We process this data set with a Data Producer and a Data Consumer. The Data Producer program (which is located out of the RPI cluster) generates a data stream of taxi ride records which are read from the dataset file and then writes them into an Apache Kafka broker. The Data Producer serves events according to their timestamps, with an adjustable speed factor. We use SpeedFactor=10 in our experiments which means that 10 s of real-world events are replayed in 1 s. To generate sufficient workload without reducing the experiment duration too much, we inject three identical records in the Flink system for each record in the dataset.

