Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

Real-Time Financial Data Processing

Diplomierende
Ferenc Csàk
Silvan Meyer

Dozierende
Nils Andri Bundi
Kurt Stockinger

SIX Financial Information ist ein Anbieter von Finanzmarktdaten und -informationen. Um eine hohe Datenqualität sicherzustellen, muss *SIX* diese Daten überprüfen. In dieser Arbeit wird das *Apache Spark* Framework darauf getestet, ob es sich für diesen Einsatz bei der Qualitätsüberprüfung eignen würde und Anomalien in den Daten in Echtzeit erkennen kann.

Bei der Implementierung wurden die zwei Komponenten von *Apache Spark*, das *Structured Streaming* und das *Spark Streaming*, verwendet. Es stellte sich heraus, dass das *Structured Streaming* für diesen Zweck nicht geeignet ist, da ein Teil der gewünschten Funktionen noch nicht enthalten ist und ein implementierter Workaround leistungsmässig sehr schlecht abschnitt. Der definierte Anomalie-Erkennungs-Algorithmus konnte erfolgreich in *Spark Streaming* implementiert werden.

Die anschliessenden Performancetests zeigten, dass die Anforderungen mit genügend Ressourcen äusserst zufriedenstellend erfüllt werden konnten. Der Anomalie-Erkennungs-Algorithmus wurde dabei auf verschiedenen Konfigurationen mit bis zu 44 *Spark Cores* verteilt auf acht *Executors* getestet. Durch die schrittweise Erhöhung der Ressourcen mit verschiedenen Parametern konnte die geeignete Einstellung für die Aufgabenstellung ermittelt werden. Bei der maximalen Anzahl von *Executors* und *Cores* konnte dadurch ein Durchsatz von 167'000 Datensätzen pro Sekunde erreicht werden.

Performance des Anomalie-Erkennungs-Algorithmus in Records per Sekunde mit einer Konfiguration von einem bis acht Executors und jeweils fünf Cores pro Executor.