STREAMLINE - Streamlined Analysis of Data at Rest and Data in Motion

Philipp M. Grulich [1], Tilmann Rabl [1,2], Volker Markl [1,2], Csaba Sidló [3], Andras Benczur [3]
[1] German Research Center for Artificial Intelligence (DFKI), [2] TU Berlin, [3] Hungarian Academy of Sciences (MTA SZTAKI)
[1]firstname.lastname@dfki.de, [2]firstname.lastname@tu-berlin.de, [3]lastname@sztaki.hu

ABSTRACT

STREAMLINE aims for improving the overall workflow of big data analytics systems. For this goal, it combines research in different areas to reduce the complexity of the work with *data at rest* and *data in motion* in a unified fashion. As a foundation STREAMLINE offers a uniform programming model on top of Apache Flink, for which it drives innovations in a wide range of areas, such as interactive *data in motion* visualization and advanced window aggregation techniques.

1. PROJECT SUMMARY

The STREAMLINE project aims to improve the workflow and usability of current big data analysis systems. Therefore it provides a uniform system, which is able to handle the analysis of *big data at rest* as well as *fast data in motion*. With this platform, STREAMLINE enables a reduction of complexity, costs, and latency.

Traditionally *batch-* and *stream-processing* were considered as two very different types of applications, but in the last years, it has been shown that the most real-world usecases required systems for both workloads. This forces companies to integrate different specialized systems, which leads not only to complex system architecturesand introduces maintenance overhead, it also introduces a high latency to the general data analysis workflow. This is also known as the problem of system and human latency in big data analysis. Even technologies that are able to combine *data in motion* and *data at rest* are currently very complex and difficult to deploy, maintain and use. Beside this many companies have a demand for much more advanced analyses, which are still hard to implement in current systems.

To reduce this complexity STREAMLINE combines research and innovations in the areas of distributed systems, data management, and machine learning. Whereby STREAMLINE's key goal is to arrive at sustainable innovation by technology transfer to an established and growing open source project. STREAMLINE focuses on innovations in the area of the following four reactive and proactive applications: customer retention, personalized recommendations, target advertisement and multilingual Web processing. To integrate the innovations into the industry, STREAMLINE partners with multiple companies.

As its system foundation STREAMLINE relies on the open source data processing system Apache Flink, which is able to handle batch and stream processing on a single pipelined execution engine [1]. On top of this STREAMLINE offers a single uniform programming model that can automatically be optimized, parallelized, and adopted to the system load, data distribution, and architecture.

Research Highlights: Cutty [2] introduces a general aggregation sharing framework for streaming windows, which outperforms previous solutions in order of magnitudes. This technique utilizes the fact that window aggregations are one of the most redundancy-prone operations in current stream processing. Cutty is also suitable for multi query aggregation sharing and non-periodic windows, such as session window, which can be used for more complex business logic. Based on this technique STREAMLINE enables higher throughput and improves the efficiency of its data processing platform.

I^2 [3] in contrast, focuses on the visualization and interactive aggregation of *data in motion*, which is a key enabler for fast and efficient real-time data analysis. It contributes an interactive development environment, which coordinates the cluster application and includes interactive stream visualization techniques. With this I^2 is able to handle advanced and adaptive aggregations directly on the cluster. As one example we provide an aggregation algorithm for timer-series data, which reduces the amount of data in a data-rate independent manner and is proven to be correct and minimal in terms of transferred data. Therefore I^2 is an important part of STREAMLINE, because it enhances the usability and accessibility of its platform.

2. ACKNOWLEDGEMENTS

This work was supported by the EU Horizon 2020 project Streamline (688191).

3. REFERENCES

[1] Carbone et al. Apache flink: Stream and batch processing in a single engine. *IEEE Data Eng. Bull.*, 38(4), 2015.
[2] Carbone et al. Cutty: Aggregate sharing for user-defined windows. In *CIKM*, pages 1201–1210, 2016.
[3] Traub et al. I2: Interactive real-time visualization for streaming data. In *EDBT*, EDBT, 2017.