Hochschule fÃijr Technik, Wirtschaft und Kultur Leipzig

FakultÃd't Informatik, Mathematik und Naturwissenschaften

Masterstudiengang Informatik

Masterarbeit

zur Erlangung der akademischen Grades

Master of Science (M.Sc.)

Untersuchung und Optimierung verteilter Geografischer Informationssysteme zur Verarbeitung Agrartechnischer Kennzahlen

Eingereicht von: Kurt Junghanns

Matrikelnummer: 59886

Leipzig 10. Oktober 2014

Erstprüfer: Prof. Dr. rer. nat. Thomas Riechert

Zweitprüfer: M. Sc. Volkmar Herbst

Abstrakt

Danksagung

Vorwort

Glossar

Computer is a programmable machine that receives input, stores and manipulates data, and provides output in a useful format

Abkürzungsverzeichnis

ACID Atomicity, Consistency, Isolation und Durability

BASE Basically Available, Soft state, Eventual consistency

GIS Geoinformationssystem

MVCC Multi Version Currency Control

Abbildungsverzeichnis

Tabellenverzeichnis

Inhaltsverzeichnis

Glossar iv

Abkürzungsverzeichnis v

Abbildungsverzeichnis vi

Tabellenverzeichnis vii

1 Einleitung 1

1.1 Motivation . . . 1
1.2 Zielsetzung . . . 1

2 Grundlagen 3

2.1 Datenbank . . . 3
2.1.1 ACID . . . 3
2.1.2 MVCC . . . 3
2.1.3 BASE . . . 3
2.1.4 weitere Begriffsdefinitionen . . . 4
2.1.5 Indexstrukturen . . . 4
2.1.6 Mehrrechner-Datenbanksystem . . . 4
2.1.7 Verteiltes Datenbanksystem . . . 4
2.1.8 Replikationsverfahren . . . 4
2.2 geografische Datenverarbeitung . . . 5
2.2.1 Bezugssysteme . . . 5
2.2.2 Datenformate . . . 5
2.2.3 GIS . . . 5

2.2.4 PostGIS . . . 5
2.2.5 GeoTools . . . 5
2.3 NoSQL . . . 5
2.3.1 Definition . . . 5
2.3.2 Kategorisierung . . . 5
2.3.3 Hadoop . . . 5
2.3.4 Accumulo . . . 5
2.3.5 NoSQL GIS . . . 6
2.3.6 MongoDB . . . 6
2.3.7 CouchDB . . . 6
2.3.8 Neo4J . . . 6
2.3.9 Rasdaman . . . 7
2.3.10 Spacebase . . . 7
2.3.11 Geomesa . . . 7
2.4 Leistungstests . . . 8

3 Ausgangsszenario 9

4 System 1 10
4.1 Aufbau . . . 10
4.2 Installation . . . 10
4.3 Datenimport . . . 10
4.4 Verarbeitung . . . 10
4.5 Schnittstelle . . . 10
4.6 Leistungstests . . . 10

5 Gegenüberstellung 11
5.1 Kosten . . . 11
5.2 Umfang . . . 11
5.3 Leistung . . . 11

6 Fazit 12
6.1 Zusammenfassung . . . 12
6.2 Wertung . . . 12

6.3 Ausblick 12

Literaturverzeichnis I

1 Einleitung

1.1 Motivation

Die Agri Con GmbH verwaltet als Akteur im Bereich „Precision Farming“ täglich mehrere Millionen geografische Punktdaten. Diese Daten werden von aktiven Landwirtschaftsmaschinen und durch die Verarbeitung durch firmeninterne und firmenexterne Mitarbeiter sowie Systeme erzeugt. Weiterhin fallen dadurch indirekt Vektor- und Rasterdaten an, welche gespeichert und anschließend verarbeitet werden müssen. Aus den Quelldaten werden Vektordaten für beispielsweise Verteilung der Grunddüngung erzeugt. Rasterdaten werden für „N-Düngung“ verwendet, was unter anderem die Biomasse, die Nährstoffaufnahme und die Nährstoffverteilung beinhaltet. Diese Menge an Daten ist essentiell für den Betrieb, weshalb diese strukturiert gespeichert und kostengünstig verarbeitet werden müssen. Nicht nur Agri Con steht vor dieser Notwendigkeit, sondern der Großteil der Unternehmen, die sich mit komplexen Geodaten beschäftigen. wie Monsanto, Google, Facebook, ESRI, OpenGEO, etc.

1.2 Zielsetzung

Eine PostgreSQL Installation auf einem Computersystem stößt bei der aktuellen Nutzung durch Agri Con an die Leistungsgrenze. Aus diesem Grund ist die Speicherung und erste Verarbeitung in ein anderes System auszulagern. Dafür sind existierende Geoinformationssystem (GIS) zu untersuchen und deren Eignung für den in Kapitel 3 beschriebenen Anwendungsfall festzustellen. Der Schwerpunkt der Untersuchung sind die Möglichkeiten und Leistungsfähigkeit der räumlichen Datenverarbeitung. Dabei werden

NoSQL und Open-Source Systeme höher gewichtet. Aus geeigneten Systemen werden bis zu 3 ausgewählt. Die Auswahl wird speziell untersucht und eine prototypische Installation[1] erstellt. Somit sollen die Systeme mit dem Ist-Stand unter den Faktoren Kosten, Funktionalität und Leistungsfähigkeit verglichen werden.

Zu Beginn werden theoretischen Grundlagen zu Datenbanken, geographischer Datenverarbeitung, NoSQL und Leistungstests festgehalten. Anschließend definiert Kapitel 3 das Ausgangsszenario, für welches die Systeme analysiert und getestet werden sollen. Die darauf folgenden Kapitel stellen die ausgewählten Systeme unter den Gesichtspunkten Aufbau, Installation, Datenimport, Verarbeitung, Schnittstelle und Leistungstest dar.

Das vorletzte Kapitel stellt die vorgestellten Systeme direkt gegenüber und führt die Daten zu Kosten, Umfang und Leistung auf. Die Thesis endet mit einer Zusammenfassung, einer Empfehlung bzw. Wertung der Ergebnisse und einem Ausblick auf die zukünftige Handhabung der räumlichen Daten bei Agri Con.

[1] Dabei kann eine Installation aus mehreren Systemen bestehen und eigens implementierte Funktionalitäten enthalten

2 Grundlagen

Computer

2.1 Datenbank

2.1.1 ACID

Atomicity, Consistency, Isolation und Durability (ACID)

2.1.2 MVCC

Multi Version Currency Control (MVCC)

2.1.3 BASE

Basically Available, Soft state, Eventual consistency (BASE)

2.1.4 weitere Begriffsdefinitionen

2.1.5 Indexstrukturen

R-Baum

B-Baum

LSM-Baum

Geohash

2.1.6 Mehrrechner-Datenbanksystem

2.1.7 Verteiltes Datenbanksystem

2.1.8 Replikationsverfahren

Synchron

Asynchron

Kaskadiert

2.2 geografische Datenverarbeitung

2.2.1 Bezugssysteme

2.2.2 Datenformate

Punkte

Vektoren

Raster

Shapefile

2.2.3 GIS

2.2.4 PostGIS

2.2.5 GeoTools

2.3 NoSQL

2.3.1 Definition

2.3.2 Kategorisierung

2.3.3 Hadoop

2.3.4 Accumulo

https://en.wikipedia.org/wiki/Apache_Accumulo

2.3.5 NoSQL GIS

2.3.6 MongoDB

2.3.7 CouchDB

2.3.8 Neo4J

2.3.9 Rasdaman

http://live.osgeo.org/de/overview/rasdaman_overview.html :
- Array-Datenbanksystem - PostgreSQL Aufsatz - Multi-Dimensionalität - eigene Anfragesprache - skalierend - unterstützt WCS Core und WCPS - Implementierte Standards: OGC WMS 1.3, WCS 2.0, WCS-T 1.4, WCPS 1.0, WPS 1.0 - Lizenz: Clients und APIs: GNU Lesser General Public License (LGPL) version 3; Server-Engine: GNU General Public License (GPL) version 3 - Unterstützte Plattformen: Linux, MacOS, Solaris - APIs: rasql, C++, Java

http://www.rasdaman.org/ :
- open-source - "extends standard relational database systems with the ability to store and retrieve multi-dimensional raster data"

http://www.rasdaman.de/ :
- "erlaubt die Ablage von unbeschränkt grossen multi-dimensionalen Arrays ("Rasterdaten") in einer konventionellen Datenbank"

2.3.10 Spacebase

http://docs.paralleluniverse.co/spacebase/ :
- serverseitig - in-memory - spatial data store - ausgelegt für viele rechner und hohen Durchsatz (real-time) - 2D und 3D Objekte mit 3D bbox - load balancing enthalten - spatial querys möglich - benötigt JVM - API für Java, Ruby, Python, Node.js, C++, Erlang - API stellt nur elementare spatial querys zur verfügung: intersect oder contains - eigene spatial querys können definiert werden

2.3.11 Geomesa

- Ingest = Import über Kommandozeile (geomesa-tools) - Ingest von shp, csv und tsv Dateien - Anderer Dateiimport mit GeoTools - Verarbeitung nur über externe Tools (Spark, geotools) - Export: csv, tsv, shp, geojson, gml

http://www.eclipse.org/community/eclipse_newsletter/2014/march/article3.php :
- open-source - build on Accumulo and Hadoop - Supporting the GeoTools API - GeoServer Plugin - geohash for indexing

https://www.locationtech.org/proposals/geomesa :
- outperforming postgis with geoserver

http://de.slideshare.net/CCRinc/location-techdc-talk2-28465214 - Verwendung fraktaler Kurven - mit Spark und Scalding wesentlich schneller als PostGIS

https://docs.google.com/presentation/d/1NO0ppk8MfDs8Q-QcUidZCSZK7YYwd9RjJoHV1V4Yq_w/edit?pli=1#slide=id.p :
-

2.4 Leistungstests

- siehe BA - in Absprache mit Prof. Riechert

3 Ausgangsszenario

4 System 1

4.1 Aufbau

4.2 Installation

4.3 Datenimport

4.4 Verarbeitung

4.5 Schnittstelle

4.6 Leistungstests

5 Gegenüberstellung

5.1 Kosten

5.2 Umfang

5.3 Leistung

6 Fazit

6.1 Zusammenfassung

6.2 Wertung

6.3 Ausblick

Literaturverzeichnis

Eidesstatliche Erklärung

Ich versichere, dass die Masterarbeit mit dem Titel „...“ nicht anderweitig als Prüfungsleistung verwendet wurde und diese Masterarbeit noch nicht veröffentlicht worden ist. Die hier vorgelegte Masterarbeit habe ich selbstständig und ohne fremde Hilfe abgefasst. Ich habe keine anderen Quellen und Hilfsmittel als die an-gegebenen benutzt. Diesen Werken wörtlich oder sinngemäß entnommene Stellen habe ich als solche gekennzeichnet.

Leipzig, 10. Oktober 2014 Unterschrift