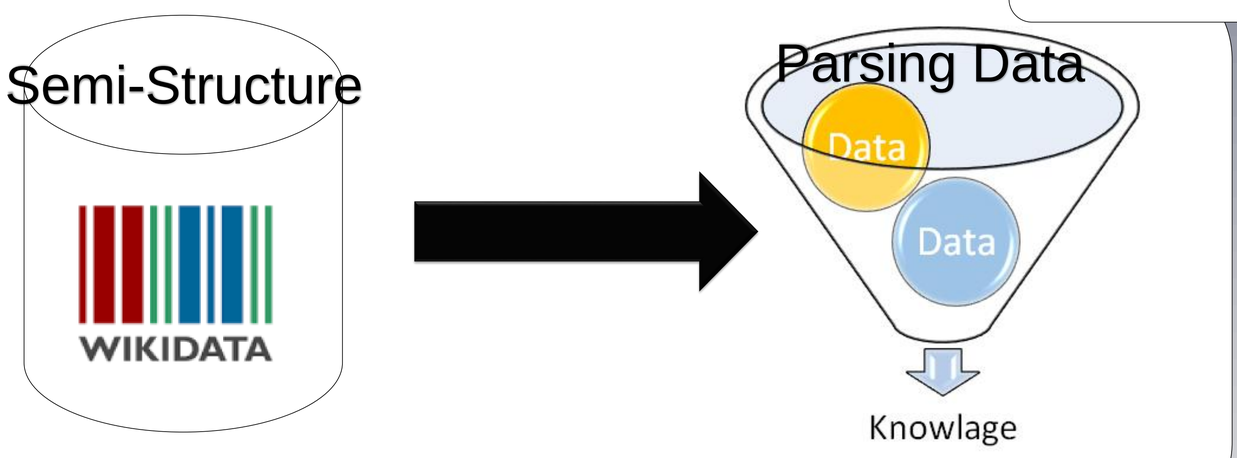
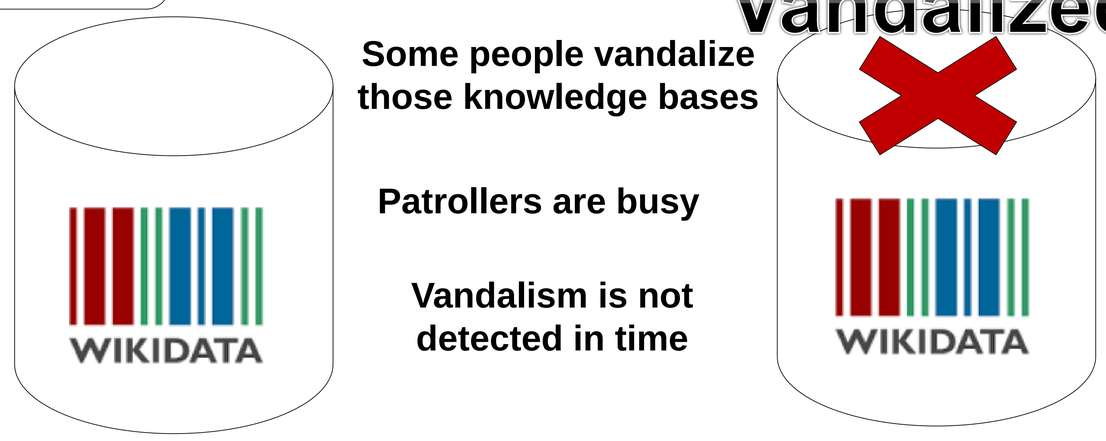
Efficient Data Parsing and Vandalism Detection on (Big) Knowledge Bases using Apache Spark and Hadoop ecosystem

Nayef Roqaya

Motivation





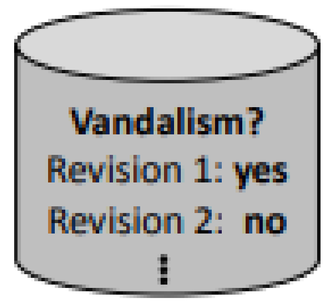
Methods

Distributed
Data storage
&&
computation environment

Distributed Data parser

Distributed Vandalism Detector

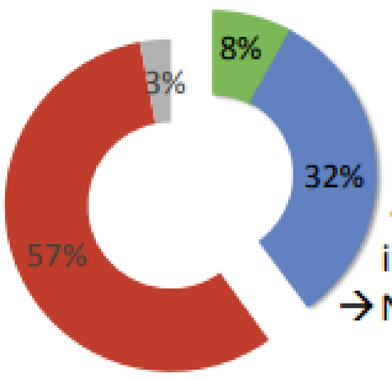
Spark ML



ROC
PR

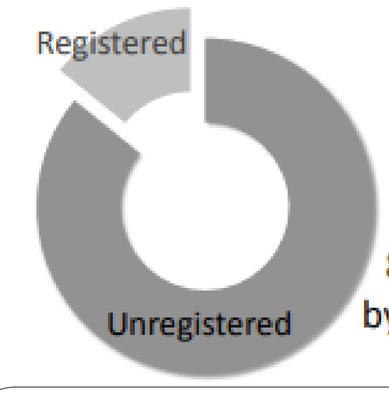
Analysis

What is vandalized?

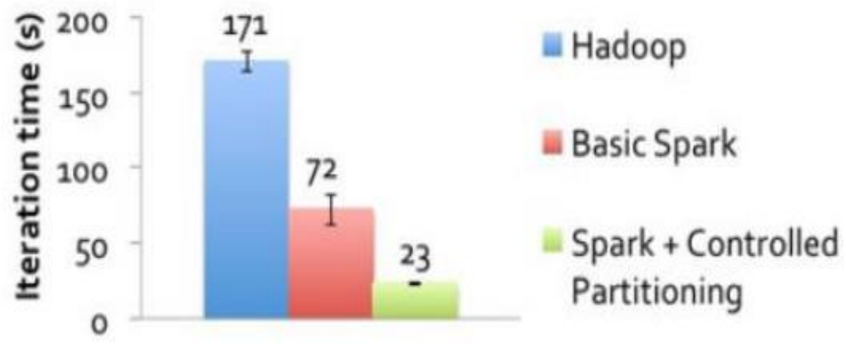


40% of vandalism in *structured* data
→ New features beneficial

Who vandalizes?



86 % of vandalism by unregistered users



Results & Limitations

- Features engineering plays a vital role in improving results of ML vandalism approach.
- By Spark with controlled partitioned techniques, the performance is efficient.

- Hadoop consumes a big size data storage.
- Streaming Data from Hadoop cannot deal with compressed files as result to need to keep **logic partitioning** and **physical partitioning** in our case study.

Conclusion

Vandalism can reduce the quality of knowledge bases.
Detect vandalism automatically is possible .

References: https://www.heindorf.me/wdvd.html - www.wsdm-cup-2017.org