An Elastic Data Stream Processing Ecosystem
for Distributed Environments

Doctoral Candidate Christoph Hochreiner, 00726292

Advisor Univ.Prof. Dr. Schahram Dustdar (TU Wien)
2nd Advisor Assistant Prof. Dr.-Ing. Stefan Schulte (TU Wien)
Reviewer Prof. Dr. Harald C. Gall (University of Zurich)
Univ.Prof. Dr. Gertrude Kappel (TU Wien)

Abstract
In the last couple of years, we have observed a trend towards an ever-growing number and volume of data streams. Up to now, these data streams were mainly originating from social media services but today the emergence of the Internet of Things (IoT) also contributes to the growth of data streams. Besides the growth of the data volume, the IoT also introduces several new challenges, like the geographically distributed locations of IoT-devices, i.e., data sources and processing capabilities, as well as a differentiation of the user base who uses Stream Processing Applications (SPAs). Previously, SPAs were only used by data stream processing experts to process large data volume primarily for social media, medical or financial purposes in a centralized setting. However, the emergence of the IoT allows a larger user base, like companies from the manufacturing domain or even individual users, to process data streams to extract valuable insights. To address these challenges, it is required to evolve the system design of today’s stream processing engines and create an ecosystem for data stream processing, which considers all aspects of designing and operating SPAs.

Therefore, we introduce the VISP Ecosystem in this thesis, which provides a holistic approach for creating SPAs and propose novel concepts to operate SPAs in a distributed environment. To improve the creation of SPAs, we present a novel description language for SPAs that supports distributed deployments as well as several non-functional aspects for SPAs that are not considered in today’s approaches. In addition to the fundamental aspects of designing and operating SPAs, we also introduce two resource provisioning approaches. These two approaches use the resource elasticity provided by the cloud computing paradigm to reduce the operational cost for running SPAs under volatile data volume. The first resource provisioning approach is threshold-based approach and can find the optimal resource configuration depending on the current data volume for the SPA. This dynamic resource provisioning approach allows this approach to outperform established fixed resource provisioning strategies regarding cost efficiency. The second approach represents an evolution of the first approach by considering additional external aspects like the billing time units to avoid any unnecessary operational overhead for updating the resource configuration. According to our evaluation, we can see that our second approach outperforms the first one for most real-world scenarios and allows for an even more cost-efficient operation of SPAs while ensuring the timely processing of data streams.