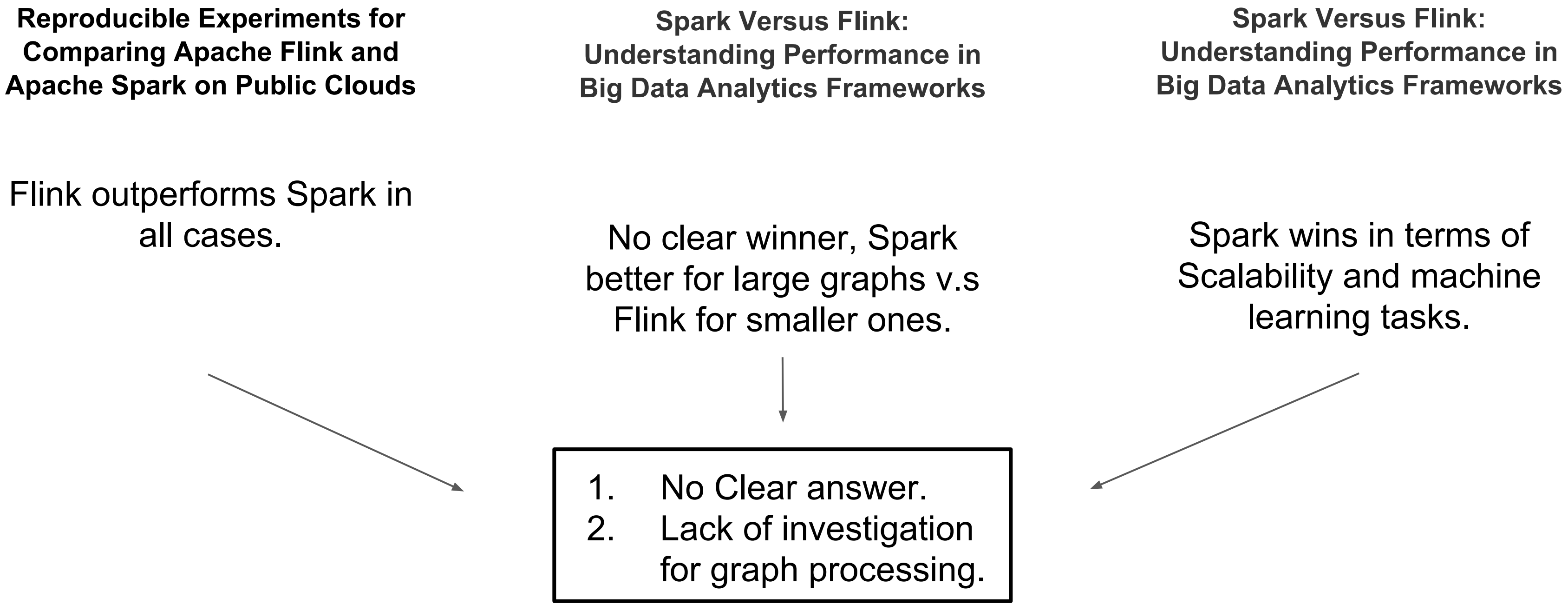
Graph processing. Which library to use

Which of the two libraries outperforms the other in terms of speed?

Main metric is execution time, but will consider CPU/Memory

Theory / Literature

Three points of views:



Reproducible Experiments for Comparing Apache Flink and Apache Spark on Public Clouds

Flink outperforms Spark in all cases.

Spark Versus Flink: Understanding Performance in Big Data Analytics Frameworks

No clear winner, Spark better for large graphs v.s Flink for smaller ones.

Spark Versus Flink: Understanding Performance in Big Data Analytics Frameworks

Spark wins in terms of Scalability and machine learning tasks.

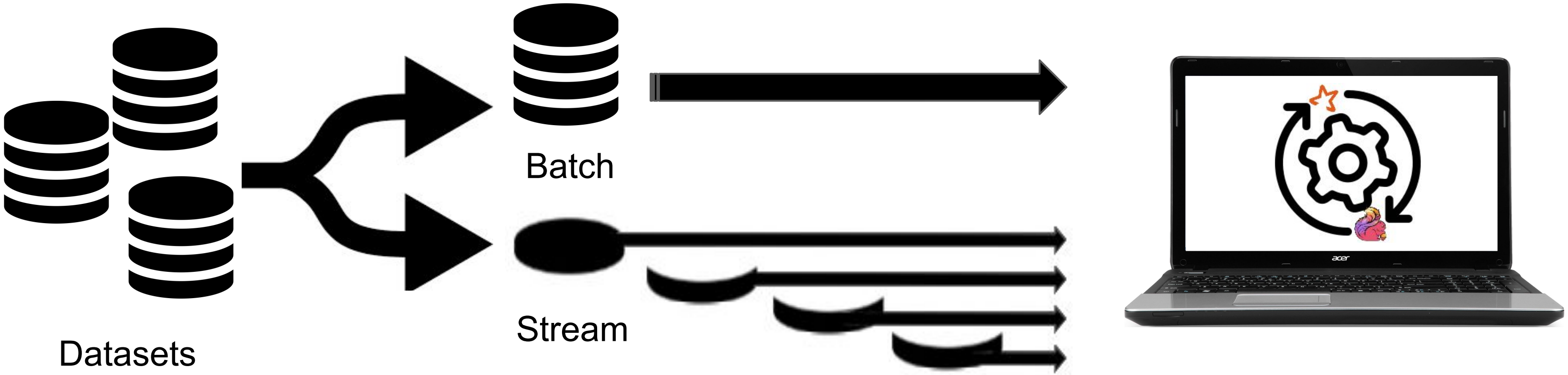
1. No Clear answer.
2. Lack of investigation for graph processing.

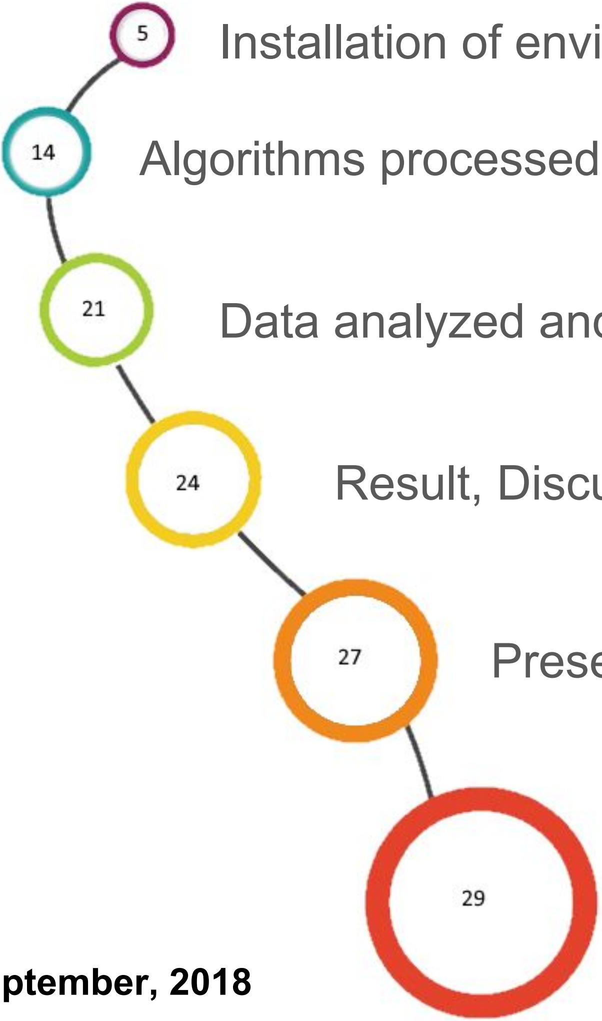
Research Methodology

Two mainstream computers with GraphX (Spark) and Gelly (Flink)

Few different datasets that vary in size

Data processed in batches or as streams





Installation of environment (Done)

Algorithms processed on graphs as batches or streams

Data analyzed and tables/figures created

Result, Discussion and Conclusion sections written in report

Presentation complete

All deliverables ready in a .zip file

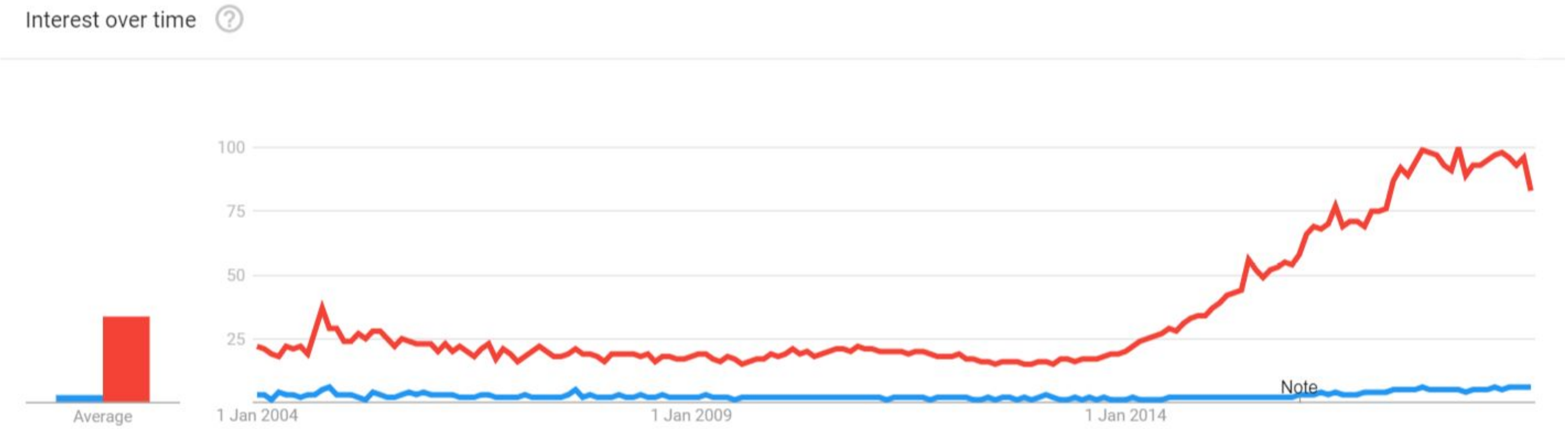
September, 2018

Risks

Former Spark Experience

Lack of Flink community and Gelly documentation

Difficulties in benchmarking



Thank you for listening!