Big data and machine learning framework for clouds and its usage for text classification

István Pintye, Eszter Kail, Péter Kacsuk
Institute for Computer Science and Control
Hungarian Academy of Sciences
Budapest, Hungary
pintye.istvan@sztaki.mta.hu

Péter Kacsuk
University of Westminster
London, UK
P.Kacsuk@westminster.ac.uk

ABSTRACT

The paper describes a big data and AI application development and execution framework that was originally developed for MTA Cloud (an OpenStack based cloud) but could be used on other clouds including Amazon, OpenStack, OpenNebula and CloudSigma. The paper explains the concept and components of the big data and AI environment and illustrates its usage by a text classification application.

Keywords—machine learning; big data; parallel and distributed execution; cloud;

REFERENCES

[1] "SZTAKI Cloud home - SZTAKI Cloud." [Online]. Available: https://cloud.sztaki.hu/en/home. [Accessed: 01-Apr-2019].

[2] "MTA Cloud | MTA Cloud." [Online]. Available: https://cloud.mta.hu/. [Accessed: 01-Apr-2019].

[3] E. Fernández-del-Castillo, D. Scardaci, and Á. L. García, "The EGI Federated Cloud e-Infrastructure," Procedia Comput. Sci., vol. 68, pp. 196–205, Jan. 2015.

[4] "Whitepapers – Amazon Web Services (AWS)." [Online]. Available: https://aws.amazon.com/whitepapers/. [Accessed: 01-Apr-2019].

[5] "Laboratory of Parallel and Distributed Systems | MTA SZTAKI." [Online]. Available: https://www.sztaki.hu/en/science/departments/lpds. [Accessed: 01-Apr-2019].

[6] J. Kovács and P. Kacsuk, "Occopus: a Multi-Cloud Orchestrator to Deploy and Manage Complex Scientific Infrastructures," J. Grid Comput., vol. 16, no. 1, pp. 19–37, Mar. 2018.

[7] "HDFS Architecture Guide." [Online]. Available: https://hadoop.apache.org/docs/current1/hdfs_design.html. [Accessed: 01-Apr-2019].

[8] "MapReduce Tutorial." [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html. [Accessed: 01-Apr-2019].

[9] "Welcome - Occopus." [Online]. Available: http://occopus.lpds.sztaki.hu/de/. [Accessed: 01-Apr-2019].

[10] "Apache SparkTM - Unified Analytics Engine for Big Data." [Online]. Available: https://spark.apache.org/. [Accessed: 01-Apr-2019].

[11] "MLlib | Apache Spark." [Online]. Available: https://spark.apache.org/mllib/. [Accessed: 01-Apr-2019].

[12] "Open source and enterprise-ready professional software for data science - RStudio." [Online]. Available: https://www.rstudio.com/. [Accessed: 01-Apr-2019].

[13] "The Jupyter Notebook — IPython." [Online]. Available: https://ipython.org/notebook.html. [Accessed: 01-Apr-2019].

[14] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[15] "Classification and regression - MLlib main guide." [Online]. Available: https://spark.apache.org/docs/latest/ml-classification-regression.html.

[16] "Ensembles - RDD-based API - Spark 2.4.0 Documentation." [Online]. Available: https://spark.apache.org/docs/latest/mllib-ensembles.html. [Accessed: 01-Apr-2019].