agricon

Exposee zur Masterarbeit

Performante Speicherung und Verarbeitung von Geodaten

Kurt Junghanns (59886)

Leipzig, der 19. September 2014

Betreuer: M. Sc. Volkmar Herbst

Inhaltsverzeichnis

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 firmenexterene Mitarbeiter sowie Systeme erzeugt. Weiterhin fallen dadurch 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.

2 Fragestellung

Täglich anfallende geographische Daten sollen in einer zu definierenden Umgebung im Gigabytebereich persistent gespeichert werden. Außerdem sollen diese Daten mit geringer Laufzeit aggregiert und verarbeitet werden. Die Verarbeitung der Daten soll anhand der geographischen Informationen mit Hilfe dafür vorgesehener Funktionen der Umgebung erfolgen. Das Ergebnis der Verarbeitung soll stets eine Menge von aufbereiteten Geodaten sein, welche von GIS direkt interpretiert werden können.

Die Daten fallen dynamisch im Tagesbetrieb als Raster- sowie Vektordaten an. Die Umgebung muss große Mengen an Daten raumbezogen ablegen können.

An die Datenhaltung werden folgende Anforderungen gestellt:

- gleichzeitige und verteilte Datenablage auf mehreren virtuellen Maschinen
- Erhalt der geographischen Informationen der Daten
- persistente Datenhaltung von statischen und dynamischen Daten
- Versionierung der dynamischen Daten
- performante Aggregation soll bereits durch intelligente Datenablage ermöglicht werden

Gespeicherte Daten sind speziell gefiltert zu aggregieren. Dabei soll die Filterung anhand von Meta- und Geoinformationen erfolgen. Neben der Datenhaltung soll auch die Verarbeitung verteilt durchführbar sein. Die Laufzeit ist eine essentielle Anforderung, wie auch der Umfang der geographischen Funktionen, wobei die Qualität der Funktionen die Laufzeit wesentlich beeinflusst.

Raumbezogene OpenSource Systeme bestehen zumeist aus PostgreSQL mit PostGIS, MariaDB, CouchDB mit GeoCouch oder MongoDB als Datenbank und einer beliebigen Middleware. Werden jedoch komplexe raumbezogene Datentypen und Funktionen benötigt, ist nur PostgreSQL mit PostGIS in solchen Systemen anzutreffen. PostgreSQL ist jedoch nicht für die verteilte Verwendung konzipiert, sodass nur vertikal skaliert werden kann.

Einerseits sind die Alternativen zu PostgreSQL zu untersuchen, andererseits die Eignung von NoSQL Systemen herauszuarbeiten.

3 Stand der Forschung

In der Agri Con GmbH werden täglich angefallene Daten nachts verarbeitet und stehen teilweise nur in verarbeiteter Form bereit. Zur Datenhaltung und -verarbeitung wird PostGreSQL mit der Erweiterung PostGIS verwendet. Auch R wird als Programmiersprache zur Verarbeitung eingesetzt.

Denkbare Systeme sind folgende:

- PostGIS mit Optimierung durch Reorganisation der Datenbank und verteilte Verwendung[1]
- MongoDB mit spatial Extension[2]
- Spacebase
- Hadoop, Hive und ArcGis[ESR12]
- Hadoop mit einer spatial Datenbank und spatial Extension
- AsterixDB mit spatial Extension

Momentan existieren komplette Systeme oder Teilsysteme für das beschriebene Szenario. Eine Lösung ist clustering von PostGIS. Eine andere ist Hadoop-GIS, ein komplettes System welches Hadoop verwendet und eine eigene Engine zur Verarbeitung bereitstellt, oder SpatialHadoop, welches einen Hadoop Aufsatz zur Verarbeitung darstellt.

4 Methodik und Vorgehen

Für dieses Szenario sind die technischen Möglichkeiten zu analysieren und zu vergleichen. Darin sollen besonders die Möglichkeiten der geographischen Speicherung und Verarbeitung der Daten und deren Beziehungen betrachtet werden. Der Vergleich soll anschließend in einer Empfehlung für das dargestellte Szenario münden. Damit die Aussagefähigkeit des Vergleiches und der Empfehlung für die Umsetzung in der Agri Con GmbH hoch ist, sind Studien in Form von empirischen Leistungsvergleichen der Laufzeit und des Funktionsumfanges durchzuführen.

Die in der Umgebung vorhandenen geographischen Funktionen sind zu nutzen und fehlende zu implementieren oder durch Erweiterungen verfügbar zu machen.

5 Zeitplan

Die Thesis wird zum ersten Oktober 2014 angemeldet und ist spätestens am 31 März abzugeben.

[1] Dazu zählt Multi-Master Replikation, Load-Balancing und clustering

[2] vorhandene oder mit GDAL Treibern ein Mapping zwischen Geometrieobjekten und BSON erstellen (plus eigenen Indexen und Analyse- und Verarbeitungsfunktionen)

6 vorläufige Gliederung

1. Einleitung
1.1 Motivation
1.2 Zielsetzung
2. Grundlagen
2.1 geografische Datenverarbeitung
2.1.1 Bezugssysteme
2.1.2 Geometriearten
2.1.3 GIS
2.1.4 PostGIS
2.2 NoSQL
2.2.1 Abgrenzung zu relationalen Systemen
2.2.2 NoSQL GIS
2.3 Leistungstests
3. Ausgangssituation
4. System 1
4.1 Überblick
4.2 Datenhaltung
4.3 Verarbeitung
4.4 Testumgebung
4.5 Zusammenfassung
5. System 2[Wik14]
6. System 3[rG14]
7. Vergleich
8. Schlussfolgerung

Literatur

[ESR12] ESRI. Mongodb example code for adding a nosql plug-in data source, Mai 2012.

[rG14] rasdaman GmbH. Rasdaman - raster data manager, 2014.

[Wik14] Wikipedia. Scidb, 2014.