PROTEUS: Scalable Online Machine Learning for Predictive Analytics and Real-Time Interactive Visualization

Bonaventura Del Monte[1], Jeyhun Karimov[1], Alireza Rezaei Mahdiraji[1], Tilmann Rabl[1,2], Volker Markl[1,2]
[1] German Research Center for Artificial Intelligence (DFKI), [2] TU Berlin
[1]firstname.lastname@dfki.de, [2]firstname.lastname@tu-berlin.de

ABSTRACT

Big data analytics is a critical and unavoidable process in any business and industrial environment. Nowadays, companies that do exploit big data's inner value get more economic revenue than the ones which do not. Once companies have determined their big data strategy, they face another serious problem: in-house designing and building of a scalable system that runs their business intelligence is difficult. The PROTEUS project aims to design, develop, and provide an open ready-to-use big data software architecture which is able to handle extremely large historical data and data streams and supports online machine learning predictive analytics and real-time interactive visualization. The overall evaluation of PROTEUS is carried out using a real industrial scenario.

1. PROJECT DESCRIPTION

PROTEUS[1] is an EU Horizon2020[2] funded research project, which has the goal to investigate and develop ready-to-use, scalable online machine learning algorithms and real-time interactive visual analytics, taking care of scalability, usability, and effectiveness. In particular, PROTEUS aims to solve the following big data challenges by surpassing the current state-of-art technologies with original contributions:

1. Handling extremely large historical data and data streams
2. Analytics on massive, high-rate, and complex data streams
3. Real-time interactive visual analytics of massive datasets, continuous unbounded streams, and learned models

PROTEUS's solutions for the challenges above are: 1) a real-time hybrid processing system built on top of Apache Flink[3] (formerly Stratosphere[4] [1]) with optimized relational algebra and linear algebra operations support through LARA declarative language [2, 3], 2) a new library for scalable online machine learning and data mining called SOLMA, and 3) investigation and development of incremental visual methods that allow end-users to efficiently explore both batch and streaming data for making well-informed decisions in real time. These three subsystems will be integrated in a single platform running in a containerized environment. Once the platform is deployed in a cluster, its life-cycle is as follows: 1) the end-user writes data analytics tasks in LARA mixing extract-transform-load and SOLMA algorithms pipelines and executes them on top of PROTEUS hybrid processing system, 2) the system continuously trains deployed machine learning models in an online fashion, 3) the visual stack queries those models and displays requested real-time predictions and statistics to end-user.

PROTEUS faces an additional challenge which deals with correct integration of machine learning solutions in big data processing systems by taking into account the principal anti-patterns and risks factors that affect this kind of interactions [4].

In addition, PROTEUS ensures the achievement of its goals through rigorous experimental testing and industrial-validated processes. The project is indeed guided by the specific requirements of the *hot strip mill* steel-making process, provided by an industrial partner of PROTEUS' consortium. Hot strip mill produces coils, whose quality is affected by several parameters (e.g. temperature, vibration intensity, tension in the rollers). Since coils are used in further production stages, they must present no defect. Predicting anomalies through the analysis of massive real-time data generated during the hot strip mill is the main target in this validation scenario.
Regardless the above validation scenario, PROTEUS platform is also applicable for general data streams analysis in other domains.

Acknowledgements. This work was supported by the EU Horizon 2020 project PROTEUS (687691).

[1]https://www.proteus-bigdata.com/
[2]https://ec.europa.eu/programmes/horizon2020/
[3]https://flink.apache.org/
[4]http://stratosphere.eu/

2. REFERENCES

[1] A. Alexandrov, R. Bergmann, et al. The stratosphere platform for big data analytics. *The VLDB Journal*, 23(6):939–964, Dec. 2014. ISSN 1066-8888.

[2] A. Alexandrov, A. Kunft, et al. Implicit parallelism through deep language embedding. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pp. 47–61. ACM, New York, NY, USA, 2015. ISBN 978-1-4503-2758-9.

[3] A. Kunft, A. Alexandrov, et al. Bridging the gap: Towards optimization across linear and relational algebra. In *Proceedings of the 3rd ACM SIGMOD Workshop on Algorithms and Systems for MapReduce and Beyond*, BeyondMR '16, pp. 1:1–1:4. ACM, New York, NY, USA, 2016. ISBN 978-1-4503-4311-4.

[4] D. Sculley, G. Holt, et al. Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. 2014.