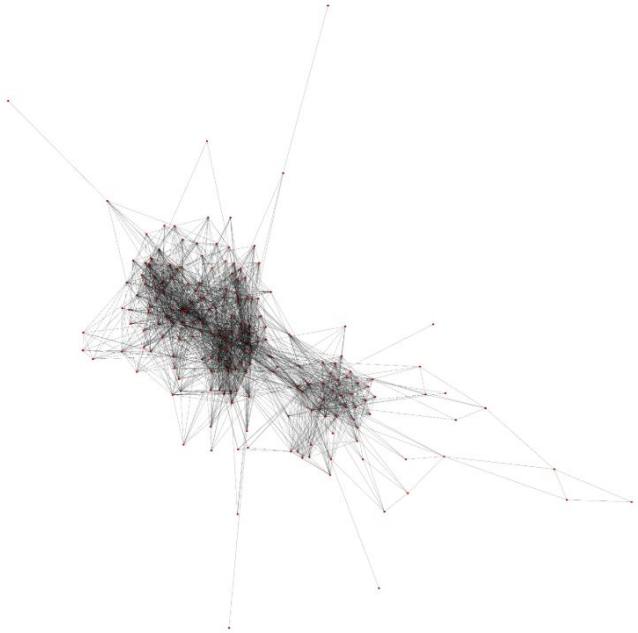
Comparing the performance of graph analysis algorithms on single node using Apache Flink Gelly and Apache Spark GraphX

Mohamed Gabr & Óttar Guðmundsson

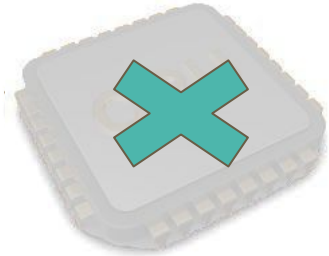
2018.10.19

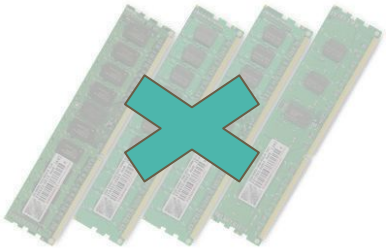


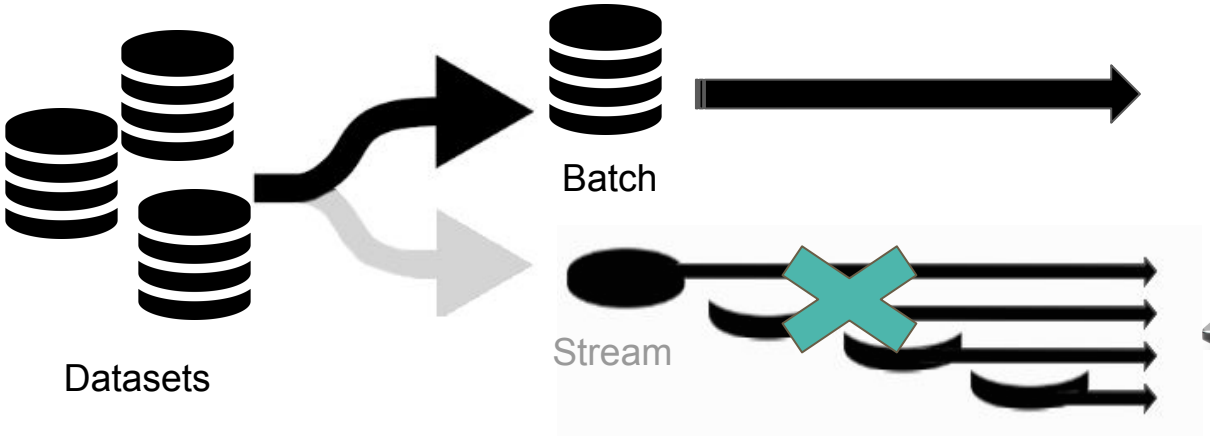
Problem statement

Original scope

Reduced scope







Background and related work

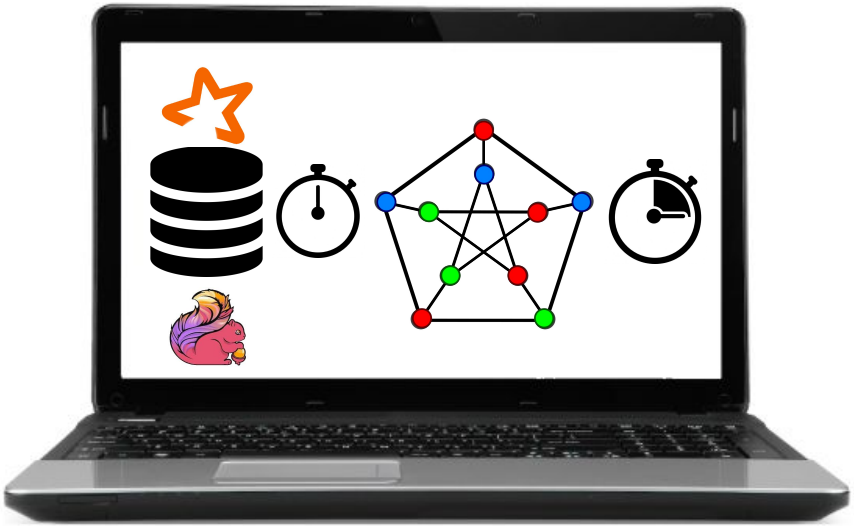
- No clear winner
- Few researches favor Spark, other Flink
- Flink is faster for smaller graphs, while Spark is faster for larger graphs
- External libraries for machine learning favors Spark

Flink
ML

Method - variables

- Three different datasets
  - Jazz Musicians - 198 nodes
  - U.Rovira i Virgili - 1.133 nodes
  - Pretty Good Privacy - 10.680 nodes
- Two different libraries
  - GraphX (Spark)
  - Gelly (Flink)
- Three algorithms
  - Connected Components
  - PageRank
  - Graph Coloring
- Environment
  - MacBook Pro Retina 2015
  - ThinkPad T450s

Method - measurements



Algorithm	PageRank			
Dataset	*Jazz Musicians*		*I rovira*	
Max iterations	Flink	Spark	Flink	Spark
10	...	...	...	...
10	...	...	...	...
10	...	...	...	...
10	...	...	...	...
10	...	...	...	...
100	...	...	...	...
100	...	...	...	...
100	...	...	...	...

	Jazz Musicians		
	Difference Significance in ns		
Max Iterations	**Graph Coloring**	**Connected Comps**	**PageRank**
10	...	...	...
100	...	...	...
200 - 400	...	...	...
1000	...	...	..

Installation of environment (**Done 2018.09.03**)

Algorithms processed on graphs as batches (**Done 2018.09.25**)

Data analyzed (**Done 2018.09.19**) and tables/figures created (**2018.09.27**)

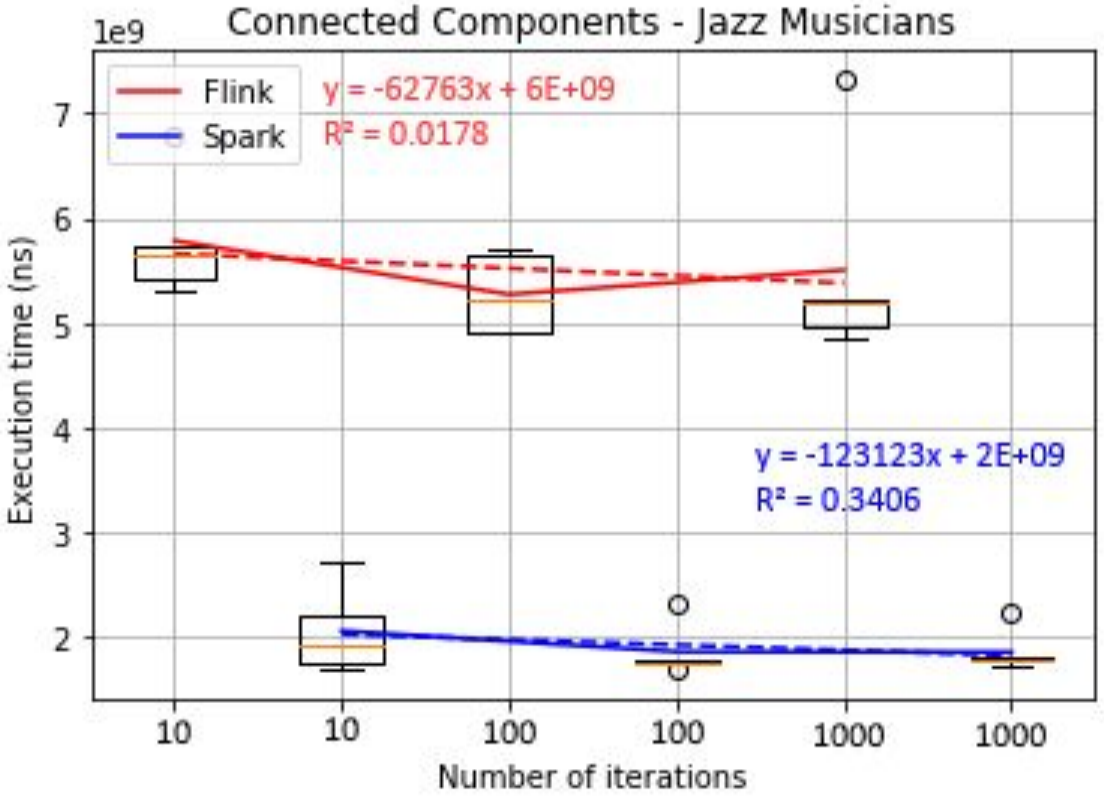
Result, Discussion and Conclusion written in report (**Done 2018.10.01**)

Presentation complete (**Done 2018.10.03**)

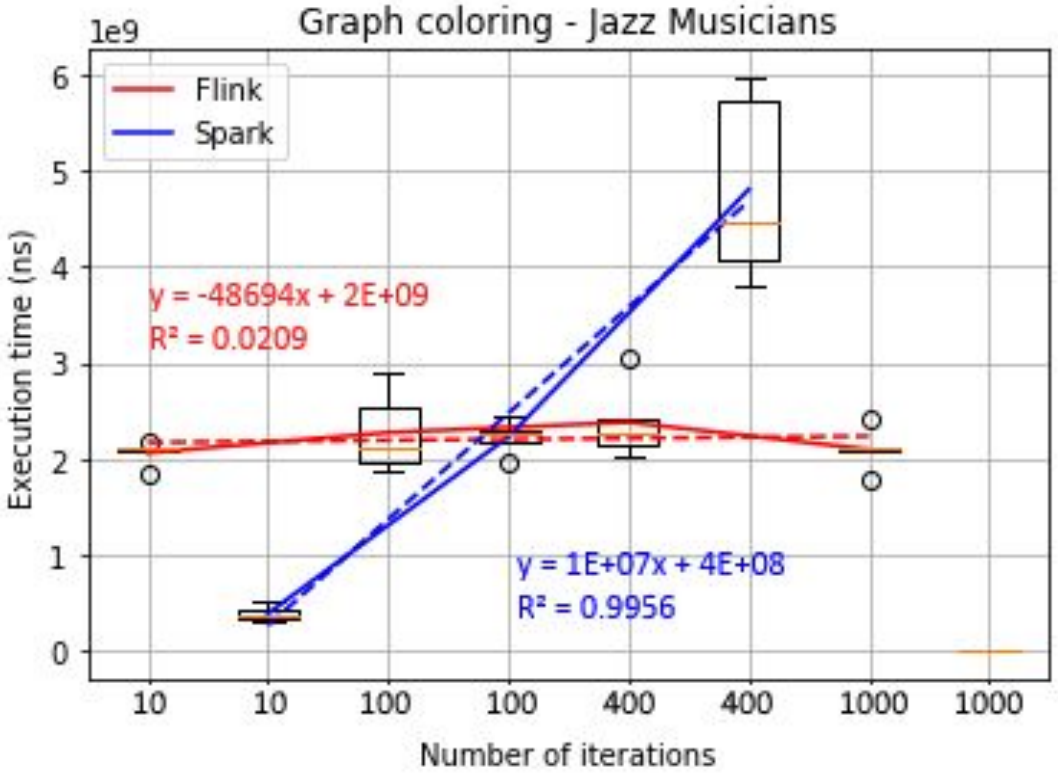
All deliverables ready in a .zip file (**Done 2018.10.03**)

September/October, 2018

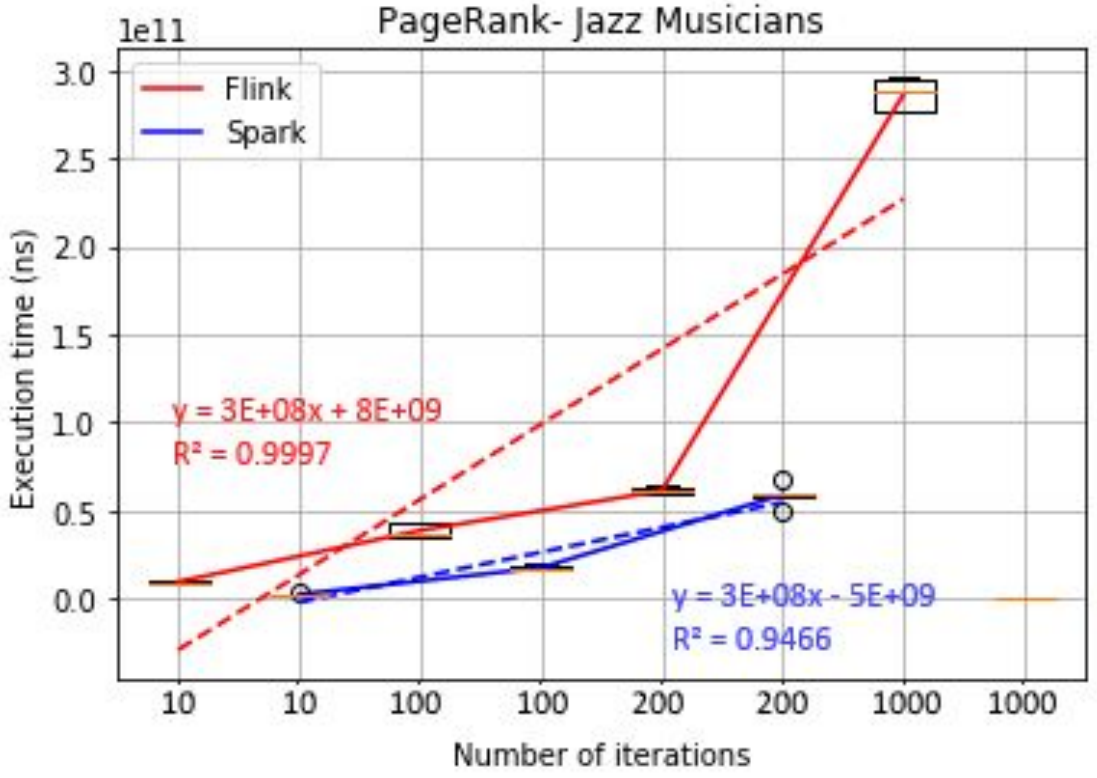
Results and analysis

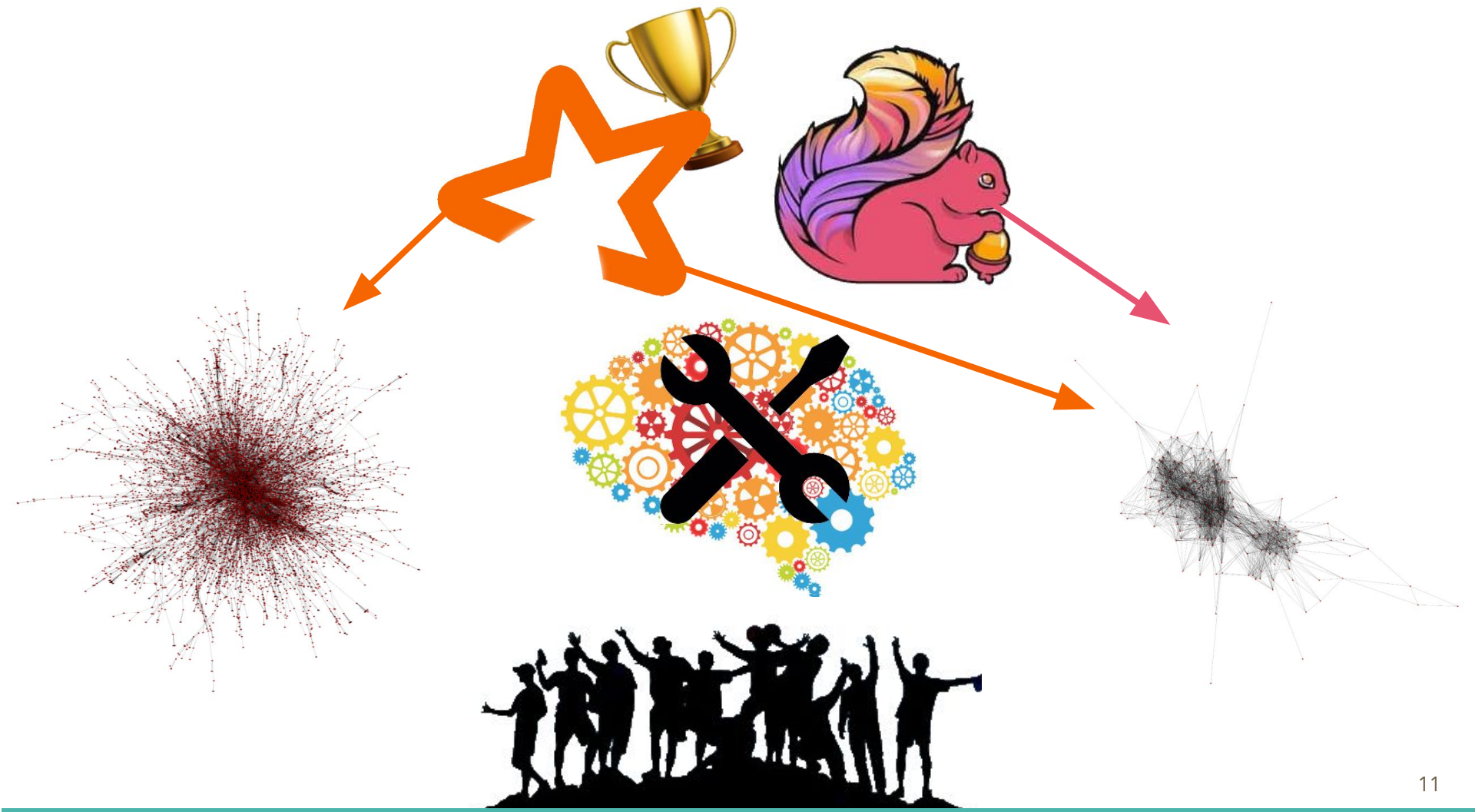


Results and analysis



Results and analysis





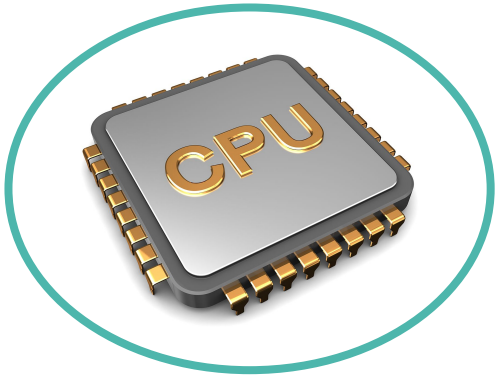
Conclusion

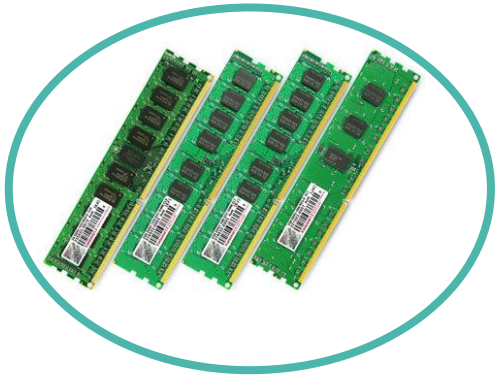
Spark is winner in batch processing for any size of graph data

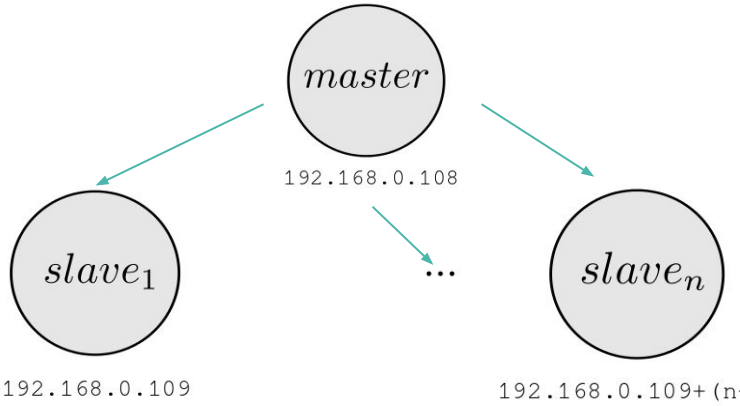
Flink manages memory better,
but Spark can be tuned to match memory efficiency

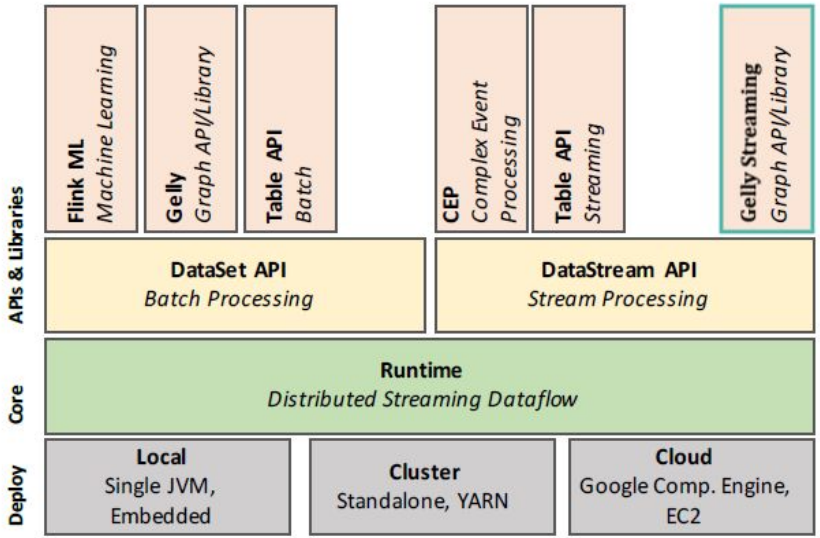
Further benchmarking research is needed on
CPU and memory between Spark and Flink

Future work


CPU




master
192.168.0.108
$slave_1$
192.168.0.109
...
$slave_n$
192.168.0.109+(n−1)


APIs & Libraries
Flink ML
Machine Learning
Gelly
Graph API/Library
Table API
Batch
CEP
Complex Event Processing
Table API
Streaming
Gelly Streaming
Graph API/Library
DataSet API
Batch Processing
DataStream API
Stream Processing
Core
Runtime
Distributed Streaming Dataflow
Deploy
Local
Single JVM, Embedded
Cluster
Standalone, YARN
Cloud
Google Comp. Engine, EC2

Thanks for your attention

Questions