BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

Exposé zur Masterarbeit im Fachbereich VI – Informatik und Medien – der Beuth Hochschule für Technik Berlin zur Erlangung des akademischen Grades **Master of Science (M.Sc.)** im Studiengang **Medieninformatik-Online (Master)**

Big Data Processing mit Apache Spark

Sascha P. Lorenz

Matrikelnummer: 501 63 21

sascha.lorenz@technik-emden.de

Berlin, 27.04.2014

1. Betreuer Prof. Dr. Stefan Edlich
Gutachter Prof. Dr. Schiemann-Lillie

1.1 Einleitung

Gegenstand dieser Arbeit sind die Grundlagen der Verarbeitung und Analyse großer Datenmengen (Big Data). Zunächst sollen verschiedene Ansätze mit Ihren Funktionsweisen sowie den Vor- und Nachteilen diskutiert werdem. Hier werden zuerst allgemeine Grundlagen zu Big Data erarbeitet. Was ist Big Data, was unterscheidet die Verarbeitung von strukturierten und unstrukturierten Daten, Relationale Datenbanken vs. noSQL, wie müssen die Quelldaten für die jeweiligen Verarbeitungen beschaffen sein, welche besonderen Herausforderungen stellen gestreamte Daten an die Verarbeitung. Besonders wird hier auf Hadoop und den Map/Reduce-Algorithmus eingegangen, um das bisher etablierte Vorgehen zu beschreiben und ein grundsätzliches Verständnis für die Domäne "Big Data Processingßu schaffen. In diesem Kontext wird das gesamte Ökosystem rund um Hadoop vorgestellt.
Für diesen Teil der Arbeit werden entsprechende Versuche besonders zu den Themen Quelldatenbeschaffenheit (sowie Beschaffung und Verfügbarmachung), Streaming vs. Persistent und dem Hadoop-Ökosystem auf dem Cluster und Lokal durchgeführt und dokumentiert. (Hinweis für Herrn Schiemann-Lillie: an der Beuth-HS Berlin wurde ein starkes Rechner-Cluster konfiguriert, um dort unter anderem diese Versuche mit Spark und dessen Ökosystem durchführen zu können)
Nachdem eine Einführung in das Thema "Big Data Processingërfolgt ist und ein entsprechend quantitativ und qualitativ brauchbarer Datensatz zur Verfügung steht, werden die Next-Generation Data-Processing Technologien betrachtet. Kernthema ist hier Apache Spark und der gesamte BDAS (Berkeley Data Analytics Stack), der von den den AMP-Labs innerhalb von Apache-Projekten um Spark herum aufgebaut wurde. Jedes der Bestandteile wird innerhalb dieser Arbeit betrachtet und in Versuchen auf dem Cluster erprobt. Zu praktisch jeder öffiziellen"BDAS-Implementierung existieren noch Alternativen. Diese werden ermittelt, gegebenenfalls erprobt und die Unterschiede und Gemeinsamkeiten dokumentiert. Besonders Laufzeiten sind in diesem Umfeld von größter Wichtigkeit. Hier werden entsprechend besonders intensive Versuche durchgeführt, zum Beispiel, in dem auf großen Datensätzen Map/Reduce- oder Logostic-Regression-Algorithmen zum Einsatz kommen. Besonders Apache flink wird hier als Alternative näher untersucht. Auch Applikationen, die auf dem eigentlichen Stack aufsetzen, werden näher betrachtet und entsprechenden Praxistests unterzogen (beispielsweise H2O für statistische Analysen).
Danach wird die API von Spark und deren Möglichkeiten mit Scala, Java und Clojure näher betrachtet und durch jeweils eigene Implementierungen untersucht.
Die Arbeit schließt mit durch verschiedene Versuchsreihen fundierte Empfehlungen für die unterschiedlichen Anforderungen im Bereich des Big-Data-Processing.