Vergleich dreier Implementationsvarianten für eine Analyse von Satellitenbildern

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Arts (B. A.)

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

eingereicht von: Robin Ellerkmann
geboren am: 25.04.1992
in: Berlin

Gutachter: Prof. Johann-Christoph Freytag, Ph.D.
Dipl.-Inf. Mathias Peters

eingereicht am: .....

verteidigt am: .....

Inhaltsverzeichnis

Kapitel 1

Einleitung

Seit einigen Jahren ist ein massiver Anstieg an Datenaufkommen zu beobachten [EMC14]. Diese Entwicklung erfordert neue Technologien und Prozesse, zum Beispiel bei der Speicherung und der Verarbeitung der Daten. Denn traditionelle Datenbanksysteme können große Datenmengen nicht immer in akzeptabler Form und Verarbeitungszeit verarbeiten [Jac09]. Ein zum Zweck der Verarbeitung großer Datenmengen entwickeltes Programmiermodell ist das Map-Reduce Paradigma, das 2004 erstmals publiziert wurde [DG08]. Dieses Paradigma sieht eine massiv parallelisierte Verarbeitung von Daten vor und wird von Datenverarbeitungssystemen wie Hadoop [Foub] und Flink [Foua] implementiert.
Im Rahmen dieser Bachelorarbeit sollen ein traditioneller Ansatz und ein massiv parallelisierbarer Ansatz bei der Verarbeitung von großen Datenmengen untersucht werden. Der Vergleich beider Ansätze wird am Beispiel eines Algorithmus zur Approximierung einer Pixelzeitreihe durchgeführt. Dieser wird im Rahmen des Projekts GeoMultiSens [GP] zur Analyse der Veränderung der Flora in einer geographischen Region genutzt. An die Analyse anschließend werden mithilfe des Algorithmus auf Basis der approximierten Werte Prognosen zur weiteren Entwicklung der Flora der untersuchten Region gestellt.
Es werden drei unterschiedliche Implementierungen des Algorithmus untersucht, die sich hinsichtlich der eingesetzten Technologien und Programmiersprachen unterscheiden. Die Methodik, die der Algorithmus implementiert, ist bei allen untersuchten Varianten identisch. Als Basis wird die bereits implementierte und in der Praxis genutzte Python-Implementation genutzt. Die zweite und dritte Variante werden in Flink implementiert. Diese beiden Varianten unterscheiden sich bezüglich der genutzten Programmiersprache. Zur Implementierung von Variante zwei wird Flinks Java-Schnittstelle genutzt, zur Umsetzung von Variante drei die Python-Schnittstelle. Schließlich werden alle drei Varianten unter identischen Bedingungen getestet. Dies bedeutet, dass sowohl die Testumgebung als auch die Testdaten identisch sein sollen. Ausgehend von den Tests und den ermittelten Ergebnissen wird eine Bewertung der drei Implementierungsvarianten des Algorithmus vorgenommen werden.

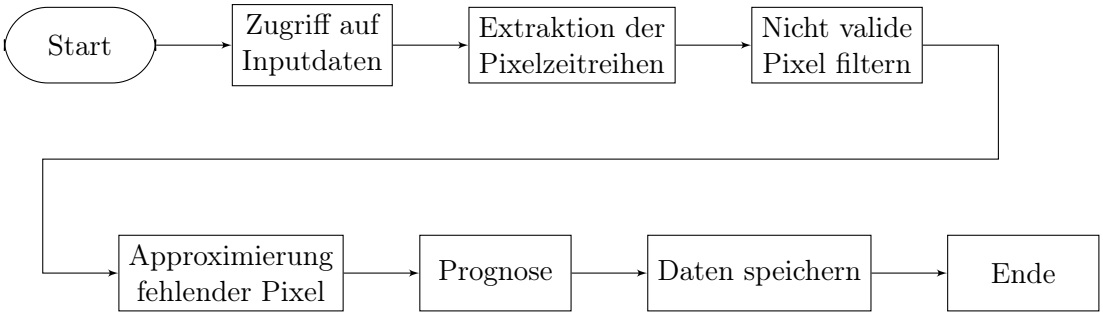
Kapitel 2

Grundlagen

Das sind die Grundlagen

2.1 Algorithmus

Beschreibung des Algorithmus, GeoMultiSens, GfZ,



2.2 Python

Python ist eine quelloffene und universell einsetzbare Programmiersprache, die seit 1989 existiert und fortwährend weiter entwickelt wird. Prägende Eigenschaften der Sprache sind unter anderem eine dynamische Typisierung von Variabeln, eine simpel gehaltene Syntax und die Erweiterbarkeit durch Module und Bibliotheken. Es ist auch möglich Python-Code durch C- beziehungsweise C++-Bibliotheken zu erweitern [Mar06]. Dies ermöglicht eine verkürzte Ausführungszeit eines Programms, insbesondere bei rechenintensiven Programmabschnitten. Ein Schwachpunkt von Python im Bezug auf die schnelle Verarbeitung großer Datenmengen ist die nicht auf automatisierte Parallelisierung ausgelegte Struktur. Daraus folgt eine unzureichende Skalierbarkeit, sobald Daten, deren Größe die Arbeitsspeichergröße der ausführenden Maschine übersteigt, verarbeitet werden müssen. (Auf weiter oben genannten Punkt der Großen Datenmengen eingehen).

2.3 Java

2.4 Flink

Kapitel 3

Evaluation

Beschreibung und Bewertung der Ergebnisse meiner Untersuchungen

Kapitel 4

Fazit

Fazit und Ausblick

Literaturverzeichnis

[DG08] Jeffrey Dean and Sanjay Ghemawat. Mapreduce. *Communications of the ACM*, 51(1):107, Jan 2008.

[EMC14] EMC². The digital universe of opportunities. Technical report, EMC², 2014.

[Foua] Apache Software Foundation. Flink website.

[Foub] Apache Software Foundation. Hadoop website.

[GP] GfZ-Potsdam. Geomultisens.

[Jac09] Adam Jacobs. The pathologies of big data. *Communications of the ACM*, 52(8):36, August 2009.

[Mar06] Alex Martelli. *Python in a Nutshell. A Desktop Quick Reference*. O'Reilly, second edition edition, 2006.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe. Weiterhin erkläre ich, eine ...arbeit in diesem Studiengebiet erstmalig einzureichen.

Berlin, den 3. Juli 2015 .....

Statement of authorship

I declare that I completed this thesis on my own and that information which has been directly or indirectly taken from other sources has been noted as such. Neither this nor a similar work has been presented to an examination committee.

Berlin, 3rd July 2015 .....