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

2.1 Grundlagen der Satellitenbildanalyse

2.1.1 Fernerkundung mithilfe des Landsat-Satellitensystems

Als Fernerkundung wird „die Gesamtheit der Verfahren zur Gewinnung von Informationen über die Erdoberfläche oder anderer nicht direkt zugänglicher Objekte durch Messung und Interpretation der von ihr ausgehenden (Energie-) Felder“ [Deu12] verstanden. Fernerkundungssatelliten verfügen über verschiedene Aufnahmesysteme, die durch multispektrale Messungen von emittierter elektromagnetischer Strahlung eine berührungsfreie Beobachtung der Erdoberfläche ermöglichen. Bei der multispektralen Messung werden von Sensoren registrierte spektrale Signaturen einzelnen Bereichen des elektromagnetischen Spektrums zugeordnet. Das Resultat sind mehrere spektrumsspezifische, simultan aufgenommene Satellitenbilder, die nur das aufgefangene Licht eines spezifischen Spektralbereichs, auch Spektralband genannt, zeigen. Die Art und Qualität der Aufnahmesensoren ist dabei abhängig vom Typ des Satelliten.

Die Ausgangsdaten für die Untersuchungen in dieser Bachelorarbeit wurden von Satelliten des Landsat-Satellitensystems aufgenommen. Der erste Landsat-Satellit Landsat 1 wurde 1972 gestartet. Seitdem wurden die Sensoren und die Satelliten kontinuierlich weiterentwickelt. Aktuell sind Landsat 7 und, im Rahmen der Landsat Data Continuity Mission, Landsat 8 im Einsatz. Landsat 8 nutzt in der aktuellen Generation zwei verschiedene Instrumente zur Fernerkundung. Den Operational Land Imager (OLI) und die Thermal Infrared Sensors (TIRS).

Der OLI erfasst emittierte elektromagnetische Strahlung im Spektralbereich von 0,433 µm bis 1,390 µm unterteilt in acht Spektralkanäle sowie einen panchromatischen Kanal. Es werden mehr als 7000 Detektoren pro Spektralband genutzt, um eine bessere Bildqualität zu bieten als frühere Systeme [MSWI04]. Neben den klassischen Farbspektren Blau, Grün und Rot nutzt Landsat-8 ein weiteres Band, das speziell für die Fernerkundung von Küsten genutzt wird. Außerdem verfügt Landsat-8 über drei Infrarotbänder, die nahes und mittleres Infrarotlicht registrieren, sowie ein weiteres Infrarotband, das auf die Beobachtung von Cirruswolken spezialisiert ist. Der panchromatische Kanal registriert elektromagnetische Strahlung mit Wellenlängen von 0,500 µm bis 0,680 µm. Dieser Spektralbereich entspricht

etwa dem des menschlichen Auges. Aufgrund des, im Vergleich zu den einzelnen Farbfrequenzbändern, breiten abgedeckten Spektralbereichs ist eine höhere Auflösung der Bilder möglich.

Die Thermal Infrared Sensors (TIRS) [Cha11] umfassen zwei Thermalkanäle. Diese erfassen im Gegensatz zu den Multispektralkanälen elektromagnetische Emissionen mit Wellenlängen zwischen 10,30 µm und 12,50 µm, also langwellige Infrarotstrahlung. Dies ist insbesondere für die Beobachtung von Wolken nützlich. Die Kantenlänge der einzelnen Pixel beträgt 100 Meter. Diese kann nachträglich auf 30 Meter angeglichen werden, um eine bessere Kompatibilität mit den Aufnahmen der Multispektralbänder zu gewährleisten.

Landsat 8 sendet pro Tag 400 Aufnahmen der Erdoberfläche, auch Szenen genannt, an die Bodenstation. Eine Aufnahme zeigt dabei eine geographische Region der Erde mit einer Ost-West-Ausdehnung von 185 Kilometer. Dies entspricht 100 nautischen Meilen. Die Nord-Süd-Ausdehnung einer Szene beträgt circa 174 Kilometer

Durchschnittlich wird jede Region der Erde alle 16 (?) Tage überflogen [IDB12].

Die von Landsat-Satelliten aufgezeichneten und übermittelten Bilder müssen jedoch vor der Durchführung von Analysen aufbereitet werden.

2.1.2 Aufbereitung und Analyse von Satellitenbildern

Die durch die Landsat-Satelliten aufgezeichneten und an die Bodenstationen übermittelten Szenen müssen vor ihrer Nutzung aufbereitet werden. Dadurch wird im Allgemeinen die Bildqualität verbessert, da externe Störfaktoren und eventuelle interne Fehlfunktionen ausgeglichen werden können. Es wird zwischen radiometrischen und die geometrischen Aufbereitungen unterschieden. Bei der radiometrischen Aufarbeitung werden digitale Werte wie zum Beispiel die Helligkeit der Szene angepasst. Eventuelle durch die Atmosphäre verursachte Verschlechterungen sollen verbessert werden, um ein genaueres Satellitenbild zu erhalten. Techniken um diese Verbesserung zu erreichen sind beispielsweise das Strahlungstransfermodell, die bildbasierte atmosphärische Korrektur und die Histogramm-Minimum-Methode. Es ist individuell von der Szene und den zur Verfügung stehenden Metadaten abhängig, mit welcher Methode die nützlichste Verbesserung erreicht werden kann.

Im Rahmen der geometrischen Aufbereitung sollen die Folgen einer eventuellen Fehlpositionierung des Satelliten korrigiert werden. Um die Szenen sinnvoll analysieren zu können, müssen sie korrekt und genau positioniert sein. Dies gilt insbesondere bei der Analyse mehrerer Szenen derselben geographischen Gegend. Um eine normierte Positionierung der Szenen zu schaffen, werden aus jeder Szene, die einen Teil der zu analysierenden geographischen Region beinhaltet, quadratische Teile der Originalszene ausgeschnitten. Dann wird jeder Pixel der Kachel auf die Zugehörigkeit zum Zielgebiet geprüft. Wenn ein Pixel relevant ist, wird er anhand seiner, aus der Position des Satelliten zum Aufnahmezeitpunkt ermittelten, Position in einem finalen Bild hinzugefügt.

Durch die zunehmend bessere Qualität von Satellitenbildern, die durch Fernerkundungsatelliten aufgezeichnet werden [MSWI04], können detailliertere Analysen getätigt werden. Dabei werden die aufbereiteten Satellitenaufnahmen als Ausgangspunkt genutzt. Im Fall der im Rahmen dieser Bachelorarbeit genutzten Analyse der Veränderung der Flora in einer geographischen Region werden die Ausgangsdaten mithilfe eines

2.2 Parallele Datenverarbeitungssysteme

Um die seit mehreren Jahren massiv ansteigenden Datenmengen [EMC14] zu verarbeiten wird zunehmend eine verteilte Verarbeitung dieser Daten populär. Dazu werden mehrere Maschinen zu einem Netzwerk, einem sogenannten Cluster, zusammengeschlossen. Diese Computer wären als einzelne Maschine nicht in der Lage ein großes beziehungsweise komplexes Problem in akzeptabler Zeit zu lösen. Die Leistungsfähigkeit des Netzwerks wird jedoch nicht über die Leistung einer einzelnen Maschine sondern primär über die Menge der zusammengeschlossenen Computer gesteuert. Dies hat mehrere Vorteile gegenüber der Verarbeitung mithilfe einzelner, besonders leistungsstarker Maschinen. Die wichtigsten Vorteile parallelisierter Systeme sind ihre Skalierbarkeit sowie die Fehlertoleranz. Falls mehr Rechenleistung benötigt wird oder wenn Teile des Netzwerks nicht funktionsfähig sind lassen sich neue Maschinen kurzfristig, meist auch im laufenden Betrieb, in das bestehende Netzwerk integrieren. Bei einzelnen, sehr leistungsstarken Computern gestaltet sich beides aufgrund der abgeschlossenen Beschaffenheit der Maschine schwierig. [Quelle]

Eine Big-Data Anwendung zeichnet sich durch drei Eigenschaften aus. Diese drei Charakteristika sind die Größe (engl. Volume), die Komplexität (eng. Variety) und die echtzeitnahe Verfügbarkeit sowie schnelle Verarbeitung (engl. Velocity) der Daten [Lan01]. Eine weiteres Merkmal ist die nicht garantierte Zuverlässigkeit und Einheitlichkeit der Daten (engl. veracity) [ZdP+12]. In den letzten Jahren hat sich das global produzierte Datenaufkommen massiv gesteigert. Insbesondere die zunehmende Zahl der Internetnutzer sowie die Verbreitung von Smartphones trägt zu dieser Entwicklung bei. Ebenso trägt die zunehmende Digitalisierung der Industrie sowie die zunehmende Verbreitung von Sensoren jeglicher Art zu diesem Anstieg bei. Aber auch in nicht kommerziellen Bereichen wächst die Datenmenge. Die Satellitenbilder der aktuellen Generation des Landsat-Satellitensystems produziert Aufnahmen, die dreimal soviel Speicherplatz benötigen wie die der vorigen Generation. Projekte wie das Sloan Digital Sky Survey produzieren täglich etwa 200 Gigabyte. Insgesamt werden die Datenmengen weiter massiv zunehmen. Für das Jahr 2020 wird eine weltweites Datenaufkommen von 44 Zettabyte prognostiziert [EMC14]. Aus dieser steigenden Datenmenge ergeben sich auch Folgen für Daten verarbeitende Dienste. Es müssen sehr viel mehr Daten auf einmal verarbeitet werden. Darüber hinaus sind die zu verarbeitenden Daten zunehmend vielfältiger und unstrukturierter. Die verarbeitenden Algorithmen und die Speicherstrukturen müssen also hinreichend auf unvollständige beziehungsweise fehlerhafte Datensätze reagieren können und diese trotzdem bestmöglich verarbeiten. [Quellen] [Beschreibung für velocity einfügen]. Wenn eine Anwendung eine Datenmenge verarbeitet, die mindestens einige der vier Kriterien nach [Lan01] erfüllt, gilt diese Anwendung als Bi-Data Anwendung.

Eigenschaften von: Big Data, DBMS, Grundlagen für Flink, Erwähnung MapReduce Prinzip

2.3 Programmierabstraktionen

2.3.1 Apache Flink

Eigenschaften + Operatoren in Flink

Apache Flink ist ein System[Framework], das auf eine massiv parallelisierte Verarbeitung Vorher einführen, 2.2.1 von großen Datenmengen spezialisiert ist. Es ging 2014 [Quelle] aus Stratosphere hervor, das seit 2010[Quelle] kooperativ von Forschern verschiedener Universitäten entwickelt wurde [ABE+14]. Seit Januar 2015 ist Flink ein Top-Level Projekt der Apache Software Foundation [Fou15].

Die Hauptkomponenten des Systems sind die Flink-Laufzeitumgebung und der Flink-Optimierer. Der Flink-Optimierer erhält einen azyklischen Graphen von Flink-Operatoren als Eingabe. Dieser wird mithilfe von Techniken der traditionellen Optimierung von relationalen Anfragen optimiert. [Weitere Details aus Stratosphere Paper?, unter welchen Gesichtspunkten wird DAG optimiert?]. Der optimierte Datenflussgraph, auch Jobgraph genannt, besteht aus mehreren, teilweise unabhängig voneinander zu verarbeitenden Arbeitsschritten. Diese können teilweise parallel bearbeitet werden [MapReduce erwähnen?]. Dieser optimierte Datenflussgraph wird an die Flink-Laufzeitumgebung weitergegeben.

Flink erweitert das Map-Reduce Paradigma um weitere Operatoren. [Operatoren beschreiben]

2.3.2 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 Variablen, 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 Architektur. Daraus resultiert 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). Bez. der Eignung zur Lösung solcher Probleme.

Kapitel 3

Beschreibung und Umsetzung des Algorithmus zur Analyse von Pixelzeitreihen

3.1 Beschreibung des Algorithmus zur Analyse von Pixelzeitreihen

Beschreibung der Vorgehensweise bei der Analyse (Zhu, SVR), Ziel der Analyse, Entwicklungsgeschichte der Analysetechnik

3.2 Umsetzung des Algorithmus mit Apache Flink

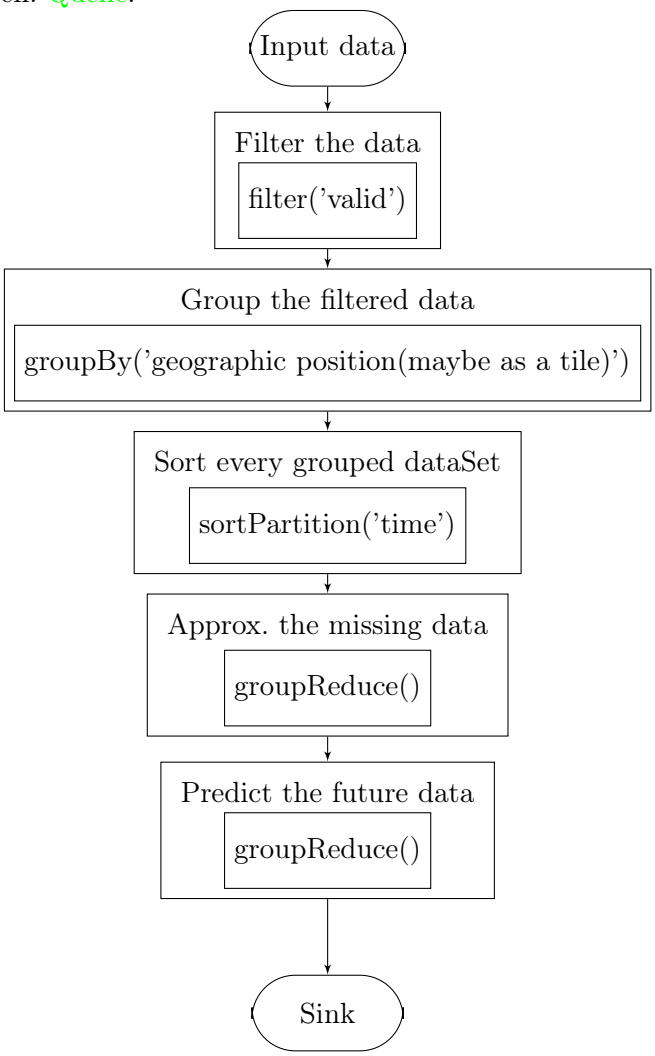
3.2.1 Nutzung der Java-Programmierschnittstelle

3.2.2 Nutzung der Python-Programmierschnittstelle

3.3 Umsetzung des Algorithmus in Python

Die Analyse von Satellitenbildern erfordert die Verarbeitung großer Mengen komplexer Rohdaten, die nahezu in Echtzeit verfügbar sind. Aufgrund dieser Charakteristika handelt es sich bei dieser Analyse um ein Big-Data Problem. Denn alle vier Kriterien, die ein solches charakterisieren sind erfüllt.
Bei der Analyse von Satellitenbildern sind die Merkmale Datengröße und Datenkomplexität sowie die schnelle Verarbeitung der Daten von Bedeutung. Abhängig von der Anzahl der genutzten Bilder sind die zu verarbeitenden Datenmengen sehr groß. Ein Bild besitzt im Regelfall abhängig vom Satellitenmodell, das die Aufnahme gemacht hat, eine Größe von 750 Megabyte bis zu 1,5 Gigabyte. Um eine Entwicklung zu untersuchen werden jedoch viele dieser Bilder in die Untersuchung mit einbezogen, so dass die zu verarbeitende Datenmenge kontinuierlich ansteigt. Dieser kontinuierliche Anstieg entsteht dadurch, dass aktuell mehrere Satelliten mit der Fernerkundung der Erde fortfahren und so in kurzen In-

tervallen neue Bilder zur Verfügung stehen, die im Rahmen der Analyse verwendet werden sollen. Quelle.



Kapitel 4

Evaluierung

4.1 Versuchsbeschreibung

Beschreibung + Begründung für meine Versuchsbedingungen

4.2 Auswertung

Beschreibung und Bewertung der Ergebnisse meiner Untersuchungen

Kapitel 5

Fazit

Fazit und Ausblick z.b. Vergleich mit anderen Untersuchungen

Literaturverzeichnis

[ABE+14] Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao, Marcus Leich, Ulf Leser, Volker Markl, and et al. The stratosphere platform for big data analytics. *The VLDB Journal*, 23(6):939,964, May 2014.

[Cha11] Anju Chaudhary. Thermal infrared sensors. *Encyclopedia of Snow, Ice and Glaciers*, page 1156, 2011.

[Deu12] Deutsches Institut für Normung e.V. *Photogrammetrie und Fernerkundung - Begriffe*, 8 2012. Rev. 3.

[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. `https://flink.apache.org/`.

[Foub] Apache Software Foundation. Hadoop website. `https://hadoop.apache.org/`.

[Fou15] Apache Software Foundation. The apache software foundation announces apache™ flink™ as a top-level project. `https://blogs.apache.org/foundation/entry/the\_apache\_software\_foundation\_announces69`, January 2015.

[GP] GfZ-Potsdam. Geomultisens website. `http://www.geomultisens.gfz-potsdam.de/`.

[IDB12] James R. Irons, John L. Dwyer, and Julia A. Barsi. The next landsat satellite: The landsat data continuity mission. *Remote Sensing of Environment*, 122:11,21, Jul 2012.

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

[Lan01] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *Application Delivery Strategies published by META Group Inc.*, Feb 2001.

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

[MSWI04] B.L. Markham, J.C. Storey, D.L. Williams, and J.R. Irons. Landsat sensor performance: history and current status. *IEEE Transactions on Geoscience and Remote Sensing*, 42(12):2691,2694, Dec 2004.

[ZdP+12] Paul Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, James Giles, and David Corrigan. *Harness the Power of Big Data The IBM Big Data Platform*. McGraw-Hill Osborne Media, 2012.

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 30. August 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, 30th August 2015 .....