

Informatik

Geodatenprozessierung mit Budget Instanzen

Semesterarbeit



Abbildung 1: Eine Tour mit Budget Instanzen [Web10]

| | |
|------------------|--------------------------------|
| Departement: | Informatik |
| Kurs: | CAS CLD FS20 – Cloud Computing |
| Autor: | Tobias Reber |
| Experte: | Jörg Thomann |
| Ansprechpartner: | Christoph Böcklin |
| Datum: | 06. 08. 2020 |

Perplexity
is the beginning of knowledge.

- Kahlil Gibran

[AD19, S. 33]

Management Summary

In der Kürze liegt die Würze [Doe00]

Inhaltsverzeichnis

| | |
|--|-----------|
| 1 Einführung | 1 |
| 1.1 GIS - Geographical Information Systems | 1 |
| 2 Ausgangslage | 2 |
| 2.1 swisstopo bei AWS | 2 |
| 2.2 Publikation von Geodaten | 2 |
| 2.3 Web-Services | 2 |
| 2.4 Exkurs 3D Daten | 3 |
| 3 Problemstellung | 4 |
| 3.1 Ist-Zustand der 3D Datenpublikation | 4 |
| 3.1.1 Prozess der Publikation | 4 |
| 3.2 Aufwand der Prozessierung | 4 |
| 3.3 Technologie Stack | 4 |
| 4 Vorgehen | 4 |
| 4.1 Arbeitsmethodik | 4 |
| 4.2 Projektplan | 5 |
| 4.3 Vorarbeiten | 5 |
| 4.3.1 AWS Account | 5 |
| 4.3.2 AWS Spot Instances | 5 |
| 4.3.3 Kubernetes testing | 5 |
| 5 Use Case | 6 |
| 5.1 Beschreibung der Technischen Komponenten | 6 |
| 5.1.1 SPOT Instanzen | 6 |
| 5.2 Abgrenzung | 6 |
| 6 Architektur | 7 |
| 6.1 Bewertungskriterien | 7 |
| 6.2 Variante 1: Spot | 7 |
| 6.3 Variante 2: Spot mit EKS | 7 |
| 6.4 Variante 3: AWS Batch | 7 |
| 7 Prototyp | 8 |
| 7.1 Realisierung | 8 |
| 8 Evaluation | 8 |
| 8.1 Erfahrungen | 8 |
| 8.2 Kritische Punkte | 8 |
| 9 Zusammenfassung | 8 |
| 10 Ausblick | 8 |
| Literaturverzeichnis | 9 |
| A Fachbegriffe und Abkürzungen | I |
| B Konfigurationen | II |
| C Kennenlernen von AWS | II |
| D Kanban | II |

| | |
|--|------------|
| E Projektplan | III |
| E.1 EFS auf EC2-Instanz mounten | III |
| E.2 Von der Instanz aus abfragen, was ihr Status ist | V |
| F Für die Semesterarbeit verwendete Software | VI |

Abbildungsverzeichnis

| | | |
|---|--|-----|
| 1 | Eine Tour mit Budget Instanzen [Web10] | 1 |
| 2 | Internetkarte des Bundes <i>map.geo.admin.ch</i> . Hier ein Ausschnitt der Saane bei Kleinbösing, das Luftaufnahmen von 1946 mit heute vergleicht. | 1 |
| 3 | Im Viewer werden zurzeit Gebäude, Bäume, Seilbahnen, Namen und das Terrain dargestellt. Um aktuell zu bleiben, müssen diese 3D Daten regelmässig nachgeführt werden. | 3 |
| 4 | AWS Architektur Spot[10a] | 5 |
| 5 | Klassisches Kanban auf <i>github.com</i> | II |
| 6 | Projektplan. Die orangen Meilensteine wurden von der BFH vorgegeben. | III |

1 Einführung

Ich arbeite als GIS-Spezialist bei der swisstopo¹, dem Bundesamt für Landestopografie, in Wabern. Wir machen Karten. Unser Team macht Internetkarten - wie Google Maps², jedoch von der Schweiz für die Schweiz; und für alle anderen auch. Unsere Internetkarte, der Viewer, erfreut sich relativ grosser Beliebtheit und beinhaltet ca. 800 Themen wie Wanderwege, Solarkataster und Luftfahrthindernisse. Lieber Leser³, falls dir map.geo.admin.ch noch kein Begriff sein sollte, kann ich dir wärmstens empfehlen, darin zu schmökern. Es gibt viel zu entdecken und es ist gratis - ein Service Public.



Abbildung 2: Internetkarte des Bundes *map.geo.admin.ch*. Hier ein Ausschnitt der Saane bei Kleinbödingen, das Luftaufnahmen von 1946 mit heute vergleicht.

1.1 GIS - Geographical Information Systems

Wie erwähnt, arbeite ich als *GIS-Spezialist*. Wobei mir der Titel *Geo-Informatiker* besser gefällt: weil er die Begriffe *Geografie* und *Informatik* vereint. *Geografie* kommt aus dem Griechischen und bedeutet Erdbeschreibung [SE04, S. 14]. *Informatik* ist die Wissenschaft von der systematischen Darstellung, Speicherung, Verarbeitung und Übertragung von Informationen [10b].

GIS ist ein Akronym für Geographical Information Systems. Es bedeutet im engsten Sinn eine Ansammlung von Computerprogrammen, die zur Bearbeitung von Karten und Geodaten verwendet werden. Geodaten sind nichts weiter als Daten mit einem räumlichen Bezug⁴. In einem weiteren Sinn deckt der Begriff GIS ein ganzes Fachgebiet ab, das sich mit Karten und Geodaten auskennt. Es ist also nicht nur ein Werkzeug, sondern ein Fachgebiet, das Kenntnisse über Datensammlung, Speicherung, Analyse und Darstellung innerhalb von vielen verschiedenen Themen mit einem räumlichen Bezug abdeckt. Typische Geodaten sind digitale Karten, Inventare und Register von Parzellen, Umweltfaktoren, Grenzen, Entwicklung, Planung etc., die einen räumlichen Bezug haben und dadurch in einem geografischen Zusammenhang analysiert und dargestellt werden können [BJB06, S. 15].

Es ist zwar weit von der klassischen Geografie zu Zeiten Alexanders von Humboldt⁵ entfernt, aber es liegt auf der Hand, dass auch die Geografie Cloud Computing nutzt.

¹www.swisstopo.ch

²maps.google.com

³Im vorliegenden Dokument wird durchwegs der männliche Singular (Leser, Benutzer) als Ansprache verwendet. Diese Ansprache bezieht sich auf beide Geschlechter sowie gegebenenfalls mehrere Personen. Sie dient lediglich der leichteren Lesbarkeit der Semesterarbeit

⁴mit Koordinaten (Nord/Ost, x/y)

⁵Ein Forschungsreisender des 19. Jh. mit einem weit über Europa hinausreichenden Wirkungsfeld. Mehrjährige Forschungsreisen führten ihn nach Lateinamerika, USA und nach Zentralasien [Keh05]

2 Ausgangslage

Hier werden das Arbeitsumfeld und die Aufgaben beschrieben, aus denen sich die Problemstellung dieser Semesterarbeit ergeben hat.

2.1 swisstopo bei AWS

Es liegt auf der Hand, dass die swisstopo in ihrer Rolle als *Geoinformationszentrum* auf Cloud Computing setzt. Die swisstopo nutzt Cloud Computing mit AWS⁶ seit mehr als 10 Jahren für den Betrieb des Geoportal des Bundes.

"Mit der BGDI⁷ unter AWS können wir derzeit ca. eine Million Internetbenutzer pro Monat bedienen. Dank AWS können wir die zur Zuordnung neuer Server benötigte Zeit erheblich verkürzen und unseren Fokus auf echte Kundenanforderungen verstärken." [Chr20].

2.2 Publikation von Geodaten

Wie bereits erwähnt, können auf dem Viewer ca. 800 Themen wie Wanderwege, Solarkataster und Luftfahrthindernisse angesehen werden. Unser Team publiziert diese Daten. Der Nachführungszyklus wie auch der Aufwand zur Aufbereitung der Daten fürs Web sind unterschiedlich. Einige Daten werden manuell aufwändig aufbereitet, andere stündlich automatisch nachgeführt.

2.3 Web-Services

Nebst der Publikation der Daten ist unser Team für den Betrieb und der Weiterentwicklung der Web-Services und des Viewers verantwortlich. Der ganze Technologie Stack wurde schon länger nicht mehr grundlegend erneuert. Zurzeit wird die gesamte Architektur analysiert und überarbeitet, um eine gestaffelte Migration auf eine neue Lösung zu ermöglichen. Einige Rahmenbedingungen dieser zukünftigen Architektur sind bereits klar: Das Geoportal des Bundes wird weiterhin in der AWS Cloud betrieben werden, die Migration wird vor allem über Microservices gestaffelt erfolgen, diese Services werden als Docker Container laufen, Amazon Elastic Kubernetes Service wird die Orchestrierung der Container übernehmen; und für Continuous Integration wird AWS Codebuild/Pipeline zum Einsatz kommen.

⁶Amazon Web Services

⁷Bundesgeodateninfrastruktur: Viewer und andere Services

2.4 Exkurs 3D Daten

Die swisstopo erfasst und aktualisiert Daten mit einem räumlichen Bezug. Diese Geodaten sind die Basis für die Ableitung in eine Vielzahl von Produkten, wie die Landeskarten 1:25'000. Nebst Karten gibt es die Produktpalette der Landschaftsmodelle. Diese geben die Objekte der Landschaft im flexiblen Vektorformat wieder. Sie bestehen aus thematischen Ebenen (Bsp. Gebäude). Jede Ebene umfasst georeferenzierte Punkt-, Linien-, Flächen- oder 3D-Objekte. Jedes Objekt enthält Attribute und Beziehungen [10d].

Zu den Landschaftsmodellen gehören Produkte wie swissTLM3D und swissBuildings3D. Im Viewer wird eine Auswahl von Themen aus eben diesen Landschaftsmodellen dargestellt: Zurzeit Gebäude, Bäume, Seilbahnen, Namen und das Terrain. Vor wenigen Jahren wurden diese 3D Daten mit einem grossen Effort medienwirksam publiziert.

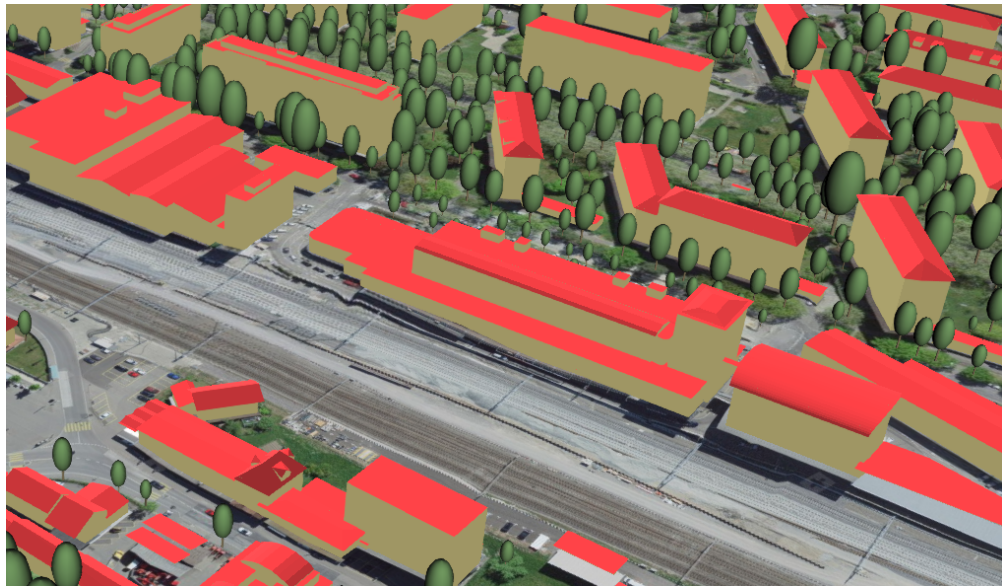


Abbildung 3: Im Viewer werden zurzeit Gebäude, Bäume, Seilbahnen, Namen und das Terrain dargestellt. Um aktuell zu bleiben, müssen diese 3D Daten regelmässig nachgeführt werden.

Seit dem Erfolg der Erstpublikation ist inzwischen etwas Zeit vergangen. Als die ersten Aktualisierungen der Daten anstanden, wurde den Beteiligten bewusst, dass sich diese nicht einfach so *auf Knopfdruck* realisieren lässt: Seit der Erstpublikation hat es personelle Wechsel gegeben und punkto Dokumentation und Automatisierungsgrad wurden Lücken identifiziert.

Es gibt immer gute Gründe für *technische Schulden*, wie in diesem Fall positive medienwirksame Reaktionen⁸. Aber früher oder später müssen diese abgebaut werden, weil es einen direkten Einfluss auf die Wartbarkeit des Produktes hat [10e].

⁸Wie Bsp. auf watson.ch oder Twitter [18]

3 Problemstellung

Es ist der swisstopo schon lange ein Anliegen, dass die Publikationsprozesse von Geodaten optimiert werden sollen. Wann immer möglich, soll der Automatisierungsgrad erhöht werden.

Besonders aufwändig erweist sich zurzeit die Publikation von 3D Daten. Die manuelle Publikation der 3D Daten benötigt eigene Tools, die auf einem performanten und somit teuren Rechner laufen müssen. Ausserdem erfordert die Bereitstellung einen hohen Koordinationsaufwand zwischen der Infrastruktur und der Entwicklung. Dabei passiert es, dass Mängel in den 3D Daten erst nach beendeter Webpublikation bemerkt werden; und der ganze Publikationsprozess muss wieder von vorne gestartet werden. Auch dem Hersteller der 3D Daten (dem Bereich Topografie) wäre es ein Anliegen, wenn er diese Daten selbst automatisch publizieren und prüfen könnte.

3.1 Ist-Zustand der 3D Datenpublikation

3.1.1 Prozess der Publikation

- ▶ Ausfindig machen des Zeitaufwandes pro Update bei uns und bei der IT
- ▶ Topo an Kogis, Kogis an IT, Kogis an Topo ...
- ▶ Bild des Prozesses, der Infra
- ▶ Manuelles Bereitstellen von EC2 Instanzen

3.2 Aufwand der Prozessierung

3.3 Technologie Stack

4 Vorgehen

In diesem Kapitel wird das Vorgehen beschrieben, wie die Arbeit geplant und realisiert wurde.

4.1 Arbeitsmethodik

Da die Arbeit nur von einem Autor umgesetzt wurde, musste bezüglich Arbeitsmethodik nur wenig definiert und koordiniert werden.

Um zu starten, hat dem Autor geholfen, dass der Experte Jörg Thomann bereits anfangs Sommer (16. Juli) einen ersten Besprechungstermin angesetzt hat. Um eine Diskussionsbasis für diesen Termin zu haben, musste sich der Autor ernsthaft mit der Semesterarbeit auseinandersetzen. Dazu hat er aus einer BFH Vorlage⁹ ein Gerüst der Arbeit erstellt und stichwortartig abgefüllt. Nach gründlicher Diskussion mit dem Experten wurde dieses Gerüst etwas angepasst und diente von da an als Basis für die Arbeit; wo sich die Kapitel zunehmend von losen Notizen zum finalen Zustand festigten.

Um die Übersicht nicht zu verlieren, wurde beschlossen, nach Kanban zu arbeiten, respektive ein Kanban Board zu verwenden. Weil für die Arbeit sowieso Repository auf github.com¹⁰ angelegt wurde, konnte dort auch gleich ein

⁹Herzlichen Dank an Andreas Habegger und Lukas Studer fürs Bereitstellen der Vorlage

¹⁰Projekt auf github.com

Kanban erstellt werden. Das Kanban Board wurde vor allem auch dafür genutzt, fortlaufend neue kleine Aufgaben aufzulisten und darauf zu achten, die Menge der angefangenen Arbeit zu begrenzen.

Dadurch konnte im Alltag (wenn gerade etwas Zeit zur Verfügung stand) eine solche Aufgabe genommen und erledigt werden. Es wurde nicht strikte nach Kanban gearbeitet, aber einige Prinzipien daraus wurden verwendet. Zum Beispiel wurde das erste Grundprinzip angewendet *"Beginne mit dem, was du gerade tust: In diesem Grundprinzip stecken zwei Dinge. Indem man mit dem beginnt, was man gerade tut, beendet man die aktuelle Arbeit, bevor etwas Neues begonnen wird. Genauso ist hier aber auch die Aussage enthalten, dass Kanban einfach eingeführt werden kann."*[10c]. Eine Momentaufnahme des Kanbans befindet sich im Anhang D.

4.2 Projektplan

Um den übergeordneten Rahmen der Arbeit nicht aus dem Fokus zu verlieren und um sich bewusst zu machen, wieviel Zeit generell bis zur Abgabe zur Verfügung steht, war es hilfreich, einen Projektplan mit Meilensteinen konsultieren zu können. Durch die Berner Fachhochschule waren die grundlegenden Meilensteine bereits vorgegeben. Der Projektplan befindet sich im Anhang E.

4.3 Vorarbeiten

4.3.1 AWS Account

- Fit werden mit AWS

4.3.2 AWS Spot Instances

- Herausfinden wie mit AWS Spot Instanzen gearbeitet werden kann

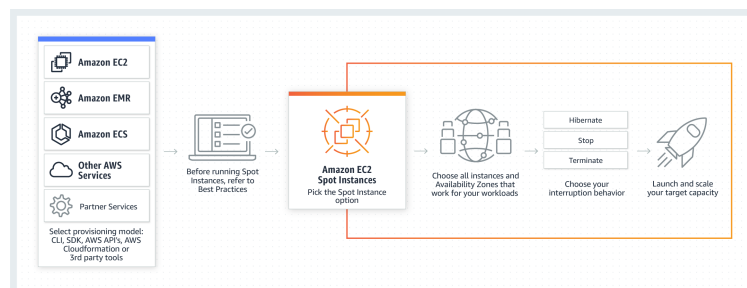


Abbildung 4: AWS Architektur Spot[10a]

4.3.3 Kubernetes testing

- Anfangs mit Minicube
- Eigenem AWS account

5 Use Case

- ▶ Automatisierung der Datenprozessierung
- ▶ Testumgebung für TopoKo
- ▶ Automatische Prozessierung via SPOT Instanzen

5.1 Beschreibung der Technischen Komponenten

5.1.1 SPOT Instanzen

5.2 Abgrenzung

Wie in der Projektskizze beschrieben wird der Fokus auf bereits dockerisierte Prozessierungstools gelegt.

Der Fokus liegt vor allem auf der Automatisierung von SPOT Instanzen im Vergleich zum Manuellen Bereitstellen.

6 Architektur

Ausfindigmachen von verschiedenen Architekturen

6.1 Bewertungskriterien

- ▶ Muss in AWS Cloud
- ▶ Zugang zu EFS
- ▶ Automatisierbar sein

6.2 Variante 1: Spot

Evtl. lässt sich eine Lösung finden ohne EKS zu verwenden

6.3 Variante 2: Spot mit EKS

Spot mit EKS

6.4 Variante 3: AWS Batch

Für das ist es da

7 Prototyp

7.1 Realisierung

8 Evaluation

8.1 Erfahrungen

8.2 Kritische Punkte

9 Zusammenfassung

10 Ausblick

Literaturverzeichnis

- [10a] (2010). Amazon EC2-Spot-Instances, Adresse: <https://aws.amazon.com/de/ec2/spot/> (besucht am 08.07.2010).
- [10b] (2010). Informatik, Adresse: <https://de.wikipedia.org/wiki/Informatik> (besucht am 16.07.2010).
- [10c] (2010). Kanban (Softwareentwicklung), Adresse: [https://de.wikipedia.org/wiki/Kanban_\(Softwareentwicklung\)](https://de.wikipedia.org/wiki/Kanban_(Softwareentwicklung)) (besucht am 19.07.2010).
- [10d] (2010). swisstopo Onlineshop, Adresse: <https://shop.swisstopo.admin.ch> (besucht am 17.07.2020).
- [10e] (2010). Technische Schulden, Adresse: https://de.wikipedia.org/wiki/Technische_Schulden (besucht am 18.07.2010).
- [18] (2018). Diese 3D-Landkarte gibt dir einen völlig neuen Blick auf die Schweiz, Adresse: <https://www.watson.ch/digital/schweiz/674619561-diese-3d-landkarte-gibt-dir-einen-voellig-neuen-blick-auf-die-schweiz> (besucht am 18.07.2010).
- [AD19] J. Arundel und J. Domingus, *Cloud Native DevOps with Kubernetes*. O'RELLY, 2019.
- [BJB06] T. Balstrøm, O. Jacobi und L. Bodum, *GIS og geodata*. Forlaget GIS og Geodata, 2006.
- [Chr20] H. Christ. (9. Juli 2020). swisstopo Fallstudie, Adresse: <https://aws.amazon.com/de/solutions/case-studies/swisstopo-hpc/>.
- [Doe00] J. Doe, *The Book without Title*. Dummy Publisher, 2100.
- [Keh05] D. Kehlmann, *Die Vermessung der Welt*. Rowohlt Verlag GmbH, 23. Sep. 2005, 304 S., ISBN: 3498035282. Adresse: https://www.ebook.de/de/product/3563778/daniel_kehlmann_die_vermessung_der_welt.html.
- [SE04] M.-H. Schertenleib und H. Egli-Broz, *Grundlagen Geografie: Aufgaben des Fachs, Erde als Himmelskörper und Kartografie : Lerntext, Aufgaben mit Lösungen und Kurztheorie*. Zürich: Compendio Bildungsmedien, 2004, ISBN: 3715591714.
- [Web10] Webpage. (2010). Keep calm and make your choice..., Adresse: <https://www.grenke-40-one.de/excursions-detail/> (besucht am 11.07.2010).

A Fachbegriffe und Abkürzungen

BGDI Bundesgeodateninfrastruktur

PaaS Plattform as a Service

21 22

31 32

41 42

51 52

61 62

71 72

81 82

91 92

101 102

B Konfigurationen

C Kennenlernen von AWS

D Kanban

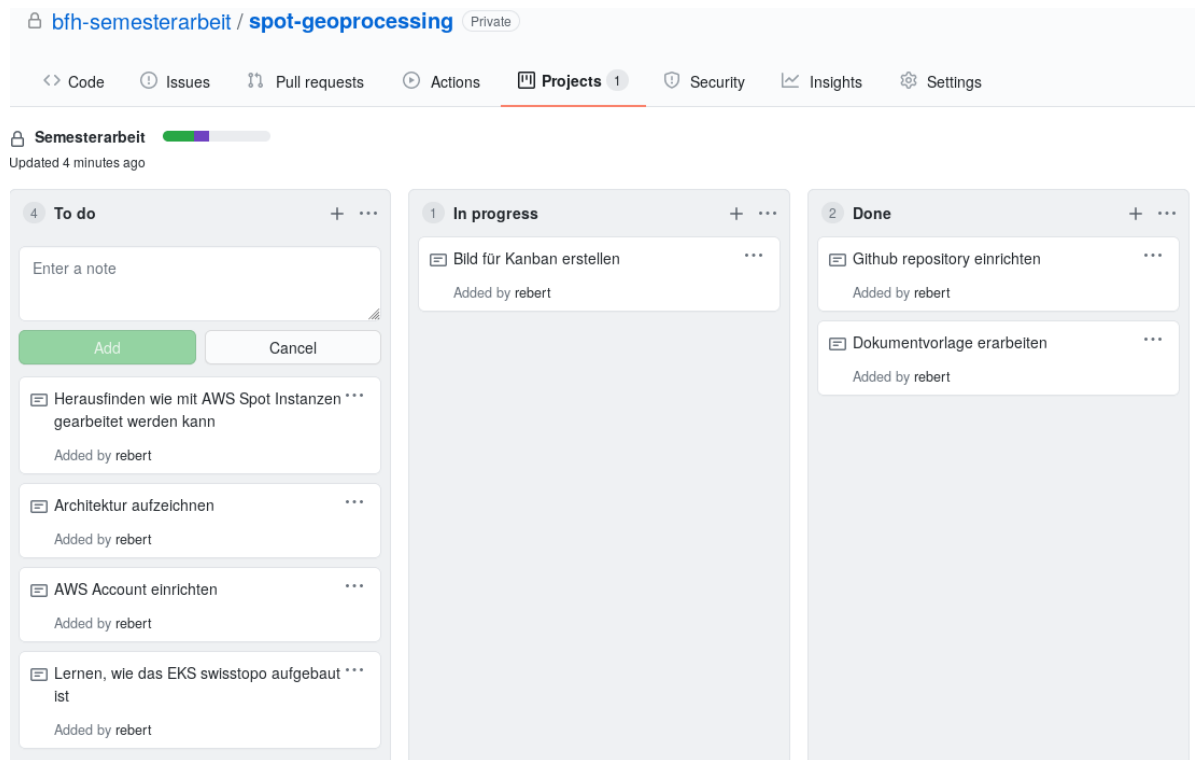


Abbildung 5: Klassisches Kanban auf *github.com*

E Projektplan

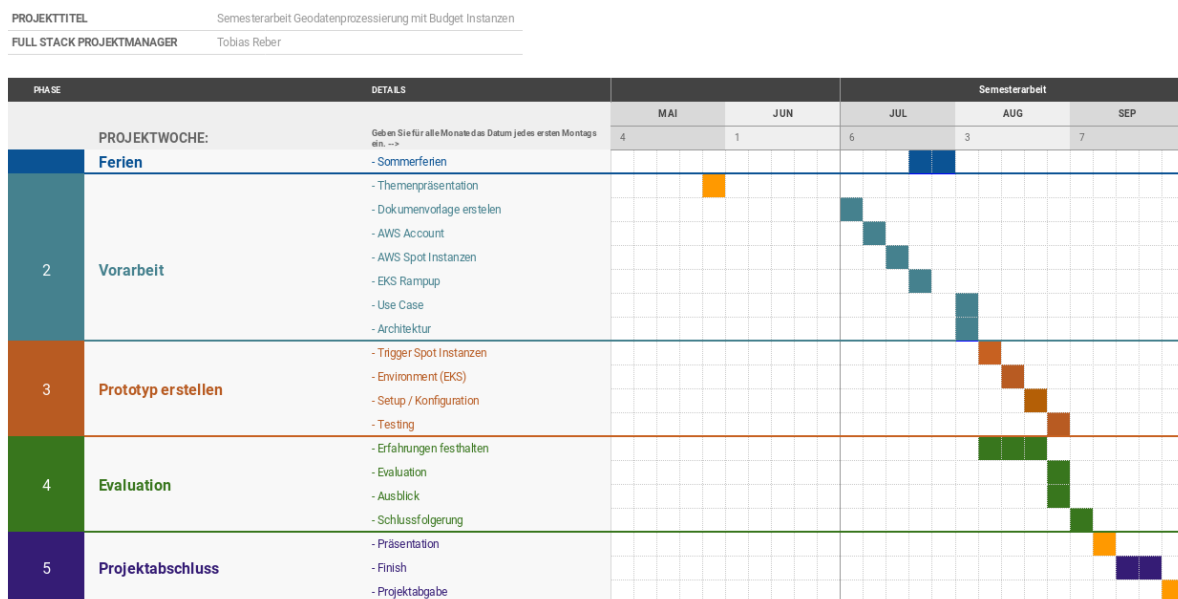


Abbildung 6: Projektplan. Die orangenen Meilensteine wurden von der BFH vorgegeben.

E.1 EFS auf EC2-Instanz mounten

Anhand einer Anleitung, einem sogenannten Walkthrough wurde via AWS CLI (AWS Command Line Interface: ein kommandozeilenorientiertes Werkzeug) ein EFS an eine EC2-Instanz gemountet und die 3D Daten wurden dorthin kopiert.

Listing 1: EFS auf EC2-Instanz mounten

```
#!/bin/bash
# from https://docs.aws.amazon.com/efs/latest/ug/wt1-create-ec2-resources.html

# security group 4 ec2
aws ec2 create-security-group \
--region eu-west-1 \
--group-name efs-dataprocessing1-ec2-sg \
--description "Amazon EFS dataprocessing 1, SG for EC2 instance" \
--vpc-id vpc-87ad55fe

# security group 4 efs
aws ec2 create-security-group \
--region eu-west-1 \
--group-name efs-dataprocessing1-mt-sg \
--description "Amazon EFS dataprocessing 1, SG for mount target" \
--vpc-id vpc-87ad55fe

# access to ec2 instance group from everywhere
aws ec2 authorize-security-group-ingress \
--group-id sg-098669727548dcedd \
--protocol tcp \
--port 22 \
--cidr 0.0.0.0/0 \
--region eu-west-1

# describe security group
aws ec2 describe-security-groups \
```

```
--region eu-west-1 \
--group-id sg-098669727548dcedd

# access to efs storage from ec2 group
aws ec2 authorize-security-group-ingress \
--group-id sg-02778494bc39601d4 \
--protocol tcp \
--port 2049 \
--source-group sg-098669727548dcedd \
--region eu-west-1

# get subnet id
aws ec2 describe-subnets \
--region eu-west-1 \
--filters "Name=vpc-id,Values=vpc-87ad55fe"

# run ec2 instance
aws ec2 run-instances \
--image-id ami-047bb4163c506cd98 \
--count 1 \
--instance-type t2.micro \
--associate-public-ip-address \
--key-name bfh_root.pem \
--security-group-ids sg-098669727548dcedd \
--subnet-id subnet-f66512ac \
--region eu-west-1

aws ec2 describe-instances \
--instance-ids i-09cb26ed64cdde683 \
--region eu-west-1

# ec2-54-75-53-151.eu-west-1.compute.amazonaws.com
# EFS =====
aws efs create-file-system \
--creation-token FileSystemForDataprocessing1 \
--tags Key=Name,Value=Dataprocessing1 \
--region eu-west-1

# create mount target
aws efs create-mount-target \
--file-system-id fs-5aceld90 \
--subnet-id subnet-f66512ac \
--security-group sg-02778494bc39601d4 \
--region eu-west-1

# On instance amazon linux (ec2-user)
sudo yum -y update
sudo reboot # dont know why... like windows
sudo yum -y install nfs-utils

# On instance ubuntu (ubuntu)
sudo apt-get update
sudo apt-get install nfs-common

mkdir ~/efs-mount-point
cd ~/efs-mount-point
sudo chmod go+rw .

sudo mount -t nfs \
-o nfsvers=4.1,rsz=1048576,wsz=1048576,hard,timeo=600,retrans=2,noresvport
fs-5aceld90.efs.eu-west-1.amazonaws.com:/ \
~/efs-mount-point

# CP from home to ec2 instance
scp -i bfh_root.pem \
/media/saibot/vortrag/bfh/buildings/*.tar \
ec2-user@54.75.53.151:/home/ec2-user/efs-mount-point/buildings/
```

E.2 Von der Instanz aus abfragen, was ihr Status ist

Von der Instanz aus kann via REST API Metadaten der Instanz abgefragt werden. Bezüglich Determinierung einer Spot-Instanz kann der Zustand *none*, *hibernate*, *stop* oder *terminate* sein. *none*, wenn nichts ansteht.

Listing 2: Status der Instanz abfragen

```
TOKEN=$(curl -X PUT "http://169.254.169.254/latest/api/token" -H  
"X-aws-ec2-metadata-token-ttl-seconds: 21600") && curl -H "X-aws-ec2-metadata-token: $TOKEN"  
http://169.254.169.254/latest/meta-data/instance-action
```

F Für die Semesterarbeit verwendete Software

- ▶ JabRef: Verwaltung des Literaturverzeichnisses (BibTeX)
- ▶ Gummi: LaTeX Editor
- ▶ AWS CLI: Für das Bereitstellen der AWS Infrastruktur
- ▶ jq: Für das Filtern von JSON (vor allem von AWS CLI Antworten)
- ▶ git: Versionsmanagement der Textdateien
- ▶ Google Spreadsheet: Für den Projektplan