Performance Benchmarking of Apache Flink and HarmonicIO with Large Message Sizes

In recent years, data stream processing has become increasingly used for large-scale data analysis. Companies like Facebook, Spotify, Twitter, LinkedIn, Google and many more are heavily relying on the software frameworks designed for the stream processing. The Distributed Computing Applications (DCA: http://www.it.uu.se/research/group/dca ) research group at the Department of Information Technology focuses on building the new methods and frameworks to address the needs of the scientific applications. We have developed the HarmonicIO stream processing framework (https://github.com/HASTE-project/HarmonicIO). The framework is designed to fulfil processing needs for scientific datasets. The conventional streaming frameworks are extremely efficient and robust for the large datasets based on small individual object size (100s of Bytes or 10s of KBs), such as small JSON objects. Whereas scientific datasets are also massive but the individual object size is significantly large (100s of KBs or MBs). Whilst being able to process relatively large individual objects, HarmonicIO also offers cloud-native execution model, auto-scaling and complete isolation for processing engines (PEs) using container technology. The essential features and initial experiments related to HarmonicIO framework are highlighted in [1]. We have also presented a detailed comparison of HarmonicIO and the stream processing in Apache Spark [2] to highlight its role and importance as a specialized streaming framework for scientific datasets.

In this project, your first task is to learn about the streaming frameworks in general; and then conduct a performance and scalability test based on Apache Flink and HarmonicIO. Apache Flink is one of the widely use streaming frameworks for the analysis of large datasets. The available literature related to Flink project shows promising results for large datasets based on small individual objects [3]. However, performance based on relatively large individual objects in a dataset is still unclear. For conducting the experiments we will use SNIC Science Cloud (SSC) resources. The findings of this project will highlight the principle differences between two streaming frameworks and help us to decide the future enhancements in the HarmonicIO framework.

Supervisors: Salman Toor (Salman.Toor@it.uu.se), Ben Blamey (Ben.Blamey@it.uu.se)
