

Apache Spark und HBase zur Speicherung und Analyse von Sensordaten

Ende Januar trafen sich in Freiburg 20 interessierte Teilnehmer aus Unternehmen und Hochschulen, zum 8. TDWI Roundtable um sich über den Einsatz von Big Data-Technologien zur Speicherung und Analyse von großen Datenmengen zu informieren.

Daniel Müller, wissenschaftlicher Mitarbeiter an der Hochschule Offenburg, stellte in seinem Vortrag dar, wie Apache Spark für den ETL-Prozess verwendet werden kann, um dann Daten einzulesen, zu transformieren und in Apache HBase zu speichern. In einem Industrieprojekt bestand die Anforderung, dass eine sehr große Anzahl von Datensätzen persistiert werden muss, wobei jeder Datensatz im Extremfall aus mehreren hundert Merkmalen bestehen kann und die Features von Datensatz zu Datensatz unterschiedlich sind. Die Datensätze enthalten sowohl Messwerte von Sensoren, die den Produktionsprozess überwachen, als auch Daten von Produkttests während des Produktionsprozesses. Nach der Speicherung ist es das Ziel, in einem weiteren Schritt die Daten auszuwerten und prädiktiv die Qualität eines Produkts mit Hilfe von Machine Learning Verfahren zu klassifizieren. Als Lösung für die Speicherung der semistrukturierten Daten wurde die NoSQL Datenbank Apache HBase gewählt. HBase ist eine skalierbare und verteilte Wide Column Store-Datenbank, die auf Apache HDFS aufsetzt. Die im CSV-Format vorliegenden Sensordaten wurden in dem Forschungsprojekt mithilfe von Apache Spark eingelesen, transformiert (ETL-Prozess) und in einer HBase Tabelle persistiert, wo die Daten nun wieder von anderen Tools, z. B. über optimierte Table Scans, ausgelesen werden können.

Im Vortrag wurde neben der allgemeinen Arbeitsweise von HBase auch das praktische Vorgehen wie die Datenuntersuchung, das Tabellendesign, die Implementierung der Spark-Jobs sowie die Stolpersteine des Projekts vorgestellt. Herr Müller ging auch auf Szenarien ein, wie die Analysen in bestehende BI-Systeme über entsprechende Schnittstellen und Tools wie Apache Drill integriert werden können.

Der nächste Roundtable in Freiburg wird am Donnerstag, 26. April, zum Thema Managed Hadoop stattfinden und wird auf XING und tdwi.eu angekündigt. Wir freuen uns auf Ihre Teilnahme.

*Prof. Dr. Stephan Trahasch,
Hochschule Offenburg*