An Experimental Survey on Big Data Frameworks

W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, E. Mephu Nguifo

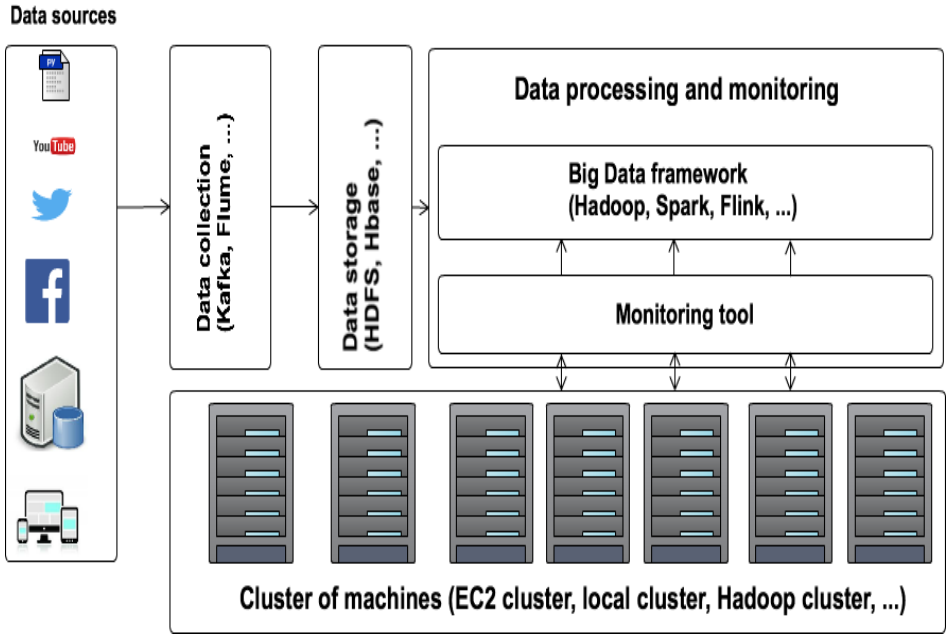
XLDB 2017
October 10-12, 2017
Casino de Royat, Allée du Pariou, 63130 Royat, FRANCE

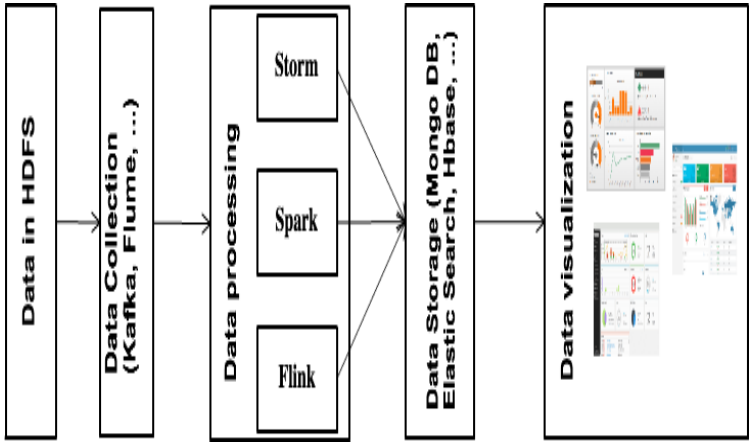
Context and motivations

Big Data problems:

- Scalability and fault tolerance requirements
- Several applications have been migrated to Big Data
- Several Big Data frameworks have been proposed

Experimental protocol





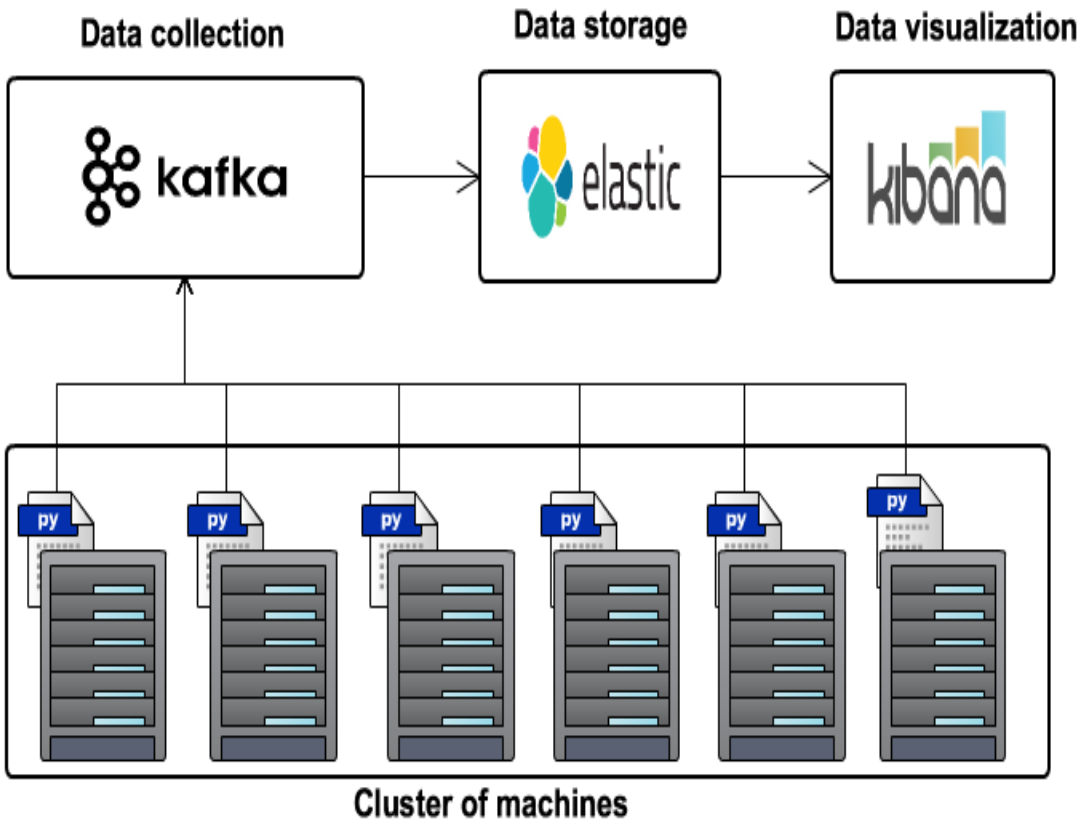
Batch Mode evaluation

- Workload: Kmeans, WordCount and PageRank.
- Frameworks: Hadoop (Mapreduce), Spark and Flink.
- Features: Scalabilty, Configuration parameters.

StreamMode evaluation

- Workload: ETL Workload.
- Frameworks: Storm, Spark and Flink.
- Features: Number of events processed.

Experimental protocol

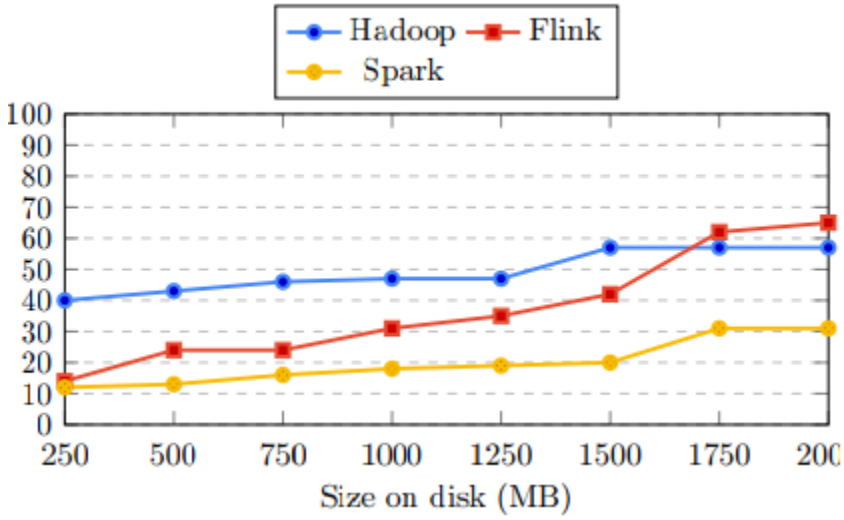


Monitoring Tool

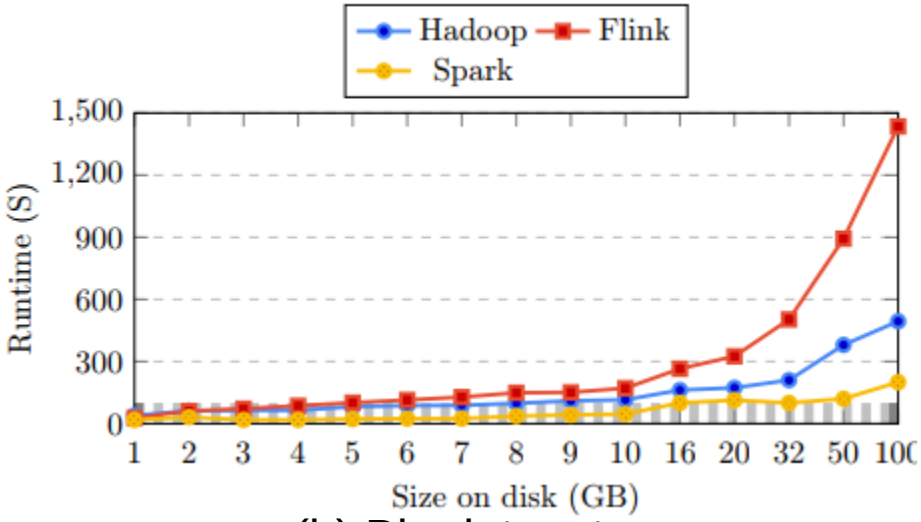
Data collection: kafka.
Data Strorage: Elastic search.
Data Visualisation: Kibana

Experimental results

Batch Mode results :



(a) Small datasets

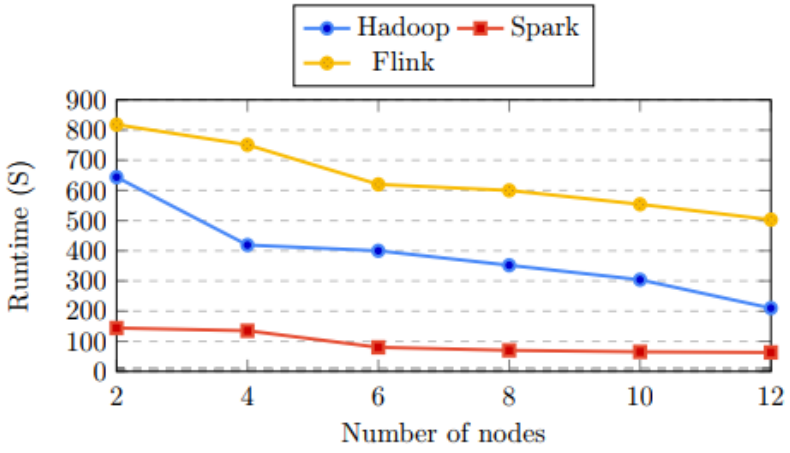


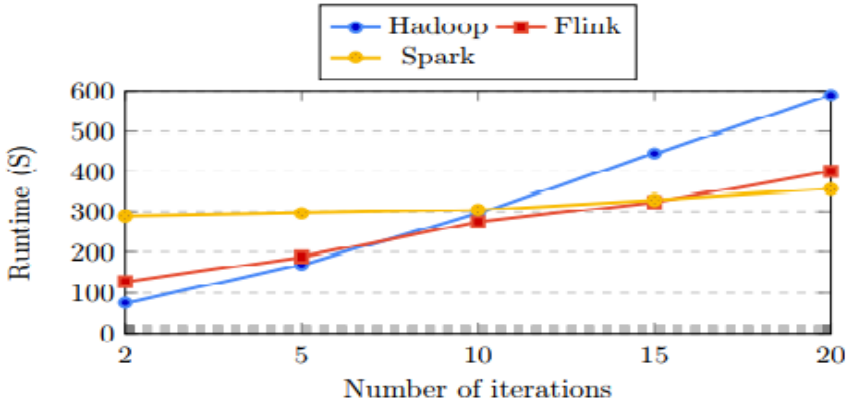
(b) Big datasets

Impact of size of data on response time, with small and big datasets.

Experimental results

Batch Mode results :



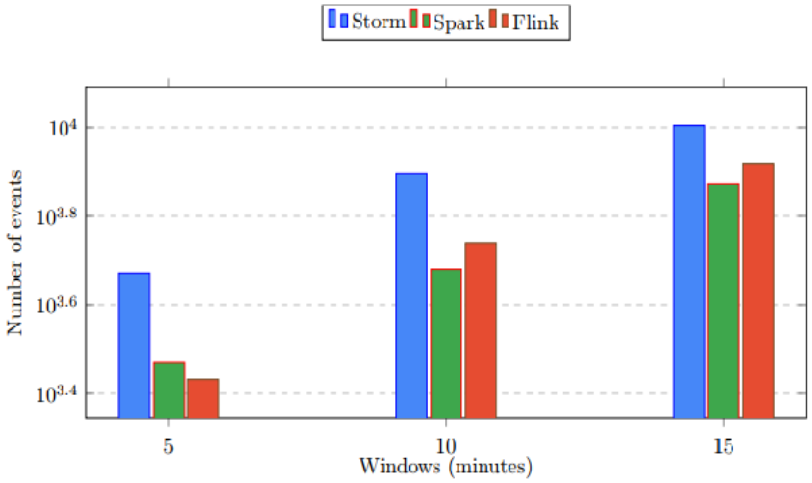


- Vary the number of machines in cluster.
- Study the scalability feature.

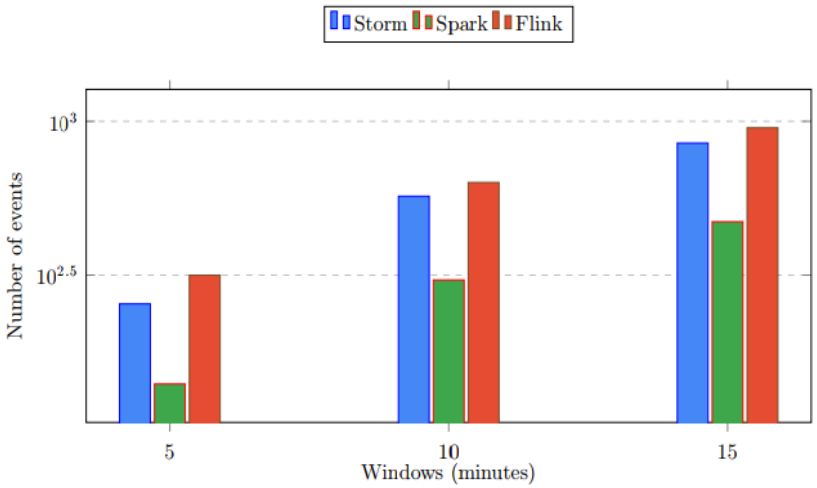
- kmeans workload
- Vary the number of iterations,

Experimental results

Stream Mode results :



Event with 100 kb



Event with 500 kb

Impact of the size of messages on the number of processed messages

- Storm performs better in the case of small datasets.
- Flink provides good results in the case of big datasets.

More at the Poster Session !